CoGrammar

# Exploratory Data Analysis (EDA)

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

# Data Science Lecture Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(FBV: Mutual Respect.)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Open Classes. You can submit these questions here: **Open Class Questions**

# **Data Science Lecture Housekeeping** cont.

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support

- Report a **safeguarding** incident: www.hyperiondev.com/safeguardreporting

- We would love your **feedback** on lectures: Feedback on Lectures

# Lecture Objectives

- Understand the **purpose and importance of EDA** in the data science workflow.
- Apply **univariate, bivariate, and multivariate analysis techniques** to explore and summarize dataset characteristics.
- Utilize Python libraries such as **pandas, Matplotlib, Seaborn, and Scikit-learn** for data manipulation, visualization, and analysis.

# Lecture Objectives

- Assess **feature importance using statistical tests and machine learning techniques** to guide further analysis and modeling.

# Introduction to EDA

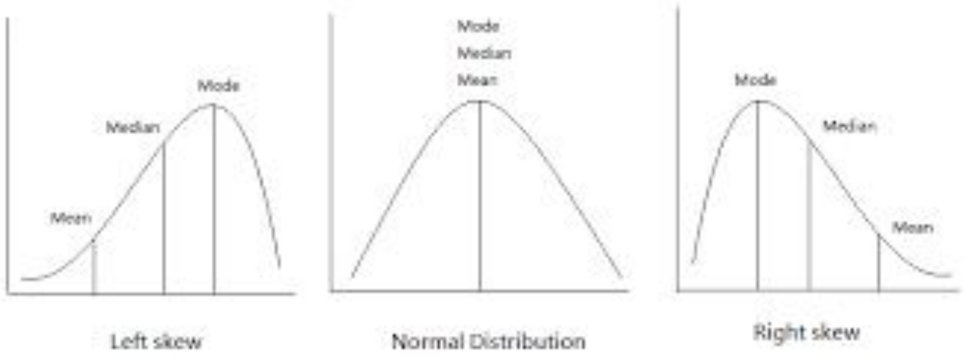
★ **Definition:** Exploratory Data Analysis (EDA) is the process of **investigating and understanding a dataset** through visual and statistical techniques.

★ **Purpose:** EDA helps to **uncover patterns, relationships, and anomalies** in the data, guiding further analysis and modeling.

★ **Importance:** EDA is a crucial step in the data science workflow, enabling **informed decision-making and hypothesis generation**.

# UA - Descriptive Statistics

- ★ **Mean:** The average value of a variable.
- ★ **Median:** The middle value when the data is sorted.
- ★ **Mode:** The most frequent value in the data.

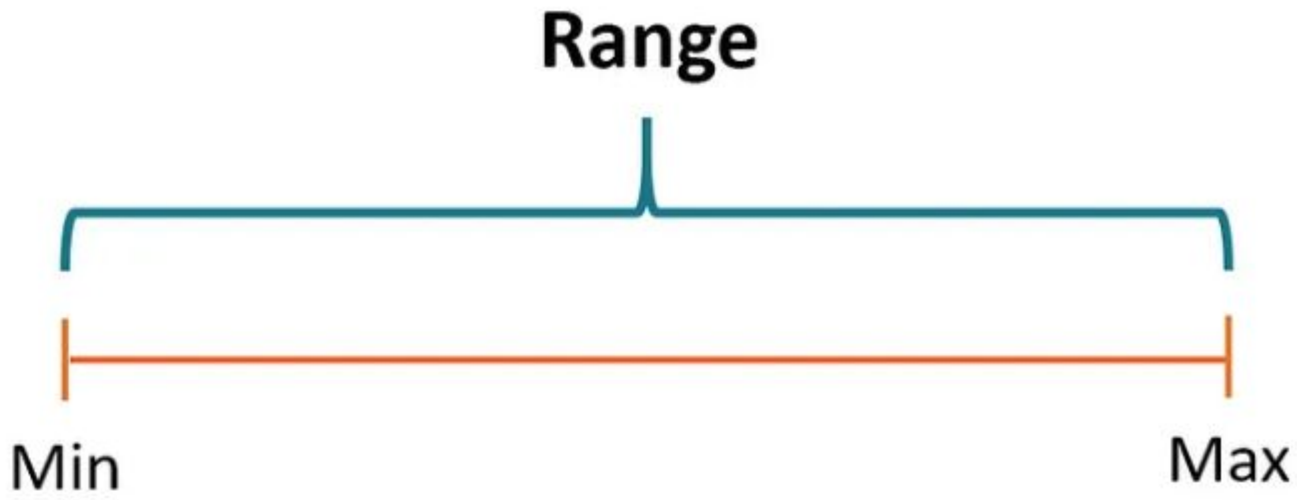


# UA - Descriptive Statistics

- ★ **Range:** The difference between the maximum and minimum values.
- ★ **Variance:** The average squared deviation from the mean.
- ★ **Standard Deviation:** The square root of the variance, indicating the spread of the data.

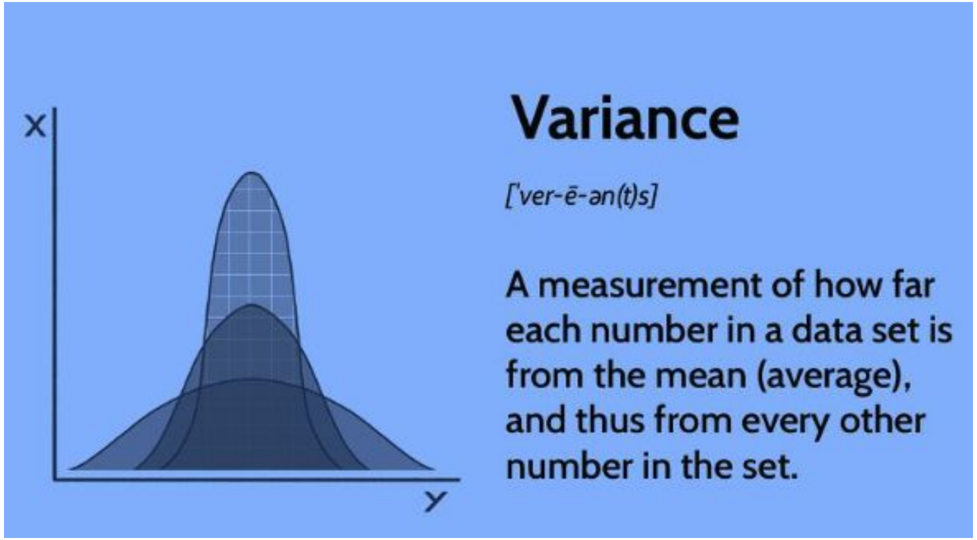
# UA - Descriptive Statistics

★ **Range:** The difference between the maximum and minimum values.



# UA - Descriptive Statistics

★ **Variance:** The average squared deviation from the mean.



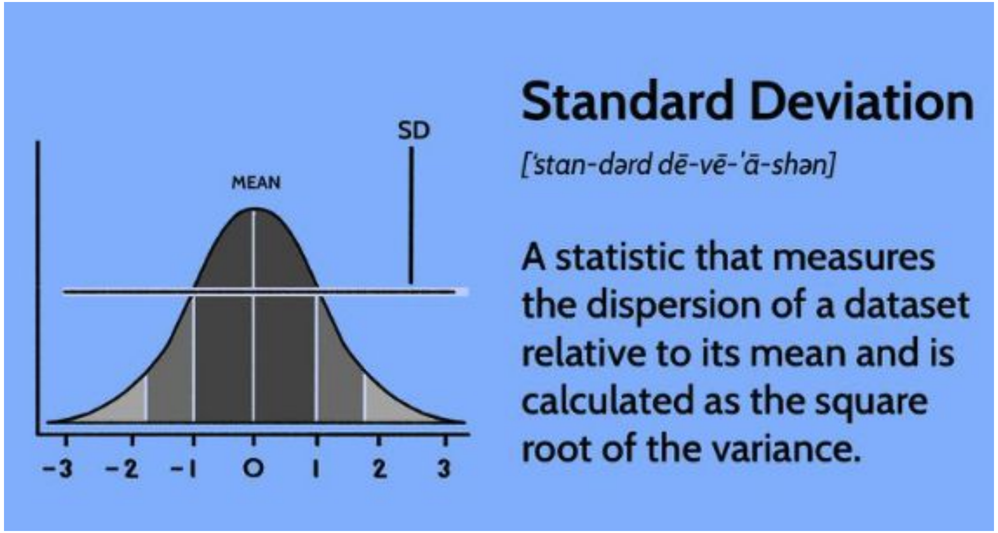
## Variance

*['ver-ē-ən(t)s]*

A measurement of how far each number in a data set is from the mean (average), and thus from every other number in the set.

