# Provide a brief explanation about possible advantages generated by business analytics in higher education sector problem solving specific to university education system.

The volume of data available is definitely on the higher side and this is expected to continue to grow. Smart use of the available data can be considered as a critical component which would help any business survive and grow from success to success. Most of the industries have already spotted the importance of using data therefore huge amounts of data are accumulated in-order generate reports and useful stats which would help the organization in the long run.

The potential of data analytics in universities is massive. As XYZ consists of more than 10,000 students, 300 lecturers, number of faculties and the university has been recognized as one of the top 5 universities in the world consecutively for the last 5 years according to the global university ranking. It is quite obvious that the education provided by the XYZ University is their main attraction point, thanks to the highly experienced panel of lecturers. Along with quality education curriculum activities is also provided by the university.

The amount of data the above mentioned few factors will give rise is absolutely massive it is the responsibility of the data scientist to use the data generated wisely in-order to maximize the opportunities and minimize future threats.

Therefore the following are some of the key factors the XYZ University has to consider;

1. Quality Education
2. Panel of Lecturers
3. Extra-curriculum activities

In-order to maintain the standard of education the university has managed to provide so far they can use a subset of data analytics called learning analytics, which uses data footprint left by the higher education students during their courses, which would help to ensure students get maximum benefit by giving importance to the following factors;

* Follow an ethical framework or policy
* Identify and improve programs in-demand and beneficiary
* Embed possible game changing programs to the universities strategic plan
* Identified the tailored needs of the university

Other factors the university should be concerned about when dealing with data are;

* Ensuring the accuracy of data
* Ensuring the security of the data
* Ensuring processes such as authentication and authorization are used to protect data.
* Other good practices

Lecturers is a another equally important asset for the university, Institutions hold data on their staff, including the number of people employed full-time and part-time, the number at each level and within each faculty, and staff equal opportunity data Institutions generally hold data on staff, including the number of staffs, the number of full-time and part-time, the hierarchy of the staff and other staff related data. These data can be used to;

* To give extra benefits to the staff
* Identify gaps
* Fulfil gaps
* Give promotions, increments and other benefits.

Other data the University will hold on includes the following;

* Student record data
  + Name
  + Age
  + Address
  + Ethnicity
  + Etc.
* Admissions and application data
* Financial data
* Alumni and historical data
* Course data
* Estates and facilities data
  + Lecture theatres
  + Lecture halls
  + Computer labs
  + Science laboratories,
  + Room capacity
  + Equipment
  + Etc.
* Etc.

This available data can also be considered as a massive asset of the University, this data would be of no use if they aren’t used rightly. Correct use of data will help the university in its day-to-day operational decision as well as long term business and strategic decisions. Following are some operations where the assistance of data will be required;

* Creating time table for semesters
* To work out the number of lecturers, tutorials and labs to schedule per module.
* The rooms and equipment required

On a strategic level as well data can be used. Which would help the university make future executive decisions based on data regarding available faculties, student enrolment and relevant historical and projected data.

# Find out whether or not lecturer salaries, lecturer years of service and lecturer years since Phd comply with normal distribution pattern using normality tests using suitable hypothesis. Include the results of the finding along with the sub conclusion using normality tests; Shapiro-Wilk Test, Anderson-Darling Test and Lilliefors-Test supported by Monte Carlo graphical simulations.

The main three data to be found out whether they comply with normal distribution pattern are;

1. Lecturer salaries
2. Lecturer years of service
3. Lecturer years since Phd

Above mentioned data will be tested normality tests; Shapiro-Wilk Test, Anderson-Darling Test and Lilliefors-Test.

Null Hypothesis would be they are normally distributed

Alternative Hypothesis would be they aren’t normally distributed

## **Test for Lecturer salaries**

### Shapiro-Wilk Test

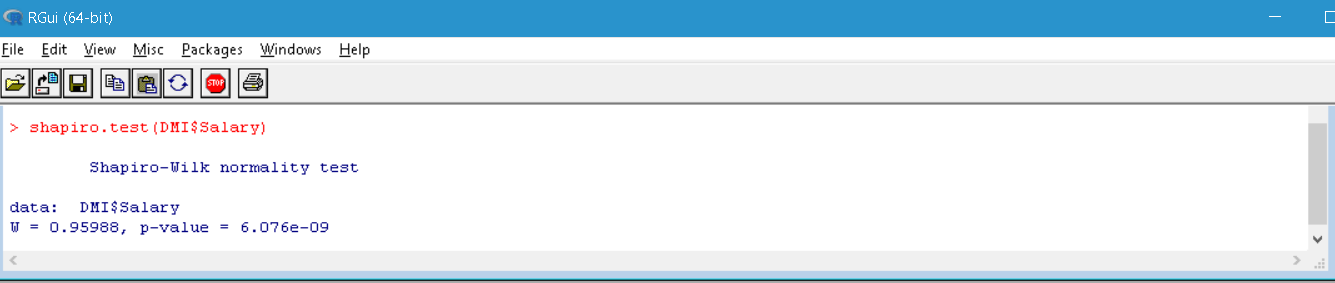


Figure Shapiro-Wilk Tes

### Anderson-Darling Test

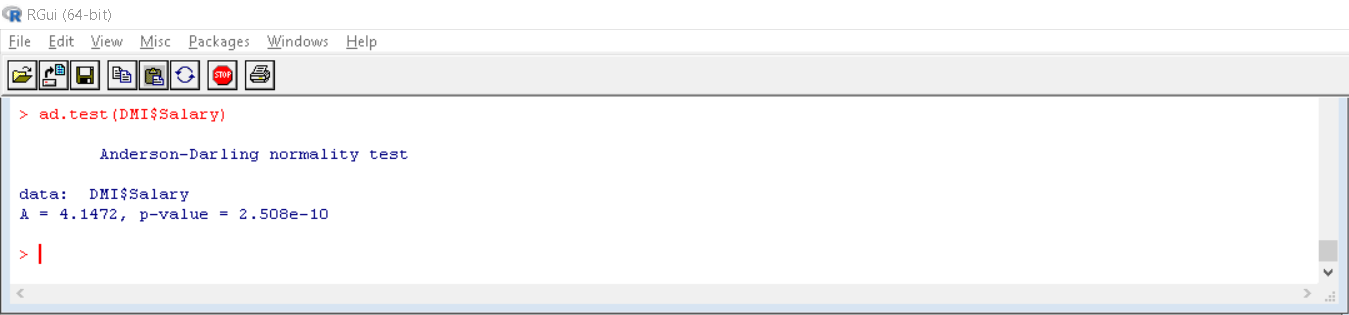


Figure Anderson-darling test

### Lilliefors-Test

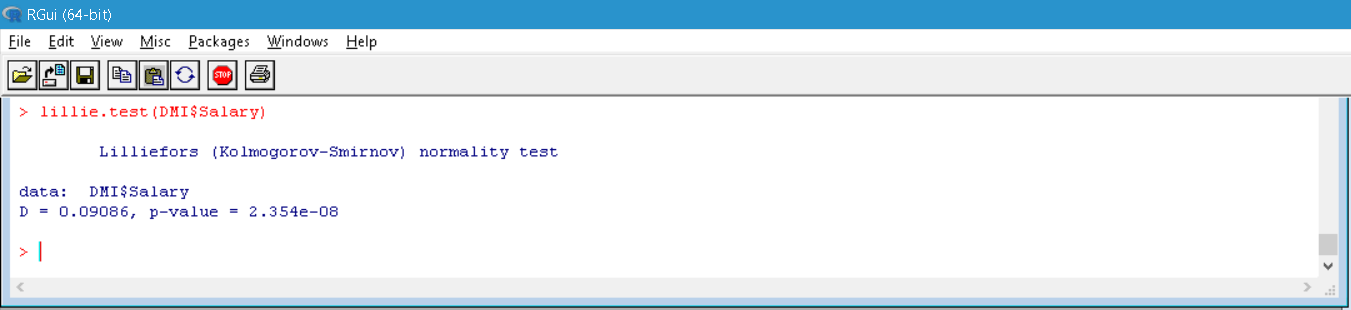


Figure Lilliefors-test

## **Lecturer years of service**

### Shapiro-Wilk Test

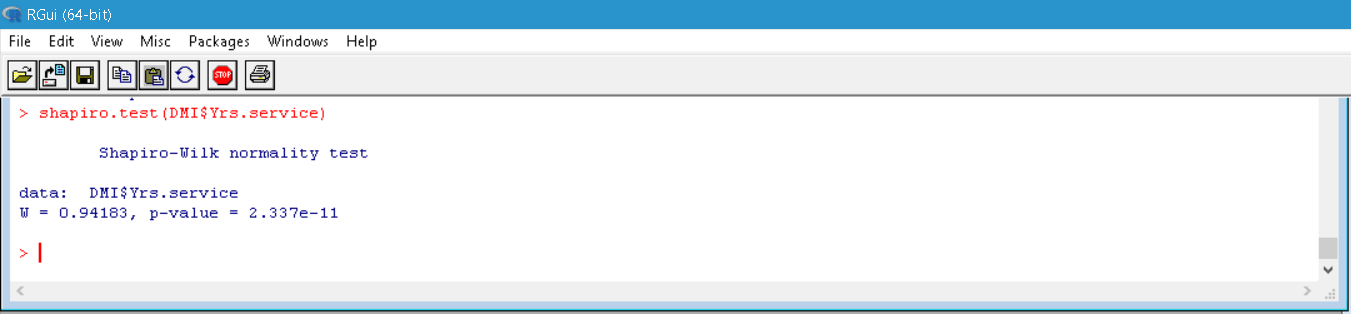


Figure Shapiro-Wilk Test

### Anderson-Darling Test

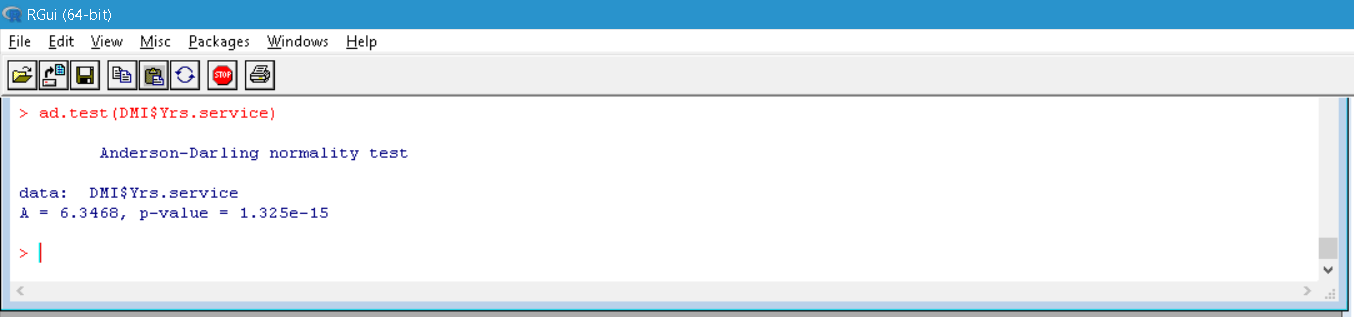


Figure Anderson-darling test

### Lilliefors-Test

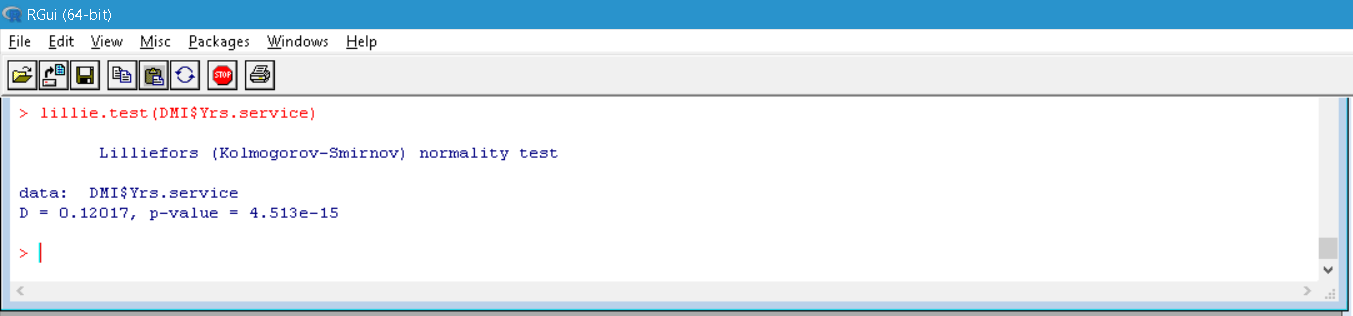


Figure Lilliefors-test

