Data Quality Report

Missing Values

- Workclass – 5.6%
  - Relatively small percentage missing
  - **Possible solution:** leave it as it is, document it in the data quality plan
- Occupation – 5.6%
  - Relatively small percentage missing
  - **Possible solution:** leave it as it is, document it in the data quality plan
- Native Country – 1.8%
  - Relatively small percentage missing
  - **Possible solution:** leave it as it is, document it in the data quality plan

Outliers

- Capital Gain – max value 99,999
  - Unusually high value when compared to mean
  - Probably invalid data
- Hours per Week – max value 99
  - Unusually high value when compared to mean, median and 3rd quartile
  - Possibly invalid data
- Capital Loss – max value 4,356
  - Unusually high value when compared to mean
  - Possibly valid data as the figure is not like Capital Gain or Hours per Week
- Age – max value 90
  - Unusually high value when compared to mean, median and 3rd quartile
  - Probably a valid outlier

Cardinality

- No obvious problems with cardinality in the dataset

Data Quality Plan

Feature	Data Quality Issue	Potential Handling Strategies
Workclass	Missing values (5.6%)	Imputation
Occupation	Missing values (5.6%)	Imputation
Native Country	Missing values (1.8%)	Imputation
Capital Gain	Outliers (high)	Clamp transformation
Capital Loss	Outliers (high)	Check, maybe clamp transformation
Age	Outliers (high)	Check, maybe clamp transformation
Hours Per Week	Outliers (high)	Clamp transformation