Calgary Air Quality Project: Data 602

Research Topic:

Introduction

Likely, few people in Calgary think about the quality of the air we breathe. Generally, it seems good. Yet, there is a national Air Quality Health Index (AQHI) with data from major urban centres across Canada including Calgary. The data is collected by the “Calgary Region Airshed Zone” and, submitted to Alberta Environment. Data stations continuously collect data with an Air Quality Health Index calculated hourly. The data is sent to Alberta Environment’s “air data warehouse”. So, there is a robust multi-level government effort to collect data, evaluate the data and, provide an Air Quality Health Index. What is the index and what parameters go into it? More salient, what is the quality of the air we breathe?

Dataset

The open Calgary web portal provides a CSV file of data. The complete file is 470,000 rows of data with multiple columns. The data provides daily measurements of air pollutants by station location. Some of the data goes back to 1980 although, the parameters collected have changed over time. The station is named and, a latitude and longitude location provided. At least eight pollutants are monitored which are particulate matter (PM2.5), ground level ozone, nitrogen dioxide, carbon monoxide, non-methane hydrocarbons, nitric oxide, total oxides of nitrogen plus, the composite Air Quality Health Index. Primarily, ground level ozone, PM2.5 and nitrogen dioxide are used in the composite Air Quality Health Index. The 10 point Air Quality Health index (AQHI) was initiated in about 2012. It does not exist prior to this.

This data is licensed under an “Open Government License” and is available at Open Calgary at: https://data.calgary.ca/Environment/Air-Quality-Index-over-time/3bxb-hhuj (https://data.calgary.ca/Environment/Air-Quality-Index-over-time/3bxb-hhuj).

Guiding Questions

1. Is there difference between the mean value of Methane when calculated using different methods,i.e. instrumental and Calibrated with Methane/Propane?

2. Is there positive linear relationship between Carbon Monoxide and Air Quality Health Index. Is the Air Quality Health Index derived from Carbon Monoxide?

3. How has Calgary's air quality changed from the 1980's to times?

DATA WRANGLING:

The data was "wrangled" in both Python and R. The air pollutants were all included in one column so the data needed to be filtered to create a files of individual pollutants. The latitude and longitude data needed parentheses stripped and the latitude and longitude placed into separate columns to create "northing" and "eastings". A 'date" column required "wrangling" to strip out the year and month for further analysis.Further work was made to aggregate the air pollutants by month, create mean monthlu values and, examine this data over multi-year time periods. This data was then exported a csv file and read into R. Or, the data file was read directly into R and "wrangled" in R.

Guiding Question 1:

Importing Data and spearating year, month and day out of date column.

```
CAirData = read.csv("Historical Air Quality.csv")
CAirData = CAirData %>%
  separate(Date, sep="/", into = c("year", "month", "day"))
head(CAirData,5)
```

	Station.Name <fctr>	year <chr>	month <chr>	day <chr>	Method <fctr>
1	Calgary Central-Inglewood	2019	08	31	Instrumental
2	Calgary Central-Inglewood	2019	08	31	Instrumental
3	Calgary Southeast	2019	08	31	Sharp (hybrid nephelometer/BAM sys)
4	Calgary Central-Inglewood	2019	08	31	Instrumental
5	Calgary Southeast	2019	08	31	Calculated

5 rows | 1-6 of 12 columns

For checking difference between mean value of methane when calculated using “Calibrated with Methane/Propane” method and mean value of methane when calculated using “Instrumental” method, - we will first filter data that only concnetrateS on these data items. - We will take a sample data for “Calibrated with Methane/Propane” method and “Instrumental” method. Measuring methane level using “Calibrated with Methane/Propane” method is introduced only in Calgary Southeast station starting July 2019. So for a fair comparison we will only concnetrate on the Calgary Southeast station and data from 2019. We will take a sample of 25 for “Calibrated with Methane/Propane” method and 25 for “Instrumental” method. - We will start with our hypotheses and use a permutation test for finding the P-Value. Based on the P-Value we can make conclusions.

```
mMethod=filter(CAirData, Parameter=='Methane' & year==2019 & Station.Name=='Calgary Southeast')
head(mMethod,5)
```

	Station.Name <fctr>	year <chr>	mo... <chr>	... <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>
1	Calgary Southeast	2019	08	31	Calibrated with Methane/Propane	Methane	2.1522
2	Calgary Southeast	2019	08	30	Calibrated with Methane/Propane	Methane	2.3870
3	Calgary Southeast	2019	08	29	Calibrated with Methane/Propane	Methane	2.0957
4	Calgary Southeast	2019	08	28	Calibrated with Methane/Propane	Methane	2.0043
5	Calgary Southeast	2019	08	27	Calibrated with Methane/Propane	Methane	2.1913

5 rows | 1-8 of 12 columns

```
tail(mMethod,5)
```

	Station.Name <fctr>	year <chr>	month <chr>	d... <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>	Units <fctr>
147	Calgary Southeast	2019	01	05	Instrumental	Methane	2.3087	ppm
148	Calgary Southeast	2019	01	04	Instrumental	Methane	2.3957	ppm
149	Calgary Southeast	2019	01	03	Instrumental	Methane	2.8130	ppm
150	Calgary Southeast	2019	01	02	Instrumental	Methane	2.3000	ppm
151	Calgary Southeast	2019	01	01	Instrumental	Methane	2.4043	ppm

5 rows | 1-9 of 12 columns

```
sampleDF_Cal=sample_n(filter(mMethod,Method=='Calibrated with Methane/Propane'),25)
sampleDF_Ins=sample_n(filter(mMethod,Method=='Instrumental'),25)
sampleDF=rbind(sampleDF_Cal,sampleDF_Ins)#head(sampleDF_NW,10)
head(sampleDF_Cal,5)
```

	Station.Name <fctr>	year <chr>	mo... <chr>	... <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>
1	Calgary Southeast	2019	08	23	Calibrated with Methane/Propane	Methane	2.1609
2	Calgary Southeast	2019	07	01	Calibrated with Methane/Propane	Methane	2.0261
3	Calgary Southeast	2019	08	21	Calibrated with Methane/Propane	Methane	2.1130

	Station.Name <fctr>	year <chr>	mo... <chr>	... <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>
4	Calgary Southeast	2019	07	16	Calibrated with Methane/Propane	Methane	2.2609
5	Calgary Southeast	2019	07	27	Calibrated with Methane/Propane	Methane	2.0435

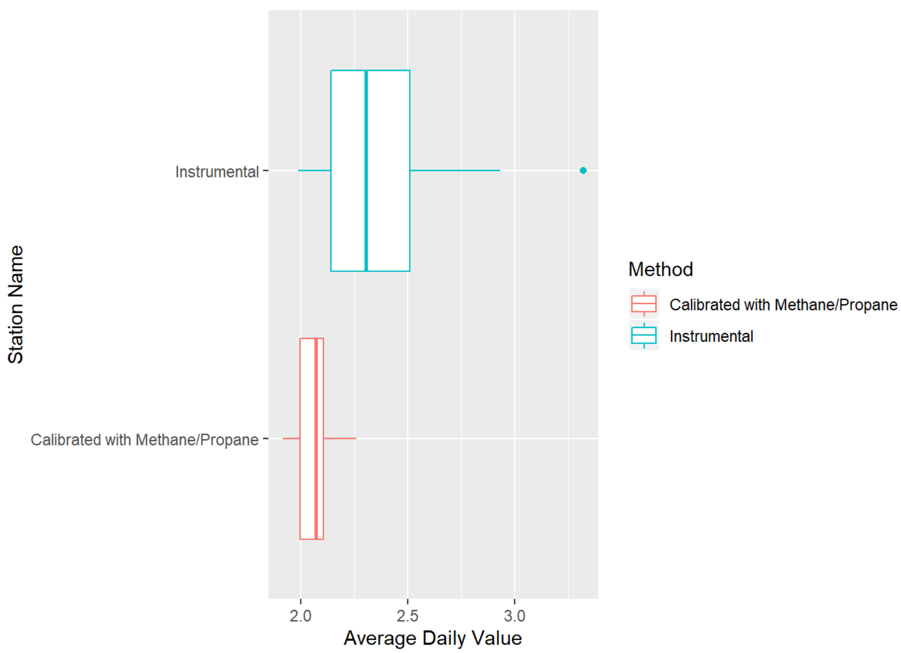
5 rows | 1-8 of 12 columns

```
head(sampleDF_Ins,5)
```

