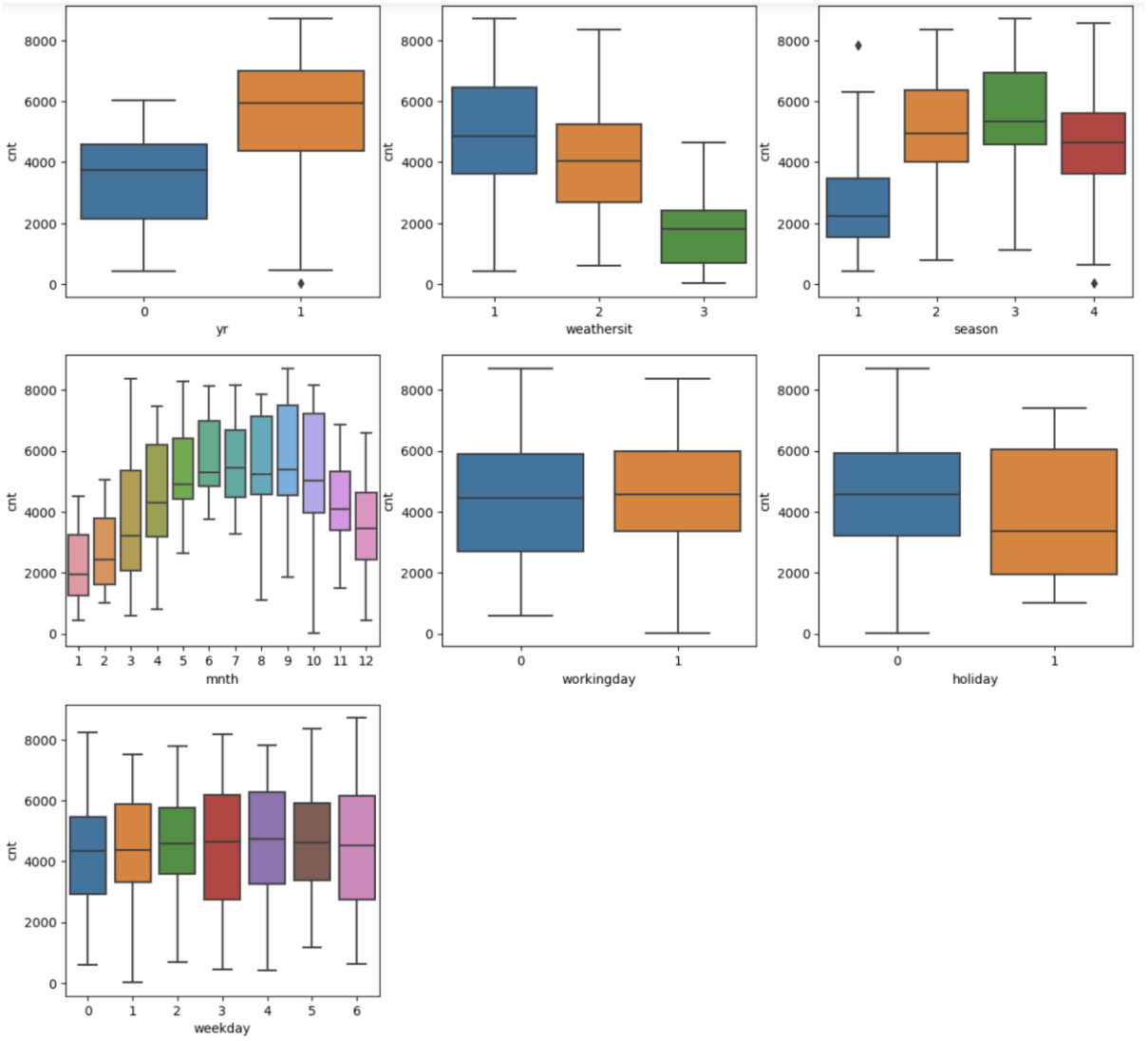
Assignment-based Subjective Questions

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

The effect of categorical variables on the dependent variable can be derived from the following:

(i) Data Visualization using box plots (before building the linear regression model):

- ***Year***: The number of bike rentals has increased from 2018 to 2019. Therefore, it can be expected to increase with each passing year.
- ***Weather situation****:* Bike rentals increase when the skies are clear.
- ***Season/Month****:* Bike rentals increase during the months of summer and fall.
- ***Working day/ holiday/ weekday****:* Bike rentals on a holiday are less but high on working days.



(ii) The coefficient of the predictors in the linear model:

The equation of the linear model built is:

$$cnt = 0.070264 + 0.539040 \times temp - 0.297362 \times lightsnow + 0.231263 \times yr + 0.146153 \times winter + 0.123609 \times sep - 0.100177 \times holiday + 0.095578 \times summer - 0.080930 \times mist + 0.058455 \times aug$$

- We can see that besides temperature (numerical variable), the significance of the categorical variables are more on the dependent variable in comparison to the numerical variables.
- The positive co-efficient of variables such as yr, seasons (summer, winter), months (August, September) indicate that these factors have a positive effect on the bike rentals. That is, the presence of these factors increases the bike demands.
- The negative co-efficient of variables such as holiday, weather situation like light snow and mist indicate that these factors have a negative effect on the bike rentals. That is, the presence of these factors reduce the bike demands.
- Light snow has the highest negative coefficient which indicates that people do not rent as many bikes when it snows lightly. This data can be used by the company to take measures such as providing reduced rates or precautionary gear to ride in light snow to increase number of rentals.

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

It is important to use *drop_first=True* during dummy variable creation in order to:

(i) *avoid multicollinearity* which makes it easier to understand the effects of individual variables in the model.

(ii) *avoid redundancy of variables* and obtain a simpler model. This allows the adjusted r-squared value to be maintained well.

Here is how it is achieved:

The key idea behind dummy encoding is that for a variable with, say, 'N' levels, you create 'N-1' new indicator variables for each of these levels.

Let's take the example of the encoding performed on weather situation in the assignment.

The variable *weathersit* had three values in the dataset, namely, clear, mist, light_snow. By using *drop_first=True*, clear has been dropped. In other words, clear is considered as the reference feature based on which the coefficients for the other weather features are to be interpreted once the model is built.

By not using drop_first = True, the variables will be encoded as shown below:

clear	mist	light_snow
0	1	0
0	1	0
1	0	0
1	0	0
1	0	0

From the definition of these weather situations, we can almost assume that clear and light_snow/mist can have a strong negative correlation.
Let's say, we obtain a final model as

Bike_demand = 0.087 + 3 x Clear – 2.7 x Light_snow

Now it is difficult to understand if the presence of clear weather situation is impacting the bike demand by three times or the absence of it; likewise for light_snow.

Here, instead of creating a new variable for ***clear***, *it* can also be interpreted as mist = 0, light_snow = 0.

In order to create n-1 dummy variables for n levels, we use the following syntax:

dummy_var = pd.get_dummies(df[c1], drop_first = True)

Now, let's say we obtain a final model as

Since *ReviewSession_No* can be easily predicted from *ReviewSession_Yes*, this situation of multicollinearity must be avoided in order to understand the effect of the individual variable on the score obtained by the student.
Let's say, we obtain a final model as we have obtained in our assignment.

Bike_demand = 0.070264 − 0.297362 × lightsnow+….

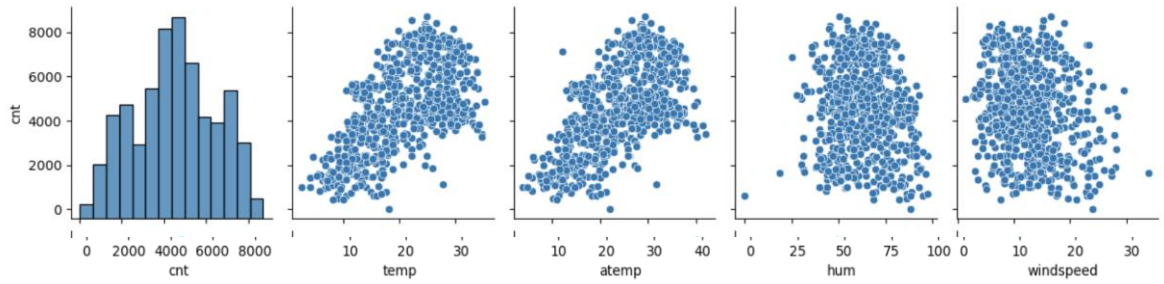
Now, the coefficient of lightsnow can be interpreted as:
During the weather situation of light snow, the bike demand is expected to decrease by approximately 0.297362 units compared to when the weather situation is clear (holding all other variables in the equation constant).

By doing this, we have a simpler model without redundant variables and multicollinearity is avoided.

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

When rounded off two two decimal places, ***Temperature (temp) and atemp*** have the ***highest correlation of 0.63*** with the target variable.

Rounding off to four decimal places, correlation of temp = 0.6270 and correlation of atemp = 0.6306.



However, since temp and atemp are highly correlated, we will drop atemp from our analysis.

Therefore, we can say ***temp has the highest correlation*** with the target variable.

It is based on the behaviour of this variable with respect to the dependent variable that I have decided that linear regression can be performed for this case study.

While building a model, even if this variable has a high VIF and a need to drop this variable arises, it has to be evaluated cautiously since this is technically a very important determinant of the outcome.

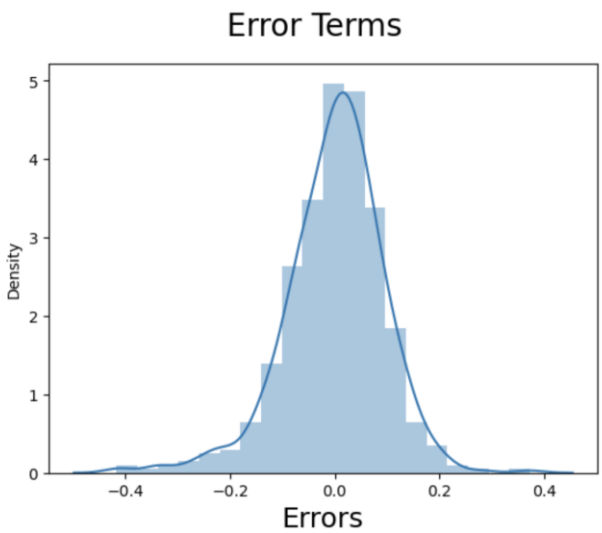
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The following are the assumptions made for linear regression:

(i) Residuals are normally distributed with mean almost equal to 0.

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

```
Text(0.5, 0, 'Errors')
```

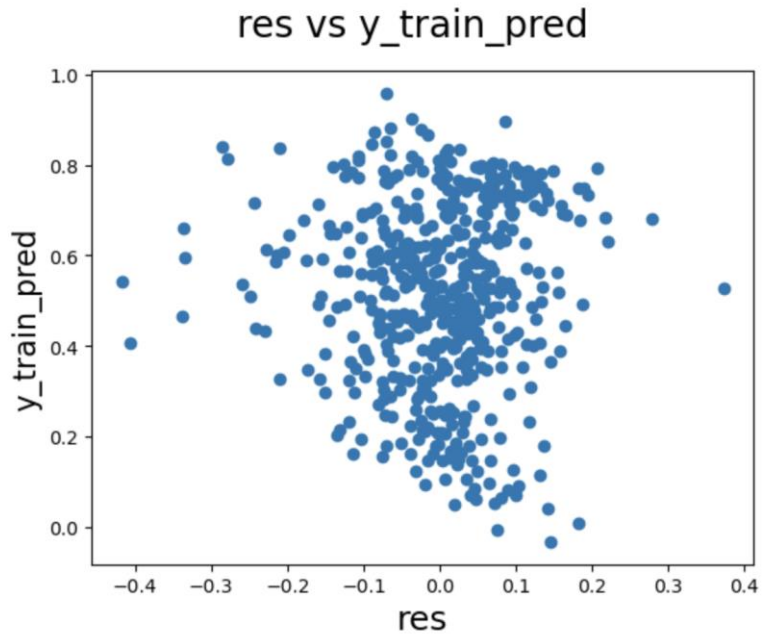


(ii) The residuals are not dependent on each other

That is, increase in the value of a residual must not affect another residual value.

```
# Plotting res and y_train_pred to understand the dependency
fig = plt.figure()
plt.scatter(res, y_train_pred)
fig.suptitle('res vs y_train_pred', fontsize = 20)              # Plot
plt.xlabel('res', fontsize = 18)                               # X-label
plt.ylabel('y_train_pred', fontsize = 16)
```

```
Text(0, 0.5, 'y_train_pred')
```

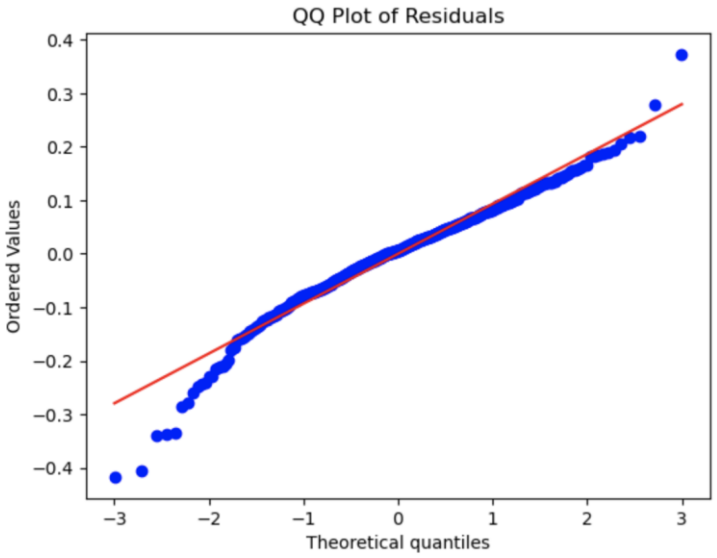


(iii) The residuals are homoscedastic

Homoscedasticity means that the model's errors (residuals) don't change in a systematic way as we make predictions across a range of values. That is, whether I'm trying to predict the bike demands when the temperature is very high or very low, errors in predictions must roughly be the same. This property is important for the reliability of regression models, as it ensures that the model's performance is consistent across different parts of the data.

This can be verified from the above two scatter plot and histograms as well. But let's plot a QQ plot and observe too. The residuals should roughly follow a straight line in the QQ plot.

```
stats.probplot(res, dist="norm", plot=plt)
plt.title("QQ Plot of Residuals")
plt.show()
```



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 equation of our best-fit model is

$$cnt = 0.070264 + 0.539040 \times temp - 0.297362 \times lightsnow + 0.231263 \times yr + 0.146153 \times winter + 0.123609 \times sep - 0.100177 \times holiday + 0.095578 \times summer - 0.080930 \times mist + 0.058455 \times aug$$

The top three features that are significant in predicting the demand for shared bikes are:

(i) **Temperature**: For every unit increase in temperature, the bike demands is expected to increase by approximately 0.539040 units (holding all other variables constant).

(ii) **Light snow**: During the weather situation of light snow, the bike demands is expected to decrease by approximately 0.297362 units compared to when the weather situation is clear (holding all other variables in the equation constant).

(iii) **Year**: Since these bike-sharing systems are gaining popularity (business-domain knowledge), the demand for these bikes is increasing every year by approximately 0.231263 units (holding all other variables constant).

General Subjective Questions

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

Regression is the most commonly used predictive analysis model for continuous variables. It falls under supervised learning and an important class of supervised learning algorithm is **linear regression.** The main objective of the linear regression

model is to find a linear relation between one or more independent variables(predictors) and a dependent variable.
There are two types of linear regression:

(i) **Simple linear regression –** where the dependent variable y is modelled as a linear relationship with one independent variable X

(ii) **Multiple linear regression –** where the dependent variable y is modelled as a linear relationship with more than one independent variable $X_1 \ldots X_n$.

Mathematical formula:

$$\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n + \epsilon}$$

where,

b_n is the change in X_n when all other coefficients are held constant

b_0 is the intercept

y is the predicted value of the dependent variables

Mathematically, the objective of linear regression is to find the values of the constant (intercept) b_0 and coefficients $b_1 \ldots b_n$ for the corresponding predictor variables $X_1 \ldots X_n$ in such a way that the r-squared value (calculated from the predicted values and the actual values of y) is maximized.

To build the model, the following steps are followed:

1. **Understand the dataset** – Get an understanding of the dependent variable and the predictors in the dataset.
2. **Visualize the data** –
   Visualize numeric variables using pairplot and categorical variables using box plots to identify is some predictors have a strong association with the outcome variable and to identify obvious multicollinearity going on in between pairs of predictors.
3. **Data preparation –**
   - Drop variables that do not add any business value or technical value while building the model.
   - Perform dummy encoding(one-hot encoding) or label encoding on the categorical variables to convert them to have numeric nature.
4. **Split the data into training and test set –**
   From the given dataset, the train-test split is either a 70-30 or a 80-20. We build the model using the training dataset and test the model using the test dataset.
5. **Rescale the training data –** Rescale the data so that the min-max endpoints of all the variables fall within the same range. If not scaled, variables that are widely spread will have smaller b coefficients and the ones that are not widely spread will have larger b coefficients. This will falsely indicate the effect the predictors have on y.
   Two scaling methods:

- Standardization

- Min-max scaling

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

6. **Build linear model –**
   - Split the training data into X and y dataset.
   - Linear model can be built using three techniques:
     - Manual feature selection
     - Recursive Feature elimination (automated)
     - A mixed approach

   To perform this, various tools are available. Python has two major libraries which help with model building, namely, sklearn and statsmodel.
   The idea behind choosing predictor variables is to ensure they have:
   - **Low p-value** – A low p-value (<0.05) means the variable is significant for model building.
   - **High VIF -** Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model and avoid multicollinearity.

$$VIF_i = 1/(1 - R_i^2)$$

7. **Residual analysis of trained data** – The distribution of error terms has to be analyzed. The error terms have to normally distributed with mean = 0.
   Error is calculated as

   ei = yi - ypred

   Residual Sum of Squares is calculated as:

$$RSS = \sum(e_i^2)$$

8. **Make predictions on the test data –** Perform scaling similar to Step 5 on the test data and make predictions using the model built above.

$$TSS = \sum(y_i - y^-)^2$$

$$R^2 = 1 - (RSS/TSS)$$

   Values lie between 0 and1

9. **Model evaluation –**
   - Model is evaluated by checking the r-squared value between the actual values and predicted values in the test set.
   - mean squared error can also be calculated to evaluate the model built.
   - Check if variance of the error terms in the test set is constant/linear.
10. **Derive equation of the best-fit model built above**

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

Anscombe's quartet, created by the British statistician Frank Anscombe in 1973 is a famous example that highlights the importance of data visualization and the limitations of relying solely on summary statistics for building linear regression models. It consists of four small datasets that have nearly identical summary statistics such as the mean, variance, correlation coefficient and linear regression line and yet they have very different distributions and appear quite distinct when graphed.
(source for all the three images: https://www.geeksforgeeks.org/anscombes-quartet/)

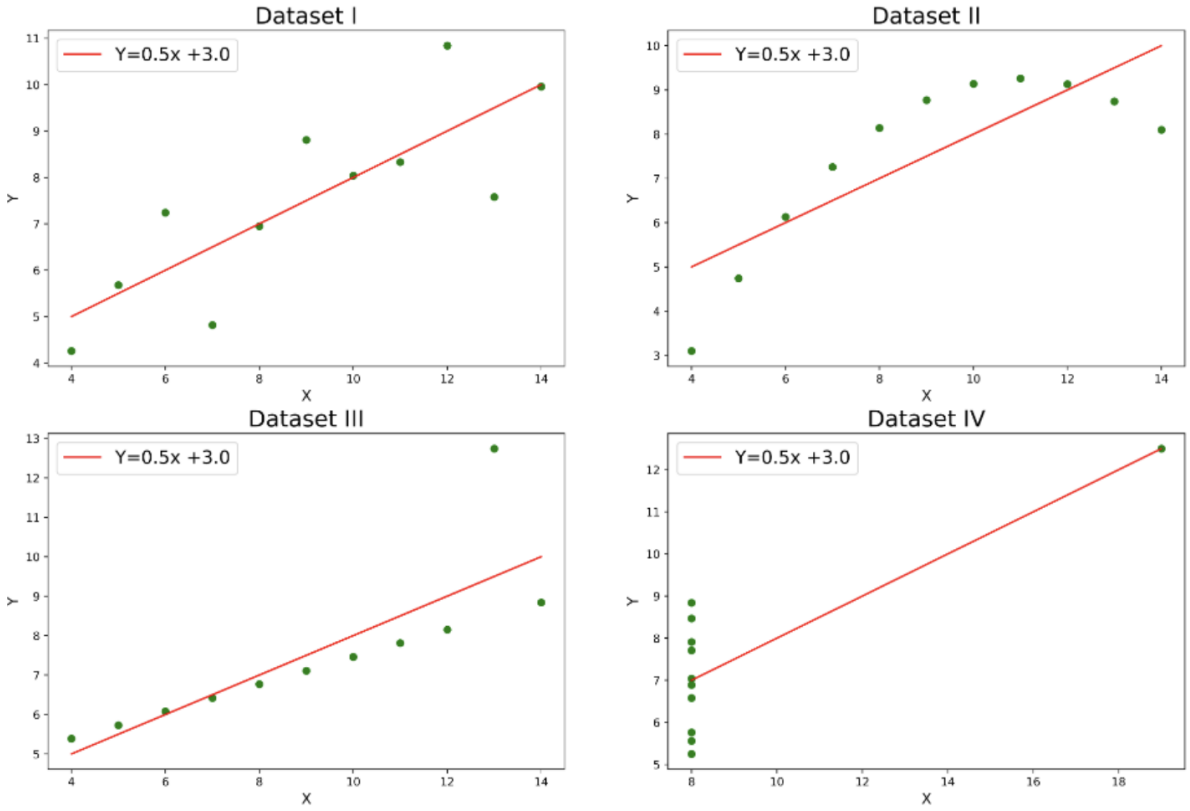
The four datasets for the quartet are shown below:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The summary statistics for the four datasets are identical as shown below:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

However, when plotted, the Anscombe's quartet plot looks as shown below:



From the plots, we can derive the following about the datasets:
Dataset 1: There is somewhat a linear relationship between x and y.
Dataset 2: There is no linear relationship between x and y.
Dataset 3: There is a linear relationship but with outliers.
Dataset 4: No linear relationship but an outlier that can completely alter the linear equation.

We can deduce the following learnings from Anscombe's quartet:

(i) **Balancing reliance on visualization and summary statistics –** Visualizing data through plots reveals patterns and outliers that are masked in summary statistics.

(ii) Influence of outliers – As seen in the case of dataset 4, a single outlier can drastically alter the correlation coefficient and thereby affect the linear regression line.

(iii) Overreliance on summary statistics – We see how although the summary statistics of all four datasets look the same, relying solely on the summary statistics is going to affect the model that will be built. Visualizing the data is a pre-requisite before deriving analysis based on summary statistics.

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

Pearson's R is also called Pearson correlation coefficient and is indicated by R. It is a statistical measure used to analyse the linear relationship between two continuous variables. It can take values between -1 and 1, with the following interpretations:

(i) Zero or close to zero: No linear relationship between the variables. Changes in one variable do not predict changes in the other.

(ii) Greater than 0 (closer to 1): Indicates a strong positive linear relationship. As one variable increases, the other tends to increase as well.

(iii) Less than 0 (closer to -1): Suggests a strong negative linear relationship. As one variable increases, the other tends to decrease.

The formula to calculate Pearson's R is:

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2 n\sum y^2 - (\sum y)^2}}$$

Example: Let's calculate the Pearson's R for the given heights and weights of students.

Student	Height X	Weight Y
A	43	79
B	24	68
C	25	84
D	33	65

Solution:

The calculated values are highlighted in yellow.

Student	Height X	Weight Y	XY	X^2	Y^2
A	43	79	3397	1849	6241
B	24	68	1632	576	4624
C	25	84	2100	625	7056
D	33	65	2145	1089	4225
$\sum$	125	296	9274	4139	22146

n = 4

Applying the formula above, we get: R = 0.10

Since the value of R is closer to 0, we can conclude that there is no strong linear relationship between height and weight of students.

Since the value is positive, we can say that as the height increases, weight also increases but the factor by which they increase is not linear.

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

While performing multiple linear regression, we will have many variables to deal with. Each variable might be on a different scale, i.e., the min-max value range for all variables will not be the same.

Scaling is the process of transforming the independent and dependent variables to a similar scale.

It is done to ensure the coefficients calculated in the model are not disproportionate and do not influence the model training incorrectly. If a model is built with variables on varying scales, variables that are widely spread will have smaller b coefficients and the ones that are not widely spread will have larger b coefficients. This will falsely indicate the effect the predictors have on the dependent variable.

The two scaling methods are:

(i) Normalized scaling: This is also called min-max scaling. This method scales the features in the range [0,1].

The formula is:

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

This scaling method is preferred when we want the variables to be within a particular range. Normalized scaling will not affect one-hot encoded or binary encoded categorical variables since these variables have values 0 and 1 (min = 0, max = 1). When we have many one-hot encoded or binary encoded categorical variables in our dataset, using normalized scaling will bring all the variables within the 0 to 1 range.

(ii) Standardized scaling: This is also called Z-score scaling. This method scales features to have a mean = 0 and standard deviation = 1.

The formula is:

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

This scaling method is less affected by outliers.

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

Let's look at the formula for VIF.

VIF = $1/(1-R^2)$

When $R^2 = 1$, VIF will become infinite.

R^2 is the factor that determines how well a variable can be explained by other variables. If R^2 for a variable $X_1 = 0.8$, it means 80% of the variation in the variable X_1 can be explained by other variables.

If $R^2 = 1$, it means all the variation in X_1 can be perfectly explained by the presence of other variables making X_1 redundant and must be carefully reviewed in comparison to other variables before model building.

This implies that when a variable has VIF = infinity, this variable is perfectly linear to one or a combination of other variables in the model. This is called perfect multicollinearity.

If VIF of a variable is equal to infinity, it can mean one of the following:

(i) The variable is a duplicate or is redundant to another variable in the dataset.
(ii) The variable has a perfect linear relationship with another variable. That is, X = nY will result in perfect linear relationship.
(iii) The variable is linearly dependent with not one but a subset of variables.

It is important to identify these variables with high VIF and address them by performing one of the following before building a model:

(i) remove them
(ii) remove the variables they are dependent on
(iii) combine variables to derive a new feature
(iv) if the dataset is small, collect more data to verify the linear dependency of the variable

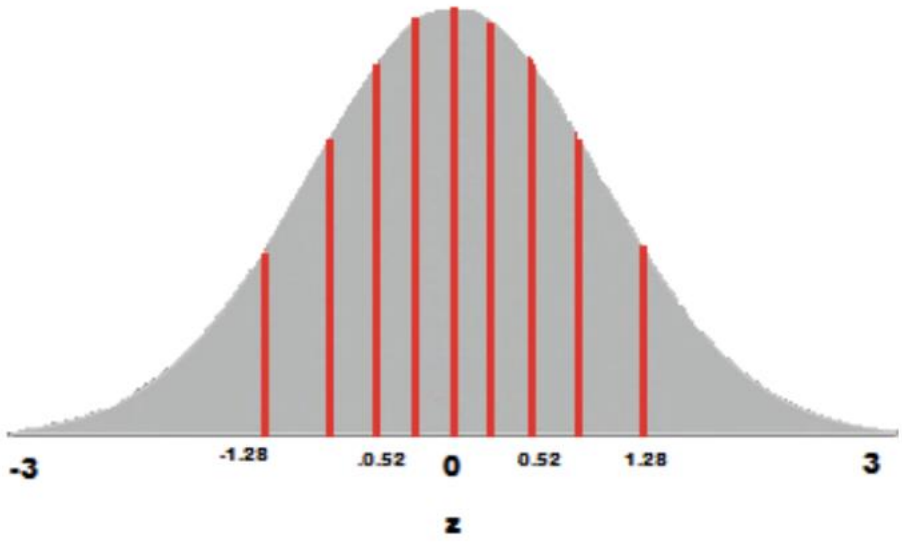
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot(quantile-quantile plot) is a graphical tool used in data analysis to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is a way to visually compare the quantiles of the observed data to the quantiles of a theoretical distribution (usually a normal distribution). We know quantile is a subset of the dataset (usually at 25th, 50th and 75th percentile).

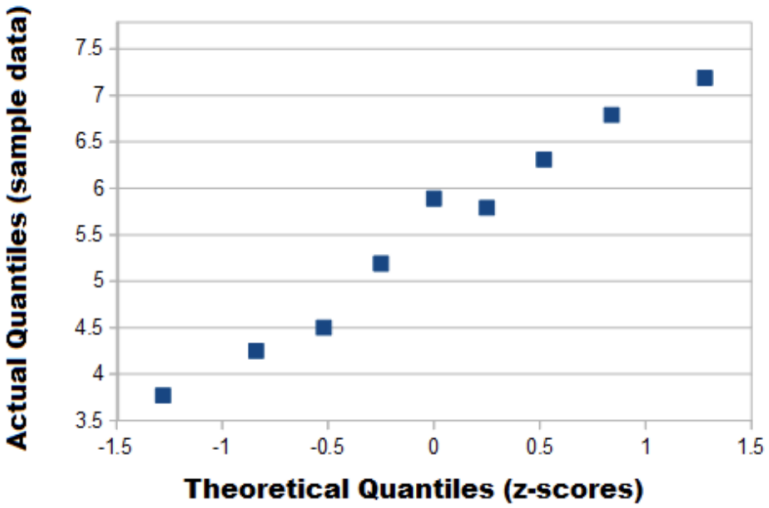
One of the main assumptions of a linear regression model is that the residuals(difference between y_{actual} and $y_{predicted}$) are normally distributed. Since a Q-Q plot can be used to assess if a dataset follows a particular theoretical distribution like normal distribution, we can use the Q-Q plots for the analysis of residuals in the following two ways:

(i) **Normality assumption of residuals and homoscedasticity:** One of the main assumptions of a linear regression model is that the residuals(difference between y_{actual} and $y_{predicted}$) are normally distributed.
We follow the below steps to identify if the residuals are normally distributed:
   - Sort the residuals in ascending order.
   - Draw a normal distribution curve, dividing the curve in n+1 equal areas. If I have 9 residuals, I will divide the curve into 10 equal areas.

- Find the z-value for the 10 segments. (Below is a sample normal distribution curve with z-values, source: https://www.statisticshowto.com/q-q-plots/)

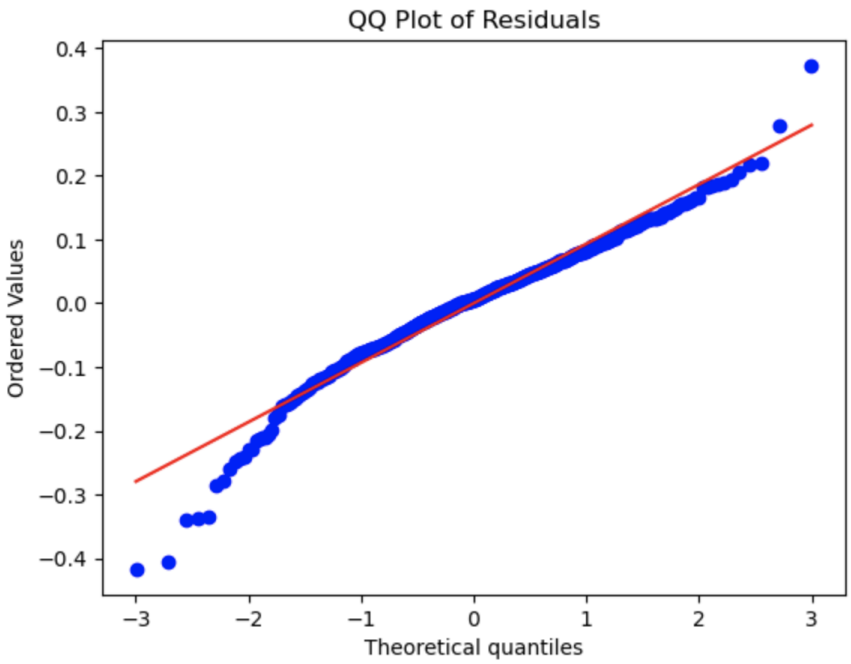


- Plot the residuals against the normal distribution cut-off points having the actual residual values on the y-axis. (Sample plotting is shown below, source: https://www.statisticshowto.com/q-q-plots/)



A normally distributed residual data set will have an almost straight line on this q-q plot.

Here's the example of a q-q plot drawn in the assignment for residuals:



(ii) **Identify outliers in the residuals:** By plotting a Q-Q plot as shown above, we will clearly be able to identify outliers in residuals. These outliers can have a significant impact on the regression model and this visualization can help us take appropriate actions.