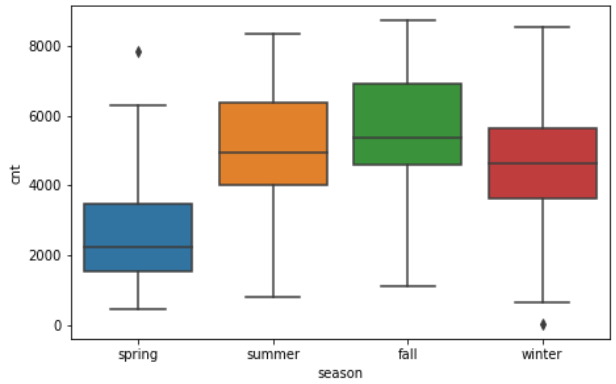
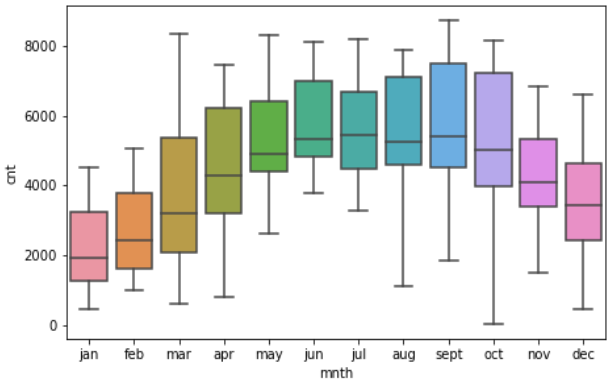
Assignment-based Subjective Questions

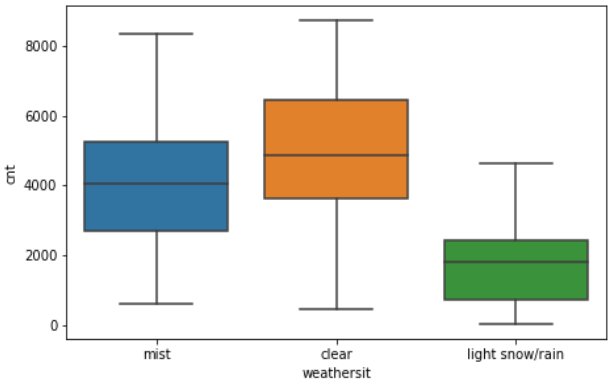
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

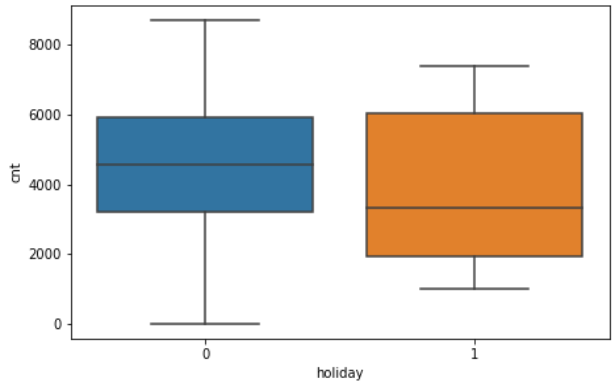
Ans. There are 6 categorical variables namely season,month(mnth),weathersit,holiday,weekday and workingday.The dependent variable is cnt(count of total rental bikes including both casual and registered).

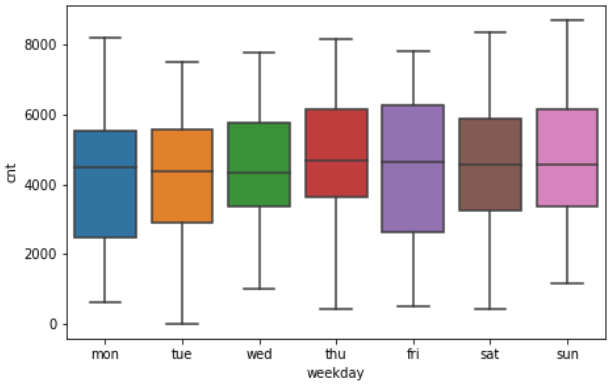
So after doing the exploratory data analysis ,we get the following boxplots.

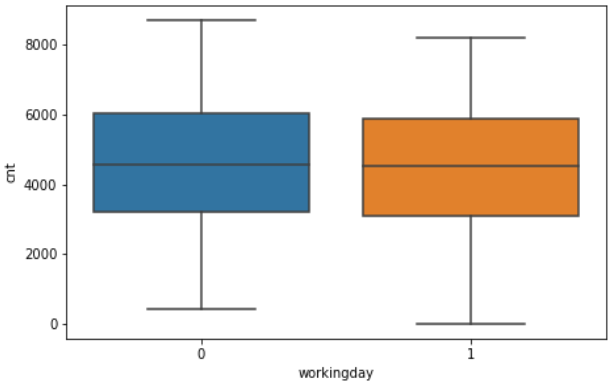












The following observations can be made-

- season: Almost 32% of the bike booking were happening 'fall' with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter . This indicates, season can be a good predictor for the dependent variable.

- mnth: Demand is continuously growing each month till June to September.September has highest demand. After September, demand is decreasing with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

- weathersit: Maximum of the bike booking were happening during ‘clear' i.e Clear, Few clouds, Partly cloudy, Partly cloudy- with a median of close to 5000 booking (for the period of 2 years). This was followed by light snow/rain . This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

- holiday: this data is clearly biased. This indicates, holiday cannot be a good predictor for the dependent variable.

- weekday: weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. It will be decided by the model if this needs to be added or not.

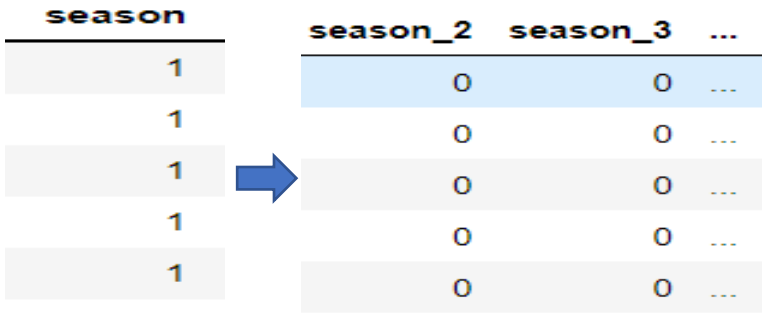
- workingday: workingday needs to be analysed for the dependent variable.

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

Ans. 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.

Dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1.

This following output was achieved when applied drop_first=True while creating dummy variables-

