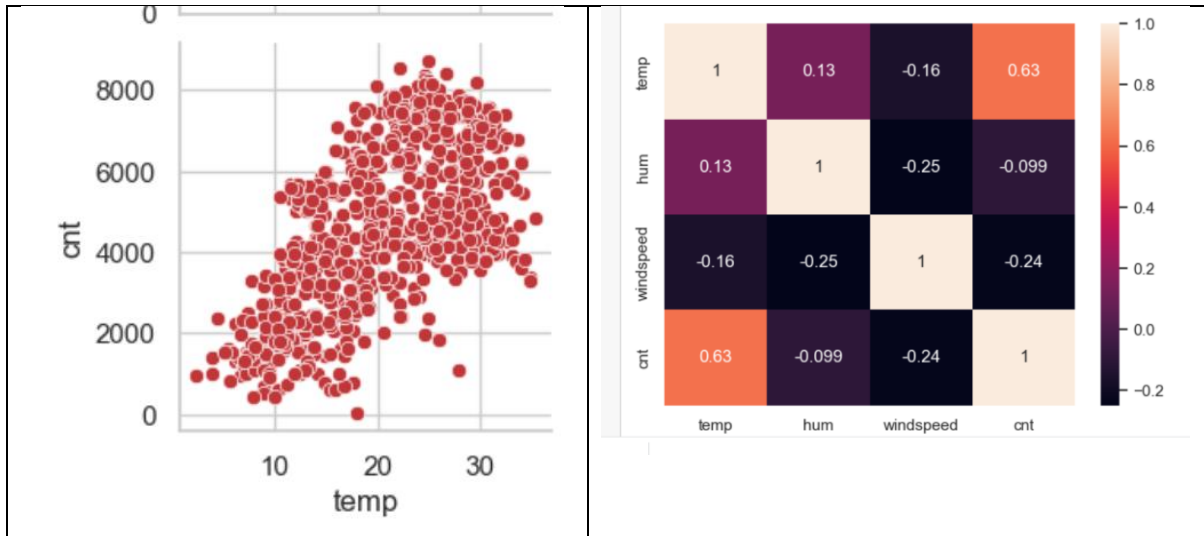


Assignment-based Subjective Questions

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

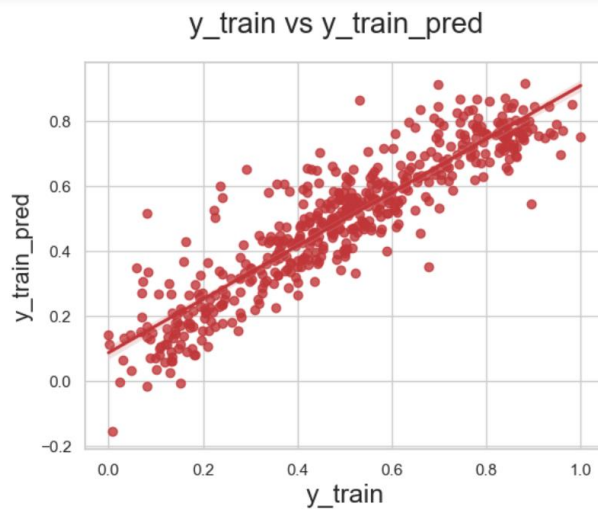
Year (yr)	Bike rental demand has been increased in significantly 2019 than the previous year.
Month(mnth)	<ul style="list-style-type: none"> ▪ Bike rental demand consistently increases from Jan-June ▪ Highest demand observed in the period from June to September. ▪ There is a down trend from October to December. ▪ Jan has the lowest demand while September has the highest demand.
Weekday	Demand between Mon-Fri slightly higher than weekends.
Holiday	There is a low demand on holidays compare to the non-holidays.
Workingday	Demand during workdays slightly higher than non-working days.
Season	Demand is high during summer and fall seasons while it's low during spring season.
Weathersit	<ul style="list-style-type: none"> ▪ Demand is high during Clear, Few clouds, Partly cloudy, Partly cloudy while demand is low during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds season ▪ No data found for the Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog category.

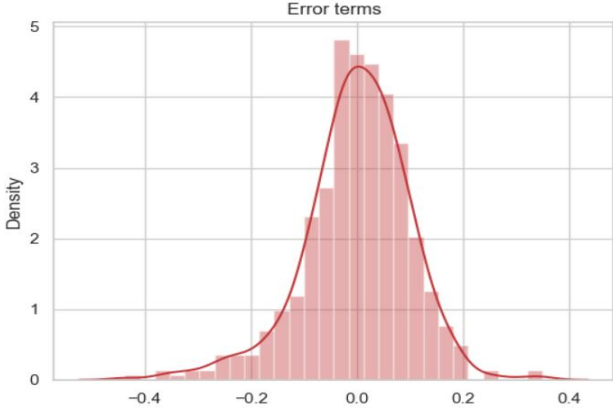
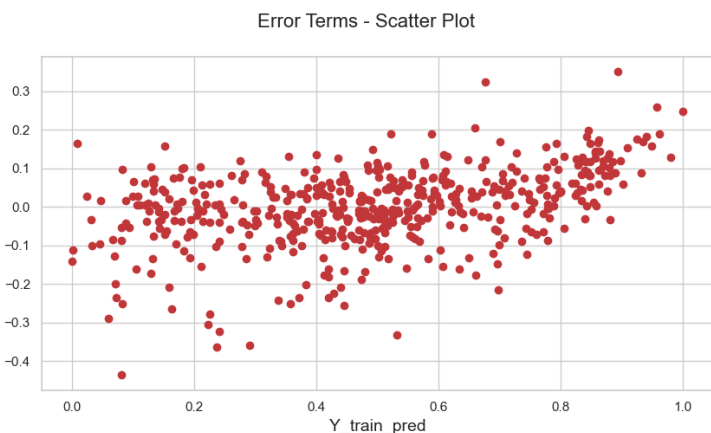
2. Why is it important to use drop_first=True during dummy variable creation?
While creating the dummy variables by default it would create n-number of columns for n different values (values/labels of the category variable).
By using the drop_first=True option we can reduce one additional column and make the outcome as 'n-1' new columns instead of 'n' new columns. Here when all n-1 columns are zero it implicitly indicates the Nth dummy variable (without having it explicitly).
This helps reduce the reduces the correlations created among dummy variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
 - "temp" (i.e Temperature) has the highest correlation with target variable "cnt".
 - Apart of this "atemp" which had high correlation with "temp" correlates well with target variable as well. But this column has been dropped and not considered for modeling as it's derived from and covered by "temp".



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
- Model is validated against the *Assumptions of simple linear regression*
 - i. Linear relationship between X and y.
 - ii. Normal distribution of error terms.
 - iii. Independence of error terms.
 - iv. Constant variance of error terms(homoscedasticity).

Points are scattered symmetrically regression line.



<p>The histogram and distribution plots shows the Error terms are normally distributed with a mean value 0.0.</p>	 <p>The plot titled 'Error terms' is a histogram with a red normal distribution curve overlaid. The x-axis ranges from -0.4 to 0.4 with major ticks at -0.4, -0.2, 0.0, 0.2, and 0.4. The y-axis is labeled 'Density' and ranges from 0 to 5 with major ticks at 0, 1, 2, 3, 4, and 5. The histogram bars are light red, and the curve is a solid red line peaking at 0.0 with a density of approximately 4.5.</p>
<p>No significant patterns observed hence the Error Terms are independent of each other.</p>	 <p>The plot titled 'Error Terms - Scatter Plot' shows residuals on the y-axis against predicted values (Y_train_pred) on the x-axis. The x-axis ranges from 0.0 to 1.0 with major ticks at 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. The y-axis is labeled 'Residual' and ranges from -0.4 to 0.3 with major ticks at -0.4, -0.3, -0.2, -0.1, 0.0, 0.1, 0.2, and 0.3. The plot contains numerous red dots scattered randomly across the area, indicating no significant patterns or trends.</p>
<p>Error terms are nearly constant and follows the 4th Linear Regression assumption of Homoscedasticity</p>	

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

- “**temp**” (i.e Temperature) is the most significant Feature (with coefficient of 0.552) which positively influence the bike rental demand.
- “**weathersit**” type 3 (i.e **Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**) – Influence the Bike rental demand negatively (with coefficient of -0.264).
- ‘**yr**’(Year) – is a significant feature variable which positively influence the model indicating that the demand for Bike rental is increasing year to year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

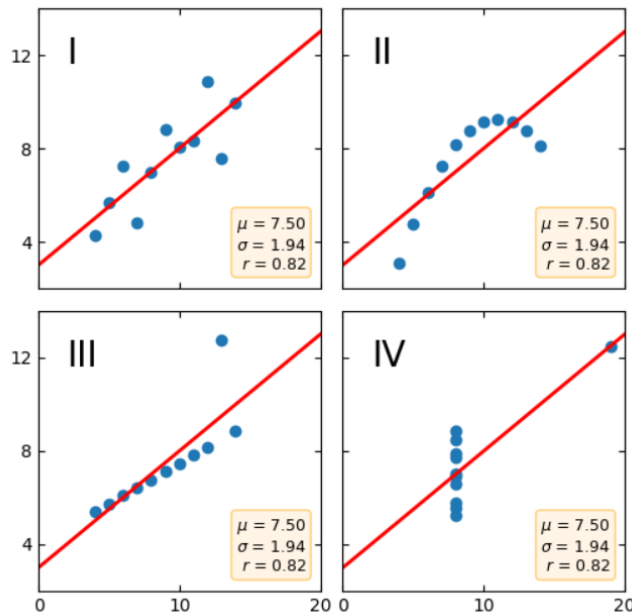
- Linear regression is a machine learning algorithm which supervised learning focuses on finding the best linear fit relationship between set of independent variables(features) and dependent variables.
- Outcome of a of a Linear Regression algorithm model is a best fitting linear equation in which Out/Target Variable described by most influencing features (independent variables)
- There are 2 type of linear regression algorithms
 - i. Simple Linear Regression Algorithm
 - Represents the relationship between a dependent variable and only one independent variable using a straight line.
 - Equation

$$Y = \beta_0 + \beta_1 X$$
 - β_0 is Intercept
 - ii. Multiple Linear Regression Algorithm
 - Represents the relationship between a dependent variable and multiple independent variables using a straight line.
 - Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X$$
 - β_0 is Intercept
 - $\beta_1, \beta_2 \dots \beta_p$ are the slope/gradient.
- Cost functions are used to identify the best fit regression line for the coefficients. Following two cost functions are used to achieve this.
 - i. Gradient descent Approach
 - ii. Differentiation
- During the regression process to find the best fit line the residuals (square of the difference between the actual target values and prediction values) are calculated and these are minimized using the method of Ordinary Least Square (OLS).
- Python provides “statsmodels” and “SKLearn” libraries which can be used to t for the linear regression.

2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.
- These 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.
-
- It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.
- For example, consider the following illustration where 4 different datasets all with identical descriptive statistics (mean = 7.50, standard deviation = 1.94 and r = 0.82) but has differently distributed.



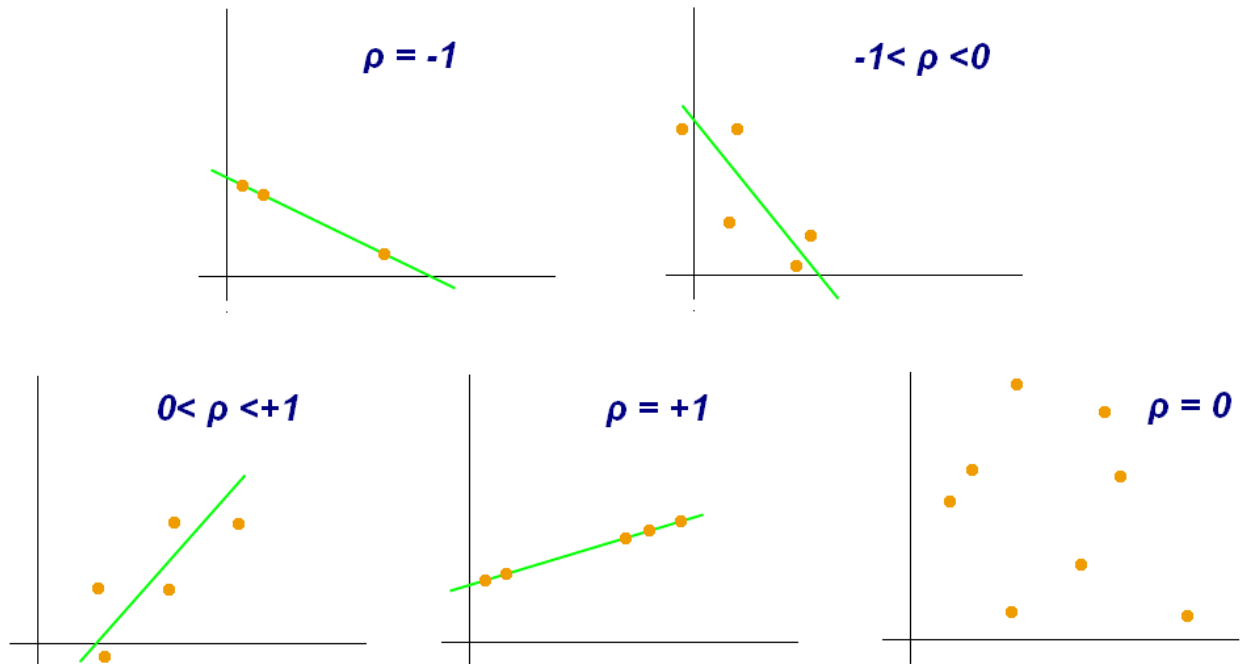
- The four datasets above can be described as:
 - **Dataset 1:** this **fits** the linear regression model pretty well.
 - **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
 - **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
 - **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- All the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

Value interpretation

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and $-.3$	Weak	Negative
Between $-.3$ and $-.5$	Moderate	Negative
Less than $-.5$	Strong	Negative



Equation	Equation details
$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$	<p>r = correlation coefficient</p> <p>x_i = values of the x-variable in a sample</p> <p>\bar{x} = mean of the values of the x-variable</p> <p>y_i = values of the y-variable in a sample</p> <p>\bar{y} = mean of the values of the y-variable</p>

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
 - What is scaling?
 - It is a data Pre-Processing step which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
 - Why is scaling performed?
 - Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account

and not units hence incorrect modelling. To solve this issue, we have to 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 t-statistic, F-statistic, p-values, R-squared, etc
- What is the difference between normalized scaling and standardized scaling?
 - Normalization/Min-Max Scaling:
 - Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.
 - Formula:
 - Standardization
 - Standardization is scaling technique in which the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
 - Formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

- VIF represented with the following equation.

$$VIF = \frac{1}{1 - R^2}$$

-

- When we get $R^2 = 1$ it would lead to $1/(1-R^2)$ infinity. This can happen when two independent variables are in perfect correlation.
- To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

- Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.
- Since this is a visual comparison, results can be subjective but useful in the understanding underlying distribution of a variable.
- A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- Q-Q plot is helpful in linear regression to validate if the given testing and training datasets are from same population with same distributions as it's important to factor in order to maintain the sanity of the model.