# BOOMBIKES

## LINEAR REGRESSION ASSIGNMENT
## SUBECTIVE QUESTIONS

PREPARED BY: VIJAY MAHAWAR

# ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

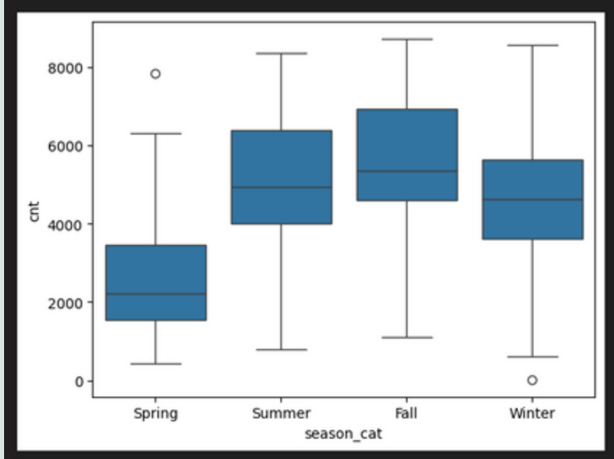
**Question 1**

**From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Categorical variables such as **season and weather** significantly influence the dependent variable, as they determine the usability and demand for shared bikes. For example, unfavorable weather conditions reduce demand. As can be seen from the **BoomBikes** can maximize their user base by taking informative decision based on the these categorical variables.

```
weathersit_cat
A    463
B    246
C     21
Name: count, dtype: int64
```

- 💡 **A - Clear, Few clouds, Partly cloudy, Partly cloudy** Weather is obviously *most popular* among the riders
- 💡 Whereas **C - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** is *least popular*




💡 **Fall season** seems to be most popular for bike riding

## Question 2

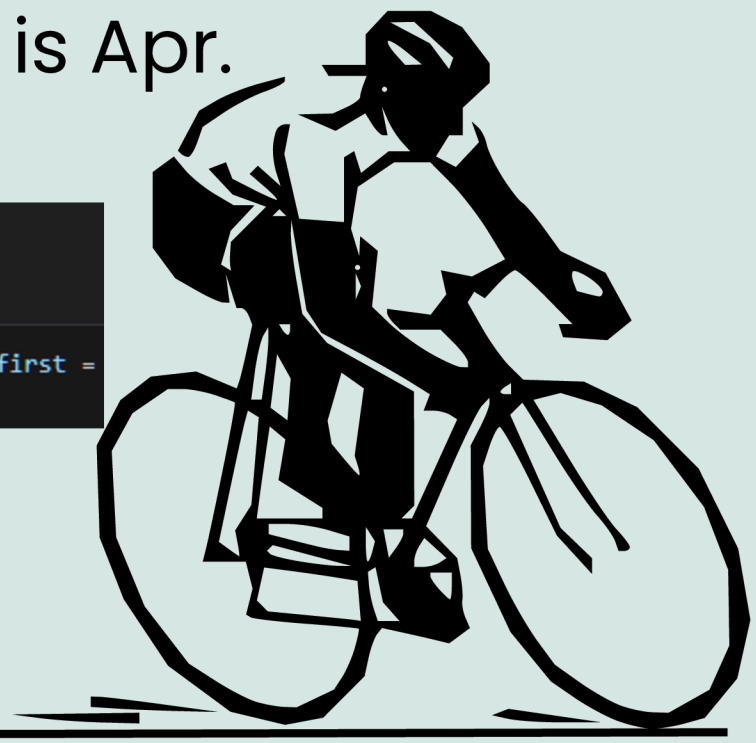
**Why is it important to use drop_first=True during dummy variable creation? (2 marks)**

Using drop_first=True avoids the dummy variable trap, which occurs when there is perfect multicollinearity among the created dummy variables. By dropping the first category, one dummy variable is omitted, ensuring that the model can interpret the effects of the remaining categories without redundancy or distortion in predictions.

For instance, mnth_cat_status would have 11 months, from May to Mar. If all are False it implies that missing month is Apr.

