Capstone Project - Perceived Mental Health

J. Albanese

April 9, 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.

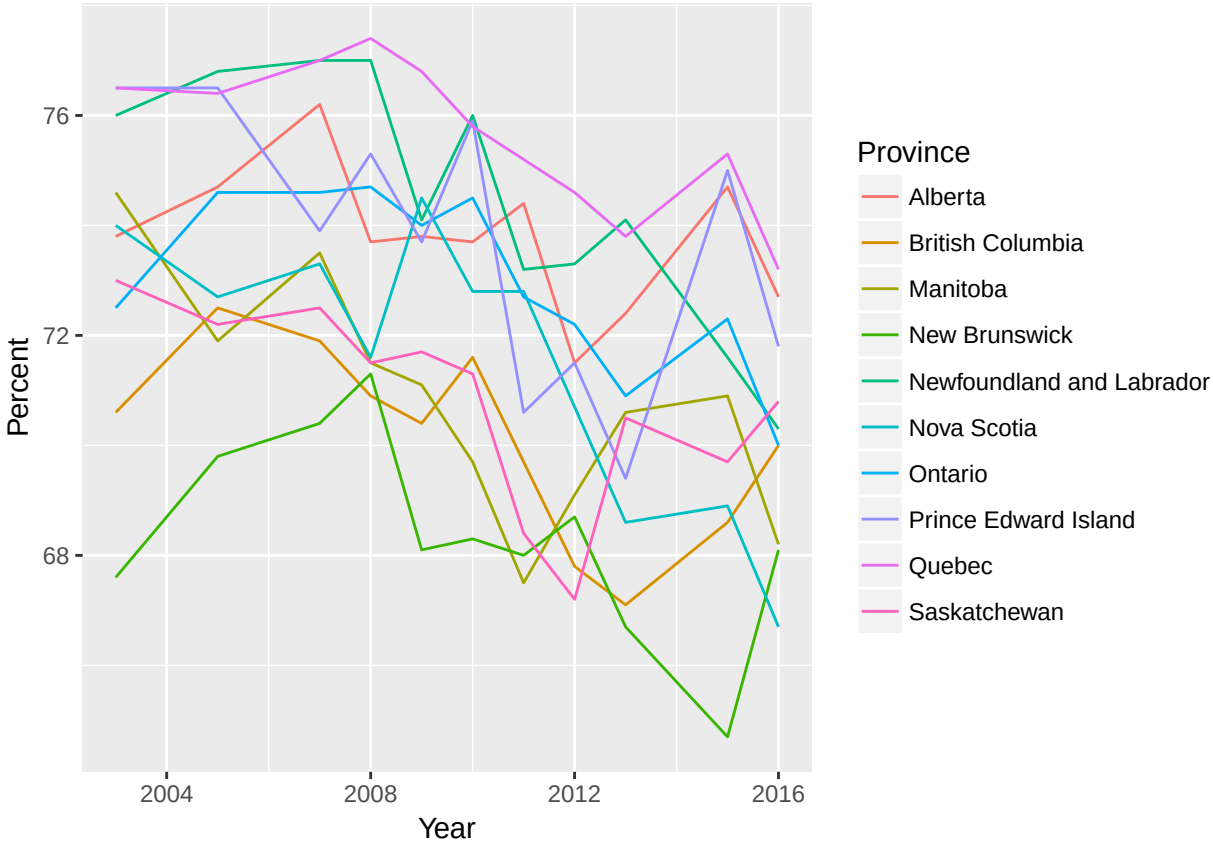
bold PERCEIVED MENTAL HEALTH, VERY GOOD OR EXCELLENT **bold**

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

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

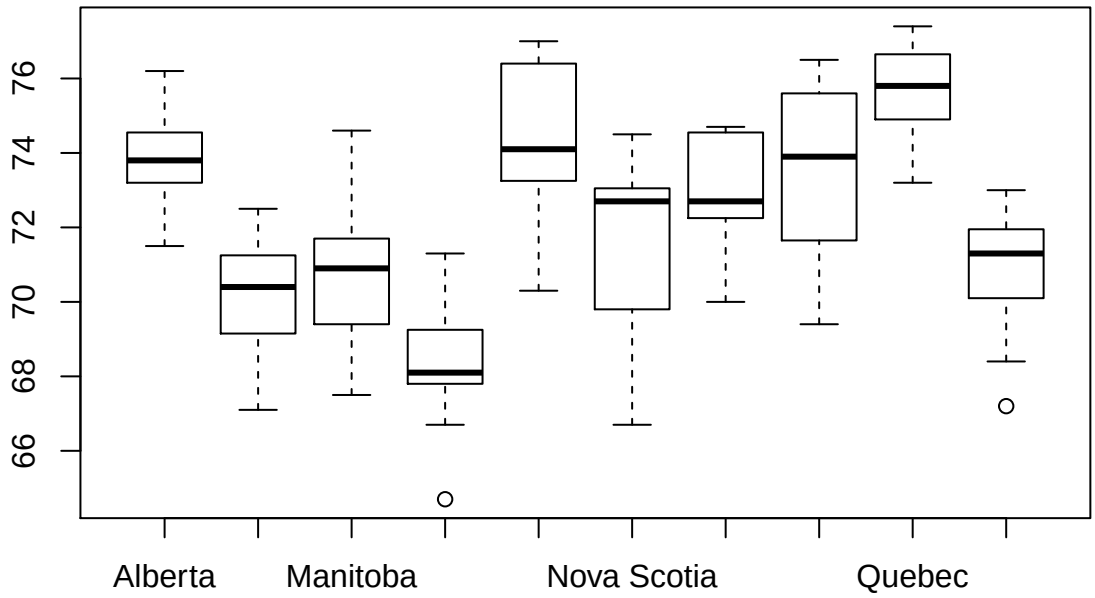
```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

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



First, must review and eliminate possible outliers from each data set.

