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?

The categorical variables present in the data set are season, weekday, weathersit, mnth, holiday, workingday. These variables have having value of numeric based on their categories. So we can use heat map to find the correlation of these variables on the dependant variable.

According to the heat map,

1. The season variable is highly correlated to mnth and cnt variable and negative correlation with weekday.
2. Weekday has no notable impact on any variable and negatively correlated to yr, temp and hum.
3. Waethersit variable is correlated to hum variable and negatively related to temp and cnt.
4. Mnth variable is highly correlated to season and negatively related to windspeed
5. Holiday has a negative effect on the target variable.

6. Workingday doesn't have any notable positive effect on any variable and has a negative effect on casual variable.

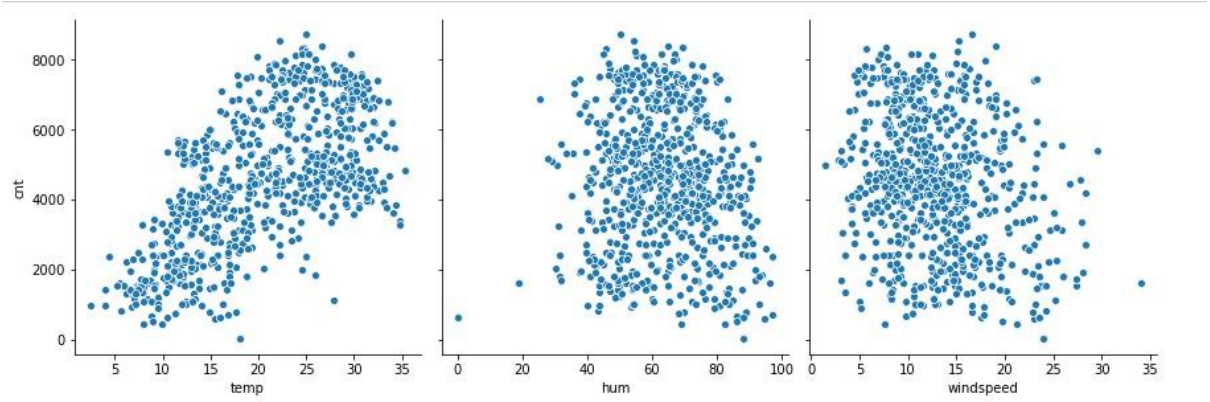
2. Why is it important to use drop_first=True during dummy variable creation?

We use drop_first = True while creating dummies because the value we are dropping in regression can be found with the remaining dummy columns itself. This is explained in detail in the example. So to avoid redundancy we are dropping a column.

Example:

If there is a categorical variable containing the colour of ball with the values of Red, Blue and Green. When we create dummies to this column so that Red, Blue and Green columns will be created. Here while creating the dummies we are using drop_first = True due to which one of the columns of Red, Blue, Green will be dropped. It is done because, if a row has Green as value the value in the dummy variable of Red and Blue would be 0 which gives it obvious that the value of Green dummy variable would be 1. So its value can be easily interpreted just the two columns. It is not necessary to have a separate column to determine its value using another variable. And it is advisable to create a dummy variable of n-1 columns which has n values. This is simply to avoid redundancy.

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



As found from the above pair plot of the numerical variables with the target variable (cnt), we can see that temp variable i.e. temperature variable has the highest correlation with the target variable.

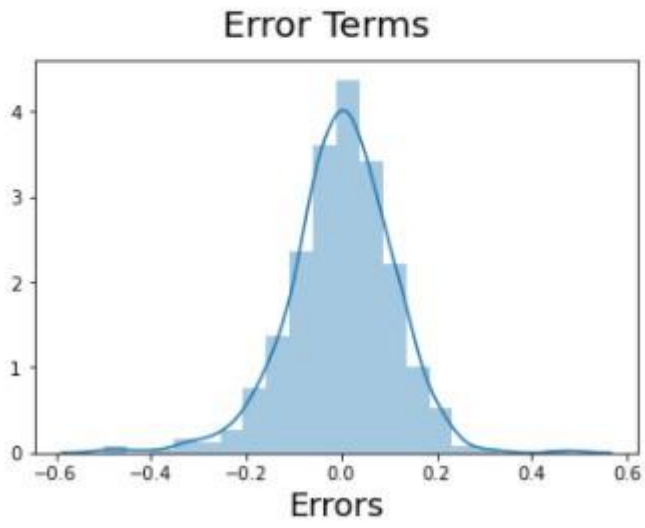
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The major assumptions of Linear Regression are,

1. X and Y should display some sort of a linear relationship. Otherwise, there is no use of fitting a linear model between them.
2. Error terms are normally distributed with mean (not X and Y)
3. Error terms are independent of each other.
4. Error terms have constant variance

Following are the assumptions done while building the linear regression model. Once the regression model is built on the training set, we use the distplot using seaborn to verify the error terms are normally distributed using the values of y_train and y_train_pred. From the below distplot it is evident that the error terms are normally distributed with mean at 0.

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
plt.show()
```



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

The equation of our best fitted line is:

cnt = (0.246788 * yr) - (0.061080* holiday) + (0.048800 * workingday) - (0.284481 * Spring) - (0.059272 * Summer) - (0.078259 * Winter) + (0.052886 * Saturday) - (0.323005 * Light Snow Rain) - (0.088213 * Mist & Cloudy) - (0.088167 * January) - (0.008956 * July) + (0.076304 * September)

Above is the equation of the best fit line for the model evolved. Through this we can find out that the top 3 features contributing significantly towards explaining the demand for shared bikes are,

1. Year (yr)
2. Working Day (workingday)
3. September

These variables are having positive relation with the cnt variable say, 1 unit increase in yr contributes 0.247 , 1 unit increate in workingday contributes 0.049 and 1 unit increase in September contributes 0.076 towards the cnt variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. Here are the types of regressions:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example:

We are running a sales promotion and expecting a certain number of count of customers to be increased now what we can do is we can look the previous promotions and plot if over on the chart when we run it and then try to see whether there is an increment into the number of customers whenever we rate the promotions and with the help of the previous historical data we try to figure it out or we try to estimate what will be the count or what will be the estimated count for our current promotion this will give us an idea to do the planning in a much better way about how many numbers of stalls maybe we need or how many increase number of employees we need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data. In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

Linear regression is used to predict a quantitative Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = mx+c$$

Where m and c given by:

m(Slope) = Change in y ⁄ Change in x (dy/dx)

c(intercept) = value of y when x=0

Here, x and y are two variables on the regression line.

m = Slope of the line. c = y-

intercept of the line.

x = Independent variable from dataset y

= Dependent variable from dataset

Use Cases of Linear Regression:

- Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
- Price Prediction – Using regression to predict the change in price of stock or product.
- Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe's quartet in detail.

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 when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

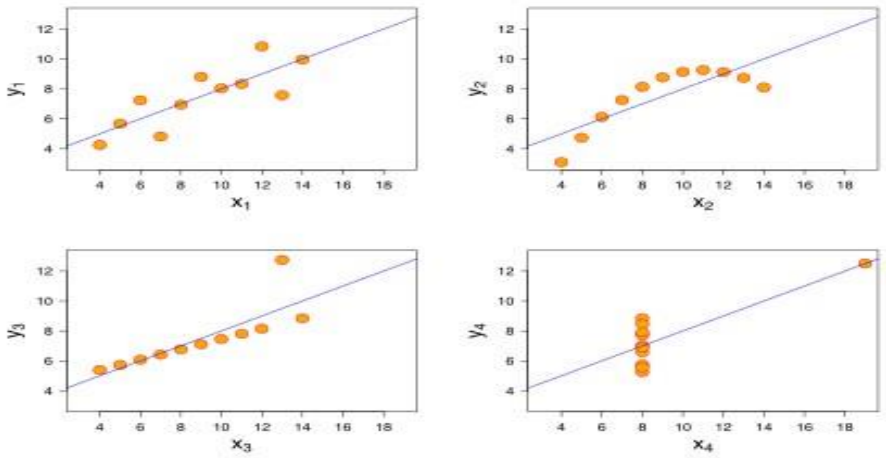
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

Quartet's Summary Stats:

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 appears 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?

The **Pearson product-moment correlation coefficient** is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we use r to represent Pearson's correlation unless otherwise noted. Pearson's r can

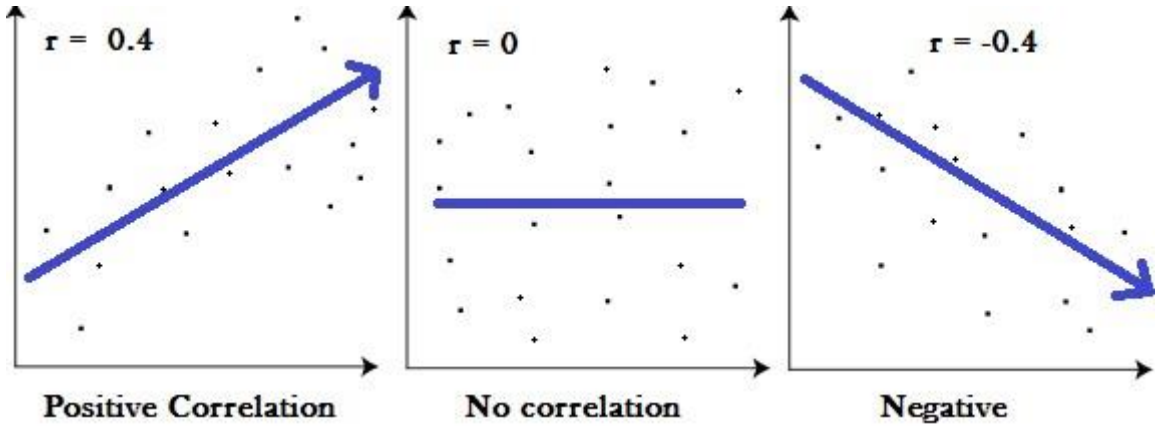
range from -1 to 1. With real data, you would not expect to get values of r of exactly -1, 0, or 1. Correlation coefficient formulas are used to find how strong a relationship is between data.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

.

The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all



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

Scaling 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 an algorithm. Most of the times, collected data set contains features highly varying in

magnitudes, units and range. If scaling is not done, the algorithm only takes magnitude in account and not units hence it leads to 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.

There are two types of scaling:

1. Normalization/Min-Max Scaling
2. Standardisation Scaling

Normalization/Min-Max Scaling:

- It brings all of the data in the range of and1 **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardised Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

The **variance inflation factor** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model. The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the *R*-

squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables.

The *variance inflation* for a variable is then computed as:

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

- 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.

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. 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.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

An **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in the scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets –

I. Come from populations with a common distribution
II. Have common location and scale
III. Have similar distributional shapes
IV. Have similar tail behaviour

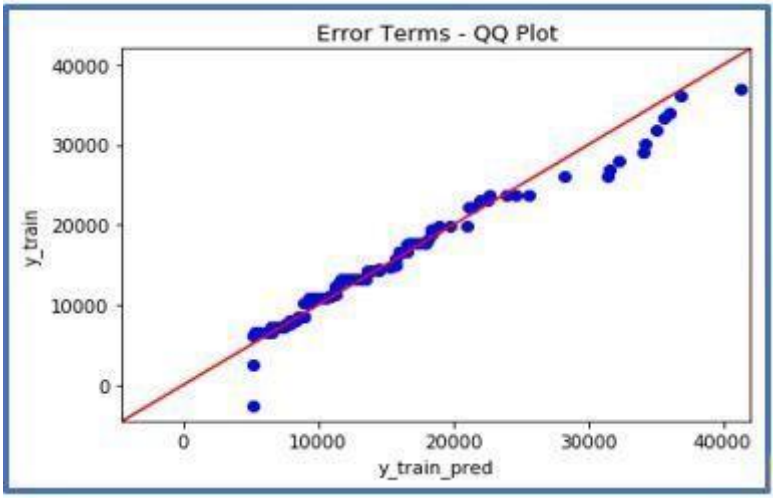
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

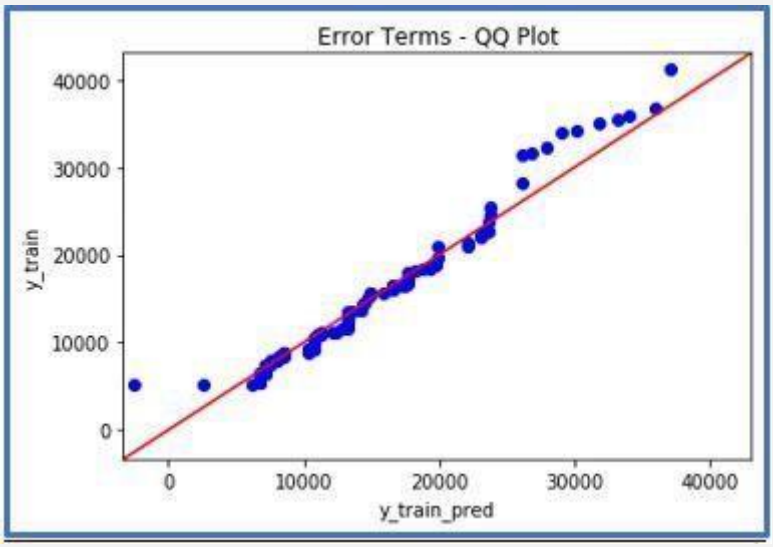
Below are the possible interpretations for two data sets.

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