Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

We have following categorical variables:

1. season
2. yr (year)
3. month
4. holiday
5. weekday
6. weekend
7. weathersit

The dependent variable is cnt(count).

Upon doing the analysis using boxplot we can see

1. Fall season has higher count(cnt) followed by summer season.
2. There is a significant rise in count(cnt) from 2018 (0) to 1 (2019).
3. count(cnt) increases from jan till oct and then we see a drop in nov and dec months.
4. days doesn't seems to affect the cnt much (though sat and wed looks to be having higher cnt)
5. When wheather is clear then count(cnt) is higher and when its Light_Snow its drastically low.
6. Also, we don't see any outliers

Boxplot from https://github.com/kc11381/Bike_Sharing_Case_Study/blob/main/bike_sharing_predictors.ipynb

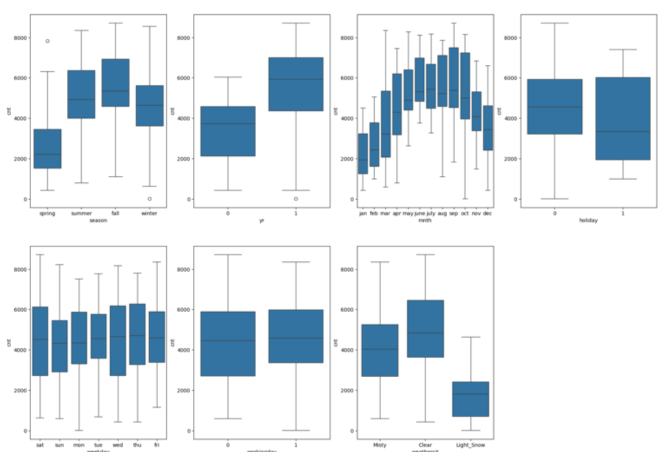


Fig. Boxplot for Cayegorical variables

Question 2. Why is it important to use drop_first=True during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

We use drop_first=True to drop unnecessary columns because we just need p-1 levels for p levels of a categorical variable.

Example- For a variable say, 'Relationship' with three levels namely, 'Single', 'In relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But you can clearly see that there is no need of defining three different levels. If you drop a level, say 'single', you would still be able to explain the three levels.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' in 1 and 'Married' is 0, that means the person is in a relationship and finally, if a 'In a relationship' is 0 and 'Married' is 1, that means that the person is married.

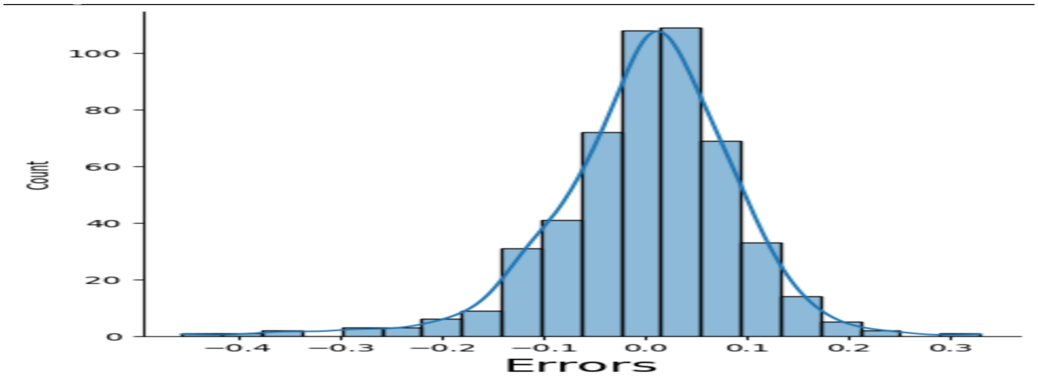
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

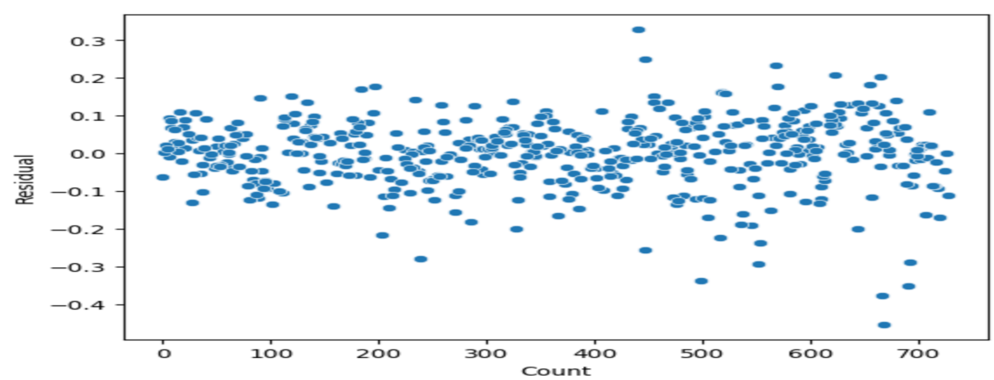
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

temp and atemp has the highest co-relation with the target variable(cnt) by looking at the pairplot. From correlation heatmap value is 0.65

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
We had to check 4 assumptions of the linear regression. They are tested as below
a. Histogram of the error terms and it came out to be a Normal distribution



b. VIF of the features of the final model < 5.0 (except temp as dropping that drastically decreased R-squared value)
c. Multicollinearity - Variables like temp is linearly related to cnt
d. Homoscedasticity – The variance of the residual is almost constant.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
Coefficients of the final model are below

yr	0.2348
workingday	0.0547
temp	0.4354
windspeed	-0.1609
spring	-0.0713
summer	0.0354

winter 0.0903
dec -0.0467
jan -0.0526
july -0.0466
nov -0.0447
sep 0.0652
sat 0.0670
Light_Snow -0.2969
Misty -0.0818

So the top 3 features are:
a. temp
b. workingday
c. yr

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Linear regression is a Supervised Machine learning algorithm in which the output variable is continuous in nature.
Idea here is to predict the value of a dependent variable based on different independent variables.
It is of 2 types:

A. Simple linear regression – Single variable is involved
   y = mx + c
   y – dependent variable which needs to be predicted
   x – independent variable
   c – intercept
   m- slope or coefficient of dependent variable x.
B. Multiple linear regression – More than one variable is involved and hence equation changes to
   y = m1x1 + m2x2+ m3x3+ ….+mnxn + c
   y – dependent variable which needs to be predicted
   xi – independent variables
   c – intercept
   mi- slope or coefficient of dependent variable xi.

To Find the dependent variable we have below steps:

Step 1. Reading, understanding and Visualizing the data

a. Here we read the data and look for its statistics values like mean, max, min etc.

b. We visualize the data using pairplot for continuous variables and boxplot for categorical variables

Step 2. Prepare the data for modelling (test/train split, rescaling etc)

a. Encoding
   a. Convert binary categorical variables to 0 and 1
   b. Convert other categorical variables to dummy variables
      i. For a variable with p levels we will need p-1 dummy variables.

b. Splitting into test and train dataset

c. Rescaling of the variable so that unit is same for all.

Step 3. Training the model

a. Here we create the object of LinearRegression.

b. To choose variables we have 3 approaches
   i . Take one by one
   ii. Take all
   iii. Use RFE (Recursive Feature elimination – need to specify the count at the start)

c. Based on VIF and p-value we eliminate the unnecessary variables.

d. Look for Adj. R-Squared value. Higher the value of it better the model.

Step 4. Residual analysis of the train data

a. Here is check for assumptions of Linear regression using
   a. Error terms should be a normal distribution
   b. Independent variables should not be related to each other
   c. Variance of the residual should be constant.

Step 5. Making predictions

a. We make predictions on test dataset using the model we created with train dataset.

Step 6. Model evaluation against test data

a. We compare the R-Squared values for train and test dataset. It should be close.

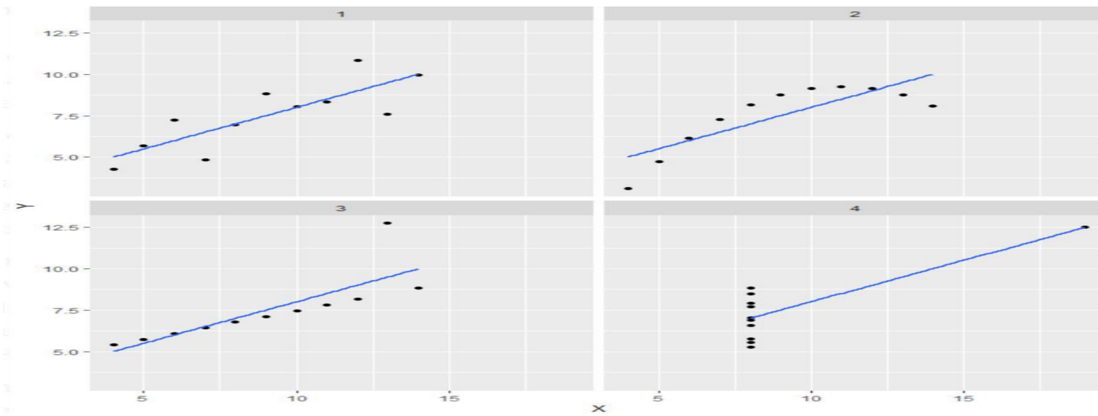
After above 6 steps we get the final model in terms of an equation.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted on scatter plots.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Use:

The quartet is still often used to illustrate the importance of looking at 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.

References:

1. https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2
2. https://www.geeksforgeeks.org/anscombes-quartet/

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's correlation coefficient or Pearson’s R is used to measure the relationship between two continuous variables. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment of the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

How is it calculated:

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Formula is:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

The positive value of Pearson’s correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other.

References:

- https://www.analyticssteps.com/blogs/pearsons-correlation-coefficient-r-in-statistics

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Before any model building, we first need to perform the test-train split and scale the features.

Scaling of variables is an important step because, a variable might be on a different scale with respect to all other numerical variables. Also, the categorical variables that we encode may take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

Scaling doesn't impact the model. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are two common ways of rescaling:

- Min-Max scaling (Normalization) --> compresses all the data between 0 and 1
    - Normalization = (x - xmin)/(xmax - xmin)
- Standardisation (mean-0, sigma-1)
    - Standardisation = (x - mu)/sigma, mu = mean

As a general rule of thumb -- we should use Min-Max scaling, as it takes care of outliers. But there are advantages of Standardisation too.
The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier)

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?   (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear

regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>