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?

According to my analysis, the categorical variables in the dataset were season, year, month, weather, holiday, and working day. As a result, I was able to infer what effect they had on the dependent variable

i. Bike bookings are highest during summer, fall, and winter. During the Spring season, there is a decline in bookings.
ii. Bookings increase year over year. Bookings are higher in 2019 compared to 2018. Next year, bookings will be even higher. As people become more conscious about their health and environment, bike rentals are becoming more popular.
iii. Average number of bike bookings per month The average number of bike bookings per month occurred between May and Sep with a median of over 4000 each month. In this case, the month is a good predictor of the dependent variable.
iv. Bike bookings are highest when the weather is clear with few clouds and partly cloudy. The median number of bike bookings for clear and cloudy weather is between 4000 and 6000 • There are fewer bike bookings when there is light snow or light rain; when there is heavy rain or heavy snow, there are no bike bookings
v. The majority of bike bookings occurred during non-holiday periods. Therefore, holiday cannot be a good predictor of dependent variables.
vi. There was a higher percentage of bike bookings on non-working days compared to working days

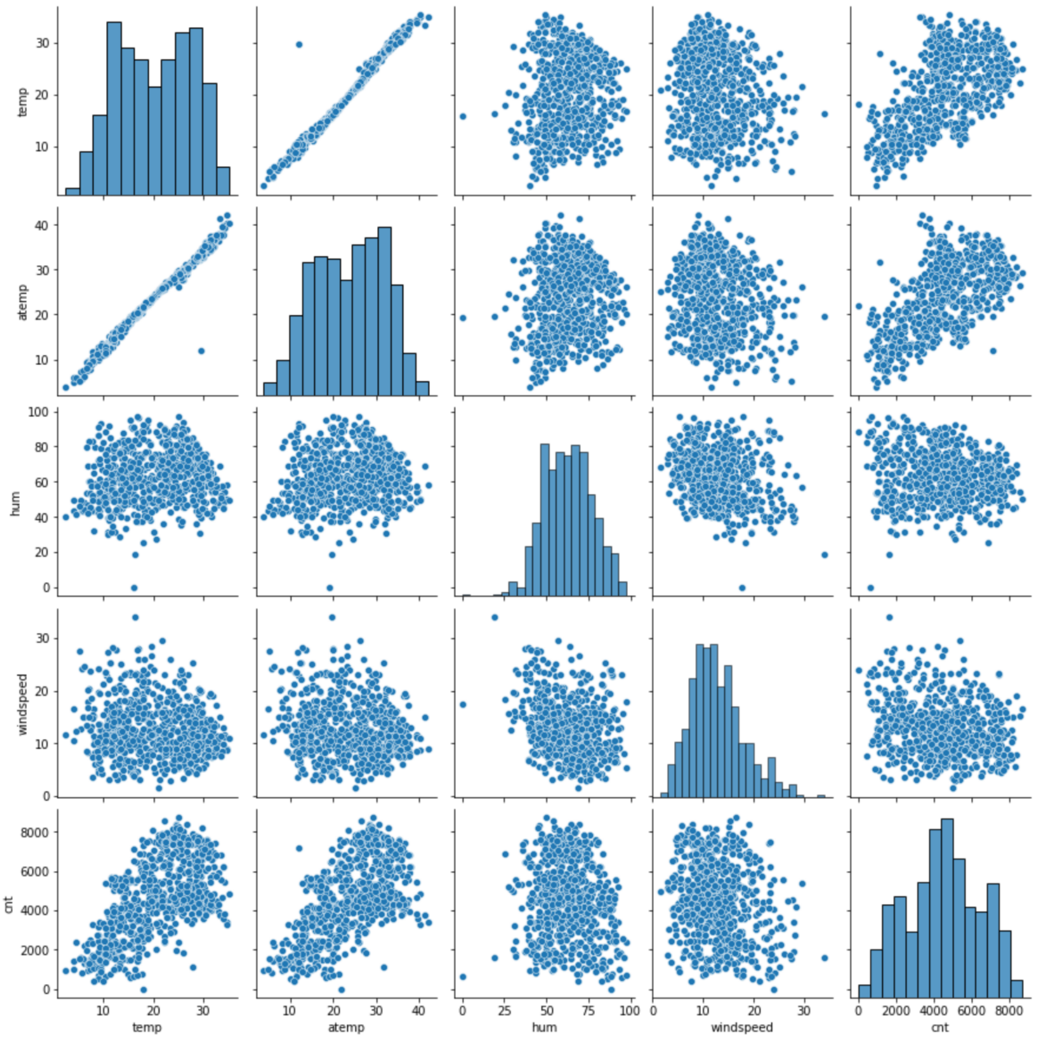
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans:

