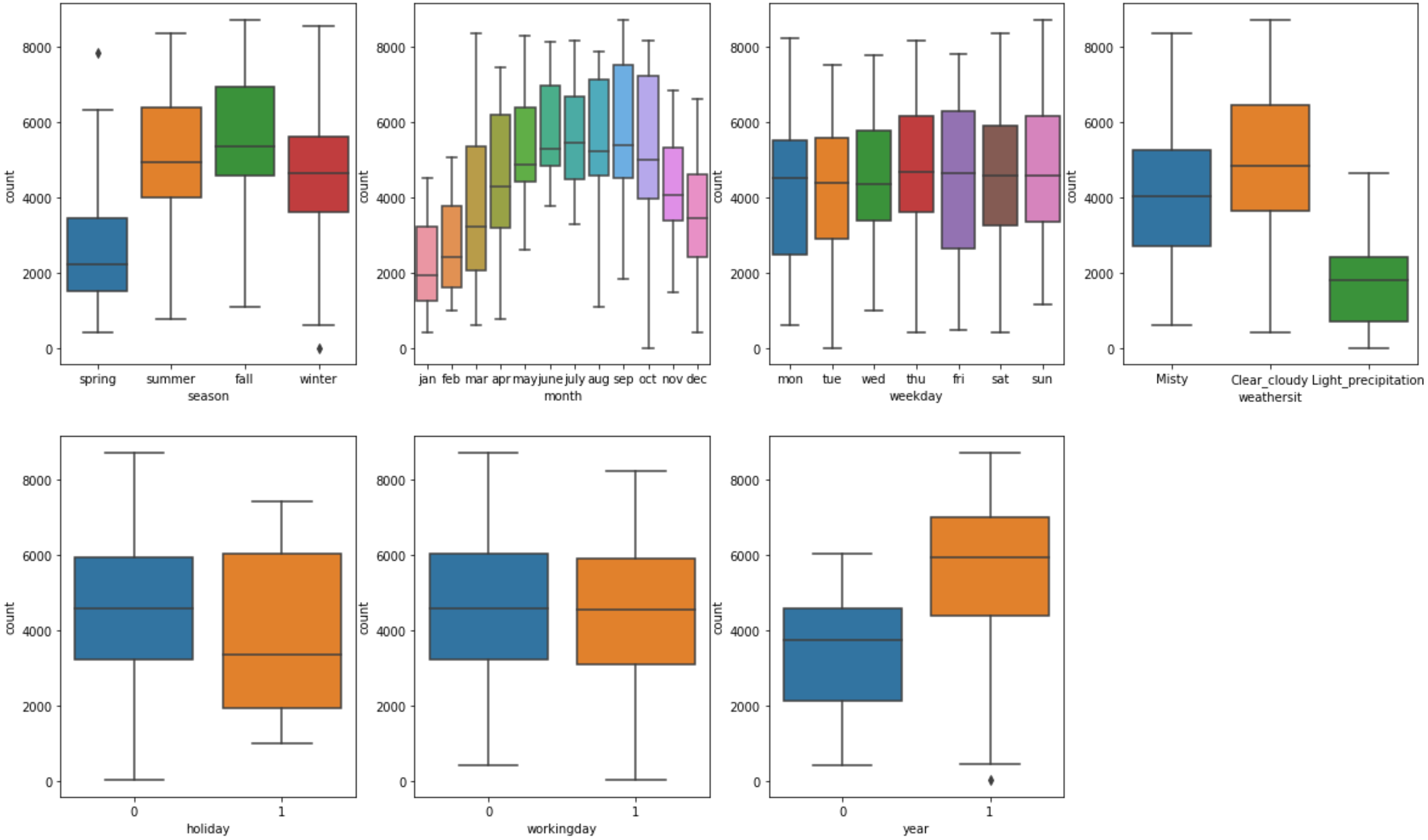
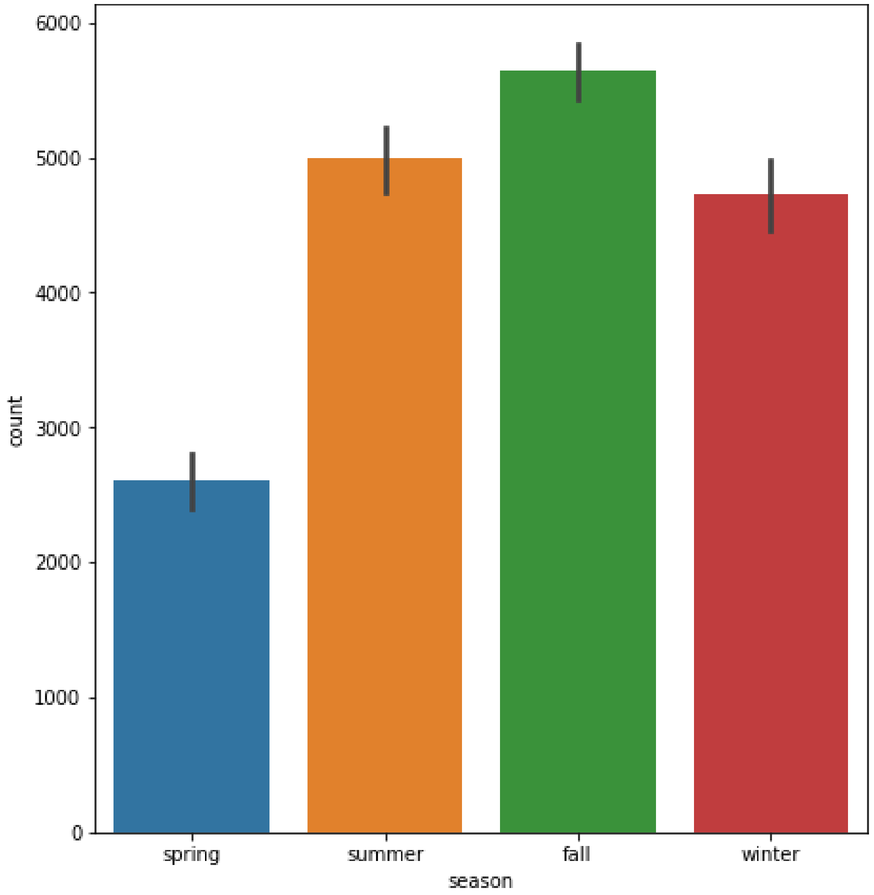
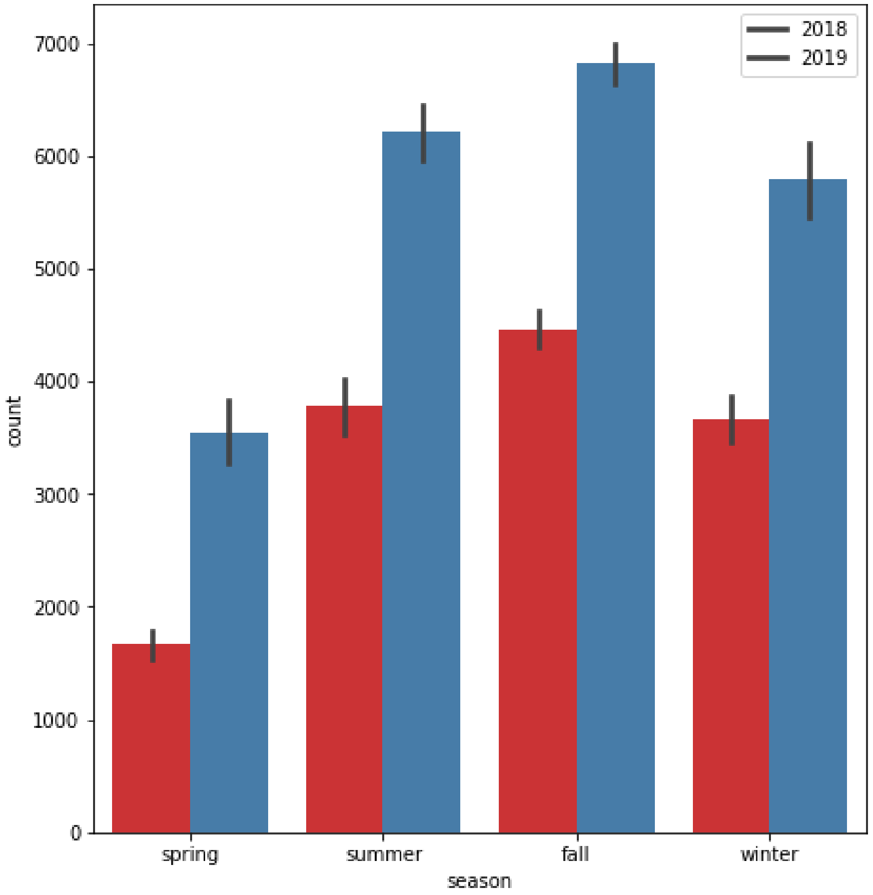
Assignment-based Subjective Questions

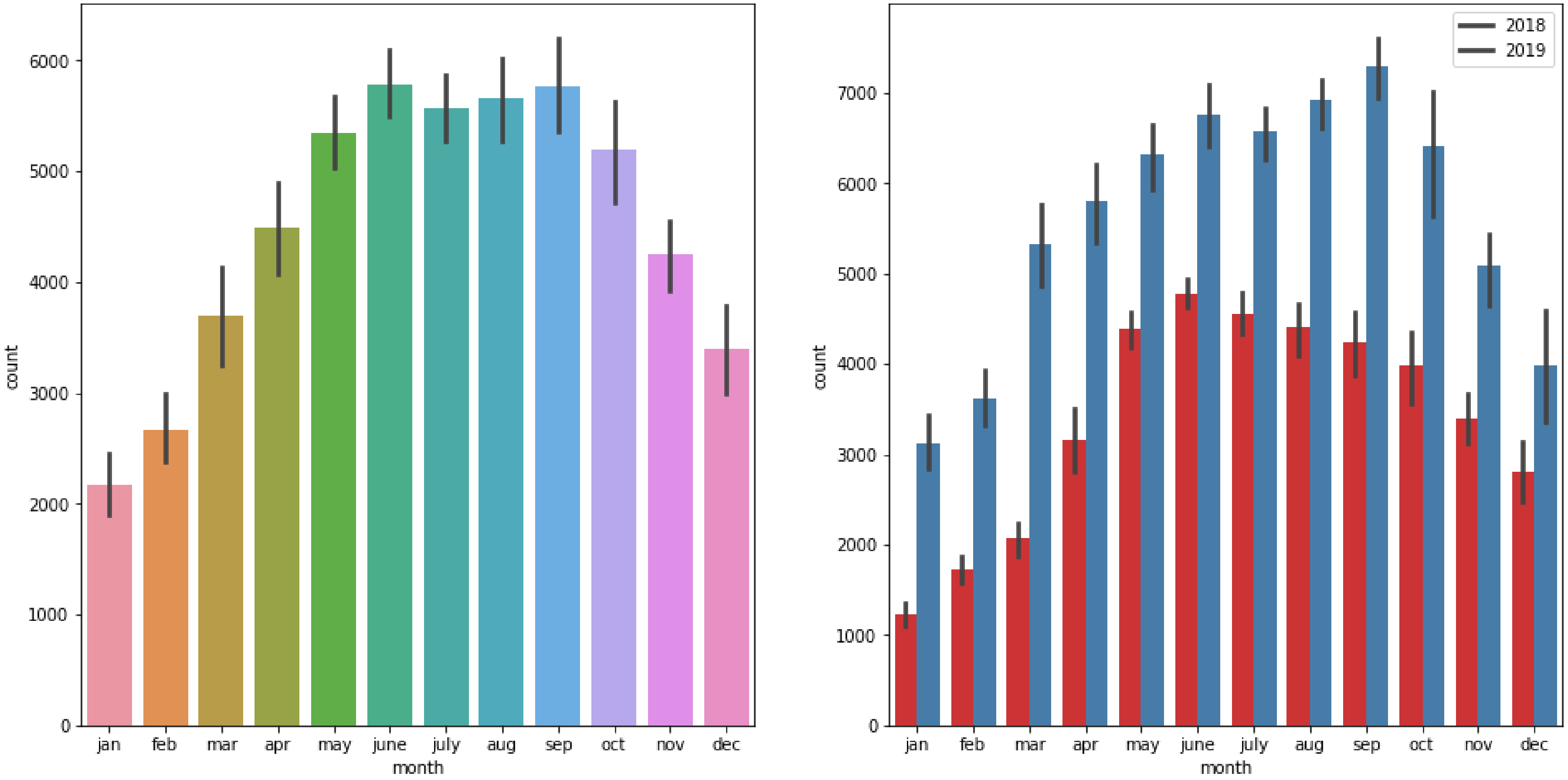
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



- 1. From the first plot we can see that the count is high in the summer and fall
- 2. It also varies a lot less when there is precipitation as people opt to not use bikes when the roads are wet
- 2. We can also see that the count has varied a lot in 2019 as compared to 2018 which shows a increasing trend for share bikes
- 3. From plot 1,2 and 4 we can assume that the bike sharing count may go higher in the months when there is no precipitation
- 4. we can also see that the count varies a lot on working days

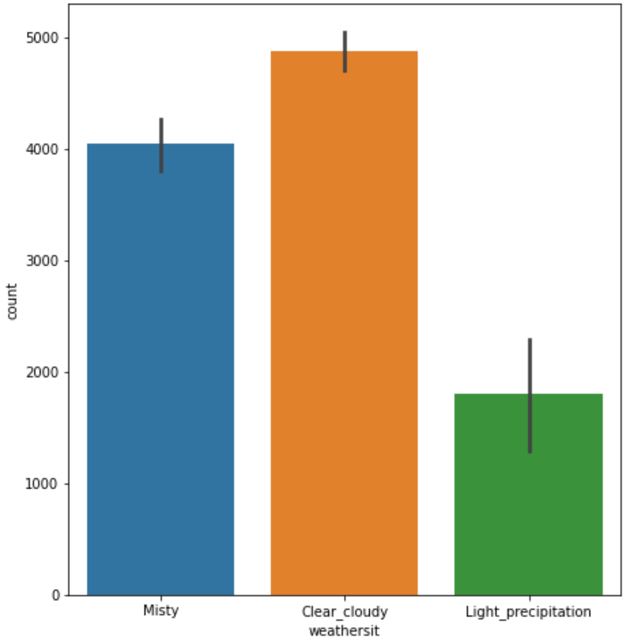

6000
5000
4000
3000
2000
1000
0
count
spring
summer
fall
winter
season

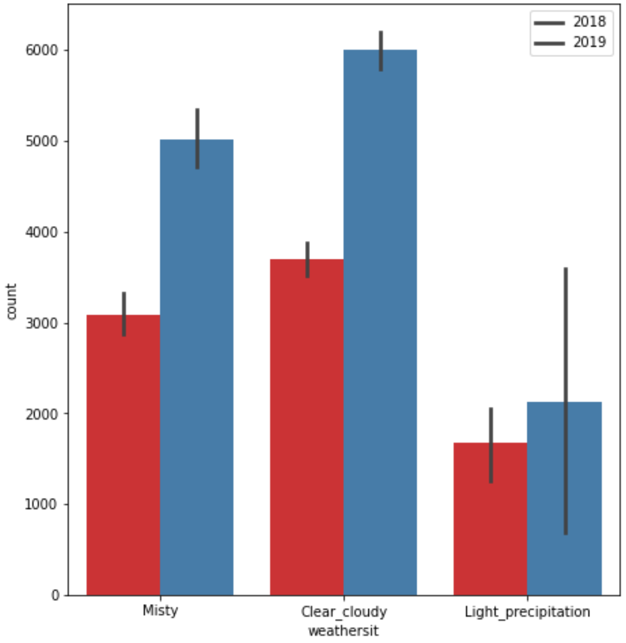

2018
2019
7000
6000
5000
4000
3000
2000
1000
0
count
spring
summer
fall
winter
season


count
0
1000
2000
3000
4000
5000
6000
jan
feb
mar
apr
may
june
july
aug
sep
oct
nov
dec
month
2018
2019
count
0
1000
2000
3000
4000
5000
6000
7000
jan
feb
mar
apr
may
june
july
aug
sep
oct
nov
dec
month

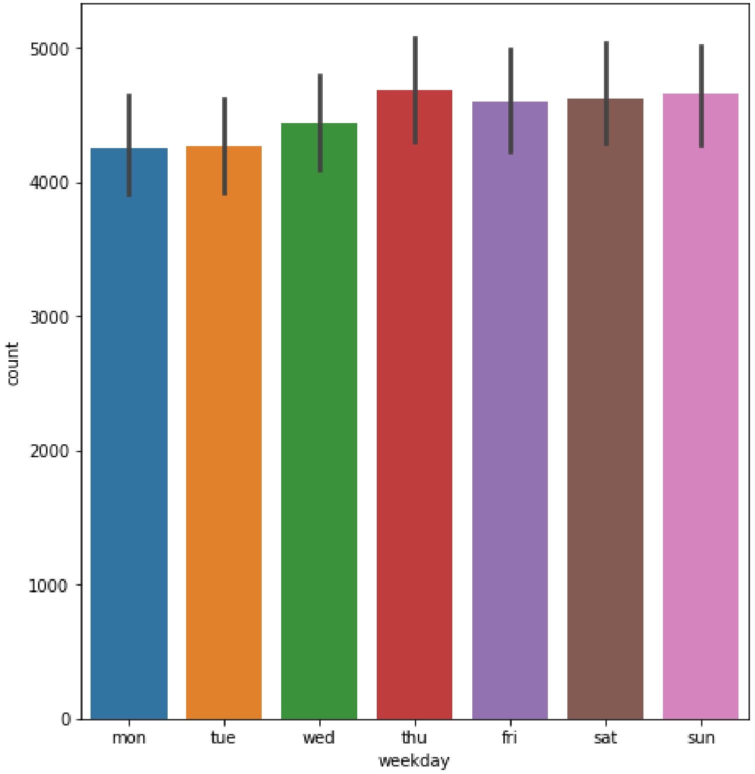
- As we infered from box plot, the most bookings for bike sharing have been done during the month of may, june, aug, sep and oct which also coincide with months of summer and fall. Hence we can conclude that most bookings happen in days when there is little or no precipitation and the bookings again go down for winter and rainy seasons as the year ends. Number of booking for each month seems to have increased from 2018 to 2019

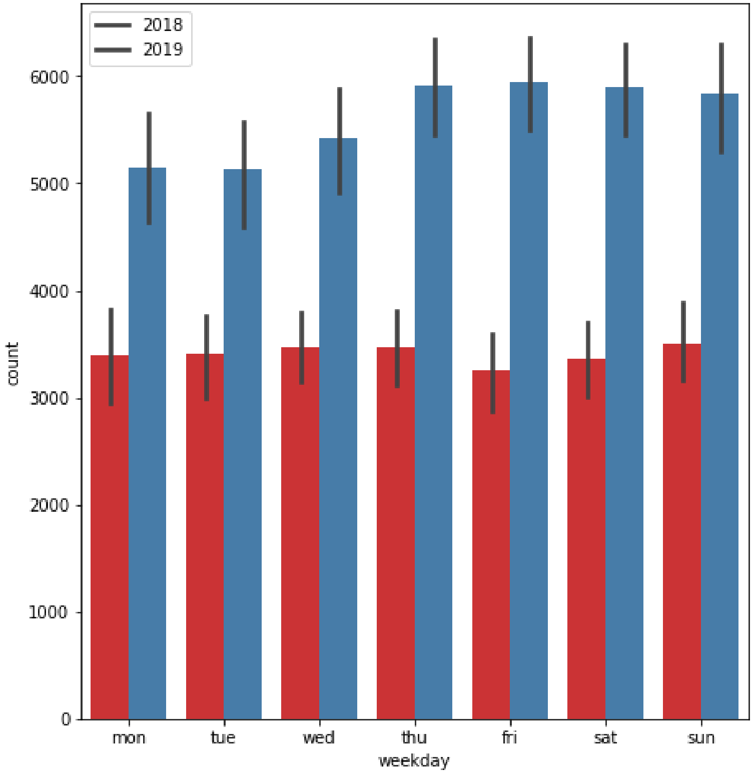
- Clear weather attracted more booking which seems obvious. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.



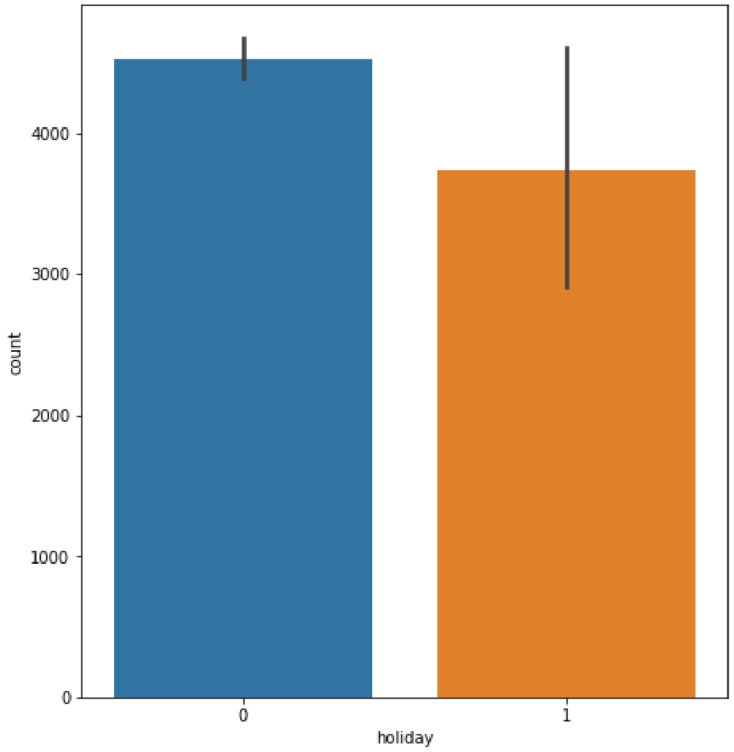


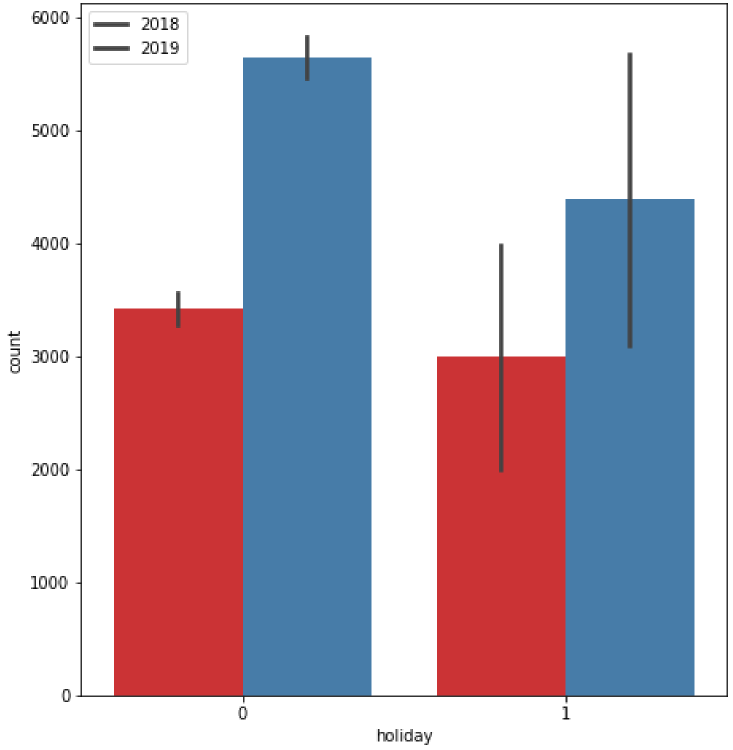
- Start of the week seems to attract lesser booking as compared to other days of the week.



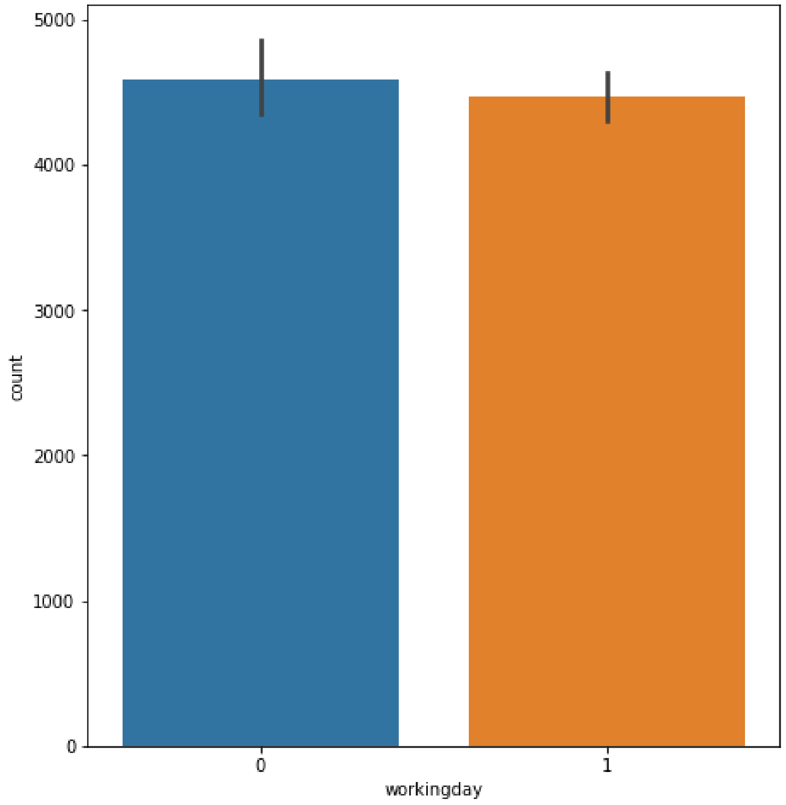


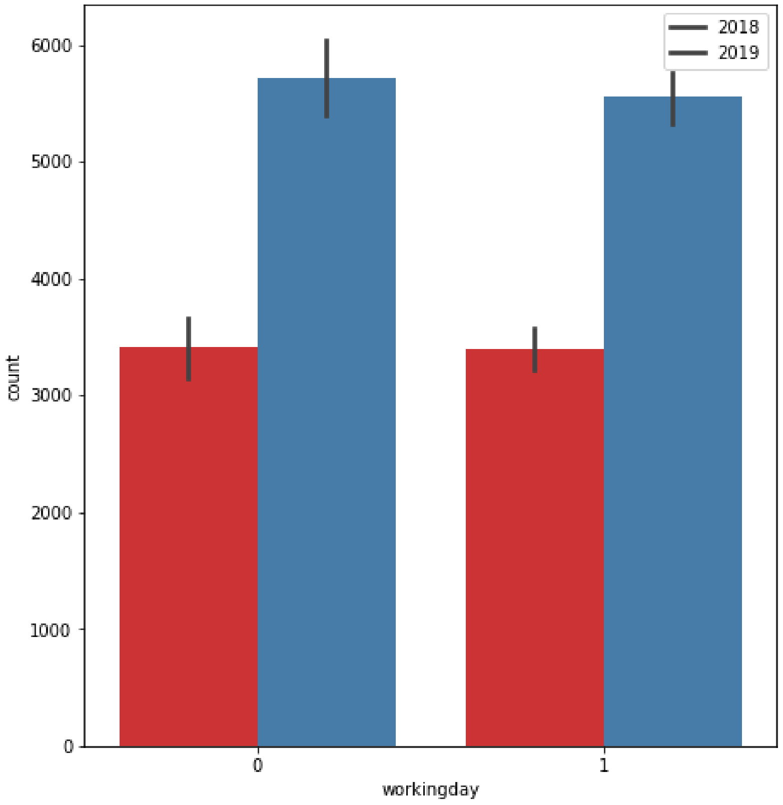
- The Bar graph below for Holiday variable shows that bookings get lower on holidays as compared to workdays as people may want to spend time at home.



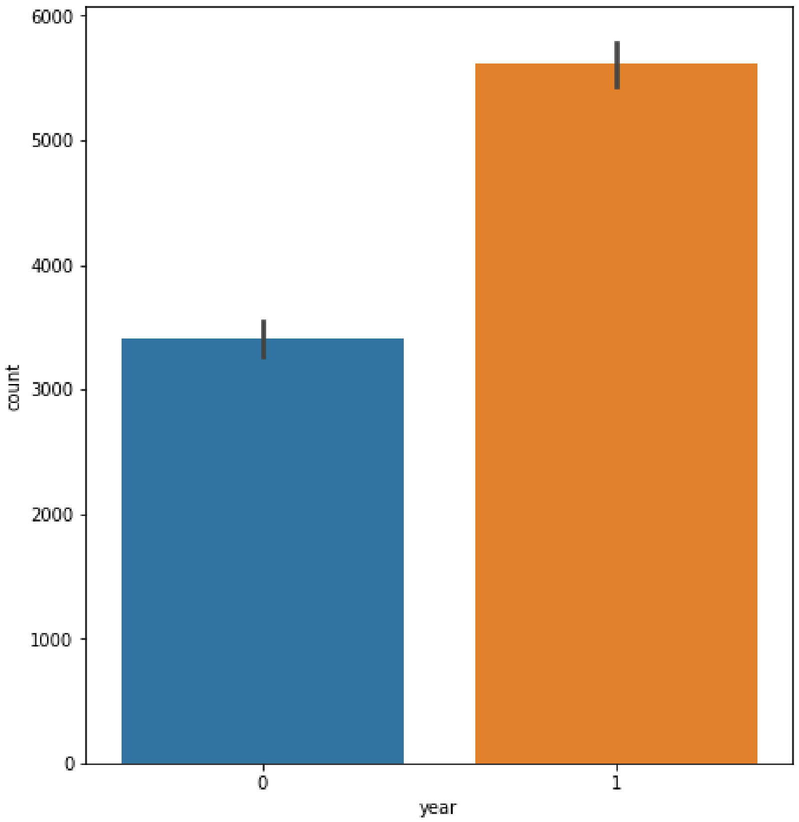


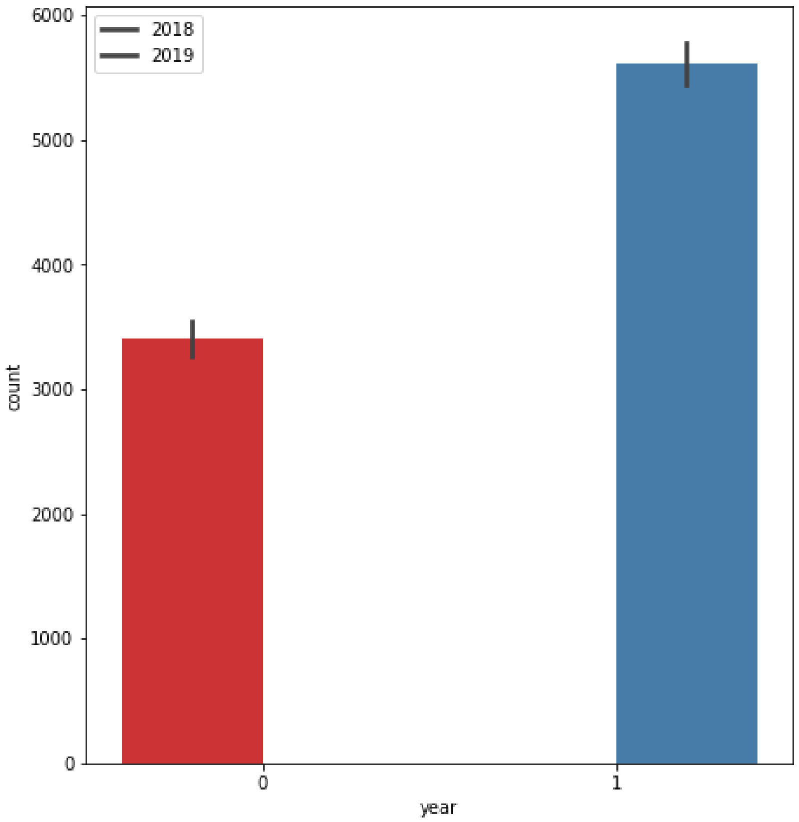
- Bookings on working and non-working days show no difference but the count increased the following year for both working and non working day.





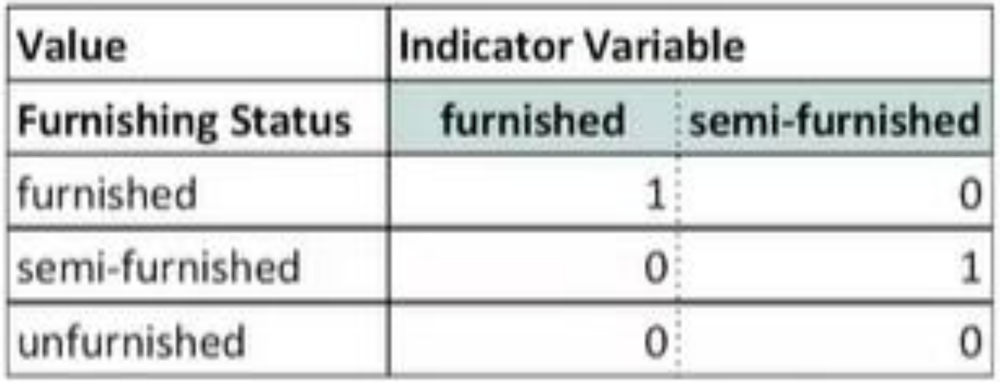
- Bookings have increased in 2019 as compared to 2018 which shows good progress in terms of business for bikesharing.





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

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

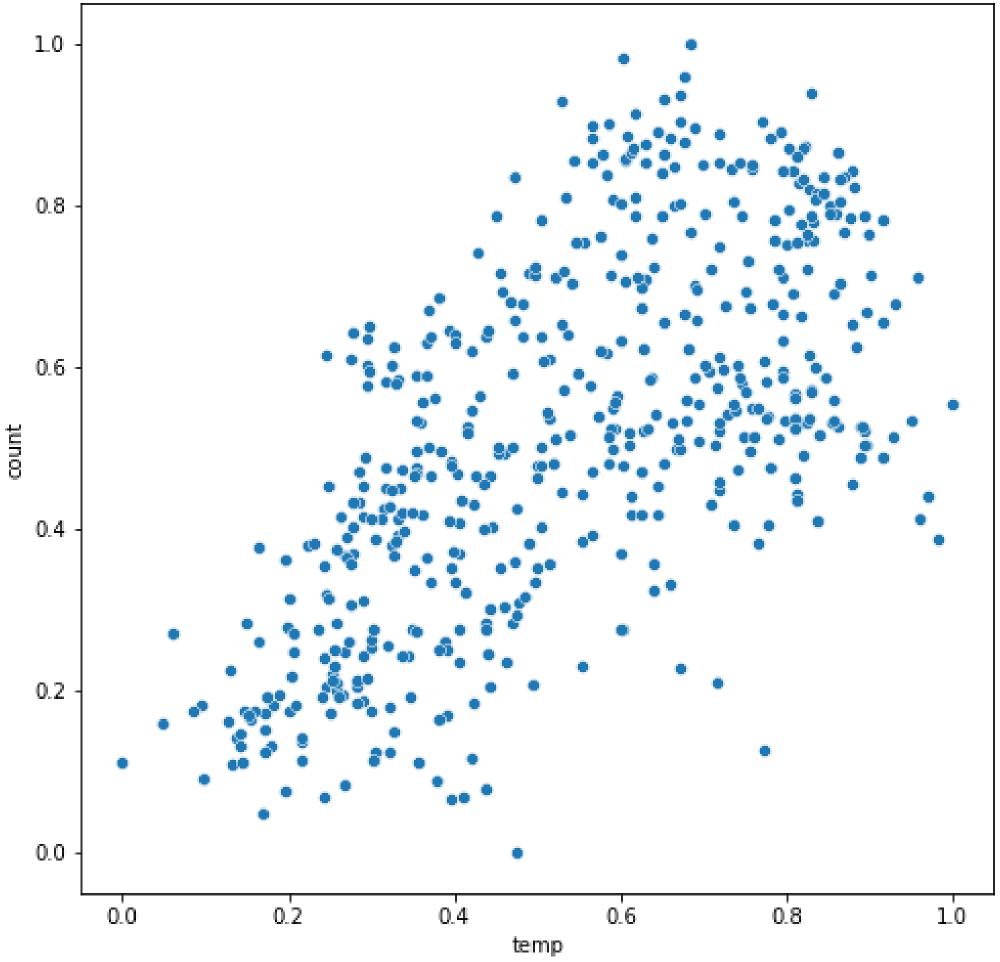


Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

- Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

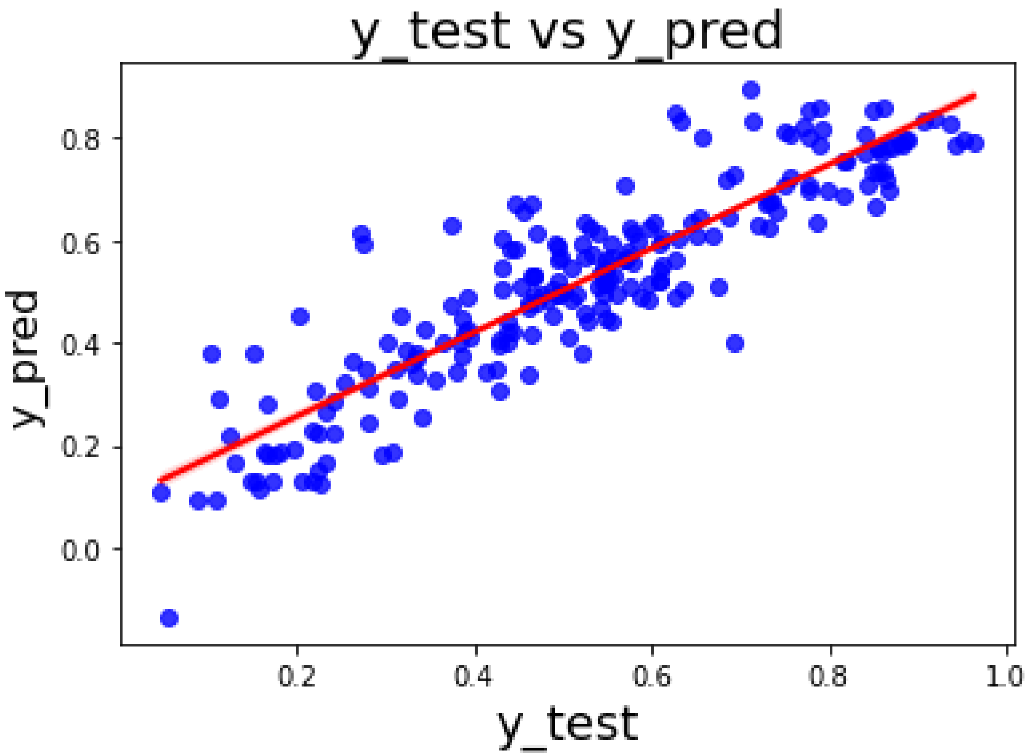
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- The plot shows high correlation between temp and count



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

- First we plotted the regression for y_test and Y_pred which showed a positive correlation.
- And the regression line was as expected as we modeled from our train data set.

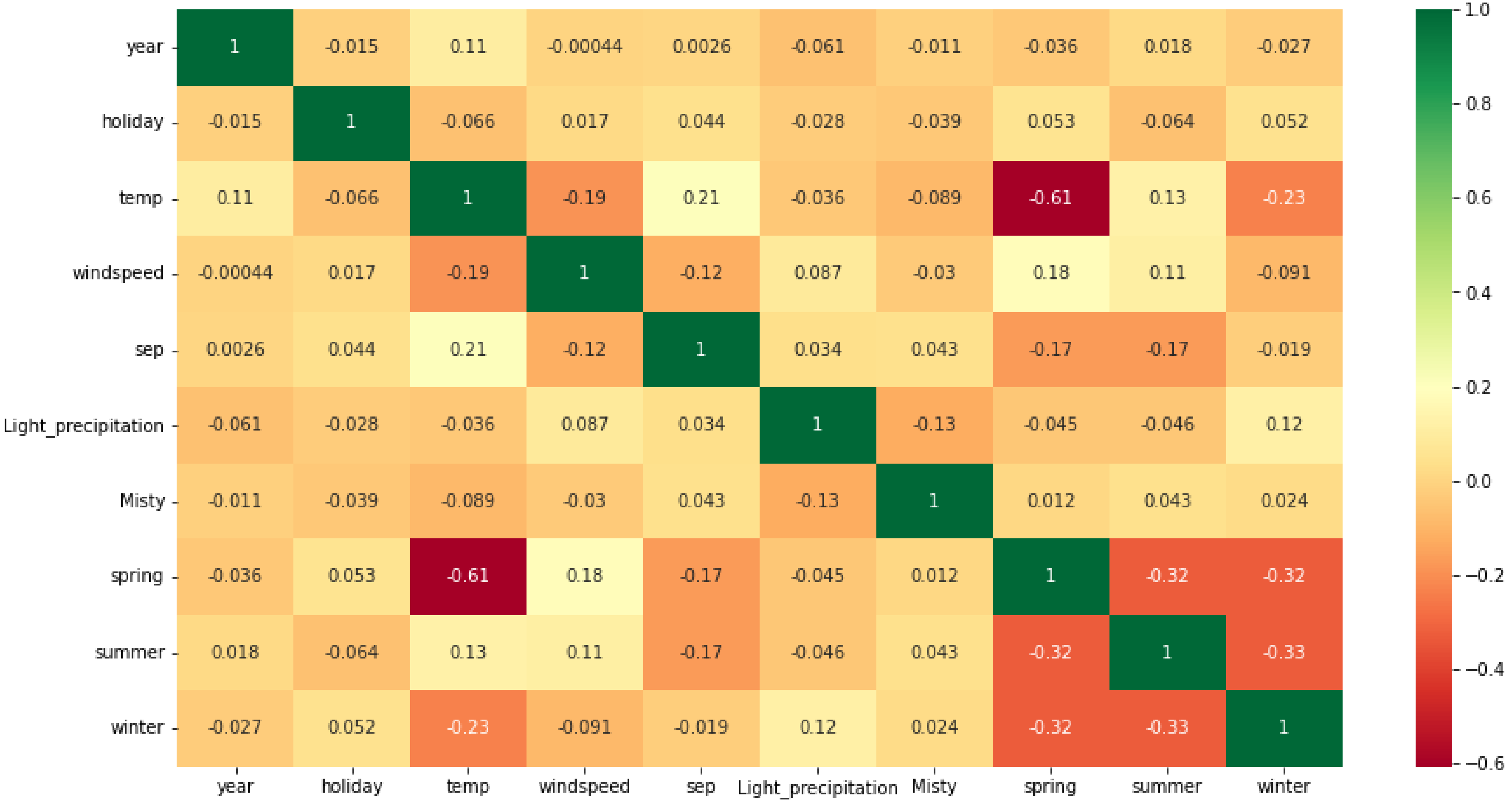


We also did comparison between training and testing data after modeling which showed the result as follows:

- Comparison between Training and Testing dataset:
- - Train dataset R^2 : 0.833
- - Test dataset R^2 : 0.8038
- - Train dataset Adjusted R^2 : 0.829
- - Test dataset Adjusted R^2 : 0.7944

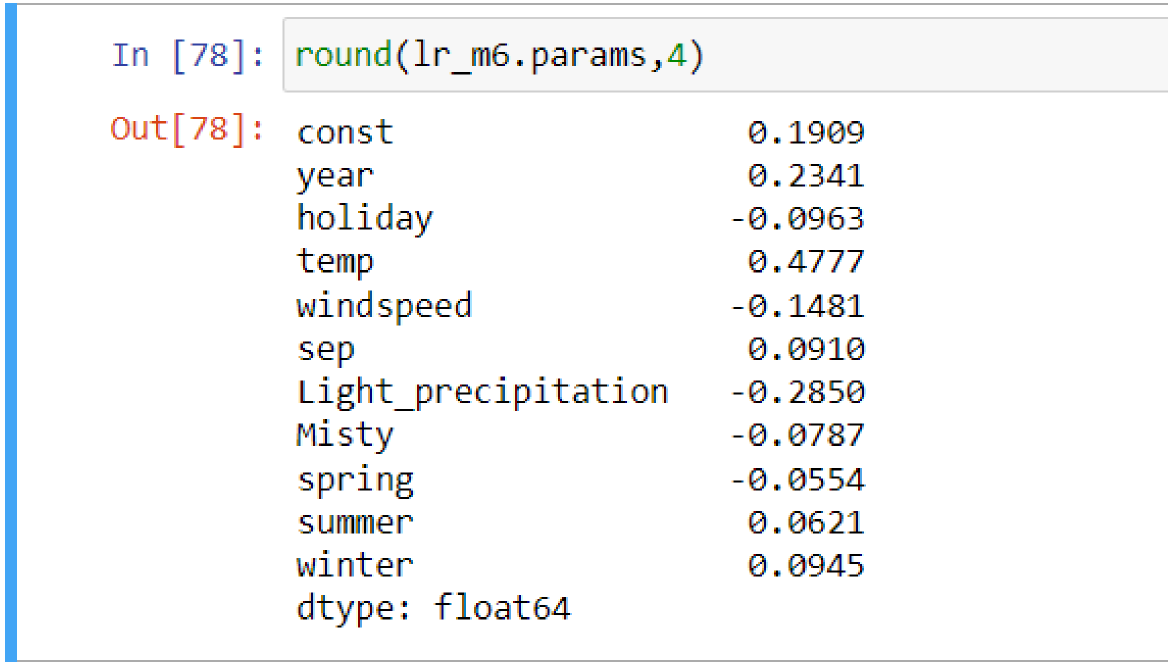
- From this we can understand that around 80% of the observed variation can be explained by the model’s inputs.

- The VIF values are all less than 5 and also there is no visible multicollinearity as seen from the heatmap



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

- The top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, year and Light_precipitation



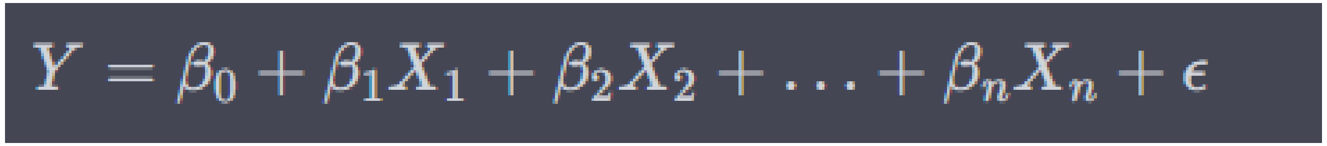
```
In [78]: round(lr_m6.params,4)

Out[78]: const                   0.1909
         year                    0.2341
         holiday                -0.0963
         temp                    0.4777
         windspeed              -0.1481
         sep                     0.0910
         Light_precipitation    -0.2850
         Misty                  -0.0787
         spring                 -0.0554
         summer                  0.0621
         winter                  0.0945
         dtype: float64
```

General Subjective Questions

Q1 - Explain the linear regression algorithm in detail.

- Linear regression is a fundamental supervised machine learning algorithm used for predicting a continuous target variable (also called the dependent variable) based on one or more input features (independent variables). It establishes a linear relationship between the input features and the target variable, allowing us to make predictions or understand the relationship between variables. Here's a detailed explanation of the linear regression algorithm:
- The Linear Regression Model
- In its simplest form, linear regression assumes that the relationship between the input features (denoted as X) and the target variable (denoted as Y) is linear and can be represented as:



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

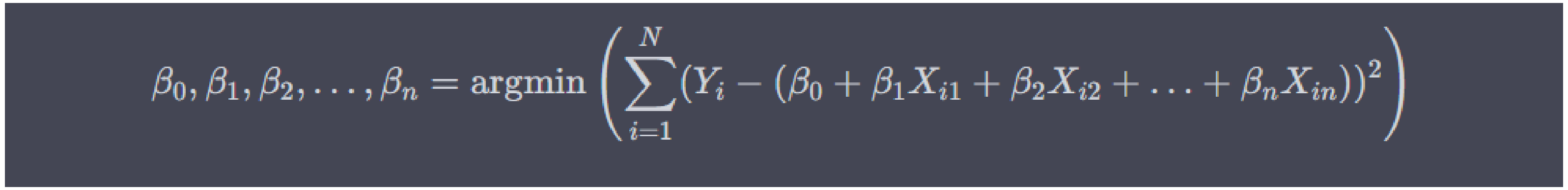
- Y is the target variable we want to predict.
- X_1,X_2,…,X_n are the input features.
- β_0 is the y-intercept or the bias term, representing the value of Y when all X values are zero.
- β_1, β_2,…, β_n are the coefficients of the respective input features, representing the change in Y for a one-unit change in the corresponding X feature while holding all other features constant.
- ϵ represents the error term, which accounts for the discrepancy between the predicted and actual values. It is assumed to be normally distributed with mean zero.

- Objective of Linear Regression
- The primary goal of linear regression is to find the values of β_o and β_1,β_2,…,β_n that minimize the error term ϵ. In other words, we want to find the best-fitting linear model that describes the relationship between the input features and the target variable.

Training the Linear Regression Model

- Training a linear regression model involves the following steps:

1. **Data Collection**: Collect a dataset that contains both the input features and the corresponding target variable.
2. **Data Preprocessing**: Clean the data by handling missing values, outliers, and scaling the features if necessary.
3. **Splitting the Data**: Divide the dataset into two parts: a training set and a test set. The training set is used to train the model, while the test set is used to evaluate its performance.
4. **Model Training**: Calculate the values of $\beta 0$ and $\beta 1, \beta 2, \ldots, \beta n$ using a technique called "least squares estimation." This technique minimizes the sum of squared errors between the predicted values and the actual values in the training data.



$$\beta_0, \beta_1, \beta_2, \ldots, \beta_n = \text{argmin}\left(\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}))^2\right)$$

1. Where *N* is the number of data points in the training set.
2. **Model Evaluation**: After training, evaluate the model's performance using various metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared (*R*^2) on the test data. These metrics help assess how well the model generalizes to unseen data.
3. **Prediction**: Once the model is trained and evaluated, it can be used to make predictions on new or unseen data by plugging in the values of the input features.

Assumptions of Linear Regression

- Linear regression makes several assumptions about the data:

1. Linearity: It assumes a linear relationship between the input features and the target variable.
2. Independence of Errors: The errors (ϵ) should be independent of each other and have constant variance (homoscedasticity).
3. Normality of Errors: The errors should follow a normal distribution.
4. No or Little Multicollinearity: The input features should not be highly correlated with each other.
5. No Autocorrelation: The errors should not be correlated with each other in a time series context.

If these assumptions are violated, the model's predictions may not be accurate, and additional techniques or adjustments may be required.

Advantages and disadvantages

Advantages

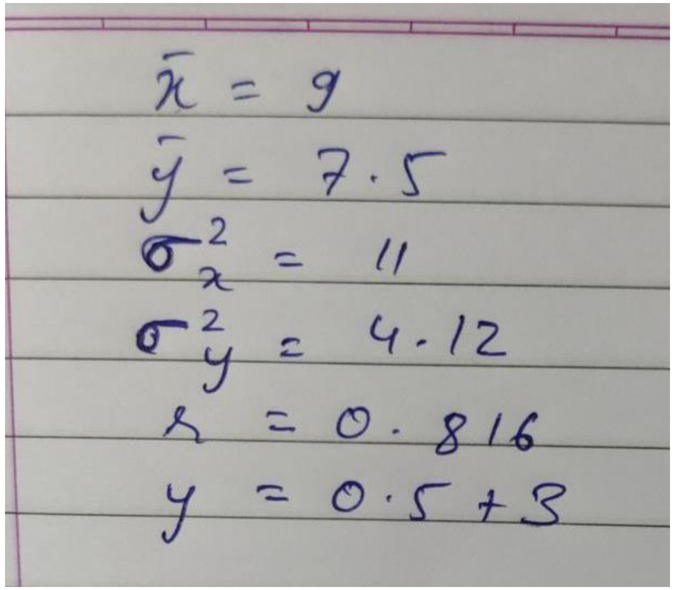
- Simplicity and interpretability.
- Well-suited for understanding relationships between variables.
- Good for making predictions when the linear relationship assumption holds.

Disadvantages

- Assumes a linear relationship, which may not always be the case.
- Sensitive to outliers.
- May not perform well if the assumptions are violated.
- Limited in handling complex, nonlinear relationships.

2. Explain the Anscombe’s quartet in detail.

- Consider the adjacent dataset. The summary statistics for that is as follows :



$\bar{x} = 9$

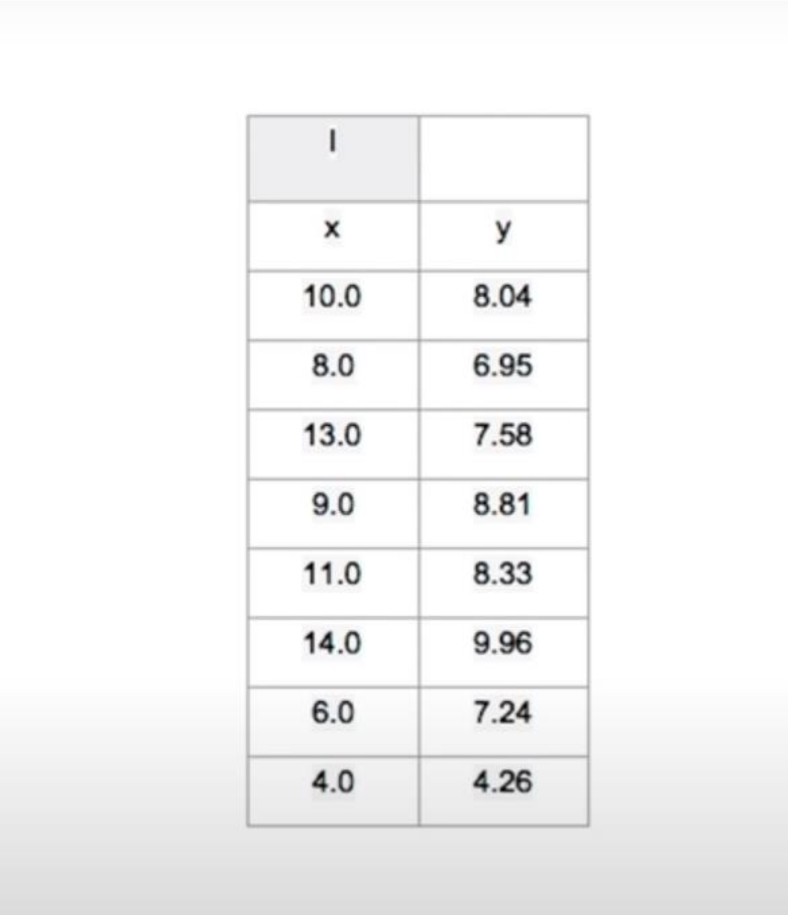
$\bar{y} = 7.5$

$\sigma_x^2 = 11$

$\sigma_y^2 = 4.12$

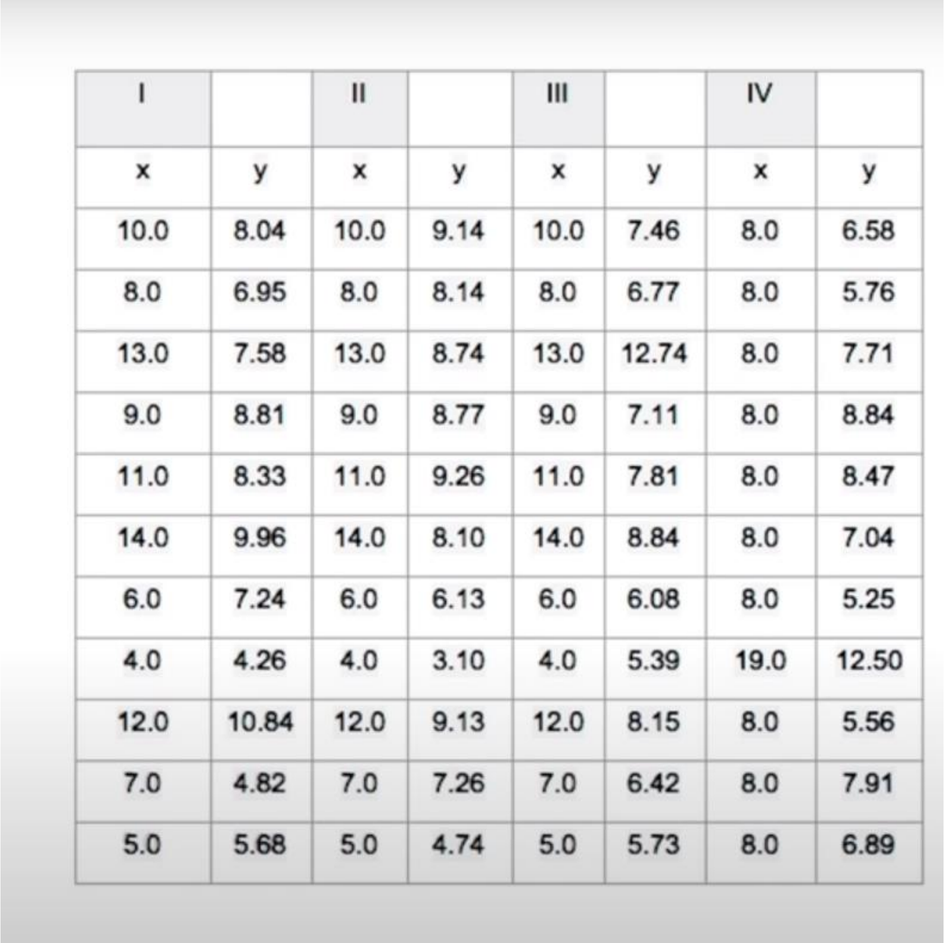
$r = 0.816$

$y = 0.5 + 3$



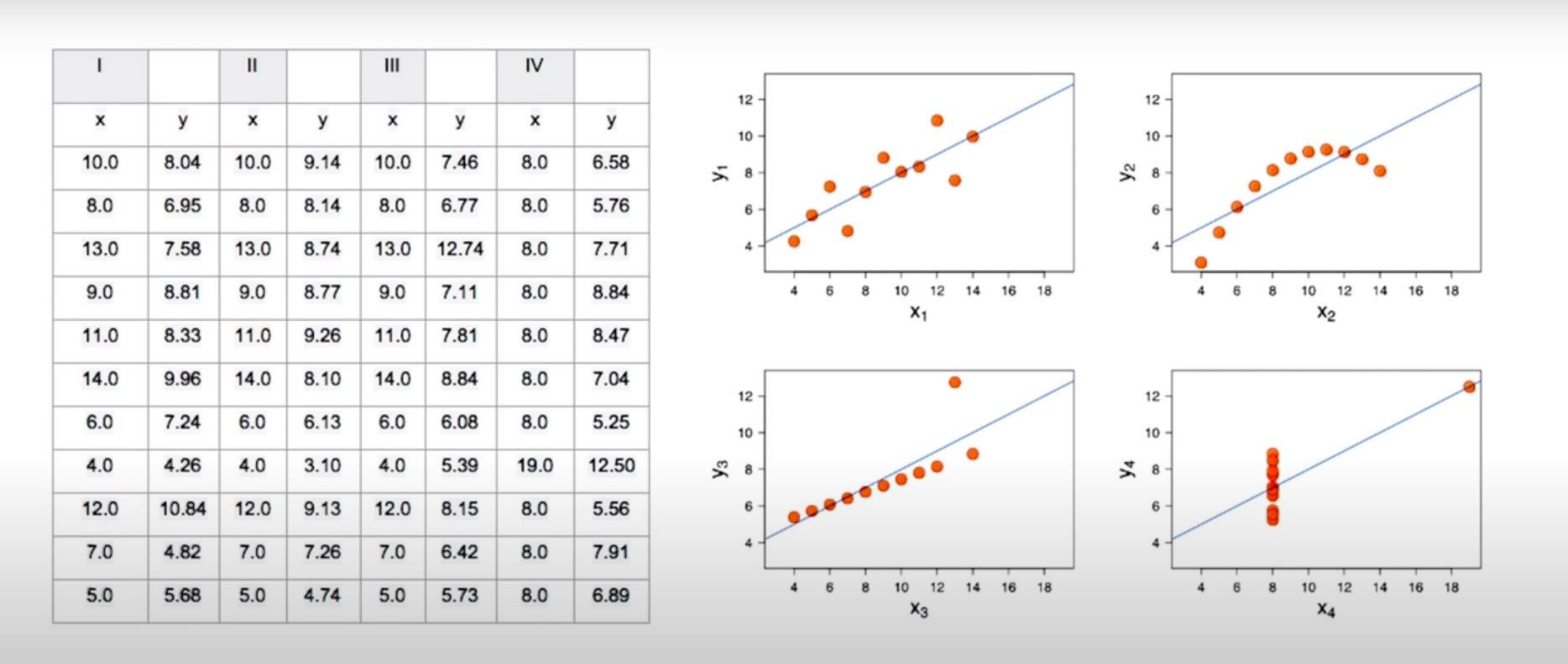
I	
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26

- Now we will take 3 more datasets which have the same summary statistics.
- Now since all of them have similar mean, standard deviation and R square, you would expect them to have similar plot when visualised.



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

The result of plotting the above datasets is as follows :



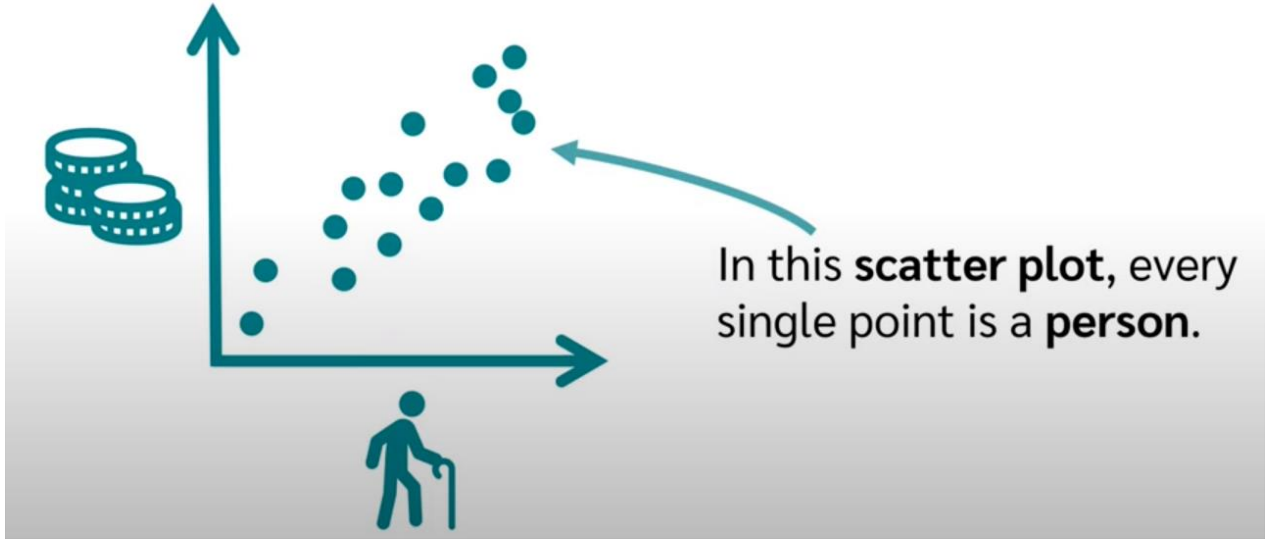
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

- The above plots show that the datasets are not similar. We can see the effect of curvature as well as outliers even though the summary statistics is the same.
- This is called as Anscombe’s Quartet and demonstrates how important it is to always plot your data rather than relying on summary statistics alone.

3. What is Pearson's R?

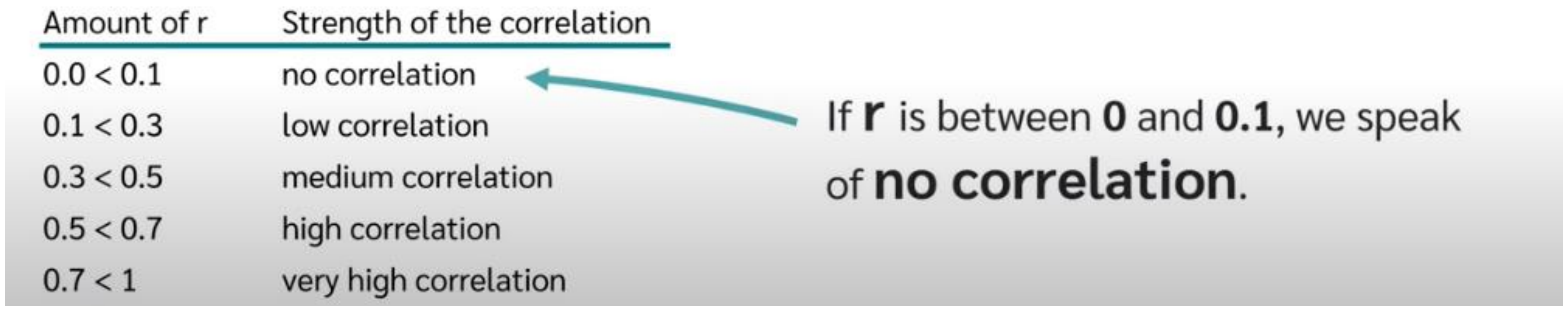
- The Pearson correlation analyses the relationship between the two variables.
- For example in this plot is there

a relationship between a person's Salary and age. If the relationship is confirmed in this example then Salary can be predicted using the age. But there must be a clear causal Relationship for this. Just because there Is a correlation you can't tell which way the relationship is going. So with the help of Pearson Correlation we can measure the linear relationship between the two variables.



In this **scatter plot**, every single point is a **person.**

- We can determine how strong the correlation is and in which direction the correlation goes. We can read both in the Pearson correlation coefficient r which is between -1 and 1 the strength of the correlation can be read in a table.
- If r is between 0.7 and 1 it is a very strong correlation

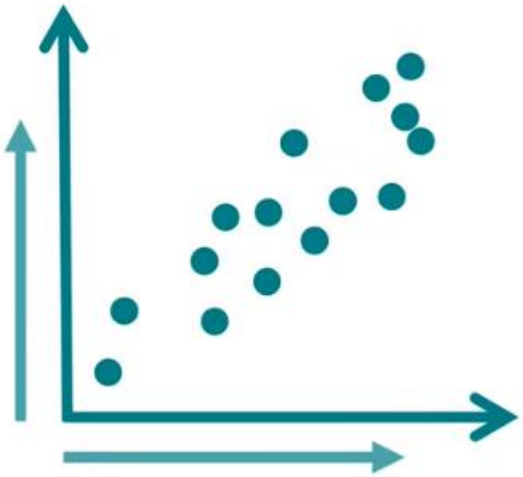


Amount of r	Strength of the correlation
0.0 < 0.1	no correlation
0.1 < 0.3	low correlation
0.3 < 0.5	medium correlation
0.5 < 0.7	high correlation
0.7 < 1	very high correlation

If **r** is between **0** and **0.1**, we speak of **no correlation**.

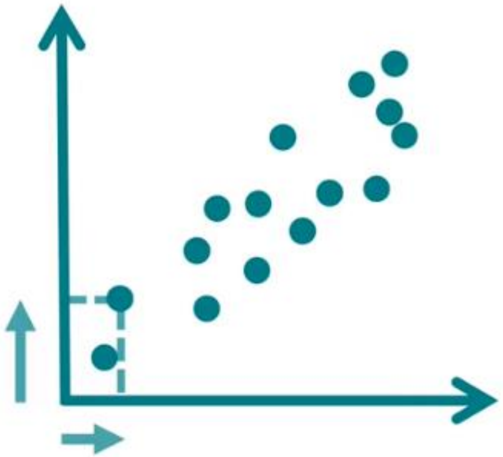
A **positive correlation** exists

when **large values** of **one variable**



go along with **large values** of the **other variable.**

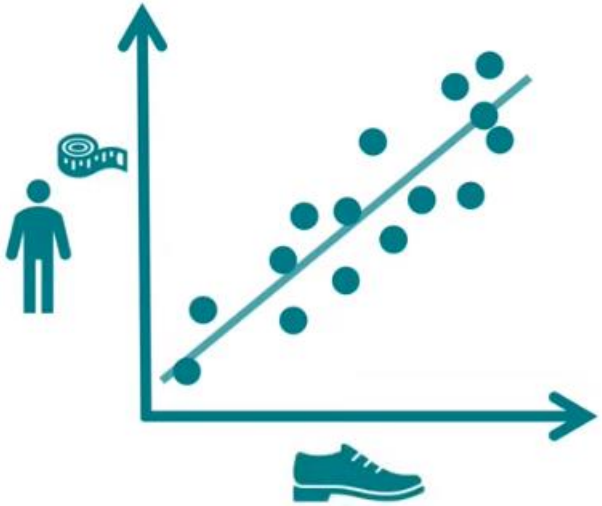
or when **small values** of **one variable**



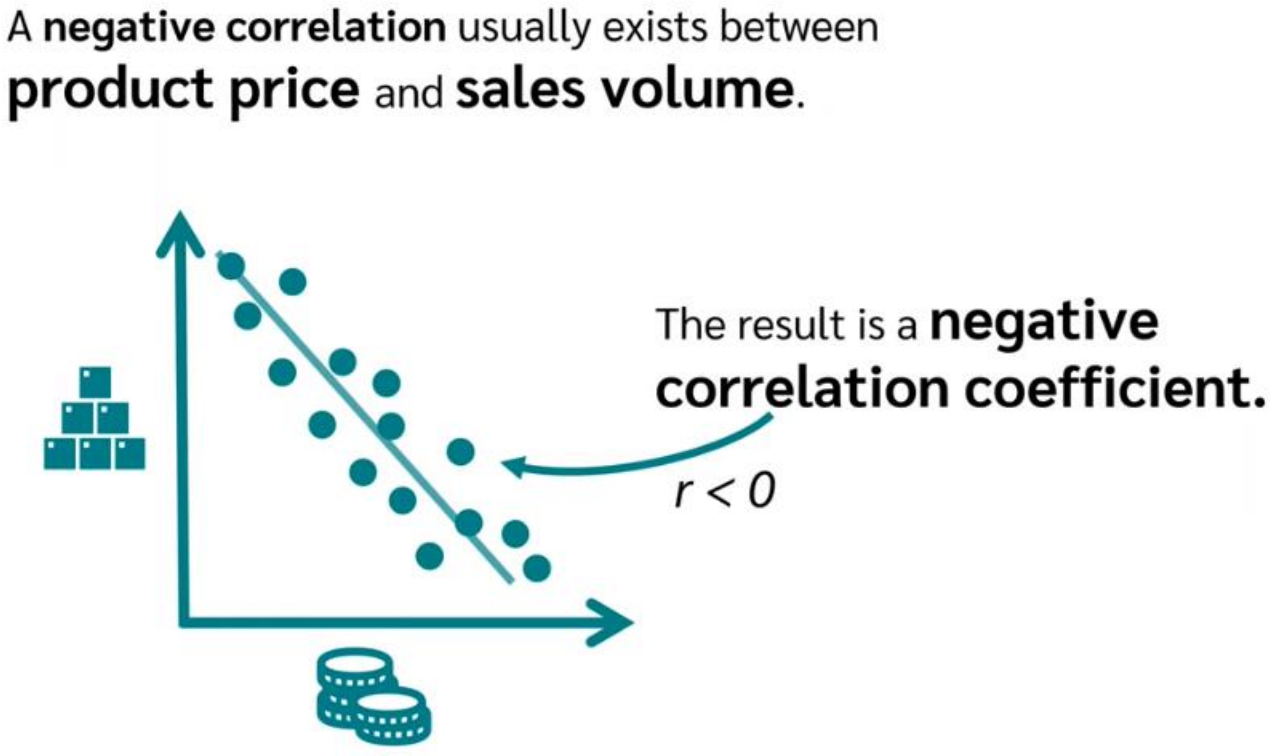
go along with **small values** of the **other variable.**

- For example body size and shoe size

A **positive correlation** is found, for example, for **body size** and **shoe size**.



- A negative correlation exists when large values of one variable go along with small values of other variable and vice versa:

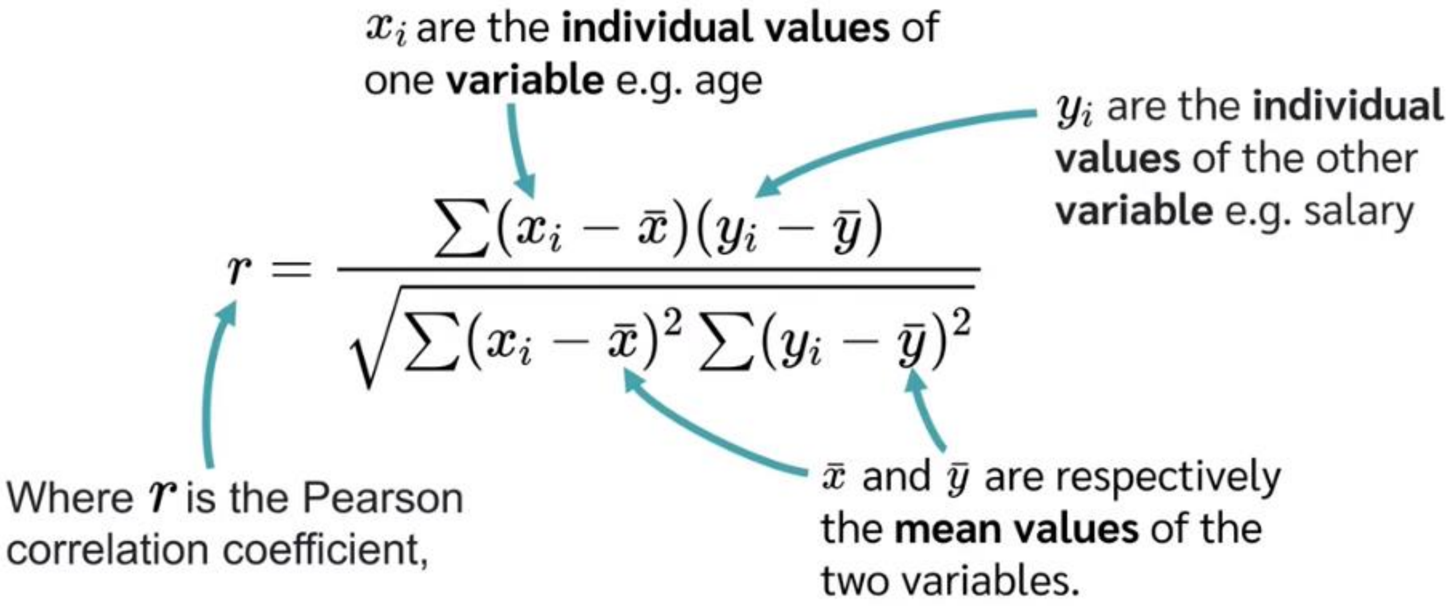


A **negative correlation** usually exists between **product price** and **sales volume**.

The result is a **negative correlation coefficient.**

$r < 0$

- The Pearson correlation is obtained by :



$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

x_i are the **individual values** of one **variable** e.g. age

y_i are the **individual values** of the other **variable** e.g. salary

Where r is the Pearson correlation coefficient,

$\bar{x}$ and $\bar{y}$ are respectively the **mean values** of the two variables.

- So the equation for our Age-Salary data will be :

$$r = \frac{\sum(Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum(Age_i - \overline{Age})^2 \cdot \sum(salary_i - \overline{salary})^2}}$$

- The expression in the denominator ensures the end result is between -1 and 1

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

- Scaling is a data preprocessing technique used in machine learning and statistics to transform the features or variables of a dataset to a similar scale or range. The primary goal of scaling is to make it easier for machine learning algorithms to work with the data, as many algorithms are sensitive to the magnitude of the input features. Scaling is particularly important when features in a dataset have different units, ranges, or variances.

Scaling is performed for several reasons:

1. Equal Weightage: Scaling ensures that all features contribute equally to the analysis, preventing features with larger scales from dominating the learning process. This is especially important for distance-based algorithms like k-means clustering or support vector machines.
2. Faster Convergence: Scaling can help iterative optimization algorithms converge more quickly because it reduces the oscillations and instability that may occur when features are on vastly different scales.
3. Improved Interpretability: Scaling makes it easier to interpret the coefficients or feature importance of models like linear regression or decision trees, as the coefficients will be in the same units as the original features.
4. Avoidance of Numerical Instabilities: Some mathematical algorithms are sensitive to large values, which can lead to numerical instability. Scaling can mitigate this issue.

There are two common methods for scaling data: normalized scaling and standardized scaling, and they differ in how they transform the data:

- Normalized Scaling (Min-Max Scaling):
    - In normalized scaling, also known as Min-Max scaling, the data is scaled to a specific range, typically between 0 and 1. The formula to normalize a feature 'x' is:
    - x_normalized = (x - min(x)) / (max(x) - min(x))
    - Normalized scaling is useful when you want to bound your features within a specific range and maintain the relationships between the data points.
- Standardized Scaling (Z-score Standardization):
    - Standardized scaling, also known as Z-score standardization, transforms the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardizing a feature 'x' is:
    - x_standardized = (x - mean(x)) / std(x)
    - Standardized scaling is useful when you want to center the data around zero and give it a unit variance. This approach assumes that the data follows a normal distribution or is at least approximately normally distributed.

- In summary, both normalized scaling and standardized scaling are techniques used to bring the features of a dataset to a common scale, but they do so in different ways. The choice between them depends on the specific requirements of your analysis and the characteristics of your data. Normalized scaling is preferred when you want to maintain the original range of the data, while standardized scaling is useful for making data more suitable for algorithms that assume a standard normal distribution.

Q5 - You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- In the context of VIF, a VIF value of infinity (or sometimes referred to as "infinite VIF") occurs when there is perfect multicollinearity in the regression model. Perfect multicollinearity means that one or more independent variables in the model can be exactly predicted from the other independent variables, leading to a situation where the correlation is so strong that it's mathematically impossible to separate the individual effects of these variables.
- Here are some common reasons why a VIF might be infinite:

1. **Redundant Variables:** One or more variables in the regression model are linear combinations of other variables in the model. For example, if you include the same variable twice in the model or if you have variables that are exact duplicates of each other, VIF will be infinite because one variable can be perfectly predicted from the other.
2. **Perfectly Predictable Outcome:** In some cases, you might have a situation where the dependent variable (the outcome) can be perfectly predicted from the independent variables. This would lead to infinite VIF values because the independent variables are perfectly multicollinear with the dependent variable.
3. **Data Issues:** Data errors or anomalies can also lead to infinite VIF values. For instance, if there are outliers or extreme values in the data, they can disrupt the calculations used to compute VIF, causing unexpected results.

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 graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. The Q-Q plot is a scatterplot that compares the quantiles (percentiles) of the observed data against the quantiles of the theoretical distribution. Each point on the Q-Q plot represents a data point from the sample and its corresponding expected value from the theoretical distribution.

- Here's how to interpret a Q-Q plot:

1. **Perfectly Straight Line (Diagonal Line):** If the points on the Q-Q plot fall along or very close to a diagonal line (the line y = x), it indicates that the observed data closely follows the theoretical distribution being tested (e.g., normal distribution). In other words, the data is approximately normally distributed if you are comparing it to a normal distribution.
2. **Deviation from the Line:** If the points deviate from the diagonal line, it suggests that the data does not conform to the theoretical distribution. The direction and degree of deviation can provide insights into how the data differs from the assumed distribution.

- The use and importance of a Q-Q plot in linear regression are as follows:

1. **Assumption Checking:** In linear regression, it is often assumed that the residuals (the differences between the observed values and the predicted values) are normally distributed with a mean of zero. Checking this assumption is crucial for the validity of regression analysis. A Q-Q plot of the residuals is a valuable tool to assess the normality assumption. If the Q-Q plot of the residuals closely follows a straight line, it indicates that the assumption of normality is reasonable.
2. **Identification of Outliers:** Q-Q plots can help identify outliers or extreme values in the dataset. Investigating these outliers can be important for understanding their impact on the regression model and deciding whether they should be addressed.
3. **Model Diagnostics:** Q-Q plots are part of a set of diagnostic plots used to evaluate the overall goodness of fit of a regression model.
4. **Model Improvement:** If the Q-Q plot reveals a substantial departure from the expected distribution, it may suggest that a different regression model or data transformation is needed to better capture the underlying patterns in the data.