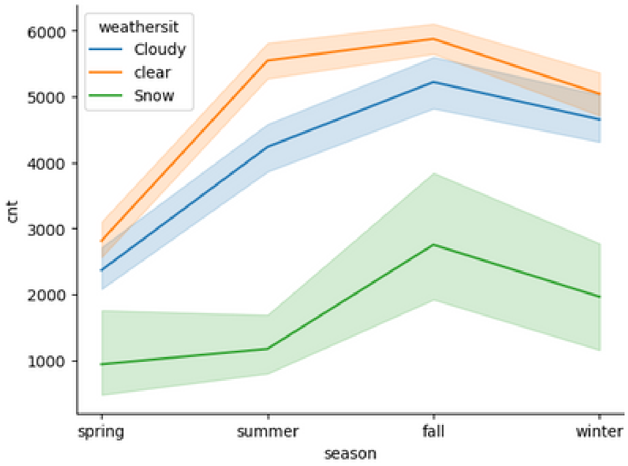
Assignment-based Subjective Questions

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

When analyzing weather situation with Season with respect to Target variable we can see it's tabular data as shown below....

season	fall	spring	summer	winter	**Total**
weathersit					
Cloudy	23.623801	32.744285	31.362884	35.944074	30.236125
Snow	1.037291	0.796355	0.381781	2.330762	1.150738
clear	75.338908	66.459360	68.255335	61.725163	**68.613137**
Total	100.000000	100.000000	100.000000	100.000000	100.000000

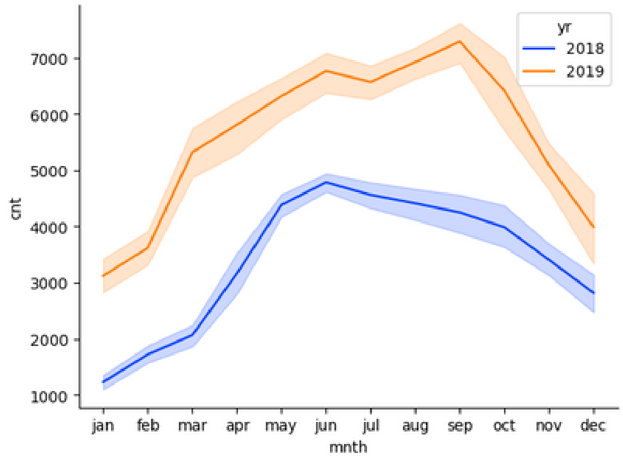
It is quite visible that ~**69%** of bike riders that is maximum among all weather situation prefer clear weather irrespective of season. The lowest count of bike riders are found in snow weather situation just merely **1.15%**



Same can be seen in this line plot ,weather situation cloudy is in between Clear and snow and talking in terms of season ,spring season is showing lowest number of riders in general...and more Bike riders in summer and fall.

Winter season showing lesser number of bike riders than summer and fall but more than spring

When analyzing month and year with respect to Target variable we can see



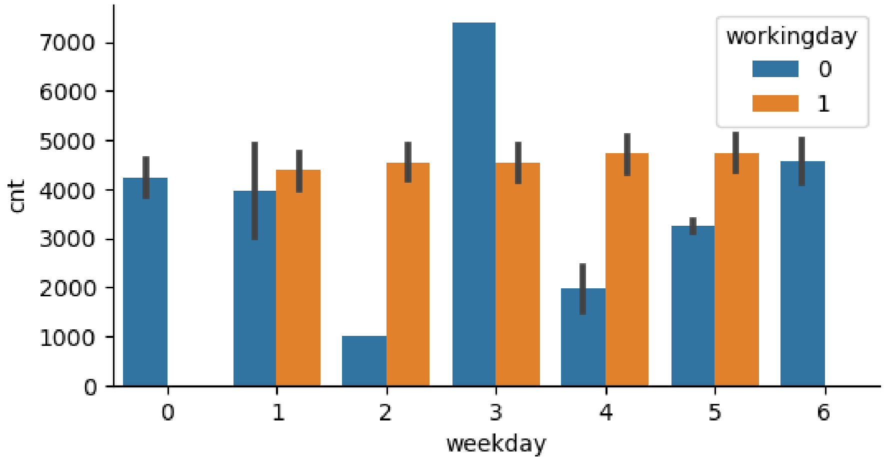
Year	total
2018	37.774584
2019	**62.225416**
total	100.000000

as the data available total riders in 2019 is **~63 %** and month wise variation can be seen on graph.

mnth	Total
Working Day	
0	30.395506
1	**69.604494**
Total	100.000000

As the tabular data is saying the working day ‘1’ shows neither weekend or holiday it’s the working day ..which is having **~70%** of the total riders.

As the tabular data is saying the working day ‘0’ shows .it’s the holiday ..having **~30%** of the total riders.



This graph shown below shows the variation of bike riders across weekdays ,with working day i.e that is particular weekday is weekend /holiday (‘0’) or it is working day (‘1’)

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

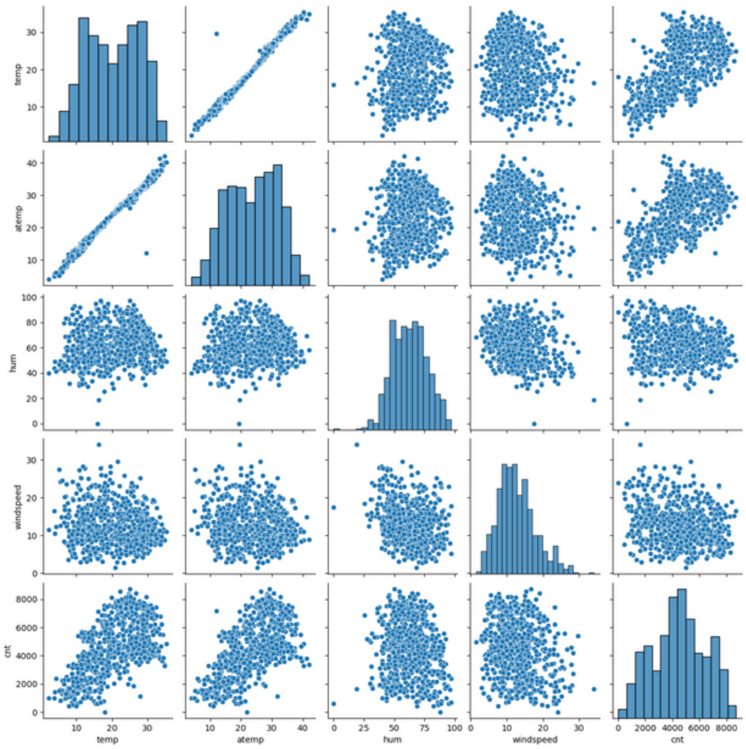
The reason for dropping one dummy variable is to avoid multicollinearity, which is a situation in which two or more independent variables in a regression model are highly correlated. In the context of dummy variables, if you include dummy variables for all categories of a categorical variable, they can be perfectly predicted from each other. This perfect correlation among dummy variables can cause issues in regression models.

