Final-Project-Part-2.R

besti

2024-12-21

```
# Loading the data
file_path <- "C:\\Users\\besti\\OneDrive\\Desktop\\Sarah Schenirer\\Intro Data Science\\Call_Center.csv
call_data <- read.csv(file_path)

# Checking the structure of the data
str(call_data)
```

```
## 'data.frame':    32941 obs. of  12 variables:
##  $ Id                      : chr  "DKK-57076809-w-055481-fU" "QGK-72219678-w-102139-KY" "GYJ-3002593
##  $ 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
##  $ Reason                  : chr  "Billing Question" "Service Outage" "Billing Question" "Billing Que
##  $ 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 ...
```

```
# Viewing the first 6 rows of data
head(call_data)
```

```
##                         Id Call.Timestamp Call.Centres.City     Channel
## 1 DKK-57076809-w-055481-fU  10-29-20 0:00       Los Angeles Call-Center
## 2 QGK-72219678-w-102139-KY   10-5-20 0:00         Baltimore     Chatbot
## 3 GYJ-30025932-A-023015-LD   10-4-20 0:00       Los Angeles Call-Center
## 4 ZJI-96807559-i-620008-m7  10-17-20 0:00       Los Angeles     Chatbot
## 5 DDU-69451719-O-176482-Fm  10-17-20 0:00       Los Angeles Call-Center
## 6 JVI-79728660-U-224285-4a  10-28-20 0:00         Baltimore Call-Center
##             City       Customer.Name           Reason Response.Time
## 1        Detroit    Analise Gairdner Billing Question    Within SLA
## 2    Spartanburg    Crichton Kidsley   Service Outage    Within SLA
## 3    Gainesville   Averill Brundrett Billing Question     Above SLA
## 4       Portland     Noreen Lafflina Billing Question    Within SLA
## 5     Fort Wayne Toma Van der Beken         Payments    Within SLA
## 6 Salt Lake City       Kaylyn Emlen Billing Question    Within SLA
##       Sentiment          State Call.Duration.In.Minutes Csat.Score
## 1       Neutral       Michigan                       17          7
## 2 Very Positive South Carolina                       23         NA
## 3      Negative        Florida                       45          NA
## 4 Very Negative         Oregon                       12           1
## 5 Very Positive        Indiana                       23          NA
## 6       Neutral           Utah                       25           5
```

```
# Find factors that impact Sentiment

# Does Call duration have an impact on sentiment?

# Factoring and changing Sentiment to numeric:
# Very Negative - -2
# Negative - -1
# Neutral - 0
# Positive - 1
# Very Positive - 2

call_data$Sentiment <- as.numeric(factor(call_data$Sentiment,
                                         levels = c("Very Negative", "Negative", "Neutral", "Positive",

# Simple Linear Regression
# Independent variable - call duration
# Dependent variable - sentiment

model <- lm(Sentiment ~ Call.Duration.In.Minutes, data = call_data)
summary(model)
```

```
##
## Call:
## lm(formula = Sentiment ~ Call.Duration.In.Minutes, data = call_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6272 -0.6203 -0.5945  0.4029  2.4072
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.3684665  0.0153734 -23.968   <2e-16 ***
## Call.Duration.In.Minutes -0.0008606  0.0005556  -1.549    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.191 on 32939 degrees of freedom
## Multiple R-squared:  7.284e-05,  Adjusted R-squared:  4.249e-05
## F-statistic:   2.4 on 1 and 32939 DF,  p-value: 0.1214
```

```
# The R=squared shows that this is not a good model
# Call duration does not have a significant impact on sentiment


# Does response time impact sentiment?

# Factoring and changing Response Time to numeric:
# Below SLA - -1
```

```r
# Within SLA - 0
# Above SLA - 1

call_data$Response.Time <- as.numeric(factor(call_data$Response.Time,
                                             levels = c("Below SLA", "Within SLA", "Above SLA"))) -2

# Simple Linear Regression
# Independent variable - response time
# Dependent variable - sentiment

model2 <- lm(Sentiment ~ Response.Time, data = call_data)
summary(model2)
```

```
##
## Call:
## lm(formula = Sentiment ~ Response.Time, data = call_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6119 -0.6119 -0.6075  0.3903  2.3925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.390268   0.006697 -58.275   <2e-16 ***
## Response.Time -0.002214   0.010953  -0.202     0.84
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.192 on 32939 degrees of freedom
## Multiple R-squared:  1.24e-06,   Adjusted R-squared:  -2.912e-05
## F-statistic: 0.04086 on 1 and 32939 DF,  p-value: 0.8398
```

```r
# The R-squared shows that this is not a good model
# Response time does not have a significant impact on sentiment

# Does channel impact sentiment?

# Simple Linear Regression
# Independent variable - channel
# Dependent variable - sentiment

model3 <- lm(Sentiment ~ Channel, data = call_data)
summary(model3)
```

```
##
## Call:
## lm(formula = Sentiment ~ Channel, data = call_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6204 -0.6139 -0.6005  0.3995  2.3995
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -0.392800   0.011552 -34.003   <2e-16 ***
## ChannelChatbot  0.006657   0.017476   0.381    0.703    
## ChannelEmail   -0.006664   0.017986  -0.371    0.711    
## ChannelWeb      0.013238   0.018691   0.708    0.479    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.192 on 32937 degrees of freedom
## Multiple R-squared:  3.404e-05,  Adjusted R-squared:  -5.704e-05 
## F-statistic: 0.3737 on 3 and 32937 DF,  p-value: 0.772
```

```
# The R-squared shows that this is not a good model
# Channel does not have a significant impact on sentiment

# Does it change when the independent variables are put together in multiple linear regression?

# Multiple Linear Regression
# Independent variables - call duration, response time, channel
# Dependent variable - sentiment

mmodel <- lm(Sentiment ~ Call.Duration.In.Minutes + Response.Time + Channel, data = call_data)
summary(mmodel)
```

```
## 
## Call:
## lm(formula = Sentiment ~ Call.Duration.In.Minutes + Response.Time + 
##     Channel, data = call_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.6396 -0.6210 -0.5898  0.4051  2.4191 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              -0.3715848  0.0181398 -20.485   <2e-16 ***
## Call.Duration.In.Minutes -0.0008587  0.0005556  -1.545    0.122    
## Response.Time            -0.0022693  0.0109531  -0.207    0.836    
## ChannelChatbot            0.0065726  0.0174767   0.376    0.707    
## ChannelEmail             -0.0065976  0.0179867  -0.367    0.714    
## ChannelWeb                0.0132451  0.0186913   0.709    0.479    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.192 on 32935 degrees of freedom
## Multiple R-squared:  0.0001078,  Adjusted R-squared:  -4.396e-05 
## F-statistic: 0.7104 on 5 and 32935 DF,  p-value: 0.6155
```

```
# Again the R-squared shows that this is not a good model
# None of the independent variables have a significant impact on sentiment

# Conclusion
# None of the variables tested have a significant impact on sentiment
```

```r
# It is still unknown what variables do impact sentiment

# Compare call duration, sentiment and csat score by Channel

# Factoring Channel
call_data$Channel <- factor(call_data$Channel)

# Summary statistics for call duration, sentiment and csat score by channel

summary_call_duration <- aggregate(Call.Duration.In.Minutes ~ Channel, data = call_data, summary)
summary_call_duration
```

```
##       Channel Call.Duration.In.Minutes.Min. Call.Duration.In.Minutes.1st Qu.
## 1 Call-Center                       5.00000                          15.00000
## 2     Chatbot                       5.00000                          15.00000
## 3       Email                       5.00000                          15.00000
## 4         Web                       5.00000                          15.00000
##   Call.Duration.In.Minutes.Median Call.Duration.In.Minutes.Mean
## 1                        25.00000                      25.04615
## 2                        25.00000                      24.91776
## 3                        25.00000                      25.09880
## 4                        25.00000                      25.02235
##   Call.Duration.In.Minutes.3rd Qu. Call.Duration.In.Minutes.Max.
## 1                         35.00000                      45.00000
## 2                         35.00000                      45.00000
## 3                         35.00000                      45.00000
## 4                         35.00000                      45.00000
```

```r
summary_sentiment <- aggregate(Sentiment ~ Channel, data = call_data, summary)
summary_sentiment
```

```
##       Channel Sentiment.Min. Sentiment.1st Qu. Sentiment.Median Sentiment.Mean
## 1 Call-Center     -2.0000000        -1.0000000       -1.0000000     -0.3928001
## 2     Chatbot     -2.0000000        -1.0000000       -1.0000000     -0.3861434
## 3       Email     -2.0000000        -1.0000000       -1.0000000     -0.3994645
## 4         Web     -2.0000000        -1.0000000       -1.0000000     -0.3795620
##   Sentiment.3rd Qu. Sentiment.Max.
## 1         0.0000000      2.0000000
## 2         0.0000000      2.0000000
## 3         0.0000000      2.0000000
## 4         0.0000000      2.0000000
```

