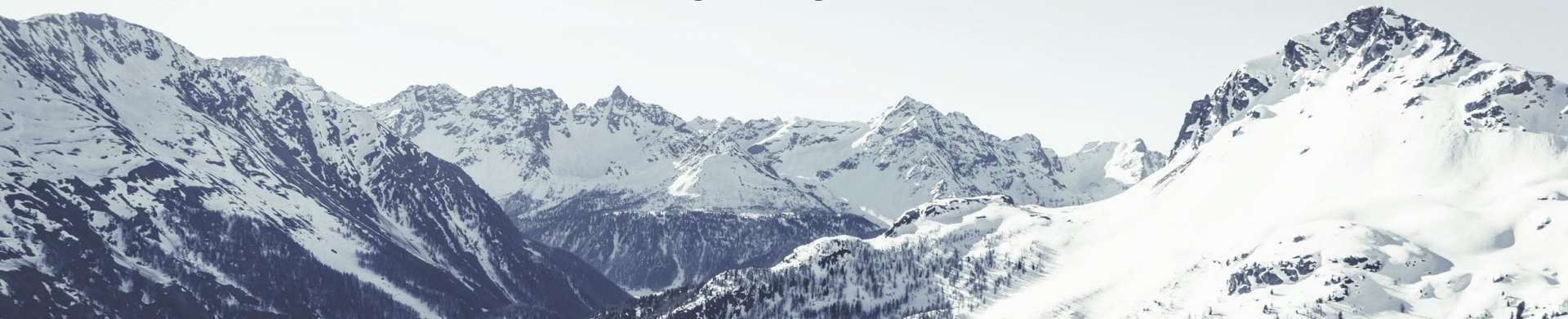




DS-GA 1007

Final Project

Group: Zexi Ye, Zian Chen, Jiayu Qiu, Xingyu Wang





Agenda

- **Problems of Interest**
- **Data Preparation**
- **Data Visualization**
 - **Patterns of Stations**
 - **The Effect of Weather**
- **Model Formulation**
- **Conclusion**

Citi Bike is the nation's largest bike share program, with 12,000 bikes and 750 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it's a fun and affordable way to get around town.



How it works?



1. **Join:** *Become an Annual Member or buy a short-term pass through the Citi Bike app.*
2. **Unlock:** *Find an available bike nearby, and get a ride code or use your member key to unlock it.*
3. **Ride:** *Take as many short rides as you want while your pass or membership is active.*
4. **Return:** *Return your bike to any station, and wait for the green light on the dock to make sure it's locked.*



Problems of Interest

- How does the number of Citi Bike riders change given different date/weather
- Focus on Data Visualization using Python
- Predict the occurrence of abnormal cases (Too many/Too few riders at a given time)

*Tools that we use: Python packages including pandas, Geoplot etc.





Data Preparation

The Dataset

The original dataset is obtained online, it contains two parts:

1. Citi Bike System Data
2. NOAA Historical Weather Data

(Part of the original dataset using `dataframe.head()`)

Citi Bike:

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gender
0	256	2017-12-01 00:00:00	2017-12-01 00:04:17	324	DeKalb Ave & Hudson Ave	40.689888	-73.981013	262	Washington Park	40.691782	-73.973730	18858	Subscriber	1981.0	1
1	325	2017-12-01 00:00:17	2017-12-01 00:05:43	470	W 20 St & 8 Ave	40.743453	-74.000040	490	8 Ave & W 33 St	40.751551	-73.993934	19306	Subscriber	1992.0	1
2	967	2017-12-01 00:00:19	2017-12-01 00:16:26	347	Greenwich St & W Houston St	40.728846	-74.008591	504	1 Ave & E 16 St	40.732219	-73.981656	28250	Subscriber	1992.0	1
3	125	2017-12-01 00:00:20	2017-12-01 00:02:26	3077	Stagg St & Union Ave	40.708771	-73.950953	3454	Leonard St & Mauger St	40.710369	-73.947060	25834	Subscriber	1988.0	1
4	451	2017-12-01 00:00:28	2017-12-01 00:08:00	368	Carmine St & 6 Ave	40.730386	-74.002150	326	E 11 St & 1 Ave	40.729538	-73.984267	14769	Subscriber	1986.0	1

Weather:

	STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	REPORTTYPE	HOURLYSKYCONDITIONS	HOURLYVISIBILITY
0	WBAN:94728	NY CITY CENTRAL PARK NY US	42.7	40.77898	-73.96925	2017-01-01 00:51	FM-15	OVC:08 55	10.00
1	WBAN:94728	NY CITY CENTRAL PARK NY US	42.7	40.77898	-73.96925	2017-01-01 01:51	FM-15	OVC:08 80	10.00
2	WBAN:94728	NY CITY CENTRAL PARK NY US	42.7	40.77898	-73.96925	2017-01-01 02:51	FM-15	BKN:07 55 OVC:08 100	10.00
3	WBAN:94728	NY CITY CENTRAL PARK NY US	42.7	40.77898	-73.96925	2017-01-01 03:51	FM-15	FEW:02 55	10.00
4	WBAN:94728	NY CITY CENTRAL PARK NY US	42.7	40.77898	-73.96925	2017-01-01 04:51	FM-15	BKN:07 50 BKN:07 60	10.00



The Final Dataset

In order to build a supervised machine learning model, we preprocessed the data

The Final Trip Data is:

	Station ID	Date	Start Hour	Departures	Arrivals	Non-business Day	DAILYMaximumDryBulbTemp	DAILYMinimumDryBulbTemp	DAILYAverageDryBulbTemp	DAILYPrecip
0	72	2017-01-01	0	0	1	True	48.0	40.0	44.0	0.0
1	72	2017-01-01	2	1	1	True	48.0	40.0	44.0	0.0
2	72	2017-01-01	3	1	0	True	48.0	40.0	44.0	0.0
3	72	2017-01-01	4	0	1	True	48.0	40.0	44.0	0.0
4	72	2017-01-01	5	2	0	True	48.0	40.0	44.0	0.0



Data Visualization

- First we would like to see what we can learn about all the stations just by using the data provided by Citi Bike
- Then we would move on to see how weather impacts the number of riders

Tools

Gain-ratio, Ranker

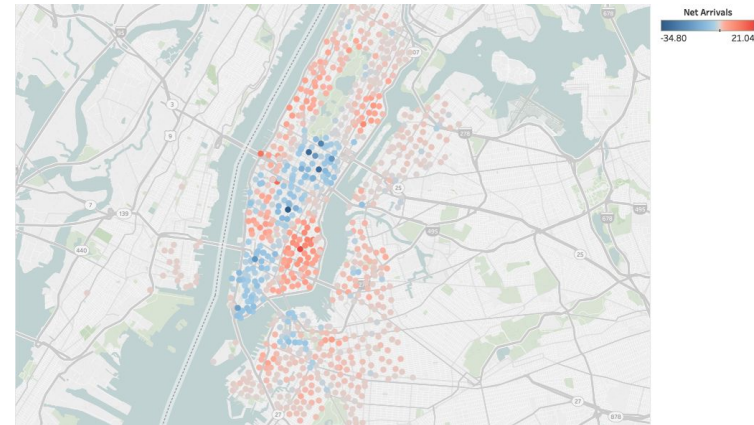
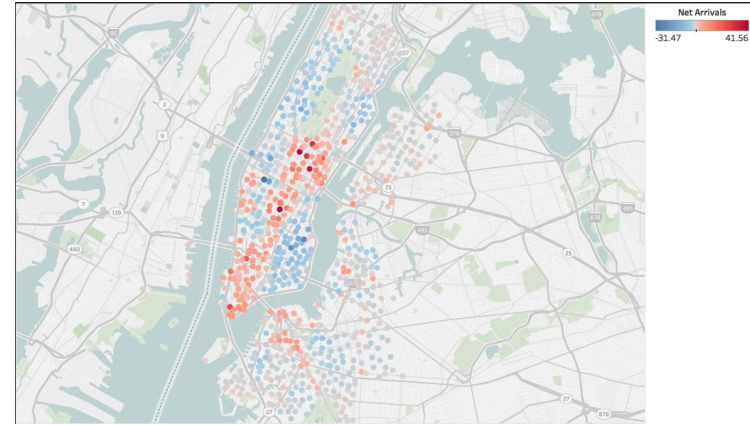
$$\text{GainR}(\text{Class}, \text{Attribute}) = (\text{H}(\text{Class}) - \text{H}(\text{Class} | \text{Attribute})) / \text{H}(\text{Attribute}).$$

