Assignment-based Subjective Questions

Q 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: All the categorical variables have some impact on the dependent variable i.e., bike bookings.

A significant increase is seen in bike bookings over the base year of 2018.

There is increase in bike bookings during summer and fall season which can be seen in the month-wise bike bookings as well. The months from May to September show increase in bookings. A similar pattern can be seen with weather situation. The bookings tend to increase on days with clear or partly cloudy sky. On the other hand, there is sharp decrease when it snows.

The bike bookings on a holiday have slightly higher mean than bookings on any other day. The mean bookings on any day of the week are more or less the same.

Q2: Why is it important to use drop_first=True during dummy variable creation?

Ans: For any categorical variable with n-levels, n-1 columns are sufficient to encode all the n levels. Therefore, while creating dummy variables, we use drop_first=True to drop the first column.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The variable temp has highest correlation with target variable.

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: The p-values of all the learned coefficients are zero suggesting a liner fit.
The residual analysis plot clearly shows normal distribution of the error terms. The error terms are centred around zero.

The distribution of error terms also shows constant variance with error terms nicely distributed on both sides of mean.

The scatter plot of predicted values to error terms shows that the error terms are randomly scattered with no specific pattern, thus we can say that the error terms are independent.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The following three features contribute significantly to the demand of shared bike

1. Temperature – it has positive impact on the bike demand. The bike demand increases by 0.52 per 1 unit increase in temperature when all other parameters are constant.
2. Year – Year also has positive impact on the bike demand.
3. Weather situation – In particular Light Snow affects negatively to the bike demand. The impact is -0.286.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Ans:

1. We start with reading, understanding the data followed by visualization of the data to understand the correlations between variables.
2. The next step is to prepare the data for building the model by converting categorical variables into dummy columns, creating new variables if required, dropping the variables which are insignificant for building model.
3. Split the data into train and test sets

4. Scale the continuous variables. We use either normalization or standardization for scaling the data. Scaling helps faster convergence of gradient descent. We scale the independent and dependent variables so that the learned coefficients are easy to interpret. Scaling however does not affect the p-value of features or the accuracy of the model.
5. Build model with all the variables or use RFE to select initial set of features to build model.
6. Eliminate the features with high p-value or with high VIF and rebuild the model.
7. Repeat the feature elimination step until all the features have p-value of zero and VIF is in an acceptable range.
8. If more than one models are built on the same data set with different sets of features, use adjusted R-squared to compare the models.
9. Residual analysis – analyse the error terms. The error term is difference between the predicted values of target variable on training set and the actual values.
10. Use the model for predictions

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet are data sets which have nearly identical statistical summary but have very different distribution and when plotted on graphs appear very different.

The quartet outlines the importance of graphs when understanding or describing the data rather than just relying on the statistical summary.

In terms of LR, what this mean is that though the statistical summary look like similar two models can fit the line in completely different ways based on the data points. This is very important when the data have outliers. The outliers can affect the line fitted by linear regression as LR will try to accommodate the outliers as well. The learned coefficients will be slightly different than compared to the line fitted on data without outliers.

Q3. What is Pearson's R?

Ans: Pearson's R is a measure of linear correlation between two variables. It's a normalized measure of covariance between two variables whose value is between -1 and 1. The covariance is measure of how changes in one variable affect the value of another variable.

For two data sets x and y with n samples, the Pearson’s R is the ratio of sum of the difference between each data point in x and y with its mean to the product of the standard deviation of x and y.

The Pearson’s R for a sample can be obtained by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

n is sample size

xi, yi are the individual sample points

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is sample mean for x and analogously for $\bar{y}$

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is process of transforming the original data points on a different, much smaller scale. Each variable in the data will have different unit and value, by way of scaling we can bring all the variables to a similar scale.

The smaller scale of the variables helps in managing and understanding the built model. The scaling does not affect the p-value of the variables or the accuracy of the model.

There are two common types of scaling

1. Normalized – Also know as MinMax scaling. The data is shifted and scaled such that its between 0 and 1.
2. Standardized – Brings all data into standard normal distribution with mean 0 and standard deviation of 1.

The differences between both these scaling is:

1. Normalization will change the sign of the data while standardization will only affect the scale of the data.
2. Outliers will be lost in normalization but not in standardization.

3. The proportion of outliers to other data points will transform to the normalized values as well. Whereas for standardization, the outliers will be transformed to manageable scale.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The infinite VIF value for a predictor means that R-squared for that predictor is 1. In other words, all the variance of that predictor was described by other predictors.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Quantile-quantile (QQ) plots are used to visualize if two data sets share the distribution and come from same population or not. Data points from both the data sets are converted to quantiles and these quantiles are plotted. If two data sets come from a population with same distribution then the points will lie approximately along the reference line which is at 45 degrees.

Sample sizes do not matter for QQ plots.

For linear regression, we can use QQ plot to find if original and predicted values of dependent variable follow same distribution. It will also mean that the error terms have constant variance.