Subjective Answers - Boom Bike Sharing Assignment

Ques 1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans –

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	241.2
Date:	Sun, 14 Aug 2022	Prob (F-statistic):	6.05e-184
Time:	16:48:36	Log-Likelihood:	488.67
No. Observations:	510	AIC:	-955.3
Df Residuals:	499	BIC:	-908.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1275	0.017	7.429	0.000	0.094	0.161
yr	0.2334	0.008	27.867	0.000	0.217	0.250
holiday	-0.0934	0.027	-3.513	0.000	-0.146	-0.041
atemp	0.5370	0.024	22.850	0.000	0.491	0.583
windspeed	-0.1319	0.026	-5.132	0.000	-0.182	-0.081
summer	0.0990	0.011	8.750	0.000	0.077	0.121
winter	0.1311	0.011	12.114	0.000	0.110	0.152
LightRain	-0.2760	0.025	-10.953	0.000	-0.326	-0.227
Mist	-0.0822	0.009	-9.184	0.000	-0.100	-0.065
Aug	0.0679	0.017	4.065	0.000	0.035	0.101
Sep	0.1209	0.017	7.226	0.000	0.088	0.154

Omnibus:	51.460	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	124.152
Skew:	-0.530	Prob(JB):	1.10e-27
Kurtosis:	5.172	Cond. No.	10.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

We can see in the above Regression result that categorical variable holiday and light rain and mist are negative and summer, winter (seasons) and Aug, Sept (Months) are positive.

Based on the overall we can infer that

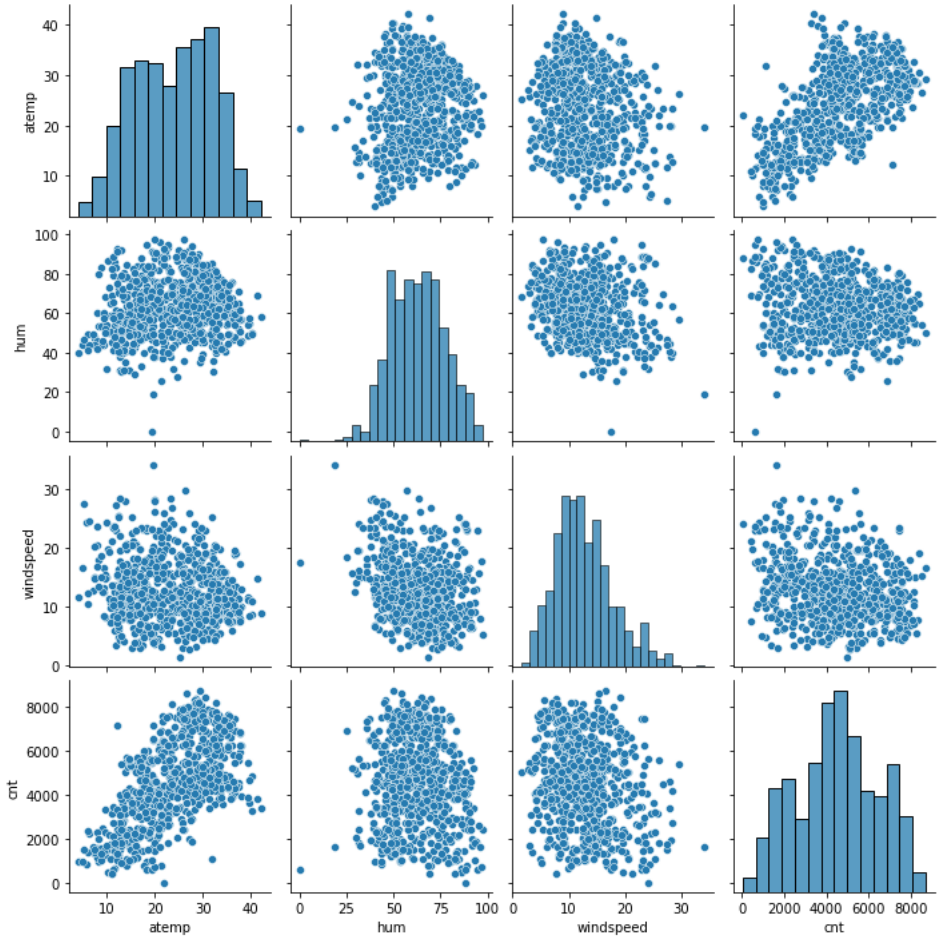
1. Summer and winter have more demands
2. Holidays - Holidays the demand becomes higher.
3. People not to prefer renting the bikes when the weather is favourable such as light rain.

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

Ans – When we use get_dummies() we get n columns based on the n discrete variables. drop_first=True helps in reducing the extra column created during the dummies method. It helps in reducing the mulicollinearity.

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

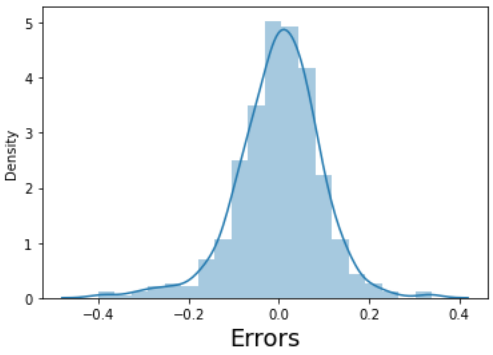
Ans –



Based on the pair plot we can see that “atemp” column has the highest correlation with the target variable.

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

Ans – Error term graph show the difference between target and predicted variable which we can see from the histogram below. Error term follow the Normal Distribution with the mean near to 0.



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

Ans – The top 3 feature which are contributing towards explaining the demand are

a) Temperature – If temperature increase the demand increase
b) Weather – If weather is not good like light rain it decreases the demand.
c) Year – Year is also the positive coefficient.

General Subjective Questions

Ques 1- Explain the linear regression algorithm in detail.

Ans –

Linear Regression is a supervised machine learning model which finds the best fits relationship linear line between the dependent and independent variables. It is of two types

a) Simple linear Regression- With only one independent variable, the model finds the linear relationship with the dependent variable.

Equation - $y = ß_0 + ß_1X_1 ...$

b) Multiple Linear Regression - With more than one independent variable and model tries to find the relationship.

Equation - $y = ß_0 + ß_1X_1 + ß_2X_2 ... ß_nX_{2n}$

A linear regression model aims to find the best fit linear line and coefficient such that error is minimised.

The assumption in linear regression is:
1) Linearity
2) Normality
3) Homoscedasticity
4) No Multicollinearity
5) Error term should be normally distributed

Ques 2- Explain the Anscombe's quartet in detail.

Ans – Anscombe's Quartet - it's an model to demonstrate the importance of the visualisation. It tell about the four (4) data set that have nearly identical descriptive statistics, still have very different distribution and appear different in graphs. This was created by Francis Anscombe to illustrate the importance of the graphs before analysing and model building.

There are 11 data points with 4 data set, all shares the same descriptive statistics (mean, variance, standard, deviation etc)

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

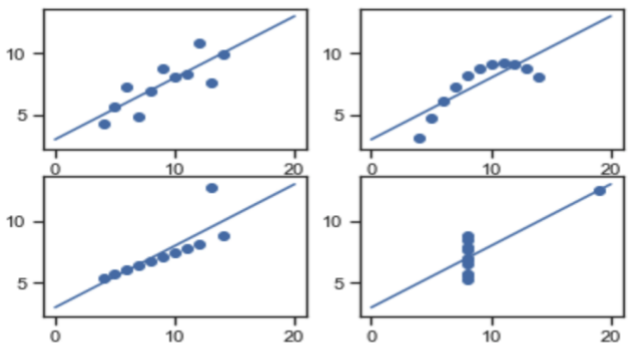
Average Value of y = 7.50

Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3



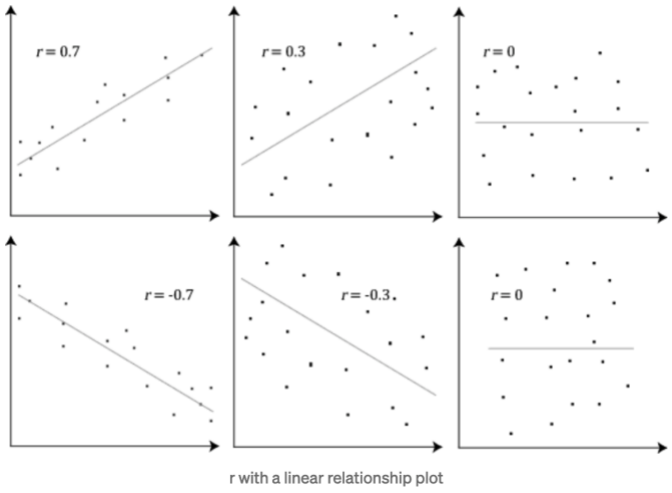
Graphical Representation of Anscombe's Quartet

Ques 3- What is Pearson's R?

Ans – The Pearson correlation coefficient (named for Karl Pearson) can be used to summarise the strength of the linear relationship between two data samples. The value of Pearson's Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation.

Properties of Pearson correlation coefficient

a) The range of r is between [-1,1].
b) The computation of r is independent of the change of origin and scale of measurement.
c) r = 1 (perfectly positive correlation), r =-1 (perfectly negative correction) r = 0 (no correlation)



r with a linear relationship plot

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

Ans – Scaling is the data pre-processing technique to standardize the independent features present in the data in a fixed range.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Techniques of Feature Scaling

1. **Normalisation / Min-Max Scaling-** Normalisation, also known as min-max scaling, is a scaling technique whereby the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

2. **Standardisation -** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

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

Ans – VIF is the ratio. If the denominator decreases the value of VIF will increase and if it reaches 0 the VIF goes to infinity. This shows the perfect corelation if VIF goes to infinity. The denominator can reach 0 when in 1-R2, R2 approaches 1. This is an ideal condition and is hard to reach naturally.

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

Ans - Q-Q plot is often called quantile plot. It is a 2D plot in which we compare the theoretical quantiles of a distribution with the sample quantiles of a dataset.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.