
# Exploratory Data Analysis

## G2M insight for Cab Investment firm

**Name:** Roger Burek-Bors
**Location:** Warsaw, Poland
**Team:** Data and Analytics
**Date:** March 11, 2021

# Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

# Executive Summary

- Descriptive, correlation and contextual analysis were made to help XYZ firm in identification of the right cab company to make investment.
- 6 hypothesis were constructed to gain knowledge on the subject and formulate final recommendation. Main topics included:
  - patterns among cities when it comes to population and cab users
  - market size and access
  - cost, income and profit per each company
  - trip patterns in particular cities per each company
  - payment method preferences
  - economic trends and seasonality
- I recommended XYZ firm to invest in Yellow Cab based on evidence collected and outcome of my analysis.

# Problem Statement

*XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry and as per their Go-to-Market (G2M) strategy they want to understand the market before taking final decision.*

**Objective:**

XYZ is interested in actionable insights to help them identify the right company to make their investment.

# Approach

1. Verification of data sets provided by XYZ, features descriptive analysis and combining into joint data frame.
2. Formulation of hypothesis.
3. Performing EDA over formulated hypothesis (correlation and contextual analysis).
4. Concluding EDA and writing recommendation.

# EDA – provided data sets

```
            City  Population   Users
0     NEW YORK NY     8405837  302149
1      CHICAGO IL     1955130  164468
2  LOS ANGELES CA     1595037  144132
3        MIAMI FL     1339155   17675
4  SILICON VALLEY     1177609   27247
Size of 'Cities': (20, 3)

   Transaction ID Date of Travel   Company        City  KM Travelled  \
0        10000011     2016-01-08  Pink Cab  ATLANTA GA         30.45
1        10000012     2016-01-06  Pink Cab  ATLANTA GA         28.62
2        10000013     2016-01-02  Pink Cab  ATLANTA GA          9.04
3        10000014     2016-01-07  Pink Cab  ATLANTA GA         33.17
4        10000015     2016-01-03  Pink Cab  ATLANTA GA          8.73

   Price Charged  Cost of Trip
0         370.95       313.635
1         358.52       334.854
2         125.20        97.632
3         377.40       351.602
4         114.62        97.776
Size of 'Cabs': (359392, 7)

   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
Size of 'Customers': (49171, 4)

   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
Size of 'Transactions': (440098, 3)
```

- XYZ provided 4 data sets:
  - City.csv
  - Cab_Data.csv
  - Customer_ID.csv
  - Transaction_IDta.csv
- There was total of 17 features in provided data sets.
- 4 features repeated across data sets and as master features were used to integrate data into one data frame.
- There was no data duplication while checking on master features in each data sets:
  - ‚City' in City.csv
  - ‚Transaction ID' in Cab_Data.csv
  - ‚Customer ID' in Customer_ID.csv
  - ‚Transaction ID' in Transaction_IDta.csv
- There was no NaN values in provided data.

# EDA – created joint data frame

**Joint DF:**

- Time period of the data: 2016-01-02 to 2018-12-31.
- Total features: 13
- Total data points: 359,392

| | Transaction ID | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip | Cost per 1km | Customer ID | Payment_Mode | Gender | Age | Income (USD/Month) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 10000011 | 2016-01-08 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.635 | 10.3 | 29290 | Card | Male | 28 | 10813 |
| 1 | 10000012 | 2016-01-06 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.854 | 11.7 | 27703 | Card | Male | 27 | 9237 |
| 2 | 10000013 | 2016-01-02 | Pink Cab | ATLANTA GA | 9.04 | 125.20 | 97.632 | 10.8 | 28712 | Cash | Male | 53 | 11242 |
| 3 | 10000014 | 2016-01-07 | Pink Cab | ATLANTA GA | 33.17 | 377.40 | 351.602 | 10.6 | 28020 | Cash | Male | 23 | 23327 |
| 4 | 10000015 | 2016-01-03 | Pink Cab | ATLANTA GA | 8.73 | 114.62 | 97.776 | 11.2 | 27182 | Card | Male | 33 | 8536 |

