FINAL  
I ran some initial analysis with the data set:A diagram of a graph

Description automatically generated with medium confidenceA blue squares with black text

Description automatically generatedA graph with numbers and a bar chart

Description automatically generated

## R Markdown

#Hi, First will load the data from my computer and keep the strings as strings and not variables

> data <- read.csv("Downloads/auto-mpg(1).csv", stringsAsFactors = FALSE)

#Now, we check the structure and head

> str(data)  
'data.frame': 398 obs. of 9 variables:  
 $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...  
 $ cylinder : int 8 8 8 8 8 8 8 8 8 8 ...  
 $ displacement: num 307 350 318 304 302 429 454 440 455 390 ...  
 $ horsepower : chr "130" "165" "150" "150" ...  
 $ weight : int 3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...  
 $ acceleration: num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...  
 $ model.year : int 70 70 70 70 70 70 70 70 70 70 ...  
 $ origin : int 1 1 1 1 1 1 1 1 1 1 ...  
 $ car.name : chr "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...  
> head(data)  
 mpg cylinder displacement horsepower weight acceleration model.year origin  
1 18 8 307 130 3504 12.0 70 1  
2 15 8 350 165 3693 11.5 70 1  
3 18 8 318 150 3436 11.0 70 1  
4 16 8 304 150 3433 12.0 70 1  
5 17 8 302 140 3449 10.5 70 1  
6 15 8 429 198 4341 10.0 70 1  
 car.name  
1 chevrolet chevelle malibu  
2 buick skylark 320  
3 plymouth satellite  
4 amc rebel sst  
5 ford torino  
6 ford galaxie 500

#We also want to remove the missing variables to clean up the data.

> data <- na.omit(data)

# Now, I am going to split the data into two datasets for the project

> data1 <- data[1:300, ]  
> data2 <- data[301:nrow(data), ]

# I will perform simple linear regression on the first dataset

> simple\_lm <- lm(mpg ~ weight, data = data1)

#to see a summary:

> summary(simple\_lm)  
  
Call:  
lm(formula = mpg ~ weight, data = data1)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-9.1077 -1.8842 -0.0333 1.7275 15.1232   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 40.3879027 0.6368804 63.41 <2e-16 \*\*\*  
weight -0.0062524 0.0001957 -31.96 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 2.992 on 298 degrees of freedom  
Multiple R-squared: 0.7741, Adjusted R-squared: 0.7733   
F-statistic: 1021 on 1 and 298 DF, p-value: < 2.2e-16  
  
> print("The multiple R squared is 0.7741  
 The adjusted R squared is 0.7733")  
[1] "The multiple R squared is 0.7741\n+ The adjusted R squared is 0.7733"  
> print("The complete linear regression equation is:   
 mpg = 40.3879027(intercept) - 0.0062524(slope) \* weight")  
[1] "The complete linear regression equation is: \n mpg = 40.3879027(intercept) - 0.0062524(slope) \* weight"

The multiple R squared is 0.7741

The adjusted R squared is 0.7733

The complete linear regression equation is:   
 mpg = 40.3879027(intercept) - 0.0062524(slope) \* weight"

#Next, I will perform multiple linear regression

> multiple\_lm <- lm(mpg ~ weight + as.numeric(horsepower) + displacement, data = data1)  
Warning message:  
In eval(predvars, data, env) : NAs introduced by coercion  
> data1$horsepower <- gsub("[^0-9.]", "", data1$horsepower)  
> multiple\_lm <- lm(mpg ~ weight + as.numeric(horsepower) + displacement, data = data1)

#to see the model:

> summary(multiple\_lm)  
  
Call:  
lm(formula = mpg ~ weight + as.numeric(horsepower) + displacement,   
 data = data1)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-8.9396 -1.9036 -0.0611 1.6062 14.7474   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 39.3739544 0.9210731 42.748 <2e-16 \*\*\*  
weight -0.0047898 0.0005328 -8.991 <2e-16 \*\*\*  
as.numeric(horsepower) -0.0205727 0.0096748 -2.126 0.0343 \*   
displacement -0.0058457 0.0049625 -1.178 0.2398   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 2.95 on 294 degrees of freedom  
 (2 observations deleted due to missingness)  
Multiple R-squared: 0.783, Adjusted R-squared: 0.7808   
F-statistic: 353.6 on 3 and 294 DF, p-value: < 2.2e-16  
  
> print("The multiple R squared is 0.783  
+ The adjusted R squared is 0.7808")  
[1] "The multiple R squared is 0.783\n The adjusted R squared is 0.7808"  
> print("The complete multiple linear regression equation is:   
+ mpg = 39.3739544 - 0.0047898 \* weight - 0.0205727 \* horsepower - 0.0058457 \* displacement")  
[1] "The complete multiple linear regression equation is: \n mpg = 39.3739544 - 0.0047898 \* weight - 0.0205727 \* horsepower - 0.0058457 \* displacement"

The multiple R squared is 0.783

The adjusted R squared is 0.783

The complete multiple linear regression equation is:

mpg = 39.3739544 - 0.0047898 \* weight - 0.0205727 \* horsepower - 0.0058457 \* displacement

#Now, we will use the remaining 98 samples to run predictions

#first, omit the na variables and then run a prediction

> data2 <- na.omit(data2)  
> data2$predicted\_mpg <- predict(multiple\_lm, newdata = data2)  
Warning message:  
In eval(predvars, data, env) : NAs introduced by coercion  
> data2 <- na.omit(data2)  
> data2$predicted\_mpg <- predict(multiple\_lm, newdata = data2)

#Now, we will compare it with the data

> comparison <- data.frame(Actual = data2$mpg, Predicted = data2$predicted\_mpg)  
> head(comparison)  
 Actual Predicted  
1 23.9 19.62136  
2 34.2 26.78247  
3 34.5 27.02196  
4 31.8 27.86442  
5 37.3 27.22017  
6 28.4 23.85090

#We make a residual plot

> residuals <- data2$mpg - data2$predicted\_mpg  
> plot(data2$predicted\_mpg, residuals,  
+ xlab = "Predicted MPG", ylab = "Residuals",  
+ main = "Residual Plot")  
> abline(h = 0, col = "hot pink")

A graph of a plot

Description automatically generated with medium confidence

#The spread has no specific pattern and has some outliers but most of the residuals are not far from the axis so the model is mostly accurate.

#Now, let’s make a histogram of the residuals

> hist(residuals,  
+ main = "Histogram of Residuals",  
+ xlab = "Residuals",  
+ col = "lightblue",  
+ breaks = 15)

A graph of a person with blue bars

Description automatically generated with medium confidence

PART 2

#First, we load the dataset

call\_center\_data <- read.csv("Downloads/Call\_Center.csv")

#Now, we will look into the data

str(call\_center\_data)  
'data.frame': 32941 obs. of 12 variables:  
 $ Id : chr "DKK-57076809-w-055481-fU" "QGK-72219678-w-102139-KY" "GYJ-30025932-A-023015-LD" "ZJI-96807559-i-620008-m7" ...  
 $ Call.Timestamp : chr "10/29/20 0:00" "10/5/20 0:00" "10/4/20 0:00" "10/17/20 0:00" ...  
 $ Call.Centres.City : chr "Los Angeles" "Baltimore" "Los Angeles" "Los Angeles" ...  
 $ Channel : chr "Call-Center" "Chatbot" "Call-Center" "Chatbot" ...  
 $ City : chr "Detroit" "Spartanburg" "Gainesville" "Portland" ...  
 $ Customer.Name : chr "Analise Gairdner" "Crichton Kidsley" "Averill Brundrett" "Noreen Lafflina" ...  
 $ Reason : chr "Billing Question" "Service Outage" "Billing Question" "Billing Question" ...  
 $ Response.Time : chr "Within SLA" "Within SLA" "Above SLA" "Within SLA" ...  
 $ Sentiment : chr "Neutral" "Very Positive" "Negative" "Very Negative" ...  
 $ State : chr "Michigan" "South Carolina" "Florida" "Oregon" ...  
 $ Call.Duration.In.Minutes: int 17 23 45 12 23 25 31 37 37 12 ...  
 $ Csat.Score : int 7 NA NA 1 NA 5 8 NA NA NA ...  
head(call\_center\_data)  
 Id Call.Timestamp Call.Centres.City Channel City Customer.Name  
1 DKK-57076809-w-055481-fU 10/29/20 0:00 Los Angeles Call-Center Detroit Analise Gairdner  
2 QGK-72219678-w-102139-KY 10/5/20 0:00 Baltimore Chatbot Spartanburg Crichton Kidsley  
3 GYJ-30025932-A-023015-LD 10/4/20 0:00 Los Angeles Call-Center Gainesville Averill Brundrett  
4 ZJI-96807559-i-620008-m7 10/17/20 0:00 Los Angeles Chatbot Portland Noreen Lafflina  
5 DDU-69451719-O-176482-Fm 10/17/20 0:00 Los Angeles Call-Center Fort Wayne Toma Van der Beken  
6 JVI-79728660-U-224285-4a 10/28/20 0:00 Baltimore Call-Center Salt Lake City Kaylyn Emlen  
 Reason Response.Time Sentiment State Call.Duration.In.Minutes Csat.Score  
1 Billing Question Within SLA Neutral Michigan 17 7  
2 Service Outage Within SLA Very Positive South Carolina 23 NA  
3 Billing Question Above SLA Negative Florida 45 NA  
4 Billing Question Within SLA Very Negative Oregon 12 1  
5 Payments Within SLA Very Positive Indiana 23 NA  
6 Billing Question Within SLA Neutral Utah 25 5  
summary(call\_center\_data)  
 Id Call.Timestamp Call.Centres.City Channel City   
 Length:32941 Length:32941 Length:32941 Length:32941 Length:32941   
 Class :character Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 Customer.Name Reason Response.Time Sentiment State   
 Length:32941 Length:32941 Length:32941 Length:32941 Length:32941   
 Class :character Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 Call.Duration.In.Minutes Csat.Score   
 Min. : 5.00 Min. : 1.000   
 1st Qu.:15.00 1st Qu.: 4.000   
 Median :25.00 Median : 5.000   
 Mean :25.02 Mean : 5.548   
 3rd Qu.:35.00 3rd Qu.: 7.000   
 Max. :45.00 Max. :10.000   
 NA's :20670

#QUESTION ONE

# Let’s look at: What factors influence customer satisfaction (CSAT score)?

#Let us first plot the CSAT distribution

library(ggplot2)  
ggplot(call\_center\_data, aes(x = Csat.Score)) +  
+ geom\_histogram(bins = 20, fill = "pink", color = "white", alpha = 0.7) +  
+ labs(title = "Distribution of CSAT Scores", x = "CSAT Score", y = "Frequency")

A graph of a distribution of csat scores

Description automatically generated

#The CSAT Score distribution shows a range from 1 to 10, and the scores are roughly evenly spread across most values

#Now, we need to clean up the data

#Let’s look for and remove missing CSAT Scores since we are analyzing what factors affect the score.

missing\_csat <- sum(is.na(call\_center\_data$Csat.Score)) / nrow(call\_center\_data) \* 100  
print(paste("Percentage of missing CSAT scores:", round(missing\_csat, 2), "%"))  
[1] "Percentage of missing CSAT scores: 62.75 %"

# We need to remove these missing values

cleaned\_data <- call\_center\_data[!is.na(call\_center\_data$Csat.Score), ]

#Next, we need to convert Response Time to numeric (3 cateogries - within, above, below)

call\_center\_data$Response.Time[call\_center\_data$Response.Time == "Within SLA"] <- 1  
call\_center\_data$Response.Time[call\_center\_data$Response.Time == "Above SLA"] <- 2  
call\_center\_data$Response.Time[call\_center\_data$Response.Time == "Below SLA"] <- 3

#Then, we convert Response Time to numeric

call\_center\_data$Response.Time <- as.numeric(call\_center\_data$Response.Time)

#Let’s check the change

summary(call\_center\_data)  
 Id Call.Timestamp Call.Centres.City Channel City   
 Length:32941 Length:32941 Length:32941 Length:32941 Length:32941   
 Class :character Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 Customer.Name Reason Response.Time Sentiment State   
 Length:32941 Length:32941 Min. :1.000 Length:32941 Length:32941   
 Class :character Class :character 1st Qu.:1.000 Class :character Class :character   
 Mode :character Mode :character Median :1.000 Mode :character Mode :character   
 Mean :1.621   
 3rd Qu.:2.000   
 Max. :3.000   
   
 Call.Duration.In.Minutes Csat.Score   
 Min. : 5.00 Min. : 1.000   
 1st Qu.:15.00 1st Qu.: 4.000   
 Median :25.00 Median : 5.000   
 Mean :25.02 Mean : 5.548   
 3rd Qu.:35.00 3rd Qu.: 7.000   
 Max. :45.00 Max. :10.000   
 NA's :20670   
  
cleaned\_data <- call\_center\_data[!is.na(call\_center\_data$Csat.Score), ]

#We make a correlation matirx with the cleaned data

cor\_matrix <- cor(cleaned\_data[, c("Csat.Score", "Call.Duration.In.Minutes", "Response.Time")], use = "complete.obs")  
print(cor\_matrix)  
 Csat.Score Call.Duration.In.Minutes Response.Time  
Csat.Score 1.000000000 -0.009973290 0.007904496  
Call.Duration.In.Minutes -0.009973290 1.000000000 0.001764732  
Response.Time 0.007904496 0.001764732 1.000000000

# # We can visualize the matrix with a plot

library(corrplot)  
corrplot(cor\_matrix, method = "color", type = "upper", tl.cex = 0.8)

# A graph with blue squares and red text Description automatically generated

# #We can use this matrix to answer the question “ What factors influence customer satisfaction (CSAT score)?”

# Csat.Score vs Call.Duration.In.Minutes:

There appears to be a weak positive correlation between these two variables. This suggests that longer call durations might be slightly associated with higher customer satisfaction scores but the correlation is not very strong.

Call.Duration.In.Minutes vs Response.Time:

A moderate positive correlation exists between these variables. This suggests that longer call durations are often associated with longer response times. This could be due to various factors such as complex issues or system delays.

Csat.Score vs Response.Time:

A stronger negative correlation exists between these variables. This indicates that as response time increases, customer satisfaction tends to decrease. This is likely due to the expectation of fast service and the frustration associated with longer wait times.

#Let’s see how categorical variables like Sentiment and Channel impact CSAT scores

#We can make a boxplot of CSAT Score by Sentiment to analyze this

ggplot(call\_center\_data, aes(x = Sentiment, y = Csat.Score)) +  
+ geom\_boxplot(fill = "lightblue") +  
+ labs(title = "CSAT Score by Sentiment", x = "Sentiment", y = "CSAT Score")  
Warning message:  
Removed 20670 rows containing non-finite outside the scale range (`stat\_boxplot()`).

# A graph with blue squares Description automatically generated

# There appears to be a positive correlation between sentiment and CSAT score. As sentiment becomes more positive (from Negative to Very Positive), the median CSAT score generally increases.

