New York Citywide Mobility Survey 2019

- Case Study -

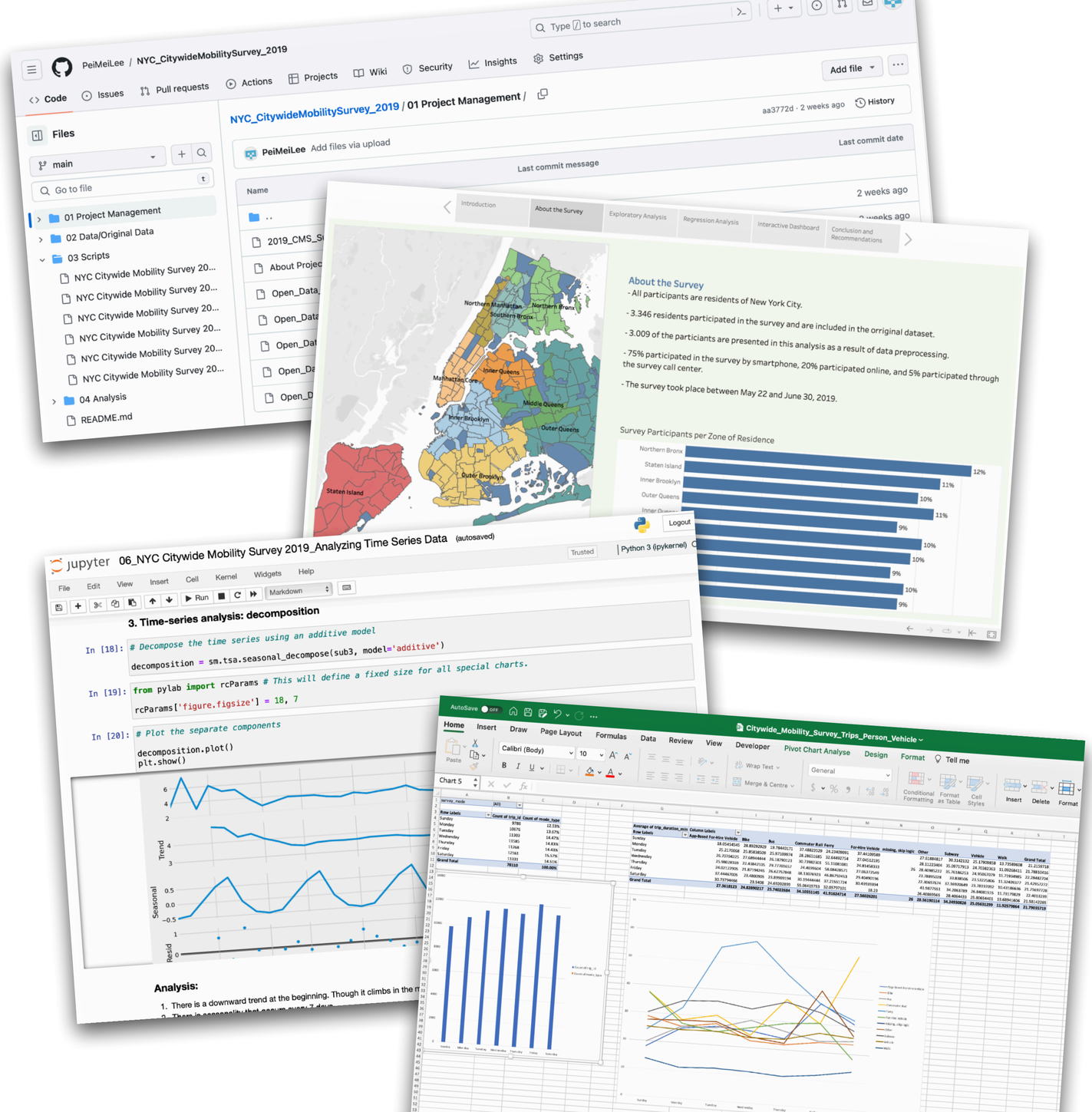
Introduction

Background

- This project was created as the final assignment for my bootcamp at CareerFoundry in 2023.
- New York City has been a fascinating city for me in many ways. Continuing my interest in developing sustainable mobility solutions, I was excited to come across this dataset.

Data

- Survey data was collected and preprocessed by a consulting company RSG for New York City Department of Transportation in 2019.
- Dataset includes 5 CSV files and 1 GeoJSON file. Data source: nyc.gov



Introduction

NEW YORK CITY DOT CITYWIDE MOBILITY SURVEY

Context

- The New York City Department of Transportation conducted an annual travel survey named the Citywide Mobility Survey between 2017 and 2020.
- It aims to assess the travel behavior, preferences, and attitudes of New York City residents towards mobility issues.
- This case study focuses solely on the survey dataset of 2019.
- All participants are residents of New York City. 3.009 of the participants are presented in this analysis as a result of data preprocessing.
- The survey took place between May and June, 2019.

Introduction

Tools Used

Jupyter

Python

Excel

Tableau

ChatGPT

GitHub

Python Libraries Used

Pandas

Numpy

Seaborn

Matplotlib

Scipy

Quandl

Statsmodels

Folium

Scikit-learn

Techniques Applied

Sourcing open data

Exploratory visual analysis

Data cleaning & wrangling

Regression analysis

Clustering analysis

Time-series analysis

Dashboard and reporting

My Work Process

Research	• Brainstorm for a topic • Source open data (trustworthiness check, initial data quality check)
Ask Questions	• Get to know data set via EDA #1 (Find anomalies, missing values…) • Define project questions & report structure
Data Preprocessing	• Perform data wrangling (Fix anomalies & missing values, choose relevant columns for analysis…) • Reflect on data limitations & data ethics
Data Analysis	• Perform EDA #2 (Confirm data quality, find trends and correlations…) • Adjust project questions • Conduct a geospatial analysis • Run a linear regression analysis and a time-series analysis
Tell the Story	• Summarize analysis results & data insights • Build a dashboard • Draw conclusions & recommendations • Produce project report & deliverables

Project Objectives

Goal #1

Discover trends and needs by studying New Yorkers' travel behavior, such as what mode of transportation is the most popular, how much time they spend on the subway everyday.

Exploratory Visual Analysis

to discover insights into the travel behavior and needs of New York residents

Linear Regression Analysis

on trip distance and trip duration to determine if the latter can be predicted by the former

Goal #2

Examine two major components of this survey, namely the trip distance and the trip duration. I'll explore the relationship and see if a linear regression model is fitting.

Interactive Dashboard

exploring how multiple variables affect trip distance and duration

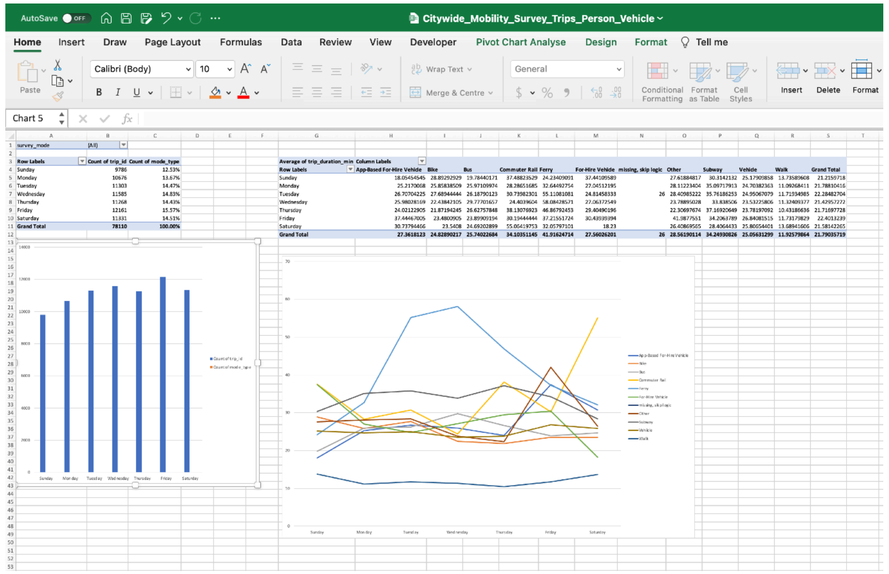
Conclusion & Recommendations

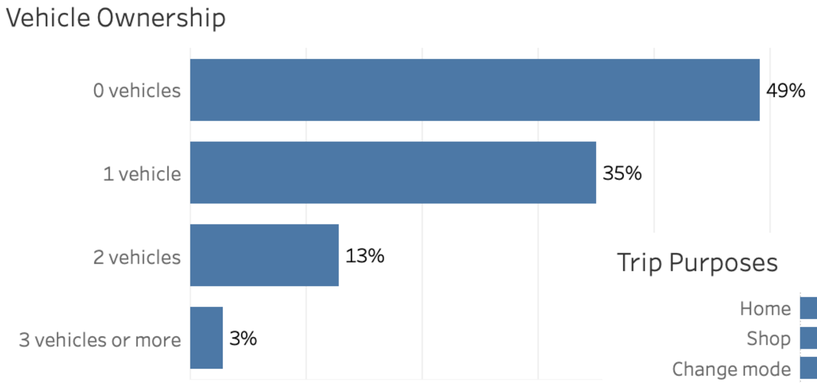
including a reflection on data limitations to wrap up my report

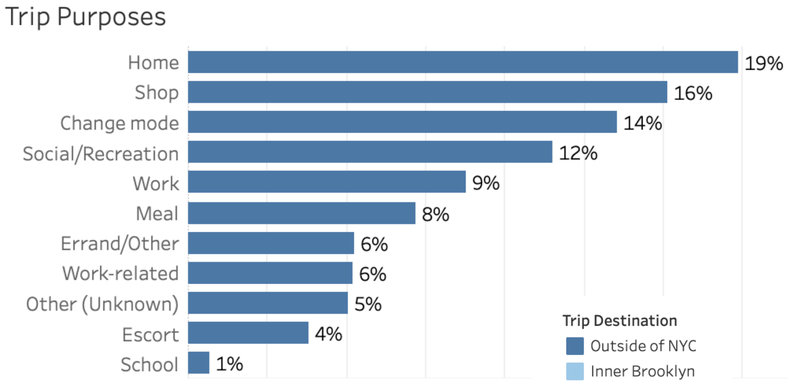
Exploratory Visual Analysis

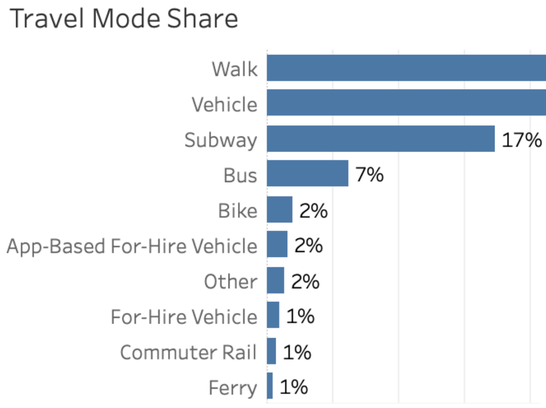
Excel has been a good companion throughout my work. I find using pivot tables and pivot charts a very efficient way to explore data!

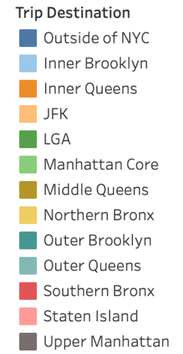
After discovering interesting trends and patterns, I turned to Tableau to start creating charts.

