Assignment-based Subjective Questions and answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- During spring people rent bikes less. Fall has the highest median rental count among all the seasons. People tend to go out more during optimal conditions.
- The number of bikes rented has been increasing over the years. The median rental count of 2019 is way too higher than that of 2018.
- Demand for bikes increases from jan to june but later from september onwards it decreases.
- During holidays people tend to rent bikes less. Probably because they do not have to travel to their working places.
- The demand for bikes is more or less same for all the weekdays.
- The median of rental count of working days and non- working days are almost similar, however the fences of non working days are slightly higher than that of working days.
- During clear weather situation the demand is high and very less during bad weather conditions like light snow/storm/rain
- During September (9th month) bike sharing is at the peak and in the beginning and ending of the year it is least.

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

While creating dummies for a categorical nominal variable with ‘n’ levels, we do not need n dummy columns to convey the information contained in the initial categorical column, we just need n-1 dummies. So we can drop one column, which is done by ‘drop_first=True’.
If we do not remove one dummy from the dummies then that will lead to redundancy of information and thereby multicollinearity.So we can drop one dummy as the other dummies will take care of the dropping dummy's information.

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

Variable ‘temp’ has the highest correlation with target variable ‘cnt’
(with a correlation coefficient of 0.63)

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

I plotted a distplot of error terms and made sure that they follow a normal curve with mean centered around zero.
I checked the VIF values of selected features to make sure that there is no trouble of multicollinearity.

Linearity of the model has been checked by evaluating a **Residuals vs. Fitted** plot.
I also made sure that there is no overfitting of the data.

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

- 'temp' - (0.548046)
- 'weathersit_light-Snow/Rain/Storm' - (-0.283837)
- 'yr' (0.232786)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a tool which can be used to train a model to predict data based on some variables. It is used to explain the relationship between independent variable and dependent variable with a straight line. With this we could predict a quantitative response Y from the predictor variable X.
Mathematically,
$Y = mX+c$
b -> slope of the line
X -> independent variable
Y -> dependent variable
C -> intercept

Initially our data set is divided into train and test,the former one is used to train the model while later is used to evaluate the predictive capacity of the model. The train dataset is then divided into features/independent and target dataset. After dividing the dataset, a linear model is fitted using the train dataset. A gradient descent algorithm is used to find the coefficients of features in the model. Gradient descent algorithm works by minimizing the cost function and finds a best fitting line for our data. When the target variable is dependent on multiple features then the predicted variable is not a straight line instead it is a hyperplane.Then this model is checked and verified over a test dataset.

Multiple linear regression follows a mathematical formula

$$yi = \beta 0 + \beta 1xi2 + \beta 2xi2 + \ldots\ldots + \beta pxip + \epsilon$$

Where **yi** is the dependent or predicted variable

β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.

β1 and **β2** are the regression coefficients representing the change in y relative to a one-unit change in **xi1** and **xi2**, respectively.

βp is the slope coefficient for each independent variable

ϵ is the model's random error (residual) term.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is four data sets, each with two variables (x and y).The property of this quartet is that these four data sets have exact mean and standard deviation for variable x and nearly same mean and std for variable y.Other descriptive statistics of these data sets are also nearly identical. However when you plot them they look entirely different. So Anscombe's quartet demonstrate the importance of plotting data while analyzing.This was constructed byFrancis Anscombe to prove that numerical descriptions alone are not enough to understand the data.
data from wikipedia

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

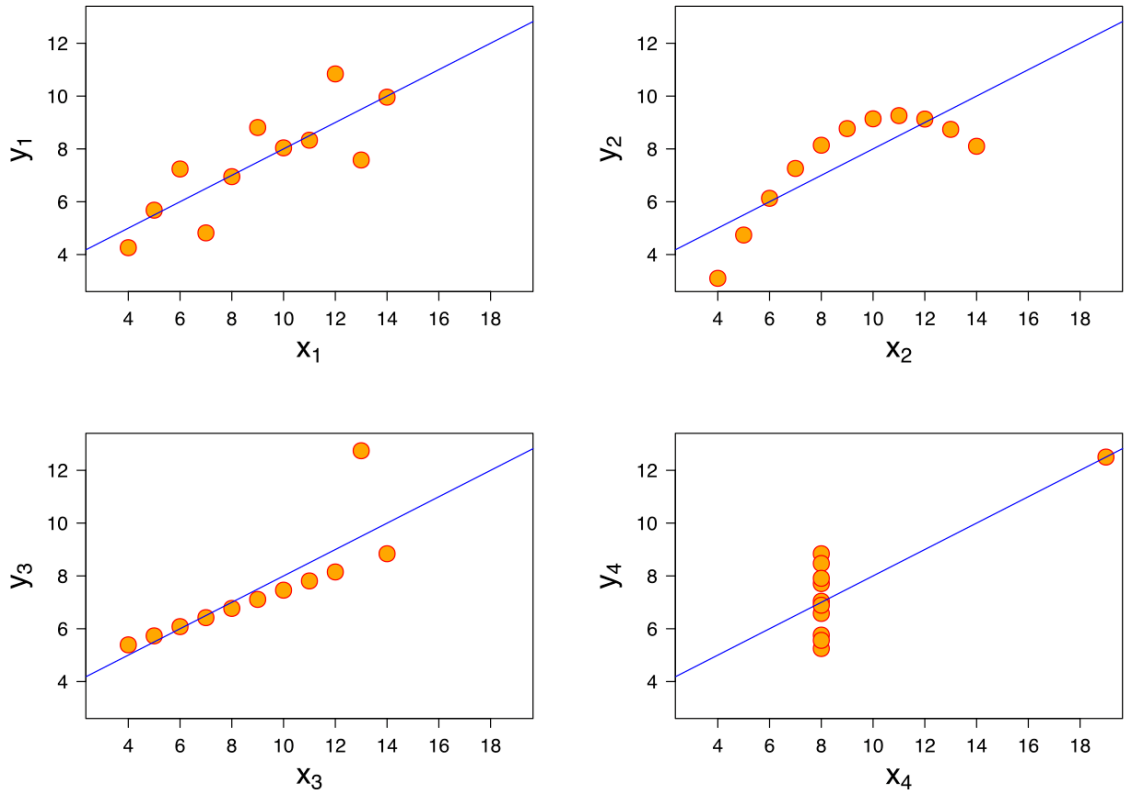

Image from wikipedia

Here first plot - variables follows a linear relationship
Second plot - non-linear relationship between variables
Third plot - one point is very far away from fitted line, follows linear relationship but different from the first plot
Fourth plot- All the x values are same except one on the right side, doesn't follow linear relationship

3. What is Pearson's R?

It is a measure of strength of association between variables.It is calculated by dividing the covariance of the variables by the product of their standard deviation. The range of Pearson's r is from -1 to +1.

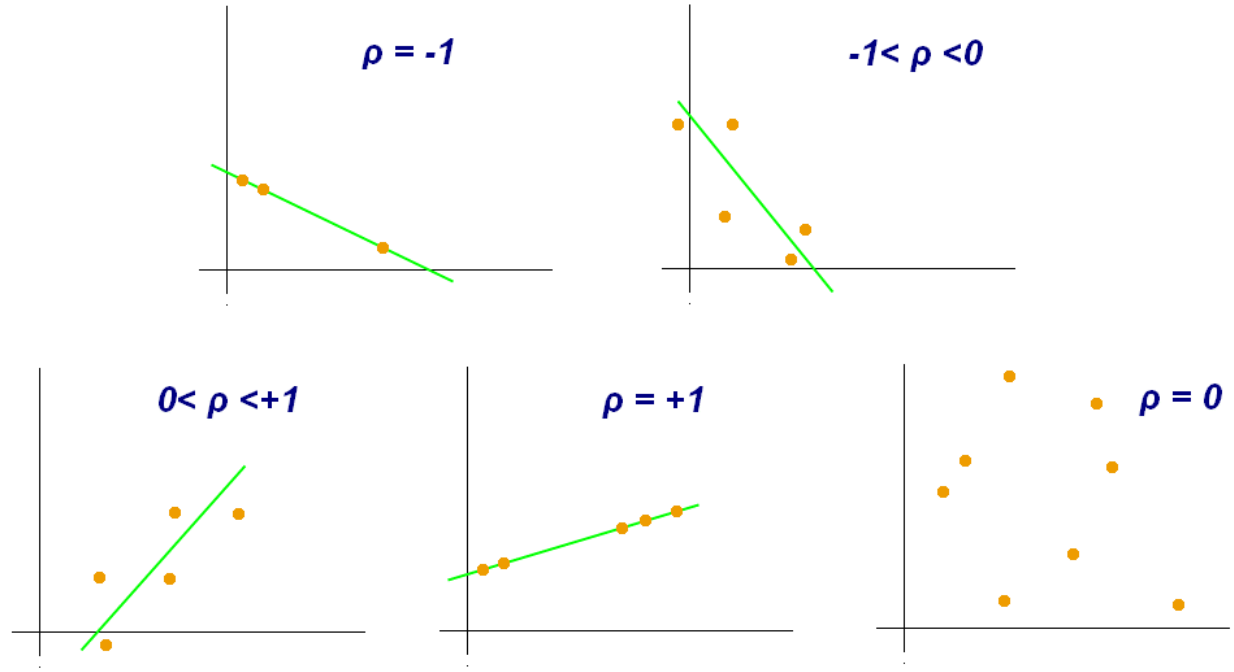

Image from wikipedia

- A pearson's r of +1 - total positive correlation between variables, means that with one variable the other variable will also increase (plot4).
- A value of 0 means that there is no correlation between the variables(plot 5)
- Pearson r of -1 means that the variables are totally negatively correlated and with one variable the other one also decreases(plot 1)

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

Scaling is process in which independent variables are normalized within a range, before building the model. Most often we have features which vary in high magnitude. If scaling is not performed on these features, then since the regression algorithm considers only the magnitude and not the units, it would lead to incorrect modeling. Model will give higher importance to greater values, irrespective of their units.In order to solve this issue we have to scale all the variables to the same level of magnitude.Scaling just affects the coefficients and none of the other parameters are affected.

Normalizaion/Min-Max scaling

—----------

Brings the data in a range of zero and one.
SKlearn.preprocessing.MinMaxScaler helps to scale the data in python.
Scaling is done using the following formula
X = (x - min(x)) /(max(x) - min(x))

- Values are bounded between zero and 1

- Outliers are also scaled

Standardized scaling
—----------
Standardization replaces the values with their z score.
It brings all of the data into a standard normal distribution.
Scaling formula
X = x-mean(x)/sd(x)

- Values are not bounded and it doesn't affect the outliers

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

Variance Inflation Factor (VIF) value is a measure of the degree of multicollinearity.
When there is a perfect correlation between features, you would see an infinite VIF value.
Ie, the formula for VIF is

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R2 value becomes one in case of perfect correlation, so the denominator becomes zero and thereby the VIF value infinite. Infinite vif shows the occurrence of multicollinearity, that is the features are totally correlated.

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

Q-Q plot is a graphical tool used to understand if two datasets are coming from the same distribution or not. Quantiles of two probability distributions are plotted against each other(scatter plot) for comparing and analyzing them.On the Q-Q plot a reference line is also plotted at an angle of 45 degree. The points on the scatter plot will align around this reference line if the datasets being compared are from the same population with the same distribution, else the points will not align along the line. A Q-Q plot is used in checking if two data sets come from populations with common distribution,whether the two sets have the same location and scale,Whether the two data sets have similar distributional shapes and if they have similar tail structure. Q-Q plots are useful in case of linear regression if you have received the train and

test datasets separately and if you want to make sure that these two datasets are coming from populations with the same distribution.

Submitted by
Akshay P