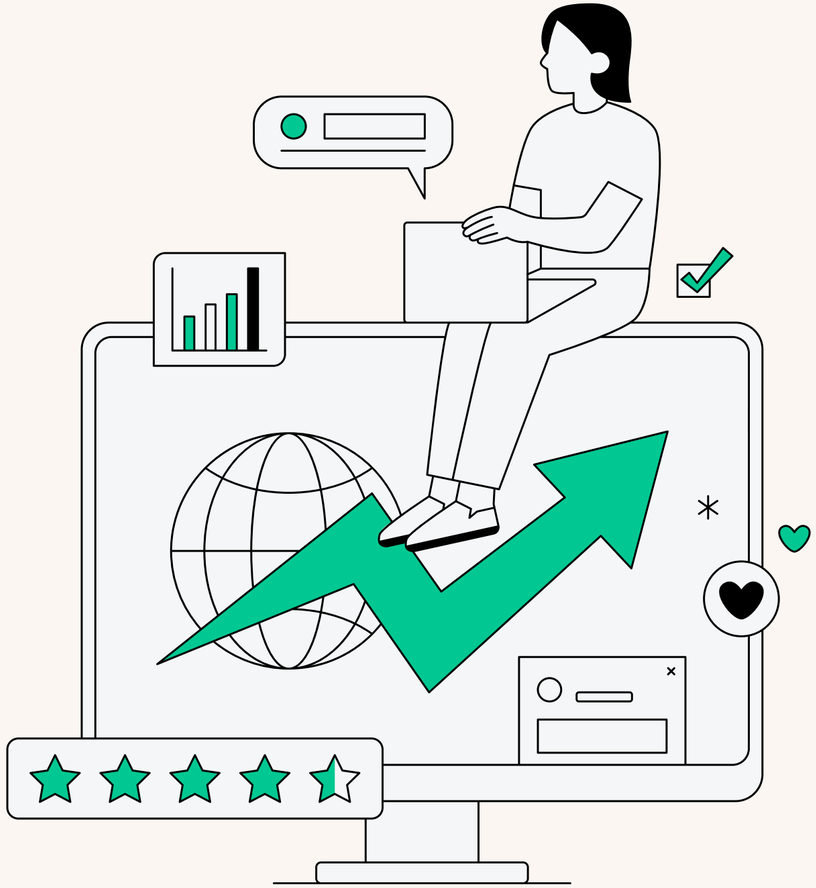
Presented by Siti Alamiah

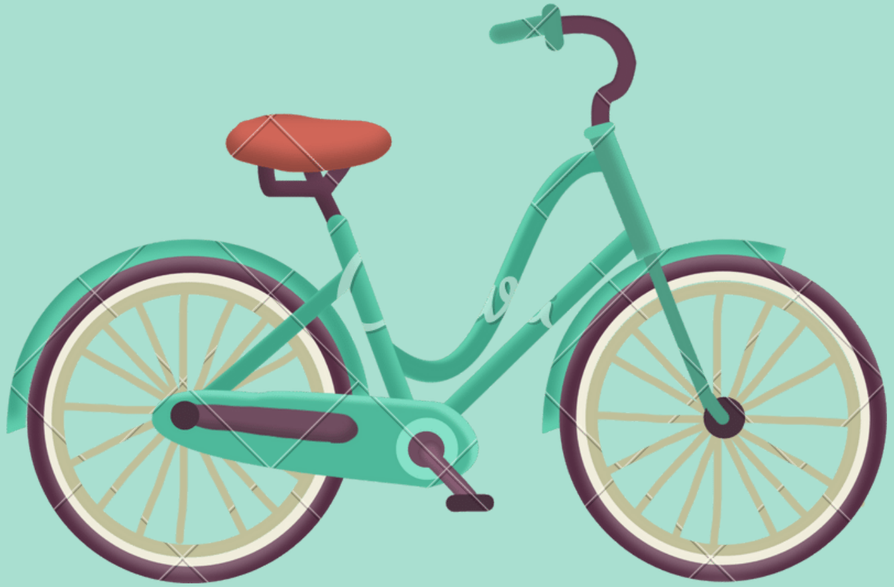
Analysis of Tren Rental Bicycle



sitialamiah@gmsil.com
Linkedin : sitialamiah

Problem of Dataset

The purpose of analyzing this dataset is to find out how citizens used rental bikes from 2011-2012. Currently, there are about more than 500 bike-sharing programs around the world consisting of more than 500 thousand bikes. Currently, there is great interest in these systems due to their important role in traffic, environmental and health issues.



The purpose of analyzing this dataset is to find out how citizens used rental bikes from 2011-2012. Currently, there are about more than 500 bike-sharing programs around the world consisting of more than 500 thousand bikes. Currently, there is great interest in these systems due to their important role in traffic, environmental and health issues.

Data Dictionary

1. dteday: Observation date.
2. season: Season of the year (1: winter, 2: spring, 3: summer, 4: fall).
3. yr: Year of the observation (0: 2011, 1: 2012).
4. month: Month of the observation (1 to 12).
5. hr: Hour of the day (0 to 23).
6. holiday: Holiday indicator (1 if it's a holiday, 0 if not).
7. weekday: Day of the week (0 to 6, representing Monday to Sunday).
8. workingday: Working day indicator (1 if it's a working day, 0 if not).
9. weathersit: Weather condition (1: clear, 2: cloudy, 3: light rain/snow, heavy rain/snow).
10. temp: Temperature in degrees Celsius.
11. atemp: "Feels like" temperature in degrees Celsius.
12. hum: Relative humidity.
13. windspeed: Wind speed.
14. casual: Number of casual (non-registered) bike rentals.
15. register: Number of registered bike rentals.
16. cnt: Total number of bike rentals (casual + registered).

Business Question

- What is the trend of hourly bike usage in Capital's bike sharing system from 2011 to 2012? Are there any particular patterns that can be identified based on the exploratory data analysis?
- What is the relationship between weather conditions (such as temperature, humidity, and wind speed) and hourly bike usage? Does the weather have a significant influence on the number of bicycles rented?
- Can we build a predictive model that can estimate hourly bicycle usage ("cnt") based on weather and seasonal information? What type of model is best suited for this dataset, and how accurate is it in predicting bicycle usage?

Methodology

- **Regression Model**

01. Import the Dataset

```
# Mengimpor Library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

02. Data Wrangling

```
# melakukan gathering dengan menggunakan google drive
from google.colab import drive
drive.mount('/content/drive')
```

```
df = pd.read_csv('/content/drive/MyDrive/idcamp_project/bike dataset/hour.csv')
df.head().style.background_gradient(cmap='Greys')
```

03. Assesing Data

```
df.info()
```

```
# Menampilkan beberapa baris pertama dan terakhir dari dataframe
# untuk melihat format dan struktur data.
df.head()
df.tail()
```

```
# Menyajikan informasi tentang tipe data, jumlah nilai yang tidak null,
# dan penggunaan memori.
df.describe()
```

```
# Menunjukkan jumlah baris dan kolom dalam dataframe.
df.shape
```

```
# Menghitung jumlah nilai null dalam setiap kolom.
df.isnull().sum()
```

```
# Melihat nilai unik pada setiap kolom untuk mendapatkan pemahaman
# tentang kategori atau klasifikasi data.
for column in df.columns:
    print(f"{column}: {df[column].nunique()} unique values")
```

```
# Menampilkan korelasi antar kolom numerik.
df.corr()
```

04. Cleaning Data

```python
# Menjelajahi kerangka data, mengidentifikasi Potensi Kesalahan,
# dan memahami tipe data
pd.set_option('display.max_columns', None)
def data_overview(df, head=5):
    print(" SHAPE ".center(125,'-'))
    print('Rows:{}'.format(df.shape[0]))
    print('Columns:{}'.format(df.shape[1]))
    print(" MISSING VALUES ".center(125,'-'))
    print(df.isnull().sum())
    print(" DUPLICATED VALUES ".center(125,'-'))
    print(df.duplicated().sum())
    print(" HEAD ".center(125,'-'))
    print(df.head(3))
    print(" DATA TYPES ".center(125,'-'))
    print(df.dtypes)

data_overview(df)
```

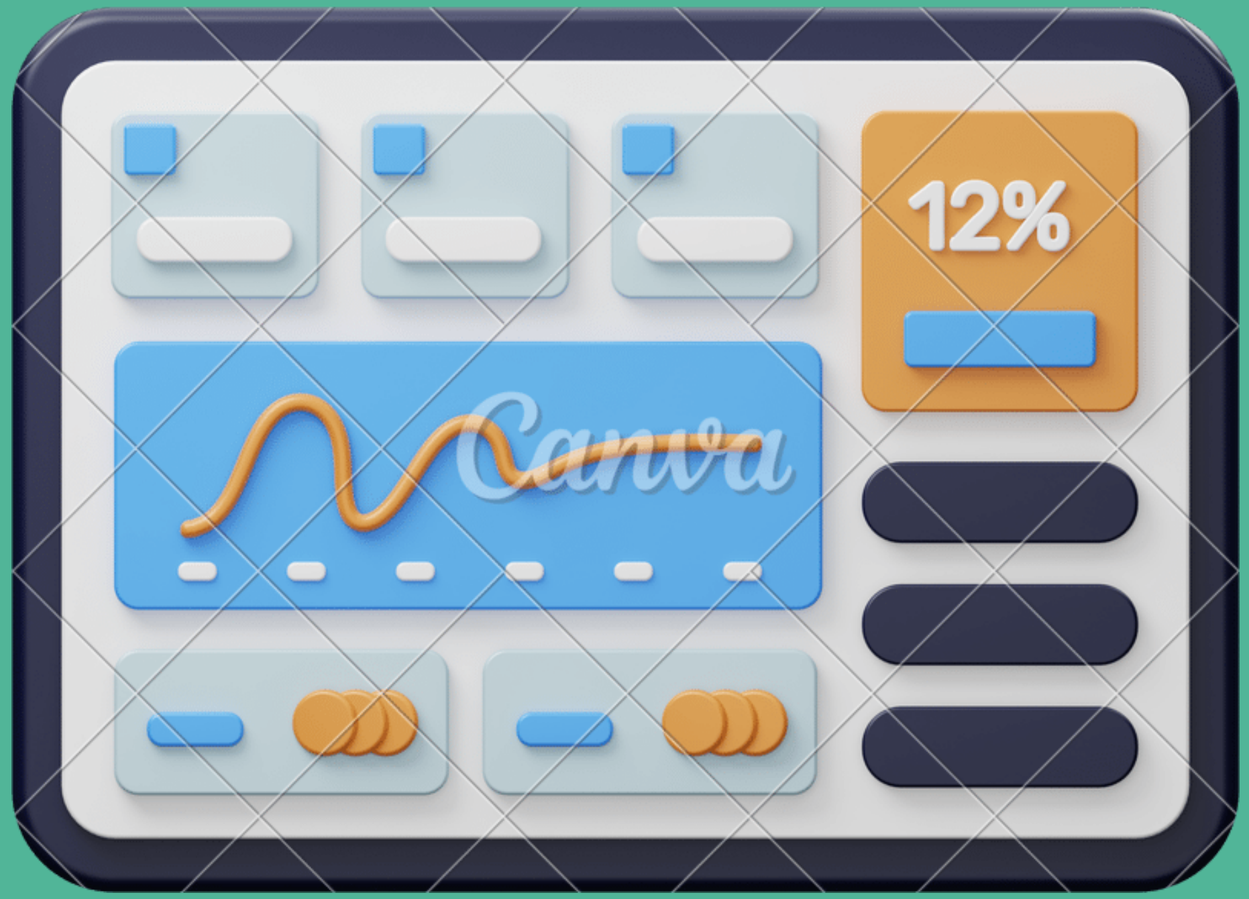
```python
# Memeriksa ouliers dalam varibale target "cnt"
Q1 = df['cnt'].quantile(0.25)
Q3 = df['cnt'].quantile(0.75)
IQR = Q3 - Q1

# menentukan batas untuk outlier
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Mengidentifikasi outlier
outliers = df[(df['cnt'] < lower_bound) | (df['cnt'] > upper_bound)]
outliers.style.background_gradient(cmap='Greys')
```

```python
# menghapus outlier
df = df[(df['cnt'] >= lower_bound) & (df['cnt'] <= upper_bound)]
print("shape after outliers removal :",df.shape)
```

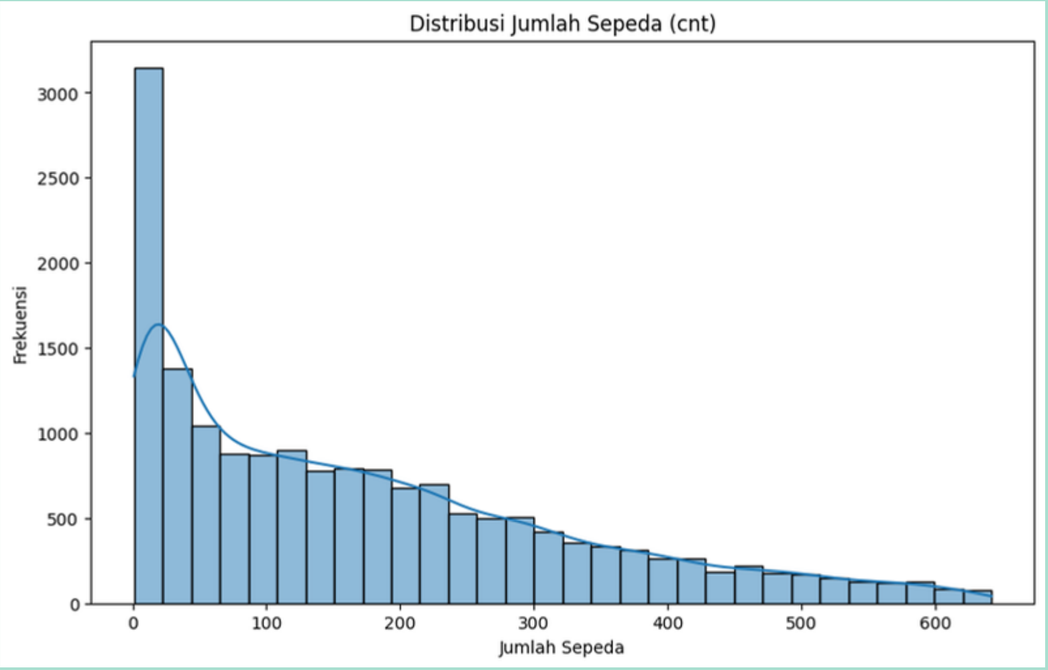
```python
# mengubah variabel diskrit "musim" menjadi tempat sampah
df = pd.get_dummies(df, columns=['season'], dtype=int)
df.head()
```

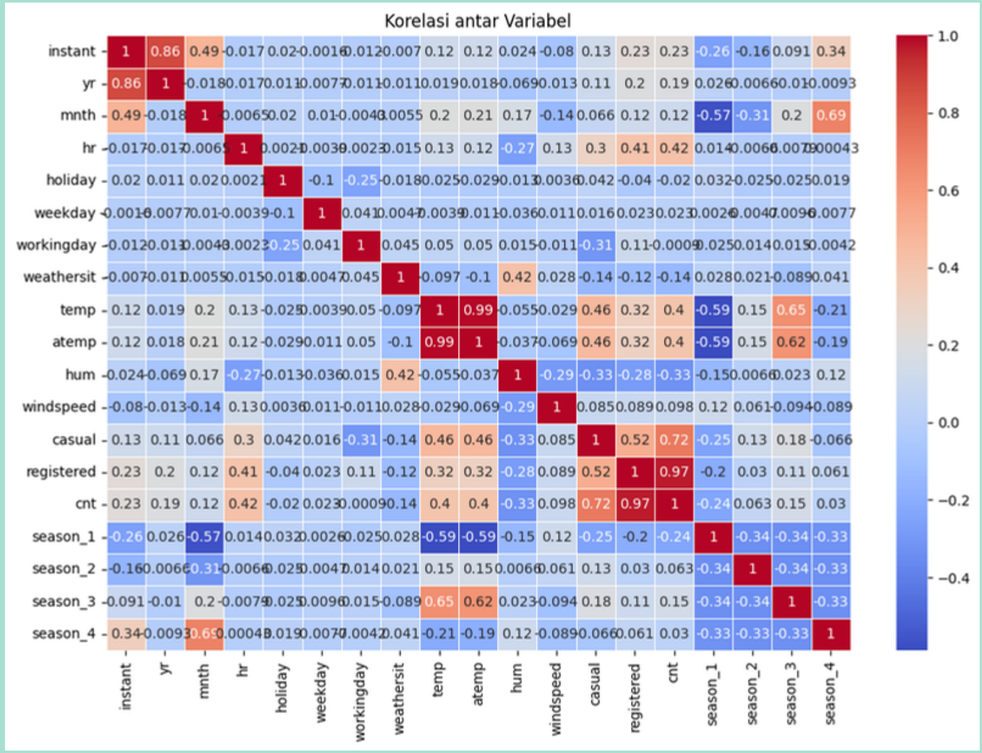


Exploratory Data Analysis

Bicycle quantity distribution

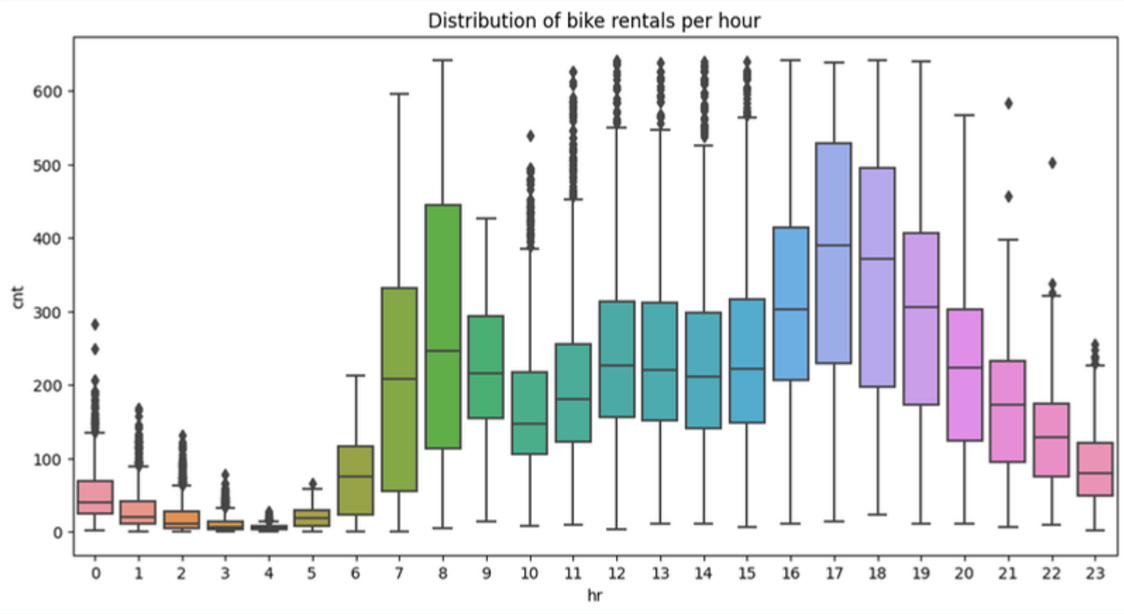
Create and display a histogram for the target variable 'cnt'. The histogram provides a visual representation of the frequency distribution of the number of bicycles in the dataset. With KDE, we also get an approximation of the distribution density curve which smooths the picture of the frequency distribution. Histograms are often used to understand data distribution patterns, such as whether the data tends to be normal, symmetrical, or has other special patterns.





Correlation between variables

It is possible to quickly visualize how closely related the numerical variables in a dataset are using a heatmap. The colors on the heatmap indicate the direction and strength of the correlation between pairs of variables, while the values in the cells provide more information about the level of correlation.

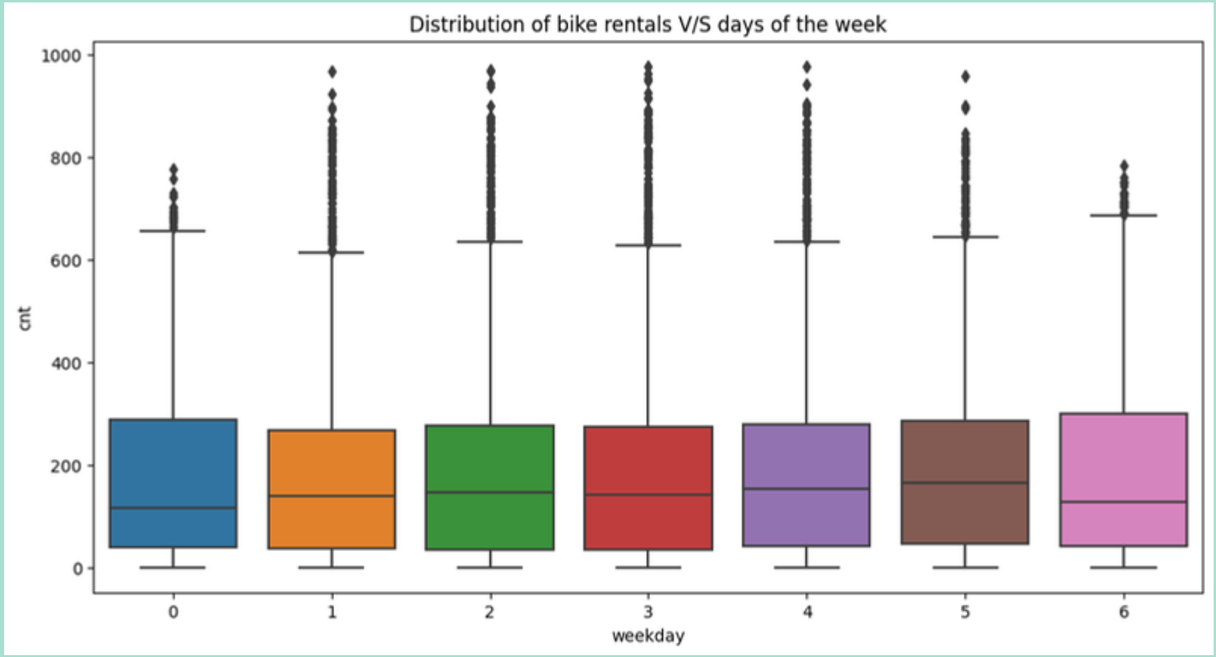


Distribution of rentals per Hour

Using the boxplot, you can see the distribution of descriptive statistics such as median, quartiles, and outliers on the number of bikes rented at each hour. The boxplot provides a visual understanding of how the distribution of 'cnt' varies at each 'hr'.

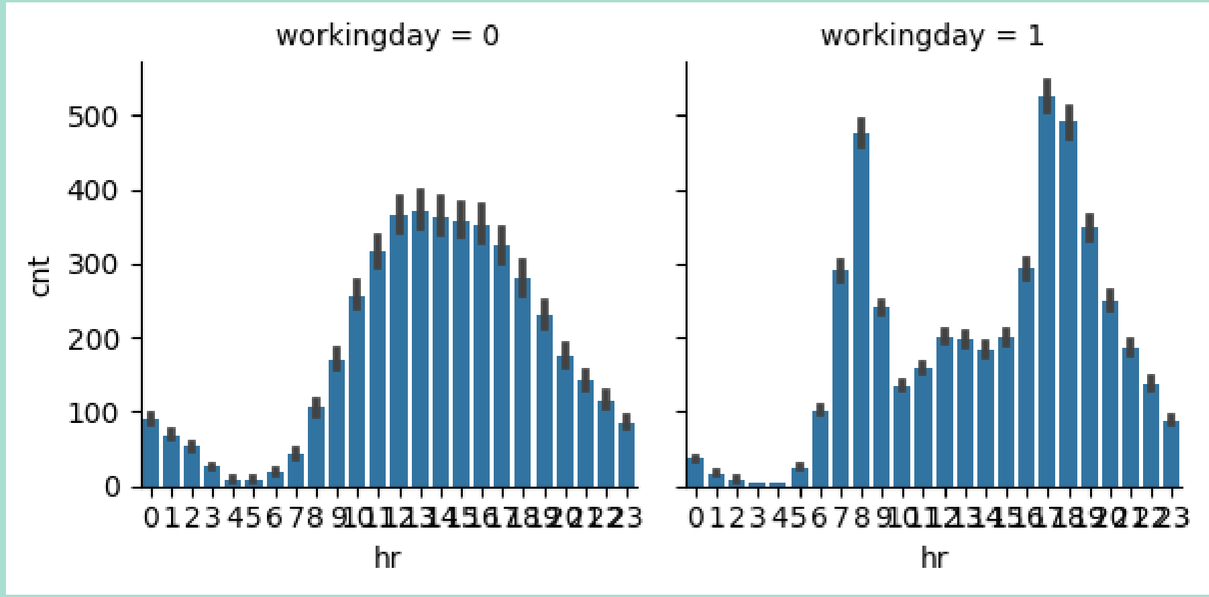
Distribution of bike rentals v/s days of the week

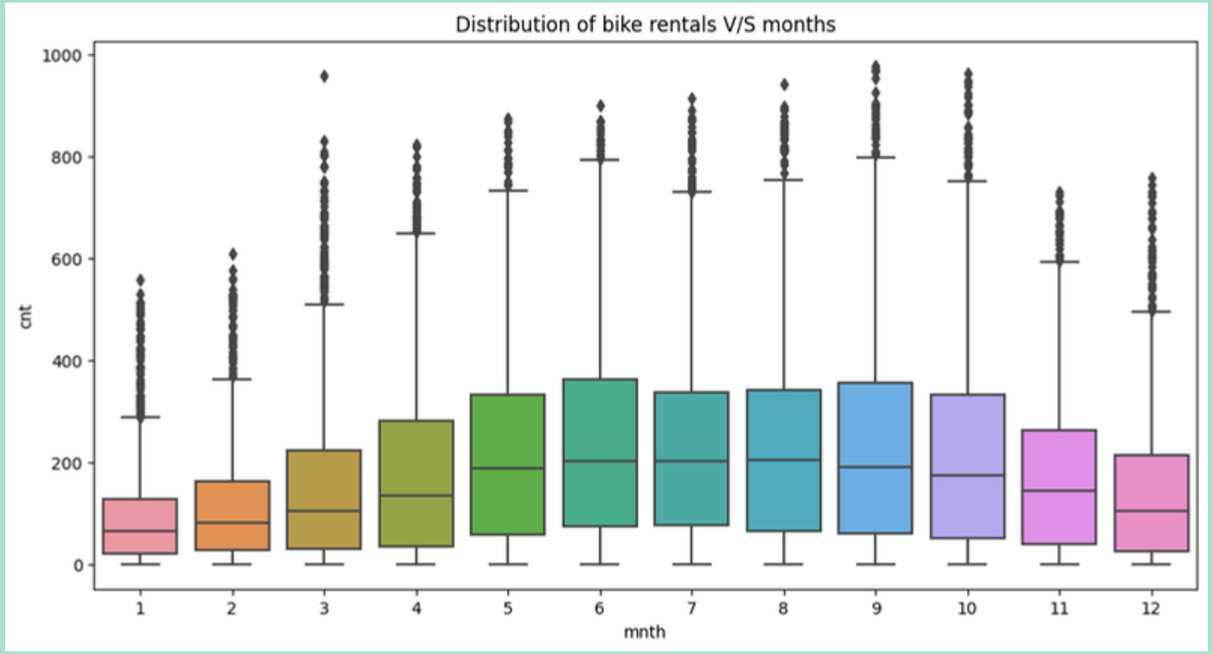
Using the boxplot, you can see the distribution of descriptive statistics such as median, quartiles, and outliers on the number of bikes rented at each hour. The boxplot provides a visual understanding of hoUsing the boxplot, you can see the distribution of descriptive statistics such as median, quartiles, and outliers on the number of bikes rented at each hour. The boxplot provides a visual understanding of how the distribution of 'cnt' varies at each 'hr'.
w the distribution of 'cnt' varies at each 'hr'.



Histogram

Using 'FacetGrid' and 'sns.barplot', you create two barplots comparing the number of rented bikes for each hour, with one barplot for working days and another for non-working days. This provides a visual representation of how the bike rental patterns change for each hour based on the working or non-working day status.

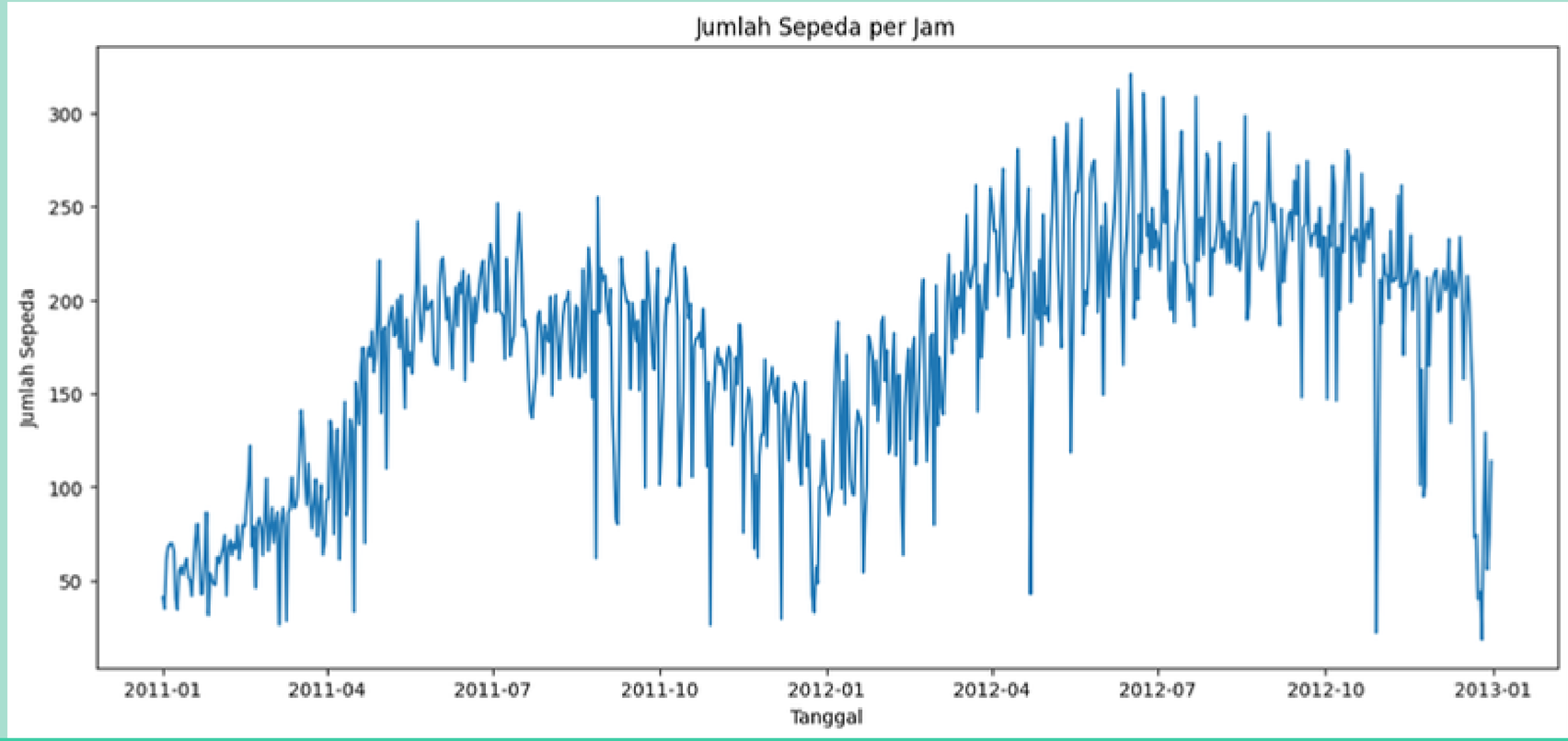




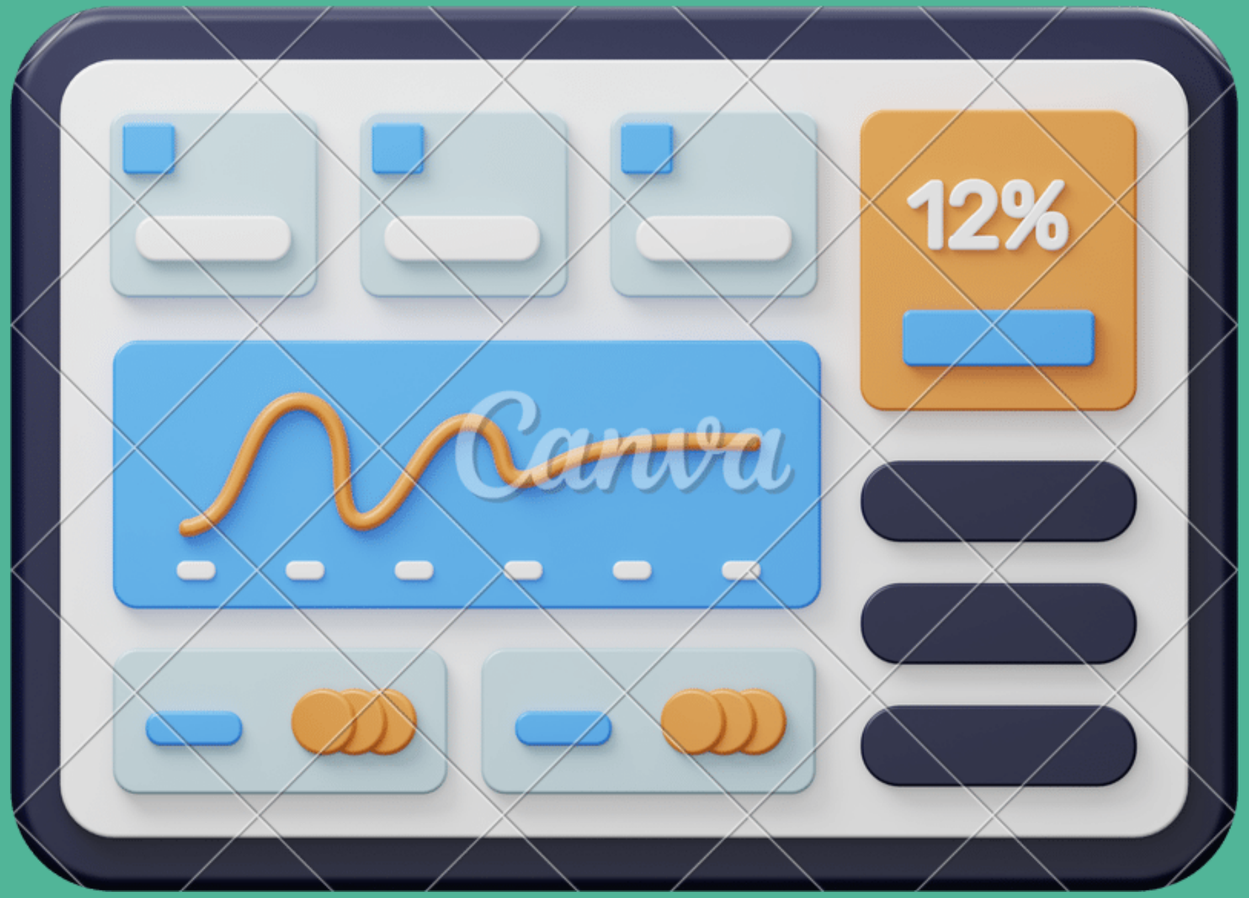
The distribution of bike rentals ('cnt') across different months ('mnth')

This code helps visualize the distribution of bike rentals across different months, showing statistics such as median, quartiles, and potential outliers for each month. It provides insights into how the bike rental patterns vary throughout the year.

Number of bicycles per hour



The provided code performs two tasks: it converts the 'dteday' column to a datetime data type and then creates a line plot to visualize the number of rented bikes ('cnt') per hour over time. This code provides a visual representation of how the number of rented bikes changes over time, specifically per hour, using a line plot. The x-axis represents dates, and the y-axis represents the count of rented bikes.

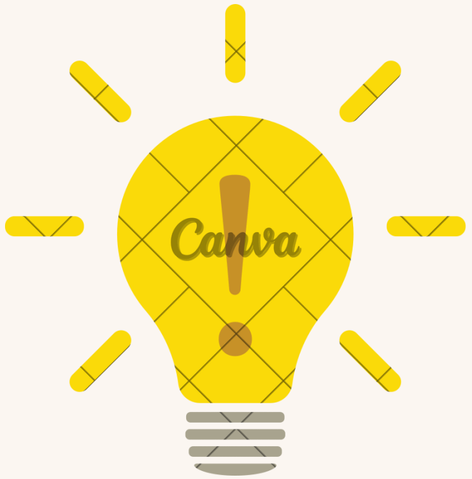


Visualization and Explanatory Analysis

The question is?

1

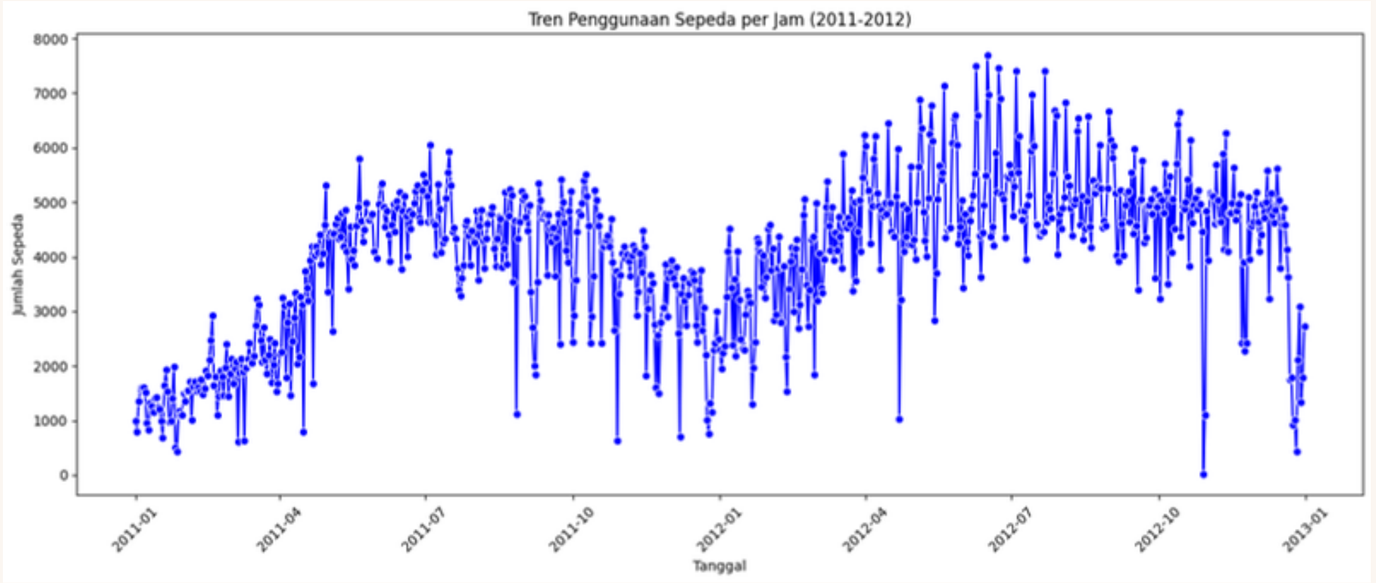
What is the trend of hourly bike usage in Capital's bike sharing system from 2011 to 2012? Are there any particular patterns that can be identified based on the exploratory data analysis?



Answer No 1

```python
# Mengelompokkan data per tanggal dan menghitung jumlah sepeda
daily_counts = df.groupby(df['dteday'].dt.date)['cnt'].sum()
```

```python
# Line plot untuk tren penggunaan sepeda per jam dari tahun 2011 hingga 2012
plt.figure(figsize=(14, 6))
sns.lineplot(x=daily_counts.index, y=daily_counts.values, marker='o', linestyle='-', color='b')
plt.title('Tren Penggunaan Sepeda per Jam (2011-2012)')
plt.xlabel('Tanggal')
plt.ylabel('Jumlah Sepeda')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



- Converting Date Data Type: Using pd.to_datetime to convert the 'dteday' column to datetime data type.
- Grouping Data: Using groupby to group the data by date and then count the number of bikes ('cnt') rented each day.
- Line Plot: Created a line plot using sns.lineplot with x-axis as date and y-axis as number of bikes rented.
- Plot Refinement: Added a title, axis labels, and rotated the date labels to make it easier to read.

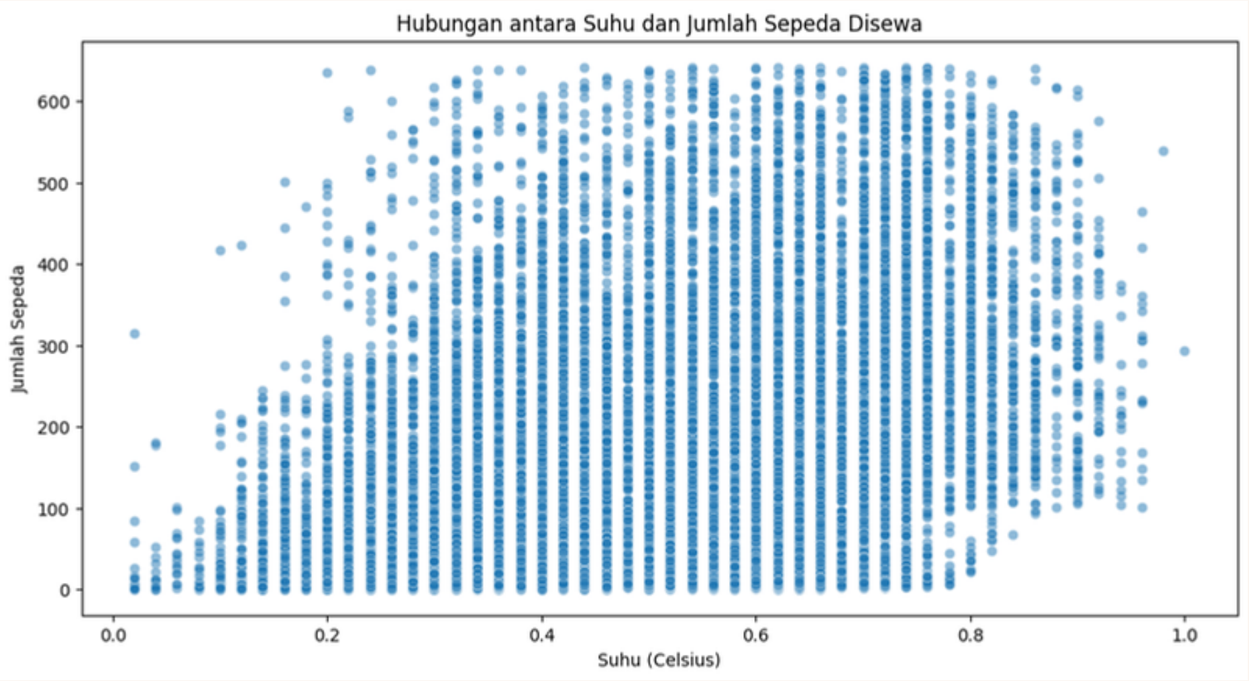
The question is?

2

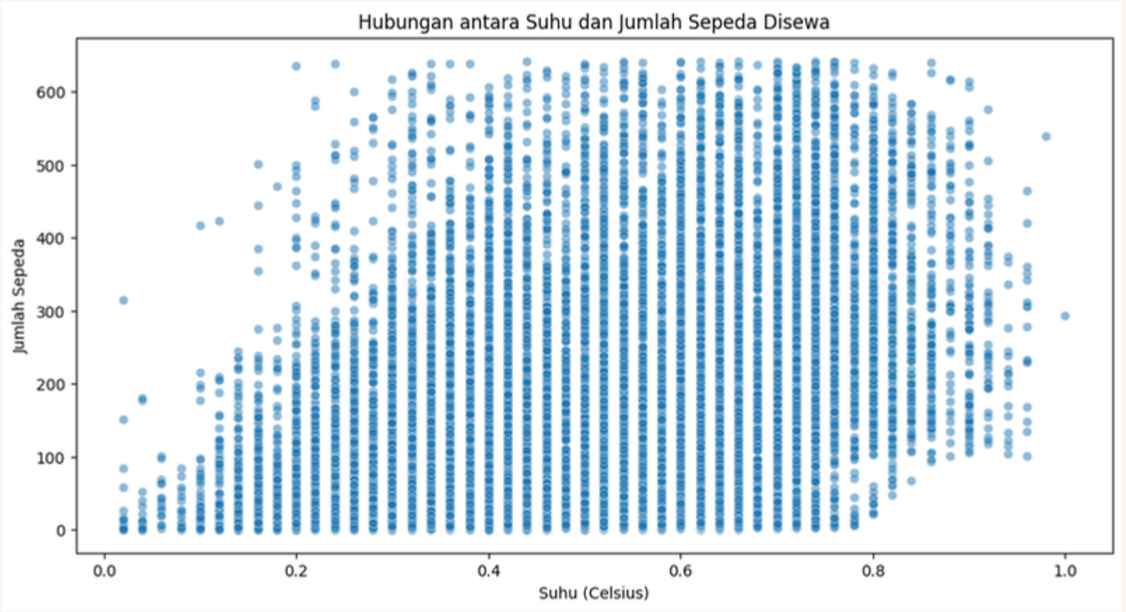
What is the relationship between weather conditions (such as temperature, humidity, and wind speed) and hourly bicycle usage? Does the weather have a significant influence on the number of bicycles rented?

Answer No 2


```python
# Scatter plot untuk suhu dan jumlah sepeda disewa
plt.figure(figsize=(12, 6))
sns.scatterplot(x='temp', y='cnt', data=df, alpha=0.5)
plt.title('Hubungan antara Suhu dan Jumlah Sepeda Disewa')
plt.xlabel('Suhu (Celsius)')
plt.ylabel('Jumlah Sepeda')
plt.show()
```

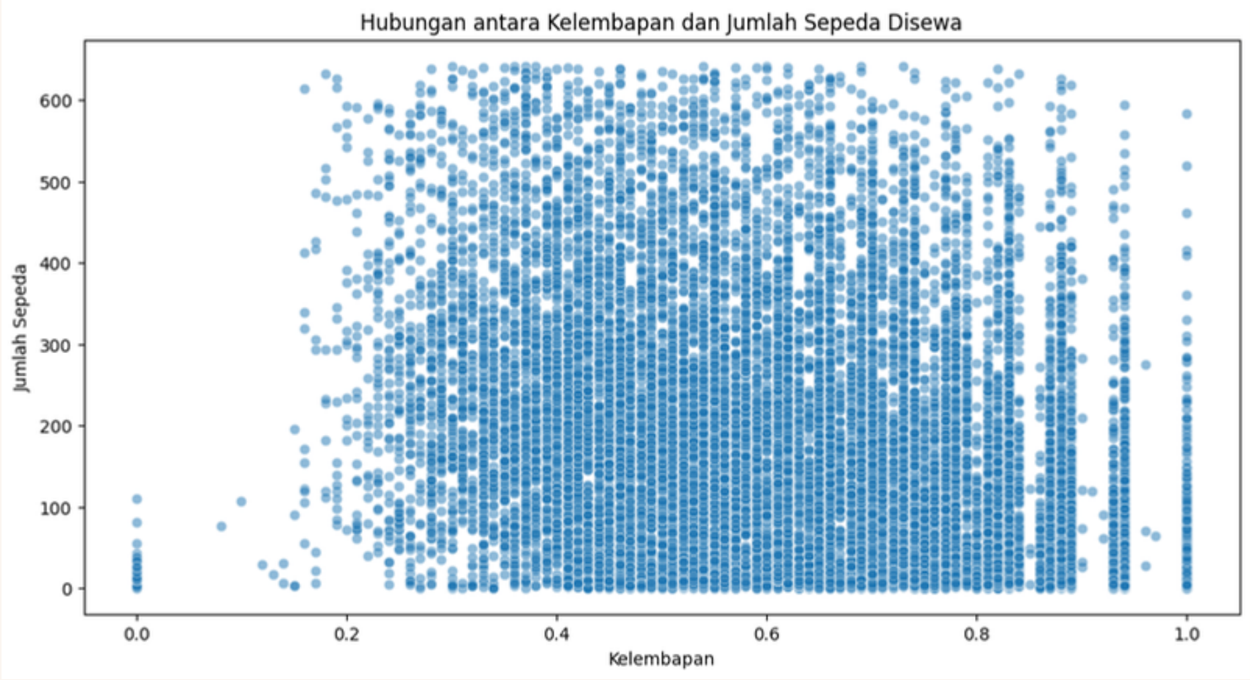



```python
# Scatter plot untuk suhu dan jumlah sepeda disewa
plt.figure(figsize=(12, 6))
sns.scatterplot(x='temp', y='cnt', data=df, alpha=0.5)
plt.title('Hubungan antara Suhu dan Jumlah Sepeda Disewa')
plt.xlabel('Suhu (Celsius)')
plt.ylabel('Jumlah Sepeda')
plt.show()
```

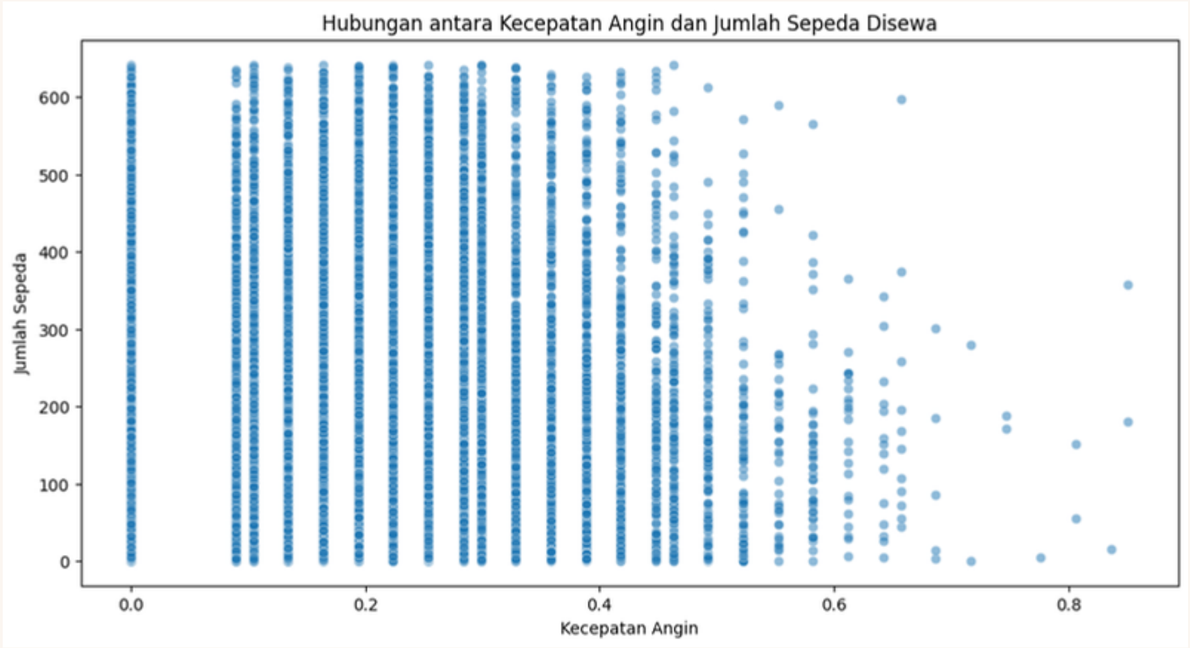



Answer No 2

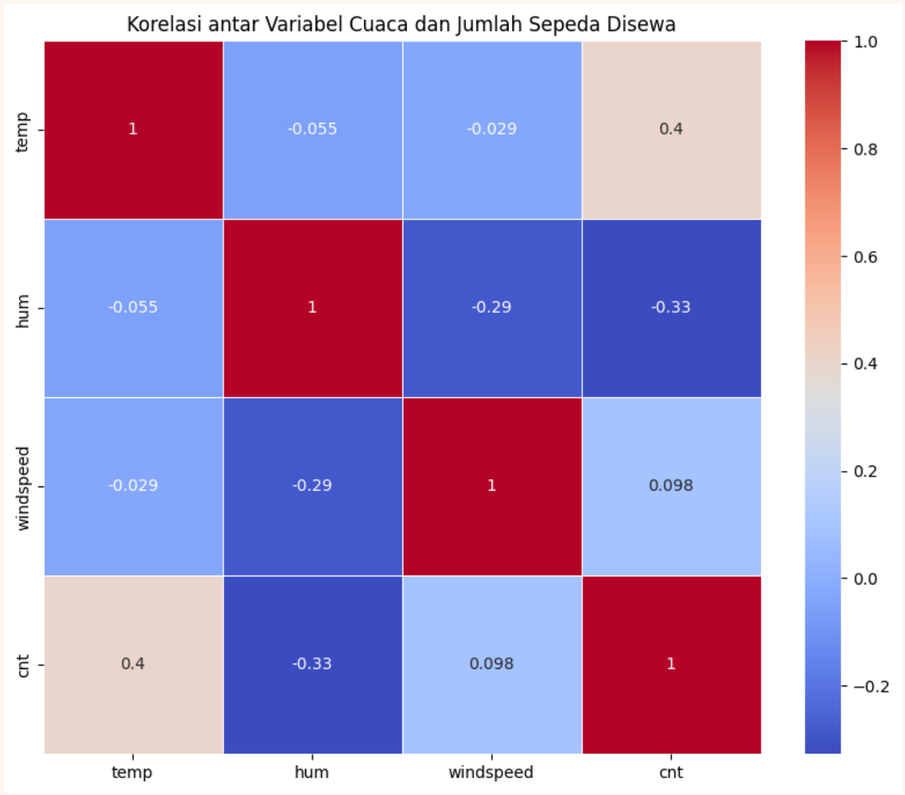
```python
# Scatter plot untuk kelembapan dan jumlah sepeda disewa
plt.figure(figsize=(12, 6))
sns.scatterplot(x='hum', y='cnt', data=df, alpha=0.5)
plt.title('Hubungan antara Kelembapan dan Jumlah Sepeda Disewa')
plt.xlabel('Kelembapan')
plt.ylabel('Jumlah Sepeda')
plt.show()
```




```python
# Scatter plot untuk kecepatan angin dan jumlah sepeda disewa
plt.figure(figsize=(12, 6))
sns.scatterplot(x='windspeed', y='cnt', data=df, alpha=0.5)
plt.title('Hubungan antara Kecepatan Angin dan Jumlah Sepeda Disewa')
plt.xlabel('Kecepatan Angin')
plt.ylabel('Jumlah Sepeda')
plt.show()
```



Answer No 2



- * Scatter Plot: Creates a scatter plot for each weather variable (temperature, humidity, and wind speed) against the number of bicycles rented (cnt).
- * Correlation Heatmap: Creates a heatmap to see the correlation between numerical variables. This gives an idea of the extent to which the weather variables are correlated with the number of bicycles rented.
- With this visualization, it can be assessed that the relationship between the weather variable and the number of bicycles rented with the Scatter plot provides a visual overview while the correlation heatmap provides more detailed information of the correlation between numerical variables. If there is a significant correlation, this may indicate that weather conditions have an influence on the number of bicycles rented.

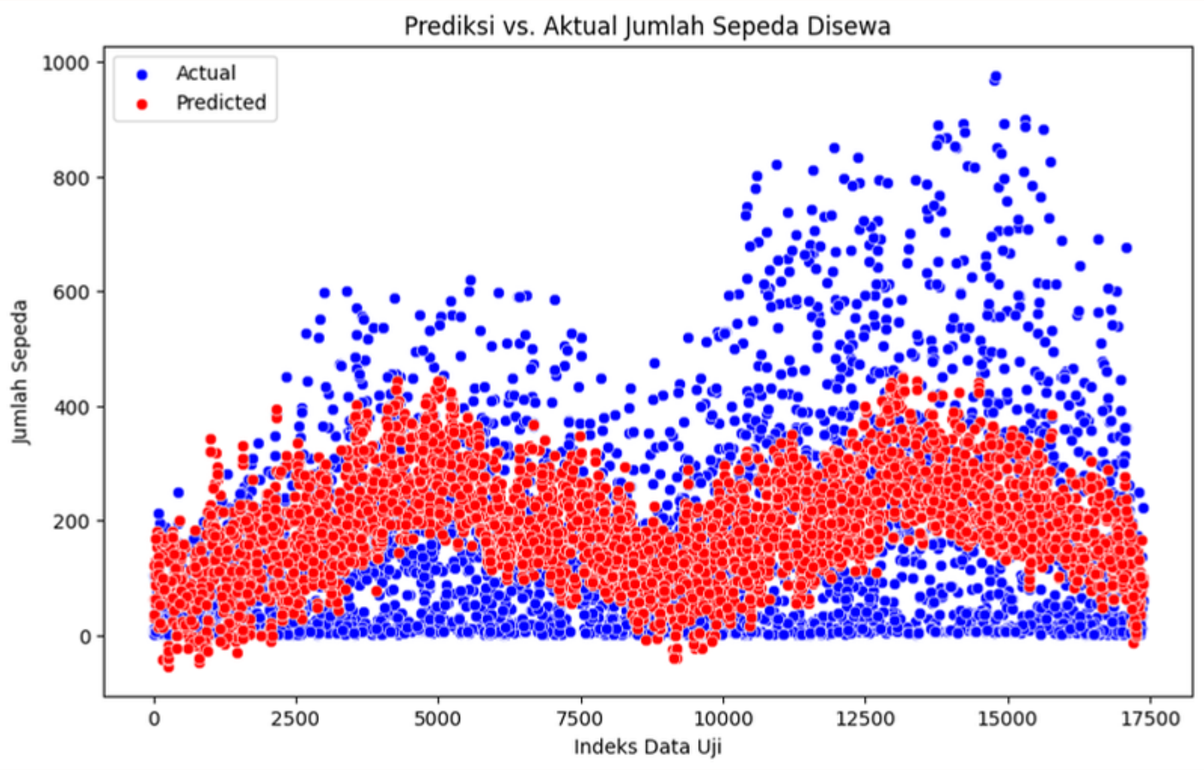
```
# Korelasi antar variabel numerik
correlation_matrix = df[['temp', 'hum', 'windspeed', 'cnt']].corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Korelasi antar Variabel Cuaca dan Jumlah Sepeda Disewa')
plt.show()
```

The question is?

3

Can we build a prediction model that can estimate hourly bicycle usage ("cnt") based on weather and seasonal information? What type of model is best suited for this dataset, and how accurate is it in predicting bicycle usage?

Answer No 3




```python
# Memilih fitur-fitur yang akan digunakan untuk prediksi
features = ['temp', 'hum', 'windspeed', 'season']

# Memisahkan variabel independen (X) dan dependen (y)
X = df[features]
y = df['cnt']

# Membagi data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Membuat model Regresi Linear
model = LinearRegression()
model.fit(X_train, y_train)

# Memprediksi data uji
y_pred = model.predict(X_test)

# Mengukur akurasi model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R-squared (R2): {r2}')

# Visualisasi prediksi vs. aktual
plt.figure(figsize=(10, 6))
sns.scatterplot(x=y_test.index, y=y_test, label='Actual', color='blue')
sns.scatterplot(x=y_test.index, y=y_pred, label='Predicted', color='red')
plt.title('Prediksi vs. Aktual Jumlah Sepeda Disewa')
plt.xlabel('Indeks Data Uji')
plt.ylabel('Jumlah Sepeda')
plt.legend()
plt.show()
```

- Memilih Fitur: Memilih fitur-fitur yang akan digunakan untuk prediksi, dalam hal ini suhu (temp), kelembapan (hum), kecepatan angin (windspeed), dan musim (season).
- Pembagian Data: Membagi data menjadi data latih dan data uji menggunakan train_test_split.
- Membuat Model: Membuat model Regresi Linear menggunakan LinearRegression() dari scikit-learn.
- Memprediksi dan Mengukur Akurasi: Memprediksi data uji dan mengukur akurasi model menggunakan Mean Squared Error (MSE) dan R-squared (R2).
- Visualisasi Prediksi vs. Aktual: Menampilkan scatter plot untuk membandingkan nilai aktual dan nilai prediksi.

Conclusion

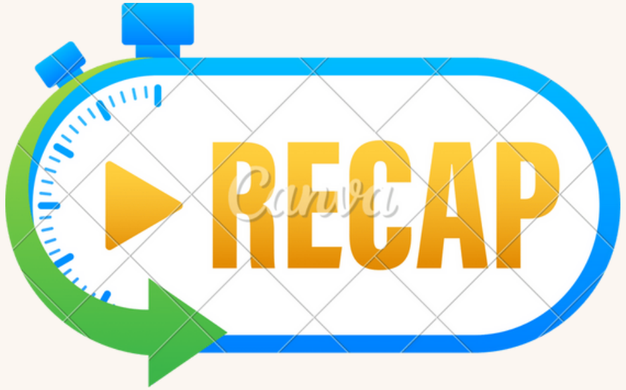
Bicycle Usage Trend (2011-2012)

Trend analysis of hourly bicycle usage from 2011 to 2012 was conducted by converting the date data to datetime data type, grouping the data by date, and constructing a line plot. The line plot provides a visual representation of how bicycle usage changed over the time period.

Relationship with Weather Conditions

To evaluate the relationship between weather conditions (temperature, humidity, and wind speed) and hourly bicycle usage, a scatter plot and correlation heatmap were analyzed. The scatter plot shows the distribution of data for each weather variable against the number of bicycles rented, while the correlation heatmap provides information on the extent to which weather variables are correlated with the number of bicycles rented.

1. Building a Prediction Model: Build a prediction model to estimate bicycle usage per hour ("cnt") based on weather and seasonal information. In the example, a Linear Regression model is used, but other regression models can also be used.
2. Features Used: Selecting features such as temperature, humidity, wind speed, and season as predictors in the model.
3. Evaluation of Model Accuracy: Measuring model accuracy using Mean Squared Error (MSE) and R-squared (R2). The results of this evaluation provide an understanding of how well the model can predict bicycle usage.
4. Predicted vs. Actual Visualization: Presents the model's predicted results compared to the actual values using a scatter plot. This visualization helps in seeing the extent to which the model can follow actual bicycle usage patterns.



The Summary

The reason for building the predictive model is that constructing a prediction model in the context of an hourly bike usage dataset can have several fundamental reasons:

1. Demand Forecasting: A predictive model allows for forecasting future bike demand based on weather conditions and seasons. This can help bike-sharing service providers anticipate surges or declines in demand, which can impact bike supply and distribution.
2. Resource Planning: By understanding bike usage patterns, service providers can plan their resources more efficiently. This includes allocating bikes in more strategic locations, managing inventory, and scheduling maintenance.
3. Service Optimization: Predictive models can assist in optimizing bike-sharing services by adjusting bike placement and quantities at various rental points based on demand estimates.
4. Improved User Experience: By predicting demand, bike-sharing services can enhance the user experience. A better understanding of when and where bikes are needed can improve bike availability and customer satisfaction.
5. Operational Efficiency: Predicting bike usage allows service providers to enhance their operational efficiency. This includes scheduling maintenance, managing bike distribution, and inventory management.
6. Data-Driven Decision-Making: Predictive models provide a foundation for data-driven decision-making. Decisions based on accurate and detailed information can optimize the performance and outcomes of bike-sharing services.
7. Understanding Influencing Factors: Predictive models also help understand how certain factors, such as weather conditions and seasons, influence bike usage. This can serve as a basis for marketing strategies or service adjustments. By building a predictive model, bike-sharing services can improve operational efficiency, enhance service quality, and respond better to changes in demand and environmental conditions.

Presented by Siti Alamiah

Thank you very much!

sitialamiah@gmsil.com
Linkedin : sitialamiah