|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 31/10/2021  Machine Learning | |
| Data Quality Report | | | | |
|  | | | | |
| Daniel Krasovski  C18357323 | |  |  |  |

## Missing Values:

There was a few missing values but not enough for it to warrant any action to be take

Chart, bar chart

Description automatically generated

Workplace had 5.6% missing

Occupation also had 5.6% missing

Native Country had 1.8% missing

## Irregular Cardinality

There seemed to be no Irregular Cardinality.

## Outliers

Age- Chart, histogram

Description automatically generated

There was a few with the age of 90 but it does not seem like an input error. Decision: keep it in as it is most likely accurate. There was no other outliers

## Handling Missing Values:

The missing values were assigned as a “?” and I let pandas deal with it automatically by specifying that they were missing values

## Data Quality issues:

There seemed to be a data quality issue with this sample as a majority of the people did not have any capital gain or capital loss. With capital gain having 91.7% zeros and capital loss with 95.4%. the way this could be solved is by getting another sample where there are less 0’s or just remove all the values with 0.

## Correlation:

There was high correlation between values as seen by Spearman’s pChart, scatter chart

Description automatically generated