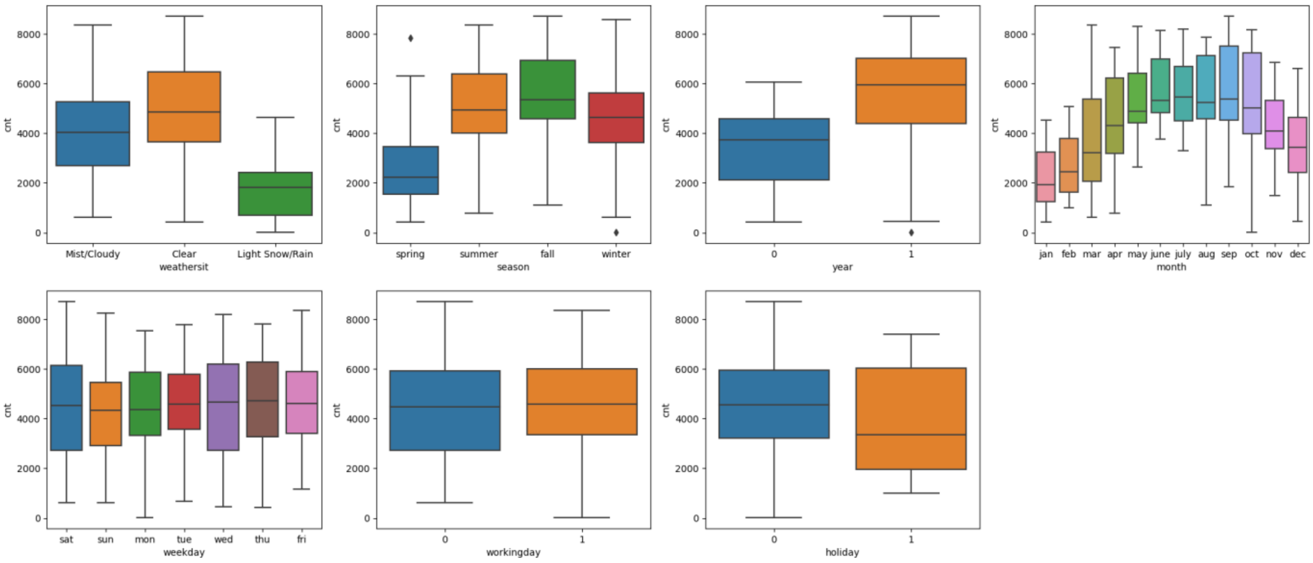
Assignment based Subjective Questions

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

Answer: Based on analysis of categorical variables following inferences were drawn:

a) Year 2019 has a greater number of bookings than 2018, which represents good progress in business.
b) Number of bookings increase till mid-year, June to September having the greatest number of bookings then it decreases by the end of year.
c) Fall season attracts most booking, followed by summer and winter.
d) If weather is clear bookings increases, if weather is misty or cloudy there's a slight decrease in number of bookings but in light rain/snow bookings are reduced drastically.
e) On holidays, there are a smaller number of bookings as compared to working days.
f) Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.



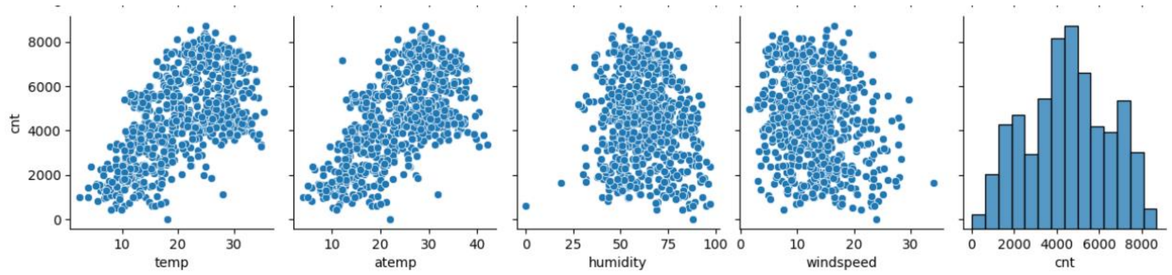
Ques 2) Why is it important to use drop_first=True during dummy variable creation?

Answer: It is important to use drop_first = True during dummy creation to avoid data redundancy, and it reduces the correlations created among dummy variables. Therefore, avoiding multicollinearity.

For a categorical feature with n levels n-1 dummies are sufficient to derive all possible combinations, for example, assume a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So, using any 3 categories we can analyse gender of the person. As, a person is either "Male", "Female" or "Other". If they are not either of these 3, their gender is "Unknown".

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

Answer: On plotting the pair-plots among numerical variables it can be seen that variable 'temp' and 'atemp' has high correlation with target variable.



Ques 4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: To validate assumptions of linear regression:

a) Normality of error terms: Built a distribution plot of error terms, to check that the error terms of model have normal distribution.
b) Multicollinearity: Used Variance Inflation Factor and heatmap to analyse relation between variables, all VIF values were <5. Hence, there's no multicollinearity in model.
c) Linear relation among variables: Plotted graphs to check for linearity.
d) Homoscedasticity: Plotted error terms against the predicted values, no definite pattern observed, therefore it can be said that model was homoscedastic.
e) Independence of residuals: Calculate Durbin-Watson score for residuals to check for autocorrelation among variables. Values was lying in range [1.5,2.5] therefore no autocorrelation.

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

Answer: Top 3 features contribution most towards demand of shared bikes are:

- Temp
- Year
- Winter

General Subjective Questions

Ques 1) Explain the linear regression algorithm in detail.

Answer: Regression is a statistical method in predictive analysis that helps to analyse and understand relation between two or more variables of interest. It helps to understands if a given feature influences and how much it influences other features.

  - Dependent variable: It's the variable we are trying to predict or forecast.
  - Independent variable: Independent variables are factors that provides information related to their relation with the dependent variable and the way it influences the dependent variable.

- Based on Dependent and independent variables we can classify linear regression as:
  - Simple linear regression: There's one dependent and one independent variable
  - Multiple linear regression: There are more than one independent variables for the model to find the relationship.

- Mathematically the simple linear relationship can be explained by equation of straight line: $y = \beta_0 + \beta_1 x$, where y is the dependent variable; x is independent variable; $\beta_1$is slope between x and y representing the effect of x on y and, β_0 is the intercept. Thus, for this regression line a unit increase in quantity of x, y increases by β_1 units.

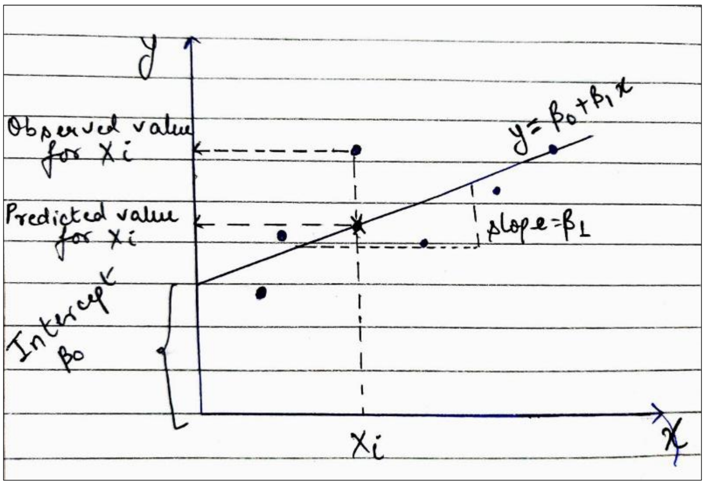
  To represent multiple linear regression mathematically the equation can be modified as:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

  In case of multiple linear regression, model fits to a hyperplane instead of line.

- There are some assumptions while implementing linear regression algorithm listed as:
  - Linearity: Relation between independent and dependent variable is linear
  - Independence: Observations are independent of each other
  - Homoscedasticity: The variance of errors is constant across all levels of independent variables.
  - Normality: Errors follow normal distribution.
  - No multicollinearity: The independent variables are not highly correlated with each other
  - No endogeneity: No relationship between errors and independent variables.

