CAPSTONE 1 / SUBMIT 4 - INFERENTIAL STATISTICS
CITIBIKE FOR AN OLDER GENERATION

At this point, you have obtained the data set for your Capstone project, cleaned and wrangled it into a form that's ready for analysis. It's now time to apply the inferential statistics techniques you have learned to explore the data. For example, are there variables that are particularly significant in terms of explaining the answer to your project question? Are there strong correlations between pairs of independent variables, or between an independent and a dependent variable?

***Submission**: Write a short report (1-2 pages) on the inferential statistics steps you performed and your findings. Check this report into your github and submit a link to it. Eventually, this report can be incorporated into your Milestone report.*

PROBLEM: As part of a health and lifestyle initiative, many want to look at how different age groups spent time more on Citibike and whether one group has longer or shorter trips than others. The goal would be to encourage older members of the community to use Citibike and show the benefits of cycling as an alternative means in traveling.

ANSWER: After looking at the box plot, one observation was how across different age groups they had the same amount of trip length. In other words, those who were young adults had a similar trip duration as those who were old adults. They differed only by a minute but they averaged around 10 minutes. This is noteworthy because this could encourage other older groups to use the bike.

The steps heading to a legit hypothesis test included:

- **Empirical cumulative distribution function** across age groups: this result showed that four of six groups were similar to box plot. Only the younger group and the unknown (customers) were outside the trip length. The idea would go to several routes:
    - Explore how the older adults would do versus the entire population.
    - Explore more into the unknown group's trends and whether to profit from them more.
    - Explore marketing for younger adults to use the Citibike provided these.

- The **hypothesis test** decided was to see whether the older adult population had the same trip length as the rest of the population. This was to expose the need for more people from this age group to leverage the Citibike system as a healthy and affordable alternative. Here, the x represents the sample mean.

$$H_o : \bar{x}_{OA} = \bar{x}_{AA}$$
$$H_A : \bar{x}_{OA} \neq \bar{x}_{AA}$$

- The first step was to use the **Kolmogorov–Smirnov test** to compare the overall sample from August 2016 from the older adults from the same sample. This two-sample test generated a test statistic of 0.0317 and a p-value of 1.141e-81. Since the older adults come from the same distribution, the values are practical.

- Creating a set of **bootstrap replications** 50 times created a sample mean with the older adult population generated a skewed-right distribution of the average mean and a 95% confidence interval of [964.617,

996.480]. In other words, this is 95% confident that the mean of older adults will reside between 964.617 seconds and 996.480 seconds. Versus the overall population, this generated a normal distribution with a 95% confidence interval of [965.132, 996.801] seconds for all riders. In practical terms, the older adult population would generate a longer trip duration than the overall population. However, they overlap significantly to realize that they are near similar. This may support not rejecting the null hypothesis. This encourages more testing, but it does lean towards the null hypothesis as true.

- **Bootstrap hypothesis testing**: Performing a permutation test between its original sample of 1.5 million trips and the older adult population sample yield a p-value of **0.998**. This may need to be looked at, but **the test FAILS to reject the null hypothesis**. In other words, out of 10,000 replications of reshuffling and checking for the mean, the replicates' mean values all were way below the observed difference of means, 77.56 seconds. Even with an absolute value provided a p-value of **0.9788**.

Further statistics: After looking at the statistics, this may now expand towards using larger data sets. From here, one may compare whether the older population now is utilizing the cycle more than the older population a year ago! This also shows that the older adults are utilizing and traveling the same length as the rest of the population. If this was a case, maybe a focus towards the young to engage them more with CitiBike would be beneficial as they could find creative ways to work with the bike share and convince their parents to subscribe them to the bike sharing service (especially at $200 a year).

Notebook w/ code: https://github.com/tiadvani/sb-capstone1/blob/master/sb_citibike_simple_inferential.ipynb