Bike Sharing – Assignment

Submitted by

Aisharya Ravichandran

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?

   - Year 1 has more demand than year 0.
   - Weather shows good correlation on dependent variable.
   - Month from April to oct has good sales.
   - Season, mnth, weathersit, workingday are good predictors of cnt.

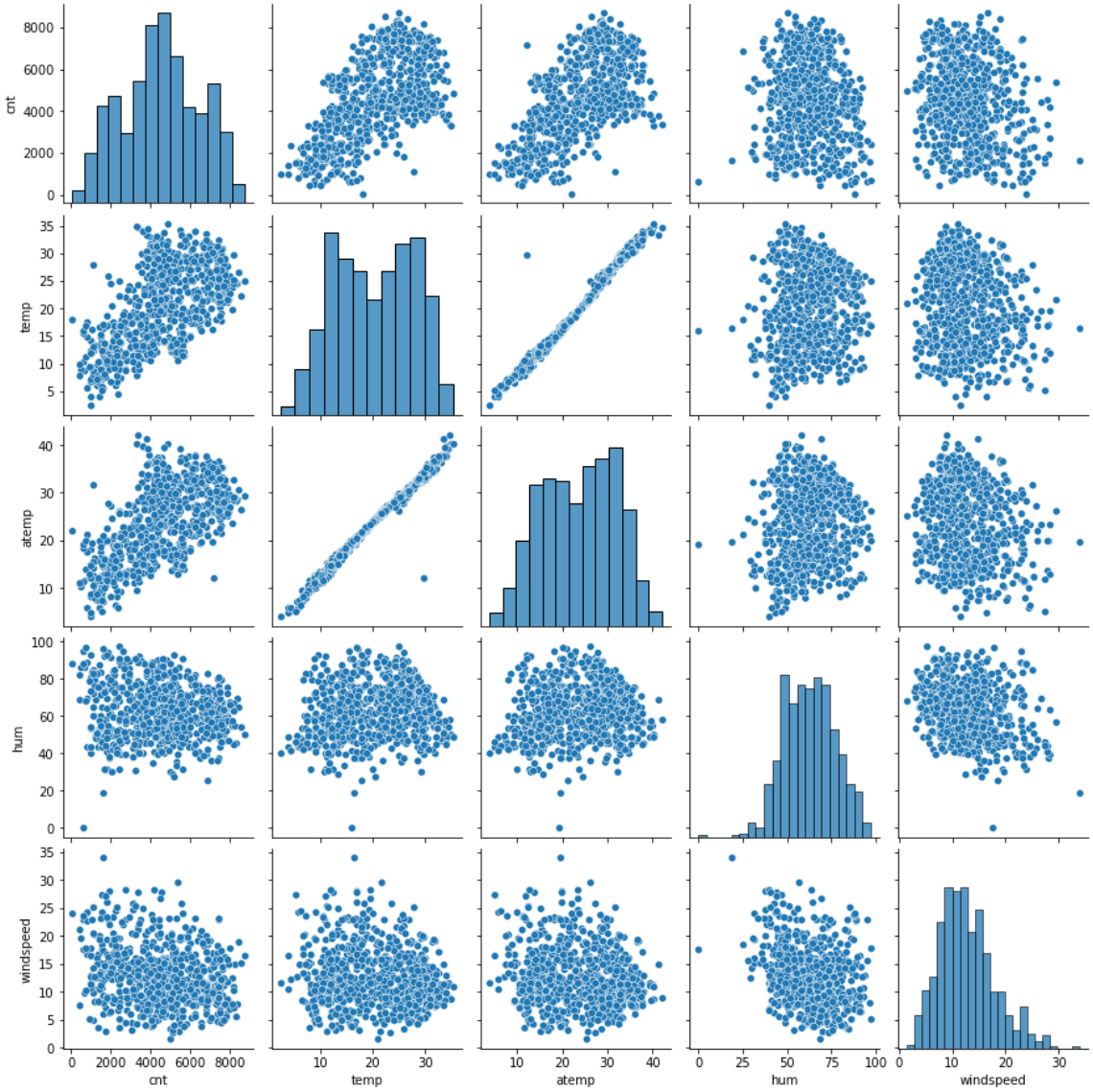
2. Why is it important to use drop_first=True during dummy variable creation?

   In order to avoid redundancy factor (Multicollinearity). drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Eg. We have a column named seasons, where we have 4 categories (summer, winter, spring and autumn). If winter, spring and autumn is false, it actually meaning that the season is summer, so summer can be dropped. (p-1) actually avoids the redundancy within the variables.

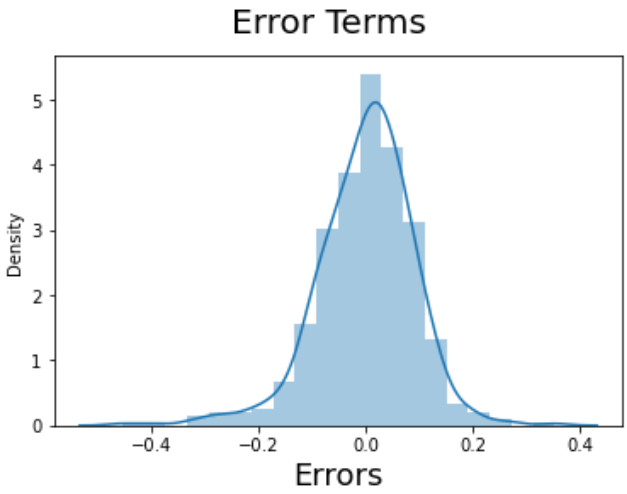
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Looking at the below pairplot, atemp and temp has the highest correlation with target variable CNT. Still atemp and temp has the highest correlation between themselves, hence we could drop either of the variable during model building.

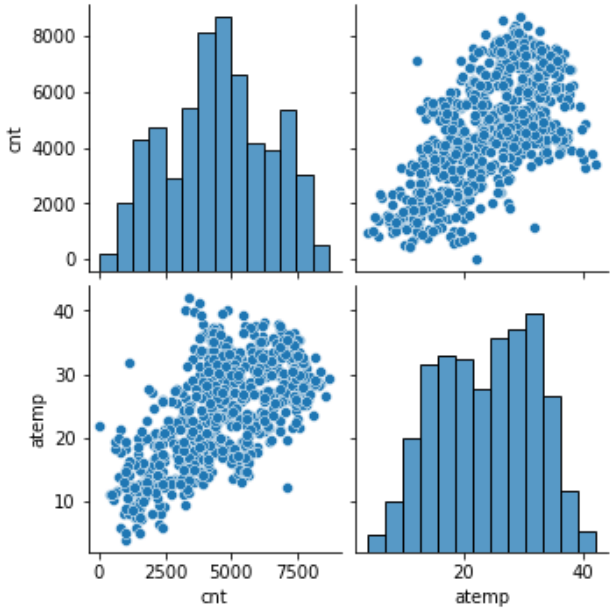


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Checking the distribution of residual erros within mean 0



- Checking the correlation between dependent and independent variables using pairplot. Which should be a straight line.



- Checking the VIF values are within 5, meaning there is no multicollinearity between predictor variables

```
get_vif(cols)
```

	Features	VIF
2	atemp	4.57
3	windspeed	3.95
5	season_Winter	2.55
4	season_Spring	2.35
0	yr	2.06
9	mnth_Nov	1.80
7	mnth_Jan	1.65
12	weathersit_Mist	1.53
6	mnth_Dec	1.46
8	mnth_July	1.35
10	mnth_Sept	1.21
11	weathersit_Light Rain/Snow	1.09
1	holiday	1.06

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

Significant variables:

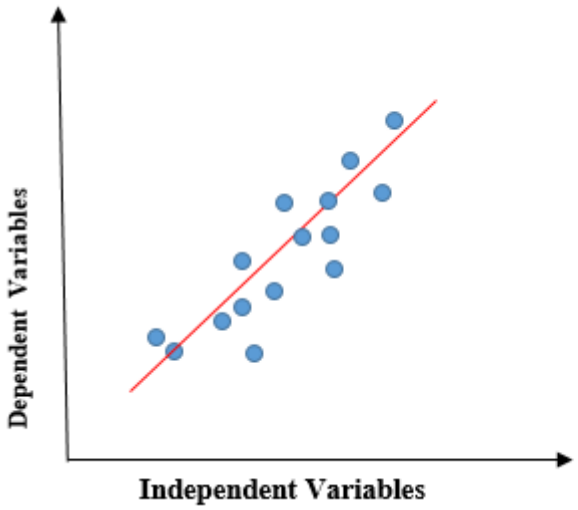
- When atemp increases the demand increases, proving that this variable has high correlation with target variable.
- Demand increases every year
- When the weathersit is bad, then the demand goes down

1. atemp
2. yr
3. season (Winter,Spring)
4. mnth (Sept, Nov, Dec, Jan, July)
5. holiday
6. windspeed
7. weathersit (Mist, Light Rain/Snow)

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear regression strives to show the relationship between dependent and independent variables using a straight line.
It is mainly based on supervised learning (Continuous variables)



When the value of X-axis increases, the values of y-axis also increases.
We can calculate the linear regression using below formula.

