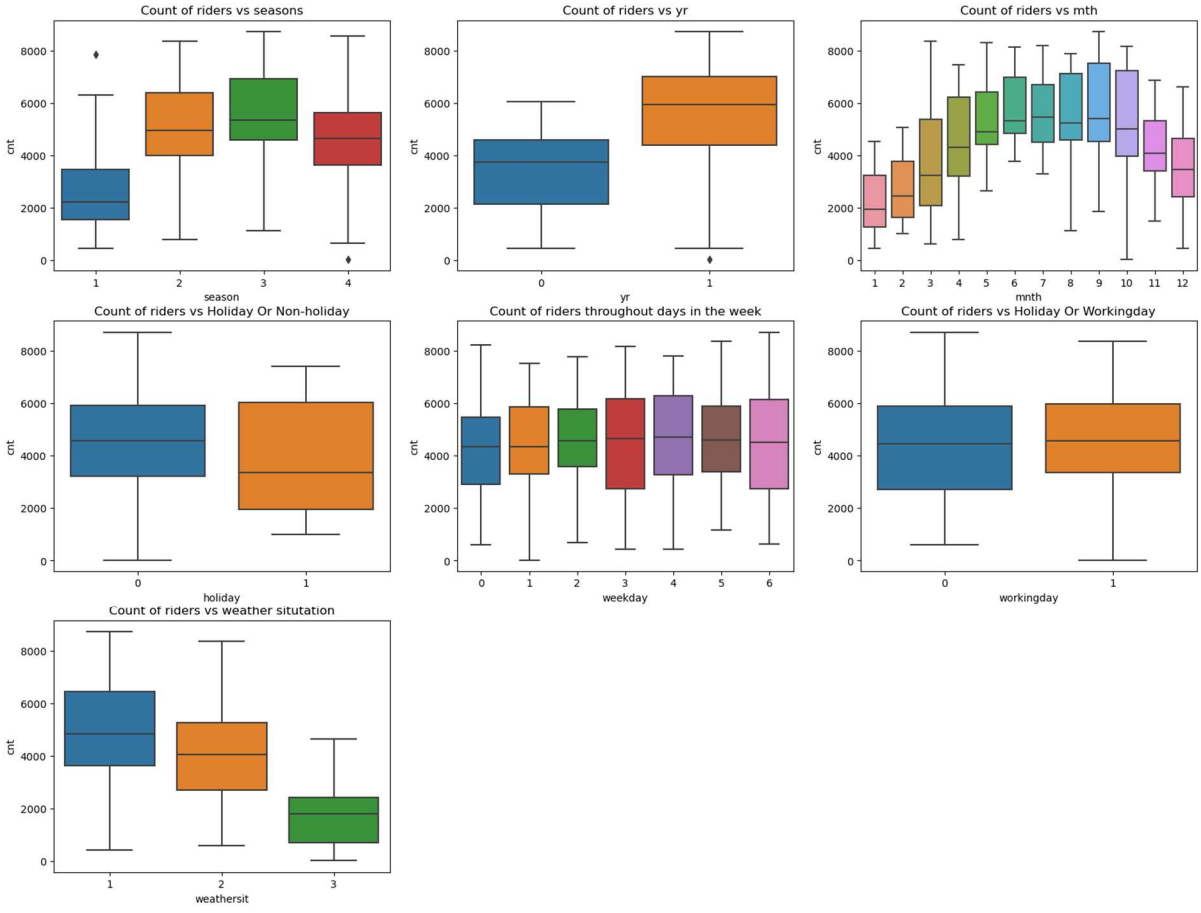
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer1:

The analysis of categorical variables is performed by plotting a scatterplot or a boxplot on x axis (considered as independent variable) and their effect on target variable '**cnt'** is analysed on y axis

- From the boxplot it can be inferred that:
    1. Overall number of **riders increased** from **2018** to **2019**.
    2. More number of people prefer to ride bikes where **weather** situation is **clear**.
    3. **Summer** and **fall** season have higher count of mean riders.
    4. As **Summer** and fall has higher number of riders it is also reflected the increased mean ridership from **August** till **October**
    5. Mean riders are higher on a no **holiday** clearly means ridership **decreases** during holidays.
    6. Average count of riders remains **constant** during **weekday** and working day.



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

Answer2:

It is important to use drop_first=**True** as we need n-1 **predictors**, since using n predictors would lead to multicollinearity and will not give use the right set of predictors which are impacting the regression. Though **multicollinearity** do not impact regression or predicted value but it will not give clear information about the predictors.

Hence, in our assignment we are creating the following dummy variables

'yr','holiday','workingday','season','mnth','weekday','weathersit'

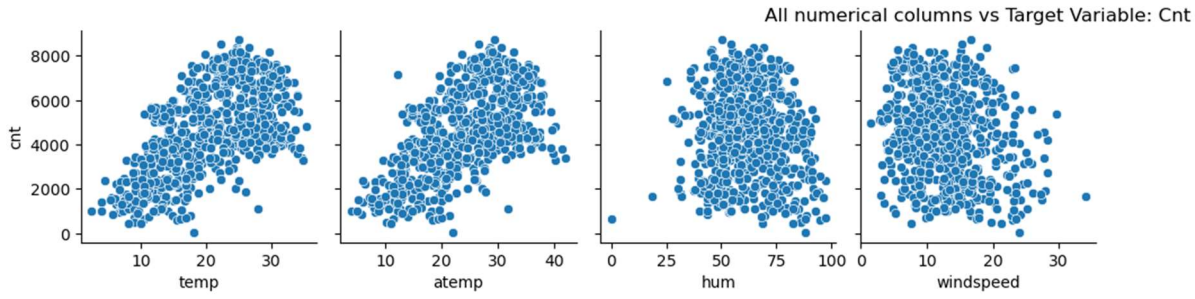
---

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer3:

When numerical variables are plotted against the target '**cnt'** variable, the pair-plot or a scatterplot is analysed to check the linear relationship, '**temp'** variable shows highest correlation with the target 'cnt'. The correlation **coefficient was 0.65**.

Thus, **temp'** variable was highly **correlated**.



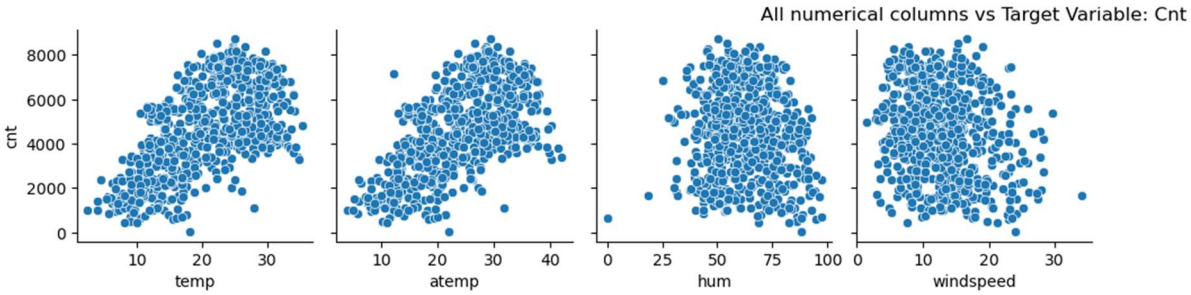
---

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Following checks were performed to validate the assumptions of Linear Regression:

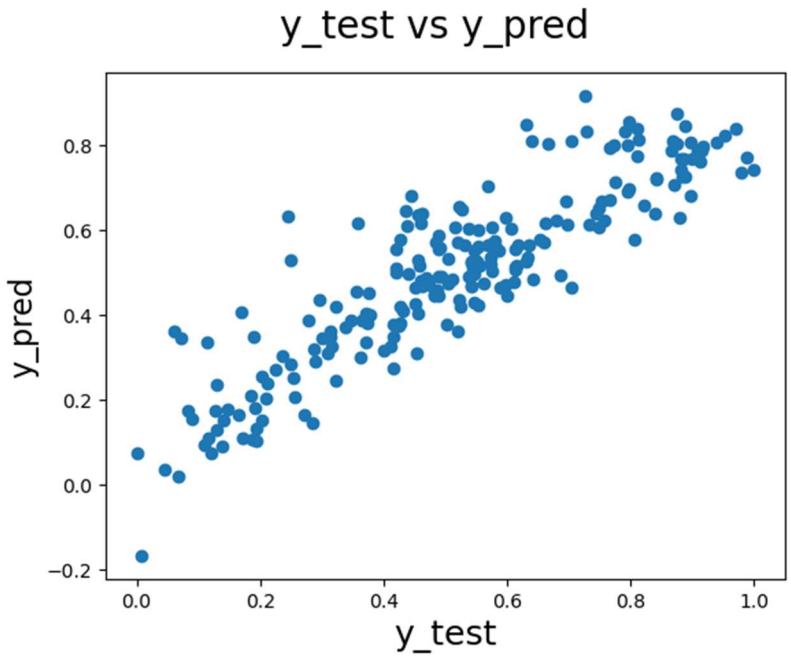
a) **The assumption about the form of the model**: It is assumed that there is a **linear relationship** between the dependent and independent variables. It is known as the '**linearity assumption'**.

Linear regression needs the relationship between the independent and dependent variables to be linear. We use a pair plot to check the relation of independent variables with the cnt variable. This clearly gives information about registered users but we need to further analyse other variables.



We can check the linearity of the data by looking at the **Residual vs Fitted plot**. Ideally, this plot would not have a pattern where the red line is approximately horizontal at zero.

Y_pred = 0.1301+0.5504X**temp**-0.1524X**windspeed**+0.2325Xyr_2019-0.0238X**workingday_working**+0.0875X**season_summer**+0.1306X**season_winter**+0.0975X**mnth_Sept**-0.0792X**weathersit_Cloudy**-0.2822X**weathersit_LightRain**



b) **Zero Mean Assumption**: Residuals are the differences between the true value and the predicted value. One of the assumptions of linear regression is that the mean of the residuals should be zero

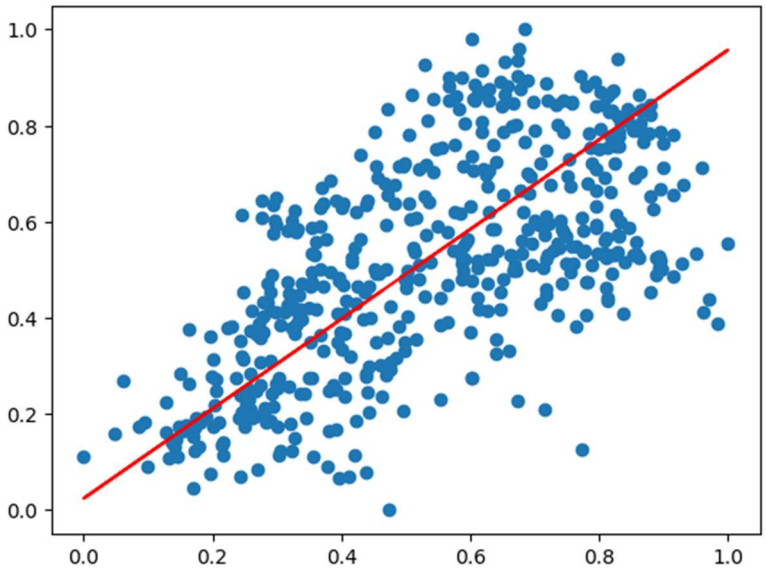
mse = mean_squared_error(y_test, y_pred_m23)

Mean_Squared_Error : 0.012195303369461065
This is coming close to zero.

c) **Check for Homoscedasticity:**

Fit a regression line through the training data using statsmodels. In statsmodels, we need to explicitly fit a constant using sm.add_constant(X) because if we don't perform this step, statsmodels fits a regression line passing through the origin, by default.

This clearly shows the dataset shows linear relationship around the mean, and we should further analyse this using other variables.

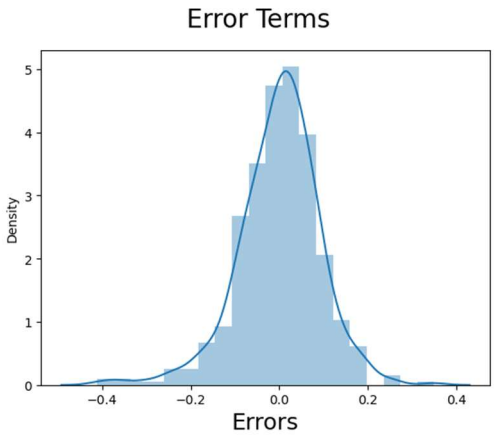


d) **Normality assumption**: error terms are normally distributed

This was validated by plotting a histogram for the delta between expected and predicted value, which should be distributed normally

Error terms = sns.distplot((y_train - y_train_cnt), bins = 20)

Clearly error terms are normally distributed.



e) **Non multicollinearity assumption.** It is assumed that independent variables should not be correlated, as this will impact the r2 value**.** Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. Hence, we drop variables having vif > 5.

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

Answer5:

Looking at the best fitted line, target variable 'cnt' will show high amount of variation, if the value of following independent variable changes out of which top 3 impacting predictors are

a) **Temp**: temperature (1 unit rise in temperature would change **0.55** times cnt)
b) **Yr_2019**: (as compared to previous year **.232** times cnt increased)
c) **Weathersit_lightrain** (1 unit rise in this variable will reduce ridership by **.282** times)

Y = 0.1301+0.5504Xtemp-0.1524Xwindspeed+0.2325Xyr_2019-
0.0238Xworkingday_working+0.0875Xseason_summer+0.1306Xseason_winter
+0.0975Xmnth_Sept-0.0792Xweathersit_Cloudy-0.2822Xweathersit_LightRain

---

<u>1.Explain the linear regression algorithm in detail. (4 marks)</u>

Answer1:

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression (SLR)

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

Y= mx+c

Multiple Linear Regression (MLR)

It is the extension of simple linear regression that predicts a response using two or more features. Mathematically we can explain it as follows −

Consider a dataset having n observations, p features i.e. independent variables and y as one response i.e. dependent variable the regression line for p features can be calculated as follows −

h(xi)=b0+b1xi1+b2xi2+...+bpxip

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is used to demonstrate the importance of data visualization which was developed to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour independent of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

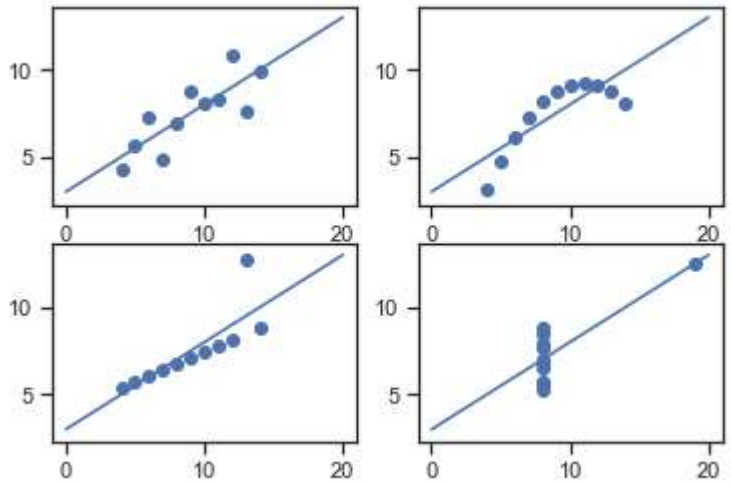
Graphically it can be shown as below

Data-set 1 — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set 2 — shows a curve shape but doesn't show a linear relationship

Data-set 3 —a tight linear relationship between *x* and *y*, except for one large outlier.

Data-set 4 —the value of *x* remains constant, except for one outlier as well.



3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient (r) is a measure of the linear association of two variables. Correlation analysis usually starts with a graphical representation of the relation of data pairs using a scatter diagram. The values of correlation coefficient vary from –1 to +1.

Positive values of correlation coefficient indicate a tendency of one variable to increase or decrease together with another variable. Negative values of correlation coefficient indicate a tendency that the increase of values of one variable is associated with the decrease of values of the other variable and vice versa. Values of correlation coefficient close to zero indicate a low association between variables, and those close to –1 or +1 indicate a strong linear association between two variables.

It can be represented mathematically as below

Where n is sample size

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process to normalize the data for a given range. It is required to bring them all in a single range. The two most popular scaling methods
are **Normalization** and **Standardized**. Normalization typically scales the values into a range of [0,1].
Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

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

(3 marks)

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.
It is denoted by the formula:

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

If VIF is infinite that mean 1-Ri2=0 , hence Ri = 1. This means variables are highly correlated which will make it more difficult to estimate the relationship between each of the independent variables and the dependent variable.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot: Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.

(3 marks)