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?

Based on the analysis we have found a model which can reasonably predict the target variable using the 6 independent variables, i.e. : yr, atemp, s_spring, w_light_snow_rain, w_misty, and windspeed.

It is important to note that year and "feels-like" temperature (atemp) have positive correlation with bike shared, and that correlates to the uptick in shared bikes in year 2019 (which is represent as 1 in the dataset) compared to 2018 (represented as 0)

The higher the temperature, the more people would like to rent more bikes because it might be conducive weather for outdoor activities.

The unfavorable weather conditions like rain, snow, and strong winds would usually discourage the shared activities, and they are negatively correlated with the target variable.

Except spring no other seasons have any major impact on outcome variable (i.e. cnt).

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

It is important to use drop_first=True because that'll generate one less dummy variable. If we generate n dummy variables, they would start correlating to each other, because each row will have only dummy variable with value 1 and others as 0, and that can become a prediction model in itself.

It would unnecessarily create the multicollinearity situation in the model. This is also known as Dummy Variable Trap.

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

Two numerical variables temp and atemp have the highest correlation with the target variable and interestingly they are multicollinear with each other (their VIF is 417)

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

Validated the assumptions of the linear regression model by analyzing following,

a. The VIF of all the columns is less than 5, which suggests that there is no multicollinearity.

	Features	VIF
5	windspeed	3.63
1	atemp	3.48
0	yr	2.02
2	s_spring	1.50
4	w_misty	1.46
3	w_light_snow_rain	1.06

b. The model's R^2 is 0.811 which is pretty good, and all the columns are significant given their p-values are less than 0.05.

[42]:

OLS Regression Results

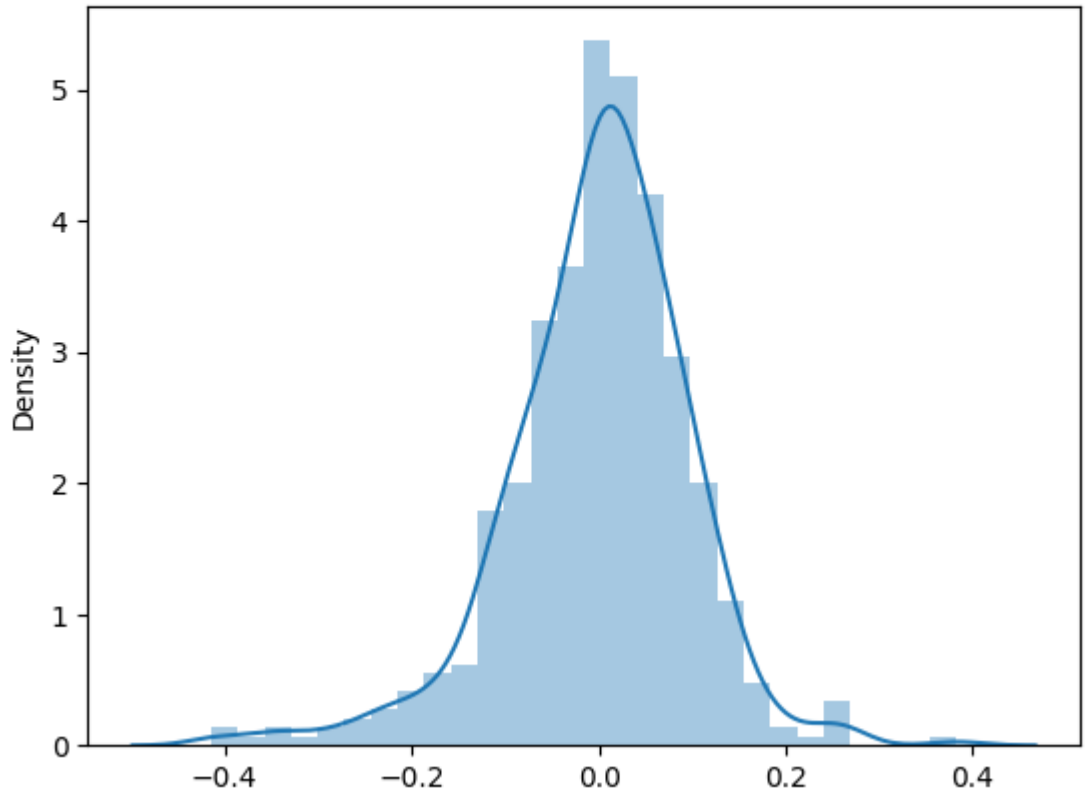
Dep. Variable:	cnt	**R-squared:**	0.811
Model:	OLS	**Adj. R-squared:**	0.809
Method:	Least Squares	**F-statistic:**	360.4
Date:	Tue, 12 Sep 2023	**Prob (F-statistic):**	1.55e-178
Time:	23:58:35	**Log-Likelihood:**	464.12
No. Observations:	510	**AIC:**	-914.2
Df Residuals:	503	**BIC:**	-884.6
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3102	0.020	15.621	0.000	0.271	0.349
yr	0.2368	0.009	27.047	0.000	0.220	0.254
atemp	0.3889	0.026	14.687	0.000	0.337	0.441
s_spring	-0.1529	0.013	-11.847	0.000	-0.178	-0.128
w_light_snow_rain	-0.2678	0.026	-10.222	0.000	-0.319	-0.216
w_misty	-0.0744	0.009	-8.022	0.000	-0.093	-0.056
windspeed	-0.1428	0.026	-5.413	0.000	-0.195	-0.091

c. Performed the residue analysis on the train dataset using following,

```
res = y_m1_train - y_m1_train_pred
sns.distplot(res)
```

```
<Axes: ylabel='Density'>
```



The error (residue) is normally distributed which means that model is fit well because for most of the predictions, the RMS is 0.

So, in summary given adequate R^2 score, the least possible number of independent variables (i.e. 6), establishing no multicollinearity in them using VIF technique, and normally distributed residue suggest that the derived model is a good fit for the data.

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

Following three features are contributing significantly,

a. Year (2018 vs 2019), the demand of shared bikes is going up significantly in 2019 compared to 2018. So the year is very positively correlated with the demand.
b. “atemp” (feels like temperature) is also positively correlated with the demand, and that might be because on sunny days, people tend to perform more outdoor activities.
c. The light rain, snow, and thunderstorms kind of weather is in negative correlation. Basically, if the weather is not clear, people choose not to share / rent the bike.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. There are two types of linear regression,

a. Simple Linear Regression : Has only one independent variable to predict target variable.

b. Multiple Linear Regression : Has more than one independent variable to predict target variable.

The goal of linear regression model is to find the linear equation using the independent variables to derive the target variable's value. The equation would be like,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

There would also be an error component represented as ε however with enough observations, the error would be normally distributed with mean = 0.

The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s). Here Beta1, Beta2, … are called coefficients to the independent variables (they represent slope of the line), and Beta0 is a constant which is known as intercept. The equation may have intercept as 0 (zero).

