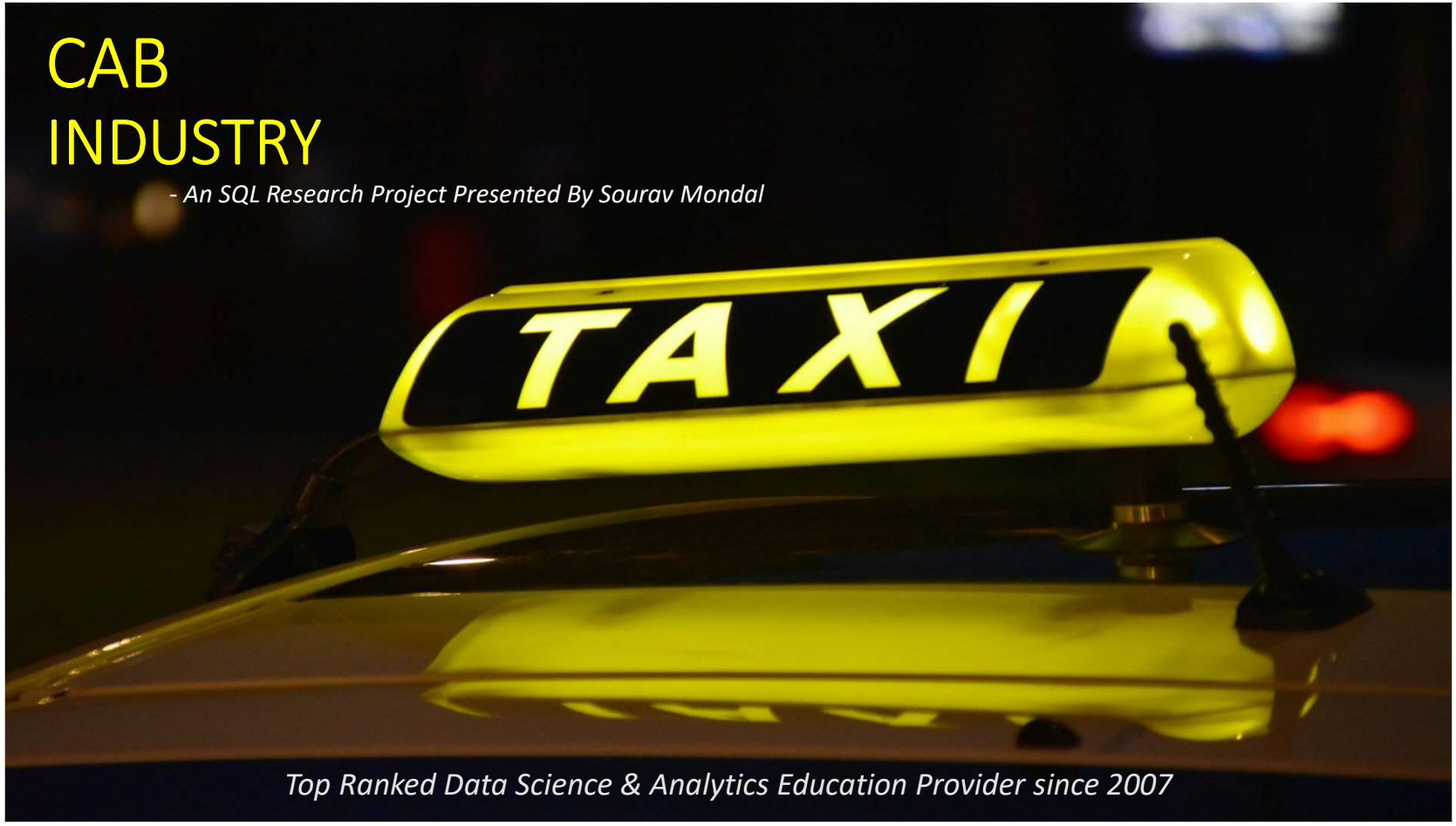
CAB INDUSTRY

- An SQL Research Project Presented By Sourav Mondal



Top Ranked Data Science & Analytics Education Provider since 2007

Project Vision

Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, an investment firm of the United states are planning for an investment in Cab industry as per their Go-to-Market strategy and they want to understand the market before taking final decision. Hence, this project is an analysis of few research topics which might help the firm in understanding the market better for their future investments.

