What Impacts the Earning of the Taxi Drivers?

Predict the Tip Using Generalised Linear Model Regression

Chenyang Dong
Student ID: 1074314

August 18, 2021

1 Introduction

In New York City (NYC), there are over ten thousand yellow taxis. The U.S. Bureau of Labor Statistics (BLS) gives the median annual wage for taxi drivers as $25,880[1]. However, this income is often hard to earn. For the taxi driver, they do not only depend on the standard fare from the trip but also count on the tips, which is unpredictable for them.

According to the trip data collected by NYC Taxi & Limousine Commission (TLC)[2], this report looks into the factors influencing the earning of the yellow-taxi driver and suggests the way of predicting the tip amount using generalised linear model regression. Besides, this report provides the recommendations for the taxi driver on selecting the pick-up region or target; at the same time to help them have a foresight of earning for the trip before accepting the order in the future.

2 Preprocessing

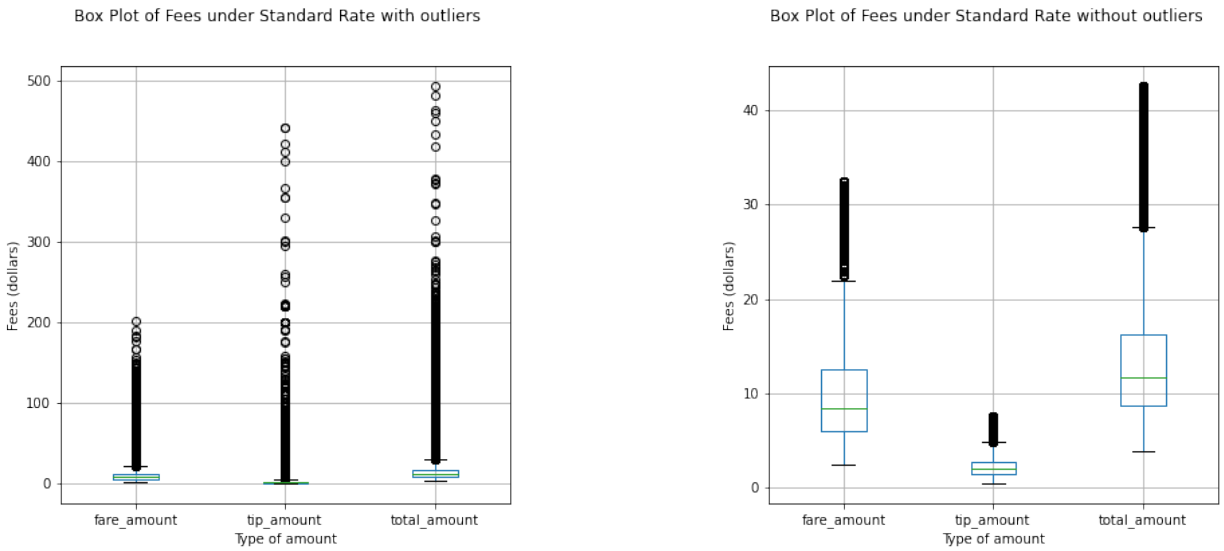
2.1 NYC TLC Dataset

From the website of NYC TLC, three-month (2018/01-03) trip records are retrieved[3]. The raw data contains 26682323 instances. For cleaning, the data dictionary for the trip records provided[3], passenger-frequently-asked-question page[4] and taxi fare page[2] are used as reference. Also, the cleaning is made based on the common sense and consideration for the analysis. It is expected to see the data size reduced a lot, which is necessary for a better analysing outcome.

1. Remove trip with invalid or empty Vendor ID.
2. Remove trip with invalid pick-up or drop-off time such as wrong year or month.
3. Remove trip with unusual duration by restricting between 1 and 120 minutes.
4. Remove trip with invalid passenger number by restricting between 1 and 6[4].
5. Remove trip with unusual distance by restricting between 0.1 and 200 miles.
6. Remove trip with unusually high average speed (greater than 80mph).
7. Remove trip not under standard rate which accounts for large proportion of the data.
8. Remove trip with invalid payment type and cash as tip is not included.

9. Remove trip with invalid fare amount or total amount[2].
10. Remove trip without tax as $0.50 tax should be automatically triggered.
11. Remove trip with unusually high total amount (greater than $500).
12. Remove trip with invalid tip amount by restricting at least 10% of total amount by research[5].

