ML lab 1 (R)

Mithesh Ramachandran (J002), B.Tech Data Science 3rd year

07/07/2020

Introduction to Statistics Package Exercises

Objectives:

1. Exploring Variables in a Dataset
2. Learn how to open and examine a dataset.
3. Practice classifying variables by their type: quantitative or categorical.
4. Learn how to handle categorical variables whose values are numerically coded.

Link to experiment: https://upscfever.com/upsc-fever/en/data/en-exercises-1.html (https://upscfever.com/upsc-fever/en/data/en-exercises-1.html)

Submission:

To be submitted on or before 09/07/20.

1. Create github repo and Upload results to repo and paste repo link in common excel sheet
2. Link to excel- https://docs.google.com/spreadsheets/d/1G0HMQkWOuAb6mk89BRauMMhhjga0OaXjTbUC-lwSu8M/edit?usp=sharing (https://docs.google.com/spreadsheets/d/1G0HMQkWOuAb6mk89BRauMMhhjga0OaXjTbUC-lwSu8M/edit?usp=sharing)

Github Repository

User ID: 259mit

https://github.com/259mit/ML (https://github.com/259mit/ML)

Statistics Package Exercise: Exploring Variables in a Dataset

Loading Dataset for questions 1 and 2: depression.csv

1

```
depression<-read.csv("/Users/home/Downloads/OneDrive_1_07-07-2020/depression.csv")
library(knitr)
kable(head(depression), caption = "first five rows of 'depression' dataset")
```

first five rows of ‘depression’ dataset

X	Hospt	Treat	Outcome	Time	AcuteT	Age	Gender
1	1	0	1	36.143	211	33	1
2	1	1	0	105.143	176	49	1
3	1	1	0	74.571	191	50	1
4	1	0	1	49.714	206	29	2
5	1	0	0	14.429	63	29	1
6	1	2	1	5.000	70	30	2

Question 1: What are the categorical variables in this dataset?

From the above image, we can see that ‘Hospt’ , ‘Treat’ , ‘Outcome’ and ‘Gender’ are categorical variables, i.e have outcomes in levels.

1. Hostp because the numbers represent codes, which are used to identify individual hospitals and place them into categories.
2. Treat because the treatment received by the patients is in the form of categories (Lithium, Imipramine, or Placebo)
3. Outcome since recurrence is in the form of two categories (Recurrence or No Recurrence)
4. Gender because the numbers represent two distinct categories: Female and Male.

Question 2: What are the quantitative variables in this dataset?

From the above image, we can see that ‘Time’ , ‘AcuteT’ and ‘Age’ are quantitative variables, i.e have outcomes in datatype int or numeric (continous)

1. Time since it can take on multiple numerical values, which have arithmetic meaning (i.e., it makes sense to add, subtract, multiply, divide, or compare the magnitude of such values)
2. Age since it can take on multiple numerical values, which represent a characteristic of the patient
3. AcuteT because it can take on multiple numerical values to represent a characteristic of the patient.

Statistics Package Exercise 2 : Tallying Data and Creating Pie Charts

Loading Dataset for question 3: friends.csv

```
friends<-read.csv("/Users/home/Downloads/OneDrive_1_07-07-2020/friends.csv")
kable(head(friends), caption = "first five rows of 'friends' dataset")
```

first five rows of 'friends' dataset

	X	Friends
	1	No difference
	2	No difference
	3	No difference
	4	No difference
	5	No difference
	6	No difference

2

```
# Identifying target variable:
names(friends)
```

```
## [1] "X"       "Friends"
```

