DATA 602 Report

An Investigation of Pipeline Releases and Their Impact on Company Stock Prices

Bhaskar, D'Ippolito, Kaur

Table of Contents

1 INTRODUCTION

Pipelines are systems of connected pipes used to transport liquids and gases (mainly oil and natural gas) across long distances to deliver their contents to refineries, ports, and other markets. There are over 840 000 km of pipelines criss-crossed throughout Canada. Pipelines utilize many sophisticated computer safety systems and visual inspections in an attempt to prevent leaks, but incidents can still occur.

In this project, we'll utilize several different statistical methods to analyze Nova Gas Transmission. Nova Gas Transmission was selected from our dataset because they have the most incidents listed in our dataset. We will be performing a comparison between Nova and Spectra Energy transmission pipeline incident locations, incident impact on share price, and a regression model to predict the number of pipeline spills in a given year.

All data and calculations can be seen in the attached Appendices at the end of the report.

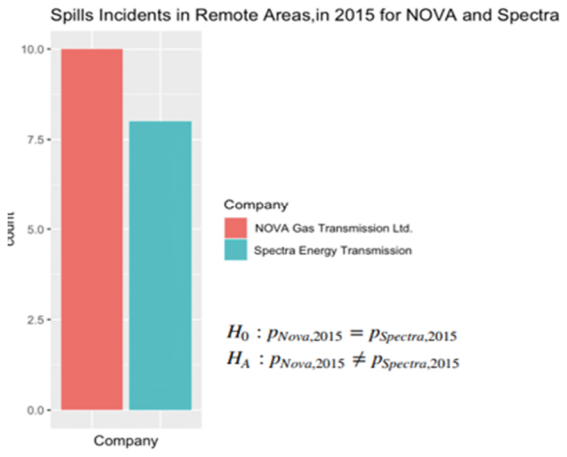
2 TOPICS OF INVESTIGATION

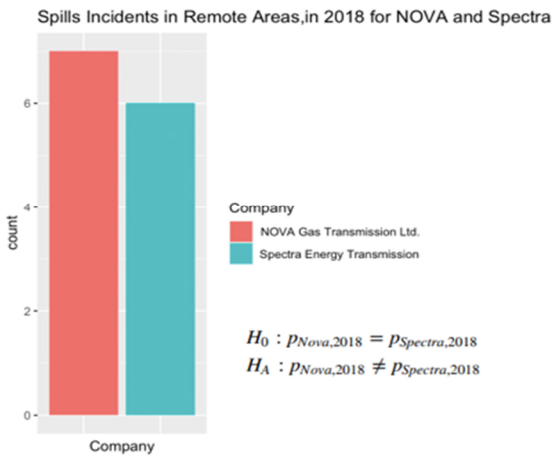
2.1 Company Pipeline Spill Location Comparison

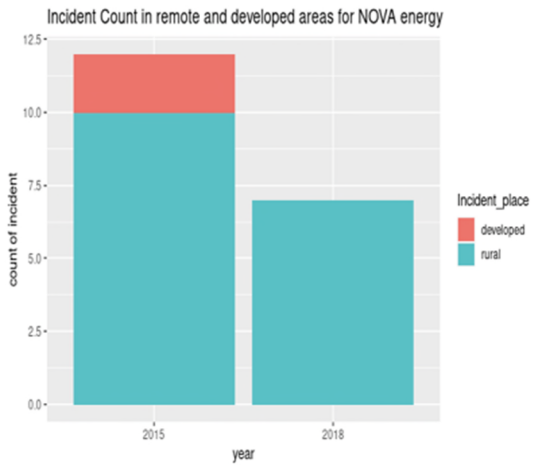
Our first major topic of investigation will be a comparison of the proportion of incidents that happened in remote locations between the companies Nova Gas Transmission and Spectra Energy Transmission. Our initial analysis discovered that these two companies are involved in many pipeline incidents that occur in Canada and provide the most data points to investigate. We will be comparing the proportion of spill incidents that happen in remote locations (instead of developed) between Nova Gas Transmission and Spectra Energy Transmission.

Our comparison will look at the years 2015 and 2018. We will create a conventional confidence interval for each year. Our null hypothesis is that there is no difference in proportion between the two companies. Our alternative hypothesis is that there is difference in proportion of pipeline spill location between the two companies. We will do a parametric comparison and calculate the probability of a Type 1 Error against an alpha of 0.05.

Data Visualizations:







Total Incident Count for Nova Gas Transmission

Total Incident Count for Spectra Transmission

Data:

For 2015 there were 12 total spill incidents for Nova and 10 of them occurred in remote locations. Spectra had 29 total spill incidents and 8 of them occurred in remote locations.

For 2018 there were 7 total spill incidents for Nova and all of them occurred in remote locations. Spectra had 15 total spill incidents and 6 of them occurred in remote locations.

Hypothesis:

First we will perform a parametric test utilizing the following formulas.

$$\hat{p} = \frac{X_{Nova} + X_{Spectra}}{n_{Nova} + n_{Spectra}} \quad and \quad Z_{obs} = \frac{\hat{p}_{Nova} - \hat{p}_{Spectra} - (p_{Nova} - p_{Spectra})}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{Nova}} + \frac{1}{n_{Spectra}}\right)}}$$

$$H_0 : p_{Nova,2015} = p_{Spectra,2015} \qquad H_0 : p_{Nova,2018} = p_{Spectra,2018}$$
$$H_A : p_{Nova,2015} \neq p_{Spectra,2015} \qquad H_A : p_{Nova,2018} \neq p_{Spectra,2018}$$

2015 Analysis:

$$\widehat{p}_{Nova2015} = \frac{10}{12} = 0.8333$$

$$\widehat{p}_{Spectra2015} = \frac{8}{29} = 0.2759$$

$$\widehat{p}_{2015} = \frac{10+8}{12+29} = 0.439$$

$$Z_{2015} = \frac{0.8333-0.2759-0}{\sqrt{0.439(1-0.439)\left(\frac{1}{12}+\frac{1}{29}\right)}} = 3.2726$$

$$P(Z > 3.2726) * 2 = 0.0011$$

Since our P value is < 0.05 we can reject the null hypothesis and support the alternative hypothesis that

$$H_A : p_{Nova,2015} \neq p_{Spectra,2015}$$

2018 Analysis:

$$\widehat{p}_{Nova2018} = \frac{7}{7} = 1$$

$$\widehat{p}_{Spectra2018} = \frac{6}{15} = 0.4$$

$$\widehat{p}_{2018} = \frac{7+6}{7+15} = 0.5909$$

$$Z_{2018} = \frac{1-0.4-0}{\sqrt{0.5909(1-0.5909)\left(\frac{1}{7}+\frac{1}{15}\right)}} = 2.666$$

$$P(Z > 2.666) * 2 = 0.0077$$

Since our P value is < 0.05 we can reject the null hypothesis and support the alternative hypothesis that

$$H_A : p_{Nova,2018} \neq p_{Spectra,2018}$$

Confidence Interval Estimation:

Confidence intervals for the difference of two proportions can be calculated using the following formulas.

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

$$(\tilde{p}1 - \tilde{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

2015 Analysis:

$$\widetilde{p}_{Nova} = \tfrac{10+1}{12+2} = 0.7857$$

$$\widetilde{p}_{Spectra} = \tfrac{12+1}{29+2} = 0.4194$$

$$(\widetilde{p}_{Nova} - \widetilde{p}_{Spectra}) \pm 1.96\sqrt{\tfrac{0.7857(1-0.7857)}{12+2} + \tfrac{0.4194(1-0.4194)}{29+2}}$$

Our 95% confidence interval for 2015 is

$$0.2276 \leq p_{Nova} - p_{Spectra} \leq 0.7632$$

2018 Analysis:

$$\widetilde{p}_{Nova} = \tfrac{7+1}{7+2} = 0.8889$$

$$\widetilde{p}_{Spectra} = \tfrac{6+1}{15+2} = 0.4118$$

$$(\widetilde{p}_{Nova} - \widetilde{p}_{Spectra}) \pm 1.96\sqrt{\tfrac{0.8889(1-0.8889)}{7+2} + \tfrac{0.4118(1-0.4118)}{15+2}}$$

Our 95% confidence interval for 2018 is

$$0.1659 \leq p_{Nova} - p_{Spectra} \leq 0.7884$$

Permutation Testing:

Our Null Hypotheses states that there is no difference in the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2015.

$$p_{Nova} - p_{Spectra} = \frac{10}{12} - \frac{8}{29} = 0.5574713$$

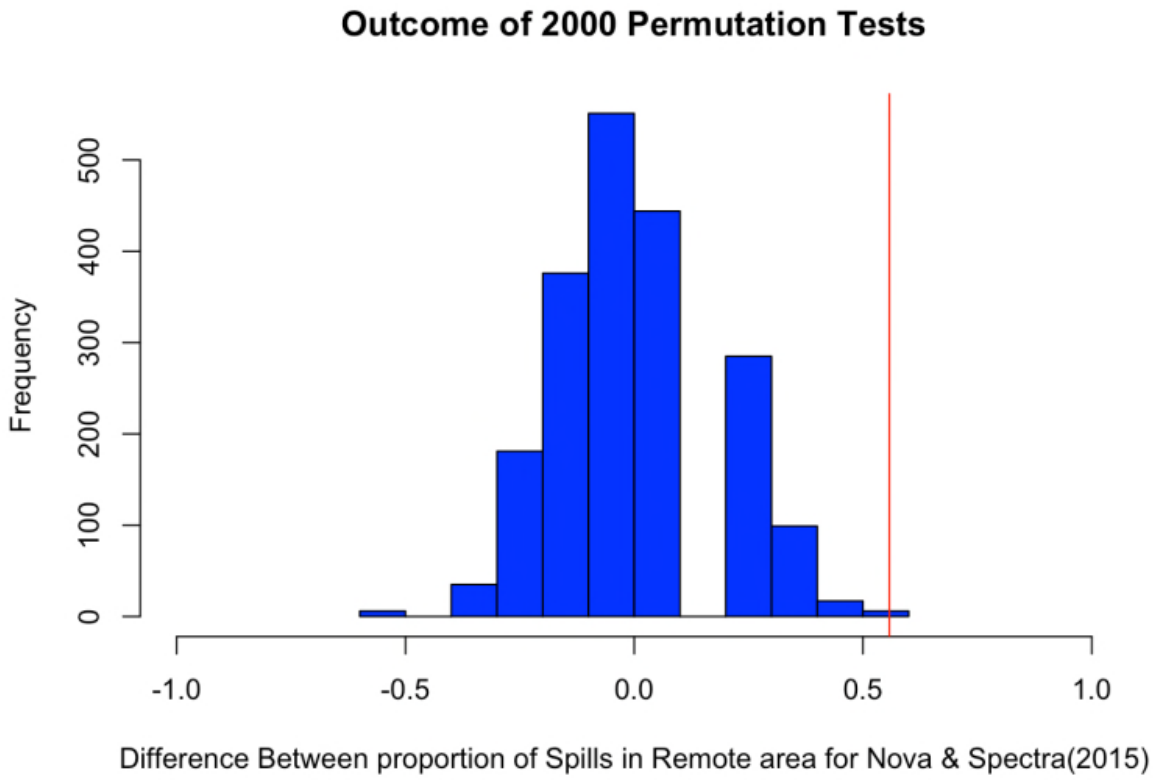
The difference between the proportions is 0.5575. Does the difference between the two sample proportions indicate that the proportion of Spill Incidents for Spectra Energy is more than NOVA Gas Transmission in 2015?

The difference between $p_{Nova} - p_{Spectra} = \frac{10}{12} - \frac{8}{29} = 0.5574713$ The question we can ask ourselves in this: is this difference between the two sample proportion indicate that proportion of Spill Incidents for Spectra Energy is more than the NOVA Gas company in 2015?

If we took all 41 cases and *randomly* of them to the 23 "NOVA Incidents" and 11 "Spectra Incidents" , there would be $\binom{41}{12}$ or $\binom{41}{29}$ = 7898654920 different *permutations* of these data and the same number of differences between $\bar{p}_{NOVA} - \bar{p}_{Spectra}$. We can generate some (not all) of these differenes by randomly assigning the 34 data values to the two different groups of 23 and 11, compute the difference $\bar{p}_{NOVA} - \bar{p}_{Spectra}$ *each* time, and see where the current observed difference of 0.5574713 lies on such a distribution of differences.

If the observed difference of 0.5574713 falls in the extreme tail of such a distribution, then such an observed difference is not likely and the null hypothesis of "proportions" for NOVA & Spectra would appear not to be the case.

Our permutation distribution looks as follows for 2015:

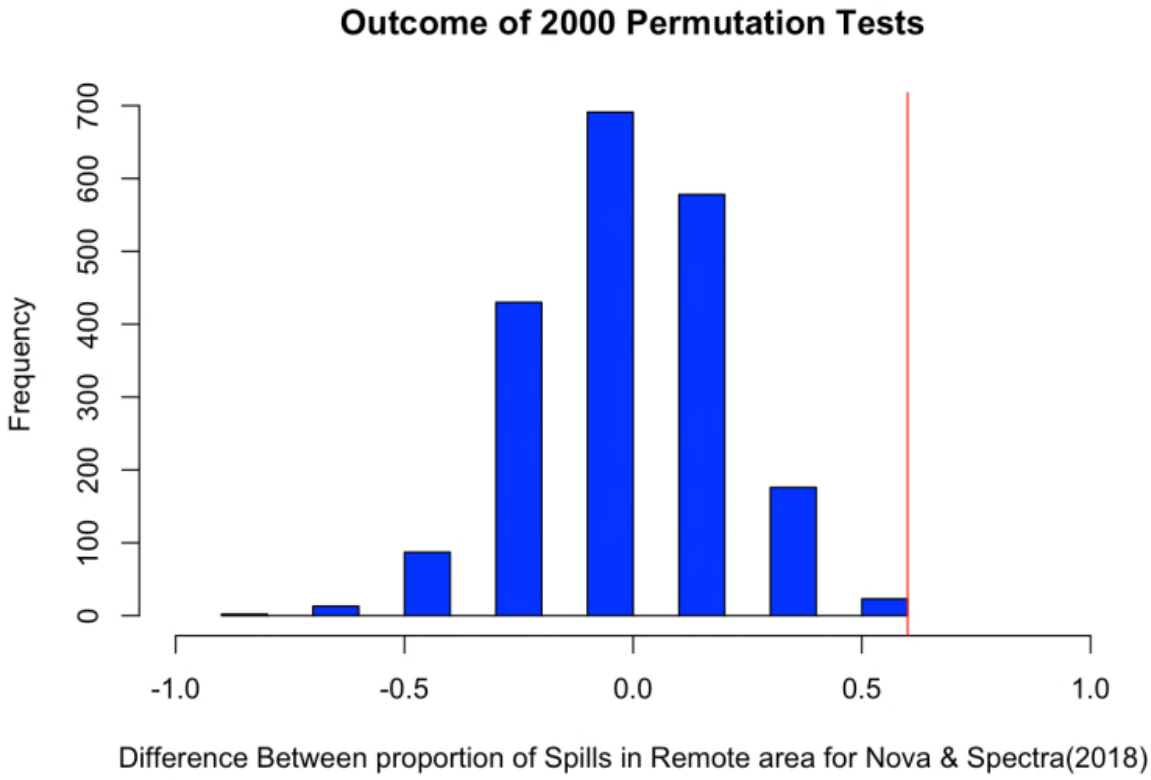


This provides an empirical P-value of 0.0005 and rejects our null hypothesis. This provides support for our alternative hypothesis that:

$$H_A : p_{Nova,2015} - p_{Spectra,2015} \neq 0$$

For 2018 our observed difference in proportions between samples is

$$\bar{p}_{NOVA} - \bar{p}_{Spectra} = 0.6$$



This provides an empirical P-value of 0.0085 and rejects our null hypothesis. This provides support for our alternative hypothesis that:

$$H_A : p_{Nova,2018} \neq p_{Spectra,2018}$$

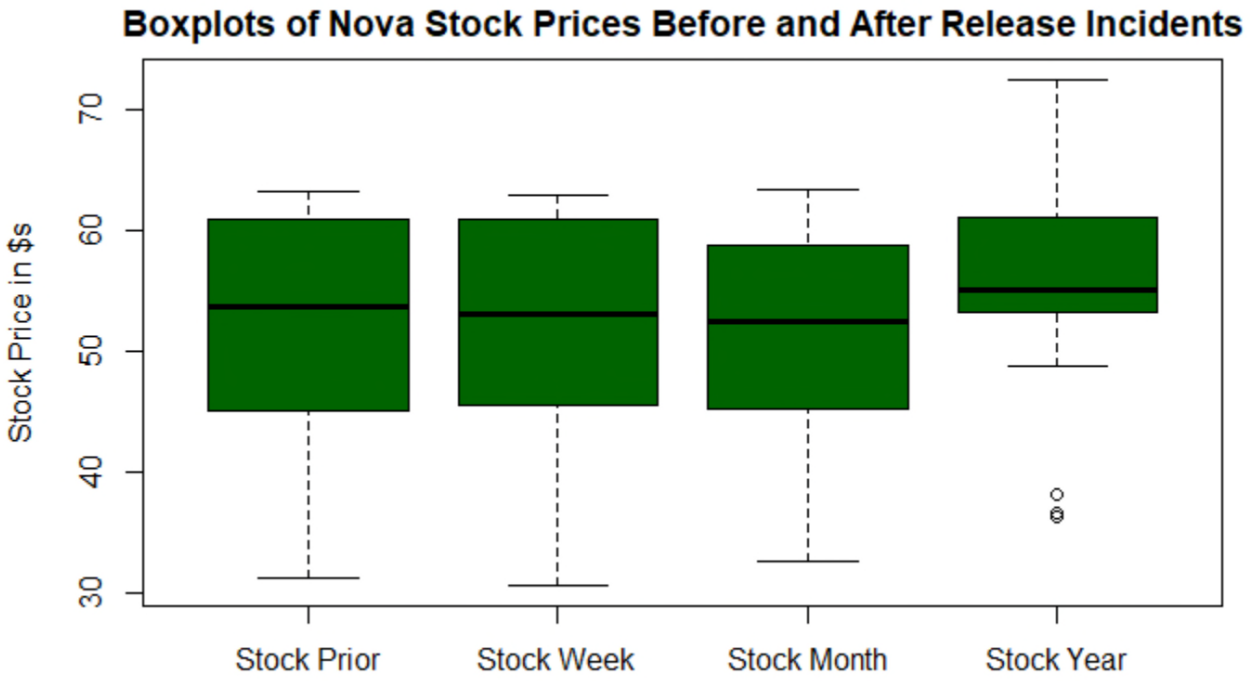
2.2 Spill Impact on Share Price

Our second topic of investigation is to analyze whether Nova Gas Transmission's stock price declines after a spill incident. We will be analyzing the difference in stock price between the day before a spill incident and 1 week, 1 month, and 1 year after. Our goal is to determine whether spill events have an impact on stock price and if so, is it short or long term.

We will be utilizing the t-test and matched pairs experimental design. Traditional and bootstrap confidence intervals will be provided.

Data:

Our data consists of spill incidents where the volume of natural gas released is > 1000 m^3 and consists of incidents from 2008 to 2019. Nova is a subsidiary of TC Energy, so our stock price data will be for the parent company. The stock price data before and after incidents can be visualized by the box plots below.

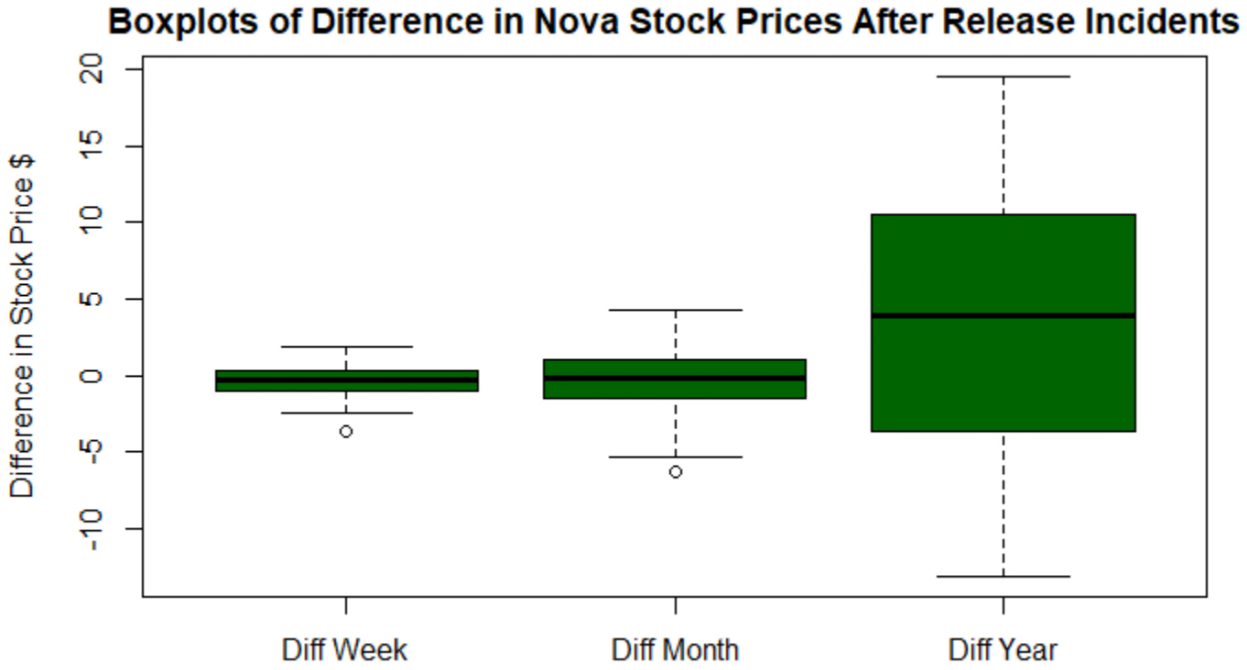


When looking at the data we can see an apparent small decline in the median from the price before the incident to the week and month after. Looking at the data is does not appear likely that spills have an impact on stock price one year later but we will perform the statistical analysis regardless.

