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?**

   Answer: i have done analysis based on categorical variable like mnth, weekdays, season, holiday, weathersit, workingday and weekday with count and year .here is what my understanding says

   - Fall season seems to have attracted more booking from 2018 to 2019.
   - Most of the bookings has been done during the month of may, june, july, aug, sep and oct.
   - When its not holiday, booking seems to be less in number which seems reasonable as on holidays
   - Clear weather attracted more booking
   - Booking seemed to be almost equal for working day or non working day

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

   Answer: Setting **drop_first =false** in **pd.dummies()** it drops the first category from the categorical variable to avoid the "dummy variable trap." This trap occurs when you include all dummy variables for a categorical feature, which can cause problems in regression model.

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

   Answer: **temp** variable has high correlation with other variables.

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

   Answer: Below are Assumptions are

   Normality of error term: Error term distributed normally.
   Multicollinearity check: there should be insignificant multicollinearity among variables
   Linearlity: Linearity should there among the variables
   Homoscedasticity :There should be no visible patterns
   Independent of residual: No autocorrelation

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

   Answer: Here are top three variables **'Temp'**,**'Winter'** and **'Light_snowrain'**.

General Subjective Questions

6) Explain the linear regression algorithm in detail.

Answer: Linear regression is a fundamental staticstical model used to build relationships with dependent variables to one or more independent variables. Its used in fields like Economics, biology, Space technology, due to its simplicity.

Formula is ***y=mx+c***

Types are

- **Simple Linear Regression**: Involves one independent variable. The model fits a straight line to the data points.
- **Multiple Linear Regression**: Involves Two or more independent variables. The model fits a straight line to the data points.

Assumptions of linear regression:

- **Normality of error term**: Error term distributed normally
- **Multicollinearity check**: there should be insignificant multicollinearity among variables
- **Linearlity**: The relationship between the dependent and independent variables should be linear
- **Homoscedasticity** :The residuals (errors) should have constant variance at all levels of x
- **Independent of residual**: Observations should be independent; errors from one observation should not influence others.

7) Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet is a collection of four datasets that illustrate the importance of data visualization in statistical analysis. Each dataset in Anscombe's Quartet consists of eleven (x, y) data points. Despite having nearly identical statistical properties, such as means, variances, and correlation coefficients, the datasets reveal different patterns

Dataset1:
Description: This dataset shows a linear relationship between x and y, fitting the linear regression model well.

Plot: Appears as a straight line.

Dataset2:

Description: This dataset exhibits a non-linear relationship. While there is a relationship between x and y, it cannot be captured by a simple linear regression.

Plot: The outlier skews the results, leading to a misleadingly low correlation coefficient.

Dataset3:

Description: Similar to Dataset I, but influenced by an outlier that significantly affects the regression line.

Plot: The outlier skews the results, leading to a misleadingly low correlation coefficient.

Dataset4:

Description: Contains one high-leverage point that dramatically increases the correlation coefficient despite the other points showing no clear relationship.

Plot: Most points are clustered with one point far away, affecting the overall analysis.

8) What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol R *r* and ranges from -1 to 1. A perfect positive linear relationship; as one variable increases, the other variable also increases proportionally.
A perfect negative linear relationship; as one variable increases, the other decreases proportionally. 0 indicates no linear relationship between the variables.

9) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is a crucial preprocessing step in data analysis and machine learning that involves transforming the range of features in a dataset to ensure they are on a similar scale.

Scaling is performed because

Uniformity: If features have different ranges, those with larger scales can disproportionately influence the results.

Convergence Speed: Algorithms that use gradient descent, such as linear regression, converge faster when features are scaled.

Outlier management: Scaling can help mitigate the impact of outliers by reducing their influence during model training.

Difference:

Feature	**Normalization**	**Standarization**
Range	[0,1]	Mean =0 ,SD=1
Outliers	High	Lower
Use case	Best for bounded distribution	Best for normal distribution

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

Answer: The occurrence of infinite Variance Inflation Factor (VIF) values typically arises from perfect multicollinearity among the independent variables in a regression model. Infinite VIF values indicate that one or more independent variables are perfectly linearly dependent on others. This means that one variable can be expressed as an exact linear combination of others, leading to redundancy in the data.

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

Answer: A Q-Q plot, or quantile-quantile plot, is a graphical tool used in statistics to compare the quantiles of two probability distributions. It helps assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. A Q-Q plot, the quantiles of one dataset are plotted against the quantiles of another dataset or a theoretical distribution. Each point on the plot corresponds to a quantile from both distributions. If the two distributions being compared are similar, the points will approximately lie on the identity line y=x. This indicates that the distributions have similar shapes, scales, and locations.