Assignment-based Subjective Questions

Q1. The categorical variables in the dataset are season, holiday, working day, weathersit, month, weekday and year. From the final model built we can analyze that most favourable seasons for biking are winters and summers. Weather conditions like Cloudy, and Light Snow/Rain seems to be impacting the business of rental bikes negatively. Among the months, starting from November to February there are less demand for the rental bikes due to cold/harsh weather forcing people to stay indoors. Holidays also seem to be impacting the business in a negative way with not much increase in demand of bikes. Compared to 2018, there has been substantial increase in the demand of rental bikes in 2019.

Q2. Using drop_first=True during dummy variable creation in the pd.get_dummies is important for avoiding multicollinearity in regression models. Multicollinearity occurs when two or more predictor variables in a model are highly correlated, leading to unreliable coefficient estimates and inflated standard errors.

Q3. temp has the highest correlation with the target variable (cnt) at 0.63. The casual and registered are part of the target variable as sum of these two column variables totals the cnt, so ignoring the correlation between cnt and registered/casual. atemp is also derived from temp which got eliminated in the model preparation step hence it is not considered.

Q4. We can validate assumption of Linear Regression after building model on the Training set by:-

1) Linear Relationship between independent and dependent variables.
2) Errors terms should be independent of one another by plotting between Residuals vs predicted values.
3) Errors terms should be normally distributed
4) Errors terms should follow homoscedasticity (constant variance)

Q5. The top 3 features that strongly contributed towards expanding demand of shared bikes are:

1) Temperature affects the business positively.
2) Yr_2019: There has been increase in demand of shared bikes in 2019 compared to 2018.
3) Season specifically Winters and summers also affects business positively.

General Subjective Questions

Q1. Linear regression is a fundamental statistical method used in machine learning to model the relationship between a dependent variable (also called the target or response variable) and one or more independent variables (also called predictors or features). It assumes that there is a linear relation between target and predictors (explanatory variables). The 2 main types of linear regression are Simple Linear Regression and Multiple Linear Regression.

Linear Regression is used when a single predictor variable is used to predict the target variable. Multiple Linear Regression is used when multiple independent variables is used to predict the target variable. Some assumptions of Linear Regression are Linearity, Independence i.e observations are independent of one another, no multicollinearity between independent variables, residuals are normally distributed and homoscedasticity (constant variance across independent variables)

The most common method for fitting a linear regression model is the OLS method. It finds the coefficients that minimize the sum of squared residuals (the differences between observed and predicted values). For model evaluation, R squared Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating better fit.

For interpreting the coefficients of linear regression model, we have the slope and the intercept terms. Slope Represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. While intercept Represents the expected value of the dependent variable when all independent variables are zero.

Q2. Anscombe's quartet is a group of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and visual characteristics. These datasets were constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing, making conclusions and to show how statistics can be misleading sometimes if not interpreted carefully. So, we should not just look at describe() function but also make it a habit of making visualizations to understand the data rather than making assumptions from statistical results.

Q3. Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree or extent to which the two variables are related to one another providing both the strength and direction of the relationship. It lies between values -1 to 1 with -1 showing negative strong relationship, 0 meaning no relationship and 1 meaning strong positive relationship between the two variables.

Q4. Scaling is a preprocessing technique used in machine learning and data analysis to adjust the range of features in a dataset. It ensures that the features have a similar scale, which is crucial for algorithms that are sensitive to the relative magnitudes of the data. It is performed for many

reasons like a) Improves Algorithm Performance b) Enhance interpretability of the results of Machine Learning model c) Avoid Bias

Normalization scales the data to a fixed range, between 0 to 1. Normalization is typically used when the data follows a non-Gaussian distribution or when the algorithm requires the features to be bounded within a specific range, such as in neural networks. On the other hand, Standardization, scales the data such that the mean of each feature becomes 0, and the standard deviation becomes 1. Standardization centers the data and makes it unit variance but doesn't compress it to a fixed range, making it suitable for data that naturally follows a Gaussian distribution.

Q5. A VIF can also be infinite in situation when two variables are perfectly correlated and have R value of -1 or 1. In such cases, the denominator in the VIF calculation becomes zero, leading to an infinite VIF value. If there is perfect multicollinearity between two or more variables, it means that one variable can be perfectly predicted using a linear combination of the others.

Q6. QQ plots or the Quantile Quantile plot is a graphical tool to compare the distribution of a dataset to a theoretical distribution, most commonly a normal distribution.

There are various use cases where QQ plots are essential

a) One key assumption of linear regression is that the residuals are normally distributed. A Q-Q plot of the residuals can be used to check this assumption. If the residuals follow a normal distribution, the points on the Q-Q plot will lie approximately on a straight line
b) A Q-Q plot can reveal outliers, which are points that deviate significantly from the theoretical distribution.
c) If the points on a Q-Q plot curve away from the reference line, it can indicate skewness in the data. This information can guide the need for data transformations like log, square etc