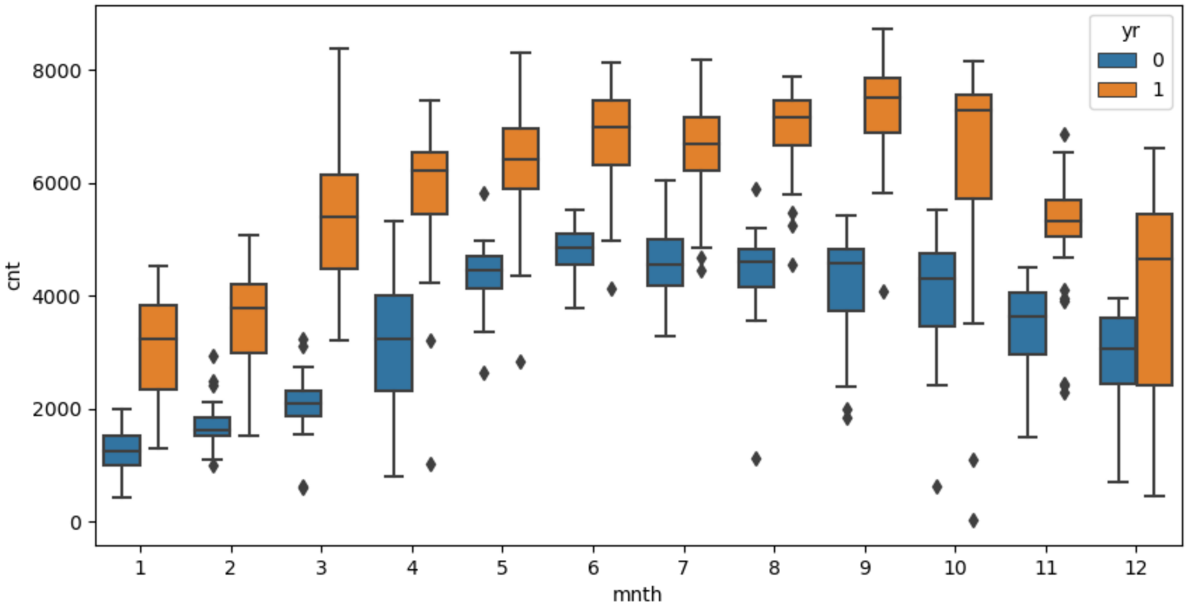
Assignment-based Subjective Questions

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



Above plot is plotted on categorical variables - mnth,yr and target variable cnt. We can clearly observe that there is impact on target variable and yr is one of the top features contri buting significantly from the final model.

Some categorical variables – dteday are not needed for the model building, this feature inf ormation is already available in other columns.

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

Encoding with drop_first=True will omit one column(first) after converting categorical variable into dummy variable, we will not lose any relevant information by doing because all points in dataset can fully explained by rest of the features (newly created dummy variables).

Each of the dummy code variables uses one degree of freedom, so n groups has n-1 degree of freedom.

If we are not encoding drop_first/drop_last = true, we need to write the code to drop the column separately. If not sone some programs will refuse to run the analysis.

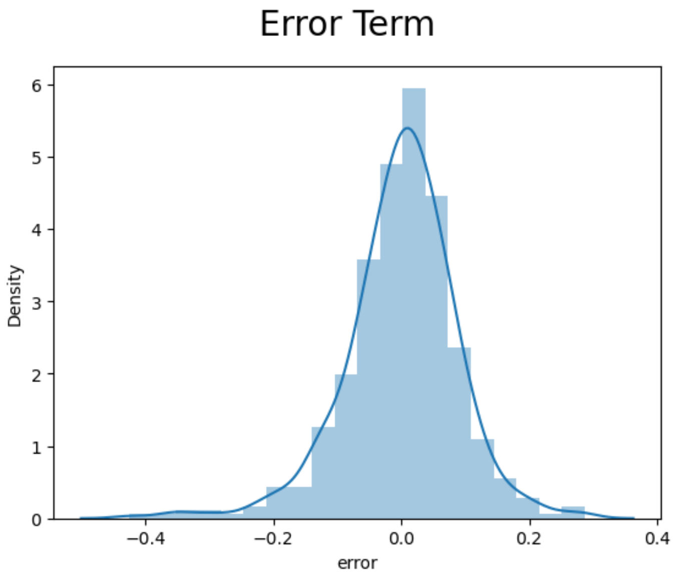
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Cnt(target variable) having highest correlation with temp and atemp -0.63 (after removing registered +casual)

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

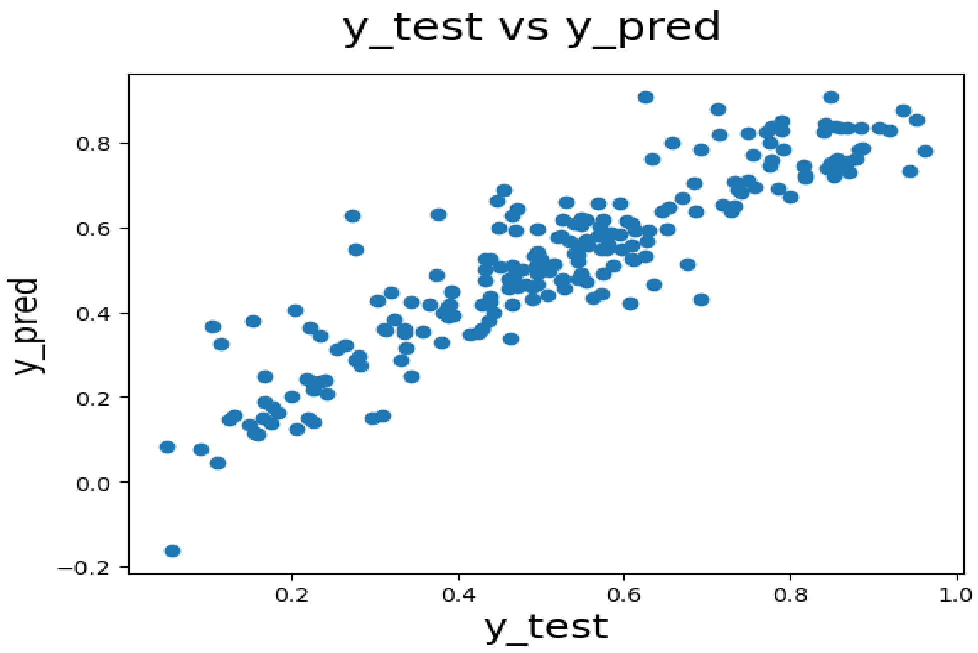
1. Checking the statistics- R square, Adj. R^2 of the model and P- values (<0.5) and VIF(<0.5) of all the features

2. By Residual Analysis: Histogram of the error term to check normality.



The graph for the error term which is normal, and mean is zero.

3.By Model evaluation: model evaluation using the test data and predicted data



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

The Top 3 Features: temp (+ve), yr_2019 (yr)(+ve), windspeed(-ve)

(writing equation n only with top 3 - we need to include rest as well)

```
cnt = 0.1765 + (0.5199 * temp) + (0.2995 * yr_2019) - (0.19 * wi
ndspeed)......
```

Positive:

1. Temp
2. Year - yr_2019
3. months - 4(april),5(may),9(sept)

Negative:

1. Weather : weather light
2. windspeed
3. hum - humidity

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is one of the machine learning algorithms-Supervised learning method. Supervised learning method uses past data with labels to build the model.

Regression model is used when the output variable to be predicted is a **Continuous Variable.**

In Linear Regression model, **relationship** between the dependent variable and independent variables is assumed to be **linear** in nature(fitting the best straight line).

Linear Regression Models are classified into two types depending on number of independent variables used to predict the target/output/dependent variable.

- Simple Linear Regression: when there is only one independent variable.
  $y = \beta 0 + \beta 1x$
- Multiple Linear Regression: When there is more than one independent variable.
  $y = \beta 0 + \beta 1x1+\ldots+ \beta nxn$

y - dependent variable
x,x1…xn - Independent variables

As Linear Regression model depends on the past dataset, we need to divide data into two parts:

- Training data – used for model to learn during modelling.
- Testing data – used to validate or evaluate the trained model for prediction -model evaluation.

The equation of the best fit regression line $y = \beta 0 + \beta 1x$ (equation of simple linear regression) can be found by minimizing the cost functions. There are two methods – Differentiation, Gradient descent method.

The strength of a linear regression model is explained by the sum of squared residuals method R^2, where R^2 = 1-(RSS/TSS)

There are some Assumptions of Linear Regression:

- linear relationship between X (independent variables) and y (dependent variable)
- Normal distribution of error terms

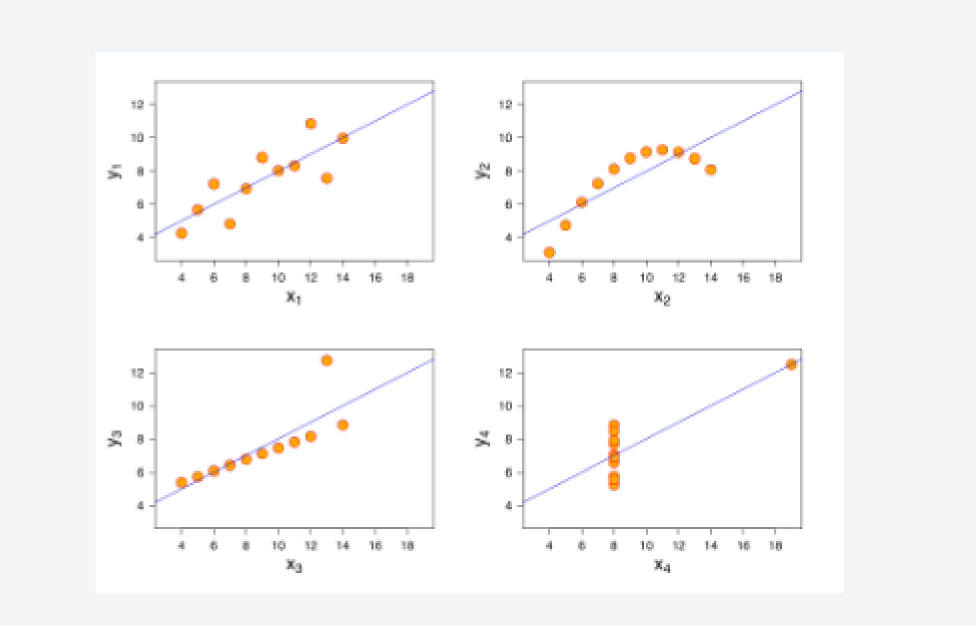
- Independence of error terms
- constant variance of error terms

Steps for Python coding:

1. Read and visualise the Dataset – check for correlation and linearity.
2. Data preparation – remove unnecessary columns and creation of dummy variable for categorial variable.
3. Split the data set into train and test sets.
4. Build Model with training dataset.
5. Residual analysis – checking for normal distribution of error terms.
6. Make Prediction - predict the target value.
7. Model evaluation - test set data vs predicted data.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a group of four dataset that are nearly identical in simple descriptive statistics (variance and mean). when you plot these data sets, they look very different from one another.



*Source of above image from Upgrad

Analysis of above plots:

X1 – fits the linear regression model
X2 – can't fit the linear regression model because the data is non linear
X3,X4- shows the outlier involves in the data set, which can't be handled by the linear regression

Anscombe's Quartet helps us to understand the importance of data visualization and how easily it is to fool a regression algorithm. So, before attempting to model any machine learning algorithm we first need to visualize the data set in order to build a well fit model.

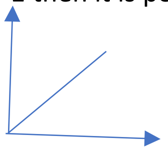
3. **What is Pearson's R? (3 marks)**

Pearson's R is most common way of measuring a linear correlation.

It is a number range between -1 to + 1 that measures the strength and direction of the relation between two variables.

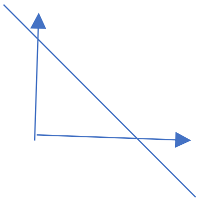
- If the value is 0 then there is no correlation between two variables.
- If the value is >0 and 1 then it is positive correlation, one variable changes the other variable changes in same direction.

  R = 1 then it is perfectly positive correlation



- If the value is <0 and -1 then it is negative correlation, one variable changes the other variable changes in opposite direction.

  R = -1 then it is perfectly negative correlation



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

Scaling: Independent variables might be in very different scales which will lead model with weird coefficients and might be difficult to interpret.

Scaling is performed to:

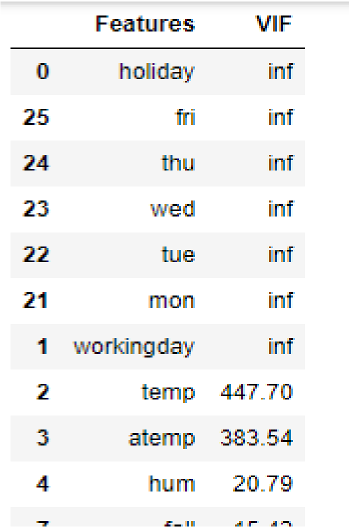
- Ease of interpretation
- Faster convergence for gradient descent method

With scaling, p-value doesn't change, model accuracy will not change, only the co-efficient will change.

Scaling Methods:

Standardization Scaling	**Normalization (Min-Max) Scaling**
1.x= (x-mean(x))/sd(x)	1.x=(x-min(x))/(max(x)-min(x))
2.Brings all the data into a standard normal distribution with mean 0 and standard deviation 1	2.Brings all the data in the range of 0 and 1

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**



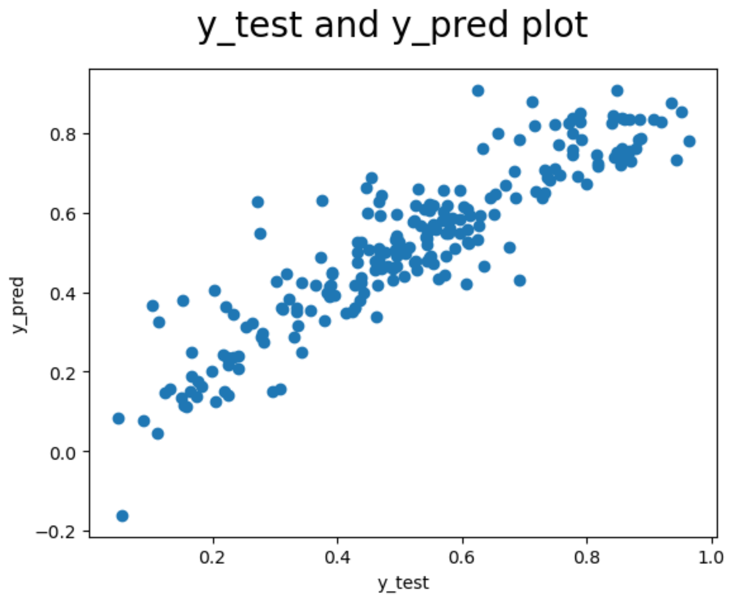
	Features	VIF
0	holiday	inf
25	fri	inf
24	thu	inf
23	wed	inf
22	tue	inf
21	mon	inf
1	workingday	inf
2	temp	447.70
3	atemp	383.54
4	hum	20.79
7	fall	15.43

Observation:

- For few features VIF value is infinite.
- VIF_i= 1/ (1- R_i^2),VIF will become infinite when R_i^2 is 1 i.e.,(1/(1-1)=1/0 = infinite)
- Which means(R^2=1) that the model can explain 100% or 1.0 of the variance which indicates the perfect fit
- If the VIF is high (>10) and the variable should be eliminated which is causing 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 or quartile-quantile is a scatter plot used to determine whether or not residuals of the model follow a normal distribution.



If plot shows rough straight line then it is normal distribution else it is not normally distributed.

If it’s not normally distributed, then there is possibility of improvement in the model.