HW9

Rhea Toves

10/26/2021

Instructions - Problem 1 is associated with your personal dataset project. Submit your work for all problems in a single .pdf compiled with knitr. The Homework Data folder has a starting .Rmd template with spaces for you to fill in your answers to the questions.

_

Problem 1 - This question is related to your personal data set project. The goal for the end of this week is to have at least one column of data uploaded as a .csv to your github repository.

(A) Access at least one column of data (with at least 5 rows), either from one of the sources you identified in problem #1 of your week 7 assignment or another source relevant to your proposed topic.

- I wanted to make sure I found datasets I would be able to easily download and access later in the process of the project. With that being said, the website Kaggle provides datasets of any variety. I was able to find a dataset, “Top 100 Most Streamed Songs on Spotify” (2021) with 10 rows and 14 columns of data.

(B) Write a brief paragraph describing why you decided on this data source, where the data comes from originally, how the chosen data will help you answer a question you are interested in (not necessarily one of the questions you wrote about in Week #7 but hopefully something related), and any processing steps you applied to the data. Submit the paragraph as a response to this question - you do not have to post this to github.

- The dataset, “Top 100 Most Streamed Songs on Spotify” (2021), provides great insight for this project. I’m interested in answering the analysis questions of (1) How many listens have there been from 2015-2021? (2) What is the top genre listened to? How does this vary in the U.S.? (3) How much revenue did Spotify generate in 2020? - and this is a good starting point. The dataset I chose discusses the top 100 songs streamed on Spotify, this can explore questions relating to the most popular genre of 2021, popular artists, songs, etc.

(C) Upload the data column to your github repo from this project (from week 5) and submit the url as a response to this problem.

- https://github.com/rheatoves/DATA-115 (https://github.com/rheatoves/DATA-115)

_

Problem 2 - In your own words, please write brief answers to the following:

(A) What are the assumptions for linear regression from a statistical perspective?

- The assumptions for linear regression are linearity, homoscedasticity, independence, and normality.

(B) What is the definition of a residual?

- Residual are the differences between what we expected based on the model and what we actually observed in the data to start with.

(C) What is the coefficient of determination?

- The coefficient of determination captures the amount of variance and how much of our values are explained in the vertical line.

(D) What is the difference between the total sum of squares and the model sum of squares?

- The total sum of squares determines the variation and the model sum of squares determines how well the model is modeling the data.

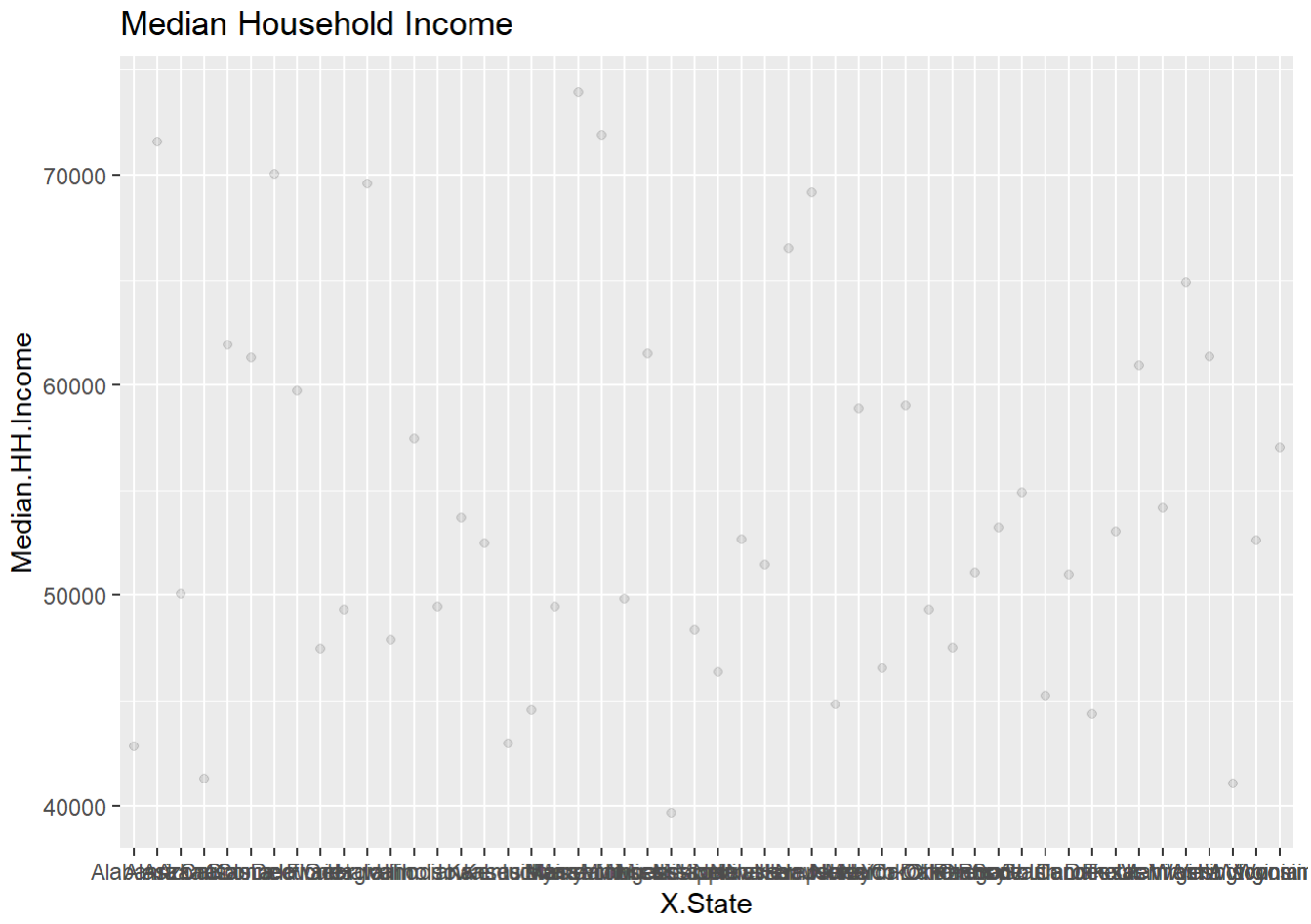
(E) What do the slope and intercept of the best fit line tell us about the data?

- The slope and intercept of the best fit line can help us predict data outside of our dataset.

_

Problem 3 - This problem checks your understanding of linear regression and diagnosis of these fits. Start by loading in the education_income.csv file as a dataframe.

(A) Make a scatterplot of the percentage of BS holders by state against the median household (HH) income.

