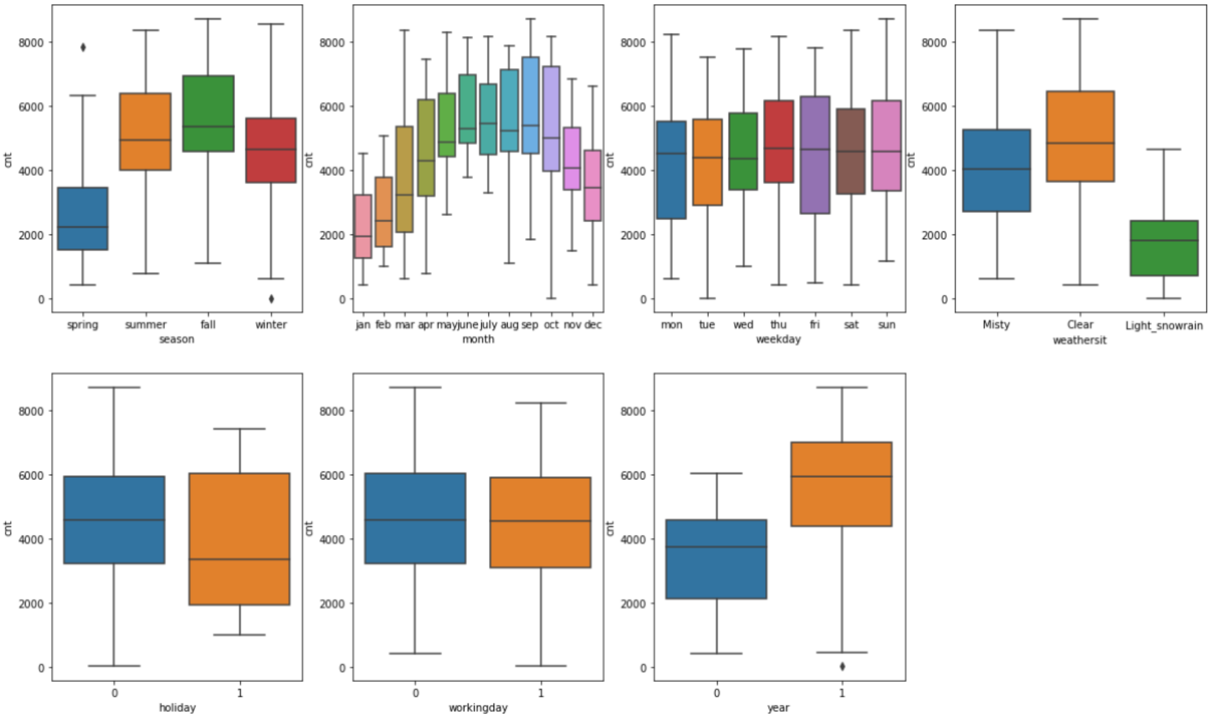
Assignment-based Subjective Questions

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

ANSWER:

Season, Year, Month, Holiday, Weekday, Weathersit are categorical variables in the dataset. From the analysis, it can be inferred that: -

1. Season: - The FALL season has the highest demand of rental bikes
2. Year: - Demand has grown for the next year i.e., 2019
3. month: - demand has continuously grown till Jun and there is fall from sept till dec
4. weekday Friday has most demand however no proper inference can be taken at this point
5. The clear weathersit has highest demand.



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

ANSWER:

By using drop_first=True we intend to drop the first level of dummy variable created. By using drop_first=True we avoid the most vital part of dummy creation that is multicollinearity.

