IST 3420 Exam 2 Sample

Fall 2017

Note: This sample shows the types of questions that will be used for Exam 2. In this sample, only 16 multiple choice and 11 short answer questions are shown as examples. The actual questions as well as the number of questions in Exam 2 will be different.

Part 1: Multiple Selection

Each question only has one best choice. Write your choice in the blank after each question.

Questions 1 through 2 are based on the following data frame:

```
str(df)
'data.frame':    10886 obs. of  10 variables:
 $ hour      : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ holiday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ workingday: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ weather   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 2 1 1 1 1 ...
 $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
 $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
 $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
 $ windspeed : num  0 0 0 0 0 ...
 $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
```

1. Which of the following tabular methods is NOT appropriate to summarize the variable “holiday”? ______________.
   A) Frequency Distribution
   B) Relative Frequency Distribution
   C) Percent Frequency Distribution
   D) Cumulative Frequency Distribution

2. Which of the following tabular methods is NOT appropriate to visualize the variable “temp”? _____________.

   A) Pie Chart
   B) Dot Plot
   C) Box Plot
   D) Density Plot

3. What is the result of the following statement? ______________

```
" abc 123 " %>% trimws %>% length
```

   A) 6
   B) 7
   C) 9
   D) 1

4. What is the result of the following statement? ______________

```
sub("[[:digit:]]","", "a1b2c3")
```

   A) A string “abc”
   B) A string “ab2c3”
   C) A string “a123”
   D) A vector with 3 elements including “a1”, “b2”, “c3”

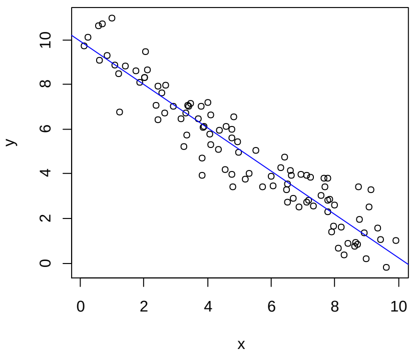
5. Which of the following statement of data manipulation is NOT true? _________.

   A) We can use $ sign to create a new variable in a dataset
   B) We can use dplyr::mutate() function to create a new variable in a dataset
   C) We can use transform() function to create a new variable in a dataset
   D) We can use select() function to create a new variable in a dataset Veracity

6. Suppose we have a string object s1 with value as "gsub$uses$regular$expressions". Which of the following statement can replace all “$” symbols in s1 as “.” ? ________.

   A) s1 <- gsub(pattern = "$", replacement = ".", s1)
   B) s1 <- gsub(pattern = "\\$", replacement = ".", s1)
   C) s1 <- sub(pattern = "$", replacement = ".", s1)
   D) s1 <- sub(pattern = "\\$", replacement = ".", s1)

7. What is the general pattern of relationship between two variables as displayed in the following graph? _______________



A) Positive linear relationship
B) Negative linear relationship
C) Non-linear relationship
D) No relationship

8. Which of the following is not a basic element in a spatial dataset? ________________
A) Longitude
B) Latitude
C) Sequence of coordinate
D) Temperature

9. Which of the following would NOT allow you to calculate a correlation? __________.
A) A negative relationship between X and Y
B) A positive relationship between X and Y
C) A curvilinear relationship between X and Y
D) A linear relationship between X and Y

10. Suppose we have a vector x containing missing values. Which of the following R statements employs the simple imputation method of replacing by mean to deal with the missing data? ______________.
A) x <- na.omit(x)
B) x <- mean(x, na.rm = TRUE)
C) x <- x %>% dplyr::recode(.missing = mean(., na.rm=TRUE))
D) None of them

11. Which one of these statistics is affected by outliers? ______________.
A) Mean
B) Standard deviation
C) Quantile
D) All of them

12. Which of the following statement about outlier is NOT correct? _____________.

    A) In boxplot rule, values beyond upper limit or lower limit are outliers
    B) Z-score method treats any value whose z-score is beyond the range of -3 and 3 as an outlier
    C) LOF (local outlier factor) algorithm will detect any value whose density is similar to its neighbors as an outlier
    D) Different methods may detect different outliers

13. The null and alternative hypotheses divide all possibilities into ______________.

    A) Two sets that overlap
    B) Two non-overlapping sets
    C) Two sets that may or may not overlap
    D) As many sets as necessary to cover all possibilities

14. Which conclusion we can get from hypothesis testing if we get a p-value as 0.04? ______________.

    A) No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
    B) Weak evidence against the null hypothesis in favor of the alternative.
    C) Moderate evidence against the null hypothesis in favor of the alternative.
    D) Strong evidence against the null hypothesis in favor of the alternative.
    E) Very strong evidence against the null hypothesis in favor of the alternative.

