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?

Seasons:- Total number of people renting the bikes increasing from spring till fall and decrease from end of fall with spring being the lowest.

Year:- Number of users for bikes has increased significantly from 2018 to 2019. Number of registered users has increased considerably compared to casual users.

Holidays:- There is a large increase in casual users during holidays.

Workingdays:- Number of registered users during the working days is significant part of the total number of users during working days

Weather:- Bad weather has a negative impact on users with clear weather attracting the most users

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

During creation of dummy variables, function get_dummies creates columns for all unique variables. As we can represent the categorical information of m variables with m-1 columns to improve efficiency, we use drop_first=True.

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

Temp and aTemp have the highest correlation with target variable

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

Model is validated using R-squared, Adj R-squared, residual analysis which is y_true – y_predicted. Residue must be normally distributed with mean 0, and there should be no patterns when residues are plotted using scatterplot.

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

atemp has positive influence as increase in temperature increases the number of people using bikes with its positive coefficient. Both light_rain and humidity have the opposite effect influencing the bike usage negatively with coefficients being negative

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithms belong to the category of supervised learning alogrithms. They take multiple independent variables as input to the model and output continuous variable. Linear regression is of two types, simple linear regression and multiple linear regression. Objective when using regression is to find the best-fit line with least mean square error. This line shows the relationship between input independent variables and output dependent variable. This line has the form

Y = (b1)x + b0 , where b1 = slope (coefficient), b0 = intercept (constant)

It assumes that independent variables and target variable are linearly related. The assumptions of linear regression are linearity, no multicollinearity, independence, homoscedasticity

2. Explain Anscombe's quartet in detail

It is a good example that shows the importance of data visualization before analyzing it.
It was developed by Francis Anscombe in 1973. The quartet has four datasets and they consist of same statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but when plotted on graphs, they show distinct patterns. Each dataset consists of 11 (x-y) data points. First dataset represents linear relationship with some variance. Second dataset has a curve shape and does not exhibit any linear relationship. Third dataset has tight linear relation between x and y with one outlier. Fourth has all x-values constant except for one outlier

3. What is Pearson's R

It measures linear relation between two variables. It is a common method used used for numeric variables. It gives both strength and direction of the correlation. Coefficient lies between –1 and 1. Positive correlation exists between two variables when increase in one is directly proportional to the increase in second variable. Negative correlation exists when two variables are inversely proportional to each other. Positive correlated values exist in first and third quadrant of the x-y plane, whereas negative correlated values exist in second and fourth quadrant. Pearson correlation relies heavily on the mean parameter for the two objects.

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

Scaling ensures all the features are on similar scale preventing dominance of values with large magnitude. Data collected for analysis has varying magnitudes, units and range. If they are not scaled to within a particular range, algorithm only takes magnitude into account and hence incorrect modelling. Scaling just affects the coefficients and no other parameters. Scaling can be done through normalization or standardization.
MinMax scaling is one of the normalization techniques, it scales the values to between 0 and 1. This takes care of the outliers as the high variance is removed. This techniques is useful for images.
Formula :- (x – min(x)) / (max(x) - min(x))

Standard scaling replaces values by their z-scores. It centers the values around mean 0 with a standard deviation of 1.
Formula :- (x – mean(x)) / std(x)

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

If there is a perfect correlation between two variables, then VIF will be infinity. This can happen if there are duplicate rows in the data. VIF for both the rows will be shown as inf

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

Q-Q plot is a quantile – quantile plot. It is used to determine if a dataset follows a certain distribution or not and if two samples came from the same population. It is a plot of quantiles of first dataset against the second dataset. Dataset needs to be sorted. A theoretical distribution is chosen against which to compare the dataset. Compute theoretical quantiles for the distribution. Plot dataset on x-axis and theoretical quantiles on y-axis. If the points on the plot fall approximately on a straight line, your dataset follows assumed distribution. Deviations from straight line indicate departure from assumed distribution.