Contents

[Introduction 2](#_Toc10809266)

[Authorisation and Purpose 2](#_Toc10809267)

[Limitations 2](#_Toc10809268)

[Scope 2](#_Toc10809269)

[Methodology 2](#_Toc10809270)

[Data Setup 2](#_Toc10809271)

[One-variable Analysis 3](#_Toc10809272)

[One-variable Analysis(1) 3](#_Toc10809273)

[One-variable Analysis(2) 4](#_Toc10809274)

[Two variable Analysis 5](#_Toc10809275)

[Two variable analysis(1) 5](#_Toc10809276)

[Two variable analysis(2) 6](#_Toc10809277)

[Advanced Analysis 7](#_Toc10809278)

[Clustering Analysis 7](#_Toc10809279)

[Clustering 7](#_Toc10809280)

[Linear regression 8](#_Toc10809281)

[Linear Regression 1 8](#_Toc10809282)

[Linear Regression 2 9](#_Toc10809283)

[Conclusion 10](#_Toc10809284)

[Reflection 11](#_Toc10809285)

ICT110 INTRODUCTION TO DATA SCIENCE

Task 2 - Data analysis report

# Introduction

## Authorisation and Purpose

The purpose of this Data analysis report for other researchers, business representatives, and government agencies. Given the dataset about Australian Death Fatalities, this document potrays the analysis of the data and prepare a report about my own work and findings.

## Limitations

This research is limited to the year 2010-2018.The only file that will be analysed is BITRE\_ARDD\_Fatalities\_Mar\_2019\_II .

## Scope

As there is no specific requirement, there is an advantage that we can dig and play with data to learn as much as we can. One variable analysis and two variable analysis is done and graph will be provided with each analysis. K-means and clustering will be explained and done to group countries through clustering. Linear regression will be explained and two regression analysis will be performed and models will be plotted.

## Methodology

Report has been provided in a .csv file and references are collected from the internet and library resources.

## Data Setup

By using the command, getwd() we can know the working directory of the system. By using the command setwd() we can set the path to the current working directory. Then we give command read.csv and assign it to the name given.

> getwd()

[1] "C:/Users/radha/OneDrive/Desktop/AssignmentICT110"

> library(cluster)

>

> Fatalities <- read.csv("BITRE\_ARDD\_Fatalities\_Mar\_2019\_II.csv")

Library(cluster) is used for cluster visualisation

# One-variable Analysis

## One-variable Analysis(1)

In this, I have taken two types of road users i.e Driver and Passenger. By using COUNTIF function on Excel, I found out that in 2010, there are 686 accidents in which Driver was the road user. And, there are 284 cases in which Passengers are road user. According to representation of this data in Bar Graph, we can clearly compare the data for Driver and Passenger.

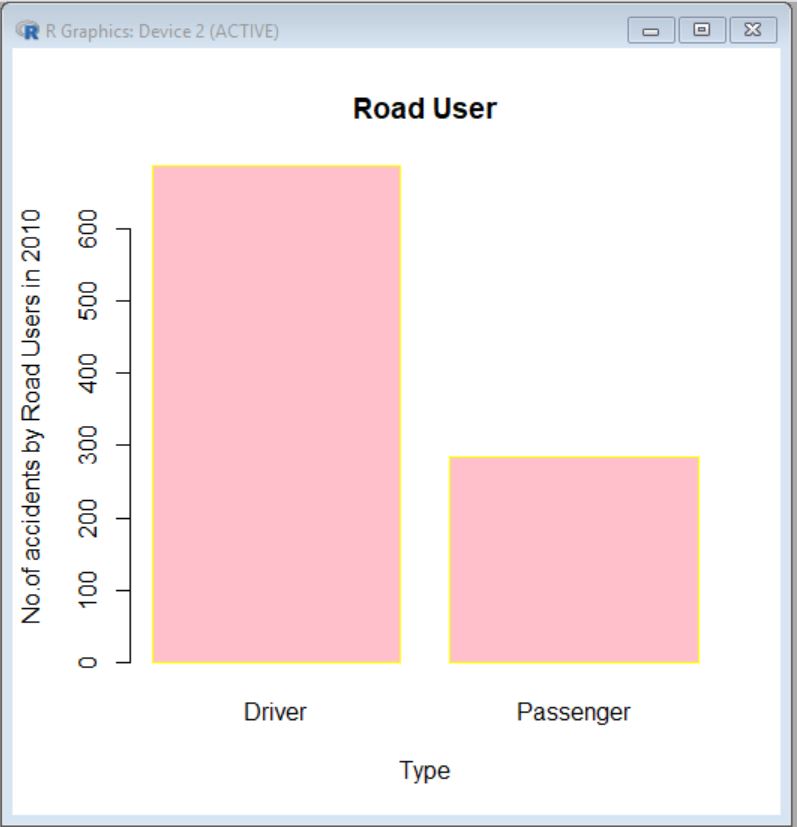
.

> H <- c(686,284)

> M <- c("Driver","Passenger")

> barplot(H,names.arg=M,xlab="Type",ylab="No.of times in 2010",col="pink",

+ main="Road User",border="yellow")



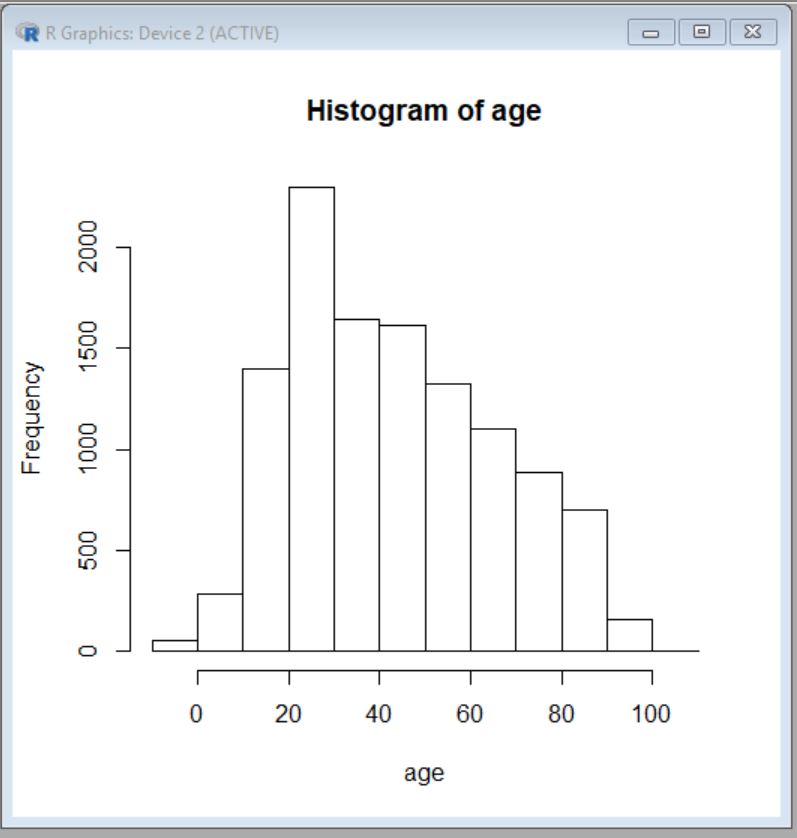
## One-variable Analysis(2)

A histogram is a visual portrayal of the dispersion of a dataset. All things considered, the state of a histogram is its most clear and instructive trademark: it enables you to effortlessly observe where a generally huge measure of the information is arranged and where there is next to less data to be found. Here from the given dataset, limiting the year of records to 2010-2018. Road fatalities happening to people who’s age is from 0-100. The records are like (e.g46,30,49, etc) According to the data, most fatalities happen to the people who’s age is 20-30. Lowest is from age 90-100.

> age <- read.csv("Book2.csv")

> hist(age)

> age <- sapply(age, as.numeric)



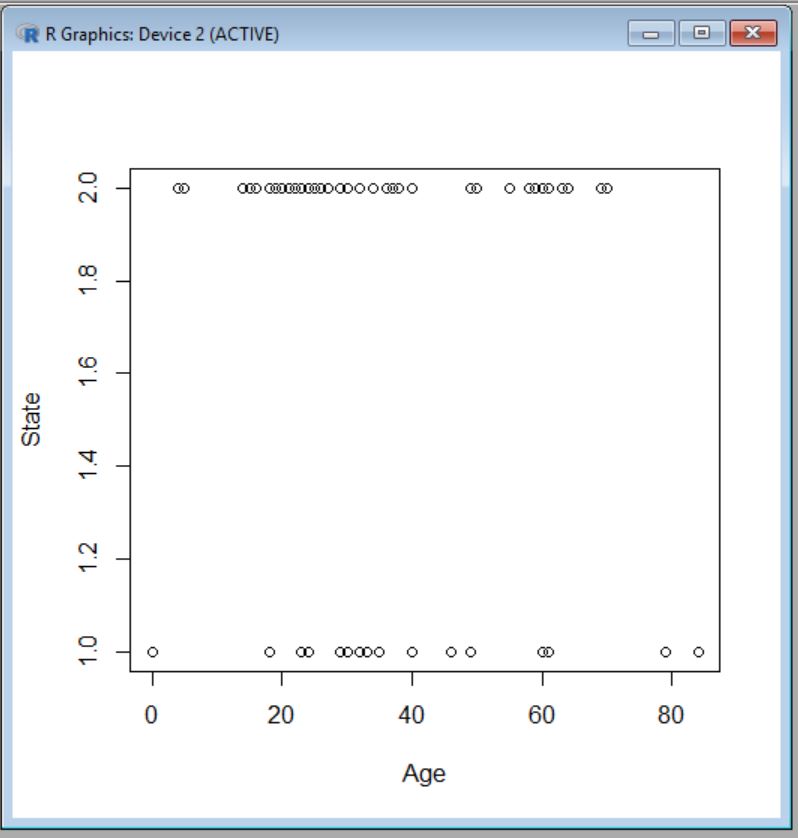
# Two variable Analysis

## Two variable analysis(1)

In this, I have taken only two states ACT and NT. R automatically converted it into 1.0 and 2.0 on the graph. I have taken Age in X-Axis. By looking at the graph, we can tell that between age 20-40, most read accident deaths have occurred in ACT as well as NT. There is only one accident in year 2010, in ACT above 80 years of age.

> graph <- read.csv("Book2-3.csv")

> plot(x=graph$Age, y=graph$State, xlab = "Age", ylab = "State")

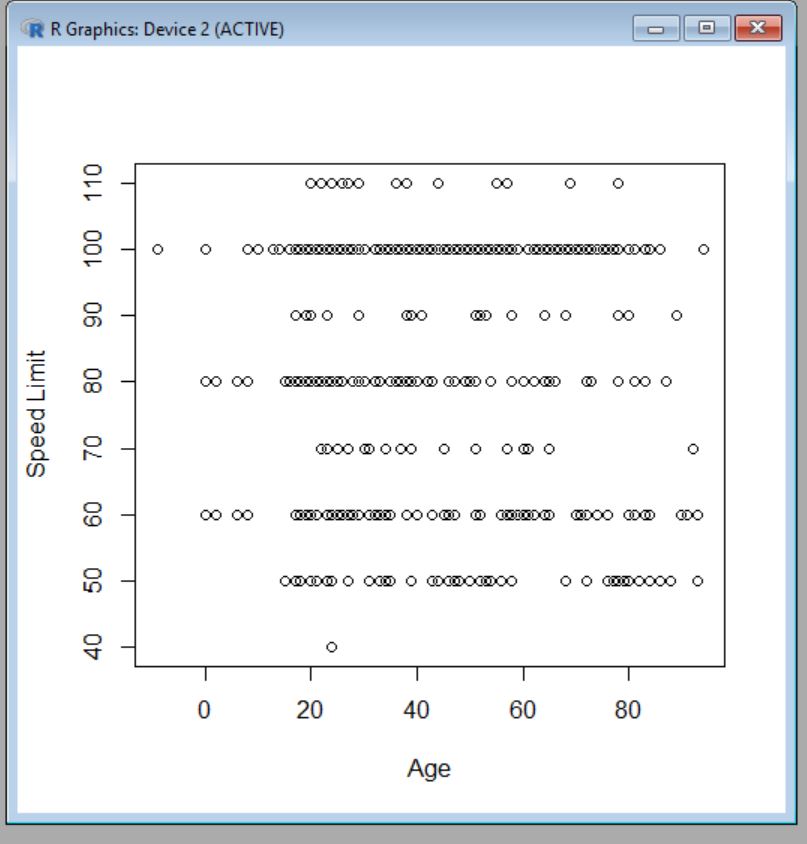


## Two variable analysis(2)

This graph is a representation of Age Vs. Speed Limit in the year 2010 in ACT and NSW. According to data representation, most of accident fatalities have happened on speed Limit 100. Least fatalities have happened on speed limit 40. In the data of age, maximum records are from age 20-60.The age is from 0-80 and Speed Limit is from 0-110.

> graph <- read.csv("RadhaGraph.csv")

> plot(x=graph$Age, y=graph$SpeedLimit, xlab="Age", ylab="Speed Limit")



# Advanced Analysis

## Clustering Analysis

The strategy for recognizing comparable gatherings of information in a dataset is called clustering. It is a standout amongst the most mainstream procedures in data science. Entities in each gathering are relatively more like elements of that gathering than those of different groups. (Anon., 2019)

K means cluster is one more method to quickly cluster large data sets. (Kaushik, 2016)

## Clustering

With clustering, grouping of Age in respect to specific attributes can be done. In this clustering, age group will be grouped based on Speed Limits and Months of year. Optimal number of cluster are found after finding k-means to the SSE. Analyzing the graph the SSE becomes insignificant at around 3 clusters. The SSE value is 76.875. after doing clustering of 3 cluster, it can be seen that one cluster is singificanlty larger in size than the others. It takes up over 2/3 of all age groups.

>data <- read.csv(“radha.csv”)

>library(cluster)

>clustplot(data.stand, k.means.fit$cluster, main=’2D representation of the Cluster solution’ , color=TRUE, shade=TRUE, labels=2, lines=0)

>data <- read.csv(“radhac.csv”)

>wssplot <- function(data, nc =15, seed=1234){

  Wss <- (nrow(data)-1\*sum(apply(data,2,var))

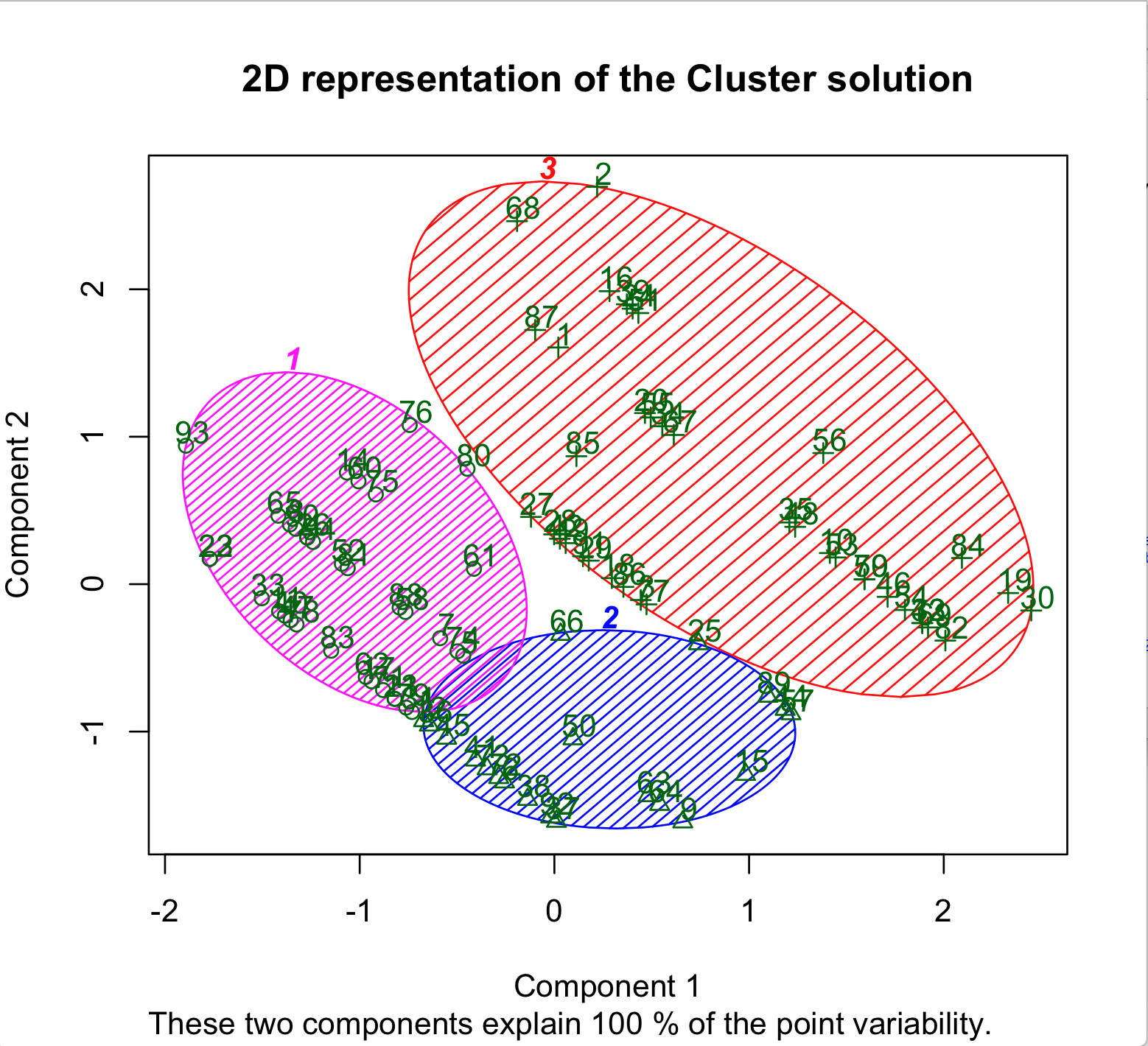
 for (I in 2:nc){

     set.seed(seed)

     wss[i] <- sum(kmeans(data, centers=i)$withinss0}

  plot(1:nc, wss, type=”b”, xlab=”Number of Clusters”, ylab=”Within groups sum of squares”)}

wssplot(data.stand, nc=6)



## Linear regression

Linear Regression tries to demonstrate the connection between two variables by fitting a direct condition to watched information. (Anon., 2018) One variable is viewed as an illustrative variable, and the other is viewed as a dependant variable. For instance, a modeler should need to relate the weight of people to their statures using a Linear regression model. (Anon., 2018)

### Linear Regression 1

According to the data represented in this linear regression model, it can be assumed that with the increasing speed limit, to an extent, the age also increases. A line straight line of best fit is created and showed on the diagram, it supposedly is slicing through the unfilled space between the lower and higher values. While the linear regression line here could be used as a loose basis to demonstrate the relationship, practically speaking it could prompt predictions with enormous variance from the base data.

> cor(data$SpeedLimit, data$Age)

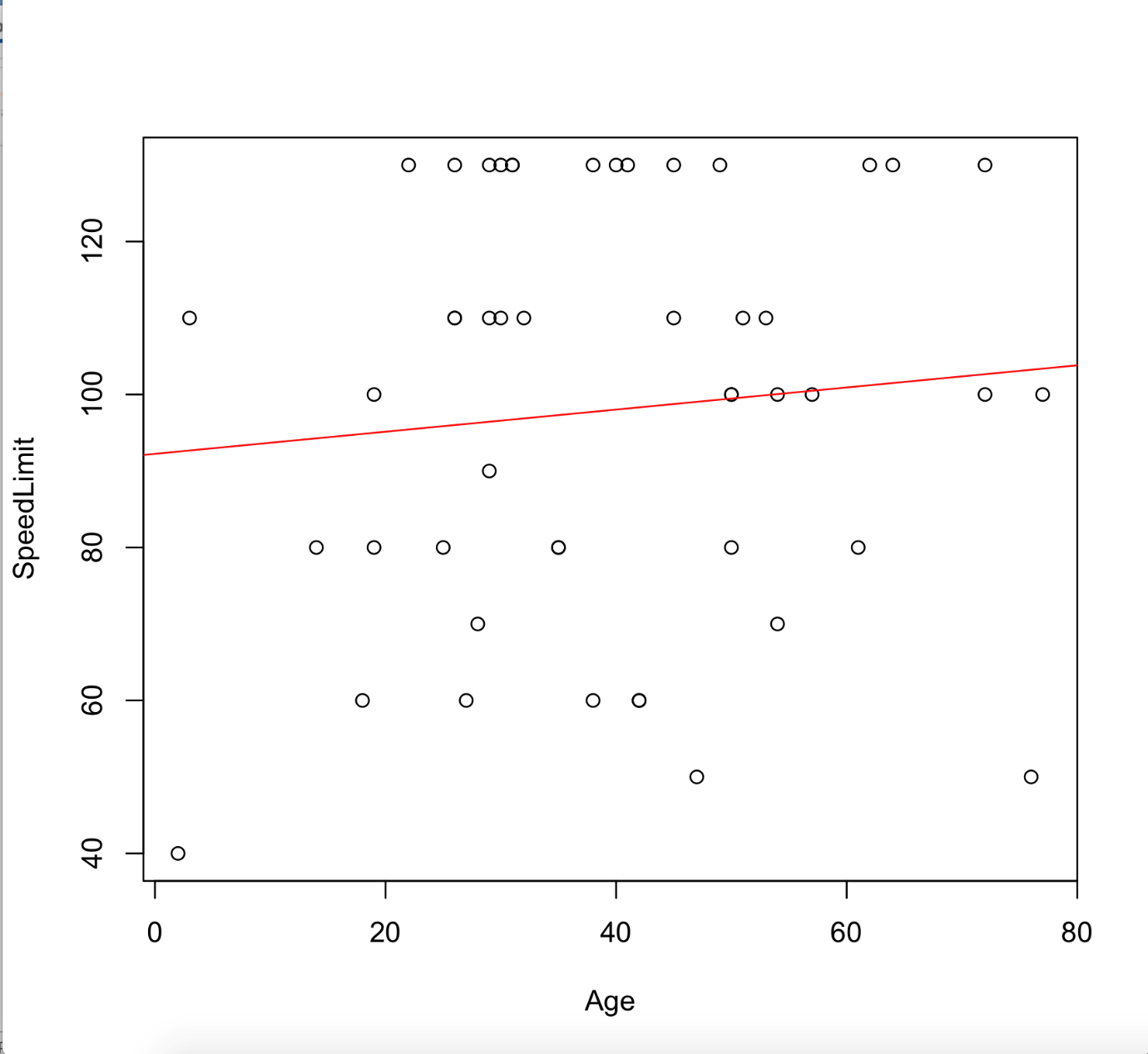
> linearMod <- lm(SpeedLimit ~ Age, data=data)

> print(linearMod)

> summary(linearMod)

> plot(SpeedLimit ~ Age, data=data)

> abline(linearMod, col="red")



### 

### Linear Regression 2

In this Linear Regression, this predicts that the age on which most of the fatalities will happen. This model predicts that most fatalities that will happen will be at which age. In this model it is shown that prediction of age at which most deaths are likely to happen is 35.

> cor(data$Month, data$Age)

> linearMod <- lm(Month ~ Age, data=data)

> print(linearMod)

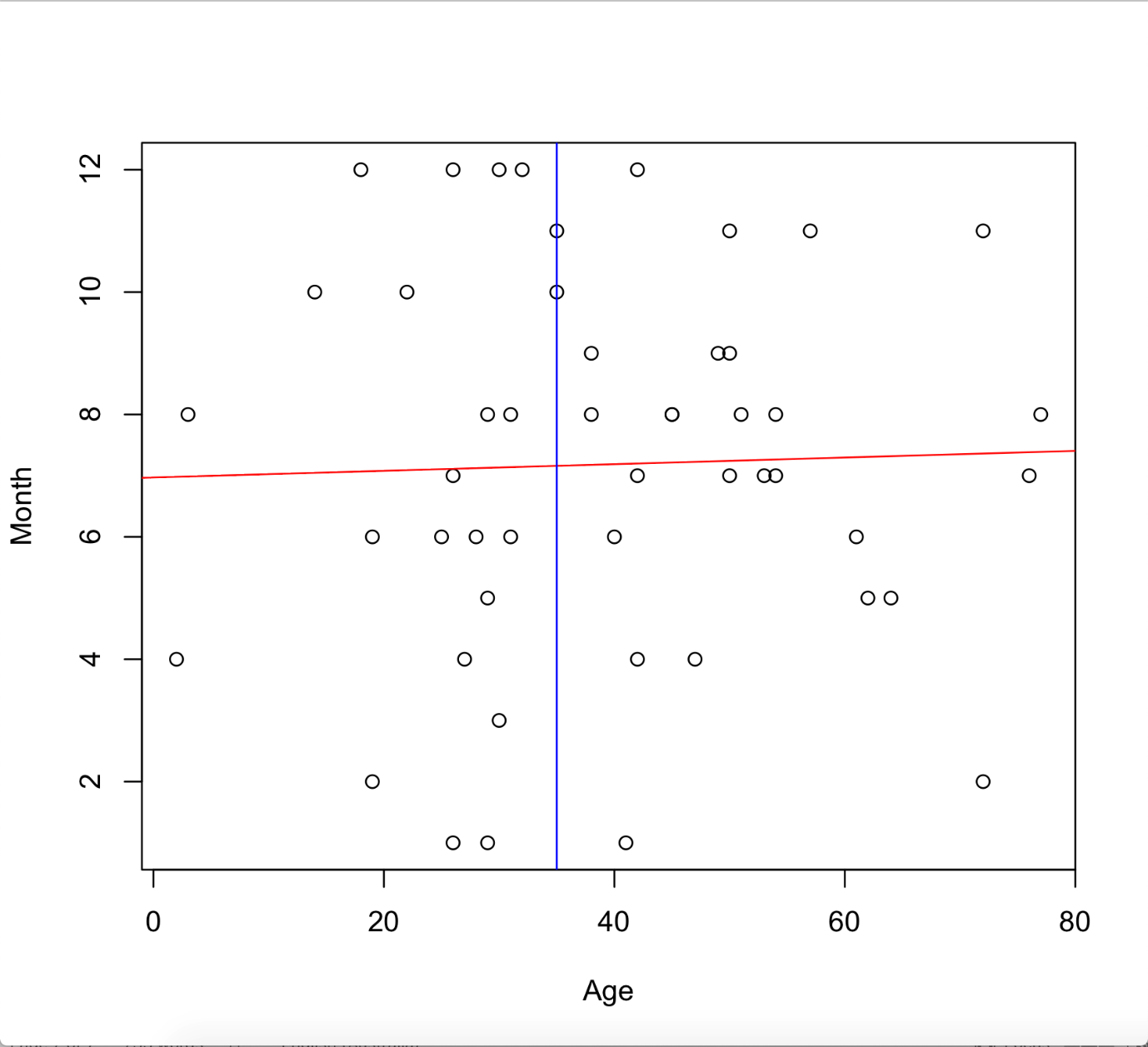
> summary(linearMod)

> plot(Month ~ Age, data=data)

> abline(linearMod, col="red")

> predict(linearMod, data.frame(Age=35))

> abline(v=35, col="blue")



# Conclusion

In this data analysis report, numerous patterns have been seen. Using one variable analysis on a barchart, frequency of driver and passengers in the records of dataset was found. Histograms could find the increase in average age of the records of dataset. Utilizing two variable via scatter plot found the pattern of two columns of dataset i.e age and state in first and Age and Speed limit in second. By clustering we can show the groups of the records. A linear trend was seen in Speed Limit Vs Age. A line of regression was observed in Month vs Age, allowing predictions to be plotted.

# Reflection

The data set i.e [BITRE\_ARDD\_Fatalities\_Mar\_2019\_II.csv](https://online.usc.edu.au/bbcswebdav/pid-1655688-dt-content-rid-9128887_2/xid-9128887_2) is a huge file with some Not Available data insufficient numeric columns. There are numerous problems in the file due to which it was required to make a new excel file to implement the commands of R for data manipulation. Due to excess data, in the scatter plot, the dots overlapped on each other which made the explanation of analysis difficult. If this data analysis is done again anytime, the main solution that can be done is make the dataset file more prepared for the analysis.