|  |
| --- |
| [Type the company name] |
| Artificial Intelligence 2 |
| Assignment 1 |

|  |
| --- |
| Christopher Jerrard-Dunne  2/15/2016 |

# Data Quality Report

Given the quality of the dataset, it is sufficient and comprehensive enough to begin further analysis of the data to draw significant data points from. However, various issues within the dataset and the analysis with the tools provided does show issues that will need to be addressed to build a more substantial and comprehensive dataset and resolve issues that many otherwise need to analysed by a data analyst.

It can be observed within the data that the Pandas library detects each object within the dataset as either an “object”, to denote non-numerical data, and “int64” to denote strictly numerical data. This variance is used fairly smartly; however certain criteria can be used or specified to gather more accurate data. Data, such as the wage bracket for each person, could be better and more accurately represented as a continuous feature than a categorical one. Likewise, the individuals educational status could be better represented using a categorical selector, such as “High School”, “University”, or “Doctorate”. Such a change should be discussed, and determined if it is possible to change.

On analysis of the data, it is determined that there are no null values within any of the data features in the dataset. However, on observation within the dataset, it can be seen that various values have been used as a substitute to a null value, and as such have not been picked up using the algorithm. This discrepancy can be commonly seen under the “native-country” feature, where a “?” value is used in place of “NaN” values to denote a lack of data. This discrepancy must be resolved to unify the data and present a more cohesive representation of data, rather than falsely inflating the data as a result. This issue is best resolved to properly format the data as best as possible.

It is my belief that if these data consistency issues can be resolved, a more comprehensive data set can be built and more accurate results can be deduced and reasoned. This would be useful for data analytics experts to build applications to predict and better serve customers for the company. These points should be considered as a group and should be resolved.