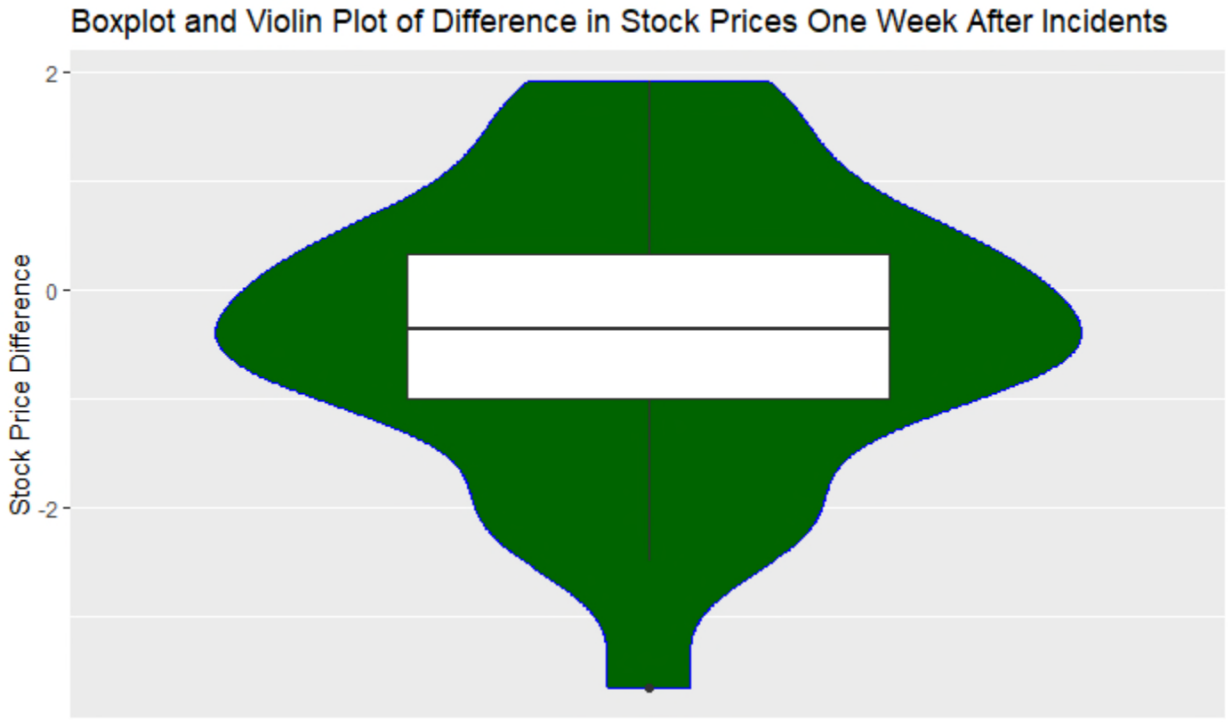
We will calculate the difference in stock price utilizing the formula below

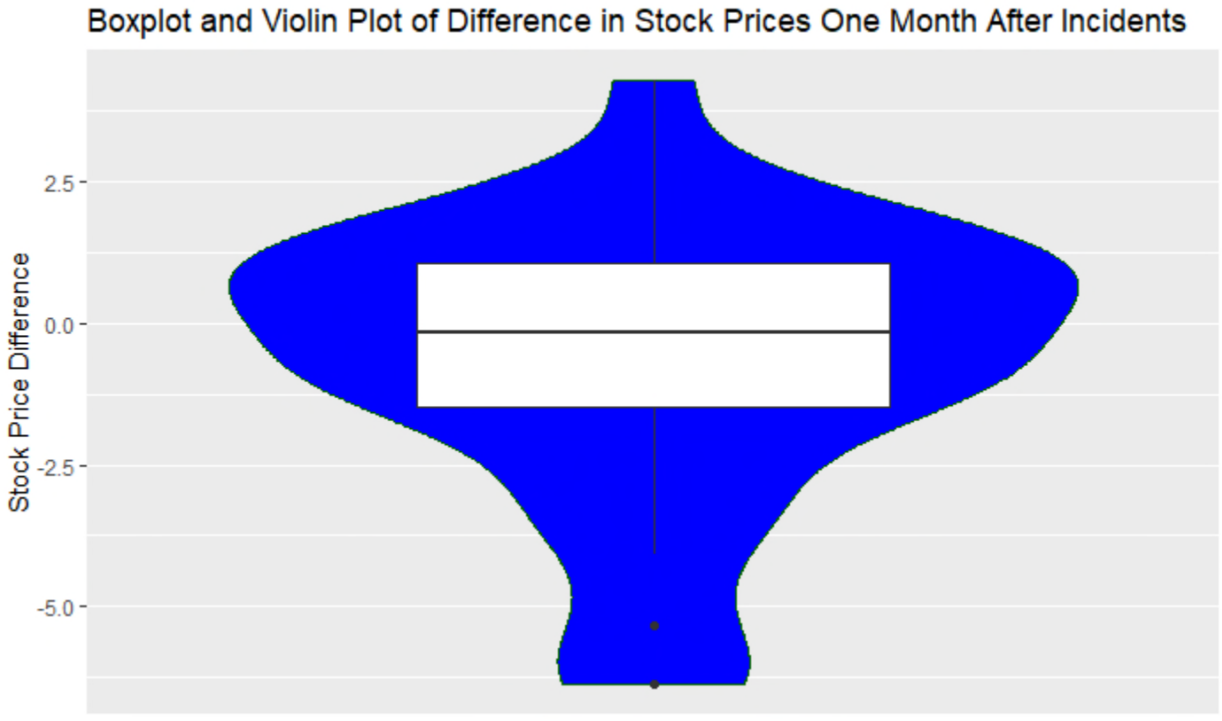
$$d_i = x_{1,i} - x_{2,i}$$

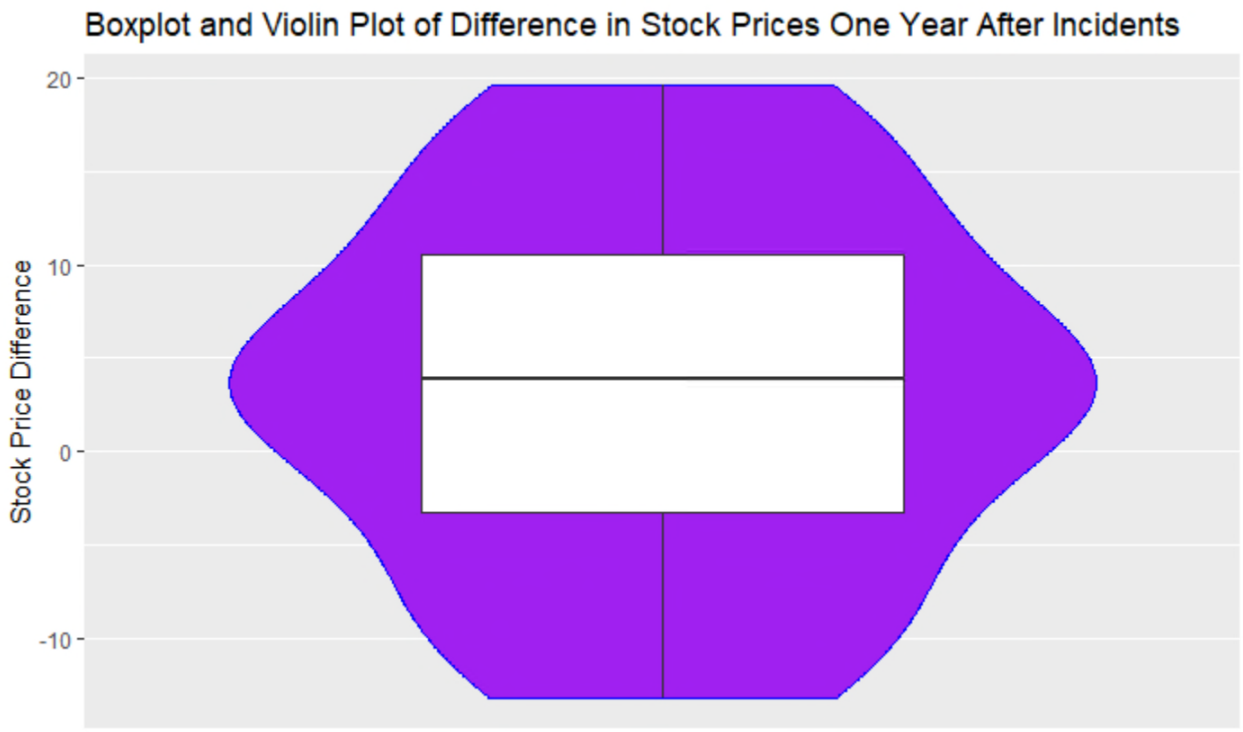
Where x1 is the stock price prior to spill incidents and x2 is the stock price a week, month, and year later. This difference in price can be visualized by the boxplots below.



Looking at the data above, we can see that the standard deviation of the samples increases as the timespan increases. We can also see that the median of the week and month differentials are just below zero. The spread in data for one year later is quite large but it appears that most of the data is above zero. The distribution of this data can be seen in more detail with the violin plots below.


Boxplot and Violin Plot of Difference in Stock Prices One Week After Incidents
Stock Price Difference
2
0
-2


Boxplot and Violin Plot of Difference in Stock Prices One Month After Incidents
Stock Price Difference
2.5
0.0
-2.5
-5.0



Analysis:

Our 3 sets of hypothesis are below:

$$\mathrm{H}_0 : \mu_{Week} - \mu_{Prior} \geq 0$$

$$\mathrm{H}_A : \mu_{Week} - \mu_{Prior} < 0$$

$$\mathrm{H}_0 : \mu_{Month} - \mu_{Prior} \geq 0$$

$$\mathrm{H}_A : \mu_{Month} - \mu_{Prior} < 0$$

$$\mathrm{H}_0 : \mu_{Year} - \mu_{Prior} \geq 0$$

$$\mathrm{H}_A : \mu_{Year} - \mu_{Prior} < 0$$

We will utilize the student's t test statistic to calculate our P values compared to an alpha of 0.05. The test statistic can be calculated using the formula below:

$$T_{obs} = \frac{\bar{d} - 0}{\frac{S_d}{\sqrt{n}}}$$

Our values for this analysis are as follows:

$$\bar{d}_{DiffWeek} = -0.3867$$

$$\bar{d}_{DiffMonth} = -0.669$$

$$\bar{d}_{DiffYear} = 3.2423$$

$$\sigma_{DiffWeek} = 1.3122$$

$$\sigma_{DiffMonth} = 2.5422$$

$$\sigma_{DiffYear} = 9.5963$$

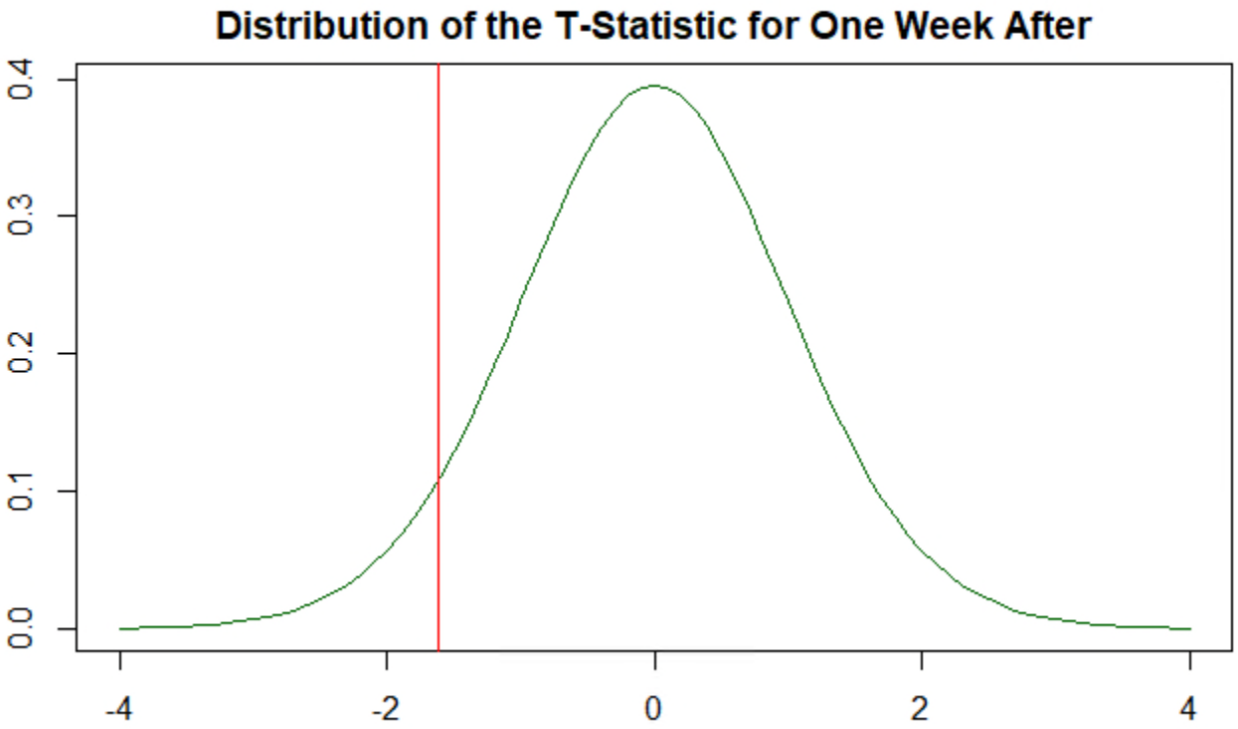
$$n = 30$$

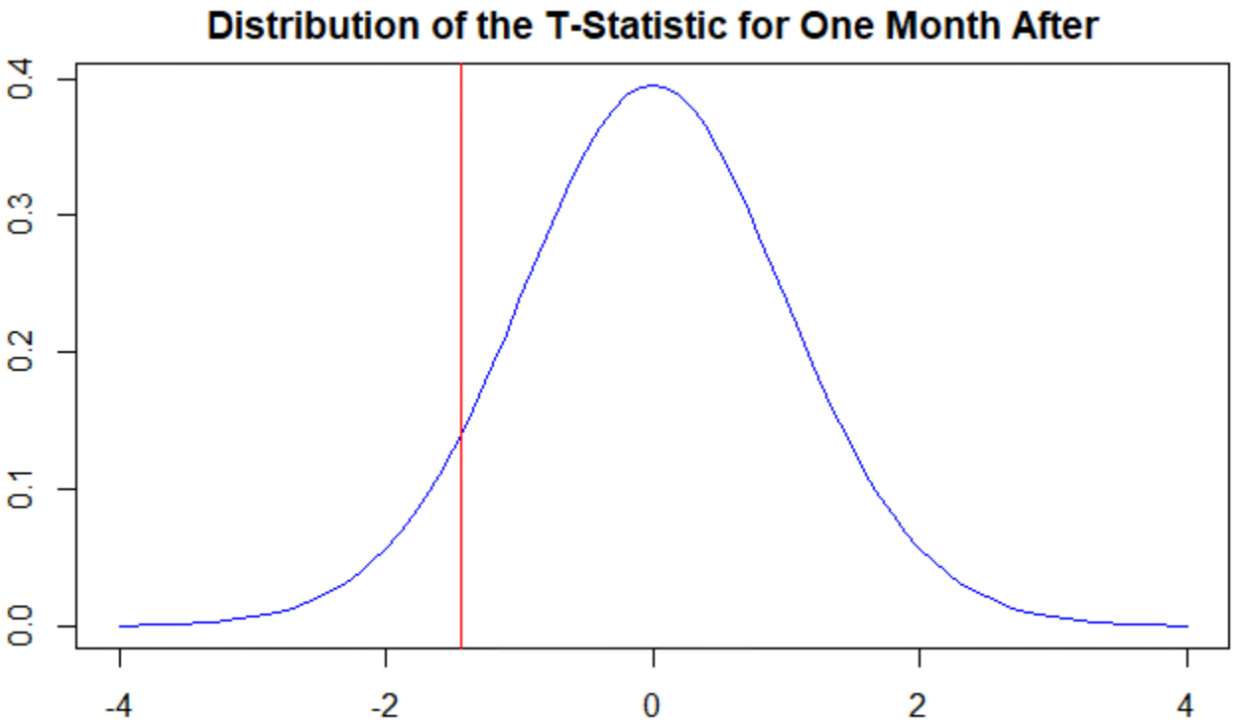
$$T_{week} = -1.6139$$

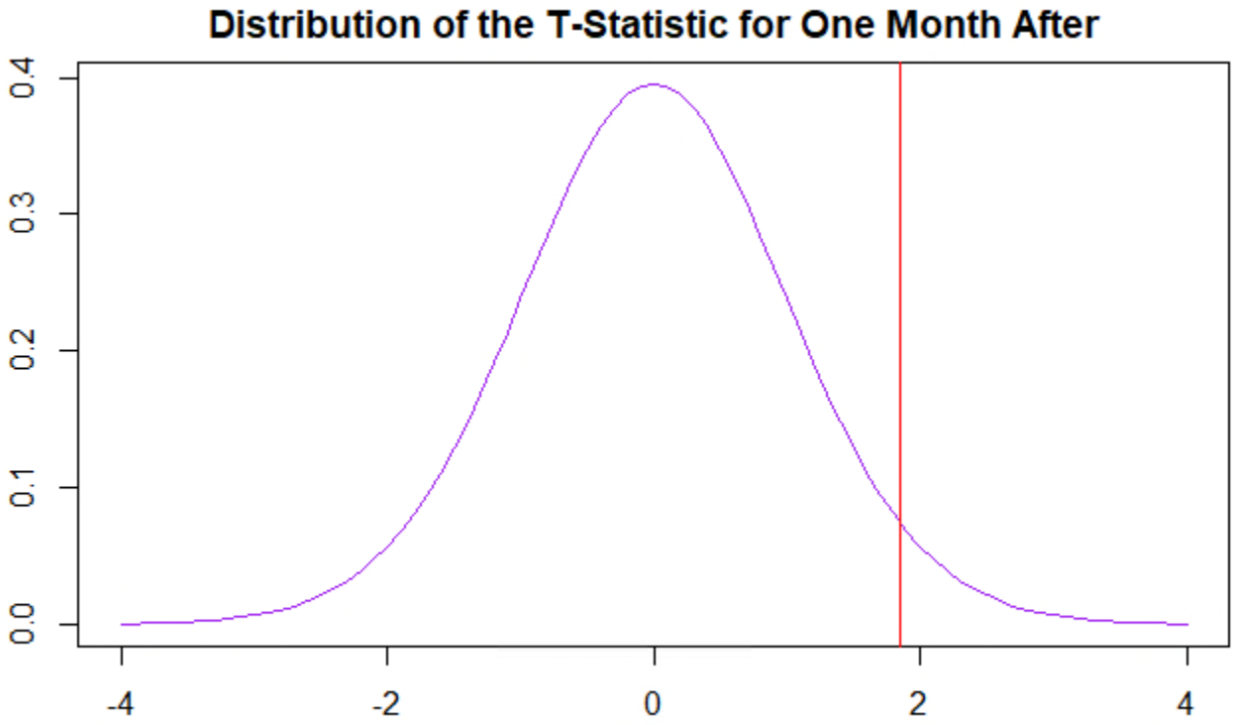
$$T_{month} = -1.4414$$

$$T_{year} = 1.8506$$

To visualize these T statistics they have been plotted on the trends below:


Distribution of the T-Statistic for One Week After
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


Distribution of the T-Statistic for One Month After
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4



Our P values are calculated as below:

$$P(T_{30-1} < T_{week}) = 0.0587$$

$$P(T_{30-1} < T_{month}) = 0.0801$$

$$P(T_{30-1} < T_{year}) = 0.9628$$

Since all of our P-values are >0.05 we must support all 3 of our null hypotheses.

$$\mathrm{H}_0 : \mu_{Week} - \mu_{Prior} \geq 0$$

$$\mathrm{H}_0 : \mu_{Month} - \mu_{Prior} \geq 0$$

$$\mathrm{H}_0 : \mu_{Year} - \mu_{Prior} \geq 0$$

The P values for the week and month differential are quite close to our alpha value of 0.05. However, more data is required if we would like to reject our null hypothesis.

Confidence Intervals:

Confidence intervals can be calculated using the formula below:

$$\bar{d} \pm t_{1-\frac{\alpha}{2},n-1}\left(\frac{S}{\sqrt{n}}\right)$$

Using the formula above our 95% confidence intervals are:

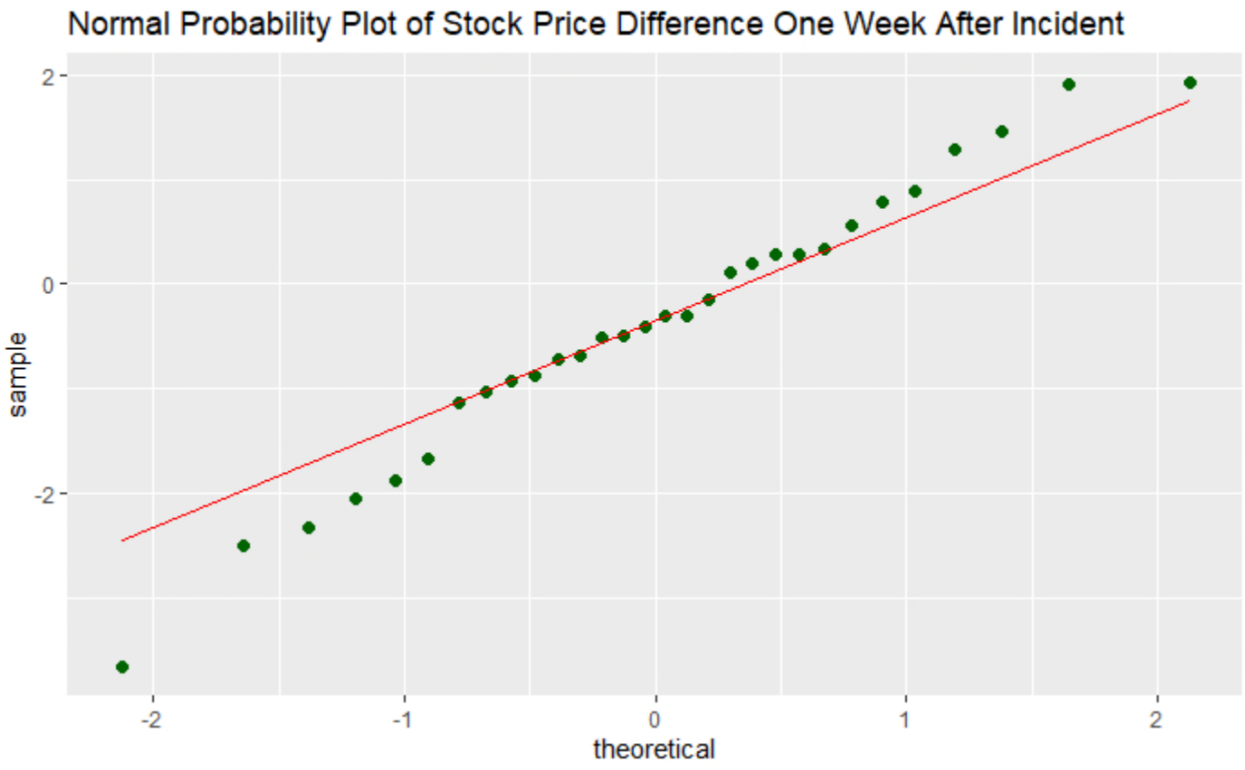
$$-0.8767 \leq \mu_{WeekDiff} \leq 0.1033$$

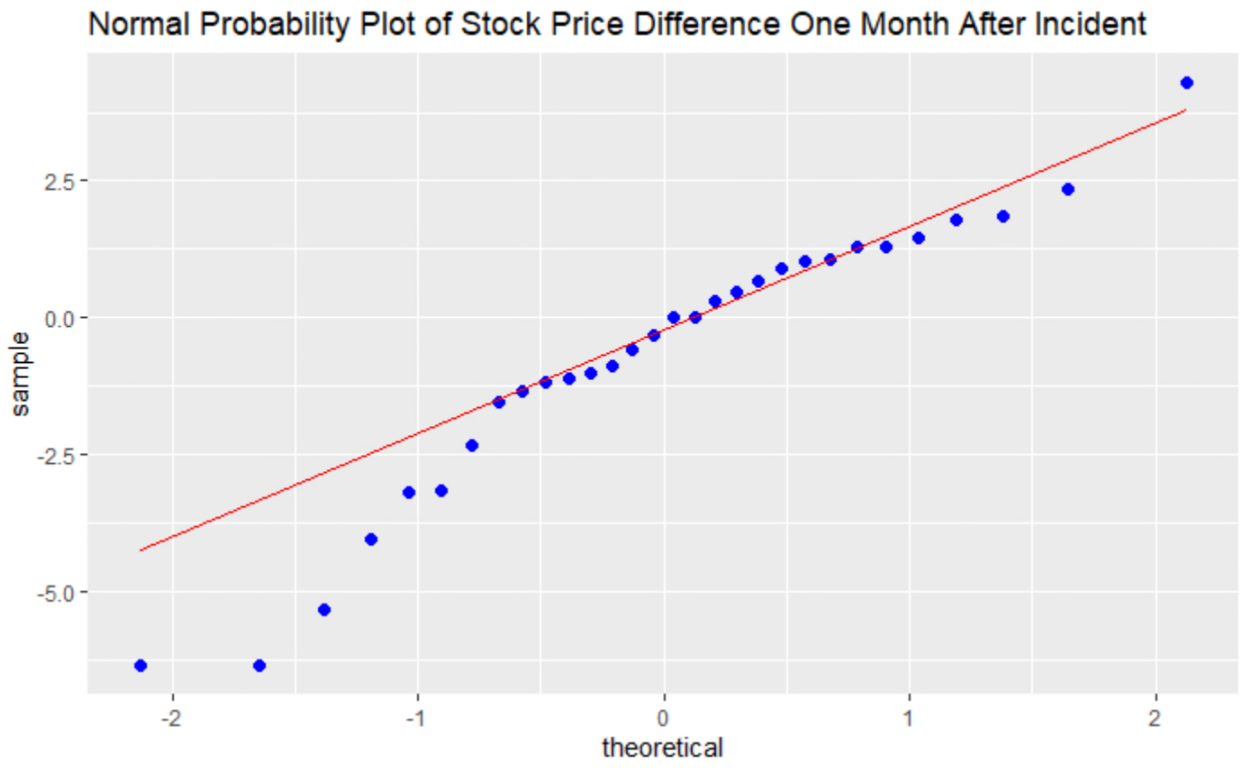
$$-1.6183 \leq \mu_{MonthDiff} \leq 0.2803$$

$$-0.341 \leq \mu_{YearDiff} \leq 6.8257$$

Bootstrap Analysis:

In order to perform a bootstrap analysis we must first create normal probability plots to verify if our data follows a Normal distribution. The normal probability plots for week, month, and year differential are below:

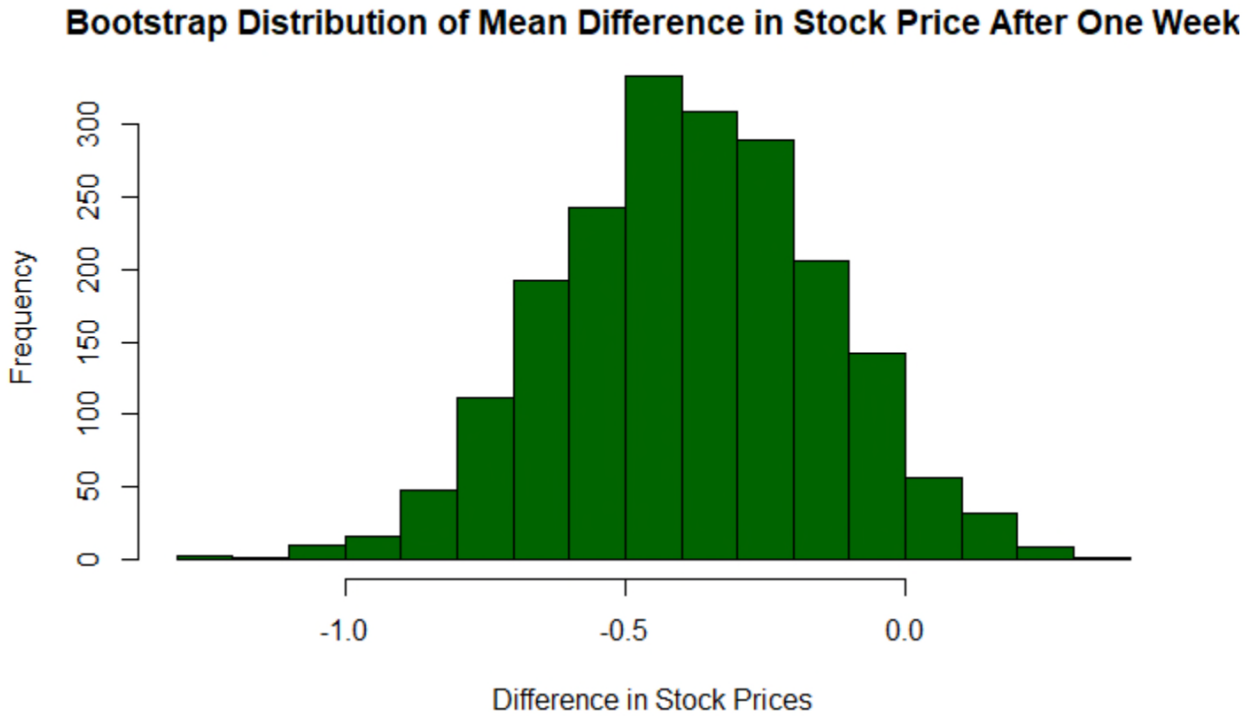


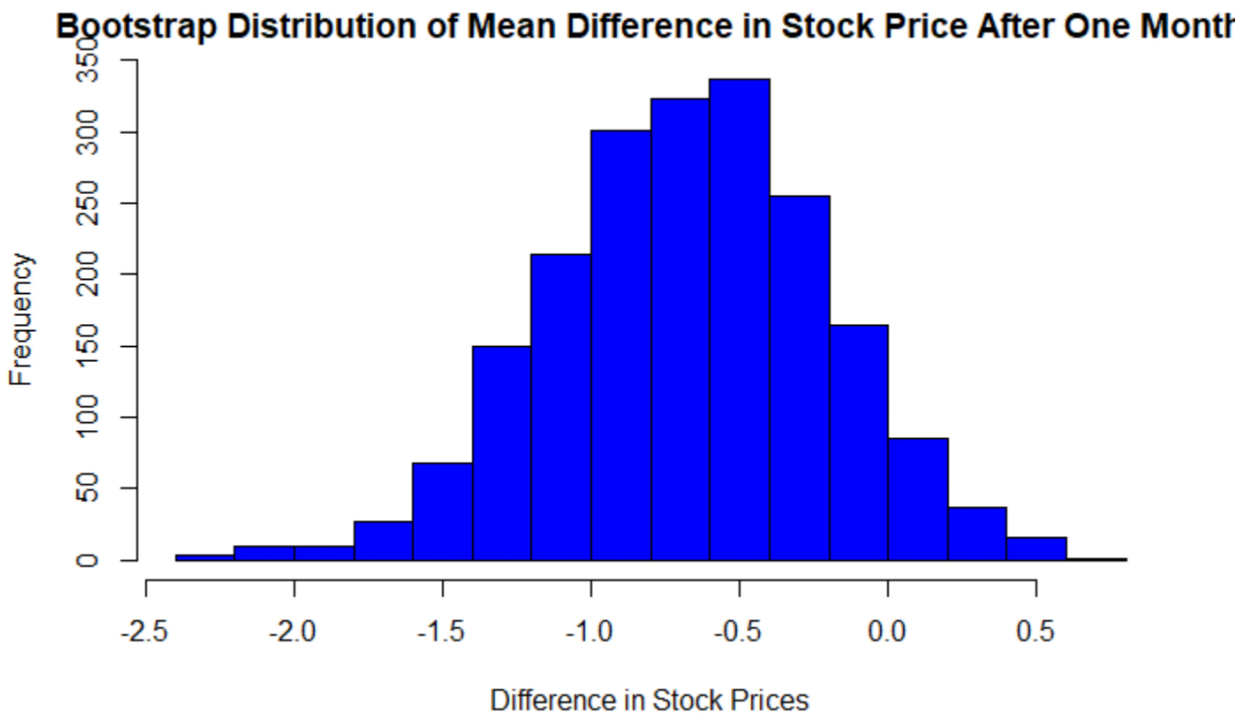

Normal Probability Plot of Stock Price Difference One Month After Incident
sample
2.5
0.0
-2.5
-5.0
-2
-1
0
1
2
theoretical

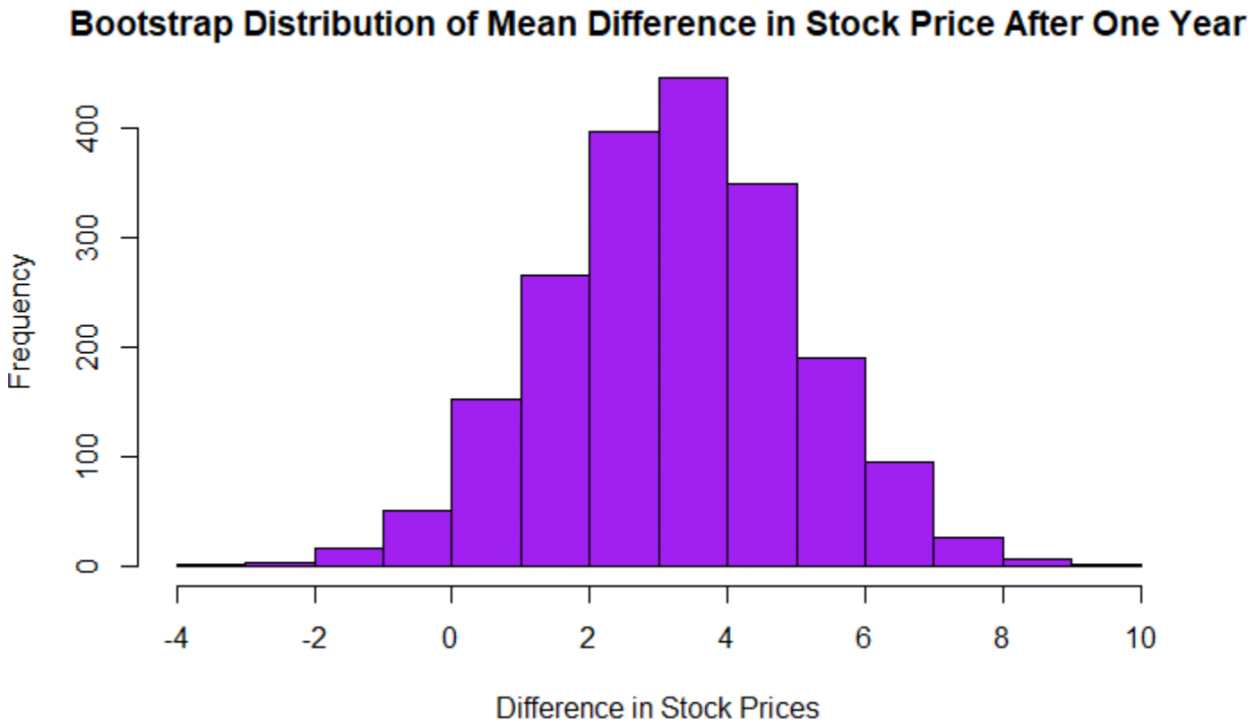

Normal Probability Plot of Stock Price Difference One Year After Incident
sample
20
10
0
-10
-20
-2
-1
0
1
2
theoretical

Based on the plots above we can conclude that our data does follow a Normal distribution. Some of the lower values for month differential do begin to stray from the Normal line but the overall shape of the data appears to be Normal.

We will create a bootstrap distribution for our data by resampling the week, month, and year data n=30 times. 2000 bootstrap trials were run and the distribution is plotted below.







Next we can create 95% confidence intervals by taking the 2.5% and 97.5% quantiles. Those bootstrap intervals are shown below:

$$-0.8401 \leq \mu_{WeekDiff} \leq 0.0817$$

$$-1.5957 \leq \mu_{MonthDiff} \leq 0.2191$$

$$-0.2188 \leq \mu_{YearDiff} \leq 6.6676$$

Since all of our bootstrap confidence intervals contain 0 we can't conclusively say that there is a difference in share price caused by spill incidents. This agrees with the previous student's t test analysis.

2.3 Annual Number of Spills Prediction

We will be using linear regression along with conventional and bootstrap techniques to predict, with 95% confidence, how many spill incidents will occur for NOVA GAS Transmission in the year 2018. Data from 2009 to 2017 will be utilized.

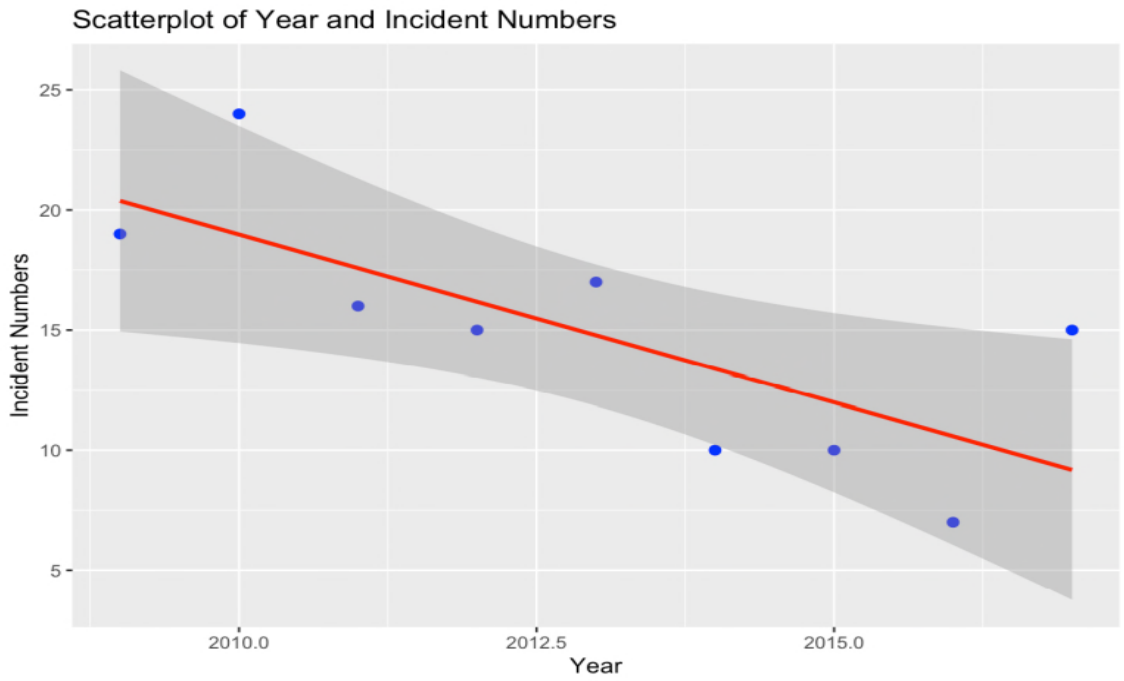
To begin our analysis we first read in the data and filter out the results to get the count of incidents for Nova gas from 2009-2017 leading to release of substance. Our derived results were set into a data frame as below.

-

Year <dbl>	IncidentNumber <dbl>
2009	19
2010	24
2011	16
2012	15
2013	17
2014	10
2015	10
2016	7
2017	15

9 rows

To start our analysis we first checked the relationship between our X and Y variables. A scatter plot of the results is plotted below.



The graph shows a downward trend as the years increase. Our next step was to quantify the relationship by deriving the correlation coefficient. The correlation coefficient is a statistic that computes the degree of linear association between variables X and Y. The formula used to compute the correlation is below:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n(\bar{X} * \bar{Y})}{(n-1)S_X S_Y} \qquad \text{where } -1 \le r \le 1$$

Our correlation coefficient was = -0.7386. We determined that the linear relationship between our X variable i.e the count of incidents and the Y variable i.e. the year of incident, is significant. Next, we created a statistical/mathematical model by utilizing linear regression.

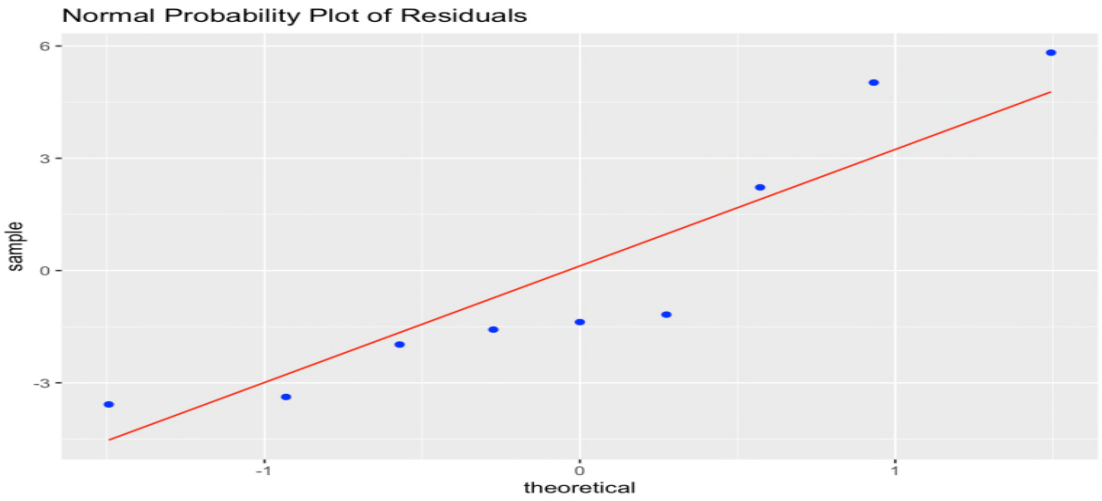
$$IncidentNumbers_i = A + B * Year_i + e_i \qquad i = 2009, 2010, \cdots, 2017$$

$$\widehat{IncidentNumbers_i} = a + b * Year_i$$

Next we must check that our linear regression model is valid. This can be confirmed by checking the two conditions below:

1. The y -variable, or commonly known as the response variable, is Normally distributed with a mean μ and standard deviation of σ.
2. For each distinct value of the x-variable (the predictor variable), the y variable has the same standard deviation σ.

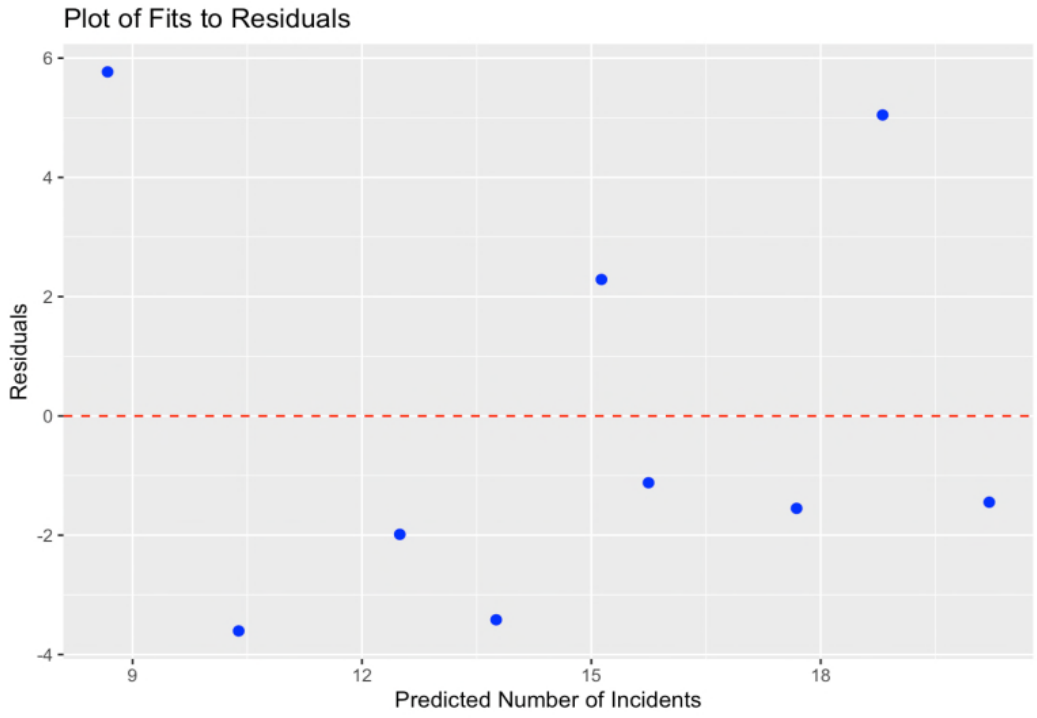
Validity Check 1 : Is the Y-variable Normally Distributed ?

- 

Our Result : Since the normal probability plot produces roughly a straight line throughout the middle of the points we concluded that the residuals of our model appear to follow a Normal distribution.

-

Validity Check 2 : For each distinct value of the x-variable, does the y variable have the same standard deviation ?

- 

Our Result : Looking at this plot we see a relatively consistent standard deviation.

After this our next step was to estimate true value of the Y-intercept term and the slope term of the model i.e. a and b of our model using lease square estimate

We computed these results in R studio and we got an estimate model to predict our results.

-

Our Estimate Model :

- $$\widehat{IncidentNumber}_i = 2832.978 - 1.400 * Year_i$$
-

Then using the R-studio we evaluated 95% confidence interval for our Incident Number

With conventional predict command we found that our Incident count with 95% confidence interval is in the following range:

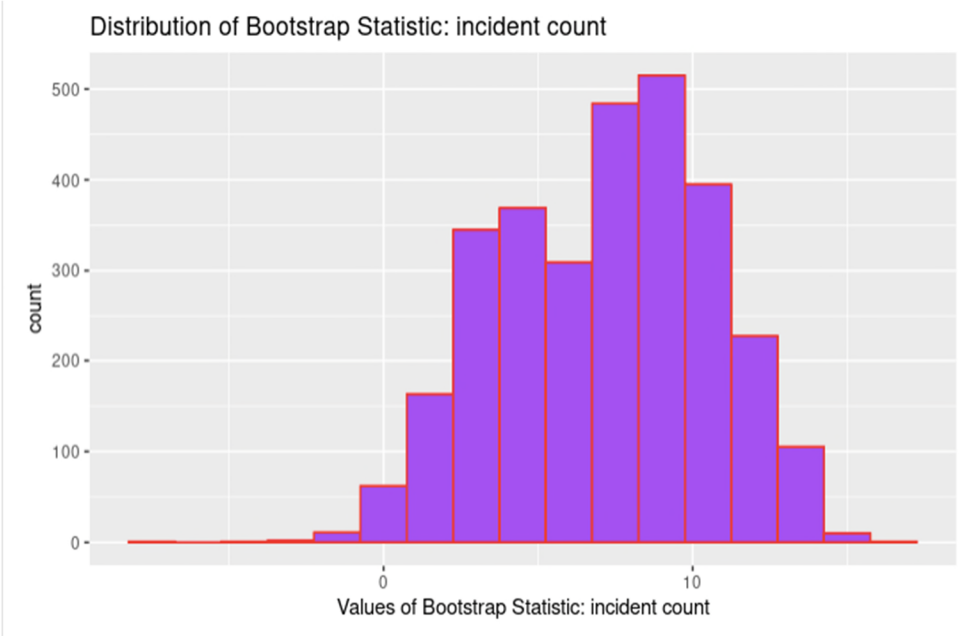
$$-3.155685 \leq IncidentNumber|_{year=2018} \leq 18.71124$$

-
-

To further analyze we then used Bootstrap strategy to find the confidence interval for the count of incidents, we used the same data frame to count our results and sampled it 3000 times to get our Bootstrap results

Our Bootstrap results were :

-



min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
-7.103448	4.351339	7.555336	9.719094	15.97674	7.174171	3.414368	3000	0

The Bootstrap 95% confidence interval was:

$$0.7330 \leq IncidentNumber|_{year=2018} \leq 13.1933$$

Further we tried to see "how well" our statistical model mimics the real-world through the computation of the **coefficient of determination**

Using R we computed **coefficient of determination:**

- $r^2 = 0.545567$

Our Final step was to test the overall linear appropriateness of the model

To perform that we used the technique of Hypotheses Testing

Our Hypotheses was defined as :

$$H0 : Incident\ Count\ CANNOT\ be\ expressed\ as\ a\ linear\ function\ of\ Year$$
$$HA : Incident\ Count\ CAN\ be\ expressed\ as\ a\ linear\ function\ of\ Year$$

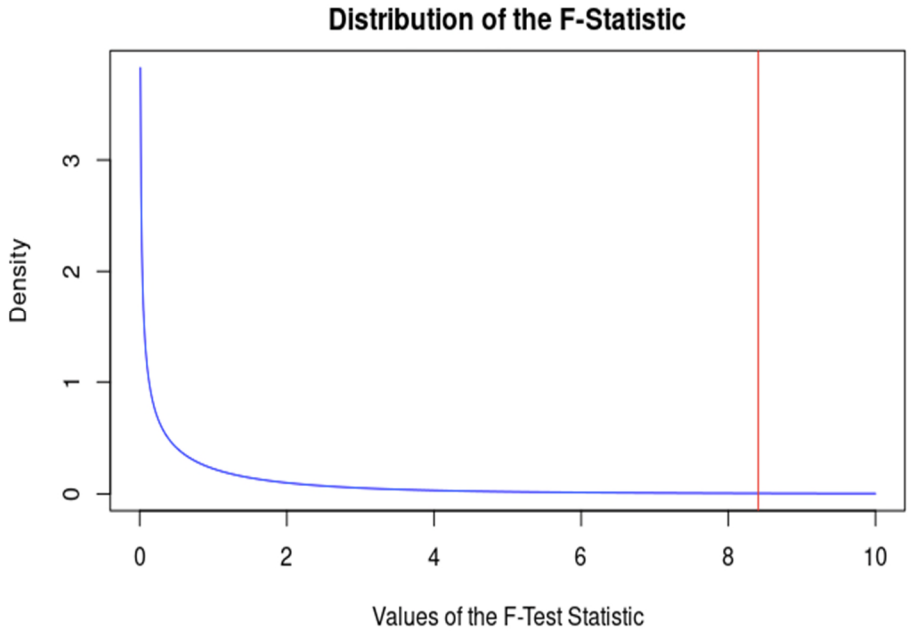
To check our hypotheses we defined our Test Statistic as :

$$F_{Obs} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \sim F_{1,n-2}$$

Using R we evaluated our Test statistic :

$$F_{Obs} = \frac{\frac{117.60}{1}}{\frac{97.96}{7}} = \frac{117.60}{13.99} = 8.406004$$

We then plotted the distribution of F-Test statistic :



Using R studio we implemented the summary command on our linear model to derive our P-value. Our P-value was :

$$P - \text{value} = P(F_{1,7} > 8.406004) = 0.02300853$$

Looking at this result we rejected our Null hypotheses and conclude that the count of Incidents can be expressed as a linear function of number of Year and our model can be used to find the count of incidents for the Year = 2018.

Finally we expressed our results as :

$$\widehat{IncidentNumber}_i = 2832.978 - 1.400 * Year_i$$
$$\widehat{IncidentNumber}_i = 2832.978 - 1.400 * 2018$$
$$\widehat{IncidentNumber}_i = 7.778$$

The actual number of Nova spill incidents for 2018 was 7.

APPENDIX A TOPIC 1 R CODE AND ANALYSIS

Data 602 - Project Task 1

Topic 1: Comparison of the proportion of spill incidents that happened in Remote locations between the Companies NOVA Gas Transmission and Spectra Energy Transmission for year 2015 and 2018.

Read .csv File

```
Pipelinedata = read.csv("/users/guldeepkaur/Downloads/ncdntcmprhnsv-eng.csv",fileEncoding = "Latin1", cl
#head(Pipelinedata,5)
```

Code to filter out data related to Companies "NOVA Gas Transmission" and "Spectra Energy Transmission" For Spill Incidents in Developed and Remote Area. Code for Data Wragling

```
PipelineComp = Pipelinedata[,c("Year","Incident Types","Company", "Land Use")]
Renamedcol1 = rename(PipelineComp, Land = "Land Use")                              # Rename Column Name
Renamedcol = rename(Renamedcol1, Incident = "Incident Types")

#Filter Incidents only related to Year 2015 and 2018
ReqYearData = select(filter(Renamedcol, Year==2015| Year==2018),c('Year','Incident','Company','Land'))

#Filter Incidents only related to Release of Substance
ReqIncData = select(filter(ReqYearData, Incident =="Release of Substance"| Incident=="Explosion, Fire, I

#Filter Incidents only related to Companies NOVA Gas and Spectra
ReqCompdataAllArea = select(filter(ReqIncData, Company=="NOVA Gas Transmission Ltd."| Company=="Westcoas

head(ReqCompdataAllArea, 5)
```

```
##   Year             Incident
## 1 2018 Release of Substance
## 2 2018 Release of Substance
## 3 2018 Release of Substance
## 4 2018 Release of Substance
## 5 2018 Release of Substance
##                                                                  Company
## 1 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 2 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 3 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 4 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 5                                              NOVA Gas Transmission Ltd.
##                          Land
## 1 Developed Land - Industrial
## 2 Developed Land - Industrial
## 3 Developed Land - Industrial
## 4                    Forests
## 5                    Forests
```

Code to filter out data related to Companies "NOVA Gas Transmission" and "Spectra Energy Transmission" For Spill Incidents in Remote Area. Code for Data Wragling

```
PipelineComp = Pipelinedata[,c("Year","Incident Types","Company", "Land Use")]
Renamedcol1 = rename(PipelineComp, Land = "Land Use")                                  # Rename Column Name
Renamedcol = rename(Renamedcol1, Incident = "Incident Types")
Remotedata1 = filter(Renamedcol, Land!="Developed Land - Industrial")              # Remove data related
Remotedata = filter(Remotedata1, Land!="Developed Land - Residential")             # Remove data related

#Filter Incidents only related to Year 2015 and 2018
ReqYearData = select(filter(Remotedata, Year==2015| Year==2018 ),c('Year','Incident','Company','Land'))


#Filter Incidents only related to Release of Substance
ReqIncData = select(filter(ReqYearData, Incident =="Release of Substance"| Incident=="Explosion, Fire, I

#Filter Incidents only related to Companies NOVA Gas and Spectra
ReqCompdata = select(filter(ReqIncData, Company=="NOVA Gas Transmission Ltd."| Company=="Westcoast Energ

head(ReqCompdata, 5)
```

```
##   Year                              Incident
## 1 2018                  Release of Substance
## 2 2018                  Release of Substance
## 3 2018                  Release of Substance
## 4 2018 Explosion, Fire, Release of Substance
## 5 2018                  Release of Substance
##                                                                        Company
## 1 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 2                                                 NOVA Gas Transmission Ltd.
## 3                                                 NOVA Gas Transmission Ltd.
## 4 Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
## 5                                                 NOVA Gas Transmission Ltd.
##                    Land
## 1               Forests
## 2               Forests
## 3               Forests
## 4               Forests
## 5 Agricultural Cropland
```

Code for Plot to show proportions of Spill incidents for both the companies in Developed and Remote area for Year 2015 and 2018:

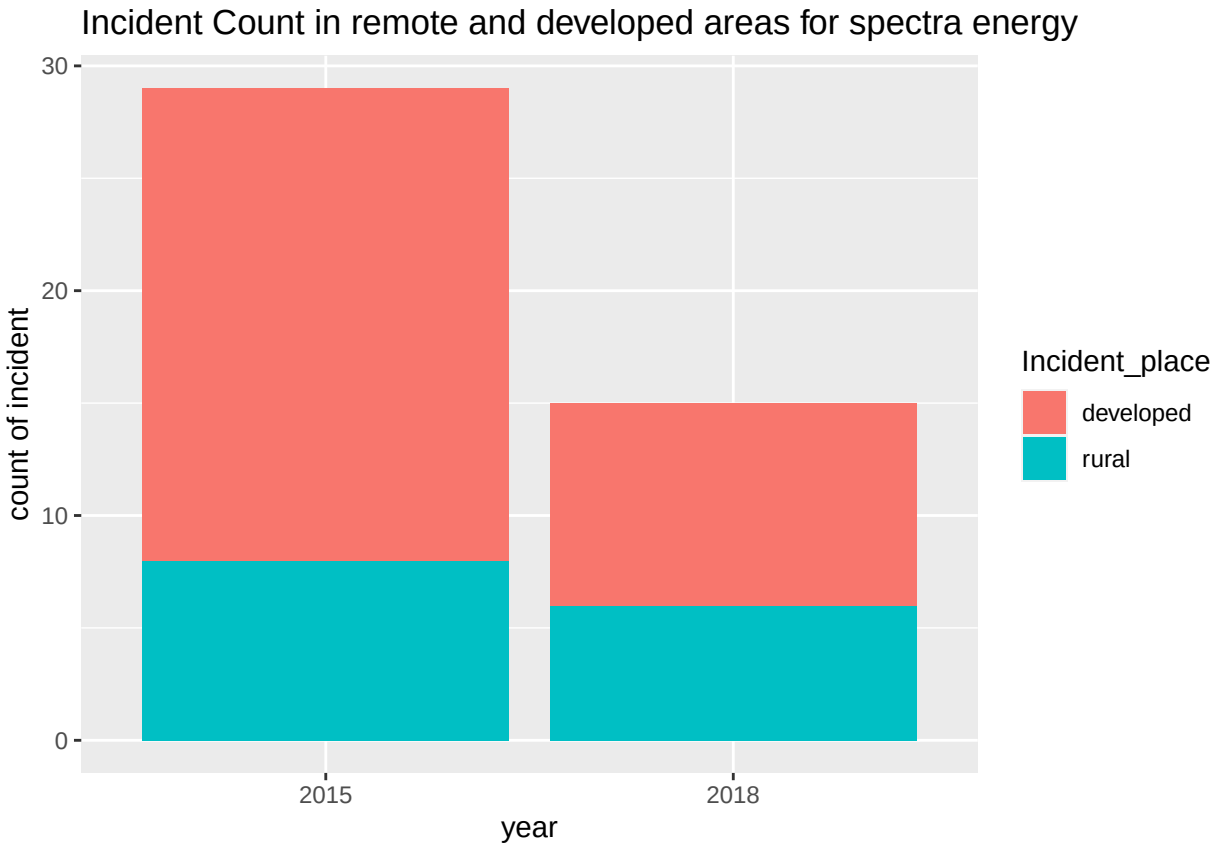
```
Ddata1 = filter(ReqCompdataAllArea, Land =="Developed Land - Industrial")              # take data relat

Ddata2 = filter(ReqCompdataAllArea, Land !="Developed Land - Industrial")              # Remove data rel

#Filter Incidents only related to Year 2015
yd_developed = select(filter(Ddata1, Year==2015),c('Year','Incident','Company','Land'))

yd_rural = select(filter(Ddata2, Year==2015),c('Year','Incident','Company','Land'))

ydNOVA = select(filter(yd_developed, Company == "NOVA Gas Transmission Ltd."),c('Year','Incident','Compa
```

```
acount = dim(ydNOVA)[1]

ydNOVA1 = select(filter(yd_rural, Company == "NOVA Gas Transmission Ltd."),c('Year','Incident','Company

acount1 = dim(ydNOVA1)[1]

ydSpec = select(filter(yd_rural, Company == "Westcoast Energy Inc., carrying on business as Spectra Ene

bcount = dim(ydSpec)[1]

ydSpec1 = select(filter(yd_developed, Company == "Westcoast Energy Inc., carrying on business as Spectra


bcount1 = dim(ydSpec1)[1]

#Filter Incidents only related to Year 2018
yd_developed1 = select(filter(Ddata1, Year==2018),c('Year','Incident','Company','Land'))

yd_rural1 = select(filter(Ddata2, Year==2018),c('Year','Incident','Company','Land'))

ydNO = select(filter(yd_developed1, Company == "NOVA Gas Transmission Ltd."),c('Year','Incident','Compa

a_count = dim(ydNO)[1]

ydNO1 = select(filter(yd_rural1, Company == "NOVA Gas Transmission Ltd."),c('Year','Incident','Company'

a_count1 = dim(ydNO1)[1]


ydSP = select(filter(yd_developed1, Company == "Westcoast Energy Inc., carrying on business as Spectra

b_count = dim(ydSP)[1]

ydSP1 = select(filter(yd_rural1, Company == "Westcoast Energy Inc., carrying on business as Spectra Ene

b_count1 = dim(ydSP1)[1]
```

Code for Plot to show proportions of Spill incidents for Spectra Energy in Developed and Remote area for Year 2015 and 2018:

```
x1 = c(rep("2015",29))
Incident_place = c(rep("rural",8), rep("developed",21))
z1 = c(rep(1,29))

xyz1 = data.frame(x1,Incident_place,z1)

x1 = c(rep("2018",15))
Incident_place= c(rep("rural",6), rep("developed",9))
z1 = c(rep(1,15))


xyz2 = data.frame(x1,Incident_place,z1)
xyz3 = rbind(xyz1,xyz2)
```

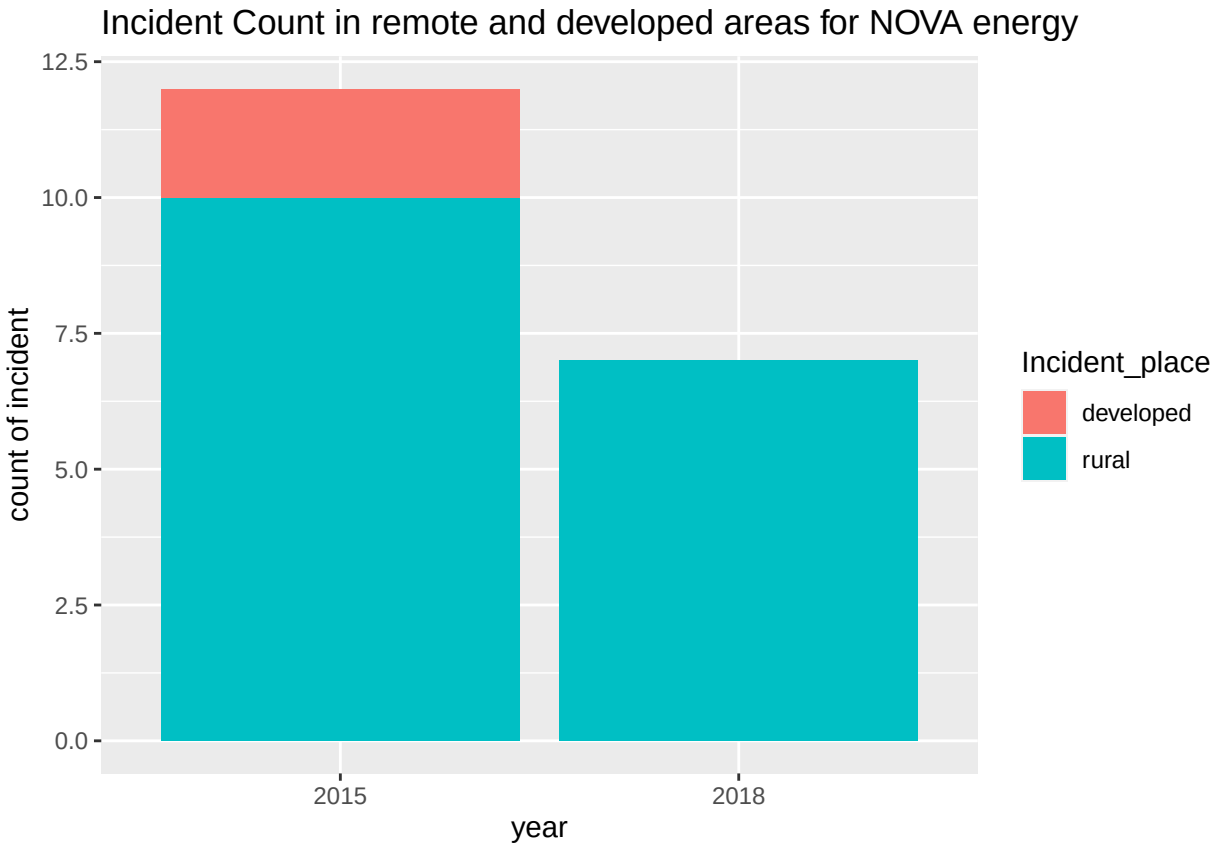
```
ggplot(data=xyz3,aes(x = x1, y=z1, fill=Incident_place)) +
    geom_bar(stat="identity", position = "stack") +
    labs(title="Incident Count in remote and developed areas for spectra energy",
         x="year", y = "count of incident")
```



Code for Plot to show proportions of Spill incidents for NOVA Energy in Developed and Remote area for Year 2015 and 2018:

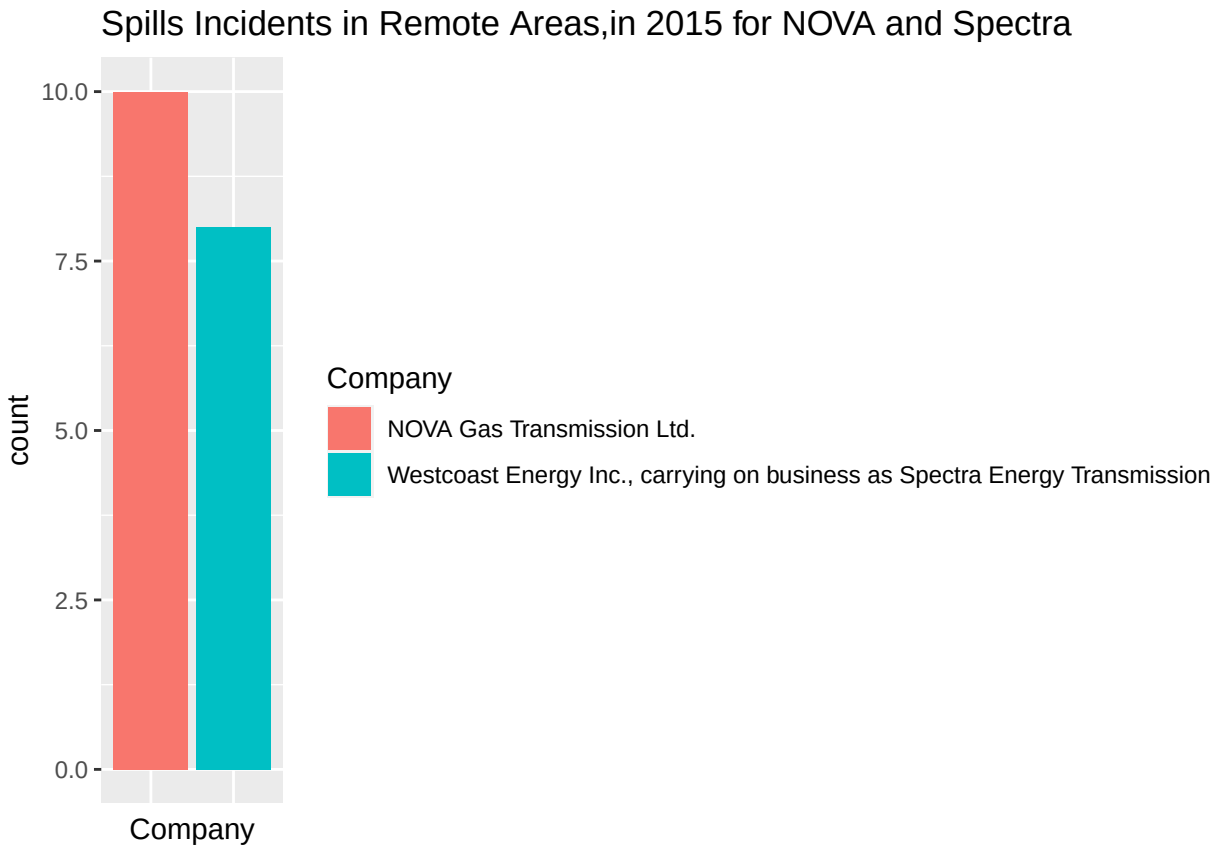
```
x11 = c(rep("2015",12))
Incident_place = c(rep("rural",10), rep("developed",2))
z11 = c(rep(1,12))

nova1 = data.frame(x11,Incident_place,z11)

x11 = c(rep("2018",7))
Incident_place= c(rep("rural",7), rep("developed",0))
z11 = c(rep(1,7))


nova2 = data.frame(x11,Incident_place,z11)
nova3 = rbind(nova1,nova2)


ggplot(data=nova3,aes(x = x11, y=z11, fill=Incident_place)) +
```

```
    geom_bar(stat="identity", position = "stack") +
    labs(title="Incident Count in remote and developed areas for NOVA energy",
         x="year", y = "count of incident")
```



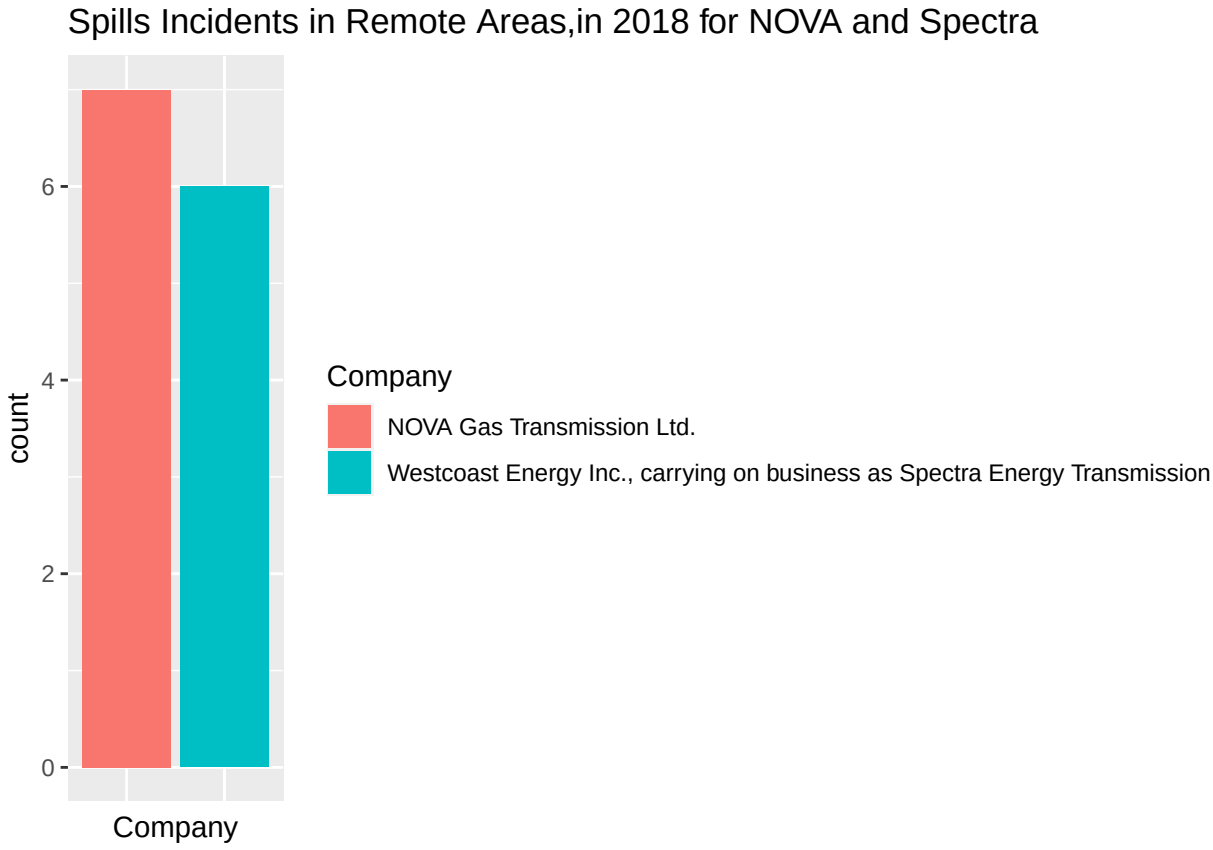
Plot to show proportions of Spill incidents in remote area for both the companies in Year 2015:

```
SpillRemotedata2015 = filter(ReqCompdata, Year == 2015)          # Filter data related to year 2015
ggplot(data=SpillRemotedata2015, aes(x = Company, fill=Company)) + geom_bar(position="dodge", na.rm=TRU
```



Plot to show proportions of Spill incidents in remote area for both the companies in Year 2018:

```
SpillRemotedata2015 = filter(ReqCompdata, Year == 2018)          # Filter data related to year 2018
ggplot(data=SpillRemotedata2015, aes(x = Company, fill=Company)) + geom_bar(position="dodge", na.rm=TRU
```



Year 2015: Lets test comparing two Population proportions i.e., $p_{Nova} = p_{Spectra}$ for Year 2015 where p_{Nova} represents proportion for Nova Gas Transmission $p_{Spectra}$ represents proportion for Spectra Energy Transmission

To test this with Z-test,we need to compute the value of the statistic that estimates the assumed "common proportion" in H_0 called as **pooled sample proportion and test statistic Z_{Obs}

$$\widehat{p} = \frac{X_{Nova} + X_{Spectra}}{n_{Nova} + n_{Spectra}} \quad and \quad Z_{obs} = \frac{\widehat{p}_{Nova} - \widehat{p}_{Spectra} - (p_{Nova} - p_{Spectra})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_{Nova}} + \frac{1}{n_{Spectra}}\right)}}$$

And we'll do above computations of the test statistic with the prop.test command in R.

```
#Total incidents reported for both companies in Remote Area in year 2015
Remotedata2015 = filter(ReqCompdataAllArea, Year == 2015)          # Filter data related to year 2015
table(Remotedata2015$Company)
```

```
##
##                                                    NOVA Gas Transmission Ltd.
##                                                                            12
## Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
##                                                                            29
```

```
#Total Spill incidents reported for both companies in Remote Area
SpillRemotedata2015 = filter(ReqCompdata, Year == 2015)          # Filter data related to year 2015
table(SpillRemotedata2015$Company)
```

```
##
##                                                 NOVA Gas Transmission Ltd.
##                                                                         10
## Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
##                                                                          8
```

Out of n = 12, we have $x_{Nova} = 10$, Spill Incidents in Remote area for Nova Gas Transmission And Out of n = 29, we have $x_{Spectra} = 8$, Spill Incidents in Remote area for Spectra Energy Transmission

We test the statistical hypotheses:

$$H_0 : p_{Nova,2015} = p_{Spectra,2015} H_A : p_{Nova,2015} \neq p_{Spectra,2015}$$

Our Null Hypotheses states that there is no difference in the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2015

The computation of the test statistic above can be completed with the prop.test command in R as below:

```
prop.test(c(10,8), c(12,29), alternative="two.sided", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c10 out of 128 out of 29
## X-squared = 10.71, df = 1, p-value = 0.001065
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2911582 0.8237843
## sample estimates:
##    prop 1    prop 2
## 0.8333333 0.2758621
```

In this R output, the test statistic is provided to be χ^2_{Obs}, this is not the value of Z_{Obs} To obtain $Z_{Obs} = \sqrt{\chi^2_1} = \sqrt{10.71} = 3.272614$

Computation of P-value using Z_{Obs} for two sided test.

```
(1-pnorm(3.272614))*2
```

```
## [1] 0.001065579
```

P-value = 0.001065579 **With the computed p-value, which comes out to be smaller than 0.05, we reject the null hypothesis, and conclude that the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2015 are not equal.**

R code to compute 95% Confidence Interval:

```
prop.test(c(10 + 1, 8 + 1), c(12 + 2, 29 + 2), conf.level=0.95, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
```

```
##
## data:  c out of c10 + 1 out of 12 + 28 + 1 out of 29 + 2
## X-squared = 9.5858, df = 1, p-value = 0.001961
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2275675 0.7632159
## sample estimates:
##    prop 1    prop 2
## 0.7857143 0.2903226
```

The 95% confidence interval for $p_{Nova} - p_{Spectra}$ has a lower bound of 0.2275675 and an upper bound of 0.7632159 Here CI is positive, it means that the proportion of Spill Incidents in remote area for Nova Gas Transmission is greater than Spectra Energy Transmission in year 2015.

Lets apply Permutation test on this data for Year 2015: Lets test our statistical hypotheses as stated below:

$$H_0 : p_{Nova,2015} - p_{Spectra,2015} = 0 H_A : p_{Nova,2015} - p_{Spectra,2015} \neq 0$$

Our Null Hypotheses states that there is no difference in the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2015

The difference between $p_{Nova} - p_{Spectra} = \frac{10}{12} - \frac{8}{29} = 0.5574713$ The question we can ask ourselves in this: is this difference between the two sample proportion indicate that proportion of Spill Incidents for Spectra Energy is more than the NOVA Gas company in 2015?

If we took all 41 cases and *randomly* of them to the 23 "NOVA Incidents" and 11 "Spectra Incidents" , there would be $\binom{41}{12}$ or $\binom{41}{29} = 7898654920$ different *permutations* of these data and the same number of differences between $\overline{p}_{NOVA} - \overline{p}_{Spectra}$. We can generate some (not all) of these differenes by randomly assigning the 34 data values to the two different groups of 23 and 11, compute the difference $\overline{p}_{NOVA} - \overline{p}_{Spectra}$ *each* time, and see where the current observed difference of 0.5574713 lies on such a distribution of differences.

If the observed difference of 0.5574713 falls in the extreme tail of such a distribution, then such an observed difference is not likely and the null hypothesis of "proportions" for NOVA & Spectra would appear not to be the case. Let's walk through the steps of conducting a **permutation test** of these data

Permutation Step 1: We pool the data together.

```
pNova = c(rep(1,10), rep(0,2))
pSpectra = c(rep(1,8), rep(0,21))
Company = c(rep("Nova Gas",12), rep("Spectra Energy",29))
Spill = c(pNova,pSpectra)
SpillCdata = data.frame(Company,Spill)
Pooleddata = SpillCdata$Spill
```

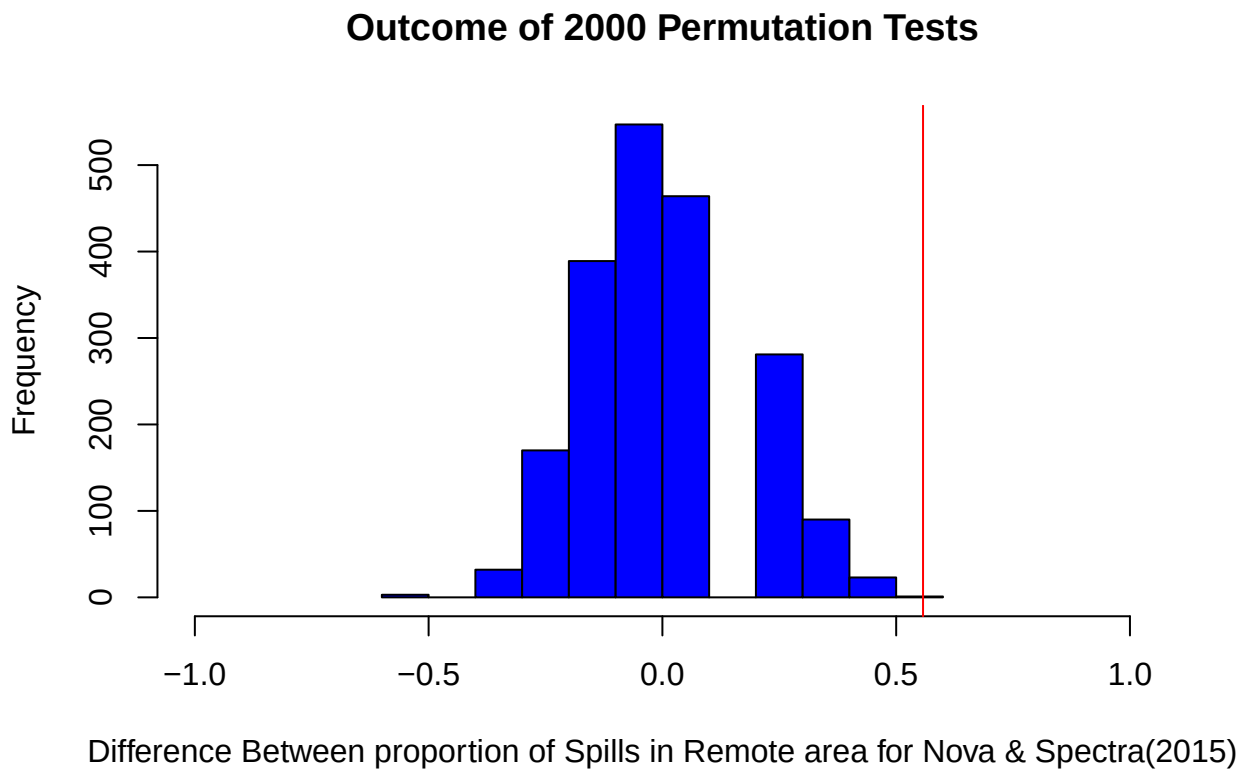
Permutation Step 2: From these data, we randomly sample $n_{Nova} = 12$ without replacement. These randomly data values will be assigned to Nova, the remaining 29 by default will be assigned to Spectra. We then compute $p_{Nova} - p_{Spectra}$, and repeat this process many times to generate a distribution of $p_{Nova} - p_{Spectra}$.

The R code that will enable this process is provided below.

```
diff = mean(~ Spill, data=filter(SpillCdata, Company=="Nova Gas")) - mean(~ Spill, data=filter(SpillCda
N = 2000                              #2000 different permutations minus the difference we have observe
outcome = numeric(N)                 #create a vector to store differences of means
for(i in 1:N)
```

```
{ index = sample(41, 12, replace=FALSE)                #randomly pick 12 nos. from 41 data
  outcome[i] = mean(Pooleddata[index]) - mean(Pooleddata[-index])       #difference between means
}
bootstrap = data.frame(outcome)           #create a data frame holding all the differences
```

```
# I used the hist command here as the outcome is a vector at this point, and I wanted to incorporate th
hist(outcome, xlim=c(-1, 1),xlab="Difference Between proportion of Spills in Remote area for Nova & Spe
abline(v = diff, col="red")
```



Code to compute P value

```
#computes P-value for two.sided test
((sum(outcome > diff)) + sum(outcome < (-1*diff)))/(N)
```

```
## [1] 0
```

The empirical P-value computed to be 0.0005. **Permutation Test, Step 3:** Because of *where* the observed difference between the samples proportions $\overline{p}_{NOVA} - \overline{p}_{Spectra} = 0.5574713$ falls from the actual, non-permutated data, occurs in the distribution of differences of sample proportions, we can see that it is not a likely outcome, implying the observed difference between these two sample proportions is smaller than expected. As a result, the null hypothesis is rejected in favour of the alternative hypothesis. We can then infer from these data that $p_{Nova,2015} \neq p_{Spectra,2015}$.

As proved above Proportions are not equal for Nova Gas and Spectra Energy. Lets test our statistical hypotheses as stated below:

$$H_0 : p_{Nova,2015} \leq p_{Spectra,2015} H_A : p_{Nova,2015} > p_{Spectra,2015}$$

Code to compute P value

```
#computes P-value for right tail test
(sum(outcome > diff))/N
```

```
## [1] 0
```

From this computed value of P for Right Tail test which is equal to 0, we reject null hypotheses and conclude that Spill Incident proportion in Remote areas are more for NOVA Gas Transmission than Spectra Energy Transmission in 2015.

Year 2018: Lets test comparing two Population proportions i.e., $p_{Nova} = p_{Spectra}$ for Year 2018 where p_{Nova} represents proportion for Nova Gas Transmission $p_{Spectra}$ represents proportion for Spectra Energy Transmission

To test this with Z-test,we need to compute the value of the statistic that estimates the assumed "common proportion" in H_0 called as **pooled sample proportion and test statistic Z_{Obs}

$$\widehat{p} = \frac{X_{Nova} + X_{Spectra}}{n_{Nova} + n_{Spectra}} \qquad and \qquad Z_{obs} = \frac{\widehat{p}_{Nova} - \widehat{p}_{Spectra} - (p_{Nova} - p_{Spectra})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_{Nova}} + \frac{1}{n_{Spectra}}\right)}}$$

And we'll do above computations of the test statistic with the prop.test command in R.

```
#Total incidents reported for both companies in Remote Area in year 2018
Remotedata2018 = filter(ReqCompdataAllArea, Year == 2018)          # Filter data related to year 2018
table(Remotedata2018$Company)
```

```
##
##                                                 NOVA Gas Transmission Ltd.
##                                                                         7
## Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
##                                                                        15
```

```
#Total Spill incidents reported for both companies in Remote Area
SpillRemotedata2018 = filter(ReqCompdata, Year == 2018)          # Filter data related to year 2018
table(SpillRemotedata2018$Company)
```

```
##
##                                                 NOVA Gas Transmission Ltd.
##                                                                         7
## Westcoast Energy Inc., carrying on business as Spectra Energy Transmission
##                                                                         6
```

Out of n = 7, we have $x_{Nova} = 7$, Spill Incidents in Remote area for Nova Gas Transmission And Out of n = 15, we have $x_{Spectra} = 6$, Spill Incidents in Remote area for Spectra Energy Transmission

We test the statistical hypotheses:

$$H_0 : p_{Nova,2018} = p_{Spectra,2018} H_A : p_{Nova,2018} \neq p_{Spectra,2018}$$

Our Null Hypotheses states that there is no difference in the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2018

The computation of the test statistic above can be completed with the prop.test command in R as below:

```r
prop.test(c(7,6), c(7,15), alternative="two.sided", correct=FALSE)
```

```
## Warning in stats::prop.test(x = x, n = n, p = p, alternative = alternative, :
## Chi-squared approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c7 out of 76 out of 15
## X-squared = 7.1077, df = 1, p-value = 0.007675
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.352082 0.847918
## sample estimates:
## prop 1 prop 2
##    1.0    0.4
```

In this R output, the test statistic is provided to be χ^2_{Obs}, this is not the value of Z_{Obs} To obtain $Z_{Obs} = \sqrt{\chi^2_1} = \sqrt{7.1077} = 2.666027$

Computation of P-value using Z_{Obs} for two sided test.

```r
(1-pnorm(2.666027))*2
```

```
## [1] 0.007675353
```

P-value = 0.007675353 **With the computed p-value, which comes out to be smaller than 0.05, we reject the null hypothesis, and conclude that the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2018 are not equal.**

R code to compute 95% Confidence Interval:

```r
prop.test(c(7+1,6+1), c(7+2,15+2), conf.level=0.95, correct=FALSE)
```

```
## Warning in stats::prop.test(x = x, n = n, p = p, alternative = alternative, :
## Chi-squared approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c7 + 1 out of 7 + 26 + 1 out of 15 + 2
## X-squared = 5.4884, df = 1, p-value = 0.01914
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1658547 0.7883937
## sample estimates:
##    prop 1    prop 2
## 0.8888889 0.4117647
```

The 95% confidence interval for $p_{Nova} - p_{Spectra}$ has a lower bound of 0.1658547 and an upper bound of 0.7883937 Here CI is positive, it means that the proportion of Spill Incidents in remote area for Nova Gas Transmission is greater than Spectra Energy Transmission in year 2018.

Lets apply Permutation test on this data for Year 2018: Lets test our statistical hypotheses as stated below:

$$H_0 : p_{Nova,2018} - p_{Spectra,2018} = 0 H_A : p_{Nova,2018} - p_{Spectra,2018} \neq 0$$

Our Null Hypotheses states that there is no difference in the proportion of Spill Incidents in remote area for Nova Gas Transmission and Spectra Energy Transmission in year 2018

The difference between $p_{Nova} - p_{Spectra} = \frac{7}{7} - \frac{6}{15} = 0.6$ The question we can ask ourselves in this: is this difference between the two sample proportion indicate that proportion of Spill Incidents for Spectra Energy is more than the NOVA Gas company in 2018?

If we took all 22 cases and *randomly* of them to the 7 "NOVA Incidents" and 15 "Spectra Incidents" , there would be $\binom{22}{7}$ or $\binom{22}{15} = 170544$ different *permutations* of these data and the same number of differences between $\overline{p}_{NOVA} - \overline{p}_{Spectra}$. We can generate some (not all) of these differenes by randomly assigning the 22 data values to the two different groups of 7 and 15, compute the difference $\overline{p}_{NOVA} - \overline{p}_{Spectra}$ *each* time, and see where the current observed difference of 0.6 lies on such a distribution of differences.

If the observed difference of 0.6 falls in the extreme tail of such a distribution, then such an observed difference is not likely and the null hypothesis of "proportions" for NOVA & Spectra would appear not to be the case. Let's walk through the steps of conducting a **permutation test** of these data **Permutation Step 1:** We pool the data together.

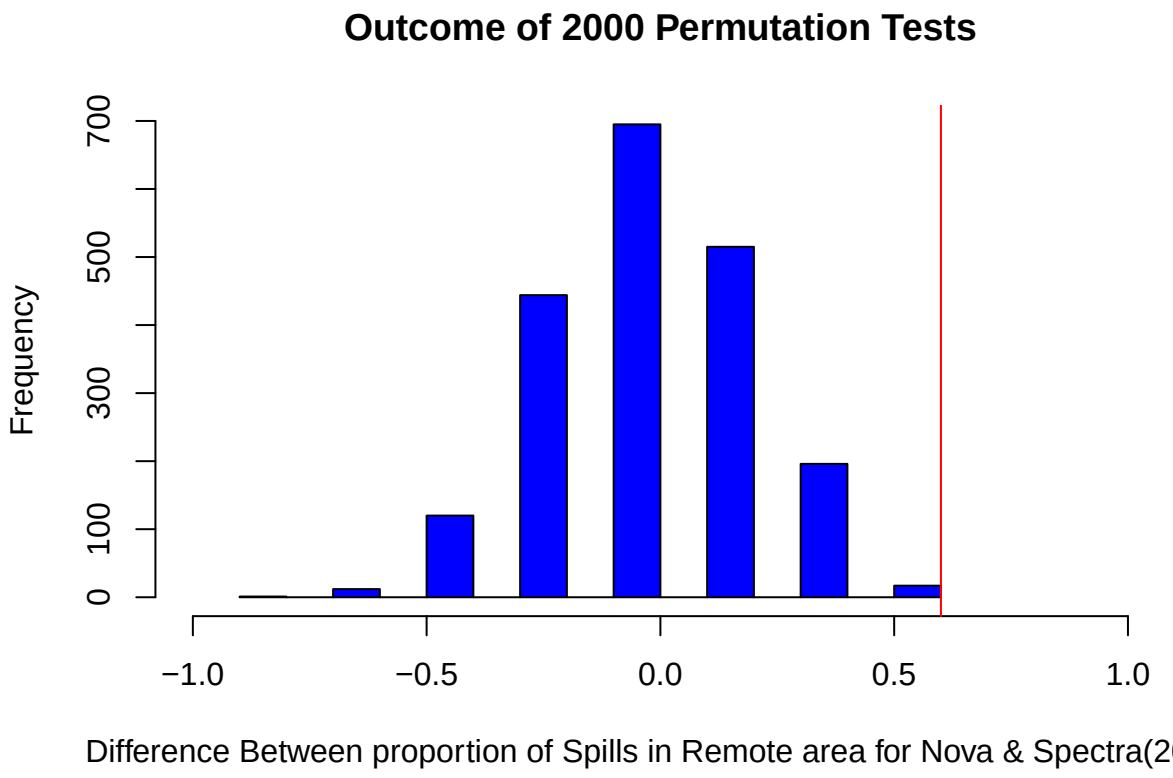
```
pNova18 = c(rep(1,7), rep(0,0))
pSpectra18 = c(rep(1,6), rep(0,9))
Company18 = c(rep("Nova Gas",7), rep("Spectra Energy",15))
Spill18 = c(pNova18,pSpectra18)
SpillCdata18 = data.frame(Company18,Spill18)

Pooleddata18 = SpillCdata18$Spill18
```

Permutation Step 2: From these data, we randomly sample $n_{Nova} = 7$ without replacement. These randomly data values will be assigned to Nova, the remaining 15 by default will be assigned to Spectra. We then compute $p_{Nova} - p_{Spectra}$, and repeat this process many times to generate a distribution of $p_{Nova} - p_{Spectra}$.

The R code that will enable this process is provided below.

```
diff18 = mean(~ Spill18, data=filter(SpillCdata18, Company18=="Nova Gas")) - mean(~ Spill18, data=filte
N = 2000                              #2000 different permutations minus the difference we have observe
outcome18 = numeric(N)                 #create a vector to store differences of means
for(i in 1:N)
{ index = sample(22, 7, replace=FALSE)               #randomly pick 12 nos. from 41 data
  outcome18[i] = mean(Pooleddata18[index]) - mean(Pooleddata18[-index])       #difference between means
}
bootstrap18 = data.frame(outcome18)          #create a data frame holding all the differences
```

```
# I used the hist command here as the outcome18 is a vector at this point, and I wanted to incorporate
hist(outcome18, xlim=c(-1, 1),xlab="Difference Between proportion of Spills in Remote area for Nova & Sp
abline(v = diff18, col="red")
```



Code to compute P value

```
#computes P-value for two.sided test
((sum(outcome18 > diff18)) + sum(outcome18 < (-1*diff18)))/(N)
```

```
## [1] 0.0055
```

The empirical P-value computed to be 0.0085. **Permutation Test, Step 3:** Because of *where* the observed difference between the samples proportions $\overline{p}_{NOVA} - \overline{p}_{Spectra} = 0.6$ falls from the actual, non-permutated data, occurs in the distribution of differences of sample proportions, we can see that it is not a likely outcome, implying the observed difference between these two sample proportions is smaller than expected. As a result, the null hypothesis is rejected in favour of the alternative hypothesis. We can then infer from these data that $p_{Nova,2018} \neq p_{Spectra,2018}$.

As proved above Proportions are not equal for Nova Gas and Spectra Energy. Lets test our statistical hypotheses as stated below:

$$H_0 : p_{Nova,2018} \leq p_{Spectra,2018} H_A : p_{Nova,2018} > p_{Spectra,2018}$$

Code to compute P value

```
#computes P-value for right tail test
(sum(outcome18 > diff18))/N
```

```
## [1] 0
```

From this computed value of P for Right Tail test which is equal to 0, we reject null hypotheses and conclude that Spill Incident proportion in Remote areas are more for NOVA Gas Transmission than Spectra Energy Transmission in 2018.

APPENDIX B TOPIC 2 R CODE AND ANALYSIS

602 Project

```
library("readxl")
```

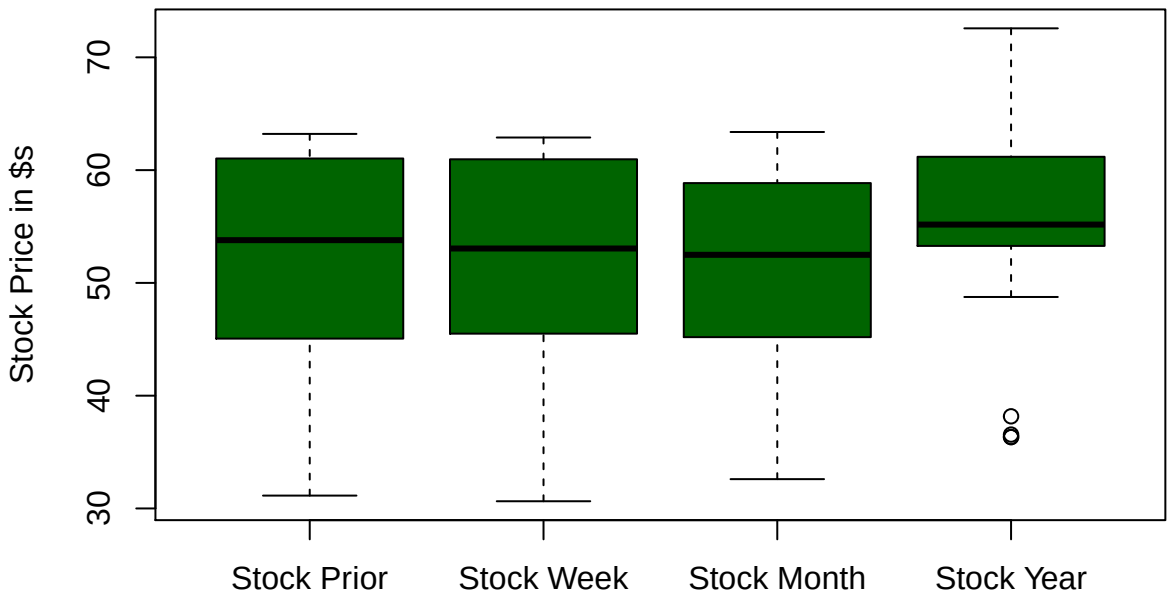
```
stock.df <- read_excel("C:/Users/Alex/Desktop/Data 602/Project/Nova Stock Info.xlsx")
stock.df
```

```
## # A tibble: 30 x 11
##    DATE  COMPANY `RELEASED SUBST~ `Approximate Vo~ Stock.Prior Stock.Week
##    <chr> <chr>   <chr>                      <dbl>       <dbl>      <dbl>
##  1 1/16~ NOVA G~ Natural Gas - S~            2866         54.6       54.8
##  2 11/2~ NOVA G~ Natural Gas - S~            8214.        53.8       53.1
##  3 10/6~ NOVA G~ Natural Gas - S~           13944         53.8       51.3
##  4 4/14~ NOVA G~ Natural Gas - S~           50622.        53.4       55.3
##  5 1/2/~ NOVA G~ Natural Gas - S~           17260         56.6       53.0
##  6 11/1~ NOVA G~ Natural Gas - S~          265457         59.6       57.2
##  7 6/12~ NOVA G~ Natural Gas - S~            5125         61.9       61.6
##  8 6/12~ NOVA G~ Natural Gas - S~            7552         61.9       61.6
##  9 11/2~ NOVA G~ Natural Gas - S~           18337         62.5       61.7
## 10 10/2~ NOVA G~ Natural Gas - S~            6474         61.9       61.2
## # ... with 20 more rows, and 5 more variables: Stock.Month <dbl>,
## #   Stock.Year <dbl>, Diff.Week <dbl>, Diff.Month <dbl>, Diff.Year <dbl>
```

Nova Gas Transmission is a subsidiary of TC Energy. Data above contains stock price for TC Energy one day prior to major gas release events and 1 week, month, and year afterwards. Our goal is to determine if release events have short and/or long term effects on a company's stock price.

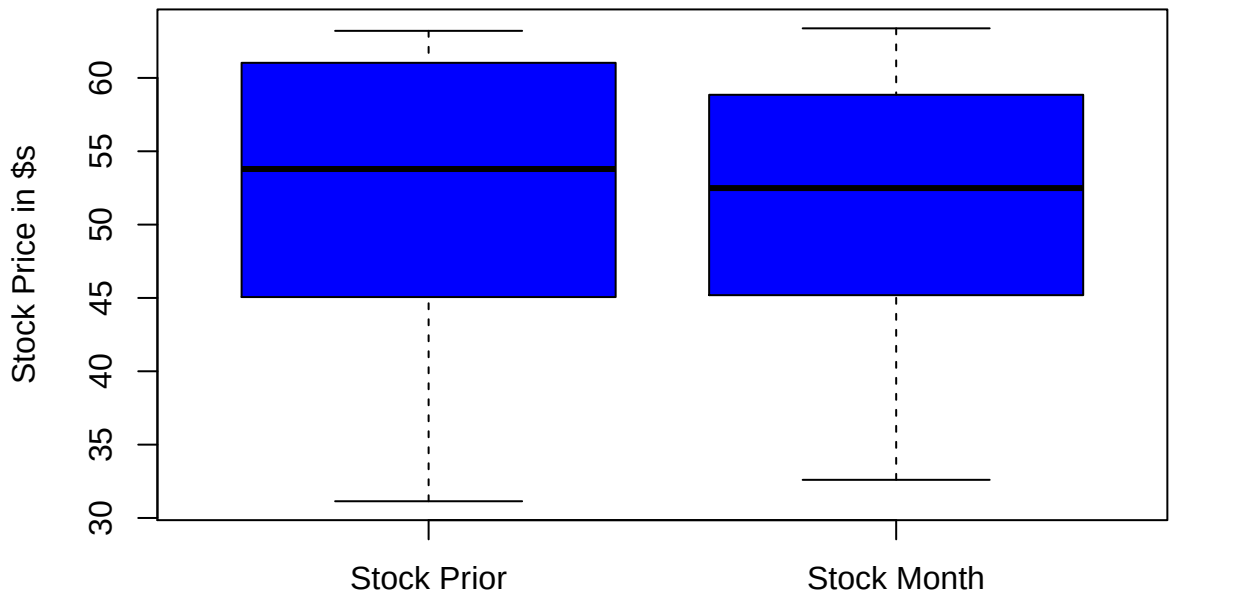
$\mathrm{H}_0 : \mu_{Week} - \mu_{Prior} \geq 0$ $\mathrm{H}_A : \mu_{Week} - \mu_{Prior} < 0$

```
boxplot(stock.df$Stock.Prior, stock.df$Stock.Week, stock.df$Stock.Month, stock.df$Stock.Year, names=c("$
```



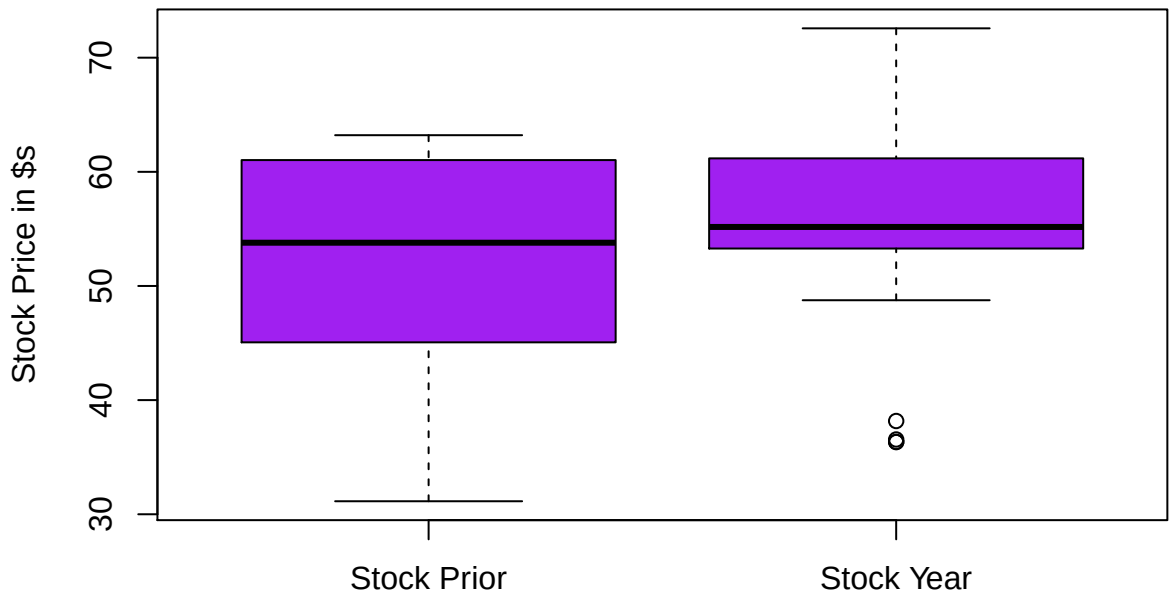
$H_0 : \mu_{Month} - \mu_{Prior} \geq 0$ $H_A : \mu_{Month} - \mu_{Prior} < 0$

```
boxplot(stock.df$Stock.Prior, stock.df$Stock.Month, names=c("Stock Prior", "Stock Month"), ylab="Stock
```

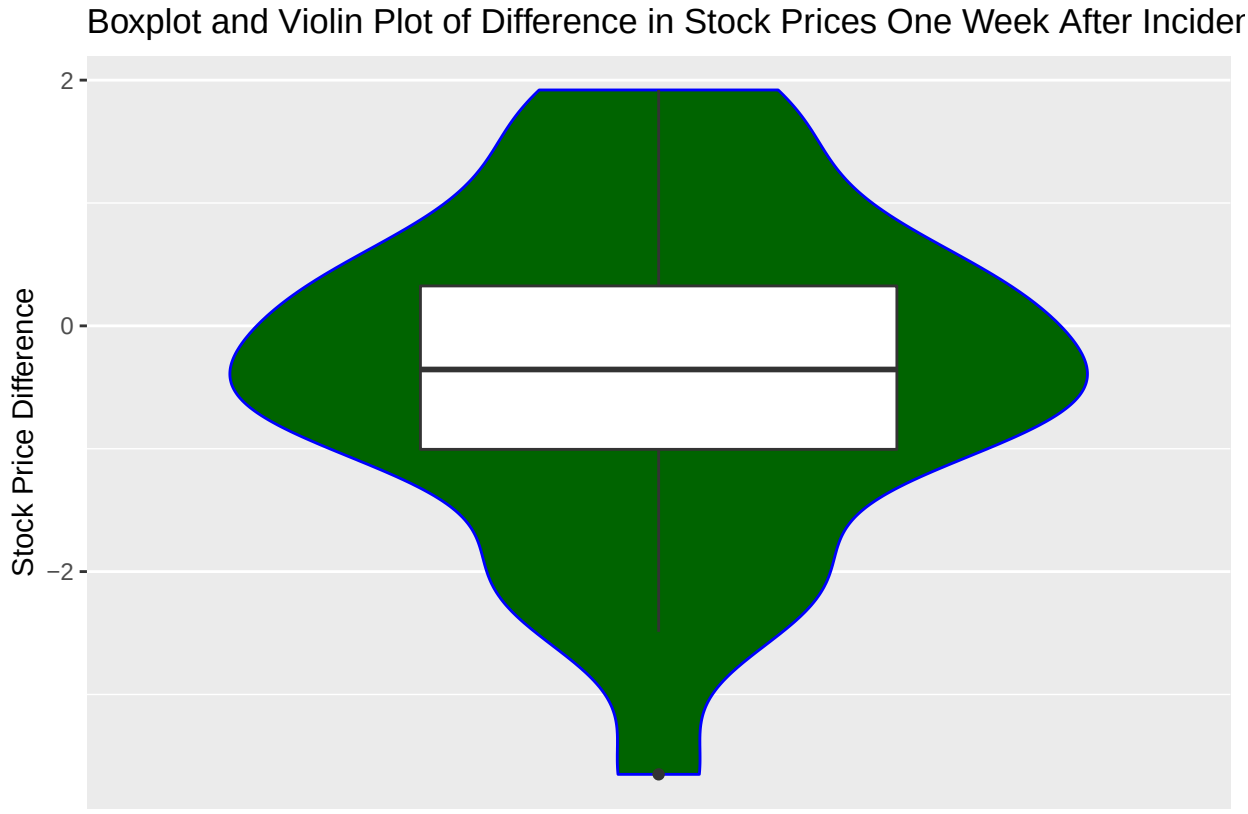


$H_0 : \mu_{Year} - \mu_{Prior} \geq 0$ $H_A : \mu_{Year} - \mu_{Prior} < 0$

```
boxplot(stock.df$Stock.Prior, stock.df$Stock.Year, names=c("Stock Prior", "Stock Year"), ylab="Stock Pr
```




```
ggplot(data=stock.df, aes(x = 'var', y = Diff.Week)) + geom_violin(col='blue', fill= 'dark green') + ge
```



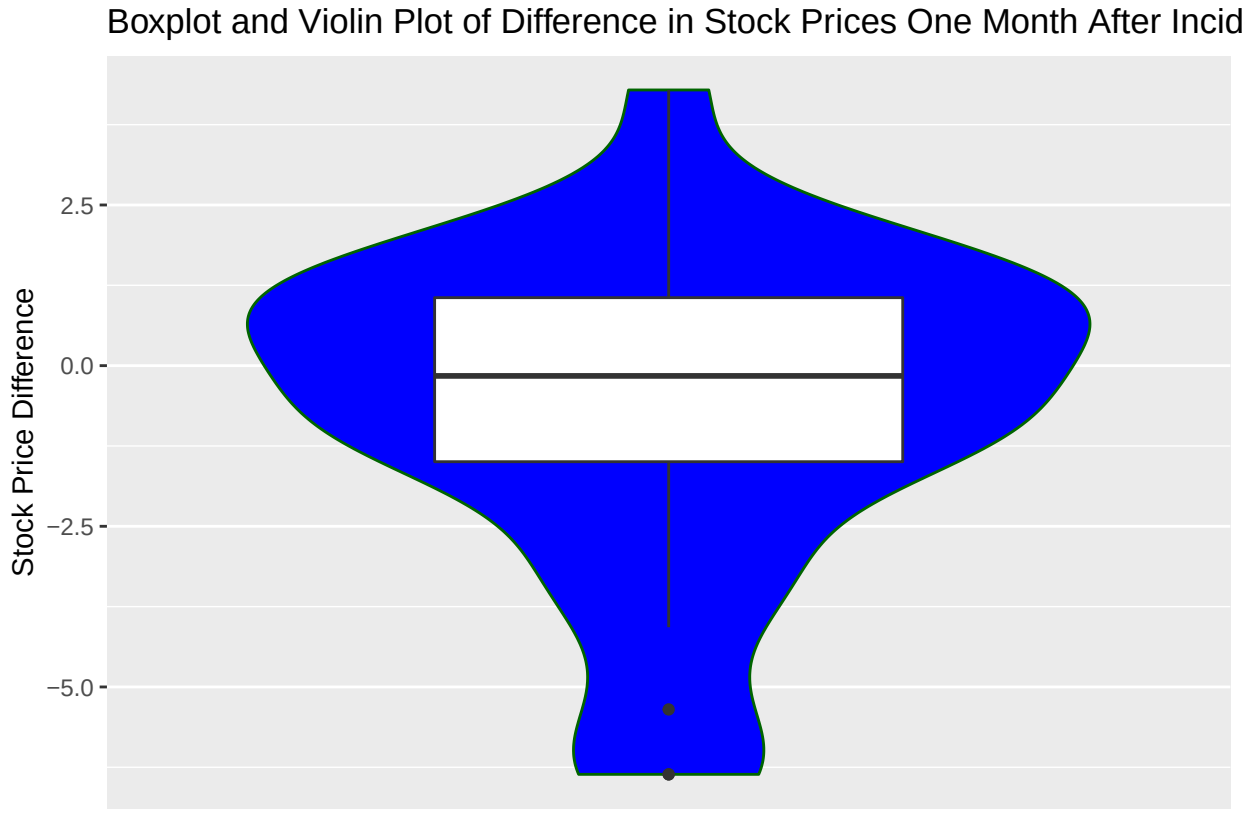

```
mean(stock.df$Diff.Week)
```

```
## [1] -0.3866667
```

```
sd(stock.df$Diff.Week)
```

```
## [1] 1.312232
```

```
ggplot(data=stock.df, aes(x = 'var', y = Diff.Month)) + geom_violin(col='dark green', fill= 'blue') + g
```



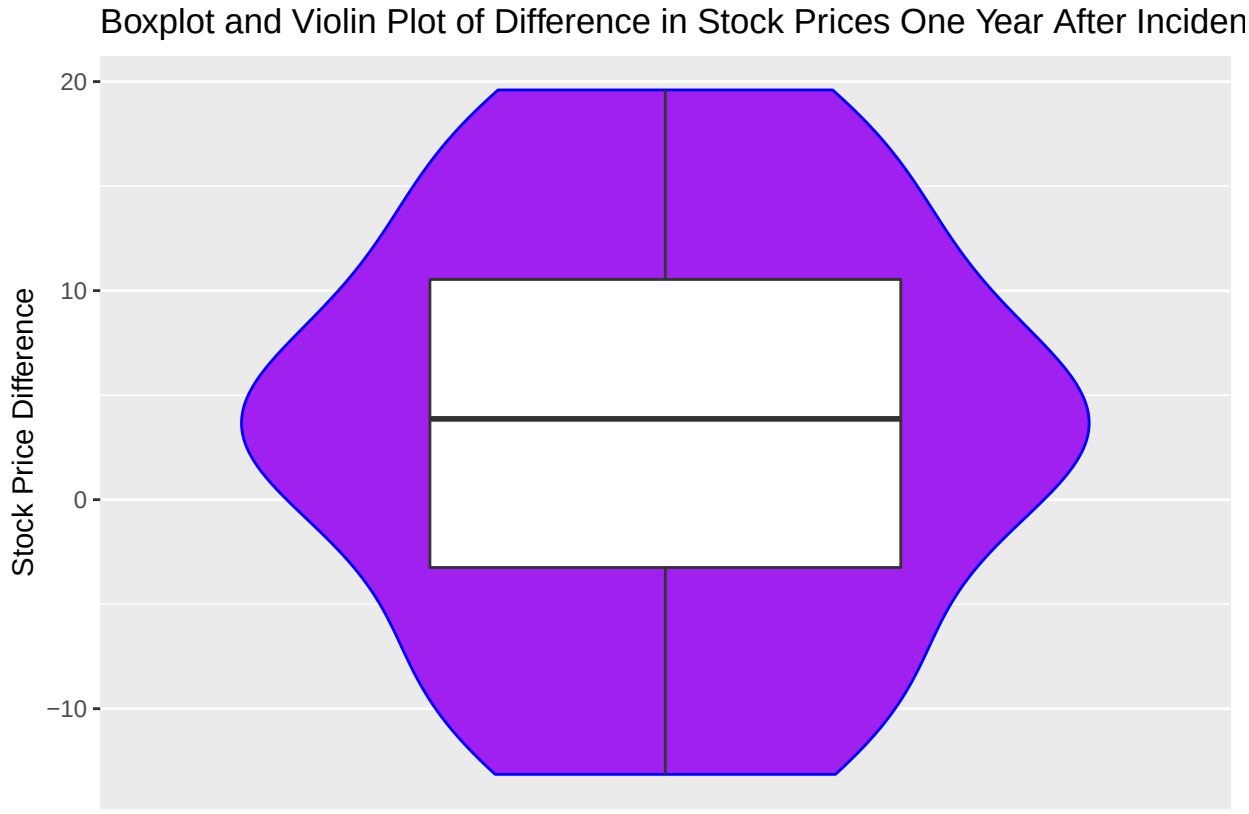

```
mean(stock.df$Diff.Month)
```

```
## [1] -0.669
```

```
sd(stock.df$Diff.Month)
```

```
## [1] 2.54224
```

```
ggplot(data=stock.df, aes(x = 'var', y = Diff.Year)) + geom_violin(col='blue', fill= 'purple') + geom_b
```



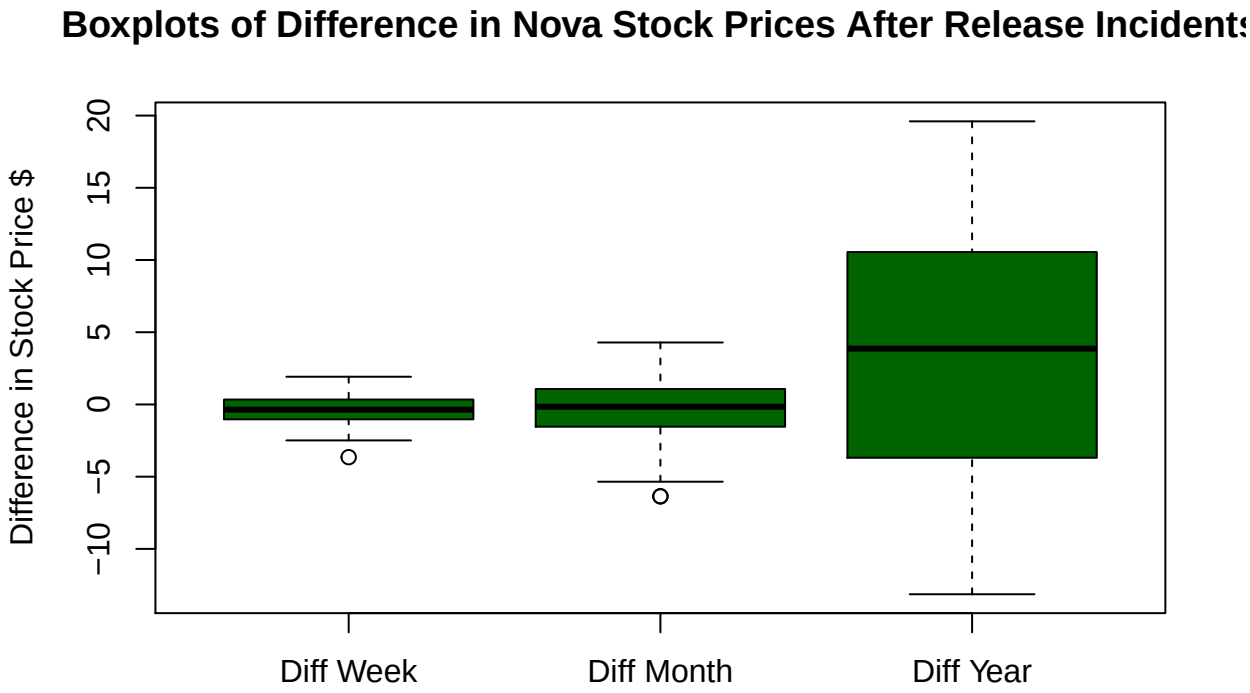

```
mean(stock.df$Diff.Year)
```

```
## [1] 3.242333
```

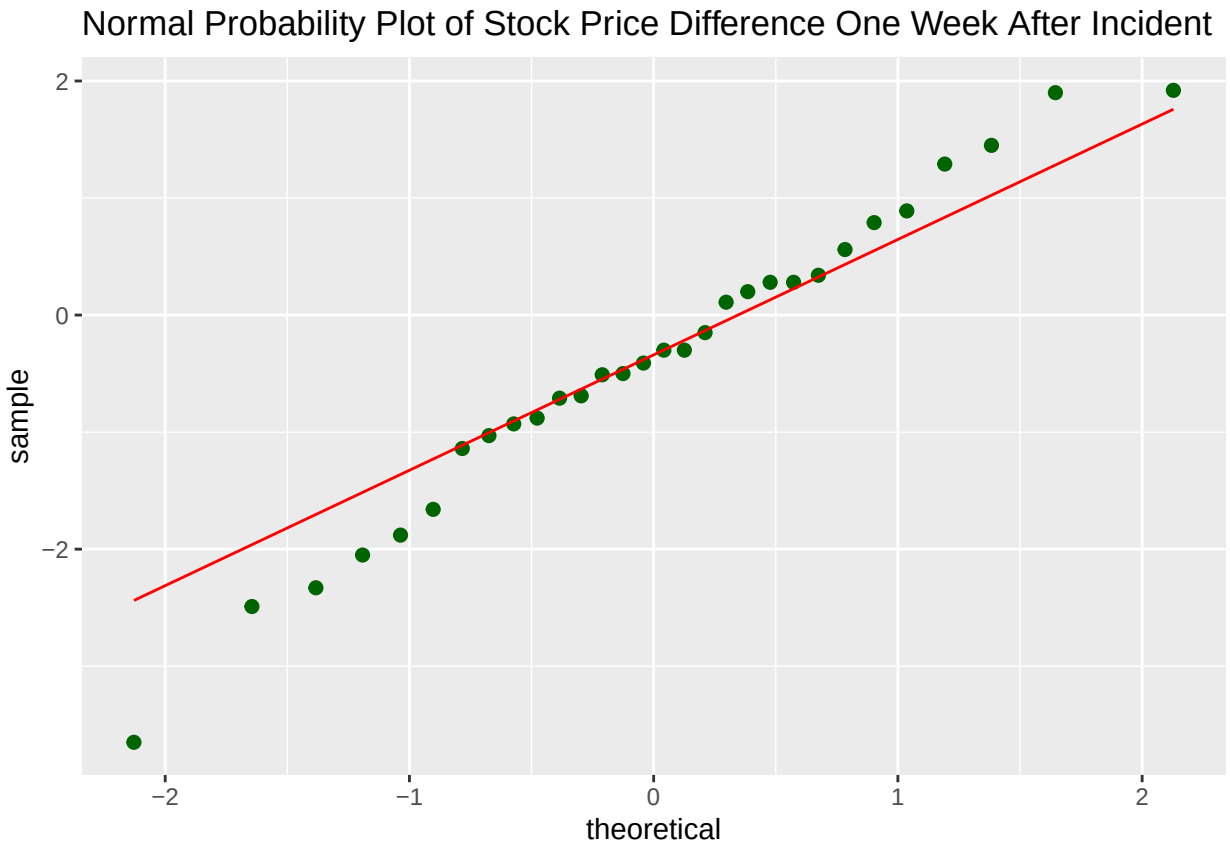
```
sd(stock.df$Diff.Year)
```

```
## [1] 9.596301
```

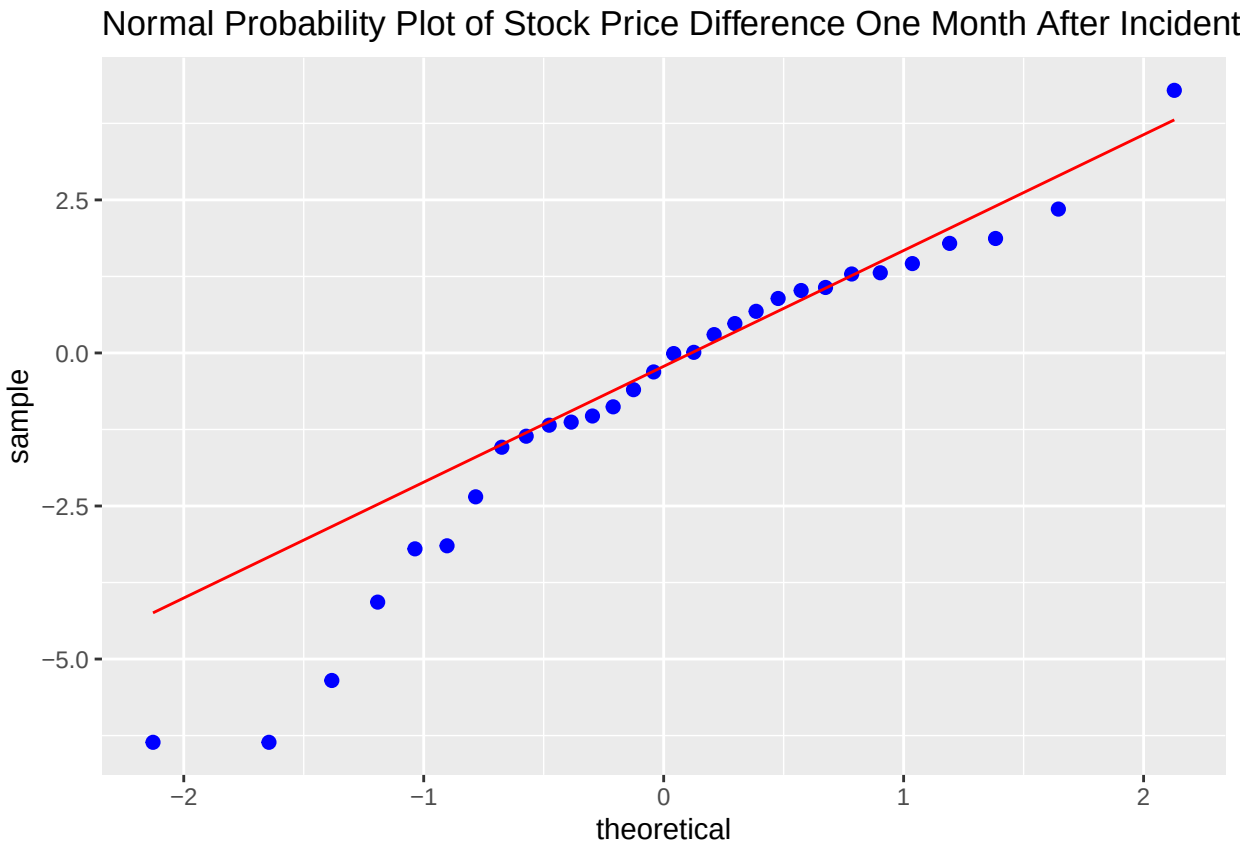
```
boxplot(stock.df$Diff.Week, stock.df$Diff.Month, stock.df$Diff.Year, names=c("Diff Week", "Diff Month",
```



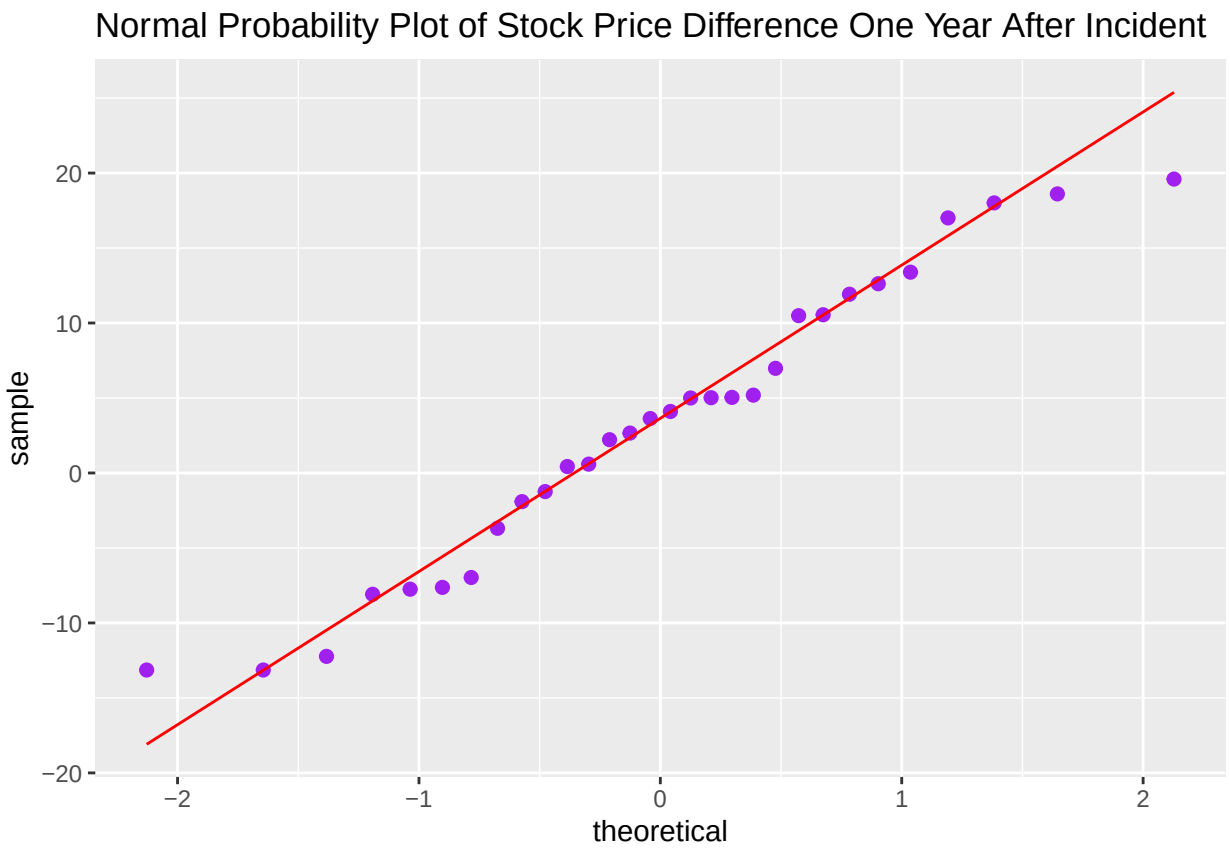

```
ggplot(data=stock.df, aes(sample = Diff.Week)) + stat_qq(size=2, col='dark green') + stat_qq_line(col='
```




```
ggplot(data=stock.df, aes(sample = Diff.Month)) + stat_qq(size=2, col='blue') + stat_qq_line(col='red')
```




```
ggplot(data=stock.df, aes(sample = Diff.Year)) + stat_qq(size=2, col='purple') + stat_qq_line(col='red'
```




```
# One Week After

mean.diffweek = mean(~Diff.Week, data = stock.df)
sd.diffweek = sd(~Diff.Week, data = stock.df)
n.diff = 30

tobs.diffweek = (mean.diffweek- 0)/(sd.diffweek/sqrt(n.diff))

tobs.diffweek
```

```
## [1] -1.613937
```

```
pt(tobs.diffweek, n.diff - 1)
```

```
## [1] 0.0586856
```

```
# One Month After

mean.diffmonth = mean(~Diff.Month, data = stock.df)
sd.diffmonth = sd(~Diff.Month, data = stock.df)
n.diff = 30

tobs.diffmonth = (mean.diffmonth- 0)/(sd.diffmonth/sqrt(n.diff))

tobs.diffmonth
```

```
## [1] -1.441353
```

```
pt(tobs.diffmonth, n.diff - 1)
```

```
## [1] 0.08009849
```

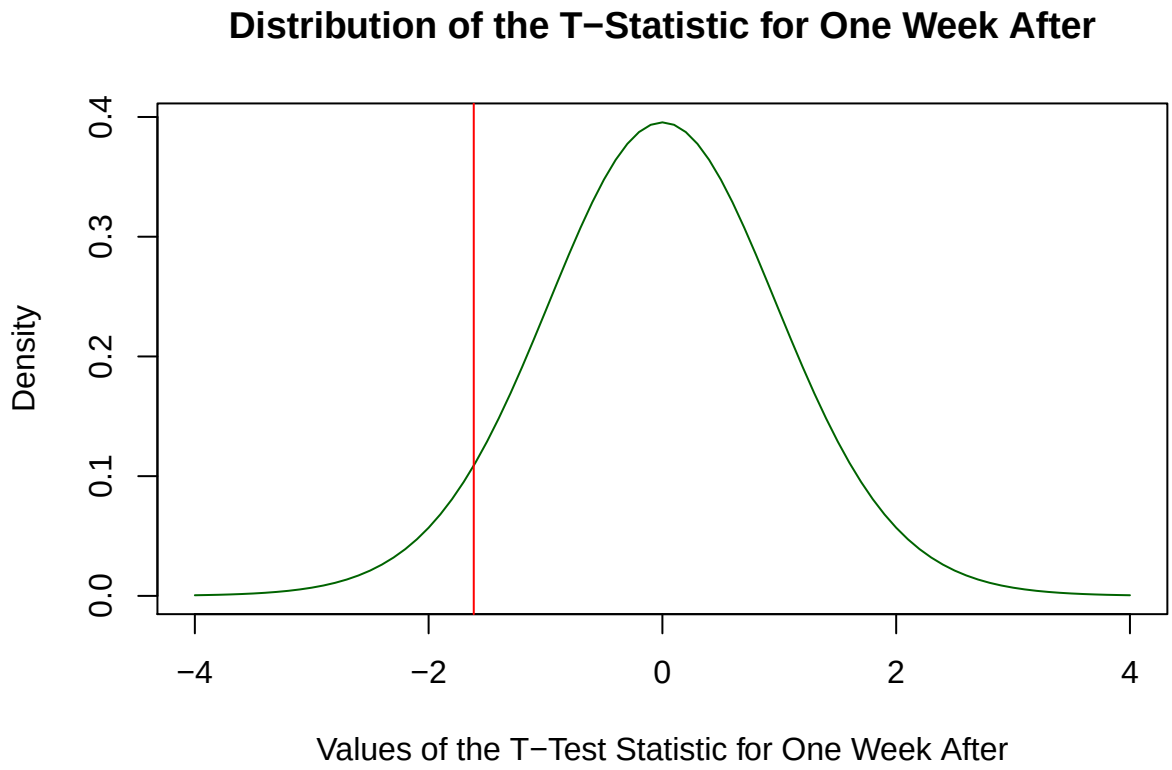
```
# One Year After

mean.diffyear = mean(~Diff.Year, data = stock.df)
sd.diffyear = sd(~Diff.Year, data = stock.df)
n.diff = 30

tobs.diffyear = (mean.diffyear- 0)/(sd.diffyear/sqrt(n.diff))

tobs.diffyear
```

```
## [1] 1.850608
```

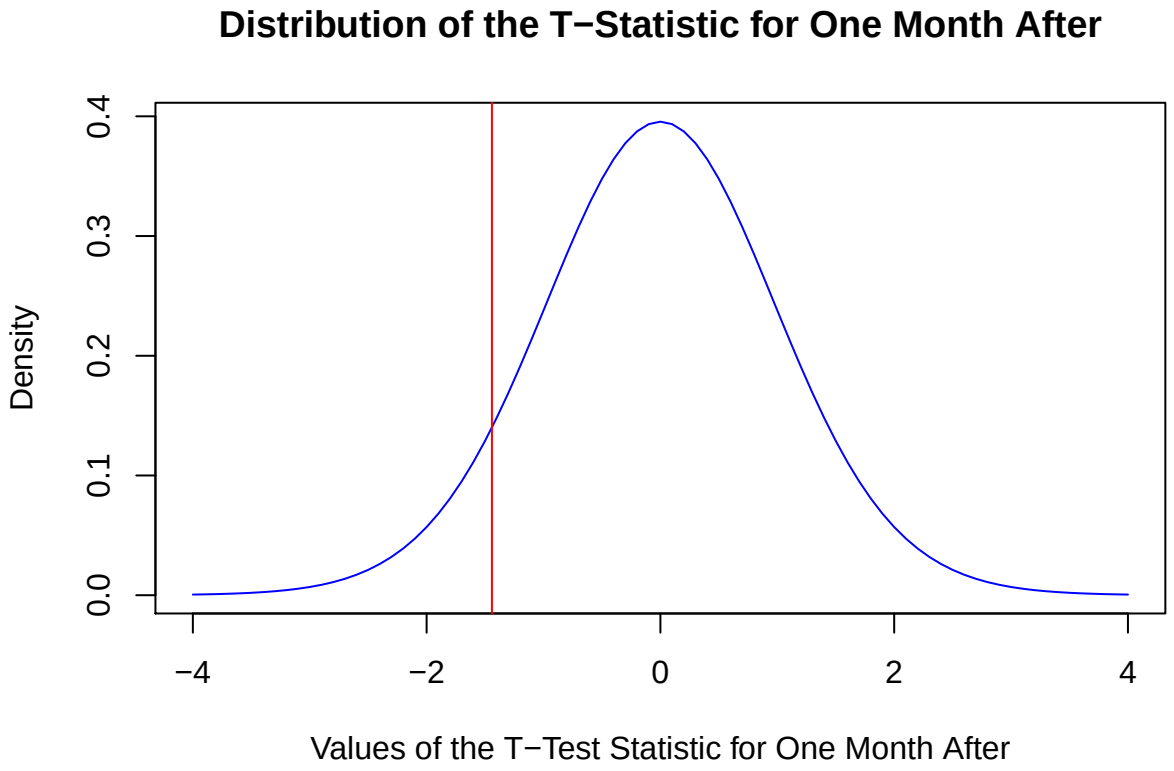
```
pt(tobs.diffyear, n.diff - 1)
```

```
## [1] 0.9627813
```

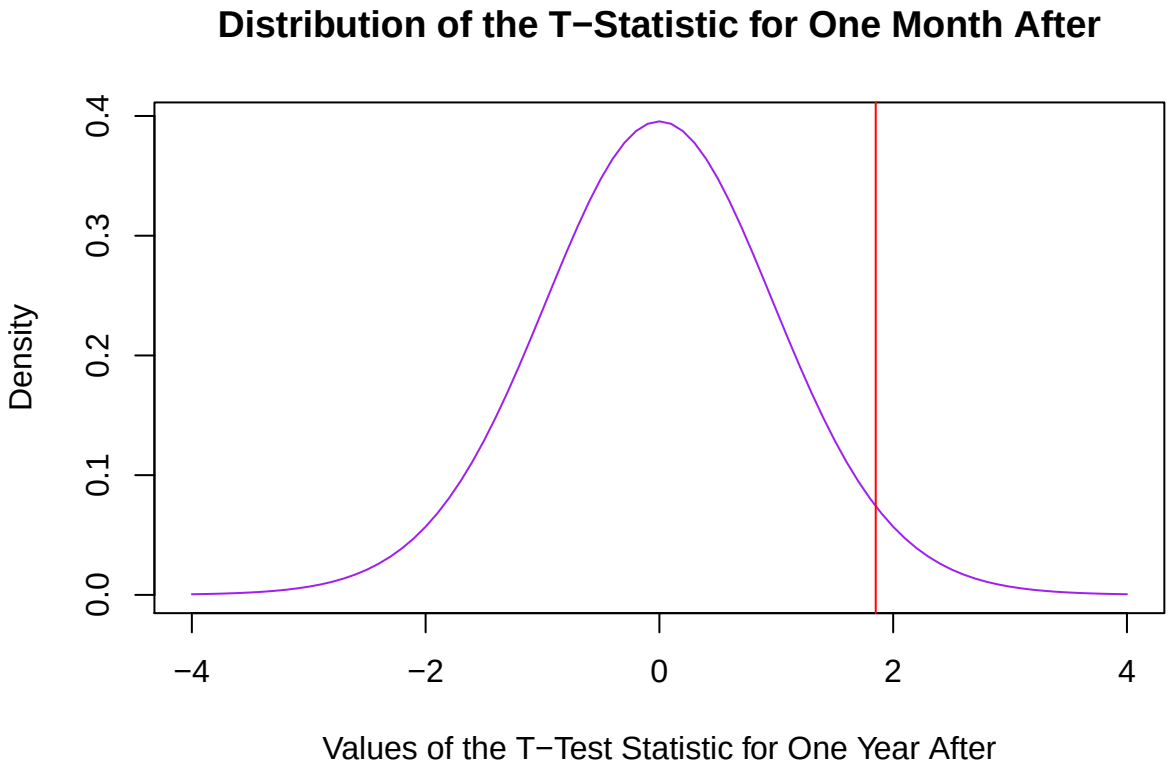
```
tvalues = seq(-4, 4, 0.1)
plot(tvalues, dt(tvalues, n.diff-1), xlab="Values of the T-Test Statistic for One Week After", ylab="Den
abline(v=tobs.diffweek, col="red")
```



```
tvalues = seq(-4, 4, 0.1)
plot(tvalues, dt(tvalues, n.diff-1), xlab="Values of the T-Test Statistic for One Month After", ylab="D
abline(v=tobs.diffmonth, col="red")
```




```
tvalues = seq(-4, 4, 0.1)
plot(tvalues, dt(tvalues, n.diff-1), xlab="Values of the T-Test Statistic for One Year After", ylab="Den
abline(v=tobs.diffyear, col="red")
```




```
# One Week After

t.test(~Diff.Week, mu=0, alternative="less", data=stock.df)
```

```
##
##  One Sample t-test
##
## data:  Diff.Week
## t = -1.6139, df = 29, p-value = 0.05869
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf 0.0204098
## sample estimates:
##  mean of x
## -0.3866667
```

```
t.test(~Diff.Week, data=stock.df)$conf
```

```
## [1] -0.8766623  0.1033290
## attr(,"conf.level")
## [1] 0.95
```

```
# One Month After

t.test(~Diff.Month, mu=0, alternative="less", data=stock.df)
```

```
##
##  One Sample t-test
##
## data:  Diff.Month
## t = -1.4414, df = 29, p-value = 0.0801
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf 0.1196452
## sample estimates:
## mean of x
##    -0.669
```

```
t.test(~Diff.Month, data=stock.df)$conf
```

```
## [1] -1.6182878  0.2802878
## attr(,"conf.level")
## [1] 0.95
```

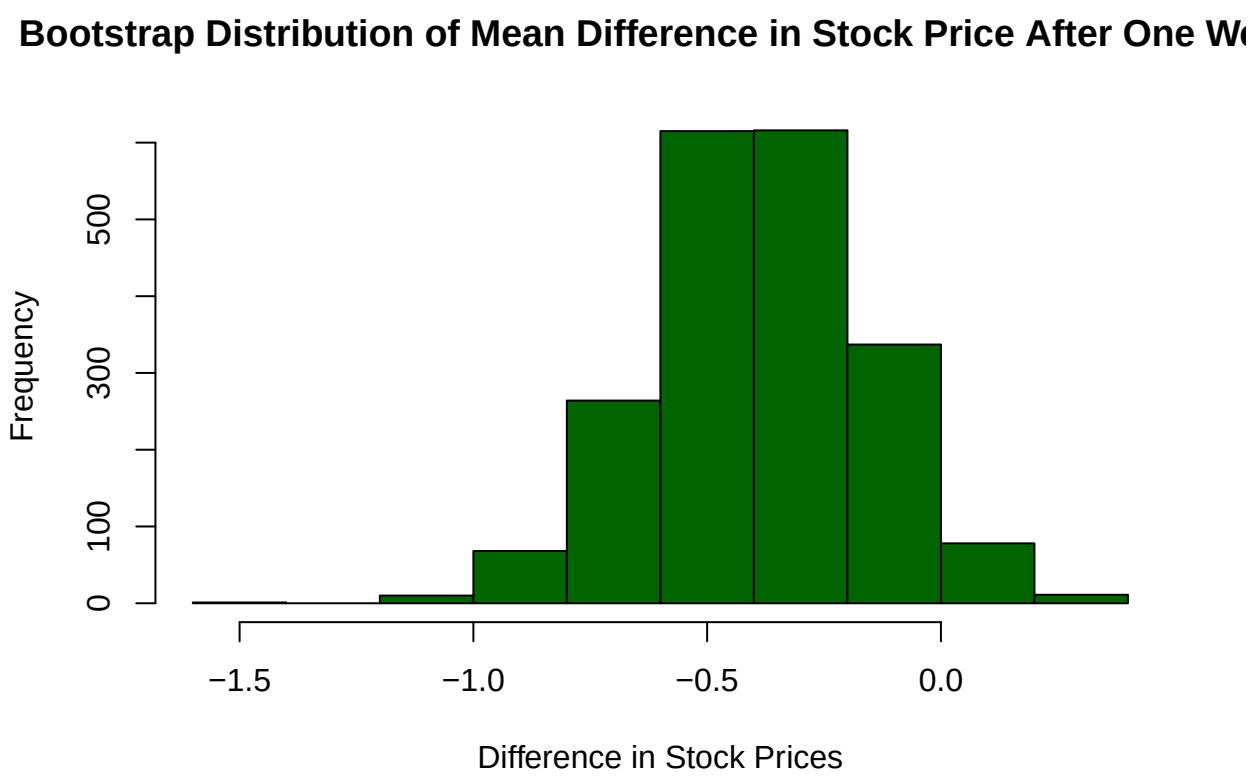
```
# One Year After

t.test(~Diff.Year, mu=0, alternative="less", data=stock.df)
```

```
##
##  One Sample t-test
##
## data:  Diff.Year
## t = 1.8506, df = 29, p-value = 0.9628
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 6.219266
## sample estimates:
## mean of x
##  3.242333
```

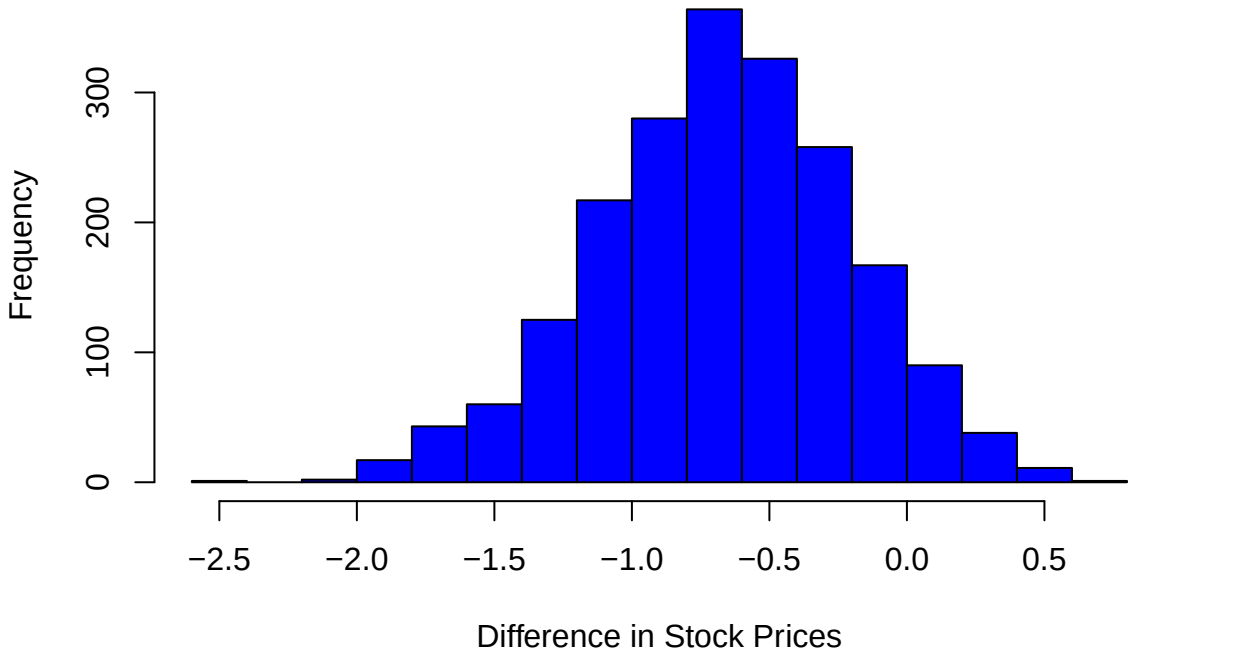
```
t.test(~Diff.Year, data=stock.df)$conf
```

```
## [1] -0.3409843  6.8256510
## attr(,"conf.level")
## [1] 0.95
```

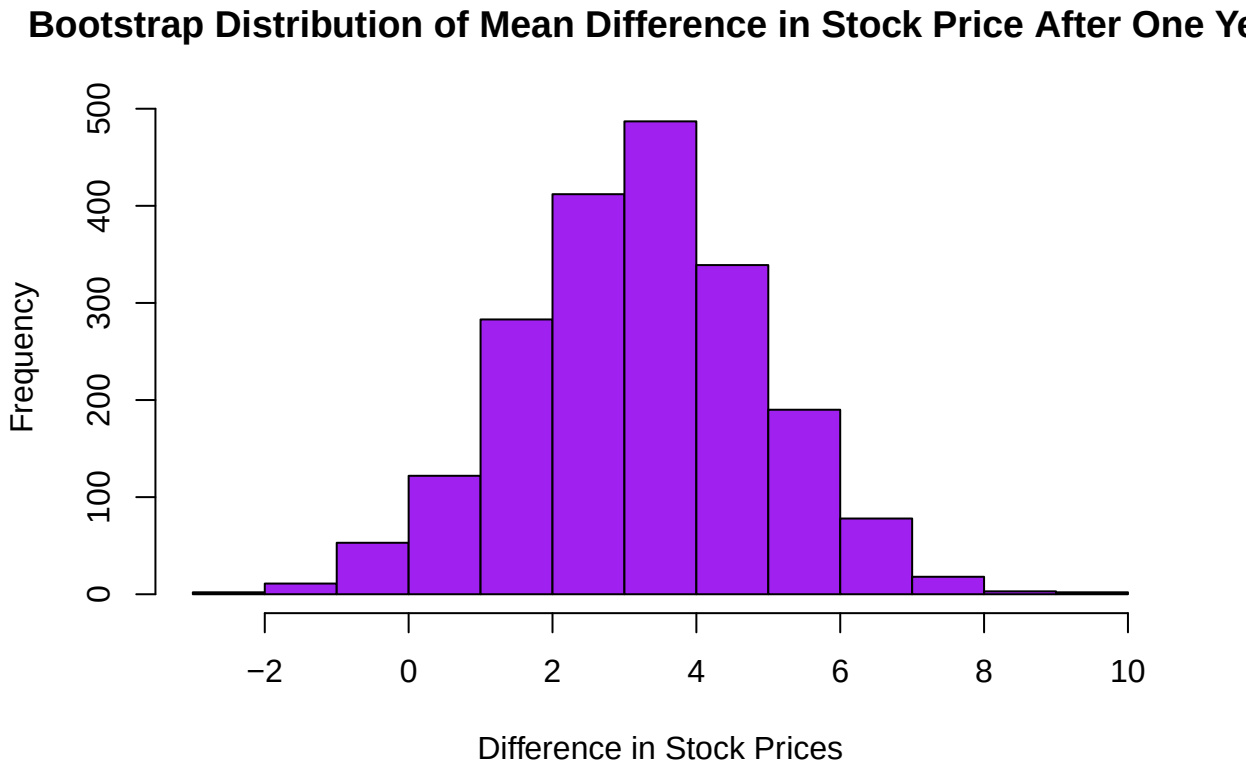
```
nsims=2000

diffmean.week = numeric(nsims)

for(i in 1:nsims)
{
  diffmean.week[i] = mean(sample(stock.df$Diff.Week, n.diff, replace=TRUE))
}

hist(diffmean.week, xlab="Difference in Stock Prices", col="dark green", main="Bootstrap Distribution o
```




```
nsims=2000

diffmean.month = numeric(nsims)

for(i in 1:nsims)
{
  diffmean.month[i] = mean(sample(stock.df$Diff.Month, n.diff, replace=TRUE))
}

hist(diffmean.month, xlab="Difference in Stock Prices", col="blue", main="Bootstrap Distribution of Mea
```




```
nsims=2000

diffmean.year = numeric(nsims)

for(i in 1:nsims)
{
  diffmean.year[i] = mean(sample(stock.df$Diff.Year, n.diff, replace=TRUE))
}

hist(diffmean.year, xlab="Difference in Stock Prices", col="purple", main="Bootstrap Distribution of Mea
```




```
qdata(diffmean.week, c(0.025, 0.975))
```

```
##        2.5%       97.5%
## -0.83811667  0.06579167
```

```
qdata(diffmean.month, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -1.629683  0.192550
```

```
qdata(diffmean.year, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -0.1777333  6.5356917
```

APPENDIX C TOPIC 3 R CODE AND ANALYSIS

R Notebook

Code

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

Hide

```
Pipelinedata = read.csv("ncdntcmprhnsv-eng.csv",fileEncoding = "Latin1", check.name
s = F)
PipelineComp = Pipelinedata[,c("Year","Incident Types","Company", "Land Use")]
# Fiter required columns from Pipelinedata
head(PipelineComp,5)
```

	Year <int>	Incident Types <fctr>
1	2019	Operation Beyond Design Limits
2	2019	Fire
3	2019	Operation Beyond Design Limits
4	2019	Fire, Release of Substance
5	2019	Operation Beyond Design Limits

5 rows | 1-3 of 4 columns

Hide

```
Renamedcol1 = rename(PipelineComp, c("Land Use" = "Land"))        # Rename Column N
ame
Renamedcol = rename(Renamedcol1, c("Incident Types" = "Incident"))



#Filter Incidents only related to Release of Substance
ReqIncData = select(filter(Renamedcol, Incident =="Release of Substance"| Incident=
="Explosion, Fire, Release of Substance" | Incident=="Release of Substance, Adverse
Environmental Effects" | Incident=="Fire, Release of Substance" | Incident=="Seriou
s Injury (NEB or TSB), Fire, Release of Substance"| Incident=="Release of Substance
, Operation Beyond Design Limits" | Incident=="Serious Injury (NEB or TSB), Release
of Substance"),c('Year','Incident','Company','Land'))

#Filter Incidents only related to Companies NOVA Gas
ReqCompdataAllArea = select(filter(ReqIncData, Company=="NOVA Gas Transmission Ltd.
"| Company=="Westcoast Energy Inc., carrying on business as Spectra Energy Transmis
sion"),c('Year','Incident','Company','Land'))

ReqCompdataAllArea = select(filter(ReqIncData, Company=="NOVA Gas Transmission Ltd.
"),c('Year','Incident','Company','Land'))

head(ReqCompdataAllArea, 5)
```

	Y…	Incident	Company	Land
	<int>	<fctr>	<fctr>	<fctr>
1	2019	Fire, Release of Substance	NOVA Gas Transmission Ltd.	Vegetative Barren
2	2019	Release of Substance	NOVA Gas Transmission Ltd.	Agricultural Cropland
3	2019	Release of Substance	NOVA Gas Transmission Ltd.	Forests
4	2018	Release of Substance	NOVA Gas Transmission Ltd.	Forests
5	2018	Release of Substance	NOVA Gas Transmission Ltd.	Forests

5 rows

Hide

```
tail(ReqCompdataAllArea, 5)
```

	Year	Incident	Company
	<int>	<fctr>	<fctr>
143	2009	Fire, Release of Substance	NOVA Gas Transmission Ltd.
144	2009	Release of Substance	NOVA Gas Transmission Ltd.

145	2009	Release of Substance	NOVA Gas Transmission Ltd.
146	2009	Release of Substance	NOVA Gas Transmission Ltd.
147	2009	Release of Substance	NOVA Gas Transmission Ltd.

5 rows | 1-4 of 4 columns

Hide

```
a = table(ReqCompdataAllArea$Year)



Year = c(2009,2010,2011,2012,2013,2014,2015,2016,2017)    # Vector with years from
2009 and 2017
IncidentNumber = c(19,24,16,15,17,10,10,7,15)                 # Incident count for
each year
c.df = data.frame(Year,IncidentNumber)               # Data Frame



gd = lm(IncidentNumber ~ Year, data=c.df)
options(scipen=999)
gd$coef
```

```
(Intercept)        Year
   2832.978      -1.400
```

Hide

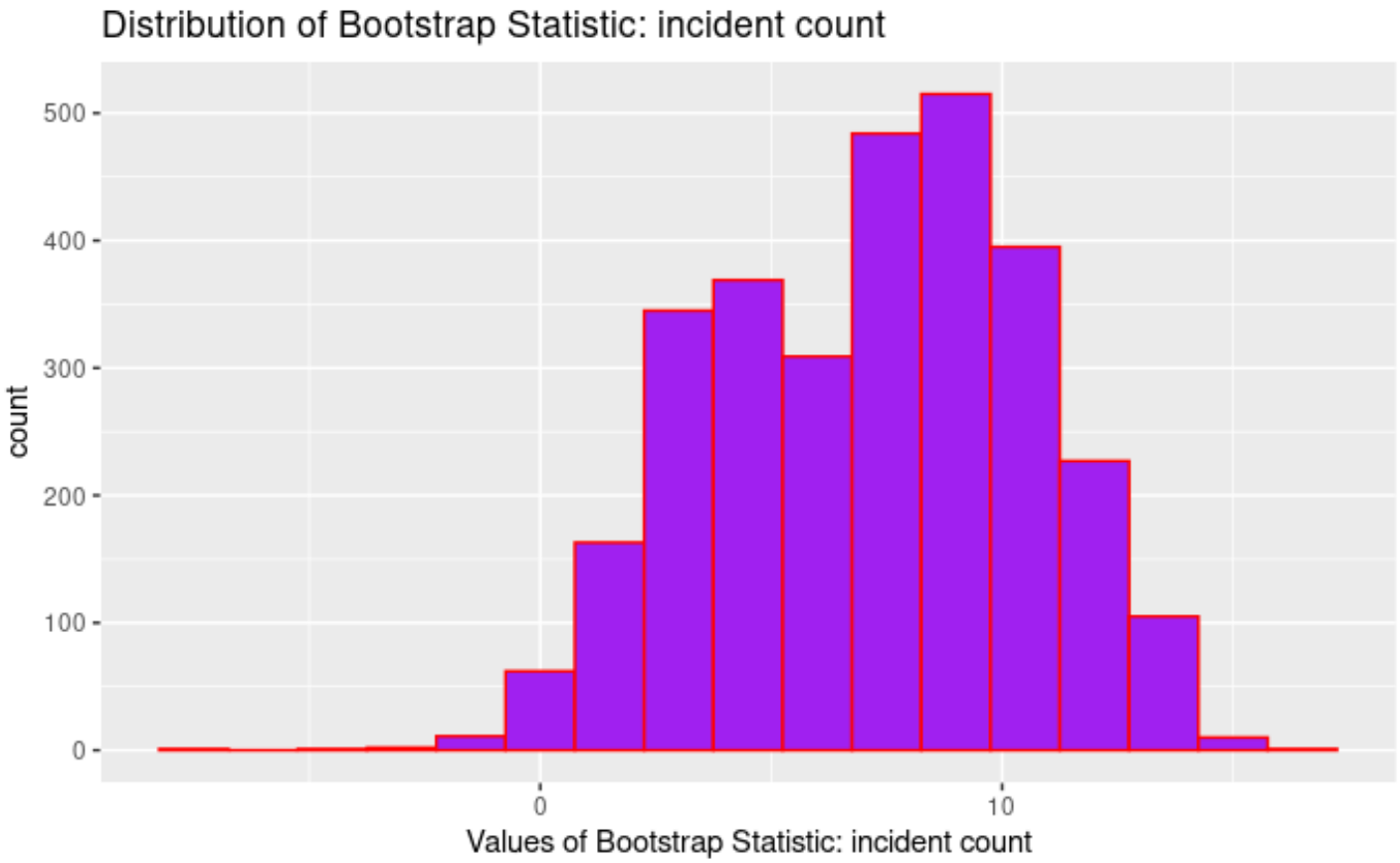
```
Nbootstraps = 3000 #resample  3000 times
cor.boot = numeric(Nbootstraps) #define a vector to be filled by the cor boot stat
a.boot = numeric(Nbootstraps) #define a vector to be filled by the a boot stat
b.boot = numeric(Nbootstraps) #define a vector to be filled by the b boot stat
ymean.boot = numeric(Nbootstraps)


nsize = dim(c.df)[1]  #set the n to be equal to the number of bivariate cases, numb
er of rows
xvalue = 2018 #set x = 60000
#start of the for loop
for(i in 1:Nbootstraps)
{   #start of the loop
    index = sample(nsize, replace=TRUE)  #randomly picks n- number between 1 and n,
assigns as index
    nova.boot = c.df[index, ] #accesses the i-th row of the SAT_2010High data frame
    #
    cor.boot[i] = cor(~IncidentNumber, ~Year, data=nova.boot) #computes correlation
for each bootstrap sample
    nova.lm = lm(IncidentNumber~Year, data=nova.boot)  #set up the linear model
    a.boot[i] = coef(nova.lm)[1] #access the computed value of a, in position 1
    b.boot[i] = coef(nova.lm)[2] #access the computed valeu of b, in position 2
    ymean.boot[i] = a.boot[i] + (b.boot[i]*xvalue)
}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
bootstrapresultsdf = data.frame(cor.boot, a.boot, b.boot, ymean.boot)
head(bootstrapresultsdf,10)
```

	cor.boot <dbl>	a.boot <dbl>	b.boot <dbl>	ymean.boot <dbl>
1	-0.9460972	3107.488	-1.537102	5.614841
2	-0.8100312	5046.722	-2.500000	1.722222
3	-0.9242393	3631.743	-1.797806	3.771160
4	-0.8761373	4129.268	-2.044586	3.292994
5	-0.7642408	2913.610	-1.439815	8.063272
6	-0.8978477	3787.944	-1.875000	4.194444
7	-0.8954851	4783.554	-2.369231	2.446154
8	-0.8964371	4482.159	-2.219626	2.953271
9	-0.8955700	5084.911	-2.518987	1.594937
10	-0.8042524	3475.067	-1.718121	7.899329

1-10 of 10 rows

Hide

```
ggplot(bootstrapresultsdf, aes(x = ymean.boot)) + geom_histogram(col="red", fill="p
urple", binwidth=1.5) + xlab("Values of Bootstrap Statistic: incident count") + ggt
itle("Distribution of Bootstrap Statistic: incident count")
```



Hide

```
favstats(~ymean.boot, data=bootstrapresultsdf)
```

	min	Q1	median	Q3	max	mean	sd	n	missing
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
	-7.103448	4.351339	7.555336	9.719094	15.97674	7.174171	3.414368	3000	0

1 row

Hide

```
qdata(~ymean.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```

```
      2.5%      97.5% 
 0.7330378 13.1933523 
```