The categorical variable is created with n-1 dummy variables when drop_first is set to True. It is sufficient to use a dummy variable of n-1 for data analysis. When creating a dummy variable, drop_first = True is necessary to reduce the number of columns created. Therefore, dummy variables result in fewer correlations. There is a constant variable (intercept) that creates a multicollinearity problem if one of the dummy variables generated from the categorical variable is not deleted. Iterative models may have convergence problems and skew the importance of variables. Dummy variables also exhibit multicollinearity when all of them are present. To control this, you lose a column

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

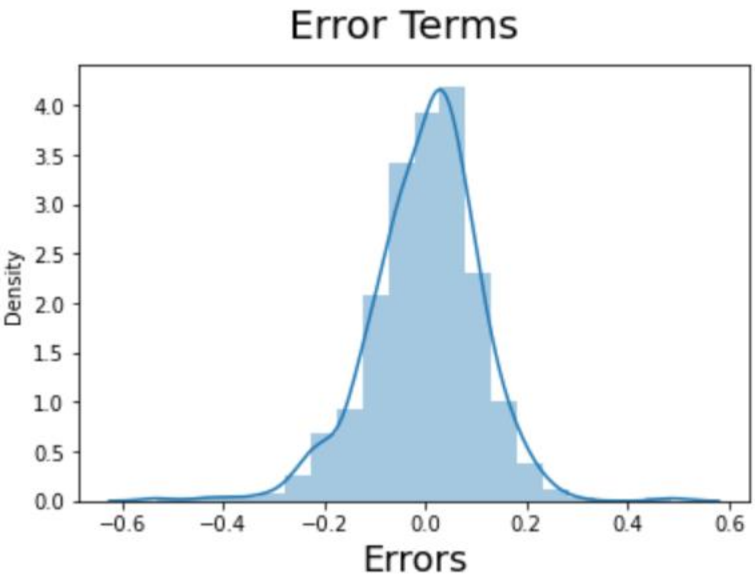
From the below graph



Pair-plots show that "temp" and "atemp" are highly correlated with target variables

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

Upon building a linear regression model based on the training data set, it is assumed that the errors are normally distributed. An analysis of residuals was conducted to support this conclusion. Remainder is the error resulting from the difference between actual and predicted y values. It is recommended that the residual distribution be normal and centered at 0 (mean = 0). By plotting the variance of the residuals, you can determine whether the residuals are normally distributed. According to the chart above, residuals are distributed around zero



After building a linear regression model on the training data set, the assumption is that the errors are normally distributed. To support this, residual analysis was performed. The remainder is the error of the difference between the actual y value and the y value predicted by the model. The residual distribution should follow a normal distribution and should be centered at 0 (mean = 0). Check this assumption about the residuals by plotting the variance of the residuals and checking whether the residuals are normally distributed. The chart above shows that the residuals are distributed around mean = 0.

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

   On the basis of the final model, the following three features significantly contribute to explaining the demand for shared bikes: Temperature (temp): As the temperature rises, so does the demand for shared bikes. As temperature increases, the number of bike bookings increases when all other variables remain constant. As the year variable increases, the demand for bikes increases. As the number of bike bookings increases by one year, all other variables remain constant. The light snow and rain discourage people from renting bikes.

General Subjective Questions

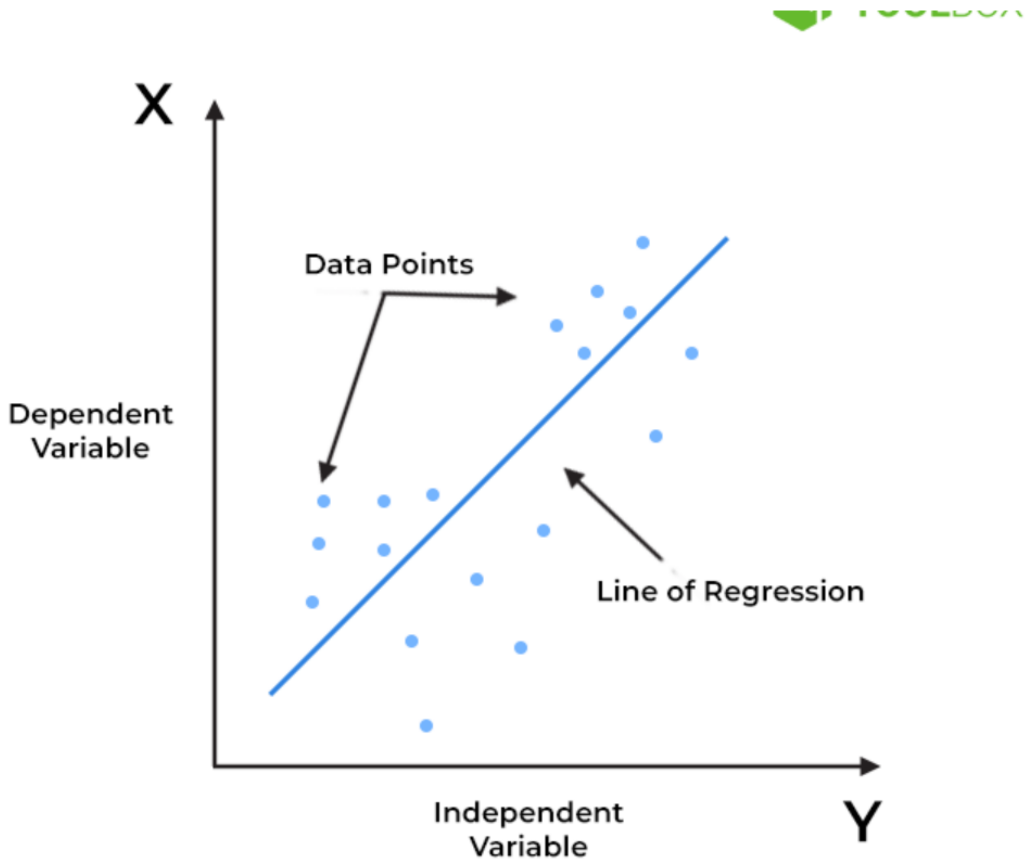
1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analysed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

A sloped straight line represents the linear regression model.

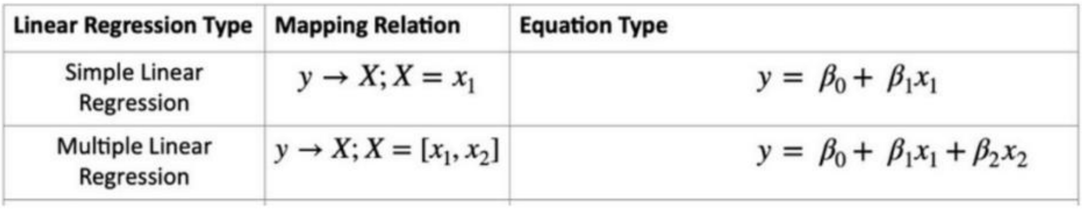


Types of Linear Regression Linear regression is of the following two types –

- Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line.
Formula: $Y=\beta0+\beta1X1 +\epsilon$

- Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables.
Formula: $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$ Where β1, β2, β3 are coefficients or slopes for the variables X1, X2 and X3 respectively and β0 is the intercept

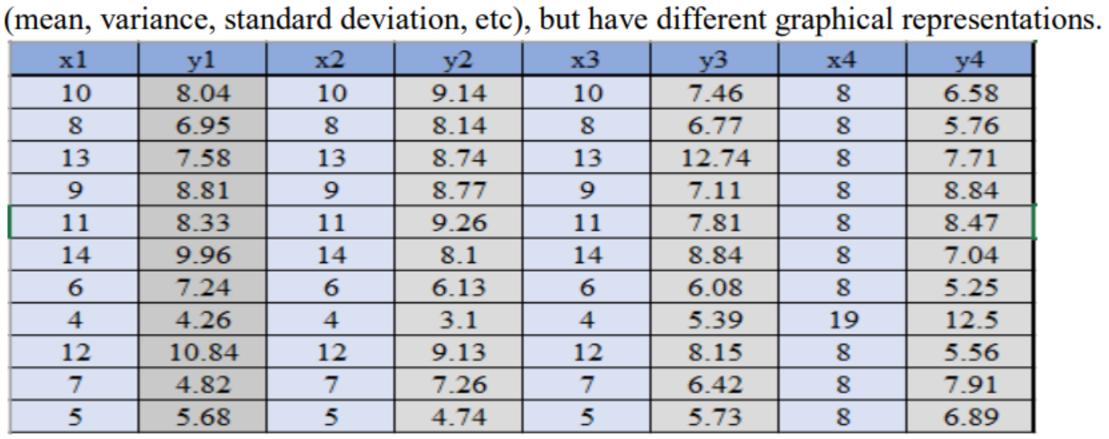
A model is a linear when it is linear in parameters relating the input to output variables. The dependency need not be linear in terms of inputs for the models to be linear. For example, all equations below are linear regression, and they define the model that represents the relationship between model parameters.



