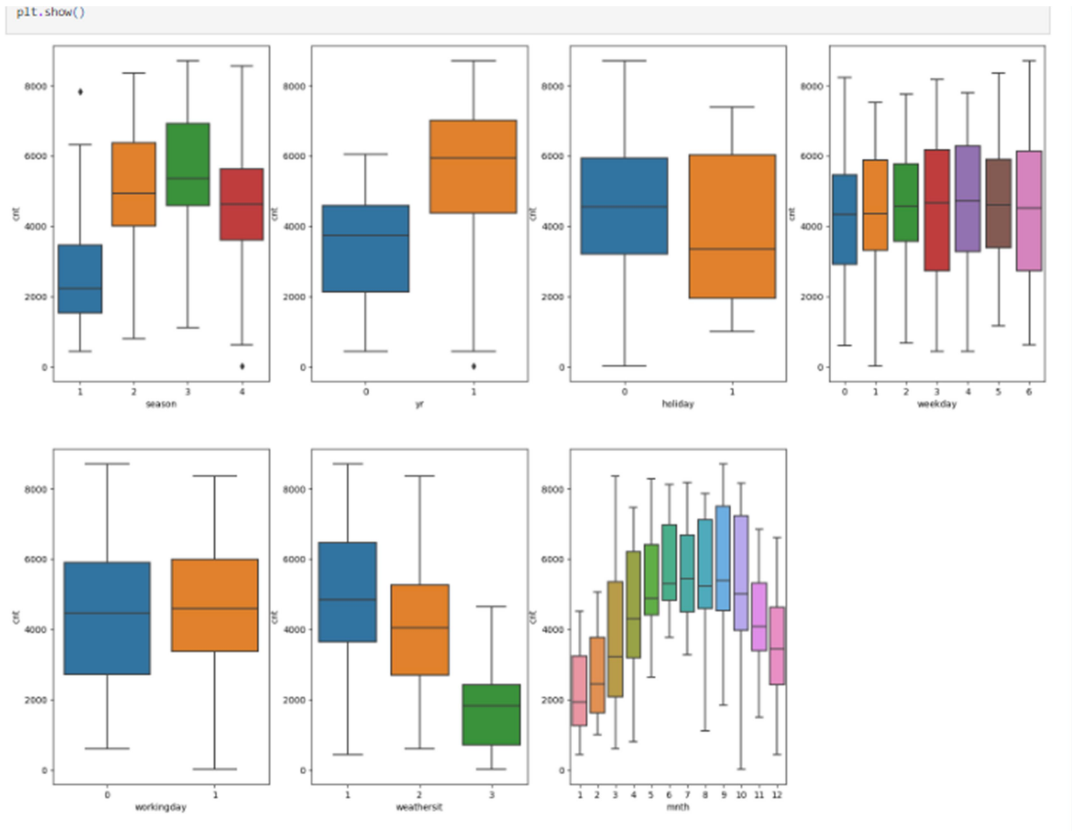
Q.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans- Here following are categorical variable –

- Season
- Yr (year)
- Holiday
- Weekday
- Workingday
- Weathersit
- Mnth

These variables have major effect on cnt variable which able to visible in below mentioned box plot.



And from upper box plots we have below mention findings-

1. We able to seen that in season3 have highest booking and in season1 have lowest booking so we consider this variable as a predictor variable in module.

2. We able to seen that in month 4,5,6,7,8,9 and 10 have more bookings respect to other months so it's also consider as a predictor variable.

3. And in holidays bike booking is less respect to non-holidays also consider as a predictor.
4. And in weekdays the bike booking is almost constant.
5. And in workingdays have more bike bookings respect to non-working days.
6. And in clear weather have more bookings respect to others.
7. we able to see that in year 2018 have lowest booking respect to 2019 so its increasing year to year

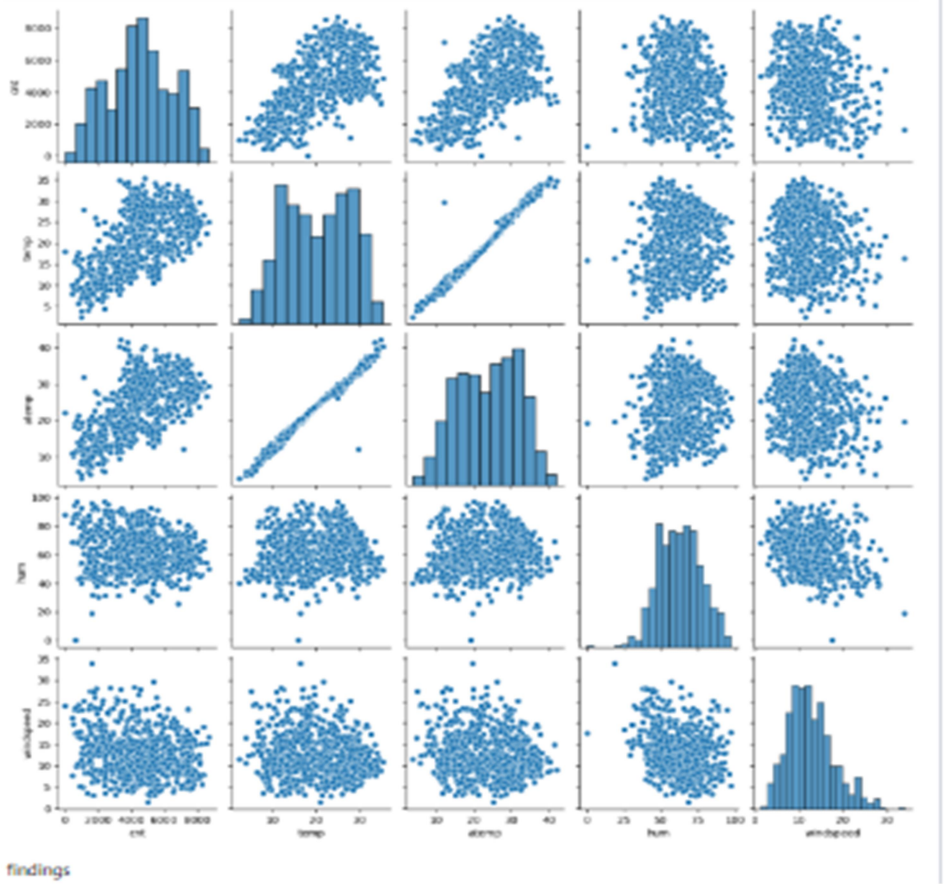
Q.2 Why is it important to use drop_first=True during dummy variable creation?
Ans- drop_first=True, It is used for dropping extra created column during dummy variable creation. And reduce the correlation between created dummy variables.

Example- If we have three category of a categorical variable then if we took 2 dummy variable its covered all three variables condition like we have three variable a, b and c so we have condition like not a , not b then it's obviously have c. so we no need to create extra column for variable c.

Q.3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans- According to pair plot mentioned below we able to seen that temp and atemp variable have highest correlation with cnt variable.

1. visualising numerical data using pairplot

```
In [20]: sns.pairplot(data = bike_df_new, vars=["cnt", "temp", "atemp", "hum","windspeed"])
         plt.show()
```



findings

Q.4 How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans- I am validate the assumption of linear regression model based on below 5 assumptions-

- Normality of error – error terms should follow normal distribution function.
- No auto-correlation
- Homoscedasticity
- Multicollinearity
- Linearity should be visible

Q.5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans- 1) temp. 2) yr (year) 3) Winter (season)

Q.1 Explain the linear regression algorithm in detail.

Ans- Linear regression is a predictive model which is defined relationship between independent variable and dependent variable. Relationship is linear its mean How the independent variable value changes(increase or decrease) according to this dependent variable value also changes(increase or decrease).

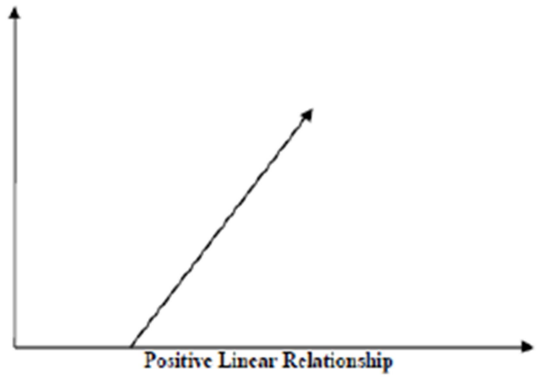
Mathematical formula-

$$Y = mX + C$$

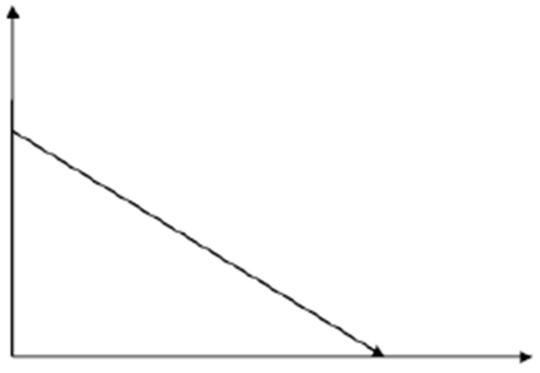
Where-

Y is the dependent variable
X is the independent variable
m is the slope of the regression line
C is a constant,

And that regression relationship can be having both type positive (if increasing dependent variable with increment in dependent variable) and negative (dependent variable decreasing with decrement in independent variable).



Positive Linear Relationship



Negative Linear Relationship

Types of liner relationship model-

1) Simple liner regression - Have only one independent variable (only one input).
2) Multiple linear regression- have more than one independent variable (have more than one input).

The goal of the linear regression is finds best fit line between dependent variable and independent variable with minimum error. For that we used RFE and VIF etc.

Q.2 Explain the Anscombe's quartet in detail.
Ans- It was developed in 1973 by statistician Francis Anscombe to describe the importance of plotting graphs and EDA before building linear regression model and showing drawbacks of model which is depending only on statically summary.

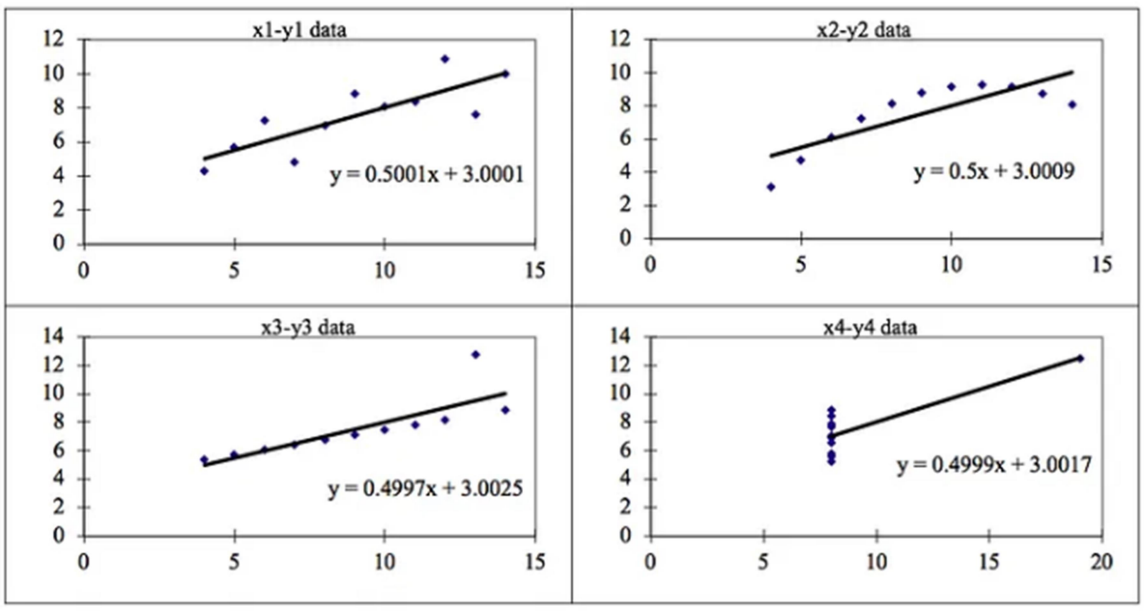
It is can be described by four data sets which have identical descriptive statically properties in terms of mean, variance, R-squared, correlation and linear regression line but have different representation when we plotting that datasets on scatter plot.

These four plots are plotting using below data-

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Image by Author

- Here mean and SD are same for all x and y and R-Squared value same for all four data sets. But when we plotting a scatter plot then these are representing different plot.



- Here first correctly fit on line.
- Second one didn't showing linear relationship between x and y. so its mean it is not fit for linear regression model.
- In 3rd one have some outliers so it's not handle by linear regression model.
- In 4th one also has outliers which one notable handle by linear regression model.

Conclusion is that any regression algorithm can be fooled so before building any model it's important to visualise the data and done EDA.

Q.3 What is Pearson's R?

Ans- It is a type of statistical parameter which is represents association or correlation between two variables. Its value lies between -1 to 1. Its value has – ve its mean when we increasing one variable then second variable goes down. And for positive value when we increasing one second also increasing. And if its vale has 0 its mean didn't have any dependency both variables on each other.

Mathematical representation-

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Ans- It is pre-processing step to fit the data in a particular range and speedup the mathematical calculation. If we didn't do scaling then here having possibility of error in model because model given weigh to higher magnitude values and neglecting lower magnitude values so its possibility that we have strong prediction variable values going neglect due to its lower magnitude values.
Difference between Normalizing Scaling and Standardize Scaling: -

- In normalized scaling us using minimum and maximum values for scaling and for Standardized scaling we used mean and standard deviation.
- Normalized scaling used when features are different scales and Standardized scaling used when we have 0 mean and 1 standard deviation.
- In normalized scaling values have in ranges -1 to 1or 0 to1. But in Standardized scaling didn't have any specific range.
- Normalized scaling affected by outliers but in Standardized scaling not have any affect by outliers.
- Normalized scaling called scaling normalized or min-max normalization whereas Standardized normalization called Z score Normalization.

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

Ans VIF basically given relationship of one independent variable with all other independent variables.
Mathematically representation-

$$VIF = 1/(1-R2)$$

If VIF value have infinite its mean variable have perfect correlation with other variables. For this problem we need to drop a variable which is causing perfect multi co-linearity as a solution. For that we need all VIF value less than 5.

Q.6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
Ans-
Q-Q plot is a probability plot, It is a plot in which one data set quantile plot draw against to the second data set quantile plot or is a pair plot of two data set quantiles against one another.

Uses-

It is used for two data sets –

- Both data sets are come from population have same distribution or not.
- Both data seta have similar type distribution shape, and have same scale etc.

Importance- In linear regression model we have two data sets test and train data sets from population for that we create Q_Q plot for checking that both are follow same distribution or not.