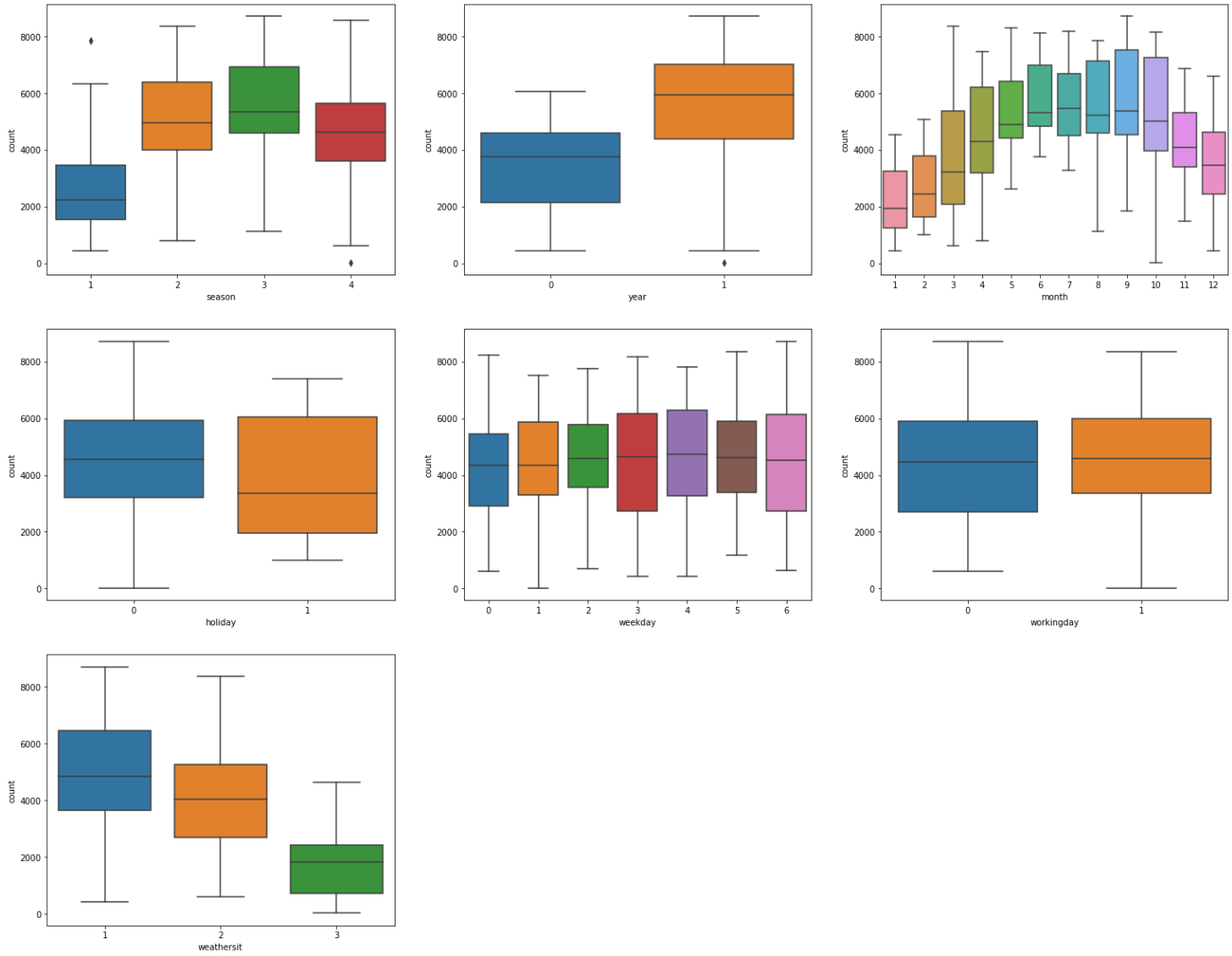
Assignment-based Subjective Questions

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

Categorical Features in Data:

'season', 'year', 'month', 'holiday', 'weekday', 'workingday','weathersit'

