The Big Apple on Citi Bikes

Zexi Ye (zy1311), Zian Chen (zc674), Jiayu Qiu (jq429) and Yuxing Wang (xw913)

Abstract— **This project investigates the traffic flow patterns of the NYC CitiBike system, examines the link between its fluctuation and external factors, and develops a predictive model that forecasts abnormal traffic flow on an hourly basis. Employing a combination of data joining, data transformation, data visualization and machine learning techniques, we provided an intuitive and comprehensive analysis and derived insightful results from our findings, laying the groundwork for future business application in an urban environment. Our statistics exhibit the fact that overall the CitiBike traffic flow follows a periodic trajectory and its level is governed by geographic locations, while its day-to-day fluctuation heavily hinges on weather. The results demonstrate strong explanatory power regarding the variation in CitiBike traffic flow, and our model is able to predict abnormal traffic flow based upon a collection of features.**

I. Introduction and Motivation

One of the most popular bicycle sharing systems in the United States, CitiBike operates 750 stations and 12,000 bikes in the New York Metropolitan Area, providing a convenient and affordable way of transportation to New Yorkers. Nevertheless, maintaining such a gigantic bike network efficiently while meeting the needs of the customers requires a profound understanding of the dynamics of bike traffic flow, which guides decision-making and optimization.

It is evident that bike traffic flow varies greatly in different neighborhoods at different time of the day, as one may observe on the streets of New York, yet what does the big picture look like if we examine the system as a whole? Additionally, how can one benefit from making more informed decisions by leveraging such knowledge? Notably, one practical issue is the anticipation of abnormal traffic flow. When traffic flow that deviates from the normal pattern occurs, what can be done to foresee and respond to the anomaly?

Undoubtedly, difficulties will arise in the course of formulating the solution. Firstly, the volume of data is tremendous. CitiBike is marked by its enormous ridership and the associated dataset covers all records in great details, making the handling of data challenging. Secondly, the data might not be in the desired form, so a systematic transformation of the data is required. Thirdly, the exploration of explanatory features can be tough. It is necessary to narrow down our candidate features into a pool with a reasonable size to pave the way for analysis.

To gain a holistic understanding of the problem, we will discover the underlying factors that dictate the pulses of Citi bikes and develop an effective model that predicts abnormality in bike traffic flow. At the end, we will conclude this project with insights into business application.

II. Methodology

A. Data Source

Fundamentally, the datasets in this project consist of two parts - the NYC CitiBike Trip Histories[1] and the National Oceanic and Atmospheric Administration (NOAA) historical weather[2]. In consideration of data volume, integrity of time period and potential concept drift, we decide to focus on historical data in the year 2017 (from 01/01/2017 to 12/31/2017), since 2017 is close to the present and the entire year's data are available. The former was retrieved directly from CitiBike's online database in a csv file, while the latter was requested from NOAA's database with the specified weather station Central Park, New York. For simplicity, in later sections we will refer to the former as the trip data and the latter, the weather data.

B. Data Cleaning

The trip data consist of detailed information about each trip that occurred in the period of interest. In this case, columns truly relevant to our analysis are information about the date/time, start station, and end station of each trip. While there do exist missing or erroneous fields in the dataset, it turns out that the aforementioned columns are all valid. Hence, we left the trip dataset as it was.

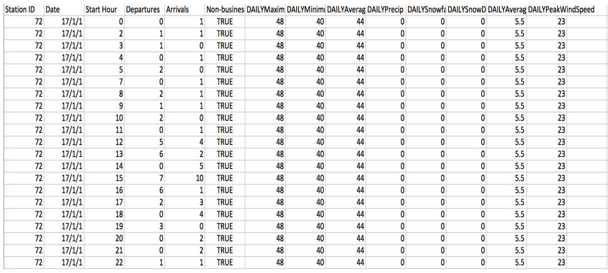
The weather data, however, do contain missing values. Note that we opted to use daily weather as our features, so we only kept records whose report type is "SOD," which aggregates each day's weather data into one row. We observed that data are missing on a few days, which is not a significant portion. Thus, we forward-filled the NA's to ensure complete coverage of all days in 2017.

C. Data Transformation and Joining

Due to the nature of our analysis, trip data had to be extensively transformed prior to any manipulation. The original trip data document the information of each historical trip and the trips are sorted in chronological order. Our desired data frame, in contrast, would follow this structure: each row has the fields Station ID, Date, Hour, Departures and Arrivals.

In order to obtain the desired form of the trip data, we applied hierarchical groupby method twice on the original data frame, by Start Station and End Station respectively, with two other tokens being Date and Hour. Thus we created two groups of data. For each group, we called count() to obtain a data frame that records the number of departures or arrivals on a station-date-hour basis. We then joined the two data frames and obtained the desired form specified above.

We performed a simple feature selection using intuition on the weather data and kept a collection of attributes that are most germane to our analysis. Additionally, using Pandas' built-in calendar, we created a column of booleans that indicates whether a day is a business day or not (either a weekend or a federal holiday) for the year 2017.



Station ID	Date	Start Hour	Departures	Arrivals	Non-busines	DAILYMaxim	DAILYMinim	DAILYAverag	DAILYPrecip	DAILYSnowfa	DAILYSnowD	DAILYAverag	DAILYPeakWindSpeed
72	17/1/1	0	0	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	2	1	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	3	1	0	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	4	0	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	5	2	0	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	7	0	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	8	2	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	9	1	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	10	2	0	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	11	0	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	12	5	4	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	13	6	2	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	14	0	5	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	15	7	10	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	16	6	1	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	17	2	3	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	18	0	4	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	19	3	0	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	20	0	2	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	21	0	2	TRUE	48	40	44	0	0	0	5.5	23
72	17/1/1	22	1	1	TRUE	48	40	44	0	0	0	5.5	23

Fig. 1. Preprocessed Data

Finally, the weather data and the business day boolean column were left joined into the transformed trip data frame on the key "Date." This data frame will be essential in the later analysis and modeling.

D. Modeling

We built a dummy variable which indicates at a certain station, whether its traffic of the hour is abnormal (not in the range of mean ± 1.96*standard deviation).We picked Start Hour, Daily Average Temperature, Business Day, Wind Speed, Daily precipitation and Daily Snow as our model features.

In order to predict abnormal traffic, we randomly split the data into training and testing dataset and used four common machine learning models,Logistic Regression, SVM (linear and RBF kernel), KNN and Decision Trees to train the training data. We calculated the R-Square score of each model and plotted ROC curves to compare their performances.

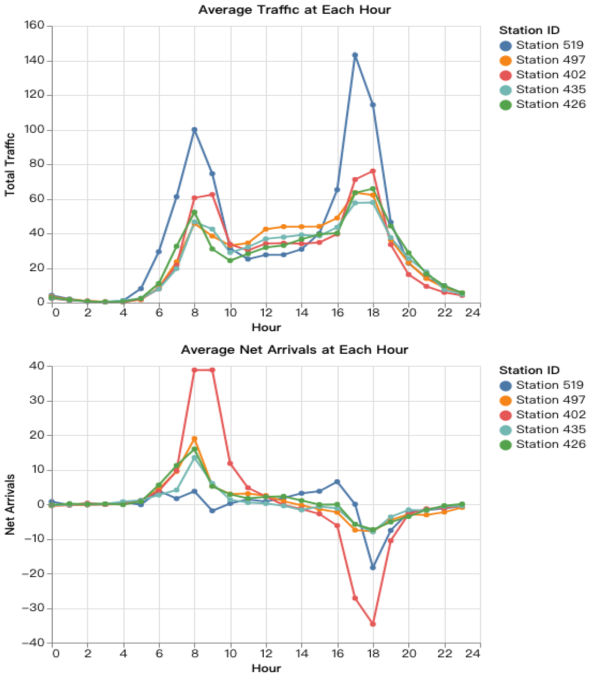


Fig. 2. Traffic flow at each hour

III. Results

We are going to divide this section into two parts, one on qualitative statistical graphs and the other on quantitative machine learning models. Looking at the Figure 2 of average net arrivals of each hour (the busiest 5 stations), it is very clear that the net arrival reaches the maximum value at around 8 am. In other words, traffic inflow of the busiest 5 stations peaks at 8 am in the morning. The traffic outflow, on the other hand, peaks at 6 pm over the day. This result aligns very well with our presumption, but is it true for every day over the week?

Figure 3 shows the comparison of the net and total traffic on business days versus non-business days. The previous pattern we found about the peak hours no longer holds in this case. Apparently, people (CitiBike users) on average start their day a few hours later on non-business days and we can clearly see that the sum of the arrivals and departures spiked at noon. To give a more detailed and spatial sense of the result, we plotted the morning rush hours and evening rush hours' net arrivals (business days) using Tableau.

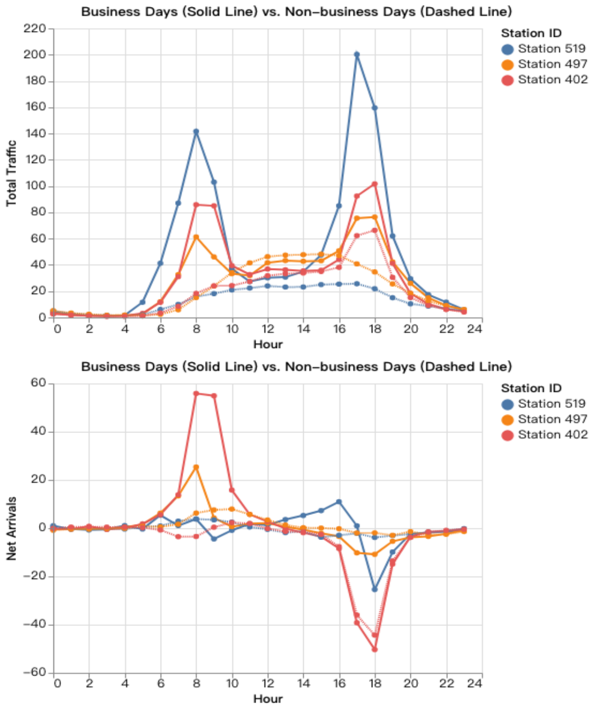


Fig. 3. Business days vs. Non-business days

As shown in Figure 4, weather also has very strong correlation with the CitiBike's usage. The biggest drop in the traffic flow this year happens to be the days when snow depth is the highest over the year. (Notice that there is a gap between the day that it snows and the day the snow starts to cover and lay thick on the ground.) Similarly, rain (measured by precipitation) reduces the overall CitiBike traffic, but not so dramatically as snow does.

The net arrivals maps illustrate an interesting pattern. According to the maps, stations can be categorized into two types: Type A is characterized by a net inflow during morning rush hours and a net outflow during evening rush hours and Type B, the other way around. Remarkably, stations of the same type tend to form clusters geographically. Note that Type A stations are primarily found along the middle strip south to the Central Park and Type B stations mostly fall on both flanks of Manhattan and the southeast. These two maps indicate the movement of population during a typical business day and, more importantly, provide general locations of business (Type A) and residential (Type B) zones in NYC.

For machine learning model, we tried to build a prediction model to foresee abnormal traffic flow in the next hour. Here abnormal traffic flow is defined as flow that is higher/lower than the mean by 1.96 standard deviations or more. Our assumption is that abnormal traffic flow is correlated with date (whether it is a business day or a holiday) and weather (precipitation, snow, etc.). Since this is a binary classification problem, we tested several models, including Logistic Regression, SVM (linear and RBF kernel), KNN, Decision Trees. Although the accuracy of those models is above 90%, all models tend to predict every future traffic flow as normal. This is not surprising since the way we define abnormal class will cause the dataset to be imbalanced as only about

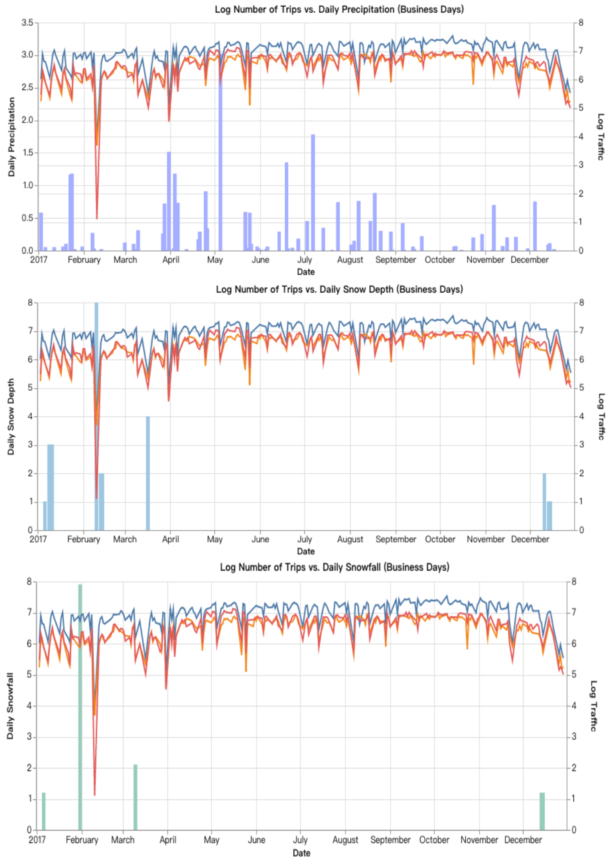


Fig. 4. Traffic flow under extreme weather conditions

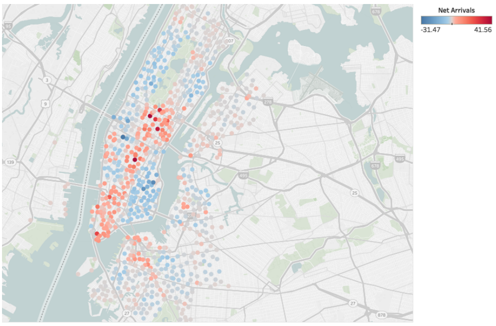


Fig. 5. Morning Rush Hours (7AM-10AM)

