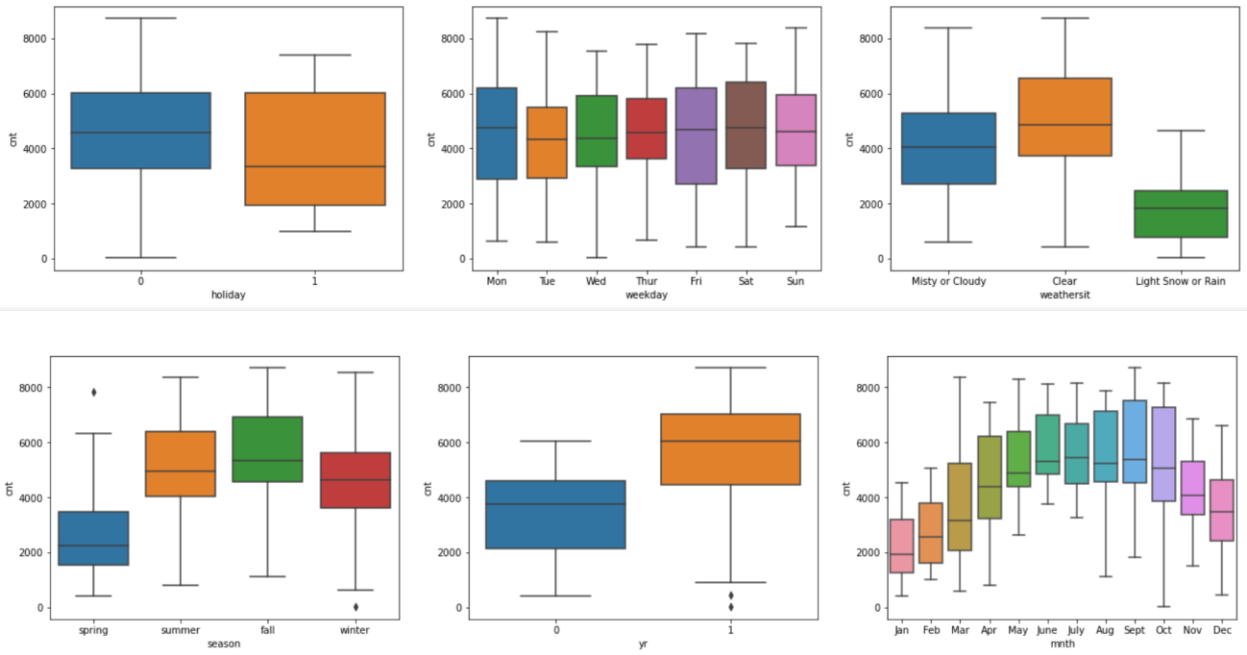
Assignment-based Subjective Questions

Question1

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

Here are the categorical variables and their effect on cnt of bike rentals,



1. Season:

While Spring, winter has relatively lower counts. Summer and fall have higher count, so Season seems an important factor

```
   ]
fall      1061129
spring     446006
summer     908121
winter     832959
```

2. Weekday:

Count seems spread across all weekdays, no specific indicator of count on weekday by the below graph. This may not be an important feature

```
Tue      438039
Wed      444917
Thur     467314
```

```
Fri     467838
Mon     471691
Sat     475195
Sun     483221
```

3. Weathersit

Clear weather has higher count of rentals. As rainfall/snow increases, count seems to decrease, hence it may be a good feature to consider

```
Clear                 2226768
Light Snow or Rain      37246
Misty or Cloudy        984201
```

4. Holiday:

Comparatively lower count on holidays observed.

```
1      78435
0    3169780
```

5. Year

Every year the count of rentals is set to increase as per the below trend, it may be an important feature in model

```
      yr
0    1229827
1    2018388
Name: cnt, dtype: int64
```

6. Month

Mid months have highest count of rentals, it is decreasing in beginning and end of year.

```
Jan     131557
Feb     141709
Dec     209287
Mar     212788
Nov     254831
Apr     264184
Oct     313698
May     331686
```

```
July     344948
Sept     345991
June     346342
Aug      351194
```

Question2

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

To convert categorical variables to numeric variables we create dummy variables. To represent n categories, we require n-1 dummy variables, however by using get_dummies we create n variables. Hence a drop_first is required to drop the first variable.

Let us check with an example:

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	1	0	0	0	0	0	0
Tuesday	0	1	0	0	0	0	0
Wednesday	0	0	1	0	0	0	0
Thursday	0	0	0	1	0	0	0
Friday	0	0	0	0	1	0	0
Saturday	0	0	0	0	0	1	0
Sunday	0	0	0	0	0	0	1

Drop_first = True will do the below:

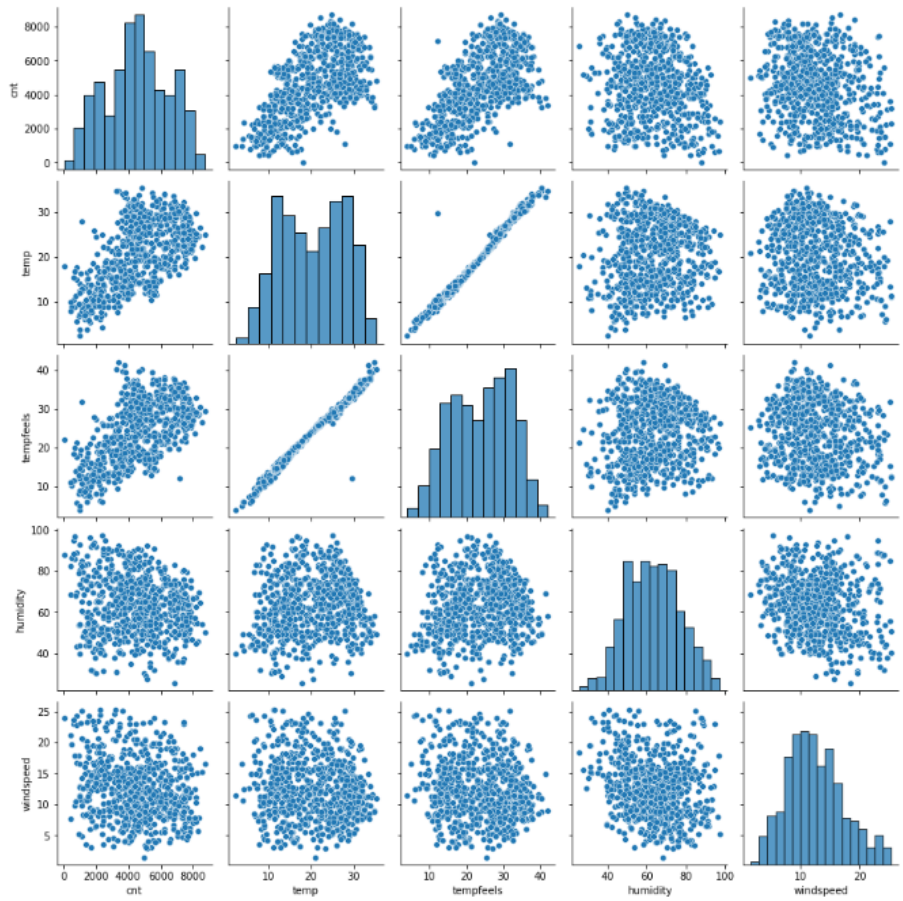
The first category can be represented as a 0 in all other categories like below:

	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	0	0	0	0	0	0
Tuesday	1	0	0	0	0	0
Wednesday	0	1	0	0	0	0
Thursday	0	0	1	0	0	0
Friday	0	0	0	1	0	0
Saturday	0	0	0	0	1	0
Sunday	0	0	0	0	0	1

Question3

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

```
In [2810]: ### pairplot to check if linear regression is a fit
           sns.pairplot(bikedata, vars=['cnt','temp','tempfeels','humidity','windspeed'])
           plt.show()
```



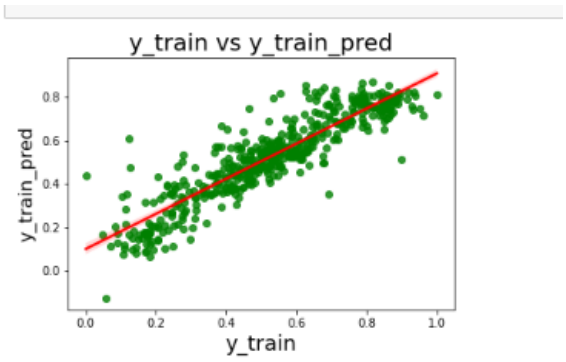
The target variable: cnt has most correlation with temp and atemp(temp feels) among all numeric variables.

