Boom Bikes Assignment: Ashish Arora – DS-C45

Assignment-based Subjective Questions:

Ques1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

For the analysis of categorical variables, I have used bar plot and bar chart. With this analysis, I was not only able to find the outliers if any but also the contribution of each category in demand of bikes. My findings are as follows.

- **Season:** In fall and summer season, the company has good demand for bikes. Spring season has the least demand.
- **Year:** In 2019, the bike docking company has more demand as compared to 2018.
- **Month:** From May to October, demand of bike is high as compared to other months. Demand tends to increase at the start of each year. Demand reaches its peak in mid of year and as the year ends, demand goes down.
- **Weather Situation:** Apparently, if the weather is clear, demand is high for bikes.
- **Weekday and Working day:** Weekday and weekend has almost same demand.
- **Holiday:** If there IS a holiday and then demand is less. Though weekday, weekend and holiday share same proportion of demands. Hence less reliable variable.

Ques2: Why is it important to use drop_first=True during dummy variable creation?

It is important to keep only k-1 dummies out of k dummies **because k-1 dummies are able to explain the k dummies.** Keeping all k dummies would add a **problem of multicollinearity.**

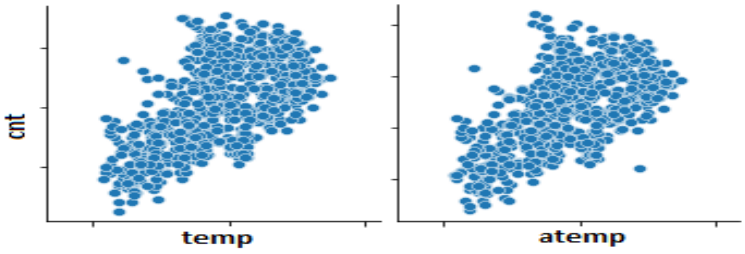
Syntax - drop_first: bool, default False

For example: In bike assignment, we have provided 4 seasons i.e. spring, winter, fall and summer. Suppose, if we create 3 dummies of 4 season i.e. season can be spring or not, season can be winter or not, season can fall or not. So when none of season is spring, winter and fall, then it is presumed that season is of summer.

Ques3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

In pair plot, both **atemp** and **temp** looked highly correlated with the target variable ***(cnt).*** In order to get the highest correlation, we used the k-Pearson measure of coefficient and found that **atemp has the highest correlation.**

	Target_var	Respone_var	corr
101	cnt	atemp	0.631649
100	cnt	temp	0.628040
105	cnt	yr	0.570481
109	cnt	spring	-0.563872



Ques4: How did you validate the assumptions of Linear Regression after building the model on the training set?

For the validation of my train model, I validated 5 major assumptions of Linear Regression.

Assumption 1: No Multicollinearity between Independent Variables.

- **Variance inflation factor** less than 4.
- **Heat map depicting correlation** among independent variables.

Assumption 2: Error should be normally distributed

- **Histogram:** Checked distribution and symmetric shape.
- **QQ-Plot:** Are the theoretical quantiles and error quantiles lying on 45 degree line.

Assumption 3: No correlation among the error terms

- **Residual vs Row number Plot:** To check the seasonality.
- **Durbin and Watson test:** If it is close to 2 or equal 2, then there is no autocorrelation.

Assumption 4: Homoscedasticity of the errors

- **Residual vs Fitted value Plot:** To check the spread among the error to know the variance.
- **Actual vs predicted Scatter curve:** To check the trend between and actual and predictive values. Is there any funnel shape which may lead to non-constant variance.

Assumption 5: Linear Relationship between response and explanatory variable.

- **Residual vs Fitted Value Plot:** To check the spread among the error to know the variance. **A nice even spread is indicative of linearity.**

Ques5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

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

1. **temp** **0.491693**
2. **yr** **0.244408**
3. **Light-Snow/Rain** **-0.311436**

General Subjective Questions

Ques1: Explain the linear regression algorithm in detail.

Linear regression is a type of predictive analysis that uses a linear equation to find a relationship between a dependent v ariable and independent variables.

When there is a ***single input variable (X)***, the method is referred to as **Simple Linear Regression.**
When there are ***multiple input variables (X_1, X_2,....X_n)***, the method is referred to as **Multiple Linear Regression.**

The linear equation assigns a **scale factor** to each input value, called a **coefficient**, and represented by **Beta (β)**.

One additional **bias coefficient** is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the **intercept**.

