1. Introduction

There has been much discussion in the media and in the political realm about the evidence for climate change in connection with the expected global warming, and many experts warn to end our dependence on fossil fuels. However, oil still plays an important role in the global economy despite the continued efforts to reduce its use and to find alternative green energy sources. Energy in the United States comes mostly from fossil fuels: in 2020, data showed that 35% of the nation's energy originates from petroleum, 10% from coal, and 34% from natural gas, nuclear power supplied 9% and renewable energy supplied 12%. For transportation, approximately 90% of the vehicles run on gasoline and diesel combined. Based on its use in fuels and countless consumer goods, it appears that oil will continue to be in high demand for the foreseeable future.

Like most commodities, the fundamental driver of oil's price is supply and demand in the market, and market sentiment toward the physical product. In addition, oil futures contracts, which are traded heavily by speculators, play a dominant role in price determination. Cyclical trends in the commodities market may also play a role. The Figue.1 shows the crude oil reserves in the world. There are many different grades of crude oils, and their prices heavily depend on the factors mentioned above, yet, they all have slightly different spot prices due to their own properties and characteristics. However, because the external influences have almost equal impact on every each one of crude oils, if two different crude oils have the same population mean over the period, then the property of each oil does not play a role in determining the price, thus it can be negligible when predicting oil price. So, my research

question is whether the property of each oil has impact on the oil price or it can be easily ignored when predicting the oil price.

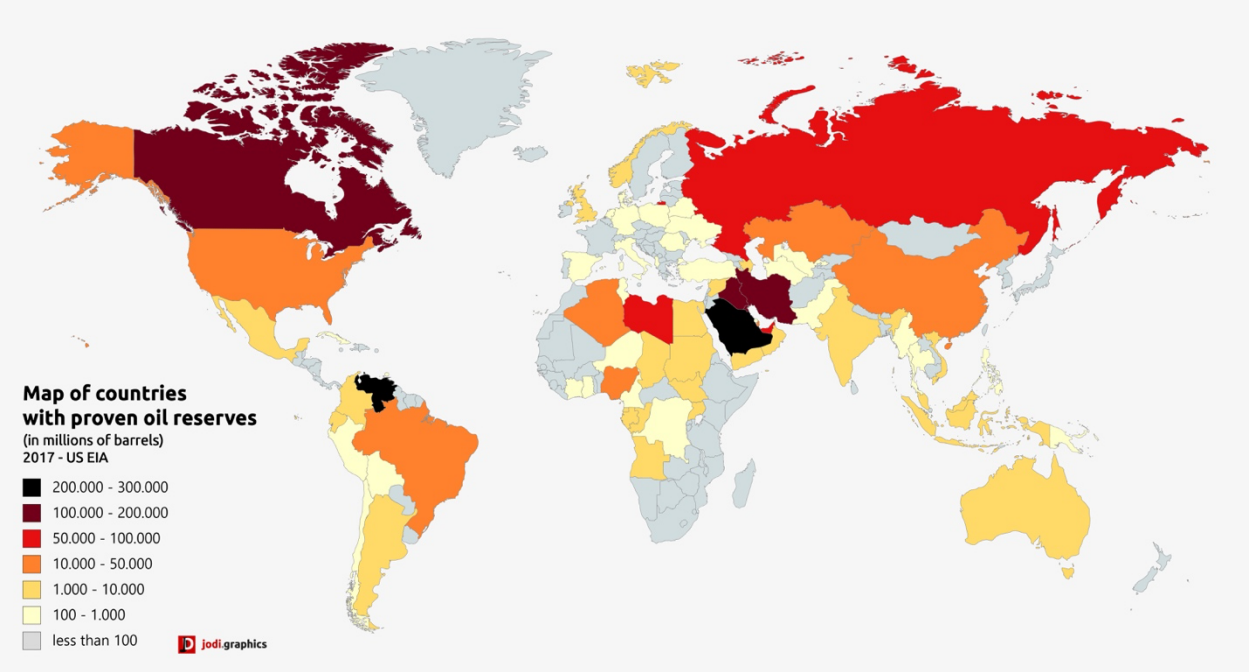


Figure 1. Map of countries with proven oil reserves

As crude oil differs in quality and availability depending on where it comes from, producers and traders need a reliable benchmark against which to judge the correct price. Two of the most widely used benchmarks are Brent crude and WTI (West Texas Intermediate) crude oil. Brent Crude oil originates from the North Sea between the Shetland Islands and Norway, and 2/3 of all oil is priced using Brent Crude as the benchmark. Brent Crude is produced near the sea, so transportation costs are significantly lower. West Texas Intermediate (WTI) origins primarily from Oklahoma, Texas, Louisiana, and North Dakota, and it accounts for the rest 1/3 of all oil pricing benchmark. It is produced in landlocked areas and make the transportation costs more onerous. I used the data of these two oil prices to answer my research question.

2. Methods / Results

The source of both Brent and WTI crude oil prices is US Energy Information Administration (Government Website). As mentioned in the Introduction, my research question is whether the difference of the property of each oil has impact on the oil price or they can be easily negligible when predicting oil prices. I assume that if two different grades of oils have the same average price over the period, I think we do not need to take their own properties into consideration when we predict or analyze the oil price, unless the property changes dramatically. My hypothesis is that even though these two crude oils have very similar trend and similar distribution over the extended period, due to their own characteristics and properties, they would have different population means over the period, so the properties of each oil cannot be negligible when predicting the prices. Hence, the null hypothesis is that "the population mean of these two oil prices are the same", and the alternative hypothesis is that "the population mean of these two oil prices are not the same". Since the population variances are unknown and there is no evidence that they are the same, I performed two separate tests for each case. The significance level α is assumed as 5% for each of the test. The computation was done by "R". Since the distribution of the both oils is almost the same, but the distribution is not known, the samples were chosen randomly using built-in function "sample()".

The oil price data are imported from US Energy Information Administration and the data type was transformed from list to data frame to draw plots and histograms with "ggplot2" library as below. The Figure.2 shows that how similar these two oil prices are.

```r
library(ggplot2)

price_brent <- read.csv("/Users/jongbaeyoon/Documents/CU Boulder/Fall,
                        2021/STAT_5000/Project/Data_cleaned/CrudePrice_Brent.csv")

price_b <- as.data.frame(price_brent);
price_b$Date <- as.Date(price_b$Date, "%d-%b-%y"); head(price_b)
ggplot(price_b,aes(x=Date, y=Price_Brent)) +
  geom_line(stat="identity") + xlab("Date") + ylab("Price") + ggtitle("Brent Crude Oil")
+ theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
        axis.title.x = element_text(color="Black", size=15, face="bold"),
        axis.title.y = element_text(color="Black", size=15, face="bold"))

price_wti <- read.csv("/Users/jongbaeyoon/Documents/CU Boulder/Fall,
                        2021/STAT_5000/Project/Data_cleaned/CrudePrice_WTI.csv")

price_w <- as.data.frame(price_wti)
price_w$Date <- as.Date(price_w$Date, "%d-%b-%y"); head(price_w)
ggplot(price_w,aes(x=Date, y=Price_WTI)) +
  geom_line(stat="identity") + xlab("Date") + ylab("Price") + ggtitle("WTI Crude Oil")
+ theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
        axis.title.x = element_text(color="Black", size=15, face="bold"),
        axis.title.y = element_text(color="Black", size=15, face="bold"))
```

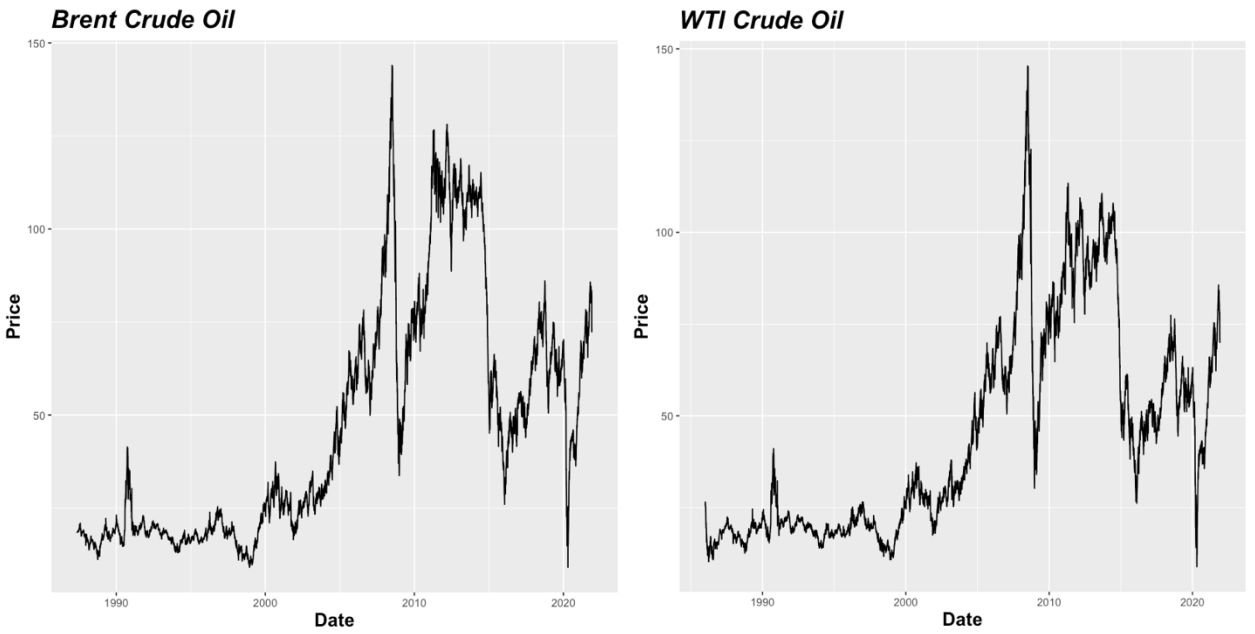


Figure2. Crude Oil Price

In addition, these two oils also have a very similar distribution as indicated in the Figure.3.

```r
ggplot(price_b,aes(x=Price_Brent)) +
  geom_histogram(color="black", fill="white", binwidth = 5)+xlab("Date") + ylab("Price")
    + ggtitle("Brent Crude Oil")
    + theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
            axis.title.x = element_text(color="Black", size=15, face="bold"),
            axis.title.y = element_text(color="Black", size=15, face="bold"))

ggplot(price_w,aes(x=Price_WTI)) +
  geom_histogram(color="black", fill="white", binwidth = 5)+xlab("Date") + ylab("Price")
    + ggtitle("WTI Crude Oil")
    + theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
            axis.title.x = element_text(color="Black", size=15, face="bold"),
            axis.title.y = element_text(color="Black", size=15, face="bold"))
```

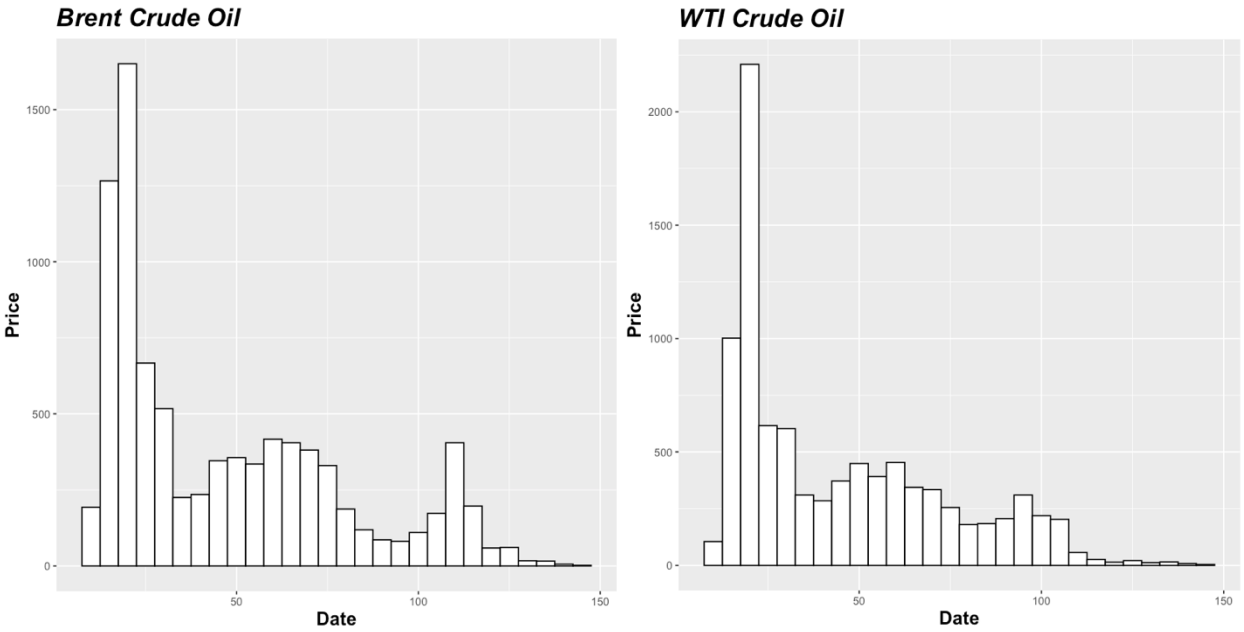


Figure.3. Distribution of the Crude Oil price

After the data is successfully imported, I transform the data from data frame to matrix and create sample datasets, sample means and variances. The size of the sample is 1,000 per each.

```
n = 1000 # size of sample

brent = data.matrix(price_b[,2]) # Brent Crude Oil price in Matrix
wti = data.matrix(price_w[,2]) # WTI Crude Oil Price in Matrix

brent_s = sample(brent, n, replace = FALSE, prob = NULL) # Brent Sampling
brent_m = mean(brent_s) # Brent Sample Mean
brent_v = var(brent_s) # Brent Sample Variance

wti_s = sample(wti, n, replace = FALSE, prob = NULL) # WTI Sampling
wti_m = mean(wti_s) # WTI Sample Mean
wti_v = var(wti_s) # WTI Sample Variance
```

As mentioned in the Introduction, population variances are unknown, and I have no clue whether the variances are the same or not, I performed two separate tests for each case with the same datasets but slightly different method. Student t-test and Welch's t-test. First, I assume that the variances are the same.

Step 1: Find the statistic.

I used t-test and pooled variance for this “two-sample” test.

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{S_p^2(\frac{1}{n_X} + \frac{1}{n_Y})}} \sim t_{n_X+n_Y-2} \text{ , where } S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

Step 2: Form the test

$$H_0 : \mu_B = \mu_W$$

$$H_1 : \mu_b \neq \mu_W$$

Reject H_0, in favor of H_1, if $\bar{X_B} - \bar{X_W} > (\mu_B - \mu_W) + C$ or $\bar{X_B} - \bar{X_W} < (\mu_B - \mu_W) - C$

Step 3: Find C

$$\alpha = P(Reject H_0; \delta_0 = \mu_B - \mu_W = 0) = P(\bar{X_B} - \bar{X_W} > C \text{ OR } \bar{X_B} - \bar{X_W} < -C) = P(\bar{X_B} - \bar{X_W} > C) + P(\bar{X_B} - \bar{X_W} < -C)$$

$$= P\left(\frac{\bar{X_B} - \bar{X_W}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}} > \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right) + P\left(\frac{\bar{X_B} - \bar{X_W}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}} < \frac{-C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right)$$

$$= P\left(T > \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right) + P\left(T > \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right) = 2P\left(T > \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right)$$

$$P\left(T > \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}}\right) = \frac{1}{2}\alpha$$

$$T_{\alpha/2,n-1} = \frac{C}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}} \text{ where } \alpha = 0.05, n = 1000$$

$$T_{0.025,999} = 1.96, S_p^2 = 944.8488, \text{ Thus, } C = 2.6943$$

Step 4: Conclusion

Reject H_0, in favor of H_1, if $\bar{X_B} - \bar{X_W} > 2.6943$ OR $\bar{X_B} - \bar{X_W} < -2.6943$

With the R built in Function t.test(), the result is as below.

```
# Student t-test: Variance assumed equal
sp = ((n-1)*brent_v + (n-1)*wti_v)/(n+n-2);sp
t1 = (brent_m - wti_m)/sqrt(sp*(1/n + 1/n));t1
p1 = 2*(1 - pt(abs(t1), n+n-2));p1
s_t = t.test(brent_s, wti_s, var.equal = TRUE, alternative = "two.sided");s_t
```

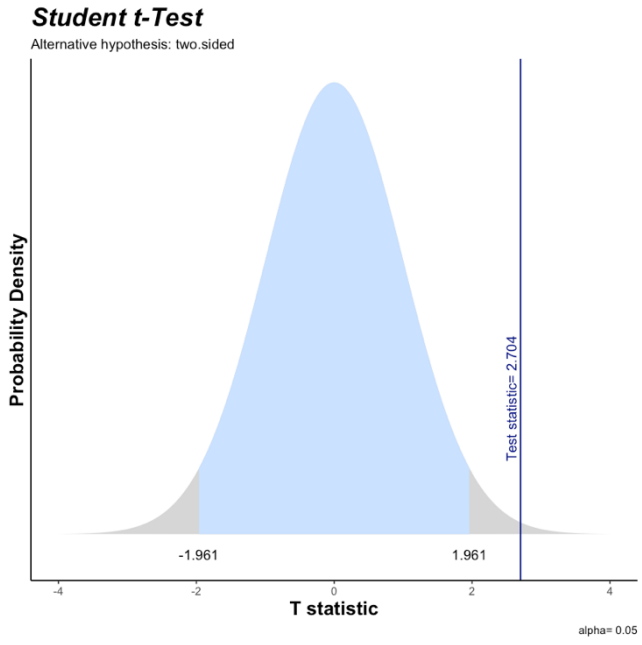
```
        Two Sample t-test

data:  brent_s and wti_s
t = 2.704, df = 1998, p-value = 0.006908
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.021212 6.413058
sample estimates:
mean of x mean of y
 48.27240  44.55527
```

p-value (=0.006908) is smaller than the significance level α(=0.05) as desired.

I used “gginference” Library to plot the t-distribution with the confidence interval, critical value, t value and p-value. The graph clearly shows that the sample mean (Brent_m – WTI-m) is in the rejection region.

```
library(gginference)
ggttest(s_t) + xlab("T statistic") + ylab("Probability Density") + ggtitle("Student t-Test")
    + theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
            axis.title.x = element_text(color="Black", size=15, face="bold"),
            axis.title.y = element_text(color="Black", size=15, face="bold"))
```



Second, I assume that the variances are NOT the same and use Welch's t-test.

Step 1: Find the statistic (Welch's T-test)

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}} \overset{approx}{\sim} t_\upsilon \text{ , where } \upsilon = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\left(\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}\right)}$$

Step 2: Form the test

$H_0 : \mu_B = \mu_W$

$H_1 : \mu_b \neq \mu_W$

Reject H_0, in favor of H_1, if $\bar{X_B} - \bar{X_W} > (\mu_B - \mu_W) + C$ or $\bar{X_B} - \bar{X_W} < (\mu_B - \mu_W) - C$

Step 3: Find C

$$\alpha = P(Reject H_0; \delta_0 = \mu_B - \mu_W = 0) = P(\bar{X_B} - \bar{X_W} > C \text{ OR } \bar{X_B} - \bar{X_W} < -C) = P(\bar{X_B} - \bar{X_W} > C) + P(\bar{X_B} - \bar{X_W} < -C)$$

$$= P\left(\frac{\bar{X_B} - \bar{X_W}}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}} > \frac{C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right) + P\left(\frac{\bar{X_B} - \bar{X_W}}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}} < \frac{-C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right)$$

$$= P\left(T > \frac{C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right) + P\left(T < \frac{-C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right) = 2P\left(T > \frac{C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right)$$

$$P\left(T > \frac{C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}}\right) = \frac{1}{2}\alpha$$

$$T_{\alpha/2,\upsilon} = \frac{C}{\sqrt{(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y})}} \text{ where } \alpha = 0.05, \upsilon = 1960.3$$

$T_{0.025,1960.3} = 1.96$, Thus, C = 1.3747 X 1.96 = 2.6944

Step 4: Conclusion

Reject H_0, in favor of H_1, if $\bar{X_B} - \bar{X_W} > 2.6944$ OR $\bar{X_B} - \bar{X_W} < -2.6944$

With the R built in Function t.test(), the result is as below.

```r
# Welch's t-test: Variance assumed not equal

v = (brent_v/n + wti_v/n)^2 / ((brent_v/n)^2/(n-1) + (wti_v/n)^2/(n-1));v
t2 = (brent_m - wti_m)/sqrt((brent_v/n + wti_v/n)); t2
p2 = 2*(1 - pt(abs(t2), n+n-2));p2
w_t = t.test(brent_s, wti_s, var.equal = FALSE, alternative = "two.sided");w_t
```

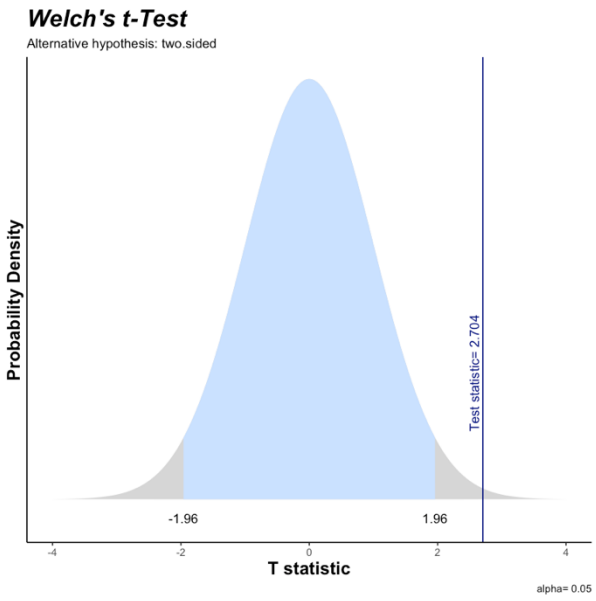
```
        Welch Two Sample t-test

data:  brent_s and wti_s
t = 2.704, df = 1972.9, p-value = 0.006909
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.021192 6.413078
sample estimates:
mean of x mean of y
 48.27240  44.55527
```

p-value(=0.006909) is smaller than significance level α (=0.05) as desired.

I plot the t-distribution with the confidence interval, critical value, t value and p-value. The plot clearly shows that the sample mean (Brent_m – WTI-m) is in the rejection region.

```r
ggttest(w_t) + xlab("T statistic") + ylab("Probability Density") + ggtitle("Welch's t-Test")
    + theme(plot.title = element_text(color="Black", size=20, face="bold.italic"),
            axis.title.x = element_text(color="Black", size=15, face="bold"),
            axis.title.y = element_text(color="Black", size=15, face="bold"))
```



3. Conclusion

Both tests reach the same conclusion that reject the null hypothesis, in favor of the alternative hypothesis which is that Brent Crude oil and WTI crude oil do not have the same population mean. In both tests, the significance level α is set as 0.05 and in each case, the p-value is smaller than α, and the sample mean is in the rejection region as desired. From the results, I conclude that even though the external influences such as supply and demand and oil futures contracts have significant impacts on crude oil prices, each oil's own property and characteristics still play a role in determining the oil price and we should take these properties into consideration when predicting the price. With this result in hand, in my future research I would focus more on the quantitative analysis about the cost impacts from each property of an oil (i.e., API gravity, contents of Sulfur, etc.) as well as the impacts from external influences.