RA Task

Andreas Kraft

University of Chicago Booth School of Business

This exercise is meant to assess your programming, statistics, and economic reasoning abilities. The task should take you about 1-2 hours, but there is no time limit.

You should submit a PDF with your answers to the questions, including any figures and tables. You should also submit a separate PDF file with your code. Your code should be well-commented, so that it is easy for someone else to understand the operations that you perform. You should code in Python or R; please explicitly state which language at the top of your code.

1 Data and Goals

There are three CSV files containing raw data on vehicle purchases in Texas, obtained from the Texas Department of Vehicles. Some variables in the data have been changed or removed to preserve anonymity, but the vehicle ID variable is a unique identifier for each vehicle. Importantly, the dataset has also been reduced to only include observations that had sales prices **between $10,500 and $13,500** and vehicle sales that **reported an odometer reading**. The three files are:

1. purchases.csv, which includes a unique vehicle ID, the acquisition cost the car dealership paid to acquire the vehicle, and the acquisition date.
2. sale.csv, which includes a unique vehicle ID, the sales price the car dealership received when selling the vehicle, and the date of the sale.
3. vehicle_information.csv, which includes the unique vehicle ID and an ID for each make, an ID for each model, and the model year.

Goal: Consumers often exhibit left-digit bias and treat numbers right below a round number significantly differently than those right above the threshold. In this task, you are going through some data cleaning and visualization steps that preceded the analysis that actually considers left-digit bias.

2 Summarizing Data and Preliminary analysis

1. Create a clearly labeled table of summary statistics that describe the data. These should include the mean, standard deviation, min, and max of the sales price, the acquisition cost, the model year, and the odometer reading.

2. There is a uniform tax rate on vehicle transactions in Texas added to the sales price. Assuming all vehicles in the dataset are priced at or above the *Standard Presumptive Values* and sold without a trade-in, generate a new variable that denotes the price of each vehicle, including the tax.

3. Generate a histogram of the sales price (including tax).

4. Generate a histogram of the number of used vehicles by model-year. Please carefully explain what you observe in this figure and explain what we can learn from it.

3 Firm Profits

1. To obtain a measure of operating profit for each vehicle, generate a variable that measures the difference between the sales price and the acquisition cost. Present a density plot that presents the operating profit for the vehicles in the sample.

2. Run a regression of the operating profit on the odometer variable. Create a clearly-labeled table with the regression output and briefly interpret the coefficients. Is this a causal effect?

4 Final Discussion

1. Suppose that you want to use this dataset—expanded to include all prices rather than just those between $10,500 and $13,500—to examine the role of left-digit bias in relation to odometer readings in vehicle transactions (e.g., for analysis in this paper: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4945815`). Based on the data analysis above, what did you learn about the data quality? Do you have any concerns, or does the data appear suitable for analysis? If you do have concerns, what additional checks would you want to perform before proceeding?