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)
Answer: We can see that 25 percentile, median, 75 percentile, 100 percentile demands in 2019 are more than 2018. Overall, demand has increased with year increase.
Season 1 (spring) has considerably low demand of rental bikes compared to other seasons. Season 3 (fall) has highest demand. Weather in fall is pleasant so it makes sense.
Demand in pleasant weather is more. As weather becomes tougher (from 1 to 3) demand decreases.
Demand on all weekdays remains almost constant. Usually, on weekend, demand should have been slightly on higher side however, that's not the case.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer: One of the categories during dummy variable creation can be self-explained by all other categories values. E.g. if gender value for Male is 0, it will be automatically mean gender is Female so no need for separate Female column. Drop_first =True will make first category as redundant category and will drop it. All other categories together will automatically explain value of this category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: Both temp and atemp (feel_temp) both have highest correlation with target variable. 0.63

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: After doing pairplot, I found out that there is linear relationship between some of the variables and dependent variable which is important for deciding if Linear Regression will be considerable for given dataset or not.
Independence between error terms is one of the assumptions which can be checked by Durbin-Watson test which has value between 0-4 always which proves independence between error terms.
No multicollinearity should be present. This can be checked by VIF. I removed all the features with high VIF so no multicollinearity holds.

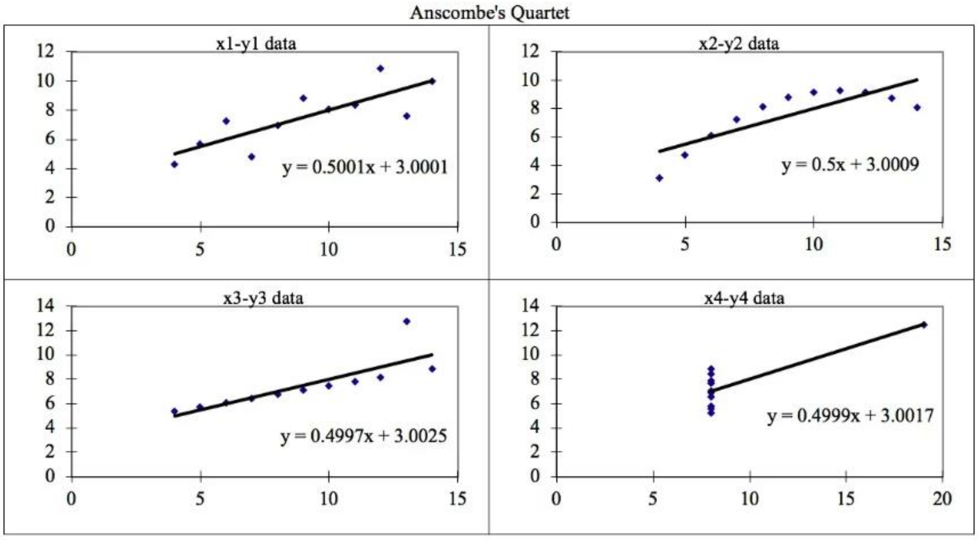
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer: Temperature has highest positive coefficient of 4857. It contributes the highest.
Weather_sit 3 has lowest coefficient (-2020). It contributes but in negative way. This makes sense because it weathersit_3 means terrible weather.
Yr_1 has second highest positive coefficient of 1957. That means 2019 has/will have highest demand.
Yr_1 has third highest positive coefficient of 1473.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer: We have train dataset which consists of dependent and independent variables. Our target is to find suitable coefficient for each independent variables such that using coefficients and a constant value, we would be able to form an expression which will give output equal (or very close) to dependent variable.
These coefficients and constant together are called parameters.
We initially assume values of these parameters and find predicted values and compare with actual value to find errors.
Using these error values as penalty, we make small changes in each parameter such that we errors in the next iteration will reduce. We use Gradient descent algorithm for the same.
Once we reduce acceptable error or no improvement in parameters, we finalize these parameters for the model. Using these parameters, we are able to predict on unknown test dataset which is in the same format as train dataset.

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

Answer: Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. 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.



This was devised to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R? (3 marks)
Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. It varies between -1 to +1. If it is tending towards +1, there is strong positive correlation which means if one variable is increasing, other will also increase in the similar proportion. If it is tending towards -1, it means if one variable is increasing, other will be decreasing linearly.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It gives strength as well as direction of relationship. As it measures only linear relationship, two variables which are highly related in non-linear way (e.g. exponential or logarithmic), this will have low value in that case. Hence, other correlations like Spearman's Rank Correlation and Kendall Rank Correlation should be checked.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer: Scaling is bringing a variable to required scale. E.g. if students are final total marks are varying between 200 to 950, they can be scaled down to required range like 0-100 where highest student will have 100 marks and lowest will have 0 marks.
In data science, scaling is performed to bring all features in same or similar scale. Sometimes, different features lie in different ranges. E.g. height may be ranging from 1.4 to 2 mtrs for an employee and salary may be ranging from 30,000 to 1,00,000. Scaling is done to bring these two features to same range.
Scaling is mainly performed to make gradient descent algorithm more efficient. Without scaling, this algorithm will take much more iterations thus resources to reach optimal point. Hence scaling is done to reach to make algorithm faster.

Normalized or min-max scaling scales variable between 0 and 1. Here, Lowest variable will have value of 0 and highest will have value of 1. This gets very highly affected if there is even single outlier present in the data. In that case, most of the values of variables will be close to 0 or 1.
Standardized scaling scales values in standard normal distribution. Here, underlying assumption is that variable follows normal distribution or is close to it. In this case, even if outlier is present, it will have value much farther from 0.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer: VIF measures multicollinearity. Its formula is 1/(1-R^2). If one the variable is well explained by combination of remaining variables, then it will have high R^2 value which will mean very high VIF. In our case, there were some variables which very well explained by other variables so they had high VIF values. However, as I had used RFE already to remove most insignificant or most multicollinear variables, I didn’t encounter infinite VIF. This will occur when one variable is completely explained by other variables and R^2 becomes 1.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer: Q-Q plots are quantile – quantile scatter plots. They are used to determine if two datasets are following similar distributions or not.
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.
It is used to check if two data sets have common location and scale or have similar distributional shapes or have similar tail behaviour or come from population with a common distribution.