LINEAR REGRESSION SUBJECTIVE QUESTIONS

HOANG NGOC TIEN – MASTER OF DATA IN DATA SCIENCE PROGRAME

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Answer: Some insights from my analysis of the categorical variables:

- ✓ Summer and fall season seem to have attracted more booking.
- ✓ Number of booking increases drastically from year 2018 to 2019
- ✓ Most of booking have been done in month 6, 7, 8, 9, especially high in September.
- ✓ Number of booking is higher in non-holiday time, as people tend to spend time for other activities in holiday.
- ✓ There is not much different in booking on different days of week
- ✓ Clear/nice weather obviously attracted more booking
- ✓ Booking is about the same on working and non-working day.

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Answer: **drop_first=True** is important to use during dummy variable creation, because of following reasons:

- ✓ **Multicollinearity Reduction**: Multicollinearity refers to a situation where two or more independent variables in a regression model are highly correlated with each other. When creating dummy variables for a categorical variable with multiple categories, it's common to introduce multicollinearity because one category's value can be predicted from the others. By setting **drop_first=True,** we eliminate one of dummy variables presenting a category. This also eliminates the multicollinearity issue.
- ✓ **Dimensionality Reduction**: When you create dummy variables for a categorical variable with "n" categories, you typically create "n-1" dummy variables. This is because the information for the "nth" category can be derived from the values of the other "n-1" categories. By dropping one of the categories, you reduce the dimensionality of your dataset, which can be beneficial when working with large datasets and can make regression analysis more computationally efficient.

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Answer: **temp/atemp** variable has the highest correlation with the target variable

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Answer: There are 5 assumptions that I did validate:

- ✓ Normality of Error Terms: using histogram of residuals to assess normality
- ✓ No or Little of Multicollinearity: calculating VIF of each independent variable. VIF values should be < 5
- ✓ Homoscedasticity: Plot residuals against predicted values (a scatterplot)
- ✓ No Autocorrelation of Error: Examine residual plots, autocorrelation plots to check for autocorrelation
- ✓ Independence of Error: Using scatterplots of each independent variable against the residuals. There should be no visible pattern of relationship

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes:

- ✓ temp: Temperature affects the demand the most as its highest coef.
- ✓ year: Year is second feature contributing significantly towards explaining the demand.
- ✓ Light_snow_rain (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) : Bad weather negatively affects the demand the most as its lowest coef (-)

GENERAL SUBJECTIVE QUESTIONS

Question 1: Explain the linear regression algorithm in detail

Answer: Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors) by fitting a linear equation to the observed data. It is commonly used for both regression (predicting a continuous numerical value) and simple classification tasks.

Simple Linear Regression

In simple linear regression, there is one independent variable (X) and one dependent variable (Y). The goal is to find a linear relationship between them. The linear regression model is represented by the equation:

$$Y = b0 + b1*X$$

Y: the dependent variable

X: the independent variable

b0: the interception

b1: the slope of the line

GENERAL SUBJECTIVE QUESTIONS

Question 1: Explain the linear regression algorithm in detail

Answer (continue):

Multiple Linear Regression

In multiple linear regression, there are multiple independent variables (X1, X2, X3, ..., Xn) and one dependent variable (Y). The model is represented as follows:

$$Y = b0 + b1*X1 + b2*X2 + \ldots + bn*Xn$$

Y: the dependent variable
X1, X2, X3 .. Xn: the independent variables
b0: the interception
b1, b2, b3, … bn: are the coefficients (slopes) associated with each independent variable

GENERAL SUBJECTIVE QUESTIONS

Question 1: Explain the linear regression algorithm in detail

Answer (continue):

Applications of Linear Regression

- ✓ Linear regression is widely used in fields such as economics, finance, healthcare, and social sciences for predictive modeling and causal analysis.
- ✓ It serves as the foundation for more advanced regression techniques and machine learning algorithms.
- ✓ It is also used for simple classification tasks when the dependent variable is categorical.

Linear regression is a straightforward and interpretable algorithm, making it a valuable tool for understanding and predicting relationships in data. However, its simplicity also means it may not capture complex, nonlinear relationships. In such cases, more advanced regression techniques may be more appropriate.

GENERAL SUBJECTIVE QUESTIONS

Question 2: Explain the Anscombe's quartet in detail

Answer: **Anscombe's quartet** is a famous statistical dataset, that were constructed in 1973 by the statistician Francis Anscombe to illustrate the importance of data visualization and the limitations of relying solely on summary statistics, such as means and variances, to understand data. The dataset consists of four subsets, or "**quartets**," of data, each with 11 data points. Despite having similar summary statistics, these subsets exhibit different patterns and relationships when visualized.

