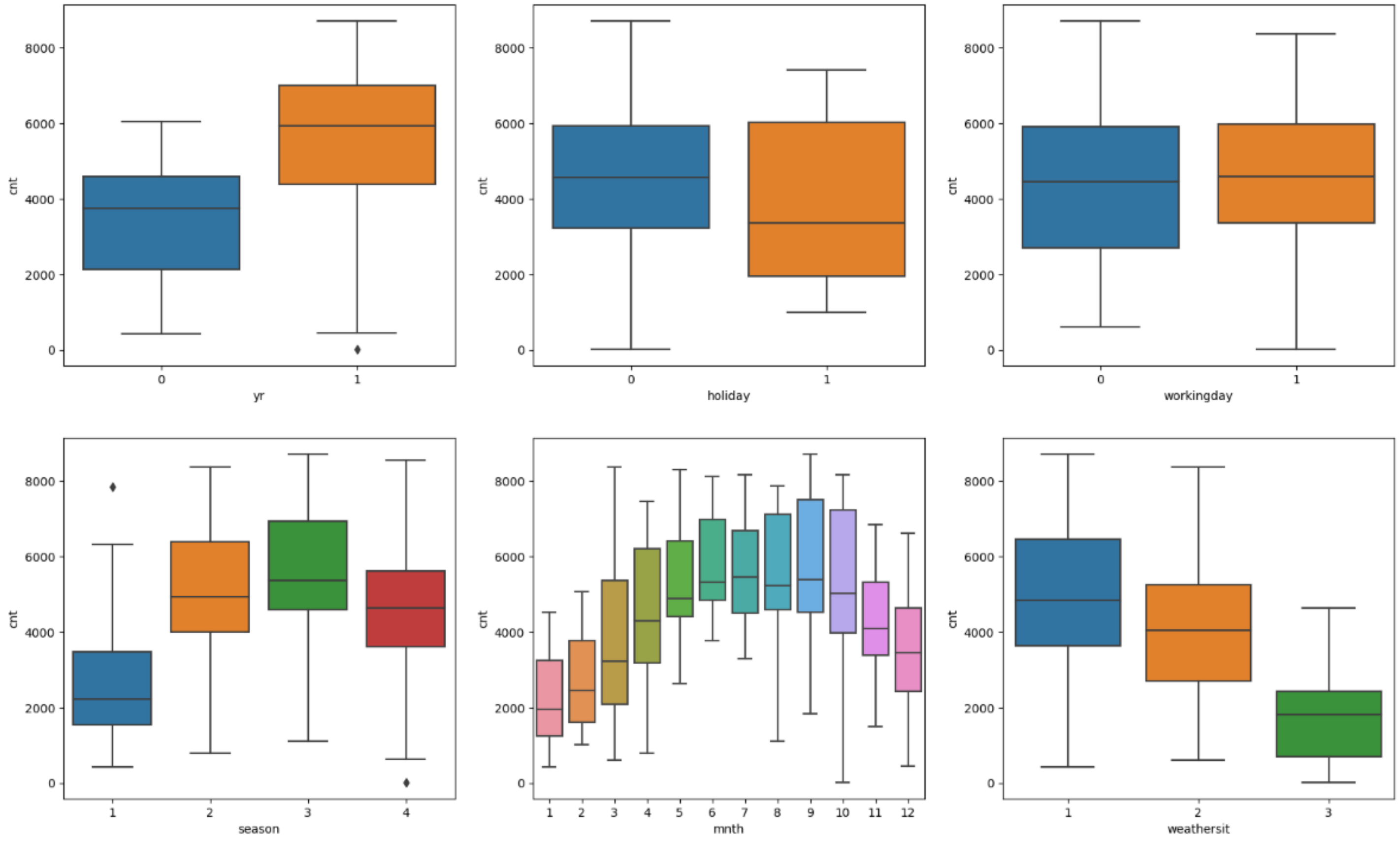
Assignment-based Subjective Questions

Biswajit Pattanayak

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

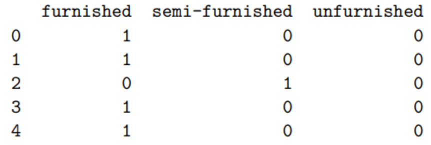
Ans –

a) 2019 had more bike counts than 2018
b) On average, during holidays the demand is less
c) Demand is highest during Fall season followed by Summer
d) From January till September, demand increases and then falls. Demand is highest in July.
e) Count is higher when it is Clear skies.
f) No significant difference between working and non working days.



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

Ans – We will answer this with an example. Suppose furnished, semi, furnished and unfurnished are converted to dummy variables.


```
   furnished  semi-furnished  unfurnished
0          1               0            0
1          1               0            0
2          0               1            0
3          1               0            0
4          1               0            0
```

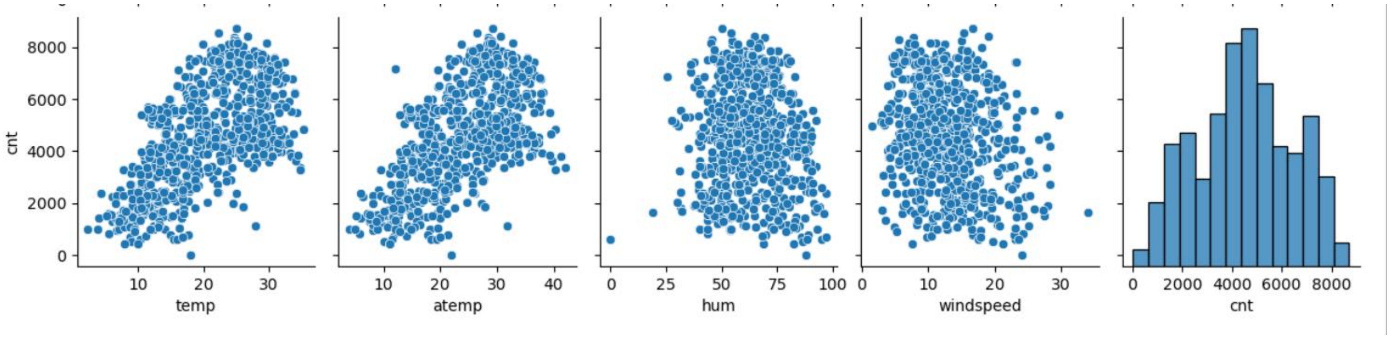
In above you don't need 3 columns. You can drop the furnished column and type can be identified with the last 2 columns -

- 00 will correspond to furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

As we can see, furnished is redundant and can be removed. 2 dummy variables can easily identify all three. Model will get over-parameterized if not dropped.

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

Ans -

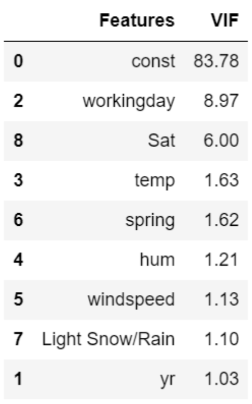


Temp/Atemp has the highest collinearity with cnt

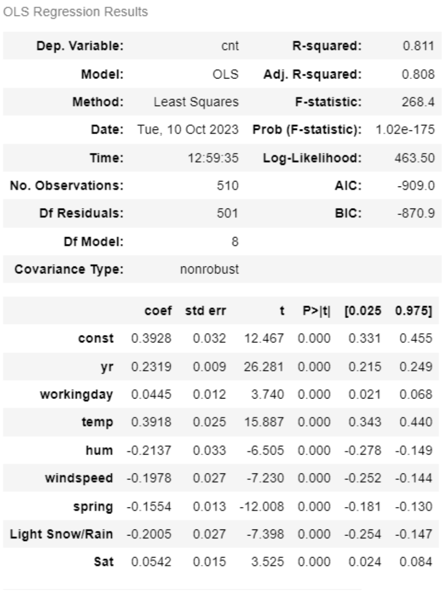
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans – After building the model by removing unwanted variables (variables with high P value, high VIF value) by using RFE + Manual method, the last values are as shown value. As we can see the P values are less than 0.05 and VIF values are < 10. Hence we can conclude the model is a good fit now.

The R squared and adjusted R squared values are also pretty high at > 80%



	Features	VIF
0	const	83.78
2	workingday	8.97
8	Sat	6.00
3	temp	1.63
6	spring	1.62
4	hum	1.21
5	windspeed	1.13
7	Light Snow/Rain	1.10
1	yr	1.03



```
OLS Regression Results
      Dep. Variable:               cnt        R-squared:      0.811
             Model:                OLS   Adj. R-squared:      0.808
            Method:      Least Squares      F-statistic:      268.4
              Date:   Tue, 10 Oct 2023 Prob (F-statistic): 1.02e-175
              Time:           12:59:35   Log-Likelihood:     463.50
 No. Observations:                510              AIC:     -909.0
      Df Residuals:                501              BIC:     -870.9
          Df Model:                  8
   Covariance Type:          nonrobust

                     coef  std err        t  P>|t|  [0.025  0.975]
           const   0.3928    0.032   12.467  0.000   0.331   0.455
              yr   0.2319    0.009   26.281  0.000   0.215   0.249
      workingday   0.0445    0.012    3.740  0.000   0.021   0.068
            temp   0.3918    0.025   15.887  0.000   0.343   0.440
             hum  -0.2137    0.033   -6.505  0.000  -0.278  -0.149
       windspeed  -0.1978    0.027   -7.230  0.000  -0.252  -0.144
          spring  -0.1554    0.013  -12.008  0.000  -0.181  -0.130
 Light Snow/Rain  -0.2005    0.027   -7.398  0.000  -0.254  -0.147
             Sat   0.0542    0.015    3.525  0.000   0.024   0.084
```

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

Ans - The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

Hence, top 3 variables as per VIF values are 'workingday', 'Saturday' and 'temperature'

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Ans - Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.
   The equation "y = mx + c" provides a straight line that represents the relationship between the dependent and independent variables. The slope 'm' of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).
   'c' is the value when x = 0 i.e. point on y axis where the Line will intersect it.
   Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y.
   X may be a single feature or multiple features representing the problem.
   For multiple regression model, equation is :
   Y = c + m1x1 + m2x2 + …. m(n)x(n)

2. **Explain the Anscombe's quartet in detail.**

   Ans - Anscombe's Quartet is the modal example to demonstrate the importance of data visualization. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

   Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process.

3. **What is Pearson's R?**

   Ans - The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

   Between 0 and 1 - Positive correlation
   When one variable changes, the other variable changes in the same direction.

   0 - No correlation

There is no relationship between the variables.

Between 0 and –1 - Negative correlation
When one variable changes, the other variable changes in the opposite direction.

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

   Ans – Scaling of variables is an important step because, as in some variable it will be on a different scale with respect to all other numerical variables, which might be taking comparatively smaller values.

   Hence, it is important to have everything on the same scale for the model to be easily interpretable.

   If we don't have comparable scales, then some coefficients will be very large compared to others and can impact model evaluation.

   There are two ways of rescaling:
   1. Normalisation or Min-Max : Between 0 and 1
   2. Standardisation(mean -0 , sigma -1) –> (x-mean)/std dev

   One advantage of Standardisation is that it doesn't compress the data between a particular range as in Min-Max. This is useful if data has extreme Outliers

   Min_max scaling –> (x-xmin)/(xmax-xmin) ---- Min-Max can take care of the Outliers

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

   Ans – It happens when some variables are able to create perfect multiple regressions on other variables
   A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

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

   Ans - A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

   A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically
   If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity 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.