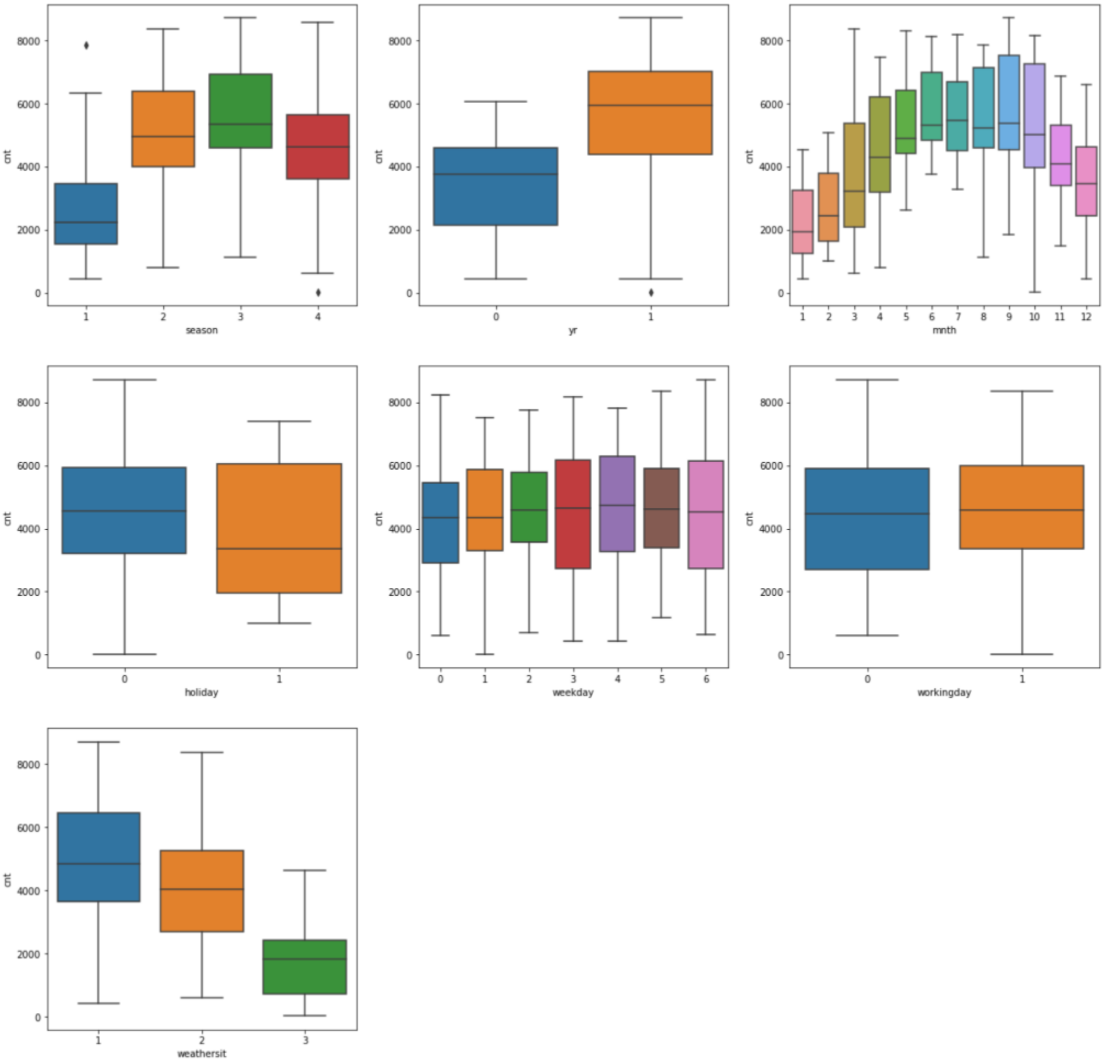
Linear Regression Subjective Questions

By Naman Kumar

All the images self-generated using jupiter notebook and not copied from elsewhere.

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?



After studying the categorical variables using boxplots. The following inferences can be concluded:

- Bike rentals increases in the summer to fall season and decrease back when the winter begins.
- There has been significant growth in rentals in 2019 from 2018.
- Weather affects the number of the rental significantly. The more the weather is clear the higher the rentals count.
- Working days, holidays or a weekend do not overall affect the rentals significantly.

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

When creating dummy variables we create columns for each of the values of the categorical variable and assign them 0 or 1. But we don't need to have columns for all the values as one value can be inferred as all other value columns being 0. Thus we only need n-1 total value columns.

Passing *drop_first=True* does the same thing for us when we use *pd.get_dummies()* to create dummy variable columns. It drops the first value column and adds the rest.

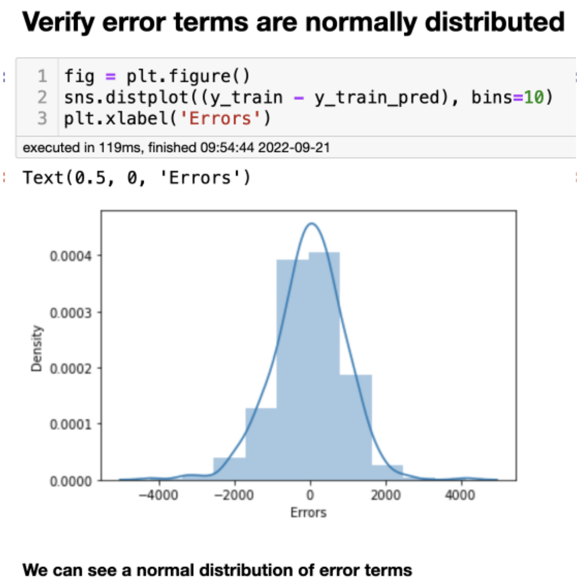
3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From the pair-plot alone, we can see that the *atemp* has the highest correlation with the *cnt* variable. This is also proved by the correlation matrix. The correlation coefficient for *atemp-cnt* was *0.65*.

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

For each of the assumptions of Linear Regression plotting was made based on *y_train* and *y_train_pred* to validate that the assumptions hold correct.

- To validate that the error terms are normally distributed distplot was made.



Verify error terms are normally distributed

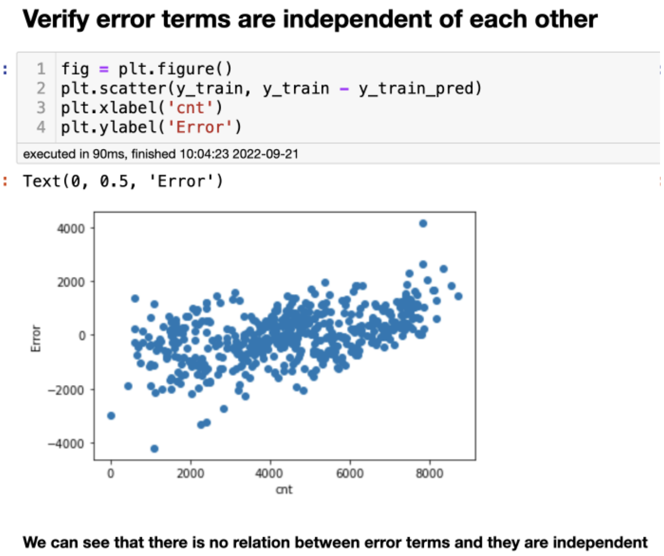
```
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins=10)
plt.xlabel('Errors')
```

executed in 119ms, finished 09:54:44 2022-09-21

```
Text(0.5, 0, 'Errors')
```

We can see a normal distribution of error terms

- To validate that the error terms are independent of each other, error terms were plotted against actual observations to find any pattern but no such pattern was found.



Verify error terms are independent of each other

```
fig = plt.figure()
plt.scatter(y_train, y_train - y_train_pred)
plt.xlabel('cnt')
plt.ylabel('Error')
```

executed in 90ms, finished 10:04:23 2022-09-21

```
Text(0, 0.5, 'Error')
```

We can see that there is no relation between error terms and they are independent

- To validate that the error terms have constant variance, scatter plots for error terms vs each of the features were created and the variance observed was constant.

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

The top 3 features contributing significantly towards the demand for shared bikes are:

- *temp* variable which tells temperature in celsius. (normalized)
- *light_snow_weather* dummy variable which is 1 for *weathersit* = 3 else 0
- *2019_year* dummy variable which is 1 for yr = 2019 else 0

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning method in which past data is used to build a model for predicting future outcomes.