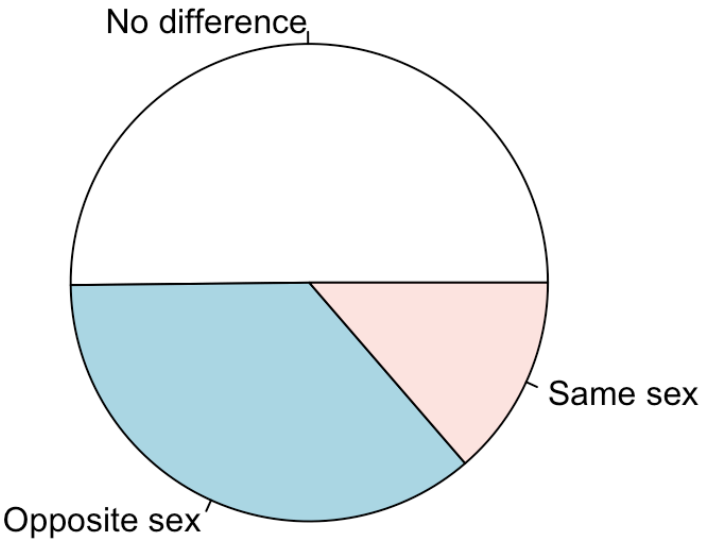
```
#Creating a pie (proportion) chart
tf<-table(friends$Friends)
#Displaying the proportion
tfp<-prop.table(tf)
head(tfp)
```

```
##
## No difference  Opposite sex      Same sex
##     0.5016667     0.3616667     0.1366667
```

```
#Displaying proportion as percentage
ptfp<-tfp*100
head(ptfp)
```

```
##
## No difference  Opposite sex      Same sex
##      50.16667      36.16667      13.66667
```

```
#plot a pie chart
pie(tf)
```



Question 3: Describe the distribution of the variable “friends” in dataset - Survey that asked 1,200 U.S. college students about their body perception

The students are NOT divided equally among the three categories. About 50% of the students find it as easy to make friends with the opposite sex as with the same sex. Among the remaining 50% of the students, the majority (36.2%) find it easier to make friends with people of the opposite sex, and the remainder (13.7%) find it easier to make friends with people of their own sex.

Statistics Package Exercise 3: Creating and Describing Histograms

Loading Dataset for question 4 to 6: actor_age.csv

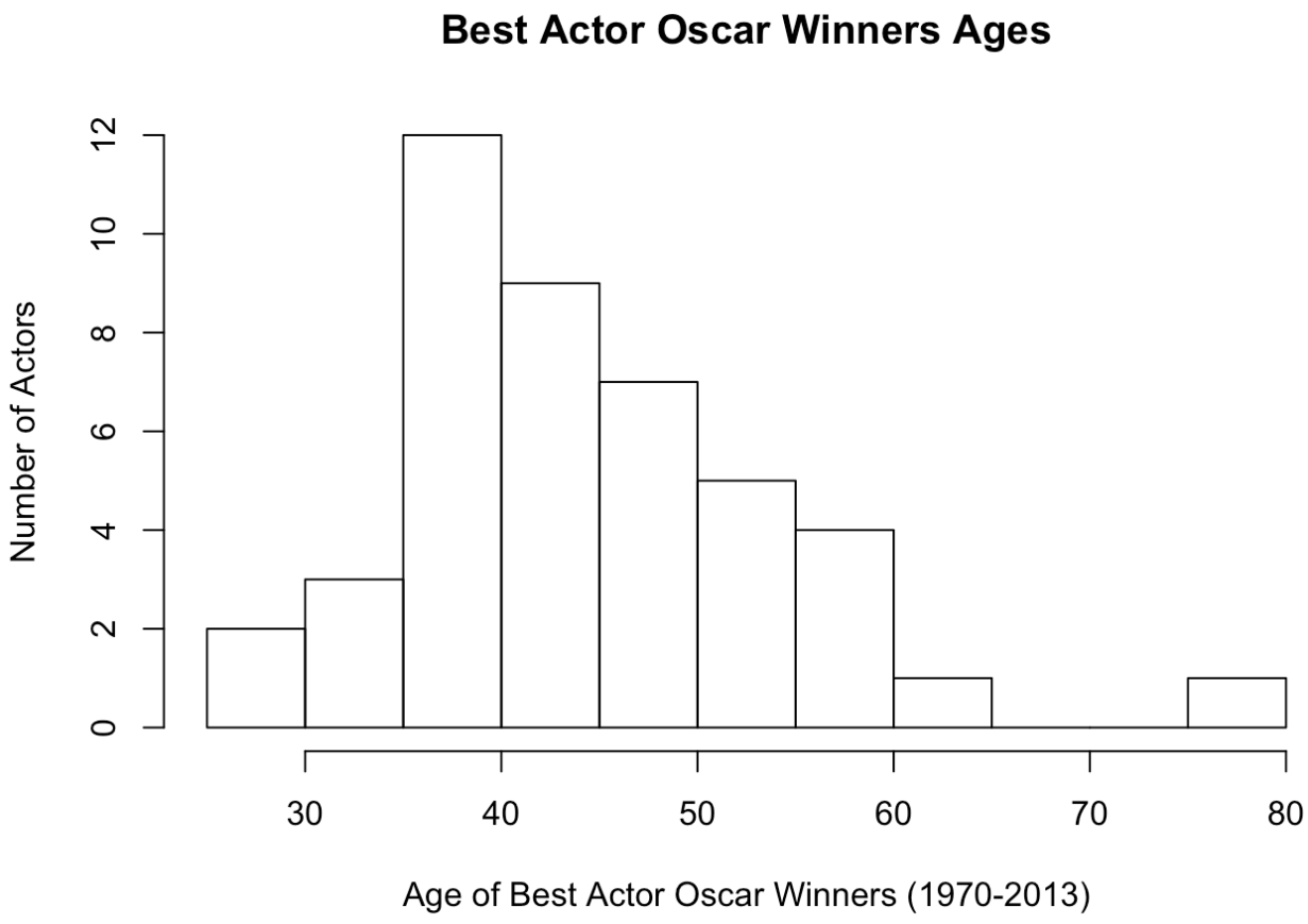
```
actors<-read.csv("/Users/home/Downloads/OneDrive_1_07-07-2020/actor_age.csv")
kable(head(actors), caption = "first five rows of 'actor_age' dataset")
```

first five rows of ‘actor_age’ dataset

X	Age
1	43
2	42
3	48
4	49
5	56
6	38

3

```
# create histogram of the actor's age
hist(actors$Age, xlab="Age of Best Actor Oscar Winners (1970-2013)", ylab="Number
of Actors", main="Best Actor Oscar Winners Ages")
```



Question 4: Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers (Dataset - Best

Actor Oscar winners (1970-2013))

1. Shape: the distribution is skewed right. This means that most actors receive the best acting Oscar at a relatively younger age (before age 48), and fewer at an older age.
2. Center: The distribution seems to be centered at around 42-43. This means that about half the actors are 42 or younger when they receive the Oscar, and about half are older.
3. Spread: The age distribution ranges from about 30 to about 75. The entire dataset is covered, then, by a range of 45 years. It should be noted, though, that there is one high outlier at around age 75, and the rest of the data ranges only from 30 to 60.
4. Outliers: As mentioned above, there is one high outlier at around age 75.

4

```
#For mean, median, 1st and 3rd quartile, min and max
summary(actors$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   38.00   43.50   44.98   50.25   76.00
```

```
#for IQR
IQR(actors$Age)
```

```
## [1] 12.25
```

Question 5: Getting information from the output:

A. How many observations are in this data set?

B. What is the mean age of the actors who won the Oscar?

C. What is the five-number summary of the distribution? (Dataset - Best Actor Oscar winners (1970-2013))

A. There are n = 44 observations in the data set (representing the age of the Best Actor Oscar winners of the 44 years from 1970 through 2013).

B. Mean = 44.98

C. The five-number summary is: min = 29, Q1 = 38, M = 43.5, Q3 = 50.5, Max = 76

Question 6: Get information from the five-number summary:

A. Half of the actors won the Oscar before what age?

B. What is the range covered by all the actors' ages?

C. What is the range covered by the middle 50% of the ages? (Dataset - Best Actor Oscar winners (1970-2013))

A. Half the actors won the Oscar before age 43.5 (the median).

B. The range covered by all the ages is: Range = Max - min = 76 - 29 = 47.

C. The range covered by the middle 50% of the ages is: IQR = Q3 - Q1 = 50.5 - 38 = 12.5

Statistics Package Exercise 4: Creating Side-by-Side Boxplots

Loading Dataset: grad_data.csv

```
grad<-read.csv("/Users/home/Downloads/OneDrive_1_07-07-2020/grad_data.csv")
kable(head(grad), caption = "first five rows of 'grad_data' dataset")
```

first five rows of 'grad_data' dataset

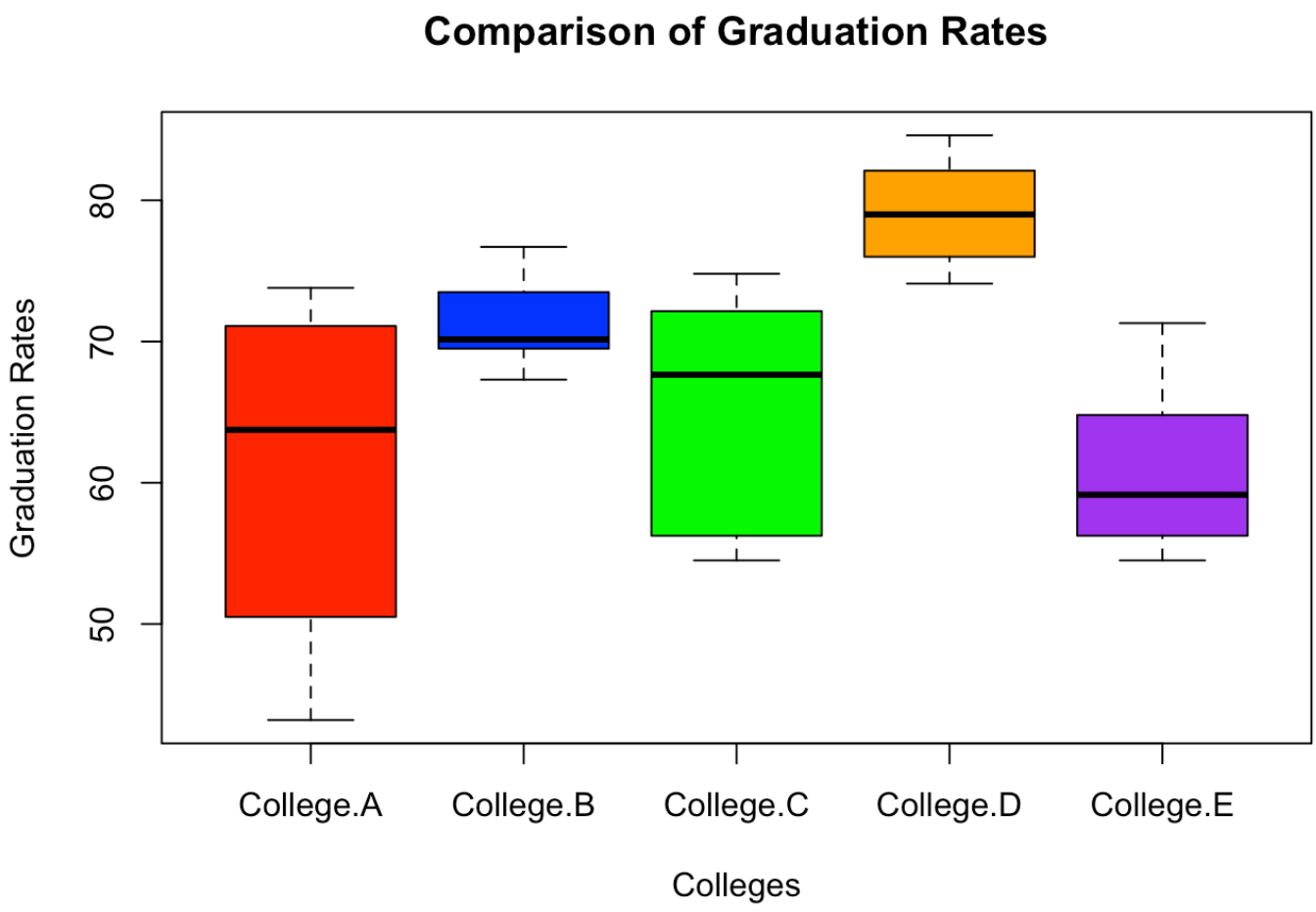
X	College.A	College.B	College.C	College.D	College.E	College.F
1	57.6	70.1	54.5	80.1	71.3	68.8
2	43.2	69.6	55.6	77.3	62.6	61.0
3	49.6	67.3	56.9	74.7	54.5	57.7
4	51.4	76.7	71.0	79.2	57.5	66.4
5	69.9	69.4	73.3	84.6	55.0	75.2
6	69.9	72.6	74.8	78.8	67.0	87.4

