Key Data Wrangling Steps for Income Census dataset

**Data Summary**

The existing data set has 42 columns

1 target column (\_Income\_category)

1 column specifying instance weights – Indicates the population colume the row represents. Not to be used as a model input but as weights

31 Categorical columns

8 Continuous columns

|  |  |  |  |
| --- | --- | --- | --- |
| **Observation** | **Description** | **Columns Impacted** | **Discussion and Actions taken** |
| **Handling Missing Values** | Following 8 columns have missing values , represented by ‘?’ | \_region\_of\_previous\_residence  \_migration\_code-change\_in\_msa  \_migration\_code-change\_in\_reg  \_migration\_code-move\_within\_reg  \_migration\_prev\_res\_in\_sunbelt  \_country\_of\_birth\_father  \_country\_of\_birth\_mother  \_country\_of\_birth\_self | I’d keep the missing value as **‘?’** for now. There is no obvious reason for missing values beyond the fact that responders forgot to enter those columns, ex. Info. Around previous residence and birth country seems missing.  **Action:** Keep the value ‘?’ as of now to review any not so obvious relationships between missing values and other parameters |
| **‘Not in Universe’ values observed** | These 14 columns have values that cannot be categorized and are bucketed into a ‘Not in Universe’ category. | \_major\_industry\_code',  '\_major\_occupation\_code',  '\_reason\_for\_unemployment’, '\_region\_of\_previous\_residence',  '\_state\_of\_previous\_residence',  '\_migration\_code-change\_in\_msa',  '\_migration\_code-change\_in\_reg',  '\_migration\_codemove\_within\_reg',  '\_migration\_prev\_res\_in\_sunbelt',  '\_family\_members\_under\_18',  '\_country\_of\_birth\_father',  '\_country\_of\_birth\_mother',  '\_country\_of\_birth\_self',  'fill\_inc\_questionnaire' | The source doesn’t provide any reasoning for this. Most likely these are driven by incomplete answers or responses that are skipped or not valid. NO obvious correlation observed between ‘Not in Universe’ values across columns, based on simple cross tab analysis  **Action:**  Plan to keep these as it is to analyze any not so obvious relations and allow for appropriate modifications |
| **Outliers** |  | No outliers observed – Data is weighted by population count for each row ~1800 counts associated with each row. No obvious outlier row observed | NO Action taken |