Oil Prices and Mental Health

August Sosick

18/11/2020

Introduction

Economics and the mental health of individuals and communities is closely interwoven. Due to COVID-19, more people are spending their days online and socially isolated. Health-care works, therapists, and those whose work involve treating individuals who are struggling mentally would find information about what individuals are searching for online and any potential correlations with other events useful as it allows them to be better prepared for increased demand for their services.

This project endeavors to determine a link between what Albertians are searching on any given month and the oil price for that period (and by extension the economy). The oil and gas sector accounts for 16% of Alberta's GDP so downturns can effect large amounts of people in varying areas.

Online search data was taken from Google Trends using a library which interfaces with their API, and webscraping was used to determine oil prices for the period between 2016 and 2020.

Determining Search Data

To begin, a dataframe with negative terms was created. The frequency or "hits" of each term within a week will be returned

```r
#Number of data points returned from 5 years.
periodCount <- 261
negativeDF <- data.frame("Date" = 1:periodCount, "Depression" = "", "Guilt" = "",
                         "Hatred" = "", "Suicide" = "","Death" = "", "Debt" = "",
                         "Angry" = "", "Shitty" = "", "Horrible" = "",
                         "Alcohol" = "", "Drugs" = "", "Sadness" = "",
                         "Divorce" = "", "Therapy" = "")

i <- 2 #skip Date column
while (i <= length(negativeDF)) {
  results <- gtrends(keyword = colnames(negativeDF)[i], geo = "CA-AB", time = "today+5-y")

  #Add the data column only once
  if(i == 2){
    negativeDF$Date <- results %>% .$interest_over_time %>% .$date
  }
  negativeDF[i] <- results %>% .$interest_over_time %>% .$hits
  i <- i + 1
}
```

We utilize the gtrendsR library to query Google trends with information we require.

- keyword: The work we wish to query. We iterate over the names of the columns from the negativeDF table as these are our search terms
- geo: CA-AB, the region of Alberta Canada
- time: today+5-y indicates we want information from today to 5 years in the past. It should be noted that longer periods of time reduce the frequency of data points you receive, from days to weeks to months.

We store all of this information into the results negativeDF dataframe. Here is a preview of the extracted data from Google trends with Date and Keyword.

```
##         Date Depression Guilt Hatred Suicide Death Debt Angry Shitty Horrible
## 1 2015-11-22         69    71      0      13    42   28    26     36       72
## 2 2015-11-29         55    54     27      12    40   32    30     46       44
## 3 2015-12-06         79    44     54      19    41   33    35     55       78
## 4 2015-12-13         70    78      0      17    39   39    40     79       52
## 5 2015-12-20         53   100      0      16    40   24    35     52       86
## 6 2015-12-27         64    26      0      17    42   33    81     76       75
##   Alcohol Drugs Sadness Divorce Therapy
## 1      49    71      33      46      61
## 2      40    62     100      55      56
## 3      45    63      28      55      69
## 4      45    77      66      55      60
## 5      60   100      53      30      53
## 6      52    70      26      69      51
```

Since the data received on search terms is from one day out of every 7 days, we will average it into a month. This is done using the aggregate function:

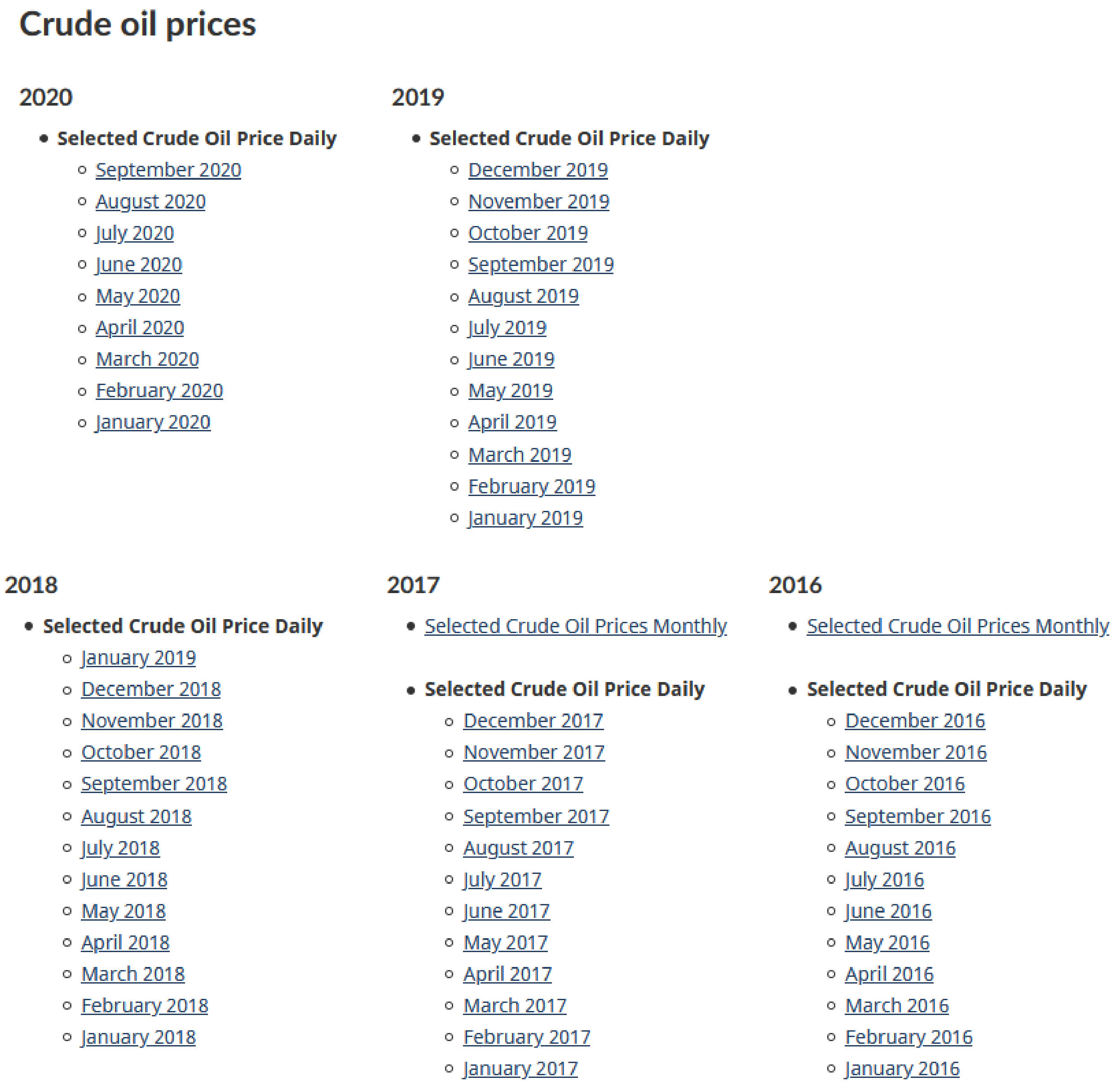
```r
#Insure that the data column type is recognized as a Date
negativeDF$Date <- as.Date(negativeDF$Date, format = "%Y-%m-%d")
#Create a new Month and Year column based off of the numeric dates in from ex: 2020-01-01
negativeDF$Month <- months(negativeDF$Date)
negativeDF$Year <- format(negativeDF$Date, format = "%y")
#Aggregate the information by finding the average hit rate for each term across each month
#and place that information into emotionsDF
emotionsDF <- aggregate(. ~ Month + Year, negativeDF, mean)
emotionsDF <- subset(emotionsDF, select = -c(Date))
```

Preview of monthly aggregated hit rate by term:

```
##      Month Year Depression Guilt Hatred Suicide Death  Debt Angry Shitty
## 1 December   15       66.5 62.00  13.50   17.25 40.50 32.25 47.75  65.50
## 2 November   15       62.0 62.50  13.50   12.50 41.00 30.00 28.00  41.00
## 3    April   16       78.0 52.75  36.75   20.00 50.75 49.00 31.50  51.50
## 4   August   16       55.5 42.75   6.75   59.00 36.00 50.50 41.75  56.25
## 5 December   16       59.0 45.25  39.00   24.25 43.00 44.50 32.00  69.00
## 6 February   16       68.0 55.75  67.75   14.50 39.25 61.00 32.25  45.75
##   Horrible Alcohol Drugs Sadness Divorce Therapy
## 1    72.75   50.50 77.50   43.25   52.25   58.25
## 2    58.00   44.50 66.50   66.50   50.50   58.50
## 3    52.50   44.75 40.00   33.50   48.75   68.00
## 4    35.50   42.25 41.75   24.00   43.00   66.00
## 5    63.50   49.50 68.25   42.25   50.50   54.75
## 6    76.25   44.00 41.00   35.75   57.25   66.25
```

Determing Oil Prices

After having used an API, we will now scrape the web for information about the price of oil. StatsCan provides tables of information grouped by year.



Crude oil prices

2020

- **Selected Crude Oil Price Daily**
  - September 2020
  - August 2020
  - July 2020
  - June 2020
  - May 2020
  - April 2020
  - March 2020
  - February 2020
  - January 2020

2019

- **Selected Crude Oil Price Daily**
  - December 2019
  - November 2019
  - October 2019
  - September 2019
  - August 2019
  - July 2019
  - June 2019
  - May 2019
  - April 2019
  - March 2019
  - February 2019
  - January 2019

2018

- **Selected Crude Oil Price Daily**
  - January 2019
  - December 2018
  - November 2018
  - October 2018
  - September 2018
  - August 2018
  - July 2018
  - June 2018
  - May 2018
  - April 2018
  - March 2018
  - February 2018
  - January 2018

2017

- Selected Crude Oil Prices Monthly

- **Selected Crude Oil Price Daily**
  - December 2017
  - November 2017
  - October 2017
  - September 2017
  - August 2017
  - July 2017
  - June 2017
  - May 2017
  - April 2017
  - March 2017
  - February 2017
  - January 2017

2016

- Selected Crude Oil Prices Monthly

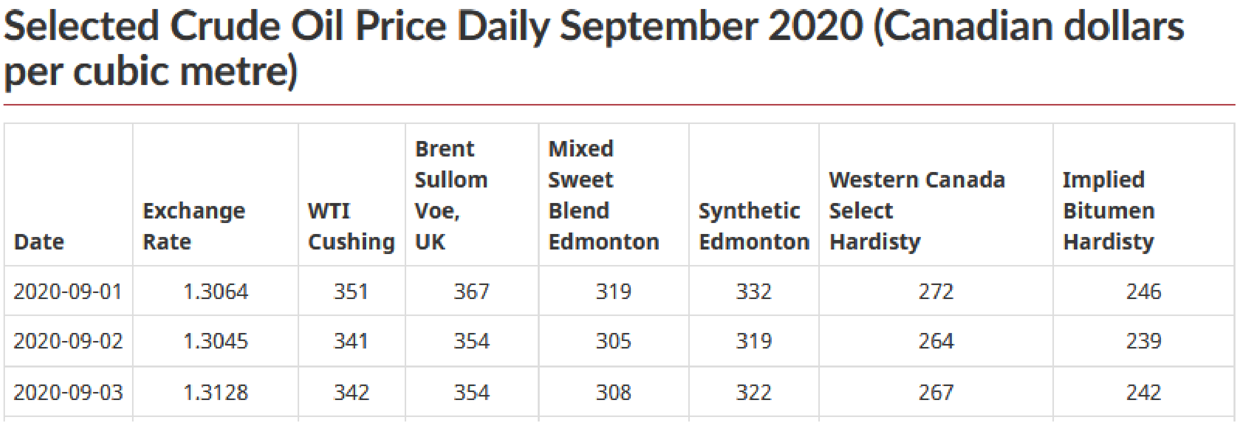
- **Selected Crude Oil Price Daily**
  - December 2016
  - November 2016
  - October 2016
  - September 2016
  - August 2016
  - July 2016
  - June 2016
  - May 2016
  - April 2016
  - March 2016
  - February 2016
  - January 2016

This is this page that we need to scrape. We'll collect all the links which redirect you to a table containing the oil prices for that month.

```
#Request the html for the webpage
#https://www.nrcan.gc.ca/our-natural-resources/energy-sources-
#distribution/clean-fossil-fuels/crude-oil/oil-pricing/18087"
urlAdd <-
"https://www.nrcan.gc.ca/our-natural-resources/energy-sources-distribution/clean-fossil-fuels/crude-oil/
web_content <- read_html(urlAdd)
```

```r
#Retrieve all the links by searching for "a" node with "href" attribtues
all_links <- web_content %>% html_nodes("a") %>% html_attr('href')
#select only links which contain "selected-crude-oil-price-daily"
#as those links contain links to the montly data.
all_links <- all_links[which(regexpr("selected-crude-oil-price-daily", all_links)>0)]
```

After having done that, we will enter each link, collect the html information for each page, and utilize a library function to extract information from the table. The tables in each link and from which we'll extract our data look like this:



Selected Crude Oil Price Daily September 2020 (Canadian dollars per cubic metre)

Date	Exchange Rate	WTI Cushing	Brent Sullom Voe, UK	Mixed Sweet Blend Edmonton	Synthetic Edmonton	Western Canada Select Hardisty	Implied Bitumen Hardisty
2020-09-01	1.3064	351	367	319	332	272	246
2020-09-02	1.3045	341	354	305	319	264	239
2020-09-03	1.3128	342	354	308	322	267	242

```r
#Create dataframe containing date and average oil prices.
oilPriceDF <- data.frame("Date"= "", "Average Oil Prices" = "")
index <- 1
#loop through all of the links we determined in the code block above
while(index < length(all_links)){
  #read the HTML page contents
  this_page_content <- read_html(all_links[index])
  #Extract the tables from the page
  tables <- html_nodes(this_page_content, "table")
  #Get the dataframe of the first table, the one pictured above,
  #and place into the oilDt variable.
  oilDt <- html_table(tables, fill = TRUE)[[1]]

  #Rename Column Names due to spaces
  colnames(oilDt) <- c("Date", "Exchange_Rate", "WTI_Crushing",
                       "Brent_Sullom_Vow_UK", "Mixed_Sweet_Blend_Edmonton",
                       "Synthetic_Edmonton", "Western_Canada_Select_Hardistry",
                       "Implied_Bitumen_Hardisty")

  #Begin to loop through each row in the table
  for(i in 1:nrow(oilDt)){
    #Do not include rows with hyphens or the final column called Average
    if(oilDt$Exchange_Rate[i] != "-" & oilDt$Date[i] != "Average"){
      #Include the current date, and the mean of the costs for each
      #of the listed crude oils.
      #Try Catch to mismatched columns in older data sets.
      tryCatch({
        oilPriceDF[nrow(oilPriceDF)+1,] <- c(oilDt$Date[i],
                  mean(c(as.numeric(oilDt$WTI_Crushing[i]),
                  as.numeric(oilDt$Brent_Sullom_Vow_UK[i]),
                  as.numeric(oilDt$Mixed_Sweet_Blend_Edmonton[i]),
                  as.numeric(oilDt$Synthetic_Edmonton[i]),
                  as.numeric(oilDt$Western_Canada_Select_Hardistry[i]),
                  as.numeric(oilDt$Implied_Bitumen_Hardisty[i])),
                  na.rm = TRUE))
      },
      warning=function(err){
        #Do nothing. Warnings are cause by additional columns in older data sets.
        #They should just be ignored and produced only NA's in table which are
        #dealt with later.
      })
    }
  }
  index <- index + 1
}
```

We continue on and type set the variables in the dataframe we extracted from each webpage again, while aggregating based off of year and month.

```
#remove inital row which is empty.
oilPriceDF <- oilPriceDF[-c(1),]
#typeset average.oil.prices variables as numerics
oilPriceDF$Average.Oil.Prices <- as.numeric(oilPriceDF$Average.Oil.Prices)
#typeset the dates as dates and indicate proper format
oilPriceDF$Date <- as.Date(oilPriceDF$Date, format = "%Y-%m-%d")
#Create month and year table
oilPriceDF$Month <- months(oilPriceDF$Date)
oilPriceDF$Year <- format(oilPriceDF$Date, format = "%y")
#Aggregate the mean of Average Oil Prices by month and year into
#the dataframe "aggregatedOilPrices"
aggregatedOilPrice <- aggregate(Average.Oil.Prices ~ Month + Year, oilPriceDF, mean)
```

