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?

Answer:
For Bike data set we had 'Season, Weather, Holiday, Month, Working Day, Weekday' categorical variables. Few had some impact while some had none. Below is detailed description on each column:

Season: There were 4 seasons namely Spring, Fall, Summer and Winter. As per or EDA plots we saw that Spring had least number of bikes rented. Fall had highest number of bikes rented followed by Summer and then winter.

Weather: There were 4 weather situations as per data dictionary: Good weather, mild turbulence, Bad weather and extreme/worse weather situation. As per data set we had data only in 3 categories: 'Good , Moderate, Bad Weather' and as per plot we see that Good weather is ideal and has best count, followed by Moderate. While there was very less count in Bad weather situation.

Holiday: As per our EDA plot we see that holiday or no holiday has less impact on count. However, we do see that on holiday the count of bike rented was less.

Month: We observe that June to September marked highest count in bike renting, probably this is summer season during that time. Lowest was encountered in the month of Jan and Feb.

Working Day: We see that usually count of bikes rented is slightly more on working day compared to holiday, however the impact is very meager. People rent bikes irrespective whether it's a working day or a holiday.

Weekday: Weekends had more bike count compared to weekdays.

So we can conclude that each column had some impact of dependent variable.

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

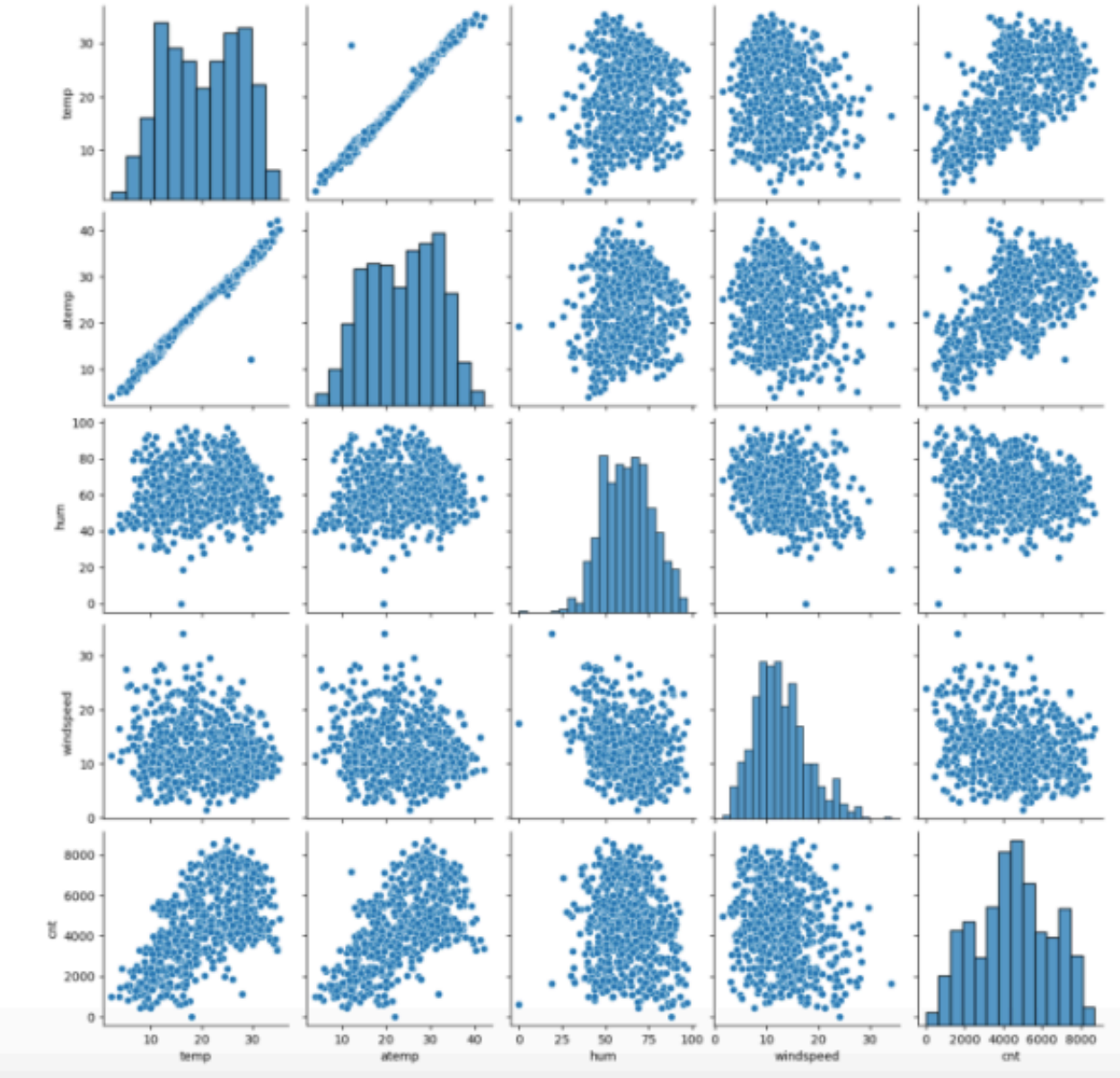
Answer:
If we have categorical variable that has 3 different values and we want to create dummy variable for same. Lets say we do not set drop_first=True, in that case it will create 3 dummy variable. It will result in redundant data and lead to multicollinearity in data set.

As we know that we can represent 3 different categories with only 2 dummy variables, when these 2 dummy variables are 0 means 3rd value holds true. This helps to remove multicollinearity with data and same can be achieved by setting drop_first=True.

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

Answer:

From plot we conclude that “temp” and “atemp” are highly correlated with count of bike rented.



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

Answer:

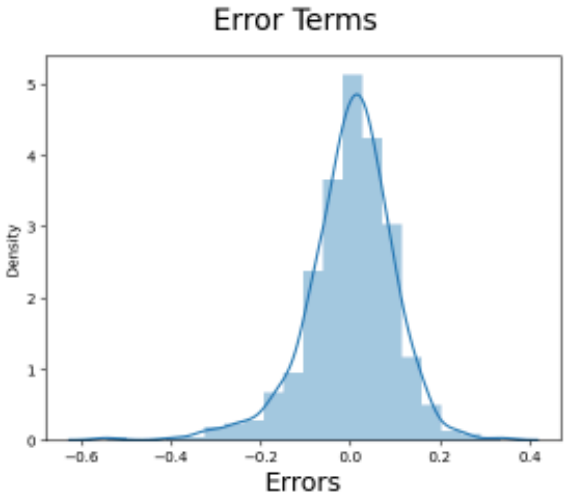
We made a distplot of residuals to validate the assumptions of Linear Regression.

As we see plot is normally distributed around mean 0 , thus helping us conclude that our assumption is correct.

Residual Analysis of the train data

```
[419]: y_train_count = lm.predict(X_train_lm)
       # Plot the histogram of the error terms
       fig = plt.figure()
       sns.distplot((y_train - y_train_count), bins = 20)
       fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
       plt.xlabel('Errors', fontsize = 18)                         # X-label

[419]: Text(0.5, 0, 'Errors')
```



Conclusion: Residual plot is Normally Distributed, thereby approving our model

---

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

Answer: According to the model below are 3 features that has significant impact of demand of shared bikes:

`Temperate:` 0.5419 - With 0.5419 increase in temp, the number of bike rentals increase by 0.5419

`Year:` 0.2364 – With 0.2364 change year , the number of bike rentals increase by 0.2364

`Windspeed:` -0.1758 – With 0.1758 decrease in windspeed , the bike rentals reduce by 0.1758

General Subjective Questions

---

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is basic regression used for continuous variables. It is based on equation of line `y=mx+c` where m = slope of line, c is the intercept.

Example of Linear regression with 2 variables, 1 is dependent variable and other is independent variable that helps determine dependent variable value. Example: We want to calculate grade of a student for maths subject, we have list of student marks and we need to learn and grade all new students. In this case

marks is out independent variable and grades is dependent variable which will be determined based on students marks.

There are 2 types of linear regression:

a. **Simple Linear Regression**: Here we have 1 independent variable that helps determine the dependent variable. It tries to find best fit line with minimum errors.

   Equation of line can be given by:
   y =mx + c + Error

   Strength of Linear regression model can be found with help of R^2 or with help of RSE.

b. **Multiple Linear Regression**: Here multiple variables contributes in the derivation of value of dependent variable. Model now fits a 'hyperplane' instead of a line.

   Equation of line can be given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

   Strength of Multiple Linear regression can be found with help of Adjusted R^2, AIC, BIC

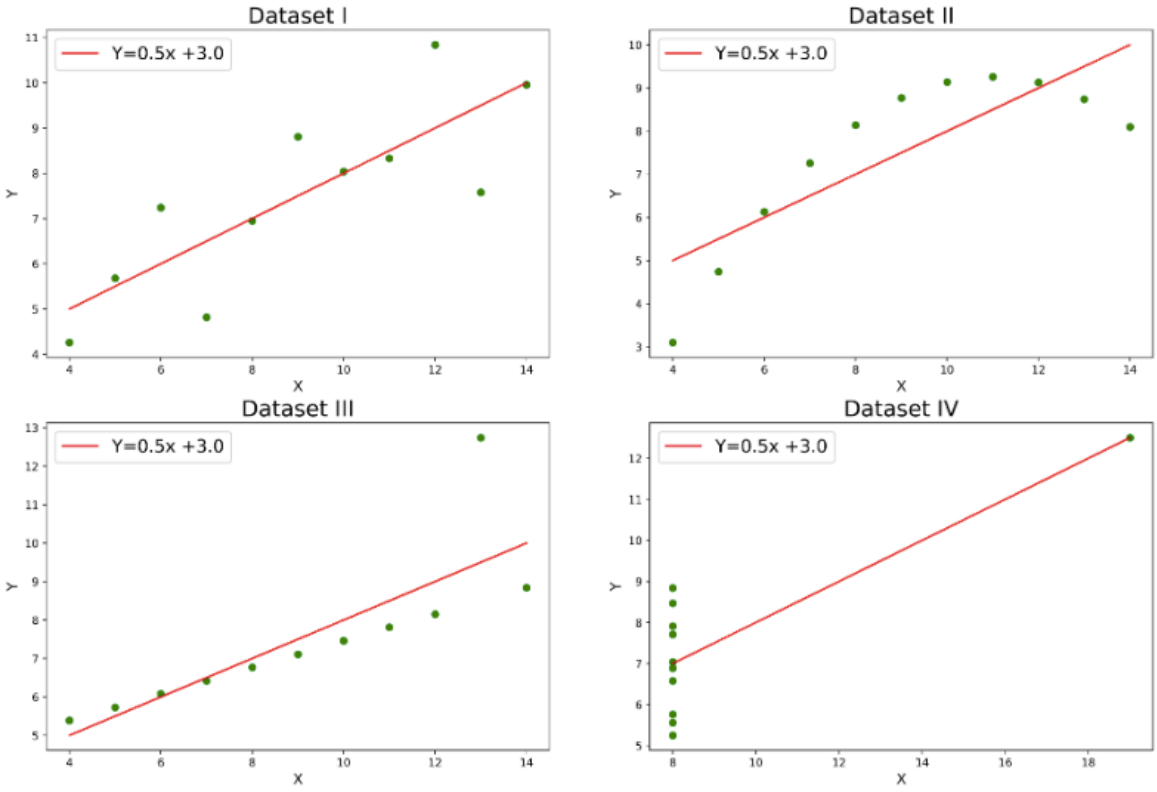
2. Explain the Anscombe's quartet in detail

Ans. Anscombe's quartet helps us understand importance of plots and graphs instead of completely relying on statistics. Anscombe's quartet was developed by statistician Francis Anscombe. He created 4 data sets with exactly same stats values however each had very different distribution when put on a graph. Its primary motive was to emphasize the importance of graphs and distribution before we analyse data, and also describes importance of small details that graph help us capture. Example: Outliers, Patterns

The four datasets of **Anscombe's quartet.**

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

Output:



Statistical Properties:

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

3. What is Pearson's R?

Ans. Pearson's correlation is used to measure strength of linear relationship between 2 variables x and y. It can have values in range of -1 to 1.

1 = Positive relationship between x and y
0 = No correlation between x and y
-1 = Complete negative relationship between x and y

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

Ans. Feature scaling is important and is done to normalize or standardize all the independent variables. Else model can have weird co-efficient and if our data to predict is in other scale than what we trained, it will generate inconsistent output. Hence, this is very important step before we train our model and is part of pre-processing.

There are 2 ways to scale features:

a. **Standardizing**: Variables are scaled in a way that their mean will be 0 and standard deviation will be 1.

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

b. **MinMax Scaling**: Variables are scales in a way that their values lie within range of 0 to 1. It is calculated based on Minimum and Maximum value of data set.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

Ans. VIF is used to determine relationship of 1 independent variable with all other independent variables. Very high VIF >10 means that variable is highly correlated and usually VIF > 5 should be handled.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Where Ri^2 :

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

Ans. QQ plot is Quantile -Quantile plot which is also known as scatter plot. It is created by plotting 2 different quantiles against each other. Here first quantile is the variable we are testing and second quantile is variable we are testing against.

Example: If we are calculating petrol price in Maharashtra follow a normal distribution. Here we are testing if the distribution of price of petrol in cities in Maharashtra is normally distributed vs quantile

from a normally distributed curve. If 2 quantile samples are sampled from same distribution, they should fall in a straight line.

Use and Importance of Q-Q plot in linear regression:

Q-Q plot is a graphical tool that helps assess if a set of data plausibility came from theoretical distribution such as a normal or exponential.

Example: If according to our statistical analysis we assume that our residuals are normally distributed, we can use QQ plot to validate our assumptions.

It helps to gain clarity on below scenarios:

1. Whether 2 datasets come from population with common distribution
2. Whether 2 datasets have common location and space
3. Whether 2 datasets have similar distributional space
4. Whether 2 datasets have similar tail behaviour