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?

I have done some analysis on categorical column using Boxplot ,Pie chart and scatterplot. Below points are inferred from their effect on the dependent variable

- Company can expand and grow their business in fall, summer and winter season mainly.
- In spring shared bike rental is lowest.
- From the month of April to October more than 80% Bike Rental happen.
- Specifically Bike Rental has most demand in the month of September.
- In 2019 the demand of shared Bike rental has increased like double, so expected like post covid also the demand can grow like this.
- There will be less demand in bad weather condition , whereas in clear weather Bike Rental has more demand
- There is no specific demands of Bike Rental sharing in weekdays or weekends

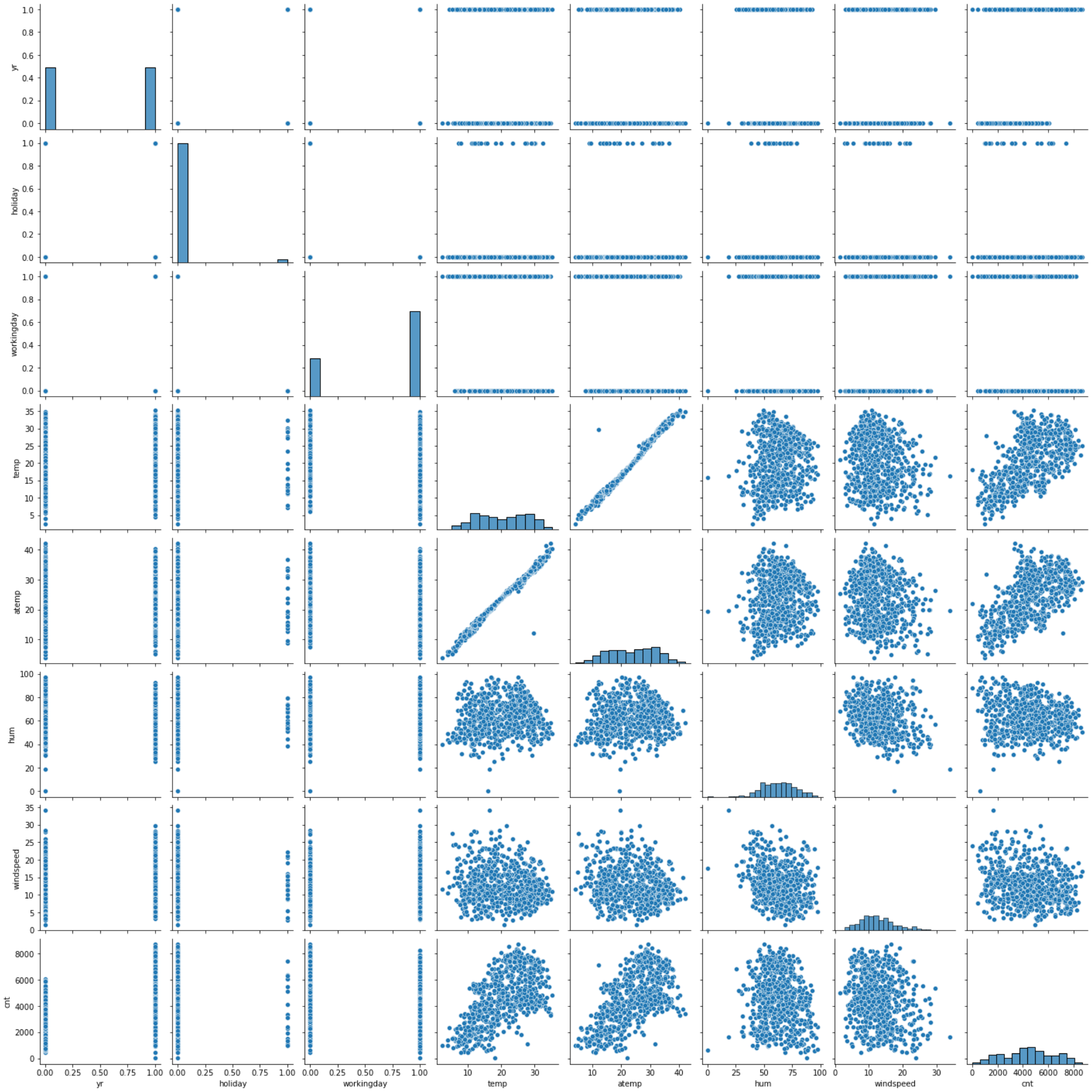
2. Why is it important to use drop_first=True during dummy variable creation?

drop first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.And also to avoid redundancy we are dropping a column. This helps the column to become linearly independent.

