

# Implement Generative AI Tools in Analytics

Applied Project

By

Sike Shou

Spring, 2023

Master of Science in Management and Systems

at the

Division of Programs in Business

School of Professional Studies

New York University

---

## **Acknowledgements**

I sincerely thank Dr.Fortino for initiating this project and for his tireless mentorship during the project. I also want to express thankfulness to Roksolana Sheverack for her contributions as the sponsor of this project. Besides, I want to express my gratitude to 8 data analysts who are willing to take the interviews and provide insights: Yabo Dong, Eric Chen, Felix Hui, Nachuan Cao, Glen Lyon, Gurneet Badi, Kate Yayla and Qitong Zhong. I also want to thank all the instructors in the Management and Systems program who I have taken courses with and learned a great deal.

## **Abstract**

**Contribution:** Initially, the study was expected to find real-world examples of how data analysts use ChatGPT in their work. But the use of ChatGPT is forbidden in many companies because of privacy issues and data security. The study then turned to finding applications of ChatGPT to improve data analysts' work without directly utilizing company data and to improve the teaching quality of the MASY program with ChatGPT.

**Background:** The main purpose of the study is to find applications of ChatGPT to improve the productivity of data analysts and improve the teaching quality of the MASY program. ChatGPT has surprised the world with its ability to perform various tasks, including writing Python/SQL code and doing large-scale data analysis. Many people argue data analysts may be replaced by ChatGPT. The paper aims to find the true situation of how data analysts use and view this tool. In addition, find out how can they use it efficiently to increase their productivity. For the MASY program with many database and analytics courses, the study also aims to find ways to integrate ChatGPT in teaching.

**Research Question:** How is ChatGPT used in data analysts' work and what are their views on this tool? What are the daily responsibilities of data analysts and how can ChatGPT boost their productivity without data leakage? How can ChatGPT be integrated in the teaching of analytics courses in the MASY program?

**Methodology:** Both quantitative and qualitative approaches were employed. Quantitative results were obtained from interviews with working data analysts. Qualitative data were obtained from the interviews, and from the research literature and practices on ChatGPT.

**Findings:** ChatGPT can perform many tasks of a data analyst but companies are reluctant to use it for analysis purposes because of privacy issues and data security. In the future, it is possible that many companies will develop their own local large language models to utilize their full potential. In addition, ChatGPT can make the preparation of teaching materials easier for professors and the studying process more efficient for students.

## Table of Contents

<b>Introduction</b> .....	1
<b>Background</b> .....	1
Background and history of ChatGPT .....	1
Technical Principles of ChatGPT .....	3
Limitations of ChatGPT.....	3
<b>Research process and methods</b> .....	4
Methodology Used in This Paper .....	4
How is ChatGPT used in this industry .....	4
<b>Analysis of the interviews of data analysts</b> .....	8
Data analysts' daily responsibilities .....	8
Data analysts' applications of ChatGPT in their work .....	10
Their views of ChatGPT's limitations.....	12
How data analysts think ChatGPT will impact their future work .....	14
<b>Examples of applications of ChatGPT in analytics work</b> .....	16
A list of useful analytics tools related to ChatGPT .....	16
Use cases of ChatGPT on data analysts' work .....	19
<b>Use Cases for Analytics Courses using ChatGPT</b> .....	27
<b>Conclusion</b> .....	42
<b>Reference</b> .....	44

## Introduction

In today's world, numerous efficient tools have emerged in every generation. In the 90s, the rising popularity of IT systems created new industries and new divisions of jobs. Generative AI is also a rising trend in recent months. Its applications varied from text to pictures, text to videos, text to PowerPoints, text to music and text to code. Although they are perceived as toys for the public, their content creation abilities are endorsed by industry professionals in advertising, technology, design, etc. Will generative AI also bring changes in industries and new jobs? How can professionals use them effectively in work?

This paper will look into Generative AI's impact on a specific position, data analyst. In particular, ChatGPT is selected as the main object for research because it has high potential in a wide range of tasks. Additionally, it will include some real-world practices of ChatGPT to illustrate how can data analysts use this tool.

## Background

### Background and history of ChatGPT

ChatGPT is the short name for Chat Generative Pre-Trained Transformer. It is an AI-powered chatbox that is able to answer the prompts in human language and generate the outputs in human languages. It can perform various tasks, such as writing coherent essays, performing quantitative analysis, teaching and tutoring and generating codes according to descriptions. It is developed and introduced to the public by the institute OpenAI. It has drawn the attention of millions of people since its introduction and quickly became a hot topic worldwide.

OpenAI is a private research institute that developed ChatGPT. OpenAI has a partnership with Microsoft. Microsoft invested in OpenAI in 2019, 2021 and signed the extended partnership contract with OpenAI in Jan 2023 (Microsoft Corporate Blogs, 2023). Microsoft plans to integrate AI features into various Microsoft products, including Microsoft 365 suits, Microsoft Teams and new Bing.

ChatGPT is a fine-tuned version of GPT 3.5. GPT 3.5 is a large language model Open AI released on March 2022. GPT 3.5 is an updated version of GPT 3, including features such as application programming interfaces, APIs on OpenAI's websites to be more easily adapted into other software or models. Before ChatGPT, OpenAI released another fine-tuned version of GPT 3.5, which is called InstructGPT on January 2022. But it was not released to the public. InstructGPT used the similar fine-tuning technique as ChatGPT called reinforcement learning from human feedback (RLHF) (Heaven, 2023). This means that a large language model can spit out anything, but the researcher will tune it and teach it what kinds of answers are better to provide to humans (Heaven,2023). That is to say, the technical capacity of these models is around the same. However, Open AI has made a more human-friendly interface for ChatGPT, including a chat interface, talk dialogues, and understanding the context. The biggest concern for developers at OpenAI during the beta stage of ChatGPT is its model's factuality because the model likes to fabricate things (Heaven, 2023). But it is a common problem in GPT 3.5, therefore, developers are pleased to say ChatGPT outperforms InstructGPT and other models in terms of safety issues and factuality issues (Heaven,2023).

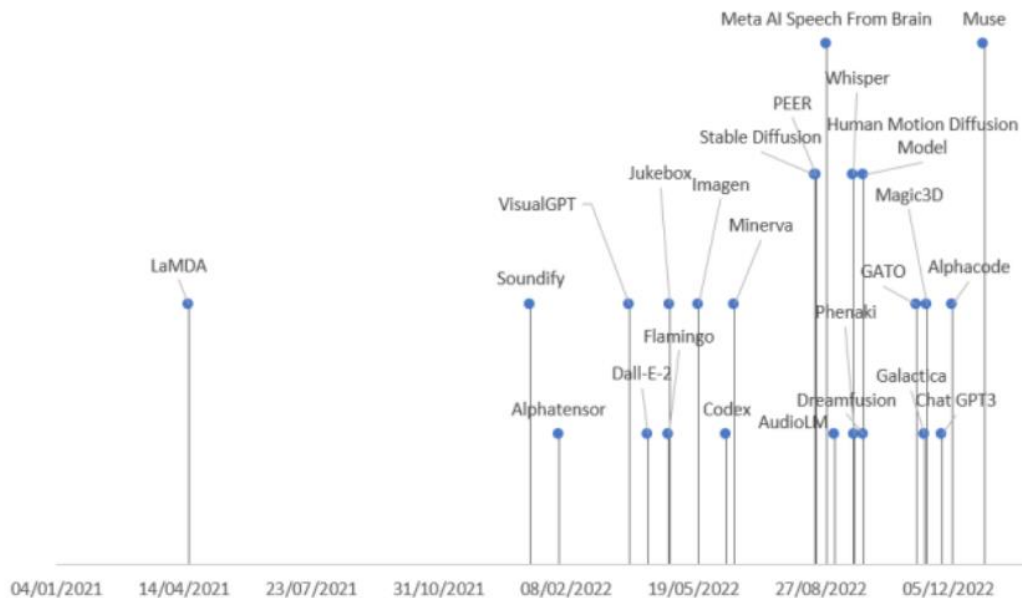
MAE (Multimodal Autoencoder) was developed by researchers at OpenAI in 2019. It is a generative model that can learn to generate images from textual descriptions and vice versa. MAE uses a multimodal autoencoder architecture that combines both text and image modalities (Zhou et.al,2021). Before ChatGPT, they launched DALL-E in 2021, which is also a model leveraging machine learning techniques and NLP principles. DALL-E uses contrastive language-image pretraining (CLIP) for capturing semantic association between image-text

pairs and GLIDE diffusion model for text-conditional image synthesis (Zhou et.al,2021). DALL-E is open to the public, which helps Open AI gathers valuable feedback and experience to be used in the development of ChatGPT.

DALL-E and ChatGPT are both Pretrained Foundation Models (PFMs). It is a type of machine learning model that has been pre-trained on large amounts of data to learn general features and patterns. Pre-trained models can be used to generate text, images and graphs, making them the ideal tool to complete tasks like machine translation, question-answering systems, topic modeling, sentiment analysis, etc. (Zhou et.al,2021). It is first used in the computer vision field then natural language processing. Pretrained Foundation Models (PFMs) itself can serve as the foundation for a lot of downstream applications without requiring training on task-specific datasets. With this advantage, researchers and creators can save time and resources if they do not want to turn to fine-tuned stage. Pre-trained models can also be fine-tuned further to achieve a specific function. Fine-tuned is a machine learning technique that leverages the pre-trained model to smaller, task-specific datasets to improve its performance on that particular task.

BERT (Bidirectional Encoder Representations from Transformers) was developed by researchers at Google AI Language in 2018. It was pre-trained on a large corpus of text data using a transformer-based architecture and achieved state-of-the-art performance on a wide range of natural language processing tasks. There are two applications of PFMs, BERT is in the category of PFMs with fine-tuning, while the other one is PFMs with zero/few-shot prompts like GPT (Zhou et.al,2021). GPT 3.5 uses an autoregressive decoder to use the first few words to predict the other words and solve the downstream tasks using fine-tuning (Zhou et.al,2021). BERT is mainly used for processing text tasks, such as recognizing the context in language and classification tasks, while GPT is more suitable for language generation tasks like question answering and writing paragraphs (Shahriar & Hayawi, 2023).

Generative AI refers to artificial intelligence that can generate new content instead of only analyzing the data. Here is a picture to demonstrate the development history of various Generative AI models. We can see the boost of the development of various generative AI in recent years.



(Gozalo-Brizuela & Garrido-Merch´an, 2023)

## Technical Principles of ChatGPT

ChatGPT marks a significant milestone in the machine learning field. Machine learning is a subfield of Artificial Intelligence where machines learn from experience with no explicit human interference. As computer systems advance, data availability increases and algorithmic improvements, deep learning emerges as a high-performing technique (Shahriar & Hayawi, 2023). It utilizes a large artificial neural network and requires more data for training. Large language models are a family of deep learning models which has strong capability in analyzing text data (Shahriar & Hayawi, 2023).

One of the most convenient and attracting feature of ChatGPT is that they can understand people's language and its context. The more information you put into it, the better defined the purposes. How does a model understand people's needs by just stating the human language but not the computer language 0 and 1? GPT simulates the learning process of the human brain. In the human brain, numerous neutrals are linked as complicated web syntax. Language is an input that intrigues a certain signal in the neutral web syntax (Space, 2023). That process is developed at human young age by observing and listening other people's conversation. Human developed a certain inclined intention to analysis the words and confirmed their judgement by communicating with others and observing others' behaviors. Machine learning model like ChatGPT, tries to simulate this process by building simulated neutral systems in computers (Space, 2023). Instead of providing an environment to observe others' conversation, engineers put massive amount of data on the internet up to 2021 and use it as materials to train ChatGPT for one year (Shahriar & Hayawi, 2023). The researchers spend half a year to analysis the response from ChatGPT to make sure it follows human logic and optimize its answer, which is called the reinforcement learning.

