**Name: Richard Kabiru Waithera**

**Student No.: 150684**

**Data Preparation Using Tableau Prep**

1. **Data: Average salaries of managers**

The data was loaded using connections to a file by specifying that it’s an excel file that needs to be connected.

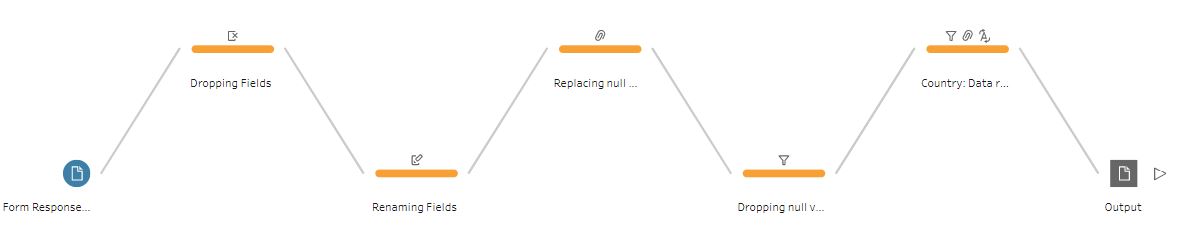
**Data understanding**

The data contains 27904 rows and 18 columns of different data types such as strings, numerics and date and time. The dataset contains null values and columns that can be converted into numerics. The field names/column names need to be summarized. This dataset therefore would require some cleaning before any insights can be gathered from it.

**Data cleaning**

1. Add a clean step by clicking add button.
2. Rename the cleaning step to …
3. Rename the fields into shorter names for easier understanding of column contents.
4. Removing and replacing the null values where necessary. Replace the null values for fields with null values making up less than 1 percent of the column values.

* In the Industry, State(USA) and Highest Level Education fields replace the null values with the string none instead of dropping
* Remove the columns with the more than 70 percent null values i.e. Other currency (if any), Income additional context, job title additional context.
* Remove the rows with null values by dropping the specific rows from the column with the highest null values. The column AVG additional income per year has 26 percent null values. Drop these rows by excluding them from the dataset.
* Exclude the remaining null values in the City, gender and race column

1. Change the data role of Country and City to be recognized as a Country/region and City respectively. Group similar values with the same meaning using manual selection such as USA and US.
2. Finally, Create an output step
3. **IMDB**

Connect tableau prep with the IMDB dataset by connecting as a text file.

**Data understanding**

The dataset has 9 fields and 9,999 rows. On the settings tab select the field operator to be automatic and the text qualifier to be “.

**Data Cleaning**

1. Add a clean step by clicking add button.
2. Rename the cleaning step to drop fields.

* Drop fields with null values exceeding 90 percent i.e. GROSS. This will aid in maintaining the number of rows that do not contain null values in the dataset.

1. Rename cleaning step to drop rows with null values.

* In the remaining columns, drop rows with null values in column with the highest number of null values i.e. RunTime.
* Drop rows with null values in the year, stars and genre field.
* The number of rows remaining is 6,956.
* The numeric columns still have null values that can be replaced.

1. Split text columns containing more than one entry.

* Use custom split and specify the separator is a comma (,)
* The genre column is split into 3 new columns the we drop the Genre column.

1. Trimming spaces

* Trim spaces in columns such as STARS and ONE-LINE
* Split the STARS field using colon (:) to get the roles field of the movies i.e Director or Star
* Remove punctuation and trim spaces in the year field and convert the data type of the column to date.

1. Group null values in the GENRE split columns and replace the null values with None. Then merge the three columns to form one Genre column with one Genre per movie.
2. Group values in the VOTES field and replace null values with 0 votes.
3. Finally, Create an output step

