

Capstone project 1 - Data Wrangling

Two excel files were provided as part of this project. Both the excel sheets had header rows (Column names). One of the excel files had “Data dictionary” set i.e. this file contains more details about the columns involved in the main data set (present in the first excel) and the other file contained the main dataset.

Listed below are the data wrangling steps conducted:

1. Both the files were uploaded into DataFrames in python.
2. Around 53646 record were present in the main data file. There were Nan entries in few of the columns. These Nan entries were replaced with 0
3. Forward / Backward filling was not performed on any missing values in columns as there was no real need at the moment.
4. To find the involvement of “Age groups” in different businesses, data from master dataframe is melted and Pivot table is created with the count of “age groups” and their success / failure rates in the businesses.
5. Also, a bar graph is created to highlight the success rate of people from different age groups.
6. Steps 4 and 5 are repeated to find the involvement of different states/territories in the business areas.
7. A sea born plot is created to analyse the personality of people that are benefitted by the Job assistance program and a pivot table indicates the personality types that are successful and those that are not.