FINAL PROJECT REPORT

IST 687 - APPLIED DATA SCIENCE

SUBMISSION DATE: May 29, 2015

SUBMITTED BY-

GARNETTE PEREIRA [672164454]
PRATIMA GANGURDE [708175933]

Table of Contents

Dataset Description

North Central Cancer Treatment Group (NCCTG) Lung Cancer Data

According to World Health Organization, Cancers figure among the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012. The number of new cases is expected to rise by about 70% over the next 2 decades. Among men, the 5 most common sites of cancer diagnosed in 2012 were lung, prostate, colorectal, stomach, and liver cancer. Among women the 5 most common sites diagnosed were breast, colorectal, lung, cervix, and stomach cancer. And the common type of cancer prevalent amongst both the sexes is lung cancer. Lung cancer is the leading cause of cancer death and the second most common cancer among both men and women in the United States. The data set **North Central Cancer Treatment Group** (NCCTG) Lung Cancer Data describes survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities. Size of the unstructured database is 229 Instances and 10 Variables.

For measuring how the patient can perform usual daily activities, we use Karnofsky Performance Scale Index and ECOG performance score. The Karnofsky Performance Scale Index allows patients to be classified as to their functional impairment. This can be used to compare effectiveness of different therapies and to assess the prognosis in individual patients. The lower the Karnofsky score, the worse the survival for most serious illnesses. The ECOG performance status is a scale used to assess how a patient's disease is progressing, assess how the disease affects the daily living abilities of the patient, and determine appropriate treatment and prognosis.

Grade 0: Fully active, able to carry on all pre-disease performance without restriction

Grade 1: Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work

Grade 2: Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours

Grade 3: Capable of only limited selfcare, confined to bed or chair more than 50% of waking hours

Grade 4: Completely disabled. Cannot carry on any selfcare. Totally confined to bed or chair

Grade 5: Dead

Metadata

URL: https://vincentarelbundock.github.io/Rdatasets/csv/survival/cancer.csv

Source: North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7, 1994.

Category: Healthcare

Demographic Indicator: Censoring status, Age, Sex, ECOG performance score, Karnofsky performance score as rated by physician, Karnofsky performance score as rated by the patient, Meal Calories and Weight Loss

Year: 1994

Dataset Variables

The variables given below are the prospective evaluations of prognostic variables from the patient-completed questionnaires in 1994 by the North Central Cancer Treatment Group.

Number of Variables: 10

Number of Instances: 229

ID	Variable	Variable Description	Data Type
1	Inst	Institution code (1-33, includes NA)	Character
2	Time	Survival time in days	Integer
3	Status	Censoring status 1=censored, 2=dead	Integer
4	Age	Age of the patient in years	Integer
5	Sex	Sex of the patient. Male=1 Female=2	Integer
6	ph.ecog	Eastern Cooperative Oncology Group (ECOG) performance score (0=good 5=dead)	Integer
7	ph.karno	Karnofsky performance score (bad=0 and good=100) rated by physician. Karnofsky Performance Status is used to determine a patient's prognosis and to measure changes in a patient’s ability to function or to determine if a patient could be included in a clinical trial. The lower the Karnofsky score, the worse the survival for most serious illnesses.	Character
8	pat.karno	Karnofsky performance score as rated by the patient.	Character
9	meal.cal	Calories that the patient consumed at meals	Character
10	wt.loss	Weight loss in the last six months	Character

Data is missing or left incomplete by the patient when they had completed the questionnaires. The variables Institution code, ECOG performance score, Karnofsky performance score as rated by physician, Karnofsky performance score as rated by the patient, Meal Calories and Weight Loss have some of the values as “NA” which needs to be cleaned and marked as “0” to make it consistent.

Audience

1) Centers for Disease Control and Prevention (CDC) can study these observations to know the effects and demonstrate to general public the side effects of smoking

2) Non- Profit Organization supporting Quit-smoking campaign can broadcast these observations demonstrate to general public the side effects of smoking

3) Physicians and Researchers working on lung cancer medication, study the data observations to learn the trend of cancer contraction and detection.

4) Prescription Drug seller, this report will visually represent survey data enabling manufacturing and marketing personnel to understand complex statistical data and make informed decisions critical to the success of their business.

5) Lung Cancer Patient, help make informed decisions on further procedures observing the past data and observation of other patients.

Questions

The following report will attempt to answer the following questions:

1. What age group is more affected by lung cancer?
2. What is the weight loss pattern in lung cancer patient based on meals consumed and survival time left?
3. What is the frequency of the censoring status based on the gender?
4. What is the probability of a lung cancer patient’s survival rate based on his age, Karnofsky Performance Scale Index as rated by physician and by patient?
5. What is meal calorie consumption trend amongst the age groups?
6. What is the probability of a lung cancer patient’s survival rate based on his ECOG performance score?
7. What is the probability of a lung cancer patient’s weight loss?
8. What is co-relation of Censoring status of a lung cancer patient and his Karnofsky Performance Scale Index as rated by physician?
9. Do men have greater Karnofsky Performance Scale Index?

Data Cleaning

Data Cleaning or Data Scrubbing is the process of amending or removing incorrect, incomplete, unformatted or duplicate data from a dataset. It is the process of transforming raw data into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability.

Data cleaning may profoundly influence the statistical statements based on the data. Typical actions like imputation or outlier handling obviously influence the results of a statistical analysis. For this reason, data cleaning should be considered a statistical operation, to be performed in a reproducible manner. The R statistical environment provides a good environment for reproducible data cleaning since all cleaning actions can be scripted and therefore reproduced.

We need to remove the errors in the data so that the data analysis process is straightforward. Errors in datasets can cause us to draw wrong conclusions. Hence, this is a very important step prior to the actual data analysis. We would need to identify incomplete or incorrect data in our dataset.

In our dataset “Cancer”, the below data needs to be cleaned:

1. The first variable should be removed from the dataset since it does not contain any useful information.
2. Variables names need to be renamed to make them more understandable.
3. The values in the variable “Sex” should be transformed into more user-friendly values such as “Male” instead of 1 and “Female” instead of 2.
4. The values in the variable “Status” should be modified to censoring status values such as “Censored” instead of 1 and “Dead” instead of 2.

Missing Data Mitigation

A missing value, represented by NA in R, is a placed in the data when the type is known but its value isn't. Therefore, we cannot perform statistical analysis on data where one or more values in the data are missing. We can choose to either omit elements from a dataset that contain missing values or to assign a value, but these missing values need to be dealt with prior to any analysis.

It will be up to the analyst to decide how to handle empty values, since a default assigning may cause unexpected or erroneous results. Another commonly encountered mistake that analysts make is confusing an NA in categorical data with the category unknown.

In our dataset, there are variables such as Institution code, ECOG performance score, Karnofsky performance score as rated by physician, Karnofsky performance score as rated by the patient, Meal Calories and Weight Loss have values as “NA”. These need to either removed or replaced with a more acceptable value such as “0”.

Data Analysis

Descriptive data analysis statistics tell you how your data look, and what the relationships are between the different variables in your data set. Descriptive data analysis statistics are used to present quantitative descriptions in a manageable form. Each descriptive statistic reduces lots of data into a simpler summary. In this dataset, we analyze the gender affected more by cancer. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. With variables that can have a large number of possible values, with relatively few people having each one, we group the raw scores into categories according to ranges of values. We analyze the dataset, to figure out the mean Karnofsky Performance Scale Index amongst the lung cancer patients at different stages of their cancer.

We figure out the mode of the dataset, which is the most frequently occurring value in the set of scores. The most frequently occurring value is the mode. We find out the mode of the age column of the lung cancer patients. This evaluation helps us to determine the age group, which is, affected the most by lung cancer.

The Standard Deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range. We try to find out the standard deviation of survival rate amongst the lung cancer patients. The standard deviation allows us to reach some conclusions about specific survival rate in our distribution. We are assuming that the distribution of survival rate is normal or bell-shaped.

Correlation measures the strength of the relationship between different variables in your data. In this particular dataset, we are correlating Censoring status of a lung cancer patient and his Karnofsky Performance Scale Index as rated by physician. This helps the researchers and the doctors examine and draw conclusion on how well a patient is responding and for predicting his life span. There is a strong relation between Karnofsky Performance Scale Index and the censoring status of a lung cancer patient.

Data Visualization

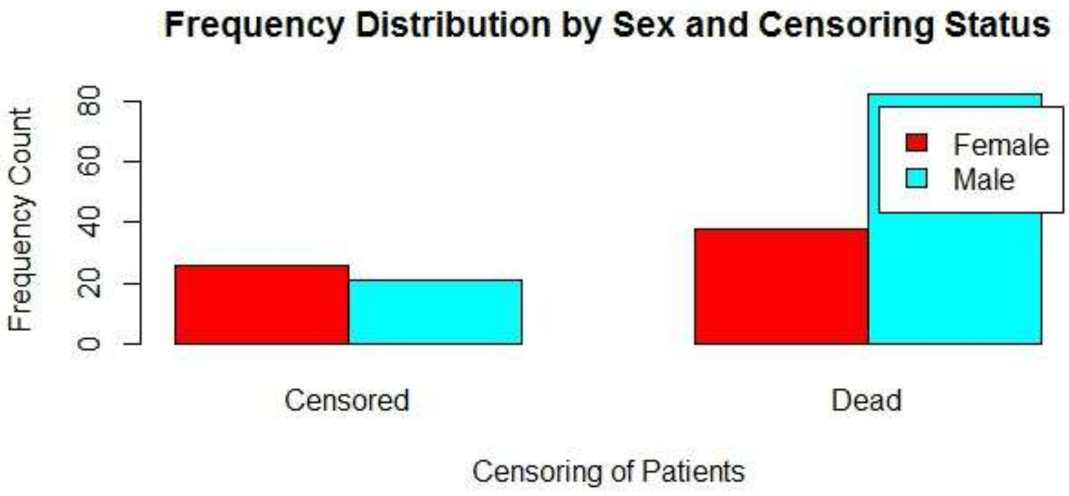
The first visualization that we have created is a frequency distribution bar chart to depict which gender is more affected by cancer. This histogram will depict gender frequency distribution against the two censoring statuses. This data visualization will help the researchers and the patients to learn more about how lung cancer affects a particular gender more.

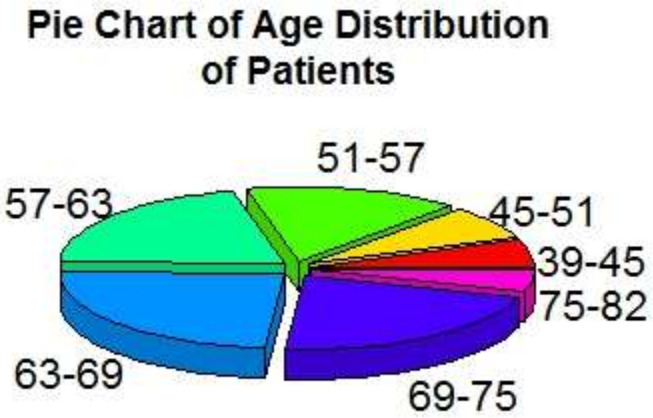
The second visualization that we have created is a 3D pie chart to describe the percentage of lung cancer patients in different age ranges. This visualization will help the researchers to analyze the reasons why a particular age group has a higher percentage of lung cancer occurrences as compared to other age groups.

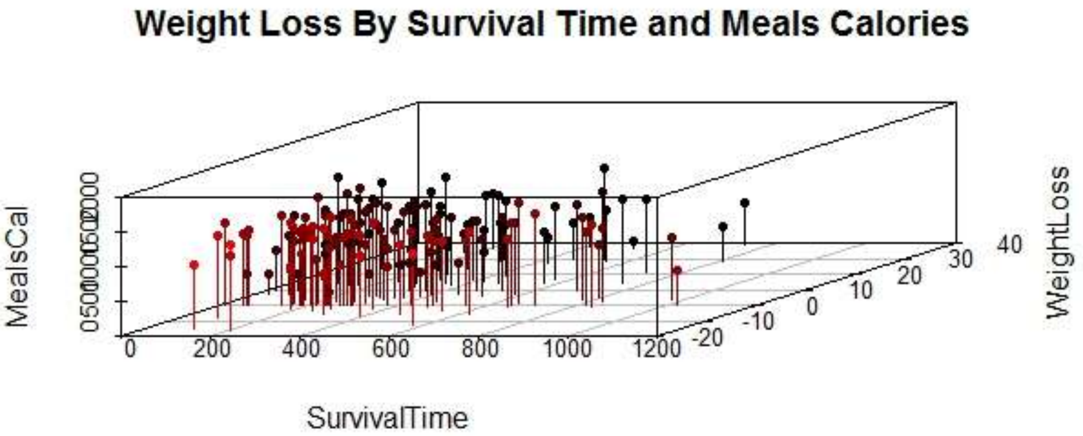
The third visualization that we have created is a 3D scatter plot showing the weight loss based on the calories that the patients consumed at meals and their survival time. This will help to monitor the patient's weight loss as compared to his meal calories intake and survival time left.

