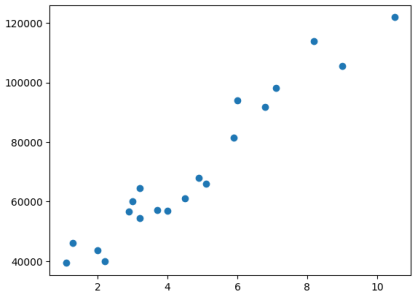
Assignment 4 Documentation

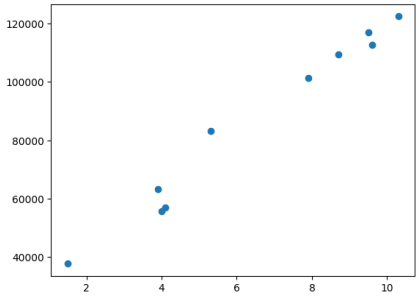
Kavyasri

700728990

1.Salary Dataset

- ✓ Read salary dataset using pandas.
- ✓ Check head to see columns and type of data.
- ✓ Check for null values using salary.isnull().sum()
- ✓ Separate X and Y variables.
- ✓ Split train test data with test=33% using train_test_split from sklearn.
- ✓ Scatter plot for train and test data seperately.
- ✓ Use LinearRegression on train dataset to fit. Then score for both train and test to get R square value.
- ✓ Predict on X_test and calculate mean_squared_error between Y_test and Y_test_pred.



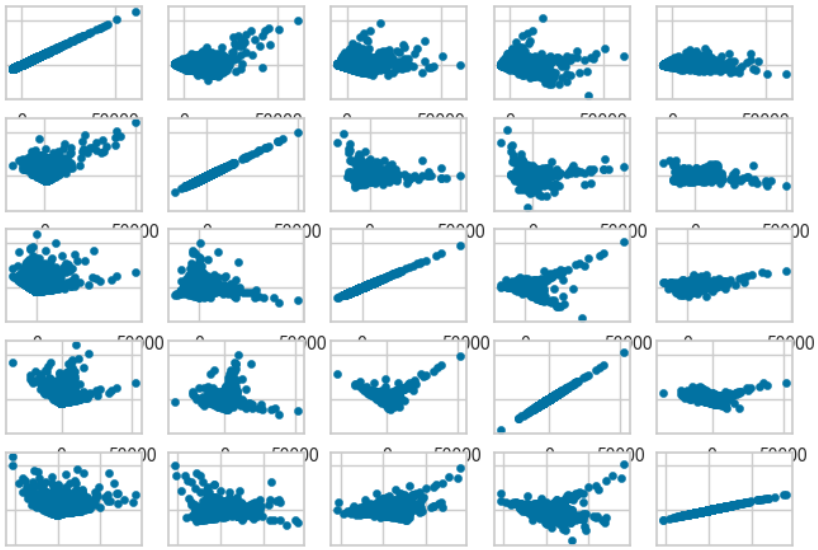


2. K Means Clustering

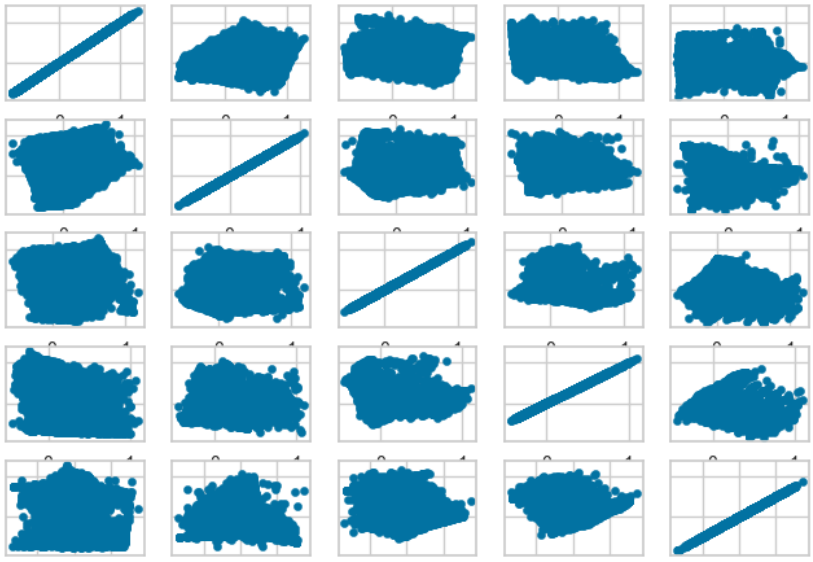
- ✓ Read salary dataset using pandas.
- ✓ Check head to see columns and type of data.
- ✓ Drop CUST_ID column.
- ✓ Check for null values using isnull().sum().
- ✓ Fill missing values with column means using fillna() method.
- ✓ Elbow method:
    - Use KMeans from sklearn fit for 1 to 15 clusters one at a time.
    - Save all sum_squared_distances or inertia_ values.
    - Plot them to find the right k value(the elbow).
- ✓ Use the k value from elbow method and fit, predict KMeans on the data.
- ✓ Calculate silhouette_score. In our case its 0.379 for k=5.
- ✓ Now do MinMaxScaling on data to bring uniformity to various columns.
- ✓ Use the k value from elbow method and fit, predict KMeans on the scaled data.
- ✓ Calculate silhouette_score. In our case its 0.319 for k=5.

Reasons for poor silhouette_score:

- ✓ **Refer to the below scatter plots of Principle components for data before and after scaling. As we can see scaling has made it even more difficult to separate clusters.**
- ✓ **Another reason could be that data is skewed and does not follow normal distribution.**



Before



After

Additional Note:

The code to experiment with PCA and plots, different scaling method is included at the end of notebook