Matriculation Number:	2	6	5	9	0

HOCHSCHULE
RHEIN-WAAL
Rhine-Waal University
of Applied Sciences

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.)	Summer Semester, 2019	Al - Samy, S M Asif

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 Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.

S M Asif Al - Samy

Matriculation # 26590

Table of Contents

List of Figures

List of Tables

1. Introduction

This report analyzes a data set of income survey data for males in the central Atlantic Region of the USA by relating wage with other information (such as the year that wage information was recorded, age and education). The relationships between wage and the other information in the data is explored by the help of **R Studio** environment. The original data were manually assembled by Steve Miller, of Open BI, from the March 2011 Supplement to Current Population Survey data which has twelve different variables. As the tasks of this assignment was to analyze only four variables (year, age, education and year), analysis is started by extracting these four variables from the entire data set. Generally, the aim is to discover appropriate information, support decision making, and suggest conclusions by inspecting, cleaning, modeling, and transforming given data.

1.1 Research Questions

It is pertinent to consider some key questions before commencing the simulation process. The questions listed below are researched on accordingly.

a) Research on the variables:

   The first and foremost research is to identify the type of variables. There are two types of data: Categorical Data and Continuous Data. Depending on these types, different kinds of statistical methods are performed for analysis. Here year, age and wage are Numeric or Continuous data and education is Categorical data.

b) Research on the language and tools:

   - What programming language and which tool should be adopted to successfully simulate the task?

     R Studio is adopted for this purpose as it’s the best platform for analyzing this data set.

c) Research on the Methods:

- Which approach should be adopted in gathering desired results?

  For Gathering the results a list of libraries and methods are used for getting some diagram such as histogram, bar diagram, pairs plot etc.

- Which method(s)should be adopted in gathering desired statistics?

  Generalized Additive Models(GAMs) with some nonlinear fitting techniques are used to fit flexible models to the data. Generalized Additive Models provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. For polynomial regression, the degree of the polynomial to use is decided by using hypothesis tests. In this regards, Analysis Of Variance (ANOVA, using an F-test) is performed.

1.2 Motivation and Goals

There are variables (such as age, calendar year and educational attainment) of males in the central Atlantic Region that affects the wages of those males. In order to find out the relationships between wages and other variables this analysis has been performed because Our first goal should be to determine whether the data provide evidence of an association between age/education/calendar year and wage. This data set is described, visualized and operated statistically by taking the help of R studio for presenting the analysis properly. Moreover, a prediction is necessary for wage to make any important decision in future. In that case GAMs model is performed to make that proper prediction. Though this is hard to explore the dataset using proper libraries, methods and diagrams, it is favorable to learn those new methods and models.

2. Methodology

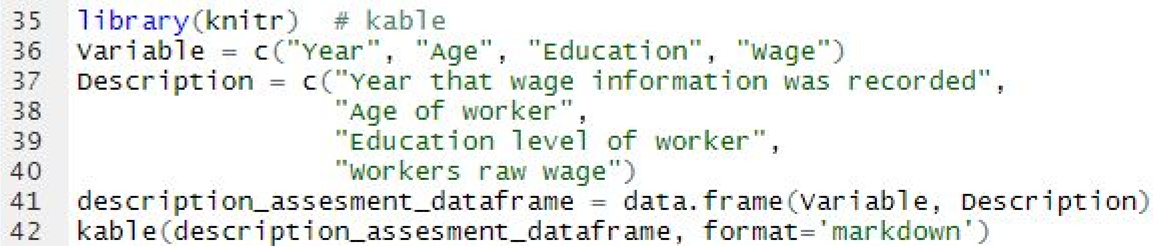
This wage data is analyzed at in exploratory data analysis. Therefore, a description of the data, visualization and statistical operations are performed using R Studio. In order to visualize the data set properly several bar plots, histograms and pairs plot is obtained using R libraries and methods. Then, generalized additive models (GAMs) is used to provide the framework for nonlinear fitting techniques. A non-linear technique is considered. This technique is used to help fit relationships between 'wage' and the other variables in the data set. Wage is the response variable. The non-linear technique has parameters that need to be tuned (ie polynomial degrees, number of cuts, degrees of freedom, etc.) The tuning parameter for each GAM is selected via cross validation. Details of how this is done is explained below.

2.1 Extraction of the Data Set

The data set is part of the ISLR package in R. It contains 3000 observations and eleven attributes or variables on workers' wages among other information. However, the task was asked to work on only three variables (year, age and education) with wage. As a result, an extraction of the columns year, age, education and wage is needed from the given data set. After running the given script a data.frame (data frame) which is named assessment_dataframe is created with my own sample of the original data. Then from that sample the required columns are selected and are assigned this changes into the same data frame. So my assesment_dataframe has now only four columns with that same sample. The following is a code snippet of extracting only four variables from the dataset.

```
assessment_dataframe <- assessment_dataframe[,c(1, 2, 5, 11)]
summary(assessment_dataframe)
View(assessment_dataframe)
```

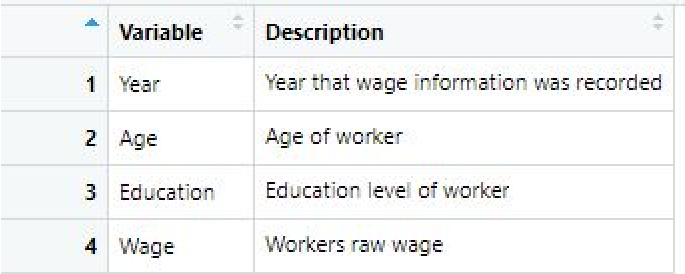
A code is also written for giving a short description of the variables as a table.



```
library(knitr)  # kable
Variable = c("Year", "Age", "Education", "Wage")
Description = c("Year that wage information was recorded",
                "Age of worker",
                "Education level of worker",
                "Workers raw wage")
description_assesment_dataframe = data.frame(Variable, Description)
kable(description_assesment_dataframe, format='markdown')
```

The output of this code is given below which represents a short description of the variables as a table.

Table 1: Variable Names and Descriptions of the Data Set

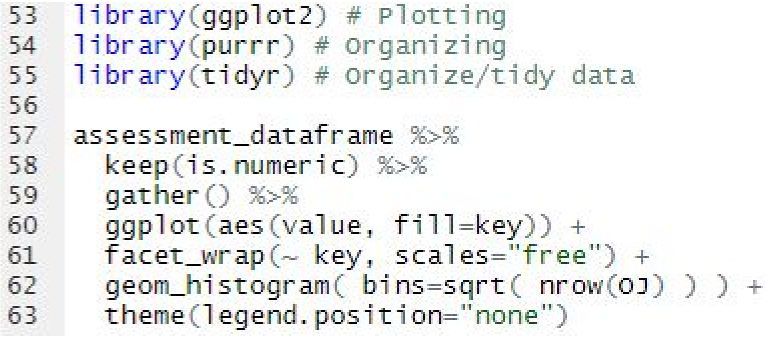


	Variable	Description
1	Year	Year that wage information was recorded
2	Age	Age of worker
3	Education	Education level of worker
4	Wage	Workers raw wage

2.2 Visualization

In order to visualize the dataset first task is to identify the types of variables. Depending on the types of variables exact libraries, methods, operations etc can be performed. In this data set year, age and wage variables have numeric or continuous type of data. On the other hand, education has categorical type of data.

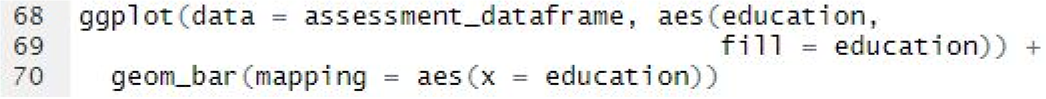
The following code snippet is for visualizing numeric or continuous variables in the data set.



```r
library(ggplot2) # Plotting
library(purrr) # Organizing
library(tidyr) # Organize/tidy data

assessment_dataframe %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value, fill=key)) +
  facet_wrap(~ key, scales="free") +
  geom_histogram( bins=sqrt( nrow(OJ) ) ) +
  theme(legend.position="none")
```

Here, some libraries are used to make a proper plot (in this case a histogram is drawn) and are also used to give some color, positioning and other features. This will give histogram of continuous data (from year, age and wage). Output of this code is described in the result section.

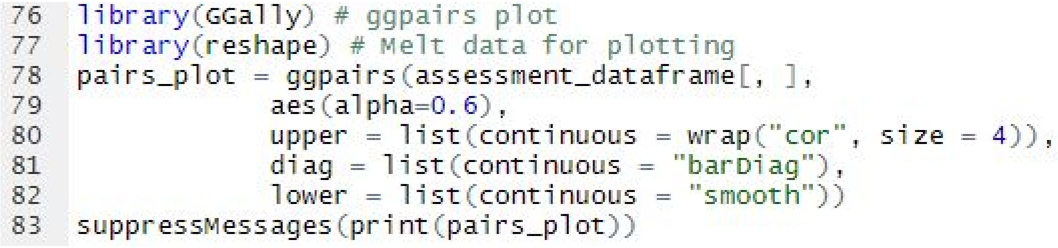
The following code snippet is for visualizing categorical variables in the dataset.



```r
ggplot(data = assessment_dataframe, aes(education,
                                        fill = education)) +
  geom_bar(mapping = aes(x = education))
```

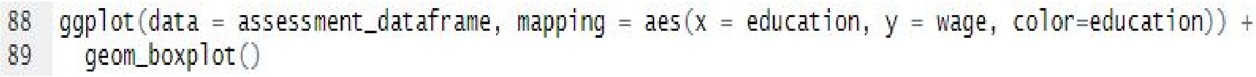
After running this code a bar plot of categorical data (from education) will appear. Output of this code is described in the result section.

The above code snippets are used to visualize each variable separately. In order to see how one variable is related to another the following code snippet has been used.



```r
library(GGally) # ggpairs plot
library(reshape) # Melt data for plotting
pairs_plot = ggpairs(assessment_dataframe[, ],
            aes(alpha=0.6),
            upper = list(continuous = wrap("cor", size = 4)),
            diag = list(continuous = "barDiag"),
            lower = list(continuous = "smooth"))
suppressMessages(print(pairs_plot))
```

After running this code a pair plot appears which shows the correlations with one another. However, wage is the response variable and its relationship to other variables is of interest in this report. The following code snippet is for visualizing the relationship with wage and other variables separately.



```
ggplot(data = assessment_dataframe, mapping = aes(x = education, y = wage, color=education)) +
  geom_boxplot()
```

Some box plots appear which show the relationships of wage and other variables separately for easier interpretation while running this code. Output of this code is described in the result section.

After visualization it is necessary to obtain some suitable statistics so that the dataset can be analyzed in a meaningful way.

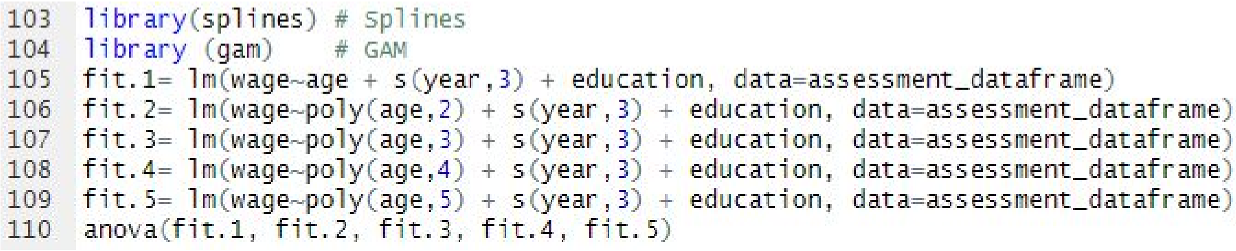
2.3 Generalized Additive Models (GAMs)

GAMs fit a non-linear function to each predictor, it can automatically model non-linear relationships that standard linear regression will miss. This means that it is not needed to manually try out many different transformations on each variable individually.

Here, one can see that two nonlinear relationships are fitted in one GAM (smooth spline of year and non-linear functions of age), but this section mainly focuses on non-linear functions for the age predictor. Many of these non-linear functions have tuning parameters to consider, as mentioned earlier. For example, step functions require certain number of intervals or cuts of the data.

For polynomial regression, it must be decided on the degree of the polynomial to use. One way to do this is by using hypothesis tests. Analysis Of Variance (Anova, using an F-test) is used in order to test the null hypothesis that a model M variance 1 is sufficient to explain the data against the alternative hypothesis that a more complex model M2 is required. In Anova, M1 and M2 must be nested models: the predictors in M1 must be a subset of the predictors in M2.

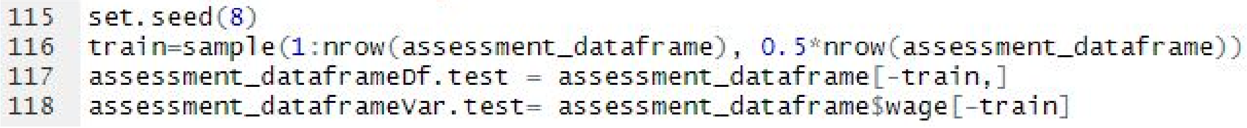
The following code snippet is used to fit five different models and sequentially compare the simpler model to the more complex model.



```
library(splines) # Splines
library (gam)    # GAM
fit.1= lm(wage~age + s(year,3) + education, data=assessment_dataframe)
fit.2= lm(wage~poly(age,2) + s(year,3) + education, data=assessment_dataframe)
fit.3= lm(wage~poly(age,3) + s(year,3) + education, data=assessment_dataframe)
fit.4= lm(wage~poly(age,4) + s(year,3) + education, data=assessment_dataframe)
fit.5= lm(wage~poly(age,5) + s(year,3) + education, data=assessment_dataframe)
anova(fit.1, fit.2, fit.3, fit.4, fit.5)
```

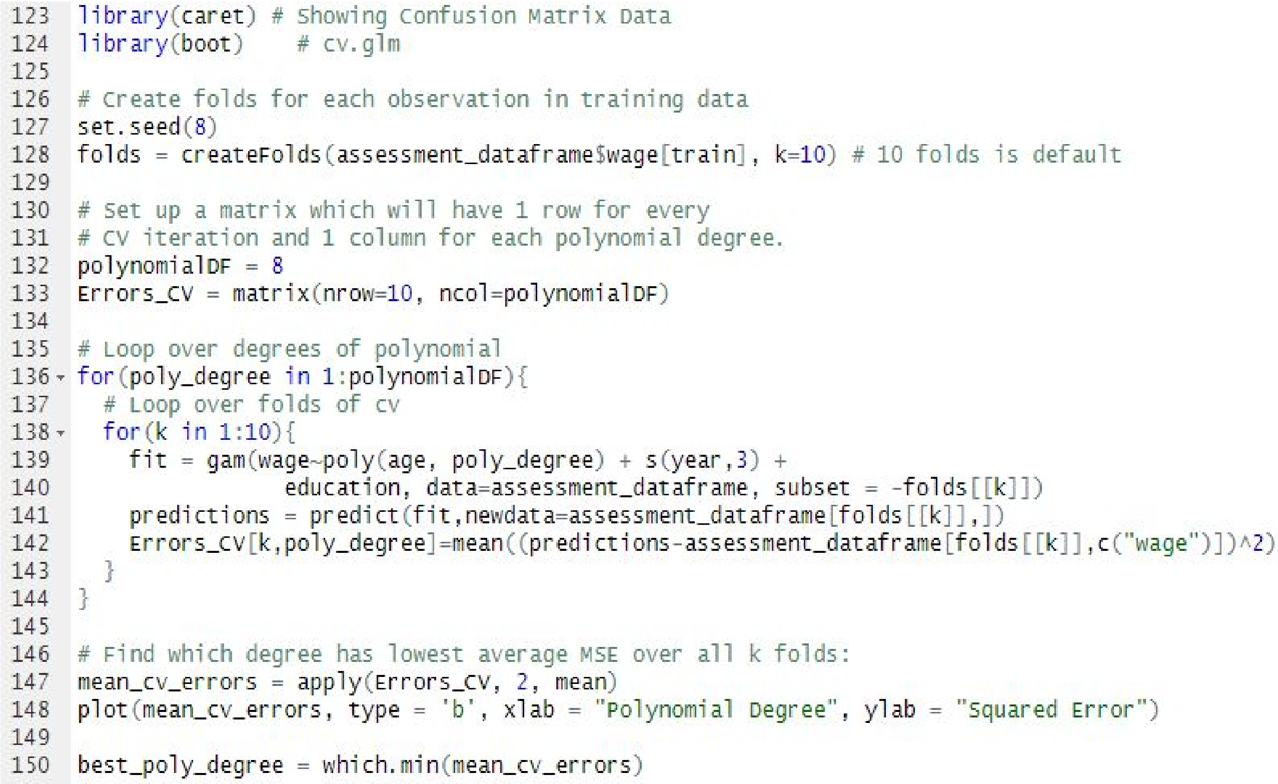
By running this code, an Anova table for polynomial age appears in the console. By analyzing that, a polynomial degree of age is determined (details with output is described in result section).

As an alternative to using hypothesis tests and ANOVA, the polynomial degree can be determined using cross-validation (CV). Before that, the dataset should be split in half. One half will be for training and tuning the model parameters, while the other half will be reserved for testing and getting the test squared error.



```
set.seed(8)
train=sample(1:nrow(assessment_dataframe), 0.5*nrow(assessment_dataframe))
assessment_dataframeDf.test = assessment_dataframe[-train,]
assessment_dataframeVar.test= assessment_dataframe$wage[-train]
```

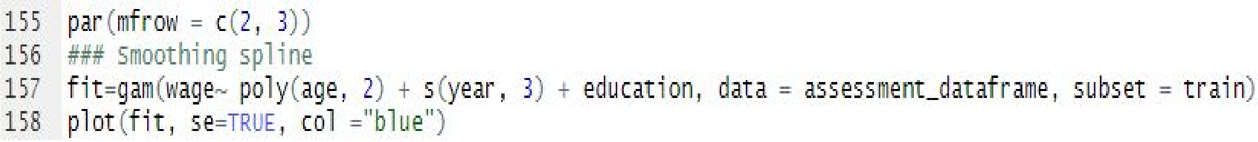
The dataset is split into two segments while running the above code snippet. After that it has been shown that how a tuning parameter is determined by cross-validation.



```
library(caret) # Showing Confusion Matrix Data
library(boot)    # cv.glm

# Create folds for each observation in training data
set.seed(8)
folds = createFolds(assessment_dataframe$wage[train], k=10) # 10 folds is default

# Set up a matrix which will have 1 row for every
# CV iteration and 1 column for each polynomial degree.
polynomialDF = 8
Errors_CV = matrix(nrow=10, ncol=polynomialDF)

# Loop over degrees of polynomial
for(poly_degree in 1:polynomialDF){
  # Loop over folds of cv
  for(k in 1:10){
    fit = gam(wage~poly(age, poly_degree) + s(year,3) +
                education, data=assessment_dataframe, subset = -folds[[k]])
    predictions = predict(fit,newdata=assessment_dataframe[folds[[k]],])
    Errors_CV[k,poly_degree]=mean((predictions-assessment_dataframe[folds[[k]],c("wage")])^2)
  }
}

# Find which degree has lowest average MSE over all k folds:
mean_cv_errors = apply(Errors_CV, 2, mean)
plot(mean_cv_errors, type = 'b', xlab = "Polynomial Degree", ylab = "Squared Error")

best_poly_degree = which.min(mean_cv_errors)
```

In this section some folds of training dataset has been created and a loop is performed over those folds in order to find the lowest average degree which is plotted in a graph (discussed in result section). Finally, the best polynomial degree is determined in this cross-validation which will be similar to the degree determined in Anova.

Finally, using the determined polynomial degree, a fixed GAM model is created to predict the relationships between wage and the other three variables.



```
par(mfrow = c(2, 3))
### Smoothing spline
fit=gam(wage~ poly(age, 2) + s(year, 3) + education, data = assessment_dataframe, subset = train)
plot(fit, se=TRUE, col ="blue")
```

The above code snippet will create some plots which defined the relationships between wage and other variables. Output of this final code snippet is also described in the result section.

This section has covered the approach of the work and the rationale behind it. The tools and methods used to do the tasks have been described in such a way that one can easily reproduce the task by reading this section.

3. Results of Analysis

After performing the tasks in R Studio, some plots are found by which the analysis has been done. An interpretation and the results of those analyses and the methods are discussed here with proper presentation of those plots in brief. Here, some predictions are also obtained by analyzing the results of Generalized Additive Models. Moreover, some errors or problems faced while performing the tasks and their effects on the results are also discussed here.

3.1 Visualization of Numeric Variables

In this section, results of the distribution of Numeric or Continuous variables are described separately via histograms.

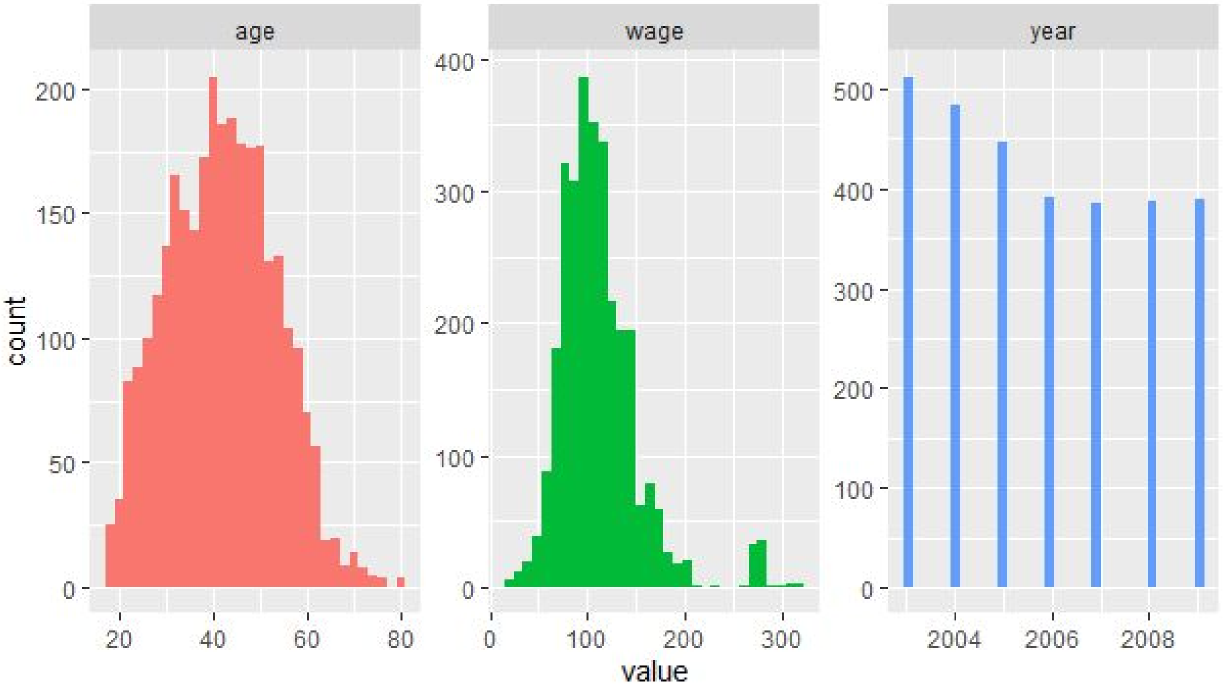


Figure 1: Visual Look at the Numeric/Continuous Variables in Data Set.

From Figure 1, it is noted that age seems to be mostly centered around the ages of 30-50 years, and then the counts begin to decline outside this range. It is interesting to note that this information set contains an employee near the age of 80. Moving on to wage, it appears that most wages are spread closer to 100 (K) with a narrower set of high wages around 250 and above. A larger number of workers recorded in the previous years (2003-2005) were observed for variable year. This amount reduces, however, until 2006, when the amount of employees recorded stays fairly constant at around 390.

3.2 Visualization of Categorical Variable

In this section, the result of distribution of categorical variable (education) is described with barplot.

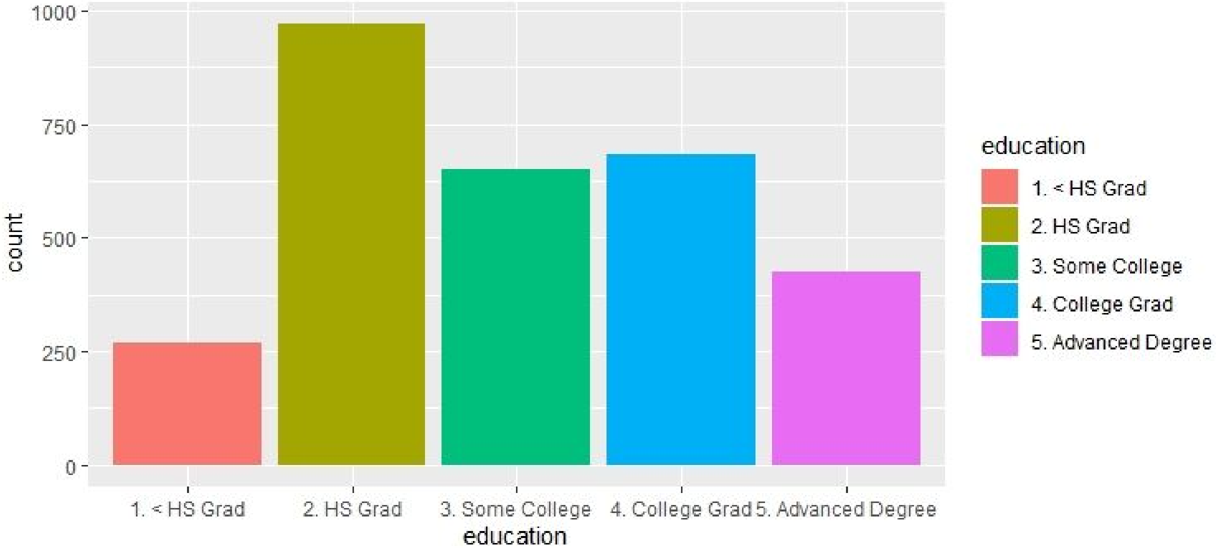


Figure 2: Visual Look at the Categorical Variable in Data Set.

From Figure 1, the distribution of education from the database is explored. Here, most of the workers either have a high school degree or college degree or something in between.

3.2 Visualization of overall Relationships

In order to see how one variable is related to another pairs plot diagram is plotted. The plot shows not only the diagram but also the correlations.

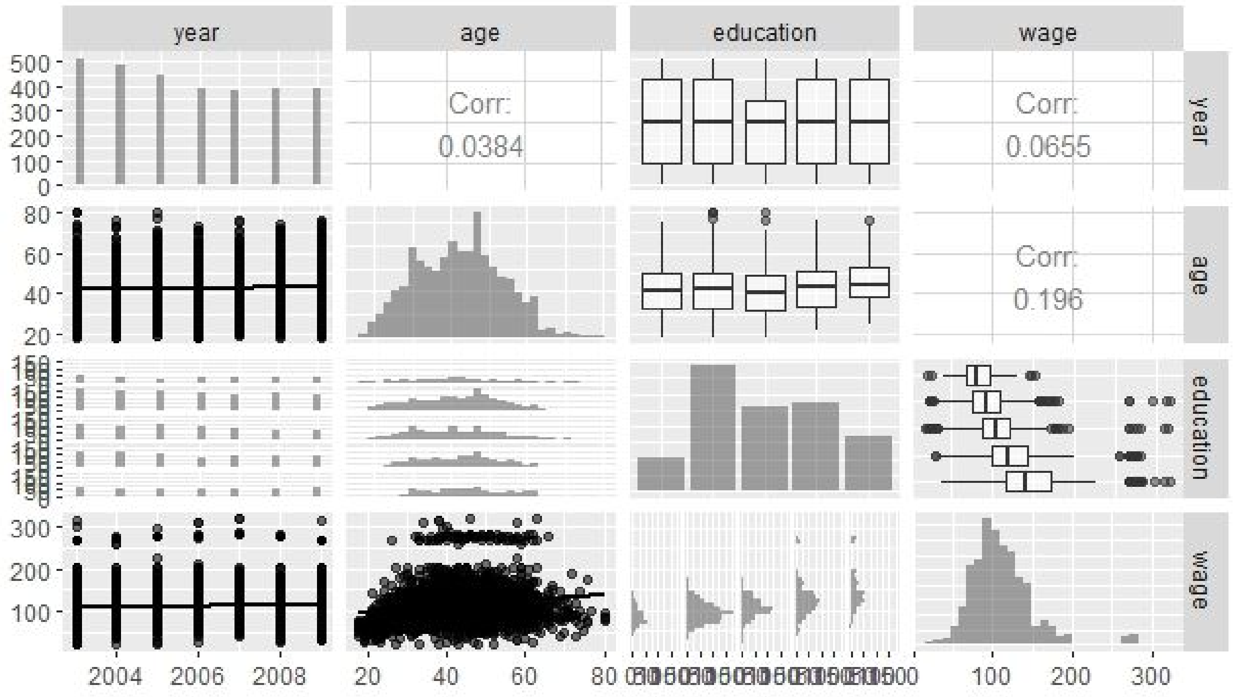


Figure 3: Visual Look at the Data Set with Pairs Plot.

From Figure 3, it is noted that none of the numeric variables (including wage) have large correlation with one another. Of course, wage is the response variable and its relationship to other variables is of interest in this report. The rightmost column and the bottom row shows how wage relates to each other variable in the data set. The right column shows year having almost no correlation with wage, while age has a small correlation value with wage. In other words, young workers tend to make slightly less money than older workers in general.

Now looking to the relationship of wage and the other categorical variables such as education, a box plot is drawn to interpret the relationship separately.

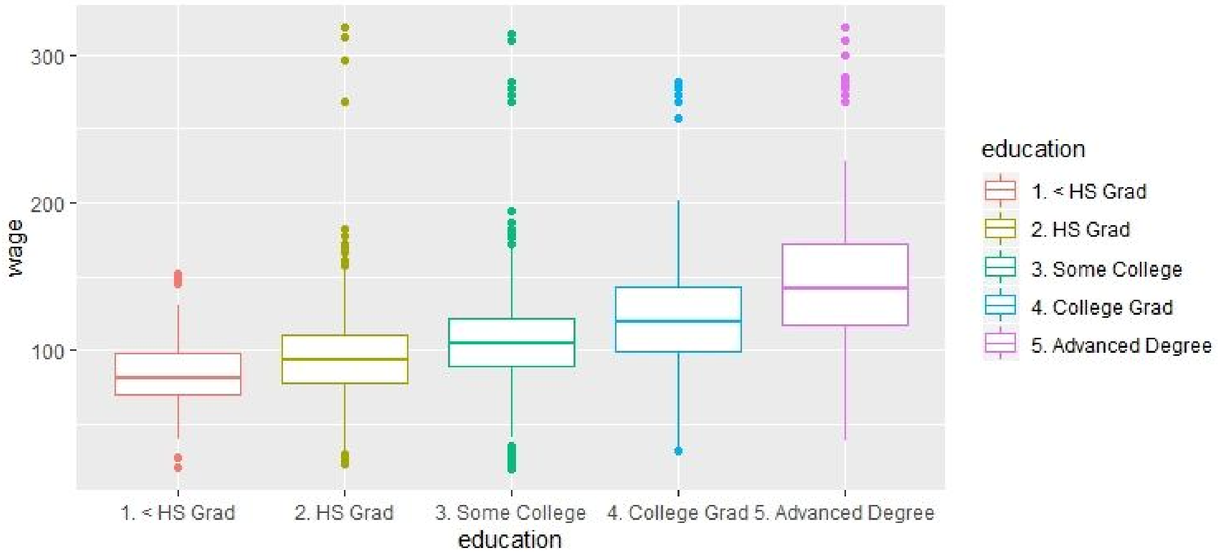


Figure 4: Visual Look at Wage Relationship with Categorical Variable (Education)

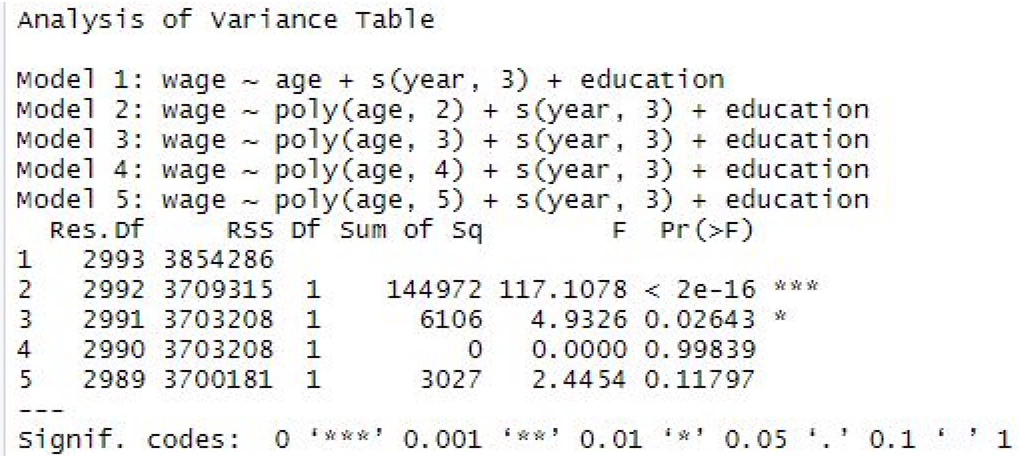
Here, it is found that wage generally increases as the workers' level of education increases. Looking to the plot, workers with an advanced degree have wages more than 250, some have more than 300. College grad workers have wage between 250 and 300. On the contrary, Most of the HS grad and less than a HS grad workers have wage less than 200.

By these visualizations the dataset has been explored. With this exploration of the dataset,the generalized additive models (GAMs) are fit to the dataset to predict some important decisions.

3.3 Results of GAMs Analysis

For GAMs (Generalized Additive Models), there is an Anova (Analysis of Variance) table and there are some figures. Statistical prediction is done by analyzing those figures and the table. Looking to the Anova table, it is used in order to find the degree for polynomial regression. This is performed by an F-test which is a hypothesis test.

Table 2: ANOVA Table for Polynomial degree for Age



```
Analysis of Variance Table

Model 1: wage ~ age + s(year, 3) + education
Model 2: wage ~ poly(age, 2) + s(year, 3) + education
Model 3: wage ~ poly(age, 3) + s(year, 3) + education
Model 4: wage ~ poly(age, 4) + s(year, 3) + education
Model 5: wage ~ poly(age, 5) + s(year, 3) + education
  Res.Df     RSS Df Sum of Sq        F  Pr(>F)    
1   2993 3854286                                  
2   2992 3709315  1    144972 117.1078 < 2e-16 ***
3   2991 3703208  1      6106   4.9326 0.02643 *  
4   2990 3703208  1         0   0.0000 0.99839    
5   2989 3700181  1      3027   2.4454 0.11797    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is proof from the anova outcomes that a GAM with a quadratic age function is better than a GAM with a linear age function. There is no proof, however, that a cubic age feature (p-value=0.02643) is required. In other words, Model 2 which has a polynomial degree of 7 is chosen based on the outcomes of this ANOVA.

Now, looking to the cross-validation (CV) to find out the best degree to use for the polynomial function of age and to show an example of how a tuning parameter is determined by CV.

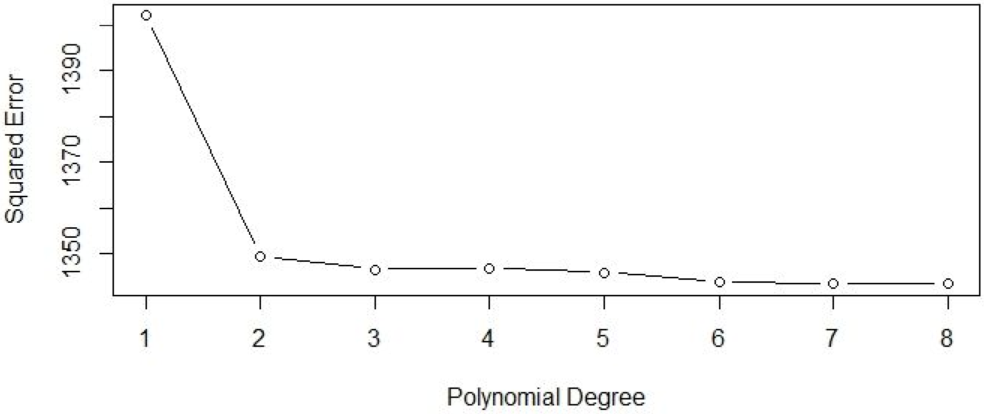


Figure 5: Squared Error via CV (for Polynomial Degree on Training Data)

From Figure 5, it is determined that a polynomial with 7 degrees of freedom produced the lowest squared error on the training data set. This reached the same conclusion as the anova method.

With the polynomial GAM having the best degree, a prediction can be described by taking a closer look at the GAM for the polynomial function of age with a degree of 7.

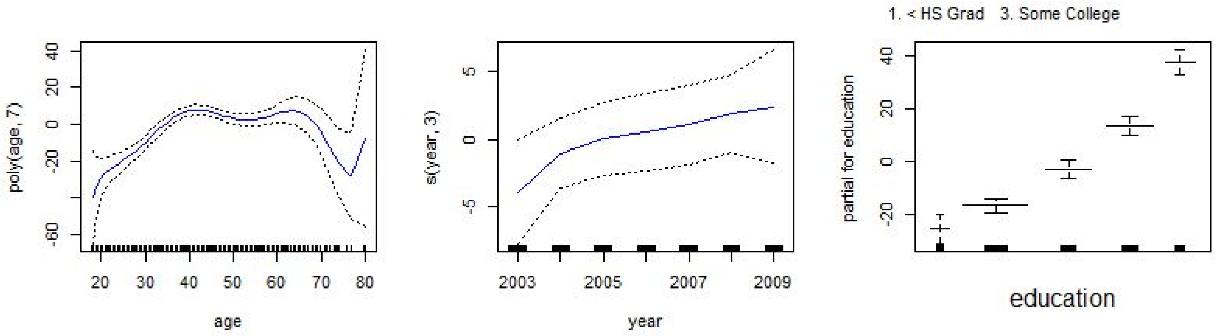


Figure 6: GAM Results with Polynomial (Degree = 2) of Age.

From Figure 6, the age plot indicates that holding the other predictors fixed, wage tends to be higher for intermediate values of age, and lowest for the very young and very old in general. However, in some cases, old people have highest wage ever. It could be they might have huge experience.

The year plot indicates that holding all the predictors fixed (aside from year), wage tends to increase in the early years and then increase less with later years and somewhere it decreases; this may be due to inflation.

The education plot indicates that holding the other variables fixed, wage tends to increase with education. In other words, the more educated a person is, the higher their salary, on average.

3.4 Findings from the results

From the results of analysis some associations between wage and other three variables (calendar year, age and education). However, there is no such strong evidence exists for those associations for wage with year and age. It is noted that education has moderately strong relationship with wage although there is no linear relationship. As a result, polynomial regression is appropriate tool which fit the data properly. In order to perform polynomial regression GAMs are used in this analysis. By using this model it can be predicted that given a certain education a wage can be predicted but with age it can be predicted slightly better than a random guess.

4. Conclusion

This report looked at a collection of wage and other information data in the Mid-Atlantic region for 3000 male employees. During exploratory data analysis, a number of observations between wage and other variables (year, age and education) were noted. For example, wage usually tends to rise as education levels rise.

Afterwards, generalized additive models (GAMs) were used to provide the framework for nonlinear fitting techniques to fit age and year to predict wage. For the predictors, a nonlinear fitting techniques was used on age, while using a smooth spline of year along with education. In order to make some predictions, the data was split into a training and testing data set. The training set consisted of half the observations, while the test set contained the rest of the observations.

This GAM was looked at in more detail with some plots and showed the behavior of each predictor while holding the other predictors constant.

References

G. James et al., An Introduction to Statistical Learning: with Applications in R,Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 1, © Springer Science+Business Media New York 2013.

Dzone.com (Dec. 08, 2014). Extract Rows?. [online] Available at: https://dzone.com/articles/learn-r-how-extract-rows [Accessed 1st August, 2019].

Datacamp.com (Mar. 11, 2019). Histogram? [online] Available at: https://dzone.com/articles/learn-r-how-extract-rows [Accessed 1st August, 2019].

Stadh.com (Not dated). ggplot2 Bar Plot? [online] Available at: http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization [Accessed 1st August, 2019].

Rdocumentation.org (Not dated). Pairs? [online] Available at: https://www.rdocumentation.org/packages/graphics/versions/3.6.1/topics/pairs [Accessed 2nd August, 2019].

Rdocumentation.org (Not dated). GAM? [online] Available at: https://www.rdocumentation.org/packages/mgcv/versions/1.8-28/topics/gam [Accessed 3rd August, 2019].

Homepages.inf.ed.ac.uk (Not dated). ANOVA in R? [online] Available at: http://homepages.inf.ed.ac.uk/bwebb/statistics/ANOVA_in_R.pdf [Accessed 3rd August, 2019].

Appendix (Code)

```
#----------------------------------------------------

# Reset R's brain

#----------------------------------------------------

rm(list=ls())



#----------------------------------------------------

# Reset graphic device

# As long as there is any dev open (exept "null device")

# close the active one!

# Caution: closes all open plots!!!!

#----------------------------------------------------

while(!is.null(dev.list()))

{

  dev.off()

}




require(ISLR)

library(ISLR)
```

```
attach(Wage)

assessment_dataframe <- Wage[sample(nrow(Wage), 3000), ]

View(assessment_dataframe)


#--------------------------------------------------------

#Extraction of year, age, education and wage from the entire dataset

#--------------------------------------------------------

assessment_dataframe <- assessment_dataframe[,c(1, 2, 5, 11)]

summary(assessment_dataframe)

View(assessment_dataframe)



#--------------------------------------------------------

#A short description of each variable (attribute) from that extracted dataset

#--------------------------------------------------------

library(knitr)  # kable

Variable = c("Year", "Age", "Education", "Wage")

Description = c("Year that wage information was recorded",

                "Age of worker",

                "Education level of worker",
            "Workers raw wage")

description_assesment_dataframe = data.frame(Variable, Description)

kable(description_assesment_dataframe, format='markdown')




#---------------------------------------------------------

#Visualizations using plots

#---------------------------------------------------------




#----------------------------------

#Visualizing Numeric/Continuous variables(year, age and wage) in the extracted Dataset

#----------------------------------

library(ggplot2) # Plotting

library(purrr) # Organizing

library(tidyr) # Organize/tidy data


assessment_dataframe %>%

  keep(is.numeric) %>%

  gather() %>%
  ggplot(aes(value, fill=key)) +
  facet_wrap(~ key, scales="free") +
  geom_histogram( bins=sqrt( nrow(OJ) ) ) +
  theme(legend.position="none")


#-----------------------------------
#Visualizing Categorical variable(education) in the extracted Dataset
#-----------------------------------
ggplot(data = assessment_dataframe, aes(education,
                                        fill = education)) +
  geom_bar(mapping = aes(x = education))



#-----------------------------------
#Visualizing realtionship between variables(wage and others 3 variables) using pairs plot in the extracted Dataset
#-----------------------------------
library(GGally) # ggpairs plot
library(reshape) # Melt data for plotting
pairs_plot = ggpairs(assessment_dataframe[, ],
        aes(alpha=0.6),
        upper = list(continuous = wrap("cor", size = 4)),
        diag = list(continuous = "barDiag"),
        lower = list(continuous = "smooth"))
suppressMessages(print(pairs_plot))


#------------------------------------------
#Visualizing realtionship between wage and categorical variables(education) using box plot in
the extracted Dataset
#------------------------------------------
ggplot(data = assessment_dataframe, mapping = aes(x = education, y = wage, color=education))
+
  geom_boxplot()



#---------------------------------------------------------------------
#Obtaining suitable Statistics and Prediction using some methods or models
#---------------------------------------------------------------------
```

```
#----------------------------------

#GAMs (Generalized Addidtive Models)

#Analysis of variance (ANOVA, using an F-test)

#Fit five different models and sequentially compare the simpler model to the more complex model

# make year a smooth spline with 3 DoF to help show GAM can combine multiple non-linear models

#----------------------------------

library(splines) # Splines

library (gam)    # GAM

fit.1= lm(wage~age + s(year,3) + education, data=assessment_dataframe)

fit.2= lm(wage~poly(age,2) + s(year,3) + education, data=assessment_dataframe)

fit.3= lm(wage~poly(age,3) + s(year,3) + education, data=assessment_dataframe)

fit.4= lm(wage~poly(age,4) + s(year,3) + education, data=assessment_dataframe)

fit.5= lm(wage~poly(age,5) + s(year,3) + education, data=assessment_dataframe)

anova(fit.1, fit.2, fit.3, fit.4, fit.5)


#----------------------------------

# split Wage in half training and testing

#----------------------------------

set.seed(8)
```

```
train=sample(1:nrow(assessment_dataframe), 0.5*nrow(assessment_dataframe))

assessment_dataframeDf.test = assessment_dataframe[-train,]

assessment_dataframeVar.test= assessment_dataframe$wage[-train]


#----------------------------------

# Determining the best degree to uge the polynomial function of age via CV(Cross-Validation)

#----------------------------------

library(caret) # Showing Confusion Matrix Data

library(boot)    # cv.glm


# Create folds for each observation in training data

set.seed(8)

folds = createFolds(assessment_dataframe$wage[train], k=10) # 10 folds is default


# Set up a matrix which will have 1 row for every

# CV iteration and 1 column for each polynomial degree.

polynomialDF = 8

Errors_CV = matrix(nrow=10, ncol=polynomialDF)
```

```
# Loop over degrees of polynomial
for(poly_degree in 1:polynomialDF){
  # Loop over folds of cv
  for(k in 1:10){
    fit = gam(wage~poly(age, poly_degree) + s(year,3) +
                education, data=assessment_dataframe, subset = -folds[[k]])
    predictions = predict(fit,newdata=assessment_dataframe[folds[[k]],])

Errors_CV[k,poly_degree]=mean((predictions-assessment_dataframe[folds[[k]],c("wage")])^2)
  }
}


# Find which degree has lowest average MSE over all k folds:
mean_cv_errors = apply(Errors_CV, 2, mean)
plot(mean_cv_errors, type = 'b', xlab = "Polynomial Degree", ylab = "Squared Error")


best_poly_degree = which.min(mean_cv_errors)


#-----------------------------------
# GAM for the polynomial function of age with a degree of 7
```

```
#-----------------------------------

par(mfrow = c(2, 3))

### Smoothing spline

fit=gam(wage~ poly(age, 7) + s(year, 3) + education, data = assessment_dataframe, subset =
train)

plot(fit, se=TRUE, col ="blue")
```