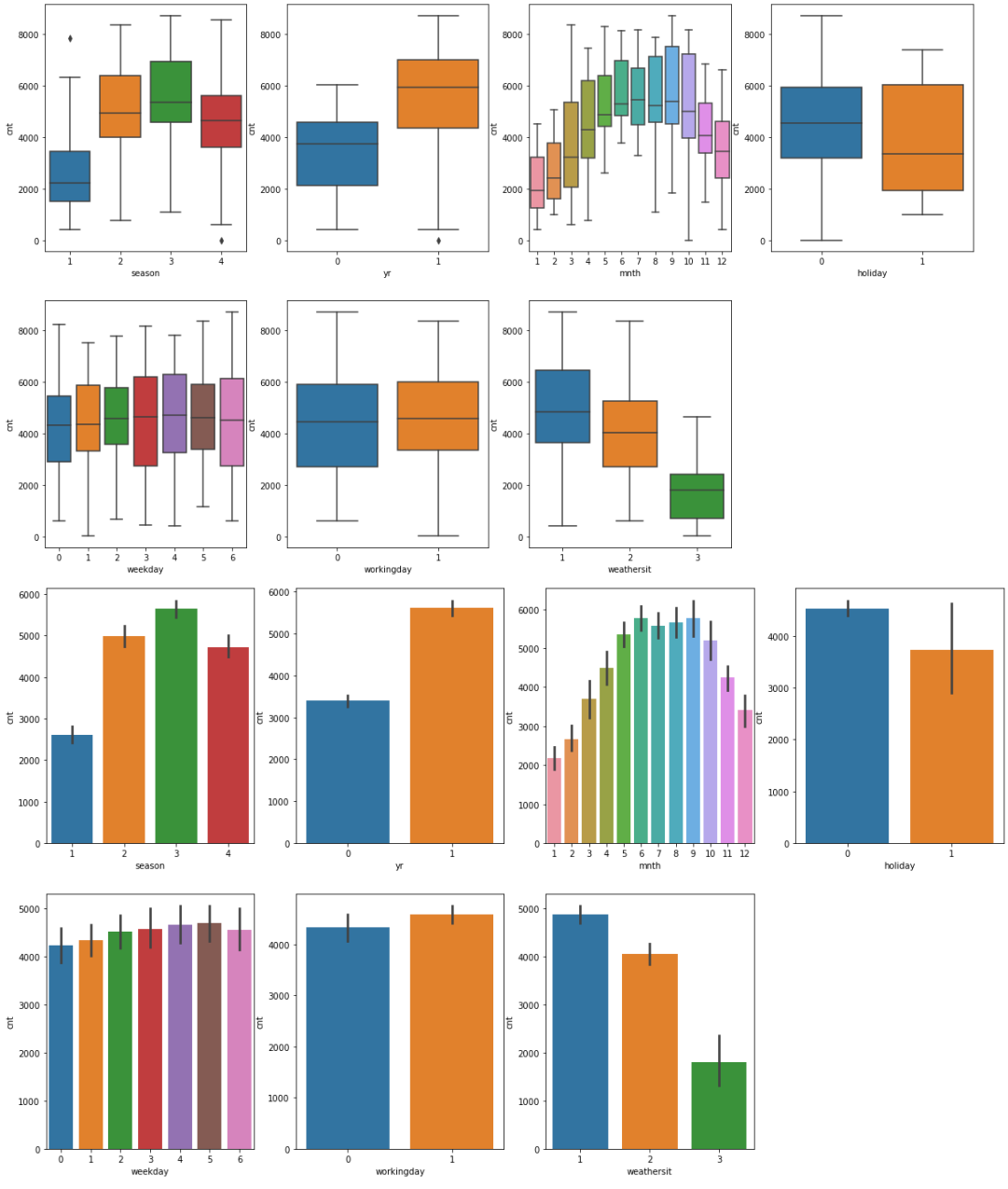
Assignment-based Subjective Questions

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

There are 7 categorical variables in the dataset provided. These are as follows:
['season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit']

For categorical variables boxplot and barplot was plotted to visualize the target variable dependency on categorical variable. Below are the images of the same.



From the above graphs we can analyse the below:

- Season 3/Fall season attracts more bike sharing amongst the user followed by season 2/summer which is aligned with median being high for these 2 seasons.
- Year 1 or 2019 has more bookings as than 2018 indicating a positive growth over years.
- The month of June and September have higher bookings.
- On non-holidays (0) there are more bookings, indicating people do not prefer to book bikes on a holiday.
- On weekday 4/Thursday, 5/Friday there are comparatively higher demand of bike bookings.
- There are slightly more bookings on a working day (1) compared to weekend(0) whereas the median is almost the same.
- Weather situation 1/ Clear, Few clouds, Partly cloudy, Partly cloudy has more demands for booking of bikes than any other weather situation. Weather situation 4/ Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog does not have any bookings in the dataset provide.

