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?**

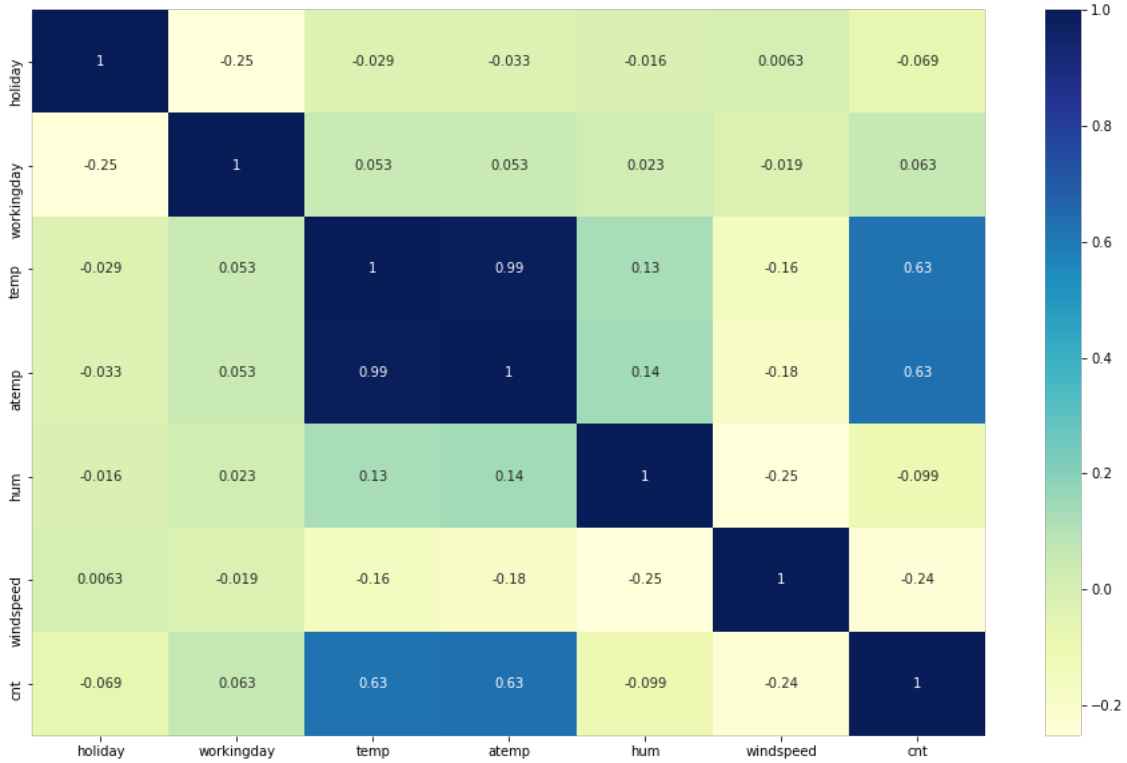
   The categorical variable in the dataset is 'Season', 'Month', 'Weekday', 'Holiday', 'Weathersit', 'Yr', 'Workingday'.

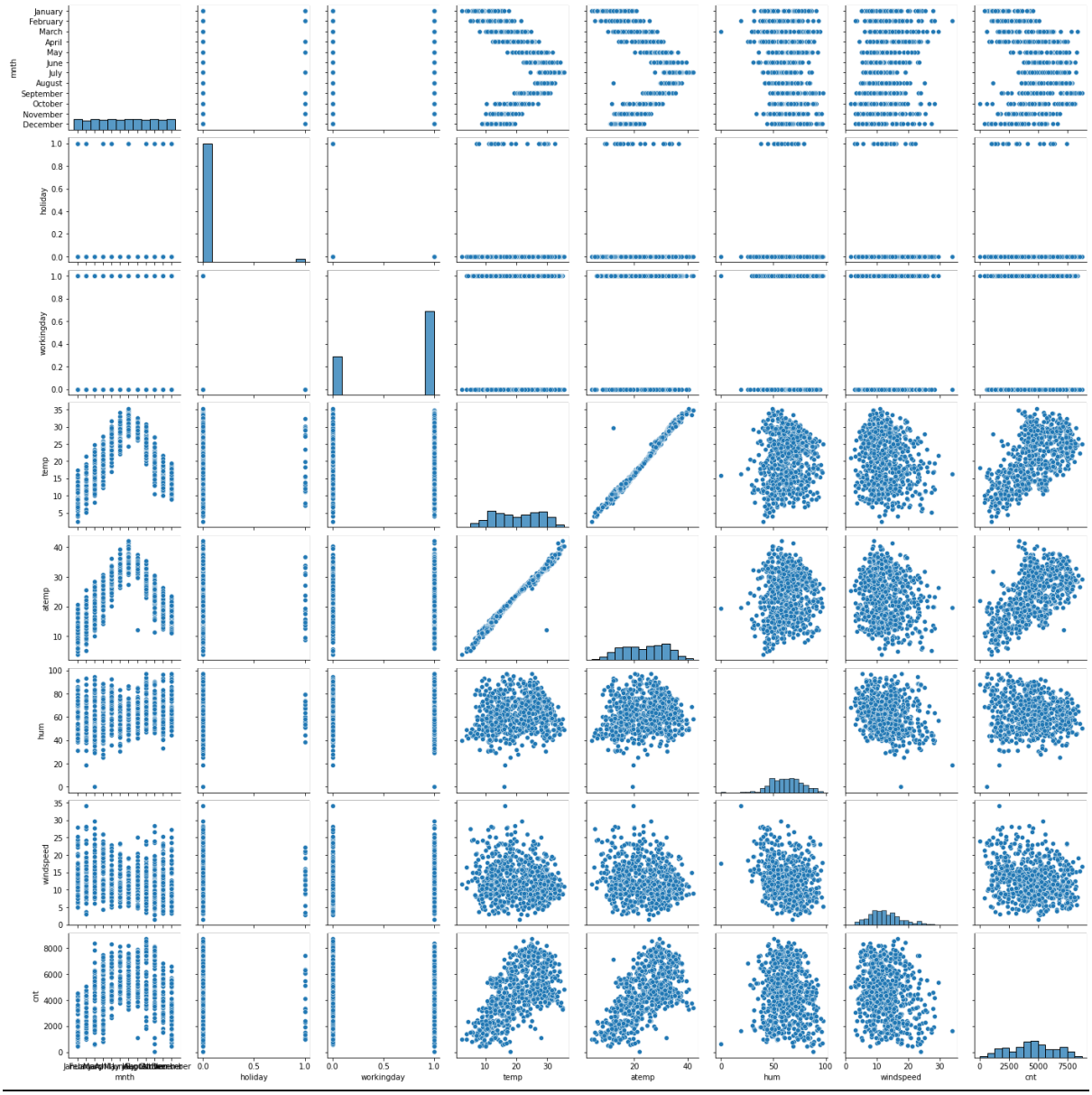
   - Season vs Cnt - The count of bike sharing is highest at Fall and Lowest at Spring
   - Year vs Cnt - The count of bike sharing is highest at 2019
   - Month vs Cnt - By Comparing overall of the month, we could see that by September there was a peak count in bike sharing
   - Holiday vs Cnt - Bike sharing were less during holidays
   - Weekday vs Cnt - During Saturday and Wednesday, There were more count of bike sharing compared to other days
   - Working day vs Cnt - There were higher bike sharing during neither weekend nor holiday
   - Weather sit vs Cnt - As we can see that there is no value for the category 'Heavy Rain' and bikers preferred to rent for ride during a clear weather condition

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

   It helps reducing the extra column created during dummy variable creation and it reduces correlation created during dummy variable.
   For example, if we have 3 categorical columns i.e., Furnished, Semi Furnished and Un Furnished and when we want to create dummy variable for these columns. If one variable is not furnished and semi furnished, then obviously its unfurnished. So, we don't need third variable for that we will use drop_first=True

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

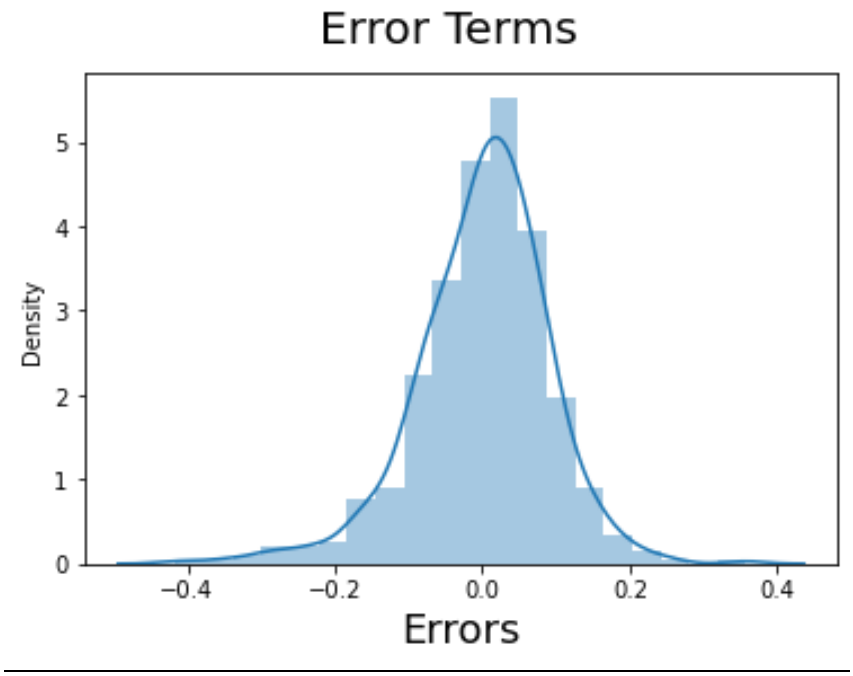




- Temp vs atemp are highly correlated ie 0.99 which is close to 1
- Temp/atemp vs count are the 2nd Highest correlated i.e 0.63
- Windspeed/Humidity/Holiday are negatively correlated when compared with count

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

Using Residual Analysis, the model should follow the normal distribution and cantered as '0' and the error term must have constant variance.



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

Features	Coefficient	Type of Coefficient
temp	0.4915	Positive
yr_2019	0.2335	Positive
weathersit_Light Snow & Rain	-0.2852	Negative

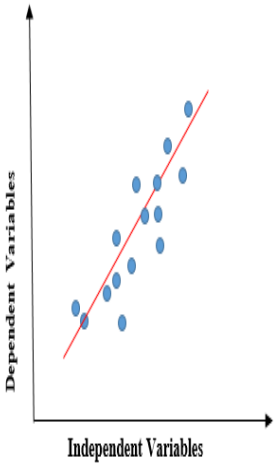
General Subjective Questions

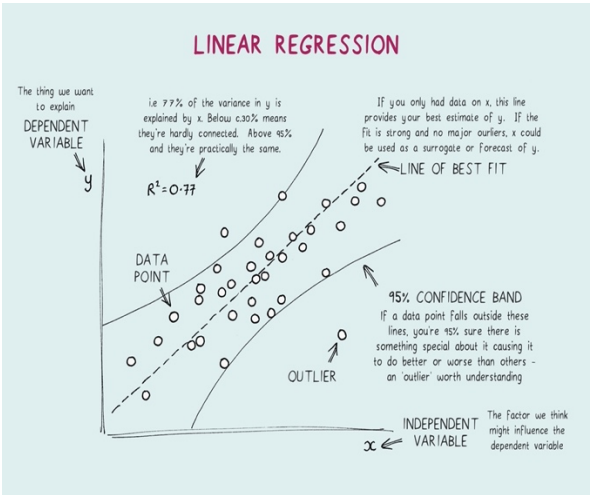
1. Explain the linear regression algorithm in detail?

It shows the linear relationship between the independent variable in X-axis And the dependent variable in Y-axis called Linear Regression.

There are two types of linear regression,

1. Simple – Single input variable
2. Multiple – Multiple input variable





The value of X increases, then the value of Y increases and the red line is referred as the best fit straight line.

To calculate best fit linear regression uses a traditional slope intercept formula, i.e

$y = mx+b \implies y = a_0+a_1x$

y= Dependent variable
X= Independent variable
a0= Intercept of the line
a1= Linear regression coefficient

2. Explain Anscombe's quartet in detail?

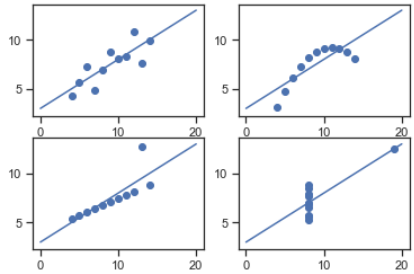
It is a modal example to demonstrate the importance of data visualization which was developed by the statistician francis Anscombe.

It comprises of 4 datasets and each dataset consist of 11 (x-y) points. To be noted that these datasets share the same mean, variance and standard deviation etc but different in graphical representation.

x1	y1	x2	y2	x3	y3	x4	y4
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

Linear regression equation, y = 0.5 x + 3

When we plot these four data across x & y co-ordinate, we get below graphical representation



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: This shows the outliers involved in the dataset which cannot be handled by linear regression model.
Dataset 4: This looks like the value of x remains constant except for one outliers which cannot be handled by linear regression model.

3. What is Pearson's R?

The Pearson's correlation coefficient Is also referred as Pearson's R. It measures linear correlation between two variable and numeric value should lies between -1.0 to +1.0

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

However Pearson's R cannot capture nonlinear relationship between two variable and cannot differentiate between dependent and independent variables.

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

Scaling helps in speeding up the calculations in an algorithm and it is a step of data pre-processing which is applied to independent variable to normalize the data within a particular range.

Collected dataset contains features highly varying in magnitude, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence it will results in incorrect modelling. To solve this issue, we have to use scaling which brings all the variables to the same level of magnitude.

Two scaling's are there

1. Normalization or Min-Max Scaling:
   It brings all the data in the range of 0 and 1.
   $$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$
2. Standardization scaling
   It brings all of the data into a standard normal distribution which has a mean zero and standard deviation one.
   $$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Note: Scaling affects the coefficients and none of the parameters like t & f statistics , P -value and r-squared. Etc. One disadvantage of normalization over standardization was it lose some information in the data especially outliers

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

If there is perfect correlation, then VIF is infinity. VIF shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which lead to 1/(1-R^2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
90An 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.

It is a graphical technique for determining if two datasets come from populations with common distributions.

It is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile means fraction of points below the given value.

If two sets come from a population with the same distribution, the points should fall approximately along the reference line.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?