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

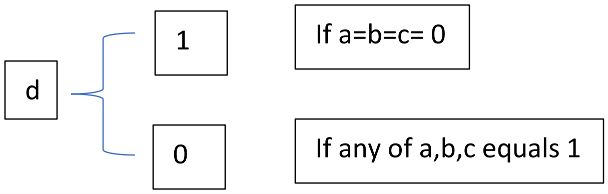
Categorical variables have had significant effect on demand of bike rentals .

a) Season : We saw that during season Spring demand of the bike rental is very low. The demand starts rising in summer and reaches maximum in Fall. Then again it starts to dip during winters. We saw sharp rise in demand from Spring to Summer
b) Year : We see demand for bike increased in year 2019
c) Months: We see the demand for bike rentals form a concave shape curve , it starts to rise from January, comes at saturation during May to October and then it starts falling
d) Weekday : We see on 6th day of the week there is high variation in demand for bike rentals
e) Weather Situation : We see that there is high demand when weather is clear and no rain . As the weather situation deters, demand for bike also dips

Some categories of these categorical variables have a positive impact , that is it becomes favorable for demand of bike rentals to increase. Whereas some decreases the demand of bike rentals.

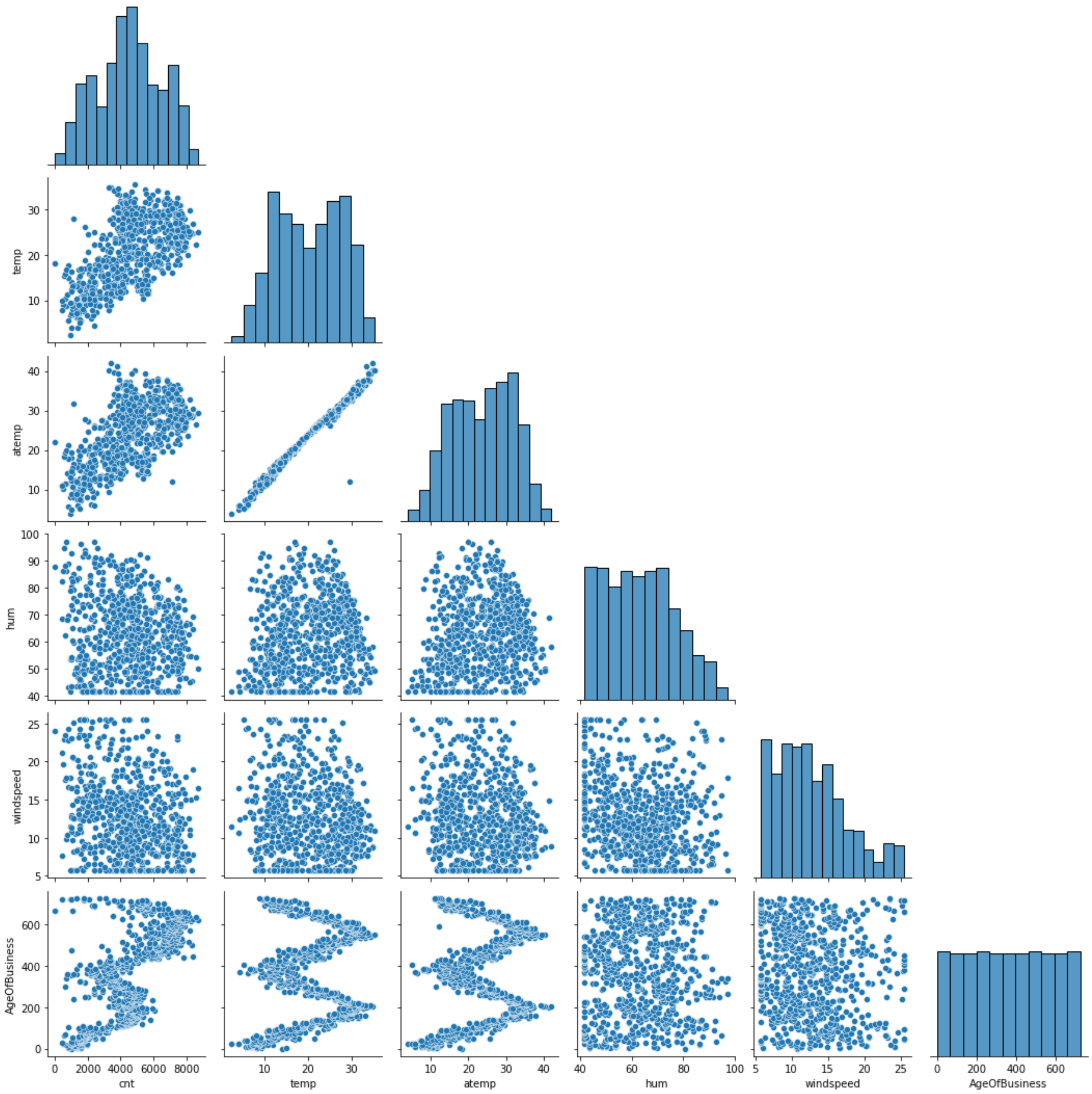
2. Why is it important to use drop_first=True during dummy variable creation?

We create dummy variables for those variables which have more than 2 levels of categories . Let say if there are 4 categories in a categorical variable ‘X’. If we create it’s dummy variable , we will create for (p – 1) levels, i.e (4-1) i.e 3 levels. When these 3 levels have values 0 it will self explain the 4rth level or category of X.  Let say categories of X are a,b,c,d. We can write :



Hence, if we do not drop a level it can cause issue of multicollinearity also.

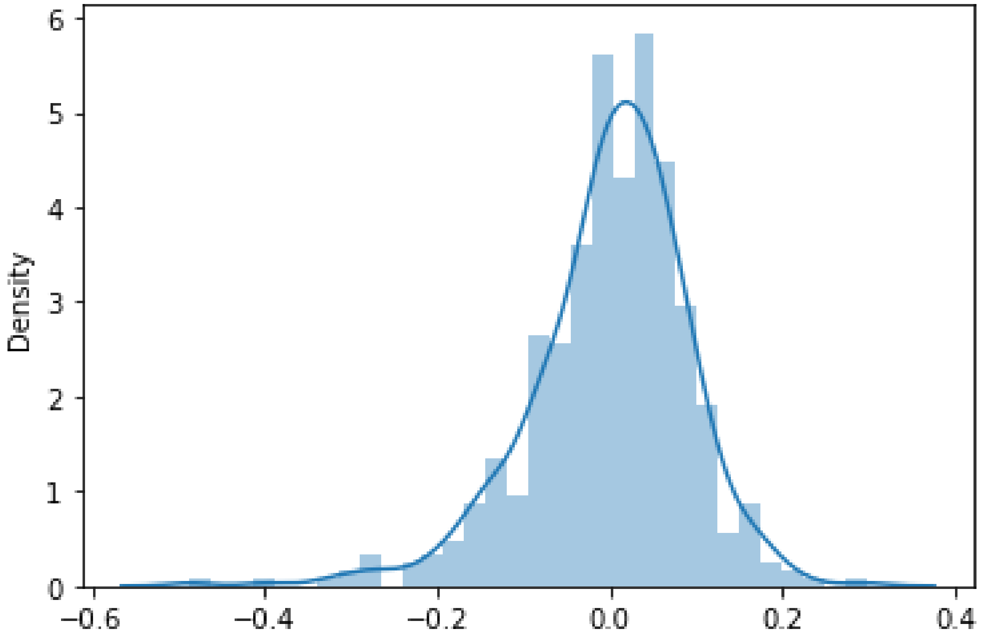
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



We can see that temp and atemp has high correlation with cnt (demand of bike rental)

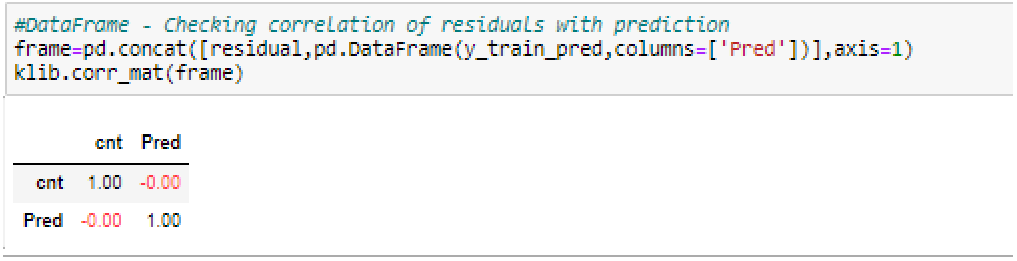
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

**Normality of Residuals**



We can see that distribution of residuals is centered at 0 and mean residual also shows that average of residual is very close to 0

**To check if the error terms are correlated with prediction**

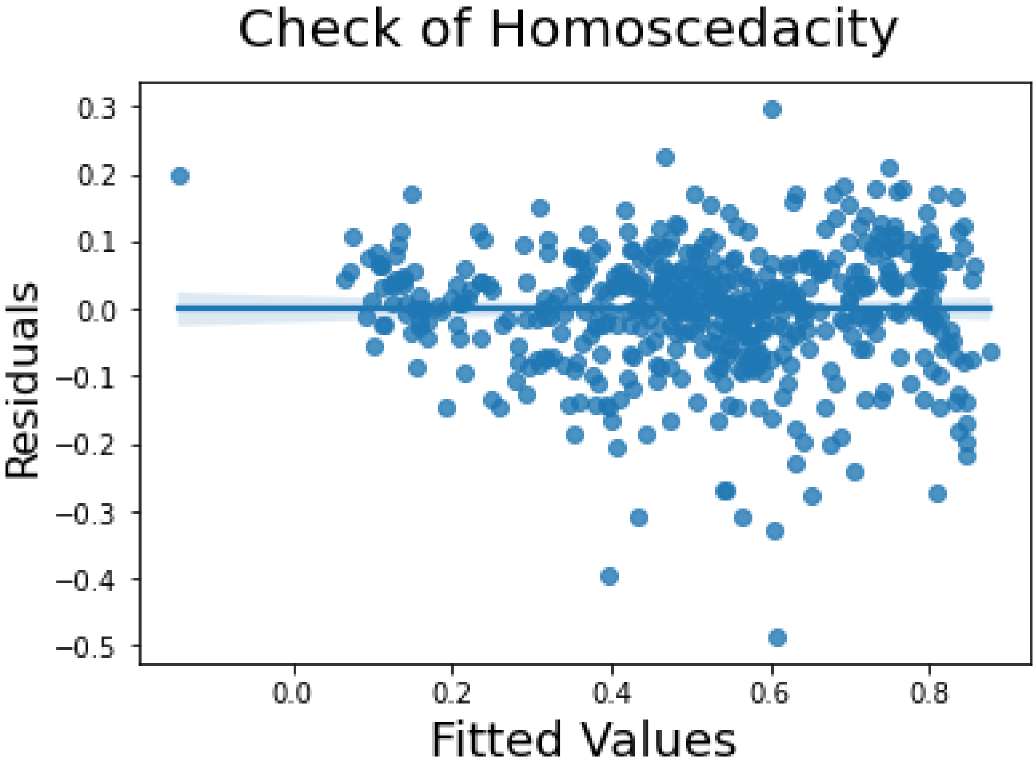


```
#DataFrame - Checking correlation of residuals with prediction
frame=pd.concat([residual,pd.DataFrame(y_train_pred,columns=['Pred'])],axis=1)
klib.corr_mat(frame)
```

|  | cnt | Pred |
|---|---|---|
| cnt | 1.00 | -0.00 |
| Pred | -0.00 | 1.00 |

We can see that error terms are un-correlated with the predicted values

**To check heteroscedasticity**



We can see that variance of error terms is almost (in majority) constant given the predictions. It means value of error terms do not vary when the value of predictor variable is changed

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

Top 3 features contributing significantly towards explaining the demand of shared bikes are :

- temp -> Having other variable constant , a unit change in temperature will casue demand for bike rental to increase by .3509 units
- weathersit_Snow_Rain_Thunder -> We saw in EDA that in weather situation like snow rain and thunder , demand for bike rental drops, hence our model validates the same. It states if the weather situation is snow + rain + thunderstorm, then the demand for bike rental will decrease by -.2927 units (having other variables constant)
- yr -> as year increases demand of bike rental also increases , we saw it during EDA. Model validates it , having other variable constant , if year is changed by a unit , bike rental demand will increase by .2414 units

6. Explain the linear regression algorithm in detail?

Linear Regression is a supervised learning technique, used to interpolate the patterns based on given data and target column being continuous , hence outputs prediction as continuous value. In it we try to fit a straight line(When you have an Independent Variable) or a hyperplane(When you have multiple Independent Variables) for the datapoints, which is used to predict value of new point, i.e for example you can predict if you spend X amount on marketing expenses, what will be the sales. It formulates a model which is in terms of mathematical equation based on the equation of straight line, i.e y=mX+c. Where y is your target variable, m is slope/coefficient, X is your independent variable and c is the intercept.

m can be expressed as tan(theta) = (y2-y1)/(x2-x1). This is slope, which signifies how strong the relationship is there between X & y. If we increase X by some unit, how much y is going to increase is dependent on slope.

Intercept c states when x equals zero what is the value of y.

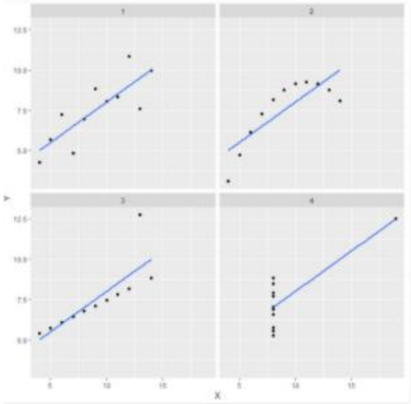
The algorithm tries to find out which line fits the data well. It select the line which gives least residuals. Residuals is difference between the actual and predicted value.

Linear Regression uses Ordinary Least Squares to identify the best fit line, it is a method which minimizes the total error squared. Which means it selects that best fit line, for the given model parameters (coefficients), which gives minimum residual sum of square.

To use such model it is favorable that the independent variable should have a linear relationship with the target variable , while independent variables should be independent of each other.

7. Explain the Anscombe’s quartet in detail.

Anscombe was a statistician, he stated his finding that states : he used 4 different sets consisting of 11 data points each. Each set showed same statistical value, i.e same mean and same standard deviation. He showed that, though it gives same statistical measure but still the distribution of all the 4 dataset is different. When he plotted all the 4 dataset, he got distribution something like shown below :



These distributions explains:

- 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.

The quartet states 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. Which makes us infer that statistical measures are the good tool to know about the dataset, but still are deceptive.

8. What is Pearson's R?

It is a measure to capture the statistical relationship or strength of association between two continuous variables. It gives the information about the magnitude of association or correlation as well as the direction of the relationship. It attempts to draw a best fit line through the data points of two variables and the coefficient states how far the data points are away from the best fit line. The range of the coefficient is between -1 to 1. The stronger the strength of association between two variables, coefficients near to 1 or -1. But it gives unreliable strength of association of there exist a non-linear relationship between two variables.

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

Scaling is bringing all the continuous variables of different units into same unit , i.e making them unitless. If scaling is not done, then the algorithm would tend to weight higher values, higher and consider smaller values as lower importance, regardless of their unit.

| | **Normalization** | **Standardization** |
|---|---|---|
| **Values Used** | Minimum and Maximum values of Features | Mean and Standard Deviation of Features |
| **Use Scenario** | When Features are of Different Scales | When we want to ensure 0 mean and unit standard deviation |
| **Range** | [0-1] or[-1-1] | Not bound |
| **Outliers** | Affected | Not Affected |
| **Distribution** | Unknown | Normal or Gaussian |

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

VIF means that variance of the model coefficient is inflated by factor of VIF value, which further mean that standard error of the coefficient is inflated by factor of square root of VIF value. The standard error of the coefficient determines the confidence interval of model coefficients. If standard error is large, then the confidence interval will also be large and model coefficient may come to be non-significant. VIF happens to be infinity when there is a perfect correlation.

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

QQ plot is a plot of two quantiles against each other. Its' purpose is to find does two sets of data follow same distribution or not? If data falls on the 45$^o$ line, it means they follow same distribution. *This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.*