STAT 440 Data Project Summary Report - Fall 2019

**Project Title**: Correlation Between Weather Around Central Park, New York and S&P 500 Stock Market Index.

**Group Members**: Joshua Immanuel, Louis Widhi Nugroho, Wilson Phurwo

**Introduction Section:**

1. Statement of Interest: Stock market is one of the indicators that can tell the overall performance of one company or country. There are many factors that can contribute to the rise and decline of the stock market. These days, weather arguable factor that can affect the behavior of the stock market. Thus, by conducting this project we hope to explore the relationship between weather and the stock market. We want to know whether bad weather can put downwards pressure on the stock prices in the market or whether a good weather can create upwards pressure on the stock prices in the market. For this project, we define bad weather as gloomy weather that has very little sunlight with colder temperatures, and good weather as weather with a lot of sunshine and warmer temperatures. We will use central park weather data that provided from National Oceanic and Atmospheric Administration (NOAA) since the location close to the wall street which is the location of New York Stock Exchange. As for the stock data, we believe that the S&P 500 stock index is a good example of the stock market and will hope to see certain price movements.
2. Information background of the data sources

* S&P 500 (GSPC) data

This dataset is taken from the Yahoo Finance website and comes from the stock index of the S&P 500 (GSPC). We used the stock ticker GSPC to indicate that the index is a price index instead of total returns index. This means that the stock index does not include dividends, which is perfect for analyzing its price volatility with the weather data. The data consists of daily stock prices from the date of 01/01/2007 to 11/14/2019. The data has more than 3000 data points and includes the 2007/2008 recession period. The link for the data is: <https://finance.yahoo.com/quote/%5EGSPC/history?period1=1167631200&period2=1573711200&interval=1d&filter=history&frequency=1d>

* NOAA’s weather station

Central Park weather station has been located inside Belvedere Castle since 1919. An automated surface observing system (ASOS) is installed in the south of the castle. However, the wind equipment is situated on the main tower. This dataset is obtained from NOAA’s website which is accessible through this link: <https://www.ncdc.noaa.gov/cdo-web/search>. However, a particular weather station has to be queried in order to download. Among all options of datasets NOAA offers, we chose the daily summaries ranging from November 2009 to November 2019 for Central Park weather station.

**Methods Section:**

1. Description of the original data files.

There are two data set which we will use for this project which is:

* S&P 500 (GSPC) has 3,241 observations in the dataset, each containing a single day’s worth of stock index price data.
* Weather data captured from NOAA’s weather station has 3,653 observations in this dataset, each containing daily record from this station located in the Central Park, New York.

1. Description of the guidelines used to validate the data

* We produced the inner joint of both original datasets by the date, called *combined*. This new dataset has 2,517 observations. The reason behind rather significant reduction of the number of observations is caused by the fact that the stock market does not operate on weekends and public holidays. However, we hope that the number of observations in the new dataset is sufficient to draw some conclusions later on.
* Several variables from the weather dataset has no or insignificant value, such as the average temperature (TAVG) and daily total sunshine (TSUN). Therefore, we drop them.
* We check for missing values for all variables using *proc means* procedure (see Appendix A), and we found that the S&P 500 price dataset does not have any missing value, but the weather dataset has several variables of which the number of missing values is rather large.

1. Description of the issue needed to be cleaned and how it will be done

* There is an issue when we want to use the weather dataset from NOAA since some variables have large number of missing values due the weather station that stopped recording the specific variables in certain range of time, such as the time of fastest mile or fastest 1-minute wind (FMTM) and peak gust time (PGTM). We assume qualitatively that these variables do not contribute significantly to the condition of the weather on any given day, thus we drop these variables.
* Some observations have missing values on several variables, but the number of these observations is insignificant. We drop these observations with missing values. The end result of this action leaves us with a new and clean dataset of 2,376 observations.

1. Description of additional data preparation
2. Description of variable to be analyzed

* S&P 500 (GSPC)

The variables of this dataset include Date which is the date of recorded price, Open which is the price of the stock when the market opened on this day , High which is the highest price the stock index has reached on this day, Low which is the lowest price the stock index has reached on this day, Close which is the price of the stock when the market closed on this day, Adj. Close which is the closing price after adjustments for all applicable splits and dividend distributions for the day, and Volume is the number of shares that has been bought and sold for the day.

* Weather Data from NOAA

There are fourteen variables that we use for this project in this dataset which are: AWND is the average wind speed, FMTM is the time of fastest 2-minute wind, PGTM is the peak gust time, PRCP is the precipitation, SNOW is the snowfall, SNWD is the snow depth, TAVG is the average temperature, TMAX is the maximum temperature, TMIN is the minimum temperature, TSUN is the daily total sunshine, WDF2 is the direction of fastest 2-minute wind, WDF5 is the direction of fastest 5-second wind, WSF2 is the fastest 2-minute wind speed, WSF5 is the fastest 5-second wind speed.

**Results:**

The following table reflects the changes in our datasets after they were trimmed and cleaned. The red-highlighted variables imply that they are removed. The yellow-highlighted variables imply that they have several missing values.

| **Variable\*** | **Label** | **N** | **N Miss** | **N** | **N Miss** | **N** | **N Miss** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Combined** | | **Trimmed** | | **Cleaned** | |
| DATE OPEN HIGH LOW CLOSE ADJ\_CLOSE VOLUME |  | 2517 2517 2517 2517 2517 2517 2517 | 0 0 0 0 0 0 0 | 2517 2517 2517 2517 2517 2517 2517 | 0 0 0 0 0 0 0 | 2376 2376 2376 2376 2376 2376 2376 | 0 0 0 0 0 0 0 |
| AWND FMTM PGTM PRCP SNOW SNWD TMAX TMIN WDF2 WDF5 WSF2 WSF5 | Average daily wind speed Time of fastest mile or fastest 1-minute wind Peak gust time Precipitation Snowfall Snow depth Maximum temperature Minimum temperature Direction of fastest 2-minute wind Direction of fastest 5-second wind Fastest 2-minute wind speed Fastest 5-second wind speed | 2399 475 468 2517 2517 2517 2517 2517 2407 2384 2408 2384 | 118 2042 2049 0 0 0 0 0 110 133 109 133 | 2399 N/A N/A 2517 2517 2517 2517 2517 2407 2384 2408 2384 | 118 N/A N/A 0 0 0 0 0 110 133 109 133 | 2376 N/A N/A 2376 2376 2376 2376 2376 2376 2376 2376 2376 | 0 N/A N/A 0 0 0 0 0 0 0 0 0 |

\*We omit the indicator variables from *weather* for brevity

We wish to select a model of which the response variable comes from the S&P 500 dataset with the predictors from the weather dataset. We choose ADJ\_CLOSE as the response variable because the price reflects the aftermath of the behavior of buyers and sellers at the daily closure of the stock exchange. As a measurement of choosing the best model, we consider the adjusted and multiple R2, RMSE, and SBC.

| **Adjusted R-Square** | **R-Square** | **Root MSE** | **SBC** | **Variables in Model** |
| --- | --- | --- | --- | --- |
| 0.3567 | 0.3605 | 458.77880 | 29225 | AWND SNOW SNWD TMAX TMIN WT01 WT03 WT05 WT07 WT13 WT16 WT18 WT19 WT22 |
| 0.3566 | 0.3607 | 458.82871 | 29232 | AWND SNOW SNWD TMAX TMIN WT01 WT03 WT05 WT07 WT13 WT16 WT17 WT18 WT19 WT22 |
| 0.3566 | 0.3606 | 458.83856 | 29232 | AWND SNOW SNWD TMAX TMIN WT01 WT02 WT03 WT05 WT07 WT13 WT16 WT18 WT19 WT22 |
| 0.3565 | 0.3601 | 458.84281 | 29218 | AWND SNWD TMAX TMIN WT01 WT03 WT05 WT07 WT13 WT16 WT18 WT19 WT22 |
| 0.3565 | 0.3606 | 458.85648 | 29232 | AWND SNOW SNWD TMAX TMIN WT01 WT03 WT04 WT05 WT07 WT13 WT16 WT18 WT19 WT22 |

However, the following ANOVA table of this selected model suggests that the regression of the model is not significant at any reasonable significance level. To show how not significant the regression is, the critical value for the F-distribution with (14,2361) degree of freedom at 10% is 1.51.

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 14 | 280157699 | 20011264 | 95.08 | <.0001 |
| **Error** | 2361 | 496938530 | 210478 |  |  |
| **Corrected Total** | 2375 | 777096229 |  |  |  |

Due to the nature of the regression of both datasets that is not well-behaved, linear regression seems to be an inappropriate method to analyze the relationship between stock price and weather. Therefore, we consider the correlation coefficient of the adjusted close price and all variables in the weather dataset. The following tables show these values.

| **Pearson Correlation Coefficients, N = 2376 Prob > |r| under H0: Rho=0** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **AWND** | **PRCP** | **SNOW** | **SNWD** | **TMAX** | **TMIN** |
| **ADJ\_CLOSE** | -0.21108 <.0001 | 0.00313 0.8789 | -0.03161 0.1235 | -0.07466 0.0003 | 0.10958 <.0001 | 0.11621 <.0001 |

The table below reflects point-biserial correlation coefficient because we are comparing continuous and categorical data.

| **Pearson Correlation Coefficients, N = 2376 Prob > |r| under H0: Rho=0** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **WT01** | **WT02** | **WT03** | **WT04** | **WT05** | **WT06** | **WT07** | **WT08** | **WT09** | **WT10\*** | **WT11** |
| **ADJ\_CLOSE** | 0.11288 <.0001 | -0.05230 0.0108 | 0.22252 <.0001 | 0.00814 0.6917 | -0.23779 <.0001 | 0.02253 0.2723 | -0.12354 <.0001 | 0.03606 0.0789 | -0.06912 0.0007 | . . | -0.02197 0.2845 |
|  | **WT12\*** | **WT13** | **WT14\*** | **WT15\*** | **WT16** | **WT17** | **WT18** | **WT19** | **WT20\*** | **WT21\*** | **WT22** |
| **ADJ\_CLOSE** | . . | -0.41422 <.0001 | . . | . . | -0.43933 <.0001 | -0.03765 0.0665 | -0.16640 <.0001 | -0.11575 <.0001 | . . | . . | -0.07990 <.0001 |

\*The empty boxes reflect that there is no record of that weather type in Central Park, New York between 2009 and 2019.

The adjusted close price of S&P 500 has an inverse relationship with the average daily wind speed in the Central Park area of New York City. The correlation coefficient is -21.11% with small p-value, which indicates that this value is significant. The average daily wind speed is a continuous predictor variable with the highest correlation coefficient. In other words, the S&P 500 price is slightly likely to go up when the wind speed is closer to zero, and vice versa. For indicator variables, mist (WT13) and rain (WT16) play a significant role in indirectly affecting the movement of the S&P 500. The point-biserial correlation coefficients are -41.42% and -43.93%, respectively, with small p-value. In brief, it is likely that S&P 500 price tends to decline when mist or rain presents on any given day.

Our decision to pick S&P 500, instead of other indices, is because S&P 500 represents 500 large companies. Whatever conclusion we can come up with, it is reasonable to assume that other stocks, both individual and indices, behave the same way as the S&P 500. Further direction of this study is to perform similar analysis to DJIA in the 1930s when trading activities were conducted physically in the stock market. If the relationship between the index price and the weather is stronger, this may imply that weather, indeed, affects trading behavior. This is our assumption because all trading activities are computerized nowadays.