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

“Data Wrangling” refers to the process of working with and cleaning up messy data.

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

Data analysis is more specifically representing data in a way that is meaningful, and data analytics is more generally the study of data

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

Machine learning deals with algorithms that allow a program to adapt and process new information, while data science is more about working with and understanding data.

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

Hypothesis, data collection, data cleaning, data organization, data visualization

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

Inconsistent data formats in the data they are working with

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

Python, Excel, R

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

EDA is important because it helps us break data into its meaningful components and analyze them.

8. What are the different methods of data collection?

Surveys, experiments, compilation of existing data online

9. Explain descriptive, predictive, and prescriptive analytics.

Descriptive analytics is used to gain insight and information from describing data
Predictive analytics is used to predict how we might expect something to occur
Prescriptive analytics is used to determine an optimal path forward

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

Impute them with the average value.

11. Explain the term Normal Distribution.

“Average” distribution, typically shows as a bell curve on a graph.

12. How do you treat outliers in a dataset?

Ignore them when talking about the general data, but examine them as a phenomenon of their own

13. What are the different types of Hypothesis testing?

Left tailed, right tailed, and two tailed.

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

Type 1 is false positives, where the result is falsely determined to be true when it is in fact actually false.
Type 2 is false negatives, where the result is falsely determined to be false when it is in fact actually true.

15. Explain univariate, bivariate, and multivariate analysis.

Univariate analysis looks at individual variables in a dataset
Bivariate analysis looks at two variables in a dataset and compares them
Multivariate analysis looks at multiple variables in a dataset and compares them as they relate to the larger set as a whole

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

Data visualization is extremely important because it is the process of putting data into visually appealing and easy to understand formats, such as charts and graphs.

17. Explain Scatterplots.

A type of graph where each data point is plotted as a point on a graph. Called a scatterplot because there are no lines connecting the data, since you often have multiple data points in the same general area, giving the plotted data a “scattered” look, hence the name.

18. Explain histograms and bar graphs.

Histograms and bar graphs both use vertical bars to graph data by frequency, but histograms work with quantitative data, and bar graphs for categorical data.

19. How is a density plot different from histograms?

Density plots show proportion of values in the range, while histograms give the counts.

20. What is Machine Learning?

The process by which algorithms are made and fed large amounts of data, to detect patterns and essentially “learn” by analyzing these patterns.

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

Mean and median, though the mean is typically more numerically accurate.

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

Maximum, minimum, Q1, Q3, and median

23. What is the difference between population and sample?

Population is a whole group, while a sample is a smaller set taken from the population

24. Explain the Interquartile range?

The spread of the middle half of the distribution

25. What is linear regression?

A method where we assume the trends of existing data will predict future or unknown points, using the line of best fit as a guide.

26. What is correlation?

Statistical relationship between two variables, not necessarily causal in nature

27. Distinguish between positive and negative correlations.

A positive correlation sees both variables increase as the other does. A negative correlation sees one variable decrease while the other increases

28. What is Range?

The space between the maximum and minimum values of a set

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

Data distributed in an “average” manner, when graphed, shows as a symmetric bell curve.

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

Regression algorithms predict values using existing data, and classification algorithms group things based on existing data.

31. What is logistic regression?

Predicting an occurrence by how close its’s measures scale on a constructed range of 0-1

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

I've always found it by using SQL or R

33. What are the advantages of R programming?

R can display graphs as you code

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

Tidyverse, Metrics

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

Catools, ggplot