Bike sharing Assignment Questions

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?

   - Company should focus on expanding business during fall season.
   - Most bookings are done during May, June, July, Aug, Sept and Oct.
   - Clear weather attracted more bookings.
   - There is more booking between Thursday-Sunday.
   - 2019 attracted more bookings.

2. Why is it important to use drop first=True during dummy variable creation?

   - drop first=True helps in reducing extra columns created during dummy variable creation.
   - It also reduces correlation among dummy variables.

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

   - Temp variable has the highest correlation with the target variable.

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

   - Validated the assumption on Linear Regression based on 5 assumptions:
     - Normality of errors- Error terms should be normally distributed.
     - Multicollinearity check- There should e insignificant multicollinearity between variables.
     - Linear relationship validation
     - Homoscedasticity- there should be no visible pattern in residual values.
     - Independence of residuals- No auto- correlation

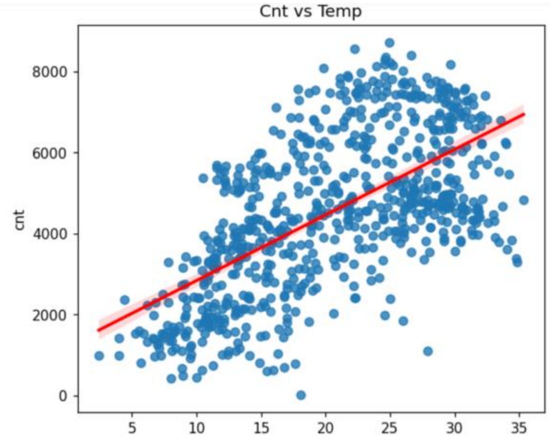
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

   - Temperature
   - Winter
   - September

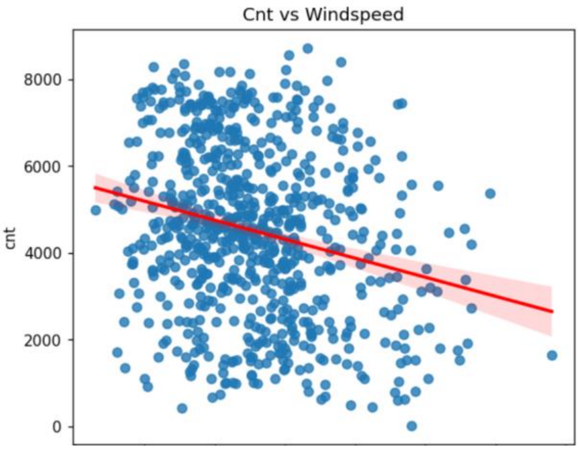
General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a statistical model that analyses the linear relationship between a dependent variable with a given set of independent variables.
- Linear relationship means that when the value of one or more independent variables increases or decreases, the value of dependent variable changes accordingly.
- Mathematically, a linear model is represented by:
  Y= mX+c
  Where,
  Y= Dependent variable
  X= Independent variable
  m=slope
  c= y-intercept
- Linear regression can be positive or negative:
  - Positive linear regression- where both dependent and independent variable increases.



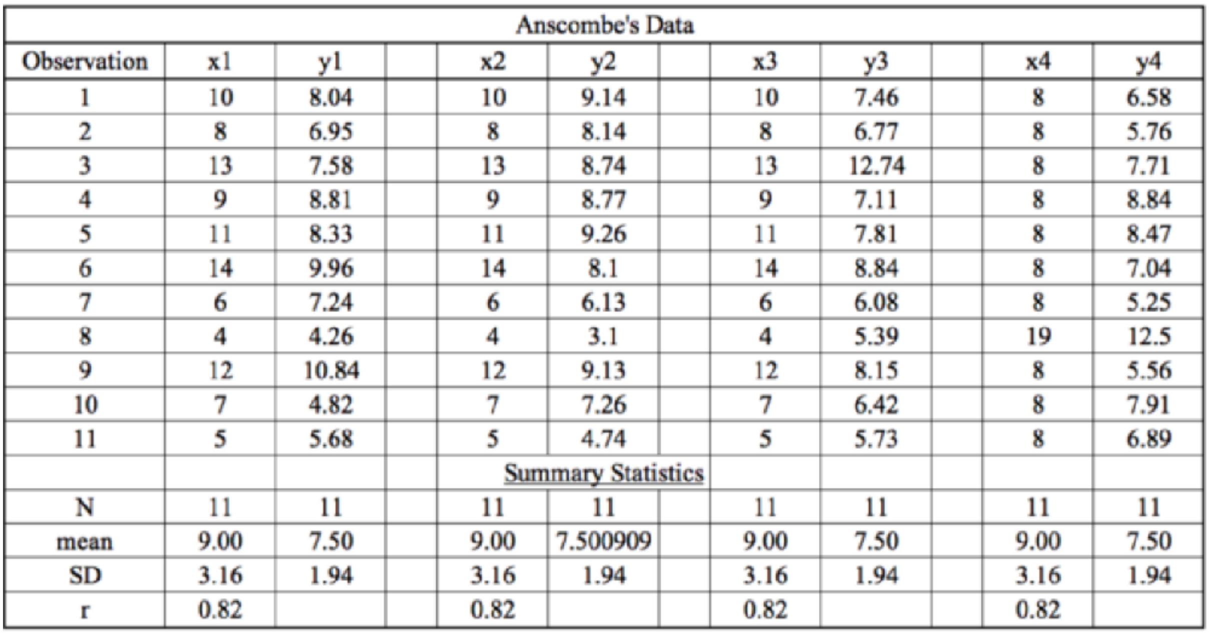
  - Negative linear regression- when independent variables increases and dependent variable decreases.



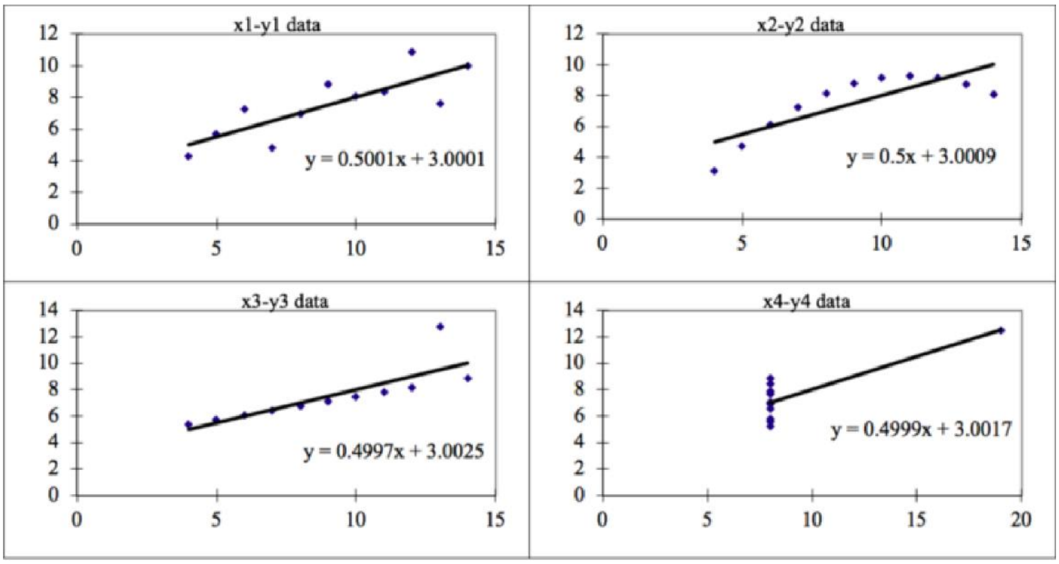
- Linear regression is of two types: Simple and Multiple.
- Assumptions made by linear regression model:
  - Multicollinearity- model assumes that there is very little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are co-dependent.
  - Auto-correlation- this occurs when there is dependency between residual errors.
  - Model assumes that relation between variables must be linear.
  - Error terms should be normally distributed.
  - There should be no visible patterns in residual values.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet was developed by statistician Francis Anscombe.
- It comprises of four datasets each containing eleven (x, y) pairs.
- The datasets share the same descriptive statistics.
- When these datasets are plotted, they look very different from each other. Anscombe's quartet intended to counter the impression among statisticians that numerical calculations are exact, but graphs are rough.

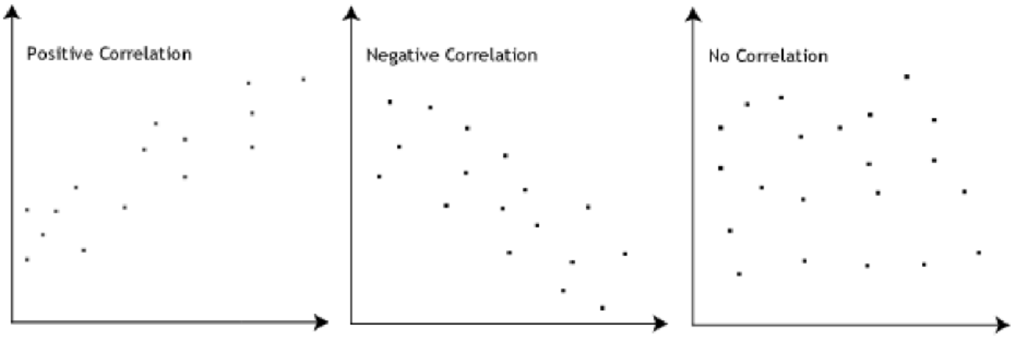


					Anscombe's Data						
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
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	



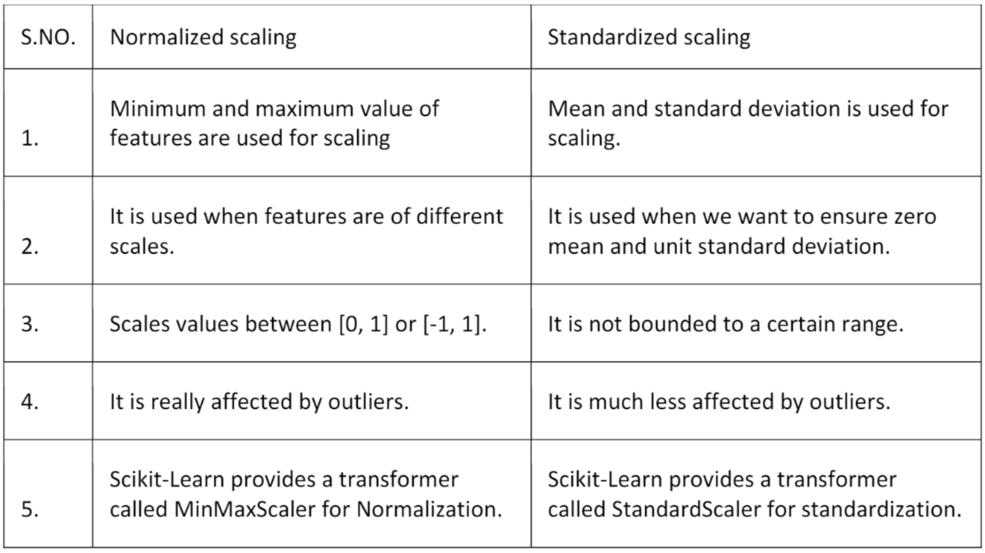
- First scatter plot appears to be in a linear relationship.
- Second, cannot fit the regression model.
- Third, relationship is linear, but should have a different regression line.
- Fourth, shows outliers, that cannot be handled by a regression line.

3. What is Pearson's R?
   - It is a numerical summary of strength of the linear association between variables.
   - If the variables go up and down together, the correlation will be positive.
   - The Pearson R ranges between +1 and -1. A value of 0 indicates that there is no association between the two variables. A value >0 indicates positive correlation.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - Scaling is a technique used to standardize independent features present in the data in a fixed range.
   - It is performed during data pre-processing.

- If scaling is not done, then machine learning algorithm tends to weigh greater values, higher and considers smaller values as lower values, irrespective of their unit.



S.NO.	Normalized scaling	Standardized scaling
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.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - If there is a perfect correlation, VIF= Inf.
   - A large VIF means there is a correlation between variables.
   - When VIF=Inf, it shows perfect correlation between two variables. Here, the R-squared=1, which will lead to 1/ (1-1)=1/0=infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   - A quartile-quartile (q-q) is a graphical technique for determining if two data sets come from populations with a common distribution.
   - A q-q plot is a plot of the quantiles of the first dataset against the quantiles of the second dataset.
   - By quantile we mean the fraction or percent of points below the given value.
   - A 45-degree reference line is also plotted. If two sets come from a population with the same distribution, points should fall along this reference line.
   - Importance of Q-plot:
     - When there are two datasets, it is desirable to know if the assumption of a common distribution is justified.
     - Id datasets differ, we can get some understanding on the same.
     - It can provide more insight than chi-squared tests.