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)

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.800
Model:	OLS	**Adj. R-squared:**	0.796
Method:	Least Squares	**F-statistic:**	221.7
Date:	Tue, 10 May 2022	**Prob (F-statistic):**	2.87e-168
Time:	22:26:09	**Log-Likelihood:**	448.84
No. Observations:	510	**AIC:**	-877.7
Df Residuals:	500	**BIC:**	-835.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0862	0.018	4.817	0.000	0.051	0.121
yr	0.2342	0.009	25.882	0.000	0.216	0.252
holiday	-0.0813	0.029	-2.835	0.005	-0.138	-0.025
atemp	0.5674	0.025	22.576	0.000	0.518	0.617
windspeed	-0.1239	0.028	-4.463	0.000	-0.178	-0.069
summer	0.0880	0.012	7.245	0.000	0.064	0.112
winter	0.1252	0.012	10.729	0.000	0.102	0.148
LightRain	-0.2454	0.027	-9.096	0.000	-0.298	-0.192
Aug	0.0538	0.018	2.993	0.003	0.018	0.089
Sep	0.1048	0.018	5.834	0.000	0.070	0.140

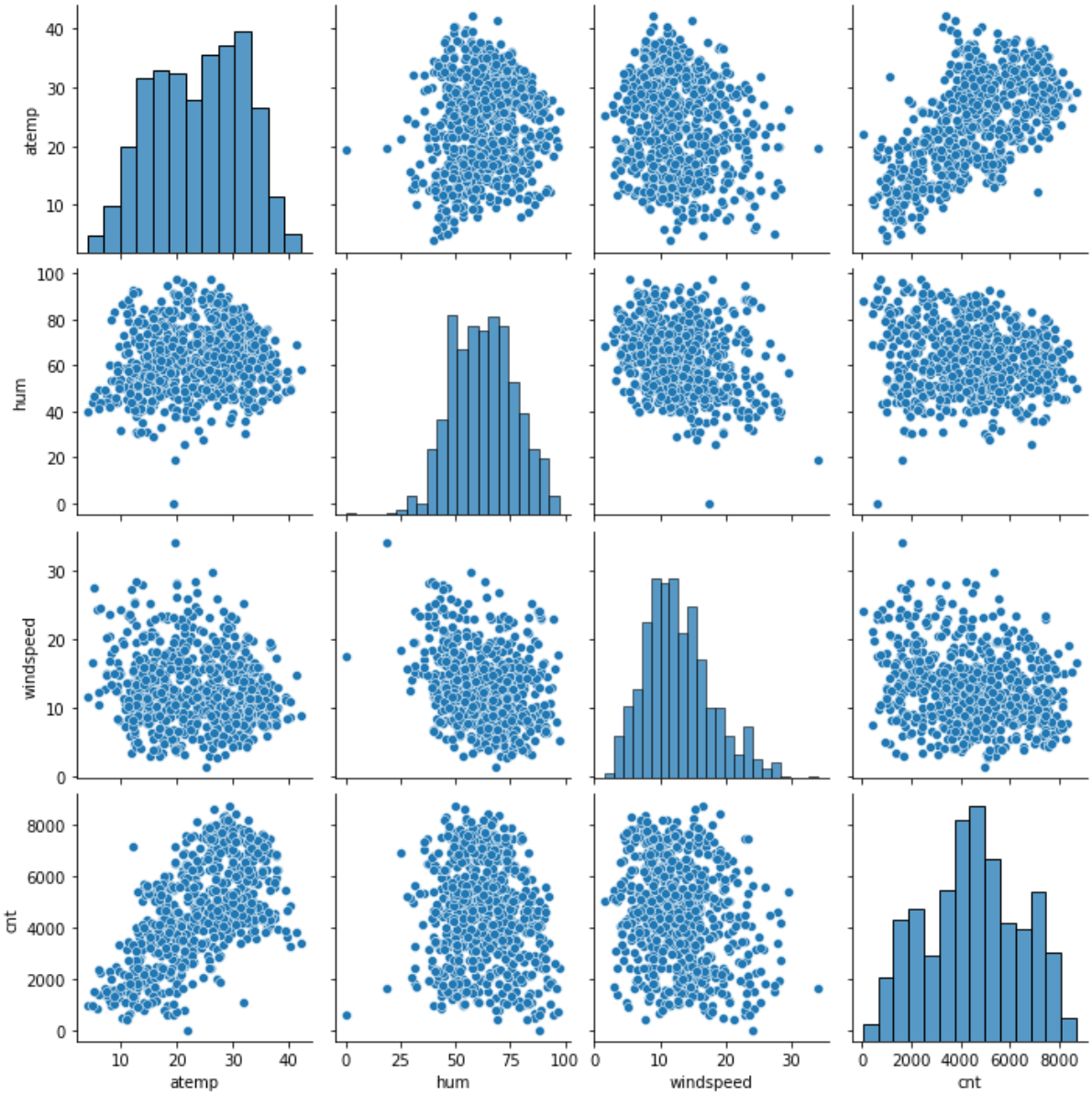
Omnibus:	58.134	**Durbin-Watson:**	1.990
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	114.823
Skew:	-0.665	**Prob(JB):**	1.17e-25
Kurtosis:	4.907	**Cond. No.**	10.1

From the final model we can see the coefficients for categorical variables holiday and LightRain(derived from weathersit) are negative whereas summer, winter (derived from season) and Aug, Sep (derived from mnth) is positive.

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

Answer: When we use the get_dummies() function we get n columns if we have n number of discreet values in the variable. But we really need n-1 columns only as n-1 columns carry the same information as n columns. So, we drop the fist column for brevity. In this case the value 0 in all the remaining columns represents the column which we deleted.

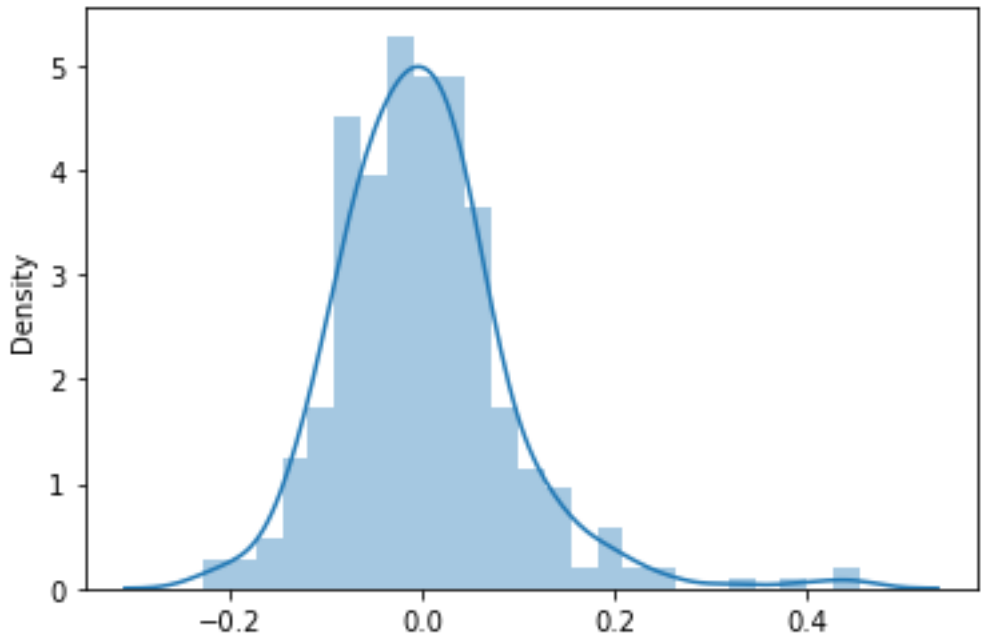
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