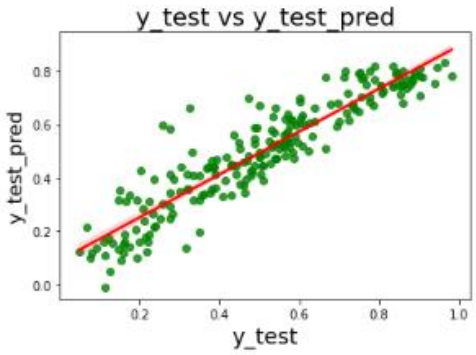
Question4

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

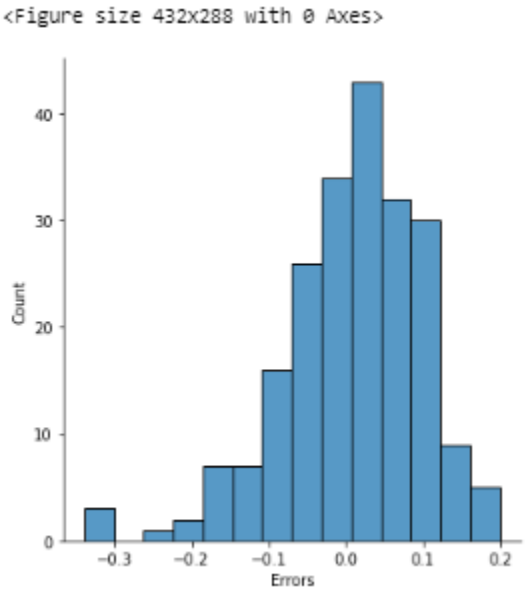
Assumptions taken into account:

1. Linear relation between predicted values and actual values





2. The errors have a mean zero and are normally distributed.



3. There is no multicollinearity between variables (VIF < 5)

	Features	VIF
2	temp	2.86
0	yr	1.96
5	Misty or Cloudy	1.45
6	July	1.39
3	spring	1.21
8	Sept	1.19
7	Oct	1.15
4	Light Snow or Rain	1.05
1	holiday	1.04

4. R2 of train and test model is almost same.

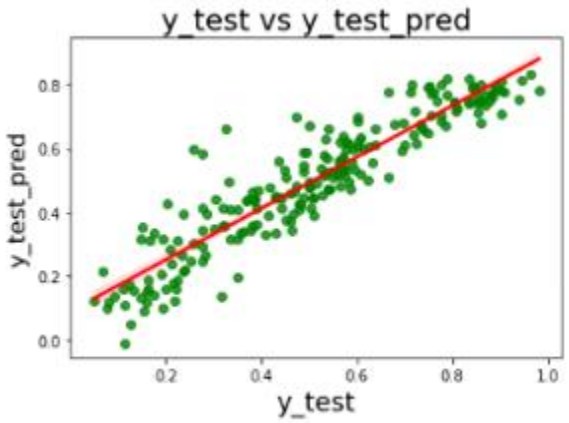
```
R2 on test dataset is: 0.8489
```

```
Adjusted R2 on test dataset is: 0.8423

R2 on train dataset is 0.809
Adjusted R2 on train dataset is: 0.805
```

5. Homoscedasticity of Residuals

There was no trend observed on distribution of residuals



6. Min Max scaling was applied to all numeric variables:
   num_vars=['temp','humidity','windspeed','cnt']
7. All categorical features were converted to numeric for linear regression using get_dummies

Question5

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

Final Model is as below:

cnt = 0.2610 + 0.2376 x yr - 0.0945 x holiday + 0.3743 x temp - 0.1525 x spring - 0.3010 x Light Snow or Rain - 0.0742 x Misty or Cloudy - 0.0557 x July + 0.0640 x Oct + 0.0585 X Sept¶

The 3 most significant feature

- Temp (Positive)
- Year (Positive)
- Light Snow or Rain (Negative)

General Subjective Questions

Question 1

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

Linear regression: It is an algorithm to describe linear relation between a dependent variable(y) with one or more independent variable(X)

Example of linear regression model equation:

$$y_i=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\varepsilon_i$$

Linear Regression is a **supervised machine learning algorithm** where the predicted output is continuous and has a constant slope.

