# 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? (Do not edit)
**Total Marks**: 3 marks (Do not edit)
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

Based on the analysis of the categorical variables in my regression model, here are the effect on the dependent variable (`cnt`, i.e., bike demand):

1. *Bike demand in the fall is the highest, indicating that this season sees peak usage.
   - Bike demand takes a dip in spring , suggesting lower demand during this season.

2. Bike demand in the year 2019 is higher compared to 2018 , indicating an upward trend in bike usage over time.

3. Bike demand is high from May to October , likely due to favorable weather conditions during these months.

4. Weather Condition:
   - Bike demand is high when the weather is clear or misty/cloudy .
   - Bike demand drops significantly when there is light rain or light snow , showing that adverse weather conditions negatively impact bike usage.

5. The demand for bikes is almost similar throughout the weekdays , suggesting no significant variation in usage between different days of the week.

6. Bike demand doesn't change significantly whether it's a working day or not , implying that both leisure and commuting purposes drive bike rentals equally.

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)
Using `drop_first=True` when creating dummy variables is important because it helps avoid a problem called the dummy variable trap. This happens when all the categories of a variable are converted into dummy variables, and one of them can be predicted from the others.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
**Total Marks:** 1 mark (Do not edit)
**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)
Temp variable has the highest correlation with the target variable

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)
After building the linear regression model on the training set, I checked its assumptions as follows:

1. Linearity: I ensured the relationship between the predictors and the target variable is linear by plotting residuals (errors) vs. predicted values. There should be no clear pattern in the plot.

2. Normality of Residuals: I checked if the residuals are normally distributed using a histogram or Q-Q plot. The residuals should look like a bell curve or follow a straight line in the Q-Q plot.

3. Homoscedasticity: I verified that the spread of residuals is consistent across all predicted values by plotting residuals vs. predicted values. There should be no funnel-shaped patterns.

4. No Multicollinearity: I checked if predictors are not highly correlated with each other using Variance Inflation Factor (VIF). High VIF values indicate multicollinearity, and such features were removed if necessary.

5. Independence of Errors: I ensured that residuals are independent using the Durbin-Watson statistic, where a value close to 2 confirms no autocorrelation.

These checks ensured that the model met all assumptions and was reliable for predictions.

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)
Based on the final model, the top 3 features contributing significantly to explaining the demand for shared bikes are:

1. T**emperature** (atemp): Higher temperatures positively influence bike demand, as favorable weather encourages outdoor activities like cycling.

2. **Year** (yr): Bike demand in the second year (e.g., 2019) is significantly higher compared to the first year (e.g., 2018), indicating an increasing trend in usage over time.

3. **Weather Conditions**: Clear or misty weather leads to higher bike demand, while adverse conditions like light rain or snow negatively impact usage.

These features have the strongest influence on bike demand based on their statistical significance and coefficients in the regression model.

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a method used to predict a dependent variable (`y`) based on one or more independent variables (`x`). It assumes a straight-line relationship between the variables and uses the equation:

$$ y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \varepsilon $$

Here:
- `β0` is the intercept.
- `β1, β2, ...` are coefficients showing how each predictor affects `y`.
- `ε` is the error (difference between actual and predicted values).

The goal is to find the best-fit line by minimizing the sum of squared errors (Ordinary Least Squares). After fitting, predictions can be made using the equation.

Assumptions:
1. The relationship between variables is linear.
2. Residuals (errors) are normally distributed and have constant variance.
3. Predictors are not highly correlated (no multicollinearity).

Evaluation:
- R-squared: Measures how well the model explains variability in `y`.
- Adjusted R-squared: Adjusts for the number of predictors.
- P-values : Show whether predictors significantly impact `y`.

Linear regression is simple, interpretable, and widely used for prediction and understanding relationships between variables.

---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a group of four datasets created by statistician Francis Anscombe

in 1973 to demonstrate the importance of visualizing data rather than relying solely on summary statistics. Each dataset contains 11 points and shares nearly identical summary statistics, including the mean, variance, correlation coefficient, and linear regression line. However, when plotted, the datasets reveal dramatically different distributions and patterns.

# Key Features of Anscombe's Quartet:
1. Identical Summary Statistics :
- Each dataset has the same mean and variance for both $$x$$ and $$y$$, the same correlation coefficient ($$r$$), and the same linear regression equation.
- Despite these similarities, the datasets are fundamentally different in structure.

2. Different Visual Patterns :
- Dataset 1 : A linear relationship that fits the regression model well.
- Dataset 2 : A linear trend with one outlier that significantly affects the regression line.
- Dataset 3 : A non-linear (quadratic) relationship that cannot be captured by a linear regression model.
- Dataset 4 : Most points are constant, with one extreme outlier that distorts the regression line.

# Purpose of Anscombe's Quartet:
- Importance of Data Visualization : It highlights how relying solely on summary statistics can be misleading. Visualizing data through scatter plots reveals patterns, outliers, and relationships that numbers alone cannot show.
- Limitations of Linear Regression : It demonstrates that linear regression is only suitable for data with a linear relationship. Non-linear patterns or outliers can invalidate the model's assumptions.
- Exploratory Data Analysis (EDA) : The quartet emphasizes the need for thorough data exploration before applying statistical or machine learning models.

Lessons from Anscombe's Quartet:
1. Always visualize your data to understand its underlying structure.
2. Summary statistics like mean, variance, and correlation are not sufficient to describe a dataset fully.
3. Outliers and non-linear relationships can significantly impact model performance and interpretation.

Anscombe's Quartet serves as a reminder that behind every dataset lies a story best understood through a combination of numerical analysis and visualization.

Sources

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is a number that tells us how strong and in what direction two continuous variables are related. It measures the linear relationship between them and ranges from -1 to +1:

- +1 : Perfect positive relationship (when one variable increases, the other also increases).
- -1 : Perfect negative relationship (when one variable increases, the other decreases).
- 0 : No linear relationship (the variables are not connected in a straight-line way).

# Example:
If we measure the relationship between study hours and exam scores:
- A Pearson's R of +0.8 means that more study hours lead to higher exam scores (strong positive relationship).
- A Pearson's R of -0.6 means that as study hours increase, exam scores decrease (negative relationship).
- A Pearson's R of 0 means study hours and exam scores have no connection.

It's important to remember that Pearson's R only shows correlation , not causation —just because two things are related doesn't mean one causes the other!

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is the process of transforming data so that it fits within a specific range or scale. It ensures that all features in a dataset have comparable magnitudes, which is particularly important when features have different units or ranges. For example, scaling can transform data values to lie between 0 and 1 or standardize them to have a mean of 0 and a standard deviation of 1.
Why is Scaling Performed?
Scaling is performed for several reasons:

1. Improves Model Performance: Many machine learning algorithms (e.g., gradient descent-based models like linear regression, logistic regression, and neural networks) are sensitive to the scale of data. Scaling ensures faster convergence and better performance.
2. Prevents Dominance by Larger Features: Features with larger ranges can dominate smaller ones in distance-based algorithms like K-Nearest Neighbors (KNN) or clustering.
3. Ensures Fair Comparisons: Scaling allows features with different units (e.g., age in years vs. income in dollars) to be compared on the same scale.
4. Improves Interpretability: Scaled data is easier to interpret and visualize.

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The value of VIF (Variance Inflation Factor) becomes infinite when there is perfect

multicollinearity among the independent variables in a regression model. This means that one variable can be completely predicted as a linear combination of other variables. In such cases, the regression model cannot distinguish the unique contribution of the collinear variable, leading to an infinite VIF value.
Reasons for Infinite VIF:

1. Duplicate or Identical Columns: If two or more columns in the dataset are identical, they will have perfect multicollinearity.
2. Linear Dependence: If one variable is a perfect linear combination of other variables, it results in infinite VIF.
3. More Variables than Observations: If the number of predictors exceeds the number of observations, it can lead to perfect multicollinearity and infinite VIF values.

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

---

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the quantiles of a dataset against the quantiles of a theoretical distribution (e.g., normal distribution). It helps determine if the data follows a specific distribution. If the data matches the theoretical distribution, the points on the Q-Q plot will align closely along a 45-degree diagonal line.

Use and Importance in Linear Regression

In linear regression, a Q-Q plot is primarily used to check the normality assumption of residuals. The normality of residuals is crucial because it ensures that statistical tests (like t-tests and F-tests) used to evaluate the model's coefficients are valid.
How It's Used:
1. Generate Residuals : After fitting the regression model, calculate the residuals (differences between observed and predicted values).
2. Create the Q-Q Plot : Plot the quantiles of these residuals against the quantiles of a normal distribution.
3. Interpretation :
   - If residuals follow a normal distribution, points will fall along a straight diagonal line.
   - Deviations from this line indicate non-normality, suggesting issues like skewness or heavy tails in the residuals.