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?**

   Season:

   - Summer & fall are most favourable season for biking.
   - Appropriate advertisement can help to increase sales during summer and fall.

   Month:

   - Bike rental demand is high during June, July, august & September.

   Weekday:

   - No significant changes during weekday.

   Weathersit:

   - Clear weather is most favourable for biking, good discounts and advertisement could work to increase sales.

   Year:

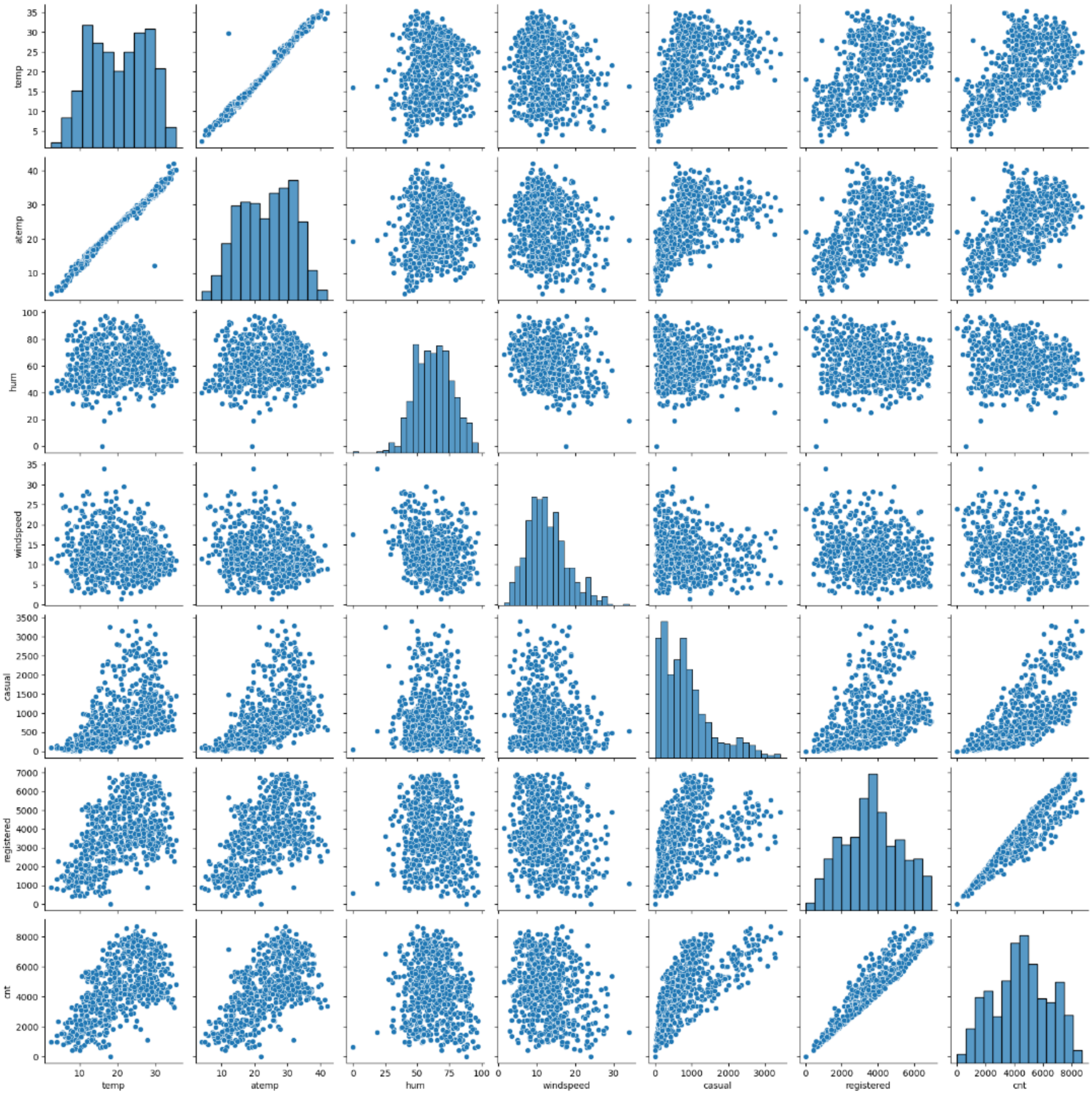
   - As 2 years of data available in sheet shows the increase in bike rental from 2018 to 2019

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

   drop_first=True helps to avoid the multicollinearity issue.

   Multicollinearity issue occurs when independent variables in regression model are highly correlated with each other, this can cause the unstable coefficient.

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

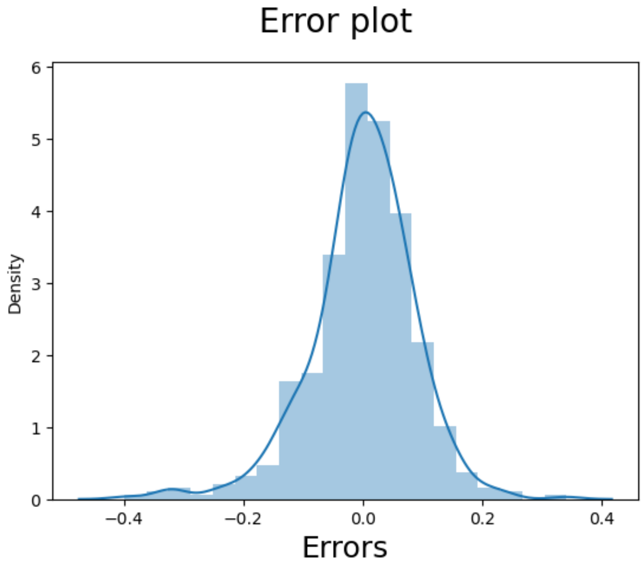


Looking at above image, we can say registered is highly correlated with target variable(count)

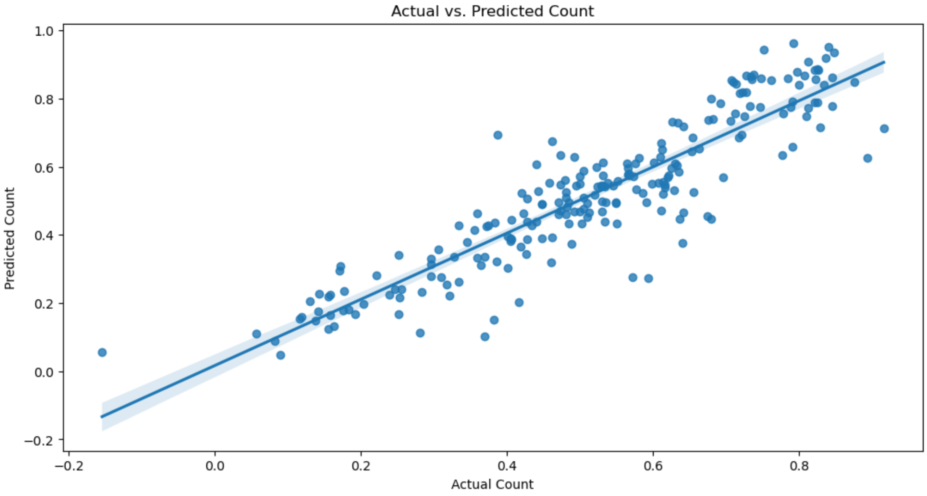
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Validated model by using 2 method below.

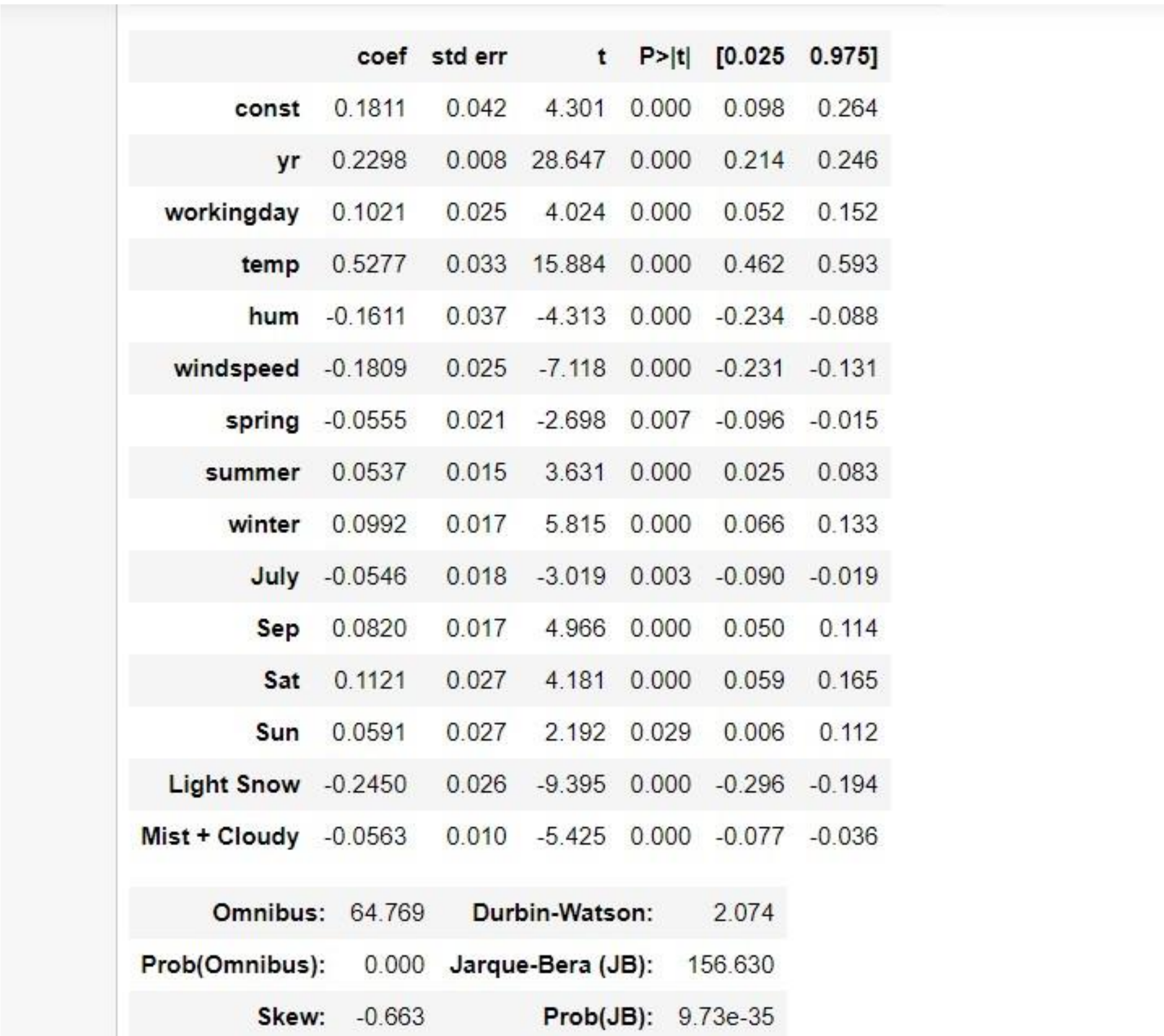
   - By checking error plot (Residuals distribution should follow normal distribution and centred around 0)



   - By checking test data plot (linear relationship between actual & predicted values)



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



	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1811	0.042	4.301	0.000	0.098	0.264
yr	0.2298	0.008	28.647	0.000	0.214	0.246
workingday	0.1021	0.025	4.024	0.000	0.052	0.152
temp	0.5277	0.033	15.884	0.000	0.462	0.593
hum	-0.1611	0.037	-4.313	0.000	-0.234	-0.088
windspeed	-0.1809	0.025	-7.118	0.000	-0.231	-0.131
spring	-0.0555	0.021	-2.698	0.007	-0.096	-0.015
summer	0.0537	0.015	3.631	0.000	0.025	0.083
winter	0.0992	0.017	5.815	0.000	0.066	0.133
July	-0.0546	0.018	-3.019	0.003	-0.090	-0.019
Sep	0.0820	0.017	4.966	0.000	0.050	0.114
Sat	0.1121	0.027	4.181	0.000	0.059	0.165
Sun	0.0591	0.027	2.192	0.029	0.006	0.112
Light Snow	-0.2450	0.026	-9.395	0.000	-0.296	-0.194
Mist + Cloudy	-0.0563	0.010	-5.425	0.000	-0.077	-0.036

Omnibus:	64.769	Durbin-Watson:	2.074
Prob(Omnibus):	0.000	Jarque-Bera (JB):	156.630
Skew:	-0.663	Prob(JB):	9.73e-35

based on final model we can say, **temp (0.528), yr (0.229)** & **Light snow (-0.2450)** are the top 3 variables that significantly contribute towards explaining rental bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is like a tool in machine learning that helps us make predictions. Imagine you're trying to predict something, like the price of a house based on its size. Linear Regression helps us find a way to draw a straight line that best fits the relationship between the size of the house (our independent variable) and its price (our dependent variable).

There are two main types of Linear Regression: Simple Linear Regression and Multiple Linear Regression.

Simple Linear Regression is used when we only have one thing we're using to make our prediction, like the size of the house in our example.

Multiple Linear Regression comes into play when we have more than one thing we're considering, like not only the size of the house but also the number of bedrooms, location, etc.

The line we draw to represent this relationship is called the regression line. If, when one thing goes up, the other thing goes up too, we call it a positive relationship. For example, as the size of the house increases, the price tends to increase as well.

On the other hand, if when one thing goes up, the other thing goes down, we call it a negative relationship. For example, as the distance from the city center increases, the price of the house tends to decrease.

So, Linear Regression helps us understand and quantify these relationships, making it easier to predict things based on the data we have.

2. Explain the Anscombe’s quartet in detail

Imagine you have four sets of data, each with 11 points. Now, when you look at the basic numbers like averages, variances, and correlations, they all seem pretty similar across these four sets. But, when you actually plot these points on a graph, you see something surprising - they all look totally different!

Anscombe's quartet is like a magic trick that shows us how relying only on numbers can sometimes deceive us. It teaches us that we shouldn't just trust the statistics blindly. Sometimes, a visual representation, like a graph, can reveal things that numbers alone can't.

3. What is Pearson’s R?

Pearson's Correlation Coefficient is like a tool we use to see how closely two things are related to each other. Imagine you're studying for a test, and you want to know if there's a connection between the number of hours you spend studying and the grade you get. Pearson's Correlation Coefficient helps you figure that out.

The value of this coefficient can range from -1 to +1.

- If the coefficient is close to +1, it means there's a strong positive relationship. So, if one thing goes up, the other thing tends to go up too. For example the bedrooms in flat and price.
- If the coefficient is close to -1, it means there's a strong negative relationship. This is the opposite scenario, where when one thing goes up, the other tends to go down. An example might be the distance from the city center increases, the price of the house tends to decrease.
- If the coefficient is close to 0, it means there's not much of a relationship at all between the two things you're comparing. It's like saying there's no clear connection between flat price and building you choose.

So, Pearson's Correlation Coefficient helps us understand how closely two things are linked, whether they move together, move in opposite directions, or don't really have much to do with each other.

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

Scaling is a preprocessing technique used in machine learning to make sure all the different features in a dataset are on a similar scale or range. This is important because when features have different units or magnitudes, it can cause issues in the model's performance. For instance, if one feature measures in the thousands and another in the tens, the model might give undue importance to the larger values.

If we don't scale the data, the model could end up incorrectly prioritizing certain features over others simply because of their scale.

Normalization and standardization are two common scaling techniques.

Normalization squishes all the data points into a range between 0 and 1, which is useful when the distribution of the data is not necessarily Gaussian.

Standardization, on the other hand, transforms the data so that it has a mean of 0 and a standard deviation of 1. It does this by replacing the values with their Z-scores, which is helpful when the data follows a Gaussian distribution.

In simpler terms, normalization is like squeezing all the values into a range, while standardization adjusts the values so they're centered around zero and have a consistent spread.

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

When two independent variables are perfectly correlated, the Variance Inflation Factor (VIF) becomes infinite. This happens because the R-squared value, which measures how well the independent variables explain the variation in the dependent variable, is equal to 1 in this scenario. Since VIF is calculated using the formula 1/(1-R^2), when R-squared equals 1, the denominator becomes zero, resulting in an infinite VIF.

This situation indicates a severe problem called multicollinearity, where independent variables are highly correlated with each other. Multicollinearity can make it challenging for the regression model to estimate the unique effects of each independent variable on the dependent variable accurately.

To address multicollinearity, one of the correlated variables may need to be removed from the model. This helps to create a more stable and reliable regression model by ensuring that each independent variable contributes unique information to the prediction of the dependent variable.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to compare the distribution of a sample dataset with a theoretical distribution, such as the normal, uniform, or exponential distribution. It allows us to visually assess whether the observed data follows a specific theoretical distribution.

By plotting the quantiles of the sample dataset against the quantiles of the theoretical distribution, we can determine if the dataset closely matches the expected distribution. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the sample distribution is similar to the theoretical distribution being tested.

Additionally, Q-Q plots can help us identify if there are systematic deviations between the sample and theoretical distributions. For example, if the points on the Q-Q plot deviate significantly from the straight line in certain regions, it indicates that the sample distribution may not follow the theoretical distribution in those areas.

Furthermore, Q-Q plots can be used to assess the normality of errors in a dataset. By plotting the quantiles of the residuals (the differences between observed and predicted values) against the quantiles of a normal distribution, we can check if the residuals are approximately normally distributed. This is important in many statistical analyses, as certain assumptions, such as the normality of errors, need to be met for reliable inference.