A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

The independent variable is also known as the **predictor variable**. And the dependent variables are also known as the **target variables**.

The equation of a straight line is **y = mx + c**, where:

- y is the target variable
- x is the independent variable
- c is the y axis intercept
- m is the slope of the line

A best-fit line signifies m amount increase in y on a unit increase of x. The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

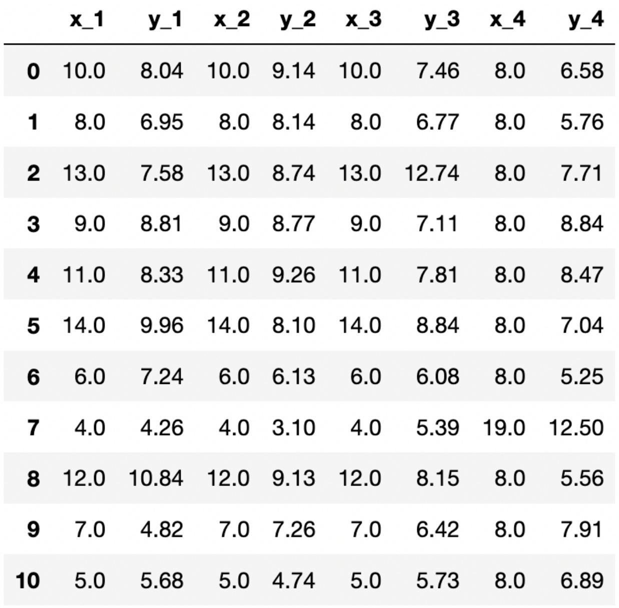
The strength of the linear regression model can be assessed using 2 metrics:

- R-square
- Residual Standard Error

2) Explain the Anscombe's quartet in detail.

Anscombe's quarter is a group of 4 datasets which are identical in description analysis but appears different when plotted using scatter plot signifying much different spread.

Consider the following data:

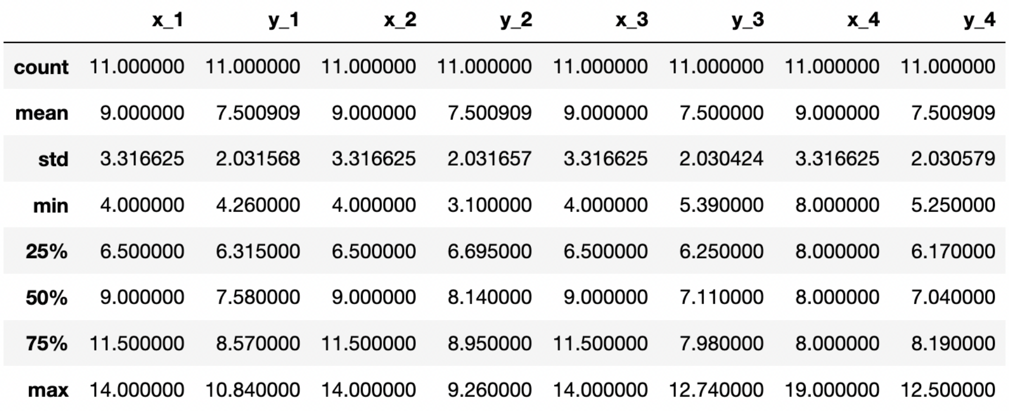


	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
0	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
1	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
2	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
3	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
4	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
5	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
7	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
8	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
9	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
10	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It contains 4 data sets:

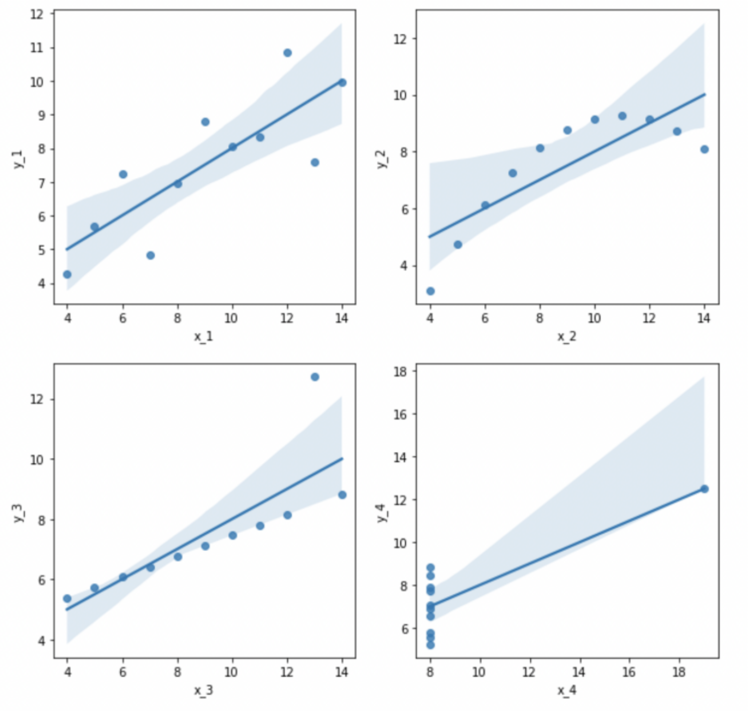
- x_1, y_1
- x_2, y_2
- x_3, y_3
- x_4, y_4

On doing the simple descriptive analysis, we find that all the metrics for the 4 datasets are almost equal.



	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	9.000000	7.500909	9.000000	7.500000	9.000000	7.500909
std	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
min	4.000000	4.260000	4.000000	3.100000	4.000000	5.390000	8.000000	5.250000
25%	6.500000	6.315000	6.500000	6.695000	6.500000	6.250000	8.000000	6.170000
50%	9.000000	7.580000	9.000000	8.140000	9.000000	7.110000	8.000000	7.040000
75%	11.500000	8.570000	11.500000	8.950000	11.500000	7.980000	8.000000	8.190000
max	14.000000	10.840000	14.000000	9.260000	14.000000	12.740000	19.000000	12.500000

This might lead us into believing that the 4 data sets are similar and we may choose to build a common linear regression model for them. But we find different results when we plot this data using scatter plot.



As we can see the 4 datasets have a very different spread even while having same descriptive stats.

- In the first dataset a positive linear relationship is visible but with a higher RSS.
- In the second plot we can see that the relationship between x & y is not linear and the linear regression line does not fit the dataset properly.
- The third dataset is almost a straight line with one anomply and the regression line fits it the best.
- The fourth dataset shows that y is independent of x and there is no relationship between them.

3) What is Pearson's R?

It is a correlation coefficient which defines the measure and direction of linear relationship between two variables. It is represented by *r* and lies between -1 and 1.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where:

- r = correlation coefficient
- x_i = values of the x-variable
- $\bar{x}$ = mean of x-variable
- y_i = values of the y-variable
- $\bar{y}$ = mean of y-variable

Meaning of values of Pearson’s R:

- Equal to 0
  - No relationship
- Between 0 and 0.5
  - Weak or moderate strength
  - Positive direction
- Greater than 0.5
  - Strong strength
  - Positive direction
- Between 0 and -0.5
  - Weak or moderate strength
  - Negative direction
- Lesser than 0.5
  - Strong strength
  - Negative direction

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

Scaling is a operation performed on features of a model to put their values in the same range.

In a multiple linear regression model where we have many independent variables, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods

The two popular methods used for scaling are:

- **Standardization:** The variables are scaled in such a way that their mean is zero and standard deviation is one.
  Formula: (x−mean(x)) / sd(x)
- **Normalization:** The variables are scaled in such a way that their values lie between 0 and 1 using the maximum and the minimum values in the data.
  Formula: (x−min(x)) / (max(x)−min(x))

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

VIF stands for Variance Inflation Factor. A large value of VIF indicates that there is a correlation between the variables. If VIF = 1, it means that all the independent variables are orthogonal to each other.

If there is a perfect correlation between independent variables then the VIF value is infinite.

VIF = 1 / (1 - R^2)

R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable. When two independent variables are perfectly correlated then the value of R^2 = 1

Thus,

VIF = 1 / (1 - 1) = 1 / 0 = infinite

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

Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

These plots are very helpful in understanding the performance of the model in linear regression as they help in determining:

- If two actual observations vs predicted values are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution