RegressionSpark

December 10, 2020

[7]:
```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://www-us.apache.org/dist/spark/spark-2.4.7/spark-2.4.
 ↪7-bin-hadoop2.7.tgz
!tar xf spark-2.4.7-bin-hadoop2.7.tgz
!pip install -q findspark
```

[8]:
```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.7-bin-hadoop2.7"
```

[9]:
```
# installing package to plot histogram
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
```

[10]:
```
!pip install pyspark_dist_explore
```

```
Requirement already satisfied: pyspark_dist_explore in /usr/local/lib/python3.6
/dist-packages (0.1.8)
Requirement already satisfied: scipy in /usr/local/lib/python3.6/dist-packages
(from pyspark_dist_explore) (1.4.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages
(from pyspark_dist_explore) (1.18.5)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.6/dist-
packages (from pyspark_dist_explore) (3.2.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.6/dist-packages
(from pyspark_dist_explore) (1.1.4)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.6
/dist-packages (from matplotlib->pyspark_dist_explore) (2.8.1)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.6
/dist-packages (from matplotlib->pyspark_dist_explore) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.6/dist-
packages (from matplotlib->pyspark_dist_explore) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/usr/local/lib/python3.6/dist-packages (from matplotlib->pyspark_dist_explore)
(2.4.7)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.6/dist-
```

```
packages (from pandas->pyspark_dist_explore) (2018.9)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.6/dist-
packages (from python-dateutil>=2.1->matplotlib->pyspark_dist_explore) (1.15.0)
```

[11]:

```
# Importing the required packages
import pyspark.sql.functions as x
from pyspark.sql.functions import col, count, isnan, when
from pyspark.sql.types import IntegerType
#below code to install files for corelation Matrix
import matplotlib.pyplot as plt
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.stat import Correlation
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import StandardScaler
from pyspark_dist_explore import hist
```

[12]:

```
#importing the Data File into Google Colab
from google.colab import files
files.upload()
```

```
Output hidden; open in https://colab.research.google.com to view.
```

[13]:

```
#Importing the SparkConf, SparkContext
from pyspark import SparkConf, SparkContext
from pyspark.sql import SQLContext
sc = SparkContext.getOrCreate();
sqlContext = SQLContext(sc)
```

[14]:

```
#creating a pyspark.sql.dataframe.DataFrame called  df by reading the data␣
 ↪previously imported
df = spark.read.csv('games.csv',inferSchema=True, header =True)
```

[15]:

```
# printing the schema of the dataframe
df.printSchema()
```

```
root
 |-- id: integer (nullable = true)
 |-- type: string (nullable = true)
 |-- name: string (nullable = true)
 |-- yearpublished: string (nullable = true)
 |-- minplayers: string (nullable = true)
 |-- maxplayers: string (nullable = true)
 |-- playingtime: integer (nullable = true)
 |-- minplaytime: integer (nullable = true)
 |-- maxplaytime: integer (nullable = true)
 |-- minage: integer (nullable = true)
 |-- users_rated: integer (nullable = true)
 |-- average_rating: double (nullable = true)
 |-- bayes_average_rating: double (nullable = true)
 |-- total_owners: integer (nullable = true)
 |-- total_traders: integer (nullable = true)
 |-- total_wanters: integer (nullable = true)
 |-- total_wishers: integer (nullable = true)
 |-- total_comments: integer (nullable = true)
 |-- total_weights: integer (nullable = true)
 |-- average_weight: double (nullable = true)
```

[16]:

```
# prionting the top 10 rows of the dataFrame
df.show(10)
```

```
+------+---------+--------------------+-------------+----------+----------+-----
------+-----------+-----------+------+-----------+--------------+---------------
-----+------------+-------------+-------------+-------------+--------------+----
---------+--------------+
|    id|     type|                name|yearpublished|minplayers|maxplayers|playi
ngtime|minplaytime|maxplaytime|minage|users_rated|average_rating|bayes_average_r
ating|total_owners|total_traders|total_wanters|total_wishers|total_comments|tota
l_weights|average_weight|
+------+---------+--------------------+-------------+----------+----------+-----
------+-----------+-----------+------+-----------+--------------+---------------
-----+------------+-------------+-------------+-------------+--------------+----
---------+--------------+
| 12333|boardgame|   Twilight Struggle|         2005|         2|         2|
180|        180|        180|    13|      20113|       8.33774|
8.22186|       26647|          372|         1219|         5865|          5347|
2562|        3.4785|
|120677|boardgame|       Terra Mystica|         2012|         2|         5|
150|         60|        150|    12|      14383|       8.28798|
8.14232|       16519|          132|         1586|         6277|          2526|
1423|        3.8939|
|102794|boardgame|Caverna: The Cave...|         2013|         1|         7|
210|         30|        210|    12|       9262|       8.28994|
8.06886|       12230|           99|         1476|         5600|          1700|
777|        3.7761|
| 25613|boardgame|Through the Ages:...|         2006|         2|         4|
240|        240|        240|    12|      13294|       8.20407|
8.05804|       14343|          362|         1084|         5075|          3378|
1642|         4.159|
|  3076|boardgame|         Puerto Rico|         2002|         2|         5|
150|         90|        150|    12|      39883|       8.14261|
8.04524|       44362|          795|          861|         5414|          9173|
5213|        3.2943|
| 31260|boardgame|            Agricola|         2007|         1|         5|
150|         30|        150|    12|      39714|       8.11957|
8.03847|       47522|         837|          958|         6402|          9310|
5065|         3.616|
|124742|boardgame|  Android: Netrunner|         2012|         2|         2|
45|         45|         45|     14|      15281|         8.1676|
7.97822|       24381|         680|          627|         3244|          3202|
1260|        3.3103|
| 96848|boardgame|Mage Knight Board...|         2011|         1|         4|
150|        150|        150|    14|      12697|       8.15901|
7.96929|       18769|         367|         1116|         5427|          2861|
1409|        4.1292|
| 84876|boardgame|The Castles of Bu...|         2011|         2|         4|
90|         30|         90|     12|      15461|      8.07879|
7.95011|       20558|         215|          929|         3681|          3244|
1176|        3.0442|
| 72125|boardgame|             Eclipse|         2011|         2|         6|
200|         60|        200|    14|      15709|       8.07933|
7.93244|       17611|         273|         1108|         5581|          3188|
1486|        3.6359|
+------+---------+--------------------+-------------+----------+----------+-----
------+-----------+-----------+------+-----------+--------------+---------------
-----+------------+-------------+-------------+-------------+--------------+----
---------+--------------+
only showing top 10 rows
```

[17]:
```
# Changing the DataType of columns minplayers,maxplayers,yearpublished from␣
↪string to Integer Type
df = df.withColumn("minplayers", df["minplayers"].cast(IntegerType()))
df = df.withColumn("maxplayers", df["maxplayers"].cast(IntegerType()))
df = df.withColumn("yearpublished", df["yearpublished"].cast(IntegerType()))
```

[18]:
```
# printing the schema of the dataframe
df.printSchema()
```

```
root
 |-- id: integer (nullable = true)
 |-- type: string (nullable = true)
 |-- name: string (nullable = true)
 |-- yearpublished: integer (nullable = true)
 |-- minplayers: integer (nullable = true)
 |-- maxplayers: integer (nullable = true)
 |-- playingtime: integer (nullable = true)
 |-- minplaytime: integer (nullable = true)
 |-- maxplaytime: integer (nullable = true)
 |-- minage: integer (nullable = true)
 |-- users_rated: integer (nullable = true)
 |-- average_rating: double (nullable = true)
 |-- bayes_average_rating: double (nullable = true)
 |-- total_owners: integer (nullable = true)
 |-- total_traders: integer (nullable = true)
 |-- total_wanters: integer (nullable = true)
 |-- total_wishers: integer (nullable = true)
 |-- total_comments: integer (nullable = true)
 |-- total_weights: integer (nullable = true)
 |-- average_weight: double (nullable = true)
```

[19]:

```
#printing the dimensions of the dataframe
print((df.count(), len(df.columns)))
```

```
(81312, 20)
```

[20]:

```
#Removing the Rows containing the users_rated below 0
df=df.filter((x.col('users_rated') > 0))
```

[21]:

```
#printing the dimensions of the dataframe after the above filtering
print((df.count(), len(df.columns)))
```

```
(56932, 20)
```

[22]:

```
#removing the following columns "bayes_average_rating", "type","name", "id"␣
 ↪from the dataframe
df=df.select([c for c in df.columns if c not in ["bayes_average_rating",␣
 ↪"type","name", "id"]])
```

[23]:

```
#printing the dimensions of the dataframe after the removal of the columns
print((df.count(), len(df.columns)))
```

```
(56932, 16)
```

[24]:

```
# prionting the top 10 rows of the dataFrame
df.show(10)
```

```
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|yearpublished|minplayers|maxplayers|playingtime|minplaytime|maxplaytime|minage|
users_rated|average_rating|total_owners|total_traders|total_wanters|total_wisher
s|total_comments|total_weights|average_weight|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|         2005|         2|         2|        180|        180|        180|    13|
20113|       8.33774|       26647|          372|         1219|         5865|
5347|         2562|        3.4785|
|         2012|         2|         5|        150|         60|        150|    12|
14383|       8.28798|       16519|          132|         1586|         6277|
2526|         1423|        3.8939|
|         2013|         1|         7|        210|         30|        210|    12|
9262|       8.28994|       12230|           99|         1476|         5600|
1700|          777|        3.7761|
|         2006|         2|         4|        240|        240|        240|    12|
13294|       8.20407|       14343|          362|         1084|         5075|
3378|         1642|         4.159|
|         2002|         2|         5|        150|         90|        150|    12|
39883|       8.14261|       44362|          795|          861|         5414|
9173|         5213|        3.2943|
|         2007|         1|         5|        150|         30|        150|    12|
39714|       8.11957|       47522|          837|          958|         6402|
9310|         5065|         3.616|
|         2012|         2|         2|         45|         45|         45|    14|
15281|        8.1676|       24381|          680|          627|         3244|
3202|         1260|        3.3103|
|         2011|         1|         4|        150|        150|        150|    14|
12697|       8.15901|       18769|          367|         1116|         5427|
2861|         1409|        4.1292|
|         2011|         2|         4|         90|         30|         90|    12|
15461|       8.07879|       20558|          215|          929|         3681|
3244|         1176|        3.0442|
|         2011|         2|         6|        200|         60|        200|    14|
15709|       8.07933|       17611|          273|         1108|         5581|
3188|         1486|        3.6359|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
only showing top 10 rows
```

[25]:

```
# printing the schema of the dataframe
df.printSchema()
```

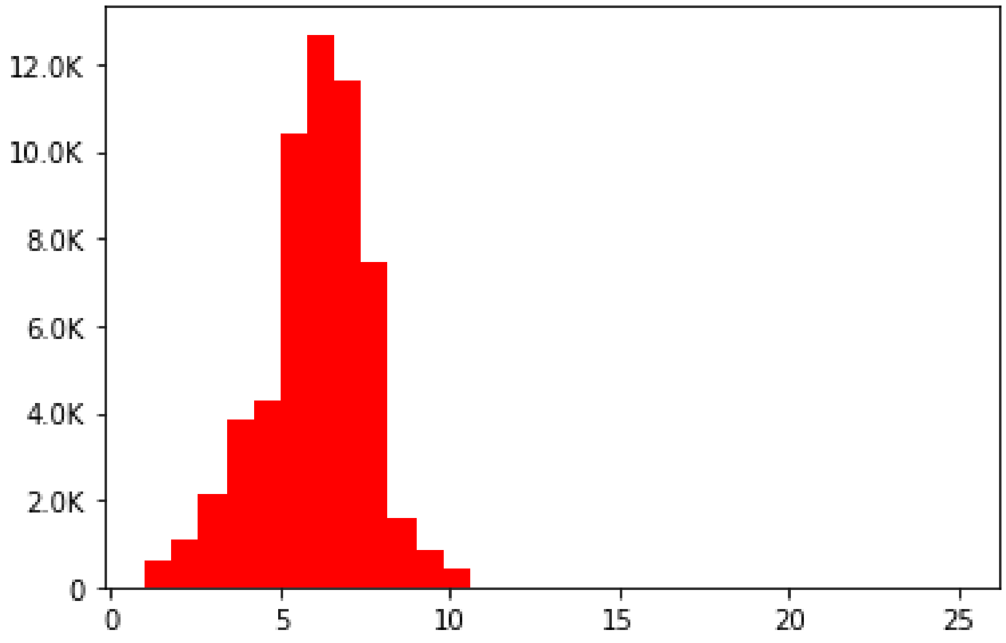
```
root
 |-- yearpublished: integer (nullable = true)
 |-- minplayers: integer (nullable = true)
 |-- maxplayers: integer (nullable = true)
 |-- playingtime: integer (nullable = true)
 |-- minplaytime: integer (nullable = true)
 |-- maxplaytime: integer (nullable = true)
 |-- minage: integer (nullable = true)
 |-- users_rated: integer (nullable = true)
 |-- average_rating: double (nullable = true)
 |-- total_owners: integer (nullable = true)
 |-- total_traders: integer (nullable = true)
 |-- total_wanters: integer (nullable = true)
 |-- total_wishers: integer (nullable = true)
 |-- total_comments: integer (nullable = true)
 |-- total_weights: integer (nullable = true)
 |-- average_weight: double (nullable = true)
```

[26]:
```
# printing the Count mean stdDev Min Max
df.describe().show()
```

```
+-------+------------------+-----------------+------------------+---------------
---+-----------------+-----------------+------------------+------------------+--
---------------+------------------+------------------+-----------------+--------
----------+-----------------+------------------+------------------+
|summary|     yearpublished|       minplayers|        maxplayers|
playingtime|      minplaytime|      maxplaytime|            minage|
users_rated|   average_rating|      total_owners|     total_traders|
total_wanters|     total_wishers|   total_comments|     total_weights|
average_weight|
+-------+------------------+-----------------+------------------+---------------
---+-----------------+-----------------+------------------+------------------+--
---------------+------------------+------------------+-----------------+--------
----------+-----------------+------------------+------------------+
|  count|             56925|            56929|             56929|
56930|            56930|            56930|             56930|             56932|
56932|             56932|             56932|            56932|
56932|            56932|             56932|             56932|
|   mean|1874.7589459815547|2.145619982785575|
5.573644364032392|59.935815914280695|57.44503776567715|59.90786931319164|
7.599279817319515|231.21197217733436|6.016053039942434|
374.2626291013841|13.158944003372444|17.99062038923628|
60.53931005409963|70.29883018337667|23.533689313567063|1.2692436924752304|
| stddev| 486.0990365309268|
16.7912994635512|50.757954719008964|406.40176883743374|393.0744116092278| 406.32
4713326961|4.9501535611699605|1363.6812580935118|1.580773749169548|1786.43241221
17926|  46.9707538877096|71.96175435037111|284.07069468630885|338.2672178431332|
138.0081582736383|1.2138650776308242|
|    min|             -3500|                0|                 0|
0|                0|                0|                 0|                 1|
1.0|                 0|                 0|                0|                 0|
0|                 0|               0.0|
|    max|              2017|             2011|             11299|
60120|            60120|            60120|               180|             53680|
25.0|             73188|              1395|             1586|              6402|
11798|              5996|               5.0|
+--------+------------------+-----------------+------------------+---------------
---+-----------------+-----------------+------------------+-----------------+--
----------------+------------------+------------------+-----------------+--------
----------+-----------------+------------------+------------------+
```

[29]:
```
#plotting the Histogram
#!pip install pyspark_dist_explore
#import pyspark_dist_explore
fig, ax = plt.subplots()
#hist(ax,dataset.select("average_rating"), bins = 30, color=['red'])
hist(ax,df.select("average_rating"), bins = 30, color=['red'])
#hist(ax, dataset.col withColumn('average_rating')
```

[29]:
```
(array([5.9600e+02, 1.0850e+03, 2.1290e+03, 3.8230e+03, 4.2890e+03,
        1.0392e+04, 1.2689e+04, 1.1625e+04, 7.4590e+03, 1.5570e+03,
        8.5600e+02, 4.3100e+02, 0.0000e+00, 0.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 1.0000e+00]),
 array([ 1. ,  1.8,  2.6,  3.4,  4.2,  5. ,  5.8,  6.6,  7.4,  8.2,  9. ,
         9.8, 10.6, 11.4, 12.2, 13. , 13.8, 14.6, 15.4, 16.2, 17. , 17.8,
        18.6, 19.4, 20.2, 21. , 21.8, 22.6, 23.4, 24.2, 25. ]),
 <a list of 30 Patch objects>)
```



[30]:
```
#Printing the count of NA values present in the DataFrame
df.select([count(when(col(c).isNull(), c)).alias(c) for c in df.columns]).show()
```

```
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|yearpublished|minplayers|maxplayers|playingtime|minplaytime|maxplaytime|minage|
users_rated|average_rating|total_owners|total_traders|total_wanters|total_wisher
s|total_comments|total_weights|average_weight|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|            7|         3|         3|          2|          2|          2|     2|
0|             0|           0|            0|            0|            0|
0|            0|             0|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
```

[31]:
```
#Dropping the NA values present in the Dataframe
df=df.dropna(how='any')
```

[32]:
```
#Printing the count of NA values present in the DataFrame after the removal of␣
 ↪NA
df.select([count(when(col(c).isNull(), c)).alias(c) for c in df.columns]).show()
```

```
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|yearpublished|minplayers|maxplayers|playingtime|minplaytime|maxplaytime|minage|
users_rated|average_rating|total_owners|total_traders|total_wanters|total_wisher
s|total_comments|total_weights|average_weight|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
|            0|         0|         0|          0|          0|          0|     0|
0|             0|           0|            0|            0|            0|
0|            0|             0|
+-------------+----------+----------+-----------+-----------+-----------+------+
-----------+--------------+------------+-------------+-------------+------------
-+--------------+-------------+--------------+
```

[33]:
```
#printing the dimensions of the dataframe after the removal of NA
print((df.count(), len(df.columns)))
```

```
(56925, 16)
```

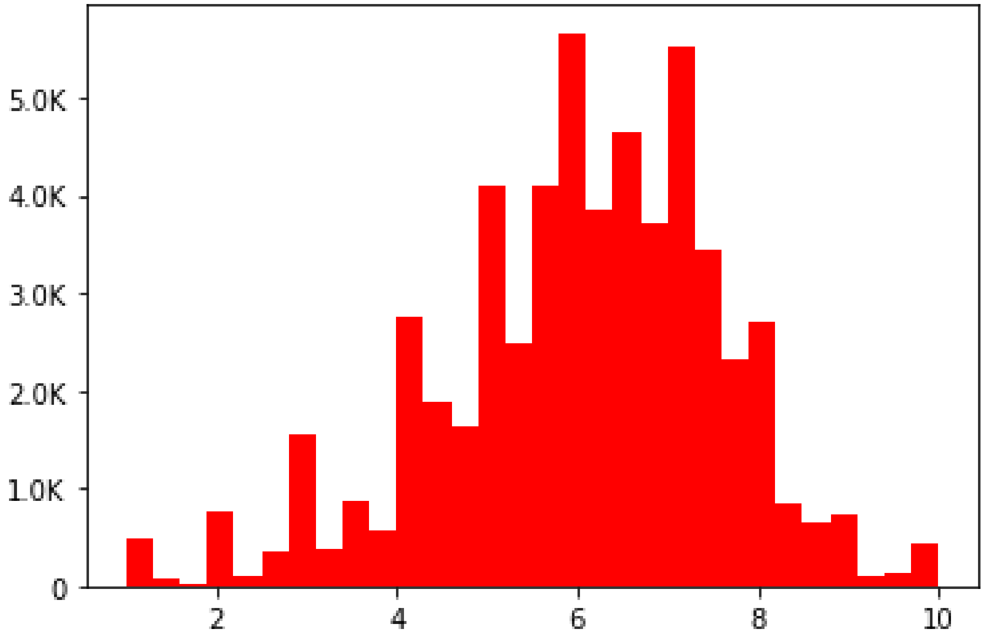
[34]:

```
#printing the Histogram after NA removal

from pyspark_dist_explore import hist
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
hist(ax,df.select("average_rating"), bins = 30, color=['red'],)
#hist(ax, dataset.col withColumn('average_rating')
```

[34]:

```
(array([ 482.,   86.,   34.,  770.,  112.,  350., 1565.,  379.,  880.,
         576., 2760., 1874., 1629., 4107., 2493., 4100., 5666., 3840.,
        4635., 3704., 5517., 3448., 2330., 2693.,  845.,  652.,  726.,
         107.,  120.,  445.]),
 array([ 1. ,  1.3,  1.6,  1.9,  2.2,  2.5,  2.8,  3.1,  3.4,  3.7,  4. ,
         4.3,  4.6,  4.9,  5.2,  5.5,  5.8,  6.1,  6.4,  6.7,  7. ,  7.3,
         7.6,  7.9,  8.2,  8.5,  8.8,  9.1,  9.4,  9.7, 10. ]),
 <a list of 30 Patch objects>)
```



[34]:

```
```

[35]:

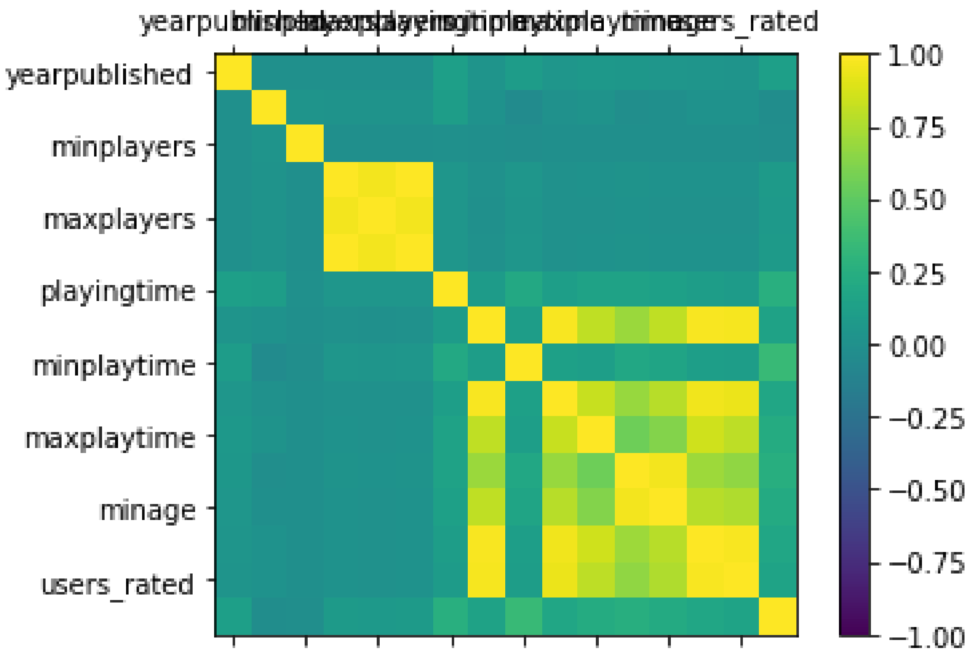
```
#Generating the Corelation Matrix
columnsData=df.columns#[c for c in dataset2.columns if c not in␣
 ↪['average_rating'] ]
vector_col = "corr_features"
assembler = VectorAssembler(inputCols=columnsData, outputCol=vector_col)
myGraph_vector = assembler.transform(df)#.select(vector_col)
matrix = Correlation.corr(myGraph_vector, vector_col)
matrix1 = Correlation.corr(myGraph_vector, vector_col).collect()[0][0]
```

```
corrmatrix = matrix1.toArray().tolist()
print("The generated Correlation Matrix is:: \n {0}".format(corrmatrix))
```

```
The generated Correlation Matrix is::
 [[1.0, 0.004372489082671985, 0.0049703856990403325, 0.006232916329865973,
0.005242466181763035, 0.006232916329865973, 0.12522219543828494,
0.037449604506840636, 0.10840383773603023, 0.04773344107604199,
0.06439133752559806, 0.062276921119165635, 0.052525157614782124,
0.043813716831314624, 0.03622198669878126, 0.12530663336643624],
[0.004372489082671985, 1.0, 0.03718879898920534, 0.02428356613412483,
0.0265323496473018, 0.02428356613412483, 0.11117982079071105,
0.020313525169298213, -0.032858165438786856, 0.014536490514991217,
0.024624729226298204, -0.008714669167627476, -0.0038365658456161987,
0.02297155625899625, 0.019011603278792415, -0.022024292273003405],
[0.0049703856990403325, 0.03718879898920534, 1.0, -0.0010206921408426179,
-0.0009504277066825062, -0.0010206921408426179, 0.004521518981639495,
-0.0008609769588981912, -0.008339907100546615, -0.0009813364242238398,
-0.002386341572889714, -0.0035900367546440105, -0.00250317090907312,
-0.0015057172595959244, -0.0014607609885830214, -0.013453775274283387],
[0.006232916329865973, 0.02428356613412483, -0.0010206921408426179, 1.0,
0.9679090330315554, 0.9999999999999998, 0.053442139957624514,
0.010913960113744634, 0.04898751699418791, 0.014539484718191427,
0.01927338845715717, 0.024295985122396398, 0.020303339878855887,
0.016896922591798955, 0.0177950776252157, 0.09090481965245623],
[0.005242466181763035, 0.0265323496473018, -0.0009504277066825062,
0.9679090330315554, 1.0, 0.9679090330315554, 0.05242629945411761,
0.005508101217625019, 0.04397794536793453, 0.008981488408351649,
0.014492718173379271, 0.01749777640721583, 0.012743733653785828,
0.010052547621304374, 0.010177700977191314, 0.0844154545516134],
[0.006232916329865973, 0.02428356613412483, -0.0010206921408426179,
0.9999999999999998, 0.9679090330315554, 1.0, 0.053442139957624514,
0.010913960113744634, 0.04898751699418791, 0.014539484718191427,
0.01927338845715717, 0.024295985122396398, 0.020303339878855887,
0.016896922591798955, 0.0177950776252157, 0.09090481965245623],
[0.12522219543828494, 0.11117982079071105, 0.004521518981639495,
0.053442139957624514, 0.05242629945411761, 0.053442139957624514, 1.0,
0.0978216192262467, 0.20988344866685532, 0.11848439964079609,
0.15220440418785858, 0.15399382162327566, 0.13739024721262805,
0.11331719837221133, 0.09701953445520294, 0.25948447050554835],
[0.037449604506840636, 0.020313525169298213, -0.0008609769588981912,
0.010913960113744634, 0.005508101217625019, 0.010913960113744634,
0.0978216192262467, 1.0, 0.11257620249092382, 0.9776642994358582,
0.8078747615566414, 0.7024387933993025, 0.8051192766502231, 0.9787082100014113,
0.975723832105259, 0.15061170841954058], [0.10840383773603023,
-0.032858165438786856, -0.008339907100546615, 0.04898751699418791,
0.04397794536793453, 0.04898751699418791, 0.20988344866685532,
0.11257620249092382, 1.0, 0.13749262146291352, 0.11947807651217442,
0.19657551824080938, 0.1713853185458103, 0.12372990572424777,
0.10970326962727345, 0.35111626853661754], [0.04773344107604199,
0.014536490514991217, -0.0009813364242238398, 0.014539484718191427,
0.008981488408351649, 0.014539484718191427, 0.11848439964079609,
0.9776642994358582, 0.13749262146291352, 1.0, 0.829879539373002,
0.6883883341927931, 0.7888253085245929, 0.9540266393812632, 0.9384231749868207,
0.18336003786744906], [0.06439133752559806, 0.024624729226298204,
-0.002386341572889714, 0.019273388457157l7, 0.014492718173379271,
0.01927338845715717, 0.1522044041878585 8, 0.807874761556 6414,
0.11947807651217442, 0.829879539373002, 1.0, 0.555853036766878,
0.631454204090 5395, 0.8565287058577824, 0.8015616561094298,
0.22733183308801955], [0.06227692111916 5635, -0.008714669167627476,
-0.0035900367546440105, 0.024295985122396398, 0.01749777640721583,
0.024295985122396398, 0.15399382162327566, 0.7024387933993025,
0.19657551824080938, 0.6883883341927931, 0.555853036766878, 1.0,
0.9610027518582096, 0.7096534539020688, 0.6676146441610739, 0.2542429847235272],
[0.05252515761478 2124, -0.0038365658456161987, -0.00250317090907312,
0.020303339878855887, 0.012743733653785828, 0.020303339878855887,
0.13739024721262805, 0.8051192766502231, 0.1713853185458103, 0.7888253085245929,
0.631454204090 5395, 0.9610027518582096, 1.0, 0.7885428287299897,
0.7646815806776168, 0.21903535783186218], [0.043813716831314624,
0.022971556258 99625, -0.0015057172595959244, 0.016896922591798955,
0.010052547621304374, 0.016896922591798955, 0.11331719837221133,
0.9787082100014113, 0.12372990572424777, 0.9540266393812632, 0.8565287058577824,
0.7096534539020688, 0.7885428287299897, 1.0, 0.9782125987665806,
0.1809582653627293], [0.03622198669878126, 0.019011603278792415,
-0.0014607609885830214, 0.0177950776252157, 0.010177700977191314,
0.0177950776252157, 0.09701953445520294, 0.975723832105259, 0.10970326962727345,
0.9384231749868207, 0.8015616561094298, 0.6676146441610739, 0.7646815806776168,
0.9782125987665806, 1.0, 0.16101794935654037], [0.12530663336643624,
-0.022024292273003405, -0.013453775274283387, 0.09090481965245623,
0.084415454551 6134, 0.09090481965245623, 0.25948447050554835,
0.15061170841954058, 0.35111626853661754, 0.18336003786744906,
0.22733183308801955, 0.2542429847235272, 0.21903535783186218,
0.1809582653627293, 0.16101794935654037, 1.0]]
```

