**Project Title:** Zillow Sum Game

**Team Members:** Joe Casino, Alex Koynoff, Maralee Capella

**Team Name:** Data JAM

**Project Description/Outline:**

* Compare historical home prices with various other variables to see if there is a correlation impacting the decrease/increase of home values.
* Questions to answer:
  + Have home prices increased over time and does higher income generally mean higher home prices?
    - Information used (state level):
      * Median home prices
      * Median household income
  + Does count and distance of various data points impact home price?
    - Data points used (zip code level):
      * Starbucks
      * McDonalds
      * Airports
      * Malls
      * Trailer Parks
  + Does demographics impact home prices?
    - Demographics used (zip code level):
      * Median age
      * Poverty rate
      * Unemployment rate
      * People with Bachelor’s Degrees

**Data gathering:**

The main sources of data that we used were Zillow, Census Bureau and Foursquare API, with the addition of a few websites for reference data.

From Zillow, we were able to download the historical home prices in an easy to use spreadsheet. For the reference variables, gathering the data from the Census Bureau’s API was not as easy, but overall, the API was easy to work with. It also has fairly relaxed limits, which made it less worrisome to run into issues with API usage limits. Foursquare was a new API that we have not used before, so it was more difficult to dissect and understand initially. Once it was set up, gathering the data points needed became easier. We had to be more careful on how data is pulled from Foursquare as it groups data points with subgroups. For example, when pulling information for airports, it could also pull its subgroups, such as number of terminals into the total count of “airports. That could overstate the count if not careful.

**Data cleanup and exploration:**

The majority of the data clean up and exploration was done in Python utilizing Jupyter Notebook and the various functionalities such as Pandas. Using Python provided more flexibility with larger datasets for example, merging different datasets and doing calculations. However, at first, we took small samples of the datasets and put them in Excel. That made it easier to utilize Excel’s chart function for faster toggle between charts to see which would better represent the data. For ease and speed when using small sample data sets, we found that Excel is still easier to work with due to its familiarity and user interface.

**Data analysis:**

During the analysis stage, we utilized Matplotlib to generate various graphs to review any possible relationships between the data points.

**Discussion:**

The findings to our questions were mixed. There was a clear indication that higher income results in higher prices. However, the results in analyzing the home price increases compared to various data points such as count of Starbucks, etc, did not clearly paint a correlation based on the initial volume of data used. Some observations from the analysis were:

* Higher income showed that in general, it meant higher income prices. This was expected and it was shown in the graphs.
* Too many zip codes were used for home prices. This did not provide clear indication that there were correlations with various variables that were used. Challenges from that were:
  + Large number of zip codes was too broad and it would need additional research and data to clean up the information. For example, there are alot of zip codes that cover a very small area, which could result in skewed data per zip code when comparing to the count of Starbucks for example. Other zip codes cover big areas, so we might not capture all the Starbuck coffee shops. For example, if a zip code that covers a very small area might pull in Starbucks shops from neighboring zip codes, while the large area zip codes might not capture all of them due to its size.
  + The large number of zip codes also provided challenges with the API limits. Due to this, we had to take a small random sample from the list of zip codes in order to stay within the limits. That could always causes issues if the random sample size does not provide a good representation of the overall dataset.
* When comparing the price increases for all the zip codes in the data sets to the various Census demographics, we were only able to see slight correlations. Based on the graphs, it could be that using such a large number of zip codes made the comparison too broad. It might be more beneficial to do these comparisons by metro area. This will allow for a more concentrated analysis that could provide better correlations.

**Data Sets used:**

* <https://www.zillow.com/research/data/> to obtain historical home value information
* Census APIs to obtain various demographics
* Foursquare APIs to obtain info on various variables, including count, distance, etc

**Other sources used:**

* Median Household Income - <https://www.americashealthrankings.org/explore/annual/measure/Medianincome/state/ALL>
* Zip Code Coordinates - <http://download.geonames.org/export/zip/>