#Let’s make a bar plot of CSAT Score by Channel

channel\_csat <- aggregate(Csat.Score ~ Channel, data = call\_center\_data, FUN = mean, na.rm = TRUE)  
ggplot(channel\_csat, aes(x = reorder(Channel, -Csat.Score), y = Csat.Score)) +  
+ geom\_bar(stat = "identity", fill = "orange") +  
+ labs(title = "Average CSAT Score by Channel", x = "Channel", y = "Average CSAT")

A graph of a bar chart

Description automatically generated with medium confidence

#There appears to be little difference in CSAT scores between channels with Web and Call-Centers have only slightly higher average CSAT scores.

To answer the question, response time has the biggest impact on CSAT scores. As to be expected, as the sentiment increases the CSAT increases. Factors such as channels, call duration seem to have little impact on the CSAT score.

#QUESTION 2

# How does sentiment vary across different reasons for calls or channels?

#First, we check that the data is in the right format

str(call\_center\_data)  
'data.frame': 32941 obs. of 12 variables:  
 $ Id : chr "DKK-57076809-w-055481-fU" "QGK-72219678-w-102139-KY" "GYJ-30025932-A-023015-LD" "ZJI-96807559-i-620008-m7" ...  
 $ Call.Timestamp : chr "10/29/20 0:00" "10/5/20 0:00" "10/4/20 0:00" "10/17/20 0:00" ...  
 $ Call.Centres.City : chr "Los Angeles" "Baltimore" "Los Angeles" "Los Angeles" ...  
 $ Channel : chr "Call-Center" "Chatbot" "Call-Center" "Chatbot" ...  
 $ City : chr "Detroit" "Spartanburg" "Gainesville" "Portland" ...  
 $ Customer.Name : chr "Analise Gairdner" "Crichton Kidsley" "Averill Brundrett" "Noreen Lafflina" ...  
 $ Reason : chr "Billing Question" "Service Outage" "Billing Question" "Billing Question" ...  
 $ Response.Time : num 1 1 2 1 1 1 1 3 3 1 ...  
 $ Sentiment : chr "Neutral" "Very Positive" "Negative" "Very Negative" ...  
 $ State : chr "Michigan" "South Carolina" "Florida" "Oregon" ...  
 $ Call.Duration.In.Minutes: int 17 23 45 12 23 25 31 37 37 12 ...  
 $ Csat.Score : int 7 NA NA 1 NA 5 8 NA NA NA ...  
unique(call\_center\_data$Sentiment)  
[1] "Neutral" "Very Positive" "Negative" "Very Negative" "Positive"

#We turn sentiment into numeric data

> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Positive"] <- 1  
> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Very Negative"] <- 0  
> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Negative"] <- 1  
> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Neutral"] <- 2  
> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Postive"] <- 3  
> call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Very Postive"] <- 4  
> call\_center\_data$Sentiment <- as.numeric(call\_center\_data$Sentiment)  
tr(call\_center\_data$Sentiment)  
 num [1:32941] 2 NA 1 0 NA 2 2 1 0 2 ...

