**Machine Learning application on Urban Planning:**

**A Case Study of Chicago Downtown Parking Structures**

Yikai (Kai) Chen

Master of Project Management Program

Department of Civil and Environmental Engineering

McCormick School of Engineering and Applied Science

Northwestern University

Evanston, Illinois

December 1, 2021

**Table of Contents:**

Abstract

Executive Summary

Preface

1. Introduction and Background
2. Urban Planning
   1. Development of Parking Planning
   2. Parking Planning in Chicago
3. Data Analysis with Machine Learning in Chicago Downtown
   1. Datasets Information
   2. Methodology Overview
   3. Data Preparation
   4. Machine Learning Model Development
   5. Results Discussion and Comparison
4. Suggestions
   1. Current Parking Structures
   2. Future Parking Structures
   3. Parking Permit Zone along Roads
5. Conclusion
6. References

**Abstract**

<TBD>

**Executive Summary**

The city of Chicago, located south of Lake Michigan, is the third largest city in the United States, one of the world's most prestigious international financial centers, and is also known as the home of the skyscraper. Chicago is the largest business downtown and one of the largest futures markets in the United States, and it boasts several world-class theaters, museums, parks, and stadiums. It attracts millions of visitors to the city each year. In addition to the constant stream of visitors, the population living in the city has exceeded one million. Chicago's venues host special events during the week, weekends, and holidays, including concerts, conferences, holiday celebrations, and sports. Due to the high volume of foot and vehicle traffic, Chicago requires excellent road planning and plenty of parking to provide smooth access and parking for the thriving city.

With thousands of people driving through the city for events such as concerts, conventions, festivals, and sports, the demand for parking near venues is increasing. It is not uncommon to see large numbers of motorists looking for parking spaces near venues where the demand exceeds the supply. This demand outstrips supply, creating parking difficulties and street congestion. As a result, for several years, Chicago has been ranked as one of the top five cities in the county for congestion.

According to Chicago city traffic statistics, an average of about 1 million motorists drive on the streets every day. Imagine if a third of those drivers spent 10 minutes looking for a parking space, there would be more than 50,000 hours of extra driving time on the city's streets. As a result of this extraordinary congestion, traffic congestion caused by congestion, and air quality degradation becomes an imminent urban problem. As a result, increasing the number of parking lots, parking capacity, and the popularity of intelligent parking has become an issue that governments and businesses are scrambling to solve.

An indispensable part of Chicago's urban planning is urban parking planning. The rational planning, supply, and demand of parking spaces are the premise and basis of the planning. Due to the rapid development of motor vehicles, the scarcity of land resources, and the imbalance between supply and demand, parking has become the primary problem of urban planning in Chicago. It is important to note that securing sustainable transportation development is the key to planning urban parking. Some of the needs that we should consider in the parking demand analysis, depending on the planning area, are the overall needs of the City of Chicago, the needs of each area, and the needs of micro-siting. According to the parking demand generation mechanism, we need to consider basic parking demand and social parking demand. And the main factors that influence the parking demand are land use, the impact of overall city land use on parking demand, population and its changes, level of socio-economic development, the standard of living and social interaction, and cultural and tourist activities. Therefore, parking planning is a very complex issue due to the multiple demand considerations.

Every year the City of Chicago invests huge amounts of assets in the construction of urban roads to improve urban traffic and dynamic traffic driving conditions. However, we still need more parking spaces to alleviate the parking difficulties caused by the growing population and the city's prosperity. Therefore, the following parking planning principles can help us to better plan our parking lots. Firstly, the parking lot should be based on the overall planning of the city, and the parking lot needs to be well laid out, meet the specifications and meet the technical requirements. Secondly, parking lots need to save the land and be appropriate to the location. Urban public parking lots should be set up close to the service objects, and the site selection should also meet the requirements of the urban environment and vehicle traffic. At the same time, the planning of parking lots should be conducive to the improvement of the urban traffic environment. Finally, in the planning of public parking lots, the corresponding parking lots should be built in strict accordance with the relevant regulations, so as to solve the problem of difficult parking for urban public parking, government agencies, and large public buildings.

**Preface**

Parking space planning is part of urban planning, a process of land usage design and regulation in urban areas (Urban Planning, 2016). Urban planning concerns itself with both the development of open land and the revitalization of existing parts of the city, thereby involving both location selection of new parking structures and relocation of existing parking structures (Fainstein, n.d.).

Chicago is one of the biggest cities in the United State since it was founded in 1830. Prosperity often means being crowded and expensive. In the past years, Chicago’s parking rates have increased about 12 percent per year, compared with 4.4 percent for the country as a whole (Diesenhouse, 2007). Old buildings in downtown Chicago restrict the city’s parking capacity, but the increasing number of vehicles requires more parking spaces. Therefore, this leads to urgency in making effective and efficient urban planning for Chicago downtown’s parking structures.

**Urban Planning**

Urban planning, also known as city planning is a technical and political process that is focused on the development and design of land use and the built environment, including air, water, and infrastructures passing into and out of urban areas, such as transportation, communications, and distribution networks and their accessibility (<https://en.wikipedia.org/wiki/Urban_planning>).

Transportation is an integral part of all urban planning. Especially after the industrial revolution, the rapid rise in population and the increase in vehicle ownership among the population have made urban transportation planning more important. In the midst of transportation planning, parking planning is an easily overlooked point. Many people think that transportation planning is just about roads and public transportation. But 95 percent of the time, cars are parked and only 5 percent of the time, they are moving (<https://usa.streetsblog.org/2016/03/10/its-true-the-typical-car-is-parked-95-percent-of-the-time/>). Good planning of parking can also divert traffic to avoid accidents and congestion caused by parking and temporary stopping. As a resident of Chicago myself, I suffer from parking problems. Parking has also become a problem for me every time I travel. Therefore, I want to use data science to help ease and solve this problem.

**Development of Parking Planning**

The history of traffic planning and regulation is short compared with other urban planning topics. Until 1900, American cities had virtually no traffic regulations. There were no federal regulations, and the existing traffic rules were almost entirely the result of ad hoc legislation by municipalities. (<https://thereader.mitpress.mit.edu/brief-cultural-history-of-the-parking-lot/>). The father of parking rules, William Phelps Eno, published "Rules of the Road" in 1909 and it was adopted by the New York City government. This marks the first traffic and parking law established in the U.S. Eno was the first person to introduce the concept of dead and live cars, and. In “The Storage of Dead Vehicles on Roadways,” Eno writes: “When vehicles are ranked, no one of them can move out of the line independently of the others unless considerable waste space is allowed for between them, whereas when they are parked, being parallel to one another, any one of them can get away without causing any other one to move.” (The Storage of Dead Vehicles on Roadways) Further, he proposed a parking hierarchy that would treat parking for cars without owners differently from parking for cars with owners. This was a precursor to the various types of road parking permits in later city planning. He also predicted the emergence of parking lots. It can be said that the modern urban parking system was established by Eno. My research today also explores the possibility of improvement under this system and does not go beyond it.

In the United States, parking facilities are divided into two main categories: on-street parking facilities and off-street parking facilities, and each category is subdivided into different subcategories. On-street parking facilities are further divided into long-stay and temporary parking. Off-street parking facilities are also divided into parking lots and parking garages. In terms of parking methods, they are also divided into self-parking and valet. As technology advances, we see changes in the middle of each category. For example, it is now possible to pay for parking by cell phone, and a lot of parking space-finding software has emerged.

**Parking Planning in Chicago**

Parking tickets per year in Chicago are a lot.

