Activity_Course 3 Automatidata project lab

April 30, 2023

0.1 Course 3 Automatidata project

Course 3 - Go Beyond the Numbers: Translate Data into Insights

1 Course 3 end-of-course project: Exploratory data analysis

In this activity, you will examine data provided and prepare it for analysis.

The purpose of this project is to conduct exploratory data analysis on a provided data set.

The goal is to clean data set and create a visualization.
This activity has 4 parts:

Part 1: Imports, links, and loading

Part 2: Data Exploration * Data cleaning

Part 3: Building visualizations

Part 4: Evaluate and share results

Follow the instructions and answer the questions below to complete the activity. Then, you will complete an Executive Summary using the questions listed on the PACE Strategy Document.

Be sure to complete this activity before moving on. The next course item will provide you with a completed exemplar to compare to your own work.

Welcome to the New York City Taxi project!

You are the newest data professional in a fictional data analytics firm: Automatidata. The team is still early into the project, having only just completed an initial plan of action and some early Python coding work.

Opening your company email, you notice a message from Luana Rodriquez, the senior data analyst at Automatidata. Luana is pleased with the work you have already completed and requests your assistance with some EDA and data visualization work for the New York City Taxi and Limousine Commission project (New York City TLC).

Recall that you have a helpful tool at your disposal! Refer to the PACE Strategy Document here to help apply your learnings, apply new problem-solving skills, and guide your approach to this project.

PACE strategy document

2 Visualize a story in Tableau and Python

In this activity, you will design a professional data visualization that tells a story, and will help someone make a data-driven decision for their business needs. Please note that this activity is optional, and will not affect your completion of the course.

Completing this activity will help you practice planning out and plotting a data visualization based on a specific business need. The structure of this activity is designed to emulate the proposals you will likely be assigned in your career as a data professional. Completing this activity will help prepare you for those career moments.

Follow the instructions and answer the question below to complete the activity. Then, you will complete an executive summary using the questions listed on the PACE Strategy Document.

Be sure to complete this activity before moving on. The next course item will provide you with a completed exemplar to compare to your own work.

2.1 PACE stages

- `[Plan](#scrollTo=psz51YkZVwtN&line=3&uniqifier=1)`
- `[Analyze](#scrollTo=mA7Mz_SnI8km&line=4&uniqifier=1)`
- `[Construct](#scrollTo=Lca9c8XON8lc&line=2&uniqifier=1)`
- `[Execute](#scrollTo=401PgchTPr4E&line=2&uniqifier=1)`

3 Pace: Plan Stage

In this stage, consider the following questions where applicable to complete your code response: 1. Identify any outliers:

- What methods are best for identifying outliers?
- How do you make the decision to keep or exclude outliers from any future models?

3.1 Step 1. Imports, links, and loading

Go to Tableau Public The following link will help you complete this activity. Keep Tableau Public open as you proceed to the next steps.

Link to supporting materials: Tableau Public: https://public.tableau.com/s/

For EDA of the data, import the data and packages that would be most helpful, such as pandas, numpy and matplotlib.

The code to read in the dataset is provided.

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import datetime as dt
import seaborn as sns
```

[2]:
```
df = pd.read_csv('2017_Yellow_Taxi_Trip_Data.csv')
```

4 pAce: Analyze Stage

In this stage, consider the following questions where applicable to complete your code response: 1. Does the data need to be restructured or converted into usable formats?

2. Is there any categorical data that needs to be converted to numerical data?

4.1 Step 2a: Data Exploration & Cleaning

Decide which columns are applicable

The first step is to assess your data. Check the Data Source page on Tableau Public to get a sense of the size, shape and makeup of the data set. Then answer these questions to yourself:

Given our scenario, which data columns are most applicable? Which data columns can I eliminate, knowing they won't solve our problem scenario?

Consider functions that help you understand and structure the data.

- head()
- describe()
- info()
- groupby()
- sortby()

What do you do about missing data (if any)?

Are there data outliers? What are they and how might you handle them?

Start by discovering, using head and size.

[3]:
```
df.head(10)
```

[3]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	\
0	24870114	2	03/25/2017 8:55:43 AM	03/25/2017 9:09:47 AM	
1	35634249	1	04/11/2017 2:53:28 PM	04/11/2017 3:19:58 PM	
2	106203690	1	12/15/2017 7:26:56 AM	12/15/2017 7:34:08 AM	
3	38942136	2	05/07/2017 1:17:59 PM	05/07/2017 1:48:14 PM	
4	30841670	2	04/15/2017 11:32:20 PM	04/15/2017 11:49:03 PM	
5	23345809	2	03/25/2017 8:34:11 PM	03/25/2017 8:42:11 PM	
6	37660487	2	05/03/2017 7:04:09 PM	05/03/2017 8:03:47 PM	

```
7    69059411         2   08/15/2017 5:41:06 PM   08/15/2017 6:03:05 PM
8     8433159         2   02/04/2017 4:17:07 PM   02/04/2017 4:29:14 PM
9    95294817         1   11/10/2017 3:20:29 PM   11/10/2017 3:40:55 PM

   passenger_count  trip_distance  RatecodeID store_and_fwd_flag  \
0                6           3.34           1                  N
1                1           1.80           1                  N
2                1           1.00           1                  N
3                1           3.70           1                  N
4                1           4.37           1                  N
5                6           2.30           1                  N
6                1          12.83           1                  N
7                1           2.98           1                  N
8                1           1.20           1                  N
9                1           1.60           1                  N

   PULocationID  DOLocationID  payment_type  fare_amount  extra  mta_tax  \
0           100           231             1         13.0    0.0      0.5
1           186            43             1         16.0    0.0      0.5
2           262           236             1          6.5    0.0      0.5
3           188            97             1         20.5    0.0      0.5
4             4           112             2         16.5    0.5      0.5
5           161           236             1          9.0    0.5      0.5
6            79           241             1         47.5    1.0      0.5
7           237           114             1         16.0    1.0      0.5
8           234           249             2          9.0    0.0      0.5
9           239           237             1         13.0    0.0      0.5

   tip_amount  tolls_amount  improvement_surcharge  total_amount
0        2.76           0.0                    0.3         16.56
1        4.00           0.0                    0.3         20.80
2        1.45           0.0                    0.3          8.75
3        6.39           0.0                    0.3         27.69
4        0.00           0.0                    0.3         17.80
5        2.06           0.0                    0.3         12.36
6        9.86           0.0                    0.3         59.16
7        1.78           0.0                    0.3         19.58
8        0.00           0.0                    0.3          9.80
9        2.75           0.0                    0.3         16.55
```

[4]:
```
print(df.size)
print(df.shape)
```