- `mnth_cat` -> `mnth_cat_status`

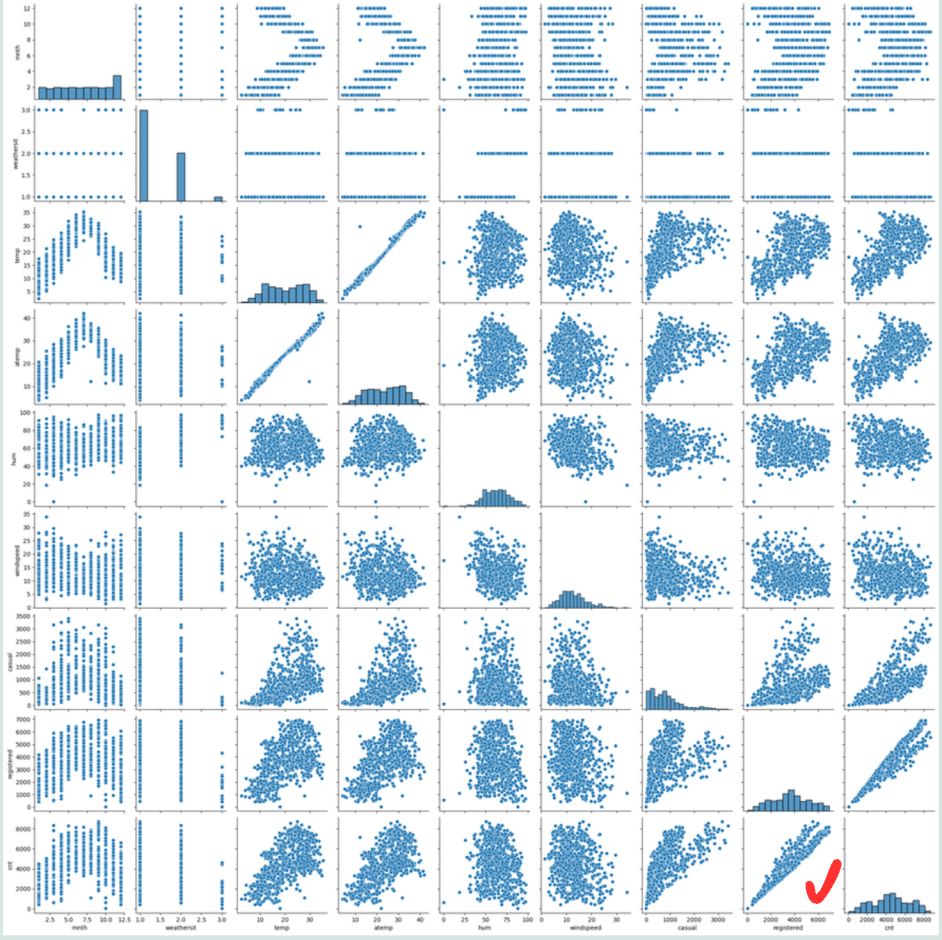
```
1  mnth_cat_status = pd.get_dummies(data=bike_df['mnth_cat'], drop_first =
2  mnth_cat_status
```



## Question 3

**Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

From the pair-plot, **Registered Users** shows the highest positive correlation with the target variable. This indicates that as Registered Users increases, the demand for shared bikes also rises, likely due to better weather conditions encouraging bike usage.

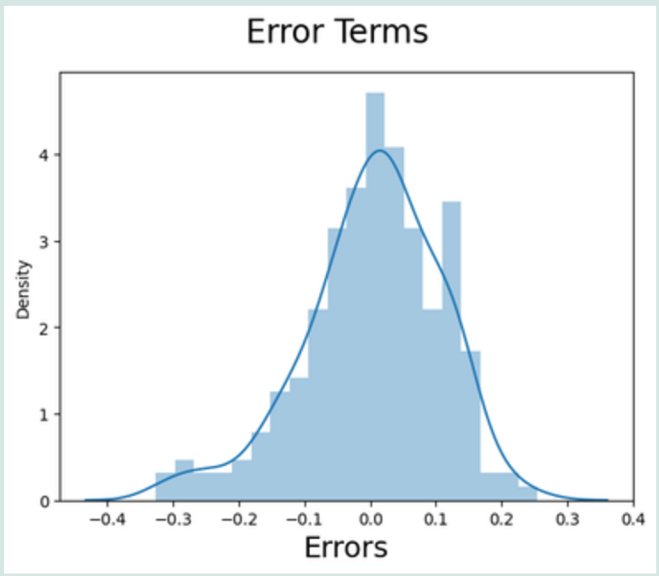



## Question 4

**How did you validate the assumptions of Linear Regression after building the model on the training set?(3 marks)**

The assumptions of Linear Regression were validated using residual diagnostics:

1. **Linearity**: Checked using scatterplots of residuals versus predicted values, ensuring no obvious patterns exist.
2. **Homoscedasticity**: Verified that residuals have constant variance through residual plots.
3. **Normality**: Assessed by plotting a Q-Q plot or histogram of residuals.
4. **Multicollinearity**: Measured using Variance Inflation Factor (VIF) to ensure predictor variables are not highly correlated. StaticMethod - model_toolkit.check_VIF(p_X_train) takes care of this.




## Question 5

**Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?(2 marks)**

The top 3 features identified from the model based on their coefficients and statistical significance are:

1.**Temperature**: Positively correlated with demand as favorable weather increases bike usage.

2.**Weather**: Indicates peak usage clear weather. This has negative relationship with the target variable

3.**Season (Summer/Spring)**: Seasonal variations show higher demand in warmer months.

```
cnt = 0.166303 + 0.126584 * season + 0.235810 * yr - 0.095565 * holiday - 0.187021 * weathersit + 0.467522 * temp - 0.148497 * windspeed
```

# GENERAL
# SUBJECTIVE QUESTIONS

## Question 1

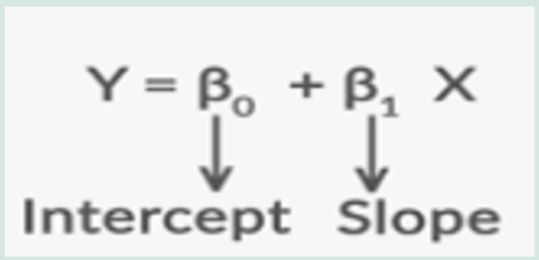
## Question **1. Explain the linear regression algorithm in detail.(4 marks)

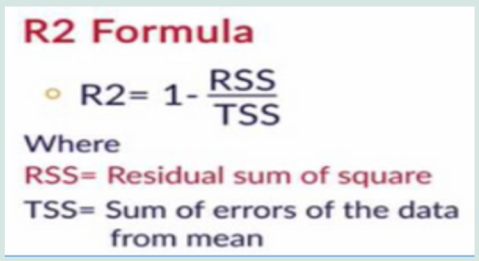
A simple linear regression is a supervised learning algorithm used to model the relationship between independent variables and a dependent variable. It assumes a linear relationship. The equation is shown at the bottom.

The Models using this algorithm minimizes the sum of squared residuals (errors) using the Ordinary Least Squares (OLS) method. Assumptions include linearity, independence, normality of residuals, and homoscedasticity.

Linear regression is widely used for prediction and inference, and it provides coefficients to interpret the impact of predictors on the target variable.

Models in Linear Regression is evaluated using R-Squared - of Coeffient of Determination. Shown below:



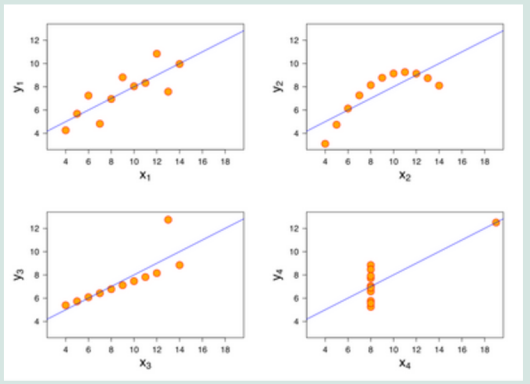




R-Squared takes values between 0 and 1. Higher the values better it is. Multi-collinearity issue in Multiple Linear Regression is solved using VIF and P-value

## Question 2

## Explain the Anscombe's quartet in detail.(3 marks)

Anscombe's quartet comprises of four datasets and have equal statistical properties, such as mean, variance, and correlation, but vastly different distributions and visual patterns. It highlights the importance of visualizing data before analysis. Despite similar numerical summaries, their scatterplots reveal (In below):




X1 has a linear relationship to y1
X2 has a parabolic relationship to y2
X3 has an outlier and its relationships to y3.
X4 has a vertical cluster with an outlier in the fourth dataset.

For all four datasets:

| Property | Value | Accuracy |
|---|---|---|
| Mean of $x$ | 9 | exact |
| Sample variance of $x$: $s_x^2$ | 11 | exact |
| Mean of $y$ | 7.50 | to 2 decimal places |
| Sample variance of $y$: $s_y^2$ | 4.125 | ±0.003 |
| Correlation between $x$ and $y$ | 0.816 | to 3 decimal places |
| Linear regression line | $y = 3.00 + 0.500x$ | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression: $R^2$ | 0.67 | to 2 decimal places |

This highlights the drawback of relying solely on summary statistics and the how visualization is imperative to understand the

(Image Source: Wikipedia)

## Question 3

## What is Pearson's R? (3 marks)

Pearson's R or Pearson correlation coefficient measures the linear correlation between two variables. It ranges from -1 to 1:
R = 1, which is Perfect positive correlation.
R = -1, which is Perfect negative correlation.
R = 0, which is No linear correlation.

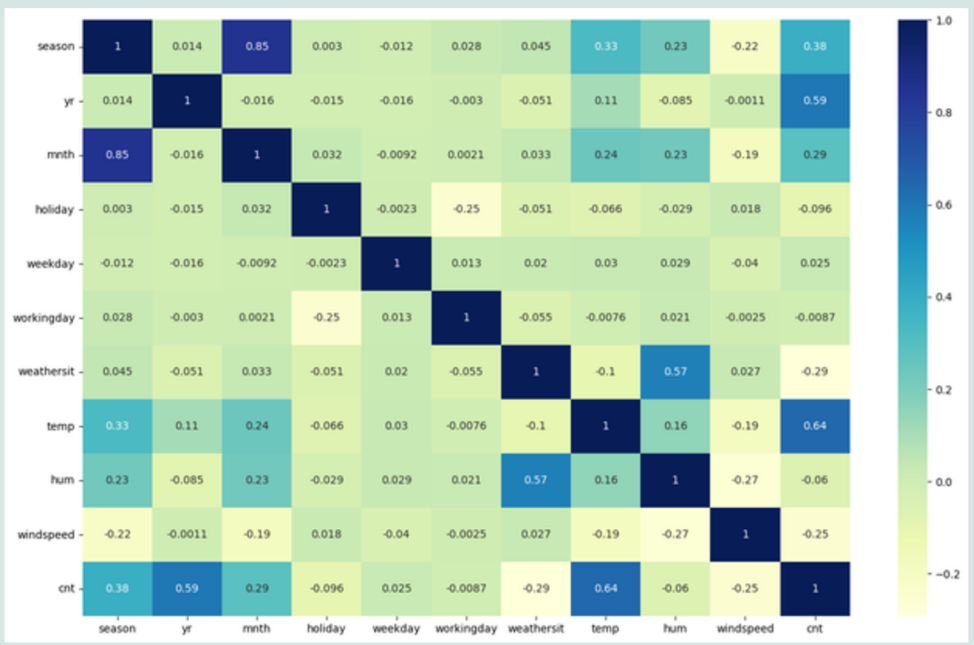
$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- $\mathrm{Cov}(X,Y)$: Covariance between $X$ and $Y$.
- $\sigma_X$: Standard deviation of $X$.
- $\sigma_Y$: Standard deviation of $Y$.

The heatmap in the BoomBike Sharing shows the correlation between various features.

In below image the tiles with darker shade means highly positive correlated and conversely, the tiles with lighter shade as highly negative correlated.




## Question 4

**What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?(3 marks)**

**Scaling**: Adjusting the range of feature values to ensure they are comparable, often necessary for machine learning algorithms sensitive to scale (e.g., gradient descent or distance-based models).
They are used mainly for below reasons:

- Prevents features with larger magnitudes from dominating.
- Speeds up convergence in optimization algorithms.
- Improves model performance and interpretability.

Types:

1. Normalized Scaling also called Min-Max scalar.
Rescales data to range between 0 and 1 or [-1, 1].
Used when absolute differences are relevant.

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where:

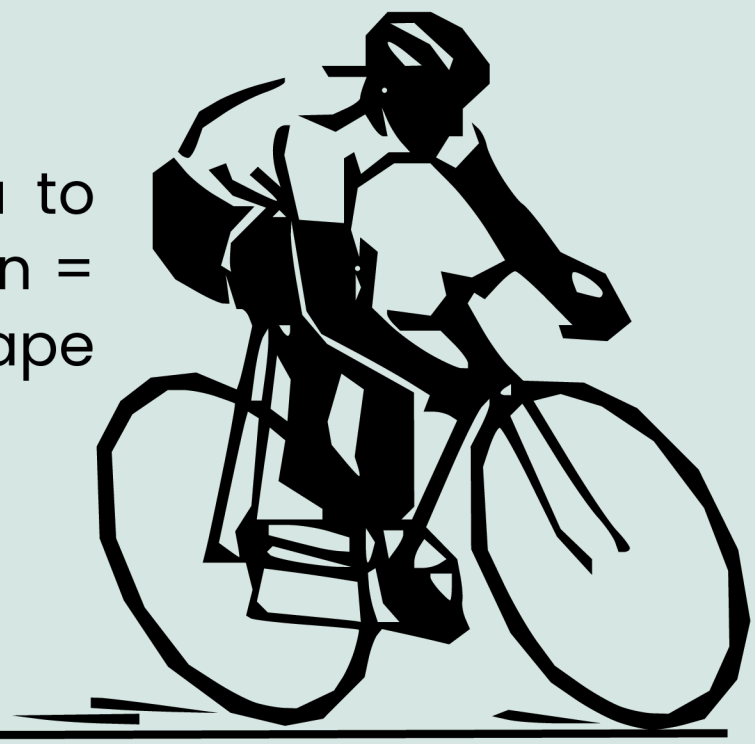
- $x$: Original value
- $x_{\min}$: Minimum value of the feature
- $x_{\max}$: Maximum value of the feature

2. Standardized Scaling rescales data to have mean = 0 and standard deviation = 1. Used when the distribution shape matters or outliers exist.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- $x$: Original value
- $\mu$: Mean of the feature
- $\sigma$: Standard deviation of the feature



## Question 5

**You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)**

VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity among predictor variables. This occurs when:
1. One predictor is a perfect linear combination of one or more other predictors.
2. The denominator of the VIF formula R-Squared becomes 0. That is R-Squared becomes 1.

For example in the boombikes dataset:
Firstly,

**cnt = Causual Users + Registered Users**

Secondly, **temp** and **atemp** are perfectly correlated.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

## Question 6

**What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)**

A Q-Q plot (Quantile-Quantile plot) compares the distribution of residuals from a model to a theoretical normal distribution. It plots the quantiles of residuals against the quantiles of a normal distribution.

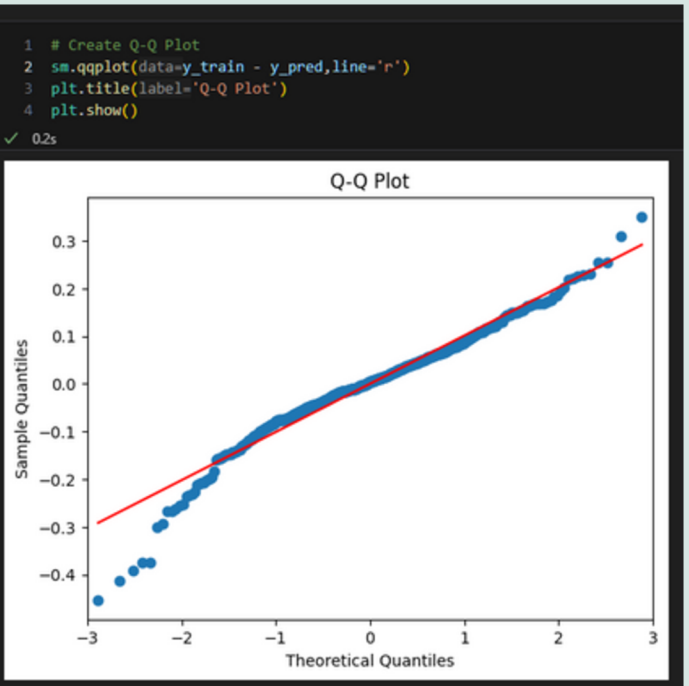
Checks if residuals are approximately normally distributed, a key assumption for linear regression.

Points aligning on the diagonal indicate normality. Deviations suggest skewness, kurtosis, or other distribution issues.

Q-Q plot helps validate assumptions and ensures the model's reliability for inference.

```
# Create Q-Q Plot
sm.qqplot(data=y_train - y_pred,line='r')
plt.title(label='Q-Q Plot')
plt.show()
```

✓ 0.2s




# THANK YOU