2.1.1 Outliers Removal



(a) Box Plot of Fees with outliers

(b) Box Plot of Fees without outliers

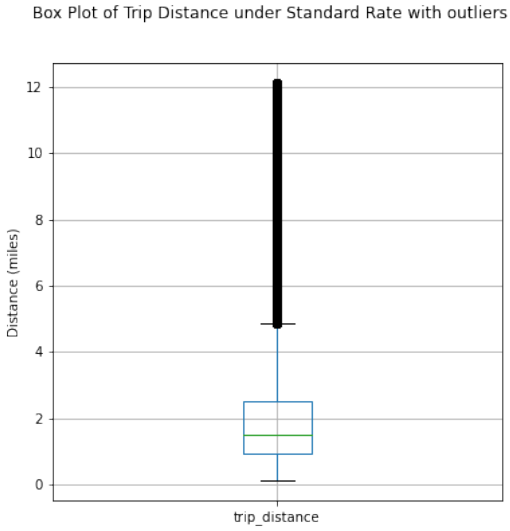
Figure 1: Box Plot of Fees for yellow-taxi trip data

Figure 1 above represents the box plot of fees amount before and after removing outliers. The outliers are removed by quartile, achieved by removing anything not in the range of (Q1 - 3 IQR) and (Q3 + 3 IQR). After that, it appears most trips having fare amount under $12.5, tip amount under $3 and total amount under $16.5 which are 75th-percentile.

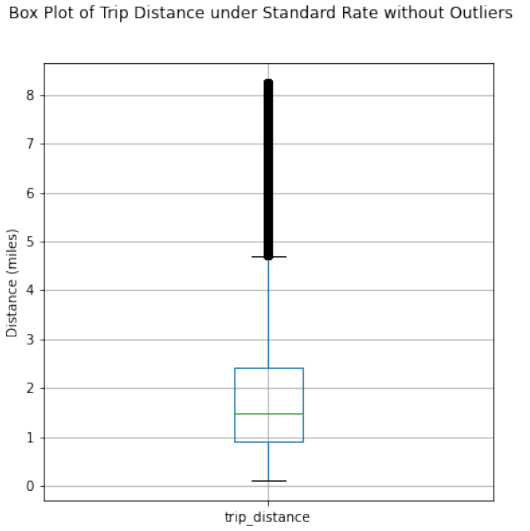
Figure 2 below represents the box plot of trip distance before and after removing outliers. The outliers are removed by quartile as the same way above. The process of removing distance outliers is after, however still using the original quartile for distance to prevent the over-reducing. After that, it appears most trips having distance under 8.5 miles.

Besides, the reason to keep 3 IQR range instead of generally-used 1.5 IQR is that in real cases, more long-distance trips should be considered; otherwise the data would be reduced too much which can only be generalised to short-distance trips.

After preprocessing and outliers removal, the amount of instance reduced to 14650614 which is much more manageable for the analysis from the aspect of data size or data quality.



(a) Box Plot of Trip Distance with outliers



(b) Box Plot of Trip Distance without outliers

Figure 2: Box Plot of Trip Distance for yellow-taxi trip data

2.2 External Climate Dataset

From the website of National Oceanic And Atmospheric Administration (NOAA), the daily summary of the climate records for the corresponding months of the trip records (90 days in total) is retrieved including date, precipitation, snow, wind, air temperature and weather types[6]. The records are collected by the station at the Central Park in Manhattan, which is the center of the city that can cover most of trips.

1. Fill 0 into instance with null wind, precipitation or snow data.
2. Fill instance with null average temperature by calculating the mean of maximum and minimum temperature.
3. Add a new column named 'WT'(Weather Type) which categorised the weather into good or bad by checking existence of any special weather type.
4. Remove the original weather type columns.