```
408582
(22699, 18)
```

Use describe…

[5]:
```
df.describe()
```

[5]:

	Unnamed: 0	VendorID	passenger_count	trip_distance
count	2.269900e+04	22699.000000	22699.000000	22699.000000
mean	5.675849e+07	1.556236	1.642319	2.913313
std	3.274493e+07	0.496838	1.285231	3.653171
min	1.212700e+04	1.000000	0.000000	0.000000
25%	2.852056e+07	1.000000	1.000000	0.990000
50%	5.673150e+07	2.000000	1.000000	1.610000
75%	8.537452e+07	2.000000	2.000000	3.060000
max	1.134863e+08	2.000000	6.000000	33.960000

	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amount
count	22699.000000	22699.000000	22699.000000	22699.000000	22699.000000
mean	1.043394	162.412353	161.527997	1.336887	13.026629
std	0.708391	66.633373	70.139691	0.496211	13.243791
min	1.000000	1.000000	1.000000	1.000000	-120.000000
25%	1.000000	114.000000	112.000000	1.000000	6.500000
50%	1.000000	162.000000	162.000000	1.000000	9.500000
75%	1.000000	233.000000	233.000000	2.000000	14.500000
max	99.000000	265.000000	265.000000	4.000000	999.990000

	extra	mta_tax	tip_amount	tolls_amount
count	22699.000000	22699.000000	22699.000000	22699.000000
mean	0.333275	0.497445	1.835781	0.312542
std	0.463097	0.039465	2.800626	1.399212
min	-1.000000	-0.500000	0.000000	0.000000
25%	0.000000	0.500000	0.000000	0.000000
50%	0.000000	0.500000	1.350000	0.000000
75%	0.500000	0.500000	2.450000	0.000000
max	4.500000	0.500000	200.000000	19.100000

	improvement_surcharge	total_amount
count	22699.000000	22699.000000
mean	0.299551	16.310502
std	0.015673	16.097295
min	-0.300000	-120.300000
25%	0.300000	8.750000
50%	0.300000	11.800000
75%	0.300000	17.800000
max	0.300000	1200.290000

And info.

[6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22699 entries, 0 to 22698
```

```
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Unnamed: 0              22699 non-null  int64
 1   VendorID                22699 non-null  int64
 2   tpep_pickup_datetime    22699 non-null  object
 3   tpep_dropoff_datetime   22699 non-null  object
 4   passenger_count         22699 non-null  int64
 5   trip_distance           22699 non-null  float64
 6   RatecodeID              22699 non-null  int64
 7   store_and_fwd_flag      22699 non-null  object
 8   PULocationID            22699 non-null  int64
 9   DOLocationID            22699 non-null  int64
 10  payment_type            22699 non-null  int64
 11  fare_amount             22699 non-null  float64
 12  extra                   22699 non-null  float64
 13  mta_tax                 22699 non-null  float64
 14  tip_amount              22699 non-null  float64
 15  tolls_amount            22699 non-null  float64
 16  improvement_surcharge   22699 non-null  float64
 17  total_amount            22699 non-null  float64
dtypes: float64(8), int64(7), object(3)
memory usage: 3.1+ MB
```

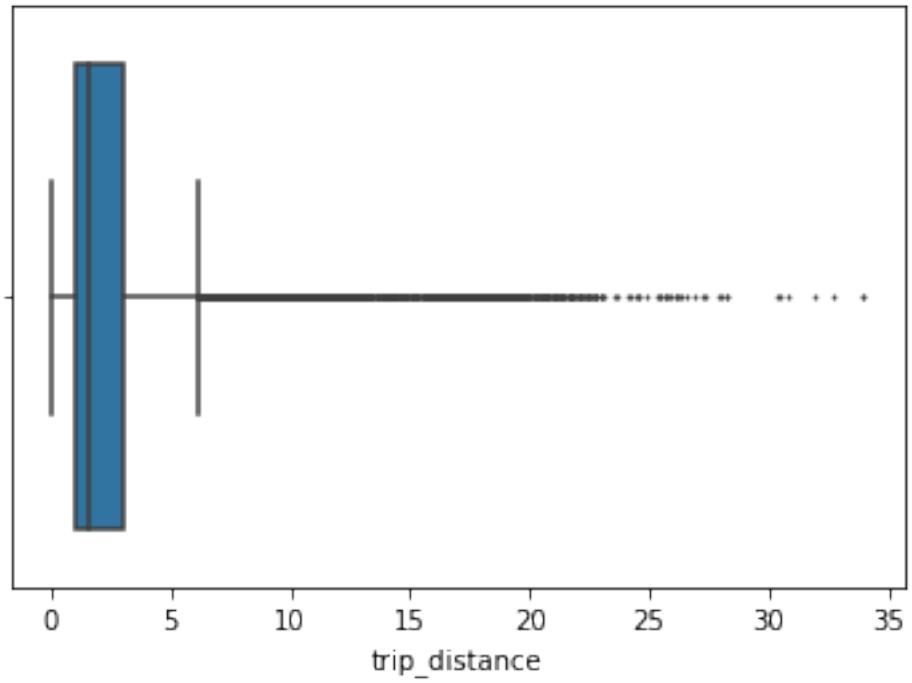
Perform a check for outliers on relevant columns sucah as trip distance and trip duration. Remember, one of the best ways to look for outliers is a box plot visualization.

Note: Remember to convert your date columns to datetime in order to derive total trip duration.

[7]:
```
df['tpep_pickup_datetime']= pd.to_datetime(df['tpep_pickup_datetime'])
df['tpep_dropoff_datetime']= pd.to_datetime(df['tpep_dropoff_datetime'])
```

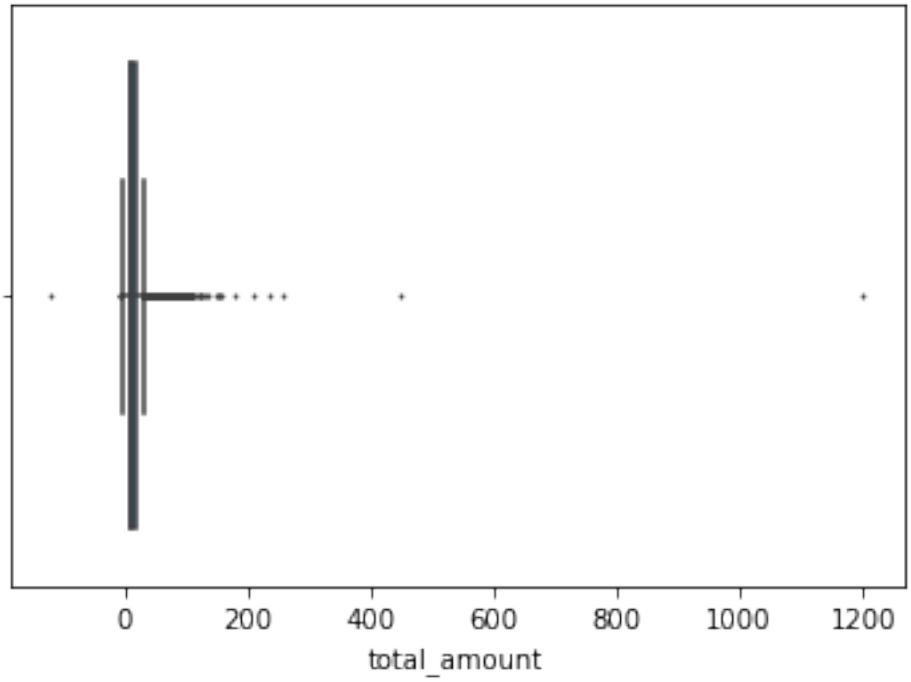
[8]:
```
sns.boxplot(data=None, x=df['trip_distance'],fliersize=1)
```

[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f35aeb08c50>
```



[9]:
```
sns.boxplot(data=None, x=df['total_amount'],fliersize=1)
```

[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7f35aea65750>



4.2 Step 2b. Assess whether dimensions and measures are correct

In Tableau, staying on the data source page, double check the data types for the applicable columns you selected on the previous step. Pay particular attention to the dimensions and measures to assure they are correct.

Review the instructions at this link to create the required Tableau visualization.

4.3 Step 2c. Select Visualization Type(s)

Select data visualization types that will help you understand and explain the data.

Now that you know which data columns you'll use, it is time to decide which data visualization makes the most sense for EDA of the TLC dataset. What type of data visualization(s) would be most helpful?

- Line graph?
- Bar chart?
- Box plot?
- Histogram?
- Heat map?
- Scatter plot?
- A geographic map?

A box plot will be helpful to determine outliers and where the bulk of the data points reside in terms of `trip_distance`*, duration and* `total_amount`

A scatter plot will be helpful to visualize the trends and patters and outliers of critical variables, such as `trip_distance` *and* `total_amount`

A bar chart will help determine average number of trips per month, weekday, weekend, etc.

4.4 paCe: Construct Stage

Consider these questions [link PACE Strategy Doc] to reflect on the Constructing stage of this task.

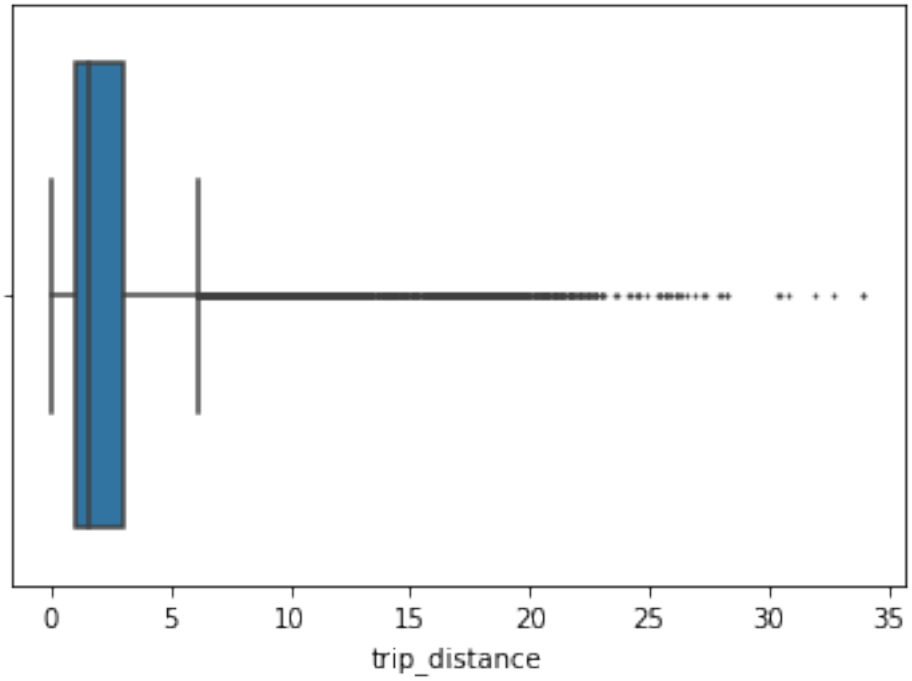
4.5 Step 3. Building visualizations

You've assessed your data, and decided on which data variables are most applicable. It's time to plot your visualization(s)!

4.5.1 Boxplots

[10]:
```
sns.boxplot(x=df['trip_distance'], fliersize=1)
```

[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f35ae872ad0>
```



