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]: Some observations are stated below:

1. Demand picks up in Spring and rises in summer and fall and then decreases in Winter
2. Demand picked up in 2019 compared to 2018
3. Demand dips on working days
4. Demand declines as weather deteriorates as expected
5. BEST FIT LINE EQUATION based on Linear Regression Modelling

   cnt = 0.0779 (const) + 0.2337 * yr + 0.0544 * workingday + 0.5409 * atemp - 0.1337 * windspeed + 0.0987 * season_2 + 0.1320 * season_4 + 0.0650 * mnth_8 + 0.1163 * mnth_9 + 0.0648 * weekday_6 - 0.0837 * weathersit_2 - 0.2802 * weathersit_3

6. BoomBikes should try to increase demand/sales in seasons (Spring and Winter) and months (Aug and Sep) as these have positive coefficients
7. Negative Weather situations decrease demand/sales as indicated by the negative coefficients of windspeed, weathersit_2 and weathersit_3

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

[Answer]: It reduces the number of dummy columns by 1 (n-1) for categorical columns (n levels). Hence it reduces the correlations created amongst the dummy variables.

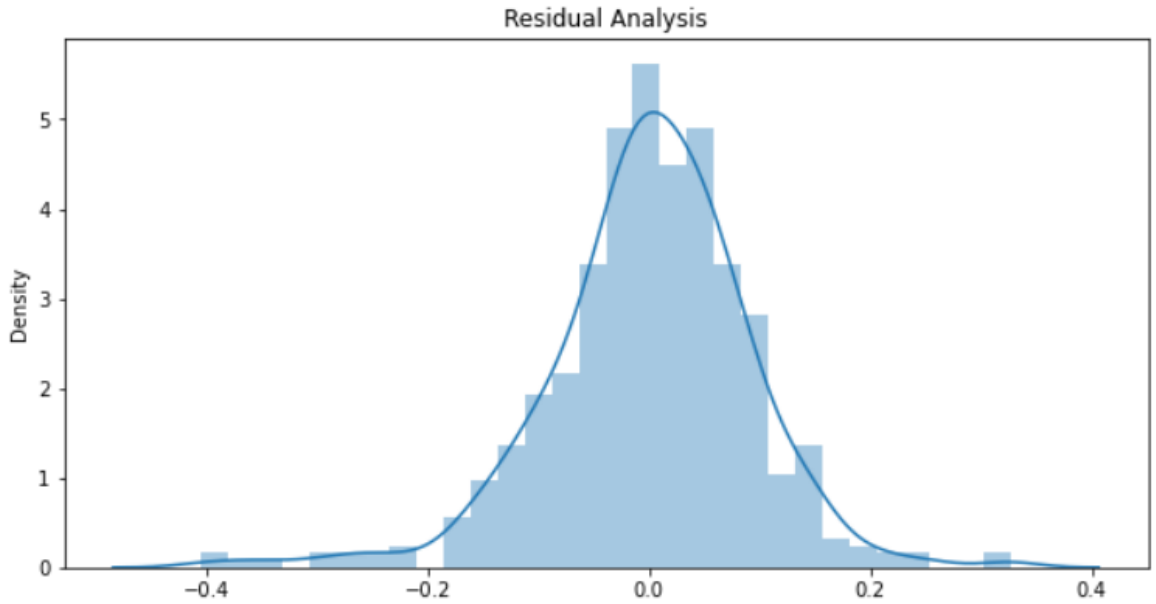
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

[Answer]: temp and atemp

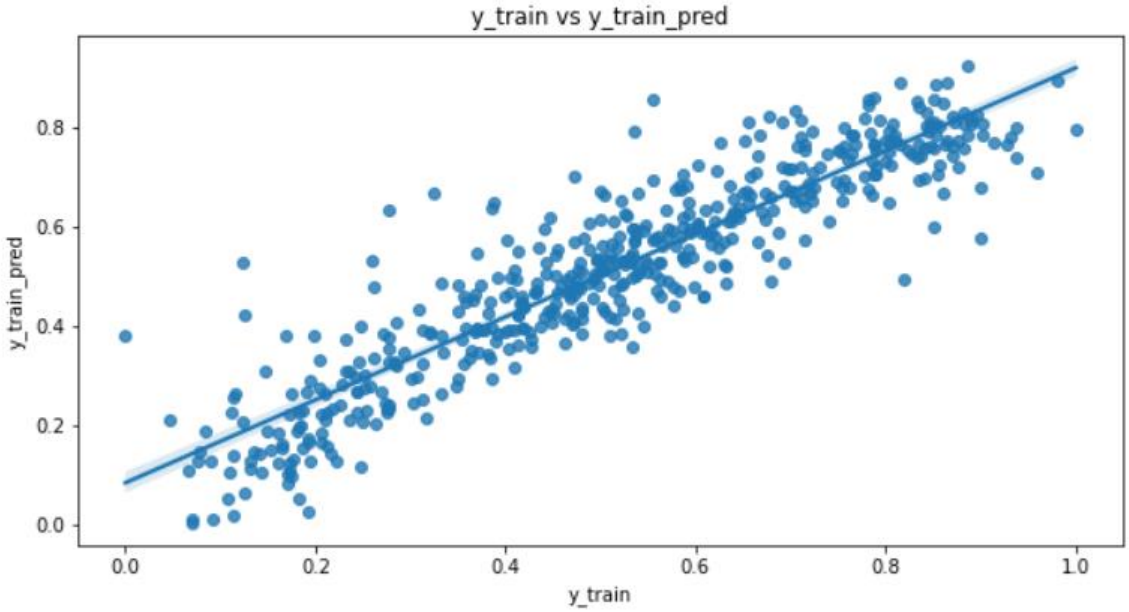
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

[Answer]: Distribution Plot on the residuals demonstrated that the error terms are normally distributed, independent and Homoscedastic

```
--------------------------------------------------------------------------------
>> Residual Analysis on Training Set
--------------------------------------------------------------------------------
```




```
--------------------------------------------------------------------------------
>> Regplot on Residual Data
--------------------------------------------------------------------------------
```



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

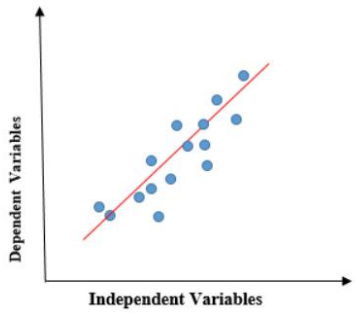
[Answer]: Feels-Like temperature (atemp), seasons (Spring and Winter) and negative weather situations

General Subjective Questions

1. Explain the linear regression algorithm in detail.

[Answer]: In ML, **linear regression algorithm** is a Supervised Learning model where the data used for training has defined labels. This method tries to fit independent predictor variables into a "cause and effect" linear relationship to predict a target value.

Linear Regression tries to fit a best possible straight line between the target and independent variables as seen in the below diagram.



For a single input variable, the linear regression is called **Simple Linear Regression** whereas if there are more than one input variable, the linear regression is called **Multiple Linear Regression**.

The line can be modelled as a linear equation => **y = b0 + b1X**

The best fit line tries to find the best possible values of b0 and b1 such that the variance can be explained. The strength of a linear regression model is best explained by R2 derived using the below formula:

$$R^2 = 1 - (RSS / TSS)$$

Where,

RSS = Residual Sum of Squares

TSS = Total Sum of Squares

2. Explain the Anscombe's quartet in detail.

[Answer]:

Anscombe's quartet consists of 4 identical datasets with similar statistical properties, but appear very different when graphed. It demonstrates the importance of graphing data before analysing it and the effect of outliers on statistical properties.

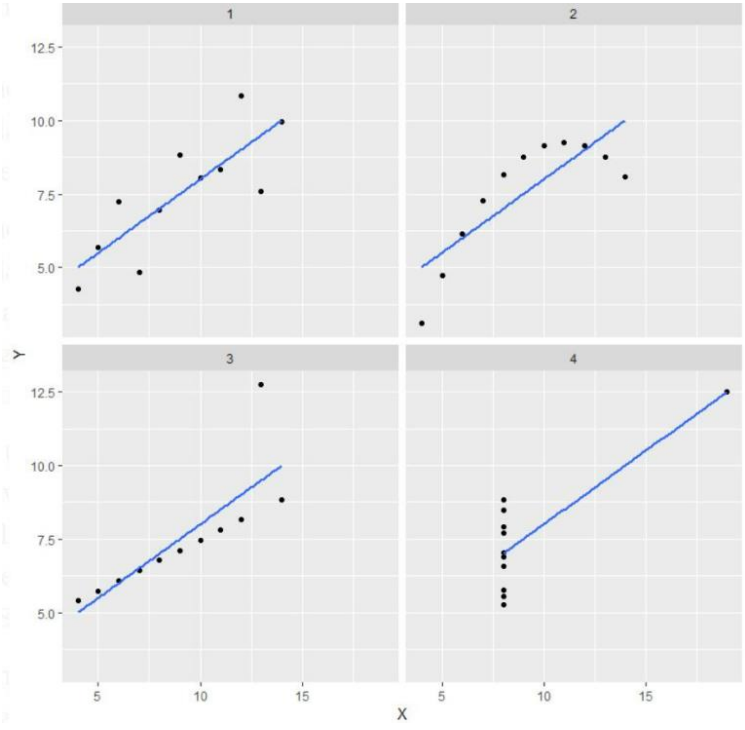
These 4 sets of data is represented below:

x1	y1	x2	y2	y3	x3	x4	y4
10	8.04	10	9.14	7.46	10	8	6.58
8	6.95	8	8.14	6.77	8	8	5.76
13	7.58	13	8.74	12.74	13	8	7.71
9	8.81	9	8.77	7.11	9	8	8.84
11	8.33	11	9.26	7.81	11	8	8.47
14	9.96	14	8.1	8.84	14	8	7.04
6	7.24	6	6.13	6.08	6	8	5.25
4	4.26	4	3.1	5.39	4	19	12.5
12	10.84	12	9.13	8.15	12	8	5.56
7	4.82	7	7.26	6.42	7	8	7.91
5	5.68	5	4.74	5.73	5	8	6.89

The mean, standard deviation, and correlation between x and y is displayed below:

set	mean(X)	sd(X)	mean(Y)	sd(Y)	corr(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

When the 4 datasets are plotted using a scatterplot, despite similar statistical properties, the graphs for the 4 datasets appear very different as seen below



3. What is Pearson's R?

[Answer]:

Pearson's correlation coefficient/Pearson's R is a measure of the linear correlation between two quantitative variables measured on an interval/ratio scale. It is the covariance of the two variables divided by the product of their standard deviations and derived using the below formula. It a popular method used for numerical variables and it assigns a value between -1 to 1 (0 for no correlation). Positive correlation indicates that if one variable increases, the other variable also increases. A Negative correlation indicates that if one variable increases, the other variable decreases.

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Certain requirements need to be considered to calculate this coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

[Answer]:

Scaling is a process in which independent variables represented by different units/ranges are converted into the same level of magnitude.

Usually collected data may features represented by different units and ranges. Since these features having different scales/magnitude may adversely impact the modelling process. For example, features represented by miles and kilometres will misrepresent the impact of these features in the modelling process specifically coefficients.

Normalized Scaling converts the data into the scale of 0 to 1. However, this may lead to some data loss/misrepresentation of outliers in the data. It is usually used in case the features are of different scales and the data distribution is unknown. One popular method used for such scenarios is Min-Max Scaling which uses the below formula:

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

Standardized Scaling converts the data into a standard normal distribution represented by the below formula. This is usually used when the feature distribution is normal/gaussian and is not impacted by outliers are the limits of the transformed data are not predefined.

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

Where sd(x) is standard deviation

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

[Answer]: VIF or Variance Inflation is represented as

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

VIF will tend to infinity if R-squared approaches 1. This means that there is perfect correlation/multicollinearity between two independent variables. In such cases, variables which are causing this needs to be dropped.

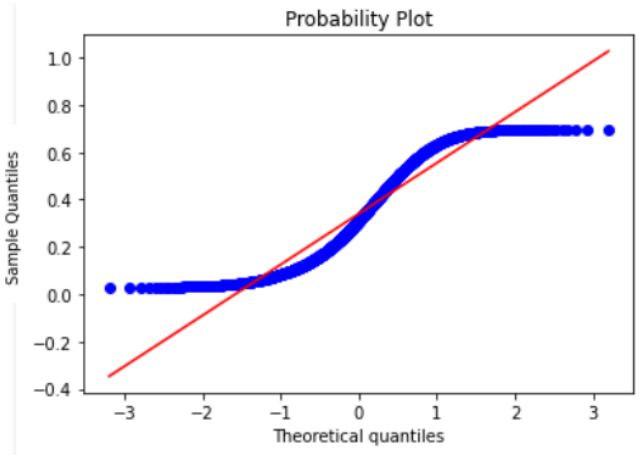
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

[Answer]: Q-Q plot (Quantile-Quantile plot) is a plot of the quantiles of two datasets. A quantile means the fraction/percentage of points below the given value i.e., 0.75 quantile is where 75% of the data fall below that value

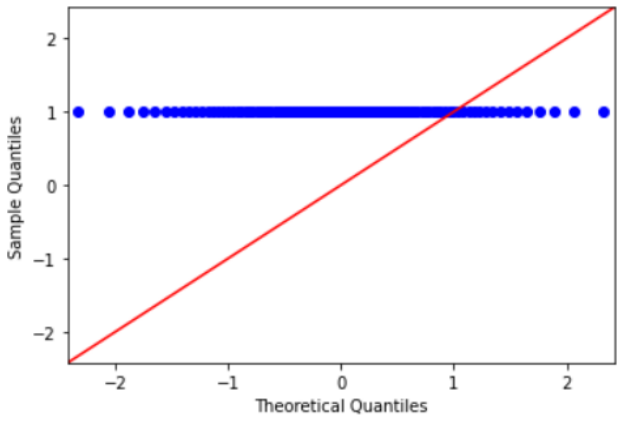
Q-Q plots can help determine if two samples are from the same population, have the same tail/distributions shape and have common location behaviour. They can be advantageous since the sample sizes need not be equal and there is no need to normalize/re-scale the datasets.

Types of Q-Q plots are:

1. Left tailed Distribution



2. Uniform Distribution



uniform distribution Q-Q plot