**Advanced Intelligence 2**

**CA 1**

**C13446122**

**James Prendergast**

I will describe my analysis regarding the dataset I was given to analyse in terms of missing values, outliers, feature cardinality and my opinions on how to address these quality issues. However, note, I only have concrete data in areas/fields that supplied an integer value as I am not able to get values such as medians nor means using a string. So, that is the reason I will be looking at the fields such as “age”, “fnlwgt”, “education-num”, “capital-gain”, “capital-loss” and “hours-per-week”. The first data analysis I see is that there is someone as 90 as that is the max value from that column while only the mean is 38.5 and the median is 37.5. So obviously this one value alone is throwing off the rest of the age field columns. I would recommend we set a limit or look at data less than the age of 50 in the future to fix this data analysis issue.

The opposite outlier data analysis issue can also be found in “education-num” where the minimum value had thrown off all the other values. The minimum “education-num” value is 1 when both its mean and median are both in the 10.

Once again, I would look at setting a minimum limit to make sure such an outlier would not have an effect on the rest of the column fields.

Then in regard to instances and feature cardinality, the field with the most unique values was “hours-per-week”, which range from 1 hour to 44 hours. This is very interesting in regards to data analysis, in regards to how to make it more unique in the future it would be best to create triggers to make sure there can be an overabundant of unique values.

The data analysis that has the best average values was “education” with the first\_quartile and third\_quartile being quite close to the mean and median.

There is also a problem in regards to missing values in some of the fields. There will needs to better requirement checks to make sure this left out values are not massively important to the rest of the data analysis.