**Supporting DF:**

- Statistical information about cities
- Total features: 4
- Total data points: 20

| | City | Population | Users | Users_100k |
|---|---|---|---|---|
| 0 | NEW YORK NY | 8405837 | 302149 | 3594.514145 |
| 1 | CHICAGO IL | 1955130 | 164468 | 8412.126048 |
| 2 | LOS ANGELES CA | 1595037 | 144132 | 9036.279409 |
| 3 | MIAMI FL | 1339155 | 17675 | 1319.862152 |
| 4 | SILICON VALLEY | 1177609 | 27247 | 2313.756094 |

# EDA – hypothsis

- Hypothesis no. 1: bigger cities have more cab users.
- Hypothesis no. 2: bigger cities provide more profit and equal opportunities to cab companies.
- Hypothesis no. 3: costs for cab companies are lower in big cities
- Hypothesis no. 4: drivers in smaller cities have shorter rides than in bigger ones.
- Hypothesis no. 5: customers like to pay by cash.
- Hypothesis no. 6: cab market is stable business and there is no seasonality in the demand.

# EDA - Hypothesis no. 1
# „cab users”

- Size of the city does not determine quantity of cab users.
- There are cities like San Francisco, Boston or Washington D.C. to have more cab users per 100k citizens than much bigger cities.
- NYC has relatively low number of cab users when calculated per 100k of population. It might be due to developed network of public transpo-rtation.
- To make best investment decision we should analyse more closely market size and market share between Yellow and Pink Can companies in: NYC, San Francisco, Boston, Washington D.S., Loas Angeles and Chicago.

**Ranking of cities – total no of users**

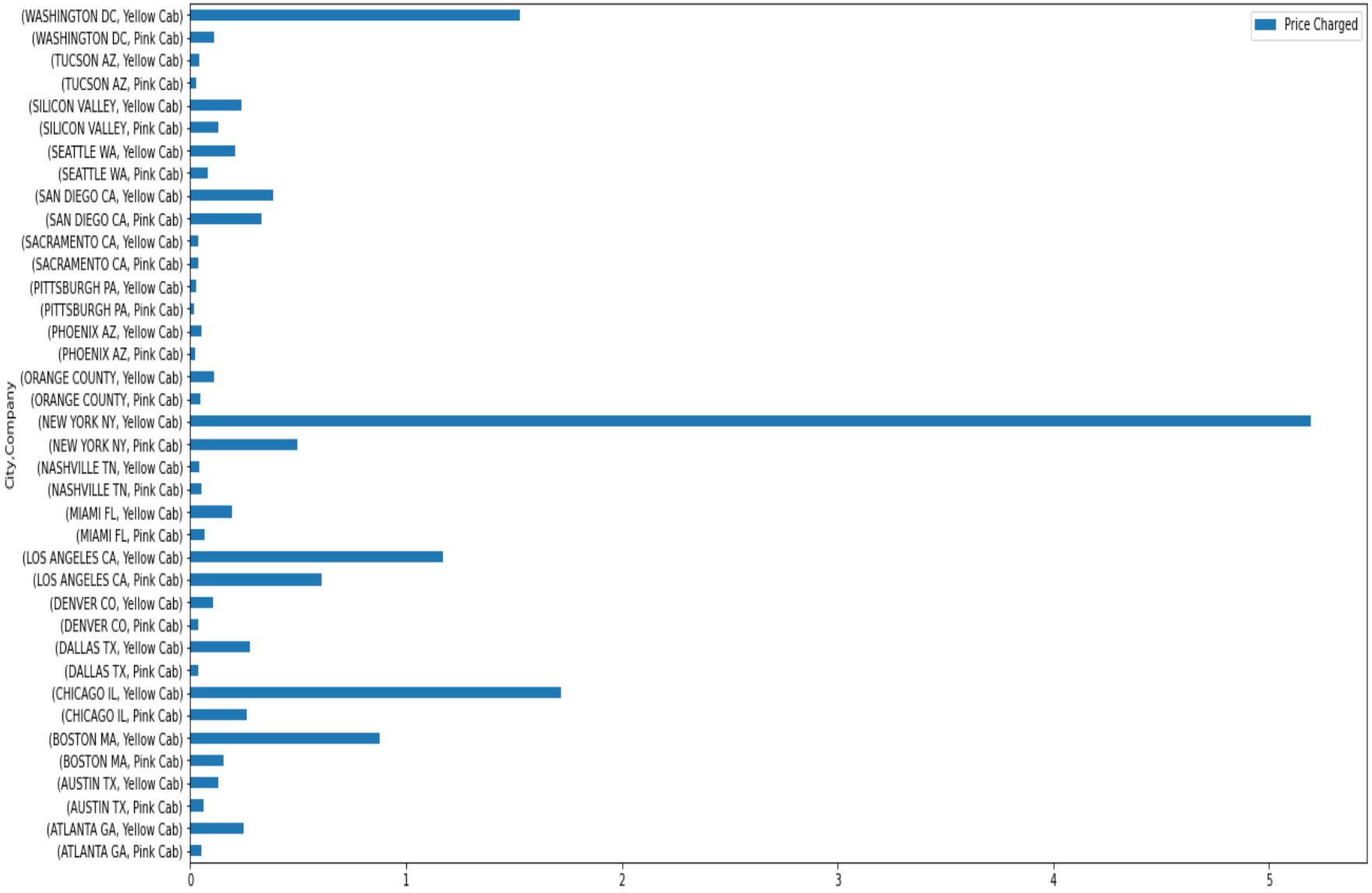
| City | Population | Users | Users_100k |
|---|---|---|---|
| NEW YORK NY | 8405837 | 302149 | 3594.514145 |
| SAN FRANCISCO CA | 629591 | 213609 | 33928.216890 |
| CHICAGO IL | 1955130 | 164468 | 8412.126048 |
| LOS ANGELES CA | 1595037 | 144132 | 9036.279409 |
| WASHINGTON DC | 418859 | 127001 | 30320.704581 |
| BOSTON MA | 248968 | 80021 | 32141.078372 |
| SAN DIEGO CA | 959307 | 69995 | 7296.412931 |
| SILICON VALLEY | 1177609 | 27247 | 2313.756094 |
| SEATTLE WA | 671238 | 25063 | 3733.847011 |
| ATLANTA GA | 814885 | 24701 | 3031.225265 |
| DALLAS TX | 942908 | 22157 | 2349.858099 |
| MIAMI FL | 1339155 | 17675 | 1319.862152 |
| AUSTIN TX | 698371 | 14978 | 2144.705321 |
| ORANGE COUNTY | 1030185 | 12994 | 1261.326849 |
| DENVER CO | 754233 | 12421 | 1646.838576 |
| NASHVILLE TN | 327225 | 9270 | 2832.913133 |
| SACRAMENTO CA | 545776 | 7044 | 1290.639383 |
| PHOENIX AZ | 943999 | 6133 | 649.682892 |
| TUCSON AZ | 631442 | 5712 | 904.596147 |
| PITTSBURGH PA | 542085 | 3643 | 672.034828 |

**Ranking of cities – users per 100k**

| City | Population | Users | Users_100k |
|---|---|---|---|
| SAN FRANCISCO CA | 629591 | 213609 | 33928.216890 |
| BOSTON MA | 248968 | 80021 | 32141.078372 |
| WASHINGTON DC | 418859 | 127001 | 30320.704581 |
| LOS ANGELES CA | 1595037 | 144132 | 9036.279409 |
| CHICAGO IL | 1955130 | 164468 | 8412.126048 |
| SAN DIEGO CA | 959307 | 69995 | 7296.412931 |
| SEATTLE WA | 671238 | 25063 | 3733.847011 |
| NEW YORK NY | 8405837 | 302149 | 3594.514145 |
| ATLANTA GA | 814885 | 24701 | 3031.225265 |
| NASHVILLE TN | 327225 | 9270 | 2832.913133 |
| DALLAS TX | 942908 | 22157 | 2349.858099 |
| SILICON VALLEY | 1177609 | 27247 | 2313.756094 |
| AUSTIN TX | 698371 | 14978 | 2144.705321 |
| DENVER CO | 754233 | 12421 | 1646.838576 |
| MIAMI FL | 1339155 | 17675 | 1319.862152 |
| SACRAMENTO CA | 545776 | 7044 | 1290.639383 |
| ORANGE COUNTY | 1030185 | 12994 | 1261.326849 |
| TUCSON AZ | 631442 | 5712 | 904.596147 |
| PITTSBURGH PA | 542085 | 3643 | 672.034828 |
| PHOENIX AZ | 943999 | 6133 | 649.682892 |

# EDA - Hypothesis no. 2
# „size of market”

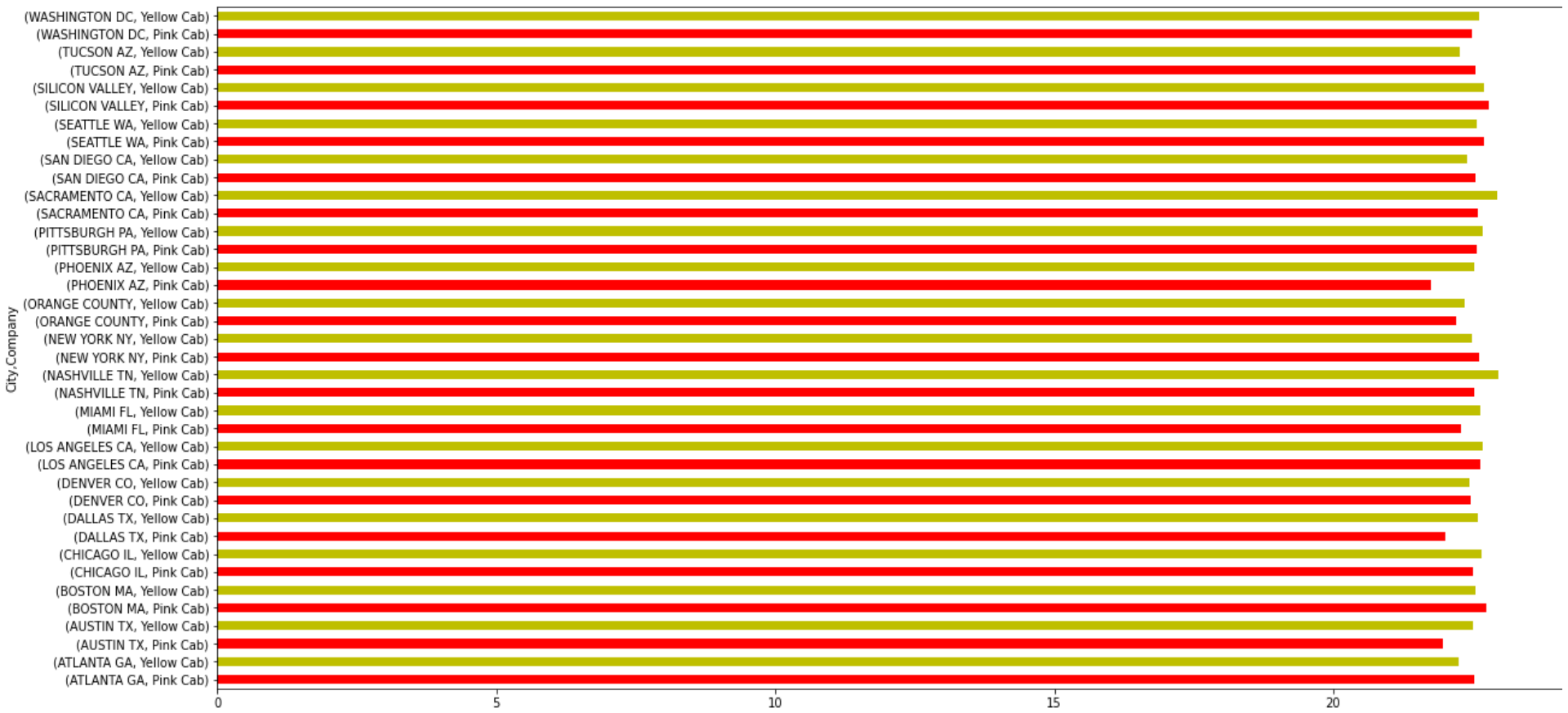
- Yellow Cub has superior market position in:
  - NYC
  - Washington D.C.
  - Miami
  - Los Angeles
  - Dallas
  - Chicago
  - Boston
  - Atlanta
- Pink Cab performs better only in Nashville.
- Equal market size is in Sacramento.
- Market value of 20 cities:
  - Total: 152 mln $ (100%)
  - Yellow Cab: 125 mln $ (82%)
  - Pink Cab: 27 mln $ (18%)

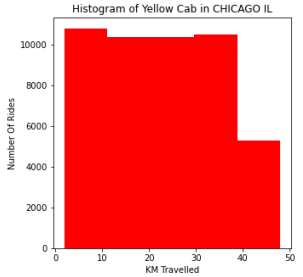


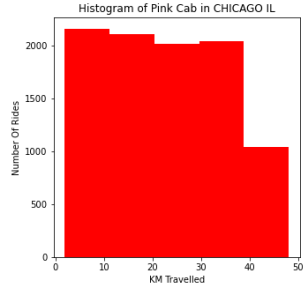
# EDA - Hypothesis no. 3
# „cost, income & profit”

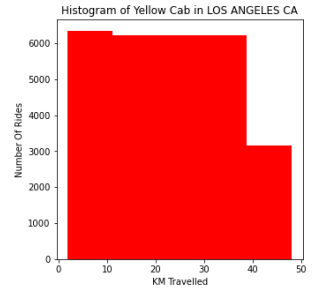
- Mean cost per trip (= all expenses like petrol, car amortisation, driver salary, dispatching, etc.):
  - Yellow Cab – 297.92 $
  - Pink Cab – 238.15 $
- Mean income per trip (= money earned on trip, excluding tip for driver):
  - Yellow Cab – 458.18 $
  - Pink Cab – 310.80 $
- Mean profit per trip:
  - Yellow Cab – 160.25 $ (35%)
  - Pink Cab – 72.65 $ (23%)

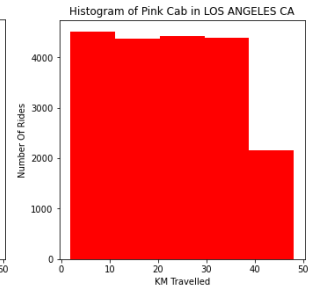
# EDA - Hypothesis no. 4
# „parameters of trips”

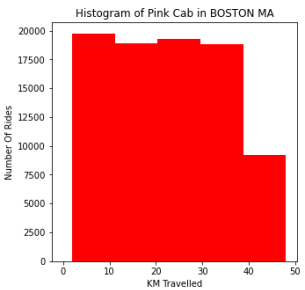


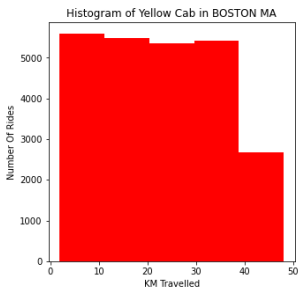



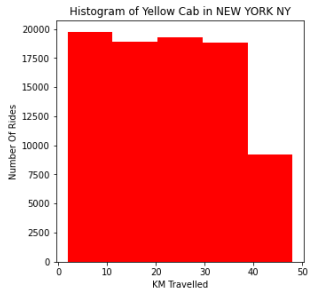



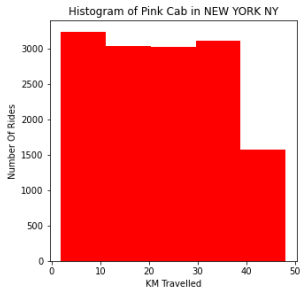



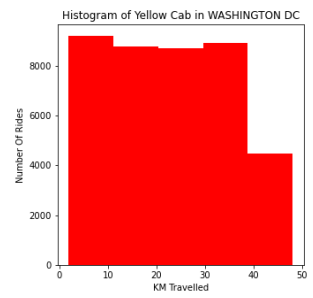



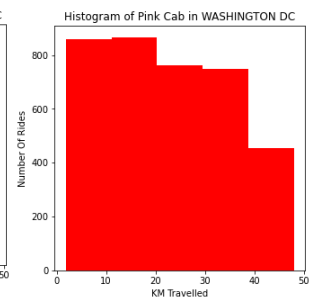



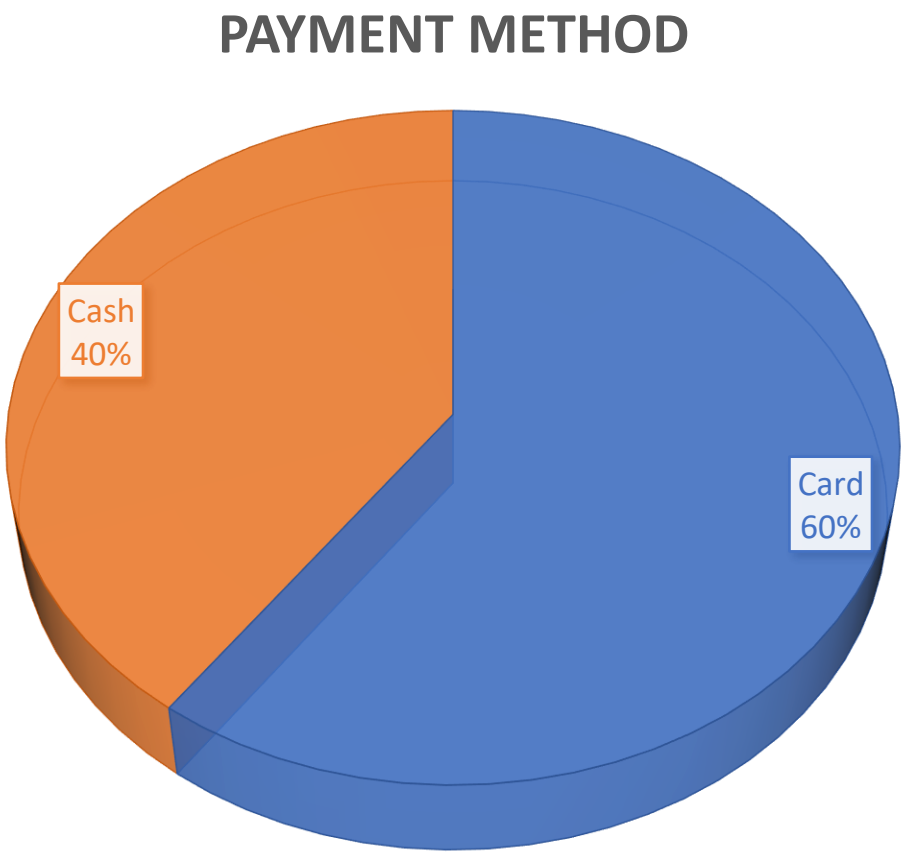



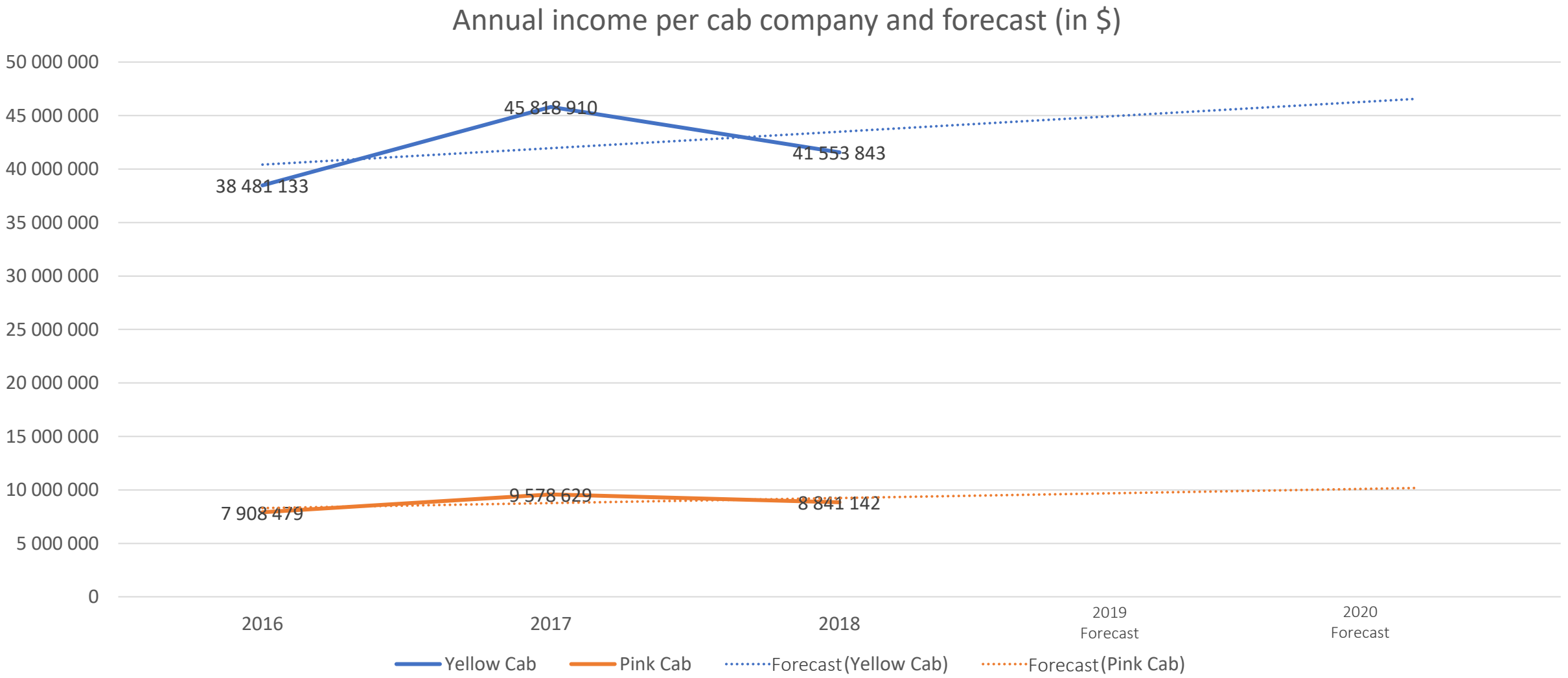



# EDA - Hypothesis no. 5 „payment methods”

- Nowadays customers likes to pay by cards.
- Quantity and value analysis indicate that card payment method scores 60%, against cash payment – 40%,
- Both companies must invest in card terminals in cabs since there is quite similar preference among their customer base.



# EDA - Hypothesis no. 6
# „annual income and trends”



# EDA Summary

- Provided data sets were verified for duplication and missing values. Most promising features were selected and few features added that broaden the expertise.
- Joint data frame was created for future modelling.
- 6 hypothesis were constructed and verified in correlation and contextual analysis.
- Final recommendation was provided based on outcome of analysis.

# Recommendations (1)

- To choose between particular cab company we should understand the market first. Market situation in enumerated below cities should be crucial to take investment decision. NYC provides the largest and outstanding market worth more than 56 million USD in analyzed period (2016-18). Cities that provided cab market bigger than 10 million USD in analyzed period (2016-18) are:
  - Chicago IL – 20 million USD
  - Los Angeles – 18 million USD
  - Washington D.C. – 16 million USD
  - Boston MA – 10 million USD
- Another important decision factor is number of users in particular city:
  - In global:
    - Boston MA – 80,021
    - Washington D.C. – 127,001
    - Los Angeles – 144,132
    - Chicago IL – 164,468
    - San Francisco – 213,609
    - NYC – 302,149
  - And per 100k citizens:
    - NYC – 3,595
    - Chicago IL – 8,412
    - Los Angeles – 9,036
    - Washington D.C. – 30,321
    - Boston MA – 32,141
    - San Francisco – 33,928

# Recommendations (2)

- Therefore, investors should choose cab company based on data and metrics for those cities.
- Yellow Cab took more than 75% market when it comes to revenue in NY City, Chicago, Los Angeles, Washington and Boston.
- Comparison of mean profit generated from all performed trip is in favor for Yellow Cab:
  - Yellow Cab – 160.25 $ (35% of income)
  - Pink Cab – 72.65 $ (23% of income)
- Both companies have similar profit trends within the year, and also while comparing year-to-year. Seasonality in demand grows bigger in second part of the year.
- Investing in Yellow Can require founds for equipping cabs with card terminals since card payments are the most popular among customers.

**Final recommendation:**

I would recommend to investors Yellow Cab as they can count on more return on investment.

# Thank You