Matriculation number:	2	6	1	2	0

Examination Assignment

Module: Data Analysis and Statistics

Exam part: Data Analysis and Statistics

Examiner: Prof. Dr. Schwind, Dipl.-Biol. Ralf Darius

Deadline for the submission: 31.08.2019, 11:59 pm

Study program	Begin of studies	Last name, First name
Information Engineering and Computer Science (M.Sc.)	Winter Semester 2018/19	Murali, Sriram

Assessment criteria and number of points that can be achieved:

Maximum number of points	Skills and Expertise	Systematic and scientific Quality	Quality of the results	Presentation of the results
100	45	15	30	10

Result:

Points	Mark	Skills and Expertise	Systematic and scientific Quality	Quality of the results	Presentation of the results

Statement of Authorship:

This report is the result of my own work. Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.

Table of Contents

1. Introduction

1.1 Background

“Census data influences decisions made from Main Street to Wall Street, in Congress and with the Federal Reserve. Not to mention, the American people who look to, and trust, the data the government releases on our nation's unemployment, state of our economy, and health insurance coverage” (Farenthold). Census plays a very vital role in identifying where a country or region is growing towards while also identifying the areas that needs to be improved. A census is taken every 10 years to mainly calculate the population of a country, in this case USA.

The data that will be analyzed in this report is an income survey data of male employees in the Central Atlantic Region of USA. The data analyzed in this report is part of the ISLR package that contains 3000 data with 11 attributes. The aim of this report is to analyze the data with respect to 4 attributes – *year, age, wage and education* and find out the correlation between all the attributes.

1.2 Motivation

Wages of an employee changes based on various factors such as academic qualification, the region of work, industry, market situation and the years of experience. The data provided will give a better understanding about which of these factors has more impact on the wages of an employee. These factors can be decided by doing an extensive analysis on the numerical and continuous variables given in the data.

With the introduction of R programming language, it has become an easy task to achieve solutions related to understanding available data and then analyzing various related aspects. R programming is highly popular as it is open source and easily accessible. This is a very obvious choice when a thorough understanding of data using various graphs and statistical methods is needed. The available packages in R can be used to perform specific functions like partitioning the data, classifying and combining data, finding hidden layers in the data, etc. Also, R has various functions which can handle many probability functions required to accomplish our task.

1.3 Research Question

The research questions are as follows:

- Is there a relationship between age/education/calendar year and wage? Our first goal should be to determine whether the data provide evidence of an association between age/education/calendar year and wage.

- How strong are the relationships? Assuming that there are relationships, we would like to know the strength of those.

- Given a certain age/education/calendar year, can we predict wage with a high level of accuracy? This would be a strong relationship. Or is a prediction of wage based on age/education/calendar year only slightly better than a random guess? This would be a weak relationship

- Which factors contribute to wage? Do all three factors — age, education and calendar year — contribute to wage, or do just one or two of the factors contribute?

- Is the relationship linear? If there is approximately a straight-line relationship between wage and age/education/calendar year, then linear regression is an appropriate tool. How well does the linear model fit the data? If the relationship is not linear, what could be considered?

- Are there interaction effects?

2. State of Art

2.1 Study System

The study system involves the income survey data of employees in the central Atlantic region of the United States. In this research, the wages of male employees in the Atlantic region is analyzed to construct a model based on factors like age, education and year. The dataset that is being analyzed consists of 3000 data on which the analysis will be carried out on 4 main attributes – age, wage, education and year. The data is collected by Census Bureau using a probability selected sample of about 60,000 occupied households. The fieldwork is conducted during the calendar week that includes the 19th of the month. The questions refer to activities during the prior week; that is, the week that includes the 12th of the month. Households from all 50 states and the District of Columbia are in the survey for 4 consecutive months, out for 8, and then return for another 4 months before leaving the sample permanently. This design ensures a high degree of continuity from one month to the next (as well as over the year). The 4-8-4 sampling scheme has the added benefit of allowing the constant replenishment of the sample without excessive burden to respondents.[10]

3. Tools and Methods

3.1 R Studio

RStudio is an open source IDE for R. It is a programming language written in C++ and uses QT Framework for the Graphical User Interface (GUI). It supports creating script files and also supports direct execution of code. RStudio supports major desktop platforms like Mac OS, Ubuntu and Windows and can also be used browser support for the IDE. (RStudio Team 2016).

3.2 Methods used

3.2.1 Exploratory Data Analysis

"Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations"[3]. Exploratory data analysis is a cyclic process where it is used to generate questions about the data being analyzed. The answers for these questions can be found out by visualizing, transforming and modeling the given data. Some of the terms defined in Exploratory Data Analysis are:

Variable – A quantity, property or a quality that can be measured.
Value – A state of a variable during the time of measurement that can change from one measurement to the other.
Observation – A set of measurements that are recorded under similar circumstances. An observation usually contains various values associated with multiple variables. An observation is also referred to as data points.
Tabular Data – It is a set of values associated with variable and observation. [4]

The following are the libraries used to analyze data through Exploratory Data Analysis:

Tidyverse:
Tidyverse is an R package which helps us to interact and perform various analysis on the given dataset. Tidyverse provides a wide range of operation on data such as transforming, visualizing and subsetting.

The tidyverse package consists of various sub packages which is extensively used for data visualization.

Ggplot2:

Ggplot2 is a sub package inside the tidyverse library which is used to visualize data in the form of graphs and charts. The data and variables are specified to the ggplot to map the graphical primitives like histogram and box plots.

To use the ggplot we will have to give the arguments as the given dataset followed by the aesthetic mapping which is usually specified as aes() and with the layers of graphical representation for your visualization such as geom_point() or geom_histogram().

Boxplot:

Boxplot is used as a function in order to show the distribution of data in the dataset. It usually divides the data into three quartiles where there is a minimum, maximum, median, first quartile and third quartile.
The general syntax for creating a boxplot is

Scatterplot:

Scatterplots are the graphical points used to denote mainly numeric values. It shows the points plotted in the Cartesian plane. Each point in the Cartesian plane represents the two variables.[3] The scatterplots are created using the plot() function. The syntax for creating the scatterplot is

Histogram:

It gives the graphical representation of variables having a continuous range. Each bar achieved from the histogram represents the peak values belonging to that range.
Histogram can be created by using the hist() function in R or by using geom_histrogram() which is used along with gglpot(). The basic syntax for creating a histogram with hist() is as follows:

FunModeling:

FunModeling is an R package which is also used for exploratory data analysis and data modeling. It is extensively used for finding correlation between data, frequency of data points and finding out unique values in the data set. It helps us to identify the null values, the close correlation between different attributes and unique values in columns. The following are some of the functions used in funModeling:

- ***freq()*** – Used to find the frequency of the categorical data using a horizontal bar graph.

- ***df_status()*** – Analyze zeros, missing values, infinity and number of unique values in data set.
- ***plot_num()*** - This function takes a dataset and plots the values of numerical variables. The non-numerical variables are not plot using this method.
- ***correlation_table()*** – The function is used to retrieve the Pearson coefficient for every numerical variable. This variable indicates which attributes are closely correlated with each other.

3.2.2 Generalized Additive Model (GAM):

The Generalized Additive Model is used to fit the non-linear functions of a dataset to every predictor in the dataset so that the non-linear relationship will be automatically modeled [8]. GAMs provide a middle ground between linear regression and neural networks used in machine learning models. Linear regression models are sometimes straight forward and the relationship between various attributes can be understood easily. At the same time, there are models which are complex which uses machine learning algorithms to find out the relationships. One set back of using these machine learning algorithms is that it needs huge amount of data to interpret and make inferences from the result. GAMs are extensively used to model non-linear relationships between two attributes. The linear model functions such as lm() doesn't perform a good analysis on the non-linear relationships. The gam() function provides a better fit to these non-linear relationship attributes and provides a better smooth and splines which takes on wide variety of shapes. It is used to analyze and find relationships between non-linear attributes of the model [7].

4. Implementation & Analysis

The implementation is done using R programming. The original data frame provided consists of 11 attributes out of which 4 attributes are filtered and analyzed. The attributes are year, age, education and wage of 3000 male employees in mid-atlantic region in USA. The major aim of this section is to analyze and answer the research questions using exploratory data analysis and generalized additive model (GAM). The exploratory data analysis, as explained earlier in the tools and methods section, makes use of boxplot, scatterplot and histogram to visualize data models. The GAM on the other hand is used to model non-linear models and find relationship between two attributes.

The required attributes for analysis are year, age, education and wage. In these 4 attributes year, age and wage are numerical variables and education is the categorical variable. The

distributions of numerical variables are displayed in the following histogram.

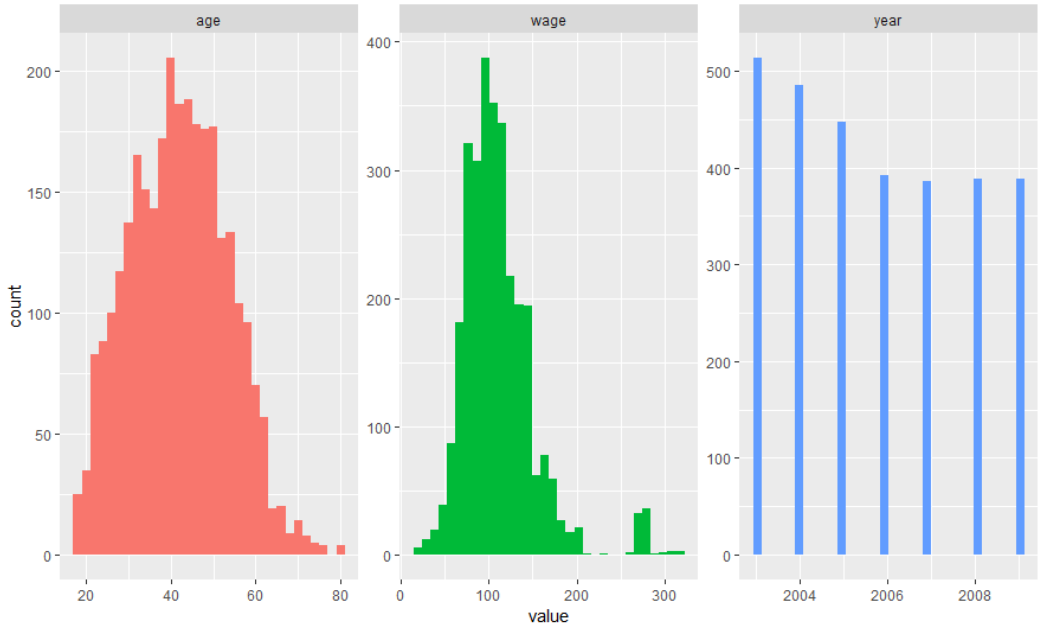

Figure:1 Distribution of Numerical variables

From Figure 1, the distribution of the data is represented using a histogram. The numerical variable is represented in x-axis and the count of the numerical variable is represented as y-axis. It can be interpreted from the chart that the age value is mostly distributed between 30 and 50 olds. The data set count seems to taper outside the age of 60. Most of the wages in the data seems to be distributed around 100k per annum and a very small amount of data distributed above 250k per annum. For the year data, there seems to be more employee data recorded from 2003-2005 which seems to decrease from 2006.

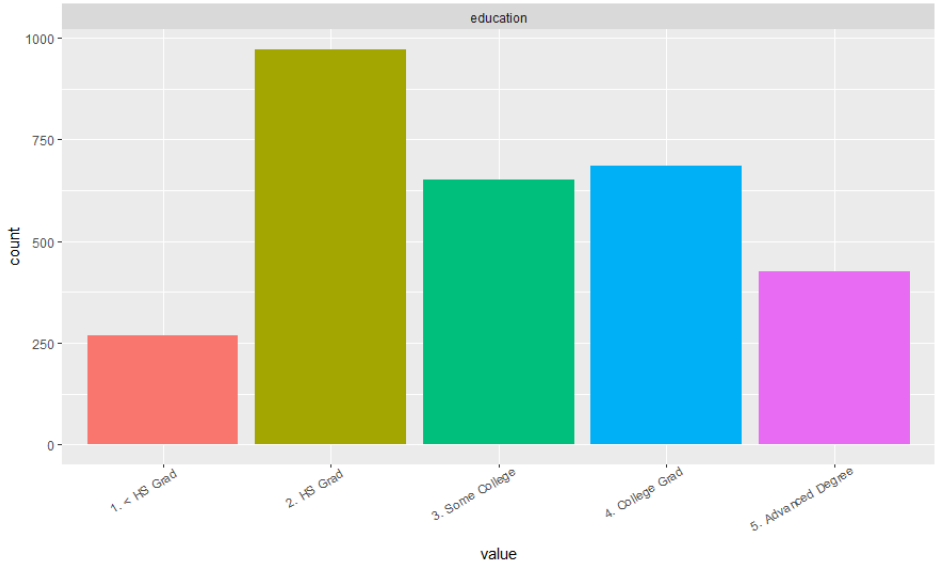

Figure:2 Distribution of Categorical variable

From Figure 2, the categorical variable Education has been distributed between 5

categorical education qualifications. Less than 10 % of the people have less than a higher secondary education qualification. Around one-third of the data distribution has a higher secondary education qualification. The rest of the data is distributed between some college degree, College graduate and an Advanced Degree.

4.1 Case 1 - Relationship between age/education/year and Wage

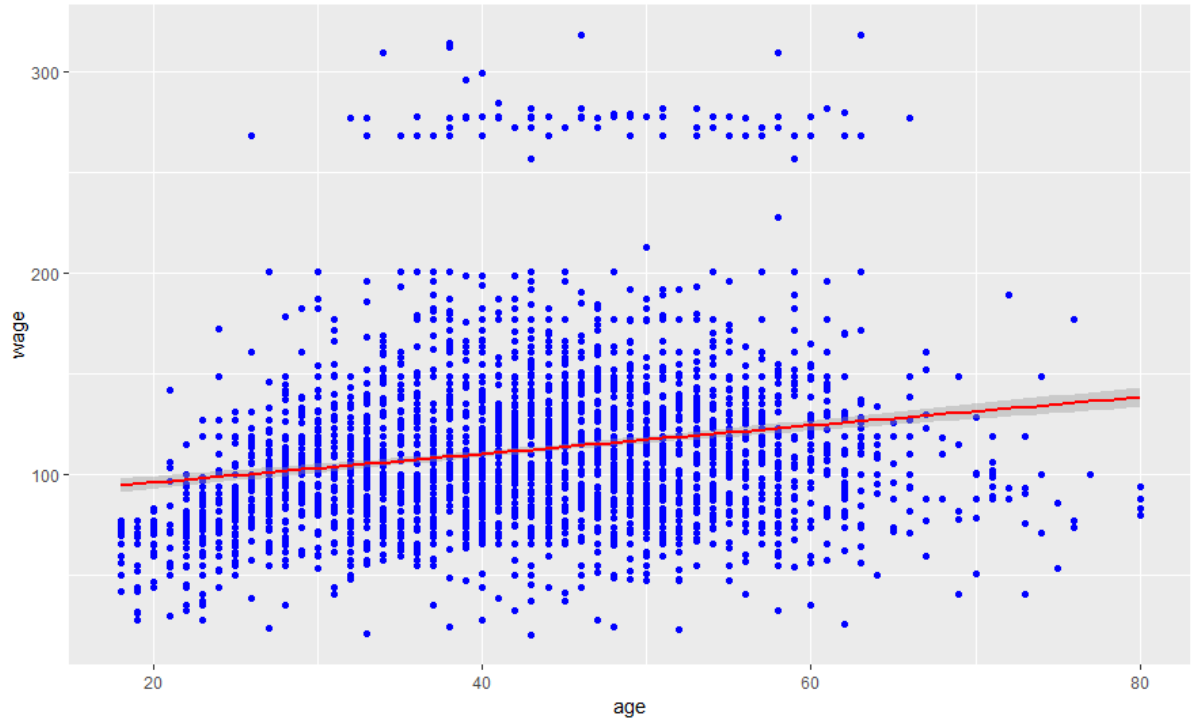

Figure 3: Scatterplot of age vs wage

The first research question attempts to find out the relationship between the age/year/education and wage. In order to find this out a linear model approach is followed and scatterplot is used to find out the relationship between various attributes with wage. In Figure 3, a scatterplot is plotted between age and wage. It is visible that most of the wage data are distributed between 30 and 60 years of age. The range of wage lies mostly between 50k and 100k for the above mentioned age range whereas only small amount of data lies between 250k and 300k. The linear regression line suggests that the wage increases with increase in age which doesn't suggest much about the linearity of the relationship.

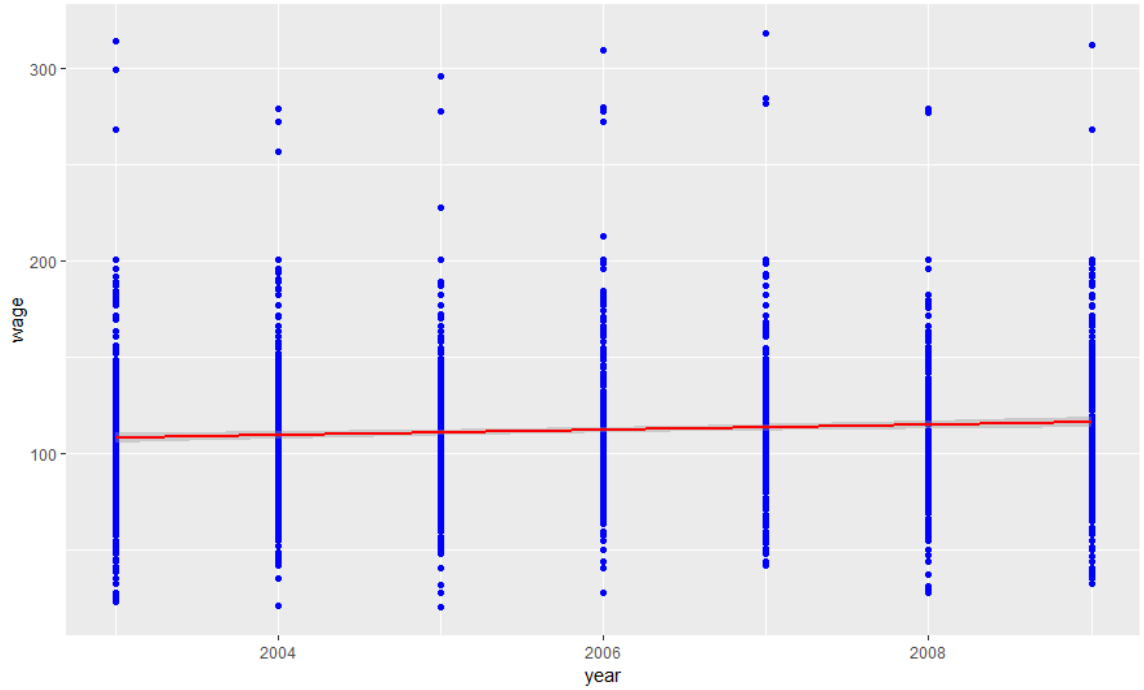

Figure 4: Scatterplot of year Vs wage

The figure 4 attempts to find a direct relationship between year and wage. The data consists of year ranging from 2003-2009 in which the wage data is equally distributed throughout the years. The linear regression line seems to suggest that the year has no or very less impact on the wage of an individual. So the year and wage seems to have no relationship.

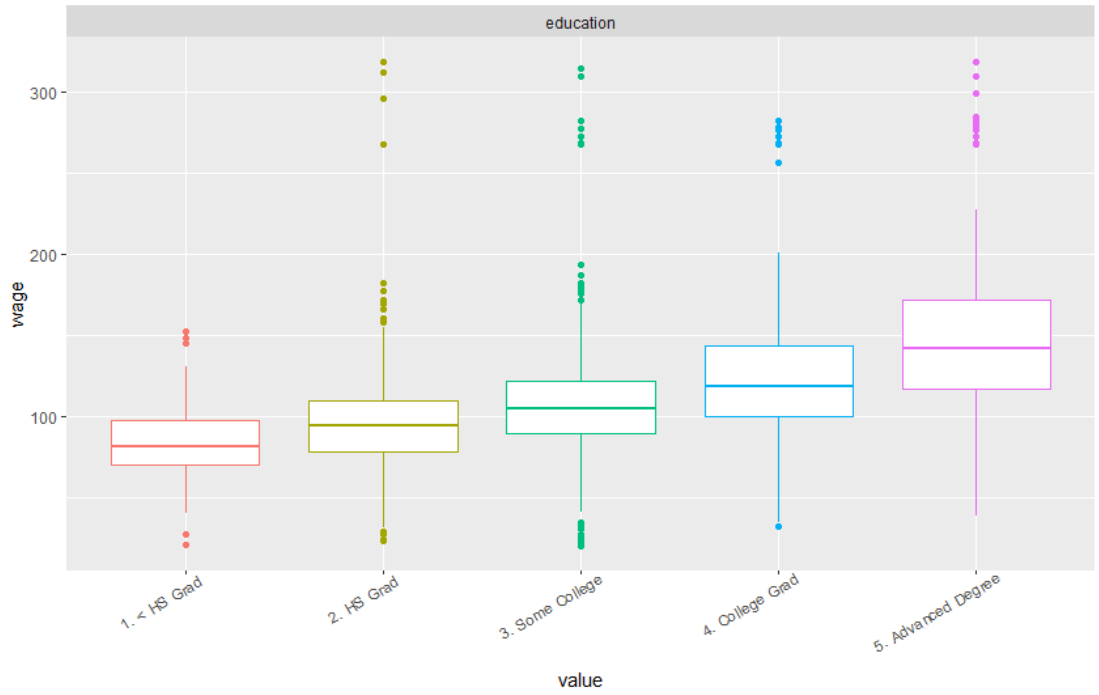

Figure 5: Boxplot of education Vs wage

The figure 5 tries to identify the relationship between education and wage using a box plot.

