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?**

Ans: The demand for bikes is high in 5^{th}-9^{th} month. So, we can concentrate on advertisements and Discounts from 10^{th} month, to increase the bike sharing services. We can also go through vehicle modifications or repair issues in During Spring (Jan-April) Season as this is the season where we have less demand, and we expect more demand right after this season (based on the previous data). The bike demand has increased from 2018 to 2019, implies we should be much more attentive in improving the services as the demand is increasing yearly.

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

Ans: During Dummy variable creation, we are creating an encoding for the respective variable. Suppose "if the variable has n levels, it is enough to create n-1 levels of encoding" because the encoding corresponding to the first level is nothing but n-1 zeros.

For example: If we have 3 levels: 100,010,001. The first level is nothing but 1 and 2 zeros i.e., 1 and (n-1) zeros here n is 3. Implies if we drop the first level, it implies that the encoding for the first level is (n-1) zeros. Since we could cover the first level even while dropping first column, we would drop the column for the sake of brevity in computing.

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

Ans: The numerical variable which has the highest correlation with the target variable is "registered" variable. This is because "cnt" variable is directly the sum of "registered" and "casual" variable. Since it directly depends on this sum, it has the highest correlation.

"temp" and "atemp" variables follow the list of highest correlation with the target variable

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

Ans: We could be able to fit a Line across the training set with a r2_score of 82% which means a linear relationship does exist between dependent and independent variable (Assumption 1 satisfied). As we can see from the Residual Analysis graph, which is a bell curve, the error terms are normally distributed, independent of each other and have constant variance.

Quantitively the performance of a Regression model is evaluation by considering: if the P value of each variable is < 0.05(with a significance level of 95%), Probability of F-statistic tending to zero, VIF values<10(preferably less than 5)

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

Ans: "temp", "yr" and "weathersit" (Specifically weathersit 3: light rain) are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: Linear Regression Algorithm assumes a linear relationship between dependant and independent variables. It also assumes that the error terms are normally distributed, independent of each other and exhibits constant variance.

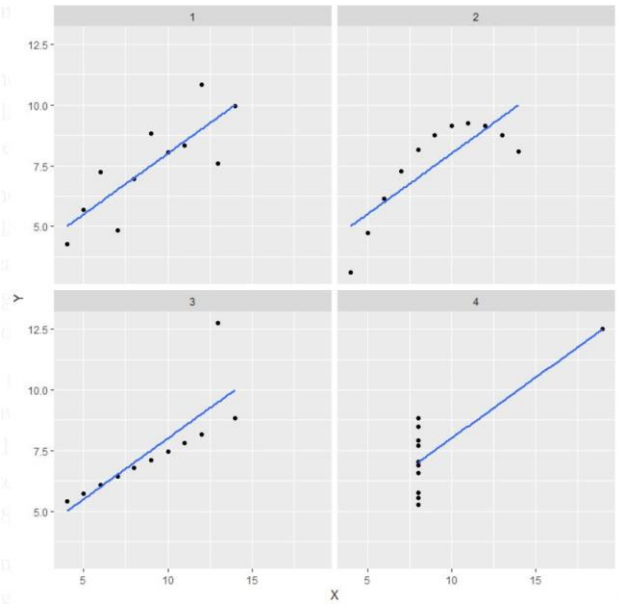
The basic algorithm flow is as follow, it starts with a random line from a single data point, sums up all the errors and move towards adding next point so as to minimize the total error. Likewise, it keeps on fitting new points by minimizing the sum of errors, Once all the points are considered, it optimizes the line by following algorithms like Least Square method, Gradient Descent to minimize the total error by fitting a line in the midst of all the datapoints in the best way.

When dimensions are high, i.e., when we have multiple variables, the line becomes a plane and so on, which are equivalents of a linear relationship in higher dimensions.

2. **Explain the Anscombe's quartet in detail.**

Ans: Anscombe's quartet consists of 4 datasets which have similar statistical properties with differently positioned points. Each dataset consists of 11 points (x, y) and the statistical properties of all 4 datasets are same as follows:

mean(x) = 9	standard deviation(x) = 3.32	
mean(y) = 7.5	standard deviation(y) = 2.03	corr (x, y) = 0.816



1. The top left graph shows a approximate fit of a linear relationship between x and y
2. The top right graph clearly shows a non linear relationship between x and y
3. The bottom left graph appears to be a perfect linear relationship except for one point which seems to be an outlier
4. The bottom right graph shows an example when one high leverage point is enough to produce a high correlation coefficient.

Although the alignment of points is different in all the 4 situations, the statistical properties of these graphs are still the same. This can be extremely useful while interpreting data before choosing a model whether to go for a Linear model or to go for high dimensional nonlinear model etc.

3. What is Pearson's R?

Ans: Pearson's R is a measure of linear correlation between two sets of data. It is the covariance of two variables divided by the product of their standard deviations. It is thus a normalized measurement of the covariance, such that the results are always between -1 and 1.

If R=1, the data is perfectly linear with a positive slope,

R=-1, the data is perfectly linear with a negative slope,

R=0, no linear relationship exists

R between 0 and 0.5, weekly positive correlation

R between -0.5 and 0, weekly negative correlation

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

Ans: Scaling is a method of bringing down all the values of a dataset into a particular sequence. We perform scaling when we have variables with different ranges of values. For example when we have values between 0 and 1 for a particular column, and values in 1000s for another column, the importance given by a model on these two columns varies because of difference in range of their unit. So to remove such discrepancies we make all the column entries under the same range of sequences. This process is called Scaling.

We have two types of Scaling:

1. Standardized Scaling
2. Normalized Scaling

Standardized scaling brings all the data into a standard normal distribution with mean zero and standard deviation 1. Normalized Scaling brings all the data in the range of 0 ands 1.

Differences:

1. Mean=0 in Standardized Scaling, Mean need not be 0 in Normalized Scaling
2. Standard Deviation =1 in Standardized Scaling, Standard Deviation need not be 1 in Normalized Scaling
3. Standardized Scaling will affect the values of dummy variables but Normalized Scaling will not affect the values of dummy variables

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

Ans: If two variables are perfectly correlated their R_Squared becomes 1. Implies 1/(1-R2) becomes infinite, implies VIF which is 1/(1-R2) becomes infinite.

To solve this problem, we need to drop one of the variables from the dataset which is causing the perfect multicollinearity.

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

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles plotted against each other. The purpose of Q-Q plots is to find out if two sets of data come from a common distribution.

If two distributions are similar, the Q-Q plot will approximately lie on the line y=x. If the distributions are linearly related the Q-Q plots approximately lie on a line but not necessarily on the line y=x.

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.