Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Data Science**	**Course Code:**	**CS481**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**60 Minutes**	**Total Marks:**	**17**
Paper Date:	**14-Apr-18**	**Weight**	**15 %**
Section:	**ALL**	**Page(s):**	**5**
Exam:	**Mid-II**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	Objective	1	2	3	Total
Marks	/ 5	/ 5	/ 4	/ 3	/ 17

Section 1

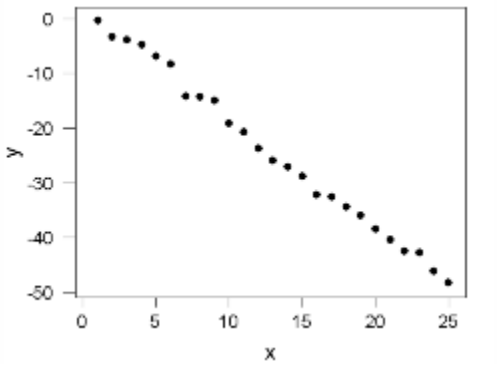
(Objective part) [points 5]

Clearly circle the correct options and explain your choice with reasoning.

Q1. Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 98% of times in the training data. Your model has 98% accuracy after taking the predictions on test data. Which of the following is true in such a case?

A) Accuracy metric is not a good idea for imbalanced class problems.
B) Accuracy metric is a good idea for imbalanced class problems.
C) Precision and recall metrics are good for imbalanced class problems.
D) Precision and recall metrics aren't good for imbalanced class problems.

Q2. Shown below is a scatterplot of Y versus X.



Which choice is most likely to be the approximate value of R^2?

A) 99.5% B) - 50.0%

C) 50.0% D) - 99.5%

Name: _____ Reg #: _____ Section: _____

Reason:

Q3. If we plot the residual, which of the following options indicate that there is margin of improvement in the model?

a) Random b) healthy plot c) Heteroscedasticity in plot d) unhealthy plot

Q4. Given the following data points

p1 = 1; p2 = 3; p3 = 8; p4 = 11; p5 = 18; p6 = 7;

If we perform hierarchical agglomerative clustering using MIN inter-cluster similarity (comparing the nearest points for each pair of clusters). Which points will be combined in the 1st iteration?

a) p1 and p2 c) p3 and p4 c) p4 and p5 d) p3 and p6

Q5. Does the correlation between two variables determines the causation? Explain your choice with an example.

a) False b) True

Section 2 (Subjective part) (marks 12)

Q1. Short Questions: (1+2+2 Marks)

a) Suppose we are given two data sets (A, B) of ice cream sales. The data set contains average temperature (x) of each day and the corresponding revenue (y) for that day. Your friend calculates the summary statistics (mean, variance, correlation coefficient, and line of fit), and tells you that both the data sets are almost identical. As a data scientist, will you believe it or will you use your data science skills to verify it. What will be your approach?

Solution:

b) [2 points] Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors (high bias or high variance) in its prediction. You can try some of the options given in the first column in order to fix the problem. Mark 'y' in the second or third column for all 6 options.

If you try	Fixes high bias	Fixes High variance
Get more training examples		
Try smaller sets of features		
Try decreasing λ		
Try increasing λ		
Try getting additional features		
Try adding polynomial features		

c) [2 points] Suppose you train a logistic regression classifier in order to predict if the aircraft engine is faulty or not. Our model predicts 1 if $h(x) > 0.7$. Given the test data ($m_{test} = 300$), we already know that 20% of the aircrafts have actually fault. On testing, our hypothesis predicted that 30% of the aircrafts have fault. Only 50% of the predicted ones (it's not 50% of the total), which actually have fault.

a. Create a table with actual number of true positive, true negative, false positive and false negative examples. Moreover, Calculate the F score.

Solution:

Name: _____ Reg #: _____ Section: _____

Q2. [4 points]: Exploratory Data Analysis

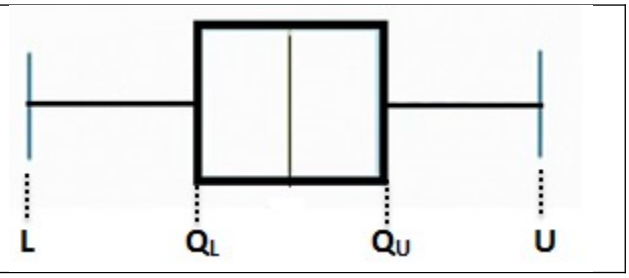
Consider a random sample of 26 grades on an easy data science exam:

100	100	99	98	97	96	95	95	95
94	93	93	92	92	91	90	90	90
89	84	80	75	68	65	50	45	

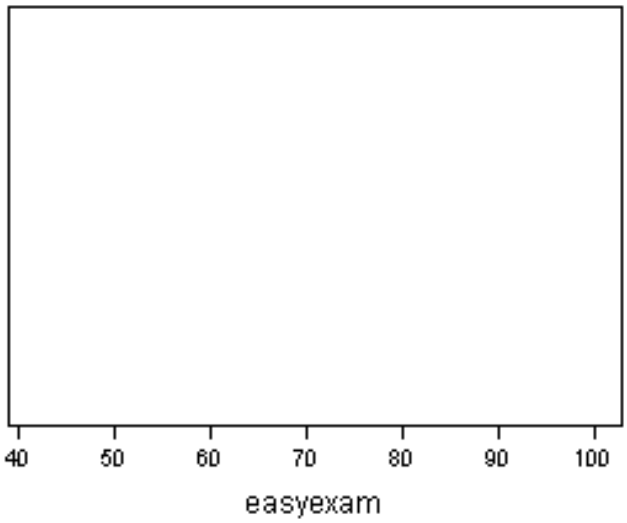
a) Draw Box plot of the data on the given figure below and suggest that the distribution of exam scores is symmetric, skewed right, or skewed left? **Moreover highlight the outliers if any.**

Note:
L (Lower limit) = Q_L – 1.5 × IQR
U (Upper limit) = Q_u + 1.5 × IQR

Where **IQR = Q_u – Q_L**



Name: _____ Reg #: _____ Section: _____



Q3: [3 points] Exploratory Data Analysis: Data Transformation

Suppose the relationship between X (independent variable) and Y (dependent variable) is represented by a power function $\mathbf{Y = 3X^2}$. The data for X is given in the table below. For the given data, we cannot best fit a linear Line.

A) Transform the data in such a way that we can fit a Linear Line using Linear Regression. Plot the transformed data on a graph and fit a linear line.
B) Find out the approximated intercept term and slope of the line.

Observation# 1	1	2	3	4	5	6	7
X	**1**	**2**	**2.5**	**3**	**3.5**	**4**	**5**