Joel T. Jayme
Prof.Poliquit
Class Group 3-MW (1:30-3:00)

BSCS-2
Spss-Activity
Group 6

Data Processing 101
Problem Set 1

Provide an appropriate response. (2 marks each)

1. A meteorologist constructs a graph showing the total precipitation in Phoenix, Arizona in each of the months of 1998. Does this involve descriptive statistics or inferential statistics?
   -Descriptive Statistics
2. Thirty of the 198 students enrolled in PD 101 were asked if they wanted Exam II to be a take-home or an in-class assessment. Twenty, or about 67%, of the students polled, indicated a preference for an in-class exam. The professor concluded that the majority of students in DP 101 would prefer an in-class examination for the second assessment. Did the professor perform a descriptive study or an inferential study?
   -Inferential Study
3. A magazine publisher mails a survey to every subscriber asking about the timeliness of its subscription service. The publisher finds that only 4% of the subscribers responded. This 4% represents what?
   -Sample

4. A magazine publisher mails a survey to every subscriber asking about the quality of its subscription service. The total number of subscribers represents what?
   -Population

Classify the data as either discrete or continuous. (2marks each)

5. The number of freshmen entering college in a certain year is 621.
   -Discrete
6. An athlete runs 100 meters in 10.7 seconds.
   -Continuous

Tables and Graphs (20 marks)

7. The preschool children at Elmwood Elementary School were asked to name their favourite colour. The results are listed below.

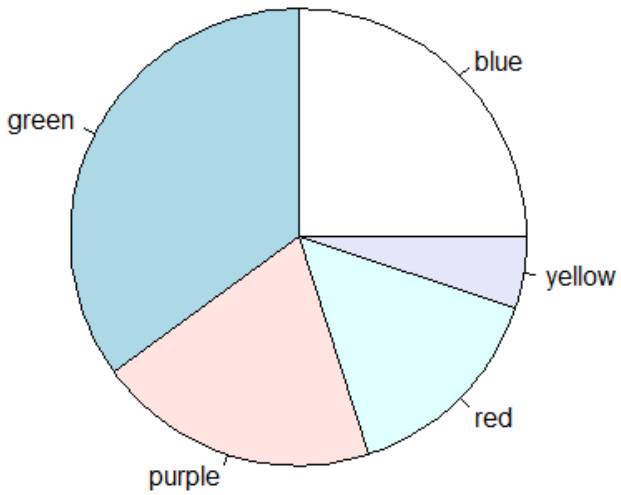
red	red	purple	blue	green
green	green	red	green	purple
green	purple	blue	blue	blue
purple	green	blue	green	yellow

a. Construct a frequency distribution, a pie graph, and a relative frequency distribution.

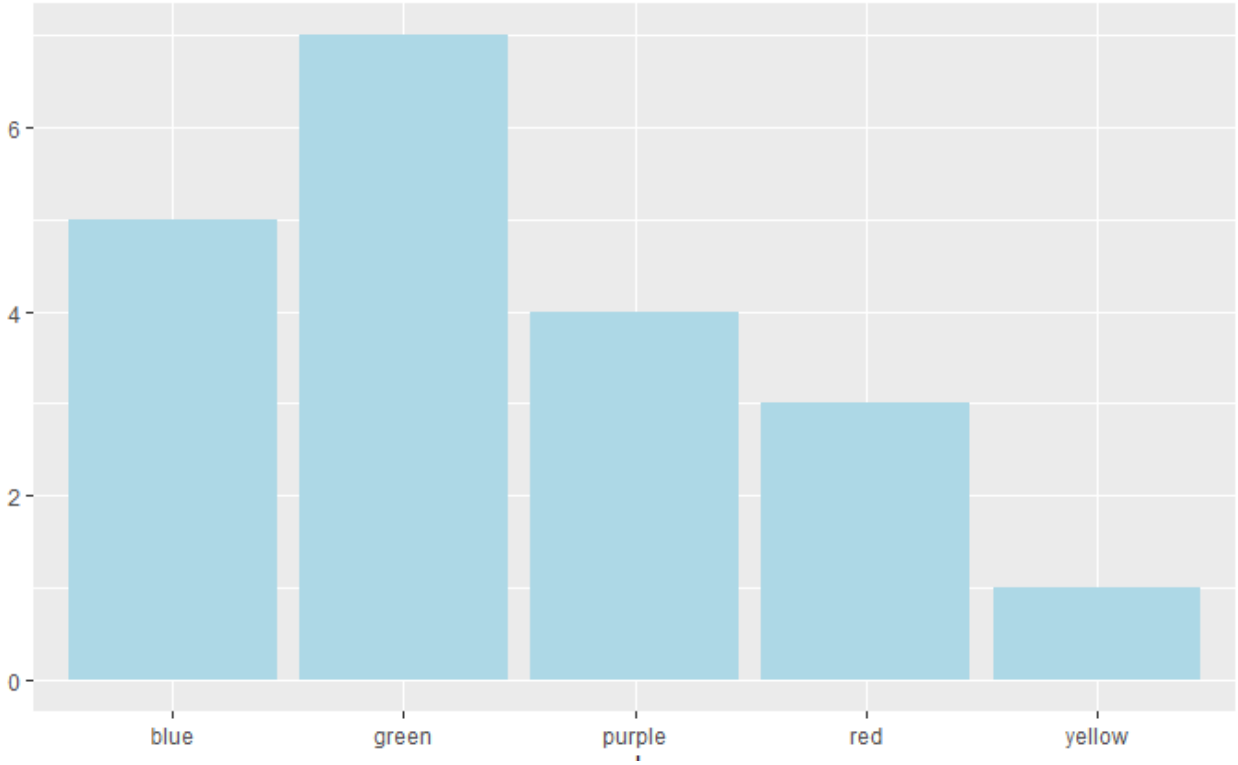
Frequency Distribution Table

```
> table(Color)
Color
  blue  green purple    red yellow
     5      7      4      3      1
```

Pie Chart



Histogram chart



b. What measure of central tendency is appropriate for the given data? Explain your answer.
-Mode, because we are finding the most frequently chosen colour.

8. A medical research team studied the ages of patients who had strokes caused by stress. The ages of 34 patients who suffered stress strokes were as follows.

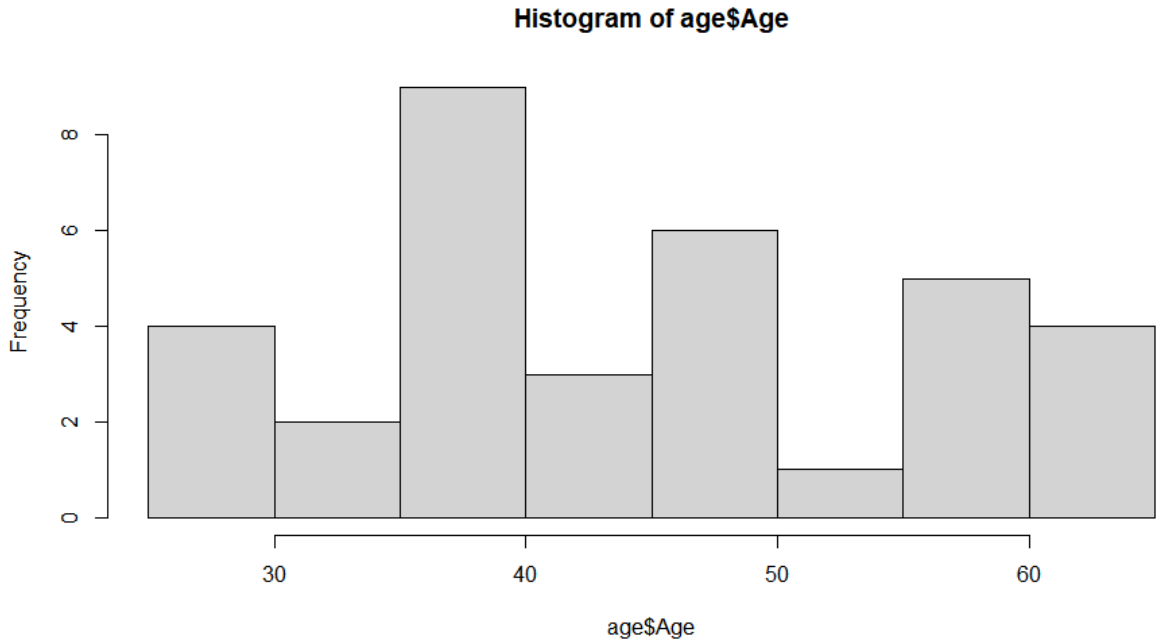
29 30 36 41 45 50 57 61 28 50 36 58
60 38 36 47 40 32 58 46 61 40 55 32
61 56 45 46 62 36 38 40 50 27

c. For the given data compute the following descriptive measures: mean, median, mode, variance, standard deviation, range, first quartile, third quartile, and IQR. (10 marks)

```
> mean(age $Age)
[1] 44.91176
> median(age $Age)
[1] 45
> mfv(age $Age)
[1] 36
> sd(age $Age)
[1] 11.0244
> range(age $Age)
[1] 27 62
> var(age $Age)
[1] 121.5374
>
> IQR(age $Age)
[1] 19.75
> quantile(age $Age)
   0%   25%   50%   75%  100%
27.00 36.00 45.00 55.75 62.00
```

Note: MFV is a feature of modeest package it is for finding the mode. Joel

d. Construct the following graphs for the data: histogram, frequency polygon, stem-and-leaf plot, and boxplot and describe the shape of the distribution. (20 marks)

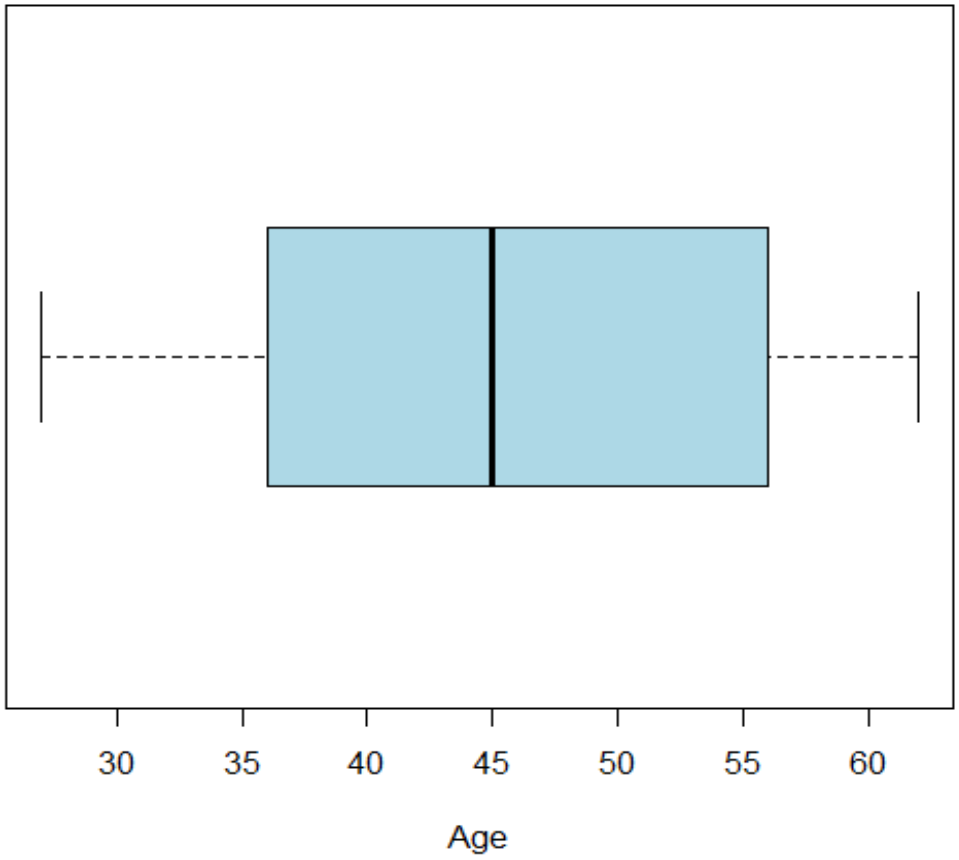


Age Stem-and-Leaf Plot

```
> stem(age $Age)

  The decimal point is 1 digit(s) to the right of the |

  2 | 789
  3 | 022
  3 | 666688
  4 | 0001
  4 | 55667
  5 | 000
  5 | 56788
  6 | 01112
```



Describe the shape of the distribution: According to the Data presented the most common age of stress related stroke is at the ages of *36, 38, 40, and 50.* The data is slightly right skewed.

DP 101 Problem Set 2

Use the **5-steps** in conducting the hypothesis testing for each of the following:

Note: note if P ≥ conF - failed to reject null, if P ≤ conF - reject null (Joel)

40 Marks

1. Laura Naples, manager of Heritage Inn, periodically collects and tabulates information about a sample of the hotel's overnight guests. This information aids her in pricing and scheduling decisions she must make. The table below lists data on 10 randomly selected hotel registrants, collected as the registrants checked out. The data listed are:

   - Number of people in the group
   - Hotel's shuttle service used: yes or no
   - Total telephone charges incurred
   - Reason for stay: business or personal

Name of Registrant	Number in Group	Shuttle Used	Telephone Charges ($)	Reason for Stay
Madam Sandler	1	yes	0.00	personal
Michelle Pepper	2	no	8.46	business
Claudia Shepler	1	no	3.20	business
Annette Rodriquez	2	no	2.90	business
Tony DiMarco	1	yes	3.12	personal
Amy Franklin	3	yes	4.65	business
Julio Roberts	2	no	6.35	personal
Edward Blackstone	4	yes	2.10	personal
Sara Goldman	1	no	1.85	business
Todd Atherton	1	no	5.80	business

   a. Before cell telephones became so common, the average telephone charge per registered group was at least $5.00. Laura suspects that the average has dropped. Test H_0: □ ≥ 5 and H_a: □ < 5 using a .05 level of significance. Use both the critical value and *p*-value approaches to hypothesis testing

```
> t.test(Charge $telephone_charge, mu = 5, alt = "greater", conf = 0.05)

        One Sample t-test

data:  Charge$telephone_charge
t = -1.4709, df = 9, p-value = 0.9123
alternative hypothesis: true mean is greater than 5
5 percent confidence interval:
 5.284932      Inf
sample estimates:
mean of x
    3.843
```

1. H_0: □ ≥ 5 and H_a: □ < 5
2. Level of significance: 0.05
3. ->z-test->Critical value = -1.96
4. P value = 0.175; p value > Level of significance
5.Failed to reject Null Hypothesis; Laura is correct with her suspicion telephone charges has risen more than $5.

b. In the past, Laura has made some important managerial decisions based on the assumption that the average number of people in a registered group is 2.5. Now she wonders if the assumption is still valid. Test the assumption with □ = .05 and use both the critical value and *p*-value approaches.

```
> t.test(Charge $Number.of.people, mu = 2.5, alt = "greater", conf = 0.05)

        One Sample t-test

data:  Charge$Number.of.people
t = -2.1433, df = 9, p-value = 0.9697
alternative hypothesis: true mean is greater than 2.5
5 percent confidence interval:
 2.398692      Inf
sample estimates:
mean of x
      1.8
```

1. H0 = □ ≥2.5, Ha =□ < 2.5
2. Level of significance: 0.05
3. ->z-test->Critical value = -1.96
4. P value= 0.9697
5. Failed to reject Null hypothesis; Laura is wrong with her assumption of 2.5.

2. A survey was recently conducted to determine if consumers spend more on computer-related purchases via the Internet or store visits. Assume a random sample of eight respondents provided the following data on their computer-related purchases during a 30-day period. Using a .05 level of significance, can we conclude that consumers spend more on computer-related purchases by way of the Internet than by visiting stores?

	Expenditures (dollars)	
Respondent	**In-Store**	**Internet**
1	132	225
2	90	24
3	119	95
4	16	55
5	85	13
6	248	105
7	64	57
8	49	0

```
> t.test(expendenture $instore, expendenture $internet, mu = 30 , conf.level = 0.05, var.equal = FALSE, alternative = "two.sided")

        Welch Two Sample t-test

data:  expendenture$instore and expendenture$internet
t = -0.003518, df = 13.993, p-value = 0.9972
alternative hypothesis: true difference in means is not equal to 30
5 percent confidence interval:
 27.60666 32.14334
sample estimates:
mean of x mean of y
  100.375    70.500

>
```

1. H0 = within the 30 day period there is no significant difference, Ha = within the 30 day period there is a significant difference

2. Level of significance: 0.05
3. ->z-test->Critical value = -1.96
4. P value= 0.9972

5. Failed to reject null hypothesis; within 30 days consumers purchase more In-store

3. Individuals were randomly assigned to three different production processes. The hourly units of production for the three processes are shown below.

Production Process		
Process 1	Process 2	Process 3
33	33	28
30	35	36
28	30	30
29	38	34

Use Excel with α = .05 to determine whether there is a significant difference in the mean hourly units of production for the three types of production processes.

```
> summary(one.way_Process)
            Df Sum Sq Mean Sq F value Pr(>F)
process      2     32  16.000   1.636  0.248
Residuals    9     88   9.778
> |
```

//The documentation of how I got the data same goes to #4//

```
Prod <- read.csv(file.choose(), header = TRUE )

//naming
process <- c(rep('process1',4),rep('process2',4),rep('process3',4))

//the rest of the integer data
limit <- c(Prod$Process.1,Prod$Process.2,Prod$Process.3)

df <- data.frame(process,limit)//creates data fream

boxplot(limit ~ process, df)//visualization of the variables

one.way_Process<-aov(limit ~ process, data = df)

summary(one.way_Process)
```

1. H0 = there is no significant difference in an hourly units of production, Ha = there is a significant difference in an hourly units of production

2. Level of significance: 0.05
3. ->z-test->Critical value = -1.96
4. P value= 0.248
//note if P > conF failed to reject null, if P < conF reject null

5. Failed to reject Null hypothesis; there is no significant difference in an hourly units of production.

4. A research organization wishes to determine whether four brands of batteries for transistor radios perform equally well. Three batteries of each type were randomly selected and installed in the three test radios. The number of hours of use for each battery is given below.

	Brand			
Radio	1	2	3	4
A	25	27	20	28
B	29	38	24	37
C	21	28	16	19

Consider the three different test radios and use Excel to carry out the analysis of variance procedure for a randomized block design. Use a .05 level of significance.

```
> summary(one.way_Radio)
            Df Sum Sq Mean Sq F value Pr(>F)
radio        2  244.7   122.3   4.893 0.0365 *
Residuals    9  225.0    25.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

```
1  res <- read.csv(file.choose(), header = TRUE)
2
3  radio <- c(rep('A',4),rep('B',4),rep('C',4))
4  battery <-c(rep(res $A),rep(res $B),rep(res $C))
5
6  df <- data.frame(radio,battery)
7
8  boxplot(battery ~ radio,df,horizontal = TRUE)
9
10 one.way_Radio <- aov(battery ~ radio, data = df)
11
12 summary(one.way_Radio)
13
14
15
```

1. 1. H0 = there is no significant difference in the battery used for the radio, Ha = there is a significant difference in the battery used for the radio.
2. Level of significance: 0.05
3. ->z-test->Critical value = -1.96
4. P value= 0.0365

//note if P > conF failed to reject null, if P < conF reject null

5. Reject Null hypothesis; there is a significant difference in battery used for the radio.