- Based on the result, the bottom 4 attributes which are irrelevant are deleted

average merit	average rank	attribute
0.058 +- 0.002	1 +- 0	18 other_club
0.039 +- 0.001	2 +- 0	16 GADGETS_FLAG
0.036 +- 0.001	3 +- 0	12 Buy_6
0.011 +- 0	4.6 +- 0.66	13 Buy_7
0.011 +- 0	4.9 +- 0.7	15 CHOCOLATE_FLAG
0.01 +- 0	5.5 +- 0.81	2 GENDER
0.006 +- 0	7 +- 0	4 LAST_PURCHASE
0.005 +- 0.001	8 +- 0	5 Num_Purchases
0.003 +- 0	9.5 +- 0.67	11 Buy_5
0.003 +- 0	9.8 +- 0.75	3 SUM
0.003 +- 0	11.6 +- 1.43	6 FIRST_ORDER
0.003 +- 0	11.8 +- 0.87	14 WINE_FLAG
0.003 +- 0	12.5 +- 0.5	20 CONTACT_CUSTOMER_SUPPORT
0.002 +- 0	13.8 +- 0.4	10 Buy_4
0.001 +- 0	15 +- 0	7 Buy_1
0.001 +- 0	16.2 +- 0.6	9 Buy_3
0 +- 0	17.1 +- 0.7	8 Buy_2
0 +- 0	17.9 +- 0.54	17 Previous_sample
0 +- 0	18.8 +- 0.4	19 affiliated_store
0 +- 0	20 +- 0	1 id

Rush hours data (in Net Arrivals)

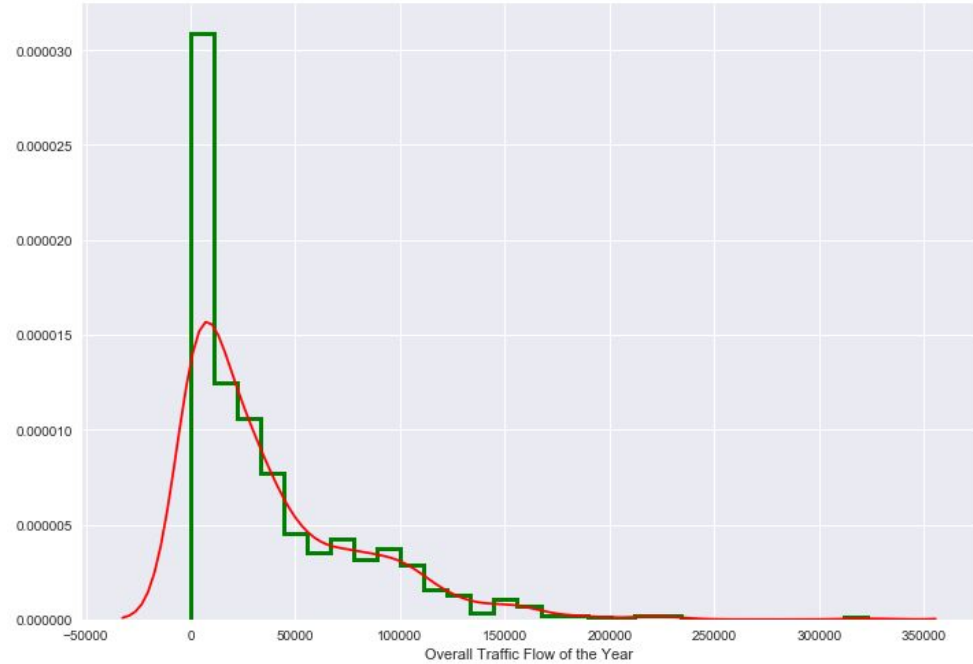
- These two figures show the number of Net Arrivals during rush hours in the morning and in the evening respectively





