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

1.A

a. Season - fall had the highest bike rentals with lowest for spring. Winter and summer had median usage from bike rentals.

b.weathersit - Bike usage was highest when skies were clear and no snow and gradually decreased as the weather moved to misty and low snow and thunderstorms

c. mnth - It peaked in month of September reducing around Dec-Jan and again increasing in mid-year during fall and summer time.

d. Holiday - Usage is more when there is no holiday

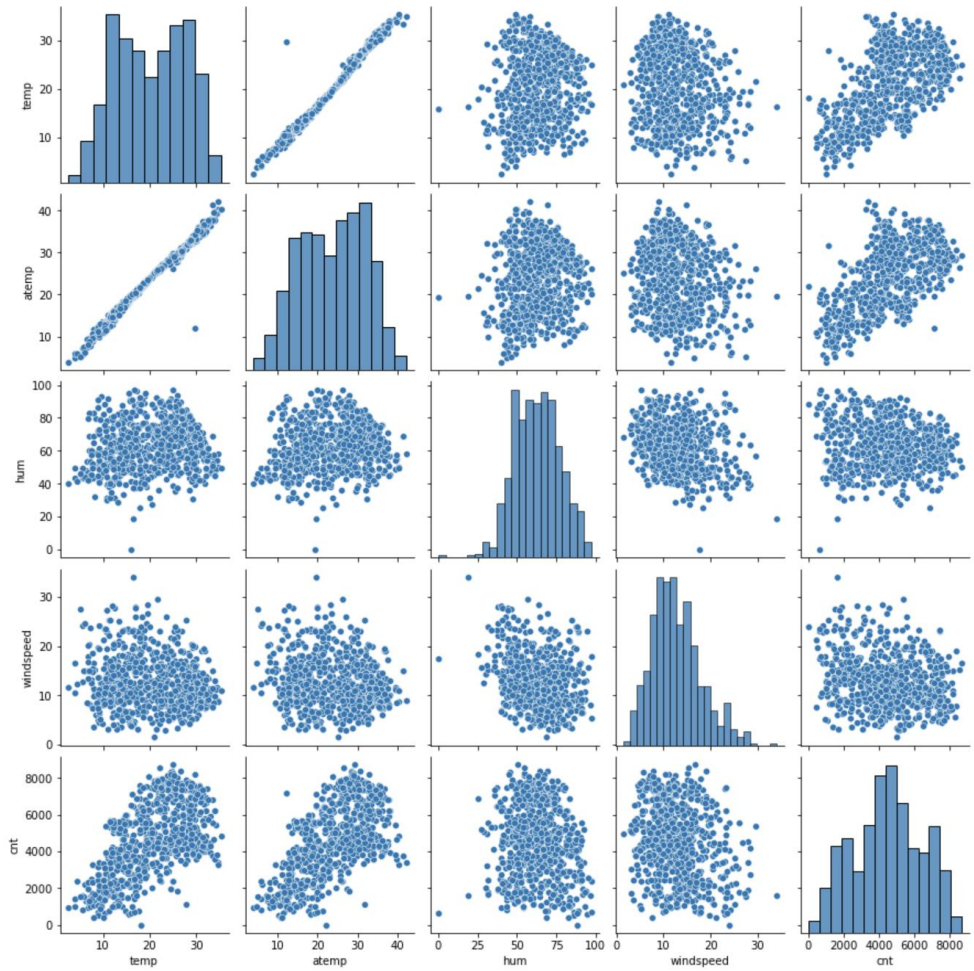
e. Yr - Usage is more in 2019 than in 2018

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

2.A It leads to increased multicollinearity.Otherwise the variables will be co-related and their VIF values will go to infinity and causing more confusion among predictors. So it is better to let the first column be dropped.

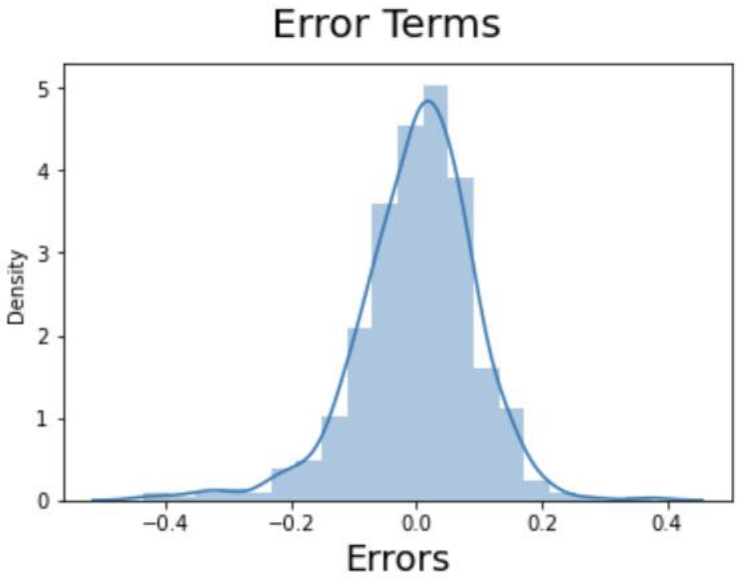
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

```
plt.show()
```



atemp and temp are highly co-relate

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



The above Residual distribution show that model is good fit as the errors on training set shows normal distribution as expected.

a. R2- 82%
b. Adjusted R2 - 81.1%
c. None of p-values greater - .04
d. Variables are Independent with less than VIF < 5

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

1. Temp - positive - 0.43
2. Year - positive - 0.23
3. Light snow - most negative - -0.30

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression =is a model that assumes a linear relationship between the input /predictor variables and the single output/target variable (y). Value for target variable (y) can be calculated from a linear combination of the input variables (x).
Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

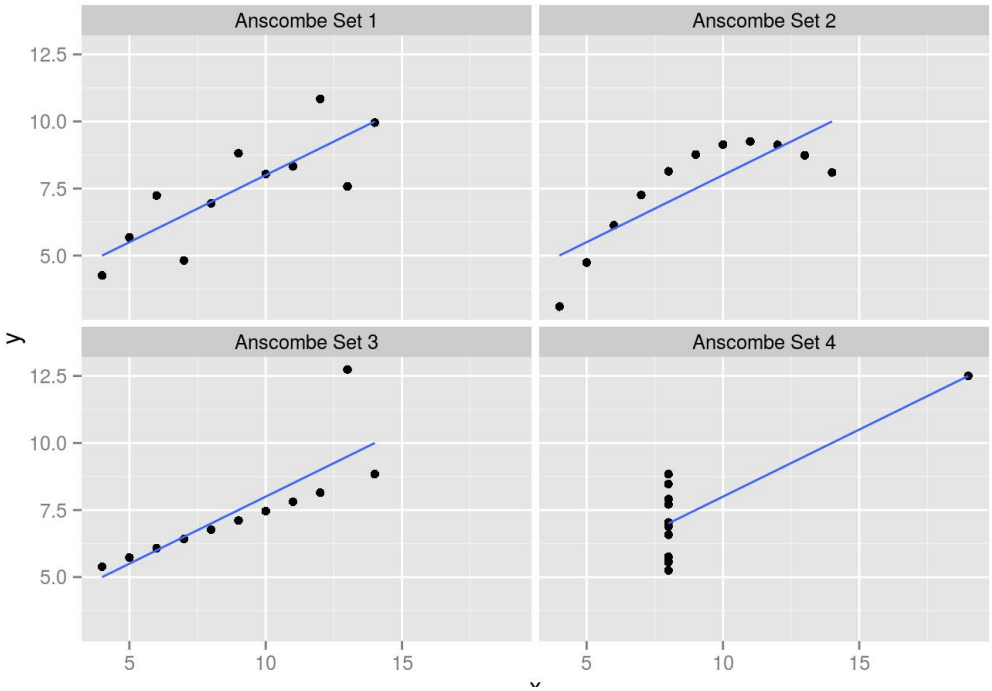
The two types of values we need to figure out is constant(intercept) and coefficients

Types of Linear Regression:

- Simple Linear Regression:
    - There is one target and one predictor variable.
    - Equation is presented by y = mx +c.
    - Values to be calculated are m = coefficient and c=constant
- Multiple Linear Regression:
    - Equation is presented by y = m1x1+m2x2+..mnxn+c
    - Several predictor variables to get target value
    - Linear Relationship is required between predictor and target variables
    - Predictor variables should be independent and should have low co-relation.

2. Explain the Anscombe's quartet in detail

a) Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
b) To illustrate importance of looking at dataset graphically before starting to analyze.
c) To overcome the inadequacy of basic statistic properties for describing realistic data sets.



The four graphs define:

a. Graph 1 defines a linear relationship.
b. Data Set 2 does not have a linear relationship
c. Data Set 3 has a linear relationship but has a outlier
d. Data Set 4 does not fit any linear but outlier keeps going of.

3.What is Pearson's R?

It is a measure of the strength of a linear association between two variables and is denoted by r. It attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit

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.

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

Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].

Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless

Min-Max Normalization(Normalized): This technique re-scales a feature or observation value with distribution value between 0 and 1. For non-Gaussian Distribution

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. For Gaussian Distribution.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. The formula for VIF is given below:

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

When R2 =1 then the denominator will be zero in above formula and hence VIF is infinity. R2 will be 1 if they are perfectly co-related. This usually happens when we don't use dropFirst = True when creating dummy variables.

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.

This is really helpful in scenarios where training data and test data come separately and not as a part of the same data set.

It is used to check if two data sets:

- Come from populations with common distribution
- Common location and scale
- Similar distributional shape
- Have similar tail behavior