Task: Perform EDA with the help of PowerBI

Task description:

XYZ is a private company in US which would to make investments in Cab industry, but before that they want to understand the market. In case of a positive decision, XYZ has to choos between two cab companies - 'Pink' and 'Yellow'.

Data description:

There are 4 data sets that contains information on 2 cab companies:

- 'Cab_Data.csv' - details of transactions
- 'Customer_ID.csv' - mapping table that contains a unique identifier which links the customer’s demographic details
- 'Transaction_ID.csv' - mapping table that contains transaction to customer mapping and payment mode
- 'City.csv' - list of US cities, their population and number of cab users

Each file (data set) represents different aspects of the customer profile.

Preparation steps:

1. Import libraries:

In [1]:
```
import numpy as np
import pandas as pd

import os
from datetime import datetime, timedelta


#!pip install powerbiclient
from powerbiclient import Report
from io import StringIO
from ipywidgets import interact
```

2. Load data:

In [2]:
```
cab_data = pd.read_csv('Cab_Data.csv', parse_dates=['Date of Travel'])
customer_ID = pd.read_csv('Customer_ID.csv')
transaction_ID = pd.read_csv('Transaction_ID.csv')
city = pd.read_csv('City.csv')
```

3. Explore tables

```
In [3]: cab_data
```

```
Out[3]:
        Transaction ID  Date of Travel     Company            City  KM Travelled  Price Charged  Cost of Trip
0             10000011           42377    Pink Cab      ATLANTA GA         30.45         370.95      313.6350
1             10000012           42375    Pink Cab      ATLANTA GA         28.62         358.52      334.8540
2             10000013           42371    Pink Cab      ATLANTA GA          9.04         125.20       97.6320
3             10000014           42376    Pink Cab      ATLANTA GA         33.17         377.40      351.6020
4             10000015           42372    Pink Cab      ATLANTA GA          8.73         114.62       97.7760
...                ...             ...         ...             ...           ...            ...           ...
359387        10440101           43108  Yellow Cab   WASHINGTON DC          4.80          69.24       63.3600
359388        10440104           43104  Yellow Cab   WASHINGTON DC          8.40         113.75      106.8480
359389        10440105           43105  Yellow Cab   WASHINGTON DC         27.75         437.07      349.6500
359390        10440106           43105  Yellow Cab   WASHINGTON DC          8.80         146.19      114.0480
359391        10440107           43102  Yellow Cab   WASHINGTON DC         12.76         191.58      177.6192

359392 rows × 7 columns
```

In [4]:
```
customer_ID
```

Out[4]:

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536
...	...	...	...	...
49166	12490	Male	33	18713
49167	14971	Male	30	15346
49168	41414	Male	38	3960
49169	41677	Male	23	19454
49170	39761	Female	32	10128

49171 rows × 4 columns

```
In [5]: transaction_ID
```

Out[5]:

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card
...	...	...	...
440093	10440104	53286	Cash
440094	10440105	52265	Cash
440095	10440106	52175	Card
440096	10440107	52917	Card
440097	10440108	51587	Card

440098 rows × 3 columns

In [6]:
```
city
```

Out[6]:

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468
2	LOS ANGELES CA	1,595,037	144,132
3	MIAMI FL	1,339,155	17,675
4	SILICON VALLEY	1,177,609	27,247
5	ORANGE COUNTY	1,030,185	12,994
6	SAN DIEGO CA	959,307	69,995
7	PHOENIX AZ	943,999	6,133
8	DALLAS TX	942,908	22,157
9	ATLANTA GA	814,885	24,701
10	DENVER CO	754,233	12,421
11	AUSTIN TX	698,371	14,978
12	SEATTLE WA	671,238	25,063
13	TUCSON AZ	631,442	5,712
14	SAN FRANCISCO CA	629,591	213,609
15	SACRAMENTO CA	545,776	7,044
16	PITTSBURGH PA	542,085	3,643
17	WASHINGTON DC	418,859	127,001
18	NASHVILLE TN	327,225	9,270
19	BOSTON MA	248,968	80,021

Comment: It is obvious, that tables *cab_data*, *transaction_ID* and *customer_ID* should be mereged for the further analysis:

- column 'Transaction ID' joins *cab_data* and *transaction_ID*
- column 'Customer ID' joins *transaction_ID* and *customer_ID*

