Analyzing Ecommerce In Pakistan

Imaan Fatima Imran

Imaan Fatima Imran 18060

Business Intelligence Final Project

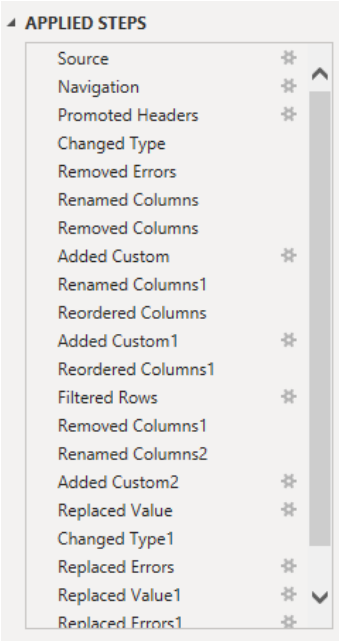
Introduction to Dataset

The dataset I have chosen for my project is Pakistan's Largest Ecommerce Dataset taken from opendata.pk. It contains over half a million records on many different fields related to online transactions. The data has been compiled with the collaboration of many online ecommerce sites, and can be very useful for new startups to analyze trends and brainstorm ideas for better engagement

This dataset initially contained 25 columns and over 500,000 rows, with columns including information on item id, order status, order creation date, stock keeping unit, price, discount, grand total, market value, category name, etc. These columns provide insight into the common trends in ecommerce, and many relationships can be observed.

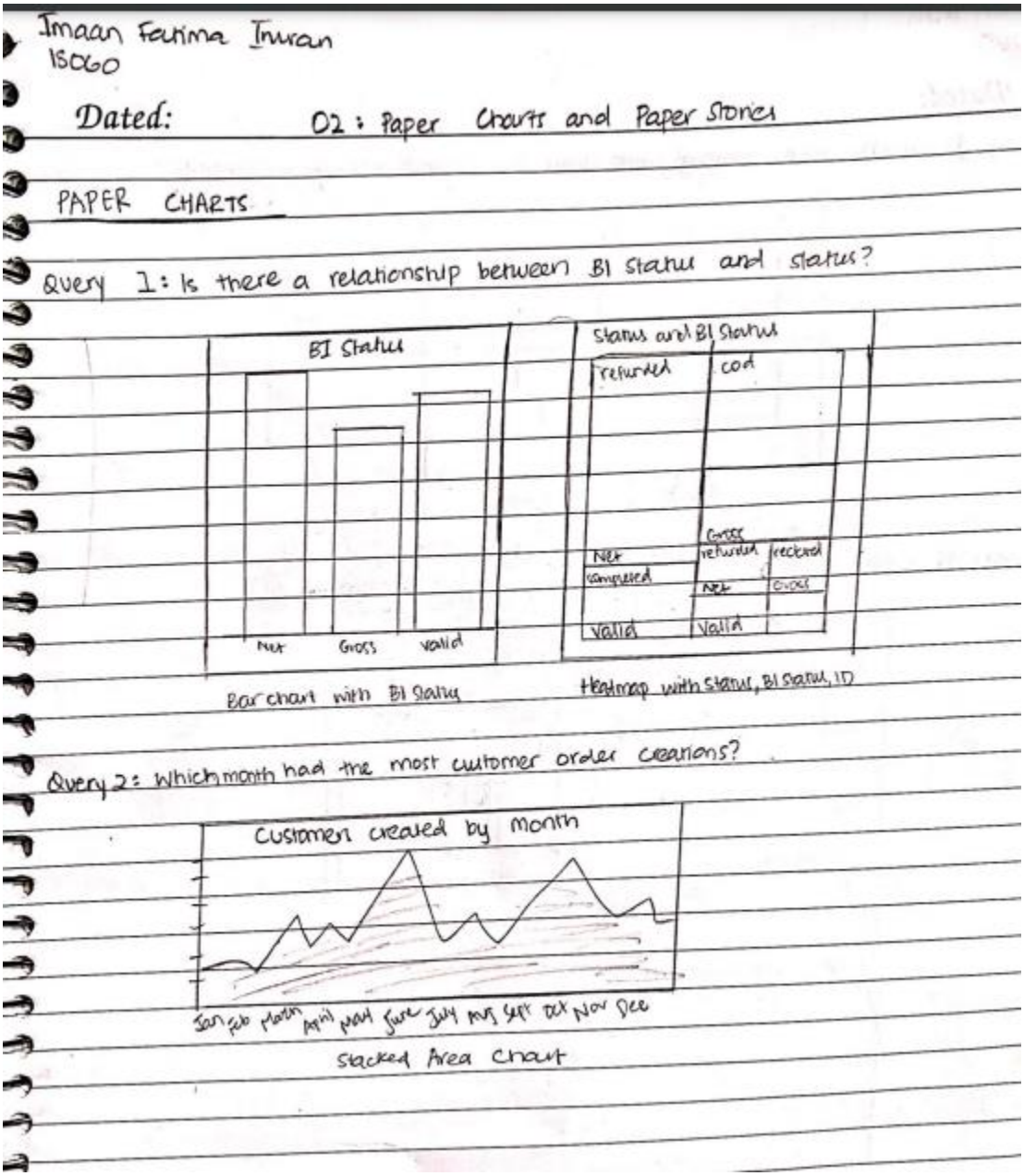
Overview of Data Wrangling

The data provided was quite well organized for the most part and did not seem messy or inaccurate at first glance. However, there were a few logical errors in the dataset that needed to be taken care of for accurate analysis to be performed. The right shows a screenshot of the BI Query Editor showing all the transformations that were done:

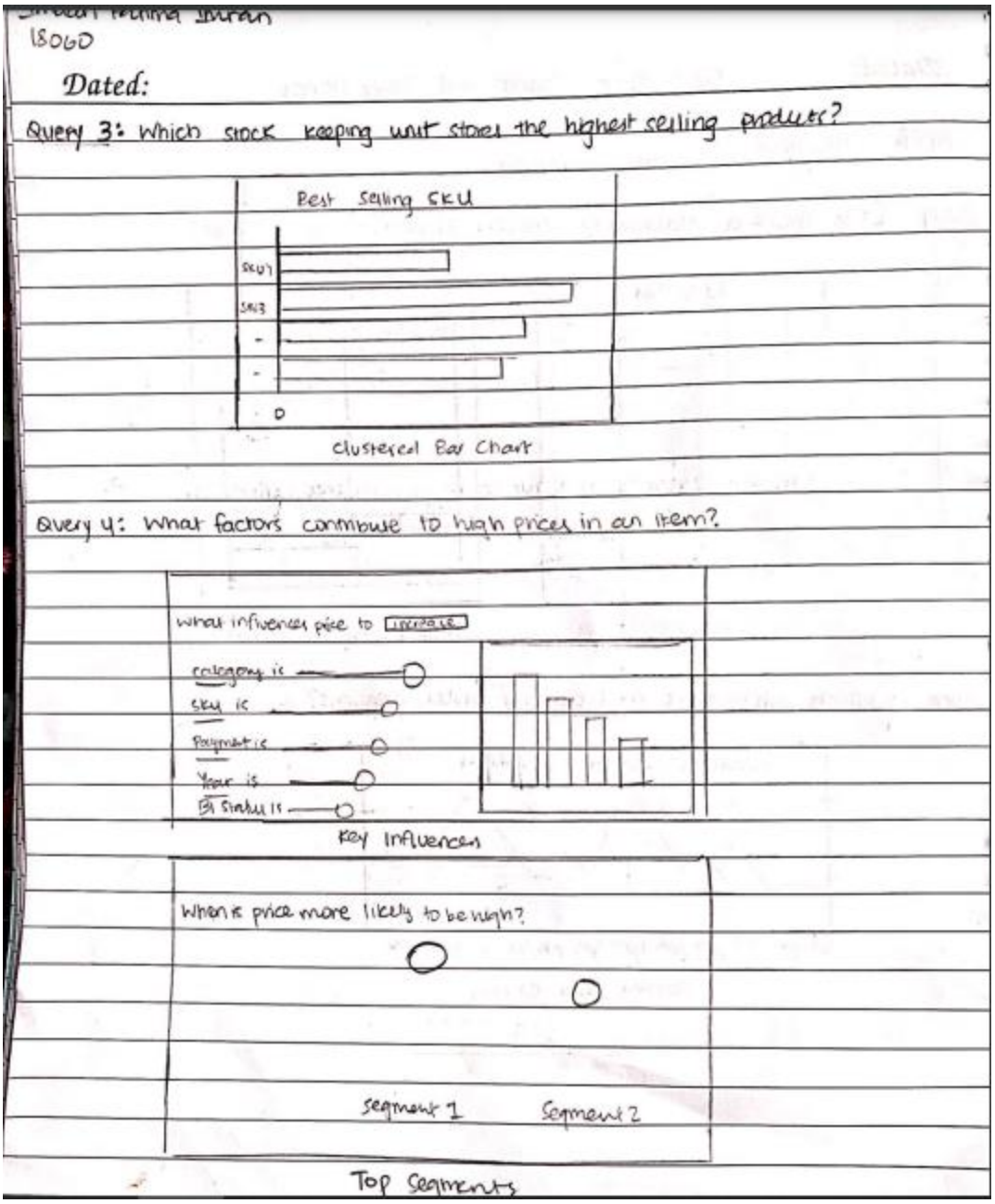


One of the most glaring issues was that of the column MV, where over 2000 rows were giving errors. As this column was not useful for analysis, it was simply dropped. Other transformations were filtering, missing and incorrect value replacement, and grouping certain columns for better data representation.

O2: Paper Charts and Stories



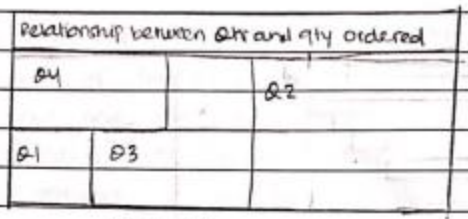
Imaan Fatima Imran 18060



Imaan Fatima Imran
18060

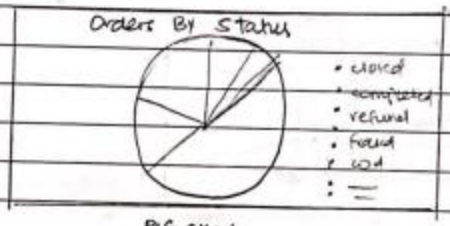
Dated:

Query 5: What is the relationship between the Quarter the customer joined and quantity ordered?



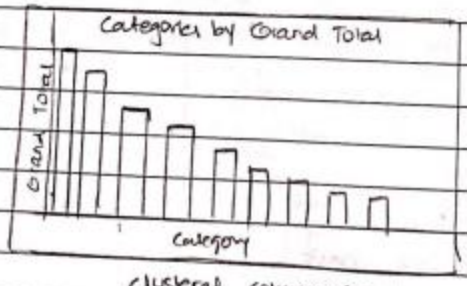
Heatmap

Query 6: Is there any relationship between refunded orders and Grand Total?



PIE Chart.

Query 7: Which categories sell the most?



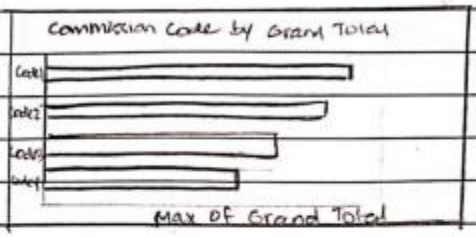
clustered column chart

Imaan Fatima Imran 18060

Imaan Imran
18060

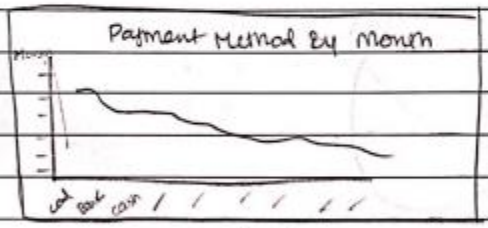
Dated:

Query 8: Which commission code earned the most sales?



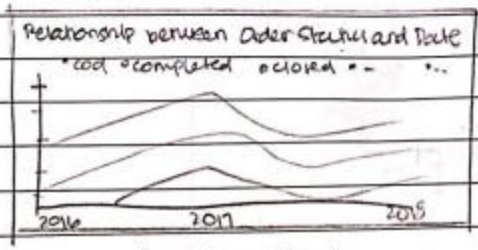
Clustered Bar Chart

Query 9: Which month was most popular for each payment method?



Line Chart

Query 10: Does working date have any effect on completion?



Area Line Chart

Imaan Fatima Imran
18060

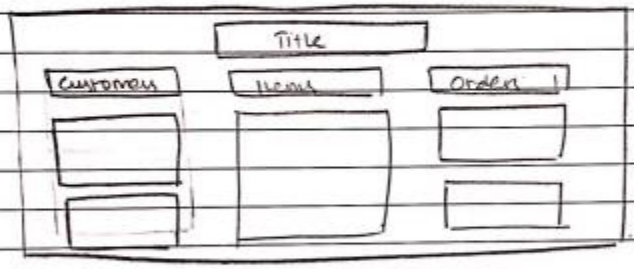
Dated:

Step 3: Brainstorming dashboards and stories

Dashboards:

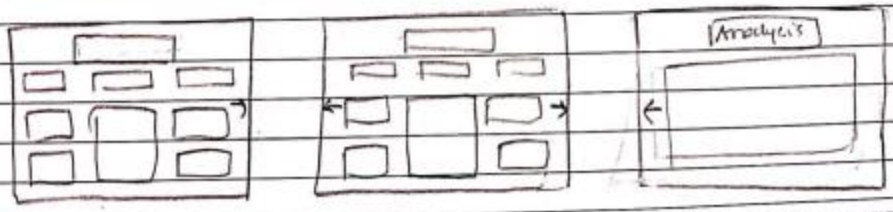
After creating paper charts, I noticed a few patterns. All the charts can fall into one of three categories: customers, items and orders. To better represent my data and its findings, I can create separate dashboards for each, with further divisions as initial statistics, further statistics, and analytics. There will be 5 charts per dashboard.

Dashboard format-



Stories

This will be a detailed analysis, so just one dashboard won't be enough. To illustrate this and tell a story, I can have multiple dashboards combined with arrow navigation enabled so the entire idea is clear.



The arrow keys shown can create story-like effect in Power BI

O3: Snapshot of Charts with Analysis

Figure 3.1.1 shows the distribution of customers by their Order Status. As shown by the chart, the majority of customers (40.07%,) have order status at *completed*, followed by *cancelled, order refunded,* and *received.* A negligible number of customers had order statuses *refund* and *cod*. This shows that although many orders are being completed, a very large amount are also refunded.

This dataset is over half a million records, so 29.27% translates to 145,000 orders being cancelled. Further analysis will uncover why.

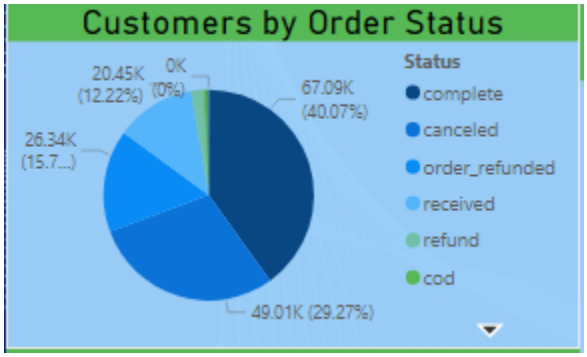


Figure 3.1. 1

Figure 3.1.2 is an initial statistic to show the earliest customer since the collection of this data. This is represented as a card and will be placed at the front page of the dashboard. We can see that the first customer was on Friday, July 01, 2016. It will be useful to keep this in mind when analyzing further statistics on longevity of customers, customer acquisition, and customer retention.

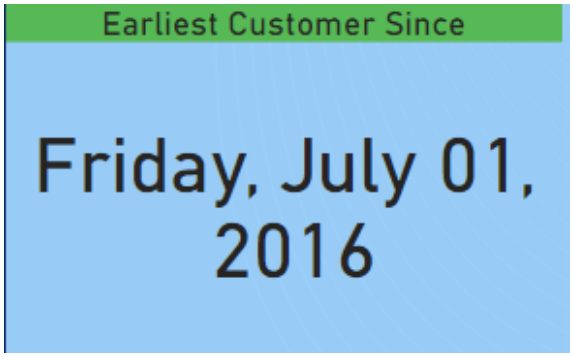


Figure 3.1. 2

Figure 3.1.3 shows the relationship between the quarter of the year that the customer joined, and the quantity purchased. This heatmap shows that Quarter 3 and Quarter 4 joining purchase more quantity of orders. This might be due to later months being holiday season, so people spend and buy more. The other quarters are slightly less popular but still evenly distributed.

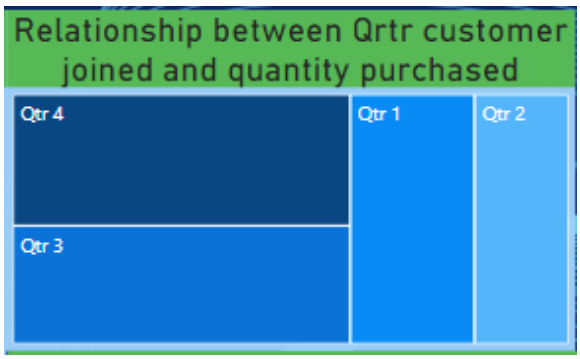


Figure 3.1. 3

Figure 3.1.4 shows a gauge representing the number of unique customers that are found to be purchasing items in the dataset. This number is 115,323 unique customers out of the total 230, 650 customers. This translates to about half of the customers being new and the rest returning, which is a good customer retention percentage.

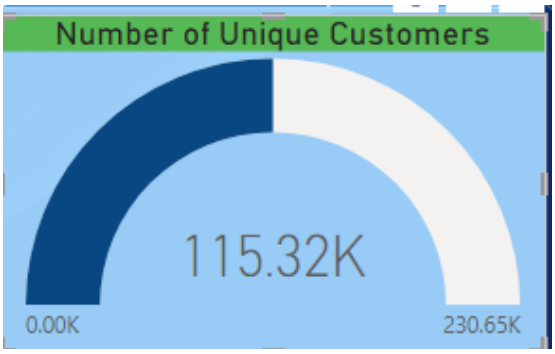


Figure 3.1. 4

Figure 3.1.5 is a stacked bar chart showing the relationship between quantity bought and year that the customer joined the site. As you can see, customer joining in the year 2017 bought the most items, even more than those that joined a year earlier. The year following, 2018 has significantly less purchases. The reason for this could be election season in Pakistan which led to greater prices and less online activity.

