Capstone Project in R - Relationship Between the COVID-19 Pandemic and Stock Prices

Alan da Silveira Fleck

25/07/2020

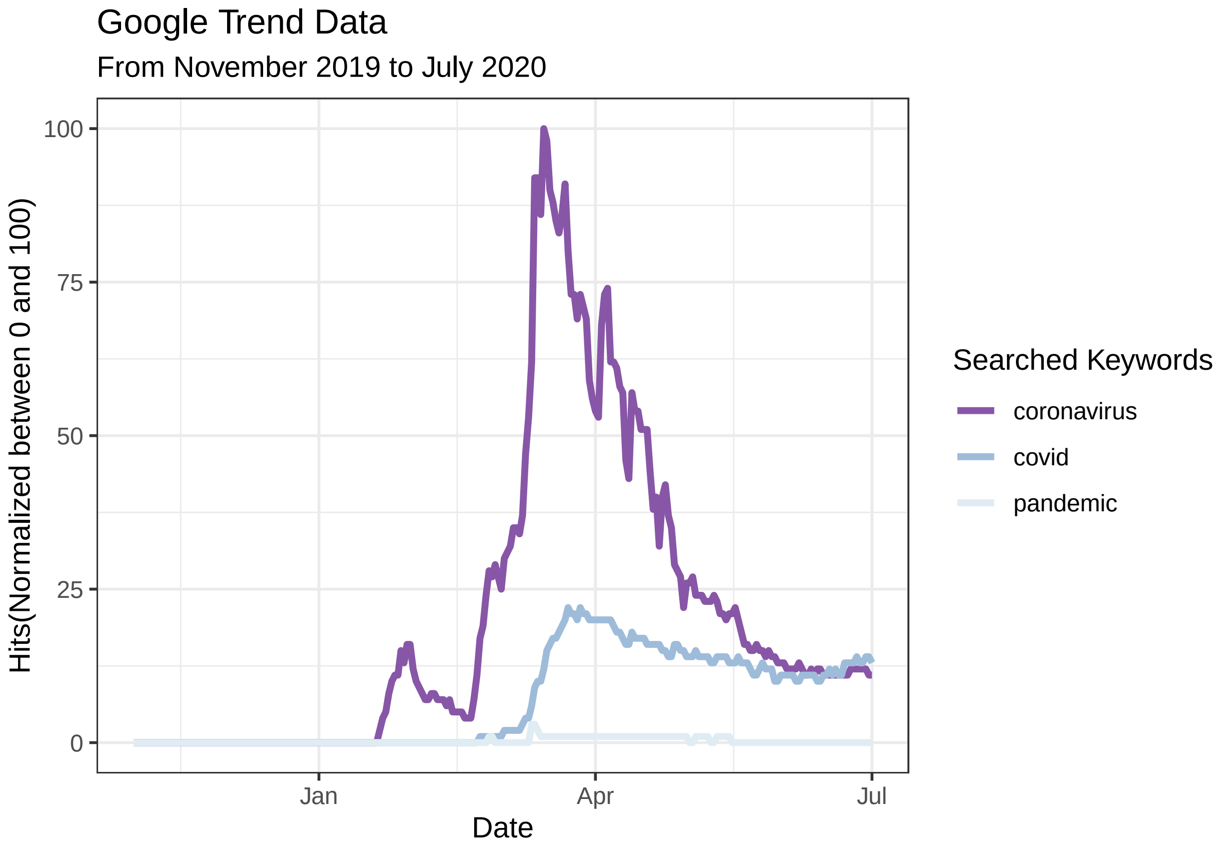
INTRODUCTION

The first case of COVID-19 was tracked back to November 17, 2019. This new disease spread rapidly around the world causing a global health crisis and huge economic disruptions. In this document, we investigated the effects that the COVID-19 pandemic had on the stock market. To answer this questions, we looked at two sources of data. Essentially, we wanted to see how the increased awareness of the spread of the disease impacted the levels of the S&P 500 Index.

ANALYSIS

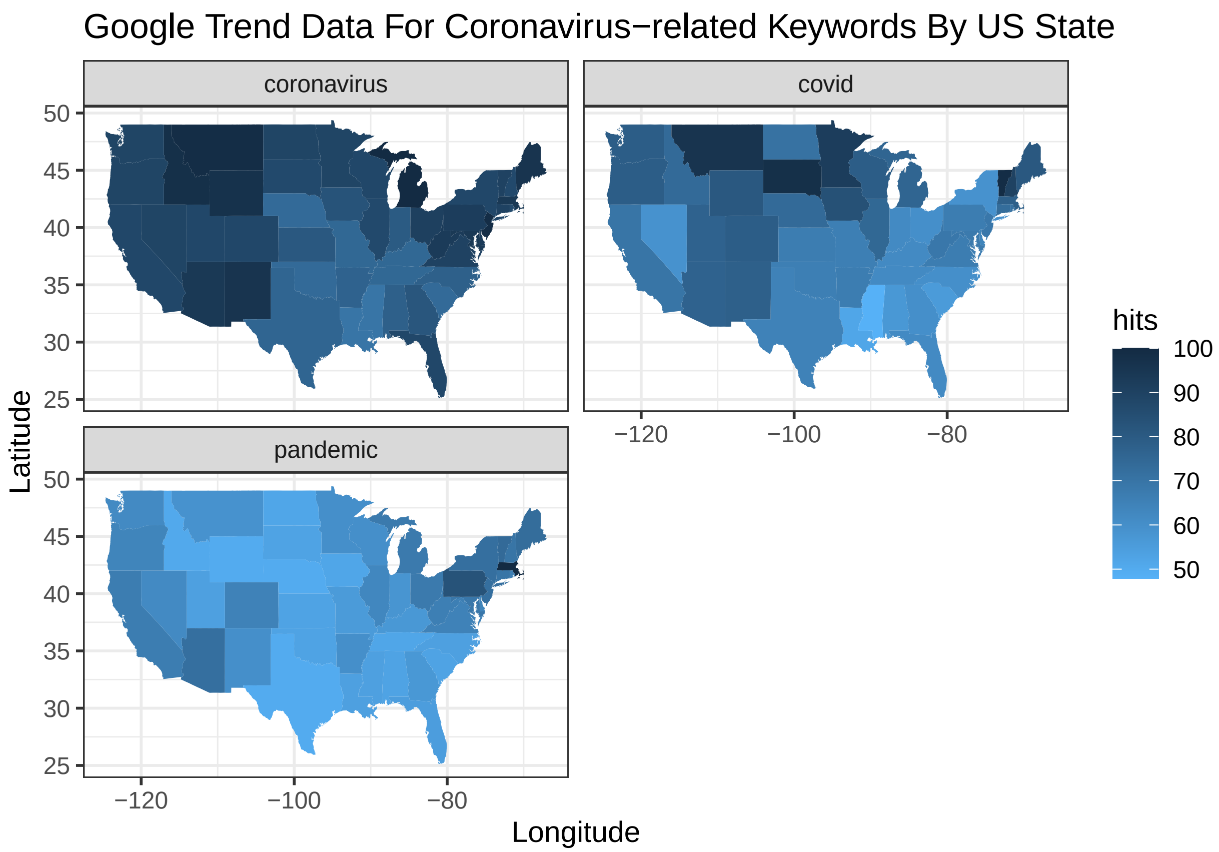
The first source of data we used was Google Trend data using the gtrendsR library. This library returns the popularity of Google Search terms. Using this data, we visualized the increased interest on the disease over time. The keywords “coronavirus”, “covid” and “pandemic” were researched and the search was extended back to November 1, 2019 to visualize the background levels of these terms. As we can see, the interest peaked between February and March 2020, and “coronavirus” was the most prevalent keyword searched among the three.

```r
library(gtrendsR)
library(dplyr)
library(tidyverse)

# Creating the dataframe. The dataframe has values "<1" for some hits which we've
# changed to "0"
world_trends <- gtrends(keyword=c("coronavirus", "covid", "pandemic"),
                time = "2019-11-01 2020-07-01")
trends_over_time <- world_trends$interest_over_time
trends_over_time <- trends_over_time %>%
  mutate(hits = as.numeric(hits)) %>% replace_na(list(hits = 0))

# Plotting the trend
trends_plot <- ggplot(trends_over_time) +
 aes(x = date, y = hits, colour = keyword) +
 geom_line(size = 1.22) +
 scale_color_brewer(palette = "BuPu", direction = -1) +
 labs(x = "Date", y = "Hits(Normalized between 0 and 100)", title = "Google Trend Data",
      subtitle = "From November 2019 to July 2020", color = "Searched Keywords")+
 theme_bw()
trends_plot
```



