## Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   We could infer the following points from analysing categorical variables with 'cnt' the dependent variable:

   - The number for bike rentals (cnt) increases in the month of June to Sept which is a summer season in US
   - The number for bike rentals (cnt) is least in Spring and highest in Fall.
   - There is not much significant effect on cnt w.r.t days of the week. Hence the rentals are almost consistent throughout the weekdays.
   - The bike rentals are consistent across working day and non-working day.
   - The bike rentals are more on a day which is not a holiday. So may be people use it as a commute option to reach their workplace.
   - The bike rentals are more when weather situation is clear.
   - The number of rentals has gone up from year 2018 to 2019.

   As per our MLR final model categorical data such as year, spring, winter, light snow/rain, mist, July are significant for the prediction.

2. Why is it important to use drop_first=True during dummy variable creation?

   'drop_first=True' is important to use while creating dummy variables as this option will **drop the first category in each categorical variable**. Thus, if a category has n levels, the total number of dummy columns created would be **n-1**. This is done to avoid the dummy variable trap, which is **the issue of multicollinearity** in linear regression models where one category can be perfectly predicted by the others.

   By default, this option is set to false. However, in situations when we need to reduce the number of columns by dropping the column not necessary, we set this option to True.

   For example,

```
import pandas as pd

# Creating a DataFrame with the categorical variable
data = {'furniture_status': ['furnished', 'semifurnished', 'unfurnished']}

df = pd.DataFrame(data)

# Creating dummy variables with drop_first=True
dummies = pd.get_dummies(df['furniture_status'], drop_first=True)

# Printing the resulting DataFrame with dummy variables
```

print(dummies)

Output:

| semifurnished | unfurnished |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 0 | 1 |

In this case, the furnished level is dropped, and dummy variables are created for semifurnished and unfurnished. The first row of the original DataFrame (furnished) is represented by [0, 0], the second row (semifurnished) is represented by [1, 0] and unfurnished by [0, 1].

## 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

**temp** and **atemp** are the numerical variables that shows the highest correlation with the target variable (cnt)

## 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Below points are considered to validate the assumptions of Linear Regression after building the model on the training set:

- **Linearity**: Residuals vs Fitted values plot validates Linearity of the final model as the residuals are randomly scattered around zero. Hence the linearity assumption holds.
- **Residual Analysis**: The distplot of residuals shows a normal distribution like graph with most of the points around the mean of 0.
- **R2 score** and **adjusted R-squared** of our prediction on train set are comparable with an approximate value of **82%** and that on test set **79%.**
- The **p-values** of all our predictor values are **0** indicating all the predictor variables in the model are **significant**.
- The **VIF (Variance inflation factor)** value all the predictor variables in the model are **less than the cut off value of 5**. Hence indicating that there is no major multicollinearity issue between the predictor variables that could impact the efficiency of the model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top 3 features based on absolute values of coefficients: Index(['temp', 'Light Snow', 'yr'], dtype='object')

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression is the **supervised Machine Learning model** in which the model finds the **best fit linear line** between the independent (predictors) and dependent variable (target variable).
- There are 2 types of linear regression – **Simple linear regression (SLR) and Multi linear regression (MLR)**
- Simple Linear Regression involves only one independent variable the model has to find the linear relationship of it with the dependent variable.
- In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
- Linear regression assumes that the relationship between the dependent variable y and the independent variable X is linear. The equation of a simple linear regression (with one predictor) is: $y=\beta0+\beta1X+\epsilon$
  where:
  y is the dependent variable.
  X is the independent variable.
  β0 is the intercept.
  β1 is the slope (coefficient) of the independent variable.
  ϵ is the error term, representing the deviation of the observed values from the true regression line.
- The objective of linear regression is to find the values of β0, β1,…, βp that minimize the sum of the squared errors (SSE) or MSE, where the error is the difference between the observed and predicted values of y.
- Methods such as **gradient descent, OLS, Mathematical Derivation and optimization of cost function** are used to minimize the error and find the coefficient for the best fit.
- Libraries such as sklearn and statsmodel in python help in LR predictions.
- **The steps involved in linear regression are as follows**:
  1. Reading and understanding the data
  2. Data Cleaning
  3. Data Visualization and EDA

4. Data pre-processing – Scaling, hot-encoding, splitting data into train and test data set.
5. Model building using sklearn or statsmodel
6. Residual analysis
7. Evaluate model on train set.
8. Improve the model
9. Prediction on test data set
10. Evaluate model on test set.
11. Validate the model

- Linear regression relies on assumptions such as
  - Linearity: The relationship between the predictors and the response variable is linear.
  - Independence: The residuals (errors) are independent.
  - Homoscedasticity: The residuals have constant variance at every level of X
  - Normality: The residuals are normally distributed.
- Following are the parameters used to evaluate model's performance and validate the assumptions:
  - R-squared: Represents the proportion of variance in the dependent variable that is predictable from the independent variables.
  - Adjusted R-squared: Adjusted for the number of predictors, providing a more accurate measure when comparing models with different numbers of predictors.
  - Residual Analysis: Check plots of residuals to validate assumptions of linearity, independence, and homoscedasticity.
  - p-values: Test the significance of individual predictors.
  - F-statistic: Tests the overall significance of the model

## 2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
- It indicates that even though the datasets have a very similar statistical summary, it is possible that these datasets have a very different distribution.
- Hence, it is important to analyse data graphically as well and not just rely on the statistics properties such as mean, variance, correlation or linear regression lines.
- Anscombe's quartet demonstrates several important principles in statistics and data analysis:
  - **Graphical Analysis**: Always visualize the data. Simple summary statistics can be misleading and do not capture the full story of the data distribution and relationships.

- **Outliers**: The presence of outliers can have a significant impact on statistical analyses and can distort relationships. Identifying and understanding outliers is crucial.
- **Nonlinearity**: Real-world data often do not follow simple linear patterns. It's essential to check for nonlinear relationships and apply appropriate models.
- **Context**: Understanding the context and the nature of the data is essential. The same statistical summary can imply very different things depending on the context.

## 3. What is Pearson's R?

- Pearson's R, also known as the Pearson correlation coefficient quantifies the **strength and direction of the linear relationship between the variables.**
- The value of R ranges between -1 and 1:
- If R=1: Perfect positive linear correlation.
- If R=−1: Perfect negative linear correlation.
- If R=0: No linear correlation, meaning there is no linear relationship between the variables. Note that R=0 does not imply no relationship at all; there could be a nonlinear relationship.
- Values of R closer to 1 or -1 indicate stronger linear relationships, while values closer to 0 indicate weaker linear relationships.
- Its library can be imported as from scipy.stats import pearsonr in python.
- It assumes that the analysed variables are linearly related, continuous and normally distributed, no outliers and data is homoscedastic.
- It is a fundamental measure in statistical analysis and is often used as a basis for further analysis, such as regression.

## 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is the process of transforming data so that it fits **within a specific range**. It is one of the steps in **data pre-processing** phase of data modelling.
- It is done to ensure that the different features contribute equally to the model, preventing any one feature from dominating due to its scale.
- Scaling is performed for the following reasons:
  1. **Data/Model Interpretability** - Scaling can make the coefficients of the model more interpretable by putting them on a common scale.
  2. **Model Efficiency** – To avoid misleading predictions and perform better and faster.
  3. **Reducing Computational Complexity** - When features are on different scales, the optimization algorithms may take longer to converge.
- Difference Between Normalized Scaling and Standardized Scaling are as follows:

1. **Min-Max Scaling** is a normalized scaling technique where data is transformed to fit within the range of [0,1] or [-1,1]
2. It's formula is x-min(x)/max(x)-min(x), where x is the original value.
   Pros - Preserves the relationships between the values, useful when the data is known to have a bounded range.
   Cons - Sensitive to outliers because they affect the min and max values.

- **Z-score Scaling** is a Standardized scaling technique that transforms the data to have a mean of 0 and a standard deviation of 1.

  Its formula is x-mean/variance, where x is the original value.

  Pros - Not sensitive to outliers as normalization, Useful when the data follows a normal distribution.

  Cons- Does not bound the values to a specific range.

## 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of VIF can be infinite in cases where there is a **perfect multicollinearity**. This occurs when there is an exact linear relationship between one predictor variable and one or more other predictor variables in the model.

The VIF for an independent variable Xi is calculated as:

VIF(Xi)= 1/(1-Ri square) where Ri square is the coefficient of determination of the regression of Xi on all the other predictors. In case of perfect multicollinearity Ri square will be 1. Therefore 1/0 would be infinite.

This can lead to model instability hence needs to be stabilized by regularization techniques.

## 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a set of data follows a particular distribution, usually the normal distribution. It compares the quantiles of the data against the quantiles of a theoretical distribution.
- Construction of a Q-Q Plot
  - Quantiles of Data: Sort the data and calculate the quantiles.
  - Quantiles of Theoretical Distribution: Calculate the corresponding quantiles of the theoretical distribution (e.g., normal distribution).
  - Plotting: Plot the data quantiles on the y-axis and the theoretical quantiles on the x-axis. If the data follows the theoretical distribution, the points should roughly form a straight line.

- Use and Importance in Linear Regression
    - In linear regression, a Q-Q plot is used primarily to check the normality of the residuals. The assumptions of linear regression include that the residuals (errors) are normally distributed.
    - The normality of residuals is helps to validate the model. Non-normal residuals may indicate problems like omitted variables, incorrect functional forms, or heteroscedasticity.
    - Validating Statistical Inferences