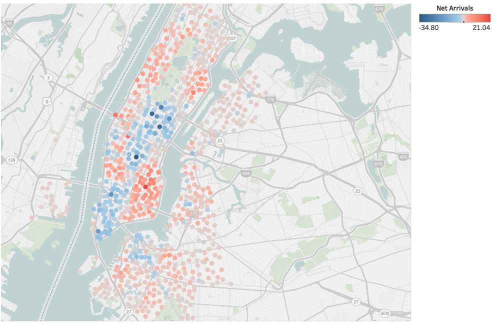


Fig. 6. Evening Rush Hours (5PM-8PM)

5%of the data points are labeled as abnormal. Among all these models, Decision Trees gave the best accuracy (AUC score=0.744) and the associated R-square is 0.2.

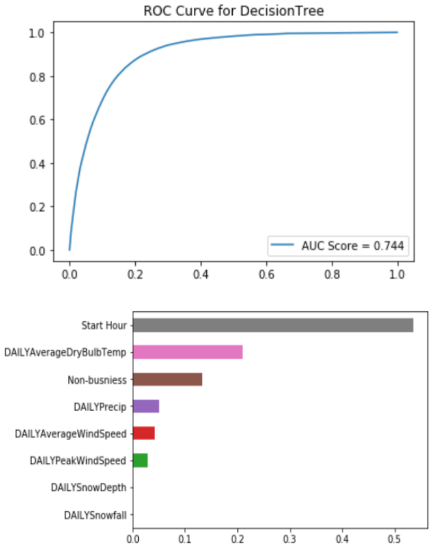


Fig. 7. Features Importance in Decision Trees

This is a common situation in machine learning model called skewed dataset and there are several popular techniques to deal with it. The three most widely used ones are upsampling, downsampling and cost-sensitive analysis (give more reward if predict a minority class correctly). Due to the emphasis on data visualization and time/page limit, we decide not to go deeper into the details.

IV. Discussion

This project aims to explore the general pattern of CitiBike traffic flow. We devoted a great amount of time to visualizing the data because we believe data visualization is the most straightforward method for accomplishing the goal in geospatial analysis. The results we got are very intuitive in terms of the weather and time's effect on the traffic flow.

Making more qualitative analysis rather than quantitative analysis may be one shortcoming of our project. However, our analysis definitely has some practical value if Citi would like to make the current overall distribution more efficient and optimize the allocation of bikes when building more stations in NYC, or in more cities in the United States.

Also, based on the reduction of rides under certain circumstances, technology companies like Uber and Lyft can predict a spike in demand when there is a decrease in CitiBikes' usage. For food trucks and other small business owners, it would be helpful for them to choose an optimal location based on the graphs we provide. For example, breakfast food trucks' owners can set Fidi or Washington Square Park as their base location, while food trucks that serve food mostly at night can move to more suburban area such as Brooklyn Williamsburg and Park Slope.

V. Conclusion

Our major finding is that the stations' traffic flows show a periodicity during business days. Stations near Midtown and Wall Street see a net arrival in rush hours in the morning while the situation reverses in the afternoon, which perfectly matches people's working patterns. However, this pattern is prone to change when New York City encounters extreme weathers like heavy snow. This gave us confidence that these features would have significant effect on the abnormality of station traffic.

Four machine learning models, Logistic Regression, SVM (linear and RBF kernel), KNN, Decision Trees, were applied in the project for traffic abnomality prediction. We picked Decision Trees as our optimal model.

For future researchers or companies interested in the fluctuation of CitiBike trips, more information that affects bike traffic may be needed or more statistical techniques need to be applied to the dataset to achieve desirable results.

References

[1] System Data. (n.d.). Retrieved December 17, 2018, from https://www.citibikenyc.com/system-data

[2] Data Access. (n.d.). Retrieved December 17, 2018, from https://www.ncdc.noaa.gov/data-access