

Assignment-based Subjective Questions

by Nitin Jayan

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

A: Trend as per categorical variables.

- **Year** – Significant increase in usage in year 2019 compared with 2018.
- **Seasonal** – The bike usage follows a similar trend in both the years.
 - ✓ Fall (season 3) has the highest usage
 - ✓ Spring (season1) has the lowest usage
 - ✓ Winter and summer enjoy almost similar usage rates.
- **Holiday** – Holiday = 0 has more usage than Holiday = 1. (boxplot)
- **Weekday** – Very consistent usage rates across all weekdays. (boxplot)
- **Monthly** – Both years have roughly similar trends in usage across months in that the usage increases till the month of June. However,
 - ✓ The usage in 2018 falls steadily after June.
 - ✓ The usage in 2019 spikes in months of Aug and Sept before falling steadily again.
- **Weathersit** –
 - ✓ Weather type 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) has the highest usage.
 - ✓ Weather type 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) also enjoys high usage rates.
 - ✓ Weather type 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) has scarce usage comparatively.
 - ✓ Weather type 4 (Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog) has no usage.
- **Working Day** – Not much of a difference depending on the values of Working Day (1 or 0). It can be observed from the box plot that median is also the same.

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

A: If you don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted.

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

A: 'Registered' variable has the highest correlation with cnt.

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

A:

- P – Value of all the variables in the final model < .05

- VIF Values of Variables ≤ 5
- Goodness of Fit: R Square = 0.816.
- Adjusted R Square = 0.812.
- F-Statistic = 221.2
- Prob (F-Statistic) = $2.97 \times 10^{-176} \approx 0$

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

- temp
- hum
- yr_1

General Subjective Questions

1. Explain the Linear Regression Algorithm in Detail.

A: - Linear Regression Algorithm is a machine learning algorithm based on supervised learning. A kind of Regression analysis, which is a technique of predictive modelling that helps you to find out the relationship between Input/ multiple input variables and the target variable. Linear Regression can be used to identify: a) Effect of featured variable to target variable b) Change in Target variable w.r.t input variables combined or keeping others constant. c) Prediction/Forecasting.

- **Simple Linear Regression:** The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.
- **Multiple Linear Regression:** Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables). The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

- **Approaches/Methodologies:**

- ✓ **BEST FIT LINE:** Minimising the expression of Residual Sum of Squares which is equal to sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

Residuals

$$Y = B_0 + B_1X \quad B_0 - \text{Intercept} \quad B_1 - \text{Slope}$$

- ✓ **Using Ordinary Least Square** method to minimize the RSS = The summation of $(Y' - B_0 - B_1X)^2$ from $l = 1$ to $l = n$

Where Y' is the predicted Target Variable.

The efficacy of Linear Regression is assessed by following two metrics, namely:

1. Rsquare = Coefficient of Determination = $(1 - \text{RSS}/\text{TSS})$

Where RSS – Residual Sum of Squares

TSS = Sum of Errors of the data from mean (Summation of $(y_{\text{pred}} - y_{\text{mean}})$).

This explains what portion of the given data variation is explained by the developed model. The value Lies between 0 – 1.

2. Performing the steps for linear regression model:

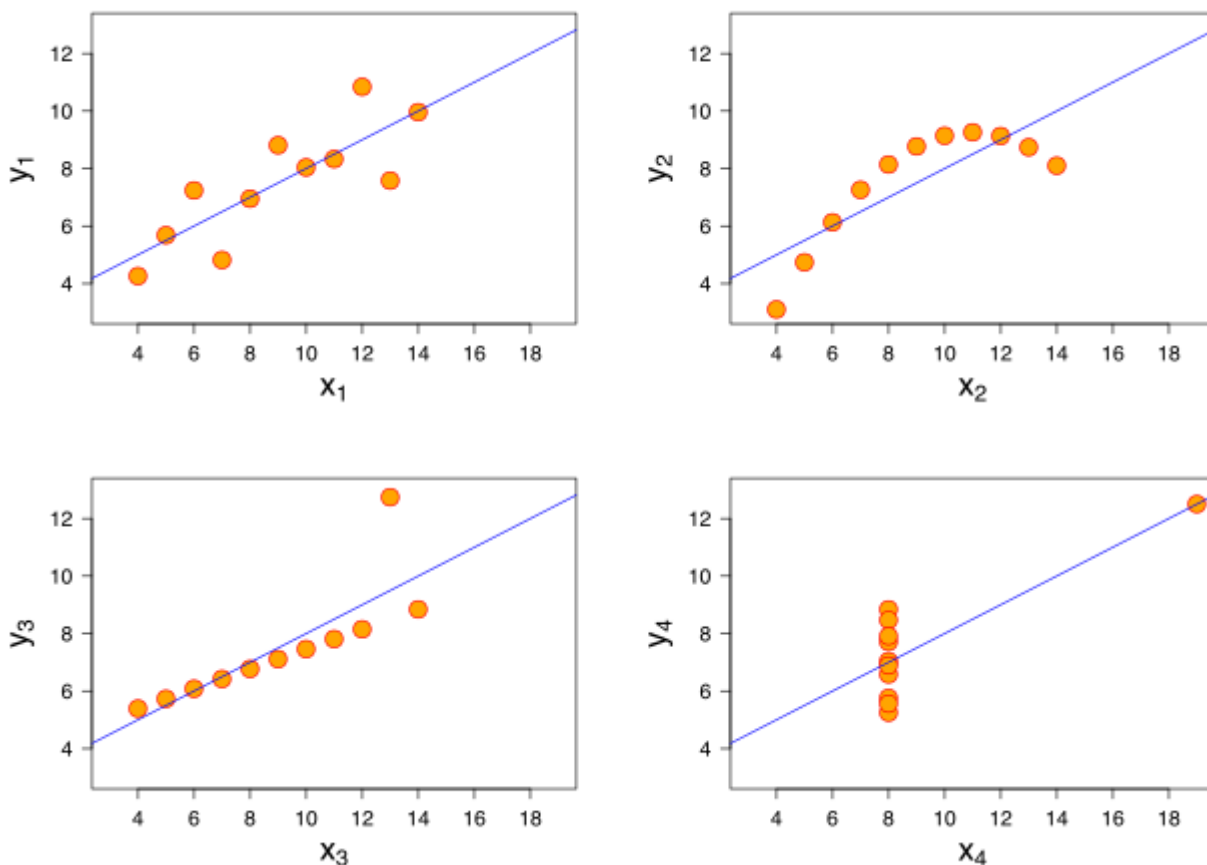
- a) Importing data set
- b) Understanding the data frame.
- c) Preparing Features and Target Variable
- d) Splitting of dataset – Training and Test data
- e) Applying model Linear Regression
- f) Coefficients Calculation
- g) Making Predictions
- h) Model Evaluation (Actual Vs. Predicted)
- i) Model Evaluation (Error Terms)
- j) Checking Mean Square Error and R Square

3. in Multiple Linear Regression

- a) Equation: $Y_{\text{Pred}} = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n$
- b) Regression Results from OLS

2. Explain the Anscombe's Quartet in Detail:

A: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet emphasizes on the importance of looking data graphically before starting to analyse according to a particular type of relationship.


3. What is Pearson's R?

A: Pearson correlation coefficient (PCC, pronounced /'piərsən/), also referred to as Pearson's R:


Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
 - -1 indicates a strong negative relationship.
 - A result of zero indicates no relationship at all.
- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
 - A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
 - Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related

Pearson Correlation Coefficient



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



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

A: Scaling is used to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.
- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

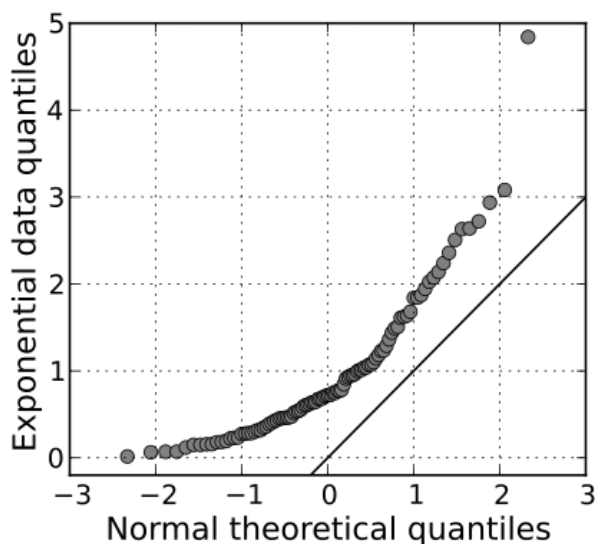
5. You Might have observed that VIF value is infinite. Why does it happen?

A: VIF i.e Variance Inflation Factor is to identify the co linearity between features within a multiple regression model. IF there is a perfect correlation then the VIF value is infinity as VIF is the ration of variance between all given models beta divided by single variance beta.

6. What is a QQ plot ? Explain the use and importance of a QQ Plot in linear Regression?

A: A Q-Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The points plotted in a Q-Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q-Q plot follows the 45° line $y = x$. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q-Q plot follows some line, but not necessarily the line $y = x$. If the general trend of the Q-Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis.



We can investigate further in three ways: a density plot, an empirical CDF plot, and a normality test. Note that one should generally do the former two after the qq plot, as it's easiest to see that there are departures from normality in a Q-Q plot, but it is sometimes easier to characterize them in density or empirical CDF plots.