Name: Nguyen Hoang Thang

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?

*** Answer:**

There are 7 categorical features ['yr','season','mnth','holiday','workingday','weekday','weathersit'], and after pre-processing them, we have found their impacts on the target variable 'cnt' as following:

- It has seasonal pattern by month and season, but not really by weekday

- It has an overall positive trend by date, month and year

- It is mostly contributed during 'workingday', 'not_holiday' and good weathersit condition such as 'Clear' and 'Mist' cloudy (the better weather condition, the higher cnt)

- Therefore, the model should include the strong relationship between the target 'cnt' and its strong predictors such as: 'year','season','month' ,'workingday' and 'weathersit'

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

*** Answer:**

It is important to use drop_first=True during dummy variable creation, because it is sufficient to let (n-1) dummy variables to interpret all (n) categorical values of a categorical feature and this practice will help to reduce total quantity of model’s features which will help to increase the R-adjusted or model’s performance eventually.

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

*** Answer:**

There are 4 numerical features ['atemp','atemp','windspeed','hum'] those we need to evaluate their relationships with the target 'cnt'. And, the pair-plot shows that **'atemp','atemp'** similarly have 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?

*** Answer:**

A good model holds assumptions on its error terms as normally distributed with center of zero, independent and not following any patterns. Therefore, plotting to visualize the error terms is conducted for this validation step. Its results are summarized as following:

- Errors are scattered, independent and not-linear correlated or not having any special patterns in the scatter plot

- Errors are approximately normally distributed fairly - centered zero in the histogram

- Therefore, the model is quite good to pass the validation and we can move ahead to predict the test data

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

*** Answer:**

Based on the final model, the top 3 features contributing significantly towards

explaining the demand of the shared bikes are: ‘temp’ (coef = 1042.6), ‘yr_2019’ (coef = 996.5) and ‘season_4_winter’ (coef = 503.9).

Remark: ‘yr_2019’ and ‘season_4_winter’ are dummy features from ‘yr’ and ‘season’ original features

General Subjective Questions

1. Explain the linear regression algorithm in detail.

*** Answer:**

Linear regression algorithm is sequential steps used to predict a continuous dependent variable as an output from inputs are one or many independent variables, by fitting the best linear relationship between the output and inputs. It is a supervised learning algorithm.

There are 2 types of linear regression: simple linear regression with 1 input variable (y= a.X + b, with a is a slope and b is an intercept), and multiple linear regression (y = a1.X1 + a2.X2 + a3.X3 + … + $a_i.X_i$ + b, where a_i is a slope of independent variable X_i and b is an intercept).

There are 3 key assumptions for a good linear regression algorithm are:

- Error terms of model output are independent from each other.
- Error terms should be scattered and not have any special patterns or having similar variances.
- Error terms should be normally distributed and centered at zero.

Typical steps to conduct the linear regression algorithm:

- Read and understand data: correlation (pair-plot) for numerical features, distributions of categorical features.
- Data exploratory analysis for the target variable (univariable)
- Data preparation and cleaning
- Categorical features handling (convert dummy variables)
- Split the dataset into train and test dataset.
- Scale to fit and transform the train dataset, transform only the test dataset (applicable to features, not the target variable)
- Features analysis and selection (based on VIF and p-value of each coef, or RFE)
- Select the best features and run the final model.
- Validate the model assumption by residuals analysis: scatter and histogram plots for errors.
- If validation passed, conduct model prediction on test dataset.
- Evaluate the model on test dataset: R2 score, correlation between y-test and y-test-predicted

2. Explain the Anscombe’s quartet in detail.

*** Answer:**

Anscombe’s quartet is a group of 4 different datasets (x,y) but they have the same descriptive statistical parameters such as mean, standard deviation, correlation coefficient, and linear regression line.

This concept is used to emphasize the **importance of looking at data visualization** to comprehensively understand data patterns/outliers, and not only depending on basis descriptive statistical parameters blindly. The following are those 4 datasets, their statistical parameters and graphs for visualization.

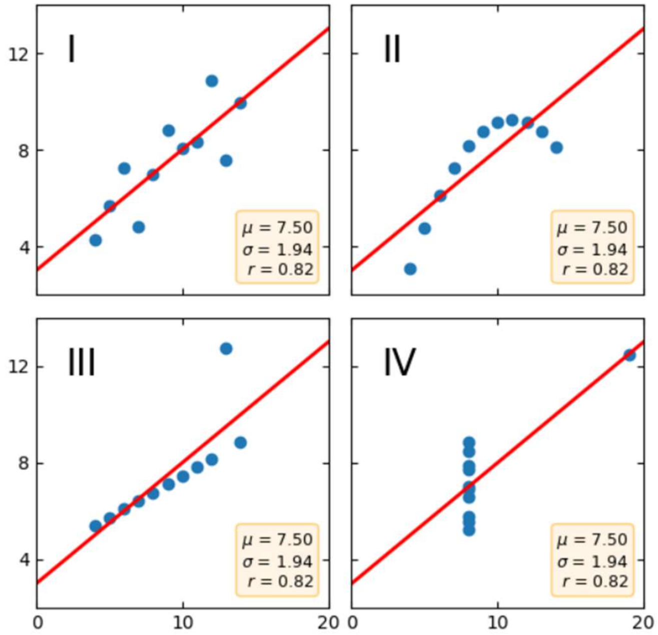
```
Source: https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

datasets = {'I': (x, y1),'II': (x, y2),'III': (x, y3),'IV': (x4, y4)}
```

All 4 datasets have the same statistical parameters as following:

- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 11.0, Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression line: y = 3.00 + 0.50 * x

But, they are actually different and visualized



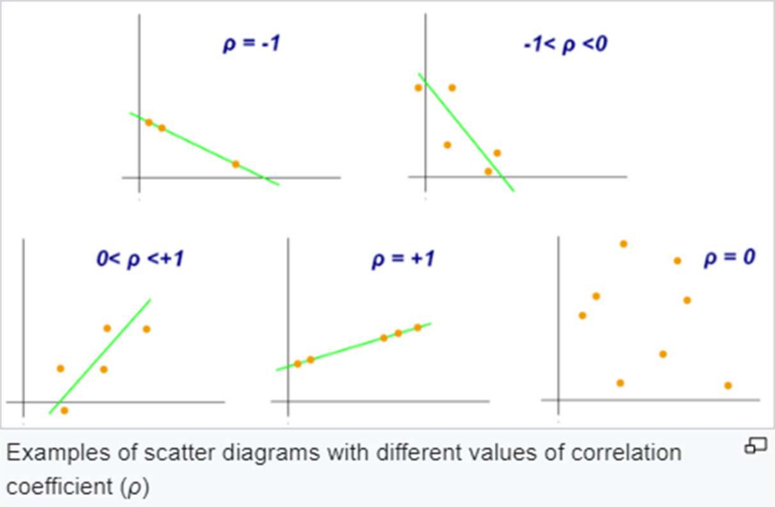
3. What is Pearson's R?

*** Answer:**

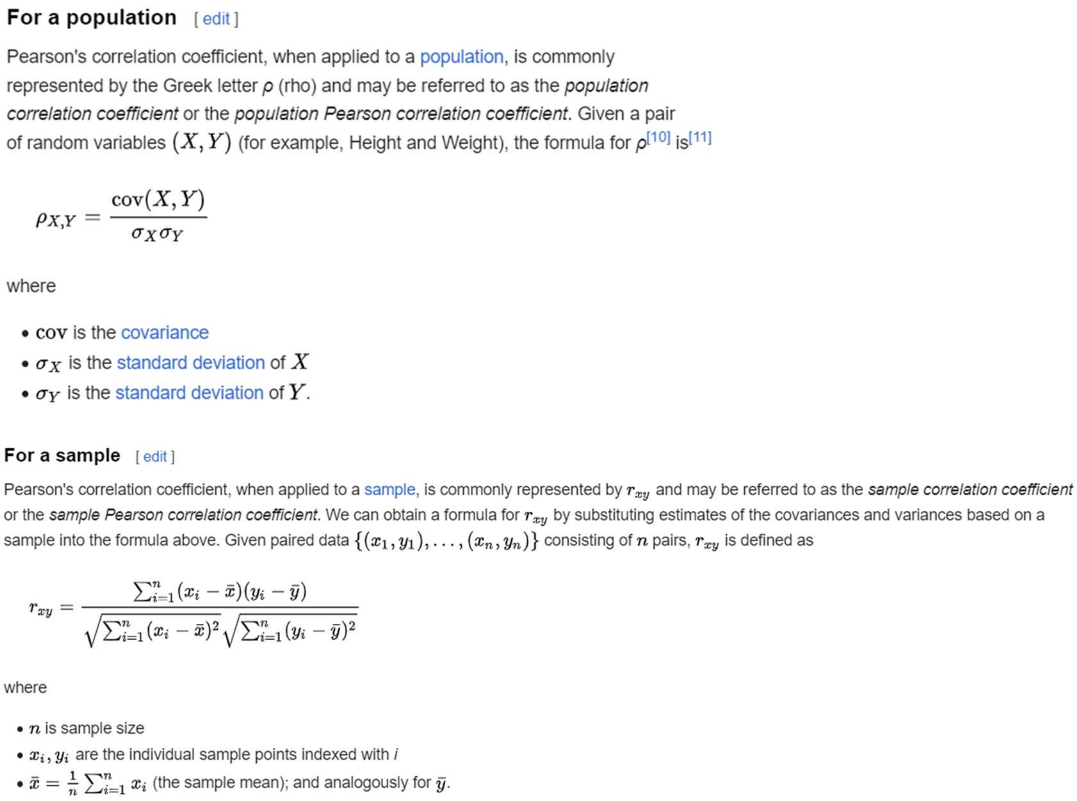
Pearson's R is Pearson correlation coefficient, the indicator measures linear correlation between 2 datasets. Its value is between -1 and +1. It is also the default correlation method calculated in both **Pandas (**dataframe.corr()**)** and **NumPy (**numpy.corrcoef()**).**

Useful examples to illustrate its values' interpretation as following:

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient



Examples of scatter diagrams with different values of correlation coefficient (ρ)



For a population [edit]

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables (X, Y) (for example, Height and Weight), the formula for ρ[10] is[11]

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

For a sample [edit]

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

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

*** Answer:**

Scaling is a method to convert raw values into the same normalized value range such as from 0 to 1 in MinMaxScaler or z-value range in standardized normal distribution (mean = 0, std. dev = 1).

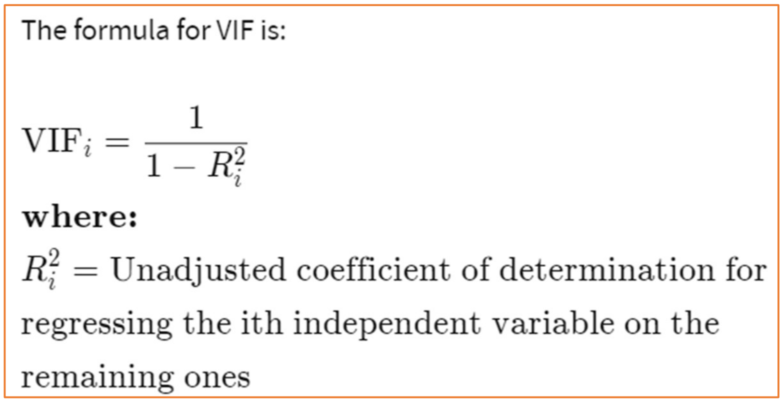
In practice, there are many various features with different levels and value ranges in 1 model. The more complex features' value ranges, the longer time the machine learning model can take to converge to the optimal model's parameters such as in gradient decent optimization algorithm. While scaling the features do not impact the model's optimal output, it can optimize the model efficiency in machine learning. Hence, it is a good practice to scale the features before developing the optimal model.

The big different between the normalized scaling and standardized scaling is the scaling range. Normalized scaling or Min-Max scaling will convert raw values into range 0 and 1 based on dataset's min and max value. While standardized scaling will convert raw values into approximated z-value in standardized normal distribution (mean = 0, std. dev = 1), and there is no specific range for z-value (scaled value)

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

*** Answer:**

Yes, I have found some dummy features of weekday, workingday, holiday etc. have VIF = inf. Just recall the formula of VIF(i) of each independent feature as below, we have found that, VIF = inf only when R2(i) = 1 or when the variances of a feature (i) are fully 100% explained by other features, because it is fully dependent on them. In this case, we absolutely should remove it from the features selection step.



The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

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

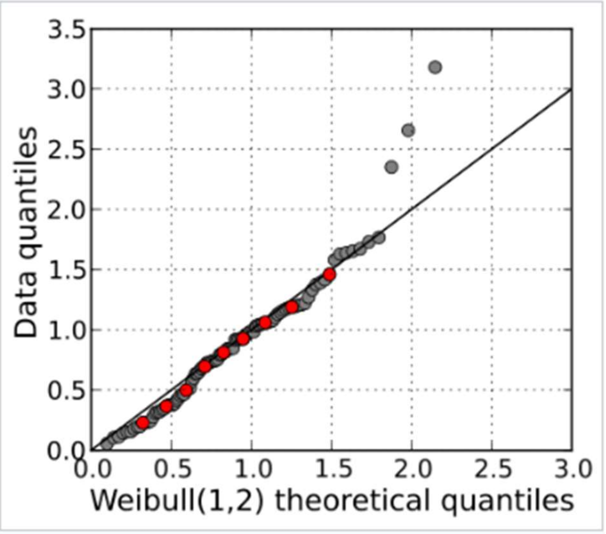
*** Answer:**

Q-Q plot is a graphical technique to determine if a dataset follows a particular theoretical distribution, such as the normal distribution. The Q-Q plot is created by plotting the quantiles of the observed dataset against the quantiles of the expected distribution (theoretical quantiles).

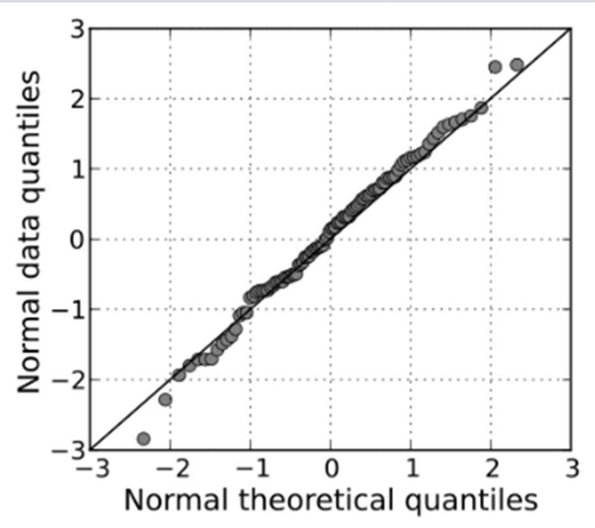
In use, if the data follows the expected distribution, the points on the Q-Q plot should fall closely along a straight line. Else, the points on the Q-Q plot deviate away from a straight line, it indicates that the distribution of the data is different from the expected distribution. Illustrative examples are given as following charts.

From this, we can use the Q-Q plot to visually validate the important assumption on error terms of the linear regression algorithm, that errors follow the normal distribution. Because this assumption is importantly required to ensure the hypothesis testing and linear regression algorithm work correctly, a good validation tool like Q-Q plot can be considered as an important tool to be used in linear regression for normality assumption validation.

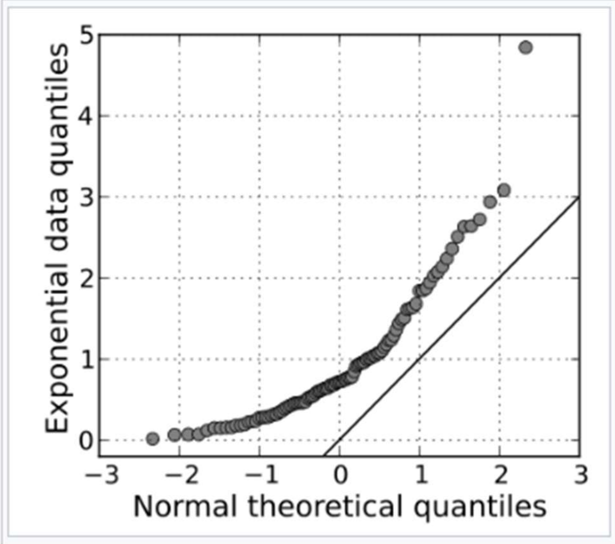
Source: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot



A Q–Q plot of a sample of data versus a Weibull distribution. The deciles of the distributions are shown in red. Three outliers are evident at the high end of the range. Otherwise, the data fit the Weibull(1,2) model well.



A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.



A normal Q–Q plot of randomly generated, independent standard exponential data, ($X \sim \mathrm{Exp}(1)$). This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal ($X \sim \mathrm{N}(0,1)$). The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7