**Introduction:**

The data set belongs to a leading online E-Commerce company. An online retail (E commerce) company wants to know the customers who are going to churn, so accordingly they can approach customer to offer some promos.

In the fast-paced world of online retail, e-commerce companies face the challenge of retaining customers. To address this, identifying at-risk customers and implementing targeted retention strategies are crucial.

**Problem Statement:**

The goal of this project is to detailed analysis of a dataset from an online retail company, revealing valuable insights on customer churn. These insights offer essential guidance for decision-making, enabling proactive measures to reduce attrition and foster long-term loyalty.

**Project Approach:**

I got this dataset from [Kaggle](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction?sort=most-comments), and it contains information such as customers' personal details, satisfaction scores, preferred payment mode, days since the last order, and cashback amount. I used SQL (Postgre SQL) to clean and analyze this dataset, and performed visualizations using Microsoft Power BI. This analysis is divided into several stages: data cleaning, data exploration, an insight section, and recommendations.

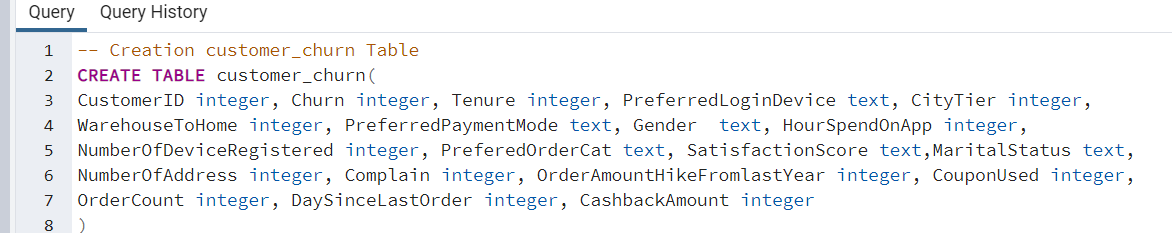
**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable** | **Discription** |
| CustomerID | Unique customer ID |
| Churn | Churn Flag |
| Tenure | Tenure of customer in organization |
| PreferredLoginDevice | Preferred login device of customer |
| CityTier | City tier |
| WarehouseToHome | Distance in between warehouse to home of customer |
| PreferredPaymentMode | Preferred payment method of customer |
| Gender | Gender of customer |
| HourSpendOnApp | Number of hours spend on mobile application or website |
| NumberOfDeviceRegistered | Total number of deceives is registered on particular customer |
| PreferedOrderCat | Preferred order category of customer in last month |
| SatisfactionScore | Satisfactory score of customers on service |
| MaritalStatus | Marital status of customer |
| NumberOfAddress | Total number of added on particular customer |
| Complain | Any complaint has been raised in last month |
| OrderAmountHikeFromlastYear | Percentage increases in order from last year |
| CouponUsed | Total number of coupons has been used in last month |
| OrderCount | Total number of orders has been places in last month |
| DaySinceLastOrder | Day Since last order by customer |
| CashbackAmount | Average cashback in last month |

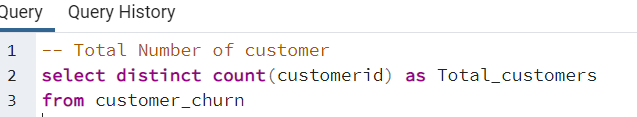
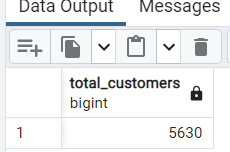
**Data Cleaning:**

Before going to analysis, it is essential to ensure the dataset is clean and reliable. The data cleaning process involves handling missing values, correcting inconsistencies, and formatting the data for analysis. In this project, we carefully cleaned the dataset to ensure the accuracy and integrity of our findings.

1. **Creation of Customer\_churn Table**

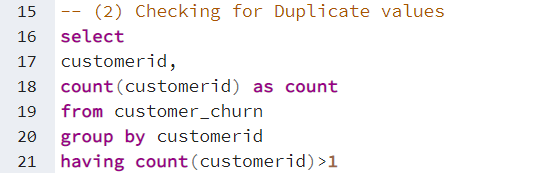
****

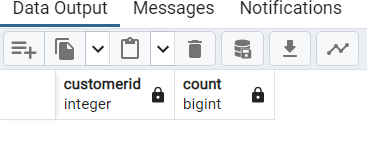
1. **Total No. of Customers**

**Findings: There are 5630 customers in given dataset**

1. **Checking for Duplicate Values**

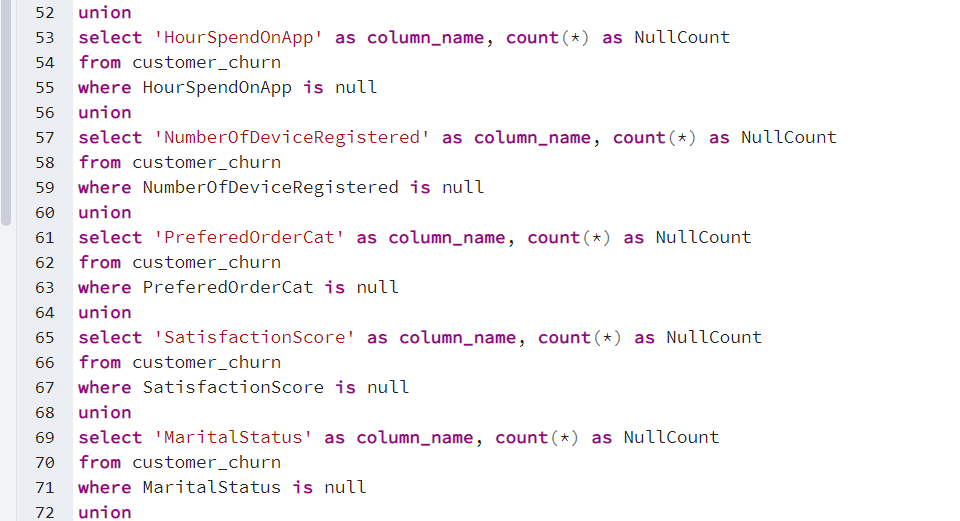
****

****

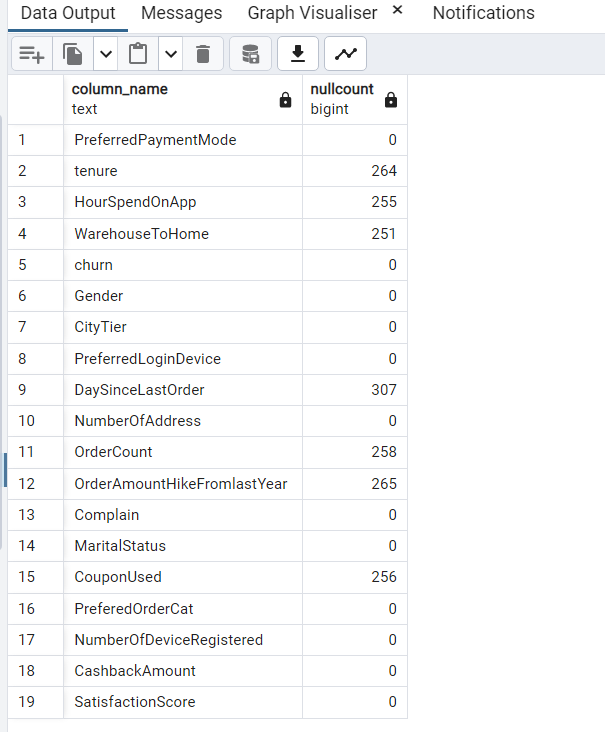
**Findings: The result showing empty table that means there is no duplicate values in the dataset**

**(4) Checking for Null Values**

****

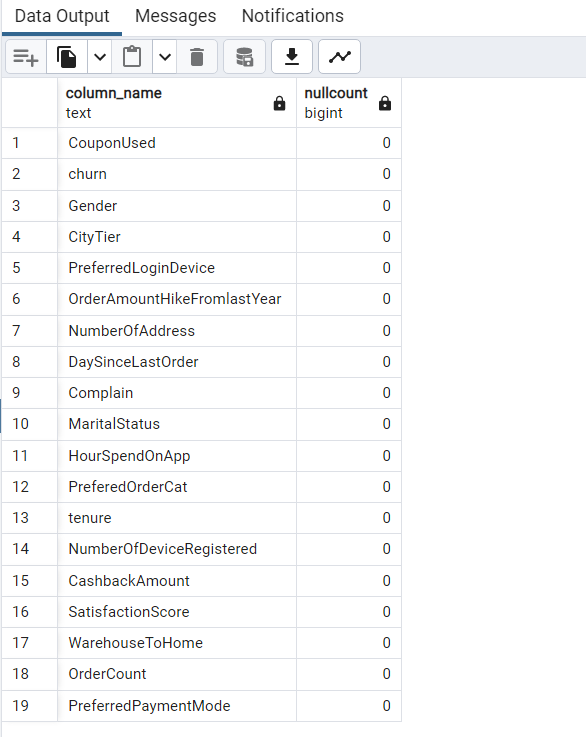
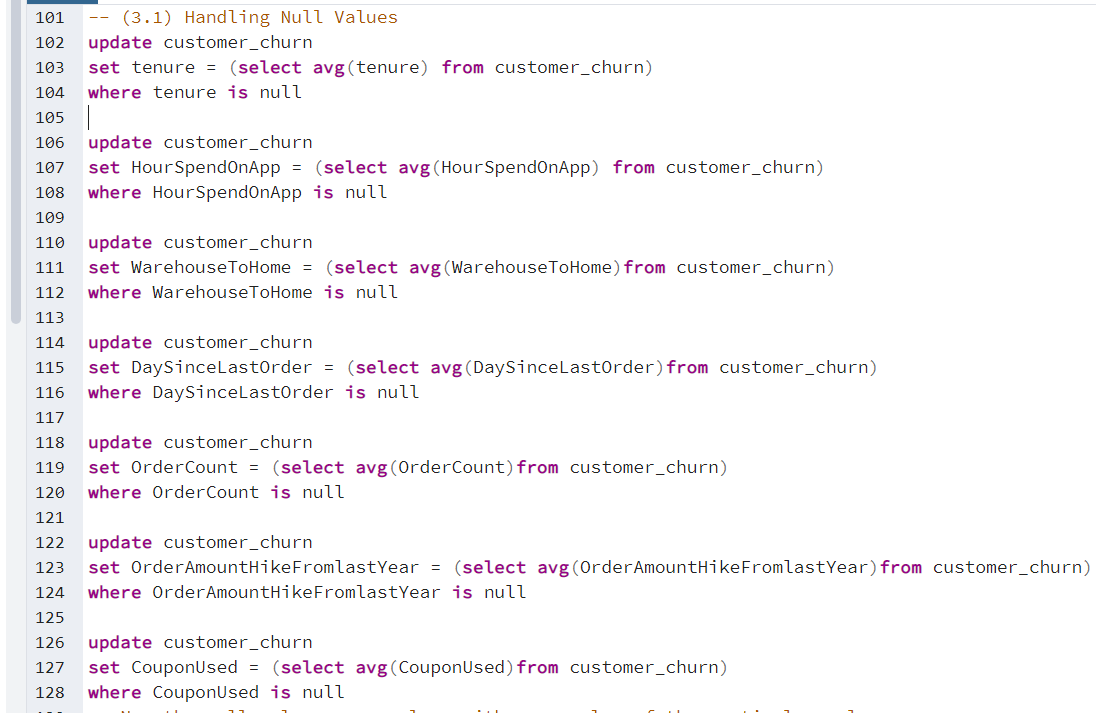