The algorithm tries to find the best fit line.

Simple linear regression – When there is one dependent variable and one independent variable

Multiple linear regression - When there is one dependent variable and multiple independent variable.

The overall goal of regression is to examine two things:

- does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

The coeffecients is determined by one of the below methods:

OLS: Ordinary Least Square procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seek to minimize.

Gradient Descent: This procedure works by starting with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved as decided for threshold.

Assumptions:

1. The residuals/errors are normally distributed

2. There is a linear relation between actual and predicted values
3. There is no multicollinearity (no 2 independent variables are correlated)
4. Homoscedasticity of Residuals (variance of errors have no trend)

Question 2

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

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

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	

The statistical information for these four data sets is **approximately similar**.

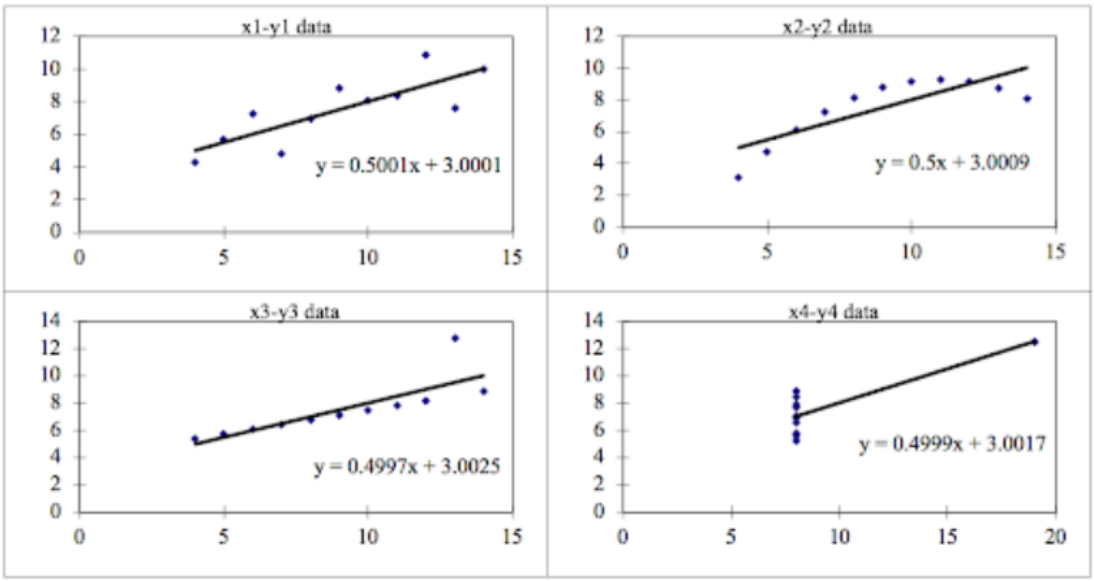
However, when these models are plotted on a scatter plot, each data set **generates a different kind of plot that isn't interpretable by any regression algorithm**

Data Set 1: fits the linear regression

Data Set 2: cannot fit the linear regression model because the data is non-linear

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model



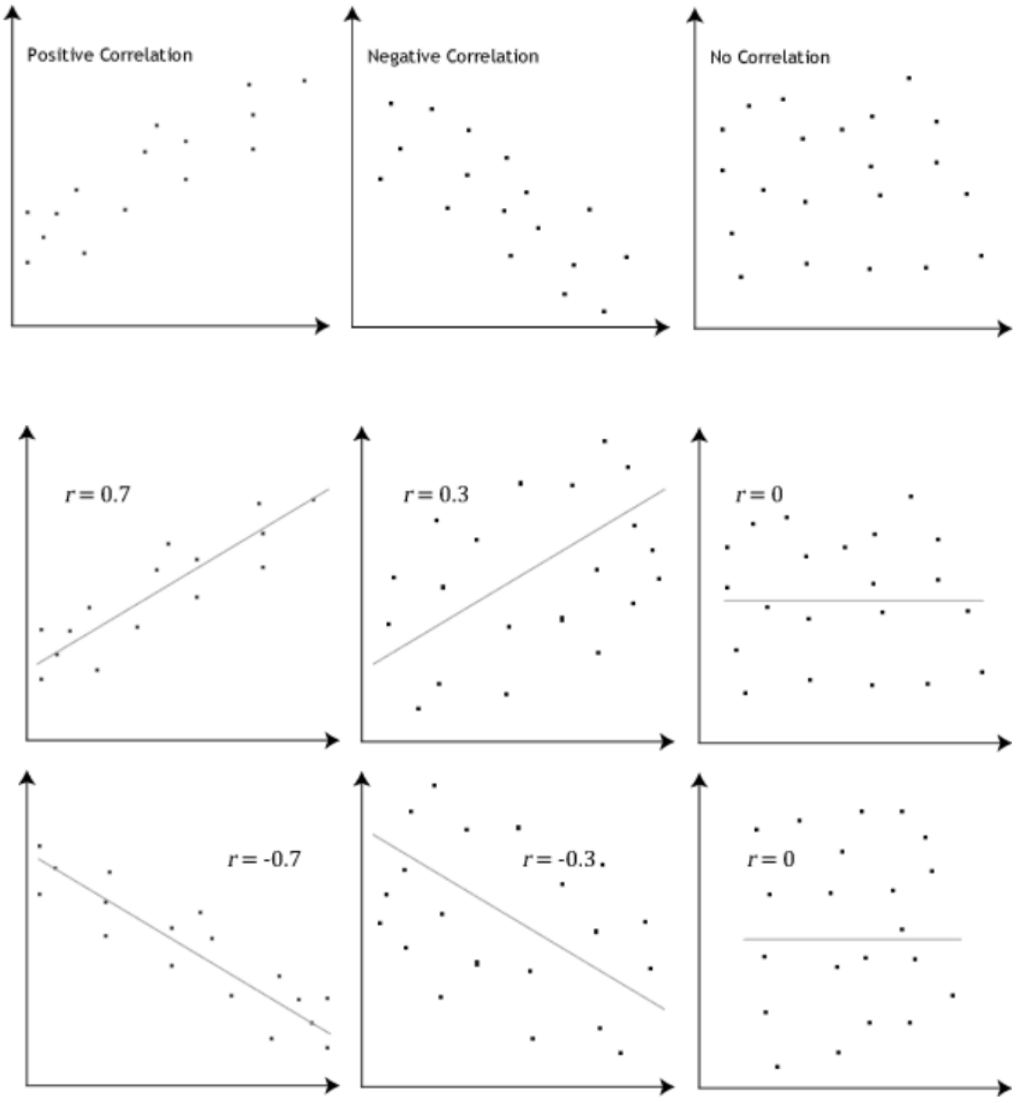
Anscombe's quartet helps us to understand **the importance of data visualization** and how easy it is to fool a regression algorithm

Question 3

What is Pearson's R? (3 marks)

The Pearson R is a measure of the strength of a linear association between two variables and is denoted by *r*. A Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit

It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



Question 4

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

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. It also helps in computation of complex algorithms. If we do not scale the units of features wont be taken into account and may result in wrong modelling.

Both Normalization and Standardization are scaling techniques

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

This scales the range to [0, 1]

Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

Standardization does not get affected by outliers because there is no predefined range of transformed features.

Differences:

a. Min and Max is used in Normalization while mean and std deviation is used in standardization
b. Normalization is used when features are of different scale, Standardization is used when we want to ensure zero mean.
c. All normalized values will always be between 0 and 1
d. Normalization is affected by outliers
e. We can use MinMaxScaler from scikit learn for normalization and StandardScaler for standardization.

Question 5

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

An 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).

As per the formula of VIF

VIF=1/(1-RSquare)

If (1-Rsquare) =0 , VIF =infninte

For (1- Rsquare) to be zero, RSquare=1 which means there is a perfect correlation between 2 independent variables.

In the bike data casestudy, we removed casual, registered as they were summing to be count. If we would not have removed it, we would have removed it during VIF calculation.

Question 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 (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

It confirms the normality of residuals in the linear regression model. If predictions are plotted across the 45 degree line we can confirm that error terms/residuals have normal distributions

Example from bikedata casestudy: