Probability Density Functions

Kalbe Digital University

KALBE Digital

Outline of Probability Density Functions

1. What is PDFs
2. Kernel Density Estimation
3. The Distribution Frameworks
   - Histogram Implementation
   - PMF Implementation
   - CDF Implementation
4. Moments
5. Skewness
6. Summary

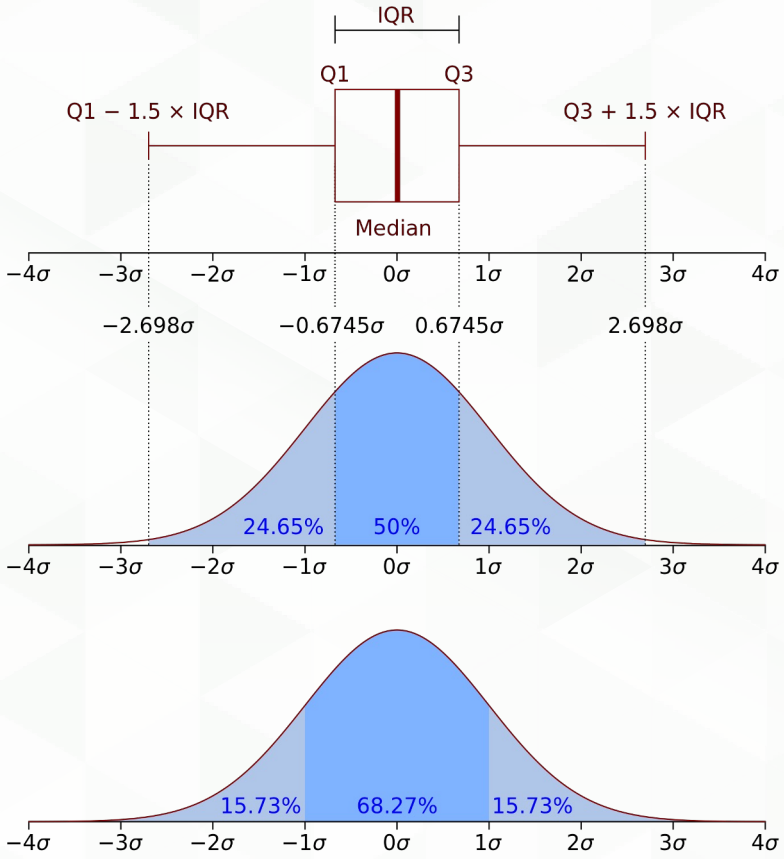
KALBE Digital

What is PDF

The PDF offers a visualization or mathematical representation of how values of a variable are distributed over its range.

Importance of PDF for Data Scientist:

- Insight into Data Distribution
- Anomaly Detection
- Assumption Checking
- Predictive Modeling
- Data Transformation Decisions



KALBE Digital

Practical Use Cases in Manufacturing

- Defect Detection
- Machine Performance
- Material Strength
- Production Time Analysis
- Inventory Management
- Maintenance Predictions
- Supplier Quality Control
- Process Variability
- Energy Consumption

PDF as valuable tool in manufacturing for analyzing & understanding continuous data distributions.

KALBE Digital

Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method to estimate the probability density function (pdf) of a continuous random variable. It provides a smoothed version of the histogram and is a way to recover the underlying distribution of a dataset.

KDE Use Cases:

- Understanding Salary Distributions
- Customer Transaction Patterns
- Fraud Detection
- Customer Onboarding Patterns
- Loan Application Patterns
- Credit Score Distribution
- User Login Patterns

FORMAL DEFINITION: Let (x_1, x_2, ..., x_n) be independent and identically distributed samples drawn from some univariate distribution with an unknown density f at any given point x. We are interested in estimating the shape of this function f. Its *kernel density estimator* is:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

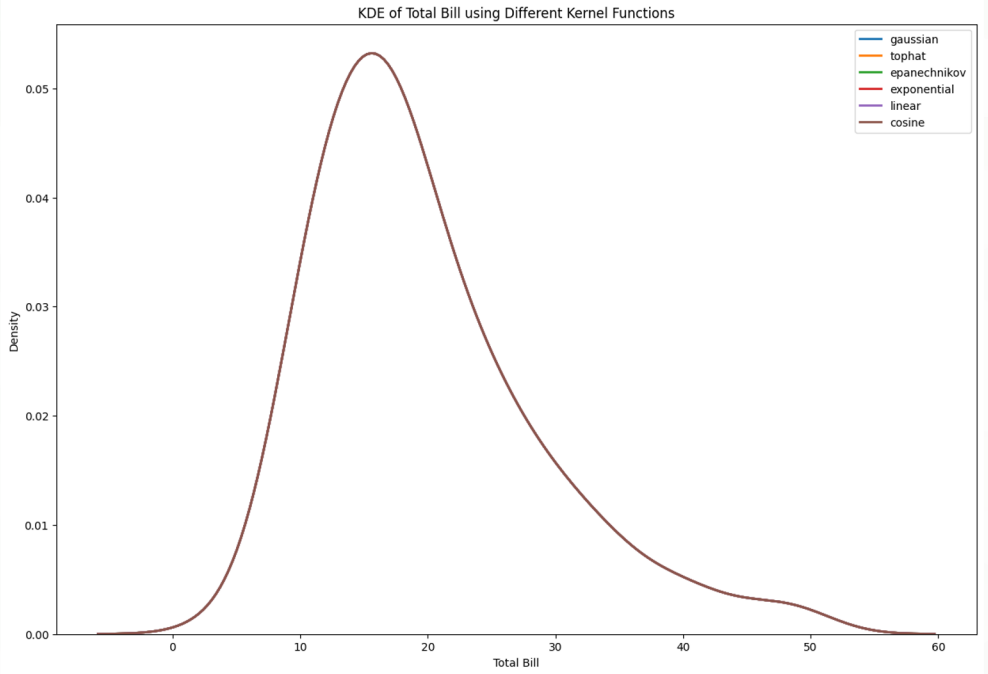
where K is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the *bandwidth*. A kernel with subscript h is called the *scaled kernel* and defined as $K_h(x) = 1/h\ K(x/h)$.

KALBE Digital

KDE with Different Kernel Functions

- **Dataset:** Tips
- **Libraries:** Seaborn

```python
# Required Libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
tips = sns.load_dataset("tips")

# Define the kernel functions to be used
kernels = ['gaussian','tophat','epanechnikov','exponential','linear','cosine']

# Plot KDE for each kernel
plt.figure(figsize=(15, 10))

for kernel in kernels:
    sns.kdeplot(tips['total_bill'], kernel=kernel, label=kernel, lw=2)

plt.title('KDE of Total Bill using Different Kernel Functions')
plt.xlabel('Total Bill')
plt.ylabel('Density')
plt.legend()
plt.show()
```

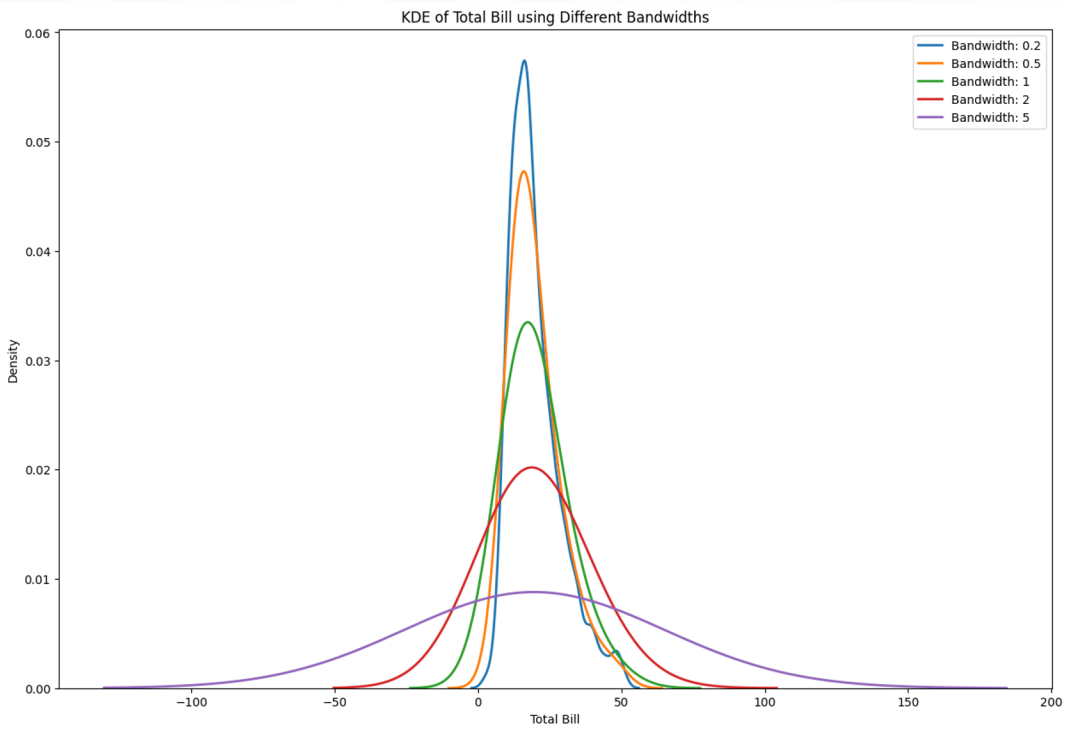


KALBE Digital

KDE with Different Bandwidth

- **Dataset:** Tips
- **Libraries:** Seaborn

```python
# Required Libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
tips = sns.load_dataset("tips")

# Define the bandwidths to be used
bandwidths = [0.2, 0.5, 1, 2, 5]

# Plot KDE for each bandwidth
plt.figure(figsize=(15, 10))

for bandwidth in bandwidths:
    sns.kdeplot(tips['total_bill'],bw=bandwidth,
                label=f"Bandwidth:{bandwidth}", lw=2)

plt.title('KDE of Total Bill using Different Bandwidths')
plt.xlabel('Total Bill')
plt.ylabel('Density')
plt.legend()
plt.show()
```



KALBE Digital

The Distribution Frameworks

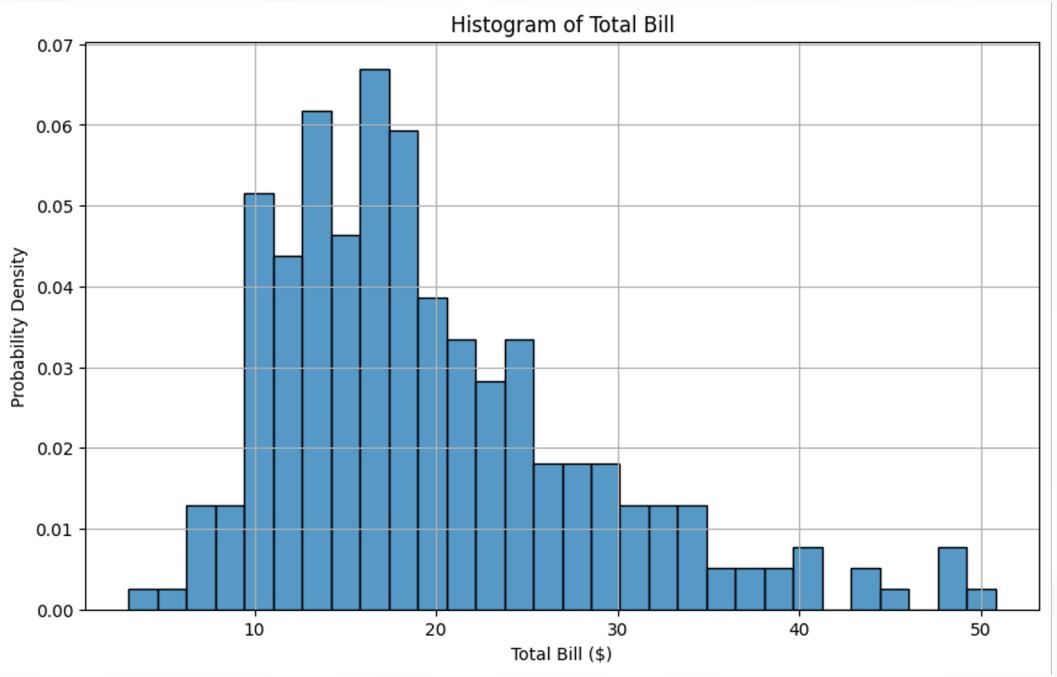
- **Standard Distributions:**
  - Normal, Exponential, Binomial, Poisson
- **Parameter Estimation:**
  - Mean and Standard Deviation
- **Fitting Distributions:**
  - Kolmogorov-Smirnov test
- **Continuous vs Discrete:**
  - Choice is based on nature of data
- **Transformations:**
  - Sometime Data does not adhere to standard distribution
- **Multivariate Distribution:**
  - Multiple variable correlation and join behaviours
- **Non-Parametric Distributions:**
  - Use KDE instead

KALBE Digital

Histogram Implementation

ChatGPT: Explain about Histogram implementation for probability density functions with sample codes in Python using Tips dataset.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Tips dataset
tips = sns.load_dataset('tips')

# Plotting the histogram for the 'total_bill' column
plt.figure(figsize=(10, 6))
sns.histplot(tips['total_bill'], bins=30, kde=False, stat="density")
plt.title('Histogram of Total Bill')
plt.xlabel('Total Bill ($)')
plt.ylabel('Probability Density')
plt.grid(True)
plt.show()
```

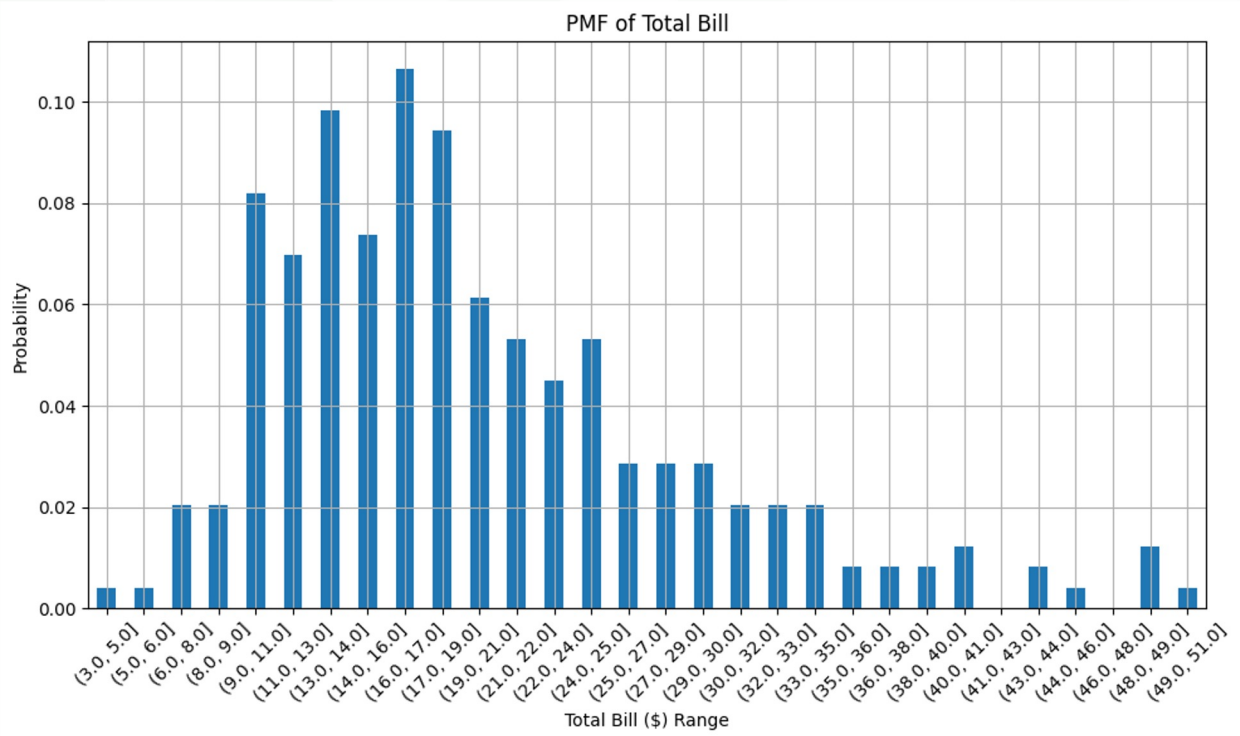


KALBE Digital

PMF Implementation

ChatGPT: Explain about PMF implementation for probability density functions with sample codes in Python using Tips dataset.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Tips dataset
tips = sns.load_dataset('tips')

# Discretize the 'total_bill' data into bins
bins = pd.cut(tips['total_bill'], bins=30, precision=0)

# Calculate the PMF for each bin
pmf = bins.value_counts(normalize=True).sort_index()

# Plotting the PMF
plt.figure(figsize=(10, 6))
pmf.plot(kind='bar')
plt.title('PMF of Total Bill')
plt.xlabel('Total Bill ($) Range')
plt.ylabel('Probability')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```

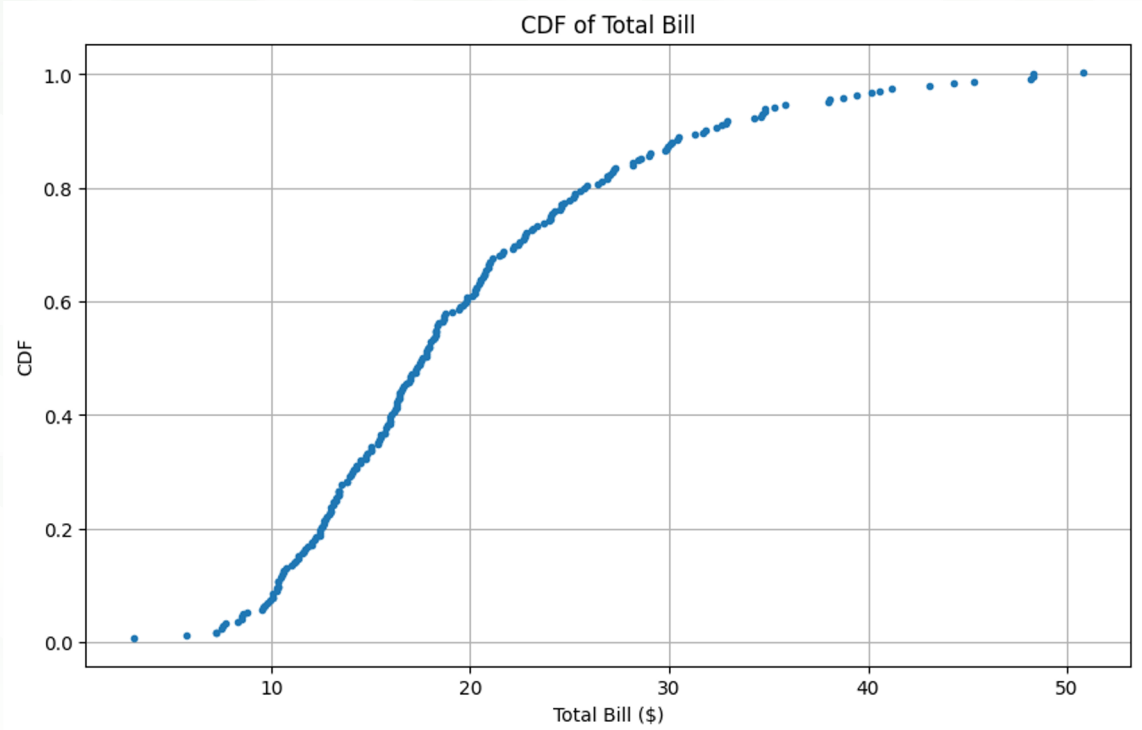


KALBE Digital

CDF Implementation

ChatGPT: Explain about CDF implementation for probability density functions with sample codes in Python using Tips dataset.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Tips dataset
tips = sns.load_dataset('tips')

# Compute the CDF
tips_sorted = tips['total_bill'].sort_values()
cdf = (1.0 + tips_sorted.rank(method='min')) / len(tips_sorted)

# Plotting the CDF
plt.figure(figsize=(10, 6))
plt.plot(tips_sorted, cdf, marker='.', linestyle='none')
plt.xlabel('Total Bill ($)')
plt.ylabel('CDF')
plt.title('CDF of Total Bill')
plt.grid(True)
plt.show()
```



KALBE Digital

Goodness of Fit Test

Determine if a sample comes from a specified distribution. In other words, it helps to determine the suitability of a distribution to a dataset.

- Kolmogorov-Smirnov test
- Anderson-Darling Test
- Shapiro-Wilk Test
- Lilliefors Test

```
Kolmogorov-Smirnov Test:
Statistic: 0.11875388157506361, P-value: 0.001866364156047776

Anderson-Darling Test:
Statistic: 5.520705531383527
At significance level 15.0: Statistic >= Critical Value. Data doesn't look normal.
At significance level 10.0: Statistic >= Critical Value. Data doesn't look normal.
At significance level 5.0: Statistic >= Critical Value. Data doesn't look normal.
At significance level 2.5: Statistic >= Critical Value. Data doesn't look normal.
At significance level 1.0: Statistic >= Critical Value. Data doesn't look normal.

Shapiro-Wilk Test:
Statistic: 0.9197186231613159, P-value: 3.3244529351605934e-10

Lilliefors Test:
Statistic: 0.11875388157506361, P-value: 0.0009999999999998899
```

```python
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns

# Load the Tips dataset
tips = sns.load_dataset('tips')

# Take the 'total_bill' column for testing
data = tips['total_bill']

# Kolmogorov-Smirnov Test
ks_statistic, ks_p_value = stats.kstest(data, 'norm', args=(data.mean(), data.std()))
print(f"Kolmogorov-Smirnov Test:\nStatistic: {ks_statistic}, P-value: {ks_p_value}\n")

# Anderson-Darling Test
result = stats.anderson(data, dist='norm')
print(f"Anderson-Darling Test:\nStatistic: {result.statistic}")
for i in range(len(result.critical_values)):
    sl, cv = result.significance_level[i], result.critical_values[i]
    if result.statistic < cv:
        print(f"At significance level {sl}: Statistic < Critical Value. Data looks normal.")
    else:
        print(f"At significance level {sl}: Statistic >= Critical Value. Data doesn't look normal.")

# Shapiro-Wilk Test
shapiro_statistic, shapiro_p_value = stats.shapiro(data)
print(f"\nShapiro-Wilk Test:\nStatistic: {shapiro_statistic}, P-value: {shapiro_p_value}\n")

# Lilliefors Test (Note: Need to install statsmodels for this)
from statsmodels.stats.diagnostic import lilliefors

lillie_statistic, lillie_p_value = lilliefors(data, dist='norm')
print(f"Lilliefors Test:\nStatistic: {lillie_statistic}, P-value: {lillie_p_value}\n")
```

KALBE Digital

Moments

The moments give an insight into the characteristics and features of a distribution. They include the following:

- Zeroth Moment
- First Moment (Mean)
- Second Moment (Variance)
- Third Moment (Skewness)
- Fourth Moment (Kurtosis)

```python
import pandas as pd
import seaborn as sns

# Load the Tips dataset
tips = sns.load_dataset('tips')

# Calculate moments of 'total_bill'
mean = tips['total_bill'].mean()
variance = tips['total_bill'].var()
skewness = tips['total_bill'].skew()
kurtosis = tips['total_bill'].kurt()

print(f"Mean (1st moment): {mean:.2f}")
print(f"Variance (2nd moment): {variance:.2f}")
print(f"Skewness (3rd moment): {skewness:.2f}")
print(f"Kurtosis (4th moment): {kurtosis:.2f}")
```

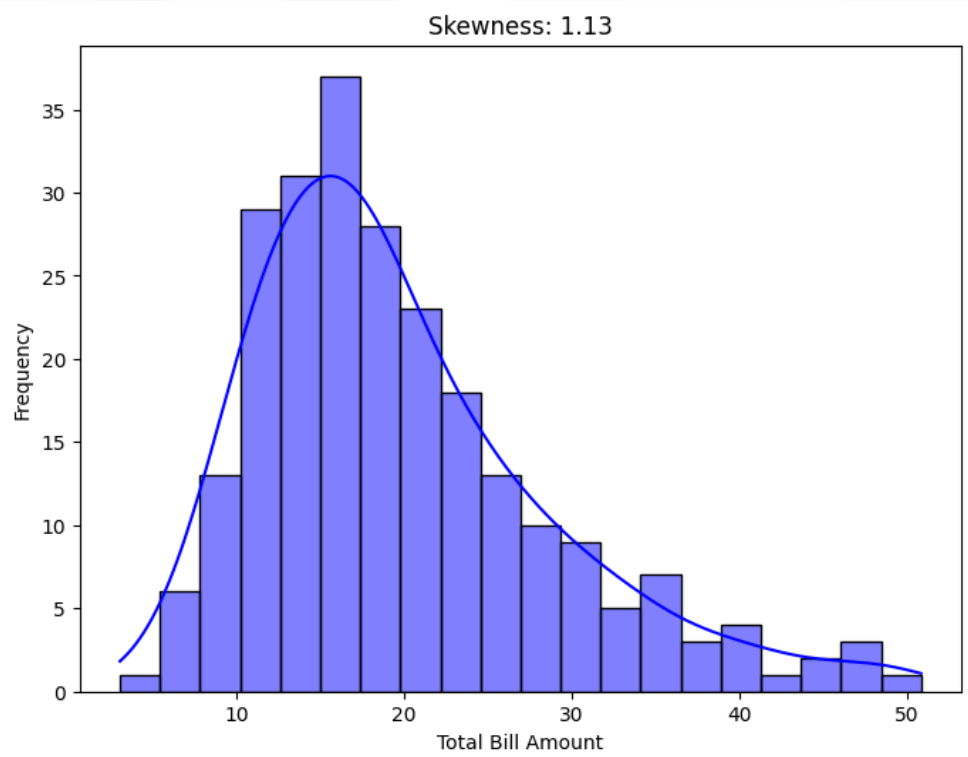
```
Mean (1st moment): 19.79
Variance (2nd moment): 79.25
Skewness (3rd moment): 1.13
Kurtosis (4th moment): 1.22
```

KALBE Digital

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. It can be used to determine the direction and degree of skew (departure from horizontal symmetry) in the data. A negative skewness indicates a distribution that is skewed towards the left, while a positive skewness indicates a distribution that is skewed to the right.

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Load the "tips" dataset from seaborn
tips = sns.load_dataset("tips")

# Calculate the skewness of the total bill column
skewness = tips["total_bill"].skew()

# Create a histogram of the total bill column
plt.figure(figsize=(8, 6))
sns.histplot(tips["total_bill"], kde=True, color='b', bins=20)

# Add labels and title
plt.xlabel('Total Bill Amount')
plt.ylabel('Frequency')
plt.title(f'Skewness: {skewness:.2f}')

# Show the plot
plt.show()
```

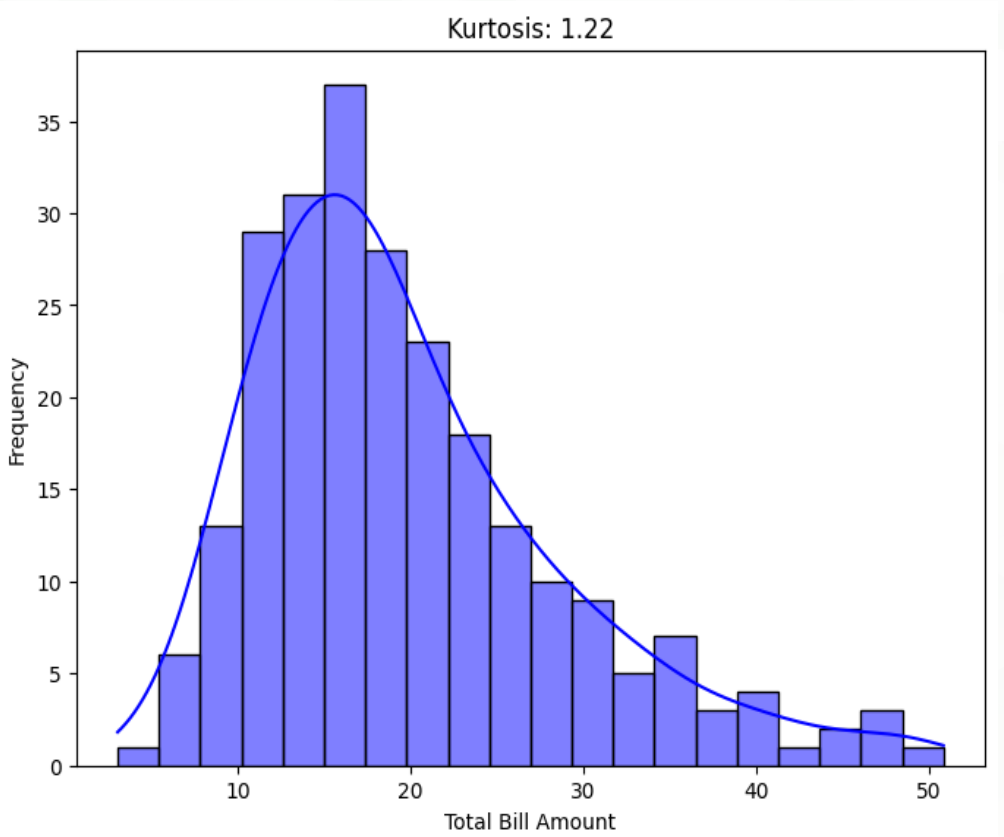


KALBE Digital

Kurtosis

Kurtosis is a statistical measure used to describe the distribution of observed data around the mean. It can be used to identify the peak and tails behavior of a distribution.

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Load the "tips" dataset from seaborn
tips = sns.load_dataset("tips")

# Calculate the kurtosis of the total bill column
kurtosis = tips["total_bill"].kurtosis()

# Create a histogram of the total bill column
plt.figure(figsize=(8, 6))
sns.histplot(tips["total_bill"], kde=True, color='b', bins=20)

# Add labels and title
plt.xlabel('Total Bill Amount')
plt.ylabel('Frequency')
plt.title(f'Kurtosis: {kurtosis:.2f}')

# Show the plot
plt.show()
```



KALBE Digital

Summary

- **Probability Density Functions (PDFs)**
  - Histogram
  - PMF
  - CDF
- **Kernel Density Estimations (KDEs)**
- **Skewness and Kurtosis**

KALBE Digital

Quiz

a) **Quiz 1: What is a Probability Density Function (PDF)?**
a) A function that gives the probability of a discrete random event.
b) A function that maps each outcome to the likelihood of observing a given range of continuous values.
c) A function that accumulates the probabilities for all possible outcomes.
d) A function that represents the distribution of a categorical variable.
Correct Answer: b) A function that maps each outcome to the likelihood of observing a given range of continuous values.

- **Quiz 2: Which of the following best describes skewness in a PDF?**
a) It measures the "tailedness" of the distribution.
b) It measures how flat or peaked a distribution is.
c) It measures the asymmetry of a distribution around its mean.
d) It measures the width of the distribution's peak.
Correct Answer: c) It measures the asymmetry of a distribution around its mean.
- **Quiz 3: What does Kernel Density Estimation (KDE) primarily help with in the context of PDFs?**
a) It helps in the estimation of the central tendency.
b) It aids in the creation of a smooth curve over the data points to estimate the PDF.
c) It helps in calculating the skewness of a distribution.
d) It assists in determining the correlation between two variables.
Correct Answer: b) It aids in the creation of a smooth curve over the data points to estimate the PDF.

Relationship Between Variables

Kalbe Digital University

KALBE Digital

Outline of Relationship Between Variables

1. Scatter Plot
2. Characterizing Relationships
3. Correlation Analysis
4. Covariance Analysis
5. Pearson’s Correlation
6. Nonlinear Relationship
7. Spearman’s Rank Correlation
8. Correlation and Causation
9. Summary

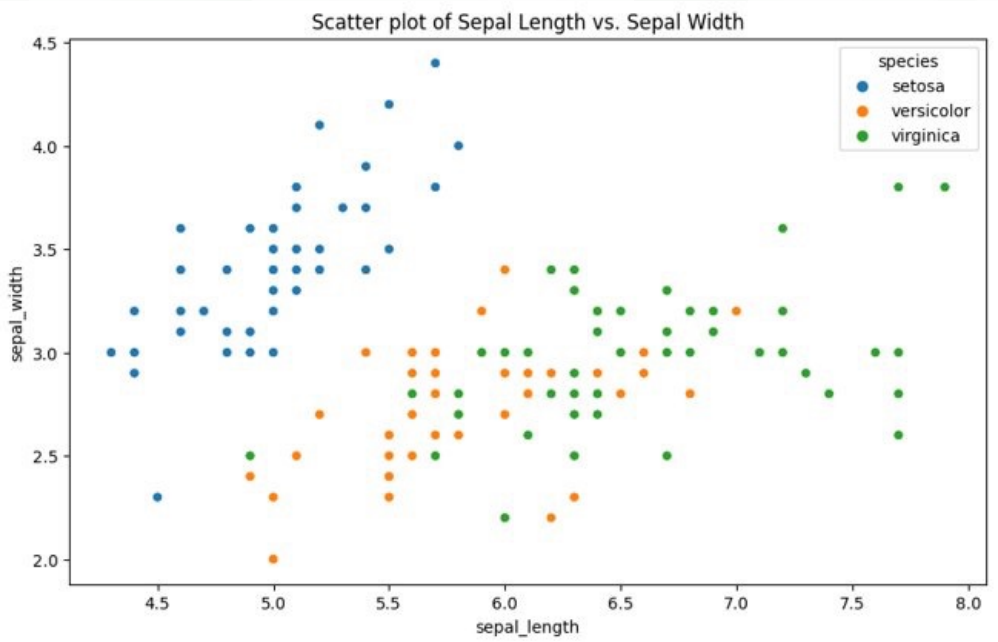
KALBE Digital

Scatter Plot

Scatter plots are a foundational tool in EDA to visualize and understand relationships between two quantitative variables.

- **Interpretation:**
  - Positive Relationship
  - Negative Relationship
  - No Relationship
  - Strength of Relationship
- **Variations and Enhancements:**
  - Color Coding, Size Variations, and Fit Lines
- **Limitation:**
  - Overplotting and Linear Focus

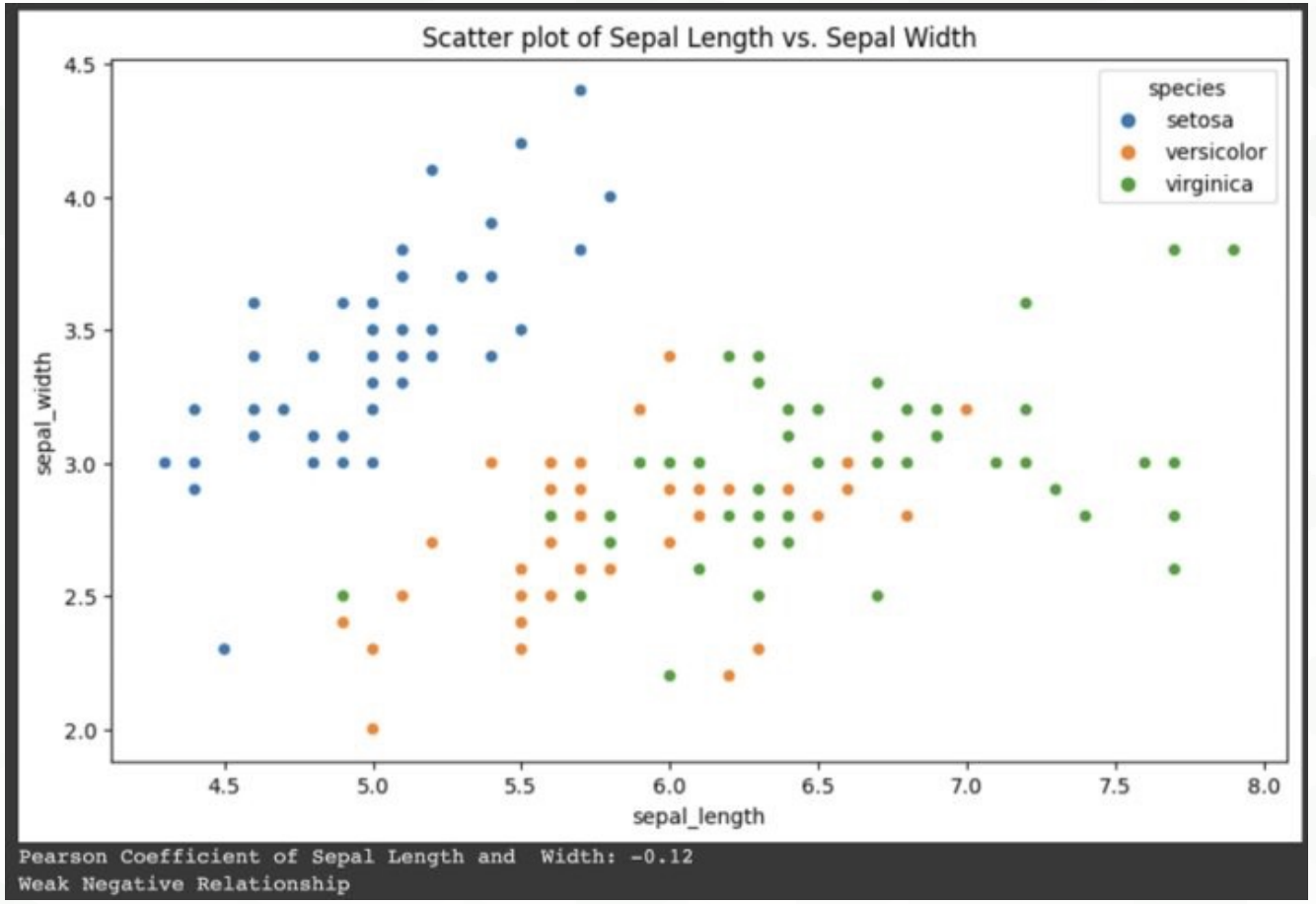
ChatGPT Prompt: Use Irish dataset, write sample code for scatter plot and analyze the result using Pandas and Seaborn.



KALBE Digital

Scatter Plot Interpretation

```python
# Import necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load the iris dataset
iris = sns.load_dataset('iris')

# Create a scatter plot using Seaborn's scatterplot function
plt.figure(figsize=(10, 6))
sns.scatterplot(x='sepal_length', y='sepal_width', hue='species', data=iris)
plt.title('Scatter plot of Sepal Length vs. Sepal Width')
plt.show()

# Calculate Pearson's correlation coefficient for sepal length and sepal width
correlation = iris['sepal_length'].corr(iris['sepal_width'])

print(f"Pearson Coefficient of Sepal Length and  Width: {correlation:.2f}")

# Interpretation
if correlation > 0.7:
    print("Strong Positive Relationship")
elif correlation > 0:
    print("Weak Positive Relationship")
elif correlation < -0.7:
    print("Strong Negative Relationship")
elif correlation < 0:
    print("Weak Negative Relationship")
else:
    print("No Relationship")
```




```
Pearson Coefficient of Sepal Length and  Width: -0.12
Weak Negative Relationship
```

KALBE Digital

Characterizing Relationship

- **Visual Methods:**
  - Scatter Plots
  - Box Plots
  - Pair Plots
  - Heatmaps
- **Correlation Techniques:**
  - Pearson
  - Spearman
  - Kendall
- **Categorical Relationships:**
  - Contingency Tables
  - Staked Bar Charts
- **Statistical Tests:**
  - T-Test and ANOVA
  - Chi-Square Test
  - Regression Analysis
  - Staked Bar Charts

ChatGPT Prompt: Use Irish dataset, write sample code for **Scatter plot** and analyze the result using Pandas and Seaborn.

ChatGPT Prompt: Use Irish dataset, write sample code for **Pearson correlation** analysis and interpret the result using Pandas and Seaborn.

ChatGPT Prompt: Use Irish dataset, write sample code for **Categorical Relationship with Contingency tables** and interpret the result using Pandas and Seaborn.

ChatGPT Prompt: Use Irish dataset, write sample code to implement **T-Test and ANOVA** then analyze the result using Pandas, Scipy and Seaborn.

KALBE Digital

Correlation Analysis

The correlation analysis is pivotal for understanding the relationship between two or more variables in a dataset.

Common Methods and Techniques:

- Pearson's Correlation Coefficient
- Spearman's Rank Correlation
- Kendal Tau

Common Methods and Techniques:

- Point-Biserial Correlation
- Phi Coefficient *
- Categorical Correlation
- Correlation Heatmaps
- Regression Analysis
- Partial Correlation
- Cross Correlation
- Spearman's Rank Correlation
- Kendal Tau

Key Considerations:

- Correlation does not imply causation
- Statistical Significance
- Outlier Identification
- Distribution

ChatGPT Prompt: Use publicly available dataset, write sample code for **Phi Coefficient** and interpret the result using Pandas, SciPy and Seaborn.

KALBE Digital

Covariance Analysis

The coveriance analysis is pivotal for understanding how two variables change together. If one variable tends to go up when the other goes up, there's a positive covariance. If one variable tends to go down when the other goes up, there's a negative covariance.

Common Methods and Techniques:

- Covariance Matrix
- Pearson's Correlation Coefficient
- Scatter Plots
- Bivariate Analysis *
- Linear Regression Analysis

Key Considerations:

- **ChatGPT Prompt:** Use publicly available dataset, write sample code for **Bivariate Analysis** and interpret the result using Pandas, SciPy and Seaborn.

ChatGPT Prompt: Use publicly available dataset, write sample code for **Bivariate Analysis** and interpret the result using Pandas, SciPy and Seaborn.

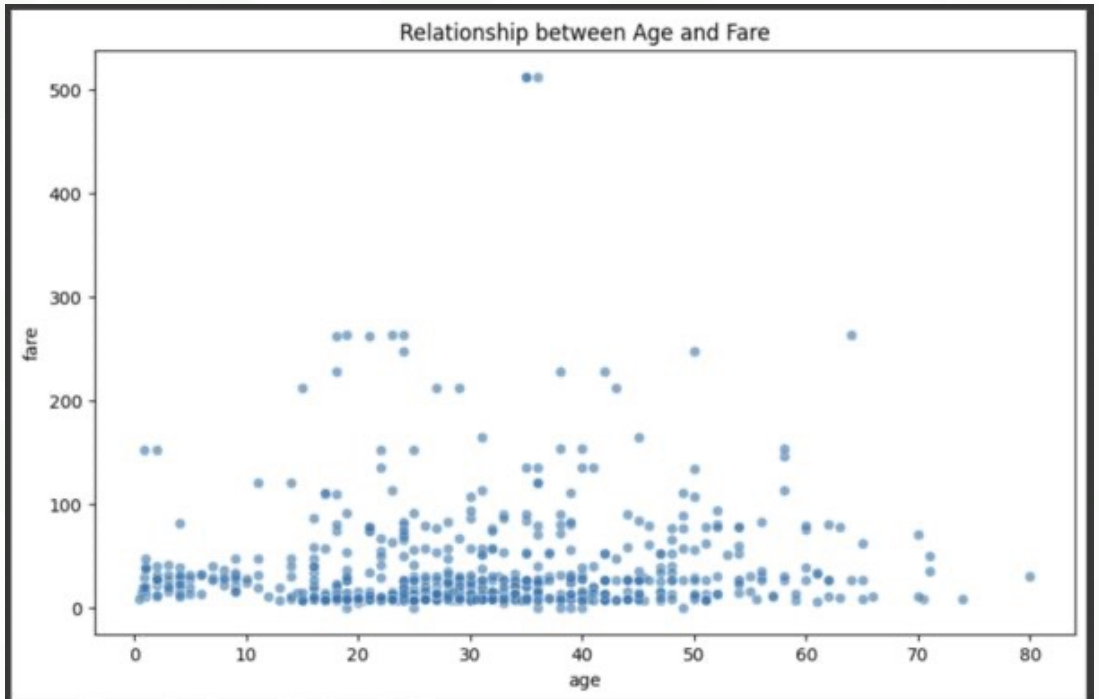
KALBE Digital

Pearson's Correlation Coefficient

Pearson's correlation coefficient, denoted as *r*, measures the linear relationship between two datasets. It's a value between -1 and 1 inclusive, where:1 indicates a perfect positive linear relationship (Wikipedia).

- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear correlation between the variables.

```python
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

# Load the Titanic dataset
titanic = sns.load_dataset('titanic')

# visually inspect the relationship between age and fare
plt.figure(figsize=(10, 6))
sns.scatterplot(x='age', y='fare', data=titanic, alpha=0.6)
plt.title("Relationship between Age and Fare")
plt.show()


# Drop rows with missing values for 'age' and 'fare'
titanic_cleaned = titanic.dropna(subset=['age', 'fare'])

# Calculate Pearson's correlation coefficient
corr_coefficient, p_value =  pearsonr(titanic_cleaned['age'],
                                      titanic_cleaned['fare'])

print(f"Pearson's correlation coefficient: {corr_coefficient:.3f}")
print(f"P-value: {p_value:.3f}")
```



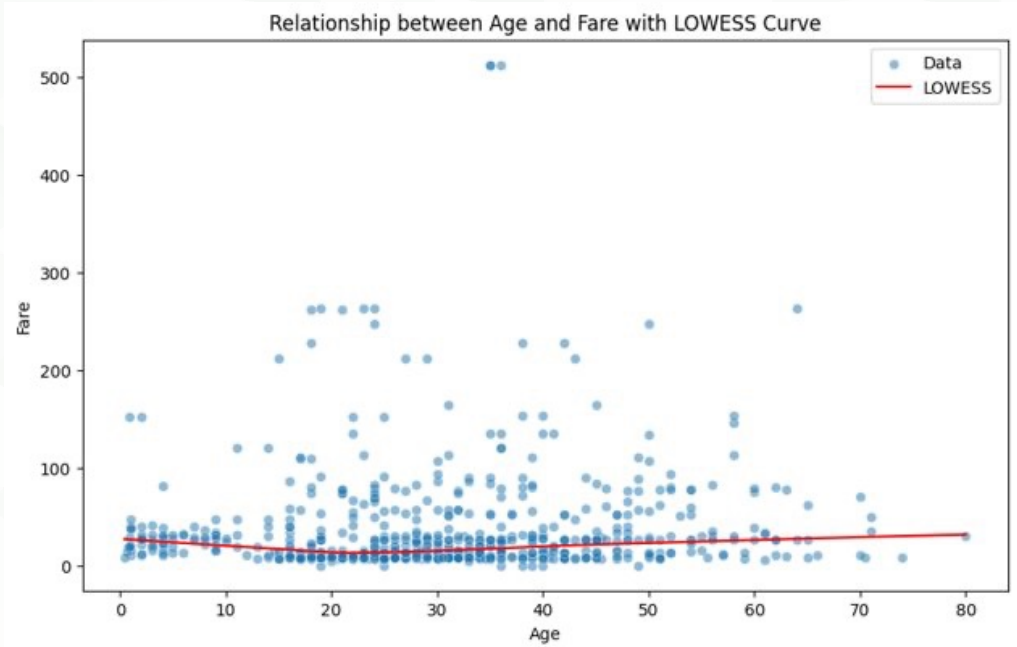
KALBE Digital

Nonlinear Relationship

Linear relationships are straightforward and well-understood, but real-world data often contains nonlinear relationships.

Common Methods and Techniques:

- Scatter Plots
- Polynomial and Spline Regression
- Transformations
- Locally Weighted Scatterplot Smoothing (LOWESS or LOESS)*
- Distance-based Methods
- Decision Trees and Random Forests
- Residual Plots
- Neural Networks
- Generalized Additive Models



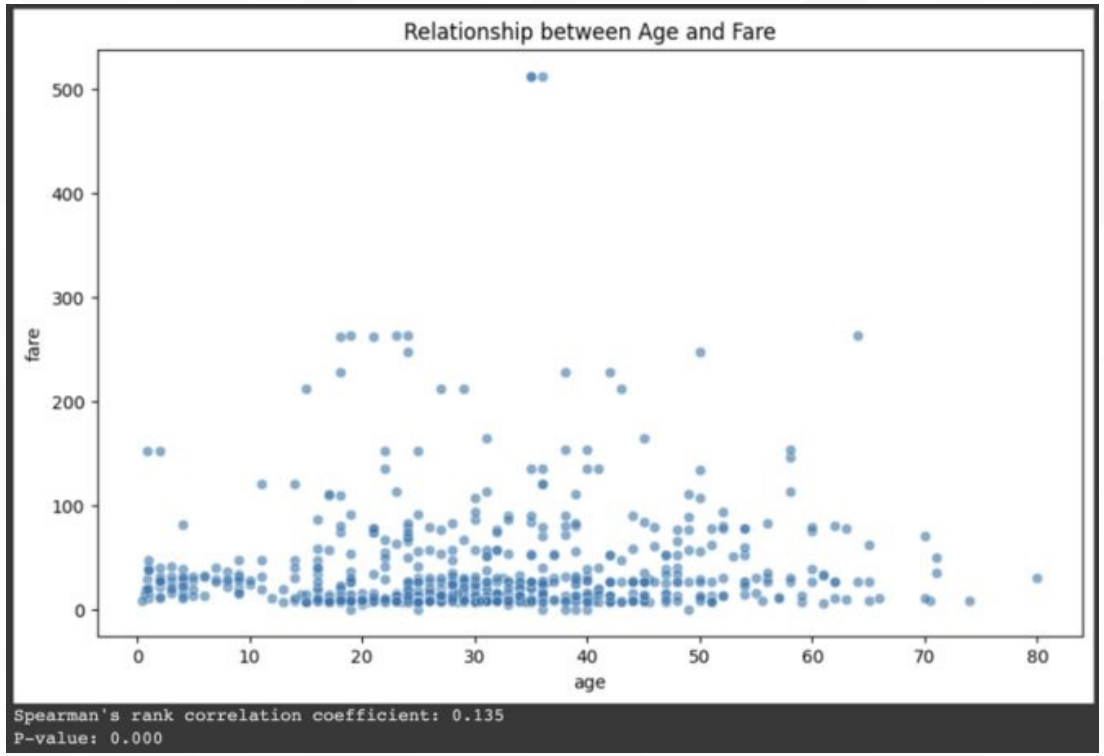
ChatGPT Prompt: Use publicly available dataset, write sample code for **LOWESS** implementation to learn nonlinear relationship and interpret the result using Pandas, SciPy and Seaborn.

KALBE Digital

Spearman's Rank Correlation

Spearman's Rank Correlation is a non-parametric measure of correlation that assesses the monotonic relationship between two variables. It's especially useful when your data isn't normally distributed or has a nonlinear relationship (Wikipedia).

- ρ=1: Perfect positive rank correlation
- ρ=−1: Perfect negative rank correlation
- ρ=0: No rank correlation

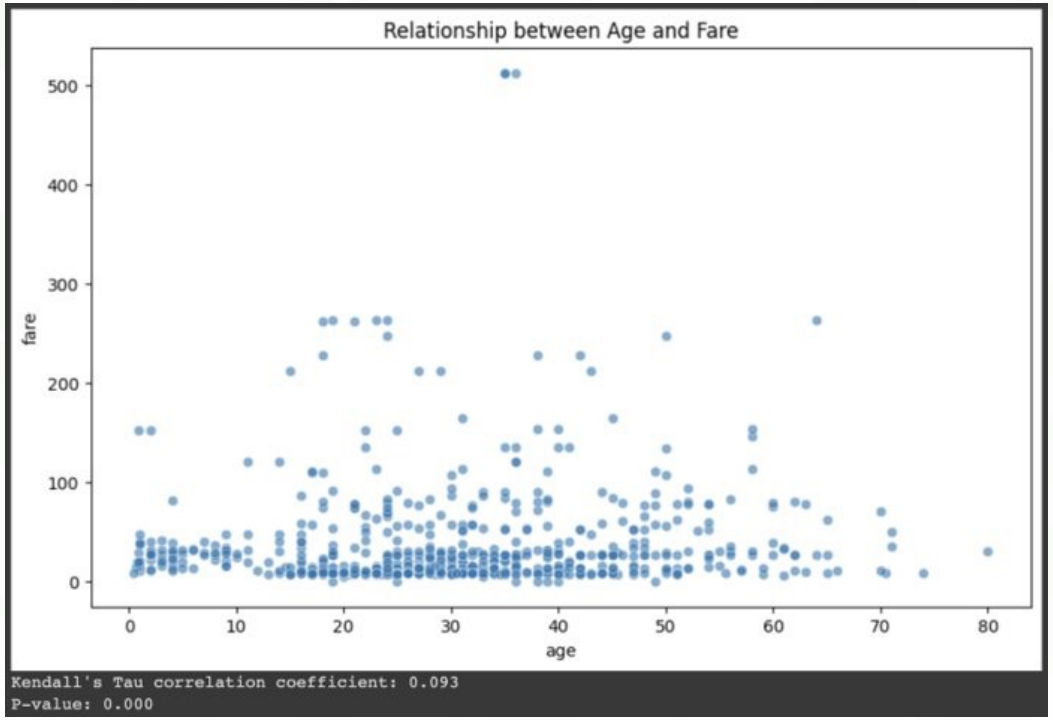


KALBE Digital

Kendall Tau Correlation

The Kendall Tau rank correlation, commonly denoted as τ (tau), is a non-parametric statistic used to measure the ordinal association between two quantities. It assesses the strength and direction of the monotonic relationship between two paired datasets (Wikipedia).

- $\tau = 1$: Perfect agreement
- $\tau = -1$: Perfect disagreement
- $\tau = 0$: No association



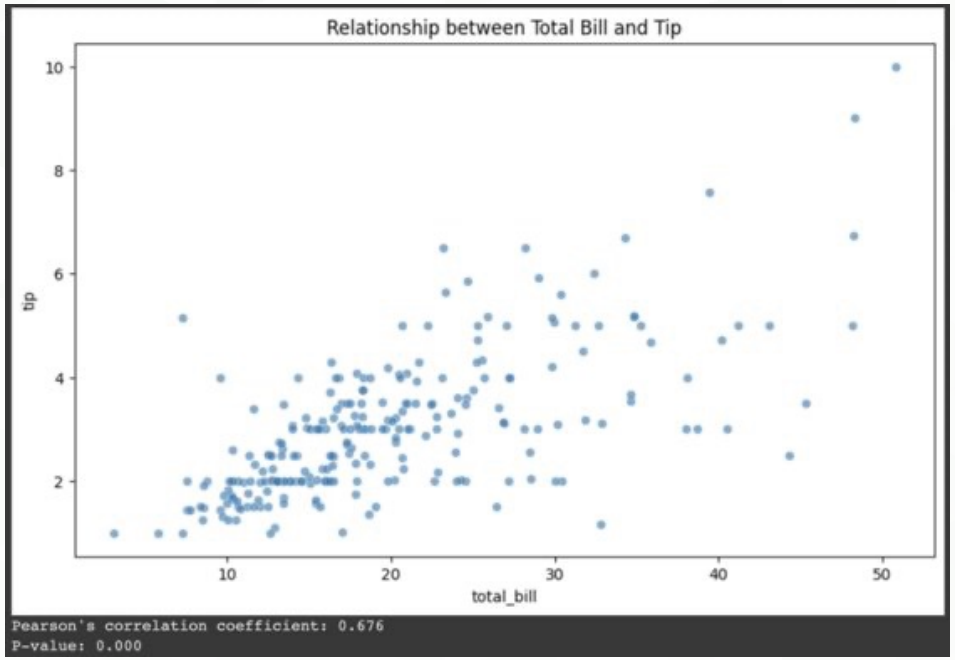
KALBE Digital

Correlation and Causation

Correlation does not imply causation:

- **Correlation** refers to a statistical measure that describes the extent to which two variables change together.
- **Causation** implies that a change in one variable is responsible for a change in another.

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
from scipy.stats import pearsonr

# Load the tips dataset
tips = sns.load_dataset('tips')

plt.figure(figsize=(10, 6))
sns.scatterplot(x='total_bill', y='tip', data=tips, alpha=0.6)
plt.title("Relationship between Total Bill and Tip")
plt.show()

# Calculate correlation
corr_coefficient, p_value = pearsonr(tips['total_bill'], tips['tip'])

print(f"Pearson's correlation coefficient: {corr_coefficient:.3f}")
print(f"P-value: {p_value:.3f}")
```



KALBE Digital

Summary

What We Have Learnt About Relationship Between Variables:

- Scatter Plot
- Correlation Analysis
- Covariance Analysis
- Pearson’s Correlation
- Spearman’s Rank Correlation
- Kendall Tau’s Correlation
- Correlation and Causation

KALBE Digital

Quizz

Quiz 1: Scatter Plots
Which of the following best describes a scatter plot?
A. A plot showing the frequency of a single variable.
B. A plot displaying the relationship between two numerical variables.
C. A plot that shows the distribution of a dataset.
D. A plot to display categories of a single variable.
Correct Answer: B
Quiz 2: Correlation Coefficients 2. Which of the following statements about Pearson's Correlation Coefficient is true?
A. It can have values only between 0 and 1.
B. It assumes a linear relationship between two variables.
C. It is best used for categorical variables.
D. It can determine the causality between two variables.
Correct Answer: B
Quiz 3: Relationship Interpretation 3. If two variables, X and Y, have a strong positive Spearman's Rank Correlation, which of the following can be concluded?
A. Changes in X cause changes in Y.
B. There is a strong linear relationship between X and Y.
C. X and Y move in the same direction, but not necessarily at a constant rate.
D. The relationship between X and Y is purely random.
Correct Answer: C