GENERAL SUBJECTIVE QUESTIONS

Question 2: Explain the Anscombe's quartet in detail

Answer (continue):

The Four Quartets:

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

GENERAL SUBJECTIVE QUESTIONS

Question 2: Explain the Anscombe's quartet in detail

Answer (continue):

Summary Statistics of The Four Quartets is about the same as bellow:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s2	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y: s^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R2	0.67	to 2 decimal places

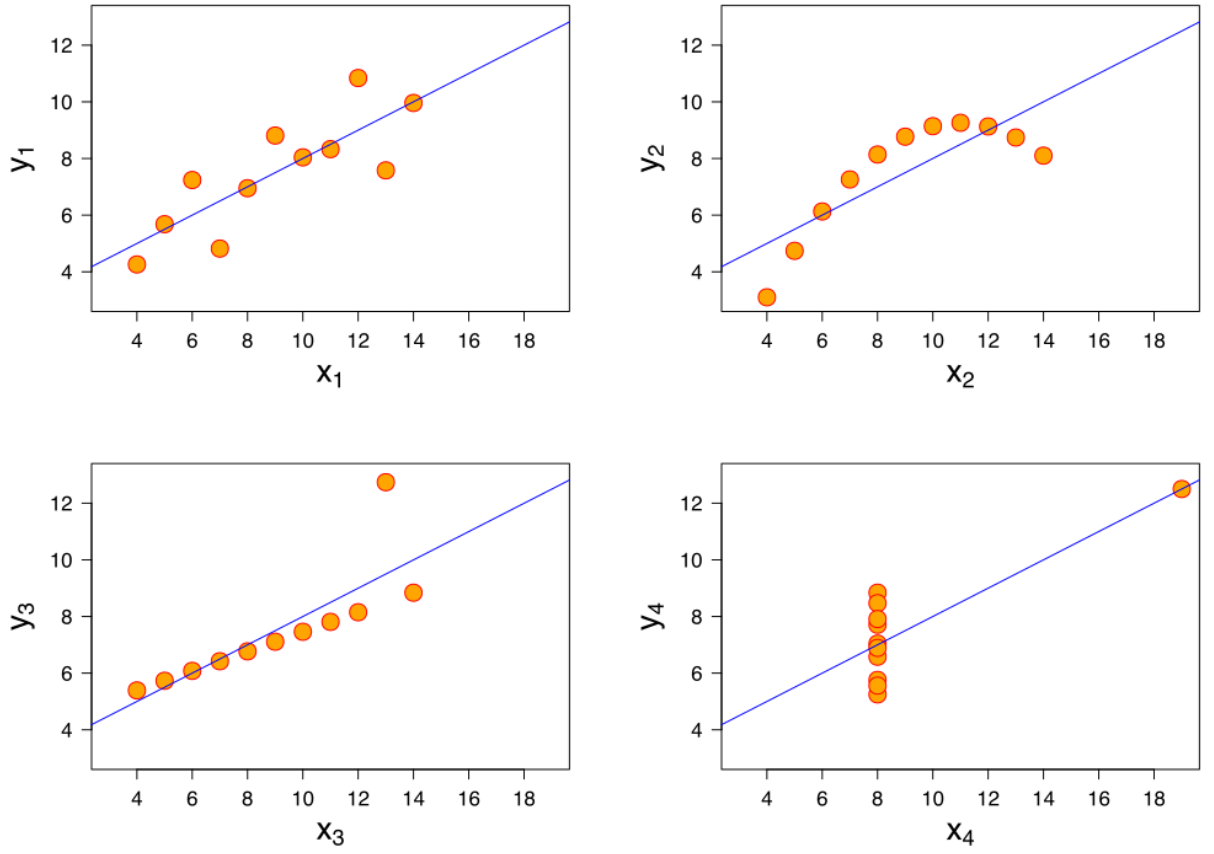
GENERAL SUBJECTIVE QUESTIONS

Question 2: Explain the Anscombe's quartet in detail

Answer (continue):

Visualisation of The Four Quartets:

- ✓ Quartet I: represents a linear relationship
- ✓ Quartet II: demonstrates a curvilinear relationship
- ✓ Quartet III: shows a linear relationship but with one outlier that significantly deviates from the linear trend
- ✓ Quartet IV: no apparent relationship between X and Y



GENERAL SUBJECTIVE QUESTIONS

Question 2: Explain the Anscombe's quartet in detail

Answer (continue):

Key Lessons from Anscombe's Quartet :

- ✓ Summary Statistics can be deceptive
- ✓ Data visualization is very important
- ✓ It's crucial to consider the context of the problem when visualizing data
- ✓ Outliers Can Have a Significant Impact
- ✓ Don't make unwarranted assumptions about data. summary statistics may not imply similar data patterns.

GENERAL SUBJECTIVE QUESTIONS

Question 3: What is the Pearson's R?

Answer: Pearson's correlation coefficient, often denoted as **R** or **Rxy,** is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It quantifies how well the variation in one variable can be explained by the variation in another variable.

- ✓ **Range**: Pearson's R can take value between -1 and 1. R = 1 indicates a perfect positive linear relationship. R = -1 indicates a perfect negative linear relationship. R = 0 means no linear relationship between the variables
- ✓ **Strength**: The absolute value of R reflect the strength of relationship. A value closer to 1 indicates strong linear relationship
- ✓ **Direction**: Sign of R (+ or -) indicates the direction of relationship (positive or negative)
- ✓ **Assumptions**: Pearson's correlation coefficient assumes that the data follows a bivariate normal distribution and that the relationship between the variables is linear.
- ✓ **Calculation**: $$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

GENERAL SUBJECTIVE QUESTIONS

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

Answer:

What is scaling?

Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the features or variables of a dataset to a standard range or distribution.

There are two common methods for scaling data: **normalization** and **standardization**.

GENERAL SUBJECTIVE QUESTIONS

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

Answer:

Normalization

Normalization (also known as min-max scaling) is a scaling technique that transforms the data so that it falls within a specific range, typically [0, 1]. The formula for normalizing a variable x is:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- ✓ X*normalized* is the normalized value of *x*.
- ✓ min(*x*) is the minimum value of *x*
- ✓ min(*x*) is the minimum value of *x*

GENERAL SUBJECTIVE QUESTIONS

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

Answer:

Standardization

Standardization (also known as z-score scaling) transforms the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardizing a variable x is:

$$x_{\text{standardized}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

- ✓ $X_{standardized}$ is the standardized value of x
- ✓ mean(x) is the mean (average) of x
- ✓ std(x) is the standard deviation of x

GENERAL SUBJECTIVE QUESTIONS

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

Answer:

Why is scaling performed?

The primary purposes of scaling are to make the data more suitable for analysis and modeling, improve the performance of certain algorithms, and ensure that the features have similar scales. Scaling is particularly important when working with numerical variables that have different units or orders of magnitude.

GENERAL SUBJECTIVE QUESTIONS

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

Answer:

The key differences between normalized scaling and standardized scaling:

- ✓ **Range**: Normalization scales data to a specific range (typically [0, 1]), while standardization centers data around 0 with a standard deviation of 1.
- ✓ **Outliers**: Standardization is less sensitive to outliers because it relies on the mean and standard deviation, which are not affected by extreme values. Normalization can be influenced by outliers.
- ✓ **Interpretability**: Standardized data does not retain the original units of measurement, making it less interpretable in the context of the original data. Normalized data preserves the original range.

GENERAL SUBJECTIVE QUESTIONS

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

Answer: A Variance Inflation Factor (VIF) can becomes infinite when there is perfect multicollinearity among the independent variables in a multiple linear regression model. In details:

- ✓ **Perfect Multicollinearity:** In a multiple linear regression model, if two or more independent variables are perfectly correlated (meaning their relationship can be expressed as a linear combination of each other), it creates an issue
- ✓ **Mathematical Problem:** The formula for calculating the VIF involves matrix inversion. When perfect multicollinearity exists, the matrix that needs to be inverted is singular, which means it doesn't have a unique inverse. As a result, the calculation breaks down, leading to an undefined or infinite result.
- ✓ **When R-squared is 1:** When perfect multicollinearity exists, the R-squared value for the regression of *Xi* on the other independent variables becomes 1. Based on formula for VIF as bellow, the value of VIF becomes infinite

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

GENERAL SUBJECTIVE QUESTIONS

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

Answer: A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. Q-Q plots help you visually compare the quantiles of the observed data to the quantiles of the theoretical distribution. The primary use and importance of Q-Q plots in linear regression are as follows:

- ✓ **Assumption Validation:** Q-Q plots are a powerful tool for visually confirming or disproving the normality assumption
- ✓ **Model Diagnostics:** By analyzing Q-Q plots, you can identify issues with the model's assumptions and potentially take corrective actions, such as transforming the data or choosing different modeling techniques.
- ✓ **Outlier Detection:** Q-Q plots can also help in identifying outliers or data points with extreme deviations from the theoretical distribution
- ✓ **Visual Interpretation:** Q-Q plots provide a simple and intuitive way to assess the normality of residuals and other characteristics of the data distribution