Data Science Project 2: PokemonGo Analytics

Group 12

A) High Level Code Explanation

1.Load_html_data.py

For web scraping we used "BeautifulSoup" , We stored the scraped data into dictionary and used that dictionary to get the data in CSV and XLSX format with the help of pandas to data frame. Also, we saved the data into json file.

2.Data_Exploration.py

With the help of 'describe' we found the count/mean/std/min/25%/50%75%/max values for each 11 variables. By using Scatter_matrix and Matplotlib we plotted the corresponding graphs. We used the numpy to generate the 'Pearsons Correlation Coefficient'.

3.ios-LinearRegression.py & android-LinearRegression.py

We used pandas,linear model, cross validation from sklearn to form the regression model and to predict the values for target class. The detail process will be explained in part c.

4.**ios-ridge.py & android-ridge.py**

We used pandas, Ridge, cross validation from sklearn.linear_model to form the ridge model and to predict the values for target class.

5.Image_Extraction.py

We used "BeautifulSoup", for scraping images and used 'set' function to store the unique images.

6.image_classify_by_tensorflow.py

We referred the code provided in the ipython notebook to get the prediction of the images.

Data Science Project 2: PokemonGo Analytics
Group 12

B.Data Exploration

1.describe()

	count	mean	std	min	\
Android_Average_rating	14291.0	4.046995e+00	7.186334e-02	3.9	
ios_all_rating	14291.0	2.029651e+05	3.342542e+04	106508.0	
ios_current_rating	14291.0	7.249607e+03	8.917629e+03	29.0	
ios_File_size	14291.0	1.973033e+02	6.675337e+01	104.0	
Android_File_size	14291.0	5.604017e+01	2.602309e+01	0.0	
Android_Rating_4	14291.0	6.523093e+05	2.023678e+05	165956.0	
Android_Rating_5	14291.0	3.283537e+06	1.083986e+06	726597.0	
Android_Rating_2	14291.0	2.214806e+05	6.151529e+04	71521.0	
Android_Rating_3	14291.0	4.072339e+05	1.196588e+05	117754.0	
Android_Total_rating	14291.0	5.286670e+06	1.693392e+06	1281802.0	
Android_Rating_1	14291.0	7.221087e+05	2.273692e+05	199974.0	

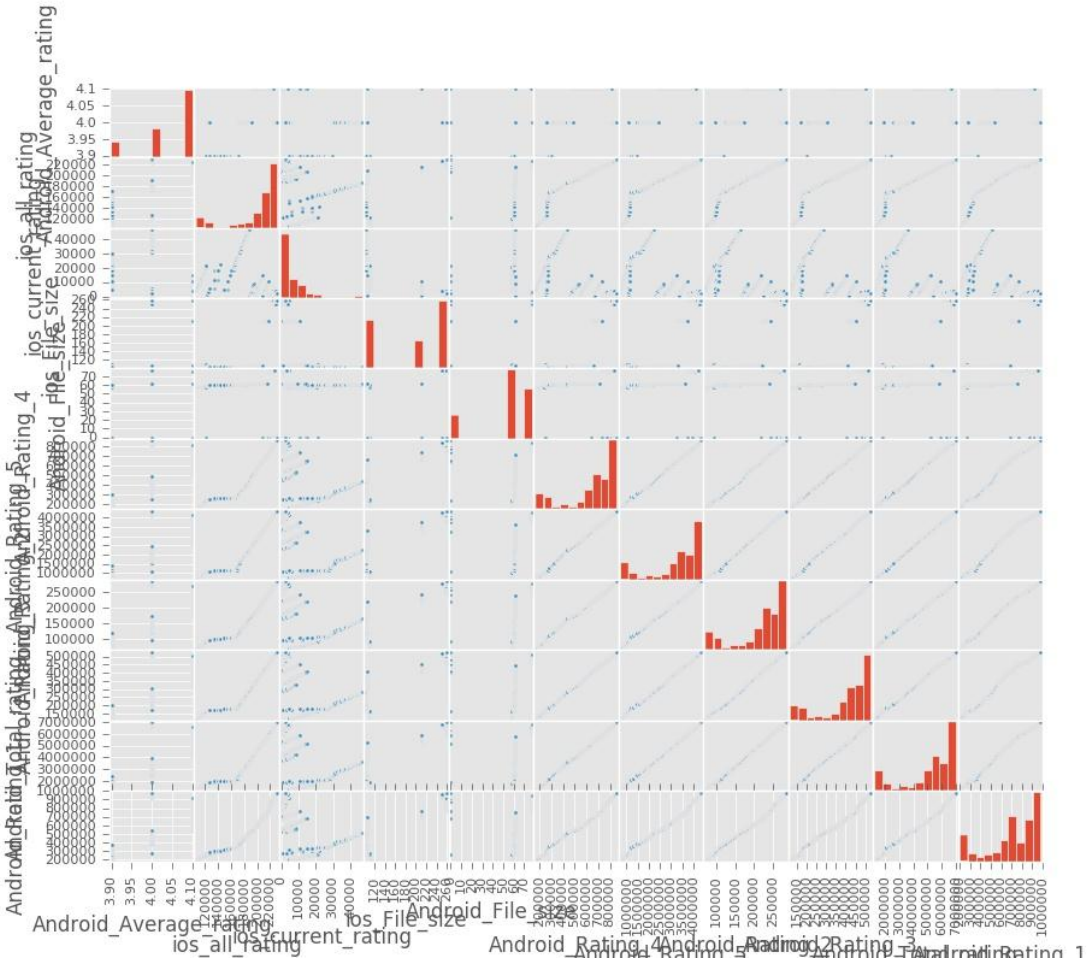
	25%	50%	75%	max
Android_Average_rating	4.0	4.1	4.1	4.1
ios_all_rating	201936.0	215355.0	223336.0	230601.0
ios_current_rating	1853.0	3612.0	9419.0	46692.0
ios_File_size	110.0	211.0	258.0	260.0
Android_File_size	58.0	61.0	76.0	77.0
Android_Rating_4	598618.0	716201.0	804331.0	856213.0
Android_Rating_5	2992284.0	3633064.0	4099775.0	4352574.0
Android_Rating_2	205261.0	240452.0	267621.0	285115.0
Android_Rating_3	375680.0	447650.0	496153.0	528687.0
Android_Total_rating	4802260.0	5790213.0	6577516.0	7005220.0
Android_Rating_1	630417.0	752846.0	909636.0	982631.0

