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?

Windspeed an humidity have a negative effect on dependent variable i.e cnt
Weather 2 and weather 3 i.e. when weather is not clear then we have less bookings
Year i.e yr column has got a positive relationship with cnt column
Month 9 i.e September have got highest bookings
Season 2,season 3 and season 4 has got good number of bookings while
Holiday has a negative effect on booking , during holidays the bookings are lesser

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

This drops the first level of categorical variables when creating dummy variables which in turn reduces collinearity, and improves the model.

Examples take three levels for a categorical variable A, B and C. If we create dummy variables without drop_first = True then there will be 3 dummy variables created
A - 1 0 0
B - 0 1 0
C - 0 0 1
But if we remove one variable using drop_first = true then
B - 1 0
C - 0 1
Which means that A will always be 0 0 , so using 2 dummy variables we create 3 levels.
The formula is : number of dummy variable = levels - 1

This has been used for all categorical variables in the assignment.

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

temp variable among the numerical variables , seems to have the highest correlation with the cnt column.

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

1. Checked p values for all independent variables, and adjusted R square.

2. Created a distribution plot of residual errors b/w y_train and y_pred to check mean is 0 and standard deviation is 1.
3. Used .predict() method on test data to calculate adjusted R square. Compared adjusted R square on test and train dataset
4. Created a distribution plot of residual errors b/w y_test and y_test_pred to validated if mean is 0 and standard deviation is 1

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

yr, temp and season(specifically season 4 i.e during heavy rain)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm which computes a linear relationship b/w a dependent variable and one or more independent features. The output variable is a real or continuous value. It is used for prediction, forecasting, time series modeling

Equation : y = B0 + B1X1 + B2X2 .. + BNXN and so on

The goal of the linear regression algo is to find the best fit line by minimizing the cost function. Cost function helps in finding the best coefficients or weights of the independent variables(B0, B1, B2 …..). Cost function is a relationship b/w actual values and predicted values. Mean squared error is a type of cost function. Gradient Descent is a method widely used to minimize the cost function.

Simple Linear Regression, Multiple Linear Regression , Logistic Regression are types of linear regression models.

2. Explain the Anscombe's quartet in detail.

This is a combination of a set of 4 dataset, having identical statistical properties in terms of mean, variance, R-squared, correlations, and linear regression lines but have different visual representations when plotted on a scatter graph. This demonstrates the importance of visualizing the data before applying various ML algorithms to build the model, and not just relying on summary statistics as it can be misleading.
The 4 types of dataset are :

1. Fits linear regression model pretty well where x has a positive correlation with y
2. Data relation bw x and y is non-linear
3. An outlier is involved, apart from this one point there is a perfect relationship b/w x and y.
4. An outlier is involved, and also there is no relationship between x and y

3. What is Pearson's R?

Pearson's correlation coefficient is a statistical measure that quantifies the strength and direction of their linear relationship between two continuous variables. The values can range from -1 to +1. Anything close to -1 depicts a strong negative relationship between variables. Values close to +1 depict a positive relationship between variables. Values around 0 represent no relation.

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

It is a step used in data processing prior to running the models using machine learning algorithms. The idea is to normalize/standardize the data within a certain range so that all the independent variables have a similar scale so that none of the variables are prioritized in model creation.

We need scaling because the datasets used during ML analysis, generally have different magnitudes, units and range. If we feed the same data to ML algos, then the algorithms may take a long time to converge and may perform poorly because they are sensitive to the scales of features. To solve this, we perform scaling to bring all features to the same level of magnitude.

Normalized scaling transforms data in the range of 0 to +1, while standardized scaling replaces the values by their Z scores. It brings all the data into a standard normal distribution which has a mean of zero and standard deviation as one. Standardized scaling performs better when there are outliers involved in the dataset.

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

Variance Inflation factor is a measure of the amount of multicollinearity in regression analysis i.e. it represents how much the independent variables are dependent on each other. We see an infinite VIF because of a strong linear relationship between variables where one predictor variable can be perfectly predicted by a linear combination of the other predictor variables. In case of perfect relation, we get R2 = 1 which then leads to 1/(1-R2) to be infinite.

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

A Q-Q plot is a graphical tool used to assess if a set of data came from some theoretical distribution such as Normal, exponential, or uniform distribution. This is done by comparing the quantiles of sorted data against the quantiles of a theoretical distribution.

If all points of Q-Q plot fall on a straight line , it suggests that the residuals follow a normal distribution. Deviation from the straight line indicates non-normality which can be caused by outliers. This graph is used to provide a visual way to inspect the distribution of residuals in linear regression assisting in identifying non-normality and guiding towards achieving fine tuning in the model.