# UA - Descriptive Statistics

- ★ **Standard Deviation:** The square root of the variance, indicating the spread of the data.



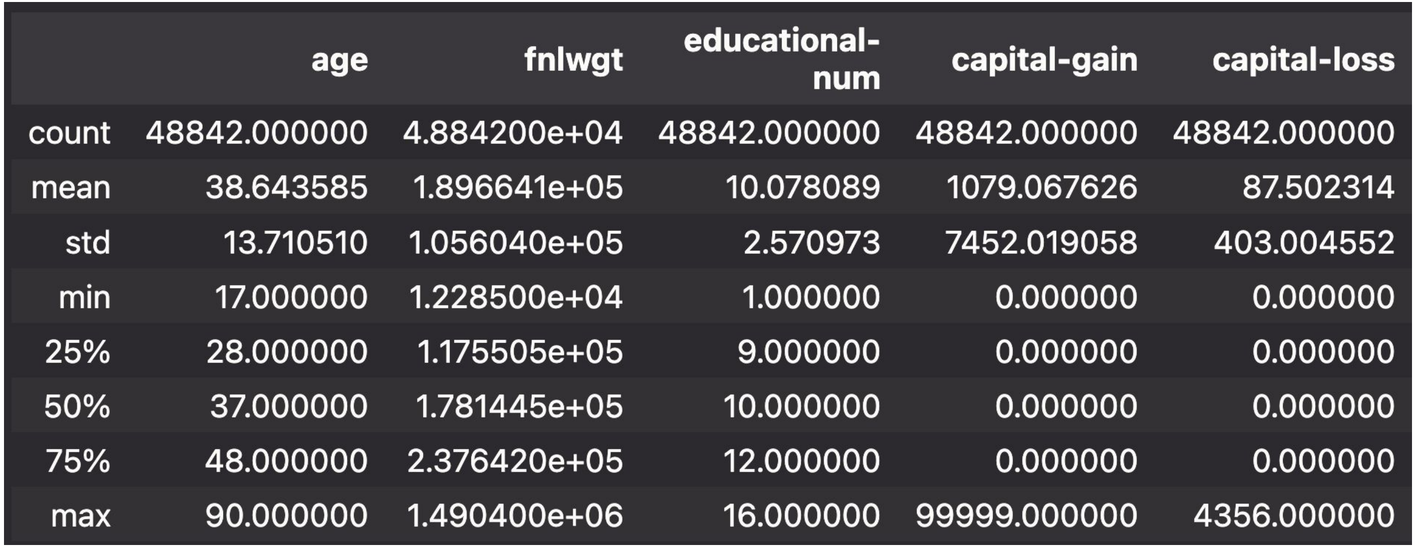
**Standard Deviation**

*['stan-dərd dē-vē-'ā-shən]*

A statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

# UA - Descriptive Statistics

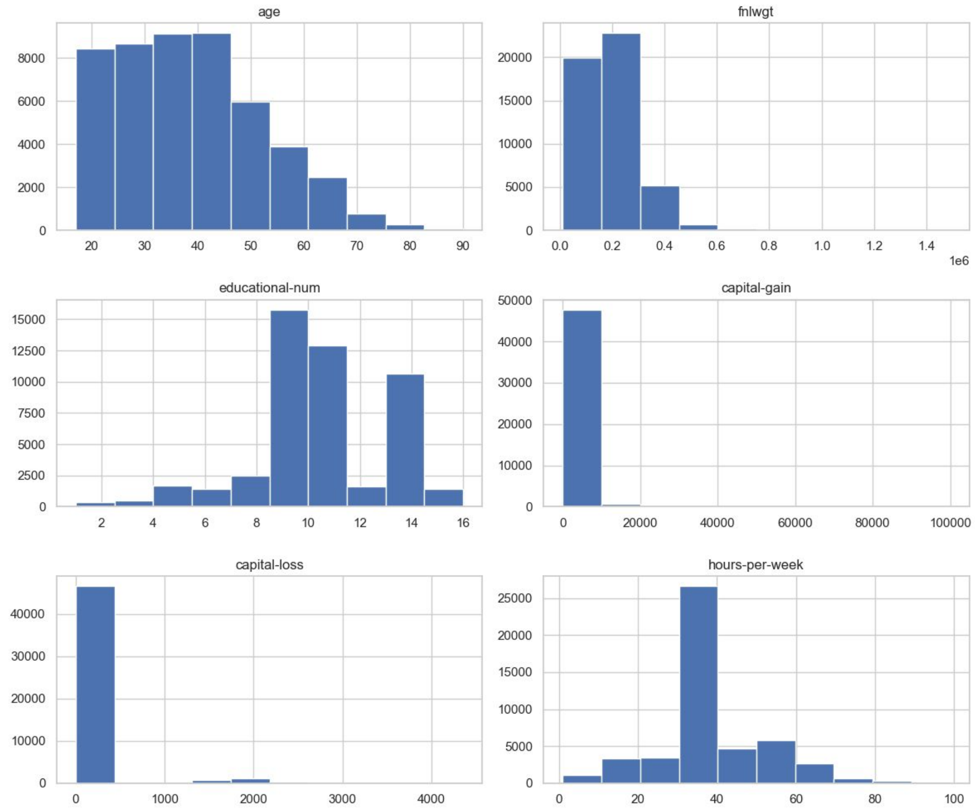
★ Use data.describe() to calculate descriptive statistics for numerical columns.



| | age | fnlwgt | educational-num | capital-gain | capital-loss |
|---|---|---|---|---|---|
| count | 48842.000000 | 4.884200e+04 | 48842.000000 | 48842.000000 | 48842.000000 |
| mean | 38.643585 | 1.896641e+05 | 10.078089 | 1079.067626 | 87.502314 |
| std | 13.710510 | 1.056040e+05 | 2.570973 | 7452.019058 | 403.004552 |
| min | 17.000000 | 1.228500e+04 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 28.000000 | 1.175505e+05 | 9.000000 | 0.000000 | 0.000000 |
| 50% | 37.000000 | 1.781445e+05 | 10.000000 | 0.000000 | 0.000000 |
| 75% | 48.000000 | 2.376420e+05 | 12.000000 | 0.000000 | 0.000000 |
| max | 90.000000 | 1.490400e+06 | 16.000000 | 99999.000000 | 4356.000000 |

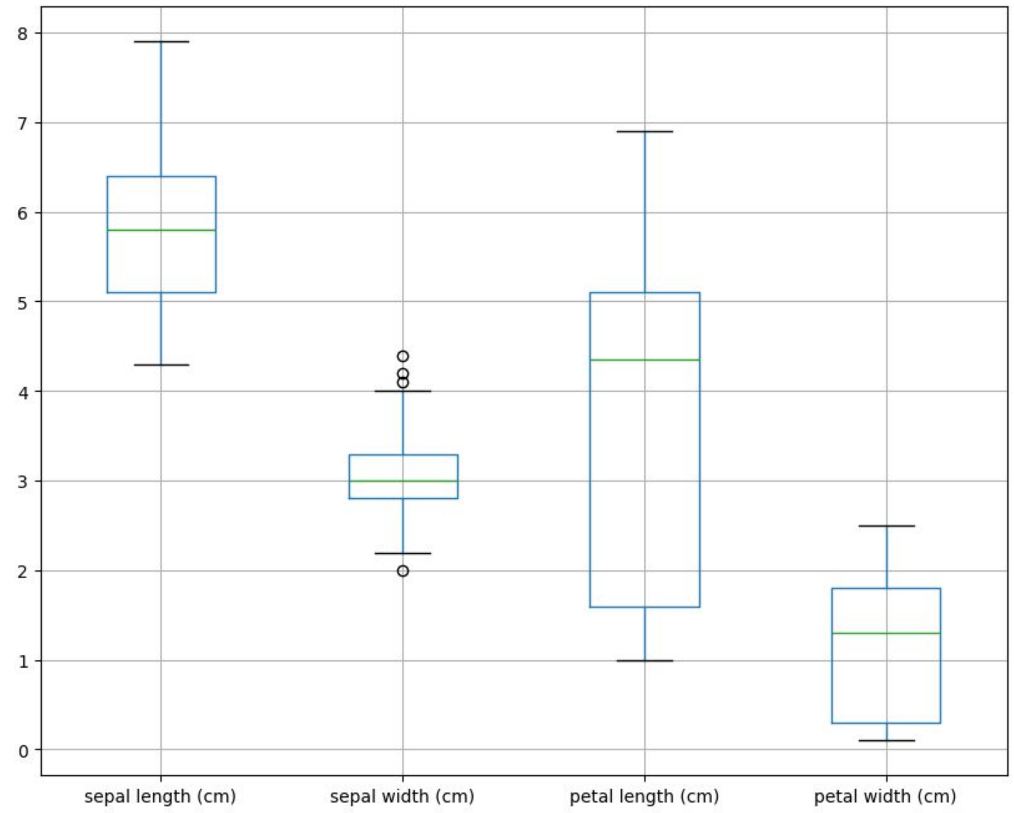
# UA - Visualization Techniques

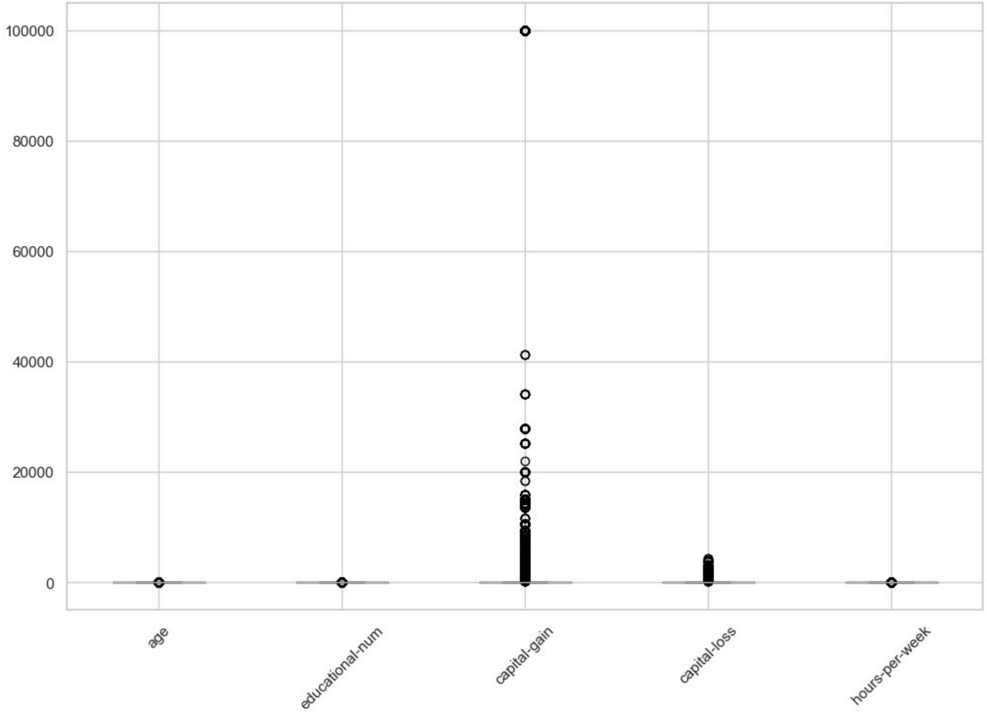
- ★ Histogram
  - ○ Visualize the distribution of a single variable.
  - ○ Use data.hist() to create histograms for numerical columns.


age
8000
6000
4000
2000
0
20
30
40
50
60
70
80
90
fnlwgt
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e6
educational-num
15000
12500
10000
7500
5000
2500
0
2
4
6
8
10
12
14
16
capital-gain
50000
40000
30000
20000
10000
0
0
20000
40000
60000
80000
100000
capital-loss
40000
30000
20000
10000
0
0
1000
2000
3000
4000
hours-per-week
25000
20000
15000
10000
5000
0
0
20
40
60
80
100

# UA - Visualization Techniques

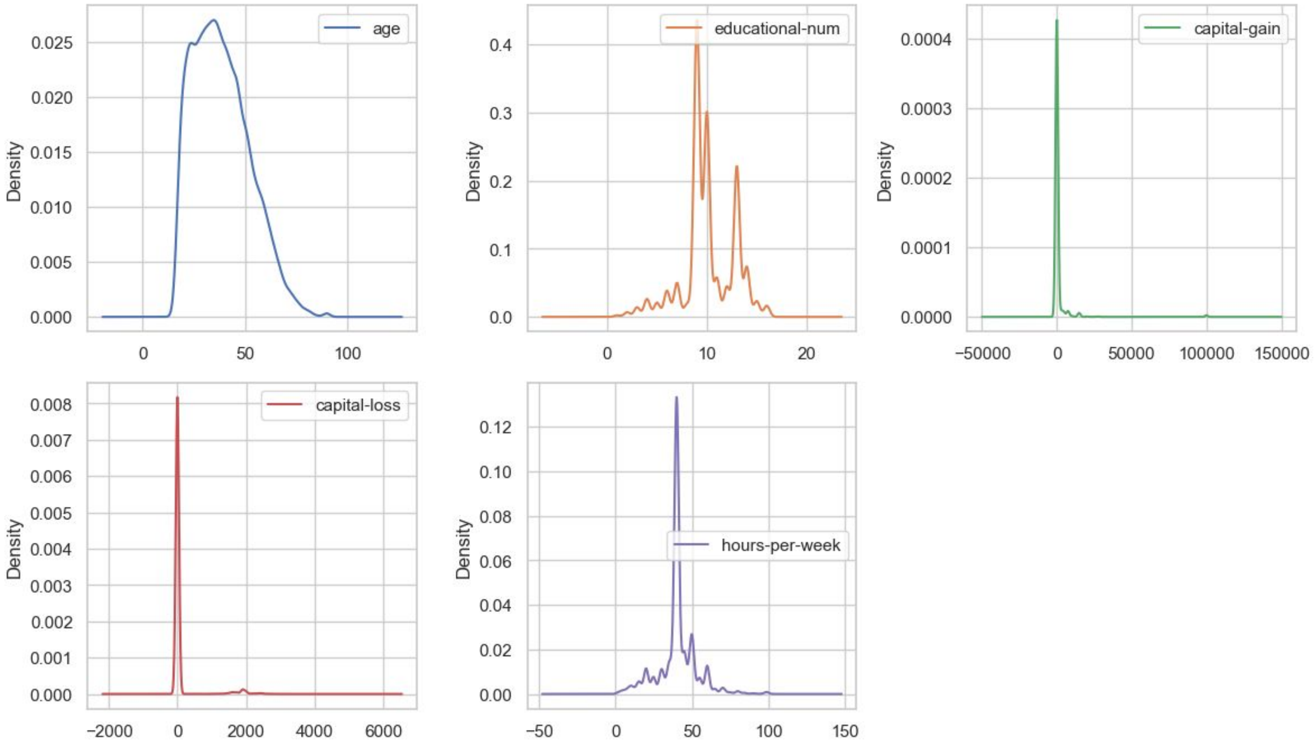
- ★ Box Plots:
  - ○ Summarize the distribution and identify outliers.
  - ○ Use data.boxplot() to create box plots for numerical columns.


8
7
6
5
4
3
2
1
0
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)


100000
80000
60000
40000
20000
0
age
educational-num
capital-gain
capital-loss
hours-per-week

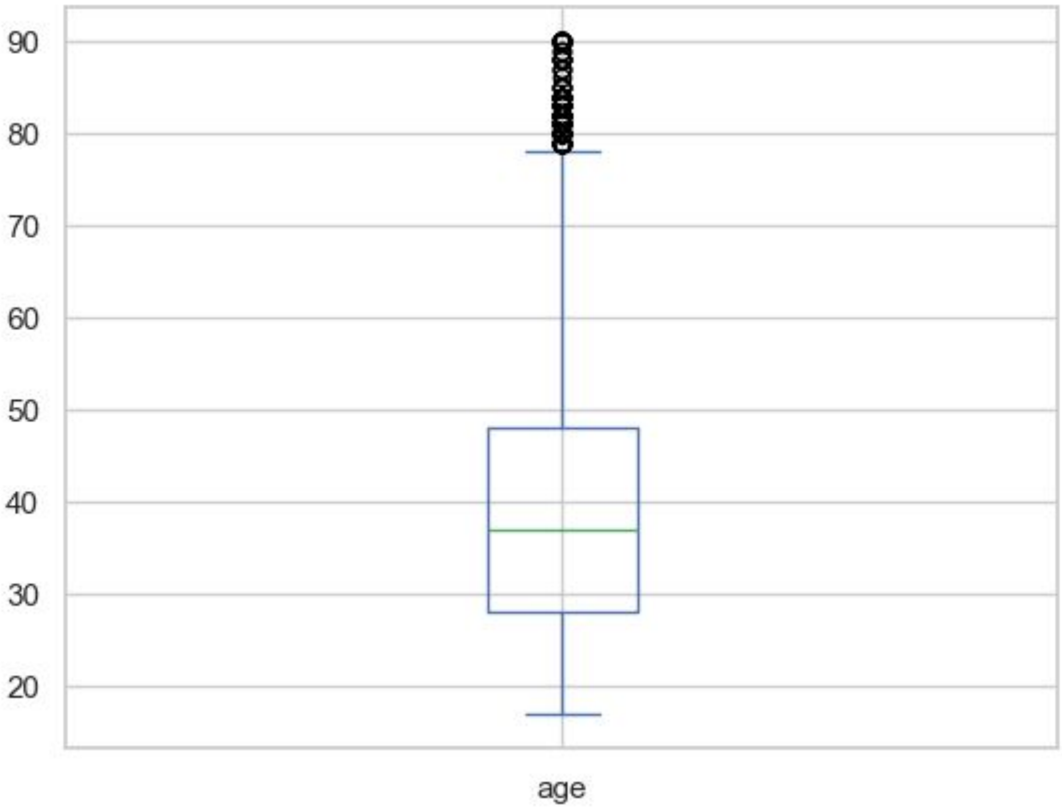
# UA - Visualization Techniques

- ★ Density Plots:
  - ○ Estimate the probability density function of a variable.
  - ○ Use data.plot(kind='density') to create density plots for numerical columns.


age
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
educational-num
Density
0.4
0.3
0.2
0.1
0.0
0
10
20
capital-gain
Density
0.0004
0.0003
0.0002
0.0001
0.0000
−50000
0
50000
100000
150000
capital-loss
Density
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−2000
0
2000
4000
6000
hours-per-week
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−50
0
50
100
150

# UA - Outliers and Missing Values

- Identifying Outliers:
  - Use box plots to visually identify outliers.
  - Calculate the interquartile range (IQR) and define the lower and upper bounds for outliers.

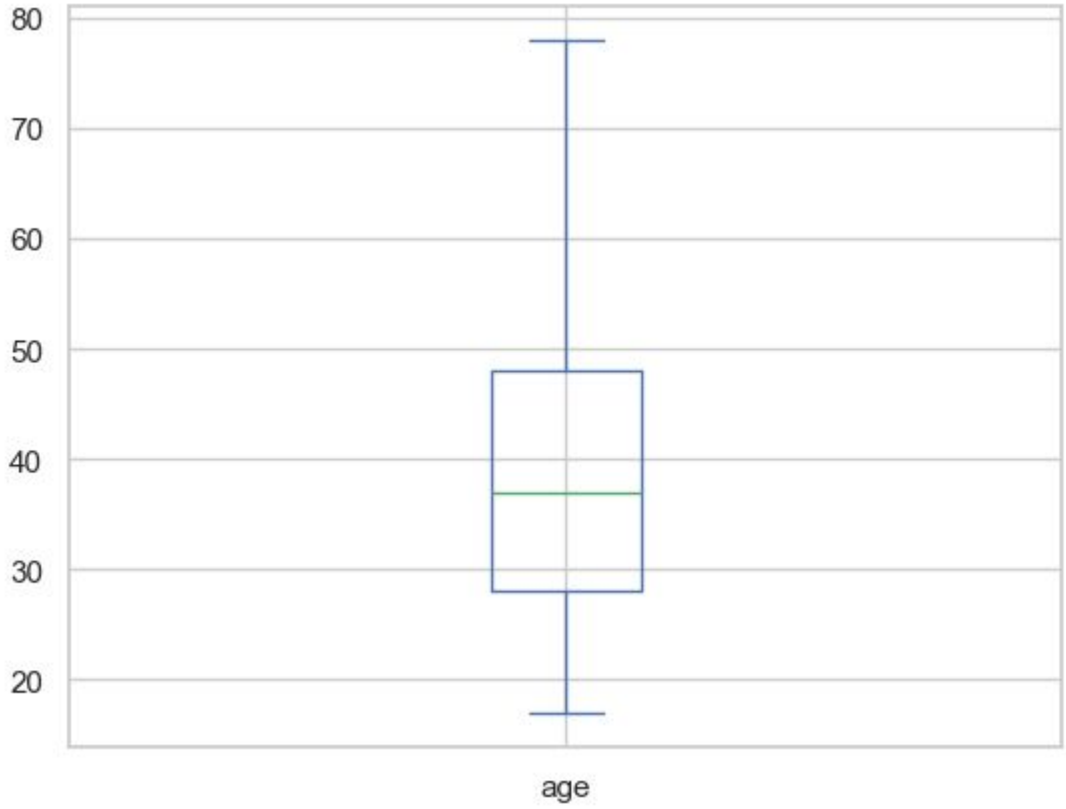

90
80
70
60
50
40
30
20
age

# UA - Outliers and Missing Values

★ Treating Outliers:

  ○ Cap the outlier values based on the lower and upper bounds.

CoGrammar



# What is the purpose of Exploratory Data Analysis (EDA)?

A. To build predictive models

B. To investigate and understand a dataset through visual and statistical techniques

C. To clean and preprocess the data

D. To deploy machine learning models

# Which of the following is NOT a measure of central tendency?

A. Mean

B. Median

C. Mode

D. Range

# Which Python function is used to calculate descriptive statistics for numerical columns?

A. data.info()

B. data.describe()

C. data.head()

D. data.tail()

# Which plot is used to visualize the distribution of a single variable?

A. Scatter plot

B. Box plot

C. Histogram

D. Heatmap

# How are outliers typically identified in a box plot?

A. Points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR

B. Points below Q1 - 2 * IQR or above Q3 + 2 * IQR

C. Points below Q1 - 3 * IQR or above Q3 + 3 * IQR

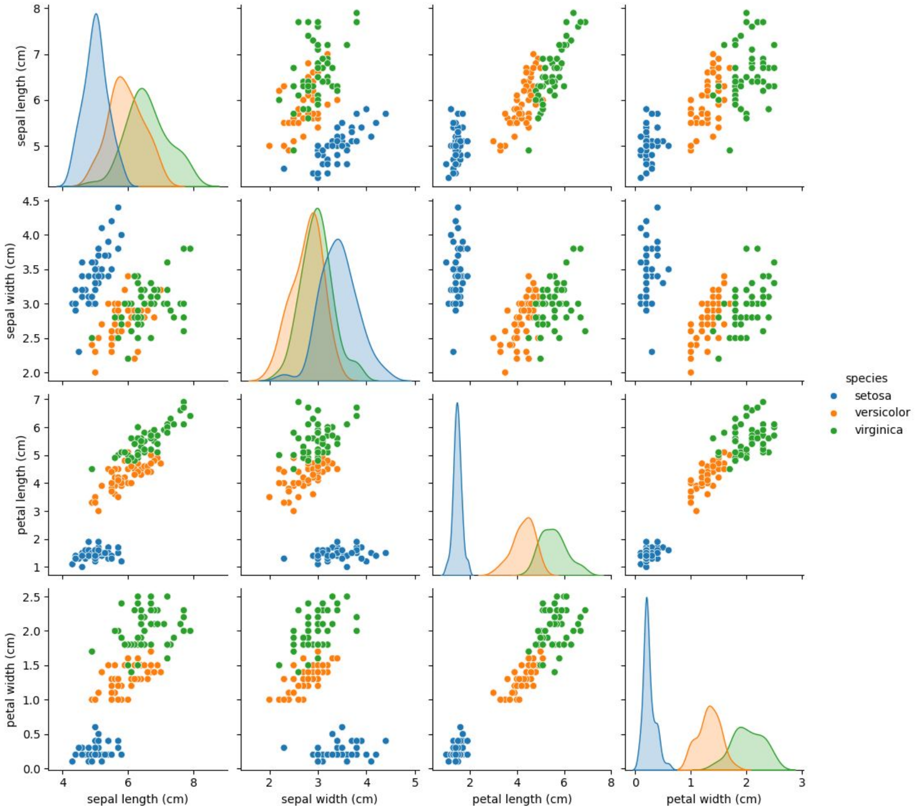
D. Points below Q1 - 0.5 * IQR or above Q3 + 0.5 * IQR

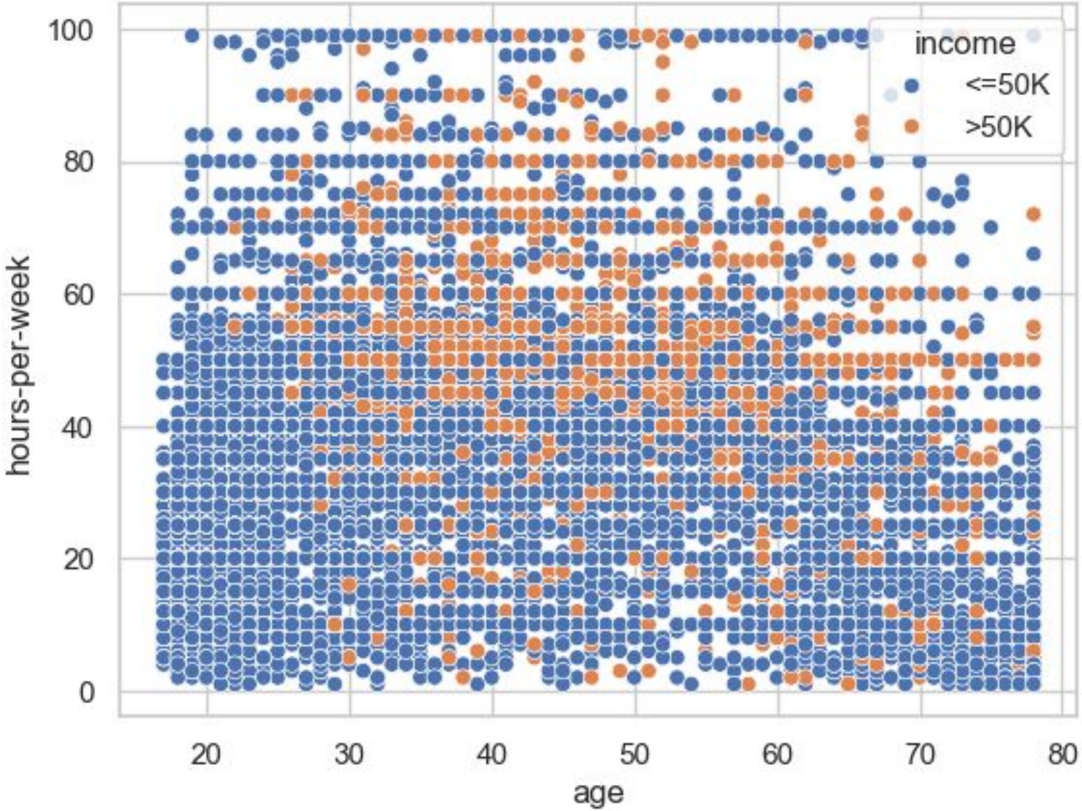
# Let’s Breathe!

Let’s take a small break before moving on to the next topic.

# Bivariate Analysis - Scatter Plots

- ★ Scatter Plots:
  - ○ Visualize the relationship between two continuous variables.
  - ○ Use sns.scatterplot() from the Seaborn library to create scatter plots.


sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
8
7
6
5
4.5
4.0
3.5
3.0
2.5
2.0
7
6
5
4
3
2
1
2.5
2.0
1.5
1.0
0.5
0.0
4
6
8
2
3
4
5
2
4
6
8
0
1
2
3
species
setosa
versicolor
virginica


100
80
60
40
20
0
hours-per-week
20
30
40
50
60
70
80
age
income
● <=50K
● >50K

# Bivariate Analysis - Scatter Plots

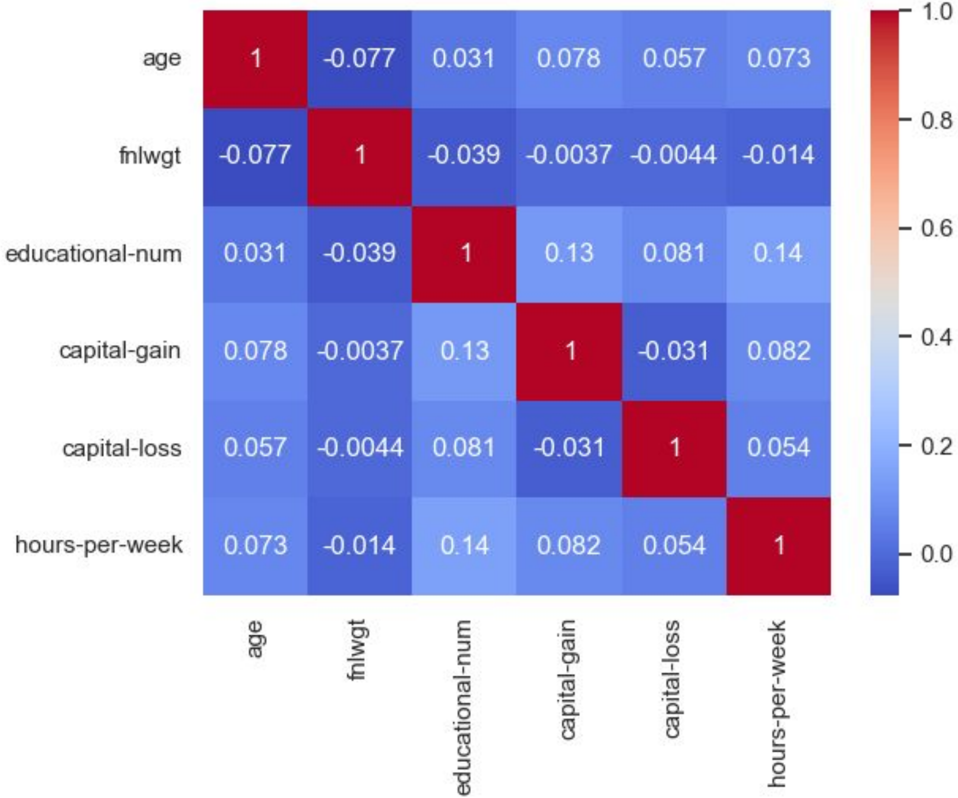
★ Interpreting Scatter Plots:

  ○ Observe the pattern and direction of the relationship between variables.

  ○ Identify clusters, outliers, or any interesting pattern

# BA - Correlation Analysis

- Correlation Matrix:
  - Calculate the correlation matrix for numerical variables using data.corr().
- Heatmap Visualization:
  - Visualize the correlation matrix using a heatmap with sns.heatmap().

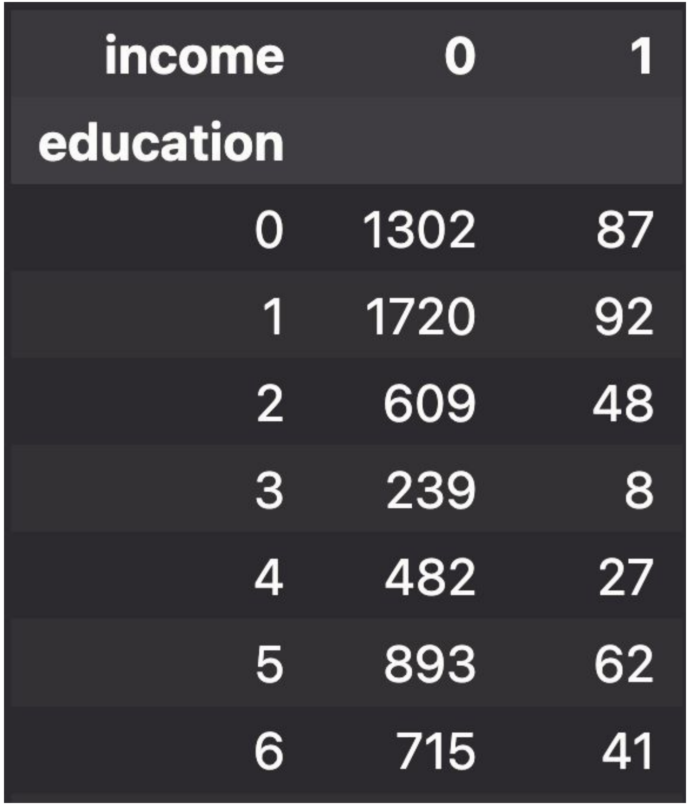

age
fnlwgt
educational-num
capital-gain
capital-loss
hours-per-week
1 -0.077 0.031 0.078 0.057 0.073
-0.077 1 -0.039 -0.0037 -0.0044 -0.014
0.031 -0.039 1 0.13 0.081 0.14
0.078 -0.0037 0.13 1 -0.031 0.082
0.057 -0.0044 0.081 -0.031 1 0.054
0.073 -0.014 0.14 0.082 0.054 1
age
fnlwgt
educational-num
capital-gain
capital-loss
hours-per-week
1.0
0.8
0.6
0.4
0.2
0.0

# BA - Correlation Analysis

- ★ Interpreting Correlation:
  - ○ Identify the strength and direction of the linear relationship between variables.
  - ○ Correlation values range from -1 to 1, with 0 indicating no linear relationship.

# BA - Contingency Tables

- ★ Contingency Tables:
  - ○ Create contingency tables to summarize the relationship between two categorical variables.
  - ○ Use pd.crosstab() to create contingency tables.



| income | 0 | 1 |
|---:|---:|---:|
| education | | |
| 0 | 1302 | 87 |
| 1 | 1720 | 92 |
| 2 | 609 | 48 |
| 3 | 239 | 8 |
| 4 | 482 | 27 |
| 5 | 893 | 62 |
| 6 | 715 | 41 |

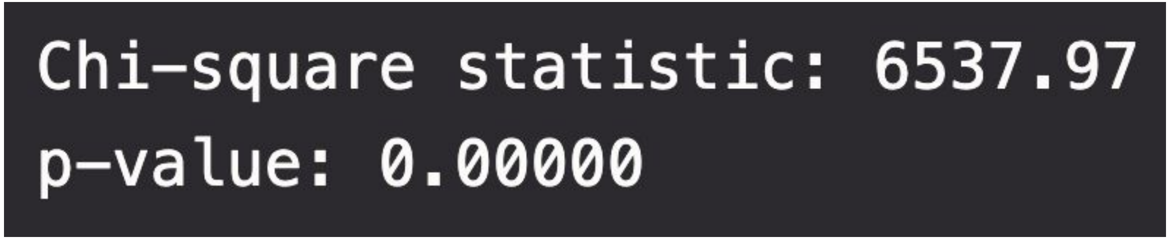
# BA - Chi-square Test

★ Chi-square Test of Independence:

- ○ Perform the chi-square test to determine if there is a significant association between categorical variables.
- ○ Use chi2_contingency() from the SciPy library to calculate the chi-square statistic and p-value.

# BA - Chi-square Test

- Interpreting Chi-square Results:
  - A low p-value (typically < 0.05) suggests a significant association between the variables.



```
Chi-square statistic: 6537.97
p-value: 0.00000
```

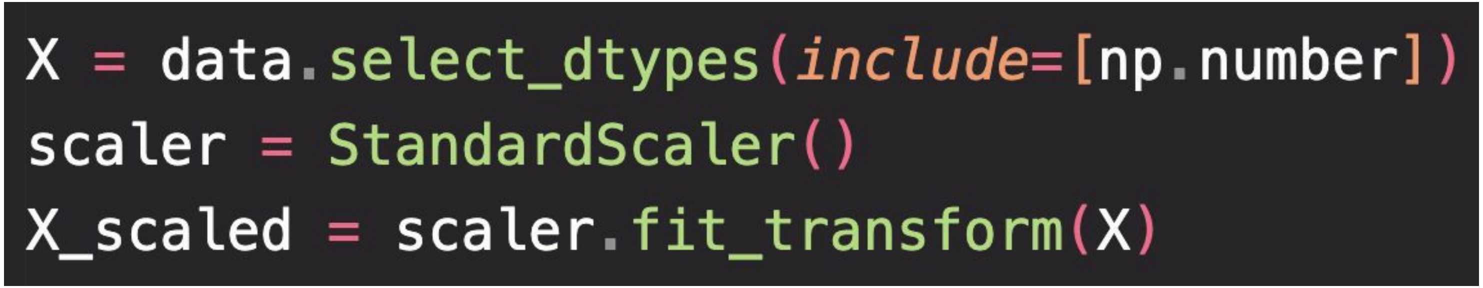
# Multivariate Analysis - PCA

- ★ Purpose of PCA:
  - ○ Reduce the dimensionality of the dataset while preserving the maximum variance.
  - ○ Transform the original features into a new set of uncorrelated features called principal components.

# Multivariate Analysis - PCA

★ Standardization:

  ○ Standardize the numerical features using StandardScaler from Scikit-learn.

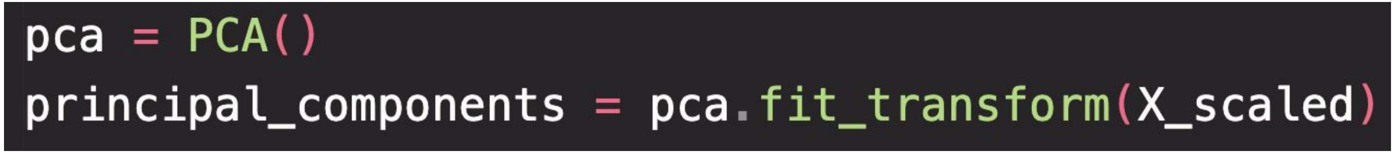


```
X = data.select_dtypes(include=[np.number])
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

# Multivariate Analysis - PCA

★ Applying PCA:

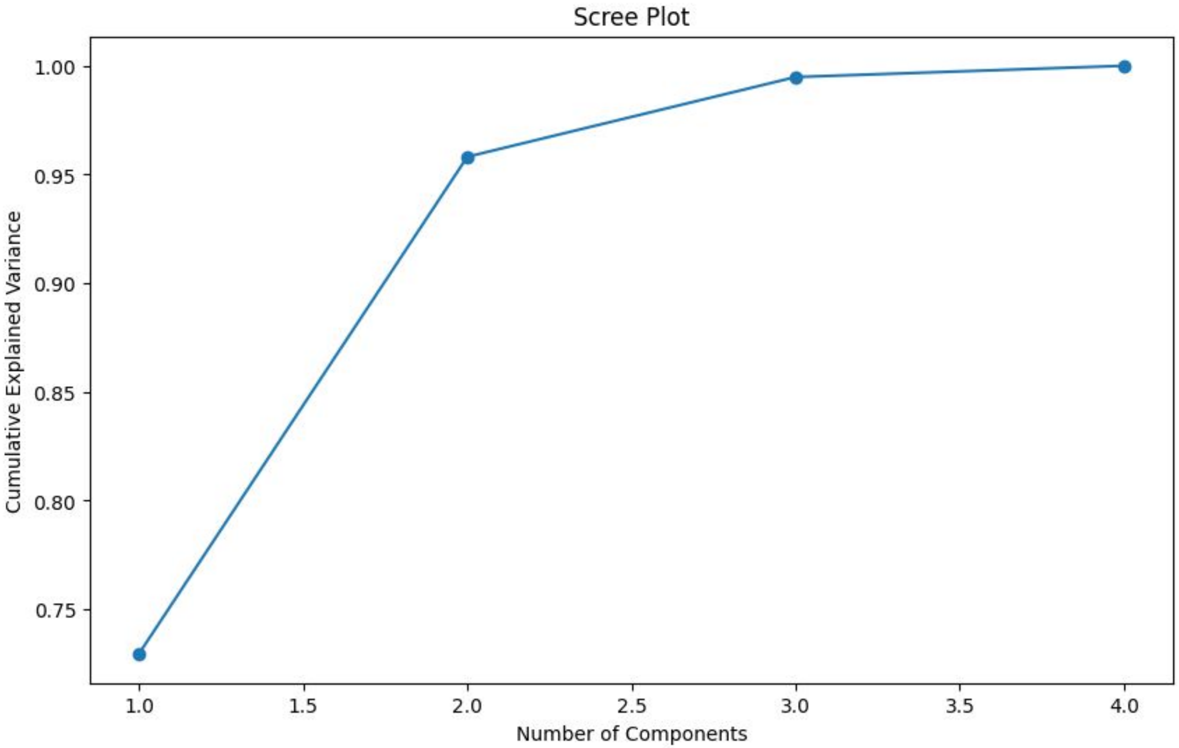
  ○ Use the PCA class from Scikit-learn to perform PCA on the scaled data.

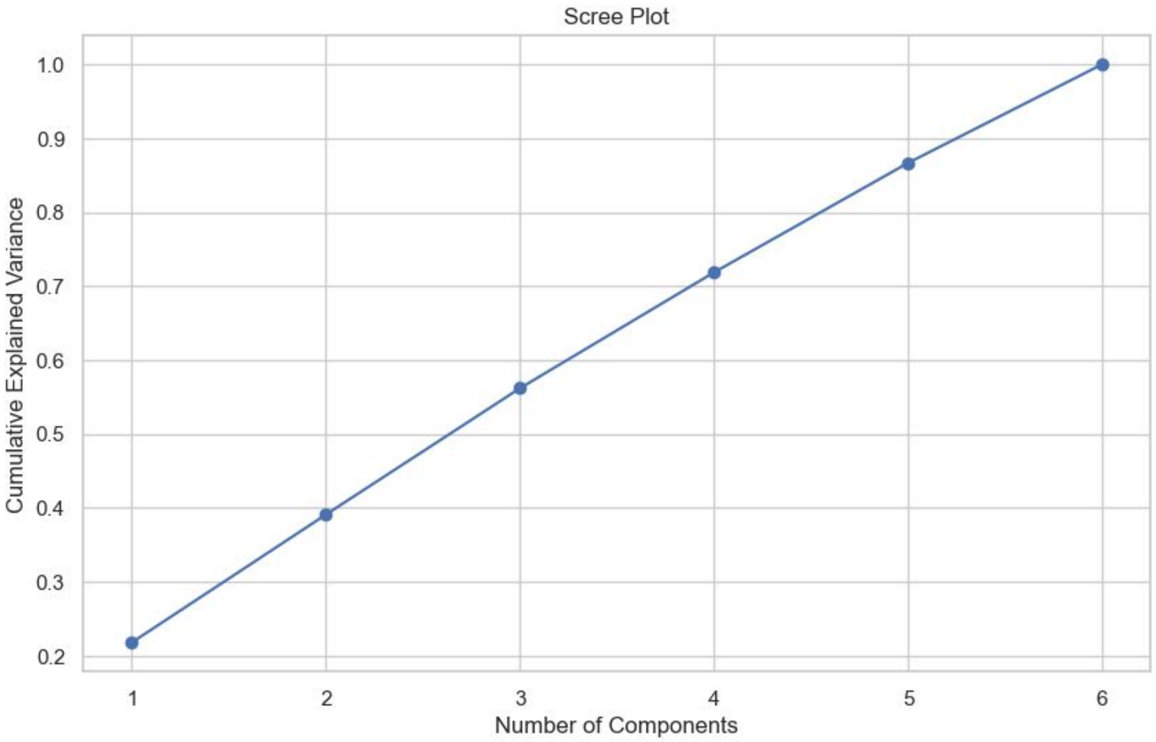


```
pca = PCA()
principal_components = pca.fit_transform(X_scaled)
```

# Multivariate Analysis - PCA

- Scree Plot:
  - Visualize the explained variance ratio of each principal component using a scree plot.


Scree Plot
Cumulative Explained Variance
1.00
0.95
0.90
0.85
0.80
0.75
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of Components


Scree Plot
Cumulative Explained Variance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1
2
3
4
5
6
Number of Components

# Multivariate Analysis - PCA

★ Interpretation:

  ○ You want to pick the smallest number of components that give the largest boost in explained variance

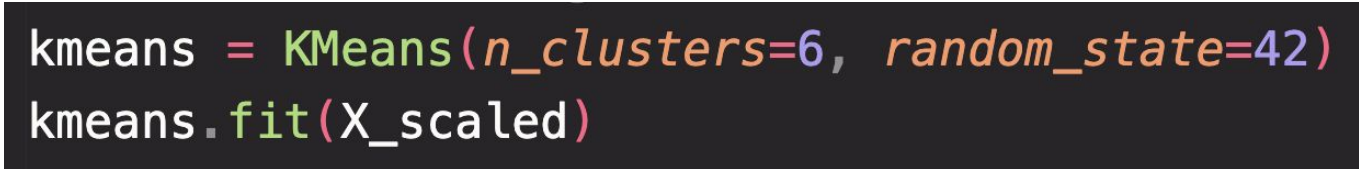
# MA - K-means Clustering

- ★ Purpose of K-means Clustering:
  - ○ Partition the data points into K clusters based on their similarity.
  - ○ Identify natural groupings or patterns in the data.

# MA - K-means Clustering

★ Applying K-means Clustering:

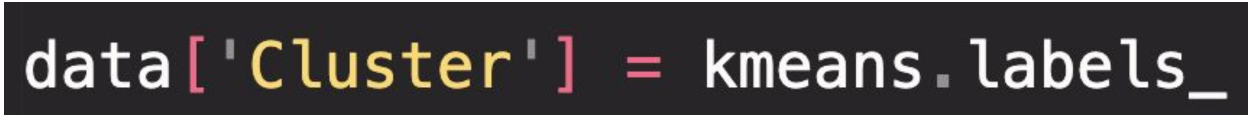
- ○ Use the KMeans class from Scikit-learn to perform k-means clustering on the scaled data.



```
kmeans = KMeans(n_clusters=6, random_state=42)
kmeans.fit(X_scaled)
```

# MA - K-means Clustering

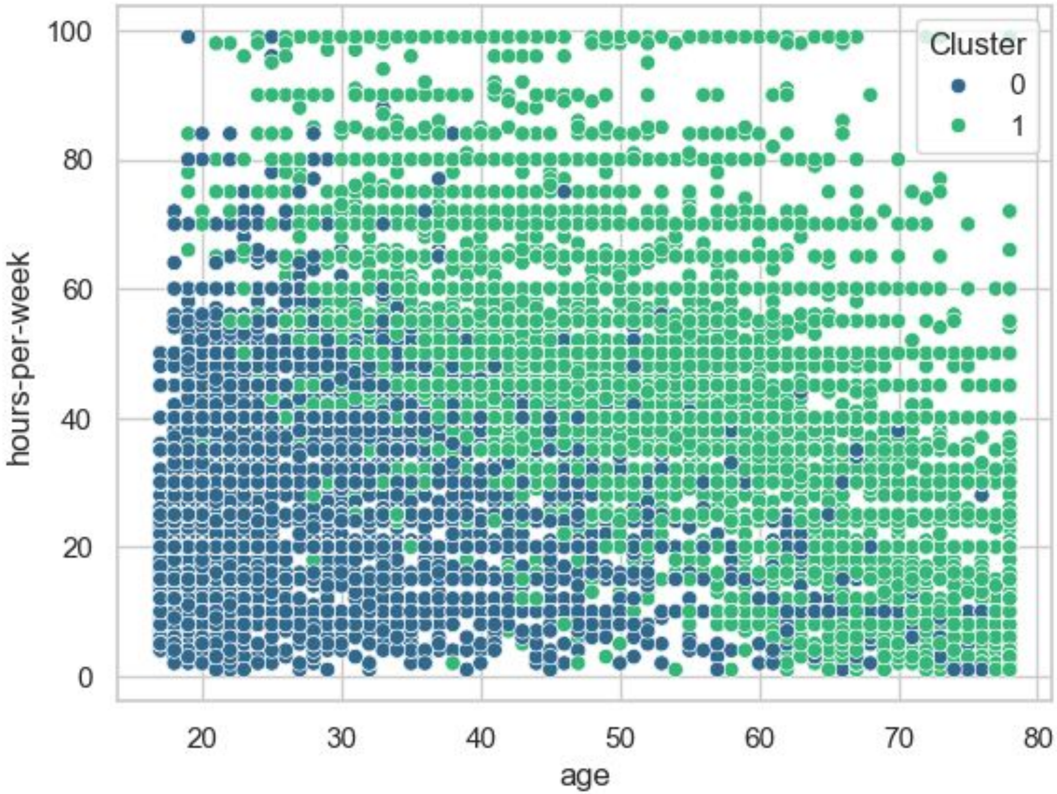
- Cluster Assignment:
  - Assign each data point to its corresponding cluster based on the k-means model.

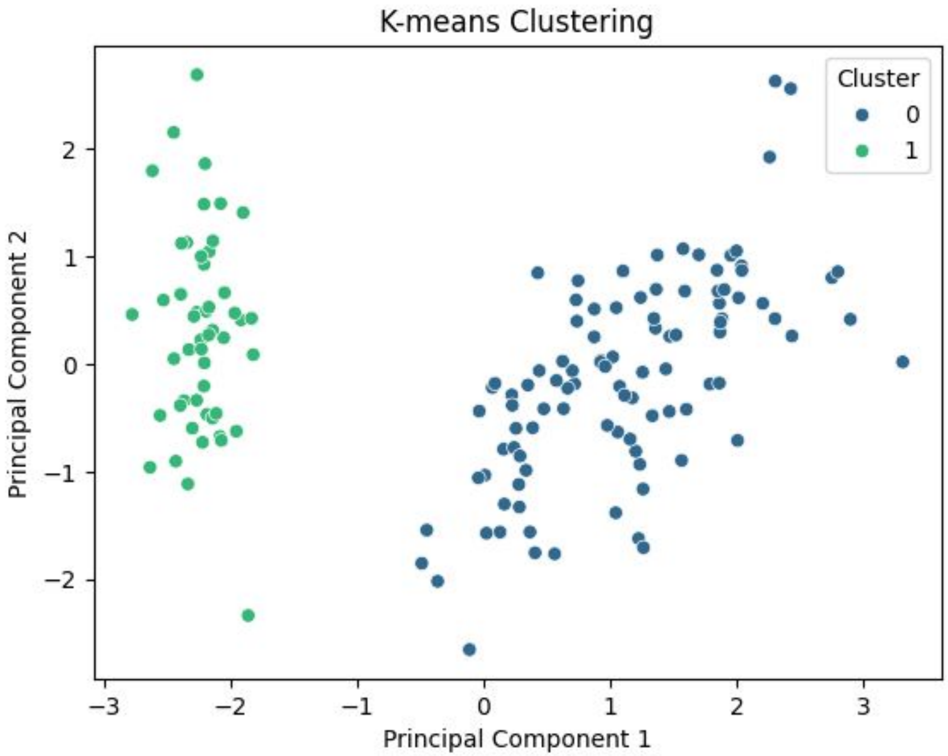


```
data['Cluster'] = kmeans.labels_
```

# MA - K-means Clustering

- ★ Visualizing Clusters:
    - ○ Use a scatter plot to visualize the clusters in the data.


Cluster
0
1
100
80
60
40
20
0
hours-per-week
20
30
40
50
60
70
80
age


K-means Clustering
Cluster
0
1
Principal Component 2
2
1
0
−1
−2
Principal Component 1
−3
−2
−1
0
1
2
3

# Feature Importance

★ Chi-square Test:

  ○ Perform the chi-square test for each categorical feature against the target variable.

  ○ Use pd.crosstab() to create contingency tables and chi2_contingency() to calculate the chi-square statistic and p-value.

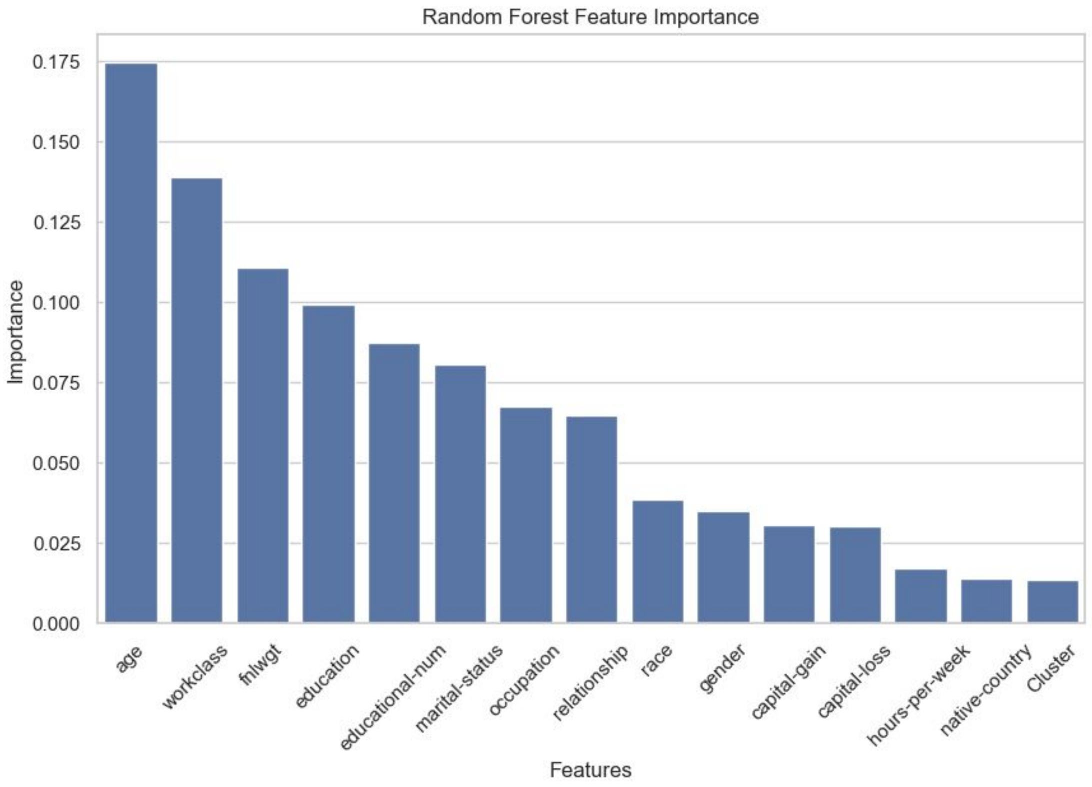
# Feature Importance

★ Interpreting Chi-square Results:

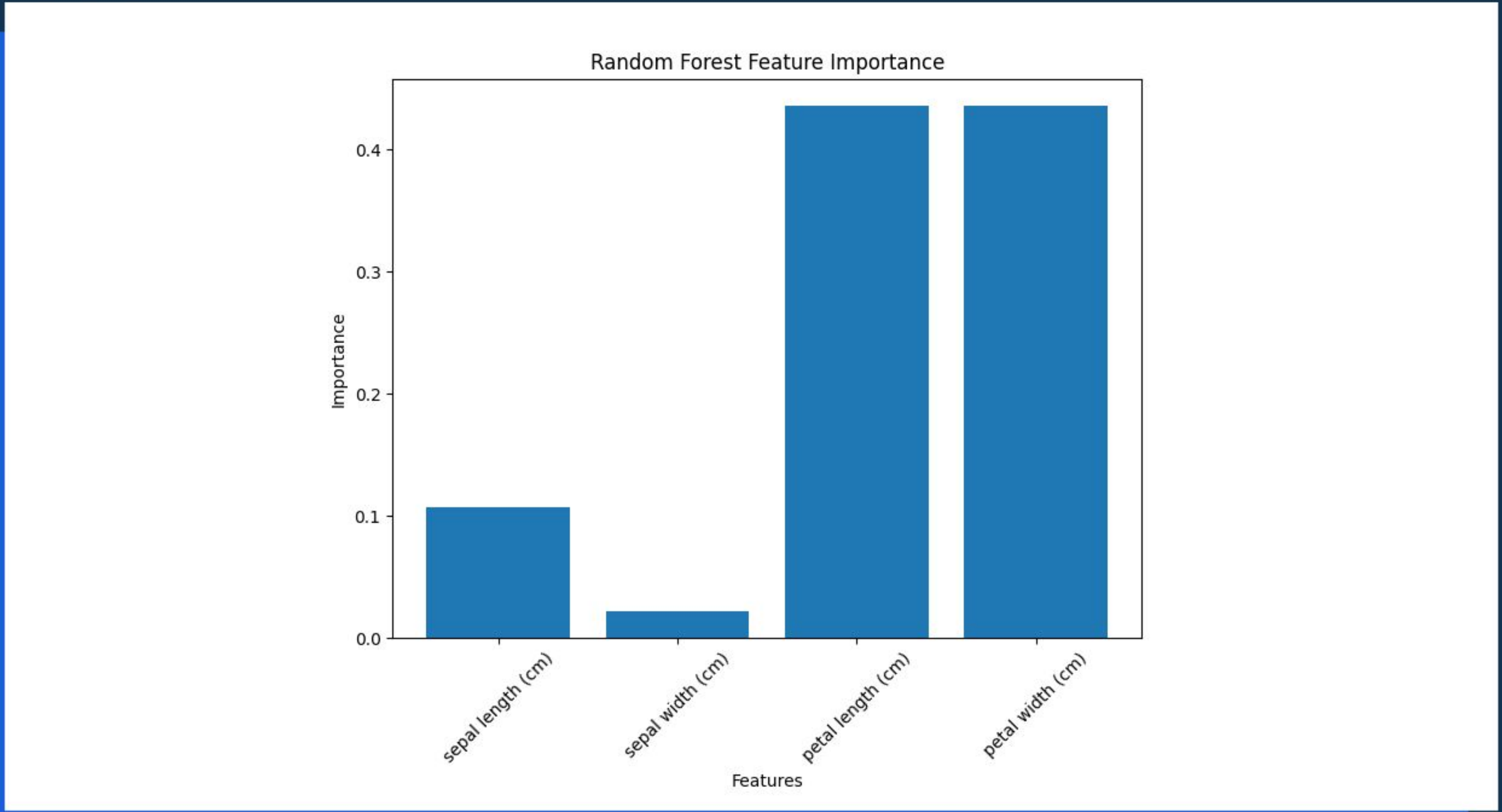
- A low p-value (typically < 0.05) suggests that the categorical feature is significantly associated with the target variable.
- The chi-square statistic measures the deviation from the expected frequencies under the assumption of independence.

# Feature Importance

- ★ Random Forest Classifier:
  - ○ Train a Random Forest classifier using the features and target variable.
  - ○ Use RandomForestClassifier from Scikit-learn to train the model.


Random Forest Feature Importance
Importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
age
workclass
fnlwgt
education
educational-num
marital-status
occupation
relationship
race
gender
capital-gain
capital-loss
hours-per-week
native-country
Cluster
Features

CoGrammar


Random Forest Feature Importance
Importance
0.0
0.1
0.2
0.3
0.4
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
Features

# What does a scatter plot visualize?

A. The distribution of a single variable

B. The relationship between two continuous variables

C. The correlation between all numerical variables

D. The clusters in the data

# What does a correlation value of 0 indicate?

A. Strong positive linear relationship
B. Strong negative linear relationship
C. No linear relationship
D. Perfect linear relationship

# Which library is used to perform the chi-square test of independence?

A. pandas

B. NumPy

C. Matplotlib

D. SciPy

# What is the purpose of Principal Component Analysis (PCA)?

A. To identify outliers in the dataset

B. To visualize clusters in the data

C. To reduce the dimensionality of the dataset while preserving maximum variance

D. To handle missing values in the dataset

# When interpreting PCA results, what should you consider when selecting the number of components?

A. Choose the components with the lowest explained variance

B. Select the largest number of components possible

C. Pick the smallest number of components that give the largest boost in explained variance

D. Ignore the explained variance and choose components randomly

CoGrammar

# Q & A SECTION

Please use this time to ask any questions relating to the topic, should you have any.

CoGrammar

# Thank you for joining!