E-Commerce Churn

Kailash Baskar 0790883

Academic Integrity

I Kailash Baskar, hereby state that I have not communicated with or gained information in any way from any person or resource that would violate the College's academic integrity policies, and that all work presented is my own. In addition, I also agree not to share my work in any way, before or after submission, that would violate the College's academic integrity policies.

Data set

https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction (https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)

Loading the data

Installing the required packages.

```
install.packages("tidyverse", repos = "https://cran.rstudio.com")
```

```
## Installing package into 'C:/Users/Kailash Baskar/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Kailash Baskar\AppData\Local\Temp\Rtmp06y6BS\downloaded_packages
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

Loading the dataset.

```
library(readxl)
E_Commerce_Dataset <- read_excel("C:\\Users\\Kailash Baskar\\Desktop\\R_DAB501\\DAB501\\Docs\\E_Com
m\\E Commerce Dataset.xlsx", sheet = "E Comm")
View(E_Commerce_Dataset)

edata <- E_Commerce_Dataset
glimpse(edata)
```

```
## Rows: 5,630
## Columns: 20
## $ CustomerID                <dbl> 50001, 50002, 50003, 50004, 50005, 50006, ~
## $ Churn                     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Tenure                    <dbl> 4, NA, NA, 0, 0, 0, NA, NA, 13, NA, 4, 11,~
## $ PreferredLoginDevice      <chr> "Mobile Phone", "Phone", "Phone", "Phone",~
## $ CityTier                  <dbl> 3, 1, 1, 3, 1, 1, 3, 1, 3, 1, 1, 1, 1, 1, ~
## $ WarehouseToHome           <dbl> 6, 8, 30, 15, 12, 22, 11, 6, 9, 31, 18, 6,~
## $ PreferredPaymentMode      <chr> "Debit Card", "UPI", "Debit Card", "Debit ~
## $ Gender                    <chr> "Female", "Male", "Male", "Male", "Male", ~
## $ HourSpendOnApp            <dbl> 3, 3, 2, 2, NA, 3, 2, 3, NA, 2, 2, 3, 2, 3~
## $ NumberOfDeviceRegistered  <dbl> 3, 4, 4, 4, 3, 5, 3, 3, 4, 5, 3, 4, 3, 4, ~
## $ PreferedOrderCat          <chr> "Laptop & Accessory", "Mobile", "Mobile", ~
## $ SatisfactionScore         <dbl> 2, 3, 3, 5, 5, 5, 2, 2, 3, 3, 3, 3, 3, 3, ~
## $ MaritalStatus             <chr> "Single", "Single", "Single", "Single", "S~
## $ NumberOfAddress           <dbl> 9, 7, 6, 8, 3, 2, 4, 3, 2, 2, 2, 10, 2, 1,~
## $ Complain                  <dbl> 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, ~
## $ OrderAmountHikeFromlastYear <dbl> 11, 15, 14, 23, 11, 22, 14, 16, 14, 12, NA~
## $ CouponUsed                <dbl> 1, 0, 0, 0, 1, 4, 0, 2, 0, 1, 9, 0, 2, 0, ~
## $ OrderCount                <dbl> 1, 1, 1, 1, 1, 6, 1, 2, 1, 1, 15, 1, 2, 1,~
## $ DaySinceLastOrder         <dbl> 5, 0, 3, 3, 3, 7, 0, 0, 2, 1, 8, 0, 2, 0, ~
## $ CashbackAmount            <dbl> 159.93, 120.90, 120.28, 134.07, 129.60, 13~
```

Univariant Analysis

Numeric Data

For the Numeric Univariant Analysis we take the variable "***CashbackAmount***" as our data.

Calculating **Mean** and **Median** for the variable.

```
summary(edata$CashbackAmount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   145.8   163.3   177.2   196.4   325.0
```

