MAST30034 Project 1

New York City Taxi Data Analysis

Haonan Zhong
Student ID: 867492

September 14, 2021

1 Introduction

Being one of the most populous cities in the United States, New York City has millions of taxi trips taken every month. This project aims to make a quantitative analysis and form a better understanding to the New York City Taxi and Limousine Commission (TLC) trip record data. In addition, make some recommendations that might improve taxi driver's income.

2 Data Selection

2.1 NYC TLC Dataset

The taxi dataset used in this project is yellow taxi trip data covering the year 2018, which records attributes such as pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Yellow taxis are taxis that allowed to respond to street hailed from a passenger in all five boroughs.

The selected timeline for this analysis will cover January and February, July and August of the year 2018 in order to compare taxis' activity in winter and summer.

2.2 Taxi Zone Dataset

The pick-up and drop-off locations are populated by numbers ranging from 1 to 263. These numbers corresponds to taxi zone.

- Taxi Zone Shapefile: contains geometric information of each taxi zones.
- Taxi Zone Look Up Table: a table that contains a list of TLC taxi zone location IDs, location names, and corresponding boroughs of each zone.

2.3 NYC Weather Dataset

The weather data are downloaded from the National Centers for Environmental Information (NCEI), which contains daily weather observations at Central Park for 2018, such as the maximum/minimum temperature, precipitation, snowfall, and wind speed. Hopefully, it will enable us to study how different weather conditions affect the usage of taxis in NYC.

Here, we assume that the entire New York City shares exactly the same daily weather condition as Central Park, and the information provided by the dataset are correct.

3 Preprocessing

The winter and summer data each containing approximately 17 million and 16 million trip records, respectively, and both have 17 columns of attributes. Luckily, after getting a glimpse of the data, none of the trip records contains any missing value.

3.1 Feature Engineering

Feature engineering is the process of transforming raw data into useful features to get the most out of your data. Down below is a list to briefly summarize what I did.

- Calculate the duration of each trip using the difference between pickup and drop-off time.
- Identify the month, day of the week, and hour for each trip. Then classify whether the trip is in workday, weekend, or, holiday.
- Identify the pickup and dropoff borough of each trip.
- Calculate the tip percentage using total amount and tip amount.

Furthermore, attributes like MTA Tax, Extra and Improvement Surcharge that are not the features we mainly focus on will be dropped to reduce the size and complexity of the data.

3.2 Data Cleansing

Records with implausible values or errors are removed based on examining the distribution of each attribute and common sense to ensure the dataset's consistency and correctness. Some examples are listed below.

- Pickup/drop-off date and time are strictly within the selected month period.
- Pickup/drop-off Location ID should be within the range of [1, 263].
- Passenger count should at least 1 and less than 7 as the maximum number of passengers allowed by law is 6.
- Trip distance should greater than 0 miles but less than 100 miles.
- Fare amount should be at least \$2.5 but at most \$250.
- Tip percentage are less than 50%
- Trip duration should be more than a minute and less than three hours.

Finally, weather data are merged together with the taxi dataset according to the date of each trip.

4 Analysis and Visualisation

4.1 Geospatial and Time Analysis

In figure 1, these maps show total numbers of taxi pickup and drop-off, respectively, in New York City from the selected timeline, where darker regions indicate more taxi activity. Notice how pickups are primarily concentrated in Manhattan, LaGuardia, and JFK, while drop-offs extend further into the outer borough. It is worth noting that Staten Island's pickup and drop-off frequency are significantly lower than other boroughs, given it is the least populated borough.

