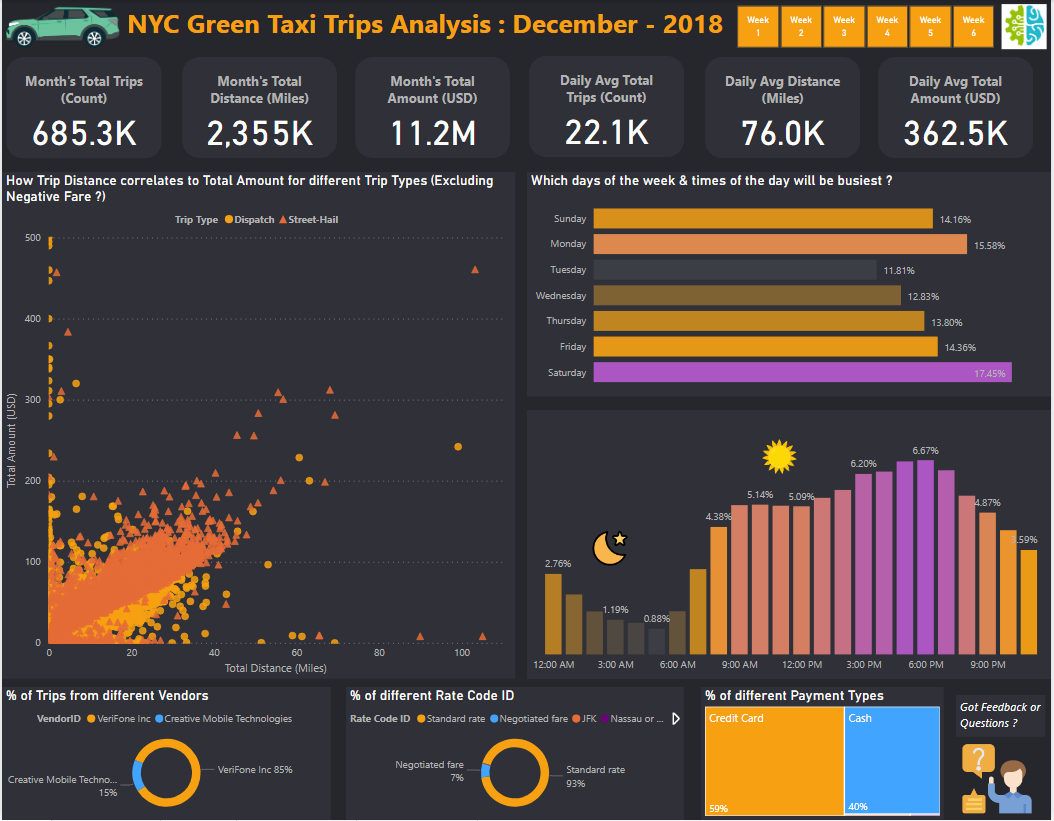
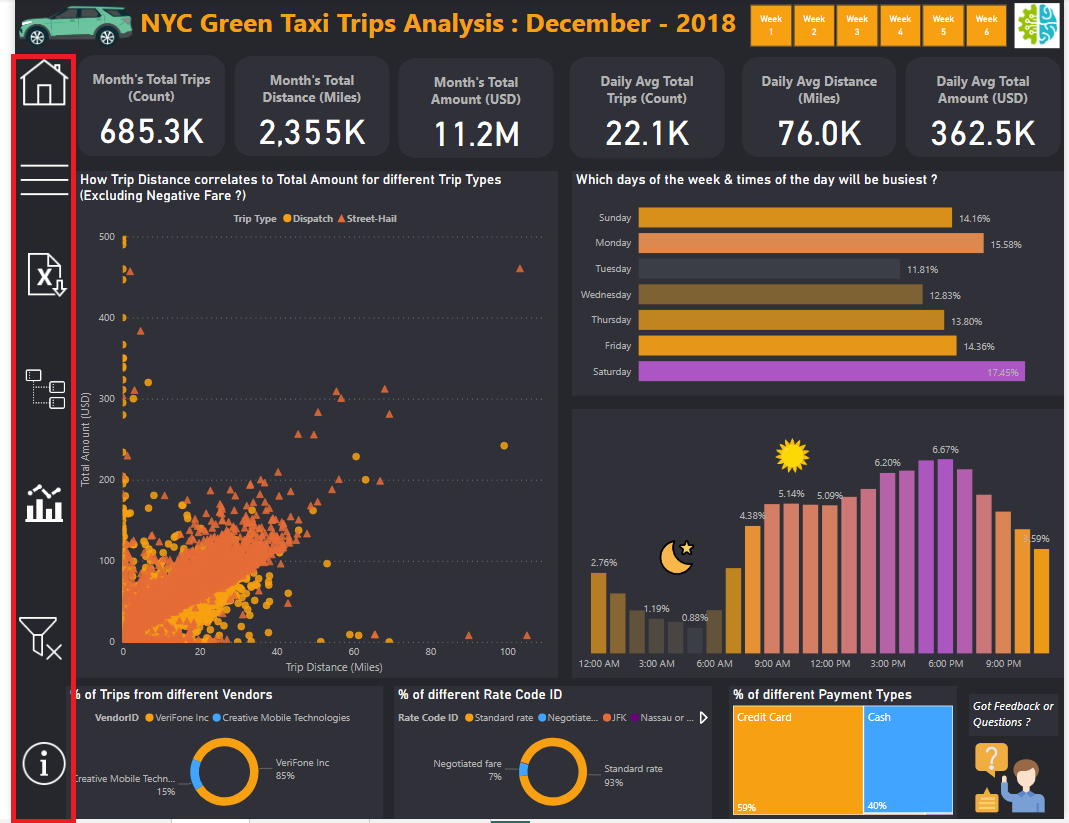
**Dataset Observations on Green Taxi Analysis**

