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?

Ans: The identified Categorical variables were

a)season

Season2 and season 4 have coefficients 0.08 and 0.12 in the final expression

b)weatersit

weathersit_2 has -0.07 co efficient, which means when this variable increases by 1 , cnt decreases by 0.07 units

c)mnth

mnth_9 has co effient of 0.094

d)weekday

weekday6 has 0.05 as its coefficient

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

It helps in reducing the extra column created during dummy variable creation. Therefore reducing correlation between dummy variable.

A simple way to make it clear is if there 3 values for a categorical variable

Value1

Value2

Value3

We can represent this with only 2 variables

Dummy variable1

Dummy variable2

Values	Dummy Variable 1	Dummy Variable 2
Value1	0	0
Value2	1	0
Value3	0	1

Here only 2 variables were used to represent 3 values reducing once column value

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

Ans: Temp and atemp variable has highest correlation with target variable(pair plots can be found in notebook)

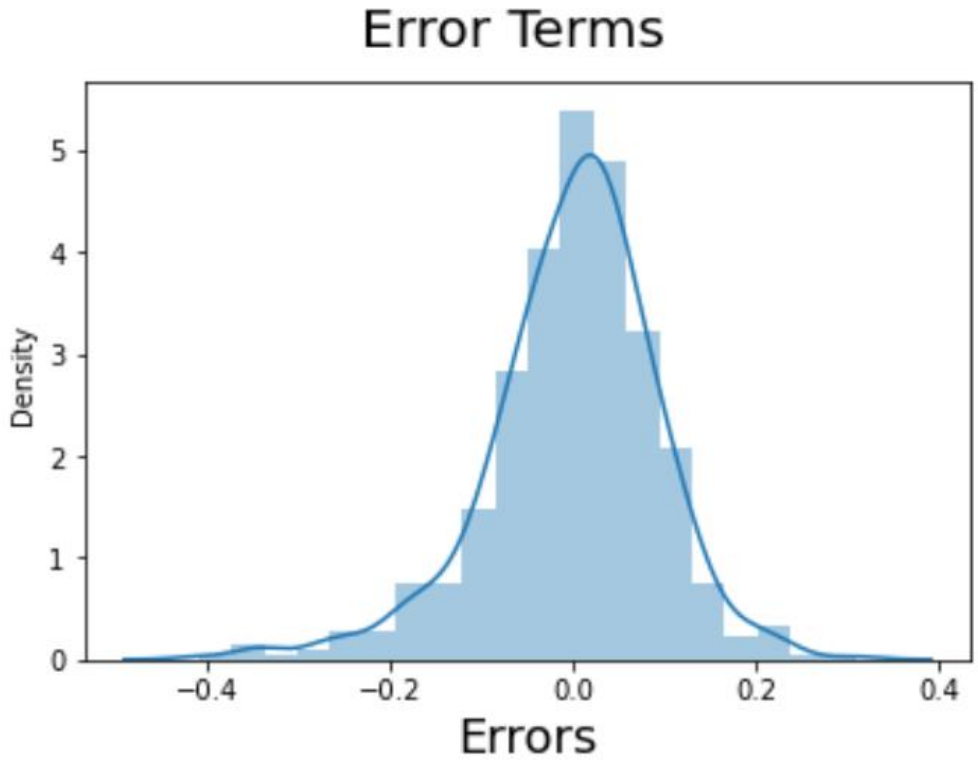
It is followed by yr variable

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

Ans: assumptions are linear regression are

1)Error terms should have a normal distribution

```
res = y_train-y_train_pred

# Plot the histogram of the error terms

fig = plt.figure()

sns.distplot((res), bins = 20)

fig.suptitle('Error Terms', fontsize = 20)              # Plot heading

plt.xlabel('Errors', fontsize = 18)
```



2)Does X and Y have linear relation ship

Pairplots were done to determine the linear relation ship between cnt and temp and cnt and yr

3)Multicollinearity : There should not be multicollinearity that should exist between predictor variables

	Features	VIF
2	temp	4.72
3	windspeed	4.02
1	workingday	4.01
0	yr	2.00
7	weekday_6	1.65
4	season_2	1.56
8	weathersit_2	1.52
5	season_4	1.38
6	mnth_9	1.20
9	weathersit_3	1.07

As VIF is less than 5, we have proved that no multicollinearity exists

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

Final Equation:

cnt = 0.084143 + (yr × 0.230846) + (workingday × 0.043203) + (temp × 0.563615) − (windspeed × 0.155191) + (season2 × 0.082706) + (season4 ×0.128744) + (mnth9 × 0.094743) + (weekday6 ×0.056909) − (weathersit2 × 0.074807) − (weathersit3 × 0.306992)

based on the values here temp, and yr have positive effect

windspeed has negative effect in explaining cnt

General Subjective Questions

1> Explain linear regression algorithm in detail

- It is a machine learning algorithm based on supervised learning
- It performs regression ask
- Target column value is predicted with the help of independent variables
- Hypothesis for linear regression goes like this

  Y=mx+c

  Where y is the dependant variable value

  x-independent variable

  m- coeofficient of x

  c: intercept value

While training the model we are given

x: input training data (univariate – one input variable(parameter))

y: labels to data (Supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x.

The model gets the best regression fit line by finding the best m and
C values. **C:** intercept **m:** coefficient of x .

Once we find the best m and c values we fit the line

There is also Cost function, where we try to come up with best fit line, by reducing error difference between predicted and true value

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y)

The idea is to start with random m and c values and then iteratively updating the values, reaching minimum cost.

2**)Explain the Anscombe's quartet in detail.**

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

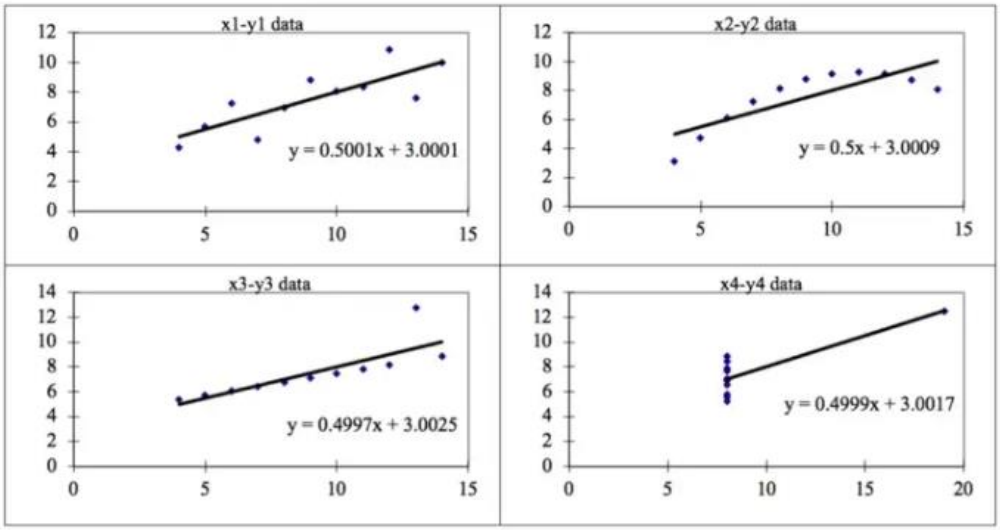


Image by Author

The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3) **What is Pearson's R**

It is the most common way of measuring a linear correlation

It can range between -1 and 1 which basically measures the strength and direction of the relationship between 2 variables

It is a positive correlation

If the value is 0: it indicates that there is no relationship between variables

If the value is 1 then if once variable increases the other variable also increases by equal proportion

If the value is between 0 and -1 It means if one variable changes, the other variable changes in opposite direction

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that ***scaling just affects the coefficients*** *and none of the other parameters like* ***t-statistic, F-statistic, p-values, R-squared****, etc.*

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and*

  1. ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale*** *helps to implement standardization in python.*

- *One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.*

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

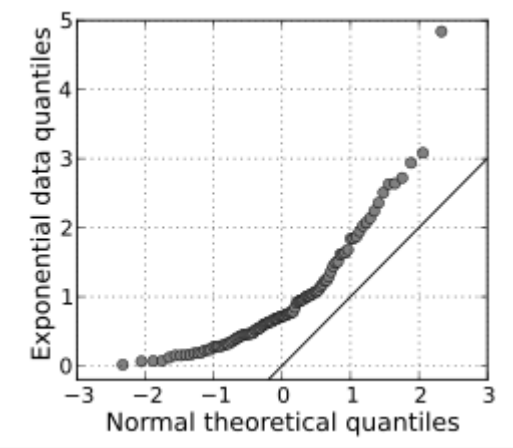
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately

lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.