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)

   there are some categorical values that affect the dependent variable significantly while others are not so significant. An example would be “season” with a value of “summer”. People like to bike during the summer rather than winter when it is cold. Thus it is advantageous to separate these categorical values through dummy variable creation

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

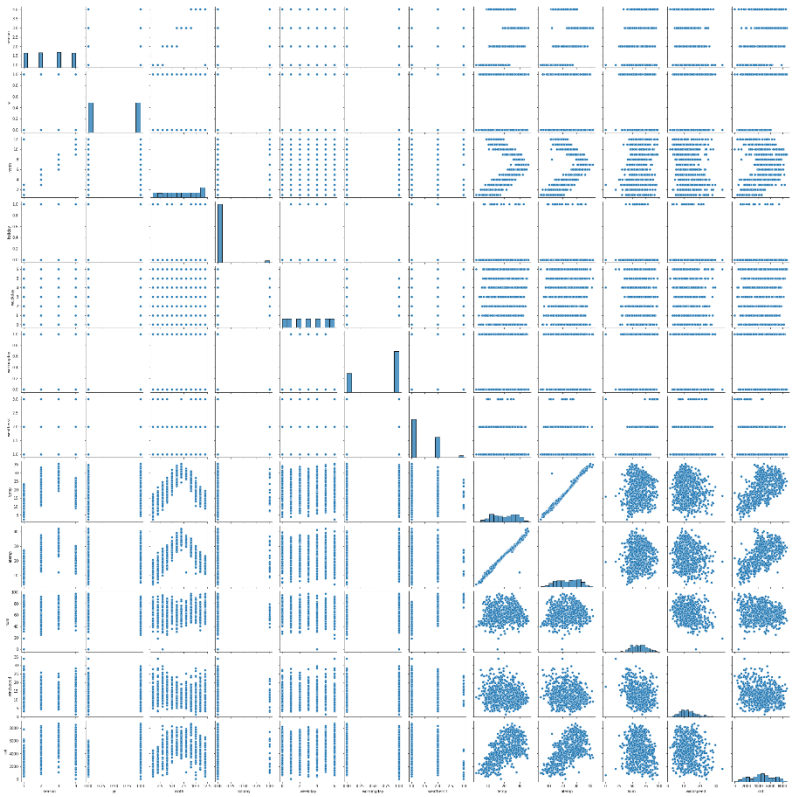
It is important because it will follow the k-1 for dummy variable creation and also reduce the collinearity of dummy variables with each other

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

Those numerical variables would be “temp” and “atemp”. Evidently as temperature increases so too does the number of bikers ride in the street.

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

Assumption: The Dependent variable and Independent variable must have a linear relationship. validated through pair plots



Assumption: No Perfect Multicollinearity
Validate through VIF checking

Assumption: Residuals must be normally distributed.
Checked distplot of final model

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

1. Weathersit

2. Weekday

3. Humidity

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.144
Model:	OLS	**Adj. R-squared:**	0.127
Method:	Least Squares	**F-statistic:**	8.386
Date:	Tue, 12 Jul 2022	**Prob (F-statistic):**	1.17e-12
Time:	22:42:02	**Log-Likelihood:**	-4547.3
No. Observations:	510	**AIC:**	9117.
Df Residuals:	499	**BIC:**	9163.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3419.5953	258.298	13.239	0.000	2912.110	3927.081
hum	45.3017	103.166	0.439	0.661	-157.392	247.995
windspeed	-474.5549	85.384	-5.558	0.000	-642.312	-306.798
weathersit2	544.8340	102.979	5.291	0.000	342.507	747.161
weathersit3	544.8340	102.979	5.291	0.000	342.507	747.161
weekday1	514.1975	306.273	1.679	0.094	-87.547	1115.942
weekday2	529.2305	308.763	1.714	0.087	-77.405	1135.866
weekday3	507.6533	294.796	1.722	0.086	-71.540	1086.847
weekday4	417.2230	308.076	1.354	0.176	-188.064	1022.510
weekday5	591.9120	313.043	1.891	0.059	-23.134	1206.958
weekday6	389.5536	297.197	1.311	0.191	-194.358	973.465
holiday1	-1374.4158	535.383	-2.567	0.011	-2426.298	-322.533

Omnibus:	20.434	**Durbin-Watson:**	1.923
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9.307
Skew:	-0.031	**Prob(JB):**	0.00953
Kurtosis:	2.341	**Cond. No.**	6.81e+16

General Subjective Questions

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

Linear Regression explains correlation of an independent variable/s with a dependent variable. It is a supervised learning method. As an independent variable increase so does the dependent variable. Linear regression explores this relation that is continuous and not discreet in value.

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

Anscombe's quartet is a situation where 4 variables in the dataset looked similar when it comes to there centralities but is very different when plotted visually. Anscombe's quartet emphasizes the importance of data visualization before model building to avoid irregularities when training the data.

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

Pearson's R measures the strength of the relationship between two variables in a linear regression

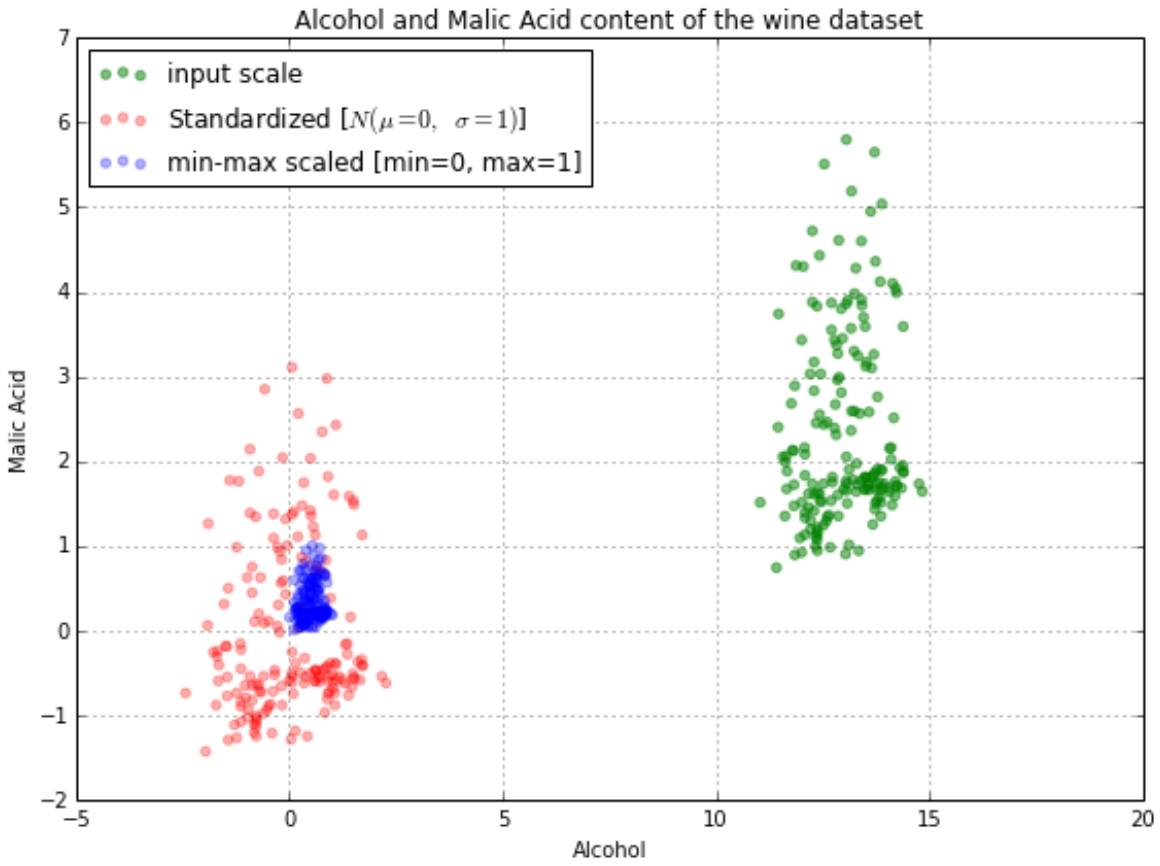
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and +1 meaning a total positive correlation. [1]

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

Scaling is performed to reduce the magnitude of calculations done in the dataset, thus reducing the processing time for training the model. It affects the coefficients and doesn't affect the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized scaling is scaling the values of the dataset to 0 and 1.
Standardized scaling is scaling the values of the dateset to mean = 0 and sigma = 1



[2]

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

(3 marks)

It means that the variables have perfect correlation between the variables and the variable needs to be removed to avoid perfect multicollinearity

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 (Quantile Plot) are probability plot that needed to be compared by plotting their quantiles on how close their distributions are. It is used in linear regression to assess normality.

Reference:

[1] https://www.sciencedirect.com/topics/computer-science/pearson-correlation
[2] https://sebastianraschka.com/Articles/2014_about_feature_scaling.html