Data Science Project 2: PokemonGo Analytics

Group 12

2. Scatter Matrix

The figure below shows a correlation of all attributes. Except the diagonal, all the other graphs shows the correlation between respective variables from the data



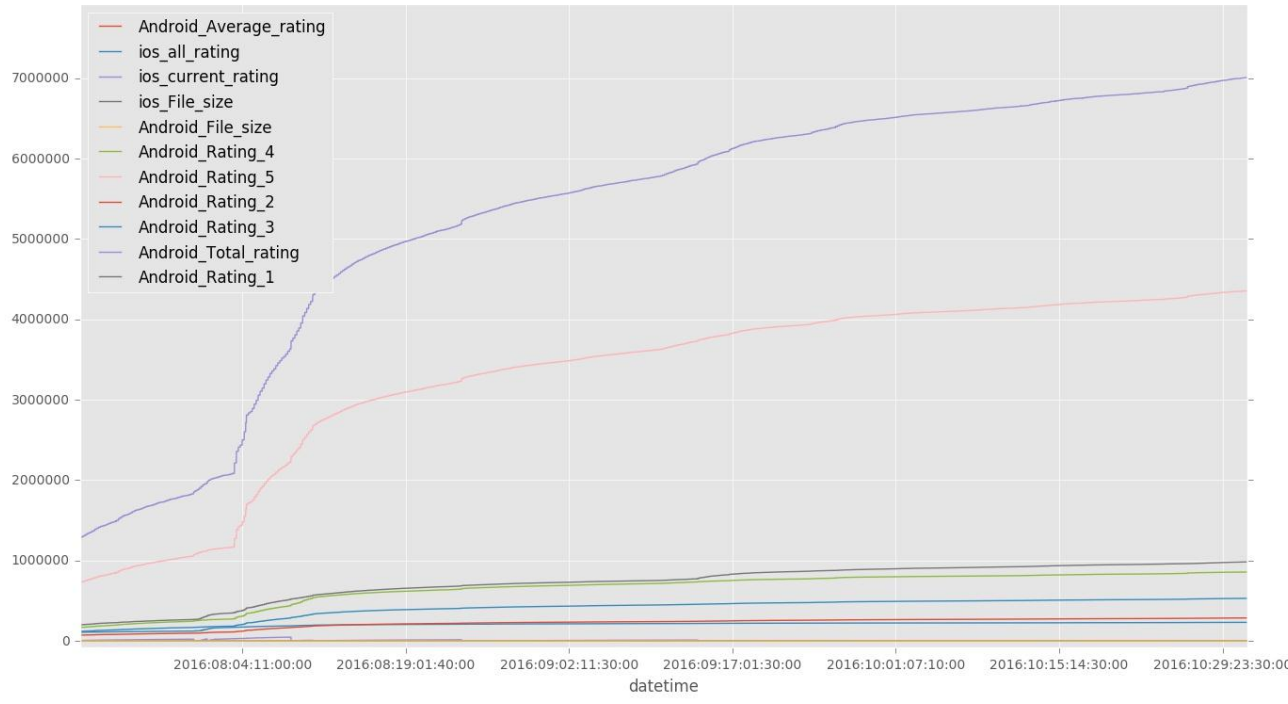
3. Pearson Correlation Coefficient

```
[[ 1.          1.          1.          ...,  0.99759971  0.99759998
   0.99760041]
 [ 1.          1.          1.          ...,  0.99759971  0.99759998
   0.99760041]
 [ 1.          1.          1.          ...,  0.99759971  0.99759998
   0.99760041]
 ...,
 [ 0.99759971  0.99759971  0.99759971 ...,  1.          1.          1.        ]
 [ 0.99759998  0.99759998  0.99759998 ...,  1.          1.          1.        ]
 [ 0.99760041  0.99760041  0.99760041 ...,  1.          1.          1.        ]]
```

