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 –

1. Count of bikes hired is greatest in clear weather followed by misty weather. The least number of bikes are hired in light snow. There are 0 bikes hired when there is heavy rain.
2. The median values of bikes hired on weekdays is almost the same. However the maximum of the bikes hired is on weekday_6.
3. Least number of bikes are hired in spring and maximum is during fall seasons.
4. Number of bikes hired was far greater in 2019 as compared to 2018.
5. Month when the demand of bikes is the highest is in September and it is lowest in the month of January.
6. Overall distribution of bikes rented on holidays is lower than the working days.
7. The median for both working days and non-working days are the same. But the max. number of bikes hired are on non-working days.

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

Ans-

When we use **drop_first = True** during the creation of dummy variables, the first categorical variable is dropped and remaining (n-1) categorical variables are used. This will be advantageous because:

there will be less number of categorical variables generated.

the correlation will be bit lower between the categorical variables as all the variables will not be having similar patterns of values.

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

Ans-

‘registered’ has highest correlation with the target variable named as cnt.

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

Ans -

The following assumptions are considered:

1. Linear Relationship between dependent and independent variables:: For this we can create a scatter plot between the independent and dependent variables.
2. Normality of the residuals:
We can plot the residual values using qq-plot. A straight line will show the normality. Else, distplot can also be used.
3. No or little multicollinearity: We can check the VIF values of the independent variables.
4. Homoscedasticity:
It is to check if error is constant along values of the dependent variable. A scatter plot and a constant line from 0 in y-axis is drawn and the deviation of error from zero-line is checked.
5. All independent variables are uncorrelated with error terms: Scatter plot is drawn between independent and residuals.
6. Observations of the error terms are uncorrelated with each other:
It is to check whether there is a correlation inside the observations of the error term. A line graph of residuals is plotted.

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

Ans-

1. Workingday
2. Year-2019
3. Weekday-6

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans-

Linear Regression is used to predict value of the dependent variable based on the values of other independent variables. It is based on the assumption that it is a linear model where the dependent variable follows a linear relation with the independent variable in the form of y = mx+c. If there are more than one independent variable, is called multiple linear regression and is in the form: y = (m1*x1) + (m2*x2) + … + (mn*xn) + c. It is mainly calculated using Ordinary Least Squares (OLS), Gradient Descent or Regularization methods.

Gradient Descent Algorithm:

First we define the cost function as

J(m,c) = sum(y-yi)^2 = sum(y-mxi-c)^2. For finding the values of m and c, we partially differentiate the equation w.r.t. m and c seperately. Then we apply the gradient descent method, where we iterate till the required precision is achived over the equations:

m_current = m_current - (learning_rate * m_gradient)

c_current = c_current - (learning_rate * c_gradient)

Learning rate is defined as low as possible so that it becomes easier in each iteration to move closer to the desired solution. Finally, we store values for each iteration in a dataframe and can find desired intercept and slope values that represents the best fit line for our training data, once the loop gets over.

2. Explain the Anscombe's quartet in detail

Ans-

It consists of four graphs which have similar descriptive statistics like mean, variance, standard deviation etc. but if we plot the graphs for checking the distribution of the data, we see it is quite different.

Example:

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

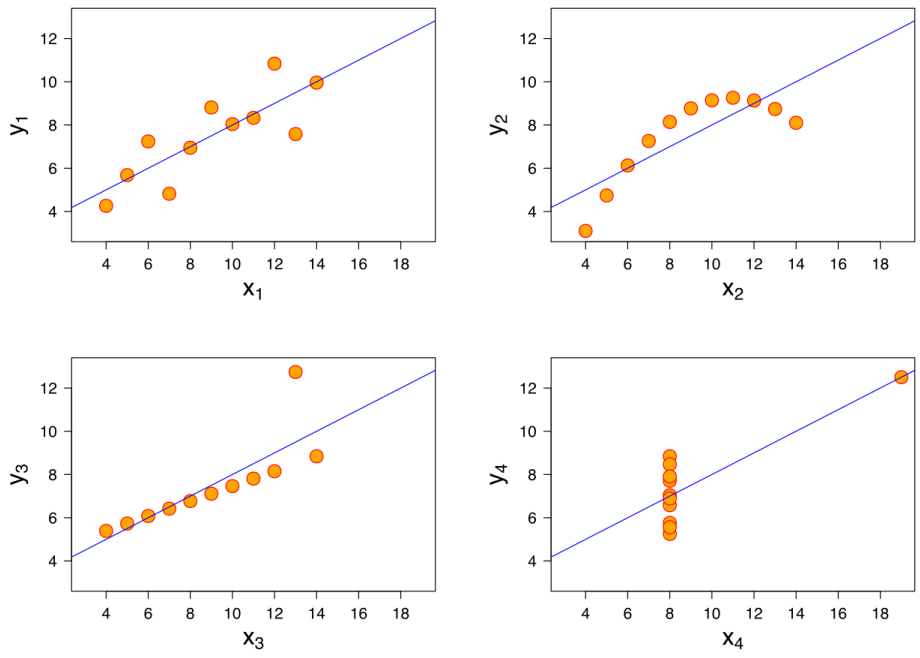

Image and data source: Wikipedia

3. What is Pearson's R?

Ans -

Pearson's R or Pearson's Correlation Coefficient is used to determine how much one variable is dependent on other variable(s). It is mostly used in the case of linear relations.

The formula for Pearson's R is

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

The values of correlation coefficient ranges from -1 to 1, where 1 indicates strong positive correlation (value of y increases with increase of x), 0 indicates no correlation and -1 indicates strong negative correlation (value of y decreases with increase of x).

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

Ans -

Scaling is the process of transforming the data to make it fit in a specific scale.

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

Normalized scaling brings all of the data in the range of 0 and 1. Standardized scaling replaces the values by their Z scores. It brings all of the data to a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans -

VIF = 1/(1-r^2)

Now, a VIF of infinity means r^2 is 1. It means correlation coefficient can be +1 or -1. This suggests that there is perfect correlation between the dependent and independent variable.

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

Ans –

Q–Q (quantile-quantile) plot is a probability plot for comparing two probability distributions by plotting the quantiles of one dataset against other dataset. The point is to understand whether the data present in two datasets has come from a common distribution.