Ashanti Benons
2/27/2025

## 1. Dataset Questions

1. The CSV file uses a | as its separator because it incorporates text that contains commas, semicolons and other symbols that are commonly already used as separators for csv files. So a | symbol prevents confusion and ensures accurate parsing of the data.

2. The data comes from the LJ speech dataset, a speech dataset consisting of short audio clips of a single speaker reading passages from 7 nonfiction books. The data set is commonly used in speech synthesis and natural language processing tasks and for training text to speech models.

3. The difference between the two columns is that the text_norm column converts numbers from the text column into words to make it suitable for processing from text to speech.

## 2. Dataframe Practice

1. The row at index 2 and the one at position 2 are not the same because in a dataframe, index and position mean two different things. The .loc[2] accesses the row at the index we labeled 2, while .iloc[2] accesses the row at position 2 in the dataframe, which is at index 9.

| | a | b |
|---|---|---|
| At index 2 → | 6 | 1 |
| At position 2 → | 3 | 99 |

2. Since .loc and .iloc use different types of indexing, row position and column name can't be combined within a single operation.

.loc is label based, where .iloc can only be used with numerical indices.

# 3. Cameras

5. Not all the columns in the original dataset appeared in the averaged Series object because pandas excludes non-numeric columns when calculating averages. This is because the .mean() method cannot compute averages for data types like strings, dates and other non-numeric types.

8. Yes, it is possible that there are missing values anyway. I feel as though they attempted to communicate this by using zeroes for certain features. I noticed that there are zeroes for certain models under columns such as zoom wide and macro focus range. I don't believe the true values are actually zero though.

This can affect the answers I gave because when placeholders are factored in with genuine data, it can dramatically skew statistical results such as averages, medians and standard deviations. The placeholders could have also led me to false conclusions regarding the improvement of certain features over time.

4. Pandas Questions

1. Series in pandas are one dimensional, consist of a single column of data, use a single index and must contain elements that share the same datatype. If mixed types are provided, pandas will upcast them to the most general type. Dataframes differ in that they are two dimensional, have multiple columns that use row and column indexing and they allow different types per column.

2. In series, indexes allow for the retrieval of values using labels. When performing arithmetic operations with other series, matching labels ensure the calculations are accurate.

In dataframes, indexes allow for slicing and subsetting to occur using the .loc[] and .loc[] methods. The .loc[] method can be used with index labels to access rows.

3. The .iloc[] attribute is position based, so it can only be used to access elements by their numerical position. The .loc attribute on the other hand is label based, so index labels have to be used to access elements.

4. In a dataframe, the .loc[] attribute is also label based, but both index labels and column names are used to access data. On the other hand, the .iloc[] attribute must be used with the integer positions for rows and columns.

5. Series are like a dictionary because they store key-value pairs where the index acts like a key and the data acts as the value. They are like lists in that they store an ordered sequence of values, index positions are 0-based and you can use slicing.

6. DataFrames are like a dictionary because they store columns as key-value pairs, where the column names act like keys and each column is a series that acts as the value. They are like 2D lists because they support slicing and indexing in the same way 2D lists do. Elements can be accessed by row and column positions like in 2D lists as well.