

Project 2: Pokémon Go! Analytics

GROUP 17
ABISHEK GANESH
MAHMOOD M NAUMANI
SRI KAVYA RAVELLA

Table of Contents

Problem Statement	2
Data Collection	2
Web Scraping.....	2
Data Organization	2
Data Exploration	2
Data Description	2
Heat Map of Correlation	3
Frequency Histogram.....	3
Scatter Matrix.....	4
Time Series Graph	5
Time Series with Missing Values	5
Time Series with Interpolated Values	6
Prediction Model.....	6
Features	6
Algorithms	6
Predicted Values	7
Deep Learning.....	7
Tensor Flow Android Results	7
Tensor Flow iOS Results	9

Problem Statement

Pokémon Go is a famous augmented reality game which became popular in summer 2016. In This project, we try to understand the reasons behind the success of the game. We aim to achieve this by the following (1) web scraping using beautiful soup (2) constructing pandas data frame (3) exploring numeric data using seaborn (4) using SK-Learn to build machine learning models to predict the review counts for the app (5) analyzing the mobile apps in screen shot images using deep learning with tensor flow.

Data Collection

For our project, we use beautiful soup to scrap data from the downloaded web pages for Pokémon go app from Google Play Store and Apple App Store. We have a total of 144 html pages on a given day for a given platform.

Web Scraping

We have converted the text data from html pages using codecs to utf-8 encoding, we have used beautiful soup to extract from android.html pages. (1) Android average rating(android_avg_rating) and the (2) number of total ratings (android_total_ratings) (3) number of ratings for 1 to 5 stars (android_ratings_1 to android_ratings_5) (4) File size in MB (android_file_size) and similarly from ios pages we have extracted number of customer ratings in current version (ios_current_ratings) (2) number of customer ratings in all versions(ios_all_ratings). (3) File size in MB (ios_file_size).

Data Organization

After collecting data from web scraping we have created a dictionary with date and time as key and the values are extracted from html and iOS pages and then we have created pandas data frame, where index is date time and columns are the 11 values scraped. We have also saved the data frame into 3 formats (json, csv, excel).

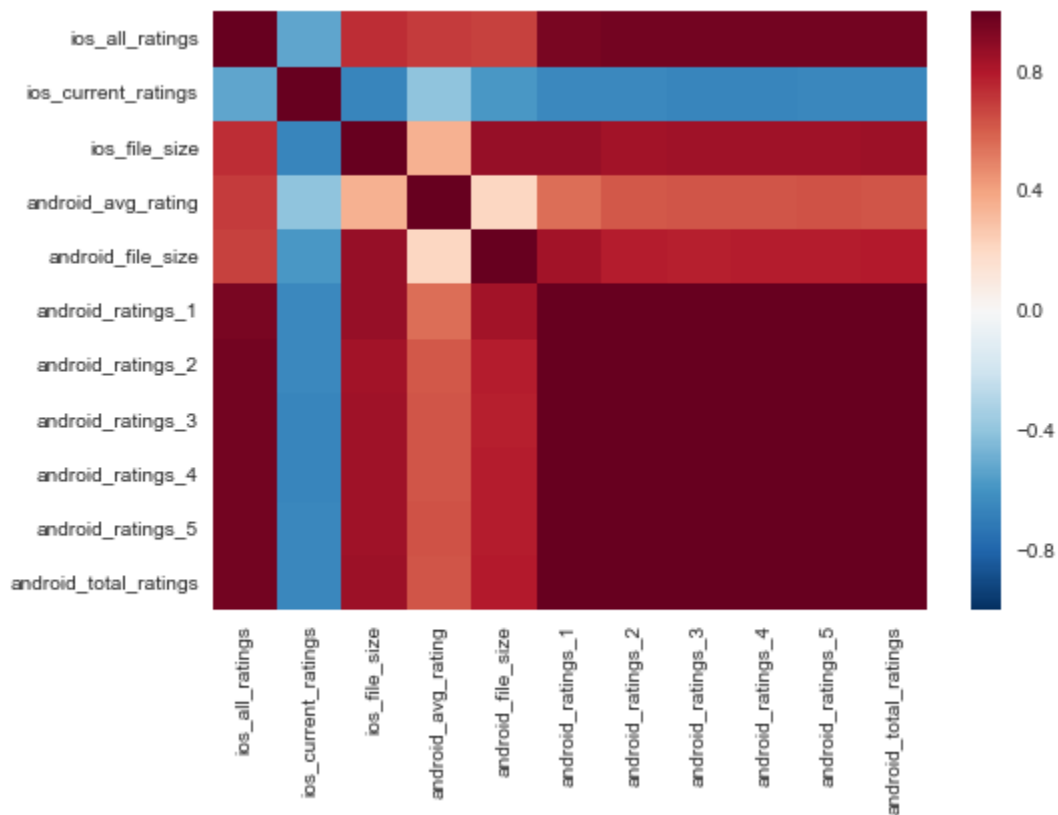
Data Exploration

Using the pandas dataframe we have used pd.describe function to check count, mean stddev, min ,max and then plotted the values to find the missing values in the data set. we used the replace function to replace missing values with np.nan and then filled np.nan values with interpolated values using linear method. We have used heat map to find the correlation between 11 values. We have found that android_ratings_1,2,3....5 are highly positively correlated with android total ratings and ios_all_ratings. Whereas ios_current_rating and android_ratings_1,2,3....5 are highly negatively correlated.

Data Description

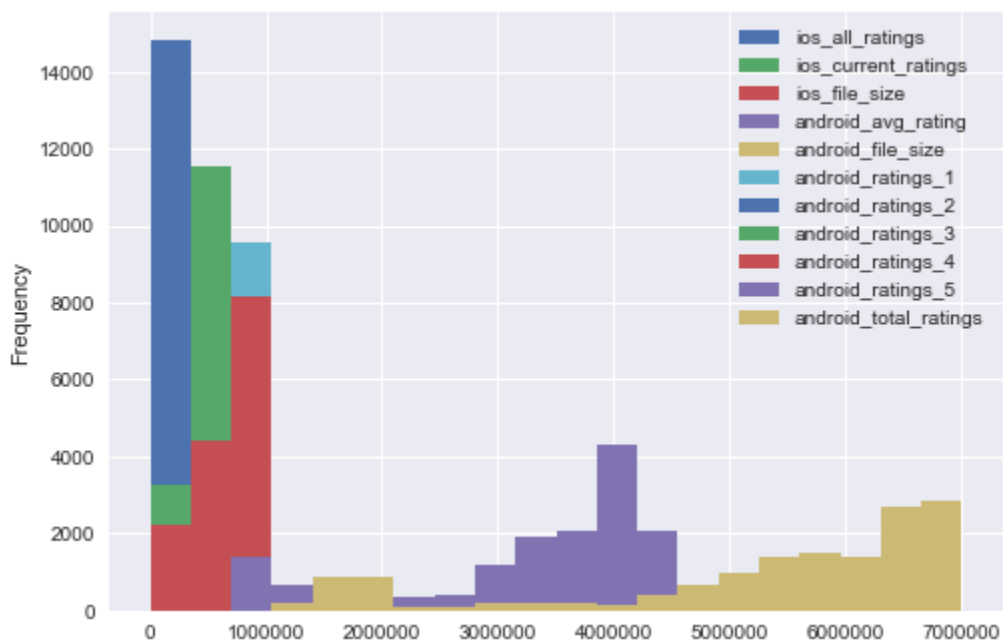
	count	mean	std	min	25%	50%	75%	max
ios_all_ratings	14810	2.03E+05	3.33E+04	106508	201533	2.15E+05	223336	230601
ios_current_ratings	14810	7.25E+03	8.93E+03	29	1815	3.61E+03	9419	46692
ios_file_size	14810	1.97E+02	6.72E+01	104	110	2.11E+02	258	260
android_avg_rating	14810	4.05E+00	7.19E-02	3.9	4	4.10E+00	4.1	4.1
android_file_size	14810	6.82E+01	8.13E+00	58	61	6.14E+01	77	77
android_ratings_1	14810	7.21E+05	2.28E+05	199974	627242	7.53E+05	909636	982631
android_ratings_2	14810	2.21E+05	6.16E+04	71521	204299	2.40E+05	267621	285115
android_ratings_3	14810	4.07E+05	1.20E+05	117754	373913	4.48E+05	496153	528687
android_ratings_4	14810	6.51E+05	2.03E+05	165956	596010	7.16E+05	804331	856213
android_ratings_5	14810	3.28E+06	1.09E+06	726597	2977746	3.63E+06	4099775	4352574
android_total_ratings	14810	5.28E+06	1.70E+06	1281802	4779210	5.79E+06	6577516	7005220

