# 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)

With the analysis done for the categorical variables in the dataset using boxplots and bar plots. The following insights can be drawn from the visualizations:

- **Seasonal Influence:** The fall season attracted the highest number of bookings. Additionally, across all seasons, booking counts showed a significant increase from 2018 to 2019.
- **Monthly Trends:** Most bookings occurred between May and October, with a noticeable upward trend in bookings from the beginning of the year to mid-year, followed by a decline towards the year's end.
- **Weather Impact**: Clear weather conditions correlated with higher booking counts, which is expected.
- **Day of the Week**: Bookings were notably higher on Fridays, Saturdays, and Sundays compared to the start of the week.
- **Holidays:** Booking counts were lower on non-holidays, which aligns with the assumption that people prefer to stay home and spend time with family on holidays.
- **Working vs. Non-Working Days**: Bookings were consistent on both working and non-working days.
- **Yearly Growth**: The year 2019 saw a considerable increase in bookings compared to 2018, indicating positive growth in business performance.

---

**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)

**drop_first=True**, this indicates it drops one Dummy Variable from all the dummy variables created for a feature. It is used to drop some of the features which are highly correlated which results in predicting of another feature.

Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

---

**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)

One can validate the assumptions of Linear Regression after building the model on the training set using following below assumptions:

- Linear relationship validation – Linearity should be visible among variables
- Normality check – Error terms should be normally distributed
- Multicollinearity – There should be insignificant multicollinearity among variables
- Homoscedasticity – There should be no visible pattern in residual values
- Independence of residuals – No auto-corelation

---

**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)
The top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. temp
2. yr
3. winter

---

# 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)

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (yy) and one or more independent variables (xx).

- **Objective**: Minimize the difference between the predicted (y^) and actual values (y) by fitting a linear equation $y=\beta0+\beta1x1+\cdots+\beta nxn+\epsilon$
- **Process**:
    1. Assumes a linear relationship between x and y.
    2. Estimates coefficients (ββ) using techniques like **Ordinary Least Squares (OLS)**, minimizing the sum of squared residuals.
    3. Makes predictions by applying the learned coefficients to new data.
- **Output**: A straight line (or hyperplane in multivariate cases) representing the best fit through the data.

It is simple, interpretable, and suitable for relationships where variables have a linear correlation.

---

**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)

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, such as mean, variance, correlation, and regression line, but differ significantly when visualized.
Anscombe's quartet highlights the limitations of relying solely on summary statistics and emphasizes the necessity of visualizing data to identify patterns, outliers, and anomalies.
Key Insights:

1. Identical Statistics: All four datasets have similar descriptive statistics, including:
    - Mean and variance of x and y.
    - Correlation between x and y.
    - Regression equation ($y = mx+c$).
2. Different Visual Patterns: When plotted:
    - Dataset 1 shows a linear relationship.
    - Dataset 2 has a non-linear relationship.
    - Dataset 3 contains an outlier that influences the regression line.
    - Dataset 4 has most points concentrated at a single value with one influential point.

---

**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)

Pearson correlation coefficient, measures the strength and direction of a linear relationship between two variables. It ranges from $-1$ to 1:

- $r = 1$: Perfect positive correlation.
- $r = -1$: Perfect negative correlation.
- $r = 0$: No linear correlation.

It is calculated as the covariance of the variables divided by the product of their standard deviations, reflecting how changes in one variable are associated with changes in 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)

Scaling is the process of transforming data to a specific range or distribution to ensure features contribute equally to a model, especially for algorithms sensitive to feature magnitudes (e.g., SVM, KNN).

Reasons for Scaling:

- Prevent features with larger magnitudes from dominating.

- Improve model performance and convergence (e.g., in gradient descent).

Difference Between Normalized and Standardized Scaling:

- Normalization: Scales data to a range, typically [0, 1] or [-1, 1]. It is suitable when data follows no particular distribution.
  $xn = x - \min(x)/(\max(x) - \min(x))$
- Standardization: Centres data around the mean with a unit variance. It is ideal when data follows a Gaussian distribution.
  $xs = x - \mu/\sigma$

Normalization focuses on range, while standardization focuses on distribution.

---

**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)

The value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between one predictor and other predictors in the dataset.
It happened due presence of highly correlated predictors in the data.

---

**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)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of data against a theoretical distribution (e.g., normal distribution). It plots the quantiles of the dataset against the quantiles of the theoretical distribution.
Use in Linear Regression:

To check if the residuals of the model follow a normal distribution, which is a key assumption of linear regression.
Interpretation:
1. points forming a straight line indicate normality
2. Deviations suggest non-normality, potentially impacting model validity.

Importance:
It helps to detect skewness, kurtosis, or outliers in the residual, ensuring the model's assumptions hold true for reliable predictions and interpretations.

---