**Findings: CouponUsed, DaysSinceLastOrder, HourSpendOnApp, OrderAmountHikeFromLastYear, OrderCount, Tenure, and WarehouseToHome all have null values present, and the number of null values present for each column can be seen in the above table.**

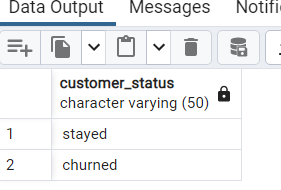
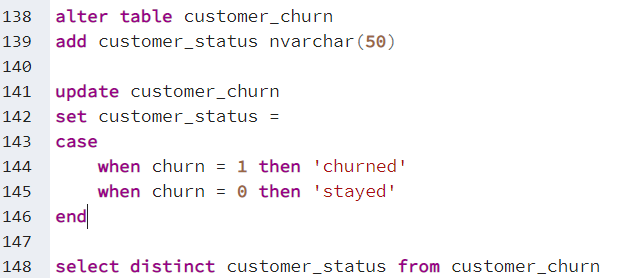
**(4) Handling Null Values**



**Findings: so, there is no null values in the dataset**

1. **Creating new column for an already existing column "Churn"**

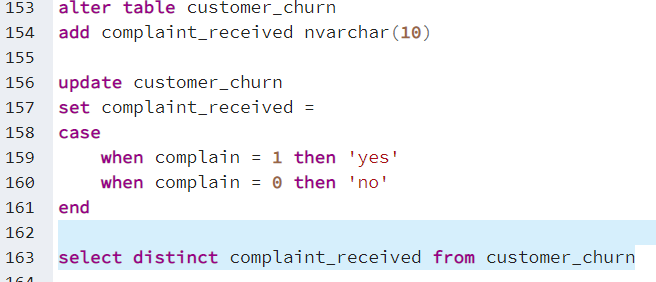
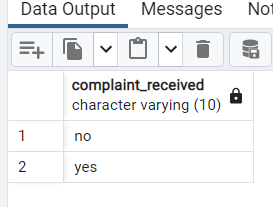
we can observe the "churn" column contained 0 and 1 values.o means that customer did not churn, while 1 means that customer churned.it is difficult to remember this,i we will create a new column called "customer\_status" that shows "stayed" or "churned



**Findings: The new column “customer\_status” has affected in table.it has two distinct values “churned”, “stayed”.**

1. **Creating new column for an already existing column "complain"**

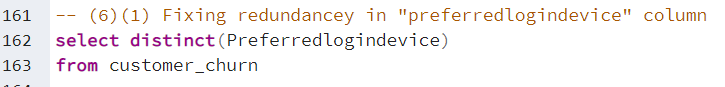
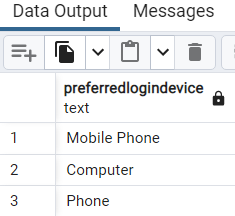
we noticed that the complaint column also contained 0 and 1. ‘0’ means that the customer did not record any complaints, while ‘1’ means the customer recorded a complaint. For clarity purposes, we will create a new column called ‘complain\_recieved’ that shows “No” when a customer did not complain and “Yes” when a customer complained.

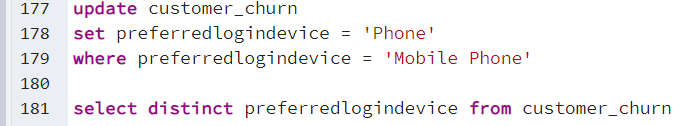
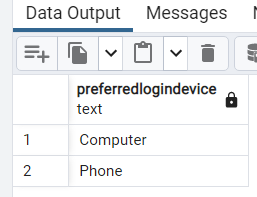
**Findings: the new columns complaint\_received affected in the database and it has two distinct values “yes”, “no”.**

1. **Checking Redundancies in each column**

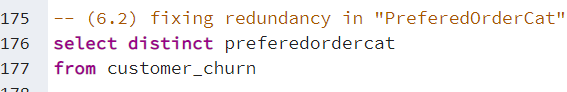
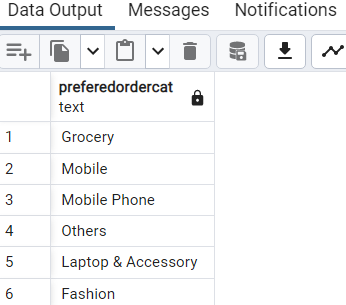
**(6.1) Checking redundancies in “PreferredLoginDevice” column**

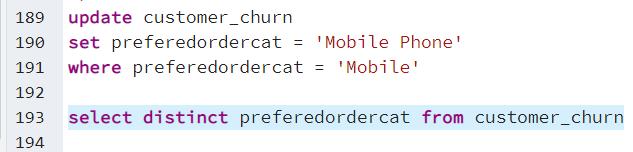
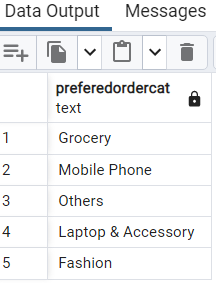
**Findings: I have observed phone and mobile phone in the same column, but they mean same thing. so, we will replace the mobile phone with phone.**

**(6.2) Checking redundancies in “preferedordercat” column**

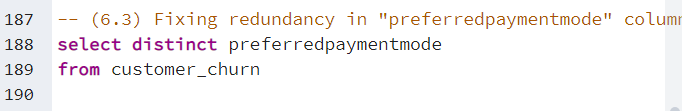
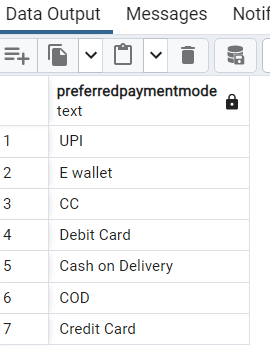
 

**Findings: I have observed mobile phone and phone appear in the column, but their meaning is same.so we will replace the phone with mobile phone.**

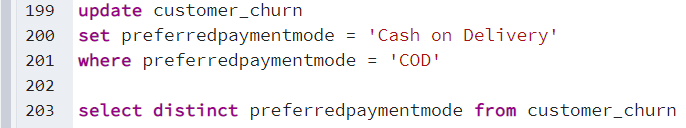
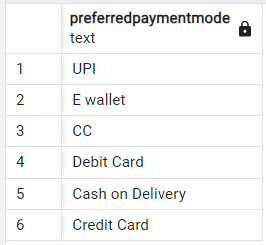
 

**Findings:** **I have observed Finally redundancy in “preferedordecat” column has been fixed**

**(6.3) Checking redundancies in “preferredpaymentmode” column**

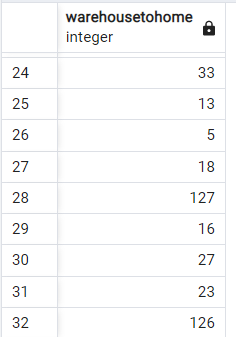
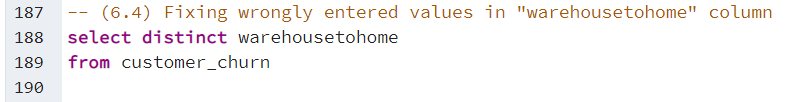
 

**Findings:** **I have observed “Cash on Delivery” and “COD” both appear in same column,but their meaning is same. So we will replace COD with Cash on Delivery.**

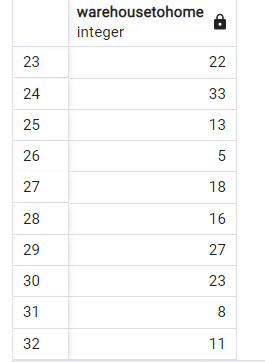
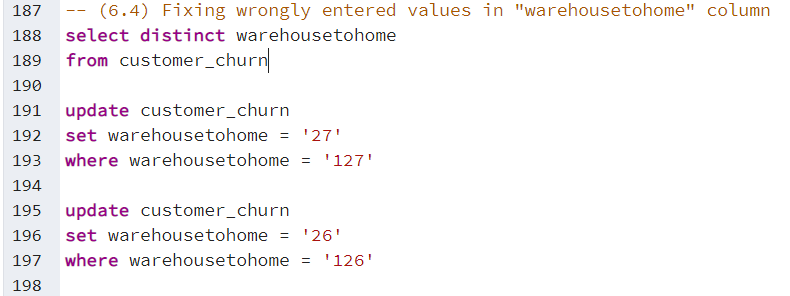
**** ****

**Findings:** **I have observed the redundancy in “preferredpaymentmode” has been fixed.**

**(6.4) Fixing wrongly entered values in “warehousetohome” column**

****

**Findings:** **I have observed 126 and 127 values. They are definitely outliers and most likely wrongly entered.to fix this we will replace those values with 26 and 27 to fall within range of the values.**

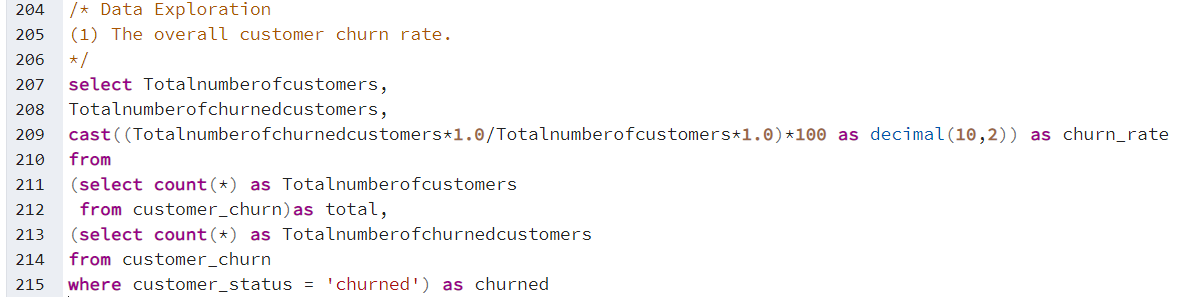
****

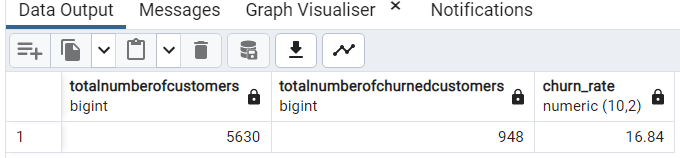
**Findings:** **I have observed 126 and 127 values are replaced with 26 and 27.**

**Now data cleaning is completed , so we will look into Data Exploration.**

**Data Exploration:**

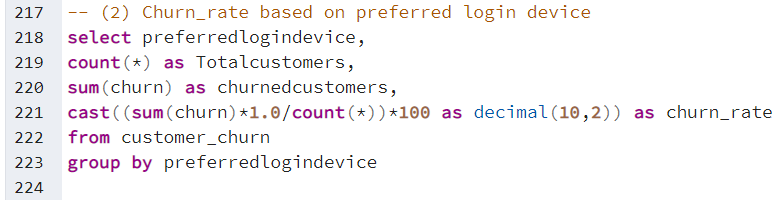
1. **The Overall customer churn rate.**

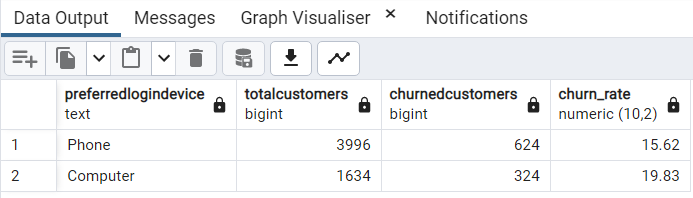
****



**Findings:** **The overall customer churn\_rate is approx 17%.**

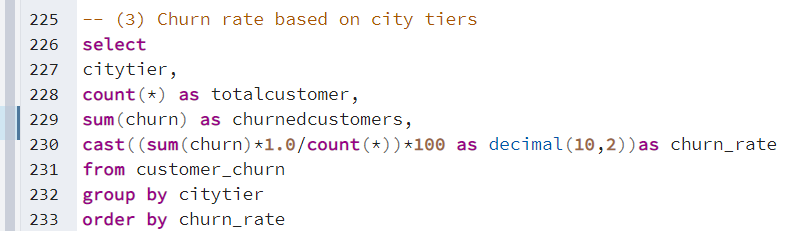
1. **Churn\_Rate based on preferred login device.**

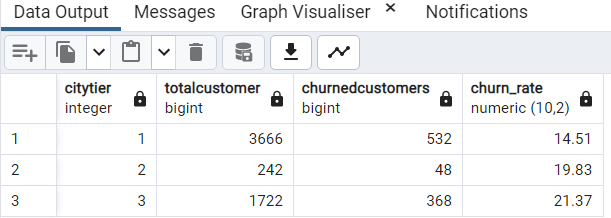
****

****

**Findings:**

1. **Customers who prefer login using computer have a slightly higher churn\_rate compared to customers who prefer login using phone.**
2. **Appox 16% customers churned who are prefer to login using phone.**
3. **Approx 20% of the customers churned who are prefer to login using computer.**
4. **Most of the customers churned who are using computer.**
5. **Churn\_Rate based on different city tiers.**

****

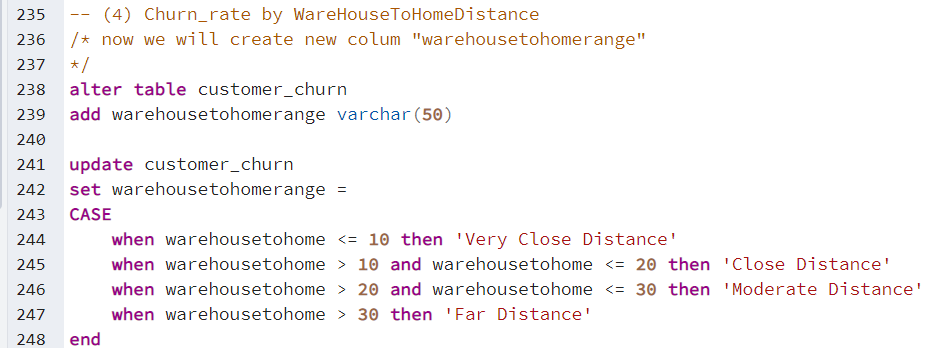
****

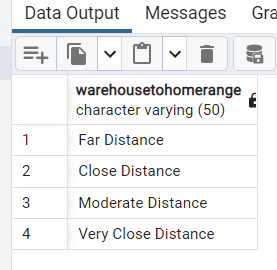
**Findings:**

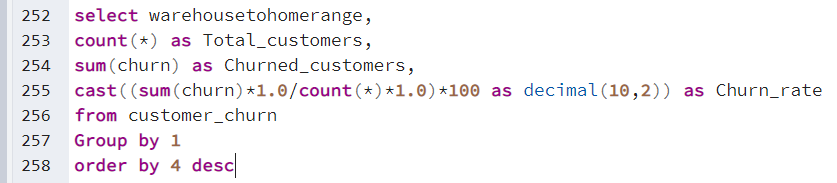
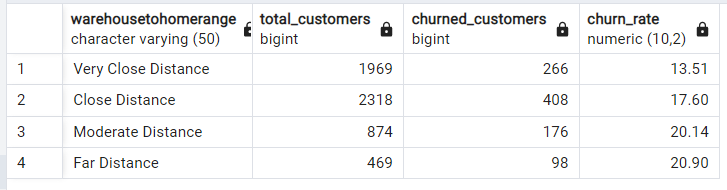
1. **Customers who prefer login using computer have a slightly higher churn\_rate compared to customers who prefer login using phone.**
2. **Most of the cutomers churned from city tier1**
3. **City tier 1 has lower churn\_rate compare to the tier 2 and tier 3 cities and city tier 3 is highest churn\_rate.**

**(4) Churn\_Rate based on warehouse to home distance.**

**We will create a new column called “WareHouseToHomeRange” that groups the distance into very close, close, moderate and far using CASE Statement**

****

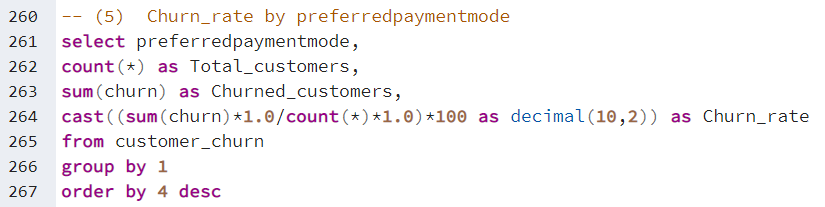
****

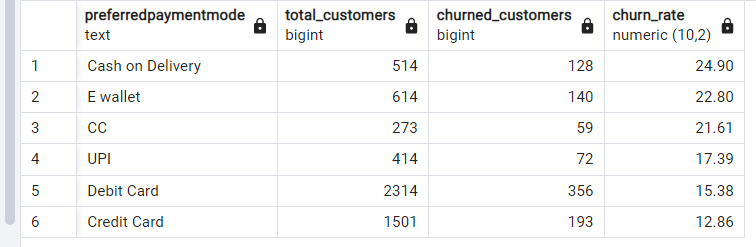
**** ****

**Findings: The distance between warehouse and customer’s home seems to have some influence on churn\_rate.**

1. **Customers who residing very close to warehouse tend to have lower churn\_rate.**
2. **Customers who residing very far to warehouse tend to have higher churn\_rate, so we will conclude that warehousetohome distance influence the churn\_rate.**

**(5) Churn\_Rate based on preferred payment mode.**



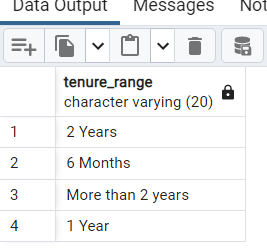
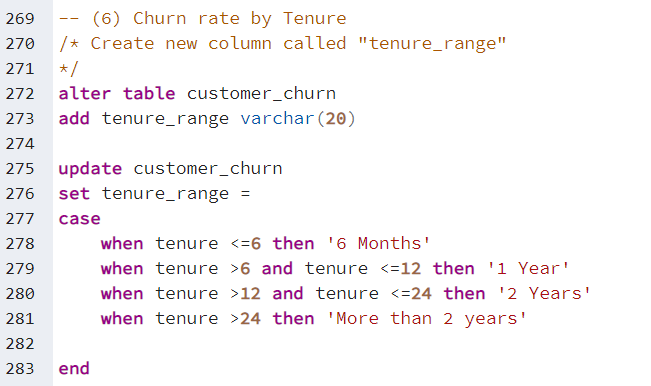


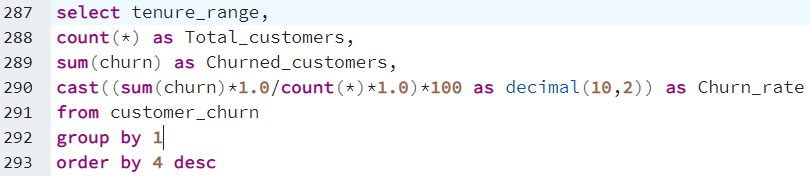
**Findings:**

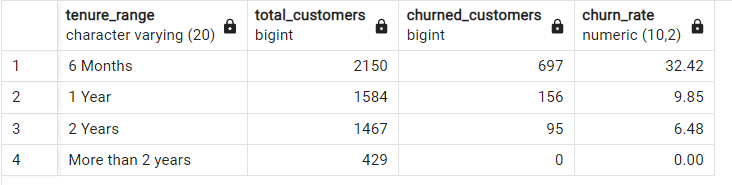
1. **Which customers had payment mode as “Cash on Delivery”, those customers having higher churn rate.**
2. **Payment mode like “Credit Card” and “Debit Card” have lower churn rate.**
3. **The preferred payment mode has some influence on churn rate.**

**(6) Churn\_Rate based on Tenure.**

First of all, I have created new column called “tenure\_range”. That groups the customers tenure into 6months, 1 year, 2 years and more than 2 years using case statement

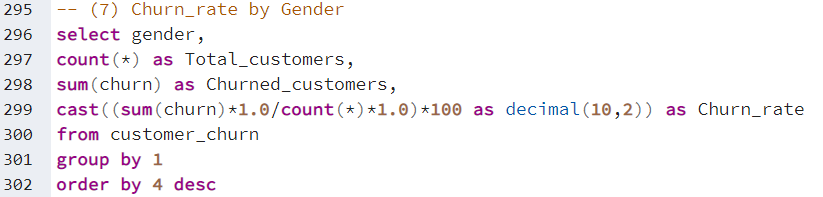
****

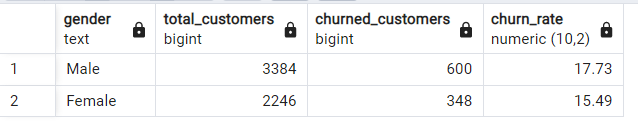
****

****

**Findings:**

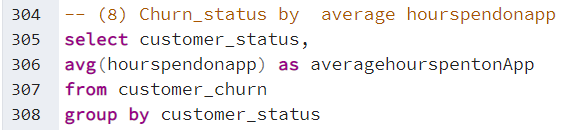
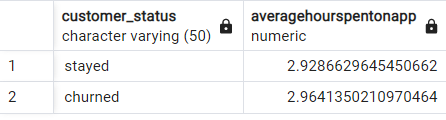
1. **Most of the customers had churned whose tenure \_range is 6 Months.**
2. **The customers who had tenure range is more than 2 years, there were not interested to churn. i.e., the churn rate Is “0”.**
3. **Churn\_Rate based on Gender.**

****

****

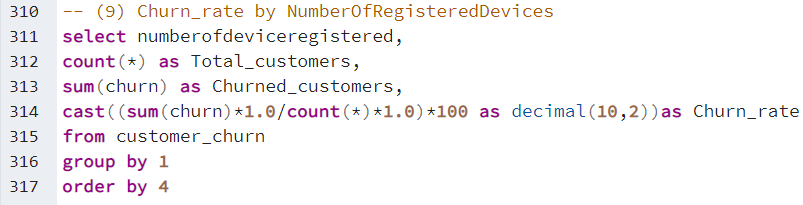
**Findings:**

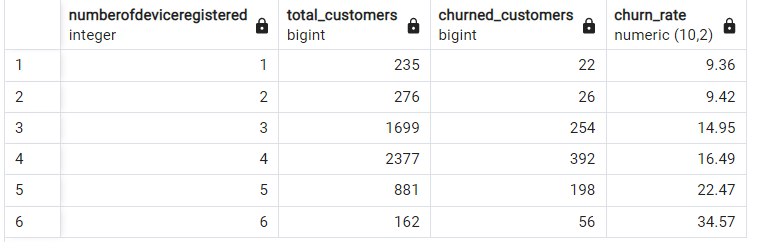
1. **Churn rate of Male customers has higher when compare to Female customer.**
2. **600 Male customers churned.**
3. **Customer\_status based on hourspentonapp.**

**** ****

**Findings:**

1. **Both Stayed and Churned customers had the same average hours spent on app.**
2. **Hourspentonapp might not be influenced churned customers.**
3. **Churn\_Rate based on Numberofregistered devices.**

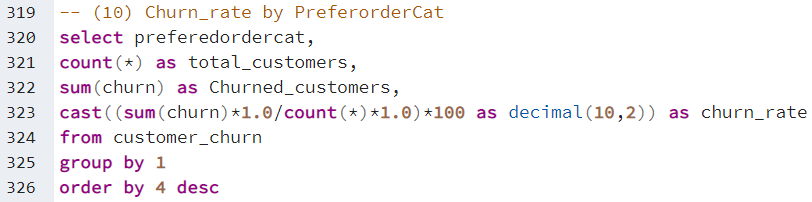
****

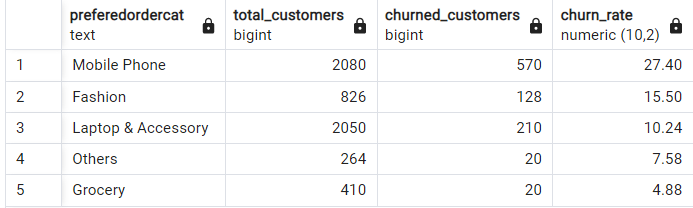
****

**Findings:**

1. **We can observe there is positive correlation between “Numberofdeviceregistered” and “Churn rate”**
2. **Customers with less registered devices more likely to low churn rate.**
3. **Numberofdeviceregistered is influence the churn rate.**

**(10) Churn\_Rate based on preferordercat.**

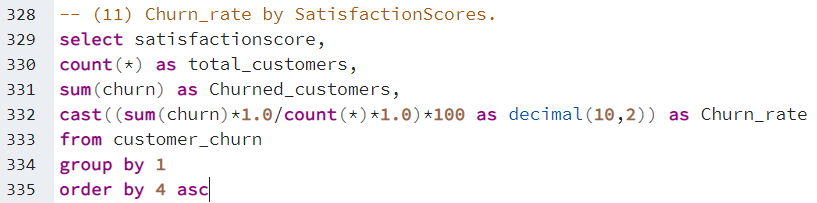
****

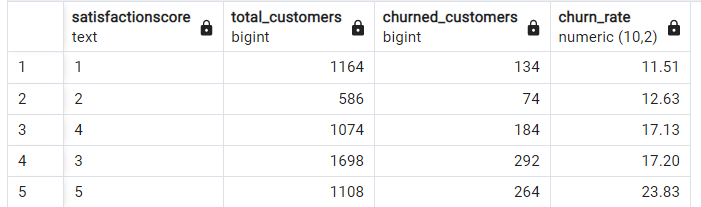
****

**Findings:**

1. **Mobile Phone category had Highest Churn\_rate compare to all other categories.**
2. **Grocery category had lowest churn\_rate.**

**(11) Churn\_Rate based on SatisfactionScores.**

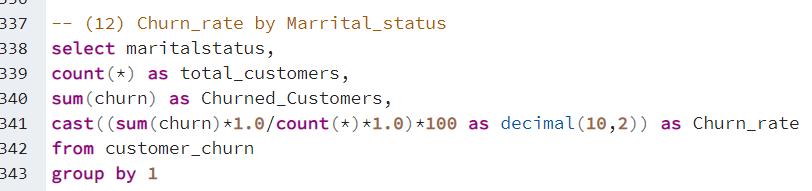
****

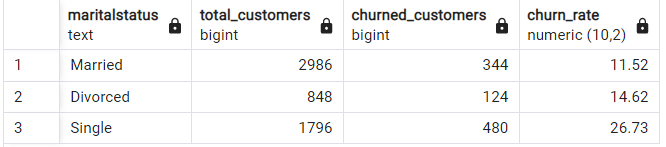
****

**Findings:**

1. **The satisfaction score ranges from 1 to 5**
2. **Most of customers had satisfaction score as 1 and 5**
3. **Those who had Higher satisfaction score as 5, had higher churn\_rate. And satisfaction score positive correlated with churn\_rate.**

**(12) Churn\_Rate based on Marital\_Status.**

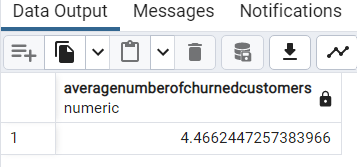
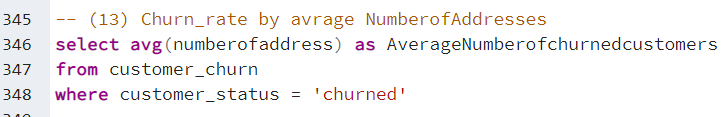
****

****

**Findings:**

1. **Single Customers have the highest churn\_rate compare to customers with other marital\_status**
2. **Married customers have the lowest churn\_rate followed by Divorced.**

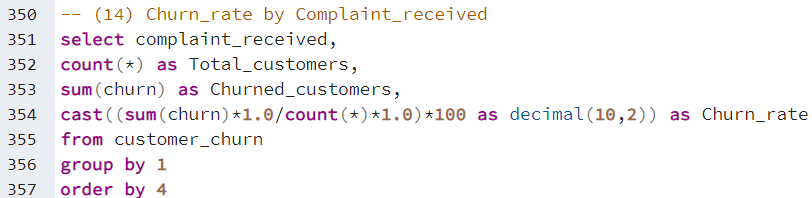
**(13) Churn\_Rate based on Numberof Address.**

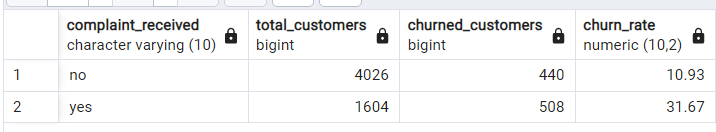
****

**Findings:**

1. **Customers churned who were 4 addresses changed on an Average.**

**(14) Churn\_Rate based on Complaints\_received.**

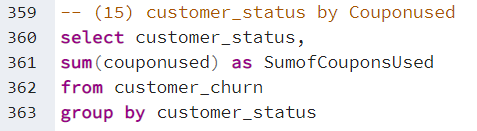
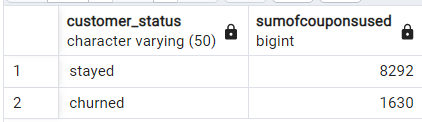
****

****

**Findings:**

1. **The customers who had raised the complaint last month they were more likely to churn.**
2. **Churn\_rate of the complaint raised is more than complaints not raised customers.**

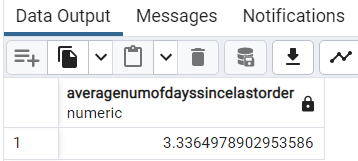
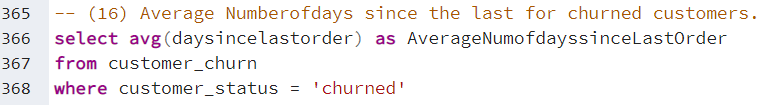
**(15) Customer\_Status by by Couponsused.**

**** ****

**Findings:**

1. **Most of the Stayed customers used coupons.**
2. **churned customers used 1630 coupons.**

**(16) Average Number of days since last order for churned customer.**

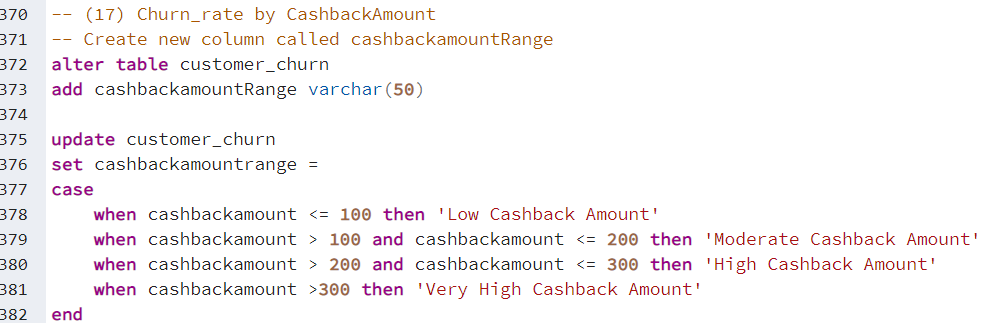
****

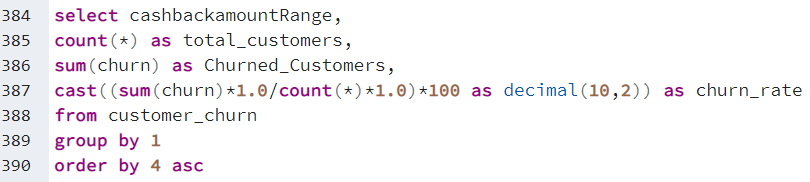
**Findings:**

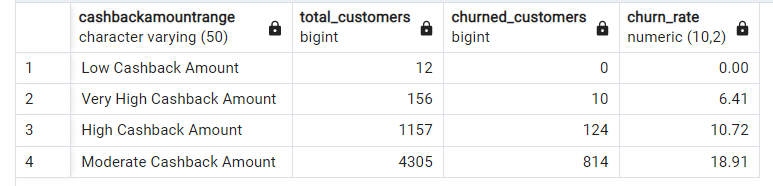
1. **the churned customers have, on average, a short period of time since their last order indicates that they recently stopped engaging with the company. By focusing on the overall customer experience, implementing targeted retention initiatives, and maintaining continuous engagement.**

**(17) Churn\_rate by CashbackAmount.**

First of all I have created new column called “CashbackAmountRange” that groups the cashbackamount into low (<100), Moderate (between 100 and 200), High(between 200 and 300), and very High (more than 300) using the CASE Statement

****

****

****

**Findings:**

1. **Customers who had moderate cashback amount had a relatively higher churn\_rate.**
2. **Those who received Higher and very High cashback amount having lower churn\_rate.**

**Key Insights of Analysis**

* **The dataset has 5630 customers.**
* **The overall churn rate is 16.84%.**
* **Customers who prefer logging in with a computer have slightly higher churn rate compare to phone users.**
* **Tier 1 cities have lower churn rates than tier 2 and tier 3 cities.**
* **Customers who residing very close to warehouse tend to have lower churn rate and Customers who residing very far to warehouse tend to have higher churn rate.**
* **“Cash on Delivery” and “E wallet” payment modes have higher churn rates, while “Credit Card” and “Debit Card” have lower churn rates.**
* **Longer tenure is associated with lower churn rates.**
* **Male customers have slightly higher churn rates than Female Customers.**
* **Both Stayed and Churned customers had the same average hours spent on app.**
* **“Mobile Phone” order category has the highest churn rate, while “Grocery” has the lowest churn rate.**
* **More registered devices correlated with higher churn rates.**
* **Highly satisfied customers (rating 5) have a relatively higher churn rate.**
* **Single customers have the highest churn rate, while married customers have lower churn rate.**
* **Churned customers have an average of 4 associated addresses.**
* **Customer complaints are prevalent among churned customers, emphasizing the importance of addressing concerns to minimize churn**.
* **Coupon usage is higher among non-churned customers.**
* **Moderate cashback amounts correspond to higher churn rates, while higher amounts leads to lower churn rates.**