The Dataset comprises of four tables :

Taxi Table – which includes details of transaction for 2 cab companies

Customers Table– This is a mapping table that contains a unique identifier which links the customer's demographic details

Transactions Table – A mapping table that contains transaction to customer mapping and payment mode

Cities Table – This contains the list of US cities, their population and number of cab users

Project Mission

- Creating a database to assign all the four tables, which are to be analysed.
- Normalising the data and to create a relational database using ER diagram.
- Data cleaning
- Analysing the following research topics :
  - Variation of income per month among different customer segments.
  - Determining the most common payment mode used by customers.
  - Cities having the highest and lowest number of taxi transactions.
  - Finding trends or patterns in taxi usage across different cities.
  - Variation of the price charged for taxi rides with distance travelled and city.
  - Identifying seasonal trends or patterns in taxi transactions.
  - Comparison of companies in terms of the number of transactions and revenue generated.
  - Identifying trends or patterns in revenue and profit over time.

Data Tables

The dataset contains a total of 4 tables, having 14 unique features.

The Cities table contains
- City: containing the name for each city.
- Population: The population of each city.
- Users: The number of Cab users in each city

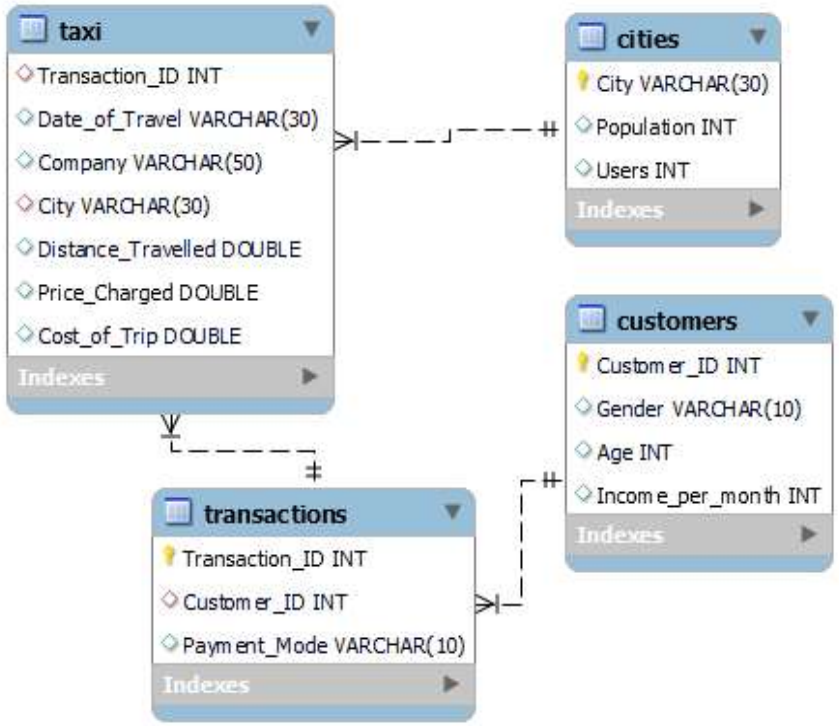
The Customers table contains
- Customer ID: Unique identifier for each customer.
- Gender: Gender of the customer.
- Age: Age of the customer.
- Income per month: The income per month of the customer.

The Transactions table contains
- Transaction ID: Unique identifier for each transaction.
- Customer ID : Links to the Customer ID in the Customers table, representing which customer made the transaction.
- Payment Mode: The mode of payment used for the transaction.

The Taxi table contains
- Transaction ID: Links to the Transaction ID in the Transactions table, indicating which transaction the taxi ride is associated with.
- Date of Travel: The date when the taxi ride occurred.
- Company: The taxi company involved in the ride.
- City : Links to the City in the Cities table, indicating the city where the taxi ride took place.
- Distance Travelled: The distance traveled during the taxi ride.
- Price Charged: The price charged for the taxi ride.
- Cost of Trip: The cost of the taxi trip..

Data Normalisation

- The Customer's table is linked to the Transactions table through the Customer ID field holding a **one-to-many** relationship.
- The Transactions table is linked to the Taxi table through the Transaction ID field, here the Transaction ID acts as a **Foreign key** to the Taxi table.
- The Cities table is linked to the Taxi table through the City field, again being the **Foreign key** to the Taxi table.