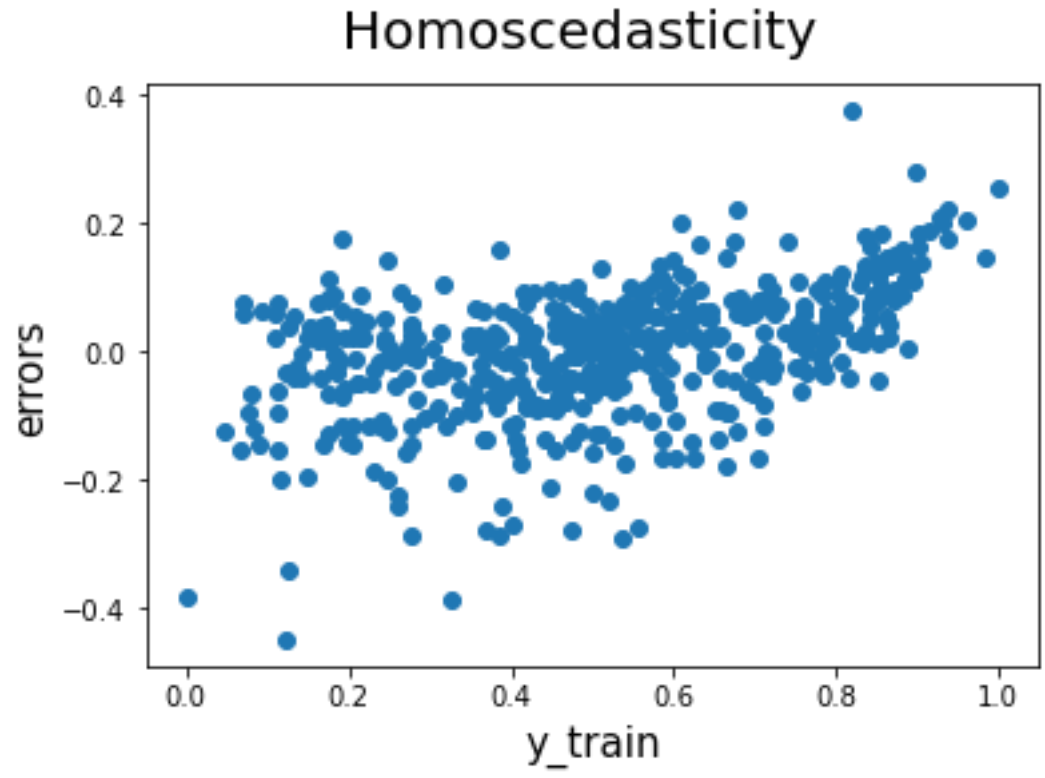
We can clearly see that atemp has the highest correlation with the target variable

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

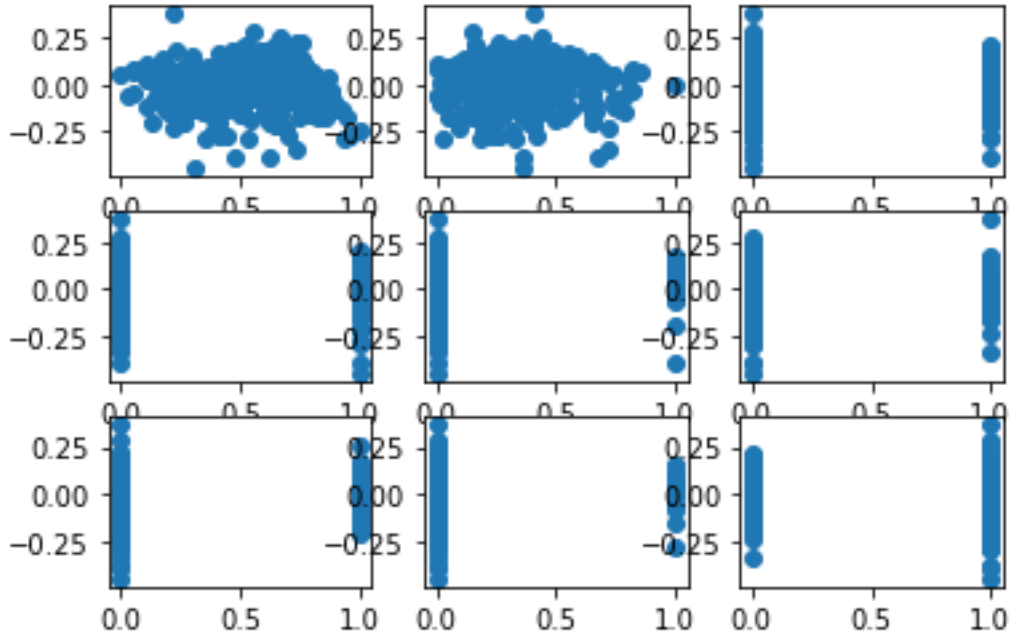
We calculated the error terms or residuals by subtracting predicted values of y by model from actual values of y. When these error terms were plotted as histogram it was following a normal distribution.



We also verified homoscedasticity using scatterplot with the residuals against the dependent variable.



Lastly, we verified that all independent variables are uncorrelated with the error term using scatterplot.



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

As from image of statsmodel summary

First is atemp with a positive coefficient meaning if atemp increases then bike demand increases.

Second is LightRain with a negative coefficient (this variable was derived from weathersit)

Third is yr with a positive coefficient.

General Subjective Questions

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

Answer: Linear Regression model is a type of Regression where we assume that the target variable has a linear relationship with one or more predictor variables. It is called Regression because we are trying to predict values of a continuous variable (target) and it is linear as the relation is that of a straight-line, $y = ß_0 + ß_1X_1 + ß_2X_2 ...$

If there are more than one independent variable to predict the value of dependent variable then it is called Multiple Linear Regression (MLR) otherwise Simple Linear Regression (SLR).

Linear regression works on the principle of fitting a straight line when dependent and independent variables are plotted on a chart. The procedure of finding the slope(s) and intercept for the fitting line requires to identify a cost function and then minimise (or maximise it).

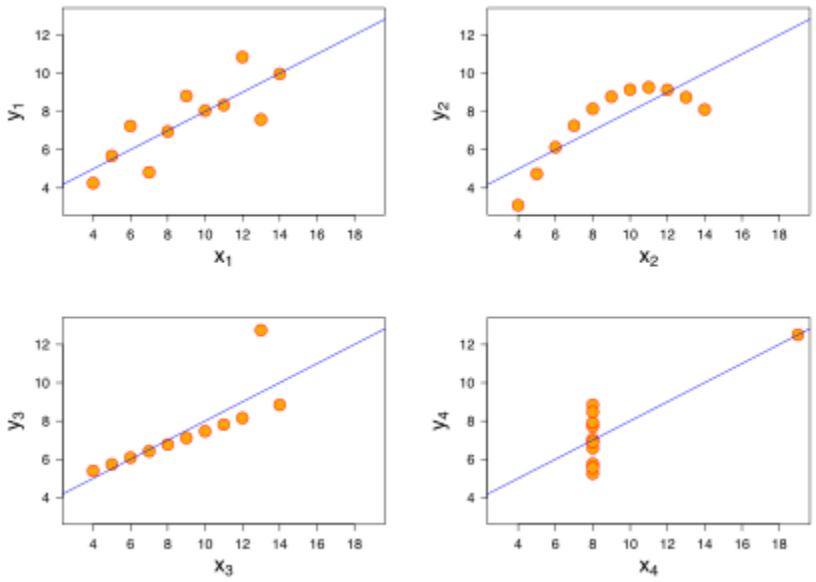
In our case we have identified Residuals - values we get after subtracting the predicted value of dependent variable from actual value of dependent variable. The cost function is the least sum of squares of residuals. After identifying the cost function next step is to minimise (or in some cases maximize) it. Differentiation can be used for this. Another way is Gradient descent method which is an iterative model of reaching the optimal solution.

These calculations are performed by python automatically and we get the coefficients and intercept.

Linear Regression has some important assumptions, the violation of which make the entire mode meaningless. These assumptions are

a) Homoscedasticity - the error is constant along the values of the dependent variable
b) Normality of the residuals - residuals should follow a normal distribution.
c) No or little Multicollinearity - explanatory variables should not be strongly correlated
d) All independent variables are uncorrelated with the error term

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



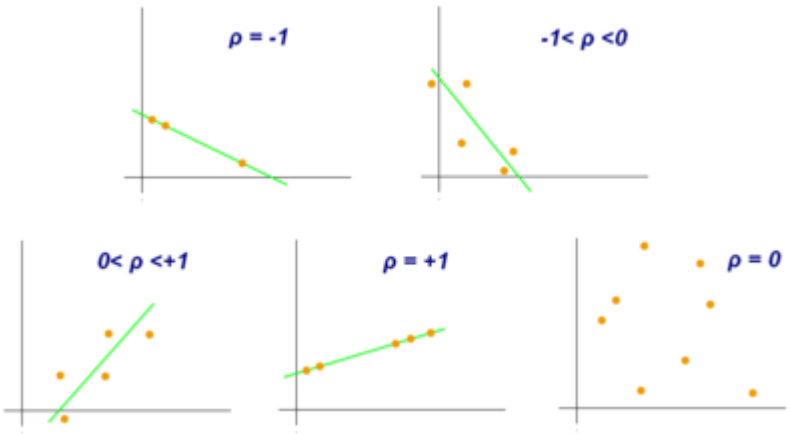
The above 4 graphs are called Anscombe's quartet. The are 11 data points used to plot each of these graphs. If you compare the data points, they all have exact same mean and standard deviation. So, if any statistician is just looking at raw data, he/she may conclude that all 4 data sets convey the same information. But when we plot them separately, we see the actual relationship between the two variables and it is totally different for each one of them.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

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

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The meaning that this ratio conveys is that it is normalising the covariance and the result would be a number between -1 and 1 where -1 signifies perfectly inverse co-variance, 0 represents no co-variance at all and 1 would mean perfect co-variance between two variables. Please note that either of these extremes would not occur naturally. Also, it should be kept in mind that Pearson's R is indicates only linear corelation between two variables and ignores any other kind of relationship as concept of co-variance only takes into account the linear relation.

Here are some examples for different values of Pearson's R



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

Answer: Scaling means to change the range in which a column has values. Scaling does not change the actual representation the numbers carry it just changes the scale on which data is represented. In normalised scaling we try to fit all values between 0 and 1 where min value is replaced by 0 and max value is replaced by 1. All other values are between 0 and 1 depending how far/near they are from min/max. Standardisation however, uses mean value and standard deviation to scale the numbers. (mean=0 and sigma=1). An example of scaling is to change the units of measurement of the data, for example, to convert a temperature from Celsius to Fahrenheit.

The benefit of standardisation is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier).

Standardization assumes that your data has a normal distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The formula for VIF is a ratio. When the denominator is decreased the value of VIF will increase. So, if denominator approaches 0, we can say VIF approaches infinity. Now when can denominator i.e., $1-R^2$, reach 0? when R^2 approaches 1. The implication of R^2 being 1 means it is perfectly explaining all the variance in the target variable. Please not that this is an ideal condition and does not occur naturally.

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 or Quantile-Quantile plot is graph used in statistics for comparing two probability distributions by plotting their quantiles against each other. By quantiles we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

It helps us to test if two data dets can be fit with the same distribution