Following are the key steps of the algorithm,

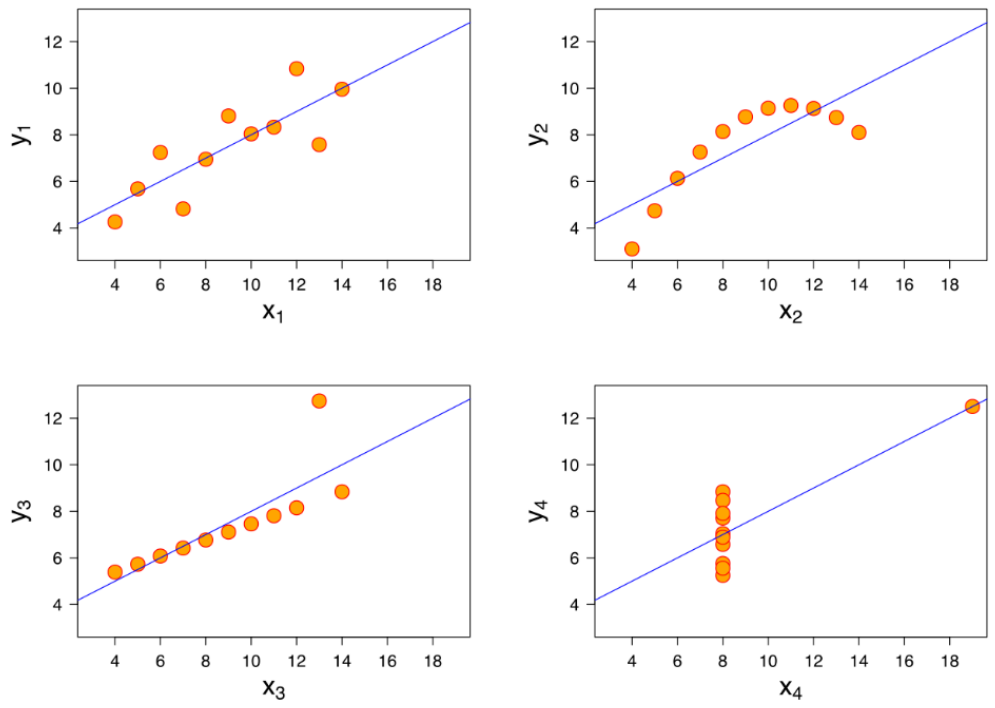
- Generate dummy variables if required.
- Split the dataset into two subsets, i.e. train dataset and test dataset.
- Scale the numeric variables using normalized / standardized scaling on train dataset.
- Derive the intercept and coefficients (slopes) using OLS (ordinary least square) method.
    - Keep enhancing the coefficients using the chosen cost function. There are two types of cost functions, Mean Squared Error (MSE), R^2, Gradient descent, etc...
- R^2 is a simple and effective cost function, wherein the algorithm starves to maximize the value. It would be between 0 to 1.
- Once the desired R^2 is achieved by choosing / removing the independent variables, in case of multiple linear regression, we need to compute VIF (Variance Inflation Factor) to find out the multicollinearity among the independent variables. If any variable has VIF more than 5 should be dropped to make the model less complex.

- Perform the residual analysis on train data set (predicted output – actual output) to ensure that error is normally distributed, and it's mean is at 0.
- Scale the test data using the scaler configured using the train dataset.
- Perform the prediction on test dataset using the built model
- Perform the residual analysis on test dataset and ensure that error follow normal distribution and the mean is at 0.
- Calculate the R^2 score using test predicted output and actual output, and ensure it is greater than 0.7. If not the model might be overfit to the train dataset and need to be revised with respect to independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is the phenomena which demonstrates how the different datasets could have similar descriptive statistical parameters (average, min, max, std deviation, etc...). It is called quartet because it takes four data sets, which have very different distributions and appear very different when graphed. These datasets were constructed in 1973 by the statistician Francis Anscombe to demonstrate this phenomenon.

In this quartet, each dataset consists of eleven (x,y) points when plotted they look completely different from each other as shown in below image,



This is to remark that we should not just rely on statistical parameters of the dataset rather they should also be examined visually to understand how similar / different they are.

3. What is Pearson's R ?

The Pearson R is a correlation coefficient, which is the most common way of measuring a linear correlation between two variables. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Following table shows the value range of r and it's meaning with respect to correlation between two variables,

Pearson R	Correlation
-1 to 0	The two variables are inversely correlated. When one goes up the other goes down, and vice versa.
0	The two variables are not correlated (in other words they are not related).
0 to 1	The two variables are positively correlated, if one goes up then other goes up and vice versa.

Another way to think of the Pearson correlation coefficient (R) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, R is negative. When the slope is positive, R is positive.

Following is the formula to compute Pearson's R

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where x is an independent variable and y is a target variable.

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

Scaling is the process of normalizing the range of independent variables of numeric nature in the dataset.

The numeric columns would have different ranges of min-max based on their units. For example, the column having values in centimeters would generally have bigger numbers compared to the column having values in meters.

It is important to perform scaling to improve the performance and accuracy of linear regression model.

The goal of feature scaling is to ensure that all features contribute equally to the prediction of the outcome variable. If the range of the features is not normalized, the algorithm will give more weight to variables with larger ranges, which may not necessarily be more important than other features. This can lead to overfitting or underfitting of the model, resulting in poor performance.

a. Normalized Scaling

   It scales the variables so that they have a range between 0 and 1. This method is useful when the range of the variables is not known or when the range varies widely between variables.

   It is computed using following formula,

```
x_scaled = (x - min(x)) / (max(x) - min(x))
```

b. Standardized scaling

   The standardized scaling is the process to transform the values of numeric variables such that they have mean to 0 and standard deviation of 1. It transforms the variables so that they are normally distributed. It is a useful technique for some algorithms like logistic regression and support vector machines.

   The formula of standardized scaling is,

```
x_scaled = (x - mean(x)) / std(x)
```

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

The VIF score for variable could be infinite in case when it is multicollinear with the other variables and some of these variables are in perfect correlation of this variable. In other words, the other variables can build the perfect multiple linear regression model for the target variable, which has infinite score.

When two variables are perfectly correlated then their R^2 score would be 1 and the VIF formula being $1 / (1 – R^2)$, would result into 1 / 0 and that's termed as infinite.

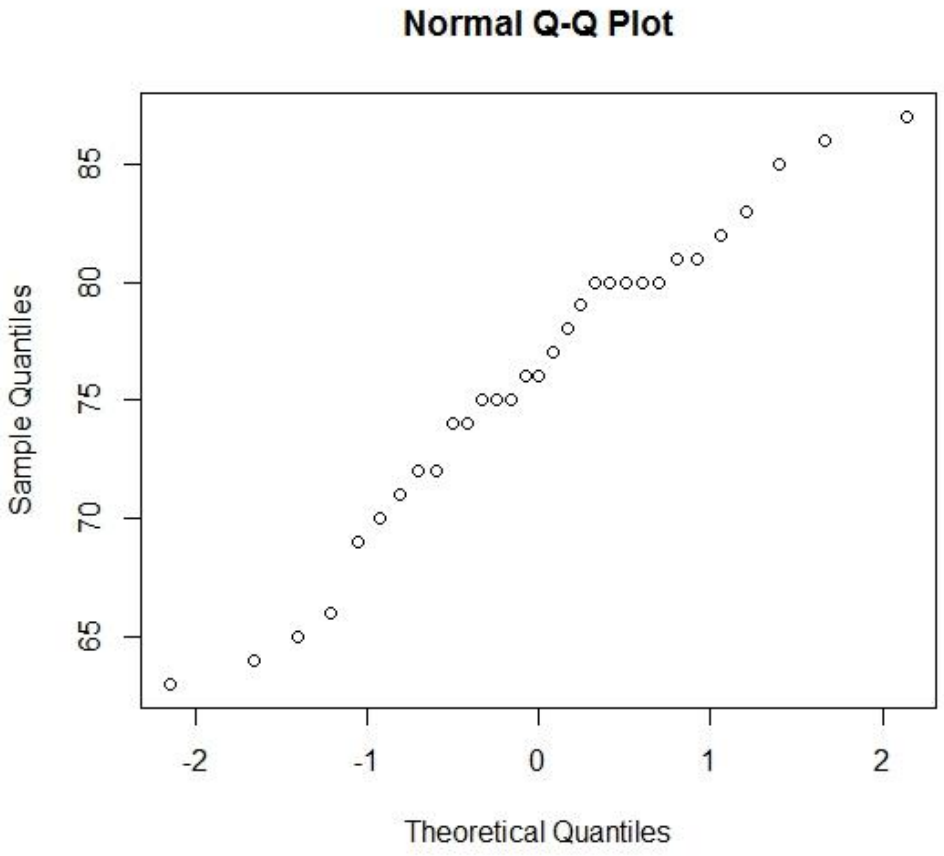
Such a scenario could occur when we have independent variables representing the same attribute but with different units. For example, distance_in_km, and distance_in_mile

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

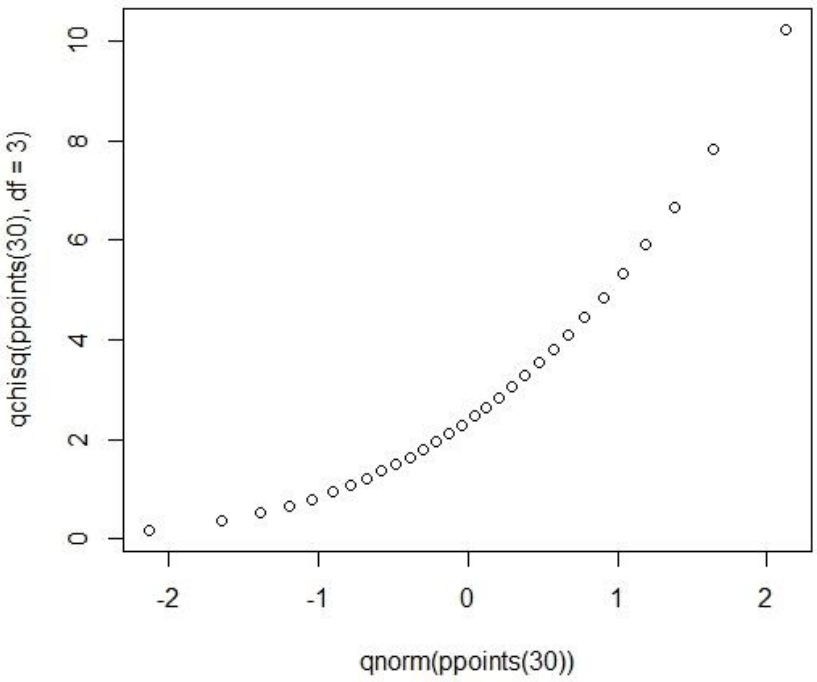
A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions. It is a graphical method to determine whether two samples' data have come from the same distribution or not.

Q-Q plot is a scatter plot between two variables which belong to the same quantile in their respective distributions.

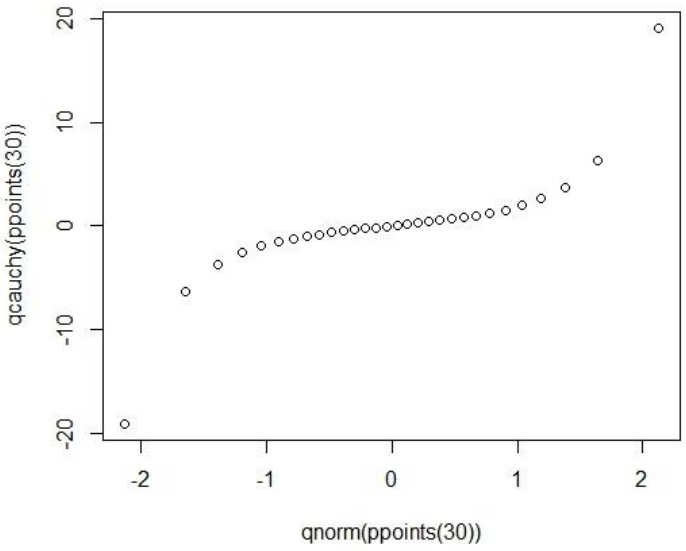
- When both distributions are perfectly normal, the Q-Q plot would appear as straight line at 45 degree,



- When the distributions are normal but the samples are skewed, the plot would appear curved,



- When the samples data has more extreme values than expected if the values have come from normal distribution, the QQ plot would appear as,



In case of linear regression, when we receive the train and test data separately, we could employ QQ plot to verify whether they have come from the same population, have same distribution, and the same location/scale, and have same tail of distribution.