

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?

Ans: The categorical variables in the data set are season, weathersit, holiday, mnth, weekday and yr.

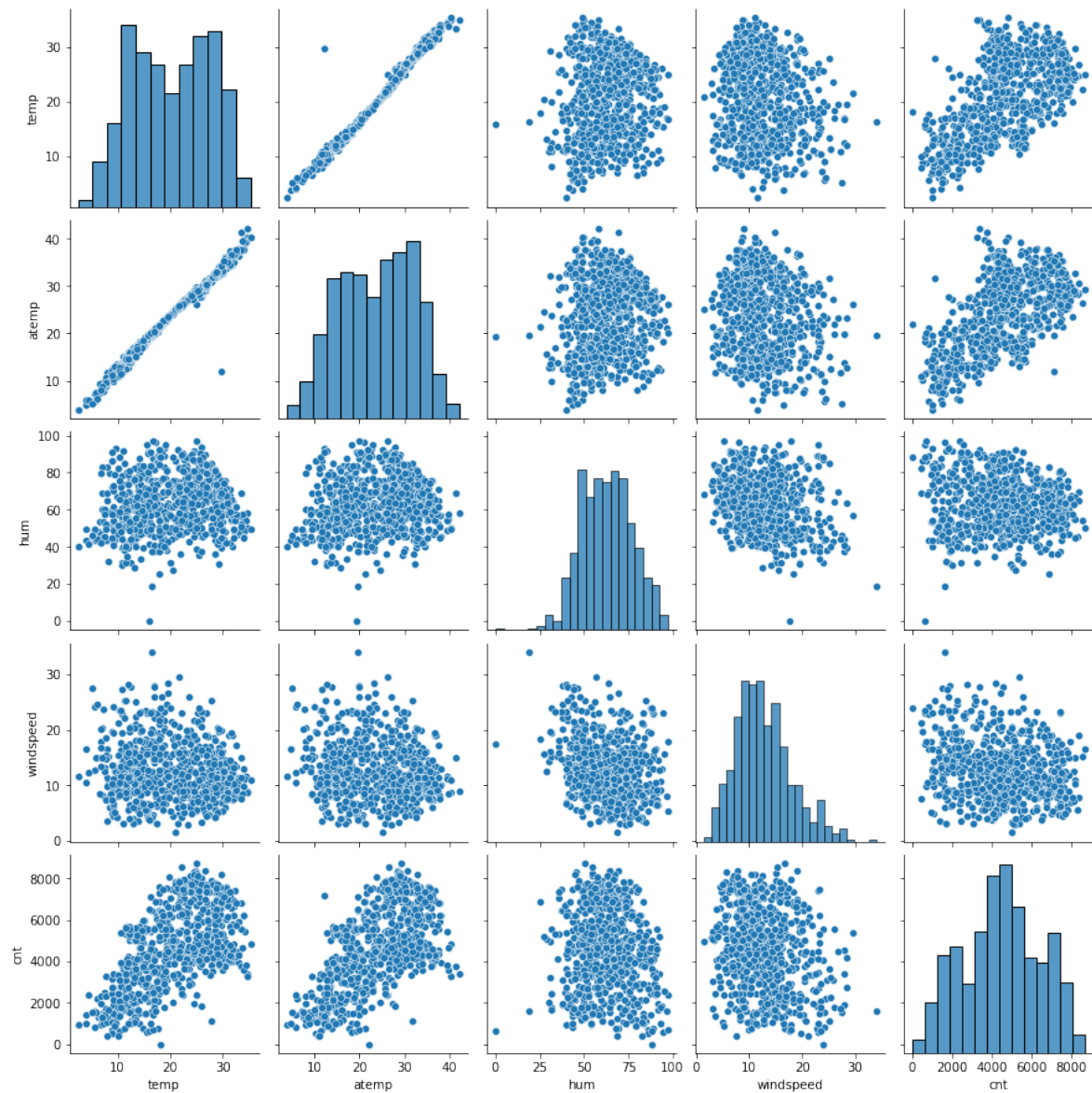
- Season: Most favourable seasons for high demand for the rentals bikes is summer and fall.
- Weathersit: Most favourable weather condition is clear few clouds.
- Holiday: consumption of bikes compared between registered and casual users, the observation is casual users are using bikes more on holiday.
- Mnth: Bike rentals are more in the months of June, July, August, September, and October.
- Weekday: Bookings look almost equal either on working day or non-working day. So no significant pattern observed with weekday.
- Yr: The demand for the bike rentals has increased from 2018 to 2019

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

Ans: It is important to drop the first column to avoid the redundancy. If the first column is not dropped while using get dummy, the dummy variables will be correlated which may affect some models adversely and effect is stronger when the cardinality is smaller. This will also lead to multicollinearity between the dummy variables and to keep this in control dropping first column is important.

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

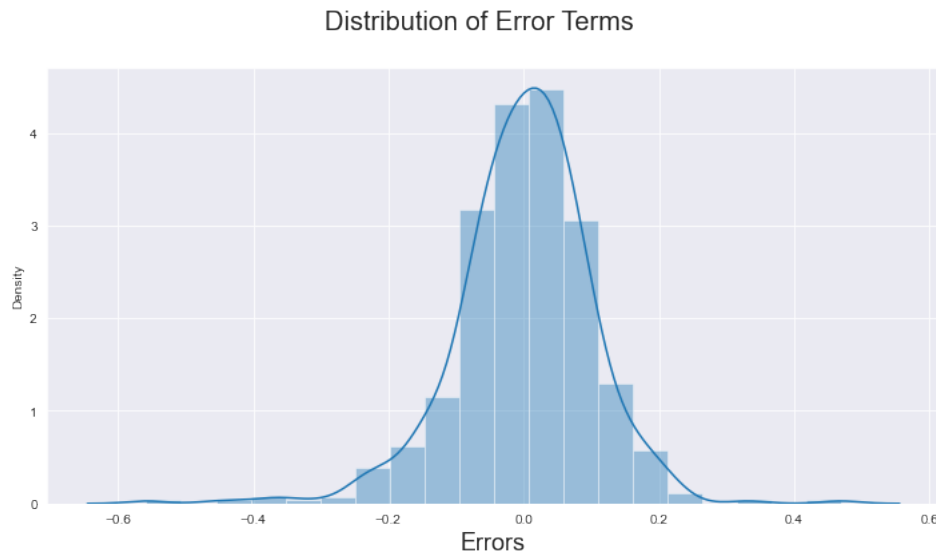
Ans:



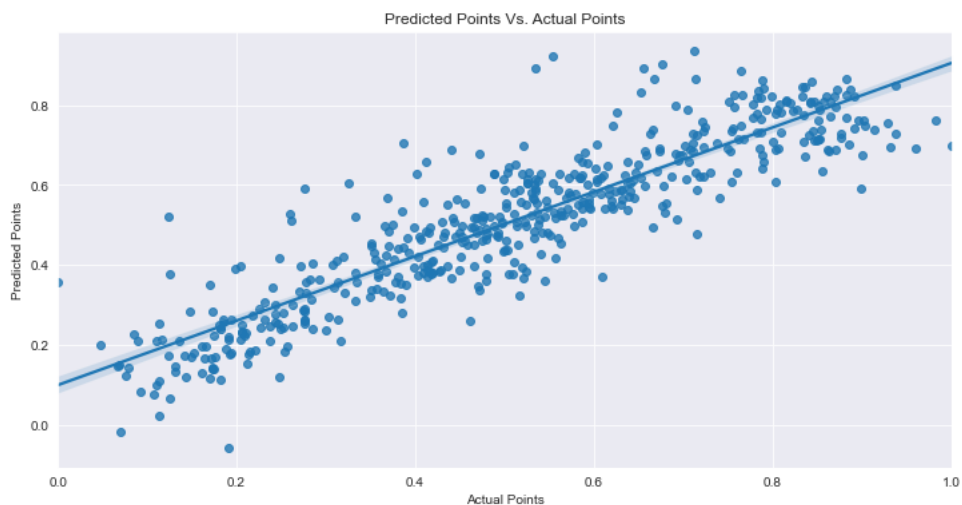
After dropping few variables to control multicollinearity, atemp and temp has the highest correlation with the target variable cnt.

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

Ans:



- Test for normal distribution of errors centred around 0 (mean=0)



- Linear relationship between the variables.
-

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

Ans: According to the model, the top three features are:

- | | | |
|------------------------------|---|--------------------|
| 1. Year | – | 0.25 coefficient |
| 2. Sep | – | 0.19 coefficient |
| 3. Weather Light rain & snow | – | - 0.32 coefficient |

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.

The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

There are 2 types of linear regression algorithms

- Simple Linear Regression – Single independent variable is used.
 - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
- Multiple Linear Regression – Multiple independent variables are used.
 - $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$ is the line equation for MLR.
- $\beta_0 = \text{value of the } Y \text{ when } X = 0 \text{ (} Y \text{ intercept)}$
- $\beta_1, \beta_2, \dots, \beta_p = \text{Slope or the gradient.}$

Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained.**

- Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
 - The straight-line equation is $Y = \beta_0 + \beta_1 X$
 - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i .
 - Now the cost function will be $J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
- The unconstrained minimization are solved using 2 methods
 - Closed form
 - Gradient descent

While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.

- $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
- OLS is used to minimize the total e^2 which is called as Residual sum of squares.
- $RSS = \sum_{i=1}^n (y_i - y_{pred})^2$

Ordinary Least Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet is defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

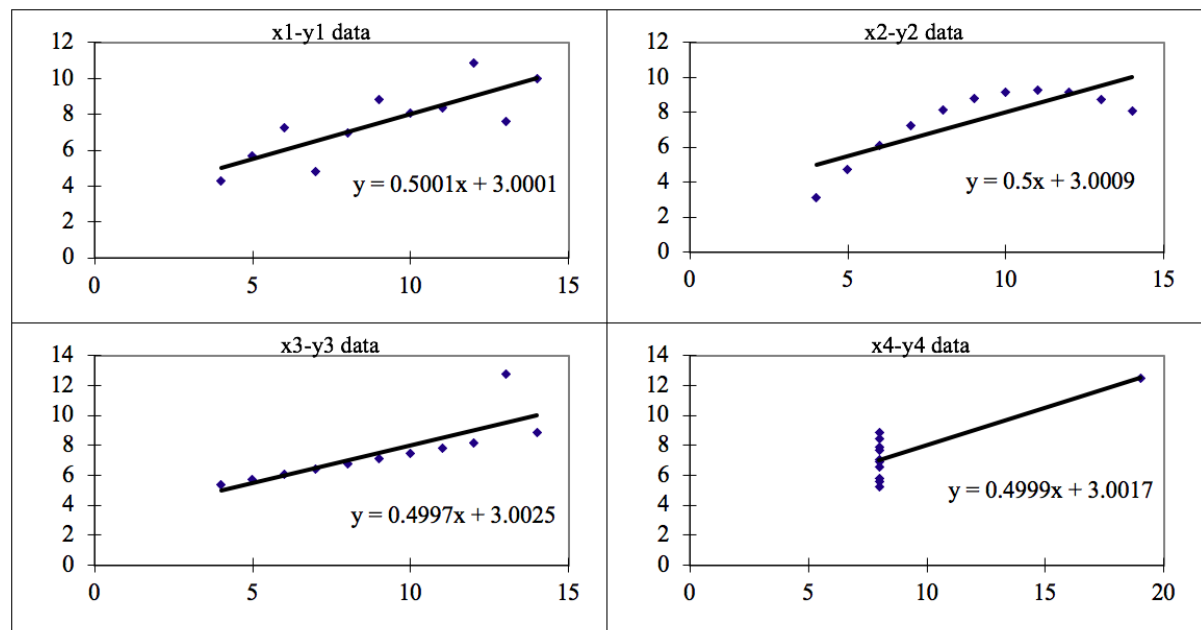
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

The statistical information for all these four datasets are approximately similar and can be computed as follows:

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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? $r = 1$ means the data is perfectly linear with a positive slope $r = -1$ means the data is perfectly linear with a negative slope $r = 0$ means there is no linear association.

$$r = \frac{n(\sum x * y) - (\sum x) * (\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] * [n\sum y^2 - (\sum y)^2]}}$$

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

Ans: The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.

We do scaling to avoid the chances of processing the data without the appropriate unit conversion which are high. Also, the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.

The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

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

Ans: VIF - the variance inflation factor -The VIF gives variance of the coefficient estimate is being inflated by collinearity. $(VIF) = 1/(1-R^2)$. If there is perfect correlation, then $VIF = \infty$. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables. So, $VIF = 1/(1-1)$ which gives $VIF = 1/0$ which results in "infinity"

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

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.