

: 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?

- Observations from above boxplots for categorical variables:
- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during Saturday
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, partly cloudy weather.
- count of bike rentals increased and became popular in year 2019 than 2018(from 'year' variable)
- count of bike rentals is more during clear weather (from 'weathersit' variable)
fall and summer are more favorable for bike rentals than spring (from 'season' variable)

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

A: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. to avoid multicollinearity (if we don't drop, dummy variables will be correlated) and affects the model adversely

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

A: By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

A: Residual errors follow normal distribution
Maintains linear relation between dependent variable (test and predicted)

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

A: The Top 3 features contributing significantly towards the demands of share bikes are:

- weathersit_Light_Snow (negative correlation).
- yr_2019(Positive correlation).
- temp (Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A: Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals Method**

- **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent 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. It fits a 'hyperplane' instead of a straight line.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet 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 before analyzing it and the effect of outliers on statistical properties.

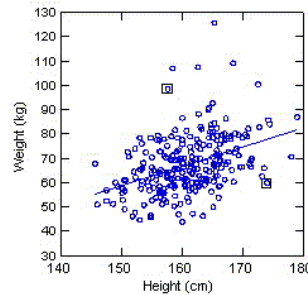
- **Simple understanding:**

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those sets of 11 data-points are given below.

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?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive, and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

- **Normalization/Min-Max Scaling:**

It brings all the data in the range of 0 and 1. sklearn. preprocessing. MinMaxScaler helps to implement normalization in python

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Standardization Scaling:**

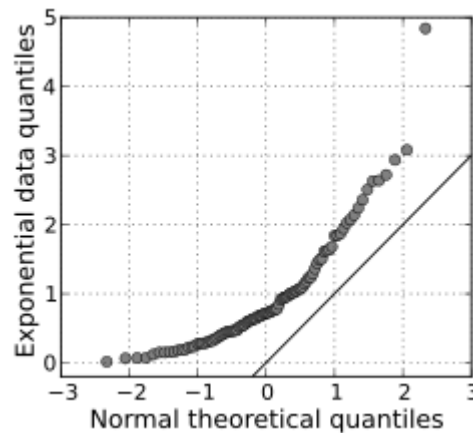
Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

$$\text{Standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

- If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation.
- We get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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



- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example,
- The median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.
- A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$.
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

Thank you