pasa-b-hw1-1

Homework 1 Analysis

Link to Github

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.

```
/var/folders/7k/7hqtt0l56y3dqfql4pb9y3jr0000gn/T/ipykernel_450/26494230.py:8: DtypeWarning:

Columns (10) have mixed types. Specify dtype option on import or set low_memory=False.

/var/folders/7k/7hqtt0l56y3dqfql4pb9y3jr0000gn/T/ipykernel_450/26494230.py:43: SettingWithCo


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guid
```

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.

```
Plan Count by Type (Original):
```

	Plan_Type	Count
0	Medicare Prescription Drug Plan	991457
1	Local PPO	704993
2	HMO/HMOPOS	479275
3	Employer/Union Only Direct Contract PDP	25630
4	Regional PPO	17578
5	PFFS	13658
6	1876 Cost	7157
7	MSA	6518
8	Medicare-Medicaid Plan HMO/HMOPOS	4130
9	National PACE	1216

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`

```
Plan Count by Type (Filtered):
+---+-----------------------------------+--------+
|   |             Plan_Type             | Count  |
+---+-----------------------------------+--------+
| 0 |  Medicare Prescription Drug Plan  | 269153 |
| 1 |            HMO/HMOPOS             | 36588  |
| 2 |             Local PPO             | 16728  |
| 3 |           Regional PPO            |  8531  |
| 4 |             1876 Cost             |  6329  |
| 5 |               PFFS                |  4232  |
| 6 | Medicare-Medicaid Plan HMO/HMOPOS |  4130  |
| 7 |           National PACE           |  1216  |
| 8 |                MSA                |  232   |
+---+-----------------------------------+--------+
```

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.

```
Average Enrollment by Type:
+---+-----------------------------------+--------------------+
|   |             Plan_Type             | Average_Enrollment |
+---+-----------------------------------+--------------------+
| 0 |             1876 Cost             | 228.1263001485884  |
| 1 |            HMO/HMOPOS             | 848.7377948436643  |
| 2 |             Local PPO             | 310.7412667319621  |
| 3 |                MSA                | 107.7927927927928  |
| 4 |  Medicare Prescription Drug Plan  | 311.75048143967064 |
| 5 | Medicare-Medicaid Plan HMO/HMOPOS |  623.963601532567  |
| 6 |           National PACE           | 139.97652582159625 |
| 7 |               PFFS                | 124.58382066276803 |
| 8 |           Regional PPO            | 201.5029878425716  |
+---+-----------------------------------+--------------------+
```

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