ChatGPT is trained on a massive foundation of data in the form of text similar to human language. It is in the category of Large Language Models (llms). ChatGPT is able to understand prompts in natural language (human language) and generate output in natural language by leveraging its extensive data storage and efficient design to understand and interpret user requests (Lund and Wang, 2023). Unlike traditional chatboxes who can only perform simple and pre-designed tasks, ChatGPT can effectively leverage its highly complex neural network to understand users' requests and perform complex tasks such as writing a business plan, teaching concepts, writing and explaining codes, etc.

ChatGPT has a multi-layer transformer network, belonging to a type of deep learning architecture, which gave ChatGPT's ability to process natural language by processing input sentences using its internal knowledge and then generating a relevant response (Majumder, 2022). It can understand the context of the conversation and generate a response in that context (Majumder, 2022). This makes the conversation with ChatGPT more like a real human because real humans can understand the different meanings of the same sentence in different contexts and give the corresponding result.

ChatGPT 3.5 was trained for one year using a large amount of data. The development of the model uses a machine-learning technique called unsupervised pretraining. It is a process where the model learns the pattern on its own through a large amount of unlabeled data without human interference. The next process is supervised fine-tuning. In this process, the model was trained on smaller sets but labeled data. The creators of the model give more detailed refinement on the products of the model. The process lasts around half a year. After ChatGPT is opened to the public, the prompts and data input by millions of users is also selected and used to refine the machine learning model.

## Limitations of ChatGPT

ChatGPT can fulfill many tasks with the user's creativity and the right prompt (meaning the

input message from the user to the chatbox for it to understand the order). However, some people also point out many problems exist in ChatGPT. One thing is the factuality issue. Even if you can use it as a search engine, many times ChatGPT will give fictional answers. In addition, its training source is up to 2021. Therefore, when you ask for information beyond 2021, it is more likely to give you fictional answers. Another thing is the model-biased issues. ChatGPT develops its capacity based on current training materials from all over the Internet. Racist, sexist, and otherwise abusive language may be in the training data (Hao, 2022). In addition, rich countries may have easier access and more presence on the Internet, making the response of ChatGPT more homogenized in the practice of rich countries (Hao, 2022). The last but most important thing is privacy issues. ChatGPT is a free research object from Open AI. At the beginning of using ChatGPT, users sign an agreement with Open AI that any data on the ChatGPT can be collected and used to improve the model. It is not recommended to put proprietorial data and private data from the organization on it because of the data leakage issues which may lead to business loss.

## Research process and methods

### Methodology Used in This Paper

Both quantitative and qualitative approaches were followed. Quantitative results were obtained from interviews with data analysts. Qualitative data were obtained from interviews, research papers and practice on ChatGPT. The data collected for the paper consisted of both primary and secondary data. Secondary information sources include collecting and researching peer-reviewed research papers, website articles, and posts on the website and social media. It was conducted at the beginning of the research to get background information and continued after the collection of primary sources. A well-structured question list was prepared and conducted with 7 data analysts in the form of a one-to-one Zoom meeting. The interview was conducted from February 2023 to March 2023. The data collected through this primary source was analyzed using keywords and frequencies, and topic modeling. Both quantitative and qualitative data were obtained through the interviews.

At the beginning of the project, I research on my own to understand what ChatGPT is, its background, and practiced ChatGPT to gain a preliminary view of the research object. After that, I started interviewing data analysts. In particular, I selected entry-level data analysts. The main reason is that they deal with data directly, which can provide great insights into what data analysts' daily work is like, what the tools they use and how to achieve their function with better efficiency. Senior managers rely more on the domain knowledge of the industry to provide strategic advice. It is a field ChatGPT may not help much.

The interview materials provide the basis for further research. The collection of data analysts' daily functions inspires me to think of how to achieve these with ChatGPT. Although ChatGPT is only a testing version, there have been various plugin available on the Internet, some of which is developed by Open AI and some of which is developed by ChatGPT users. It can give me a great hand on how to accomplish the certain function.

Finally, based on the interview transcripts collected and the syllabus of MASY program, create use cases using ChatGPT. Perfect the prompts to achieve a certain function. Make a screenshot of prompts and results to record in the paper.

### How is ChatGPT used in this industry

With the popularity of ChatGPT and other large language models, some people argue that this technology can improve the efficiency of the work and generate more robot-related positions,



while some are worried about it will replace the human. According to Zarifhonarvar (2023), ChatGPT and other technology alike can be attributed to labor-saving technology (LSTs). It can be historically traced to the industrial revolution and the invention of electricity and information technology in the 20<sup>th</sup> century. The main characteristics of LSTs include the labor-saving effect, the size of which depends on the skill, industry and country (Montobbio et al. 2022, cited in Zarifhonarvar, 2023).

As for the impact of ChatGPT in labor market, it is possible to increase productivity, which leads to higher economic growth and higher wages. It is also possible to replace some repetitive work resulting higher unemployment rate.

Zarifhonarvar (2023) performs a text-mining approach that extracts 20 most-mentioned keywords from the tasks of each position under the International standard classification of occupations (ISCO) and compares them with the tasks that ChatGPT can potentially automate. Based on the number of tasks that ChatGPT can automate, each position is classified into ‘Full Impact’, ‘Partial Impact’ and ‘No Impact’. Here is a picture showing the count of the number of positions in each category in ISCO’s impact level.

**Table 2 - ISCO Classification and ChatGPT Impact (Major Label)**

Major	Full Impact	No Impact	Partial Impact
<b>Professionals</b>	95	22	9
<b>Technicians and Associate Professionals</b>	60	34	16
<b>Managers</b>	20	21	6
<b>Clerical, service and sales workers</b>	8	20	14
<b>Craft and Related Trades Workers</b>	8	33	45
<b>Plant and Machine Operators and Assemblers</b>	5	19	34
<b>Skilled agricultural and trades workers</b>	4	24	3
<b>Services And Sales Workers</b>	3	37	18
<b>Armed Forces Occupations</b>	0	0	10
<b>Elementary Occupations</b>	0	16	35

(Zarifhonarvar ,2023)

For the position ‘Data Analyst’, its ISCO code is 2511.2, which is fall below the category of ‘Professionals’. We can see that it is the category that may be influenced most by ChatGPT. Therefore, data analyst needs to learn how to work with ChatGPT to increase their productivity.

Based on this method, I have collected the main responsibilities for data analyst from ISCO definition and asked ChatGPT which tasks can be assisted by ChatGPT. I categorize the ChatGPT’s impact on the tasks of data analysts to full impact, partial impact and no impact. In addition, I asked ChatGPT to mark the tasks it can complete but may violate data privacy.

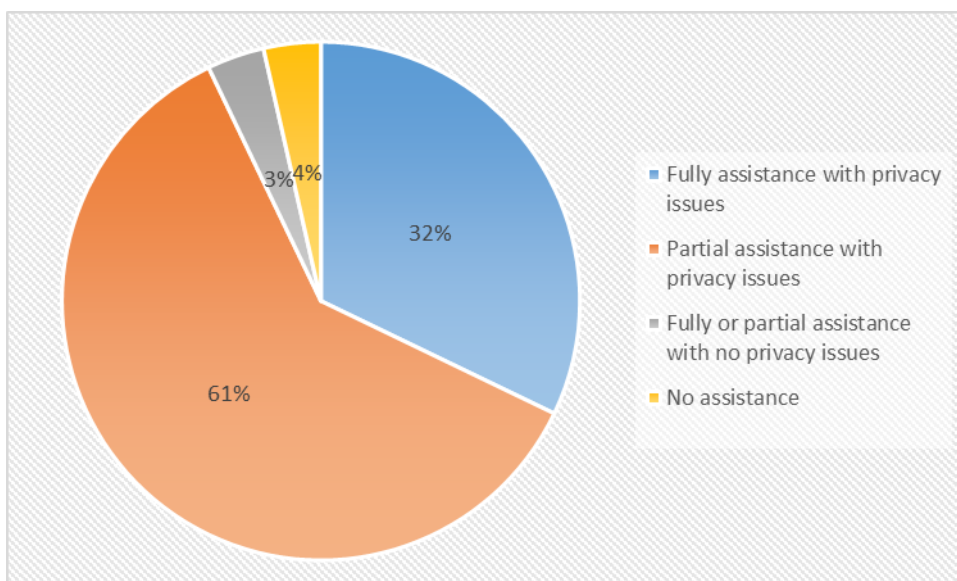
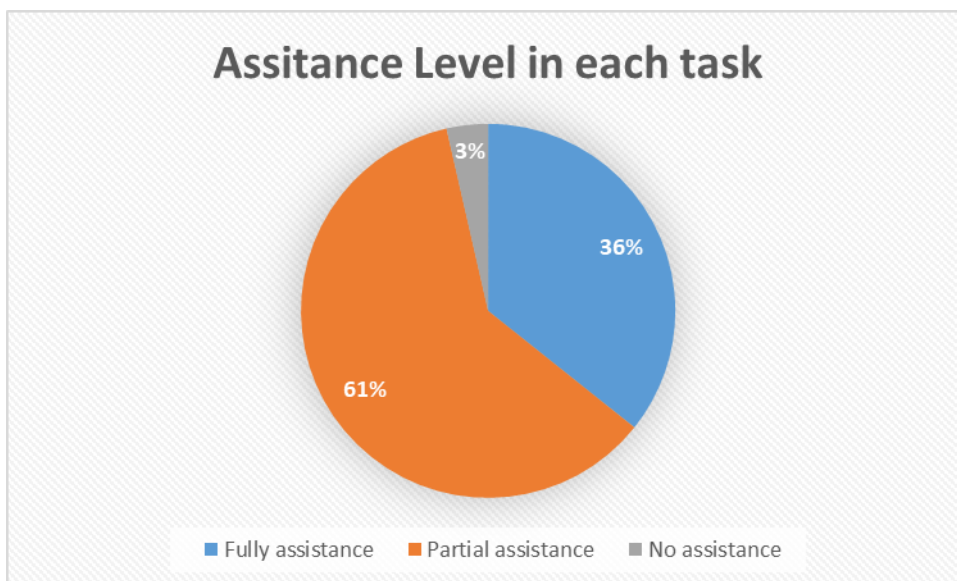
SH Hi chatgpt, I want you to assess the following task and mark each one as fully assistance(means you can perform this task automatelly with little or nearly no human interfere), partially assistance (means you can do part of task with human hand-to-hand mentoring), no assistance( you can perform little duty of this task). Some companies are not comfortable to directly analyze data on ChatGPT because of privacy issues. So for each assitance, if there is privacy concerns, please mark and show it. List: Define Data Quality

Here are the answers from ChatGPT:

1. Define Data Quality Criteria - fully assistance

2. Manage Data - fully assistance, but potential privacy concerns depending on the specific data and how it's managed
3. Normalise Data - fully assistance, but potential privacy concerns depending on the specific data and how it's normalized
4. Handle Data Samples - fully assistance, but potential privacy concerns depending on the specific data and how it's handled
5. Unstructured Data - partially assistance, but potential privacy concerns depending on the specific data and how it's analyzed
6. Execute Analytical Mathematical Calculations - fully assistance, but potential privacy concerns depending on the specific data and how it's analyzed
7. Integrate ICT Data - partially assistance, but potential privacy concerns depending on the specific data and how it's integrated
8. Establish Data Processes - partially assistance, but potential privacy concerns depending on the specific data and how the processes are established
9. Business Intelligence - partially assistance, but potential privacy concerns depending on the specific data and how it's used for business purposes
10. Information Confidentiality - fully assistance, but potential privacy concerns depending on the specific data and how it's handled
11. Documentation Types - partially assistance, but potential privacy concerns depending on the specific data and how it's documented
12. Perform Data Cleansing - fully assistance, but potential privacy concerns depending on the specific data and how it's cleansed
13. Analyse Big Data - partially assistance, but potential privacy concerns depending on the specific data and how it's analyzed
14. Information Extraction - partially assistance, but potential privacy concerns depending on the specific data and how it's extracted
15. Resource Description Framework Query Language - no assistance
16. Information Structure - partially assistance, but potential privacy concerns depending on the specific data and how it's structured
17. Query Languages - partially assistance, but potential privacy concerns depending on the specific data and how it's queried
18. Implement Data Quality Processes - partially assistance, but potential privacy concerns depending on the specific data and how the processes are implemented
19. Data Models - partially assistance, but potential privacy concerns depending on the specific data and how the models are used
20. Information Categorisation - partially assistance, but potential privacy concerns depending on the specific data and how it's categorized
21. Apply Statistical Analysis Techniques - fully assistance, but potential privacy concerns depending on the specific data and how it's analyzed
22. Collect ICT Data - partially assistance, but potential privacy concerns depending on the specific data and how it's collected

- 23. Data Mining - partially assistance, but potential privacy concerns depending on the specific data and how it's mined
- 24. Perform Data Mining - partially assistance, but potential privacy concerns depending on the specific data and how it's mined
- 25. Interpret Current Data - fully assistance, but potential privacy concerns depending on the specific data and how it's interpreted
- 26. Visual Presentation Techniques - partially assistance, but potential privacy concerns depending on the specific data and how it's presented
- 27. Data Quality Assessment - partially assistance, but potential privacy concerns depending on the specific data and how it's assessed
- 28. Statistics - fully assistance, but potential privacy concerns depending on the specific data and how it's analyzed.



We can see that many tasks related to data handling and cleaning, statistics and calculations can be assisted by ChatGPT. Because Open AI collect users' input as training materials for their next model generation, many companies refuse to directly use it for data analytics. But

ChatGPT can assist in writing queries, generating sample data and providing insights such as which Python library to use. These can also boost data analysts' productivity. For certain objective tasks, such as defining the data quality criteria or business intelligence, its' output depends on the individual company, ChatGPT may not be very useful in these areas.

## Analysis of the interviews of data analysts

### Data analysts' daily responsibilities

Here is a list of daily responsibilities and time allocation for the different interviewee.

1. Communicate with the client to figure out the business needs (30%), pull the data from the data pool, clean, analyze and visualize it (40%), make presentations to the client (30%)
2. Write SQL code or construct relational databases (time depends on the project)
3. Extract the data from data pool to Excel (30%)  
Build new auto visualization software (70%)
4. Monitor the accounts, read the auto-generated reports about which accounts are suspicious (20%), find these accounts and review their activity (60%), improve the monitoring system (20%)
5. Build data pipeline using API from the website to draw data by writing Python code (80%)  
Visualize the data from the company's database and build dashboards (20%)
6. Research the client's company, decide the risk level and generate reports (use some Power BI)

In order to find things in common in their talk, I perform a keyword extraction on a file that integrates their conversation about their work duties. Here are the top 20 keywords and their frequencies:

('data', 83), ('need', 45), ('know', 45), ('time', 36), ('usually', 29), ('dashboard', 20), ('see', 20), ('day', 20), ('mean', 20), ('cause', 20), ('work', 19), ('would', 19), ('project', 19), ('much', 18), ('us', 18), ('team', 17), ('problem', 17), ('people', 17), ('system', 16), ('na', 16)



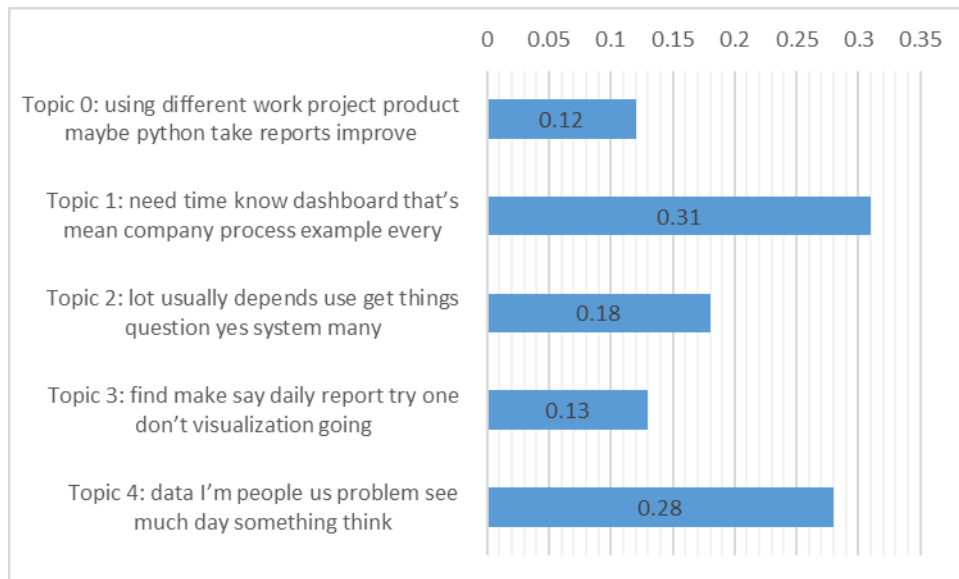
Interpretations of the word cloud: Some words with the highest frequencies are ‘know’ ‘need’ ‘data’, ‘project’ and ‘time’. It means data analysts usually take much time to figure out the clients’ needs in each project.

Topic modeling results for analyzing the materials:

The weight means how important this topic is in the document. The higher the weight, the more important of this topic is in the document.

Overall Dominant Topic: Topic 1 with a weight of 0.31

- Topic 0: using different work project product maybe python take reports improve 0.12
- Topic 1: need time know dashboard that’s mean company process example every 0.31
- Topic 2: lot usually depends use get things question yes system many 0.18
- Topic 3: find make say daily report try one don’t visualization going 0.13
- Topic 4: data I’m people us problem see much day something think 0.28



The most common topic is ‘need time know dashboard that’s mean company process example every’. It means building the dashboard and following the company process takes much time in their daily responsibility.

### Data analysts’ applications of ChatGPT in their work

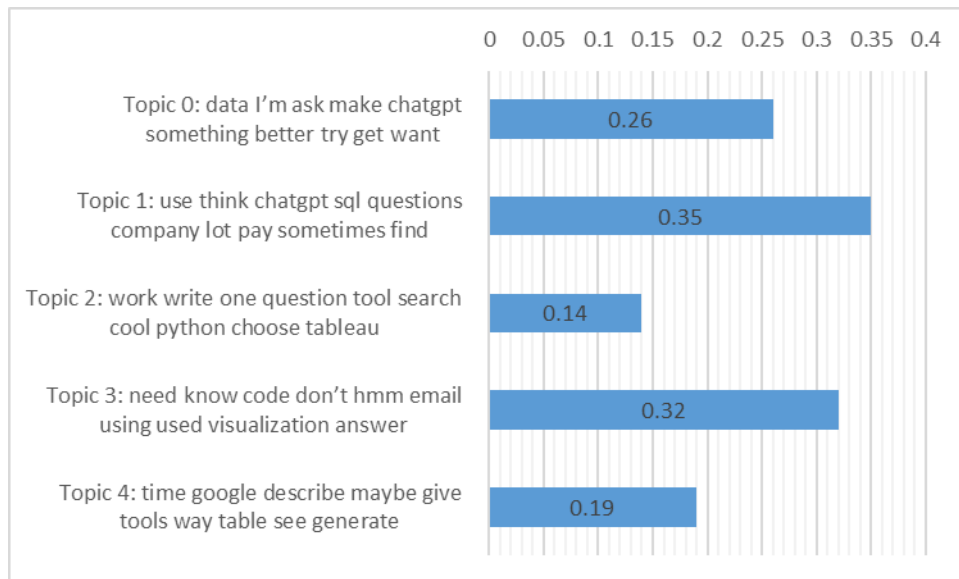
In companies across finance, technology and healthcare industries where interviewees are, there are policies restricting the use of ChatGPT. The main reason behind that is the data privacy issues. Their customer data is sensitive and they do not want it to be compromised by ChatGPT, the open-sourced platform. In addition, I have noticed the main information source of ChatGPT is the interviewee’s conversations with colleagues and news on social media. Only one big accounting firm has the training resources of ChatGPT provided to the employee. Without proper guidance and training, many people may not realize the potential of ChatGPT and do not know how to use it properly.

Some people have some experience with ChatGPT, mainly in writing emails. According to them, the style of email ChatGPT writes is too formal, sometimes they need to adjust the prompt to make it more casual. Therefore, I think it emphasizes the importance of a guidance handbook to the employee. It may contain several prompts of email styles. Whenever they want to write a certain type of email, they can just copy and paste the prompt from the handbook, adjust a bit and the work will be done by ChatGPT. The such handbook can truly save the efforts of asking ChatGPT several times to adjust the context of emails.

In order to find things in common in their talk, I perform a keyword extraction on a file that integrates their conversation about their work application of ChatGPT in their work. Here are the top 20 keywords and their frequencies:

('chatgpt', 62), ('use', 47), ('think', 38), ('data', 38), ('know', 35), ('write', 30), ('work', 29), ('used', 28), ('need', 26), ('code', 23), ('sql', 22), ('lot', 21), ('time', 21), ('re', 20), ('could', 20), ('information', 19), ('something', 18), ('would', 17), ('make', 16), ('question', 16)





The most common topic is 'Topic 1: use think chatgpt sql questions company lot pay sometimes find'. It means data analysts mostly use ChatGPT to answer questions about SQL. The software they use at work is those paid by companies.

### Their views of ChatGPT's limitations

Here is a list of reasons of ChatGPT's limitations according to working data analysts:

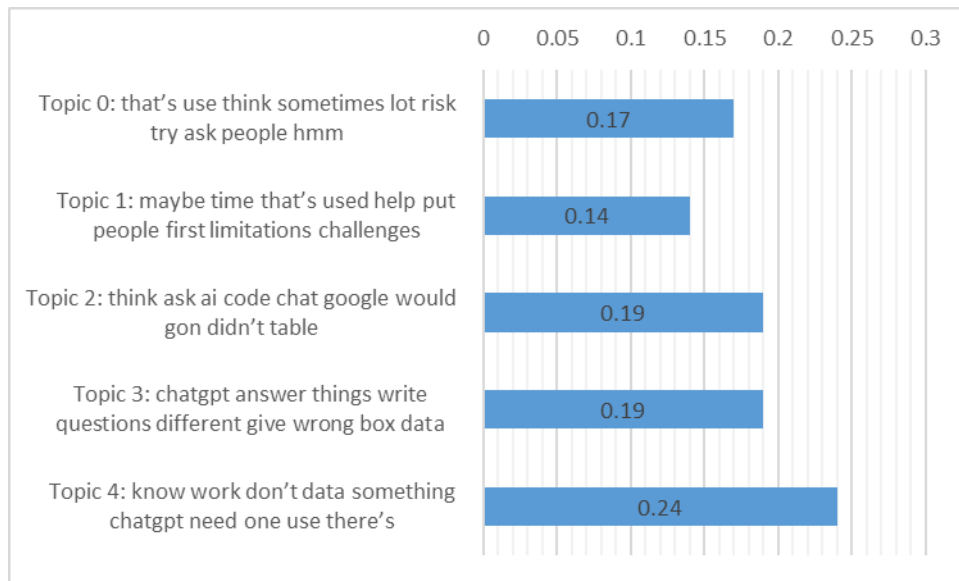
1. Privacy issue (customer data, confidential database schema)
2. Sometimes its information is not accurate
3. They spend most of their time contacting the client's company/PM and figuring out the business needs. The actual amount of time they spend on coding is not very long. But to ask ChatGPT to code, they need to define very specific and complex prompts, which is time-consuming.
4. For learning code, they would use Google as a search engine, because they can see multiple people replying to Stack Overflow and learn different methods. But ChatGPT only gives one solution.
5. ChatGPT can give a broad general data report, but one of DA's responsibilities is to explain why they select this data/ why it is important and use it to persuade managers.
6. Coding quality.
7. While ChatGPT can replace some repetitive work, some of these works are already done by the contract to low-labor-cost countries.

In order to find things in common in their talk, I perform a keyword extraction on a file that integrates their conversation about their view of the limitation of ChatGPT. Here are the top 20 keywords and their frequencies:

('think', 31), ('chatgpt', 25), ('use', 20), ('know', 17), ('work', 16), ('data', 16), ('ask', 14), ('people', 13), ('answer', 13), ('time', 11), ('ai', 10), ('something', 10), ('maybe', 10), ('mean', 10), ('one', 10), ('questions', 10), ('things', 9), ('code', 9), ('using', 8), ('help', 8)







The most common topic is ‘know work don’t data something ChatGPT need one use there’s’. It means data analysts do not put work data on it. They also need time to familiarize with the use of ChatGPT.

### How data analysts think ChatGPT will impact their future work

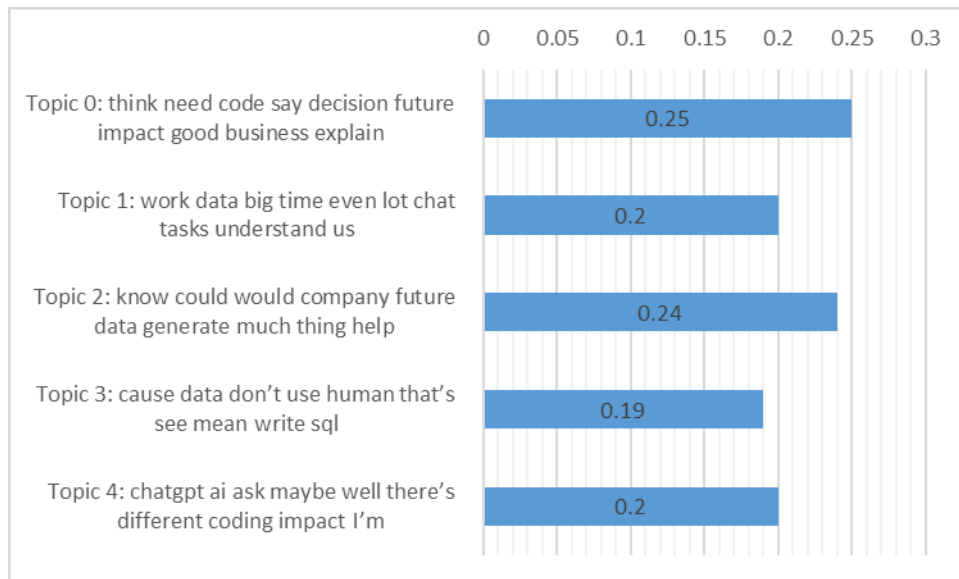
In conclusion, because of its privacy issues, many companies are unwilling to adopt ChatGPT in their day-to-day work. However, we would see its potential in administrative work, such as organizing meetings and writing emails. In some cases, companies can use ChatGPT to create mock data for training purposes for new employees, which can protect their customer data privacy. In the future, companies who have the technical ability will try to develop their own local third-party “ChatGPT” to exploit its analytics ability while protecting the customer data security. In addition, as different prompt requests can create different results, for those companies who wish to use ChatGPT or local third-party “ChatGPT”, they will try to create a tutorial with prewritten prompts for the employees to teach them how to leverage this tool efficiently. It is my observation that one of my interviewee’s companies has already created tutorials of ChatGPT on their internal training platform for their employee.

In conclusion, I do not think ChatGPT can truly replace data analysts in their daily work. The reason is a large part of their work is to understand the context of the problem. They can only give a prompt to ChatGPT when they fully understand the request from the customers and build the logic of the code they would like to write. This part of the work involves lots of thinking and communication that cannot be done by the ChatGPT.

However, I do think the ability to leverage ChatGPT efficiently can create an efficiency difference in their day-to-day work. Lots of work involving recording down the materials, writing emails, or some other administration work. This kind of work can be easily done by ChatGPT. Even if they mainly deal with the data, as ChatGPT has evolved to ChatGPT 4, its API can be used in Excel and PowerPoint. Using ChatGPT can also save a lot of time in pulling data into Excel, cleaning data in Excel and making PowerPoint to visualize it. Therefore, as I mentioned in previous paragraphs, companies should try to make detailed guidance on how to write prompts in ChatGPT and which areas of work can be covered by ChatGPT. Maybe in the future, ChatGPT will become a tool every position uses like Excel. ChatGPT will not replace humans, but people who have great use of ChatGPT will replace people who cannot use ChatGPT efficiently.

In order to find things in common in their talk, I perform a keyword extraction on a file that





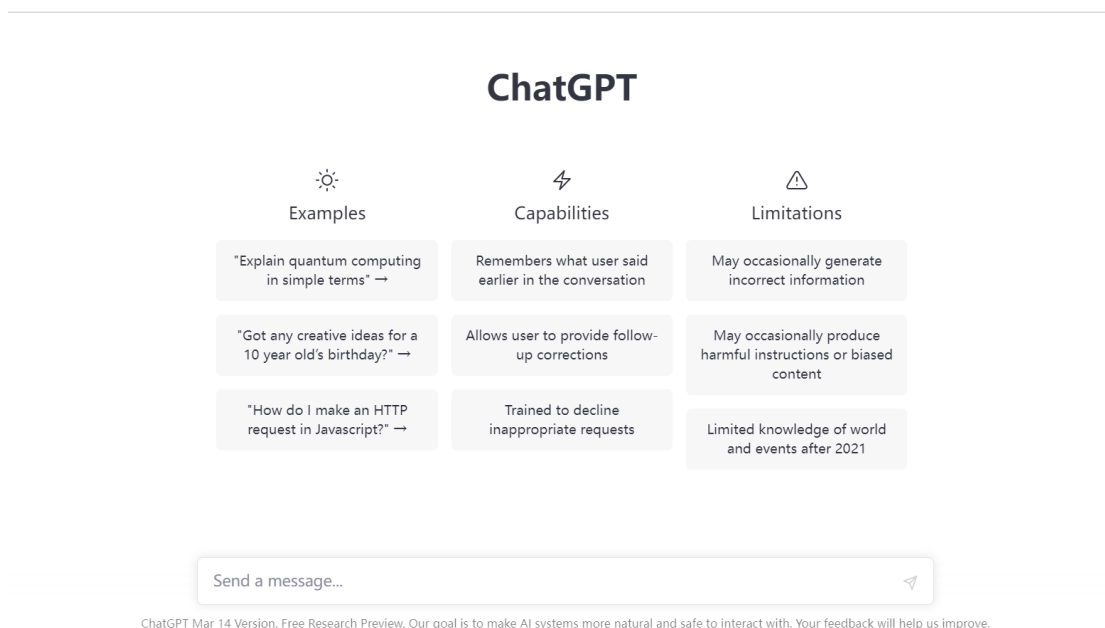
The most common topic is 'Topic 0: think need code say decision future impact good business explain'. It means ChatGPT can be applied to write code and make good explanations about business in the future.

## Examples of applications of ChatGPT in analytics work

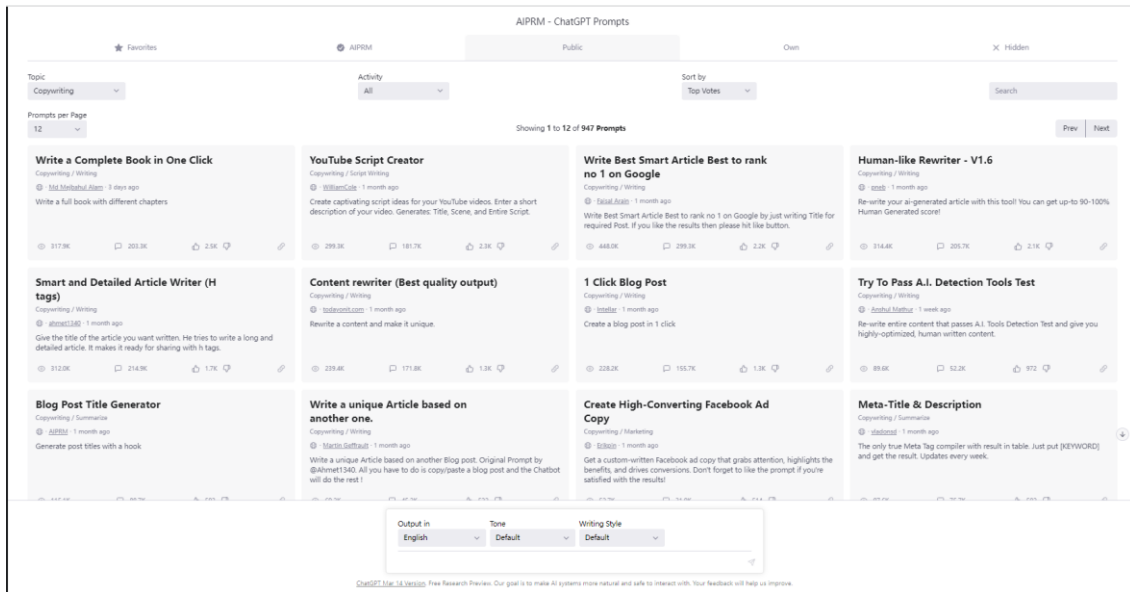
A list of useful analytics tools related to ChatGPT

1. Plugins for writing the prompt
  - AIPRM is a chrome plugin. It has a great collection of high-quality prompts. It is also a great community that provides the environment for posting prompts and communicating on how to improve it.

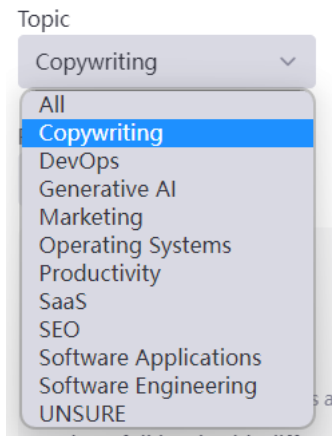
Before installing AIPRM, the interface for ChatGPT is like this:



After installing AIPRM, the interface is like:

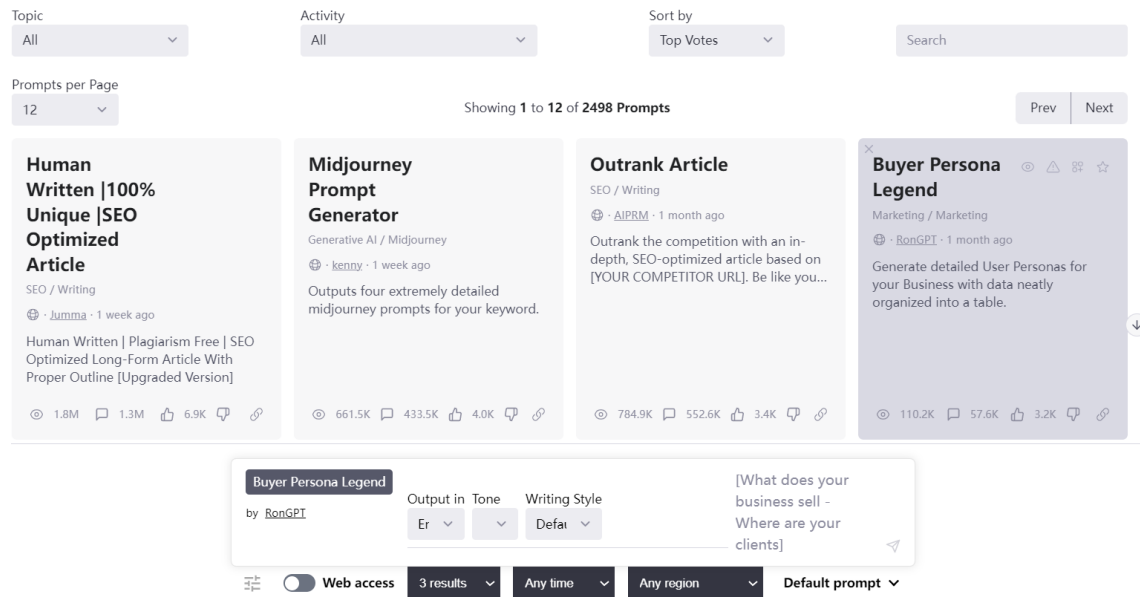


You can select the topic on the left upside and search for the topic on the right upside.



You can use the already written prompts to satisfy your requirement.

For instance, I click Buyer Persona Legend and input 'soft drink' 'young people'.



The plugin generates a whole detailed prompt using my keyword.

SH You are a marketing researcher that speaks and writes fluent English. Your task is to generate a detailed USER PERSONA for a soft drink , young people business in a specific [MARKET].

Structure your response in 4 separate tables.  
 Above the first table write "USER PERSONA [BUSINESS]" and replace [BUSINESS] by soft drink , young people.  
 Here is the required format for each table:

Table 01: DEMOGRAPHICS  
 2 columns and 7 rows  
 Column 1 = Data points (Name, Age, Occupation, Annual income, Marital status, Family situation, Location)  
 Column 2 = Answers for each data point in Column 1 based on the specific [MARKET]

Table 02: USER DESCRIPTION  
 A summary of the user persona in no more than 240 characters.

Table 03: PSYCHOGRAPHICS  
 2 columns and 9 rows  
 Column 1 = Data points (Personal characteristics, Hobbies, Interests, Personal aspirations, Professional goals, Pains, Main challenges, Needs, Dreams)  
 Column 2 = Answers for each data point in Column 1 based on the specific [MARKET]

Table 04: SHOPPING BEHAVIORS  
 2 columns and 8 rows  
 Column 1 = Data points (Budget, Shopping frequency, Preferred channels, Online behavior, Search terms, Preferred brands, Triggers, Barriers)  
 Column 2 = Answers for each data point in Column 1 based on the specific [MARKET]

Please make sure that your response is structured in 4 separate tables and has a separate row for each data point. Do not provide bullet points.  
 At the end of your response, include these 3 sentences:  
 🙏 Thanks for using this Prompt.  
 📺 Learn more with our ChatGPT tutorials on YouTube: <http://bit.ly/rondiplayai>  
 🏠 RonGPT

Please provide all responses in English

- Prompt Storm is another tool that has preset ChatGPT prompts.
- 2. Browse AI- monitor the website and capture data on the website
- 3. DurableAI website builder- build website
- 4. Some low or no-code AI platforms for analytics (Marr, 2022)
- Datarobot- Prep modeling data, validate and govern models

- Nanonets – extract structured or semi-structured data from documents
  - Base64AI- extract text, data, photos & more from all types of documents
  - ObviouslyAI – perform time series analysis and prediction model by simply uploading the file
  - Crystal – Connect different data sources (Salesforce, Google Sheets, Oracle, Azure SQL, SaaS Applications, etc.) and get real-time insights with Crystal, the GPT for numbers platform.
5. AI tools for research
    - Ask Your PDF – upload PDF and ask questions to GPT about this article
  6. AutoGPT- GPT that can learn from its experience and grow

## Use cases of ChatGPT on data analysts' work

In conclusion I found that many companies are reluctant for data analysts to directly put data on ChatGPT because of privacy issues. In addition, data analysts also complain about using ChatGPT for data analysis purposes. One thing they concern about is the word limitation of the prompt, which prevents them from inputting large-scale data. Another thing is that ChatGPT is often out of capacity during work hour.

Inspired by these limitations, I discovered ways of avoiding compromising data, limitations on input and avoiding high traffic time using ChatGPT. That is to use ChatGPT to write Python code to analyze the local file. They only need ChatGPT to produce Python script once, which can avoid the high traffic time of ChatGPT. In addition, because the Python script executed a local file, it also solves the problem of input limitation and data security. Besides, once they write a Python script for a specific purpose, they can reuse it for different files. So, the effort only needs to be taken at the beginning with the assistance of ChatGPT. These are just examples of how to use it.

During the real work of data analysts, I am sure they can invent more use of ChatGPT in various work aspects. And I do believe in the near future, when companies have designed their localized version of Large Language Models, they can directly import their data to it for analysis. The use cases I have generated below are focused on the processes to acquire data from the website or doing preliminary research on the client's company using open-sourced information on the Internet.

The ChatGPT version used in this paper is ChatGPT 3.5, the version without updating to ChatGPT 4.

The code created from ChatGPT can be found at: <https://github.com/NicoleShou/ChatGPT-in-analytics-project/tree/main>

Use case 1: Analysts need to work on figuring out the business needs before starting the analysis. Therefore, it is useful to give an overall tech report of the business you want to look at. I find the prompts in AIPRM and specify my needs to include main products, the technology used in the product, revenue/profit situation in recent years, trends, staff number and business model risk.

SH

As an analyst with comprehensive experience in software development, technology and business operations write about

Give me a tech business report of the company Intuit. It includes its main products, technology used to support the products, revenue /profit margin in recent years, whether it has a rising profit trend or decreasing trend, staff number and how it changes between the years, any risks in its business models..

In professional business English. The summary should be detailed and firstly describe the topic and then reflect on it.





#### Introduction:



This tech business report will provide an overview of the company Intuit, its main products, the technology used to support those products, and its revenue and profit margin trends in recent years. It will also examine changes in the number of staff and potential risks in its business models.

#### Company Overview:

Intuit is a technology-driven company that provides financial management and tax preparation software for individuals and small businesses. The company was founded in 1983 and has its headquarters in Mountain View, California.

#### Products:

Intuit's main products include QuickBooks, TurboTax, and Mint. QuickBooks is a cloud-based accounting software that helps small businesses manage their finances, including payroll, inventory, and expenses. TurboTax is a tax preparation software that assists individuals and small businesses in preparing and filing their tax returns. Mint is a personal finance management software that enables users to track their expenses and manage their budgets.

#### Technology:

Intuit's products are powered by cloud-based technology that enables users to access their financial data from anywhere with an internet connection. The company also uses artificial intelligence and machine learning to provide personalized financial insights to its users.

#### Revenue and Profit Margin:

Intuit's revenue has steadily increased in recent years, from \$6.1 billion in 2018 to \$7.7 billion in 2021. Its profit margin has also been consistently high, at around 20% in the past three years.

#### Staff Number:

Intuit has seen a steady increase in its number of employees, from 8,000 in 2018 to 10,000 in 2021.

#### Profit Trend:

Intuit's profit trend has been consistently rising in recent years. This is primarily due to the increasing demand for its products and the shift towards cloud-based software.

#### Risks:

One potential risk to Intuit's business model is increased competition in the financial management software market. There are several companies that offer similar products and services, and Intuit will need to continue innovating and improving its offerings to remain competitive. Additionally, changes in tax laws and regulations could impact the demand for its tax preparation software.

#### Conclusion:

Intuit is a technology-driven company that provides financial management and tax preparation software for individuals and small businesses. Its products are powered by cloud-based technology and artificial intelligence, and the company has seen a steady increase in revenue and profit margin in recent years. However, there are potential risks to its business model, including competition and changes in tax laws and regulations. Overall, Intuit appears to be well-positioned for continued growth and success in the future.

But because ChatGPT's data is only up to 2021, therefore we cannot see the analysis after 2021.

Therefore, I need to introduce a Chrome extension called web ChatGPT. It allows ChatGPT to connect to the Internet and search for real-time results using the search engine Duck duck go.

After installing Web ChatGPT, and I add in my prompt that the data should be up to 2023, it automatically searches on the website, which is reflected in my prompt.

Web search results:  

[1] "Put simply, a Business Analyst (BA) evaluates business data to improve decision-making within the organization. They essentially act as a bridge between management or stakeholders (the decision-makers) and production (the decision implementers). Their goals are to maximize profits, streamline production, and increase efficiency."  
URL: <https://www.codecademy.com/resources/blog/what-is-a-business-analyst-in-software-development/>

[2] "Some job titles similar to business systems analyst, along with their average annual base salary in the US, include: Management analyst: \$69,881. IT business analyst: \$77,029. Computer systems analyst: \$73,715. Information systems manager: \$94,755. \*All salary data is sourced from Glassdoor as of August 2022."  
URL: <https://www.coursera.org/articles/business-systems-analyst>

[3] "Business analysts use data to form business insights and recommend changes in businesses and other organizations. Business analysts can identify issues in virtually any part of an organization, including IT processes, organizational structures, or staff development. As businesses seek to increase efficiency and reduce costs, business analytics ..."  
URL: <https://www.coursera.org/articles/what-does-a-business-analyst-do-and-how-to-become-one>

Current date: 4/1/2023

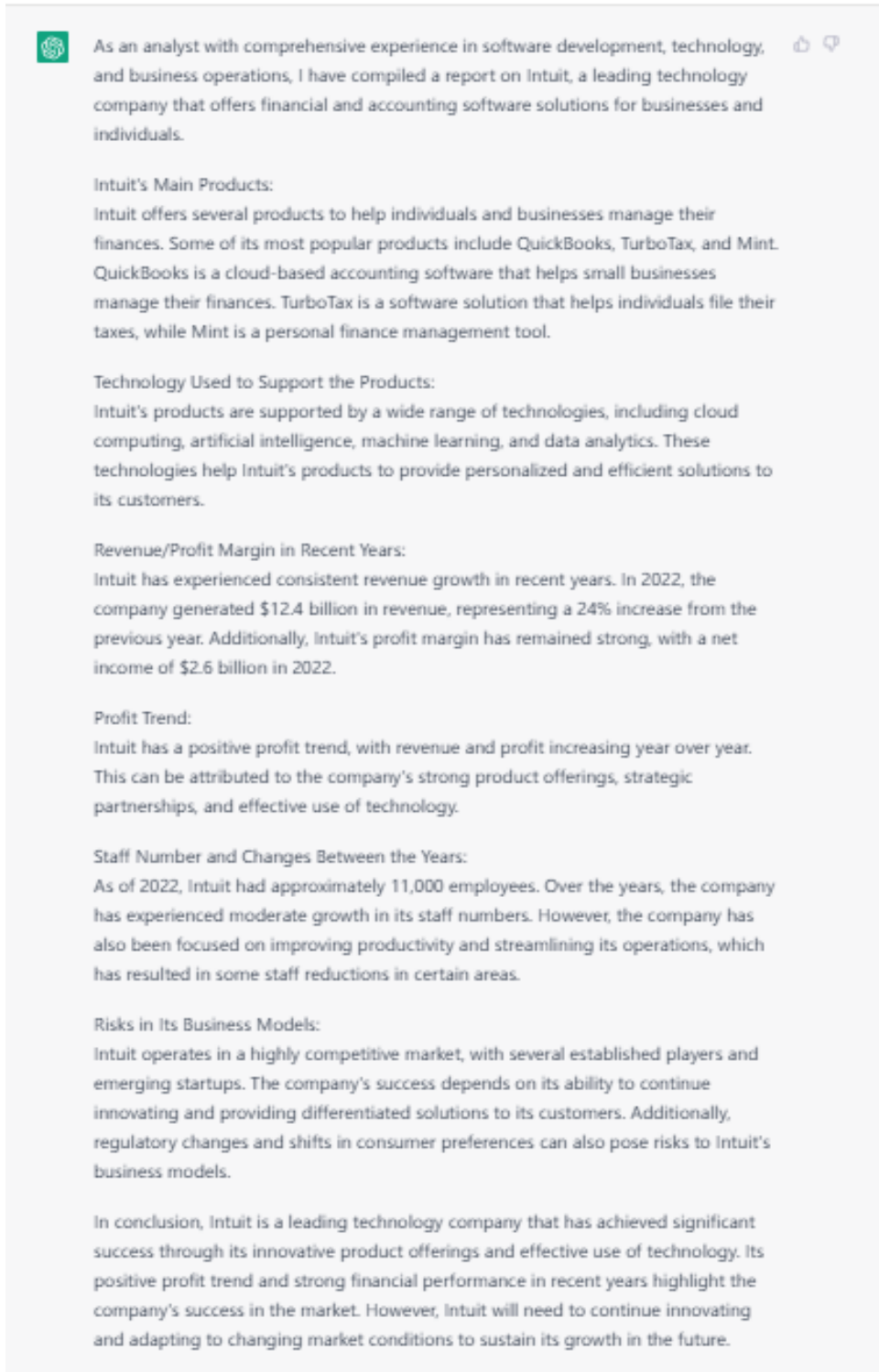
Instructions: Using the provided web search results, write a comprehensive reply to the given query. Make sure to cite results using `[[number](URL)]` notation after the reference. If the provided search results refer to multiple subjects with the same name, write separate answers for each subject.

Query: As an analyst with comprehensive experience in software development, technology an business operations write about

Give me a tech business report of the company Intuit. It includes its main products, technology used to support the products, revenue /profit margin in recent years, whether it has a rising profit trend or decreasing trend, staff number and how it changes between the years, any risks in its business models..

In professional business English. The summary should be detailed and firstly describe the topic and then reflect on it. The data should be up to 2023.

The final report is like this. We can notice slight changes in the output. In Revenue/Profit Margin in Recent Years section and Staff Number and Changes Between the Years section. The new generated report's data is up to 2022. ChatGPT may search the financial statement of 2022 online and integrate it into the analysis.



Use case 2: Use ChatGPT to extract data from website.



Businesses may need to extract outside data from a public website. It is a process that needs lots of repetitive work of writing Python crawling script. I test on ChatGPT to extract data for an IMDB TV review page. I enrich my prompt to ask the ChatGPT to save the extracted data to a csv file with the separate column as Rating, Review Time, Review Title, User and Review Content. The data is much cleaner and more prepared for further analysis. ChatGPT gave me a Python script for this function.

SH

write me a python web crawler for this website:  
[https://www.imdb.com/title/tt3581920/reviews/?ref\\_=tt\\_ov\\_rt](https://www.imdb.com/title/tt3581920/reviews/?ref_=tt_ov_rt). Include the rating(enclose the rating value in quotes before writing it to the csv file), review time, review title, user and review content. Import the result to a csv file.

Please write in English language.



Sure, here's the updated Python web crawler code that includes the rating(enclosing the rating value in quotes before writing it to the csv file), review time, review title, user, and review content, and imports the result to a csv file:  

```
python Copy code

import requests
from bs4 import BeautifulSoup
import csv

# specify the URL to scrape
url = "https://www.imdb.com/title/tt3581920/reviews/?ref_=tt_ov_rt"

# send a GET request to the URL
response = requests.get(url)

# create a BeautifulSoup object to parse the HTML content
soup = BeautifulSoup(response.content, "html.parser")

# find all the review containers on the page
reviews = soup.find_all("div", class_="lister-item-content")

# create a list to store the extracted data
data = []
```

```

# loop through each review and extract the desired information
for review in reviews:
    rating = '"' + review.find("span", class_="rating-other-user-rating").text.strip()
    title = review.find("a", class_="title").text.strip()
    user = review.find("span", class_="display-name-link").text.strip()
    date = review.find("span", class_="review-date").text.strip()
    content = review.find("div", class_="text").text.strip().replace("\n", " ")

    # append the extracted data to the list
    data.append([rating, date, title, user, content])

# write the extracted data to a csv file
with open("imdb_reviews.csv", "w", encoding="utf-8", newline="") as f:
    writer = csv.writer(f)
    writer.writerow(["Rating", "Review Time", "Review Title", "User", "Review Content"])
    writer.writerows(data)

print("Data extraction complete. Check imdb_reviews.csv file for the results.")

```

This script should extract the desired information and save it to a CSV file. Note that we're enclosing the rating value in double quotes before writing it to the CSV file to prevent it from being automatically converted to a date format by the CSV writer.

The output is saved to a CSV file with the exact columns I asked.

Rating	Review Time	Review Title	User	Review Content
"7/10"	13 March	Peaked to Leungzy	I'll start with the negatives:My opinion after Episode 8 was that you cannot tell the story of TLOU in such short episodes, that was after a 50 min episode.For the	
"7/10"	13 March	Doesn't q matthewr	Season 1:While I may feel disappointment by this television adaptation. I am grateful that it introduced the greatest story gaming has ever told to a wider audien	
"8/10"	16 Januar	Great first eve_dollu****	Update after watching the full series ****Well I think overall Id give this a 7 or 8, the opening episode was definitely a 9 the rest of the show would be gettin	
"7/10"	15 March	Good but gallagherk	I know everyone is obsessed with this show, and I can see why. The acting is really good, not only by the main actors but all supporting actors.I didn't play the gar	
"6/10"	18 March	Just OK	knifemagr HBO, of course, deviates and inflates things at every change they get, dedicating an entire episode to a gay couple's love story (which wasn't in the game), or dedi	
"6/10"	14 March	Great ada rockdude	Let me start with saying this is one of the best video game adaptations. There's a lot to love in this series that thankfully stayed true to the game while creating sp	
"9/10"	13 March	First Seas Mickipark	The first season of The Last of Us is a masterpiece of post-apocalyptic storytelling that will leave you wanting more. There is hardly anything to complain about, e	
"7/10"	30 Januar	"I think I Gioioe6	I'm not here to comment on the relationship to the video game this show is adapted from, I don't play games like that and don't know or care anything about the	
"6/10"	30 Januar	Get a Grip skinzilla	So far it's a fairly generic but acceptable plot. The infected are more comical than they are fearsome. The acting is decent, but not a single performance has stood	
"4/10"	18 Februa	Another Gdk777	It started off interesting, only to turn out to be another standard series that gradually got worse and worse as time went on.The beginning was solid, the charact	
"8/10"	28 Januar	Good So Ftstarr-991	The Last Of Us is off to a good start, and as far as video game adaptations go this is a solid adaptation. So far. It's too early to call this a masterpiece or give it a hi	
"8/10"	16 Januar	That's hovrapaelmi	First of all, I'd like to make it clear that I'm writing this review having only watched the episode and never having played the game or researched anything about it	
"5/10"	15 March	The high r Silcone54	To start I gave this movie a 5, which to me means its good enough to watch but your life won't change if you don't. 6-9 scales up from you should give it a try to y	
"7/10"	14 March	Great star annajane	r The pilot of the 'Last of Us' is one of the best pilots I've ever watched. It is a seriously exciting, terrifying and emotional episode of television.The world presented	
"5/10"	19 March	What was halcyoned	They wasted an unforgivably long amount of time exploring the lives of the supporting characters, while dedicating only half of the season to Joel and Ellie. The s	
"7/10"	25 March	Interestinjuraj-jerko	Disclaimer: Never played the game.I kept reading how great this series is and looked forward to it. I watched all of the episodes and I do not share the enthusias	
"10/10"	17 Januar	The Last o OzzzzSzzz	"The Last of Us," the highly-anticipated HBO series based on the critically acclaimed video game of the same name, is a truly remarkable viewing experience that i	
"6/10"	14 March	Overhypeitmbhuges	This is the most overhyped show I've ever watched. It's solid. Very predictable, nothing special. A lot of it watched like levels of a computer game, which it was, w	
"6/10"	4 Februan	Way overraajp2012	The amount of reviews here saying "it's the best thing I've ever seen" is absolutely hilarious and after the pilot episode everyone decided it was worthy of a 10/1	
"4/10"	6 Februan	Started oushalloswe	Started off pretty good, spot on to the details of the game, as well as some interesting back story etc. However the last 2 episodes 3 and 4 are boring... Nothing ha	

Use case 3: After finding the data on the website, transform it to an SQL table.

I find the prompts in AIPRM.

SH

[Ignore all previous instructions] You are a SQL expert with over 20 years of experience, your job is to transform the data I provide you into organised, clean, coherent, and readable tables. You must always add an auto-increment column as the first column of the table. You will always ask me which column I want to have in my table and their name. You will always ask me if I want you to display the full table in your message.



You MUST send a visual table in your message as an SQL expert, you know how important it is to see what the table looks like before executing SQL queries. Then, you MUST send the SQL queries that will allow me to fully recreate the table I sent you, including the query to create the table and columns, and the query to populate the columns with all the values. All output should be in English.

Date	Sales (in units)	Revenue (in currency)
April 1, 2023	50	\$1,000
April 2, 2023	62	\$1,240
April 3, 2023	45	\$900
April 4, 2023	37	\$740
April 5, 2023	55	\$1,100
April 6, 2023	48	\$960
April 7, 2023	60	\$1,200



Understood. Here is the table `sales` with auto-increment column `id` as the first column:

id	date	sales	revenue
1	April 1, 2023	50	\$1,000
2	April 2, 2023	62	\$1,240
3	April 3, 2023	45	\$900
4	April 4, 2023	37	\$740
5	April 5, 2023	55	\$1,100
6	April 6, 2023	48	\$960
7	April 7, 2023	60	\$1,200

And here are the SQL queries to create and populate the table:

```
SQL Copy code
CREATE TABLE sales (
  id INT AUTO_INCREMENT PRIMARY KEY,
  date DATE,
  sales INT,
  revenue VARCHAR(10)
);

INSERT INTO sales (date, sales, revenue)
VALUES ('2023-04-01', 50, '$1,000'),
       ('2023-04-02', 62, '$1,240'),
       ('2023-04-03', 45, '$900'),
       ('2023-04-04', 37, '$740'),
       ('2023-04-05', 55, '$1,100'),
       ('2023-04-06', 48, '$960'),
       ('2023-04-07', 60, '$1,200');
```



Let me know if you have any questions or need further assistance!

Sometimes the data file is too large to type in manually. Often it is stored in a CSV file. We can also ask GPT how to import a CSV file to the MySQL database.

SS

Hi ChatGpt, you are a data analyst working for a large firm. I have a question for you. I have a data set in the form of csv. How can I import into MySQL database?



Hello! I can certainly help you with that. To import a CSV file into a MySQL database, you can follow these steps:  

1. Login to your MySQL server using a client tool such as MySQL Workbench or the command line.
2. Create a new database or select the database where you want to import the data.
3. Create a new table that matches the structure of the CSV file. You can use the CREATE TABLE statement to define the column names and data types.
4. Use the LOAD DATA INFILE statement to import the CSV file into the MySQL database.

The syntax for the command is as follows:

```
vbnet Copy code  
  
LOAD DATA INFILE 'file.csv'  
INTO TABLE table_name  
FIELDS TERMINATED BY ','  
ENCLOSED BY '"'  
LINES TERMINATED BY '\n'  
IGNORE 1 ROWS;
```

- Replace `'file.csv'` with the path to your CSV file.
  - Replace `'table_name'` with the name of the table you created in step 3.
  - The `'FIELDS TERMINATED BY ',''` statement indicates that the fields in the CSV file are separated by commas. If your file is separated by a different character, you can change this value accordingly.
  - The `'ENCLOSED BY '"''` statement indicates that fields in the CSV file are enclosed in double quotes. If your file uses a different character to enclose fields, you can change this value accordingly.
  - The `'LINES TERMINATED BY '\n''` statement indicates that each row in the CSV file is terminated by a new line character. If your file uses a different character to terminate rows, you can change this value accordingly.
  - The `'IGNORE 1 ROWS'` statement indicates that the first row of the CSV file should be ignored, assuming that it contains the column headers.
5. Once you execute the command, the data from the CSV file will be imported into the MySQL database.

I hope this helps! Let me know if you have any further questions.

## Use Cases for Analytics Courses using ChatGPT

While ChatGPT can help data analysts with various aspects of their work, according to the

study it can also contribute effectively to peer tutoring. Pardos and Bhandari found that 70% of the hints offered by ChatGPT in elementary and intermediate Algebra topics could result in positive learning gains for students (Shahriar & Hayawi, 2023). Although students are not allowed to directly use ChatGPT for homework, it can provide great insights and help them learn.

I have created use cases for three analytics courses I have taken in MASY program: Text Data Mining course, Database Design & Management Course and Data Mining and Data Warehousing course.

Use case for Text Data Mining course:

When I was doing keyword extraction and word cloud for my survey, I thought about how to do it using ChatGPT. At first, I was able to create separate Python script for the extraction of keywords and generate a word cloud. After modifying the prompts and adding some stop words, I was able to use ChatGPT to generate one Python script that can do both keyword extraction and word cloud creation. It reads the local file as input materials. When I want to analyze different files, I only need to modify 'path/to/your/file.txt' this part to create different results. In addition, because it is a Zoom script that includes a timestamp and multiple punctuations, I also told ChatGPT to remove them from the keyword results by identifying them as stop words. The keyword generated should not contain words like 'yeah' and timestamp. I can customize the keywords by including them in the prompt I give to ChatGPT. It has made my work much more efficiently.

**SH** Write me a python script that extract keywords and generate a wordcloud for data mining. The result should include printed 20 keywords with frequency and a word cloud image. Include timestamp and punctuations in stopwords. Also include 'like', 'oh', 'right', "'m", "'s", 'yeah', 'okay', 'also', 'thank', "n't", 'yup', 'sure' as stopword. The subject for analyzing is a local txt file.





Sure, here's a Python script that extracts keywords and generates a wordcloud for data mining:



```
python Copy code

import re
import nltk
from nltk.corpus import stopwords
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from datetime import datetime

# Define the stopwords
stopwords = set(stopwords.words('english'))
stopwords.update(['like', 'oh', 'right', "m", "s", 'yeah', 'okay', 'also', 'thank', 'n

# Read the local txt file
with open('path/to/your/file.txt', 'r') as file:
    data = file.read()

# Tokenize the text and remove stopwords
tokens = nltk.word_tokenize(data.lower())
filtered_tokens = [token for token in tokens if token not in stopwords]

# Count the frequency of each word
freq_dist = nltk.FreqDist(filtered_tokens)
```

```
# Print the top 20 most frequent words
print(f"Top 20 most frequent words: {freq_dist.most_common(20)}")

# Generate the word cloud image
wordcloud = WordCloud(width=800, height=800, background_color='white', stopwords=st

# Plot the word cloud image
plt.figure(figsize=(8,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()

# Add timestamp
print("Script executed on:", datetime.now())
```

Make sure to replace ``path/to/your/file.txt`` with the actual path to your local txt file. The script reads the file, tokenizes the text, removes the stopwords (including the specified additional ones), counts the frequency of each word, prints the top 20 most frequent words, generates the word cloud image based on the word frequency, and plots the image using `matplotlib`. Finally, the script prints the timestamp when it was executed.

Example output:

```
View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)
Top 20 most frequent words: [('data', 83), ('need', 45), ('know', 45), ('time', 36), ('usually', 29), ('dashboard', 20), ('see', 20), ('day', 20), ('mean', 20), ('cause', 20), ('work', 19), ('would', 19), ('project', 19), ('much', 18), ('us', 18), ('team', 17), ('problem', 17), ('people', 17), ('system', 16), ('na', 16)]
```



In addition, I asked ChatGPT to write a Python script that can do the topic modeling for me. I input stopwords in my prompts and ask it to remove timestamps. In addition, I asked it to calculate the weight of topics (how important this topic is to the whole document) and generate a bar graph to visualize the results.

```
SH Write me a Python script to achieve the following function: Extract 5 most common topics for 10 words each for a local zoom script. The zoom transcript has time stamp like [hh:mm:ss]. Include timestamp and punctuations in stopwords. Also include words like 'like', 'oh', 'right', '"m"', 's', 'yeah', 'okay', 'also', 'thank', 'n't', 'yup', 'sure' as stopword. Extracts dominant topic for whole text and shows the overall weight of the topic in a nicely formatted output. Print 10 words for each topic and an overall weight of the topic following 10 words please. I do not want to print all sentences. Make a simple bar chart to visualize the topics and weight.
```



Sure, here's a Python script that uses the Natural Language Toolkit (NLTK) library and Gensim library to extract 5 most common topics for 10 words each for a local zoom script, as per your requirements:



```
python Copy code

import re
import string
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from gensim import corpora, models

# Load stop words and add custom ones
stop_words = set(stopwords.words('english'))
custom_stopwords = {'like', 'oh', 'right', "'m", "'s", 'yeah', 'okay', 'also'}
stop_words = stop_words.union(custom_stopwords)
```

```
# Load zoom transcript
with open('zoom_transcript.txt', 'r') as file:
    transcript = file.read()

# Clean up transcript
transcript = re.sub('\[.*?\]', '', transcript) # Remove timestamps
transcript = transcript.translate(str.maketrans('', '', string.punctuation))
transcript = transcript.lower() # Convert to lowercase

# Tokenize transcript
tokens = word_tokenize(transcript)

# Remove stop words
filtered_tokens = [token for token in tokens if token not in stop_words]

# Create bag of words
dictionary = corpora.Dictionary([filtered_tokens])
corpus = [dictionary.doc2bow([token]) for token in filtered_tokens]

# Train LDA model
lda_model = models.LdaModel(corpus, num_topics=5, id2word=dictionary)
```

```

# Extract dominant topic and overall weight
dominant_topic = None
dominant_weight = 0
for topic, weight in lda_model.show_topics(formatted=False):
    topic_weight = sum([w for _, w in weight])
    if topic_weight > dominant_weight:
        dominant_topic = topic
        dominant_weight = topic_weight

# Print dominant topic and overall weight
print(f"Overall Dominant Topic: Topic {dominant_topic} with a weight of {dominant_weight}")

# Print topics and weights
topics_df = pd.DataFrame(columns=['Topic', 'Weight'])
for i, (topic, weight) in enumerate(lda_model.show_topics(formatted=False)):
    topic_weight = sum([w for _, w in weight])
    topics_df.loc[i] = [f"Topic {topic}", topic_weight]
    print(f"Topic {topic}: {topic_weight:.2f}")
    print(" ".join([w[0] for w in weight[:10]]))
    print()

```

```

# Create bar chart
topics_df.set_index('Topic', inplace=True)
topics_df.plot(kind='bar', legend=False)
plt.title('Topic Distribution')
plt.xlabel('Topic')
plt.ylabel('Weight')
plt.show()

```

Here's what this script does:

1. Loads the stop words and adds custom ones based on your requirements.
2. Loads the zoom transcript from a file, removes timestamps and punctuations, and converts all text to lowercase.
3. Tokenizes the transcript and removes stop words.
4. Creates a bag of words using the filtered tokens.
5. Trains an LDA (Latent Dirichlet Allocation) model with 5 topics using the bag of words.
6. Extracts the dominant topic and overall weight based on the weights of each topic.
7. Prints the dominant topic and overall weight, as well as the top 10 words for each topic.
8. Creates a bar chart showing the distribution of topic weights.

Example output:

Overall Dominant Topic: Topic 0 with a weight of 0.25

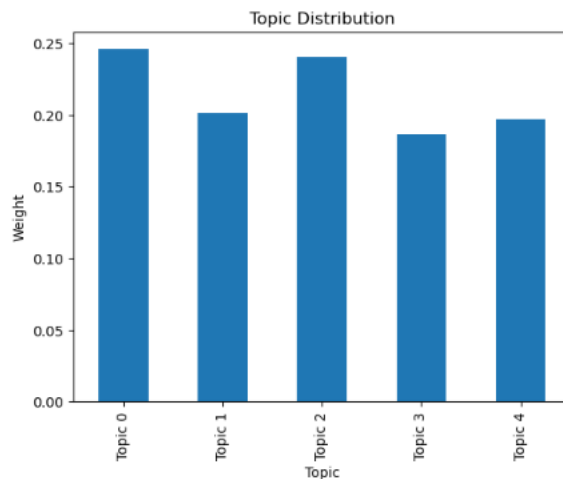
Topic 0: 0.25  
think need code say decision future impact good business explain

Topic 1: 0.20  
work data big time even lot chat tasks understand us

Topic 2: 0.24  
know could would company future data generate much thing help

Topic 3: 0.19  
cause data dont use human thats see mean write sql

Topic 4: 0.20  
chatgpt ai ask maybe well theres different coding impact im



Use case for Database Design & Management Course:

ChatGPT can help professors create teaching materials. It can generate mock databases from scratch and add explanations while designing the database. It will make the preparation of class much easier because professors can just modify the content ChatGPT has generated instead of doing it by hand from scratch.

ss

You are a professor with 20 years teaching experience for a database course. You are creating teaching materials for your course. Please create a database schema for a soft drink business in order to increase revenue with at least 8 tables. For each table, list all the columns in the table and identify the primary keys and foreign key. Please explain clearly the relationship between tables (is it one to one or one to many) and why it is designed so. Please also explain how it can increase the revenue.

Please write in English language.



Sure, I'd be happy to help you create a database schema for a soft drink business.



The schema is designed to increase revenue by providing the business with a better way to manage their products, customers, sales, and inventory.

The following are the 8 tables and their respective columns, primary keys and foreign keys:

1. Products table:

Columns: product\_id, product\_name, product\_type, price, quantity, date\_added

Primary Key: product\_id

2. Ingredients table:

Columns: ingredient\_id, ingredient\_name, supplier\_id, price\_per\_unit

Primary Key: ingredient\_id

Foreign Key: supplier\_id

3. Suppliers table:

Columns: supplier\_id, supplier\_name, supplier\_address, supplier\_phone

Primary Key: supplier\_id

4. Customers table:

Columns: customer\_id, customer\_name, customer\_address, customer\_phone

Primary Key: customer\_id

5. Sales table:

Columns: sale\_id, customer\_id, product\_id, quantity, sale\_date

Primary Key: sale\_id

Foreign Keys: customer\_id, product\_id

6. Orders table:

Columns: order\_id, supplier\_id, ingredient\_id, quantity, order\_date

Primary Key: order\_id

Foreign Keys: supplier\_id, ingredient\_id

7. Inventory table:

Columns: product\_id, ingredient\_id, quantity

Primary Key: product\_id, ingredient\_id

Foreign Keys: product\_id, ingredient\_id

8. Recipes table:

Columns: product\_id, ingredient\_id, quantity

Primary Key: product\_id, ingredient\_id

Foreign Keys: product\_id, ingredient\_id

Explanation of the relationships between tables:

- One product can have many ingredients, and one ingredient can be used in many products, so there is a many-to-many relationship between the Products and Ingredients tables. This is managed through the Inventory table, which acts as a junction table.

- One supplier can supply many ingredients, but each ingredient can only be supplied by one supplier. Therefore, there is a one-to-many relationship between the Suppliers and Ingredients tables.
- One customer can make many purchases, but each purchase can only be made by one customer. Similarly, one product can be sold to many customers, but each sale is for only one product. Therefore, there is a many-to-many relationship between the Customers and Products tables, which is managed through the Sales table, acting as a junction table.
- One supplier can make many orders, but each order is made to only one supplier. Similarly, one ingredient can be ordered from many suppliers, but each order is for only one ingredient. Therefore, there is a many-to-many relationship between the Suppliers and Ingredients tables, which is managed through the Orders table, acting as a junction table.
- One product can have many recipes, but each recipe is for only one product. Similarly, one ingredient can be used in many recipes, but each recipe is for only one ingredient. Therefore, there is a many-to-many relationship between the Products and Ingredients tables, which is managed through the Recipes table, acting as a junction table.

How this database schema can increase revenue:

By using this database schema, the soft drink business can keep track of their products, inventory, customers, and sales more efficiently. They can analyze their sales data to identify popular products, which can inform their marketing and production strategies. They can also manage their inventory more effectively, ensuring that they always have enough ingredients to meet demand without overstocking. This can help reduce waste and improve profitability. Additionally, by managing their suppliers more effectively, they may be able to negotiate better prices or find new suppliers that offer better deals. All of these benefits can help the soft drink business increase revenue over time.

After creating the data schema, we can use ChatGPT to write the data dictionary for us:

ss For the database you have generated in our conversation, create data dictionary for exactly each column in each table. Do not add new columns and miss columns. Include data type and description. Include different columns from same table in one table. Put 'column name' , ' data type','description' as header for each table and construct like a sheet.

Please write in English language.



Certainly! Here is the data dictionary for each column in each table, presented as a sheet:



Table name: Products

Column name	Data type	Description
product_id	integer	Unique identifier for each product
product_name	varchar(255)	Name of the product
product_type	varchar(255)	Type of the product (e.g. soda, juice, energy)
price	decimal(10,2)	Price of the product
quantity	integer	Number of units in stock for the product
date_added	date	Date the product was added to the database

Table name: Ingredients

Column name	Data type	Description
ingredient_id	integer	Unique identifier for each ingredient
ingredient_name	varchar(255)	Name of the ingredient
supplier_id	integer	Foreign key referencing the supplier that supplies the ingredient
price_per_unit	decimal(10,2)	Price per unit of the ingredient

Table name: Suppliers

Column name	Data type	Description
supplier_id	integer	Unique identifier for each supplier
supplier_name	varchar(255)	Name of the supplier
supplier_address	varchar(255)	Address of the supplier
supplier_phone	varchar(20)	Phone number of the supplier

Table name: Customers

Column name	Data type	Description
customer_id	integer	Unique identifier for each customer
customer_name	varchar(255)	Name of the customer
customer_address	varchar(255)	Address of the customer
customer_phone	varchar(20)	Phone number of the customer



Table name: Sales

Column name	Data type	Description
sale_id	integer	Unique identifier for each sale
customer_id	integer	Foreign key referencing the customer who made the purchase
product_id	integer	Foreign key referencing the product that was purchased
quantity	integer	Number of units of the product that were sold
sale_date	date	Date the sale was made

Table name: Orders

Column name	Data type	Description
order_id	integer	Unique identifier for each order
supplier_id	integer	Foreign key referencing the supplier who supplies the ingredient
ingredient_id	integer	Foreign key referencing the ingredient that was ordered
quantity	integer	Number of units of the ingredient that were ordered
order_date	date	Date the order was made

Table name: Inventory

Column name	Data type	Description
product_id	integer	Foreign key referencing the product in inventory
ingredient_id	integer	Foreign key referencing the ingredient in inventory
quantity	integer	Number of units of the ingredient used in the product

Table name: Recipes

Column name	Data type	Description
product_id	integer	Foreign key referencing the product the recipe belongs to
ingredient_id	integer	Foreign key referencing the ingredient used in the recipe
quantity	decimal(10,2)	Quantity of

After clarifying the data type of each column, we can ask ChatGPT to generate DDL code for use to build this database in Oracle.

**ss** For the database you have generated in our conversation, create an DDL code that can construct the database using oracle. Be sure to add explanation to your code.



Certainly! Here is the DDL code to create the soft drink business database using Oracle:



```
scss Copy code
--Create the Products table
CREATE TABLE Products (
  product_id INTEGER PRIMARY KEY,
  product_name VARCHAR2(255),
  product_type VARCHAR2(255),
  price NUMBER(10,2),
  quantity INTEGER,
  date_added DATE
);

--Create the Ingredients table
CREATE TABLE Ingredients (
  ingredient_id INTEGER PRIMARY KEY,
  ingredient_name VARCHAR2(255),
  supplier_id INTEGER,
  price_per_unit NUMBER(10,2),
  FOREIGN KEY (supplier_id) REFERENCES Suppliers(supplier_id)
);

--Create the Suppliers table
CREATE TABLE Suppliers (
  supplier_id INTEGER PRIMARY KEY,
  supplier_name VARCHAR2(255),
  supplier_address VARCHAR2(255),
  supplier_phone VARCHAR2(20)
);
```

```

--Create the Customers table
CREATE TABLE Customers (
  customer_id INTEGER PRIMARY KEY,
  customer_name VARCHAR2(255),
  customer_address VARCHAR2(255),
  customer_phone VARCHAR2(20)
);

--Create the Sales table
CREATE TABLE Sales (
  sale_id INTEGER PRIMARY KEY,
  customer_id INTEGER,
  product_id INTEGER,
  quantity INTEGER,
  sale_date DATE,
  FOREIGN KEY (customer_id) REFERENCES Customers(customer_id),
  FOREIGN KEY (product_id) REFERENCES Products(product_id)
);

--Create the Orders table
CREATE TABLE Orders (
  order_id INTEGER PRIMARY KEY,
  supplier_id INTEGER,
  ingredient_id INTEGER,
  quantity INTEGER,
  order_date DATE,
  FOREIGN KEY (supplier_id) REFERENCES Suppliers(supplier_id),
  FOREIGN KEY (ingredient_id) REFERENCES Ingredients(ingredient_id)
);

```

```

--Create the Inventory table
CREATE TABLE Inventory (
  product_id INTEGER,
  ingredient_id INTEGER,
  quantity INTEGER,
  FOREIGN KEY (product_id) REFERENCES Products(product_id),
  FOREIGN KEY (ingredient_id) REFERENCES Ingredients(ingredient_id)
);

--Create the Recipes table
CREATE TABLE Recipes (
  product_id INTEGER,
  ingredient_id INTEGER,
  quantity NUMBER(10,2),
  FOREIGN KEY (product_id) REFERENCES Products(product_id),
  FOREIGN KEY (ingredient_id) REFERENCES Ingredients(ingredient_id)
);

```

In this code, we first create the Products table with its columns and primary key. Then, we create the Ingredients table with its columns, primary key, and a foreign key referencing the Suppliers table. We then create the Suppliers and Customers tables with their respective columns and primary keys.

Next, we create the Sales and Orders tables with their respective columns, primary keys, and foreign keys referencing the Customers, Products, Suppliers, and Ingredients tables as needed. Finally, we create the Inventory and Recipes tables with their respective columns and foreign keys referencing the Products and Ingredients tables.

Overall, this DDL code creates the soft drink business database with all the necessary tables, columns, primary keys, and foreign keys to store the data needed to manage the business and increase revenue.

Use case for data mining and data warehousing course:

Before each class, students can use ChatGPT to generate explanations of concepts in the syllabus which will be taught during the class. It can help students prepare for the class.

SH

Hi ChatGPT, you are a professor with 20 years experience teaching data warehousing course. Please write a coherent story about these concepts below or simulate a data warehousing process that integrate the following concepts to help students learn better:•

Data Warehouse Schemas

- Dimensions and Dimension Tables
- Facts and Fact Tables
- The Star Schema
- The Snowflake Schema
- Degenerate and Junk Dimensions
- The Data Warehouse Bus Architecture
- Conformed Dimensions and Standard Facts
- Data Granularity
- Changing Dimensions
- Components of a Data Warehouse
- Source Systems, Staging Area, Presentation, Access Tools
- Building the Data Matrix
- The Four Steps Process
- Multiple Fact Tables in a single Data Mart
- Chain, Heterogeneous, Transaction/Snapshot & Aggregate Facts
- Fact and Dimension Table Detail
- Identifying Source for each Fact & Dimension
- Mapping from Source to Target



As a seasoned data warehousing professor, I'm often asked by my students how all the different concepts fit together. To help them better understand, I often tell them the story of a simulated data warehousing process.



Let's imagine a fictional company, XYZ Corp, that is looking to build a data warehouse to help with business intelligence reporting. As with any data warehousing project, the first step is to identify the business requirements and decide on the appropriate data warehouse schema.

After careful analysis, it's decided that the best schema for XYZ Corp's needs is the star schema. This schema has a central fact table surrounded by dimension tables. The fact table contains the measures or metrics of interest, while the dimension tables contain the descriptive attributes.

Next, the dimensions and dimension tables must be defined. These are the categories or aspects of the business that are important for reporting purposes. Examples might include time, geography, product, and customer.

Once the dimensions are defined, the fact tables and fact table detail must be identified. The fact tables are the tables that contain the metrics or measures of interest, and the fact table detail is the level of detail at which the facts are captured. For example, sales might be captured at the daily, weekly, or monthly level.

Now that the dimensions and fact tables have been defined, it's time to build out the star schema. This involves identifying the source systems, staging area, presentation layer, and access tools that will be used. The source systems are the systems that the data is coming from, and the staging area is where the data is temporarily stored and transformed before being loaded into the presentation layer.

As the data is loaded into the presentation layer, the data matrix is built. This is a mapping of the source systems to the target data warehouse tables. It's important to identify the source for each fact and dimension, and map it from source to target.

As the data is being loaded, it's important to consider data granularity. This refers to the level of detail at which the data is captured. In some cases, it may be necessary to capture more detailed data to provide better insight into the business.

As the data warehouse is being built, it's important to consider changing dimensions. These are dimensions that change over time, such as product hierarchies or organizational structures. It's important to design the data warehouse in a way that accommodates these changes.

To ensure scalability and flexibility, it's also important to consider the data warehouse bus architecture. This architecture allows for multiple fact tables in a single data mart, and allows for the integration of heterogeneous and transaction/snapshot facts.

As the data warehouse is being built, it's important to consider the various components, such as degenerate and junk dimensions, the snowflake schema, and conformed dimensions and standard facts. These components can help ensure that the data is accurate and consistent across the data warehouse.

Finally, as the data warehouse is completed, it's important to consider the four-step process of loading, transforming, aggregating, and presenting the data. This process ensures that the data is properly structured and formatted for reporting purposes.

Overall, building a data warehouse is a complex process that involves careful planning and consideration of many different concepts. By understanding these concepts and how they fit together, students can gain a better understanding of the data warehousing process and become more effective at building data warehouses that meet the needs of their organizations.

## Conclusion

According to the interview results above, the conclusion can be drawn that at the current stage, ChatGPT cannot replace data analysts. Companies concern mostly about data privacy and

security problems. In addition, data analysts spend much time communicating and defining business needs. Knowing the business needs is a prerequisite to writing quality prompts for ChatGPT. Therefore, ChatGPT cannot perform the duties of data analysts undependably.

In the near future, large organizations may develop their own local large language models to utilize their data processing ability. It can increase their productivity. They still need data analysts who have a deep understanding of business and technology terms to write quality prompts for LLMs.

Additionally, I did a few practices on ChatGPT to find use cases to support data analysts' tasks. Without compromising data privacy, I mainly use it to write a business report and Python script for web crawling, text data mining and data visualization. I also generated examples of how professors and students in MASY program can utilize this tool. I believe it has much potential to be discovered and can be helpful to human work.

## Reference

1. Gozalo-Brizuela, R. & Garrido-Merchan, E.C. (2023). ChatGPT is not all you need. A State of the Art Review of large Generative AI models. arXiv:2301.04655 [cs.LG]. <https://doi.org/10.48550/arXiv.2301.04655>
2. Hao, K. (2022, January 10). *We read the paper that forced Timnit Gebru out of google. here's what it says.* MIT Technology Review. Retrieved April 14, 2023, from <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>
3. Heaven, W. D. (2023, March 3). Inside the story of how ChatGPT was built at OpenAI. MIT Technology Review. <https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/>
4. Majumder, E. (2022, December 18). *Chatgpt-what is it and how does it work exactly?* Medium. Retrieved April 14, 2023, from <https://medium.com/geekculture/chatgpt-what-is-it-and-how-does-it-work-exactly-62e7010524d3>
5. Marr, B. (2022, December 13). *The 10 best examples of low-code and no-code AI.* Forbes. Retrieved April 14, 2023, from <https://www.forbes.com/sites/bernardmarr/2022/12/12/the-10-best-examples-of-low-code-and-no-code-ai/?sh=350daab174b5>
6. Microsoft Corporate Blogs. (2023, February 7). *Microsoft and OpenAI Extend Partnership.* The Official Microsoft Blog. Retrieved April 14, 2023, from <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/>
7. Lund, B., & Ting, W. (2023). Chatting about chatgpt: How may AI and GPT Impact Academia and libraries? SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.4333415>
8. Shahriar, S. & Hayawi, K. (2023). Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations [Preprint]. arXiv. <https://arxiv.org/abs/2302.13817>
9. Space, K. (2023). *How ChatGPT Works Technically For Beginners.* YouTube. Retrieved March 18, 2023, from <https://www.youtube.com/watch?v=uCIa6V4uF84>.
10. Zarifhonarvar, A. (2023). Economics of chatgpt: A labor market view on the occupational impact of Artificial Intelligence. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.4350925>
11. Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L.



(2021). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8), 3453-3473. <https://doi.org/10.48550/arXiv.2302.09419>