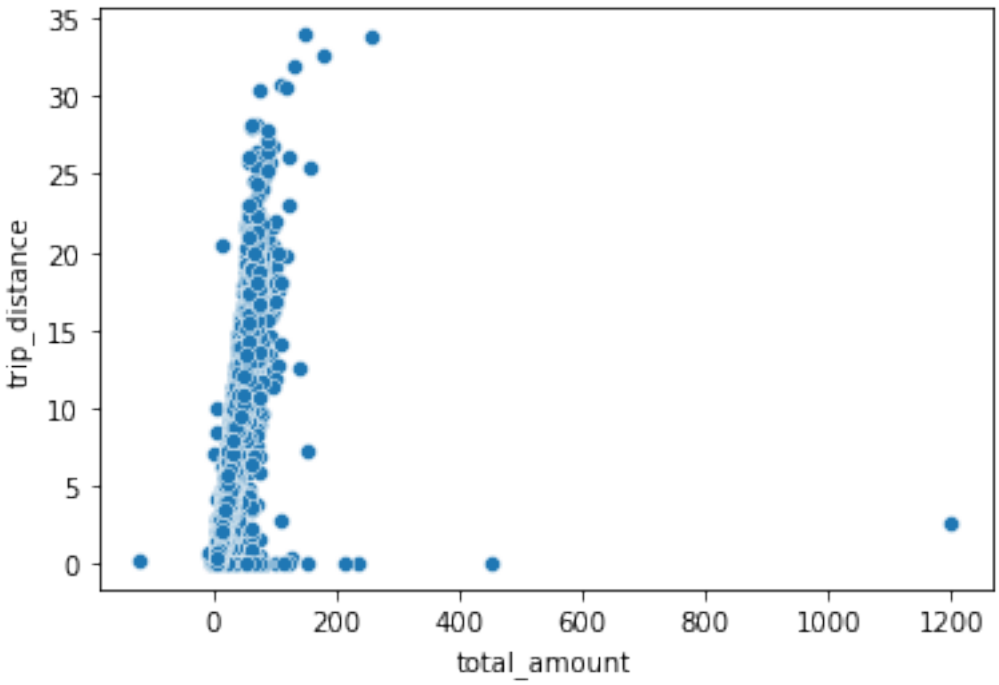
4.5.2 Scatter plot

Remove those trips with costs associated but with a trip distance = to "0."

[11]:
```
df_2= df['trip_distance'].loc[~(df==0).all(axis=1)]
```

[12]:
```
sns.scatterplot(x=df['total_amount'],y=df_2)
```

[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f35ae7e0e50>
```



You can do a scatterplot in Tableau Public as well, which can be easier to manipulate and present. If you'd like step by step instructions, you can review the following link:

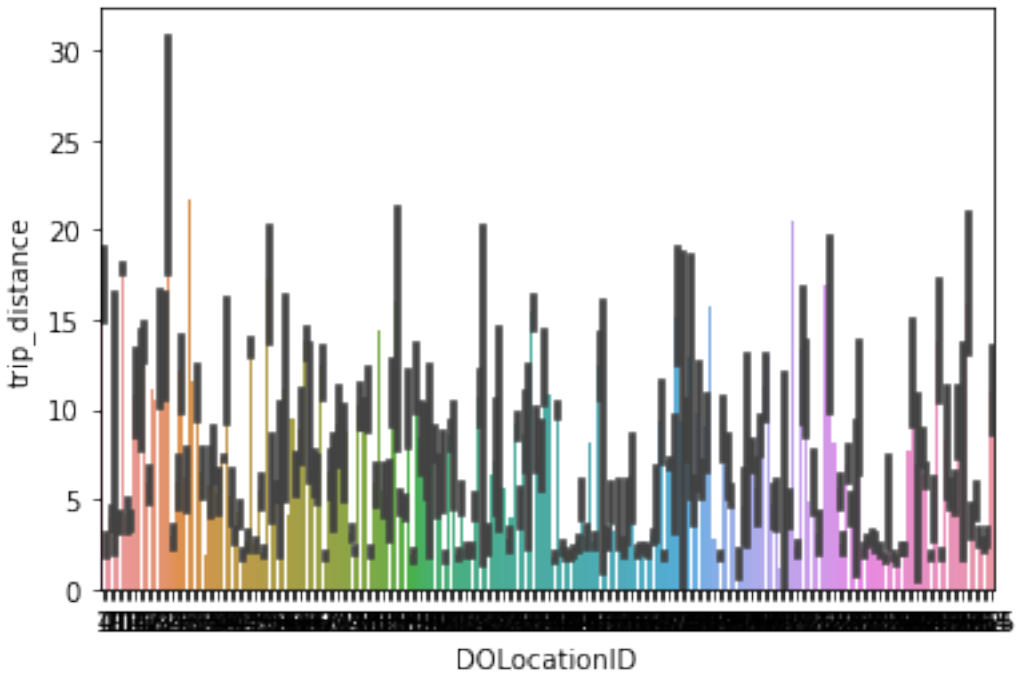
Tableau visualization guidelines

[13]:
```
df.groupby('DOLocationID')['trip_distance'].mean()
```

[13]:
```
DOLocationID
1      17.027353
4       2.436634
7       3.964944
9       9.305000
10      3.750000
          …
261     4.935897
262     2.866897
263     2.501951
264     2.928783
265    11.039000
Name: trip_distance, Length: 216, dtype: float64
```

[21]:
```
sns.barplot(data=df, x=df['DOLocationID'], y=df['trip_distance'])
```

[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f35aa2c1e90>
```



4.6 pacE: Execute Stage

Consider the questions in the Execute section of the PACE strategy document to reflect on the Execute stage of this task.

4.7 Step 4a. Results and Evaluation

Having built visualizations in Tableau and in Python, what have you learned about the dataset? What other questions have your visualizations uncovered that you should pursue?

Pro tip: Put yourself in your client's perspective, what would they want to know?

Use the following code fields to pursue any additional EDA based on the visualizations you've laready plotted. Also use the space to make sure your visualizations are clean, easily understandable, and accessible.

Ask yourself: Did you consider color, contrast, emphasis, and labeling?

I have learned how the step of removing outliers and nonsense values is essential to obtain accurate results in a project

My client would likely want to know the reason why I deleted the outliers and the benefits of applying cleanning methods to the data frame.

[22]:
```
df['trip_duration'] = (df['tpep_dropoff_datetime']-df['tpep_pickup_datetime'])
```

[23]:
```
df.head(10)
```

[23]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount
0	24870114	2	2017-03-25 08:55:43	2017-03-25 09:09:47	6	3.34	1	N	100	231	1	13.0	0.0	0.5	2.76	0.0	0.3	16.56
1	35634249	1	2017-04-11 14:53:28	2017-04-11 15:19:58	1	1.80	1	N	186	43	1	16.0	0.0	0.5	4.00	0.0	0.3	20.80
2	106203690	1	2017-12-15 07:26:56	2017-12-15 07:34:08	1	1.00	1	N	262	236	1	6.5	0.0	0.5	1.45	0.0	0.3	8.75
3	38942136	2	2017-05-07 13:17:59	2017-05-07 13:48:14	1	3.70	1	N	188	97	1	20.5	0.0	0.5	6.39	0.0	0.3	27.69
4	30841670	2	2017-04-15 23:32:20	2017-04-15 23:49:03	1	4.37	1	N	4	112	2	16.5	0.5	0.5	0.00	0.0	0.3	17.80
5	23345809	2	2017-03-25 20:34:11	2017-03-25 20:42:11	6	2.30	1	N	161	236	1	9.0	0.5	0.5	2.06	0.0	0.3	12.36
6	37660487	2	2017-05-03 19:04:09	2017-05-03 20:03:47	1	12.83	1	N	79	241	1	47.5	1.0	0.5	9.86	0.0	0.3	59.16
7	69059411	2	2017-08-15 17:41:06	2017-08-15 18:03:05	1	2.98	1	N	237	114	1	16.0	1.0	0.5	1.78	0.0	0.3	19.58
8	8433159	2	2017-02-04 16:17:07	2017-02-04 16:29:14	1	1.20	1	N	234	249	2	9.0	0.0	0.5				
9	95294817	1	2017-11-10 15:20:29	2017-11-10 15:40:55	1	1.60	1	N	239	237	1	13.0	0.0	0.5				

```
8         0.00           0.0                    0.3          9.80
9         2.75           0.0                    0.3         16.55
```

```
    trip_duration
0 0 days 00:14:04
1 0 days 00:26:30
2 0 days 00:07:12
3 0 days 00:30:15
4 0 days 00:16:43
5 0 days 00:08:00
6 0 days 00:59:38
7 0 days 00:21:59
8 0 days 00:12:07
9 0 days 00:20:26
```

4.8 Step 4b. Conclusion

Make it professional and presentable

You have visualized the data you need to share with the director now. Remember, the goal of a data visualization is for an audience member to glean the information on the chart in mere seconds.

Questions to ask yourself for reflection: Why is it important to conduct Exploratory Data Analysis? Why would we need to create a visual map of the NYC Taxi rides? Why would this be useful?

EDA is important because it improves the future development of the predictive model.

Visualizations helped me understand the distribution of the data and values that negatively affects the structure of the database, additionally, they made me realize of the missing values.