Overview

Business Problem: Using Kaggle datasets, determine whether Apple Store apps receive better reviews than Google Play apps, or vice versa.

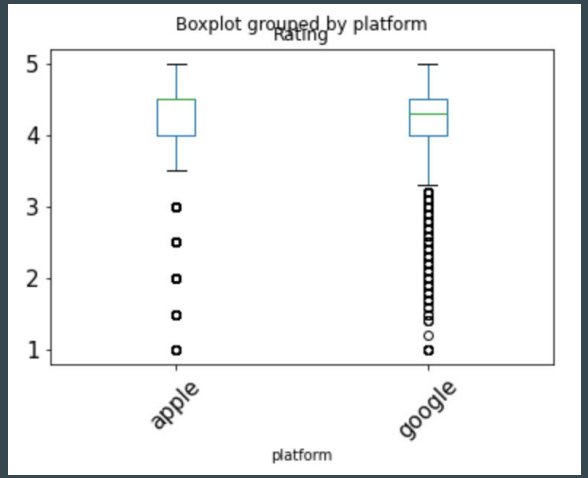
Step 1: Cleaning/Transforming

- Loaded Kaggle datasets into Python using the pandas library
- Selected relevant columns of data: ratings, reviews, categories, and prices
- Checked and fixed data types
- Added a ‘platform’ column to both dataframes and changed column names to match in preparation to join
- Joined the two data sets by appending one to the other
- Eliminated the NaN values
- Filtered only those apps with 1+ reviews

Step 2: Visualization

- Summarized the ‘Rating’ data analytically and visually, grouped by ‘platform’
  - `df.groupby(by='platform')['Rating'].describe()`
  - `df.boxplot(by='platform', column ='Rating')`
- Results: the observed difference between Apple and Google Play reviews is 0.14206. Modeling will determine whether this is significant and/or due to the platform.

	count	mean	std	min	25%	50%	75%	max
platform								
apple	6268.0	4.049697	0.726943	1.0	4.0	4.5	4.5	5.0
google	9366.0	4.191757	0.515219	1.0	4.0	4.3	4.5	5.0



Step 3: Modeling

- Hypothesis formulation:
  - Hnull: the observed difference in ratings is due to chance
  - Halternative: the observed difference in ratings is due to the platform
- Analytical and visual analysis of distribution: Using the `stats.normaltest()` method and a histogram, we can see that the distribution of both data sets is non-normal
- Permutation test (a non-parametric test)
  - Method: compare permutation difference (calculated by shuffling ratings column and keeping platform column the same 10000 times, calculating average of difference in means) to observed difference - use significance level of 0.05
  - Results: The null hypothesis is false; the platform impacts ratings

Findings

The difference between average ratings for Google Play apps and Apple apps is statistically significant. Our official business recommendation is to strike a deal with Google Play Apps.