

# Bike - sharing Machine Learning





# Content

- 1 Business Problem Understanding**
- 2 Data Understanding**
- 3 Data Preprocessing**
- 4 Modelling**
- 5 Conclusion & Recomendation**

# 1

# Pendahuluan

## 1.1 Context

Sistem Bike-Sharing adalah persewaan sepeda tradisional di mana seluruh proses, mulai dari keanggotaan, persewaan, dan pengembalian, telah menjadi otomatis. Melalui sistem ini, pengguna dapat dengan mudah menyewa sepeda dari posisi tertentu dan kembali lagi di posisi lain.

Saat ini, minat besar ada pada sistem ini ada karena peran pentingnya dalam masalah lalu lintas, lingkungan, dan kesehatan. Fitur ini mengubah sistem bike-sharing menjadi jaringan sensor virtual yang dapat digunakan untuk merasakan mobilitas di dalam kota.

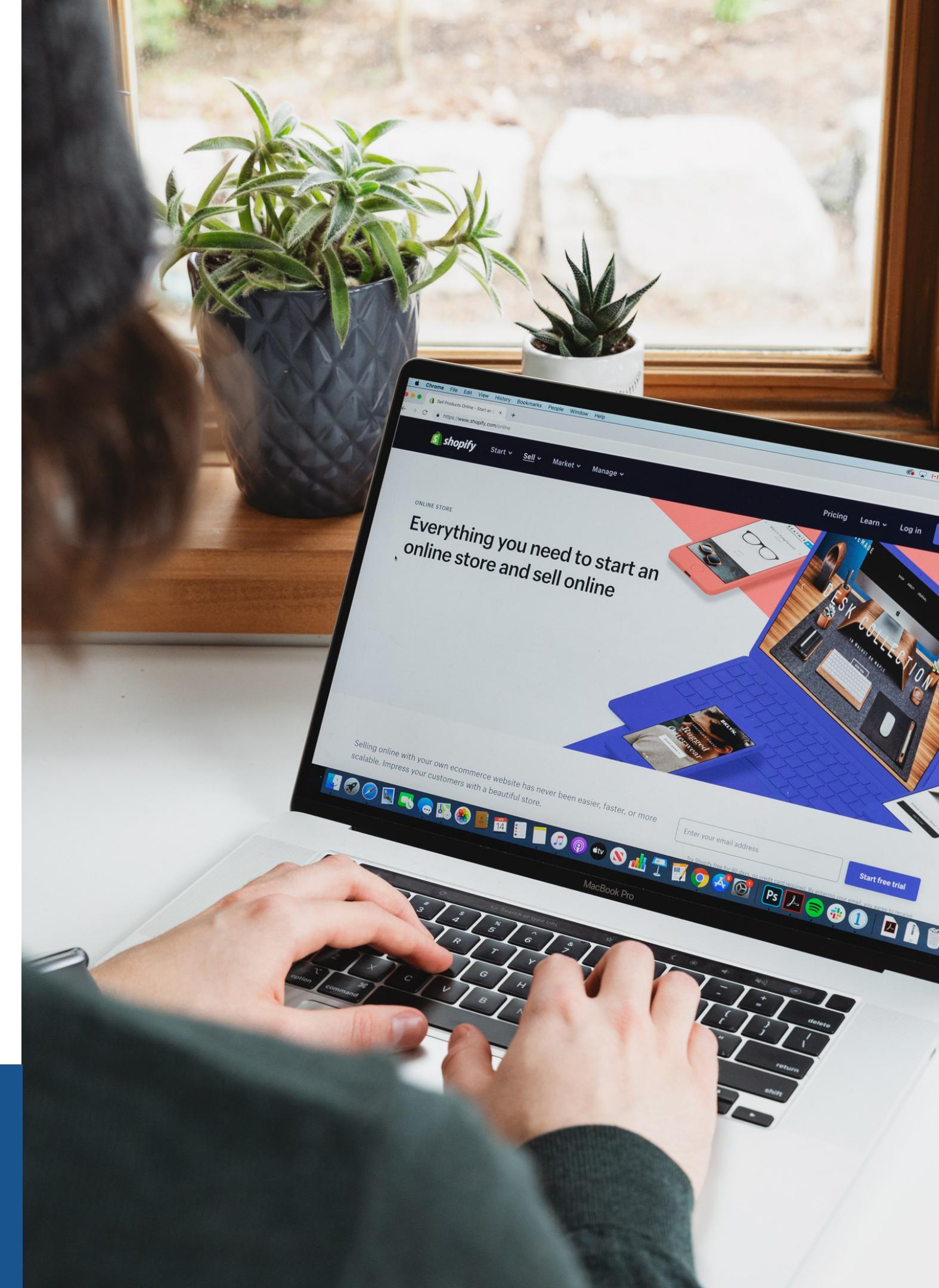


# 1

# Pendahuluan

## 1.2 Tujuan

Penting untuk membuat sepeda sewaan tersedia dan dapat diakses oleh publik pada waktu yang tepat karena mengurangi waktu tunggu. Bagian penting adalah prediksi jumlah sepeda yang dibutuhkan setiap jam untuk pasokan sepeda sewaan yang stabil.



# 1

# Pendahuluan

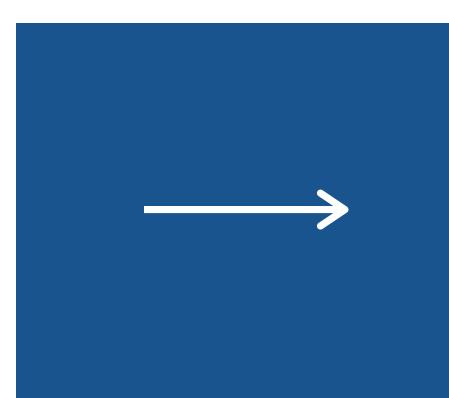
## 1.3 Action & Value

Action :

Mengalokasikan jumlah sepeda yang sesuai dengan hasil prediksi

Value :

Perusahaan bike-sharing dapat memiliki tingkat efisiensi dalam operasional bisnisnya



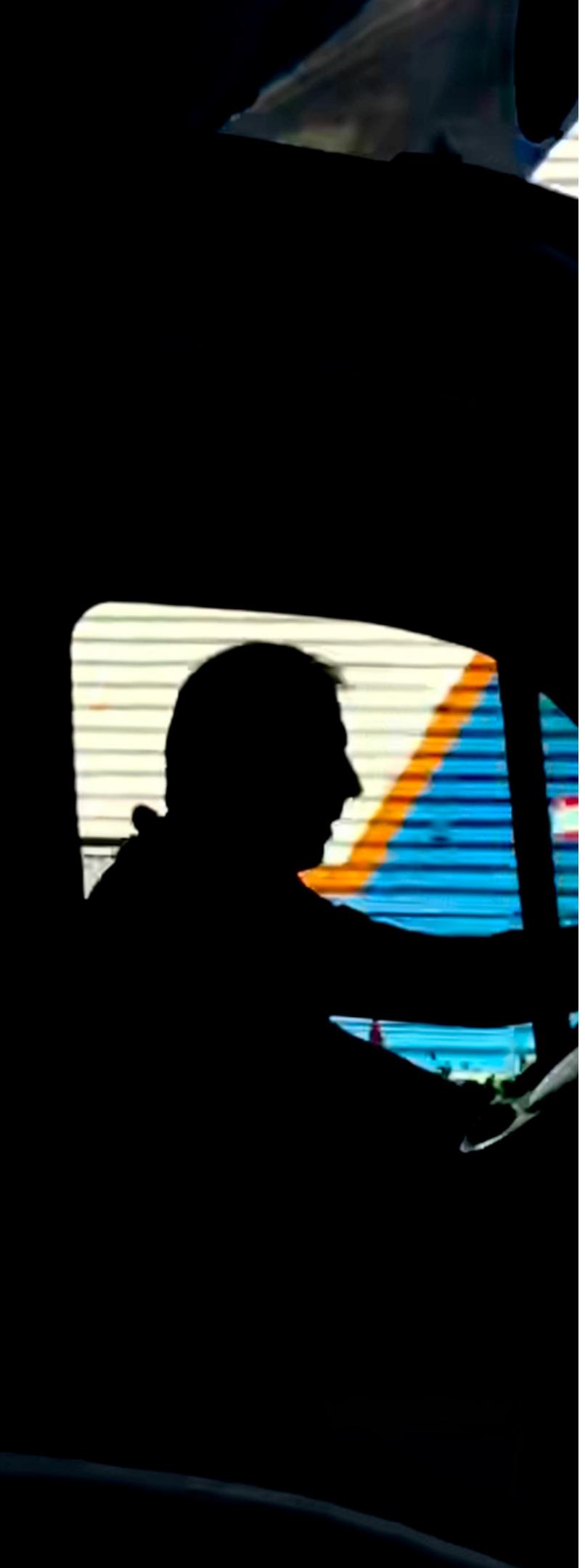
# 1

# Pendahuluan

## 1.4 Analytics Approach

Melakukan analisa terhadap data untuk dapat menemukan pola dari fitur-fitur yang tersedia. Diantaranya membedakan satu kondisi dengan kondisi yang lain & bagaimana setiap kondisi mempengaruhi jumlah pasokan sepeda yang tersedia. Selanjutnya akan dibuat model regresi yang bertujuan untuk menentukan jumlah unit sepeda yang perlu disediakan oleh perusahaan





# 1 Pendahuluan

## 1.5 EVALUATION METRICS

MAE	MAPE	R2
Selisih absolut antara hasil aktual dan hasil prediksi,	Rataan persentase error yang dihasilkan oleh model regresi	Melihat seberapa signifikan variabel independen mempengaruhi variabel dependen.



# Data Understanding



# Informasi Dataset

Attribute	Data Type	Description
dteday	Object	Tanggal
hum	float64	Kelembaban / humidity yang dinormalisasi. Values dibagi 100
weathersit	int64	cuaca(*)
holiday	int64	Libur / tidak libur
season	int64	season (1: winter, 2: spring, 3: summer, 4: fall)
temp	float64	Temperatur dalam celcius
atemp	float64	Normalisasi feeling dalam celcius
hr	int64	hour / jam (0 - 23)
casual	int64	Jumlah casual users
registered	int64	Jumlah registered users
cnt	int64	Jumlah pengguna (casual&registered)

(\*) 1: Clear, Few clouds, Partly cloudy, Partly cloudy  
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

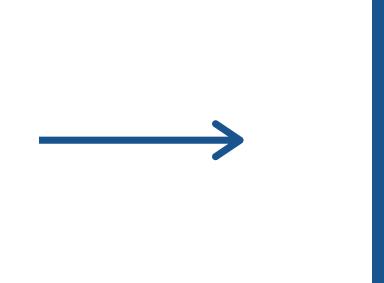
## Bentuk Data

- Data memiliki 11 kolom dan 12.165 baris
- Data memiliki value NaN sebesar 0%
- Data memiliki duplikat sebesar 0%
- Data kategorikal terdiri atas weathersit, holiday, dan season. Sedangkan sisanya berupa numerikal



# Data Preprocessing

---



# 3 Data Preprocessing

## 1. Penyesuaian Kolom dan Value

dteday	hum	weathersit	holiday	season	atemp	temp	hr	casual	registered	cnt
2011-12-09	0.62	1	0	4	0.3485	0.36	16	24	226	250
2012-06-17	0.64	1	0	2	0.5152	0.54	4	2	16	18
2011-01-01	0.62	1	0	4	0.3485	0.36	16	24	226	250

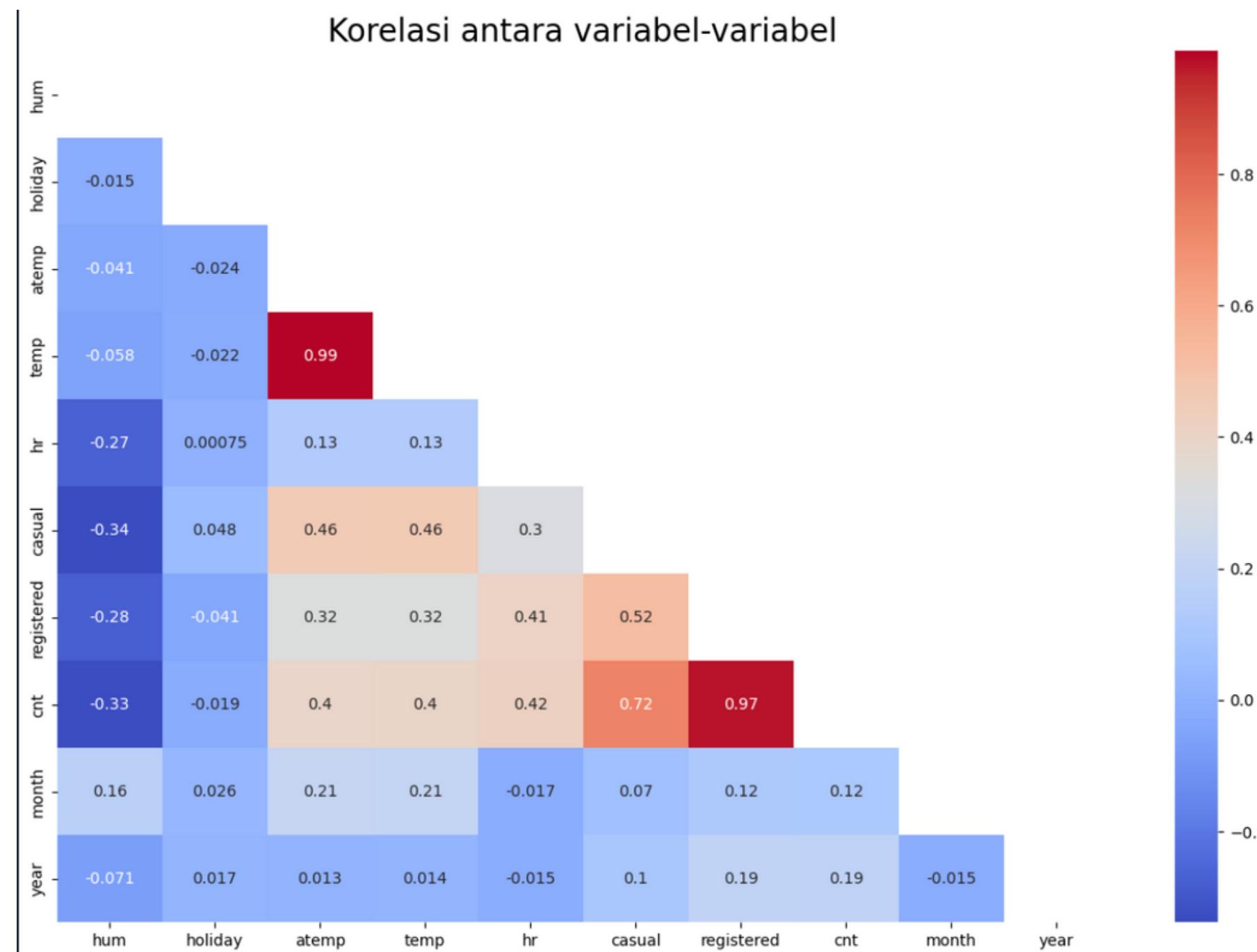
## 2. Mengubah & memisah tipe data 'Date'

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11813 entries, 0 to 12164
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   dteday          11813 non-null   datetime64[ns]
```

# 3

# Data Preprocessing

## 3. Memeriksa Korelasi Data



Menghapus kolom:

- Atemp
- Registered
- Casual
- dteday

# 3

# Data Preprocessing

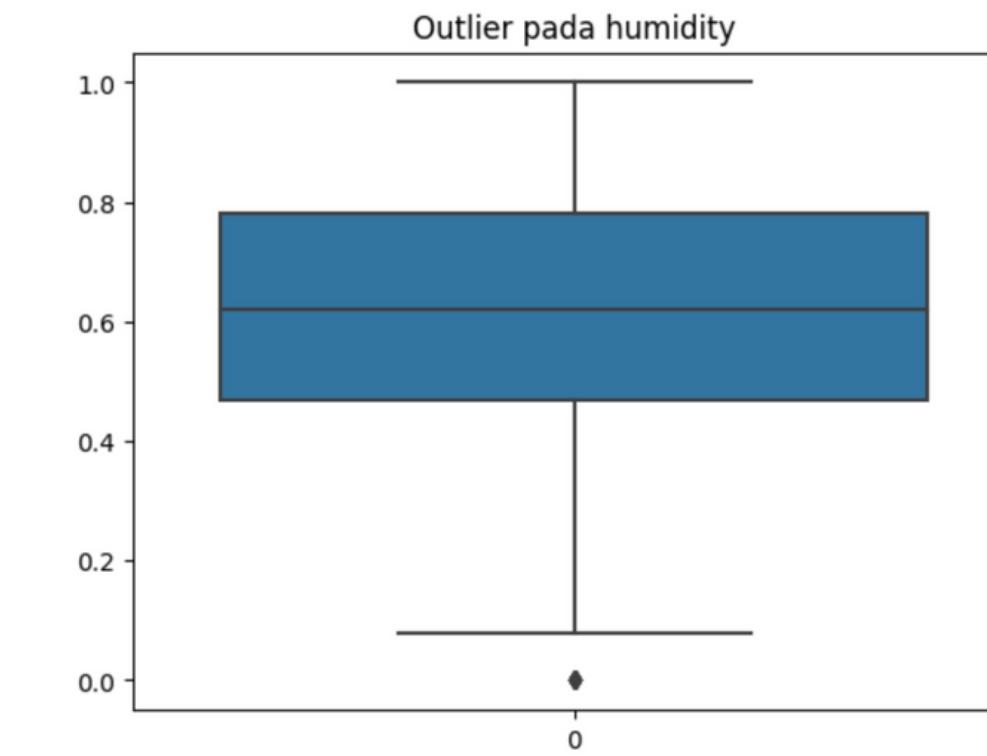
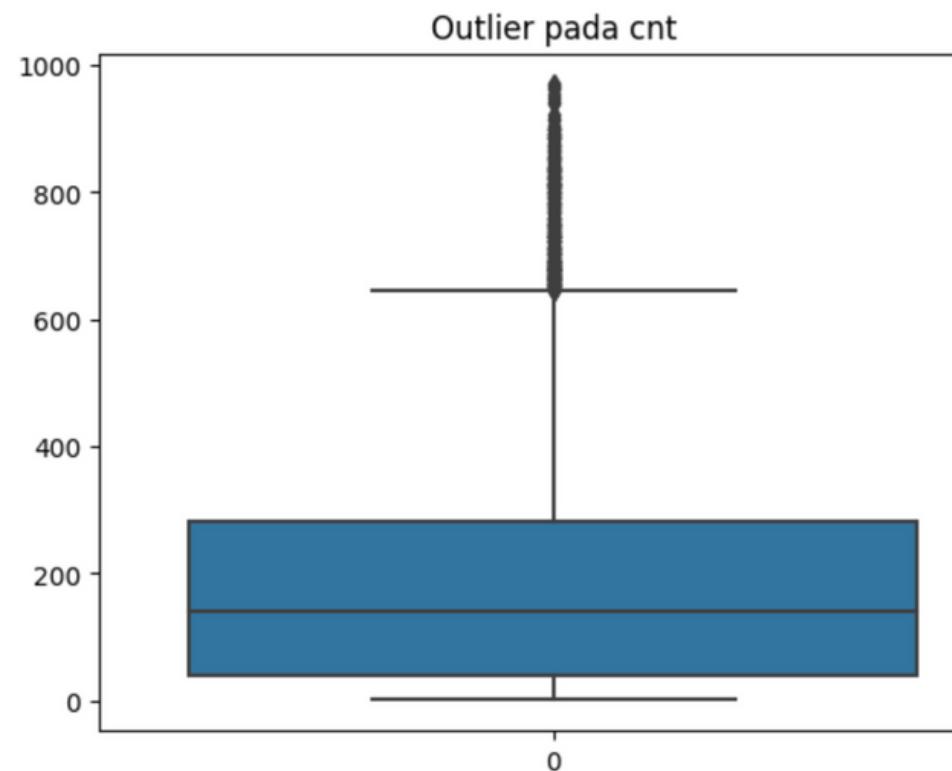
## 4. Memeriksa Outlier

Count:

- Limit atas : 645
- Terdapat 338 baris data lebih dari 645

Humidity:

- Terdapat 14 baris data yang memiliki  $\text{humidity} = 0$
- Berdasarkan domain knowledge,  $\text{humidity} \neq 0$



## 5. Data Bersih

	hum	weathersit	holiday	season	atemp	hr	cnt	month	year	day
0	0.62	Clear	0	Winter	0.3485	16	250	12	2011	Friday
1	0.64	Clear	0	Summer	0.5152	4	18	6	2012	Sunday
2	0.53	Clear	0	Summer	0.6212	23	107	6	2011	Wednesday
3	0.87	Mist	0	Summer	0.3485	8	145	3	2012	Saturday
4	0.55	Clear	0	Fall	0.6970	18	857	7	2012	Tuesday

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12150 entries, 0 to 12164
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   hum         12150 non-null   float64
 1   weathersit  12150 non-null   object 
 2   holiday     12150 non-null   int64  
 3   season      12150 non-null   object 
 4   atemp       12150 non-null   float64
 5   hr          12150 non-null   int64  
 6   cnt         12150 non-null   int64  
 7   month       12150 non-null   int64  
 8   year        12150 non-null   int64  
 9   day         12150 non-null   object 
dtypes: float64(2), int64(5), object(3)
memory usage: 1.0+ MB
```

# ③

# Data Preprocessing

## 4. Encoding

- OneHotEncoding:
  - Season,
  - Weather,
  - Year
- Binary Encoder :
  - Day

## 5. Train Test Split

- 70% Train Set
- 30% Test Set

## 6. Scaler

- Standard Scaler



# Model Implementation





# Modelling

## Base Model

- Linear Regression
- KNN Regressor
- DecisionTree Regressor

## Ensemble Method

- Random Forest Regressor
- XGBoost Regressor

	Model	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	R2
0	Linear Regression	-108.367818	3.436361	-1.377933	0.034738	0.179721
1	KNN Regressor	-76.812760	1.735999	-1.121603	0.040184	0.533720
2	DecisionTree Regressor	-44.994748	3.445480	-0.475126	0.037745	0.814921
3	RandomForest Regressor	-34.562219	2.553333	-0.337930	0.028628	0.899405
4	XGBoost Regressor	-29.823606	1.489769	-0.279614	0.016727	0.928073

Berdasarkan hasil skor evaluation metrics dari setiap model dengan menggunakan K-Fold Cross Validation, didapatkan bahwa model **XGBoost** memiliki score terbaik. Dengan masing-masing nilai metrics **MAE** (29.82), **MAPE** (27%), **R-Squared** (0.928)

# XGBoost

XGBoost adalah salah satu model dalam supervised machine learning. Objective Function dari XGBoost mengandung fungsi *loss* dan fungsi regularisasi. Fungsi tersebut merupakan perbedaan antara data aktual dan data prediksi. Model ini merupakan tipe ***Similar Type Ensemble Method.***

Di dalam algoritma ini, Decision Tree dibuat dalam bentuk yang sekuensial, yaitu bobot memegang peranan penting dalam XGBoost. Bobot diterapkan kepada semua variabel independen yang nantinya akan dimasukkan ke dalam Decision Tree untuk memprediksi hasil. Bobot dari variabel yang diprediksi akan ditambahkan dan variabel ini akan dimasukkan ke dalam Decision Tree berikutnya.

# Predict to Test (XGBoost)

	MAE	MAPE	r2
XGB	28.618234	0.448199	0.93827

Terlihat bahwa ketika melakukan prediksi pada Test Set, XGBoost tetap memiliki performa yang baik (tidak underfitting / overfitting). Di mana nilai **MAE** mengalami penurunan (dari 29.8 ke 28.6) dan **MAPE** tidak berbeda jauh dengan train set, walaupun mengalami peningkatan (dari 0.27 ke 0.44). Sedangkan nilai **R-Squared** mengalami peningkatan (dari 0.92 ke 0.93)

# Hyperparameter Tuning (RandomizedSearchCV)

PARAMETER	VALUE
max_depth	[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
n_estimators	range 100 - 300
colsample_bytree	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
subsample	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
reg_alpha	list(np.logspace(-1, 1, 10))
reg_lambda	list(np.logspace(-1, 1, 10))
gamma	[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]

# Hyperparameter Tuning Result

## Hyperparameter Tuning Result

PARAMETER	VALUE
max_depth	10
learning_rate	0.1
n_estimators	185
colsample_bytree	0.9
subsample	0.4
reg_alpha	1.291549665014884
reg_lambda	1.291549665014884
gamma	6

## XGBoost test sebelum tuning

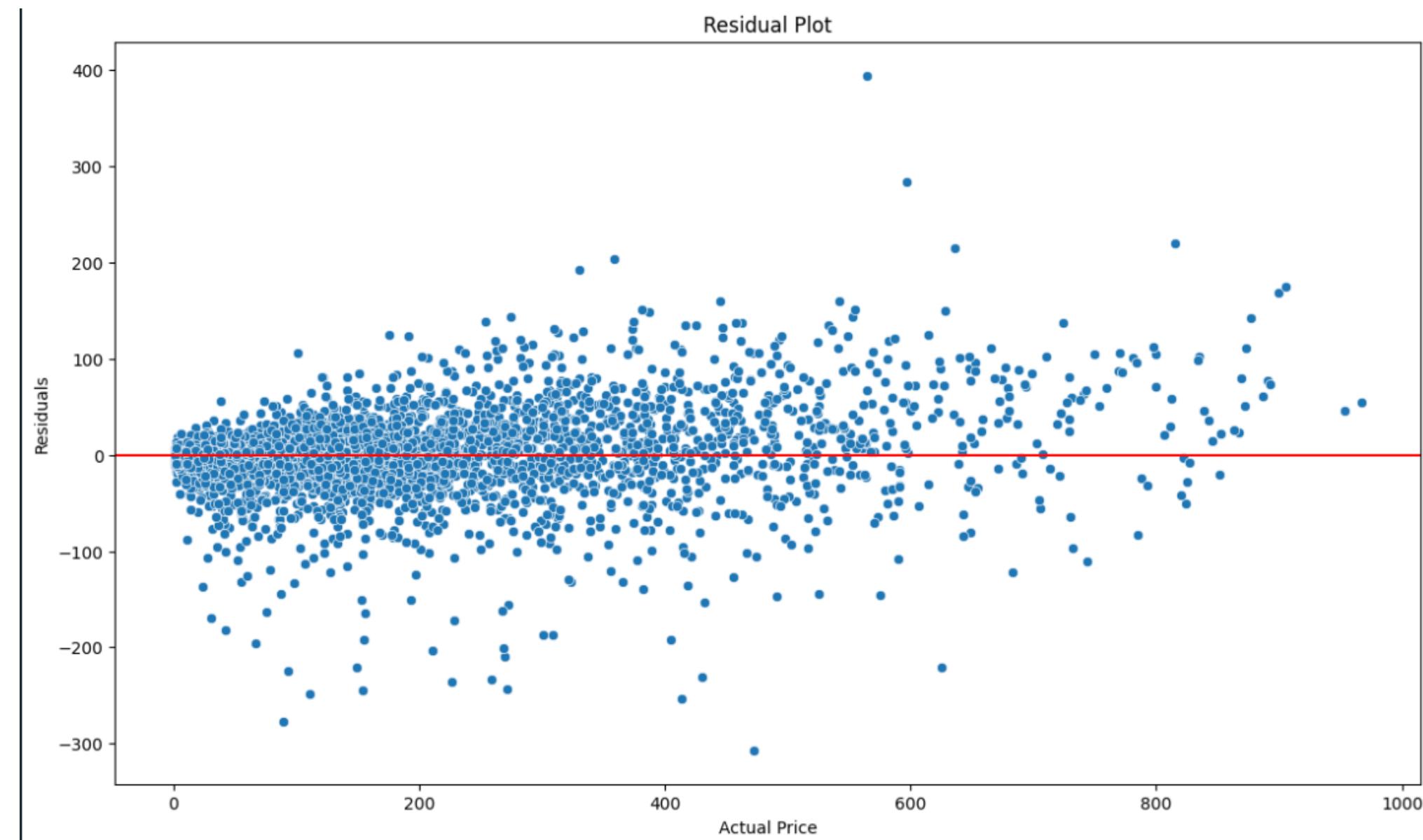
	MAE	MAPE	r2
XGB	28.618234	0.448199	0.93827

## XGBoost test setelah tuning

	MAE	MAPE	r2
XGB	27.766109	0.388605	0.941186

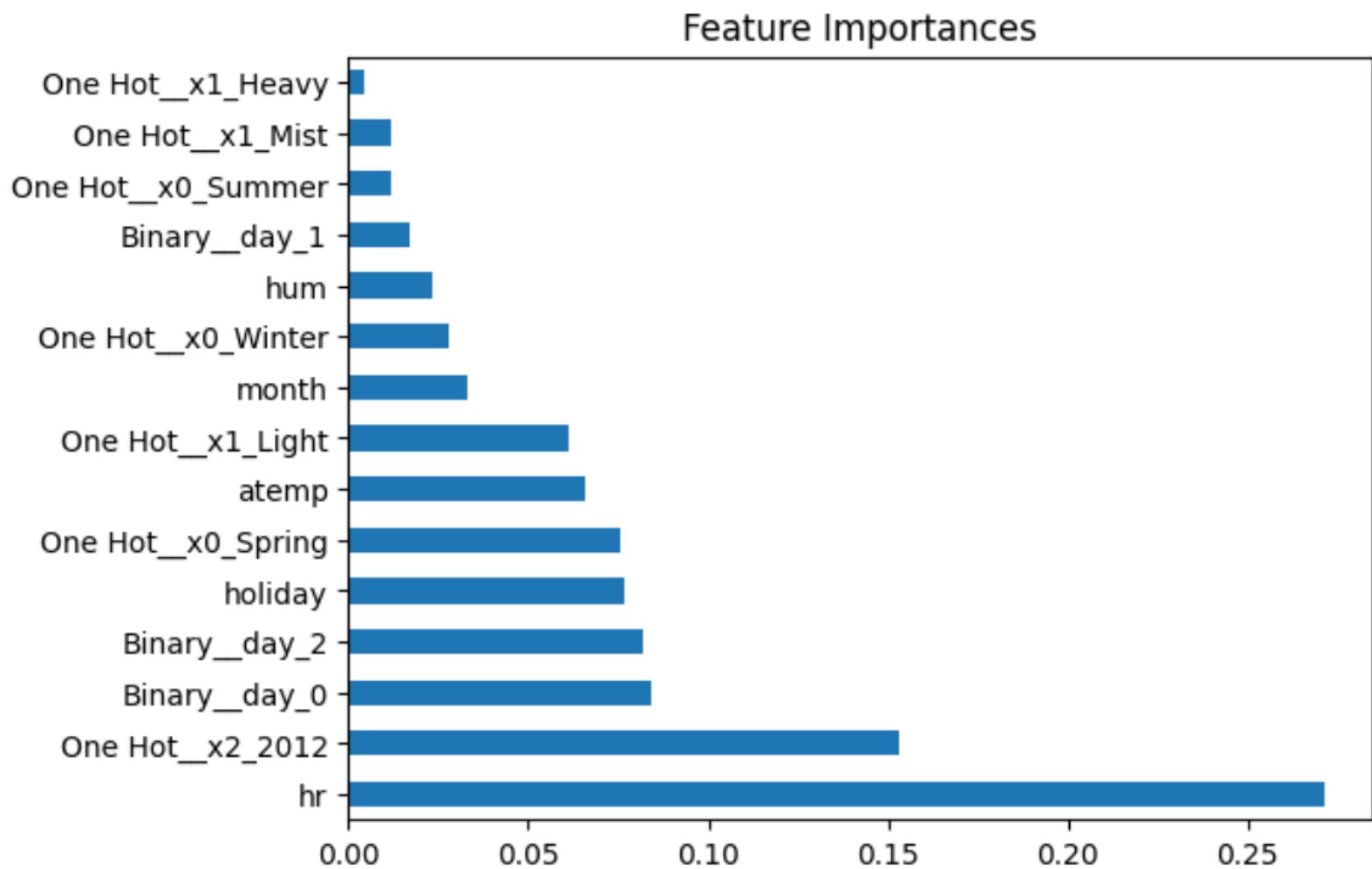
# RESIDUAL ANALYSIS

- Terlihat bahwa hasil prediksi yang diperoleh cukup baik. Namun masih terdapat kemungkinan bias.
- Adanya nilai-nilai error yang besar membuat perbedaan yang cukup signifikan antara nilai MAPE dan MAE
- Terdapat harga aktual yang rendah tapi diprediksi jauh lebih tinggi (overestimation), dan juga sebaliknya (underestimation)



# FEATURE IMPORTANCE

- Berdasarkan pemodelan yang telah dilakukan, fitur 'Hour', 'one\_hot\_x2\_2012', dan 'binary\_day\_0' merupakan fitur yang paling berpengaruh terhadap 'Count'.



# CONCLUSION



- Nilai R<sup>2</sup> yang dihasilkan setelah hyperparameter tuning adalah 94%. Sehingga dapat disimpulkan bahwa 93% data yang diobservasi pada target terlingkup pada model regresi.
- Nilai MAPE yang dihasilkan setelah hyperparameter tuning adalah 38%. Sehingga dapat disimpulkan bahwa jika model akan digunakan untuk memperkirakan penyewa sepeda pada rentang nilai sesuai model yang dilatih (maksimal Count 970), maka perkiraan penyewa dapat meleset kurang lebih 38orang dari total aktualnya.
- Namun dapat terjadi kesalahan lebih jauh juga karena masih terdapat bias yang terlihat dari visualisasi data prediksi dan aktual.

# Recomendation



- Bias dalam prediksi dihasilkan karena terbatasnya feature pada dataset yang berkaitan dengan Target (jumlah unit sepeda yang disewa) atau yang mampu merepresentasikan keadaan dimana calon pelanggan memutuskan untuk menggunakan jasa peminjamanan seperti lokasi stasiun sepeda terhadap lokasi perkantoran, sekolah, tempat wisata, ruang publik, dll
- *Penambahan data. Jika terdapat data lebih banyak dan lebih dari dua tahun, dapat dianalisis lebih baik lagi keterkaitan tahun dengan Count. Dapat juga digunakan model yang lebih kompleks untuk proses pemodelannya untuk dibandingkan dan dicari model dengan error paling sedikit.*

# Thankyou

