## Table of Contents

### General Information

In cities, bike-sharing programs are becoming a must-have form of mobility. Nonetheless, the COVID-19 epidemic had a major impact on this industry's profitability. US bike-sharing company BoomBikes is attempting to forecast post-pandemic demand for rented bikes. They will be able to better meet future demand and streamline their business plan with the aid of this approach.

Business Issue
Finding the variables that affect the demand for bike sharing is the project's main goal. Through the identification of these variables and their implications for bike rentals, BoomBikes can enhance their profitability and get ready for the post-lockdown market.

### Model Development and Evaluation

### Preparing Data

Converting variables of a category To prevent any bias in ranking, columns such as season and weathersit (weather condition) are transformed into dummy variables.
**Goal variable:** The prediction objective is the total bike rentals, or cnt column.

### Choosing Features

Analyse exploratory data (EDA) to find any possible relationships between variables and the demand for bikes.
When appropriate, normalise features using feature scaling approaches (normalization/standardization).

**Model Construction Algorithm:** Regression using Multiple Linear Models
Seaborn, Matplotlib, NumPy, Scikit-learn, and Pandas are the libraries.
Data splitting: Partition the dataset (often 80%–20%) into training and testing sets.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Split data into features and target variable
X = df.drop(columns=['cnt', 'casual', 'registered'])
y = df['cnt']
```

```
# Split dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Model training
lr = LinearRegression()
lr.fit(X_train, y_train)

# Predictions
y_pred = lr.predict(X_test)

# Evaluate model
r2 = r2_score(y_test, y_pred)
print(f'R-squared on test set: {r2}')
```

### In conclusion

First, weather, holiday status, and temperature all have a big impact on how much demand there is for shared bikes.
**Conclusion 2**: The year (2018 or 2019) and seasonal variations also play a major role in the demand for bikes.
**Conclusion 3**: The model fits the data for demand prediction fairly well, as indicated by the R-squared value (the coefficient of determination).
**Conclusion 4**: Non-linear models can be investigated further, and ensemble methods for prediction can be employed.

## Technologies Used

- Python - 3.8
- Pandas - 1.1.3
- NumPy - 1.19.2
- Scikit-learn - 0.23.2
- Matplotlib - 3.3.2
- Seaborn - 0.11.0

### Recognitions

BoomBikes and their desire to comprehend post-pandemic bike demand served as the impetus for this endeavour.
We would especially like to thank DataScience Projects for their thoughtful data science approaches.

# Contact

Created by [@SarbjeetParija] - feel free to contact me!

## Assignment-based Subjective Questions:

**1. The Impact of Categorical Variables on the Dependent Variable**
Season and weather were examples of categorical variables that were transformed into dummy variables. These factors have a big impact on bike rentals. For instance, the weather has a significant effect on bike usage since inclement weather, such as prolonged rain, reduces demand. In a similar vein, greater rental counts are influenced by seasonal factors (such as summer and autumn).

**2. When creating a dummy variable, why not use drop_first=True?**
The drop_first=True argument removes one level of the category variable in order to prevent multicollinearity. If there are four seasons, for instance, we only need three dummy variables because it is possible to deduce the fourth from the lack of the other three.

**3. Pair Plot's Highest Correlation**
Temperature (temp) and the goal variable (cnt) have the most positive association, according to the pair plot. This makes sense because pleasant weather (warmer temps) tends to encourage people to rent bikes.

**4. Verifying Linear Regression Assumptions**
Scatter plots between predicted values and residuals were used to verify linearity.
Normality: To determine whether the residuals follow a normal distribution, Q-Q plots were used to plot the data.
Multicollinearity: To find multicollinearity between variables, the Variance Inflation Factor, or VIF, was utilised.

### 5. Top 3 Features Contributing to Demand

- **Temperature (`temp`)**
- **Year (`yr`)**
- **Weather Situation (`weathersit`)**

## General Subjective Questions:

### 1. Linear Regression Algorithm

Linear regression is a statistical method to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The equation takes the form Y=β0+β1X1+β2X2+...+βnXnY = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_nY=β0+β1X1+β2X2+...+βnXn, where YYY is the predicted output, XXX are the input features, and β\betaβ are the coefficients.

## 2. Anscombe's Quartet

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation), but appear very different when graphed. It highlights the importance of visualizing data before analysis, as different datasets can have the same statistical properties but vastly different distributions.

## 3. Pearson's R

Pearson's correlation coefficient (Pearson's R) measures the strength and direction of the linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

## 4. Scaling in Machine Learning

Scaling is a technique used to standardize the range of independent variables. It is important because many machine learning algorithms perform better when features are on the same scale. Standardization scales the data to have a mean of 0 and a standard deviation of 1, while normalization scales the data to be between 0 and 1.

## 5. Infinite VIF

VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity, meaning one variable is a perfect linear combination of another. This leads to singular matrices that cannot be inverted, causing issues in regression analysis.

## 6. Q-Q Plot

A Q-Q plot compares the distribution of the residuals to a normal distribution. It is used in linear regression to check the assumption that the residuals are normally distributed. A straight line in the Q-Q plot indicates that the residuals are normally distributed.