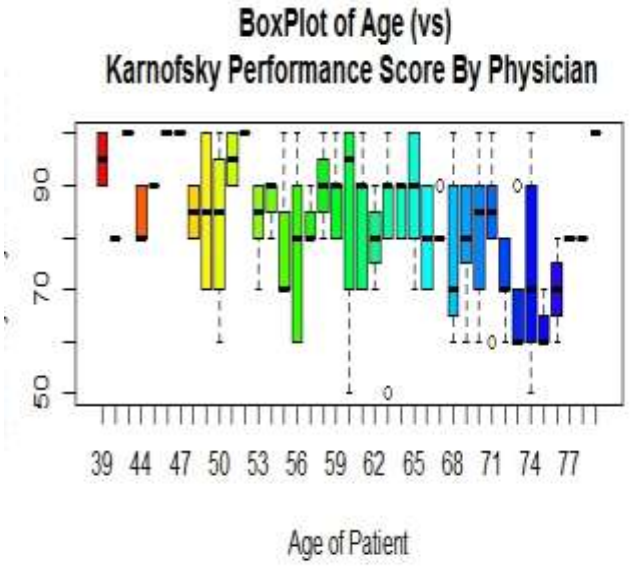
Lastly, we also created a box-plot for the Karnofsky performance score as rated by the patient for the different age groups and a box-plot for the Karnofsky performance score as rated by the physician for the different age groups. We have combined both these box plots and compared them. This will help to show how different age groups respond to the medicines and treatment provided to them and thus will help in further treatment.

Visualization Screenshots

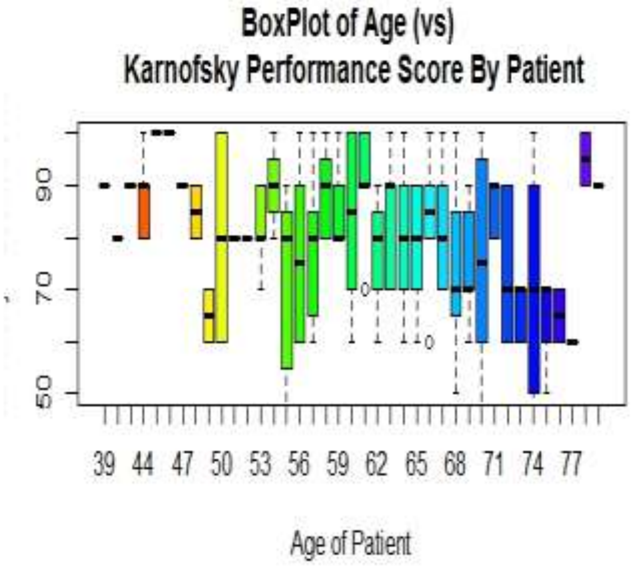



Pie Chart of Age Distribution
of Patients
51-57
57-63
45-51
39-45
75-82
63-69
69-75


Weight Loss By Survival Time and Meals Calories
MealsCal
0500100015002000
0
200
400
600
800
1000
1200
SurvivalTime
-20
-10
0
10
20
30
40
WeightLoss



BoxPlot of Age (vs)
Karnofsky Performance Score By Physician



BoxPlot of Age (vs)
Karnofsky Performance Score By Patient

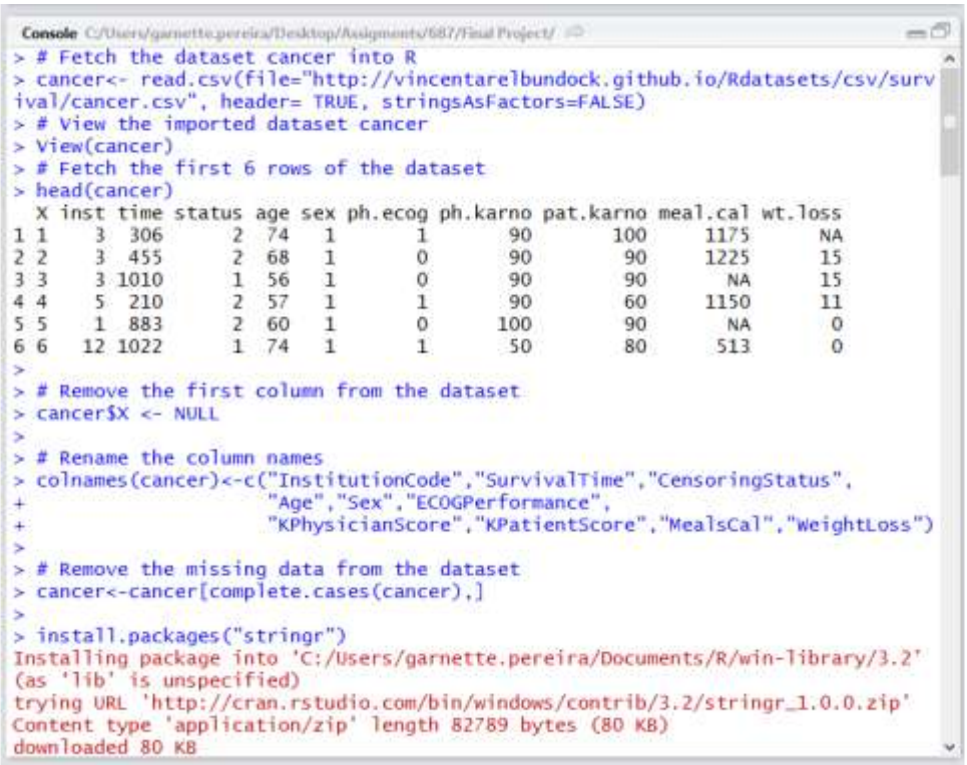
R Script

```
# Fetch the dataset cancer into R
cancer<-
read.csv(file="http://vincentarelbundock.github.io/Rdatasets/csv/survival/cancer.csv",
header= TRUE, stringsAsFactors=FALSE)
# View the imported dataset cancer
View(cancer)
# Fetch the first 6 rows of the dataset
head(cancer)
# Remove the first column from the dataset
cancer$X <- NULL
# Rename the column names
colnames(cancer)<-c("InstitutionCode","SurvivalTime","CensoringStatus",
                "Age","Sex","ECOGPerformance",
                "KPhysicianScore","KPatientScore","MealsCal","WeightLoss")
# Remove the missing data from the dataset
cancer<-cancer[complete.cases(cancer),]
install.packages("stringr")
library(stringr)
# Replace the values of the column Sex
cancer$Sex<- str_replace(cancer$Sex,"1","Male")
cancer$Sex<- str_replace(cancer$Sex,"2","Female")
# Replace the values of the column Censoring Status
cancer$CensoringStatus<- str_replace(cancer$CensoringStatus,"1","Censored")
cancer$CensoringStatus<- str_replace(cancer$CensoringStatus,"2","Dead")
# Fetch the first 6 rows of the dataset
head(cancer)
install.packages("sqldf")
library(sqldf)
```

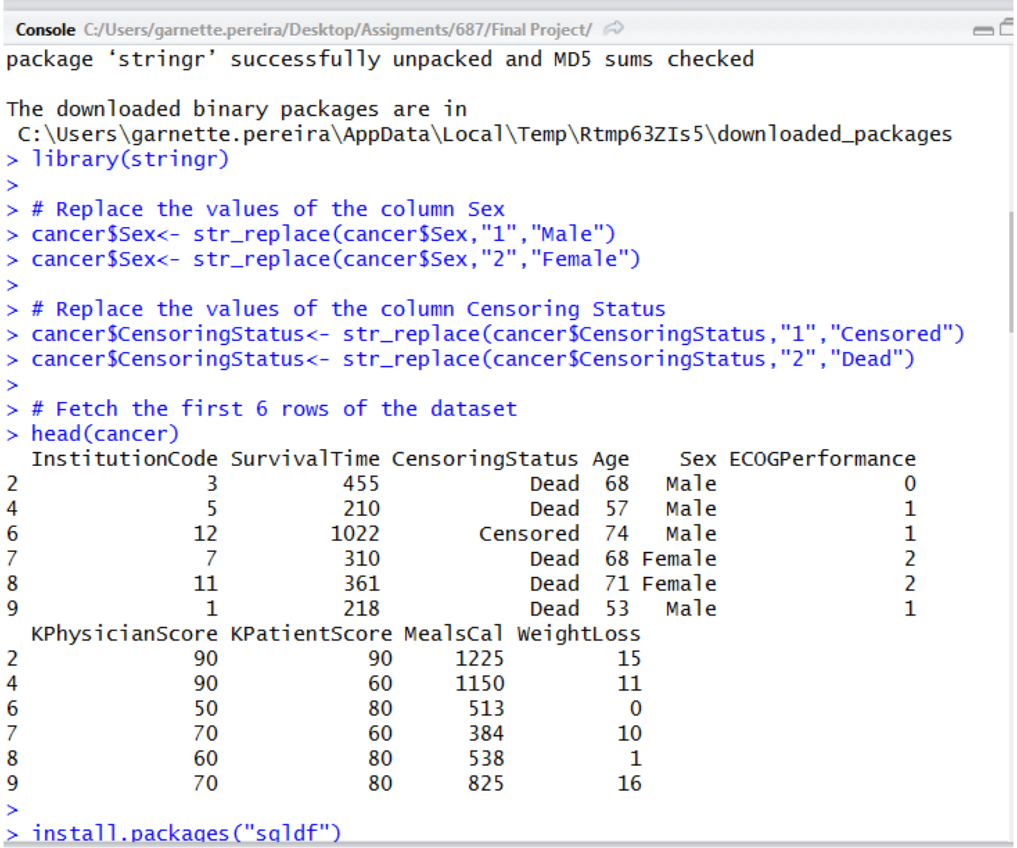
```
install.packages("tcltk2")
library(tcltk2)
# Create a table of counts of the combination of Sex and Censoring Status
counts <- table(cancer$Sex,cancer$CensoringStatus)
# Create a barplot depicting the Frequency Distribution by Sex and Censoring Status
barplot(counts, main="Frequency Distribution by Sex and Censoring Status",
        xlab="Censoring of Patients",ylab="Frequency Count",
        col=rainbow(2),legend = rownames(counts), beside=TRUE)
install.packages("plotrix")
library(plotrix)
# Create Age Groups
cancer$age<- cut(cancer$Age,7)
# Create a data frame of the Age column
age<-as.data.frame(table(cancer$Age))
# Create a label for the age groups
label <- c("39-45", "45-51", "51-57","57-63","63-69","69-75", "75-82")
# Create a 3D Pie Chart showing the Age Distribution of the patients
pie3D(age$Freq,labels=label,explode=0.1, radius=1.5,labelcex =1.3,
      main="Pie Chart of Age Distribution \n of Patients")
install.packages("scatterplot3d")
library(scatterplot3d)
# Attach the dataset cancer
attach(cancer)
# Create a 3D Scatter Plot showing the Weight Loss By Survival Time and Meals
Calories
scatterplot3d(SurvivalTime,WeightLoss,MealsCal, pch=20,type="h",
              highlight.3d=TRUE,ylim=c(-20,40),
              zlim=c(0,2000),main="Weight Loss By Survival Time and Meals Calories")
# Create Age Groups
cancer$age<- cut(cancer$Age,7)
# Create data frame of the Age column
```

```
age<-as.data.frame(table(cancer$age))
# Combine the frequencies of the age groups
agegroup<-age$Freq
# Combine the two box plots arranged in 1 row and 2 columns
par(mfrow=c(1,2))
# Create a box plot of the Age Vs Karnofsky Performance Score By Physician
boxplot(KPhysicianScore~Age,col=rainbow(50),
        main="BoxPlot of Age (vs) \n Karnofsky Performance Score By Physician",ylim=c(50,100),
        xlab="Age of Patient", ylab=" Karnofsky Physician Score")
# Create a box plot of the Age Vs Karnofsky Performance Score By Patient
boxplot(KPatientScore~Age,col=rainbow(50),
        main="BoxPlot of Age (vs) \n Karnofsky Performance Score By Patient",ylim=c(50,100),
        xlab="Age of Patient", ylab=" Karnofsky Patient Score")
```

Console Output Screenshots

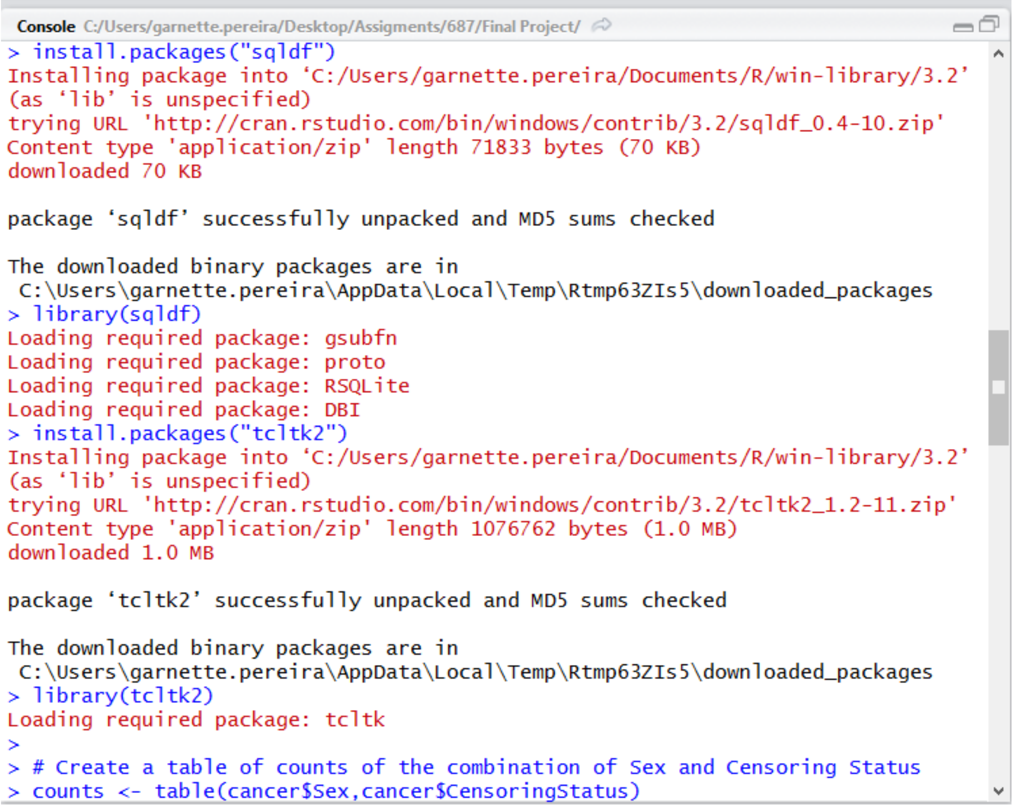

Console C:/Users/garnette.pereira/Desktop/Assigments/687/Final Project/

```
> # Fetch the dataset cancer into R
> cancer<- read.csv(file="http://vincentarelbundock.github.io/Rdatasets/csv/surv
ival/cancer.csv", header= TRUE, stringsAsFactors=FALSE)
> # View the imported dataset cancer
> View(cancer)
> # Fetch the first 6 rows of the dataset
> head(cancer)
  X inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1 1    3  306      2  74   1       1       90       100     1175      NA
2 2    3  455      2  68   1       0       90        90     1225      15
3 3    3 1010      1  56   1       0       90        90       NA      15
4 4    5  210      2  57   1       1       90        60     1150      11
5 5    1  883      2  60   1       0      100        90       NA       0
6 6   12 1022      1  74   1       1       50        80      513       0
>
> # Remove the first column from the dataset
> cancer$X <- NULL
>
> # Rename the column names
> colnames(cancer)<-c("InstitutionCode","SurvivalTime","CensoringStatus",
+                     "Age","Sex","ECOGPerformance",
+                     "KPhysicianScore","KPatientScore","MealsCal","WeightLoss")
>
> # Remove the missing data from the dataset
> cancer<-cancer[complete.cases(cancer),]
>
> install.packages("stringr")
Installing package into 'C:/Users/garnette.pereira/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/stringr_1.0.0.zip'
Content type 'application/zip' length 82789 bytes (80 KB)
downloaded 80 KB
```


Console C:/Users/garnette.pereira/Desktop/Assigments/687/Final Project/

```
package ‘stringr’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\garnette.pereira\AppData\Local\Temp\Rtmp63ZIs5\downloaded_packages
> library(stringr)
>
> # Replace the values of the column Sex
> cancer$Sex<- str_replace(cancer$Sex,"1","Male")
> cancer$Sex<- str_replace(cancer$Sex,"2","Female")
>
> # Replace the values of the column Censoring Status
> cancer$CensoringStatus<- str_replace(cancer$CensoringStatus,"1","Censored")
> cancer$CensoringStatus<- str_replace(cancer$CensoringStatus,"2","Dead")
>
> # Fetch the first 6 rows of the dataset
> head(cancer)
  InstitutionCode SurvivalTime CensoringStatus Age    Sex ECOGPerformance
2               3          455            Dead  68   Male               0
4               5          210            Dead  57   Male               1
6              12         1022        Censored  74   Male               1
7               7          310            Dead  68 Female               2
8              11          361            Dead  71 Female               2
9               1          218            Dead  53   Male               1
  KPhysicianScore KPatientScore MealsCal WeightLoss
2              90            90     1225         15
4              90            60     1150         11
6              50            80      513          0
7              70            60      384         10
8              60            80      538          1
9              70            80      825         16
>
> install.packages("sqldf")
```


Console C:/Users/garnette.pereira/Desktop/Assigments/687/Final Project/

```
> install.packages("sqldf")
Installing package into ‘C:/Users/garnette.pereira/Documents/R/win-library/3.2’
(as ‘lib’ is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/sqldf_0.4-10.zip'
Content type 'application/zip' length 71833 bytes (70 KB)
downloaded 70 KB

package ‘sqldf’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\garnette.pereira\AppData\Local\Temp\Rtmp63ZIs5\downloaded_packages
> library(sqldf)
Loading required package: gsubfn
Loading required package: proto
Loading required package: RSQLite
Loading required package: DBI
> install.packages("tcltk2")
Installing package into ‘C:/Users/garnette.pereira/Documents/R/win-library/3.2’
(as ‘lib’ is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/tcltk2_1.2-11.zip'
Content type 'application/zip' length 1076762 bytes (1.0 MB)
downloaded 1.0 MB

package ‘tcltk2’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\garnette.pereira\AppData\Local\Temp\Rtmp63ZIs5\downloaded_packages
> library(tcltk2)
Loading required package: tcltk
>
> # Create a table of counts of the combination of Sex and Censoring Status
> counts <- table(cancer$Sex,cancer$CensoringStatus)
```

Conclusion

The purpose of this Data Analysis project is successfully completed. As we have visually represented the age groups of the lung cancer patients, the censoring status based on the patient’s gender, the weight loss based on the survival time and calories consumed at meals and finally, the Karnofsky performance score as rated by the patient for the different age groups as compared to Karnofsky performance score as rated by the physician for the different age groups. These visualizations and data analysis will greatly benefit the lung cancer patients and also, for all research purposes for the cure of Lung Cancer.

References

1] World Health Organization “http://www.who.int/mediacentre/factsheets/fs297/en/” n.d retrieved on May 15, 2015

2] MIT Library “http://web.mit.edu/r_v3.0.1/lib/R/library/survival/html/lung.html” n.d retrieved on May 15, 2015

3] The National Cancer Institute “http://ncctg.mayo.edu/ “n.d retrieved on May 15, 2015

4] Radiopaedia. Org “http://radiopaedia.org/articles/ecog-performance-status” n.d retrieved on May 15, 2015

5] National Cancer Institute “http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=44156 “ n.d retrieved on May 15, 2015

6] An introduction to data cleaning with R, Edwin de Jonge and Mark van der Loo