Assignment-based Subjective Questions

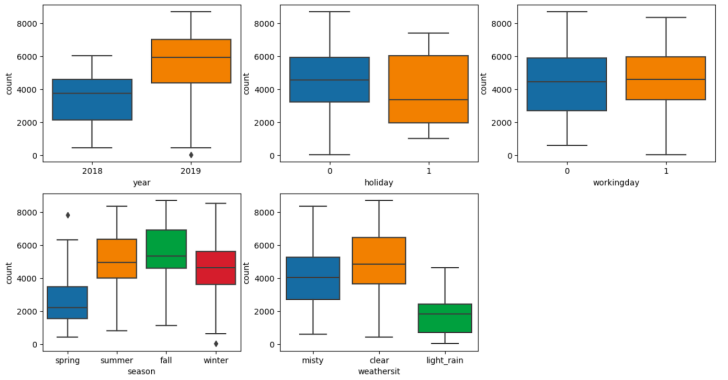
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

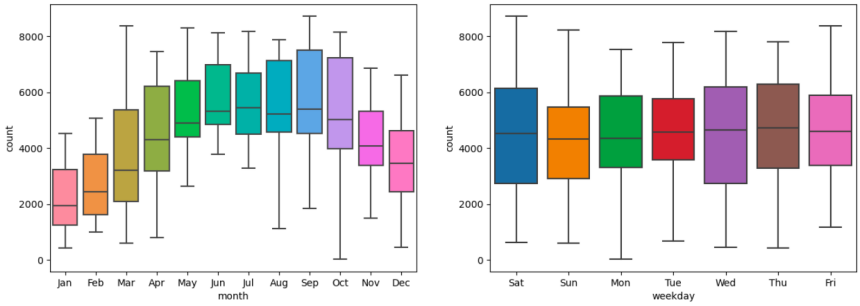
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

There are few categorical variables like, year, holiday, working day, season, weather, month, week day. Below are the effects of each on dependent variable count.

1. yr – Number of rentals in 2019 were more than 2018. Even if combined with other categories like weathersit, season, holiday, 2019 was having higher bike rentals than 2018.
2. weathersit – Clear weather was having highest bike rentals. Also, there are no samples with highest rain which means that its rarely occurring event
3. season – Fall season have highest number of bike rentals
4. month – Sep month was able to attract more booking than any other month.
5. holiday – Bike rentals shows negative effect in case of holidays, possibly people want to spend time with family or there are other modes of travel more preferred.
6. workingday – There is no clear sign that working day have any effect on number of bookings
7. weekday – There is no clear sign or any pattern that bookings are dependent on weekday





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)

If number of values are p, then number of dummy variables required is p-1, as last category value could be explained by combination of rest of the categorical values.

In case of multiple categorical values and features, it becomes difficult to identify correlation, build model and select features with unwanted variables. Hence, drop_first=True helps reducing this extra work, helps keeping model simple and also remove unwanted variables.

Below table with 3 categories can help with reason to drop column.

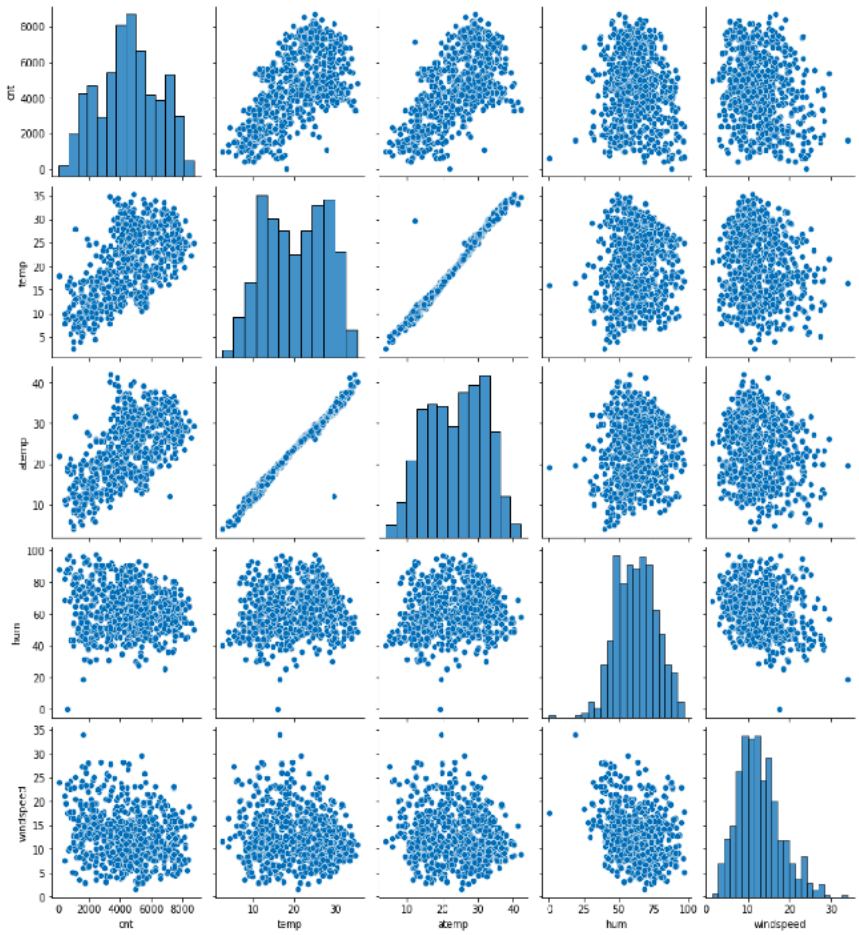
	Without dropping Catg 3				After dropping Catg 3		
	Catg 1	Catg 2	Catg 3		Catg 1	Catg 2	
Catg 1	1	0	0		1	0	
Catg 2	0	1	0		0	1	
Catg 3	0	0	1		0	0	-> 0 0 is able to explain that data is related to Catg 3

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" and "atemp" both having almost similar correlation with "cnt" target variable which can be seen using below pairplot.



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)

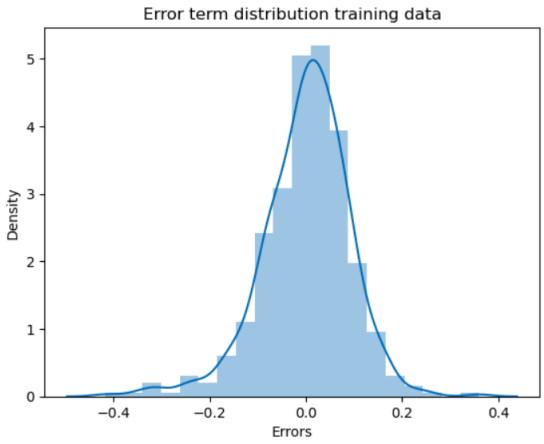
Below checks are executed to run through linear regression assumptions-

A. Linear relationship

Check whether there is linear relationship with independent and dependent variable using pair plot for continuous numeric variables as shown in above plot and as present in Jupyter notebook.

B. Error term mean is zero.

After error term distribution plot as shown below, it is clear that error term mean is around 0.



C. Normal distribution of error terms

From above plot, it is clear that error term is having normal distribution based on symmetry

D. Homoscedacity

Homoscedacity is to check error term distribution against predictor variable. Homoscedacity is checked against temperature and windspeed continuous variables. Variance of error terms doesn't seem to be changing with change in predictor variable. Please refer Jupyter notebook for more details.

E. Multicollinearity

Multicollinearity was present after automatic selection of features, like humidity, Jul and was removed after getting variance_inflation_factor (VIF) score of each feature. VIF function is taken from statsmodels.

