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)

Based on the categorical variable analysis, below are the interpretations on their effect on dependent variable (cnt) :

- Temperate shows a strong linear association (& correlation) with bike rental count. This also aligns with business understanding as higher temperatures make it uncomfortable to ride bike.
- Bike rental business seems to be growing significantly showing ~60% higher sales in 2019. Sales majorly spiked from Q2 in 2019
- Fall & Summer season seem to have high rental counts followed by spring. Same trend is also observed from the monthly split of rental counts.
- In terms of weather, clear weather shows highest rentals as opposed to snow/rain/thunderstorm which showed very low rentals. This is self-explanatory as clear weather makes it comfortable to ride
- Working day or non-working day does not seem to have significant impact on rentals.
- Holiday shows low impact with less rentals on holidays but the impact is not very significant

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

drop_first=true in get_dummies method indicates to drop the first category of categorical variable. This is because only (n-1) columns are required to identify all categories of a categorical variable having n unique values. Reducing an extra variable input to our model is a positive as it reduces the correlation among dummy variables.

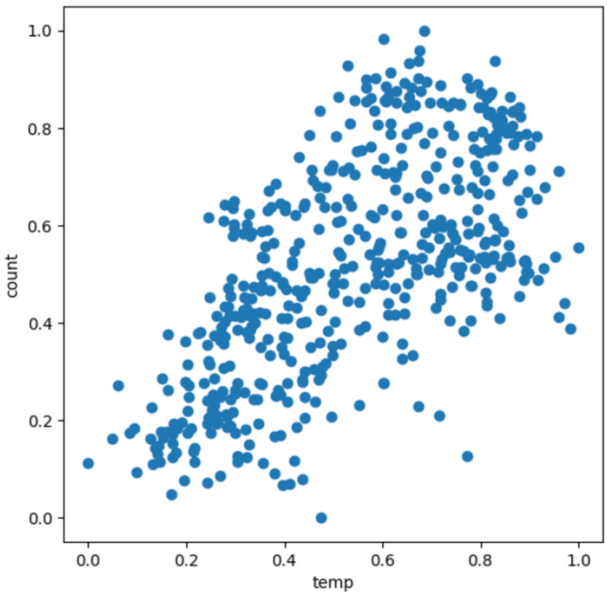
For ex : Consider a categorical variable with values ‘furnished’, ‘semi-furnished’, ‘unfurnished’.

There are 3 categories but we only need 2 dummy variables to define those :

- 01 indicating unfurnished
- 10 indicating semi-furnished
- 00 indicating furnished

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

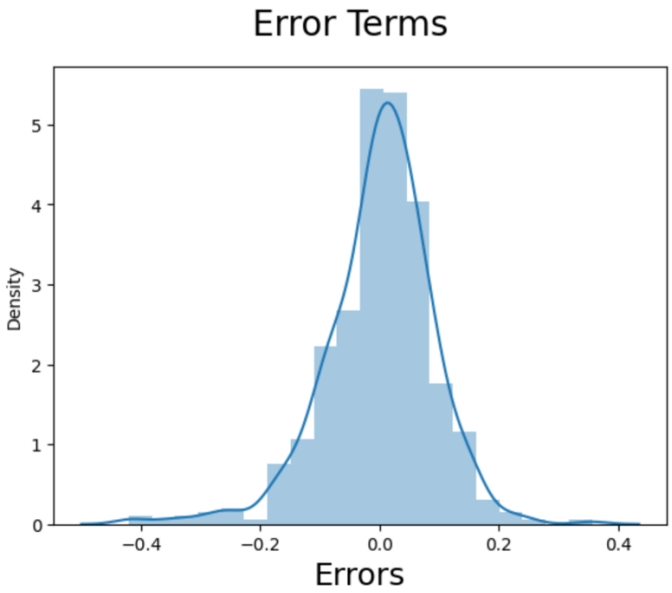
'temp' variable has highest correlation with the target variable



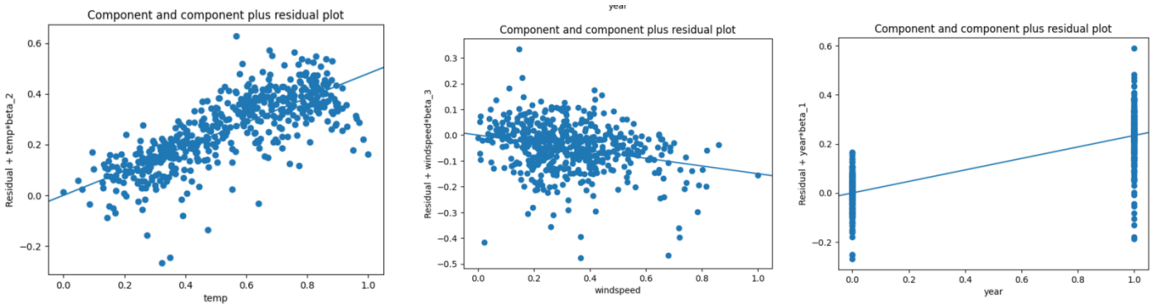
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Below steps were followed to validate assumptions of my final linear regression model –

i. Normal Distribution of Error terms : Verified as shown in below graph -

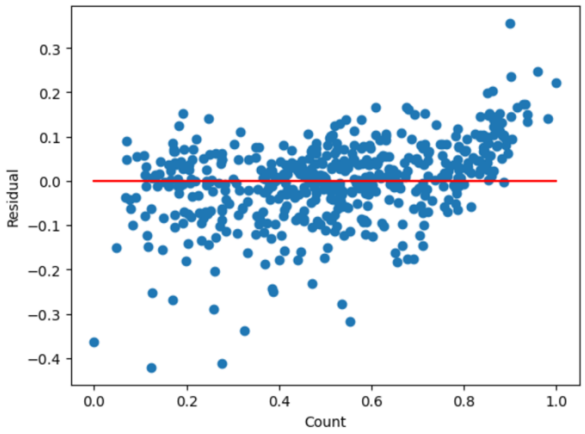


ii. Linearity : Verified we are not trying to fit a linear model into a non-linear, no-additive data. Shown in below graphs -



iii. Multicollinearity : Verified the predictor variables are not highly correlated (should have VIF < 5)

iv. Homoscedasticity : Verified almost constant variance in error terms



v. Independence & No endogeneity : Verified there is no relationship between errors and independent variables

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

Top 3 features impacting demand of shared bikes are –

1. Temp
2. Year
3. lightSnowRain (weather)

General Subjective Questions

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

Linear regression is a statistical model which estimates the linear relationship between the outcome variable (dependent variable) with 1 or more predictor variable (aka independent variables). When predictor variable is 1, it is called *simple linear regression* & when predictor variables are >1, it is called multiple linear regression.

Statistically, a linear relationship can be depicted as –

$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

where

- b_0 is called the coefficient (i.e. value of Y when all independent variables are 0)
- b_1, b_2, .. b_n are the coefficients of independent variables. To understand the meaning of coefficients, we can further say that 'if X1 value changes by 1 unit while all other variables stay constant, the value of Y will increase by b1'
- n is number of independent variables. For simple linear regression, n=1
- coefficients can be positive (>0) or negative (<0) indicating whether dependent variable will increase or decrease with change in given independent variable

For a linear regression model to be applicable for predicting data, below assumptions must hold true:

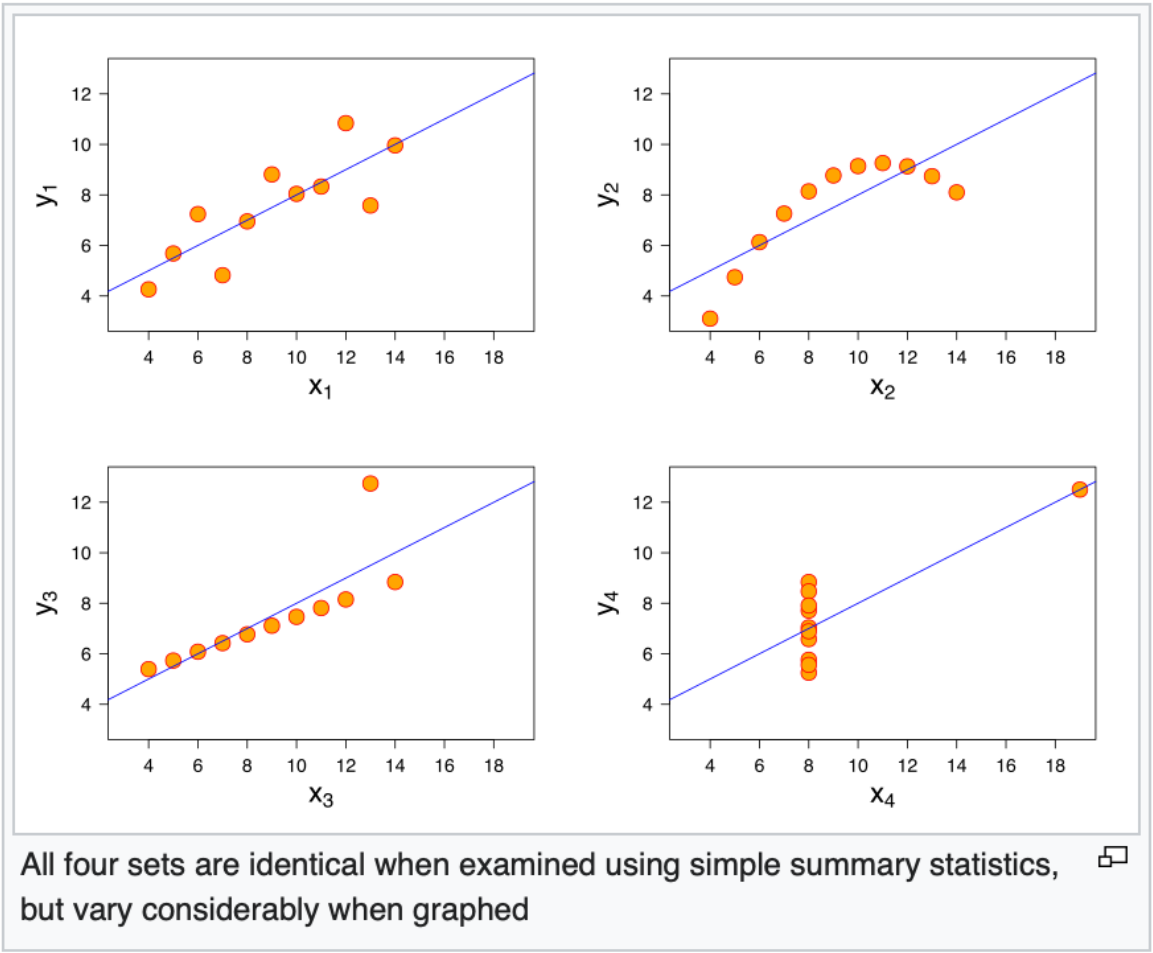
i. Normal Distribution of Error terms : All error terms should be normally distributed as shown & validated in Q3 above.
ii. Linearity : This means that should be able to validate a linear relationship between dependent & independent variables i.e. not try to fit a linear model into a non-linear, no-additive data.
iii. Multicollinearity : The independent variables should not be highly correlated
iv. Homoscedasticity : Error terms should be distributed with a constant variance
v. Independence & No endogeneity : There should be no clear relationship between errors and independent/dependent variables

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

Anscombe's quartet are a set of 4 datasets having almost identical statistics (mean, standard deviation, regression line) but appear very different when graphed due to their unique data distributions. Each of these datasets contain 11 datapoints.

The idea is to demonstrate the need for graphically visualising data instead of only relying on statistical properties. Looking at the data gives a lot of useful insights, one such insight being whether data is linearly related or not. Without validating linear relationship from diagrams, a linear regression association should not be assumed. For ex : See below the graphical representation of anscombe's quartlet

1. First dataset graph shows a simple linear relationship among 2 correlated variables
2. Second dataset graph shows some relationship but it is not linear. Linear regression might not be appropriately applicable here
3. Third dataset graph shows a linear relationship with a significant outlier, hence the linear regression fit line is different than dataset 1.
4. Forth dataset graph indicates that one strong outlier is enough to produce a high correlation

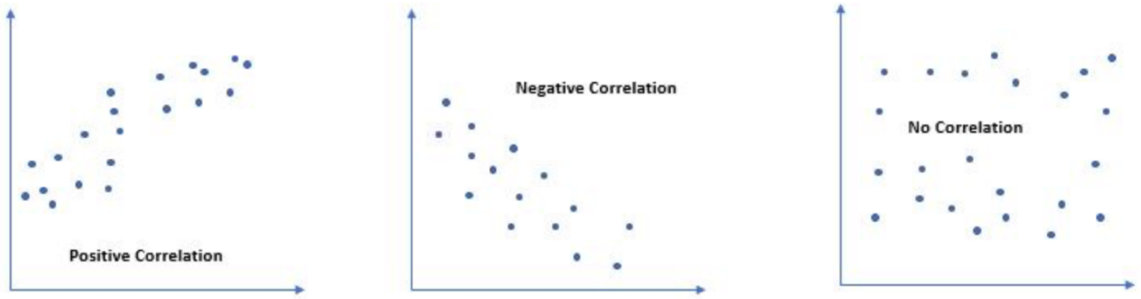


Above graphs taken from [wikipedia]

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

Pearson's R or Pearson's correlation coefficient indicates how strongly 2 variables are correlated or dependent on each other. For ex – if increase/decrease of 1 variable leads to increase/decrease of other variable, they are correlated.

Pearson's R value is a numerical value ranging between -1 to +1 where negative correlation (-1 to 0) between X & Y indicates that increase in X will lead to decrease in Y & vice-versa. Similarly, a positive correlation (0 to +1) between X & Y indicates that increase in X will lead to increase in Y & vice-versa. A value of -0.5 to 0.5 indicates weak correlation whereas a value of >0.5 or <-0.5 indicate that X and Y are strongly correlated (i.e. dependent).



Pearson's R has high significance in linear regression model as it helps us identify

1. Independent variable having strong correlation with outcome variable
2. Avoid multicollinearity by identifying independent variables which are strongly correlated among each other

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

Scaling is a method to normalise the all variables' data range into a given range. This is done during data preparation stage before creating the model to ensure all variables have data lying in the same range.

This makes the coefficients of all independent variables comparable and easy to understand in terms of relative impact of these variables on dependent variable. For ex – lets say we have independent variable age (0-100), salary (10000-100000) & travel time (0-40). We can see that all 3 variables have different data range. Scaling them would fit them into a fixed data range say 0 to 1 depending on scaling method we use.

A model with data scaling performs much better than a model without data scaling.

Difference between normalised scaling and standardised scaling-

Normalised Scaling (min-max scaling)	**Standardised Scaling**
Scaling is done using min & max of given variable range	Scaling is done using mean & standard deviation of given variable dataset
Formula : (x – xmin) / (xmax – xmin)	Formula : (x – mu) / sigma
Scales values between 0 to 1	No fixed outcome range
Highly affected by outliers	Less affected by outliers

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

VIF is a measure of multicollinearity in regression analysis i.e. how strongly an independent variable is correlated with other independent variables. A VIF value infinite indicates a perfect correlation i.e. R2 = 1 leading to VIF=Infinite as VIF = $1/(1-R^2)$. This means we need to drop one of the variables leading to perfect collinearity (having infinite VIF) to avoid multicollinearity in our model.

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 is a scatterplot which plots 2 sets of quantiles against each other and helps identify if both came from the same population with a common distribution. Quantile mean the fraction/percentage of points below the given value. For ex – 0.3 quantile is the point at which 30% of the data falls below it.

Use of Q-Q Plot :
The pattern of points in a Q-Q plot is used to compare 2 distributions. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q Plot :
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.