Data Science Project 2: PokemonGo Analytics
Group 12

4.Time Series graph

The graph below describes all the eleven attributes with respect to date and time. Where X axis represents datetime.



C)Predictive Modelling

Below are the predictive models and results associated with them. We have stored datetime column in nanoseconds(int) since we have to perform linear regression on data and time which is possible only if it is in numeric format(int/float).

Below are the results for the prediction models for the date 2016/11/01 11:50 pm. Based on our predictive model linear regression performed well in predicting ratings for both ios and android.

1.Android Linear Regression

('Coefficients: \n', array([[3.26530919e-18, -6.05710712e-09, 1.11574338e-11,
1.00000000e+00, 1.00000000e+00, 1.00000000e+00,
1.00000000e+00, 1.00000000e+00]]))

('intercept', array([-4.77023423e-06]))

root Mean squared error: 0.00

Variance score: 1.00

-4.77023422718e-06 3.26530919107e-18 -6.05710712232e-09 1.11574337951e-11 1.0 1.0 1.0 1.0 1.0

('android predicted value for Android_Total_rating :', 5.6037083884271822e-08)

2. ANDROID RIDGE

('Coefficients: \n', array([[7.84944321e-17, 1.54290728e-06, 5.59835830e-08,

1.00000000e+00, 1.00000000e+00, 1.00000000e+00,

1.00000000e+00, 1.00000000e+00]]))

('intercept', array([-0.00012628]))

root Mean squared error: 0.00

Variance score: 1.00

-0.000126278959215 7.84944321416e-17 1.54290727733e-06 5.59835830279e-08 0.999999999925 1.0 1.00000000002 1.00000000011 0.999999999992

('android predicted value for Android_Total_rating :', -1.0260719055451778e-05)

3. IOS RIDGE

('Coefficients: \n', array([[1.23049506e-05, -3.75487069e+01, -1.24535912e-01]]))

('intercept', array([-17918178.13325238]))

root Mean squared error: 20172.31

Variance score: 0.67

-17918178.1333 1.23049505958e-05 -37.5487068836 -0.124535911931

('ios predicted value for ios_all_rating :', 269082.72611317039)

4.IOS Linear Regression

('Coefficients: \n', array([[1.23049507e-05, -3.75487096e+01, -1.24535915e-01]]))

('intercept', array([-17918178.22341248]))

root Mean squared error: 20172.31

Variance score: 0.67

-17918178.2234 1.23049506573e-05 -37.5487095746 -0.124535915085

('ios predicted value for ios_all_rating :', 269082.72695503384)

D) Deep Learning

The number of unique screenshots for iOS and Android are 17 & 5 i.e total 22 unique images.

The probabilities obtained were trained on a very small dataset due to time constraint. If we increase the train data set we can obtain much better results.

Below are the tags/probabilities for each image.

1.

http://a3.mzstatic.com/us/r30/Purple30/v4/4c/8d/0a/4c8d0a13-c73b-2972-b37d-3601de8bdef5/sc1024x768.jpeg
web site, website, internet site, site (score = 0.89077)
menu (score = 0.00364)
monitor (score = 0.00185)
screen, CRT screen (score = 0.00184)
analog clock (score = 0.00177)

2.

http://a2.mzstatic.com/us/r30/Purple18/v4/12/1f/5b/121f5bb3-f00a-13e2-6ce8-ccc9bdb96cc9/screen696x696.jpeg
web site, website, internet site, site (score = 0.42241)
comic book (score = 0.03248)
carousel, carrousel, merry-go-round, roundabout, whirligig (score = 0.02089)
fountain (score = 0.01781)
safety pin (score = 0.01440)

3.

http://a5.mzstatic.com/us/r30/Purple30/v4/ef/1d/b1/ef1db155-870f-cd2d-f93b-2728741a0168/screen696x696.jpeg
web site, website, internet site, site (score = 0.60886)
television, television system (score = 0.05665)
monitor (score = 0.01996)
notebook, notebook computer (score = 0.01607)
iPod (score = 0.01180)

4.

http://a4.mzstatic.com/us/r30/Purple20/v4/b2/aa/a1/b2aaa1a3-e66a-745c-cd57-48b527290c48/screen696x696.jpeg

aircraft carrier, carrier, flattop, attack aircraft carrier (score = 0.09968)pole (score = 0.03657)
wing (score = 0.02655)
lakeside, lakeshore (score = 0.02437)
magnetic compass (score = 0.02396)

5.

http://a4.mzstatic.com/us/r30/Purple60/v4/bc/08/db/bc08dbd0-3ccf-70b8-5f8f-3f76cc4b7b31/screen322x572.jpeg
comic book (score = 0.19361)
maze, labyrinth (score = 0.19330)
web site, website, internet site, site (score = 0.05236)
monitor (score = 0.02957)
book jacket, dust cover, dust jacket, dust wrapper (score = 0.02767)

6.

http://a2.mzstatic.com/us/r30/Purple60/v4/a6/d9/e4/a6d9e449-e9ca-dba0-99c8-09bb5d8fa3c3/screen480x480.jpeg
laptop, laptop computer (score = 0.49859)
web site, website, internet site, site (score = 0.10646)|
monitor (score = 0.06384)
screen, CRT screen (score = 0.02985)
notebook, notebook computer (score = 0.02801)

7.

http://a3.mzstatic.com/us/r30/Purple60/v4/65/4e/b6/654eb6d4-96ba-351d-8be5-a0a267f7e9ed/sc1024x768.jpeg
web site, website, internet site, site (score = 0.11637)
laptop, laptop computer (score = 0.08080)
notebook, notebook computer (score = 0.05349)
joystick (score = 0.04791)
monitor (score = 0.04169)

8.

http://a3.mzstatic.com/us/r30/Purple20/v4/90/87/61/908761f2-9f46-4a88-26e1-e79790eebcfb/screen322x572.jpeg
fountain (score = 0.20303)
carousel, carrousel, merry-go-round, roundabout, whirligig (score = 0.08314)
comic book (score = 0.05171)

toyshop (score = 0.03343)
monitor (score = 0.03227)

9.

http://a2.mzstatic.com/us/r30/Purple60/v4/c7/c6/00/c7c60057-6358-33c1-cb00-707bef81883c/screen696x696.jpeg
space shuttle (score = 0.23042)
joystick (score = 0.05992)
racer, race car, racing car (score = 0.05626)
scoreboard (score = 0.04957)
airliner (score = 0.04576)

10.

http://a2.mzstatic.com/us/r30/Purple20/v4/b0/97/2e/b0972e89-068f-fb49-0f73-25c94f118765/screen322x572.jpeg

ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash barrel, trash bin (score = 0.15498)
joystick (score = 0.06405)
cannon (score = 0.03585)
maraca (score = 0.02727)
pedestal, plinth, footstall (score = 0.02715)

11.

http://a3.mzstatic.com/us/r30/Purple30/v4/40/4b/6a/404b6a60-1563-df95-f90c-3f0c8093e55a/sc1024x768.jpeg

web site, website, internet site, site (score = 0.22753)
envelope (score = 0.09163)
Band Aid (score = 0.03712)
pinwheel (score = 0.02946)
airship, dirigible (score = 0.02486)

12.

http://a1.mzstatic.com/us/r30/Purple18/v4/20/b8/51/20b851a8-9c3a-0659-e20d-804c9150ac96/screen322x572.jpeg

web site, website, internet site, site (score = 0.36619)
safety pin (score = 0.02004)
sunglasses, dark glasses, shades (score = 0.01677)
toilet seat (score = 0.01562)
washer, automatic washer, washing machine (score = 0.01438)

13.

http://a5.mzstatic.com/us/r30/Purple60/v4/24/67/26/24672654-14d4-1dd0-57e6-9804f74a9303/sc1024x768.jpeg

web site, website, internet site, site (score = 0.88357)
menu (score = 0.00803)
slot, one-armed bandit (score = 0.00404)
washer, automatic washer, washing machine (score = 0.00371)
hand-held computer, hand-held microcomputer (score = 0.00296)

14.

http://a2.mzstatic.com/us/r30/Purple18/v4/07/6e/dc/076edcea-e44e-6c6c-f136-7048ce01cf4e/screen696x696.jpeg

web site, website, internet site, site (score = 0.12342)
maze, labyrinth (score = 0.07149)
comic book (score = 0.04789)
joystick (score = 0.04421)
television, television system (score = 0.03758)

15.

http://a3.mzstatic.com/us/r30/Purple18/v4/f1/bd/c9/f1bdc989-acb2-b76a-5922-c5d044f2b59c/screen480x480.jpeg

web site, website, internet site, site (score = 0.94092)
analog clock (score = 0.00367)
envelope (score = 0.00291)
monitor (score = 0.00225)
screen, CRT screen (score = 0.00217)

16.

http://a4.mzstatic.com/us/r30/Purple20/v4/7c/6f/35/7c6f3593-2ef1-838b-4a83-253e1da2e2cb/sc1024x768.jpeg

web site, website, internet site, site (score = 0.36779)
envelope (score = 0.16914)
binder, ring-binder (score = 0.05812)
tray (score = 0.01764)
monitor (score = 0.01721)

17.

http://a1.mzstatic.com/us/r30/Purple20/v4/9e/5f/33/9e5f3303-8649-e870-51a2-e20650865a28/screen322x572.jpeg

web site, website, internet site, site (score = 0.58624)
monitor (score = 0.07197)
television, television system (score = 0.05955)
comic book (score = 0.04756)

18.

//lh3.googleusercontent.com/22ySaopy8gQQelxKpUMUP56i9kAhnoONR4RmjEZ1AyvWqbO-ae_kO8Hi1zIqBfqNjFk=h310

web site, website, internet site, site (score = 0.62170)
television, television system (score = 0.08697)
monitor (score = 0.04946)
screen, CRT screen (score = 0.03223)
hand-held computer, hand-held microcomputer (score = 0.02868)
teapot (score = 0.01425)

19.

//lh3.googleusercontent.com/UJsqbNSI3dFLNeVw0qGYdDNz3uvzrKOw9r0DHQ0KZigwrKfFyiLSFjSTkI_DBdYz2yt-=h310

web site, website, internet site, site (score = 0.49719)
monitor (score = 0.07830)
notebook, notebook computer (score = 0.05803)
iPod (score = 0.03292)
desktop computer (score = 0.02498)

20.

//lh3.googleusercontent.com/dq_t7Is81-gkHYxKfAQ7PuLQBR-Qrte-7S1DsKFZnhaZATpibMSiw3aCrJzYik1x3IV5=h310

screen, CRT screen (score = 0.12488)
desktop computer (score = 0.07624)
web site, website, internet site, site (score = 0.05729)
television, television system (score = 0.02949)

21.

//lh3.googleusercontent.com/S97uuMKpsZXzQ70_lf-535aUHwN3pw98veYbHR0CduoNCD7nt9QuBqTPXrk916mNwnJ1=h310

lawn mower, mower (score = 0.17193)
golf ball (score = 0.11031)
croquet ball (score = 0.08029)
mountain tent (score = 0.03014)
bow (score = 0.02765)

22.

//lh3.googleusercontent.com/J8kUfrUigeTuBZYHVp5XRKlxmOaOl5g1oXT6EFdVon8xYPoUvkW1N_e05O7-hnqk7UQ=h310

web site, website, internet site, site (score = 0.59586)
comic book (score = 0.03351)
iPod (score = 0.02989)
screen, CRT screen (score = 0.02483)
television, television system (score = 0.02017)

References

1.Codes from blackboard

https://www.dropbox.com/sh/cpkokzf1ko0p53t/AAB5qL9AX8qgLl4HjmDgA2zia?dl=0

2. http://scikit-learn.org/stable/

3. https://www.tensorflow.org/