LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: **James Connolly**

Assessment Title: **Hypothesis**

Work to be submitted to: **08/01/19**

Date for submission of work: **James Connolly**

Place and time for submitting work: ____________________

To be completed by the Student

Student's Name: **PRATEEK PARASHER**

Class: **MSc Big Data Analytics**

Subject/Module: **DATA SCIENCE**

Word Count (where applicable): ____________________

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: **PRATEEK PARASHER** Date: **08/01/19**

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

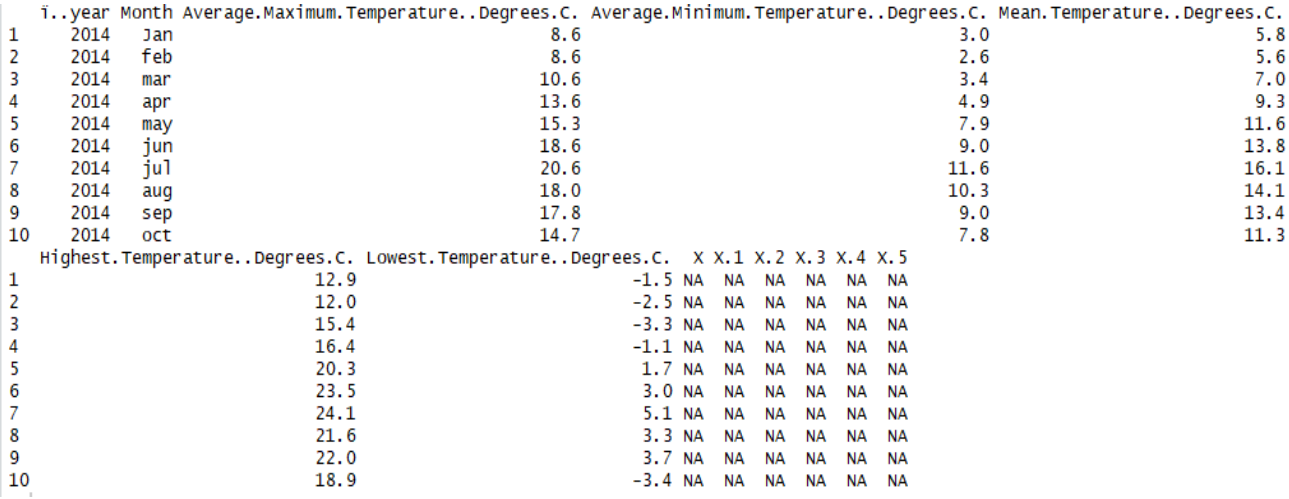
Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

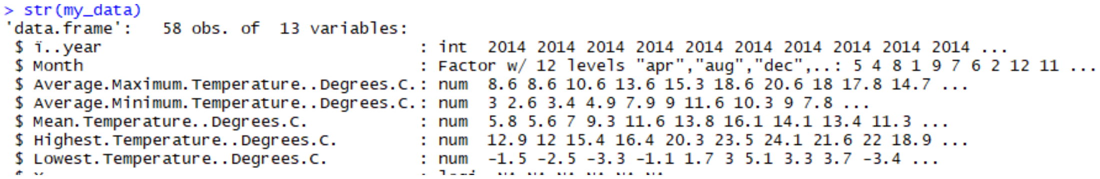
Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

ABSTRACT: - Using statistical analysis I am interested to examine my dataset, I want to explore the mean of lowest temp of Dublin through **one sample t-test** & normal distribution test through **Shapiro-Wilk test**:. Then I want to see power analysis of test for comparing means

DATA DESCRIPTION



```
   ï..year Month Average.Maximum.Temperature..Degrees.C. Average.Minimum.Temperature..Degrees.C. Mean.Temperature..Degrees.C.
1     2014   Jan                                     8.6                                     3.0                          5.8
2     2014   feb                                     8.6                                     2.6                          5.6
3     2014   mar                                    10.6                                     3.4                          7.0
4     2014   apr                                    13.6                                     4.9                          9.3
5     2014   may                                    15.3                                     7.9                         11.6
6     2014   jun                                    18.6                                     9.0                         13.8
7     2014   jul                                    20.6                                    11.6                         16.1
8     2014   aug                                    18.0                                    10.3                         14.1
9     2014   sep                                    17.8                                     9.0                         13.4
10    2014   oct                                    14.7                                     7.8                         11.3
   Highest.Temperature..Degrees.C. Lowest.Temperature..Degrees.C.  X X.1 X.2 X.3 X.4 X.5
1                             12.9                           -1.5 NA  NA  NA  NA  NA  NA
2                             12.0                           -2.5 NA  NA  NA  NA  NA  NA
3                             15.4                           -3.3 NA  NA  NA  NA  NA  NA
4                             16.4                           -1.1 NA  NA  NA  NA  NA  NA
5                             20.3                            1.7 NA  NA  NA  NA  NA  NA
6                             23.5                            3.0 NA  NA  NA  NA  NA  NA
7                             24.1                            5.1 NA  NA  NA  NA  NA  NA
8                             21.6                            3.3 NA  NA  NA  NA  NA  NA
9                             22.0                            3.7 NA  NA  NA  NA  NA  NA
10                            18.9                           -3.4 NA  NA  NA  NA  NA  NA
```



```
> str(my_data)
'data.frame':   58 obs. of  13 variables:
 $ ï..year                                : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ Month                                  : Factor w/ 12 levels "apr","aug","dec",..: 5 4 8 1 9 7 6 2 12 11 ...
 $ Average.Maximum.Temperature..Degrees.C.: num  8.6 8.6 10.6 13.6 15.3 18.6 20.6 18 17.8 14.7 ...
 $ Average.Minimum.Temperature..Degrees.C.: num  3 2.6 3.4 4.9 7.9 9 11.6 10.3 9 7.8 ...
 $ Mean.Temperature..Degrees.C.           : num  5.8 5.6 7 9.3 11.6 13.8 16.1 14.1 13.4 11.3 ...
 $ Highest.Temperature..Degrees.C.        : num  12.9 12 15.4 16.4 20.3 23.5 24.1 21.6 22 18.9 ...
 $ Lowest.Temperature..Degrees.C.         : num  -1.5 -2.5 -3.3 -1.1 1.7 3 5.1 3.3 3.7 -3.4 ...
```

Dataset containing year 2014-2018 data of Dublin temperature

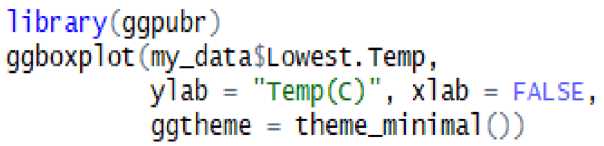
Type of data -> continuous
No. of sample -> one – sample
Hypothesis testing -> one t-test

HYPOTHESIS TESTING

H0 = Mean temperature of lowest temperature of Dublin equal to theoretical mean
H1 = Mean temperature of lowest temperature of Dublin **NOT equal** to theoretical mean

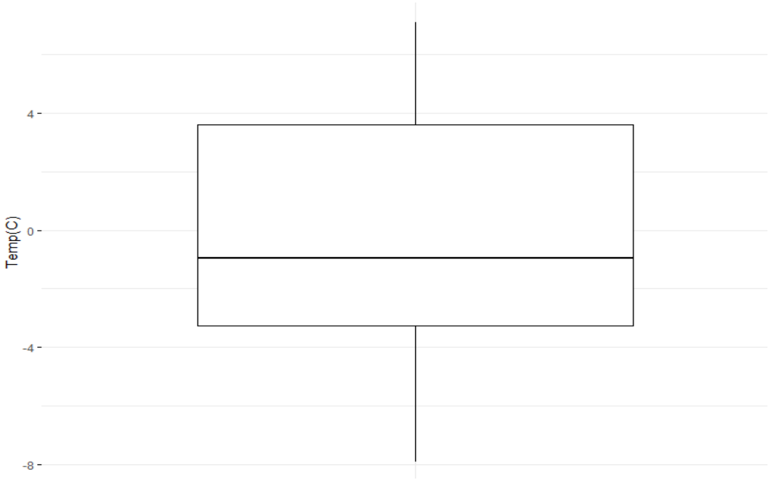
probability that IF the null hypothesis were true, sampling variation would produce an estimate that is further away from the hypothesized value than my data estimate, predetermined cutoff (0.05) is called the significance level of the test that's why I am using 0.05 value t-test is used to compare the mean of one sample to a known standard (or theoretical/hypothetical) mean (μ).

HYPOTHESIS TESTING



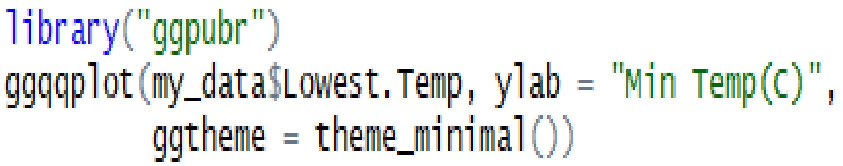
```
library(ggpubr)
ggboxplot(my_data$Lowest.Temp,
          ylab = "Temp(C)", xlab = FALSE,
          ggtheme = theme_minimal())
```

Visualize Dublin Temp data using box plots

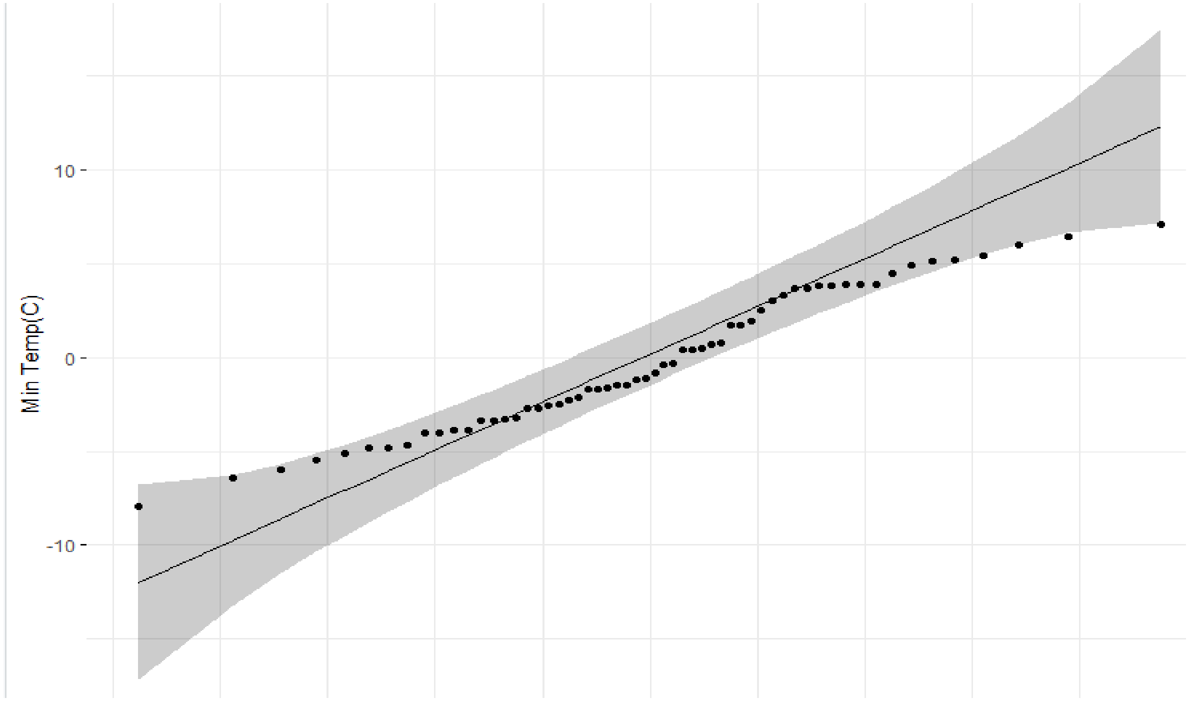


Preliminary test to check one-sample t-test assumptions

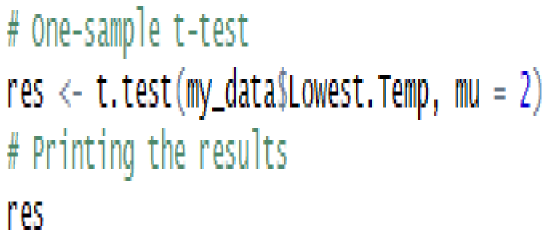
- **Visual inspection** of the data normality using **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution.



```
library("ggpubr")
ggqqplot(my_data$Lowest.Temp, ylab = "Min Temp(C)",
         ggtheme = theme_minimal())
```



From the normality plots, we conclude that the data may come from normal distributions.



```
# One-sample t-test
res <- t.test(my_data$Lowest.Temp, mu = 2)
# Printing the results
res
```

```
data:  my_data$Lowest.Temp
t = -4.4746, df = 57, p-value = 3.712e-05
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 -1.2144901  0.7731108
sample estimates:
 mean of x
-0.2206897
```

The p-value of the test is 3.712e-05, which is less than the significance level alpha = 0.05. We can conclude that the mean min.temp of the dublin is significantly different from 2 degree C with a p-value = 3.712e-05

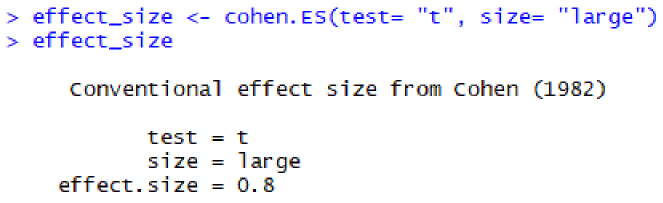
POWER ANALYSIS

maximise the power of statistical tests while maintaining an acceptable significance level and employing as small a sample size as possible.

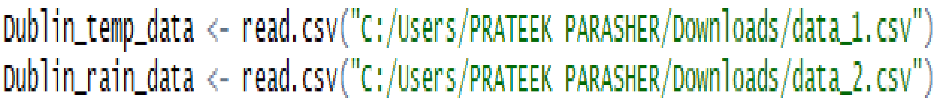
pwr.t.test(n=, d=, sig.level=, power=, type=, alternative=)

Where ...

- **n** is the sample size
- **d** is the effect size defined as the standardised mean difference
- **sig.level** is the significance level (0.05 is the default).
- **power** is the power level.
- **type** is a two-sample t-test ("two.sample"), one-sample t-test ("one.sample"),or dependent sample t-test ("paired"). A two-sample test is the default.
- **alternative** indicates whether the statistical test is two-sided ("two.sided")or one-sided ("less" or "greater"). A two-sided test is the default.



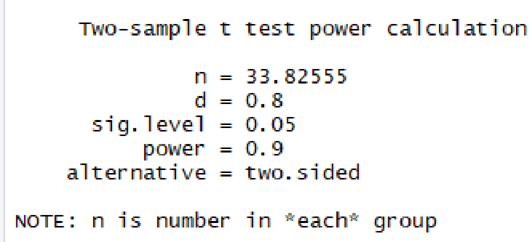
```
> effect_size <- cohen.ES(test= "t", size= "large")
> effect_size

     Conventional effect size from Cohen (1982) 

           test = t
           size = large
    effect.size = 0.8
```



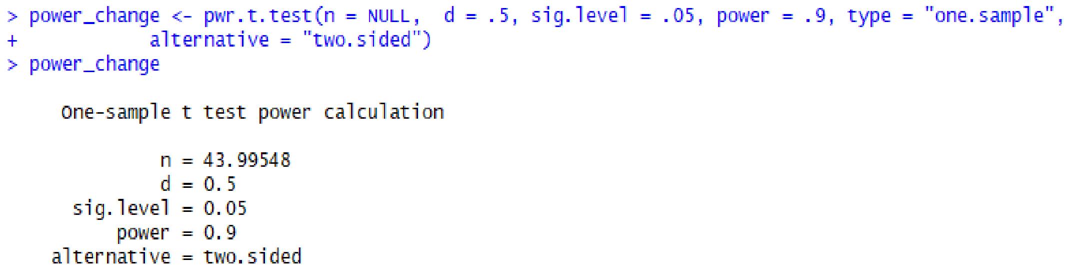
```
Dublin_temp_data <- read.csv("C:/Users/PRATEEK PARASHER/Downloads/data_1.csv")
Dublin_rain_data <- read.csv("C:/Users/PRATEEK PARASHER/Downloads/data_2.csv")
```

Power_change<- pwr.t.test(n = NULL, d = .8, sig.level = .05, power = .9, type = "two.sample",

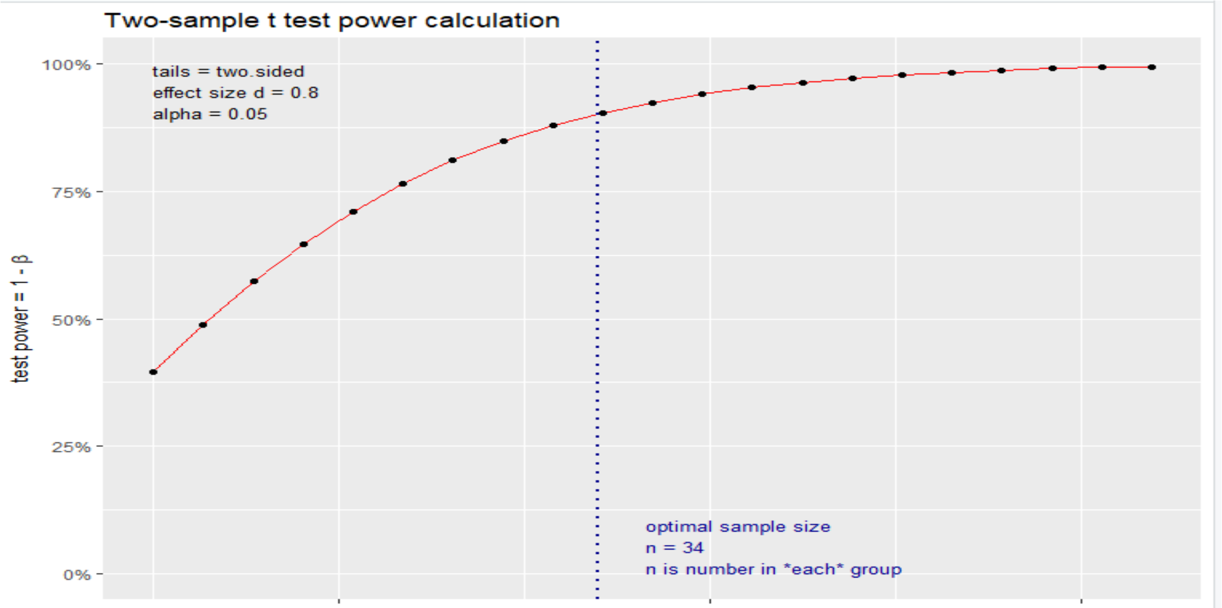
alternative = "two.sided")
plot(power_change)

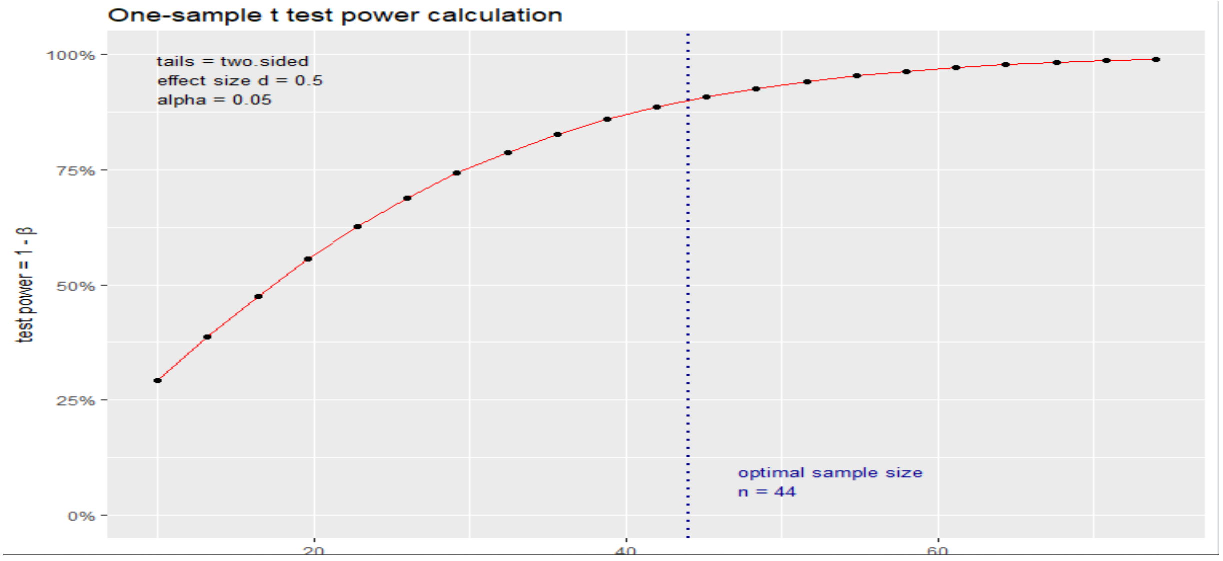


```
     Two-sample t test power calculation 

              n = 33.82555
              d = 0.8
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```



```
> power_change <- pwr.t.test(n = NULL,  d = .5, sig.level = .05, power = .9, type = "one.sample",
+            alternative = "two.sided")
> power_change

     One-sample t test power calculation 

              n = 43.99548
              d = 0.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```





Conclusion:- From this hypothesis test, I am concluding that the mean min.temp of the Dublin is significantly different from 2 degree C. By this result my initial null hypothesis false and I will continue my prediction and further findings with alternative hypothesis.

GitHub Link :- https://github.com/prateekparasher/web_scrap