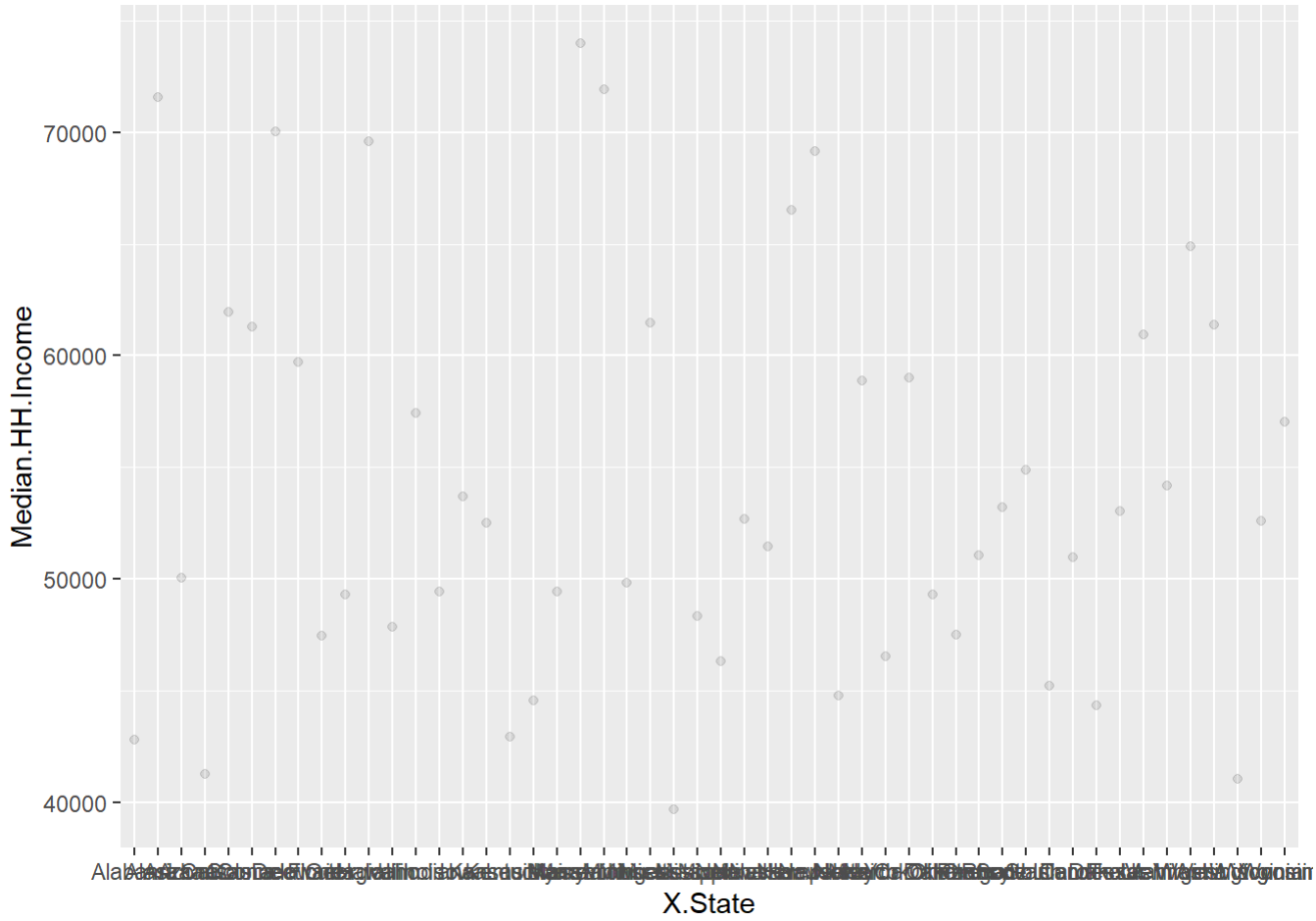


(B) Based on this plot, do you think linear regression is appropriate to attempt?

- The data does not seem to be flowing together, with this graph the data points are more sporadic and everywhere. Therefore, no I do not think that linear regression is appropriate to attempt.

(C) Fit a simple linear model to this pair of columns.

(D) Overlay the best fit line on the scatterplot from part (A).



(E) Write the linear equation estimated by your fit.

- ŷ = bX + a

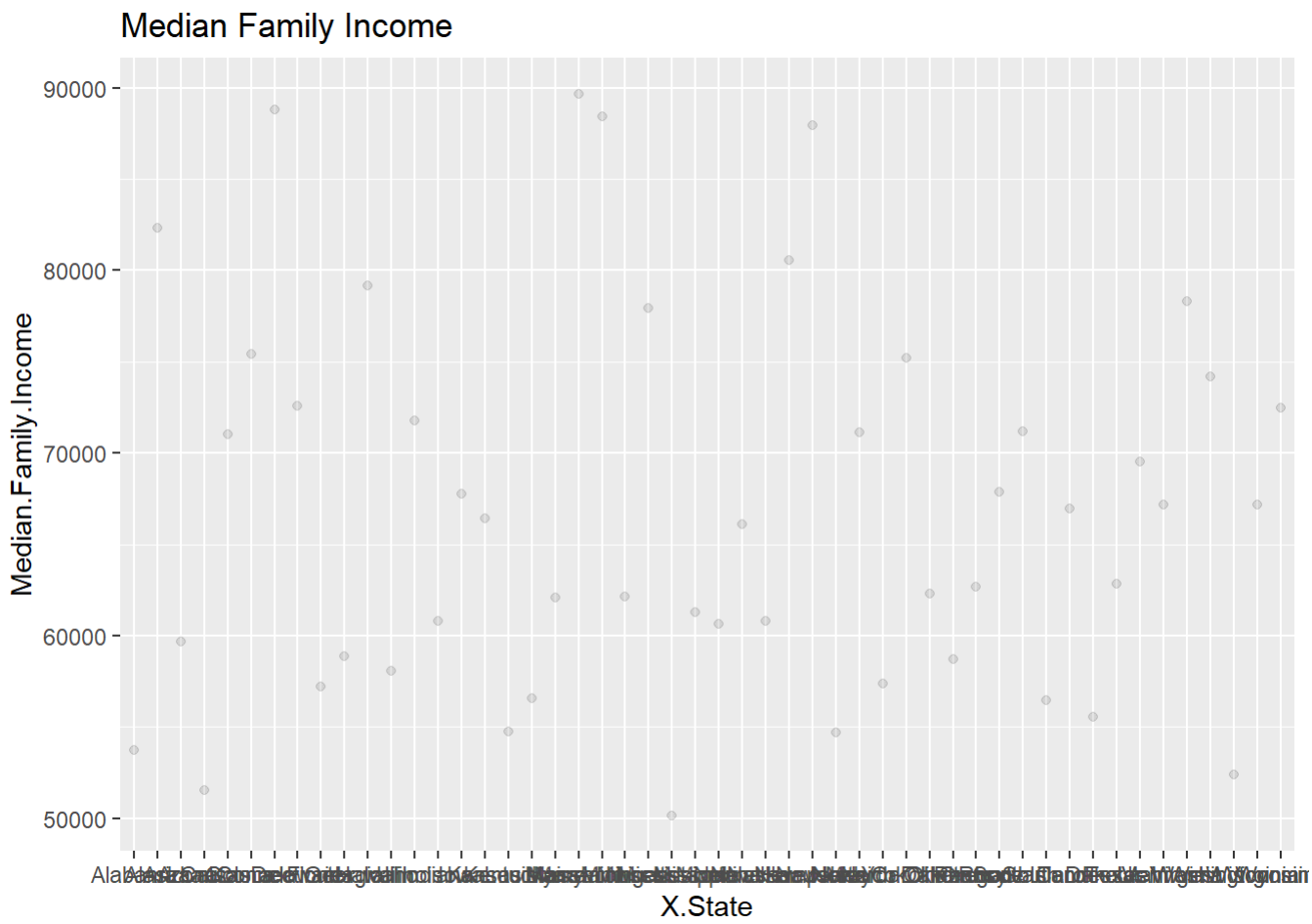
(F) Write the coefficient of determination for your fit.

- R^2 = (educationdf$Median.HH.Income - mean)/(educationdf$Median.HH.Income - mean)^2

(G) Make a QQ plot of the residuals to check if they are normally distributed. What do you conclude?

(H) Plot the residuals against the fitted values. Do you notice anything concerning?

(I) Choose two other columns from the dataframe and repeat steps (a)-(h).



(J) Do you observe any differences in the main results for these new columns?

- When looking at the median family income graph looks similar to the median household income graph.

–

Problem 4 - Read the following exploratory data analysis project about traffic and COVID: https://arxiv.org/pdf/2009.04612.pdf (https://arxiv.org/pdf/2009.04612.pdf) and write a paragraph summarizing your responses to the following questions.

- This article analyzes data with the question, “How politically involved was the spread of COVID-19?”. The author does a great job in diving into the importance of the initial map at the start of the article. There are other points to consider, like travel patterns in each state (which the author looks into). The data does seem to be public, and we could reproduce the cleaning and processing steps in another analysis. We dive into clusters of information and analyzing this information. The final conclusion of the article speak on wrapping up the information discussed previously and conclude that there are a lot of factors that are similar between the Red and Blue states.