By adding geo = “US” we were able to get trend data at the state level by looking at the interest_by_region data frame. We also used the maps library to help create maps of the United States. Both “coronavirus” and “covid-19” keywords were less searched in the central and southeast states compared to the other regions of the US.

```r
library(maps)
region_trends <-  gtrends(keyword=c("coronavirus", "covid", "pandemic"), geo = "US",
                          time = "2019-11-01 2020-07-01")
states <- region_trends$interest_by_region
states <- states %>% mutate(location = tolower(location))

# Getting the map data
states_map <- map_data("state")

# Plotting the map
state_plot <- states %>% ggplot(aes(map_id = location)) +
  geom_map(aes(fill = hits), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat) +
  facet_wrap(~ keyword, nrow = 2) +
  labs(title = "Google Trend Data For Coronavirus-related Keywords By US State",
       x = "Longitude", y = "Latitude") +
  scale_fill_continuous(name = "Hits (Scaled to 100)")+
  scale_fill_gradient(low="#56B1F7", high="#132B43")+
  theme_bw()
state_plot
```



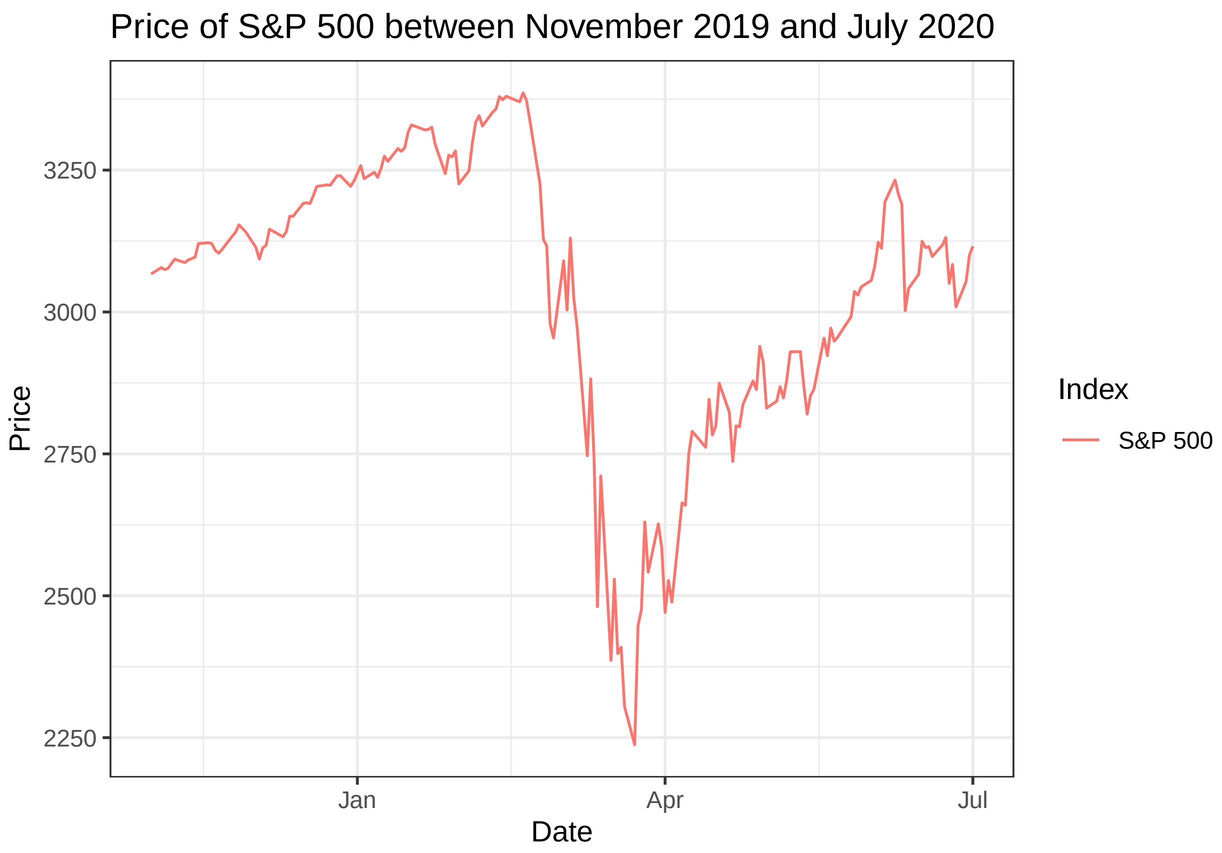
STOCK MARKET PRICE

In this next step, we retrieved data from the S&P 500 Index as a broader indicator of the US stock market performance. As we can see in the figure below, this index fell more than 35% between the months of February and March 2020.

```
library(quantmod)
# Getting the stock data
getSymbols("^GSPC")
```

```
## [1] "^GSPC"
```

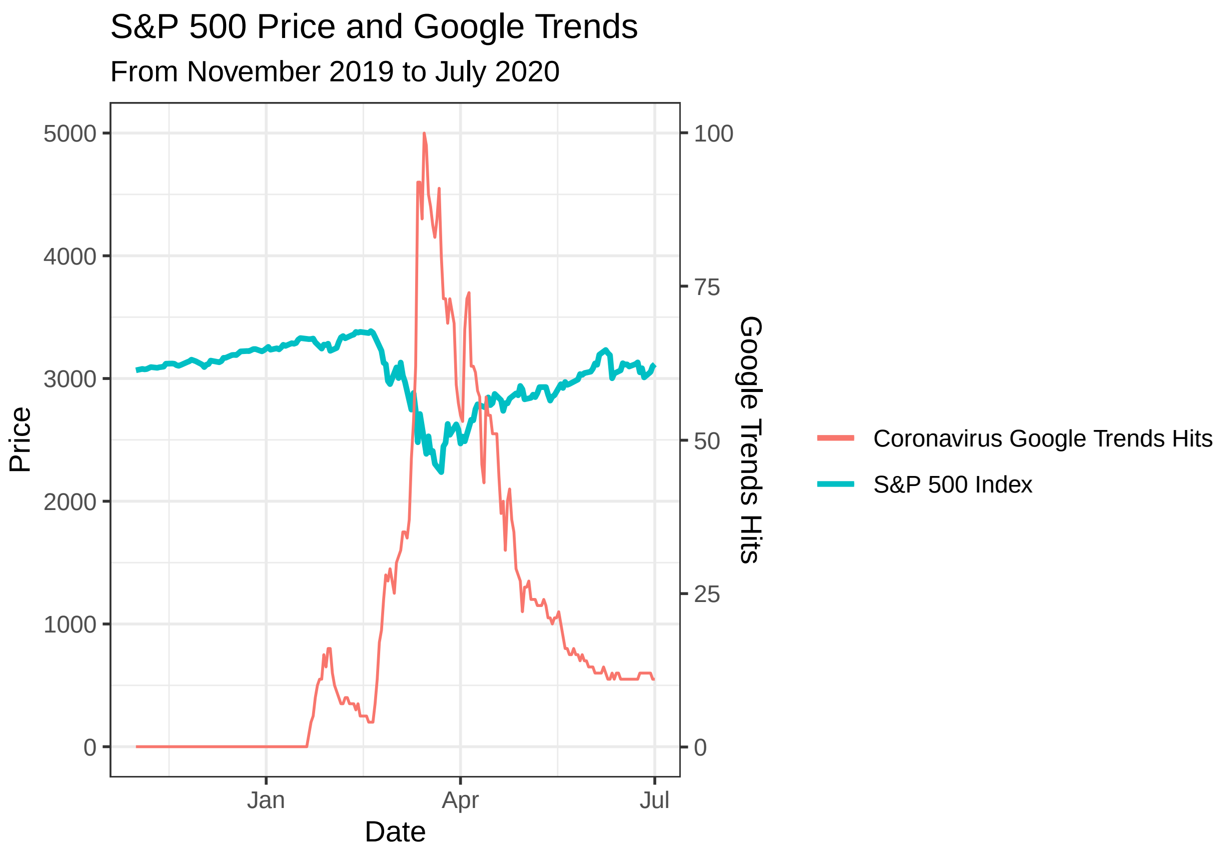
```
# Merging the stock data into one data frame
stocks <- data.frame("S&P 500"= GSPC$GSPC.Close, "Date" = as.Date(row.names(as.data.frame(GSPC))))

# Reshaping the data frame so one column contains the type of stock and  filtering the date.
recent_stocks <- stocks %>%
  filter(Date >= "2019-11-01" & Date <= "2020-07-01") %>%
  gather(key = "stock", value = "value", -Date)

# Graphing the stock prices
recent_stocks %>% ggplot() + geom_line(aes(x = Date, y= value, color = stock)) +
  scale_color_discrete(name = "Index", labels = "S&P 500") +
  labs(title = "Price of S&P 500 between November 2019 and July 2020", y = "Price") + theme_bw()
```



COMBINING THE DATA

As we can see in the combined analysis, the search of the word “coronavirus” on Google peaked at the same time that the price of the S&P 500 index reached its lowest level. This suggests that an increased level of fear caused by the awareness about the disease resulted in a market sell-off. Interestingly, the reduction of the number of searches on Google coincided with the recovery on the market’s value.

```
# Plotting stock and trends on the same graph for S&P500 and coronavirus.
# Note that we adjust the scale of the Trends axis

ggplot(recent_stocks) +
 aes(x = Date, y = value, colour = "S&P 500 Index") +
 geom_line(size = 1L) +
 scale_color_hue() +
 labs(color = "") +
 theme_bw()+
geom_line(filter(trends_over_time, keyword == "coronavirus"),
 mapping = aes(as.Date(date), hits*50, color = "Coronavirus Google Trends Hits")) +
scale_y_continuous(name = "Price", sec.axis = sec_axis(~./50, name="Google Trends Hits")) +
labs(title = "S&P 500 Price and Google Trends", subtitle = "From November 2019 to July 2020")
```



LINEAR REGRESSION

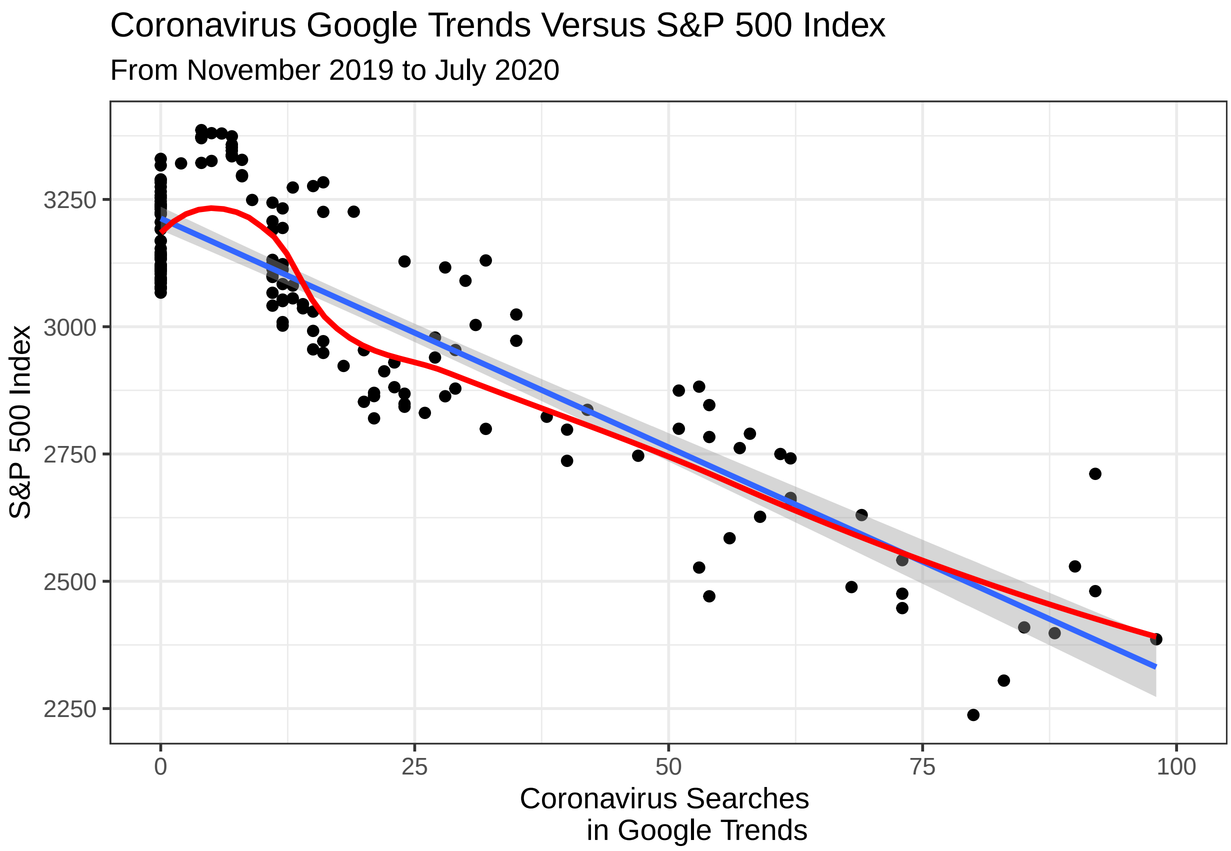
The next analysis shows a linear regression associating the S&P 500 price (dependent variable) and the search of coronavirus on Google Trends (independent variable). There is a negative and significant relationship between the two variables. Each point increase on Google Trends hits resulted in a decrease of around 9 points on the S&P 500 index. Our independent variable explains 78.7% of the variation on the S&P 500 price between November 2019 and July 2020. Interestingly, our model best fits the data after around 25 hits on Google Trends, suggesting that other factors were influencing the stock market price when the awareness about the disease (and fear) was lower.

```
# Joining the tables
trends_over_time2 <- trends_over_time %>%
  rename(Date = date) %>%
  filter(keyword == "coronavirus") %>%
  mutate(Date=as.Date(Date, format = "%Y.%m.%d"))

trends_and_stocks <- recent_stocks %>%
  full_join(trends_over_time2)

#Linear regression model
library(jtools)
model <- lm(value~hits, data = trends_and_stocks, na.rm = TRUE)
summ(model, confint = TRUE, digits = 3)
```

```
## MODEL INFO:
```

```
## Observations: 167 (77 missing obs. deleted)
## Dependent Variable: value
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,165) = 608.463, p = 0.000
## R² = 0.787
## Adj. R² = 0.785
##
## Standard errors: OLS
## ------------------------------------------------------------------
##                          Est.       2.5%      97.5%     t val.       p
## ----------------- ---------- ---------- ---------- ---------- -------
## (Intercept)          3212.738   3189.753   3235.723   275.979   0.000
## hits                   -8.993     -9.713     -8.273   -24.667   0.000
## ------------------------------------------------------------------
```

```
plot <- ggplot(trends_and_stocks, aes(x = hits, y = value)) +
  geom_point(na.rm=TRUE) +
  geom_smooth(method = "lm") +
  geom_smooth(se = FALSE, color = "red", na.rm = TRUE) +
  labs(title = "Coronavirus Google Trends Versus S&P 500 Index",
       subtitle = "From November 2019 to July 2020", x = "Coronavirus Searches
       in Google Trends", y = "S&P 500 Index")+
  theme_bw()
plot
```



CONCLUSION AND FUTURE WORKS

In this project we were able to compare the impacts of the coronavirus pandemic on the price of the stock market. Google Trends were used as a general surrogate of the levels of awareness (and fear) about the pandemic, while the S&P 500 index was used as an indicator of the stock market performance. Although market prices are extremely difficult to predict and generally driven by complex fundamental and technical indicators, we showed that fear and public awareness of the coronavirus pandemic was an important predictor of the 2020 market crash. If we were to continue this work there are a few different routes we could take. First, we could retrieve data from individual stocks that may have been affected or benefited from the global health crisis. In addition, the complexity of the model could be increased by adding other co-variables. Specific time frames such as pre-pandemic, crisis and recovery phases could also be investigated in the future.