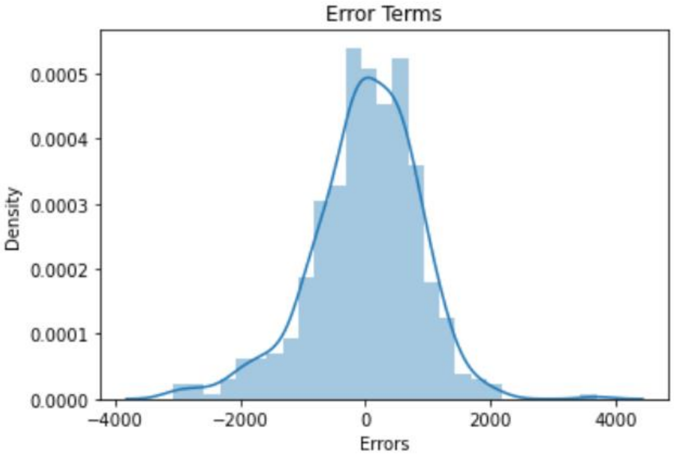
If we do not drop one of the dummy variables, then there would be multicollinearity issue can arise.

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

Temp (or atemp) are the numerical variables, which 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?



Above diagram shows -Errors are normally distributed with mean value zero, and that is fine. So we can conclude like below-

- Error terms are normally distributed with mean zero
- Error Terms do not follow any pattern
- Multicollinearity check using VIF(s).
- Ensured not overfitting by looking the R-Square value and Adjusted R-Square

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

Temperature, humidity and Windspeed/Weather these are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, rather than trying to classify them into categories. It is a part of regression analysis. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

Simple Linear Regression - Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y = mx + b$$

where,

y = dependent variable

x = independent variable m =

intercept of the line

b = linear regression coefficient

When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

Assumptions of Simple Linear Regression –
There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

Multiple Linear Regression - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z)=w1x + w2y + w3z$$

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation.

Assumptions of Multiple Linear Regression –

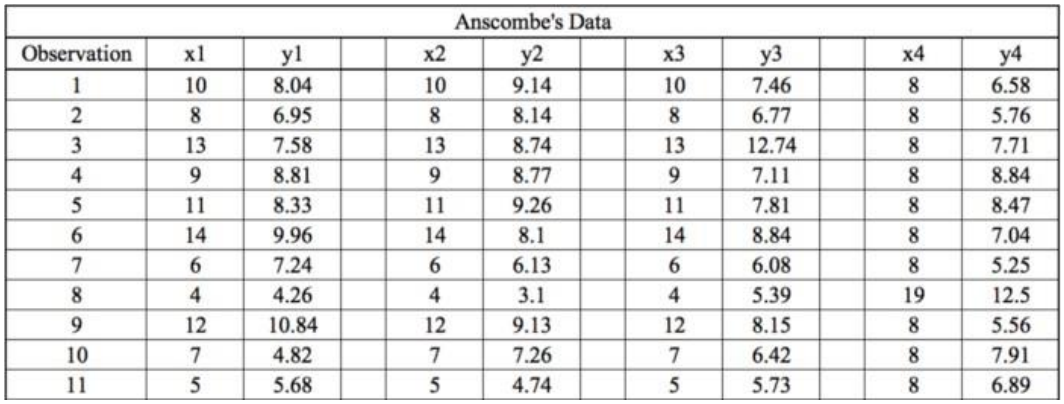
Multiple linear regression analysis makes five key assumptions:

1. **Linear relationship:** There exists a linear relationship between each predictor variable and the response variable.
2. **No Multicollinearity:** None of the predictor variables are highly correlated with each other.
3. **Independence:** The observations are independent**.**
4. **Homoscedasticity:** The residuals have constant variance at every point in the linear model.
5. **Multivariate Normality:** The residuals of the model are normally distributed.

2. Explain the Anscombe's quartet in detail.

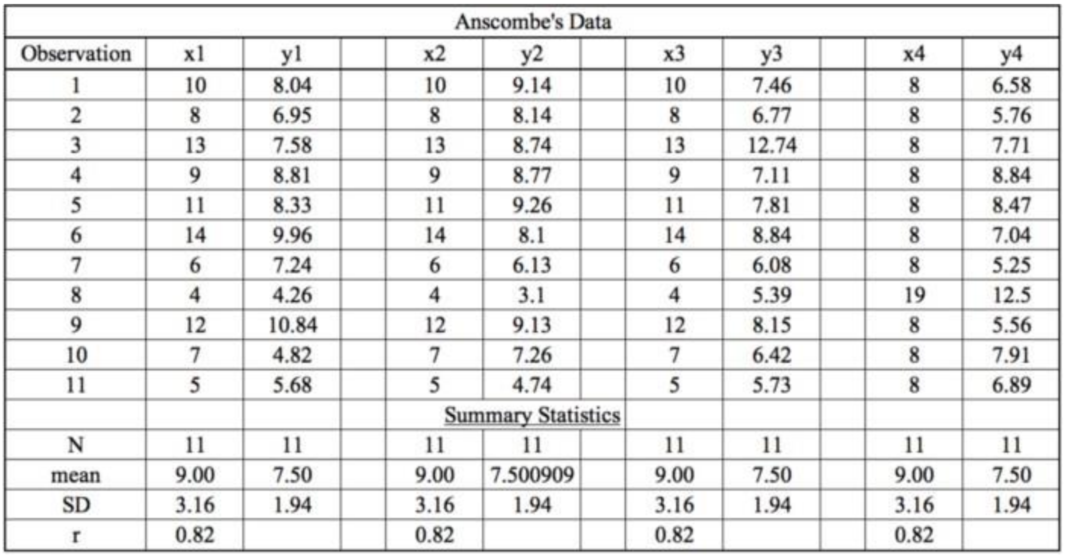
Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe’s quartet intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

