# Capstone Project - Perceived Mental Health

*J. Albanese*

*March 30, 2018*

Data will be analyzed in two sets - Perceived Mental Health, Very Good or Excellent (%) and Perceived Mental Health, Fair or Poor (%). Results will be analyzed in parallel to determine historic mental health trends.

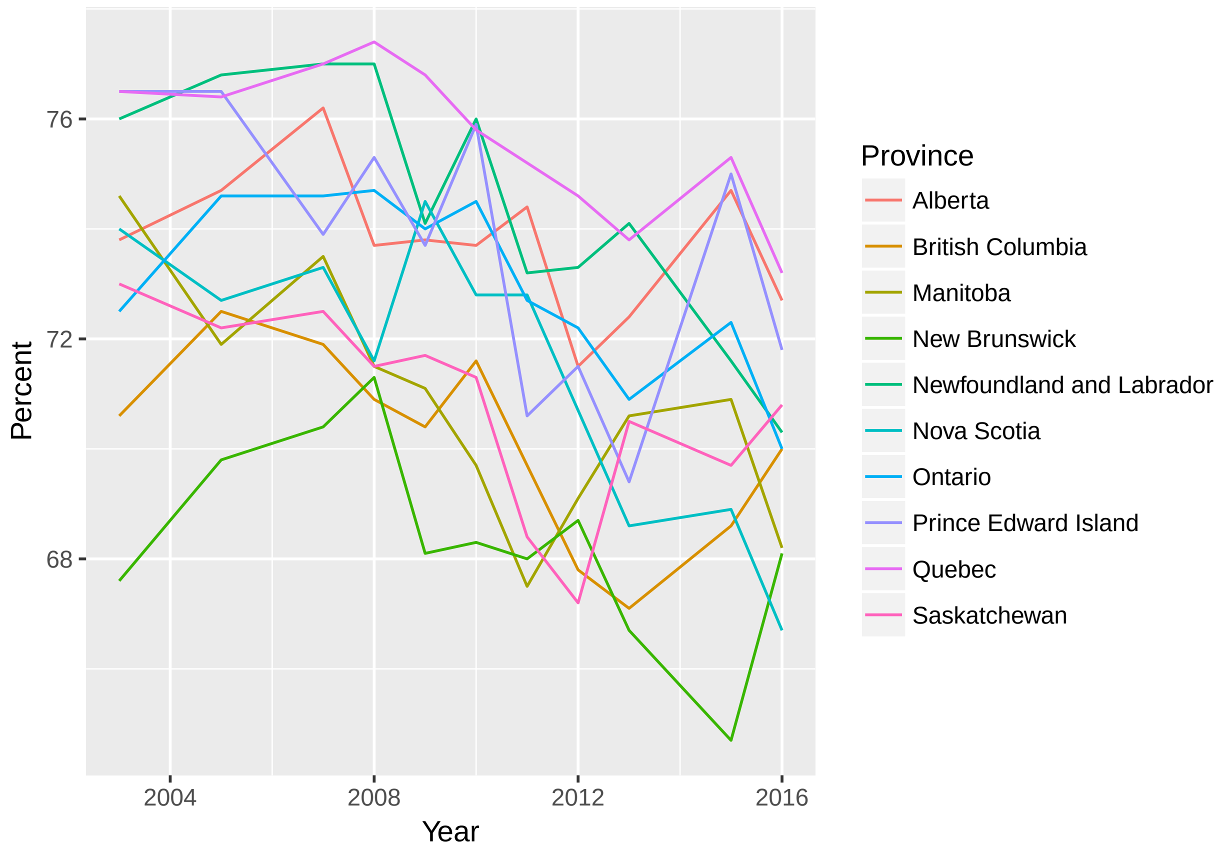
**bold** PERCEIVED MENTAL HEALTH, VERY GOOD OR EXCELLENT **bold**

Import Very Good or Excellent data to analyze and represent visually.

```
library('ggplot2')
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
Good<-read.csv('C:/Users/alba67300/Documents/ZOther/School/CKME136 - Data Analytics Capstone Project/CKM
ggplot(data=Good,aes(x=Year,y=Total),) + geom_line(aes(colour=Name)) + labs(y = "Percent") + labs(colou
```




Seperate into individual files by province to determine if a parametric or non-parametric analysis will be conducted. Evaluate normality of the individual datasets and determine if the variances of each dataset can be considered statistically equal to each other.

```
BCG<-Good[Good$Name=='British Columbia',]
AG<-Good[Good$Name=='Alberta',]
SG<-Good[Good$Name=='Saskatchewan',]
```

```
MG<-Good[Good$Name=='Manitoba',]
OG<-Good[Good$Name=='Ontario',]
QG<-Good[Good$Name=='Quebec',]
NBG<-Good[Good$Name=='New Brunswick',]
NSG<-Good[Good$Name=='Nova Scotia',]
PEIG<-Good[Good$Name=='Prince Edward Island',]
NG<-Good[Good$Name=='Newfoundland and Labrador',]
```

Perform Shapiro-Wilk test to determine normality. Ho = data is normally distributed Ha = data is not normally distributed alpha = 0.05

```
shapiro.test(BCG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BCG$Total
## W = 0.96154, p-value = 0.7907
```

```
shapiro.test(AG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  AG$Total
## W = 0.96403, p-value = 0.8208
```

```
shapiro.test(SG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SG$Total
## W = 0.93044, p-value = 0.4153
```

```
shapiro.test(MG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  MG$Total
## W = 0.97843, p-value = 0.9568
```

```
shapiro.test(OG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OG$Total
## W = 0.88913, p-value = 0.1356
```

```
shapiro.test(QG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  QG$Total
## W = 0.94826, p-value = 0.6217
```

```r
shapiro.test(NBG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NBG$Total
## W = 0.95754, p-value = 0.7406
```

```r
shapiro.test(NSG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NSG$Total
## W = 0.91383, p-value = 0.2705
```

```r
shapiro.test(PEIG$Total)
```

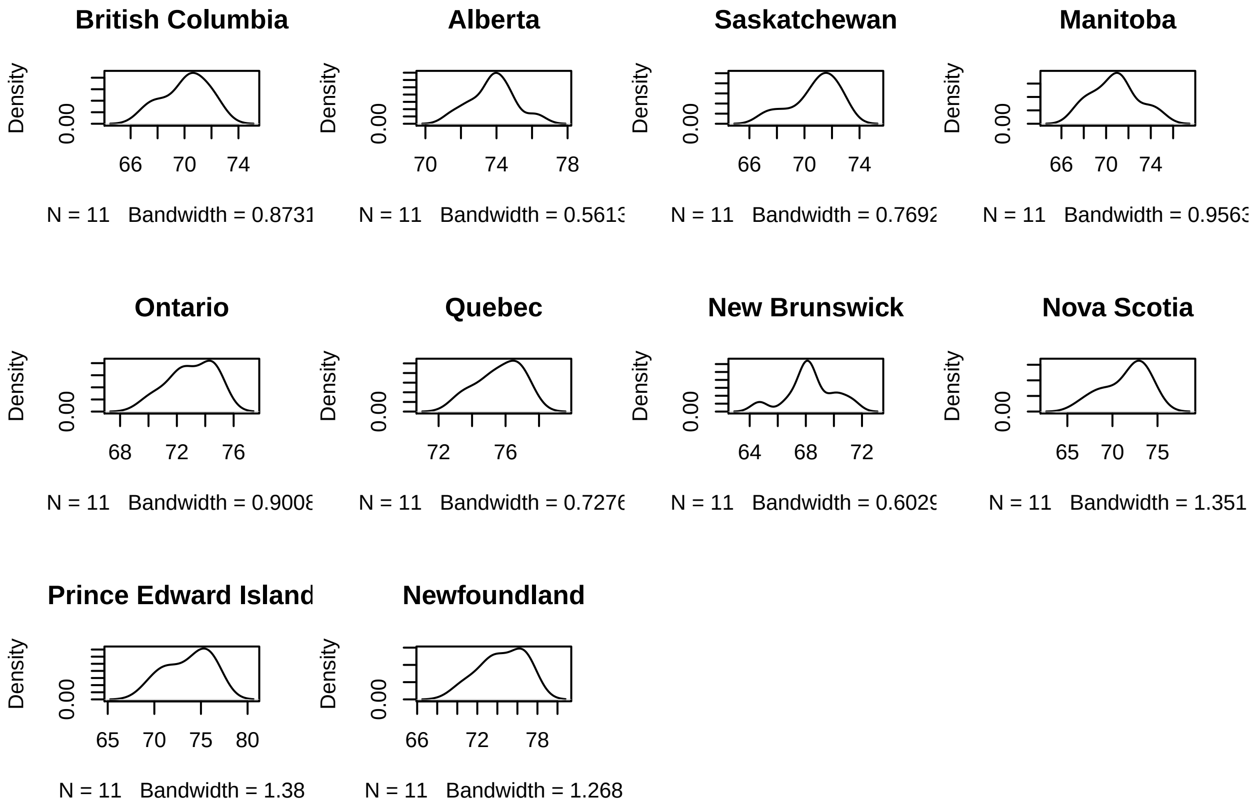
```
##
##  Shapiro-Wilk normality test
##
## data:  PEIG$Total
## W = 0.92179, p-value = 0.3338
```

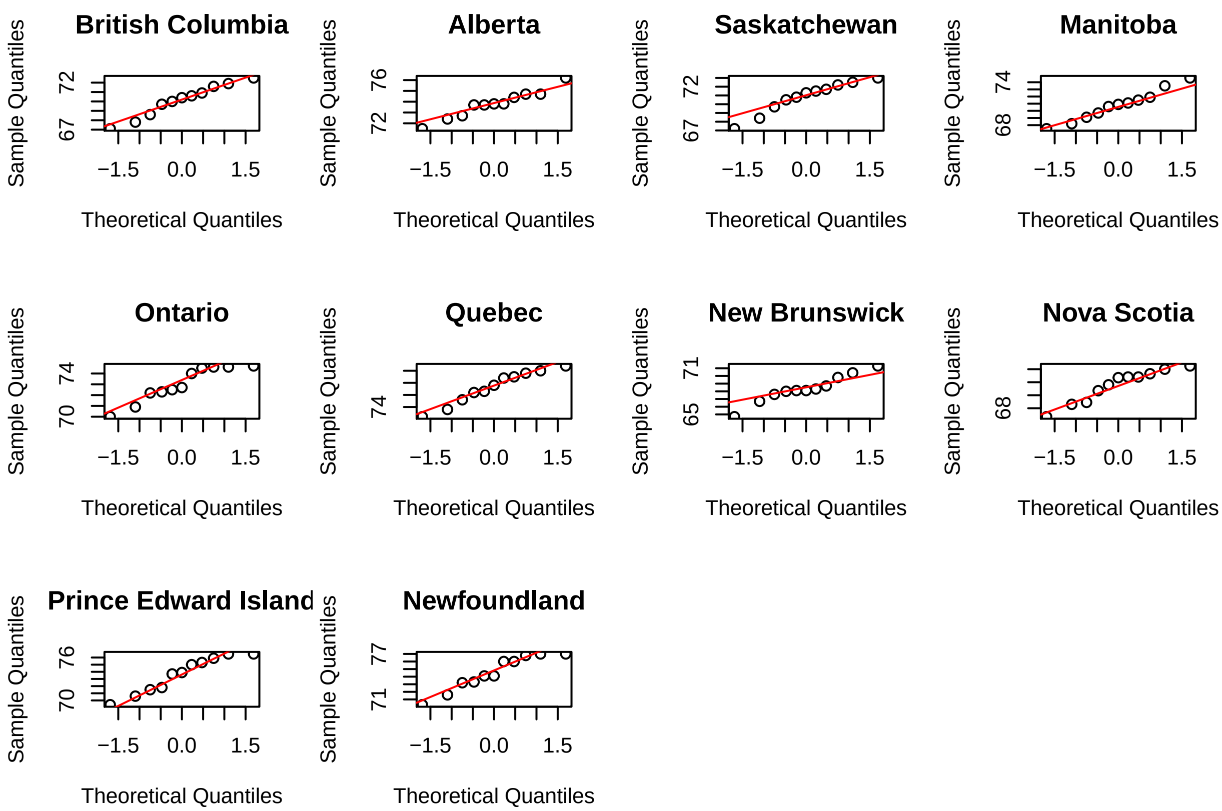
```r
shapiro.test(NG$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NG$Total
## W = 0.91275, p-value = 0.2628
```

Based on p-values for each provincial data set as calculated using the Shapiro-Wilk test, each null hypothesis can't be rejected and all datasets are assumed to be normally distributed.

Graphically represent the data to visually confirm data is sufficiently normally distributed.


British Columbia
Density
0.00
66 70 74
N = 11 Bandwidth = 0.8731
Alberta
Density
0.00
70 74 78
N = 11 Bandwidth = 0.561
Saskatchewan
Density
0.00
66 70 74
N = 11 Bandwidth = 0.769
Manitoba
Density
0.00
66 70 74
N = 11 Bandwidth = 0.956
Ontario
Density
0.00
68 72 76
N = 11 Bandwidth = 0.900
Quebec
Density
0.00
72 76
N = 11 Bandwidth = 0.727
New Brunswick
Density
0.00
64 68 72
N = 11 Bandwidth = 0.602
Nova Scotia
Density
0.00
65 70 75
N = 11 Bandwidth = 1.351
Prince Edward Island
Density
0.00
65 70 75 80
N = 11 Bandwidth = 1.38
Newfoundland
Density
0.00
66 72 78
N = 11 Bandwidth = 1.268




Confirm if dataset variances can be considered equal. Calculate variances of each dataset, determine if variances create a normally distributed dataset and compare, in turn, each variance to the mean of the remaining dataset using an independent two-tailed t-distribution (as it is a small sample size). Ho = variance is equal to mean Ha = variance is not equal to mean alpha = 0.05

```
GoodVars<-c(var(BCG$Total),var(AG$Total),var(SG$Total),var(MG$Total),var(OG$Total),var(QG$Total),var(NB
t.test(GoodVars[-1],mu=GoodVars[1])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-1]
## t = 1.6009, df = 8, p-value = 0.1481
## alternative hypothesis: true mean is not equal to 2.894
## 95 percent confidence interval:
##  2.482403 5.174688
## sample estimates:
## mean of x
##  3.828545
```

```
t.test(GoodVars[-2],mu=GoodVars[2])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-2]
## t = 4.4161, df = 8, p-value = 0.002238
## alternative hypothesis: true mean is not equal to 1.621636
```

```
## 95 percent confidence interval:
##  2.743704 5.196135
## sample estimates:
## mean of x
##  3.969919
```

```r
t.test(GoodVars[-3],mu=GoodVars[3])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-3]
## t = 1.1506, df = 8, p-value = 0.2831
## alternative hypothesis: true mean is not equal to 3.126
## 95 percent confidence interval:
##  2.446409 5.159126
## sample estimates:
## mean of x
##  3.802768
```

```r
t.test(GoodVars[-4],mu=GoodVars[4])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-4]
## t = -1.5136, df = 8, p-value = 0.1686
## alternative hypothesis: true mean is not equal to 4.531636
## 95 percent confidence interval:
##  2.298222 4.994950
## sample estimates:
## mean of x
##  3.646586
```

```r
t.test(GoodVars[-5],mu=GoodVars[5])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-5]
## t = 2.1609, df = 8, p-value = 0.0627
## alternative hypothesis: true mean is not equal to 2.614
## 95 percent confidence interval:
##  2.530366 5.188947
## sample estimates:
## mean of x
##  3.859657
```

```r
t.test(GoodVars[-6],mu=GoodVars[6])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-6]
## t = 3.8769, df = 8, p-value = 0.004695
## alternative hypothesis: true mean is not equal to 1.836545
## 95 percent confidence interval:
```

```
##  2.691305 5.200775
## sample estimates:
## mean of x
##   3.94604
```

```r
t.test(GoodVars[-7],mu=GoodVars[7])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-7]
## t = 1.0117, df = 8, p-value = 0.3413
## alternative hypothesis: true mean is not equal to 3.198545
## 95 percent confidence interval:
##  2.435837 5.153577
## sample estimates:
## mean of x
##  3.794707
```

```r
t.test(GoodVars[-8],mu=GoodVars[8])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-8]
## t = -5.4434, df = 8, p-value = 0.0006135
## alternative hypothesis: true mean is not equal to 6.216909
## 95 percent confidence interval:
##  2.291132 4.627535
## sample estimates:
## mean of x
##  3.459333
```

```r
t.test(GoodVars[-9],mu=GoodVars[9])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-9]
## t = -5.195, df = 8, p-value = 0.0008274
## alternative hypothesis: true mean is not equal to 6.132727
## 95 percent confidence interval:
##  2.286146 4.651228
## sample estimates:
## mean of x
##  3.468687
```

```r
t.test(GoodVars[-10],mu=GoodVars[10])
```

```
##
##  One Sample t-test
##
## data:  GoodVars[-10]
## t = -2.8381, df = 8, p-value = 0.02188
## alternative hypothesis: true mean is not equal to 5.178909
## 95 percent confidence interval:
##  2.271178 4.878156
```

```
## sample estimates:
## mean of x
##  3.574667
```

Looking at the results of the t-tests for the variances, there are five instances were the null hypothesis can be rejected because the p-value is less than the alpha value of 0.05. These rejections indicate that the variances are not statistically equal so a non-parametric test is required to analyze the original dataset.

A Friedman test will be performed to compare the ten (10) different populations and determine if at least two (2) of the ten (10) distributions differ. The years are the blocks (b) and the provinces are the treatments (k). Ho = Provincial data sets are all the same Ha = at least two of the dataset distributions differ alpha = 0.05

```r
GoodDF<-data.frame(Year=as.factor(Good$Year),Province=Good$Name,Perc=Good$Total)
friedman.test(Perc~Province|Year, data=GoodDF)
```

```
##
##  Friedman rank sum test
##
## data:  Perc and Province and Year
## Friedman chi-squared = 74.456, df = 9, p-value = 2.023e-12
```

Based on the results of the Friedman test, at least two (2) of the Provincial populations differ. In order to determine which, posthoc analysis using the Nemenyi method will be used.

```r
library('PMCMR')
```

```
## Warning: package 'PMCMR' was built under R version 3.3.3
```

```
## PMCMR is superseded by PMCMRplus and will be no longer maintained. You may wish to install PMCMRplus
```

```r
posthoc.friedman.nemenyi.test(Perc~Province|Year,data=GoodDF)
```

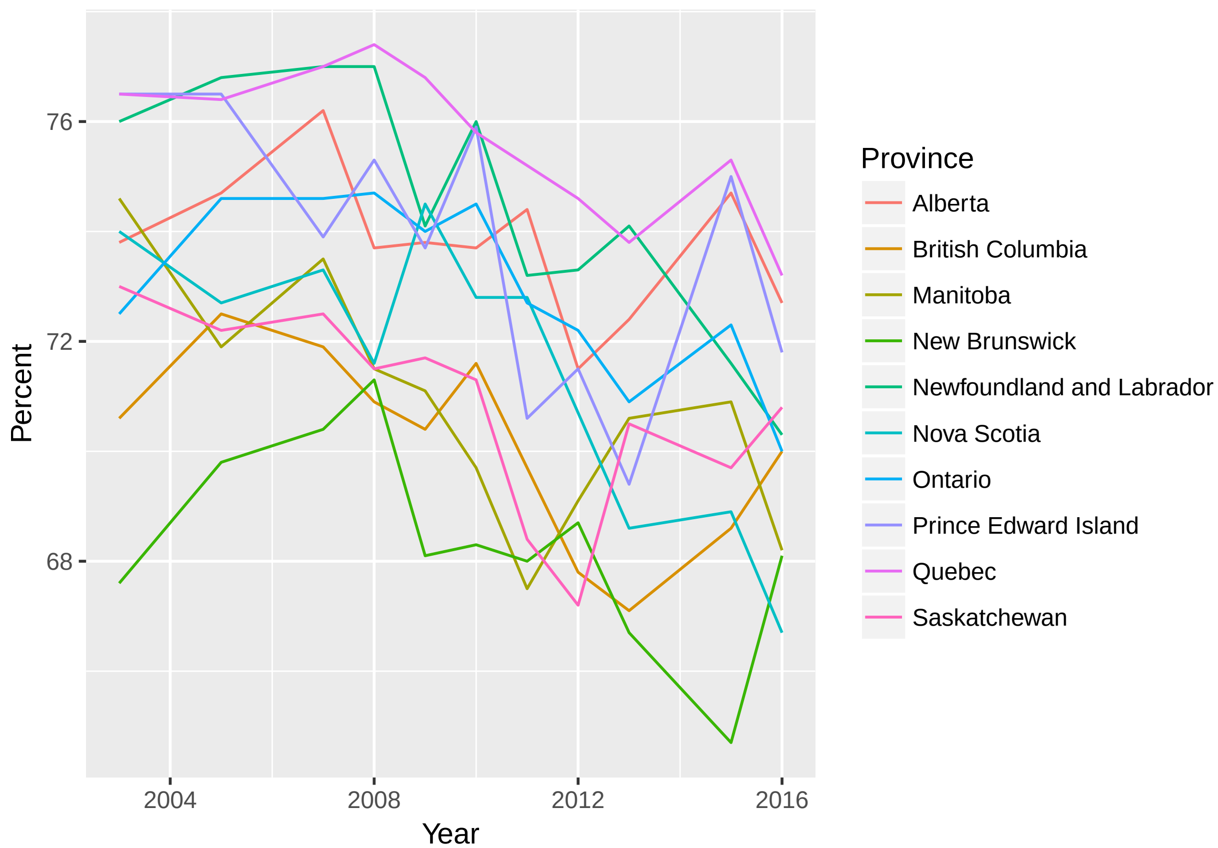
```
##
##  Pairwise comparisons using Nemenyi multiple comparison test
##              with q approximation for unreplicated blocked data
##
## data:  Perc and Province and Year
##
##                           Alberta British Columbia Manitoba New Brunswick
## British Columbia          0.01993 -                -        -
## Manitoba                  0.21480 0.99785          -        -
## New Brunswick             0.00046 0.99473          0.73825  -
## Newfoundland and Labrador 0.98862 0.00028          0.00945  2.2e-06
## Nova Scotia               0.73825 0.82073          0.99846  0.21480
## Ontario                   0.99979 0.13066          0.61913  0.00638
## Prince Edward Island      1.00000 0.01767          0.19868  0.00039
## Quebec                    0.73825 6.9e-06          0.00046  2.6e-08
## Saskatchewan              0.18340 0.99892          1.00000  0.78130
##                           Newfoundland and Labrador Nova Scotia Ontario
## British Columbia          -                         -           -
## Manitoba                  -                         -           -
## New Brunswick             -                         -           -
## Newfoundland and Labrador -                         -           -
## Nova Scotia               0.11947                   -           -
## Ontario                   0.80150                   0.97790     -
## Prince Edward Island      0.99108                   0.71556     0.99967
## Quebec                    0.99925                   0.01220     0.30805
## Saskatchewan              0.00728                   0.99706     0.56891
```

```
##                             Prince Edward Island Quebec
## British Columbia            -                    -
## Manitoba                    -                    -
## New Brunswick               -                    -
## Newfoundland and Labrador -                    -
## Nova Scotia                 -                    -
## Ontario                     -                    -
## Prince Edward Island        -                    -
## Quebec                      0.76019              -
## Saskatchewan                0.16898              0.00033
##
## P value adjustment method: none
```

**bold** PERCEIVED MENTAL HEALTH, FAIR OR POOR **bold**

Import Fair or Poor data to analyze and represent visually.

```
Poor<-read.csv('C:/Users/alba67300/Documents/ZOther/School/CKME136 - Data Analytics Capstone Project/CKI
ggplot(data=Good,aes(x=Year,y=Total),) + geom_line(aes(colour=Name)) + labs(y = "Percent") + labs(colou
```




Seperate into individual files by province to determine if a parametric or non-parametric analysis will be conducted. Evaluate normality of the individual datasets and determine if the variances of each dataset can be considered statistically equal to each other.

```
BCP<-Poor[Poor$Name=='British Columbia',]
AP<-Poor[Poor$Name=='Alberta',]
SP<-Poor[Poor$Name=='Saskatchewan',]
MP<-Poor[Poor$Name=='Manitoba',]
```

```r
OP<-Poor[Poor$Name=='Ontario',]
QP<-Poor[Poor$Name=='Quebec',]
NBP<-Poor[Poor$Name=='New Brunswick',]
NSP<-Poor[Poor$Name=='Nova Scotia',]
PEIP<-Poor[Poor$Name=='Prince Edward Island',]
NP<-Poor[Poor$Name=='Newfoundland and Labrador',]
```

Perform Shapiro-Wilk test to determine normality. Ho = data is normally distributed Ha = data is not normally distributed alpha = 0.05

```r
shapiro.test(BCP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BCP$Total
## W = 0.85055, p-value = 0.04338
```

```r
shapiro.test(AP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  AP$Total
## W = 0.83767, p-value = 0.02944
```

```r
shapiro.test(SP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SP$Total
## W = 0.8893, p-value = 0.1363
```

```r
shapiro.test(MP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  MP$Total
## W = 0.96436, p-value = 0.8246
```

```r
shapiro.test(OP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OP$Total
## W = 0.94393, p-value = 0.5678
```

```r
shapiro.test(QP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  QP$Total
## W = 0.89924, p-value = 0.1809
```

```r
shapiro.test(NBP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NBP$Total
## W = 0.92138, p-value = 0.3303
```

```r
shapiro.test(NSP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NSP$Total
## W = 0.93877, p-value = 0.5062
```

```r
shapiro.test(PEIP$Total)
```

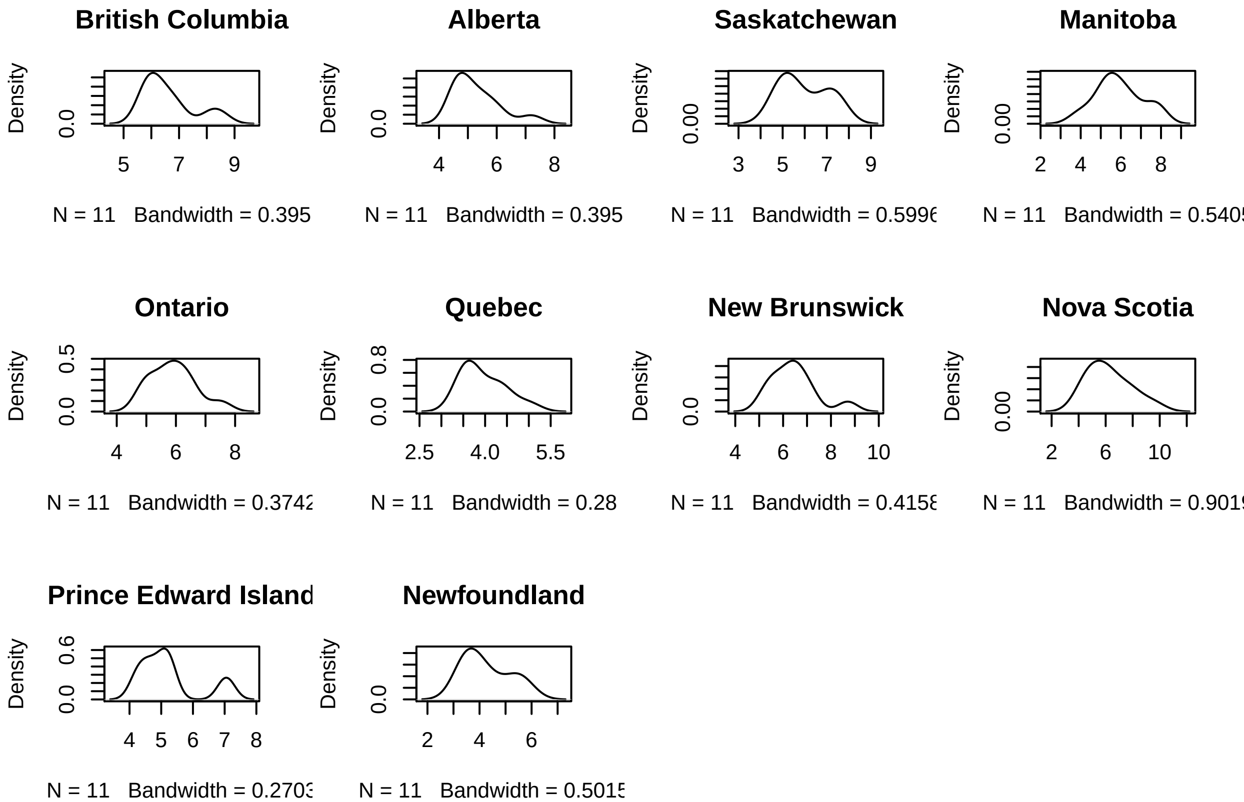
```
##
##  Shapiro-Wilk normality test
##
## data:  PEIP$Total
## W = 0.81838, p-value = 0.01642
```

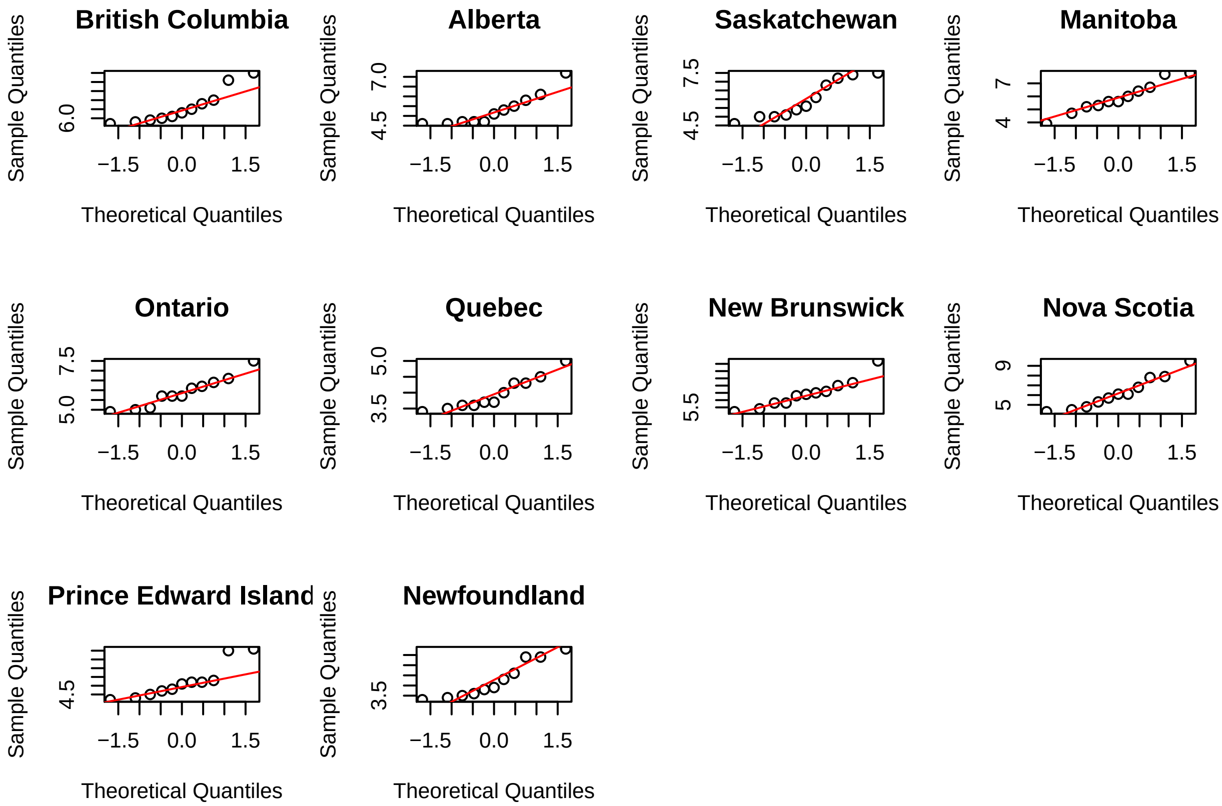
```r
shapiro.test(NP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NP$Total
## W = 0.88035, p-value = 0.1051
```

Based on p-values for each provincial data set as calculated using the Shapiro-Wilk test, not all data sets are normally distributed so non-parametric statistical comparisons will be used.

Graphically represent the data to visually confirm not all data is sufficiently normally distributed.


British Columbia
Density
0.0
5
7
9
N = 11 Bandwidth = 0.395
Alberta
Density
0.0
4
6
8
N = 11 Bandwidth = 0.395
Saskatchewan
Density
0.00
3
5
7
9
N = 11 Bandwidth = 0.599
Manitoba
Density
0.00
2
4
6
8
N = 11 Bandwidth = 0.540
Ontario
Density
0.0
0.5
4
6
8
N = 11 Bandwidth = 0.374
Quebec
Density
0.0
0.8
2.5
4.0
5.5
N = 11 Bandwidth = 0.28
New Brunswick
Density
0.0
4
6
8
10
N = 11 Bandwidth = 0.415
Nova Scotia
Density
0.00
2
6
10
N = 11 Bandwidth = 0.901
Prince Edward Island
Density
0.0
0.6
4
5
6
7
8
N = 11 Bandwidth = 0.270
Newfoundland
Density
0.0
2
4
6
N = 11 Bandwidth = 0.501




A Friedman test will be performed to compare the ten (10) different populations and determine if at least two (2) of the ten (10) distributions differ. The years are the blocks (b) and the provinces are the treatments (k). Ho = Provincial data sets are all the same Ha = at least two of the dataset distributions differ alpha = 0.05

```
PoorDF<-data.frame(Year=as.factor(Poor$Year),Province=Poor$Name,Perc=Poor$Total)
friedman.test(Perc~Province|Year, data=PoorDF)
```

```
##
##  Friedman rank sum test
##
## data:  Perc and Province and Year
## Friedman chi-squared = 62.706, df = 9, p-value = 4.022e-10
```

Based on the results of the Friedman test, at least two (2) of the Provincial populations differ. In order to determine which, posthoc analysis using the Nemenyi method will be used.

```
posthoc.friedman.nemenyi.test(Perc~Province|Year,data=PoorDF)
```

```
##
##  Pairwise comparisons using Nemenyi multiple comparison test
##              with q approximation for unreplicated blocked data
##
## data:  Perc and Province and Year
##
##                           Alberta British Columbia Manitoba New Brunswick
## British Columbia          0.02523 -                -        -
## Manitoba                  0.92506 0.61913          -        -
## New Brunswick             0.26828 0.99706          0.98564  -
```

```
## Newfoundland and Labrador 0.61913 3.9e-06         0.02523  0.00033
## Nova Scotia               0.66828 0.90084         0.99997  0.99987
## Ontario                   0.83894 0.76019         1.00000  0.99706
## Prince Edward Island      1.00000 0.02244         0.91350  0.24962
## Quebec                    0.37330 5.2e-07         0.00728  6.2e-05
## Saskatchewan              0.83894 0.76019         1.00000  0.99706
##                           Newfoundland and Labrador Nova Scotia Ontario
## British Columbia          -                         -           -
## Manitoba                  -                         -           -
## New Brunswick             -                         -           -
## Newfoundland and Labrador -                         -           -
## Nova Scotia               0.00424                   -           -
## Ontario                   0.01220                   1.00000     -
## Prince Edward Island      0.64389                   0.64389     0.82073
## Quebec                    1.00000                   0.00099     0.00321
## Saskatchewan              0.01220                   1.00000     1.00000
##                           Prince Edward Island Quebec
## British Columbia          -                    -
## Manitoba                  -                    -
## New Brunswick             -                    -
## Newfoundland and Labrador -                    -
## Nova Scotia               -                    -
## Ontario                   -                    -
## Prince Edward Island      -                    -
## Quebec                    0.39635              -
## Saskatchewan              0.82073              0.00321
##
## P value adjustment method: none
```

Lastly, the data sets will be analyzed to determine if there has been a significant increase or decrease in perceived mental health over the last 13 years.

```
G2003<-Good[Good$Year==2003,]
G2016<-Good[Good$Year==2016,]
P2003<-Poor[Poor$Year==2003,]
P2016<-Poor[Poor$Year==2016,]
shapiro.test(G2003$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  G2003$Total
## W = 0.91117, p-value = 0.2891
```

```
shapiro.test(G2016$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  G2016$Total
## W = 0.9638, p-value = 0.8282
```

```
shapiro.test(P2003$Total)
```

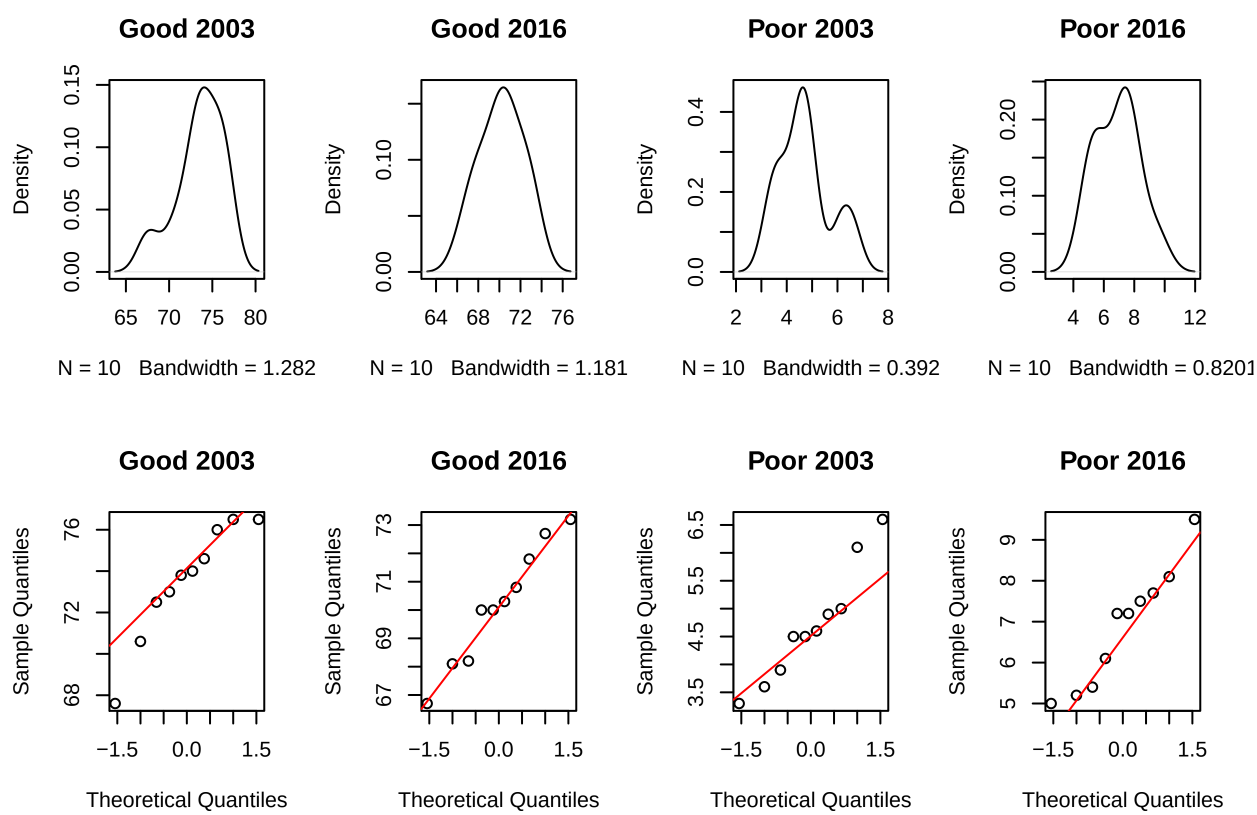
```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  P2003$Total
## W = 0.94037, p-value = 0.5572
```

```r
shapiro.test(P2016$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  P2016$Total
## W = 0.94013, p-value = 0.5545
```

```r
par(mfrow=c(2,4))
plot(density(G2003$Total), main = "Good 2003")
plot(density(G2016$Total), main = "Good 2016")
plot(density(P2003$Total), main = "Poor 2003")
plot(density(P2016$Total), main = "Poor 2016")
qqnorm(G2003$Total, main = "Good 2003")
qqline(G2003$Total,col=2)
qqnorm(G2016$Total, main = "Good 2016")
qqline(G2016$Total,col=2)
qqnorm(P2003$Total, main = "Poor 2003")
qqline(P2003$Total,col=2)
qqnorm(P2016$Total, main = "Poor 2016")
qqline(P2016$Total,col=2)
```




Both the Perceived Mental Health Good and Poor show that they are normally distributed so variance calculations will be performed to determine if parametric or non-parametric tests should be performed.

```r
var(G2003$Total)
```

```
## [1] 7.807667
```

```r
var(G2016$Total)
```

```
## [1] 4.324
```

```r
var(P2003$Total)
```

```
## [1] 1.066667
```

```r
var(P2016$Total)
```

```
## [1] 2.085444
```

Based on that none of the variances are equal, non-parametric comparisons will be used on the Good and Poor data sets.

```r
wilcox.test(G2003$Total,G2016$Total,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  G2003$Total and G2016$Total
## V = 54, p-value = 0.003906
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(P2003$Total,P2016$Total,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  P2003$Total and P2016$Total
## V = 0, p-value = 0.001953
## alternative hypothesis: true location shift is not equal to 0
```