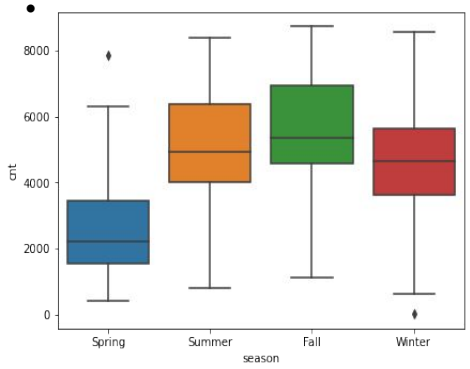
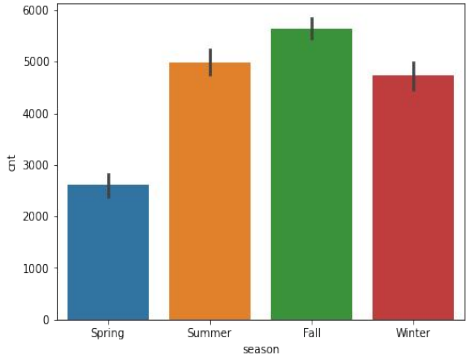
BIKE SHARING ASSIGNMENT

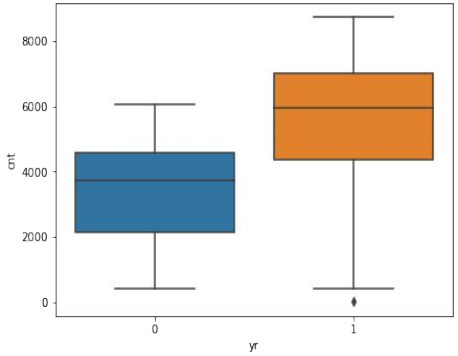
SUBMISSION

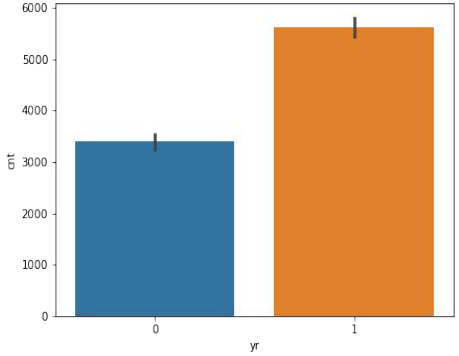
Name: Saumya Bisht

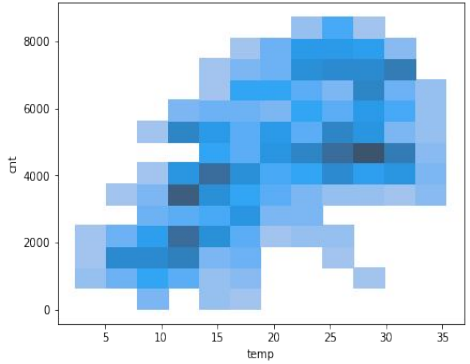
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

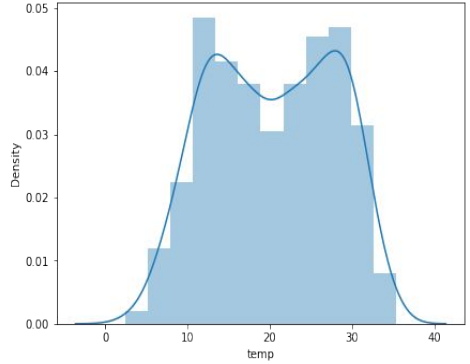








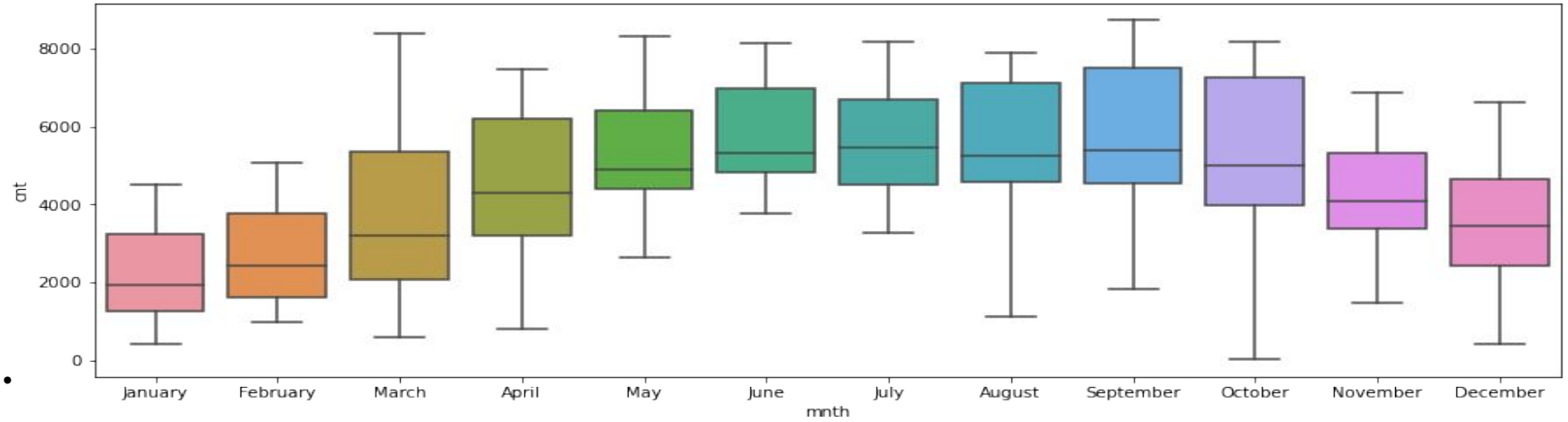


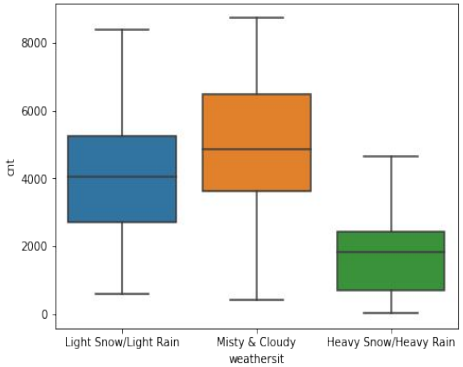


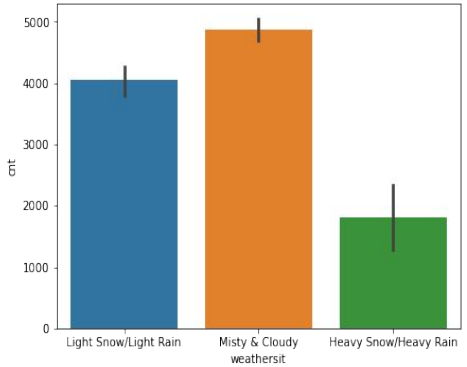
We can infer from these categorical variables that :-

- There is higher demand during the fall season.
- The year 2019 have significant demand compared to 2018.
- The demand rises with the rise in temperature.

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







We can infer from these categorical variables that :-

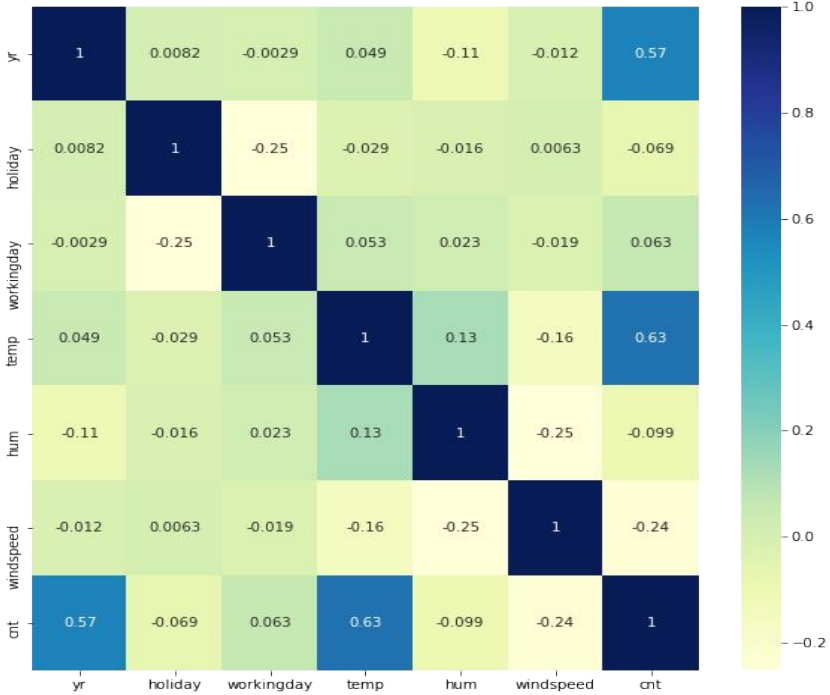
- There is demand increase as the year progresses and peaks at September.
- The demand is highest for Misty and Cloudy weather followed by Light Snow/ Light Rain.

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

Let’s say we have 3 types of values in Categorical column as Spring, Summer, Fall and Winter and we want to create dummy variable for that column. When we create dummy variable it create a column for each category and put the value 1 if it was there in that row or 0 if it was not present in that row. If value in the Spring column is 1 we know it’s Spring and others must be 0, If value in the Summer column is 1 we know it’s Summer and others must be, If value in the Fall column is 1 we know it’s Fall and others must be 0 but when all value of each column is 0 we can say that it must be Winter. So we do not need 4th variable to identify the Winter.

Thus, drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

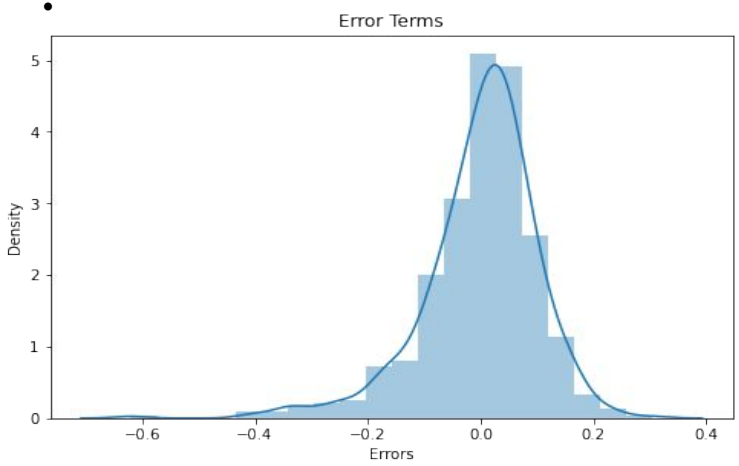
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

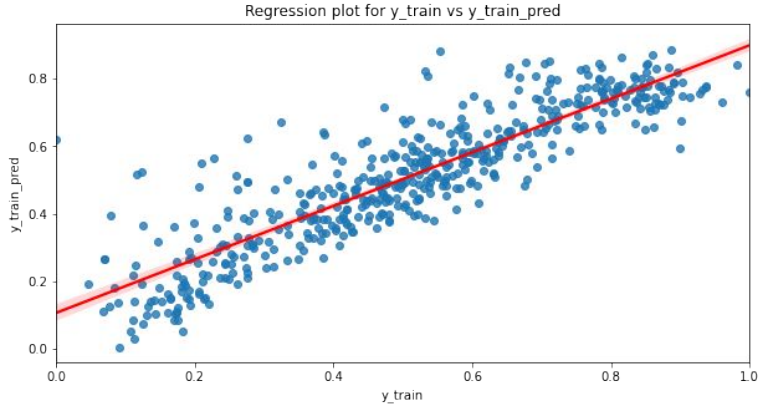
- 

From the pair plot we can see that temperature has the highest correlation with the target variable.

	yr	holiday	workingday	temp	hum	windspeed	cnt
yr	1.000000	0.008195	-0.002945	0.048789	-0.112547	-0.011624	0.569728
holiday	0.008195	1.000000	-0.252948	-0.028764	-0.015662	0.006257	-0.068764
workingday	-0.002945	-0.252948	1.000000	0.053470	0.023202	-0.018666	0.062542
temp	0.048789	-0.028764	0.053470	1.000000	0.128565	-0.158186	0.627044
hum	-0.112547	-0.015662	0.023202	0.128565	1.000000	-0.248506	-0.098543
windspeed	-0.011624	0.006257	-0.018666	-0.158186	-0.248506	1.000000	-0.235132
cnt	0.569728	-0.068764	0.062542	0.627044	-0.098543	-0.235132	1.000000

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





Here we can see on the top that the error terms are normally distributed with mean zero.

In our final model has an R- squared of 0.791 on the training set and 0.769 on the test set with a low Prob (F-statistic) tells that our model fits a linear line.

In our final we have selected features with VIF < 5. Thus, reducing multicollinearity.

To verify that the observations are not auto-correlated, we can use the Durbin-Watson test. The closer it is to 2, the less auto-correlation there is between the various variables. For our model the value is 2.058. Thus, error terms are independent of each other.

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

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	210.3
Date:	Mon, 08 Mar 2021	Prob (F-statistic):	1.01e-163
Time:	14:24:22	Log-Likelihood:	437.74
No. Observations:	510	AIC:	-855.5
Df Residuals:	500	BIC:	-813.1
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0788	0.021	3.760	0.000	0.038	0.120
yr	0.2389	0.009	25.860	0.000	0.221	0.257
workingday	0.0482	0.013	3.840	0.000	0.024	0.073
temp	0.5515	0.022	24.815	0.000	0.508	0.595
windspeed	-0.1839	0.028	-6.553	0.000	-0.239	-0.129
September	0.0874	0.018	4.936	0.000	0.053	0.122
Monday	0.0593	0.016	3.665	0.000	0.028	0.091
Light Snow/Light Rain	-0.0677	0.010	-6.920	0.000	-0.087	-0.048
Summer	0.0875	0.012	7.528	0.000	0.065	0.110
Winter	0.1174	0.012	10.159	0.000	0.095	0.140

Omnibus:	137.384	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	459.564
Skew:	-1.229	Prob(JB):	1.61e-100
Kurtosis:	6.947	Cond. No.	11.5

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Based on the final model we can see the top 3 features which are contributing significantly towards explaining the demand are :-

1) temp - Temperature,
2) yr- Year, and
3) windspeed - Wind Speed.

Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

1) Simple Linear Regression uses traditional slope-intercept form, where *m is the slope coefficient* and *c* is our intercept. *x is our independent variable* and *y* represents our dependent variable.

   $y = mx + c$

2) Multiple linear regression is a statistical technique that uses several independent variables to predict the outcome of a dependent variable. Here y is our dependent variable, xp is our independent variable, β0 is the intercept, βp is coefficient of each independent variable and ϵ the error term in the model.

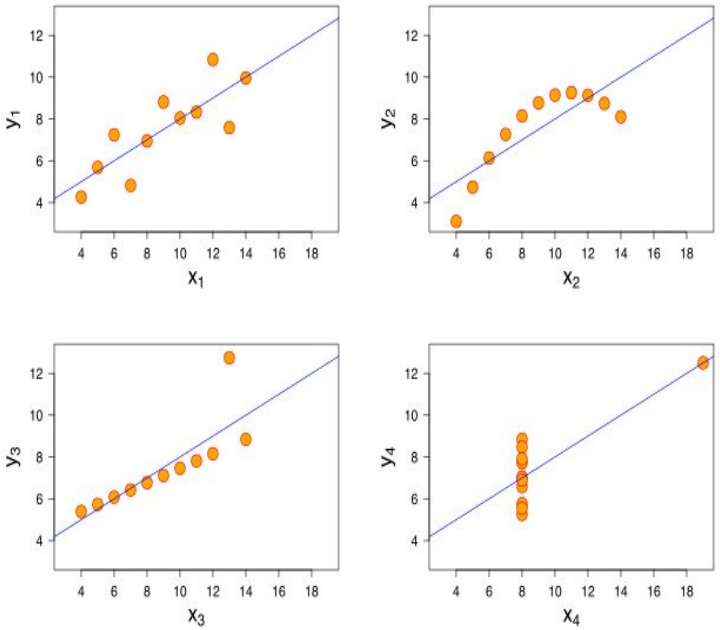
   $y = \beta 0 + \beta 1x1 + \beta 2xi + ... + \beta pxp + \epsilon$

   The Linear Regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables,
- Error terms have *constant variance* (homoscedasticity),
- Error terms are *independent* of each other,
- Error terms are *normally distributed* with mean zero.

Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analyzing it and the effect of outliers on statistical properties.



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you 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 which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is still often used to 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

What is Pearson's R?

Pearson's R or Pearson correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables. Pearson correlation coefficient 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

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

Scaling or Feature scaling is a method used to normalize the range of independent variables or features of data. It is a data preprocessing step and performed before building a model. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization.Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

There are two method for scaling

1) Min-Max Scaling :-

It is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$ where x is an original value, x' is the normalized value.

2) Standardization :-

Standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance.The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

$$x' = \frac{x - \bar{x}}{\sigma}$$ Where x is the original feature vector, $\bar{x} = \text{average}(x)$ is the mean of that feature vector, and σ is its standard deviation.

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

The variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate standard deviation) of an estimated regression coefficient is increased because of collinearity. The formula for VIF is given as:

$$VIF = \frac{1}{1 - R^2}$$

Where R^2 is the coefficient of determination.

When the the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared or the coefficient of determination = 1, which lead to 1/(1-1) which is infinite. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

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) X-values < Y-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