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

We convert the categorical variables to dummies using the pandas method get_dummies which by default, returns one hot encoding of the categorical variable. The keyword argument drop_first=True is used which returns P-1 columns. This drop_first=True is important during dummy variable creation as it removes redundancy and the problem of multicollinearity.

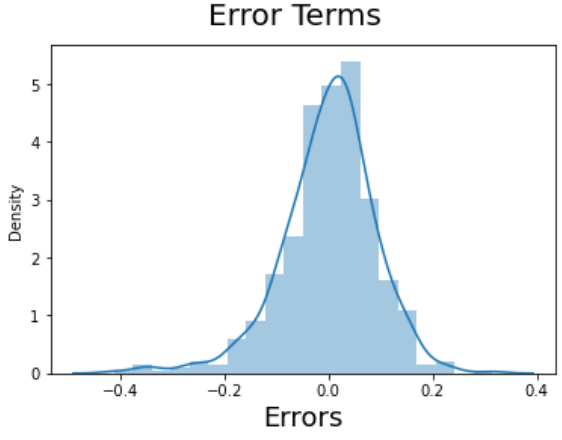
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

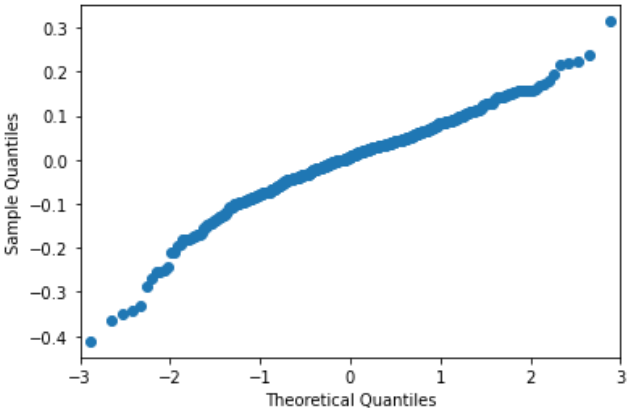
temp has highest correlation with the target variable. As temp and atemp are highly correlated with each other, so atemp also has similar correlation with the target variable.

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

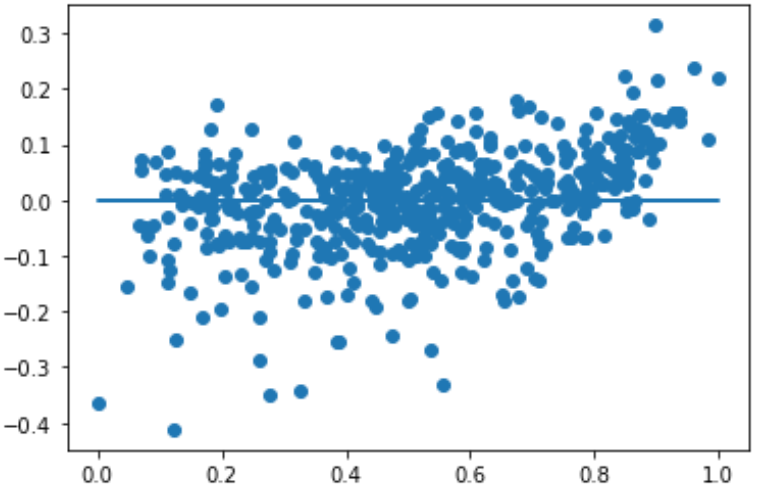
After building the model on the training set, below validations were made:

- ❖ There is a linear relationship between predictor and target variable.
- ❖ Error terms are normally distributed. Below is the graph for same.





- ❖ No or little Multicollinearity between the predictor variables. Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated.
- ❖ Error terms have constant variance (homoscedasticity).



- ❖ Error terms are independent of each other. Little to no autocorrelation exists between the residuals.

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

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

- Temperature: A coefficient value of '0.549892' indicated that a unit increase in temp variable increases the bike hire numbers by 0.549892 units.
- Weather situation 3 or Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds: A coefficient value of '-0.287090' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.287090 units.
- Year: A coefficient value of '0.233139' indicated that a unit increase in yr variable increases the bike hire numbers by 0.233139 units.

General Subjective Questions

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

Linear regression is one of the most basic machine learning algorithms. It is a supervised machine learning method which attempts to explain the relationship between dependent and independent variables using a straight line. The independent variable is known as predictor variable and the dependent variable is known as the output variable. The algorithm aims to find the best-fit line that minimizes the difference between the predicted values and the actual target values. The equation of the best fit line

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where:

$\hat{y}$: *dependent variable (predicted value)*

β_0 : *estimated intercept*

$\beta_k X_k$: *estimated slope coefficient*

Can be formed by minimising the cost function (RSS in this case), using the Ordinary Least Squares Method)

Linear regression models can be classified into two types based upon the number of independent variables:

- Simple Linear Regression refers to the method used when there is only one independent variable,
- Multi-Linear Regression refers to the method used when there is more than one independent variable.

The strength of linear regression model is mainly explained by R-squared, where

R-squared = 1- (RSS/TSS)

RSS: Residual Sum of Squares

TSS: Total Sum of Squares

Since, we are making inferences on the population using a sample, the assumption that the variables are linearly dependent is not enough to generalise the results obtained on the sample to the population, which is much larger in size than the sample. Thus, we need to have certain assumptions in place in order to make the inferences. Below are the assumptions of the multi linear regression model:

- There should be a linear relationship between the predictor variables and the output variable.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance.
- Little to no Multicollinearity exists between the predictor variables. This assumes that the predictors used in the regression are not correlated with each other.

Apply the trained model to make predictions on new, unseen/test data. The r-squared for the test data should not be overfitting or underfitting. A variance of 0.05 is accepted in the values or r-squared of train and test data.

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

Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

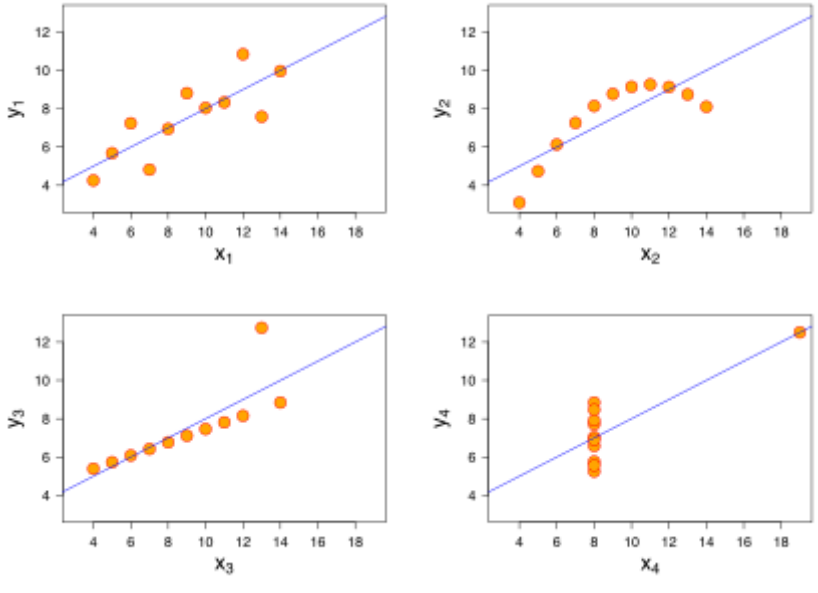
Below four datasets were used by Anscombe where we see that for all four datasets the simple descriptive statistics is identical.

	I		II		III		IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	*9*	*7.5*	*9*	*7.5*	*9*	*7.5*	*9*	*7.5*
Sample Variance	*11*	*4.127*	*11*	*4.128*	*11*	*4.123*	*11*	*4.123*
Correlation between x and y	*0.816*		*0.816*		*0.816*		*0.817*	

Linear Regression Line		*y = 3.00 +0.500 x*		*y = 3.00 +0.500 x*		*y = 3.00 +0.500 x*		*y = 3.00 +0.500 x*
R-squared	*0.67*		*0.67*		*0.67*		*0.67*	

However, when we plot the graph, we find that the four dataset has totally different behaviour.

- The first (top left) scatter plot appears to be a simple linear relationship.
- The second graph (top right) there is a non-linear relationship.
- In the third graph (bottom left), the modelled relationship is linear, but the calculated model has an outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Pearson's r also known as Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

Alternative formula is as below:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

where

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above and:
- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y.

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

Scaling is the process of transforming the numerical features of a dataset to a common scale. It involves applying a mathematical transformation to ensure that all features have similar ranges or distributions. It is important because:

- ❖ Prevents the features with larger scales to dominate the learning algorithm or analysis, ensuring that all features contribute equally to the learning process.
- ❖ Enables distance-based algorithms by avoiding biases due to difference in scales.
- ❖ Improves the convergence speed and stability of optimization algorithms

There are two commonly used techniques for scaling: Standarized scaling and normalized scaling. Difference between the two techniques is as below:

Normalized scaling	Standardized Scaling
It is used to transform features to be on a similar scale.	It scales the features to have zero mean and unit variance
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Formula to derive the new values is: X_new = (X - X_min)/(X_max - X_min)	Formula to derive the new values is: X_new = (X - mean)/Std
It is used when features are of different scales and the distribution is unknown.	It is used when we want to ensure zero mean and unit standard deviation and the feature distribution is Normal or Gaussian.

Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is affected by outliers.	It is not much affected by outliers.

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

Variance inflation factor or VIF measures the strength of the correlation between the independent variables in regression analysis. If there is a perfect correlation between variables then the VIF is infinite. This happens because the independent variable has a high collinear relationship to the other variables.

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 also known as Quantile-Quantile plots is a graphical plotting of the quantiles of two distributions with respect to each other. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. The advantage of Q-Q plot is:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

The Q-Q plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.