## **Lecturer years since Phd**

### Shapiro-Wilk Test

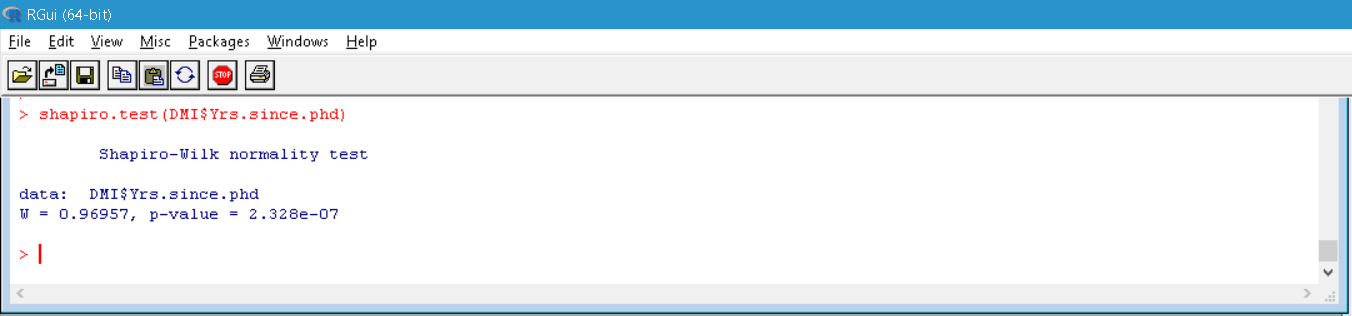


Figure Shapiro-Wilk test

### Anderson-Darling Test

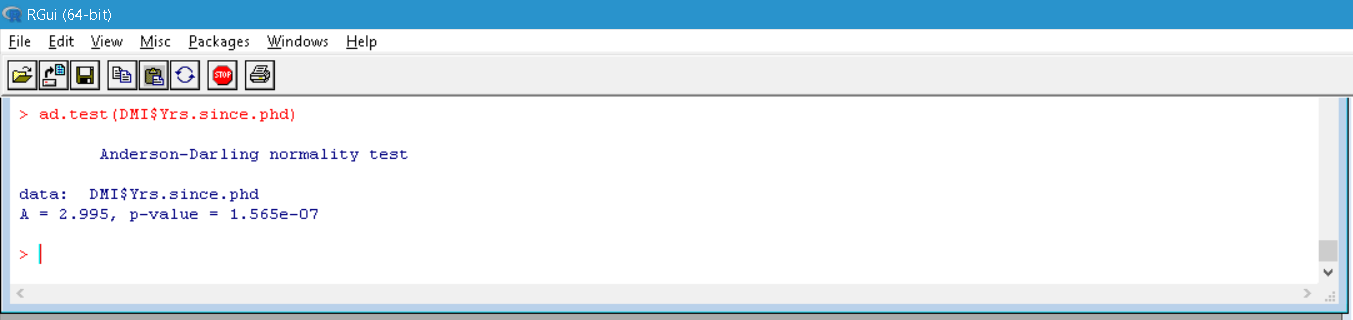


Figure Anderson-darling test

### Lilliefors-Test

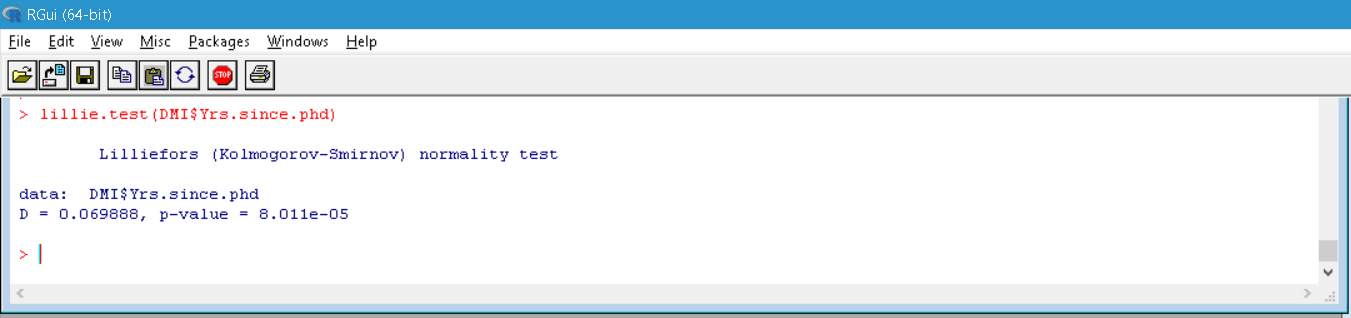


Figure Lilliefors-test

Furthermore to illustrate that the datasets aren’t normally distributed, some histograms for data sets lecturer salary, lecturer years of service and lecturer years since PhD are given below.

#### Histogram of lecturer salary

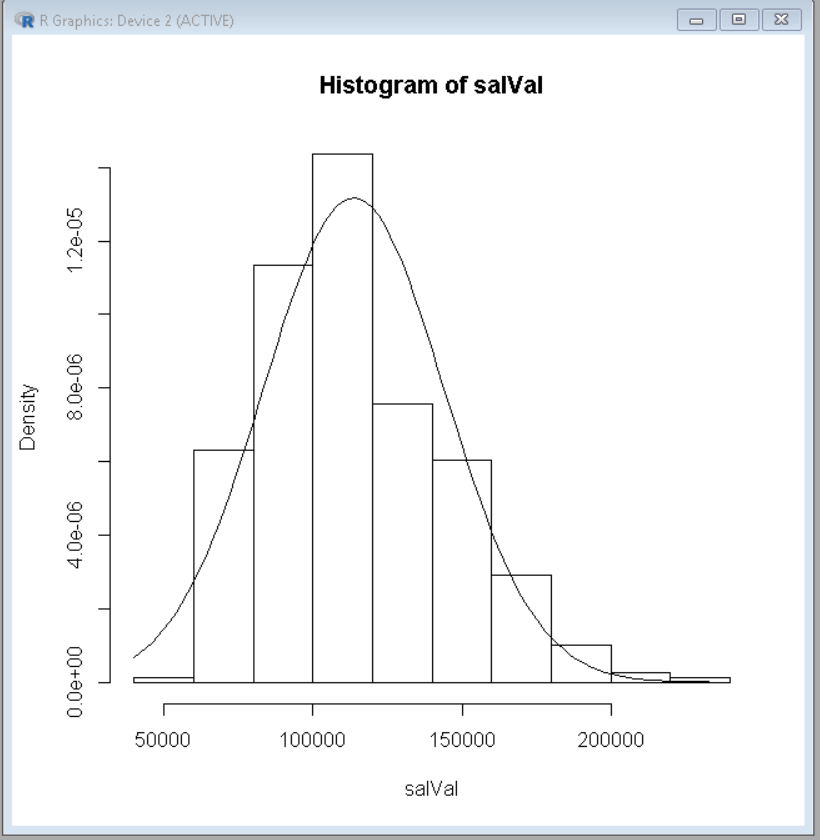


Figure lecturer salary histogram

#### Histogram of lecturer years of service

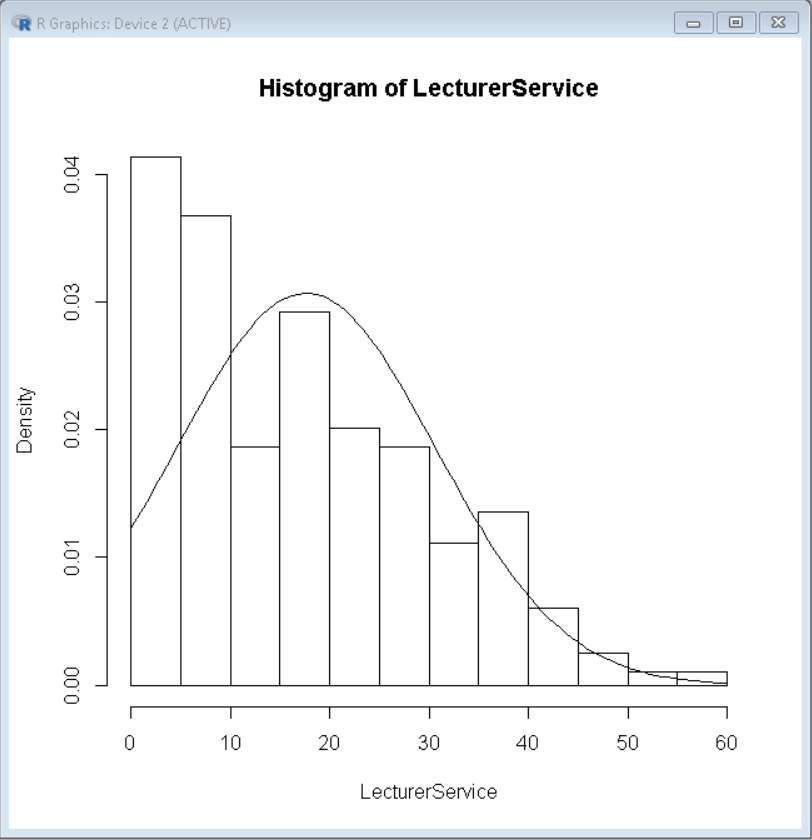
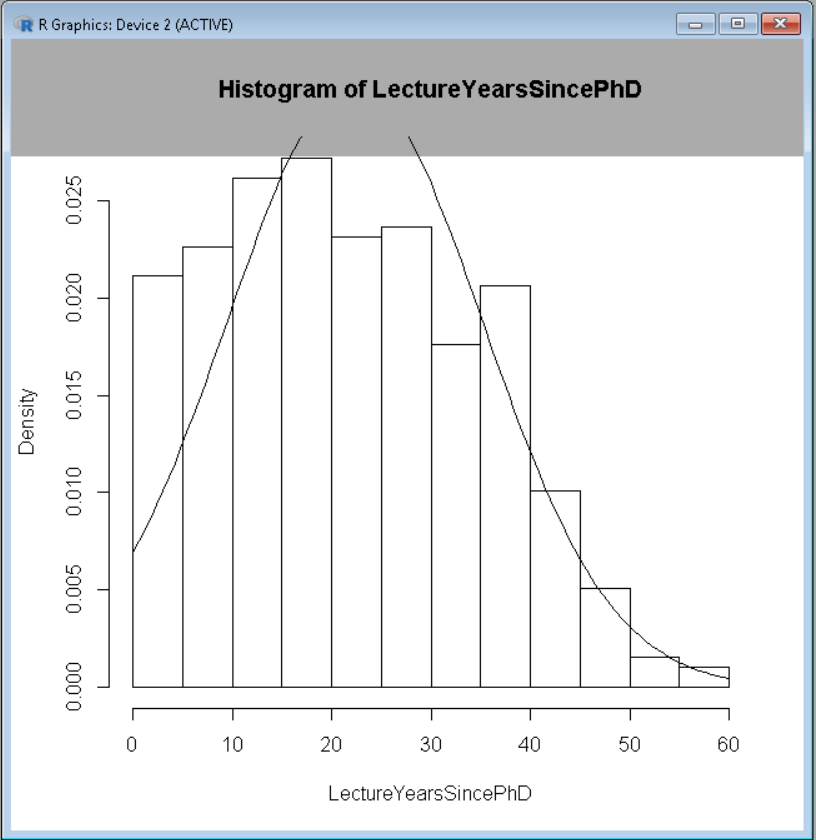


Figure lecturer years of service histogram

#### Histogram of lecturer years since PhD



## **Conclusion**

Null Hypothesis = H0 = normal distribution

Alternative Hypothesis = H1 = aren’t normal distribution

After conducting Shapiro-Wilk test, Anderson-darling test and Lilliefors-test these are the p values obtained for lecturer salaries, lecturer years of service and lecturer years since Phd;

**Lecturer salaries**

* 6.076e-09 = 6.076\*10-9  (Shapiro-Wilk Test value)
* 2.508e-10 = 2.508\*10-10 (Anderson-Darling Test value)
* 2.354e-08 = 2.354\*10-8 (Lilliefors-Test value)

**Lecturer years of service**

* 2.337e-11 = 2.337\*10-11 (Shapiro-Wilk Testvalue)
* 1.325e-15 = 1.325\*10-15 (Anderson-Darling Test value)
* 4.513e-15 = 4.513\*10-15 (Lilliefors-Test value)

**Lecturer years since Phd**

* 2.328e-07 = 2.328\*10-7 (Shapiro-Wilk Test value)
* 1.565e-07 = 1.565\*10-7 (Anderson-Darling Test value)
* 8.011e-05 = 8.011\*10-5 (Lilliefors-Test value)

Considering the chosen alpha level is 0.05, therefore;

P < 0.05 🡪 statistically significant difference.

P > 0.05 🡪 no statistically significant difference.

If we consider the p values we’ve obtained for the lecturer salaries, lecturer years of service and lecturer years since Phd from the different tests they are all below 0.05. So there is less than 5% possibility that the null hypothesis (normal distribution) is true. The histograms above too clearly illustrates that that the data sets aren’t normally distributed. So the null hypothesis of each dataset could be rejected in favour of the alternative hypothesis.

# Clearly justify whether or not lecturers’ salary mean values and salary variances are different with respect to their positions using one way Analysis of Variance (**ANOVA**)and **boxplot** analysis

## **ANOVA – Analysis of variance**

ANOVA is a statistical method used to compare the means of two of more groups.

*Hypothesis in ANOVA*

The mean of the different groups of the population will be equal is the null hypothesis of ANOVA. If we have two groups;

µ = µ

Considering µ to be the mean of the group.

### **ANOVA one way analysis of variance for Salary vs Position**

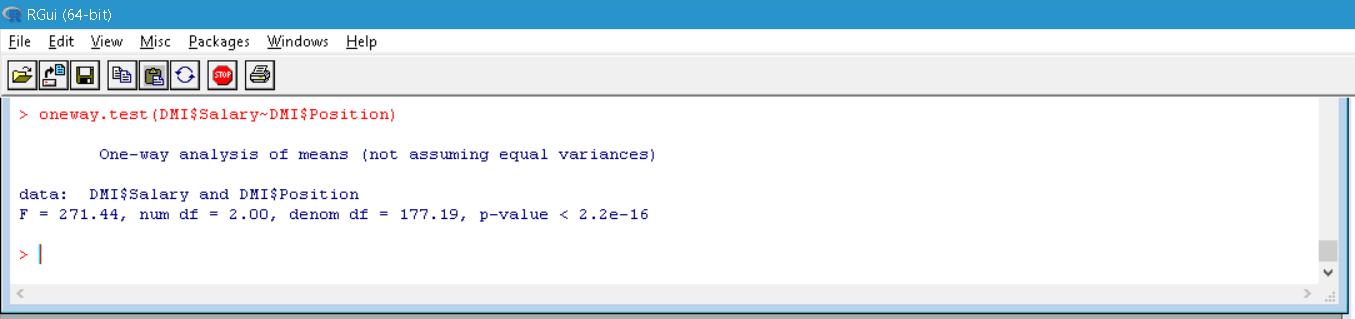


Figure ANOVA

For any hypothesis test there should be a significance level, assuming a significance level of 5%.

Significance level = α = 0.05

**Following are the values got form the ANOVA test done on Salary vs Position.**

* F = 271.44
* Numerator Degree of freedom = 2.00
* Denominator Degree of freedom = 177.19
* P-Value < 2.2e.16 = 2.2\*10-16

Using the numerator Degree of freedom and the denominator Degree of freedom we can find the critical F statistic value where α = 0.05.

Critical F statistic Value = 2.9957.

## **Boxplot analysis for salary vs position**

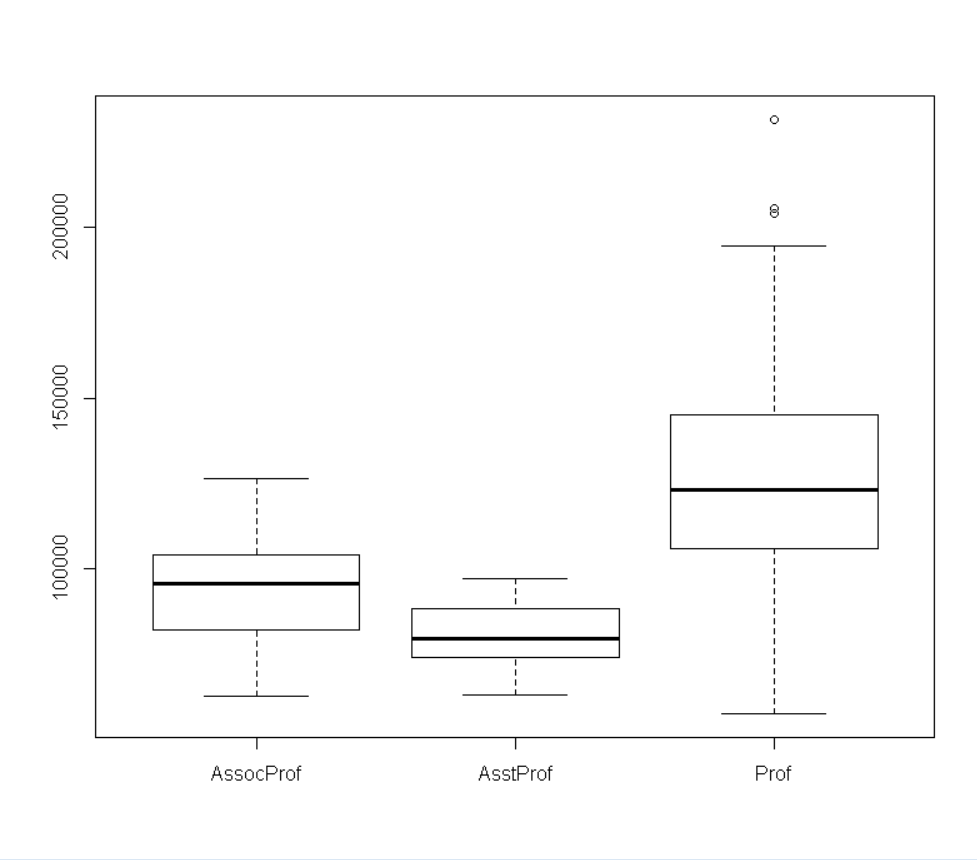


Figure Boxplot Analysis

Boxplot analysis is a good method to display the mean, spread of data, first quartile range, second quartile range, third quartile range and fourth quartile range. Above image shows the boxplot diagram for salary vs position of the university.

Following information can be observed clearly from the above diagram;

* Median of Associate Professor, Assistant Professor and Professor
* Range of Associate Professor, Assistant Professor and Professor
* First quartile range, second quartile range, third quartile range and fourth quartile range of Associate Professor, Assistant Professor and Professor
* Difference in salary for the different positions.

## **Conclusion**

Null Hypothesis = H0 = means are equal

Alternative Hypothesis = H1 = means aren’t equal

Null hypothesis of an ANOVA analysis of variance is that the mean of a population of different groups is equal. In-order for the null hypothesis of the ANOVA one way analysis of variance done to lecturers’ salary mean values and salary variances with respect to their positions to be true, the mean should be equal. The following are the values derived from the test;

* F = 271.44
* Numerator Degree of freedom = 2.00
* Denominator Degree of freedom = 177.19
* P-Value < 2.2e.16 = 2.2\*10-16
* Critical F statistic Value = 2.9957

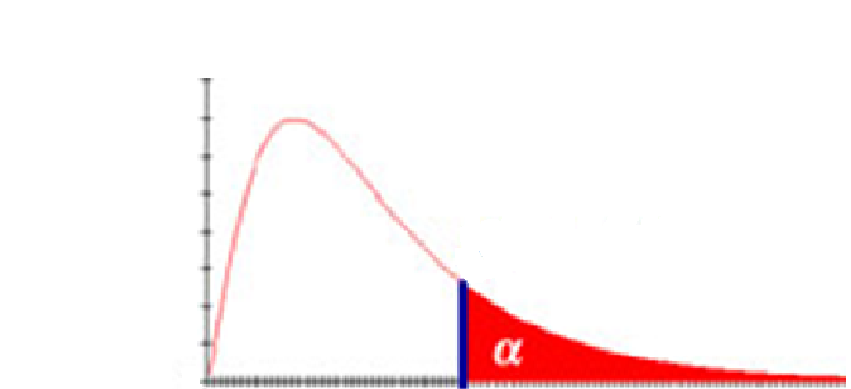
Considering the significance level to be 0.05;

P < 0.05 🡪 statistically significant difference.

P > 0.05 🡪 no statistically significant difference.

The P-value we’ve derived is 2.2\*10-16 in another words the probability of getting the null hypothesis true is less than 5%.

The critical F statistic value obtained using numerator’s degree of freedom and denominator’s degree of freedom is 2.9957, which is nowhere near the F value 271.44. Which makes it possible to reject the null hypothesis. Below F distribution graph shows clearly how far away the F value is from the critical F value.



271.44

2.9957

Figure F distribution graph

Furthermore Boxplot analysis for salary vs position clearly shows that the range and median of the different positions are significantly different. So it is clear that there is a considerable difference between the salaries of the lecturers’ of the different positions.

# Clearly justify whether or not there are relationship can be identify between lecturer salary –lecturer years of service, lecturer salary- lecturer years since Phd and lecturer salary-lecturer rank using correlation analysis supported by suitable hypothesis.

## **Correlation**

Correlation models are used to find the relationship between two variables. Outcomes of a correlation model are;

* Positive correlation (coefficient; ῥ = 1)
* Negative Correlation (coefficient; ῥ = -1)
* No Correlation (coefficient; ῥ = 0)

Whenever there is a positive linear relationship, when x variable gets larger y variable too gets larger and vice versa.

Whenever there is a negative linear relationship, when x variable gets larger y variable too gets smaller and vice versa.

Whenever there isn’t a linear model that does describe the relation between the variables that well, we consider there is no correlation.

**The null hypothesis and alternative hypothesis in correlation;**

H0 = ῥ = 0

H1 = ῥ ≠ 0

## **Correlation between lecturer salary – lecturer years of service**

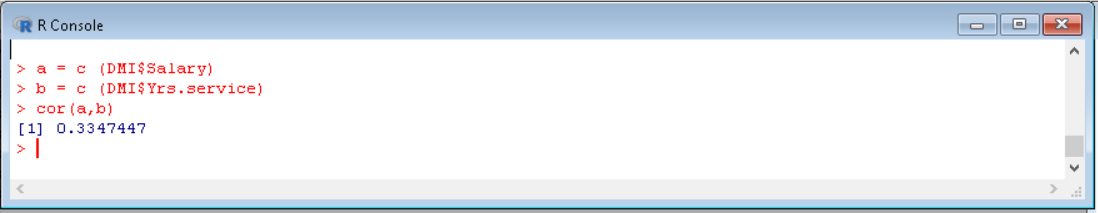


Figure Correlation between Salary & Yrs. service

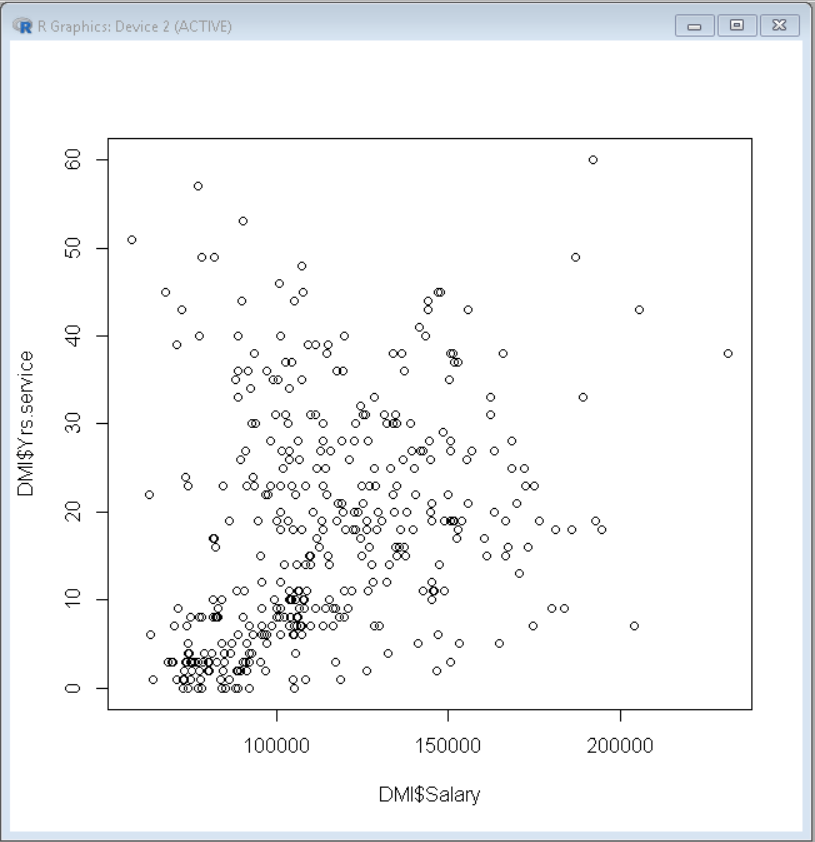


Figure Scatter plot of Salary & Yrs. Service

## **Lecturer salary- lecturer years since Phd**

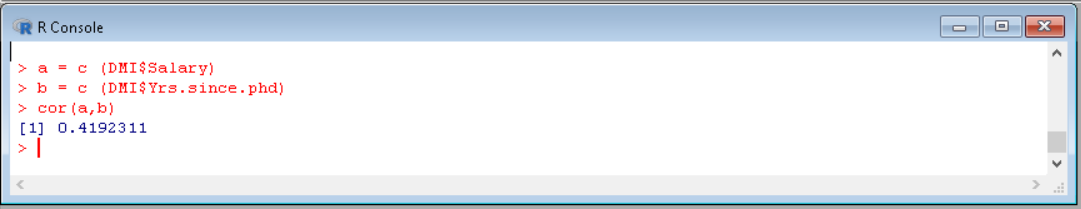


Figure Correlation between Salary & Yrs. since PhD

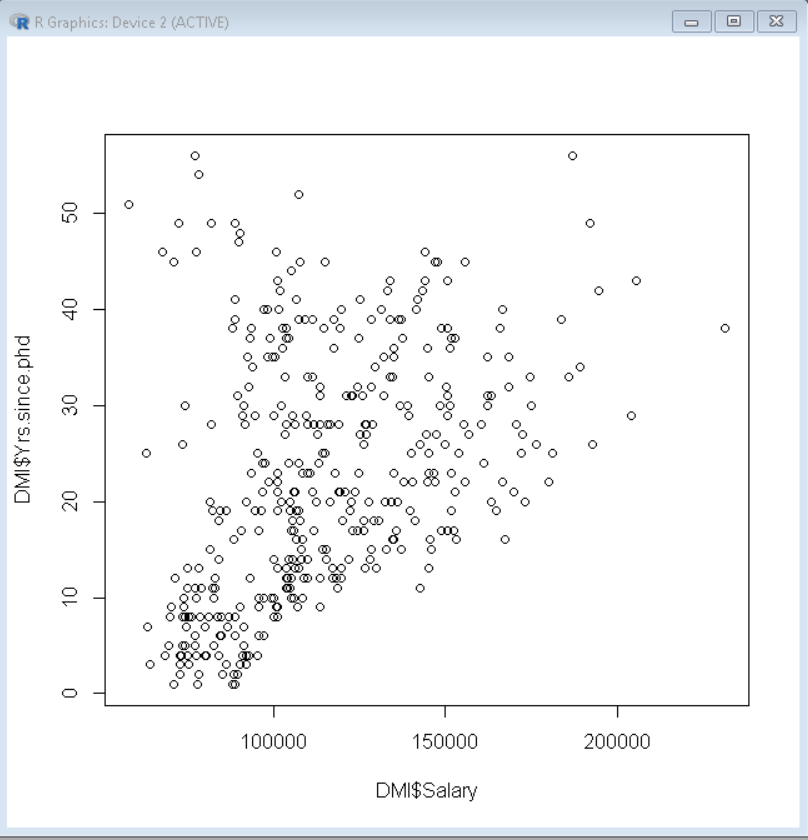


Figure Scatter plot of Salary & Yrs. Since PhD

## **Lecturer salary-lecturer rank**

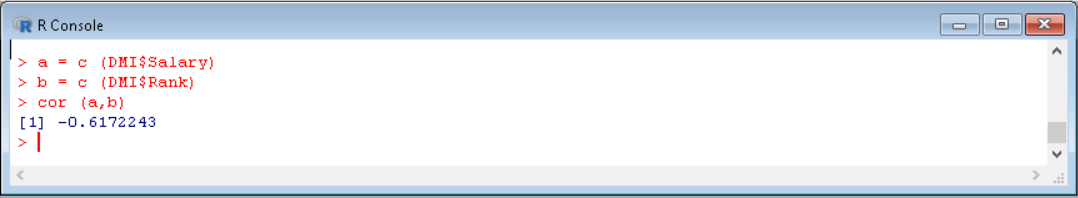


Figure Correlation between Salary & Rank

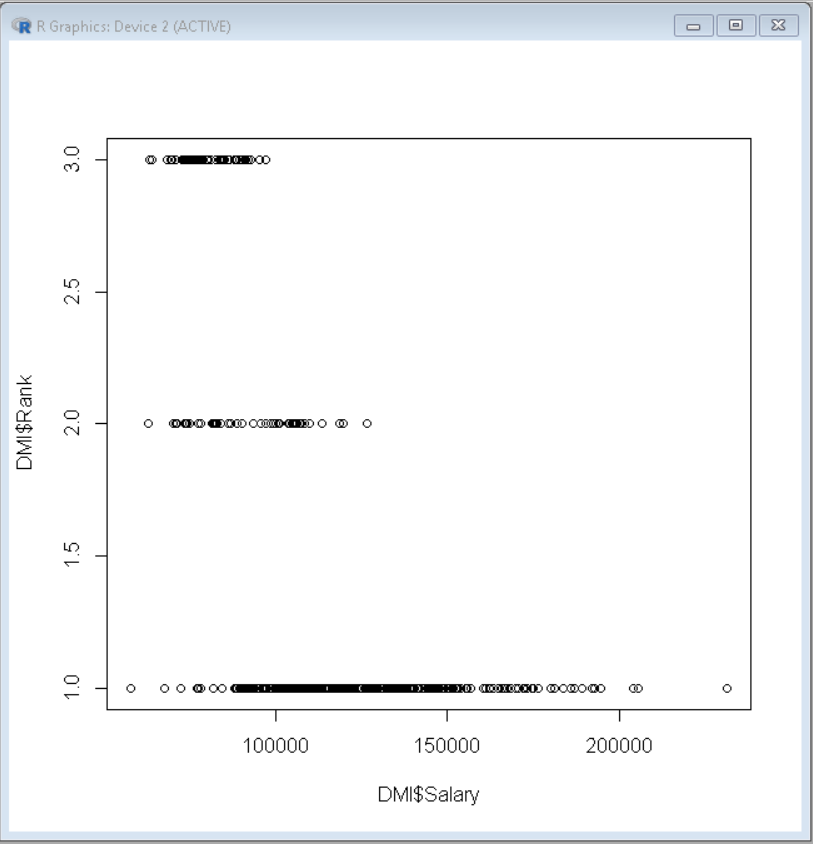


Figure Scatter plot of Salary & Rank

## **Conclusion**

* Positive correlation (coefficient; ῥ = 1)
* Negative Correlation (coefficient; ῥ = -1)
* No Correlation (coefficient; ῥ = 0)

If we consider the correlation between lecturer salary and years of service the correlation coefficient value is 0.3347447. Which means there is slight positive linear relationship between the lecturer salary and years of service.

If we consider the correlation between lecturer salary and lecturer years since Phd the correlation coefficient value is 0.4192311. Which means there is better positive linear relationship between the lecturer salary and lecturer years since Phd.

If we consider the correlation between lecturer salary and lecturer rank the correlation coefficient value is -0.6172243. Which means there is quite a significant negative linear relationship between the lecturer salary and lecturer rank.

Furthermore the respective scatter plot diagrams of lecturer salary - years of service, lecturer salary - lecturer years since Phd and lecturer salary - lecturer rank above clearly illustrates the point.

As none of the above correlation coefficients are equal to zero, we can reject the null hypothesis of each of the scenarios in favour of the alternative hypothesis.

# Using the finding of the research conducted, develop a well explained statistical model to predict precisely the salary required to be offered based on lecturer rank, lecturer years of service and lecturer years since Phd using statistical regression analysis followed by explanation of the resulting model.

## **Statistic regression analysis**

Statistic regression models are used to find relationship between two variables. Whenever there is a cluster of data and there is a possibility of either showing a positive linear relationship or negative linear relationship by fitting a regression line the null hypothesis is naturally rejected because the null hypothesis states that the slope is equal to zero, when plotting regression line the vertical distance between the line and the scattered points is the error, the line that best minimizes the squared distances between the points and the slope should be found. Y= mx + b is the equation that’s used eventually to draw the slope after the constants m and b are found.

This way the best fitting regression line could be plot on a graph to show the relationship between two variables with the least distance between the regression line and points possible.

### **Statistical regression analysis for salary and lecturer years since Phd**

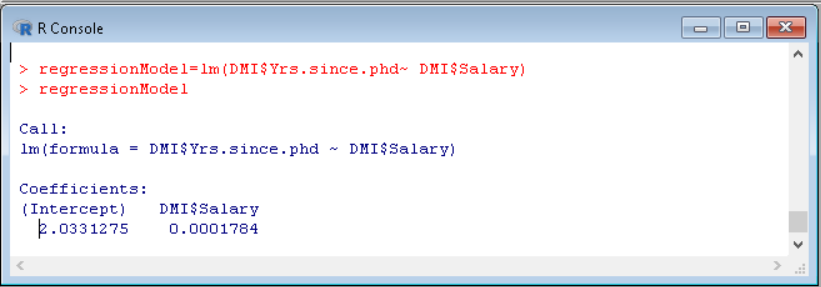


Figure Regression

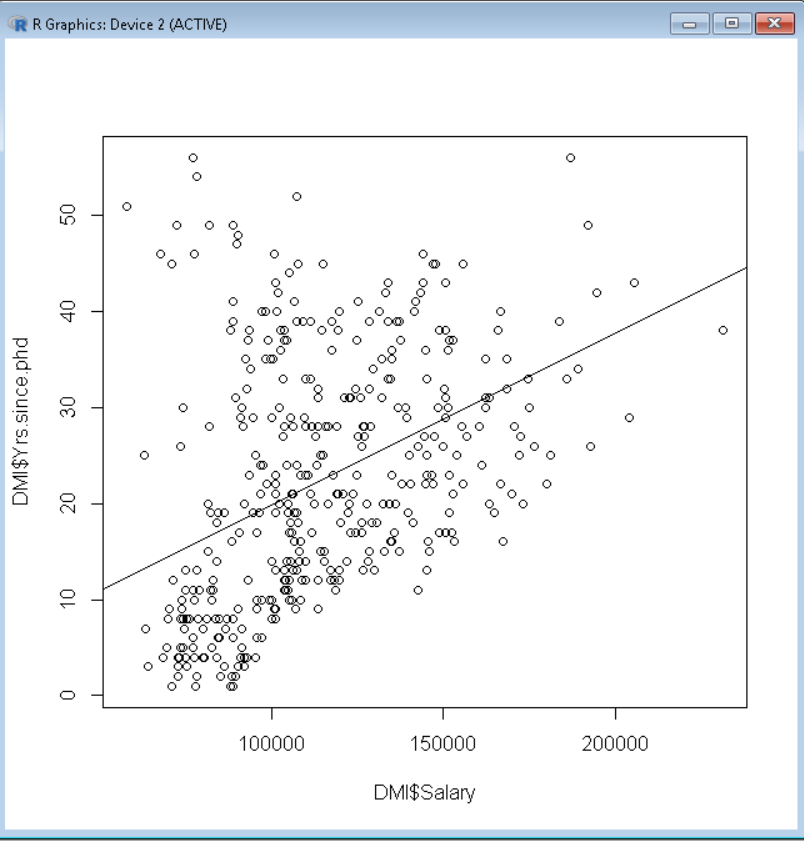


Figure Scatter plot with regression line

Statistical regression analysis for salary and lecturer years since Phd

* Intercept : 2.0331275
* Slope.       : 0.0001784

If we consider the lecturer salary vs the lecturer years since PhD. It is fairly obvious that there is a positive linear relationship, considering that the Slope = 0.0001784. Furthermore the regression line makes it fairly obvious.

### **Statistical regression analysis for salary and lecturer years of service**

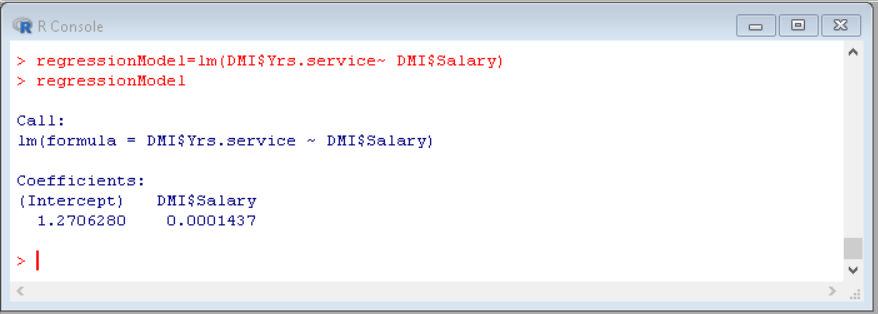


Figure Regression model

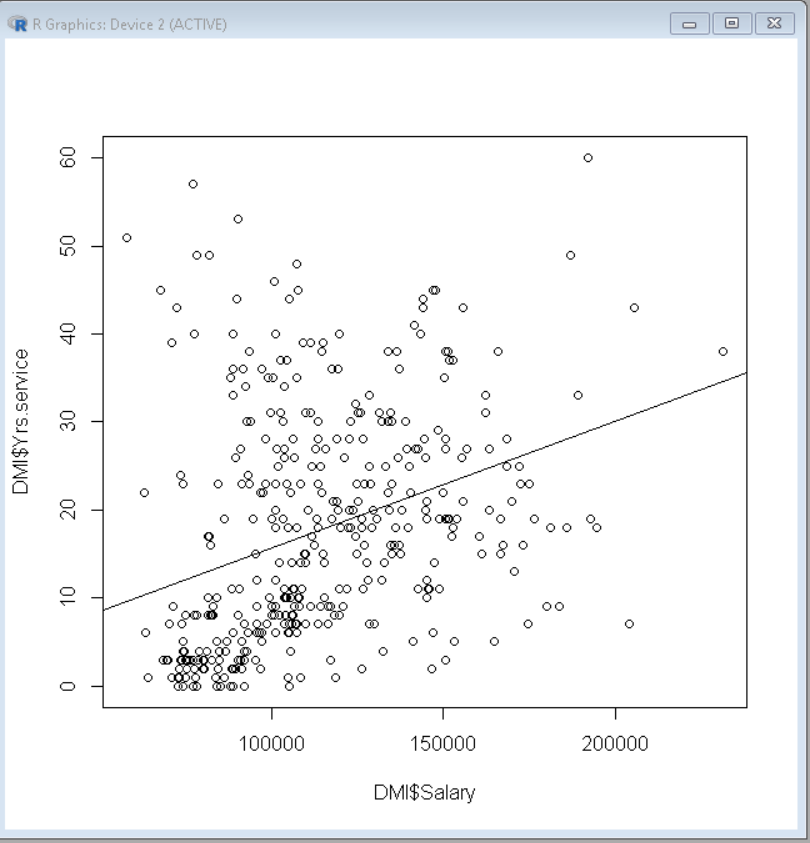


Figure Scatter plot with regression line

Statistical regression analysis for salary and lecturer years of service

* Intercept : 1.2706280
* Slope.       : 0.0001437

If we consider the lecturer salary vs the lecturer years of service. It is fairly obvious that there is a positive linear relationship, considering that the Slope = 0.0001437. Furthermore the regression line makes it fairly obvious.

### **Statistical regression analysis for salary and lecturer rank**

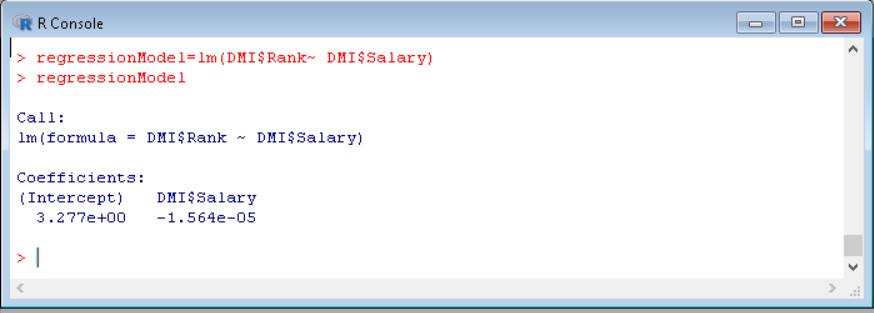


Figure Regression model

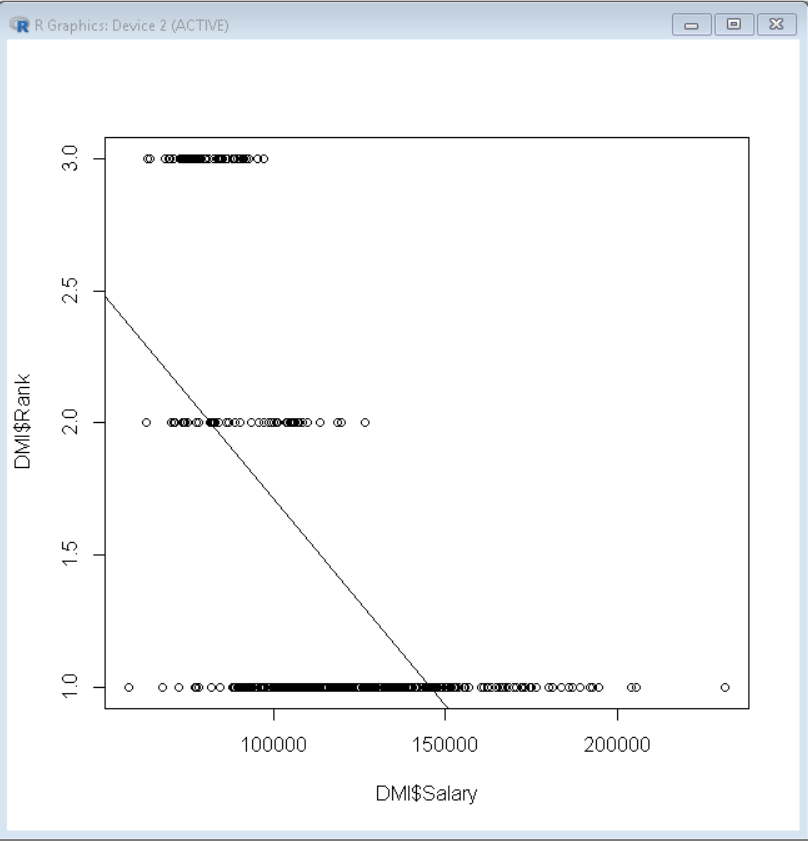


Figure Scatter plot with regression line

Statistical regression analysis for salary and lecturer rank

* Intercept : 3.277e+00
* Slope.       : -1.564e-05

If we consider the lecturer salary vs the lecturer rank. It is fairly obvious that there is a negative linear relationship, considering that the

Slope = -1.564e-05 Furthermore the regression line makes it fairly obvious.

### **The salary required to be offered based on lecturer rank, lecturer years of service and lecturer years since Phd**

The mean of means should give us the reasonable salary to be given to the lecturer based on lecturer rank, lecturer years of service and lecturer years since Phd.

# Provide final conclusion(s) of the research followed by recommendations based on the findings of data analysis

## Conclusion

The potential of business intelligence is massive, especially in a university such as this. The amount of data that is in circulation at present is as high as it has ever been, the number of benefits the university could capitalize on due to the availability of data is quite high. Some of the most important advantages the university should capitalize on are;

* Creating time table for semesters
* To work out the number of lecturers, tutorials and labs to schedule per module.
* The rooms and equipment required
* Admissions and application data
* Financial data
* Alumni and historical data
* Etc.

Even though the availability of data has the potential of benefitting the university massively it could well damage the integrity or cause other kinds of problem as well. So when it comes to dealing with data it is always recommended to use the best practices such as;

* Ensuring the accuracy of data
* Ensuring the security of the data
* Ensuring processes such as authentication and authorization are used to protect data.
* Etc.

The fact that the data sets lecturer salary, lecturer years of service and lecturer years since PhD aren’t normally distributed was proved by rejecting the null hypothesis of different tests such as;

* Shapiro-Wilk Test
* Anderson-Darling Test
* Lilliefors-Tests

Data is often analysed in-order to make predictions, generate reports, observe a current situation in a simple understandable way by representing them in a graph and etc.

The mean of the different groups of the population will be equal is the null hypothesis of ANOVA. But the mean of the different groups calculated aren’t equal to each other, so the null hypothesis was rejected in favour of the alternative hypothesis. Which means that there is a definite variation in salary with respect to the position. Calculation of these values could well be a daunting task but thanks to the available software calculations have become very easy and accurate.

Whenever a linear line can be fit into a graph with scattered points we assume that there is a linear relationship between the two variables. Range of the coefficient value of the correlation.

(-1 – 0 – +1)

Whenever the value of coefficient is equal to 0, there is no correlation between the two variables. If the coefficient is equal to -1 or has a negative value then there is a negative linear relationship and if the coefficient is equal to +1 or has a positive value then there is a positive linear relationship. Below are some correlation values that we’ve found.

* Correlation between lecturer salary – lecturer years of service
  + 0.3347447
* Lecturer salary- lecturer years since Phd
  + 0.4192311
* Lecturer salary-lecturer rank
  + -0.6172243

If we consider the above values we’ve derived none of them are equal to zero so we could reject the null hypothesis of the correlation in favour of the alternative hypothesis. Both correlation between lecturer salary – lecturer years of service and lecturer salary – lecturer years since Phd have a positive linear relationship but the relationship between lecturer salaries – lecturer rank has a negative linear relationship.

Regression analysis done to find the relationship between lecturer salary – lecturer years of service, lecturer salary – lecturer years since Phd and lecturer salaries – lecturer rank produce the following results;

* lecturer salary – lecturer years of service
  + Intercept =1.2706280, Slope = 0.0001437
* lecturer salary – lecturer years since Phd
  + Intercept =2.0331275, Slope = 0.0001784
* lecturer salaries – lecturer rank
  + Intercept =3.277e+00, Slope = -1.564e-05

Above findings emphasises further on the finding using correlations, here the slope value been means there is a positive linear relationship if the slope value is negative there is a negative linear relationship between the variables.

Thanks to the software available we have being able to maximize the use of the huge amount of data available. This trend is expected to stay and this could well be what determines failure and success in future.