4. Merging tables:

```
In [7]: merged_data = pd.merge(cab_data, transaction_ID, on = 'Transaction ID', how = 'inner')
        merged_data = pd.merge(merged_data, customer_ID, on = 'Customer ID', how = 'inner')
```

Cleaning data:

1.Check data types

```
In [8]: merged_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 359391
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   Transaction ID      359392 non-null  int64  
 1   Date of Travel      359392 non-null  object 
 2   Company             359392 non-null  object 
 3   City                359392 non-null  object 
 4   KM Travelled        359392 non-null  float64
 5   Price Charged       359392 non-null  float64
 6   Cost of Trip        359392 non-null  float64
 7   Customer ID         359392 non-null  int64  
 8   Payment_Mode        359392 non-null  object 
 9   Gender              359392 non-null  object 
 10  Age                 359392 non-null  int64  
 11  Income (USD/Month)  359392 non-null  int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 35.6+ MB
```

All data types are correct expet of the *Date of Travel*. This column still reflect the quantity of days from 1900 till today.

Here is how to change it:

In [9]:
```python
date = pd.to_datetime('1900-01-01') #count from this date
merged_data['Date of Travel'] = pd.to_numeric(merged_data['Date of Travel']) #covert column to numeric
merged_data['Date of Travel'] = merged_data['Date of Travel'].apply(lambda x:  date + timedelta(x)) #calculate dates
```

In [10]:
```python
merged_data
```

Out[10]:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)
0	10000011	2016-01-10	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813
1	10351127	2018-07-23	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813
2	10412921	2018-11-25	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813
3	10000012	2016-01-08	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237
4	10320494	2018-04-23	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237
...	...	...	...	...	...	...	...	...	...	...	...	...
359387	10439790	2018-01-09	Yellow Cab	SEATTLE WA	16.66	261.18	213.9144	38520	Card	Female	42	19417
359388	10439799	2018-01-05	Yellow Cab	SILICON VALLEY	13.72	277.97	172.8720	12490	Cash	Male	33	18713
359389	10439838	2018-01-06	Yellow Cab	TUCSON AZ	19.00	303.77	232.5600	41414	Card	Male	38	3960
359390	10439840	2018-01-08	Yellow Cab	TUCSON AZ	5.60	92.42	70.5600	41677	Cash	Male	23	19454
359391	10439846	2018-01-06	Yellow Cab	TUCSON AZ	13.30	244.65	180.3480	39761	Card	Female	32	10128

359392 rows × 12 columns

2. Check and remove duplicates

In [11]:
```
duplicates_indecies = np.where(merged_data.duplicated() == True)[0]

if len(duplicates_indecies) == 0:
    print("There is no duplicates")
else:
    print("Indecies:", duplicates_indecies)
```

```
There is no duplicates
```

3. Check and remove N/A

In [12]:
```
bool_na = merged_data.isnull().values.any()

if bool_na == True:
    print("There are missed values in dataset")
else:
    print("There are no missed values")
```

```
There are no missed values
```

Exploratory Data Analysis with Power BI

Download merged data as .csv to visualize in Power BI:

In [13]:
```
import os
os.makedirs('data/', exist_ok = True)
merged_data.to_csv('data/merged.csv')
city.to_csv('data/city.csv')
```

Setting parametrs for PowerBI:

In [14]:
```
# Import the DeviceCodeLoginAuthentication class to authenticate against Power BI
from powerbiclient.authentication import DeviceCodeLoginAuthentication

# Initiate device authentication
device_auth = DeviceCodeLoginAuthentication()
access_token = device_auth.get_access_token()
```

```
Performing interactive authentication. Please follow the instructions on the terminal.
 To sign in, use a web browser to open the page https://microsoft.com/devicelogin (https://microsoft.com/devicelogin) and enter t
he code DL7GW7SYJ to authenticate.
You have logged in.
Interactive authentication successfully completed.
```

Report data:

In [15]:
```
group_id = "30c9ae19-4d19-4b04-b186-602205d30f72"
report_id = "eb2d3182-2996-4593-ab85-941ffc05a76b"
```

Load report:

