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?
Answer: The final multiple linear regression model contains many predictor variables that are categorical in nature, some of which are coded as dummy variables. Spring and winter are classified as seasonal categories and dummy coded. Weathersit_2 and Weathersit_3 belong to the weathersit category and are dummy coded. Similarly, the month variable is grouped into the month category and dummy coded. From the image above, we can conclude that these variables are statistically significant and explain well the variance of the model.

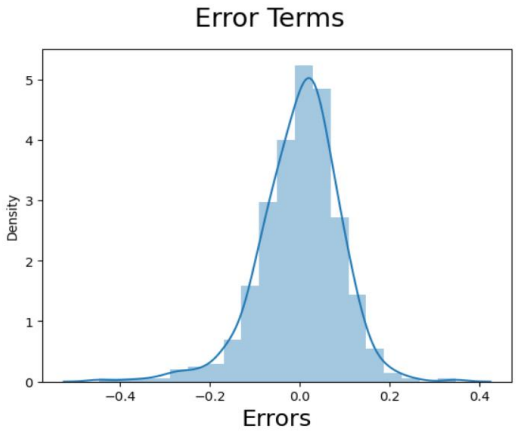
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2466	0.032	7.679	0.000	0.184	0.310
season_Spring	-0.0716	0.021	-3.337	0.001	-0.114	-0.029
season_Summer	0.0333	0.015	2.148	0.032	0.003	0.064
season_Winter	0.0887	0.018	4.951	0.000	0.053	0.124
mnth_Dec	-0.0445	0.018	-2.520	0.012	-0.079	-0.010
mnth_Jan	-0.0503	0.018	-2.738	0.006	-0.086	-0.014
mnth_Jul	-0.0504	0.018	-2.725	0.007	-0.087	-0.014
mnth_Nov	-0.0419	0.019	-2.198	0.028	-0.079	-0.004
mnth_Sep	0.0682	0.017	3.992	0.000	0.035	0.102
weathersit_Light Snow & Rain	-0.2929	0.025	-11.908	0.000	-0.341	-0.245
weathersit_Mist & Cloudy	-0.0814	0.009	-9.359	0.000	-0.099	-0.064
yr	0.2343	0.008	28.709	0.000	0.218	0.250
holiday	-0.0919	0.026	-3.533	0.000	-0.143	-0.041
temp	0.4377	0.036	12.083	0.000	0.366	0.509
windspeed	-0.1586	0.025	-6.290	0.000	-0.208	-0.109

2. Why is it important to use drop_first=True during dummy variable creation?
Answer: Using Drop_first=True is important as it helps reduce extra columns created when creating dummy variables. This reduces the correlations generated between dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: Both atemp and temp have the same correlation with the target variable 0.63, which is the highest among numeric variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer:
Residual analysis:

We need to check whether the error term is also normally distributed (this is actually one of the main assumptions of linear regression). Plotting the histogram of the error term gives:
The residuals follow a normal distribution with mean 0. All fine.



• **Linear relationship between predictor and target variables:**
This is because all predictors are statistically significant. Also, the R-squared value for the training set is 0.813. This means that the variance in the data is explained by all these predictors.

• **The error terms are independent of each other.**
Handled correctly in the model. Predictors are independent of each other. The VIF (variance inflation coefficient) is less than 5 for all predictor variables, so there is no multicollinearity problem.

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

1) **temp** (coef: 0.437655)
2) **const** (coef: 0.246635)
3) **yr** (coef: 0.234287)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Answer: Linear regression is one of the most fundamental algorithms in the machine learning world that belongs to supervised learning. It basically runs a regression task. Regression models predict dependent (target) values based on

independent variables. It is mainly used to explore the relationship between variables and predictions. Different regression models differ based on the type of relationship between dependent and independent variables considered and the number of independent variables used.

2. Explain the Anscombe's quartet in detail.
Answer: Anscombe's quartet consists of a set of four data sets that share identical descriptive statistical properties in terms of mean, variance, R-squared, correlation, and linear regression line, but plotting a scatterplot on a graph Plot is different.

3. What is Pearson's R?
Answer: The Pearson correlation coefficient (r) is the most common way to measure linear correlation. This is a number between -1 and 1 that represents the strength and direction of the relationship between the two variables.

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

- Scaling is a data pre-processing step applied to the independent variables to normalize the data within the specified range. It also helps speed up the computation of algorithms.
- In most cases, collected datasets contain features that vary widely in magnitudes, units and range. Without scaling, the algorithm only considers magnitude, not units, resulting in inaccurate modelling. To solve this problem, we need to perform scaling so that all variables have the same number of digits. It is important to note that scaling only affects coefficients, not other parameters such as t-statistics, f-statistics, p-values, r-squared, etc.
- **Normalization:** This method, also known as min-max scaling. This consists of rescaling the feature space to scale the space in [0,1]. The general formula for normalization is:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

  Where max(x) and min(x) are the maximum and minimum feature values, respectively.
- **Standardization:** Standardization replaces values with z-values. Fit all data to a standard normal distribution with mean (μ) 0 and standard deviation 1 (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Sklearn.preprocessing.scale helps implement standardization in Python. The disadvantage of normalization over standardization is that some information in the data is lost, especially information about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: For perfect correlation, VIF = infinity. This shows perfect correlation between the two independent variables. Perfect correlation yields R2 = 1, which is 1/(1-R2) infinity. To fix this problem, one of the variables responsible for this full multicollinearity needs to be removed from the dataset.
An infinite VIF value indicates that the corresponding variable can be exactly represented by a linear combination of other variables (which also have infinite VIFs).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: A Q-Q plots is also known as a quantile-quantile plots. As the name suggests, the quantiles of the sampling distribution are plotted against the quantiles of the theoretical distribution. Q-Q charts are used to find distribution types of random variables such as Gaussian, uniform, exponential, and even Pareto distribution. You can tell the distribution type from the power of the Q-Q chart just by looking at the plot.