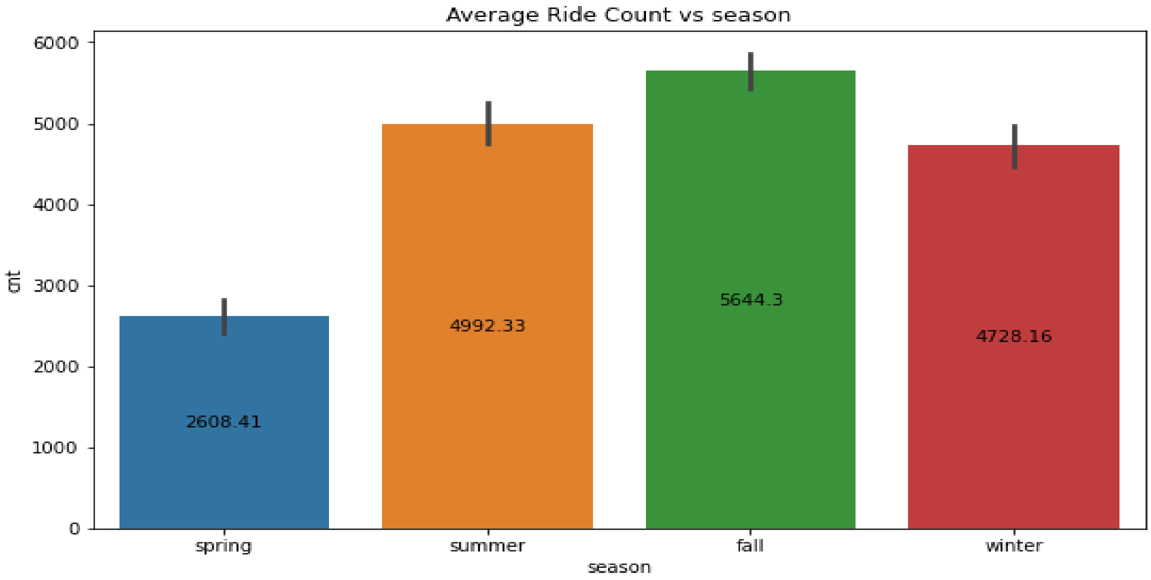
# Assignment-based Subjective Questions

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

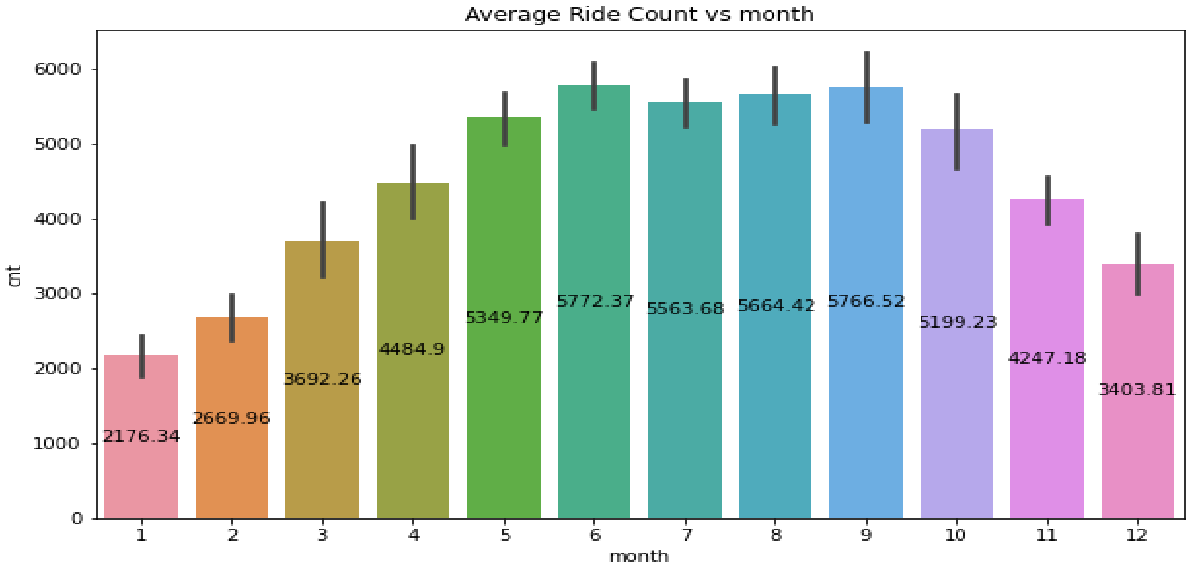
**Answer:** I have done analysis on categorical variables using boxplot and bar plot.

Below are the few points we can infer from the visualizations: -

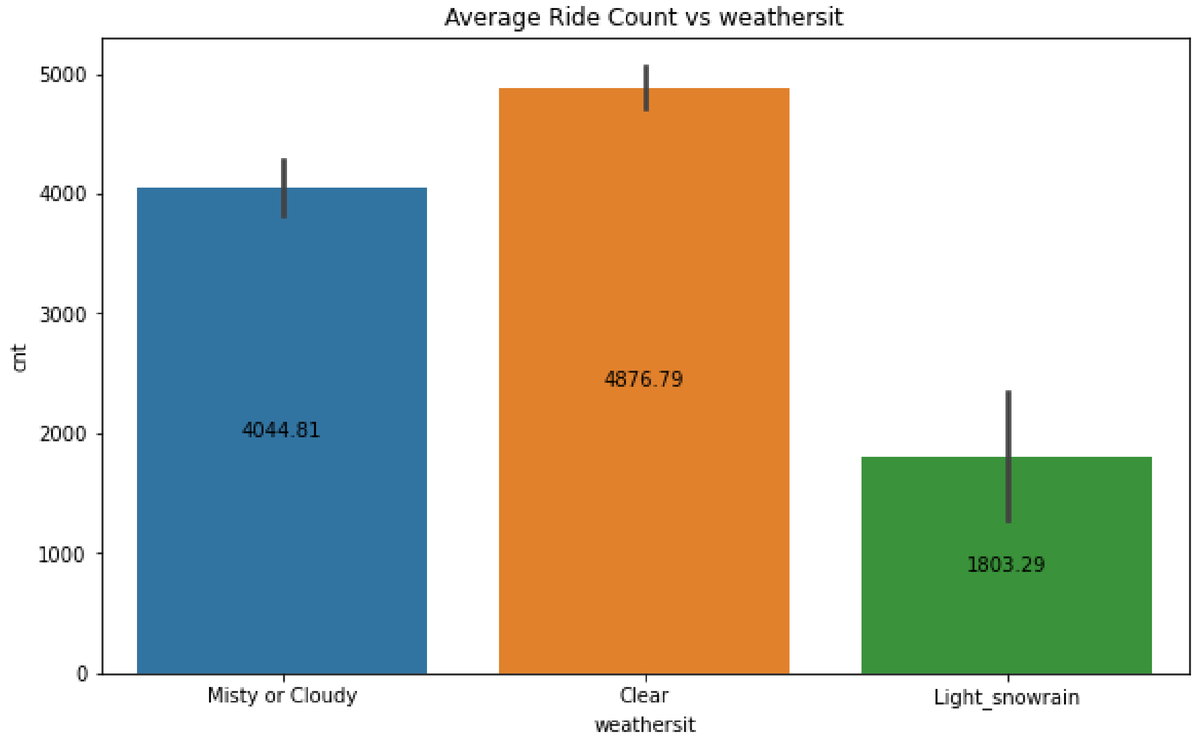
1.Ride Count significant increase in fall and summer



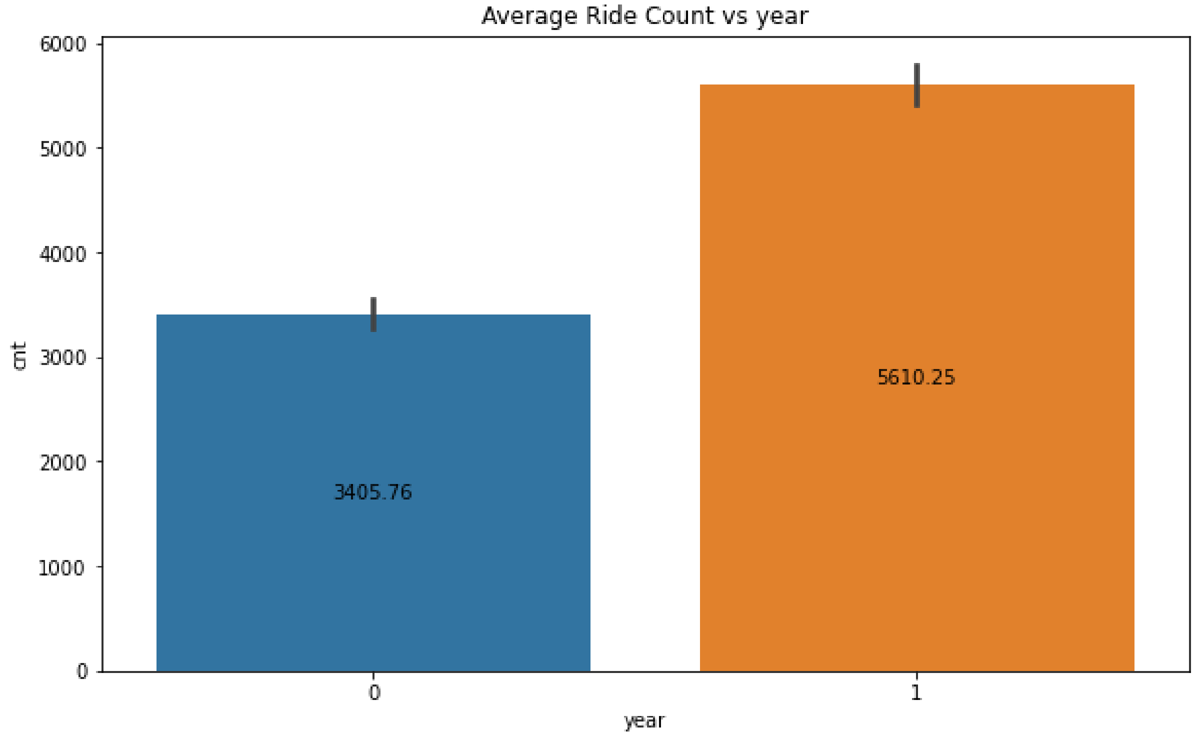
2.Ride seems to be increased from march to October and decreased at the end of the year.



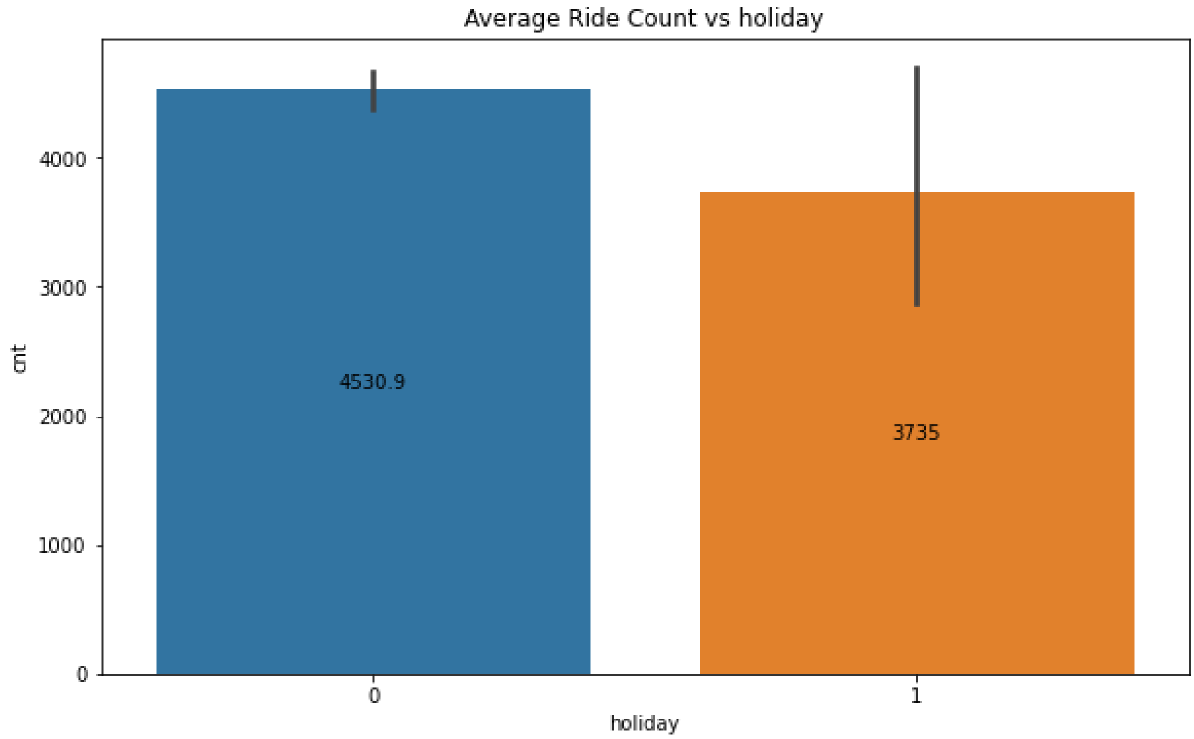
3.Ride count increase in clear weather condition.



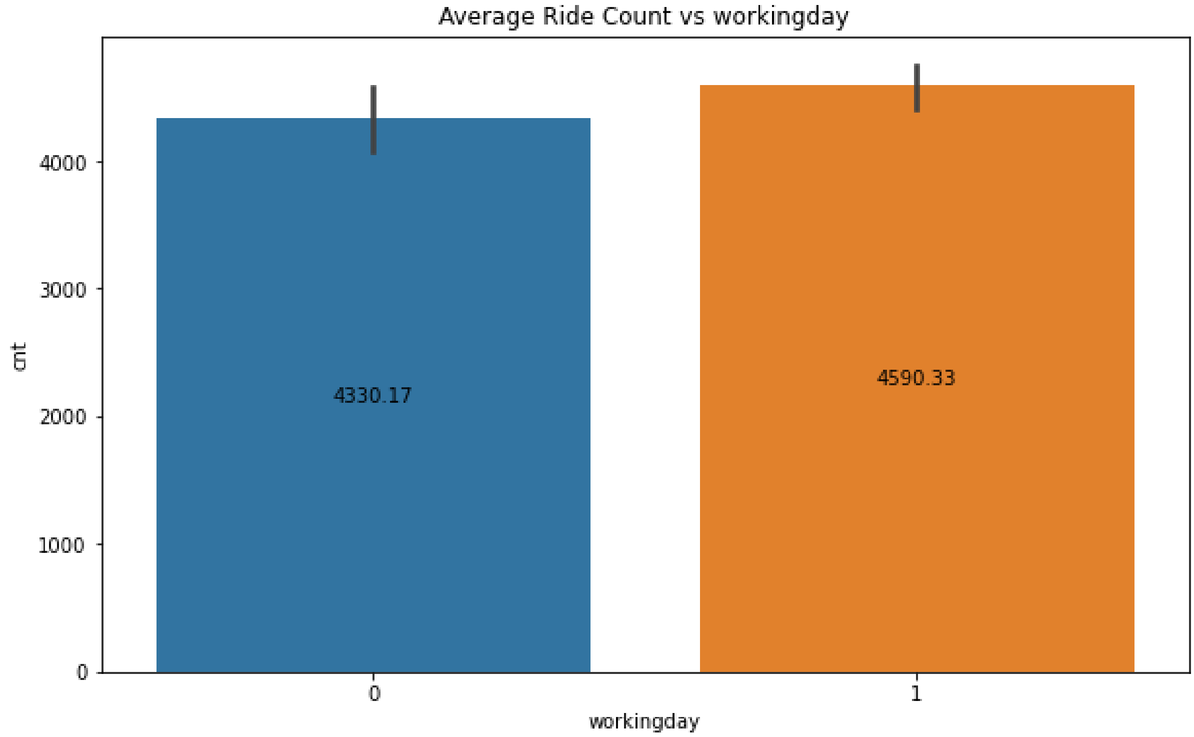
4.Ride count increased drastically from 2018 to 2019.



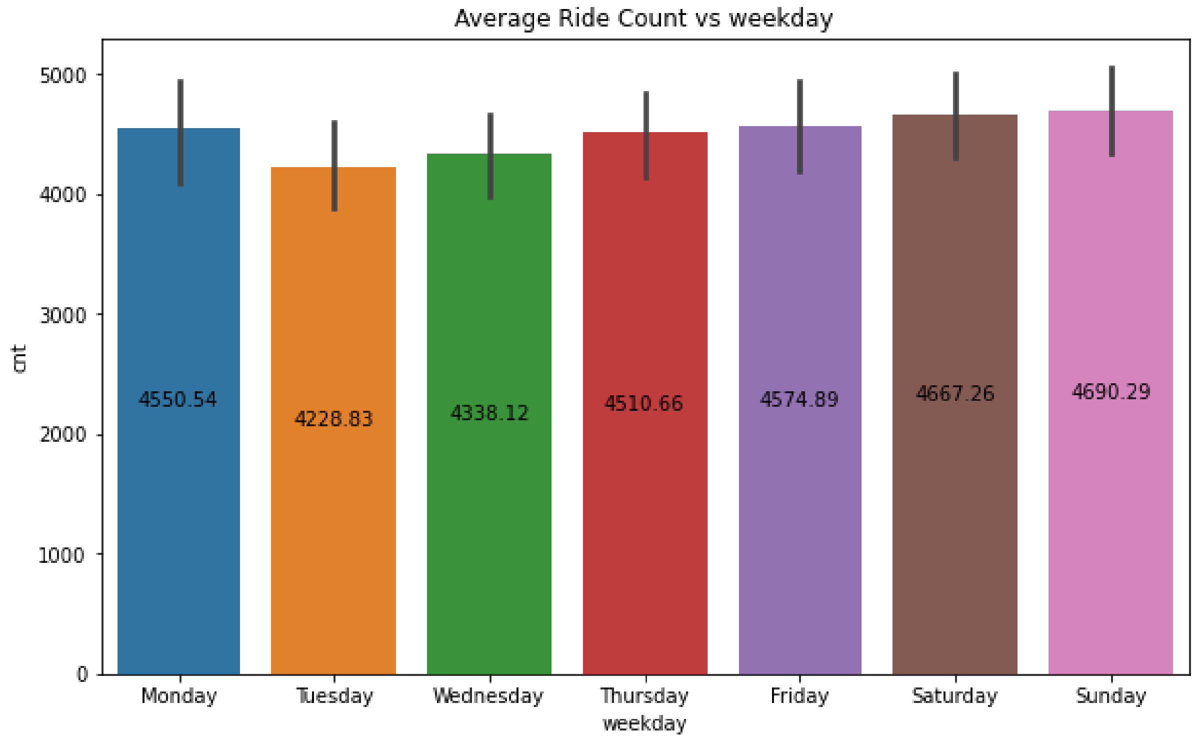
5.Ride Count reduced during holiday.



6.Booking seems to be almost equal either on working day or non-working day.



7.Booking is less on Tuesday and Wednesday as compared to other days.

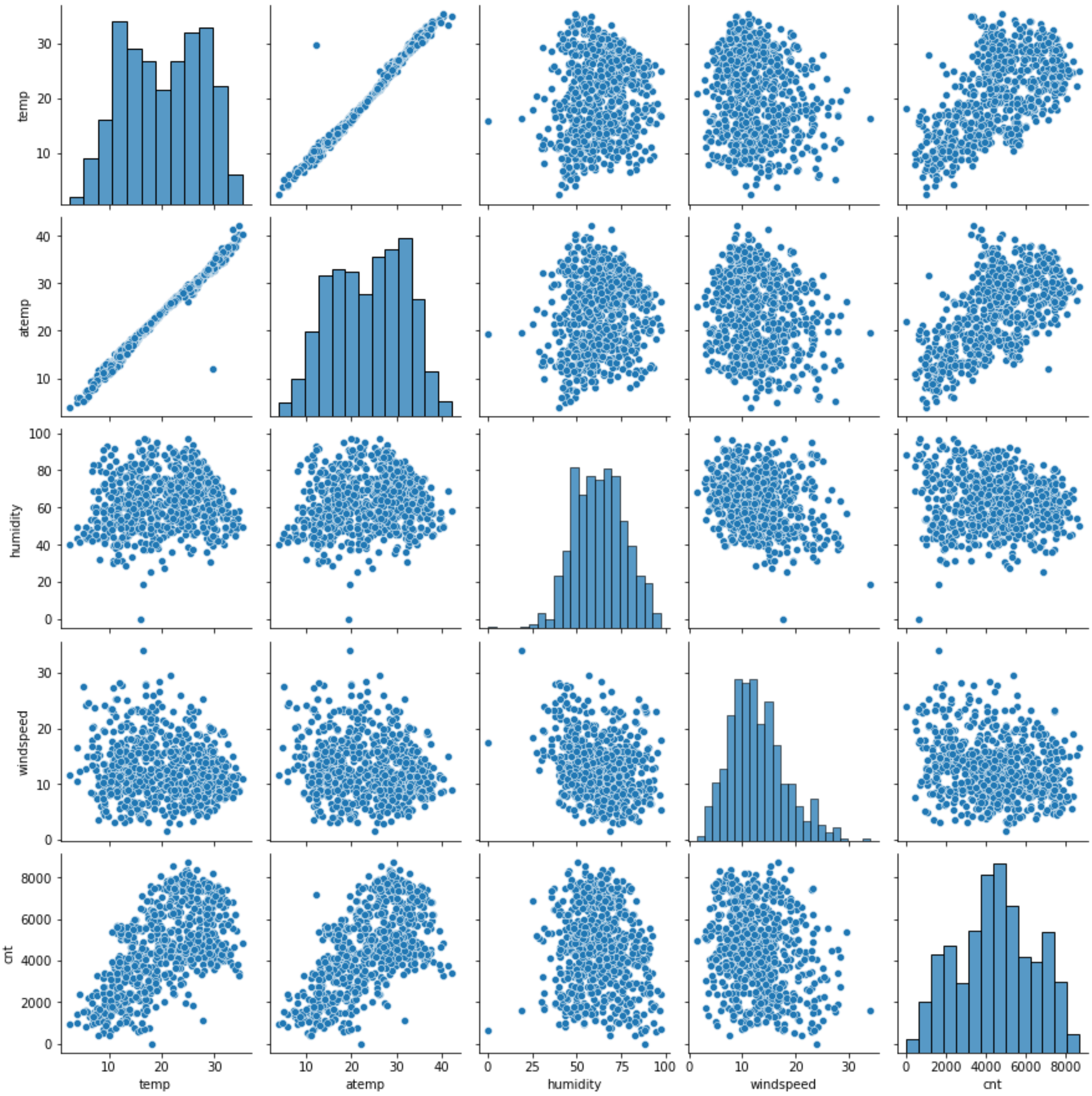


## 2. Why is it important to use drop_first=True during dummy variable creation?

- If we don’t drop the first column, then the dummy variables will be correlated. This may affect some models adversely and effect is stronger when cardinality is smaller.
- drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces correlations created among dummy variables.
- Let’s say we have 3 types of values of categorical column, and we want to create dummy variable for that column
- We have furnishing status categorical column has 3 categories like furnished, unfurnished and semi-furnished.
- Now, if we create dummy variables, we don’t need 3 categories – we can drop the furnished column, as the type can be identified with last two unfurnished and semi-furnished columns
- Syntax: -
  drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

Using the below pair-plot it can be seen that,” temp”, and “atemp” are the two numerical variables which are highly correlated with the target variable(cnt). “atemp” will be removed due to high VIF.
“temp” variable has the highest correlation with the target variable.



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

I have validated the assumptions of Linear regression Model based on below 5 assumptions: -

- ✓ Normality of the error terms
  - ○ Error terms should be normally distributed
- ✓ Multicollinearity
  - ○ There should be insignificant multicollinearity among variables
- ✓ Linear relationship exists
  - ○ Linear relationship between the variables
- ✓ Homoscedasticity
  - ○ There should be no visible patterns in residual values
- ✓ Independence of residuals
  - ○ No auto correlation

**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 are:

- temp
- Year
- Light snowrain

# General subjective Questions

1. **Explain the linear regression algorithm in detail**
   Linear Regression may be defined as the statistical model that analyses the linear relationship between a dependent variable and independent variables. Linear relationship shows change (increase or decrease) in independent variables, also the dependent variable will change.
   Mathematically linear Regression expressed as
   Y = mx + c this is the standard equation
   m is slope of the regression line which represents the effect X has on Y.
   c is a constant, knows as y-intercept. If X = 0, Y would be equal to c

generally, we have,
$Y = \beta 0 + \beta 1X1$
Where Y is dependent variable
β0 is the intercept
β1 is the co-efficient of X1(independent variable)
X1 is independent variable

Furthermore, the linear relationship can be positive or negative in nature as explained below: -

- Positive Linear Relationship: -
  A linear relationship positive when independent variables increase, the dependent variable also increases.
- Negative Linear Relationship: -
  A linear relationship positive when independent variables increase, the dependent variable decrease.
- No Linear Relation: -
  when there is no linear relation exists between dependent and independent variables.

Linear Regression is of the following two types: -

- Simple Linear Regression – only one independent variable
- Multiple Linear Regression – multiple independent variables

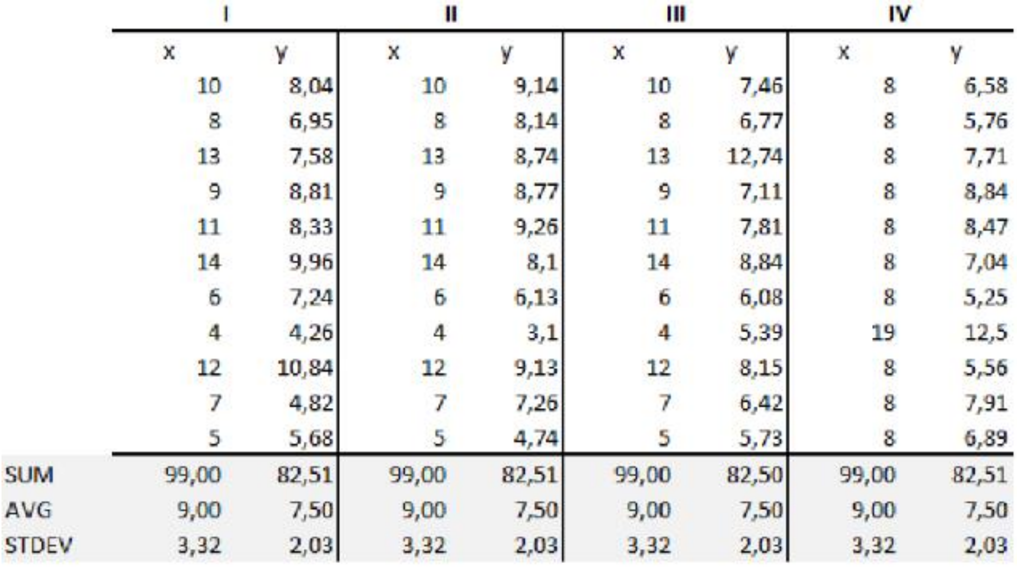
Assumptions of Linear Regression

- ✓ Normality of the error terms
  - Error terms should be normally distributed
- ✓ Multicollinearity
  - Linear Regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features dependency in them.
- ✓ Linear relationship exists
  - Linear relationship should exist between the dependent and independent variables.
- ✓ Homoscedasticity
  - There should be no visible patterns in residual values

- ✓ Auto correlation
  - o Assumes that there is very little or no auto-correlation in the data.it occurs when there is dependency between residual errors.

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

Anscombe’s quartet was developed by statistician Francis Anscombe. It compromises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they are the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

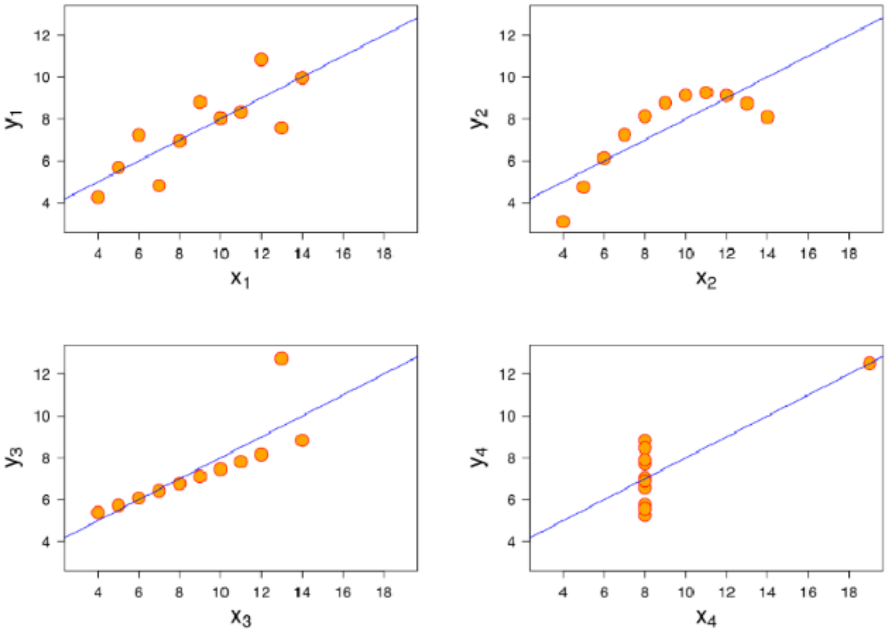


| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| | 11 | 8,33 | 11 | 9,26 | 11 | 7,81 | 8 | 8,47 |
| | 14 | 9,96 | 14 | 8,1 | 14 | 8,84 | 8 | 7,04 |
| | 6 | 7,24 | 6 | 6,13 | 6 | 6,08 | 8 | 5,25 |
| | 4 | 4,26 | 4 | 3,1 | 4 | 5,39 | 19 | 12,5 |
| | 12 | 10,84 | 12 | 9,13 | 12 | 8,15 | 8 | 5,56 |
| | 7 | 4,82 | 7 | 7,26 | 7 | 6,42 | 8 | 7,91 |
| | 5 | 5,68 | 5 | 4,74 | 5 | 5,73 | 8 | 6,89 |
| SUM | 99,00 | 82,51 | 99,00 | 82,51 | 99,00 | 82,50 | 99,00 | 82,51 |
| AVG | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 |
| STDEV | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 |

The summary statistics shows that means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation coefficient between x and y is 0.816 for each dataset

When we plot these four datasets on x/y coordinate plane, we can observe that they show the same linear regression lines as well, but each dataset is telling a different story.



- Dataset 1 appears to have clean and well-fitting linear models.
- Dataset 2 is not distributed normally.
- Dataset 3 distribution is linear, but calculated regression is thrown off by an outlier.
- Dataset 4 shows that one outlier is enough to produce high correlation coefficient.

This quartet emphasizes the importance of visualization in data analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

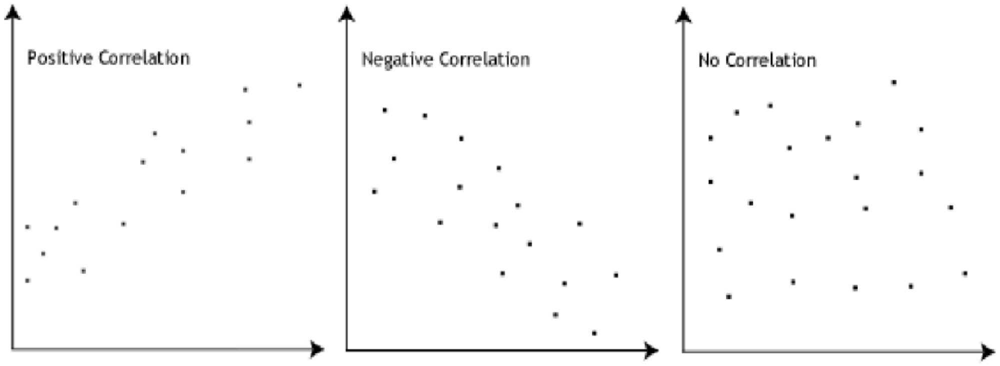
### 3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient also referred as Pearson’s r, or the bivariate correlation, is a measure of linear correlation between two sets of data. It is covariance of two variables, divided by their standard deviations, values vary between -1 to +1.

The Pearson correlation coefficient varies between -1 to + 1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in same direction).
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different direction).
- r = 0 means there is no linear association.

- r > 0 < 5 means there is weak association.
- r > 5 < 8 means there is a moderate association.
- r > 8 means there is a strong association.



Pearson r Formula

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

Where,

r – correlation coefficient
xi – values of the x-variable in the sample.
x bar – mean of values of the x -variable
y bar - mean of values of the y -variable
yi – values of the x-variable in the sample.

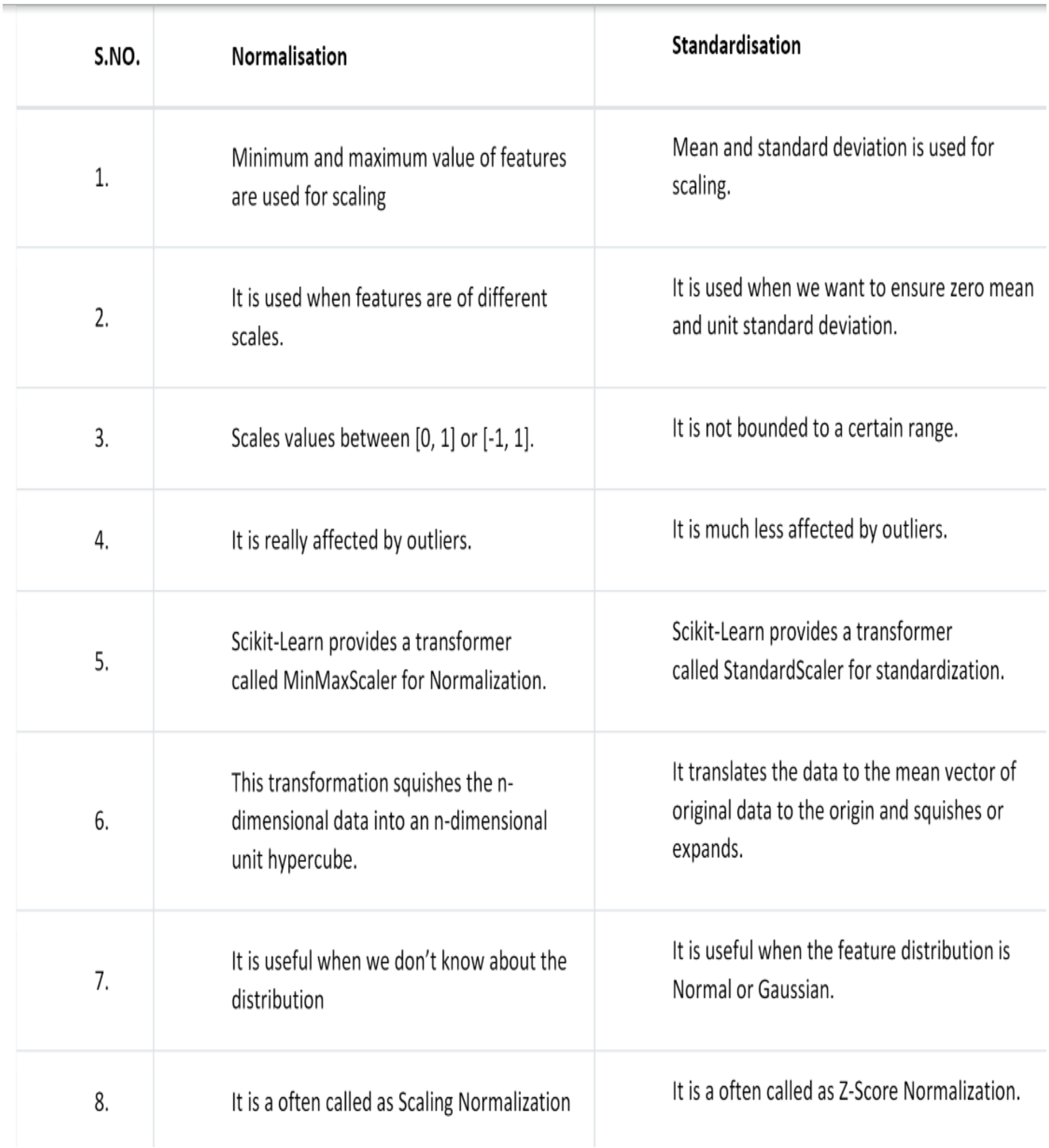
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   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 algorithm.

   Most of the times, collected dataset contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes the magnitude in account and not units hence incorrect modelling.

To solve this issue, we must 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 p-value, t-statistic, F-statistic, R-squared etc.

Normalization typically means rescales the value into a range of [0,1].
Standardization means rescales data to have mean 0 and standard deviation of 1.



| S.NO. | Normalisation | Standardisation |
|---|---|---|
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 6. | This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| 7. | It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| 8. | It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

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

If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) = 1, which leads to 1/(1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this multicollinearity.

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

The quantile-quantile(q-q) plot is a graphical technique for determining if two datasets come from the populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3(3%) quantile is the point at which 30% percent of data fall below and 70% fall above that value. A 45-degree reference line is plotted. If 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 datasets have come from the 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 datasets to obtain the 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.