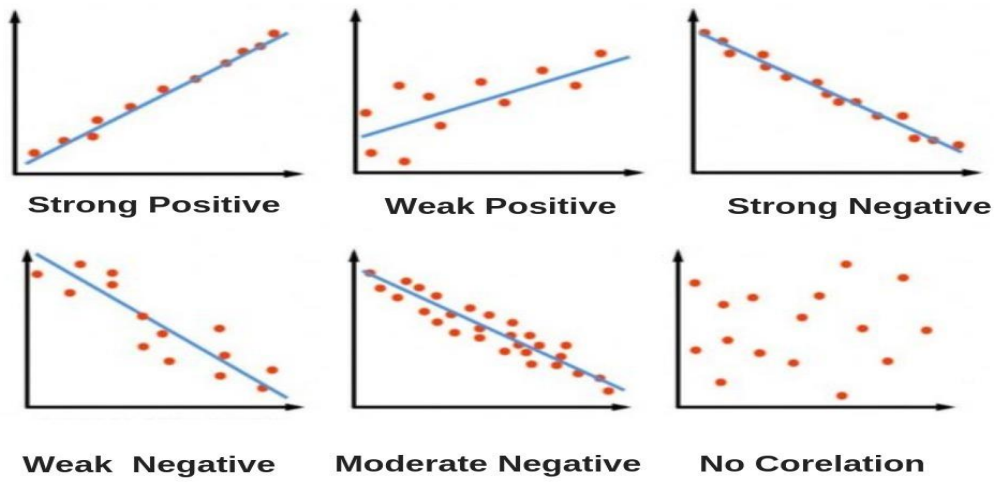
For example, in a **Simple Linear Regression** problem, the form of the model would be:

$$Y = \beta_o + \beta_1.X$$

In higher dimensions,(i.e. In **Multiple Linear Regression**) when we have more than one input variable (X_1, X_2,....X_n), the line is called a plane or a hyper-plane.

$$Y = \beta_o + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 + ... + \beta_n.X_n$$

Using this function, Linear Regression tries to find the ***Best-Fit Line*** **running through most of the data points** while ignoring the noise and outliers and with the minimum cost.



Minimum the loss, better the prediction model.

Ques2: Explain the Anscombe's quartet in detail.

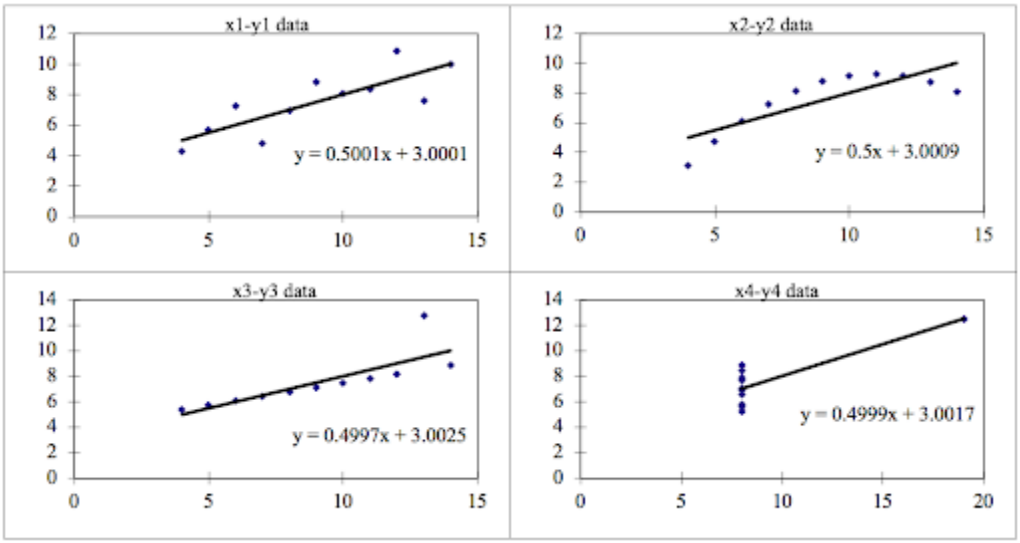
Anscombe's quartet was constructed in **1973** by statistician **Francis Anscombe** to illustrate the importance of plotting data before you analyze it and build your model.

To illustrate this, he used four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when he plotted those data sets, they came out very different from one another.

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	

The statistical information for these four data sets are approximately similar.

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:

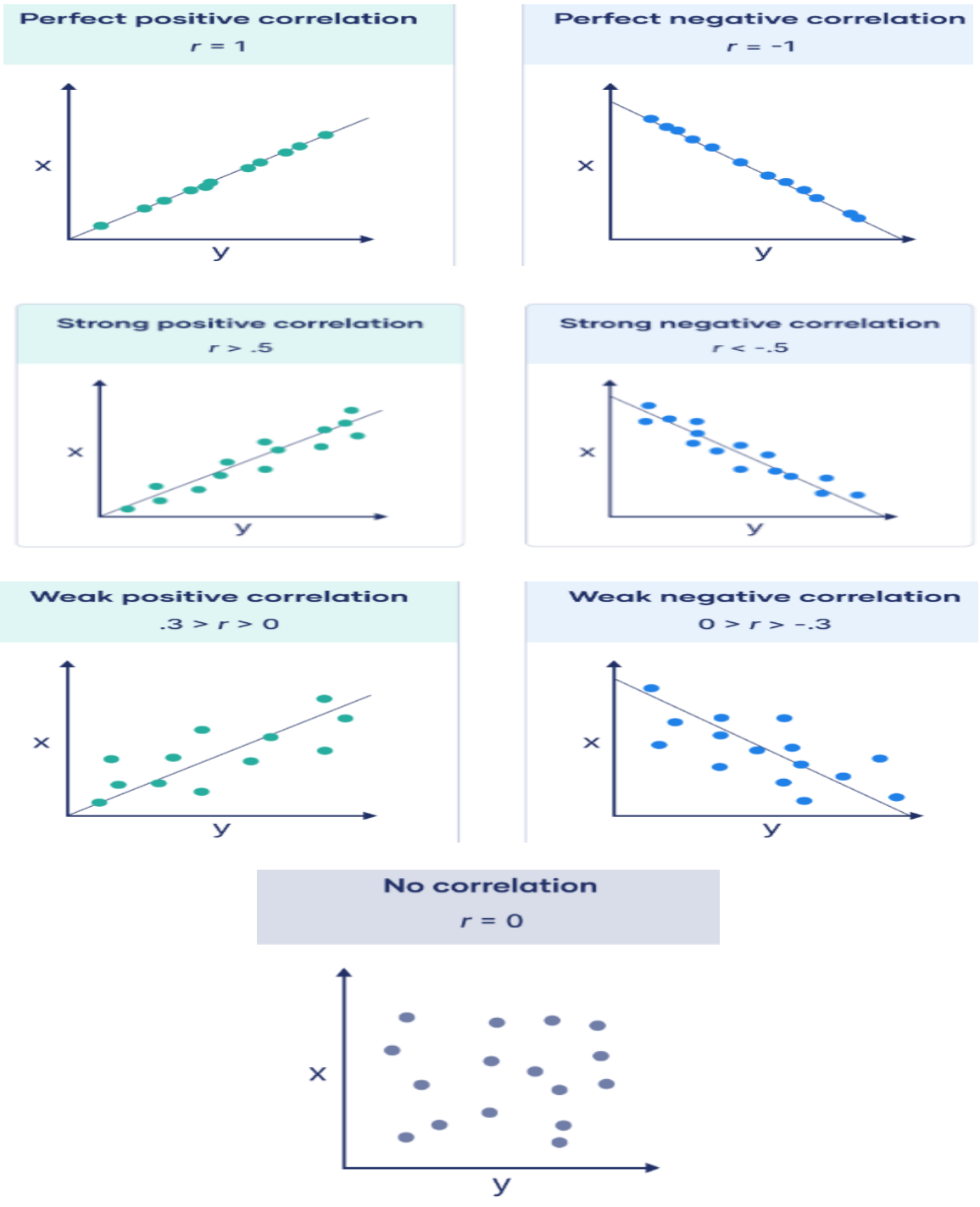


This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Ques3: What is Pearson's R?

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. It is also known as correlation of coefficient. The given image will illustrate the strength and directions using range.



Assumptions of Pearson correlation coefficient:

- Both variables are quantitative
- The variables are normally distributed:
- The data have no outliers
- The relationship is linear

Ques4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a step of data Pre-Processing which is applied to variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

It is performed because most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provide a transformer called MinMaxScaler for Normalization.	Scikit-Learn provide a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

Ques5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

For example in bike assignment, VIF value is very high for atemp and temp because they almost have perfect correlation.

Ques6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

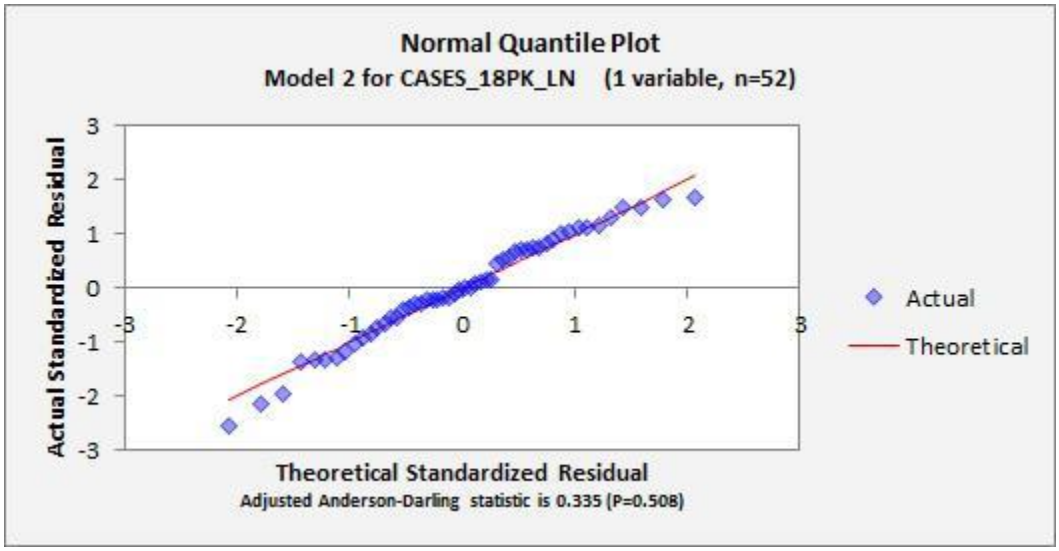
It graphically analyzes and compares two probability distributions by plotting their quantiles against each other.

By comparing the quantiles, we can assess whether or not a set of data potentially came from some theoretical distribution. Here, theoretical distribution can be any. But Q-Q plots are generally made to check the type of distribution (Gaussian Distribution, Uniform Distribution, Exponential Distribution, or even Pareto Distribution, etc.)

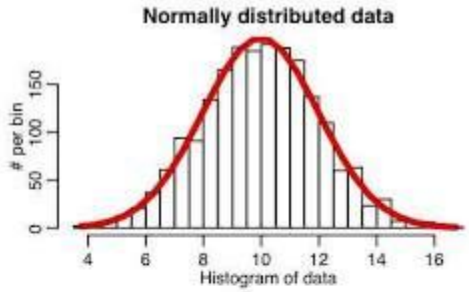
Hence for checking normality, our theoretical distribution will be standardized normal distribution as it is also centered on 0.

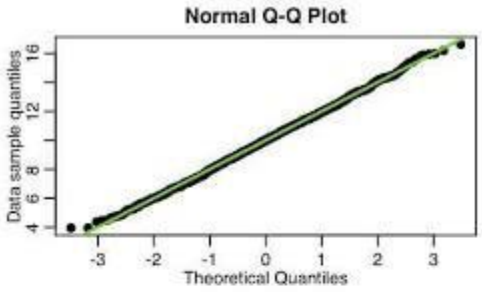
So if the quantiles of residual distribution and quantiles of the standard normal distribution are the same, then the theoretical points of the Q-Q plot will perfectly lie on a straight line y = x.

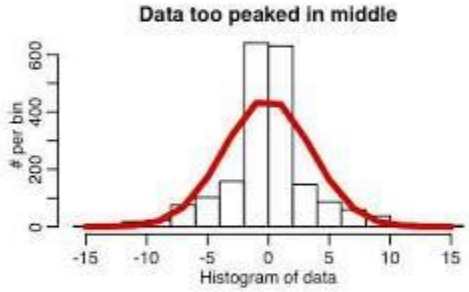
Theoretical points are also called the z-score. QQ-plots find the z-score for both the residuals and Z-distribution and plot a scatter plot of it and then fit a line y=x.

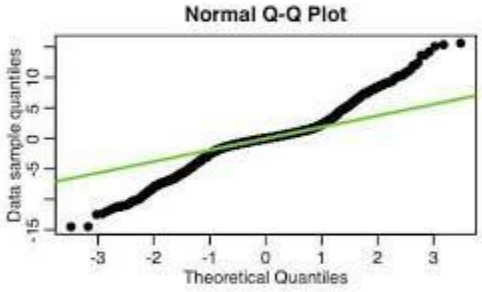


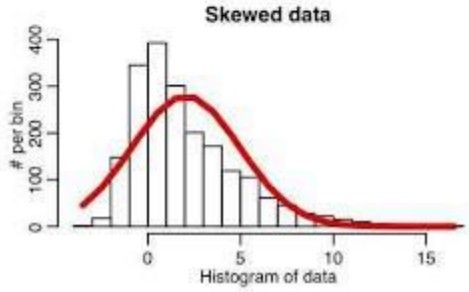
If the theoretical variates do not lie on a straight line then it is clearly understandable that the residual plot is not normally distributed, which states that the model will have Outliers or Fewer Observations or there is a problem of multicollinearity.

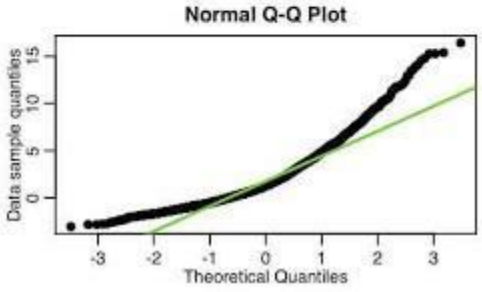












- If S-shaped then data too peaked in the middle.
- If Right-skewed, then concaves upward.
- If Left-skewed, then concaves downward.