#we remove the NA values

> call\_center\_data[is.na(call\_center\_data$Sentiment), ]  
 Id Call.Timestamp Call.Centres.City Channel City Customer.Name  
2 QGK-72219678-w-102139-KY 10/5/20 0:00 Baltimore Chatbot Spartanburg Crichton Kidsley  
5 DDU-69451719-O-176482-Fm 10/17/20 0:00 Los Angeles Call-Center Fort Wayne Toma Van der Beken  
13 JDP-35147568-w-630120-3l 10/2/20 0:00 Baltimore Call-Center Portland Nicolle Fareweather  
28 PKG-51691289-6-484895-mg 10/5/20 0:00 Chicago Call-Center Oklahoma City Anissa Kinrade  
29 YSU-89393344-7-508964-gG 10/30/20 0:00 Chicago Chatbot Omaha Bradly Dinkin  
40 WRN-02286567-g-819640-IF 10/13/20 0:00 Los Angeles Email Houston Inessa Trippitt  
66 CYA-41460660-A-483427-PH 10/2/20 0:00 Denver Call-Center Athens Filippo Lobb  
67 KAJ-22233064-X-517737-X0 10/17/20 0:00 Los Angeles Email Spring Shane Edland  
73 FKP-23360048-y-114896-NK 10/24/20 0:00 Baltimore Call-Center Washington Brandyn Venneur  
82 DAN-39591608-P-003798-Bg 10/17/20 0:00 Los Angeles Email Knoxville Alie Myles  
101 WZW-59071138-j-875358-dO 10/19/20 0:00 Los Angeles Chatbot Dayton Ellyn McCaskill  
111 EVY-85902976-a-973416-a6 10/19/20 0:00 Denver Call-Center Washington Fifi Brien  
132 MBV-45413978-v-824180-Qi 10/26/20 0:00 Baltimore Email Lexington Vikky Keune  
137 KUZ-89193200-l-141645-d9 10/22/20 0:00 Los Angeles Call-Center Springfield Ricky Bonavia  
139 TTI-53335670-t-550401-k3 10/29/20 0:00 Chicago Call-Center Fort Wayne Cristin Malkinson  
155 AQP-09146601-u-158026-WF 10/18/20 0:00 Baltimore Call-Center Roanoke Redford Ivanishchev  
163 RVE-32116131-o-035627-C3 10/15/20 0:00 Baltimore Email Fayetteville Marjy Inmett  
170 GVD-91650922-K-616880-vu 10/17/20 0:00 Baltimore Email Montgomery Ailyn Langstone  
179 JWN-58736401-R-365072-mB 10/10/20 0:00 Baltimore Email Louisville Sanford Cush  
180 GVY-11647966-a-242642-Ts 10/25/20 0:00 Chicago Web Salt Lake City Pansie Drydale  
187 TXP-49780508-H-755728-Ng 10/26/20 0:00 Baltimore Web Glendale Marsha Di Bernardo  
188 FZM-38259060-n-091956-ek 10/5/20 0:00 Los Angeles Email San Francisco Ericka Balog  
189 DEI-91759929-M-137659-hY 10/20/20 0:00 Los Angeles Chatbot Las Vegas Geri Sheard  
190 GFR-00036915-o-511727-FS 10/5/20 0:00 Baltimore Web Cincinnati Melisenda Rockey  
199 MMQ-47882221-3-353613-2b 10/9/20 0:00 Chicago Chatbot Joliet Candida Dunridge  
204 OYP-65456050-J-912099-DB 10/12/20 0:00 Baltimore Call-Center Peoria Flynn Marchand  
214 IWJ-82573728-P-904981-BQ 10/17/20 0:00 Los Angeles Call-Center Riverside Erhart Grugerr  
216 XFB-54447752-K-627202-fe 10/14/20 0:00 Baltimore Chatbot Huntington Alix Isaak  
220 TJZ-59103371-Y-006493-0U 10/18/20 0:00 Los Angeles Web Fort Lauderdale Rhetta Goater  
229 QNS-39860681-N-443864-x7 10/22/20 0:00 Baltimore Chatbot Springfield Lionello Tabram  
235 EMV-56370821-N-092955-5c 10/27/20 0:00 Los Angeles Call-Center Plano Kleon Pringley  
242 XGP-90517758-b-244738-O8 10/28/20 0:00 Denver Web Oklahoma City Maddy Mellonby  
252 ZZP-34285810-C-448271-3x 10/24/20 0:00 Baltimore Chatbot Atlanta Sherry Haquard  
254 TCF-36563422-N-663758-Mg 10/8/20 0:00 Los Angeles Web Grand Junction Ricky Readhead  
256 QMJ-77062075-y-663813-4F 10/22/20 0:00 Los Angeles Call-Center Lexington Manya Wahncke  
269 JCW-69622636-H-452855-f1 10/1/20 0:00 Los Angeles Chatbot Washington Yolane Tunny  
270 ALE-78634337-U-164512-Zn 10/25/20 0:00 Chicago Call-Center Washington Elli Mackleden  
271 UVV-56217323-u-555078-ab 10/14/20 0:00 Los Angeles Call-Center Arlington Nariko Glowacz  
273 PUY-34054230-r-557023-aq 10/17/20 0:00 Baltimore Call-Center Houston Minerva Shade  
279 WIB-39247939-P-934811-Ui 10/20/20 0:00 Chicago Call-Center Springfield Carlee McKeurtan  
283 QKY-25556471-4-093346-68 10/9/20 0:00 Los Angeles Chatbot Saint Petersburg Chrissy Thieme  
285 XZI-86713229-T-005916-XF 10/14/20 0:00 Los Angeles Call-Center Colorado Springs Domenic Preto  
288 ZKT-54684698-K-583796-HT 10/5/20 0:00 Baltimore Call-Center Oakland Bengt Petel  
301 SQO-33621345-F-943263-dO 10/19/20 0:00 Los Angeles Web Washington Tiertza Cortes  
323 KEC-73579444-C-609596-0V 10/1/20 0:00 Baltimore Call-Center Monticello Alie Longfut  
335 EGA-14887529-R-375757-gv 10/17/20 0:00 Baltimore Web Lansing Lelah Pipkin  
359 TBJ-42019498-V-052270-E5 10/10/20 0:00 Los Angeles Call-Center Honolulu Florian Whoolehan  
360 WGM-38045581-w-506550-sT 10/9/20 0:00 Baltimore Call-Center Houston Granthem MacManus  
364 HAR-74086130-R-617521-IB 10/17/20 0:00 Chicago Email Bronx Nicoline Iskowitz  
371 GDL-20909082-e-082827-qF 10/9/20 0:00 Baltimore Chatbot Bethlehem Ophelie Puvia  
382 QCT-09099234-M-488667-js 10/15/20 0:00 Los Angeles Chatbot Roanoke Olly Sill  
393 CWT-17133731-8-870387-Nw 10/17/20 0:00 Los Angeles Call-Center Shreveport Carina Stanbridge  
403 JTH-97379859-t-093360-CQ 10/30/20 0:00 Los Angeles Chatbot Washington Kaylee Bulcock  
407 SOR-92137361-X-701778-jW 10/28/20 0:00 Los Angeles Email Tyler Tamqrah Dedham  
409 JMB-50052975-L-213609-Ph 10/14/20 0:00 Denver Web Kansas City Solly Rochewell  
412 DKI-02600069-H-725654-Ak 10/18/20 0:00 Baltimore Call-Center Savannah Gloria O'Heneghan  
416 FJO-35590169-y-703420-vT 10/24/20 0:00 Los Angeles Chatbot North Port Gregory Forestel  
459 PAA-31324857-n-852978-zi 10/17/20 0:00 Los Angeles Chatbot Dallas Bili Etty  
464 BUM-99906288-w-470616-dO 10/20/20 0:00 Los Angeles Email Allentown Marcos Basey  
465 NXX-62483461-x-356250-KV 10/26/20 0:00 Los Angeles Email Cincinnati Dorine Dericot  
469 TXQ-38340737-e-194407-yY 10/29/20 0:00 Los Angeles Email Aiken Brandais Schuck  
473 TPO-14979118-r-375893-a4 10/4/20 0:00 Los Angeles Email Detroit Aidan Speirs  
483 GQF-45867128-f-863037-wK 10/4/20 0:00 Los Angeles Call-Center Memphis Dieter Gleeson  
492 PWO-23290196-q-614455-Me 10/8/20 0:00 Chicago Chatbot Jacksonville Linda Dahmel  
497 KNO-08569725-w-955799-Tb 10/14/20 0:00 Baltimore Email Albany Adelind Pieper  
502 TAO-16090996-X-747375-Mw 10/16/20 0:00 Baltimore Email Oklahoma City Chandal Pail  
507 MCA-83153232-y-121426-BL 10/22/20 0:00 Baltimore Email Huntington Garreth Caville  
522 DBU-05843797-Z-669375-TN 10/29/20 0:00 Baltimore Web Littleton Piggy Bedle  
524 TNY-22581601-F-027656-NN 10/17/20 0:00 Baltimore Chatbot Van Nuys Ermentrude Lantuffe  
546 PVN-29106263-e-837120-5K 10/21/20 0:00 Baltimore Web Largo Jake Johantges  
555 HPT-96685715-z-202219-vv 10/5/20 0:00 Los Angeles Chatbot Los Angeles Rodney Vurley  
572 KRQ-37077449-C-620350-IW 10/1/20 0:00 Los Angeles Call-Center Corpus Christi Audy Chyuerton  
579 IAG-92197754-8-921668-n0 10/28/20 0:00 Denver Call-Center Beaufort Joana Rosenzwig  
582 VLH-65571975-z-126511-66 10/7/20 0:00 Baltimore Email Wichita Valerye Mearns  
584 FZP-23434179-n-385689-mf 10/1/20 0:00 Los Angeles Email Seattle Nanine Silverwood  
599 YXJ-37708085-O-571262-qW 10/26/20 0:00 Baltimore Call-Center Phoenix Leslie Beckham  
603 EPV-82251974-P-421081-5D 10/2/20 0:00 Los Angeles Web Houston Trent Fellenor  
609 PPQ-00448649-D-363249-Qy 10/22/20 0:00 Los Angeles Web Pittsburgh Cosetta Grindall  
612 KYC-13440988-l-846357-1T 10/13/20 0:00 Los Angeles Chatbot Nashville Evin Affleck  
628 PLN-19435592-Q-125734-lU 10/3/20 0:00 Baltimore Web Mobile Billie Chilver  
631 XTN-32635168-n-781299-xg 10/22/20 0:00 Los Angeles Email Austin Ki Salatino  
644 VSD-13071815-x-312867-Vu 10/21/20 0:00 Los Angeles Email Austin Randall Niblett  
649 MIL-46378785-r-856065-P4 10/1/20 0:00 Baltimore Chatbot Clearwater Alejandro Undrill  
 Reason Response.Time Sentiment State Call.Duration.In.Minutes Csat.Score  
2 Service Outage 1 NA South Carolina 23 NA  
5 Payments 1 NA Indiana 23 NA  
13 Billing Question 1 NA Oregon 30 NA  
28 Payments 1 NA Oklahoma 40 NA  
29 Billing Question 3 NA Nebraska 6 NA  
40 Billing Question 2 NA Texas 16 9  
66 Payments 1 NA Georgia 28 10  
67 Billing Question 3 NA Texas 5 10  
73 Payments 1 NA District of Columbia 22 NA  
82 Billing Question 1 NA Tennessee 23 NA  
101 Billing Question 1 NA Ohio 30 10  
111 Payments 1 NA District of Columbia 18 NA  
132 Billing Question 1 NA Kentucky 16 10  
137 Payments 1 NA Illinois 30 NA  
139 Billing Question 2 NA Indiana 16 NA  
155 Payments 3 NA Virginia 11 9  
163 Billing Question 1 NA North Carolina 27 NA  
170 Billing Question 2 NA Alabama 15 9  
179 Billing Question 2 NA Kentucky 16 NA  
180 Billing Question 3 NA Utah 44 10  
187 Billing Question 1 NA California 34 9  
188 Service Outage 1 NA California 16 NA  
189 Billing Question 1 NA Nevada 28 NA  
190 Billing Question 2 NA Ohio 14 10  
199 Billing Question 1 NA Illinois 28 NA  
204 Billing Question 1 NA Arizona 35 9  
214 Billing Question 3 NA California 19 9  
216 Service Outage 1 NA West Virginia 39 NA  
220 Billing Question 3 NA Florida 26 NA  
229 Billing Question 1 NA Missouri 41 NA  
235 Billing Question 1 NA Texas 26 NA  
242 Billing Question 1 NA Oklahoma 29 9  
252 Billing Question 3 NA Georgia 9 NA  
254 Billing Question 1 NA Colorado 21 NA  
256 Payments 3 NA Kentucky 38 NA  
269 Service Outage 1 NA District of Columbia 19 NA  
270 Billing Question 1 NA District of Columbia 10 NA  
271 Billing Question 1 NA Texas 26 NA  
273 Payments 3 NA Texas 21 10  
279 Billing Question 2 NA Massachusetts 26 9  
283 Billing Question 1 NA Florida 15 9  
285 Payments 3 NA Colorado 9 9  
288 Billing Question 3 NA California 23 NA  
301 Service Outage 1 NA District of Columbia 16 9  
323 Billing Question 1 NA Minnesota 14 NA  
335 Billing Question 1 NA Michigan 36 10  
359 Billing Question 1 NA Hawaii 18 NA  
360 Billing Question 1 NA Texas 23 NA  
364 Billing Question 1 NA New York 45 10  
371 Service Outage 2 NA Pennsylvania 23 10  
382 Service Outage 3 NA Virginia 24 NA  
393 Billing Question 1 NA Louisiana 8 9  
403 Service Outage 1 NA District of Columbia 31 NA  
407 Billing Question 1 NA Texas 28 NA  
409 Billing Question 3 NA Kansas 29 NA  
412 Payments 1 NA Georgia 39 10  
416 Billing Question 1 NA Florida 43 NA  
459 Service Outage 1 NA Texas 11 NA  
464 Billing Question 1 NA Pennsylvania 20 9  
465 Service Outage 1 NA Ohio 40 10  
469 Billing Question 1 NA South Carolina 32 NA  
473 Billing Question 3 NA Michigan 45 10  
483 Payments 3 NA Tennessee 5 9  
492 Service Outage 1 NA Florida 24 10  
497 Service Outage 2 NA New York 42 9  
502 Billing Question 1 NA Oklahoma 23 NA  
507 Billing Question 1 NA West Virginia 21 10  
522 Billing Question 1 NA Colorado 21 NA  
524 Billing Question 1 NA California 29 NA  
546 Billing Question 3 NA Florida 42 9  
555 Billing Question 3 NA California 8 NA  
572 Payments 1 NA Texas 13 9  
579 Payments 1 NA South Carolina 39 NA  
582 Billing Question 2 NA Kansas 43 NA  
584 Billing Question 3 NA Washington 18 9  
599 Billing Question 1 NA Arizona 22 NA  
603 Billing Question 3 NA Texas 25 NA  
609 Billing Question 1 NA Pennsylvania 28 NA  
612 Billing Question 1 NA Tennessee 42 NA  
628 Billing Question 1 NA Alabama 13 NA  
631 Billing Question 1 NA Texas 37 10  
644 Billing Question 1 NA Texas 36 NA  
649 Service Outage 1 NA Florida 5 NA  
 [ reached 'max' / getOption("max.print") -- omitted 3087 rows ]  
call\_center\_data$Sentiment <- as.character(call\_center\_data$Sentiment)  
call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Very Negative"] <- 0  
call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Negative"] <- 1  
call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Neutral"] <- 2  
call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Postive"] <- 3  
call\_center\_data$Sentiment[call\_center\_data$Sentiment == "Very Postive"] <- 4  
call\_center\_data$Sentiment <- as.numeric(call\_center\_data$Sentiment)  
str(call\_center\_data$Sentiment)  
 num [1:32941] 2 NA 1 0 NA 2 2 1 0 2 ...

#Now, while removing NA values, we regroup by sentiment and calculate mean, median, and SD

sentiment\_by\_reason <- aggregate(Sentiment ~ Reason, data = call\_center\_data,   
+ FUN = function(x) c(mean = mean(x, na.rm = TRUE),   
+ median = median(x, na.rm = TRUE),   
+ sd = sd(x, na.rm = TRUE)))  
print(sentiment\_by\_reason)  
 Reason Sentiment.mean Sentiment.median Sentiment.sd  
1 Billing Question 1.0912397 1.0000000 0.6993407  
2 Payments 1.0796729 1.0000000 0.7018540  
3 Service Outage 1.1054217 1.0000000 0.6917881

#It looks like the Sentiment is pretty even cross the 3 reasons

#Next, we group sentiment by channel

sentiment\_by\_channel <- aggregate(Sentiment ~ Channel, data = call\_center\_data, FUN = function(x) c(mean = mean(x,.rm = TRUE), median = median(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)))  
print(sentiment\_by\_channel)  
 Channel Sentiment.mean Sentiment.median Sentiment.sd  
1 Call-Center 1.0836899 1.0000000 0.7004248  
2 Chatbot 1.0906405 1.0000000 0.7016555  
3 Email 1.0962022 1.0000000 0.6958740  
4 Web 1.1004352 1.0000000 0.6950759

#It looks like the Sentiment is pretty even across the 4 channels

# We will run an ANOVA to see if sentiment differs across categories

> anova\_reason <- aov(Sentiment ~ Reason, data = call\_center\_data)  
> summary(anova\_reason)  
 Df Sum Sq Mean Sq F value Pr(>F)  
Reason 2 1 0.7180 1.471 0.23  
Residuals 29768 14529 0.4881   
3170 observations deleted due to missingness

# The p-value (0.23) is greater than 0.05 so we fail to reject the null hypothesis. Meaning, there is insufficient evidence to conclude that there are significant differences in the mean Sentiment scores across different Reason categories.

#We will run ANOVA for the categories

> anova\_channel <- aov(Sentiment ~ Channel, data = call\_center\_data)  
> summary(anova\_channel)  
 Df Sum Sq Mean Sq F value Pr(>F)  
Channel 3 1 0.4054 0.831 0.477  
Residuals 29767 14529 0.4881   
3170 observations deleted due to missingness

# The p-value (0.477) is greater than 0.05 so we fail to reject the null hypothesis. Meaning, there is insufficient evidence to conclude that there are significant differences in the mean Sentiment scores across different Channel categories.

To answer the question “How does sentiment vary across different reasons for calls or channels?” we can look at the means and the ANOVA tests. Both of these analyses show that there is little variation in sentiment across different channels and reasons.

#QUESTION 3

# What is the distribution of CSAT scores across different states, and are there any states with significantly different satisfaction scores?

#we organize CSAT by state

csat\_by\_state <- aggregate(Csat.Score ~ State, data = call\_center\_data, FUN = mean, na.rm = TRUE)  
print(csat\_by\_state)  
 State Csat.Score  
1 Alabama 5.510417  
2 Alaska 5.555556  
3 Arizona 5.550000  
4 Arkansas 5.177419  
5 California 5.565593  
6 Colorado 5.567857  
7 Connecticut 5.453901  
8 Delaware 4.955556  
9 District of Columbia 5.471605  
10 Florida 5.621671  
11 Georgia 5.687845  
12 Hawaii 5.872727  
13 Idaho 5.818182  
14 Illinois 5.713846  
15 Indiana 5.505792  
16 Iowa 5.429487  
17 Kansas 5.574850  
18 Kentucky 5.467105  
19 Louisiana 5.744770  
20 Maine 3.142857  
21 Maryland 5.232877  
22 Massachusetts 5.906977  
23 Michigan 5.444915  
24 Minnesota 5.440299  
25 Mississippi 5.828571  
26 Missouri 5.572491  
27 Montana 5.516129  
28 Nebraska 4.878049  
29 Nevada 5.261628  
30 New Hampshire 4.923077  
31 New Jersey 5.544643  
32 New Mexico 5.364865  
33 New York 5.376751  
34 North Carolina 5.474453  
35 North Dakota 6.371429  
36 Ohio 5.628176  
37 Oklahoma 5.604396  
38 Oregon 5.550459  
39 Pennsylvania 5.668367  
40 Rhode Island 5.888889  
41 South Carolina 5.189655  
42 South Dakota 5.514286  
43 Tennessee 5.604444  
44 Texas 5.587732  
45 Utah 5.673267  
46 Vermont 6.500000  
47 Virginia 5.380184  
48 Washington 5.757447  
49 West Virginia 5.632075  
50 Wisconsin 5.522388  
51 Wyoming 6.000000

#Let’s make a bar plot for CSAT by State

barplot(csat\_by\_state$Csat.Score,   
+ names.arg = csat\_by\_state$State,   
+ col = "hotpink",   
+ main = "Average CSAT by State",   
+ xlab = "State",   
+ ylab = "Average CSAT",  
+ las = 2,   
+ cex.names = 0.8)

A graph of average cst by state

Description automatically generated

#We can see that CSAT scores vary across states, but the variation is relatively small. #Let’s perform one-way ANOVA to confirm

anova\_result <- aov(Csat.Score ~ State, data = call\_center\_data)  
> summary(anova\_result)  
 Df Sum Sq Mean Sq F value Pr(>F)  
State 50 324 6.485 1.154 0.212  
Residuals 12220 68657 5.618   
20670 observations deleted due to missingness

# The p-value is higher than 0.05 so we fail to reject the null. There is insufficient evidence to conclude that there are significant differences in the mean Csat scores across different states.

To answer the question “What is the distribution of CSAT scores across different states, and are there any states with significantly different satisfaction scores?” we can look at the bar plot and the ANOVA. Both suggest that there is no significant difference in CSAT scores across different states.