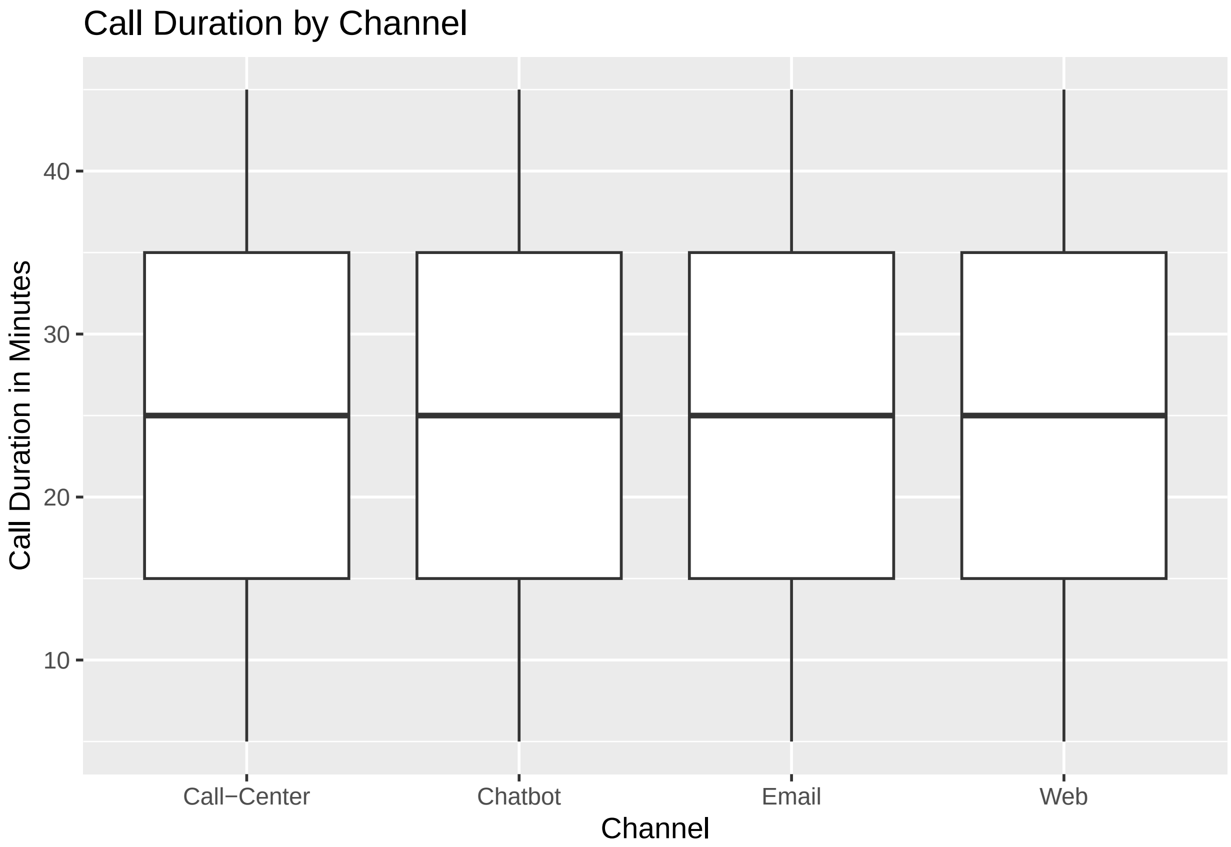
```r
summary_csat_score <- aggregate(Csat.Score ~ Channel, data = call_data, summary)
summary_csat_score
```

```
##       Channel Csat.Score.Min. Csat.Score.1st Qu. Csat.Score.Median
## 1 Call-Center        1.000000           4.000000          6.000000
## 2     Chatbot        1.000000           4.000000          5.000000
## 3       Email        1.000000           4.000000          5.000000
## 4         Web        1.000000           4.000000          6.000000
##   Csat.Score.Mean Csat.Score.3rd Qu. Csat.Score.Max.
## 1        5.613310           8.000000       10.000000
## 2        5.492470           7.000000       10.000000
## 3        5.481720           7.000000       10.000000
## 4        5.591726           7.000000       10.000000
```

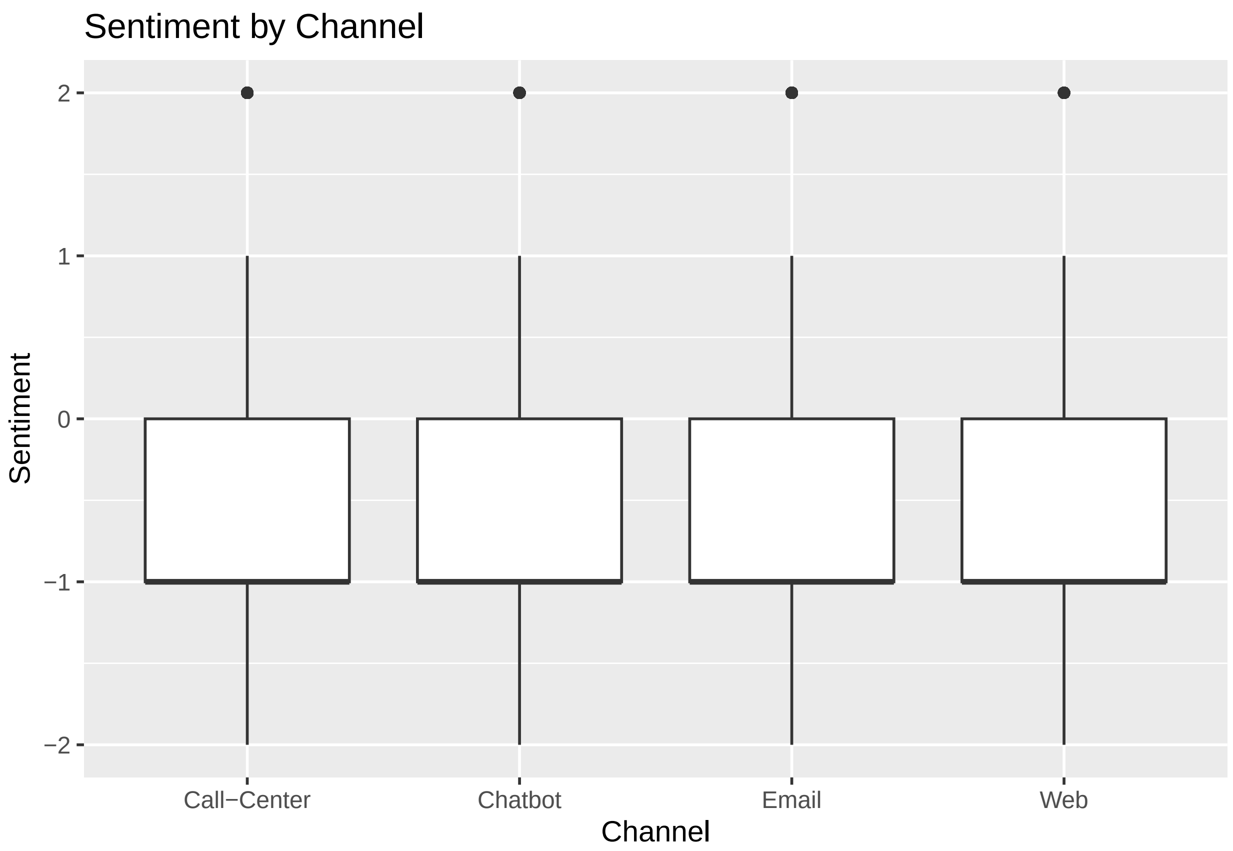
```
# Boxplots

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```
# Call duration by Channel
ggplot(call_data, aes(x = Channel, y = Call.Duration.In.Minutes)) + geom_boxplot() + labs(title = "Call
```

