The Effect of Weather, Population, and Property Price on the Demand for Yellow Taxis in New York City

Haiyang(Henry) Huang
Student ID: 1071147
Github repo with commit

October 15, 2022

1 Introduction

New York City, the most populous mega-city in the world [1], is famous for its fast-paced lifestyle and busy traffic. Under such a background, the demand for efficient transportation is enormous. The New York City Taxi and Limousine Commission (TLC) has constantly updated trip records for multiple types of taxi and for-hire vehicles since 2009. This report will be based on the yellow taxi data in 2019[2], which contains a total number of 84,598,444 rows and 19 attributes.

The research goal of the report is to investigate the **effect of weather**, **regional population**, and **regional price of the property** on the demand for yellow taxis in New York. This report will assume that all the datasets are correct and validated.

2 Dataset

2.1 Taxi Type

As one of the most iconic and traditional taxis in New York, **the medallion(yellow) taxi is the subject of this study** because it is flexible for the driver to pick up a passenger in any borough.

2.2 Date Range

New York has been affected by the COVID-19 pandemic since March 1st, 2020 [3]. As a result, the volume of trips declined drastically. To avoid the unexpected influence of the pandemic, **data from Jan 2019 to Dec 2019 are chosen to be the training set**, and **data from Jan 2020 to Feb 2020 are chosen to be the testing set** [2].

2.3 External Data

There are 3 external datasets employed in this research. The first one is the Population By Neighborhood Tabulation Areas (NTA) from the NYC OpenData [4], which contains useful information on populations in each NTA. This dataset also comes with a shapefile of NTA, which can help us to estimate the population in each taxi zone. The second one is the Annualized Sales Update data from the Department of Finance, NYC [5]. It provides information on property sale records in NYC which allows us to determine the average price of property per square foot for each zone. The last dataset is the daily weather in New York provided by Visual Crossing Corporation [6].

2.4 Hypothesis

Our hypothesis is that population will have a positive effect on the number of trips because more customers mean a larger market. Also, the property price is reflective of one's social-economic background, which can tell an interesting story about the tendencies of taking taxis of people from different financial backgrounds. Finally, we believe that weather will have a certain effect on the number of trips. For example, unexpected precipitation may increase the demand for taxis.

3 Target Audience

The target audiences of this report are Medallion taxi drivers, taxi companies, and New York City Taxi and Limousine Commission.

4 Preprocessing

4.0.1 Taxi data

1. Drop the attributes that we are not interested in.
2. Drop the rows with null passenger count and RatecodeID.
3. Remove data that are not in 2019/2020 depending on which dataset we are dealing with.
4. Remove outliers that are 3 standard deviations away from the mean (z-score method).
5. RateCodeID should be one of the integers in the range of 1-6.
6. Remove trips that are not paid by credit card or cash.
7. Remove rows with zero Passenger count.
8. Remove trips that start or end at zone 264 and 265 as they are unknown zone.
9. Extract the pick-up and drop-off month in case of cross-month trips.
10. Aggregate the number of trips by (PULocationID, PUMonth) and (DOLocationID, DOMonth)

4.1 Property price data

1. Remove outliers that are 1.5 IQR away from Q1 or Q3
2. Select rows that are dated before February 29, 2020 and after January 1, 2019
3. Remove rows with sale price less than $9999
4. Remove rows with gross square feet less than 9 sq. ft
5. Retain rows with building class starting with A or B or C or D or H where these are the building class type for accommodations (e.g. A refer to ONE FAMILY DWELLINGS)
6. Retain rows where the building class is unchanged throughout the life of the building because building class might have an effect on price.
7. Aggregate the mean price of properties in each neighborhood.
8. Link the property price data with taxi zones data based on the similarity between neighborhood name and service zone name.

4.2 Population data

- Assume the population in each neighborhood is evenly distributed, use the shapefiles to compute the overlapping area of taxi zones and neighborhoods, then sum all the segments of the population for each taxi zone.

$$Population_i = \sum_j Population_j * P(TaxiZone_i \cap Neighborhood_j)$$

4.3 Weather data

- Select Attributes that are thought to be relevant such as "feels like" temperature, precipitation, and snow. The rationale behind choosing "feels like" temperature rather than the actual temperature is because "feels like" conveys a more direct message of how people feel about the temperature, which may impact their decision of whether or not to hail a cab.
- Group the data by month and take the mean of selected attributes.

5 Preliminary Data Analysis

5.1 Monthly trip count

Apparently, from Figure 1, there is a trend of downturn between June and August, indicating that the number of trips can be negatively correlated with temperatures. But **there is no causal relationship between temperature and number of trips in a month**. In addition to weather data, we need more data to explain the phenomenon, for example, important events in New York City during that time.

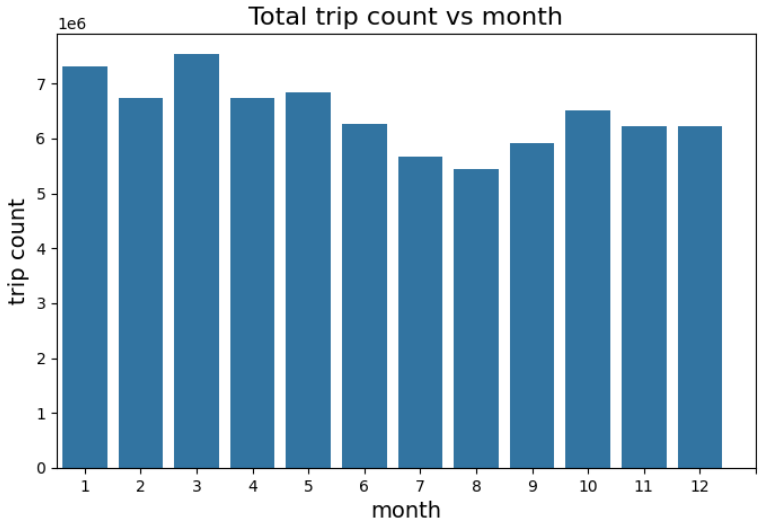

Figure 1: Number of Trips vs. Month

5.2 Distribution of total trip count

As we can see from Figure 2, the total number of trip counts per location is extremely right-skewed, which implies a log transformation is suitable.

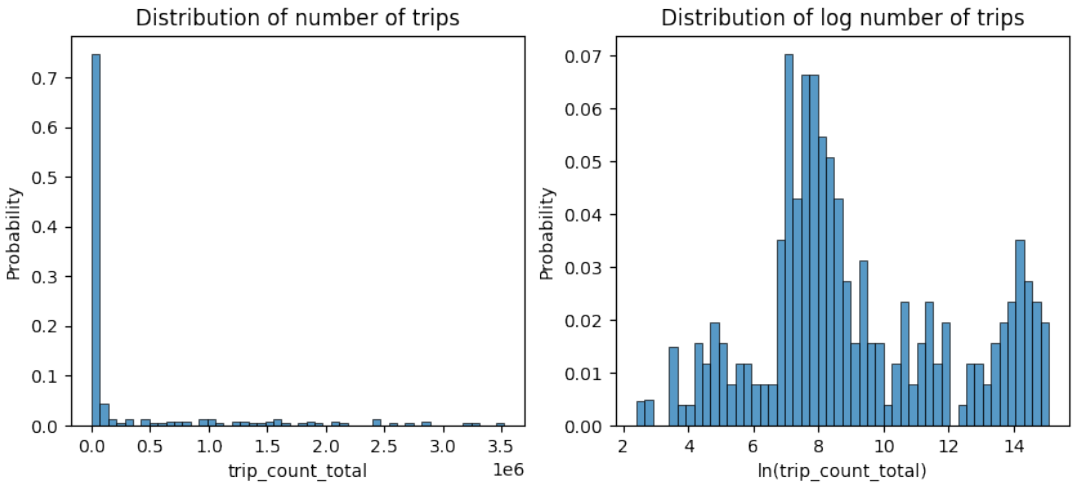

Figure 2: Distribution of Number of Trips vs. Log-transformed

5.3 Correlation between features

As shown in Figure 3, we can see that the correlation between price per square foot and price per unit is relatively high compared to other attributes. Intuitively it makes sense because the gross square feet of a unit may not vary significantly. In contrast, the correlation between population and density is not as high as the former because there are taxi zones in Manhattan that have a small geographical area but are densely populated.

Some highly correlated attributes are depicted in Figure 4, such as average feelslike, feelslike min, and feelslike max. We will retain densely the average feelslike since it is more representative of the temperature in a month. In addition, precipitation, cloud cover, precipitation cover, and visibility are highly correlated to each other. As stated in [7], the precipitation cover is the amount of time and the cloud cover is the percentage of the sky covered by cloud, precipitation is retained because intuitively the heaviness of rain has a greater impact on people's decisions of whether to take a cab or not. Moreover, visibility may contain some extra information about fog, smog, or some such which is not covered by other attributes.