Distribution of the traffic flow

→ We choose data from Year 2017



Focus on large stations

→ Since there are too many stations, we only choose the largest stations to plot

Top 5 Stations



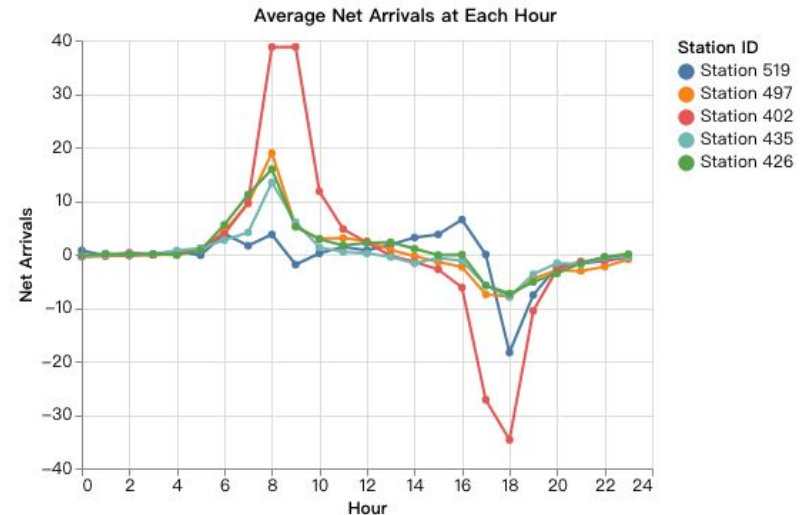
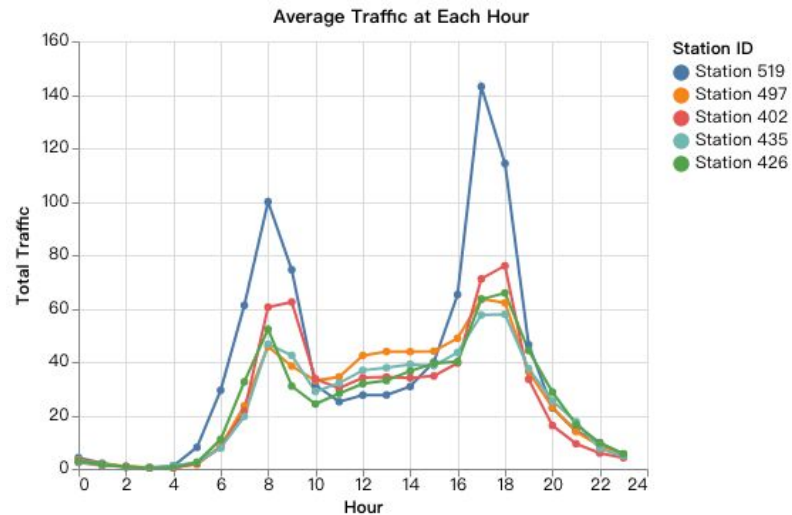
Map based on Start Station Longitude and Start Station Latitude. Size shows details about Number. The marks are labeled by Start Station Name.

The five largest stations and their corresponding station numbers are:

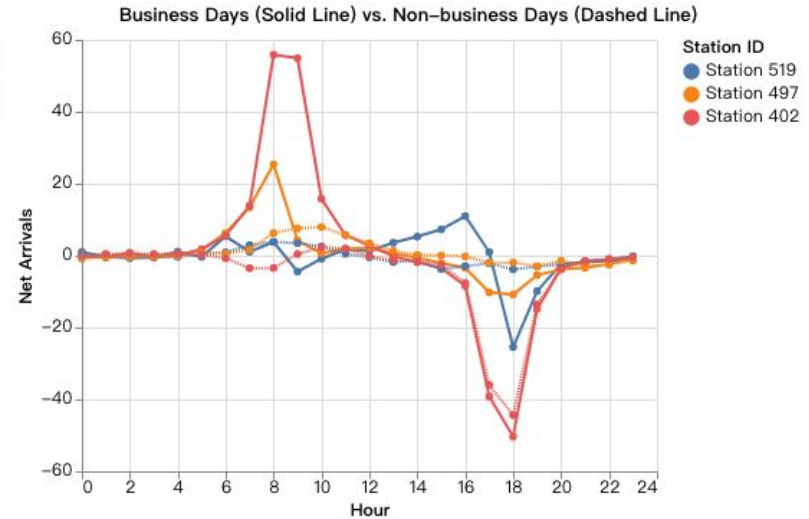
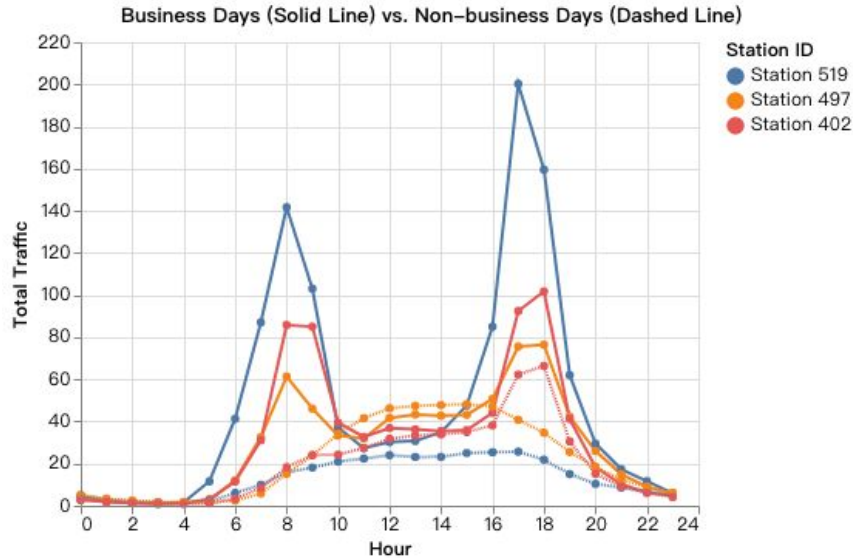
1. Pershing Square North ↔ 519
2. E 17 St & Broadway ↔ 497
3. Broadway & E 22 St ↔ 402
4. W 21 St & 6 Ave ↔ 435
5. West St & Chambers St ↔ 426



Average number at Each Hour

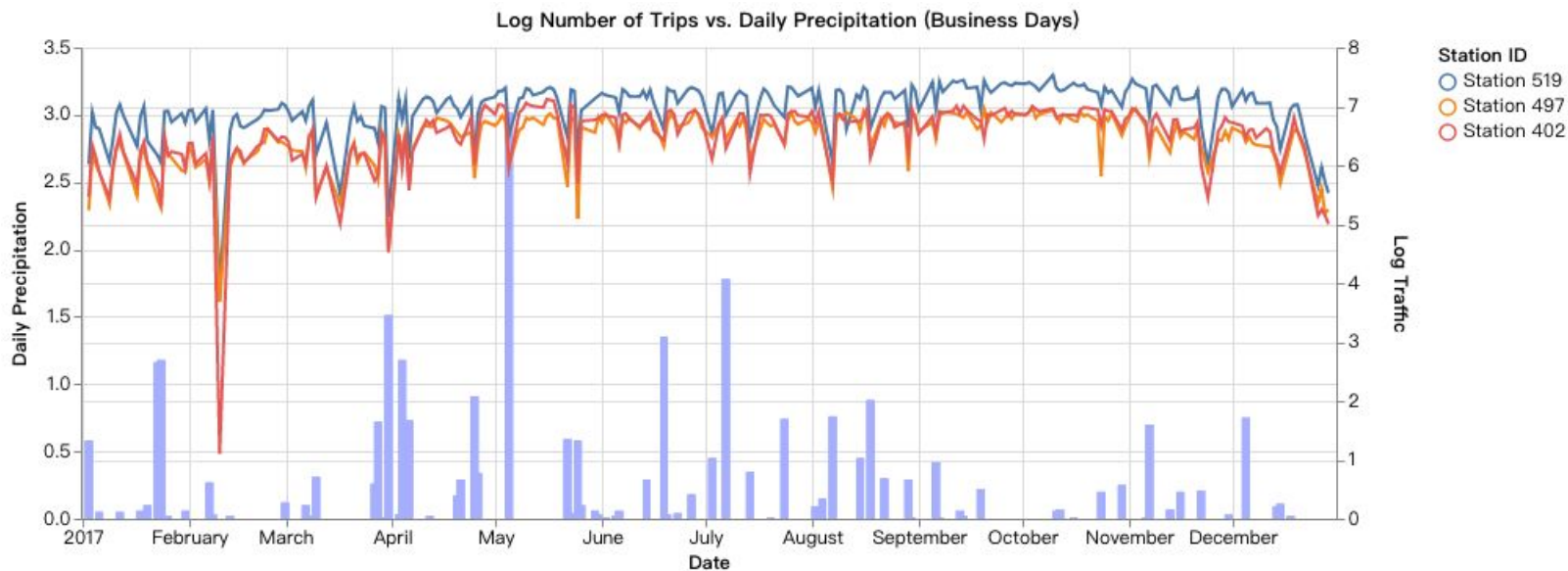


Business Days vs. Non-Business Days



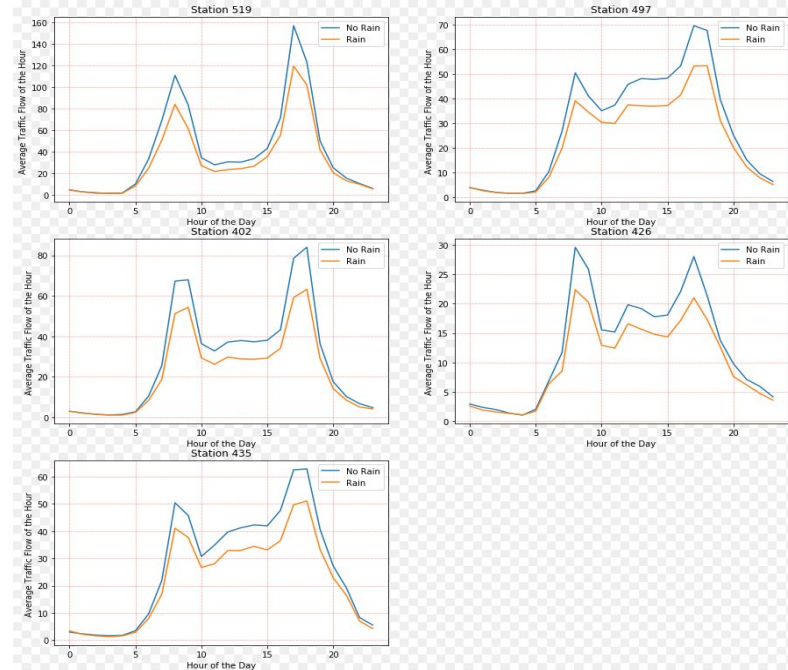
The Effect of precipitation

