Assignment-based Subjective Questions

Q 1. From your analysis of the categorical variables from the data-set, what could you infer about their effect on the dependent variable?

Ans:
Categorical variables identified were - season, yr, mnth, holiday, weekday, workingday, weathersit.

Conclusion from the dependent variables:
Season - In summer and fall season, demand seems to be better than other seasons.
Yr - Fact that demand is higher in 2019 than 2018 shows that demand is growing year on year.
Mnth - May to October months have higher demand than other months in year.
Holiday - Demand is lower on holidays than other non-working days.
Weekday - There is no remarkable variation in demand over the weekdays.
Workingday - Working days have higher demand for bikes than the non-working days.
Weathersit - Clear and partly cloudy days have better demand for bikes than the days when light or heavy rain occurs.

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

Ans:
It is important to note that the first column that is being dropped can be easily reconstructed using the remaining dummy variables so there is no loss of the information.
Second, it helps to reduce the correlation created among the dummy variables.
Third, extra columns that are not really required should be dropped otherwise it increases the complexity of the model.

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

Ans:
Temperature (temp / atemp) has the highest correlation with the target variable I.e. cnt (demand).

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

Ans:
First, error terms were calculated by subtracting the actual values of the training data set and predicted values of target variable. Then, a distribution plot was drawn for these error terms and it was found that they are nominally distributed. It indicates that assumption was true for linear regression.

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

Ans:
Top 3 features are: atemp, yr and season.

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail. (4 marks)

Ans: A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line. A general equation of a straight line is fitted during linear regression referred as Best-Fit line by minimizing the RSS and finding the optimal value of coefficients.

The equation of a simple best fit regression line is $Y = \beta_0 + \beta_1 X$.
The strength of a linear regression model is mainly explained by R^2 or Adjusted R^2. However, the statistical importance is measured by the F-stat.

Below assumptions should also hold true for a linear model -

1. There is a linear relationship between independent variable and target variable otherwise, there is no use of fitting a linear model between them.
2. Error terms are normally distributed with mean zero.
3. Error terms are independent of each other.
4. Error terms have constant variance (homoscedasticity).

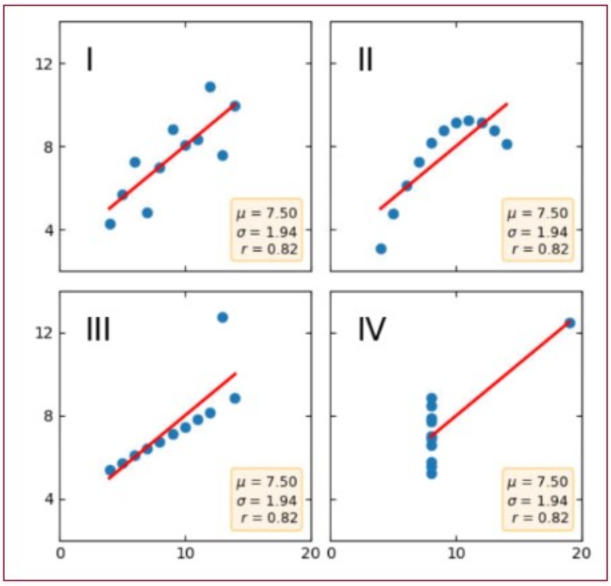
Selecting significant independent variables and multicollinearity:
Significant independent variables can be found manually by starting with most important variable identified using pair plotting and categorical variable analysis. From there on, keep adding next important variable and with the help of summary, observe and adjust based on R-squared, Adj R-squared, p-value and F-stat.
Automatic techniques like RFE can also be used. Calculate VIF to address the issue of multicollinearity.

Perform a hypothesis test on the beta coefficient -
Once you have fitted a straight line on the data, you need to ask, “Is this straight line a significant fit for the data?” Or simply, is the beta coefficient significant to the extent that it is helping in explaining the variance in the data plotted?

Assessing the Model fit -
After you have determined that the coefficient is significant, using p-values, you need some other metrics to determine whether the overall model fit is significant. To do that, you need to look at a parameter called the F-statistic.
So,the parameters to assess a model are:
1. t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Q2. Explain the Anscombe’s quartet in detail? (3 marks)

Ans:
Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analyzing data. To do this he created 4 data sets which would produce nearly identical statistical measures, as it is evident from below four graphs.



Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

So – the moral here is always use graphical analysis alongside statistical measures.

Q 3. What is Pearson's R? (3 marks)

Ans:
Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Pearson correlation is used to test the strength of linear relationships. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Q 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:
Scaling:
Scaling is performed to bring all the variables values to the same scale.

Why it is performed:

Without scaling, variables which have high value will get higher coefficient as model will think that this variable is much more significant than the other variables. While this is not true. Fact is just that the scale of measurement for that variable happens to be different from the other ones. E.g. in our data here, humidity variable mostly has values from 40 to 80 with a mean of 62. While atemp variable has values mostly between 16 and 42 with a mean of 23. So hum variable has higher values than atemp. For model, it would mean that humidity is more important determinant variable of the demand than the atemp while it is not true as evident from our analysis. It just happens to be the case that measurement scale of humidity is different from temperature. So to resolve this anomaly and not to let model affect negatively, scaling is performed.

Difference between normalized scaling and standardized scaling:
1. To start with, the formula used to calculate them is different.

Normalized scaling: X_new = (X - X_min)/(X_max - X_min)

Standardized scaling: X_new = (X - mean)/Std

2. Normalization scaling typically means rescale the values into a range of [0,1]. Standardization scaling typically means rescale data to have a mean of 0 and a standard deviation of 1.

Q 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:
For for calculating VIF is: 1 / (1 - R^2)
So you will have VIF = infinity when R^2 will equal to 1.
R^2 being equal to 1 means that your model is able to explain 100% variation of the target variable. Perfect linear equation has been found, which is generally not the case in reality.

Q 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
Use: The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set’s quantiles along the x-axis and plotting the second data set’s quantiles along the y-axis is how the plot is constructed. In practice, many data sets are compared to the normal distribution. The normal distribution is the base distribution and its quantiles are plotted along the x-axis as the “Theoretical Quantiles” while the sample quantiles are plotted along the y-axis as the “Sample Quantiles”.

Importance in linear regression:
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.