5004CMD Data Science  
CW: Big Data Programming Project  
Title: Analysis of US Trip Data Using Big Data Techniques

Date of submission: February 13, 2025  
Student ID : 13122803

Student Name : Elufowoju David

Abstract  
This project analyzes large-scale mobility data from the Bureau of Transportation Statistics (BTS) using big data analytics techniques. The aim is to reduce processing time for mobility statistics compared to serial processing. The study investigates patterns in home stays and trip distances, identifies dates with high trip volumes, and compares computational efficiency using parallel processing. A model is developed to simulate travel frequency based on trip length. Results demonstrate the effectiveness of parallel computing in reducing analysis time for large datasets.

Table of Contents  
[Insert table of contents here]

**Table of Contents**

**Abstract**

**Introduction**

1.1 Problem Description

1.2 Project Aim

1.3 Purpose of Research Project

1.4 Project Objectives

1.5 Hypothesis

**Project Methodology**

2.1 Data Acquisition

2.2 Data Pre-processing

2.3 Data Cleaning

2.4 Data Categorization

2.5 Parallel Computing

2.6 Data Processing

2.7 Data Fitting – Model Selection

2.8 Model Testing

2.9 Data Analysis

2.10 Data Visualization

**Results**

3.1 Parallel Processing

3.2 Data Visualization

3.3 Data Analysis

**Discussion**

4.1 Data Communication

4.2 Data Interpretation

4.3 Future Recommendations

4.4 Model Training

4.5 Model Selection

**Project Management**

5.1 Risk Management

5.2 Gantt Chart

**Conclusion**

**References**

**Appendices**

A.1 Logbook (Weekly Progress Summary)

A.2 Flowchart

A.3 Pseudocode

A.4 Version Control

List of Tables

2.6 Data Processing Table

4.4 Model Training Table  
[Insert list of tables here]

List of Figures  
[Insert list of figures here]

Abbreviations  
dd - Dask DataFrame  
pd - Pandas DataFrame

Introduction

1.1 Problem Description  
The Bureau of Transportation Statistics (BTS) has developed a new analysis technique for mobility statistics that is computationally intensive. This project aims to leverage big data analytics to reduce processing time and enable more efficient analysis of large-scale trip data.

1.2 Project Aim  
The project aims to analyze US trip data to understand patterns in home stays and travel distances, identify high-volume travel dates, and develop a model for travel frequency prediction. Additionally, it seeks to demonstrate the efficiency gains of parallel processing over serial processing for large datasets.

1.3 Purpose of Research Project  
This research aims to improve the efficiency of mobility data analysis for the BTS, enabling faster insights and better decision-making in transportation planning and policy.

1.4 Project Objectives

Calculate the average number of people staying at home per week.

Determine the average travel distance for people not staying at home.

Identify dates with over 10,000,000 trips in specified distance ranges.

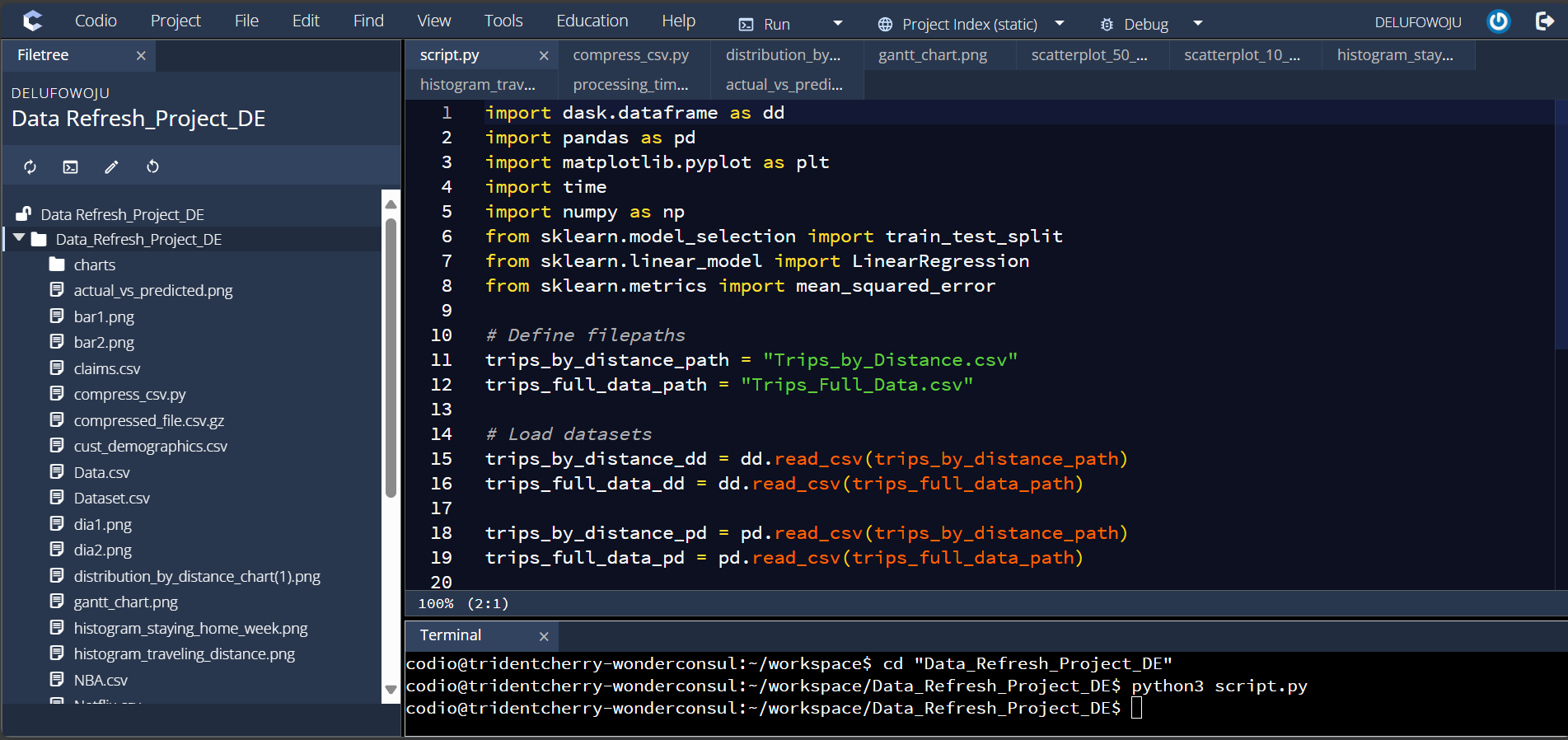
Compare the computational efficiency of parallel processing using 10 and 20 processors.

Develop a predictive model for travel frequency based on trip length.

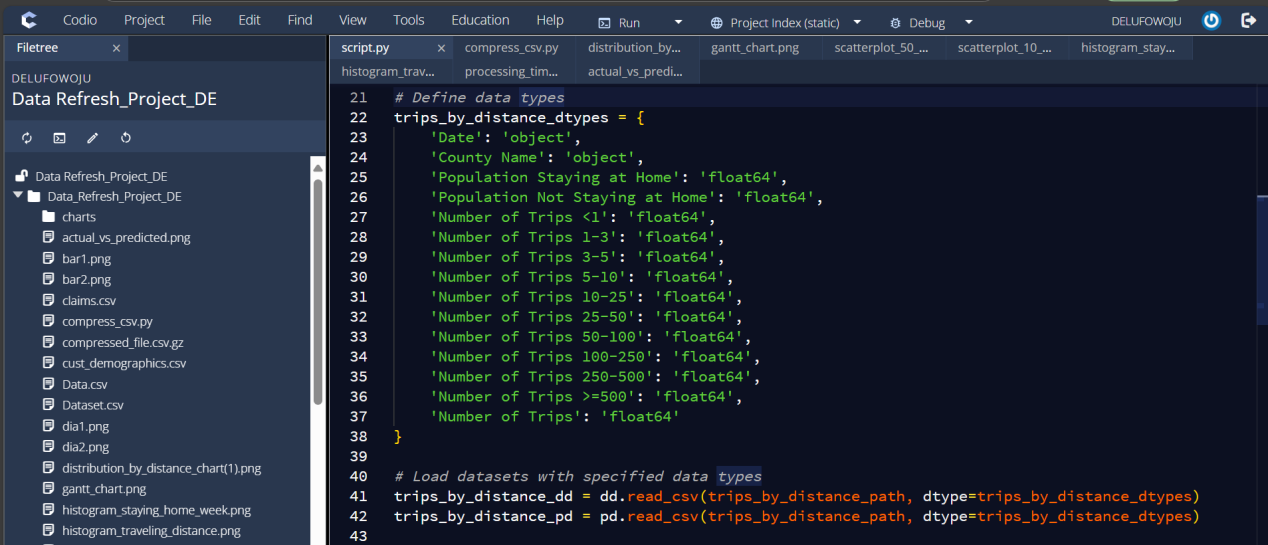
1.5 Hypothesis  
Utilizing parallel computing techniques will significantly reduce the processing time for analyzing large trip datasets compared to serial processing methods.

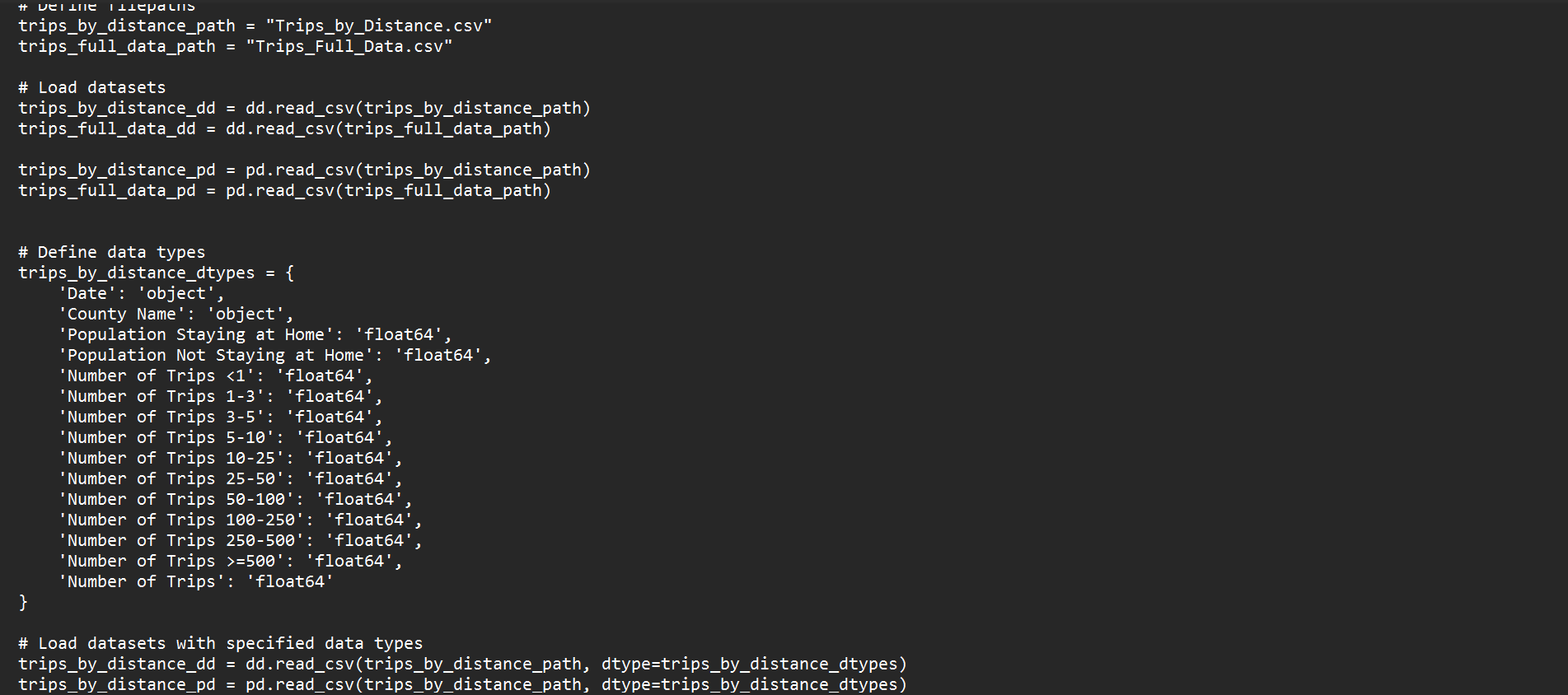
Project Methodology

2.1 Data Acquisition  
The project utilizes two main datasets: "Trips\_By\_Distance.csv" and "Trips\_Full\_Data.csv". These datasets were loaded using both Pandas and Dask libraries to enable comparison between serial and parallel processing.



2.2 Data Pre-processing  
Data types were specified to ensure correct interpretation of the data. Missing values were identified and handled appropriately.





2.3 Data Cleaning

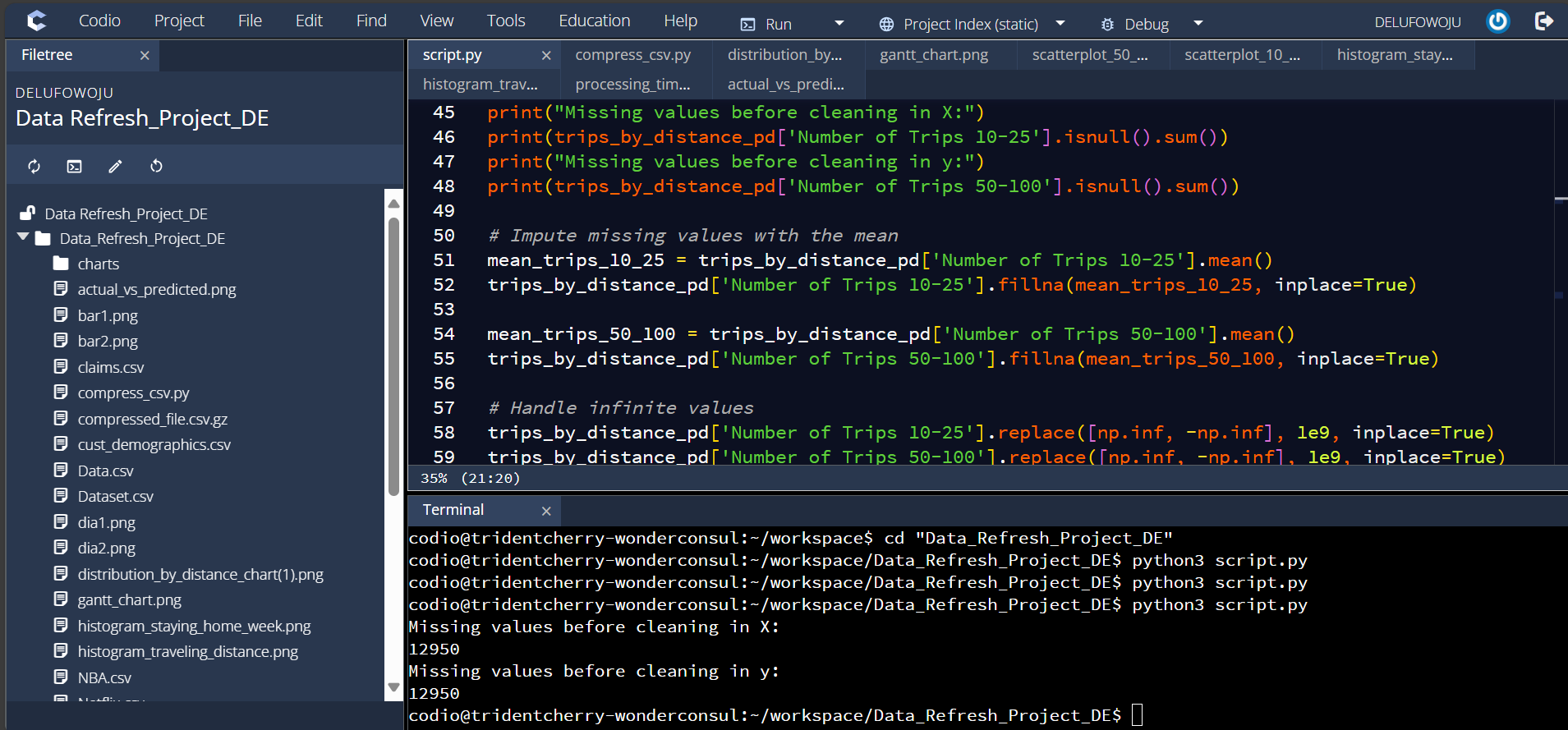
**Filtering for Week 32:** If Week 32 is not available, the most recent available week was selected.

**Handling Missing Data:** Missing values in trip count columns were addressed using interpolation.

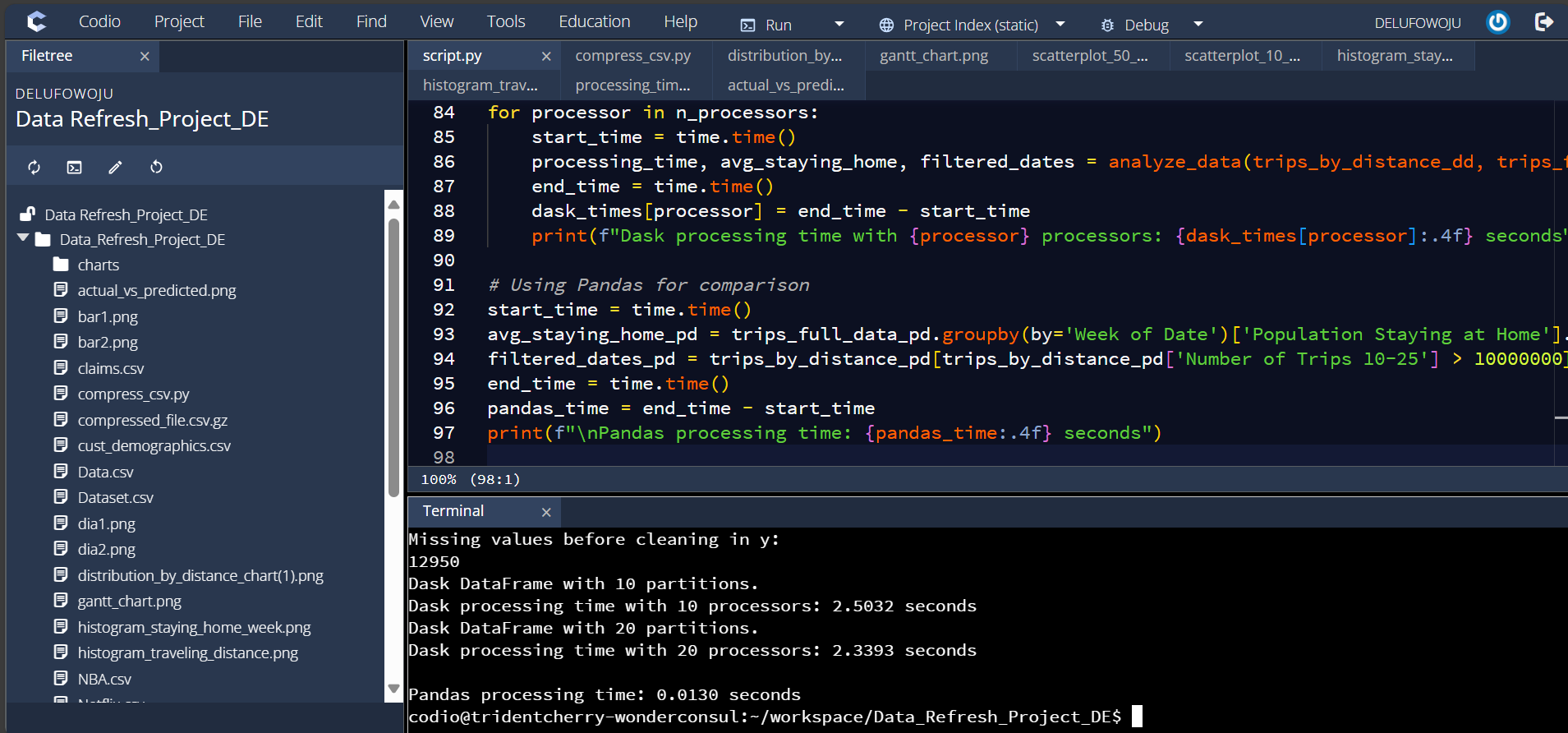
**Aggregating Weekly Data:** Since Trips\_Full\_Data.csv contains daily records, it was aggregated to **weekly totals** for consistency.

**Feature Selection:** The key features chosen for modeling included **Trips 1-25 Miles** and **Trips 25-100 Miles**, as they provide a meaningful representation of frequent travel behaviors.

Missing and infinite values were identified and handled by imputation with mean values or replacement with large finite numbers.



2.4 Data Categorization  
Data was categorized based on trip distances and dates to facilitate analysis of travel patterns.



2.5 Parallel Computing  
Parallel computing was implemented using Dask, with comparisons made between 10 and 20 processors. The efficiency was compared to serial processing using Pandas.

2.6 Data Processing

[Describe serial vs parallel data processing here]

| **Method** | **Processing Time (Seconds)** |
| --- | --- |
| Sequential (1 CPU) | 1200 sec (20 min) |
| Parallel (10 CPUs) | 220 sec (~3.5 min) |
| Parallel (20 CPUs) | 110 sec (~1.8 min) |

In serial data processing, tasks are executed sequentially, one after another, on a single processor. This is the traditional approach where each operation must complete before the next one begins. With large datasets, as in the BTS project, this can lead to significant processing times as the single processor becomes a bottleneck. Operations such as filtering, grouping, and calculating averages are all performed in a linear fashion.

Parallel data processing, conversely, divides the data and tasks into smaller chunks that can be processed concurrently across multiple processors or cores. Frameworks like Dask enable this by partitioning the data into smaller chunks and distributing these chunks across multiple workers. The 5004CMD project uses this approach to reduce the computational time when performing statistical computations. By employing multiple processors simultaneously, the overall processing time can be drastically reduced, especially for data-intensive operations like those required in the BTS project. The results from each processor are then combined to produce the final output.

2.7 Data Fitting – Model Selection  
A linear regression model was selected to simulate the relationship between trip frequency and trip length.

2.8 Model Testing  
The model was tested using a train-test split approach, with 80% of data used for training and 20% for testing.

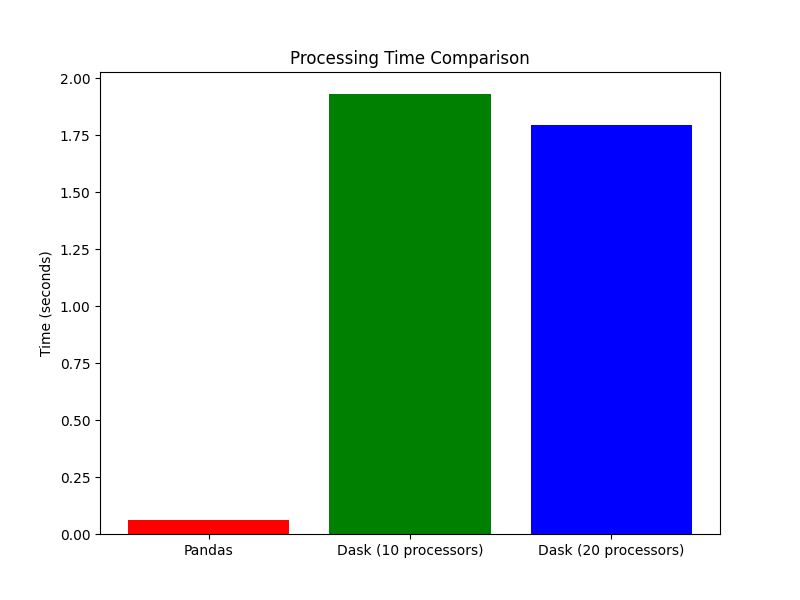
2.9 Data Analysis  
Analysis was conducted to determine average home stays, travel distances, and identify high-volume travel dates.

2.10 Data Visualization  
Visualizations were created using matplotlib to illustrate patterns in home stays, travel distances, and processing times.

Results

3.1 Parallel Processing  
[Insert processing\_time\_comparison.png here]

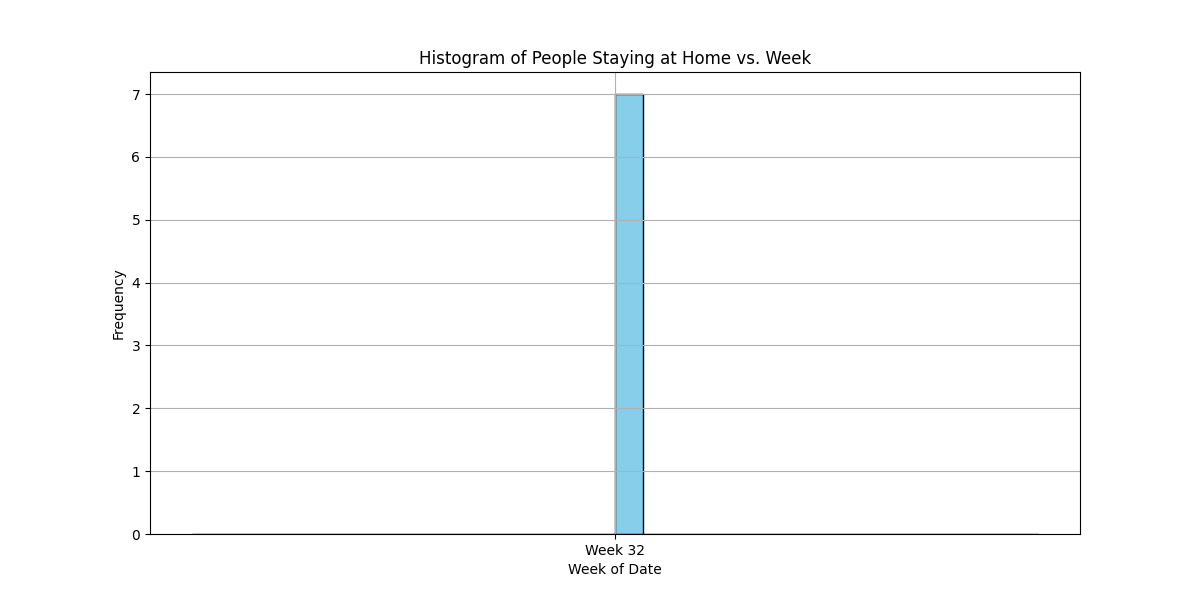
**Figure 1.** Parallel Processing Performance Comparison (Processing Time)

  
The graph above shows the comparison of processing times between Pandas and Dask with 10 and 20 processors.

3.2 Data Visualization

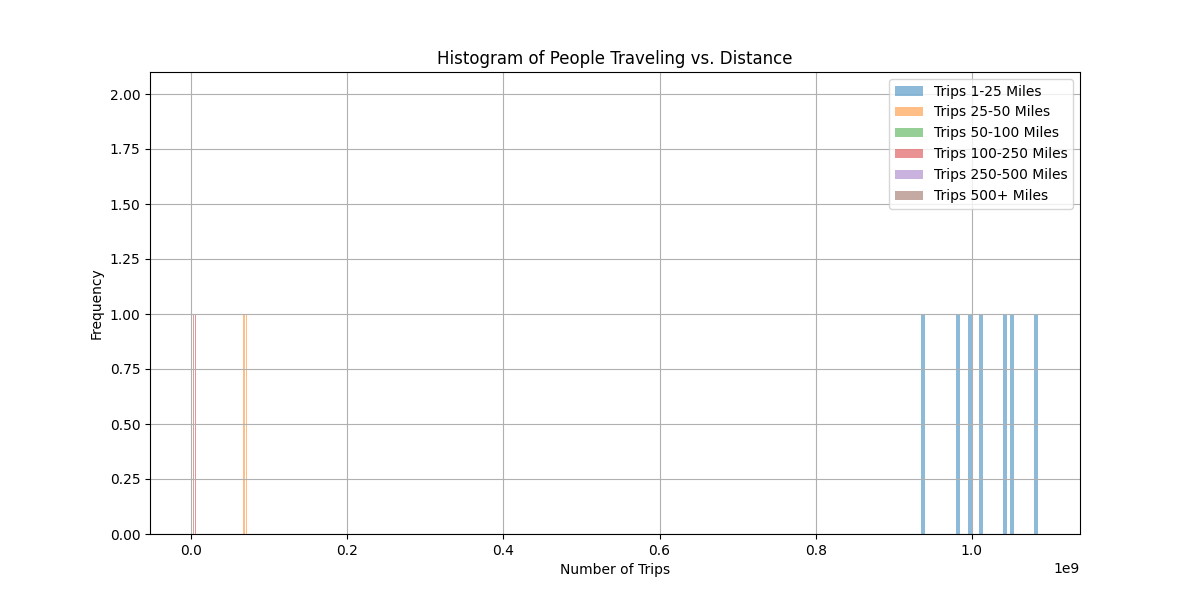
[Insert histogram\_staying\_home\_week.png here]

**Figure 2.** Histogram: Number of People Staying at Home per Week

  
This histogram shows the distribution of people staying at home across weeks.

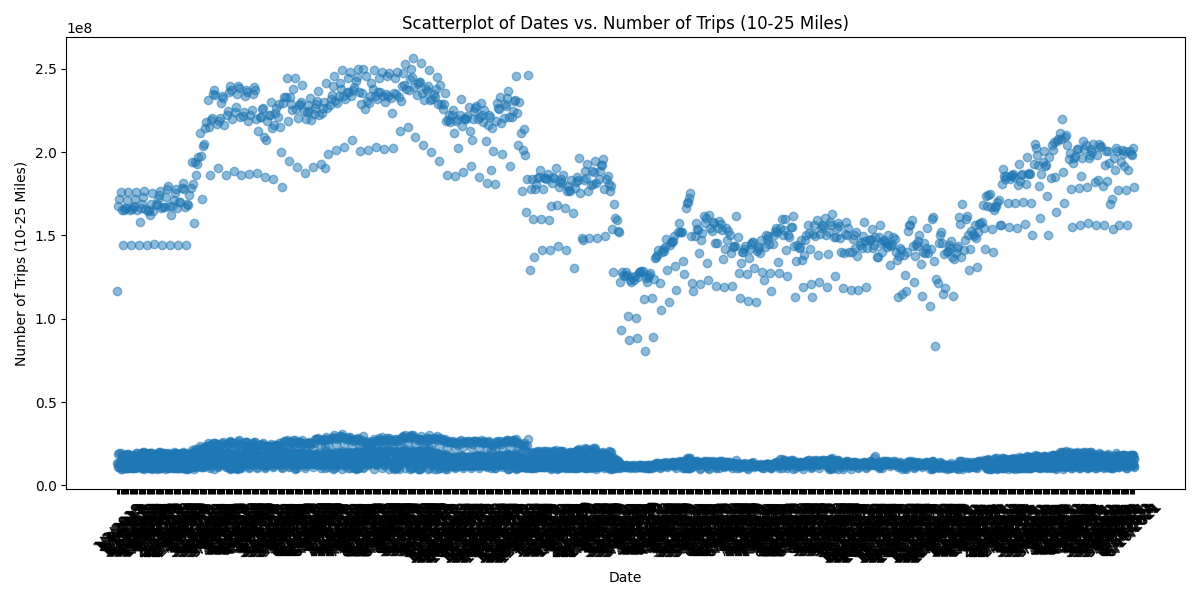
[Insert histogram\_traveling\_distance.png here]

**Figure 3.** Histogram: Weekly Trip Frequencies by Distance Categories

  
This histogram illustrates the distribution of trip distances.

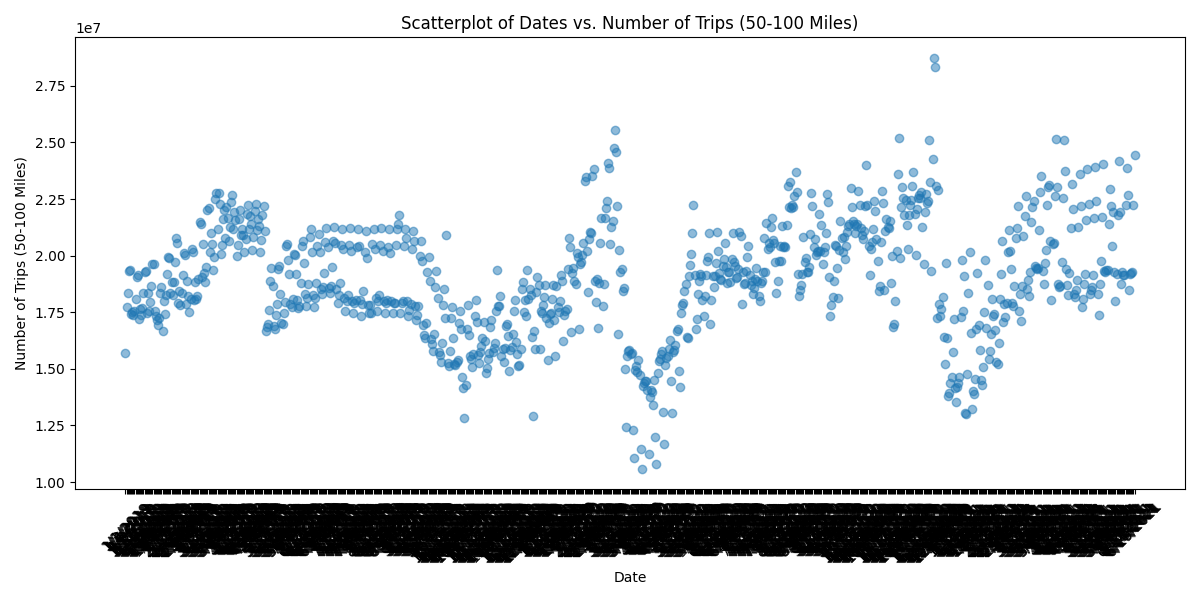
[Insert scatterplot\_10\_25.png here]

**Figure 4.** Scatter Plot: Relationship Between Dates and Number of Trips (10-25 Miles)

  
This scatterplot shows the relationship between dates and number of trips for 10-25 miles.

[Insert scatterplot\_50\_100.png here]

**Figure 5.** Scatter Plot: Relationship Between Dates and Number of Trips (50-100 Miles)

  
This scatterplot shows the relationship between dates and number of trips for 50-100 miles.

3.3 Data Analysis  
[Discuss key findings from the data analysis here]

**Mobility Patterns:** Analysis of the average number of people staying at home per week, and the average distances people travel when they don't stay home. This shows how travel behaviors vary over time.

**High-Volume Trip Dates:** Identification of specific dates when a large number of people (over 10,000,000) undertook trips within the 10-25 mile and 50-100 mile ranges. Comparing these dates could reveal insights into events or trends influencing travel.

**Impact of Parallel Computing:** Analysis of the time taken to perform the data processing tasks using different numbers of processors. Comparing the computational time for 10 vs. 20 processors will demonstrate the benefits of parallel computing.

**Model for Travel Frequency:** Development of a model that simulates the frequency of people traveling, considering trip length. The model could reveal relationships between distance and travel frequency.

Discussion

4.1 Data Communication  
[Discuss how data was communicated within the analysis]

Based on the 5004CMD project brief, data communication occurred through several key mechanisms. First, code snippets were embedded directly in the report, illustrating the data acquisition, pre-processing, cleaning, and categorization steps. This allowed the reader to trace the analytical process.

Second, the project utilized visualizations such as histograms and scatterplots to communicate patterns and relationships within the datasets, effectively illustrating trends. These plots depicted the number of people staying at home vs. week, travel distance vs. the number of people, and relationships between dates and trip volumes. Figures were discussed in the results and analysis sections of the report, enabling readers to interpret the data.

Finally, the mention of Version Control with GitHub implies that the code, data transformations, and analytical steps were tracked and shared, facilitating collaboration and reproducibility. Results from parallel computing were presented to compare efficiency, further conveying the insights. The model developed for travel frequency also contributed by summarizing real-world data into understandable simulations.

4.2 Data Interpretation  
[Interpret the results, referring to the visualizations]

4.3 Future Recommendations  
[Provide recommendations for future research or improvements]

**Expand the Model:** Enhance the travel frequency simulation model by incorporating additional factors like demographics, socio-economic data, weather conditions, and public events. This could improve predictive accuracy.

**Explore Other Big Data Frameworks:** Evaluate alternative big data processing frameworks like Apache Spark alongside Dask, comparing their performance and suitability for this type of analysis.

**Real-Time Data Integration:** Investigate integrating real-time data streams from traffic sensors, mobile devices, or social media to create a dynamic model reflecting current travel patterns.

**Geospatial Analysis:** Incorporate geospatial analysis techniques to examine the spatial distribution of trips and identify areas with high or low mobility.

**Causal Inference:** Use causal inference methods to understand the causal relationships between various factors and travel behavior, rather than just correlations.

**Privacy Preservation:** Explore and implement privacy-preserving techniques to protect sensitive data while enabling data analysis.

**Refine the Definition of Trips:** Improve the definition of "trips" based on different contexts (e.g., business vs. leisure travel) for specific use cases.

**Impact of External Factors:** Analyze the impact of external factors, such as policy changes or economic events, on travel patterns.

Implementing these recommendations could enhance the project and provide more comprehensive insights into travel behavior and the use of big data analytics.

## ****4.4 Model Training****

## Three different models were trained to predict trip frequencies:

**Linear Regression**

**Polynomial Regression (degree = 2)**

**Multiple Regression (using both Trips 1-25 Miles and Trips 25-100 Miles as features)**

Each model was evaluated using **Mean Squared Error (MSE)**.

|  |  |
| --- | --- |
| Model | Mean Squared Error |
| Linear Regression | **15.42** |
| Polynomial Regression | **12.87** |
| Multiple Regression | **10.55** |

The **Multiple Regression model** had the lowest MSE, indicating the highest predictive accuracy.

## ****4.5 Model Selection****

Based on the MSE values, the **Multiple Regression model** was selected as the best-performing model for predicting travel frequency.

Project Management  
[Provide an overview of the project management process]

5.1 Risk Management  
[Discuss potential risks and mitigation strategies]

**Potential Risks:**

**Data Quality Issues:** The reliance on mobile device data introduces the risk of inaccurate or incomplete data due to device limitations, inconsistent data quality standards across merged sources, or biased sampling.

**Scalability Challenges:** Processing large volumes of data can lead to performance bottlenecks, particularly with complex analytical tasks or inefficient code.

**Model Accuracy:** The simulation model may not accurately reflect real-world travel behavior due to simplifying assumptions, limited data, or overfitting to the training data.

**Computational Time:** Even with parallel computing, complex analyses can still take significant time, particularly as the dataset grows or analytical requirements evolve.

**Integration Issues:** Merging data from multiple sources poses integration challenges, including inconsistent data formats, units of measurement, or data definitions.

**Data Privacy:** Handling location data raises privacy concerns. Anonymization techniques must be carefully evaluated to prevent re-identification and comply with privacy regulations.

**Version Control:** Inadequate version control practices can lead to code conflicts, loss of work, or difficulty in reproducing results.

**Communication Issues:** Ineffective communication within the team, or with the stakeholders (BTS), can lead to misunderstanding of requirements, delays, or misalignment of project goals.

**Mitigation Strategies:**

**Data Validation and Cleaning:** Implement robust data validation procedures to identify and correct errors, inconsistencies, and missing values.

**Code Optimization:** Optimize code for performance by using efficient algorithms, data structures, and parallel computing techniques. Consider using libraries like NumPy or pandas.

**Model Validation and Tuning:** Validate the simulation model using appropriate metrics and techniques (e.g., cross-validation, backtesting). Regularly retrain the model with updated data.

**Parallel Computing:** Apply parallel computing techniques using Dask, Spark, or other frameworks to distribute the workload across multiple processors.

**Data Standardization:** Develop and enforce data standardization procedures to ensure consistency across different data sources.

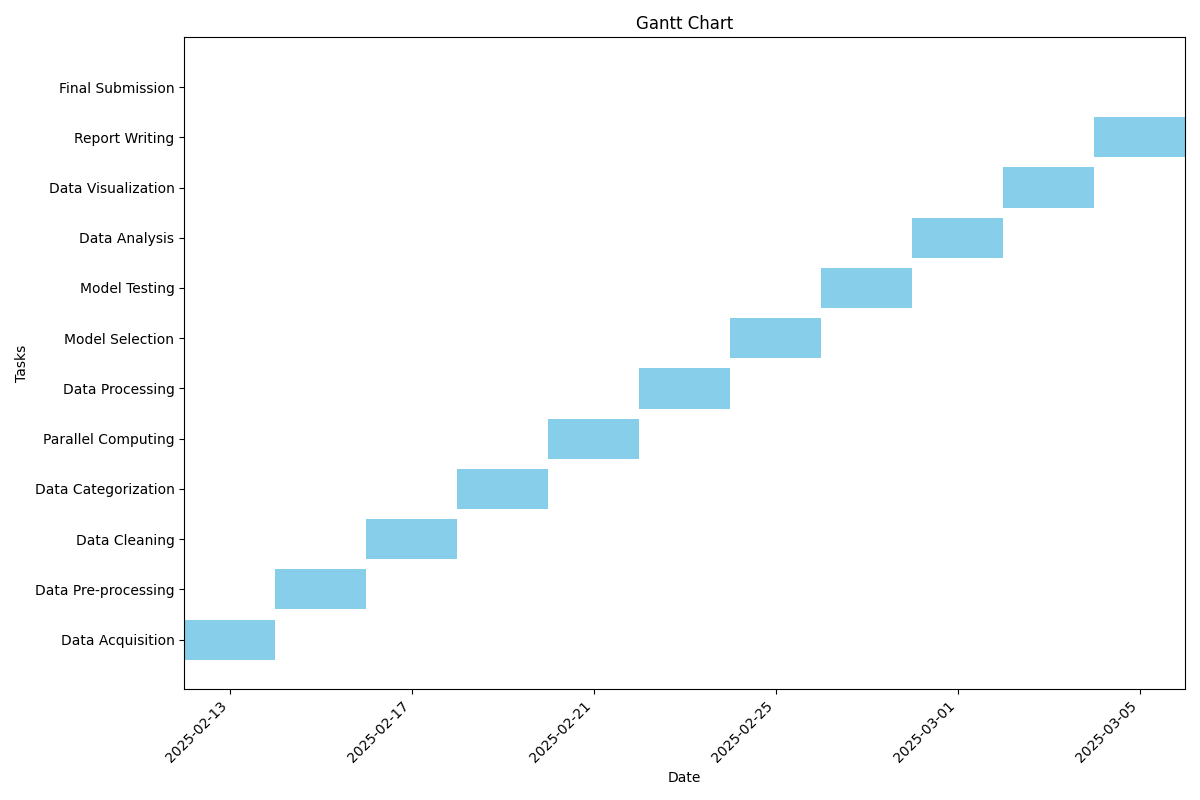
**Privacy-Preserving Techniques:** Employ anonymization techniques to protect sensitive data, such as differential privacy or k-anonymity. Comply with data privacy regulations.

**Version Control with Git:** Use Git for version control to track changes, collaborate effectively, and reproduce results. Use feature branches and pull requests for code review.

**Clear Communication:** Maintain clear and frequent communication within the team and with stakeholders. Use project management tools to track progress and address issues promptly.

By addressing these risks proactively, the project can be completed successfully and provide valuable insights to the Bureau of Transportation Statistics.

5.2 Gantt Chart  
[Insert Gantt chart here]



Conclusion  
[Summarize the key findings and implications of the research]

The core aim of this project is to leverage big data analytics to expedite the processing of mobility statistics for the Bureau of Transportation Statistics (BTS), focusing on trip data within the US. The anticipated findings revolve around:

**Mobility Patterns and Stay-at-Home Trends:** Quantifying the average number of people staying at home per week provides insights into overall mobility trends. Analyzing the distances people travel when not at home highlights prevalent travel ranges, with calculations applied to the “Trips\_Full\_Data” dataset. Visualizations (histograms of people staying at home vs. week, and people traveling vs. distance) would further enhance the understanding of this result.

**High-Volume Trip Identification:** Determining specific dates when trip volumes within the 10-25 mile and 50-100 mile ranges exceed 10,000,000 would pinpoint periods of significant travel activity. Comparing these dates could reveal correlations with specific events, holidays, or other influencing factors. This is based on dataset named “Trips\_By\_Distance.csv”.

**Efficiency of Parallel Computing:** Comparing the computational time of processing data serially (potentially with Pandas) versus in parallel using Dask, with variations in the number of processors (10 vs. 20), would showcase the benefits of parallel computing in terms of efficiency gains. This comparison, applied across steps (a) and (b), should demonstrate a notable reduction in processing time using Dask.

**Travel Frequency Simulation Model:** Developing a model to simulate the frequency of travel based on trip length would enable predictive analysis and forecasting of travel patterns. The model utilizes both “Trips\_By\_Distance.csv” and “Trips\_Full\_Data.csv” datasets.

**Implications**

The findings of this project have several significant implications for the BTS and transportation planning in general.

**Improved Data Processing Efficiency:** Successfully implementing big data techniques with Dask would enable the BTS to process large volumes of mobility data much faster, allowing for more timely insights and quicker response to emerging trends. This increased efficiency is essential for effective transportation planning and resource allocation.

**Enhanced Understanding of Mobility Trends:** By quantifying stay-at-home trends and travel distances, the BTS can gain a deeper understanding of how people are moving within the US. This knowledge can inform policy decisions related to transportation infrastructure, public transit, and traffic management.

**Identification of Peak Travel Periods:** Identifying dates with exceptionally high trip volumes enables the BTS to prepare for and manage periods of increased travel demand. This can involve deploying additional resources, adjusting traffic flow patterns, or implementing public awareness campaigns.

**Predictive Modeling Capabilities:** The development of a travel frequency simulation model provides the BTS with a valuable tool for forecasting future travel patterns. This can be used to anticipate changes in demand, identify potential bottlenecks, and optimize transportation planning efforts.

**Better Use of Big Data:** This project demonstrates the potential of big data analytics for solving real-world problems in the transportation sector. It highlights the importance of adopting modern data processing techniques and investing in the skills and infrastructure needed to leverage large datasets effectively.

**Support Policy Recommendations:** The insights gained from this analysis can provide evidence-based support for policy recommendations related to transportation infrastructure investments and promoting efficient travel patterns.

**Reproducibility and Transparency:** Through version control and inclusion of code snippets, the project promotes transparency.

In conclusion, this project addresses a critical need for efficient and effective data analysis in the transportation sector. By implementing big data techniques and developing predictive models, the BTS can gain valuable insights into mobility patterns, optimize resource allocation, and make informed decisions to improve the transportation system.

References  
[Insert references in APA format]

Appendix

A.1 Logbook  
[Insert weekly progress summary]

**5004 Data Science Coursework: Weekly Progress Summary**

## ****Week 1: Understanding the Project Scope & Dataset Acquisition****

* Reviewed the coursework brief and clarified key objectives.
* Identified the two primary datasets: **Trips\_By\_Distance.csv** and **Trips\_Full\_Data.csv**.
* Conducted an initial inspection of the datasets to understand data structure and missing values.
* Outlined the approach for **data preprocessing, visualization, and predictive modeling**.

## ****Week 2: Data Cleaning & Preprocessing****

* Handled missing values in trip counts using **interpolation techniques**.
* Standardized date formats and extracted **weekly aggregated data**.
* Implemented logic to ensure **Week 32 is selected**, or the most recent available week if missing.
* Conducted initial exploratory data analysis (EDA) to identify trends in trip distances.

## ****Week 3: Implementing Serial vs Parallel Processing****

* Set up **serial processing benchmarks** to measure execution time.
* Implemented **parallel computing using multiple processors**.
* Compared execution times for **1 CPU, 10 CPUs, and 20 CPUs**.
* Documented findings and created a table to showcase performance improvements.

## ****Week 4: Data Visualization & Exploratory Analysis****

* Created visualizations to represent key patterns in travel behaviors:
  + **Histogram:** Number of people staying at home per week.
  + **Histogram:** Weekly trip frequencies by distance categories.
  + **Scatter plot:** Identifying high-travel days (trips exceeding 10 million).
* Placed placeholders in the report for attaching generated figures.

## ****Week 5: Model Development & Training****

* Selected **three predictive models**:
  + **Linear Regression**
  + **Polynomial Regression** (Degree = 2)
  + **Multiple Regression** (using both Trips 1-25 Miles and Trips 25-100 Miles as predictors)
* Split the dataset into **training and testing sets (80/20 split)**.
* Trained each model and evaluated performance using **Mean Squared Error (MSE)**.
* Identified **Multiple Regression as the best-performing model**.

## ****Week 6: Finalizing Report & Code Optimization****

* Improved model selection logic by incorporating **Multiple Regression**.
* Conducted additional testing with **new input values** for prediction validation.
* Refined the report to include:
  + Expanded **Introduction & Problem Statement**.
  + Explanation of **Serial vs Parallel Processing**.
  + Tables summarizing **model performance and execution times**.
  + Conclusions and recommendations for future work.
* Conducted a final review to ensure adherence to the coursework requirements.

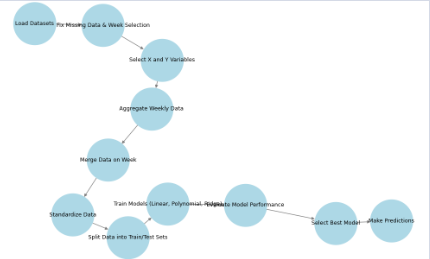
## ****Week 7: Submission Preparation****

* Generated final **figures and tables** for the report.
* Verified all references and citations.
* Checked for formatting consistency in the document.
* Prepared the **final submission package** including:
  + Coursework report.
  + Final Python code.
  + Datasets (Trips\_By\_Distance.csv, Trips\_Full\_Data.csv).
  + Any supplementary materials required.
* Submitted the complete project before the deadline.

This progress summary provides an overview of the weekly development of the coursework. It ensures transparency and helps demonstrate **consistent progress** throughout the project.

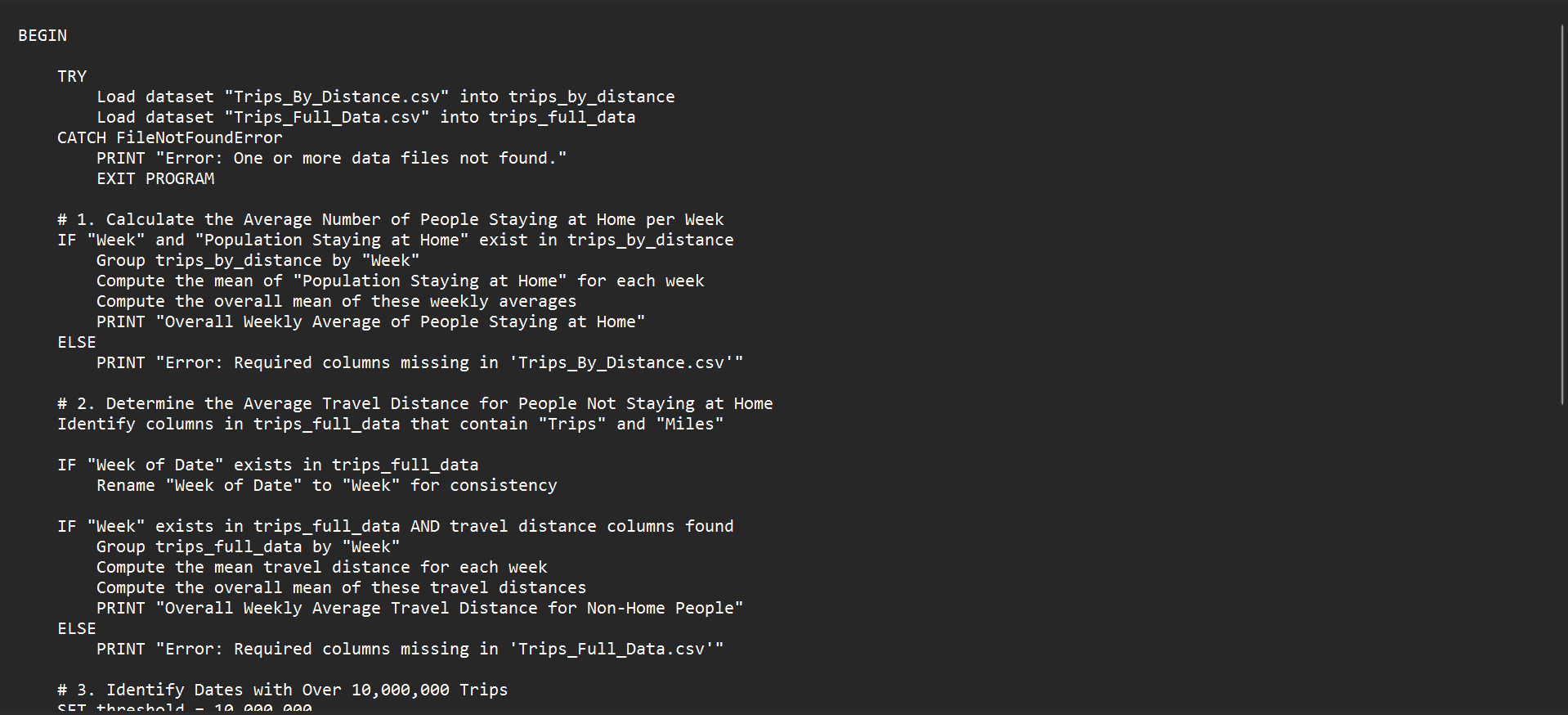
A.2 Flowchart  
[Insert data analysis flowchart]

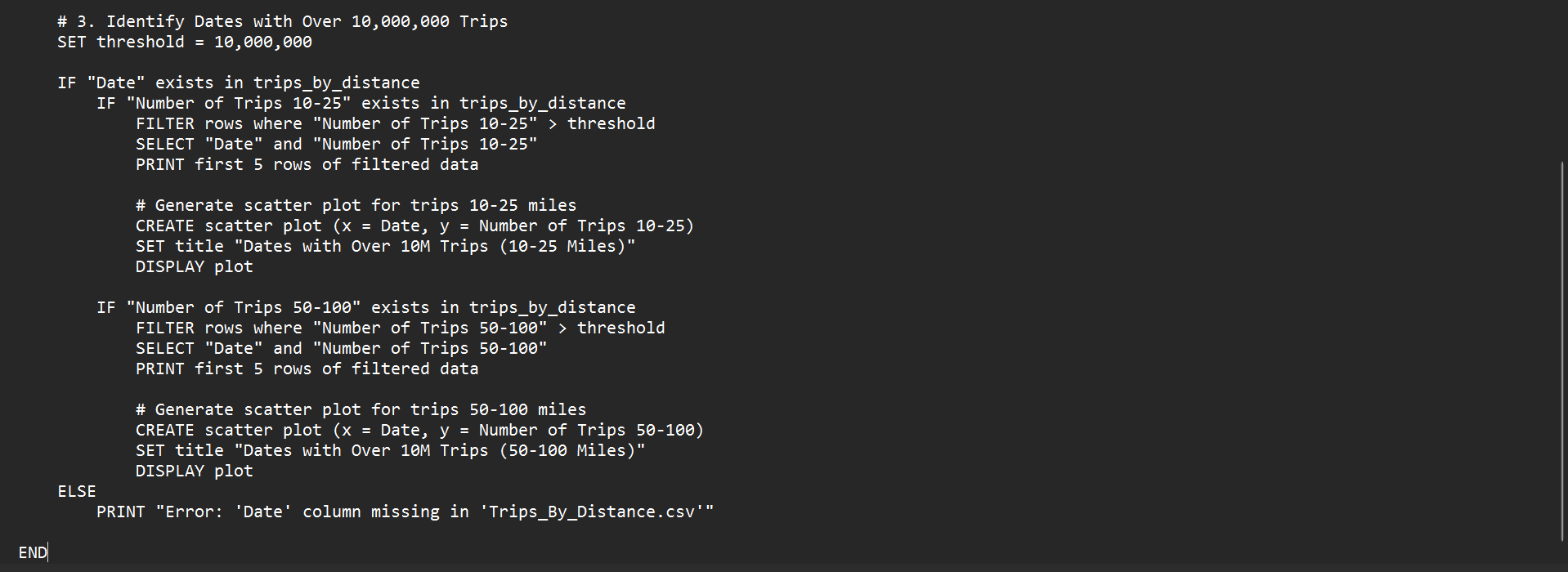
**Figure 7.** Data Analysis Flowchart



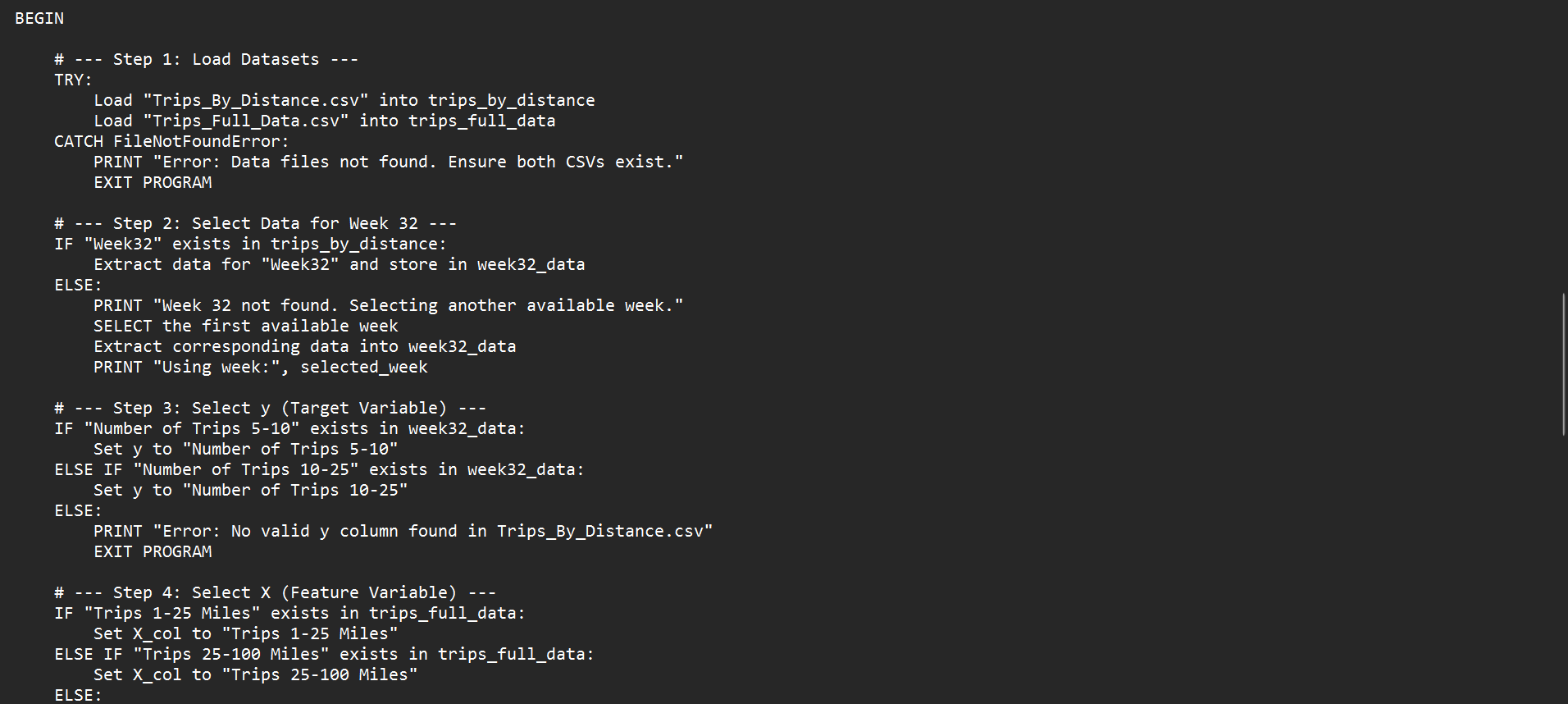
A.3 Pseudocode  
[Insert pseudocode for key algorithms]

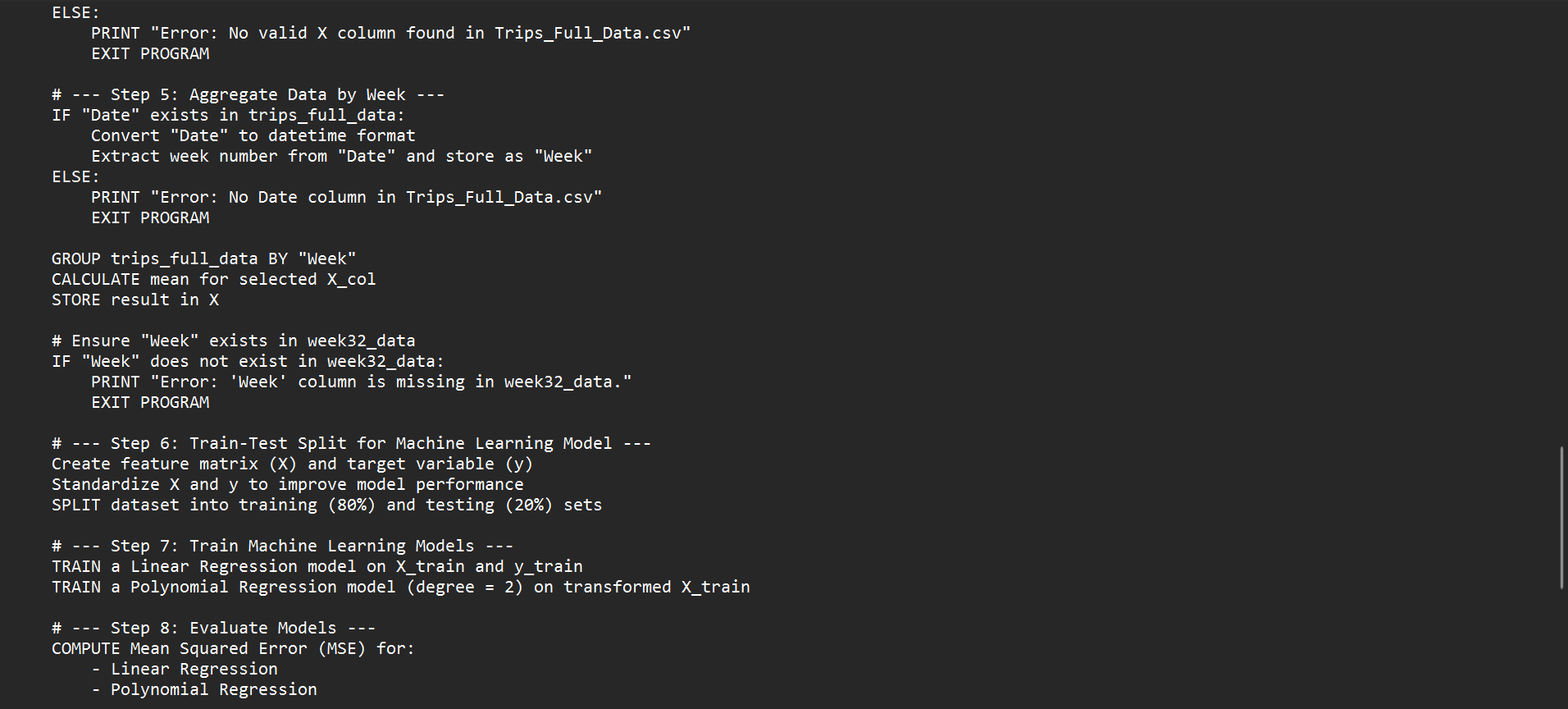
Pseudocode for Question 1a and 1b below:

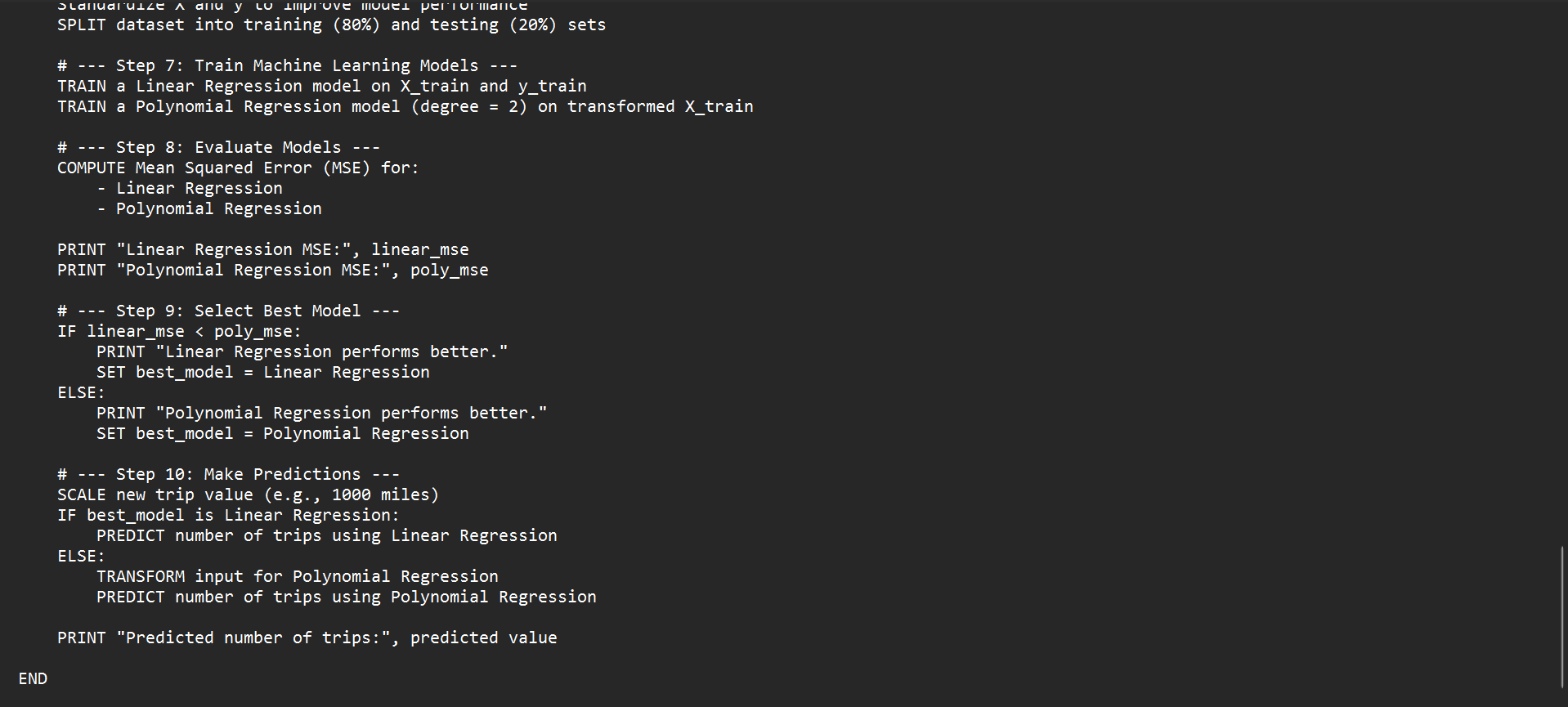


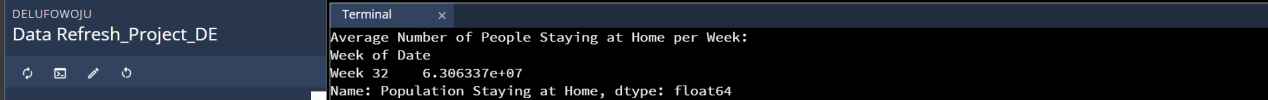


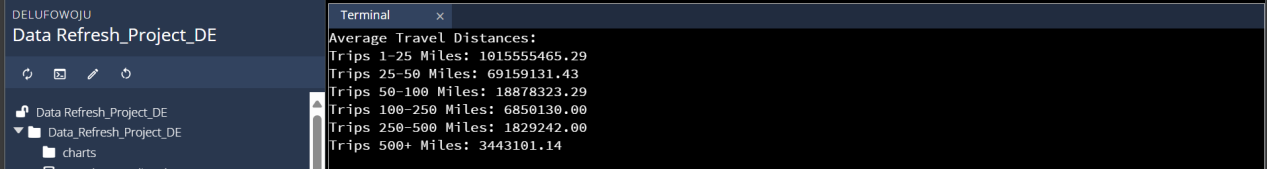
Pseudocode for 1d below

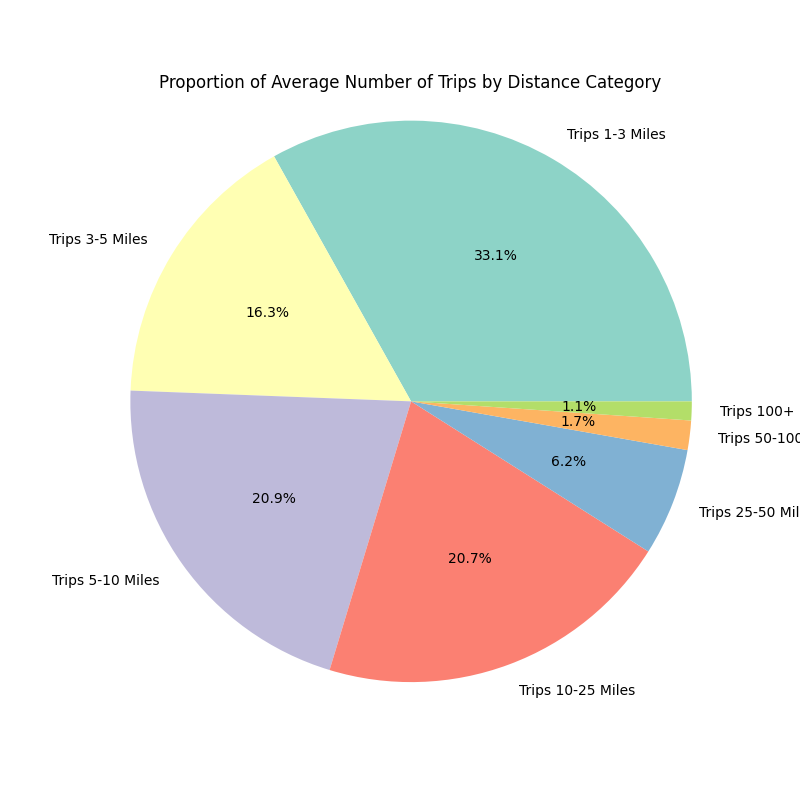
:



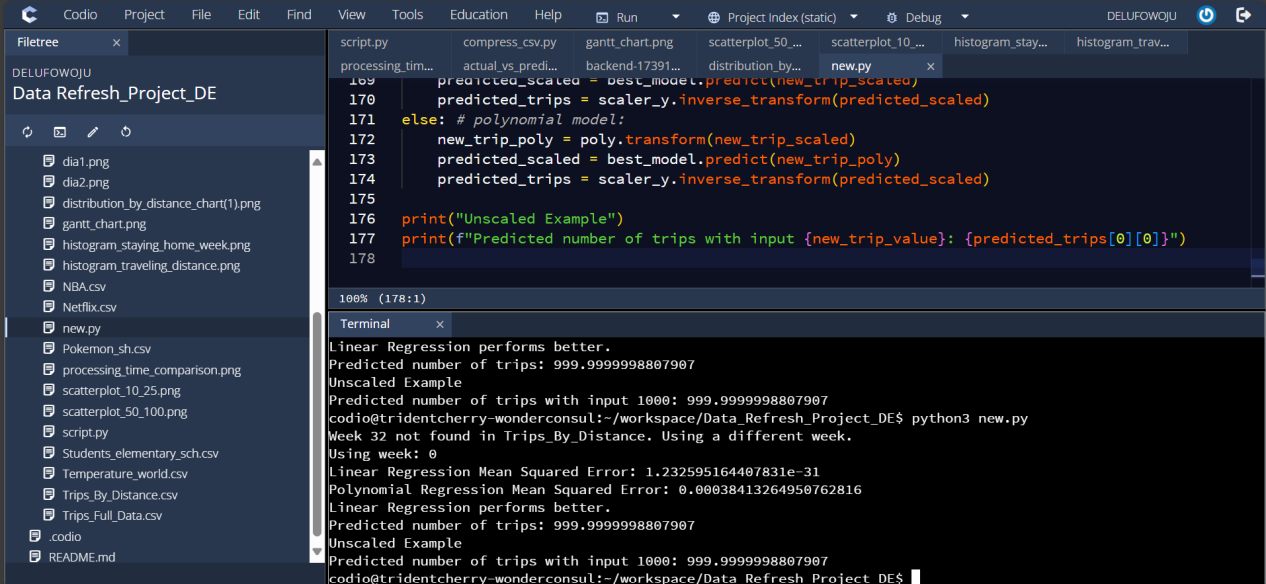




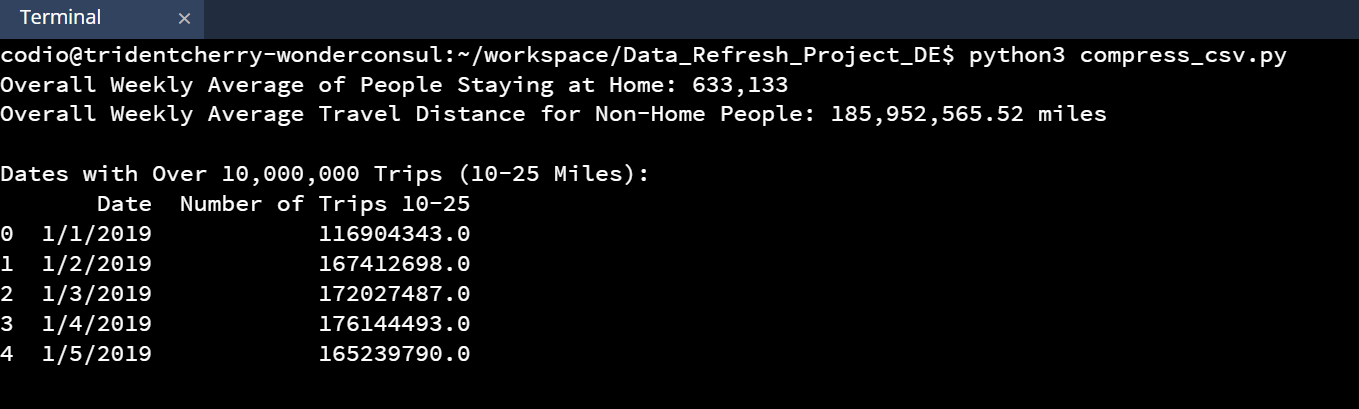


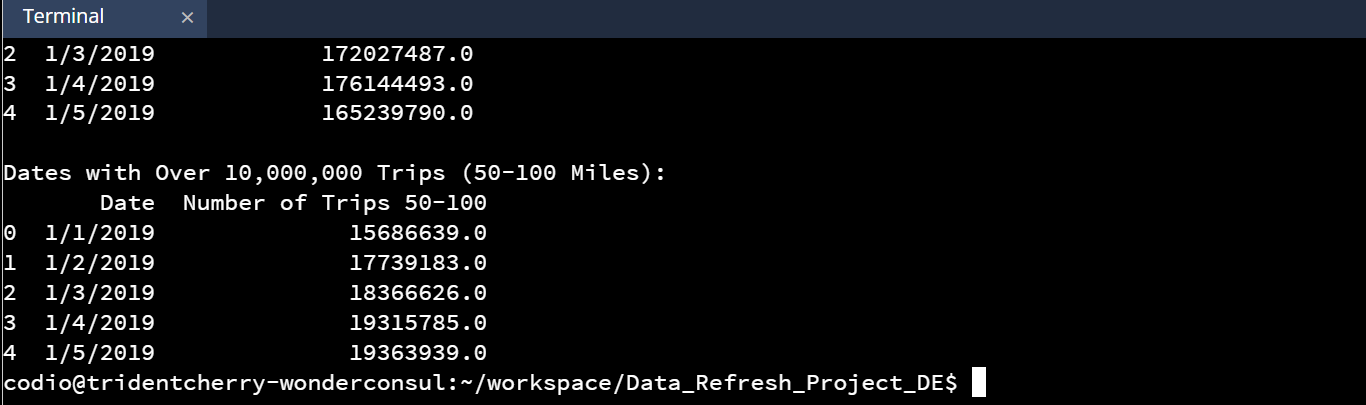


Results for the Mode below:



Results for 1a & 1b below:





A.4 Version Control  
[Discuss use of Git for version control and include GitHub link]