Assignment-based:

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

Ans: From the dataset we have, can see that:

- Bike Rentals are more during the Fall season and then in summer
- Bike Rentals are more in the year 2019 compared to 2018
- Bike Rentals are more in partly cloudy weather
- Bike Rentals are more on Saturday, Wednesday and Thursday
- Bike rentals are more in clear weather then in Light snow
- Bike Rentals are same on working day and holidays
- Bike rentals are more in August, September and October months

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

Ans: When using **pandas.get_dummies**, there is a parameter i.e. **drop_first** so that whether to get m-1 dummies out of n categorical levels by removing the first level.

It is important to use **drop_first = True,** If we don't use "drop_first" we will get a redundant feature. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

For example:

If we have a variable gender, we don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female.

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

Ans: From Pair- plot we can see that **temperature** has the highest correlation with the target variable **'count'**.

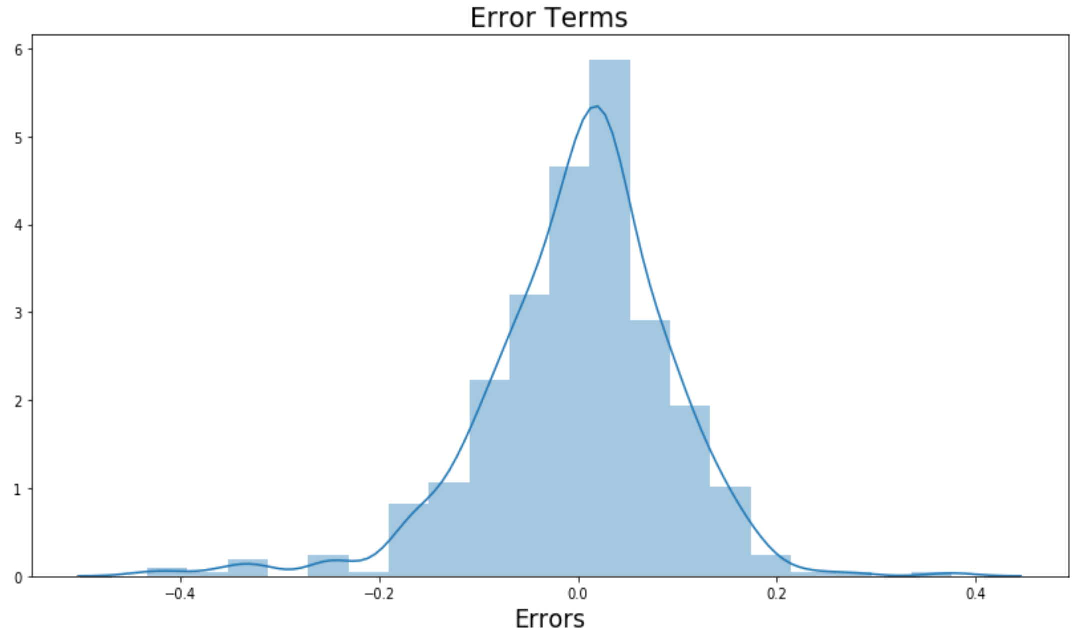
4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: Validating the assumptions:

- Before making the predictions, we need to check that the model is reliable. For that, we need to perform a residual analysis of the error terms
- Residuals are the difference between any data point and the regression line, they are sometimes called “errors.” Error in this context doesn’t mean that there’s something wrong with the analysis; it just means that there is some unexplained difference. In other words, the residual is the error that isn’t explained by the regression line.

  Residual = Observed – Predicted

- Used Distribution plot on the residuals and see if it is normally distributed. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.
- This assumes homoscedasticity, which is the same variance within our error terms. Heteroscedasticity, the violation of homoscedasticity, occurs when we don’t have an even variance across the error terms.
- We can see that the residuals center on zero, which indicate that the model’s predictions are correct on average rather than systematically too high or low. Regression also assumes that the residuals follow a normal distribution
- Then made the predictions on the test set, and evaluated the model based on the predictions.

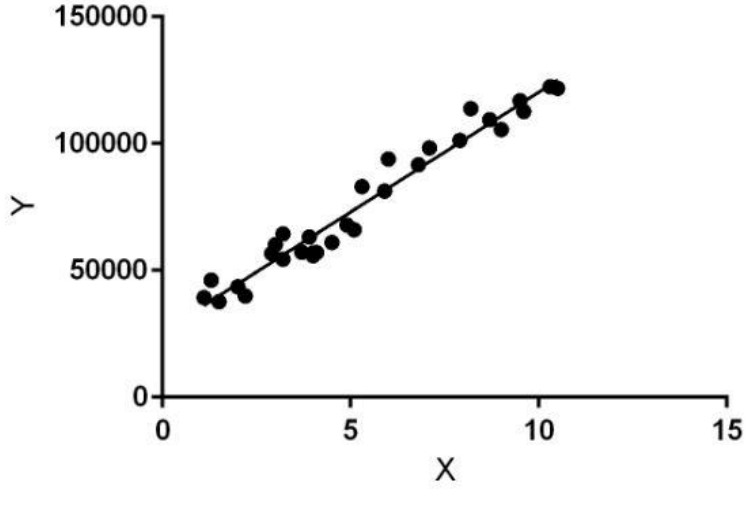


5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: In the final model,
   - We got all the p values within the acceptable range**.** A low p-value (< 0.05) indicates that you can reject the null hypothesis. And all **the p values we got are less than 0.05.**
   - A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity**. In our case, with values less than 5**, we are in good shape, and can proceed with our regression.
   - R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. **And we have the R-square value of 0.826 or 82.6%**
   - The **adjusted R-squared** adjusts for the number of terms in the model
     We have **adjusted R-squared value of 82%**
   - Also we can see in the residual analysis, the error terms are normally distributed and centered at 0, which indicates that the models predictions are correct.

.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.
   Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$\boldsymbol{y = \beta 0 + \beta 1 X}$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best $\beta 1$and $\beta 2$ values.
$\beta 1$: Slope(coefficient of x)
$\beta 0$: intercept

Implementing Linear Regression
The process takes place in the following steps:

- Loading the Data
- Exploring the Data
- Slicing The Data
- Train and Split Data
- Generate The Model
- Evaluate The accuracy

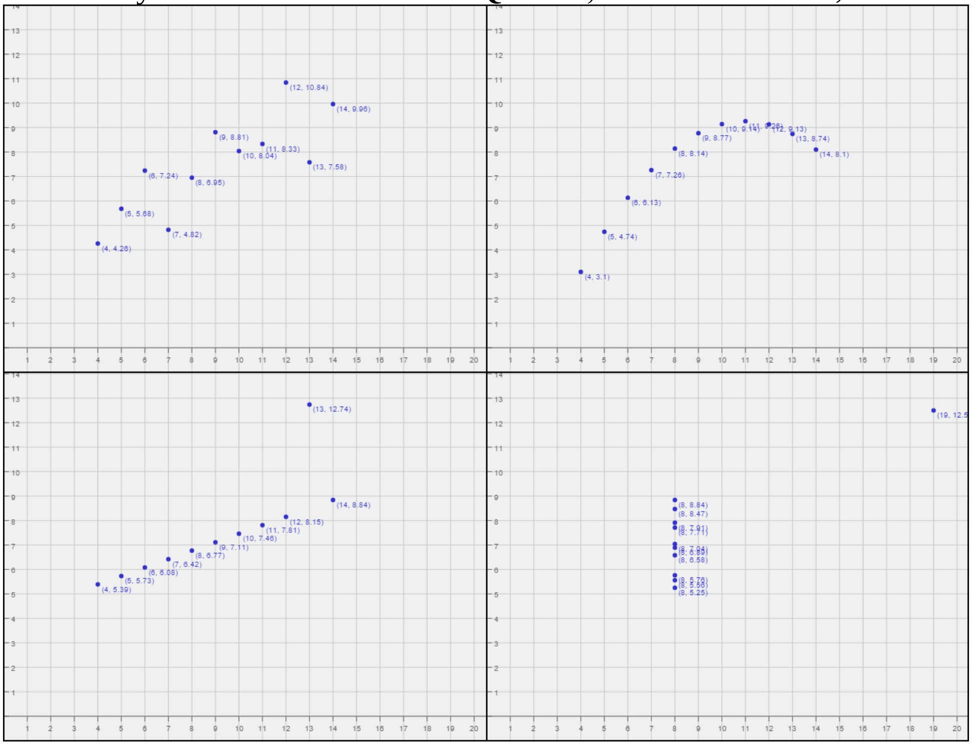
Linear Regression Algorithm works on:

Ordinary Least Square:
Ordinary Least Squares (OLS), also known as Ordinary least squares regression or least squared errors regression is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters for a linear function, the goal of which is to minimize the sum of the squares of the difference of the observed variables and the dependent variables i.e. it tries to attain a relationship between them.

.**R Square Method – Goodness of Fit**
R–squared value is the statistical measure to show how close the data are to the fitted regression line

2) Explain the Anscombe’s quartet in detail.

Ans: Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. **Francis Anscombe realized this in 1973** and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as **“Anscombe's Quartet,”** are shown below,



Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even linear regressions cannot accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.

Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets
If the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict how to analyze your data. For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably should be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analyzing it.

Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

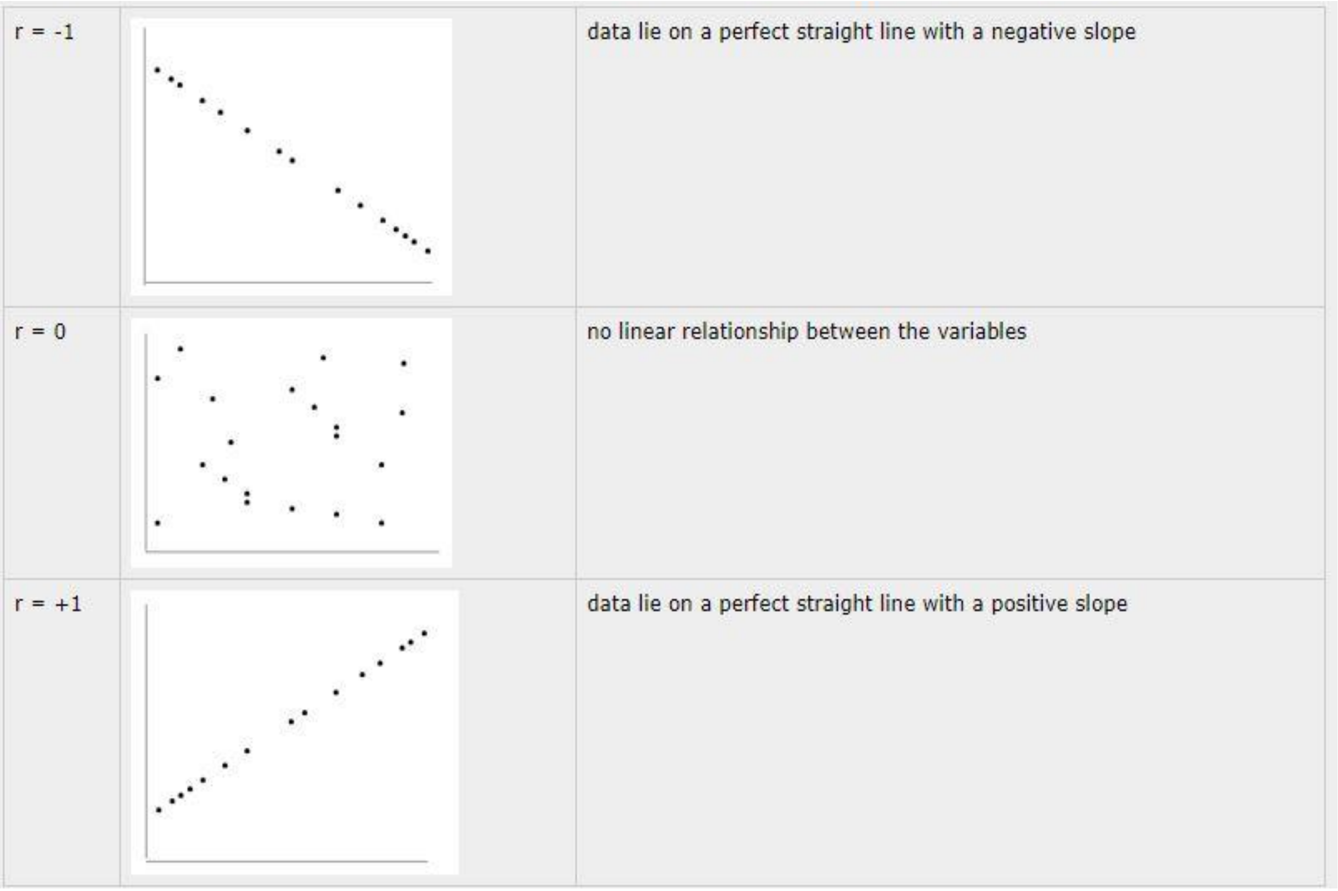
3) What is Pearson’s R?
Ans :
Pearson's Correlation Coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:



r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression.

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson’s correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods

Scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

We can scale the features using two very popular method:

1) Normalization : is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
   Formula:

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

   - When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
   - On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
   - If the value of X is between the minimum and the maximum value, the

2) Standardization: is another scaling technique where 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:

$$X' = \frac{X - \mu}{\sigma}$$

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

- If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is 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 the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.
- Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. **As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.**
- 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.
Ans:

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the

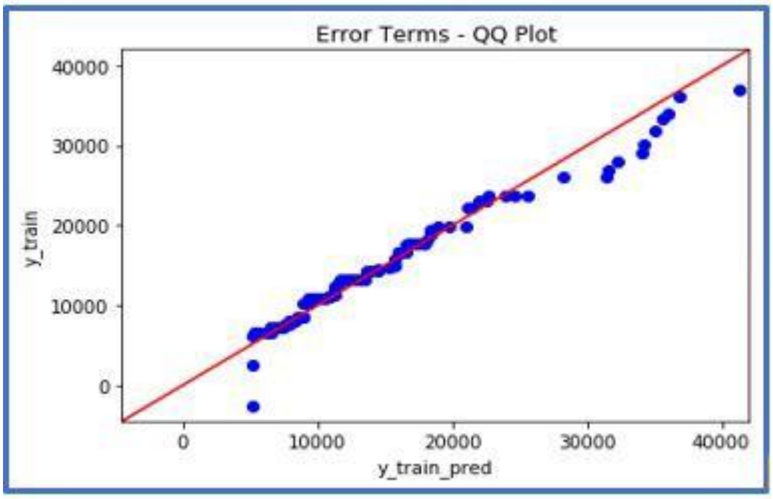
given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

- A 45-degree reference line is also plotted. If the 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 data sets have come from populations with different distributions.

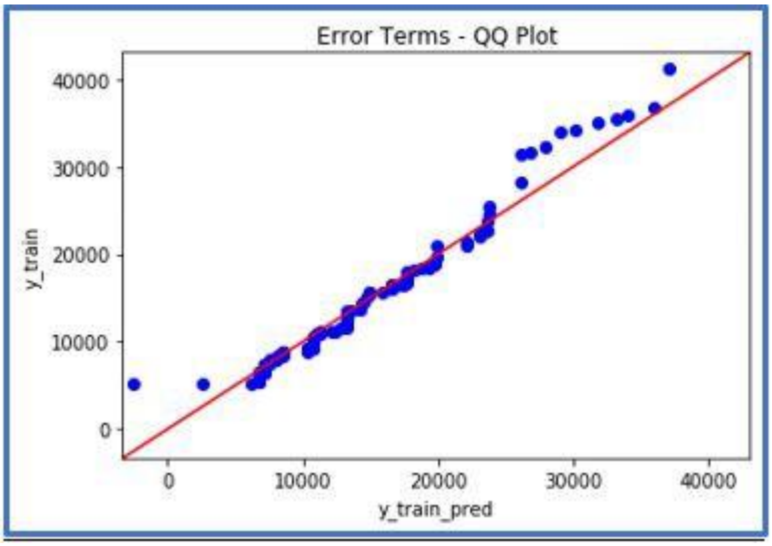
The advantages of the q-q plot are:

The sample sizes do not need to be equal.
Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis