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 in Data Analytics is cleaning, organizing, and restructuring raw data to make it worthwhile for analytics or machine learning. Data wrangling is essential to ensuring the accuracy and reliability of data, providing meaningful insights, and making informed business decisions.

2. What are the differences between data analysis and data analytics?

   Data analysis involves extracting data in a way that can benefit a decision-maker. On the other hand, data analytics uses analytical tools and techniques to find new insights and make predictions. For example, many businesses today hire data analytics to find new ways to be ahead of their competitors.

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

   Machine learning focuses on understanding and building models that use data to improve performance or inform predictions. On the other hand, data science studies data and finds ways to extract meaning from it.

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

   I believe the first step in any analytics project is figuring out the problem you are trying to solve or the question you are trying to answer. Next, you should obtain and understand the data. Then, you should prepare your data by implementing data cleaning or wrangling in your routine. In addition, performing an EDA can help you summarize and identify outliers or patterns in the model to help you build a model. Later, you should perform data modeling and use data visualizations to present your work to others.

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

   Data analysts encounter various problems during the analysis process. For example, data quality issues like incomplete or inaccurate data can cause errors, outliers, and inconsistency, impacting the analysis results. In addition, data exploration, like exploratory data analysis (EDA), can be challenging when identifying patterns, trends,

and outliers with large datasets. Additionally, statistical challenges like sampling bias in the data sample can affect the results. Also, choosing an appropriate model can be challenging due to avoiding underfitting and overfitting. Lastly, communication in general, like communicating and interpreting results, can be challenging to ensure that the results are accurate and appropriate.

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

   Data analysis and visualization:
   - Python – Pandas for data manipulation, NumPy for numerical operations, and Matplotlib/Seaborn for data visualization.
   - R - Statistical programming language with several built-in data analysis and visualization packages. Also, R is used heavily for statistical modeling and hypothesis testing.
   - Tableau and Microsoft PowerPoint – can be helpful for presentation purposes while communicating your data with others.
   - Jupyter Notebooks– creates formatted documents that combine live code, equations, visualizations, and narrative text.
   - GitHub- Useful for tracking source code and collaborative work.

7. What is the significance of Exploratory Data Analysis (EDA)?

   Exploratory Data Analysis (EDA) is vital because it involves examining and understanding a dataset's structure, patterns, and characteristics before modeling. EDA includes data cleaning, visualization, and summary statistics to identify outliers, patterns, and relationships, which sets the groundwork for meaningful interpretation and effective communication of data findings.

8. What are the different methods of data collection?

   A few different methods of data collection:
   - Surveys and Questionnaires - involve gathering information from respondents through structured questions.
   - Secondary data analysis is a collection of data by others for more extraordinary interpretation and analysis.
   - Internet and social media data - Data collection from online platforms and social media
   - Biometric data collection - Physiological data collection, such as heart rate or brain activity

9. Explain descriptive, predictive, and prescriptive analytics.

   - Descriptive analytics provides a baseline understanding of historical performance, helping organizations identify trends, patterns, and insights from past data.
   - Predictive analytics helps organizations anticipate future events or trends, enabling them to make proactive decisions and take preventive actions.

- Prescriptive analytics helps organizations make better decisions by suggesting the most effective actions to achieve desired outcomes. It considers constraints, resources, and potential uncertainties.

10. How can you handle missing values in a dataset?

    - Identify missing values – use is.na() in R or isnull() in Python to identify missing values in the dataset.
    - Impute missing values - imputation methods include using the mean, median, or mode for numerical data or the most frequent category for categorical data to replace missing values.

11. Explain the term Normal Distribution.

    A normal Gaussian or bell curve is a symmetric probability distribution that forms a characteristic bell-shaped curve. The normal distribution is a fundamental concept in statistics, widely used for modeling and analyzing various phenomena where data tends to cluster around a central value.

12. How do you treat outliers in a dataset?

    One way to treat outliers is by removing them if they result from errors or extreme values to reduce their impact on statistical analyses.

13. What are the different types of Hypothesis testing?

    There are various types of hypothesis testing, including t-tests, chi-square tests, ANOVA, and regression analysis, each designed for specific scenarios to assess whether observed differences or relationships in data are statistically significant.

14. Explain the Type I and Type II errors in Statistics.

    Type I error occurs when a true null hypothesis is rejected incorrectly, while Type II error occurs when a false null hypothesis is not rejected. Balancing these errors is crucial in hypothesis testing to maintain the reliability of statistical conclusions.

15. Explain univariate, bivariate, and multivariate analysis.

    Univariate analysis examines a single variable, bivariate analysis explores relationships between two variables, and multivariate analysis considers interactions among multiple variables, offering a more comprehensive understanding of complex data patterns.

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

Data visualization involves representing data graphically to facilitate understanding and interpretation. It is crucial in data analytics to convey insights, trends, and patterns effectively to both technical and non-technical audiences.

17. Explain Scatterplots.

Scatterplots display the relationship between two continuous variables, with each point representing a data observation. They help visualize patterns, trends, and potential correlations in a dataset.

18. Explain histograms and bar graphs.

Histograms represent a continuous variable's distribution, displaying the data frequency within predefined intervals. Bar graphs, on the other hand, illustrate the distribution of categorical variables by representing the frequency or proportion of each category with distinct bars.

19. How is a density plot different from histograms?

Density Plot:

- Smooth representation of the distribution of a continuous variable
- It provides a more continuous view of the data distribution.
- Useful for visualizing the underlying probability distribution.

Histograms:

- Represents the distribution of a variable by dividing it into discrete bins.
- Shows the frequency or count within each bin.
- It can be less smooth and more sensitive to bin width.

20. What is Machine Learning?

Machine Learning is a field of artificial intelligence focused on developing algorithms that enable computers to learn patterns and make predictions. It involves training models on large datasets, allowing them to identify patterns, relationships, and trends. Machine Learning applications range from image and speech recognition to recommendation systems, and they play a pivotal role in automating tasks and extracting valuable insights from complex datasets.

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

- Mean: Suitable for symmetric distributions without extreme values.
- Median: Appropriate for skewed distributions or in the presence of outliers.
- Mode: Useful for identifying the most frequently occurring value in categorical or discrete data.

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

The five-number summary in statistics includes the minimum and maximum values, the first quartile (Q1), the median (Q2), and the third quartile (Q3), providing a concise summary of the dataset's central tendency and spread.

23. What is the difference between population and sample?

In statistics, a population refers to the entire group of individuals, items, or data points that share a common characteristic and are of interest to a particular study. A sample, on the other hand, is a subset of the population selected for analysis. Sampling aims to draw valid inferences about the population based on observations from the sample. The main distinctions lie in the scope (entire group vs. subset) and the practicality of studying the population, making sampling a valuable technique for statistical analysis.

24. Explain the Interquartile range.

- The Interquartile Range (IQR) is a statistical measure describing a dataset's spread or dispersion.
- It calculates the difference between a dataset's third quartile (Q3) and the first quartile (Q1).
- IQR is commonly used in box-and-whisker plots to visually represent a dataset's central tendency and spread, highlighting the range where the middle 50% of the data lies.

25. What is linear regression?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values. This linear model allows for predicting the dependent variable based on the values of the independent variables, making it a fundamental tool in statistical modeling and predictive analytics.

26. What is correlation?

Correlation is a statistical concept that measures the degree of association or relationship between two variables. It quantifies the strength and direction of the linear relationship: a correlation coefficient close to 1 signifies a robust positive relationship, relatively close to -1 indicates a strong negative relationship and around 0 suggests no linear correlation.

27. Distinguish between positive and negative correlations.

Positive correlation - an increase in the value of one variable is associated with an increase in the value of the other variable. The positive correlation coefficient, approaching +1, indicates a direct and proportional relationship.

Negative correlation - an increase in the value of one variable is associated with a decrease in the value of the other variable. The correlation coefficient is negative, approaching -1, indicating an inverse relationship where one variable decreases as the other increases.

28. What is Range?

    Range in statistics represents a dataset's spread or dispersion. It is calculated as the difference between the maximum and minimum values within the dataset, providing a simple but sensitive indication of the variability of the data.

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

    The normal distribution, also known as the Gaussian or bell curve, is a symmetric probability distribution characterized by a bell-shaped curve.

    Key characteristics include:

    - Symmetry - The distribution is symmetric around its mean, median, and mode.
    - Bell-shaped curve – The graph forms a smooth, bell-shaped curve, with the highest point at the mean.
    - 68-95-99.7 Rule: Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations

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

    Regression algorithms can predict continuous numerical values, such as house prices. In contrast, classification algorithms can categorize data into distinct classes or groups, such as spam or non-spam emails. Regression output is a range of values, while classification produces discrete class labels. Examples of regression algorithms include Linear Regression, while classification algorithms include Logistic Regression, Decision Trees, and Support Vector Machines. Evaluation metrics differ as well, with regression using measures like Mean Squared Error and classification using metrics such as accuracy, precision, recall, and F1 score.

31. What is logistic regression?

    Logistic regression is a statistical model used for binary classification tasks, predicting the probability that an observation belongs to a specific category. It employs the logistic function to transform a linear combination of input features into a probability and can classify observations into binary outcomes.

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

    The Root Mean Square Error (RMSE) and Mean Square Error (MSE) are measures commonly used to assess the accuracy of a predictive model.

Mean Square Error (MSE):
The MSE is calculated as the average of the squared differences between the observed (actual) values and the predicted values. The formula gives it:
MSE=1n∑i=1n(yi−y^i)2*MSE*=*n*1∑*i*=1*n*(*yi*−*y*^*i*)2
Where:

- n*n* is the number of observations.
- yi*yi* is the observed value.
- y^i*y*^*i* is the predicted value.

R can calculate it by using the mean() function.

Root Mean Square Error (RMSE):
The RMSE is the square root of the MSE. It measures the average magnitude of the errors in the predicted values. The formula is:
RMSE=MSE*RMSE*=*MSE*
R can calculate RMSE using the sqrt() function.

The lower the values of MSE and RMSE, the better the model performance.

33. What are the advantages of R programming?

R programming is a free, open-source language that companies can use without a license. In addition, R is applicable in data analysis, including machine learning and statistical analysis. Lastly, R is compatible with any operating system, which is also a plus.

34. Name a few packages used for data manipulation in R programming.
    - Dplyr
    - Data.table
    - Tidyr
    - Matrix
    - stringr

35. Name a few packages used for data visualization in R programming.
    - ggplot2
    - leaflet
    - lattice
    - Shiny
    - Plotly