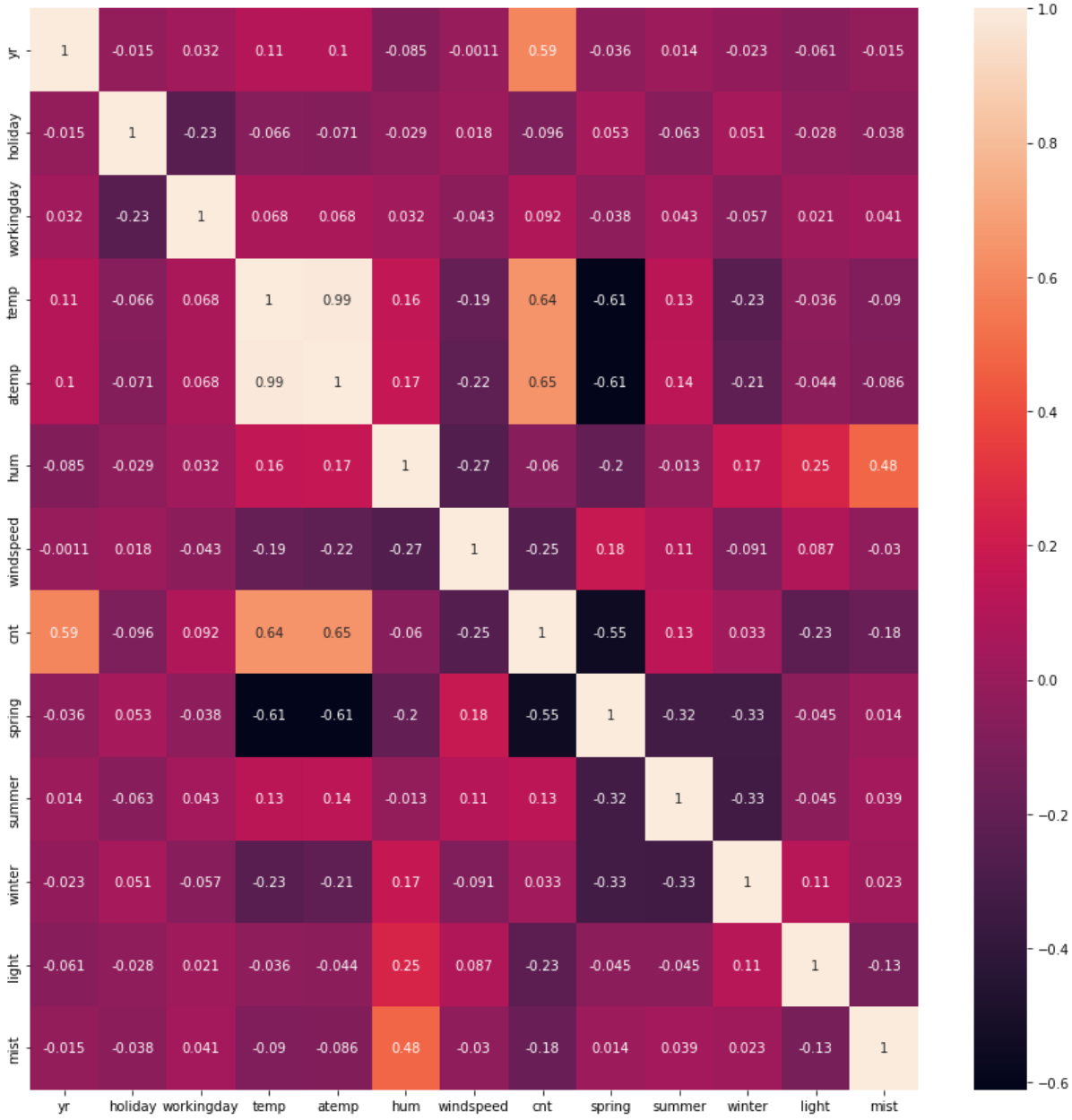
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

There are a couple of categorical variables in the dataset. Namely, season,mnth,yr,weekday, workingday and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same.



From the equation that is derived based on the model,

cnt = 0.235 x yr + 0.018 x workingday - 0.072 x holiday + 0.462 x atemp - 0.136 x windspeed - 0.103 x spring + 0.023 x summer + 0.057 x winter - 0.278 x light - 0.078 x mist

the two most important feature variables are winter and summer.

The dependent variable increases 0.278 times per one unit of winter.

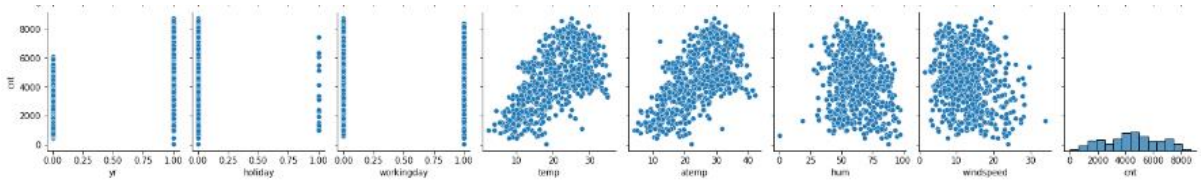
The dependent variable increases 0.057 times per one unit of summer.

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

The intension behind dummy variable is that for a categorical variable with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.
Eg: If there are 3 levels, the drop_first will drop the first column (3-1 =2 columns)

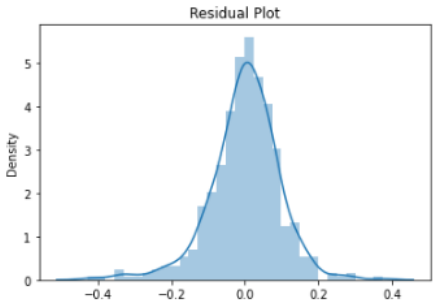
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

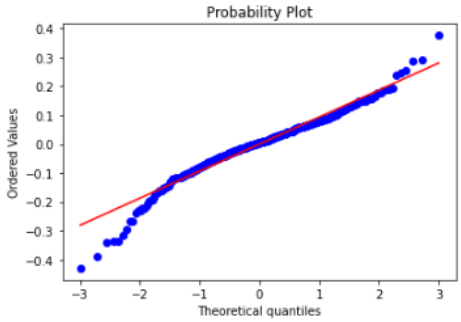
The variable ‘temp’ and ‘atemp’ have a highest correlation of 0.63 with the target variable, ‘cnt’.



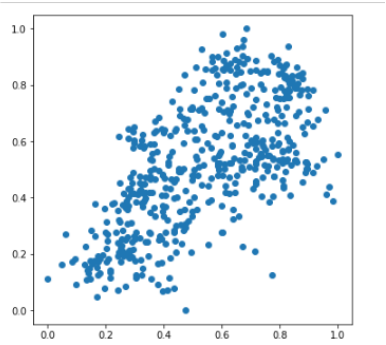
4. How did you validate the assumptions of Linear Regression after building the model on the training set?(3 marks)

i. To validate the assumption if the error terms or residuals are normally distributed, I have plotted a histogram of error terms and provided a q-q plot. This resulted in a normally distributed curve, hence proving the assumption.





ii. Assumption of linear relationship between feature and target variables are provided using a scatter plot which shows a linear relationship among the variables.
Eg: the below plot shows the relationship between the variable ‘atemp’ and target variable ‘cnt’



iii. The correlation matrix shows that there is minimal collinearity between the variables. Hence proving the assumption of no multicollinearity between features.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

According to the model we have obtained, the 3 features contributing to the demand of shared bikes are,

- atemp with a coefficient of 0.462,
- season winter with a coefficient of 0.057
- season summer with a coefficient of 0.023

	coef	std err	t	P>|t|
const	0.2307	0.030	7.654	0.000
yr	0.2349	0.009	27.336	0.000
holiday	-0.0727	0.028	-2.608	0.009
workingday	0.0181	0.009	1.923	0.055
atemp	0.4620	0.035	13.204	0.000
windspeed	-0.1357	0.026	-5.154	0.000
spring	-0.1033	0.020	-5.128	0.000
summer	0.0226	0.014	1.652	0.099
winter	0.0569	0.016	3.518	0.000
light	-0.2773	0.026	-10.742	0.000
mist	-0.0776	0.009	-8.506	0.000

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.

When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y= Dependent Variable.

x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient.

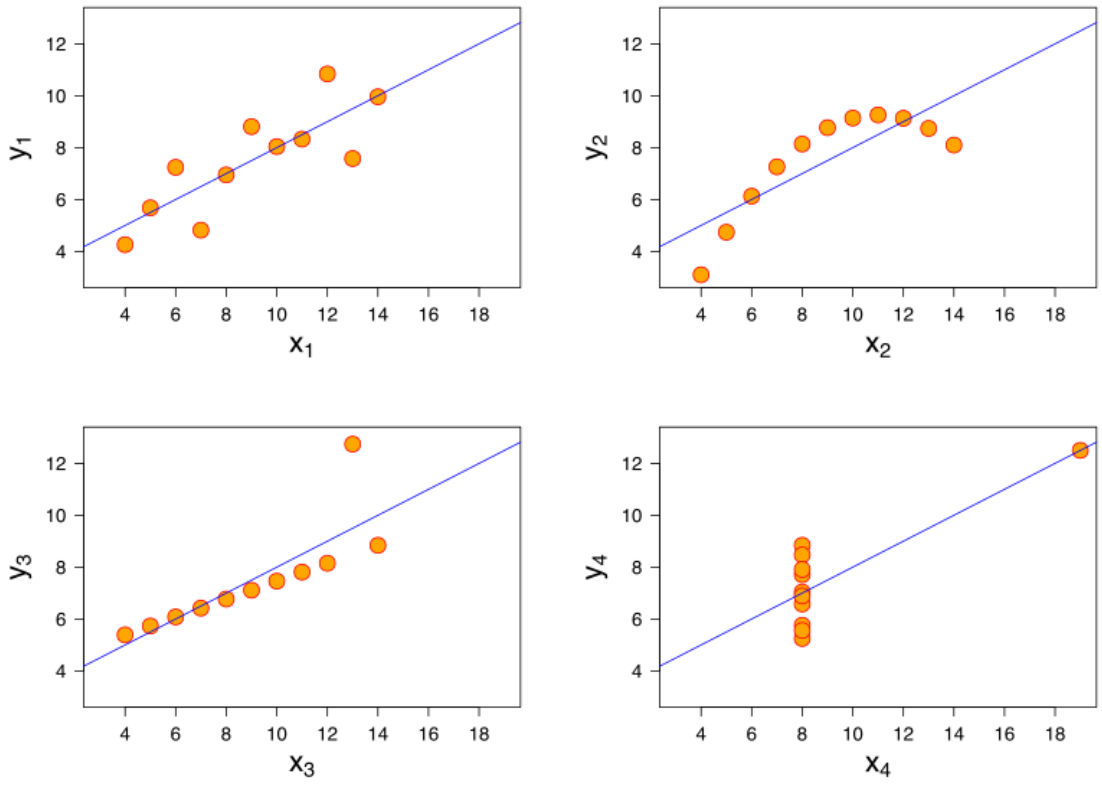
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best

fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



The four datasets 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 visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R? (3 marks)

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where,

Xi – x-variable in a sample

Yi – y-variable in a sample

Xbar – mean of x values

Ybar – mean of y values

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

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.

You can scale the features using two very popular method:

Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

sklearn.preprocessing.scale helps to implement standardization in python.

Normalising/MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

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? (3 marks)

   VIF(VarianceInflationFactor) basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

   A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

   A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distribution are similar or not. If they are quite similar you can expect the QQ plot to be more linear.

The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.