Assignment-based Subjective Questions

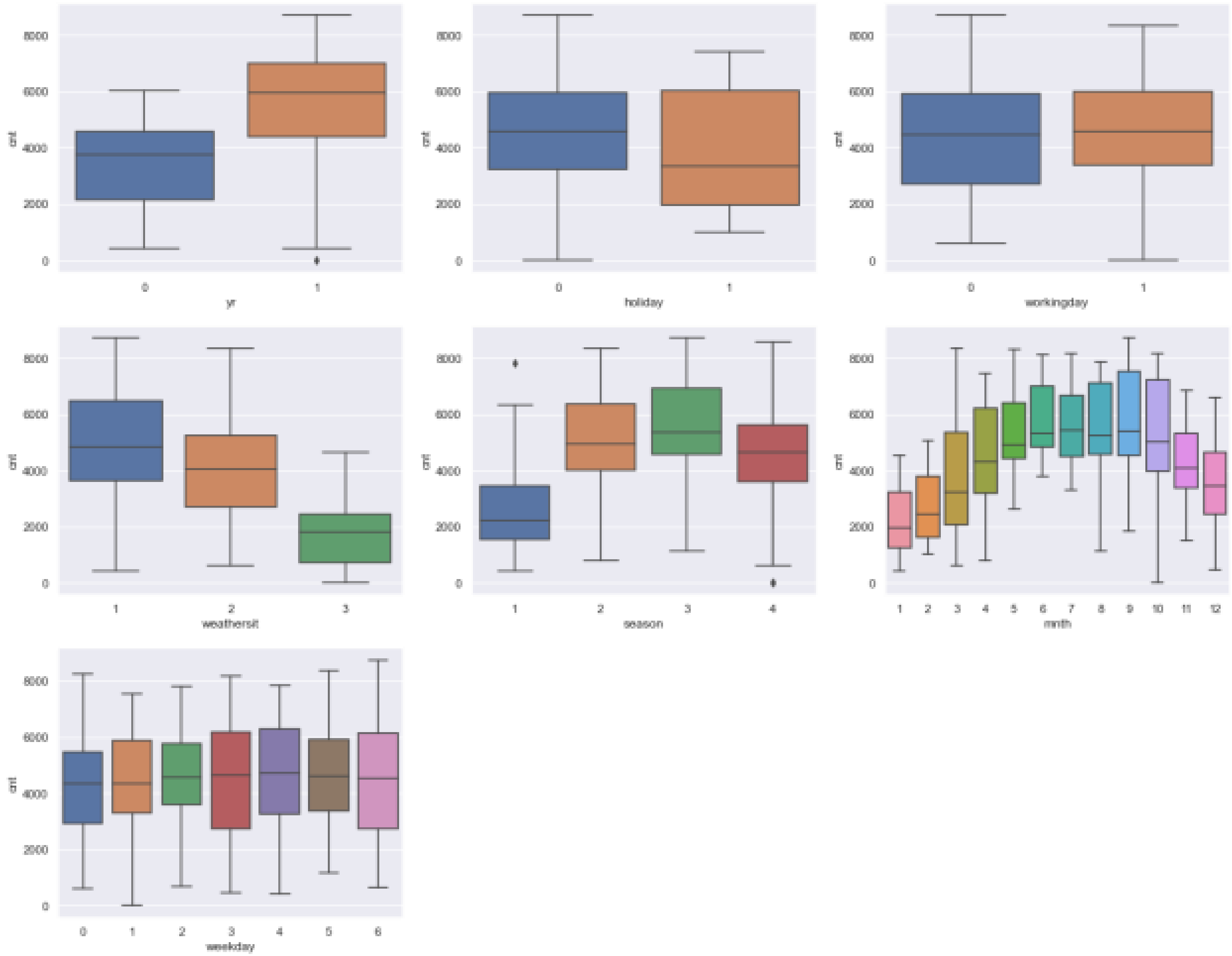
(Completed by Prayag Sanjay)

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

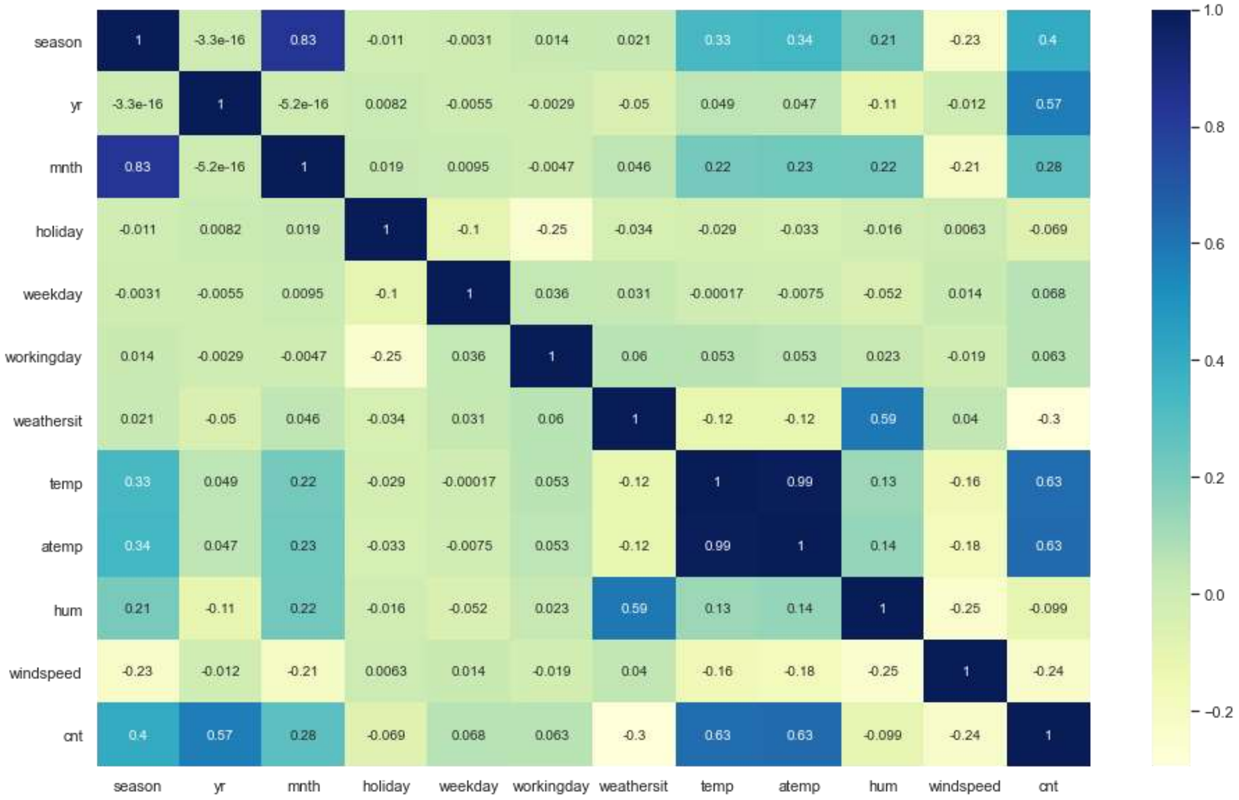
A 1 :

To understand their effect on the dependent variable (cnt), we plotted a boxplot of these categorical variables.

We got following plots



We also created a heatmap



Based upon above, we can conclude following about the effect of categorical variables on the target (cnt)

Year - There are more rental in year 2019 as compared to 2018, showing that there is scope for increase in rental on year-to-year basis.

Heatmap also shows a strong positive correlation with rental number, meaning rental number increase year on year.

Holiday – Average number of rentals are more on non-holidays as compared to holidays.

This might mean that people rent more for business purpose such as going to work.

Heatmap also shows a slight negative correlation, meaning rental number decline on holiday.

Working day – Average number of rentals working day being slightly higher than non-working day.

Heatmap also shows a slight positive correlation, meaning rental number increase on working day.

Weather Situation- Expectedly the number of rentals went down during adverse weather condition number 3 which is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

Heatmap also shows a slight negative correlation, meaning rental number decline on adverse weather days,

Season- There is a marked reduction in bike rentals in season 1 (spring), where it is relatively high in summer, fall and surprisingly high in winter as well. This shows that biker population bike all year along.

Correlation index is 0.4 which shows that season has strong bearing on the rental numbers.

Weekday- Average number of rentals seems uniform across the week with a slight increase on day 6 (Sunday

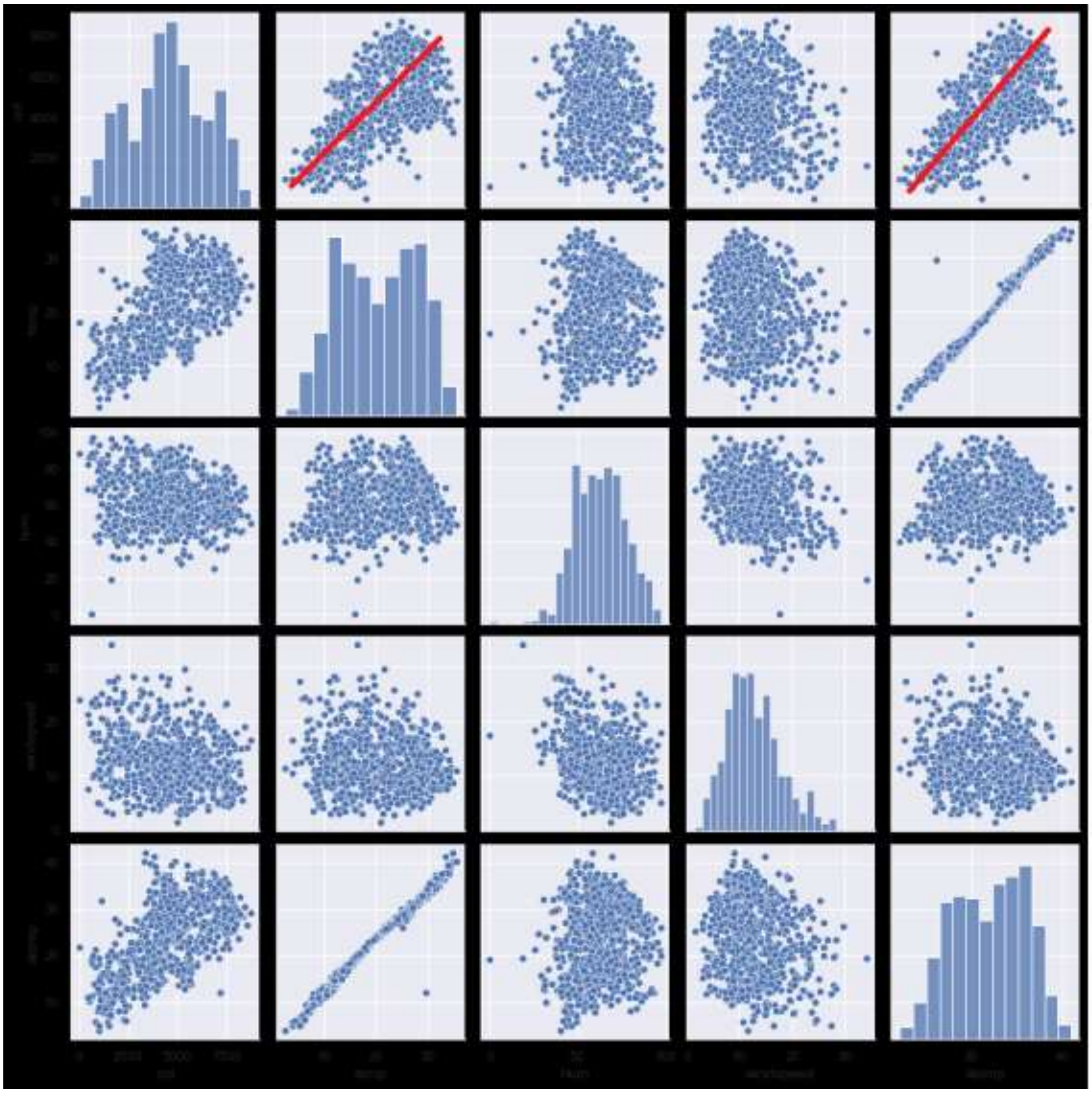
Correlation index is 0.0.68 which shows that weekday has marginal effect on rental, with Sunday being strongest day.

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

A 2 : When you have a k levels (or values) of dummy variable, and if you create k dummy variables for each of the level then, then one level (or variable) will have 0 for all other levels (variables). Including this variable then just adds redundant information and thus causes multicollinearity. Hence, we use **drop_first=True** in **get_dummies()** function to drop this variable and thus having k – 1 variable only and **eliminating multicollinearity**.

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

A 3 : Here is the pair-plot of numerical variables.



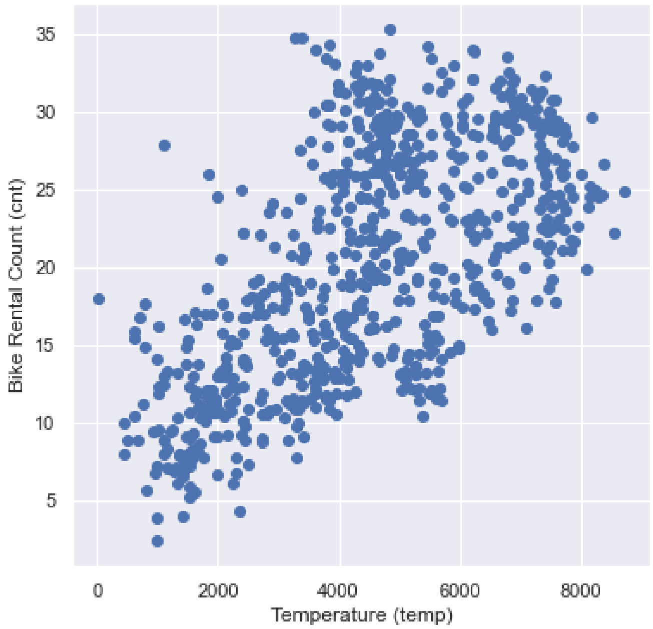
From this we can see that that **temp** and **atemp** variables show almost linear increase between them and cnt. So, we can conclude that **temp** and **atemp** **has the highest correlation** with the target variable (cnt).

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

A 4 : We validated the assumptions of the Linear Regression by plotting error terms etc. on the training set as described below.

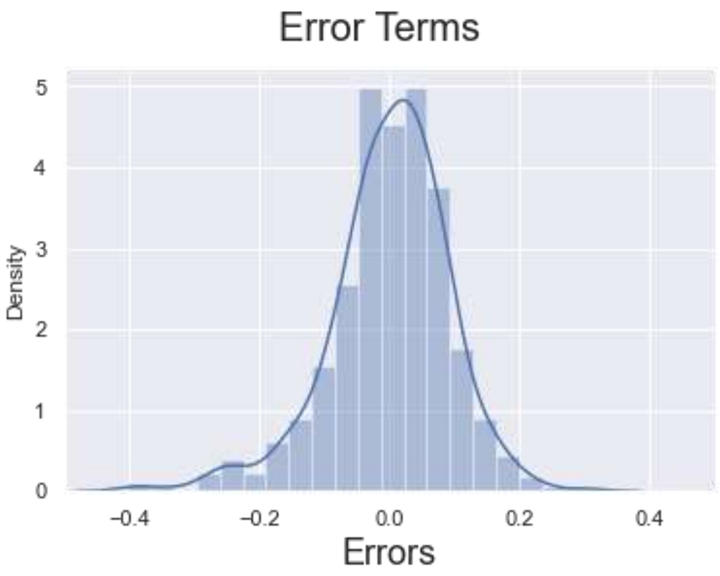
Assumption 1: **There is a linear relationship between X and Y**

To verify this, we plotted a scatter plot between one of the significant variable temp (X) and bike rentals (Y). We can clearly see a straight-line kind of linear relationship between the two, thus validating the assumption.



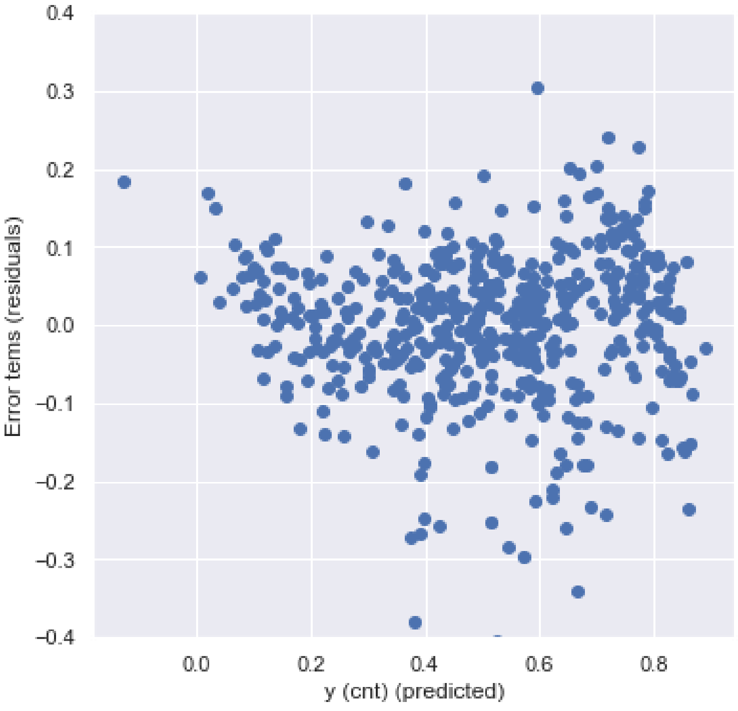
Assumption 2: Errors terms are normally distributed

To verify this, we plotted a histogram of the error terms. The following histogram shows that a normal distribution of error terms, hence proving the first assumption of linear regression.



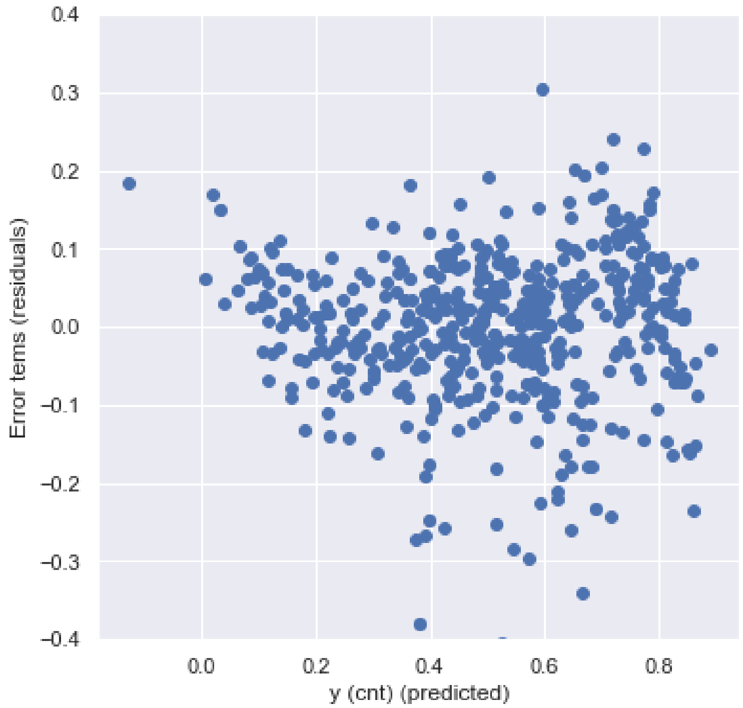
Assumption 3: Error terms are independent of each other:

To verify this, we plotted a scatter plot between the error terms and prediction of count of rental (y_pred). Below plot shows that error terms do not show any constant pattern (increase or decreasing) and are randomly distributed and thus independent of each other.



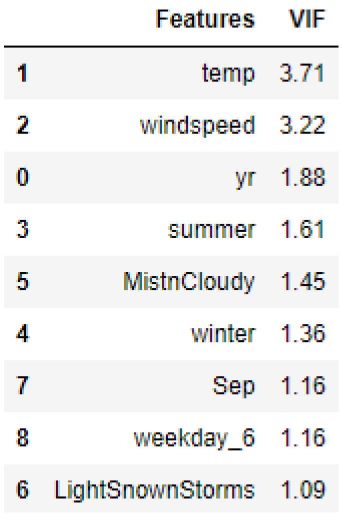
Assumption 4: Error terms have constant variance (homoscedasticity):

To verify this, we plotted a scatter plot between the error terms and prediction of count of rental (y_pred). Following plot shows that the variance does not changes as the error values change. Also, the variance does not follow any pattern as the error terms change.



Additional considerations for multiple linear regression

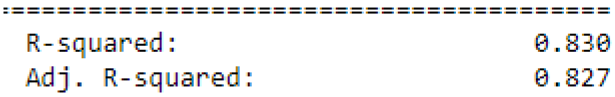
Here is VIF from the final model.



	Features	VIF
1	temp	3.71
2	windspeed	3.22
0	yr	1.88
3	summer	1.61
5	MistnCloudy	1.45
4	winter	1.36
7	Sep	1.16
8	weekday_6	1.16
6	LightSnownStorms	1.09

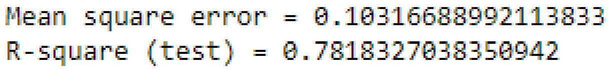
And here are R squares

Training dataset



```
==============================================
 R-squared:                         0.830
 Adj. R-squared:                    0.827
```

Test dataset



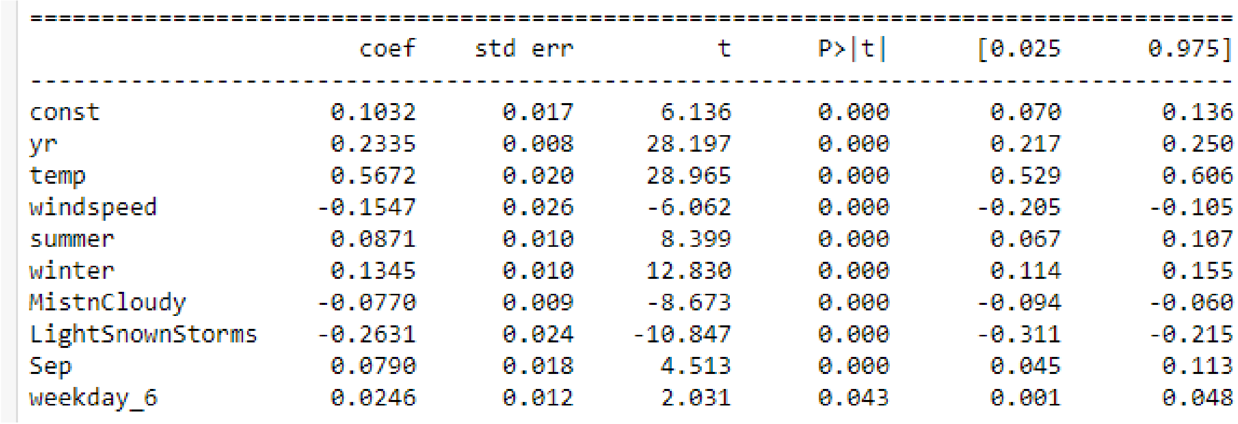
```
Mean square error = 0.10316688992113833
R-square (test) = 0.7818327038350942
```

1. Since the VIF values of independent variables is less than 5, we are confident that requirement of multicollinearity is satisfied.

2. R-square on test dataset is 0.78 while it was 0.83 on the training dataset. Since they are quite close it shows that model has generalized well on the test dataset and there is no over or under fitting.

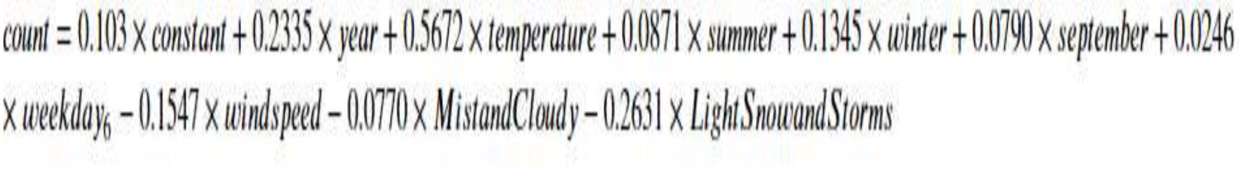
Q5 : Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A 5: Here is the metrics from the final model.



```
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const             0.1032      0.017      6.136      0.000       0.070       0.136
yr                0.2335      0.008     28.197      0.000       0.217       0.250
temp              0.5672      0.020     28.965      0.000       0.529       0.606
windspeed        -0.1547      0.026     -6.062      0.000      -0.205      -0.105
summer            0.0871      0.010      8.399      0.000       0.067       0.107
winter            0.1345      0.010     12.830      0.000       0.114       0.155
MistnCloudy      -0.0770      0.009     -8.673      0.000      -0.094      -0.060
LightSnownStorms -0.2631      0.024    -10.847      0.000      -0.311      -0.215
Sep               0.0790      0.018      4.513      0.000       0.045       0.113
weekday_6         0.0246      0.012      2.031      0.043       0.001       0.048
```

And here is the final model equation



$$count = 0.103 \times constant + 0.2335 \times year + 0.5672 \times temperature + 0.0871 \times summer + 0.1345 \times winter + 0.0790 \times september + 0.0246 \times weekday_6 - 0.1547 \times windspeed - 0.0770 \times MistandCloudy - 0.2631 \times LightSnowandStorms$$

Based on above data we can conclude that top three 3 features explaining the demand of the bikes are

- **Temperature of the day** (Coeff 0.5672) - This has got big impact of the demand. Higher the temperature higher is the demand.

- **Year** (Coeff 0.2335) - Model shows that there is natural increase in demand from one year to another. Probably with growing population and awareness for health. So, company can plan to increase production to meet this.
-
- **Weathersit (Adverse weather days)** (Coeff -0.2631) - On bad weather days particularly Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, demand is very low.

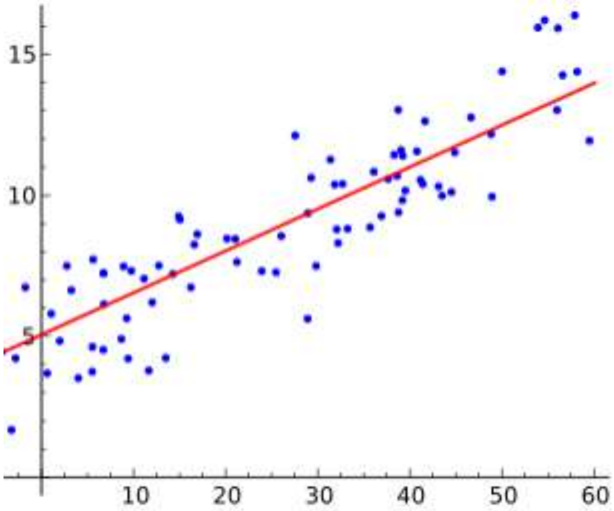
General Subjective Questions

Q 1 : Explain the Linear Regression algorithm in detail.

A 1 : Definition : It is a form of predictive modelling technique that models the relationship between a scalar dependent variable (or target variable) and the independent variables also known as predictors. More formally it is defined as a Machine Learning algorithm that finds the best fitting linear relationship between the target and the predictor variables.

In Machine Learning parlance it is one of the **supervised type of learning** algorithm as the target labelling is known.

Following figure shows a fitted line in linear regression which has one independent variable.



Types : Based on the independent variables involved there are two types

Simple Linear Regression - There is only one independent variable and one target variable.

Multiple Linear Regression – There are more than one independent variable and one target variable.

Mathematical Formulation :

Linear Regression model is expressed as a following equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

Here,

Yi is a vector of target variables

β is vector of coefficient of each predictor which measure how much variance it causes.
Xi, is vector of independent variables.
ε are error terms

Method:

Following method is commonly used in parameter estimations (coefficient)

Ordinary Least Squares Method – Here Residual sum of squares is minimized which is sum of square of error of each prediction .

Gradient descent is optimization algorithm used to minimize the cost function by iteratively moving along the direction of steepest descent.

Steps

Broad steps to do linear regression are

- Perform cleaning of data
- To exploratory data analysis of data. Use visualization
- Divide the data in test and training sets
- Do modelling on the training set.
- Iterate till variables are statistically coefficient, there is no multicollinearity.
- Check the linear regression assumptions on the training data.
- Predict using the best model on the test data.
- Check that there is under or over fitting using metrics such adjusted R-square.

Assumptions in Linear Regression :

A linear regression model will be only valid if it satisfies following assumptions

- There is a linear relationship between target and predictors
- Errors terms are normally distributed
- Error terms are independent of each other
- Error terms are independent of each other

Shortcomings of Linear Regression:

- It is sensitive to outliers .
- It models the linear relationships only .
- A few assumptions are required to make the inference as given above.
- It only shows relationship, i.e., correlation and NOT **causation .**
- Linear regression means ‘interpolation’ of data but not necessarily ‘extrapolation’.

Use cases :

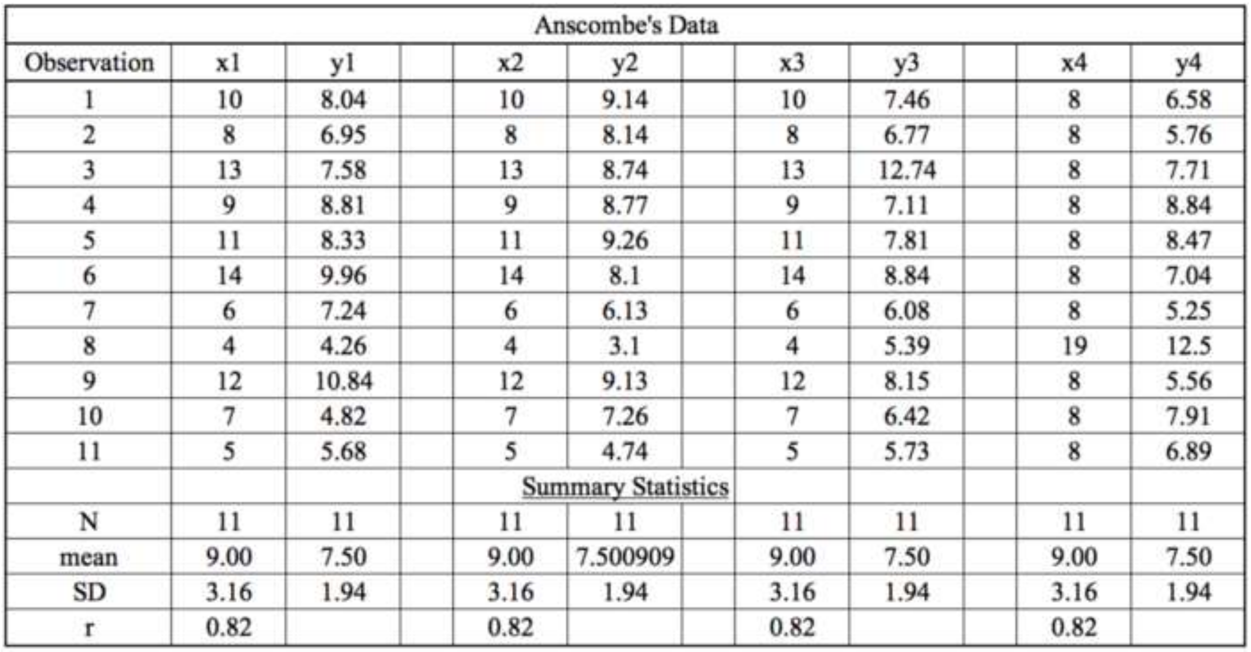
- To analyze marketing strategies effect. Also, on the pricing , sale and promotions of products.
- To assess the risk in financial domain (such as banks) or insurance sector.
- It can be used to gain insights on the consumer behaviour.

Q 2: Explain the Anscombe’s quartet in detail.

A 2: Anscombe’s Quartet is a group of four data sets which are nearly identical in simple descriptive statistics as mean , standard deviation but there is certain behavior in the dataset that fools the regression model i.e. They have very different distributions and appear differently when plotted on scatter plots.

It was built by the **statistician Francis Anscombe** to show **that it is important to of plotting the graphs before analyzing and model building.**

He used following four datasets and computed the summary statistics as below.

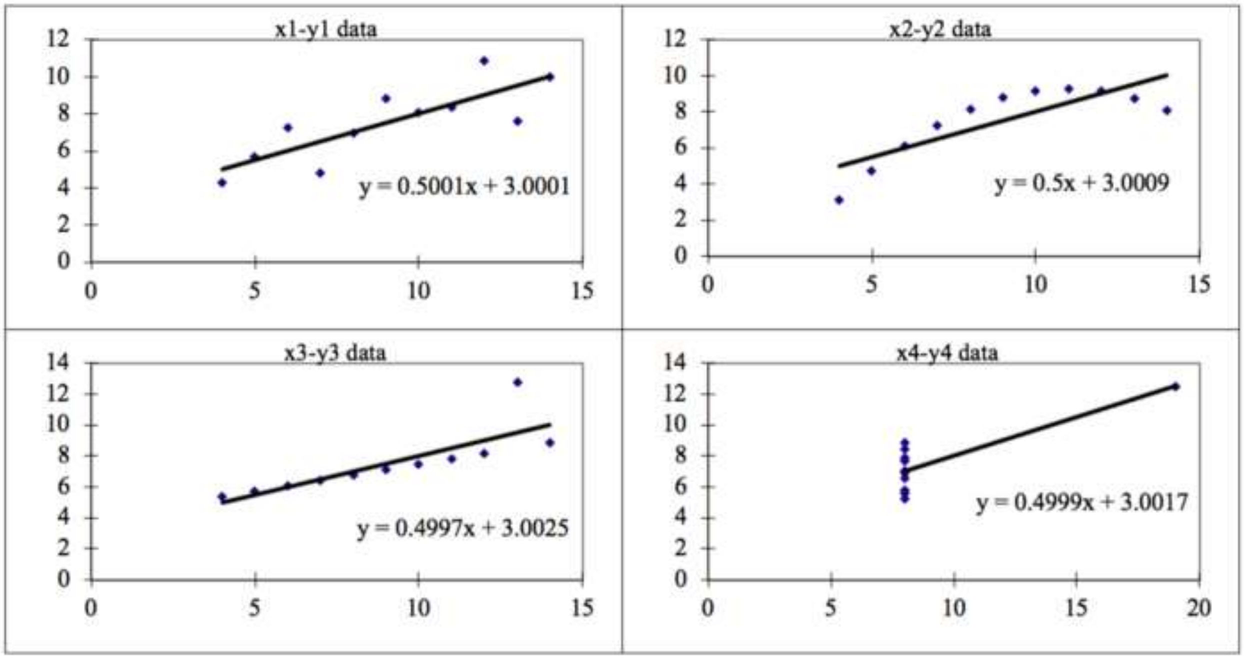


Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

One can see the summary statistics such as mean and standard deviation is same in all four cases.

But when they are plotted as scatter plots, they all generate different kind of plots which is not interpretable by linear regression.



The four datasets show following

Dataset 1: this appears to be a good fit and hence a good linear regression model.

Dataset 2: this looks like that data is non-linear and this shows that linear regression should not be used.

Dataset 3: this seem to be a good fit but there are there are outliers which cannot be handled by the best fit model.

Dataset 4: this shows that there are outliers which cannot be handled by the best fit model and thus fit is unsuitable.

With above we can see that how linear regression model was fooled by the data given. Unless above scatter plot were not plotted, we would think model fits correctly on dataset. **So, data visualization is very important part of Linear regression.**

Q 3 : What is Pearson’s R?

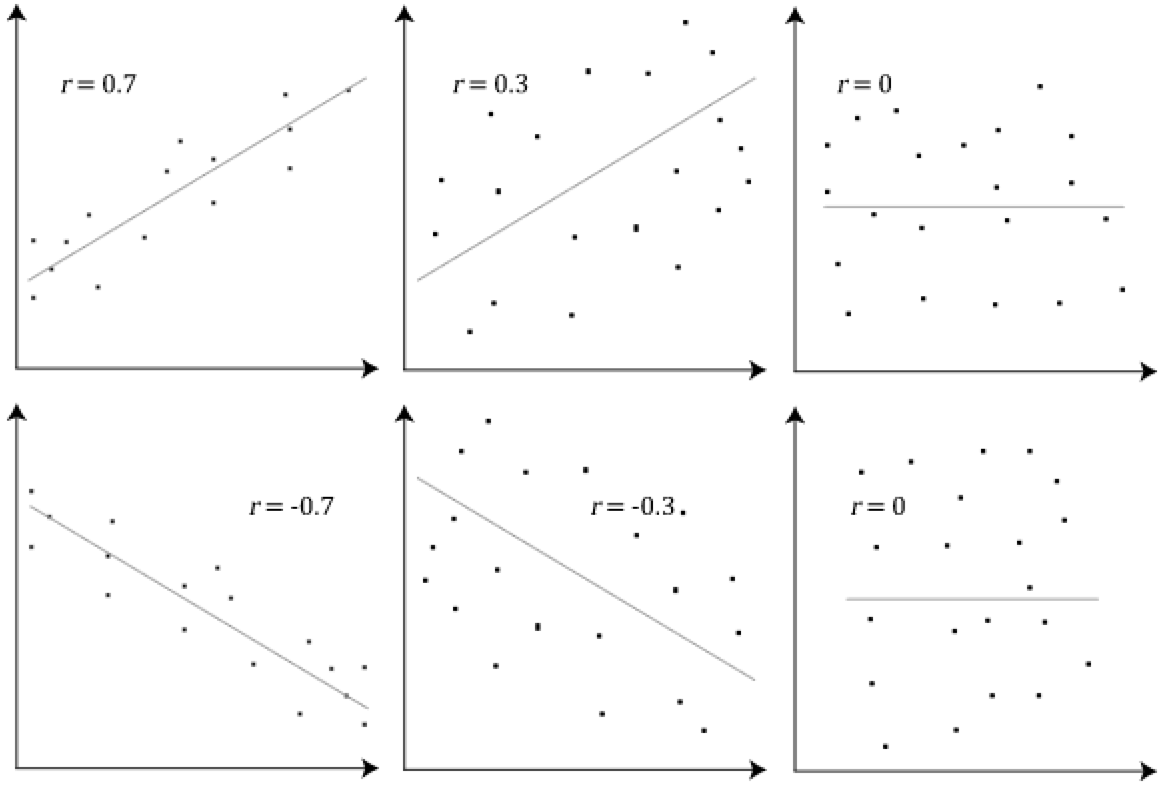
A 3 : Pearson's R or the Pearson correlation coefficient (PCC), also referred to the Pearson product-moment correlation coefficient (PPMCC), is a measure of linear correlation between two sets of data. It is measurement of the correlation between two variables , such that it always has a value between −1 and 1.

Following are interpretation of different values of Pearson's R.

R > 0 but <= 1 means both variables change in the same direction, when one increases, other increases too.(images 1 and 2)

R < 0 but >= -1 means both variables change in reverse directions. i.e., when increases other decreases. (images 4 and 5)

R = 0 means that variables are independent of each other. (images in 3rd column, 3 and 6)
This is illustrated by following images



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

R (r) =correlation coefficient
xi = =values of the x-variable in a sample
x̄ =mean of the values of the x-variable
y bar =values of the y-variable in a sample

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

A 4: Scaling is the method used in machine learning data preparation to bring the independent variables in the same range for e.g., 0 to 1 or -1 to 1.

Scaling is done because the data can have different range of values for e.g., dollars, rupees, feet , cms. The machine learning model normally behave better when the data is scaled in the same range. So scaling is done primarily for following two reasons

- It makes easier to compare and interpret the beta coefficients.
- Machine learning algorithms like gradient descent converge faster as step size are uniform each variable.

Normalized Scaling	Standardized Scaling
Values are scaled in a such a manner that they are in range 0 to 1	Values are scaled in a such a manner that they are in range -1 to 1. Basically, values are centered around mean of 0 and standard deviation of 1.
This is used when data distribution does not follow Gaussian distribution.	This is used when data distribution follows Gaussian distribution.
Formula is $x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$	Formula is $z = \frac{x_i - \mu}{\sigma}$

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

A 5 : Formula for VIF is

$$VIF_i = \frac{1}{1 - R_i^2}$$

If the VIF value is infinity for a variable, then it means that denominator (1- R ^ 2) is zero. This can be only possible if R ^ 2 is 1.

A value of R^2 being one means that

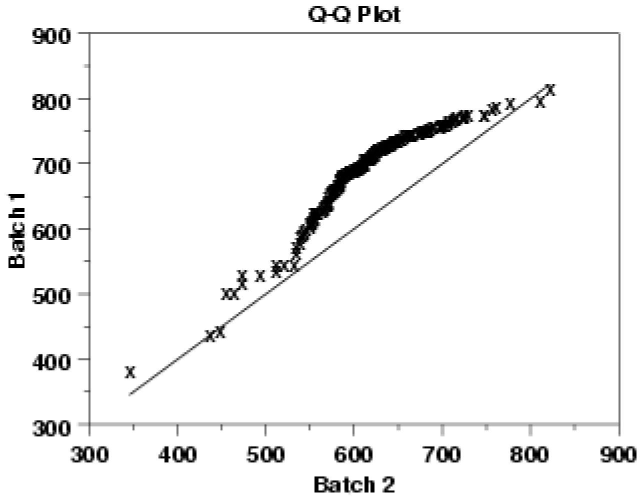
- This implies that other variables are explaining are 100% of variation of this variable.
- In other word the corresponding variable may be expressed exactly by a linear combination of other variables.
- So, this is a perfect case of a variable showing multicollinearity and such a variable should be dropped.

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

A 6 : The quantile-quantile (Q-Q) plot is which used for determining whether two data sets used have come from populations showing a common distribution.

A q-q plot is made by plotting the quantiles of the first data set against the quantiles of the second data set.

Following image shows a Q-Q plot.



Use and importance in machine learning

- This is particularly useful when training and test data are received separately or even data is given in multiple parts and to assess that they belong to same population. This is possible as it checks that

  Data has come from populations with a common distribution
  Data has common location and scale
  Data has same distributional shapes
  Data has same tail characteristics

- It is also used to validate the assumption of normally distributed residuals.