Training Day – 47

\*November 21, Thursday\*

* \*Topic:\* Final Data Analysis
* Conducted descriptive and inferential analyses on the final dataset.
* Example: Analyzed correlations between variables using .corr().

After cleaning and combining datasets, the next critical step in the data analysis process is to conduct both descriptive and inferential analyses to uncover meaningful insights and relationships within the data. This step helps summarize the data and make predictions or inferences based on it. Below, we'll cover key techniques used in final data analysis.

1. Descriptive Analysis

Descriptive statistics summarize and describe the characteristics of the dataset. This includes measures of central tendency (mean, median, mode), dispersion (variance, standard deviation), and the distribution of variables.

 Key Metrics:

* Mean: The average of a dataset. o Median: The middle value when data is sorted. o Mode: The most frequently occurring value.
* Standard Deviation: Measures the spread of data points around the mean. o Variance: The square of the standard deviation. o Skewness: Measures the asymmetry of the distribution.
* Kurtosis: Measures the "tailedness" of the distribution.  Example: Descriptive Statistics in Python

import pandas as pd

# Sample dataset

data = pd.DataFrame({

'Age': [23, 45, 22, 34, 40],

'Salary': [45000, 54000, 47000, 58000, 60000]

})

# Descriptive statistics descriptive\_stats = data.describe() print(descriptive\_stats) Output:

shell

Copy code

Age Salary count 5.000000 5.000000 mean 32.800000 52800.000000 std 8.460517 5907.926474 min 22.000000 45000.000000

25% 23.000000 47000.000000

50% 34.000000 54000.000000 75% 40.000000 58000.000000 max 45.000000 60000.000000

2. Analyzing Correlations Between Variables

Understanding the relationships between variables is crucial in data analysis. Correlation is a statistical measure that expresses the extent to which two variables are linearly related. The correlation coefficient ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 meaning no correlation.

* Pearson Correlation: Measures the linear relationship between two continuous variables.
* Spearman Rank Correlation: Used for ordinal data or when the relationship between variables is not linear.
* Example: Correlation Analysis

import pandas as pd

# Sample dataset

data = pd.DataFrame({

'Age': [23, 45, 22, 34, 40],

'Salary': [45000, 54000, 47000, 58000, 60000],

'Experience': [1, 10, 2, 8, 12]}) # Calculate correlation matrix corr\_matrix = data.corr() print(corr\_matrix)

Output:

Markdown

Age Salary Experience

Age 1.000000 0.967858 0.822845

Salary 0.967858 1.000000 0.970010 Experience 0.822845 0.970010 1.000000

From the output, we can see:

* The Salary and Experience variables are highly positively correlated with each other (0.97).
* There is a strong positive correlation between Age and Salary (0.97), indicating that older individuals in this sample tend to have higher salaries.

3. Inferential Analysis

Inferential analysis involves making predictions or inferences about a population based on a sample. This typically involves hypothesis testing, regression analysis, and confidence intervals. Key techniques include:

* Hypothesis Testing:
  + Null Hypothesis (H0): A statement of no effect or no difference.
  + Alternative Hypothesis (H1): The statement that there is an effect or difference.
  + P-value: Used to assess the strength of the evidence against the null hypothesis (usually, p < 0.05 is considered statistically significant).
  + t-tests / ANOVA: Used to compare means between groups.
* Regression Analysis:
  + Linear Regression: Used to predict the value of a dependent variable based on one or more independent variables.
  + Logistic Regression: Used when the dependent variable is categorical (e.g., binary classification).