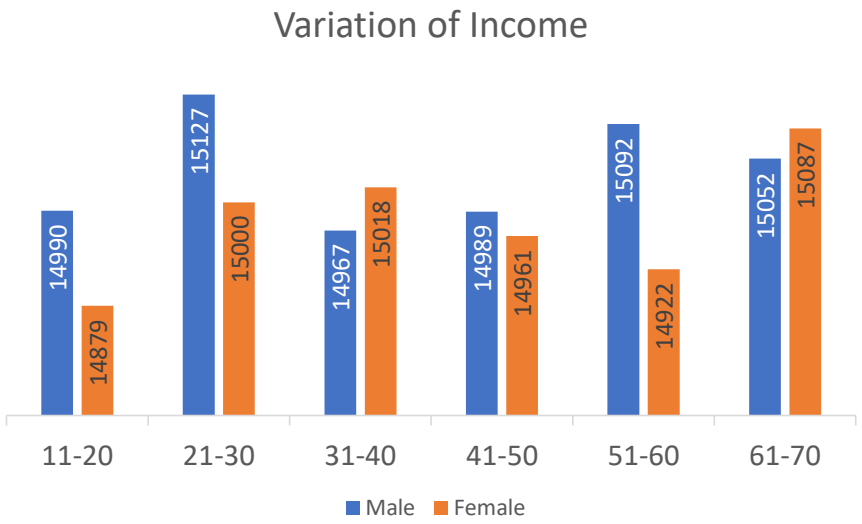
Data Cleaning

- initial dataset underwent thorough validation to ensure completeness and accuracy and no missing values were detected during the data validation process.
- The data column in the Taxi table, initially stored as text, was converted to a more appropriate data type using the string to date function.

Variation of Income among Customers

From the data chart it can be concluded that :

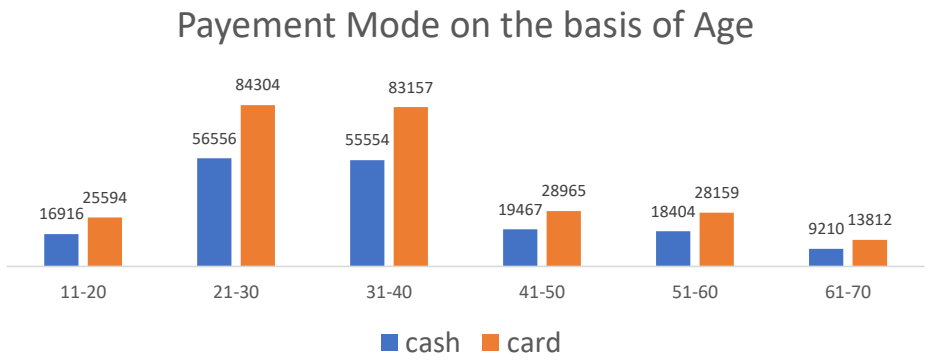
- Males with age between 21-30 has the **maximum** salary, that is approximately 15,127 dollars per month on average.
- Females with age between 61-70 has the **maximum** salary, which is approximately 15087 dollars per month on average.

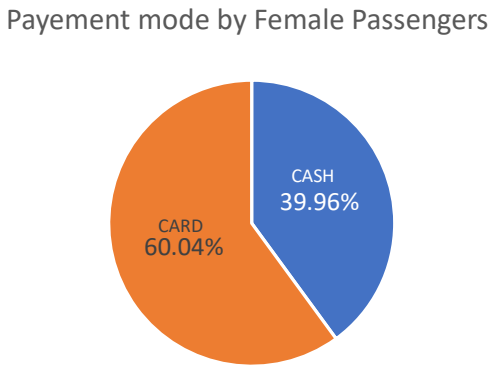


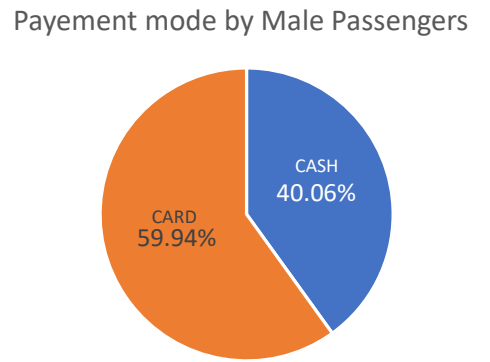
Query for Variation of Income among Customers

```
5  •  with CTE1 as
6        (select *, case
7                when age between 0 and 10 then "0-10"
8                when age between 11 and 20 then "11-20"
9                when age between 21 and 30 then "21-30"
10               when age between 31 and 40 then "31-40"
11               when age between 41 and 50 then "41-50"
12               when age between 51 and 60 then "51-60"
13               when age between 61 and 70 then "61-70"
14               else "Above 70"
15               end as age_group
16       from customers)
17   Select age_group,
18   sum(if(gender="Male", Income_per_month,0))/sum(if(gender="Male", 1,0)) as Male,
19   sum(if(gender="Female", Income_per_month,0))/sum(if(gender="Female", 1,0)) as Female
20   from CTE1
21   group by age_group
22   order by age_group;
```

Payment Mode used by Customers







From the data table it can be stated that:

- Customers prefer cards more, which indicates the era of cashless transactions.
- Male and Female both customers often **prefer cards over cash approximately 60% of the time**.

Query for Payment Mode used by Customers

```sql
create view customers1 as
    (select *, case
            when age between 0 and 10 then "0-10"
            when age between 11 and 20 then "11-20"
            when age between 21 and 30 then "21-30"
            when age between 31 and 40 then "31-40"
            when age between 41 and 50 then "41-50"
            when age between 51 and 60 then "51-60"
            when age between 61 and 70 then "61-70"
            else "Above 70"
            end as age_group
    from customers); -- creating a view that consists age group.
```

```sql
-- based on age group
select c1.age_group,
sum(if(t.payment_mode="cash",1,0))as cash,
sum(if(t.payment_mode="card",1,0))as card
from customers1 as c1 inner join transactions as t
on c1.customer_ID = t.customer_ID
group by c1.age_group
order by c1.age_group;

-- based on gender
select c.gender,
sum(if(t.payment_mode="cash",1,0))/count(*) as cash,
sum(if(t.payment_mode="card",1,0))/count( *)as card
from customers as c inner join transactions as t
on c.customer_ID = t.customer_ID
group by c.gender;
```

Taxi Transaction on each Cities

Counting the number of transactions and grouping the data based on each cites from Taxi table gives the number of transaction on each cities.

- The **maximum** Taxi transactions is seen in the New York city which is approximately **99885.**
- The **minimum** Taxi transaction is seen in Pittsburgh having a number of only **1313** transactions.

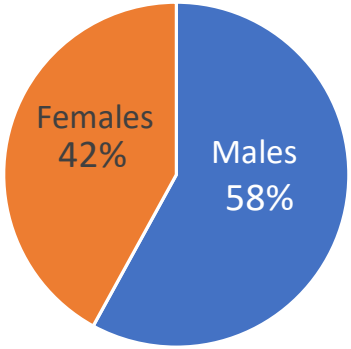
Which clearly indicates Ney York city having more population and tourist attractions. Also, New York city is a global financial and business hub with a diverse economy, attracting a large number of business travelers and professionals who may prefer taxis for their convenience and efficiency.

Query for Taxi Transaction on each Cities

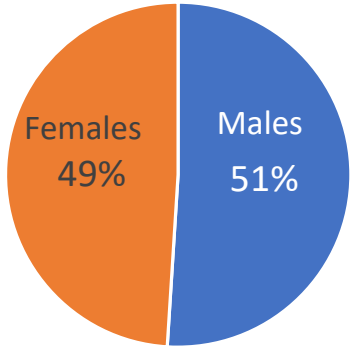
```
56 •   with CTE2 as
57         (select city, count(transaction_ID) as `transaction counts`
58         from taxi
59         group by city)
60     select city, `transaction counts`,
61     rank() over (order by `transaction counts` desc) as ranking
62     from CTE2;
```

