**Data Set Analysis and Question Forming**

**By: Marcus Fardal, Scott Rippke, Patrick Turvin, Josh Wheaton**

**Date: 11/18/2015**

Here is a list of the different open data sources our group has found:

* Quandl.com
* Socrata.com
* Data.gov
* Google.com/publicdata/directory
* zillow.com

While searching through these open data sources we’ve found a variety of data sets that we took interested in. These data sets include:

1. National Marriage and Divorce trends in the United States. This is actually two data sets that we are going to compare, one data sets contains the total number of divorces per total population, while the other one contains the total number of marriages per total population. We thought this data set was very interesting because marriages ending in divorce is a very common thing in our world today and we thought this was something a lot of people could relate to. These data sets only deal with the total number of marriages and the total number of divorces per year.
2. National home sales by area. This data set contains the yearly annual total sales of home prices in the United States and the four regions (Northeast, Midwest, South, and West). We thought this data set was interesting because it gave us a general idea of which areas were popular by homeseekers. This data set contains the total annual home sales data for the past six years.
3. iPhone app submissions and their delay of approval. This data set contains three separate data sets: one telling us how many apps are submitted per month, one telling us how many apps are currently active (available for download) on the marketplace, and the last one stating the average amount of time an app is waiting before it gets approved. This data set contains information from the past six years. This data set interested us because we are living in an age where technology is always growing and the amount of new apps submitted a day is a clear sign of this.

Of these data sets we decided to focus on the iPhone app submissions for three reasons:

1. The first reason is because of the large quantity of data it had. When we were searching for data sets to use we found a lot of data sets that had less than ten cells in the excel document, while this data set had over 100 cells. For this reason we thought that if we chose the larger data set we could gain more accurate information in order to answer the questions we decided to focus on.
2. The second reason we chose this data set was because all the information was broken into three different data sets. We thought it was interesting that the creator of this data set decided to break this information into three different parts when they could’ve easily combined all of them into one. Along with this we also were curious with whether or not we would be able to use pandas to iterate through all three data sets and try to compare all the information. We still don’t know the exact syntax we would have to use to get this to work, but we are searching for it online and will hopefully be able to figure it out before part 2 of this project.
3. The final reason we chose this data set was because we thought this data set provided us with more interesting questions then the other data sets did. For this data set we we could do questions along the line of which season had the most app submissions or do developers create apps more in the spring than they do in the fall? Along with this we were also hoping to discover more about the application lifecycle. Some questions we came up with because of this were: Is there a limit to the number of apps that can be placed on the store? What is the average life expectancy of an app? And how many apps must be removed from the apps store, before a new app can be added?

As of right now we have decided to focus on two questions for the iPhone app submission data set:

1. The first question is: “How does the quantity of apps submitted change as the years progress?” Knowing, and understanding, this submission rate trend over time could help Apple gain the ability to strategize for future changes/improvements to the App Store. In order to do this we will calculate and assign total app submissions into their respective months and visualize this data over time.
2. The second question we are focusing on is: “Which month (throughout time/on average) has the highest number of apps submitted to the App Store?” We will use the information (total app submissions per month) found in the previous question and bring monthly data together to form our answers. This information will help Apple plan to allocate resources for approval processing and visualize/fill holes in the App Store where new apps are needed.

In order to answer these questions we will have to break down each question and manipulate the data to use in our calculations.

1. For our first question we will have to subtract the previous month’s “total app submissions” from the current month’s total. We will have to repeat this process for each month so we will have to use a for loop within our code. This will give us the number of app submissions within the current month for each month. After we have the total number of app submissions per month, year by year, we can compare this data over time and visualize the trend.
2. For our second question we can utilize the data (total app submissions per month) found in the previous question. We will take this data and sum each month together for the past 7 years. After, we will average this monthly data using numpy functions.

Once the actual coding process beings we will take the url of the data set from Quandl and run it through anaconda with this line of code:

“response = urlopen('https://www.quandl.com/api/v3/datasets/148APPS/APPS\_SUB.csv')”.

Once the data set is properly ran through anaconda we will transfer the information from the data set into a data frame with this line of code: “df = pd.read\_csv(response)”. Once the data set is read into the data frame we will use the data from the columns “Month” and “# Total” to answer our two questions stated above. If all goes according to plan we will be able to answer our two questions with ease, and be ready to present our findings to the class.