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? (3 marks)**

I used Box plot to study their effect on the dependent variable ('cnt') from Independent catagorical variables.

The inference that I could derive were:

season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during 'weathersit1 This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' This indicates, workingday can be a good predictor for the dependent variable

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

Answer:

Using drop_first = True is crucial because it minimizes the generation of redundant columns during dummy variable creation, ultimately mitigating the correlation between these dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:

'temp' variable has the highest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer:
I have assessed the validity of the Linear Regression Model based on the following five assumptions:

Normality of Error Terms: This assumption ensures that the errors or residuals, which represent the differences between the observed and predicted values, follow a normal distribution. It's important because many statistical tests and confidence intervals assume normally distributed errors.

Multicollinearity Check: Multicollinearity occurs when two or more independent variables in the model are highly correlated with each other. High multicollinearity can make it challenging to interpret the individual impact of each variable and may lead to unstable coefficient estimates. Ensuring low multicollinearity is important for model stability.

Linear Relationship Validation: Linear regression assumes that the relationship between the independent variables and the dependent variable is linear. This assumption should be checked to confirm that the model is appropriate for the data. Non-linear relationships may require different modeling techniques.

Homoscedasticity: Homoscedasticity means that the variance of the residuals should be constant across all levels of the independent variables. If there is heteroscedasticity (varying variance), it can affect the reliability of standard errors and confidence intervals, leading to incorrect statistical inferences.

Independence of Residuals: This assumption requires that the residuals are independent of each other, meaning that the value of the residual for one observation should not depend on the value of the residual for another observation. Independence is crucial for the validity of statistical tests and predictions.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

mnth_9
mnth_8
weathersit_3

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a statistical model used to analyze the linear association between a dependent variable and a set of independent variables. The concept of a linear relationship implies that changes (either increases or decreases) in one or more independent variables correspond to corresponding changes in the dependent variable.

Mathematically, this relationship can be expressed using the following equation:

Y = mX + c

In this equation:

Y represents the dependent variable under prediction.
X stands for the independent variable used for making predictions.
m signifies the slope of the regression line, indicating the impact of X on Y.
c represents a constant known as the Y-intercept. When X = 0, Y is equal to c.

Linear regression is of the following two types –
Simple Linear Regression
Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

Multi-collinearity –
Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation –
Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

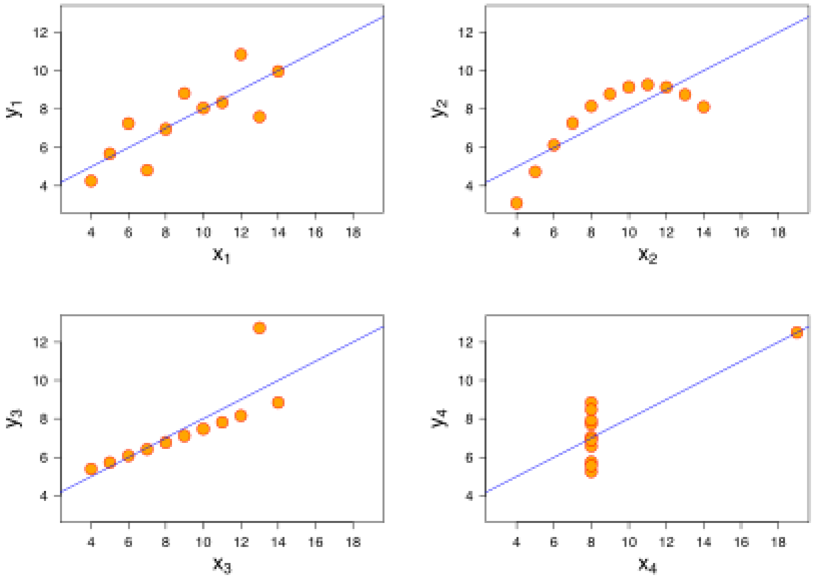
Relationship between variables –
Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms –
Error terms should be normally distributed

Homoscedasticity –
There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)

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. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

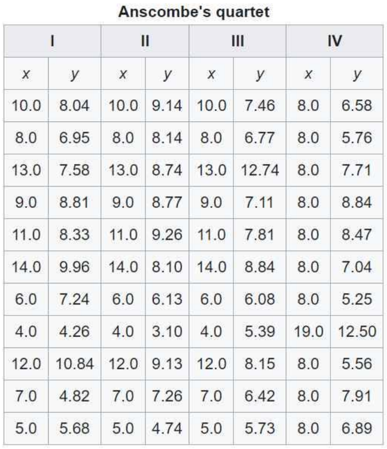


All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
2. For the second graph (top right), 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.
3. In the third graph (bottom left), the modelled relationship 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.
4. 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 is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets.

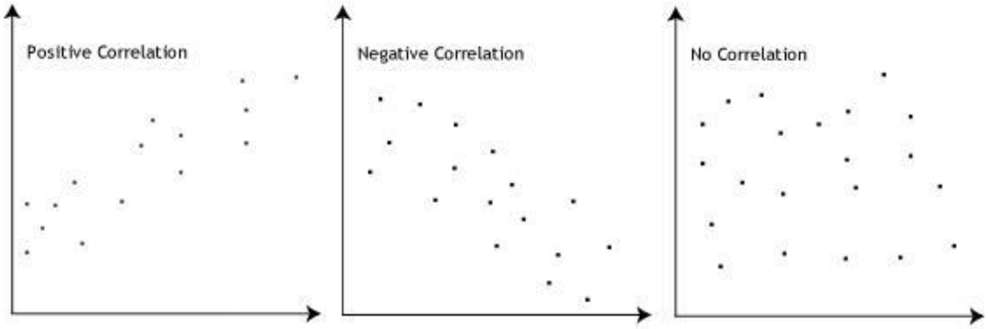


Anscombe's quartet

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

3. What is Pearson’s R? (3 marks)

The Pearson correlation coefficient, denoted as "r," can assume values within the range of +1 to -1. A value of 0 suggests that there is no discernible relationship between the two variables. A positive value indicates a positive association, meaning that as one variable's value rises, so does the other's. Conversely, a negative value indicates a negative association, signifying that as one variable's value increases, the other's decreases. This is visually depicted in the diagram below:



Pearson's r is a quantitative measure that reflects the strength of the linear relationship between two variables. When both variables tend to move in the same direction, the correlation coefficient is positive. Conversely, when one variable tends to increase as the other decreases, the correlation coefficient is negative

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling (3 marks)

Feature Scaling is a data preprocessing technique aimed at standardizing independent features within a consistent range. Its purpose is to address the challenge of dealing with variables that exhibit widely varying magnitudes, values, or units. When feature scaling is omitted, machine learning algorithms tend to assign greater weight to larger values and treat smaller values as though they are less significant, irrespective of the underlying units.

For example, in the absence of feature scaling, an algorithm might erroneously consider a value of 3000 meters as greater than 5 kilometers, which is clearly not accurate. In such a scenario, the algorithm is prone to making incorrect predictions. Therefore, Feature Scaling is applied to bring all values to a uniform magnitude, effectively mitigating this issue.

Feature scaling is of two types
Normalized Scaling(Min-Max scaling) and Standardized Scaling:

Here's a comparison between Normalized Scaling and Standardized Scaling:

Basis for Scaling:
Normalized Scaling: It uses the minimum and maximum values of features for scaling.
Standardized Scaling: It uses the mean and standard deviation of features for scaling.

Use Case:
Normalized Scaling: Applied when features have different scales and need to be brought to a common scale.
Standardized Scaling: Used when you want to ensure that the data has a mean of zero and a unit standard deviation.

Scaling Range:
Normalized Scaling: Scales values between [0, 1] or sometimes [-1, 1].
Standardized Scaling: It is not bounded to a specific range; it maintains the original distribution.

Outlier Sensitivity:
Normalized Scaling: Highly affected by outliers in the data.
Standardized Scaling: Much less affected by outliers.

Scikit-Learn Transformers:
Normalized Scaling: Scikit-Learn provides a transformer called MinMaxScaler for normalization.
Standardized Scaling: Scikit-Learn provides a transformer called StandardScaler for standardization.

5 **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

The phenomenon of obtaining an infinite value for the Variance Inflation Factor (VIF) typically occurs when there is perfect multicollinearity in the dataset. Perfect multicollinearity happens when one or more independent variables can be perfectly predicted from a combination of other independent variables in the regression model. This results in a mathematical issue when calculating the VIF. Here's why it happens:

Perfect Multicollinearity: When perfect multicollinearity exists, it means that one or more independent variables can be expressed as exact linear combinations of other variables. This leads to an R^2 value of 1 in the formula above because the model can perfectly explain the variance in those variables.

Resulting in Infinity: When R^2 is equal to 1, the denominator in the VIF formula becomes zero (1 - 1 = 0), which causes VIF to become undefined, or mathematically, it goes to infinity (1 / 0 = ∞).

6 **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3marks)**

A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is a plot of the quantiles (ordered values) of the dataset against the quantiles of the chosen theoretical distribution. Here's the use and importance of a Q-Q plot in the context of linear regression:

Use and Importance:
a. Normality Assumption Check: In linear regression, one of the fundamental assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. Deviations from normality can affect the validity of statistical inference, hypothesis testing, and the reliability of model predictions. A Q-Q plot is used to assess whether the residuals from a regression model approximate a normal distribution.

b. Visual Assessment: Q-Q plots provide a visual comparison between the observed data and a theoretical normal distribution. If the data points in the plot closely follow a straight line, it suggests that the residuals are approximately normally distributed. Deviations from a straight line indicate departures from normality.

c. Identifying Skewness and Outliers: Q-Q plots can reveal skewness in the data. If the plot exhibits curvature at the ends or bends away from a straight line, it may indicate skewness. Outliers in the data can also be detected as points that deviate significantly from the straight line.

d. Model Assessment and Improvement: Assessing normality of residuals is crucial for model assessment. If the Q-Q plot reveals deviations from normality, it may signal a need for model improvement or the need to explore alternative modeling techniques.

e. Data Transformation: If the Q-Q plot indicates non-normality, it may suggest the need for data transformation techniques, such as log transformation or Box-Cox transformation, to make the data more closely adhere to a normal distribution.

f. Statistical Inference: Departures from normality can impact the reliability of p-values and confidence intervals. By confirming that residuals are normally distributed through a Q-Q plot, you can have more confidence in the results of hypothesis tests and parameter estimates.