Question Bank for Data Science

Unit-1: Introduction

1. What is Data Science? What is its relationship to Statistics?
2. What is Data Science? With neat diagram explain Data Science Process.
3. Who is Data Scientist? What are typical job duties for data scientists?
4. What are applications of Data Science?
5. What are Top 10 Challenges to Practicing Data Science at Work?
6. Compare study of Data Science with Databases. What is the role of SQL in data science?
7. Define Scientific Computing. What are the capabilities of Computational Scientist? What are applications of Scientific Computing?
8. Explain Data Modeling Approaches.
9. Explain Statistical Data Modeling techniques.
10. Explain Bonferroni's Principle with suitable example.
11. Explain Data Visualization Techniques available for Data Scientist.

Unit -2: Data Preprocessing

1. Why do we need to pre-process the Data?
2. Describe the possible negative effects of proceeding directly to mine data that has not been pre-processed.
3. What are four ways to handle missing data in dataset? Of the four methods for handling missing data, which method is preferred?
4. What is an outlier? Why do we need to treat outliers carefully?
5. Explain graphical methods for identifying outliers.
6. Explain measures of center and spread.
7. Explain why data analysts need to normalize their numeric variables.
8. Explain Min-Max normalization, Z-Score Standardization and Decimal Scaling data transformation techniques.
9. For the stock price data given below, find Min-Max normalization stock price for all the stock prices

10	7	20	12	75	15	9	18	4	12	8	14

10. For the stock price data given below, find Z-Score Standardization stock price for all the stock prices

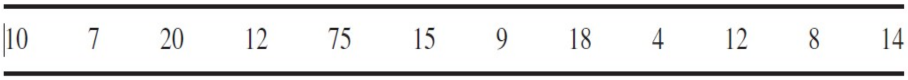
10	7	20	12	75	15	9	18	4	12	8	14

11. For the stock price data given below, find the decimal scaling stock price for all the stock prices

10	7	20	12	75	15	9	18	4	12	8	14

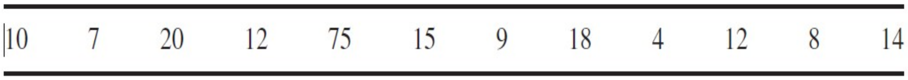
12. Explain transformations that can be applied to achieve normalization in the data? Why normalized dataset is preferred?
13. Explain numerical methods to identify outliers in the dataset.
14. For the stock price data given below, do the following
    a. Identify the outlier.
    b. Verify that this value is an outlier, using the Z-score method
    c. Verify that this value is an outlier, using the Interquartile Range (IQR)

method.



10	7	20	12	75	15	9	18	4	12	8	14

15. For the stock price data given below, identify all possible stock prices that would be outliers, using:
    a. The Z-score method.
    b. The IQR method.



10	7	20	12	75	15	9	18	4	12	8	14

16. What is flag variable? What is it's use?
17. Explain techniques for Binning Numerical Variables.
18. Explain why we might not want to remove a variable that had 90% or more missing values.
19. Explain why we might not want to remove a variable just because it is highly correlated with another variable.
20. How to remove duplicate records from dataset?

Unit-3: Exploratory Data Analysis

1. What is Exploratory Data Analysis (EDA)? What are objectives of EDA?
2. What is Exploratory Data Analysis (EDA)? Why do we need to perform EDA?
3. Explain the difference between EDA and hypothesis testing, and why analysts may prefer Exploratory Data Analysis (EDA) when doing data science project.
4. What is contingency table? Why do we use contingency tables, instead of just presenting the graphical results?
5. What is the difference between taking row percentages and taking column percentages in a contingency table?
6. What type of histogram is useful for examining the relationship between a numerical predictor and the target?
7. How histogram can be used to examine relationship between a numerical predictor and the target?
8. With suitable examples explain how Exploratory Data Analysis (EDA) would help to uncover the anomaly in training data.
9. Explain the objective and the methods of binning numerical variables.
10. Explain the objective and the method of binning based on predictive value.
11. Explain method of binning based on predictive value.
12. What are correlated variables? Describe the possible consequences of allowing correlated variables to remain in the model.
13. What are advantages of deriving new variables from predictor variables? How to assess usefulness of new derived variables in predicting the target variable using Exploratory Data Analysis (EDA)?
14. How Exploratory Data Analysis (EDA) can be used to investigate correlated predictor variables?

Unit 4: Unstructured Data Mining

1. What are applications of applications of text categorization?
2. What is difference between single label and multilabel categorization, document-pivoted and category-pivoted categorization, hard and soft categorization?
3. Explain techniques for document representation.
4. What is Feature Selection? Explain methods of Feature Selection used in text categorization.
5. How dimensionality can be reduced by using feature extraction?
6. Explain knowledge engineering approach to Text Categorization (TC).
7. Explain any two Machine Learning approaches to Text Categorization (TC).
8. Explain Text Categorization (TC) using Probabilistic Classifiers and Bayesian Logistic Regression.
9. Explain Text Categorization (TC) using Decision Tree Classifiers and Decision Rule Classifiers.
10. Explain Text Categorization (TC) using Bayesian Logistic Regression (BLR), Decision Tree Classifiers and Decision Rule Classifiers.
11. Explain Text Categorization (TC) using Regression Method, Rocchio Method, and Neural Networks.
12. Explain Text Categorization (TC) using Example-Based Classifiers and Support Vector Machines.
13. Explain committees of classifiers. What are advantages of using classifier committees?
14. How unlabeled data can be used to improve classification?
15. Explain ways of incorporating knowledge from unlabeled documents to improve classification.
16. How to evaluate performance of text classifiers?
17.
18. What are applications of text clustering?
19. Define general clustering problem. Explain most popular Similarity Measure metrics used in clustering algorithms.

Unit 4: Unstructured Data Mining

20. Explain document clustering algorithms.
21. Explain Hierarchical Agglomerative Clustering (HAC) algorithm for clustering text documents.
22. Explain Hierarchical Agglomerative Clustering (HAC) algorithm for clustering text documents. How to dendrogram are used in the clustering process?
23. How documents are represented for text clustering?
24. Explain dimension reduction using Latent Semantic Indexing (LSI).
25. Explain data abstraction in text clustering and evaluation of text clustering algorithms.

Unit-5 : Social Network Analysis

1. What is a Social Network? How Social Network can be modeled?
2. How Social-Network Graph can be clustered by applying Standard clustering methods?
3. What is Betweenness? Explain Girvan-Newman Algorithm to calculate the Betweenness.
4. What is Betweenness? How Betweenness can be used to Find Communities in Social Network Graph.
5. How to discover Communities in Social-Network Graph directly?
6. How Social-Network Graph can be partitioned to identify Communities?
7. How to find overlapping communities in Social Network Graph?
8. Explain Affiliation-Graph Model to find overlapping communities in Social-Network Graph.
9. Why triangles in Social-Network Graph are counted? Explain algorithm for finding triangles in Social Network Graph.

Unit 6: Model Evaluation Techniques

1. Why do we need to evaluate our models before model deployment?
2. What is the minimum descriptive length principle, and how does it represent the principle of Occam's razor?
3. Why do we not use the average deviation as a model evaluation measure? How is the square root of the Mean Square Error (MSE) interpreted?
4. What might be a drawback of evaluation measures based on squared error? How might we avoid this?
5. Explain model evaluation techniques for the estimation and prediction tasks.
6. Explain model evaluation measures for the classification task.
7. Explain classification evaluation measures accuracy, overall error rate, sensitivity and specificity.
8. What is the difference between the total predicted negative and the total actually negative?
9. What is the relationship between accuracy and overall error rate?
10. Explain classification evaluation measures false-positive rate and false-negative rate, proportions of true positives, true negatives, false positives, and false negatives.
11. Explain how misclassification cost can be adjusted to reflect real world concerns.
12. With suitable example explain decision cost/benefit analysis.
13. Explain use of lift charts and gains charts to compare model performance.