Lauren Byrne - C16452764

Type of Data

This dataset includes 16 column features. 9 of these columns are categorical, 6 are continuous, and one is an ID column so will be classified as neither categorical nor continuous. A column is continuous if the values are numeric for example age. There can be an infinite number of values between two continuous values. A categorical value is a distinct category for example gender.

The continuous columns in this dataset are 'age', 'fnlwgt', 'education-num', 'capital-gain', 'captial-loss' and 'hours-per-week'.

The categorical columns are 'race', 'relationship', 'education', 'sex', 'workclass', 'marital-status', 'occupation', 'target', 'native-country'.

Data Quality Issues

There are many potential data quality issues including missing values, outliers, and feature cardinality. I will identify the quality issues for this dataset and suggest what could be done to handle these issues. These issues will be summarised in the data quality plan below.

Missing Values

The columns that contain missing values with the percentage of missing values in each column are 'workclass' – 5.60763, 'occupation' – 5.63025 and 'native-country' – 1.79056.

I suggest **Imputation** to deal with this issue. Imputation replaces missing values with a value that is a measure of the central tendencies of that feature. For categorical data, imputation can be used with the mode as the estimated value. For the column 'native-country', because it only has a missing percentage of 1.79056, another possible solution could be to leave the missing values as is, and document it in the data quality plan.

Outliers

Examining the data shows that the column 'fnlwgt' has a very high maximum value at 1484705 especially compared to its median and 3rd Quartile and this could affect the performance of the model.

I suggest using a technique called **clamp threshold** to handle these outliers. This works by setting a lower and a higher threshold and removing values that are outside of that threshold.

Other interesting outliers are within the columns 'capital-gain' and 'capital-loss'. Both have a large range of values between the 1st Quartile and 3rd Quartile, but have zero values in the Min, Median and Max columns. With no missing values, I wonder if the zero values in the column are invalid data. This is something that will be reported in the data quality plan.

Feature Cardinality:

The following features have a cardinality below 10: 'workclass', 'martital-status', 'relationship', 'race', 'sex', 'target', which is a low cardinality. However, further inspection these columns would have a naturally low cardinality, so this is ok to leave as is.

*data quality plan on next page

Data Quality Plan

The above information has been summarised in this Data Quality Plan

Feature	Data Quality Issue	Potential Handling Strategies
workclass	Missing values (5.60763)	Imputation (mode: 'Private')
occupation	Missing values (5.63025)	Imputation (mode: 'Prof-speciality')
native-country	Missing values (1.79056)	Imputation (mode: 'United-States') or leave as is
fnlwgt	Outliers (high)	Clamp Transformation (threshold to clamp should be discussed with business)
capital-gain	Potential Missing Values in mean, median and mode	Discuss with business on importance of columns to help decide on a decision, potentially drop column if invalid data
Capital-loss	Potential Missing Values in mean, median and mode	Discuss with business on importance of columns to help decide on a decision, potentially drop column if invalid data