```
bxplt_Good<-boxplot(Total~Name,data=Good)
```

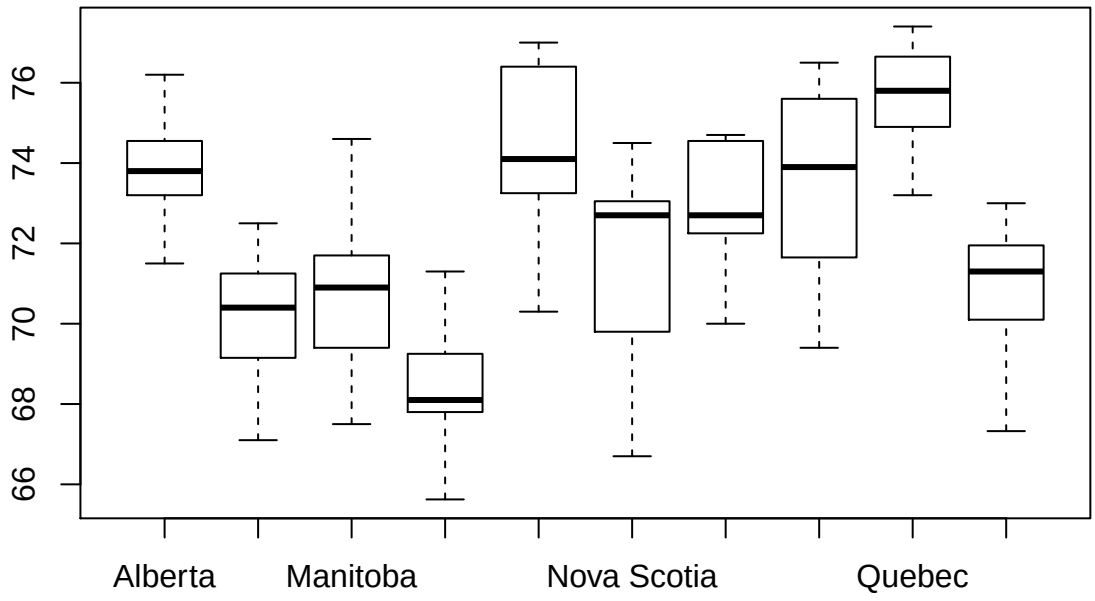



```
bxplt_Good
```

```
## $stats
##       [,1]  [,2] [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
## [1,] 71.50 67.10 67.5 66.70 70.30 66.70 70.00 69.40 73.20 68.40
## [2,] 73.20 69.15 69.4 67.80 73.25 69.80 72.25 71.65 74.90 70.10
## [3,] 73.80 70.40 70.9 68.10 74.10 72.70 72.70 73.90 75.80 71.30
## [4,] 74.55 71.25 71.7 69.25 76.40 73.05 74.55 75.60 76.65 71.95
## [5,] 76.20 72.50 74.6 71.30 77.00 74.50 74.70 76.50 77.40 73.00
##
## $n
##  [1] 11 11 11 11 11 11 11 11 11 11
##
## $conf
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 73.15688 69.39959 69.80431 67.40924 72.59938 71.15174 71.60431
## [2,] 74.44312 71.40041 71.99569 68.79076 75.60062 74.24826 73.79569
##          [,8]     [,9]    [,10]
## [1,] 72.01827 74.96632 70.41868
## [2,] 75.78173 76.63368 72.18132
##
## $out
## [1] 64.7 67.2
##
## $group
## [1]  4 10
```

```
##
## $names
##  [1] "Alberta"                   "British Columbia"
##  [3] "Manitoba"                  "New Brunswick"
##  [5] "Newfoundland and Labrador" "Nova Scotia"
##  [7] "Ontario"                   "Prince Edward Island"
##  [9] "Quebec"                    "Saskatchewan"
```

```
G<-subset(Good,select=-c(Code))
G$Total[(G$Total==64.7&G$Name == "New Brunswick")] <- quantile(G$Total[G$Name=="New Brunswick"],0.25)-1
G$Total[(G$Total==67.2 & G$Name == "Saskatchewan")] <- quantile(G$Total[G$Name=="Saskatchewan"],0.25)-1
boxplot(Total~Name,data=G)
```



Separate 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<-G[G$Name=='British Columbia',]
AG<-G[G$Name=='Alberta',]
SG<-G[G$Name=='Saskatchewan',]
MG<-G[G$Name=='Manitoba',]
OG<-G[G$Name=='Ontario',]
QG<-G[G$Name=='Quebec',]
NBG<-G[G$Name=='New Brunswick',]
NSG<-G[G$Name=='Nova Scotia',]
PEIG<-G[G$Name=='Prince Edward Island',]
NG<-G[G$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 = 0.93288, p-value = 0.4407
```

```
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
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  NBG$Total
## W = 0.96643, p-value = 0.8484
```

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

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

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

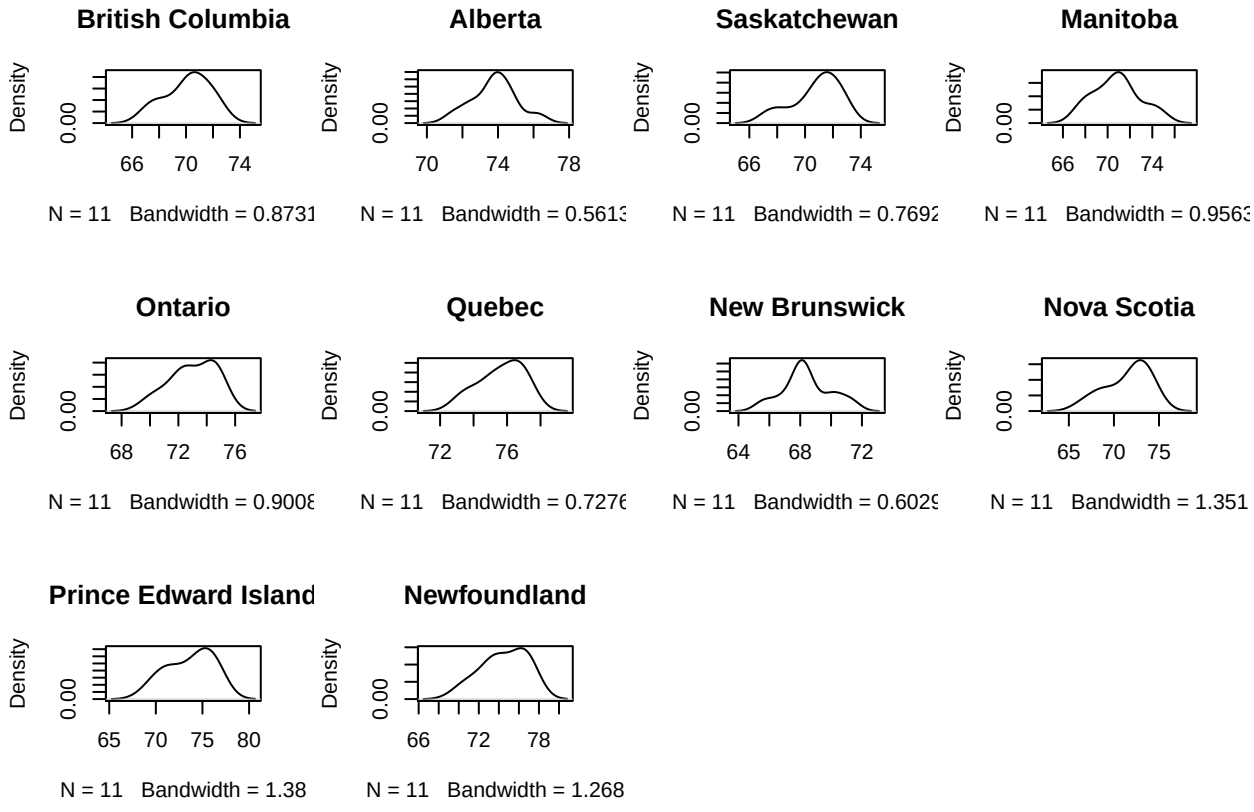
```
##
##  Shapiro-Wilk normality test
##
## data:  PEIG$Total
## W = 0.92179, p-value = 0.3338
```

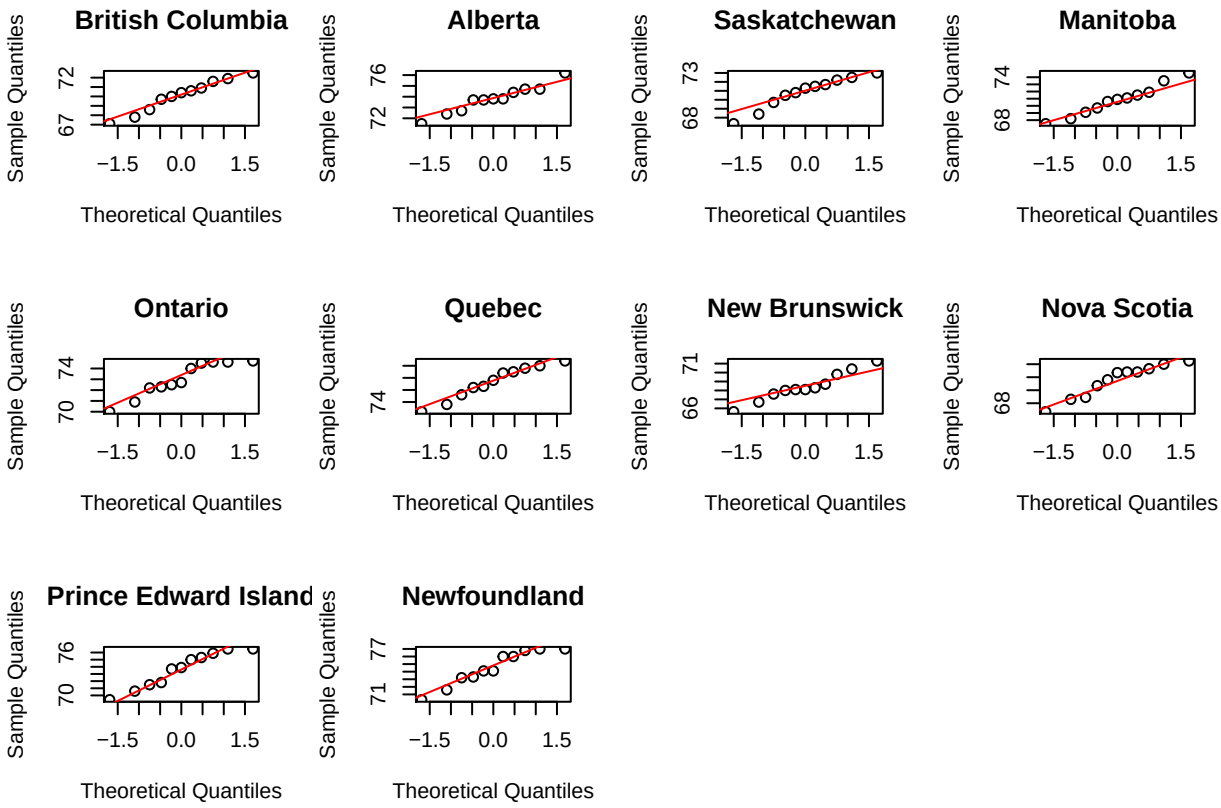
```
shapiro.test(NG$Total)
```

```
##
##  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.





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 = 1.4366, df = 8, p-value = 0.1888
## alternative hypothesis: true mean is not equal to 2.894
## 95 percent confidence interval:
##  2.374359 5.130838
## sample estimates:
## mean of x
##  3.752598
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-2]
## t = 4.1357, df = 8, p-value = 0.003273
## alternative hypothesis: true mean is not equal to 1.621636
```

```
## 95 percent confidence interval:
##  2.626959 5.160986
## sample estimates:
## mean of x
##  3.893972
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-3]
## t = 1.1649, df = 8, p-value = 0.2776
## alternative hypothesis: true mean is not equal to 3.03742
## 95 percent confidence interval:
##  2.352450 5.120875
## sample estimates:
## mean of x
##  3.736663
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-4]
## t = -1.6132, df = 8, p-value = 0.1454
## alternative hypothesis: true mean is not equal to 4.531636
## 95 percent confidence interval:
##  2.196900 4.944378
## sample estimates:
## mean of x
##  3.570639
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-5]
## t = 1.9791, df = 8, p-value = 0.08317
## alternative hypothesis: true mean is not equal to 2.614
## 95 percent confidence interval:
##  2.420772 5.146647
## sample estimates:
## mean of x
##   3.78371
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-6]
## t = 3.6248, df = 8, p-value = 0.006736
## alternative hypothesis: true mean is not equal to 1.836545
## 95 percent confidence interval:
```

```
##  2.576390 5.163796
## sample estimates:
## mean of x
##  3.870093
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-7]
## t = 1.9996, df = 8, p-value = 0.08057
## alternative hypothesis: true mean is not equal to 2.603602
## 95 percent confidence interval:
##  2.42259 5.14714
## sample estimates:
## mean of x
##  3.784865
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-8]
## t = -5.4933, df = 8, p-value = 0.0005783
## alternative hypothesis: true mean is not equal to 6.216909
## 95 percent confidence interval:
##  2.193907 4.572866
## sample estimates:
## mean of x
##  3.383386
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-9]
## t = -5.248, df = 8, p-value = 0.0007757
## alternative hypothesis: true mean is not equal to 6.132727
## 95 percent confidence interval:
##  2.188782 4.596698
## sample estimates:
## mean of x
##   3.39274
```

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

```
##
##  One Sample t-test
##
## data:  GoodVars[-10]
## t = -2.9198, df = 8, p-value = 0.0193
## alternative hypothesis: true mean is not equal to 5.178909
## 95 percent confidence interval:
##  2.171732 4.825707
```

```
## sample estimates:
## mean of x
##   3.49872
```

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 different populations and determine if at least two 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

```
GoodDF<-data.frame(Year=as.factor(G$Year),Province=G$Name,Perc=G$Total,Item=G$Item)

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, where the p-value is less than the alpha value, at least two (2) of the Provincial populations differ. In order to determine which, posthoc analysis using the Nemenyi method will be used.

