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? (3 marks)

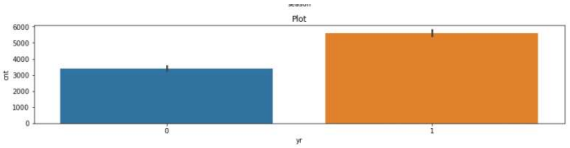
Answer:

Below is the equation of line and if we corelate the impact of

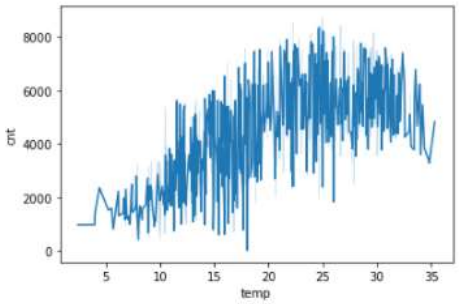
*ypred= 0.2177 + yr*0.2287 - holiday*0.0973 + temp*0.6058 - hum*0.1419 - windspeed*0.1724 + winter*0.1125 - Mist_Few_clouds*0.0485 - Light_Snow_Rain_Thunderstorm*0.2382*

Below is the impact of categorical variables on dataset

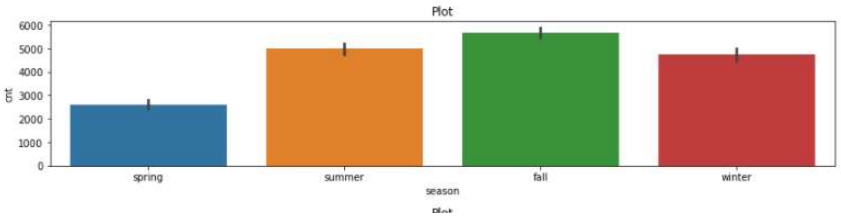
1. As year progresses, we do see the increase in ypred which is also evident from the chart below



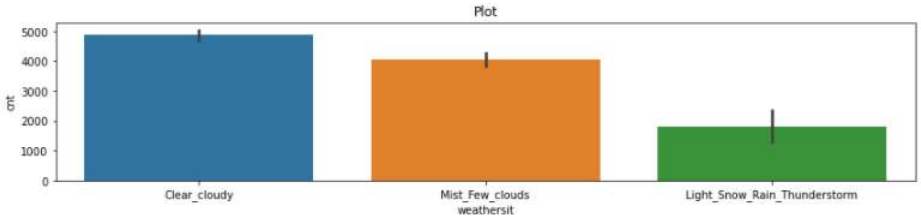
2. Temperature also causes an impact on dataset which is evident on increase in number with increase in temp



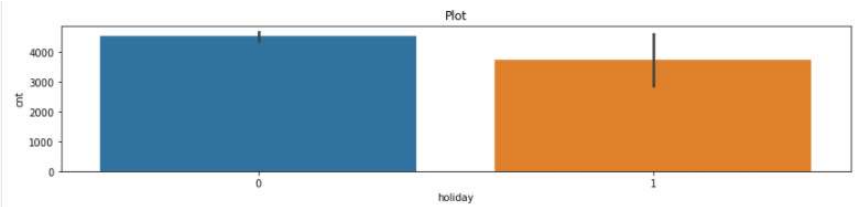
3. Winter also has causes increase in sales



4. As evident from plot and graph the riders decrease with increase in any form of precipitation which can be seen from graph for mist and thunderstorm,



5. There is a dip in ridership whenever there is a holiday



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

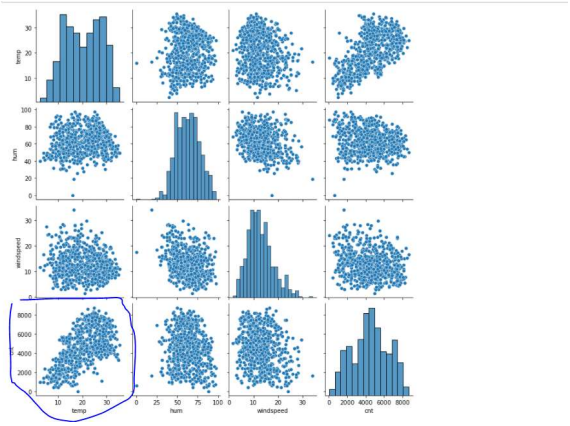
Answer:

Drop_first= True or dummy encoding basically reduces the number of dummy variables by 1 without loosing any information. It also reduces corelation created between variables. Another advantage is we reduce the number of variables that machine learning algorithm needs to learn

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

Answer:

Looking at pair plot we see temp having high corelation with cnt

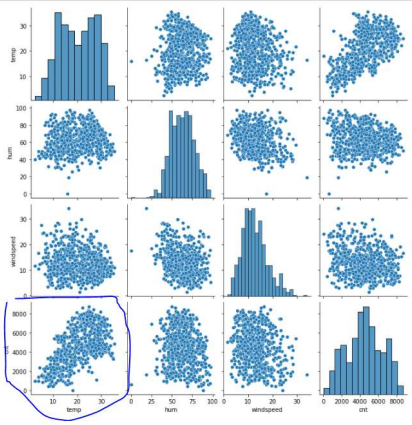


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

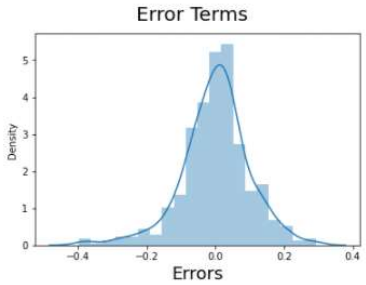
Answer:

There are 5 basic assumptions of linear regression

1. Linear relationship
   If we see the scatter plot the relation between temp and cnt appears to be linear. If we see other variables the don't seem to be linear



2. Multivariate normality.
   This assumption requires that residuals are normally distributed. If we see the residual plot it is normally distributed



3. No or little multicollinearity.
   This assumption is used to determine the relationship between independent variables. We need to ensure that there is less corelation between independent variables. We can use VIF or Corelation matrix to derive this. Scores less than 5 for VIF are desired.

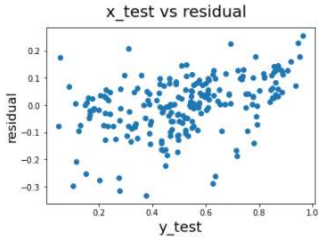
```
: #check the VIF again
  generateVIF(X_train_rfe)
```

	Features	VIF
0	const	44.77
4	hum	1.85
7	Mist_Few_clouds	1.55
8	Light_Snow_Rain_Thunderstorm	1.23
3	temp	1.22
5	windspeed	1.16
6	winter	1.14
1	yr	1.03
2	holiday	1.01

Observation

- above VIF seems to be in good state.const has high vif lets leave it

4. Homoscedasticity. This assumption need that error terms are having constant variance. It should not be increasing or decreasing like a cone but equally spread. As seen in below scatter plot the variance is evenly spread



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

Answer:

Based on analysis temperature/year and season(Winter) contributes more towards shared bikes

General Subjective Questions

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

Answer:

Linear regression is one of the supervised learning algorithms that works by fitting a line or a plane give a set of attributes or predictors. To illustrate this let's take an example where a team wants to determine influence of marketing methods like tv, radio on sales number. The basic idea over here is to identify an equation that helps to determine the predicted value here sales wrt to given input sales methods like tv/ radio etc also called independent variables

Equation for this plane or line is generally written by Y=B0+B1X1+B2X2+B3X3+….BnXn. Intent of regression is to identify B0 and B1…Bn

We typically use techniques like differentiation or gradient descent to identify coefficients. In addition to this we strive to achieve 4 principles of regression Homoscedasticity, no multicollinearity, multivariate normality

The effectiveness of algorithm is usually derived by methods like RSS. We try to get least RSS

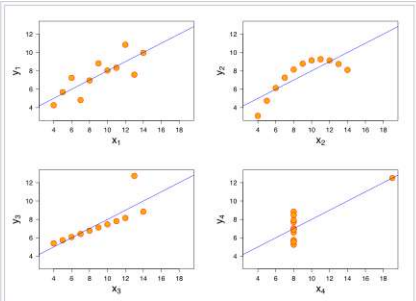
For any successful linear regression, we perform following steps

1. Understanding data
2. Performing analysis and plotting
3. Preparing data by scaling or dummy variables
4. Splitting data into train and test set(70:30 or 80:20)
5. Building model by leveraging VIF and p-Values either through top-down or bottom up approach on train set
6. Analysing the residuals
7. Testing the model on test set
8. Validating the model

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

Answer:

Anscombe's quartet shows that a dataset with similar statistical properties can still be different when graphed. As per wilipedia *"**Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed."* *Citation*.



Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Image ref : Importance of Data Visualization — Anscombe's Quartet Way. | by Sparsh Gupta | Towards Data Science

If we see above data we observe the summary stats are nearly identical for all datasets but their plots are varying.

There are some observations of this

- Data can be non linear or linear
- There can be outliers which can or cannot be handled by linear regression model

This brings an important conclusion that plot of data is needed before right model is picked for a given dataset

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

Answer:

Pearson coefficient is the measure of linear corelation between two sets of data. The value of this typically lies between -1 and 1. A value of 0 means no correlation where any thing greater than 0 means a positive tendency of increase with increase on other variable. A negative value means tendency of decrease with increase in value of other variable

covariances and variances based on a sample into the formula abo

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$

where:

The value of corelation is generally derived using corr function in python.

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

Answer:

Scaling is the process of bringing all that variables in uniform or comparable measuring scale. This is needed specially to avoid coefficients swinging on extreme end, that leads to difficulty in interpretation of model.

This help in speeding up Beta derivation using gradient descent.

Normalized scaling or Min Max scaling tries to fit data in [0 and 1] scale by doing

$(x-x_{min})/(x_{max}-x_{min})$

Standardized scaling scales value in such a way that mean lies at 0. It is computed by

(x-mean(x))/standard deviation(x)

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

Answer:

An infinite VIF means that the corresponding variable can be expressed linearly by other variables

Take this formula, if variables are highly corelated R2 becomes 1. This causes denominator to become 0 and hence infinite

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

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

Answer:

QQ plot or the quantiles plot is a scatter plot created by plotting 2 quantiles against each other. It helps us in identifying the normality of a distribution. If a distribution is normal then it follows a straight line. This is especially significant to validate the assumption that residual follow normal distribution. Also, it also can be used to confirm that data comes from same distribution