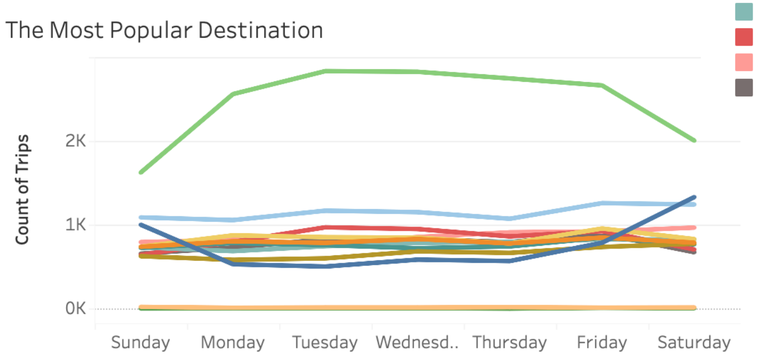






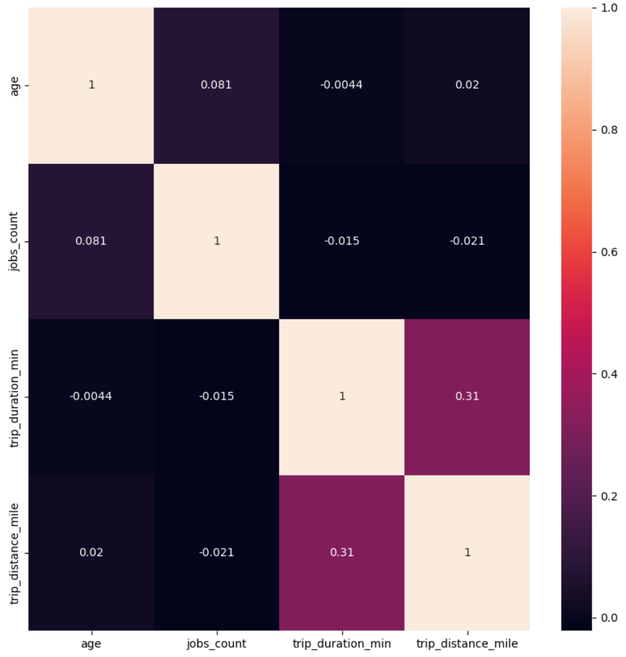






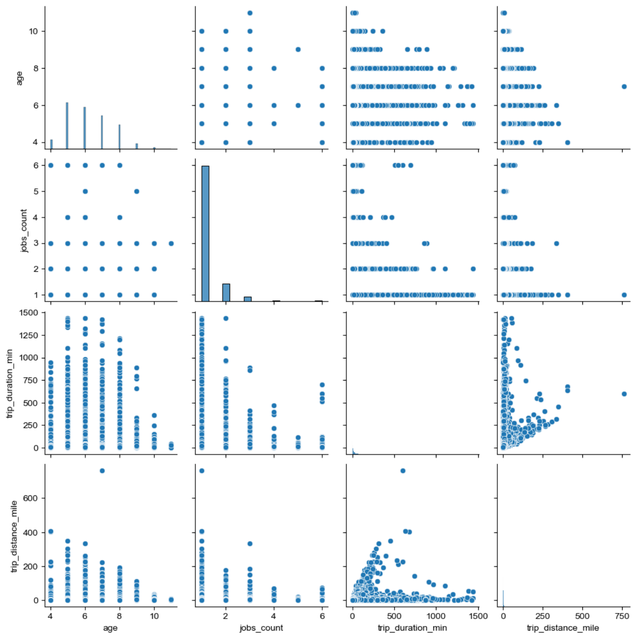
Exploratory Visual Analysis

Furthermore, I used Python to do some more exploratory work.



Heat Map

For investigating how strong a correlation between two numeric variables is, if any.



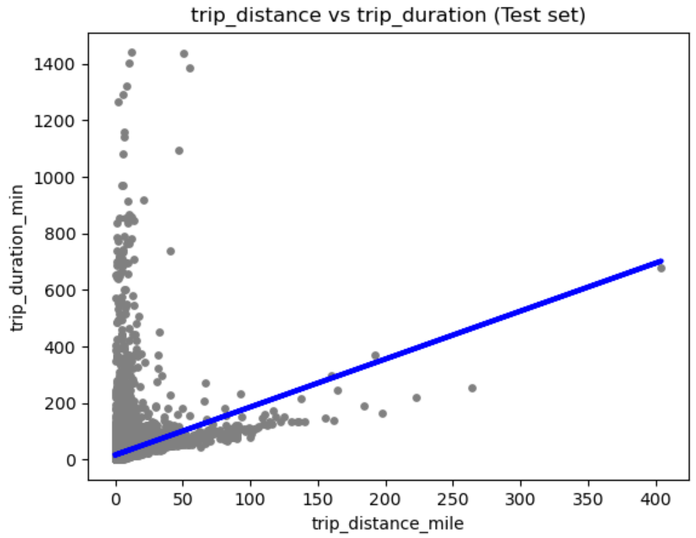
Pair Plot

to combine scatterplot and histogram, cross examining relationships between variables- both categorical and numeric.