→ In general, rain stops people from riding Citi Bike



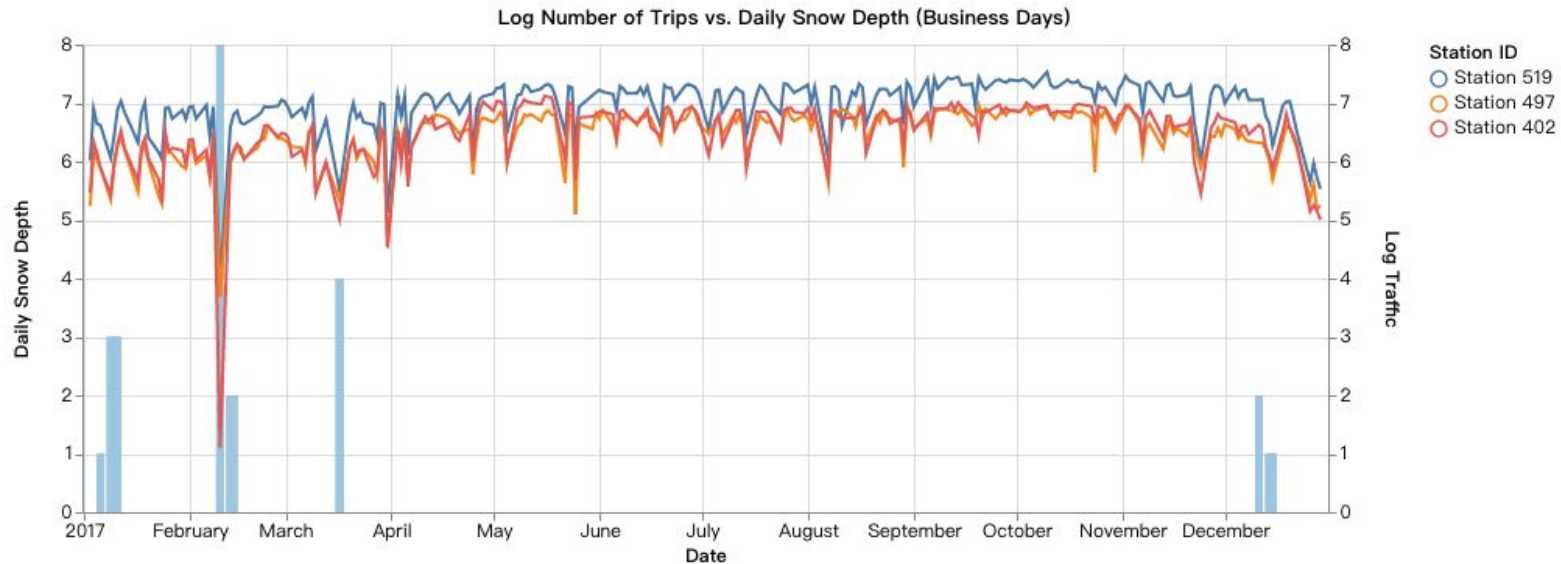
The Effect of precipitation

→ It seems that the percentage of decrease in traffic flow is quite similar among the five stations