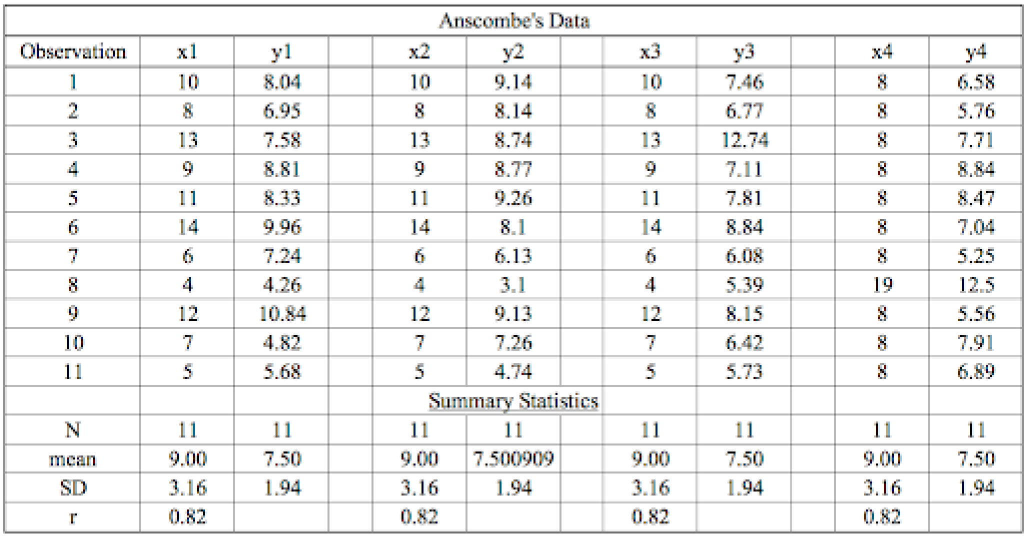
- In order to analyze performance of linear regression model following evaluation metrics can be used:
  a) R-Squared(R2) or Coefficient of determination
  b) Residual Standard Error and Root Mean Squared error.



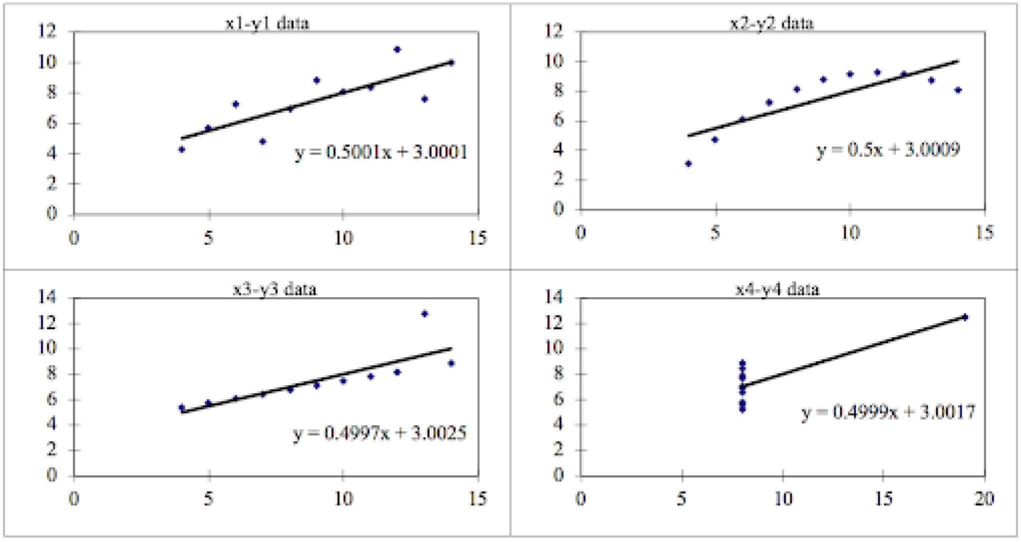
Ques 2) Explain the Anscombe’s quartet in detail

Anscombe’s quartet was designed in 1973 by statistician Francis Anscombe to illustrate the significance of plotting and visualizing data before building model. It is a group of 4 datasets that are nearly identical in descriptive statistics (mean, variance) for each value of x and y in all four datasets, but when plotted they are very different from each other.

Consider the dataset shown on figure below, the mean, variance and other descriptive statistics are almost same, however when plotted on scatter plot each dataset generates a different kind of plot.



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	



from above shown scatter plot, following inferences can be drawn:

a) Dataset 1 fits linear regression pretty well

b) Dataset 2 data is not linear and therefore, can’t fit linear regression

c) Dataset 3 distribution is linear but has outliers, which cannot be handled by linear regression

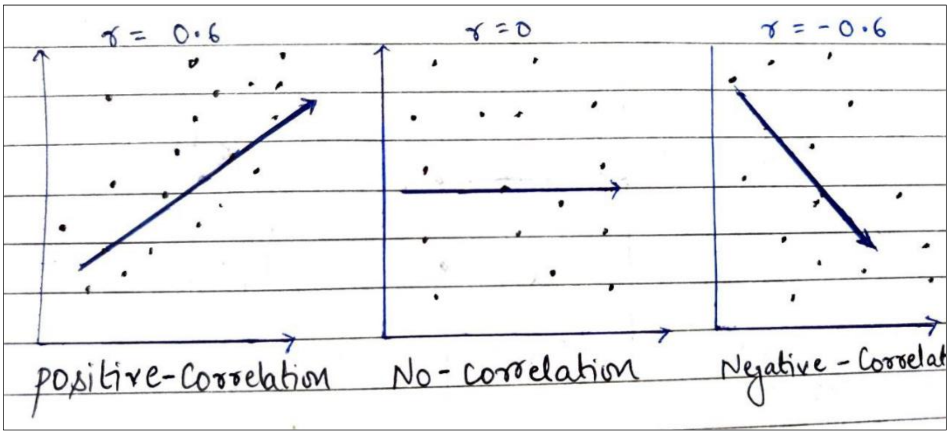
d) Dataset 4 also has outliers and represent that few outliers can produce high correlations, hence this dataset can’t be handled by linear regression.

Thus, data features must be plotted to understand data distributions that can help to analyse anomalies present in data.

Ques 3) What is Pearson’s R?

Pearson’s R also known as Pearson’s Correlation Coefficient is most common way to measure linear relation, it’s a number in range [-1,1] that helps in measuring strength and direction of relation between variables.

An absolute value of exactly 1 implies that a linear equation can be used to describe relationship between x and y, with all data points lying on a line, whereas the sign of correlation coefficient implies the direction of correlation. A positive value implies that as value of x increases value of y increases too and vice versa. Similarly, value 0 implies there’s no association between the variables.



Ques 4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature scaling is preprocessing technique used to transform feature values to a similar scale, in order to ensure that all features contribute to model equally and avoid bias of features. It’s a necessary step as dataset may have features of different measurements, ranges etc. If feature scaling is not implemented machine learning model will tend to consider features with higher values of more significance and vice versa irrespective of units of measurement.

Normalization	**Standardization**
Rescales values to a range between 0 and 1	Centres data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

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

Variance Inflation Factor is used to detect multicollinearity among variables in a dataset, multicollinearity implies that we can predict one independent variable with another independent variable. Multicollinearity becomes a big problem in linear regression when using Ordinary Least Squares algorithm because the estimated regression coefficients may become large and unpredictable, leading to unreliable inferences about the effects of the predictor variables on the response variable.

VIF score of an independent variable represents how well the variable is explained by other independent variables.

$$VIF = \frac{1}{1 - R^2}$$

A high value of R^2 means that the variable is highly correlated with the other variables. So as R2 approaches 1 VIF tends towards infinity indicating high multicollinearity among independent variables, and if R2=1 then VIF is equal to infinity. So, VIF is infinite implies there a very strong multicollinearity among independent variables.

Ques 6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Quantile-Quantile plot or QQ plot is plot of quantiles of a sample distribution against quantiles of a theoretical distribution. It helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of another dataset. By a quantile, we mean the fraction of points. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. 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 to conclude that the data points belong to different distributions.

QQ plots is very useful to determine:

- 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

In linear regression it's assumed that errors are normally distributed, if this assumption fails then confidence intervals become too wide or too narrow, and once confidence interval is unstable it causes difficulty in coefficient estimation. Here, Q-Q plot can be of high significance as it can be used to perform statistical test of normality.