Data Glacier Assignment (Week 2)
By: Amir shahcheraghian
Introduction to Project:

US-based XYZ is a private company. Considering the remarkable growth of the Cab Industry in the last few years and the fact that there have been multiple key players in the market, it plans to invest in the Cab Industry and, as per their Go-to-Market (G2M) strategy, they intend to learn more about the market before making a final decision.

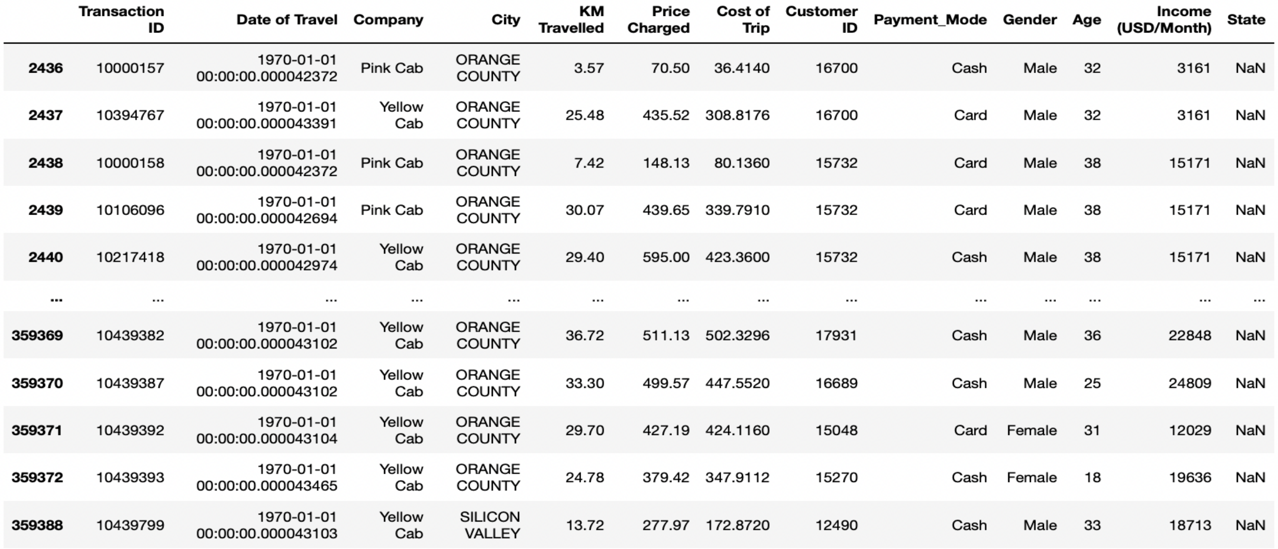
By providing actionable insights, the project aims to assist XYZ company in making an informed investment decision.

Five parts make up the analysis:

- **Dataset Evaluation**
- **Visualizing Demographics**
- **Visualizing Trips**
- **Customer Income Analyse**
- **Conclusion**

Dataset Evaluation

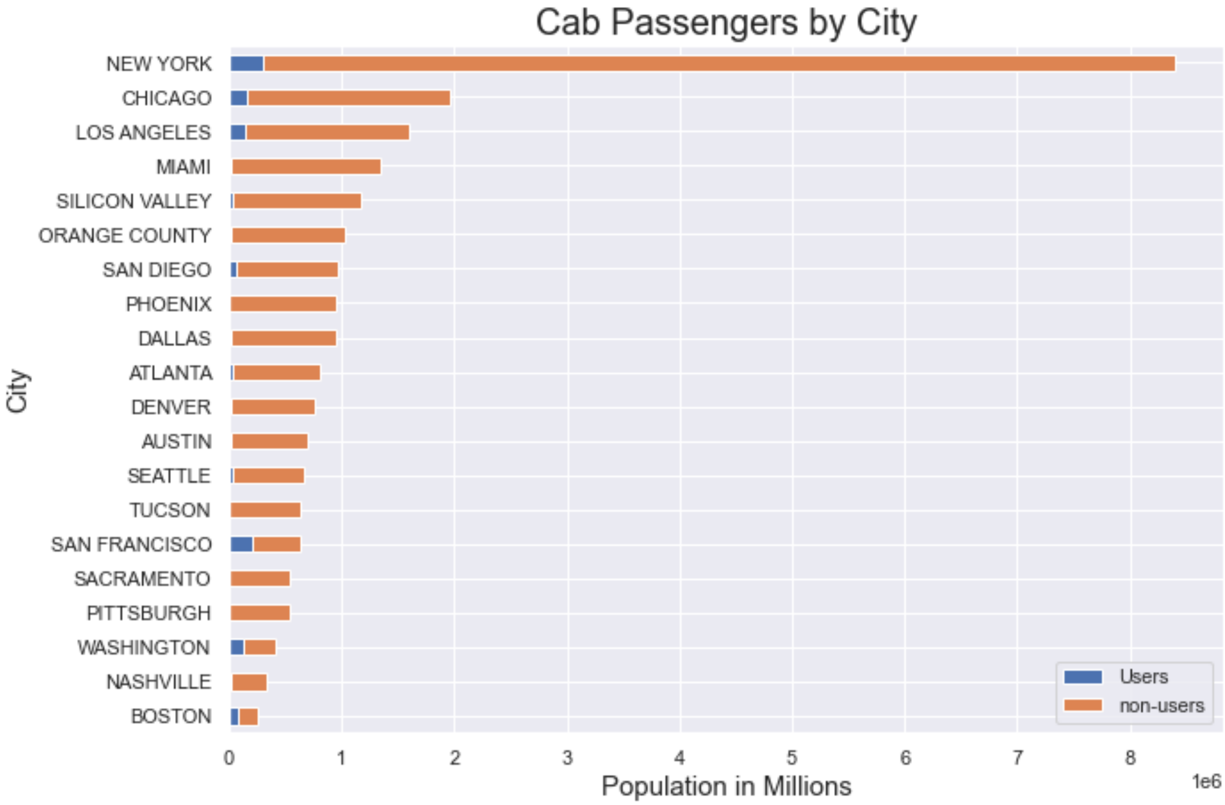
- We have been provided 4 individual data sets. Time period of data is from 02/01/2016 to 31/12/2018.
- Below is the list of datasets which are provided for the analysis:
- Cab_Data.csv: this file includes details of transaction for 2 cab companies.
- Customer_ID.csv: this is a mapping table that contains a unique identifier which links the customer's demographic details.
- Transaction_ID.csv: this is a mapping table that contains transaction to customer mapping and payment mode.
- City.csv: this file contains list of US cities, their population and number of cab users.

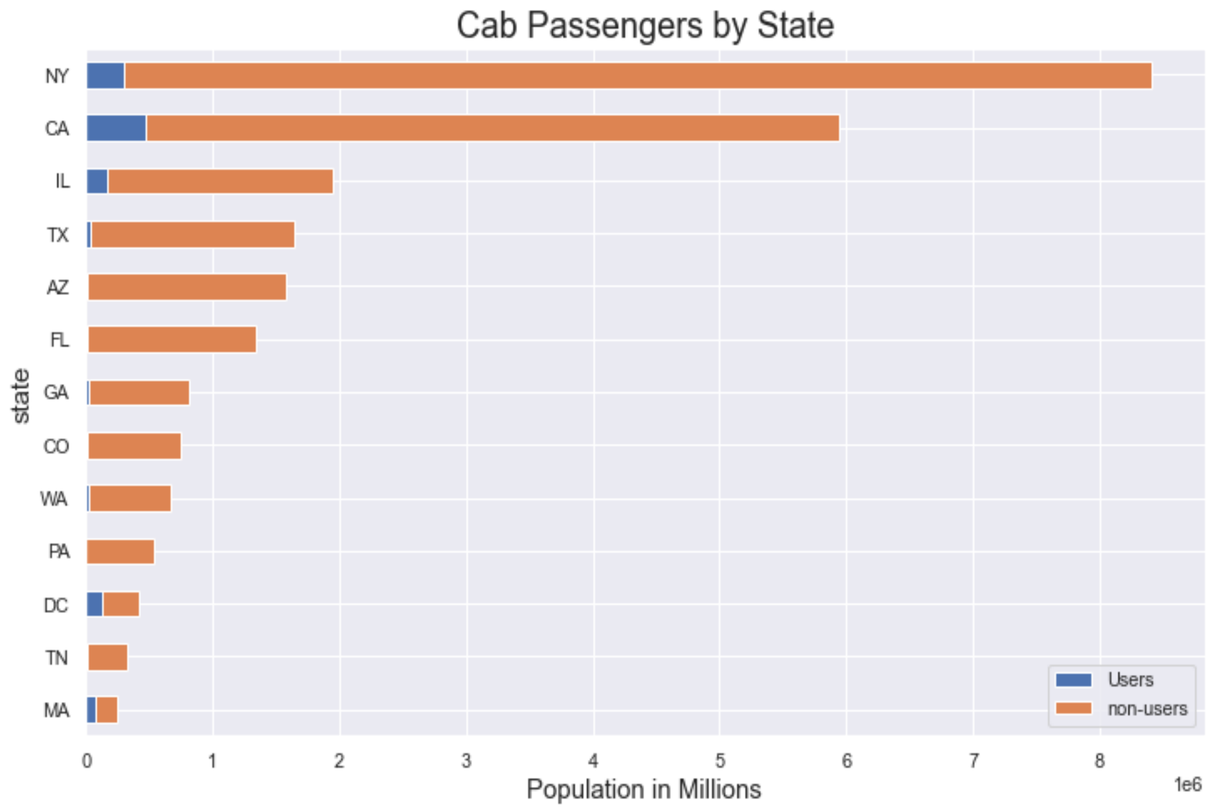


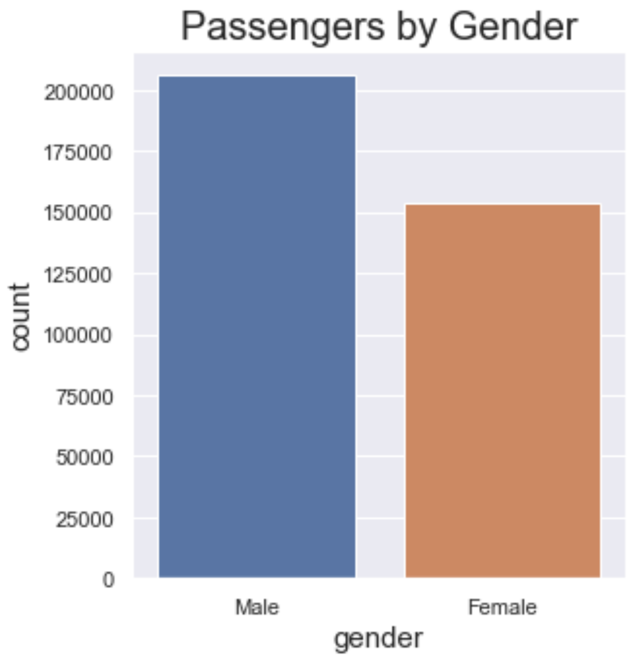
	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	State
2436	10000157	1970-01-01 00:00:00.000042372	Pink Cab	ORANGE COUNTY	3.57	70.50	36.4140	16700	Cash	Male	32	3161	NaN
2437	10394767	1970-01-01 00:00:00.000043391	Yellow Cab	ORANGE COUNTY	25.48	435.52	308.8176	16700	Card	Male	32	3161	NaN
2438	10000158	1970-01-01 00:00:00.000042372	Pink Cab	ORANGE COUNTY	7.42	148.13	80.1360	15732	Card	Male	38	15171	NaN
2439	10106096	1970-01-01 00:00:00.000042694	Pink Cab	ORANGE COUNTY	30.07	439.65	339.7910	15732	Card	Male	38	15171	NaN
2440	10217418	1970-01-01 00:00:00.000042974	Yellow Cab	ORANGE COUNTY	29.40	595.00	423.3600	15732	Cash	Male	38	15171	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
359369	10439382	1970-01-01 00:00:00.000043102	Yellow Cab	ORANGE COUNTY	36.72	511.13	502.3296	17931	Cash	Male	36	22848	NaN
359370	10439387	1970-01-01 00:00:00.000043102	Yellow Cab	ORANGE COUNTY	33.30	499.57	447.5520	16689	Cash	Male	25	24809	NaN
359371	10439392	1970-01-01 00:00:00.000043104	Yellow Cab	ORANGE COUNTY	29.70	427.19	424.1160	15048	Card	Female	31	12029	NaN
359372	10439393	1970-01-01 00:00:00.000043465	Yellow Cab	ORANGE COUNTY	24.78	379.42	347.9112	15270	Cash	Female	18	19636	NaN
359388	10439799	1970-01-01 00:00:00.000043103	Yellow Cab	SILICON VALLEY	13.72	277.97	172.8720	12490	Cash	Male	33	18713	NaN

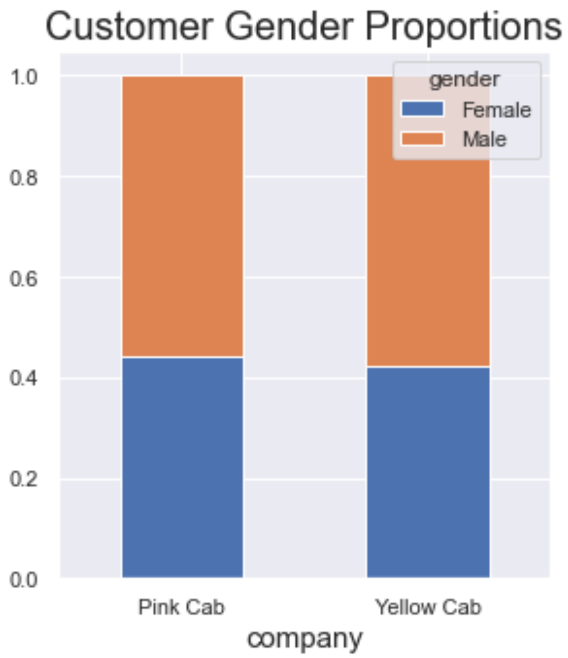
Visualizing demographic

The total population of both users and non-users of cab services by state and City.

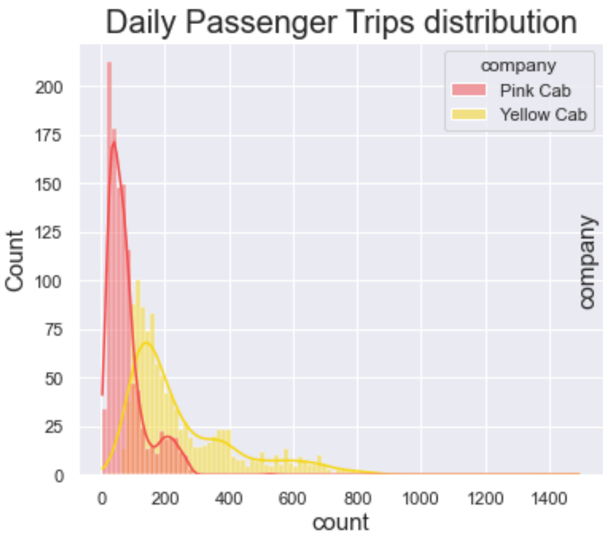


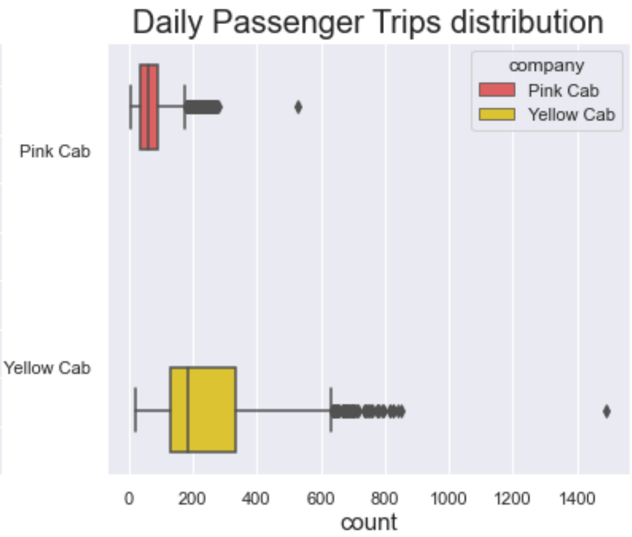




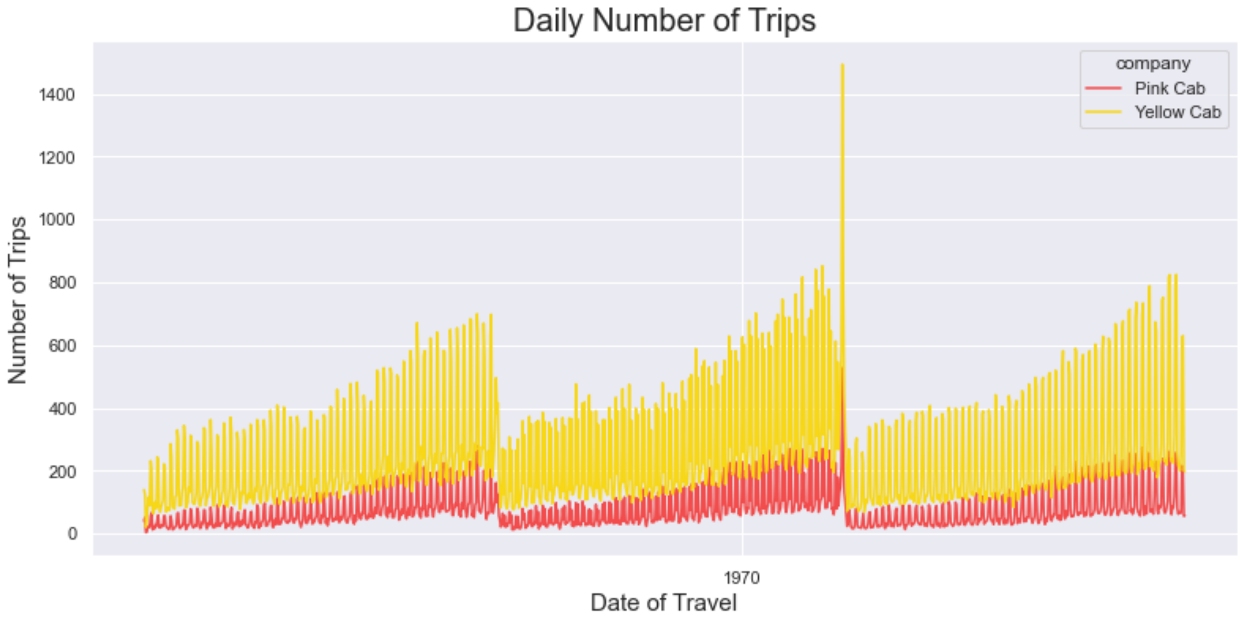


The plots depicts the distribution of daily trips by both Cab companies. Yellow Cab has a higher median trips compared to Pink Cab . Both distributions are skewed to the right, signifying that greater number of trips on some days are rarer.



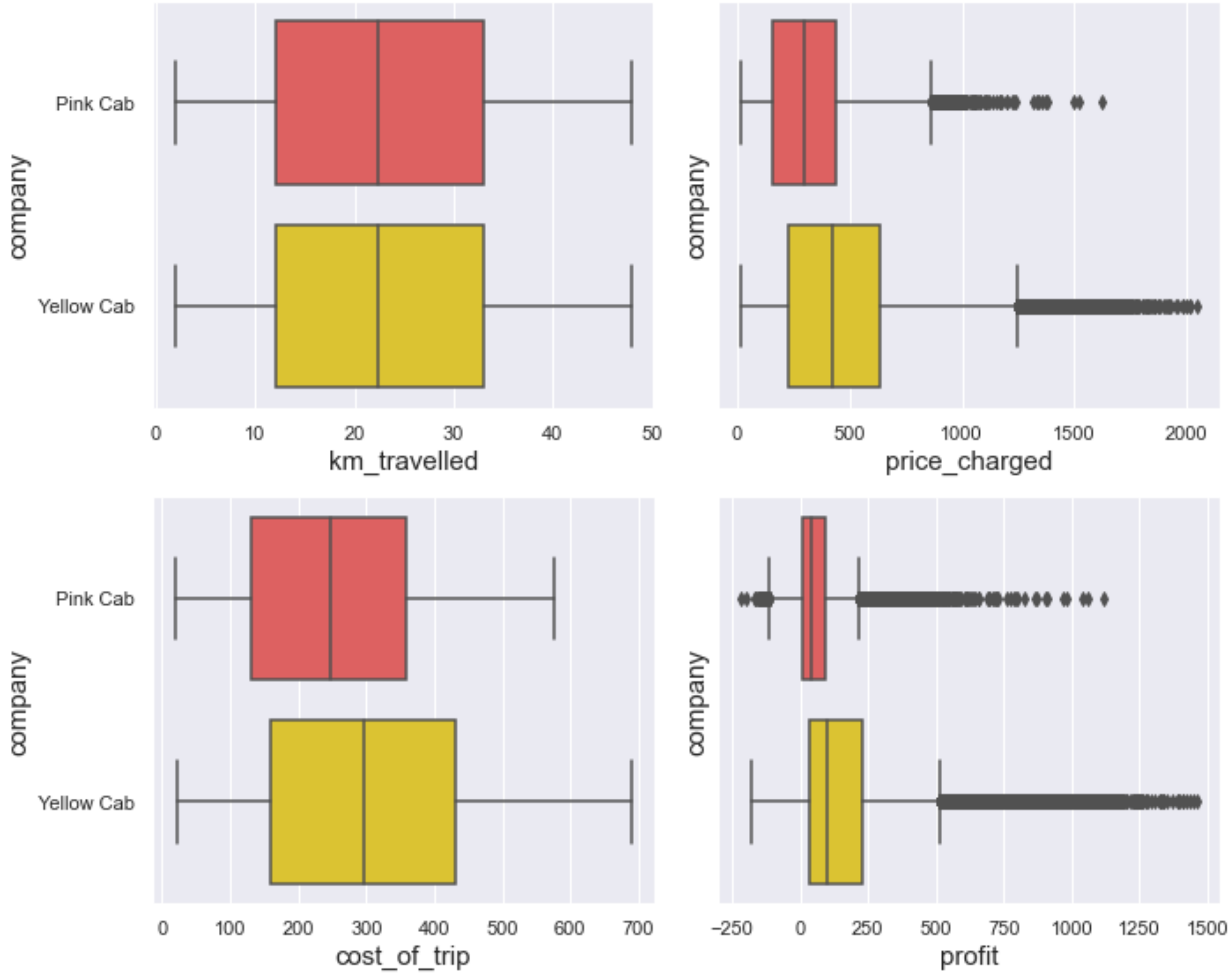


The plot displays daily trips made by both Cab companies from beginning of 2016 till the end of 2018. There is a clear seasonality on a weekly, monthly and yearly level for both Cab companies. Both Cab companies follows the same patterns.
On a monthly level, there is a clear upward trend. On new year, the daily trips dips down to the lowest again. But on a yearly level, the trend seems to be almost uniform.
Yellow Cab makes significantly more trips on any given day compared to Pink Cab . The highest reported trips for both Cab companies was on 5th of January, 2018.

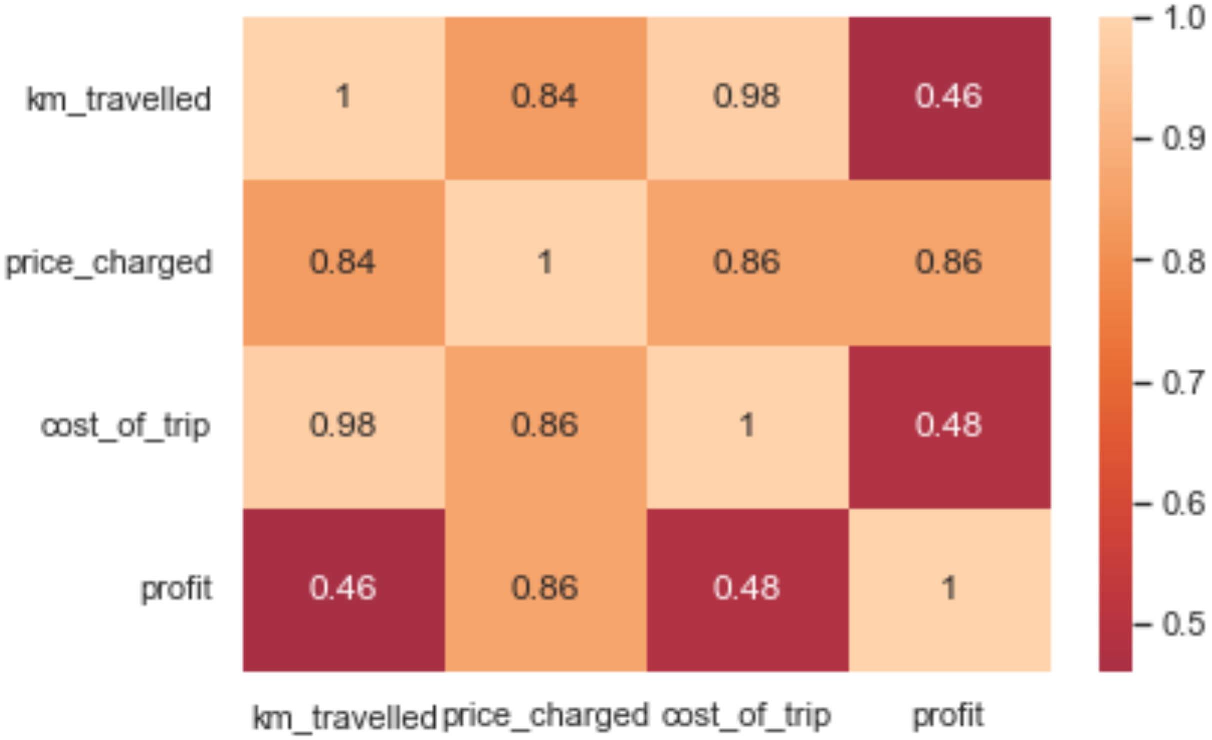


- A plot shows the distribution of trip-related features. There is a uniform distribution in distance traveled, cab expenses. Profit is the only variable with a Gaussian distribution that has a right-skewed distribution.
- Both the profit and price charged columns have outliers on the right. The median distance traveled by both cab companies is the same. Overall, Yellow Cab's expenses are higher. Pink Cab offers a lower median price than its competitors. Its profit margin is considerably higher than its competitors.

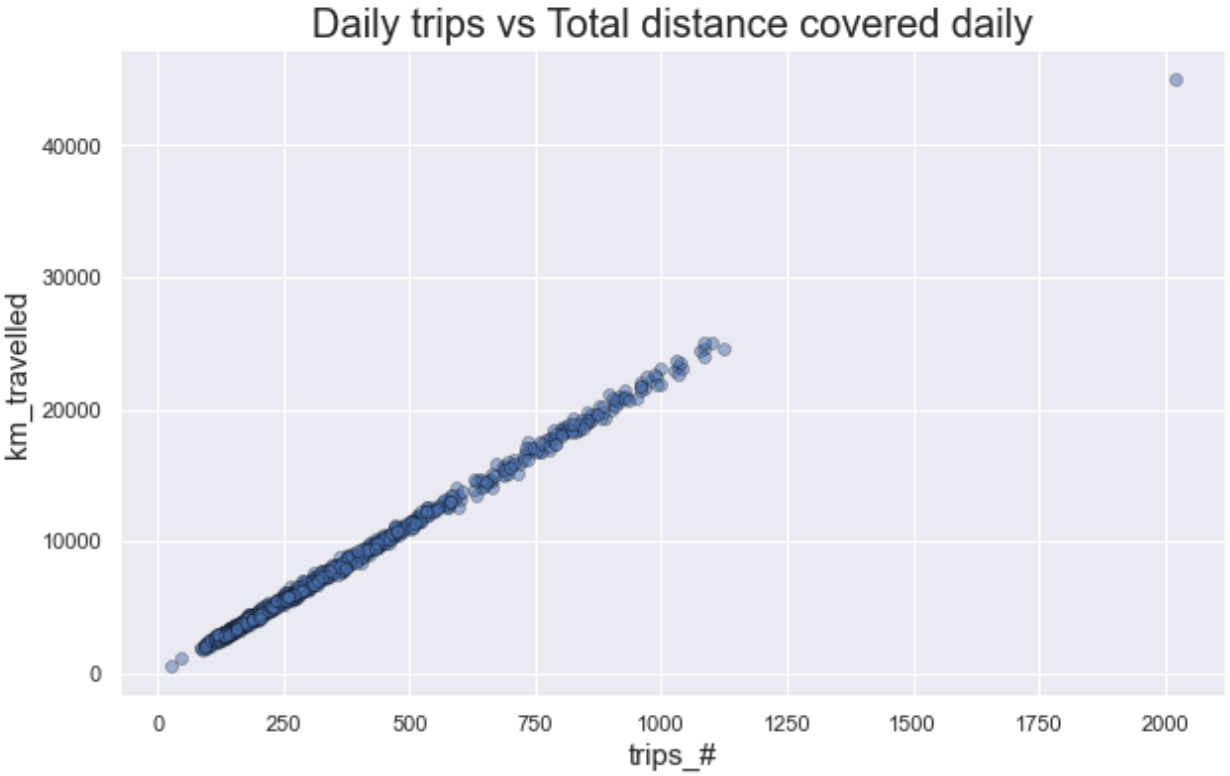
- In the left-hand plot of the profit box plots, both Cab companies have made some losses. In the next session I will examine this in more detail.
- A possible explanation for the outliers in the price charged variable could be that the Cab companies offer luxury or high-end trips under the 'Premium' service. This hypothesis will be tested by capping the price range according to the inter-quantile range of both Cab companies' price charged variable.
- The customer who calls for a premium taxi is richer, and they would travel almost any distance in a premium taxi.



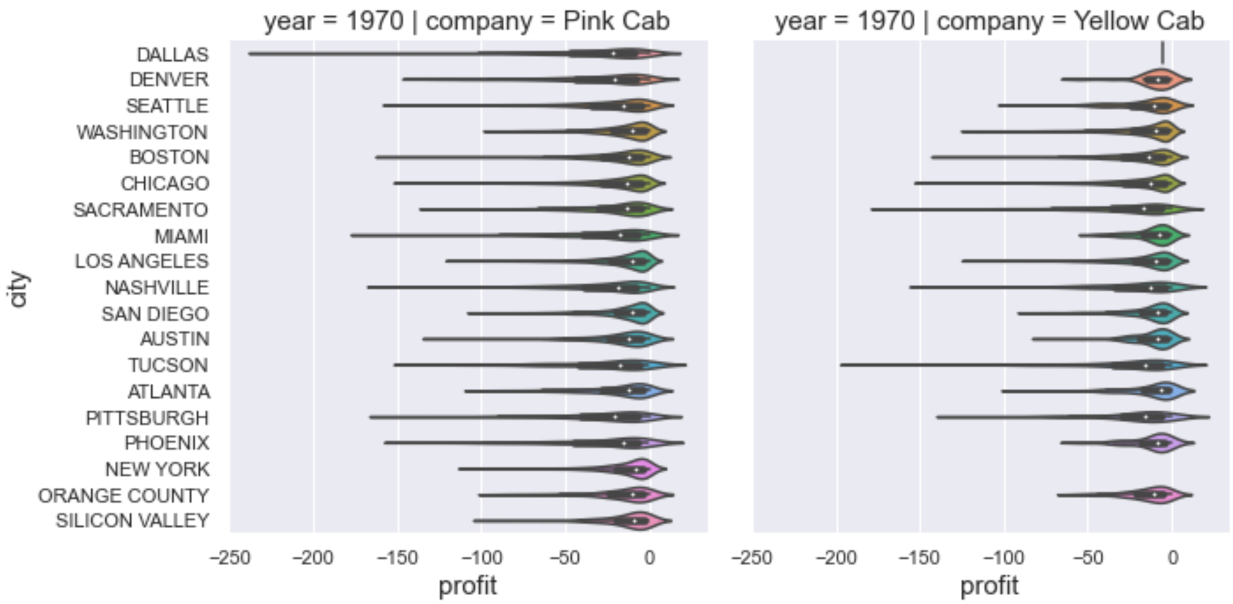
It is obvious from the plot that all features are correlated. Both Cab expenses and Cab fares increase as the distance of travel increases. Cab expenses are highly correlated with distance traveled.



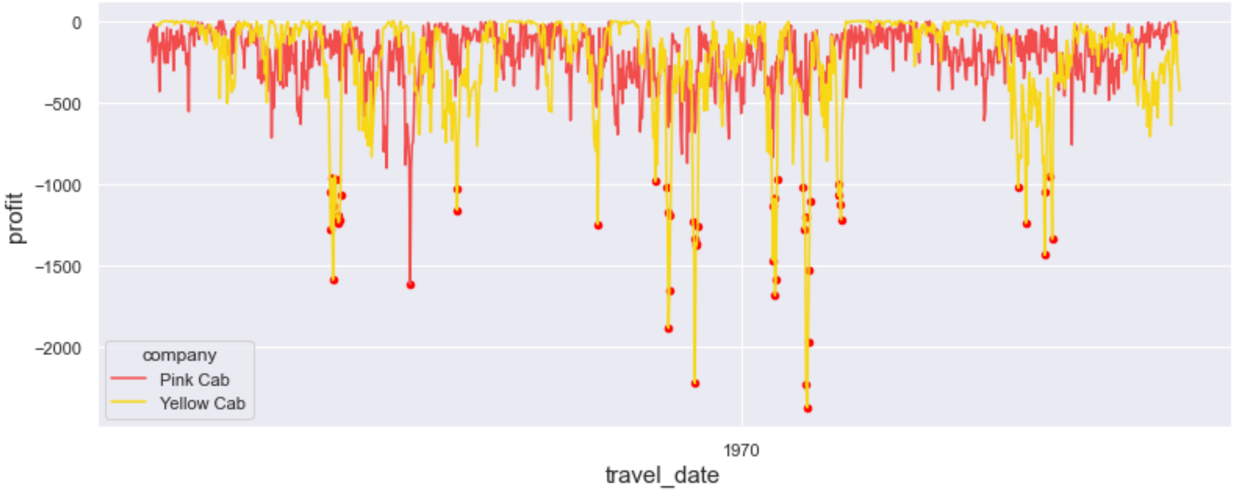
There is a perfect correlation between number of trips and total distance traveled in a day.



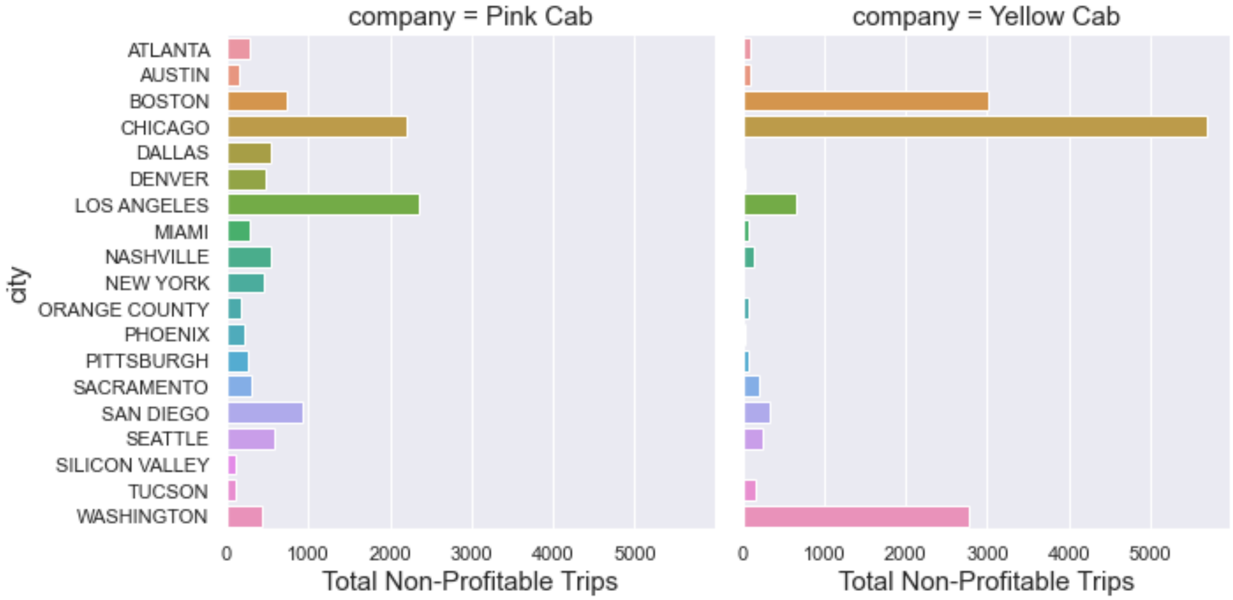
Pink Cab has had a higher frequency of losses than Yellow Cab across all cities for all three years, even though both companies had trips that did not profit. Pink Cab may be affected by these losses in the long run.



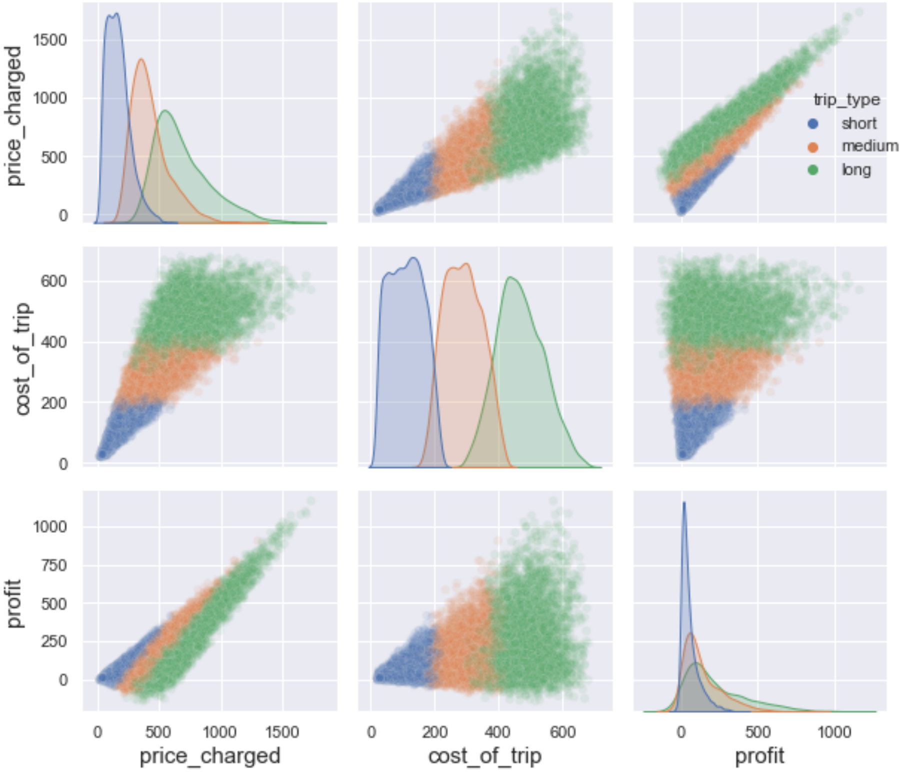
By aggregating the losses at a daily level, the plot shows a time-line of trips that only caused losses. Each year, Yellow Cab has experienced total losses. What's apparent is a pattern. During certain time periods at particular months, there are clusters of losses, most pronounced in July and August.



According to the data, the most number of non-profit trips made by Yellow Cab was on Chicago, Boston, Washington and Los Angeles. For Pink Cab , its mostly on Chicago and Los Angeles.

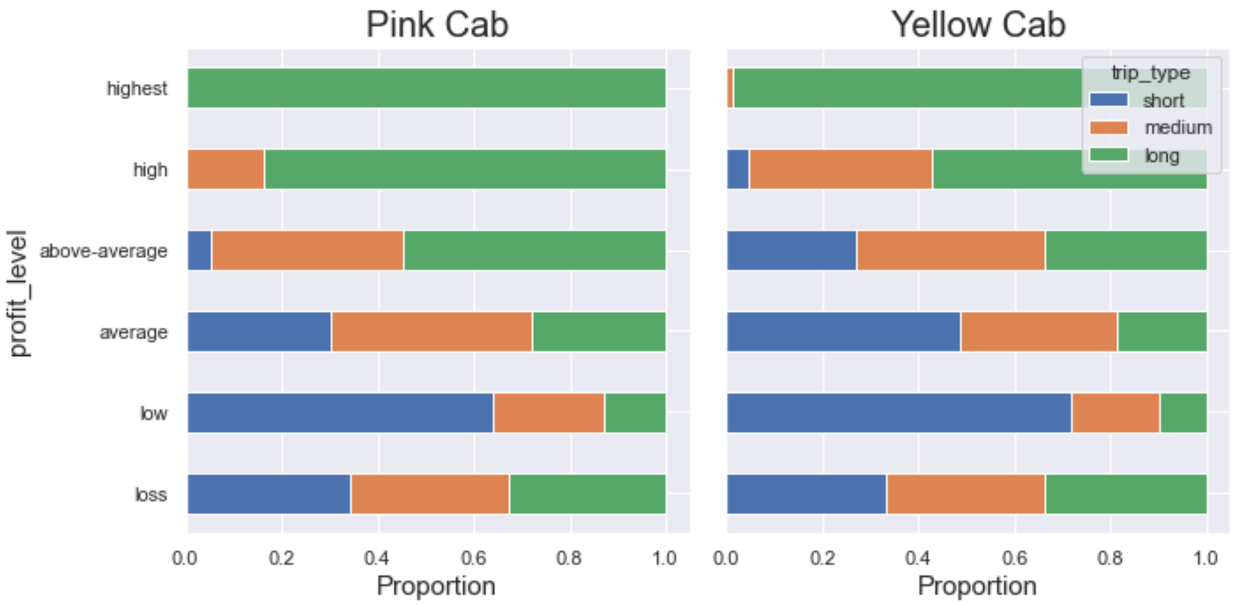


With increasing distances, the price, profit, and costs increase, although the variability in all variables increases.

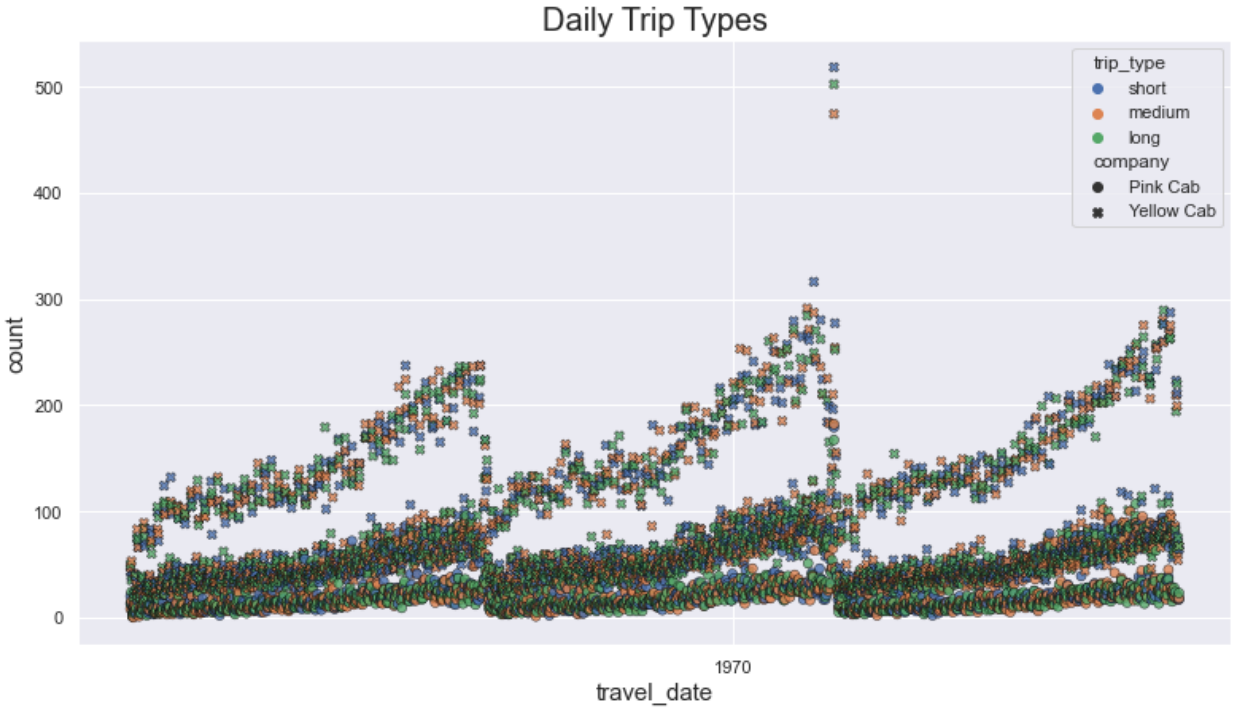


For both companies, the plots show the proportion of trip intervals that contribute to each level of profit. There is no difference in trip durations for the loss category (profit*= 0), so it may not be distance that is the main reason both Cab companies are losing money. The highest profit comes from long trips.

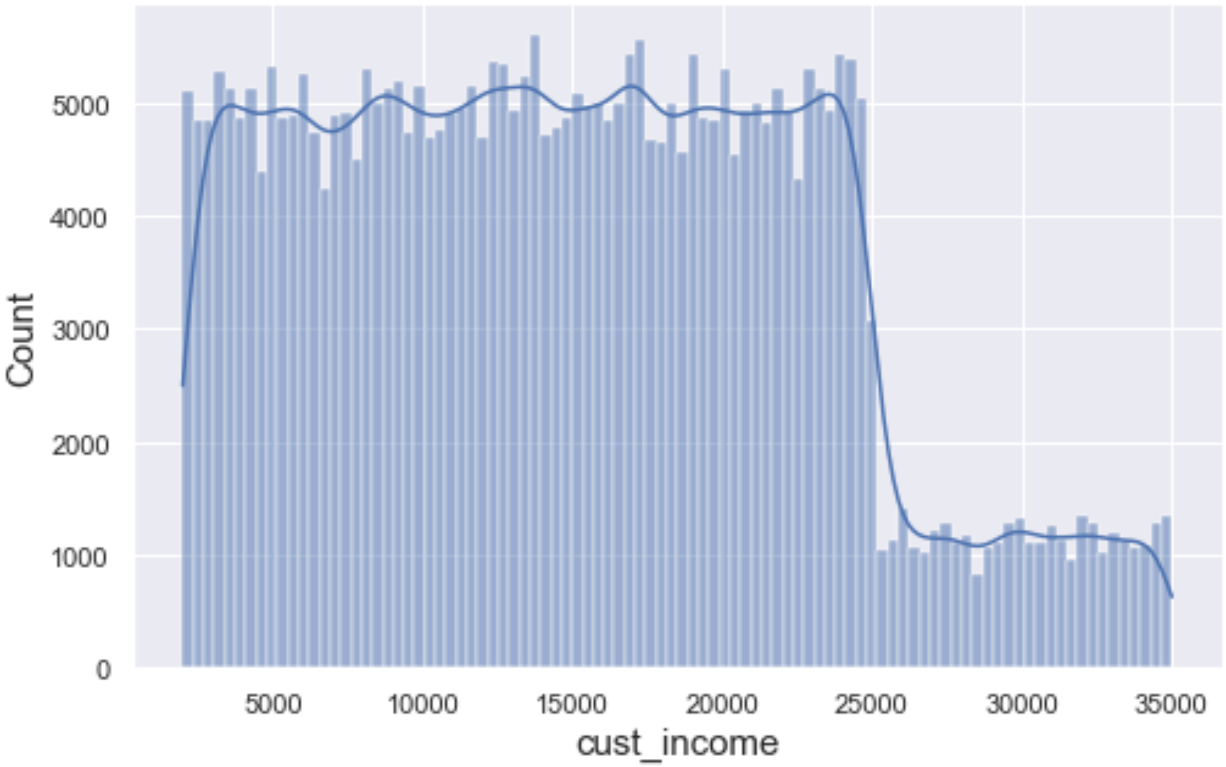
The advantage of Yellow Cab is that it makes better profits from shorter trips than its rival.



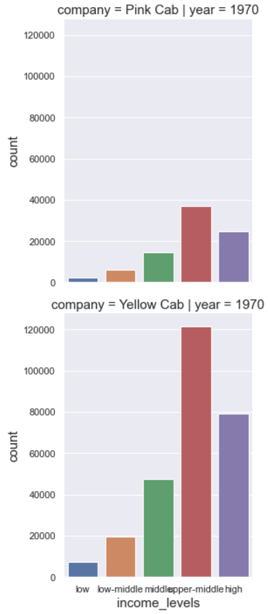
All trip types appear to be equally frequent on a daily basis. There are no glaring exceptions. There was a high number of trips on the 5th of January 2018, according to the data, so all types of trips were taken on that day.



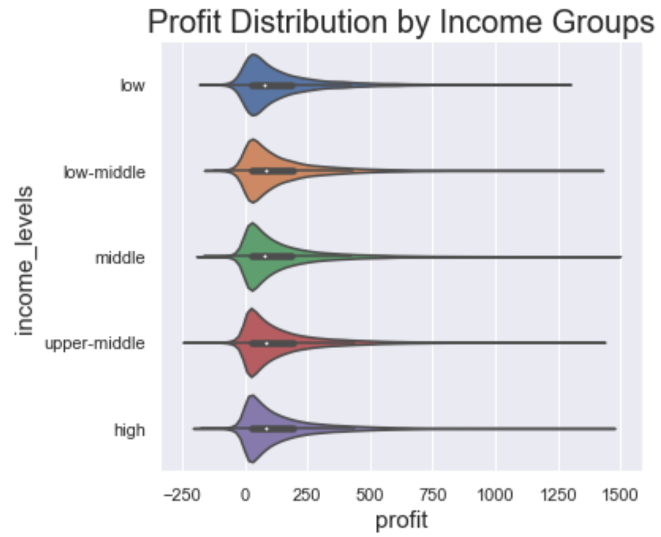
Like the age of a customer, the income of a customer follows a two-phase uniform distribution, meaning the probability of finding customers in all ranges of salaries below 25000 $ is equal to the probability of finding customers in higher income ranges.

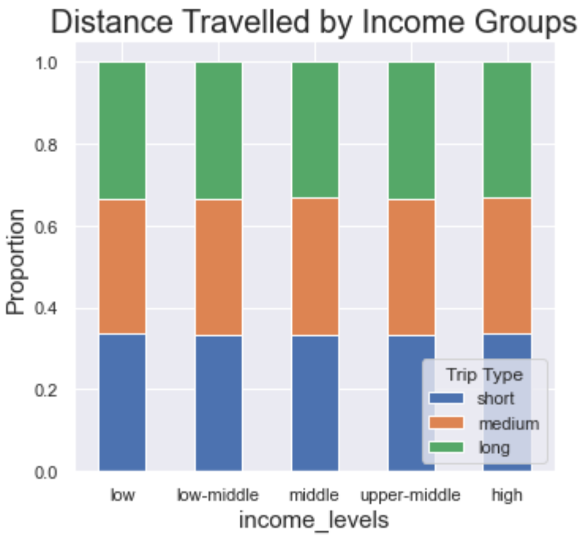


After binning income, we can see that most of the passengers belong to upper-middle class for both Cab companies, followed by high income class. Yellow Cab have higher proportion of passengers. For both companies, there is a slight growth in passengers from 2016 to 2017, but then stagnated/dipped slightly below from 2017 to 2018.


company = Pink Cab | year = 1970

company = Yellow Cab | year = 1970





Conclusion

After analyzing all the variables in the dataset, here is a summary of my analysis:

Both Cab company's financial performance is mainly based on profit. Profit is derived from the difference of the price charged and cost of trip for each trip. Both variables are highly correlated with the distance traveled for each trip. And the total distance traveled in a day is positively correlated with total number of daily trips.

There is weekly, monthly, and quarterly seasonality on the number of rides in each time. The number of cab rides are higher during December and at their lowest during February.

Yellow Cab has higher coverage on cities and has higher loyal customers compared to Pink Cab. Moreover, Yellow Cab seems to perform well almost on all cities and can make significantly higher profits compared to its rival. In conclusion, we can measure a company's performance by looking at the total number of daily trips.

In the next section, I will include extra datasets with the full dataset to see other factors that can affect both company's mode of operations.