Linear Regression Analysis

- Putting what I learned about supervised machine learning, I performed a regression analysis to see if we can predict the **trip duration** by the **trip distance**-the independent variable.

- **My hypothesis:** The longer the trip distance is, the longer the trip lasts in time.



jupyter 04_NYC Citywide Mobility Survey 2019_Regression Analysis Last Checkpoint: 2 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [16]:
```
# Create objects that contain the model summary statistics.

rmse = mean_squared_error(y_test, y_predicted) # This is the mean squared error
r2 = r2_score(y_test, y_predicted) # This is the R2 score.
```

Check the model performance statistics—MSE and R2 score.

In [17]:
```
# Print the model summary statistics. This is where we evaluate the performance of the model.

print('Slope:' ,regression.coef_)
print('Mean squared error: ', rmse)
print('R2 score: ', r2)
```

```
Slope: [[1.69735762]]
Mean squared error:  2737.7102838105175
R2 score:  0.08922481043501418
```

In [18]:
```
y_predicted
```

Out[18]:
```
array([[34.00253571],
       [17.19869529],
       [19.4052602 ],
       ...,
       [21.6118251 ],
       [17.02895953],
       [17.02895953]])
```

In [19]:
```
# Create a dataframe comparing the actual and predicted values of y.

data = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_predicted.flatten()})
data.head(30)
```

Out[19]:

	Actual	Predicted
0	36.9	34.002536
1	9.0	17.198695

- Based on the statistical results, we can say the relationship between trip distance and duration doesn't follow a single, straight regression line.

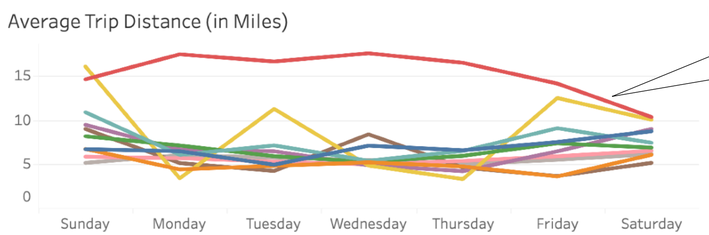
- **Conclusion:** My hypothesis can be ruled out. To accurately represent the data, new hypotheses need to be formed to find the best fitting model.

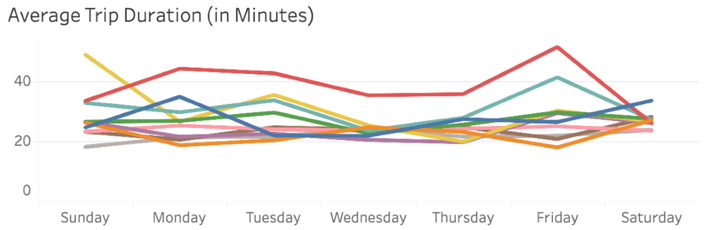
Interactive Dashboard

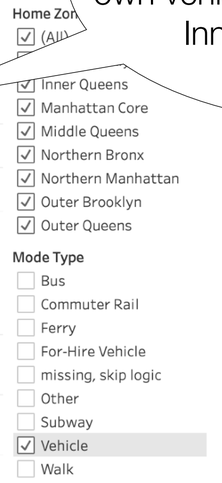
So, what affects trip distance and trip duration?

- As it turns out, factors like mode of transportation, where one lives and the day of week all have some influence on the trip distance and duration.

