Bike Sharing Analysis Assignment - ML1

EPGP in ML & AI - April 2023

Agenda

This document intends to answer below questions where assignment based subjective questions are answered using **Bike Sharing Analysis** assignment.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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)
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Problem Statement

A US bike-sharing provider BoomBikes aspires to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

- Which variables are significant in predicting the demand for shared bikes
- How well those variables describe the bike demands

Business Goal

Requirement is to model the demand for shared bikes with the available independent variables, to understand how exactly the demands vary with different features. And use the model to understand the demand dynamics of a new market.

General Subjective Questions

General Subjective Questions - Answer (Q1)

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

A Linear Regression algorithm involves building an equation like below:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... b_nx_n ;$$

where y is the target variable, $x_{1,}$ $x_{2,}$ x_3 etc. are independent variable and $b_{1,}$ $b_{2,}$ $b_{3.}$ etc. are coefficients for independent variables and b_0 is the intercept which corresponds to the value of y when all independent variables are zero.

Once we have the equation as mentioned above we can predict the values of y based on the values of independent variables.

<u>Cost Function</u>

The model gets the best regression fit line by finding the best values for b0, b1, b2, b3 ... bn . To find the best fit line we use cost function, where Cost Function (C) can be written as:

Cost Function (C) = $[(y_{pred(0)} – y_{(0)})^2 + (y_{pred(1)} – y_{(1)})^2 + (y_{pred(2)} – y_{(2)})^2 ...+ (y_{pred(n)} – y_{(n)})^2]/n.$

As we can see cost function is the error or difference between the predicted value $y_{pred(i)}$ and the true value $y_{(i)}$. Or more strictly speaking Mean Squared Error between the predicted value and actual value.

It is important to mention the assumption taken for applying Linear Regression:

- Target variable is linearly dependent on independent variables
- Multivariate Normality
- No or little multicollinearity
- No autocorrelation
- Homoscedasticity

General Subjective Questions - Answer (Q1)

Achieving best-fit regression line

To achieve the best-fit regression line, our ideal target is to ensure Cost Function (C) is zero. That means there is no difference between the predicted values and actual values of target variable. But realistically speaking our target is to minimize the Cost Function (C). To achieve the realistic target, we need to derive b_0, b_1, b_2, b_3 ... b_n such that Cost Function (C) is minimal.

Gradient Descent

To update b_0, b_1, b_2, b_3 ... b_n in order to minimize the Cost Function and achieve the best-fit regression line the model uses Gradient Descent. The idea is to start with random values for b_0, b_1, b_2, b_3 ... b_n and then iteratively update the values reaching minimum cost.

A Gradient Descent is the derivative that defines the effects on outputs of the function with a small variation in inputs.

The algorithm starts with assuming certain values for b_0, b_1, b_2, b_3 ... b_n (usually starts with 0) and calculate the Mean Squared Error (MSE). Then we reduce the values of b_0, b_1, b_2, b_3 ... b_n by some amount (called the learning rate). We will notice a decrease in MSE. We will keep repeating this steps until our Cost Function (i.e., MSE) is a very small value or 0.

Please note, if we give very high value for learning rate we may jump over the goal (as pointed out in the image on right). If we give very low value of learning rate, we will reach the goal slowly.

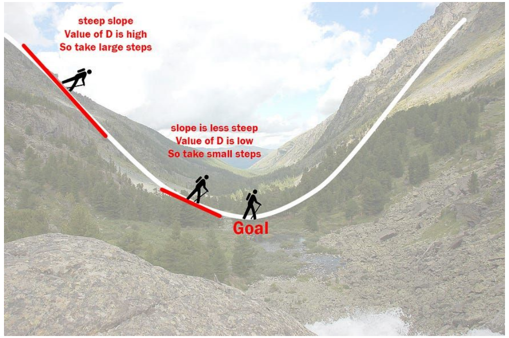


Image Courtesy: https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c8700931

General Subjective Questions - Answer (Q1)

If we differentiate Cost Function (C) w.r.t. b_0 , we will have below result:

$$D_{b0} = 2\,[(y_{pred(0)} - y_{(0)}) + (y_{pred(1)} - y_{(1)}) + (y_{pred(2)} - y_{(2)}) \ldots + (y_{pred(n)} - y_{(n)})]/n$$

If we differentiate Cost Function (C) w.r.t. b_1 , we will have below result:

$$D_{b1} = 2\,[(y_{pred(0)} - y_{(0)}) + (y_{pred(1)} - y_{(1)}) + (y_{pred(2)} - y_{(2)}) \ldots + (y_{pred(n)} - y_{(n)})]*x_1/n$$

$b0 = b0 - L*D_{b0}$

$b1 = b1 - L*D_{b1}$

We keep repeating this step until MSE is very small or 0.

The values of b_0, b_1, b_2, b_3 … b_n corresponding to the situation where MSE is very small or 0 is what we are looking for.

We can now use b_0, b_1, b_2, b_3 … b_n to make predictions using our model.

General Subjective Questions - Answer (Q2)

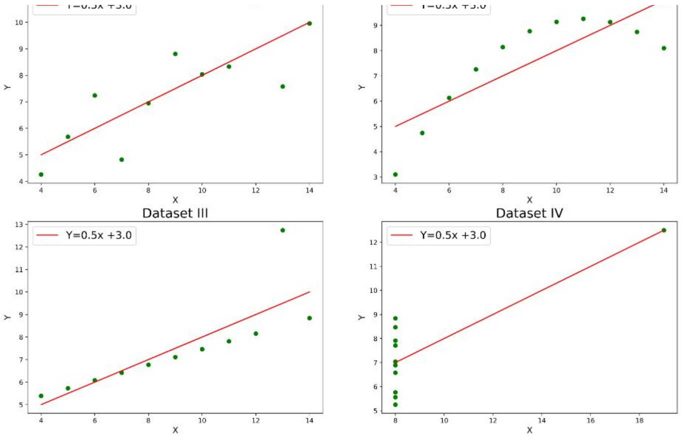
Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four dataset created by Francis Anscombe in 1973. The intention was to demonstrate the importance of visualizing data before performing statistical analysis on the data. Below is the dataset from here, representing Anscombe's quartet of datasets (Fig. 1):

Fig. 1

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Fig. 2



For the datasets: mean, variance, correlation, linear regression slope and linear regression intercept all have same values, but data visualizations looks as in Fig. 2.

Image Courtesy: https://www.geeksforgeeks.org/anscombes-quartet/

This clearly shows that data visualization is extremely important before interpreting any summary statistics.

General Subjective Questions - Answer (Q3 & Q4)

What is Pearson's R? (3 marks)

Pearson's R is a statistic that measures the linear relationship between two continuous variables. The numerical value of Pearson's R lies between -1 to 1. A negative value indicates inverse correlation i.e., if X increases Y will decrease. A positive value indicates direct correlation i.e., if X increases Y will also increase. If the Pearson's R correlation coefficient value is 0, then there is no relationship between the variables.

Pearson's correlation coefficient is used when:

- When both variables are quantitative.
- Variables are normally distributed.
- Data has no outliers.
- Variables have a linear relationship.

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

VIF (Value Inflation Factor) is a measure of correlation of a variable with other independent variables in multiple linear regression. If the value of VIF is high for a variable it indicates that variable has a strong correlation with other variables. If the correlation is perfect, then VIF value can be infinite.
In numerical terms VIF = $1/(1 – R^2)$, if R^2 is 1 then VIF will reach infinity.

General Subjective Questions - Answer (Q5)

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

A	B
20	14098
35	18109
40	10092

Scaling is a technique by which all continuous variables are brought to the same level of numerical representation. For example, if we have two variables A and B having values as shown on right side.

Scaling A and B will ensure that scaled values are comparable.

Scaling is performed for ease of data interpretation, faster convergence of gradient descent. Also, scaling only effects coefficients and none of the other parameters.

Standardized Scaling ensures that values for a variable are centered at 0 and has a standard deviation of 1. Formula for standardized scaling:

$x_i = (x_i - x_{mean})/x_{sd}$

Normalized Scaling ensures that values for a variable lie between 0 and 1. Formula for normalized scaling:

$x_i = (x_i - x_{min})/(x_{max} - x_{min})$

Only drawback with normalized scaling is that it will normalize outliers as well, which might not be desirable sometimes.

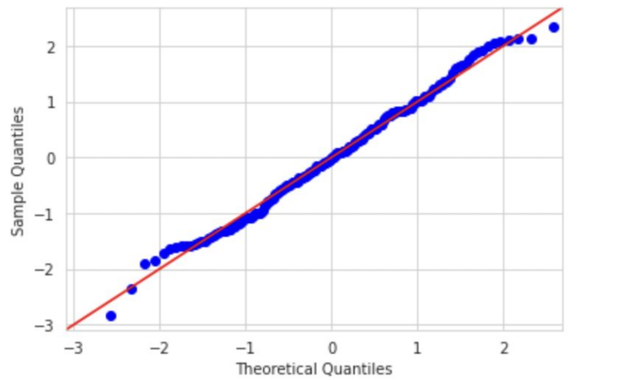
General Subjective Questions - Answer (Q6)

What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
The Q-Q plot or Quantile-Quantile plot is used to assess the distribution of a given population dataset. It can also help in understanding if two datasets are coming from the populations with same kind of distribution. A Q-Q plot is a plot of quantiles or percentiles of the first dataset against the percentiles of the second dataset.

For linear regression, if train data and test data is received separately then we can use Q-Q plot to assert if both datasets are from populations with same distribution.

Q-Q plot is advantageous:

- As it can be used with different sample sizes also
- Many distributional aspects like shifts in location, shift in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- For linear regression, q-q plot can be used to verify if residual error terms follows a normal distribution
- To determine skewness of distribution



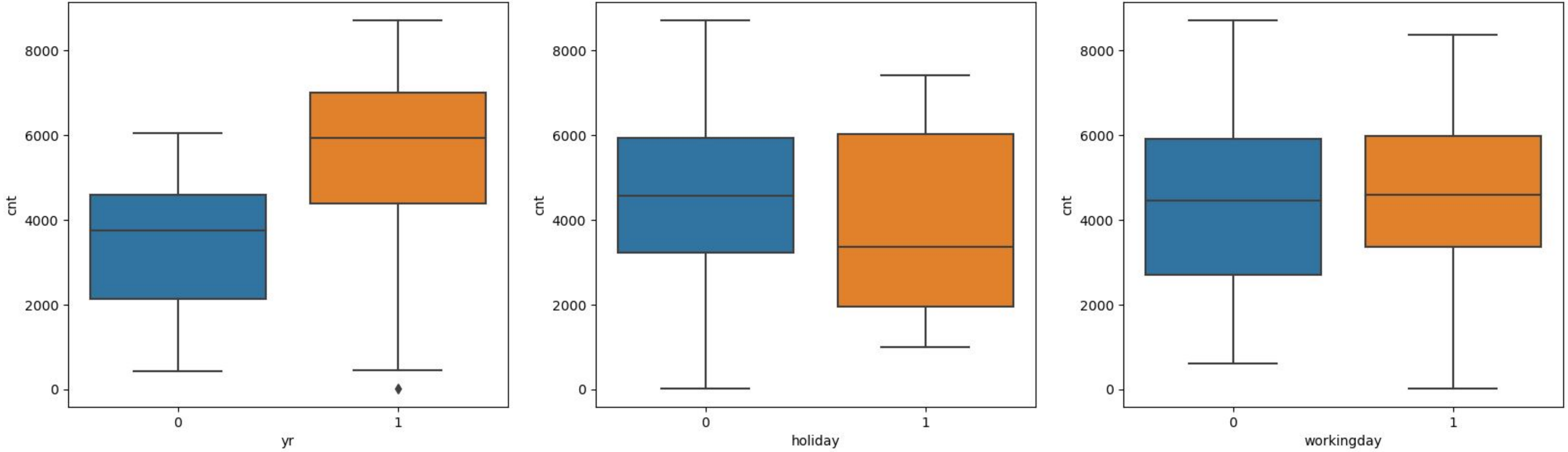
In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

Assignment based Subjective Questions

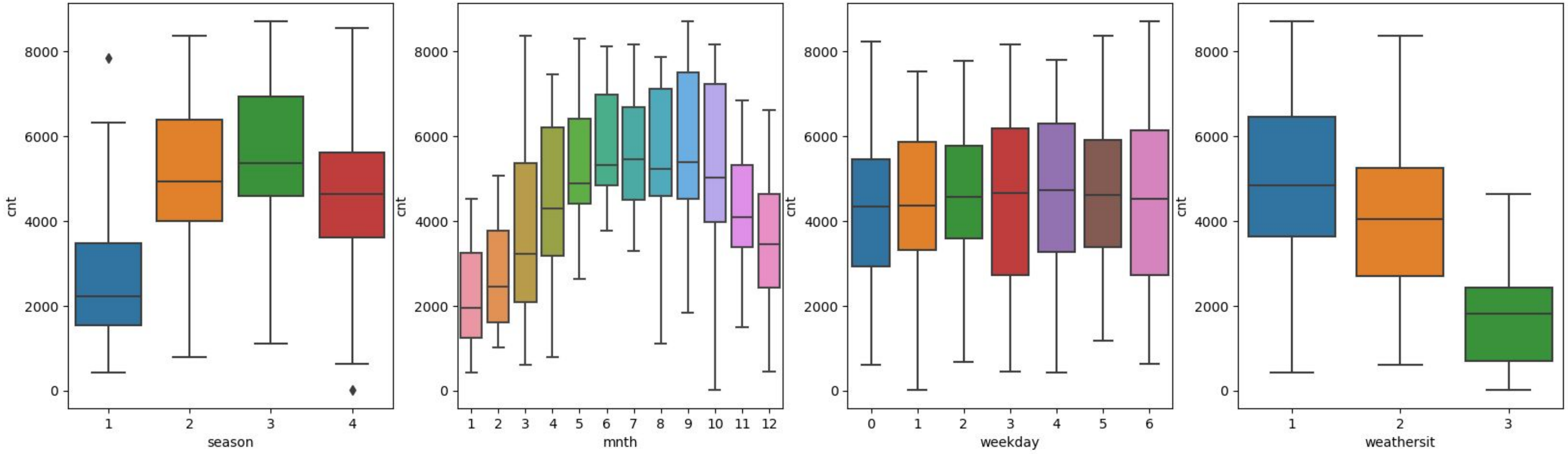
Assignment based Subjective Questions - Answer (Q1)

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
We have 7 categorical variables namely, season, yr (year), mnth (month), holiday, weekday, workingday, weathersit.

Below is the boxplot for `yr`, `holiday` and `workingday`



Assignment based Subjective Questions - Answer (Q1)



Following are the observations:

1. Count is significantly lower for spring season compared to other seasons
2. There is an outlier under spring season and under winter season
3. September has the highest interquartile range compared to other months
4. There is no data for heavy rain weather
5. For light rain weather, 75 percentile is lower than 25 percentile for clear and misty weather
6. Year 2019 has more count than in year 2018. And year 2019 has one outlier.

Assignment based Subjective Questions - Answer (Q2)

Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Let's consider with an example from the current assignment, taking `weekday` variable. So, for 7 days in the week below is how data will look like:

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

If we view the first row, Sunday can be represented by all the other variables having value as 0. So, if we drop first column representing Sunday then also whenever Sunday must be represented all other weekday values will be 0. This also ensures that while running p-value and VIF analysis this redundant column is not included.

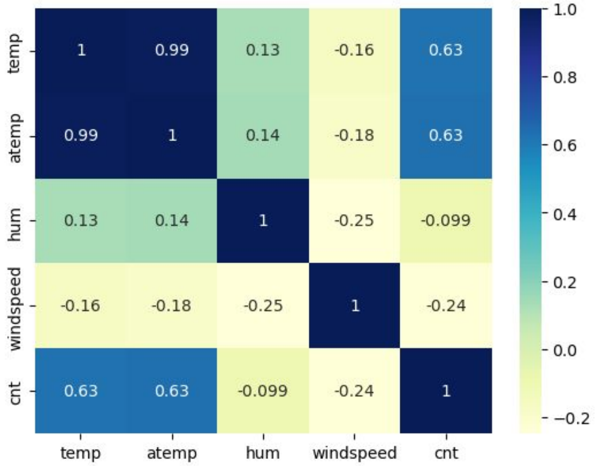
Please note, it is not necessary to remove first column only it can be any column ideally. Any column can be dropped for analysis if we are able to represent the values of the dropped column with the remaining dummy columns.

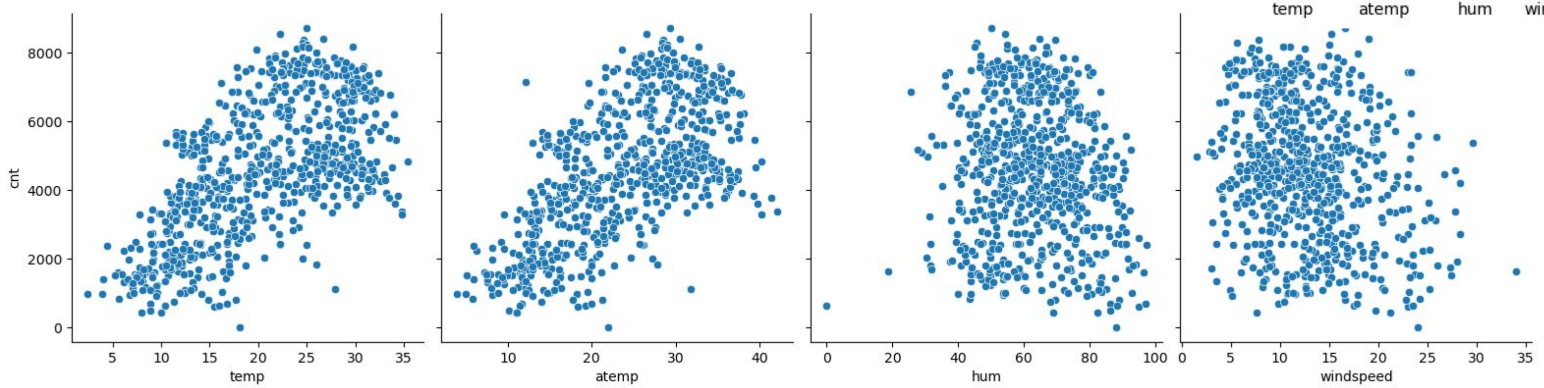
Assignment based Subjective Questions - Answer (Q3)

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

temp and **atemp** variables has highest correlation as is evident from the below pair-plot and heat map.

But it also needs to be mentioned that scatter plot for temp and atemp looks very similar. Also in the heat map it is visible that temp and atemp are highly correlated.





Assignment based Subjective Questions - Answer (Q4)

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

The assumptions of Linear Regression are:

- Target variable is linearly dependent on independent variables: By checking scatter plots it becomes easy to visualize a linear relationship for both continuous and categorical variables. Fig. 1
- Multivariate Normality: By checking the normal distribution of the residual errors. Fig 2
- No or little multicollinearity: By checking the correlation heat map and by checking the VIF value and ensuring that the value <= 5 (please check screenshot on next slide (slide # 18)
- No autocorrelation: By using VIF and p-value to keep only variables not showing multicollinearity.
- Homoscedasticity: By visualizing scatter plot between y_test and residual on test data. Fig 3.

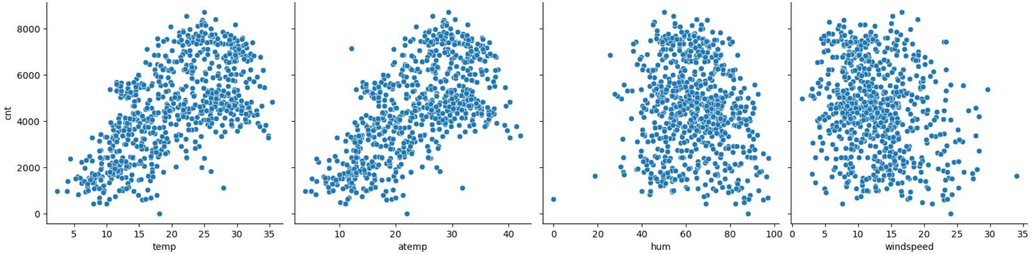


Fig. 1

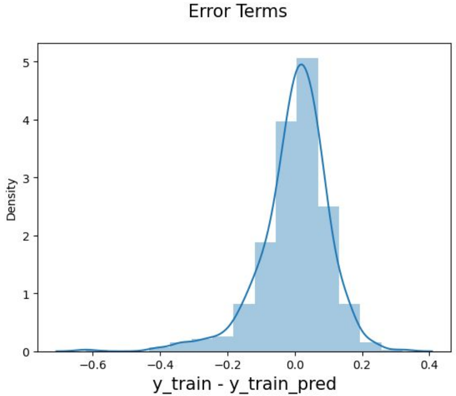


Fig. 2

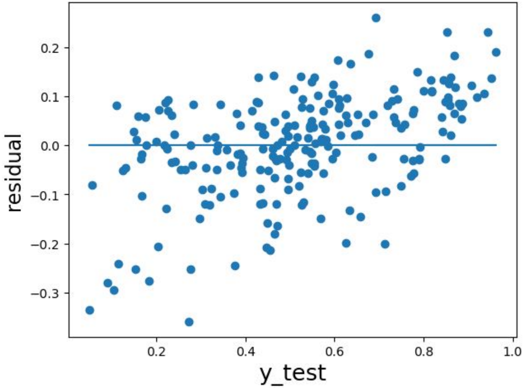


Fig. 3

Assignment based Subjective Questions - Answer (Q5)

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

The top 3 features are:

- **Temperature**: Has a major impact on the demand, as the temp gets more pleasant that is when temperature increases it corresponds increase in bike demands.
- **Year**: We observe an increase in demand in year 2019 compared to year 2018.
- **Wind Speed**: Affects the target variable inversely. That means the demand for bikes comes when wind speed is high.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.790
Method:	Least Squares	F-statistic:	193.0
Date:	Tue, 11 Jul 2023	Prob (F-statistic):	2.29e-164
Time:	23:57:08	Log-Likelihood:	442.03
No. Observations:	510	AIC:	-862.1
Df Residuals:	499	BIC:	-815.5
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0850	0.021	4.067	0.000	0.044	0.126
yr	0.2385	0.009	26.014	0.000	0.221	0.257
workingday	0.0466	0.012	3.740	0.000	0.022	0.071
temp	0.5208	0.024	21.308	0.000	0.473	0.569
windspeed	-0.1802	0.028	-6.461	0.000	-0.235	-0.125
season_summer	0.1005	0.012	8.124	0.000	0.076	0.125
season_winter	0.1252	0.012	10.628	0.000	0.102	0.148
mnth_aug	0.0540	0.019	2.911	0.004	0.018	0.090
mnth_sep	0.1026	0.018	5.595	0.000	0.067	0.139
weekday_sat	0.0570	0.016	3.548	0.000	0.025	0.089
weathersit_misty	-0.0704	0.010	-7.223	0.000	-0.090	-0.051

Omnibus:	136.973	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	488.719
Skew:	-1.201	Prob(JB):	7.52e-107
Kurtosis:	7.150	Cond. No.	11.8

Thank You