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

The categorical variables in the dataset are season, yr, mnth, holiday, weekday, workingday, weathersit, casual, registered:

1. The bookings are less in the month of spring when compared with other seasons
2. The bookings show an increasing trend from 2018 to year 2019.
3. Months Jun to Sep in both the years in general show high booking counts. Jan shows the lowest demand month.
4. Booking counts are less in holidays in comparison to not being holiday.
5. The bookings is almost consistent throughout the weekdays. There is no significant change in bike demand with working day and non-working day.
6. The bike demand is high when weather is clear however demand is less in case of Light snow.
7. Casual bookings are lower during Dec-Feb, while Registered users are almost consistent throughout the year.

==========

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

Ans : The "drop_first=True" is important to use, as it helps in reducing the extra column created during dummy variable creation.

It reduces the correlations created among dummy variables.

For eg : we have 3 types of values in a season column and we want to create dummy variable for that column. If one variable is not Fall, Spring or Summer, then It is obvious winter.

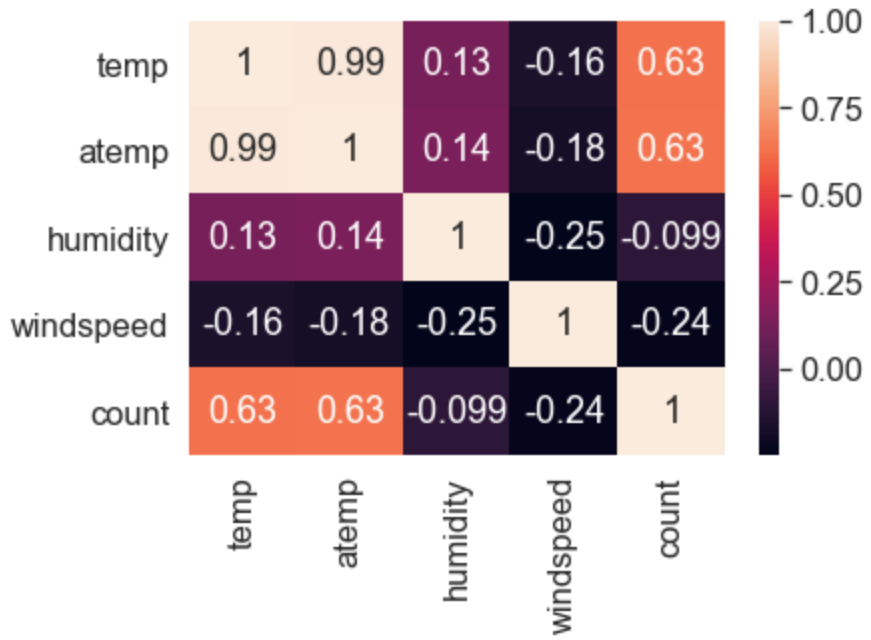
So we do not need 4th variable to identify the "winter" values.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

==========

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

Temperature(temp) has highest correlation with value of 0.63 with target bookings(count) variable.

