ACADGILD

SESSION 6: Visualization & Plotting

Assignment 1

Table of Contents

1. Introduction

This assignment will help you understand the concepts learnt in the session.

2. Objective

This assignment will test your skills on Visualization and Plotting operations in R.

3. Prerequisites

Not applicable.

4. Associated Data Files

Not applicable.

5. Problem Statement

1. Import the Titanic Dataset from the following link: https://drive.google.com/file/d/1JTJCjdGuUxzKXYlwOavwovB01k6FWg3r/view?ts=5b42ea10

Perform the below operations:

a. Pre-process the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

b. `library(readr) # Reading in data`
c. `library(dplyr) # Data manipulation`
d. `library(tibble) # Data manipulation`
e. `library(ggplot2) # Data visualization`
f. `library(ggthemes) # Data visualization`
g. `library(RColorBrewer) # Data visualization`

```
library(readr)
Titanic3 <- read_csv("C:/Users/archana/Desktop/archana/GITS/titanic3.csv")
str(Titanic3)
head(Titanic3)
tail(Titanic3)
str(Titanic3$name) # check structure, as only charecter vectors can be split using
 strsplit function
Titanic3$name<-as.character(Titanic3$name)
str(Titanic3$name)
#telling R to call rbind, on two charecters split by strsplit.
#in strsplit, as the data has many " ", and all breaks in many pieces
# hence, using sub() {and not gsub()}, which replaces only first pattern
# so, sub changes first space in ; and the strsplit splits along ; and then rbind
binds along colums, which is called by do.call
namessplit<-do.call(rbind,strsplit(sub(" ",";",Titanic3$name),";"))
head(namessplit)
#converting the charecters to data frame and naming the columns
namessplit<-data.frame(namessplit)
names(namessplit)<-c("family_name", "name")
head(namessplit)
str(namessplit)
#getting title separated from first name
Title<-do.call(rbind,strsplit(sub(" ",";",namessplit$name),";"))
head(Title)
Title<-data.frame(Title)
names(Title)<-c("title", "first_name")
head(Title)
str(Title)
head(Title)
#merging the rownames in titanic survival data to form new data set
#similar to text to columns in excel
#tried merge function which didnt work as expected, but cbind is simpler and gives
 right data.
str(Titanic3)
TitanicData<-cbind(namessplit,Titanic3)
head(TitanicData)
View(TitanicData)
str(TitanicData)
```

```
TitanicData<-cbind(Title,TitanicData)
head(TitanicData)
View(TitanicData)
#graphical representation of the data in various forms
#barplot -No. of passangers by Family name

familyname<-table(TitanicData$family_name)
View(familyname)
barplot(familyname,main = "survival as per family name", xlab = "family_name", yla
b = "count",col ="red")

#barplot -No. of passangers by Title

Title<-table(Title)
Title
View(Title)
barplot(Title,xlab = "Title", ylab = "No. of Passangers",
        main = "survival as per Title" , col = c("blue", "red"), las=3)
text(Title, 0,table(Title), pos = 3, srt = 90)
```

Convert Variable into Factors

- Convert **Pclass**, **Survived** and **Sex** Variables into Factors using the *mutate* function
- Keep **Age** numeric

```
titanic <- titanic %>%
  mutate(Pclass = factor(Pclass),
         Survived = factor(Survived),
         Sex = factor(Sex))
```

Look at the Total Survival Rate

- Use *table* function to look at the Survival rates
- Convert table to a **tibble** (similar to dataframe) to use later when plotting with ggplot2 to add text to the graph

- Use *rename* function to rename default Column names

```
survival <- table(titanic$Survived) %>%
  as_tibble() %>%
  rename(Survived = Var1, Count = n)

survival
```

Look at the Total Survival Rate Proportion

- Use *prop.table* to get the proportion

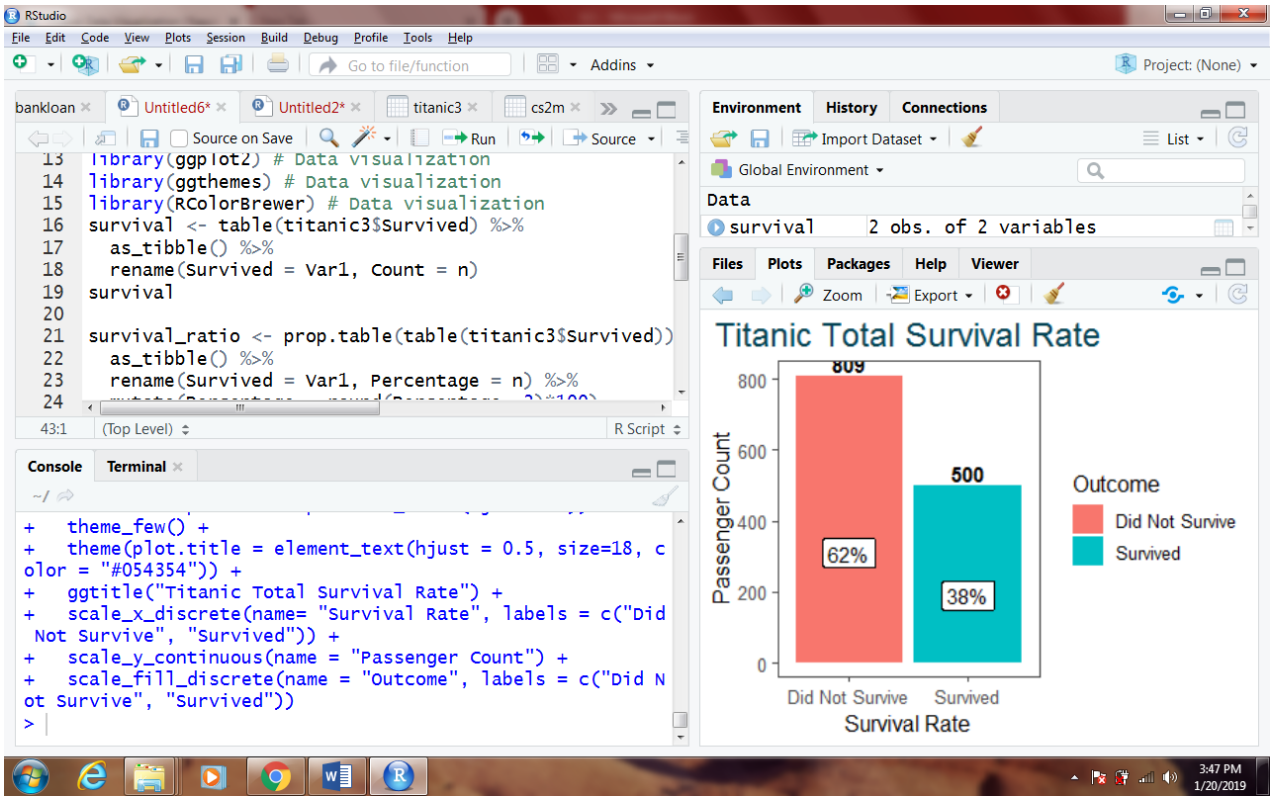
```
survival_ratio <- prop.table(table(titanic$Survived)) %>%
  as_tibble() %>%
  rename(Survived = Var1, Percentage = n) %>%
  mutate(Percentage = round(Percentage, 2)*100)

survival_ratio
```

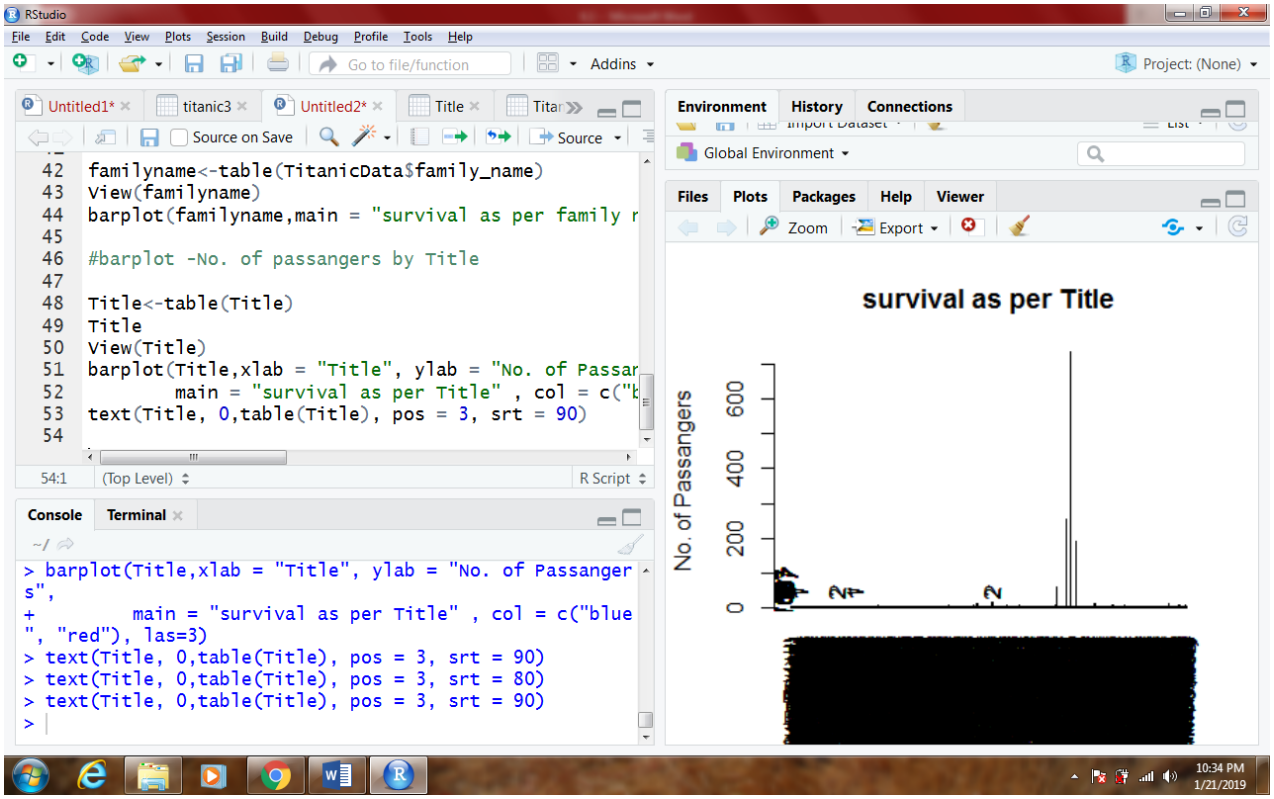
Plot the Total Survival Rate

- Using a barplot with *theme_few* theme from the **ggthemes** package
- Add some styling to the plot: Center the Title and color, Edit the Legends
- Use tibble of survival data in *geom_text* to add the Count to the plot
- Use tibble of survival data ratio in *geom_label* to add the Percentages to the bars

```
titanic %>%
  ggplot() +
  geom_bar(aes(x = Survived, fill = Survived)) +
  geom_text(data = survival,
            aes(x = Survived, y = Count, label = Count),
            position = position_dodge(width=0.9),
            vjust=-0.25,
            fontface = "bold") +
  geom_label(data = survival_ratio,
             aes(x = Survived, y = Percentage, label = paste0(Percentage, "%"), gr
oup = Survived),
             position = position_stack(vjust = 5)) +
  theme_few() +
  theme(plot.title = element_text(hjust = 0.5, size=18, color = "#054354")) +
  ggtitle("Titanic Total Survival Rate") +
```

```
  scale_x_discrete(name= "Survival Rate", labels = c("Did Not Survive", "Survived"
)) +
  scale_y_continuous(name = "Passenger Count") +
  scale_fill_discrete(name = "Outcome", labels = c("Did Not Survive", "Survived"))
```





h. Represent the proportion of people survived by family size using a graph.

Look at Survival Based on Family Size

Add a Varibale for Family Size

- Combine **SibSp** and **Parch** variables together and add 1 (for self)
- Use the *mutate* function to add **FamilySize** to the dataset

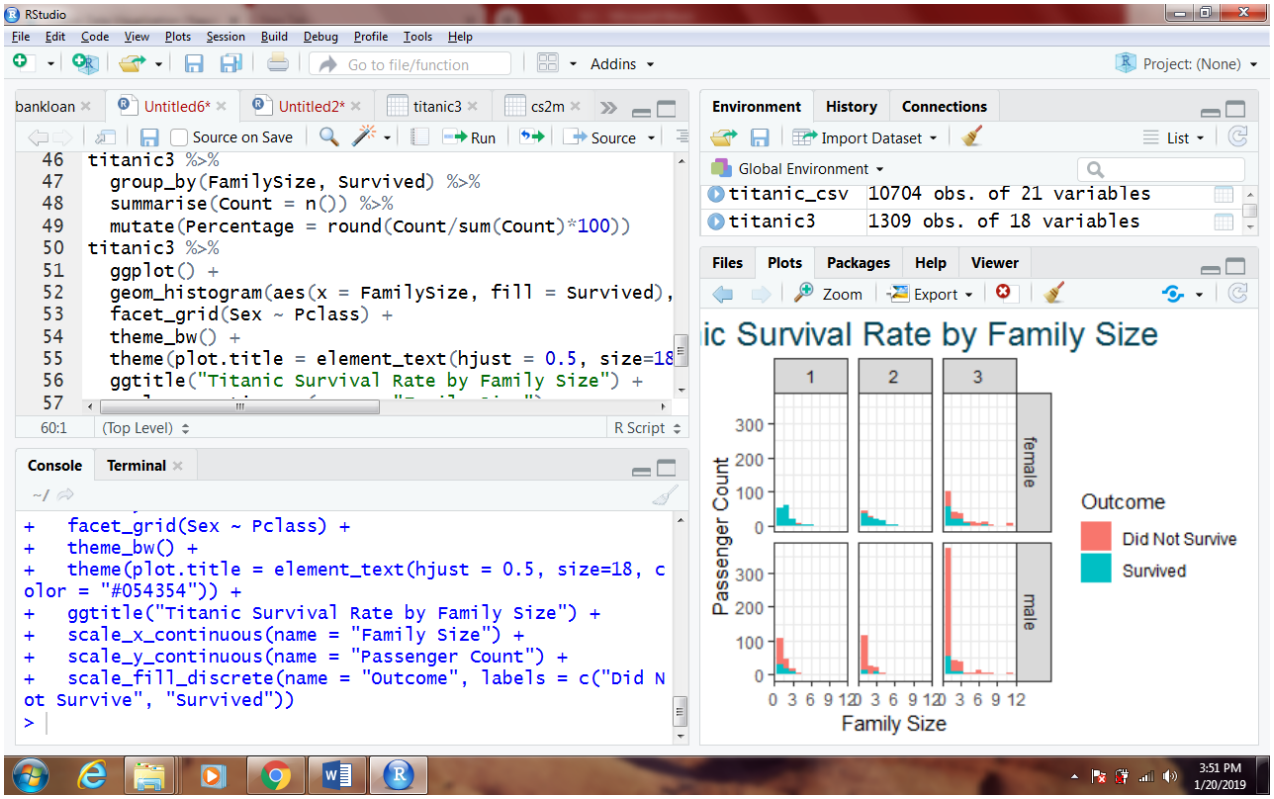
```
In [24]:
titanic <- titanic %>%
  mutate(FamilySize = 1 + SibSp + Parch)
```

Look at Survival Rate Based on Family Size

```
In [25]:
titanic %>%
  group_by(FamilySize, Survived) %>%
  summarise(Count = n()) %>%
```

Plot Survival Rate Based on Family Size

```
In [26]:
titanic %>%
  ggplot() +
  geom_histogram(aes(x = FamilySize, fill = Survived), binwidth = 1) +
  facet_grid(Sex ~ Pclass) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, size=18, color = "#054354")) +
  ggtitle("Titanic Survival Rate by Family Size") +
  scale_x_continuous(name = "Family Size") +
  scale_y_continuous(name = "Passenger Count") +
  scale_fill_discrete(name = "Outcome", labels = c("Did Not Survive", "Survived"))

  mutate(Percentage = round(Count/sum(Count)*100))
```



One more way:

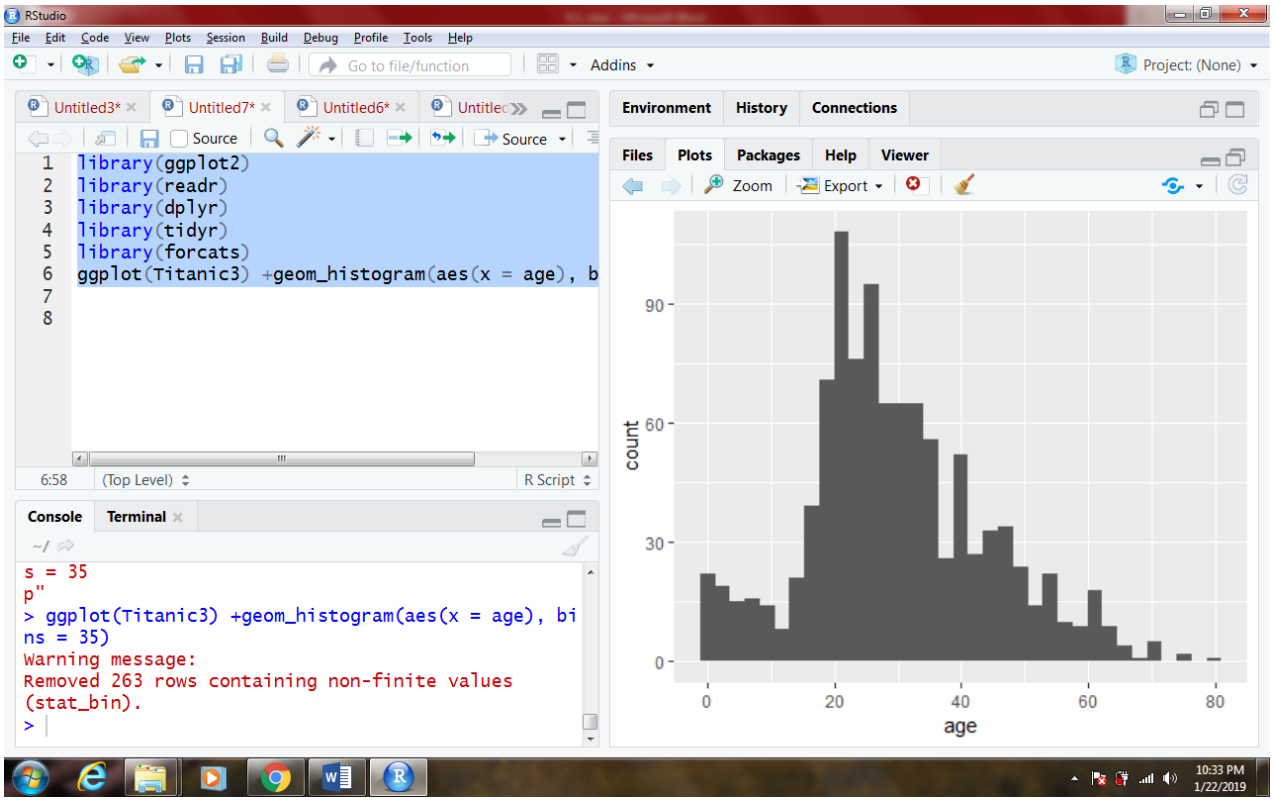
```
View(Titanic3Data)
SurvivedTitle<-table(TitanicData$Survived, TitanicData$title)
#survived is 0, first row. we will take only that
p<-SurvivedTitle[1,]

#barplot of survived numbers per title
barplot(p,xlab = "Title", ylab = "survived",
        main= "Survival as per title", col=rainbow(length(p)))
#pie chaart showing proportion of survival title wise

pie_chart<-pie(p, main = "Pie-Chart of Titles survived", col = rainbow(length(p)) )
legend("topright", names(p), cex= 0.5, fill = rainbow(length(p)))
```

i. Impute the missing values in Age variable using Mice library, create two different graphs showing Age distribution before and after imputation

Age distribution before Imputation

```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(forcats)
ggplot(Titanic3) +geom_histogram(aes(x = age), bins = 35)
```

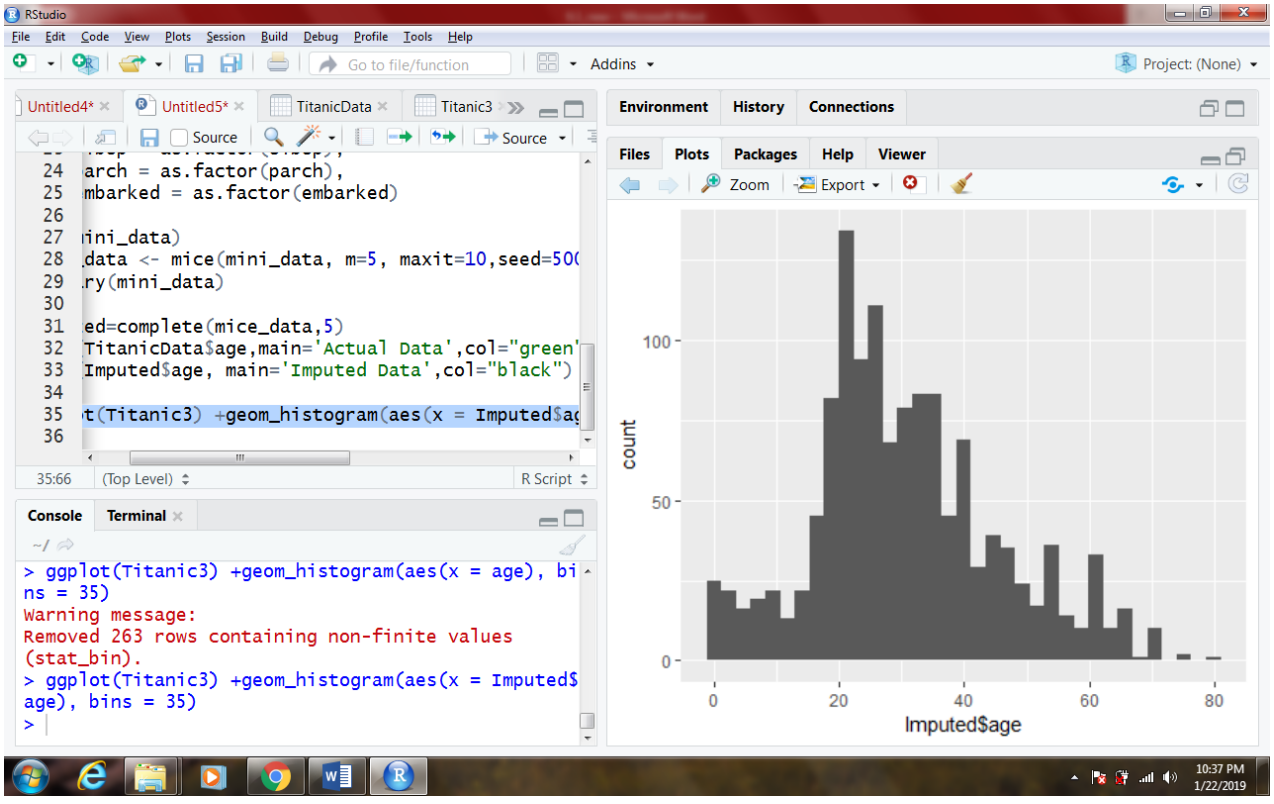
```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(forcats)
ggplot(Titanic3) +geom_histogram(aes(x = age), bins = 35)
```



Age distribution after imputation

```
library(mice)
sum(is.na(TitanicData$age))
str(TitanicData)

#Removing columns 1,2,3,4,5,7,12,13,14,16,17,18

mini_data <- TitanicData[-c(1,2,3,4,5,7,12,13,14,16,17,18)]
View(mini_data)

md.pattern(mini_data)

library(dplyr)
mini_data <- mini_data %>%
```

```
  mutate(
    survived = as.factor(survived),
    sex = as.factor(sex),
    age = as.numeric(age),
    sibsp = as.factor(sibsp),
    parch = as.factor(parch),
    embarked = as.factor(embarked)
  )
str(mini_data)
mice_data <- mice(mini_data, m=5, maxit=10,seed=500)
summary(mini_data)

Imputed=complete(mice_data,5)
hist(TitanicData$age,  main='Actual Data',col="green")
hist(Imputed$age, main='Imputed Data',col="black")

ggplot(Titanic3) +geom_histogram(aes(x = Imputed$age), bins = 35)
```



6. Expected Format

1. R file should be submitted where applicable.
2. R file should be in PDF or in .r format
3. Proper screenshots of the outputs should be submitted as well
4. The r codes, if submitted in any other format, will be subjected to deduction in marks

Note: Your solution will not be entertained if it is any other format, e.g., .zip, .doc, .rtf etc.

7. Approximate Time to Complete Task

30 mins.