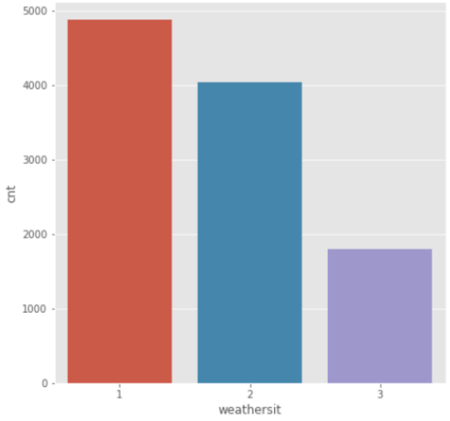
# Bike Sharing Assignment

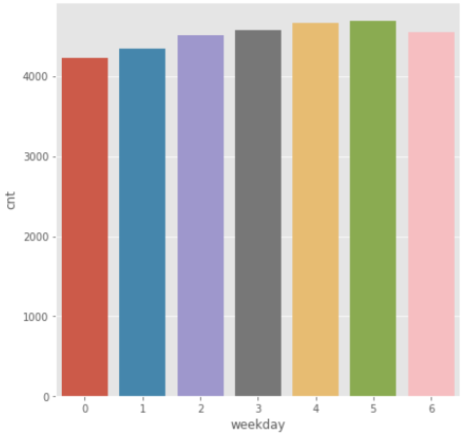
## Nikhil Vasant Khedkar ML-C33

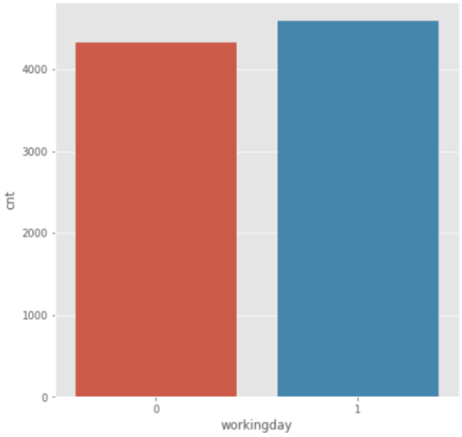
## Assignment-based Subjective Questions

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

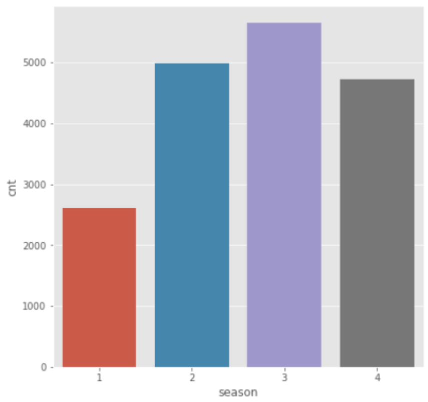


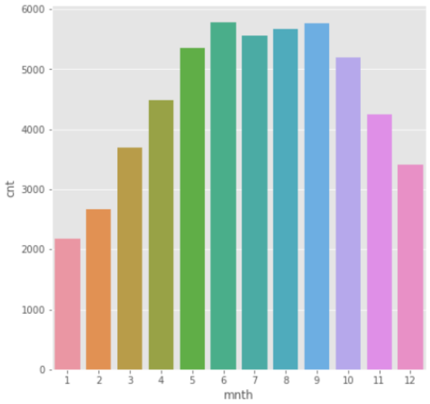
weathersit – As the weather gets worse, the average count drops. This was treated as an ordinal categorical variable, because this indicates weather conditions which get progressively worse. Dummy variables were not created.



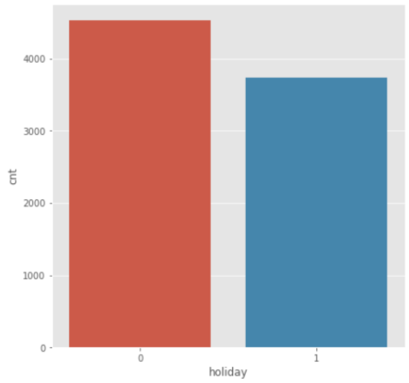


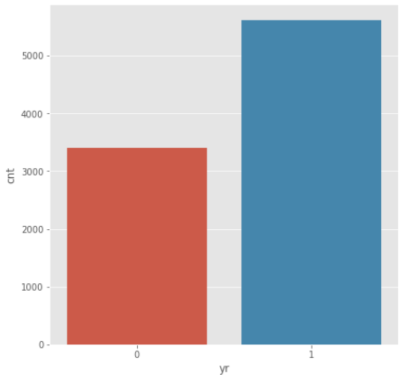
weekday and workingday does not seem to have much effect on the target variable – they were dropped based on VIF and p-values.





Season and month do influence cnt – both in similar manner. Season and month themselves have 0.83 positive correlation.





Holiday seems to have negative impact – bookings are less on holiday.

Year seems to have positive impact – as year increases, bookings increase. This can be treated as ordinal categorical variable, because years are in increasing order and popularity is increasing as years increase. So, dummy variable need not be created for yr.

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

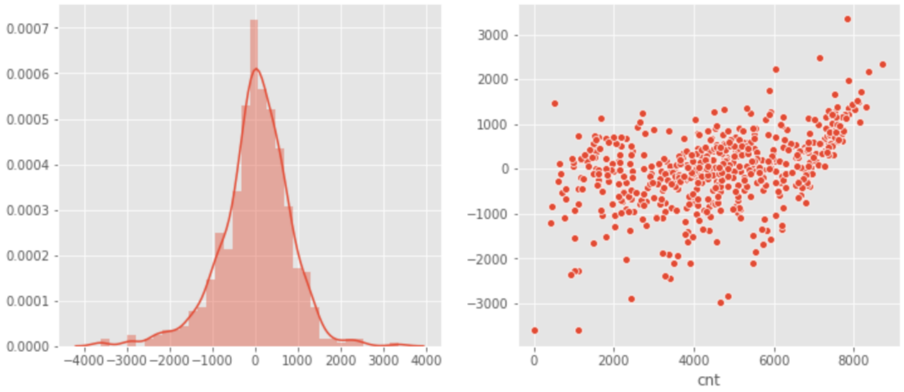
When drop_first=True is used n levels in each categorical variable are converted to n-1 columns. If it is not used, n levels are converted to n columns. This creates 1 extra un-necessary feature for each categorical variable. This feature is unnecessary because it can be represented using the other features.

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

The variable “atemp” has the highest correlation of 0.63 with target variable.

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

We got a high adjusted R-squared of 0.81 for our linear regression model which confirms linear relationship between X and y.



Plotted distplot of residuals and scatterplot of error vs actual values.

Error terms are symmetrically distributed around 0 confirm normal distribution

They seem independent with no visible pattern.

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

Features with highest coefficients make big changes in the demand. Top features are:

| **Name** | **Description** | **Coefficient** |
|---|---|---|
| temp | Daily temperature | 3688.85 |
| yr | Year- this shows popularity | 2079.63 |
| weather_3 | Thunderstorm snow and rain | -2323.50***** |

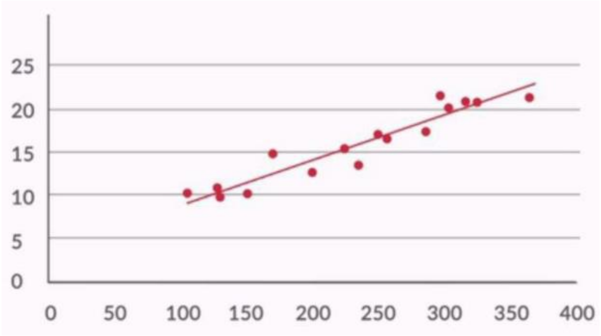
***Bad weather has a high negative impact on bookings**

## General Subjective Questions 1.

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

Linear regression algorithm works as follows:

1. It tries to fit a linear equation to the available data.



2. If X represents features and y represents the target values – linear regression tries to find the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

3. The values of beta are decided such that the error for each term with respect to the predicted value is minimized. This is called residual sum of squares. This is the cost function for linear regression. It is given by:

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

4. The above equation is used as the objective function for gradient descent optimization algorithm. Gradient decent finds values of beta which gives us the
   a. best fit coefficients
   b. with least errors,
   c. for the chosen set of features
5. The significance of the features can then be checked using p-values and VIF. Other methods such as RFE can also be used. This gives us the final list of features.
6. The goodness of fit itself can be evaluated using adjusted R-squared.

- $R2 = 1 - \frac{RSS}{TSS}$

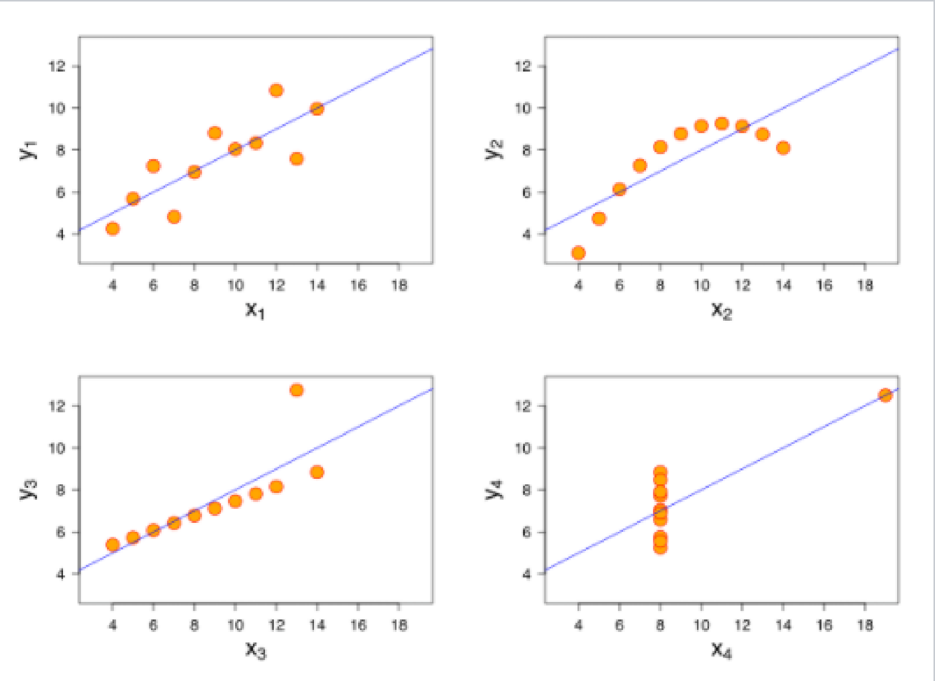
Where
RSS= Residual sum of square
TSS= Sum of errors of the data
from mean

7. Finally, error terms are analyzed to ensure that they are:
    a. Normally distributed
    b. Have zero mean
    c. Constant Variance
    d. Independent of each other
8. Other assumptions in linear regression are
    a. Target variable has a linear relationship with features.
    b. Estimators are independent of each other – no multicollinearity
    c. Estimators are measured without error

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

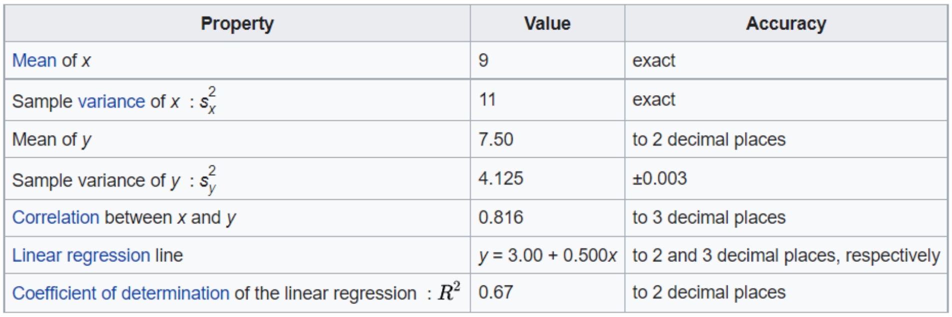
Ascomb's quartet has four data sets. They were constructed to show the importance of visualizing data, rather than relying on the descriptive statistics only to make conclusions about the data.

The following figure from wikipedia shows this:



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The descriptive statistics for this are:



| Property | Value | Accuracy |
|---|---|---|
| Mean of $x$ | 9 | exact |
| Sample variance of $x$ : $s_x^2$ | 11 | exact |
| Mean of $y$ | 7.50 | to 2 decimal places |
| Sample variance of $y$ : $s_y^2$ | 4.125 | ±0.003 |
| Correlation between $x$ and $y$ | 0.816 | to 3 decimal places |
| Linear regression line | $y = 3.00 + 0.500x$ | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression : $R^2$ | 0.67 | to 2 decimal places |

All four datasets have identical descriptive statistics. Unless we visualize the data sets we cannot know the difference between them.

They show us the importance of using EDA before directly jumping to using ML methods to perform analysis.

***Both the above diagrams are taken from wikipedia page https://en.wikipedia.org/wiki/Anscombe%27s_quartet***

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

Pearson's R is the ratio of Covariance between two variables and the product of their standard deviations:

Pearson's R (X, Y) = Cov(X,Y) / Stdev(X) * StDev(Y)

1. The value is between –1 to 1.
2. It shows positive and negative correlation.
3. Value of 1 show perfect positive correlation and –1 shows perfect negative correlation.
4. Value of 0 shows no correlation between variables
5. Unlike covariance – it is insensitive to scale of the variables. This is because covariance is divided by product of standard deviation.
6. It can be used in regression to eliminate features which have high covariance – thus removing redundant features and making our model simpler.

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

Scaling is used to normalize the range of input features of a dataset. Input features can have ranges which are very different from each other.

Eg: Age can be range 0-120, and income can be range 10,000 - 20,00,00,000. When ranges are so different, gradient descent finds it difficult to converge to the optimal solution.

Hence scaling is applied to enable gradient descent to converge faster.

| **Normalized Scaling** | **Standardized scaling** |
|---|---|
| Xnew = (X - X_min)/(X_max - X_min) | Xnew = (X - mean)/StDev |
| Also called minMax scaling.<br>Range of Xnew is [0, 1] | New distribution will have mean 0 and StDev 1 – hence called standardization |
| Shape of Xnew is same as shape of X but scaled down | Shape of Xnew is different from X because mean and stDev are different |
| Affected by outliers – because new shape is same | More robust to outliers – because new shape is different |

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

A value of infinite for VIF shows perfect correlation – that means the values of these variables are being completely explained by other variables in the feature set. Features with infinite VIF should be dropped and the effect on the model needs to be checked.

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

Q-Q plots are used to see if a dataset fits a theoretical distribution such as a normal distribution, uniform distribution etc.

In this method we plot quantiles of the dataset at various levels (0%, 5%, 10% … 100%) against the quantiles of the theoretical distribution we want to compare against. If the two plots coincide – then the distributions are the same.

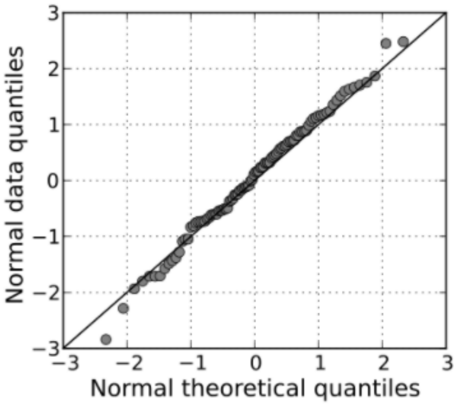


**Figure taken from wikipedia**

They can be used in linear regression to:

1. Check if error terms are normally distributed.
2. Check if test and train data have same distribution after splitting or the test-train data come from different sources.