Week 1 Exercises: Statistics for Data Science

Part 1: Library Circulation

The file physical_circulation.csv contains the total physical circulation (checkouts and renewals) at each branch of the Nashville Public Library.

Read in this file as a dataframe named *circulation*.

1. Find the following:

- mean circulation
- median circulation
- min, max, and range
- the 10% trimmed mean
- standard deviation
- 25th and 75th percentile and the interquartile range

2. Plot the distribution of circulation numbers. Describe the shape of the distribution. Is is symmetric? left-skewed? right-skewed? How many modes does it have?

3. Would you consider any branches to be outliers?

Part 2: Metro Government Salaries

The file General_Government_Employees_Titles_and_Base_Annual_Salaries.csv contains the base annual salary of all Metro Government employees.

Read in this dataset as a dataframe named *salaries*.

1. What percentage of employees are full time?

2. Create a bar chart showing the number of employees per department. Which has the largest number of employees?

3. What is the most common job title for metro employees?

4. What are the mean and median salaries?

5. Plot the distribution of the dataset via both a histogram and boxplot. Is the data symmetric? skewed? How many modes does it have?

6. Create a boxplot showing the distribution of salaries across the different categories of employment status. What do you notice? Which employment statuses have higher salaries on average? What can you say about the variability of salaries across employment statuses?

7. Find the standard deviation of salaries, and use this to compute z-scores for each observation.

8. Find the interquartile range of salaries.

9. Use either your answer for 7 or 8 to hunt down any observations you might consider to be outliers? What do you find?

Part 3: Conceptual

1. In what scenarios might the mean of a dataset be significantly lower than the median? Can you come up with any examples of such a distribution?
2. Can you come up with an example of a distribution that has a median of 0 but a nonzero mean?
3. Consider the distribution of number of Twitter followers per Twitter account. Would you expect the interquartile range of the number of Twitter followers to be (a) about half as large as the range (b) almost as large as the range (c) much smaller than the range?
4. You are analyzing the housing market in Davidson County. You notice that there are some very expensive homes and are afraid that including these homes might skew your analysis. You are considering dropping the top 1% and bottom 1% of the observations prior to your analysis. What impacts might this have on your analysis? What alternatives do you have is you do not wish to drop any observations?
5. You are analyzing daily stock market returns. You are considering dropping the top 1% and bottom 1% of the observations prior to your analysis. What impacts might this have on your analysis?
6. Say you are interested in studying commute times. You are looking at the daily commute times for two different people. Person A commutes from Murfreesboro into Nashville. Person B commutes in the opposite direction, from Nashville to Murfreesboro. If you look at the distribution of their daily commute times over a one-year period, which would you expect to have a larger standard deviation and why? Assume that they both leave for work around rush hour every morning.