Homework 2-2

Sarina Tan

The link to my repository:
https://github.com/sarina-tan/HLTH470hw2/tree/main

1. How many hospitals filed more than one report in the same year? Show your answer as a line graph of the number of hospitals over time.

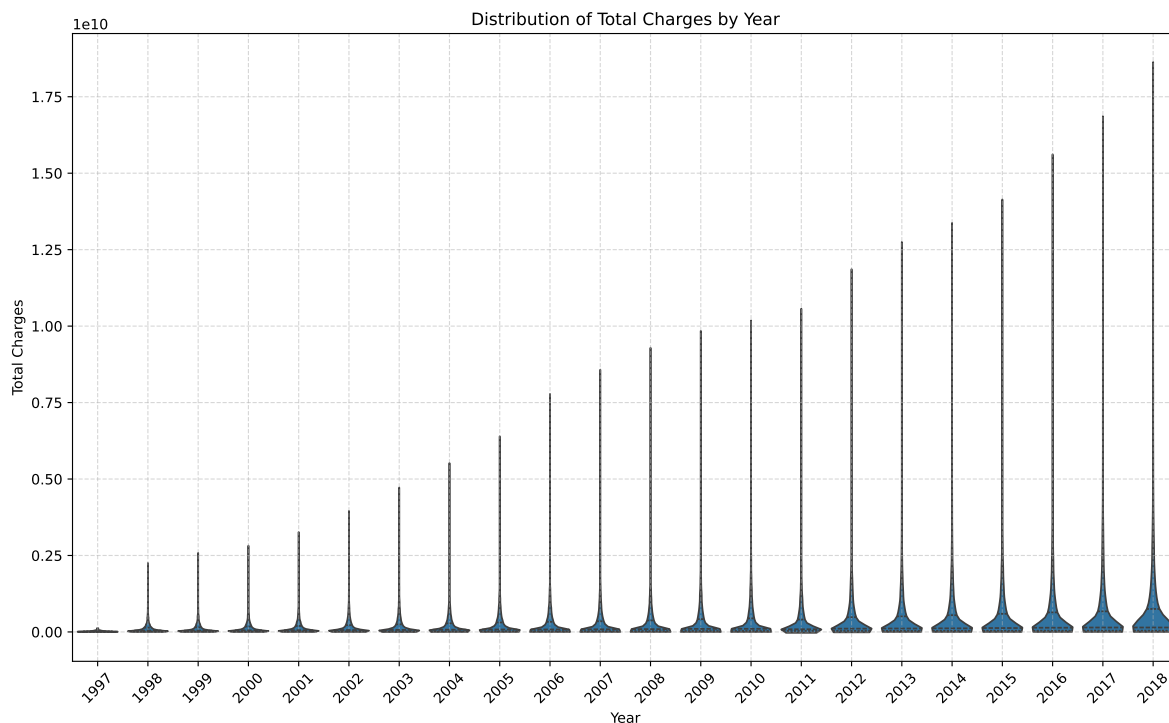Number of Hospitals Filing More Than One Report per Year
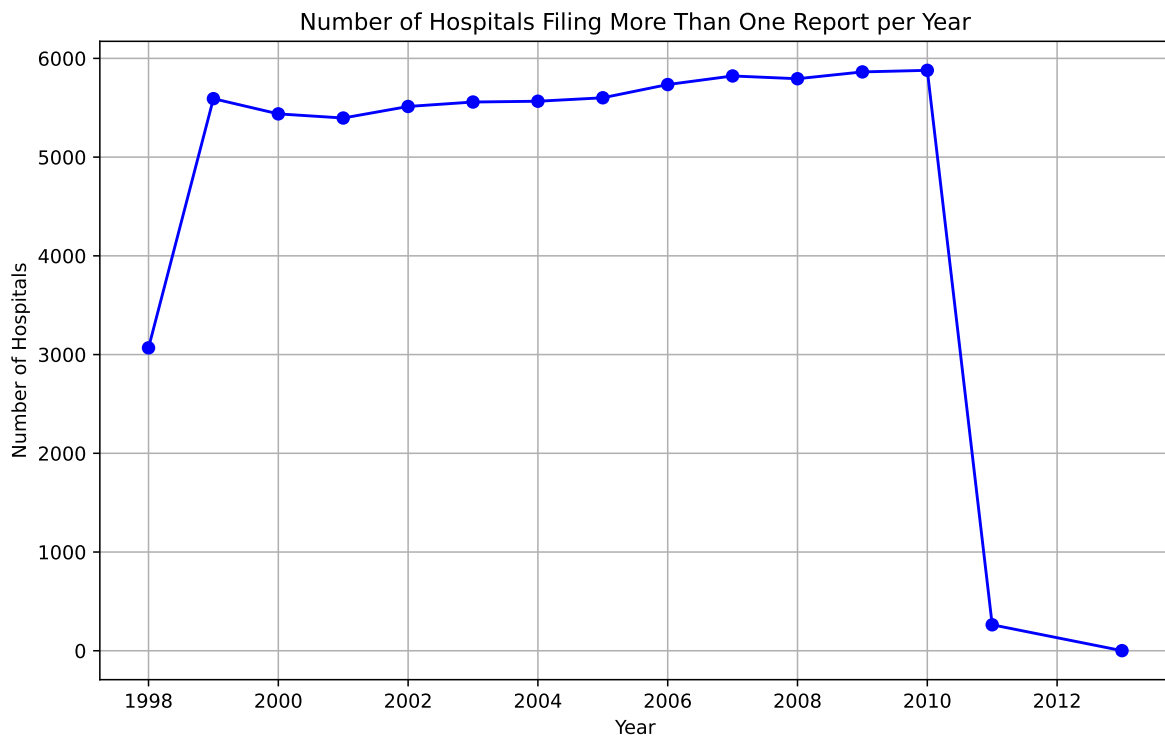
2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

71089
9323

3. What is the distribution of total charges (tot_charges in the data) in each year?

Distribution of Total Charges by Year

5. Calculate the average price among penalized versus non-penalized hospitals.

```
penalty
False    9386.853470
True     9914.349854
Name: estimated_price, dtype: float64
```

6. Split hospitals into quartiles based on bed size. To do this, create 4 new indicator variables, where each variable is set to 1 if the hospital's bed size falls into the relevant quartile. Provide a table of the average price among treated/control groups for each quartile.

```
penalty              False        True
bed_quartile
1             7604.578635   6611.755708
2             8376.977522   8965.532347
3             9679.023341  10554.510254
4            11916.116003  12435.120588
```

Table: Average Price by Treatment Status for Each Bed Size Quartile

Bed Quartile	Control (No Penalty)	Treated (Penalty)
1	7604.58	6611.76
2	8376.98	8965.53
3	9679.02	10554.51
4	11916.12	12435.12

7. Find the average treatment effect using each of the following estimators, and present your results in a single table:

Nearest neighbor matching (1-to-1) with inverse variance distance based on quartiles of bed size

Nearest neighbor matching (1-to-1) with Mahalanobis distance based on quartiles of bed size

Inverse propensity weighting, where the propensity scores are based on quartiles of bed size

Simple linear regression, adjusting for quartiles of bed size using dummy variables and appropriate interactions as discussed in class

Nearest neighbor matching (1-to-1) with inverse variance distance based on quartiles of bed size

```
Nearest Neighbor Matching ATE: 1042.10
```

Nearest neighbor matching (1-to-1) with Mahalanobis distance based on quartiles of bed size

```
Nearest Neighbor Matching (Mahalanobis) ATE: 4415.49
```

Inverse propensity weighting, where the propensity scores are based on quartiles of bed size

```
Mean Price (Treated): 11169.65
Mean Price (Control): 9420.51
ATE (IPW): 1749.14
```

Simple linear regression, adjusting for quartiles of bed size using dummy variables and appropriate interactions as discussed in class

Final summary table

	Estimator	ATE Estimator
0	Nearest Neighbor Matching	1042.102151
1	Mahalanobis Distance Matching	4415.493659
2	Inverse Propensity Weighting	1749.135702

8. With these different treatment effect estimators, are the results similar, identical, very different?

With these different treatment effect estimators, the results from nearest neighbor matching and inverse propensity weightinge are similar in magnitude but not identical.

Nearest neighbor matching vs Mahalanobis distance matching: I would have that these methods should yield comparable estimates since both match hospitals based on bed quartiles, but Mahalanobis distance accounts for correlations between variables, which probably led to the drastic difference.

Inverse Propensity Weighting adjusted for differences across all hospitals rather than selecting specific matches, so it might produced a slightly different ATE from nearest neighbor matching

9. Do you think you've estimated a causal effect of the penalty? Why or why not? (just a couple of sentences)

Overall, I do not think I have estimated a causal effect of the penalty. Hospitals receiving penalties may differ systematically from those that do not. Even though matching and regression techniques to control for bed size were used, unobserved confounders could still bias the results. A more rigorous causal analysis would probably require an instrumental variable or a randomized design.

10. Briefly describe your experience working with these data (just a few sentences). Tell me one thing you learned and one thing that really aggravated or surprised you.

My experience working with this data was a bit frustrating. The data took a long time to load onto my laptop as well as processing to make the new cleaned csv files. One thing that I learned is that with a lot of data, there are also a lot of blanks that need to be filled in and/or removed while merging files together. While I was able to make the final HCRIS data pretty smoothly, it was aggravating to then see that there were still blanks and spots that said NaN that made me unable to analyze it.