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? (3 marks)

Total no of users of bikes is affected by categorical variables in following ways:

1. With categorical variable season and month we can see a similar overall pattern with dependent variable total users. Spring time or months of January, February and March having least demand.
2. On a holiday we can see that median users are less than on a day when it is not a holiday.
3. With categorical variable weekday we can see that median users on a given day are nearly the same. With highest median users on weekday 4 and lowest median users in weekday 0.
4. Weathersit variable boxplot shows that total users are more on a clear/ partly cloudy day and considerably low on a day with rain.
5. Yr variable boxplot with total users shows that with time the total users for the bike company is increasing with a 50% increase in median users in 2019.

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

While creating dummy columns or dummy encoding we are basically creating a new column for different values of a given categorical variable. If a categorical variable has 3 categories then we are creating 2 columns as it is important to remove the first columns to avoid having multicollinearity in the dataset. Since we can explain a categorical variable of n levels with n-1 dummy columns. Hence removing first column eradicates multicollinearity from dataset. Otherwise the coefficient of features will swing a lot and their signs can also invert and thus pvalue will become unreliable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temp and atemp variables has the highest correlation with target variable cnt. With correlation coefficient of 0.63
That is both temp and atemp are positively correlated with demand. Hence an increase in demand will mostly be observed with when we have higher temperature or in places where we have higher temperature.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

For Validating the assumption of Linear Regression Model:

- **Linear Relationship .**

We have seen from correlation and pairplot that temp has a good linear relationship with cnt. Also variables like windspeed has linear relationship (inverse).

- **Homoscedasticity**

Once we have built the model we have predicted values for target variable from training set. After predicting we have plotted the residual or error terms with respect to index value of y_train and we have seen from the plot that the variance is constant or bounded and not increasing or decreasing. That is error terms does not vary much as the predictor variables changes

- **Absence of Multicollinearity .**

We have plotted correlation heatmap on the X_train dataframe used for predictions (after rfe and manual feature elimination) and we have seen that there are no highly correlated features present in the dataset.

- **Independence of residuals**

Once we have built the model we have predicted values for target variable using training features. After predicting we have plotted the residual or error terms on a scatter plot with respect to variable y_train and we can see that there are no patterns that are seen in the error distribution as the y_train changes and is random.

- **Normality of Errors**

By plotting residuals on a quantile-quantile plot we can see that the residuals line up with 45 degree line and hence have a normal distribution.

We have also plotted the distribution of residuals on a histogram and we have observed that they are normally distributed around mean = 0.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

On the basis of final model Temp, light rain and yr are top 3 variables that are contributing significantly towards explaining the demand of shared bikes.
If temp is high for a given day then the demand will be impacted in a positive manner.
The demand will also increase with time.
On days when rain is expected we can expect lesser demand.
September month will have considerably higher demand
On days/ areas where windspeed is usually less and higher temperature with no rain the demand will be higher
When temp will be low in winter season and we have rain or snow we can expect demand to be very low and usually during month of December or early January.

General Subjective Questions

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

Linear regression is a supervised learning method where the target variable is continuous in nature. Linear regression is based on following assumptions that:

- There is a linear relationship between features and label,
- The error terms are normally distributed with mean zero,
- Error terms are independent of each other and
- The errors are homoscedastic that there is a constant variance in error terms.

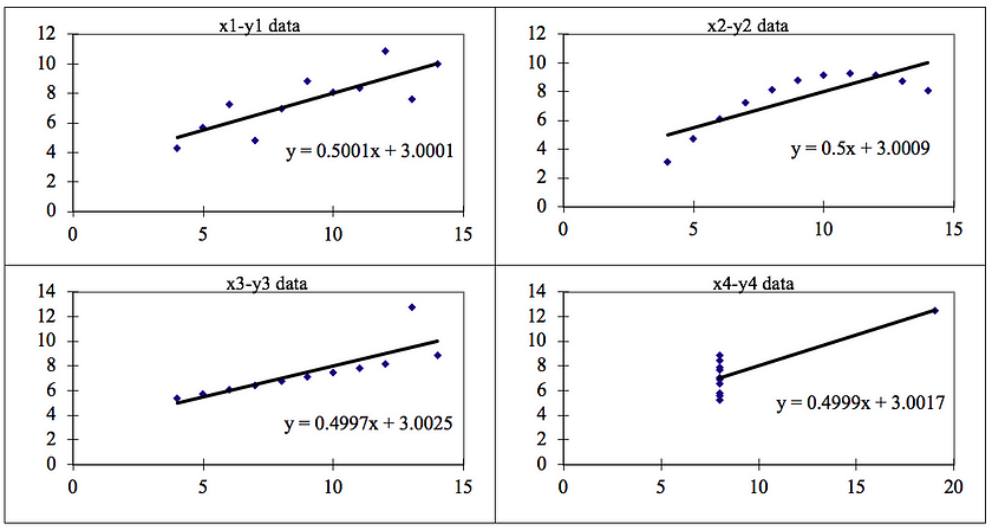
In linear regression the null hypothesis is Bi=0. If p value is less than 0.05 then we can reject the Null hypothesis.

Mathematically we can find this linear function and its coefficients by minimizing the sum of squares of error terms (also called as cost function) where error is the difference between the actual values and predicted. In simple linear regression this linear function is a straight line, that is predicted values lies on a straight line and in case of multiple linear regression this linear function is a plane or hyperplane and all predicted values lies on a plane in the hyperspace.

The linear regression model is evaluated on the basis of RMSE, Rsquare and F statistics.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe quartet is a collection of four data sets which have same descriptive statistics and model equation but very different distributions. This quartet establishes the importance of plotting the distribution and checking whether the data is interpretable by a linear regression model or not. This was first demonstrated by statistician named Francis Anscombe in 1973 to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties



Dataset in Plot 1 is explainable by linear regression model and it fits well with model.
Whereas dataset in plot 2, 3 and 4 have considerable difference in distribution and cannot be explained by linear regression model. In some cases we may have to use a non-linear method to find the model that fits and in some cases removing outliers will make it a good fit with linear regression model.

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

Pearson's R is the measure of linear correlation between two variables. If there is a tendency among variables to increase or decrease together or vice versa, then pearson's correlation coefficient R is a measure of that tendency. Its value lies between -1 to +1. If Pearson's R is less than zero then correlation is negative and variables are linearly but inversely related and if R is greater than zero then correlation is positive and variable are linearly related. If R is near zero then there is very less correlation among the variables if |R| is near or 1 then there is a high correlation.

Rsquared is a measure of explained variance of data by a linear regression model and it is square of R. Its values lies between 0 and 1. It is also called coefficient of determination.

R squared= (1-(RSS/TSS)), where RSS is residual sum of square of errors and TSS is sum of squared of errors from the mean.

If we want to check Pearson's R in python we can do this by using corr() for a dataframe. Corr() will provide the pairwise correlation R for each column in dataframe. To check R squared for a linear model we can use model summary or use function r2 score() or use model property rsquared.

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

Scaling is basically bringing all the required features for model building on a same scale before the model is trained.
Since in real world datasets there are features which have completely different scales, magnitudes and still impact the target variables.

Scaling is performed for mainly two reasons:

- It helps in simplifying the understanding of coefficients of features and their interpretation. Thus, Scaling is important so that the predictor's coefficients are measured on the same level and are comparable to find their true impact properly.
- And it helps to achieve faster convergence of gradient descent and makes the model to find the best fit equation faster.

Difference between Normalized and Standardized scaling:

- Normalized scaling:

Also called min-max scaling. In this type of feature scaling all the features are brought in range of 0 and the formula used by the minmax scaler in background is as follows:

$$X=(x-min(x))/(max(x)-min(x))$$

Normalized type of scaling does not affect the values of dummy variables. It also helps keep the outliers between 0 and 1.

- Standardized scaling:

Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. The formula used by standardisation scaler is as follows:

$$X= (x-mean (x))/ sd (x)$$

Standardize scaling will affect the values of dummy variables

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

VIF is Variance inflation factor. Sometimes one variable may be explained by a combination of other features in the dataset. VIF is a measure of how much a feature is explained by other feature variables. It helps with measuring and detecting multicollinearity in the dataset. If a variable has VIF less than 5 then it is considered to be okay and this variable need not be

removed. If VIF for a variable is above 10 then it is considered high and should be avoided and removed. If VIF is between 5 and 10 then it can be okay and we can inspect it further.

VIF is calculated mathematically as follows:

$VIF_i=1/(1-R_i^2)$, Hence if Ri =1 then VIF tends to infinity.

If VIF for a variable is infinite then it means it is getting completely explained by other variables in the data. And it is perfectly correlated with other variables. Hence we can proceed to removing it from the training dataset.

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

Q-Q plot or Quantile – Quantile plot. Q-Q plot is an extremely useful tool to determine the normality of the data or how much the data is deviated from normality. This plot represents the quantile of standard normal distribution along x-axis and quantile of the obtained data on the y-axis. If the samples are from the same population then the points in the plot will line up with the reference line which is a 45 degree line which can also be plotted.

The Quantile-Quantile plot is used for determining whether two samples are from the same population, have the same tail, and the same normal distribution shape.

For linear regression machine learning model one of our assumptions before building model is that the error terms are normally distributed with mean equal to zero. Hence once we have built the model and fitted on the training dataset we can check and verify whether the residuals that is error values between actual and predicted is normally distributed around zero or not. To check this we can simply make use of qqplot and pass the data as residuals = (y_train-y_train_pred) and see if the points line up on 45 degree line.