$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.

Once the we fit the line, we have to find the whether it is the best fit line. In order achieve it, we calculate the R-squared value using RSS and TSS.

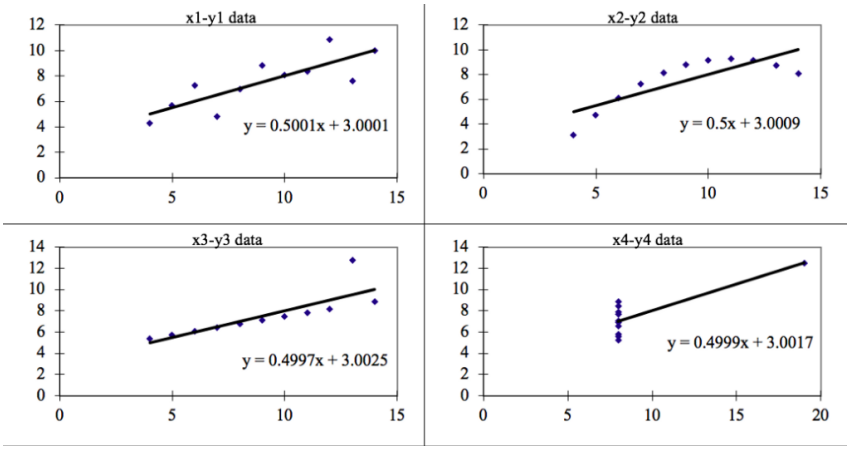
RSS: Residual sum of squares
TSS: Total sum of squares

Formula:

R^2= 1-RSS/TSS

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet can be defined as group of four datasets which Are nearly identical with statistical data, but then when plotted in a scatterplot shows very different distribution and appear differently as shown below.



This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them. We should plot the data in order to handle the anomalies 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.
Except 1st dataset, all other datasets cannot be handled by linear regression model.

3. What is Pearson's R?

Pearson's r is a measure of linear correlation between two sets of data. It is the most common method used for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation. 1 is

total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

It is a data pre-processing which is applied to independent variables to fix it within a particular range. It is important to perform scaling because we will have variables with different scales which leads to build incorrect model.
We perform scaling in 2 methods:

1. Normalized scaling. (MinMaxScaler)
   It fits the numerical data between 0 and 1. Majorly used to handle the outliers.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

2. Standardized scaling
   It fits the data between mean(mu)=0 and sigma(Standard deviation)=1

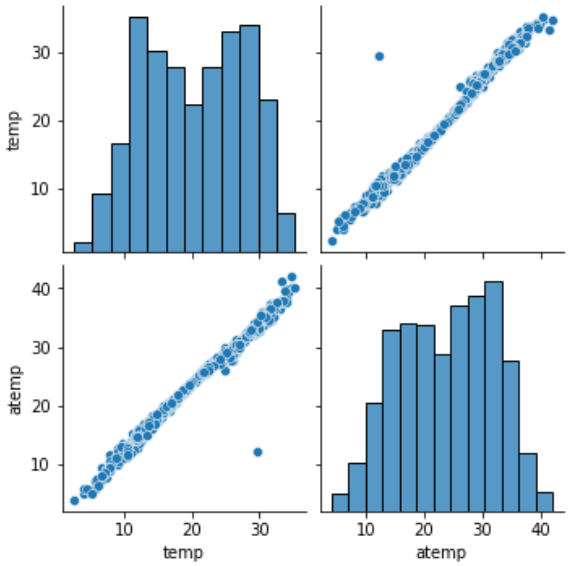
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

   One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

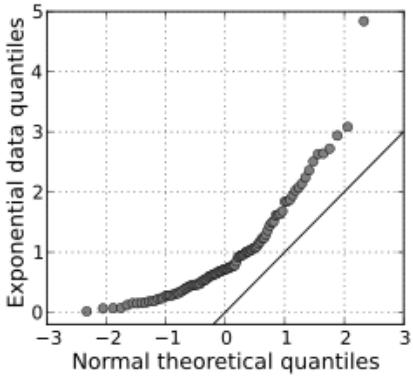
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

For example: In this assignment, we have temp and atemp which are highly correlated to each other leading to multicollinearity. Hence, we dropped temp to solve this problem. As shown below,



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

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.



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.