<TBD Chicago parking system, Chicago parking situation, Chicago ticket system, Chicago ticket situation>

**Data Analysis with Machine Learning in Chicago Downtown**

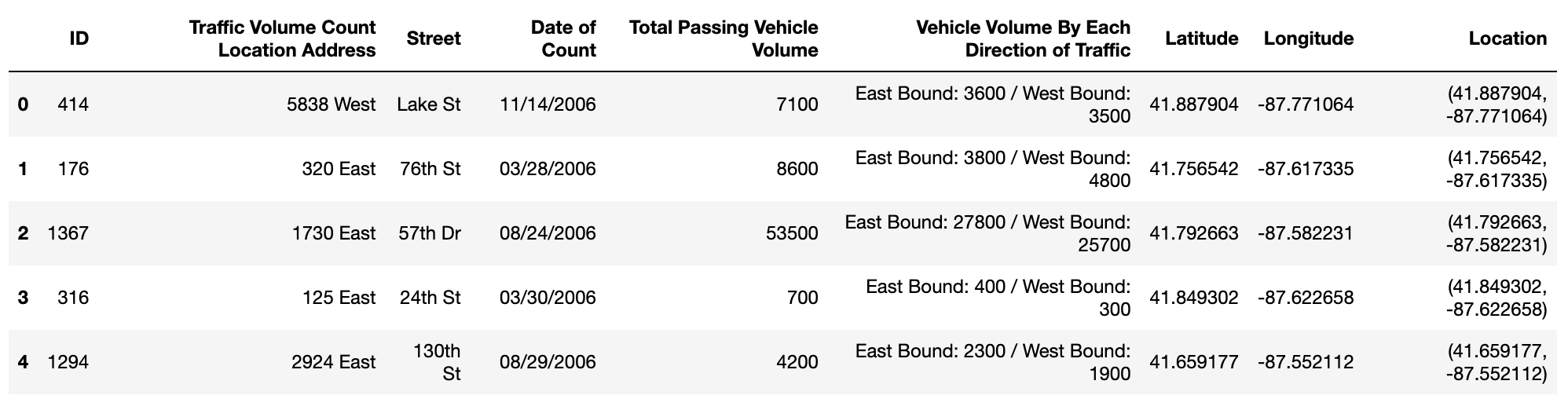
In my Capstone Project, I want to investigate solutions to the shortage of parking spaces in Chicago downtown. I am going to use the River North area as a case study because River North is the busiest district in Chicago’s downtown and faces the most severe crisis in parking space shortage. I will use machine learning to analyze the parking shortage issue. Machine learning is a methodology that applies computing algorithms to analyze sample data to achieve certain goals. I will use two methods, classification and regression (Matijosaitiene, McDowald, Juneja, 2019) to predict and visualize outcomes from machine learning based on the Parking Permit Zones Data, Traffic Data, and Parking Violation Tickets Data. To model parking in the Michigan Avenue area, I am going to apply machine learning techniques to simulate the parking flow based on traffic data and parking violation data. Results of the modeling would show the insufficiency of parking spaces in both time-based and space-based.

From the space-based simulation, I would like to discuss where potential choices for building new parking facilities are and how big the parking facilities are. I would also discuss the potential relocation or expansion of current parking facilities based on this simulation. From the time-based simulation, I would like to discuss the effectiveness of existing parking facilities over time.

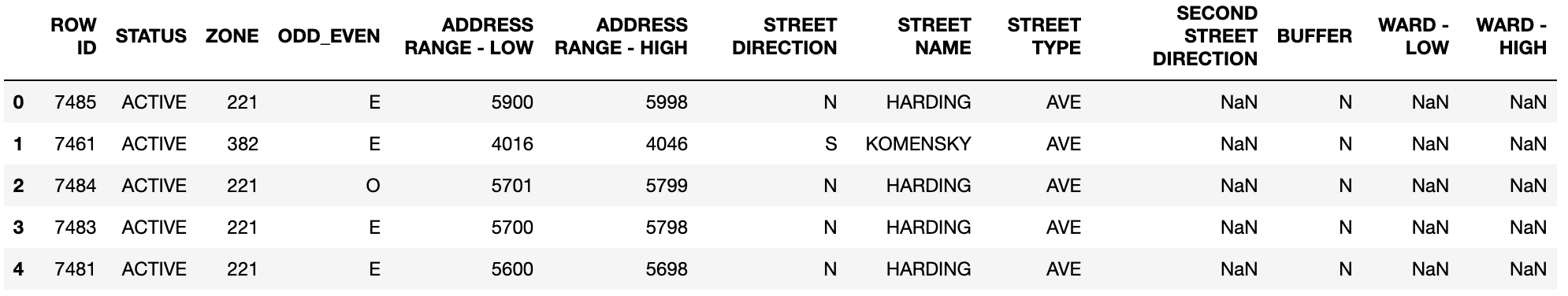
This study will show the significance of machine learning and data analysis techniques in urban planning. The results of the study may potentially support the Chicago government’s future decision-making and facilitate sustainable parking planning (Matijosaitiene, McDowald, Juneja, 2019). I am also a driver in Chicago and suffer from the poor parking situation in downtown Chicago; thus, I want to contribute to changing this situation.

**Datasets Information**

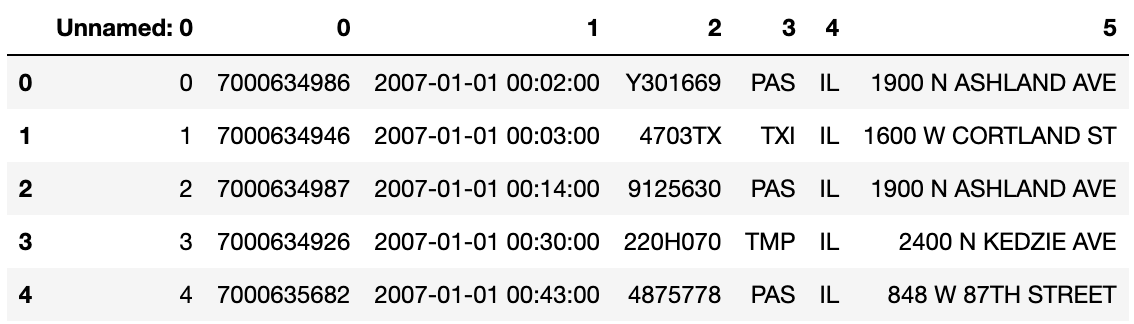
For the traffic data, I get Average Daily Traffic counts from the Chicago government’s dataset (Chicago Metropolitan Agency for Planning). Average Daily Traffic (ADT) counts are analogous to a census count of vehicles on city streets. These counts provide a close approximation to the actual number of vehicles passing through a given location on an average weekday (<https://data.cityofchicago.org/Transportation/Average-Daily-Traffic-Counts/pfsx-4n4m>).



From the above preview of data, we can see that the dataset contains Date, Location, Latitude, Longitude and Traffic Volume at that day. Date and Location will be used by me to aggregate with other datasets. Latitude and Longitude will be used to help locate the spots on the map. Traffic Volume is the main concerning element here. I believe Traffic Volume is highly correlated to the parking situation at the location. Therefore, I would like to aggregate the Volume with Parking Permit Zone numbers to predict tickets and parking situation in this location.

For the data on parking permit zones along the road, I will get it from the Chicago government’s data portal (Chicago Data Portal). ​​This dataset contains all those street segments (individually uniquely identified by Record ID) that have been designated as belonging to a “Residential Parking Zone.” (<https://data.cityofchicago.org/Transportation/Parking-Permit-Zones/u9xt-hiju>) I downloaded data from the data portal. 

From the above preview of data, we can see that the dataset contains Address and Zone location. I will need to combine the Street Direction, the Street Name and the Street Type together to retrieve the comprehensive address information matching the Address data in Traffic table. I will need to aggregate the zone id to get parking permit spaces on each street. Then, I can get the variable Parking Permit Count to be used to predict parking situation.

For the ticket data, I get it from ProPublica Data Store’s City of Chicago tickets dataset. This dataset provides details on all parking and vehicle compliance tickets issued in Chicago from January 1, 1996 to May 14, 2018 (<https://www.propublica.org/datastore/dataset/chicago-parking-ticket-data>). Here is the preview of the dataset. 

The dataset is huge because it contains tickets information from 1966 to 2018. I will first filter out based on the date. Then, I will aggregate the tickets cound on each day for each street to later join on with previous data tables.

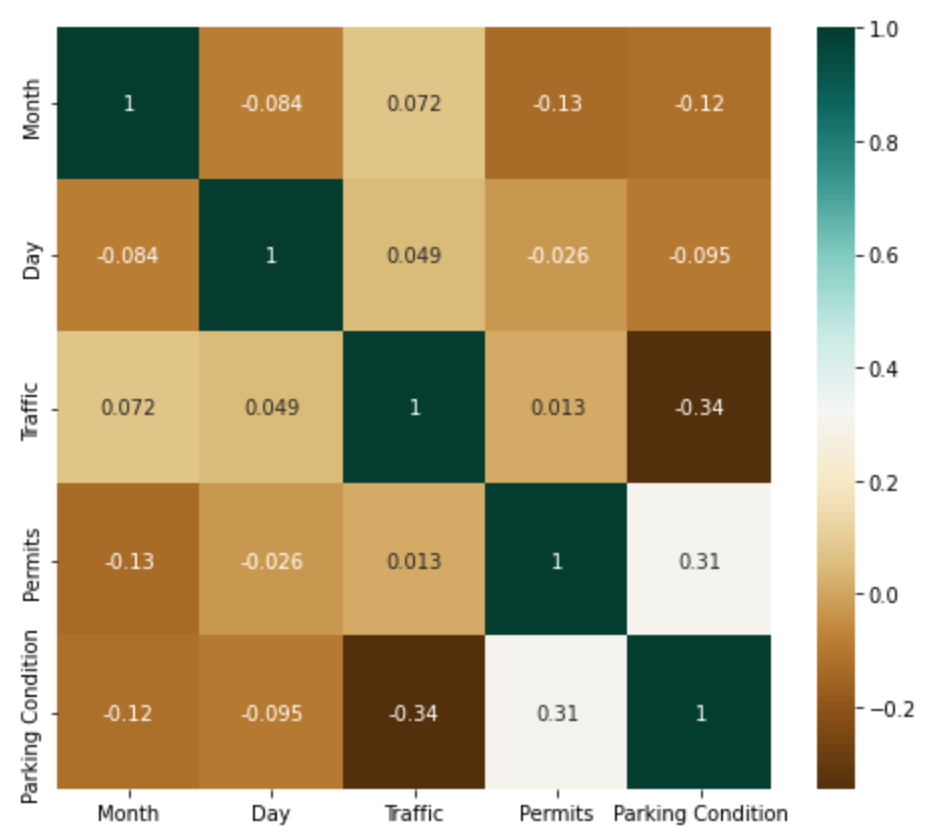
I will also use records from some of the parking facilities’ websites. In addition, I will also use vacant buildings or spaces from Chicago Data Portal.

**Methodology Overview**

Before conducting any model development, I will first conduct Data Preparation process. In data preparation, I will first clean all the data to squeeze data in the same time period and location area. Then, I will aggregate zones and tickets to get counts in each location. Finally, I will join all the tables together to form the final dataframce can be use to train and test data.

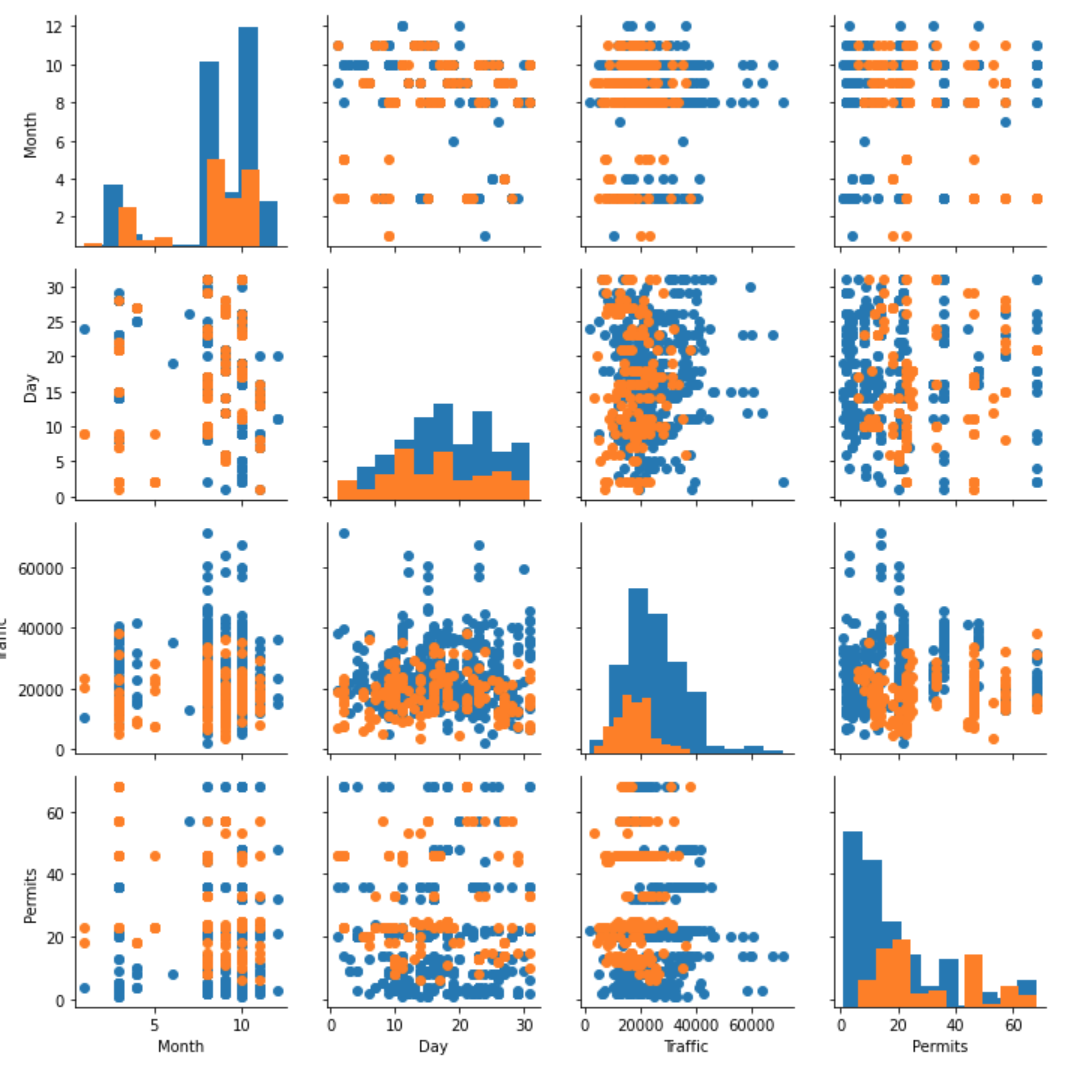
I also created a new variable called Parking Condition. This variable represents the parking situation in this location. Moreover, this variable is derived from this location’s parking space versus this location’s parking tickets. If tickets number overwhelm the parking space number, I mark parking condition as 0 which means the parking condition is bad. On the other hand, if there are more space counts than the tickets number, I mark the variable as 1 to show it’s good.

After data is cleaned and intially prepared, I make exploratory data analysis on the data to investigate relationships between each variables. The first methodology I took is correlation matrix. Correlation matrix is a matrix showing correlations between each pair of variables.



