Alexander Kyrkjeeide

Individual Project 5
DS160
Introduction to Data Science
Fall 2023

Data Science Questions (70 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 **IP5_XXX.pdf**, where **XXX** are your initials. Also create a GitHub repository titled **IP5_XXX** to which you can **push your pdf file along with the Word file.** Show your best work and keep the document for your future journey.

1. **Define the term 'Data Wrangling in Data Analytics:**

   Data wrangling, or data munging, is the process of cleaning, structuring, and organizing raw data into a format suitable for analysis. This involves tasks such as handling missing data, converting data types, dealing with outliers, and merging datasets.

2. **What are the differences between data analysis and data analytics?**
   - Data analysis is a broader term, encompassing the process of inspecting, cleaning, transforming, and modeling data to discover information and draw conclusions.
   - Data analytics is a more specific term, focusing on using tools and techniques to analyze data, often with the goal of extracting actionable insights and making informed decisions.

3. **What are the differences between machine learning and data science?**

   Data science includes a range of activities, while machine learning is a subset that involves using algorithms to enable systems to perform tasks without explicit programming.

4. **What are the various steps involved in any analytics project?**

   Steps include defining the problem, data collection, cleaning, exploratory data analysis, feature engineering, model building, evaluation, deployment, and monitoring.

5. **What are the common problems that data analysts encounter during analysis?**

   Common problems include dealing with missing data, handling outliers, addressing inconsistent data, and interpreting data accurately.

6. **Which technical tools have you used for analysis and presentation purposes?**

I, as a text-based AI model, don't use tools directly. However, analysts often use tools like Python (pandas, scikit-learn), R, SQL, Tableau, Excel, and Jupyter Notebooks.

7. **What is the significance of Exploratory Data Analysis (EDA)?**
   EDA is crucial for understanding data characteristics before in-depth analysis. It involves summarizing main features, often with statistical graphics, to uncover patterns and relationships.
8. **What are the different methods of data collection?**
   Methods include surveys, experiments, observational studies, interviews, sensor data, and web scraping.
9. **Explain descriptive, predictive, and prescriptive analytics.**
   - Descriptive Analytics: Describes what has happened.
   - Predictive Analytics: Predicts what is likely to happen.
   - Prescriptive Analytics: Recommends actions to achieve a desired outcome.
10. **How can you handle missing values in a dataset?**
    Handling missing values can involve removing rows, imputing with mean/median, or using advanced imputation techniques like k-nearest neighbors.
11. **Explain the term Normal Distribution.**
    A normal distribution is a symmetric, bell-shaped probability distribution characterized by a mean and standard deviation. It follows the 68-95-99.7 rule, where most data falls within one, two, and three standard deviations of the mean.
12. **How do you treat outliers in a dataset?**
    Treatment may involve identification using statistical methods, transformation of data, or removal if they are data errors.
13. **What are the different types of Hypothesis testing?**
    Types include t-test, Chi-squared test, ANOVA, Z-test, and Wilcoxon signed-rank test.
14. **Explain the Type I and Type II errors in Statistics?**
    - Type I Error: Incorrectly rejecting a true null hypothesis (false positive).
    - Type II Error: Failing to reject a false null hypothesis (false negative).
15. **Explain univariate, bivariate, and multivariate analysis.**
    Univariate: Analyzing one variable at a time.
    Bivariate: Analyzing the relationship between two variables.

Multivariate: Analyzing the relationship between three or more variables simultaneously.

16. **Explain Data Visualization and its importance in data analytics?**

    Data visualization is the representation of data graphically to discover insights. It aids in understanding patterns, trends, and outliers.

17. **Explain Scatterplots.**

    Scatterplots are graphical representations of the relationship between two continuous variables, with one variable on the x-axis and the other on the y-axis.

18. **Explain histograms and bar graphs.**

    Histograms: Show the distribution of a continuous variable using bins.

    Bar Graphs: Display the distribution of a categorical variable using bars.

19. **How is a density plot different from histograms?**

    Density plots provide a smoothed representation of the distribution, showing the probability density function, while histograms represent discrete bins.

20. **What is Machine Learning?**

    Machine learning is a field of artificial intelligence focused on developing algorithms that enable systems to learn from data and make predictions or decisions without explicit programming.

21. **Explain which central tendency measures to be used on a particular data set?**

    Use mean for symmetrical distributions and median for skewed distributions or in the presence of outliers.

22. **What is the five-number summary in statistics?**

    The five-number summary includes the minimum, first quartile, median, third quartile, and maximum of a dataset.

23. **What is the difference between population and sample?**

    Population: Entire set of individuals or data.

    Sample: Subset selected for analysis.

24. **Explain the Interquartile range?**

    The interquartile range is the range covered by the middle 50% of the data, calculated as the difference between the third quartile and the first quartile.

25. **What is linear regression?**

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the data.

26. **What is correlation?**

    Correlation measures the strength and direction of a linear relationship between two variables.

27. **Distinguish between positive and negative correlations.**

    Positive Correlation: Both variables increase or decrease together.

    Negative Correlation: One variable increases while the other decreases, and vice versa.

28. **What is Range?**

    Range is the difference between the maximum and minimum values in a dataset, providing a measure of data spread.

29. **What is the normal distribution, and explain its characteristics?**

    A normal distribution is symmetric and bell-shaped, characterized by mean and standard deviation. It follows the 68-95-99.7 rule.

30. **What are the differences between regression and classification algorithms?**

    Regression predicts a continuous output, while classification predicts a categorical output.

31. **What is logistic regression?**

    Logistic regression is used for binary classification problems, predicting the probability of an instance belonging to a particular class.

32. **How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?**

    RMSE is the square root of the mean squared differences between predicted and actual values, while MSE is the mean of these squared differences.

33. **What are the advantages of R programming?**

    R is open-source, has a vast ecosystem of packages, and is widely used for statistical computing and graphics.

34. **Name a few packages used for data manipulation in R programming?**

    dplyr, tidyr, data.table are popular packages for data manipulation in R.

35. **Name a few packages used for data visualization in R programming?**

ggplot2, lattice, and plotly are commonly used packages for data visualization in R.