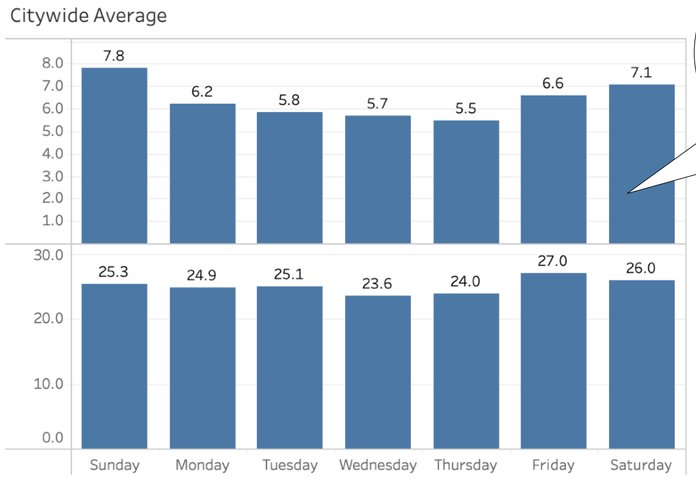
- I built a dashboard for demonstrating all sorts of insights on Tableau.



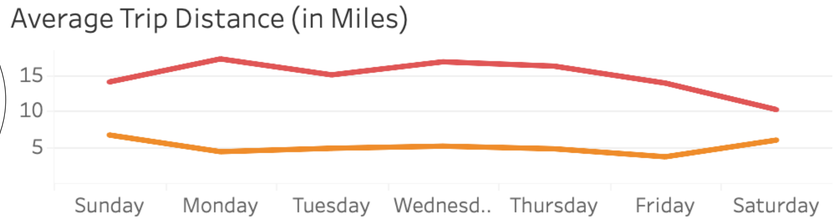


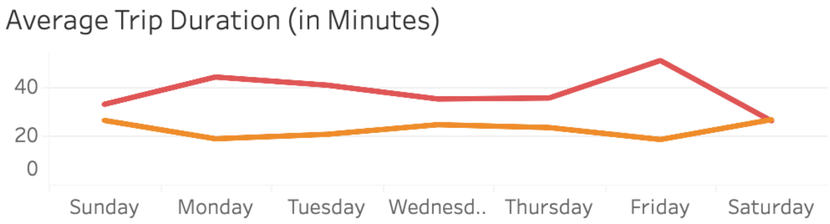


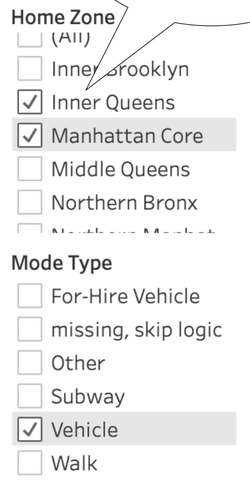
People living in Manhattan Core travel the longest distances in own vehicle whereas residents from Inner Queens travel the shortest!



New Yorkers spend longer time and travel longer distances in own vehicle during weekend than during the week!







Filter to have a closer look!

Conclusion and Recommendations

Conclusion

- The relationship between trip distance and trip duration cannot be explained solely by linear correlation.
- Factors such as mode of transportation, day of the week and where one lives all have an influence on how long the trip is, both in distance and in time.

NEW YORK CITY DOT CITYWIDE MOBILITY SURVEY

Recommendations

- Conduct a dedicated survey focused on tourists' mobility preferences (presence of many tourists in New York).
- Conduct the same survey in the other three seasons to understand the year-round transportation needs and preferences.
- Prioritize efforts to improve conditions for pedestrians, as walking is the most frequently chosen mode.
- Allocate more budget for infrastructure maintenance in Manhattan Core, as it receives the most arrivals among all zones.

Retrospective: What went well? What was challenging?

A Structured Approach

I applied what I learned in the bootcamp and followed analysis processes with modifications where I saw fit. This turned out to be a very efficient and reassuring method. I am proud of the result.

Finding Solutions

When conducting analyses in Python, I encountered my share of errors and learned to find solutions by utilizing the power of community like Stack Overflow or tools like ChatGPT.

Handling Stress

It can be stressful to make commands run. It requires patience and perseverance to find exciting relationships between variables.
Lesson learned -> Take short breaks to refresh mind and relax strained eyes.

The Big Picture

Seeing many trends and patterns reveal themselves, it was easy to be distracted from answering the key questions.
Lesson learned -> Go back to the project brief for orientation and refocus.

Thank you.