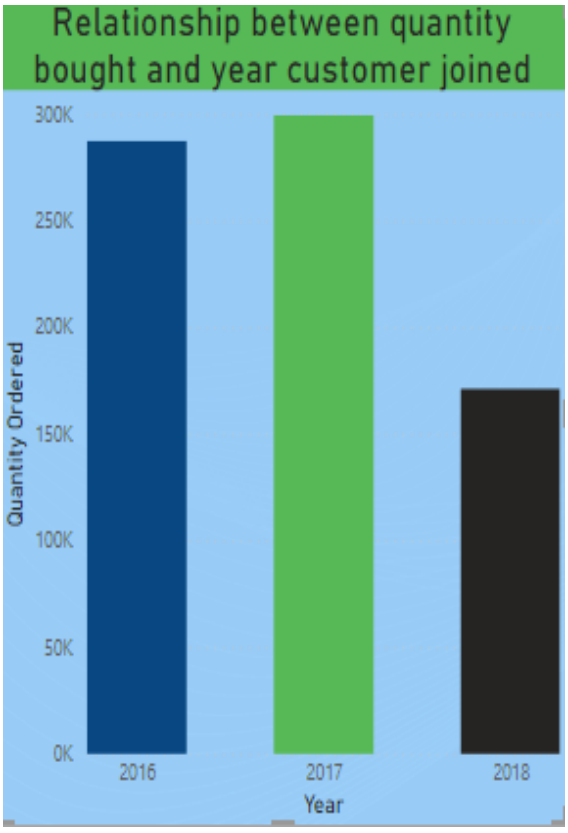


Figure 3.1. 5

Figure 3.1.6 is a stacked bar chart showing the distribution of first-time customers by year. The chart shows that the year 2018 experienced the largest number of customers joining. This may be due to a greater increase in ecommerce with greater access to internet across Pakistan. However, customer joining does not always mean more quantity bought, these customers browse more and buy less as shown in Figure 3.1.5

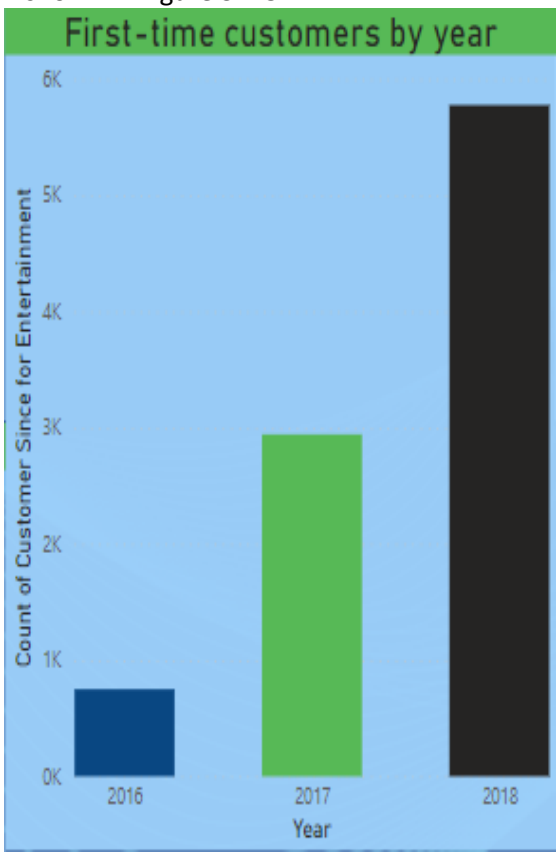


Figure 3.1. 6

Figure 3.1.7 shows the distribution of discounts that customers have taken, with the overwhelming majority of customers making purchases with no discounts, at over 74.25%. The rest of the discount amounts are all less than 2%, making up the remaining portion. This may be a reflection of the dataset, where that many discounts just weren't available or it could show that Pakistanis buy anyway regardless of whether a discount is offered.

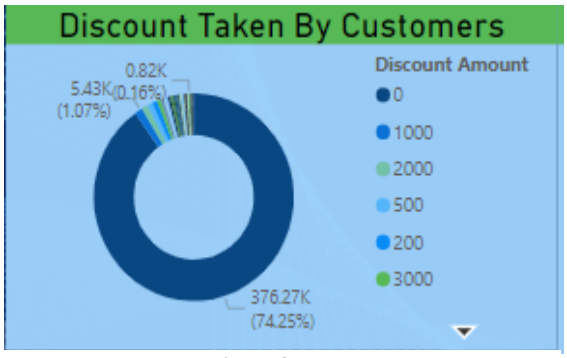


Figure 3.1. 7

Figure 3.1.8 builds upon Figure 3.1.3 where the first customer to join's date was shown as Friday July 1, 2016. The last customer since joined Wednesday, August 1, 2018 a little over 2 years later. This is a reflection of the duration of the dataset and also shows that customers kept joining the site, so customer acquisition was pretty good.

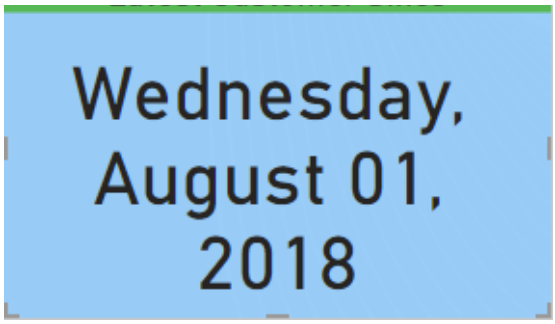


Figure 3.1.8

Figure 3.1.9 is a pie chart showing the distribution of customers by payment method they chose to pay with. The majority of customers opted for cod, at 46.65%, while Payaxis and Easypay followed. These are local payment tools used by people all over the country, but mostly cash on delivery was chosen. This must be due to internet banking still not being trusted and fear of fraud. Proper internet banking like bankalfalah and ubl are also options opted for but are less common.

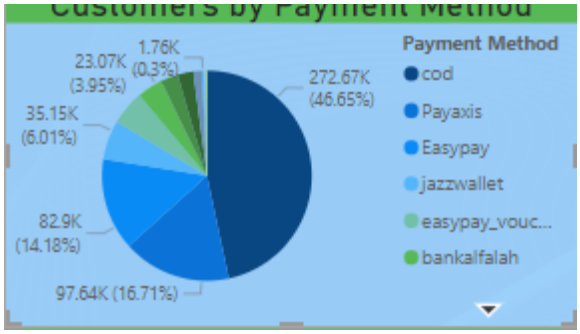


Figure 3.1. 9

Figure 3.1.10 is an area chart showing the customers created by month. The most customers across the three years have joined in the winter months of November and December, declining sharpy at the new year. This may be due to this being the season with the most discounts and holiday offers so more customers join and browse products.

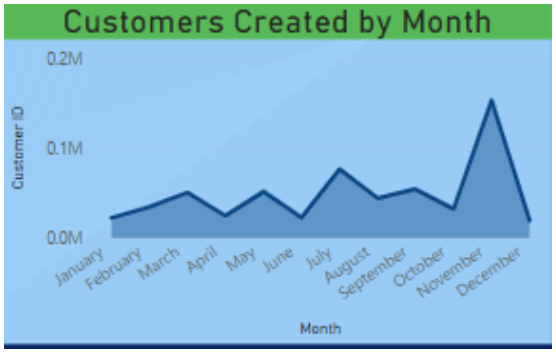


Figure 3.1.10

Figure 3.2.1 shows the distribution of items by their working date. We can see that most items were purchased in the fourth quarter of the year, followed by the second, first and, third quarters respectively. This can be due to holiday season towards the end of the year, where discounts are more rampant and shopping activity is increased. The third quarter is usually during summer vacation so it would have less activity.

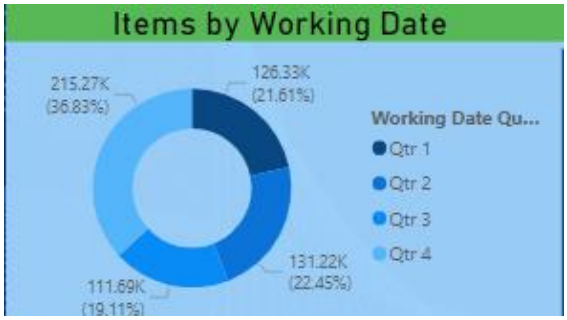


Figure 3.2.1

Figure 3.2.2 is a representation of which stock keeping unit (SKU) has the best performance in terms of item sales. This is one of our BI queries and is illustrated with a clustered bar chart. The SKU BAGGEM5A70 shows the most sales, so it is possible that this SKU stores expensive items like electronics.

Figure 3.2.2

Figure 3.2.3 is a clustered column chart showing the relationship between the quantity bought and the year Item was bought. The figure shows that the year 2017 saw the highest quantity of items being bought, which might be due to lower prices resulting in customers wanting to buy more.

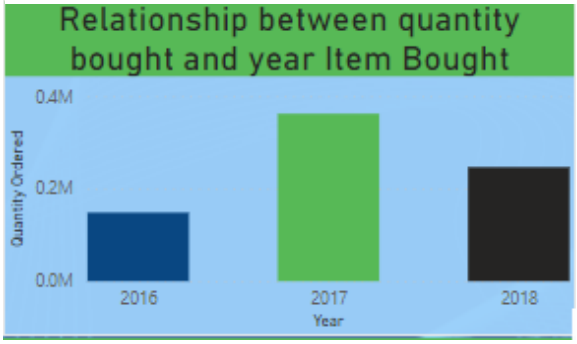


Figure 3.2.3

Figure 3.2.4 is a funnel chart expressing item ID by the month that it was purchased. The months are represented as numbers, with each number corresponding to its specific month. The chart shows that the month 11 (November) experiences the most item sales, at 62,000 of the whole dataset. This may be due to sales such as Black Friday, Thanksgiving, all taking place in November leading to more sales.

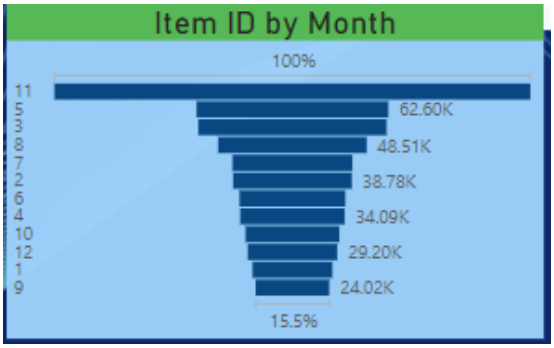


Figure 3.2.4

Figure 3.2.5 answers the BI query what is the relationship between BI status and Status? We have observed this relationship using a treemap across Item ID. The figure shows that the orders which are complete are all net, canceled are gross, and received and refunded are valid. This shows that those orders which are complete are being registered as net in the accounting software while those that are cancelled are at some loss, so they are gross. Valid orders are mostly those that are still pending.

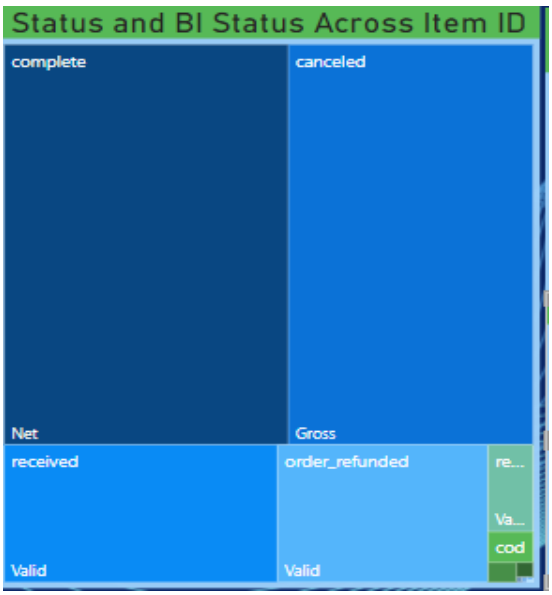


Figure 3.2. 5

Figure 3.3.1 shows the distribution of orders by their payment status. We can see that the majority of orders are complete, followed by cancelled, and then refunded. Cancelled orders as seen in previous charts are usually of a lower price and have cod selected as payment method. The popularity of cod is what motivates people to cancel easily, and result in such high cancellations in the dataset.

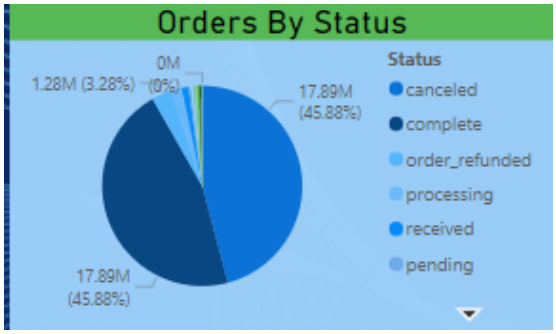


Figure 3.3. 1

Figure 3.3.2 is an area chart showing the relationship between order status and working date, one of our BI queries. This shows that canceled orders are highest in 2017, and decline slowly in 2018, compared to completed orders which peak in 2017 and decline sharply in 2018. This is alarming and shows that the website needs to work on their order fulfillment.

Figure 3.3. 2

Figure 3.3.3 is a 100% stacked bar chart showing commission code and quantity ordered by the sales. The chart shows that the majority of products do not have any commission code associated with them, but R-KHS 102986 appear the most in the grand total.

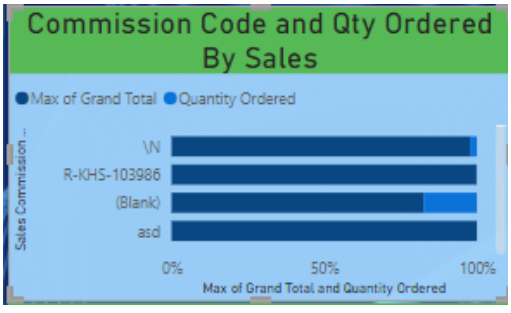


Figure 3.3. 3

Figure 3.3.4 is a line graph representing the payment method that is most common for each month. The graph shows that cod is most opted for in summer months while winter months see more internet banking. This may be due to weather conditions or overall interest.

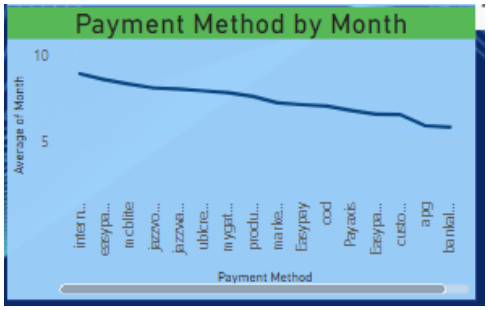


Figure 3.3. 3

Figure 3.3.5 shows a bar graph representing the distribution of categories by grand total purchased. By far, the most popular category is Beauty and Grooming, followed by mobiles and tablets and appliances. This must be due to these items being more expensive typically. As for beauty and grooming, it is very convenient to shop online for such items as going out and purchasing such a small item can be a hassle if the items are commonplace and also sell out quickly.

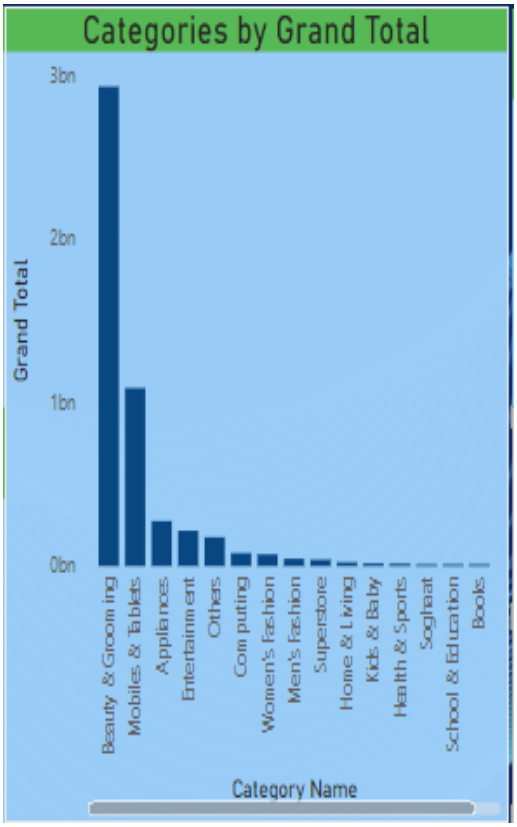
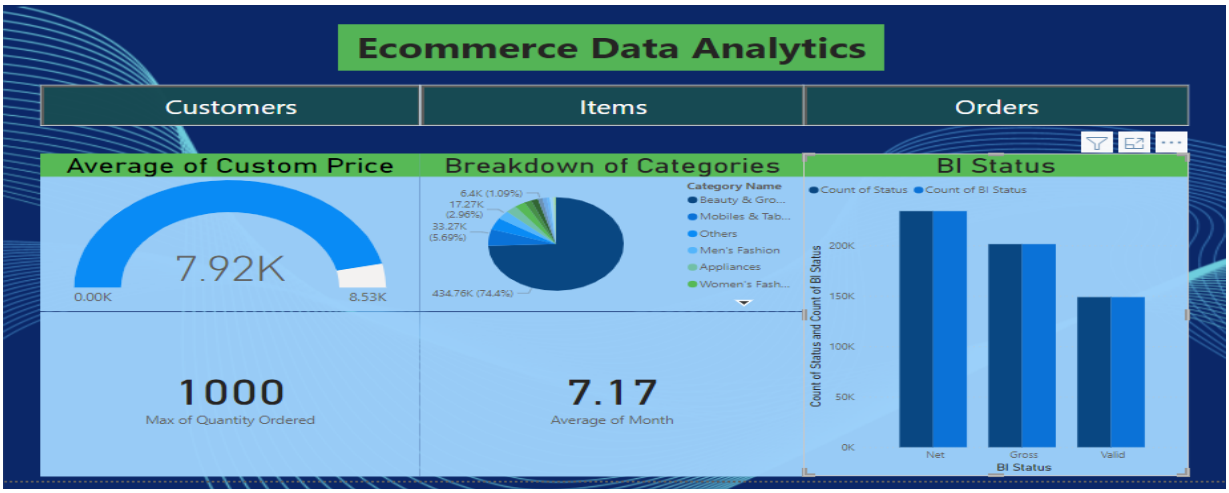


Figure 3.3. 5

Imaan Fatima Imran 18060

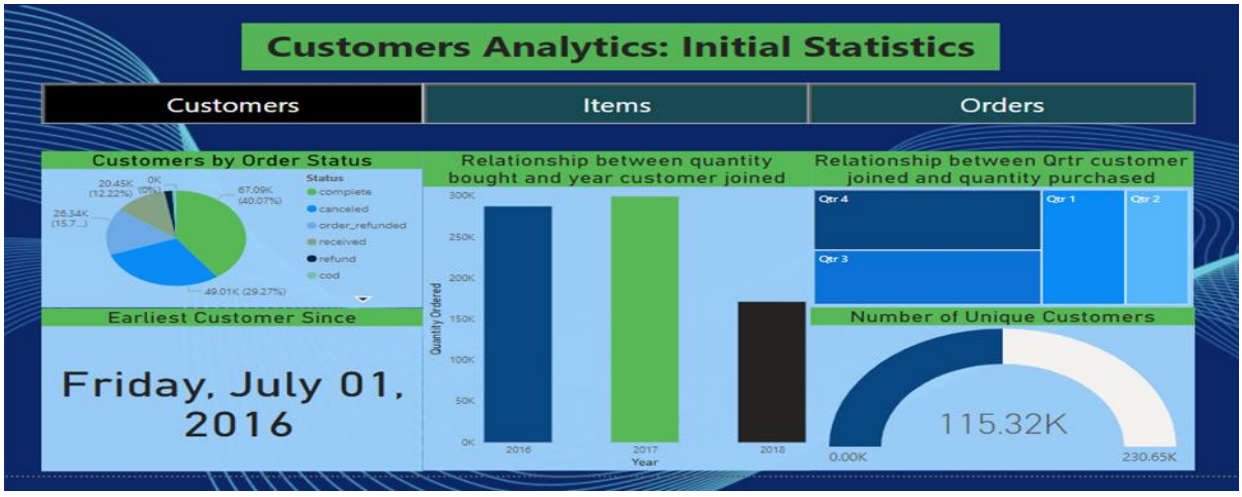
O4: Snapshot of Dashboards with Analysis

Dashboard 1: Ecommerce Data Analytics



The above figure is a dashboard showing the overview of the Ecommerce Dataset. There is average of custom price, breakdown of categories, distribution of BI status, maximum of quantity ordered, and the average month. We can analyze form this dashboard that beauty is the most popular category, July-August months are the busiest, and the highest quantity of order is at 1000 items. This dashboard on every click drills the data down to show representation for each specific record

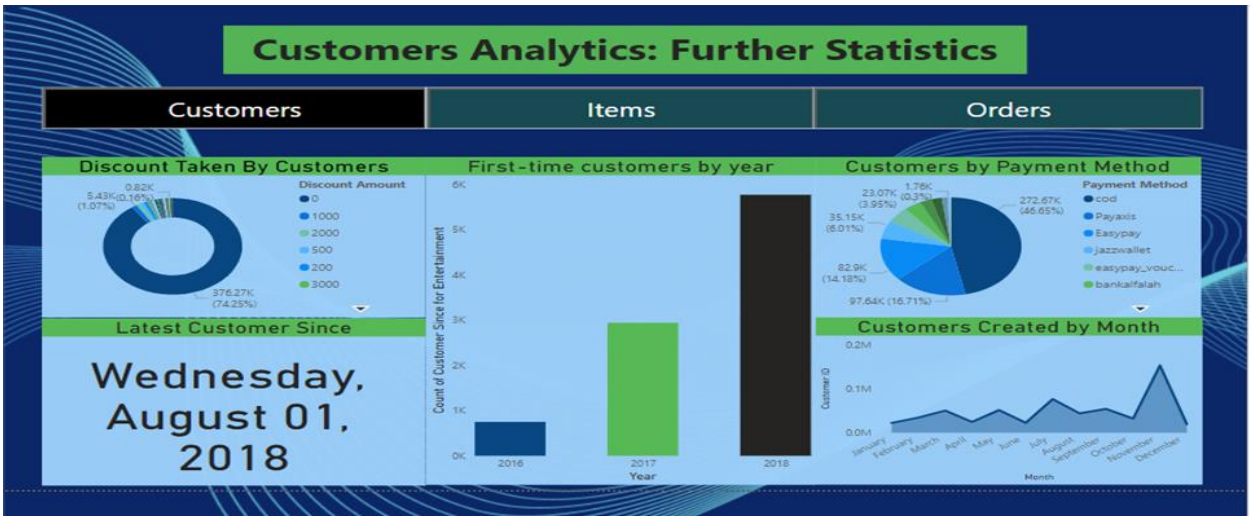
Dashboard 2: Customers Analytics: Initial Statistics



The above figure is a dashboard showing the initial statistics of customer data. We can see that customers usually complete their orders, but there are also high number of cancellations, The quarters 3 and 4 have the most online traffic, and 2017 was a popular joining year. Unique customers make up about half of the dataset, showing good customer retention and acquisition. This dashboard gives *superficial figures while the following dashboard is more detailed*

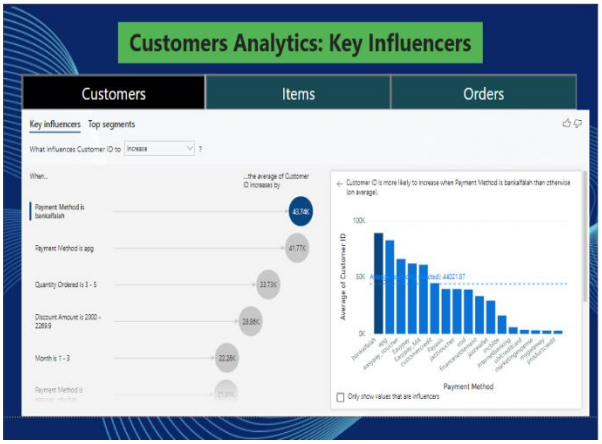
Imaan Fatima Imran 18060

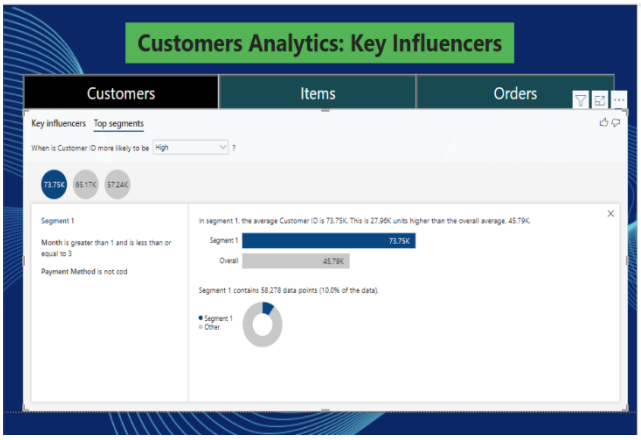
Dashboard 3: Customers Analytics: Further Statistics



The above figure is a dashboard showing further statistics on analytics related to the customer. We can see that mostly 0 discounted orders are popular, the last joining date was August 1, 2018, the months of November and December are most frequent for customer joining, and cod is the overwhelming majority choice of customers. All this information shows that Pakistani customers are not very trusting of ecommerce, but there was an improvement in 2018, and the holiday season and offers definitely attract them the most.

Dashboard 4: Customers Analytics: Key Influencers

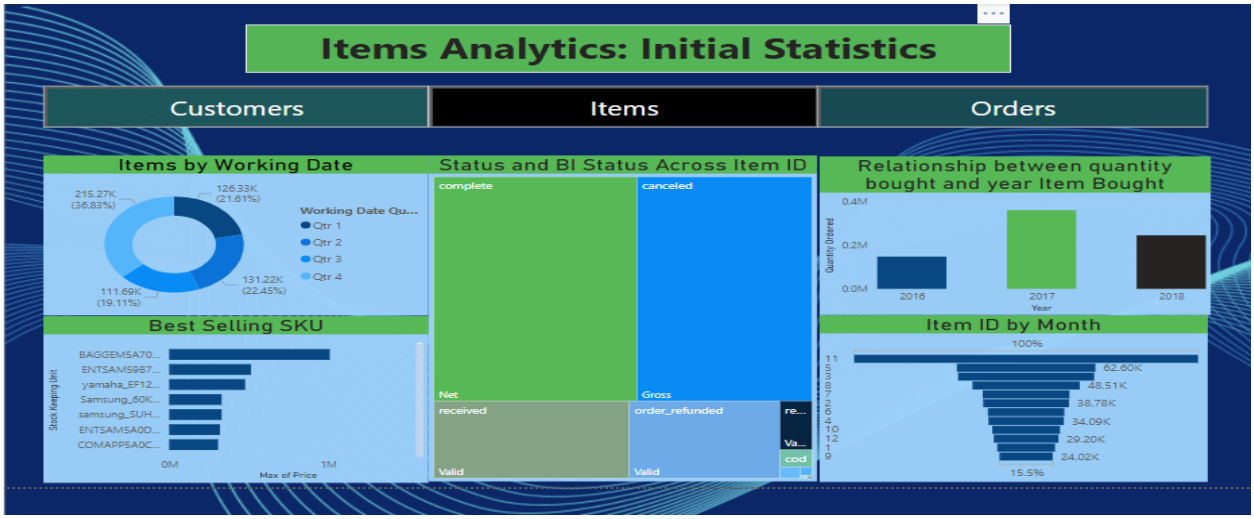




The above figure is dashboard showing the key influencers that affect customer ID increasing, decreasing, as well as the top segments. With each click, the chart drills down to show the affect that each field has on the other. It is a very informative chart showing distribution of data points for many different fields, such as payment method is alfalah, discount amount range, quantity ordered, early months of the year, etc. This shows the various customer acquisition strategies a company can focus on.

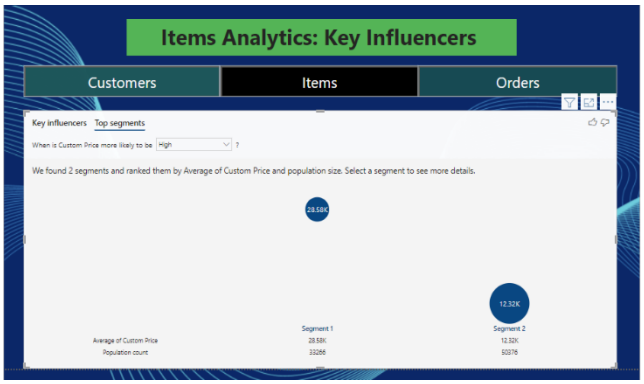
Imaan Fatima Imran 18060

Dashboard 5: Items Analytics: Initial Statistics



The above figure is a dashboard showing the initial statistics for the item analytics. The dashboard has a total of 5 graphs, which are a donut chart, treemap, clustered bar chart, column chart, funnel chart. The dashboard shows that items are mostly bought in th month of November, in the year 2016, and have a complete or canceled status. They are bought in the fourth quarter of the year and the best selling SKU is also shown. This dashboard is interactive and can be drilled down on each click.

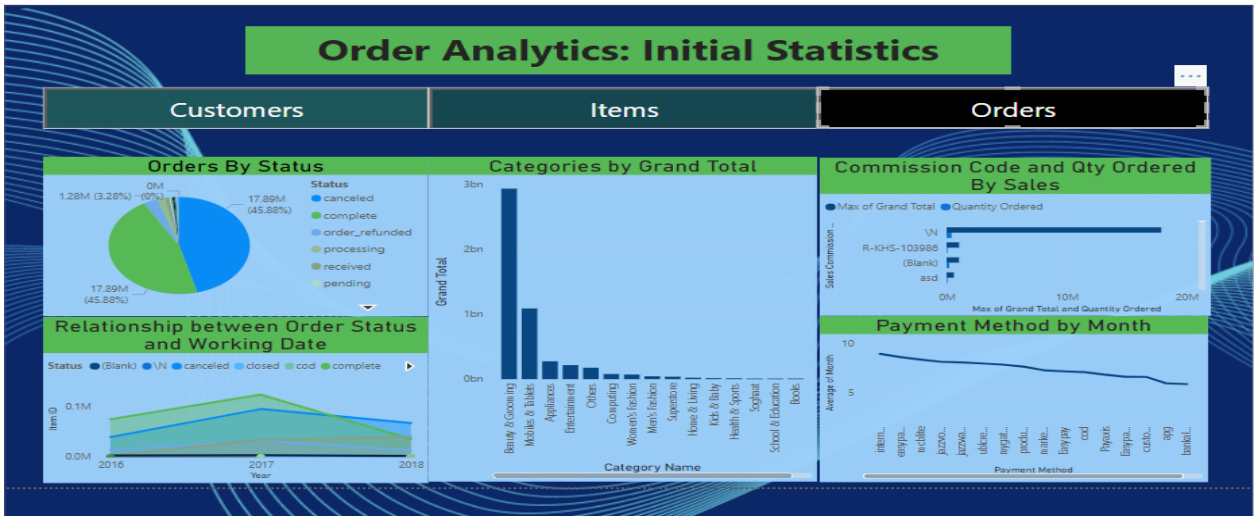
Dashboard 6 : Items Analytics: Key Influencers



The above figure shows the two views of the dashboard Item Analytics: Key Influencers. This dashboard shows the key influencers that cause price of items to increase or decrease, identifying category mobiles and tablets, payment method alfalah, year 2018, and easypay_voucher. This shows that the customers of alfalah are most inclined to buy expensive products and 2018 may have resulted in less purchases, but the prices were quite high, possibly due to election season in Paksitan. The second view shows the top segments, identifying the different segments where price may decrease or increase.

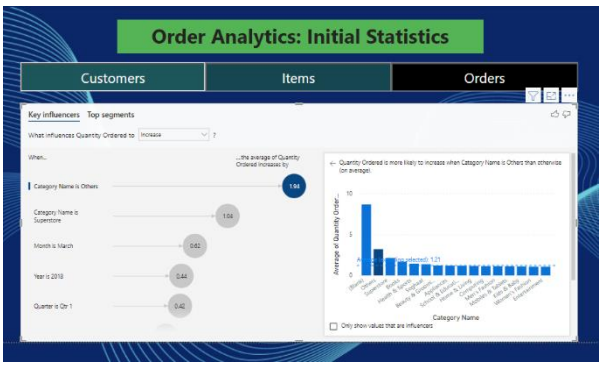
Imaan Fatima Imran 18060

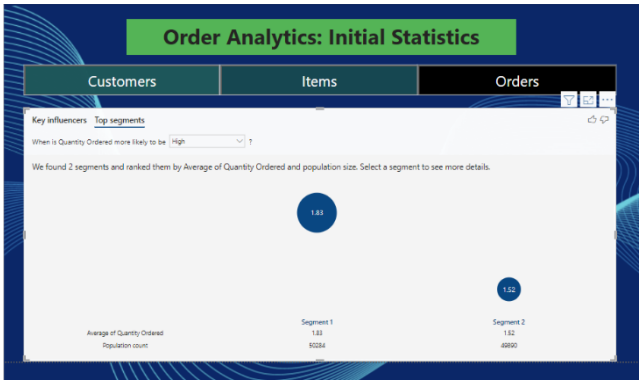
Dashboard 7: Orders Analytics: Initial Statistics



The above figure is a dashboard showing order analytics dashboard showing its statistics. This dashboard comprises of five charts, a pie chart, area graph, line graph, clustered bar chart, and column chart. The dashboard can be accessed by clicking on orders button at the top and with each click all charts on the dashboard drill down and show the effects of a particular field. WE can see that beauty and grooming are the most popular orders, November is the most busy season, most orders do not have commission codes, and working date does have an affect on order status.

Dashboard 8: Orders Analytics: Key Influencers



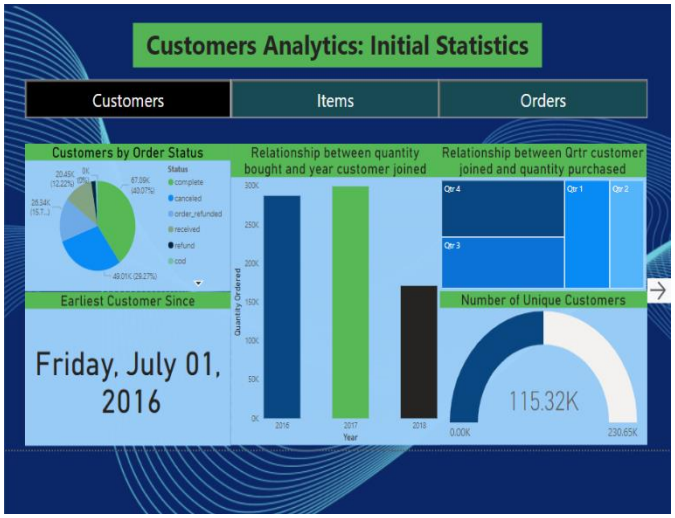


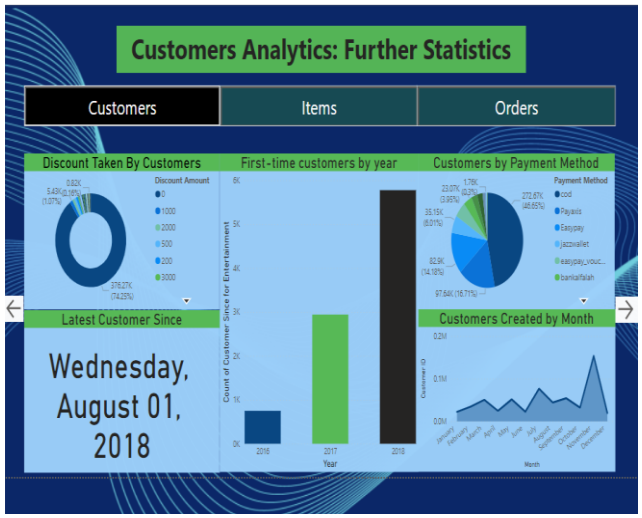
The above figure is a dashboard showing the key influencers for order analytics. The first view shows what causes quantity of orders to increase or decrease, identifying the month of march, year 2018, category superstore, and quarter 1. The analysis for this can be that the superstores usually have people buying in bulk, with ration or home supplies, all bought more in the beginning of the year. The top segments are also shown in the second view. This is interesting as although beauty and grooming and mobile and tablets are most popular categories in terms of sale, the quantity ordered is always less, which may be due to their high prices.

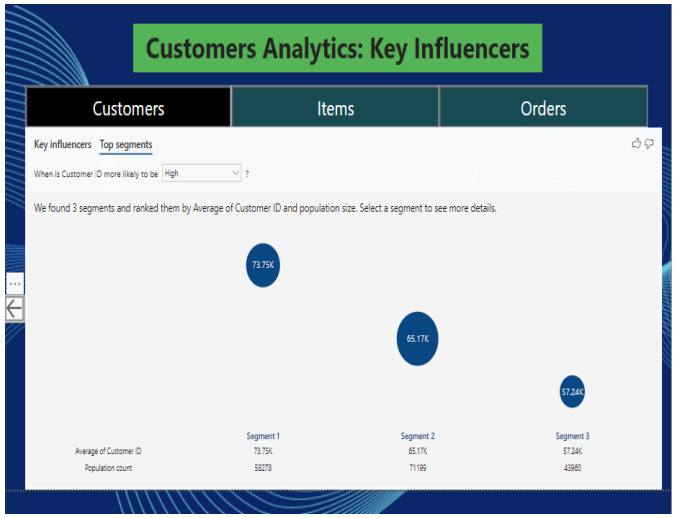
Imaan Fatima Imran 18060

O5: Snapshot of Stories with Analysis

Story 1: Customers Analytics



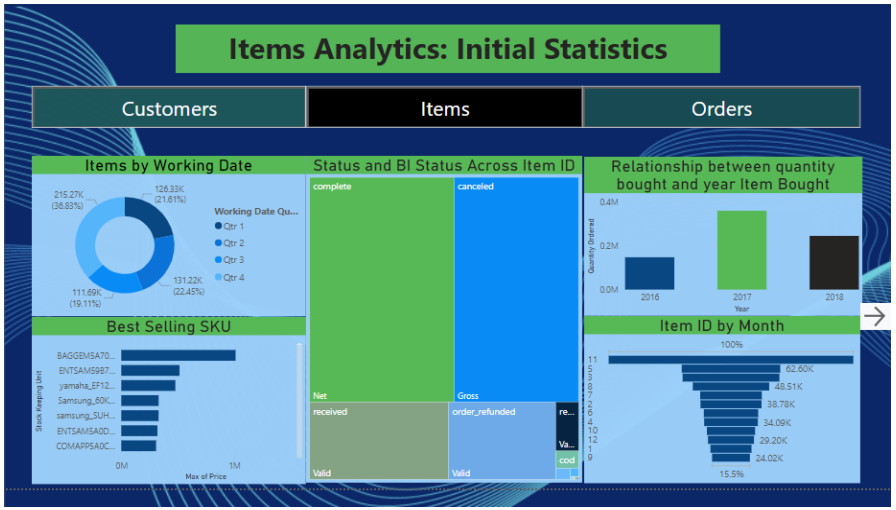


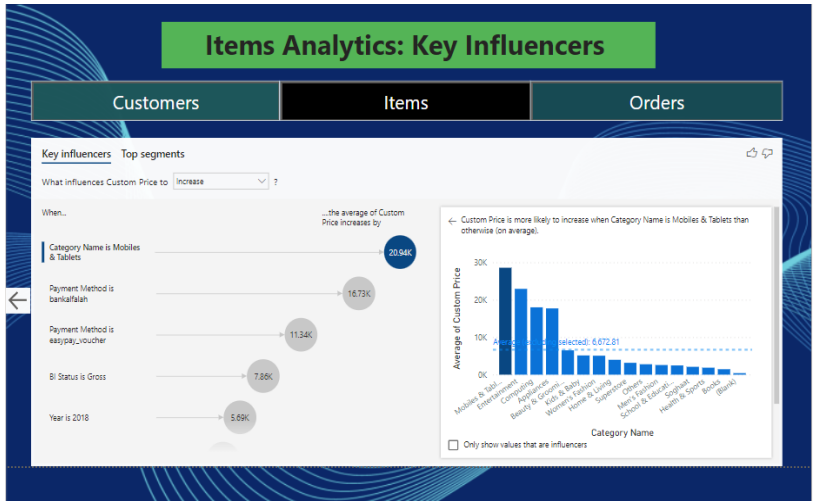


The above shows the story formed by the three dashboards. The arrows when clicked go to the previous and next pages, This story shows a complete overview of the customer's analytics from initial statistics, to further statistics, to key influencers that affect it. The analysis shows that the first customer joined on Friday, July 01 2016, and last on Wednesday, August 01, 2018 a total of 2 years approximately. It also shows that the majority of customers opt for cod, have net BI status, complete orders, and prefer winter months to do their shopping.

Imaan Fatima Imran 18060

Story 2: Items Analytics

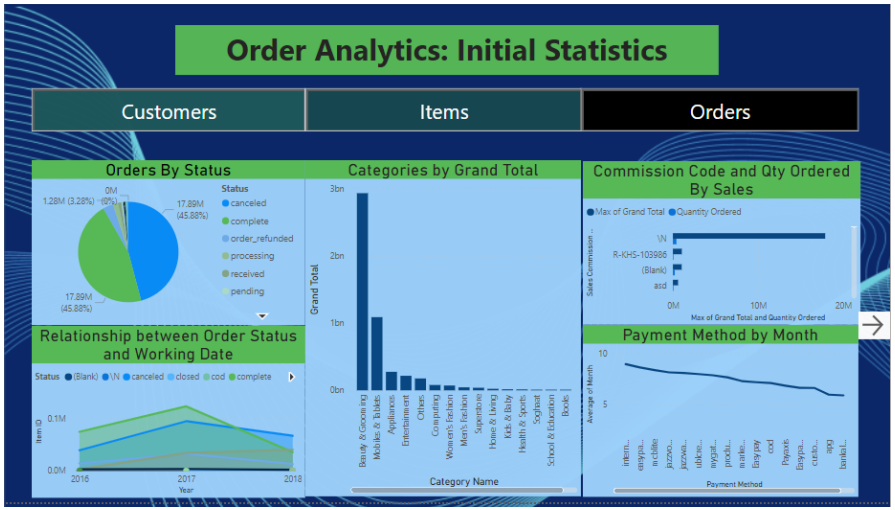


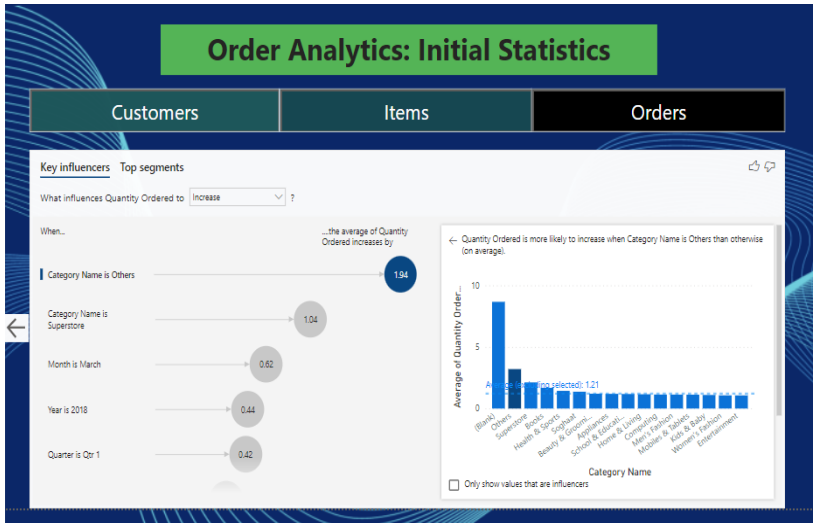


The above shows the story formed by the two dashboards. This can be accessed by the arrows on the right and left. The story shows a complete overview of the items analytics from initial statistics, to key influencers and top segments. It is completely interactive and each click drills down to give a complete picture.

Imaan Fatima Imran 18060

Story 3: Order Analytics





The above shows the story formed by the two dashboards. This can be accessed by the arrows on the right and left. The story shows a complete overview of the order analytics from initial statistics, to key influencers and top segments. It is completely interactive and each click drills down to give a complete picture.

Imaan Fatima Imran 18060

O6: Overall Analysis

Query 1: Is there a relationship between status and BI status?

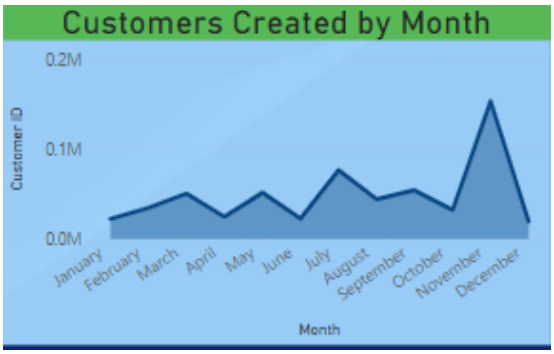
This query was answered by multiple charts, indicating that yes there is a relationship, with the following obersvations:

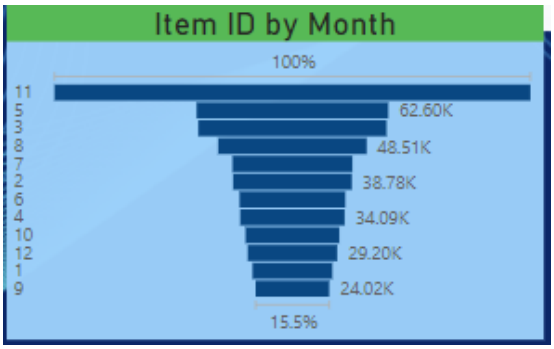
- All transactions marked as either 'complete' or 'closed', are marked as 'Net' for 'BI Status'
- All transactions marked as 'received','paid','cod','exchanged' or refunded in any way are marked as 'Valid'
- All transactions marked as either 'canceled' or incomplete in any way are marked in 'Gross' category

This shows that BI status is an accounting related check, with closed and complete orders being Net or counted from profit. Those which are pending or resulted in an exchange are Valid, so they have not finished. Gross transactions are those that have been at some loss.

Query 2 Which month had the most order creations?

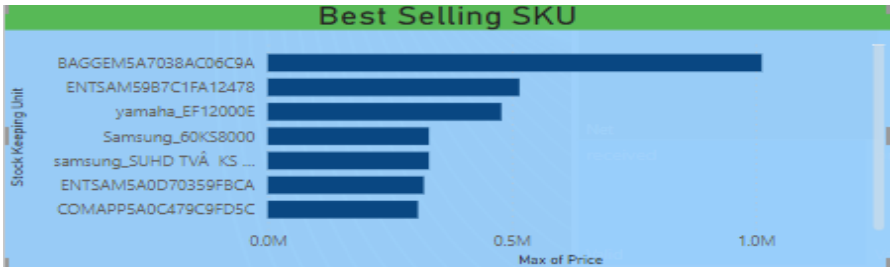
The answer to this query is month 11, or November. This was illustrated by the following charts:





Query 3 Which stock keeping unit stores the highest selling products?

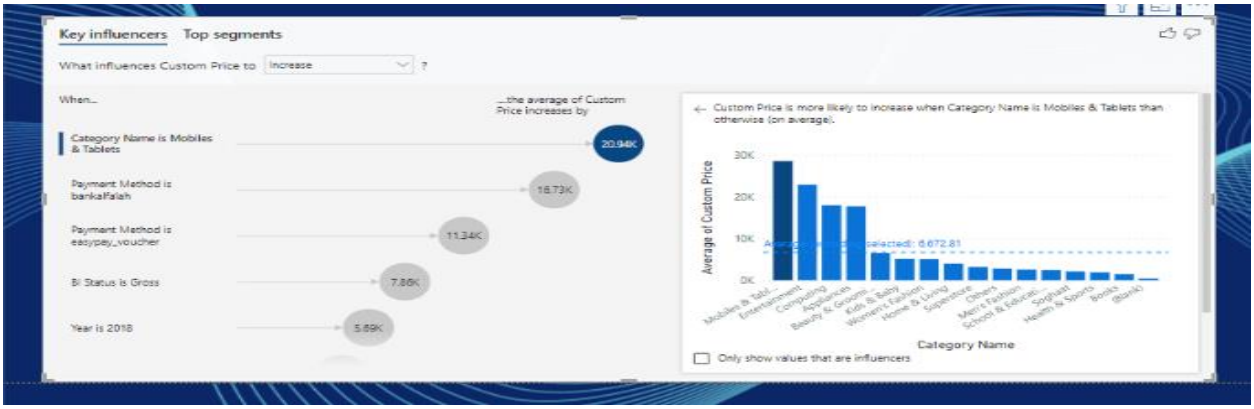
The answer to this query is the SKU. BAGGEM5A7038AC06C9A This is illustrated below:



Imaan Fatima Imran 18060

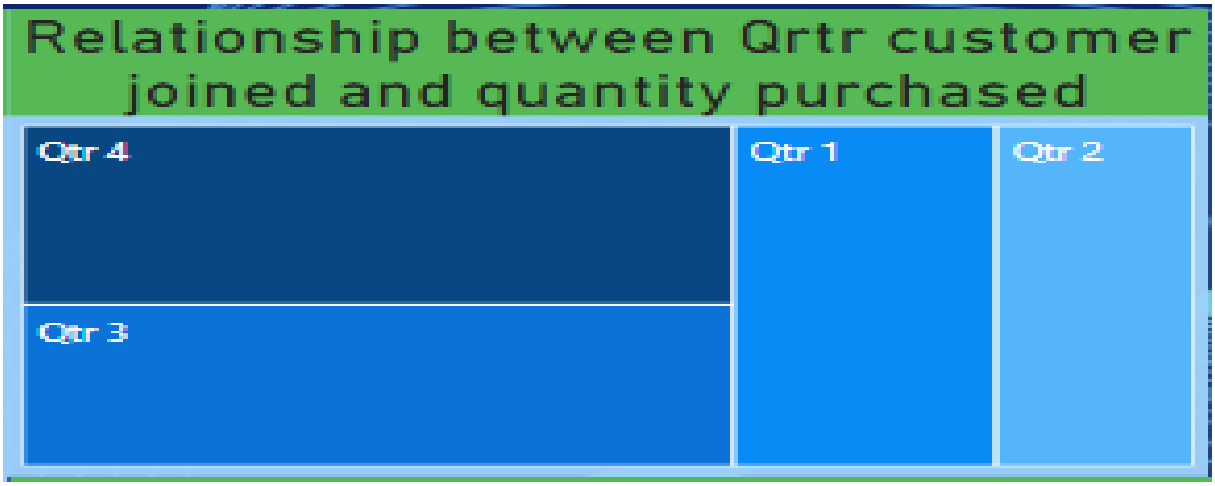
Query 4: What factors contribute to high prices in an item?

The answer to this query is the category mobiles and tablets, payment method alfalah, year 2018, and easypay_voucher. This is illustrated by the following chart:



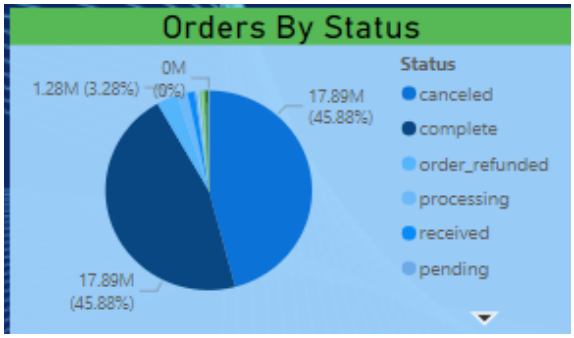
Query 5: What is relationship between Qrtr customer joined and quantity ordered?

The answer to this query is that the Quarter 3 and Quarter 4 joining purchase more quantity of orders..



Query 6: Is there any relationship between refunded orders and high grand total??

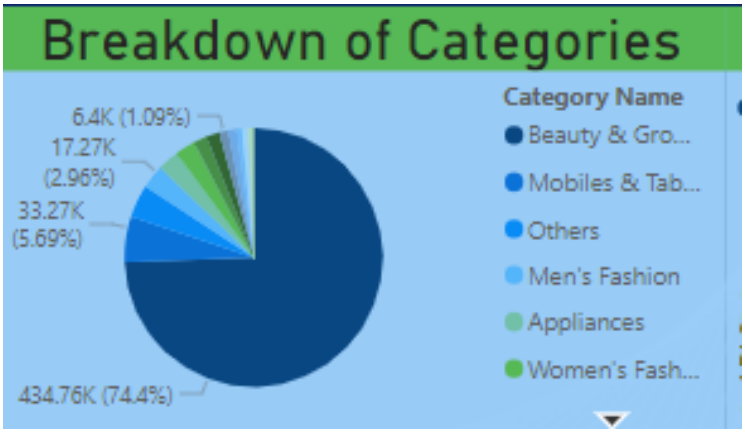
The answer to this query is that that most orders which are complete, followed by cancelled, and then refunded have high grand totals. The higher the grand total, the higher the refunded total as well.

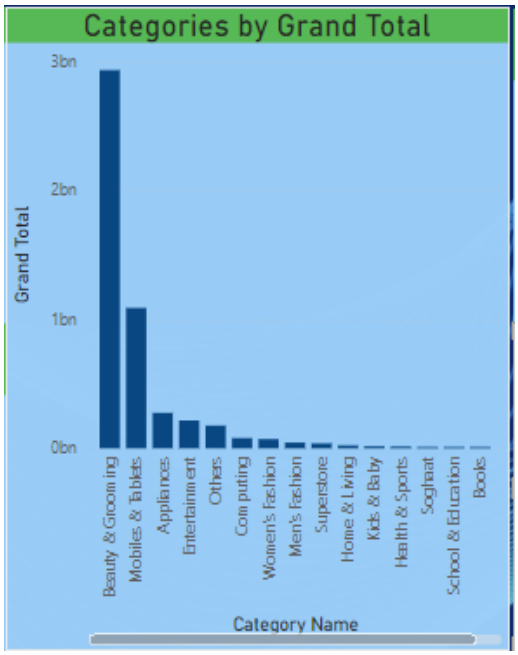


Imaan Fatima Imran 18060

Query 7: Which categories sell the most?

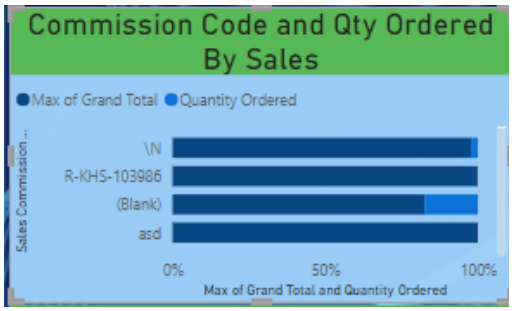
The answer to this query is Beauty and Grooming followed by Mobile and Accessories. This is illustrated by the following charts:





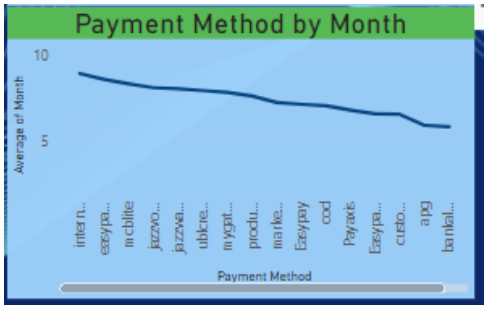
Query 8: Which commission code earned the most sales?

The answer to this query is firstly that those without commission actually earn the most, but when It comes to commission codes, the code R-KHS 10298 is the most successful. This is illustrated in the following chart:



Query 9: Which month was most popular for each payment method?

The answer to this query is that cod is most opted for in summer months while winter months see more internet banking. This is illustrated in the following chart:



Query 10: Does working date have any affect on order completion?

The answer to this query is that canceled orders are highest in 2017, and decline slowly in 2018, compared to completed orders which peak in 2017 and decline sharply in 2018.. This is illustrated by the following chart: