Individual Project 9
DS160-02
Introduction to Data Science
Spring 2023

Data Science Questions (35 points)

Goal: This project aims to do a basic knowledge check that we covered in this class.

Instructions: For this project, create a pdf script titled **IP9_XXX.pdf**, where **XXX** are your initials. Also create a GitHub repository titled **IP9_XXX** to which you can **push your pdf file along with the Word file.**

1. Define the term 'Data Wrangling in Data Analytics.
   - This is the term that refers to the process of cleaning, formatting, and transforming data to make sure it is consistent, accurate, and easier to work with.
2. What are the differences between data analysis and data analytics?
   - Data Analytics is a term that describes the overall process of getting the data, cleaning it, performing analysis, and interpreting it to get desired information. Data Analysis is a specific step in the overall analytics process.
3. What are the differences between machine learning and data science?
   - Data Science is a field; it is the entire process of collecting, cleaning, processing, analysis, and interpretation.
   - Machine Learning is a subset of Data Science that uses algorithms and models to learn from past data to make decisions, predictions, and interpretations about future data.
4. What are the various steps involved in any analytics project?
   - Defining the problem(s), collection, cleaning, EDA, modeling.
5. What are the common problems that data analysts encounter during analysis?
   - Missing values or inconsistent data, communication problems, bias.
6. Which technical tools have you used for analysis and presentation purposes?
   - Python, R, and SQL (Languages)
   - Excel and Tableu (Visualization)
   - Pandas, Scikit-Learn, ggplot, caTools, Numpy, Seaborn, etc. (Packages)
7. What is the significance of Exploratory Data Analysis (EDA)?
   - It is a major step in the analysis process because it allows analysts to do visualization techniques, understand the data, and eventually communicate what is in the data, and what is happening. The communication from this can lead to better informed analysis moving forward.
8. What are the different methods of data collection?
   - Interviews, surveys, experiments, etc.
9. Explain descriptive, predictive, and prescriptive analytics.
   - Descriptive Analytics is what is focused on what has happened in the past, and how we can identify patterns from it.

- Predictive Analytics is what is focused on what can happen in the future (machine learning, as an example).
- Prescriptive Analytics is focused on what should be done to achieve a certain outcome. It can use the other two analytics to help aid in decision making.

10. How can you handle missing values in a dataset?
    - Imputation, regression, or simply just deleting the null values.
11. Explain the term Normal Distribution.
    - It is a probability distribution that is symmetric about the mean, appearing visually on a graph to be perfectly evenly distributed on both sides.
12. How do you treat outliers in a dataset?
    - You can delete the outlier from the dataset, you can also impute the missing data.
13. What are the different types of Hypothesis testing?
    - A/B Testing, Null Hypothesis
14. Explain the Type I and Type II errors in Statistics?
    - Type I Error is a false positive; it happens when the null hypothesis is true, but is rejected.
    - Type II Error is a false negative; it happens when the null hypothesis is false, but fails to be rejected.
15. Explain univariate, bivariate, and multivariate analysis.
    - Univariate analysis is the study of a single variable. Bivariate is the study of a relationship between two variables. Multivariate analysis is the analysis of two or more variable.
16. Explain Data Visualization and its importance in data analytics?
    - Data Visualization is relatively self-descriptive; it is the process or ability to visually see changes made to our dataset(s). This process can happen through graphs, charts, plots, etc. It gives a visual context to our information and is easier to understand for both analysts and non-analysis.
17. Explain Scatterplots.
    - Scatterplots use dots to compare (or represent) values of two different numeric values. This type of plot is beneficial for showing relationships between two numeric values.
18. Explain histograms and bar graphs.
    - Histograms break the range of values of a variable into classes and displays the count(or percent) of the observations that fall in each class.
    - Bar graphs display the distribution of a categorical variable (whereas histograms are for quantitative variables).
19. How is a density plot different from histograms?
    - A density plot uses individual dots on a plot, whereas histograms use boxes/rectangle shapes.
20. What is Machine Learning?
    - An area of artificial intelligence that builds methods that teaches machines to improve performance, responses, or tasks (teaching machines to “learn”).
21. Explain which central tendency measures to be used on a particular data set?

-

22. What is the five-number summary in statistics?
    - Minimum, 1st Quartile, Median, 3rd Quartile, Maximum.
23. What is the difference between population and sample?
    - Population is a (large) group of individuals about whom we have questions.
    - A Sample is a subgroup of individuals from a population in which we will collect data from.
24. Explain the Interquartile range?
    - IRQ is the middle 50% of a dataset.
25. What is linear regression?
    - Statistical technique where the score of a variable Y is predicted from the score of a second variable X (the X variable is the predictor variable).
26. What is correlation?
    - Correlation explains how one or more variables are related to each other.
27. Distinguish between positive and negative correlations.
    - Positive correlation means that as one variable increases, the other variable increase. Negative correlation means that as one variable increases, the other variable decreases.
28. What is Range?
    - The difference between the largest and smallest values in a dataset.
29. What is the normal distribution, and explain its characteristics?
    - A normal distribution is a bell-shaped curve that is symmetrical. A characteristic of normal distribution are Z-scores, mean, median, mode, and standard deviation.
30. What are the differences between the regression and classification algorithms
    - Regression algorithms help to predict numerical values, whereas classification algorithms are used to predict categorical values.
31. What is logistic regression?
    - A process of modeling the probability of a discrete outcome given an input variable.
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    - RMSE is found by calculating the predicted values of the dependent variable, calculating the residuals, squaring the residuals, and taking the square root of of the mean squared error.
    - MSE is found by the same process, however you do not take the square root.
33. What are the advantages of R programming?
    - Integration, cross-platform, versatility for data analysis.
34. Name a few packages used for data manipulation in R programming?
    - Reshape2, tidyr
35. Name a few packages used for data visualization in R programming?
    - Ggplot2, leaflet, shiny