Heat Map of Correlation

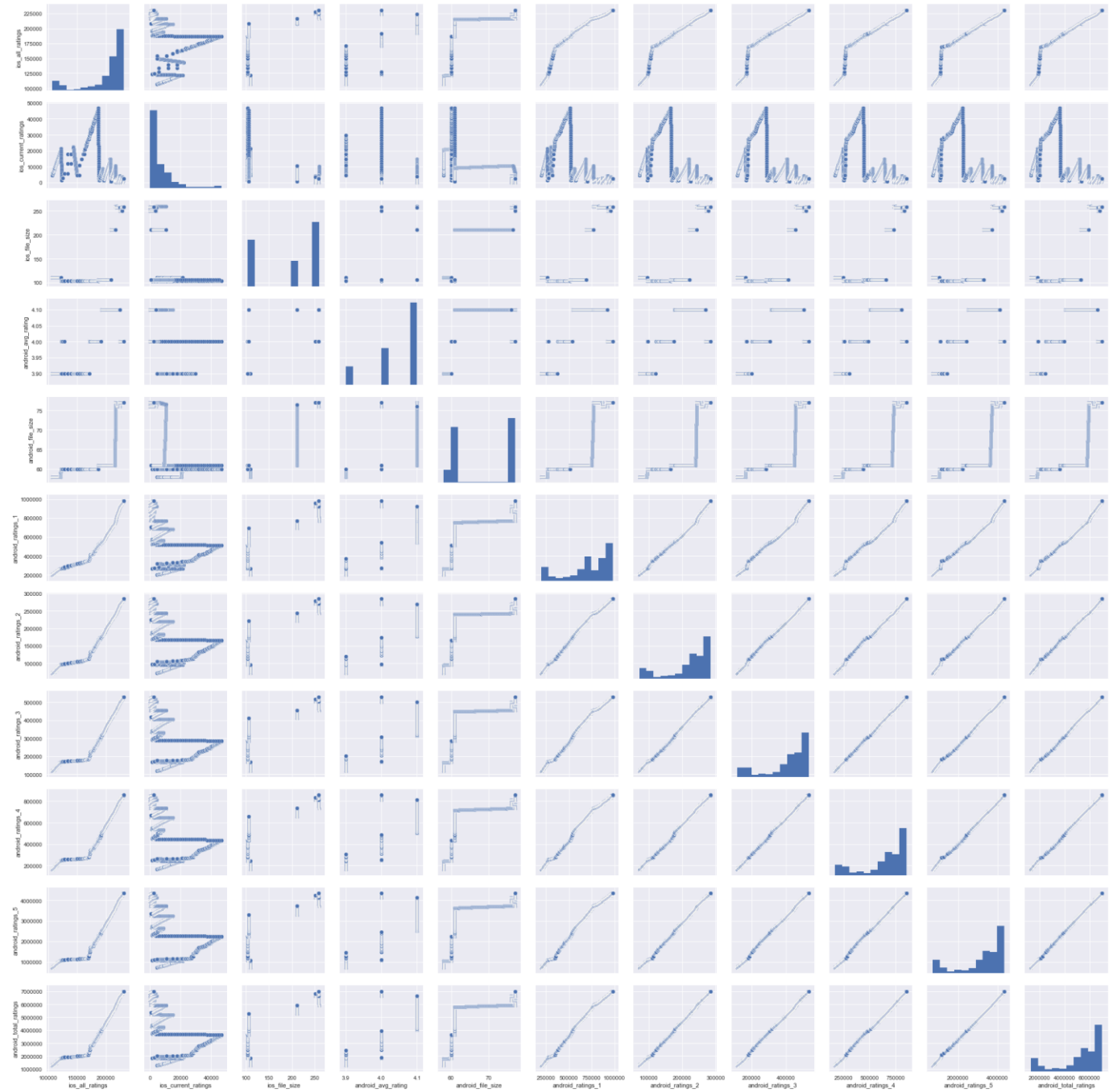


We have plotted a histogram with 11 values to find the frequency of all the values and we can see that the ios_all_ratings have highest frequency we use pair plot to find the correlation between the different combination of 11 values.

Frequency Histogram

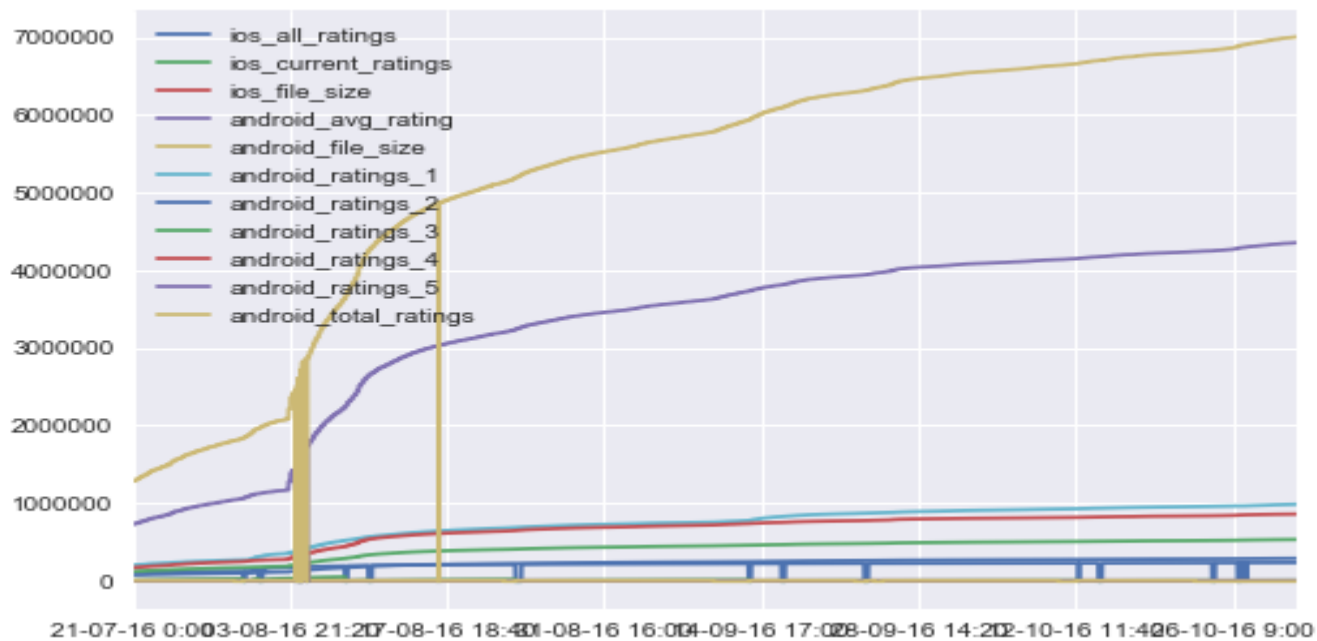


Scatter Matrix



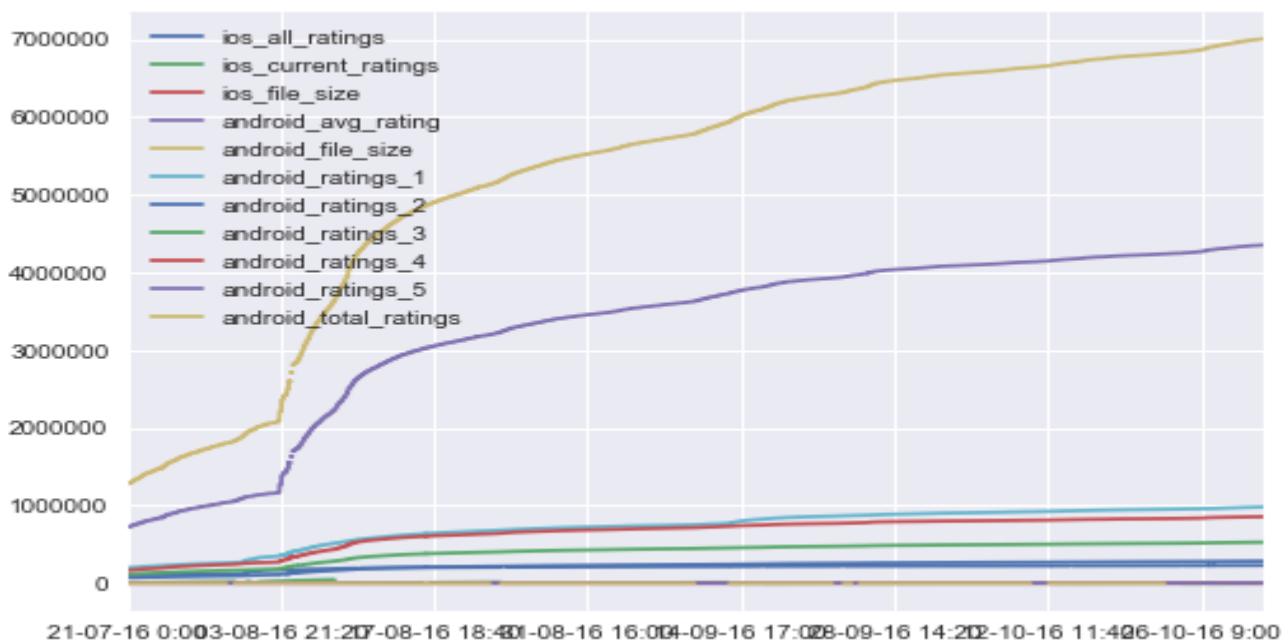
Using matplotlib we have created time series graph for each of 11 variables.

Time Series Graph



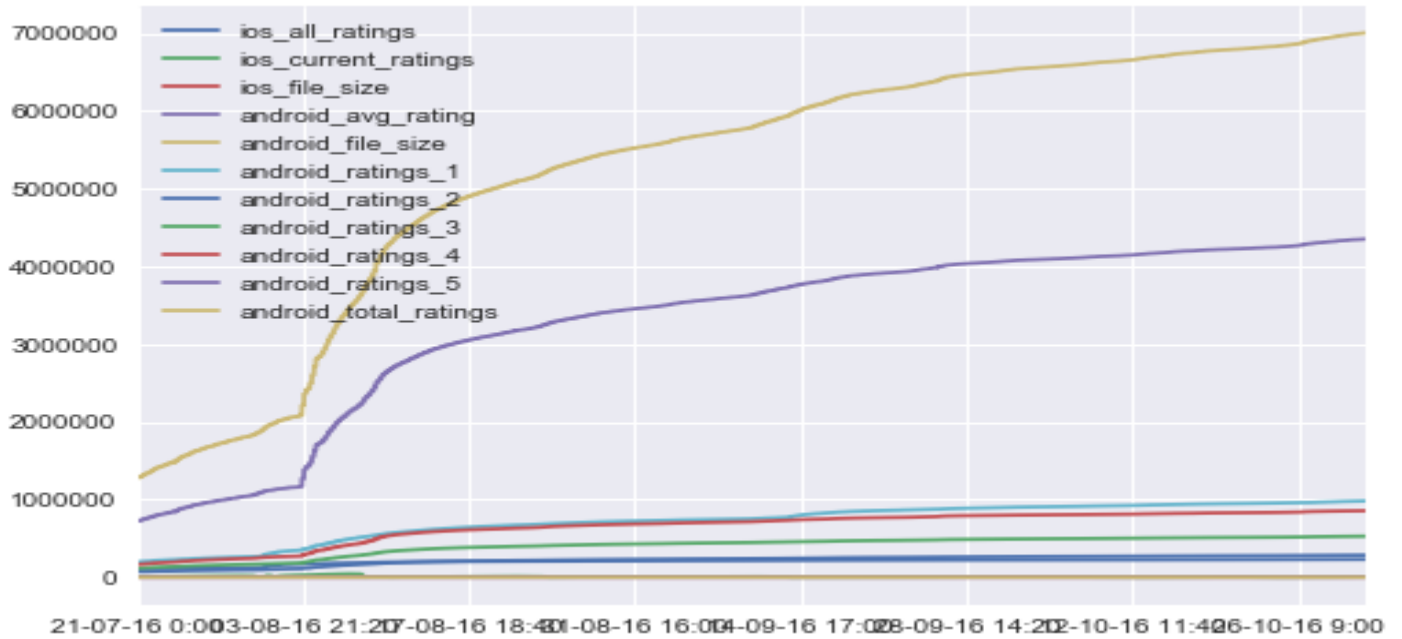
Before Interpolation of Missing Values:

Time Series with Missing Values



After Linear Interpolation of Missing Values:

Time Series with Interpolated Values



Prediction Model

Features

For predicting the target variable, we are using the following attributes as features:

ios_all_ratings, android_file_size, ios_file_size, android_avg_rating, android_total_ratings

Algorithms

We are using the following four algorithms:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Random forest Regression

The Mean absolute error for the runs are as follows:

Android	Mean Absolute Error
Linear	205275.1847
Ridge	356858.6326
Lasso	205274.398
Random Forest	3667.943633

IOS	Mean Absolute Error
Linear	4302.401043
Ridge	9256.803563
Lasso	4302.990047
Random Forest	67.80573149

Based on the above runs for different algorithms, it can be seen that Random Forest Regressor is the generating the best model. Therefore, the predictions for ios_all_ratings and android_total_ratings are as follows

Predicted Values

Android: android_total_ratings: **7002757.5204255246**

Ios: ios_all_ratings: **218071.30351426356**




Deep Learning


Using BeautifulSoup we extracted the links for the screens shots that were given on the web page for both Google Play Store and iOS App Store.

Using tensor flow we have extracted the image tags with the corresponding probabilities for the images, the probabilities gives are the how likely the tags are close to describing the actual object.

Tensor Flow Android Results







No of Unique images : 10






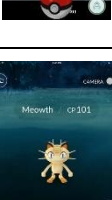


Tensor Flow Analysis of Android Screen Shots		
Srno	Image	Tensor Flow Predications
1		web site, website, internet site, site0.8835709095001221 menu 0.008027342148125172 slot, one-armed bandit 0.0040437160059809685 washer, automatic washer, washing machine 0.0037059905007481575 hand-held computer, hand-held microcomputer 0.0029643995221704245
2		aircraft carrier, carrier, flattop, attack aircraft carrier 0.0996822640299797 pole 0.03657464310526848 wing 0.026553139090538025 lakeside, lakeshore 0.024369344115257263 magnetic compass 0.023960651829838753
3		web site, website, internet site, site0.22753271460533142 envelope 0.0916254073381424 Band Aid 0.03712096065282822 pinwheel 0.029456576332449913 airship, dirigible 0.024857910349965096



4		web site, website, internet site, site0.3677922785282135 envelope 0.16913515329360962 binder, ring-binder 0.05812011659145355 tray 0.017636913806200027 monitor 0.017210323363542557
5		web site, website, internet site, site0.8907706141471863 menu 0.003637630958110094 monitor 0.0018525996711105108 screen, CRT screen 0.0018418238032609224 analog clock 0.001773565192706883
6		Could Not Anlayze due to Error in TensorFlow
7		Could Not Anlayze due to Error in TensorFlow
8		Could Not Anlayze due to Error in TensorFlow
9		Could Not Anlayze due to Error in TensorFlow
10		Could Not Anlayze due to Error in TensorFlow

Tensor Flow iOS Results

Number of Unique Images: 17

Tensor Flow Analysis of iOS Screen Shots		
Sr	Image	Tensor Flow Predictions
1		web site, website, internet site, site0.8835709095001221 menu 0.008027342148125172 slot, one-armed bandit 0.0040437160059809685 washer, automatic washer, washing machine 0.0037059905007481575 hand-held computer, hand-held microcomputer 0.0029643995221704245
2		aircraft carrier, carrier, flattop, attack aircraft carrier 0.0996822640299797 pole 0.03657464310526848 wing 0.026553139090538025 lakeside, lakeshore 0.024369344115257263 magnetic compass 0.023960651829838753
3		web site, website, internet site, site0.22753271460533142 envelope 0.0916254073381424 Band Aid 0.03712096065282822 pinwheel 0.029456576332449913 airship, dirigible 0.024857910349965096
4		web site, website, internet site, site0.3677922785282135 envelope 0.16913515329360962 binder, ring-binder 0.05812011659145355 tray 0.017636913806200027 monitor 0.017210323363542557
5		web site, website, internet site, site0.8907706141471863 menu 0.003637630958110094 monitor 0.0018525996711105108 screen, CRT screen 0.0018418238032609224 analog clock 0.001773565192706883
6		web site, website, internet site, site0.11636681109666824 laptop, laptop computer 0.08079587668180466 notebook, notebook computer 0.05348573625087738 joystick 0.04790787771344185 monitor 0.04169078171253204
7		web site, website, internet site, site0.4224100708961487 comic book 0.032477572560310364 carousel, carrousel, merry-go-round, roundabout, whirligig 0.02088969573378563 fountain 0.017811322584748268 safety pin 0.014400486834347248

8		web site, website, internet site, site0.6088579297065735 television, television system 0.05664973333477974 monitor 0.019958246499300003 notebook, notebook computer0.01607217825949192 iPod 0.011798117309808731
9		web site, website, internet site, site0.12342061847448349 maze, labyrinth 0.07148625701665878 comic book 0.04789267107844353 joystick 0.04420960694551468 television, television system 0.03757670894265175
10		space shuttle 0.23042458295822144 joystick 0.05992179736495018 racer, race car, racing car0.05625780299305916 scoreboard 0.0495721809566021 airliner 0.0457567498087883
11		comic book 0.19361543655395508 maze, labyrinth 0.19329755008220673 web site, website, internet site, site0.05235739052295685 monitor 0.029567204415798187 book jacket, dust cover, dust jacket, dust wrapper 0.027674075216054916
12		ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash barrel, trash bin 0.1549760401248932 joystick 0.06404902786016464 cannon 0.03585105016827583 maraca 0.027268292382359505 pedestal, plinth, footstall0.02715473063290119
13		laptop, laptop computer 0.4985915422439575 web site, website, internet site, site0.10645920038223267 monitor 0.0638403445482254 screen, CRT screen 0.029848407953977585 notebook, notebook computer0.028014272451400757
14		web site, website, internet site, site0.36618542671203613 safety pin 0.020038092508912086 sunglasses, dark glasses, shades 0.01677463762462139 toilet seat 0.015619936399161816 washer, automatic washer, washing machine 0.014380021952092648
15		fountain 0.20302684605121613 carousel, carrousel, merry-go-round, roundabout, whirligig 0.0831361934542656 comic book 0.05170503258705139 toyshop 0.03342542424798012 monitor 0.03227033093571663

16		web site, website, internet site, site0.9409149885177612 analog clock 0.0036712682340294123 envelope 0.0029094445053488016 monitor 0.002251121448352933 screen, CRT screen 0.00216930010356009
17		web site, website, internet site, site0.5862423181533813 monitor 0.07197437435388565 television, television system 0.05955282226204872 comic book 0.04756326600909233 teapot 0.014249777421355247