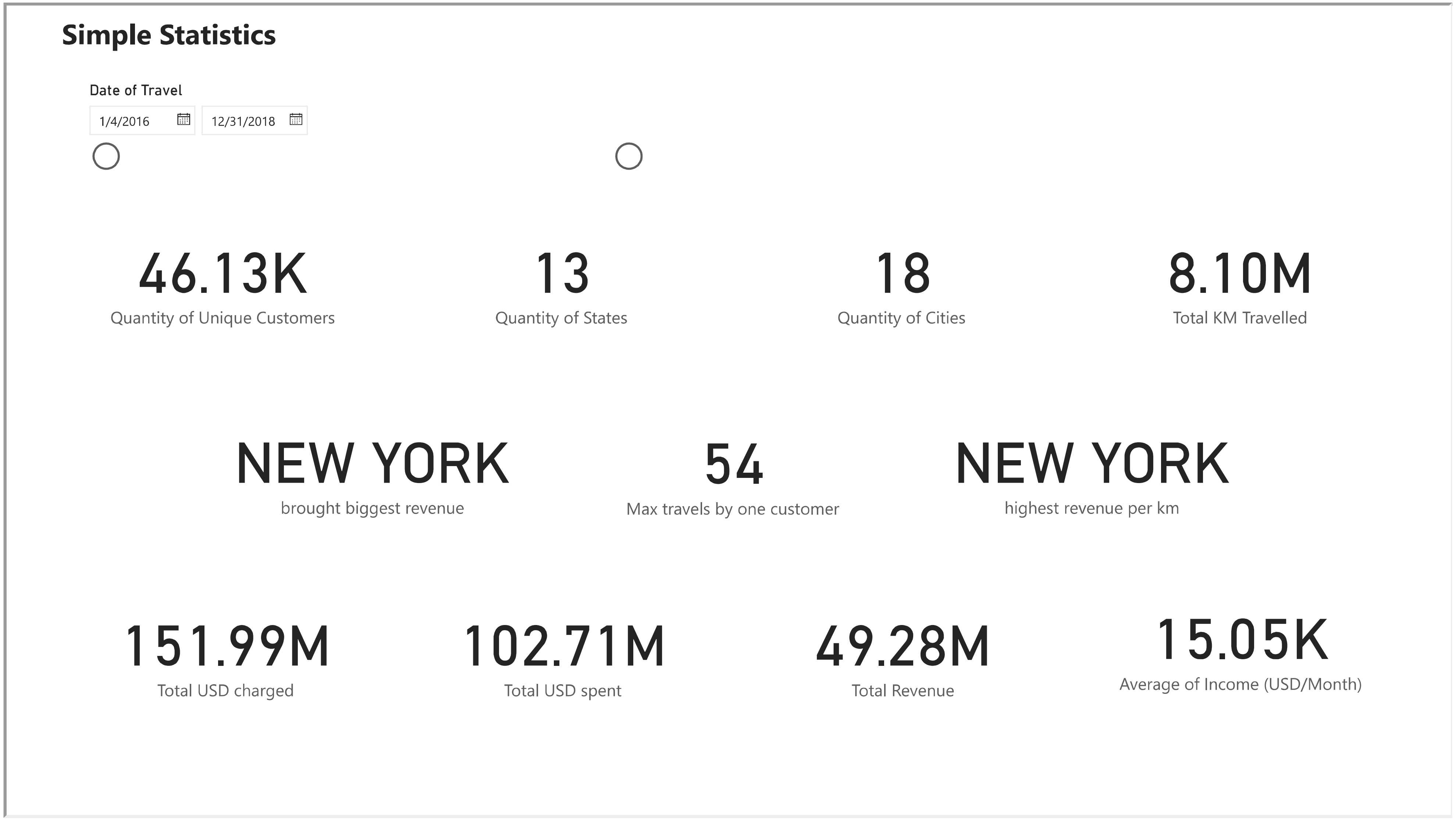
In [16]:
```
report = Report(group_id=group_id, report_id=report_id, access_token = access_token)
```

Click on visuals and walk through tabs.

Comments are in presentation.

- Simple statistics

```
In [17]: report
```



Simple Statistics

Date of Travel

1/4/2016 12/31/2018

46.13K
Quantity of Unique Customers

13
Quantity of States

18
Quantity of Cities

8.10M
Total KM Travelled

NEW YORK
brought biggest revenue

54
Max travels by one customer

NEW YORK
highest revenue per km

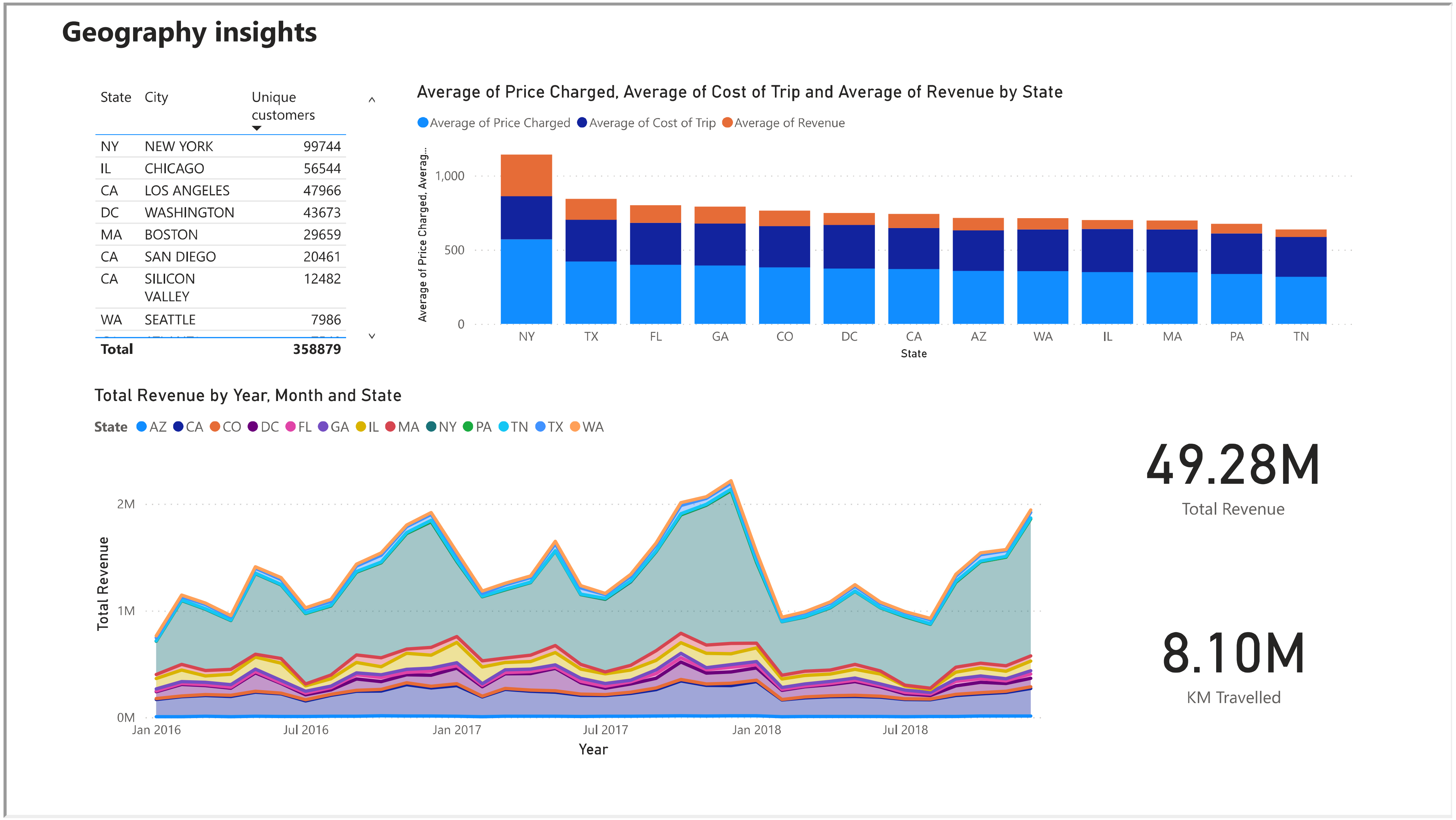
151.99M
Total USD charged

102.71M
Total USD spent

49.28M
Total Revenue

15.05K
Average of Income (USD/Month)

```
In [18]: report
```



Geography insights

State	City	Unique customers
NY	NEW YORK	99744
IL	CHICAGO	56544
CA	LOS ANGELES	47966
DC	WASHINGTON	43673
MA	BOSTON	29659
CA	SAN DIEGO	20461
CA	SILICON VALLEY	12482
WA	SEATTLE	7986
Total		**358879**

Average of Price Charged, Average of Cost of Trip and Average of Revenue by State

● Average of Price Charged ● Average of Cost of Trip ● Average of Revenue
Average of Price Charged, Averag...
1,000
500
0
NY
TX
FL
GA
CO
DC
CA
AZ
WA
IL
MA
PA
TN
State

Total Revenue by Year, Month and State

State ● AZ ● CA ● CO ● DC ● FL ● GA ● IL ● MA ● NY ● PA ● TN ● TX ● WA
Total Revenue
2M
1M
0M
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Jul 2018
Year

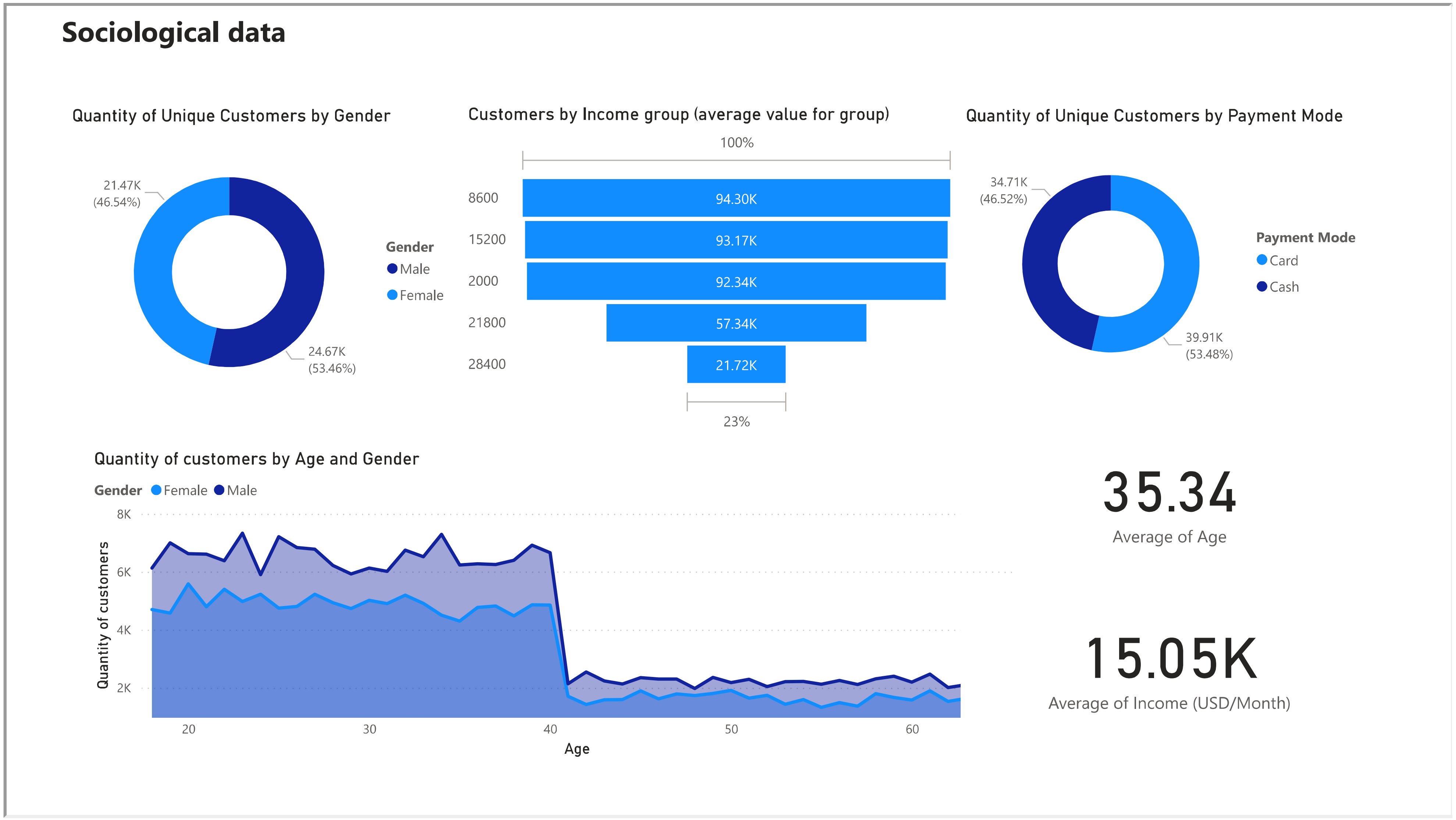
49.28M

Total Revenue

8.10M

KM Travelled

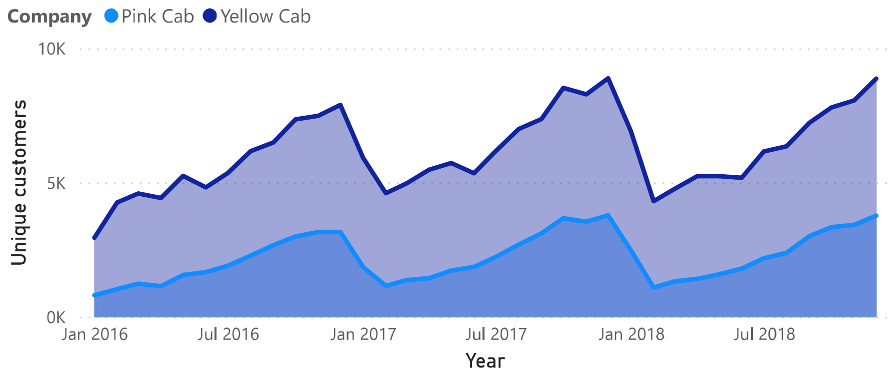
```
In [19]: report
```


Sociological data
Quantity of Unique Customers by Gender
21.47K (46.54%)
24.67K (53.46%)
Gender
● Male
● Female
Customers by Income group (average value for group)
100%
8600
15200
2000
21800
28400
94.30K
93.17K
92.34K
57.34K
21.72K
23%
Quantity of Unique Customers by Payment Mode
34.71K (46.52%)
39.91K (53.48%)
Payment Mode
● Card
● Cash
Quantity of customers by Age and Gender
Gender ● Female ● Male
8K
6K
4K
2K
Quantity of customers
20
30
40
50
60
Age
35.34
Average of Age
15.05K
Average of Income (USD/Month)

```
In [20]: report
```

Data by Company

Unique customers by Year, Month and Company



8.10M

KM Travelled

46.13K

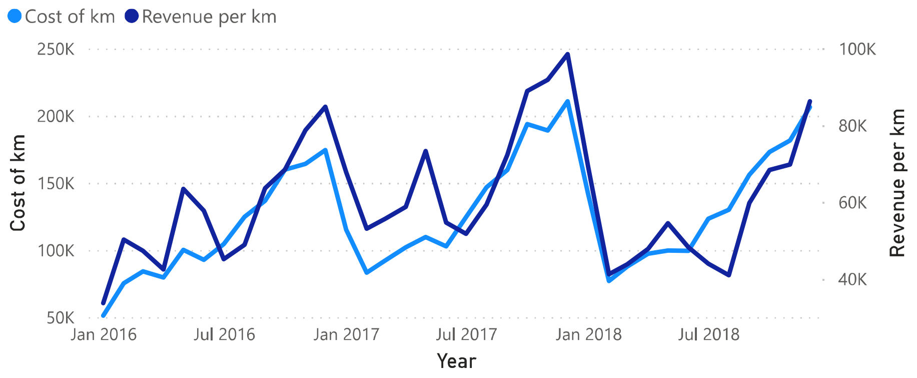
Unique customers

Company

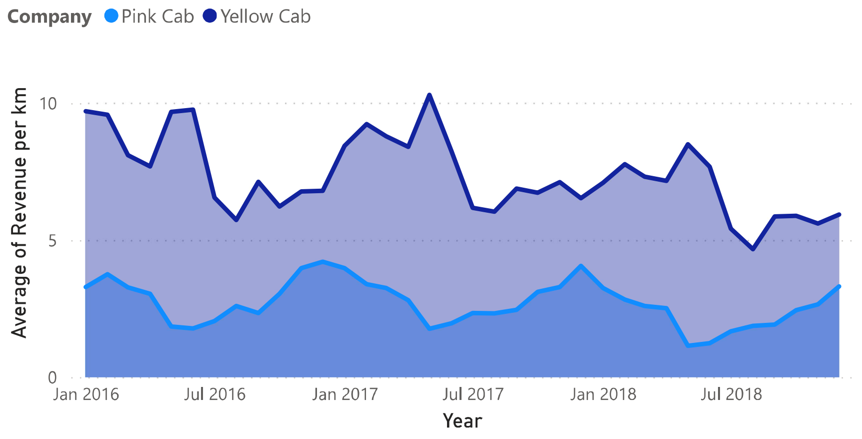
☐ Pink Cab

☐ Yellow Cab

Cost of km and Revenue per km by Year and Month



Average of Revenue per km by Year, Month and Company



19.00

Min of Cost of Trip

691.20

Max of Cost of Trip