```
cash_mean = mean(edata$CashbackAmount)
cash_mean
```

```
## [1] 177.223
```

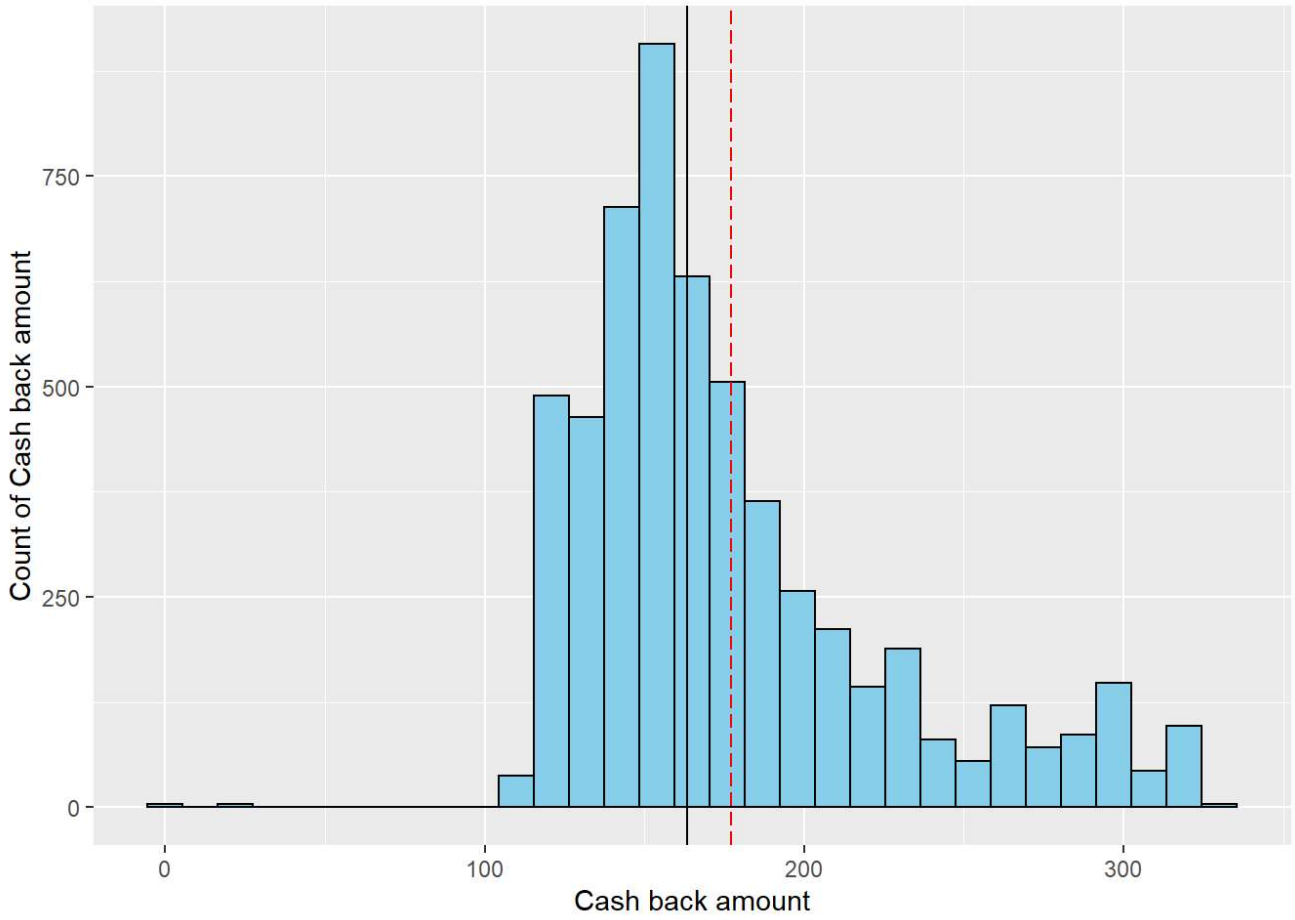
```
cash_median = median(edata$CashbackAmount)
cash_median
```

```
## [1] 163.28
```

- ***Create an appropriate plot to visualize the distribution of this variable.***

Visualizing the variable "***CashbackAmount***" distribution.

```
ggplot(edata, aes(CashbackAmount)) + geom_histogram(binwidth = 11, color="black", fill="skyblue") +
geom_vline(xintercept = cash_mean, linetype="longdash", color="red")  + geom_vline(xintercept = cas
h_median, color="black") + labs(x="Cash back amount", y="Count of Cash back amount")
```



- ***Consider any outliers present in the data. If present, specify the criteria used to identify them and provide a logical explanation for how you handled them.***

Checking for possible outliers

```
t1 <- edata %>% filter(CashbackAmount <= 0)
t1
```

CustomerID	Ch…	Ten…	PreferredLoginDevice	CityTier	WarehouseToHo…	PreferredPaymentMo
<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<chr>
50102	0	10	Computer	3	10	E wallet

CustomerID <dbl>	Ch… <dbl>	Ten… <dbl>	PreferredLoginDevice <chr>	CityTier <dbl>	WarehouseToHo… <dbl>	PreferredPaymentMo <chr>
51027	0	1	Mobile Phone	1	33	Credit Card
51177	0	30	Computer	3	8	Credit Card
51256	0	8	Mobile Phone	3	24	Credit Card

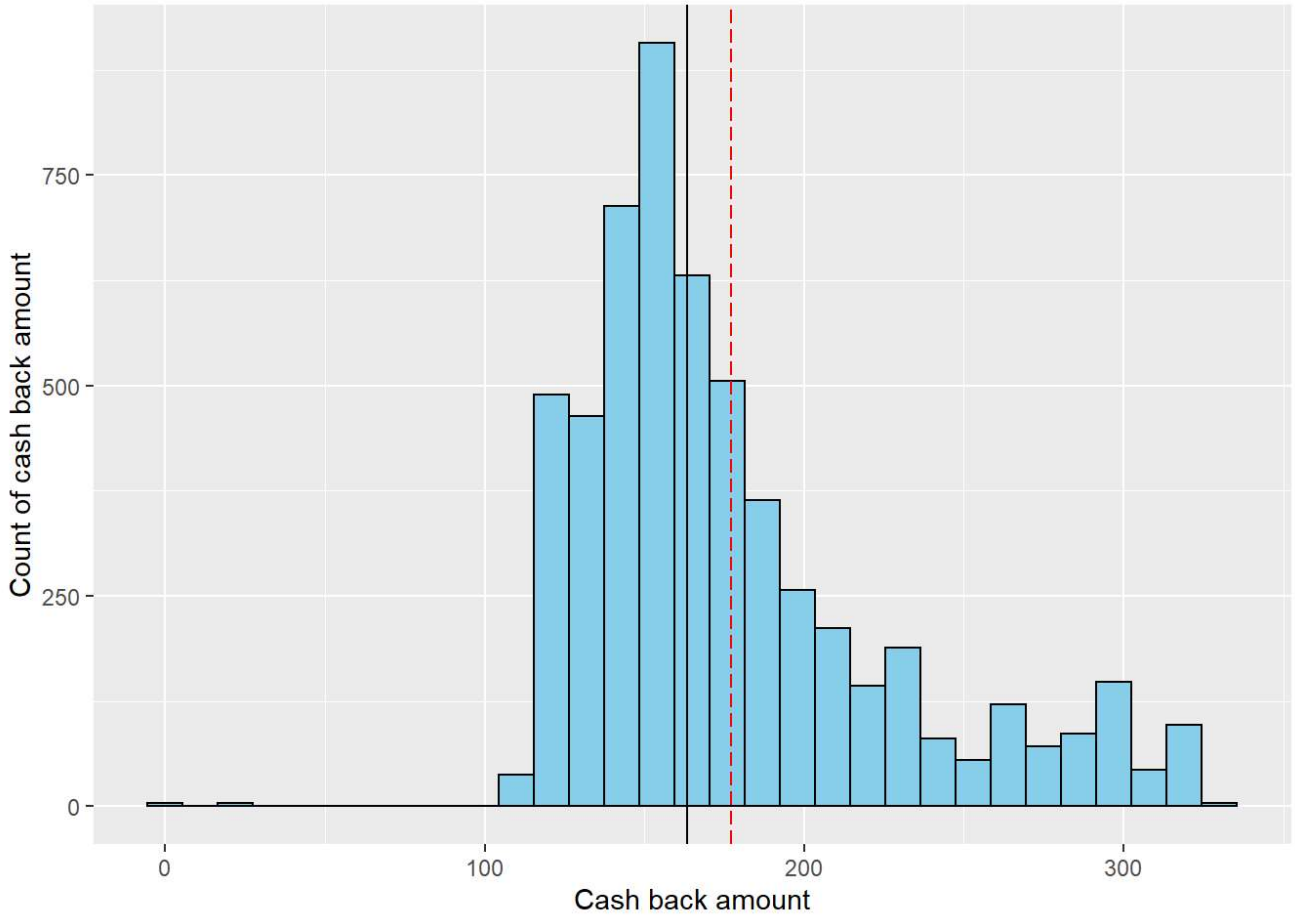
4 rows | 1-8 of 20 columns

There are 4 entries with 0 cashbackamount value with multiple coupons used, this might be due to a different type of coupon used on the order which does not involve cashback.

- ***Describe the shape and skewness of the distribution.***

Shape and Skewness

```
ggplot(edata, aes(CashbackAmount)) + geom_histogram(binwidth = 11, color="black", fill="skyblue") +
geom_vline(xintercept = cash_mean, linetype="longdash", color="red")  + geom_vline(xintercept = cas
h_median, color="black")  + labs(x="Cash back amount", y="Count of cash back amount")
```




```
summary(edata$CashbackAmount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   145.8   163.3   177.2   196.4   325.0
```

On looking at the summary data and the plot looks like the data is right skewed and as it has few high bars and as it not uniform it looks slightly like a multimodel modality.

- ***Based on your answer to the previous question, decide if it is appropriate to apply a transformation to your data. If no, explain why not. If yes, name the transformation applied and visualize the transformed distribution***

As the data is not extremely skewed, this data sample does not need any Transformation.

- ***Choose and calculate an appropriate measure of central tendency***

As the data is skewed Mean is the right measure to calculate central tendency.

- ***Explain why you chose this as your measure of central tendency. Provide supporting evidence for your choice***

```
summary(edata$CashbackAmount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   145.8   163.3   177.2   196.4   325.0
```

The data is right skewed i.e there are very less data lesser than that of Median and as the distribution is not uniform the variance and standard deviation calculation would not result in the proper central tendency value.

- ***Choose and calculate a measure of spread that is appropriate for your chosen measure of central tendency. Explain why you chose this as your measure of spread.***

The spread of the variable is mainly towards the median area where most of the observations live within an Interquatile range of 50.

```
summary(edata$CashbackAmount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   145.8   163.3   177.2   196.4   325.0
```

```
IQR(edata$CashbackAmount)
```

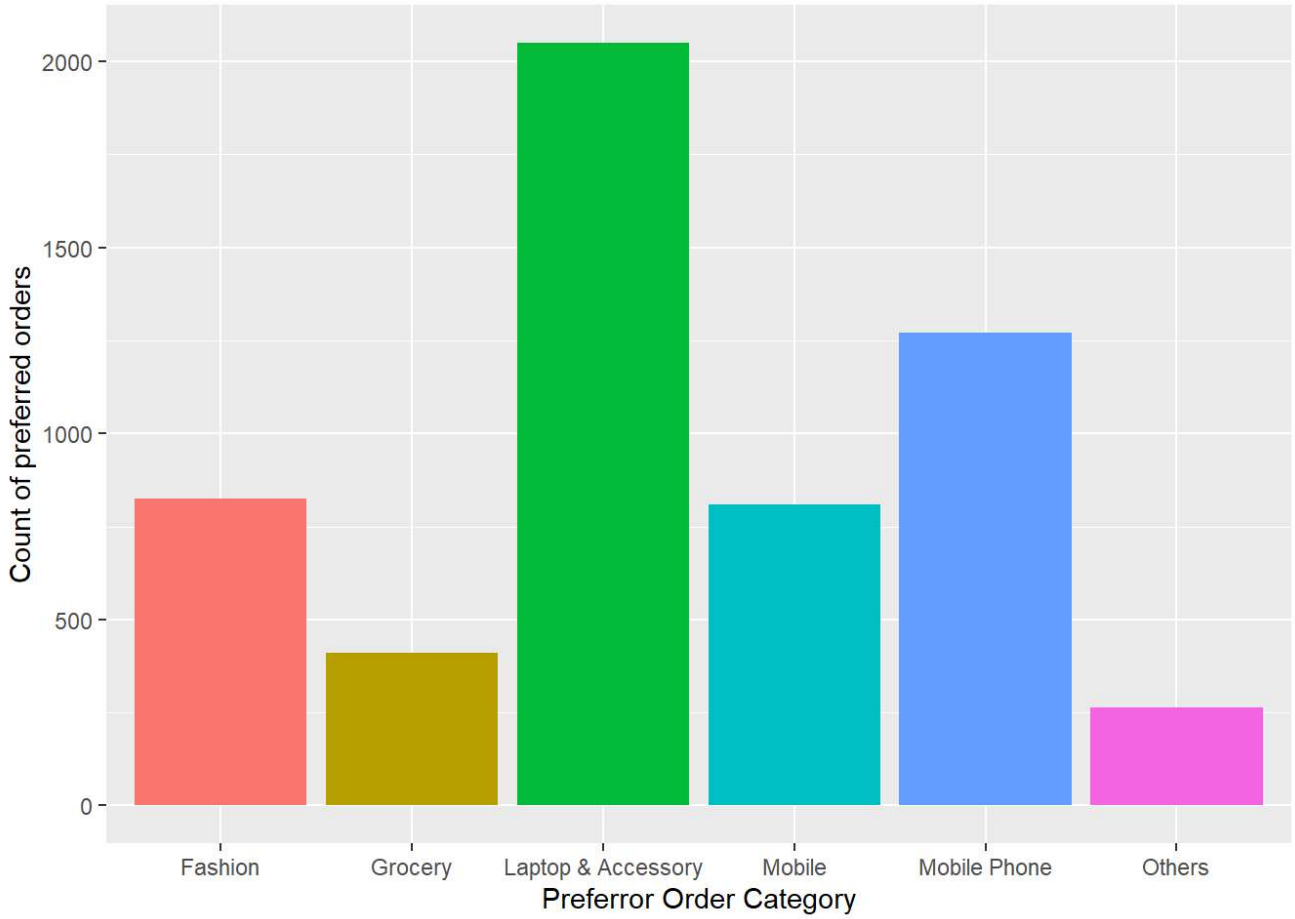
```
## [1] 50.6225
```

Categorical variable

For this exercise we take the variable "***PreferedOrderCat***" as our data.

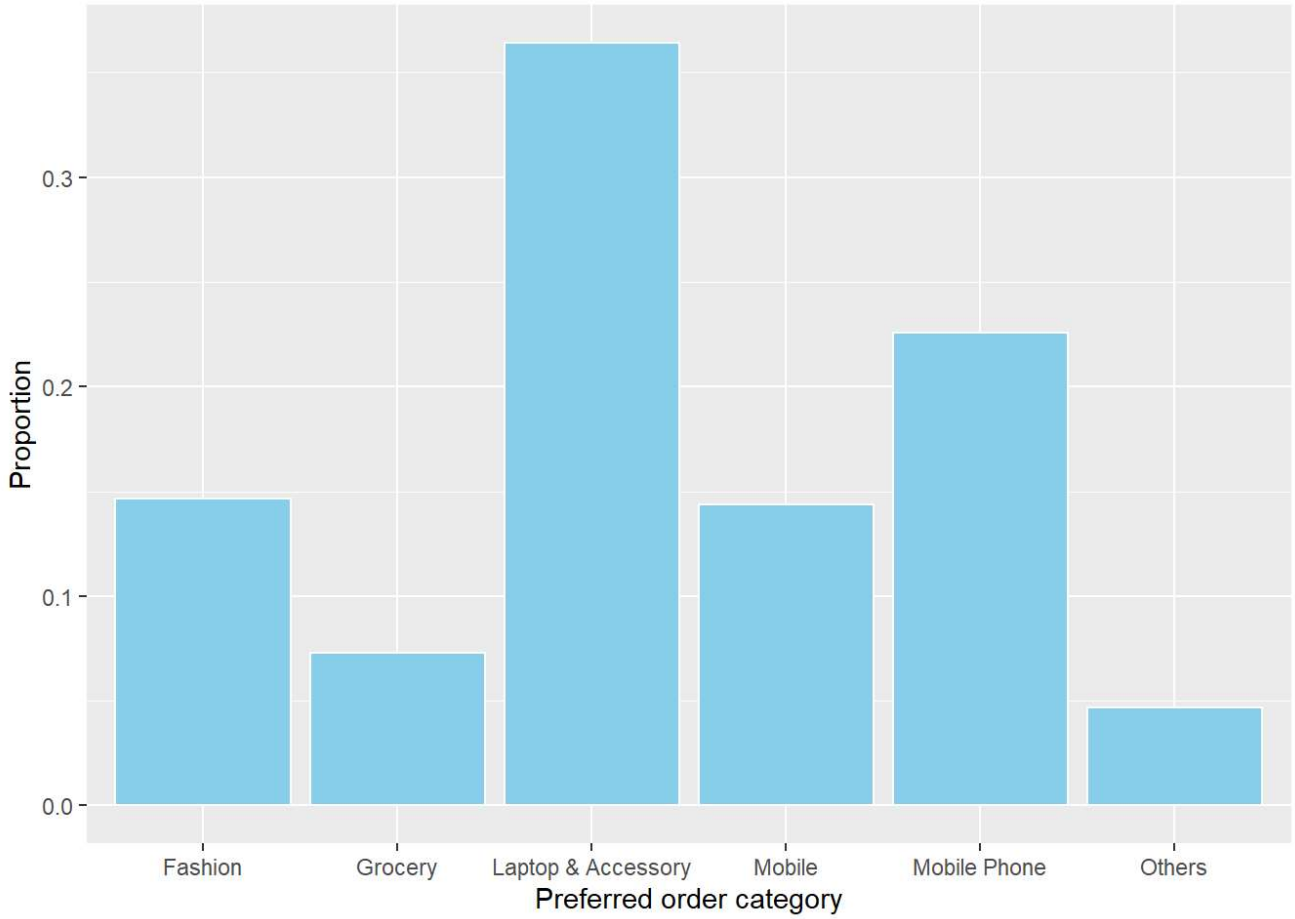
- ***Create an appropriate plot to visualize the distribution of counts for this variable***

```
ggplot(edata, aes(PreferedOrderCat)) + geom_bar(aes(fill=PreferedOrderCat), show.legend=FALSE) + la
bs(x="Preferror Order Category", y="Count of preferred orders")
```



- ***Create an appropriate plot to visualize the distribution of proportions for this variable.***

```
ggplot(edata) + geom_bar(aes(PreferedOrderCat, y=..prop.., group=1), stat='count', fill="skyblue",
 color="white") + labs(x="Preferred order category", y="Proportion")
```



- ***Discuss any unusual observations for this variable.***

There are no major unusual observation recorded for this variable.

```
examine <- edata %>% select(PreferedOrderCat) %>% group_by(PreferedOrderCat) %>% summarise(count = n()) %>% mutate(prop = count / sum(count)) %>% mutate(percent_of_total = (count / sum(count) * 100 ))

examine
```

PreferedOrderCat <chr>	count <int>	prop <dbl>	percent_of_total <dbl>
Fashion	826	0.14671403	14.671403
Grocery	410	0.07282416	7.282416
Laptop & Accessory	2050	0.36412078	36.412078
Mobile	809	0.14369449	14.369449
Mobile Phone	1271	0.22575488	22.575488
Others	264	0.04689165	4.689165

6 rows

- ***Discuss if there are too few/too many unique values.***

```
unique(edata$PreferedOrderCat)
```

```
## [1] "Laptop & Accessory" "Mobile"            "Mobile Phone"
## [4] "Others"             "Fashion"           "Grocery"
```

In total there are 6 unique values available for this variable it is considered as a good number of unique value for this variable.

Bivariate Analysis

Let's select and clean the data.

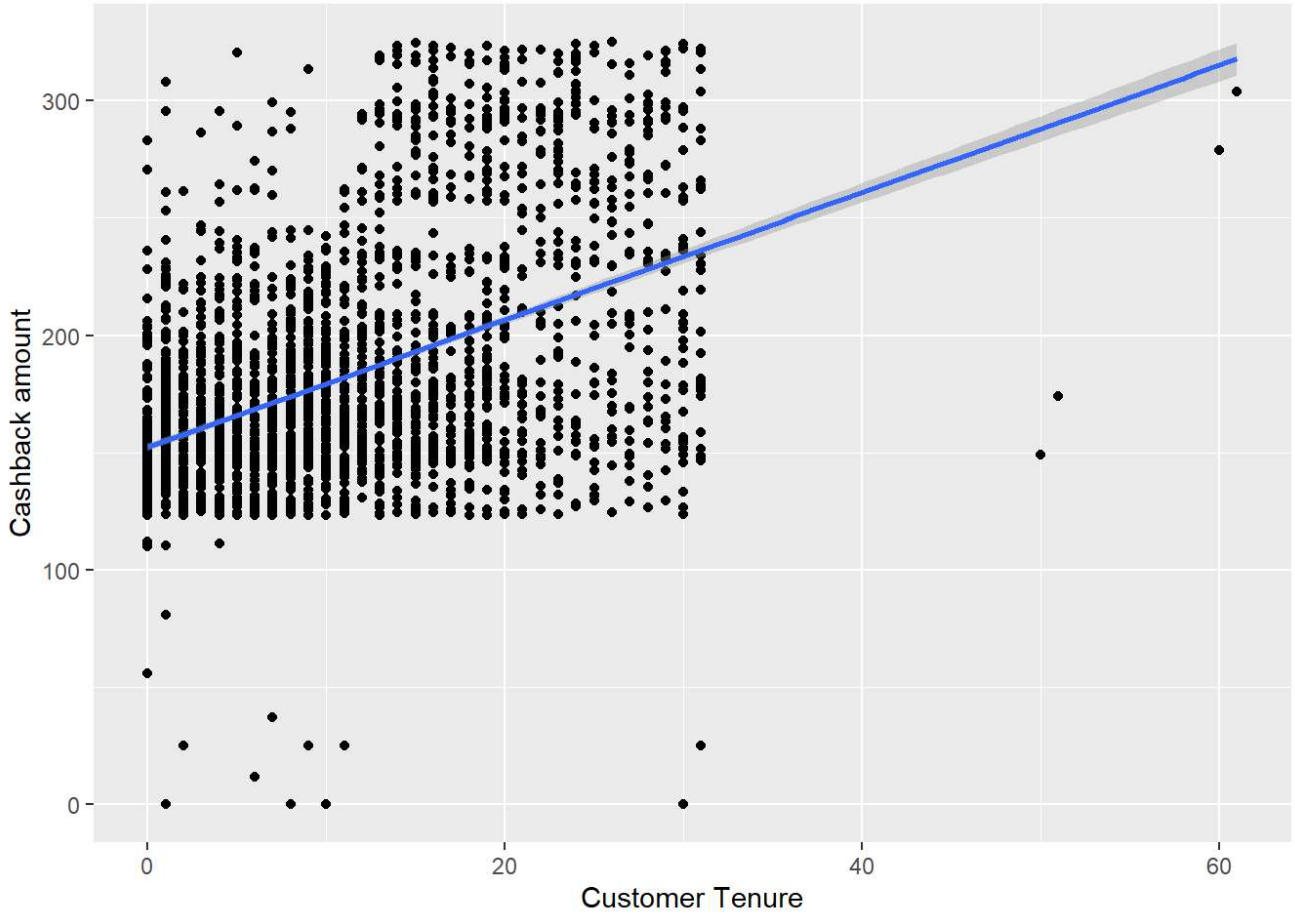
Four this exercise the variables ***MaritalStatus, CashbackAmount and Tenure*** are selected.

```
bi_c <- edata %>% select(MaritalStatus, CashbackAmount, Tenure) %>% drop_na(MaritalStatus, Cashback
Amount, Tenure)
```

Two numerical data

- ***Create an appropriate plot to visualize the relationship between the two variables.***

```
ggplot(bi_c, aes(Tenure, CashbackAmount)) + geom_point() + geom_smooth(method="lm", formula = y ~
 x) + labs(x="Customer Tenure", y="Cashback amount")
```



- ***Describe the form, direction, and strength of the observed relationship. Include both qualitative and quantitative measures, as appropriate.***

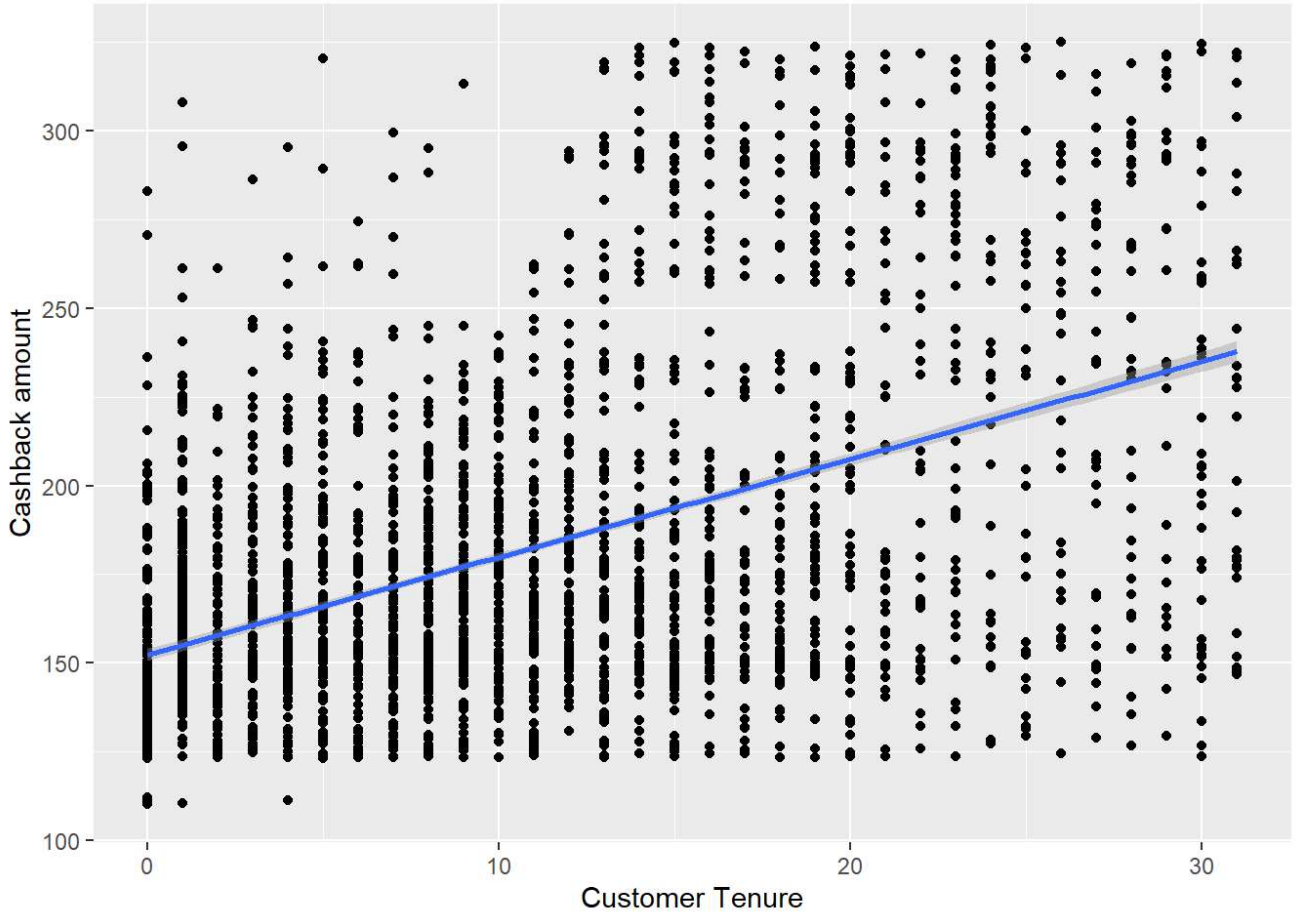
Positive, Strong and Moderately Linear.

The points have a positive trend and though there are few outliers, the points are not spread completely around the plot. Most observations stays close together and as most of the points in lower tenure period has closer observations than higher tenure the graph is moderately linear.

Clearing out outliers for better visualization

```
bi_1 <- bi_c %>% select(CashbackAmount, Tenure) %>% filter(between(Tenure, 0, 40)) %>% filter(betwe
en(CashbackAmount, 100, 400))

ggplot(bi_1, aes(Tenure, CashbackAmount)) + geom_point() + geom_smooth(method="lm", formula= y ~ x)
+ labs(x="Customer Tenure", y="Cashback amount")
```



- ***Explain what this relationship means in the context of the data***

This means that the data between Tenure and CashbackAmount has strong relationship.

As the customer Tenure increases the CashbackAmount claimed by the customer also increases.

- ***Describe the variability that you observe in the plot and how that corresponds to the strength you calculated in #2 above.***

```
mean(bi_c$Tenure)
```

```
## [1] 10.1899
```

```
sd(bi_c$Tenure)
```

```
## [1] 8.557241
```

```
mean(bi_c$CashbackAmount)
```

```
## [1] 180.015
```

```
sd(bi_c$CashbackAmount)
```

```
## [1] 48.72242
```

```
summary(bi_c$CashbackAmount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   147.7   165.4   180.0   199.5   325.0
```