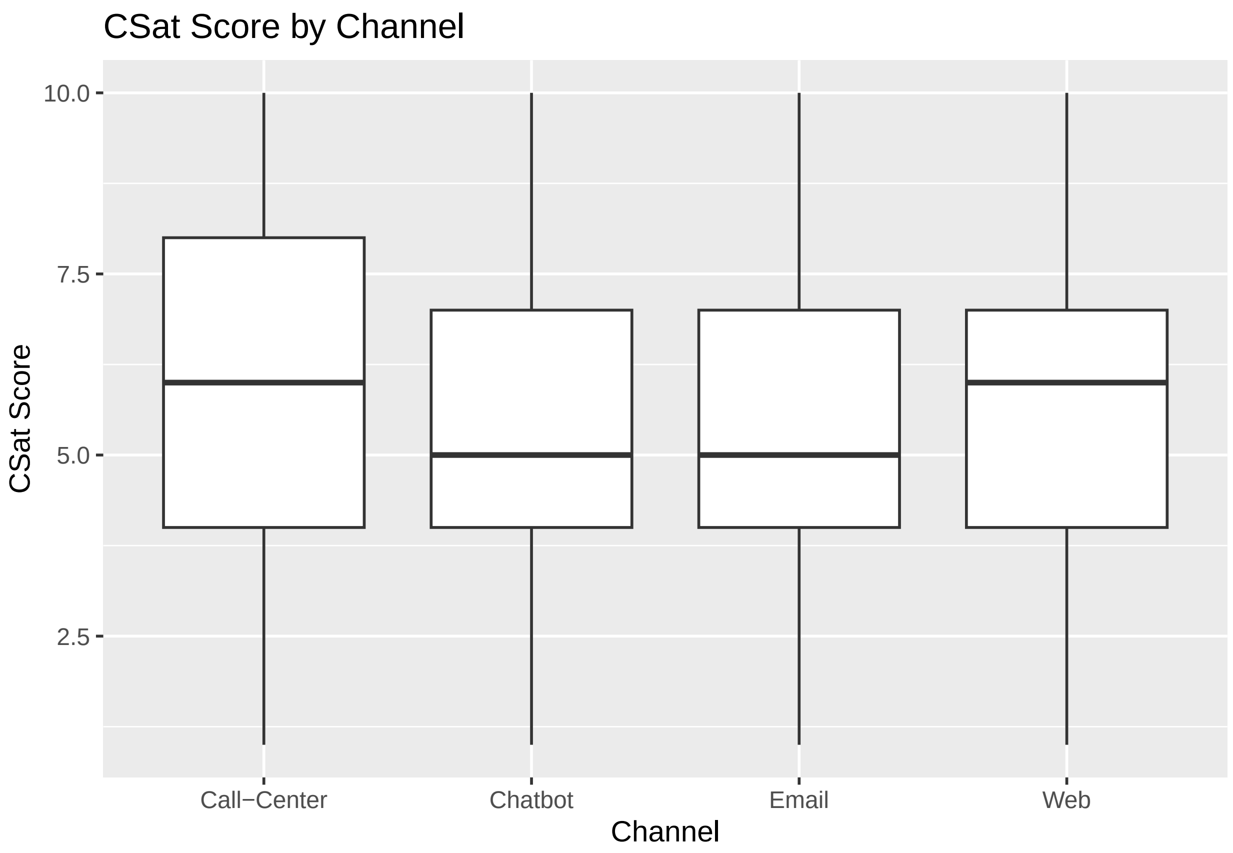


```
# Call duration is the same across the different channels

# Sentiment by Channel
ggplot(call_data, aes(x = Channel, y = Sentiment)) + geom_boxplot() + labs(title = "Sentiment by Channel
```



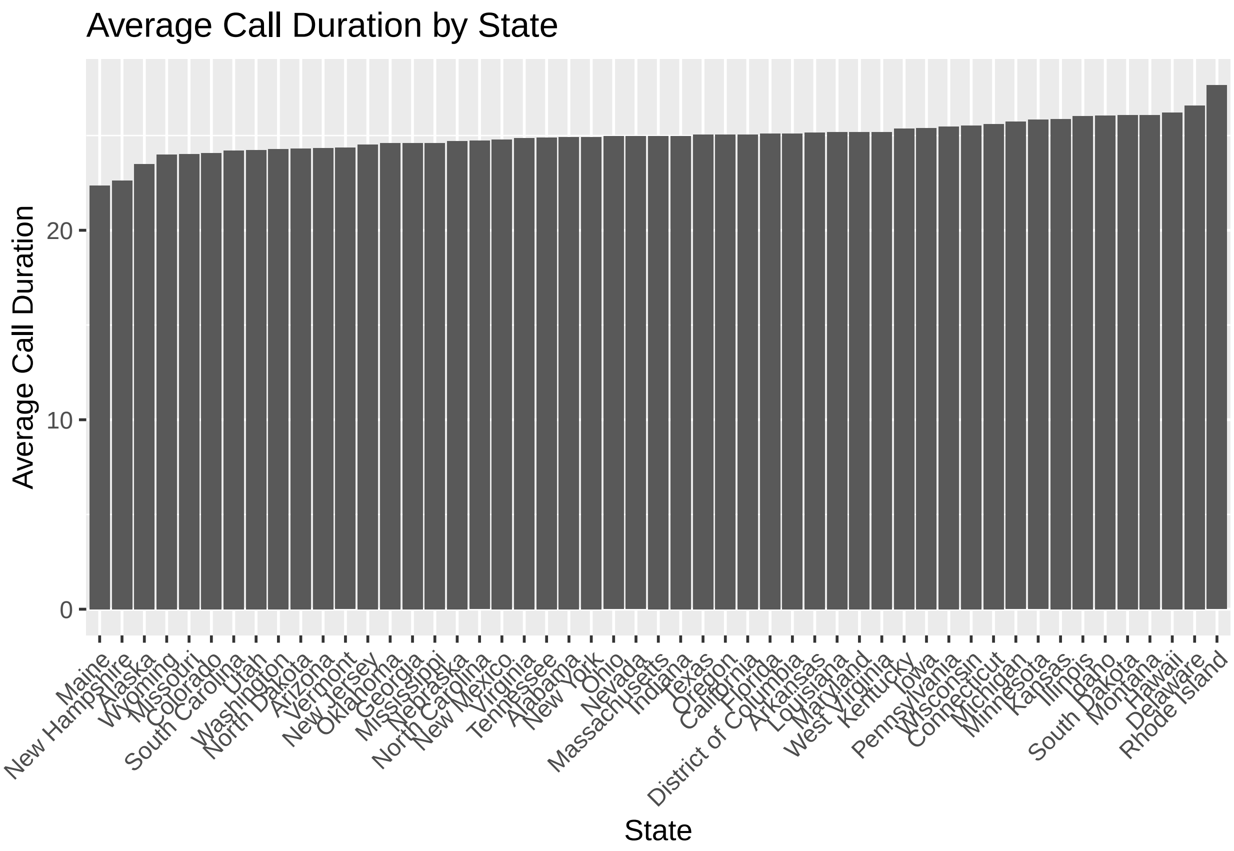
```r
# Sentiment is the same across the different channels

# Csat score by Channel
ggplot(call_data, aes(x = Channel, y = Csat.Score)) + geom_boxplot() + labs(title = "CSat Score by Chann
```

```
## Warning: Removed 20670 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

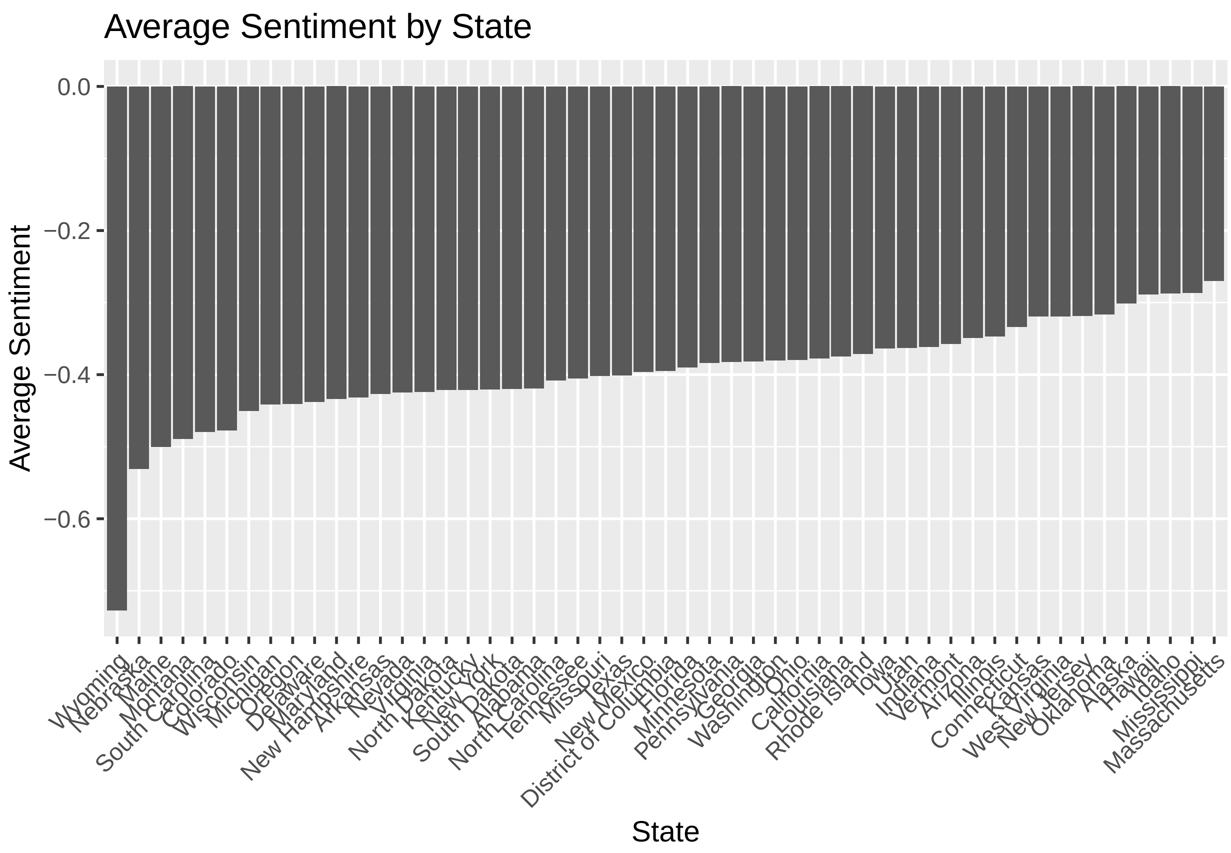


```r
# Call-Center and Web have the same median but call-center has a wider spread which means that there is
# Csat score for chatbot and email are the same
# the Csat score column has a lot of null values which were removed in the boxplot

# Conclusion
# Call duration and sentiment are the same across the different channels
# Csat scores are different for different channels but many rows of data have null values for Csat

# Compare call duration, sentiment and csat score by state

# Factoring State
call_data$State <- factor(call_data$State)

# Mean call duration by state
call_duration_by_state <- aggregate(Call.Duration.In.Minutes ~ State, data = call_data, mean)

# Mean sentiment by state
sentiment_by_state <- aggregate(Sentiment ~ State, data = call_data, mean)

# Mean csat score by state
csat_by_state <- aggregate(Csat.Score ~ State, data = call_data, mean, na.rm=TRUE)
# removed null values

# Bar plots

# Call duration by state
```

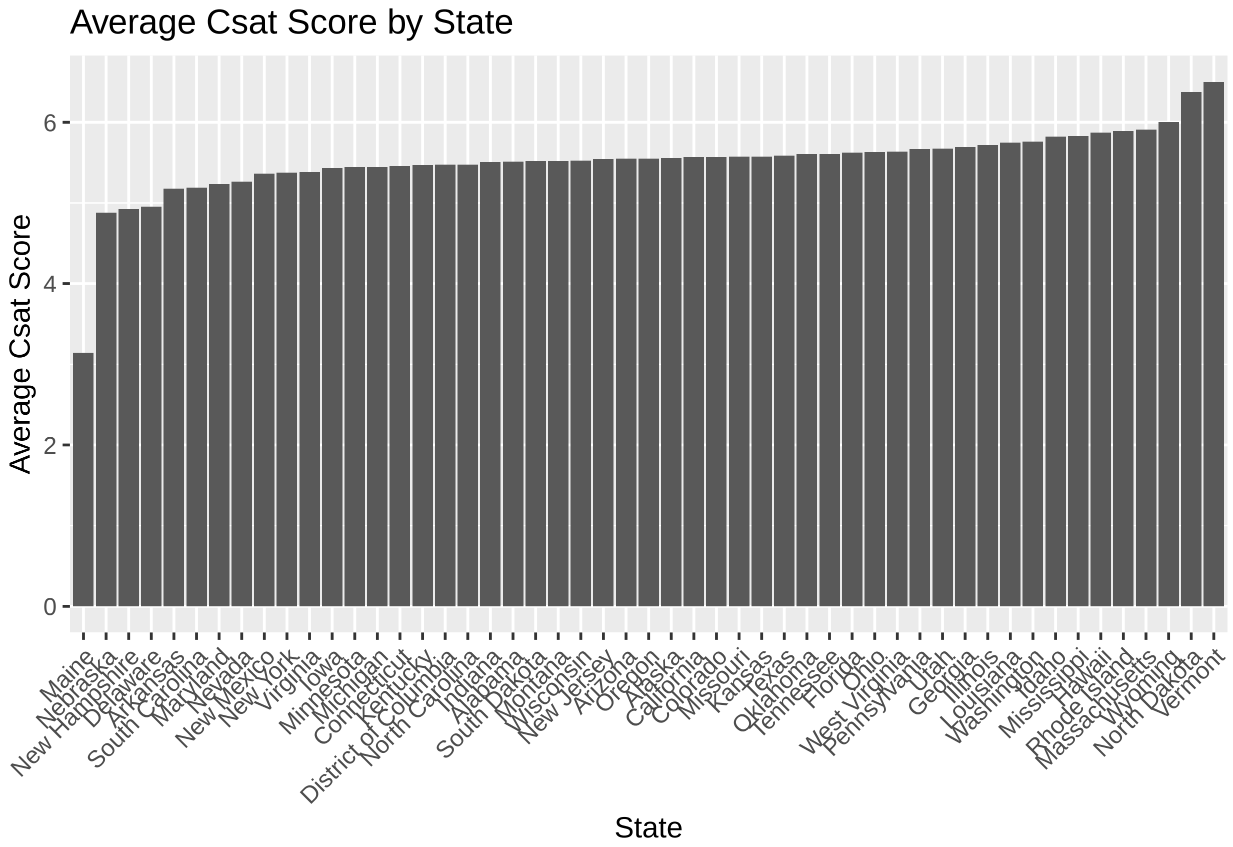
```r
ggplot(call_duration_by_state, aes(x = reorder(State, Call.Duration.In.Minutes), y = Call.Duration.In.Mi
```



```r
# Sentiment by State
ggplot(sentiment_by_state, aes(x = reorder(State, Sentiment), y = Sentiment)) + geom_bar(stat = "identi
```



```r
# Csat by state
ggplot(csat_by_state, aes(x = reorder(State, Csat.Score), y = Csat.Score)) + geom_bar(stat = "identity"
```



```
# Conclusion
# Average sentiment changes by state with Wyoming having the lowest average sentiment by far.
# Average call duration is pretty much the same across all the states with slight variation
# Average csat scores are also pretty much the same across the states with the exception of Maine with
```