**About Dataset –**

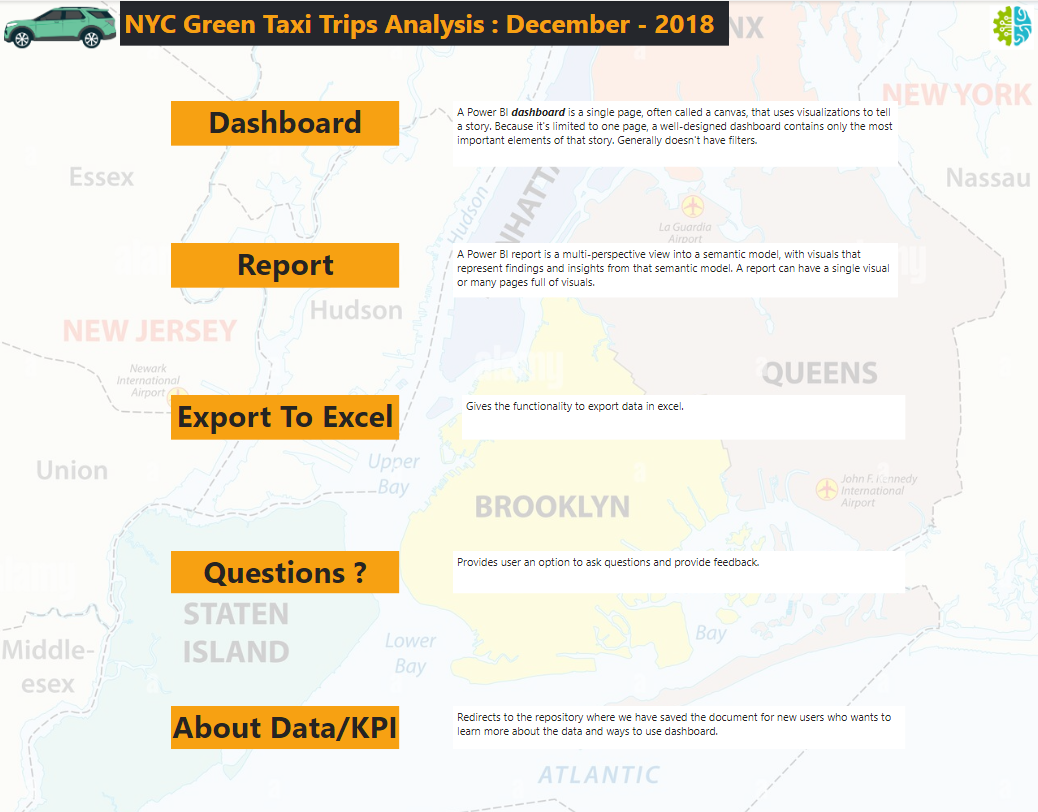
* Trip Records: Each record in the dataset represents a single taxi trip, encompassing various details such as pick-up and drop-off times, locations, distances, fares, number of passengers, and additional information.
* Date Table: The date table included in the dataset encompasses a fiscal calendar spanning the years 2008 to 2019. This table is utilized by the NYC Taxi system and includes fields containing specific details for each date, such as the fiscal year, quarter, month, and week.

