New York City TLC Project - Preliminary Data Analysis

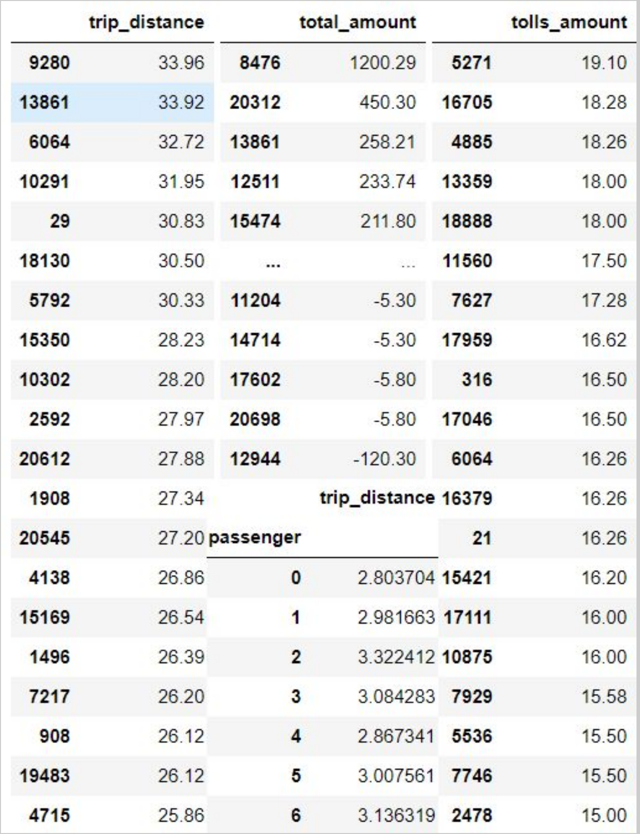
By **Automatidata**

Overview

At this stage of this project, a preliminary analysis of the TLC data is performed to understand the key data variable and verify the suitability of information for generating insights and predictive model.

Key Insights

- This dataset includes variables that seems relevant in building predictive model.
- Unusual value (-negative amount) is noticed for total_amount for some records.
- The passenger_count does not seems correlated with trip_distance.



	trip_distance
9280	33.96
13861	33.92
6064	32.72
10291	31.95
29	30.83
18130	30.50
5792	30.33
15350	28.23
10302	28.20
2592	27.97
20612	27.88
1908	27.34
20545	27.20
4138	26.86
15169	26.54
1496	26.39
7217	26.20
908	26.12
19483	26.12
4715	25.86

	total_amount
8476	1200.29
20312	450.30
13861	258.21
12511	233.74
15474	211.80
...	...
11204	-5.30
14714	-5.30
17602	-5.80
20698	-5.80
12944	-120.30

	trip_distance
passenger	
0	2.803704
1	2.981663
2	3.322412
3	3.084283
4	2.867341
5	3.007561
6	3.136319

	tolls_amount
5271	19.10
16705	18.28
4885	18.26
13359	18.00
18888	18.00
11560	17.50
7627	17.28
17959	16.62
316	16.50
17046	16.50
6064	16.26
16379	16.26
21	16.26
15421	16.20
17111	16.00
10875	16.00
7929	15.58
5536	15.50
7746	15.50
2478	15.00

Details

- We have investigated a 0.02% sample data of TLC complete dataset.
- Explored dataset to find any unusual value.
- trip_distance, total_amount, and tolls_amount seems most useful to build model(s) for predicting trip_duration.
- Investigate the relationship between two chosen variables.
- Built up the basic understanding of TLC dataset for future exploratory data analysis, insights, visualizations, and models.
- We will perform EDA as next step of this project

Next Steps

1. Perform Exploratory Data Analysis on Complete Dataset.
2. Use Data Cleaning and Manipulation for predictive modeling.
3. Perform Data Wrangling on data variables like datetime that support predictive modeling.
4. We can further analyse variables correlation to find best variables for predictive model(s)