Data Analytics in Software Engineering

Assignment 1

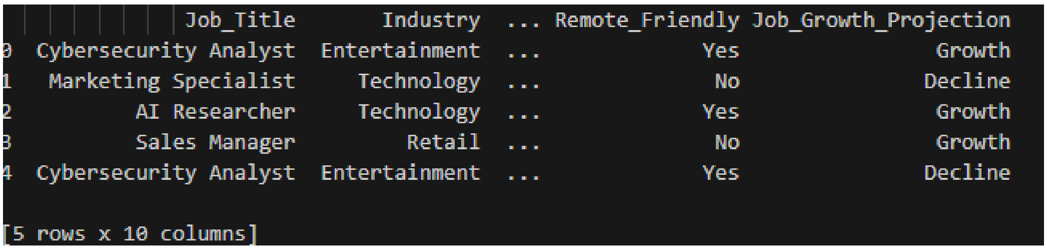
Assignment: Principal Component Analysis (PCA) in Software Engineering

Dataset:

I have selected the **AI Job Market Insights** dataset for PCA assignment. This data is about the job market insights. It has 10 attributes and 500 transactions. Link of Dataset

Step 1: Load and Explore the Data

Load the dataset and view its basic structure to understand the attributes.

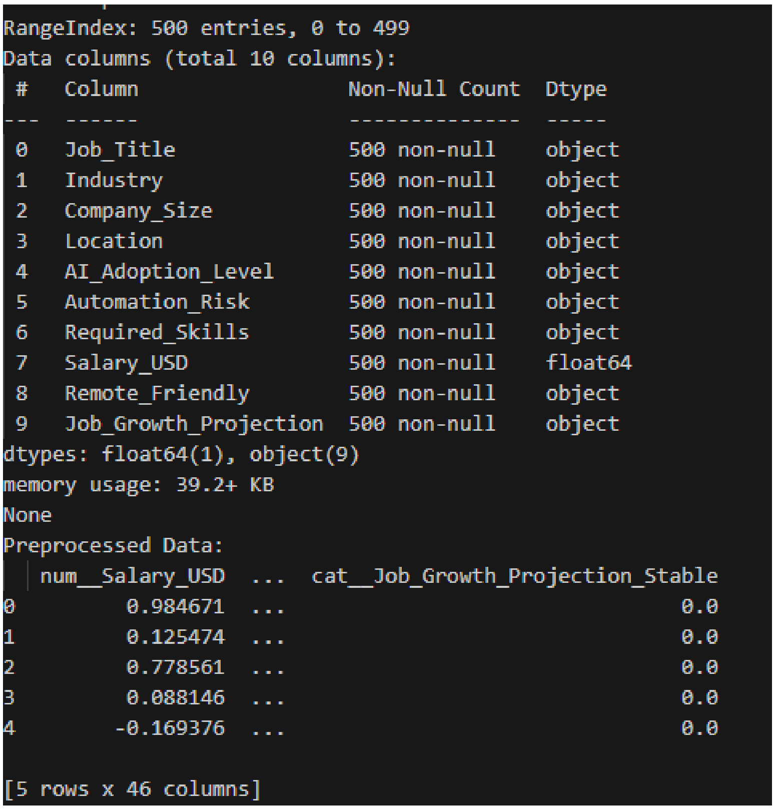


```
              Job_Title       Industry  ... Remote_Friendly Job_Growth_Projection
0  Cybersecurity Analyst  Entertainment  ...             Yes                Growth
1   Marketing Specialist     Technology  ...              No               Decline
2          AI Researcher     Technology  ...             Yes                Growth
3          Sales Manager         Retail  ...              No                Growth
4  Cybersecurity Analyst  Entertainment  ...             Yes               Decline

[5 rows x 10 columns]
```

Step 2: Preprocessing

Encoding Categorical Variables: PCA requires numerical data, so we need to encode categorical variables (e.g., Job_Title, Industry, Company_Size).

Standardization: PCA also requires standardized data (mean = 0, variance = 1), so we'll scale numerical features.



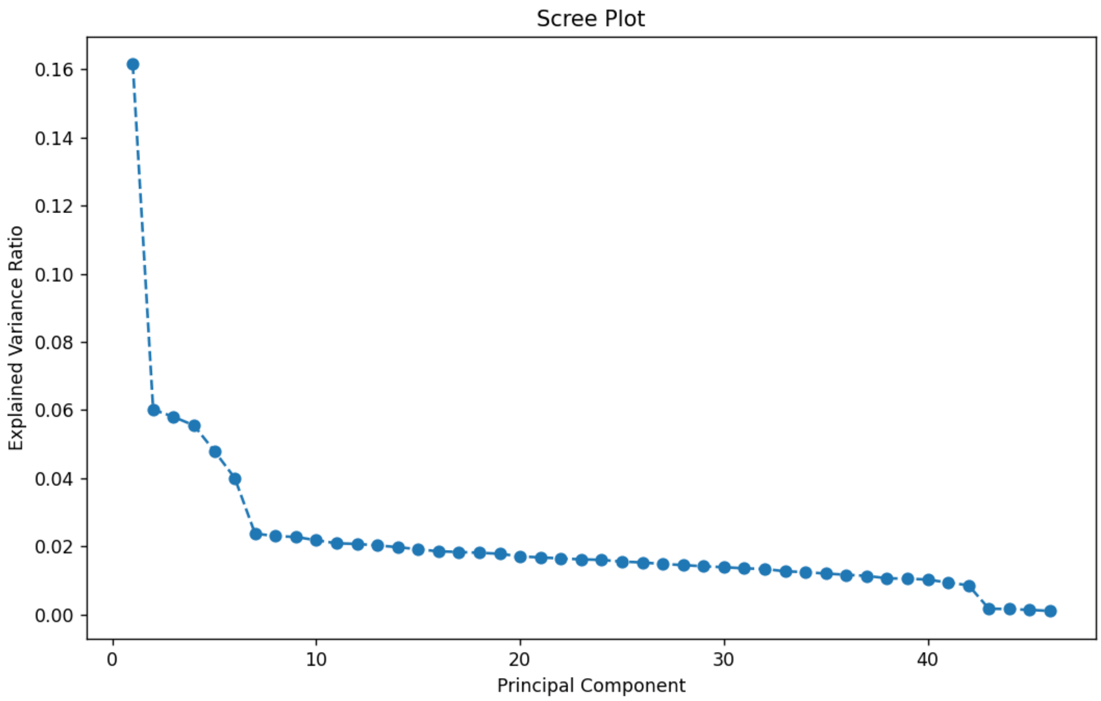
```
RangeIndex: 500 entries, 0 to 499
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Job_Title             500 non-null    object 
 1   Industry              500 non-null    object 
 2   Company_Size          500 non-null    object 
 3   Location              500 non-null    object 
 4   AI_Adoption_Level     500 non-null    object 
 5   Automation_Risk       500 non-null    object 
 6   Required_Skills       500 non-null    object 
 7   Salary_USD            500 non-null    float64
 8   Remote_Friendly       500 non-null    object 
 9   Job_Growth_Projection  500 non-null    object 
dtypes: float64(1), object(9)
memory usage: 39.2+ KB
None
Preprocessed Data:
   num__Salary_USD  ...  cat__Job_Growth_Projection_Stable
0         0.984671  ...                                0.0
1         0.125474  ...                                0.0
2         0.778561  ...                                0.0
3         0.088146  ...                                0.0
4        -0.169376  ...                                0.0

[5 rows x 46 columns]
```

Step 3: Normalize the data (only for numeric columns)

Step 4: Apply PCA and Visualize Explained Variance

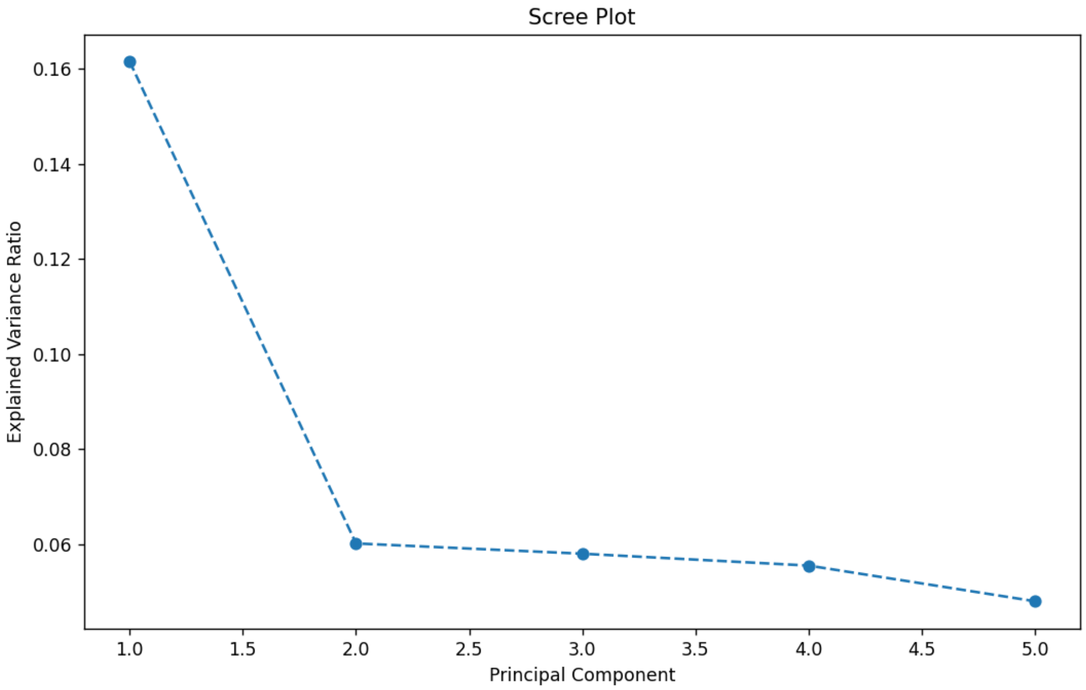
Fit PCA: We'll start by fitting PCA on the preprocessed dataset to see how many components explain most of the variance.

Scree Plot: This plot shows how much variance each component explains, helping to decide the number of components to keep.

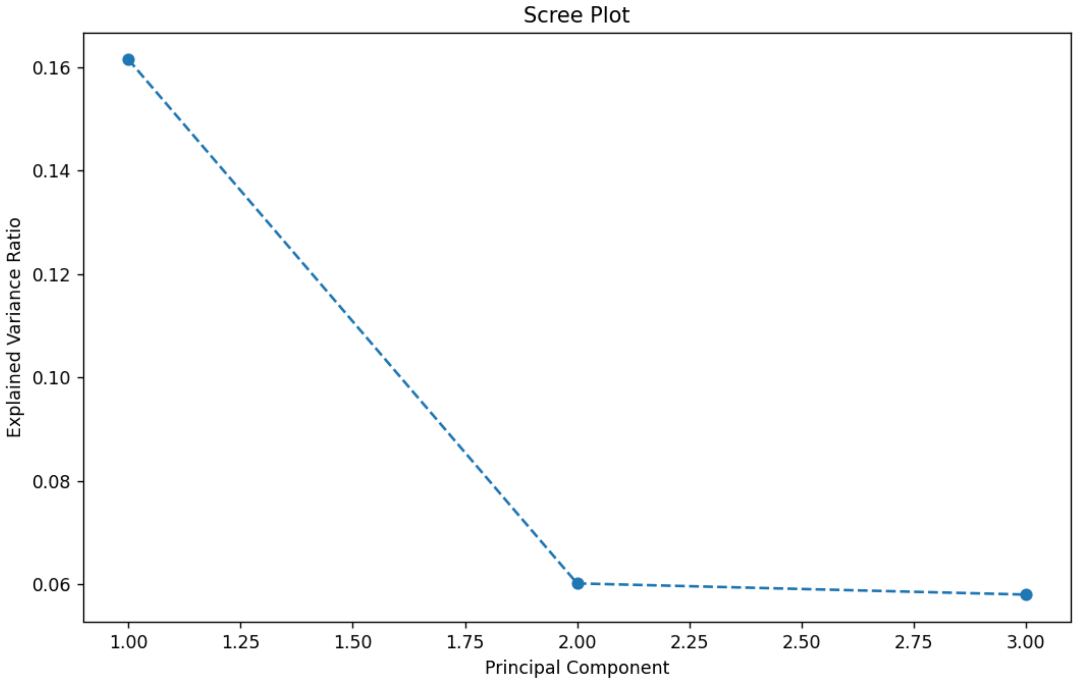


Step 5: Feature Reduction

Based on the scree plot, selected the 5 components.

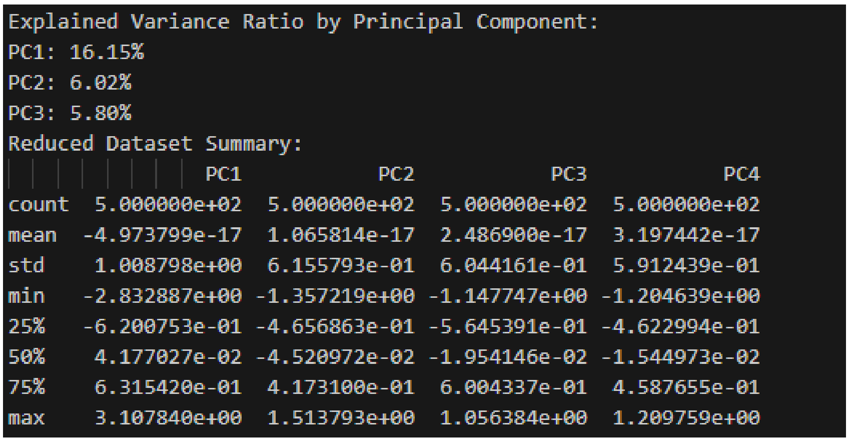


Now, these results are optimal with 3 PCs.



Step 6: Analyze and Interpret Principal Components

To understand which features contribute most to each principal component, examine the loadings (the coefficients of each feature in each component).



```
Explained Variance Ratio by Principal Component:
PC1: 16.15%
PC2: 6.02%
PC3: 5.80%
Reduced Dataset Summary:
                PC1           PC2           PC3           PC4
count  5.000000e+02  5.000000e+02  5.000000e+02  5.000000e+02
mean  -4.973799e-17  1.065814e-17  2.486900e-17  3.197442e-17
std    1.008798e+00  6.155793e-01  6.044161e-01  5.912439e-01
min   -2.832887e+00 -1.357219e+00 -1.147747e+00 -1.204639e+00
25%   -6.200753e-01 -4.656863e-01 -5.645391e-01 -4.622994e-01
50%    4.177027e-02 -4.520972e-02 -1.954146e-02 -1.544973e-02
75%    6.315420e-01  4.173100e-01  6.004337e-01  4.587655e-01
max    3.107840e+00  1.513793e+00  1.056384e+00  1.209759e+00
```