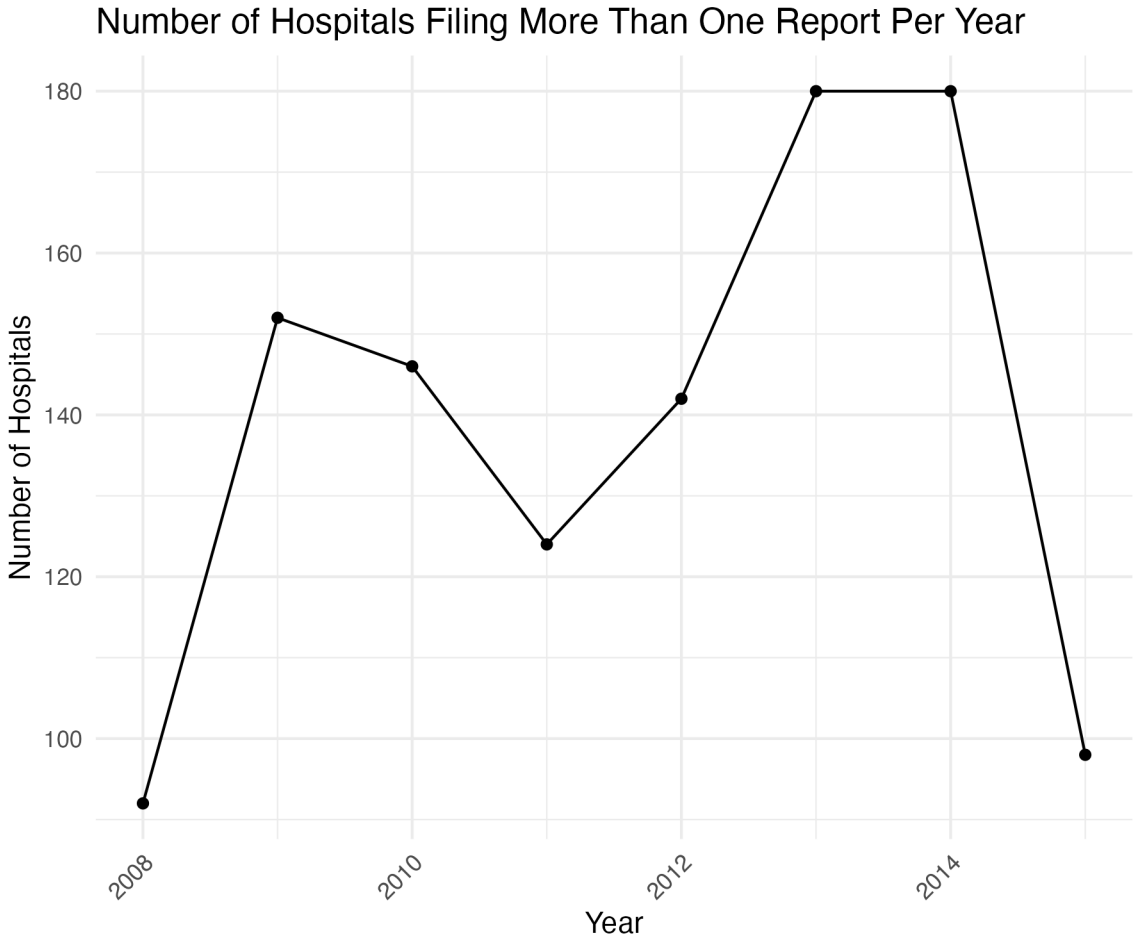
Homework 2

Sammy Ramacher

Building the Data

Answer the following based on our initial, simplified dataset of enrollments, plan types, and service areas:

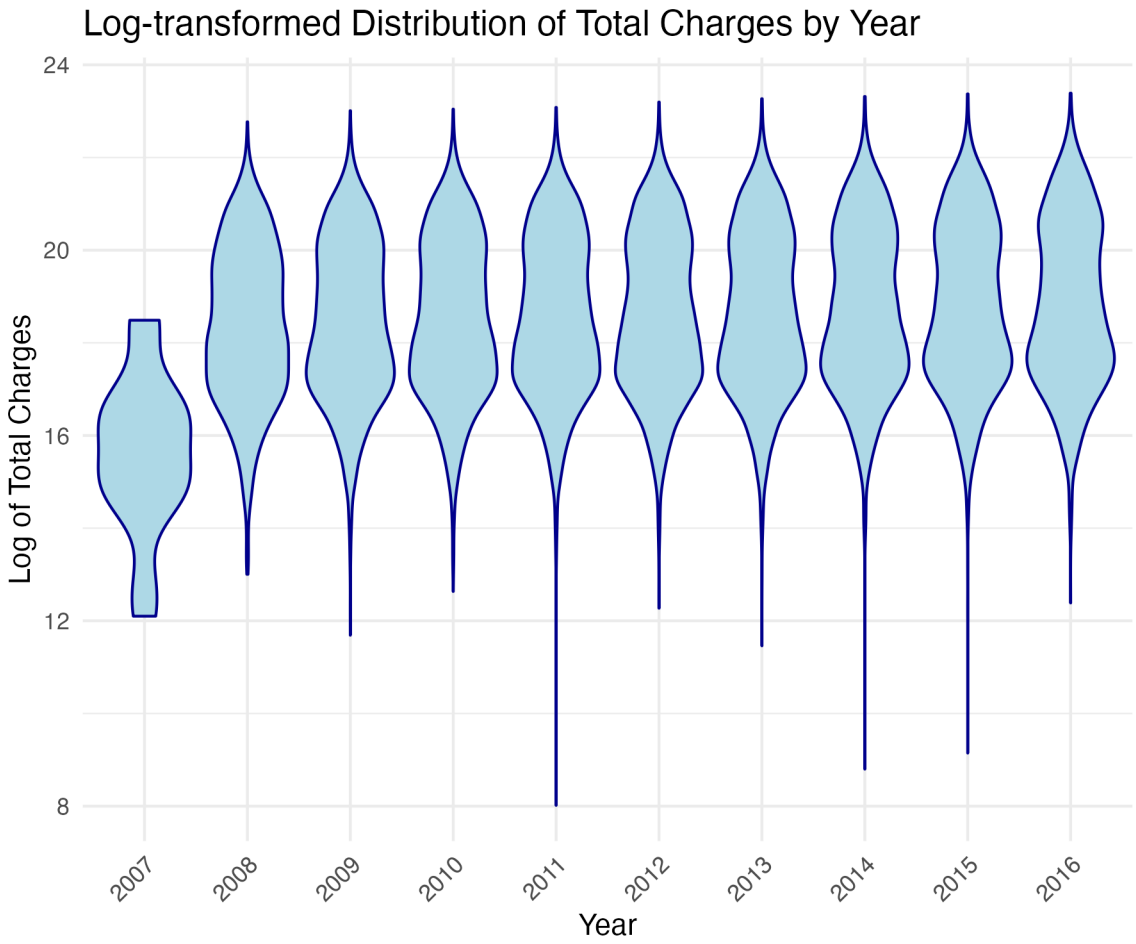
1. How many hospitals filed more than one report in the same year? Show your answer as a line graph of the number of hospitals over time.



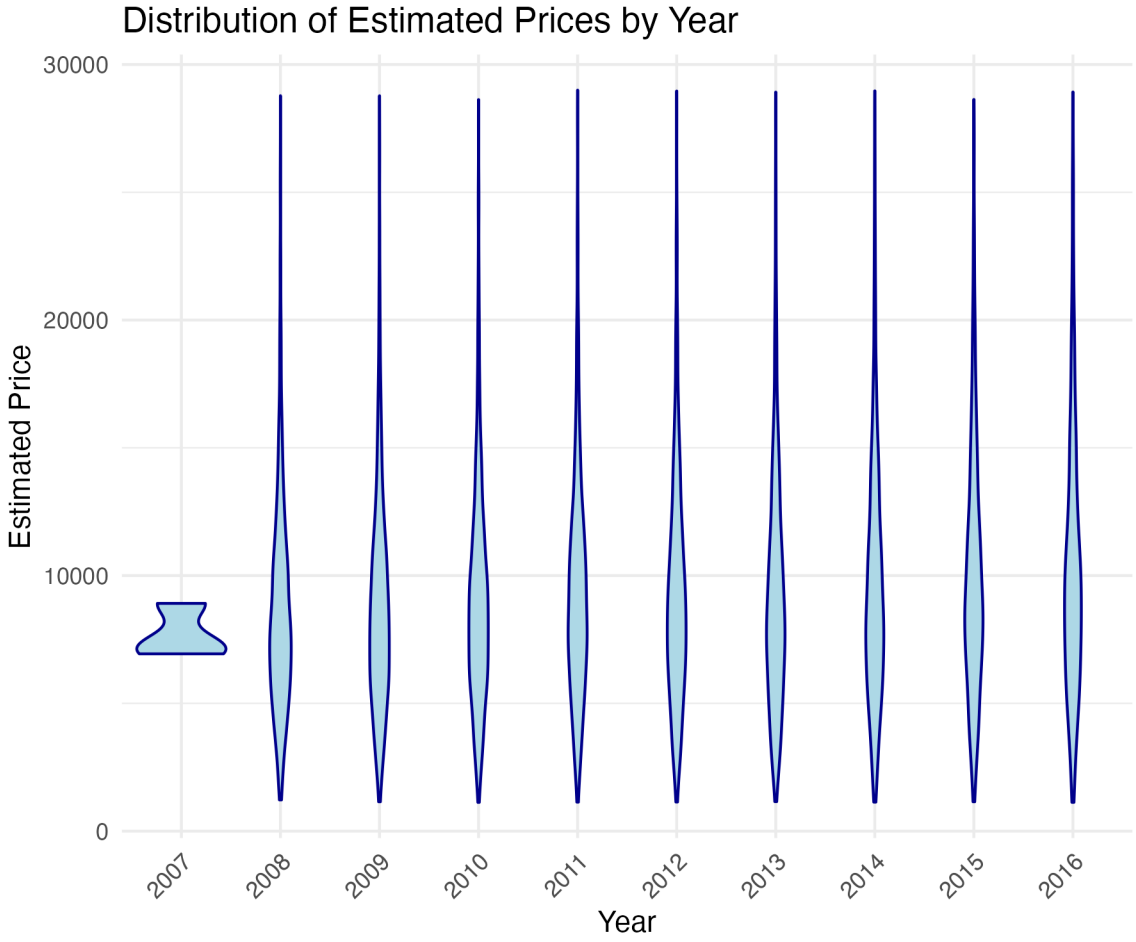
2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

year	num_unique_providers
2,007	16
2,008	3,525
2,009	6,100
2,010	6,103
2,011	6,097
2,012	6,140
2,013	6,066
2,014	6,064
2,015	6,042
2,016	2,650

3. What is the distribution of total charges (tot_charges in the data) in each year? Show your results with a "violin" plot, with charges on the y-axis and years on the x-axis. For a nice tutorial on violin plots, look at Violin Plots with ggplot2.



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.

The average price among penalized hospitals is 9685.11. The average price among non-penalized hospitals is 9323.93.

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.

	bed_quartile	penalty_FALSE	penalty_TRUE
1	Q1	7562.219	7530.446
2	Q2	8336.514	9592.664
3	Q3	9563.395	11026.711
4	Q4	11933.160	12882.861

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

	Method	ATE	SE
1	Nearest Neighbor Matching (IV)	459.1942	231.6536
2	Nearest Neighbor Matching (Mahalanobis)	459.1942	231.6536
3	Inverse Propensity Weighting	905.2288	NA
4	Simple Linear Regression	NA	184.8561

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

The results are the exact same between the first two groups (nearest neighbor matching with inverse variance and nearest neighbor matching Mahalanobis). However, I was not able to obtain a standard error value for inverse propensity weighting, or an average treatment effect for the simple linear regression model. The ATE for inverse propensity weighting is mugh igher than the nearest neighbor matching, and the SE for the simple linear regression is slightly lower.

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

I do not think I have fully estimated the causal effect of the penalty because I did not control for enough confounding variables when running the models. There are likely to be other factors influencing price, so running analyses of only penalty status against price will not produce a causal effect.

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.

Working with this data definitely has a learning curve. Because I am also using this dataset for my thesis analysis, I had some experience working with it previously, but even just the results of the clean data set can be difficult to produce. One thing I learned was the importance of matching data types when trying to merge data sets, such as making sure that everything is a string or numeric value so that it is read properly while merging datsets. One thing that was especially aggravating was properly cleaning the data so that it combined the alpha/numeric files with the report files, while still including data values for the alpha/numeric variables. I was stuck on this issue for a long time, both for this assignment and my thesis. Overall, I have a much better understanding of running econometric analyses on large, complicated datasets, which is helpful for any type of data analysis or research that I pursue in the future.