```
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
```

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

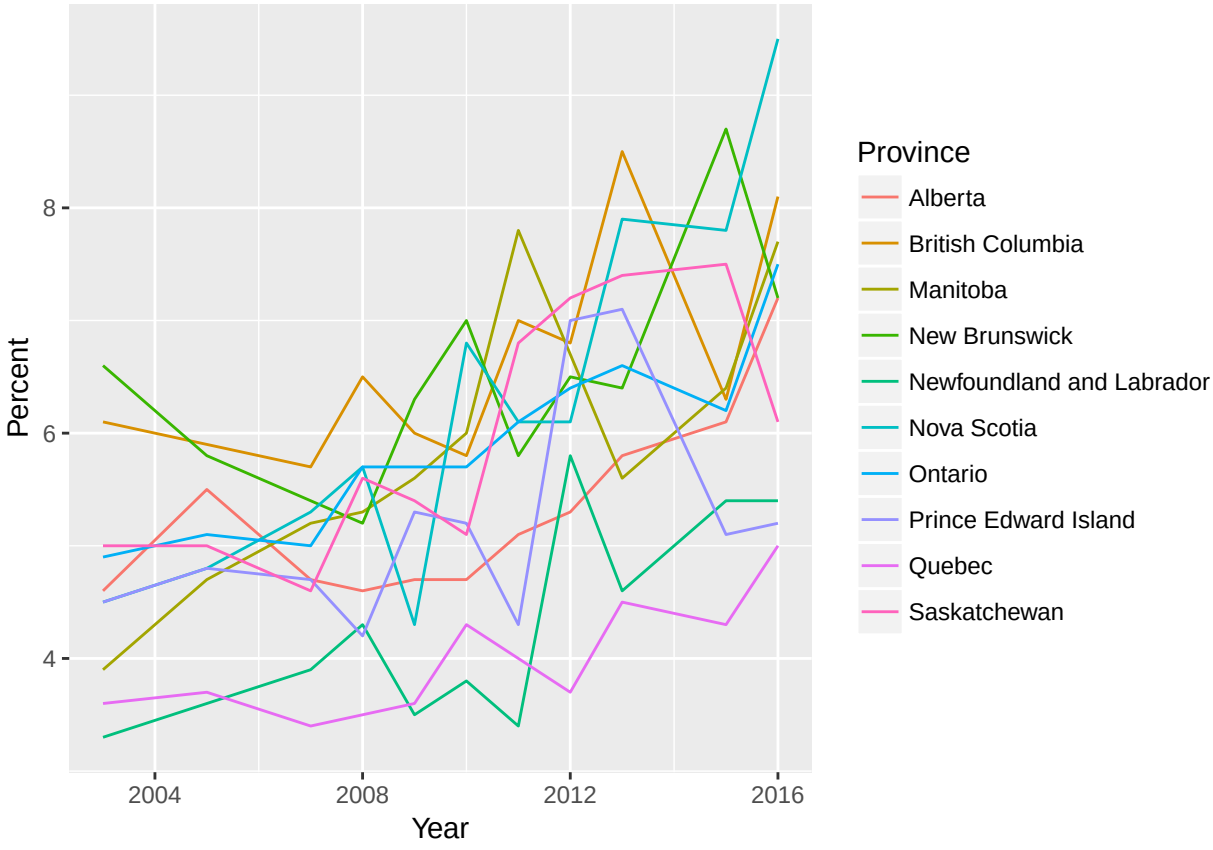
```
##
##  Pairwise comparisons using Nemenyi multiple comparison test
##              with q approximation for unreplicated blocked data
##
## data:  Perc and Province and Year
##
##                           Alberta British Columbia Manitoba New Brunswick
## British Columbia          0.01993 -                -        -
## 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**

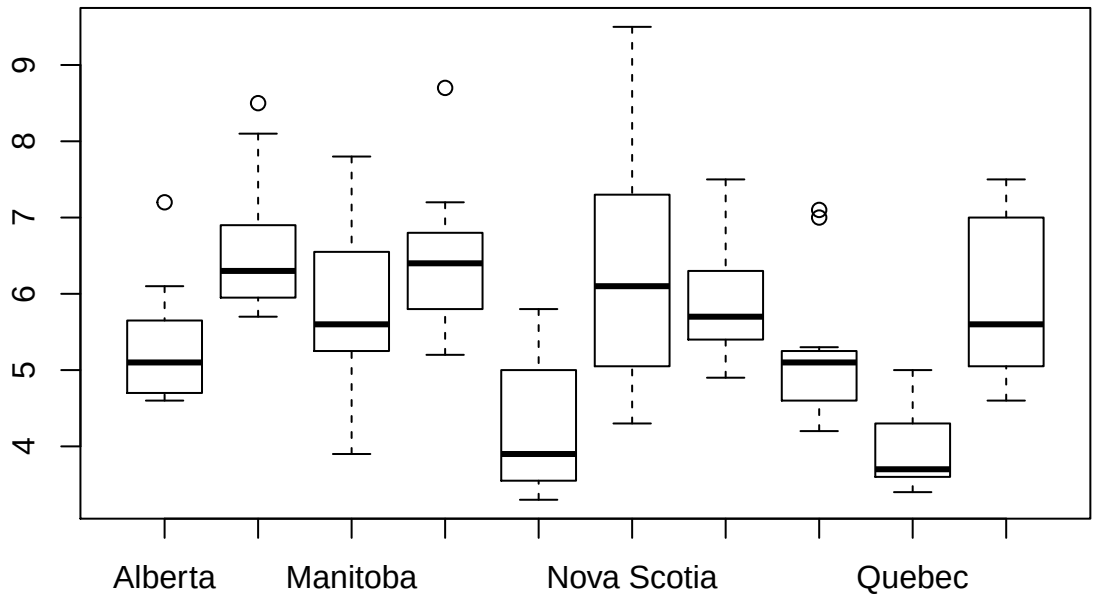
Import Fair or Poor data to analyze and represent visually.

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



First, must review and eliminate possible outliers from each data set.

```
bxplt_Poor<-boxplot(Total~Name,data=Poor)
```

