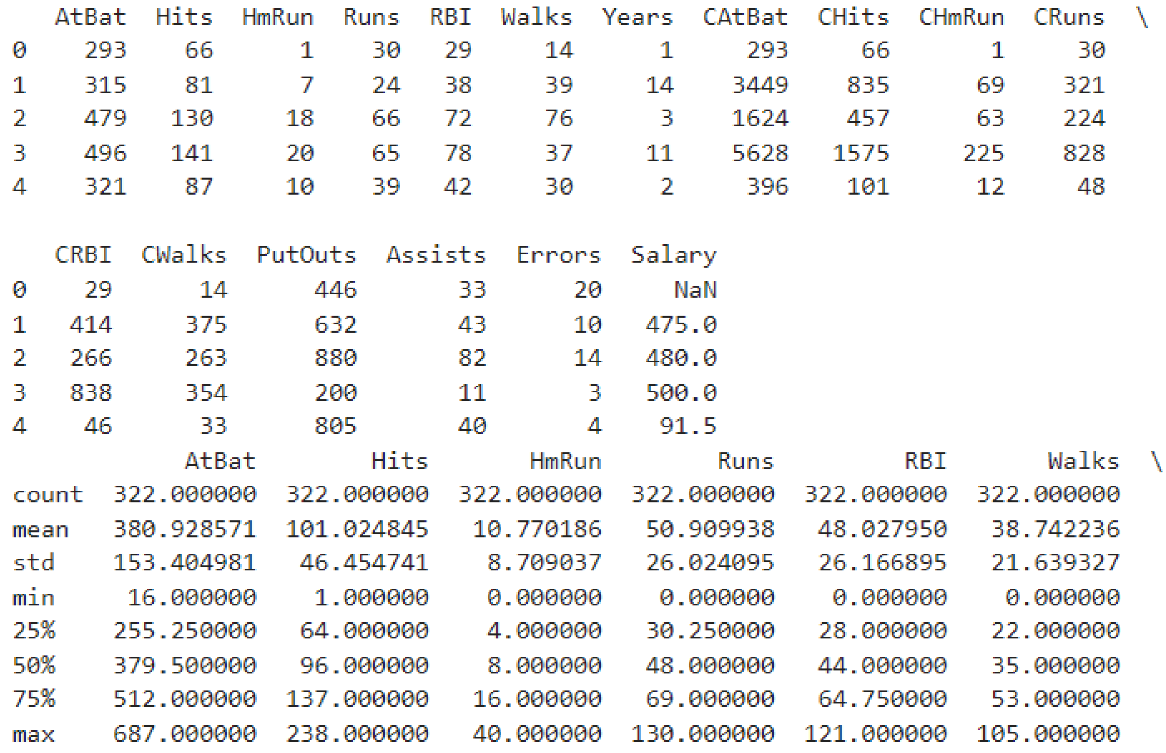
Documentation 1B:

1. Introduction:

In this analysis, we explore the application of Principal Component Analysis (PCA) and linear regression to predict salary values in a baseball dataset. The code follows a structured approach, including data loading, preprocessing, exploratory data analysis, model implementation, and results analysis.

Here is the following data description:



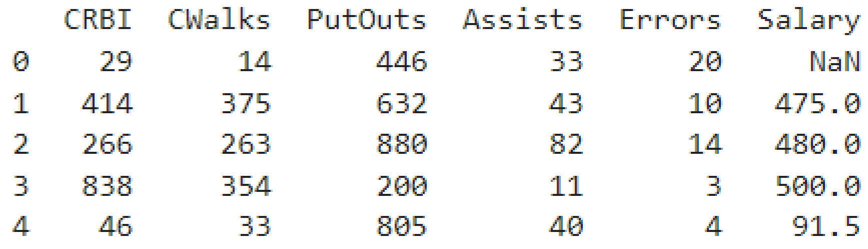
```
   AtBat  Hits  HmRun  Runs  RBI  Walks  Years  CAtBat  CHits  CHmRun  CRuns  \
0    293    66      1    30   29     14      1     293     66       1     30
1    315    81      7    24   38     39     14    3449    835      69    321
2    479   130     18    66   72     76      3    1624    457      63    224
3    496   141     20    65   78     37     11    5628   1575     225    828
4    321    87     10    39   42     30      2     396    101      12     48

   CRBI  CWalks  PutOuts  Assists  Errors  Salary
0    29      14      446       33      20     NaN
1   414     375      632       43      10   475.0
2   266     263      880       82      14   480.0
3   838     354      200       11       3   500.0
4    46      33      805       40       4    91.5
            AtBat        Hits       HmRun        Runs         RBI       Walks  \
count  322.000000  322.000000  322.000000  322.000000  322.000000  322.000000
mean   380.928571  101.024845   10.770186   50.909938   48.027950   38.742236
std    153.404981   46.454741    8.709037   26.024095   26.166895   21.639327
min     16.000000    1.000000    0.000000    0.000000    0.000000    0.000000
25%    255.250000   64.000000    4.000000   30.250000   28.000000   22.000000
50%    379.500000   96.000000    8.000000   48.000000   44.000000   35.000000
75%    512.000000  137.000000   16.000000   69.000000   64.750000   53.000000
max    687.000000  238.000000   40.000000  130.000000  121.000000  105.000000
```

2. Data Loading and Preprocessing

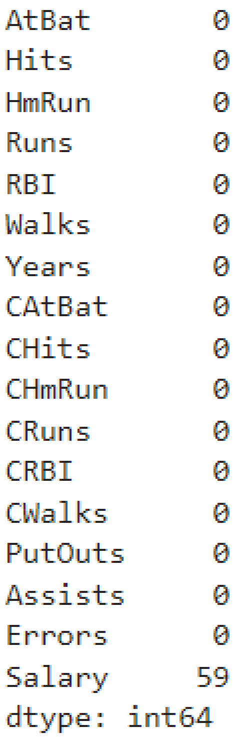
The dataset is loaded from an Excel file (Hitters.xlsx), and categorical columns are dropped. To handle missing values, entries with NaN in the 'Salary' feature are removed.

Here in the below we can see that Salary has NAN:



```
   CRBI  CWalks  PutOuts  Assists  Errors  Salary
0    29      14      446       33      20     NaN
1   414     375      632       43      10   475.0
2   266     263      880       82      14   480.0
3   838     354      200       11       3   500.0
4    46      33      805       40       4    91.5
```

The number of NAN are depicted below :

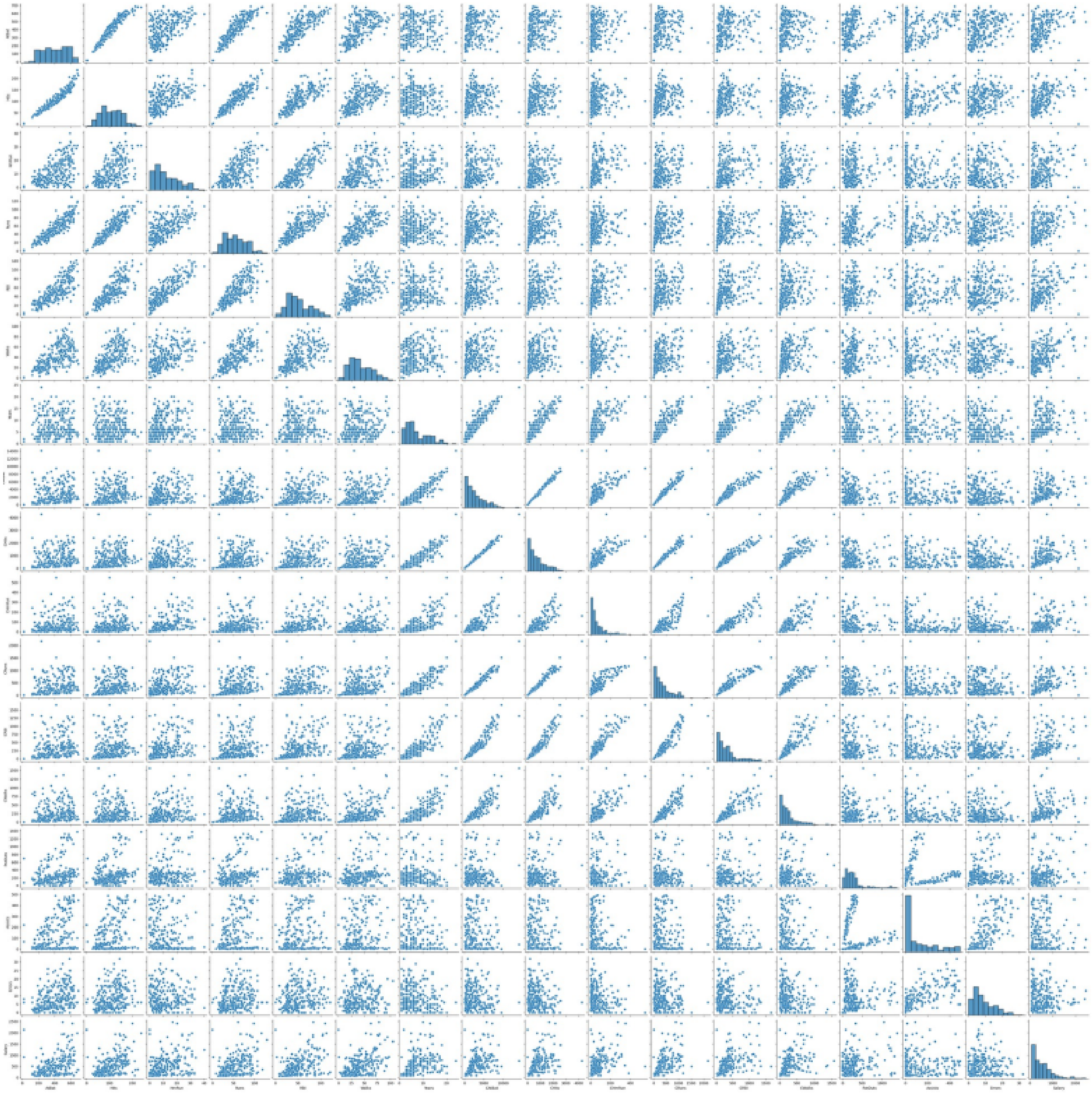


```
AtBat       0
Hits        0
HmRun       0
Runs        0
RBI         0
Walks       0
Years       0
CAtBat      0
CHits       0
CHmRun      0
CRuns       0
CRBI        0
CWalks      0
PutOuts     0
Assists     0
Errors      0
Salary     59
dtype: int64
```

3. Exploratory Data Analysis (EDA)

Basic statistics and a pair plot are generated to gain insights into the dataset's structure and relationships between variables.

Here is the sns pair plot :

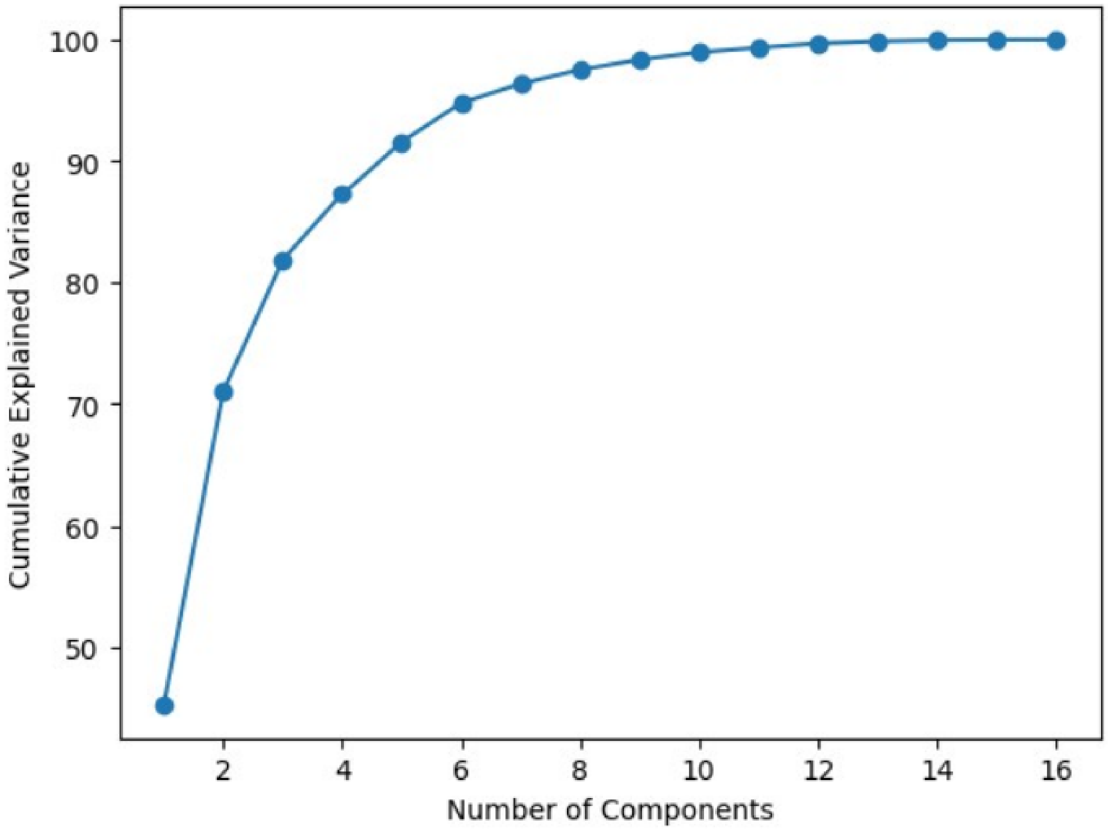


4. Principal Component Analysis (PCA)

- Features are standardized to ensure uniformity.

- The covariance matrix, eigenvalues, and eigenvectors are calculated.
- Eigenvalues and eigenvectors are sorted to determine the principal components.
- Explained variance ratio and cumulative explained variance are computed.
- The number of components necessary for efficient prediction is identified through cumulative explained variance.

Here is the plot to determine the establishment of number of components vs Cumulative Variance

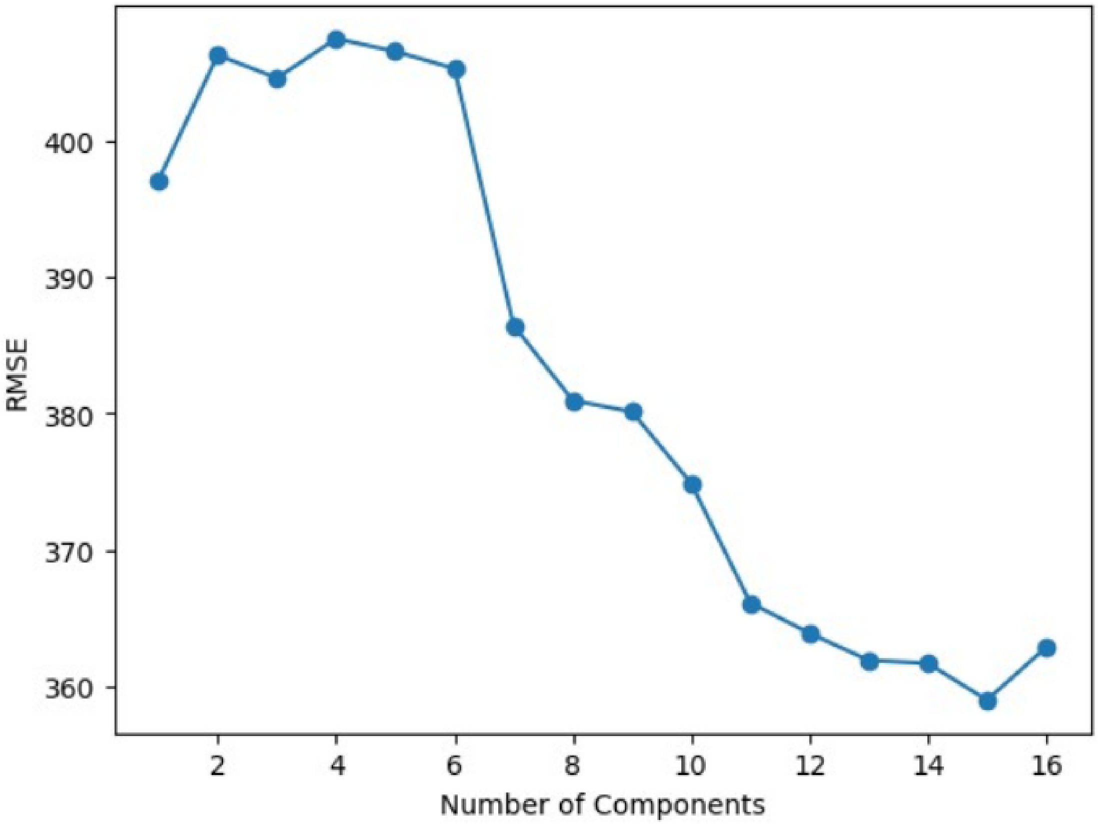


5. Number of Components vs RMSE

A loop iterates over different numbers of principal components, and for each iteration:

- Training and test data are transformed using PCA.
- Linear regression is applied.

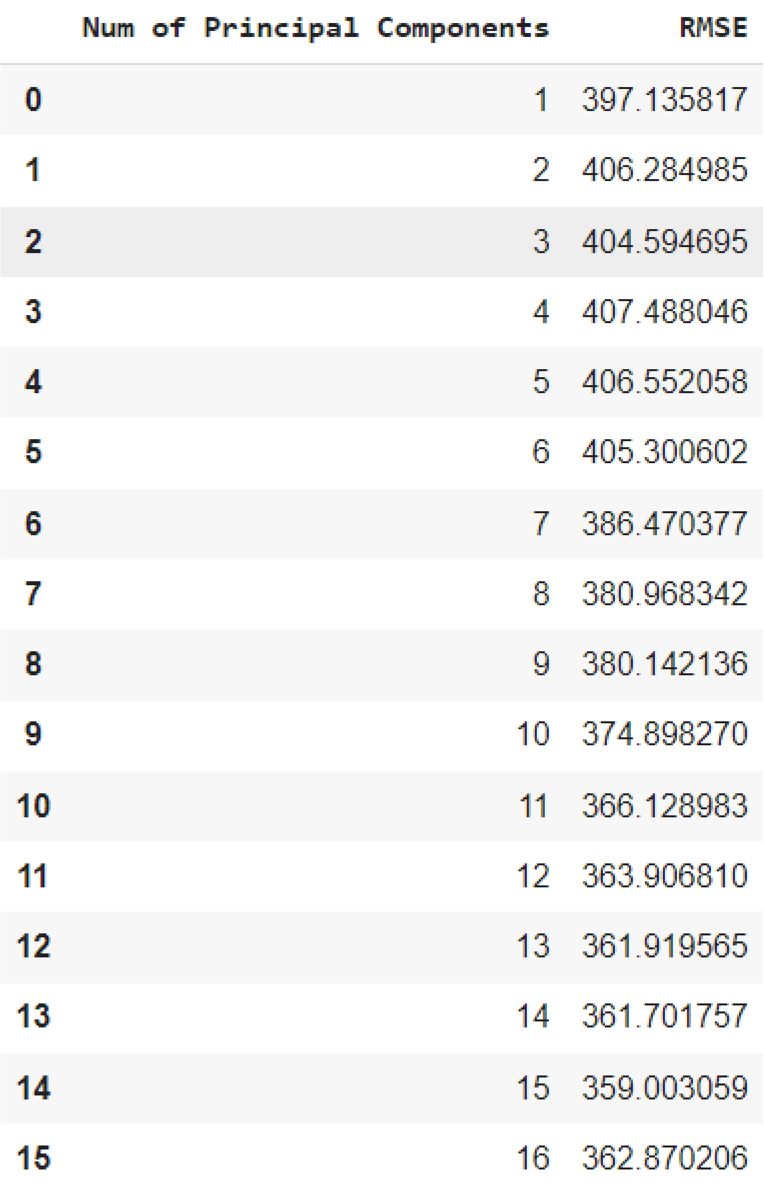
- Root Mean Squared Error (RMSE) is calculated for each iteration.



6. Results Analysis

- The cumulative explained variance and RMSE vs number of components are presented graphically.

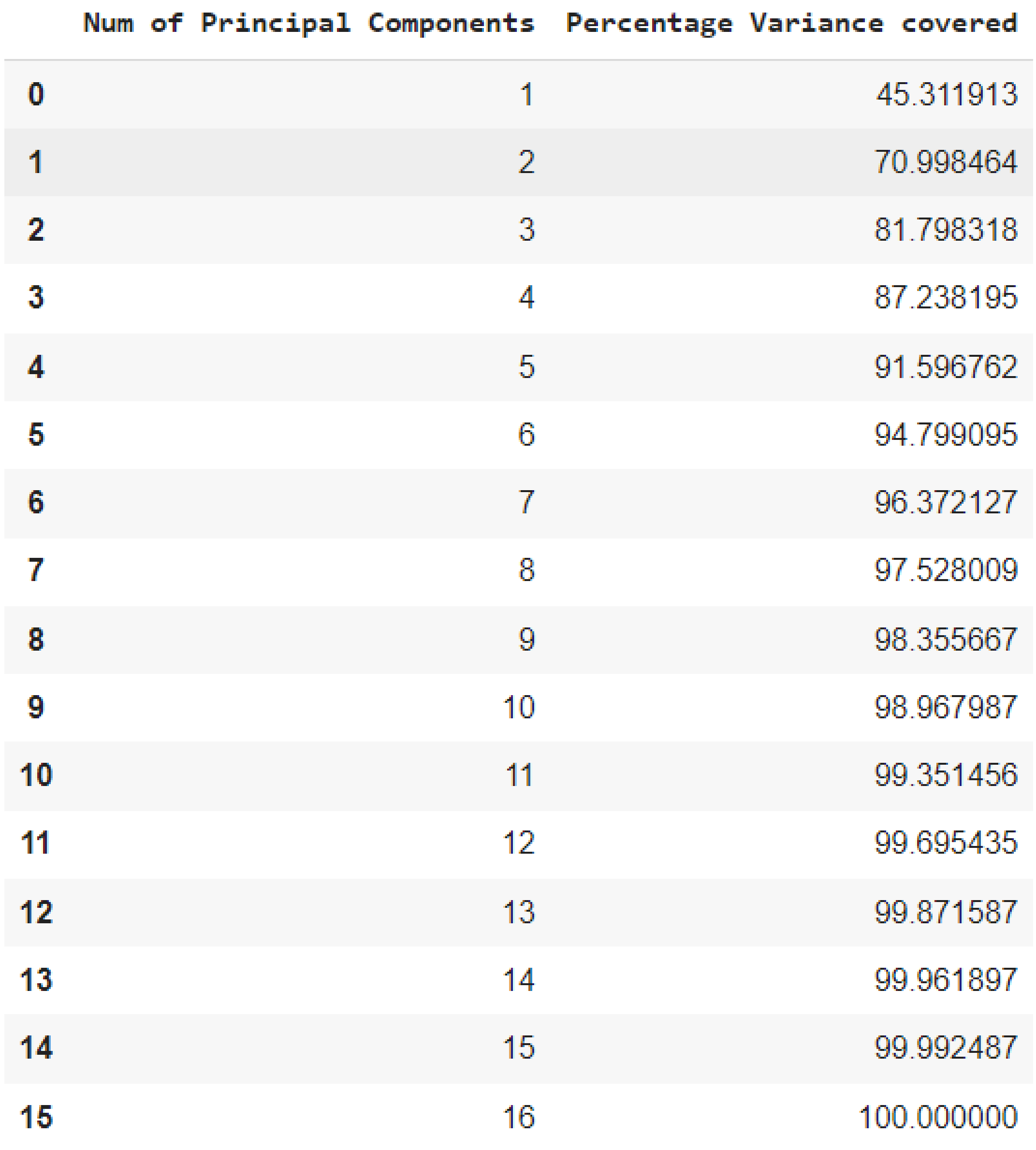
The optimal number of components is selected based on minimal RMSE.



	Num of Principal Components	RMSE
0	1	397.135817
1	2	406.284985
2	3	404.594695
3	4	407.488046
4	5	406.552058
5	6	405.300602
6	7	386.470377
7	8	380.968342
8	9	380.142136
9	10	374.898270
10	11	366.128983
11	12	363.906810
12	13	361.919565
13	14	361.701757
14	15	359.003059
15	16	362.870206

- From the above table we have decided to take the optimal number of components but there might be a question of not taking into account lower RMSE when the number of features are 12 ,13,14.

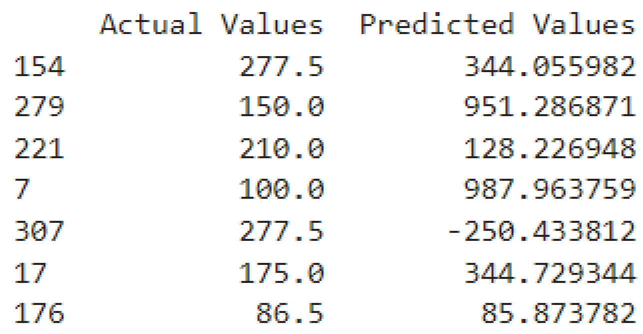
- But the reason for taking the optimal no of components to be taken as 15 is described in the below table:



	Num of Principal Components	Percentage Variance covered
0	1	45.311913
1	2	70.998464
2	3	81.798318
3	4	87.238195
4	5	91.596762
5	6	94.799095
6	7	96.372127
7	8	97.528009
8	9	98.355667
9	10	98.967987
10	11	99.351456
11	12	99.695435
12	13	99.871587
13	14	99.961897
14	15	99.992487
15	16	100.000000

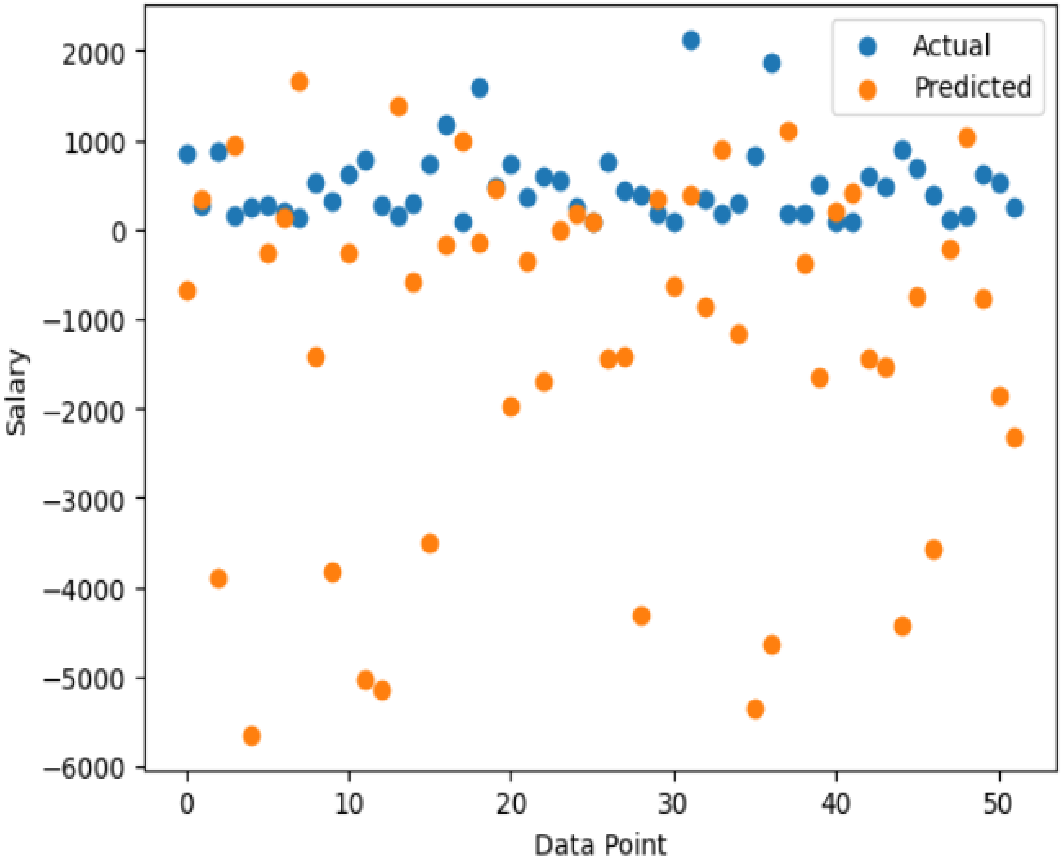
- Since the encapsulated variance is maximum in case of 15 features and RMSE being close enough wrt to the number of features being 12 ,we therefore consider 15 as the optimal number of features.

- Upon mounting Linear Regression model and running it through the testing data here is the tabulation of Actual vs Predicted Values:



```
     Actual Values  Predicted Values
154          277.5        344.055982
279          150.0        951.286871
221          210.0        128.226948
7            100.0        987.963759
307          277.5       -250.433812
17           175.0        344.729344
176           86.5         85.873782
```

- Here is the Corresponding plot to depict the Actual vs Predicted values :


Actual
Predicted
Salary
Data Point