LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: **James Connolly**

Assessment Title: **Prediction**

Work to be submitted to: **James Connolly**

Date for submission of work: **29-08-2018**

Place and time for submitting work: **3:00 pm**

To be completed by the Student

Student's Name: **PRATEEK PARASHER**

Class: **Msc in Big Data**

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: **29/8/2018**

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
Working with data and playing with that always interesting and I always try to find more out of the data i n this project initially my aim was finding correlation between two dataset & other hidden pattern throug h visualization and investing deeper into data in this project I worked on environmental I always want to k now how much pollution contribute for I found one dataset initially which was not in a condition to do an y testing or investigation so I clean the data pre-process that than used hypothesis testing and then the p rediction which is described in a documentation.

Data Pre-processing: –
During this pre-processing Labs which help a lot for my actual dataset in my lab practical 1 I cre ated data frame and examine total missing values after that removed those missing value in practi cal 1 lab I used diabetes dataset later in lab practical 2 I used cleaning postdata and I replaced all the missing value with suitable title and categorized the factor. After all these pre-processing it's time for use hypotheses analysis for my dataset

Web Scraping: - Twitter is a great source for sentiment data and social media mining furthermo re it is quite easy to get significant amounts of data to be able to scrape data from Twitter. Last y ear we all experienced the different climate conditions and multiple climate warning announced by govt. this lab I scrapped the data from twitter regarding climate warning like heatwaves warni ng, snow warning etc. after that I visualize the word cloud of scraped data from twitter.

Data Analysis The null hypothesis and alternate hypothesis are defined as-
H0 = The road accident is not related to the red weather warnings
H1 = The road accident is related to the red weather warnings
This data analysis task examines the relationship between in red weather warnings & road accid ents of Ireland. . I applied power test, t-test , correlation test. Correlation test is used to find the sample si ze to perform the power analysis with effect size of 0.5 with 80% certainty and no more than a 5% chance of inaccuracy. Then the sample of 29 records is used to do the power analysis. p-value is significantly h igh from 0.05. Therefore, this proves that the null hypothesis is true. It was seen that 29 data valu es are needed in each sample to have 80% of chance to reject the null hypothesis by determining the small p-value, which could reject the null hypothesis The road accident is not related to the re d weather warnings.

Predictive Modelling As, the null hypothesis in the above step was rejected therefore, road acci dent is not related to the red weather warnings. Further, carrying the research now, the relation b etween the accidents and red warning months in Ireland was found. To proceed, correlation betw een the accidents and red warning months in Ireland was calculated which came out to be -0.24 68042, which is a not a strong negative correlation, also the scatter graph was plotted To, procee d and get a understanding of predictive modelling the linear model was chosen to find the small c orrelation that still exists

Results

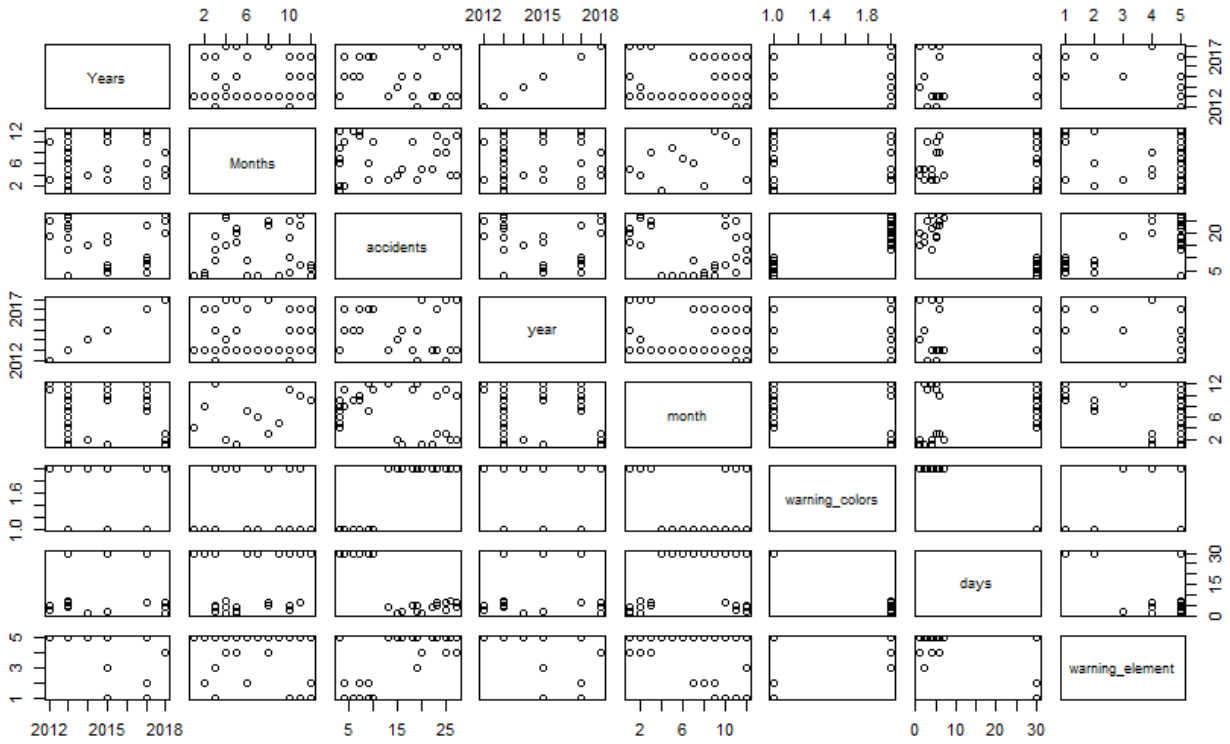

Fig 1 :- visualizing data

If we see diagonal line start from years , months , accidents ….. & warning_element, lets start with years if we see vertically down it's showing years 2012 , 2015 , 2018 then months if we see vertically up and left that showing 2, 6 …. & 12 same with accidents , warning_colors, days, warning_element.

Validation

```
cor(road_data$accidents, warning_data$month)
[1] -0.2468042
```

```
> AIC(linearMod)
[1] 212.0835
> AIC(polynomialMod)
[1] 202.5569
> BIC(linearMod)
[1] 216.1854
> BIC(polynomialMod)
[1] 208.0261
```

AIC and BIC linear and polynomial model values are high with correlation value -0.246

```
> lm_min_max_accuracy <- mean(apply(lm_actual_preds, 1, min) / apply(lm_actual_preds, 1, max))
> lm_min_max_accuracy
[1] 0.5360418
> pl_min_max_accuracy <- mean(apply(pl_actual_preds, 1, min) / apply(pl_actual_preds, 1, max))
> pl_min_max_accuracy
[1] 0.5214372
> lm_mape <- mean(abs(lm_actual_preds$predicted - lm_actual_preds$actuals) / lm_actual_preds$actuals)
> lm_mape
```

Linear and polynomial accuracy

```
> summary(lr_model)

Call:
lm(formula = accidents ~ month, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.164  -9.227   2.234   7.705  14.040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.2998     4.0171   4.555 0.000172 ***
month        -0.5339     0.5061  -1.055 0.303427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.466 on 21 degrees of freedom
Multiple R-squared:  0.05033,   Adjusted R-squared:  0.005108
F-statistic: 1.113 on 1 and 21 DF,  p-value: 0.3034
```

Summary Linear

```
> summary(pl_model)

Call:
lm(formula = accidents ~ month + I(month^2), data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.638  -5.090  -2.872   7.378  15.265

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.6942     5.1956   6.100 5.82e-06 ***
month        -7.3585     2.0869  -3.526  0.00212 **
I(month^2)    0.5363     0.1607   3.337  0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.774 on 20 degrees of freedom
Multiple R-squared:   0.39,     Adjusted R-squared:  0.329
F-statistic: 6.393 on 2 and 20 DF,  p-value: 0.007134
```

Conclusion

The assumption that the red warning weather data depends on road accidents in same months is failed with linear regression the p-value in linear regression is 0.303 and ideally should less that 0.05 adjusted r-squared which is inversely proportion with p value we got 0.005 which is very low ideally one should be 0.70 but on other hand with polynomial regression it seems that p value which is 0.007 which too low. all the parameter not supporting the relationship and any correlation but 2 degree polynomial model are better than linear model In future work we can try with increasing the degree of polynomial chances we might get better results.

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