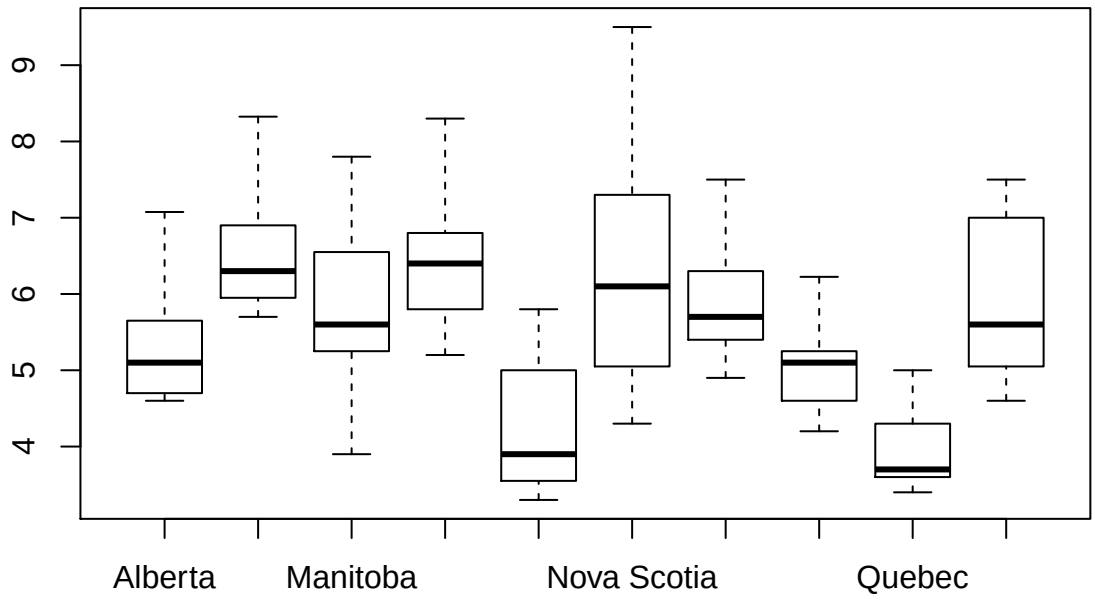



```
bxplt_Poor
```

```
## $stats
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 4.60 5.70 3.90  5.2 3.30 4.30  4.9 4.20  3.4  4.60
## [2,] 4.70 5.95 5.25  5.8 3.55 5.05  5.4 4.60  3.6  5.05
## [3,] 5.10 6.30 5.60  6.4 3.90 6.10  5.7 5.10  3.7  5.60
## [4,] 5.65 6.90 6.55  6.8 5.00 7.30  6.3 5.25  4.3  7.00
## [5,] 6.10 8.10 7.80  7.2 5.80 9.50  7.5 5.30  5.0  7.50
##
## $n
##  [1] 11 11 11 11 11 11 11 11 11 11
##
## $conf
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 4.647431 5.847431 4.980696 5.923612 3.209238 5.028127 5.271251
## [2,] 5.552569 6.752569 6.219304 6.876388 4.590762 7.171873 6.128749
##          [,8]     [,9]    [,10]
## [1,] 4.790348 3.366528 4.671044
## [2,] 5.409652 4.033472 6.528956
##
## $out
## [1] 7.2 8.5 8.7 7.0 7.1
##
```

```
## $group
## [1] 1 2 4 8 8
##
## $names
##  [1] "Alberta"                   "British Columbia"
##  [3] "Manitoba"                  "New Brunswick"
##  [5] "Newfoundland and Labrador" "Nova Scotia"
##  [7] "Ontario"                   "Prince Edward Island"
##  [9] "Quebec"                    "Saskatchewan"
```

```
P<-subset(Poor,select=-c(ï..Code))
P$Total[(P$Total==7.2& P$Name == "Alberta")] <- quantile(P$Total[P$Name=="Alberta"],0.75)+1.5*IQR(P$Tota
P$Total[(P$Total==8.5& P$Name == "British Columbia")] <- quantile(P$Total[P$Name=="British Columbia"],0
P$Total[(P$Total==8.7& P$Name == "New Brunswick")] <- quantile(P$Total[P$Name=="New Brunswick"],0.75)+1
P$Total[((P$Total==7.0|P$Total==7.1)&P$Name == "Prince Edward Island")] <- quantile(P$Total[P$Name=="Pri
boxplot(Total~Name,data=P)
```



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<-P[P$Name=='British Columbia',]
AP<-P[P$Name=='Alberta',]
SP<-P[P$Name=='Saskatchewan',]
MP<-P[P$Name=='Manitoba',]
OP<-P[P$Name=='Ontario',]
QP<-P[P$Name=='Quebec',]
```

```
NBP<-P[P$Name=='New Brunswick',]
NSP<-P[P$Name=='Nova Scotia',]
PEIP<-P[P$Name=='Prince Edward Island',]
NP<-P[P$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(BCP$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BCP$Total
## W = 0.85245, p-value = 0.04593
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  AP$Total
## W = 0.84749, p-value = 0.03958
```

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

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

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

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

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

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

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

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

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

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  NBP$Total
## W = 0.95375, p-value = 0.6921
```

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

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

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

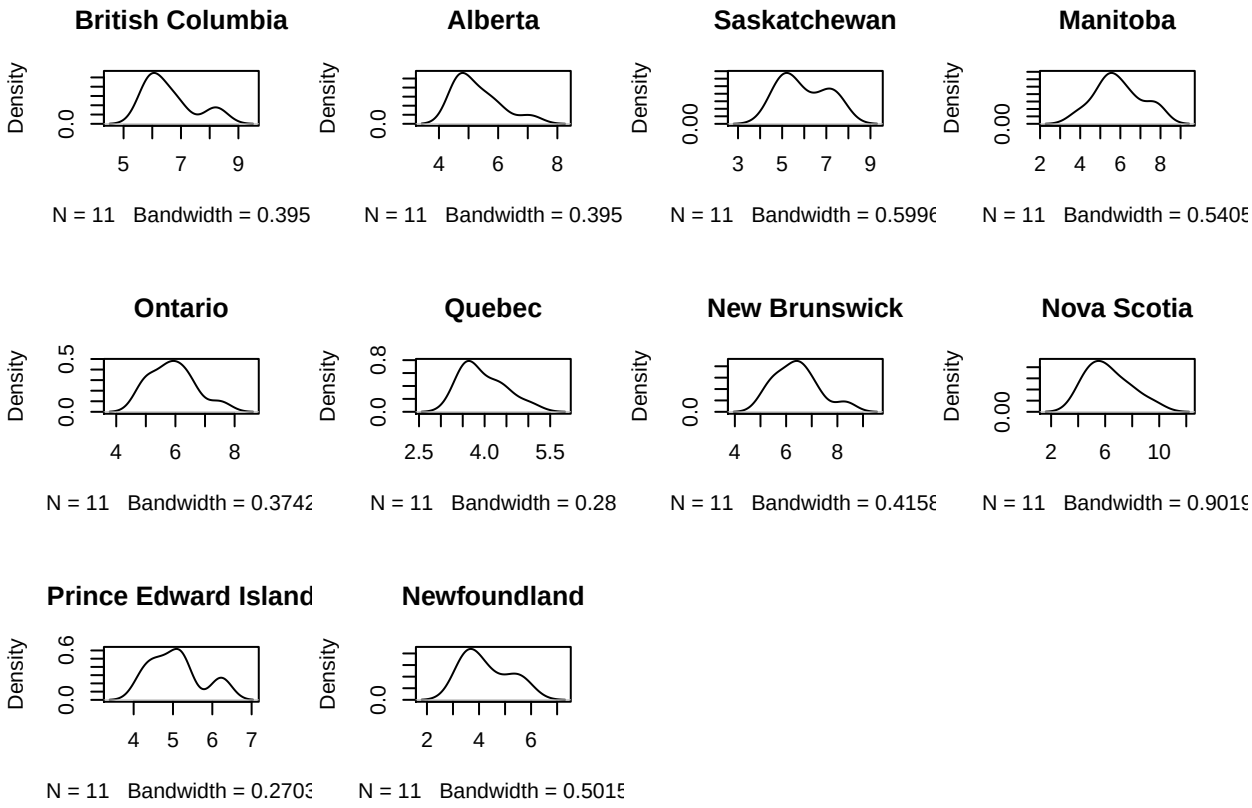
```
##
##  Shapiro-Wilk normality test
##
## data:  PEIP$Total
## W = 0.90794, p-value = 0.2305
```

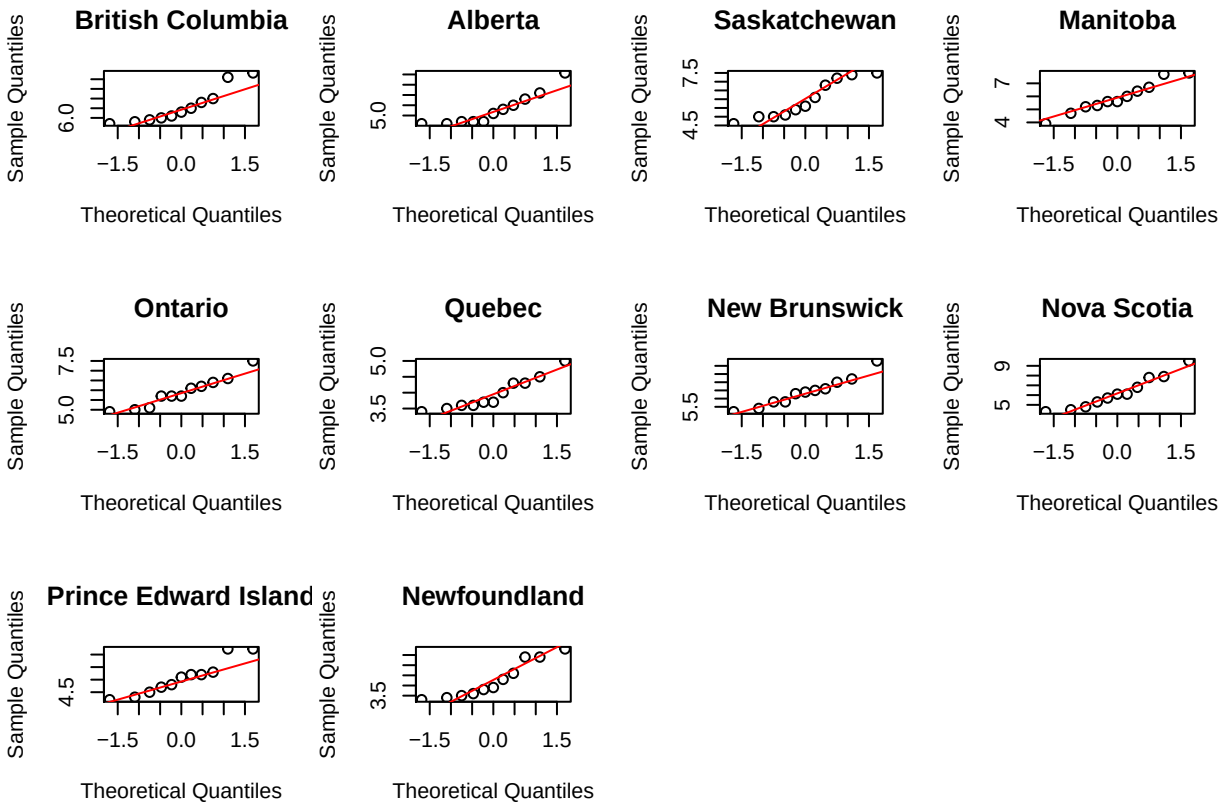
```
shapiro.test(NP$Total)
```

```
##
##  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
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
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 different populations and determine if at least two of the ten 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(P$Year),Province=P$Name,Perc=P$Total,Item=P$Item)

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

```
##
##  Friedman rank sum test
##
## data:  Perc and Province and Year
## Friedman chi-squared = 67.246, df = 9, p-value = 5.265e-11
```

Based on the results of the Friedman test, with a p-value less than the alpha threshold, 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         0.01767 -                -        -
```

```
## Manitoba                  0.88704 0.61913         -        -
## New Brunswick             0.16898 0.99892         0.97299  -
## Newfoundland and Labrador 0.64389 2.6e-06         0.01993  0.00015
## Nova Scotia               0.64389 0.87212         0.99999  0.99892
## Ontario                   0.73825 0.80150         1.00000  0.99603
## Prince Edward Island      0.99999 0.00369         0.64389  0.05471
## Quebec                    0.39635 3.5e-07         0.00557  2.5e-05
## Saskatchewan              0.82073 0.71556         1.00000  0.98862
##                           Newfoundland and Labrador Nova Scotia Ontario
## British Columbia          -                         -           -
## Manitoba                  -                         -           -
## New Brunswick             -                         -           -
## Newfoundland and Labrador -                         -           -
## Nova Scotia               0.00424                   -           -
## Ontario                   0.00728                   1.00000     -
## Prince Edward Island      0.88704                   0.35086     0.44406
## Quebec                    1.00000                   0.00099     0.00181
## 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                    0.69221              -
## Saskatchewan              0.54366              0.00321
##
## P value adjustment method: none
```

bold YEARLY TRENDS **bold**

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<-G[G$Year==2003,]
G2016<-G[G$Year==2016,]
P2003<-P[P$Year==2003,]
P2016<-P[P$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
```

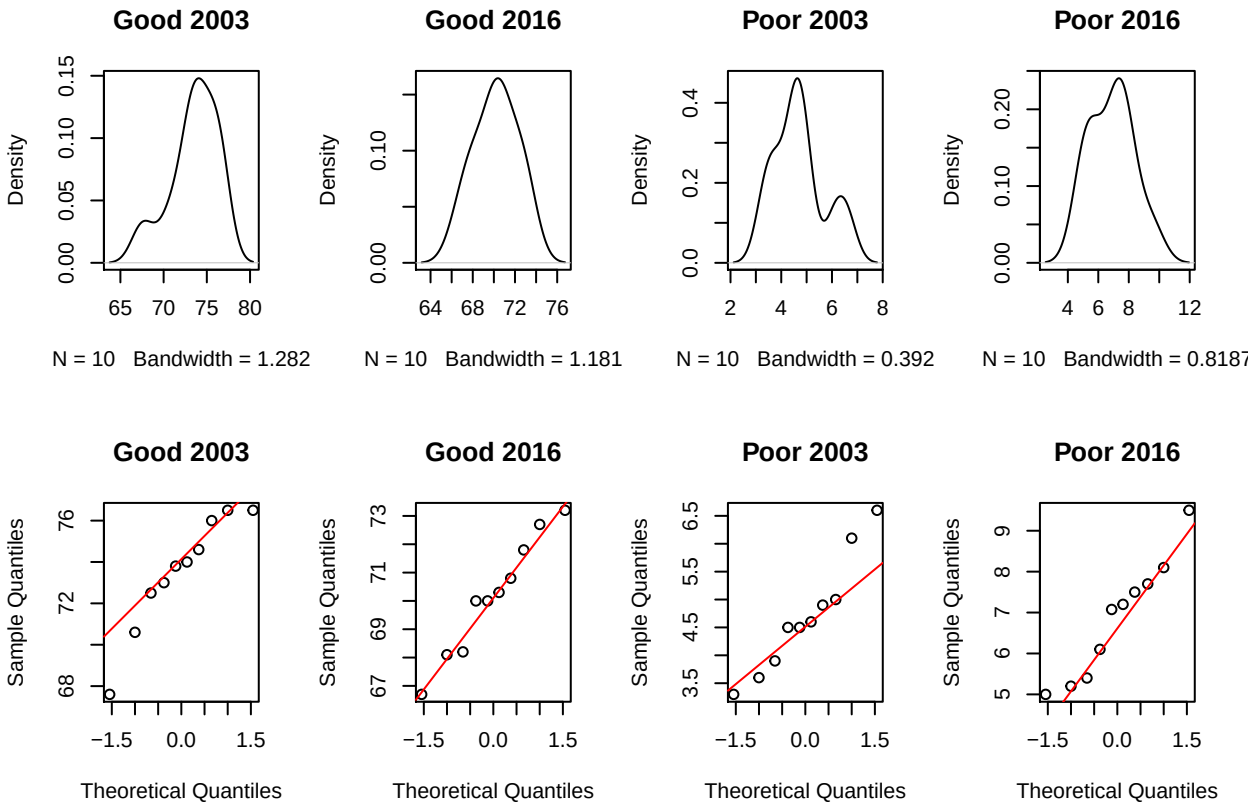
```
shapiro.test(P2003$Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  P2003$Total
## W = 0.94037, p-value = 0.5572
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  P2016$Total
## W = 0.94557, p-value = 0.6165
```

```
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.

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

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

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

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

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

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

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

```
## [1] 2.078396
```

As none of the variances are equal, non-parametric comparisons will be used on the Good and Poor data sets. Ho = Yearly data sets are the same Ha = Datasets differ alpha = 0.05

```
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
```

```
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
```

As both p-values are below the threshold of 0.05, the null hypothesis can be rejected and it is determined that there is statistical significance between the values in 2003 and the values in 2016. Modelling these trends in order to predict the percentage of provincial populations that will classify themselves into one of the two categories will be beneficial.

First we will combine the data in order to create a single model for the entire data set. From the combined data set a training and test data set will be created in order to create and evaluate the models. An 80/20 split of the data will be used.

```
All<-rbind(G,P)
All_Index<-sample(1:nrow(All),0.8*nrow(All))
All_Train<-All[All_Index,]
All_Test<-All[-All_Index,]
```

In order to determine which of the parameters should be included in the multiple linear regression model, a stepwise analysis of each model type will be conducted to determine prameter significance.

```
library(MASS)
full<-lm(Total~Name+Item+Year,data=All)
null<-lm(Total~1,data=All)
stepF<-stepAIC(null,scope=list(lower=null,upper=full),direction="forward",trace=TRUE)
```

```
## Start:  AIC=1545.75
## Total ~ 1
##
##        Df Sum of Sq    RSS     AIC
## + Item  1    244362   1041  345.85
## <none>              245403 1545.75
## + Year  1        17 245386 1547.73
## + Name  9       111 245292 1563.65
##
## Step:  AIC=345.85
## Total ~ Item
##
##        Df Sum of Sq     RSS    AIC
## + Name  9   110.857  929.67 339.06
## + Year  1    17.149 1023.38 344.19
## <none>              1040.53 345.85
##
## Step:  AIC=339.06
## Total ~ Item + Name
##
##        Df Sum of Sq    RSS    AIC
## + Year  1    17.149 912.52 336.97
## <none>              929.67 339.06
##
## Step:  AIC=336.97
```

```
## Total ~ Item + Name + Year
```

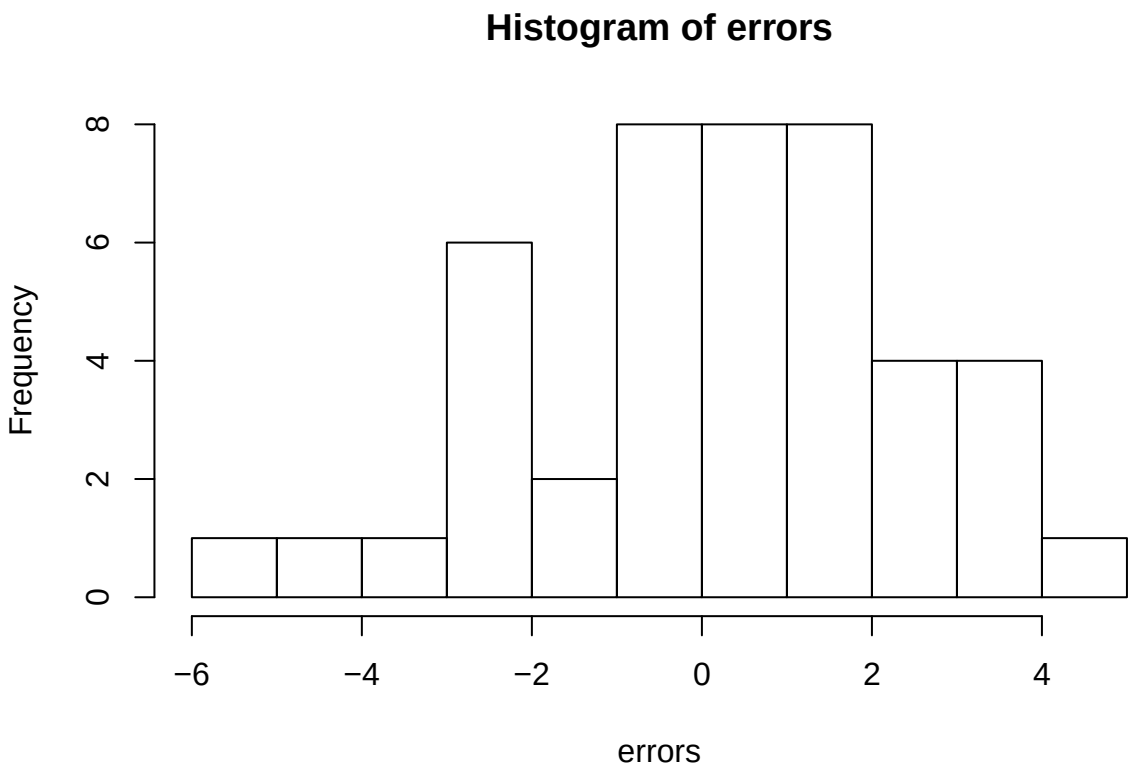
```
summary(stepF)
```

```
##
## Call:
## lm(formula = Total ~ Item + Name + Year, data = All)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0677 -1.5599 -0.0114  1.4784  4.5823
##
## Coefficients:
##                                              Estimate Std. Error
## (Intercept)                                 218.67810   73.75435
## ItemPerceived mental health, fair or poor (%) -66.65545    0.28243
## NameBritish Columbia                         -1.18864    0.63153
## NameManitoba                                 -1.19432    0.63153
## NameNew Brunswick                            -2.12045    0.63153
## NameNewfoundland and Labrador                -0.15341    0.63153
## NameNova Scotia                              -0.65341    0.63153
## NameOntario                                  -0.08523    0.63153
## NamePrince Edward Island                     -0.17841    0.63153
## NameQuebec                                    0.26477    0.63153
## NameSaskatchewan                             -1.14318    0.63153
## Year                                         -0.07255    0.03669
##                                              t value Pr(>|t|)
## (Intercept)                                    2.965 0.003381 **
## ItemPerceived mental health, fair or poor (%) -236.008  < 2e-16 ***
## NameBritish Columbia                          -1.882 0.061211 .
## NameManitoba                                  -1.891 0.059994 .
## NameNew Brunswick                             -3.358 0.000935 ***
## NameNewfoundland and Labrador                 -0.243 0.808309
## NameNova Scotia                               -1.035 0.302036
## NameOntario                                   -0.135 0.892779
## NamePrince Edward Island                      -0.283 0.777839
## NameQuebec                                     0.419 0.675462
## NameSaskatchewan                              -1.810 0.071711 .
## Year                                          -1.977 0.049354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.095 on 208 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9961
## F-statistic:  5066 on 11 and 208 DF,  p-value: < 2.2e-16
```

The model will be expected to evaluated at a 90% confidence interval. With this threshold in place, all parameters (Year, Name, Item) are deemed to be significant as per the stepwise AIC evaluation. The AIC value for the mlr inclusive of all parameters is smallest and therefore the best fit for the data.

Using this data, the model is trained and tested.

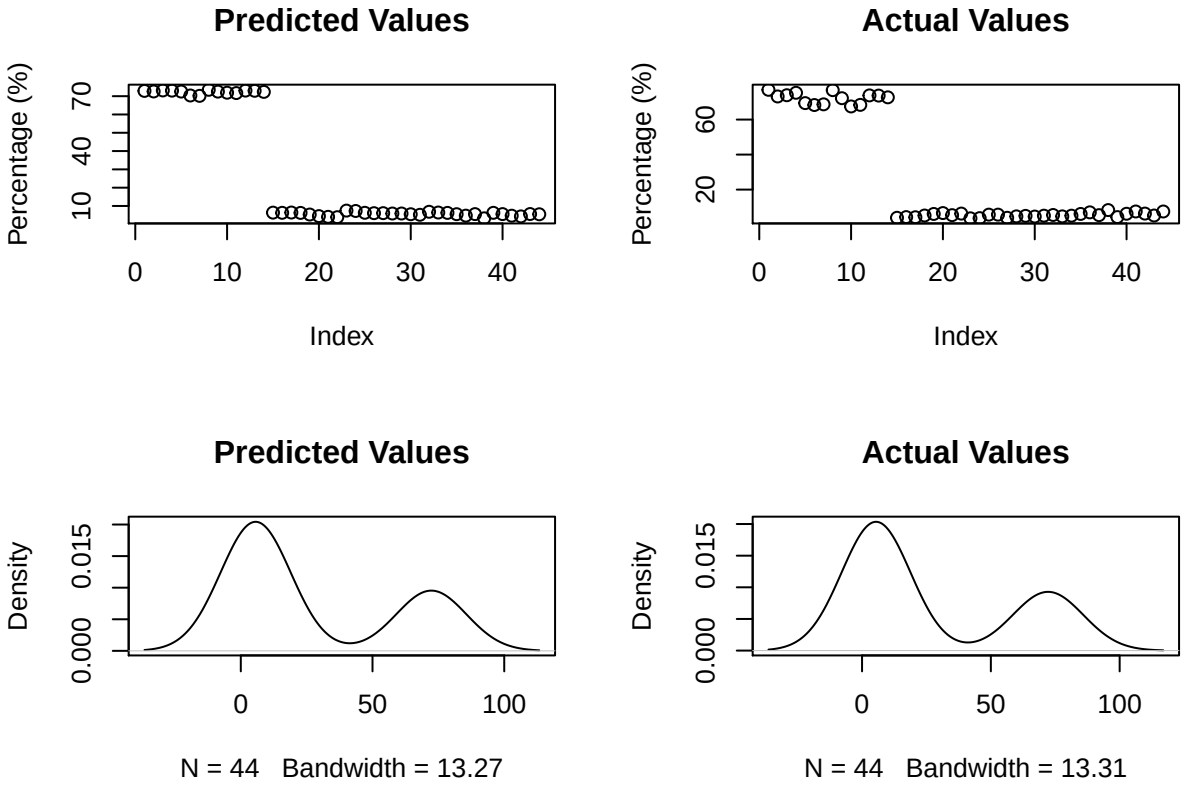
```
model_mlr<-lm(Total~Name+Item+Year,data=All_Train)
predict<-predict(model_mlr,interval="prediction",newdata=All_Test)
errors<-predict[,"fit"]-All_Test$Total
hist(errors)
```



The distribution of errors has a roughly normally distributed shape indicating a good fit. The calculated values using the model and the actual values from the test set will be comapred to determine if they are equal.

Ho = mean of values from model is equal to the mean of values from test set. Ha = two sets of values aren't equal alpha = 0.05

```
par(mfrow = c(2,2))
plot(predict[,"fit"], main = "Predicted Values", ylab = "Percentage (%)")
plot(All_Test$Total, main = "Actual Values", ylab = "Percentage (%)")
plot(density(predict[,"fit"]),main = "Predicted Values")
plot(density(All_Test$Total), main = "Actual Values")
```



Clearly data is not parametric, so a Wilcoxon Rank Sum test will be used to validate null hypothesis. The data is paired as the values have come from the same parameters.

```
wilcox.test(predict[,"fit"],All_Test$Total,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  predict[, "fit"] and All_Test$Total
## V = 546, p-value = 0.5592
## alternative hypothesis: true location shift is not equal to 0
```

As the p-value is greated than 0 .05, the null hypothesis holds and the actual values for the data are concluded to be statistically equal to the predicted values from the model. Therefore, the model is a good fit.