## 3. Literature Review

Anand (2021) employs a machine learning method to train a model for predicting taxi trip durations. He leverages calendar-related features for making predictions. Akimova (2021) does the similar work on subway data. Zhong (2021) also does exploratory analysis of New York taxi trip, but he introduces weather as an important feature having effect on taxi trip.

Deng et al. (2020) apply principal component analysis to derive synthesized measures from multiple distinct variables. This synthesis method generates weights for each original variable based on the data itself.

Our project use not only the calendar-related features but also the weather features to predict the busyness. We use principal component analysis to sythesize two public transportation busyness into a single busynes measure for POIs.

## 4. Methodology

The data team's responsibility is to generate predictions for the busyness measure of points of interest (POIs) or of taxi zones. To execute this task, we utilize machine learning methods to train regression models that the backend team will employ for the specific predictions.

**Descriptive and Target features.** In the regression model, the descriptive features are divided into two groups: the calendar-related features and the weather-related ones, and the target features encompass the busyness levels of both POIs and taxi zones. This approach follows Anand (2021), Akimova (2021), and Zhong (2021).

**The busyness of POIs.** The busyness of POIs is a composite measure of the busyness of two transportation modes: taxi zones and subway stations. Specifically, it is the first principal component derived from the two fundamental busyness measures of taxi zones and subway stations. This approach follows Deng et al. (2020). A comprehensive explanation of principal component analysis can be found in Chapter 3 of Joseph et al. (2010).

**The busyness of taxi zones and of subway stations.** The taxi zone busyness is simply measured by the sum of the numbers of the passengers picked or dropped in the taxi zone within a specified time window. The subway station busyness is simply measured by the sum of the numbers of the passengers surveyed at the station within a specified time window. A POI is matched with the busyness of the taxi zone in which this POI lies and is matched with the busyness of the subway station that is most closed to this POI.

## 5. Data Analytics and Visualization

### 5.1 Data Origin

**The POI list.** Firstly, we require a list of POIs in Manhattan, which we extract from the nodes data provided by https://overpass-turbo.eu/. Within this dataset, we filter the nodes labeled as 'tourism'. The resultant POI list includes POI IDs, latitude and longitude coordinates, names, and types (museum or viewpoint) for each POI.

**Time Range.** To assess the busyness measure of POIs, we consider two modes of transportation: taxis and subways. Consequently, data pertaining to these two transportation modes is necessary. The dataset spans from January 1, 2022, to April 30, 2023, a period closely proximate to the current date. The busyness measure for taxi zones is straightforwardly derived from the taxi dataset.

**Transportation data.** The predictive model's input features encompass two types: date/time features and weather features. The date/time features can be derived from the date/time column in the raw transportation datasets. These datasets, namely the taxi trip dataset and subway ridership dataset, are obtainable from the official website of the New York government. In the taxi trip dataset, the features used sequentially are date-time, event type (pickup or drop-off), event location (represented in latitude and longitude coordinates), and passenger number. In the subway ridership dataset, these features include date-time, station ID, rider number, and station location (represented in latitude and longitude coordinates).

**Taxi zone list.** The taxi zone list is also required to correlate the POIs with the taxi zones. This dataset can be downloaded from the NYC Open website. It is in geojson format, and the 'geometry' column can be matched with the coordinates of the POIs directly.

**Weather data.** The weather features are sourced from a weather data repository. In this project, we obtained the historical weather dataset from http://www.openmeteo.com/, as it imposes no limitations on the quantity of historical weather data downloads. The weather features included in our predictive models encompass Weather Code for labeling the weather type, Temperature (measured at 10 meters above ground), Wind Speed (measured at 10 meters above ground), and Precipitation (including rain and snow). The choice of these features derives from simulations of travel choices undertaken by regular individuals.

### 5.2 Data Preparation Strategy

Before the data is input into the predictive model, it undergoes pre-processing through essential steps. The most pivotal process is the generation of the busyness measure for POIs.

**Taxi zone busyness.** The process strategy involves creating the taxi busyness measure for each POI within every time window in the historical timeframe. This leads to the creation of a taxi busyness table indexed by a combination of taxi zone ID (extracted from the raw taxi trip dataset) and time window (indicated by the window's start time). Our time window length is set at 1 hour. This transformation is achieved by grouping the rows in the raw taxi trip dataset based on time window and taxi zone ID, followed by summing the passenger numbers. The passenger number is derived from the sum of pickups and drop-offs, as the passenger count for each direction reflects the historical busyness before or after a specific time point. Given that our busyness concept pertains to a time window rather than a specific time point, this approach aligns with our objectives. Thus the passenger number for a POI within a time window is used as the busyness value.