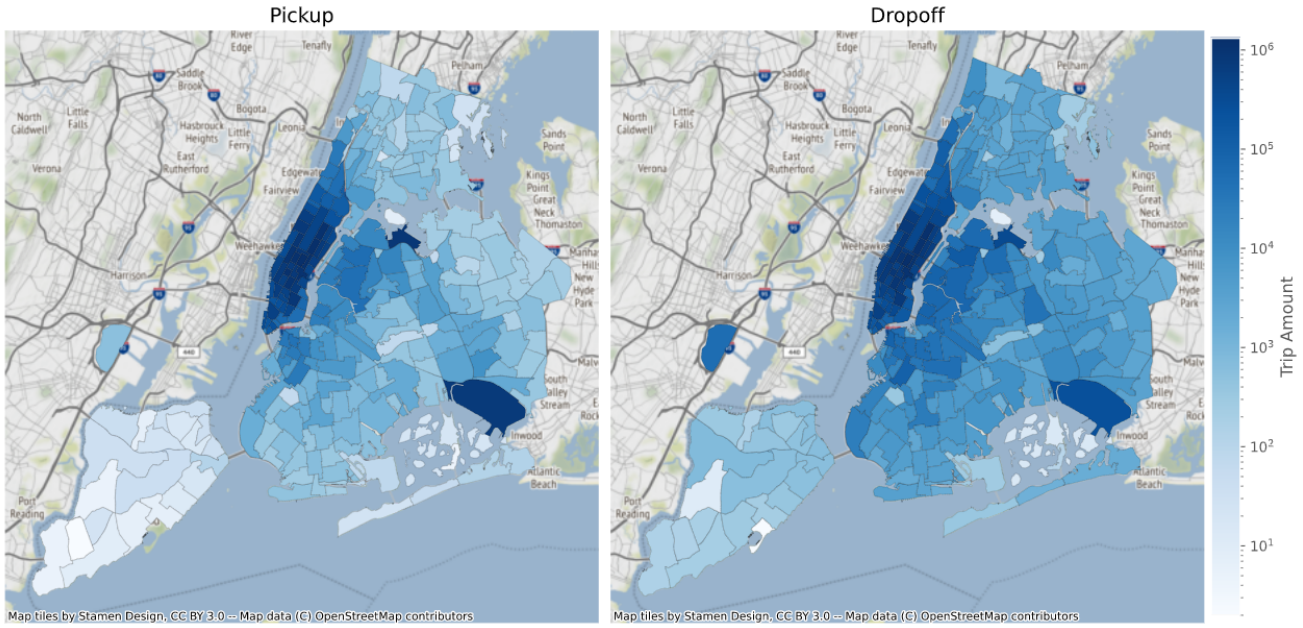

Figure 1: Taxi Pickup Frequency at Different Times of the Day

As shown in figure 2, evening rush hour during weekdays appears to be the busiest. Midnight to 6 am have the lowest pickup frequency on average during the weekdays. Moreover, we observed a clear distinction between weekdays and weekends at night, as the period with the lowest pickup volume was almost pushed back by two hours on the weekend; the reason could be people often enjoy more nightlife over the weekend. Another point worth mentioning is that winter has more pickup than summer during the morning and evening rush hour, which might be due to the cold weather. Thus people are more willing to take taxis to the office and home.

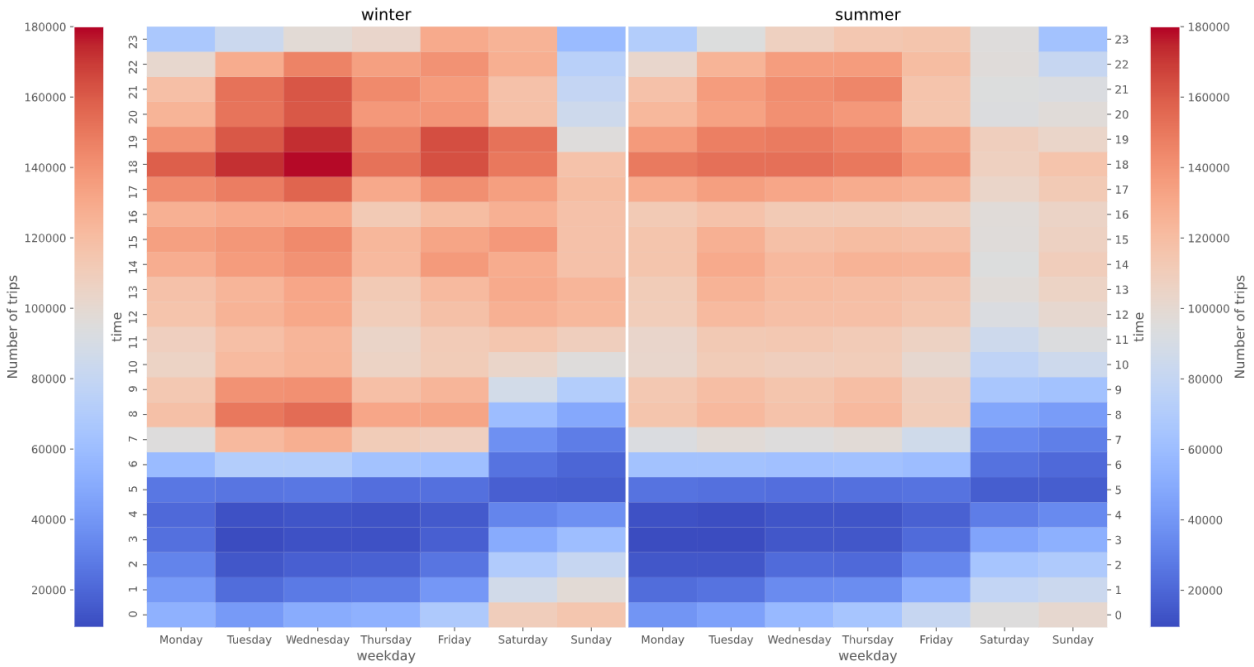

Figure 2: Taxi Pickup Frequency at Different Times of the Day

4.2 Weather Impact

Next, the impact of weather is considered, the following scatter plots shows the number of trips taken versus different weather condition. Summer trips will not be considered for the bottom two plots, as there is no snow during summer.

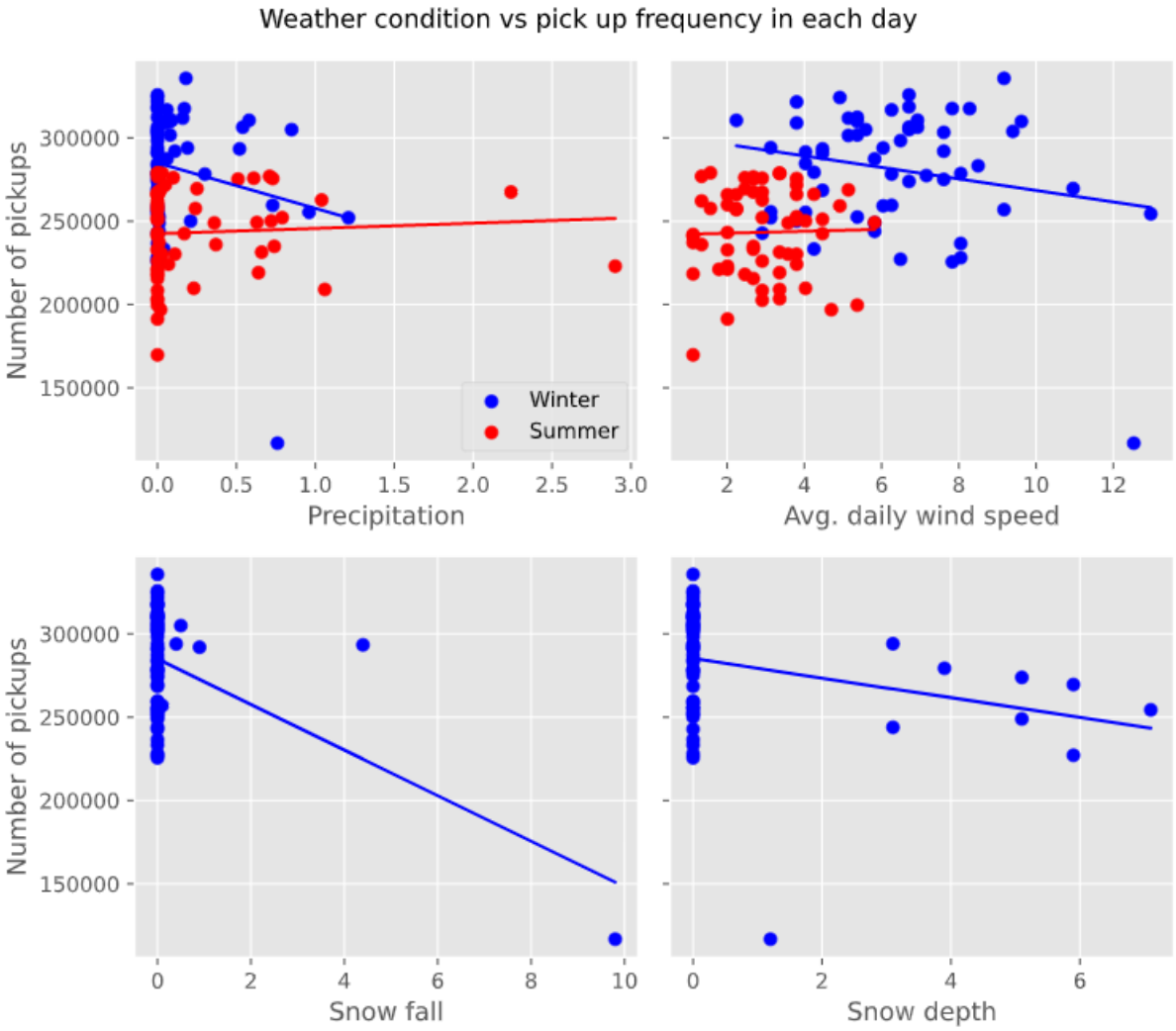

Figure 3: Taxi Pickup Frequency in Different Weather Conditions

From figure 3, we observed a slight increase in daily trip amount as precipitation increases during the summer, while winter is the opposite. Wind speed alone does not seem to affect the daily pickup frequency a lot. On the other hand, snow depth/snowfall appears to have a negative impact on daily ridership. As the plot shows, due to the January 4, 2018 blizzard in NYC, Central Park reported 9.8 inches of snow, caused the daily ridership dropped significantly to around 120,000.

However, recall that we assumed all taxi zones in NYC share the exact weather condition as Central Park; Hence, it could impact the analysis.

4.3 Tipping Behaviour

One limitation of this analysis is that only trips with credit card payment will be considered, since tip amount on cash payment are not recorded.

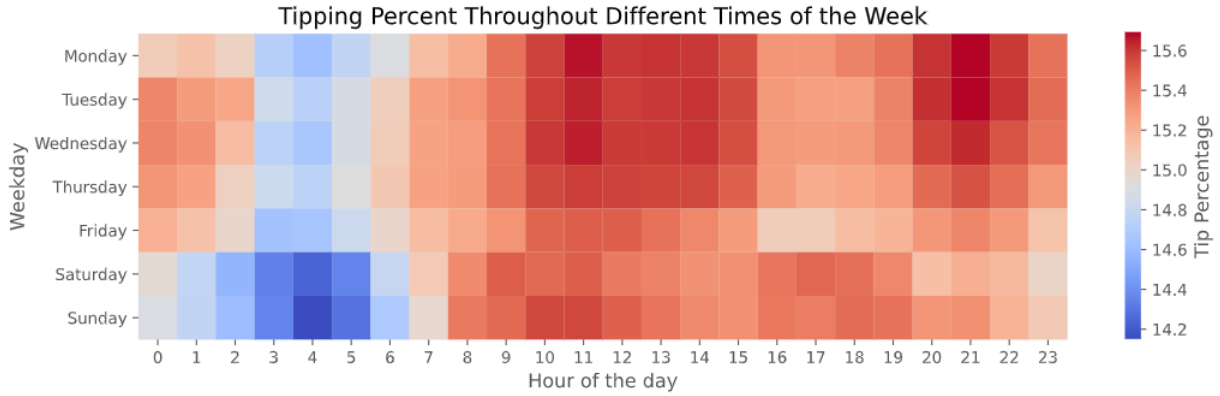

Figure 4: Tip Percentage at Different Times of the Day

As depicted in Figure 4, we observed a clear difference between weekdays and weekends, suggesting that passengers tend to tip differently depending on the trip. Another point worth mentioning is that passengers tend to tip a smaller percentage towards the total fare in the early morning between 3 to 6 am. However, the average tip percentage increases gradually as the day progress, with a slight decrease between 16:00 and 19:00 on weekdays, and reached the highest with another obvious surge between 20:00 to 22:00.

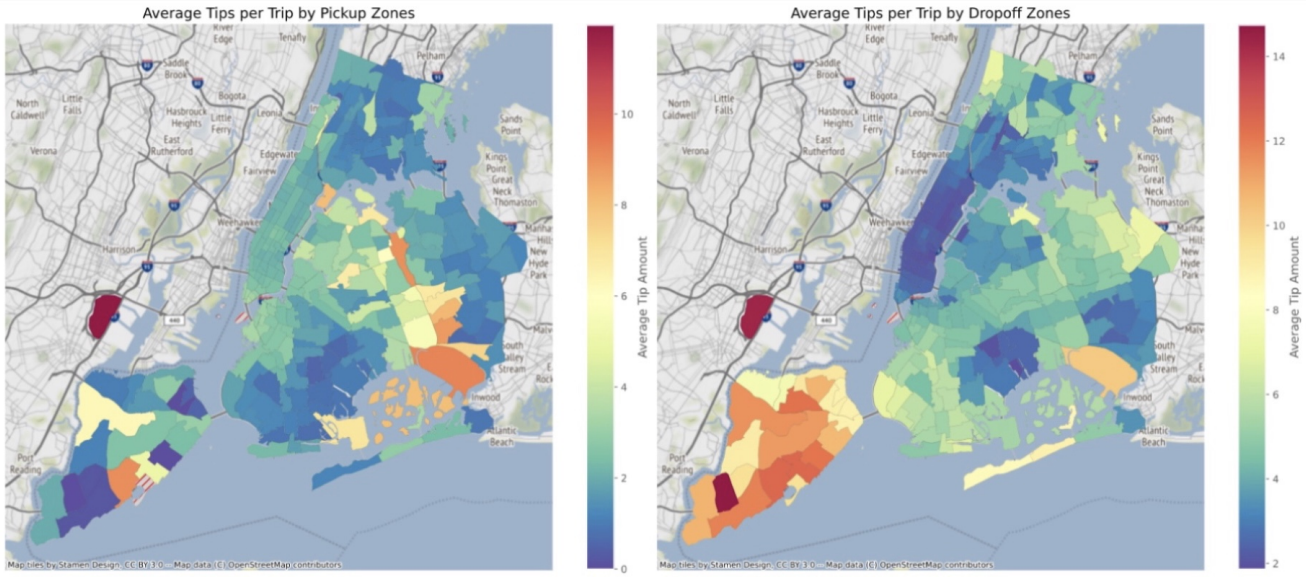

Figure 5: Average Tips per Trip by Pick Up Zones

As Figure 5 highlighted, on average, passengers picked up from Newark Airport, JFK, and Flushing Meadows Corona Park are more likely to give a higher amount of tips. Contrarily, passengers dropped off at Great Kills and Oakwood on Staten Island tend to give a higher amount of tips, which well reflects Staten Island's status as one of the most well-off boroughs. However, since some zones only have a few trips, the statistics might be affected by the insufficient amount of samples.

4.4 Correlation Between Attributes

Then, we will investigate the correlation between continuous variables. As shown in figure 6, not surprisingly, the correlation coefficient between attributes like "fare amount," "tip amount," "total amount," "trip distance," and "duration" indicates there is a positive relationship between these attributes. Contrarily, the correlation between other variables is not that significant, except for a strong negative correlation between wind speed and daily average temperature.