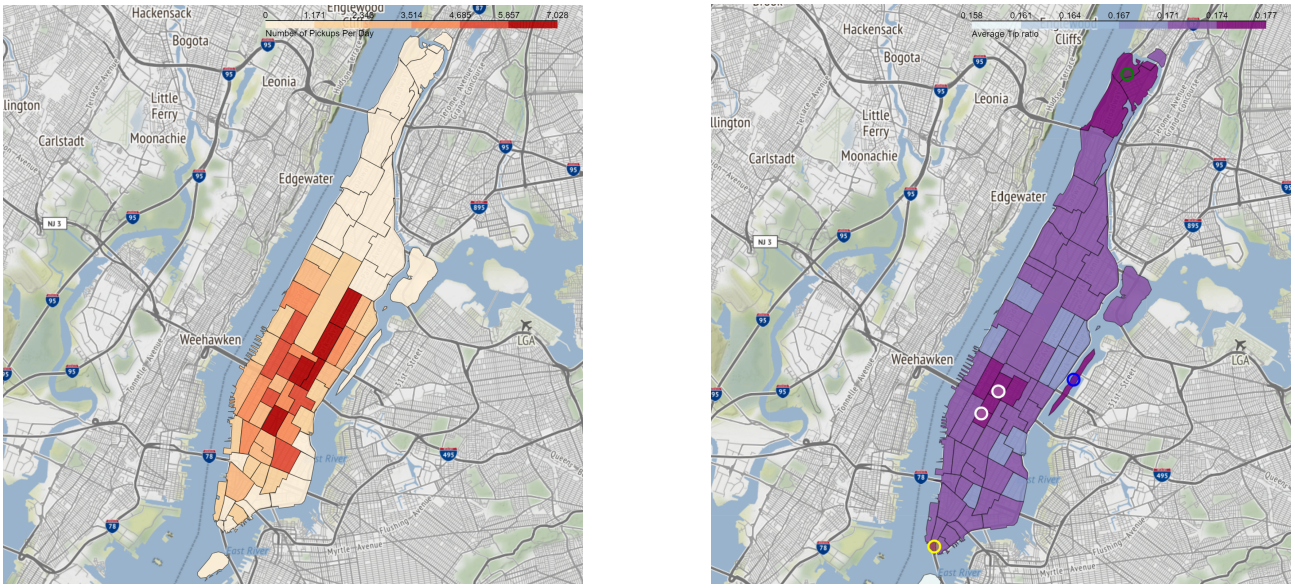
2.3 External Hotel Dataset

From the website of NYC Open Data, the data of hotels is retrieved of 5519 instances[7]. Among all attributes, the most important is location. Therefore for the geospatial visualisation, it needs to unify with trip records. Besides, as the data does not provide the year hotels built, it is assumed that all of them were built before 2018.

1. Remove any other borough code to only focus on hotels in Manhattan for analysis.
2. Add location ID column by using longitude and latitude of the hotel to find corresponding location ID in the shape file of the taxi zones.

3 Preliminary Analysis

3.1 Geospatial Visualisation



(a) Average number of pickups in Manhattan

(b) Average tip ratio of the trip in Manhattan

Figure 3: Visualisation of yellow-taxi trip data in Manhattan

Most of yellow-taxi trips pick up in Manhattan, so the analysis would be focused there. Figure 3 stands for the visualisation of average number of pickups and average tip ratio. It does not specifically show the relationship between amount of trips and average tip ratio. This is predictable as analysing on the average tip ratio instead of total tip amount, thus more investigation is needed.

In Figure 3(b), five spots are circled. They symbolise five places in top five highest tip ratio region, which are probable reasons for their regions having higher tip ratio than others. Two white circles respectively represent Times Square and Madison Square Garden, which one is the most popular tourist attractions and the other is world's most famous arena. The yellow circle represents Financial District which is one of New York's most expensive neighborhoods. The green circle represents InWood Hill Park which is one of biggest family attractions there. Besides, as it is relatively remote far to the center of city, the trip distance would be increased which may increase the tip as well. The blue circle represents Roosevelt Island which have much fewer taxi trips that can be shown in Figure 3(a) than other places due to the special location; therefore it is not representative. From the observations, it can be concluded that the tip ratio of a trip might be influenced by the departure or purpose of the passenger, wealth of the passenger and trip distance.

Before geospatial visualisation, it is assumed that region with more hotels may carry more trips with higher average tip ratio, as commonly believed there picks up more tourists or people in business trip whom could be more willing to give the tip. By comparing Figure 3 and Figure 4, the assumed relationship between number of hotel in the region and number of trips or tip ratio is presented to a certain extent. However, there still has some regions showing a weak relationship or not following the pattern. For example, for the region with greatest amount of hotel, it has much less pickups per day. For those regions, here gives some guesses. As the visualisation only considers the amount of hotel in each region, it does not represent the amount of guests in total; there might be more small hotels with

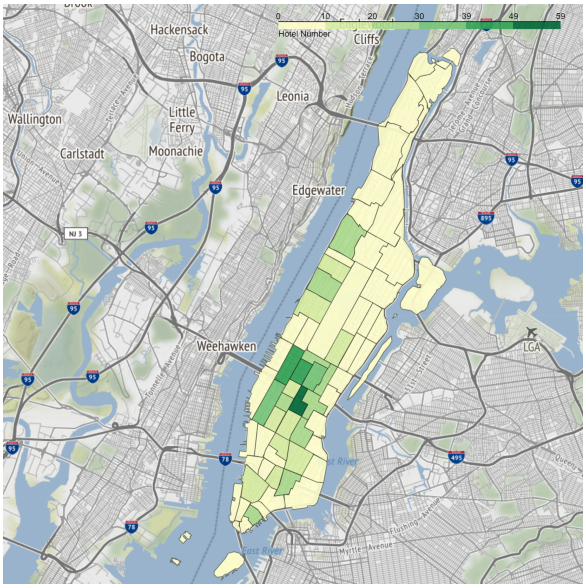


Figure 4: Visualisation of number of hotels in Manhattan

much less amount of beds. In addition, hotel also has different types, as people who stays at luxury hotel might be more likely to take taxi or tip the driver.

3.2 Attribute Analysis

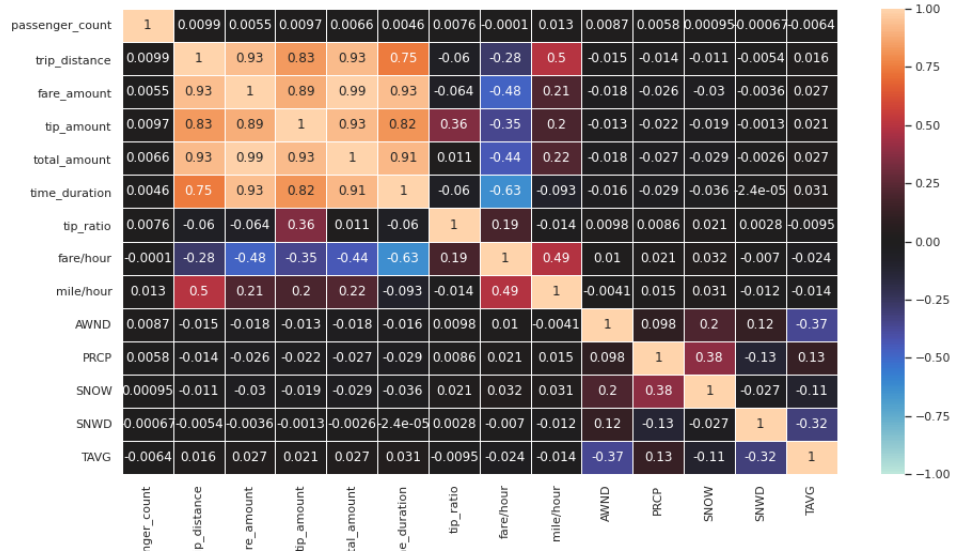


Figure 5: Correlation Heatmap

From Figure 5, the correlation heatmap mainly represents the significantly weak relationship between trip data with external climate data and strong relationship among most trip data. Therefore, the focus of relationship analysis move within the trip attributes.

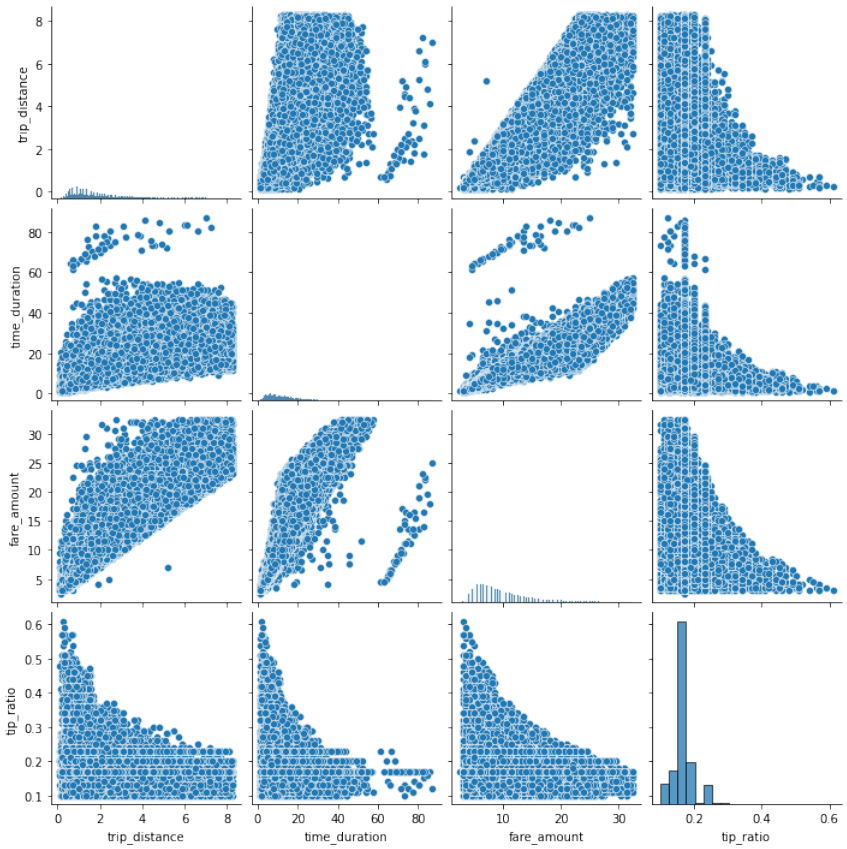


Figure 6: Pair Plot Between Some Trip Data

Figure 6 stands for the relationship between distance, duration, fare amount and tip ratio. From the observation, there seems to have negative association between tip ratio and any other attribute. However, there also found constant lines on the graphs. The most probable reason for that is for many people, they might give the tip in a quite fixed way such as following the minimum-percentage required[5] or rounding up the fee. Refer back to Figure 1(b), tip amount is quite concentrated shown through short range between Q1 and Q3, which strengthen the idea in a certain way. Besides, among trip distance, time duration and fare amount, they have relatively obvious linear associations.

4 Statistical Modelling

Before modelling, the data is split with grouping by date to randomly choose 80 percent of data as train data and the rest as test data for validation test.

4.1 Distribution Selection

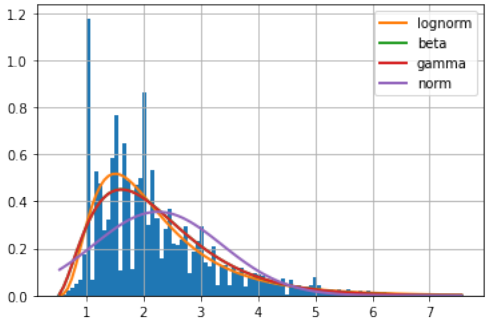


Figure 7: Fitter Model

Before fitting to the model, a fitter function is used from Scipy library to find the best-fit distribution for tip amount shown in Figure 7, which lognorm distribution comes out from several options[8]. After selecting distribution, a Gaussian generalized linear model with log link is ready to apply to the data.

4.2 Model

The Generalised Linear Model:

$$Y_i = \beta_0 + \beta x_i + \epsilon_i \tag{1}$$

In this model, there are three components: one is random component which refers to probability distribution of response variable Y_i ; one is systematic component which specifies the explanatory variables(x_1, x_2...) ; one is link function which specifies the link between the other two.

Back to the case, Y_i is the tip amount; x_i begin to be all other attributes including trip data and climate data except for the datetime and total amount which is because of the direct association (ratio) between tip and total amount; link function is log link.

For feature selection, all insignificant variables are removed and it gives lower AIC and BIC which means the model gets better, as AIC presents the score that the model might overfit, whereas BIC presents the score that it might underfit[9]. Then, the interaction between fare amount and trip distance is included in the model which may help improve the model, as they have large main effects[10] and also strong association between them are shown in Figure 6. After that, it gives the final model.

4.3 Results

By fitting to this model, it gives 0.782 of R^2 and 0.273 of MSE for validation test; 0.670 of R^2 and 0.434 of MSE for prediction on the future data of same months in 2019.

From Table 1, by checking the parameter estimate, it is shown that fare amount and trip distance have the greatest impact on tip amount.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0973	0.003	-32.361	0.000
day_type[T.Weekend]	-0.0222	0.001	-15.309	0.000
passenger_count	0.0022	0.000	6.076	0.000
fare_amount	0.0941	0.001	123.532	0.000
trip_distance	0.1713	0.002	105.096	0.000
fare_amount:trip_distance	-0.0077	4.04e-05	-190.345	0.000
time_duration	-0.0101	0.000	-30.492	0.000
mph	-0.0107	0.000	-45.074	0.000
SNOW	0.0016	0.000	3.538	0.000
SNWD	0.0007	0.000	2.655	0.000

Table 1: GLM Results Table

4.4 Discussion

The model is fitted to predict the future data in 2019, which is also restricted to same months of the train data, therefore to avoid other factors due to the month difference influencing the result. Those data is used after they have taken exactly the same preprocessing as the train data.

To ensure the reliability of distribution selection, other family options suitable to the data such as Gamma and inverse Gaussian are also tested by checking the R^2 and MSE, which all give relatively worse results.

For R^2, it indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, which means the higher the R^2, the better the model fits the data. Generally speaking, the value of R^2 shows a better performance of the model than the expectation. However, for the reason that R^2 of prediction gets decreased, there might be existing inherent variability of data affecting it. Also, as the increase of the data size, there might be some non-linear relationship pattern in the predicting data which influence the result as well. Hence, some non-linear model could be implemented such as trees or neural network.

In the future, more feature engineering could be considered to help with analysis, such as the time of a day. Instead of using regression, classification can also be used for more findings, such as predicting the tip level. Also, the model only fits on the small sample of the full dataset due to hardware performance, which is far less than the amount of data for prediction. If possible in the further investigation, more data could be trained to have a better performed model.

5 Recommendations

- Taxi drivers may choose to find the passenger in the region with places such as sightseeing spots, venue and arena where is more likely to attract tourists or audiences; and it is more probable for them to take taxi after visiting or finishing the event.
- May choose to find the passenger in the region with more hotels, especially for those luxury hotel. Not only because there would be more guests willing to take the taxi, but also it is more possible for people staying in there give more tips.
- May choose to find the passenger in the expensive neighborhood, especially those with prestigious

apartments. With a higher chance, there would pick more wealthy passengers and receive higher amount of tips.

- May choose to take more short-distance trips for having tipped in a higher proportion of total amount, as passenger maybe tip to round up the fee or follow certain tip percentage. However, may consider more on the trip distance for higher amount of the tip.

6 Conclusion

With comparison between visualisation of attributes, either geospatial or pair plot, the association between attributes is clearly represented. Some assumptions are primarily observed and some are rebutted, which all give intriguing findings on the taxi trip, especially on the aspect of the tip.

To predict the tip amount, the data is fitted to a Gaussian generalized linear model with log link. With this model, it gives a not bad prediction, which comes to believe that the driver could give thought to predicting the probable tip might be earned before accepting the trip, which for the purpose of the report, to help them have a better understanding of the factors influencing on their earning especially on the tips.

References

[1] Taxi Drivers and Chauffeurs - Bureau of Labor Statistics
https://www.bls.gov/ooh/transportation-and-material-moving/taxi-drivers-and-chauffeurs.html

[2] Taxi Fare - TLC.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[3] TLC Trip Record Data - TLC
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[4] Passenger Frequently Asked Questions - TLC
https://www1.nyc.gov/site/tlc/passengers/passenger-frequently-asked-questions.page

[5] A Guide to Tipping in New York City
https://www.tripsavvy.com/guide-to-tipping-in-new-york-city-4177115

[6] Daily Summaries Station Details: NY CITY CENTRAL PARK - Climate Data Online - National Climatic Data Center (NCDC) - NOAA
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[7] Hotels Properties Citywide - NYC Open Data
https://data.cityofnewyork.us/City-Government/Hotels-Properties-Citywide/tjus-cn27

[8] Finding the Best Distribution that Fits Your Data using Python's Fitter Library - Towards Data Science
https://towardsdatascience.com/finding-the-best-distribution-that-fits-your-data-using-pythons-fitter-library-319a5a0972e9

[9] Is there any reason to prefer the AIC or BIC over the other - StackExchange
https://stats.stackexchange.com/questions/577/is-there-any-reason-to-prefer-the-aic-or-bic-over-the-other

[10] Why and When to Include Interactions in a Regression Model - Quantifying Health
https://quantifyinghealth.com/why-and-when-to-include-interactions-in-a-regression-model/