Executive Summary

Understand The Data

Stage Overview

This stage aims to investigate and understand the data provided, use a data frame constructed within Python, perform a cursory inspection of the provided dataset, and inform team members of my findings.

Key Findings

- The data includes 22,699 rows and 17 columns
- The two most important variables for training our model are the trip distance and the tip amount
- The data types of date-time variables are objects.
- We just have the IDs of the start and the end location of each trip. We don't have their Latitude and longitude, or any other useful information about them.
- Validity issues in the data:
  - Negative amount of money in columns related to total amount
  - Observation in Ratecode & extra columns with values are not in their intervals
- Everything else in the data is good.

Visual Details

Variables Selection

5. Select the two most important variables for training our model

```
In [14]: df.corr()[["total_amount"]]
Out[14]:
```

	total_amount
VendorID	0.000687
passenger_count	0.007724
trip_distance	0.767182
RatecodeID	0.226681
PULocationID	-0.050302
DOLocationID	-0.062068
payment_type	-0.118319
fare_amount	0.987303
extra	0.104406
mta_tax	-0.199457
tip_amount	0.770913
tolls_amount	0.584475
improvement_surcharge	0.043972
total_amount	1.000000

```
In [15]: df.groupby("store_and_fwd_flag").agg(["median", "mean"])[["total_amount"]]
Out[15]:
```

	total_amount	
	median	mean
store_and_fwd_flag		
N	11.80	16.301617
Y	13.55	18.338788

Note: The two most important variabel for training our model are trip distance and the tip_amount

Validity Issues

	RatecodeID	fare_amount	extra	mta_tax	improvement_surcharge
count	22699.000000	22699.000000	22699.000000	22699.000000	22699.000000
mean	1.043394	13.026629	0.333275	0.497445	0.299551
std	0.708391	13.243791	0.463097	0.039465	0.015673
min	1.000000	-120.000000	-1.000000	-0.500000	-0.300000
25%	1.000000	6.500000	0.000000	0.500000	0.300000
50%	1.000000	9.500000	0.000000	0.500000	0.300000
75%	1.000000	14.500000	0.500000	0.500000	0.300000
max	99.000000	999.990000	4.500000	0.500000	0.300000

Next Steps

- Modify the structure of date-time variables to the appropriate structure and split them into month, day, and hour columns.
- Check if we can access the values of ID locations or not
- Clean the data from validity issues
- Analyze the data.