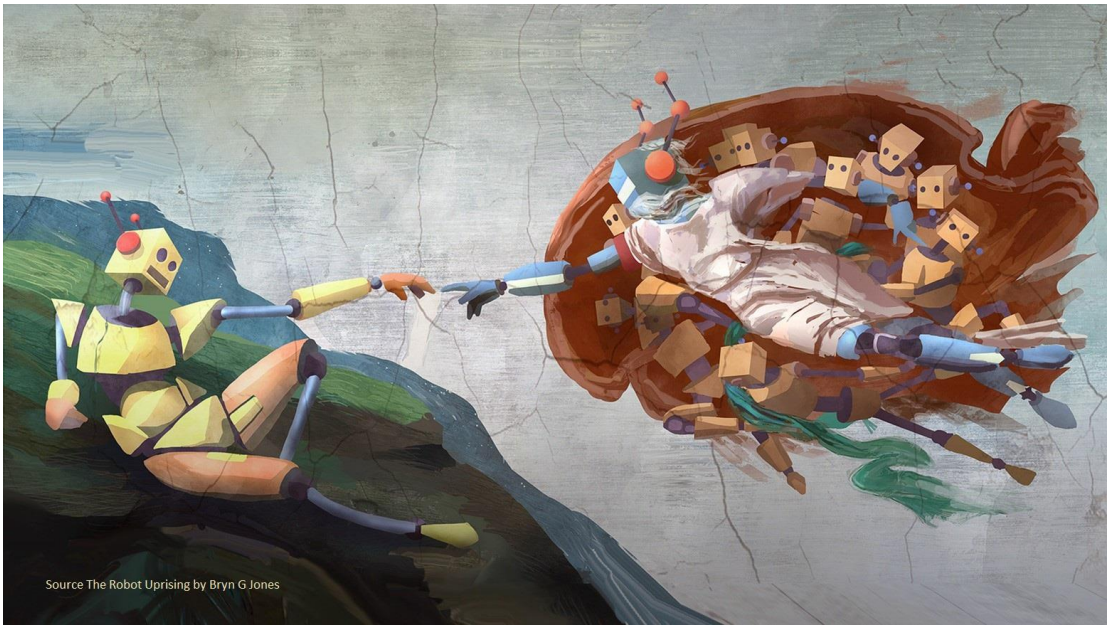
# PRINCIPAL COMPONENT ANALYSIS (PCA)

## USING PYSPARK

DR. ALVIN ANG



# CONTENTS

# I. PCA WITH PYSPARK

Refer here: https://www.alvinang.sg/s/Principal-Component-Analysis-PCA-with-Python-by-Dr-Alvin-Ang.pdf

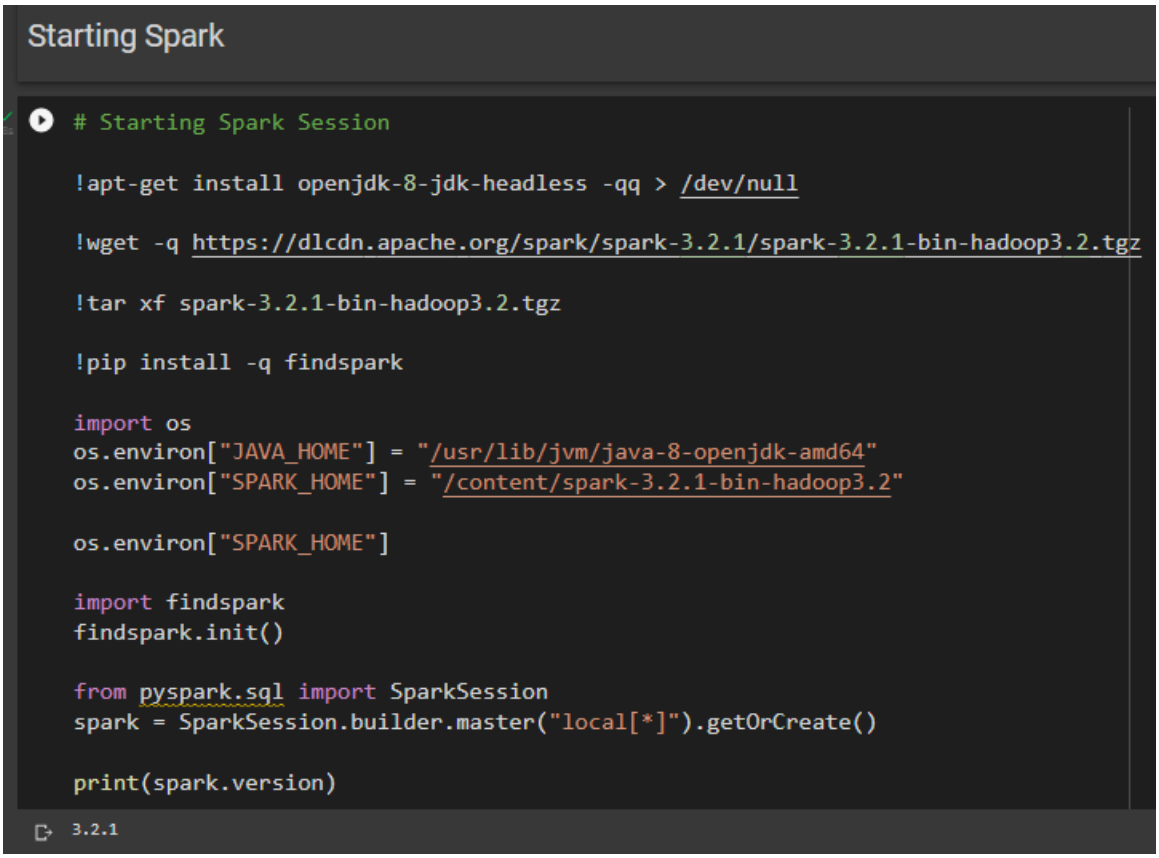
Most of the code is taken from: https://www.amazon.com/Learn-PySpark-Python-based-Machine-Learning/dp/1484249607

Dataset here: https://www.alvinang.sg/s/transformations.csv

Ipynb file here: https://www.alvinang.sg/s/PCA_with_PySpark.ipynb

## A. START A SPARK SESSION

Refer here: https://www.alvinang.sg/s/How_To_Start_A_Spark_Session.ipynb



## B. IMPORTING THE CSV

Import the CSV

```
from pyspark import SparkFiles

url = 'https://www.alvinang.sg/s/transformations.csv'
spark.sparkContext.addFile(url)

df = spark.read.csv(SparkFiles.get("transformations.csv"), \
                    header=True, inferSchema=True)
df
```

```
DataFrame[Col 1: int, Col 2: int, Col 3: double, Col 4: int, Col 5: int, Col 6: int, Col 7: int, Col 8: int, label: double]
```

```
[3] df.count()

20
```

The dataframe has 20 rows.

```
df.show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+---------+
|Col 1|Col 2|Col 3|Col 4|Col 5|Col 6|Col 7|Col 8|    label|
+-----+-----+-----+-----+-----+-----+-----+-----+---------+
|    3|   32|  9.0|    3|    3|   17|    2|    5|0.1111111|
|    3|   27| 13.0|    3|    1|   14|    3|    4|3.2307692|
|    4|   22|  2.5|    0|    1|   16|    3|    5|1.3999996|
|    4|   37| 16.5|    4|    3|   16|    5|    5|0.7272727|
|    5|   27|  9.0|    1|    1|   14|    3|    4| 4.666666|
|    4|   27|  9.0|    0|    2|   14|    3|    4| 4.666666|
|    5|   37| 23.0|    6|    2|   12|    5|    4|0.8521735|
|    5|   37| 23.0|    6|    2|   12|    2|    3| 1.826086|
|    3|   22|  2.5|    0|    2|   12|    3|    3|4.7999992|
|    3|   27|  6.0|    0|    1|   16|    3|    5| 1.333333|
|    2|   27|  6.0|    2|    1|   16|    3|    5|3.2666645|
|    5|   27|  6.0|    2|    3|   14|    3|    5| 2.041666|
|    3|   37| 16.5|    6|    1|   12|    2|    3|0.4848484|
|    5|   27|  6.0|    0|    2|   14|    3|    2|      2.0|
|    4|   22|  6.0|    1|    1|   14|    4|    4|3.2666645|
|    4|   37|  9.0|    2|    2|   14|    3|    6|1.3611107|
|    4|   27|  6.0|    1|    1|   12|    3|    5|      2.0|
|    1|   37| 23.0|    6|    4|   14|    5|    2| 1.826086|
|    2|   42| 23.0|    2|    2|   20|    4|    4| 1.826086|
|    4|   37|  6.0|    0|    2|   16|    5|    4| 2.041666|
+-----+-----+-----+-----+-----+-----+-----+-----+---------+
```

## C. TRANSFORM THE DATASET (CREATE FEATURES COLUMN)

Transform the Dataset using VectorAssembler

```
[6] from pyspark.ml.feature import VectorAssembler

    assembler = VectorAssembler(inputCols=[col for col in df.columns \
                                           if col !='label'], outputCol="features")
    df_new=assembler.transform(df)
```

```
df_new.show()
```

| Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 | Col 7 | Col 8 | label | features |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 32 | 9.0 | 3 | 3 | 17 | 2 | 5 | 0.1111111 | [3.0,32.0,9.0,3.0...] |
| 3 | 27 | 13.0 | 3 | 1 | 14 | 3 | 4 | 3.2307692 | [3.0,27.0,13.0,3....] |
| 4 | 22 | 2.5 | 0 | 1 | 16 | 3 | 5 | 1.3999996 | [4.0,22.0,2.5,0.0...] |
| 4 | 37 | 16.5 | 4 | 3 | 16 | 5 | 5 | 0.7272727 | [4.0,37.0,16.5,4....] |
| 5 | 27 | 9.0 | 1 | 1 | 14 | 3 | 4 | 4.666666 | [5.0,27.0,9.0,1.0...] |
| 4 | 27 | 9.0 | 0 | 2 | 14 | 3 | 4 | 4.666666 | [4.0,27.0,9.0,0.0...] |
| 5 | 37 | 23.0 | 6 | 2 | 12 | 5 | 4 | 0.8521735 | [5.0,37.0,23.0,6....] |
| 5 | 37 | 23.0 | 6 | 2 | 12 | 2 | 3 | 1.826086 | [5.0,37.0,23.0,6....] |
| 3 | 22 | 2.5 | 0 | 2 | 12 | 3 | 3 | 4.7999992 | [3.0,22.0,2.5,0.0...] |
| 3 | 27 | 6.0 | 0 | 1 | 16 | 3 | 5 | 1.333333 | [3.0,27.0,6.0,0.0...] |
| 2 | 27 | 6.0 | 2 | 1 | 16 | 3 | 5 | 3.2666645 | [2.0,27.0,6.0,2.0...] |
| 5 | 27 | 6.0 | 2 | 3 | 14 | 3 | 5 | 2.041666 | [5.0,27.0,6.0,2.0...] |
| 3 | 37 | 16.5 | 6 | 1 | 12 | 2 | 3 | 0.4848484 | [3.0,37.0,16.5,6....] |
| 5 | 27 | 6.0 | 0 | 2 | 14 | 3 | 2 | 2.0 | [5.0,27.0,6.0,0.0...] |
| 4 | 22 | 6.0 | 1 | 1 | 14 | 4 | 4 | 3.2666645 | [4.0,22.0,6.0,1.0...] |
| 4 | 37 | 9.0 | 2 | 2 | 14 | 3 | 6 | 1.3611107 | [4.0,37.0,9.0,2.0...] |
| 4 | 27 | 6.0 | 1 | 1 | 12 | 3 | 5 | 2.0 | [4.0,27.0,6.0,1.0...] |
| 1 | 37 | 23.0 | 6 | 4 | 14 | 5 | 2 | 1.826086 | [1.0,37.0,23.0,6....] |
| 2 | 42 | 23.0 | 2 | 2 | 20 | 4 | 4 | 1.826086 | [2.0,42.0,23.0,2....] |
| 4 | 37 | 6.0 | 0 | 2 | 16 | 5 | 4 | 2.041666 | [4.0,37.0,6.0,0.0...] |

you took columns 1 to 8 and created a new column called 'features' and put it in

## D. PCA

```
PCA

[10] from pyspark.ml.feature import PCA

     pca = PCA(k=2, inputCol="features", outputCol="pca_features")
     pca_model=pca.fit(df_new)
     pca_model

     PCAModel: uid=PCA_c4a8356357d2, k=2
```

- We import PCA, then fit the "df_new" into PCA

```
pca_comp = pca_model.transform(df_new).select("pca_features")
pca_comp.show(truncate=False)
```

```
+-----------------------------------------+
|pca_features                             |
+-----------------------------------------+
|[27.65346473555165,-24.198755612681367] |
|[27.45461471499118,-17.263948562463476] |
|[15.824784697285162,-20.909113408315598]|
|[36.677119951007285,-23.21276720859636] |
|[24.007458610134332,-19.804794167612606]|
|[23.875630131196115,-20.111114858821267]|
|[41.84916285084273,-17.568727131347273] |
|[41.775463011554464,-17.294959292241384]|
|[15.885864619415862,-19.18017614977016] |
|[21.609730723090987,-22.59830342342499] |
|[22.030221747574853,-22.14831313216692] |
|[22.005821687966737,-21.37005458883427] |
|[36.90487201745985,-20.77397738718488]  |
|[21.661524932598155,-21.377784686846688]|
|[18.681908524349062,-17.956106467352154]|
|[30.47653176501122,-26.952352443132913] |
|[21.715975562523198,-20.82483328827104] |
|[42.12619884551877,-18.220017596993195] |
|[44.364043272864755,-25.101538406970164]|
|[27.99452158837228,-29.567600829764075] |
+-----------------------------------------+
```

we used the previous df_new['features'] column (which comprises of 8 columns of data)

and PCA it into 2 columns

in other words, we compressed 8 columns into 2 columns

The 2 columns are known as Principal Component 1 and Principal Component 2

that's why there are 2 columns in pca_features

## ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.