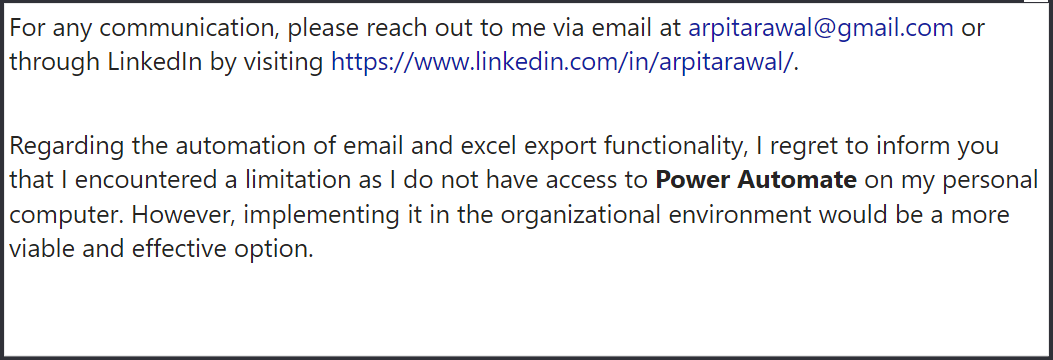
**Report Snapshot –**

****

****

****

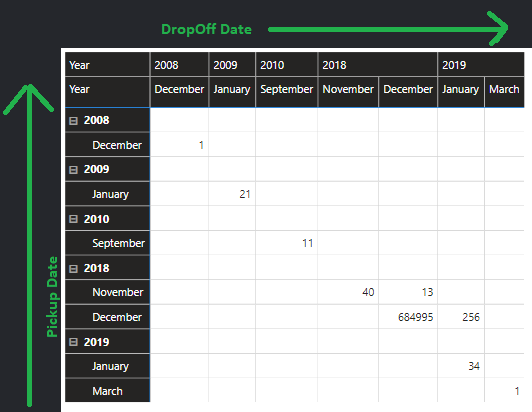
****

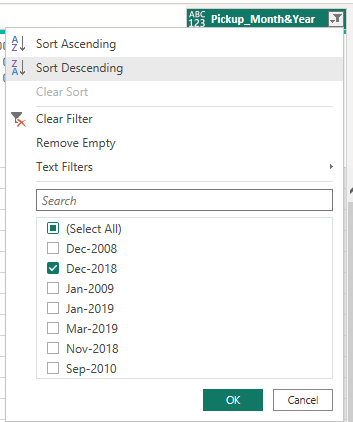
****

**Data Quality check–**

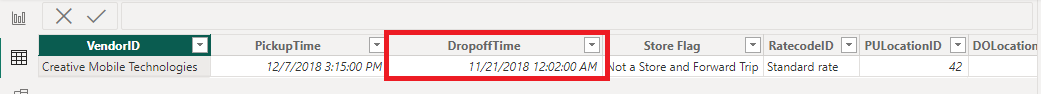
1. Outliers Analysis: In the course of examining outliers, it has come to my attention that there are approximately 121 records in which the Pickup/Drop-off date deviates from December 2018. As of now, there is insufficient evidence to justify replacing these 121 records with December 2018 as their pickup and drop-off date.

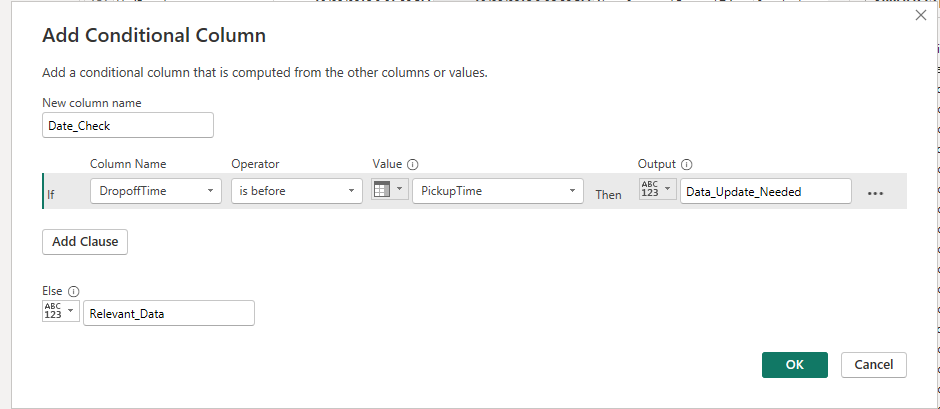
Pending the acquisition of substantial evidence, I have opted to exclude these outliers from our analysis. This approach ensures the integrity and reliability of our current analysis until further clarification is obtained.

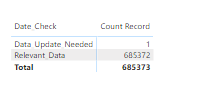
****

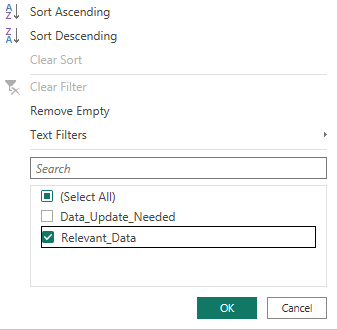
****

1. Data Integrity Issue: A single record has been identified wherein the Drop-off Time precedes the Pickup Time. In light of this anomaly, I have opted to exclude this particular record from our analysis until further details are available to guide the necessary update.

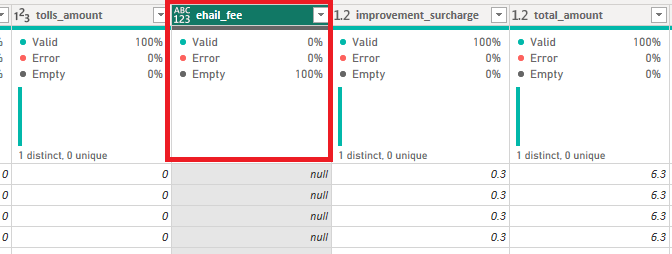
It is worth noting that a simple swapping of times may not be a suitable solution in this case, as a substantial date difference of 16 days persists. This discrepancy underscores the presence of an underlying date-related issue that warrants further investigation and correction.

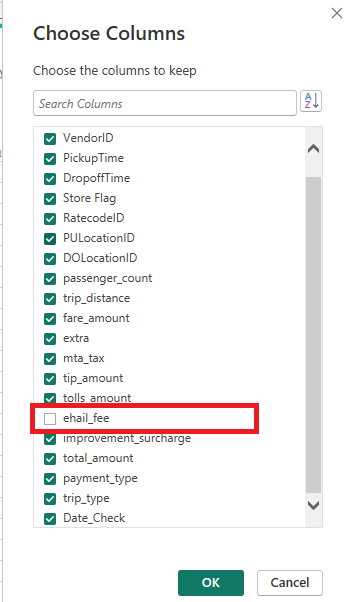




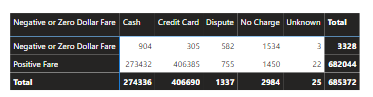


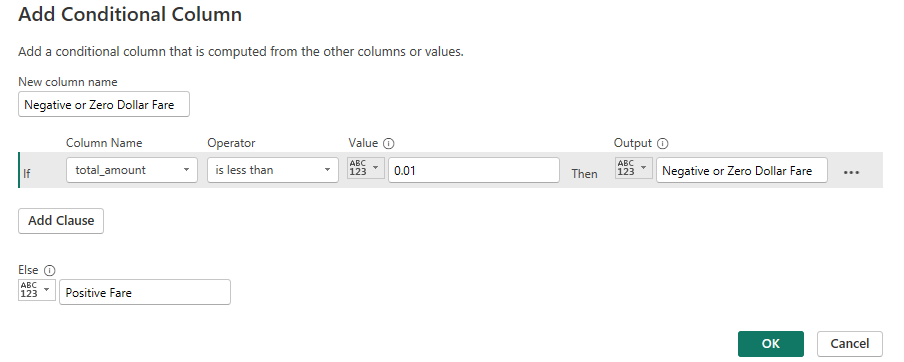
3) In the column quality check, we found out that ehail\_fee is completely blank, so for now we have removed it from the view, we have not deleted it. Considering if data in this column will show up in the future.





4) To find the records with Negative fare and zero dollar value, we create a calculated column. We found out that we had around 3328 records with negative value. Further clarification is required to understand the negative value input and update the values accordingly.





**Assumptions –**

1. Let’s assume any trips with no recorded passengers had 1 passenger. We had 1362 trips, where recorded passengers were 0.

**Code –**

1. Code to create Date Table –

Date =

ADDCOLUMNS (

CALENDAR (DATE (2018, 12, 1), DATE (2018, 12, 31)),

"Year", YEAR([Date]),

"MonthNumber", MONTH([Date]),

"Quarter", QUARTER([Date]),

"DayOfWeek", WEEKDAY([Date]), "Month", Format ([Date],"mmmm"), "Day", FORMAT([Date],"dddd") , "WeekNumber", WEEKNUM([Date])

)

1. Total number of unique service days (considering if there are strick/off on some days)

Total\_Number\_of\_days = DISTINCTCOUNT('Sheet1'[Pickup Date])

Daily Avg Total Trip = Avg Total Trip = [Count\_No\_of\_Trips] / [Total\_Number\_of\_days]

Count\_No\_of\_Trips = COUNT(Sheet1[VendorID])

Daily Avg Total Distance =

DIVIDE(SUM('Sheet1'[trip\_distance]), [Total\_Number\_of\_days])

Daily Avg Total Amount =

DIVIDE(SUM('Sheet1'[total\_amount]), [Total\_Number\_of\_days])

1. Trip Time = DATEDIFF(Sheet1[PickupTime], Sheet1[DropoffTime], MINUTE)

Rest of the code was written in Power Query using M.