Lab 06

Jyosna Philip

2024-01-03

# About the dataset

This dataset contains online shopping data. It includes categorical variables like Gender,Coupon Status and Location.

online\_shopping <- read.csv("C:/Users/jyosn/Downloads/archive (2)/online\_shopping.csv")  
  
  
online\_shopping<- na.omit(online\_shopping)  
  
  
dim(online\_shopping)

## [1] 52524 21

The dataset has 21 columns and 52524 entries.

summary(online\_shopping)

## X CustomerID Gender Location   
## Min. : 0 Min. :12346 Length:52524 Length:52524   
## 1st Qu.:13171 1st Qu.:13869 Class :character Class :character   
## Median :26362 Median :15311 Mode :character Mode :character   
## Mean :26372 Mean :15345   
## 3rd Qu.:39557 3rd Qu.:16996   
## Max. :52923 Max. :18283   
## Tenure\_Months Transaction\_ID Transaction\_Date Product\_SKU   
## Min. : 2.00 Min. :16679 Length:52524 Length:52524   
## 1st Qu.:15.00 1st Qu.:25404 Class :character Class :character   
## Median :27.00 Median :32678 Mode :character Mode :character   
## Mean :26.14 Mean :32445   
## 3rd Qu.:37.00 3rd Qu.:39174   
## Max. :50.00 Max. :48497   
## Product\_Description Product\_Category Quantity Avg\_Price   
## Length:52524 Length:52524 Min. : 1.0 Min. : 0.39   
## Class :character Class :character 1st Qu.: 1.0 1st Qu.: 5.70   
## Mode :character Mode :character Median : 1.0 Median : 16.99   
## Mean : 4.5 Mean : 52.43   
## 3rd Qu.: 2.0 3rd Qu.:119.00   
## Max. :900.0 Max. :355.74   
## Delivery\_Charges Coupon\_Status GST Date   
## Min. : 0.00 Length:52524 Min. :0.0500 Length:52524   
## 1st Qu.: 6.00 Class :character 1st Qu.:0.1000 Class :character   
## Median : 6.00 Mode :character Median :0.1800 Mode :character   
## Mean : 10.48 Mean :0.1374   
## 3rd Qu.: 6.50 3rd Qu.:0.1800   
## Max. :521.36 Max. :0.1800   
## Offline\_Spend Online\_Spend Month Coupon\_Code   
## Min. : 500 Min. : 320.2 Min. : 1.000 Length:52524   
## 1st Qu.:2500 1st Qu.:1252.6 1st Qu.: 4.000 Class :character   
## Median :3000 Median :1837.9 Median : 7.000 Mode :character   
## Mean :2832 Mean :1894.2 Mean : 6.666   
## 3rd Qu.:3500 3rd Qu.:2425.3 3rd Qu.: 9.000   
## Max. :5000 Max. :4556.9 Max. :12.000   
## Discount\_pct   
## Min. :10.00   
## 1st Qu.:10.00   
## Median :20.00   
## Mean :19.95   
## 3rd Qu.:30.00   
## Max. :30.00

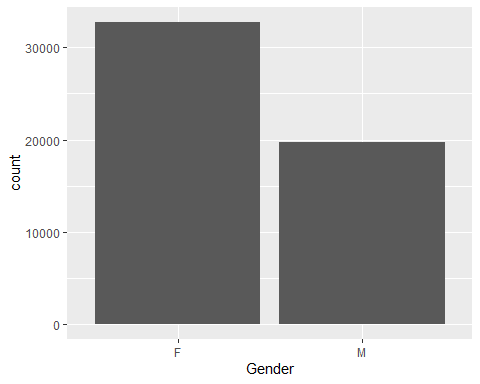
We can see that the continuous variables are Avg\_Price, Delivery\_Charges,GST, Online\_Spend,Offline\_Spend. On the other hand we have categorical variables Gender, Location, Tenure\_Months, Coupon\_Status. Avg\_Price ranges from 0.39 to 355.74. Delivery\_Charges ranges from 0 to 521.36. GST ranges from 0.05 to 0.18 Offline\_Spend ranges from 500 to 5000. Online\_Spend ranges from 320.2 to 4556.9.

5.AIM:Perform the visualization for all categorical and target variables

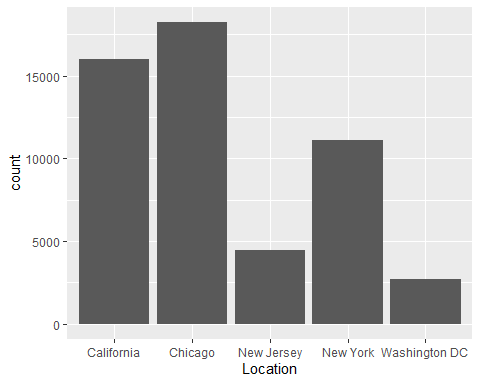
Categorical variables are Gender, Location, Tenure\_Months, Coupon\_Status

library(ggplot2)

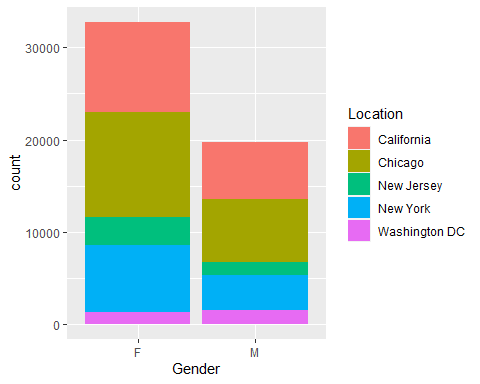
#visualisation for categorical variable : Gender  
ggplot(online\_shopping,aes(x=Gender))+geom\_bar()



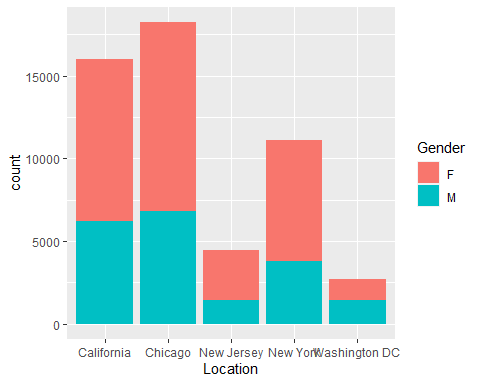
#visualisation for categorical variable : Location  
ggplot(online\_shopping,aes(x=Location))+geom\_bar()



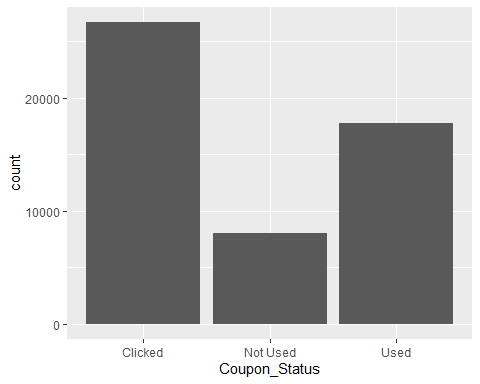
#Visualisation for gender and Location  
ggplot(online\_shopping,aes(x=Gender,fill=Location))+geom\_bar()



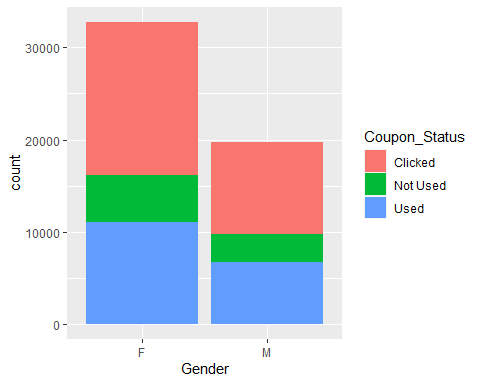
ggplot(online\_shopping,aes(x=Location,fill=Gender))+geom\_bar()



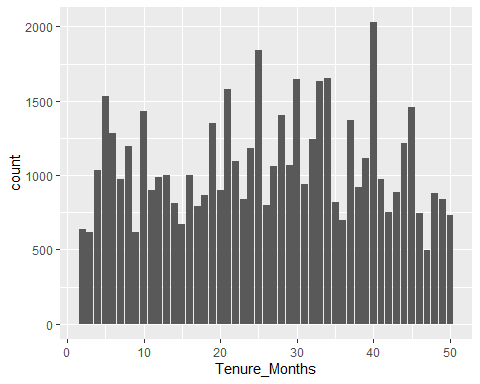
#visualisation for categorical variable : Coupon Status  
ggplot(online\_shopping,aes(x=Coupon\_Status))+geom\_bar()



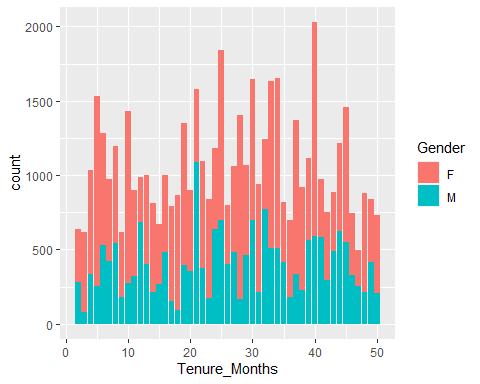
#visualisation for coupon status and gender  
ggplot(online\_shopping,aes(x=Gender,fill=Coupon\_Status))+geom\_bar()



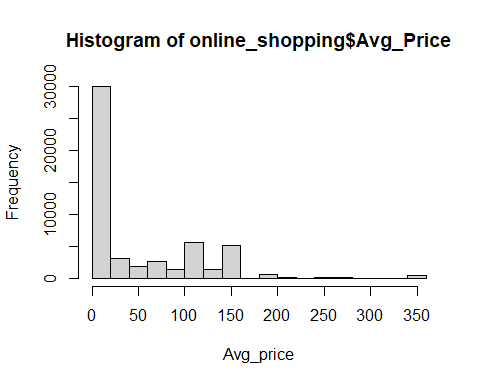
#visualisation for categorical variable :Tenure\_Months  
ggplot(online\_shopping,aes(x=Tenure\_Months))+geom\_bar()



#visualisation for Tenure\_Months and Gender  
ggplot(online\_shopping,aes(x=Tenure\_Months,fill=Gender))+geom\_bar()



#Visualisation for target variable: Avg\_Price  
hist(online\_shopping$Avg\_Price,xlab="Avg\_price",ylab="Frequency")



1. AIM: Perform the single proportion test when σ is known and interpret the results.

Our target variable is average\_price

mean(online\_shopping$Avg\_Price)

## [1] 52.42909

sd(online\_shopping$Avg\_Price)

## [1] 64.15665

Hence, the mean and standard deviation of avg\_price is 52.4 and 64.1 respectively.

s=sample(online\_shopping$Avg\_Price,100,replace = FALSE)  
mean(s)

## [1] 61.8229

The mean of sample is 48.77.

sd\_s=sd(s)  
sd\_s

## [1] 75.93352

The standard deviation of the sample is 63.62.

TWO-SIDED HYPOTHESIS TEST H0:mu=51 H1: mu!=51

library(BSDA)

## Loading required package: lattice

##   
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':  
##   
## Orange

z.test(x=s,alternative="two.sided",mu=51,sigma.x=sd\_s,conf.level=0.95)

##   
## One-sample z-Test  
##   
## data: s  
## z = 1.4253, p-value = 0.1541  
## alternative hypothesis: true mean is not equal to 51  
## 95 percent confidence interval:  
## 46.9402 76.7056  
## sample estimates:  
## mean of x   
## 61.8229

p-value is greater than 0.05. Hence we fail to reject H0. We can conclude that population mean is 51.

Conclusion: The population mean of Avg\_Price is 51.

1. Let’s consider an assertion that states 30% of students prefer online learning. A survey was conducted among a random sample of 600 students, revealing that 180 students favour online learning. Can we validate this claim at a 5% significance level based on the sample data?

AIM: To carry out single proportion test when variance is unknown Let P0=Proportion of students favouring online learning

H0: P0=0.3 H1: P0 != 0.3

n=600  
x=180  
prop.test(x,n,p=0.3,alternative = "two.sided",conf.level = 0.05)

##   
## 1-sample proportions test without continuity correction  
##   
## data: x out of n, null probability 0.3  
## X-squared = 0, df = 1, p-value = 1  
## alternative hypothesis: true p is not equal to 0.3  
## 5 percent confidence interval:  
## 0.2988282 0.3011744  
## sample estimates:  
## p   
## 0.3

Here p-value =1 p-value>0.05,Hence we fail to reject H0.

Conclusion: We can validate the claim that 30% students prefer online learning. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_