**Analysis of Dataset Features**

**Categorical Features:**

1. Missing Values: **Work Class: 5.608%, Occupation: 5.63% and Native-Country: 1.791%**

* A possible solution for the above three values would be to leave the data as it is, as the missing percentage of each feature is below 10%.
* Another possible solution could be to use imputation to replace the missing values with the mode, most frequent, value.
* In my opinion, the data should be left as it is as the missing percentage is under 10% of over 30,000 records.

1. Irregular Cardinality:

* There are no issues with irregular cardinality in the categorical features however, if the business would like a more specific target number, they could change the target feature to a continuous feature. This will allow for higher cardinality.

1. Outliers:

* There are no issues of outliers in the categorical features

**Continuous Features:**

1. Missing Values: **Capital-gain: 91.739% and Capital-loss: 95.378%**

* There is a very high percentage of missing values for the two above features, both over 90%. Therefore, there are a few solutions available for missing values.
* 1. The features could be removed from the ABT. This would be the approach I would recommend most as there is over 90% of the data in these features missing.
* 2. Complete case analysis could be carried out. However, with this approach most of the data will be lost so I would therefore not recommend this approach.
* 3. Lastly a missing feature could be derived. However, this would not be useful if the features relied on another target feature. These features only rely on each other so this approach would not be useful, and I would not recommend it.

1. Irregular Cardinality:

* There are no issues with irregular cardinality in the continuous features.

1. Outliers:

* **Fnlwgt: max = 1,484,705 3rd Qtr. = 237,319 median = 178,385**
* **Capital-gain: max = 99,999 3rd Qtr. = 0 median = 0**
* **Capital-loss: max = 4,356 3rd Qtr. = 0 median = 0**
* All three of the above features have a high maximum value compared to the median and 3rd Qtr. values. However, capital gain and capital loss have a high outlier because of a high percentage of missing instances, meaning both their median and 3rd Qtr. are 0.
* One option would be to remove the affected row from the ABT, however, that would cause a significant loss in data.
* Another option, which would be the option that I would recommend, would be to use a clamp transformation on capital loss and capital gain. Method 2 would be the best out of the two methods available in this case because, these two features have a mean and standard deviation whereas the 1st and 3rd quartiles, needed for method 1, are 0.
* For the fnlwgt feature, it would be best to document the outlier and handle it later with the business, to ensure the values are correct.