The box plot is used in the above case to identify whether education has a linear relationship with wage. The mean wage for an employee who has less than HS grad seems to be well below 100K per annum. The mean wage seems to increase subsequently for the HS Graduate and other categorical values. From the above plotting it is evident that the mean value of wage seems to increase with increase in the educational qualification and has a linear relationship.

4.2 Case 2 – Strength of relationship

The strength of relationship between any two attributes is usually found out by correlation factor. The correlation between any two attributes varies between 0 and 1, where 0 suggests weak correlation and 1 suggests strong correlation.

In R programming, pair plotting is used to find out the correlation of two or more attributes in a dataset. But pair plotting will only be able to find correlation between two numeric variables. In the dataset provided, education is a categorical variable rather than a numerical variable. In order to find the correlation between education and wage, education attribute will be converted into a numeric variable.

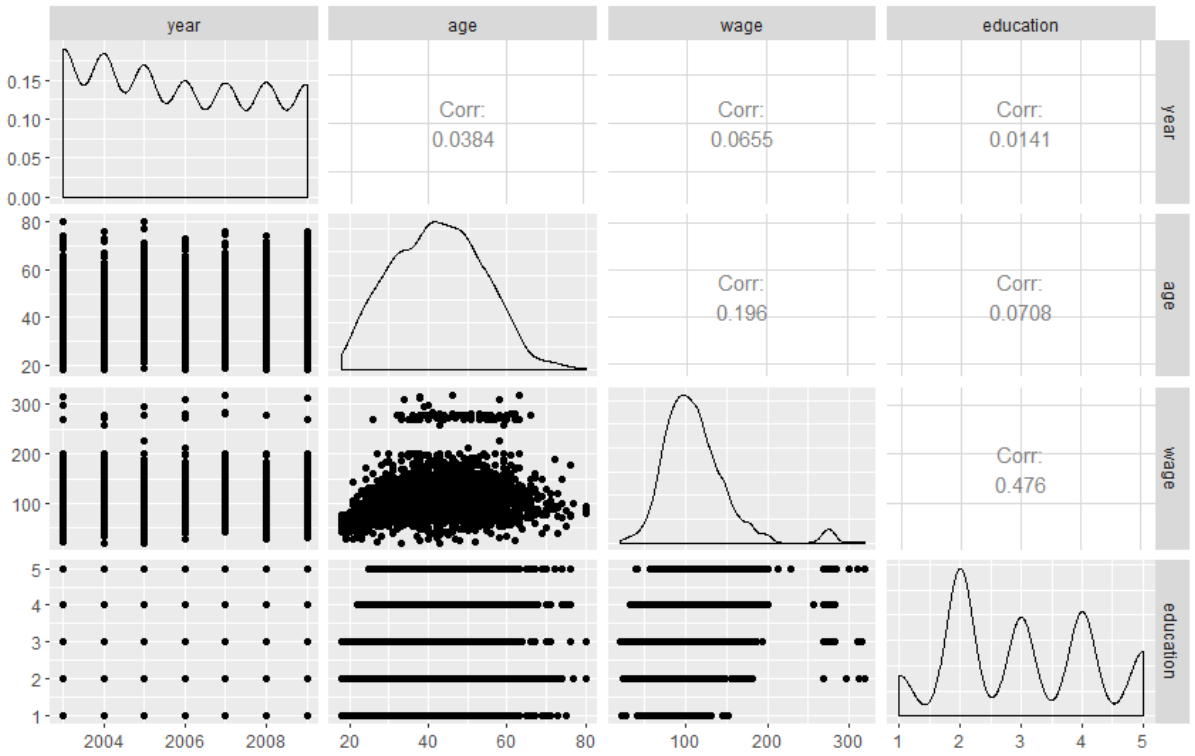

Figure 6: Pair Plot

In Figure 6, a pair plotting has been done between all the attributes in the dataset. Although all the attributes have a corresponding correlation value, the main aim is to find the correlation between wage and the other 3 attributes. The pair plot shows that the correlation coefficient (R) is higher between wage and education at 0.476. This shows a positive correlation between wage and education. Age and wage has the next highest correlation coefficient of 0.196 whereas year and wage has the lowest coefficient of 0.0655.

In order to further investigate the strength of relationship, Pearson correlation formula is used which calculates the correlation coefficient and the significance level (p-value). The Pearson coefficient formula is:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

x,y – vector of length n
m_x – means of variable x.
m_y – means of variable y.

The p-value is determined from the above function by finding the distribution of above data set. The p-value of less than 5% implies that the correlation between two attributes is significant [9]. If the value is more or less equal to 0, the correlation is more significant. The below table mentions the correlation coefficient and the significance value of wage attribute with the other attributes.

Attributes	***Correlation Coefficient(R)***	***Significance Value(p)***
Education	0.48	$2.2e^{-16}$
Age	0.2	$2.2e^{-16}$
Year	0.066	0.00033

Table 1: Pearson coefficient and significance value

From the above table it is evident that the age and education has a very strong relationship with wage and the year attribute has weak relationship with wage.

4.3 Case 3 – Accuracy of prediction

In this section, the level of prediction accuracy is calculated for the variables age, year and education. Generalized Additive Model (GAM) is used to fit the non-linear functions to the predictor which helps to model non-linear relationship with attributes which does not fit into the linear regression. With respect to the data set, it is evident that the education can be predicted using a linear regression. But the other two parameters, age and year seems to

have a non-linear relationships which can fit into one GAM i.e. smooth spline of year and non-linear functions of age. For these non-linear relationships polynomial regression needs to be used. In order to use polynomial regression, the degree of polynomial to use needs to be decided. ANOVA (analysis of variance) is used in this case to test the null hypothesis of a model M with a variance of 1 which is sufficient to explain the data. The variance will be analyzed with 5 models in which the year is represented as a cubic function and age will be represented with an increase in polynomial degree for every model starting from 1 to 5. With this model, the sum of square error and p-value are calculated to find which polynomial degree fits best to predict the non-linear relationship.[8]

Degree of Models	***Sum of Square Error***	***P-value***
1	-	-
2	144972	0.0000000000000002
3	6106	0.026
4	0	0.998
5	3027	0.118

Table 2: SSE and P-value for degree of polynomials

From the above table, it is evident that the p-value is less and the sum of square error is more for the model of polynomial degree 2. The sum of square error or the test squared error is significant as a metric for comparison of non-linear models. Hence, with the annova result set we can conclude that the Model 2 with polynomial degree 2 is preferred over other models.

As an alternate to determining the polynomial degree using anova, cross-validation (CV) method is used. For this purpose, the data set is divided into equal halves as training and test data set and test squared error for each model is calculated. The data is split into two halves of 1500 data each in training and testing model and finally the test squared error for both the models are compared.

The best degree to use for the polynomial degree of age is calculated using a cross validation method. By performing cross validation method we get mean squared error which is used to determine the better polynomial degree to use.

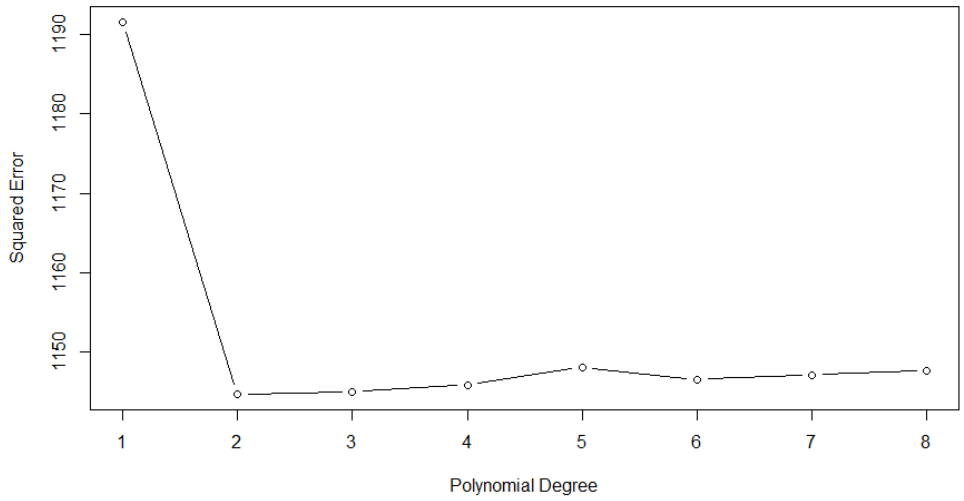

Figure 7: Polynomial Degree Vs Squared Error

In Figure 7, a graph is plotted with polynomial degree against the mean square error value. The graph clearly denotes that the mean square error is low for polynomial degree 2 when compared to all the other degrees. This brings us to the same conclusion as reached with the anova method that the polynomial degree 2 can be used to predict the non-linear relationship of age.

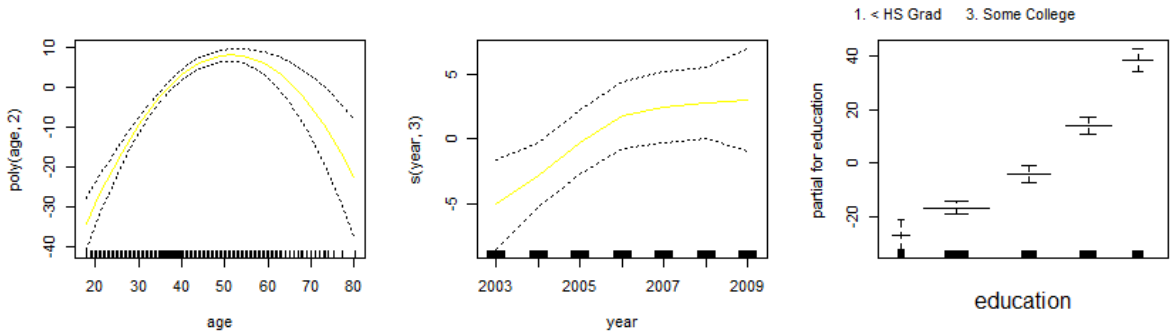

Figure 8: Prediction line for age,year and education

Figure 8 displays the prediction line for wage with respect to all the other attributes. The

plot for age indicates that keeping other predictors fixed, the wage seems to low for very young and old persons whereas the wage increases for middle aged persons around 40-60. Also the confidence interval is close to the actual line of prediction. Hence it can be said that age has a strong relationship with wage. The year plot indicates that the wage seems to increase until a certain year and then seems to increase less as the year moves on. The confidence interval for year is also not very close to the actual line of prediction. So the year has less effect on the prediction of wage and hence has a weak relationship. As discussed earlier in other cases, the education seems to have a big impact on the wage. The wage seems to increase with increase in educational qualification. Therefore the education has a strong relationship with wage.

4.4 Case 4 - Factors contributing to wage

The wage data which is numeric has been compared with various attributes such as age, year and education in the previous cases in order to find the strength, correlation and linearity of the model. In order to find out the factors influencing the wage, the significance value of each attribute needs to be calculated with respect to wage. The significance value (p) of 0 denotes that the attribute is more significant in factor contributing to the wage.

Attributes	***Significance Value(p)***
Education	$2.2e^{-16}$
Age	$2.2e^{-16}$
Year	0.00033

Table 3: p-value of attributes with wage

From the table above, the value is p is almost equal to 0 for the attributes education and age. This proves that the education and age attribute values have significant effect on the increase and decrease of wage. On the contrary, the year attribute has a considerably higher p-value. Hence it can be concluded from the above inference that the education and age play a major factor in deciding the wage.

