Solar Wind compare

elisabeth wolfert

1/30/2021

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#Load Common R Libraries to examine the data set.  
  
library("car")

library("dplyr")

library("ggplot2")  
library("ggpubr")  
library("tidyverse")

library("broom")  
library("gplots")

library("coin")

library("multcompView")  
library("BSDA")

library("DescTools")

library("effsize")  
library("AICcmodavg")  
library("rcompanion")

#Load Data File and Filter to Pull Required Data for the data test.  
  
SolarWind\_PRE = read.csv("WindSolar\_State9.csv",header = TRUE)

# Filter ony required data to standardize the comparisons.  
  
SolarWind <- filter(SolarWind\_PRE, #SolarWind\_PRE$State == c('California','Arizona'),  
 SolarWind\_PRE$Type %in% c('Wind','Photo'),  
 SolarWind\_PRE$Com\_Type %in%   
 c('Commercial','Residential','Industrial'),  
 SolarWind\_PRE$Total > 0)

# Load data set to see that the data displays properly and filters correctly.  
head(SolarWind)

## State Type Com\_Type Total  
## 1 Alaska Photo Commercial 2.00  
## 2 Alaska Photo Industrial 1.01  
## 3 Alaska Photo Residential 8.00  
## 4 Arkansas Photo Commercial 9.00  
## 5 Arkansas Photo Residential 70.00  
## 6 Arizona Photo Commercial 161.00

#Display Data in Dataframe agian showing more values.  
  
str(SolarWind)

## 'data.frame': 26051 obs. of 4 variables:  
## $ State : chr "Alaska" "Alaska" "Alaska" "Arkansas" ...  
## $ Type : chr "Photo" "Photo" "Photo" "Photo" ...  
## $ Com\_Type: chr "Commercial" "Industrial" "Residential" "Commercial" ...  
## $ Total : num 2 1.01 8 9 70 ...

#Sum Total Data field  
  
sum(SolarWind$Total)

## [1] 104131049

#Summary of Data Frame  
  
summary(SolarWind)

## State Type Com\_Type Total   
## Length:26051 Length:26051 Length:26051 Min. : 0   
## Class :character Class :character Class :character 1st Qu.: 10   
## Mode :character Mode :character Mode :character Median : 52   
## Mean : 3997   
## 3rd Qu.: 375   
## Max. :1041063

#Fields in Data Frame summary data.  
  
names(SolarWind)

## [1] "State" "Type" "Com\_Type" "Total"

## Including Plots

You can also embed plots, for example:

# Data Discovery  
# Change Data Types to Factors  
  
SolarWind$Type <- as.factor(SolarWind$Type)  
SolarWind$Com\_Type <- as.factor(SolarWind$Com\_Type)  
SolarWind$Total <- as.numeric(SolarWind$Total)  
  
is.factor(SolarWind$State)

## [1] FALSE

is.factor(SolarWind$Type)

## [1] TRUE

is.factor(SolarWind$Com\_Type)

## [1] TRUE

is.numeric(SolarWind$Total)

## [1] TRUE

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

# Table Data Types to Show Values by Type  
  
table(SolarWind$Type)

##   
## Photo Wind   
## 14499 11552

table(SolarWind$Com\_Type)

##   
## Commercial Industrial Residential   
## 10234 5376 10441

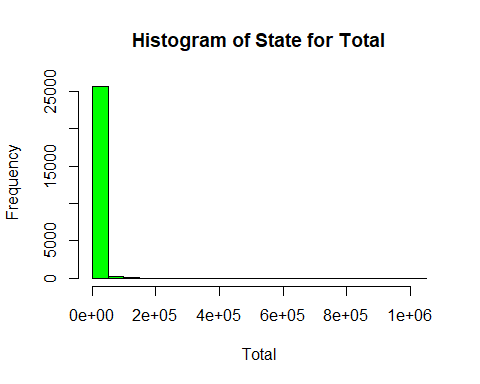
# Display values with Grouping for Summary data for Type  
# Hypothesis  
# We are performing a test to see which type of energy type is used more: Wind or Solar. Below is a grouping of the Type field showing the two energy types of Photo(Solar) or Wind energy. The analysis will be performed using a 5% significance level.  
  
# H0: There will be no difference between the mean types for Wind or Photo(Solar) energy usage.  
  
# H1: There will be a difference between the mean types for Wind or Photo(Solar) energy usage.  
  
group\_by(SolarWind, Type) %>% summarise(count=n(),mean=mean(SolarWind$Total),sd=sd(SolarWind$Total),.groups='drop')

## # A tibble: 2 x 4  
## Type count mean sd  
## <fct> <int> <dbl> <dbl>  
## 1 Photo 14499 3997. 36587.  
## 2 Wind 11552 3997. 36587.

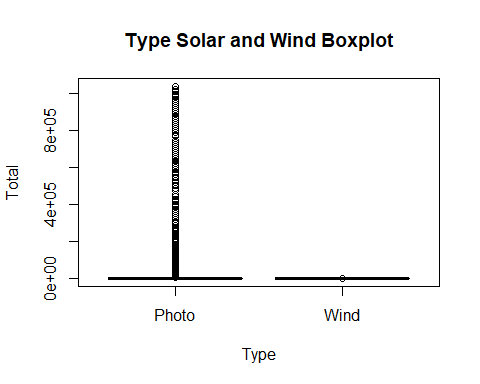
# Display values with Grouping for Summary data for Com Type  
  
# Hypothesis  
# We are performing a test to see which type of renewable energy source types including; Commercial, Residential or Industrial renewable energy usage is used more often. Below is a grouping of the Com Type field showing the three energy sources of Commercial, Residential or Industrial energy. The analysis will be performed using a 5% significance level.  
  
# H0: There will be no difference between the mean types for Commercial, Residential or Industrial renewable energy usage.  
  
# H1: There will be a difference between the mean types for Commercial, Residential or Industrial renewable energy usage.   
  
#The count value shows different field counts but the mean and standard deviations are the same.  
  
group\_by(SolarWind, SolarWind$Com\_Type) %>% summarise(count=n(),mean=mean(SolarWind$Total),sd=sd(SolarWind$Total),.groups='drop')

## # A tibble: 3 x 4  
## `SolarWind$Com\_Type` count mean sd  
## <fct> <int> <dbl> <dbl>  
## 1 Commercial 10234 3997. 36587.  
## 2 Industrial 5376 3997. 36587.  
## 3 Residential 10441 3997. 36587.

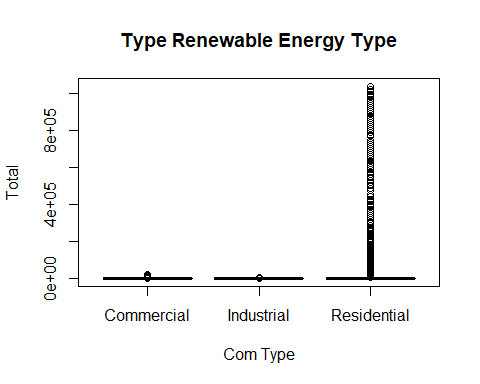
# Create a Histogram on the Total field in data file  
#   
# The data in the file has a distribution that leans to the left  
  
hist(SolarWind$Total, col=c("green"),main = "Histogram of State for Total", xlab="Total")



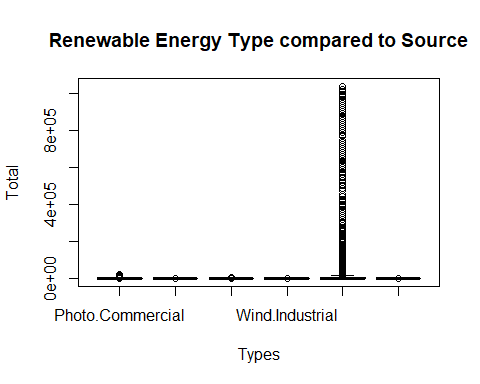
# Box plot comparing the types fo Wind and Solar  
# Photo is another term for Solar. The Box plot shows that Solar is the dominant energy source.  
  
boxplot(SolarWind$Total ~ SolarWind$Type, col=c("purple","red"),main="Type Solar and Wind Boxplot", ylab = "Total",xlab = "Type")



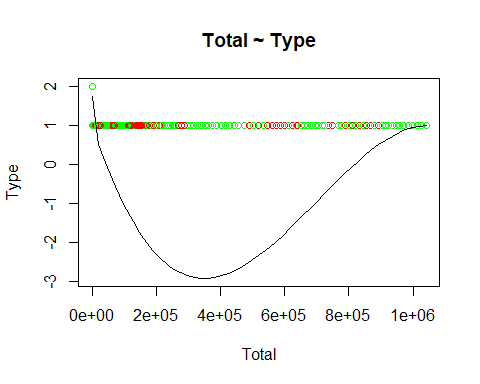
# Created a Box plot to compare energy sources of Commercial, Residential and Industrial. The box plot shows that the data is driven by Residential sources.  
  
  
boxplot(SolarWind$Total ~ SolarWind$Com\_Type, col=c("purple","red"),main="Type Renewable Energy Type", ylab = "Total",xlab = "Com Type")



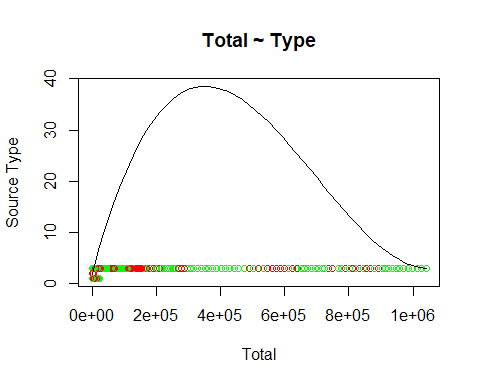
# Created a box plot comparing Solar Type to Com Types. The Box plot is leaning toward Solar and Residential Data. The chart is hard to read becuase of the size but the chart should be leaning toward Residential and Solar.  
  
boxplot(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, col=c("purple","red","Green"),main="Renewable Energy Type compared to Source", ylab = "Total",xlab = "Types")



# Scatter plot for Total and Type of Wind or Photo(Solar) energy.  
  
scatter.smooth(x=SolarWind$Total, y=SolarWind$Type, col=c("Green","Red"), main="Total ~ Type",ylab = "Type",xlab= "Total") # scatterplot



# Smooth Scatter plot for Total and Com Type including Residential, Commercial or Industrial energy sources.  
  
scatter.smooth(x=SolarWind$Total, y=SolarWind$Com\_Type, col=c("Green","Red"), main="Total ~ Type",ylab = "Source Type",xlab= "Total") # scatterplot



# Created GLM model for Total and Type  
  
glm(formula = SolarWind$Type ~ SolarWind$Total, family = binomial,   
 data = SolarWind)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Call: glm(formula = SolarWind$Type ~ SolarWind$Total, family = binomial,   
## data = SolarWind)  
##   
## Coefficients:  
## (Intercept) SolarWind$Total   
## 0.775173 -0.007306   
##   
## Degrees of Freedom: 26050 Total (i.e. Null); 26049 Residual  
## Null Deviance: 35780   
## Residual Deviance: 25780 AIC: 25790

# Fit the GLM model and displayed the summary  
  
lrfit <- glm(SolarWind$Com\_Type ~ SolarWind$Total,family=binomial,data=SolarWind)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

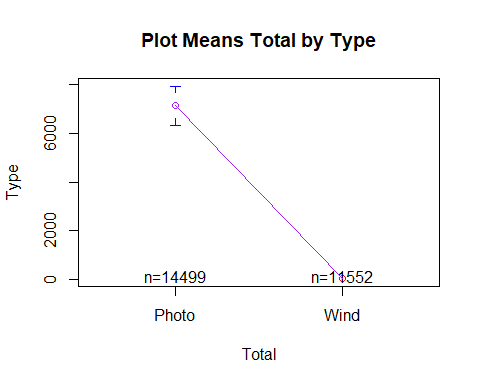
summary(lrfit)

##   
## Call:  
## glm(formula = SolarWind$Com\_Type ~ SolarWind$Total, family = binomial,   
## data = SolarWind)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5434 -1.3099 0.9769 1.0497 1.0514   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.039e-01 1.345e-02 22.59 <2e-16 \*\*\*  
## SolarWind$Total 1.232e-04 6.631e-06 18.58 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 34909 on 26050 degrees of freedom  
## Residual deviance: 33830 on 26049 degrees of freedom  
## AIC: 33834  
##   
## Number of Fisher Scoring iterations: 8

# Plot the means for Total and Type showing that Photo(Solar) dominates the data. The plot shows that the means for Photo(Solar) power has a higher significance.  
  
plotmeans(SolarWind$Total ~ SolarWind$Type, col=c("purple","red","Green"),main="Plot Means Total by Type", ylab = "Type",xlab = "Total")

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped

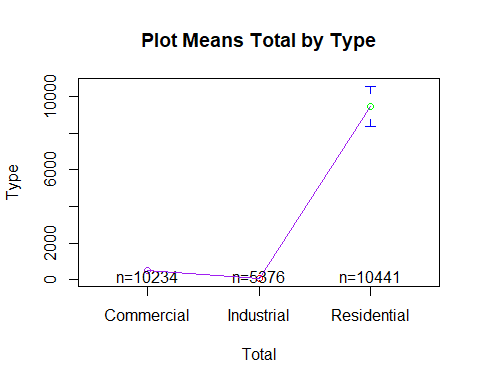
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped



# Plot the means for Total and Com Type showing that Residential dominates the data. Plotting the means for energy source shows that the Residential source data has a higher level of significance in the data.  
  
plotmeans(SolarWind$Total ~ SolarWind$Com\_Type, col=c("purple","red","Green"),main="Plot Means Total by Type", ylab = "Type",xlab = "Total")

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped  
  
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped  
  
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped



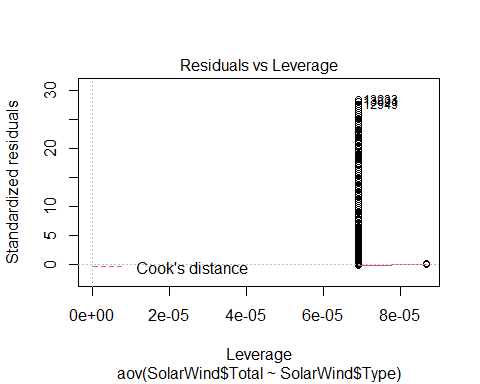
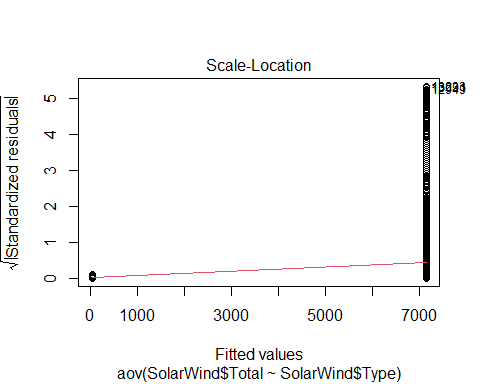
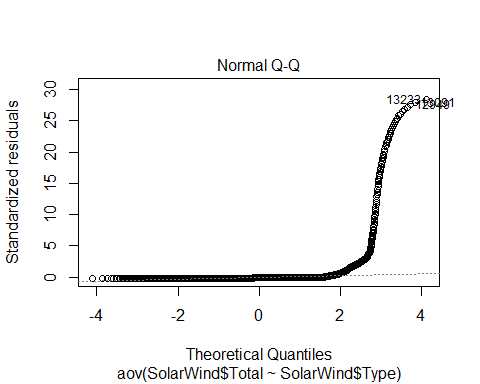
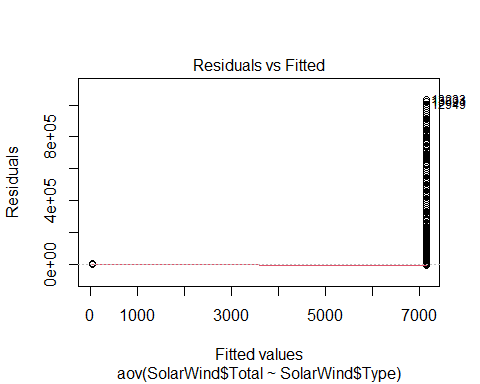
# Ran a one way ANOVA test with the p value less than 0.05 with (P > 2.89e-14) giving the level of significance. Now that the data has been evaluated we run the one way ANOVA test and see that the P value shows a 95 percent significance with the p value (p<=0.05) so we can not accept the null hypothesis. The null hypothesis is rejected and we can see that there is a significant difference between the two types of Wind and Photo(Solar) energy types.  
  
#According to the output of the test for energy type being (F(1,627)=60.61,p=2.89e-14) we see a high level of significance between energy types.  
  