5.3.1 Distribution of Attributes

The demographic attributes are slightly right-skewed, which may imply that a log transformation can be applied 5.

5.3.2 Final Attributes

The final set of attributes for weather model are "feelslike", "precip", "windspeed", "visibility", "snow" and "snowdepth". For demographic model, the final set of attributes are "Price_per_square_feet", "Population_By_LocationID", "Density_per_hectare", "ln_Price_per_square_feet", "ln_Population" and "ln_Density_per_hectare".

6 Geospatial Visualisation

As we can see in the map 6, the number of pickups in a location is very similar to the number of dropoffs, which is not surprising. The most popular pick-up and drop-off locations are around

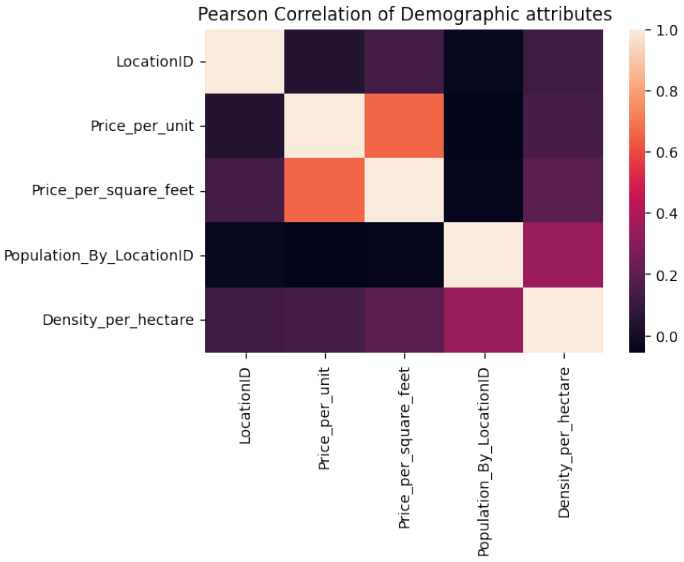

Figure 3: Correlations between Demographic Attributes

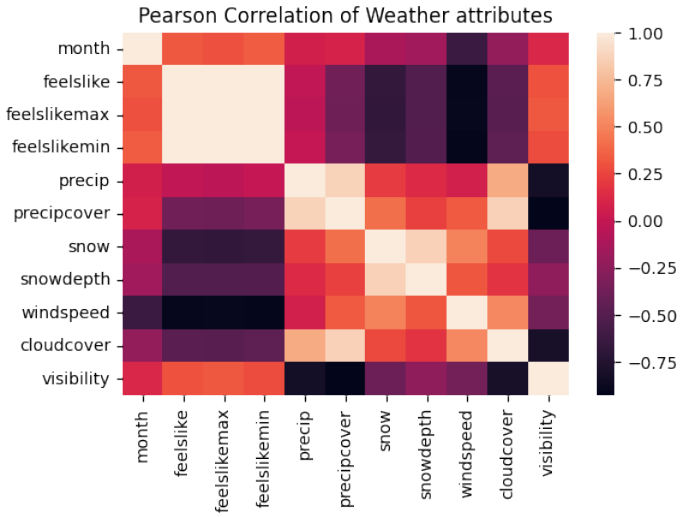

Figure 4: Correlations between Weather Attributes

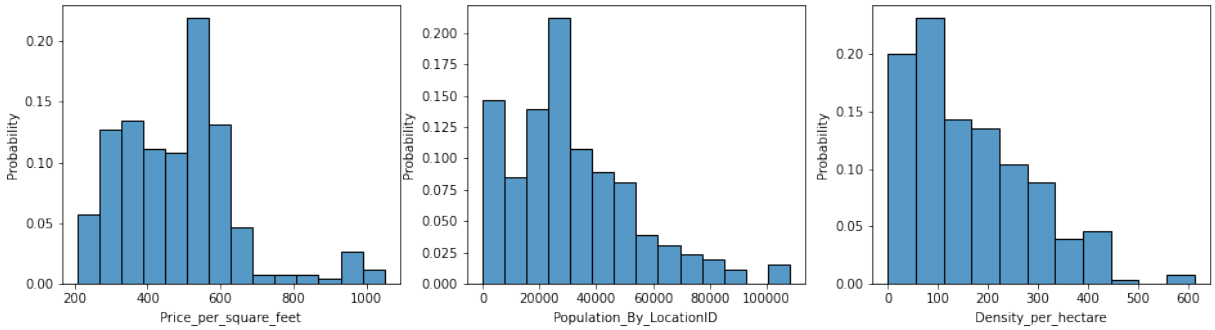

Figure 5: Distribution of Demographic Attributes

Manhattan, which suggests a positive relationship between the number of trips and the density of pick-up or drop-off locations. In addition, the number of pickups in JFK and Laguardia airports is relatively higher than their neighboring zones. For the purpose of this report, the three airports are removed because the number of pickups and dropoffs is affected by their usage.

7 Modelling

In this report, we will be using two contrasting models, one is the Ordinary Least Square model, and another one is Lasso Penalised regression. The reasons for choosing these two models are because they are highly interpretable and have low time complexity. These two models should be adequate to show some of the relationships between the response variables and covariates. Since they are both linear models, the linearity between the response variable and the covariates is assumed. This is checked through visualization in Figure 8. In addition, we assume that there is no imperfect collinearity and independence of errors as it will affect the accuracy of our model parameters, which will lead to a false interpretation. To eliminate uninformative attributes, we make use of the step-wise selection method

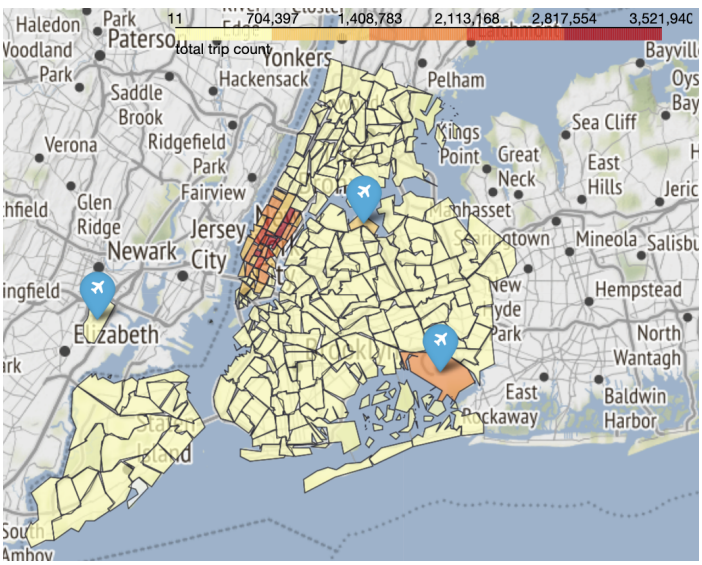

Figure 6: Number of Pickup for all Locations in 2019

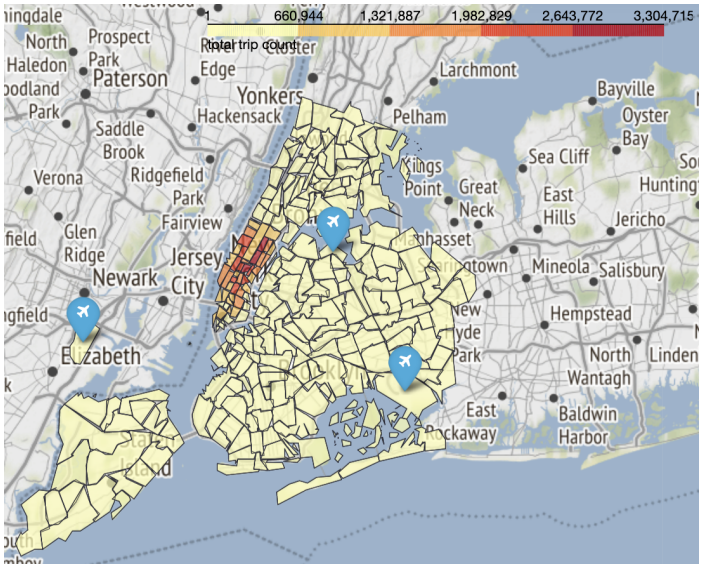

Figure 7: Number of Dropoffs for all Locations in 2019

with AIC as the metric. The evaluation metric used are ANOVA table, R^2, and Root Mean Square Error. The advantage of using RMSE is that it has the same scale as the data, which can help us understand how well our regressions perform on a similar scale. Moreover, HC2 standard errors are used in ANOVA, which can avoid the effect of heteroskedasticity.

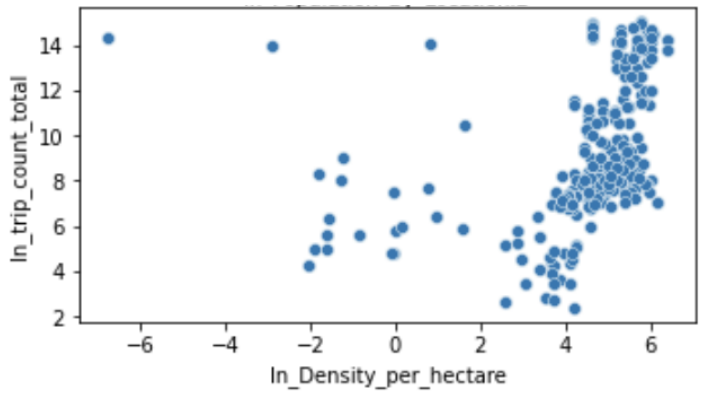

Figure 8: Sample Linearity check of Demographic Attributes

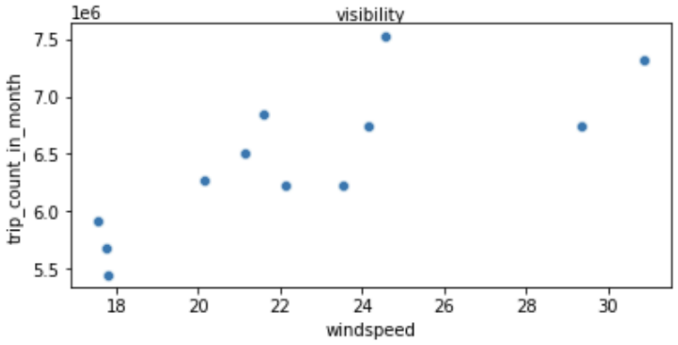

Figure 9: Sample Linearity check of Weather Attributes

7.1 Effect of Weather

7.1.1 OLS Model

Our initial covariates are “feelslike”, “precip”, “windspeed”, “visibility”, “snow” and “snowdepth”. After a step-wise selection is performed, “feelslike” and “precip” are removed. The ANOVA table shows that all of the remaining covariates are statistically significant from 0 at 0.001 significance level11. As shown in figure 10, visibility and snow have a negative effect on the number of trips. One possible explanation could be that bad weather condition makes driver harder to find a customer because people tend to stay indoors until the weather is clear. An interesting fact is that snow depth has a

positive effect on the number of trips, this may happen when snow is deep and other transportation like riding bikes or walking are not suitable. The R^2 of the model is 0.939 which is much higher than the baseline model with a 0 R squares (the result of the baseline model is now shown here due to the limitations of report length), however, the RMSE of 1.597e+06 is very high as well, meaning that our model is not very predictive.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	7.632e+06	1.62e+06	4.719	0.000	4.46e+06	1.08e+07
windspeed	1.277e+05	1.53e+04	8.360	0.000	9.77e+04	1.58e+05
visibility	-2.716e+05	1.17e+05	-2.319	0.020	-5.01e+05	-4.21e+04
snow	-3.012e+06	4.91e+05	-6.135	0.000	-3.97e+06	-2.05e+06
snowdepth	1.61e+06	2.27e+05	7.081	0.000	1.16e+06	2.06e+06

Figure 10: OLS Model Coefficient of Weather

ANOVA TABLE

	df	sum_sq	mean_sq	F	PR(>F)
windspeed	1.0	2.570332e+12	2.570332e+12	69.129954	0.000071
visibility	1.0	8.711548e+10	8.711548e+10	2.343000	0.169702
snow	1.0	3.617224e+09	3.617224e+09	0.097286	0.764196
snowdepth	1.0	1.342217e+12	1.342217e+12	36.099391	0.000538
Residual	7.0	2.602681e+11	3.718116e+10	NaN	NaN

RMSE = 1.597e+06

Figure 11: ANOVA Table of Weather

7.1.2 Lasso Model

The final Lasso regression model is as followed: $trip\_count\_in\_month = 1.03\times10^6 - 7.78\times10^2 feelslike - 7.70\times10^4 precip + 1.20\times10^5 windspeed - 4.25\times10^5 visibility - 3.08\times10^6 snow + 1.63\times10^6 snowdepth$. Noting that the coefficients of the Lasso model are quite different from the OLS model, this is because the Lasso regression does not shrink any of the covariates to zero, which implies that there is no imperfect multicollinearity between covariates. The R^2 of Lasso is 0.9478, which is higher than the OLS model. However, the RMSE has increased from 1.597e+06 to 1.622e+06, which indicates a trend of overfitting.

7.2 Pickup and Dropoff Model

From table 2 and table 4, we can see that all three covariates are statistically significant from 0 at 0.01 significance level. Amongst the three covariates, the log of density per hectare has the highest mean sum of squares, which means that there is a large variation in the number of pickups/dropoffs due to the population density of a zone. Keeping other covariates unchanged, the pickup model indicates that a 1% increase in density per hectare is expected to result in a 2.1149% increase in the pickup number in a zone. For the dropoff model, this number rises to 2.5174%. It is worth noting that, a 1% increase in total population in a zone is associated with a 1.5140% and 1.8432% decrease in the number of pickups and dropoffs, respectively. This suggests that our models are biased toward Manhattan, which is the most densely populated borough in New York City. In addition, both models show a trend of increase in the number of trips as the property price increase, but the p-values are relatively high compared to the other two variables. Therefore, we cannot make a confident conclusion about the effect of the property price. The R^2 is 0.559 for the pickup model and 0.513 for the dropoff model, which means

Attribute	Coef	std err	z	P>‖z‖
Intercept	-1.8982	2.676	-0.709	0.478
ln_Price_per_square_feet	2.6515	0.401	6.617	0.000
ln_Population_By_LocationID	-1.5140	0.144	-10.507	0.000
ln_Density_per_hectare	2.1149	0.164	12.880	0.000

Table 1: Coefficient of Pickup Model

Attribute	df	sum_sq	mean_sq	F	PR>(F)
ln_Price_per_square_feet	1	429.055	429.055	103.215	1.55e-20
ln_Population_By_LocationID	1	37.956	37.956	9.131	2.77e-03
ln_Density_per_hectare	1	857.187	857.187	206.207	1.54e-34
Residual	251	1043.389	4.157		

Table 2: ANOVA Table of Pickup Model

Attribute	Coef	std err	z	P>‖z‖
Intercept	5.7480	4.198	1.369	0.171
ln_Price_per_square_feet	1.2721	0.623	2.042	0.041
ln_Population_By_LocationID	-1.8432	0.171	-10.809	0.000
ln_Density_per_hectare	2.5174	0.186	13.537	0.000

Table 3: Coefficient of Dropoff Model

Attribute	df	sum_sq	mean_sq	F	PR>(F)
ln_Price_per_square_feet	1	72.672	72.672	13.796	2.513e-04
ln_Population_By_LocationID	1	36.537	36.537	6.936	8.973e-03
ln_Density_per_hectare	1	1278.341	1278.341	242.685	8.973e-03
Residual	250	1316.872	5.267		

Table 4: ANOVA Table of Dropoff Model

that our model is able to explain around 50% of variations in the response variable. The RMSE for the pickup model is 3.179, which suggests that our model is not particularly useful for prediction. The RMSE for the dropoff model is 2.445. In comparison to the pickup model, the dropoff model is able to predict more accurate results.

8 Recommendations

The first recommendation is for the taxi company. Based on our findings, regions with high population density have higher demand. This information can be used to allocate taxis for higher profitability, or plan the location of a new branch office to make shifts smoother. The second recommendation is for medallion taxi drivers and TLC. As the model suggests, bad weather condition like heavy snow or low visibility is associated with a decrease in the number of trips, which will have a detrimental effect on taxi drivers' income. TLC is encouraged to find a solution to cope with weather contingencies, for example, by providing this information to taxi drivers and advising them to re-arrange their working time in a more efficient way.

References

[1] Wikipedia. *New York City*. https://en.wikipedia.org/wiki/New_York_City. Accessed: 2022-08.

[2] New York City Taxi and Limousine Commission. *TLC trip record data*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08.

[3] Alexandra Kerr. *A Historical Timeline of COVID-19 in New York City*. https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986. Accessed: 2022-08.

[4] City of New York. *New York City Population By Neighborhood Tabulation Areas*. https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp. Accessed: 2022-08.

[5] City of New York. *Annualized Sales Update*. https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page. Accessed: 2022-08.

[6] Visual Crossing Corporation. *Weather Query Builder*. https://www.visualcrossing.com/weather/weather-data-services. Accessed: 2022-08.

[7] Visual Crossing Corporation. *Weather Data Documentation*. https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/. Accessed: 2022-08.