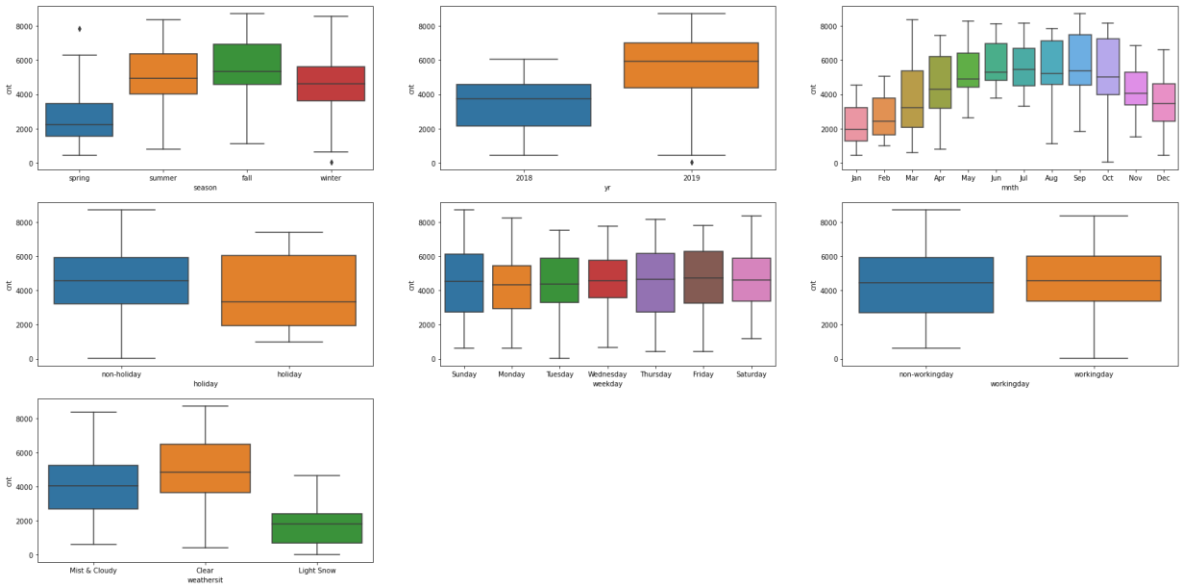
Assignment-based Subjective Questions

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



- **Season**: Demand (Cnt) is higher in Summer, fall and Winter than spring
- **Year** (yr): There is an increase in demand rom 2018 to 2019
- **Month** (mnth): There is increase in demand in mid of the year between April and Oct
- **Holiday**: There is high demand on non-holiday
- **Weekday**: No much we can infer from weekday as demand looks constant on all days.
- **Workingday**: Demand is constant on working day also.
- **Weathersit**: Demand is highest when weather is clear. Also demand of bike is high during Mist and cloudy but low when there is snow weather situation.

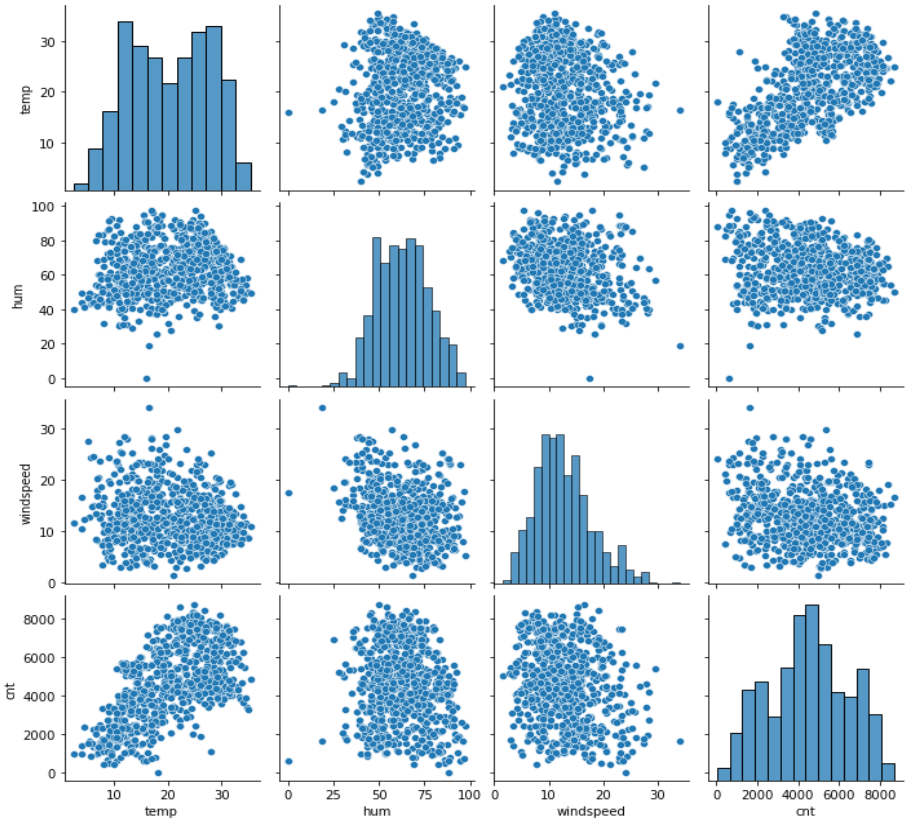
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- We do not need all k columns/variables as we can represent same information by k-1 variables, so we will delete one variable/column. It reduces one column which can analysing the correlation. Deleted column to be used as base.
  Example in case of season: So we can drop fall column as it can be indicated by 000.
  - 000 will correspond to fall
  - 100 will correspond to spring
  - 010 will correspond to summer
  - 001 will correspond to winter

  So k-1 dummy variables, where K is level for categorical variable

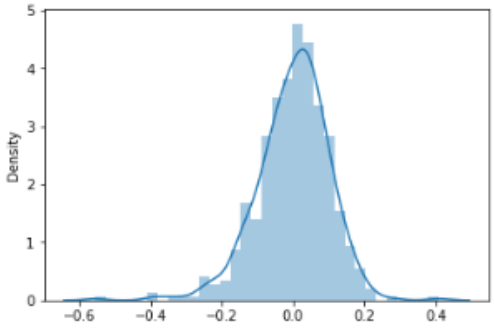
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- Temp has the highest correlation with the target cnt variable



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

- By checking the pair plot if linear relationship exists or not between target (Cnt) and predictor variables.
- The residuals of the model are normally distributed which confirms the normality. And centered around zero



- Variable selection and removal are done checking p-value and VIF.

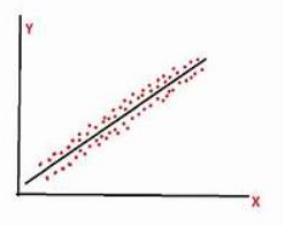
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Spring
- 2019
- Light Snow

General Subjective Questions

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

- Linear regression is a supervised machine learning algorithm.
- Target variable has linear relationship with predictor variables.
- Mathematically it can be represented as
    - Y = a + bx where a – intercept, b is coefficient of predictor variable x and x is independent variable and Y is dependent target variable.

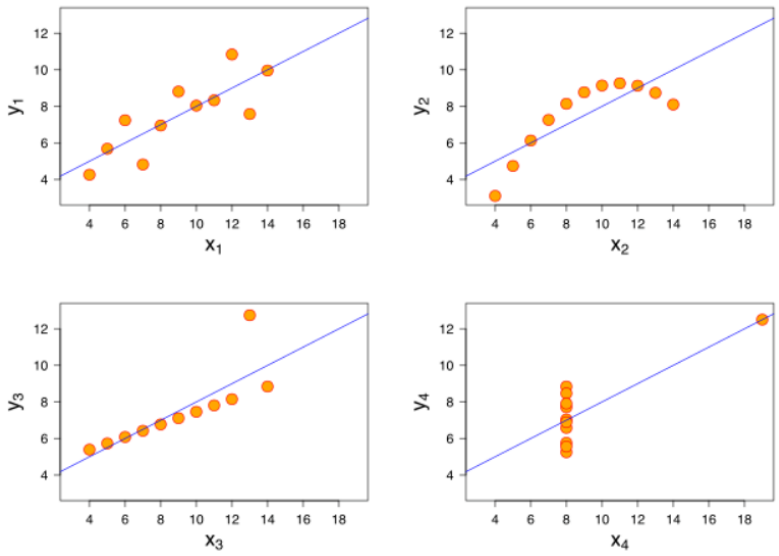


- In this linear regression machine learning algorithm, we train the model using training dataset using significant variables to make prediction when new data is given to the model.

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

- Anscombe's quartet comprises 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 analyzing it, and the effect of outliers and other influential observations on statistical properties.
- For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s^2_x	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y : s^2_y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places



All 4 sets are identical when examined using summary statistics but vary considerably when graphed

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one 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.

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

- Pearson's R is Pearson correlation coefficient (PCC) which 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.
- The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1.[15] A value of 0 implies that there is no linear dependency between the variables.
- Pearson correlation coefficient is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient.

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

- Scaling is very important pre-processing steps.
- If variables are at very different scale, interpreting coefficient is very difficult (coefficient of lower value variable will be larger than coefficient of variable with larger value, also we cannot say field having larger coefficient is strong predictor.)
- So in any model, for coefficients to be interpreted, all variable must be at comparable scale. so that coefficients are comparable)
- Mathematical reason if scaled between 0 and 1, the optimization behind the scene is faster, gradient descent function tries to minimize the cost function, so minimization routine becomes much more fast.

a) Min-Max scaling (normalisation): Between 0 and 1
b) Standardisation (mean - 0, sigma - 1)

Additional info and differences:

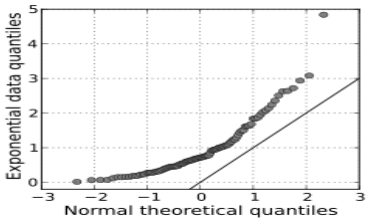
- normalisation: (x-x_min)/(x_max - x_min) - data is compressed between 0 and 1
- standardisation: (x-mu)/Sigma; (mu - mean; sigma is standard deviation --> mean will be zero and std dev will be 1)

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

- When VIF is infinite then there is perfect correlation between two variables (multicollinearity).
- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of 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 (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. A below Q Q plot showing the 45 degree reference line:



- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.