Problem 1

1. True, although Pandas can also be used for unstructured data
2. False
3. True
4. True
5. True

Problem 2

1. A, as this function provides basic summary statistics for the dataset
2. B
3. D
4. A
5. D

Problem 3

1.

a. The code organizes the data from the ‘question.csv’ file into a dictionary called ‘data_dict’, grouping values based on the “Group” and “Category” columns. The code then transforms it into a new DataFrame called ‘df_new’ after performing some data manipulation on the set. The data manipulation done has each column represent a category (A,B,C,D,E) and each row represent a group, with corresponding counts of occurrences for each category within each group.The code then calculates and prints the number of rows in ‘df_new’ where the value in the ‘D_count’ column is equal to 0.

b. Yes. A file contains the simplified code

2.

a. File with code is attached, called “T12611201_Midterm1_Q3P2a.py”. I chose to replace empty age with the average age for the dataset

b. Looking at the results, all models have Jack dying and Rose surviving except for SVC, which has both of them dying

c. I think the difference in results is because each model handles the data and uses different classification methods, I mean that’s why different models exist — to provide different methods of classifying the data. For the discrepancy between SVC and the other models specifically, the

difference may stem from model complexity, handling of outliers, kernel choice and parameters, the data's distribution, and hyperparameter tuning.

3.

a.
The average lifespan was 1008 hours, meaning each motor will usually last 1008 hours on average according to the dataset. This statistic gives you a single value that summarizes the central tendency of the data, and is the point around which the entire dataset tends to cluster

b.
In code

c.
5 motors have reached or exceeded the expected lifespan

d.
The standard deviation is 42.374 hours. This statistic gives you a good idea of the level of dispersion and variability in the dataset. It also allows you to find statistical outliers by finding values in the dataset 3 standard deviations from the mean

e.
The dataset appears to follow a normal distribution according to the Shapiro-Wilk test with a 0.05 significance level in Scipy. After running the code, we obtain a Shapiro-Wilk test statistic of 0.9171733260154724 and a p-value of 0.33398160338401794. Since the p-value is more than the alpha value used, the lifespan of the batch of motors appears to follow a normal distribution

Problem IV

1.
Data cleaning is extremely important as it allows you to identify and correct errors or inconsistencies in a dataset. The accuracy and validity of any sort of data analysis is heavily dependent on the quality of the data being used, hence the need for data cleaning to improve the quality and reliability of a dataset as much as possible.

Examples of data cleaning include:
(1) Removing duplicates
This may be used, on, say, data obtained from a store's customer purchase history. Duplicates may arise due to data entry errors from canceled transactions. Or, if you are looking to standardize each customer, multiple entries from the same customer. Removing duplicates ensures each customer is only represented once in the dataset

(2) Handling Missing/NA values
Oftentimes in athlete performance tests, values are either missing or marked as NA if they failed to complete the test or performed it in an invalid manner

2.
Overfitting is an issue where the model learns the training data “too well”, capturing noise or random fluctuations in the data rather than the underlying pattern. This can result in a model which performs well on the training data but fails to generalize unseen data in real-world testing, ultimately culminating in worsened performance.

Some methods to avoid overfitting include cross-validation (multiple subsets for training/testing), regularization, feature selection (retain most important features of a dataset, discard irrelevant or redundant ones), ensemble methods, and early stopping (halt training process when performance on a validation set begins to degrade)

3.
Decision trees essentially break down a task into a “tree”, where each node represents a decision based on the value of a feature and each branch represents the outcome of that decision. Finally, each leaf node represents the final decision/classification

Advantages of decision trees:
- Easy to understand and interpret
- Can handle non-linear relationships
- Can handle mixed data types

Disadvantages of decision trees:
- Prone to overfitting
- Less stable
- Lack of global optimality

Example use-case:
- Credit report assessment, or a medical diagnosis

Random forest uses *multiple* decision trees during training and combines their predictions through averaging or voting to make final prediction. Each decision tree in the random forest is trained on a random subset of the training data (bootstrap sampling) and a random subset of features (feature bagging), introducing randomness to improve generalization performance and reduce overfitting.

Advantages of random forest:

- Improved generalization
- Less sensitive to noise or irrelevant features
- Can provide measures of feature importance

Disadvantages of random forest:
- Too complex
- Less interpretable due to complexity
- Larger memory usage

4.
Matplotlib pros:
- Flexibility
- Widespread adoption
- Integration

Matplotlib cons:
- Steep learning curve
- Verbose/difficult syntax
- Default aesthetics may be unappealing

Suitable Usage Methods:
- Matplotlib is suitable can be used to create many static plots
- It is particularly useful when fine-grained control over plot elements or customizations is required.
- Matplotlib can also serve as the foundation for building more specialized visualization libraries or frameworks

Seaborn Pros:
- High level interface
- Built in themes, color palettes, good aesthetics overall
- Wide variety of statistical plots

Seaborn Cons:
- Less variety
- Limited scope
- Slower performance when handling more complex tasks

Suitable Usage Methods:
- Seaborn is ideal for quickly creating attractive statistical visualizations, especially when exploring relationships between variables or analyzing distributions.
- Canbe well-suited for tasks such as data exploration, hypothesis testing, and communicating insights to a broader audience.

- Seaborn can be used in conjunction with Matplotlib to leverage the strengths of both libraries, combining Seaborn's high-level interface with Matplotlib's flexibility for customization.

5.