Final Project part 2

The dataset Call_Center.csv contains information for close to 33000 calls to a call center. Information included in this file includes: a timestamp, call-centers city, channel, city, customer name, reason, response time, sentiment, state, call duration in minutes,csat (customer satisfaction) score. In this project will make a few analysis of how different factors affect customer satisfaction. All the visualizations for this project are in Tableau

Load and clean data set

```
#load the library to read excel spreadsheets
library(readxl)

## Warning: package 'readxl' was built under R version 4.4.2

#load the dataframe
df <- read_excel("C:/Users/filte/Downloads/Call_Center.xlsx")

# Check the first few rows of the dataset
head(df)

## # A tibble: 6 × 12
##   Id       `Call Timestamp`    `Call-Centres City` Channel City  `Customer
Name`
##   <chr>    <dttm>              <chr>               <chr>   <chr> <chr>
## 1 DKK-570… 2020-10-29 00:00:00 Los Angeles         Call-C… Detr… Analise G
airdn…
## 2 QGK-722… 2020-10-05 00:00:00 Baltimore           Chatbot Spar… Crichton
Kidsl…
## 3 GYJ-300… 2020-10-04 00:00:00 Los Angeles         Call-C… Gain… Averill B
rundr…
## 4 ZJI-968… 2020-10-17 00:00:00 Los Angeles         Chatbot Port… Noreen La
fflina
## 5 DDU-694… 2020-10-17 00:00:00 Los Angeles         Call-C… Fort… Toma Van
der B…
## 6 JVI-797… 2020-10-28 00:00:00 Baltimore           Call-C… Salt… Kaylyn Em
len
## # ℹ 6 more variables: Reason <chr>, `Response Time` <chr>, Sentiment <chr>
,
## #   State <chr>, `Call Duration In Minutes` <dbl>, `Csat Score` <dbl>

# Check for missing values
summary(df)

##       Id            Call Timestamp                    Call-Centres City
##  Length:32941       Min.   :2020-10-01 00:00:00.00   Length:32941
##  Class :character   1st Qu.:2020-10-08 00:00:00.00   Class :character
##  Mode  :character   Median :2020-10-16 00:00:00.00   Mode  :character
##                     Mean   :2020-10-15 12:51:15.31
```

```
##                     3rd Qu.:2020-10-23 00:00:00.00
##                     Max.   :2020-10-31 00:00:00.00
##
##    Channel              City           Customer Name         Reason
##  Length:32941       Length:32941       Length:32941       Length:32941
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Response Time       Sentiment            State
##  Length:32941       Length:32941       Length:32941
##  Class :character   Class :character   Class :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
```

```
colSums(is.na(df))
```

```
##                       Id           Call Timestamp        Call-Centres City
##                        0                        0                        0
##                  Channel                     City            Customer Name
##                        0                        0                        0
##                   Reason            Response Time                Sentiment
##                        0                        0                        0
##                    State Call Duration In Minutes               Csat Score
##                        0                        0                    20670
```

```
df <- na.omit(df)
#View column names
colnames(df)
```

```
##  [1] "Id"                      "Call Timestamp"
##  [3] "Call-Centres City"       "Channel"
##  [5] "City"                    "Customer Name"
##  [7] "Reason"                  "Response Time"
##  [9] "Sentiment"               "State"
## [11] "Call Duration In Minutes" "Csat Score"
```

Quesiton 1: The relationship between call duration, customer satisfacion and State

#Calculate avearge call duration and Csat score by state

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Calculate summary statistics for Call Duration and Csat Score by State
summary_stats <- df %>%
  group_by(State) %>%
  summarise(
    mean_call_duration = mean(`Call Duration In Minutes`, na.rm = TRUE),
    median_call_duration = median(`Call Duration In Minutes`, na.rm = TRUE),
    sd_call_duration = sd(`Call Duration In Minutes`, na.rm = TRUE),
    mean_csat_score = mean(`Csat Score`, na.rm = TRUE),
    median_csat_score = median(`Csat Score`, na.rm = TRUE),
    sd_csat_score = sd(`Csat Score`, na.rm = TRUE),
    count = n()  # Count of observations in each state
  )

# Print the summary statistics for both Call Duration and Csat Score by State
print(summary_stats)
```

```
## # A tibble: 51 × 8
##    State                mean_call_duration median_call_duration sd_call_du
ration
##    <chr>                             <dbl>                <dbl>
<dbl>
##  1 Alabama                            24.1                   24
11.7
##  2 Alaska                             22.3                   21
11.1
##  3 Arizona                            25.3                   25
11.8
##  4 Arkansas                           24.3                   22
11.6
##  5 California                         25.2                   26
11.9
##  6 Colorado                           24.7                   24
12.0
```

```
##  7 Connecticut                          25.2                    26
12.3
##  8 Delaware                             27.6                    29
11.9
##  9 District of Columbia                 24.7                    24
12.0
## 10 Florida                              25.4                    26
11.8
## # ℹ 41 more rows
## # ℹ 4 more variables: mean_csat_score <dbl>, median_csat_score <dbl>,
## #   sd_csat_score <dbl>, count <int>
```

conclusion States with the longest call duration(like Idaho and Montana) need more time to resolve their issues but have the biggest customer satisfaction(Visualization in Tablua)

Question 2 analyze how sentiment and channal impact cat score ##alculate average CSAT by sentiment

```
# Calculate average CSAT by sentiment
sentiment_csat <- tapply(df$'Csat Score', df$Sentiment, mean, na.rm = TRUE)

# Convert to data frame for easy visualization
sentiment_csat_df <- data.frame(
  Sentiment = names(sentiment_csat),
  Avg_CSAT = sentiment_csat
)

# Print the result
print(sentiment_csat_df)
```

```
##                   Sentiment Avg_CSAT
## Negative           Negative 4.528131
## Neutral             Neutral 6.473039
## Positive           Positive 7.993298
## Very Negative Very Negative 2.457381
## Very Positive Very Positive 9.493484
```

We can conclude that a positve senitment has a highest costamer satisfaction score
#Calculate CSTAT by Channal

```
# Calculate average CSAT by channel
channel_csat <- tapply(df$"Csat Score", df$Channel, mean, na.rm = TRUE)

# Convert to data frame for easy visualization
channel_csat_df <- data.frame(
  Channel = names(channel_csat),
  Avg_CSAT = channel_csat
)
```

```r
# Print the result
print(channel_csat_df)
```

```
##                 Channel Avg_CSAT
## Call-Center Call-Center 5.613310
## Chatbot         Chatbot 5.492470
## Email             Email 5.481720
## Web                 Web 5.591726
```

We can conclude that call-center is the most effective for costumer satisfaction and email is the least effective. (We will visualize both conclusions in Tablua)

#Question 3: Is there corrilation between Response time, Reason, and Costumer Satisfaction

Calculate correlation between Reason and Customer Satisfaction

```r
df$reason_numeric <- as.numeric(factor(df$Reason))
correlation <- cor(df$reason_numeric, df$'Csat Score')
print(correlation)
```

```
## [1] 0.00213682
```

##There is very small positive correlation between reason and costomer satisfaction # Calculate correlation between Red and Customer Satisfaction

```r
df$response_time_numeric <- as.numeric(factor(df$'Response Time'))
correlation2 <- cor(df$response_time_numeric, df$'Csat Score')
print(correlation2)
```

```
## [1] -0.01259633
```

##There is very small negitive correlation between reason and costomer satisfaction