Analysis of New York City Citywide Mobility Survey Data 2019

The New York City Department of Transportation (NYC DOT) conducted an annual travel survey named the Citywide Mobility Survey (CMS) between 2017 and 2020. It aims to assess the travel behavior, preferences, and attitudes of New York City residents. In the year 2020, it was modified to a three-part study investigating transportation impacts from COVID-19 in May, July and October 2020, due to the COVID-19 pandemic. Given the time constraints for my final project, I decided to focus solely on the survey dataset of 2019.

I. Data Source

Data source: It's open source data collected by a consulting company named Resource Systems Group, Inc. (RSG), which was contracted by the New York City government for the CMS project.

Data citation: There are five data files made available on nyc.gov, including 2019 Citywide Mobility Survey 1) Person Table, 2) Household Table, 3) Trips Table, 4) Day Table and 5) Vehicle Table. My analysis will mainly focus on the Trips Table. All data files were accessed on November 8th, 2023. It's important to note that the data have been preprocessed by the RSG consulting company for their analysis purposes. The Geojson file for plotting the survey zones was accessed on Nov. 14th, 2023.

Data collection methods: The travel survey data was collected annually via three channels: 1) Smartphone app 2) Online website 3) Call center. Participants who engaged online or through the call center undertook a one-day retrospective travel diary, while smartphone participants recorded a real-time seven-day travel diary. However, not all smartphone participants completed all seven days of travel diary. Data from partially complete days has been included in the dataset and has been marked as incomplete in respective columns.

Reason for choosing this data set: 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 the project called "Citywide Mobility Survey (CMS)" initiated by the New York City Department of Transportation (NYC DOT). There are five csv files that offer a great abundance of data regarding New York city residents' transportation needs, preferences and their opinions on some current transportation topics in their city. After examining all datasets, I could already come up with some interesting questions I would like to answer.

II. Data Profile

Data cleaning

Various data cleaning steps have been carried out in the Jupyter Notebook. Here is a summary of the cleaning process.

- **Check for mixed-type data**: Five variables contain mixed-type data. All were modified to single data types.
- **Check for unclear variable names**: Seven variables have been renamed to contain information about the variables respectively.
- **Create a new dataframe containing only necessary columns**: As there are way more variables (77) than needed in the original dataset, I dropped 38 unnecessary variables and saved a new dataframe containing only 39 variables I find necessary for my analysis.
- **Check for abnormal values**:

```
In [153]: # Set display options to show all columns without truncation
          pd.set_option('display.max_columns', None)

          # Check descriptive statistics
          trips_trimmed.describe()
```

Out[153]:

	hh_id	trip_weight	survey_mode	person_id	day_num	travel_date_dow	trip_id	trip_num	leg_num	survey_complete
count	8.545900e+04	85459.000000	85459.000000	8.545900e+04	85459.000000	85459.000000	8.545900e+04	85459.000000	85459.000000	85459.000000
mean	2.596713e+07	320.889582	1.027779	2.596713e+09	3.859348	3.964697	2.596713e+12	22.657040	25.497315	0.925110
std	3.308693e+07	1570.015939	0.183516	3.308693e+09	2.032533	1.951837	3.308693e+12	17.619213	153.468402	0.263215
min	1.900024e+07	0.000000	1.000000	1.900024e+09	1.000000	1.000000	1.900024e+12	1.000000	1.000000	0.000000
25%	1.925253e+07	13.400000	1.000000	1.925253e+09	2.000000	2.000000	1.925253e+12	9.000000	1.000000	1.000000
50%	1.951766e+07	30.700000	1.000000	1.951766e+09	4.000000	4.000000	1.951766e+12	19.000000	1.000000	1.000000
75%	1.979350e+07	102.300000	1.000000	1.979351e+09	6.000000	6.000000	1.979351e+12	32.000000	1.000000	1.000000
max	1.999892e+08	94812.000000	3.000000	1.999892e+10	7.000000	7.000000	1.999892e+13	178.000000	995.000000	1.000000

    - These columns have an abnormally wide range of 1 - 995: 'mode_type', 'leg_num', 'mode_type', 'mode_type_detailed', 'sustainable_mode'. According to the Survey User Guide, 995 is code for "a value or response not required under the circumstances."

    - These columns have an abnormally wide range of 1 - 997: 'mode_2', 'mode_3', 'mode_4'. According to the Data Dictionary, 997 is code for "Other."
- **Check for missing values**: The variable 'trip_distance_mile' contains 87 NaN values. Given that this accounts for a negligible portion of the 85,459 records, these NaN values have been imputed with the mean value.
- **Check for duplicates**: No duplicates are found.

Data limitations

- **Statistical limitation**: The sample size is 3,346 records that met the survey completion criteria. As of July 1st, 2019, the population of New York City was estimated to be 8,336,817 by the U.S. Census Bureau. With a confidence level of 95%, we have a 2% margin of error. This allows us to expect the survey results to reflect the views from the overall population.
- **Scope of survey participants**: The survey exclusively includes New York City residents, rendering it unrepresentative of the diverse needs and behaviors of tourists in the city. Given that approximately 66.6 million people visited New York in 2019, conducting a dedicated survey focused on tourists' mobility preferences would provide valuable insights into this unique demographic. Such an approach would enhance our understanding of the city's overall mobility landscape.

- **Seasonality**: All trips took place in May or June. This limits our observations to only the participants' travel behavior and needs in Summer. In order to achieve a comprehensive understanding of the year-round transportation needs and preferences of the New York residents, further data should be gathered in the other three seasons.
- **Bias**: The survey dataset was weighted to address sampling and geographical bias, ensuring its representativeness of the sample region. Detailed weighting methodology is summarized in the Survey User Guide.

Data ethics

- There is no personal identifiable information (PII) present in the dataset. So this ensures that no data can be traced back to any individuals.
- I also noticed that the values for both columns "depart_time" and "arrive_time" have been reduced to only the year-month-day part. This eliminates my concerns for personal privacy intrusion.

III. Define questions

Below are the initial questions I would like to answer in my analysis. This list of questions will be adjusted as the analysis work progresses.

- What are the most popular means of transportation?

- What is the average travel time to work?
- Is there a relationship between age and travel behavior?
- Are there discernible travel behavior trends between different zones within the city?
- Is there a relationship between the transportation mode used and the day of week?
- Is there variation in household car ownership across different zones?