**Work Sample**

**Instructions:**

Using the data provided, complete each of the two tasks below. Python, R & Power BI are our tools of choice to complete the tasks. Your email back to NBC should include your created report, the code (Python/R) used to clean and analyze the data, and a word document with a paragraph as asked for in each task below.

Context:

Managers of the Green Taxi are interested in figuring out the trends of tips given during the trips taken in Jan 2020 & Jan 2019. They have asked you to look into this for them.

Data to use:

NYC Taxi & Limousine Commission Trip Data –

<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Jan 2020 - Green Taxi Trip Records (CSV)

Jan 2019- Green Taxi Trip Records (CSV)

Task 1:

After looking at the data, write a paragraph detailing any possible issues/anomalies that you might see in the data and explain how you would choose to deal with them in your analytics.

**Answer:** Both the dataset has New York Green Taxi trip data contains information from previous years, which does not correspond to the Jan 2019 and Jan 2020. For example, the ‘Jan 2020 - Green Taxi Trip Records’ includes New year eve (Dec 3tst 2019) trip information. The problem is severe in the case of ‘Jan 2019 - Green Taxi Trip Records’, which includes trip information of 1/1/2009, 9/23/2010, 12/4/2018, 12/5/2018, 12/7/2018, 12/21/2018, and 12/31/2018. Furthermore, a large number of ‘RateCodeID’, ‘passenger\_count’, and ‘trip\_type’ corresponding to the 2020 trip information were missing.

Our objective here is to compare the trips trend between January of 2019 and 2020. Therefore, we dropped the tip information of dates other than January of 2019 and 2020. However, the trips that started on December 31 and ended in the early hours of January were included in the analysis. The filter function in R was used for this purpose.

Data cleaning: While R by default dropped those missing observations, we utilize, filter function to remove the error observation ‘99’.

The corresponding R codes that were used to the data analytics were attached.

Task 2:

Create a report displaying major factors effecting tips, trends, similarities, or differences that you find about probability of tips being given and the possible amounts. Analyze how these factors and trends changed between 2019 & 2020. Write a paragraph summarizing the major points of your report.

**Answer:** Based on the frequency of trips over Jan 2019 and 2020 (as shown in figure 1a, b), we found the following similarities in the Green taxi use trends in New York:

1. On both years early morning on January 1, there was a significant increase in Green taxi use
2. Each week Sunday afternoon and evening, the green taxi demand was substantially low
3. The taxi ride demand was high every weekday during the rush hours between 5-10 am and 3-8 pm.

Differences:

The overall number of Green taxi ride was relatively low in 2020 compared to 2019.

While the average amount of fare including, toll, tax, improvement surcharge, rush hours, and overnight charges plus tip (only paid by credit card) remain the same for 2019 and 2020 in case of the street-hailed trip, in 2019 dispatched taxi revenue was higher (as shown in figure 2a, b).

With the higher volume of passengers, the total amount of fares increases in 2019 and 2020. In 2019 the green taxi fare is the highest when there were 9 passengers (as shown in figure 3a, b). In contrast, in 2020, the taxi fare was maximum when there were 7 passengers.

The total daily trip distance covered by green taxi in Jan 2019 was longer than the 2020’s trips (as shown in figure 4a, b).

In both years, the highest fare was received for the JFK, Newark, Nassau or Westchester as end point of the ride, while negotiated and group ride fares were the lowest (as shown in figure 5a, b).

The following regression equation is estimated to determine the factors effecting trip average fare:

total\_amountt= a0 + a1 dayt + a2 hourt + a3 trip\_distance t+ a4 passenger\_countt + a5 trip\_typet + a5 RatecodeIDt + et

t (2019, 2020)

In 2019, days, hours, trip distance, number of passengers, end of the trip (RatecodeID) positively impacted the average total fare. However, when the trip was dispatched, the average total fare decreased by $7.29.

In 2020, the significant difference was that the average total fare decreases as the number of passengers increase, while trip type (street-hailed or dispatched) still negatively impacted the fare, but at a substantially lower rate of $1.72.

**Table 1. 2019 average total fare estimation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Average fare** | **Estimate** | **Std. Error** | **t value** | Pr(>|t|) |
| (Intercept) | 10.888 | 0.074 | 146.94 | <2e-16 |
| day | 0.013 | 0.001 | 19.68 | <2e-16 |
| hour | 0.033 | 0.001 | 32.42 | <2e-16 |
| trip\_distance | 2.993 | 0.002 | 1976.78 | <2e-16 |
| passenger\_count | 0.070 | 0.006 | 12.07 | <2e-16 |
| trip\_type | -7.287 | 0.093 | -78.12 | <2e-16 |
| RatecodeID | 1.521 | 0.023 | 65.18 | <2e-16 |
|  |  |  |  |  |
| **n** |  |  |  | 630831 |
| **R-squared** |  |  |  | 0.8754 |

**Table 2. 2020 average fare estimation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | 6.643 | 0.134 | 49.5 | < 2e-16 |
| day | 0.021 | 0.001 | 16.25 | < 2e-16 |
| hour | 0.015 | 0.002 | 7.91 | 2.59E-15 |
| trip\_distance | 3.074 | 0.004 | 831.39 | < 2e-16 |
| passenger\_count | -0.133 | 0.012 | -11.07 | < 2e-16 |
| trip\_type | -1.721 | 0.157 | -10.96 | < 2e-16 |
| RatecodeID | 1.577 | 0.037 | 42.95 | < 2e-16 |
|  |  |  |  |  |
| **n** |  |  |  | 331685 |
| **R-squared** |  |  |  | 0.6809 |

A picture containing timeline

Description automatically generatedA picture containing timeline

Description automatically generated

**Figure 1. Total number of Green Taxi ride in New York, a) January 2019 b) February 2020**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 2. Total average Green Taxi daily fare income by trip type, a) 2019 b) 2020**

Trip type 1: Trip was a street-hail

Trip type 2: A dispatch that was automatically assigned based on the metered rate in use but can be altered by the driver.

Chart, bar chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generated

**Figure 3. Total average daily fare by the number of passenger a) 2019 b) 2020**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Figure 4. Total daily trip distance a) 2019 b) 2020**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 5. Total average daily fare by the end of the trip code (RatecodeID) a) 2019 b) 2020**

**RatecodeID**

1= Standard rate

2=JFK

3=Newark

4=Nassau or Westchester

5=Negotiated fare

6=Group ride