CIND 123 Data Analytics - Basic Methods
Midterm Review
Instructor: Dr. Tamer Abdou

mmmm dd, yyyy
hh:mm
Duration: xx hrs

You are allowed to use text books, notes, and calculators

Last name: _____

First name: _____

Student ID#: _____

There are **8** parts
and **100** marks total
with **10** bonus marks.

This exam paper should have 9 pages, including this cover page.

Part (I)	Introduction to Statistics	/15
Part(II)	Describing Data with Numeric Measures	/16
Part(III)	Describing Data with Graphs	/12
Part(IV)	Introduction to Probability	/12
Part(V)	Basics of R Language	/20
Part(VI)	Data Structures in R	/15
Part(VII)	Statistical Graphics in R	/10
Part(VIII)	Statements and Functions in R	/10
Total		/110

Instructions. (15 points) *Part I: Introduction to Statistics*
Identify the following variables as qualitative or quantitative. Classify the quantitative variables as discrete or continuous.

(3pts) **1.** Brand of a mobile phone
Solution:

(3pts) **2.** Whether or not a subject has disease X
Solution:

(3pts) **3.** Number of votes a political candidate receives
Solution:

(3pts) **4.** The pounds of sugar consumed by a person in a week
Solution:

(3pts) **5.** Number of persons on a fight from Toronto to Montréal
Solution:

Instructions. (16 points) *Part II: Describing Data with Numeric Measures*
For the following dataset: `10,6,2,7,100`

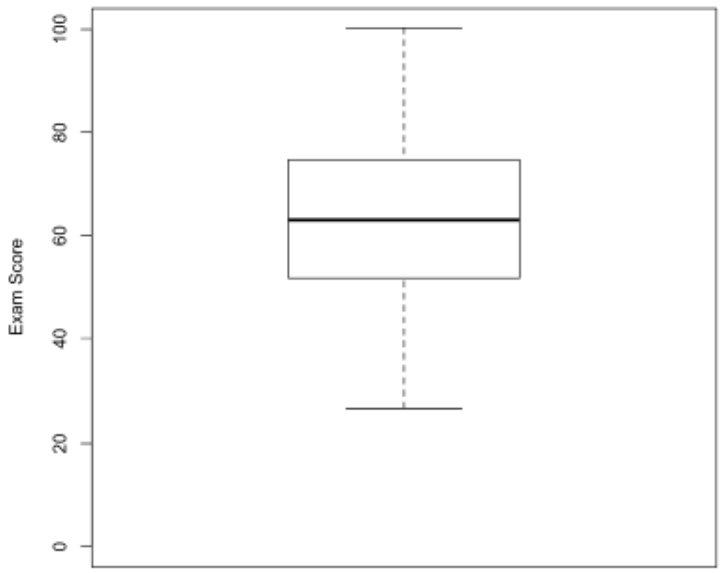
(4pts) **1.** Calculate the mean and median.
Solution:

(4pts) **2.** Which is a better measure of central tendency, the mean or median? Why?
Solution:

(4pts) **3.** If you added the same constant number, `K`, to each value in the data set, would the standard deviation change?
Solution:

(4pts) **4.** Will removing the lowest score affect the standard deviation?
Solution:

Instructions. (12 points) *Part III: Describing Data with Graphs*
The exam scores (out of 100 points) for all students taking a data analytics course are used to construct the following boxplot.



(4pts) **1.** Is the distribution of scores fairly symmetric? Why?
Solution:

(4pts) **2.** What is the approximate range of the data represented in the box-and-whisker plot?
Solution:

(4pts) **3.** What is the approximate percent of students scored above 65? Why?
Solution:

Instructions. (12 points) *Part VI: Introduction to Probability*
The numbers `1, 2, 3 and 4` are written on four pieces of paper. Suppose two of the pieces of paper are randomly selected from a hat WITHOUT replacement

(4pts) **1.** List the 12 outcomes in the sample space, S. (Hint: Use a tree diagram)
Solution:

(4pts) **2.** Give the probability of each element in S. Note all 12 outcomes in S are equally likely.
Solution:

(4pts) **3.** A coach is selecting 5 players to start in the game out of a team with a total of 12 players.
How many different ways can the coach select his/her 5 players?
Solution:

Instructions. (20 points) *Part V: Basics of R Language*
In each case below, write down the response (if any) that you would see in the R console window, if the given commands were typed into the console.

(4pts) **1.**

```
> x <- c(1, 1, 2, 3, 5, 8, 13)
> x[x > 4]
```

Solution:

(4pts) **2.**

```
> c(1, 1, 2, 3, NULL, 8, 13)
```

Solution:

(4pts) **3.**

```
> 8 %% 4 - 4 %% 8
```

Solution:

(4pts) **4.**

```
> 3 + 2 * 4 ^ 2
```

Solution:

(4pts) **5.**

```
x <- 1:3
x <= 2
```

Solution:

Instructions. (15 points) *Part VI: Data Structures in R*

The data in `thiamin.txt`, refer to 6 samples of each of 2 different types of cereal grain. Thiamin content was measured in each sample as follows:

```
grain content
WHEAT 5.2
WHEAT 4.5
WHEAT 6.0
WHEAT 6.1
WHEAT 6.7
WHEAT 5.8
OATS 8.3
OATS 6.1
OATS 7.8
OATS 7.0
OATS 5.5
OATS 7.2
```

Write the line(s) of R code that are required to

(5pts) **1.** Read the data from the file into a data frame `thiamin`.
Solution:

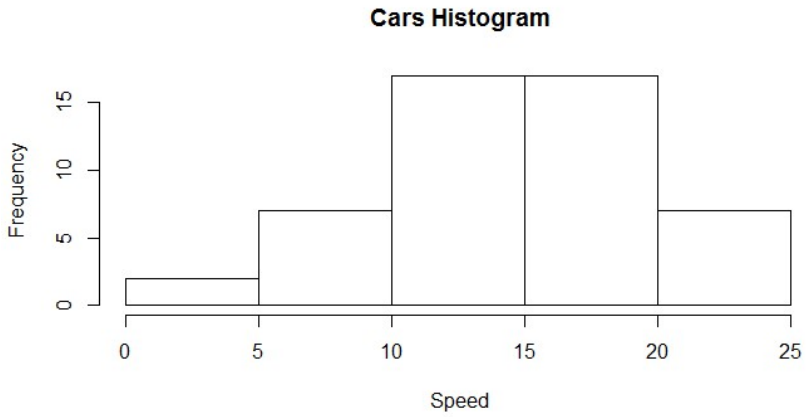
(5pts) **2.** Computes the median of `content`.
Solution:

(5pts) **3.** Computes the average of the thiamin `content` measurements in the sample of `OATS`.
Solution:

Instructions. (10 points) *Part VII: Statistical Graphics in R*
Write down the code required to produce the plots below, based on data in the built-in `cars` data frame, for which you will need the following information.

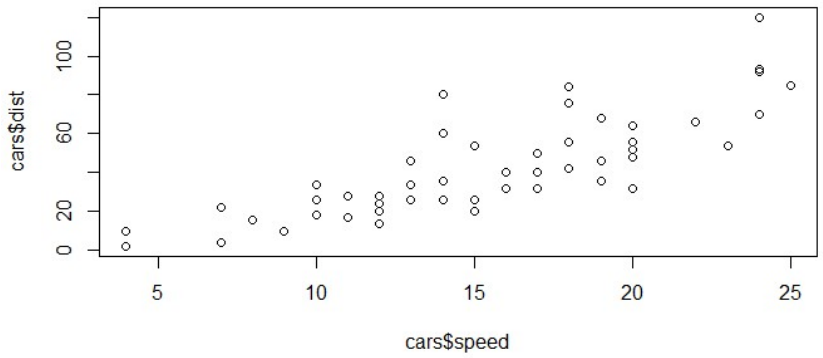
```
> names(cars)
[1] "speed" "dist"
```

(5pts) **1.**



Solution:

(5pts) **2.**

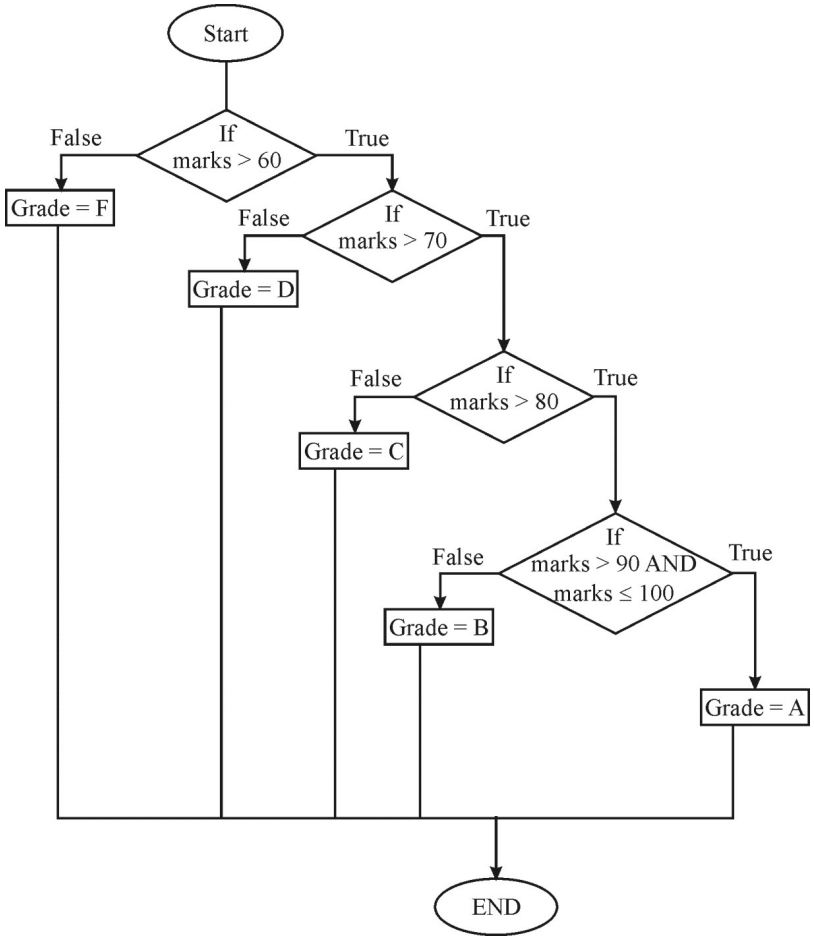


Solution:

Instructions. (10 points) *Part VIII: Statements and Functions in R*

Write an R function that reads a number as a student mark and determine the correspondent grade. The following is an algorithm for this program in a flow-chart diagram.

(10^{pts}) **1.**



Solution:

This is the end of questions.
Dr. Tamer Abdou