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)

   [Ans]: Categorical variables have a significant effect (either negative or positive) on the dependent variable

Out[76]:

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.797
Model:	OLS	**Adj. R-squared:**	0.792
Method:	Least Squares	**F-statistic:**	162.7
Date:	Mon, 11 Sep 2023	**Prob (F-statistic):**	2.00e-163
Time:	16:06:58	**Log-Likelihood:**	445.62
No. Observations:	510	**AIC:**	-865.2
Df Residuals:	497	**BIC:**	-810.2
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4459	0.018	25.388	0.000	0.411	0.480
yr	0.2461	0.009	27.027	0.000	0.228	0.264
workingday	0.0571	0.012	4.600	0.000	0.033	0.081
windspeed	-0.1926	0.028	-6.836	0.000	-0.248	-0.137
season_spring	-0.2376	0.014	-16.537	0.000	-0.266	-0.209
season_summer	-0.0385	0.013	-3.070	0.002	-0.063	-0.014
mnth_dec	-0.1183	0.017	-6.880	0.000	-0.152	-0.085
mnth_jan	-0.1232	0.020	-6.315	0.000	-0.162	-0.085
mnth_nov	-0.1122	0.018	-6.376	0.000	-0.147	-0.078
mnth_sep	0.0563	0.018	3.116	0.002	0.021	0.092
weekday_sat	0.0664	0.016	4.146	0.000	0.035	0.098
weathersit_clear	0.0890	0.010	9.183	0.000	0.070	0.108
weathersit_light_snow	-0.2312	0.028	-8.291	0.000	-0.286	-0.176

Omnibus:	70.521	**Durbin-Watson:**	1.965
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	185.014
Skew:	-0.692	**Prob(JB):**	6.68e-41
Kurtosis:	5.606	**Cond. No.**	10.7

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   [Ans]: Number of dummy variables needed to represent an N number of dummy variables is N-1
   Ex:

Dummy-1	Dummt-2
1	0
0	1

	Dummt-1&2
Dummy-1	0
Dummy-2	1

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   [Ans]: “temp” and ‘atemp”

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   [Ans]: In the final model calculated the R2 value with test data and compared it with training set data R2 value, both were very close.

   In the final model we got a pretty reasonable R2 value (0.7852321) on the test set and it is very close to the train set R2 value (0.797)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   [Ans]:
   1st: yr – year,
   2nd: spring - season,
   3rd: weathersit – low snow,
   4th: windspeed

General Subjective Questions

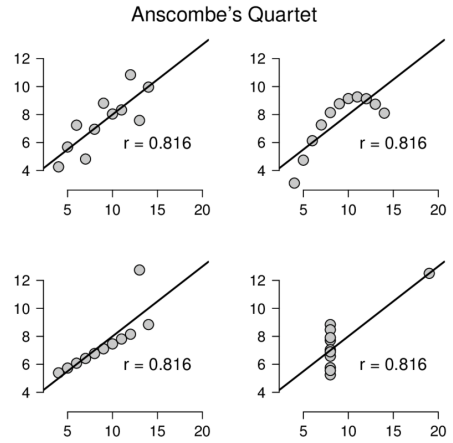
1. Explain the linear regression algorithm in detail. (4 marks)
   [Ans] Linear regression is an algorithm that finds the possible linear relationship between an independent variable and a dependent variable to predict the future possible relation between those same independent variable and the dependent variable.

   It is a supervised machine-learning algorithm that learns from the past datasets and used to predict the future data.

   It assumes that

   1. The independent and dependent variables have a linear relationship with each other - Linearity
   2. The observations in the dataset are independent of each other - Independence
   3. Across all levels of the independent variable(s), the variance of the errors is constant - Homoscedasticity
   4. The errors in the model are normally distributed - Normality
   5. There is no high correlation between the independent variables - No multicollinearity

2. Explain the Anscombe’s quartet in detail. (3 marks)
   [Ans]: Anscombe's quartet comprises four data sets that have closely similar simple descriptive statistics, but have very different distributions and look very different when graphed.



3. What is Pearson's R? (3 marks)
   [Ans]: Pearson's r is a numerical description of the strength of linear association between variables. It is the ratio between the covariance of two variables and the product of their standard deviations. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The result always has a value between −1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   [Ans]:
   a) It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range (ex: 0 to 1) . It also helps in speeding up the calculations in an algorithm.
   b) Most of the times, the collected data set contains information highly varying in magnitudes, units and range. So we do scaling to bring all the variables to the same level of magnitude.
   c)
   Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.
   Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks) '
   [Ans]: When there is a perfect correlation between variables, then VIF becomes infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks
   [Ans] : The Quantile-Quantile plots are used to plot quantiles of a sample distribution against quantiles of a theoretical distribution. This is used to find if a dataset follows any particular type of probability distribution like normal, uniform, exponential.