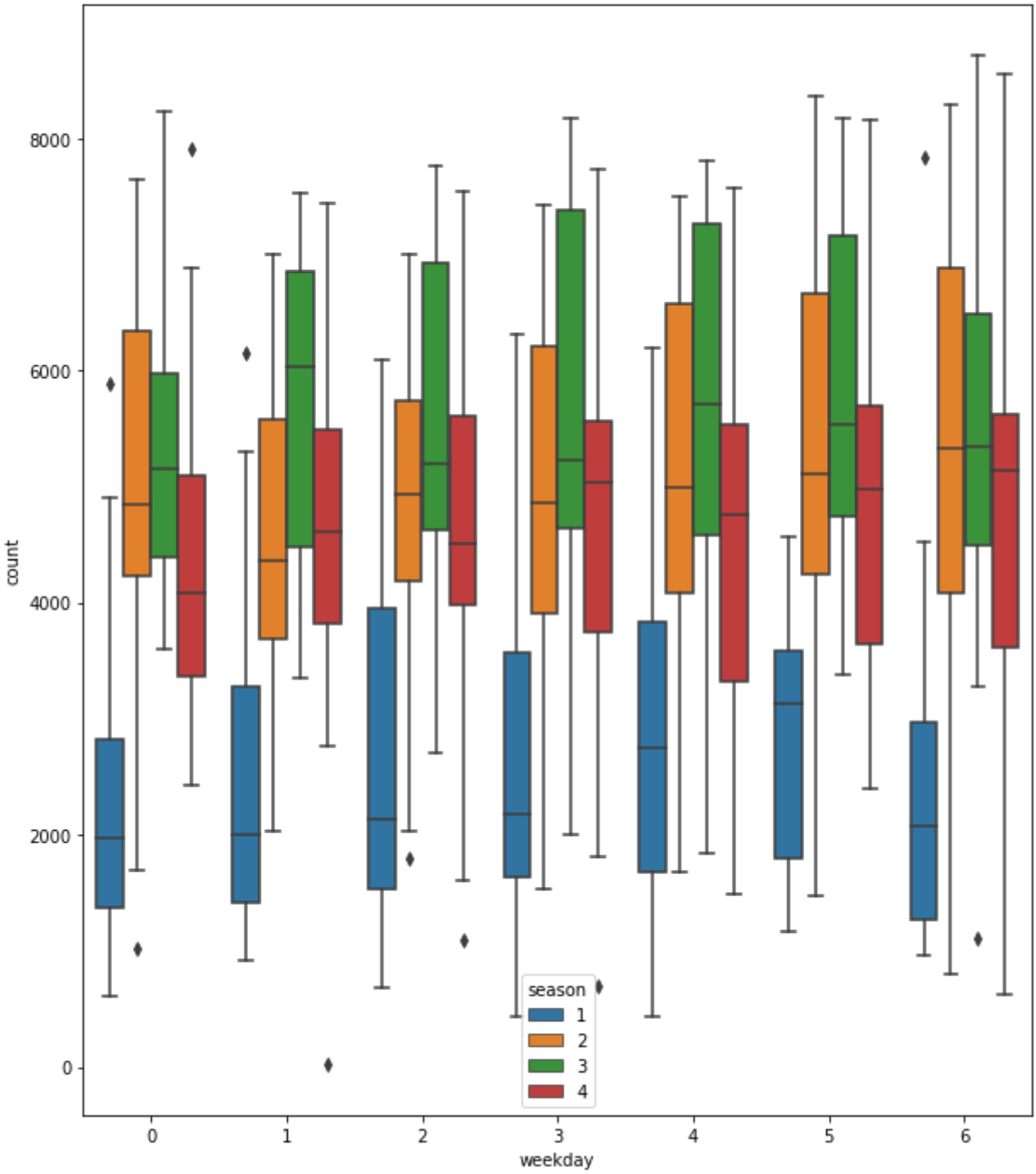
Analysis of categorical variable with target variable count



Findings:

1. We can see season 2 and season 3 has more number of bike rentals than season 1 and season 4.
2. We can see year 2019 is having more number of bike rentals than 2018.
3. We can see Clear weather is having more rentals compare to other categories.
4. Mid of the year is having more rentals than start and end of the year.
5. Working day does not have much impact on the target variable.
6. The median of bike rented on holiday is less than the bike rented on non-holiday.

Analysis of working day with target variable count considering hue as season



Findings:

1. We can see season 1 i.e Spring season is having less number of bike rentals compared to other seasons

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

drop = True is important as it help in reducing the extra column created during creation of dummy variable. Hence it reduces the corrleation created among dummy variable.

E.g. Lets say we have categorical variable name ***Investment*** which has three categories as *short term, medium term and long term* . So, if the variable is not long term and medium term then obviously its short term. So, we dont need third variable to identify short term Combination could be:

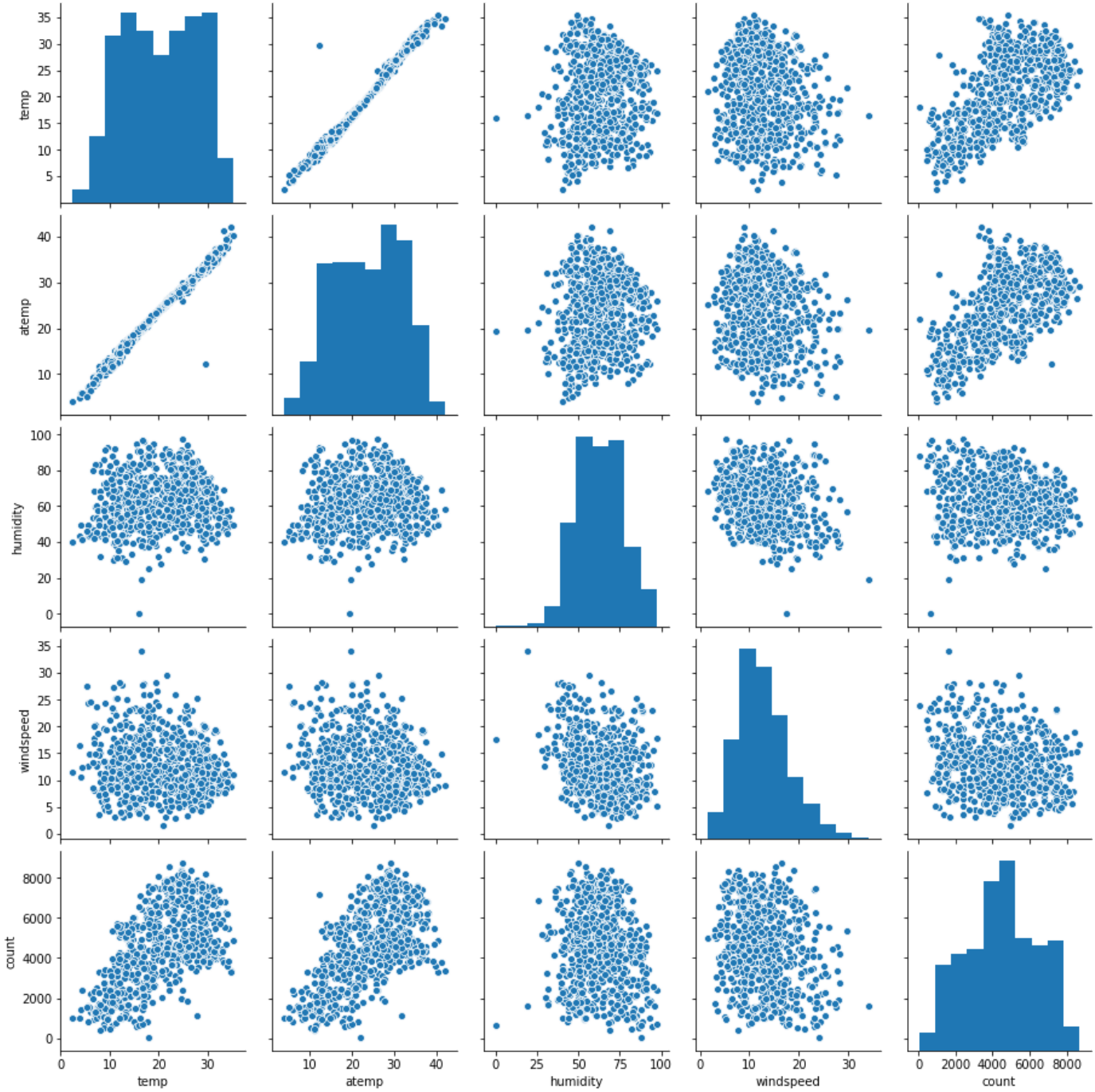
00 => Short term 10 => medium term 01 => long term

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

Numerical variable in dataset:

'temp', 'atemp', 'humidity', 'windspeed'

Pair plot:



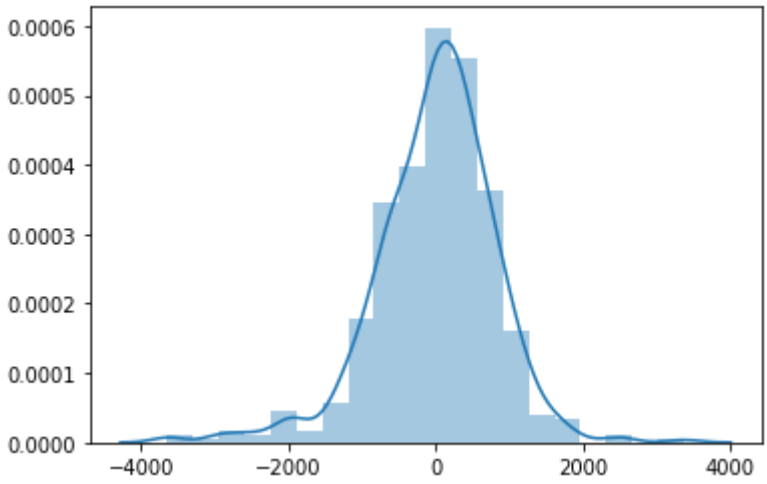
Findings:

1. We can see from the plot both temp and atemp is having highest(almost equal) correlation witht he target variable.
2. As temp and atemp are highly correlated, so in our modelling we have dropped temp as atemp(feel like temperature) is making more sense from user point of view.

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

Assumptions

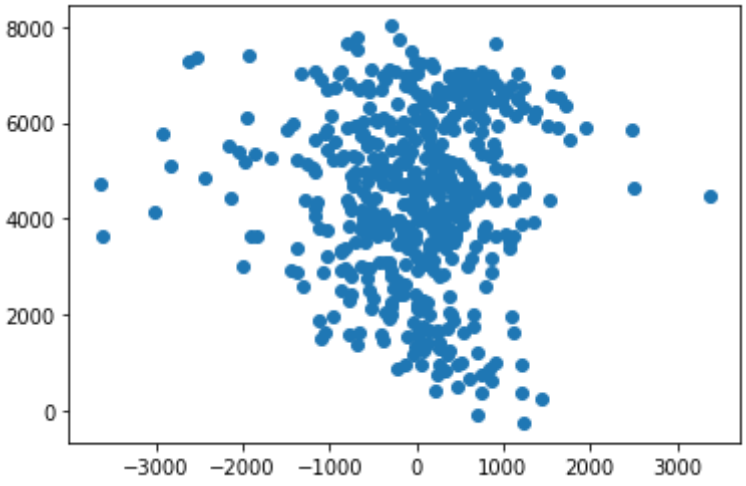
1. Error terms are normally distributed: So we plotted a displot of residual term



2. No perfect multicollinearity: For this we checked the final model VIF values

Features VIF
workingday 2.75
weathersit_2 2.27
season_4 2.09
year 1.88
humidity 1.82
season_2 1.63
month_10 1.63
atemp 1.55
month_8 1.49
weekday_6 1.41
month_9 1.28
weathersit_3 1.28
windspeed 1.19

3. No Heteroskedasticity: plot a scatter plot between residual and fitted values and if the plot is looking like a funnel shape then heteroskedasticity is present. But in our data its not there.



5. Based on the final model, which are the top 3 features contributing

significantly towards explaining the demand of the shared bikes?

Importance Features

4 important features are: atemp, year, season_4 and month_9

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. The main idea of regression is to examine two things. First, does a set of predictor variables do a good job in predicting an outcome (dependent) variable? The second thing is which variables are significant predictors of the outcome variable?

A linear regression line equation is written in the form of:

Y = c + mX

where Y is the target variable, c is intercept, m is slope and X is data point.

Types of Linear Regression
1. Simple Linear Regression
2. Multiple Linear Regression

Simple linear regression

In Simple linear regresion we are going to use one independent variable for the prediction of traget variable. The equation of regression line in this case is:
Y = c + mX

Multiple Linear regression

In multiple linear regression we are going to use multiple independent variable for the prediction of target variable. The regression line equation in this case is: Y = c + m1X1 + m2X2 + m3X3 + ... + mnXn
where m1, m2, ... mn are the coefficient/slope

Assumptions of Linear Regression

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

1.R^2 or Coefficient of Determination

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

where RSS: Residual sum of squares, TSS: Sum of errors of data from mean

2. Residual Standard Error (RSE)

It is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

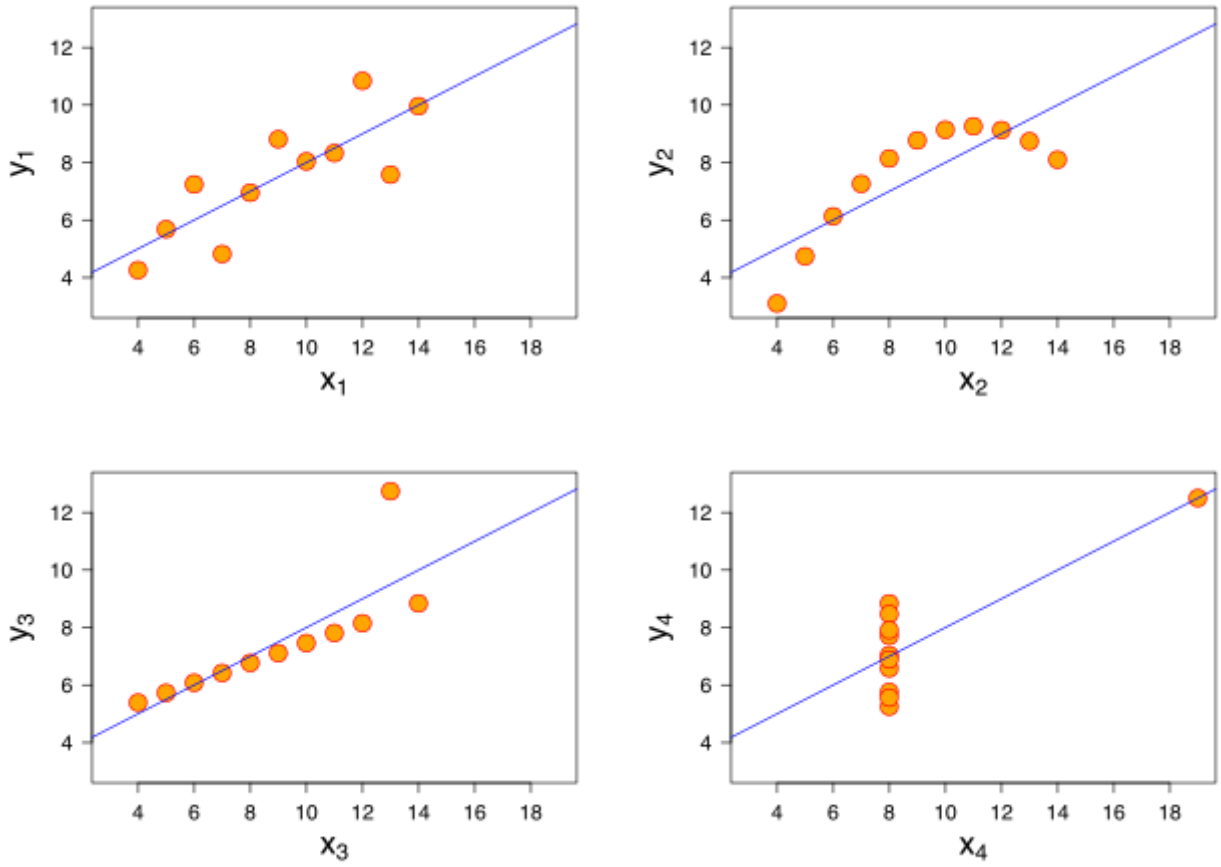
RSS = (sum(y_true - y_pred))2

Statistical point

1. Null hypothesis: all the coefficient are 0
2. Alternate hypothesis: coefficient are not 0
3. If we fail to reject null hypothesis then those variables are not significant.

2. Explain the Anscombe’s quartet in detail.|

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



for all four dataset

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
2. The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

How is the Correlation coefficient calculated?

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

The positive value of Pearson’s correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other.
The negative value of Pearson’s correlation coefficient implies that if we change either of these variables, there will be a negative effect on the other.

The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

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

Scaling is a technique to standardize/normalize the independent features present in the data in a fixed range. It is performed during the data pre-processing. Let suppose we have data in which 2 columns are there which are in different unit and different range, but machine will not understand the difference as those are just the numbers to machine. So we have to perform scaling to make them into same scale so that model can interpret from it. One more reason to perform scaling is that it help in gradient descent to converge easily.

Types of Scaling

1. Standardization
2. Normalization

Standardization
What it does to your variable is centering the data to a mean of 0 and standard deviation of 1.

Normalization

It also brings the data to same scale and brings the data to 0 to 1 range.

Difference

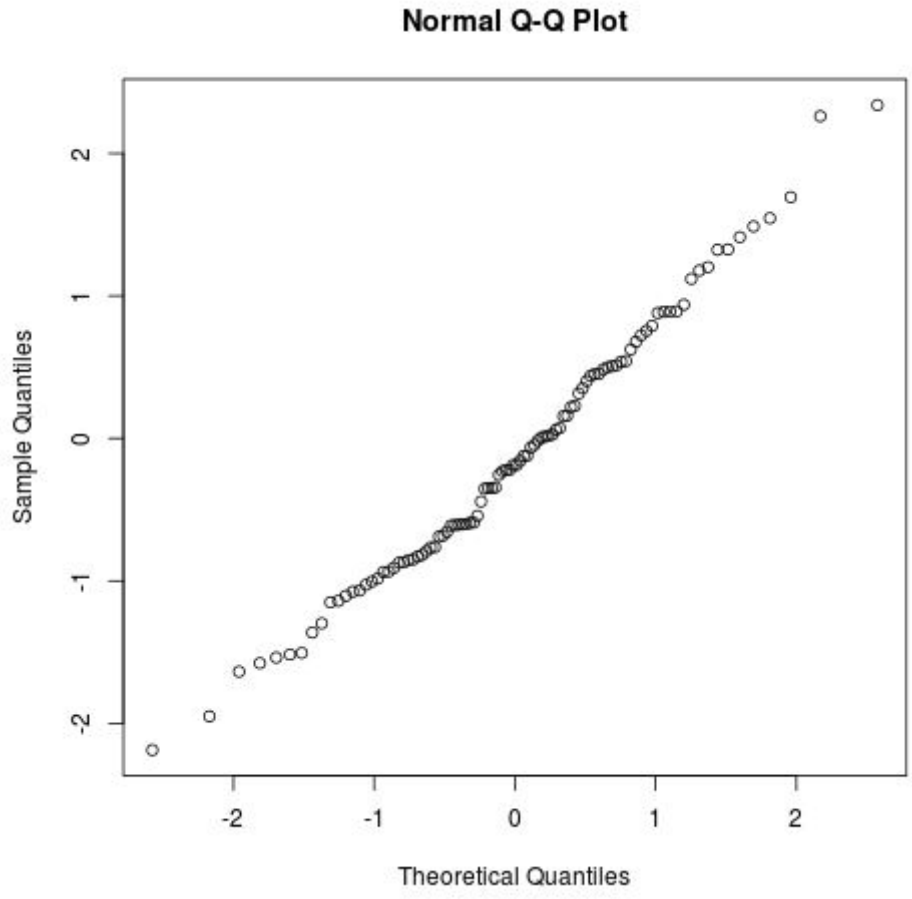
1. Minimum and maximum value of features are used in normalization but in standardization, Mean and standard deviation is used for scaling.
2. Normalization is used when feature are of different scale but standardization is used when we want to ensure zero mean and unit standard deviation.
3. Normalization scales values between [0, 1] or [-1, 1]. Standardization is not bounded to a certain range.
4. Normalization is really affected by outliers. Standardization is much less affected by outliers.

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

Yes, we have observed in given case study that few variables are having VIF as infinity. This is happening because the R2 value of those variable while calculating VIF is coming as 1. So as per the VIF formula in the denominator (1 - R2), if R2 is 1 the whole term will be 0, and if we have denominator 0, that leads to infinity.

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

A Q-Q plot, short for “quantile-quantile” plot, is a type of plot that we can use to determine whether or not a set of data potentially came from some theoretical distribution.



Although a Q-Q plot isn’t a formal statistical test, it does provide an easy way to visually check whether a dataset follows a normal distribution, and if not, how this assumption is violated and which data points potentially cause this violation.

In Linear regression we can check if the residual are following normal distribution or not using Q-Q plot.

In [ ]: