# Proyek Analisis Data: Bike Sharing Dataset


- **Nama:** Sylvie Lee
- **Email:** sylvielee273@gmail.com
- **ID Dicoding:** sylvielee


## Menentukan Pertanyaan Bisnis

1. Bagaimana pengaruh hari kerja terhadap rata-rata jumlah pengguna bike sharing harian?
2. Bagaimana pengaruh kondisi cuaca terhadap rata-rata jumlah pengguna bike sharing harian?
3. Bagaimana hubungan antara musim dan rata-rata jumlah pengguna bike sharing harian?
4. Bagaimana pola peningkatan dan penurunan rata-rata jumlah pengguna bike sharing harian dalam kurun waktu tertentu?

## Import Semua Packages/Library yang Digunakan

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from statistics import mean, median, mode, StatisticsError
import math
from scipy.stats import norm
import seaborn as sns
from babel.numbers import format_currency
sns.set(style='dark')
```

# Data Wrangling

## Gathering Data

### 1. Importing The Dataset

Note: I only use day.csv for analysis data

```
df1 = pd.read_csv("/content/drive/MyDrive/Bike-sharing-dataset.zip (Unzipped F
```

```
df1.head()
```

| | instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersi |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 6 | 0 | |
| **1** | 2 | 2011-01-02 | 1 | 0 | 1 | 0 | 0 | 0 | |
| **2** | 3 | 2011-01-03 | 1 | 0 | 1 | 0 | 1 | 1 | |
| **3** | 4 | 2011-01-04 | 1 | 0 | 1 | 0 | 2 | 1 | |
| **4** | 5 | 2011-01-05 | 1 | 0 | 1 | 0 | 3 | 1 | |

### 2. Dropping Unused Column

```
df1 = df1.drop(["temp", "atemp", "hum", "windspeed"], axis=1)
df1.head()
```

| | instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersi |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 6 | 0 | |
| **1** | 2 | 2011-01-02 | 1 | 0 | 1 | 0 | 0 | 0 | |
| **2** | 3 | 2011-01-03 | 1 | 0 | 1 | 0 | 1 | 1 | |
| | | 2011- | | | | | | | |

## 3. Rename and Convert for better understanding

```
#Rename the columns
df1.rename(columns={'instant':'user_id','dteday':'datetime','yr':'year','mnth'
                    'hum':'humidity','cnt':'total_count'},inplace=True)
df1.head()
```

| | user_id | datetime | season | year | month | holiday | weekday | workingday | weat |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 6 | 0 | |
| **1** | 2 | 2011-01-02 | 1 | 0 | 1 | 0 | 0 | 0 | |
| **2** | 3 | 2011-01-03 | 1 | 0 | 1 | 0 | 1 | 1 | |
| **3** | 4 | 2011-01-04 | 1 | 0 | 1 | 0 | 2 | 1 | |
| **4** | 5 | 2011-01-05 | 1 | 0 | 1 | 0 | 3 | 1 | |

```
# Data convertion for understanding

# Converting 'season' : 1:Winter, 2:Spring, 3:Summer, 4:Fall
df1.season.replace((1,2,3,4), ('Winter','Spring','Summer','Fall'), inplace=Tru

# Converting 'year' : 0:2011, 1:2012
df1.year.replace((0,1), (2011,2012), inplace=True)
```

```
# Converting 'month' :  1:Jan, 2:Feb, 3:Mar, 4:Apr, 5:May, 6:Jun, 7:Jul, 8:Aug
df1.month.replace((1,2,3,4,5,6,7,8,9,10,11,12),('Jan','Feb','Mar','Apr','May',

# Converting 'weathersit' : 1:Clear, 2:Misty, 3:Light_RainSnow 4:Heavy_RainSno
df1.weather_condition.replace((1,2,3,4), ('Clear','Misty','Light_RainSnow','He

# Comverting'weekday' : 0:Sun, 1:Mon, 2:Tue, 3:Wed, 4:Thu, 5:Fri, 6:Sat
df1.weekday.replace((0,1,2,3,4,5,6), ('Sunday','Monday','Tuesday','Wednesday',

# Converting 'workingday' : 0:No, 1:Yes
df1.workingday.replace((0,1), ('No', 'Yes'), inplace=True)

# Converting 'holiday' : 0:No, 1:Yes
df1.holiday.replace((0,1), ('No', 'Yes'), inplace=True)

df1.head(100)
```

| | user_id | datetime | season | year | month | holiday | weekday | workingday | w |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2011-01-01 | Winter | 2011 | Jan | No | Saturday | No | |
| **1** | 2 | 2011-01-02 | Winter | 2011 | Jan | No | Sunday | No | |
| **2** | 3 | 2011-01-03 | Winter | 2011 | Jan | No | Monday | Yes | |
| **3** | 4 | 2011-01-04 | Winter | 2011 | Jan | No | Tuesday | Yes | |
| **4** | 5 | 2011-01-05 | Winter | 2011 | Jan | No | Wednesday | Yes | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **95** | 96 | 2011-04-06 | Spring | 2011 | Apr | No | Wednesday | Yes | |
| **96** | 97 | 2011-04-07 | Spring | 2011 | Apr | No | Thursday | Yes | |
| **97** | 98 | 2011-04-08 | Spring | 2011 | Apr | No | Friday | Yes | |
| **98** | 99 | 2011-04-09 | Spring | 2011 | Apr | No | Saturday | No | |
| **99** | 100 | 2011-04-10 | Spring | 2011 | Apr | No | Sunday | No | |

100 rows × 12 columns

## Assessing Data

## 1. Missing Value

```
df1.isnull().sum()
```

```
user_id             0
datetime            0
season              0
year                0
month               0
holiday             0
weekday             0
workingday          0
weather_condition    0
casual              0
registered          0
total_count         0
dtype: int64
```

## 2. Duplicated Data

```
df1.duplicated().sum()
```

```
0
```

## 3. Data info

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 731 entries, 0 to 730
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   user_id            731 non-null    int64 
 1   datetime           731 non-null    object
 2   season             731 non-null    object
 3   year               731 non-null    int64 
 4   month              731 non-null    object
 5   holiday            731 non-null    object
 6   weekday            731 non-null    object
 7   workingday         731 non-null    object
 8   weather_condition  731 non-null    object
 9   casual             731 non-null    int64 
 10  registered         731 non-null    int64 
 11  total_count        731 non-null    int64 
dtypes: int64(5), object(7)
memory usage: 68.7+ KB
```

## 4. Cleaning Data

### 1. Convert Datatype

```
df1['user_id'] = df1.user_id.astype('category')
df1['datetime'] = pd.to_datetime(df1['datetime'])
df1['season'] = df1.season.astype('category')
df1['month'] = df1.month.astype('category')
df1['holiday'] = df1.holiday.astype('category')
df1['weekday'] = df1.weekday.astype('category')
df1['workingday'] = df1.workingday.astype('category')
df1['weather_condition'] = df1.weather_condition.astype('category')
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 731 entries, 0 to 730
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype   
---  ------             --------------  -----   
 0   user_id            731 non-null    category
 1   datetime           731 non-null    datetime64[ns]
 2   season             731 non-null    category
 3   year               731 non-null    int64   
 4   month              731 non-null    category
 5   holiday            731 non-null    category
 6   weekday            731 non-null    category
 7   workingday         731 non-null    category
 8   weather_condition  731 non-null    category
 9   casual             731 non-null    int64   
 10  registered         731 non-null    int64   
 11  total_count        731 non-null    int64   
dtypes: category(7), datetime64[ns](1), int64(4)
memory usage: 57.6 KB
```

```
df1.to_csv("all_data.csv", index=False)
```

## Exploratory Data Analysis (EDA)

### Explore Data Statistic

```
df1.describe()
```

| | year | casual | registered | total_count |
|---|---|---|---|---|
| **count** | 731.000000 | 731.000000 | 731.000000 | 731.000000 |
| **mean** | 2011.500684 | 848.176471 | 3656.172367 | 4504.348837 |
| **std** | 0.500342 | 686.622488 | 1560.256377 | 1937.211452 |
| **min** | 2011.000000 | 2.000000 | 20.000000 | 22.000000 |
| **25%** | 2011.000000 | 315.500000 | 2497.000000 | 3152.000000 |
| **50%** | 2012.000000 | 713.000000 | 3662.000000 | 4548.000000 |
| **75%** | 2012.000000 | 1096.000000 | 4776.500000 | 5956.000000 |
| **max** | 2012.000000 | 3410.000000 | 6946.000000 | 8714.000000 |

```
df1.describe(include='all')
```

```
<ipython-input-37-f89e5509e305>:1: FutureWarning: Treating datetime data as
  df1.describe(include='all')
```

| | user_id | datetime | season | year | month | holiday | weekday | worki |
|---|---|---|---|---|---|---|---|---|
| **count** | 731.0 | 731 | 731 | 731.000000 | 731 | 731 | 731 | |
| **unique** | 731.0 | 731 | 4 | NaN | 12 | 2 | 7 | |
| **top** | 1.0 | 2011-01-01 00:00:00 | Summer | NaN | Aug | No | Monday | |
| **freq** | 1.0 | 1 | 188 | NaN | 62 | 710 | 105 | |
| **first** | NaN | 2011-01-01 00:00:00 | NaN | NaN | NaN | NaN | NaN | |
| **last** | NaN | 2012-12-31 00:00:00 | NaN | NaN | NaN | NaN | NaN | |
| **mean** | NaN | NaN | NaN | 2011.500684 | NaN | NaN | NaN | |
| **std** | NaN | NaN | NaN | 0.500342 | NaN | NaN | NaN | |
| **min** | NaN | NaN | NaN | 2011.000000 | NaN | NaN | NaN | |
| **25%** | NaN | NaN | NaN | 2011.000000 | NaN | NaN | NaN | |
| **50%** | NaN | NaN | NaN | 2012.000000 | NaN | NaN | NaN | |
| **75%** | NaN | NaN | NaN | 2012.000000 | NaN | NaN | NaN | |
| **max** | NaN | NaN | NaN | 2012.000000 | NaN | NaN | NaN | |

Berdasarkan data di atas, kita dapat mengetahui bahwa:

- Pengguna Kasual Bike Sharing berjumlah 2 hingga 3410 per hari pada tahun 2011 dan 2012
- Pengguna Register Bike Share berjmlah 20 hingga 6946 per hari pada tahun 2011 dan 2012
- Jumlah keseluruhan pengguna (Kasual dan Register) Bike Sharing berjumlah 22 hingga 8714 per hari pada tahun 2011 dan 2012

## 2. Grouping Data

```
import pandas as pd
import matplotlib.pyplot as plt

# Assuming 'result' is the DataFrame obtained from the previous aggregation
df1.groupby(by="datetime").agg({
    "casual": ["max", "min", "mean"],
    "registered": ["max", "min", "mean"],
    "total_count": ["max", "min", "mean"]
})
```

| | casual | | | registered | | | total_count | | |
|---|---|---|---|---|---|---|---|---|---|
| | max | min | mean | max | min | mean | max | min | mean |
| datetime | | | | | | | | | |
| 2011-01-01 | 331 | 331 | 331.0 | 654 | 654 | 654.0 | 985 | 985 | 985.0 |
| 2011-01-02 | 131 | 131 | 131.0 | 670 | 670 | 670.0 | 801 | 801 | 801.0 |
| 2011-01-03 | 120 | 120 | 120.0 | 1229 | 1229 | 1229.0 | 1349 | 1349 | 1349.0 |
| 2011-01-04 | 108 | 108 | 108.0 | 1454 | 1454 | 1454.0 | 1562 | 1562 | 1562.0 |
| 2011-01-05 | 82 | 82 | 82.0 | 1518 | 1518 | 1518.0 | 1600 | 1600 | 1600.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2012-12-27 | 247 | 247 | 247.0 | 1867 | 1867 | 1867.0 | 2114 | 2114 | 2114.0 |
| 2012-12-28 | 644 | 644 | 644.0 | 2451 | 2451 | 2451.0 | 3095 | 3095 | 3095.0 |
| 2012-12-29 | 159 | 159 | 159.0 | 1182 | 1182 | 1182.0 | 1341 | 1341 | 1341.0 |
| 2012-12-30 | 364 | 364 | 364.0 | 1432 | 1432 | 1432.0 | 1796 | 1796 | 1796.0 |
| 2012-12-31 | 439 | 439 | 439.0 | 2290 | 2290 | 2290.0 | 2729 | 2729 | 2729.0 |

731 rows × 9 columns

```
df1.groupby('weather_condition')['total_count'].mean().reset_index().sort_valu
```

| | weather_condition | total_count |
|---|---|---|
| 1 | Light_RainSnow | 1803.285714 |
| 2 | Misty | 4035.862348 |
| 0 | Clear | 4876.786177 |

```
df1.groupby('workingday')['total_count'].mean().reset_index().sort_values("tot
```

| | workingday | total_count |
|---|---|---|
| 0 | No | 4330.168831 |
| 1 | Yes | 4584.820000 |

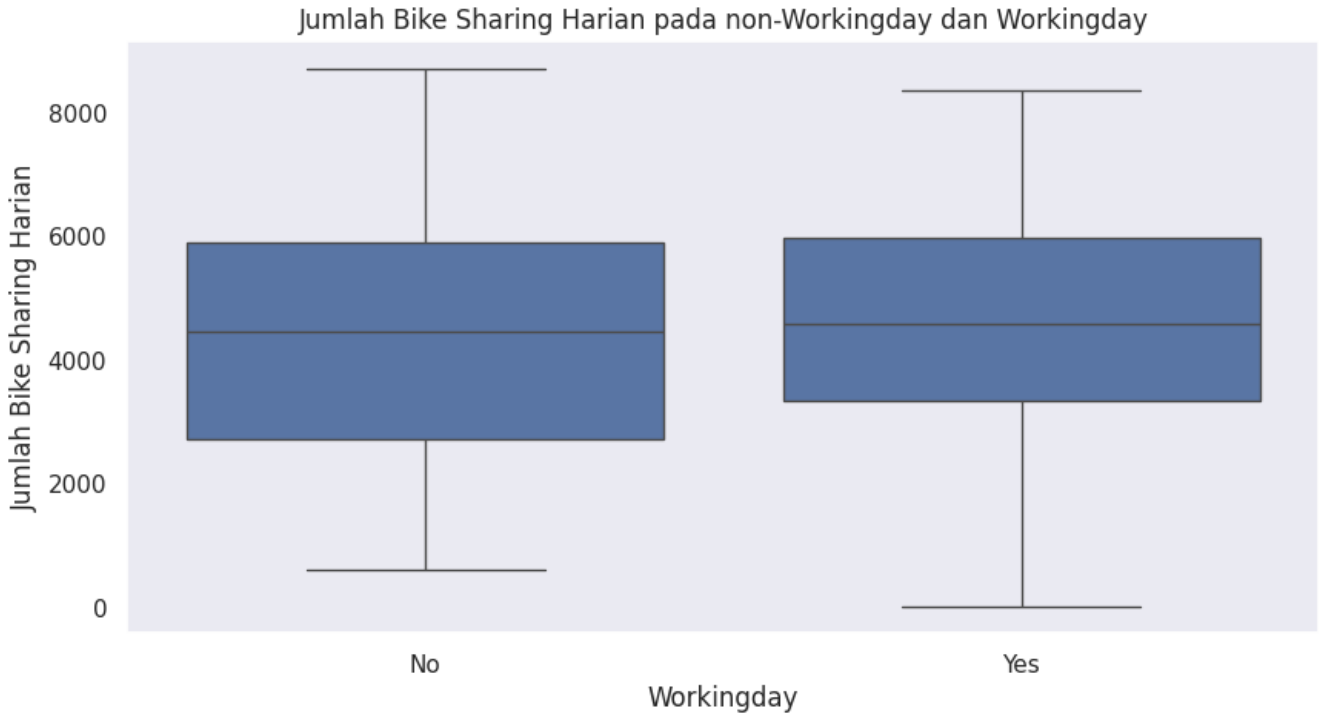
```
df1.groupby('season')['total_count'].mean().reset_index().sort_values("total_c
```

| | season | total_count |
|---|---|---|
| 3 | Winter | 2604.132597 |
| 0 | Fall | 4728.162921 |
| 1 | Spring | 4992.331522 |
| 2 | Summer | 5644.303191 |

## Visualization & Explanatory Analysis

### Pertanyaan 1: Bagaimana pengaruh hari kerja terhadap rata-rata jumlah pengguna keseluruhan (kasual dan register) bike sharing harian?

```
plt.figure(figsize=(10, 5))
sns.boxplot(x="workingday", y="total_count", data=df1)
plt.title("Jumlah Bike Sharing Harian pada non-Workingday dan Workingday")
plt.xlabel("Workingday")
plt.ylabel("Jumlah Bike Sharing Harian")
plt.show()
```




Berdasarkan visualisasi di atas, kita dapat mengetahui bahwa rata-rata jumlah pengguna bike-sharing harian pada **hari kerja** lebih besar daripada **hari non-kerja**.

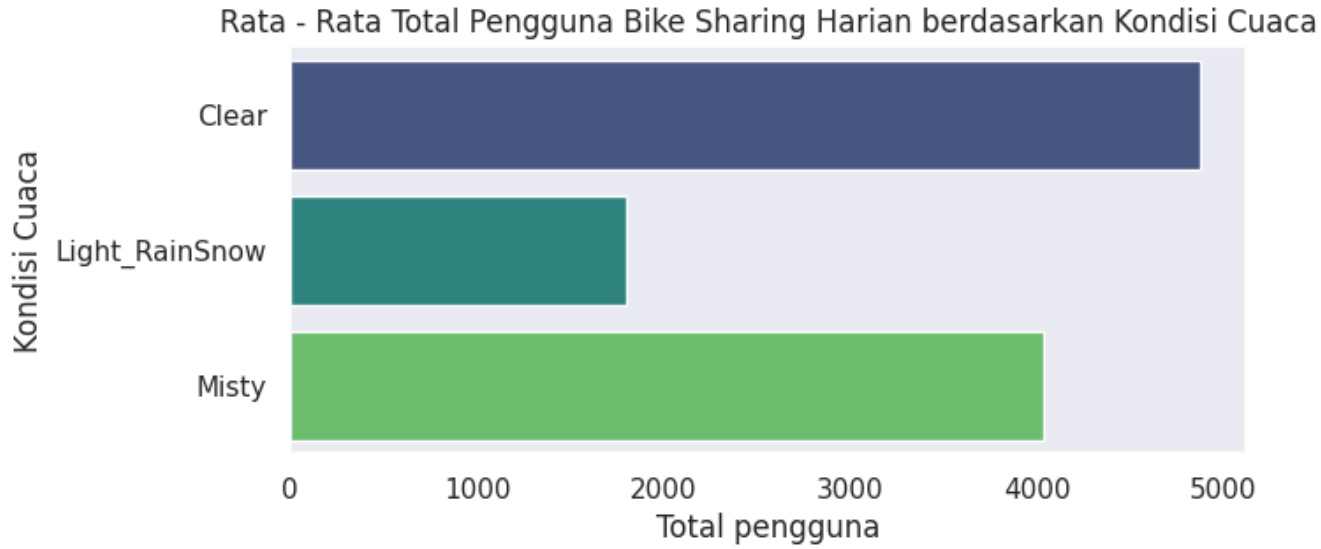
## Pertanyaan 2: Bagaimana pengaruh kondisi cuaca terhadap rata-rata jumlah pengguna keseluruhan (kasual dan register)bike sharing?

```
weather = df1.groupby('weather_condition')['total_count'].mean().reset_index()

plt.figure(figsize=(7, 3))
sns.barplot(x='total_count', y='weather_condition', hue='weather_condition', d

plt.title('Rata - Rata Total Pengguna Bike Sharing Harian berdasarkan Kondisi
plt.xlabel('Total pengguna')
plt.ylabel('Kondisi Cuaca')
```

```
Text(0, 0.5, 'Kondisi Cuaca')
```

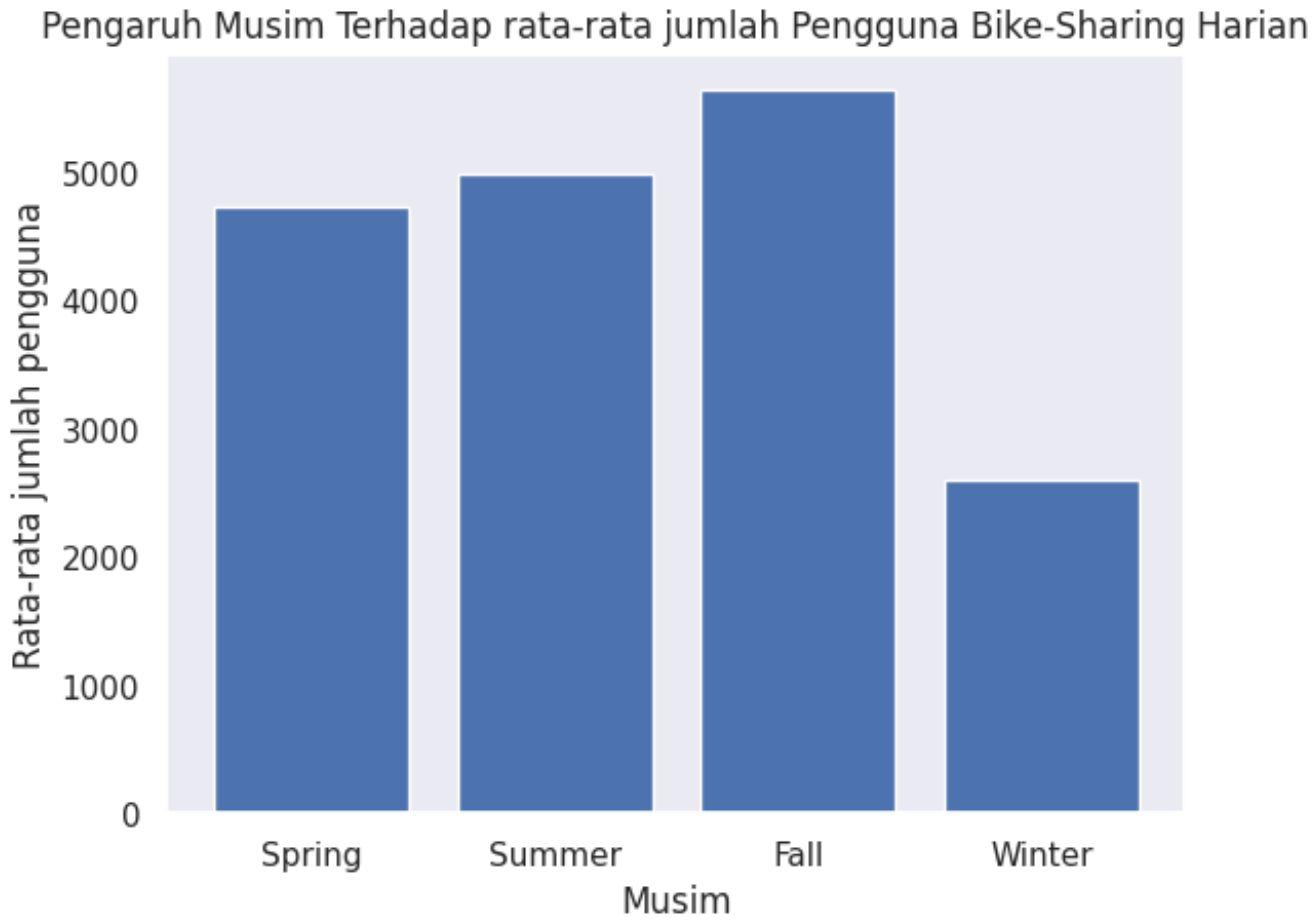



Berdasarkan visualisasi di atas, kita dapat mengetahui bahwa rata-rata jumlah pengguna bike-sharing harian paling banyak terjadi pada saat kondisi cuaca **clear** dan paling sedikit terjadi saat kondisi cuaca **Light_Rainsnow**.

Kondisi cuaca sangat mempengaruhi tingkat pengguna bike-sharing harian. Makin baik kondisi cuaca, maka total pengguna bike-sharing akan makin meningkat, sebaliknya Makin buruk kondisi cuaca, maka total pengguna bike-sharing akan makin menurun.

## Pertanyaan 3: Bagaimana hubungan antara musim dan rata-rata jumlah pengguna keseluruhan (kasual dan register) bike sharing?

```
season = df1.groupby('season')['total_count'].mean()
names = ['Spring', 'Summer', 'Fall', 'Winter']
plt.bar(names, season)
plt.xlabel('Musim')
plt.ylabel('Rata-rata jumlah pengguna')
plt.title('Pengaruh Musim Terhadap rata-rata jumlah Pengguna Bike-Sharing Hari
plt.show()
```




Berdasarkan visualisasi di atas, kita dapat mengetahui bahwa urutan musim dengan rata-rata jumlah pengguna bike-sharing harian dari yang paling banyak hingga yang paling sedikit secara berurutan adalah Fall, Summer, Spring, dan Winter.

## 4. Bagaimana Pola Peningkatan dan Penurunan rata-rata jumlah Pengguna Bike-Sharing Harian dalam kurun waktu tertentu?

```
# Set the Seaborn style to whitegrid
sns.set_style('whitegrid')
```

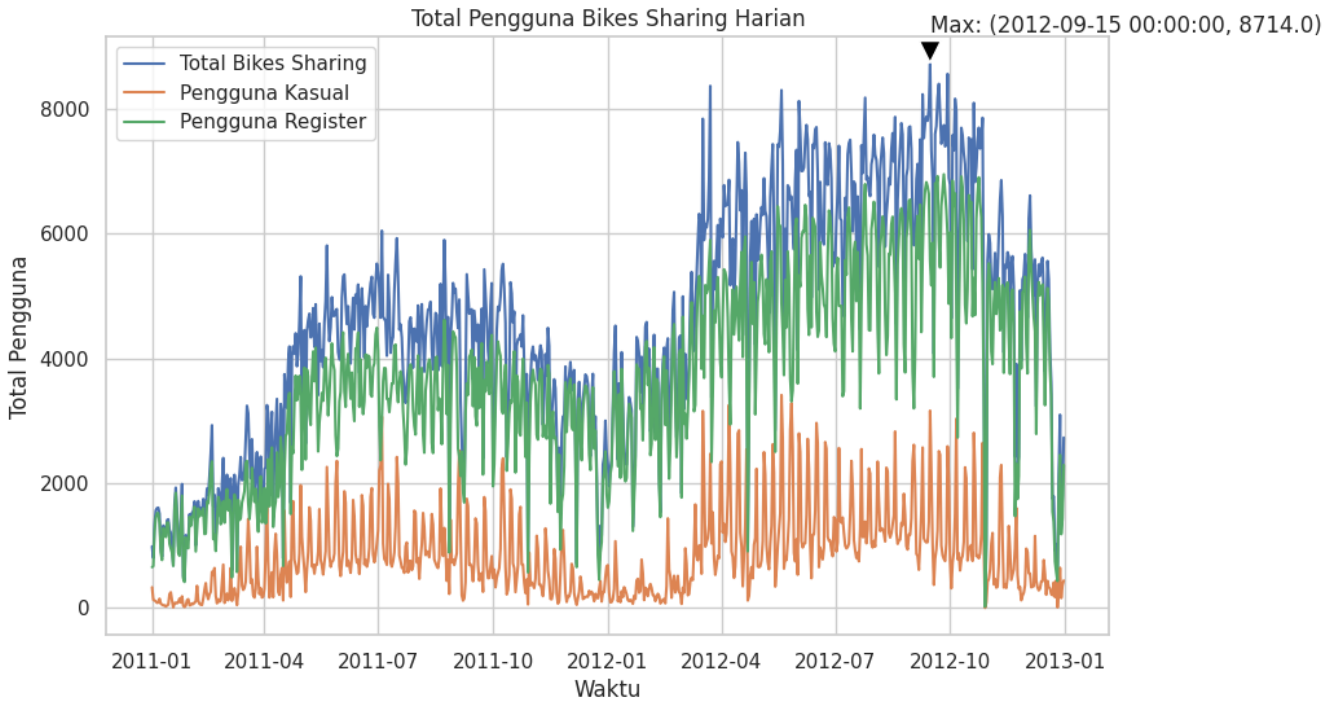
```
# Assuming 'cnt', 'casual', and 'registered' are the columns you want to mean
count_by_day = df1.groupby('datetime')[['total_count', 'casual', 'registered']

plt.figure(figsize=(10, 6))  # Adjust the figure size as needed

# Use 'count_by_day.index' for x-axis and respective columns for y-axis
plt.plot(count_by_day.index, count_by_day['total_count'], label='Total Bikes S
plt.plot(count_by_day.index, count_by_day['casual'], label='Pengguna Kasual')
plt.plot(count_by_day.index, count_by_day['registered'], label='Pengguna Regis

# Add max values for x and y axes
max_x = count_by_day['total_count'].idxmax()
max_y = count_by_day['total_count'].max()
plt.annotate(f'Max: ({max_x}, {max_y})', xy=(max_x, max_y), xytext=(max_x, max

plt.xlabel('Waktu')
plt.ylabel('Total Pengguna')
plt.title('Total Pengguna Bikes Sharing Harian')
plt.legend()
plt.grid(True)
plt.show()
```




Berdasarkan visualisasi di atas, kita dapat mengetahui bahwa terdapat pola rata-rata jumlah penggunaan bike-sharing lebih besar terjadi pada Bulan **Mei** hingga Bulan **Oktober** per tahunnya.

## Conclusion

A. Pertanyaan 1: rata-rata jumlah pengguna bike-sharing harian pada hari kerja lebih besar daripada hari non-kerja.

> Peningkatan jumlah penyewaan pada hari kerja kemungkinan menunjukkan adanya kebutuhan masyarakat untuk melaksanakan kegiatan sehari-hari seperti bekerja, bersekolah, atau berbelanja. Pemilihan sepeda sebagai sarana transportasi mungkin menjadi pilihan yang diminati untuk perjalanan singkat, terutama pada hari-hari ketika orang memiliki kegiatan rutin mereka.

B. Pertanyaan 2: rata-rata jumlah pengguna bike-sharing harian paling banyak terjadi pada saat kondisi cuaca clear dan paling sedikit terjadi saat kondisi cuaca Light_Rainsnow.

> Kondisi cuaca sangat mempengaruhi tingkat pengguna bike-sharing harian. Makin baik kondisi cuaca, maka total pengguna bike-sharing akan makin meningkat. Sebaliknya, makin buruk kondisi cuaca, maka total pengguna bike-sharing akan makin menurun.

> Cuaca yang cerah cenderung menciptakan suasana yang lebih menyenangkan untuk melakukan kegiatan di luar ruangan, seperti bersepeda. Pada saat cuaca cerah, orang lebih condong untuk menggunakan sepeda karena kelangsungan aktivitas tersebut tidak terhambat oleh hujan atau kondisi cuaca yang tidak mendukung.

C. Pertanyaan 3: Urutan musim dengan rata-rata jumlah pengguna bike-sharing harian dari yang paling banyak hingga yang paling sedikit secara berurutan adalah Fall, Summer, Spring, dan Winter.

> Kondisi di atas kemungkinan terjadi karena:
>
> - Musim fall sering dihubungkan dengan suhu yang nyaman, pemandangan yang indah karena daun-daun berguguran, dan kondisi cuaca yang stabil. Semua faktor ini dapat membuat orang lebih termotivasi untuk bersepeda, baik untuk kegiatan rekreasi maupun transportasi.
> - Musim dingin dapat menjadi tantangan bagi aktivitas sepeda karena cuaca yang dingin, bersalju, atau berawan. Kondisi ini dapat mengurangi minat orang untuk menggunakan sepeda, terutama untuk perjalanan.

D. Pertanyaan 4: Pola rata-rata jumlah penggunaan bike-sharing lebih besar terjadi pada Bulan **Mei** hingga Bulan **Oktober** per tahunnya.

> Pola ini bersesuaian dengan rata-rata jumlah penggunaan bike sharing terbesar yang terjadi pada musim fall dan summer. Bulan Mei hingga Bulan Oktober berada dalam rentang kedua musim tersebut.