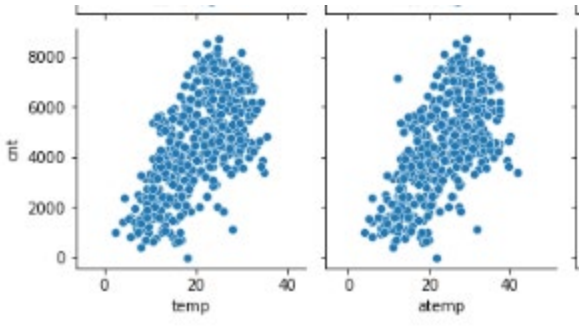
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   Answer: The following are the inferences of the analysis on the categorical variables on the target variable:
   - Season: Season_3 shows the maximum bookings with the median of greater than 5000, followed by season_2, season_4 and season_1 (where season_1:spring, season_2:summer, season_3:fall, season_4:winter). It is a good predictor variable to be considered in the model.
   - Mnth: As per the box plot the bookings are above 4000 for the months 4, 5, 6, 7, 8, 9,10.
   - Weathersit: Maximum bookings are done in weather1 with a median of around 5000, followed by weather2, weather3. It is a good predictor variable to be considered in the model.(weathersit :
     - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
     - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
     - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
     - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
   - Holiday: Majority of the booking about 97.2% is happening on non-holiday. The data doesn't seem to be distributed so it can't be considered as a good predictor.
   - Weekday: There is no clear difference in trend on all the days of the week.
   - Workingday: Bookings done on a working day have a median of greater than 5000. It can be a good predictor variable to be considered in the model.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer: The drop_first=True is used to help in reducing the extra column being created during dummy variable creation, which could cause correlations among other variables. If the categorical variable has n-levels, then we need to use n-1 columns to represent the dummy variables.

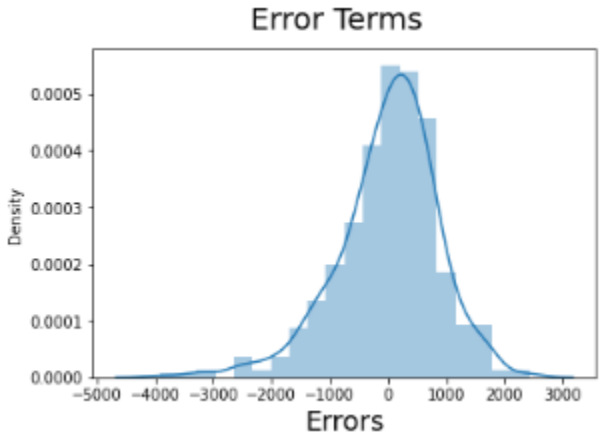
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer: In the pair plot we can see that there is a linear relation between the variables temp, atemp and the target variable cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer: Below are the assumptions of Linear Regression and the results on the training dataset:
   a). Linearity: The dependent and independent variables should show a linear relationship.



b). Homoscedasticity: The variance of the error term has to be same across all values of the independent variable.

c). Normal Error: The error term should be normally distributed



d). No Autocorrelation of residual: This is assumption is applicable to time series data. Autocorrelation means the current value of Yt is dependent on historic value of Yt-n with n as lag period.

e). No Multi-Collinearity: Multi-Collinearity is a phenomenon when two or more independent variables are highly correlated. Variance Inflation Factor (VIF) can be used to identify multi collinearity. A high value of VIF indicates high collinearity.

VIF values of the Bike sharing case study

	Features	VIF
1	temp	3.83
2	windspeed	2.95
0	yr	2.00
3	season_2	1.54
4	season_4	1.34
5	mnth_9	1.19
6	weathersit_3	1.06

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

Answer:

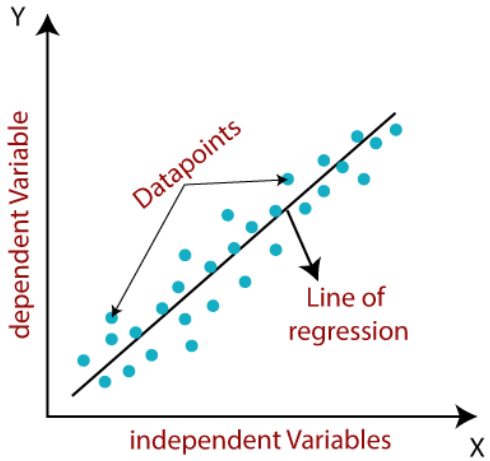
Temperature (temp) - A coefficient value of '4975.19' indicated that a unit increase in temp variable increases the bike hire numbers by 4975.19 units.

Weather Situation 3 (weathersit_3) - A coefficient value of '-2180.0'' indicated that, with resp ect to Weathersit, a unit increase in Weathersit3 variable decreases the bike hire numbers by 218 0.08 units, where weathersit_3 refers to Light Snow, Light Rain + Thunderstorm + Scattered clo uds, Light Rain + Scattered clouds

Year (yr) - A coefficient value of '2033.34'indicated that a unit increase in yr variable increases the bike hire numbers by 2033.34units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Answer: Linear regression is a machine learning algorithm based on supervised learning, it shows a linear relationship between a dependent (y) and one or more independent (y) variables. Simple Linear Regression is represented by the equation y=mx+c where m is the slope and c is the intercept.



It helps in finding how the value of the dependent variable is changing according to the value of the independent variable.

Two types of Linear regressions:

1. Simple Linear Regression
2. Multiple Linear Regression

The strength of the linear regression model can be assessed using 2 metrics: R^2 or Coefficient of Determination and Residual Standard Error (RSE)

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1.

$R^2 = 1 - (RSS / TSS)$
Where, RSS: Residual Sum of squares, TSS: Sum of errors of data from mean

Assumptions of simple linear regression are:
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Hypothesis testing of the co-efficients: The p-value is used to determine the significance of the variable.
Parameters to assess a model are:
1. t-statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not.
2. F statistic: Used to assess whether the overall model fit is significant or not. The higher the value of F statistic, the more significant a model turns out to be.
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit.

Multiple Linear Regression is used to understand the relationship between one dependent variable and several independent variables.

Multicollinearity: Multicollinearity is the effect of having related predictors in the multiple linear regression model that is one variable related to more than one other independent variable. It can be determined by the correlation matrix , VIF(Variance Inflation Factor). Higher the VIF value greater is the correlation. The variables with high VIF can be dropped one by one to reduce the multicollinearity

Feature Scaling: The independent variables in a model, might be on different scales which will lead a model to result in unexpected coefficients that might be difficult to interpret. So we use the Standardizing or the Min Max Scaling approach to scale the variables.

Creating dummy variable for the categorical variables is essential to build a good model. For categorical variables with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one.

Adjusted R-sq penalizes the models based on the no. of variables present in it. If a variable is added and the adjusted R-sq drops then that variable is insignificant for the model and it has to be dropped.

One of the automated options of Feature selection is Recursive Feature Elimination (RFE). In this approach, the RFE is applied repeatedly on the train dataset to identify the

best model. In the process the variables with high p-value and high VIF values are dropped one by one and we arrive at a final model. Then the test dataset is used to make the final predictions.

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

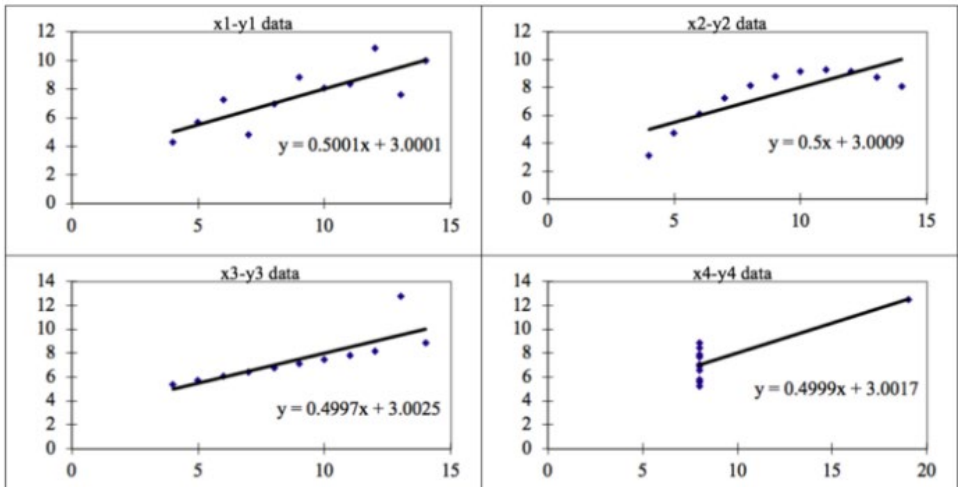
Answer: Anscombe's Quartet can be defined as a group of four data sets which seem to be identical in simple descriptive statistics, but they turn out to be different when they are plotted in a graph.

It describes the importance of data visualization before proceeding with the model building. The statistical data may be misleading since it does not display the distribution or the anomalies of the data.

The statistical info as shown in the screenshot looks similar:

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 below screenshot shows the distributed data when plotted in a scatter plot.



Conclusion: Dataset1 shows a good linear regression fit.

Dataset2 could not fit the linear regression model well

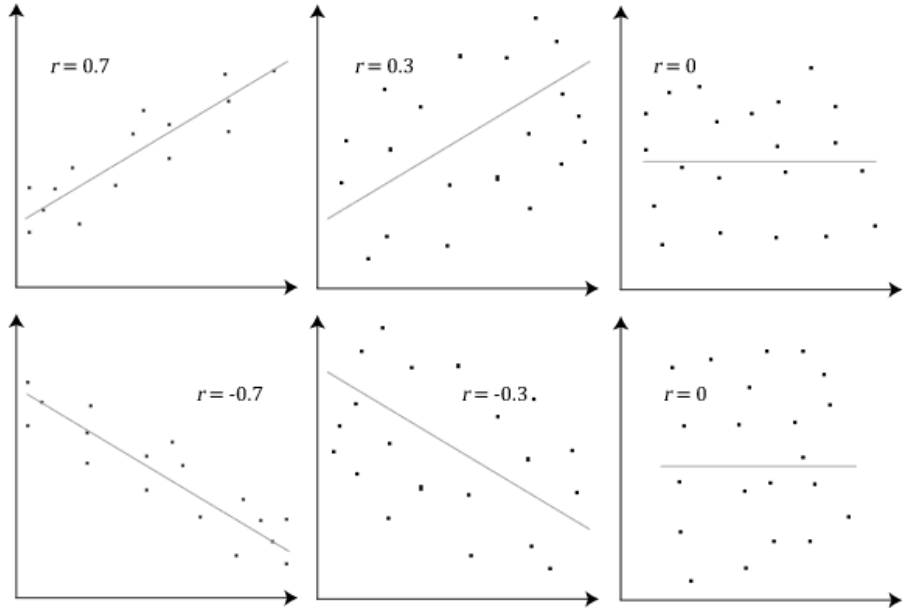
Dataset3 and Dataset4 show outliers and can't be handled in a linear regression model.

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

Answer: The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*.

The Pearson correlation coefficient, *r*, 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 correlation and a value lesser than 0 indicates negative correlation.

The below screenshot shows the Pearson correlation coefficient value of two variables and the strength of the associated variables:



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

   Feature scaling refers to the process of converting the variable into the same range of values. The independent variables in a model, might be on different scales which will lead a model to result in unexpected coefficients that might be difficult to interpret. There are two types:

Normalized scaling: The feature variables are mapped to a minimum value of 0 and a maximum value of 1.

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized scaling: The feature variables are not set to 0 or 1 but a calculated based on the formula with the mean of 0 and standard deviation of 1.

$$z = \frac{x - \mu}{\sigma}$$

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

Answer: The VIF value is infinite for the variables that have a perfect correlation. For such variables the Rsq value will be 1 which will lead to 1/(1-Rsq). The infinite values for VIF was observed in the Bike sharing case study as shown below:

	Features	VIF
1	holiday	inf
2	workingday	inf
9	weekday_1	inf
10	weekday_2	inf
11	weekday_3	inf
12	weekday_4	inf
13	weekday_5	inf
4	hum	10.94
3	temp	7.20
5	windspeed	3.58
0	yr	2.03
7	season_4	1.64
6	season_2	1.58
8	mnth_9	1.21
14	weathersit_3	1.10

On dropping the variable with high p-value the perfect correlation had changed and hence the VIF values were modified to other than infinity.

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

Answer: Q-Q plots are also known as Quantile-Quantile plots, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. It helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. In probability distributions, the data is represented in charts where the x-axis represents the possible values of the sample and the y-axis represents the probability of occurrence.

The machine learning models are built based on the distributions which helps us achieve the best models. The Q-Q plots are used to help us understand the data visually and is useful to determine the following:

- If two populations are of the same distribution.
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

Q-Q plot for a normal distribution is as below: