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? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Weather Situation has direct impact on Bike Rentals. No rentals recorded during extreme weather i.e 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog and less during 3: Light Snow

Season also impacts the bike rentals. Summer and Fall seasons have recorded higher rental count than other two seasons.

Yr also shows that bike rental demand is increasing year on year.

Higher number of bike rental during a workingday

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

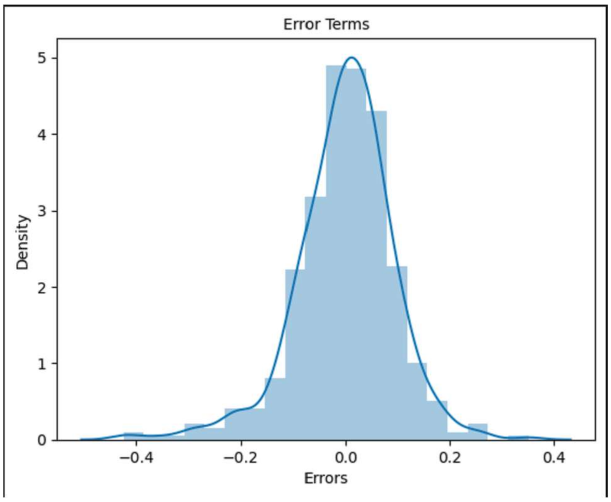
We can describe/represent a categorial variable with N-1 dummy variables where N is the number of unique values of a column. This won't lead to loss of data representation and dropping the first column while creating dummy variable helps avoids Dummy Variable Trap which is, a perfect multicollinearity between these predictors.

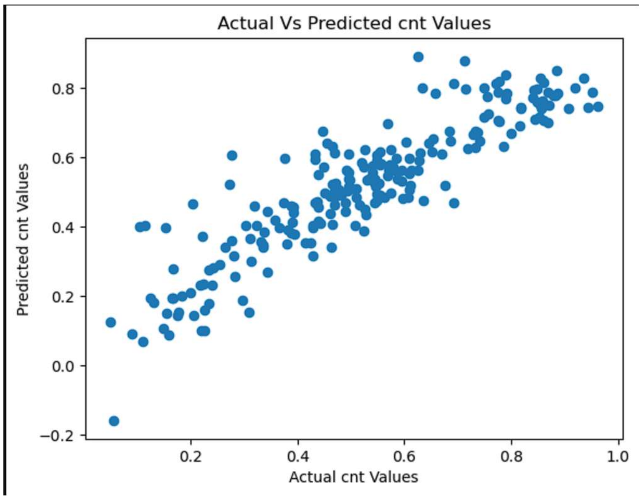
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temp/atemp has the highest correlation with target variable

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

By completing the residual analysis of error terms. Plot of the residuals, the difference between y and y_pred should be normally distributed, centered around mean of zero and the regression plot should show linear relationship.





Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Temp – As temperature rises, demand for renting bikes increases when all other variables remain constant.

Yr – rental bike demand is increasing year on year

Weathersit – Weather such as Light_Snow (3) decreases the demand for bike rentals. Extreme weather (4) has recorded no bike rentals at all. When weather situation is clear (1) or mist (2), demand for bike rentals as soared.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method to model the relationship between a target (dependent) variable and one or more predictors (independent variables).

There are two types of Linear Regression.

Simple Linear Regression – Models the relationship between a dependent variable and an independent variable.

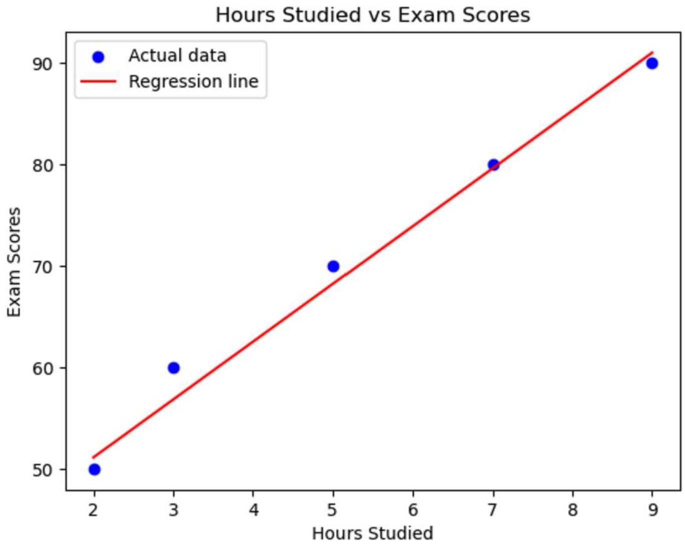
Multiple Linear Regression – Models the relationship between a target variable and two or more predictor variables.

It is mathematically represented as

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \varepsilon$

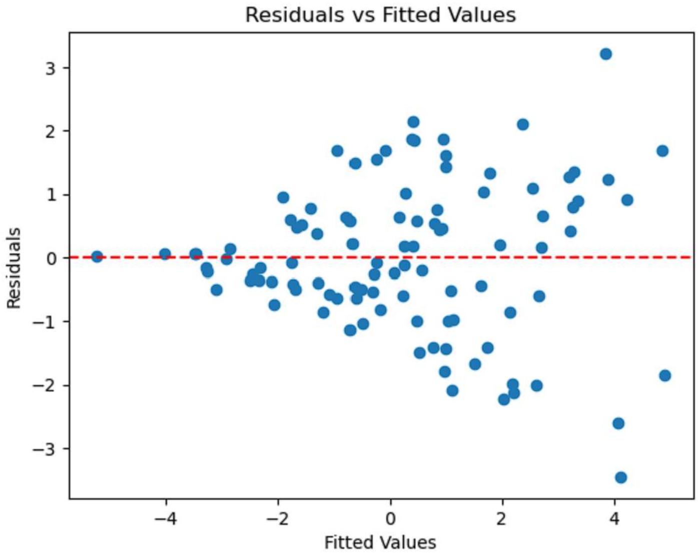
Y = dependent / target variable
$X_1, X_2 ..$ = independent / predictor variables
β_0 = Intercept
$\beta_1, \beta_2 .. \beta_n$ = coefficients (slopes)
ε = Error terms (residuals)

Objective of linear regression algorithm is to find the best fit line, in case of simple linear regression and hyperplane in case of multiple linear regression that minimizes the difference (error residuals) between actual and predicted value.

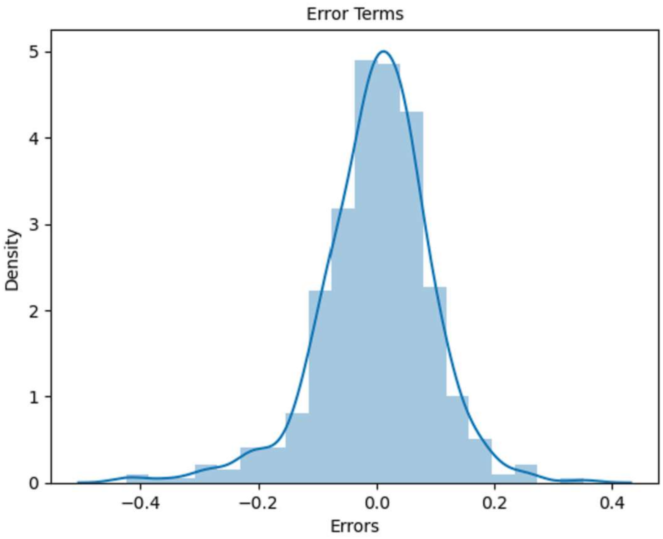


It has certain assumptions and they are

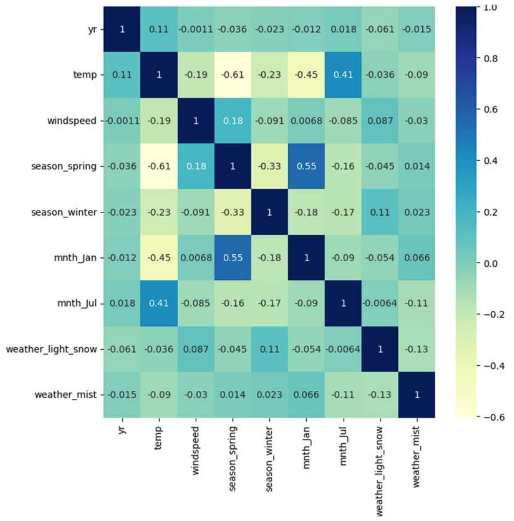
1) Assumes Linear relationship between dependent and independent variables
2) Observations are independent of each other
3) Homoscedasticity i.e. constant variance of residuals. If this is violated, residual plot will show a funnel shape for the residuals.



4) Residual analysis should show normal distribution of error terms



5) No multicollinearity, means independent variables should not be highly correlated with each other.



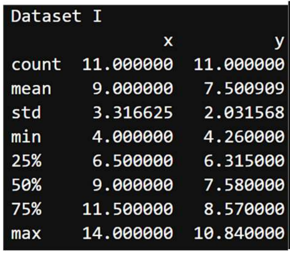
	yr	temp	windspeed	season_spring	season_winter	mnth_Jan	mnth_Jul	weather_light_snow	weather_mist
yr	1	0.11	-0.0011	-0.036	-0.023	-0.012	0.018	-0.061	-0.015
temp	0.11	1	-0.19	-0.61	-0.23	-0.45	0.41	-0.036	-0.09
windspeed	-0.0011	-0.19	1	0.18	-0.091	0.0068	-0.085	0.087	-0.03
season_spring	-0.036	-0.61	0.18	1	-0.33	0.55	-0.16	-0.045	0.014
season_winter	-0.023	-0.23	-0.091	-0.33	1	-0.18	-0.17	0.11	0.023
mnth_Jan	-0.012	-0.45	0.0068	0.55	-0.18	1	-0.09	-0.054	0.066
mnth_Jul	0.018	0.41	-0.085	-0.16	-0.17	-0.09	1	-0.0064	-0.11
weather_light_snow	-0.061	-0.036	0.087	-0.045	0.11	-0.054	-0.0064	1	-0.13
weather_mist	-0.015	-0.09	-0.03	0.014	0.023	0.066	-0.11	-0.13	1

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

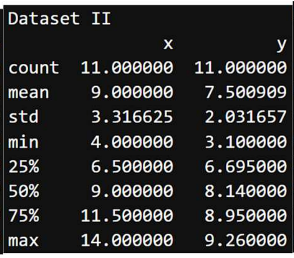
Anscombe's quartet is a set of 4 datasets with different data points that have nearly identical summary statistics. These datasets were created by statistician Francis Anscombe in 1973 to demonstrate that relying solely on numerical summary statistics alone is not a right approach. It is used as an example to underscore the importance of Exploratory Data Analysis (EDA)/Data Visualization rather than merely relying on numerical summary statistics.

Key Points:

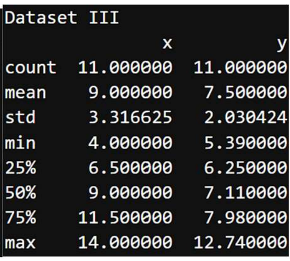
All 4 data sets have same mean, variance, R2(correlation coefficient) and linear relationship. Numerical summary statistics tell that they are nearly identical.



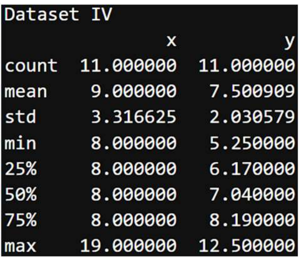
```
Dataset I
               x          y
count  11.000000  11.000000
mean    9.000000   7.500909
std     3.316625   2.031568
min     4.000000   4.260000
25%     6.500000   6.315000
50%     9.000000   7.580000
75%    11.500000   8.570000
max    14.000000  10.840000
```



```
Dataset II
               x          y
count  11.000000  11.000000
mean    9.000000   7.500909
std     3.316625   2.031657
min     4.000000   3.100000
25%     6.500000   6.695000
50%     9.000000   8.140000
75%    11.500000   8.950000
max    14.000000   9.260000
```



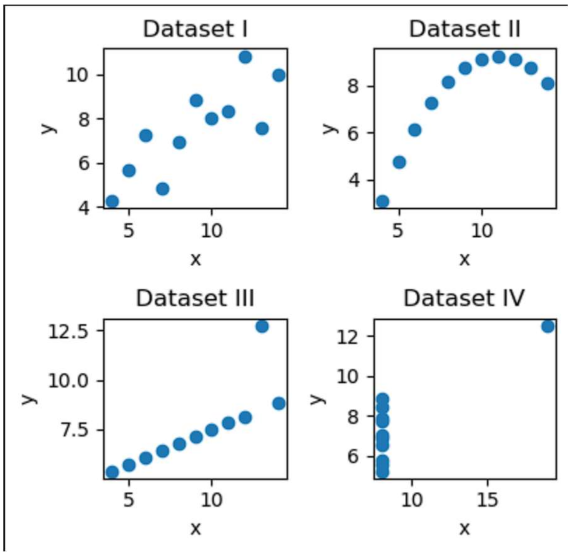
```
Dataset III
               x          y
count  11.000000  11.000000
mean    9.000000   7.500000
std     3.316625   2.030424
min     4.000000   5.390000
25%     6.500000   6.250000
50%     9.000000   7.110000
75%    11.500000   7.980000
max    14.000000  12.740000
```



```
Dataset IV
               x          y
count  11.000000  11.000000
mean    9.000000   7.500909
std     3.316625   2.030579
min     8.000000   5.250000
25%     8.000000   6.170000
50%     8.000000   7.040000
75%     8.000000   8.190000
max    19.000000  12.500000
```

However, visualization of data tells a different story about each data set i.e they are not at all similar.

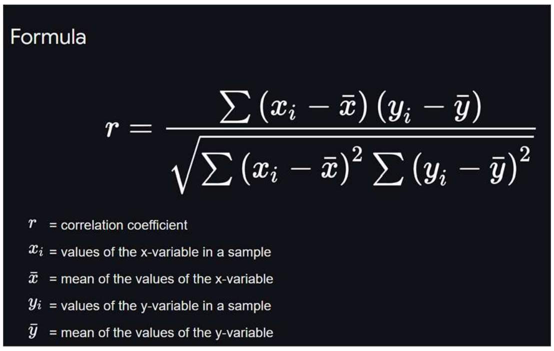
Visual representation of Anscombe's quartet using sns module's built-in Anscombe dataset.



Dataset1 – Perfect Linear Relationship
Dataset2 – Non-linear Relationship
Dataset3 – Liner Relationship with an outlier
Dataset4 – Vertical line with an outlier

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R or Pearson's correlation coefficient is a measure of linear correlation between two variables.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

It quantifies the strength and the direction of linear relationship (positive/negative) between variables.

R value ranges between -1 and 1 where 1 signifies perfect positive linear correlation and -1 signifies perfect negative linear correlation.

If R is 0 then there is no linear correlation at all.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Data is represented as quantitative (numerical) and qualitative (string values). Linear regression can only interpret the numerical data. Therefore, we convert qualitative data into numerical representation such as 0 and 1 for binary values and also use encoding using dummy variables to represent non binary categorical values, for example, 0001.

We now have numerical representation of all data. However, they are on different scales i.e numerical variables often have continuous and non-binary numbers. In order to represent all variable on same scale, we have to perform scaling i.e convert all variables to same scale using the methods such as MinMaxScaling or Standard Scaling.

The process of transforming data to represent them on same scale is called scaling and it is performed so that data analysis is not skewed by larger values.

Normalized Scaling – This is the technique used to adjusts the values such that they fall between 0 and 1 without distorting the associations between the data points. We can use MinMaxScaling to achieve this. Formula for MinMaxScaling is

X(scaled) = X – Xmin / Xmax – Xmin

Xmin – lowest value of X
Xmin – highest value of X

Standard Scaling – Standardized scaling also called z score scaling transforms data to have mean of 0 and standard deviation of 1. Formula for standardized scaling is

Xstd = X – μ/ SD(X)

Xstd – Standardised value of X
X – is the original value
SD(X) – Standard Deviation of X

Key Differences

Normalisation method scales features between 0 and 1 whereas standardization centers data around mean and scales as per standard deviation.

Nomalisation method may fail to deal with outliers whereas standardization is very good at dealing with outliers

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF (Variation Inflator Factor) is a measure of multicollinearity between different variables or predictors. Formula is

$$\mathrm{VIF}_{\mathrm{i}} = \frac{1}{1 - R_i^2}$$

If R2 is 1, then the VIF becomes infinity.

If R2 is 1 means that a particular variable can be totally explained by all other predictor variables. It’s a perfect multicollinearity situation i.e one variable can be predicted/explained by all other predictor variables and can be removed from our analysis or model building.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

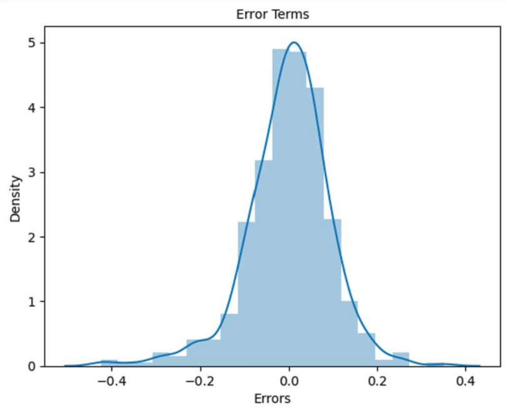
A Q-Q (Qunatile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theorical distribution such as normal distribution / uniform distribution.

For example, we could plot our actual data on Y axis and theoretical data for normal distribution on X-axis and see if there is perfect liner relationship between the two datasets. If there is then our data follows a normal distribution.

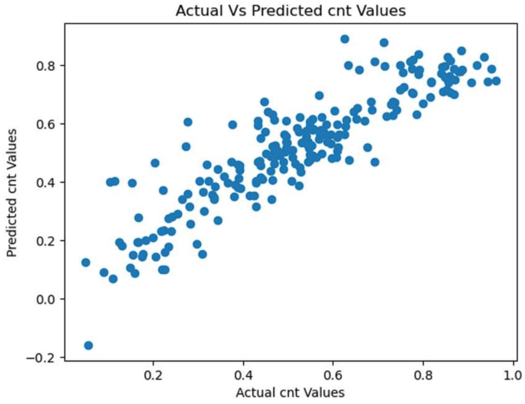
If there isn't a near perfect linear relationship then it may be following another distribution for example, uniform distribution.

Importance of a Q-Q plot in Linear Regression:

Q-Q plots are used to assess the normality. For example, we plot the error terms of the built lr model to check if its normally distributed. Residuals being normally distributed is one of the key assumptions of Linear Regression.



Used to detect deviation. For example, we plot actual y vs predicted y and expectation is that we see linear relationship with no patterns/deviations.



We use Q-Q plots to validate the Linear Regression model we have built on the basis that error residuals follow a normal distribution.