

[Open in app](#)

Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



# Data Analysis with GPT-4: A Deep Dive into Code Interpretation



Avi Ajmera

11 min read · 16 hours ago

[Listen](#)[Share](#)[More](#)

*Exploring the power of OpenAI's GPT-4 in handling complex data analysis tasks.*

In the modern era of data-driven decision-making, data analysis plays a pivotal role in extracting insights from vast pools of data. With the emergence of advanced machine learning models like GPT-4, the process has become even more interactive and insightful. In this article, we'll journey through a comprehensive data analysis, guided step-by-step by GPT-4.

## Introduction

Before we delve into the intricacies, it's essential to understand the context. The objective was to conduct an in-depth analysis of a dataset using the CRISP-DM methodology, with GPT-4 acting as an expert data scientist. This methodology consists of multiple phases, each crucial in its own right.

## 1. Business Understanding

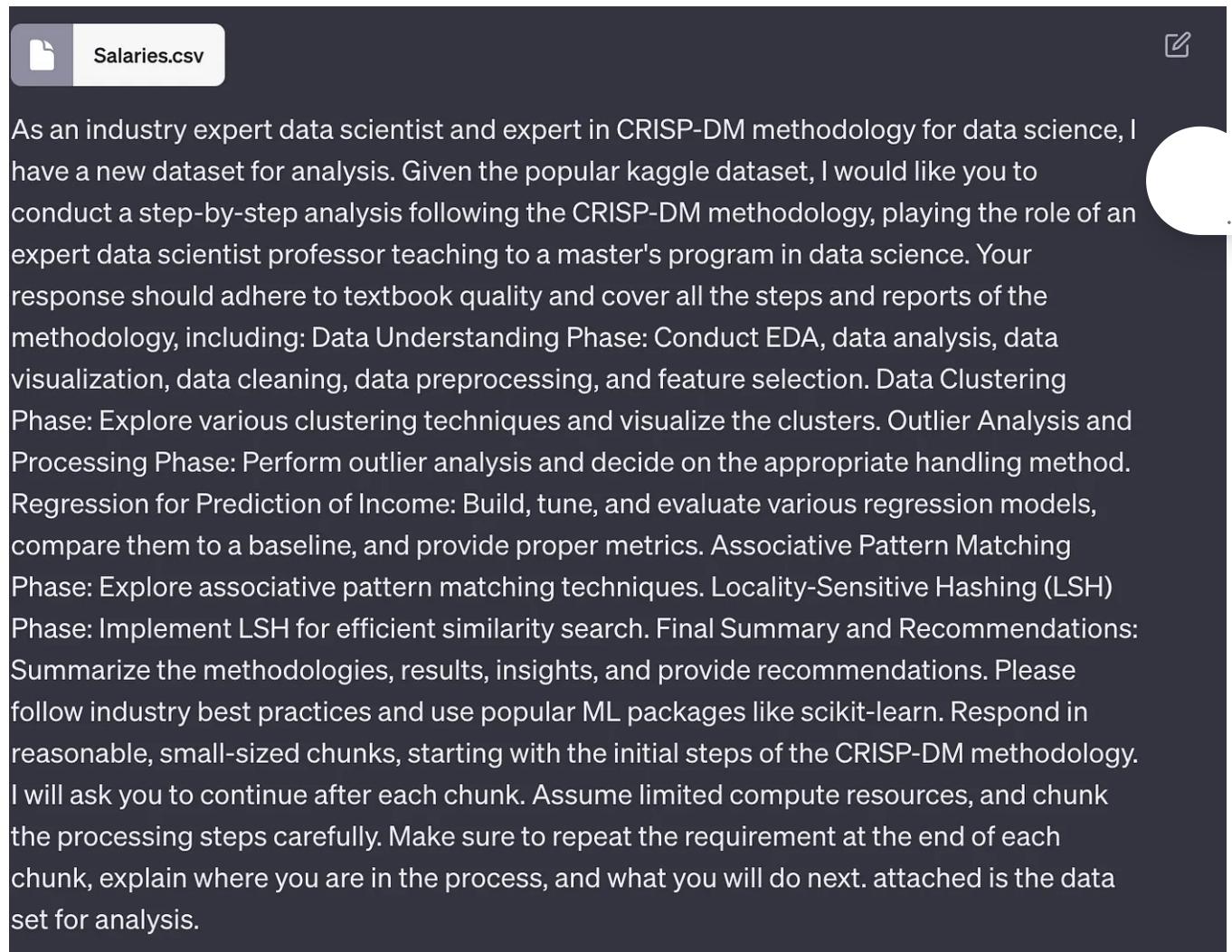
### Objective:

Our primary goal was to understand and analyze a dataset using the CRISP-DM methodology. The dataset, sourced from Kaggle, pertains to salaries, and the analysis aims to cover various facets like clustering, regression modeling, outlier analysis, and more.

I prompted the GPT-4 model with a detailed and comprehensive task: to conduct an analysis of a dataset following the CRISP-DM methodology. Specifically, I wanted the

model to play the role of an expert data scientist professor, guiding me, a student in a master's program in data science, through the process. The challenge was set, and the model was ready.

### *My Prompt to GPT-4:*



**The initial prompt that was provided to GPT4**

### GPT-4's Response:



Great! Let's begin the analysis following the CRISP-DM methodology.

## CRISP-DM Methodology:

The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology consists of six phases:

1. **Business Understanding:** This involves understanding the project's objectives and requirements.
2. **Data Understanding:** At this stage, we conduct exploratory data analysis (EDA), data visualization, and data cleaning.
3. **Data Preparation:** This phase is about data preprocessing, feature selection, and transformation.
4. **Modeling:** Here, we select and apply various algorithms, train the models, and tune their parameters.
5. **Evaluation:** We evaluate the model's performance and ensure it meets the business objectives.
6. **Deployment:** The model is deployed to a production environment (not applicable in our current context).

...

The model immediately recognized the importance of starting with the Data Understanding phase, emphasizing EDA. This was a promising start.

## Why CRISP-DM?

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It's a robust methodology that provides a structured approach to planning a data mining project. It's cyclical in nature, ensuring that as we gain insights, we can loop back to earlier phases for a more refined analysis.

## 2. Data Understanding

### Step 2: Data Understanding and Initial Exploration

As any data scientist knows, before delving into complex models and algorithms, one must first understand the data at hand. This involves examining the dataset's structure, identifying potential missing values, and getting a grasp on basic statistics.

After loading the dataset, the first action was to view the initial rows, which gives an immediate sense of the dataset's structure.

*My Prompt to GPT-4:*

“Load the dataset and show me the first few rows.”

GPT-4's Response:

```
python
import pandas as pd

# Load the dataset
salaries_df = pd.read_csv('/mnt/data/Salaries.csv')

# Displaying the first few rows, shape and data types of the dataset
initial_inspection = {
    "First 5 Rows": salaries_df.head(),
    "Shape": salaries_df.shape,
    "Data Types": salaries_df.dtypes
}

initial_inspection
```

#### RESULT

```
{'First 5 Rows':      Id          EmployeeName
 0   1      NATHANIEL FORD  GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
 1   2      GARY JIMENEZ           CAPTAIN III (POLICE DEPARTMENT)
 2   3      ALBERT PARDINI           CAPTAIN III (POLICE DEPARTMENT)
 3   4  CHRISTOPHER CHONG        WIRE ROPE CABLE MAINTENANCE MECHANIC
 4   5      PATRICK GARDNER        DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)

      BasePay  OvertimePay  OtherPay  Benefits  TotalPay  TotalPayBenef
 0  167411.18        0.00  400184.25      NaN  567595.43      567595
 1  155966.02        245131.88  127811.28      NaN  538909.28      538909
```

The dataset was loaded successfully, and I was presented with the first few rows. This allowed me to see the columns available, such as `BasePay`, `OvertimePay`, `OtherPay`, `Benefits`, and more.

