UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER I

WQD7001 : Principles of Data Science

Jan 2019 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

**ANSWER SCHEME**

(This question paper consists of 11 questions on 3 printed pages)

1. State **THREE (3)** types of data properties that you can investigate in exploratory data analysis (EDA).

(3 marks)

**Answer:**

*The center of the data*

*The spread among the members of the data*

*The skewness of the data*

*The probability distribution the data follows*

*The correlation among the elements in the dataset*

*Whether or not the parameters of the data are constant over time*

*The presence of outliers in the data*

1. Why skipping EDA is a bad idea?

(4 marks)

**Answer:**

*Exploratory Data Analysis is valuable to data science projects since it allows to get closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and data scientists but can be very informative about a particular business.*

1. Which language is more suitable for text analytics? R or Python?

Defend your answer.

(3 marks)

**Answer:**

*Since Python consists of a rich library called Pandas which allows the analysts to use high-level data analysis tools as well as data structures, while R lacks this feature. Hence Python will more suitable for text analytics.*

1. Data scientist requires skills in statistics.

Differentiate between descriptive statistics and inferential statistics.

(5 marks)

**Answer:**

***Descriptive statistics*** *consist of methods for organizing and summarizing information.*

***Descriptive statistics*** *includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.*

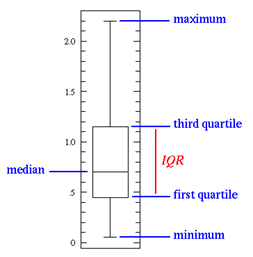
***Inferential statistics*** *consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population.*

***Inferential statistics*** *includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.*

1. Illustrate the five-number summary using a boxplot.

(5 marks)

**Answer:**



1. Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points.

Structure and explain the **THREE (3)** steps for linear regression.

(6 marks)

**Answer:**

*Firstly, a scatter plot should be used to analyze the data and check for directionality and correlation of data. Look for i) linear or non-linear pattern of the data and ii) deviations from the pattern (outliers). If the pattern is non-linear, consider a transformation. If there are outliers, you may consider removing them only IF there is a non-statistical reason to do so. This step enables the data scientist to formulate the model.*

*The second step of regression analysis is to fit the regression line. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot (for constant standard deviation assumption) and normal probability plot (for normality assumption). Once a “good-fitting” model is determined, write the equation of the least-squares regression line. Include the standard errors of the estimates, the estimate of , and R-squared.*

*The last step for the linear regression analysis is the test of significance, which is evaluating the validity and usefulness of the model. Determine if the explanatory variable is a significant predictor of the response variable by performing a t-test or F-test. Include a confidence interval for the estimate of the regression coefficient (slope).*

1. Discuss the **FIVE (5)** questions that machine learning can help answer, and for each question indicate the related algorithm.

(5 marks)

**Answer:**

*Is this A or B? (Classification algorithms)*

*How much / many? (Regression algorithms)*

*Is this weird? (Anomaly detection algorithms)*

*How is this organized? (Clustering algorithms)*

*What should I do now? (Reinforcement learning algorithms)*

1. In enhancing reproducibility, what is your advice or principles in terms of facilitating reproducibility and enabling discoverability?

(4 marks)

**Answer:**

*To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.*

*To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.*

1. DJ Patil defines a Data Product as “… a product that facilitates an end goal through the use of data”.

Provide **TWO (2)** broad groups of data products function.

(5 marks)

**Answer:**

*5 broad groups of data products functions: raw data, derived data, algorithms, decision support and automated decision-making.*

***Raw data****. Starting with raw data, we are collecting and making available data as it is (perhaps we’re doing some small processing or cleansing steps). The user can then choose to use the data as appropriate, but most of the work is done on the user’s side.*

***Derived data****. In providing users with derived data, we are doing some of the processing on our side. We could, in the case of customer data, add additional attributes like assigning a customer segment to each customer, or we could add their likelihood of clicking on an ad or of buying a product from a certain category.*

***Algorithms.*** *Next we have algorithms, or algorithms-as-a-service. We are given some data, we run it through the algorithm — be that machine learning or otherwise — and we return information or insights. A good example would be Google Image: the user uploads a picture, and receives a set of images that are the same or similar to the one uploaded. Behind the scenes, the product extracts features, classifies the image and matches it to stored images, returning the ones that are most similar.*

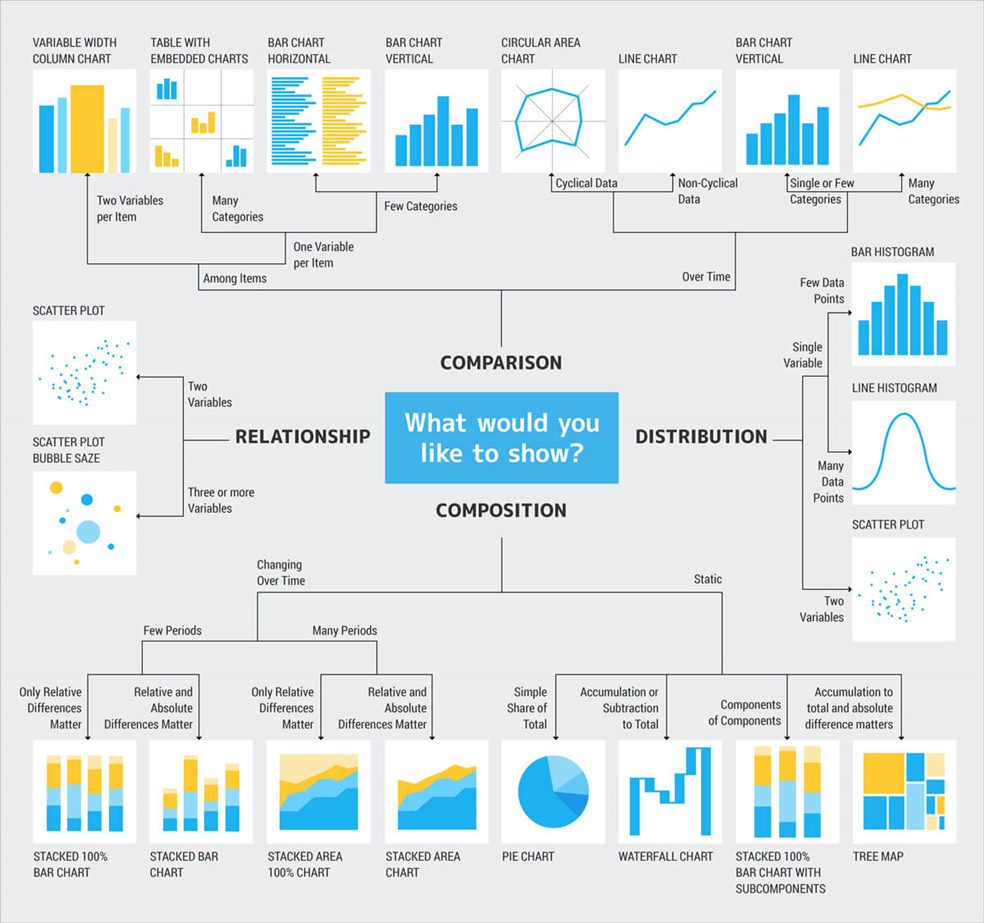
***Decision support.*** *Here we are looking to provide information to the user to help them with decision-making but we are not taking the decision ourselves. Analytics dashboards such as Google Analytics, Flurry, or WGSN would fall into this category. We are doing most of the heavy lifting on our side; our intention is to give the user relevant information in an easy-to-digest format to allow them to take better decisions. In the case of Google Analytics, that could mean changing the editorial strategy, addressing leaks in the conversion funnel, or doubling down on a given product strategy. The important thing to remember here is as follows: while we have taken design-decisions in data collection, derivation of new data, in choosing what data to display and how to display it, the user is still tasked with interpreting the data themselves. They are in control of the decision to act (or not act) on that data.*

***Automated decision-making****. Here we outsource all of the intelligence within a given domain. Netflix product recommendations or Spotify’s Discover Weekly would be common examples. Self-driving cars or automated drones are more physical manifestations of this closed decision-loop*.

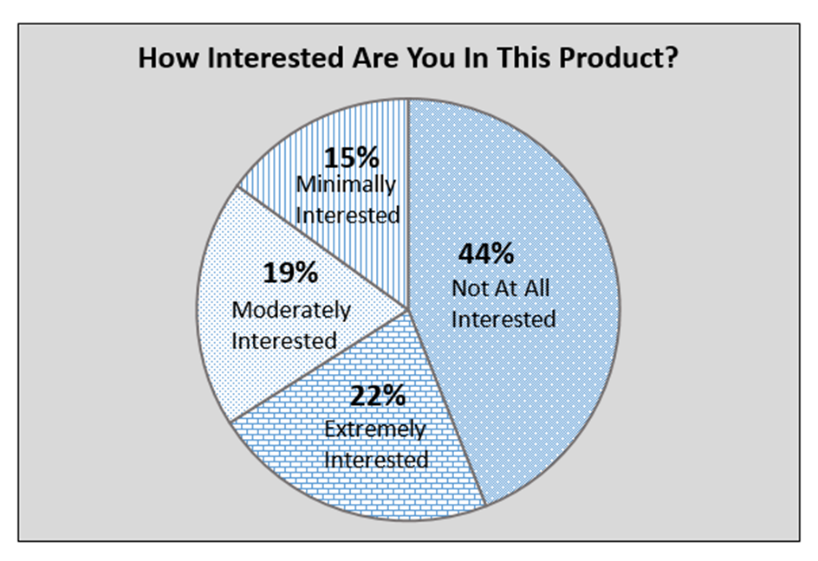
1. Summarize using a diagram, potential data storytelling visuals by grouping them based on relationship, comparison, distribution, or composition.

(6 marks)

**Answer:**

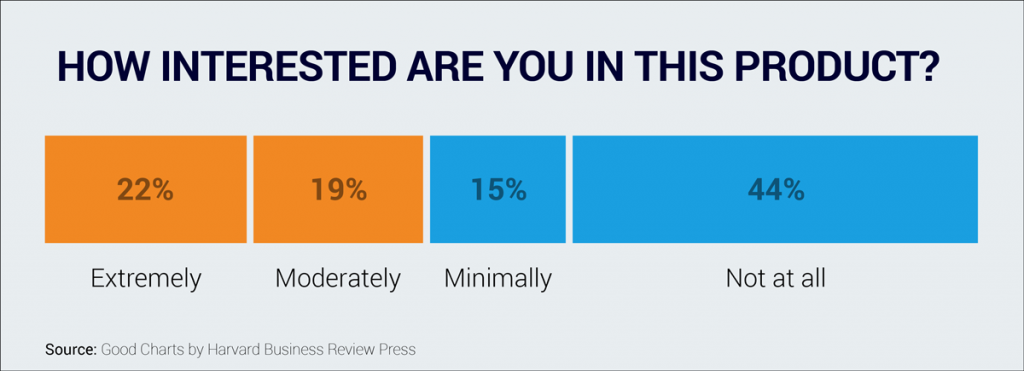


1. Evaluate the pie chart below and suggest a way that could improve the data storytelling. Depict your suggestion.



(4 marks)

**Answer:**



**END**