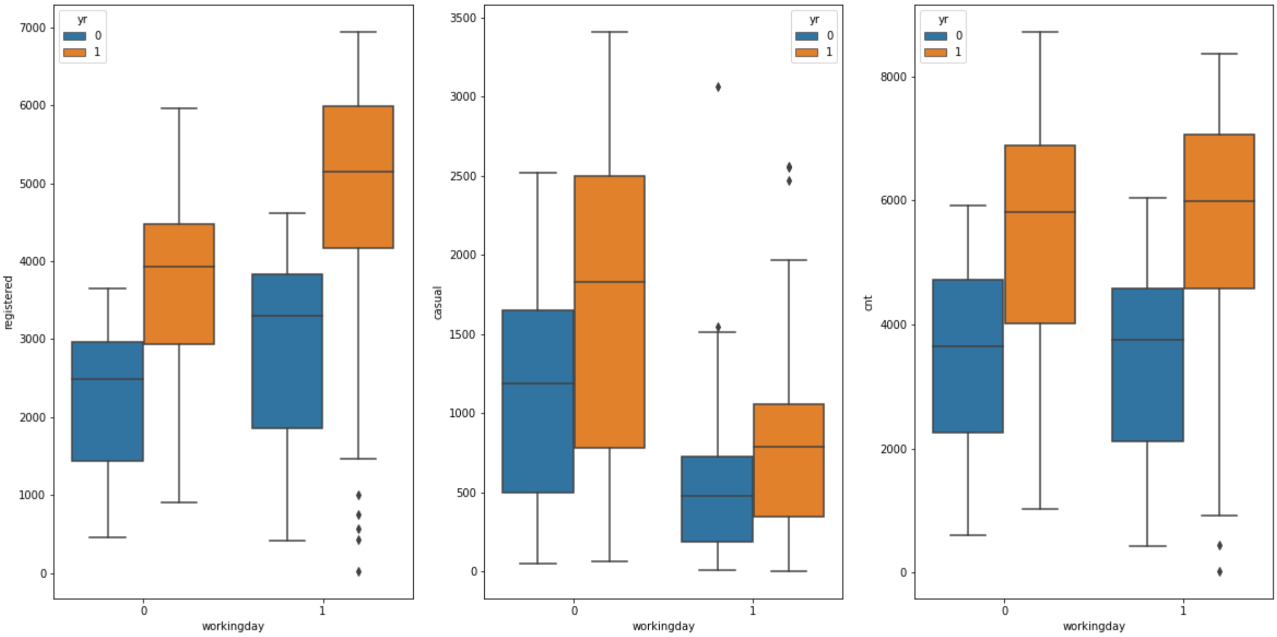
Assignment-based Subjective Questions

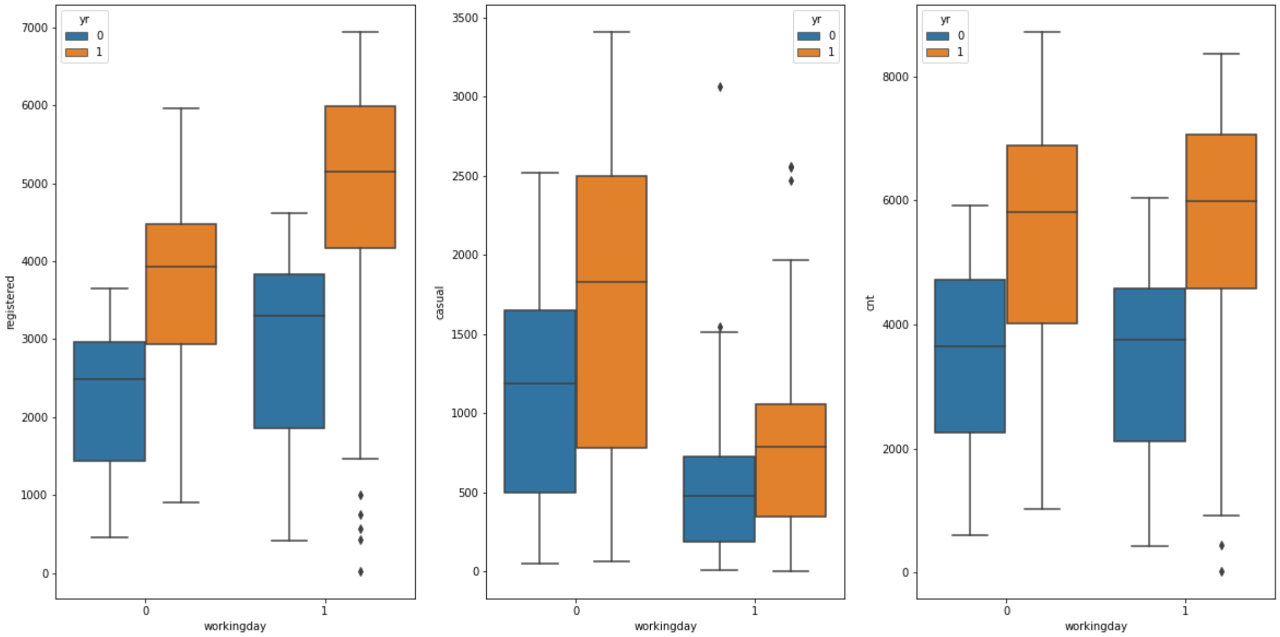
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

The original data had a total of 16 variables, of which except three (casual, registered and cnt) rest all could be categorised as categorical variables. Continuous variables like date, temp, atemp, hum, windspeed were also considered as categorical as they are perceptional. Human beings don't look at them as an absolute number but as a condition. Whether day is Sunday or not, temperature is hot or not or its too humid or not. Hence these variables were considered as categorical, binned and evaluated. The summary of the bivariate analysis is as tabulated below:-



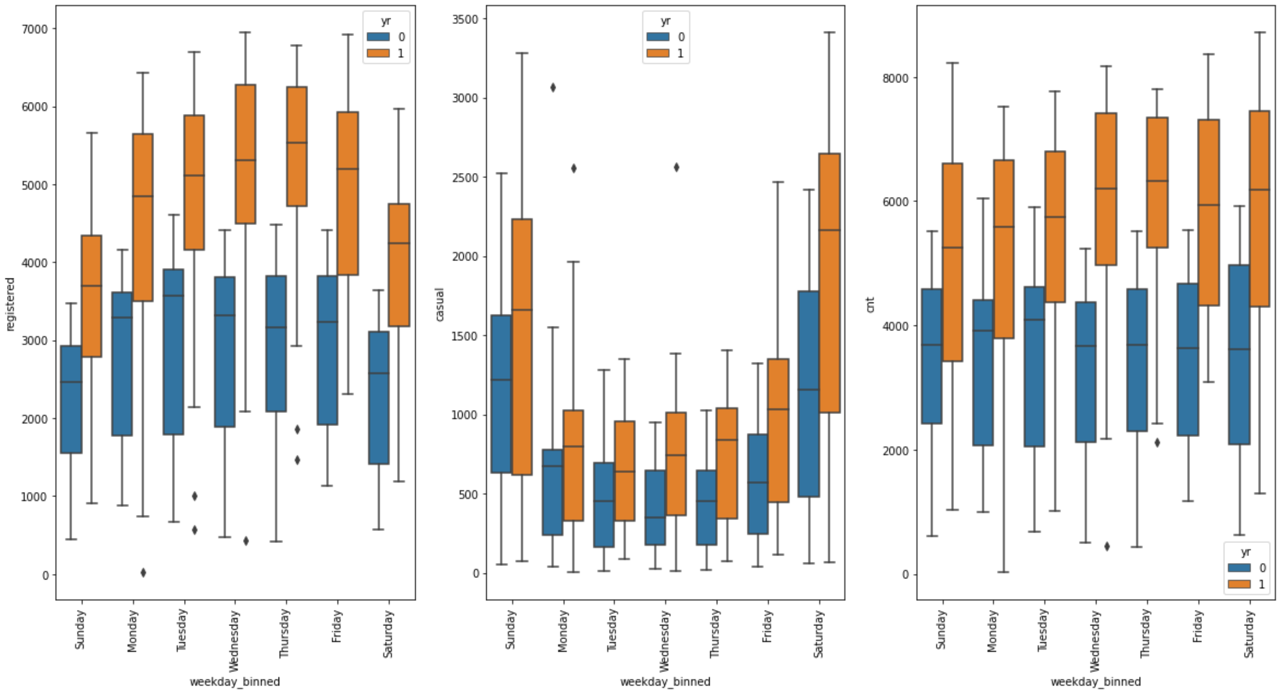
Plot 1: Weather Situation vs Dependent Variable(s)

Inference: Rental demand is higher when the weather is fair (clear or mildly cloudy)



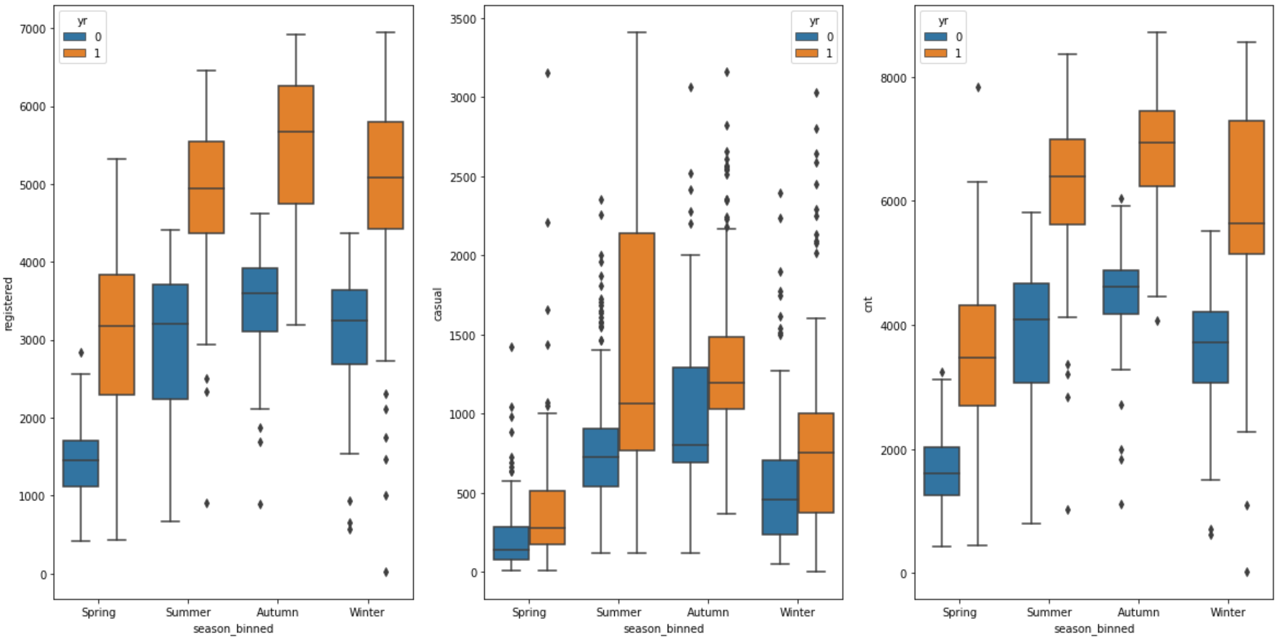
Plot 2: Working Day vs Dependent Variable(s)

Inference: Not much effect on overall demand, however surge observed in casual rentals on non-working days



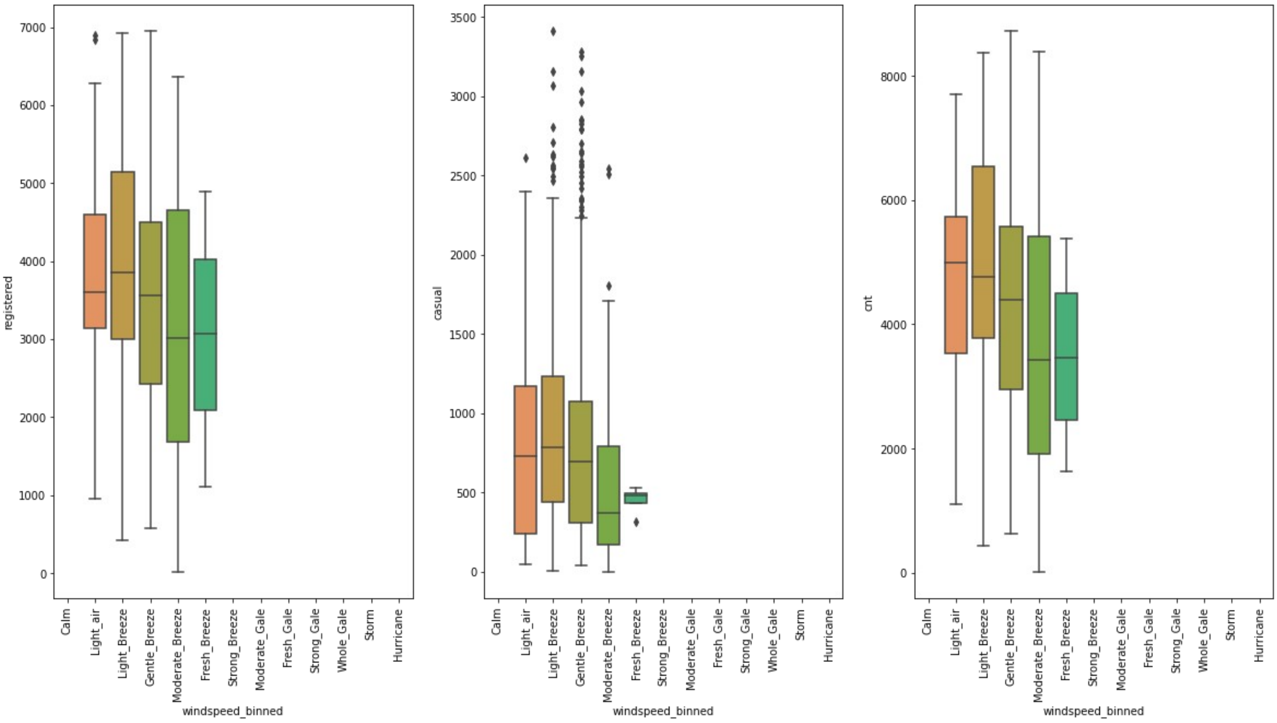
Plot 3: Day of the week vs Dependent Variable(s)

Inference: Not much effect on overall demand, however surge observed in casual rentals on Saturday and Sunday (which incidentally are also non-working days)



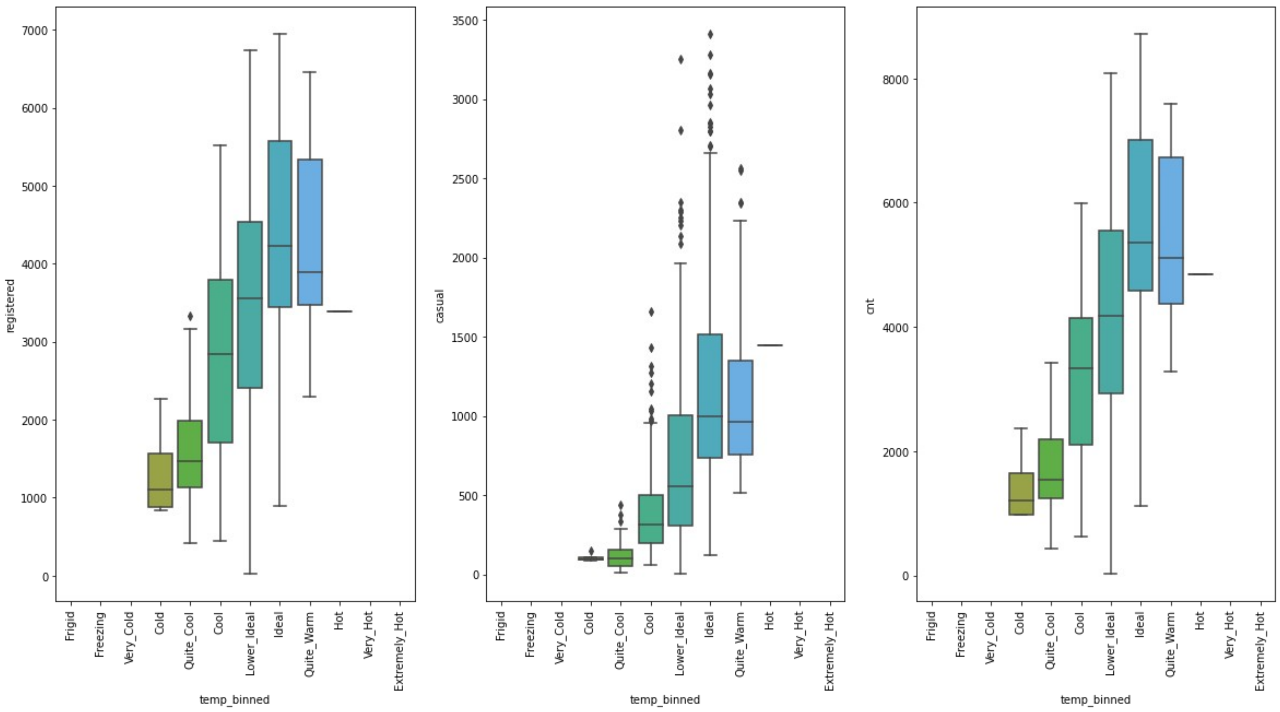
Plot 4: Seasons vs Dependent Variable(s)

Inference: Significant increase observed in Summer, Autumn and Winter season. Significant surge observed in casual hiring in Summer season.



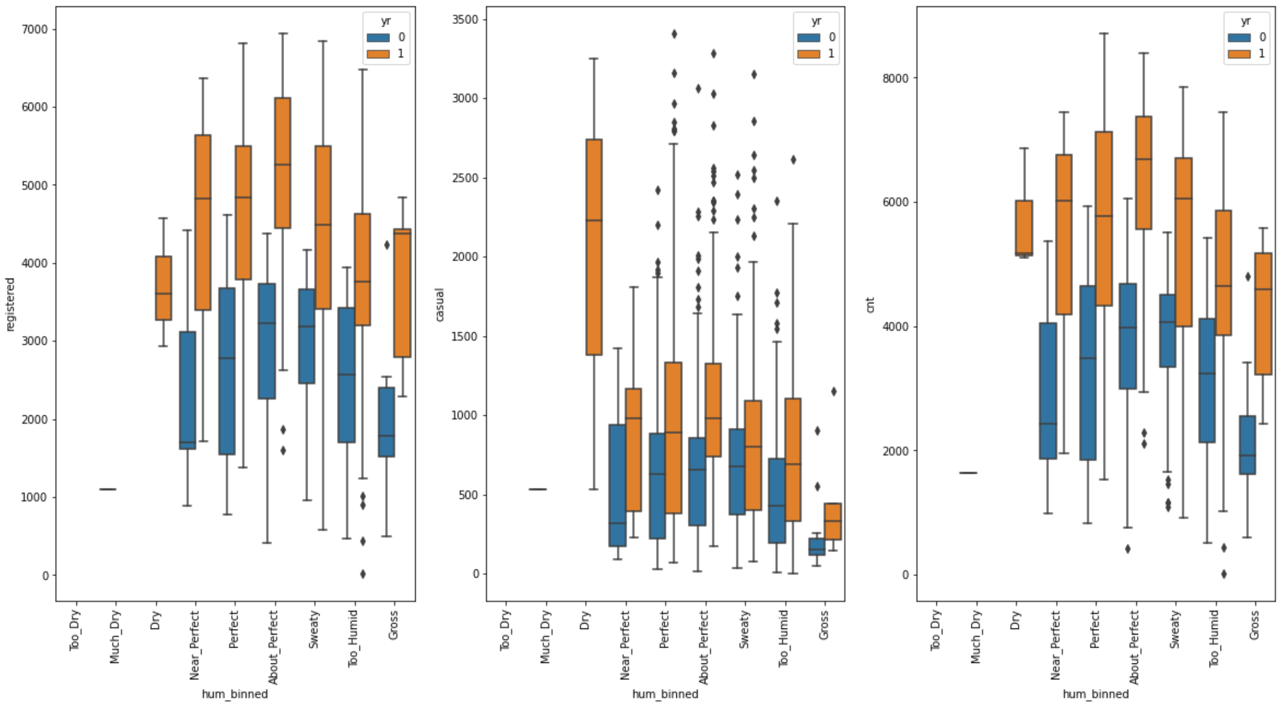
Plot 5: Wind speed vs Dependent Variable(s)

Inference: Increase in wind speed has a detrimental effect on number of bikes hired. Noticeable dip observed in casual rider-ship when wind speeds exceed 19 KmpH



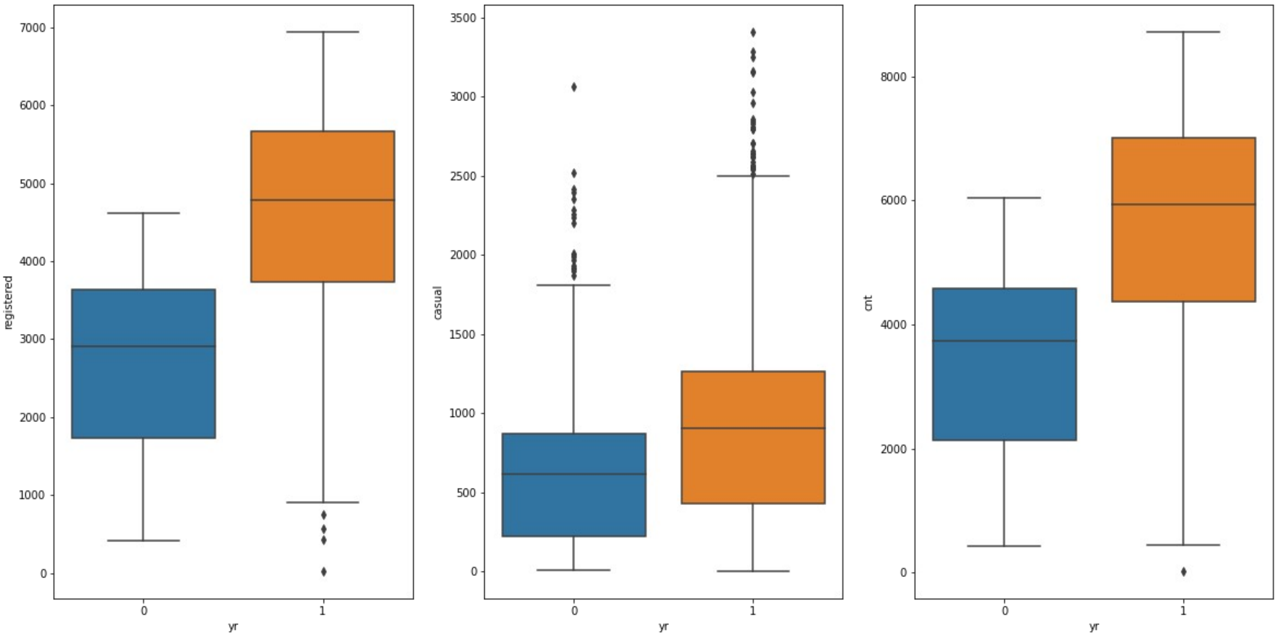
Plot 6: Temperature vs Dependent Variable(s)

Inference: Temperature is directly correlated rider-ship till its quite warm (<35 deg C). Beyond which there are hardly any rentals.



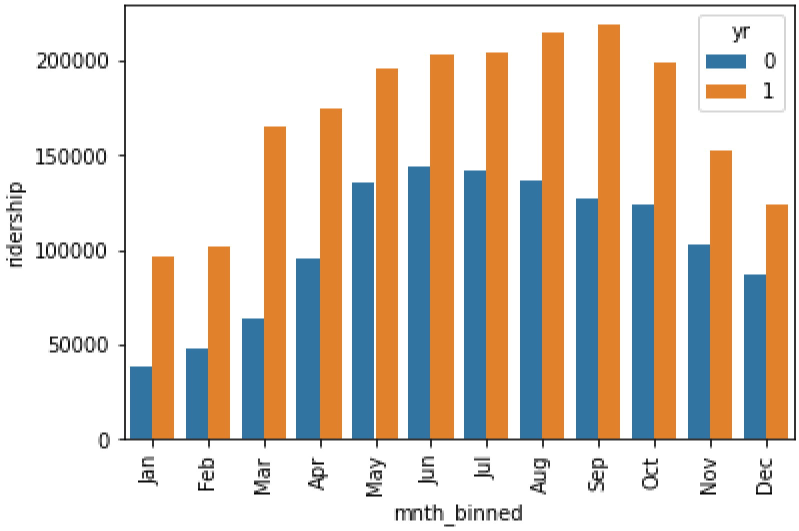
Plot 7: Humidity vs Dependent Variable(s)

Inference: Rentals are observed only during periods when humidity is within comfort levels, lower and beyond which they taper off.



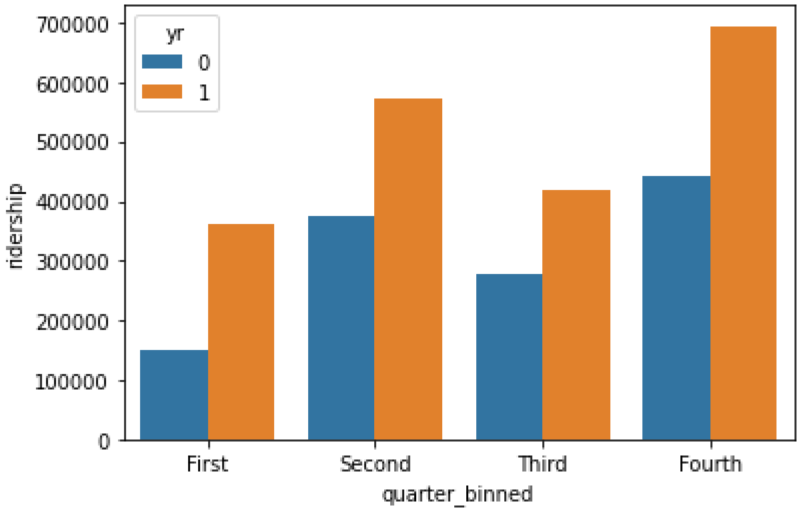
Plot 8: Year vs Dependent Variable(S)

Inference: Rental has shown growth YoY. However, data is sparse to draw any meaningful inferences.



Plot 9: Months vs Dependent Variable(S)

Inference: Rental demand is observed to increase as winter reduces and peaking towards autumn after which it reduces as Winter approaches.



Plot 10: Quarter vs Dependent Variable(s)

Inference: Earnings show an upswing in the Second and Fourth Quarter.

Based on the above inference, we can make following general inferences on the influence of categorical variables on the dependent variables:-

(a) Categorical variables do have significant impact on the dependent variables.
(b) In the given data set, weather conditions appear to be the driving factors in determining the strength of demand.
(c) Some categorical variables like day of the week or working day may not have a signifiant impact on the overall rider ship, however, they do significantly affect the sub-group classification.

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

Using binary encoding scheme, ***n*** unique categorical variables can be completely described using ***n-1*** binary variables. When creating dummy variables for a categorical column, the function generates ***n*** columns. However, since all the ***n v***ariables can be described by ***n-1*** variables***,*** the ***drop_fist=True*** parameter is selected so that post creation of columns the head column(first) is automatically dropped and the user does not have to write an extra line of code to implement it.

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

Variable ***registered*** has the highest correlation with target variable ***cnt***.

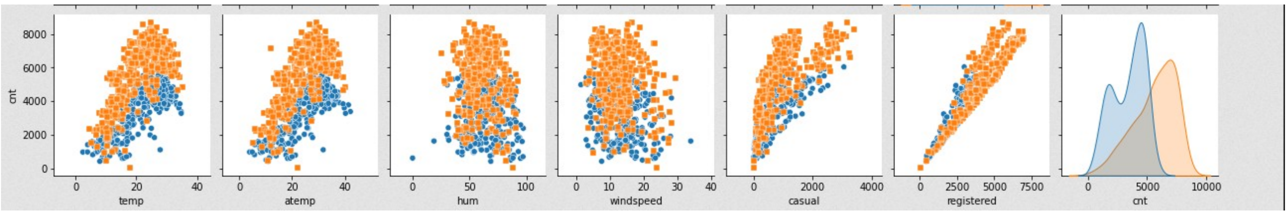


Illustration 1: Pairplot of numerical variables with target variable `cnt`

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

Based on the model parameters generated on the training dataset, the model coefficients were applied on the test dataset. Thereafter, following evaluation metrics were applied to validate assumptions

(a) Plot of predicted vs actual was plotted to see the spread of data.

(b) Q-Q plot was charted to observe the distribution of errors and residuals. The plot tells whether the errors are normally distributed or not.

(c) Histogram of error terms (predicted-actual) was plotted to observe the trend of error distribution or not.

(d) Regression plot of predicted vs residuals was charted to observe pattern and confirm that residuals are independent of each other.

(f) Line plot of actual vs projected values was plotted to see closeness of fit of the developed model.

(g) Finally, model (test and training set) was evaluated on numerical metrics like R-squared, Mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error.

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

Consequent to the analysis of the final model and the coefficients value, the top three features (as per original data set) are registered, temperature and working day (not in a particular order). The model metrics and coefficient values deduced are

Iteration 1: Model 2: Dependent Variable <u>Not Scaled</u>

R-squared for train data:	0.97
R-squared for test data:	0.95
Mean Squared Error for train data	112666.492
Mean Absolute Error for train data	250.039
Mean Squared Error for test data	175299.653
Mean Absolute Error for test data	290.521
Mean Absolute Percentage Error for test data	0.096
Mean Absolute Percentage Error for train data	0.099

Final Parameters for Model 2 based on RFE+VIF analysis are :

const	800.214639
workingday	-856.643319
temp	1015.550242
hum	-468.549067
windspeed	-430.695640
registered	7977.926665
Tuesday	-166.486956
Wednesday	-192.740461
Thursday	-121.406911
Sep	182.969600
Oct	383.600291
Nov	212.423360
Summer	252.801051
Winter	-219.449704

Iteration 2: Model 2: Dependent Variable Scaled

R-squared for train data:	0.97
R-squared for test data:	0.95
Mean Squared Error for train data	0.001
Mean Absolute Error for train data	0.029
Mean Squared Error for test data	0.002
Mean Absolute Error for test data	0.033
Mean Absolute Percentage Error for test data	0.098
Mean Absolute Percentage Error for train data	3.464

Final Parameters for Model 2 based on RFE+VIF analysis are :

const	0.089532
workingday	-0.098555
temp	0.116837
hum	-0.053906
windspeed	-0.049551
registered	0.917847
Tuesday	-0.019154
Wednesday	-0.022174
Thursday	-0.013968
Sep	0.021050
Oct	0.044133
Nov	0.024439
Summer	0.029084
Winter	-0.025247

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression (LR) is a supervised learning algorithm that uses a line as a function that approximately characterises all the data points in given set. The LR graphically shows as a line that passes through / near all the data points in such a manner that the vertical distance between the data points and the line is minimum. If the number of variables in the data set other than dependent variable is one, then LR is called as Simple Linear Regression (SLR) else in case of multiple variables its called as Multiple Linear Regression (MLR).

Mathematically LR is defined as **$y=ß_0+ß_1X$,** where y is the dependent variable and X is an independent variable. $ß_0$ is referred to as the intercept of line or constant. $ß_1$ is called as the coefficient.

Since it is not possible for all the points to be fitted on a straight line, the aim of the Linear Regression is to minimise the Cost Function (CF). CF is the difference between actual value and the predicted value. In case of LR, CF is defined as Root Mean Squared Error (RMSE) or the mean Mean Squared Error (MSE). This method is also called as Ordinary Least Square (OLS) Method and has been applied in the instant study.

$$MSE=\frac{1}{n}\sum_{i=1}^{i=n}(\hat{y}i-yi)^2$$

To minimise $ß_0$ and $ß_1$ values Gradient Descent (GD) method is used. In this method coefficient values are selected randomly and thereafter iteratively updated to arrive at the minimum value for the cost function. The GD algorithm calculates the next point by calculating the gradient at the current position and scales it by the learning rate and later subtracts it from the current position. The learning rate therefore determines the step size. Too less would imply a slowly iterating algorithm. A big step would lead to overstepping the minima, thereby giving false minima.

Once the minima is established, the value of the coefficients, $ß_0$ and $ß_1$, is calculated to deduce the predictor function. In case of MLR, there would be number of $ß_1$ corresponding to the number of independent variables.

2. Explain the Anscombe's quartet in detail.

Francis Anscombe devised the Anscombe's quartet in 1973 to highlight the pitfalls of undertaking statistical analysis of data set using only purely statistical means especially classical statistical methods like mean, median and mode. For this purpose, he deduced a four unique type of sample data sets that had identical/ similar statistical characteristics like mean, median, mode, standard deviation etc., but had substantially different graphical distribution (when plotted on a chart). With these charts and associated data sets he highlighted the requirement and importance of graphical assessment of data along with numerical statistical assessment to give a more better contextual understanding of the data.

3. What is Pearson's R?

Pearson R, also called as Pearson's Correlation Coefficient, is a numerical value that indicates the strength (numerical value) and direction (positive or negative) relationship between two variables. The variables may be continuous or categorical. The value of the coefficient can vary from +1 to -1 with value 0 indication no correlation between two given variables. Based on the degree of correlation, an analyst can ascertain whether the given two variables are independent or not, which is a significant whilst undertaking regression modelling.

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

Scaling of variables is a method to transform data over a defined scale (min-max range).

Since value of a variable has a large impact on the value of a coefficient, especially, in a linear model; in a multivariable data set with a large number of variables across different value range, retaining original values will lead to large swings in the coefficients associated with that variable. This will lead to difficulty in undertaking a realistic assessment of importance of a particular variable based on the size of its coefficient. Scaling compresses/expands variable data over the defined range and if the same scaling limits are applied to all variables, this will lead to generation of coefficients in a linear model that can be compared easily.

In normalised scaling the variable values are scaled between limits 0 and 1, commonly referred to as Min-Max scaling. Whereas, in standardised scaling the variable values are scaled to have a mean of 0 and a standard deviation of 1.

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

The formula for VIF is VIF= $\frac{1}{1-R^2}$. Therefore, VIF = ∞, IFF, R^2=1. R^2=1 will occur when there is perfect correlation between variables. i.e, the variable being evaluated can be precisely defined by other variable.

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

Q-Q plot or Quantile-Quantile plot, is a plot of quantiles of two sets on an XY chart. The chart is used to evaluate distribution of two data sets. If distribution of both the data sets are similar, the Q-Q plot will follow a line. Therefore, based on the plot one can analyse, whether distribution of sets is similar or not.

In a linear regression, the Q-Q plot is used to reconfirm the distribution of the residuals. They are also used to find the skewness and kurtosis of the distribution. It is also used to validate the basic assumption of a linear regression, i.e, its linear.