SType\_av= aov(SolarWind$Total ~ SolarWind$Type, data = SolarWind)  
summary(SType\_av)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SolarWind$Type 1 3.244e+11 3.244e+11 244.6 <2e-16 \*\*\*  
## Residuals 26049 3.455e+13 1.326e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

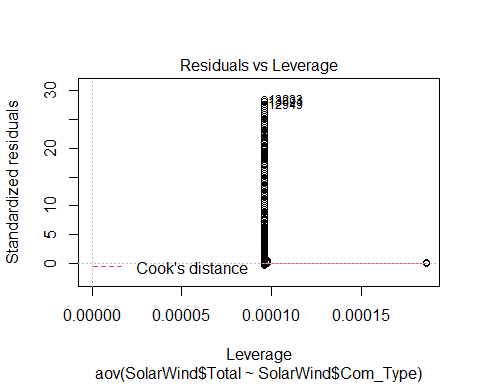
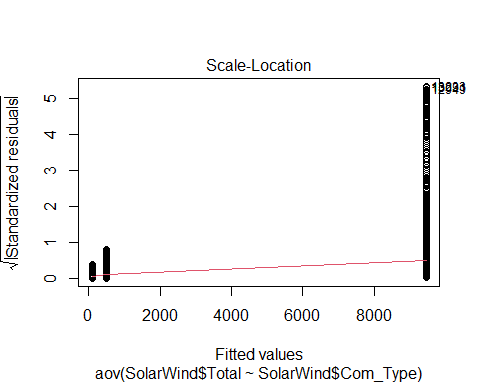
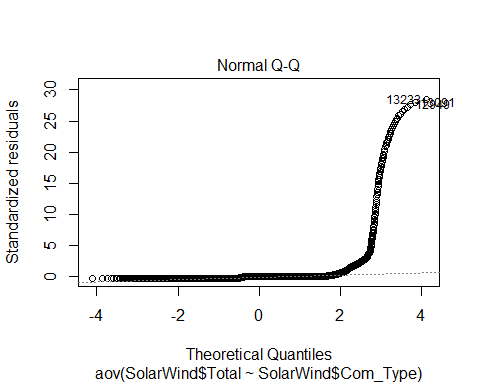
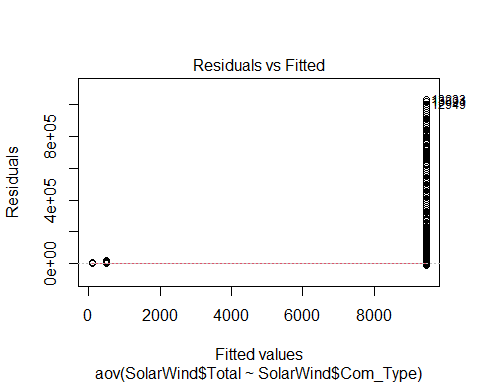
# Ran a one way ANOVA test with the p value less than 0.05 with (P > <2e-16) giving the level of significance. Now that the data has been evaluated we run the one way ANOVA test and see that the P value shows a 95 percent significance with the p value (p<=0.05) so we can not accept the null hypothesis. The null hypothesis is rejected and we can see that there is a significant difference between the two types of Wind and Photo(Solar) energy types.  
  
#According to the output of the test for energy type being (F(2,626)=67.29,p=<2e-16) we see a high level of significance between energy types.  
  
WCom\_Type\_av= aov(SolarWind$Total ~ SolarWind$Com\_Type, data = SolarWind)  
summary(WCom\_Type\_av)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SolarWind$Com\_Type 2 5.200e+11 2.600e+11 197.2 <2e-16 \*\*\*  
## Residuals 26048 3.435e+13 1.319e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Plot the residuals for the ANOVA test by Type energy source.  
  
plot(SType\_av)



# Plot the residuals for the ANOVA test by Com Type energy source.  
  
plot(WCom\_Type\_av)



# Run a pairwise test on the data for energy Type when we reject the null hypothesis.  
  
pairwise.t.test(SolarWind$Total, SolarWind$Type,p.adjust.method = "BH",pool.sd = FALSE)

##   
## Pairwise comparisons using t tests with non-pooled SD   
##   
## data: SolarWind$Total and SolarWind$Type   
##   
## Photo   
## Wind <2e-16  
##   
## P value adjustment method: BH

# Run a pairwise test on the data for energy Com Type when we reject the null hypothesis.  
  
pairwise.t.test(SolarWind$Total, SolarWind$Com\_Type,p.adjust.method = "BH",pool.sd = FALSE)

##   
## Pairwise comparisons using t tests with non-pooled SD   
##   
## data: SolarWind$Total and SolarWind$Com\_Type   
##   
## Commercial Industrial  
## Industrial <2e-16 -   
## Residential <2e-16 <2e-16   
##   
## P value adjustment method: BH

# Run the levene test of the Type Wind or Photo(Solar) data test.  
  
leveneTest(SolarWind$Total ~SolarWind$Type,data=SolarWind)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 241.31 < 2.2e-16 \*\*\*  
## 26049   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Run the levene test of the Com Type Residential, Commercial or Industrial data test.  
  
leveneTest(SolarWind$Total ~SolarWind$Com\_Type,data=SolarWind)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 2 195.65 < 2.2e-16 \*\*\*  
## 26048   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

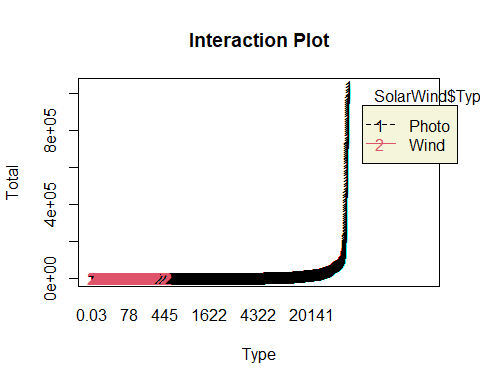
# Run the oneway test of the Type Wind or Photo(Solar) data test.  
  
oneway.test(SolarWind$Total ~SolarWind$Type,data=SolarWind)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: SolarWind$Total and SolarWind$Type  
## F = 306.99, num df = 1, denom df = 14498, p-value < 2.2e-16

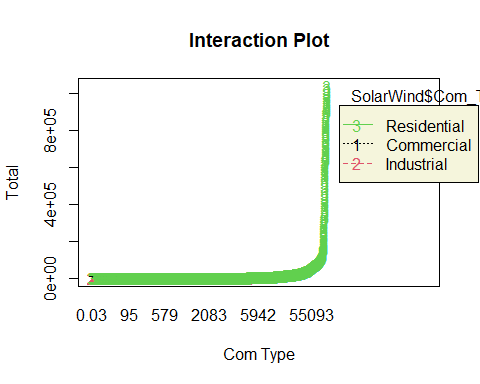
# Run the one way test of the Com Type Residential, Commercial or Industrial data test.  
oneway.test(SolarWind$Total ~SolarWind$Com\_Type,data=SolarWind)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: SolarWind$Total and SolarWind$Com\_Type  
## F = 403, num df = 2, denom df = 15545, p-value < 2.2e-16

# Run the interaction Plot test of the Type Wind or Photo(Solar) data test. The plot shows the higher level of sinificance for Photo(Solar) energy.  
  
interaction.plot(SolarWind$Total, SolarWind$Type, SolarWind$Total,type="b",col=c(1:3),leg.bty="o",leg.bg="beige",lwd=2,xlab="Type",ylab="Total",main="Interaction Plot")



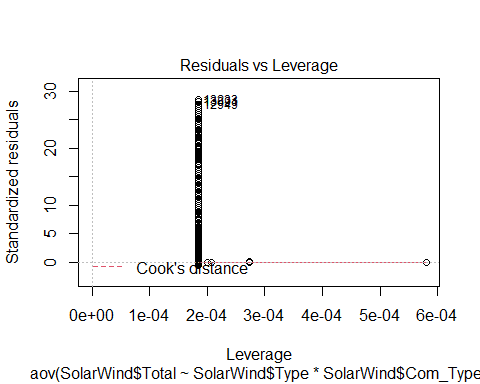
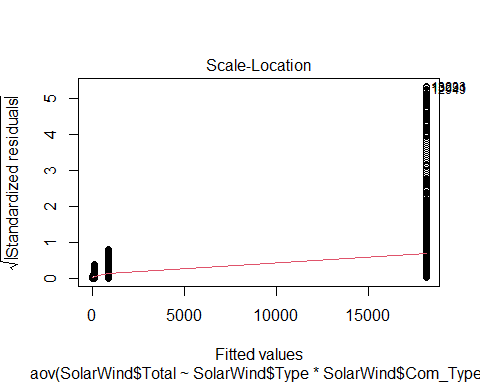
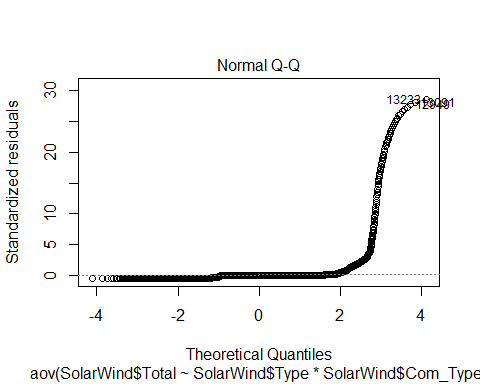
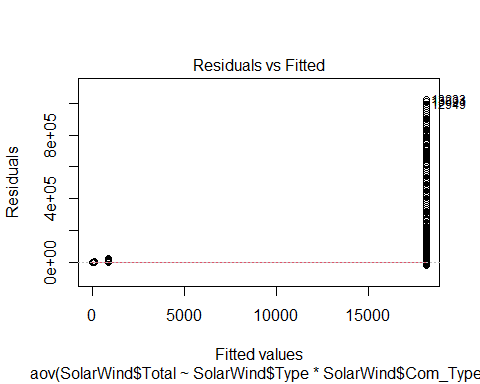
# Run the interaction plot test of the Com Type Residential, Commercial or Industrial data test. We cans ee in the test that Residential Solar energy is the has a higher level of significance.  
  