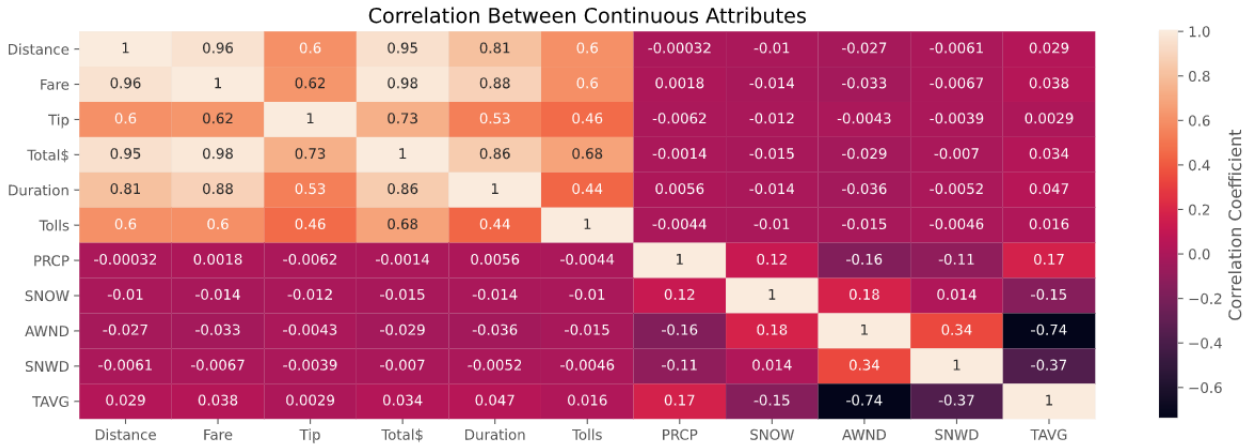

Figure 6: Correlation between Continuous Attributes

5 Statistical Modelling

5.1 Ordinary Least Squares Regression and Gradient Boost Regression

This section will build two regression model to predict the tip amount for eligible taxi trips. Model fit is evaluated using the train-test split method to randomly split sampled data into 80% training set and 20% testing set.

Initial feature selection is done by inspecting the correlation plot to drop redundant predictors to avoid overfitting. Features like weather data that are uncorrelated with the response will be ignored as they might masked the effect of other predictors. Additional feature selection will be using t-test.

5.2 Evaluation and Results

The fitness of the models are evaluated in terms of **Root Mean Square Error** and **Coefficient of determination $\mathbf{R^2}$**. Result of our models are compared against the baseline model that only contains the intercept, which simply used the mean as the predicted value.

Goodness-of-fit statistics				
Model	Train Set RMSE	Test Set RMSE	Train Set $\mathbf{R^2}$	Test Set $\mathbf{R^2}$
Baseline	2.342990	2.338763	0	0
Linear Regression	1.294810	1.287552	0.694599	0.696919
GBR	1.300274	1.309557	0.693590	0.690436

Table 1: Model Performance Statistics

Compared to the baseline, the final fitted linear model explained an additional 69% variance in tip amount; and scored approximately 1.3 for RMSE; Given that the average tip amount is around $2.59, most predictions are about $1.3 away from the actual tip amount. A more advanced gradient boost regression was also applied to predict the tip amount. Gradient boosting minimises the loss by iteratively modeling the residuals. However, it did not achieve good performance either.

5.3 Discussion

Overall, the model performed poorly as the prediction error is quite large, so the model is not very suitable for predictions. After inspecting the diagnostic plot, it suggests that the residuals are not normally distributed and show significant signs of heteroscedasticity indicates the violation of the linear model assumption. The performance of the model is affected due to the heavily right-skewed distribution of tip amount, as a large portion of tip amount are clustered at a very low value between $0 and $7.

6 Recommendations

From the visualisation section, weekdays are usually the busiest, especially at the evening rush hours; late-night pickup frequency is generally higher over the weekend, possibly due to people's late-night activity. As for tipping, passengers tend to tip less generously in the early morning and evening rush hours, and tipping percentages are higher during the day between 10 am to 3 pm and night between 8 pm to 10 pm. On average, passengers picked up from Newark and JFK often give a higher tip amount. Hence, taxi drivers could use the above information to improve their earning.

7 Conclusion

In conclusion, through some exploratory analysis of yellow taxi trip records, the relationship between some attributes has become a lot more clearer. The assistance of visualisation plots has been beneficial; it allowed us to gain a deeper insight into taxi usage on a weekly and hourly basis. We also manage to build a few models to predict the tip amount, though it did not perform well. Furthermore, for future references, with a more detailed weather data, perhaps we could measure how different weather conditions affect taxi activity more precisely.

References

[1] Taxi Fare - TLC. (2021). *Taxi Fare*. NYC TLC. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

[2] Wikipedia contributors. (2020). *Demographics of Staten Island*. Wikipedia. https://en.wikipedia.org/wiki/Demographics_of_Staten_Island

[3] Wikipedia contributors. (2021). *Staten Island*. Wikipedia. https://en.wikipedia.org/wiki/Staten_Island

[4] National Weather Service. (2018). *January 4, 2018 Blizzard*. https://www.weather.gov/okx/Blizzard_Jan42018

[5] Wikipedia contributors. (2021). *January 2018 North American blizzard*. Wikipedia. https://en.wikipedia.org/wiki/January_2018_North_American_blizzard

[6] Wikipedia contributors. (2021). *Gradient boosting*. Wikipedia. https://en.wikipedia.org/wiki/Gradient_boosting