```
summary(bi_c$Tenure)
```

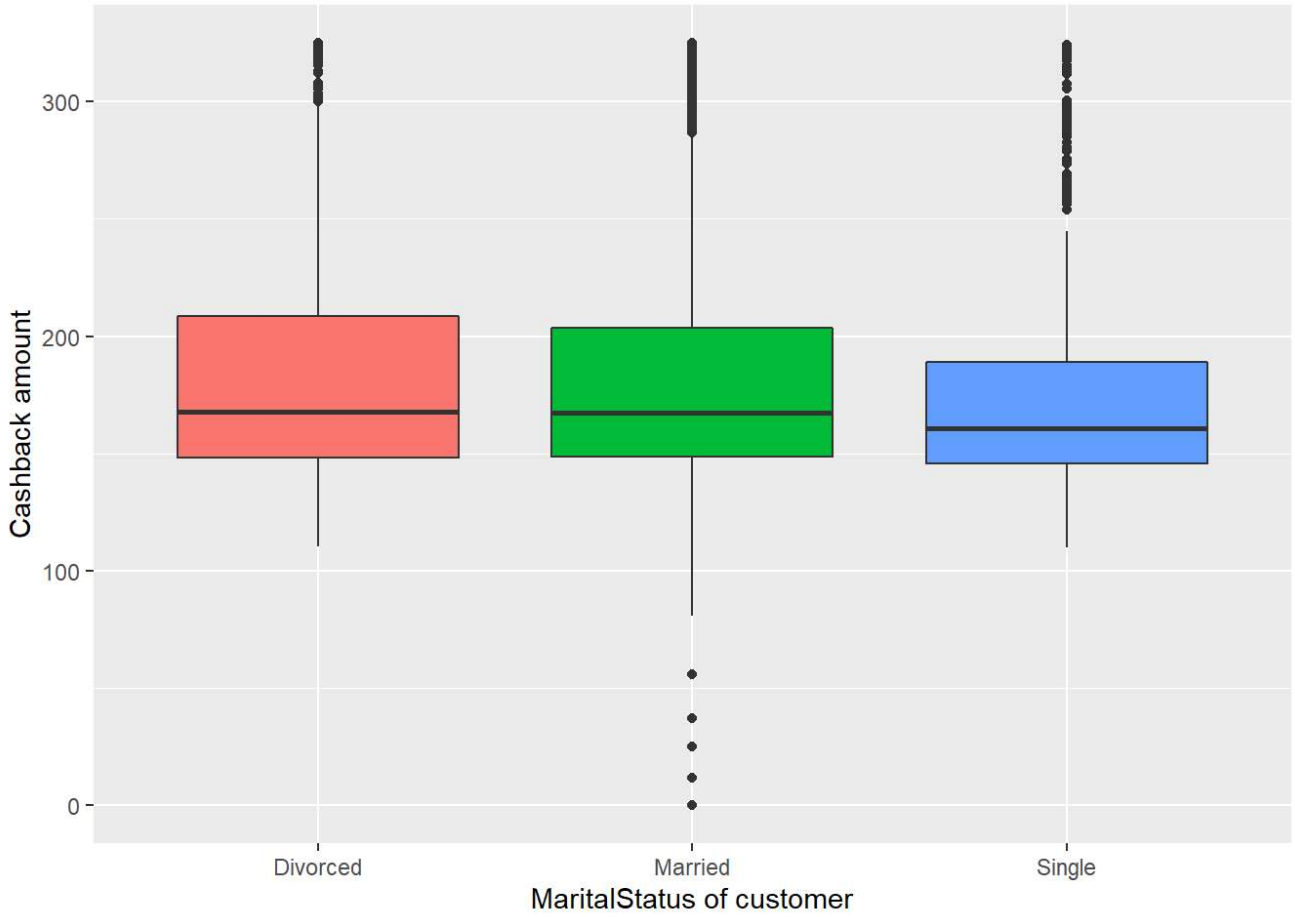
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    9.00   10.19   16.00   61.00
```

When comparing the Median and the IQR range for both the variables it is clear that CashbackAmount variable has a strong relationship with the Tenure and they both are positively related to each other.

One Numeric and one categorical

- ***Create an appropriate plot to visualize the relationship between the two variables.***

```
ggplot(bi_c, aes(MaritalStatus, CashbackAmount)) + geom_boxplot(aes(fill=MaritalStatus), show.legen
d = FALSE) + labs(x="MaritalStatus of customer", y="Cashback amount")
```



- ***Describe the form, direction, and strength of the observed relationship. Include both qualitative and quantitative measures, as appropriate.***

Though there are few suspected outliers the median lies closer to the middle of the IQR range for all the MaritalStatus types of people.
The form appears to be strong as all the category has the median almost closer to the middle of IQR and have a linear and has a neither positive nor negative direction considering the median.

- ***Explain what this relationship means in the context of the data***

From the plot and the context of data we could see that the Divorsed users had got the maximum average of cashback amount and also has the highest IQR of cashback among the other category of users.

- ***Describe the variability that you observe in the plot and how that corresponds to the strength you calculated in #2 above.***

```
tapply(bi_c$CashbackAmount, bi_c$MaritalStatus, mean)
```

```
## Divorced  Married   Single
## 185.2877 182.1564 173.8984
```

```
tapply(bi_c$CashbackAmount, bi_c$MaritalStatus, sd)
```

```
## Divorced  Married   Single
## 51.69184 50.03167 44.25621
```

```
tapply(bi_c$CashbackAmount, bi_c$MaritalStatus, IQR)
```

```
## Divorced  Married   Single
##  60.6800  55.2075  43.0700
```

The mean CashbackAmount for the Divorced peopleis higher when compared to the other two user category also the measure of spread is large on that user category.

Though the means for other user vary slightly they all lie around the middle of the IQR range and thus exhibit a strong relationship. Though single users has the lease when compared to others two category they have the closest measure of spread when compared to other two user category.

References

https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf (https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf)

https://lms.stclaircollege.ca/bbcswebdav/pid-2273873-dt-content-rid-26842374_1/courses/DAB501-21F-001002003004/R4DS_Chapter_7b_soln.html (https://lms.stclaircollege.ca/bbcswebdav/pid-2273873-dt-content-rid-26842374_1/courses/DAB501-21F-001002003004/R4DS_Chapter_7b_soln.html)

https://lms.stclaircollege.ca/webapps/blackboard/execute/content/file?cmd=view&content_id=_2263843_1&course_id=_49673_1 (https://lms.stclaircollege.ca/webapps/blackboard/execute/content/file?cmd=view&content_id=_2263843_1&course_id=_49673_1)

https://runestone.academy/runestone/books/published/ac1/scatter_plots_and_correlation/describing_scatter_plots.html (https://runestone.academy/runestone/books/published/ac1/scatter_plots_and_correlation/describing_scatter_plots.html)