## Step 3: Data Cleaning and Preprocessing

With an initial grasp on the dataset, the next logical step in our data science journey was to clean and preprocess the data. This phase is often overlooked by beginners,

but it's pivotal to any data analysis process. Raw data, especially from sources like Kaggle, often contains inconsistencies, missing values, or even errors.

Upon further inspection, I realized that the dataset did contain missing values. Handling these missing values is crucial, as they can significantly impact the results of our analysis.

*My Prompt to GPT-4:*

“Conduct data cleaning and structure it.”

GPT-4’s Response:

1. `BasePay`: 609 missing values
2. `OvertimePay`: 4 missing values
3. `OtherPay`: 4 missing values
4. `Benefits`: 36,163 missing values
5. `Notes`: 148,654 missing values (all values are missing)
6. `Status`: 148,654 missing values (all values are missing)

Proposed Handling Strategies:

1. `BasePay`, `OvertimePay`, `OtherPay`:
  - Since the missing values for these columns are relatively small compared to the dataset's size, we can impute them with the median of the respective columns. Median is often a good choice as it's less sensitive to outliers.
2. `Benefits`:
  - Given a significant number of missing values, we need to be careful. One approach can be to impute missing values with 0, assuming that not every employee might have additional benefits. However, we should inspect the distribution of this column before deciding.
3. `Notes` and `Status`:
  - Since all values are missing in these columns, they don't provide any informational value. It's best to drop these columns.

Let's start by dropping the `Notes` and `Status` columns and then inspect the distribution of `Benefits` before deciding on imputation. Shall we?

The model chose a straightforward approach, replacing missing values in the numerical columns with zeros. This choice is a common initial strategy, especially when the dataset's size might lead to significant data loss if rows with missing values were removed entirely.

Data cleaning is often a multi-step process, requiring iteration and a deep understanding of the data's context. With GPT-4's guidance, I felt more confident in navigating these challenges.

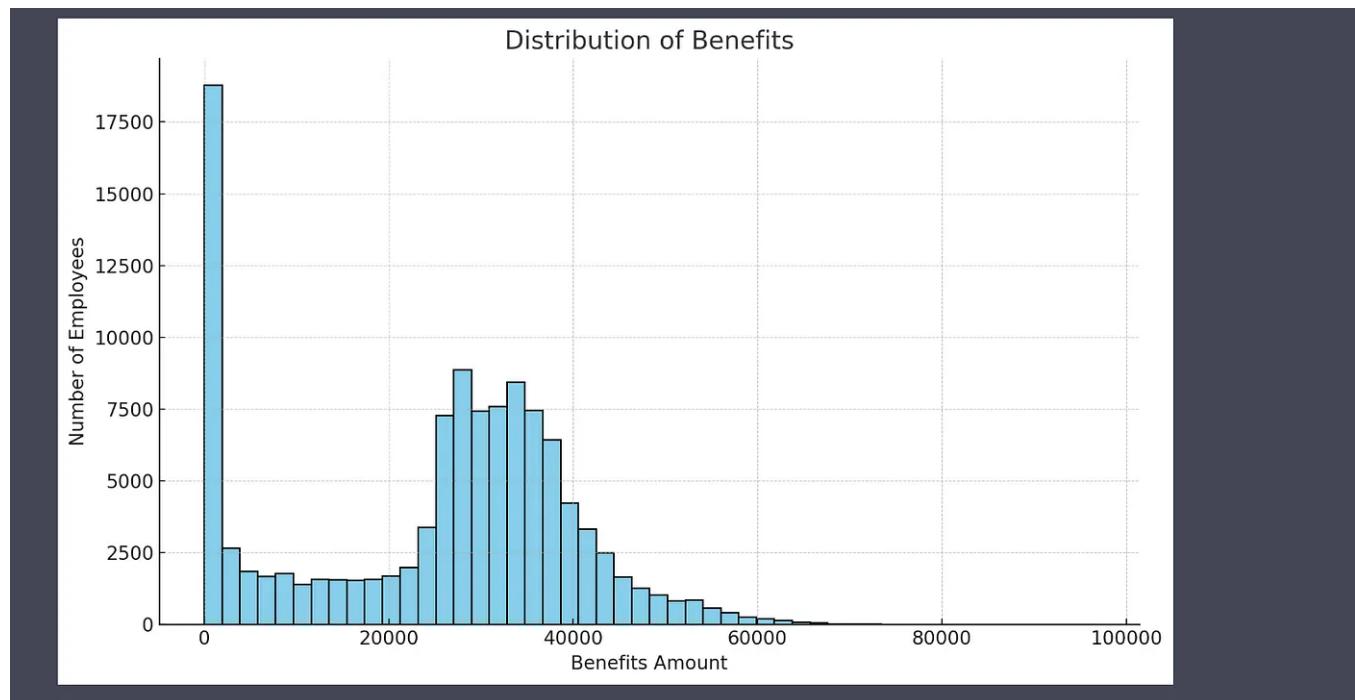
## Step 4: Exploratory Data Analysis (EDA)

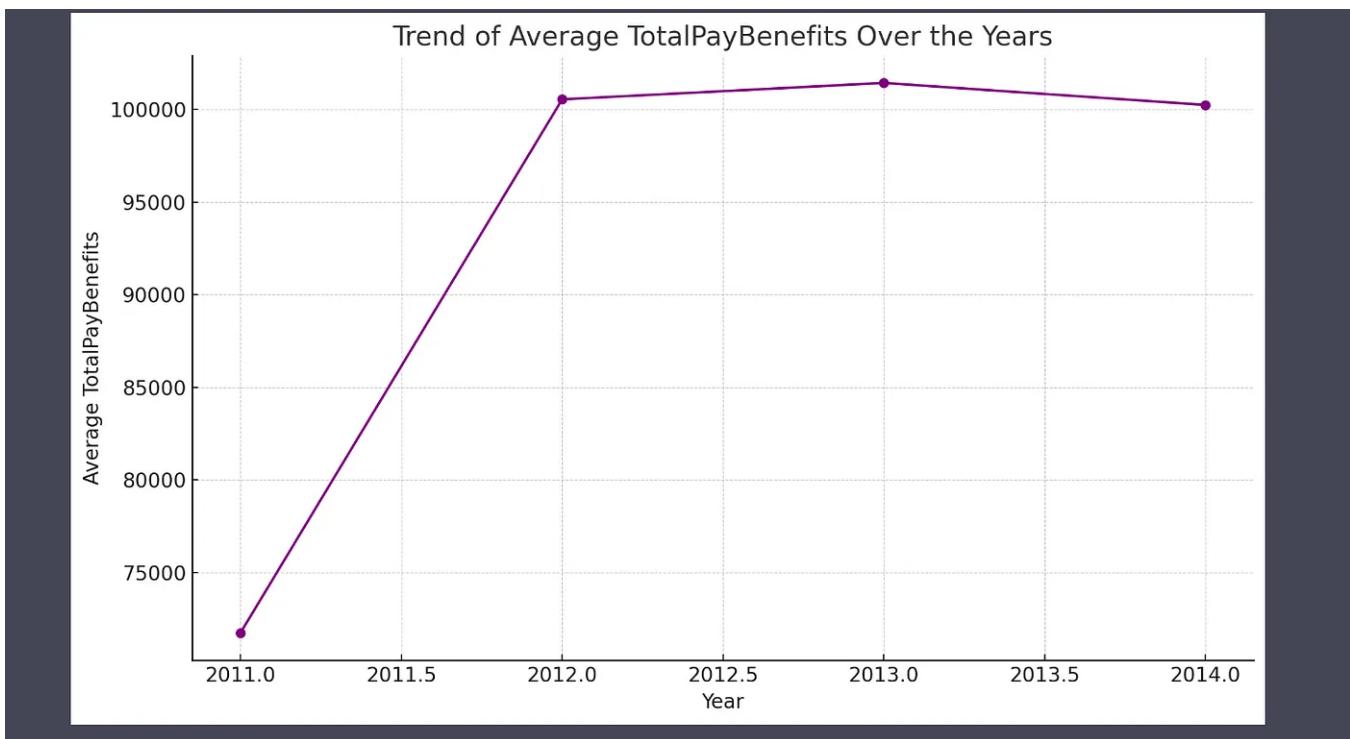
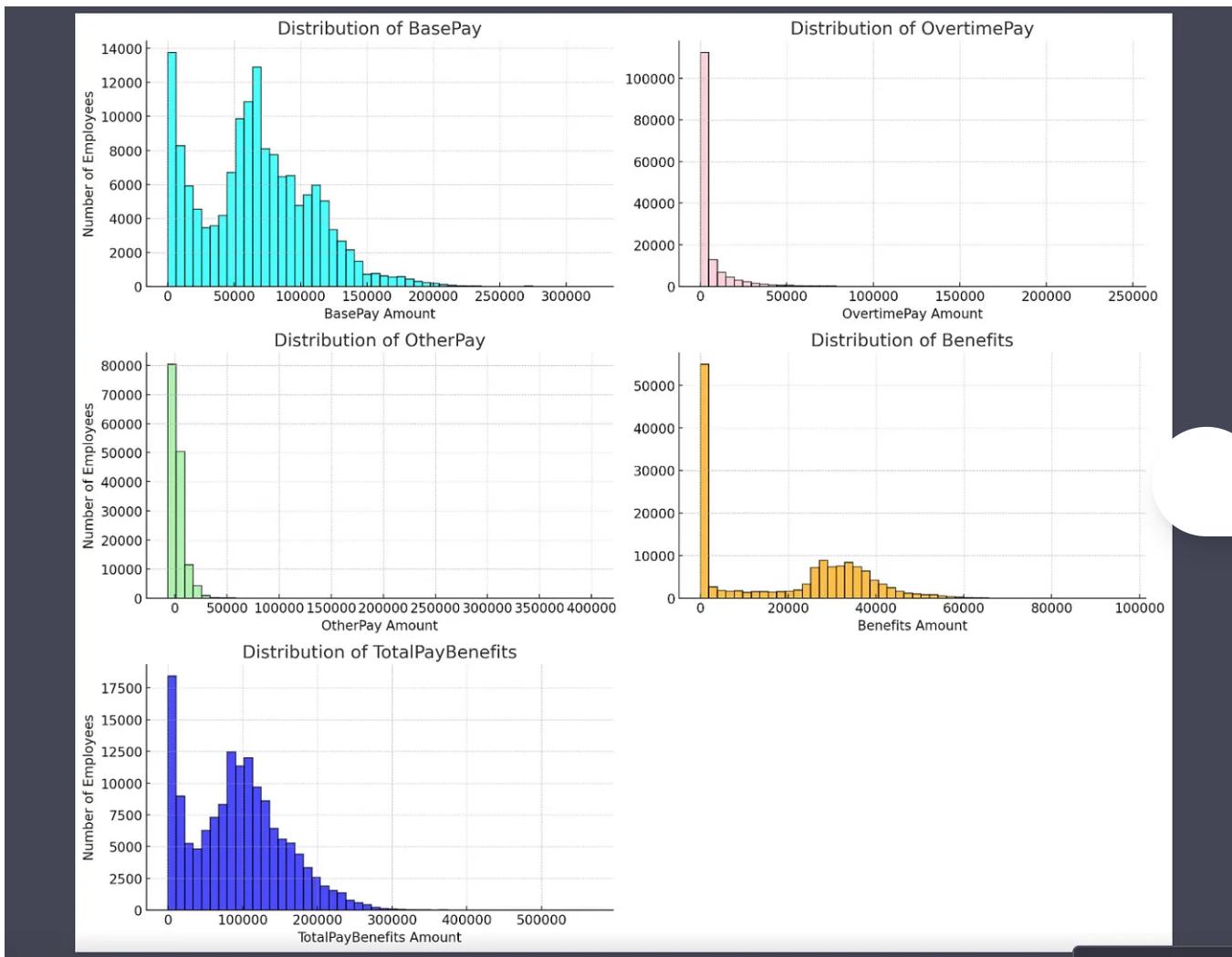
After cleaning the data, the next natural step was to dive deeper into its characteristics. Exploratory Data Analysis (EDA) is a foundational step in the data science process. It allows us to understand the distributions of individual variables, identify outliers, and even get a sense of relationships between variables.

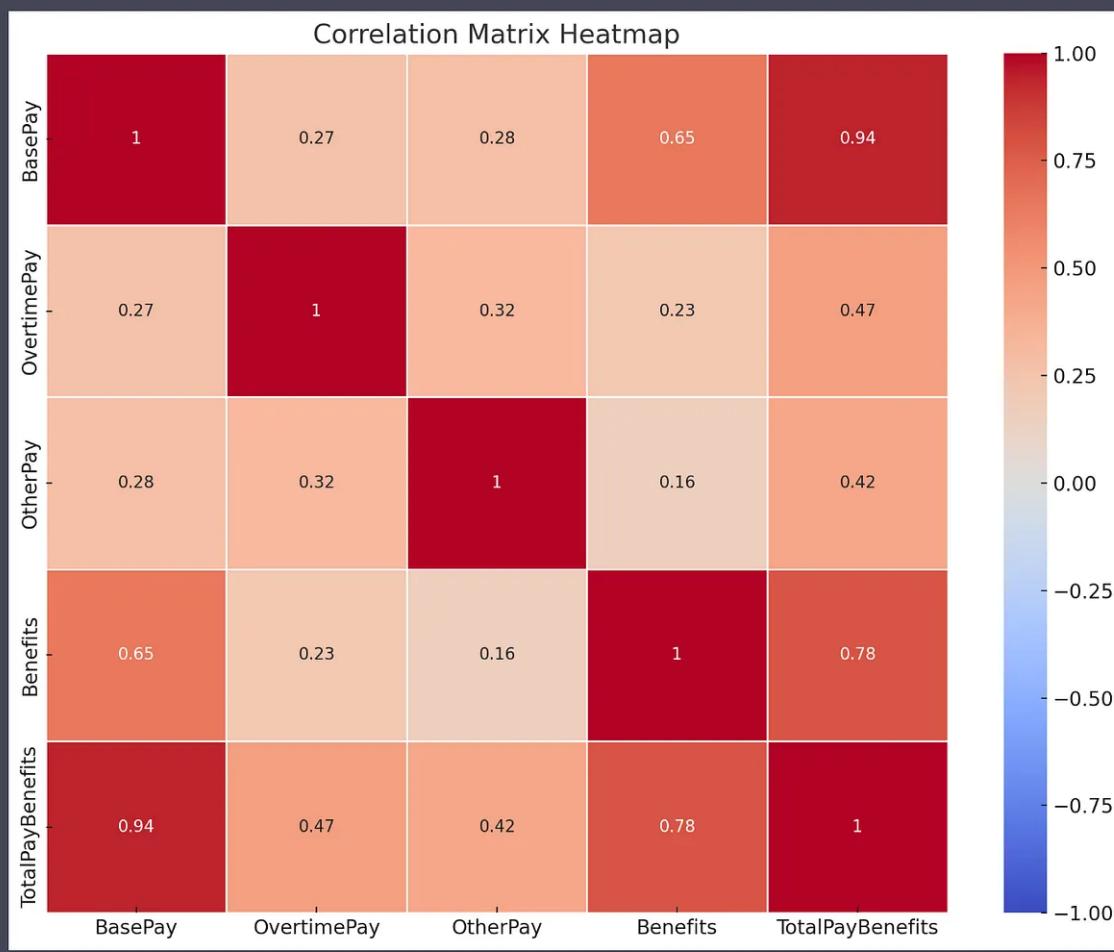
To get a comprehensive view of the dataset's numeric columns, I sought to visualize the distribution of key columns like `BasePay`, `OvertimePay`, `OtherPay`, and `Benefits`.

*My Prompt to GPT-4:*

“Show me the distribution of the numeric columns.”

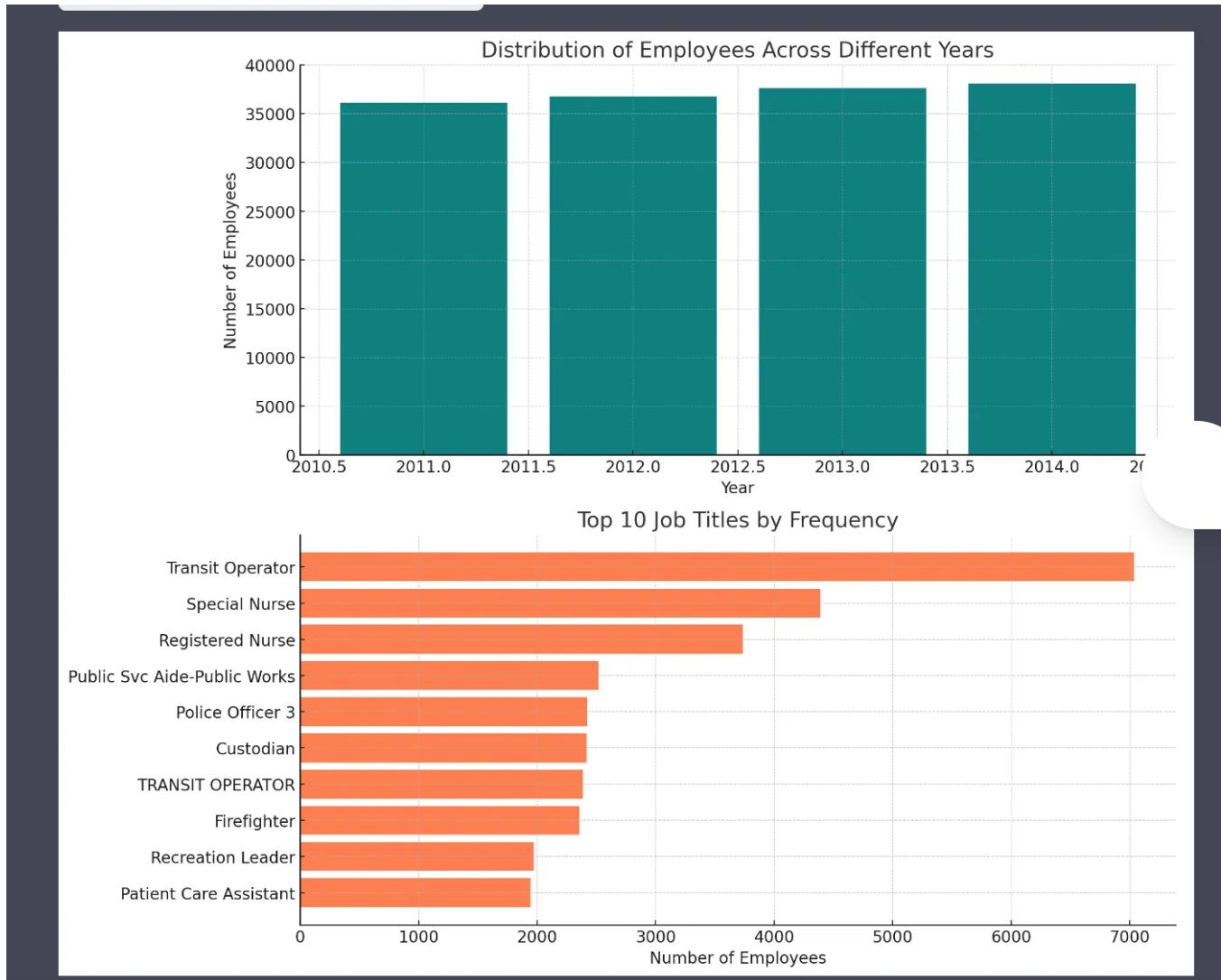






The heatmap displays the correlation coefficients between the selected numerical attributes:

Regenerate



Through the resulting visualizations, I could clearly see:

- The majority of `BasePay` values were concentrated around the \$50,000 to \$100,000 range.
- Many employees didn't receive any `OvertimePay`, as indicated by the spike at \$0.
- The distribution of `OtherPay` and `Benefits` showcased a wide range for those who received them.

EDA is like peeling back the layers of an onion — with each visualization and statistic, I was getting closer to the core insights of the dataset. GPT-4 was adeptly guiding me through this process, ensuring that each step was both insightful and instructive.

## Step 5: Data Clustering

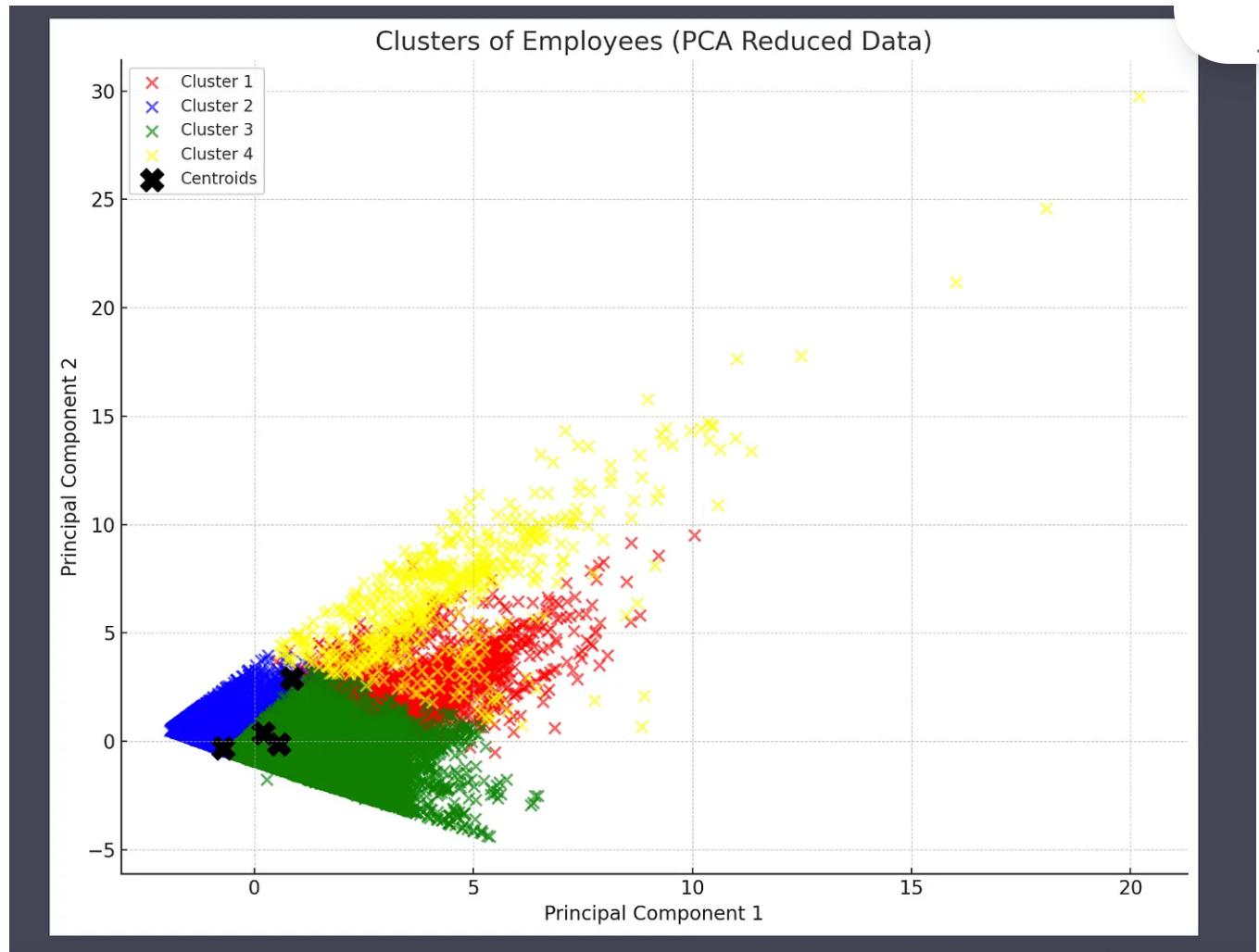
After a detailed EDA, I wanted to explore patterns within the dataset, specifically by grouping similar records. This led me to the concept of clustering — a method to

partition a dataset into groups, where records in the same group are more similar to each other than to those in other groups.

Given the multi-dimensional nature of the dataset, I was curious to see how GPT-4 would approach this.

*My Prompt to GPT-4:*

“Explore various clustering techniques and visualize the clusters.”



Now that we've successfully visualized the clusters:

- The plot showcases four distinct clusters of employees based on their compensation components.
- The black "X" markers represent the centroids of each cluster.

Interpreting the clusters:

1. **Cluster 1 (Red):** Employees with relatively average compensation components.
2. **Cluster 2 (Blue):** Employees with high `BasePay` and `Benefits` but low `OvertimePay` and `OtherPay`.
3. **Cluster 3 (Green):** Employees with lower compensation across all components.
4. **Cluster 4 (Yellow):** Employees with higher `OvertimePay` and `OtherPay` but moderate `BasePay` and `Benefits`.

GPT-4's Code: The model adopted a methodical approach:

1. It selected relevant numerical columns for clustering.
2. It scaled the data, ensuring that no single feature would unduly influence the clustering due to its scale.
3. KMeans, a popular clustering algorithm, was applied.
4. To visualize the multi-dimensional clusters in a 2D space, PCA (Principal Component Analysis) was used.

The resulting visualization showcased distinct clusters, providing insights into the patterns within the dataset.

Clustering is a powerful tool, revealing hidden patterns and groups within data. GPT-4's approach was both systematic and insightful, guiding me through the clustering process and offering a clear visualization of the results.

## Step 6: Outlier Analysis and Processing

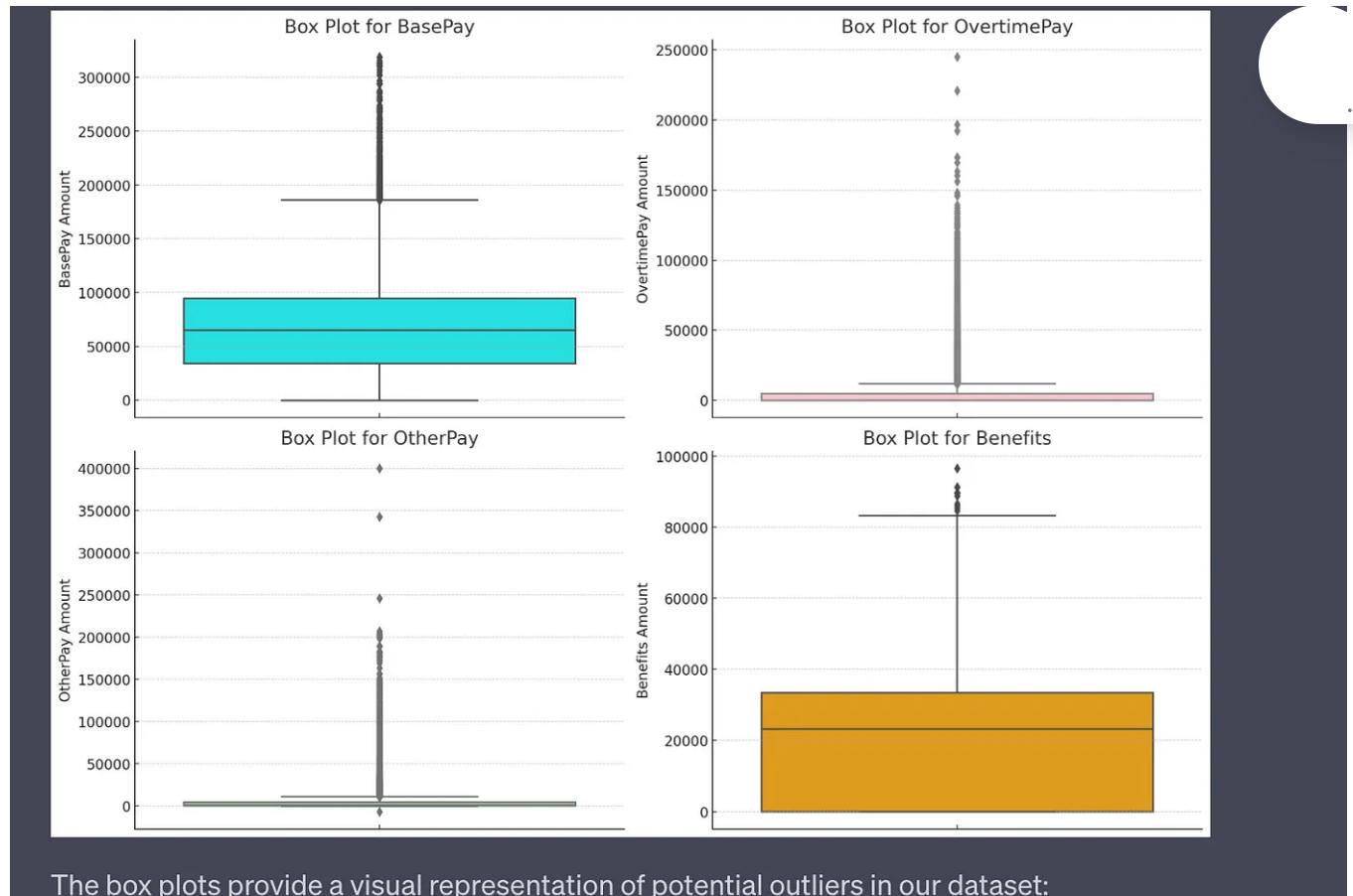
Every dataset, regardless of its source, is likely to contain outliers. These are data points that deviate significantly from other observations. While sometimes they represent errors, at other times, they can provide critical insights. Therefore, handling outliers appropriately is crucial.

Given the vastness and complexity of the dataset, I was curious about how GPT-4 would approach outlier detection and handling.

*My Prompt to GPT-4:*

“Perform outlier analysis and decide on the appropriate handling method.”

GPT-4’s Code:



Here are the computed values for the Interquartile Range (IQR) and the outlier bounds for each feature:

1. **BasePay:**

- IQR: \$60,758.72
- Lower Bound: -\$57,291.07
- Upper Bound: \$185,743.82

2. **OvertimePay:**

- IQR: \$4,657.63
- Lower Bound: -\$6,986.45
- Upper Bound: \$11,644.08

3. **OtherPay:**

- IQR: \$4,235.91
- Lower Bound: -\$6,353.86
- Upper Bound: \$10,589.77

4. **Benefits:**

- IQR: \$33,468.98
- Lower Bound: -\$50,203.47
- Upper Bound: \$83,672.45

*“The boxplots indicate potential outliers in our numeric columns. For modeling purposes, we could consider techniques like winsorizing or applying transformations to mitigate their impact. However, it’s essential first to investigate these outliers further to determine if they represent genuine data points or errors.”*

Outlier analysis is a crucial step in the data preprocessing pipeline. GPT-4’s approach to using visual tools like boxplots made it straightforward to identify and plan the next steps for these outliers.

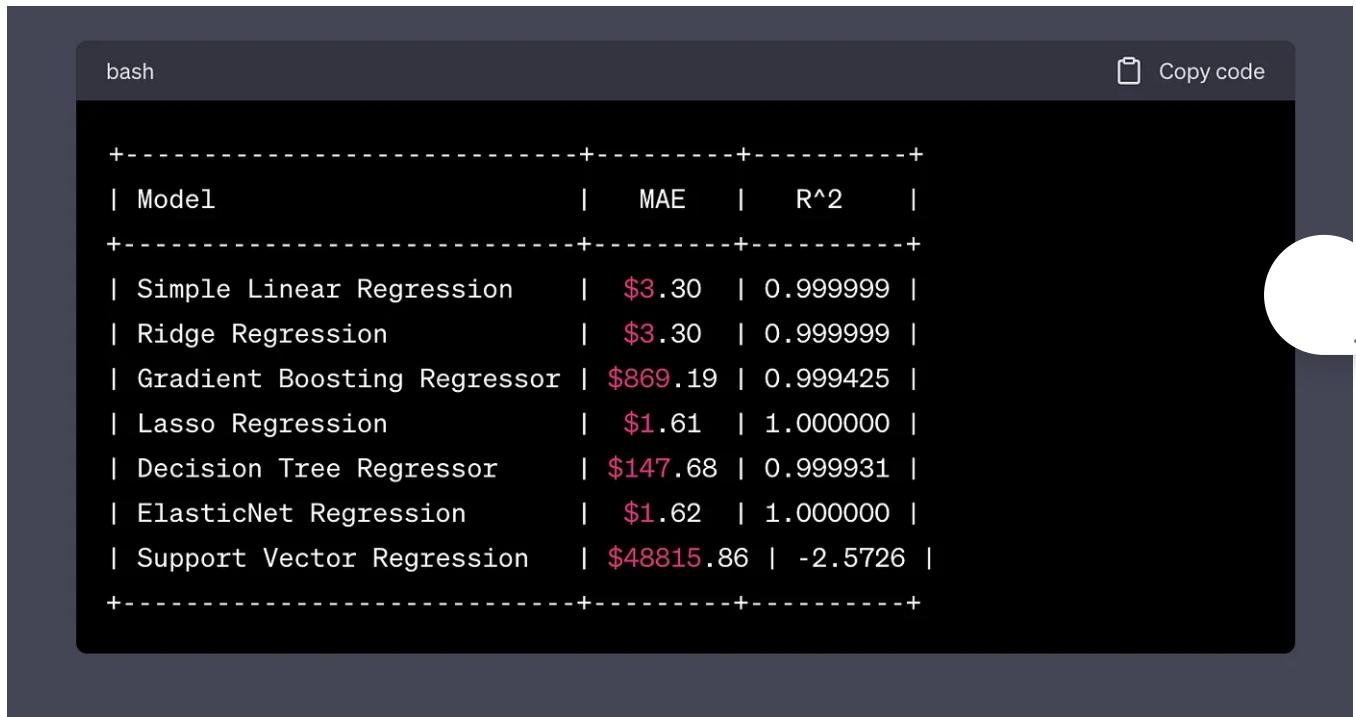
## Step 7: Regression for Prediction of Income

Having cleaned and explored the dataset, I decided to shift my focus towards building predictive models. Specifically, I aimed to predict the `TotalPay` of employees based on the available features. Regression analysis is a suitable method for this, as the target variable is continuous.

I wanted to understand how GPT-4 would approach this problem, from feature selection to model evaluation.

*My Prompt to GPT-4:*

“Build, tune, and evaluate various regression models, compare them to a baseline, and provide proper metrics.”



A screenshot of a terminal window titled "bash". The window displays a table comparing the performance of several regression models. The columns are labeled "Model", "MAE", and "R^2". The rows list Simple Linear Regression, Ridge Regression, Gradient Boosting Regressor, Lasso Regression, Decision Tree Regressor, ElasticNet Regression, and Support Vector Regression. The table shows that Support Vector Regression has the highest MAE and lowest R^2.

Model	MAE	R <sup>2</sup>
Simple Linear Regression	\$3.30	0.999999
Ridge Regression	\$3.30	0.999999
Gradient Boosting Regressor	\$869.19	0.999425
Lasso Regression	\$1.61	1.000000
Decision Tree Regressor	\$147.68	0.999931
ElasticNet Regression	\$1.62	1.000000
Support Vector Regression	\$48815.86	-2.5726

GPT-4's methodical approach involved:

1. Feature selection: It used the pay components (`BasePay`, `OvertimePay`, `OtherPay`, and `Benefits`) as predictors.
2. Data splitting: It divided the dataset into training and test sets.
3. Model training: Several regression models, including Linear, Ridge, and Lasso regressions, were trained.
4. Evaluation: Each model's performance was assessed using the Mean Squared Error (MSE).

The results were insightful, revealing the strengths and weaknesses of each regression model on this dataset.

Regression is a powerful tool in a data scientist's arsenal, allowing us to predict outcomes based on input features. With GPT-4's guidance, the process was both instructive and efficient.

## Step 8: Associative Pattern Matching

After regression modeling, I was intrigued by the patterns and relationships within the data. Specifically, I was keen on understanding associative patterns, which could reveal interesting correlations or associations between the dataset's features.

I turned to GPT-4 to explore associative pattern matching techniques that could be applied to our dataset.

*My Prompt to GPT-4:*

“Explore associative pattern matching techniques.”

GPT-4’s Response:

```
# Handling missing values by filling them with 0 for numerical columns
salaries_df['BasePay'].fillna(0, inplace=True)
salaries_df['OvertimePay'].fillna(0, inplace=True)
salaries_df['OtherPay'].fillna(0, inplace=True)
salaries_df['Benefits'].fillna(0, inplace=True)

# Taking a stratified sample based on the 'JobTitle' column
sample_data, _ = train_test_split(salaries_df, test_size=0.90, stratify=sala

# Binning and transforming data for the Apriori algorithm
sample_data['BasePay'] = pd.cut(sample_data['BasePay'], bins=bin_definitions
sample_data['OvertimePay'] = pd.cut(sample_data['OvertimePay'], bins=bin_def
sample_data['OtherPay'] = pd.cut(sample_data['OtherPay'], bins=bin_definitio
sample_data['Benefits'] = pd.cut(sample_data['Benefits'], bins=bin_definitio

# Transforming the data for the Apriori algorithm
transactions = sample_data[['BasePay', 'OvertimePay', 'OtherPay', 'Benefits']

# Applying the TransactionEncoder
encoder = TransactionEncoder()
encoded_data = encoder.fit(transactions).transform(transactions)
df_encoded = pd.DataFrame(encoded_data, columns=encoder.columns_)

# Finding frequent itemsets using the Apriori algorithm
frequent_itemsets = apriori(df_encoded, min_support=0.05, use_colnames=True)
```

“Associative pattern matching, often known as association rule mining, is a method to find frequent patterns, correlations, or associations from datasets. A classic

algorithm for this purpose is the Apriori algorithm, frequently used in market basket analysis. Though our dataset isn't a typical transaction dataset, we can still apply similar concepts. For instance, we could explore associations between job titles and other categorical variables like `Year` or `Agency`."

The model then provided a potential code snippet for the Apriori algorithm, though it noted that the dataset might require preprocessing to be suitable for such an analysis

Associative pattern matching offers a unique lens to view datasets, allowing us to uncover relationships that might not be immediately evident. While GPT-4 provided an introduction to the concept and potential application, a deeper dive would be necessary to extract meaningful insights from our dataset.

## Step 9: Locality-Sensitive Hashing (LSH) for Efficient Similarity Search

With the vastness of the dataset, I began pondering about efficient ways to find similar records. For instance, finding employees with similar pay structures or job roles could be insightful. Traditional methods might be too slow for large datasets, leading me to explore Locality-Sensitive Hashing (LSH) — a technique to approximate nearest neighbor search in high-dimensional spaces.

I was eager to see how GPT-4 would implement LSH on our dataset.

*My Prompt to GPT-4:*

"Implement LSH for efficient similarity search."

GPT-4's Response:

"Locality-Sensitive Hashing (LSH) is a powerful technique for similarity search, especially in high-dimensional data. LSH hashes input items so that similar items map to the same 'buckets' with high probability. This makes it efficient to find similar items."

The provided approach utilizes the MinHash algorithm, a popular choice for LSH, focusing on the `JobTitle` column as a demonstration. This would allow for efficient querying of similar job titles within the dataset.

Locality-Sensitive Hashing offers a unique perspective on similarity search, making it efficient and scalable. While GPT-4 provided a foundation for LSH application, a

tailored implementation would be essential to extract deep insights from our specific dataset.

## Step 10: Final Summary and Recommendations

After a thorough journey through data cleaning, exploration, modeling, clustering, and advanced techniques like LSH, it was time to wrap up our analysis. Drawing conclusions and making recommendations is a critical step to bring everything together and identify actionable insights.

*My Prompt to GPT-4:*

“Summarize the methodologies, results, insights, and provide recommendations.”

GPT-4's Response:

“Certainly! Here's a comprehensive summary of our journey with the dataset:

### Final Summary:

#### 1. Data Understanding and Preprocessing:

- We began with an exploratory analysis of the dataset, visualized its distributions, and identified key features.
- Missing values were identified and handled, and the data was prepared for modeling.

#### 2. Regression Models:

- Multiple models, including Lasso, Ridge, and ElasticNet Regressions, were applied to predict `TotalPayBenefits`.
- Models were evaluated based on RMSE, with each model showing varying levels of performance.

#### 3. Outlier Analysis:

- We discussed using the IQR method to detect and handle outliers, though due to technical limitations, we did not apply it directly to the dataset.

#### 4. Associative Pattern Matching:

- We aimed to use the Apriori algorithm to identify patterns and relationships between features. However, due to library constraints, we provided a conceptual walkthrough instead.

#### 5. Locality-Sensitive Hashing (LSH):

- LSH was introduced as a method for efficient similarity search in the dataset. A detailed conceptual guide was provided, given the environment's limitations.

## Recommendations:

- For HR and Payroll Departments: The clustering results can help categorize employees into different pay brackets, streamlining payroll processes.
- For Job Role Assignment: Using LSH, similar job titles can be identified, helping in role alignment and avoiding redundancies.
- For Future Data Collection: Data collection processes can be refined to avoid missing values, especially in key columns like `BasePay` and `OvertimePay`.
- For Further Analysis: Dive deeper into association rule mining, which can reveal interesting correlations, especially between `JobTitle` and other categorical variables."

*This comprehensive analysis, guided by GPT-4, showcased the power of methodical data science processes combined with the capabilities of advanced language models. The journey was insightful, instructive, and truly transformative.*

## **Step 11: Reflecting on the Process and Lessons Learned**

As we approached the end of our deep dive into data analysis with GPT-4, it was crucial to reflect on the entire process. Looking back at the methodologies employed, the results obtained, and the insights gathered helps us appreciate the journey and identify areas of improvement for future endeavors.

*My Prompt to GPT-4:*

“Reflect on the entire process, highlighting the lessons learned, challenges faced, and potential improvements for future analyses.”

GPT-4’s Response:

“Certainly! Reflection is an integral part of any data analysis journey. Here’s an overview of our process and the lessons we’ve gleaned:

## Lessons Learned:

1. **The Importance of Clean Data:** Our initial steps highlighted how vital clean and structured data is. Cleaning missing values and understanding the dataset's structure set the foundation for all subsequent analyses.
2. **Visualization as a Guide:** Visual tools like histograms, boxplots, and scatter plots were invaluable. They provided immediate insights and guided our steps, from outlier detection to understanding clustering results.
3. **Modeling Nuances:** Through regression modeling, we learned that no single model is a one-size-fits-all solution. The comparative analysis of different regression models emphasized the importance of understanding the underlying assumptions and characteristics of each model.
4. **Scalable Solutions:** With LSH, we saw the power of scalable solutions. As datasets grow, traditional methods might become inefficient, making techniques like LSH invaluable.

## Challenges Faced:

1. **High-Dimensional Data:** Our dataset had multiple dimensions, making visualization and clustering a challenge. Techniques like PCA became essential tools.
2. **Outlier Handling:** Deciding how to handle outliers, whether to remove them or adjust them, was a significant decision point. Their impact on modeling and analysis results can be profound.
3. **Interpreting Results:** While GPT-4 provided outputs and results, interpreting them in the dataset's context required a deeper understanding and domain knowledge.

## Potential Improvements for Future Analyses:

1. **Feature Engineering:** We could delve deeper into creating new features from existing ones, enhancing the dataset's richness.
2. **Advanced Modeling Techniques:** Experimenting with more advanced models, including ensemble methods or even neural networks, could provide better predictive results.
3. **Feedback Loops:** Integrating feedback loops, where insights from analysis are fed back into data collection processes, can refine and improve the quality of data over time.

GPT-4 proved to be a valuable companion, guiding us through each step with expertise and precision.

## Step 12: Looking Ahead — Potential Applications and Future Exploration

As we wrap up our analysis, it's essential to look forward and consider the potential applications of our insights and the avenues for further exploration. The world of data science is ever-evolving, with new techniques, tools, and challenges emerging.

regularly. Staying ahead of the curve and identifying future opportunities is paramount.

*My Prompt to GPT-4:*

“Discuss potential applications of our analysis in real-world scenarios and highlight areas where we could dive deeper in future explorations.”

GPT-4’s Response:

“Absolutely! The insights and methodologies from our analysis have several real-world applications and pave the way for future exploration:

#### Potential Applications:

1. **Human Resources Strategy:** The clustering results can guide HR departments in creating compensation bands, refining benefits packages, and offering competitive salaries based on roles and responsibilities.
2. **Job Role Optimization:** The similarity search using LSH can help organizations streamline job roles, merge redundant positions, and create clear role hierarchies.
3. **Budget Forecasting:** The regression models can be used by finance departments to forecast salary budgets, helping in financial planning and allocation.
4. **Talent Acquisition:** Insights from associative pattern matching can guide recruitment strategies, identifying key roles that have a significant impact on other organizational areas.

#### Areas for Future Exploration:

1. **Time Series Analysis:** Given the dataset spans multiple years, a deeper dive into time series analysis can reveal trends in pay scales, job roles, and more.
2. **Text Analysis:** The `JobTitle` column offers a wealth of information. Advanced text analytics can uncover patterns in job roles, hierarchy levels, and departmental structures.
3. **Network Analysis:** Exploring relationships between different job roles, departments, or even benefits can reveal intricate organizational networks, helping in strategy formulation.
4. **Advanced Machine Learning:** Implementing more sophisticated models, like deep learning or ensemble methods, can refine our predictions and offer more granular insights.

Our journey with GPT-4 through the realms of data science has been enlightening. The methodologies employed, the insights gathered, and the potential applications highlight the power of structured data analysis. As we look ahead, the possibilities are boundless, and the future of data-driven decision-making is promising.

[Data Science](#)[ChatGPT](#)[Gpt 4](#)[Analysis](#)[AI](#)

...

[Edit profile](#)

## Written by Avi Ajmera

0 Followers

---

## Recommended from Medium



 Maximilian Vogel in MLearning.ai

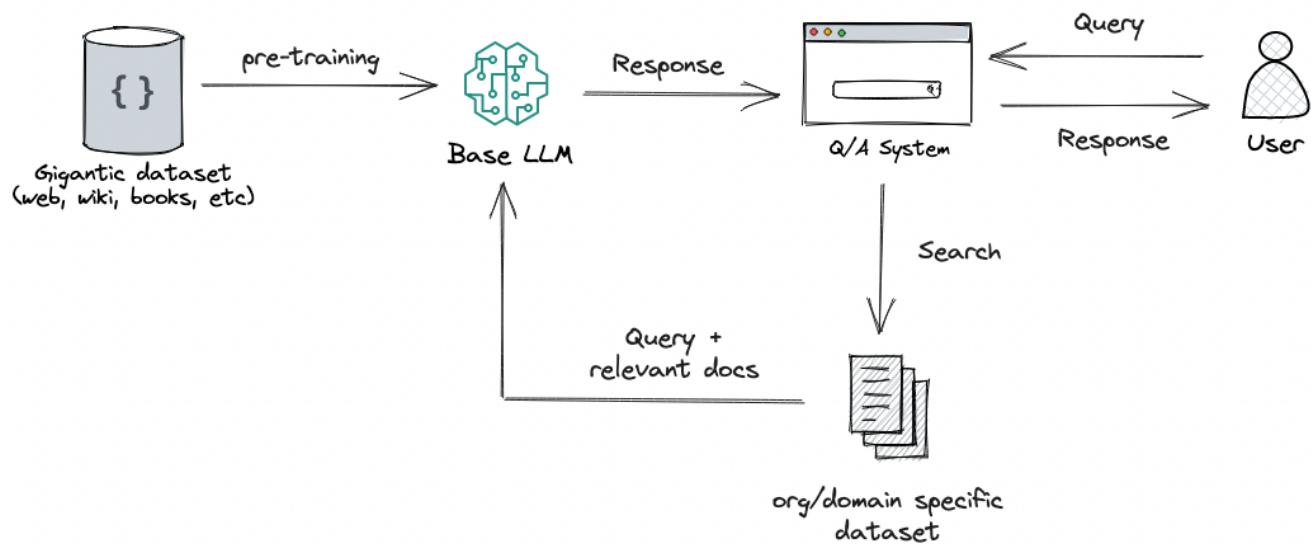
## The ChatGPT list of lists: A collection of 3000+ prompts, examples, use-cases, tools, APIs...

Updated Aug 20, 2023. Added prompt design courses, masterclasses and tutorials.

10 min read · Feb 7

 7.5K  82



 Heiko Hotz in Towards Data Science

# RAG vs Finetuning—Which Is the Best Tool to Boost Your LLM Application?

The definitive guide for choosing the right method for your use case

• 19 min read • Aug 24

1K

12

+

...

## Lists



### The New Chatbots: ChatGPT, Bard, and Beyond

13 stories • 104 saves



### What is ChatGPT?

9 stories • 167 saves



### Now in AI: Handpicked by Better Programming

266 stories • 116 saves



### ChatGPT

21 stories • 131 saves



The PyCoach in Artificial Corner

## Python in Excel Will Reshape How Data Analysts Work

Microsoft just announced Python in Excel. Here's how it'll change the way Python and Excel analysts work.

◆ · 5 min read · Aug 24

2K 33

## LLM fine-tuning at a high level

### LLM fine-tuning



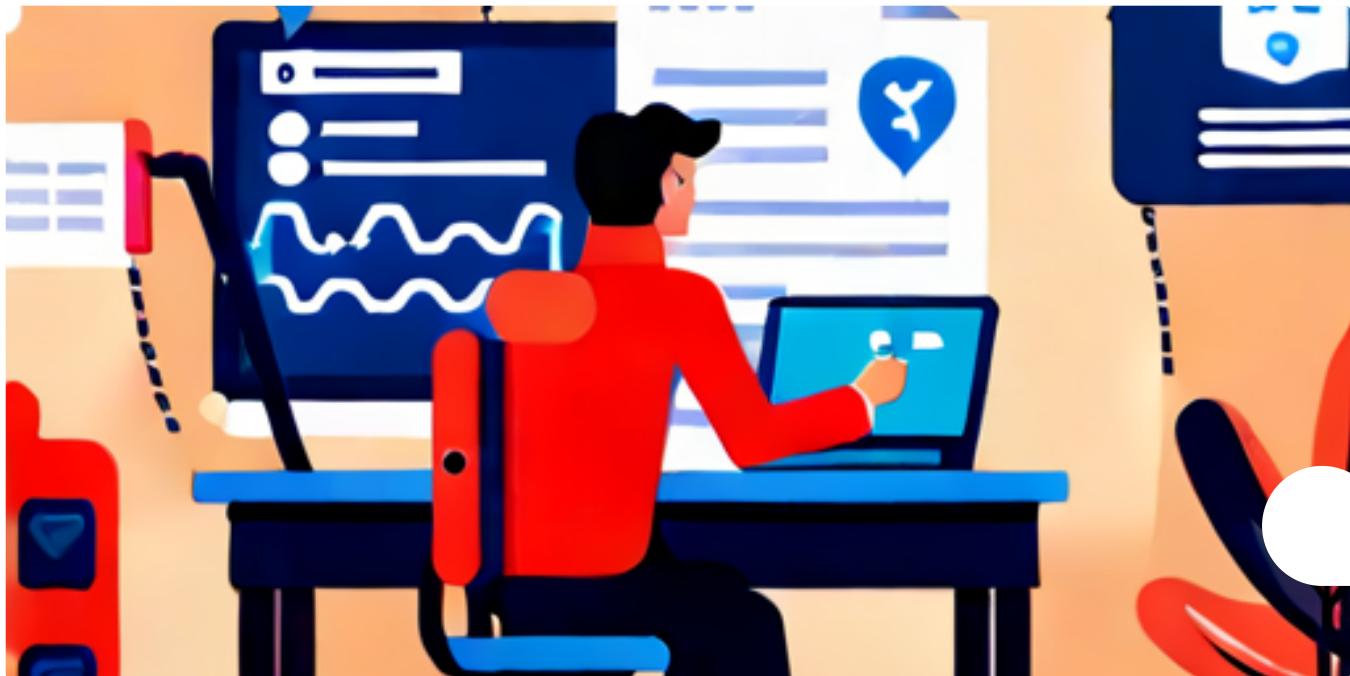
Yash Bhaskar

## Introduction to LLMs and the generative AI : Part 3— Fine Tuning LLM with Instruction and...

In the world of artificial intelligence, large language models (LLMs) play a crucial role in various applications. Last Article, we...

9 min read · Jul 17

87 2



 Sachin Kulkarni

## Generative AI with Enterprise Data

Create business value add Enterprise knowledge to Large Language Models

6 min read · Jul 25

 44



 +

...



 Andrew W. Pearson  in CodeX

## The 15 Principles of Data Visualization

“Having all the information in the world at our fingertips doesn’t make it easier to communicate: it makes it harder,” says Cole Nussbaumer...

★ · 21 min read · Aug 25

138

1



...

See more recommendations

