pasa-b-hw1-1

Homework 1 Analysis

Step 1: Importing the data

First we import the necessary python packages. `Pandas` is a commonly used Python package for managing data, `os` will be used for saving files to their respective folders, and `tabulate` will be used for creating tables in terminal to confirm our work.

Once the packages are installed we turn the .csv files into dataframes:

```
enrollment_df = pd.read_csv(enrollment_file_path)
```

```
contract_df = pd.read_csv(contract_file_path, encoding='latin1')
```

Step 2: Merging and Cleaning the data

Once that is finished, an inner merge is conducted on Contract Number/Contract ID and Plan ID:

```
merged_df = enrollment_df.merge(
    contract_df,
    left_on=["Contract_Number", "Plan_ID"],
    right_on=["Contract_ID", "Plan_ID"],
    how="inner"
)
```

After the datasets are merged, redundant columns are dropped and the result is saved to the `output` sub-folder in the `data` folder.

Step 3: Questions and Tables

Now that the dataset is prepared, the assigned questions can be answered and necessary tables can be created.

1. Counting Plans by Type

To create this column, the dataframe is filtered to reflect `Plan_Type` and the pandas command `.value_counts()` is appended.

2. Excluding Unnecessary columns

In order to exclude SNP, EGHP, and 800 series plans, another dataframe is created exluding these columns. This methods ensure that the data is not lost in case it is required later. The table is then regenerated to reflect the redction in `Plan_Types`

3. Average Enrollment by Type

A table displaying average enrollment by type is created. A new dataframe is created with the columns `Plan_Type` and `Average_Enrollment`. The average enrollment column is created by averaging the enrollment by plan type.

Lastly, new csv files are created and added to the output folder.