Names: David Nguyen, Jason Nguyen, Ryan Tomas

Questions for Assessing Proficiency in Data, Data Preprocessing, and Data Exploration

1. Conceptual Questions

1. **What are the key differences between structured, semi-structured, and unstructured data?**
    a. Structured data: organized data and stored in a relational database
    b. Semi-structured data: not fully organized data but doesn't have a relational database
    c. Unstructured data: no organized data and often requires processing to extract information
2. **Can you explain the difference between categorical and numerical data? Give examples.**
    a. Categorical data represents categories or labels with no inherent numerical meaning. Examples include car brand names or genders, which have nothing to do with numerical meaning.
    b. Numerical data: are measurable quantities that can be discrete like many students.
3. **What challenges do you face when working with time-series data?**
    a. Challenges we can face when working with time-series data include irregular timestamps, such as missing or unevenly spaced data.
4. **What are some common ways to handle missing data? Which method would you use if the dataset has 30% missing values?**
    a. A way to handle missing values is by implementing mean/median/mode or dropping the rows/columns with missing values. However, if the null values are random we can use predictive methods.
4. **How do you detect and handle outliers in a dataset?**
    a. There are a few ways to handle outliers in a dataset we could cap/floor values to see where the average data is in. Also, we can use a z-score method to detect the outliers or we can use boxplots/scatterplots to visually see the outliers.
5. **Why is feature scaling important in machine learning? When would you use Min-Max Scaling vs. Standardization?**
    a. We can't just give all the data to the machine but only key data that will help it learn. We would use min-max scaling for neural networks and use standardization for linear regression.
6. **How do you encode categorical variables in a dataset? What are the advantages and disadvantages of one-hot encoding vs. label encoding?**
    a. The way we encode categorical variables in a dataset is to convert the variables to numerical format. To do so we would have to use one-hot encoding and label encoding. The advantage of one-hot is that it prevents false ordinal relationships but increases dimensionality. Label encoding is simple, and memory-efficient but introduces ordinal relationships.
7. **Explain feature selection. How do you decide which features to keep and which to drop?**
    a. The best way to remove features is to remove irrelevant or redundant features to improve model performance.

9. **How would you check for multicollinearity in a dataset? How would you address it?**
    a. This occurs when many features are highly correlated and we can address it by removing one of the correlated features or combining correlated features into a single meaningful feature.
10. **What is the difference between Pearson and Spearman correlation?**
    a. Pearson measures the linear relationship between two variables but Spearman measures the monotonic relationship between two variables.
11. **Suppose you have a highly imbalanced dataset. What steps would you take before training a model?**
    a. The step we would take before training a model is to understand the data and then use data resampling techniques to make it balances
12. **Explain the significance of PCA in dimensionality reduction. When should you use it?**
    a. This is a technique for reducing the dimensionality of a dataset while preserving as much variance as possible and we would use it in a high-dimensional dataset.
13. **What statistical tests would you use to check if two groups have significantly different means?**
    a. When two groups have significantly different means we would have to do a version of the t-test which depends on the value of the dataset.

2. Scenario-Based Questions

14. **You are given a dataset with missing values in several columns, including age, income, and education level. How would you handle these missing values?**

You handle using the mean, mode, and median. You can decide which to use depending on the missing values. You should use mean for normally distributed data, median for skewed numerical data, and mode for categorical data. This method of using the mode retains the most frequent category.

15. **You notice that some data points in your dataset have extremely high or low values. How would you determine if these are genuine data points or outliers? What methods would you use to handle them?**

    You can compare the existing values to see if the values are either too high or low by looking at the distribution through tools like boxplots, histograms, or even z-scores. By cross-referencing these values with related data, it can ensure the anomalies are valued before the user proceeds to address them.

16. **You are given a dataset with transaction data, including purchase amount, date, and customer ID. How would you create new features to improve a machine learning model's performance?**

Some features we would create or introduce to improve the machine learning performance are average purchase amount, purchase frequency, and time since the last purchase (relating to purchase frequency). This would capture customer behavior patterns and improve machine learning performance by highlighting the patterns and relationships between customers and products.

17. **You are working on a fraud detection dataset where only 2% of transactions are fraudulent. What techniques would you use to ensure your model does not simply predict the majority class?**

    With a dataset with only 2% of transactions are fraudulent we would have to use a resampling technique since fraud cases are rare and we can balance the dataset. Also, we can create features that can improve the detection that can detect anomalies.

18. **You have a dataset with 500 features. Your model is overfitting, and training time is high. How would you reduce the number of features while maintaining performance?**

You have to set the features that are not too complex but are not too simple. To simplify it we can do aggregation (group similar feature), get rid of irrelevant features, We can do subset selection with (relationship or domain “we don’t know what he said”) knowledge,

19. **You are given a new dataset for a classification problem. What steps would you take to understand the dataset before building a model?**

For a new dataset, we would need a starting point which would be examining the structure of columns, data types,and size, and addressing any missing values. We would also utilize tools like histograms and other plots to identify patterns and outliers to understand the relationships within the data. After analyzing this data and validating how it aligns with our expectations, we can start building the model.