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 is the process of transforming raw data into a more suitable format for analysis. It involves tasks such as cleaning and transforming to address issues like missing values and inconsistencies. The goal is to create a well-organized and accurate dataset, facilitating effective analysis and interpretation.
2. What are the differences between data analysis and data analytics?
   Data analysis is the broad term for looking at the data and coming to conclusions whereas data analytics uses programming to analyze large datasets more accurately.
3. What are the differences between machine learning and data science?
   Data science is the practice of studying data and what it means, while machine learning is the discipline of creating the technology that we use to help us analyze data.
4. What are the various steps involved in any analytics project?
   Identify the problem, collect data, analyze, interpret results
5. What are the common problems that data analysts encounter during analysis?
   Missing values, null values, lack of context in data, or lack of amount of data are all common problems that occur during analysis.
6. Which technical tools have you used for analysis and presentation purposes?
   Python and R for analysis, Tableau for presentation.
7. What is the significance of Exploratory Data Analysis (EDA)?
   The EDA helps give analysts a full understanding of the dataset. EDAs explore the characteristics of the data, importance of certain variables, and distribution spreads.
8. What are the different methods of data collection?
   Surveys, interviews, and experiments are all different methods of data collection.
9. Explain descriptive, predictive, and prescriptive analytics.
   - Descriptive – summary of data
   - Predictive – taking this data to infer what may happen in the future
   - Prescriptive – taking descriptive analytics and determining what to do going forward with what we now know
10. How can you handle missing values in a dataset?

Imputing the median/mean value into missing values.

11. Explain the term Normal Distribution.
    Normal Distribution means there is a bell-shaped distribution curve with few or no outliers.
12. How do you treat outliers in a dataset?
    You should delete them or replace them with the median value.
13. What are the different types of Hypothesis testing?
    The different types of hypothesis testing are simple and composite testing.
14. Explain the Type I and Type II errors in Statistics?
    Type I errors are when the null hypothesis is rejected even though it was true. Type II errors are when false null hypotheses are not rejected.
15. Explain univariate, bivariate, and multivariate analysis.
    - Univariate – Examining one variable at a time
    - Bivariate – Examining two variables at a time
    - Multivariate – Examining three or more variables at a time
16. Explain Data Visualization and its importance in data analytics?
    Data visualization helps us translate our findings to colleagues or whoever may need analysis on what we are investigating.
17. Explain Scatterplots.
    Scatterplots are graphs between to numerical variables that are used to show correlation.
18. Explain histograms and bar graphs.
    Histograms' x-axis contains all numbers while bar graphs' x-axis contains specific numbers.
19. How is a density plot different from histograms?
    Density plots are continuous lines while histograms contain several bins.
20. What is Machine Learning?
    Machine learning is the discipline of creating the technology that we use to help us analyze data.
21. Explain which central tendency measures to be used on a particular data set?
    Median, mean, or mode.
22. What is the five-number summary in statistics?
    Min, IQ1, Median, IQ3, Max
23. What is the difference between population and sample?
    Population is every piece of data while sample is a randomized portion of data.
24. Explain the Interquartile range?
    Middle 50% of data.
25. What is linear regression?
    Linear regression is finding the correlation between the x and y value.
26. What is correlation?
    Correlation is the measurable relationship between two variables.
27. Distinguish between positive and negative correlations.
    Positive – When one goes up, the other goes up, and vice versa
    Negative – When one goes up, the other goes down, and vice versa

28. What is Range?
    Max – Min = Range
29. What is the normal distribution, and explain its characteristics?
    Normal distribution is when the distribution is symmetrical down the middle.
30. What are the differences between the regression and classification algorithms?
    Regression – predicts a continuous numerical variable
    Classification – predicts a category that an object with qualities of the input belong to
31. What is logistic regression?
    Logistic regression uses mathematics to find the relationship between two data factors.
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    RMSE = np.sqrt((np.mean((actual – predicted)**2))
    MSE = np.mean((actual – predicted)**2)
33. What are the advantages of R programming?
    R is best at the statistical tests necessary for predictive modeling and statistical analysis because of its pre-programmed functions.
34. Name a few packages used for data manipulation in R programming?
    Tidyr, data.table, stringr
35. Name a few packages used for data visualization in R programming?
    Ggplot2, lattice, shiny