

**Final Report: Exploratory Statistical Analysis On Los Angeles Payroll Data**

**Submitted to:**

Prof. Marija Bezbradica

School of Computing

Dublin City University

**Report Prepared By:**

Bharathvaj Devarajan

**16212388** bharathvaj.devarajan2@mail.dcu.ie

Subject : Statistical Analysis - CA660

**December 14, 2016**

**Declaration:**

In submitting this project, we declare that the project material, which we now submit, is our own work. We make this declaration in the knowledge that a breach of the rules pertaining to project submission may carry serious consequences. We also regard that both of us have contributed equally to the work presented.

**Introduction:**

This report aims to present an exploratory research done on the Los Angeles Payroll Dataset to better understand the Statistical Methods and use them in answering real life problems that the data poses. The Hypothesis Testing using different techniques like Normal Distribution, Students T-Test and ANOVA has been performed on various features of the dataset to evaluate the validity of the various assumptions presumed as Null hypothesis. The report also presents the work done to predict the Annual Salary and Average Health Cost making use of Linear regression and RandomForest (Decision trees) algorithms.

**Dataset:**

Los Angeles Payroll Data of four subsequent years starting from 2013 to 2016 under the Finance/Business Division has been taken for performing the statistical analysis. The dataset is self-explanatory giving an insight into the payroll of the employees in different Departments segmented as per their job titles over the four years. The motivation behind selecting this dataset is that it poses a lot of questions which are explored while working on this project, using the statistical methods. The dataset tells about the employment type, Hourly or Event rate of different personnel and their projected salaries, it depicts the Base pay, Quarterly pay and the annual pay of the individual and what kind of Bonuses, overtime pay and benefits they receive. The data also tells about the Job class and the paygrade of the employees.

**Technologies Used:**

Python 3.5:

To perform all the Hypothesis Tests, and to implement and run the tests Anaconda’s Jupyter Notebook is used.

Open refine:

Before starting to work on the dataset, the first step is to explore and fully understand what the data

Is. After comprehending the dataset we cleansed the data and removed the null values and the outliers, formatting the dataset. For data cleansing, Open Refine has been utilized to cleanse the dataset.

**Questions Explored:**

1. Using the Hypothesis testing with Normal distribution, determine whether the Annual Salaries of the employees increase in the Financial year 2016 as compared to 2015?
2. Using the Hypothesis testing with Student’s T-Distribution, check if the Average Health Cost of the Employees increases over a period of one year: 2014-2015?
3. Using ANOVA, find if in three subsequent years 2014, 2015, 2016 – does the mean of the Base salary of the Electricians remains the same?
4. Will the Linear Regressor be able to predict the Average Benefit cost based on Annual and quarterly Payments?
5. Will the Random forest Regressor making use of the Decision Trees be able to predict the Annual salary based on the Job title?

**Statistical Data Analysis**

**Normal Distribution:**

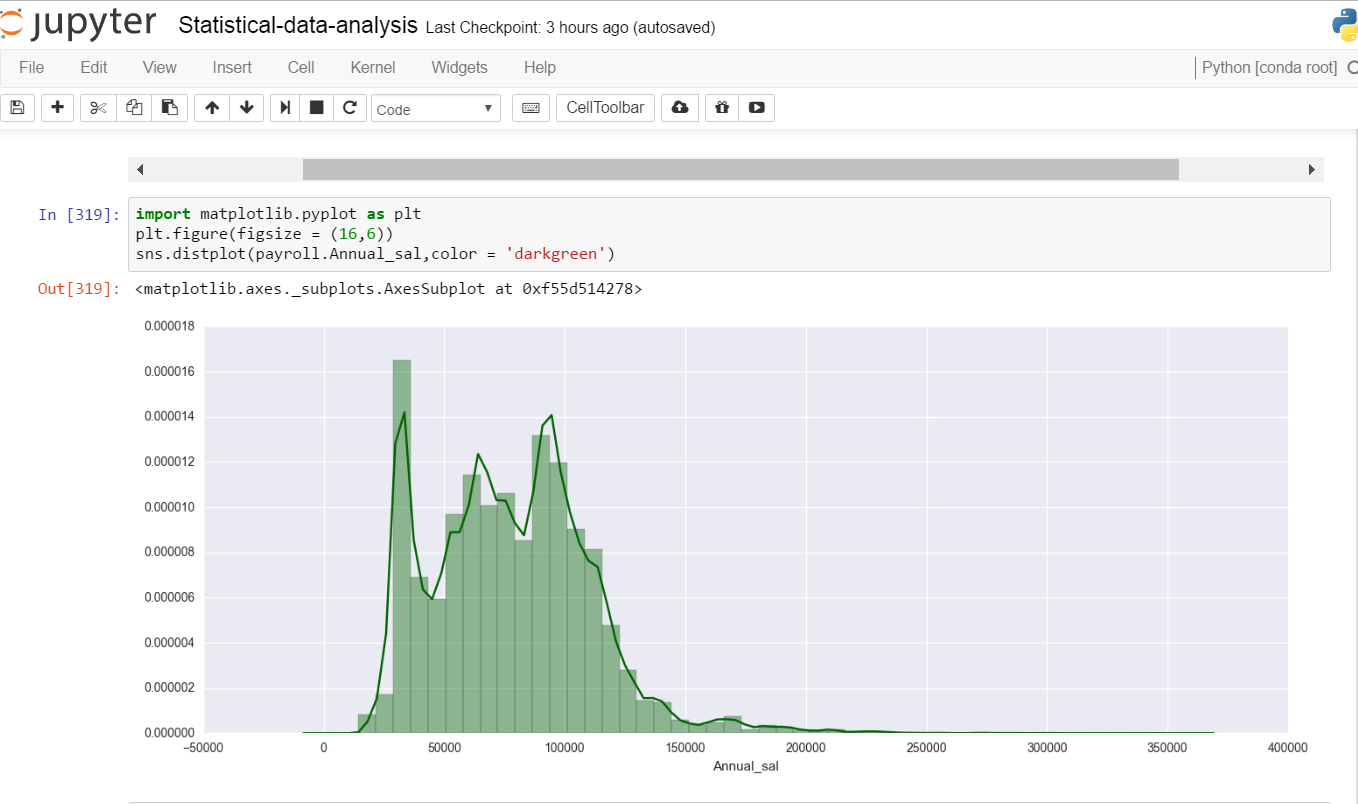
Normal distribution often called as Gaussian distribution or the Bell curve is the one the most common distributions used widely by the statisticians. The data is normally distributed and the three kind of averages a Normal distribution makes use of are:

Mean - Average of the values

Median - The middle value

Mode - Most occurred value

In our dataset, we have made use of the Normal Distribution test statistics in the Hypothesis Testing to answer the following question: whether the Annual Salaries of the employees increase in the Financial year 2016 as compared to 2015?



**Null Hypothesis H0:** Annual salaries of the employees increase in the year 2016 as compared to 2015.

**Alternate Hypothesis H1:** Annual salaries does not increase.

**Experiment:**

*Step1:* To check if the Payroll Data’s annual salaries follow the Normal Distribution a Histogram for all

the salaries over the four years 2013-2016 is plotted using the SEABORN library of Python. The

below histogram shows that the Annual Salaries do follow the Normal distribution curve and

Normal distribution Test statistics can be applied to prove the Hypothesis right or wrong.

*Step2:* After removing the null values from the Annual salary column, the payroll salaries are divided into two smaller datasets: Annual salary for the year 2015 and Annual salary for the year 2016.

*Step3:* The 2015 annual salaries are taken as the population and a 10% chunk of annual salaries from

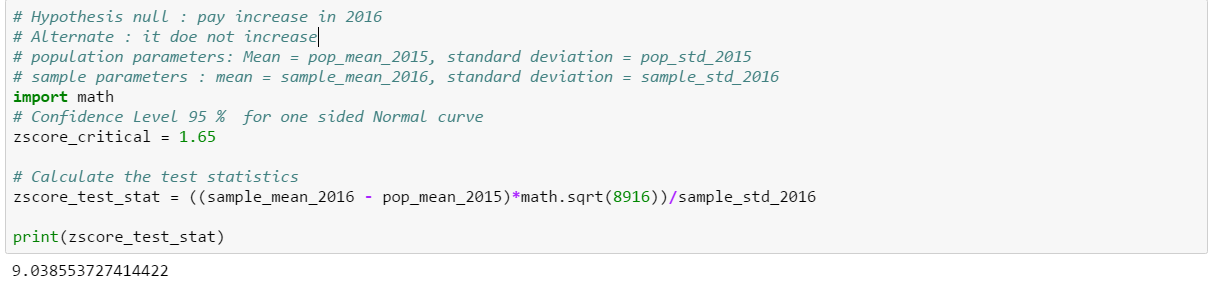
2016 is taken as the sample.

*Step4:* Population i.e 2015’s Annual Salary’s Mean and Standard deviation is calculated.

*Step5:* Sample i.e 2016’s Annual Salary’s Mean and Standard deviation is calculated.

*Step6:* With the 95% confidence, the critical value is set to 1.65 as this a one-sided distribution.

*Step7:* Based on all the above parameters, the Z\_Score is calculated.

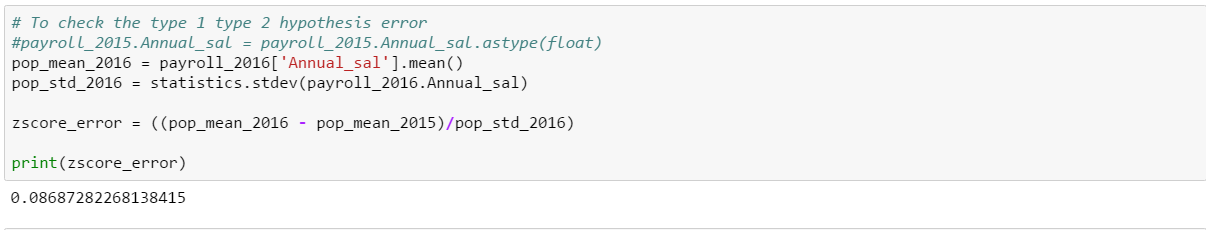


**Result**

As the Z\_Score is way beyond the test statistics and falls in the rejection region. The Null Hypothesis is Rejected and it is observed that the Annual Salaries does not increase in the year 2016.

**Type I Error:**

As a part of the analysis, we calculated the population mean of 2016 to check the true mean and standard deviation to compute the test statistics, it was observed that the test statistics from the population mean fell within the confidence interval, which denotes that the null hypothesis cannot be rejected. This kind of error is termed as a Type 1 Error in statistics. So it can be concluded that the sample we took for our analysis resulted in a Type 1 error as a consequence of which the Null hypothesis cannot be rejected.



**Student’s T Distribution:**

It is a statistical test that follows the T distribution and are used to check if two datasets are significantly different from one another. They make use of the degrees of freedoms to determine the T critical values which can be used to run hypothesis tests. This distribution is particularly useful in determining the trends in small datasets. A correction in the denominator is required by subtracting 1 to the sample size while calculating the standard deviation , this correction is called Bessel’s correction.

Using T distribution, we hope to answer whether the average health cost of an employee increases with year.

**Null Hypothesis H0**: The average health cost remains constant over the years.

**Alternate Hypothesis H1**: The average health cost increases with year.

**Experiment**:

*Step 1:* Following the same methodology like in case of Normal distribution, the payroll data is divided into two smaller datasets, Annual salary for the year 2013 and Annual salary for the year 2014.

*Step 2:* The 2013 annual salaries are taken as the population and a small chunk of 30 entries is taken as the sample from the 2014 population.

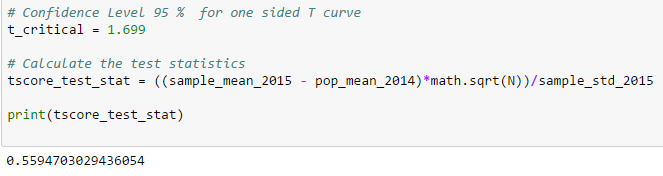
*Step 3:* Mean and standard deviation are then calculated for the sample and population data.

*Step 4:* Following a 95 % confidence interval, we calculated the t critical value as 1.699 (as it is a one-sided test).

*Step 5:* Test Statistics is then calculated making use of the parameters calculated in step 3.

*Step 6:* A comparison is then made between the Test statistics and the T critical value to decide whether to hold or reject the null hypothesis.

**Result:** The T statistics was evaluated as 0.5594703029436054 ~ 0.56, which is within the confidence interval, so the Null hypothesis is accepted and thus we could answer that the health cost remains constant over the years.



**ANOVA (*An*alysis *O*f *Va*riance):**

ANOVA is an acronym for Analysis Of Variance, and as the name suggests, it is the comparison of the variances. One-way Anova comes into picture when the variances of 3 or more populations are to be compared and the multiple T-Tests cannot be used as the Type I with 95% confidence compounds to .857 resulting into 14.3% which is not a good way to compute the Hypothesis Tets. One-Way ANOVA makes use of the F Distribution which can be defined as square of the Variances Between / square of the Variances Within.

Variance Between + Variance Within = Total Variance.

In our dataset to know whether the mean of the Base salary of the Electricians remains the same over the years 2014, 2015 and 2016 we have made use of the ANOVA.

**Null Hypothesis H0:** The Mean of the Base Salary of the Electricians remain the same over three years.

**Alternate Hypothesis H1:** The Mean Varies.

**Steps:**

*Step1:* The Dataset is again segregated into three smaller datasets as per the years.

*Step2:* These three datasets containing all the Job titles are again broken down into three datasets

where the Job Title is – ELECTRICIAN.

*Step3:* Three samples from these distinct datasets are taken with the sample size of 35 to further the

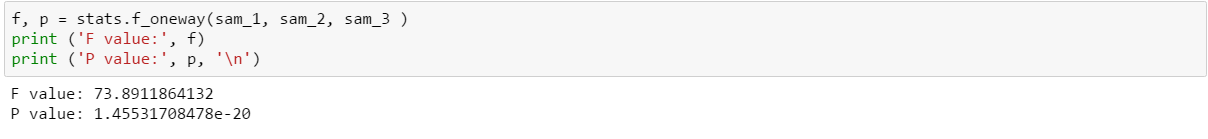
analysis.

*Step4:* These 3 samples for which the Hypothesis testing using ANOVA is to be carried out now contains

35 entries each for the Basic Pay of the Electricians.

*Step5:* F\_ONEWAY function of the STATS library in Python is used to calculate the P-Value and the

F-Ratio.



**Result**

As the P value <= 0.05 (95% confidence), the Null Hypothesis is Rejected which concluded that the mean of the salaries of the Electricians changes over the years.

**Linear Regression:**

It is a method used to predict the outcome of the dependent variable based on the values of the independent variables.

It can be represented by the formula:

Y = bX + a.

Where Y is the dependent variable, X denotes the independent variable.

a is the Y intercept and b is the slope of the line.

In our report, we made use of the Linear Regressor to predict the “Average Benefit Cost” based on the annual and quarterly payments.

**Null Hypothesis H0**: The Average Benefit cost remains constant over the years (No Linear relationship).

**Alternate Hypothesis H1**: The Average cost increases with increase in yearly and quarterly payments (Linear relationship).

**Experiment**:

Step 1: In python, Import the LinearRegression() function from Sklearn library.

Step 2: Clean the data as per requirement and split the data into training and testing, the split ratio should be 7:3.

Step 3: Train the Model using the training data.

Step 4: Compute the accuracy of the model using Sklearn.metrics library, in our case the accuracy was 0.51704167228296 / 0.5202426725733107 for training and testing respectively.

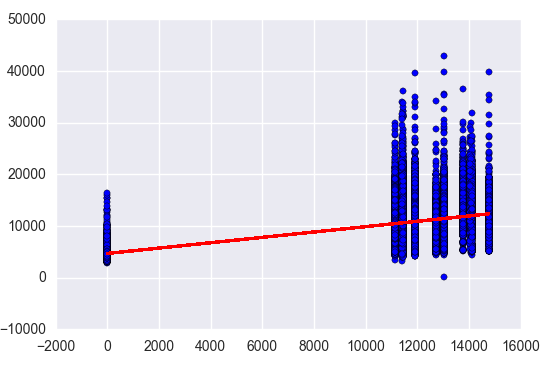
Step 5: Calculate the best-fit line for visualizing the split between the data points.

Step 6: Plot the scatter plot with the regression line.

**Result:** With only 52 % accuracy, there is not enough to suggest that a linear relationship exists between the variables, hence the Null hypothesis is accepted.

**Plot:** The clustering around “0 benefits” suggests that people with lower salary are not entitled for benefits. Since more than 250000 data points are plotted together, we can see overlapping clusters.

The line segregates the points in a way in which the sum of the mean square distance from the line for all data points is minimum.



**Random Forest Regression**

It makes use of many randomized Decision trees for predicting real valued numbers based on the independent variables. A decision tree follows a top-down approach to cluster the data instances subsets which contains similar results, but in a regression tree the target value is predicted by creating a model using the independent variables. Once the model is fitted the training data instances are then split at several nodes. The node with the least “Sum of Squared Error(SSE)” is finally selected.

We made use of Random Forest Regression to predict the Annual Salary based on the Job Title.

**Null Hypothesis H0**: The Annual Salary cannot be predicted based on the Job Title.

**Alternate Hypothesis H1**: The average salary can be predicted based on the Job Title.

**Experiment**:

Step 1: In python, Import the RandomForestRegression() function from Sklearn library.

Step 2: Since the Job title is a text field, vectorize it using extraction text feature of sklearn library.

Step 3: Clean the data as per requirement and split the data into training and testing, the split ratio should be 7:3.

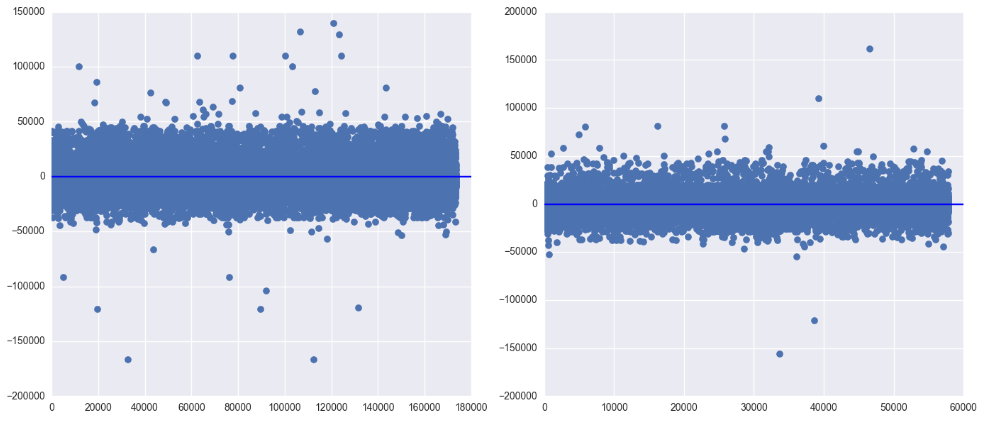
Step 4: Train the Model using the training data.

Step 5: Compute the accuracy of the model using Sklearn.metrics library, in our case the accuracy was 0.959925561813541 / 0.9581456436879675 for training and testing respectively.

Step 6: Plot the scatter plot for the test and train models.

**Result:** With 96 % accuracy, we can safely state that the model works reasonably well for predicting the Annual salary based on the Job title.

**Plot**: The plot shows the scatter plot of the data points with training and testing data. We can clearly visualize that there is a uniform distribution along the line which substantiates the claim that this model can be used to predict the Annual salary based on the Job Title and that there is a strong correlation between these variables.



**Challenges Faced:**

1. The data was not clean and required a lot of cleansing to remove the insignificant fields and null values.
2. Outlier analysis was required before conducting the experiments to avoid erroneous results.
3. Due to the large size of the data, a single sample test may not always return the correct result.
4. Selecting the features to train the models for Linear regression and Random forests (Decision Trees) require deep knowledge about the problem.

**Conclusion:**

In our report, we have done an explanatory analysis using statistical methods to predict patterns in the payroll data of Los Angeles. Running Hypothesis testing using techniques such as Normal, Student-T, Anova, Linear Regression and Random Forests helped us answer key questions in the datasets which can be used to solve real world problems. With further improvements, these techniques can be used to develop an application which can be leveraged to decide different components of pay such as allowances, bonus, health benefits etc.