We now will combine the information about Oil we determined through scraping StatsCan, with the information we determined about word searches from Google Trends into one dataframe to make our regression and plotting easier in the coming steps.

```
#create an OilPrice column in the dataframe containing our averaged monthly hits by search
#term
emotionsDF$OilPrice <- NA
#Search for the same month and year across the two dataframes and
#combine where a match is determined.
for(i in 1:nrow(aggregatedOilPrice)){
  for(j in 1:nrow(emotionsDF)){
    if(aggregatedOilPrice$Month[i] == emotionsDF$Month[j] &
        aggregatedOilPrice$Year[i] == emotionsDF$Year[j]){
      emotionsDF$OilPrice[j] <- aggregatedOilPrice$Average.Oil.Prices[i]
    }
  }
}
#remove Na's which are where no month year matches were found
emotionsDF <- na.omit(emotionsDF)
```

Summary of Table:

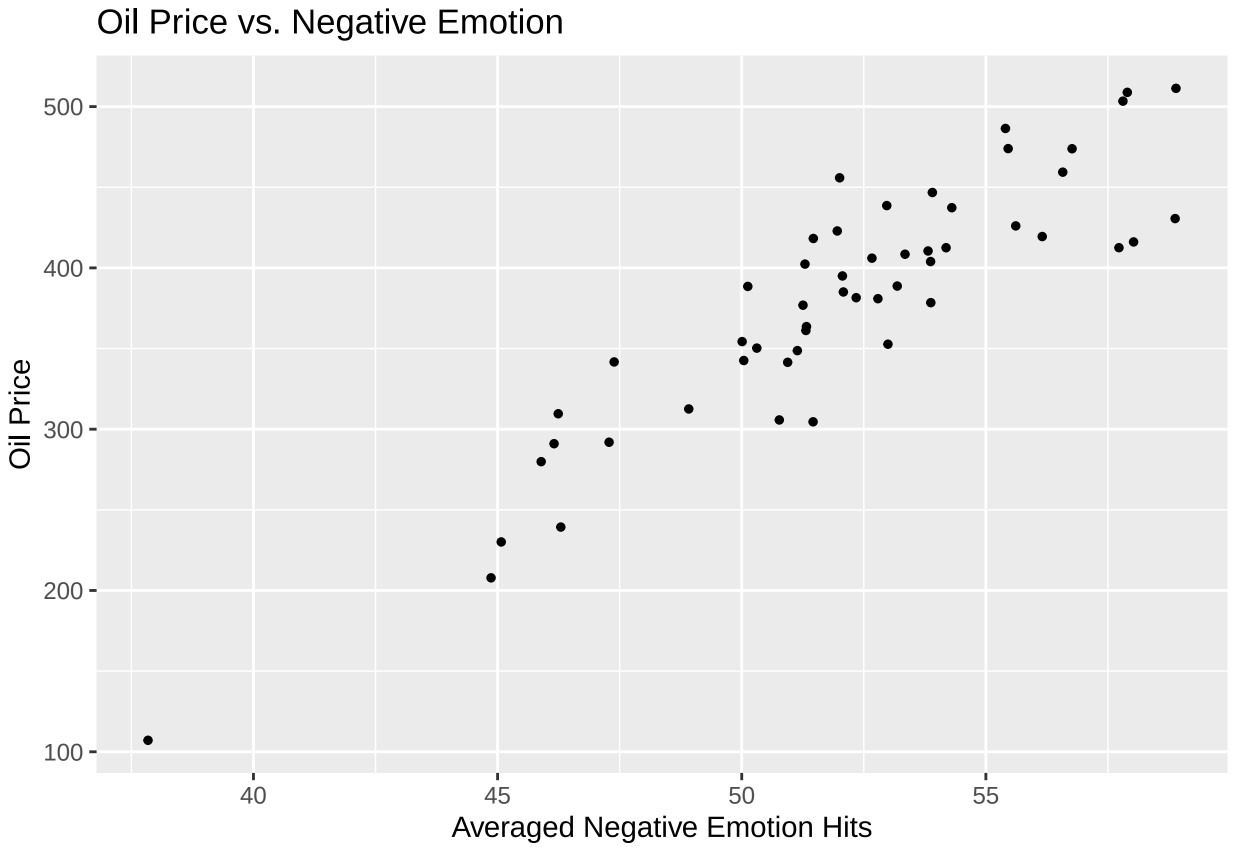
```
##       Month Year Depression Guilt Hatred Suicide Death  Debt Angry Shitty
## 3     April   16      78.00 52.75  36.75   20.00 50.75 49.00 31.50  51.50
## 5  December   16      59.00 45.25  39.00   24.25 43.00 44.50 32.00  69.00
## 8      July   16      53.80 60.60  30.00   40.80 36.40 47.00 34.80  51.60
## 9      June   16      50.25 52.00  31.25   16.00 41.75 42.25 43.25  38.75
## 10    March   16      63.25 49.00  59.25   15.50 41.00 53.25 34.00  49.25
## 11      May   16      60.00 44.60  15.00   13.20 39.60 42.80 68.20  55.60
##    Horrible Alcohol Drugs Sadness Divorce Therapy OilPrice
## 3     52.50   44.75 40.00   33.50   48.75   68.00 305.7222
## 5     63.50   49.50 68.25   42.25   50.50   54.75 304.5795
## 8     43.80   45.00 41.40   34.20   46.80   59.20 341.4583
## 9     56.50   44.75 41.25   39.50   43.75   54.25 376.8788
## 10    58.75   46.00 39.50   35.25   45.75   64.50 239.3106
## 11    58.20   40.80 38.00   30.80   44.80   60.80 363.6190
```

Regression and Data Analysis

Dot Plot

We know have one dataframe containing monthly hit rates on emotionally negative terms and issues and the average price of crude oil for that period. In addition to the other columns, we created an Average Negative Emotion column which averages the hit rate per month for our terms. We'll compare that to the price of crude below:

```
g <- ggplot(emotionsDF, aes(AvgNegEmotion, OilPrice))
g + geom_jitter(width = .5, size=1) +
      labs(y="Oil Price",
           x="Averaged Negative Emotion Hits",
           title="Oil Price vs. Negative Emotion")
```



As you can see from above, both the hit rate for negative emotion and the oil price do go together in general. However, upon examination we see that in fact it is **when oil price are high we see higher negative emotion searches!** While this does re-enforce the claim that oil prices have bearing on people's emotions, a person's first idea might be that when oil is low and therefore the economy is worse off you would see more negative searches in Alberta however that appears to not be the case.

Linear Regression

It's obvious from the above graph that there is some variance from the imaginary trend line. We'll take a closer look at the numbers that describe this relationship with a linear regression model.

```
regressionModel <- lm(OilPrice ~ AvgNegEmotion, data = emotionsDF)
summary(regressionModel)
```

```
##
## Call:
## lm(formula = OilPrice ~ AvgNegEmotion, data = emotionsDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.355 -18.187   4.189  19.968  69.350
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -539.550     59.157  -9.121 3.24e-12 ***
## AvgNegEmotion   17.656      1.134  15.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.03 on 50 degrees of freedom
## Multiple R-squared:  0.829,  Adjusted R-squared:  0.8256
## F-statistic: 242.4 on 1 and 50 DF,  p-value: < 2.2e-16
```

The H0 hypothesis originally was that there is not correlation and therefore we can not estimate oil prices based off of the frequency of negative terms searched online by Alberta residents. There is reason enough to reject that hypothesis in favour of H1 which says there is relationship between negative online searched and oil prices. Reason: * P value is less than 2.2e-16. That indicates it's highly unlikely the relationship we see here is due to chance alone

Additionally.

- The average across the search terms chosen to represent negative emotion can explain up to 83.86% of the movement of the price of crude based off of the adjusted R-squared.
- The F-statistic is 266

Disclaimer: Depending on the day, the data may not match the summary of the regression or plot above. I've provided a globalEnviroment.RData file in the ZIP to replicate my results. There is a provided script file that matches the code within this RMD file in order to verify my results.

Conclusion

Alberta is a province highly dependent on the oil and gas industry- when that sector hurts, so do the people and economy of this province. Additionally, it is no secret that during COVID-19, the up-tick in mental health issues for everyone has risen sharply which means that examining and understanding where individuals go in the digital space when confronting these issues is more critical and pertinent than ever. This mini-project's intent was to demonstrate that there is a connection between O&G and mental well-being within Alberta which it succeeded in doing. Additionally, further examination determine that higher oil prices comes along with more negative online searching by Albertians which may have not been the initial intuition an individual may have about this particular idea.