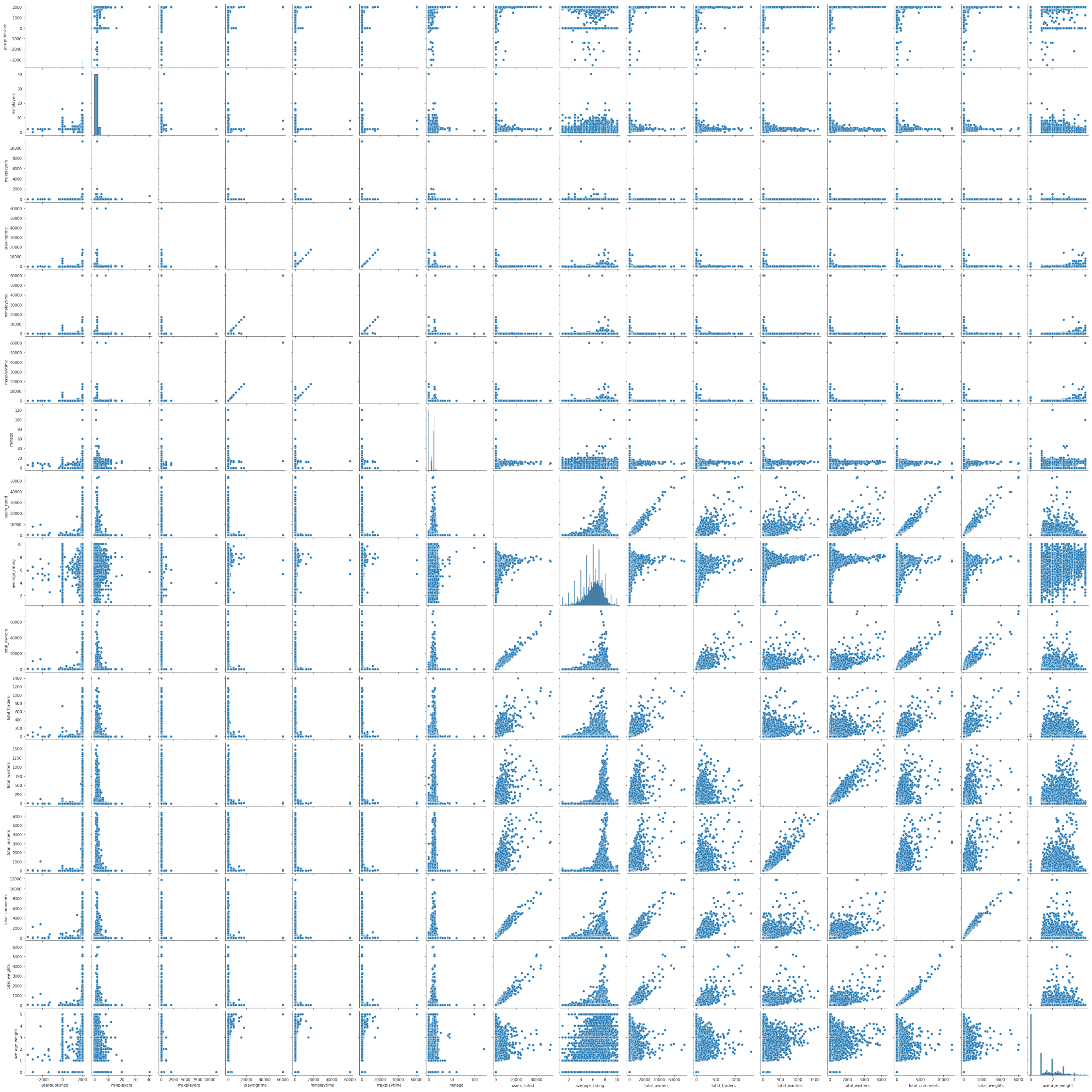
[36]:

```python
# Generating the HeatMap of the Data
figure=plt.figure(1)
figure1=figure.add_subplot(111)
figure1.set_xticklabels(['']+columnsData)
figure1.set_yticklabels(['']+columnsData)
figure.colorbar(figure1.matshow(corrmatrix,vmax=1,vmin=-1))
plt.show()
```

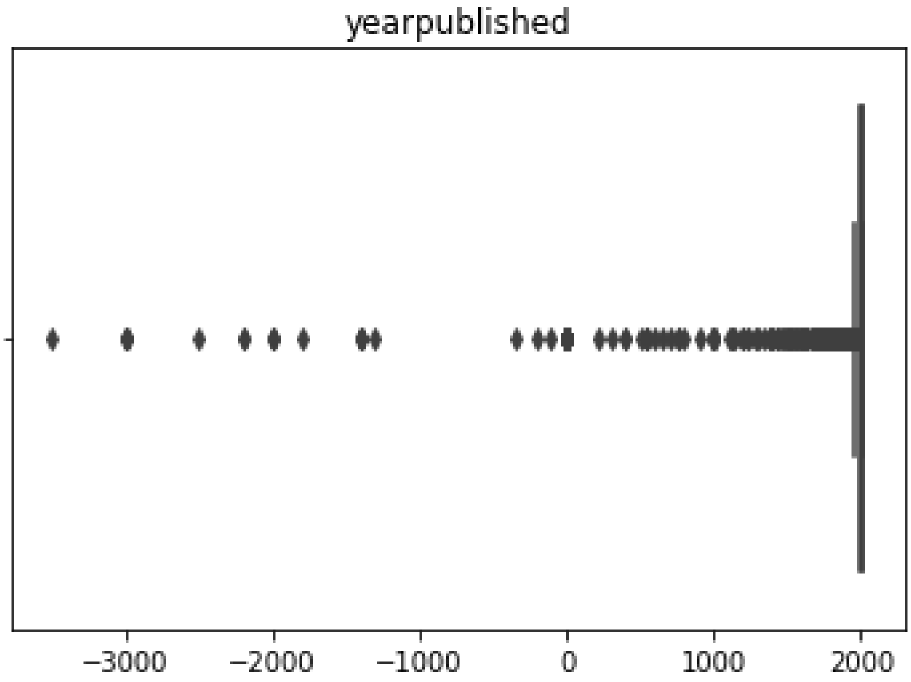


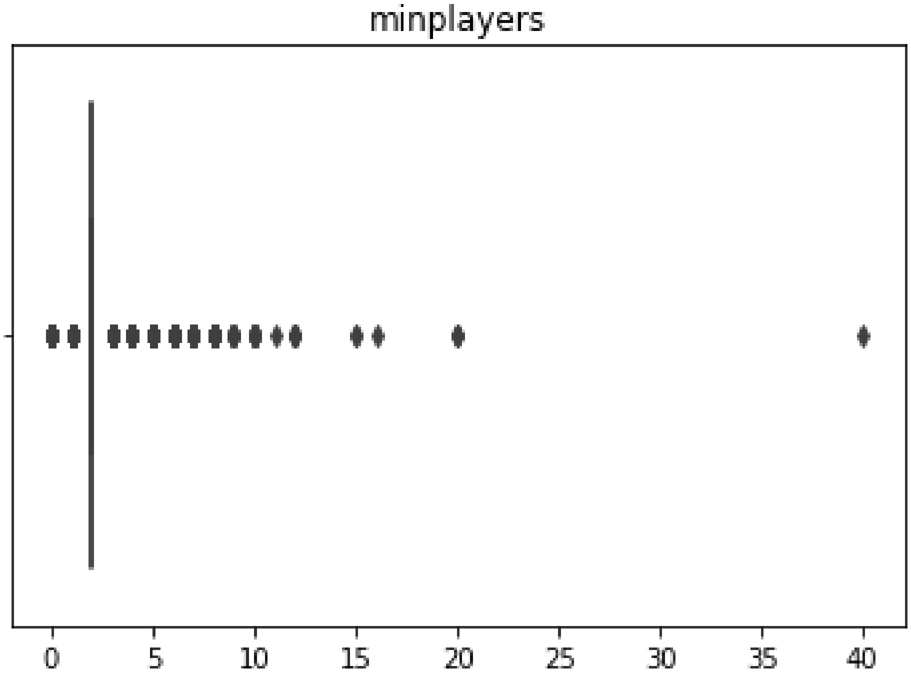
[37]:
```
# Plotting the pairplot
import seaborn as sns
sns.pairplot(df.toPandas())
```

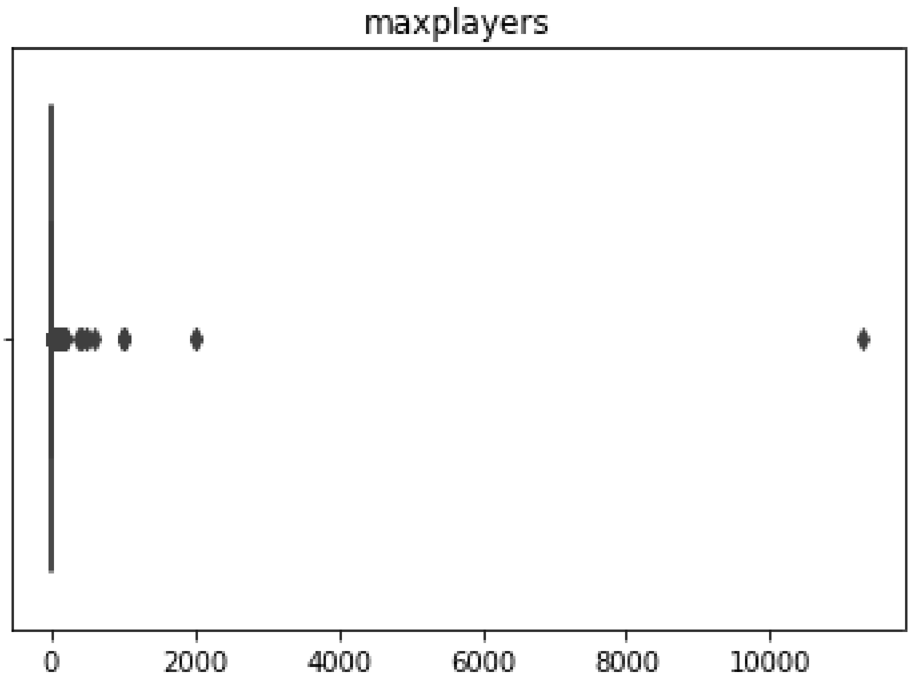
[37]:
```
<seaborn.axisgrid.PairGrid at 0x7f3f5ca5c630>
```

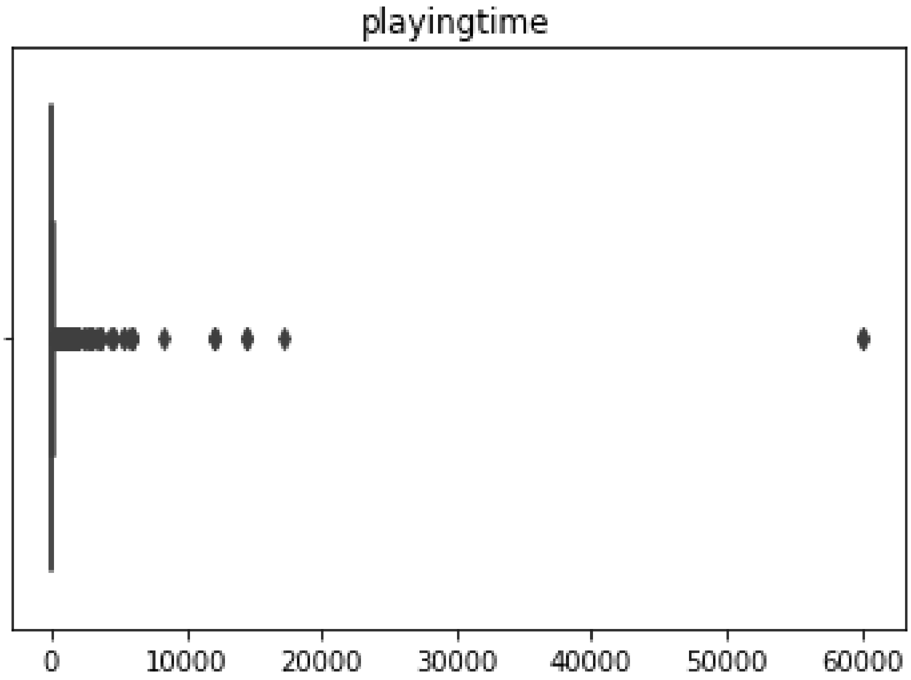


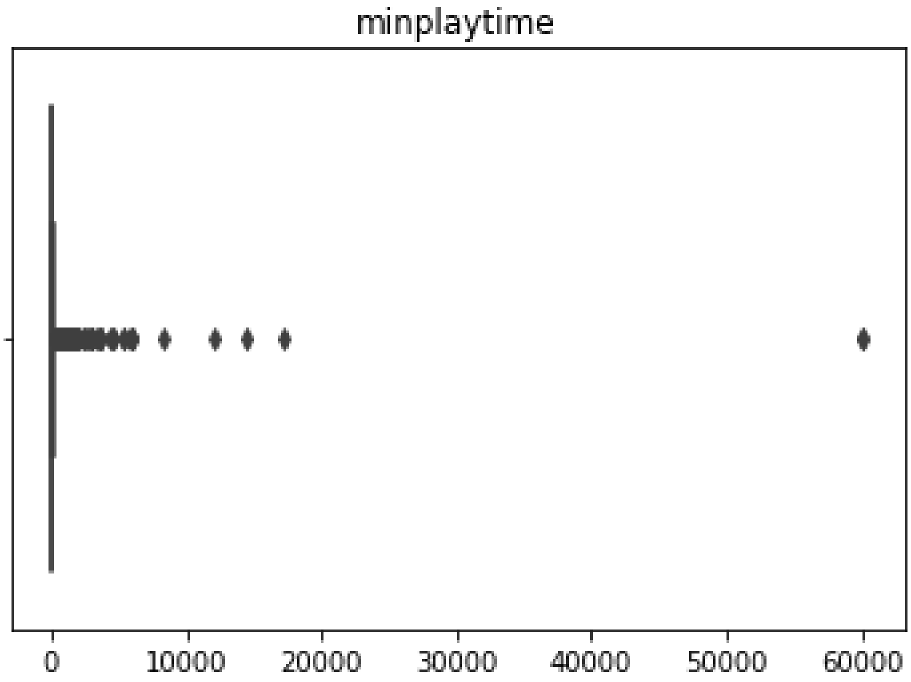
[38]:
```
#Plotting the Box plots
for i in columnsData:
  l = df.select(i).collect()
  l = [r[0] for r in l]
  plt.figure()
  b = sns.boxplot(x = l)
  plt.title(i)
  plt.show()
```

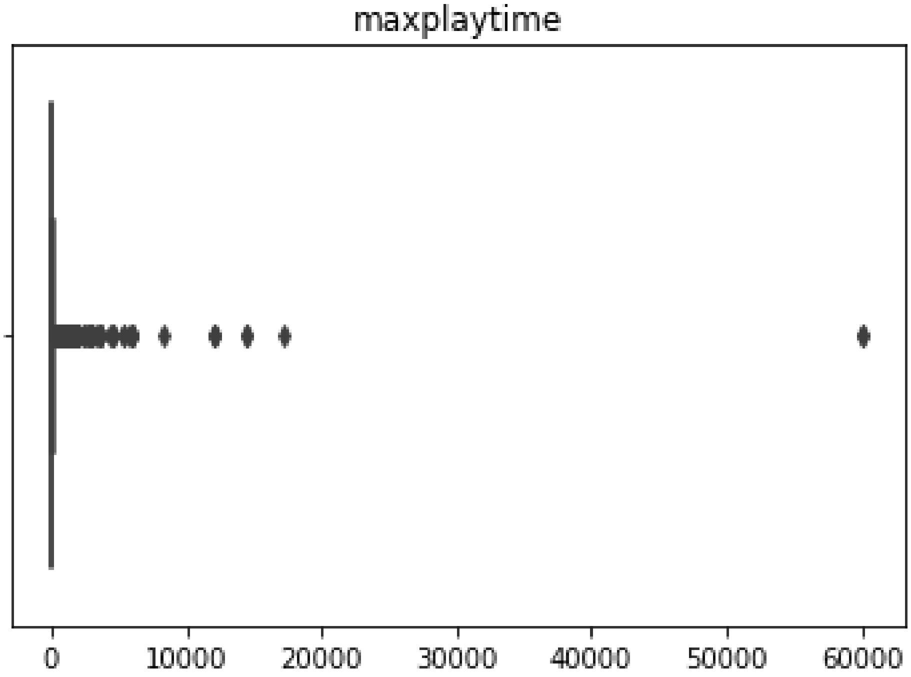

yearpublished
−3000
−2000
−1000
0
1000
2000

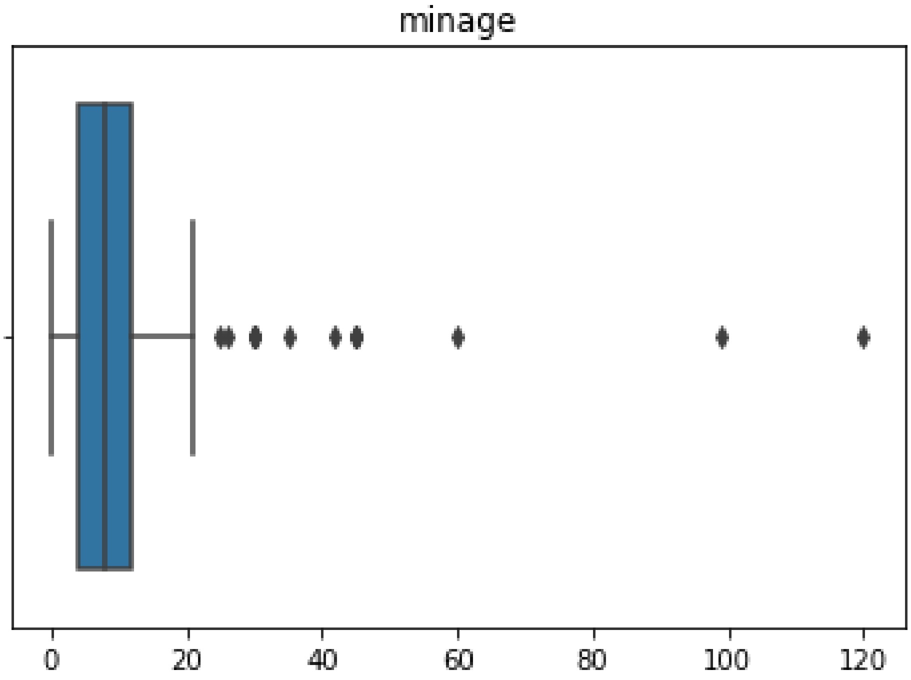

minplayers
0
5
10
15
20
25
30
35
40

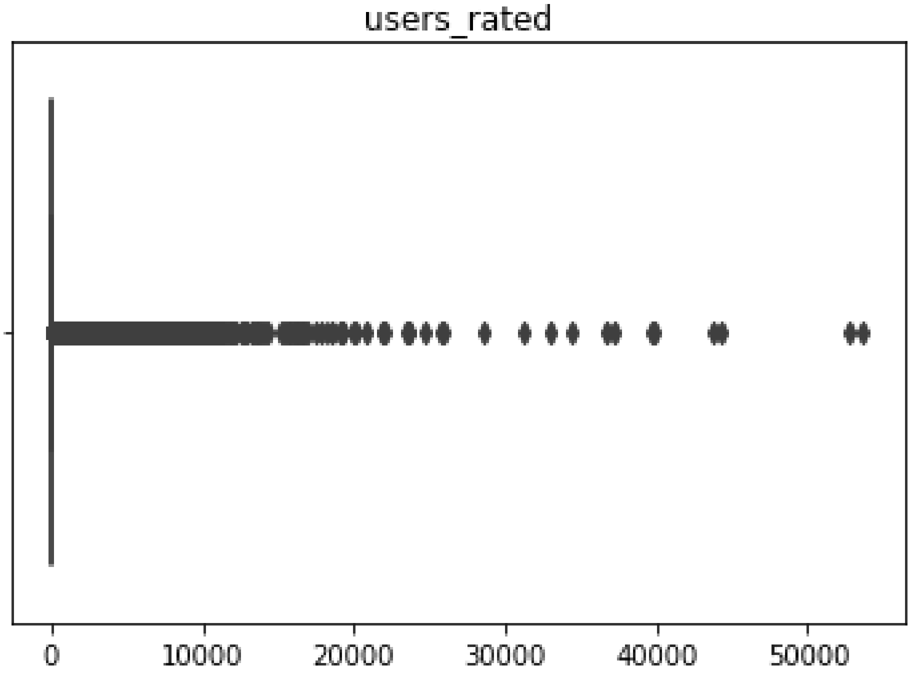

maxplayers
0
2000
4000
6000
8000
10000

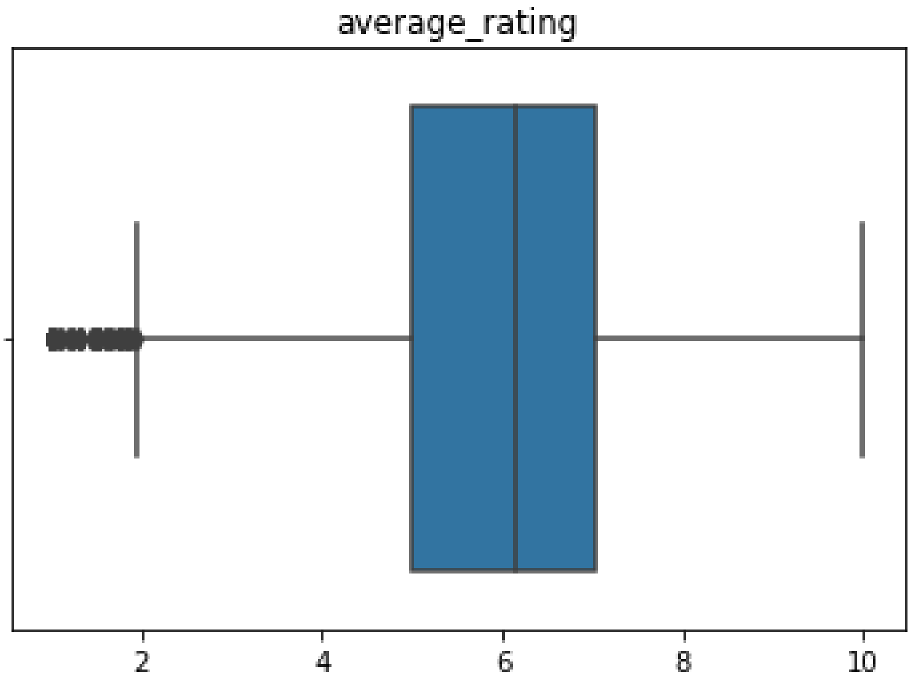

playingtime
0
10000
20000
30000
40000
50000
60000

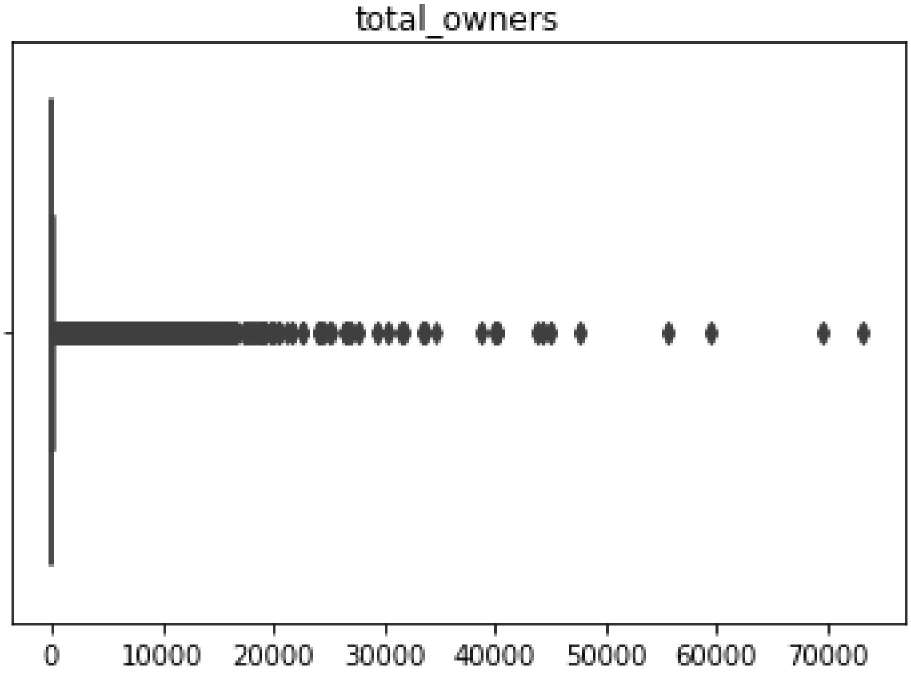

minplaytime
0
10000
20000
30000
40000
50000
60000

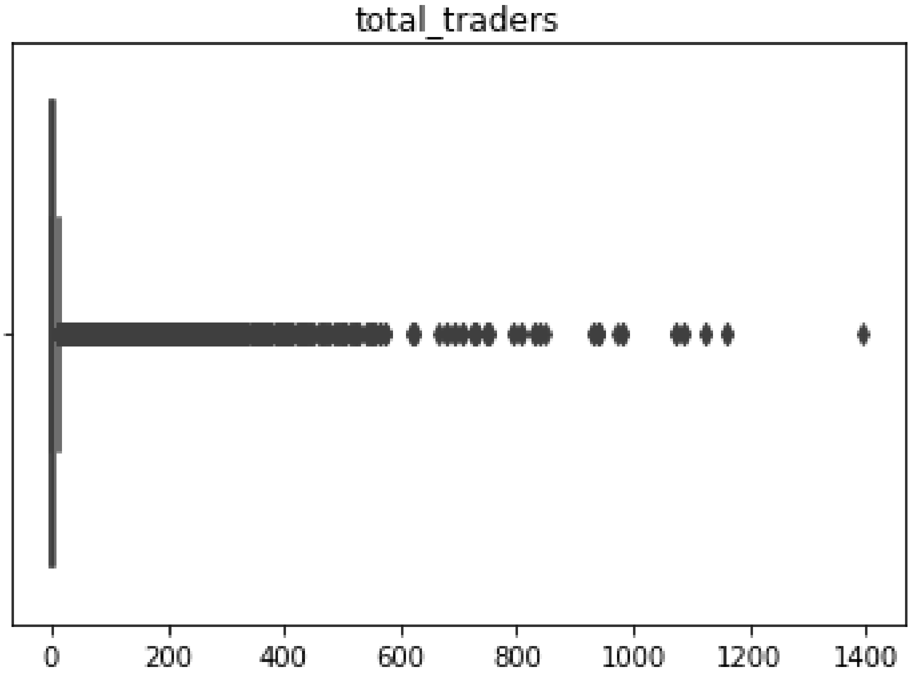

maxplaytime
0
10000
20000
30000
40000
50000
60000

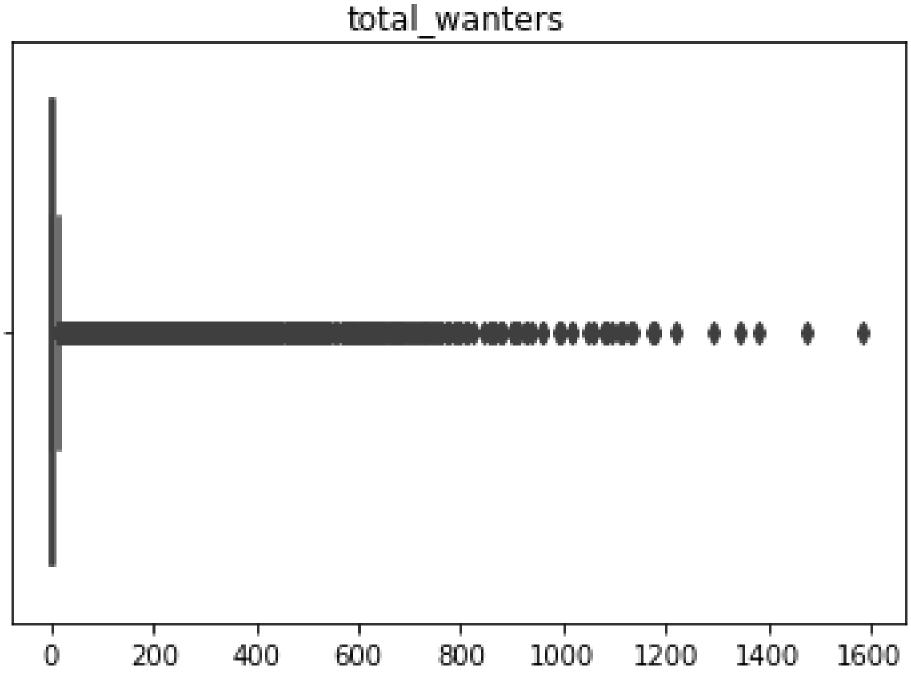

minage
0
20
40
60
80
100
120

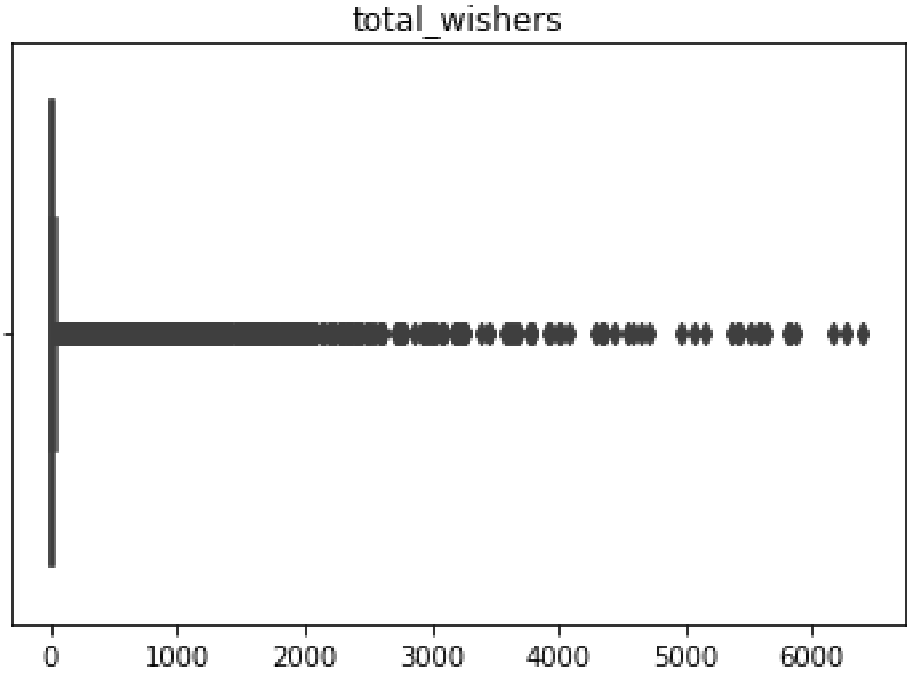

users_rated
0
10000
20000
30000
40000
50000

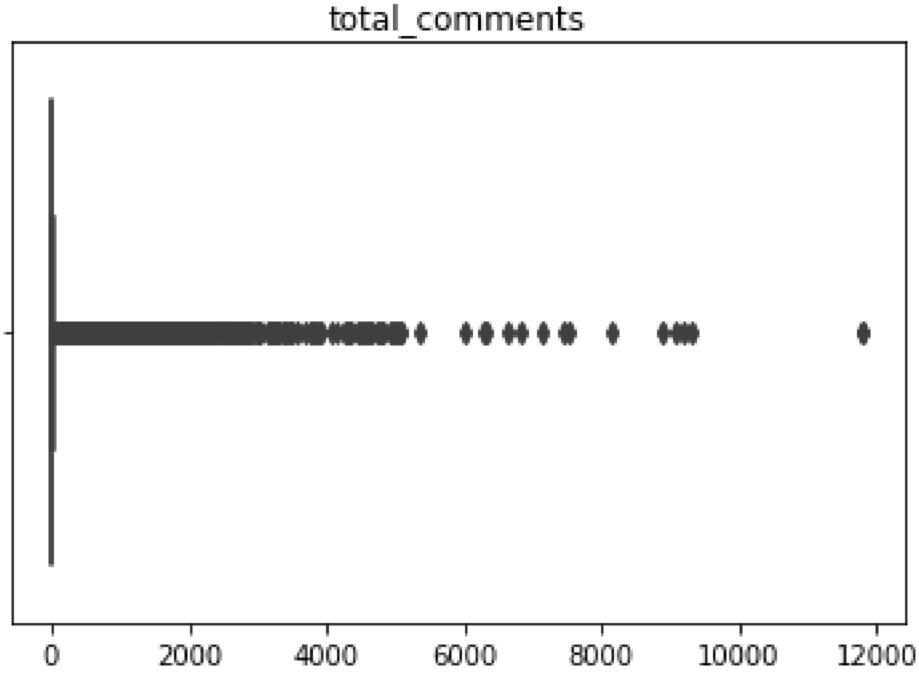

average_rating
2
4
6
8
10

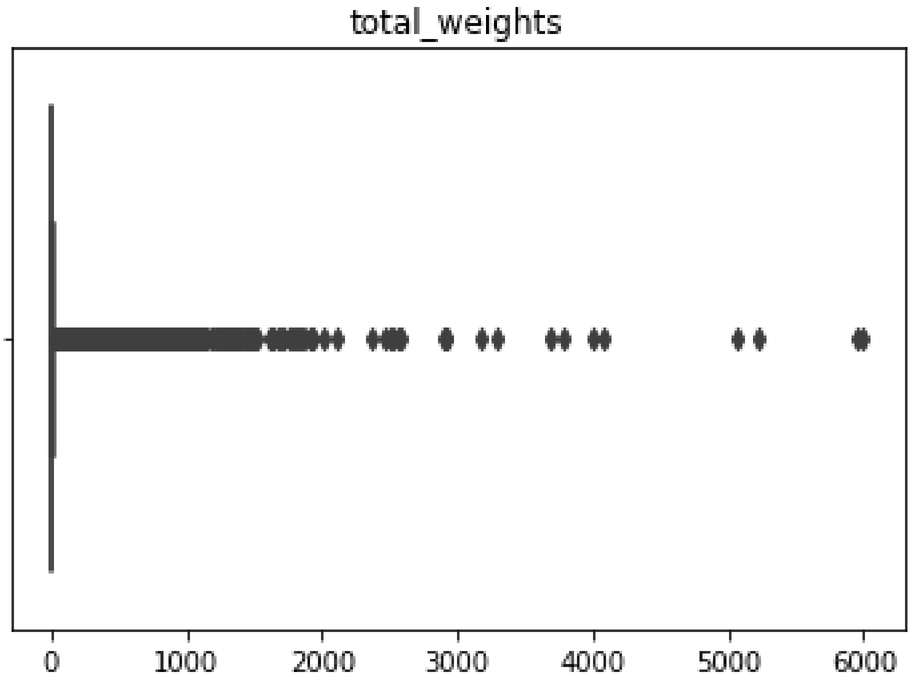

total_owners
0
10000
20000
30000
40000
50000
60000
70000

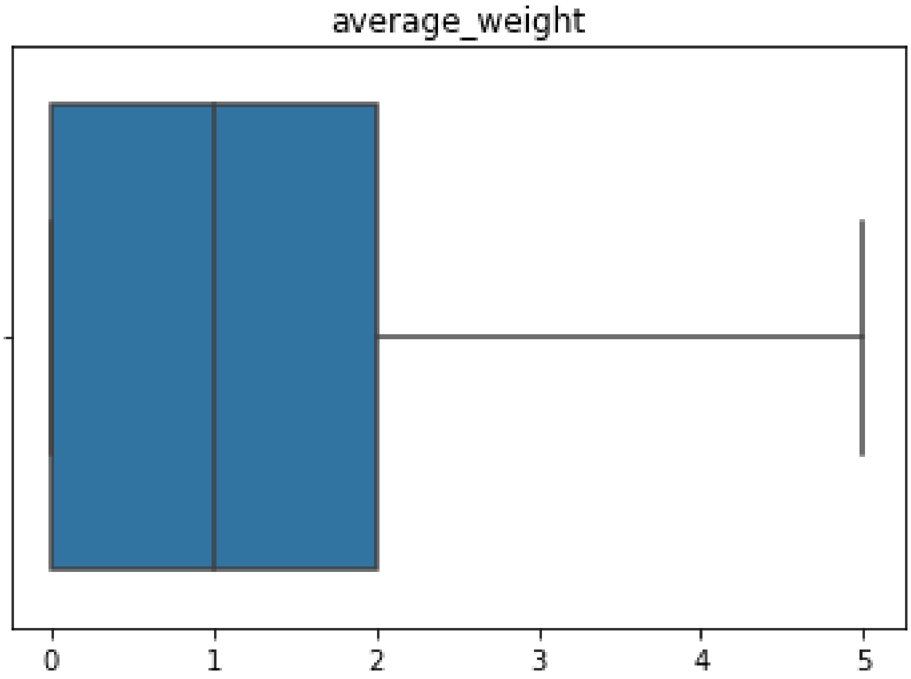

total_traders
0
200
400
600
800
1000
1200
1400


total_wanters
0
200
400
600
800
1000
1200
1400
1600


total_wishers
0
1000
2000
3000
4000
5000
6000


total_comments
0
2000
4000
6000
8000
10000
12000





[39]:
```
#Input all the features into single vectors
```

```
#standard Scaling all the feature vector to handle the outlier observed in the␣
 ↪Boxplots
columnsReg=[c for c in df.columns if c not in␣
 ↪['average_rating','yearpublished']]
assembler = VectorAssembler(inputCols=columnsReg, outputCol = 'Features')
output = assembler.transform(df)
scaled = StandardScaler(inputCol = 'Features',outputCol =␣
 ↪'scaled_features',withStd = True,withMean = True)
final_df = scaled.fit(output).transform(output)
final_df.select('Features','scaled_features').show(1)
finalized_data = final_df.select("Features","average_rating","scaled_features")
finalized_data.show()
```

```
+--------------------+--------------------+
|            Features|     scaled_features|
+--------------------+--------------------+
|[2.0,2.0,180.0,18...|[-0.0063462794810...|
+--------------------+--------------------+
only showing top 1 row

+--------------------+--------------+--------------------+
|            Features|average_rating|     scaled_features|
+--------------------+--------------+--------------------+
|[2.0,2.0,180.0,18...|       8.33774|[-0.0063462794810...|
|[2.0,5.0,150.0,60...|       8.28798|[-0.0063462794810...|
|[1.0,7.0,210.0,30...|       8.28994|[-1.2694999838810...|
|[2.0,4.0,240.0,24...|       8.20407|[-0.0063462794810...|
|[2.0,5.0,150.0,90...|       8.14261|[-0.0063462794810...|
|[1.0,5.0,150.0,30...|       8.11957|[-1.2694999838810...|
|[2.0,2.0,45.0,45...|        8.1676|[-0.0063462794810...|
|[1.0,4.0,150.0,15...|       8.15901|[-1.2694999838810...|
|[2.0,4.0,90.0,30...|       8.07879|[-0.0063462794810...|
|[2.0,6.0,200.0,60...|       8.07933|[-0.0063462794810...|
|[2.0,6.0,120.0,12...|        7.9888|[-0.0063462794810...|
|[2.0,5.0,90.0,90...|       8.43944|[-0.0063462794810...|
|[2.0,4.0,150.0,15...|       8.35044|[-0.0063462794810...|
|[1.0,4.0,180.0,90...|       8.09283|[-1.2694999838810...|
|[1.0,5.0,200.0,10...|       7.99115|[-1.2694999838810...|
|[3.0,4.0,180.0,12...|       8.03071|[1.25680742491897...|
|[2.0,4.0,90.0,90...|       7.98673|[-0.0063462794810...|
|[2.0,5.0,210.0,45...|       8.05776|[-0.0063462794810...|
|[2.0,7.0,30.0,30...|       7.87047|[-0.0063462794810...|
|[2.0,5.0,150.0,60...|       7.89829|[-0.0063462794810...|
+--------------------+--------------+--------------------+
only showing top 20 rows
```

[40]:
```
#test Train split the Data into 80:20 Ratio
train,test = finalized_data.randomSplit([0.8,0.2],seed=42)
regressor = LinearRegression(featuresCol = 'scaled_features', labelCol =␣
↪'average_rating')

#Fitting the Training sent to linearRegression Model
regressor = regressor.fit(train)

#Model Building
pred = regressor.evaluate(test)

#Predicting
pred.predictions.show()
```

```
+--------------------+--------------+--------------------+------------------+
|            Features|average_rating|     scaled_features|        prediction|
+--------------------+--------------+--------------------+------------------+
|(14,[0,1,2,3,4,5,...|           3.0|[-1.2694999838810...| 5.357054248980058|
|(14,[0,1,2,3,4,5,...|           2.0|[-1.2694999838810...| 5.472450147939093|
|(14,[0,1,2,3,4,5,...|           4.0|[-1.2694999838810...| 5.472450147939093|
|(14,[0,1,2,3,4,5,...|           7.0|[-0.0063462794810...| 5.395723754488879|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.434594767362343|
|(14,[0,1,2,3,4,5,...|           2.0|[-0.0063462794810...|  5.39565472239216|
|(14,[0,1,2,3,4,5,...|           4.0|[-0.0063462794810...| 5.434334384004851|
|(14,[0,1,2,3,4,5,...|           1.0|[-0.0063462794810...| 5.511311004708688|
|(14,[0,1,2,3,4,5,...|           5.0|[-0.0063462794810...| 5.511311004708688|
|(14,[0,1,2,3,4,5,...|           1.0|[-0.0063462794810...|  5.39622877617448|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.434717086526398|
|(14,[0,1,2,3,4,5,...|           5.0|[-0.0063462794810...| 5.512267761012554|
|(14,[0,1,2,3,4,5,...|           6.0|[-0.0063462794810...| 5.356844869100723|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...|5.3959072332349605|
|(14,[0,1,2,3,4,5,...|           5.0|[-0.0063462794810...| 5.511433323872742|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.513469155644529|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.513469155644529|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.513469155644529|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.513469155644529|
|(14,[0,1,2,3,4,5,...|           3.0|[-0.0063462794810...| 5.513469155644529|
+--------------------+--------------+--------------------+------------------+
only showing top 20 rows
```

[41]:
```
#calculating the summary to find out the Root mean suared error and Rsquared to␣
↪check the model accuracy
Summary = regressor.summary
print("RMSE: %f" % Summary.rootMeanSquaredError)
print("r2: %f" % Summary.r2)
```

```
RMSE: 1.452548
r2: 0.155437
```

[45]:
```
!sudo apt-get install texlive-xetex texlive-fonts-recommended␣
 ↪texlive-generic-recommended

!jupyter nbconvert --to pdf
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-texgyre
  javascript-common libcupsfilters1 libcupsimage2 libgs9 libgs9-common
  libijs-0.35 libjbig2dec0 libjs-jquery libkpathsea6 libpotrace0 libptexenc1
  libruby2.5 libsynctex1 libtexlua52 libtexluajit2 libzzip-0-13 lmodern
  poppler-data preview-latex-style rake ruby ruby-did-you-mean ruby-minitest
  ruby-net-telnet ruby-power-assert ruby-test-unit ruby2.5
  rubygems-integration t1utils tex-common tex-gyre texlive-base
  texlive-binaries texlive-latex-base texlive-latex-extra
  texlive-latex-recommended texlive-pictures texlive-plain-generic tipa
Suggested packages:
  fonts-noto apache2 | lighttpd | httpd poppler-utils ghostscript
  fonts-japanese-mincho | fonts-ipafont-mincho fonts-japanese-gothic
  | fonts-ipafont-gothic fonts-arphic-ukai fonts-arphic-uming fonts-nanum ri
  ruby-dev bundler debhelper gv | postscript-viewer perl-tk xpdf-reader
  | pdf-viewer texlive-fonts-recommended-doc texlive-latex-base-doc
  python-pygments icc-profiles libfile-which-perl
  libspreadsheet-parseexcel-perl texlive-latex-extra-doc
  texlive-latex-recommended-doc texlive-pstricks dot2tex prerex ruby-tcltk
  | libtcltk-ruby texlive-pictures-doc vprerex
The following NEW packages will be installed:
  fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-texgyre
  javascript-common libcupsfilters1 libcupsimage2 libgs9 libgs9-common
  libijs-0.35 libjbig2dec0 libjs-jquery libkpathsea6 libpotrace0 libptexenc1
  libruby2.5 libsynctex1 libtexlua52 libtexluajit2 libzzip-0-13 lmodern
  poppler-data preview-latex-style rake ruby ruby-did-you-mean ruby-minitest
  ruby-net-telnet ruby-power-assert ruby-test-unit ruby2.5
  rubygems-integration t1utils tex-common tex-gyre texlive-base
  texlive-binaries texlive-fonts-recommended texlive-generic-recommended
  texlive-latex-base texlive-latex-extra texlive-latex-recommended
  texlive-pictures texlive-plain-generic texlive-xetex tipa
0 upgraded, 47 newly installed, 0 to remove and 14 not upgraded.
Need to get 146 MB of archives.
After this operation, 460 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-droid-fallback
all 1:6.0.1r16-1.1 [1,805 kB]
```

```
Get:2 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-lato all 2.0-2
[2,698 kB]
Get:3 http://archive.ubuntu.com/ubuntu bionic/main amd64 poppler-data all
0.4.8-2 [1,479 kB]
Get:4 http://archive.ubuntu.com/ubuntu bionic/main amd64 tex-common all 6.09
[33.0 kB]
Get:5 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-lmodern all
2.004.5-3 [4,551 kB]
Get:6 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-noto-mono all
20171026-2 [75.5 kB]
Get:7 http://archive.ubuntu.com/ubuntu bionic/universe amd64 fonts-texgyre all
20160520-1 [8,761 kB]
Get:8 http://archive.ubuntu.com/ubuntu bionic/main amd64 javascript-common all
11 [6,066 B]
Get:9 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libcupsfilters1
amd64 1.20.2-0ubuntu3.1 [108 kB]
Get:10 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libcupsimage2
amd64 2.2.7-1ubuntu2.8 [18.6 kB]
Get:11 http://archive.ubuntu.com/ubuntu bionic/main amd64 libijs-0.35 amd64
0.35-13 [15.5 kB]
Get:12 http://archive.ubuntu.com/ubuntu bionic/main amd64 libjbig2dec0 amd64
0.13-6 [55.9 kB]
Get:13 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libgs9-common
all 9.26~dfsg+0-0ubuntu0.18.04.13 [5,092 kB]
Get:14 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libgs9 amd64
9.26~dfsg+0-0ubuntu0.18.04.13 [2,263 kB]
Get:15 http://archive.ubuntu.com/ubuntu bionic/main amd64 libjs-jquery all
3.2.1-1 [152 kB]
Get:16 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libkpathsea6
amd64 2017.20170613.44572-8ubuntu0.1 [54.9 kB]
Get:17 http://archive.ubuntu.com/ubuntu bionic/main amd64 libpotrace0 amd64
1.14-2 [17.4 kB]
Get:18 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libptexenc1
amd64 2017.20170613.44572-8ubuntu0.1 [34.5 kB]
Get:19 http://archive.ubuntu.com/ubuntu bionic/main amd64 rubygems-integration
all 1.11 [4,994 B]
Get:20 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 ruby2.5 amd64
2.5.1-1ubuntu1.7 [48.6 kB]
Get:21 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby amd64 1:2.5.1
[5,712 B]
Get:22 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 rake all
12.3.1-1ubuntu0.1 [44.9 kB]
Get:23 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-did-you-mean all
1.2.0-2 [9,700 B]
Get:24 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-minitest all
5.10.3-1 [38.6 kB]
Get:25 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-net-telnet all
0.1.1-2 [12.6 kB]
```

```
Get:26 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-power-assert all
0.3.0-1 [7,952 B]
Get:27 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-test-unit all
3.2.5-1 [61.1 kB]
Get:28 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libruby2.5
amd64 2.5.1-1ubuntu1.7 [3,068 kB]
Get:29 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libsynctex1
amd64 2017.20170613.44572-8ubuntu0.1 [41.4 kB]
Get:30 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libtexlua52
amd64 2017.20170613.44572-8ubuntu0.1 [91.2 kB]
Get:31 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libtexluajit2
amd64 2017.20170613.44572-8ubuntu0.1 [230 kB]
Get:32 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libzzip-0-13
amd64 0.13.62-3.1ubuntu0.18.04.1 [26.0 kB]
Get:33 http://archive.ubuntu.com/ubuntu bionic/main amd64 lmodern all 2.004.5-3
[9,631 kB]
Get:34 http://archive.ubuntu.com/ubuntu bionic/main amd64 preview-latex-style
all 11.91-1ubuntu1 [185 kB]
Get:35 http://archive.ubuntu.com/ubuntu bionic/main amd64 t1utils amd64 1.41-2
[56.0 kB]
Get:36 http://archive.ubuntu.com/ubuntu bionic/universe amd64 tex-gyre all
20160520-1 [4,998 kB]
Get:37 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 texlive-
binaries amd64 2017.20170613.44572-8ubuntu0.1 [8,179 kB]
Get:38 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-base all
2017.20180305-1 [18.7 MB]
Get:39 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-fonts-
recommended all 2017.20180305-1 [5,262 kB]
Get:40 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-plain-
generic all 2017.20180305-2 [23.6 MB]
Get:41 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-generic-
recommended all 2017.20180305-1 [15.9 kB]
Get:42 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-latex-base all
2017.20180305-1 [951 kB]
Get:43 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-latex-
recommended all 2017.20180305-1 [14.9 MB]
Get:44 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-pictures
all 2017.20180305-1 [4,026 kB]
Get:45 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-latex-
extra all 2017.20180305-2 [10.6 MB]
Get:46 http://archive.ubuntu.com/ubuntu bionic/universe amd64 tipa all 2:1.3-20
[2,978 kB]
Get:47 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-xetex all
2017.20180305-1 [10.7 MB]
Fetched 146 MB in 8s (17.5 MB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76,
<> line 47.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package fonts-droid-fallback.
(Reading database ... 145208 files and directories currently installed.)
Preparing to unpack .../00-fonts-droid-fallback_1%3a6.0.1r16-1.1_all.deb ...
Unpacking fonts-droid-fallback (1:6.0.1r16-1.1) ...
Selecting previously unselected package fonts-lato.
Preparing to unpack .../01-fonts-lato_2.0-2_all.deb ...
Unpacking fonts-lato (2.0-2) ...
Selecting previously unselected package poppler-data.
Preparing to unpack .../02-poppler-data_0.4.8-2_all.deb ...
Unpacking poppler-data (0.4.8-2) ...
Selecting previously unselected package tex-common.
Preparing to unpack .../03-tex-common_6.09_all.deb ...
Unpacking tex-common (6.09) ...
Selecting previously unselected package fonts-lmodern.
Preparing to unpack .../04-fonts-lmodern_2.004.5-3_all.deb ...
Unpacking fonts-lmodern (2.004.5-3) ...
Selecting previously unselected package fonts-noto-mono.
Preparing to unpack .../05-fonts-noto-mono_20171026-2_all.deb ...
Unpacking fonts-noto-mono (20171026-2) ...
Selecting previously unselected package fonts-texgyre.
Preparing to unpack .../06-fonts-texgyre_20160520-1_all.deb ...
Unpacking fonts-texgyre (20160520-1) ...
Selecting previously unselected package javascript-common.
Preparing to unpack .../07-javascript-common_11_all.deb ...
Unpacking javascript-common (11) ...
Selecting previously unselected package libcupsfilters1:amd64.
Preparing to unpack .../08-libcupsfilters1_1.20.2-0ubuntu3.1_amd64.deb ...
Unpacking libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) ...
Selecting previously unselected package libcupsimage2:amd64.
Preparing to unpack .../09-libcupsimage2_2.2.7-1ubuntu2.8_amd64.deb ...
Unpacking libcupsimage2:amd64 (2.2.7-1ubuntu2.8) ...
Selecting previously unselected package libijs-0.35:amd64.
Preparing to unpack .../10-libijs-0.35_0.35-13_amd64.deb ...
Unpacking libijs-0.35:amd64 (0.35-13) ...
Selecting previously unselected package libjbig2dec0:amd64.
Preparing to unpack .../11-libjbig2dec0_0.13-6_amd64.deb ...
Unpacking libjbig2dec0:amd64 (0.13-6) ...
Selecting previously unselected package libgs9-common.
Preparing to unpack .../12-libgs9-common_9.26~dfsg+0-0ubuntu0.18.04.13_all.deb
...
Unpacking libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.13) ...
Selecting previously unselected package libgs9:amd64.
```

```
Preparing to unpack .../13-libgs9_9.26~dfsg+0-0ubuntu0.18.04.13_amd64.deb ...
Unpacking libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.13) ...
Selecting previously unselected package libjs-jquery.
Preparing to unpack .../14-libjs-jquery_3.2.1-1_all.deb ...
Unpacking libjs-jquery (3.2.1-1) ...
Selecting previously unselected package libkpathsea6:amd64.
Preparing to unpack .../15-libkpathsea6_2017.20170613.44572-8ubuntu0.1_amd64.deb
...
Unpacking libkpathsea6:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package libpotrace0.
Preparing to unpack .../16-libpotrace0_1.14-2_amd64.deb ...
Unpacking libpotrace0 (1.14-2) ...
Selecting previously unselected package libptexenc1:amd64.
Preparing to unpack .../17-libptexenc1_2017.20170613.44572-8ubuntu0.1_amd64.deb
...
Unpacking libptexenc1:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package rubygems-integration.
Preparing to unpack .../18-rubygems-integration_1.11_all.deb ...
Unpacking rubygems-integration (1.11) ...
Selecting previously unselected package ruby2.5.
Preparing to unpack .../19-ruby2.5_2.5.1-1ubuntu1.7_amd64.deb ...
Unpacking ruby2.5 (2.5.1-1ubuntu1.7) ...
Selecting previously unselected package ruby.
Preparing to unpack .../20-ruby_1%3a2.5.1_amd64.deb ...
Unpacking ruby (1:2.5.1) ...
Selecting previously unselected package rake.
Preparing to unpack .../21-rake_12.3.1-1ubuntu0.1_all.deb ...
Unpacking rake (12.3.1-1ubuntu0.1) ...
Selecting previously unselected package ruby-did-you-mean.
Preparing to unpack .../22-ruby-did-you-mean_1.2.0-2_all.deb ...
Unpacking ruby-did-you-mean (1.2.0-2) ...
Selecting previously unselected package ruby-minitest.
Preparing to unpack .../23-ruby-minitest_5.10.3-1_all.deb ...
Unpacking ruby-minitest (5.10.3-1) ...
Selecting previously unselected package ruby-net-telnet.
Preparing to unpack .../24-ruby-net-telnet_0.1.1-2_all.deb ...
Unpacking ruby-net-telnet (0.1.1-2) ...
Selecting previously unselected package ruby-power-assert.
Preparing to unpack .../25-ruby-power-assert_0.3.0-1_all.deb ...
Unpacking ruby-power-assert (0.3.0-1) ...
Selecting previously unselected package ruby-test-unit.
Preparing to unpack .../26-ruby-test-unit_3.2.5-1_all.deb ...
Unpacking ruby-test-unit (3.2.5-1) ...
Selecting previously unselected package libruby2.5:amd64.
Preparing to unpack .../27-libruby2.5_2.5.1-1ubuntu1.7_amd64.deb ...
Unpacking libruby2.5:amd64 (2.5.1-1ubuntu1.7) ...
Selecting previously unselected package libsynctex1:amd64.
Preparing to unpack .../28-libsynctex1_2017.20170613.44572-8ubuntu0.1_amd64.deb
```

```
...
Unpacking libsynctex1:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package libtexlua52:amd64.
Preparing to unpack .../29-libtexlua52_2017.20170613.44572-8ubuntu0.1_amd64.deb
...
Unpacking libtexlua52:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package libtexluajit2:amd64.
Preparing to unpack
.../30-libtexluajit2_2017.20170613.44572-8ubuntu0.1_amd64.deb ...
Unpacking libtexluajit2:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package libzzip-0-13:amd64.
Preparing to unpack .../31-libzzip-0-13_0.13.62-3.1ubuntu0.18.04.1_amd64.deb ...
Unpacking libzzip-0-13:amd64 (0.13.62-3.1ubuntu0.18.04.1) ...
Selecting previously unselected package lmodern.
Preparing to unpack .../32-lmodern_2.004.5-3_all.deb ...
Unpacking lmodern (2.004.5-3) ...
Selecting previously unselected package preview-latex-style.
Preparing to unpack .../33-preview-latex-style_11.91-1ubuntu1_all.deb ...
Unpacking preview-latex-style (11.91-1ubuntu1) ...
Selecting previously unselected package t1utils.
Preparing to unpack .../34-t1utils_1.41-2_amd64.deb ...
Unpacking t1utils (1.41-2) ...
Selecting previously unselected package tex-gyre.
Preparing to unpack .../35-tex-gyre_20160520-1_all.deb ...
Unpacking tex-gyre (20160520-1) ...
Selecting previously unselected package texlive-binaries.
Preparing to unpack .../36-texlive-
binaries_2017.20170613.44572-8ubuntu0.1_amd64.deb ...
Unpacking texlive-binaries (2017.20170613.44572-8ubuntu0.1) ...
Selecting previously unselected package texlive-base.
Preparing to unpack .../37-texlive-base_2017.20180305-1_all.deb ...
Unpacking texlive-base (2017.20180305-1) ...
Selecting previously unselected package texlive-fonts-recommended.
Preparing to unpack .../38-texlive-fonts-recommended_2017.20180305-1_all.deb ...
Unpacking texlive-fonts-recommended (2017.20180305-1) ...
Selecting previously unselected package texlive-plain-generic.
Preparing to unpack .../39-texlive-plain-generic_2017.20180305-2_all.deb ...
Unpacking texlive-plain-generic (2017.20180305-2) ...
Selecting previously unselected package texlive-generic-recommended.
Preparing to unpack .../40-texlive-generic-recommended_2017.20180305-1_all.deb
...
Unpacking texlive-generic-recommended (2017.20180305-1) ...
Selecting previously unselected package texlive-latex-base.
Preparing to unpack .../41-texlive-latex-base_2017.20180305-1_all.deb ...
Unpacking texlive-latex-base (2017.20180305-1) ...
Selecting previously unselected package texlive-latex-recommended.
Preparing to unpack .../42-texlive-latex-recommended_2017.20180305-1_all.deb ...
Unpacking texlive-latex-recommended (2017.20180305-1) ...
```

```
Selecting previously unselected package texlive-pictures.
Preparing to unpack .../43-texlive-pictures_2017.20180305-1_all.deb ...
Unpacking texlive-pictures (2017.20180305-1) ...
Selecting previously unselected package texlive-latex-extra.
Preparing to unpack .../44-texlive-latex-extra_2017.20180305-2_all.deb ...
Unpacking texlive-latex-extra (2017.20180305-2) ...
Selecting previously unselected package tipa.
Preparing to unpack .../45-tipa_2%3a1.3-20_all.deb ...
Unpacking tipa (2:1.3-20) ...
Selecting previously unselected package texlive-xetex.
Preparing to unpack .../46-texlive-xetex_2017.20180305-1_all.deb ...
Unpacking texlive-xetex (2017.20180305-1) ...
Setting up libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.13) ...
Setting up libkpathsea6:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Setting up libjs-jquery (3.2.1-1) ...
Setting up libtexlua52:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Setting up fonts-droid-fallback (1:6.0.1r16-1.1) ...
Setting up libsynctex1:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Setting up libptexenc1:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Setting up tex-common (6.09) ...
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line
76.)
debconf: falling back to frontend: Readline
update-language: texlive-base not installed and configured, doing nothing!
Setting up poppler-data (0.4.8-2) ...
Setting up tex-gyre (20160520-1) ...
Setting up preview-latex-style (11.91-1ubuntu1) ...
Setting up fonts-texgyre (20160520-1) ...
Setting up fonts-noto-mono (20171026-2) ...
Setting up fonts-lato (2.0-2) ...
Setting up libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) ...
Setting up libcupsimage2:amd64 (2.2.7-1ubuntu2.8) ...
Setting up libjbig2dec0:amd64 (0.13-6) ...
Setting up ruby-did-you-mean (1.2.0-2) ...
Setting up t1utils (1.41-2) ...
Setting up ruby-net-telnet (0.1.1-2) ...
Setting up libijs-0.35:amd64 (0.35-13) ...
Setting up rubygems-integration (1.11) ...
Setting up libpotrace0 (1.14-2) ...
Setting up javascript-common (11) ...
Setting up ruby-minitest (5.10.3-1) ...
Setting up libzzip-0-13:amd64 (0.13.62-3.1ubuntu0.18.04.1) ...
Setting up libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.13) ...
Setting up libtexluajit2:amd64 (2017.20170613.44572-8ubuntu0.1) ...
Setting up fonts-lmodern (2.004.5-3) ...
Setting up ruby-power-assert (0.3.0-1) ...
Setting up texlive-binaries (2017.20170613.44572-8ubuntu0.1) ...
update-alternatives: using /usr/bin/xdvi-xaw to provide /usr/bin/xdvi.bin
(xdvi.bin) in auto mode
update-alternatives: using /usr/bin/bibtex.original to provide /usr/bin/bibtex
(bibtex) in auto mode
Setting up texlive-base (2017.20180305-1) ...
mktexlsr: Updating /var/lib/texmf/ls-R-TEXLIVEDIST...
mktexlsr: Updating /var/lib/texmf/ls-R-TEXMFMAIN...
mktexlsr: Updating /var/lib/texmf/ls-R...
mktexlsr: Done.
tl-paper: setting paper size for dvips to a4: /var/lib/texmf/dvips/config
/config-paper.ps
tl-paper: setting paper size for dvipdfmx to a4: /var/lib/texmf/dvipdfmx
/dvipdfmx-paper.cfg
tl-paper: setting paper size for xdvi to a4: /var/lib/texmf/xdvi/XDvi-paper
tl-paper: setting paper size for pdftex to a4:
/var/lib/texmf/tex/generic/config/pdftexconfig.tex
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line
76.)
debconf: falling back to frontend: Readline
Setting up texlive-fonts-recommended (2017.20180305-1) ...
Setting up texlive-plain-generic (2017.20180305-2) ...
Setting up texlive-generic-recommended (2017.20180305-1) ...
Setting up texlive-latex-base (2017.20180305-1) ...
Setting up lmodern (2.004.5-3) ...
Setting up texlive-latex-recommended (2017.20180305-1) ...
Setting up texlive-pictures (2017.20180305-1) ...
Setting up tipa (2:1.3-20) ...
Regenerating '/var/lib/texmf/fmtutil.cnf-DEBIAN'... done.
Regenerating '/var/lib/texmf/fmtutil.cnf-TEXLIVEDIST'... done.
update-fmtutil has updated the following file(s):
        /var/lib/texmf/fmtutil.cnf-DEBIAN
        /var/lib/texmf/fmtutil.cnf-TEXLIVEDIST
If you want to activate the changes in the above file(s),
you should run fmtutil-sys or fmtutil.
Setting up texlive-latex-extra (2017.20180305-2) ...
Setting up texlive-xetex (2017.20180305-1) ...
Setting up ruby2.5 (2.5.1-1ubuntu1.7) ...
Setting up ruby (1:2.5.1) ...
Setting up ruby-test-unit (3.2.5-1) ...
Setting up rake (12.3.1-1ubuntu0.1) ...
Setting up libruby2.5:amd64 (2.5.1-1ubuntu1.7) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for libc-bin (2.27-3ubuntu1.2) ...
/sbin/ldconfig.real: /usr/local/lib/python3.6/dist-
packages/ideep4py/lib/libmkldnn.so.0 is not a symbolic link
```

```
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
Processing triggers for fontconfig (2.12.6-0ubuntu2) ...
Processing triggers for tex-common (6.09) ...
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line
76.)
debconf: falling back to frontend: Readline
Running updmap-sys. This may take some time... done.
Running mktexlsr /var/lib/texmf ... done.
Building format(s) --all.
        This may take some time... done.
This application is used to convert notebook files (*.ipynb) to various other
formats.

WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
-------

Arguments that take values are actually convenience aliases to full
Configurables, whose aliases are listed on the help line. For more information
on full configurables, see '--help-all'.

--execute
    Execute the notebook prior to export.
--allow-errors
    Continue notebook execution even if one of the cells throws an error and
include the error message in the cell output (the default behaviour is to abort
conversion). This flag is only relevant if '--execute' was specified, too.
--no-input
    Exclude input cells and output prompts from converted document.
    This mode is ideal for generating code-free reports.
--stdout
    Write notebook output to stdout instead of files.
--stdin
    read a single notebook file from stdin. Write the resulting notebook with
default basename 'notebook.*'
--inplace
    Run nbconvert in place, overwriting the existing notebook (only
    relevant when converting to notebook format)
-y
    Answer yes to any questions instead of prompting.
--clear-output
    Clear output of current file and save in place,
    overwriting the existing notebook.
--debug
```

```
    set log level to logging.DEBUG (maximize logging output)
--no-prompt
    Exclude input and output prompts from converted document.
--generate-config
    generate default config file
--nbformat=<Enum> (NotebookExporter.nbformat_version)
    Default: 4
    Choices: [1, 2, 3, 4]
    The nbformat version to write. Use this to downgrade notebooks.
--output-dir=<Unicode> (FilesWriter.build_directory)
    Default: ''
    Directory to write output(s) to. Defaults to output to the directory of each
    notebook. To recover previous default behaviour (outputting to the current
    working directory) use . as the flag value.
--writer=<DottedObjectName> (NbConvertApp.writer_class)
    Default: 'FilesWriter'
    Writer class used to write the  results of the conversion
--log-level=<Enum> (Application.log_level)
    Default: 30
    Choices: (0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR',
'CRITICAL')
    Set the log level by value or name.
--reveal-prefix=<Unicode> (SlidesExporter.reveal_url_prefix)
    Default: u''
    The URL prefix for reveal.js (version 3.x). This defaults to the reveal CDN,
    but can be any url pointing to a copy  of reveal.js.
    For speaker notes to work, this must be a relative path to a local  copy of
    reveal.js: e.g., "reveal.js".
    If a relative path is given, it must be a subdirectory of the current
    directory (from which the server is run).
    See the usage documentation
    (https://nbconvert.readthedocs.io/en/latest/usage.html#reveal-js-html-
    slideshow) for more details.
--to=<Unicode> (NbConvertApp.export_format)
    Default: 'html'
    The export format to be used, either one of the built-in formats
    ['asciidoc', 'custom', 'html', 'latex', 'markdown', 'notebook', 'pdf',
    'python', 'rst', 'script', 'slides'] or a dotted object name that represents
    the import path for an `Exporter` class
--template=<Unicode> (TemplateExporter.template_file)
    Default: u''
    Name of the template file to use
--output=<Unicode> (NbConvertApp.output_base)
    Default: ''
    overwrite base name use for output files. can only be used when converting
    one notebook at a time.
--post=<DottedOrNone> (NbConvertApp.postprocessor_class)
    Default: u''
    PostProcessor class used to write the results of the conversion
--config=<Unicode> (JupyterApp.config_file)
    Default: u''
    Full path of a config file.

To see all available configurables, use `--help-all`

Examples
--------

    The simplest way to use nbconvert is

    > jupyter nbconvert mynotebook.ipynb

    which will convert mynotebook.ipynb to the default format (probably HTML).

    You can specify the export format with `--to`.
    Options include ['asciidoc', 'custom', 'html', 'latex', 'markdown',
'notebook', 'pdf', 'python', 'rst', 'script', 'slides'].

    > jupyter nbconvert --to latex mynotebook.ipynb

    Both HTML and LaTeX support multiple output templates. LaTeX includes
    'base', 'article' and 'report'.  HTML includes 'basic' and 'full'. You
    can specify the flavor of the format used.

    > jupyter nbconvert --to html --template basic mynotebook.ipynb

    You can also pipe the output to stdout, rather than a file

    > jupyter nbconvert mynotebook.ipynb --stdout

    PDF is generated via latex

    > jupyter nbconvert mynotebook.ipynb --to pdf

    You can get (and serve) a Reveal.js-powered slideshow

    > jupyter nbconvert myslides.ipynb --to slides --post serve

    Multiple notebooks can be given at the command line in a couple of
    different ways:

    > jupyter nbconvert notebook*.ipynb
    > jupyter nbconvert notebook1.ipynb notebook2.ipynb

    or you can specify the notebooks list in a config file, containing::
        c.NbConvertApp.notebooks = ["my_notebook.ipynb"]

    > jupyter nbconvert --config mycfg.py
```

[47]:

```
!jupyter nbconvert --to pdf RegressionSpark.ipynb
```

```
[NbConvertApp] WARNING | pattern u'RegressionSpark.ipynb' matched no files
This application is used to convert notebook files (*.ipynb) to various other
formats.

WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
-------

Arguments that take values are actually convenience aliases to full
Configurables, whose aliases are listed on the help line. For more information
on full configurables, see '--help-all'.

--execute
    Execute the notebook prior to export.
--allow-errors
    Continue notebook execution even if one of the cells throws an error and
include the error message in the cell output (the default behaviour is to abort
conversion). This flag is only relevant if '--execute' was specified, too.
--no-input
    Exclude input cells and output prompts from converted document.
    This mode is ideal for generating code-free reports.
--stdout
    Write notebook output to stdout instead of files.
--stdin
    read a single notebook file from stdin. Write the resulting notebook with
default basename 'notebook.*'
--inplace
    Run nbconvert in place, overwriting the existing notebook (only
    relevant when converting to notebook format)
-y
    Answer yes to any questions instead of prompting.
--clear-output
    Clear output of current file and save in place,
    overwriting the existing notebook.
--debug
    set log level to logging.DEBUG (maximize logging output)
--no-prompt
    Exclude input and output prompts from converted document.
--generate-config
```

```
    generate default config file
--nbformat=<Enum> (NotebookExporter.nbformat_version)
    Default: 4
    Choices: [1, 2, 3, 4]
    The nbformat version to write. Use this to downgrade notebooks.
--output-dir=<Unicode> (FilesWriter.build_directory)
    Default: ''
    Directory to write output(s) to. Defaults to output to the directory of each
    notebook. To recover previous default behaviour (outputting to the current
    working directory) use . as the flag value.
--writer=<DottedObjectName> (NbConvertApp.writer_class)
    Default: 'FilesWriter'
    Writer class used to write the  results of the conversion
--log-level=<Enum> (Application.log_level)
    Default: 30
    Choices: (0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR',
'CRITICAL')
    Set the log level by value or name.
--reveal-prefix=<Unicode> (SlidesExporter.reveal_url_prefix)
    Default: u''
    The URL prefix for reveal.js (version 3.x). This defaults to the reveal CDN,
    but can be any url pointing to a copy  of reveal.js.
    For speaker notes to work, this must be a relative path to a local  copy of
    reveal.js: e.g., "reveal.js".
    If a relative path is given, it must be a subdirectory of the current
    directory (from which the server is run).
    See the usage documentation
    (https://nbconvert.readthedocs.io/en/latest/usage.html#reveal-js-html-
    slideshow) for more details.
--to=<Unicode> (NbConvertApp.export_format)
    Default: 'html'
    The export format to be used, either one of the built-in formats
    ['asciidoc', 'custom', 'html', 'latex', 'markdown', 'notebook', 'pdf',
    'python', 'rst', 'script', 'slides'] or a dotted object name that represents
    the import path for an `Exporter` class
--template=<Unicode> (TemplateExporter.template_file)
    Default: u''
    Name of the template file to use
--output=<Unicode> (NbConvertApp.output_base)
    Default: ''
    overwrite base name use for output files. can only be used when converting
    one notebook at a time.
--post=<DottedOrNone> (NbConvertApp.postprocessor_class)
    Default: u''
    PostProcessor class used to write the results of the conversion
--config=<Unicode> (JupyterApp.config_file)
    Default: u''
    Full path of a config file.
```

```
To see all available configurables, use `--help-all`

Examples
--------

    The simplest way to use nbconvert is

    > jupyter nbconvert mynotebook.ipynb

    which will convert mynotebook.ipynb to the default format (probably HTML).

    You can specify the export format with `--to`.
    Options include ['asciidoc', 'custom', 'html', 'latex', 'markdown',
'notebook', 'pdf', 'python', 'rst', 'script', 'slides'].

    > jupyter nbconvert --to latex mynotebook.ipynb

    Both HTML and LaTeX support multiple output templates. LaTeX includes
    'base', 'article' and 'report'.  HTML includes 'basic' and 'full'. You
    can specify the flavor of the format used.

    > jupyter nbconvert --to html --template basic mynotebook.ipynb

    You can also pipe the output to stdout, rather than a file

    > jupyter nbconvert mynotebook.ipynb --stdout

    PDF is generated via latex

    > jupyter nbconvert mynotebook.ipynb --to pdf

    You can get (and serve) a Reveal.js-powered slideshow

    > jupyter nbconvert myslides.ipynb --to slides --post serve

    Multiple notebooks can be given at the command line in a couple of
    different ways:

    > jupyter nbconvert notebook*.ipynb
    > jupyter nbconvert notebook1.ipynb notebook2.ipynb

    or you can specify the notebooks list in a config file, containing::

        c.NbConvertApp.notebooks = ["my_notebook.ipynb"]

    > jupyter nbconvert --config mycfg.py
```

[49]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```