We use Syntax of dummy=pd.get_dummies

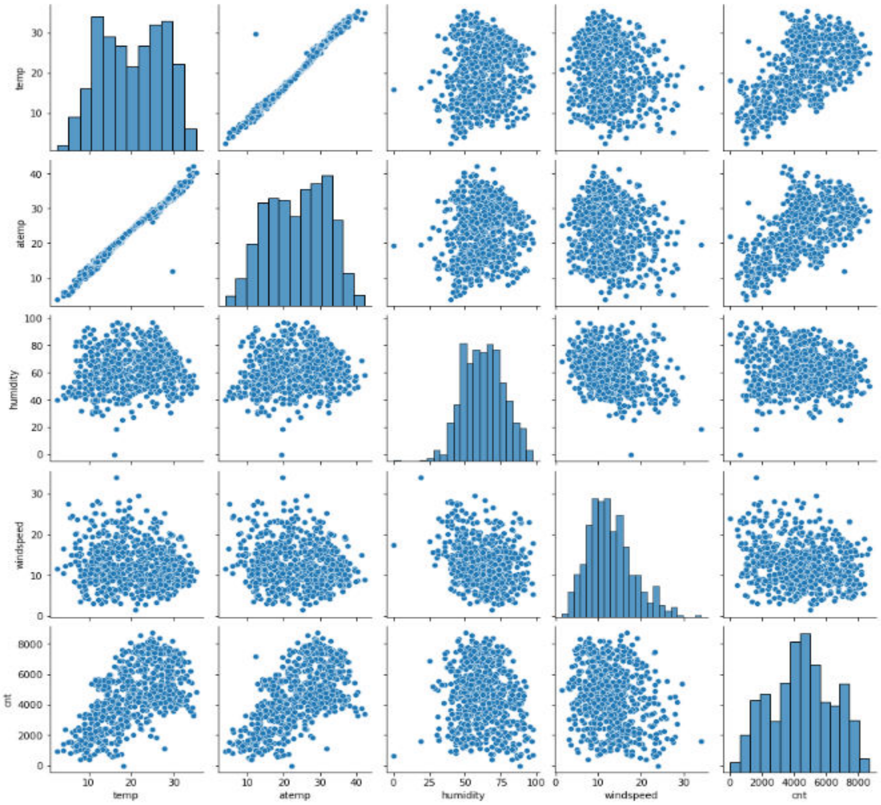
Let us say we have 3 types of values in categorical column example as shown below

Travel type	Train	Bike	Plane
Plane	0	0	1

Bike	0	1	0
Train	1	0	0

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

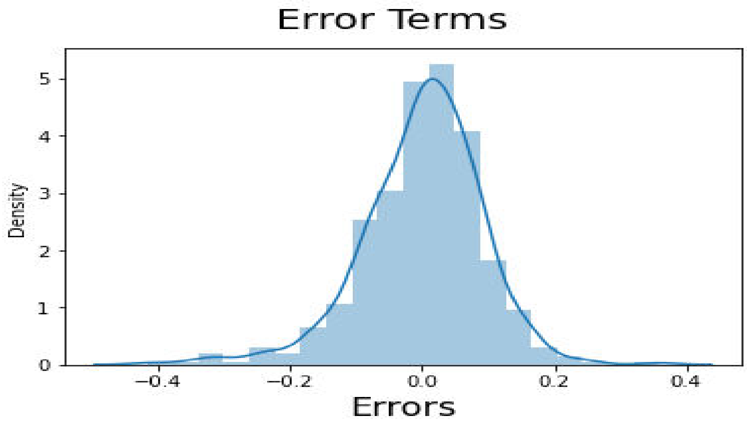
ANSWER:



As per the pair plot –the highest correlation with target variable is clearly “temp” and “atemp”.

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

ANSWER:



The distribution of residuals should be normal and centred around 0. (This mean is 0). We test the residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not.

The residuals are scattered around mean = 0 as seen in the diagram above.

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

ANSWER:

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.4777' indicates that a unit increase in temp variable increases the bike hire numbers by 0.4777 units.
- windspeed: - A coefficient value of -0.1481' indicates that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1481 units.
- Year (yr) - A coefficient value of '0.2341' indicates that a unit increase in year variable increases the bike hire numbers by 0.2341 units.

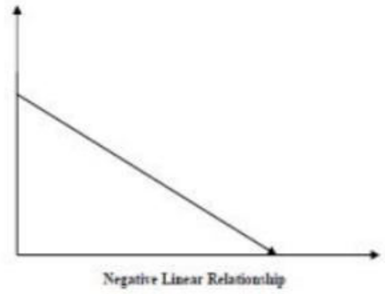
General Subjective Questions

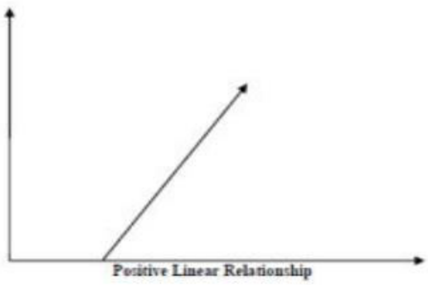
1. **Explain the linear regression algorithm in detail.?**

ANSWER:

Linear regression is one of the most fundamental algorithms in Machine learning. Basically it performs regression tasks. Based on the provided data the model predicts a dependent variable (TARGET) on other independent variable which has a direct impact on the target variable. It is mostly used for finding out the relationship between variables and forecasting. The regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression. There are Simple Linear Regression and Multivariate Linear Regression. The linear regression model is represented by the equation of a straight line: y = mx + b. Here, y represents the predicted target variable, x is the input feature, m is the slope of the line, and b is the y-intercept (the value of y when x = 0).

After analyzing the data and identifying if there are any null values and cleaning the data till, we can do appropriate analysis we will perform EDA on the data. After EDA we split the data into training data (which will be used to train a model) and test data (which will be used to check how close is our model to the actual output). As the model is prepared, we will check for p value determine if the results are statistically significant or not and we will check the VIF for the magnitude of multicollinearity in the model, and dropping the variables accordingly till we get the perfect model. After that we do residual analysis check and check if the curve must be a normal curve. Finally, the conclusion drawn from the model will provide valuable insights/predictions of the data set.

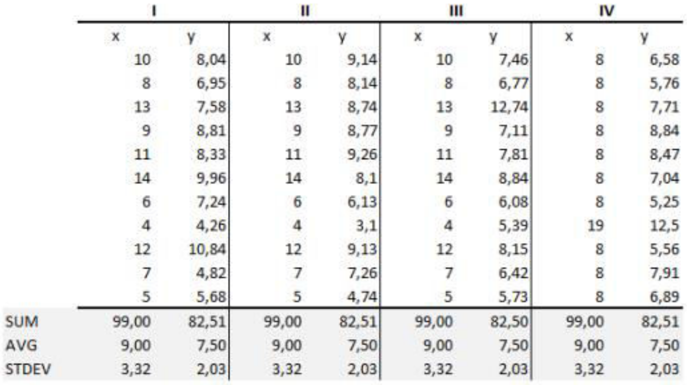




2. Explain the Anscombe’s quartet in detail.?

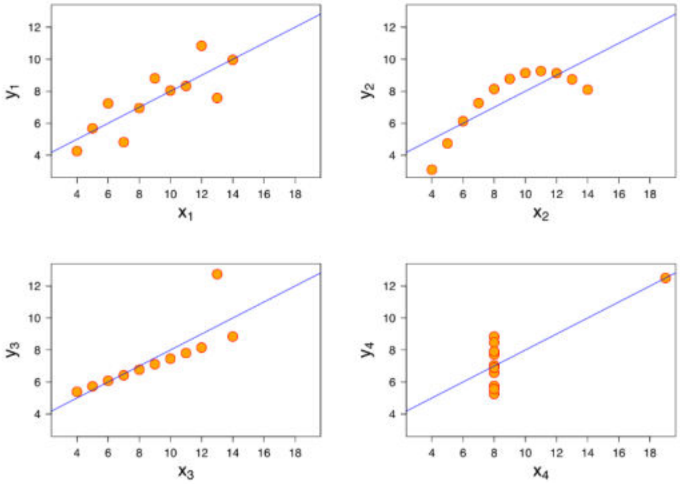
ANSWER:

Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize 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 show that the 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 (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a 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?

ANSWER:

It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, Pearson’s R is also known as Pearson’s correlation coefficient is a measure of strength of co relation between two variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

The value of Pearson’s always lies between -1 to +1. It is widely used in statistics and data analysis to access the strength and direction of linear relationship between variables.

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

ANSWERS:

Scaling is a term used in linear regression which refers to transforming the input features to similar scale or range. It involves adjusting the values of variable to common scale which will help to improve the performance and interpretation of the model.

Scaling is performed because, scaling allows fair comparison of variables with different units and scales or range without proper scaling the values of variables may differ which will have a great impact on the model. Scaling is essential for visualization data when using plots or charts to ensure visual representation is accurate. Lastly it enhances the reliability and effectiveness of the analysis

There are 2 types of scaling “NORMALIZED SCALING” AND “STANDARDIZATION SCALING”

STANDARDIZATION SCALING: - Standardization scales the variable to have a mean of zero and a standard deviation of one. Standardization maintains the distribution shape of variables and is useful when the variables have different units or different scales of measurement

NORMALIZED SCALING: - Normalization Scaling or Min Max Scaling is rescaling the variables to a specific range, typically between 0 and 1. Normalized scaling compresses the variable values into specific range and is useful when maintaining the original scale.

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

ANSWER:

In some cases, The VIF is infinity, because a large value of VIF indicates that there is correlation between variables, if there is perfect correlation then VIF = INFINITY. Therefore, VIF if observed as infinity. This variable must be dropped in order to successfully run the model. The VIF is calculated as VIF = 1 / (1 - R1^2)

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

ANSWER:

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential.

The use and importance are as follows:

- Linear regression assumes that the residuals (difference between observed and predicted values follow a normal distribution. This plot allows us to visualize if the plot is normally distributed or not

- Outliers are data points that deviate significantly this helps in identifying outliers

- This plot provides visual tool to evaluate the adequacy of the model by examining the residual distribution properties

- This plot serves as a diagnostic tool in linear regression analysis.