Data Glacier

Your Deep Learning Partner

G2M Case Study

21-Jan-2023

Problem Statement

- XYZ is a private equity 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.

- Objective: Provide actionable insights to help XYZ firm identify the best company in which to invest.

The analysis can be broadly divided into the following parts:

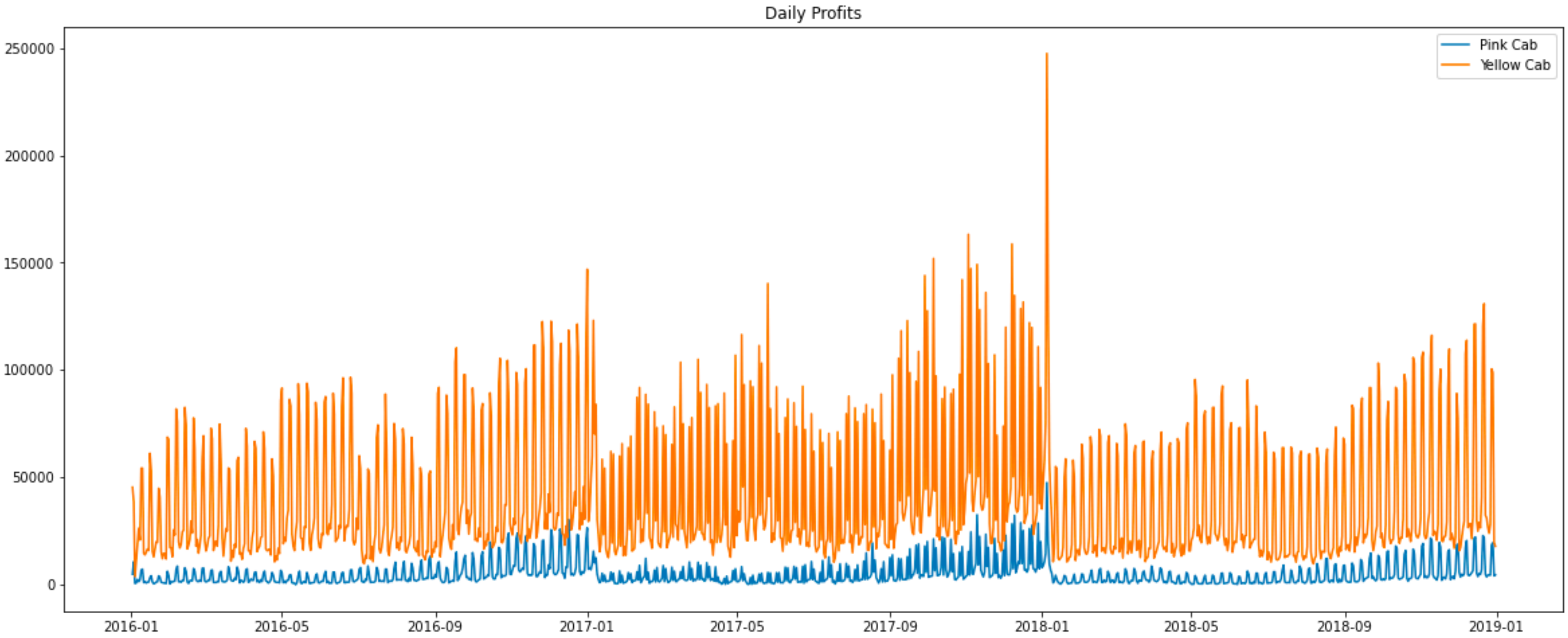
- Cleaning data and building fresh database
- Analysing monthly sales counts and profits
- Analysing customer profile for each company
- Forecasting profits and recommendations

Exploratory Data Analysis (EDA)

- Extract data from .csv and join tables as required
- Create new data frame with required (16) features: ['Transaction ID', 'Date of Travel', 'Company', 'City', 'KM Travelled', 'Price Charged', 'Cost of Trip', 'Customer ID', ‘Payment_Mode', 'Gender', 'Age', 'Income (USD/Month)', 'Profit', ‘Population (city)’, ‘Users (city)’, ‘Proportion of cab users (by city)’]
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total observations: 359,392
- Profit per trip can be found as: Price Charged - Cost of Trip.
- Proportion of users can be found as: Users/Population.
- Since span of data is only 3 years, it would be more useful to examine smaller slices, such as months or weeks.

