INSY 5378 DATA SCIENCE PROGRAMMING APPROACH

GROUP 7

POKEMON GO

**Team Members:**

Salman Mohammed-1001398053 [salman.mohammed@mavs.uta.edu](mailto:salman.mohammed@mavs.uta.edu) Information Systems

Sujatha Sivakumar-1001359237 [sujatha.sivakumar@mavs.uta.edu](mailto:sujatha.sivakumar@mavs.uta.edu) Information Systems

Tapan Patel-1001450804 [tapansaileshbha.patel@mavs.uta.edu](mailto:tapansaileshbha.patel@mavs.uta.edu) Information Systems

Table of Contents .

**Introduction …………………………………………………………………………………….. 3**

**Web Scraping …………………………………………………………………………………… 3**

**Data Organization ……………………………………………………………………………. 3**

**Data Exploration ………………………………………………………………………………. 4**

**Prediction and Modelling …………………………………………………………………. 6**

**Deep Learning …………………………………………………………………………………… 9**

**Insights ……………………………………………………………………………………………… 15**

# Introduction

Pokemon Go is a location based augmented reality game. It was developed for IOS and Android devices by the collaborative efforts of Niantic and Nintendo in July 2016. The App is only for selective countries. The game deploys the usage of GPS to capture, battle and train virtual creatures.

In Our project, we have extracted the ratings for the app in App store and play store (IOS and Android) from HTML links, created pandas data frames, explored the data and performed prediction modelling using Linear Regression, Ridge, Lasso and Deep learning using Tensor Flow.

# Web Scraping

We have scrapped web data from HTML links using the Python module Beautiful Soup. The HTML links were obtained from 3 month’s data of the App’s App store and Play store web pages. The links were downloaded every 10 minutes. However, while extracting the data we found that few of the links were corrupted. In such cases, we replaced the unknown values with zeros for convenience.

The data consists of 12 parameters inclusive of date. They are:

* Date
* Android Average Ratings
* Android File Size
* Android Ratings 1
* Android Ratings 2
* Android Ratings 3
* Android Ratings 4
* Android Ratings 5
* Android Total Ratings
* IOS All Ratings
* IOS Current Ratings
* IOS File Size

# Data Organization

The Data organization step involves three steps:

* The extracted data from the HTML links were converted into a nested dictionary with Date and Time as key and a dictionary of parameters as value.
* The dictionary was then converted into a pandas data frame for the ease of exploration.
* The data frame was finally saved into 3 file formats viz – JSON, CSV and XLSX

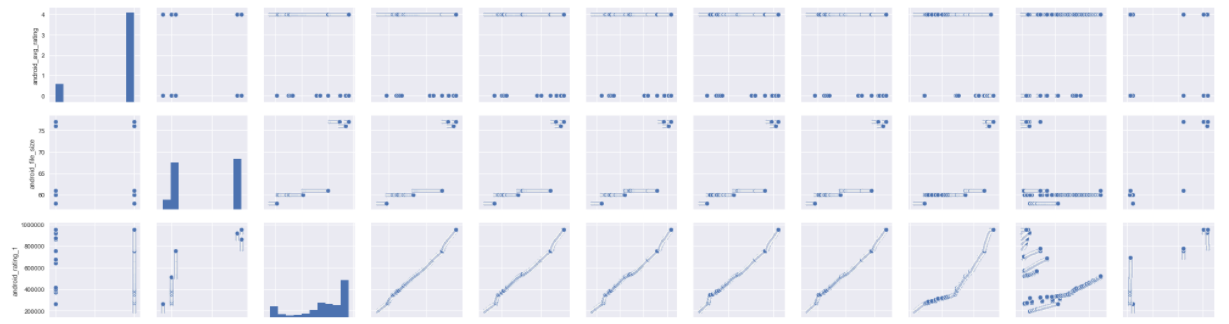
# Data Exploration

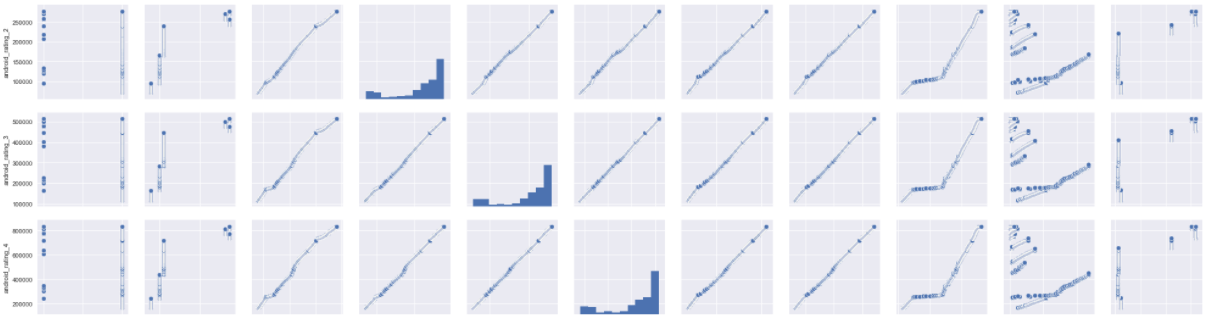
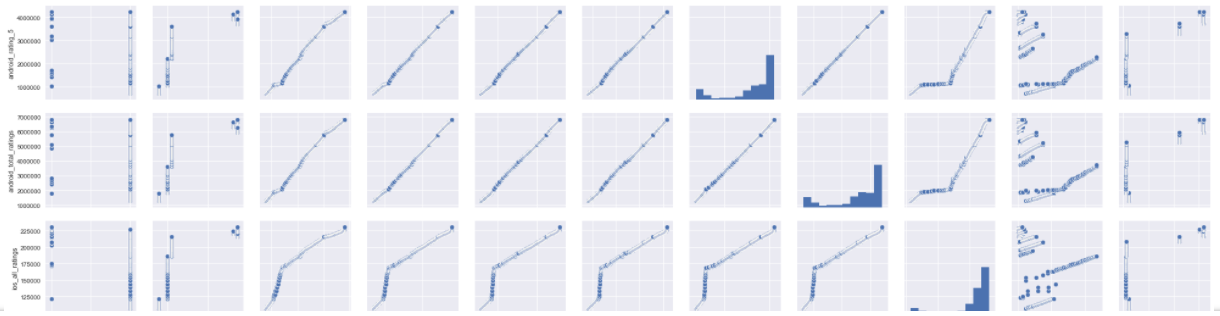
We explored the data using data using various pandas methods.

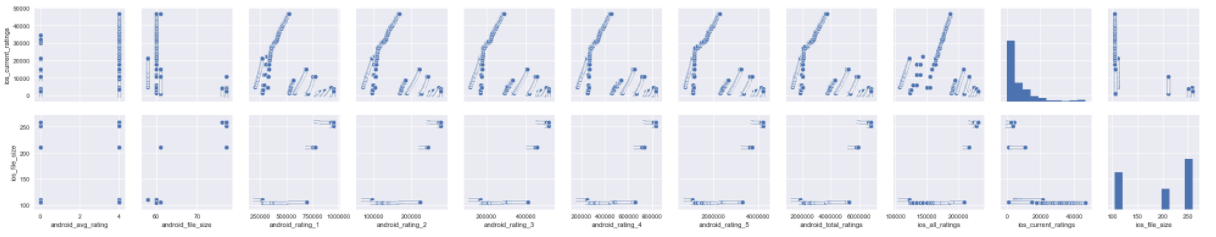
Firstly We used describe() method to compute various statistical coefficients like count, mean, std, min, 25%, 50%, 75%,max values.

|  | **android\_avg\_rating** | **android\_file\_size** | **android\_rating\_1** | **android\_rating\_2** | **android\_rating\_3** | **android\_rating\_4** | **android\_rating\_5** | **android\_total\_ratings** | **ios\_all\_ratings** | **ios\_current\_ratings** | **ios\_file\_size** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 14810.000000 | 14810.000000 | 14810.000000 | 14810.000000 | 14810.000000 | 14810.000000 | 1.481000e+04 | 1.481000e+04 | 14810.000000 | 14810.000000 | 14810.000000 |
| **mean** | 3.324781 | 67.968265 | 719444.100068 | 220767.765834 | 405772.282444 | 649749.918163 | 3.270630e+06 | 5.266365e+06 | 202847.324916 | 7428.748751 | 196.614382 |
| **std** | 1.498368 | 8.191596 | 226316.494405 | 61262.582764 | 119192.281381 | 201520.679210 | 1.080574e+06 | 1.687374e+06 | 33368.236315 | 9113.271931 | 67.086982 |
| **min** | 0.000000 | 58.000000 | 199974.000000 | 71521.000000 | 117754.000000 | 165956.000000 | 7.265970e+05 | 1.281802e+06 | 106508.000000 | 29.000000 | 104.000000 |
| **25%** | 4.000000 | 61.000000 | 627242.000000 | 204299.000000 | 373913.000000 | 596010.000000 | 2.977746e+06 | 4.779210e+06 | 201533.000000 | 1865.000000 | 110.000000 |
| **50%** | 4.000000 | 61.000000 | 752214.000000 | 240268.000000 | 447252.000000 | 715378.000000 | 3.627830e+06 | 5.782942e+06 | 215355.000000 | 3676.000000 | 211.000000 |
| **75%** | 4.000000 | 77.000000 | 909636.000000 | 267621.000000 | 496153.000000 | 804331.000000 | 4.099775e+06 | 6.577516e+06 | 223336.000000 | 9609.000000 | 258.000000 |
| **max** | 4.000000 | 77.000000 | 952604.000000 | 277695.000000 | 514286.000000 | 832416.000000 | 4.239138e+06 | 6.816139e+06 | 230601.000000 | 46692.000000 | 260.000000 |

We plotted a scatter matrix using the seaborn module to analyze the correlation among the variables.





We used the numpy.corrcoef() method to compute the Pearson correlation Coefficient for the most correlated variable pairs.

* The Pearson correlation coefficient between android\_total\_ratings and

ios\_all\_ratings is 0.9633397998128864

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_1 is

0.9947230545173792

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_2 is

0.9996621330458791

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_3 is

0.999555743094971

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_4 is

0.9997143438105867

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_5 is

0.9998462721425666

* The Pearson correlation coefficient between android\_total\_ratings and android\_rating\_1 is

0.9947230545173792

* The Pearson correlation coefficient between android\_rating\_1 and ios\_all\_ratings is

0.9508203679637299

* The Pearson correlation coefficient between android\_rating\_2 and ios\_all\_ratings is

0.9682256523498077

* The Pearson correlation coefficient between android\_rating\_3 and ios\_all\_ratings is

0.9634898886164692

* The Pearson correlation coefficient between android\_rating\_4 and ios\_all\_ratings is

0.9627893079479812

* The Pearson correlation coefficient between android\_rating\_5 and ios\_all\_ratings is

0.964 4412257905655

Finally, Time series for the variables was plotted using Matplotlib.

# Prediction And Modelling

We developed two prediction models for the android and ios parameters. For this we have used Linear Regression, Ridge and Lasso from Scikit Learn to predict the values of android\_total\_ratings and ios\_all\_ratings. We have used cross validation for splitting the data into training and testing sets. The following are the results:

From the above results we predict the values android\_total\_ratings and ios\_all\_ratings of for the date ios\_all\_ratings and android\_total\_ratings for 2016/11/01 11:50 PM.

# Deep Learning

# Insights