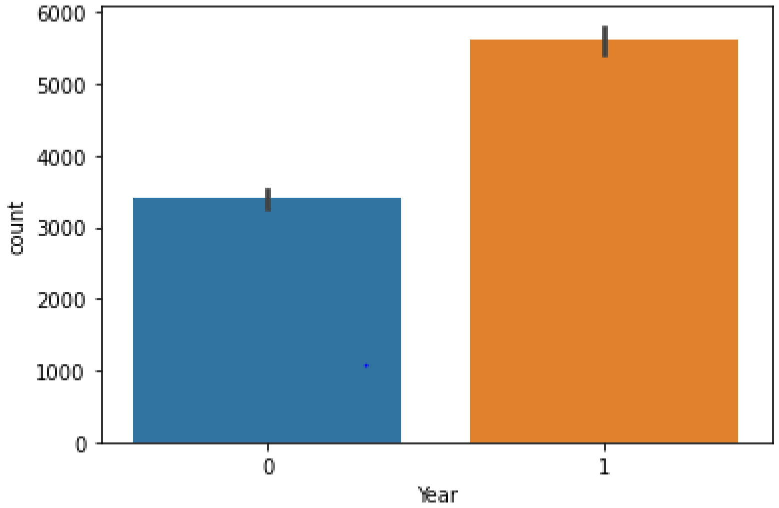
Assignment-based Subjective Questions

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

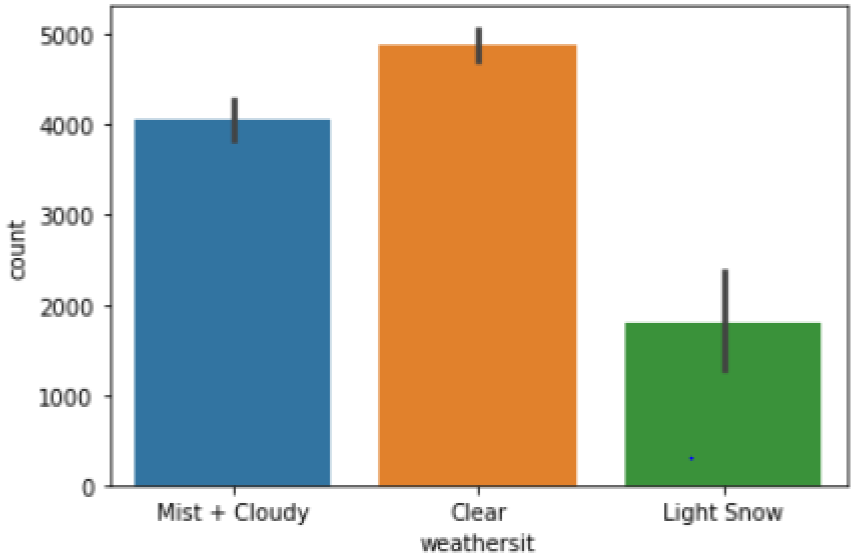
Ans.

Followings are the categorical variable in given dataset:

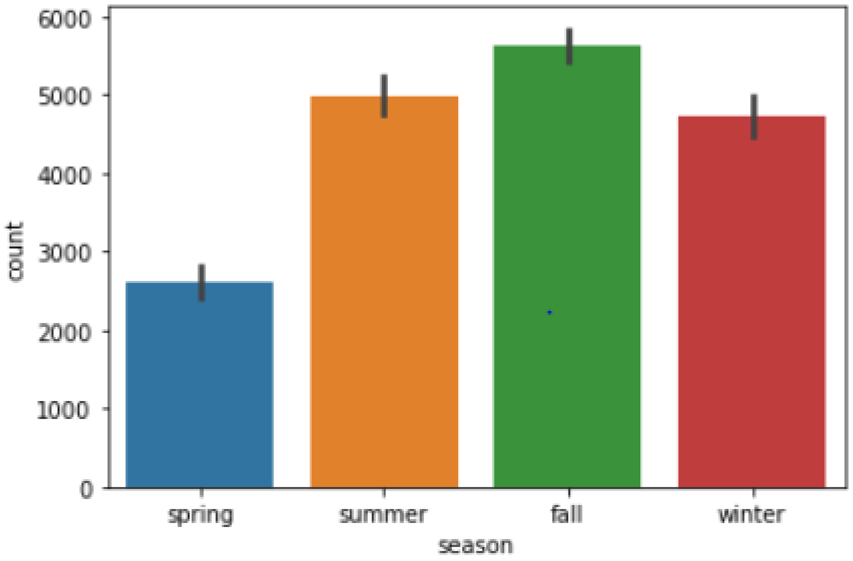
1.Year: count of bike rentals in year 2019 is more than in year 2018.

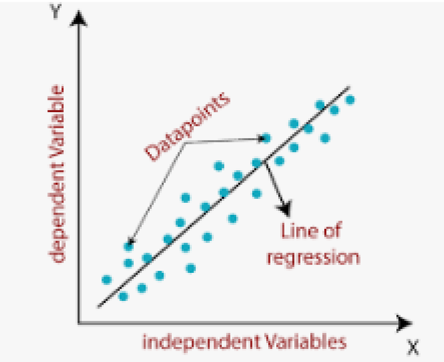
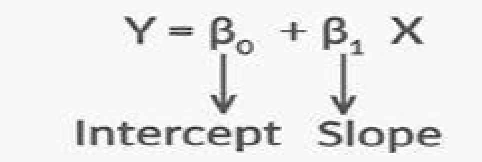


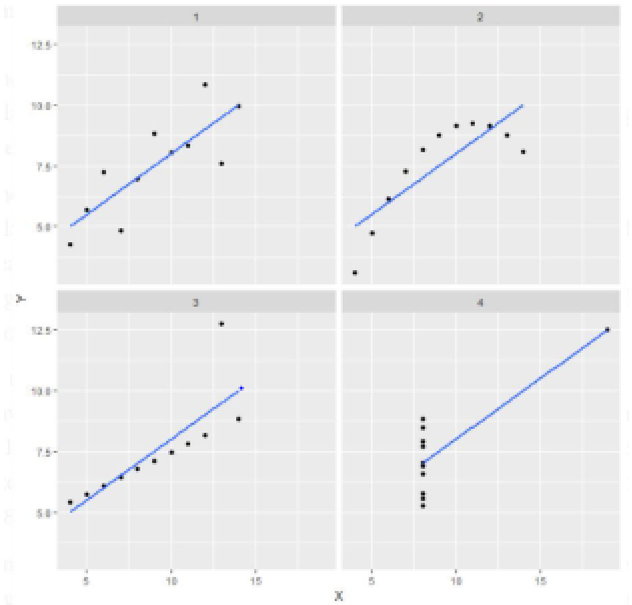
2. Weathersit: maximum no of people takes bike in partly cloudy weather.

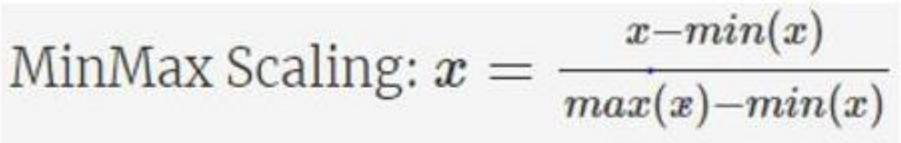
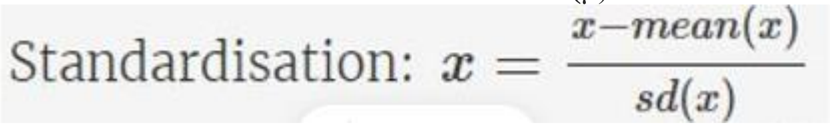


3.Season: no of people took bike on rent in fall season compared to summer.

	
Q-2	Why is it important to use drop_first =True during dummy variable creation?
Ans.	When we will create dummy variables, it also increase no of columns, drop_first eliminate this extra columns hence it will reduce correlations. For example, we consider categorical variable season in which there are 4 category {1:winter,2:summer,3:spring,4:fall} 1 to 3 is false it is clear that it is fall season which is unnecessary.
Q-3	Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans.	Temperature
Q-4	How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans.	**.** Plot pairplot to check linear relationship between dependent variable and predictor. If there is nonlinear relationship between variables use exponential, log plot 2. there is no correlation between the consecutive error terms of the time series data ,If conditions are not met than remove outliers(over differenced variables),add lags in variables. 3.No multicollinearity, Check VIF terms and remove terms that are not necessary Create Scatter plot of variables to check correlation 4. Homoscedasticity Error terms should have constant variance over scale.Check homoscedsticity plot scatter plot of error terms vs fitted values. 5.Using statistical test ,check error terms re normally distributed or not
Q-5 Ans.	Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? Temperature, spring season, Mist +cloudy and Lightsnow weather

	General Subjective Questions
Q-1	Explain linear regression model in detail.
Ans.	There are many different kinds of machine learning algorithms to discover specific patterns in huge amount of data that lead to actionable insights. There are two types of learning Supervised learning and unsupervised learning Regression is a technique used to define the relationship between independent variables and a dependent variable . It's used to predict future outcomes. there are two types of regression analysis techniques in machine learning as follow. Linear regression and logistic regression Linear regression fall under Supervised learning. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.  Y dependent Variable Datapoints Line of regression independent Variables X Equation for linear regression: $Y = \underset{\substack{\downarrow \\ \text{Intercept}}}{\beta_0} + \underset{\substack{\downarrow \\ \text{Slope}}}{\beta_1} X$ This equation is for simple line,if we increase no of varibales it will convet into hyperplane. The strength of the linear regression model can be characterizing by two terms: 1.r square Value of R square lie between 0 to 1. the higher the R-squared, the better the model fits your data. $R^2 = 1- (RSS/TSS)$
Q-2	Explain the Anscombe's quartet in detail.
Ans.	Anscombe's quartet consists of four datasets that have nearly identical simple

	statistical properties, but when we plot it ,it appears very different .  • In the first one(top left) we can say from scatter plot that there seems to be a linear relationship between x and y. • In the second one(top right) we can conclude that there is a non-linear relationship between x and y. • In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one(outlier) • Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.
Q-3	What is Pearson's R**?**
Ans.	Pearson correlation coefficient is a the measurement of the strength of the relationship between two variables and their correlation with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes. For example: • Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases. • Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa.
Q-4	What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.	We perform a scaling in prepressing step. It is applied to independent variables(predictor variables) to normalize the data within a particular range**.** In real time scenario, collected data set have magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients Normalization/Min-Max Scaling: It will convert all data I range between 0 to 1.  MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$ Standardization Scaling: 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).  Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
Q-5	You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans.	VIF= 1/(1-R2) ,VIF is infinite means r2 is 1.means perfectly linear relationship between independent and dependent varibles.or we can say that a perfect correlation between two independent variables.
Q-6	What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans.	When we are talking about Q-Q plot, we concentrate on Y-X line which is also known

	as 45 degree line.it shows that data comes from same distribution. If training and test datasets comes from different source,Q-Q plot help us to confirm both dataset belongs to population having same distribution.