interaction.plot(SolarWind$Total, SolarWind$Com\_Type, SolarWind$Total,type="b",col=c(1:3),leg.bty="o",leg.bg="beige",lwd=2,xlab="Com Type",ylab="Total",main="Interaction Plot")



# Now we run a two way ANOVA test to see if there is any difference in significance between the types of energy sources for Wind and Photo(Solar) energy or the source of that energy of Residential, Industrial of Commercial.  
  
# H0: There will be no difference between the mean types for Wind or Photo(Solar) energy usage.  
  
# H1: There will be a difference between the mean types for Wind or Photo(Solar) energy usage.  
  
# H0: There will be no difference between the mean types for Residential, Commercial and Industrial energy usage.  
  
# H1: There will be a difference between the mean types for Residential, Commercial and Industrial usage.  
  
SolarWind\_av2= aov(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, data = SolarWind)  
  
#Summary data for the two way ANOVA test.  
  
summary(SolarWind\_av2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SolarWind$Type 1 3.244e+11 3.244e+11 252.2 <2e-16 \*\*\*  
## SolarWind$Com\_Type 2 5.808e+11 2.904e+11 225.8 <2e-16 \*\*\*  
## SolarWind$Type:SolarWind$Com\_Type 2 4.654e+11 2.327e+11 180.9 <2e-16 \*\*\*  
## Residuals 26045 3.350e+13 1.286e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Plot the summary data for the two way ANOVA test.  
  
plot(SolarWind\_av2)

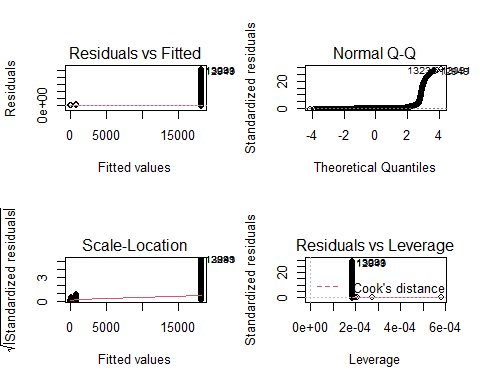


# Now we run a an interaction chart on two way ANOVA data test to see if there is any difference in significance between the types of energy sources for Wind and Photo(Solar) energy or the source of that energy of Residential, Industrial of Commercial.  
  
interaction<- aov(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, data = SolarWind)  
summary(interaction)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SolarWind$Type 1 3.244e+11 3.244e+11 252.2 <2e-16 \*\*\*  
## SolarWind$Com\_Type 2 5.808e+11 2.904e+11 225.8 <2e-16 \*\*\*  
## SolarWind$Type:SolarWind$Com\_Type 2 4.654e+11 2.327e+11 180.9 <2e-16 \*\*\*  
## Residuals 26045 3.350e+13 1.286e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

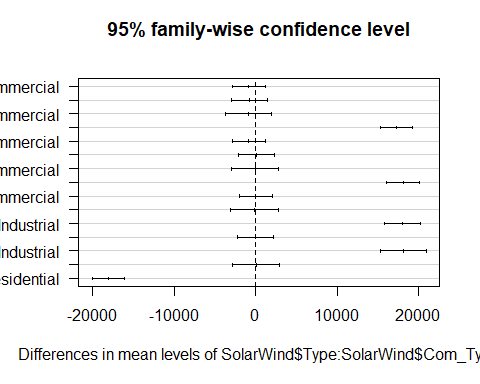
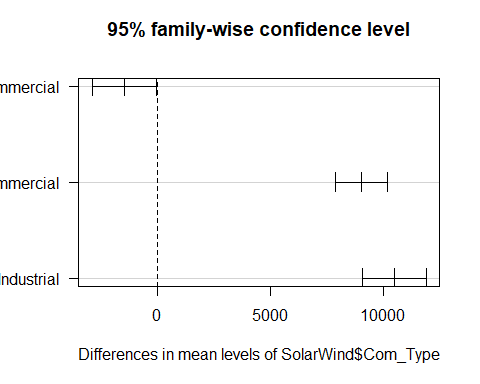
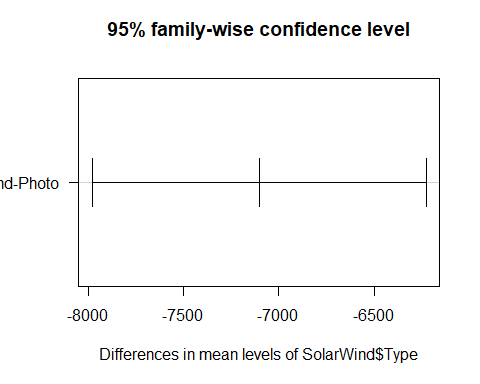
#model.names <- c("SState\_av", "WState\_av","SolarWind\_av2", "interaction")  
#aictab(model.set, modnames = model.names)

# Plot the residuals for the two way ANOVA test.  
  
par(mfrow=c(2,2))  
plot(SolarWind\_av2)



par(mfrow=c(1,1))

# Run the tukey two way test on the two way ANOVA data.  
  
tukey.two.way <- aov(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, data = SolarWind)  
  
# Run the tukey two way test on the two way ANOVA data. The plots show that there is a 95 percent confidence and significance level in the result of the tukey tests.  
  
tukey.plot.aov<-aov(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, data = SolarWind)  
  
tukey.plot.test<-TukeyHSD(tukey.plot.aov)  
plot(tukey.plot.test, las = 1)



#Display the results of the tukey two way test on the two way ANOVA data.  
  
tukey.two.way

## Call:  
## aov(formula = SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type,   
## data = SolarWind)  
##   
## Terms:  
## SolarWind$Type SolarWind$Com\_Type  
## Sum of Squares 3.243722e+11 5.808391e+11  
## Deg. of Freedom 1 2  
## SolarWind$Type:SolarWind$Com\_Type Residuals  
## Sum of Squares 4.653860e+11 3.349956e+13  
## Deg. of Freedom 2 26045  
##   
## Residual standard error: 35863.89  
## Estimated effects may be unbalanced

#Run the linear regresion model on the results of the tukey two way test on the two way ANOVA data.  
  
lm\_output <- lm(SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type, data = SolarWind)  
summary(lm\_output)

##   
## Call:  
## lm(formula = SolarWind$Total ~ SolarWind$Type \* SolarWind$Com\_Type,   
## data = SolarWind)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18121 -766 -53 -3 1022941   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 884.8 487.2 1.816  
## SolarWind$TypeWind -860.0 710.3 -1.211  
## SolarWind$Com\_TypeIndustrial -757.7 767.8 -0.987  
## SolarWind$Com\_TypeResidential 17236.7 688.7 25.028  
## SolarWind$TypeWind:SolarWind$Com\_TypeIndustrial 737.1 1266.0 0.582  
## SolarWind$TypeWind:SolarWind$Com\_TypeResidential -17185.5 999.0 -17.203  
## Pr(>|t|)   
## (Intercept) 0.0693 .   
## SolarWind$TypeWind 0.2260   
## SolarWind$Com\_TypeIndustrial 0.3238   
## SolarWind$Com\_TypeResidential <2e-16 \*\*\*  
## SolarWind$TypeWind:SolarWind$Com\_TypeIndustrial 0.5604   
## SolarWind$TypeWind:SolarWind$Com\_TypeResidential <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35860 on 26045 degrees of freedom  
## Multiple R-squared: 0.03931, Adjusted R-squared: 0.03912   
## F-statistic: 213.1 on 5 and 26045 DF, p-value: < 2.2e-16

#Plot the results of the tukey linear regresion model of the two way test on the two way ANOVA data. The P value is less than (p<0.5) with a value of 2.2e-16.  
  
plot(lm\_output)

