Project: Clean and Analyze Employee Exit Surveys

Guided Project Goal

Project Goal:

- Data aggregation
- Combining of data
- Transformation data
- String manipulation
- Working with missing or duplicated data

Dataset description:

- In this project we will be working with surveys from the employees of the Department of Education, Training and Employment (DETE) and the Technical and Further Education (TAFE) Institute in Queensland, Australia.
- The surveys provide value information regarding the reasons why employees resign or retire.
- We'll pretend our stakeholders want us to combine the results for *both* surveys to answer the following question:

  - Are employees who only worked for the institutes for a short period of time resigning due to some kind of dissatisfaction? What about employees who have been there longer?

Dataset Description

DETE Survey

- ID: an id used to identify the participant of the survey
- SeparationType: The reason why the person's employment ended
- Cease Date: The year or month the person's employment ended
- DETE Start Date: The year the person began employment with the DETE

TAFE Survey

- Record ID: an id used to identify the participant of the survey
- Reason for ceasing employment: The reason why the person's employment ended
- LengthofServiceOverall. Overall Length of Service at Institude (in years): The length of the person's employment(in years)

Identify Missing Values and Drop Unnecessary Columns

Instructions

- Import the pandas and numpy libraries
- Read the dete_survey.csv CSV file into pandas
- Read the tafe_survey.csv CSV file into pandas
- Use the DataFrame.info() and DataFrame.head() methods to print information about both dataframes, as well as the first few rows

Fillna, drop unused columns

- Replace ‘Not Stated’ by NaN
- Use DataFrame.drop() to drop the following columns for tafe_survey: tafe_survey.columns[17:66]
- Use DataFrame.drop() to drop the following columns form dete_survey: dete_survey[28:49]

Clean Column Names

Instructions

- Rename the remaining columns inn dete_survey dataframe
  - Make all the capitalization lowercase
  - Remove any trailing whitespace from the end of the strings
  - Replace spaces with underscores('_')
- As an example, Cease Date should be updated to cease_date
- Update column names to match the names in dete_survey_updated

Filter Data

- Use the Series.value_counts() method to review the unique values in the separationtype column in both dete_survey_updated and tafe_survey_updated
- In each of dataframes, select only the data for survey respondents who have a Resignation separation type
  - Remember that the dete_survey_updated dataframe contains three Resignation separation types. We want to select all of them
  - Use the DataFrame.copy() method on the result to avoid the SettingWithCopyWarning
  - Assign the result for dete_survey_updated to dete_resignations
  - Assign the result for tafe_survey_updated to tafe_resignations

Verify the Data

- Check the years in each dataframe for logical inconsistencies
  - First, clean the cease_date column in dete_resignations
    - Use the Series.value_counts() method to view the unique values in the cease_date column
    - Use vectorized string methods to extract the year
    - User the Series.astype() method to convert the type to a float
  - User the Series.value_counts() to check the values in the cease_date and dete_start_date and dete_resignations and the cease_date column in tafe_resignation

Create a New Column

- Create a new column named institute_service column in dete_resignations
  - Subtract the dete_start_date from the cease_date. Assign the result to a new column named institute_service

Identify Dissatisfied Employees

- View the values in the ‘Contributing Factors. Dissatisfaction’ and ‘Contributing Factors. Job Dissatisfaction’ in the tafe_resignations dataframe
- Update the values in these columns so that each contains only True, False, or NaN values
  - Write a function named update_vals that makes the following changes:
    - If the value is NaN, return np.nan
    - If the value is ‘-’, return False
    - For any other value, return True
  - Use the DataFrame.applymap() method to apply the function above to the ‘Contributing Factors. Dissatisfaction’ and ‘Contributing Factors. Job Dissatisfaction’ in the tafe_resignations dataframe
  - Use the df.any() method as described above to create a dissatisfied column in BOTH the tafe_resignations and dete_resignations dataframes
  - Use the df.copy() method to create a copy of the results, named dete_resignations_up and tafe_resignations_up

Combine the Data

- Add a column named institute to dete_resignations_up. Each row contain the value DETE
- Add a column named institute to tafe_resignations_up. Each row contain the value TAFE
- Combine the dataframes. Assign the result to combined
- Drop all columns with less than 500 non null values. Assign the result to combined_updated

Clean the Service Column

- Extract the years of service from each value in the institute_service column
  - Change the type to ‘str’
  - Use vectorized string methods to extract the years of service from each pattern
  - Change the type to ‘float’
- Map each year value to one of the career stages
  - Create a function that maps each year value to one of the career stages
  - Use the Series.apply() to apply the function to the institute_service column. Assign the result to a new column named service_cat
- Calculate the percentage of employees who resigned due to dissatisfaction in each category