4.5 Case 5 – Is Relationship Linear

Relationship between two attributes are said to be linear if the value of one attribute increases or decreases steadily. The relationship of wage with each attribute is examined below:

Education & Wage:

As seen in Case 1, the wage of the employee increases with increase in the education qualification. The boxplot also indicates that the quartile range increases with increase in education. In order to check whether a linear regression line can fit into a straight line relationship between education and wage, the education attribute is converted into a numeric and the relationship is tested.

Based on the results of boxplot, it is evident that there is a straight line relationship/linear relationship between education and wage. Hence the linear regression is an appropriate tool for education.

Age & Wage:

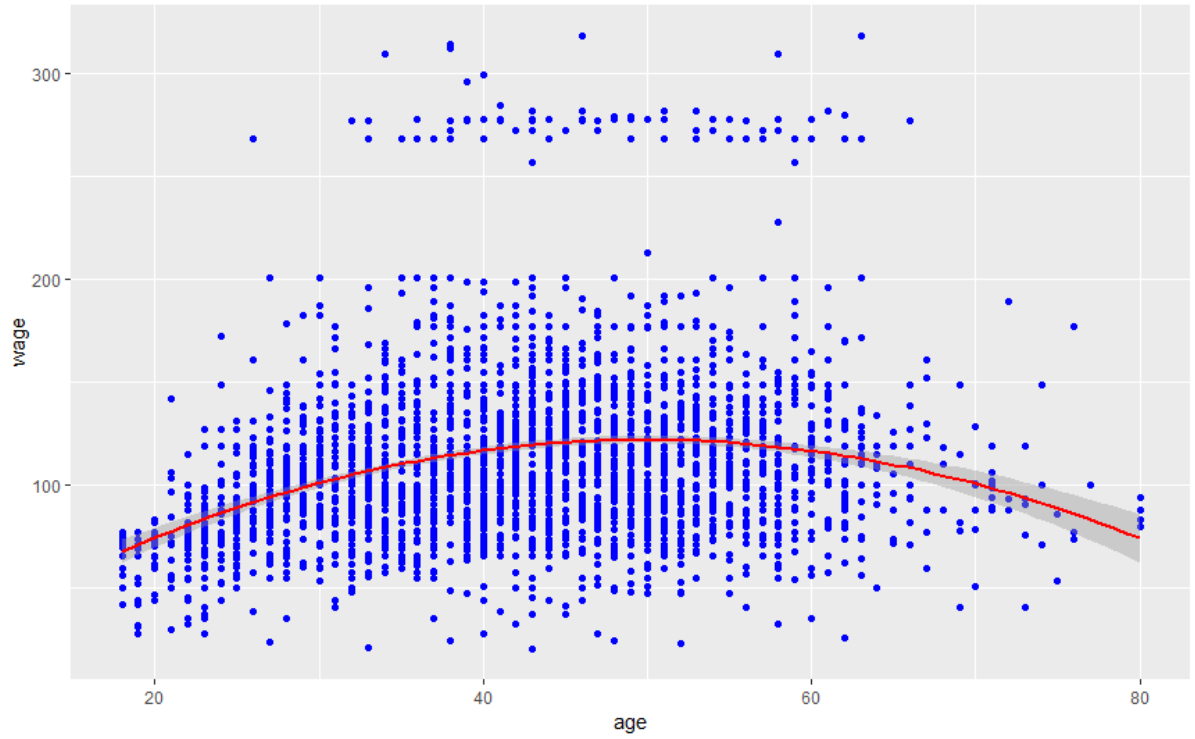

Figure 9: Polynomial Regression curve for age vs wage

In earlier cases, in order to define a relationship between age and wage**,** the relationship was tried to fit into a linear model. Since linear model was not able to fit and prove the significance between age and wage, a polynomial regression model was computed. In case 3, the polynomial degree was computed for age which was degree 2. The above graph shows that when using a polynomial regression of degree 2, the age tends to increase steadily and is in peak during a period of 40-60 years and seems to decrease. The relationship is not a linear one; hence linear regression cannot be used in this case. Instead,

polynomial regression should be used in order to fit the model.

Year & Wage:

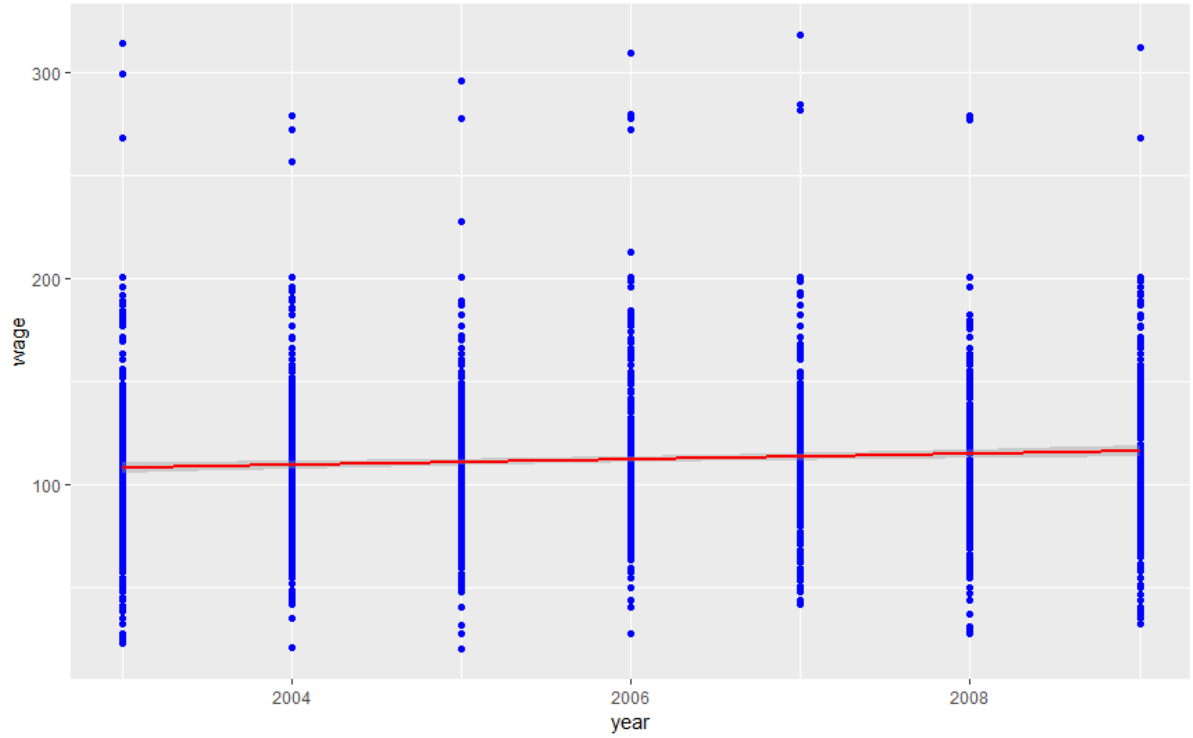

Figure 10: Scatterplot year vs wage

In Figure 10, a linear regression line was drawn for the year attribute. The line suggests that the linear regression cannot be used to fit year. The year value seems to have no or a very negligible effect on the change in wage. Trying to fit a polynomial regression line also seems to have the same result. Hence it can be concluded that there is a non-linear relationship between wage and year.

4.6 Case 6 – Interaction Effect

An interaction effect happens when the effect of one variable depends on the value of another variable. In standard linear regression model, an additive relationship is assumed between the predictor and response. An additive model is easy to interpret because the effect of each predictor on the response is unrelated to the values of the other predictors. However, the additive assumption may be unrealistic for certain data sets. So, we are going to check whether this data set contains interaction effect or not by removing additive model.

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

A third predictor will be added for the interaction effect called an interaction term which is computed by the product of X_1 and X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

The resulting wage data would look like the following:

$$\text{wage} = \beta_0 + (\beta_1 \text{ x AGE}) + (\beta_2 \text{ x EDUCATION}) + (\beta_3 \text{ x (AGE x EDUCATION)})$$

Factor	Estimate	P-Value
Age	0.56869	<2e-16
education2. HS Grad	11.43865	4.1e-06
education3. Some College	24.16700	< 2e-16
education4. College Grad	39.76677	< 2e-16
education5. Advanced Degree	64.98656	< 2e-16

Table 4: Coefficient estimate without interaction terms

Factor	Estimate	P-Value
Age	0.38316	0.02789
education2. HS Grad	7.24752	0.40429
education3. Some College	0.86720	0.92497
education4. College Grad	36.83562	9.36e-05
education5. Advanced Degree	61.13781	2.31e-08
age:education2. HS Grad	0.10113	0.61100
age:education3. Some College	0.56573	0.00789
age:education4. College Grad	0.07277	0.73490
age:education5. Advanced Degree	0.09876	0.68463

Table 5: Coefficient estimate with interaction terms

From the tables 4 and 5, it is evident that the models including the interaction term doesn't have a superior value. Here, the p-value for the interaction term (Age x Education) is not low, indicating that there is not strong evidence for β3 is not equal to 0. It is clear that relationship has no interaction effect.

5. Interpretation and Discussions

The practical task in this report aimed at studying the effect of attributes like age, education and year on the wage of an employee. The following interpretations are made based on the complete analysis:

Education & Wage: The Exploratory analysis done using box plot suggests that the relationship has been linear i.e. the wage increases with increase in educational qualification. Also, both these attributes have the highest positive correlation of 0.48 compared to all other interaction which proves that education has a strong relationship with wage.

Age & Wage: The relationship between age and wage has been analyzed using exploratory data analysis and generalized additive model (GAM). Using exploratory data analysis, a linear model was tried to fit in the relationship between both the attributes which seems to suggest that the relationship is not linear. Hence a polynomial regression of degree 2 has been used to find the relationship. The polynomial curve proves that the employees between 40 and 60 years of age seems to have a higher salaries that those in very young and very old ages.

Year & Wage: The relationship between year and wage was neither linear nor non-linear based on the graphs, correlation and interaction effects. The year attribute seems to have no effect on the wage of the employee with very less or no change in the employee salary based on the year.

References

[1] - https://www.brainyquote.com/quotes/blake_farenthold_766706?src=t_census
[2] - RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/. Accessed on 21st August,2019
[3] – What is Exploratory Data Analysis: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15.March 2018
[4] – What is exploratory data analysis: https://r4ds.had.co.nz/exploratory-data-analysis.html
[5] - https://www.analyticsvidhya.com/blog/2019/05/beginner-guide-tidyverse-most-powerful-collection-r-packages-data-science/
[6] – Funmodeling : https://blog.datascienceheroes.com/exploratory-data-analysis-data-preparation-with-funmodeling/
[7] – Generative Additive Models - https://noamross.github.io/gams-in-r-course/chapter1
[8] - Generative Additive Models -https://www.kaggle.com/mrshih/mid-atlantic-wages-and-gams/report
[9] – Correlation - http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
[10] –Current Population Survey - https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html