Ecommerce Data Analysis

Data Overview:

The data selected, contains details about customer purchases on a particular ecommerce site for a span of 5 years. Some of the important features like, date time stamp, net cost of the purchase, discount amount, location of purchase, method of payment, type of the purchase have been included.

Given below is a screenshot of sample data.

	CID	TID	Gender	Age Group	Purchase Date	Product Category	Discount Availed	Discount Name	Discount Amount (INR)	Gross Amount	Net Amount	Purchase Method	Location
0	943146	5876328741	Female	25-45	30/08/2023 20:27:08	Electronics	Yes	FESTIVE50	64.30	725.304000	661.004000	Credit Card	Ahmedabad
1	180079	1018503182	Male	25-45	23/02/2024 09:33:46	Electronics	Yes	SEASONALOFFER21	175.19	4638.991875	4463.801875	Credit Card	Bangalore
2	337580	3814082218	Other	60 and above	06/03/2022 09:09:50	Clothing	Yes	SEASONALOFFER21	211.54	1986.372575	1774.832575	Credit Card	Delhi
3	180333	1395204173	Other	60 and above	04/11/2020 04:41:57	Sports & Fitness	No	NaN	0.00	5695.612650	5695.612650	Debit Card	Delhi
4	447553	8009390577	Male	18-25	31/05/2022 17:00:32	Sports & Fitness	Yes	WELCOME5	439.92	2292.651500	1852.731500	Credit Card	Delhi

Data Exploration:

The data had 55000 entries with 13 features.

Some basic data quality checks have been conducted.

Check 1: Check for missing/null values

```
CID                          0
TID                          0
Gender                       0
Age Group                    0
Purchase Date                0
Product Category             0
Discount Availed             0
Discount Name            27585
Discount Amount (INR)        0
Gross Amount                 0
Net Amount                   0
Purchase Method              0
Location                     0
```

Only the discount name has missing values, upon checking these rows, it's been concluded that these are purchases done without any discount value.

Check 2: Check for erroneous entries

The Net Amount and Gross Amount features were checked for values <= 0.

There are 613 (~1%) entries have been found with negative Net Amount values. These entries have been deleted before proceeding to further analysis.

Feature Engineering:

Created the following new features on top of existing features, to extract trends and other important inferences.

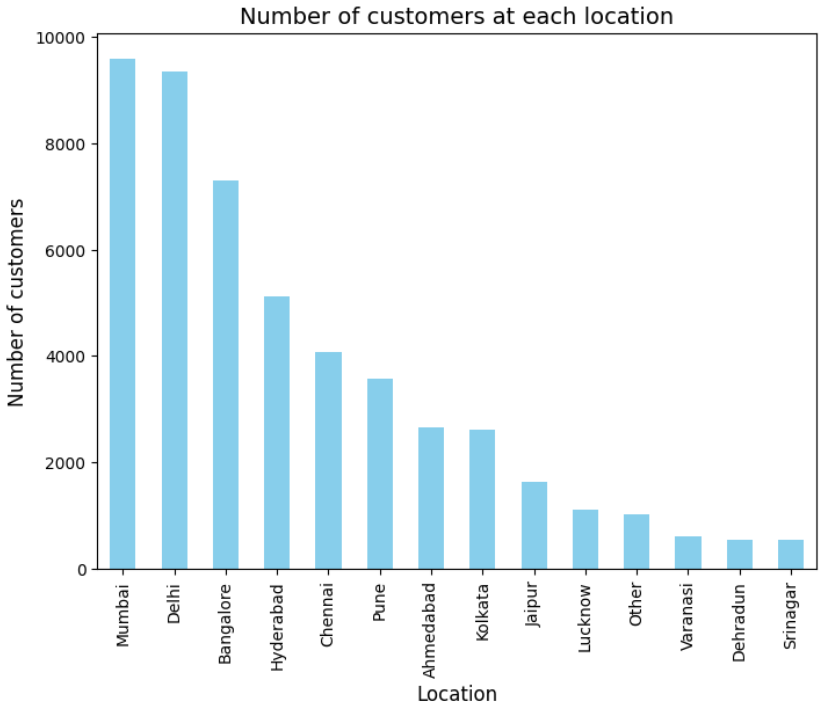
1. *Weekend* – A binary variable which indicates whether a purchase was made on a weekday (Not Saturday/Sunday) or weekend (Saturday/Sunday)
2. *Location2* – A binary variable which indicates whether a purchase is made in a metro city or not.
   Metro cities - 'Bangalore', 'Delhi', 'Chennai', 'Hyderabad', 'Mumbai', 'Kolkata'
   Non metro cities – Cities apart from the above ones.
3. *Year, month, day* - These variables extracted from the date time stamp, indicate the year, month and purchase date respectively.
4. *Day_night* – A binary variable which based on the time of the day of purchase. If the purchase is between 6 am and 6 pm, it's a 'day' purchase, else it's a 'night' purchase.
5. *Month progression* – It's a variable that tracks if a purchase is made in the beginning, middle or end of the month.
   'beginning' – If the purchase is made on or before the 7th of a month.
   'end' – If the purchase is made after 23rd of a month.
   'middle' – For the days that don't fall in the above category.
6. *Discount percent* – This variable indicates the discount percentage of a purchase. It's calculated by the formula below.
   data['Discount percent'] = (data['Discount Amount (INR)']/data['Gross Amount'])*100

Data Analysis:

Customer distribution by location:

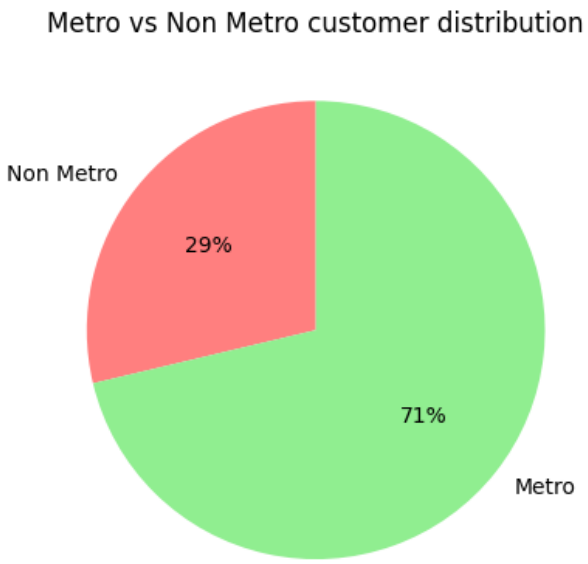
There are a total of 28921 distinct customers in all the transactions.

Their distribution across various cities is as shown below.



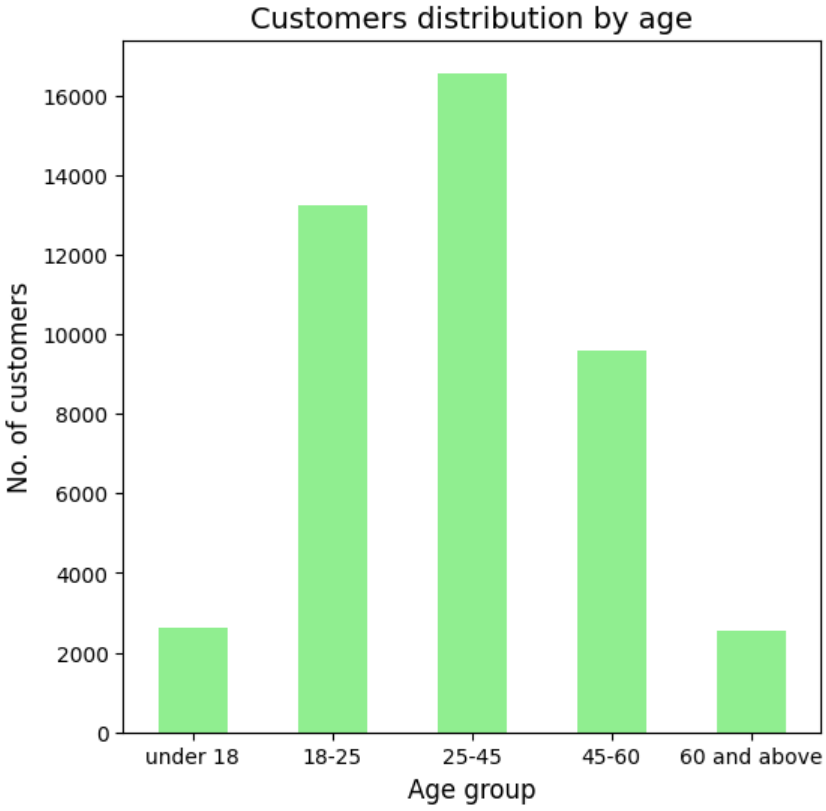
The metro cities are leading the category as shown in the above graph.

The pie chart below indicates the distribution of metro/non-metro category.



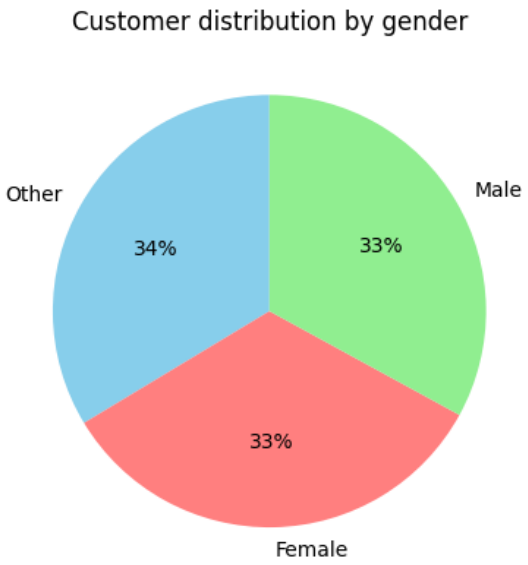
Customer distribution by age:

The plot below indicates the customer distribution by age.



It clearly shows the majority of the purchases are by customers with 25-45 age group. And as a whole majority of the chunk comes from 18-60 age group which is to be expected.

Customer distribution by gender:



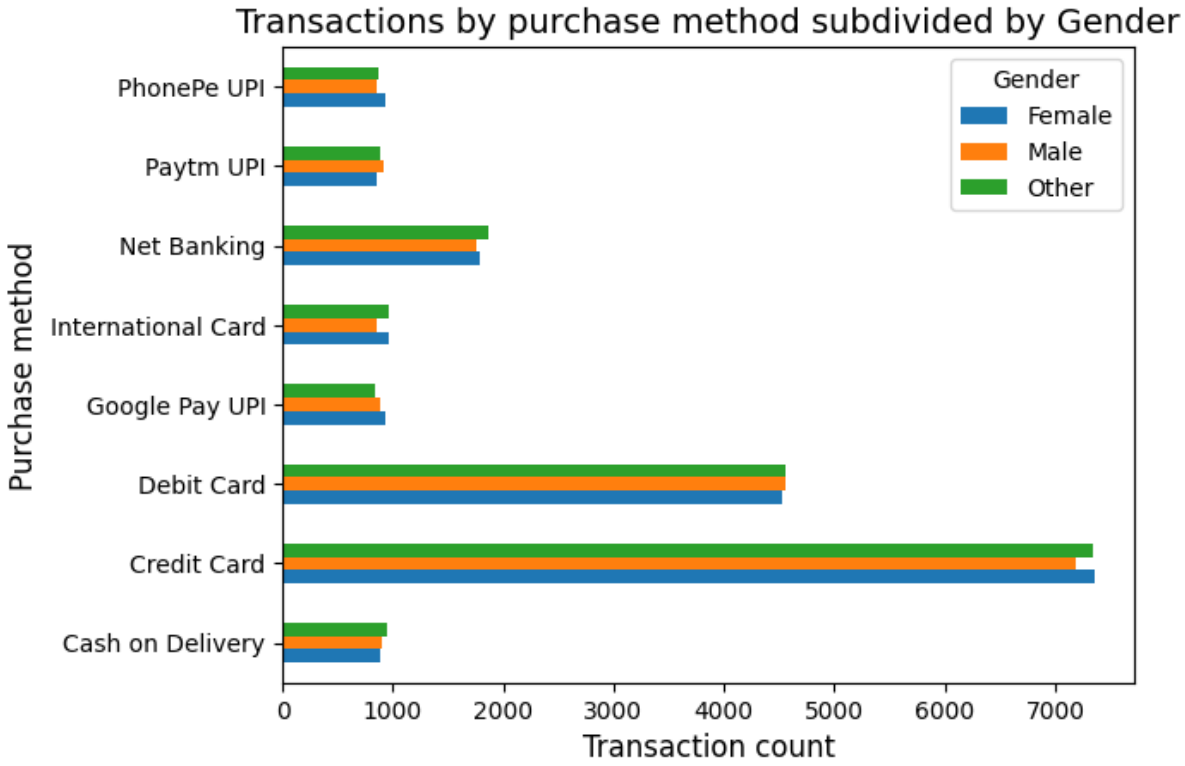
The data shows that the gender plays no role in the purchases. This needs to be further analysed to see if the transaction volume is different or is there a difference in the purchase of any particular category of product.

Gender distribution by product category:

Product Category	Beauty and Health	Books	Clothing	Electronics	Home & Kitchen	Other	Pet Care	Sports & Fitness	Toys & Games
Gender									
Female	33%	32%	34%	32%	33%	35%	34%	33%	33%
Male	33%	34%	32%	33%	31%	31%	32%	32%	30%
Other	32%	33%	33%	33%	34%	33%	32%	33%	36%

As shown in the table above, the distribution did not vary across the product categories as well.

Mode of payment distribution:



As shown above it can be observed that majority of the transactions are being done using credit card followed by debit card.

Age group distribution by product category:

Age Group	**18-25**	**25-45**	**45-60**	**60 and above**	**under 18**
Product Category					
Beauty and Health	30%	39%	20%	4%	5%
Books	28%	39%	20%	5%	5%
Clothing	30%	39%	20%	4%	4%
Electronics	30%	40%	19%	4%	5%
Home & Kitchen	29%	40%	19%	4%	5%
Other	28%	40%	21%	4%	4%
Pet Care	30%	39%	19%	5%	4%
Sports & Fitness	28%	40%	21%	5%	4%
Toys & Games	29%	42%	19%	4%	4%

There is no change in the age group distribution across multiple product categories. This indicates that there is no correlation between both the variables.

Age Group by discount availed:

Checks have been done to see if there is any variation in the discount availed, by age group

Given below are the details of average discount availed per each age group category.

```
Age Group
under 18         6.84%
18-25            7.19%
25-45            7.05%
45-60            7.22%
60 and above     7.52%
```

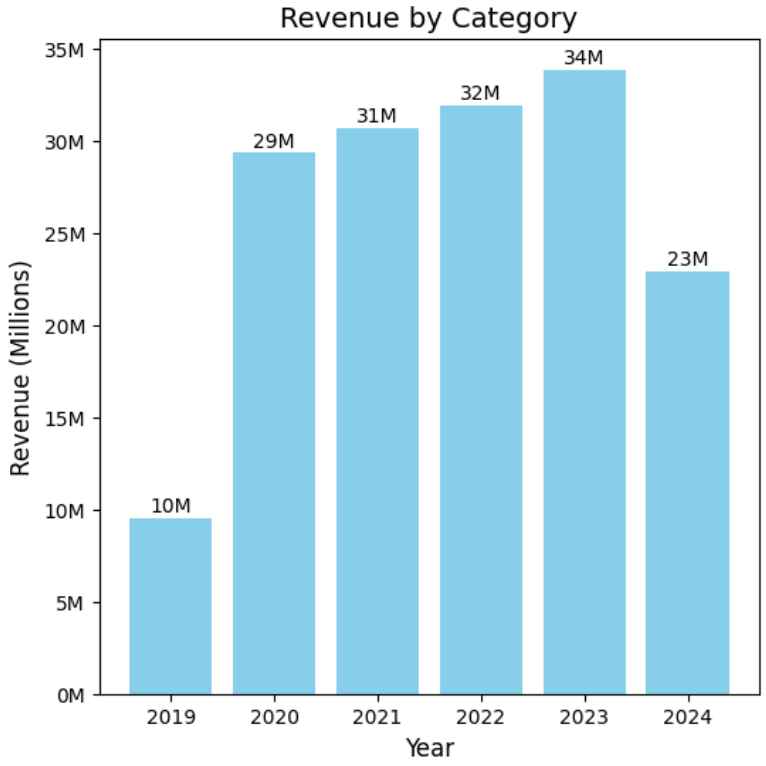
As shown above, there is not much difference between the categories.

Another statistic has been extracted to check the percentage of transactions in the which discount was availed by each category.

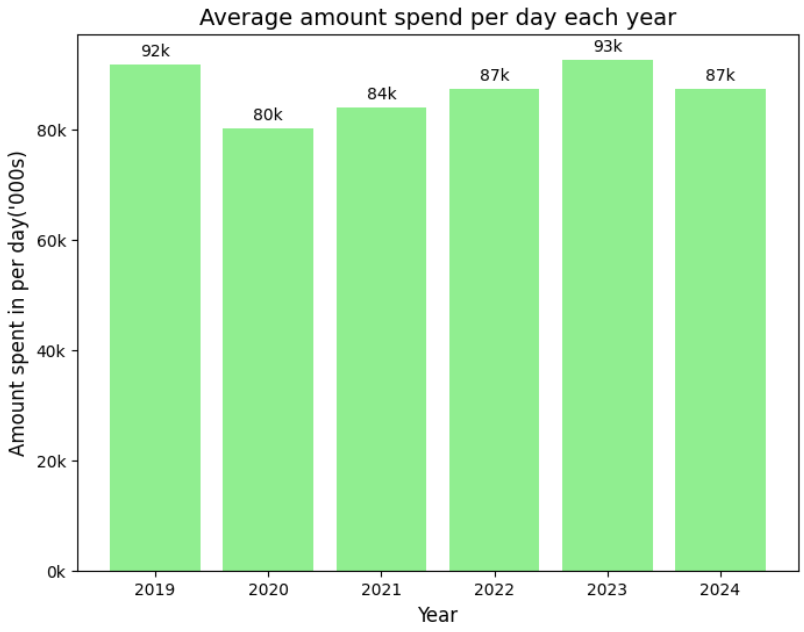
	Discount_availed	**total_transactions**	**ratio**
Age Group			
under 18	1322	2731	48.4%
18-25	7953	16236	49.0%
25-45	10747	21789	49.3%
45-60	5455	10977	49.7%
60 and above	1325	2654	49.9%

As shown in the table above, there is not much difference there as well.

Revenue trend across years:



The graph above gives the total revenue each year. It can be observed that 2019 and 2024 are kind of outliers, and there is a clear increasing trend from 2020-23. For further analysis, the average per day spend of each year has been extracted.



From the above graph, we don't have the complete data for 2019 and 2024. For the years 2020-23 a trend like total revenue can be observed.

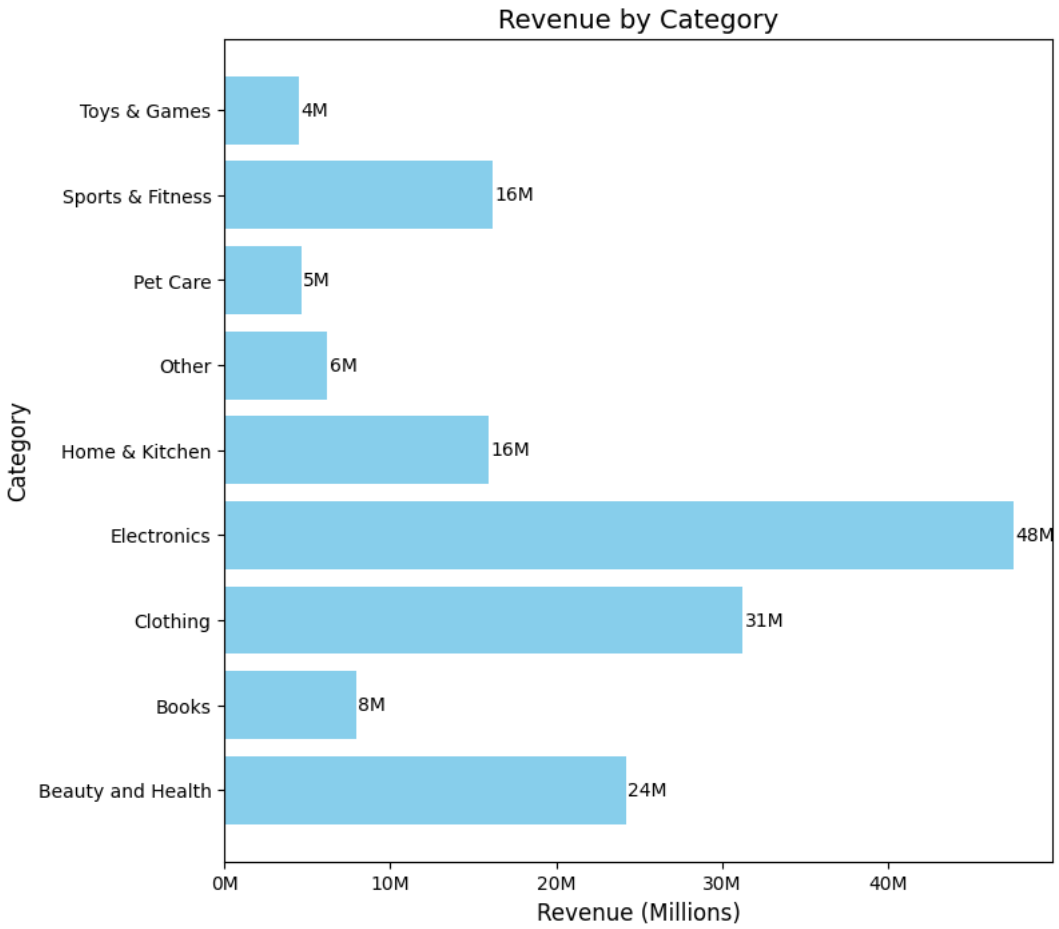
Day vs Night:

There is not much difference in the average amount spent day vs night.

Day / Night	Average amount spent ($)	% of transactions
Day	2909.81	50.3%
Night	2909.39	49.7%

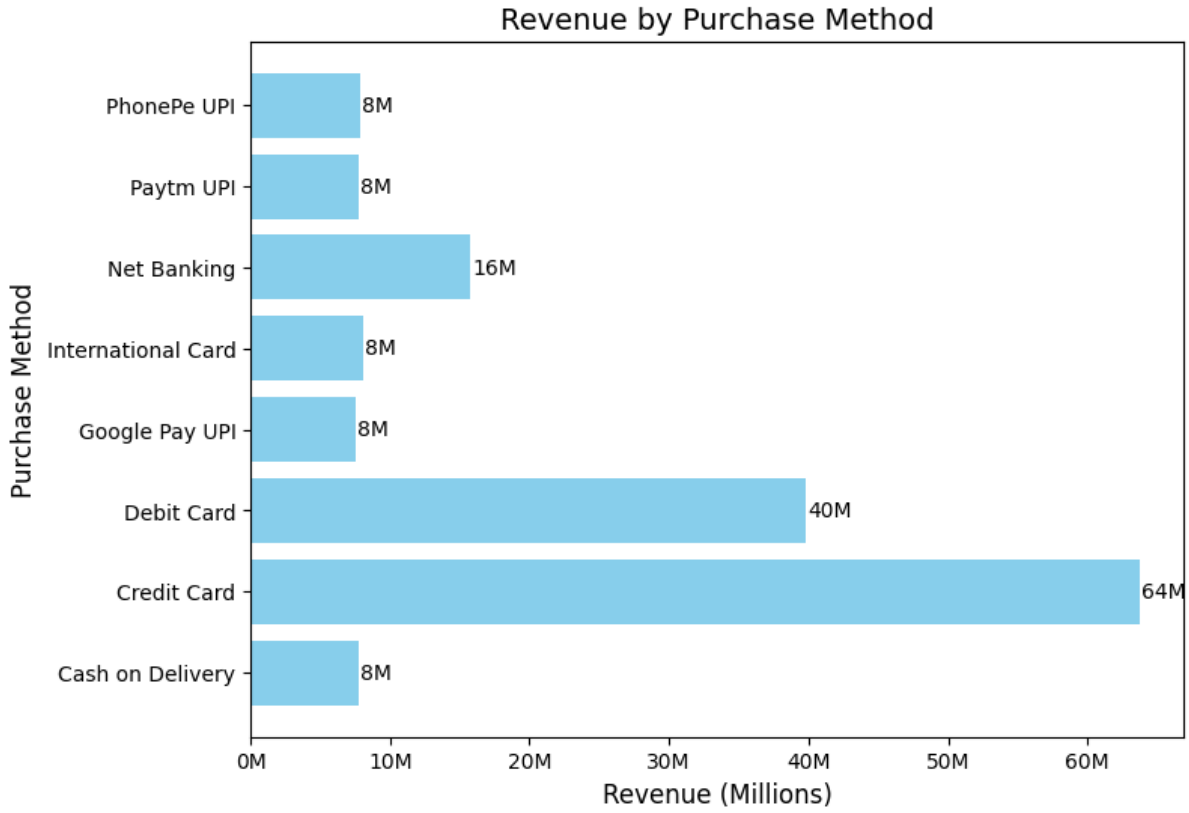
Similar trends have been observed for start, middle and end of the month.

Revenue by product category:

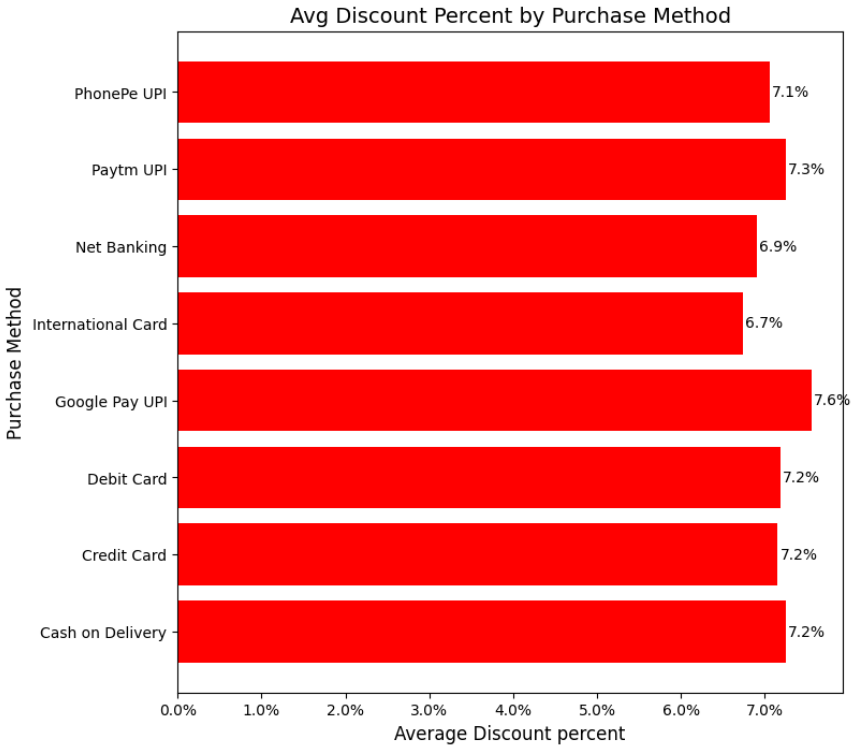


As shown above, there is clear trend in the revenue by product category with Electronics being the highest earner followed by Clothing, Beauty and Health.

Revenue and discount percentage by payment mode:



It can be seen that majority of the spend is happening using cards as compared to other payment modes.



The average discount availed in each payment mode is almost the same.

Hypothesis Testing:

Test 1: Impact of gender on transaction amount:

Null Hypothesis: The average transaction size does not differ across genders

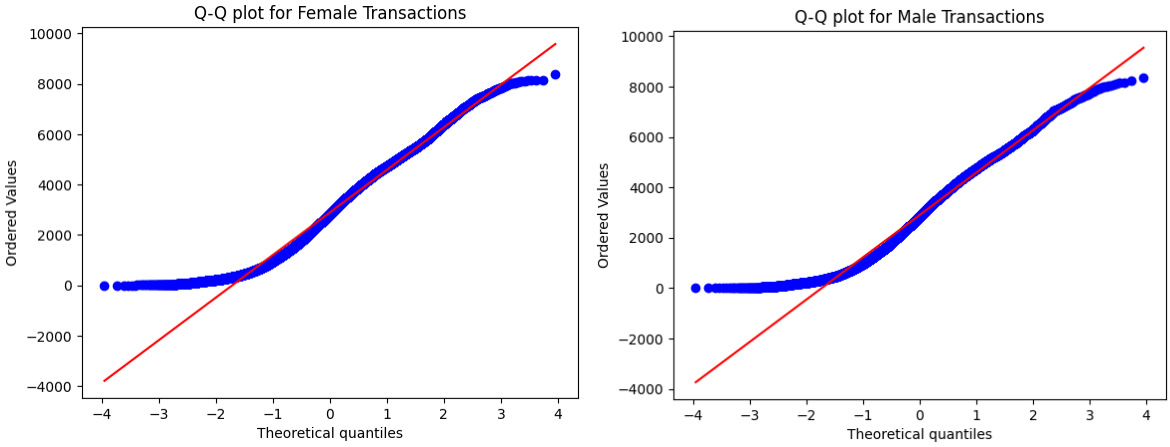
Alternate Hypothesis: The average transaction size differs across genders

For this analysis, considered only the male and female genders.

The male transactions and female transactions have been separated into male_data and female_data respectively.

The normality of both these groups has been checked, in order to determine the hypothesis test to be used. To check the same, the Kolmogorov-Smirnov (KS) test has been performed along with the Q-Q plot.

```
K-S test p-value for male transactions: 6.574907714072008e-46
K-S test p-value for female transactions: 1.4330011795450575e-49
```



From the KS test and the Q-Q plots, it can be concluded that the distributions are not normal.

Levene's test has been conducted to check the equality of variances in both the plots.

```
Levene's Test for equality of variances: statistic=0.562942313232851, p-value=0.45308205056813966
```

From the above p-value, it can be taken that the variances of both the distributions are equal.

As the distributions are not normal, Mann-Whitney U test has been used to test the hypothesis above.

```python
# As the data is not normal, using mannwhitneyu test

# Perform Mann-Whitney U test (non-parametric test)
u_stat, p_value = stats.mannwhitneyu(male_data, female_data, alternative='two-sided')

# Output the U-statistic and p-value
print(f"U-statistic: {u_stat}")
print(f"P-value: {p_value}")

# Interpretation of the results
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant difference between the transaction sizes of males and females.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference between the transaction sizes of males and females.")
```

```
U-statistic: 163705964.5
P-value: 0.5254452011918256
Fail to reject the null hypothesis: There is no significant difference between the transaction sizes of males and females.
```

As the p-value is > 0.05, we fail to reject the null hypothesis and conclude that there is no difference in transaction sizes between males and females.

Test 2: Impact of age group on product category preferences

Null hypothesis: There is no significant relationship between age group and product category

Alternate hypothesis: There is significant relationship between age group and product category

As both are categorical variables, chi-square test of independence has been performed.

```
Chi-Square Statistic: 31.63613654893272
P-value: 0.48489280711859517
Degrees of Freedom: 32
Expected Frequencies:
[[2464.34221413  816.1734238  3233.94539136 4884.50239947 1620.70428595
   639.74383584  479.13619063 1646.07909978  451.37315903]
 [3307.19096475 1095.31921231 4340.01208009 6555.08886315 2175.01390038
   858.54757571  643.00926692 2209.06735065  605.75078603]
 [1666.11754647  551.80682884 2186.43868939 3302.36405759 1095.7422362
   432.52451873  323.93926858 1112.89789839  305.16895582]
 [ 402.83100741  133.41489694  528.63334988  798.43984776  264.92665527
   104.57502712   78.32147388  269.07452148   73.78322025]
 [ 414.51826723  137.28563811  543.97048927  821.60483204  272.61292221
   107.6090426    80.59379999  276.88112968   75.92387887]]
Fail to reject the null hypothesis: There is no significant relationship between age
group and product category.
```

From the p-value above, we fail to reject the null hypothesis and conclude that there is no significant relationship between age group and product category.

Test 3: Purchase method preference by location:

Null hypothesis: There is no significant relationship between purchase method and location

Alternate hypothesis: There is significant relationship between purchase method and location.

As both are categorical variables, chi-square test of independence has been performed.

```
Chi-Square Statistic: 7.267788244223159
P-value: 0.40154059457434416
Degrees of Freedom: 7
Expected Frequencies:
[[ 2131.74811996   603.25188004]
 [17046.97006638  4824.02993362]
 [10632.23989189  3008.76010811]
 [ 2056.14315921   581.85684079]
 [ 2169.94031662   614.05968338]
 [ 4222.96574549  1195.03425451]
 [ 2061.59918731   583.40081269]
 [ 2069.39351316   585.60648684]]
Fail to reject the null hypothesis: There is no significant relationship between
purchase method and location.
```

From the p-value above, we fail to reject the null hypothesis and conclude that there is no significant relationship between purchase method and location.

Conclusion:

1. The majority chunk of customers belongs to the metro cities in the ratio of about 3:1
2. The age group distribution shows a spike in the customers of age group 25-45
3. The gender distribution is equal in the entire data.
4. The product category by gender distribution there is no correlation between both the variables.
5. The mode of payment trends shows a large chunk of customer spend is by cards.
6. There is no correlation between age group and product category.
7. Revenue trend across years reveals that there is a constant increase in the transaction from 2020-23.
8. Electronic products contribute to highest revenue followed by Clothing, Beauty & Health categories.

Next Steps:

A model can be built on this data to predict the net amount spent in a transaction.