5

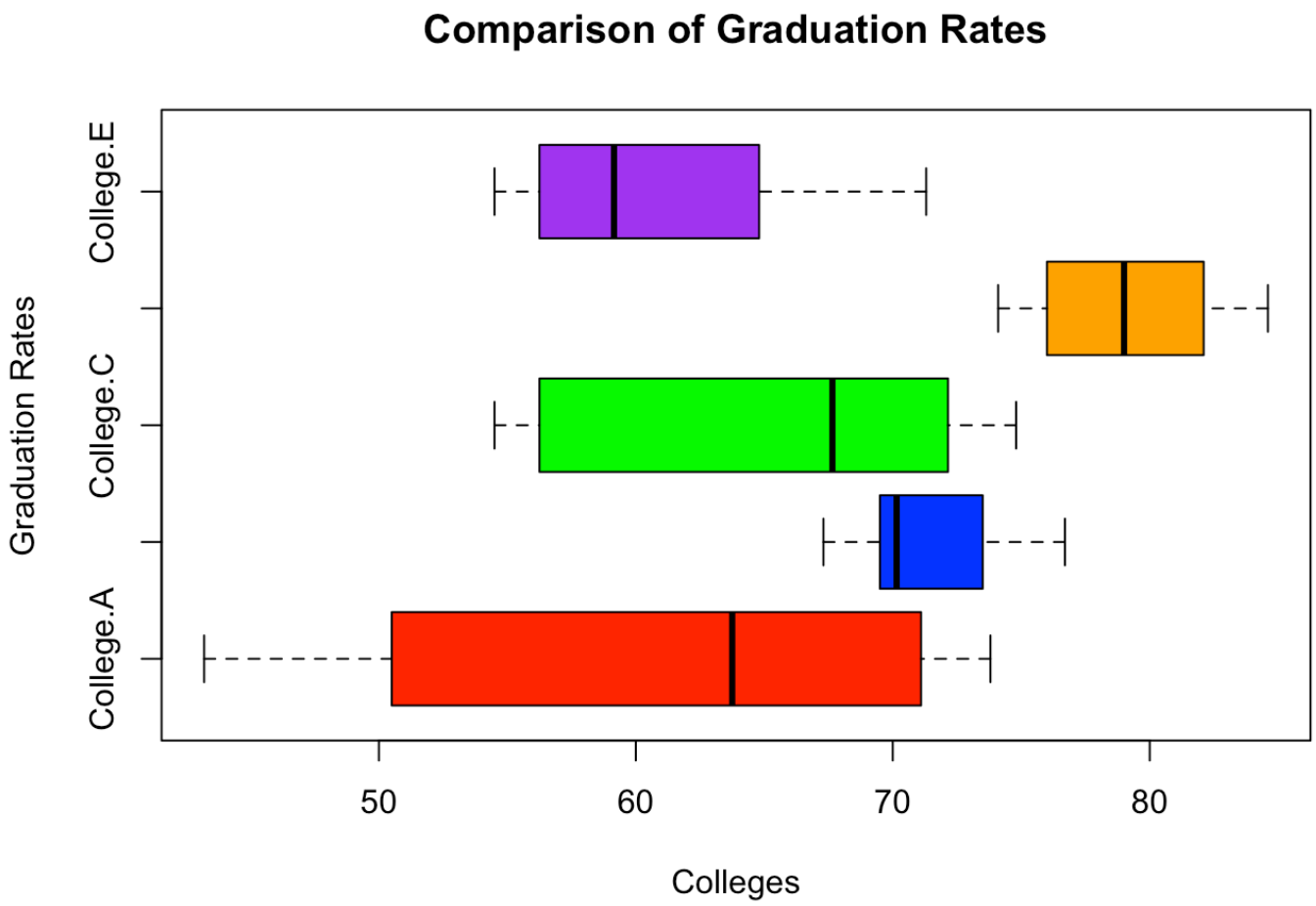
```
# summary table of grad_data
summary(grad)
```

```
##        X          College.A       College.B       College.C       College.D
##  Min.   :1.00   Min.   :43.20   Min.   :67.30   Min.   :54.50   Min.   :74.10
##  1st Qu.:2.75   1st Qu.:50.95   1st Qu.:69.55   1st Qu.:56.58   1st Qu.:76.65
##  Median :4.50   Median :63.75   Median :70.15   Median :67.65   Median :79.00
##  Mean   :4.50   Mean   :60.96   Mean   :71.29   Mean   :65.17   Mean   :79.11
##  3rd Qu.:6.25   3rd Qu.:70.50   3rd Qu.:73.05   3rd Qu.:71.58   3rd Qu.:81.10
##  Max.   :8.00   Max.   :73.80   Max.   :76.70   Max.   :74.80   Max.   :84.60
##    College.E       College.F
##  Min.   :54.50   Min.   :57.70
##  1st Qu.:56.88   1st Qu.:65.05
##  Median :59.15   Median :72.00
##  Mean   :60.77   Mean   :72.78
##  3rd Qu.:63.70   3rd Qu.:81.28
##  Max.   :71.30   Max.   :87.40
```

```
#boxplot of grad_data
boxplot(grad[,2:6], col = c("red", "blue", "green", "orange", "purple"), xlab="Col
leges",ylab ="Graduation Rates", main="Comparison of Graduation Rates")
legend(95,1, legend = c("College A","College B","College C", "College D", "College
E", "College F"))
```



```
#Horizontal
#boxplot of grad_data
boxplot(grad[,2:6], col = c("red", "blue", "green", "orange", "purple"), xlab="Col
leges",ylab ="Graduation Rates", main="Comparison of Graduation Rates", horizontal
= TRUE)
legend(95,1, legend = c("College A","College B","College C", "College D", "College
E", "College F"))
```



Statistics Package Exercise 5:Calculating the Standard Deviation

Loading Dataset for question 7 and 8: ratings.csv

```
ratings<-read.csv("/Users/home/Downloads/OneDrive_1_07-07-2020/ratings.csv")
kable(head(ratings), caption = "first five rows of 'ratings' dataset")
```

first five rows of ‘ratings’ dataset

X	Class.I	Class.II	Class.III
1	1	1	1
2	1	1	2
3	5	1	3
4	5	1	4
5	5	1	5
6	5	1	6

6

```
#find standard deviation of all columns in dataframe 'ratings'
kable(as.table(sapply(ratings[,2:4], sd)), col.names = c("Variable", "Std. dev"))
```

Variable	Std. dev
Class.I	1.568929
Class.II	4.000000
Class.III	2.631174

Question 7: What are the standard deviations of the three rating distributions? Was your intuition correct? (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

Here are the three standard deviations:

Class I: 1.569

Class II: 4.000

Class III: 2.631

Note that through this example, we also learn that the number of distinct values represented in a histogram does not necessarily indicate greater variability.

Question 8: Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a “typical” or “average” distance between the data points and their mean. Judging from the table and the histograms, which class would have the largest standard deviation, and which one would have the smallest standard deviation? Explain your reasoning (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

In class I, almost all the ratings are 5, which is also the mean. The average distance between the observations and the mean, then, would be very small.

In class II most of the observations are far from the mean (at 1 or 9). The average distance between the observations and the mean in this case would be larger.

Class III is the case where some of the observations are close to the mean, and some are far, so the average distance between the observations and the mean would be somewhere in between class I and II.

Ranking of Std. Dev from smallest to largest -
Class I < Class III < Class II

End of assignment. Return to top.

By Mithesh R, J002. ML