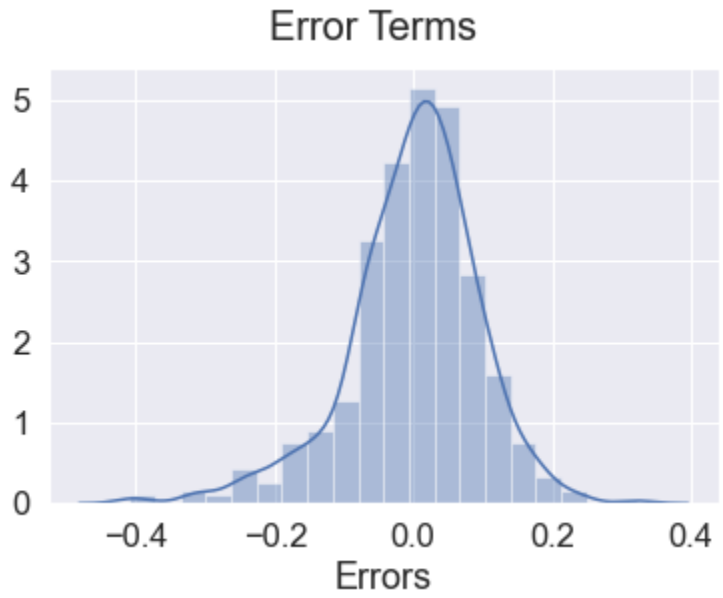


==========

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

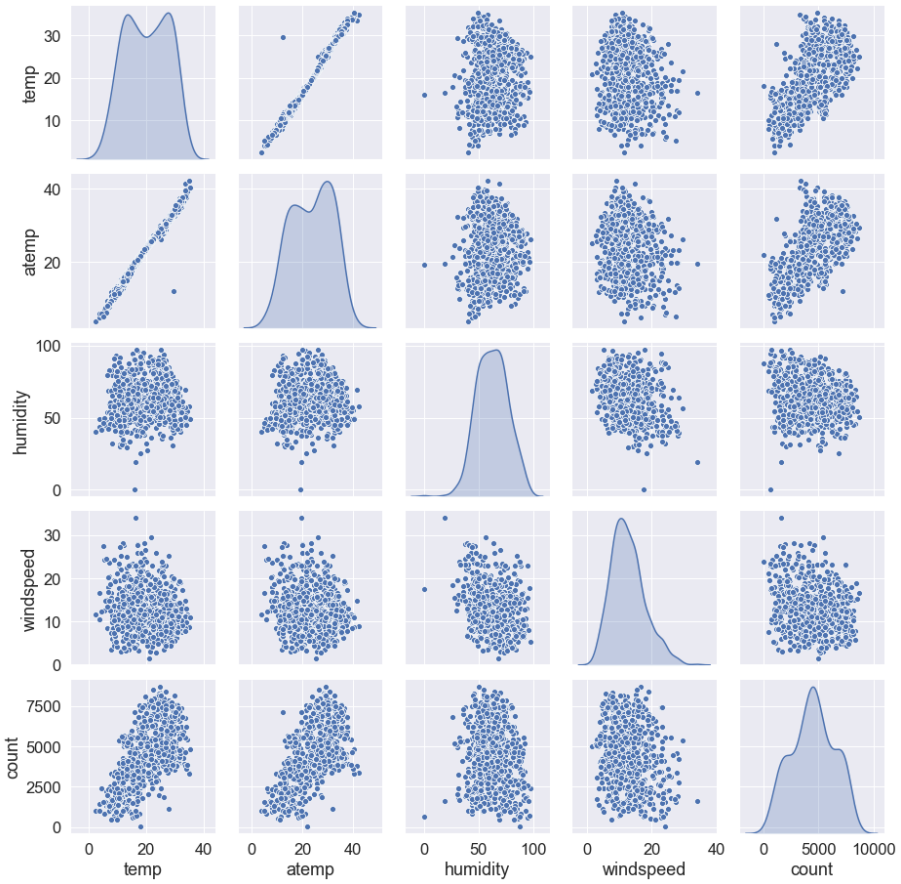
1. Error terms are normally distributed with mean zero (not X, Y)

first find the y_predicted of the X_train using the finalized model, then find (y_train - y_train_predicted)and plot a distplot of errors



2. There is a linear relationship between X and Y

using pairplot of the numerical variables : 'temp', 'atemp', 'humidity', 'windspeed','count' we can identify if they are positively/negatively correlated.



3. There is No Multicollinearity between the predictor variables

Check for the VIF values of the feature variables. From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 10.

	Features	VIF
1	temp	3.89
2	windspeed	3.46
0	year	1.98
3	season_2	1.56
7	weathersit_2	1.48
4	season_4	1.36
5	month_9	1.19
6	weekday_2	1.18
8	weathersit_3	1.07

```
### Conclusion :

From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 10.
```

==========

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

The Best line fit equation =

count = const *0.1269 + year * 0.2310 + temp * 0.5628 -windspeed * 0.1537 + season_2 * 0.0823 + season_4 * 0.1270 + month_9 * 0.0958 -weekday_2 * 0.0331 -weathersit_2 * 0.0732 -weathersit_3 * 0.3050

The top 5 predictor variables that influences the bookings and should be considered while planning are.

- Temperature (temp) =0.5628
- Weather Situation 3 (weathersit_3) = -0. 3050
- Year (yr) = 0. 2310
- season_4 = 0.1270
- month_9 = 0.0958

A positive coefficient value indicated that a unit increase in the variable, increases the bookings counts by "coeff" units.

A negative coefficient value indicated that a unit increase in temp variable, decreases bookings counts by "coeff" units.

General Subjective Questions

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

1.1 Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables.

1.2 The linear regression model provides a sloped straight line representing the relationship between the variables.

1.3 Mathematically, we can represent a linear regression as:

y= a0+a1x+ random_error

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

random_error = random error

1.4 The values for x and y variables are training datasets for Linear Regression model representation.

1.5 We have to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

1.6 The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

Different coefficients (a0, a1) gives the different line, and the cost function uses these coefficients for the best fit line.

1.7 we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values.

MSE=sum(yi-(a1x1+….a2xi+a0))^2)/N

Where,

N=Total number of observation

Yi = Actual value

(a1xi+...a2xi+a0)= Predicted value.

1.8 The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

1.9 We use the R-squared method to find if our model is good.

The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

where,

R^2 = explained variance/total variance

1.10 Assumptions of Linear Regression:

a)Linear relationship between the features and target.

b)Small or no multicollinearity between the features.

c)Homoscedasticity is a situation when the error term is the same for all the values of independent variables.

d)Normal distribution of error terms.

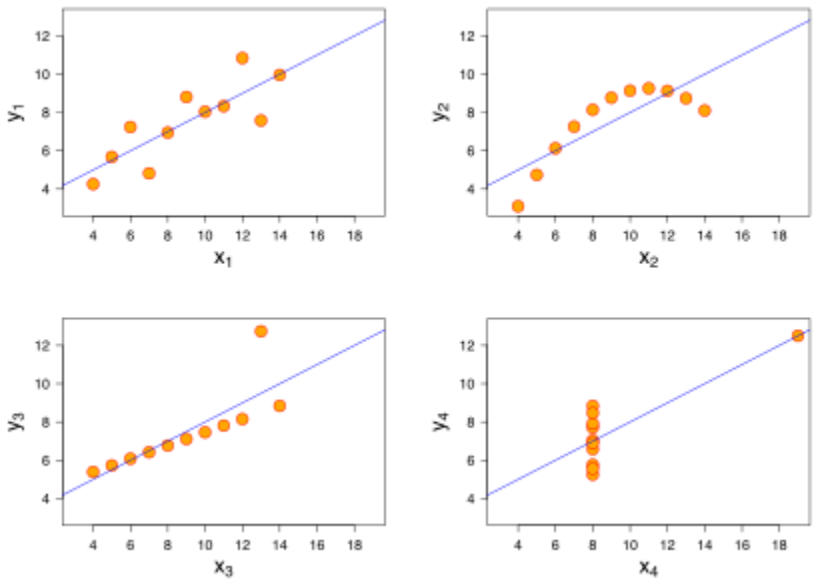
e)The linear regression model assumes no autocorrelation in error terms.

==========

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Mean of x, Sample variance of x ,Mean of y,Sample variance of y,Correlation between x and y ,Linear regression line coefficients are all same for all four datasets.



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) is not distributed normally; 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 distribution 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.

We should analyze a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

==========

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

Pearson's correlation is utilized to see find linear relationship between two numerical variables. Your research hypothesis would represent that by stating that one score affects the other in a certain way. The correlation is affected by the size and sign of the r.
What is Pearson Correlation?

However it is not able to tell the difference between dependent variables and independent variables.

For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. In addition, the value does not give any information about the slope of the line; it only tells you whether there is a relationship.

==========

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

1. Feature scaling is a method used to normalize the range of independent variables or features of data.
2. In data processing, it is also known as data normalization and is generally performed during the data preparation step.
3. We need to scale the variables in our dataset because some machine learning algorithms are sensitive to feature scaling
4. Linear regression uses gradient descent as an optimization technique require data to be scaled.
5. The presence of feature value X in the formula will affect the step size of the gradient descent.
6. The difference in ranges of features will cause different step sizes for each feature.
7. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.
8. Having features on a similar scale can help the gradient descent converge more quickly towards minima.
9. Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitude.
10. This will impact the performance of the machine learning algorithm and it may be biased towards one feature.
11. Normalization :
    - It is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
    - Formula :

      X'=(X-Xmin)/(Xmax-Xmin)

      Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

      When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0

      On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

12. Standardization:
    - Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.
    - This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

X'=(X-mu)/sigma

13. We should use normalization when we don't know anything about the data, or the data distribution does not follow a Gaussian distribution. Standardization is used mostly in cases where the data follows a Gaussian distribution. But data should not have outliers.

==========

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

VIF=1/(1-R^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.

In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features. When R2 reaches 1, VIF reaches infinity

Here we try to remove features for which VIF > 5.

==========

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.

==========