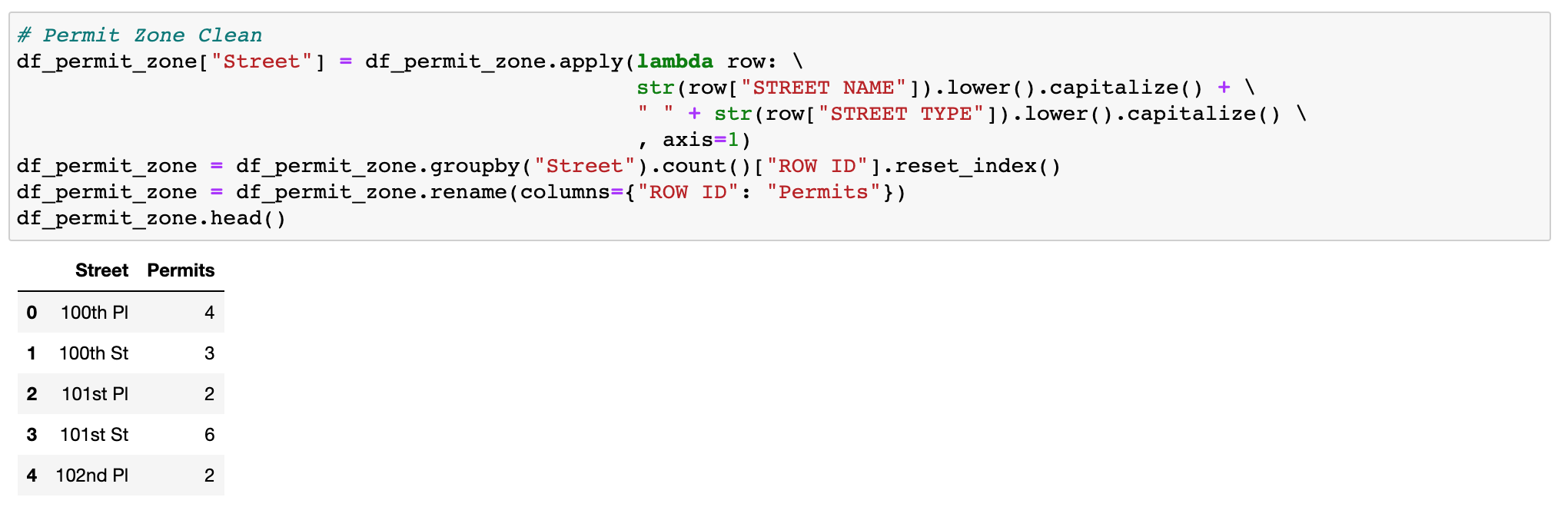
As we can see, the variable parking condition is more correlated to Permits count and Traffic volume than the other variables. Therefore, these two variables will be the two main variables as inputs for model training and predictions.

The next analysis I did is that I plot pair girds of all the variables. This plot can help us visualize relations between variables by ploting out their correlated distributions.



**Data Preparation**

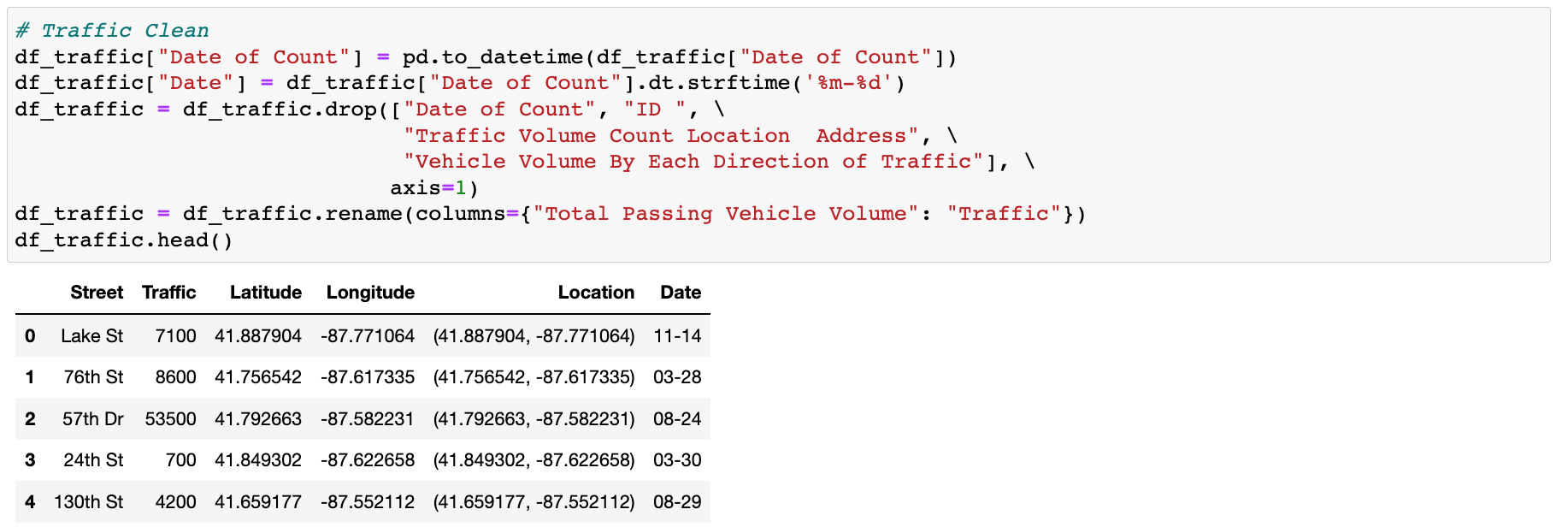
In the data preparation stage, I clean and engineer my data using Python libraries including Pandas and Numpy. Both libraries are open source tools can be used to do data analysis and data manipulation. Pandas is more focused on table data and numpy is more focused on series data.



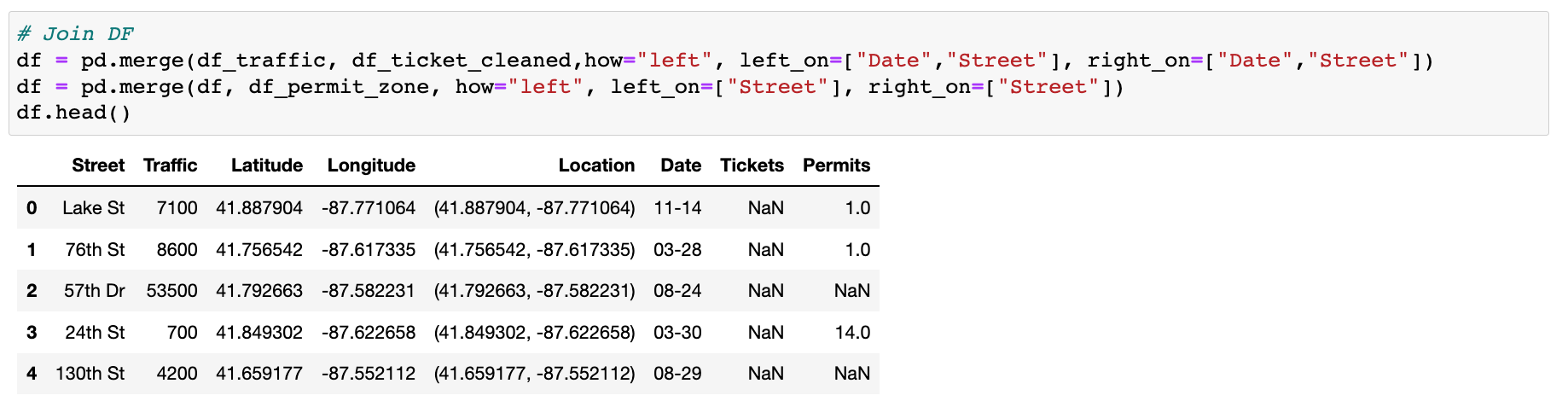
The main steps I took here for parking permit zones are street name transformation and permit zones aggegregation. Street name transformation is done through Pandas apply function which can apply a self-defined transformation to all the rows among one or several columns. The aggregation is done by groupby function. This is a Pandas built-in aggregation function can be interacted with other operation functions like count, max, and mean. I choose count here as the most important point here is number of parking spaces in one parking permit zone.



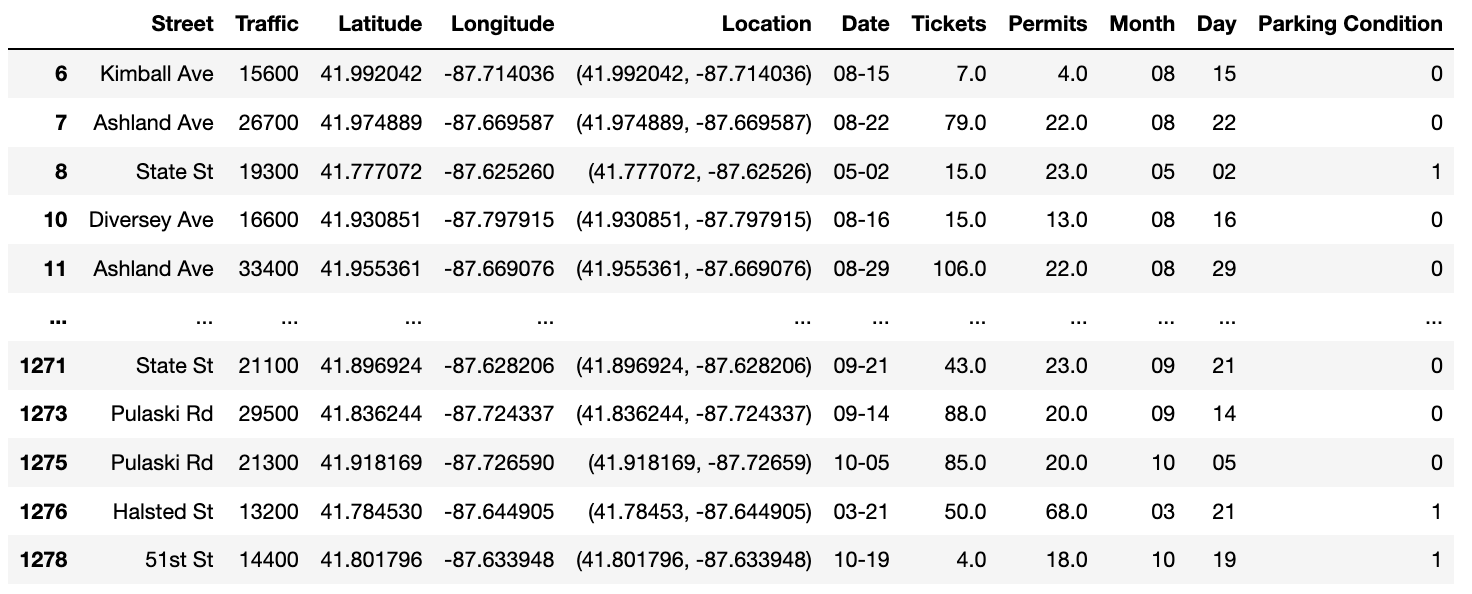
Here I also did address transformation to gain the street name like the previous dataset. One thing worth noting is that I am not only aggregating the dataset with Street but also on Date because I believe date is another essential element that influencing the parking situation. Therefore, I standardlized the original date column into Python datetime object so that I can utilize the date in the following analysis.

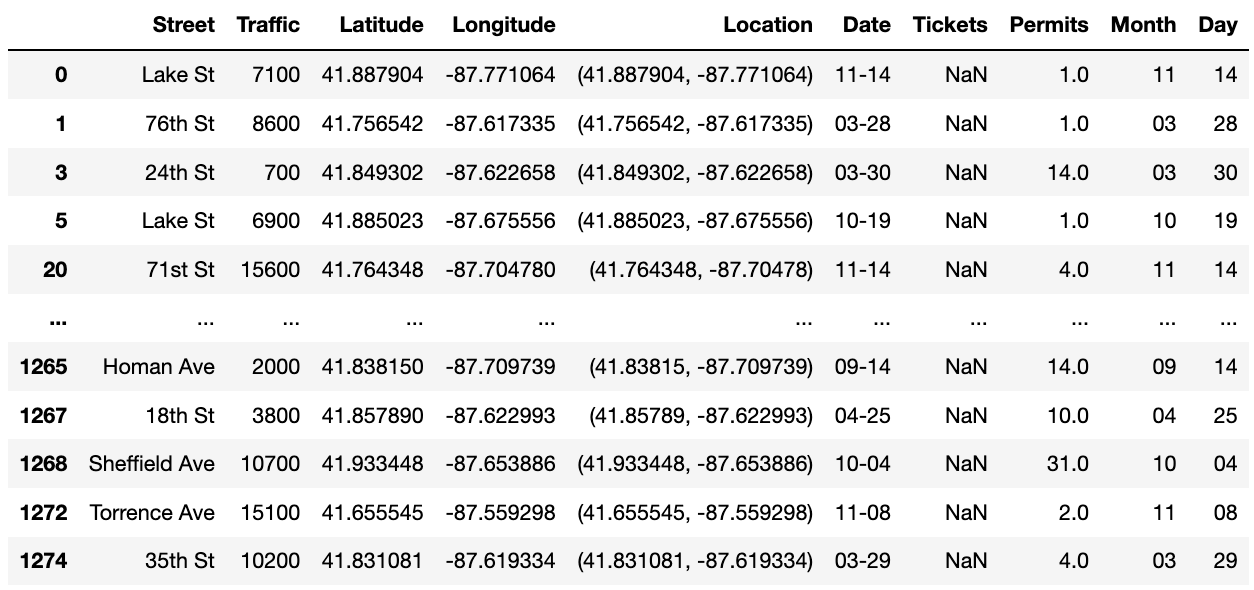


Traffic data is the most well formed one among all three. Therefore, I mainly just did some renaming and filtering with drop function to standarize the dataset. I didn’t transform Street here because I transform the previous two datasets based on the Traffic dataset.



After joining the three datasets, I find there are discrepancy between the essential variables: Traffic, Tickets, and Permits. Therefore, I did drop those null values after when creating the training and testing dataset. Training dataset look below after final trimming. It contains both inputs like Month, Day, Traffic, and Permits. It also contains the target variable Parking Condition. The target variable, as discussed above, is derived by comparing parking spaces versus tickets number to see whether this street’s parking situation is above or below average during certain amount of time. The criteria can vary based on the design. For example, I can also use a threshold of ticket to evaluate the parking condition of this area. The test dataset does not contain target variable because we will need predict the target variable in test dataset with trained models. The difference can be seen by the below two previews of train dataset and test dataset.





**Machine Learning Model Development**

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks (<https://en.wikipedia.org/wiki/Machine_learning>). The tasks here is to predict parking condition for unevaluated areas with their traffic and parking space values. Generally, the models will pick train data and develop coefficient from inputs variables and target variables. Then, we can use those pre-trained or calculated coefficients to predict or calculate target variables with input data from test dataset.



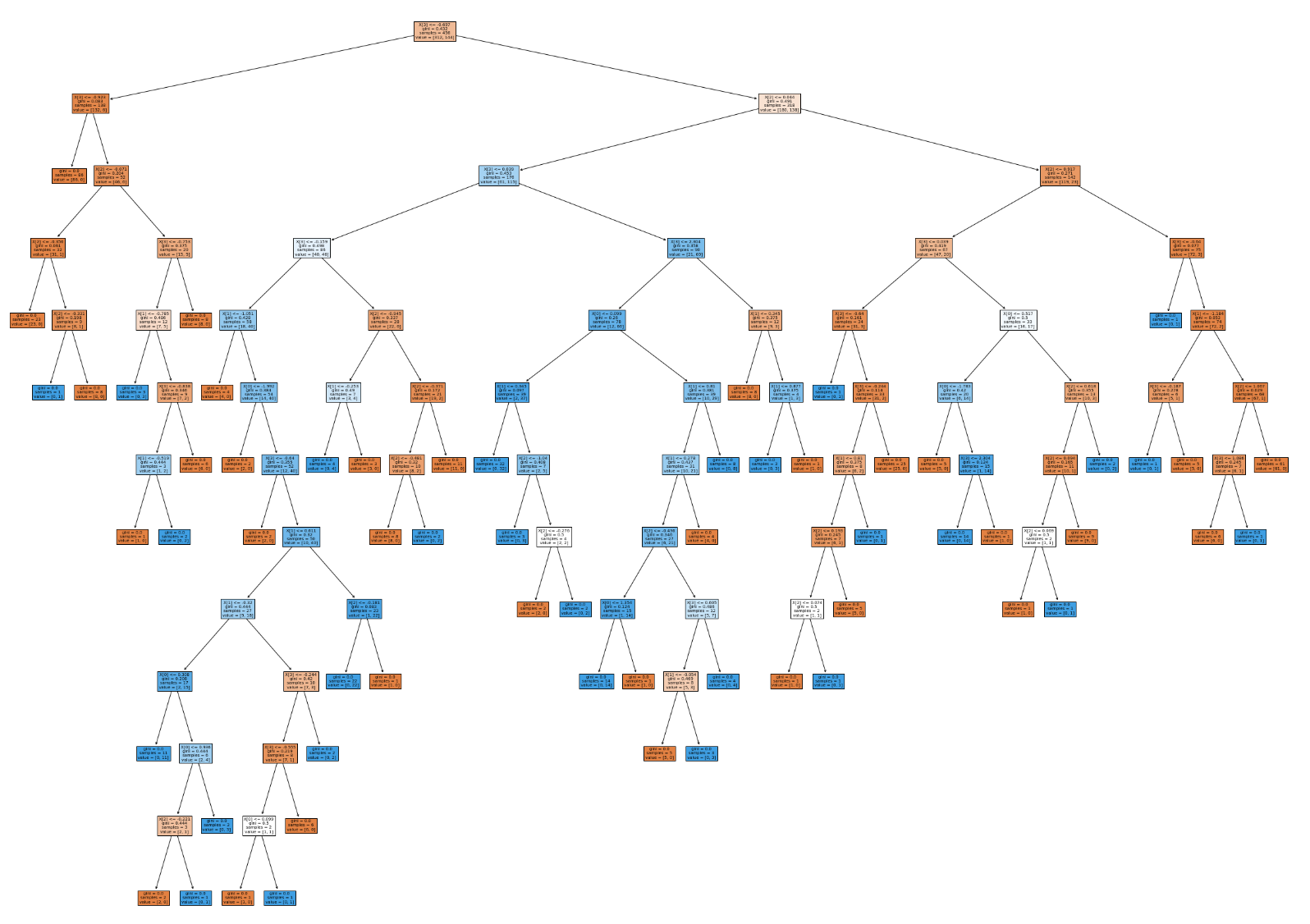
Here I am splitting dataset into X and y. Variable X stands for input variables in both train and test dataset. Variable y stands for target variables in train dataset. Further, I rename them as X\_train, y\_train, and X\_test. And I extract 20 percent data from X\_train and y\_train as X\_valid and y\_valid to help me evaluate our model and prediction. The test dataset is blind, which means I do not know the true target variable even after I predict. Therefore, the validation data, which have both input data and target data but is not involved in training process, can be used to calculate accuracy for validation.

Then, I also did standardization on the dataset. The reason why I did this is because our data varies a lot in the absolute values. The Traffic data is much larger than the rest data. This could cause uneven weights when calculating coefficient. Therefore, I map all variables against the maximum and minimum value in each data through standardrization to minimize the range of data but keep the distribution.



I am using classifiers here to do the supervised model training and predicting because the target variable of our dataset is a binary variable which only has 0 and 1. If I am going to predict some other variables like Ticket number, I will need to use regressor. Classifier algorithms employ sophisticated mathematical and statistical methods to generate predictions about the likelihood of a data input being classified in a given way (<https://c3.ai/glossary/data-science/classifier/#:~:text=What%20is%20a%20Classifier%3F,%E2%80%9D%20or%20%E2%80%9Cperson%E2%80%9D>).).

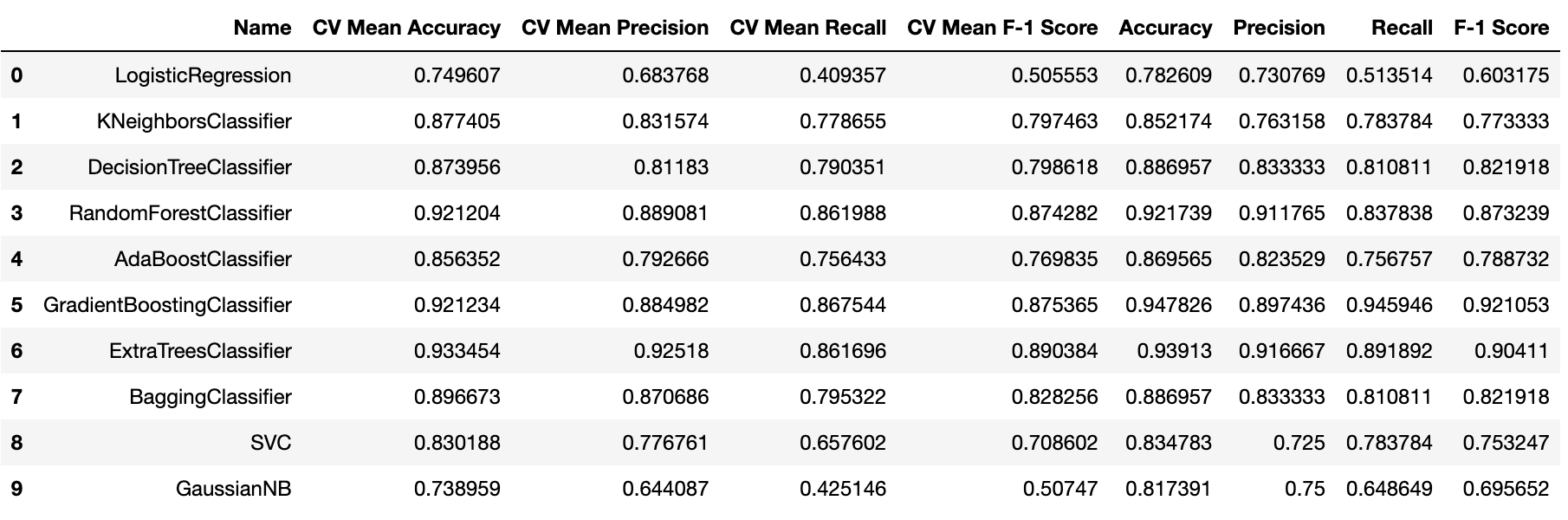
I choose a bunch of classifiers here. Logistic Classifier is mainly calculating probability each row towards 1 or 0 and assigning for the larger probability label. K-Neighbour Classifier is classify each label based on their k nearest neightbours so it’s a similarity classification process. SVC, Support Vector Machine Classifier, means to create a vector (line in 2D and plane in 3D) to separate rows into two classes. All the rest classifiers are Tree Based algorithms. Tree Based classifiers usually classify classes based on the entropy difference between children nodes and the parent node. In another word, the tree classifiers will try to differentiate the elements or inputs as much as they can. Here is an image of the actual structure of our Decision Tree Model.



For the rest classifiers, they are all based on Decision Tree. However, all of them, like Random Forest and Gradient Boosting, are formed by a bunch of trees. The difference lies in how they combine those trees together. Random Forest tools the weighted average of each tree’s result as the final output. Gradient Boosting uses Gradient algorithm to refine coefficients in tree in each round of training to minimize the loss. Ada Boost weighted the inputs for trees in each round to strength those wrongly classified data to reach a better result.

In the model evaluation, I uses Accuracy, Precision, Recall, and F1 Score. All testing requirements are calculated from the four essential criterias: False Positive, False Negative, True Positive and True Negative. In this case, False Positive means that rrong detect good parking condition as bad. False Negative means the model didn’t catch bad parking conditions. True Positive means the model prediction catches bad parking conditions. True Negative means the model successfully identify good parking conditions. Based on these information, Accuracy is (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative). Accuracy is very balance and therefore the most common used test score method. However, it cannot show the model’s bias between False Positive and False Negative. Therefore, I also uses Precision and Recall. Precision, which is True Positive / (True Positive + False Positive), is related to whether we pull false alarms on good parking conditions. Recall, which is True Positive / (True Positive + False Negative), is related to how many bad parking conditions we catched among all the streets. F1 = 2 \* (Precision \* Recall) / (Precision + Recall). I also calculated F1 because I want to increase both precision and recall to increase the reliability of model’s prediction.

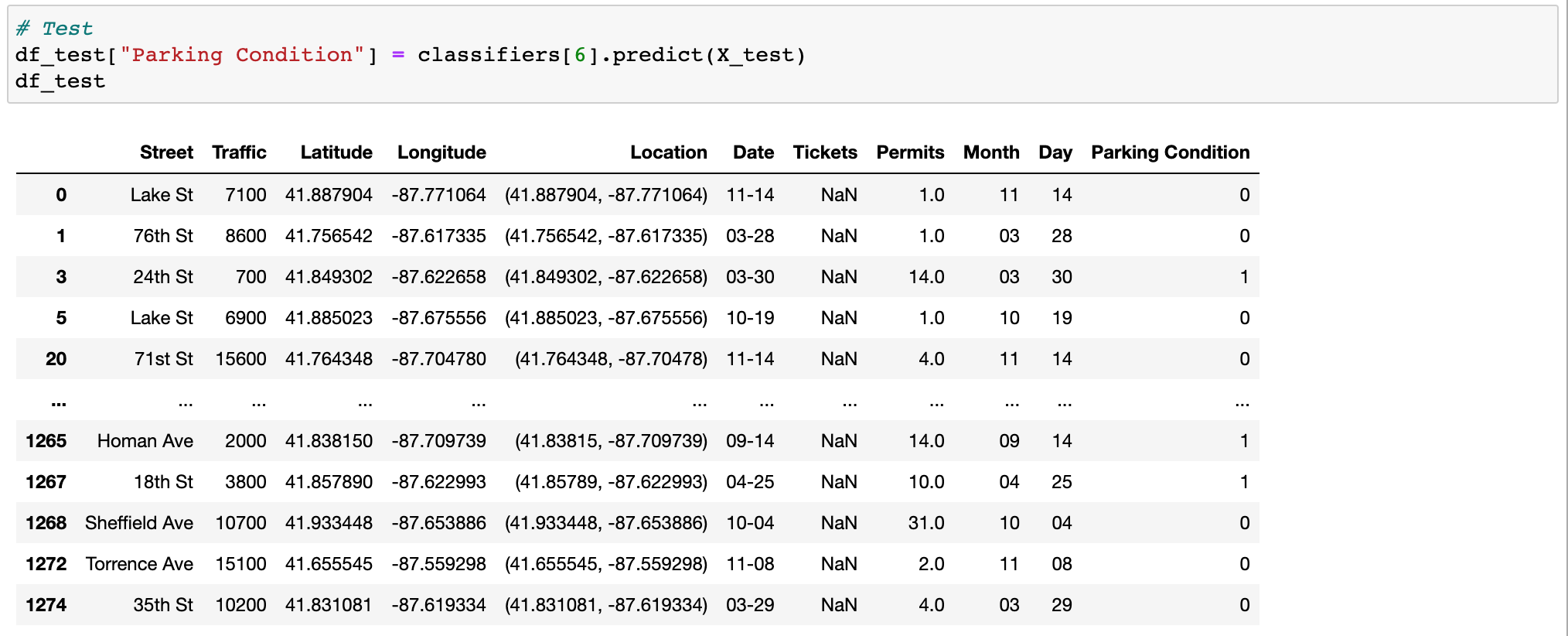
Another technology involved here is the cross validation. Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set (<https://en.wikipedia.org/wiki/Cross-validation_(statistics)>). This is similar to my intention in splitting out the validation dataset. However, with more variance in the cross validation. I can identify the variance in the model prediction results and avoid overfitting.



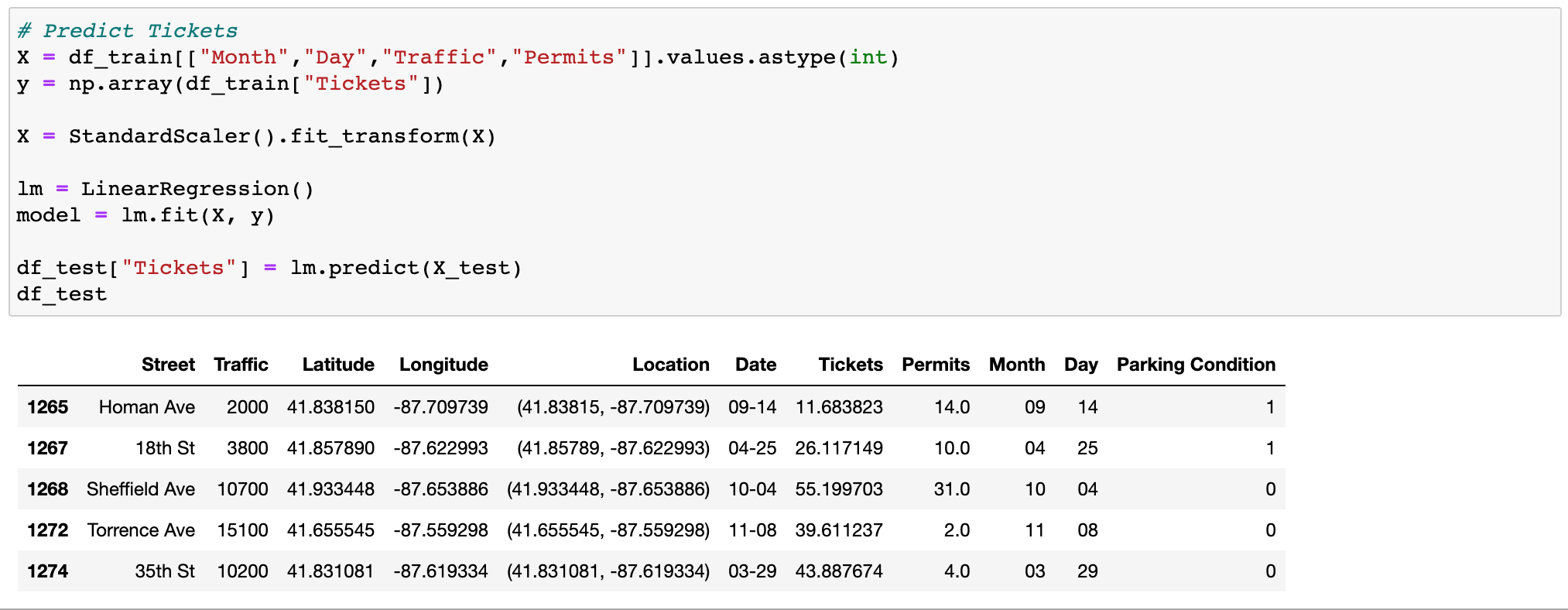
From the results, the tree algorithms are doing better than the other types fo classifiers. This could be the reason that our data is pretty complicated and tree algorithms tend to have higher complexity to handle more complicated data. Among tree and forest classifiers, the best performer is ExtraTreeClassifier. ExtraTreeClassifier implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html>). Also because of the complexity of the data, decreasing in bias often companies the tradeoff of increasing variance. ExtraTreeClassifier helps avoid overfit thus decreasing the variance. Both Accuracy and F1 score of ExtraTreeClassifier are over 90 percent and this is a very good result. Therefore, I choose to use ExtraTreeClassifier as my final predicting model.

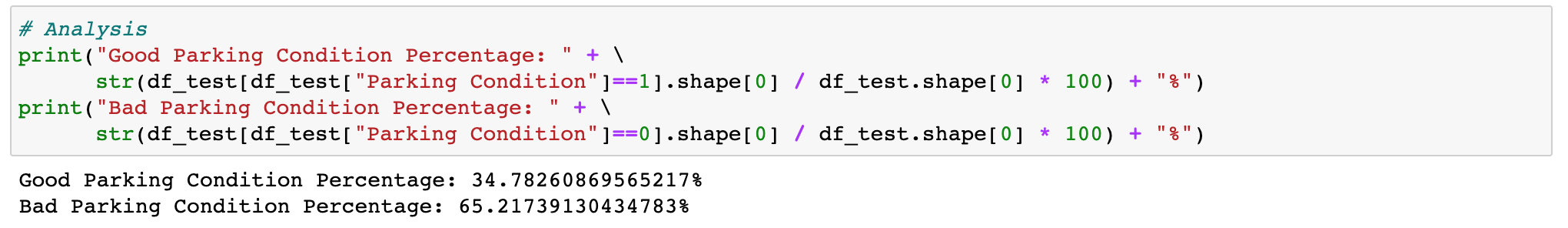
**Results Discussion and Comparison**

Here is the prediction result of ExtraTreeClassifier on the test dataset.



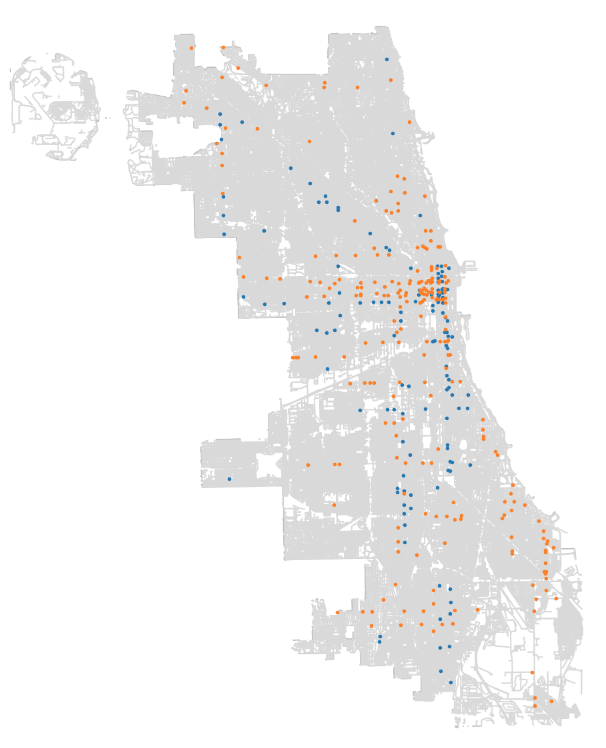
Furthermore, I use regressor to predict Ticket numbers in test dataset to help analyze.

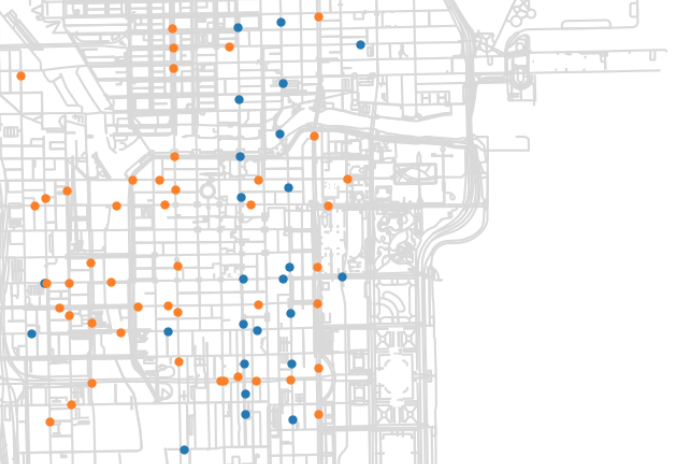


Apparently, the overall condition is bad from the percentage of good and bad parking conditions calculated below. 

The Bad Parking Condition is almost twice as many as those good parking conditions.

<TBD Geographic mapping to do location analysis>





**Suggestions**

**Current Parking Structures**

<TBD>

**Future Parking Structures**

<TBD>

**Parking Permit Zone along Roads**

<TBD>

**Conclusion**

<TBD>

**References**

*Chicago parking rates accelerate: [Chicago Edition] - ProQuest*. (n.d.). Retrieved October 26, 2021, from http://www.proquest.com/docview/420619063?pq-origsite=primo&accountid=12861

*City of Chicago | Data Portal*. (n.d.). Chicago. Retrieved October 26, 2021, from https://data.cityofchicago.org/browse?tags=parking

Matijosaitiene, I., McDowald, A., & Juneja, V. (2019). Predicting Safe Parking Spaces: A Machine Learning Approach to Geospatial Urban and Crime Data. *Sustainability (Basel, Switzerland)*, *11*(10), 2848-. https://doi.org/10.3390/su11102848

*Traffic Data—CMAP*. (n.d.). Retrieved October 26, 2021, from https://www.cmap.illinois.gov/data/transportation/traffic

Urban Planning. (2016). In *Urban Planning.* Cogitatio Press.

*Urban planning | Definition, History, Examples, Importance, & Facts*. (n.d.). Encyclopedia Britannica. Retrieved October 26, 2021, from https://www.britannica.com/topic/urban-planning

**Additional References**

[*(Barton-Aschman Associates & Chicago Central Area Committee, 1965)*. (n.d.). Retrieved October 12, 2021, from](https://www.zotero.org/google-docs/?nBBy1t) <https://search.library.northwestern.edu/discovery/fulldisplay?docid=cdi_gale_infotracacademiconefile_A649017550&context=PC&vid=01NWU_INST:NULVNEW&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,%22machine%20learning%22%20urban%20planning%20location&offset=10>

[Bandyopadhyay, M., Rout, M., & Satapathy, S. C. (2021). *Machine learning approaches for urban computing*. Springer.](https://www.zotero.org/google-docs/?nBBy1t)

[*Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning—Northwestern University*. (n.d.). Retrieved October 12, 2021, from](https://www.zotero.org/google-docs/?nBBy1t) <https://search.library.northwestern.edu/discovery/fulldisplay?docid=cdi_crossref_primary_10_1016_j_jtrangeo_2019_102479&context=PC&vid=01NWU_INST:NULVNEW&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,%22machine%20learning%22%20urban%20planning%20location&offset=10>

[Milusheva, S., Marty, R., Bedoya, G., Williams, S., Resor, E., & Legovini, A. (2021). Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PloS One, 16*(2), e0244317–e0244317.](https://www.zotero.org/google-docs/?nBBy1t) <https://doi.org/10.1371/journal.pone.0244317>