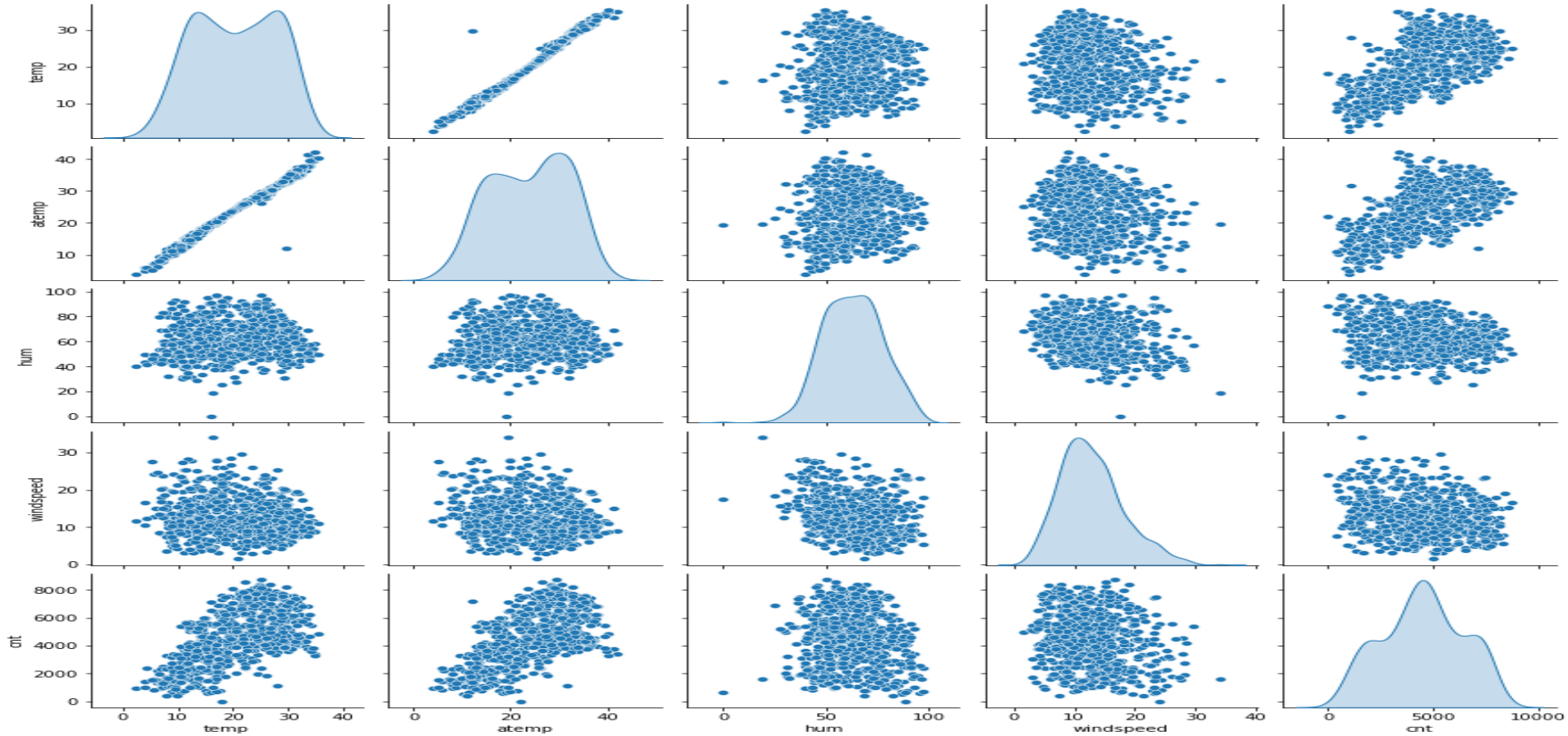


season
1
1
1
1
1

season_2	season_3	...
0	0	...
0	0	...
0	0	...
0	0	...
0	0	...

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

Ans.



Looking at the pairplot ,atemp and temp has highest correlation with target variable i.e cnt.

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

Ans. Linear Regression makes certain assumptions about the data and provides predictions based on that.

The assumptions-

- Assumption 1:The Dependent variable and Independent variable must have a linear relationship.
  A simple pairplot of the dataframe can help us see if the Independent variables exhibit linear relationship with the Dependent Variable
- Assumption 2: No Autocorrelation in residuals.
  Use Durbin-Watson Test.
  DW = 2 would be the ideal case here (no autocorrelation)
  0 < DW < 2 -> positive autocorrelation
  2 < DW < 4 -> negative autocorrelation
  statsmodels' linear regression summary gives us the DW value amongst other useful insights.

  In our final model (model 13) we have got Durbin-Watson: 2.026, which seems to be very close to the ideal case.

- Assumption 3: No Heteroskedasticity.
  Residual vs Fitted values plot can tell if Heteroskedasticity is present or not.
  If the plot shows a funnel shape pattern, then we say that Heteroskedasticity is present.
- Assumption 4: No Perfect Multicollinearity.
  Common way to check would be by calculating VIF (Variance Inflation Factor) values.
  If VIF=1, Very Less Multicollinearity

VIF<5, Moderate Multicollinearity

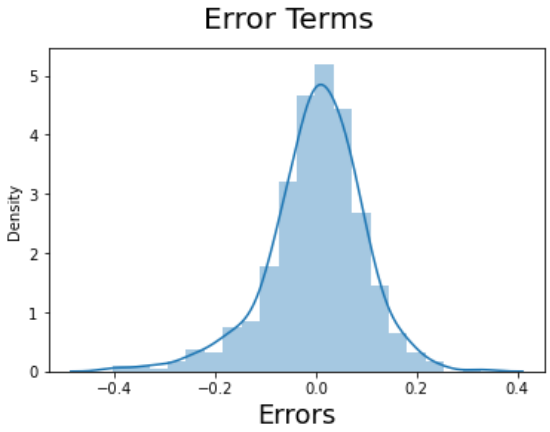
VIF>5 , Extreme Multicollinearity (This is what we have to avoid)

	Features	VIF
1	temp	3.85
2	windspeed	3.43
0	yr	1.98
3	season_2	1.56
6	weathersit_2	1.48
4	season_4	1.35
5	mnth_9	1.19
7	weathersit_3	1.07

So we have moderate multicollinearity.

- Assumption 5: Residuals must be normally distributed.

  Use Distribution plot on the residuals and see if it is normally distributed.



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

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

Ans. As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.5614' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5614 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3022' indicated that a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3022 units.
- Year (yr) - A coefficient value of '0.2309' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2309 units.

So, it's suggested to consider these variables utmost importance while planning, to achieve maximum Booking.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. Linear Regression algorithm can be explained as –

- Machine learning algorithm based on supervised learning.
- It performs a regression task.
- It is used for predictive analysis and shows the relationship between the continuous variables.

- It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).
- It gives a sloped straight line describing the relationship within the variables.

- $y = \theta_1 + \theta_2.x$

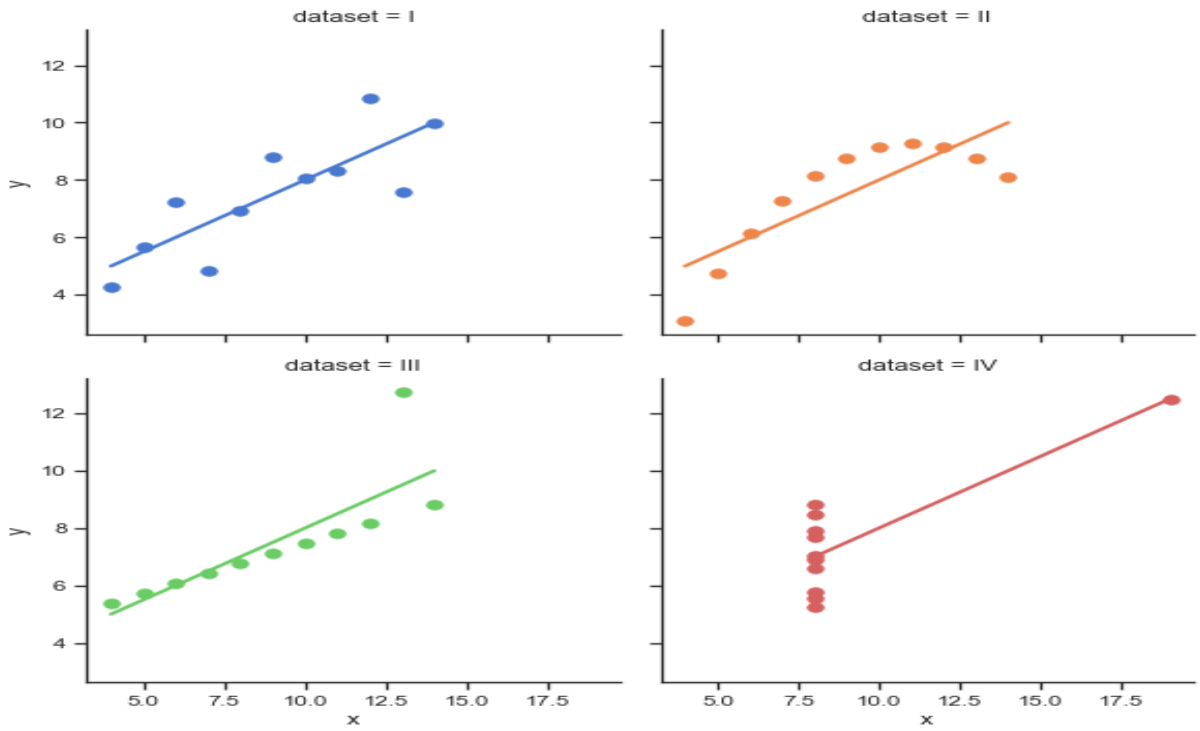
  x: input training data (univariate – one input variable(parameter))
  y: labels to data (supervised learning)
  θ1: intercept
  θ2: coefficient of x
- It is of two types-
  - ❖ Simple linear regression- there is a single input variable (x).
  - ❖ multiple linear regression -there is more than one input variable.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before it is analysed and build the model. These four data sets ,as shown in figure ,have nearly the same

statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

- Purpose of Anscombe's Quartet-
  - ❖ It tells the importance of visualizing data before applying various algorithms to build models because it suggests that the data features must be plotted to see the distribution of the samples that can help to identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)



As we see the real relationships in the datasets start to emerge-

- Dataset I - consists of a set of points that appear to follow a rough linear relationship with some variance.
- Dataset II - it fits a neat curve but doesn't follow a linear relationship.
- Dataset III - looks like a tight linear relationship between x and y, except for one large outlier.
- Dataset IV -looks like x remains constant, except for one outlier as well.

3. What is Pearson’s R?

Ans. Pearson’s R –

- Also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. It has a numerical value that lies between -1.0 and +1.0.
- It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
- It is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
- Named after Karl Pearson.
- How is the Correlation coefficient calculated?

  Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age. So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r.

  There are certain requirements for Pearson's Correlation Coefficient:

  - ❖ Scale of measurement should be interval or ratio
  - ❖ Variables should be approximately normally distributed
  - ❖ The association should be linear
  - ❖ There should be no outliers in the data

The formula given is:

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Ans. Scaling -

It means that one is transforming data so that it fits within a specific scale, like 0-100 or 0-1.

- Example- look at the prices of some products in both Rupees and US Dollars. One US Dollar is worth about approximately 75 rupees . But if you don't scale your prices methods like SVM or KNN will consider a difference in price of 1 Rupees as important as a difference of 1 US Dollar. This clearly doesn't fit with our intuitions of the world.

By scaling the variables, one can help compare different variables on equal footing.

Normalized vs. standardized scaling-

- Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans. 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.

Importance of Q-Q plot:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

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