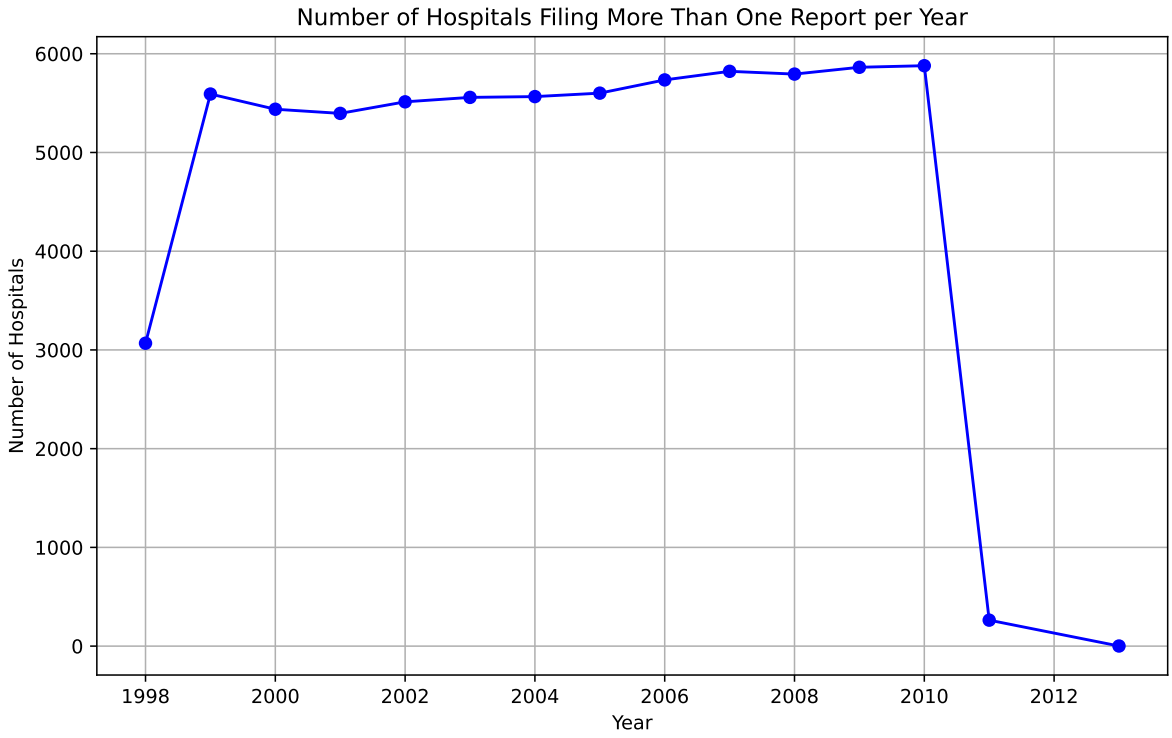
Homework 2-3

Sarina Tan

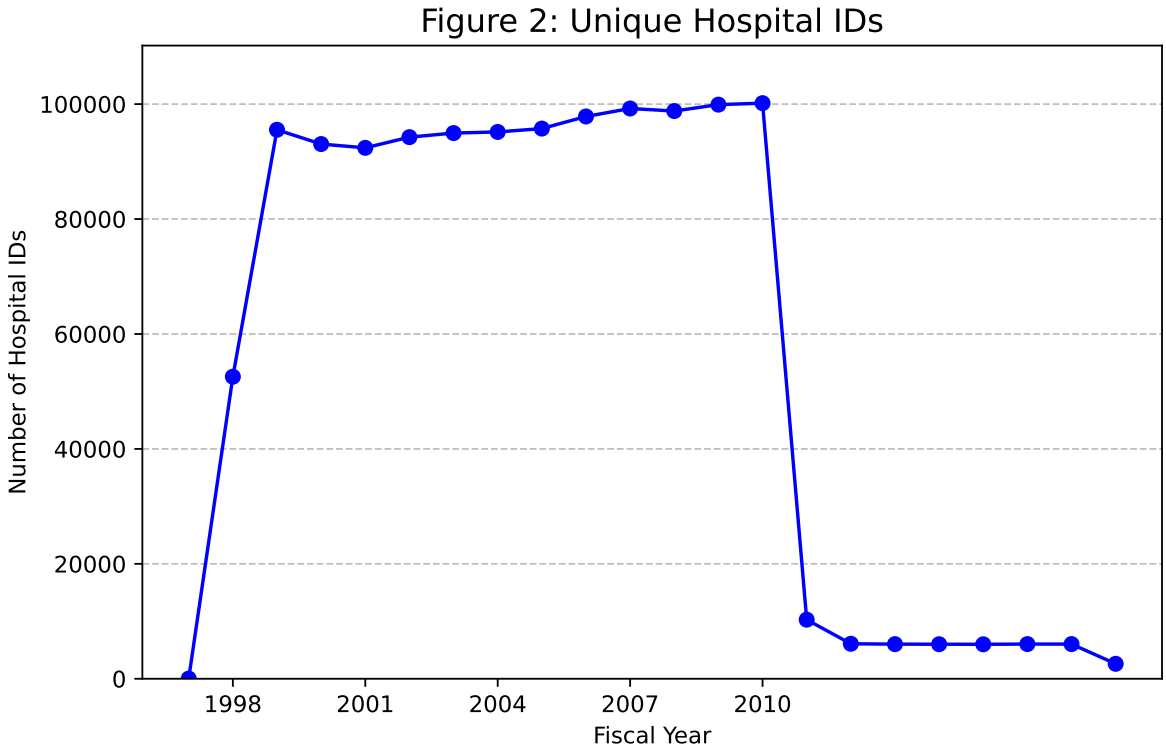
The link to my repository:
https://github.com/sarina-tan/HLTH470hw2/tree/main

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.

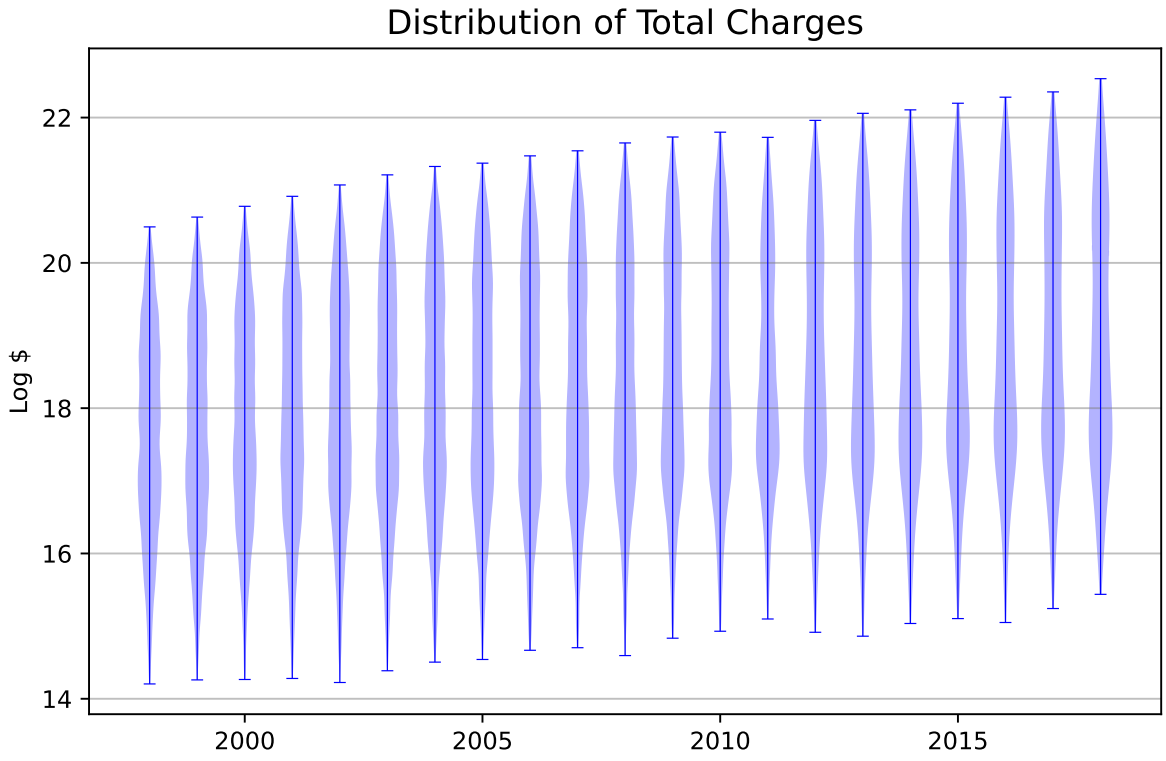


2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

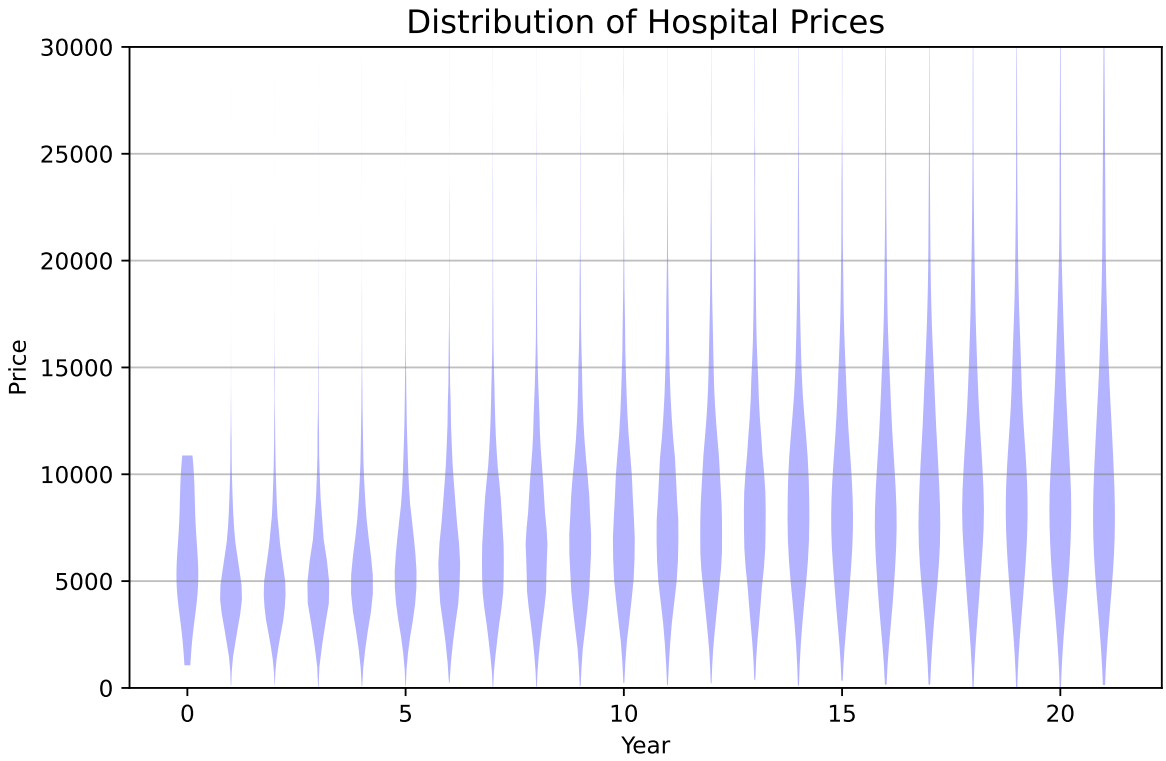
```
The number of unique hospital IDs is: 9323
```



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



4. What is the distribution of estimated prices in each year? Again present your results with a violin plot, and recall our formula for estimating prices from class. Be sure to do something about outliers and/or negative prices in the data.



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.

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

```
<causalinference.causal.CausalModel object at 0x147c198d0>
```

Average Treatment Effect Estimates

```
|     |   NN-INV |   NN-MAH |     IPW |     OLS |
|-----|----------|----------|---------|---------|
| ATE |  246.915 |  246.915 | 246.915 | 246.915 |
| SE  |  493.912 |  493.912 | 444.988 | 444.769 |
```

	NN-INV	NN-MAH	IPW	OLS
ATE	246.915	246.915	246.915	246.915
SE	493.912	493.912	444.988	444.769

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

With these different treatment effect estimators, the results of the average treatment effect (ATE) were the same for all four estimators. The standard error across the four estimators slightly vary, but are still close. Nearest neighbor matching with inverse variance and nearest neighbor matching with Mahalanobis distance resulted in the same average treatment effect and standard error.

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.