Homework 2

[ECON 4753] — *University of Arkansas*

Data Analysis Project

This assignment will summarize the first part of the course. We have primarily studied how to use forecasting methods for cross-sectional data (i.e. data about individuals measured at a point in time). For this assignment, you will be asked to perform a lightly guided analysis of your choice of one of 5 possible datasets. I will give you some prompts to make sure you use the methods we covered in the class.

You should think of this as a project where you are trying to learn something about your dataset. A key takeaway from this class will be to learn how to formulate questions about data available to you and use data to try and answer your question. That is, you are going to be use a dataset about a topic that (may) interest you and I want you to try and learn something new using data.

You will be graded, in part, on how professional your results look. All graphs should be labelled clearly and you should write in complete sentences in your writeup. I have provided a template that will hide your code in your results document that you submit (in the real world, your boss is not looking for your code).

Semi-guided questions

Please work through these questions and write up your results. See the template for an example of how I want this to look.

1. *Exploring a Single Variable:*

   Choose one continuous variable from the dataset and explore its distribution. What can you tell me about its distribution? Try visualizing the data using a histogram and/or calculate the mean and variance.

2. *Investigating Relationships Between Two Variables:*

   Pick two continuous variables from the dataset: one outcome variable and a predictor variable. You should have a motivation for predicting one variable using another (do not just pick two random variables). You should write out a motivation stating your rationale for the variables you are choosing.

   How do the two variables relate to one another? Create a scatter plot to visualize their relationship and/or estimate the correlation. What sort of relationship do you observe? Do you think a linear approximation does a good job? If you feel it's appropriate, run a bivariate regression to quantify the relationship. What insights can you draw?

3. *Reflecting on previous question*

   In class we talked about the importance of multivariate regression. Specifically, we were nervous that the regression coefficient from a bivariate regression might 'pick up on' the influence of other variables that are correlated with our X variable. For example, we discussed regressing wages on years of schooling and talked about the challenge that arises in the this bivariate regression because years of schooling are often correlated with family income.

   I want you to reflect on the regression you ran for question 2 and discuss some potential variables that might be correlated with X *and* have an effect on the outcome variable. Please discuss how you think that other variable might influence your bivariate regression coefficient.

4. *Analyzing a Binary Indicator Variable:*

   Identify a binary indicator variable in your dataset (e.g., gender, yes/no decision) and a continuous outcome variable to predict. How does the outcome variable differ based on this indicator? Compare the means by value of the indicator (0 or 1) using a bivariate regression where the outcome variable is regressed on the indicator variable. What conclusions can you draw?

In this exercise, you can *create* a binary variable using your data. For example, it is common to create a binary variable for being above or below the median value:

`df$above_median_x = df$x >= median(df$x, na.rm = TRUE)`.

5. *Analyzing a Discrete Variable with Multiple Categories:*

   Select a discrete variable with multiple categories (e.g., education level or income bracket). How does the outcome variable vary across these categories? You can create multiple histograms (one per value of the variable) or run a regression using separate indicator variables for each category. What differences, if any, do you observe between categories?

Datasets

You have the choice of one of five datasets. You can see the brief overview of each as well as the *data dictionary* that describes each variable in the dataset by clicking on the links.

1. AirBnB listings in Nashville is a scrapped dataset from the website Inside Airbnb. It contains information about the listing as well as reviews received. I downloaded the current "snapshot" for Nashville, Tennessee.
2. The NBA 2K25 dataset contains information about all the players in the NBA 2K25 video game. It contains info on the player (team, salary, years in nba) and their ratings.
3. College Scorecard is a dataset created by the US Education Department. The dataset was created to allow for potential college students to easily understand the cost-benefit of different colleges.
4. iMDB collects user reviews of movies as well as some basic info of the film.
5. The CPS contains survey data on workers collected by the US Government. This has characteristics of workers and their wages.