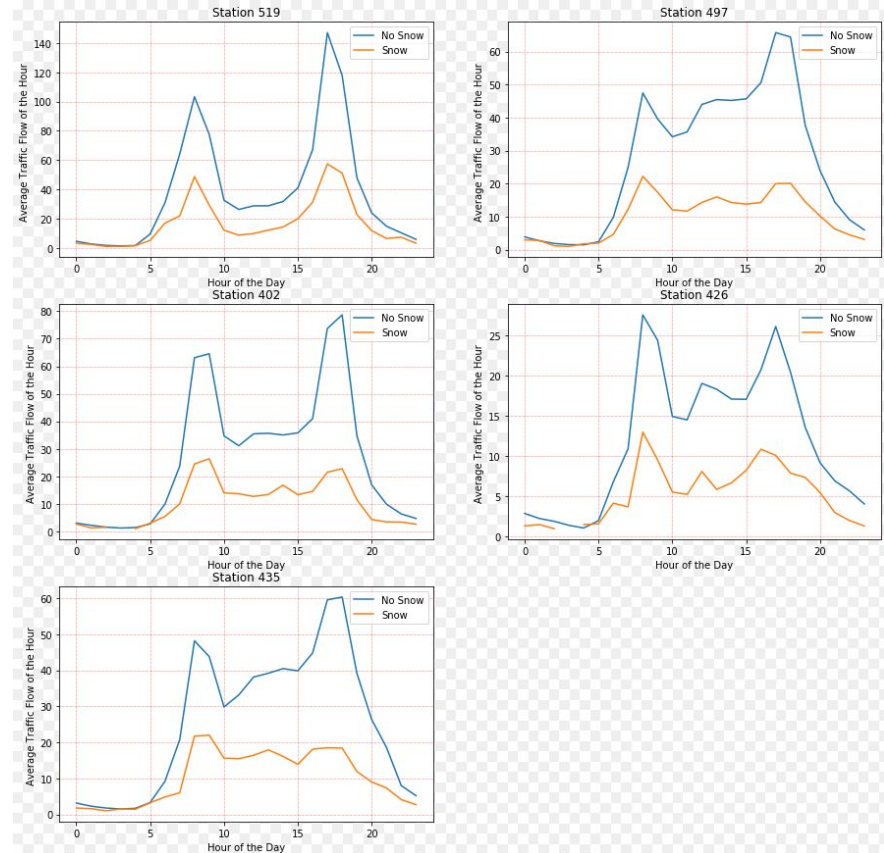
The Effect of snow

→ Snow is different from rain as it might have a long-term effect on several days



The Effect of snow

→ Snow fall shows similar pattern but the influence is much stronger





Machine Learning Part: Model Formulation



Methodology and Models



- Binary Classification
- Target Variable: Whether there will be abnormal traffic flow in the next hour
- Definition of abnormal: traffic flow is far away from its mean by more than 1.96 standard deviation
- Features:
 - Precipitation at hour t
 - Snowfall at hour t
 - Snow depth at hour t
 - Time t
 - Business day or not



Models and Results

→ We tried Logistic Regression and SVM

Logistic regression

	Predicted (a)	Predicted (b)	
	3606 90.15%	38 0.95%	Actual (a): 0
	304 7.6%	52 1.3%	Actual (b): 1
Classification Accuracy: 91.45%			

SVM with RBF Kernel

Test error for this model is: 0.083
 Validation error for this model is: 0.092
 $\begin{bmatrix} 3632 & 0 \\ 368 & 0 \end{bmatrix}$

→ In terms of accuracy, the results seem to be really good, but is it really the case?

Model Analysis



- Logistic Regression is correctly predicting some abnormal situations thus it's better than SVM in this case
- However, both models are generally predicting almost everything as 0
- The training set has 94% of data points label as negative(0), only 6% of them are labeled as positive (1)
- We have encountered the problem of skewed dataset in this case !

Possible solutions



- There are several popular tricks to deal with skewed dataset:
 - ◆ Undersampling
 - ◆ Oversampling
 - ◆ Cost sensitive analysis
- However, we will not discuss in detail due to limited time and purpose of this project

Conclusion



Why do we do this project?

- Great application of what we learned in this course e.g dealing with dataframe, SQL-like operation, data visualization using Python
- Although we just covered the largest stations on a hour basis, it's not hard to generalize to more stations on different time basis using the same technique
- The ability to predict abnormal traffic flow has great potential business value: companies like Uber can use it to improve their response time, shopping malls and restaurants can make adjustment accordingly based on future customer flow (related to traffic flow)





Q&A?

