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?

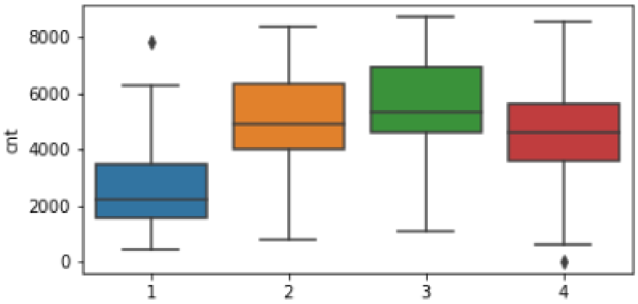
   Dependent variable “cnt” is related to all the predictor as per the below equation:

   cnt = 0.1451 + yr(0.2330) - holiday(0.0910) + temp(0.4042) -hum(0.1287) - windspeed(0.1776) - spring(0.1483) - july(0.0802) + september(0.0584) + misty(0.1811) + clear(0.2418)

   Categorical variables from the dataset:

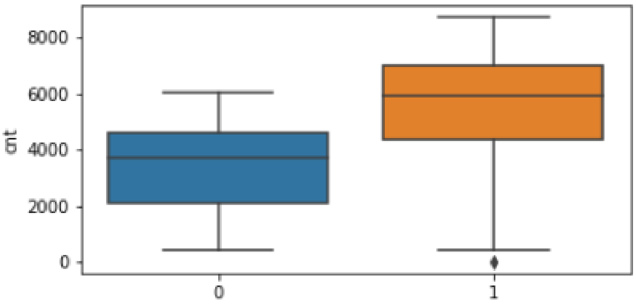
   a. Season:

      Spring season has negative impact on the users of bikes when rest of the predictors are kept constant. This has linear coefficient of -0.1483. Other categories of summer fall and winter doesn’t have influence on the bike usage. This can be visualized also



   b. Year:

      Bike users increase from year 2018 to 2019 therefore has the positive impact. This is true from the data visualization and also from the model building activity. Year categorical variable has the linear coefficient of 0.2330



   c. Month

      Bike users seem to dip in the month of July. It has a negative coefficient of -0.802. This was not captured in data visualization directly.

      Bike users increase in the month of September if the remaining predictors are kept constant. This has the positive coefficient of 0.0584

   d. Holiday

      A holiday has negative impact on bike users with a coefficient of -0.0910. This is intuitively acceptable as holiday leads lesser necessity of bikes for business purposes

   e. Weekday

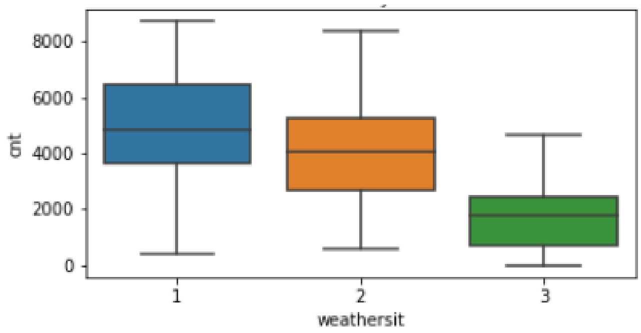
      Days of the week doesn’t seem to have impact directly on the users of bikes. Important factor is holiday/working day that working weekdays and weekends

   f. Working day

      This is opposite of holiday and inverse explanation of the holiday holds

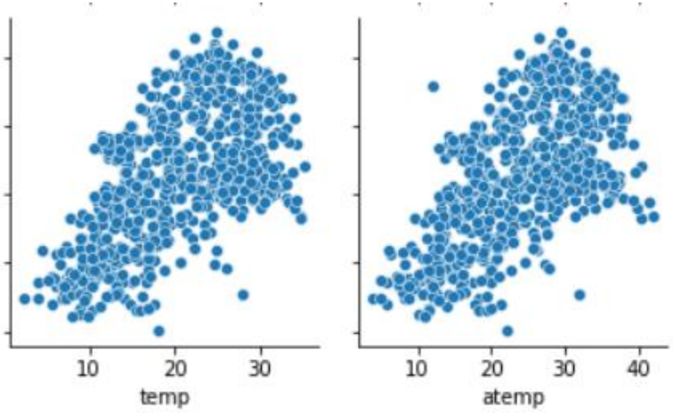
   g. Weather situation

      Misty and clear weather situation attracts more bikers. This is also plausible with data visualization

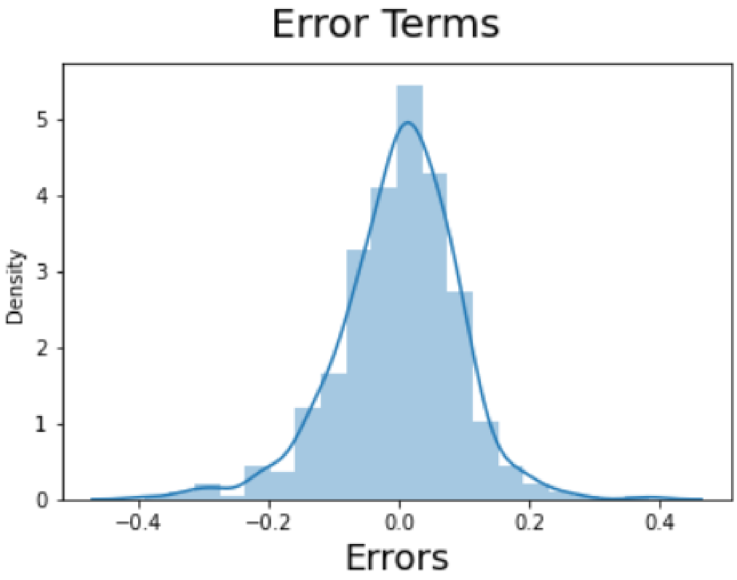


2. Why is it important to use **drop_first=True** during dummy variable creation?
   Lesser predictor variables to build the model are advantageous in terms of computational runtime and simplicity. Dropping one of the dummy variables created out of a categorical variable is an effective way of reducing predictor dummy variable without losing the information. Formula for creating dummy variable is (n-1) dummy variables for a categorical variable of n levels.
   Underlying fact with this formula is that nth level of categorical variable can be represented by values of remaining n-1 levels by n-1 dummy variables

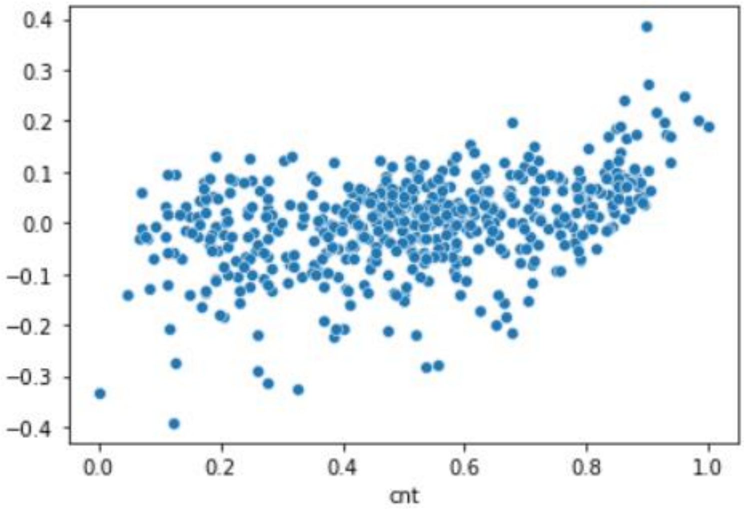
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Feeling temperature in Celsius or temperature in Celsius has the highest linear co-relation with target variable cnt



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Linear regression assumptions
   1. Zero mean of the error terms
      This is validated by plotting the residual errors distribution to know the centre of the distribution to be around zero. Example from the fitted model look like below



2. Normal distribution
   Approach is same as that of zero mean. Plot of residual errors are observed for normal distribution. Example plot from the fitter model is same as above
3. Error term independency
   This is observed by plotting the error terms over y actual. No pattern in the plot will be considered to prove independency



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

   Model equation:
   cnt = 0.1451 + yr(0.2330) - holiday(0.0910) + temp(0.4042) -hum(0.1287) - windspeed(0.1776) - spring(0.1483) - july(0.0802) + september(0.0584) + misty(0.1811) + clear(0.2418)

   Top 3 predictor variable from the equation based on the magnitude of the coefficient are
   1. Temperature → Positive impact. Warmer the weather higher the bike users
   2. Year → Positive impact. Bike users increase year by year provided all the other conditions remain same as the previous years absorbed in dataset
   3. Clear weather situation → Positive impact. Clear weather attracts more bikers

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Linear regression algorithm is used to establish the linear relationship between the predictor and target variable. There are two types of linear regression
   a. Simple linear regression
      This is linear relationship between a predictor and target variable explained by straight line equation
      Y=mX+c where,
      Y is the target variable to be predicted
      X is the predictor variable and
      C is the intercept
      M is the slope of the straight line
      This equation can be interpreted as the Y changes by m units for ever unit change in X
   b. Multiple linear regression
      This is linear relationship between multiple predictors and target variables explained by straight plane fit
      $Y=c+m_1X_1+m_2X_2+\ldots\ldots+m_nX_n$ where
      Y is the target variable to be predicted
      Xn is the nth predictor variable and
      C is the intercept
      Mn is the coefficient of the Xn predictor variable
      This equation can be interpreted as Y changes by m_n units for every unit change in Xn where n is from 1 to n → reference 1

      For this algorithm to fit or predict well below linearity assumption is made between the predictor and target variable

      1. Variance in the target data is symmetric across the predictor points in other words error terms from the fitted line/plane have zero mean and equal distribution
      2. Each of the target data points are independent from its previous data points, in other words error terms from the fitted line/plane doesn't show any pattern

      For multiple linear regression equation to interpret it is necessary to have no co-linear predictor variable for the explanation made in reference 1 above to hold good.

      Linear regression algorithm fits a line by minimizing the cost function of Residual sum of squares (RSS) as shown in the equation below. This achieved typically by gradient decent algorithm which runs iteratively to achieve minima of the error terms
      RSS = $\sum_{i=1}^{n}(Yi - Zi)^\wedge 2$
      i→ number of the data rows
      Yi→actual target variable in the dataset for the ith row
      Zi → predicted target variable from the linearly fitted module for the ith row

      Finally R2 an relative factor is used to understand the data explanation capacity/coverage of the fitted model to the given data set. Higher the R2 vale better is the coverage of the fitted model. This is provided by the following formula
      R2 = 1-(RSS/TSS)
      Where RSS is the residual sum square and TSS is the total sum of squares

      R2 equal to around 1 typically means an over fitted module

2. What is Pearson’s R?
   Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Below is the formula

$$\rho = \frac{cov(X,Y)}{\sigma x * \sigma y}$$

   Where cov(X,Y) is the co-variance between X and Y
   σ is the standard deviation

   Value of the pearson’s R varies between -1 to +1. -1 means a high negative correlation and +1 means high positive correlation. 0 means no correlation between data points X and Y

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a. Scaling means changing the range of the data without altering the information
   b. Scaling is performed in linear regression context to get better
      - computing time during learning a model and predicting
      - Model coefficient derived from the scaled predictor variables will be easy to interpret. For example in minmax scaling, higher coefficient value means greater value in prediction of the target variable
   c. Difference between normalized and standardized scaling

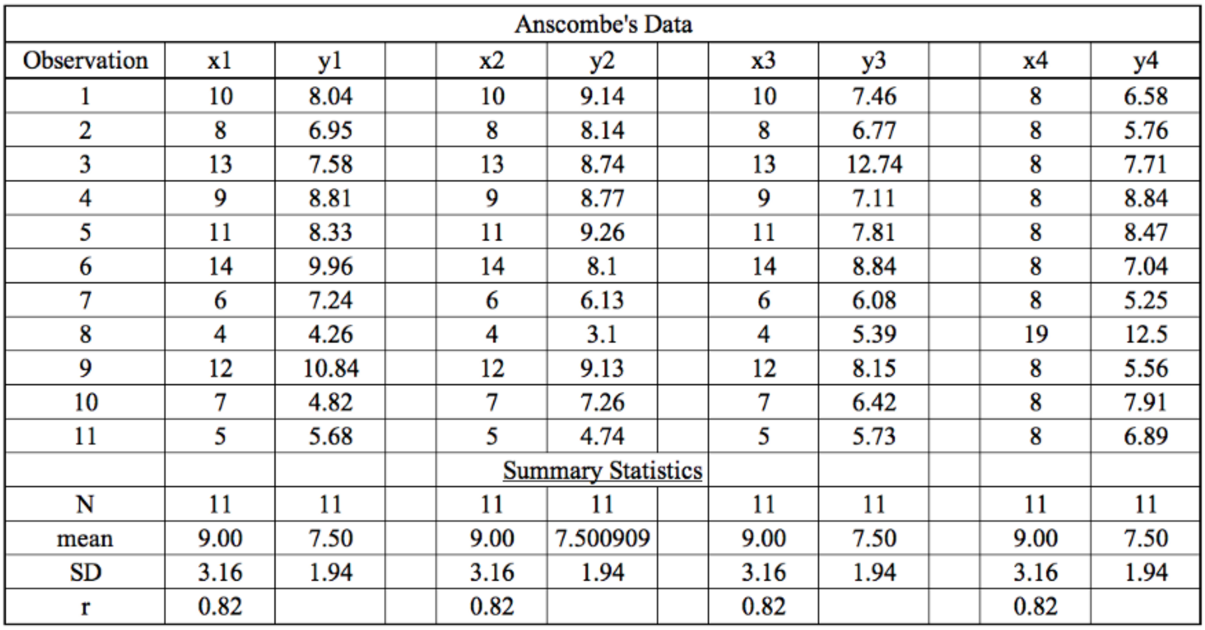
S.NO.	**Normalisation**	**Standardisation**
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Yes, this was observed in the assignment model. This can happen when the predictor variable whose VIF is infinite, can be expressed by other predictor variable very well that a fitted model between the predictor variable whose VIF is infinite and the rest of the predictor variables has a R2 value close to 1
   VIF is given by the following formula
$$VIF = \frac{1}{1 - R^\wedge 2}$$
   Where R^2 is the R^2 value calculated for the linear model fitted between predictor variable under focus with rest of the predictor variables.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. 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.

   Q-Q plot can be used in linear regression to know if predicted target variable from the model and the actual values of target variable have the same distribution for the given set of the input variables. If they have the same distribution which means the error is less in the predicted model and the resulting QQ plot will have points approximately on the reference point.

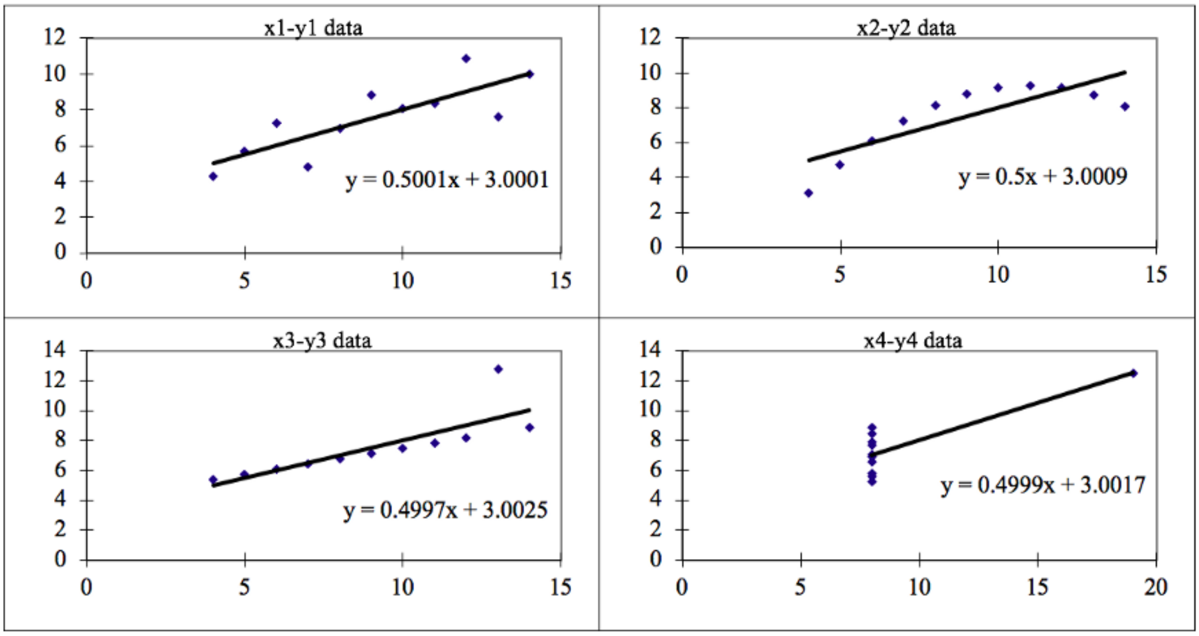
6. Explain the Anscombe's quartet in detail.
   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
   It was constructed in 1973 by statistician Francis Anscombe to illustrate
   the importance of plotting the graphs before analyzing and model building



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	

Above table provides the Anscombe's data points as 4 instances where the summary statistics remain exactly same. Typical inference from such descriptive statistics is that a model built using X1 and Y1 can be used to predict the y2 y3 and y4 using x2 x3 and x4.

However this is not true entirely because of the data distribution. This can be understood by visualization of the data distribution



From the above picture it is evident that fitted lines are similar in all the four datasets. However the error terms or the explanation of the fitted model on the given data is different in different datasets and hence the summary statistics is not sufficient alone to decide correctness of the predicted model.

Dataset1 → model fit is decent

Dataset2 → linear model is fit to a non linear relation

Dataset3 → outlier effect can be seen at the right end of the graph

Dataset4 → outlier influence is heavy on the fitted model. Any input values after ~7 is predicted by the straight line where as the dataset doesn’t have any points between ~7 and ~19 that explains the linear relation ship