Report on Analysis of Medicare Physician and Provider Data (2013 & 2014)

1. Data Preparation and Cleaning

Dataset Loading and Combination

- Two datasets, representing Medicare Physician and data for 2013 and 2014, were loaded into DataFrames (df_2013 and df_2014) using pandas.read_csv.
- A new column year was added to each dataset to distinguish between the two years.
- The datasets were combined using pd.concat into a single DataFrame (df).

Missing and Duplicate Values Check

- **Missing Values**: A check for missing values was performed using isnull() and sum().
- **Duplicates**: The number of duplicate rows in the combined dataset was identified using df.duplicated().sum().

2. Data Filtering

- Only providers with MD or DO credentials were selected.
- Data was further restricted to providers from the states Rhode Island (RI) and New Hampshire (NH) using filtering conditions.

---

3. Summary Statistics of Charges

The summary statistics for total submitted charges (tot_sbmtd_chrg) and Medicare allowed charges (tot_mdcr_alowd_amt) were calculated:

- Mean and standard deviation were aggregated for each state (RI and NH).

Visualization:

- A bar chart compared the Mean Submitted Charges and Mean Allowed Charges for each state.

---

4. Specialties Analysis

Top 3 Most Common Specialties

- The top 3 most common physician specialties were identified .
- Data was filtered to include only these top 3 specialties.

Proportion Analysis

- The count of doctors for each specialty in each state was calculated.
- Proportions were computed by dividing the count of doctors for each specialty by the total number of doctors in the state.

Visualization:

- A grouped bar plot displayed the proportion of doctors in the top 3 specialties within each state.

---

5. Regression Analysis

A linear regression model was built to predict the Medicare allowed charges (tot_mdcr_alowd_amt) based on various independent variables.

Steps:

1. Categorical variables (state, specialty, and year) were one-hot encoded using pd.get_dummies (with drop_first=True to avoid multicollinearity).
2. Irrelevant or redundant columns were dropped to focus on key predictors.
3. A constant term was added to the model for the intercept.
4. Boolean columns were converted to integers for compatibility with the regression model.
5. The model was fitted using the Ordinary Least Squares (OLS) method from statsmodels.
6. Regression summary output was printed, showing:
    - Coefficients
    - R-squared value
    - Statistical significance of predictors

Specialty with the Highest Charges:

- The specialty with the highest Medicare allowed charges was identified based on the coefficient values of the specialty variables.

---

6. Correlation Analysis Between 2013 and 2014

To examine the relationship between charges submitted in 2013 and allowed charges in 2014:

- The datasets (2013 and 2014) were merged on the provider NPI (National Provider Identifier).

- Correlation between tot_sbmtd_chrg (2013) and tot_mdcr_alowd_amt (2014) was computed.

Visualization:

- A line graph compared Total Submitted Charges (2013) and Total Allowed Charges (2014) across providers.