**Exploratory Data Analysis (EDA)**

**Missing Values**:

A screenshot of a computer

Description automatically generated

Replaced with mean.

**Outlier Summary**

**A screenshot of a computer

Description automatically generated**

**High Outlier Counts**: Variables like CouponUsed, OrderCount, and DaySinceLastOrder have a significant number of outliers.

**Low or No Outliers:** Variables such as CityTier, NumberOfDeviceRegistered, and SatisfactionScore have no outliers.

**Consistency:** Variables and their Z-score counterparts show consistent outlier detection results.

**Cap and Floor**

I implemented capping and flooring to mitigate the influence of extreme outliers on our analysis. By setting a maximum and minimum threshold for our data, I aimed to reduce the impact of these extreme values, ensuring they don't disproportionately affect our statistical measures and model performance while still retaining the data points within a reasonable range. This approach helps maintain the integrity of the dataset and improves the robustness of our analyses and predictions.

**Removing Outliers**

A screenshot of a computer

Description automatically generated

**Charts**

* **Distribution of Churn**

A chart with a number of bars

Description automatically generated with medium confidence

* A graph with red and blue squares

  Description automatically generated**Correlation Matrix**

Interpretations:

1. **Tenure and Churn**: I've identified a moderate negative correlation (-0.33) between Tenure and Churn. This is our most significant lever for reducing churn. Customers who stay with us longer are less likely to leave, so I recommend we prioritize strategies to increase customer tenure.
2. **Cashback's Impact**: I've noticed that CashbackAmount has a moderate positive correlation with Tenure (0.46) and a slight negative correlation with Churn (-0.14). This suggests our cashback program could be an effective tool in reducing churn by encouraging longer tenure.
3. **Order Frequency**: I see that OrderCount has a slight negative correlation with Churn (-0.02). While small, this indicates that customers who order more frequently are somewhat less likely to churn. Encouraging repeat purchases could help reduce churn.
4. **Coupon Usage**: Interestingly, CouponUsed shows almost no correlation with Churn (-0.00). However, it's strongly correlated with OrderCount (0.67), which in turn slightly reduces churn. This suggests that while coupons don't directly impact churn, they could indirectly help by increasing order frequency.
5. **Satisfaction and Churn**: I've observed a weak positive correlation (0.10) between SatisfactionScore and Churn. This counterintuitive result suggests we may need to reassess how we measure customer satisfaction to ensure it aligns with reducing churn.
6. **Complaints and Churn**: There's a weak positive correlation (0.25) between Complain and Churn. This indicates that addressing customer complaints promptly and effectively could be crucial in our churn reduction efforts.
7. **App Engagement**: HourSpendOnApp shows a very weak positive correlation (0.01) with Churn. This suggests that while app engagement isn't strongly linked to churn, there might be an opportunity to use the app more effectively for retention.

* **Bar Charts for Categorical Variables**

A group of graphs with numbers and text

Description automatically generated with medium confidence

Interpretations:

1. When analyzing the **Preferred Login Device vs. Churn** graph, I notice that a significant number of users who prefer to log in using a mobile phone tend not to churn. Specifically, there are 2,190 users who continue to use the service without churning, compared to 324 who churn. In contrast, for users who prefer logging in via computer, 1,172 do not churn, while 285 do. This indicates a higher retention rate among mobile phone users. Interestingly, no users who prefer logging in via a computer churn, which could suggest high satisfaction within this group.
2. Looking at the **Preferred Payment Mode vs. Churn** graph, I observe that users preferring to pay with a debit card show the highest number of non-churners, with 1,783 not churning and 315 churning. Credit card users follow, with 1,174 non-churners and 174 churners. On the other hand, users who prefer cash on delivery exhibit the lowest retention rate, with only 113 not churning and 22 churning. This suggests that electronic payment methods, particularly debit cards, are associated with higher retention rates among users.
3. From the **Gender vs. Churn** graph, it is evident that male users are more likely to churn compared to female users. Specifically, 2,515 males do not churn, whereas 532 males churn. In comparison, 1,717 females do not churn, and only 309 females churn. This indicates a higher churn rate among male users, suggesting that females are generally more satisfied or engaged with the service.
4. I notice that the "**Laptop & Accessory**" category has the highest number of customers, with 1653 non-churned and 194 churned users. The "Mobile phone" category follows as the second most popular, showing 831 non-churned and 299 churned customers. Interestingly, the "Fashion" category has a notable difference between non-churned 639 and churned 124 customers, suggesting it might be an area where customer retention is relatively strong.
5. Looking at the "**MaritalStatus vs Churn**" graph, I observe that married customers form the largest group, with 2370 non-churned and 302 churned individuals. Single customers come next, showing 1138 non-churned and 415 churned customers. The divorced category is the smallest, with 724 non-churned customers and 124 churned ones. It's worth noting that the churn rate appears to be proportionally higher among single customers compared to married ones, which could be an interesting point for further analysis.

Therefore, customers who prefer logging in via mobile phones and those using debit cards for payments show higher retention rates, indicating their satisfaction with the service. Gender-wise, female users tend to stay with the service longer than male users. When looking at preferred order categories, "Laptop & Accessory" and "Mobile phone" have the highest customer base, with the "Fashion" category showing strong retention. Marital status also plays a role in churn, with married customers having a lower churn rate compared to single ones. These patterns suggest that targeted strategies to enhance user experience for mobile login users, electronic payment users, and single customers could effectively reduce churn rates.

* **Bar Charts for Numerical Variables**

A group of blue and green rectangular shapes

Description automatically generated

**Interpretations:**

**Mean Tenure by Churn:** The graph indicates that customers who have not churned exhibit a substantially higher mean tenure compared to those who have churned. This suggests that customers with longer tenures are more likely to remain with the service, highlighting the importance of customer retention efforts over time.

**Mean CityTier by Churn:** The mean CityTier for both churned and non-churned customers is nearly identical, indicating that the geographical tier of a city does not significantly influence customer churn. This suggests that other factors beyond geographic segmentation might be more critical in understanding customer retention.

**Mean WarehouseToHome by Churn:** The average distance from the warehouse to the customer's home is very similar for both churned and non-churned customers. This indicates that logistical factors, such as the proximity of the warehouse to the customer's residence, do not play a significant role in customer churn.

**Mean HourSpendOnApp by Churn:** The mean number of hours spent on the app is almost the same for both churned and non-churned customers, suggesting that the duration of app usage does not have a strong correlation with customer churn. This implies that app engagement metrics may need to be examined in conjunction with other variables to understand churn better.

**Mean NumberOfDeviceRegistered by Churn:** There is no substantial difference in the mean number of devices registered between churned and non-churned customers, indicating that the number of devices a customer registers does not significantly impact their likelihood of churning.

**Mean SatisfactionScore by Churn:** The satisfaction score is fairly consistent between both groups, implying that overall satisfaction levels, as measured, do not significantly differentiate between customers who churn and those who remain. This suggests that other aspects of customer experience might be influencing churn more strongly.

**Mean NumberOfAddress by Churn:** The mean number of addresses maintained by customers is similar regardless of their churn status, indicating that the number of addresses associated with a customer's account is not a significant factor in predicting churn.

**Mean Complain by Churn:** The data reveals that churned customers have a higher mean number of complaints compared to non-churned customers. This suggests a strong correlation between customer complaints and churn, indicating that addressing customer grievances promptly and effectively could be critical in reducing churn rates.

**Mean OrderAmountHikeFromlastYear by Churn:** The average increase in order amount from the previous year is similar for both groups. This indicates that changes in order value over time do not significantly influence whether customers decide to churn.

**Mean CouponUsed by Churn:** The mean number of coupons used is nearly the same for both churned and non-churned customers. This suggests that the usage of promotional coupons does not have a strong impact on customer retention or churn.

**Mean OrderCount by Churn**: The mean order count is comparable between both groups, indicating that the total number of orders placed does not significantly influence customer churn. This implies that frequency of orders alone may not be sufficient to predict churn behavior.

**Mean DaySinceLastOrder by Churn:** Customers who have churned have a lower mean number of days since their last order compared to those who have not churned. This might indicate that recent purchasing activity is not necessarily a deterrent to churn, and other factors such as customer satisfaction or competitive offerings might be at play.

**Mean CashbackAmount by Churn:** The mean cashback amount received by churned customers is slightly lower than that of non-churned customers, suggesting that cashback incentives may have a minor but notable impact on customer retention. This indicates that financial incentives could play a role in customer loyalty strategies.

* **Histogram For Continuous Variables**

A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

**Interpretations:**

The **distribution of tenure** indicates that the majority of customers have a tenure of fewer than 10 years, with a steep decline in frequency as tenure increases. This suggests that the company has a relatively young customer base with a high turnover rate in the initial years, while a smaller segment of loyal customers maintains a longer relationship with the company.

The **distribution of the WarehouseToHome** variable shows a concentration of customers within a short distance from the warehouse, primarily within 20 units of distance. There are a few outliers at much greater distances, but the majority of customers reside close to the warehouse, which could imply a regional business focus or an efficient local delivery system.

The **distribution of HourSpendOnApp** is characterized by several distinct peaks, indicating that customers tend to spend their time on the app in specific durations, notably around 2, 3, and 4 hours. These peaks suggest common usage patterns or behaviors among customers, which might be tied to specific app features or services that engage users for set periods.

The **distribution of OrderAmountHikeFromLastYear** reveals that most customers have experienced a moderate increase in order amounts, predominantly between 12 and 18 units. The frequency decreases steadily as the order amount hike increases, indicating that significant hikes are less common and the majority of customers see incremental increases in their order amounts year over year.

The **distribution of CouponUsed** shows that the majority of customers use between 0 to 2 coupons, with a sharp decline in frequency as the number of coupons used increases. This suggests that while coupon usage is prevalent, extensive use of coupons is rare, indicating either limited availability or lower perceived value of additional coupons.

The **distribution of OrderCount** is heavily skewed towards lower counts, with the majority of customers placing between 1 to 4 orders. The frequency sharply declines as the order count increases, highlighting that while occasional orders are common, frequent ordering is much less typical among customers.

The **distribution of DaySinceLastOrder** suggests that most customers have placed an order within the last 10 days, with frequency sharply decreasing as the days since the last order increase. This pattern indicates a high level of recent engagement, although there is a long tail of customers who have not ordered in a significant amount of time.

The **distribution of CashbackAmount** shows a concentration around 100 to 150 units, with the frequency tapering off as the cashback amount increases. This indicates that while a moderate amount of cashback is common, higher cashback amounts are less frequent, possibly reflecting company policy on cashback rewards or customer spending patterns.

* **Distribution Bar Charts**

A group of blue squares

Description automatically generated

**Interpretations:**

In analyzing the **distribution of churn**, I observe that the majority of customers have not churned, indicating a relatively stable customer base. However, a significant minority has churned, which suggests there are areas for improvement in customer retention strategies.

The **distribution of internet service** types reveals that the majority of customers use fiber optic services, followed by DSL, with a smaller portion using no internet service. This highlights the preference for high-speed internet among customers.

Examining the **payment methods**, I see that electronic check is the most common, followed by mailed check, bank transfer, and credit card. This suggests a diverse range of payment preferences among customers, with a notable inclination towards electronic transactions.

The **contract distribution** shows that month-to-month contracts are the most prevalent, followed by one-year and two-year contracts. This indicates a preference for flexibility among customers, with fewer committing to long-term contracts.

The **gender distribution** is fairly balanced, with a slight majority of male customers. This balance suggests that the services offered appeal equally to both genders, without significant bias.

In **the partner distribution**, a larger portion of customers do not have partners compared to those who do. This could imply that the services are particularly appealing to single individuals or those without partners.

The **distribution of dependents** shows that the majority of customers do not have dependents. This might indicate that the services are more attractive to individuals or couples without children.

The **marital status distribution** reveals that a significant portion of customers are married, followed by those who are single, divorced, or widowed. This suggests that the services cater well to a diverse range of marital statuses, with a strong appeal to married individuals.

* **Correlation between Numeric and Categorical variables (ANOVA)**

**A screenshot of a computer

Description automatically generated**

1. **Tenure**: The ANOVA results indicate a highly significant effect of Tenure on Churn, with an F-Statistic of 618.95 and an extremely low P-Value (5.22e-129). This suggests that the duration of a customer's engagement with the service strongly influences their likelihood to churn.
2. **CityTier**: The analysis shows that CityTier has a significant impact on Churn, evidenced by an F-Statistic of 49.28 and a P-Value of (2.51e-12). This suggests that customers from different city tiers exhibit varying propensities to churn.
3. **WarehouseToHome**: The significant F-Statistic of 24.17 and a low P-Value (9.11e-07) for WarehouseToHome indicate that the distance between the warehouse and the customer's home significantly affects their likelihood of churning.
4. **HourSpendOnApp**: The results for HourSpendOnApp show an F-Statistic of 0.91 and a P-Value of 0.34, suggesting that the amount of time a customer spends on the app does not have a significant impact on their likelihood to churn.
5. **NumberOfDeviceRegistered**: With an F-Statistic of 69.94 and a very low P-Value (7.81e-17), NumberOfDeviceRegistered is found to have a strong association with Churn, indicating that the number of devices registered by a customer is a significant predictor of churn.
6. **SatisfactionScore**: The analysis reveals a significant relationship between SatisfactionScore and Churn, with an F-Statistic of 52.28 and a P-Value of (5.54e-13). This highlights that customer satisfaction scores are a strong determinant of churn likelihood.
7. **NumberOfAddress**: The F-Statistic of 13.76 and a P-Value of (2.10e-04) indicate a significant effect of NumberOfAddress on Churn, suggesting that the number of addresses associated with a customer influences their propensity to churn.
8. **Complain**: The extremely high F-Statistic of 334.32 and a very low P-Value (2.23e-72) suggest that the number of complaints is a highly significant predictor of Churn, indicating a strong relationship between the frequency of complaints and the likelihood of churning.
9. **OrderAmountHikeFromlastYear**: The F-Statistic of 2.86 and a P-Value of 0.09 suggest that OrderAmountHikeFromlastYear does not have a significant effect on Churn, indicating that changes in order amounts over the past year are not a strong predictor of churn.
10. **CouponUsed**: The analysis shows an F-Statistic of 0.09 and a P-Value of 0.76, indicating that the use of coupons does not significantly affect Churn. This suggests that coupon usage is not a major factor influencing customer churn.
11. **OrderCount**: With an F-Statistic of 2.26 and a P-Value of 0.13, OrderCount does not have a significant effect on Churn, suggesting that the number of orders placed by a customer is not a strong predictor of their likelihood to churn.
12. **DaySinceLastOrder**: The high F-Statistic of 111.46 and a very low P-Value (8.69e-26) indicate a significant effect of DaySinceLastOrder on Churn, showing that the time elapsed since a customer's last order is a strong predictor of their likelihood to churn.
13. **CashbackAmount**: The significant F-Statistic of 105.95 and a low P-Value (1.32e-24) reveal that CashbackAmount has a strong impact on Churn, indicating that the amount of cashback received by customers is a significant factor in predicting churn.

* **PCA Analysis**

A screenshot of a computer

Description automatically generated

A graph with a red line and blue line

Description automatically generated

The PCA analysis has provided us with the variance explained by each principal component, which helps in understanding the dimensionality reduction of our dataset. The first principal component (PC1) explains approximately 18.96% of the variance, while the second principal component (PC2) accounts for about 10.36%. The cumulative explained variance ratio reveals that PC1 and PC2 together capture around 29.32% of the total variance. As we include more components, the cumulative variance increases, with all 13 components collectively accounting for 100% of the variance.