VIF score of features is below 5 on final model, well within range. Please refer Jupyter notebook for final VIF score.

	Features	VIF
0	year	2.066336
1	holiday	1.042493
2	temp	3.837505
3	windspeed	4.592815
4	spring	1.994767
5	summer	1.892565
6	winter	1.634327
7	light_rain	1.080748
8	misty	1.543395
9	Sep	1.228923

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)

Based on final model, below are top 3 features contributing significantly either positively or negatively. Values are taken based on coefficient values of each feature.

feature	coefficient
temp	0.478
Yr (year)	0.234
light_rain weather	-0.286

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)

<Your answer for Question 6 goes here>

Linear regression is supervised machine learning algorithm that can explain linear relationship between set of independent variables with dependent variable and explain any change in independent variables, what impact it could have on dependent variable.

Multiple linear equation can be represented as below

$y = c + m1X1 + m2X2 + m3X3 \ldots$

where y is dependent variable, c is intercept, m1, m2, m3 are coefficients of independent variable and X1, X2, X3 are independent variables. There could be positive relationship, negative or no relationship. Linear regression is useful if relationship is either positive or negative.

Coefficients are calculated based on OLS (Ordinary Least of Squares) which can result in minimum residual sum of squares.

There are certain assumptions made by Simple Linear regression

A. Linear relationship between independent and dependent variable
B. Error terms mean should be 0
C. Error terms should be normally distributed
D. Error terms should be independent of each other
E. Error terms variance should not follow any pattern

In addition to above, Multiple Linear regression have below new considerations-

A. Overfit, where model becomes too complex and remembers all values of training data set, but fails to predict output on test data set as it doesn't give generic output
B. Multicollinearity, where individual or set of independent variables have high

---

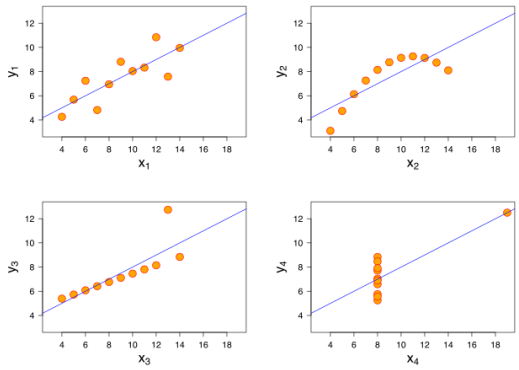
correlation with each other. In such case adding these variables just increases complexity

C. Feature selection based on automated approach, p value, VIF score and business params increases significance of model that can predict with minimum errors on new data or test data.

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)

<Your answer for Question 7 goes here>

Anscombe's quartet was developed by Anscombe. There are 4 dataset possesses same properties, like mean, variance, correlation, linear regression line, coefficients, but distribution is different.



Property	Value
Mean of x	9
Sample variance of x: s2	11
Mean of y	7.5
Sample variance of y: s2	4.125
Correlation between x and y	0.816
Linear regression line	Y = 3.00 + 0.500x
Coefficient of determination of the linear regression: R2	0.67

A. Graph 1 looks linear
B. Graph 2 is not distributed normally and there is relationship between them and its not linear
C. Graph 3 distribution is linear, but regression line is different because outlier is impacting the result.
D. Graph 4 where 1 high point is enough to create correlation, even though rest of the data points doesn't show any correlation.

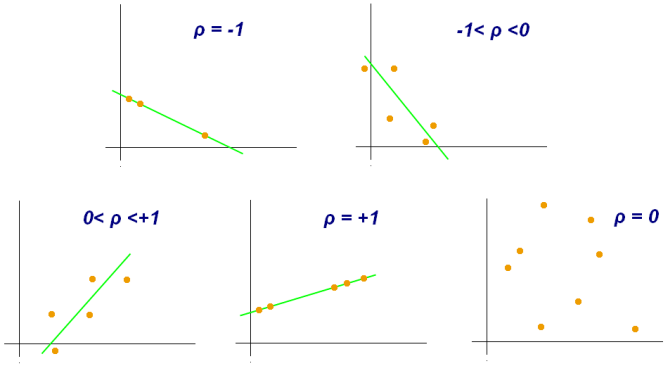
Illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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)

<Your answer for Question 8 goes here>

Pearson's R is correlation coefficient that measures linear relation between 2 variables. Value describes the direction and strength of relationship between 2 numeric variables. Pearson's R value is in the range between -1 to 1.

Below graph shows the slope for multiple values of Pearson's R.



Below table gives direction and strength of relationship.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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)

<Your answer for Question 9 goes here>

Scaling is a method to normalize or standardize data. It is pre-processing step and applied on training data set, not on test data set.

In real world scenarios, data from different variable could vary drastically which could increase problems interpreting it. If we reduce this magnitude and fit dataset into comparable data with bounded range or fixed range, then interpretation also becomes easy. Also, scaling doesn't impact model, it could impact only coefficient of independent variables. Without scaling, coefficient values of independent variable could vary a lot.

There are mainly 2 scaling features present as Min-Max Scaling and standardized scaling. Difference is as follows-

	Normalized scaling	Standardized scaling
1	Minimum and maximum values are used	Mean and standard deviation is used
2	Scaled values are between 0 and 1	Scaled values are not bounded to a certain range
3	Can be affected by outliers	Much less affected by outliers
4	Formula = (x - xmin) / (xmax - xmin)	Formula = (x – xmean) / std
5	If data is not normally distributed or mean is not 0	If data is normally distributed and mean is 0

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)

<Your answer for Question 10 goes here>

VIF could become infinite in only 1 case where there are no residuals present. As per given below formula, R2 if is 1, VIF will become infinite. R2=1 means there is perfect correlation exists between variables.

$$VIF_i = \frac{1}{1-R_i^2}$$

We calculate VIF to check multicollinearity issue where independent variable can be explained by set of independent variables showing strong relationship or not. Higher VIF means higher chances of variable becoming redundant.

To solve infinite VIF, we need to drop one of the variable which is causing this multicollinearity.

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)

<Your answer for Question 11 goes here>

Q-Q plot is a quantile-quantile plot where quantiles of the first data set are compared against quantiles of the second data set. Normally it is checked against theoretical data vs observed data.

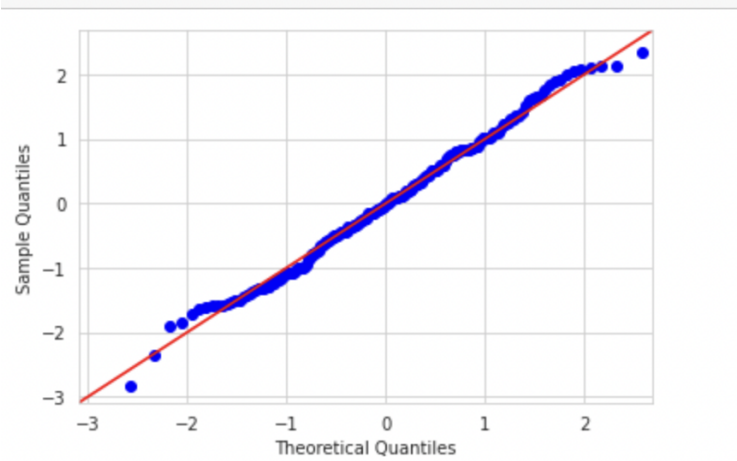
It helps to check below considerations-

A. come from populations with a common distribution
B. have common location and scale
C. have similar distributional shapes
D. have similar tail behavior

There are 4 interpretations

A. Similar distribution

If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis



B. y values < X values

If y-quantiles are lower than the x-quantiles.

C. y values > X values

If x-quantiles are lower than the y-quantiles.

D. Different distribution

If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

---