dataframe$Satisfaction <- as.numeric(as.character(t(dataframe$Satisfaction)))

dataframe$Satisfaction[is.na(dataframe$Satisfaction)] <- 3.5

cor(dataframe$Shopping.Amount.at.Airport, dataframe$Satisfaction)

clim = c(min(Shopping.Amount.at.Airport),max(Shopping.Amount.at.Airport))

plot(Shopping.Amount.at.Airport, Satisfaction, xlim=clim)

df$southeast<-as.factor(trimws(df$Airline.Name)=='Southeast Airlines Co.')

* Fill missing values(with mean, median)
* Try a model to predict the value of the missing variables
* Create dummy variables
* Outlier detection (zscore, lower fence upper fence)
* and treatment(winsorize, trimming)
* EDA ignoring missing values

Jay : 2 - 6

Eashani : 7 - 10

Akshita: 11-14

Yesaswi : 15-18

Rohan: 19-23

Tasks: Missing Values, Outliers, EDA (Southeast vs. Other airlines and Overall), correlation

**Attributes Name:**

1.   **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer?

a.       5 means higher satisfied, and 1 is lowest level of satisfaction.

2.   **Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.

3.   **Age** – the specific customer’s age. That is starting from 15 to 85 years old.

4.   **Gender** – male or female.

5.   **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.

6.   **Year of First Flight** – this attributes shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.

7.   **No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.

8.   **Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many time that customer fly with other Airlines.

9.   **Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation

10.   **No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefits such as, discounts.

11.   **Shopping Amount at Airport** – showing the costumer’s result of how many products have been purchased. The range of shopping amount is from 0 to 875.

12.   **Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.

13.   **Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.

14.   **Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.

15.   **Flight date** – all of these data are abbreviate the passenger’s flight date travel, which were since 2014 and only in January, February, and March.

16.   **Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and FlyToSun Airlines Inc. This attribute provide what airline name that passenger have been used.   *Average satisfaction by airline*

17.   **Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.

18.   **Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage are starting from 0 until 1128 minutes.

19.   **Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang  of delayed minutes in this data are starting from 0 until 1115 minutes.

20.   **Flight Cancelled** – occurs when the airline does not operates the flight at all, and that is for a certain reason.

21.   **Flight time in minutes** – indicate to period time to the destination.

22.   **Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.

23.   **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

Code:

#using this for Age

plot<-ggplot(df,aes(Age,Satisfaction))+geom\_count()

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$age,df$Satisfaction)

#using this for Airline Status

plot1<-ggplot(df,aes(Airline.Status,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot1+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Satisfaction,as.numeric(df$Airline.Status))

#using this for gender

plot2<-ggplot(df,aes(Gender,Satisfaction))+facet\_grid(southeast ~ .)+geom\_count()

plot2+stat\_summary(aes(y=Satisfaction,fill=southeast),fun.y="mean",size=2,colour='red',geom="point")

cor(df$Satisfaction,as.numeric(df$Gender))

#using this for price sensitivity

plot3<-ggplot(df,aes(Price.Sensitivity,Satisfaction))+facet\_grid(southeast ~ .)+geom\_count()

plot3+stat\_summary(aes(y=Satisfaction),fun.y="mean",size=2,colour='red',geom="point")

cor(df$Satisfaction,df$Price.Sensitivity)

#using this for year of first flight

unique(df$Year.of.First.Flight)

plot5<-ggplot(df,aes(as.factor(Year.of.First.Flight),Satisfaction))+facet\_grid(southeast ~ .)+geom\_count()

plot5+stat\_summary(aes(y=Satisfaction),fun.y="mean",size=2,colour='red',geom="point")

cor(df$Satisfaction,df$Year.of.First.Flight)

#comparative statistics

plot4<-ggplot(df,aes(southeast))

plot4+stat\_summary(aes(y=Satisfaction),fun.y="mean",geom="bar")

plot4+stat\_summary(aes(y=Age),fun.y="mean",geom="bar")

plot4+stat\_summary(aes(y=Price.Sensitivity),fun.y="mean",geom="bar")

#generating proportion plots

plot6<-ggplot(df)+geom\_bar(mapping=aes(x=Satisfaction,fill=southeast),position="fill",width=0.4)

plot6

ggplot(df)+geom\_bar(mapping=aes(x=Gender,fill=southeast),position="fill",width=0.4)

ggplot(df)+geom\_bar(mapping=aes(x=Airline.Status,fill=southeast),position="fill",width=0.4)

ggplot(df)+geom\_bar(mapping=aes(x=Year.of.First.Flight,color=southeast,fill=Airline.Status),position="fill",width=0.4)

ggplot(df)+geom\_bar(mapping=aes(x=Price.Sensitivity,fill=southeast),position="fill",width=0.4)

ggplot(df)+geom\_bar(mapping=aes(x=Age,fill=southeast),position="fill",width=0.4)

F=sat$Flight.date

F

sat$Satisfaction=as.numeric((as.character((t(sat$Satisfaction)))))

class(sat$Flight.date)

sat$Flight.date=as.Date(sat$Flight.date, format="%m/%d/%Y")

class(sat$Flight.date)

plot(sat$Flight.date,sat$Satisfaction)

sat$week\_days=weekdays(sat$Flight.date)

plot(as.factor(sat$week\_days),sat$Satisfaction)

class(sat$week\_days)

head(sat$Flight.date)

install.packages("ggplot2")

library(ggplot2)

fD=ggplot(sat, aes(x=Flight.date, y=Satisfaction))

fD=fD+ geom\_line(aes(color = Flight.date))

fD=fD+ ggtitle("Flight Date")

fD

fD1=ggplot(sat, aes(x=week\_days, y=Satisfaction))

fD1=fD1+ geom\_count()

fD1=fD1+ ggtitle("Flight weekdays")

fD1

fD2=ggplot(sat, aes(x=Airline.Name, y=Satisfaction))

fD2=fD2+ geom\_point()

fD2=fD2+theme(axis.text.x = element\_text(angle = 90, hjust = 1))

fD2=fD2+ ggtitle("Airline Satisfaction")

fD2

fD3=ggplot(sat, aes(x=sat$Scheduled.Departure.Hour, y=Satisfaction))

fD3=fD3+ geom\_point()

fD3=fD3+ ggtitle("Scheduled departure hour")

fD3

dep=hist(sat$Departure.Delay.in.Minutes)

#avgDepDelay=mean(sat$Departure.Delay.in.Minutes)

#avgDepDelay

fD4=ggplot(sat, aes(x=Departure.Delay.in.Minutes, y=Satisfaction))

fD4=fD4+geom\_count()

fD4=fD4+ ggtitle("Departure delay")

fD4

# 19 - 23 - Rohan

plot<-ggplot(df,aes(Flight.Distance,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Flight.Distance,df$Satisfaction)

plot<-ggplot(df,aes(Arrival.Delay.in.Minutes,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Flight.Arrival.Delay.in.Minutes,df$Satisfaction)

plot<-ggplot(df,aes(Departure.Delay.in.Minutes,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Departure.Delay.in.Minutes,df$Satisfaction)

plot<-ggplot(df,aes(Flight.cancelled,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Flight.cancelled,df$Satisfaction)

plot<-ggplot(df,aes(Arrival.Delay.greater.5.Mins,Satisfaction))+geom\_count()+facet\_grid(southeast ~ .)

plot+stat\_summary(fun.y = "mean", colour = "red", size = 2, geom = "point")

cor(df$Arrival.Delay.greater.5.Mins,df$Satisfaction)

11-14

#Class variable

ggplot(dataframe, aes(x = Class,y = Satisfaction))+facet\_grid(southeast ~ .)+geom\_count()+

  stat\_summary(aes(y=Satisfaction),fun.y="mean",size=2,colour='red',geom="point")

#For airlines other than Southeast,

#Business class has highest rating (average) at 3.5, followed by Eco and then Eco plus

#The same trend is followed by Southeast airlines

#Day of month

ggplot(dataframe , aes( x=Day.of.Month, y=Satisfaction)) + facet\_grid(. ~ southeast)+

  stat\_summary(fun.y = "mean", geom="col", color = "red", fill = "white")

cor(dataframe$Satisfaction, dataframe$Day.of.Month)

#For airlines other than Southeast,

#the satisfaction value averages around 3.3 for almost all days of month

#For Southeast airlines, the satisfaction value differs for different days.

#For the 4th, 7th, 10th days, it is 3.5 with values lesser than that for the other days

#The correlation between the two variables is 1.3% which is not very high

#Shopping amount at Airport

ggplot(dataframe , aes( x=Shopping.Amount.at.Airport, y=Satisfaction)) + facet\_grid(. ~ southeast)+

  stat\_summary\_bin(fun.y = "mean", geom="col", binwidth = 100, color = "red", fill = "white")

cor(dataframe$Shopping.Amount.at.Airport, dataframe$Satisfaction)

#For airlines other than Southeast,

#Shopping amount ranges from 0 to 900

#with customers who spend 700-800 units rating the Satisfaction at 4 (max) and

#customers who spend 800-900 units rating the Satisfaction at 3 (min)

#For Southeast airlines,

#Shopping amount ranges from 0 to 600

#with customers who spend 500-600 units rating the Satisfaction at 4 (max) and

#customers who spend 400-500 units rating the satisfaction at around 3.2 (min)

#This gives us an insight that customers do not spend more than 600 units on

#Shopping at airport when they are a Southeast customer

#The correlation between the two variables is 1.7% which is not very high

#Eating and drinking at airport

ggplot(dataframe , aes( x=Eating.and.Drinking.at.Airport, y=Satisfaction)) + facet\_grid(. ~ southeast)+

  stat\_summary\_bin(fun.y = "mean", geom="col", binwidth = 100, color = "red", fill = "white")

cor(dataframe$Satisfaction, dataframe$Eating.and.Drinking.at.Airport)

#For airlines other than Southeast,

#Eating and Drinking amount ranges from 0 to 900

#with customers who spend 700-800 units rating the Satisfaction at 3.8 (max) and

#customers who spend 500-600 units rating the Satisfaction at 3 (min)

#For Southeast airlines,

#Eating and Drinking amount ranges from 0 to 800

#with customers who spend 700-800 units rating the Satisfaction at 5 (max) and

#customers who spend 200-300 units rating the satisfaction at around 3.3 (min)

#The 750 value could be an outlier (since we are getting an average rating of 5)

#The correlation between the two variables is 0.01% which is not very high