These four plots can be defined as follows:



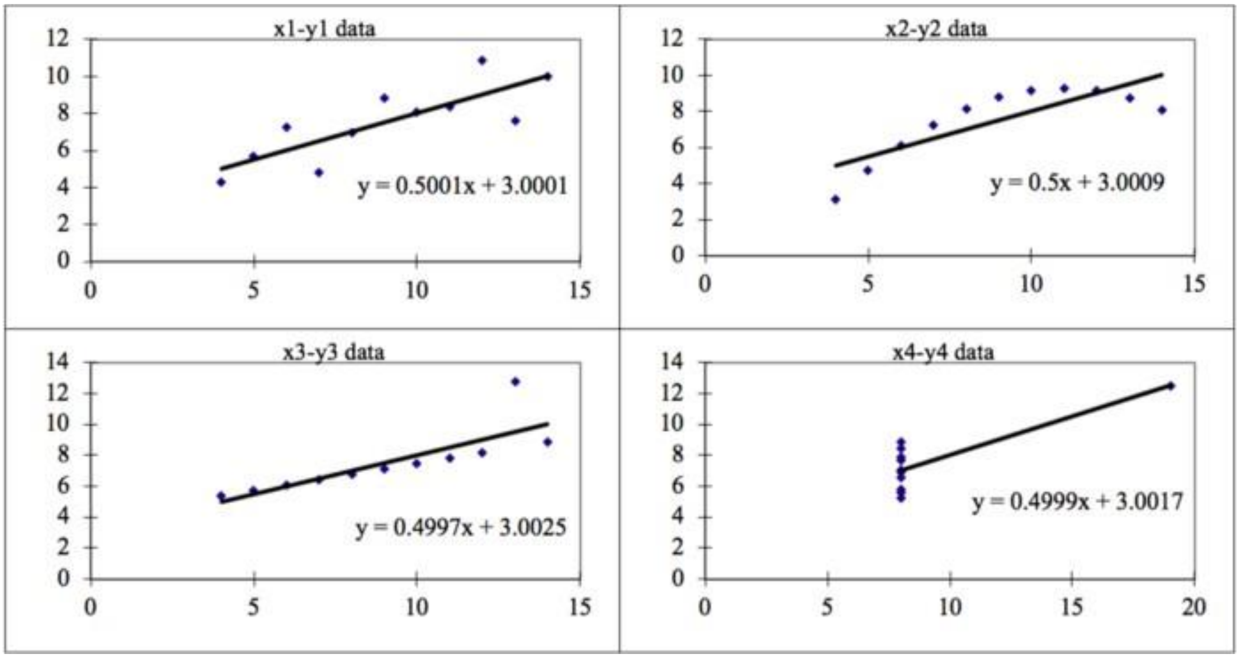
					Anscombe's Data						
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25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

The statistical information for all these four datasets are approximately similar and can be computed as follows:



					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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



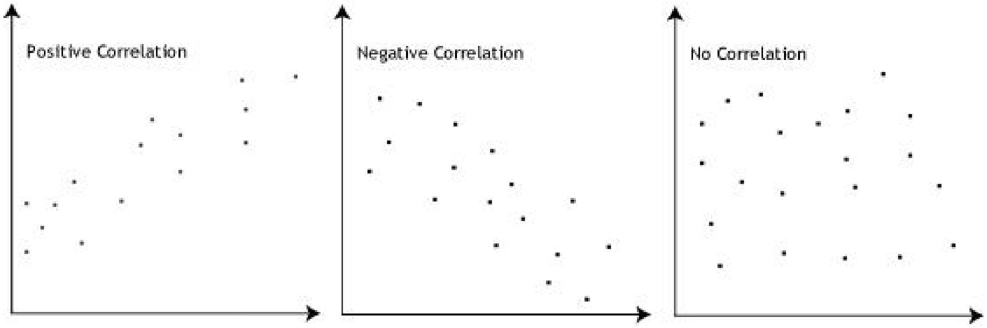
The four datasets can be described as:

1. The first scatter plot (top left) appears to be a simple linear relationship,
2. The second graph (top right); cannot fit the linear regression model because the data is non-linear
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Finally, the fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It 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. It shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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?

Variance Inflation Factor (VIF) is an index that provides a measure of how much the variance ofan estimated regression coefficient increases due to collinearity. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. The higher the VIF value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. If there is perfect correlation, then **VIF** = **infinity**. An infinite VIF value means that the variable is exactly linear combination of other variable. If the 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 one. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

Quantile - Quantile (Q-Q) 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 distribution. Also, it helps to determine if two data sets come from populations with a common distribution. 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. The power of Q-Q plots lies in their ability to summarize any distribution visually.

The advantages of the Q-Q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested.

Q-Q plot is very useful to determine:

1. If two populations are of the same distribution
2. 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.
3. Skewness of distribution