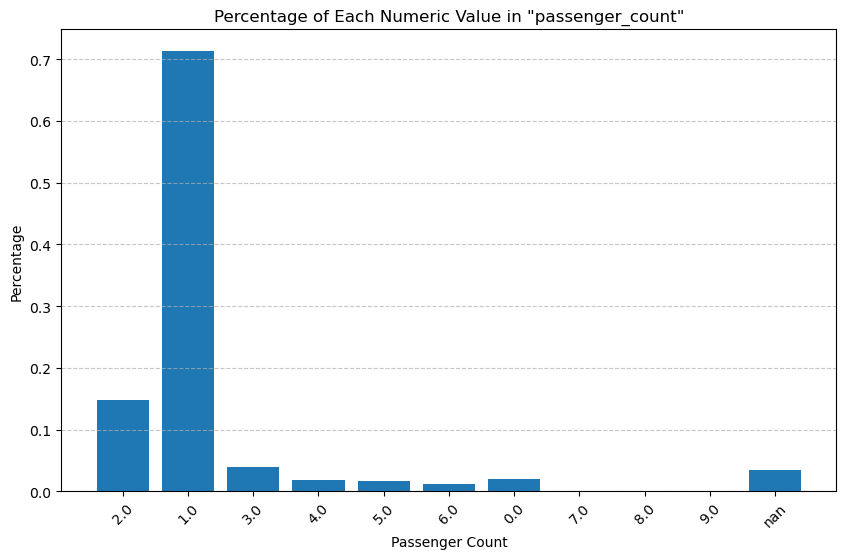
**Subway station busyness.** The subway busyness measure is constructed in the similar way, resulting in the generation of a subway busyness table indexed by a combination of subway station ID (extracted from the raw subway ridership dataset) and time window (indicated by the window's start time).

**Matching POIs with taxi zones and with subway stations.** To assign the busyness measures of taxi zones or subway stations to the POIs, we construct a POI-zone table to establish a correspondence between the POIs and taxi zones, and a POI-station table to establish a correspondence between the POIs and subway stations. Utilizing these two tables, we proceed to allocate the respective busyness measures of taxi zones or subway stations to the relevant POIs. Following the normalization of these busyness values, we calculate their principal components to generate a synthesized busyness measure for the POIs.

### 5.3 Data Cleaning

There are two issues in data cleaning: the presence of incorrectly recorded year numbers, such as 2003 or 2004 in the dataset for the year 2022, and the occurrence of missing values in the passenger numbers.

The first issue can be readily resolved by verifying whether the value of the year attribute matches the correct one. For the second issue, we address the missing values by replacing them with the most frequent value to minimize the information loss. The distribution of the values is as follows:



Obviously, replacing the missing values with 1 minimize the information loss.

### 5.4 The POI’s Busyness

**Constructing the POI’s busyness.** Firstly, the busyness measures of taxis and of subways should have been standardized to have a mean of 0 and a standard deviation of 1 by the function StandardScaler() from scikit-learn library. Then, inputting the busyness columns of taxi zones and of subway stations into the function PCA() from the sklearn.decomposition library, we obtain the first principal component from the output and use it as the POI’s busyness. The implementation of this method using the scikit-learn library can be referenced to the official tutorial page at <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA>

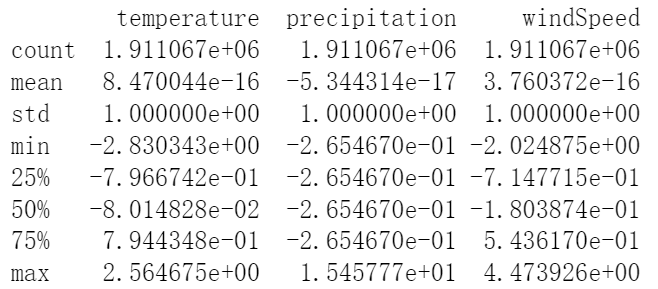
**Validating the POI’s busyness.** To assess the validity of using the principal component method to construct POI busyness, we calculated the Explained Variance Ratio of 2 principal components. The results are as follows:

|  |
| --- |
| Explained Variance Ratio - Principal Component 1: 64.37% |
| Explained Variance Ratio - Principal Component 2: 35.63% |

This result indicates that the first principal component can account for a significant portion of the busyness in the two transportation modes. Thus, employing a single principal component as the synthesized busyness measure is justified.

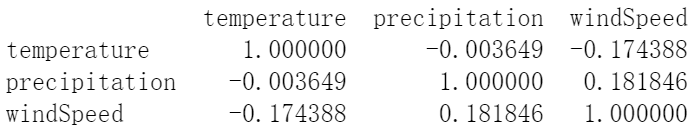
### 5.5 The Resulted Data

The descriptive statistics of the standardized continuous features are as follows:



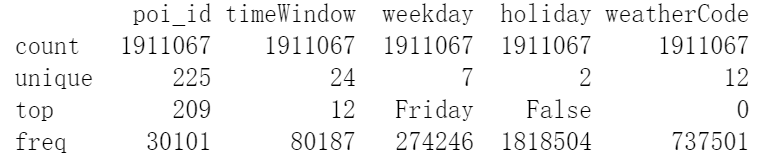
The values of each continuous feature have been standardized to have a mean of 0 and a standard deviation of 1 by the scikit-learn function StandardScaler(). The values of temperature and precipitation both fall within a range of 3 standard deviations, while those of windSpeed fall within a range of 4 standard deviations. This observation indicates the absence of any abnormal outlier points.

The correlation matrix of standardized continuous features is as follows:



The correlation coefficients are all below 0.2, indicating the absence of co-linearity among the features. Hence, these three weather features can effectively capture distinct essential decision factors.

The descriptive statistics of categorical features are as follows:



The count of unique values for each feature aligns with real-world expectations. For instance, we have a total of 225 POIs, 24 time windows in a day, 7 days in a week, binary values for holiday status, and 12 distinct weather descriptions.

In conclusion, the data inputted into the regression model is of high quality.

## 6. Evaluation and Results

To bring down the model training time to an acceptable level while experimenting with various regression models, we utilize a small subset of the data for model comparison. We randomly select one-tenth of the rows from the entire dataset for model comparison. The dataset is split into training and test data at an 80:20 ratio.

Given that the target feature is continuous, we employ the mean squared error (MSE) as the evaluation metric to assess the accuracy of the model's predictions. Linear regression, being both straightforward to implement and highly efficient, serves as our benchmark model. We consider adopting an alternative model for our application only if it demonstrates a notably superior level of accuracy.

|  |  |
| --- | --- |
| model | mse |
| Linear regression | 0.5064777639900107 |
| Ridge Regression (alpha=1.0) | 0.506521404669609 |
| K-Nearest Neighbors Regression (n\_neighbors=20) | 0.5384099811070161 |
| Random forerst regression(n\_estimators = 40, max\_depth = 40) | 0.31464794977337357 |
| XGBoost Regression | 0.34592513510308964 |

Clearly, Random Forest Regression (n\_estimators = 40, max\_depth = 40) yields the highest accuracy that is significantly higher than that of linear regression model. As a result, our application employs the trained Random Forest Regression (n\_estimators = 40, max\_depth = 40) to predict the busyness of POIs. When applied to the entire dataset, the trained Random Forest Regression model produces a final MSE of 0.28593275930093004.

References

[1] D. Akimova. "A Geospatial Analysis of the NYC Subway in R," June 25, 2021 [Online] Available: https://www.kaggle.com/code/dashaa/a-geospatial-analysis-of-the-nyc-subway-in-r/report

[2] S. Anand. "Exploratory Data Analysis on NYC Taxi Trip Duration Dataset," January 9, 2021 [Online] Available: <https://www.analyticsvidhya.com/blog/2021/01/exploratory-data-analysis-on-nyc-taxi-trip-duration-dataset/>

[3] Deng C, Dong X, Wang H, et al. A data-driven framework for walkability measurement with open data: A case study of triple cities, New York. ISPRS International Journal of Geo-Information, 2020, 9(1): 36.

[4] F.H. Joseph, J.B. Barry, E.A. Rolph, and E.A. Rolph, "Multivariate data analysis," Pearson Prentice Hall. 2010.

[5] H. Zhong. "Exploratory Data Analysis of New York City Taxi," June 25, 2021 [Online] Available: https://medium.com/@haonanzhong/new-york-city-taxi-data-analysis-286e08b174a1