Taxi usage among cities : Gender wise

Taxi Usage for the Top 5 Cities



Taxi Usage for the Bottom 5 Cities



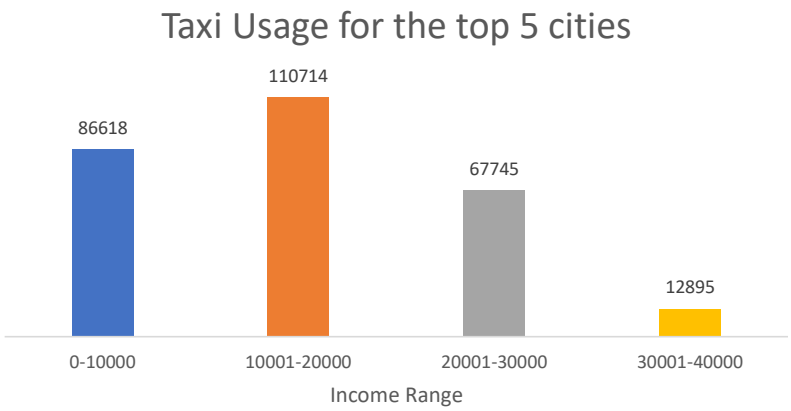
Considering the **Taxi usage** of the Top 5 and the bottom 5 countries, it is inferred that the transactions for both the genders on both the criteria is almost the same with Males taking the lead.

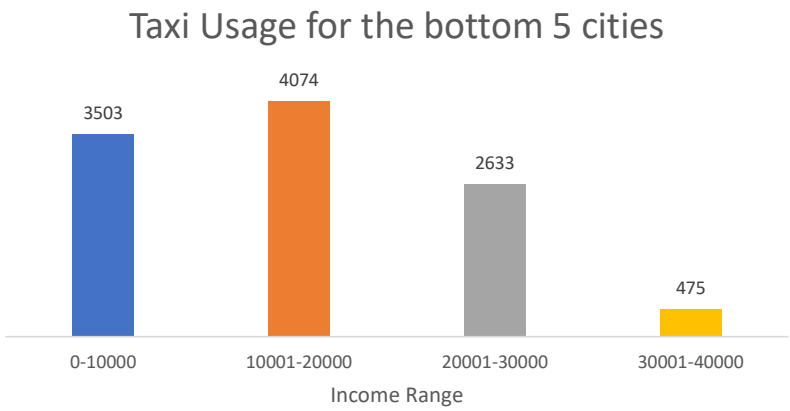
Query for Taxi usage among cities : Gender wise

```sql
-- Finding pattern in genders in top 5 taxi transaction countries
with CTE3 as
    (select t1.city, sum(if(c.gender="male",1,0)) as Males,
    sum(if(c.gender="female",1,0)) as Females, count(*) as total,
    rank() over (order by count(*) desc) as ranking
    from customers as c
    inner join transactions as t on c.customer_ID = t.customer_ID
    inner join taxi as t1 on t.transaction_ID = t1.transaction_Id
    group by t1.city)
select round(avg(Males/(Males+Females)),2) as Males,
round(Avg(Females/(Males+Females)),2) as Females
from CTE3
where ranking <=5;
```

```sql
-- Finding pattern in genders in bottom 5 taxi transaction countries
with CTE3 as
    (select t1.city, sum(if(c.gender="male",1,0)) as Males,
    sum(if(c.gender="female",1,0)) as Females, count(*) as total,
    rank() over (order by count(*) asc) as ranking
    from customers as c
    inner join transactions as t on c.customer_ID = t.customer_ID
    inner join taxi as t1 on t.transaction_ID = t1.transaction_Id
    group by t1.city)
select round(avg(Males/(Males+Females)),2) as Males,
round(Avg(Females/(Males+Females)),2) as Females
from CTE3
where ranking <=5;
```

Taxi usage among cities : Income wise





Among the 2 data charts which represents the **taxi transaction of the top 5 and the bottom 5 cities**, a similar pattern of taxi usage based on income range is observed. Customers having salary up to **20 thousands per month prefer cabs.** The reason why the number lessens for higher salary could be because individuals with higher salary would have their own personal car and would avail lesser cabs.

Query for Taxi usage among cities : Income wise

```sql
create view customers2 as
(select *, case
            when Income_per_month between 0 and 10000 then "0-10000"
            when Income_per_month between 10001 and 20000 then "10001-20000"
            when Income_per_month between 20001 and 30000 then "20001-30000"
            when Income_per_month between 30001 and 40000 then "30001-40000"
            else "Above 40000"
            end as income_range
from customers); -- creating a view that consists income range
```

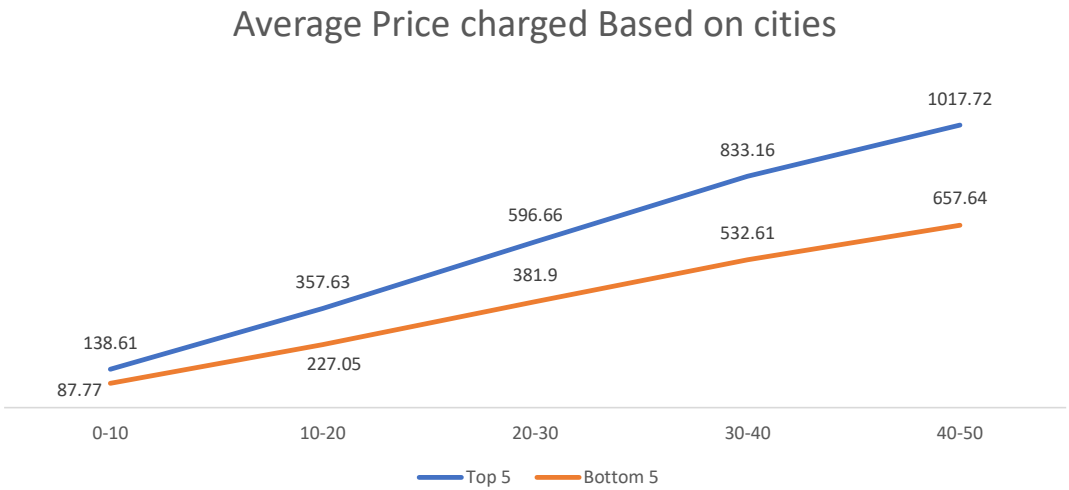
```sql
-- finding pattern in income range for the top 5 countries in taxi usage
with CTE4 as
    (select city, count(*), rank() over (order by count(*) desc) as ranking
    from taxi
    group by city)

select C2.income_range, count(*) as transactions
from CTE4 as C1 inner join taxi as t1 on C1.city = t1.city
inner join transactions as t2 on t1.transaction_ID = t2.transaction_ID
inner join customers2 as C2 on t2.customer_ID = C2.customer_ID
where ranking <= 5
group by C2.income_range
order by C2.income_range;
```

```sql
-- finding pattern in income range for the bottom 5 countries in taxi usage
with CTE5 as
    (select city, count(*), rank() over (order by count(*) asc) as ranking
    from taxi
    group by city)

select C2.income_range, count(*) as transactions
from CTE5 as C1 inner join taxi as t1 on C1.city = t1.city
inner join transactions as t2 on t1.transaction_ID = t2.transaction_ID
inner join customers2 as C2 on t2.customer_ID = C2.customer_ID
where ranking <= 5
group by C2.income_range
order by C2.income_range;
```

Variation of Price charged with Distance Travelled



A linear growth is detected from the charts with the top 5 and bottom 5 cities based on **average price charged**, which suggests that the pricing structure follows a simple linear relationship, where the fare charged is directly proportional to the distance traveled. Also, it is clear, that the **rise in average price charged** for the top 5 cities is **steeper** than the bottom 5 cities which infers the heavy demand of cabs in the top 5 cities.

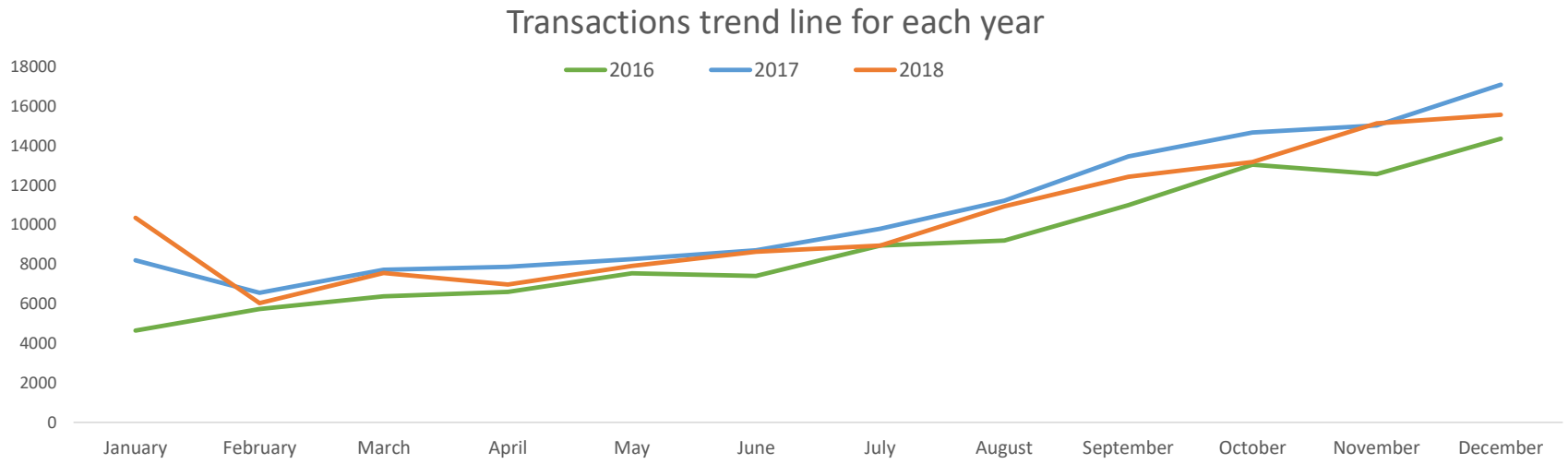
Query for Variation of Price charged with Distance Travelled

```sql
create view taxi1 as
(Select *, case
                when Distance_travelled between 0 and 10 then "0-10"
                when Distance_travelled between 10.001 and 20 then "10-20"
                when Distance_travelled between 20.001 and 30 then "20-30"
                when Distance_travelled between 30.001 and 40 then "30-40"
                when Distance_travelled between 40.001 and 50 then "40-50"
                else "Above 50"
                end as distance_range
from taxi); -- creating a view that consists income range
```

```sql
-- for top 5 countries based on average price charged
with CTE6 as
    (select city, avg(price_charged),
    rank() over (order by avg(price_charged) desc) as ranking
    from taxi
    group by city)
select t.distance_range, round(avg(t.price_charged),2) as `avg price charged`
from CTE6 as C inner join taxi1 t
on c.city = t.city
where ranking <=5
group by t.distance_range
order by t.distance_range;
```

```sql
-- for bottom 5 countries based on average price charged
with CTE7 as
    (select city, avg(price_charged),
    rank() over (order by avg(price_charged) asc) as ranking
    from taxi
    group by city)
select t.distance_range, round(avg(t.price_charged),2) as `avg price charged`
from CTE7 as C inner join taxi1 t
on c.city = t.city
where ranking <=5
group by t.distance_range
order by t.distance_range;
```

Seasonal Trends in Taxi Transactions

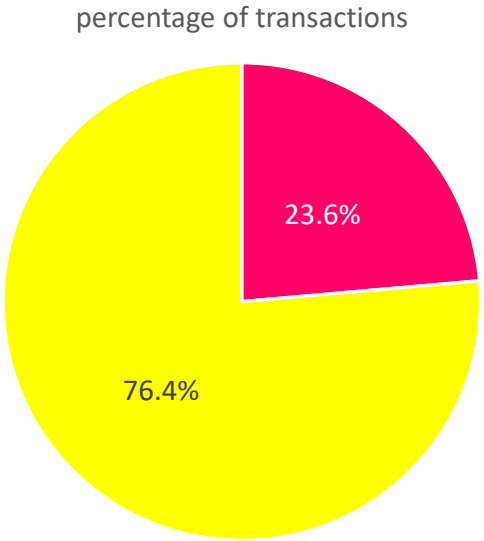


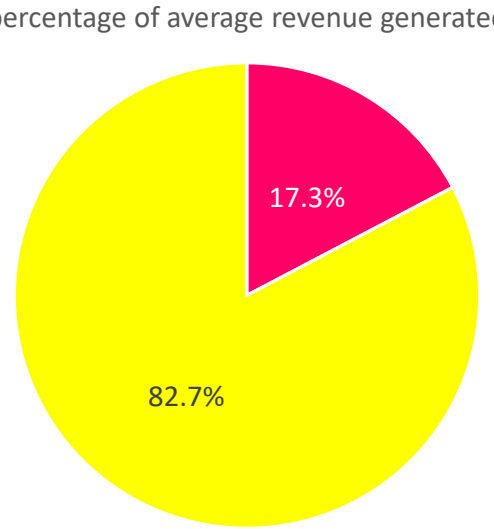
Here the Taxi transactions peaks in the months of **November and December** which reflects a combination of increased travel, tourism, shopping, social activities, and weather-related factors associated with the holiday season. It is also evident that **the number of transactions decreased all over in the year 2018 from 2017**, which indicates the concern of environmental issues and also Ride sharing services.

Query for Seasonal Trends in Taxi Transactions

```
172 •   select month(New_DOT) as slno, monthname(New_DOT) as Months,
173     sum(if(year(New_Dot)=2016,1,0)) as "2016",
174     sum(if(year(New_Dot)=2017,1,0)) as "2017",
175     sum(if(year(New_Dot)=2018,1,0)) as "2018"
176     from taxi
177     group by monthname(New_DOT), month(New_DOT)
178     order by month(New_DOT);
```

Comparison of CAB Companies



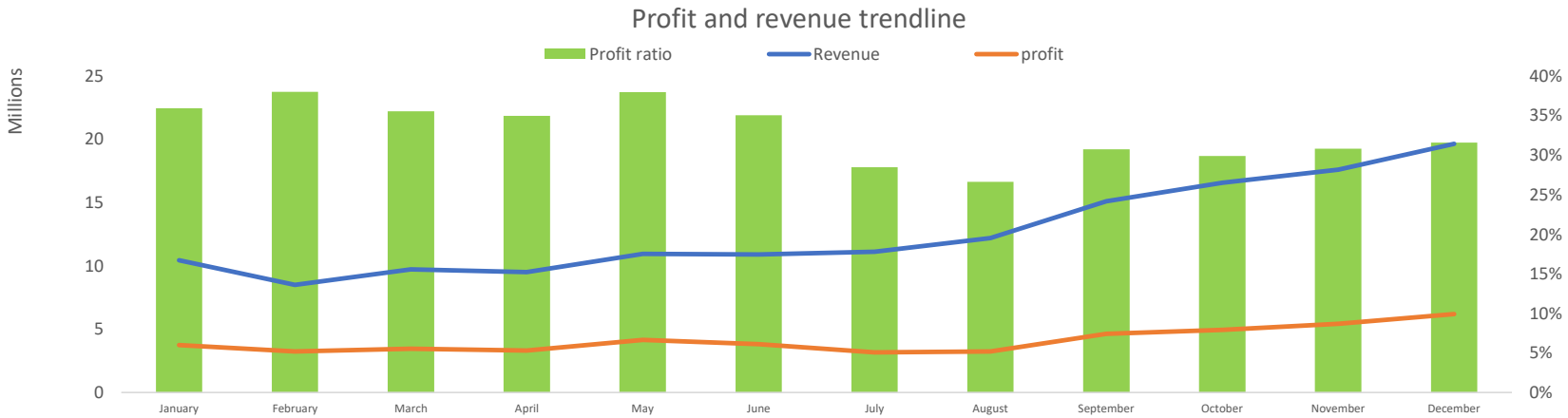


Yellow cabs have more demand than the Pink cab, as the **percentage of transaction and the average revenue generated is significantly high**.

Query for Comparison of CAB Companies

```
181 •   select company,
182     round(count(transaction_ID)/(select count(transaction_ID) from taxi),3)
183     as `percentage of transactions`,
184     round(sum(Price_charged)/(select sum(Price_charged) from taxi),3)
185     as  `percentage of average revenue generated`
186     from taxi
187     group by company;
```

Revenue and Profit Trendline



As it is quite evident from the previous analysis that the transactions peaks in the month of November and December, the revenue also increases at that point of the year, but the profit remains the same throughout the year. The profit ratios for the total year however decreases in the year ending.

Query for Revenue and Profit Trendline

```
190 •   select month(New_DOT) as slno, monthname(New_DOT) as Months,
191     Round(sum(price_charged),2) as Revenue,
192     round(sum(price_charged-cost_of_trip),2) as profit
193     from taxi
194     group by monthname(New_DOT), month(New_DOT)
195     order by month(New_DOT);
```

THANK YOU