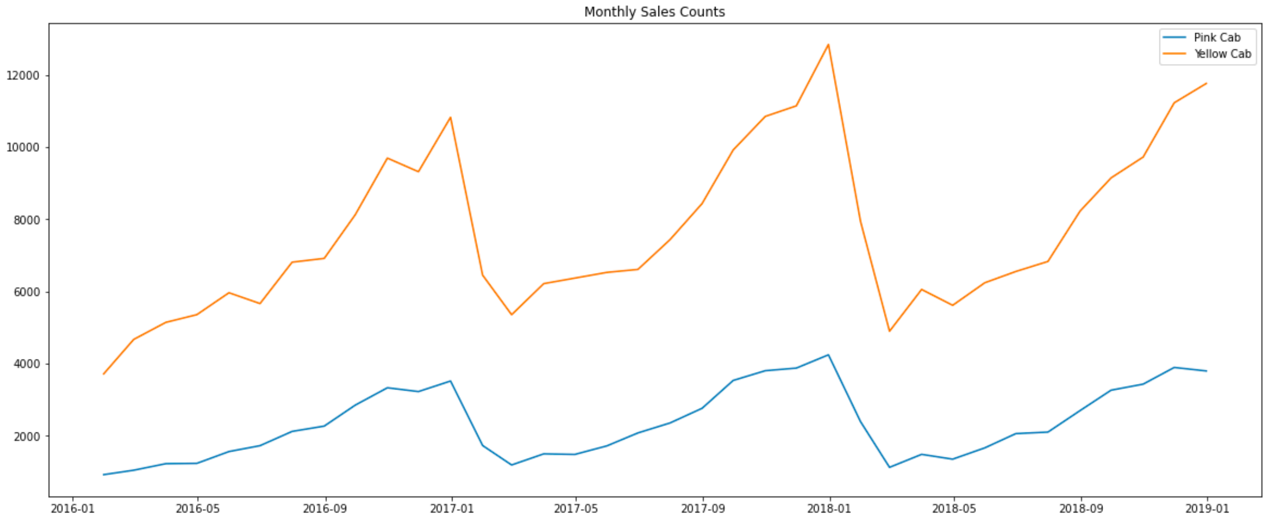
EDA Visualisations

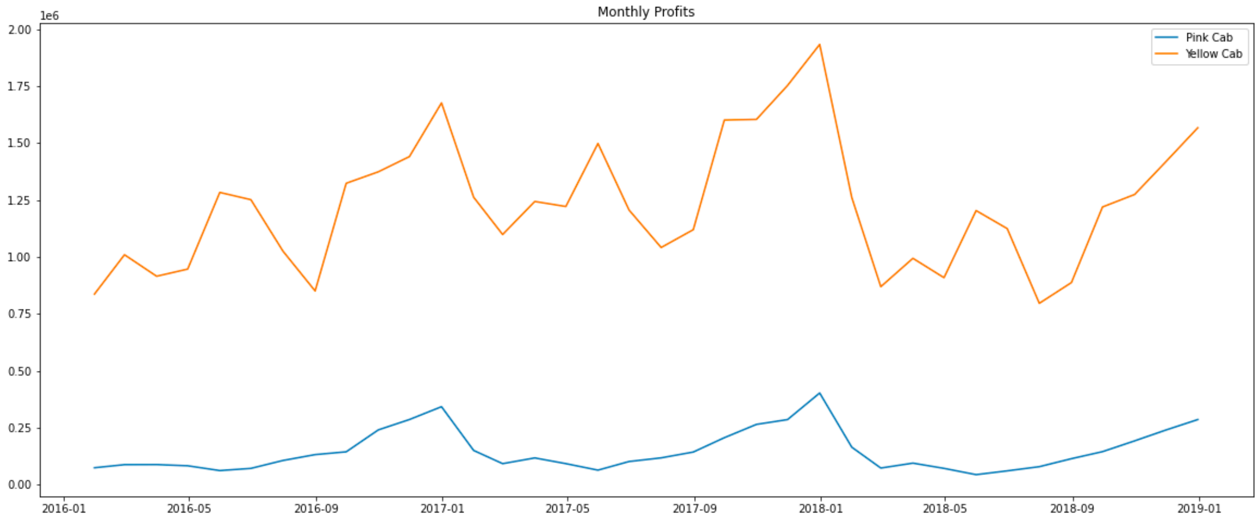
From an investment perspective, it is common to analyse the profits first. Shown below are the daily profits reported by each Cab company. As we can see, the Yellow cab company has enjoyed higher profits for the duration of this 3 year time window. There is a vague trace of some kind of seasonality.



EDA Visualisations - Monthly

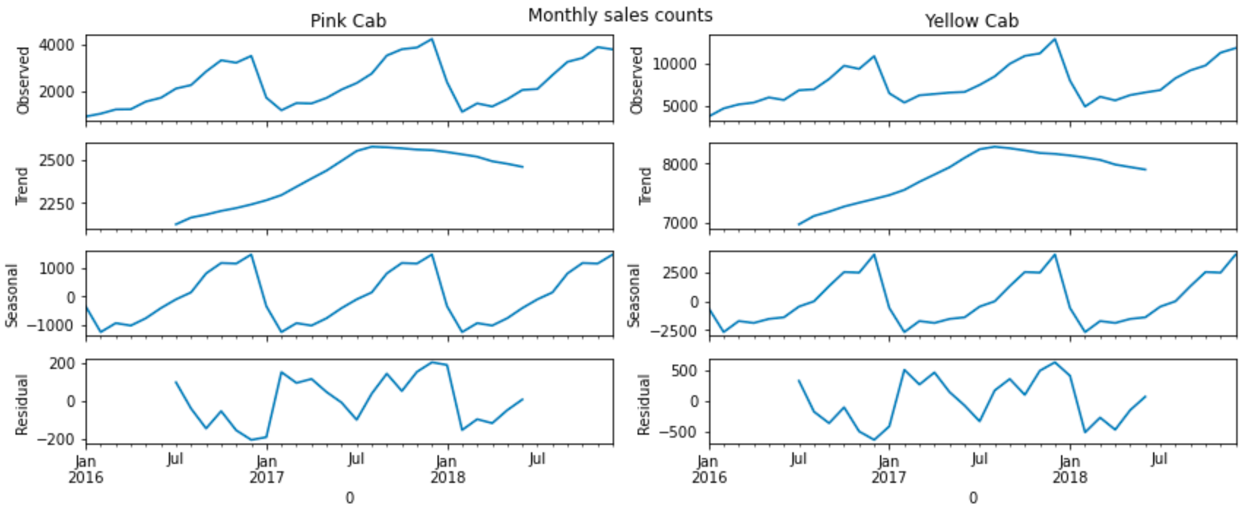
- The plots below show monthly sales counts (left) and monthly profits (right).
- It is clear that Yellow Cab company is leading in both sales count and profits.
- There is a seasonal trend that is easier to spot when looking at the monthly sales counts.
- The number of sales (for both companies) seems to peak around the end of the year (month 12).
- This coincides with the holiday period in the USA (location of data).
- This seasonality is easier to spot when examining monthly data.

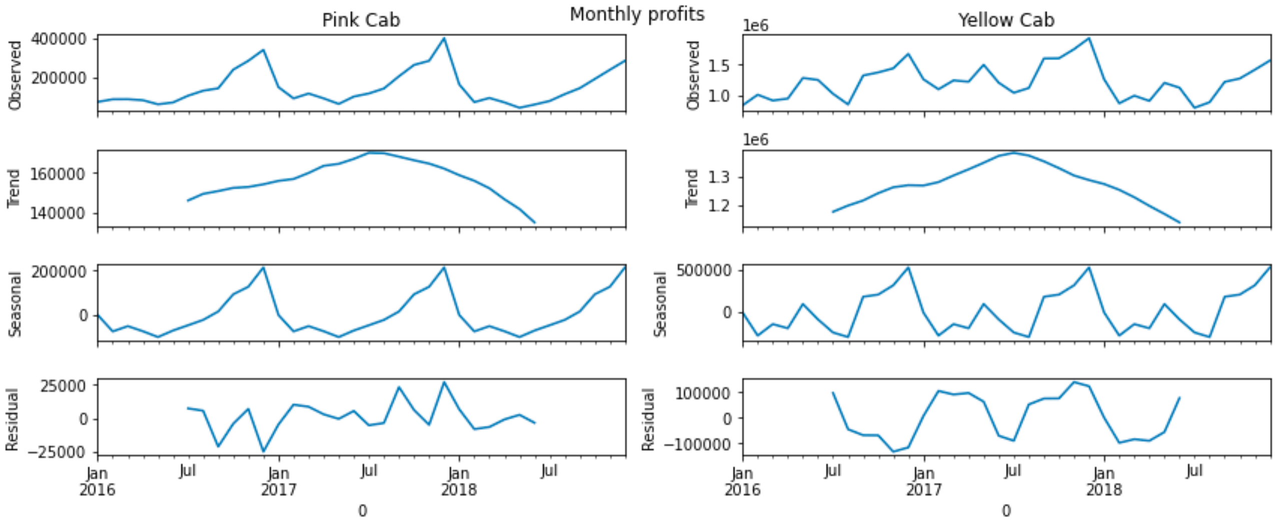




Time Series Analysis

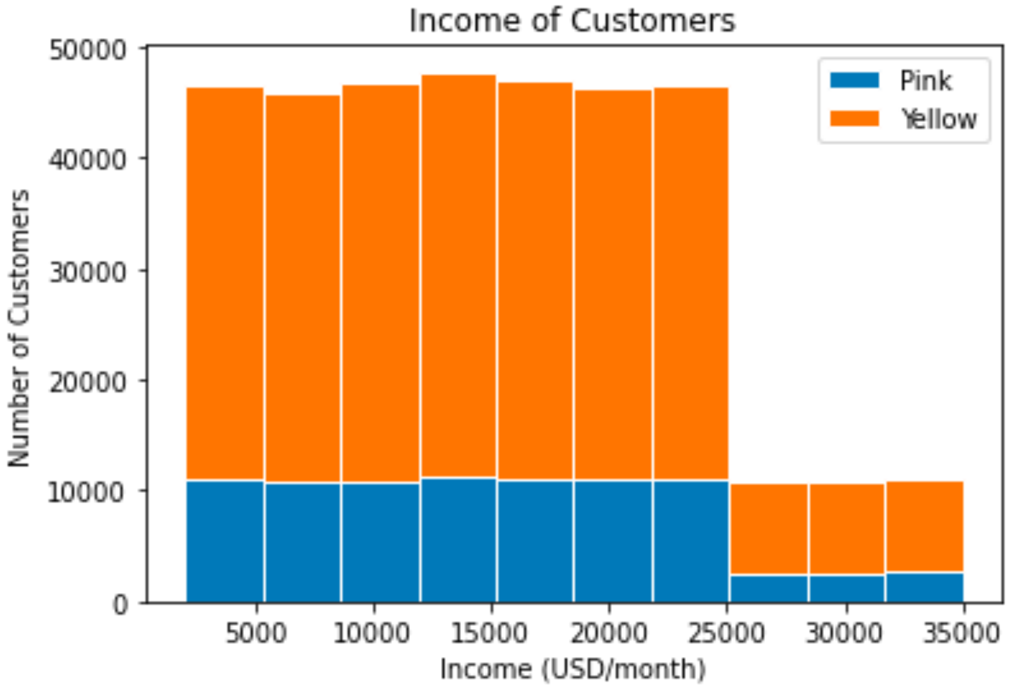
- The ‘*statsmodels*’ package allows us to decompose a time series and view its components in isolation.
- The plots below the components: Observed = Trend + Seasonal + Residual
- The plot on the left shows monthly sales counts and the one on the right shows monthly profits.
- Both plots show very clear seasonality. The peaks in sales coincide with the winter holidays.
- Upon closer examination, it can be seen that there is a difference between the seasonal components of the two companies’ monthly profits. The Yellow Cab company has a second, smaller peak in profits that occurs each year around the month of May, which is spring break period. However this peak is not seen in the profits of the Pink Cab company.

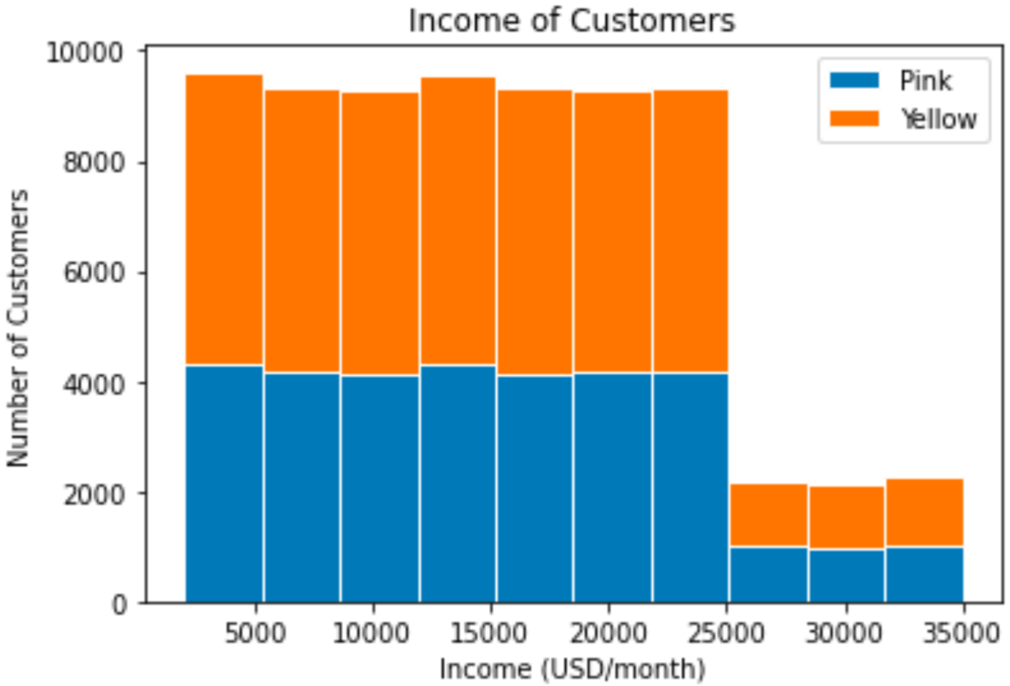




Customer Profile - Income

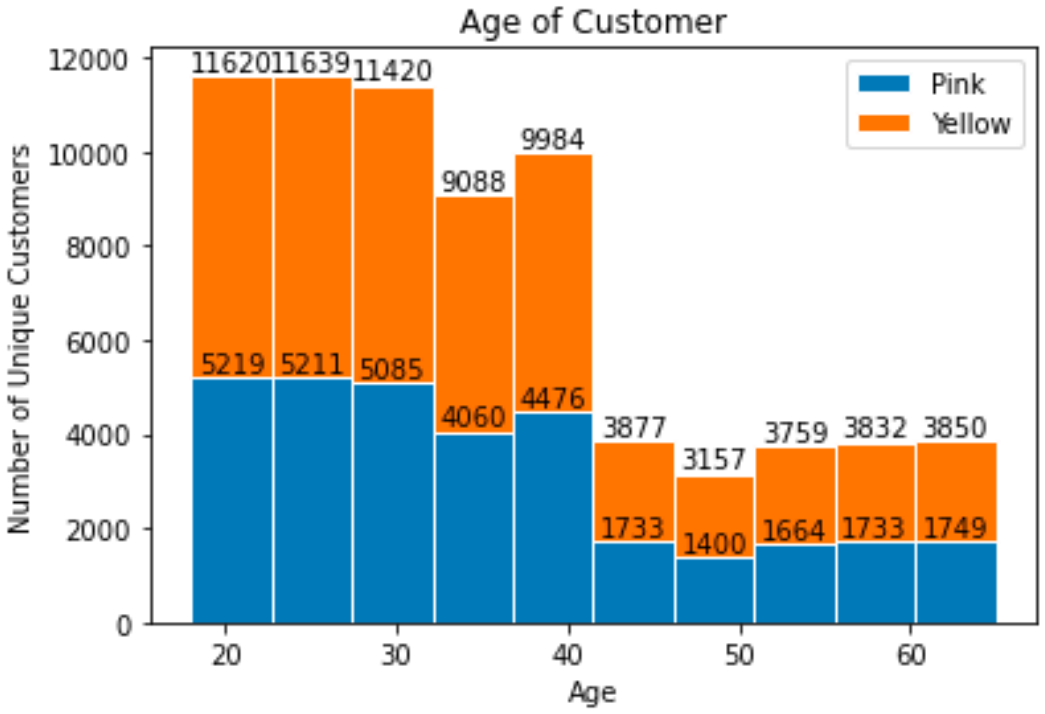
- The plots below show customers grouped by income for each company.
- The plot on the left includes data from all transactions. It can be seen that the Yellow Cab company has approximately 4 times as many customers as the Pink Cab company for every income bracket.
- The plot on the right show customer incomes, but repeat customers have been removed. This paints a clearer picture of the customer base. The Yellow Cab company has approximately twice as many customers in every income bracket.

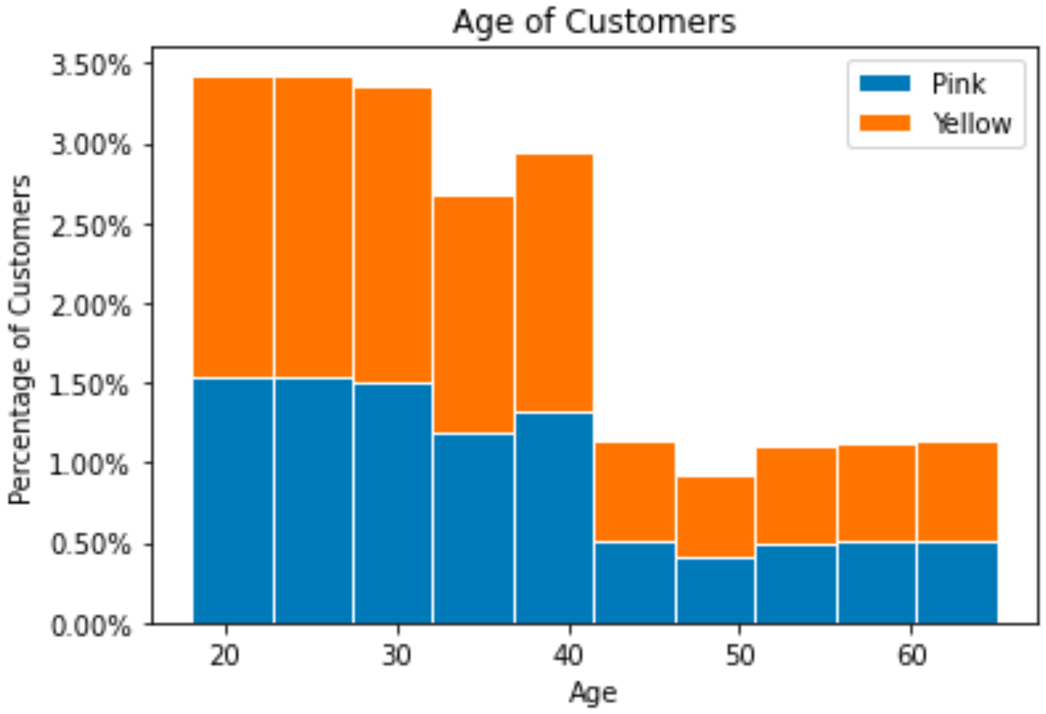




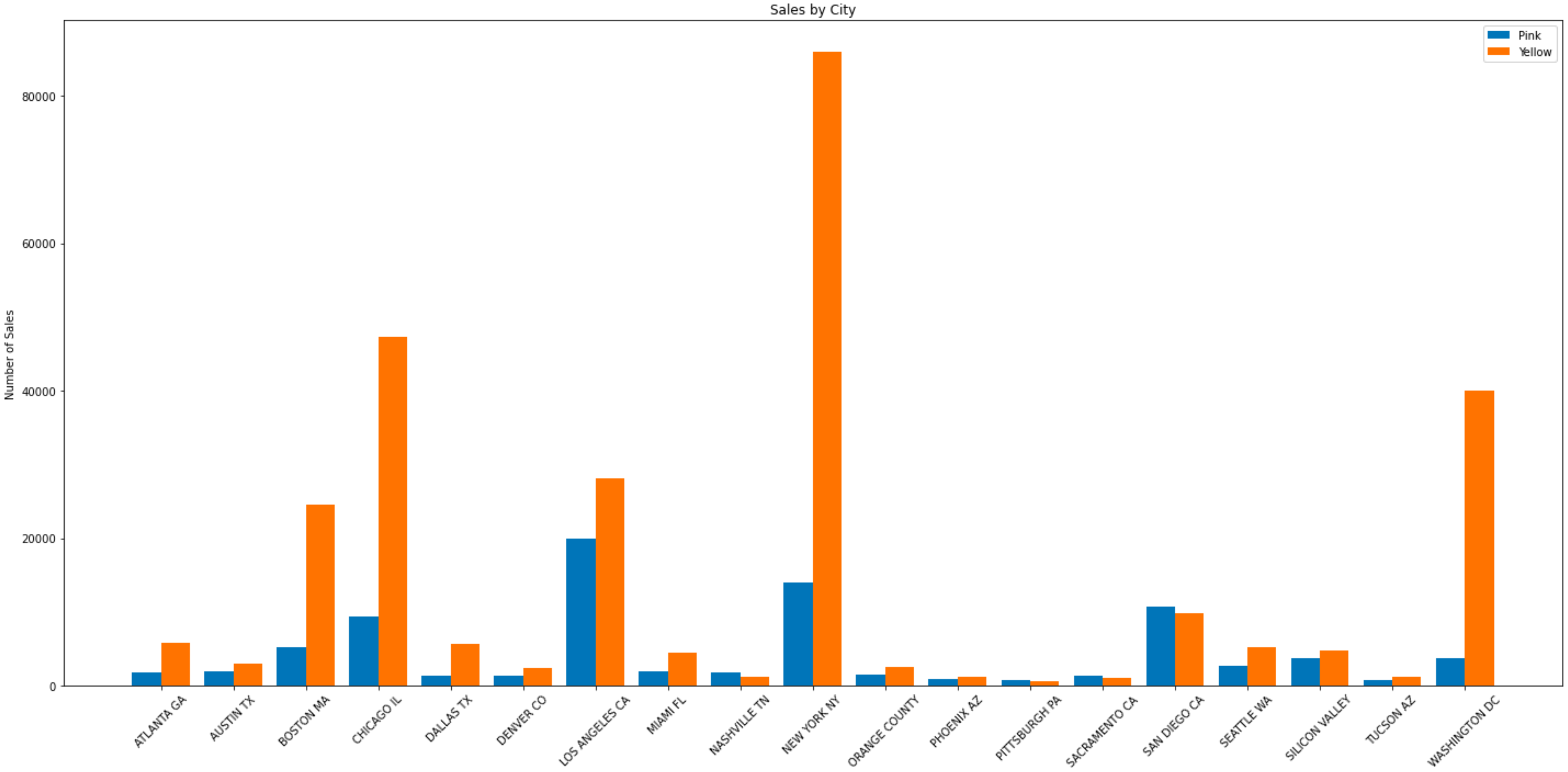
Customer Profile - Age

- The plot on the left shows a comparison of the raw numbers of customers for each air group.
- It can be seen that the Yellow Cab company has at least twice as many customers in each age group.
- The plot on the right shows the relative densities of the (unique) customers by age group.
- Here, it is confirmed that the Yellow Cab company does indeed have *at least* twice as many customers in every age group.



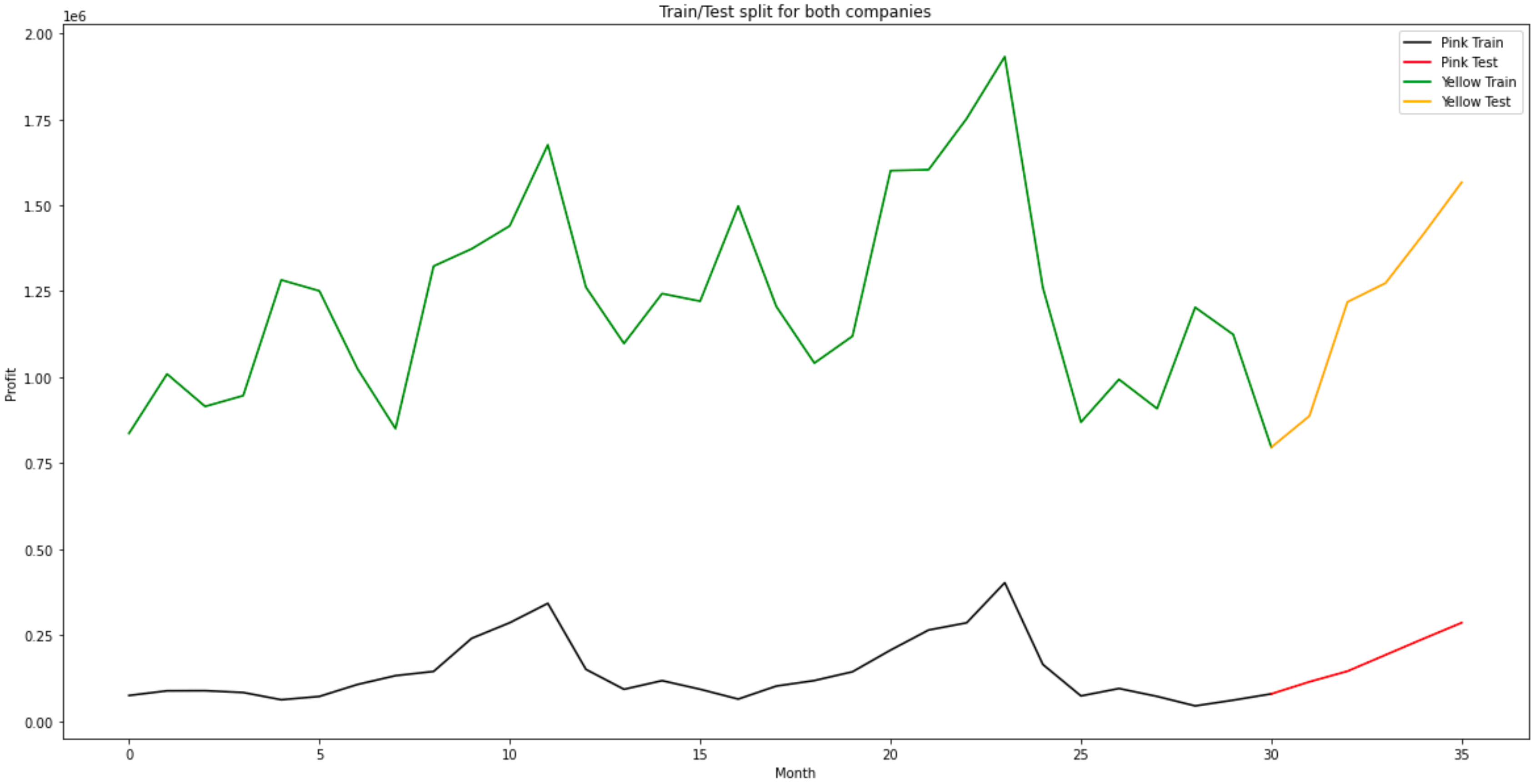


Sales by City



- The plot shows number of sales by city.
- As expected, Yellow Cab company’s sales numbers are a lot higher.
- However, the Pink Cab company wins by sales counts in 4 locations, and these are: 'NASHVILLE TN', 'PITTSBURGH PA', 'SACRAMENTO CA’ and 'SAN DIEGO CA’.
- Two of these locations are in California, this information might be useful to an investor with more specific interests, such as an inclination to invest in local (CA) branches.
- It is worth noting, however, that Yellow Cab still has higher profits in these locations, just not higher sales numbers.

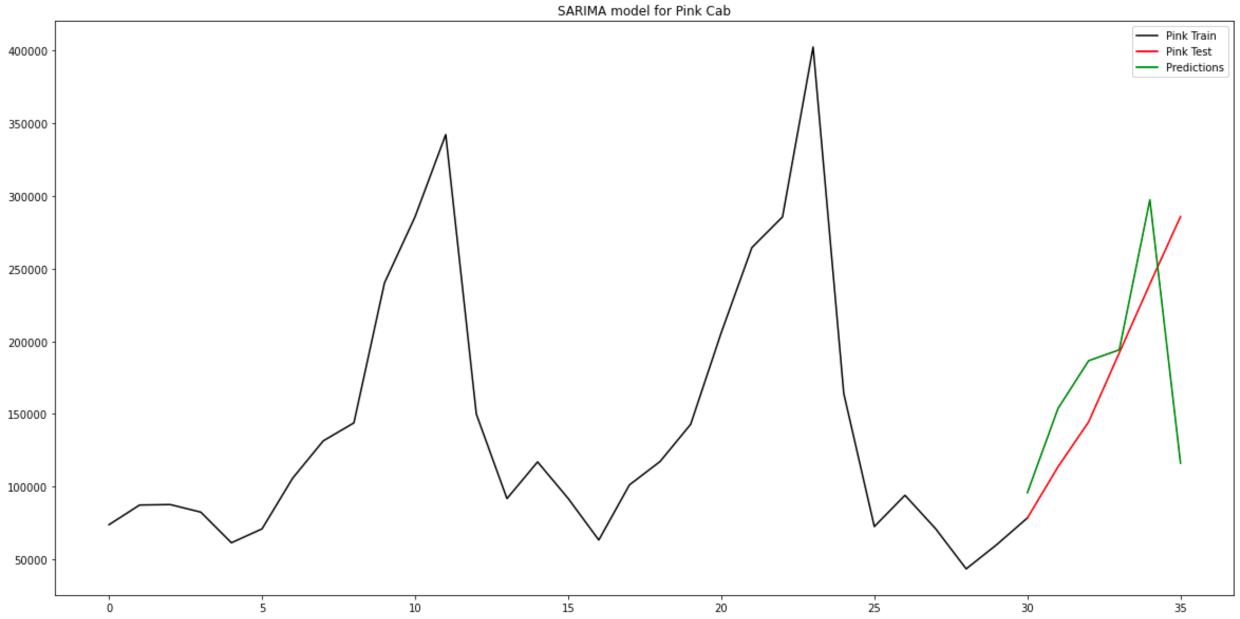
Time Series Analysis - Model Selection

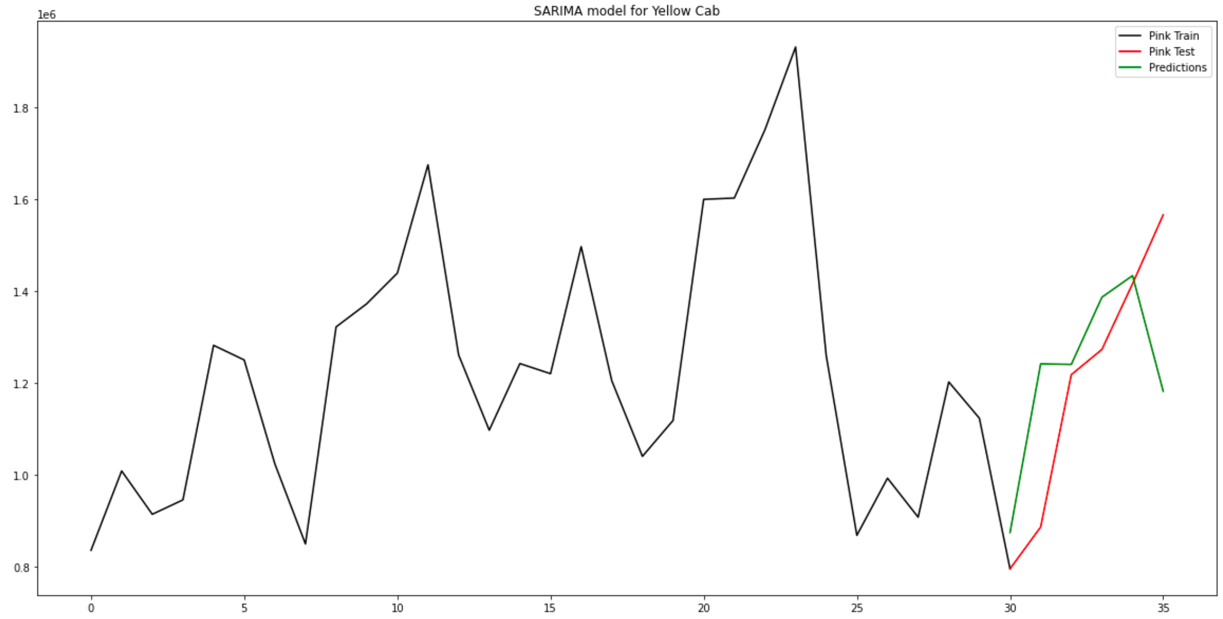


- We have seen in the initial ‘Time Series Analysis’ slide how the data can be broken down into its 3 components: Observed = Trend + Seasonal + Residue.
- Time Series are generally have AR and MA components (ARMA process). We can include the procedure for differencing to achieve stationarity and also account for the seasonal affects (SARIMA process). Once again, we can use the ‘*statsmodels*’ package to build and fit time series models to our data.
- The data for each company can be split into training and test sets. This helps in selecting the best model.

Time Series Analysis - Model Selection

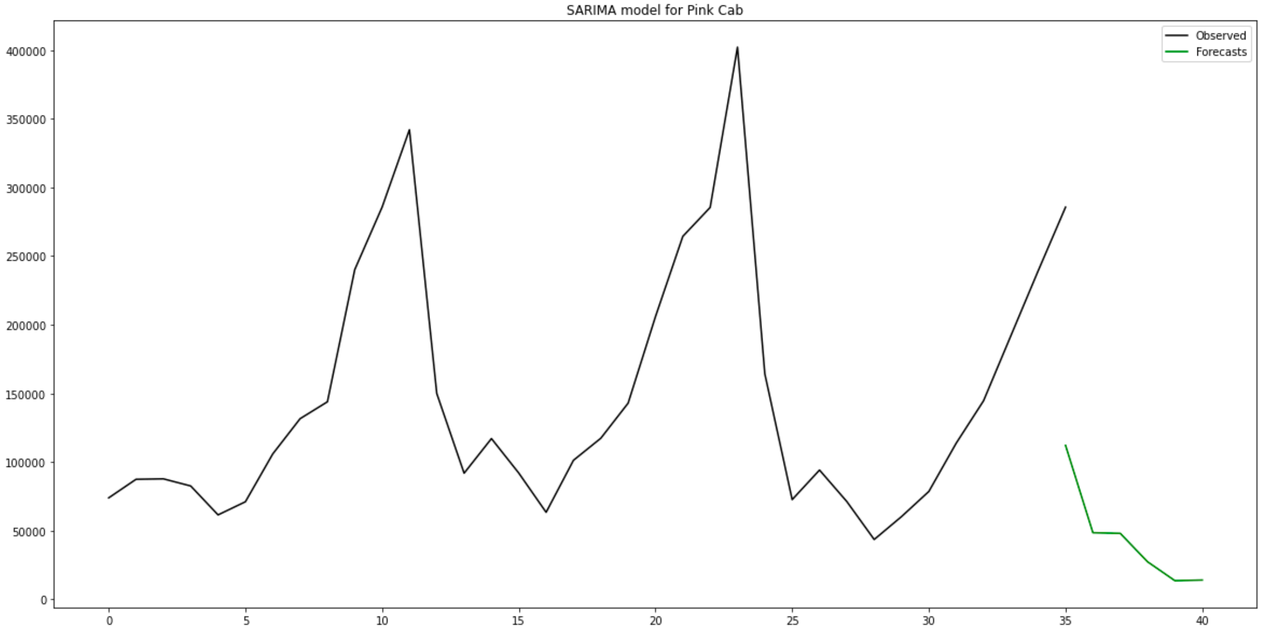
- Plots below show SARIMA models for both companies profits. The model was trained on monthly data i.e; 36 data points split in to 30 for training and the rest for testing.
- In both cases the model seems to want to follow the downward trend seen when we examined the seasonal decomposition of the time series. However, the reality seems to be slightly different and this might indicate a change in the trend that will be seen in the future.
- It is important to choose the model based not only on the error (RMSE) but also AIC and BIC scores. This is because highly complex models will consume additional computational resources.

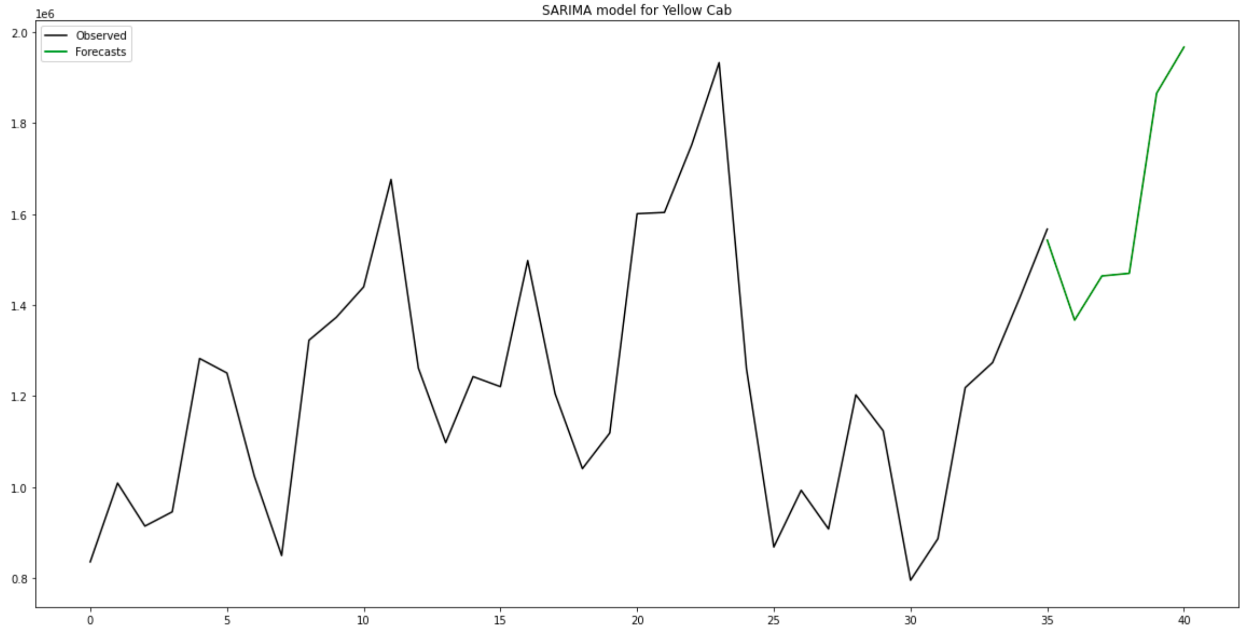




Time Series Analysis - Forecasting

- Plots below show 6 month forecasts for both companies. While Yellow Cab forecast shows large increase in future profits, Pink Cab profits are only declining.
- Models of different seasonal orders were compared during the model selection phase (previous slide). All these models generate similar forecasts. (Model details and comparison in python notebook)





Recommendations

- Pink Cab and Yellow Cab data has been analysed and the companies have been compared.
- Profits: Yellow cab has higher daily and monthly profits.
- Sales Counts: Yellow cab has higher sales counts.
- Customer Income: Yellow cab has more (about 2x) customers from every income bracket.
- Customer Age: Yellow cab has more customers from each age bracket.
- Other Demographics: There is more demographical data provided about the customers, but it is not useful for our analysis. In fact it is not clear whether any factor, aside from customer income, has an tangible effect on the cab companies' profits.
- City: There are 4 cities in which Pink Cab has higher sales counts (['NASHVILLE TN', 'PITTSBURGH PA', 'SACRAMENTO CA', 'SAN DIEGO CA']). Since two of these cities are in California, it might be worth considering Pink Cab if an investor is interested in a particular state (CA). Even though the sales counts for Pink Cab are higher in these cities, Yellow Cab still has higher profits. This might suggest that they have better business practices and are an overall more profitable company.

- Models: The models selected for both companies are similar. The non-seasonal components are the same. This is not too surprising as they are both cab companies operating in the same market. However, the seasonal AR and MA components are a bit different for the 2.
- Both companies have a seasonality of 12 months. This has been seen in the plots and has been confirmed by seasonal decomposition.
- When the models are used to generate forecasts, Yellow Cab profits are shown to increase whereas Pink Cab is in decline.

- Conclusion: Based on all the above information, it is recommended that the investor pick the Yellow Cab company.

Thank You

Data Glacier

Your Deep Learning Partner