Linear Regression Type	Mapping Relation	Equation Type
Simple Linear Regression	$y \rightarrow X; X = x_1$	$y = \beta_0 + \beta_1 x_1$
Multiple Linear Regression	$y \rightarrow X; X = [x_1, x_2]$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

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

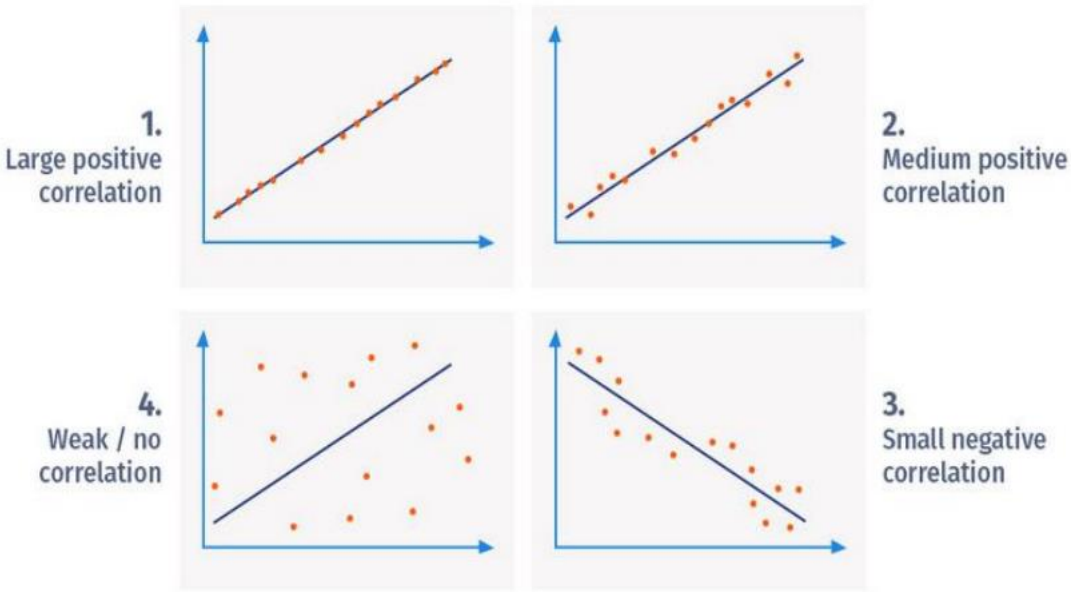
Anscombe's Quartet is a modal example of the importance of data visualization to demonstrate the importance of plotting data before analyzing it for its statistical properties. It consists of 4 data sets, each data set of 11 (x, y) points. The main thing to analyze for these data sets is that they all have the same descriptive statistics (mean, variance, standard deviation, etc), but have different graphical representations



(mean, variance, standard deviation, etc), but have different graphical representations.

x1	y1	x2	y2	x3	y3	x4	y4
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

3. What is Pearson’s R? (3 marks)
 Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables move up and down together, the correlation coefficient will be positive.
 Pearson's R measures the strength of a linear relationship between two variables. Simply put, Pearson's correlation coefficient calculates the effect of changing one variable when the other changes.



The Pearson's correlation coefficient varies between -1 and +1 where: i. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
ii. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
iii. r = 0 means there is no linear association iv. r > 0 < 5 means there is a weak association
v. r > 5 < 8 means there is a moderate association
The formula for Pearson’s R is Here, R = Correlation coefficient =values of the x-variable in a sample =mean of the values of the x-variable =values of the y-variable in a sample =mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data preprocessing to handle highly varying magnitudes or values or units.
It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. Why scaling is performed because, most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modeling.
To solve this problem, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. Normalized scaling:
i. It is a process where the variables are scaled in the range of 0 and 1.
ii. It is also called as MinMaxScaling
iii. In python, sklearn.preprocessing.MinMaxScaler helps to implement normalization
MinMaxScaling: X = X – Min(x) / Max(x) – Min(x)

Standardized scaling:
i. It is a process where the variables are scaled in a way that each data has it's mean as 0 and a standard deviation of 1
ii. In python, sklearn.preprocessing.scale helps to implement standardization iii. One disadvantage of normalization over standardization is that some information in the data is lost, especially regarding outliers.
Standardization: X = X – Min(x) / SD(X)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
The Variance Inflation Factor(VIF) is a measure of collinearity among predictor variables within a multiple regression. The formula of VIF is 1/(1-R^2). o Here R^2 denotes that how much variable is co-related to other variables. o When R^2 = 1 then VIF = Infinity. o That means when there is a perfect co-relation then VIF will be infinity. If there is perfect correlation, then VIF = infinity. Where R is the R-square value of that independent variable which we want to check how well the independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives us VIF = 1/0 which results the VIF value as infinity

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

Q-Q (Quantile-Quantile) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
It can be used to identify
If 2 samples are similarly distributed or not based on the fit- line passing closely w.r.t to the plotted points or not
It can explain if the distribution scale is similar or not depending on the angle or slope of the fit-line
It can also be used to explain what kind of distribution best fits the sample data by fitting q-q plot between quantiles of the dataset and quantiles of different distribution(uniform/normal etc)
It is used to check the below scenarios: - If two data sets come from populations with a common distribution
-If two data sets have common location and scale - If two data sets have similar distributional sha