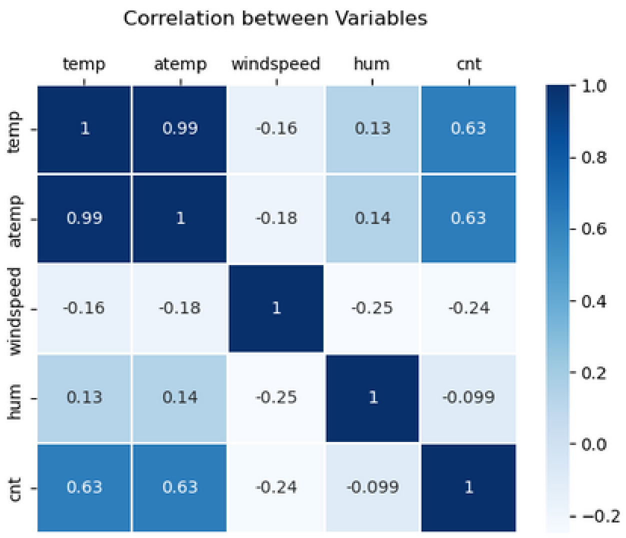
For example ,let’s say that we are creating dummy variable for Gender column in a dataset , having two categories male and female without any ordinal order, now after creating dummy variables it create two column Gender_male & Gender_female ,it is so obvious that where the Gender_male is **‘0’** there Gender_female automatically becomes **‘1’** .Thats create strong multicollinearity in the dataset which affect the regression.

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

From pair plot and heatmap ,temp and atemp variable is showing very strong correlation with target variable ,although temp are atemp are similar in context.

Heatmap and pairplot is shared for reference ,having strong correlation with target variable also by statsmodel ,it’s coefficient is **+0.402482** showing postive correlation with target variable



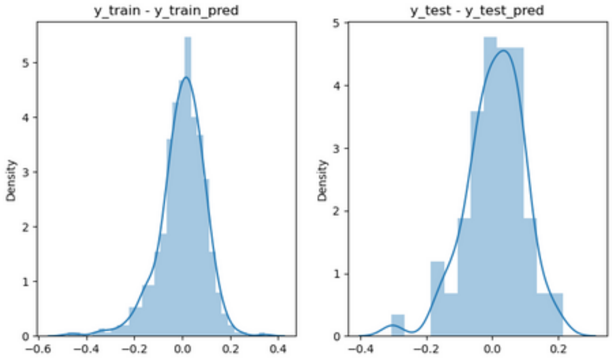


How did you validate the assumptions of Linear Regression after building the model on the training set?

Validation of Linear regression after model building is:-

1) Linearity :- The relationship between the independent variable(s) and the dependent variable is linear.

2**) Normality of residuals** :- The residuals should be normally distributed..

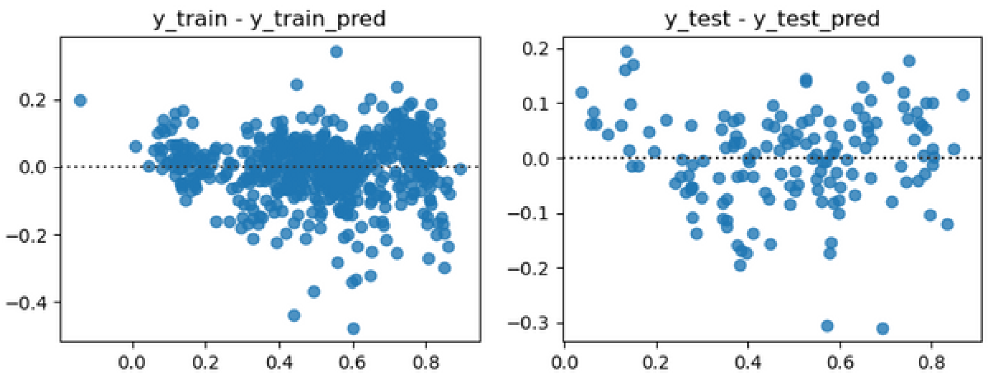


3)No perfect multicollinearity :- Independent variables should not be perfectly correlated with each other.

	Features	VIF
0	temp	4.05
10	weathersit_clear	2.80
2	season_winter	2.35
11	yr_2019	2.06
6	mnth_nov	1.74
1	season_spring	1.50
4	mnth_jul	1.39
3	mnth_dec	1.33
7	mnth_sep	1.22
5	mnth_mar	1.16
8	weekday_6	1.16
9	weathersit_Snow	1.10

visible from tabular data there is no observe multicollinearity between variables as the VIF score for each of them are below threshold of 5.

4) **Homoscedasticity (Constant Variance):-** The variance of the residuals should be constant across all levels of the independent variable(s).



5) **No autocorrelation of residuals** :- The residuals should not show a systematic pattern over time

Durbin-Watson: 2.077 , A value close to 2 suggests no autocorrelation

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

weathersit_Snow	-0.224837
season_spring	-0.143225
mnth_nov	-0.079456
mnth_dec	-0.057408
mnth_jul	-0.048639
weekday_6	0.023337
mnth_sep	0.048386
mnth_mar	0.059198
weathersit_clear	0.075835
season_winter	0.089199
const	0.141463
yr_2019	0.231771
temp	0.426293

As we can see from tabular data shared:-

1) Temp : The coefficient of temperature is **+0.426** that means as the temp rises number of bike rider increases.

2) Weathersit_Snow : The coefficient of weather situation is **-0.2248** that means as the snow weather situation occurs number of bike riders descreases.

3)season_spring :- The coefficient is negative for spring season as it is also visible from graph that bikers do not prefer spring season.

General Subjective Questions

Explain the linear regression algorithm in detail

In the most simple words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.
The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables

Linear Regression is of two types:
1) **Simple Linear Regression** :- Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
2) **Multiple Linear regression** :- Multiple Linear Regression is where there are more than one independent variables for the model to find the relationship.

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$\text{y}= b_o + b_1 x$$

equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4...,bn are coefficients or slopes of the independent variables x1,x2,x3,x4...,xn and y is the dependent variable

$$\text{y}= b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \;\ldots. + b_n x_n$$

Our Model tries to predict values based on training data which is extracted from dataset for training and try to predict value on similar set of data called test data extracted from original dataset

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Here Training refers to finding the coefficients bo,b1 mentioned above such that the error of estimation should me minimum,and we use **Odinary Least Square** method for coefficients estimation .

Also to Minimize the error we use **Gradient Descent** method to arrive at best fit line of model.

There are some assumptions of Linear Regression:-
1)Linearity: The relationship between the independent variables and the dependent variable should be linear.

2) Independence of Residuals: The residuals (the differences between observed and predicted values) should be independent. There should be no patterns or trends in the residuals.

3) Non-Homoscedasticity The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X

4) Independence/No Multicollinearity: The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated

5) The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.

The Accuracy od LInear regression models is done by various method:-

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):

The mean squared error is a commonly used metric that measures the average squared difference between predicted and actual values. The root mean squared error is the square root of the mean squared error and is in the same units as the dependent variable.

R-squared

R-squared represents the proportion of the variance in the dependent variable that is explained by the independent variables

Adjusted R-squared:

Adjusted R-squared takes into account the number of predictors in the model and adjusts the R-squared value accordingly. It penalizes models with excessive predictors.

Explain the Anscombe's quartet in detail.

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".

This is the 4 dataset:-

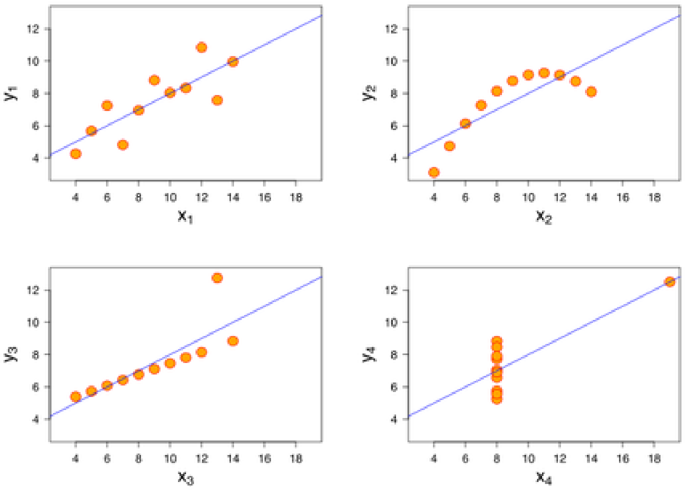
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

The basic thing to analyze about these datasets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

Property	**Value**	**Accuracy**
Mean of x	9	exact
Sample variance of x: s2x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s2y	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

But if we plot each dataset we get this four graphs as follows

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — while a relationship between the two variables is obvious, it is not linear,
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.



What is Pearson's R?

Pearson's correlation coefficient, often denoted by r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics and data analysis.

It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The value of r ranges from -1 to 1

r=1 indicates a perfect positive linear relationship (as X increases, Y increases proportionally).

r=−1 indicates a perfect negative linear relationship (as X increases, Y decreases proportionally).

r=0 indicates no linear correlation between the variables.

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

What

Scaling in the context of data preprocessing refers to the process of transforming the numerical values of variables to a standardized range or distribution. The primary goal of scaling is to ensure that all variables contribute equally to the analysis and modeling processes

Why

The machine learning models provide weights to the input variables according to their data points and inferences for output. In that case, if the difference between the data points is so high, the model will need to provide the larger weight to the points and in final results, the model with a large weight value is often unstable. This means the model can produce poor results or can perform poorly during learning..

Difference between normalized scaling and standardized scaling?

Normalization/Min-Max Scaling

Normalization or Min-Max Scaling is the simplest method and consists in rescaling the range of features to scale the range in [0, 1]

The general formula for a min-max is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized Scaling (Z-score Normalization):

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

The general formula for a Standardize scaling is given as:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Differences between Normalized Scaling and Standardized Scaling:

Normalisation is suitable to use when the data does not follow Gaussian Distribution principles. On the other hand, standardisation is beneficial in cases where the dataset follows the Gaussian distribution.

Normalization is affected by Outliers , whereas Standardisation is not affected by the outliers in the dataset as it does not have any bounding range.

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

The main reason for infinite VIF values is perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model can be exactly predicted by a linear combination of the other variables. In other words, there is a perfect linear relationship among the independent variables.

A high VIF suggests that the corresponding variable is highly correlated with other variables in the model

A VIF value is considered problematic when it becomes extremely large, sometimes even infinite. This happens when the tolerance (the reciprocal of VIF) becomes very close to or equal to zero. The tolerance is calculated as:

$$VIF\ x_i = \frac{1}{Tolerance} = \frac{1}{1 - R_i^2}$$

the coefficient of determination (R-square) that is derived by regressing a variable against every other variable. An infinite VIF results from the tolerance getting close to or equal to 0 if the R-square is closer to 1 (high collinearity).

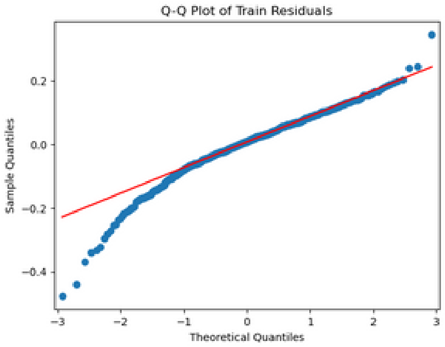
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line.

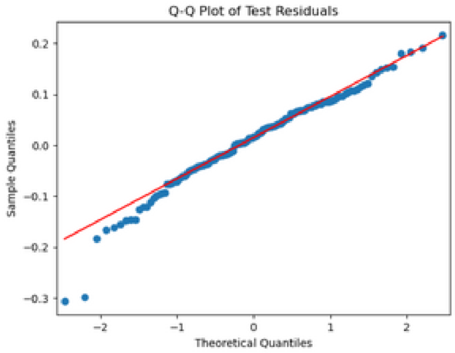
A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, we can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. we can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model.

In our bike sharing Linear regression model the Q-Q plot is:-

y_train - y_train_pred



y_test - y_test_pred



If the points in the Q-Q plot approximately fall along a straight line, it suggests that the residuals are close to being normally distributed.