15. When should we use multiple linear regression model? ______________.

    A) There is not enough data to carry out simple linear regression analysis.
    B) The relationship between the dependent variable and the independent variables cannot be described by a linear function.
    C) The dependent variable depends on more than one independent variable.
    D) The dependent variable is categorical.

16. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied by using least squares regression. The following regression equation was obtained from this study:

    $y= -0.0127 + 0.0180x$

    The above equation implies that:

    A) Each beer consumed increases blood alcohol by 1.27%
    B) On average it takes 1.8 beers to increase blood alcohol content by 1%
    C) Each beer consumed increases blood alcohol by an average of amount of 1.8%
    D) Each beer consumed increases blood alcohol by exactly 0.018

Part 2: Short Answer

Write your answer in the blank or box after each question.

1. What is the result of statement: strsplit("A B 1 2"," ")?

____

2. Rewrite the following R script by using pipe operator.

```
cyl <- mtcars$cyl
unique_cyl <- unique(cyl)
unique_cyl_sorted <- sort(unique_cyl, decreasing = TRUE)
print(unique_cyl_sorted)
```

Write your answer in the following box. Eliminate all temporary variables.

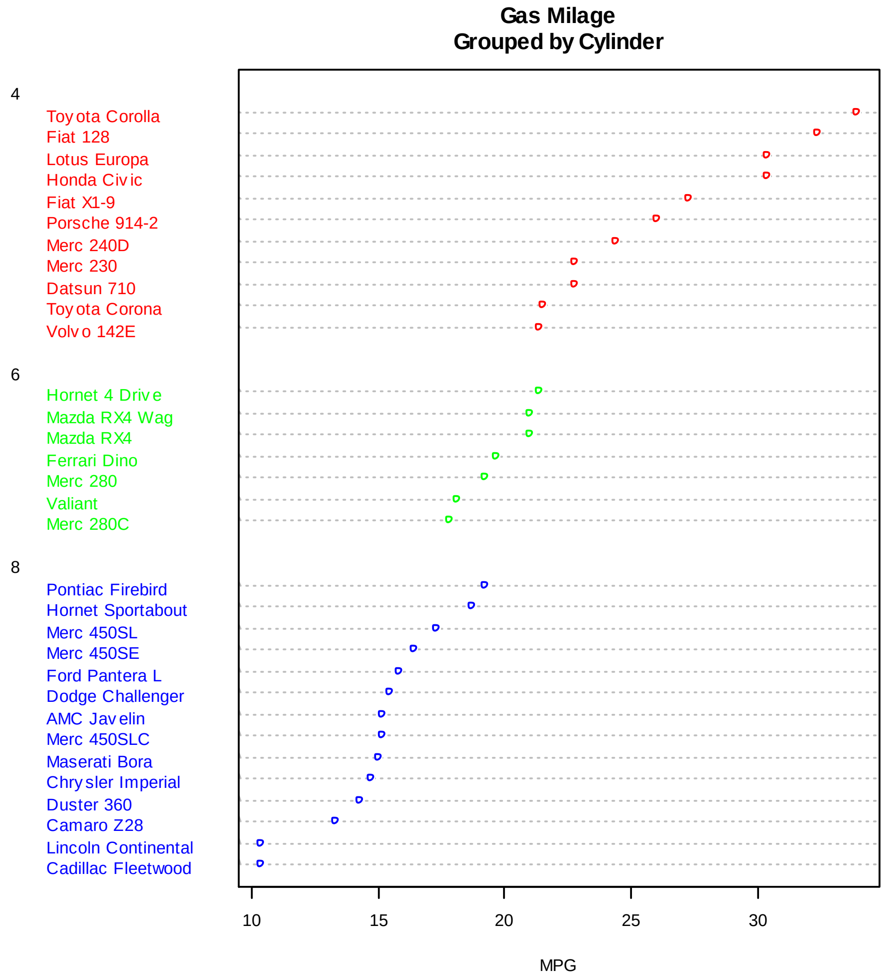
3. What is the difference between histograms and density plots?

Your answer:

____

____

5. What conclusion can you get from the following dot plot of mtcars dataset?

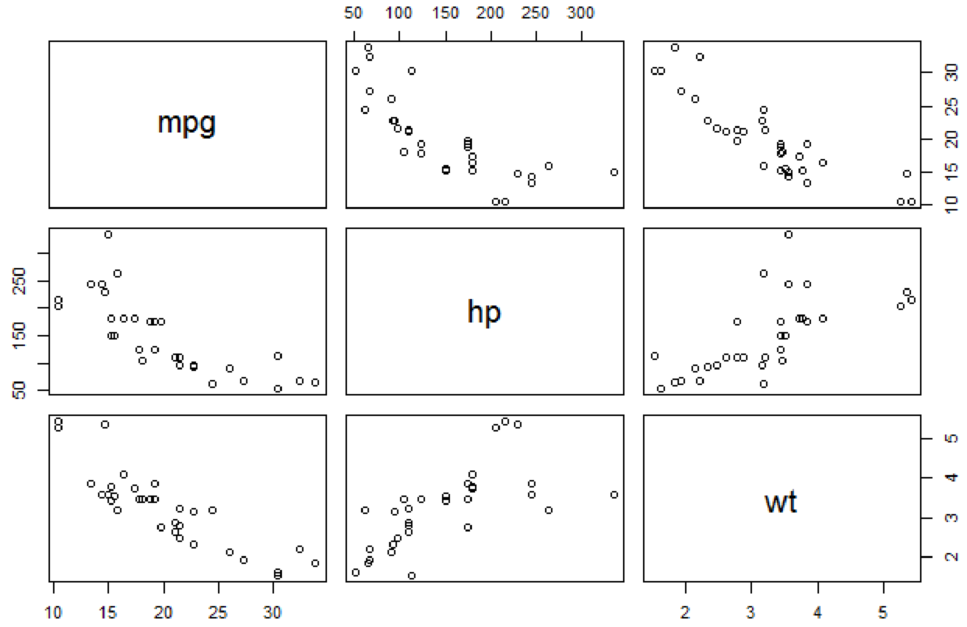


Your answer:

____

____

6. Based on the scatterplot matrix shown below, what is your explanation on the relationship between mpg and weight (wt)?



Your answer:

7. The school principal wants to test if it is true what teachers say – that high school juniors use the computer an average 3.2 hours a day. What are our null and alternative hypotheses? Specify them in the following box.

If we instead want to test if the average daily computer use is at least 3.2 hours, what are the null and alternative hypotheses?

Questions 8 through 9 are based on the following regression analysis:

Below is a regression analysis of used Toyota Corolla during late summer of 2004 in the Netherland.

```
Linear Regression
============================================================================
                                     Dependent variable:
                    --------------------------------------------------------
                                            Price
                                 (1)                           (2)
----------------------------------------------------------------------------
Age                         -122.0145***                  -122.1299***
                              (2.6022)                      (2.5963)

KM                           -0.0162***                    -0.0163***
                              (0.0013)                      (0.0013)

FuelTypeDiesel              3,390.0770***                 3,387.6750***
                             (518.7954)                    (509.0436)

FuelTypePetrol              1,120.6760***                 1,112.1620***
                             (332.3653)                    (331.6909)

HP                           60.8133***                    60.8932***
                              (5.7559)                      (5.6387)

MetColor                      57.1598
                             (74.9390)

Automatic                     330.2509*                     330.4641*
                             (157.0956)                    (156.1795)

CC                           -4.1744***                    -4.1682***
                              (0.5453)                      (0.5369)

Doors                         -7.7763
```

```
                             (40.0643)

Weight                       20.0094***                    19.9383***
                              (1.2033)                      (1.1259)

Constant                   -3,801.3610**                 -3,718.3640**
                           (1,304.0820)                  (1,261.4050)


-----------------------------------------------------------------------------
Observations                    1,436                         1,436
R2                             0.8693                        0.8693
Adjusted R2                    0.8684                        0.8685
Residual Std. Error   1,315.7140 (df = 1425)       1,315.0700 (df = 1427)
F Statistic        947.9742*** (df = 10; 1425) 1,186.0530*** (df = 8; 1427)
=============================================================================
Note:                                         *p<0.05; **p<0.01; ***p<0.001
```

8. In the first model as shown in column (1)

   a. What is the effect of Age on Price? Is it statistically significant?

   b. What is the coefficient estimate for dummy variable “FuelTypeDiesel”? What does it mean?

   c. List all independent variables that does not have statistically significant effect on Price?

9. Evaluate the simplicity (i.e., number of IVs) and usefulness (ie.e, adjusted R2) of the two models. Which model is best to predict/explain the price of used Toyota Corolla? Explain your reasons.

10. The structure and content of a dataset “stu_course” dataset is shown below.

```
> str(stu_course)
'data.frame':   4 obs. of  3 variables:
 $ name  : Factor w/ 3 levels "Helen","Mike",..: 2 3 1 1
 $ course: Factor w/ 2 levels "IST3420","IST5001": 1 2 1 2
 $ score : num  80 70 90 80
>
> print(stu_course)
   name  course score
1  Mike IST3420    80
2 Sarah IST5001    70
3 Helen IST3420    90
4 Helen IST5001    80
```

What is the result of the following R statement?

```
stu_course %>% dplyr::group_by(course) %>% dplyr::summarize(max(score))
```

Write the result in the following box.

11. Assume we have the following dataset:

X	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Y	4.5	4.6	3.1	5.2	3.9	4.8	3.8	4.2	4.3

Can we use X to predict/explain Y through a regression analysis? Why?