## 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)

---

- Season could be a key predictor for the dependent variable as season 3 (fall) saw a median of approx 5000+ bookings
- Holiday cannot be a good predictor as approx 98% bookings happened on non-holidays
- Month can be good predictor as the data indicates bookings in months 5,6,7,8 & 9 have a median of greater than 4000 bookings
- Weathersit shows some trend towards bookings as Clear weather (indicated by 1) with a median approx 5000, account for 69% bookings

---

**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)

---

- When creating dummy variables (for e.g., creating n dummy variables for a categorical variable with n levels), the dummy variables become perfectly correlated (multicollinearity) with each other, causing problems in regression analysis
- Dropping the first variable will essentially facilitate use the remaining variables as a reference against which the others are compared, effectively eliminating redundancy and allowing for proper interpretation of the model coefficients

---

**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)

---

- By checking the linear relationship between x & y
- By finding out whether the error terms are normally distributed
- By checking if there is no autocorrelation (using Durbin-Watson value – equal to 2)
- By validating if no multicollinearity (VIF) among predictor variables

---

**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)

---

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.5772' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5772 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2769' indicated that (with respect to weathersit_1, a unit increase in weathersit_3 variable decreases the bike hire numbers by -0.2769 units.
- Year (yr) - A coefficient value of '0.2334' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2334

---

## 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>

---

- Regression Analysis Algorithms are used to determine the strength of relationship between variables
- Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors)
- Linear regression shows the linear relationship, which means it findshow the value of the dependent variable is changing according to the value of the independent variable
- If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression
- Linear regression comprises a straight line that splits the data points on a scatterplot. The goal of linear regression is to split the data in a way that minimizes the distance between the regression line and all data points on the scatterplot
- If one were to draw a vertical line from the regression line to each data point on the graph, the aggregate distance of each point would equate to the smallest possible distance to the regression line

---

**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 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
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths
- Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line
- Anscombe's quartet is used to show the importance of EDA and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone

---

**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>

- The Pearson correlation coefficient (the full name is Pearson's Product Moment correlation - PPMC), often symbolized as (R), is a widely used metric for assessing linear relationships between two variables
- It yields a value ranging from –1 to 1, indicating both the magnitude and direction of the correlation. A change in one variable is mirrored by a corresponding change in the other variable in the same direction
- Pearson's correlation helps in measuring the correlation strength (given by coefficient r-value between -1 and +1) and the existence (given by p-value) of a linear correlation relationship between the two variables and if the outcome is significant, it confirms that the correlation exists

---

**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 (also known as Feature Scaling) is a data preprocessing technique

in which the values of independent variables in a dataset are transformed to a similar range, ensuring that all features contribute equally to a machine learning model

- When features in a dataset have vastly different ranges, without scaling, the feature with a larger range could disproportionately influence the model's predictions, leading to inaccurate results
- So there is a need for scaling because of 2 reasons – Ease of interpretation and Faster convergence for gradient descent methods
- Difference between Normalized Scaling
    - In Standardized scaling, the variables are scaled in such a way that their mean is zero and standard deviation is one. In normalized scaling (MinMax scaling), the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data
    - Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation
    - Normalized scaling scales values between [0, 1] or [-1, 1], whereas standardized scaling is not bound by any range
    - Normalized scaling is very much affected by outliers, whereas Standardized scaling is impacted less by outliers

---

**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>

- When there is a perfect correlation between 2 independent variables, then VIF value is infinite
- To explain further, we get R-squared (R2) =1 when there is perfect correlation, which lead to 1/ (1-R2) and this works out to be infinity

---

**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>

- Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other
- The quantile-quantile (q-q plot) plot helps in determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not
- It is used for the following purposes
    - Assessing Distributional Assumptions: Q-Q plots are frequently used to visually inspect whether a dataset follows a specific probability distribution, such as the

normal distribution. By comparing the quantiles of the observed data to the quantiles of the assumed distribution, deviations from the assumed distribution can be detected. This is crucial in many statistical analyses, where the validity of distributional assumptions impacts the accuracy of statistical inferences.

- Detecting Outliers: Outliers are data points that deviate significantly from the rest of the dataset. Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution. Outliers may appear as points that deviate from the expected straight line in the plot.
- Comparing Distributions: Q-Q plots can be used to compare two datasets to see if they come from the same distribution. This is achieved by plotting the quantiles of one dataset against the quantiles of another dataset. If the points fall approximately along a straight line, it suggests that the two datasets are drawn from the same distribution.
- Assessing Normality: Q-Q plots are particularly useful for assessing the normality of a dataset. If the data points in the plot closely follow a straight line, it indicates that the dataset is approximately normally distributed. Deviations from the line suggest departures from normality, which may require further investigation or non-parametric statistical techniques.
- Model Validation: In fields like econometrics and machine learning, Q-Q plots are used to validate predictive models. By comparing the quantiles of observed responses with the quantiles predicted by a model, one can assess how well the model fits the data. Deviations from the expected pattern may indicate areas where the model needs improvement.
- Quality Control: Q-Q plots are employed in quality control processes to monitor the distribution of measured or observed values over time or across different batches. Departures from expected patterns in the plot may signal changes in the underlying processes, prompting further investigation