	Station.Name <fctr>	year <chr>	month <chr>	day <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>	Units <fctr>
1	Calgary Southeast	2019	01	12	Instrumental	Methane	2.1826	ppm
2	Calgary Southeast	2019	03	10	Instrumental	Methane	2.4957	ppm
3	Calgary Southeast	2019	02	11	Instrumental	Methane	2.7304	ppm
4	Calgary Southeast	2019	03	19	Instrumental	Methane	2.8957	ppm
5	Calgary Southeast	2019	02	15	Instrumental	Methane	2.3217	ppm

5 rows | 1-9 of 12 columns

```
ggplot(data=sampleDF, aes(x=Method, y=Average.Daily.Value, color=Method))+coord_flip()+ geom_boxplot()+xlab("Station Name")+
ylab("Average Daily Value")
```



Hypothesis :

Lets assume average methane level IS EQUAL when calculated using two different methods

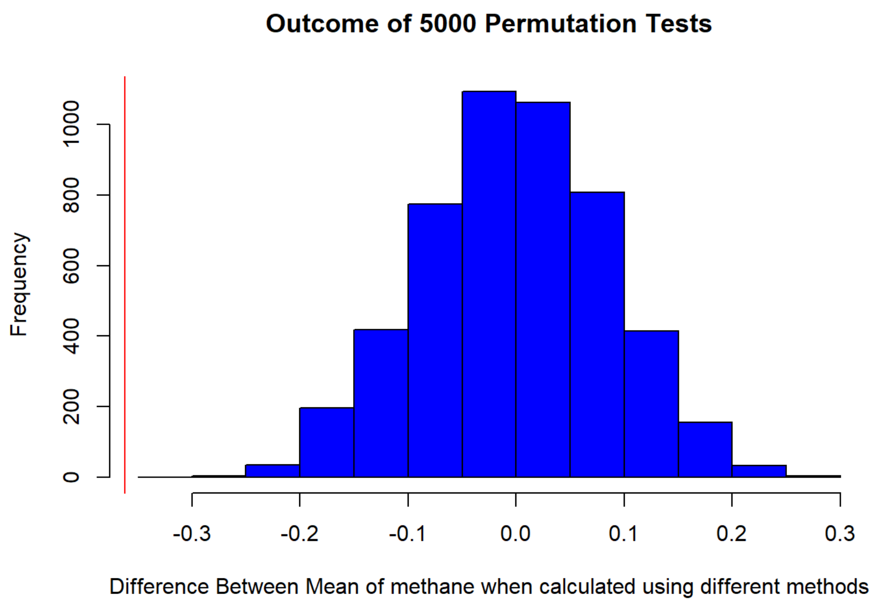
$H_0 : \mu_{Calibrate} = \mu_{Ins}$

Alternative is average methane level IS Less when calculated using "Calibrated with Methane/Propane" method than "Instrumental" me

$H_A : \mu_{Calibrate} < \mu_{Ins}$

```
obsdiff = mean(sampleDF_Cal$Average.Daily.Value)-mean(sampleDF_Ins$Average.Daily.Value)#,data=sampleDF_SW)
N = (5000 - 1)
outcome = numeric(N)
for(i in 1:N)
{ index = sample(50, 25, replace=FALSE)
  outcome[i] = mean(sampleDF$Average.Daily.Value[index]) - mean(sampleDF$Average.Daily.Value[-index]) #difference between me
ans
}
```

```
hist(outcome, xlab="Difference Between Mean of methane when calculated using different methods", ylab="Frequency", main="Out
come of 5000 Permutation Tests", col='blue')
abline(v = obsdiff, col="red")
```




```
pvalue=(sum(outcome< obsdiff))/N

cat("The computed empirical P-value is", pvalue)
```

```
## The computed empirical P-value is 0
```

In our permutation test, the empirical P-value is computed to be approximately 0. This p-value does not suggest any evidence in support of the null hypotheses $\mu_{Calibrate} = \mu_{Ins}$

We can not accept the null hypotheses. $\mu_{Calibrate} = \mu_{Ins}$, so we reject the null hypotheses. That being said “average methane level IS EQUAL when calculated using two different methods” is rejected.

We will accept alternative hypothesis $H_A : mu_{Calibrate} < \mu_{Ins}$

Conclusion: From this data it can be inferred that average methane level IS Less when calculated using “Calibrated with Methane/Propane” method than “Instrumental” method.

Guiding Question 2:

For checking the positive linear relationship between Carbon Monoxide and Air Quality Health Index, - we will first filter data that only concnetrate on these data items. - We will take a sample of year 2018 data from Calgary Central-Inglewood station to research. - We will start with hypotheses and than bootstraping method for finding a.boot and b.boot values, establish model and find 95% confidence interval with P-Value

```
caqhi=filter(CAirData, Parameter=='Air Quality Index' & year==2018 & Station.Name=="Calgary Central-Inglewood")
head(caqhi)
```

	Station.Name <fctr>	year <chr>	mo… <chr>	… <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>	Units <fctr>
1	Calgary Central-Inglewood	2018	12	31	Calculated	Air Quality Index	2.3429	null
2	Calgary Central-Inglewood	2018	12	30	Calculated	Air Quality Index	2.1238	null
3	Calgary Central-Inglewood	2018	12	29	Calculated	Air Quality Index	2.5843	null
4	Calgary Central-Inglewood	2018	12	28	Calculated	Air Quality Index	2.7089	null
5	Calgary Central-Inglewood	2018	12	27	Calculated	Air Quality Index	2.2307	null
6	Calgary Central-Inglewood	2018	12	26	Calculated	Air Quality Index	3.1056	null

6 rows | 1-9 of 12 columns

```
cco=filter(CAirData, Parameter=='Carbon Monoxide' & year==2018  & Station.Name=="Calgary Central-Inglewood")
head(cco)
```

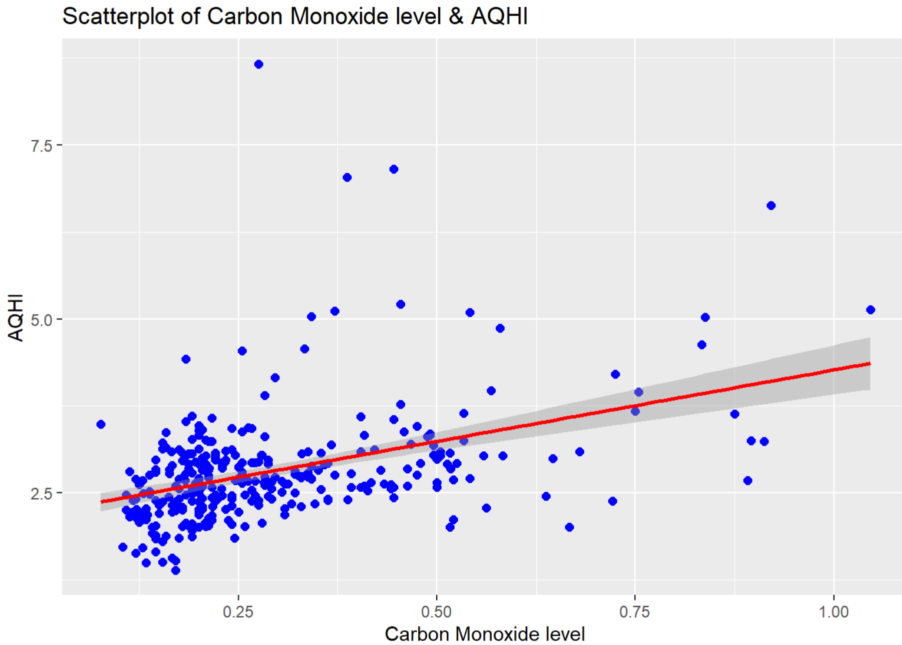
	Station.Name <fctr>	year <chr>	mo… <chr>	… <chr>	Method <fctr>	Parameter <fctr>	Average.Daily.Value <dbl>	Units <fctr>
1	Calgary Central-Inglewood	2018	12	31	Instrumental	Carbon Monoxide	0.3458	ppm
2	Calgary Central-Inglewood	2018	12	30	Instrumental	Carbon Monoxide	0.1333	ppm
3	Calgary Central-Inglewood	2018	12	29	Instrumental	Carbon Monoxide	0.4042	ppm
4	Calgary Central-Inglewood	2018	12	28	Instrumental	Carbon Monoxide	0.3417	ppm
5	Calgary Central-Inglewood	2018	12	27	Instrumental	Carbon Monoxide	0.2500	ppm
6	Calgary Central-Inglewood	2018	12	26	Instrumental	Carbon Monoxide	0.2333	ppm

6 rows | 1-9 of 12 columns

```
max.len<-max(length(caqhi$Average.Daily.Value),length(cco$Average.Daily.Value))
AQHI = c(caqhi$Average.Daily.Value, rep(NA, max.len - length(caqhi$Average.Daily.Value)))
CO = c(cco$Average.Daily.Value, rep(NA, max.len - length(cco$Average.Daily.Value)))
co_aqhi=data.frame(AQHI,CO)
co_aqhi = drop_na(co_aqhi)
```

Visulizing relationship using scatterplot.

```
ggplot(data=co_aqhi, aes(x = CO, y = AQHI)) + geom_point(col="blue", size=2, position="jitter") + xlab("Carbon Monoxide leve
l") + ylab("AQHI") + ggtitle("Scatterplot of Carbon Monoxide level & AQHI")+ stat_smooth(method="lm", col='red')
```



The statisticall hypotheses is $H_0 : B \leq 0$ Slope to be negative, Air Quality Health Index CAN NOT be expressed as a positive linear function of Carbon Monoxide level $H_1 : B > 0$ Slope greater than zero, Air Quality Health Index CAN be expressed as a positive linear function of Carbon Monoxide level

Value of F_{Obs} From R

```
airmodel=lm(AQHI~CO, co_aqhi )
coefficients(summary(airmodel))
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.217203 0.08295912 26.726449 1.618647e-83
## CO          2.051036 0.24790047  8.273628 3.546171e-15
```

Here $T_{Obs} = 8.273628$ and $P - Value = 3.546171e - 15$

Conclusion: This P-Value = 3.546171e-15 is less than 0.05. There are not enough evidence in support of null hypotheses. Our null hypotheses that "Air Quality Health Index CAN NOT be expressed as a positive linear function of Carbon Monoxide level" do not hold.

Alternative hypothesis "Air Quality Health Index CAN be expressed as a positive linear function of Carbon Monoxide level" is True.

We can conclude that Air Quality Health Index can be modeled by a positive linear function of Carbon Monoxide level.

95% confidence interval for intercept and slope of linear model of AQHI and CO:

95% confidence interval for A and B is calculated using follwoing R function.

```
confint(airmodel)
```

```
##                2.5 %   97.5 %
## (Intercept) 2.053989 2.380417
## CO          1.563316 2.538757
```

95% confidence interval for intercept A is: $2.053989 \leq B \leq 2.380417$ 95% confidence interval for slope B is: $1.563316 \leq B \leq 2.538757$

Computing mean value of AQHI when CO level is 0.82 with 95% confidence interval:

95% confidence interval for mean AQHI when CO value is 0.82 can be computed using following R function.

```
predict(airmodel, newdata=data.frame(CO=0.82), interval="conf", conf.level=0.95)
```

```
##        fit      lwr      upr
## 1 3.899052 3.628807 4.169298
```

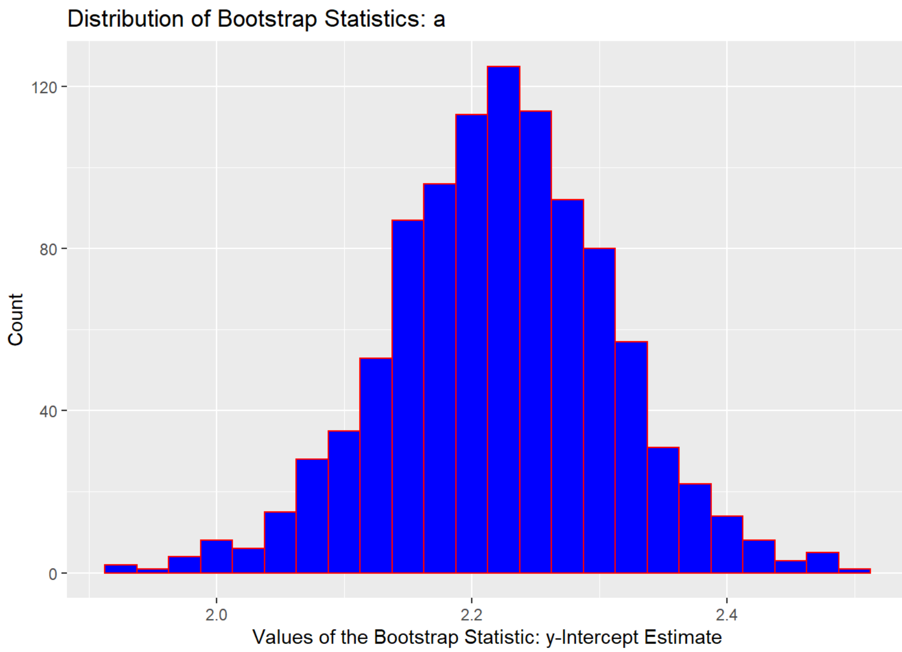
95% confidence interval for the mean AQHI when the CO=0.82 is: $3.628807 \leq \mu_{AQHI|CO=0.82} \leq 4.169298$

```
Nbootstraps_air = 1000
a.boot = numeric(Nbootstraps_air)
b.boot = numeric(Nbootstraps_air)
```

```
nsize_air = dim(co_aqhi)[1]
for(i in 1:Nbootstraps_air)
{
    index = sample(nsize_air, replace=TRUE)
    air.boot = co_aqhi[index, ]
    air.lm = lm(AQHI~CO, data=air.boot)
    a.boot[i] = coef(air.lm)[1]
    b.boot[i] = coef(air.lm)[2]

}
    bootstrapresultsdf_air = data.frame(a.boot, b.boot)
```

```
ggplot(bootstrapresultsdf_air, aes(x = a.boot)) + geom_histogram(col="red", fill="blue", binwidth=0.025) + xlab("Values of t
he Bootstrap Statistic: y-Intercept Estimate") + ylab("Count") + ggtitle("Distribution of Bootstrap Statistics: a")
```




```
qdata(~a.boot, c(0.025, 0.975), data=bootstrapresultsdf_air)
```

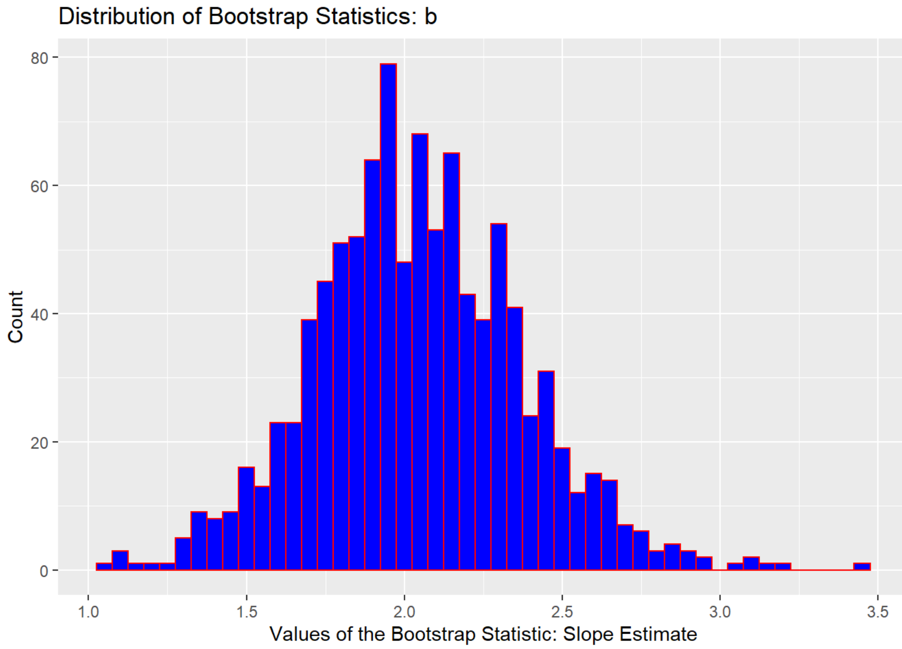
	quantile <dbl>	p <dbl>
2.5%	2.045055	0.025
97.5%	2.394003	0.975

2 rows

Mean of a_{boot} is

```
aboot=mean_(~a.boot,data=bootstrapresultsdf_air)
```

```
ggplot(bootstrapresultsdf_air, aes(x = b.boot)) + geom_histogram(col="red", fill="blue", binwidth=0.05) + xlab("Values of th
e Bootstrap Statistic: Slope Estimate") + ylab("Count") + ggtitle("Distribution of Bootstrap Statistics: b")
```




```
qdata(~b.boot, c(0.025, 0.975), data=bootstrapresultsdf_air)
```

	quantile <dbl>	p <dbl>
2.5%	1.413466	0.025
97.5%	2.717940	0.975

2 rows

Mean of b_{boot} is

```
bboot=mean_(~b.boot,data=bootstrapresultsdf_air)
```

Using the means of a_{boot} and b_{boot}, our estimate of the model is

```
cat("AQHI=",aboot,"+",bboot,"*CO")
```

```
## AQHI= 2.221804 + 2.04394 *CO
```

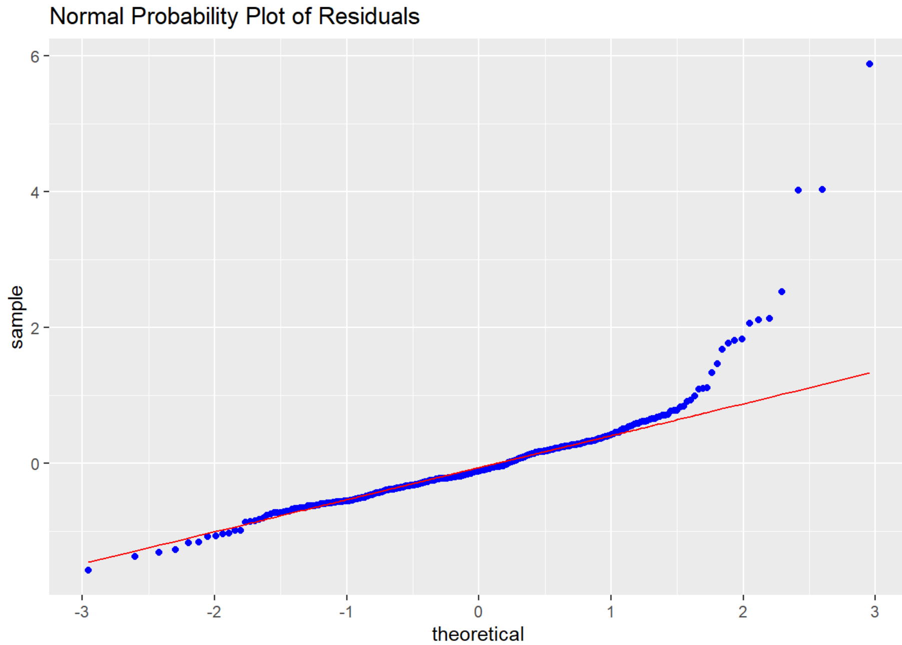
Checking Normality and homoscedasticity condition:

```
predictreturn = airmodel$fitted.values
eisreturn = airmodel$residuals
diagnosticdf = data.frame(predictreturn, eisreturn)
favstats(predictreturn)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
2.373492	2.593158	2.691505	2.932912	4.362176	2.814207	0.3391914	322	0

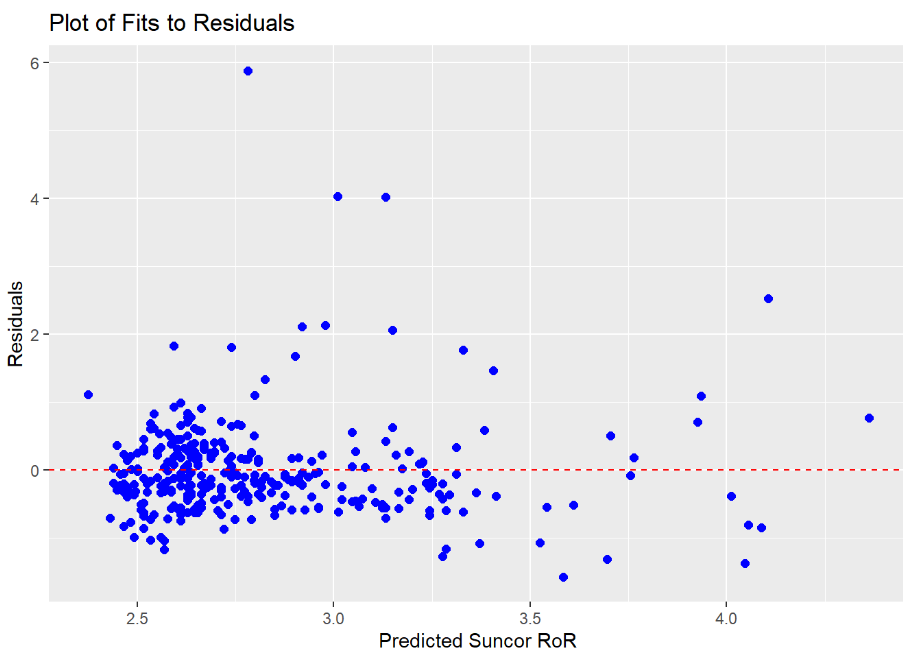
1 row

```
ggplot(data=diagnosticdf, aes(sample = eisreturn)) +  stat_qq(col='blue') + stat_qqline(col='red') + ggtitle("Normal Probabi
lity Plot of Residuals")
```



To inspect the homoscedasticity condition, we plot the fitted.values with the residuals.

```
ggplot(diagnosticdf, aes(x = predictreturn, y = eisreturn)) +  geom_point(size=2, col='blue', position="jitter") + xlab("Predicted Suncor RoR") + ylab("Residuals") + ggtitle("Plot of Fits to Residuals") + geom_hline(yintercept=0, color="red", linetype="dashed")
```



Above visulizations holds both conditions of modeling to be true. We can construct the above model.

1.The AQHI, or commonly known as the response variable, is Normally distributed with a mean $\mu = 2.814207$ and standard deviation of $\sigma = 0.339191$. 2. The homoscedasticity test: a visulisation of the plot of fits to residual shows that data are equally and symatrically plotted around fit line. data visulisation do not suggest differece in variance between Carbon Monoxide and Air Quality Health Index.

Guiding Question 3:

GROUND LEVEL OZONE:

Ground Level Ozone is one of the air pollutants comprising the Air Quality Health Index. This pollutant is examined to determine if there are changes from the 1980's to the most recent 5 year period.

Introduction:

Calgary's air quality is examined and comparisons made between the time frame 1985 to 1989 and, 2014 to 2018. Three pollutants comprise the Air Quality Health Index. These are Ground Level Ozone, Nitrogen Dioxide and Particulate Matter smaller than 2.5 microns. In particular, Ground Level Ozone and Nitrogen Dioxide are examined to see if there has been a change from the late 1980's to the last 5 years. Particulate Matter smaller than 2.5 microns only has data going back to 1997. So, not the time frame that Ground Level Ozone and Nitrogen Dioxide cover.

In terms of hypotheses:

Hnull: mean differences of ozone levels = 0, i.e. no change in Ground Level Ozone Halternate: mean differences of ozone levels are not equal to zero, i.e. there is a change in Ground Level Ozone

```
Ozone = read.csv('ozone combo.csv')

DFozone = data.frame(Ozone)

head(DFozone)
```

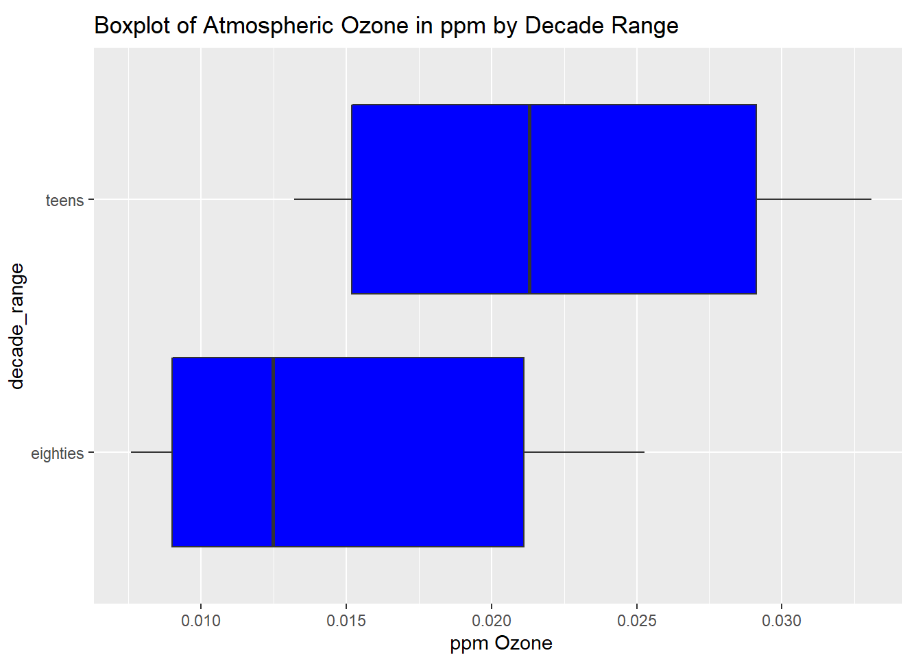
	X	Parameter	MONTH	ppm_eighties	ppm.2015	decade_range	ppm_ozone
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>
1	0	Ozone	Jan	0.007898710	0.01321828	eighties	0.007898710
2	1	Ozone	Feb	0.009364602	0.01792281	eighties	0.009364602
3	2	Ozone	Mar	0.012819079	0.02426483	eighties	0.012819079
4	3	Ozone	Apr	0.020621333	0.03111852	eighties	0.020621333
5	4	Ozone	May	0.025279839	0.03308064	eighties	0.025279839
6	5	Ozone	Jun	0.023970000	0.03266583	eighties	0.023970000

6 rows

```
neighties=12
nteens=12
```

Create a boxplot of the Ground Level Ozone levels in pppm:

```
ggplot(data=DFozone, aes(x = decade_range, y = ppm_ozone)) + geom_boxplot(fill='blue') + xlab("decade_range") + ylab("ppm Oz
one") + coord_flip() +ggtitle("Boxplot of Atmospheric Ozone in ppm by Decade Range")
```



There are seasonal i.e. month to month differences in Ozone levels. Ozone levels are higher in winter months. Since there is a match of levels by month a pair wise difference was created:

```
DFozone = DFozone %>%
  mutate(Diff = ppm_eighties-ppm.2015)
head(DFozone, 4)
```

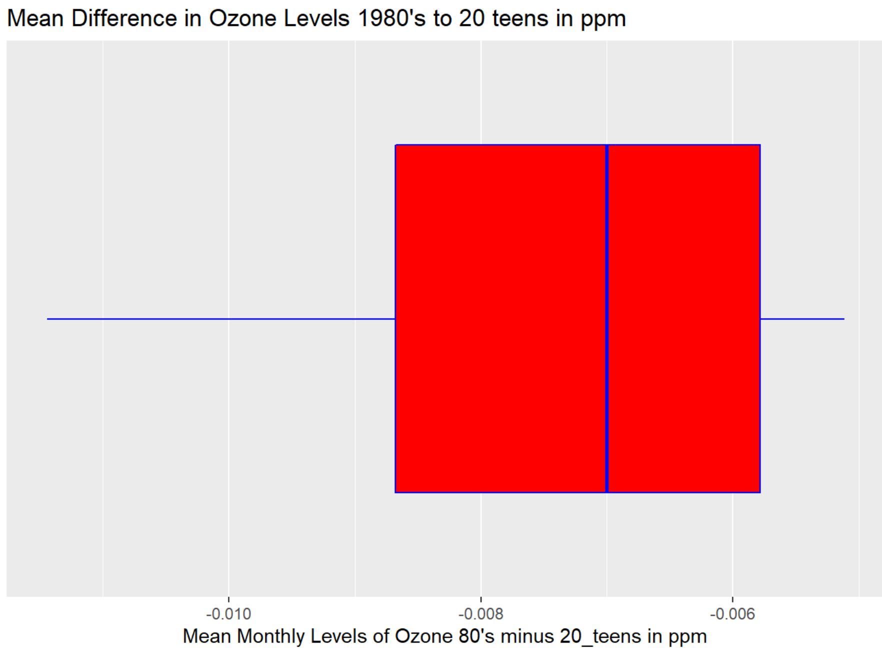
	X	Parameter	MON...	ppm_eighties	ppm.2015	decade_range	ppm_ozone	Diff
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>	<dbl>
1	0	Ozone	Jan	0.007898710	0.01321828	eighties	0.007898710	-0.005319570
2	1	Ozone	Feb	0.009364602	0.01792281	eighties	0.009364602	-0.008558205
3	2	Ozone	Mar	0.012819079	0.02426483	eighties	0.012819079	-0.011445756
4	3	Ozone	Apr	0.020621333	0.03111852	eighties	0.020621333	-0.010497186

4 rows

A box plot of the mean monthly differences between the 80's and 20_teens is created:

```
ggplot(data=DFozone, aes(x = "var", y = Diff)) + geom_boxplot(col='blue', fill= 'red') + xlab("") + ylab("Mean Monthly Level
s of Ozone 80's minus 20_teens in ppm") + scale_x_discrete(breaks=NULL) + coord_flip() + ggtitle("Mean Difference in Ozone L
evels 1980's to 20 teens in ppm")
```

```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```

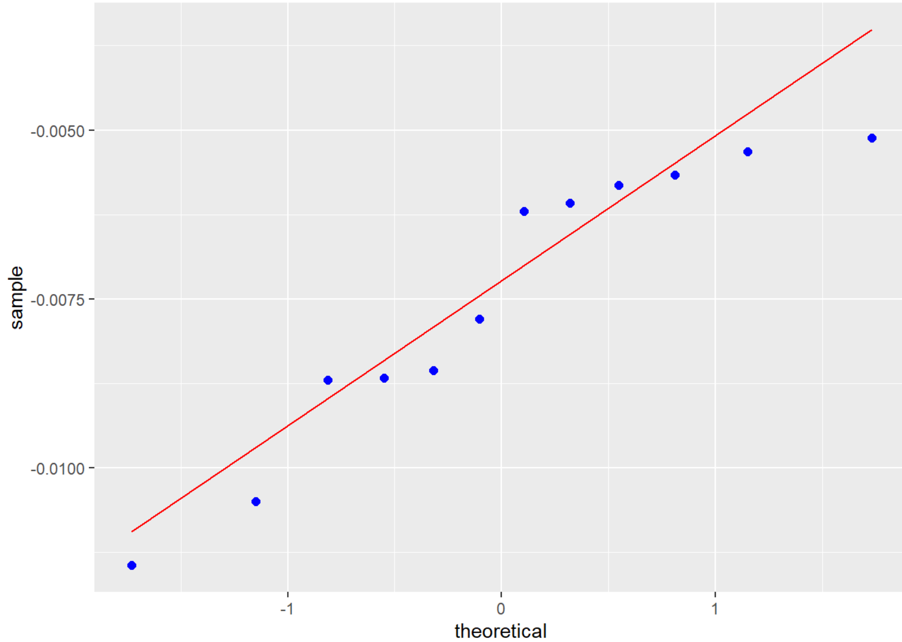


To examine whether a t.test is appropriate a normal probability plot is created. The plot demonstrates a weak normal probability distribution.

```
ggplot(data=DFozone, aes(sample = Diff)) + stat_qq(size=2, col='blue') + stat_qq_line(col='red')
```

```
## Warning: Removed 12 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 12 rows containing non-finite values (stat_qq_line).
```




```
favstats(~ Diff, data = DFozone)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
-0.01144576	-0.008679811	-0.007003369	-0.005783685	-0.005114754	-0.007490318	0.002107844	12

1 row | 1-9 of 10 columns

The mean difference is calculated as -0.007490.

Because there is a weak justification of a normal probability distribution in the difference of means a t-test is used to calculate a 95% confidence interval and a p-value.

The hypotheses are again:

Hnull: mean differences of ozone levels = 0, i.e. no change in Ground Level Ozone Halternate: mean differences of ozone levels are not equal to zero, i.e. there is a change in Ground Level Ozone

So a “two sided” test is calculated:

```
t.test(~ Diff, mu=0, alternative="two.sided", conf.level = 0.95, data = DFozone)
```

```
##
##  One Sample t-test
##
## data:  Diff
## t = -12.31, df = 11, p-value = 8.947e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.008829578 -0.006151058
## sample estimates:
##    mean of x
## -0.007490318
```

The p-value is much less than <0.05 so Hnull is rejected. Halternate is accepted: the mean differences of ozone levels between the 80’s and the last 5 years is different. The mean of the difference in ozone levels is -0.007490. Ozone levels have become worse in the last forty years. The 95% confidence interval is -0.008830 <= Mean Difference <= -0.006151.

Since the condition of a normal distribution of the mean differences is weak a bootstrap simulation of the mean differences is calculated:

```
nsims = 1000 #the number of simulations
meaneighties = numeric(nsims) #hold the mean of each resampling from eighties ppm
meanteens = numeric(nsims) #hold the mean of each resampling from teens ppm
diffmeans = numeric(nsims) #hold the difference between the sample means
eighties = filter(DFozone, decade_range=="eighties")  #filters out all eighties ppm from data frame
teens = filter(DFozone, decade_range=="teens")  #filters out all teens ppm from data frame
head(eighties, 3)# will just check verify stripping out eighties
```

	X	Parameter	MON…	ppm_eighties	ppm.2015	decade_range	ppm_ozone	Diff
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>	<dbl>
1	0	Ozone	Jan	0.007898710	0.01321828	eighties	0.007898710	-0.005319570
2	1	Ozone	Feb	0.009364602	0.01792281	eighties	0.009364602	-0.008558205
3	2	Ozone	Mar	0.012819079	0.02426483	eighties	0.012819079	-0.011445756

3 rows

```
head(teens, 3) #vice versa
```

	X	Parameter	MONTH	ppm_eighties	ppm.2015	decade_range	ppm_ozone	Diff
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>	<dbl>
1	NA			NA	NA	teens	0.01321828	NA
2	NA			NA	NA	teens	0.01792281	NA
3	NA			NA	NA	teens	0.02426483	NA

3 rows

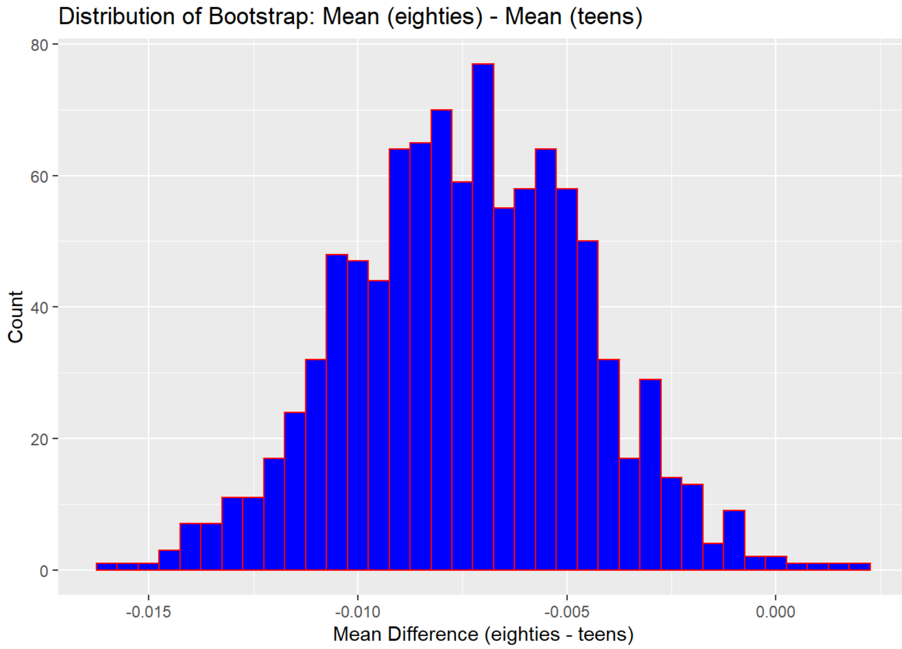
```
for(i in 1:nsims)
{   meaneighties[i] = mean(sample(eighties$ppm_ozone,neighties, replace=TRUE))  #computes the mean of 12 resampled eighties
 ozone
    meanteens[i] = mean(sample(teens$ppm_ozone,nteens,  replace=TRUE))  #computes the mean of 12 resampled teens ppm
    diffmeans[i] = meaneighties[i] - meanteens[i]  #computes the difference between the sample means
}
bootstrapOzone = data.frame(meaneighties, meanteens, diffmeans)  #create a data frame holding all the means
```

```
bootstrapOzone
```

meaneighties	meanteens	diffmeans
<dbl>	<dbl>	<dbl>
0.013918453	0.02209400	-8.175544e-03
0.011667608	0.01960439	-7.936780e-03
0.016451387	0.02294673	-6.495343e-03
0.015502975	0.02078051	-5.277534e-03
0.013679708	0.01992952	-6.249809e-03
0.012650311	0.01877561	-6.125297e-03
0.017209404	0.02036257	-3.153164e-03
0.018847890	0.02057947	-1.731585e-03
0.013318321	0.02192737	-8.609045e-03
0.017969915	0.01898090	-1.010989e-03

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

```
ggplot(data=bootstrapOzone, aes(x = diffmeans)) + geom_histogram(fill='blue', col='red', binwidth=.0005) + xlab("Mean Differ
ence (eighties - teens)") + ylab("Count") + ggtitle("Distribution of Bootstrap: Mean (eighties) - Mean (teens)")
```



The bootstrap distribution is approximately normally distributed.

```
favstats(bootstrapOzone$diffmeans)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
-0.01615778	-0.009302247	-0.007326395	-0.005421471	0.002203009	-0.007376473	0.002833725	1000

1 row | 1-9 of 10 columns

```
quantile(diffmeans, c(0.025, 0.975), data=bootstrapOzone)
```

```
##         2.5%        97.5%
## -0.013101514 -0.001966179
```

A mean difference of -0.007494 is derived from the bootstrap simulation with a 95% confidence interval of: -0.01294<= Mean Difference Ozone <= -0.001910

Ground Level Ozone levels have increased over the last forty years.

The values derived from the bootstrap simulation agree with the values calculated from the t-test. For instance, the mean difference from the t-test is -0.007490 compared to -0.007494 from the simulation. The 95% confidence intervals are:

-0.008830 <= Mean Difference Ozone <= -0.006151 t-test -0.01294<= Mean Difference Ozone <= -0.001910 simulation

This apparently provides some justification in using the t-test.

NITROGEN DIOXIDE:

Nitrogen Dioxide is one of the air pollutants comprising the Air Quality Health Index. This pollutant is examined to determine if there are changes from the 1980's to the most recent 5 year period.

The hypotheses are:

Hnull: mean differences of N2 levels = 0, i.e. no change in Nitrogen Dioxide Halternate: mean differences of N2 levels are not equal to zero, i.e. there is a change in Nitrogen Dioxide

```
N2 = read.csv('N2 combo.csv')

DFN2 = data.frame(N2)

head(DFN2)
```

	X <int>	Parameter <fctr>	Station.Name <fctr>	Northing <dbl>	Easting <dbl>	MO... <fctr>	ppm_eighties <dbl>	X.1 <lgl>	ppm.2015 <dbl>
1	0	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Jan	0.04111032	NA	0.02910430
2	1	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Feb	0.04371964	NA	0.02173860
3	2	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Mar	0.03987712	NA	0.01835591

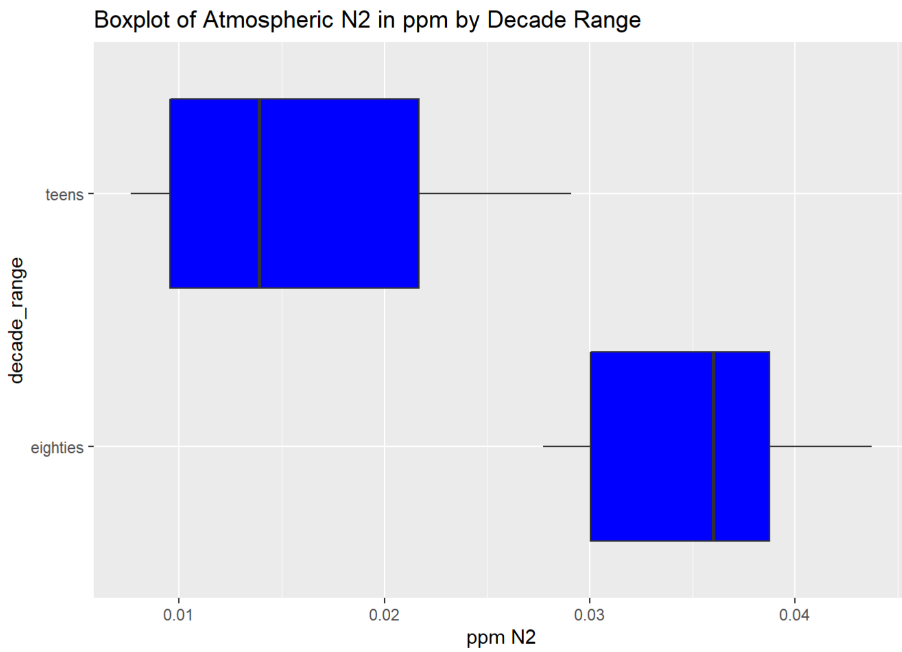
	X <int>	Parameter <fctr>	Station.Name <fctr>	Northing <dbl>	Easting <dbl>	MO... <fctr>	ppm_eighties <dbl>	X.1 <lgl>	ppm.2015 <dbl>
4	3	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Apr	0.03645442	NA	0.01160734
5	4	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	May	0.03136529	NA	0.00910000
6	5	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Jun	0.02930833	NA	0.00764661

6 rows | 1-10 of 12 columns

```
neighties=12
nteens=12
```

Create a boxplot of the Nitrogen Dioxide levels in pppm:

```
ggplot(data=DFN2, aes(x = decade_range, y = ppm_N2)) + geom_boxplot(fill='blue') + xlab("decade_range") + ylab("ppm N2") + c
oord_flip() +ggtitle("Boxplot of Atmospheric N2 in ppm by Decade Range")
```



There are seasonal i.e. month to month differences in Nitrogen Dioxide. Nitrogen Dioxide levels are higher in winter months. Since there is a match of levels by month a pair wise difference was created:

```
DFN2 = DFN2 %>%
  mutate(DiffN2 = ppm_eighties-ppm.2015)
head(DFN2, 4)
```

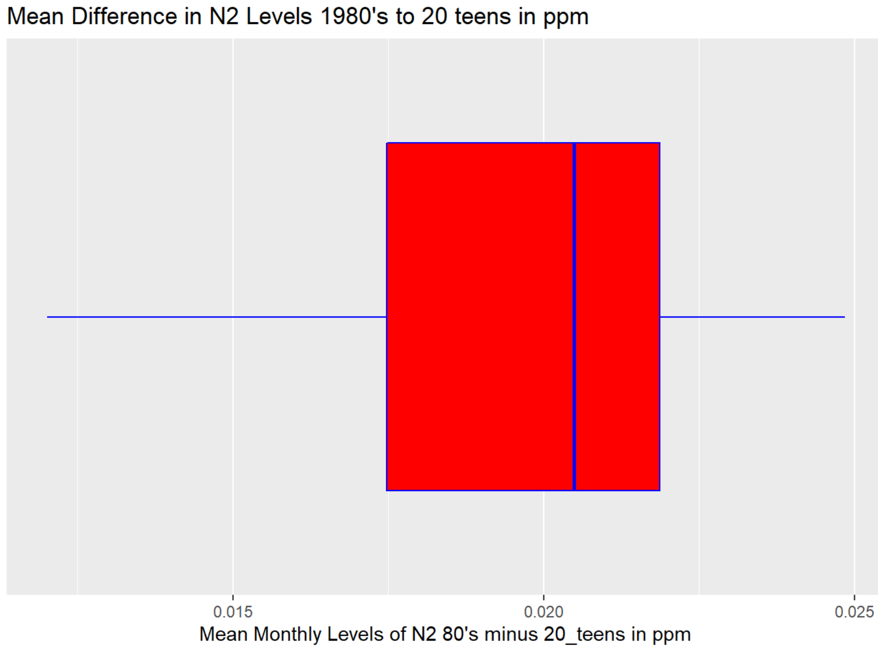
	X <int>	Parameter <fctr>	Station.Name <fctr>	Northing <dbl>	Easting <dbl>	MO... <fctr>	ppm_eighties <dbl>	X.1 <lgl>	ppm.2015 <dbl>
1	0	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Jan	0.04111032	NA	0.02910430
2	1	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Feb	0.04371964	NA	0.02173860
3	2	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Mar	0.03987712	NA	0.01835591
4	3	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Apr	0.03645442	NA	0.01160734

4 rows | 1-10 of 13 columns

A box plot of the mean monthly differences in Nitrogen Dioxide between the 80's and 20_teens is created:

```
ggplot(data=DFN2, aes(x = "var", y = DiffN2)) + geom_boxplot(col='blue', fill= 'red') + xlab("") + ylab("Mean Monthly Levels
of N2 80's minus 20_teens in ppm") + scale_x_discrete(breaks=NULL) + coord_flip() + ggtitle("Mean Difference in N2 Levels 19
80's to 20 teens in ppm")
```

```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```

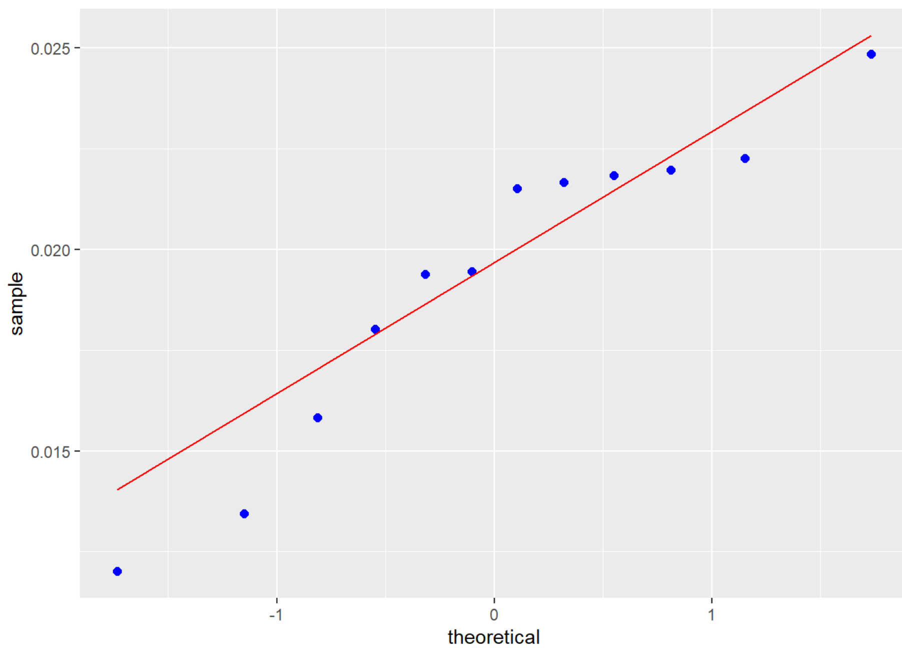


To examine whether a t.test is appropriate a normal probability plot is created. The plot demonstrates a weak normal probability distribution.

```
ggplot(data=DFN2, aes(sample = DiffN2)) + stat_qq(size=2, col='blue') + stat_qq_line(col='red')
```

```
## Warning: Removed 12 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 12 rows containing non-finite values (stat_qq_line).
```




```
favstats(~ DiffN2, data = DFN2)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
0.01200602	0.01747676	0.02048923	0.02186704	0.02484708	0.01935403	0.003870591	12	12

1 row

The mean difference is calculated as 0.01935

Because there is a weak justification of a normal probability distribution in the difference of means, a t-test is used to calculate a 95% confidence interval and a p-value.

The hypotheses are again:

Hnull: mean differences of Nitrogen Dioxide levels = 0, i.e. no change in Nitrogen Dioxide levels Halternate: mean differences of Nitrogen Dioxide levels are not equal to zero, i.e. there is a change in Nitrogen Dioxide levels

So a “two sided” test is calculated:

```
t.test(~ DiffN2, mu=0, alternative="two.sided", conf.level = 0.95, data = DFN2)
```

```
##
##  One Sample t-test
##
## data:  DiffN2
## t = 17.321, df = 11, p-value = 2.484e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.01689477 0.02181328
## sample estimates:
##  mean of x
## 0.01935403
```

The p-value is much less than <0.05 (p=2.48E-9) so Hnull is rejected. Halternate is accepted: the mean differences of Nitrogen Dioxide levels between the 80’s and the last 5 years is different. The mean of the difference in Nitrogen Dioxide levels is 0.01935403. Nitrogen Dioxide levels have become better in the last forty years. The 95% confidence interval is 0.016895 <= Mean Difference <= 0.021813.

Since the condition of a normal distribution of the mean differences is weak a bootstrap simulation of the mean differences is calculated:

```
nsims = 1000 #the number of simulations
meaneighties = numeric(nsims) #hold the mean of each resampling from eighties ppm
meanteens = numeric(nsims) #hold the mean of each resampling from teens ppm
diffmeans = numeric(nsims) #hold the difference between the sample means
eighties = filter(DFN2, decade_range=="eighties")  #filters out all eighties ppm from data frame
teens = filter(DFN2, decade_range=="teens")  #filters out all teens ppm from data frame
head(eighties, 3)# will just check verify stripping out eighties
```

	X	Parameter	Station.Name	Northing	Easting	MO…	ppm_eighties	X.1	ppm.2015
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>	<lgl>	<dbl>
1	0	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Jan	0.04111032	NA	0.02910430
2	1	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Feb	0.04371964	NA	0.02173860
3	2	Nitrogen Dioxide	Calgary Central	51.04715	-114.0731	Mar	0.03987712	NA	0.01835591

3 rows | 1-10 of 13 columns

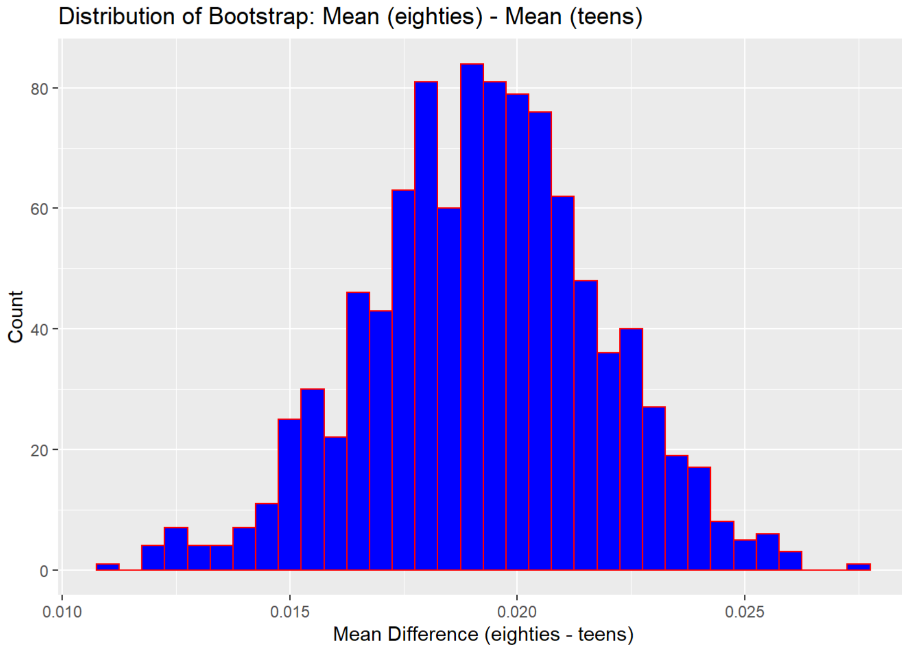
```
head(teens, 3) #vice versa
```

	X	Parameter	Station.Name	Northing	Easting	MONTH	ppm_eighties	X.1	ppm.2015
	<int>	<fctr>	<fctr>	<dbl>	<dbl>	<fctr>	<dbl>	<lgl>	<dbl>
1	NA			NA	NA		NA	NA	NA
2	NA			NA	NA		NA	NA	NA
3	NA			NA	NA		NA	NA	NA

3 rows | 1-10 of 13 columns

```
for(i in 1:nsims)
{   meaneighties[i] = mean(sample(eighties$ppm_N2,neighties, replace=TRUE))  #computes the mean of 12 resampled eighties ozone
    meanteens[i] = mean(sample(teens$ppm_N2,nteens,  replace=TRUE))  #computes the mean of 12 resampled teens ppm
    diffmeans[i] = meaneighties[i] - meanteens[i]  #computes the difference between the sample means
}
bootstrapN2 = data.frame(meaneighties, meanteens, diffmeans)  #create a data frame holding all the means
```

```
ggplot(data=bootstrapN2, aes(x = diffmeans)) + geom_histogram(fill='blue', col='red', binwidth=.0005) + xlab("Mean Difference (eighties - teens)") + ylab("Count") + ggtitle("Distribution of Bootstrap: Mean (eighties) - Mean (teens)")
```



The bootstrap distribution is approximately normally distributed.

```
favstats(bootstrapN2$diffmeans)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.01089324	0.01763069	0.01931336	0.02092554	0.02732461	0.01925912	0.002561722	1000	0

1 row

The mean of the bootstrap simulation is 0.01933 which compares closely to the difference of means test of 0.01935.

```
quantile(diffmeans, c(0.025, 0.975), data=bootstrapN2)
```

```
##       2.5%      97.5%
## 0.01421343 0.02407124
```

A mean difference of 0.01933 is derived from the bootstrap simulation with a 95% confidence interval of: 0.01490<= Mean Difference Ozone <= 0.02422

Nitrogen Dioxide levels have increased over the last forty years.

The values derived from the bootstrap simulation agree with the values calculated from the t-test. For instance, the mean difference from the t-test is 0.01935 compared to 0.01933 from the simulation. The 95% confidence intervals are:

0.016895 <= Mean Difference Nitrogen Dioxide<= 0.021813. t-test -0.01490<= Mean Difference Nitrogen Dioxide <= 0.02422 simulation

This apparently provides some justification in using the t-test.

Nitrogen Dioxide levels have decreased over the past 40 years.

DISCUSSION OF RESULTS:

Primarily, air pollutants in Calgary are created by vehicle exhaust. There are secondary sources from wood burning fireplaces and light industry. Three pollutants comprise the Air Quality Health Index: Particulate matter smaller than 2.5 microns; Ground Level Ozone; Nitrogen Dioxide. It is important to note that these levels are generally below threshold levels to trigger a Air Quality Alert. The threshold level for Nitrogen Dioxide is 0.159 ppm and typical levels over the past 5 years are around 0.02 to 0.03 ppm. For Ozone the threshold level to trigger an Air Quality Alert is 0.076 ppm and typical levels the last 5 years are around 0.03 ppm. Particulate Matter smaller than 2.5 microns can exceed the threshold level of 80 g/m3 when smoke from distant forest fires drifts into Calgary.

CONCLUSIONS:

The data demonstrates that there have been changes over the past forty years in air pollutants. Nitrogen Dioxide was demonstrated to have decreased over forty years by a mean of 0.024ppm. However, Ground Level Ozone has increased by a mean value 0.0075. The concentrations of these pollutants are not great enough to create a Air Quality Health Alert.

Further,it can be inferred that the average methane level IS Less when calculated using “Calibrated with Methane/Propane” method rather than the “Instrumental” method.

Last, we can conclude that Air Quality Health Index can be modeled as a positive linear function of Carbon Monoxide level.

Further Work:

An examination on why have Nitrogen Dioxide levels decreased over time while Ground Level Ozone have increased particularly, since these are mostly generated by vehicle traffic.