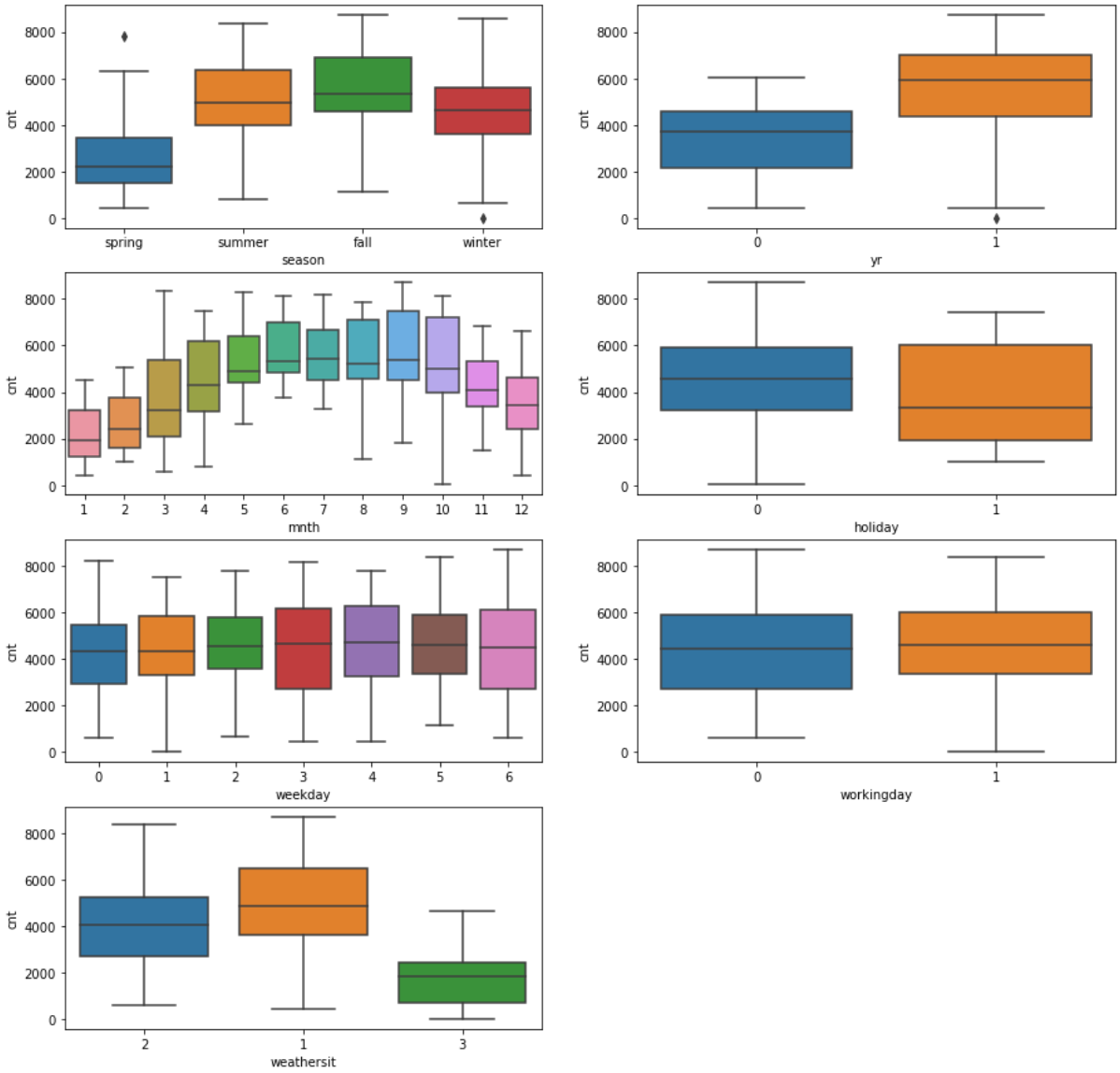
Answers to Assignment-based Subjective Questions

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



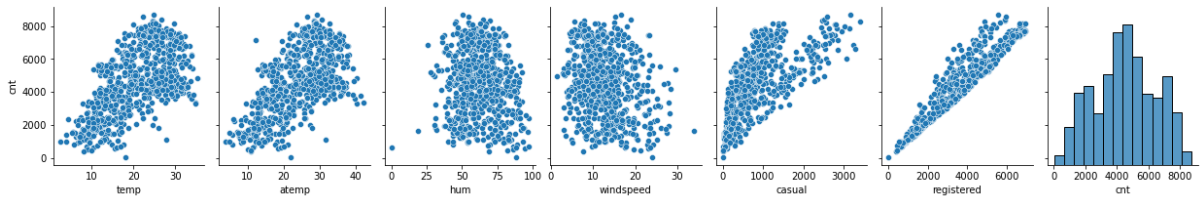
1. Fall is the season with highest average usage of bikes followed by summer, winter and spring
2. There is a clear indication of the growth of bike users from 2018 to 2019.
3. Average usage of the bikes across the months from May to October remains more or less the same with different IQR. There is a clear pattern of increase in bike usage from the beginning of the year till mid of year and then lowers towards the end of the year. The year end bike usage doesn't come down to the year beginning which is also a clear indication of bike usage across years.
4. Average usage of bikes is less on holidays.
5. Average usage of bikes remains the same across the week days with varying IQR.
6. More average usage of bikes with less IQR during working days.

7. There is more usage of bikes during Clear, Few clouds, Partly cloudy, Partly cloudy(Weather Situation 1). People don't prefer to ride a bike during rains and snow.

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

A categorical variable with **k** categories can be encoded with **k-1** variables. Having the information of **k-1** variables can generate the information of the k^{th} variable. Hence having a redundant column which is related with other k-1 dummy variables results in multicollinearity issue. This goes against one of the assumption of the multiple linear regression of no multicollinearity. Hence it is important to drop a category while creating dummy variables using get_dummies function of the pandas with an argument of **drop_first=True,** which otherwise is False and creates **k** dummy variables.

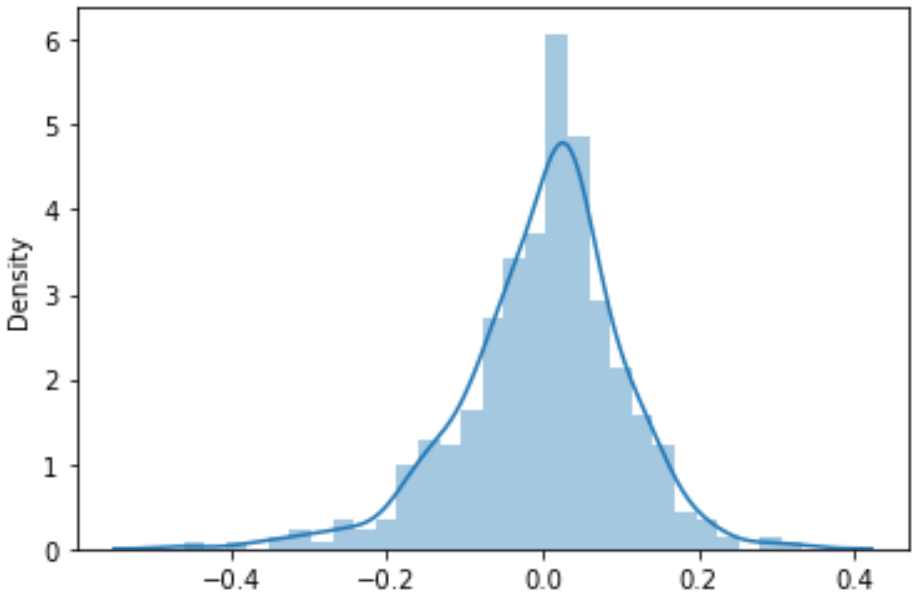
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



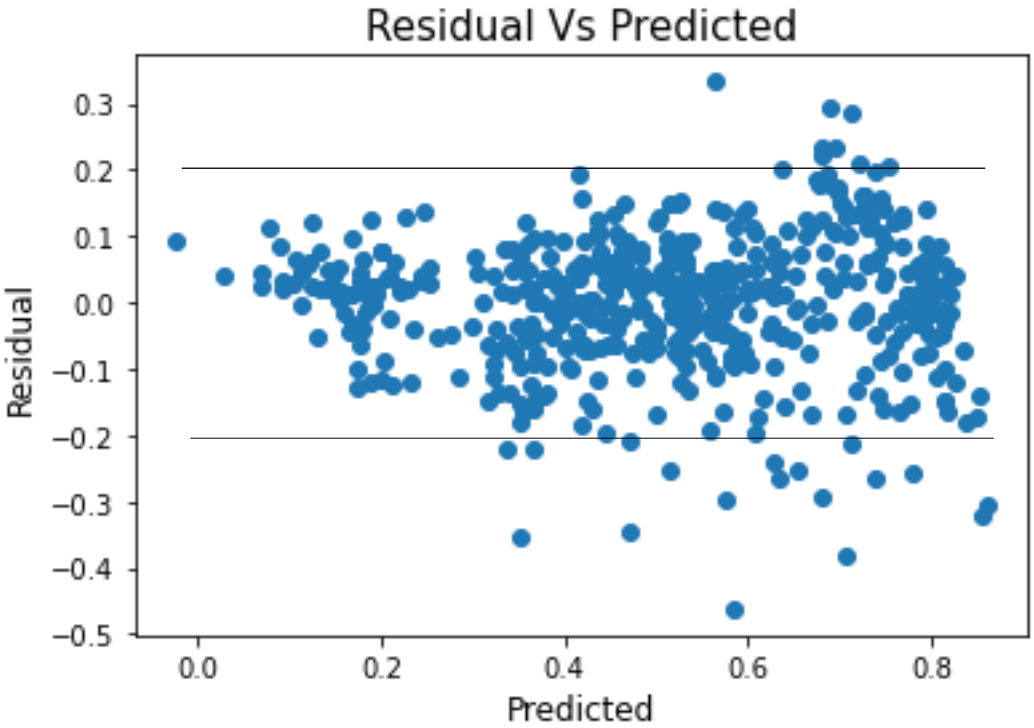
The variables registered has the highest correlation with the cnt which is the dependent variable. This is natural as cnt is derived from casual and registered. If we leave those two variables, then temp and atemp are the next highly correlated with cnt.

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

Residual plot



1. Residuals are normally distributed
2. Residuals are centred at 0.
3. Constant variance (Homoscedasticity), except for some outliers, much of the data variance is between -0.2 to +0.2. There is no pattern in this plot and are evenly distributed around 0.



4. No multicollinearity in the data – VIF of all the selected variables are within 5.

	Features	VIF
3	windspeed	3.64
2	temp	3.37
0	yr	2.02
4	season_spring	1.50
5	weathersit_3	1.05
1	holiday	1.03

5. Linearity relationship – This is established from the residual vs predicted plot which is a random scatter and there is no curvy pattern which indicates a non-linear relationship.

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

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2794	0.020	14.253	0.000	0.241	0.318
yr	0.2365	0.009	25.641	0.000	0.218	0.255
holiday	-0.0720	0.029	-2.475	0.014	-0.129	-0.015
temp	0.3843	0.026	14.868	0.000	0.334	0.435
windspeed	-0.1492	0.028	-5.392	0.000	-0.204	-0.095
season_spring	-0.1466	0.014	-10.841	0.000	-0.173	-0.120
weathersit_3	-0.2451	0.027	-8.965	0.000	-0.299	-0.191

Coefficients explains the change in the dependent variable with unit change in the contributing independent variable, keeping all other variables constant.

From the coefficients of the selected variables, temp contributes the highest as it has highest coefficient value(0.3843) followed by year(yr) which has the next highest coefficient value of 0.2365. The above two variables explains the prediction of bike usage in the positive terms. The next best explainer is the negative term i.e, weather situation of Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, which reduces the usage of bikes.

Answers to General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning method to predict a continuous variable. As the name suggests there is a linear relationship between the independent variable(s) with the dependent variable. This algorithm is used for interpolation of the data for predicting the dependent variable on the independent variable(s) inside the range of the data the model was built on and is not for extrapolation. Linear regression models can be classified into 2 types.

Simple Linear Regression – Where number of independent variables is 1.
Multiple Linear Regression – Where the number of independent variables is >1.

Linear regression in machine learning is used for modelling in which the machine learns from the data and develops a functional relationship between the independent variable(s) and dependent variable.

The equation of the regression line for a simple linear regression is

$$Y = \beta_0 + \beta_1 X$$

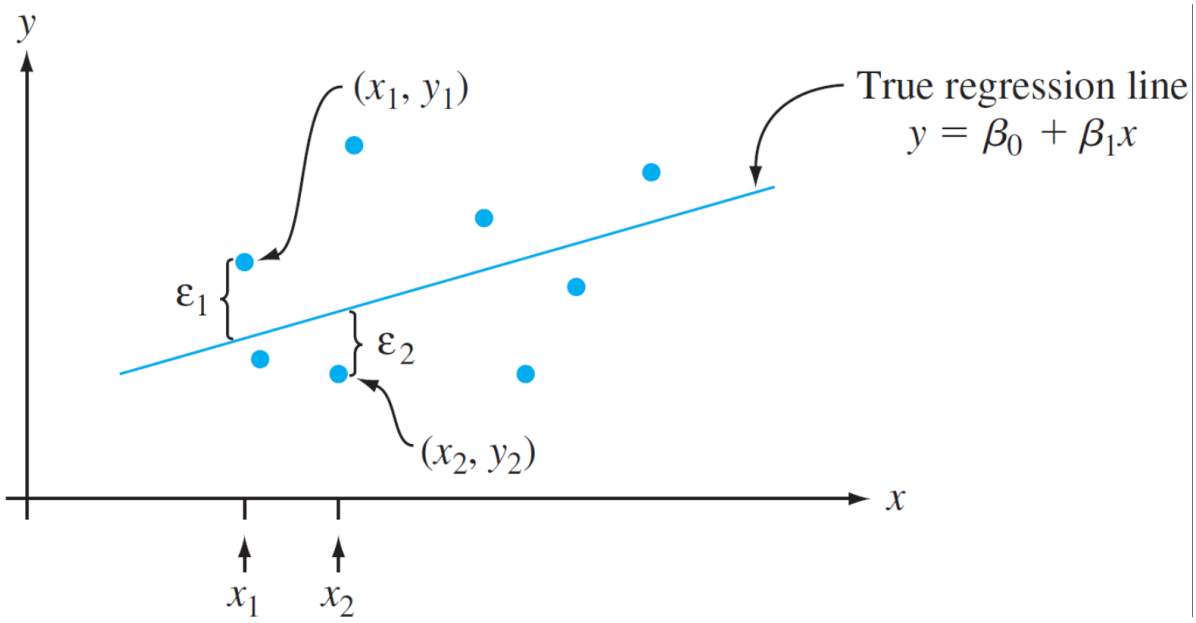
where X is an independent variable and Y is a dependent variable.

The equation for regression line in Multiple Linear Regression is

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p$$

Where $X_1, X_2, ... X_p$ are the independent variables and Y is the dependent variable.

Fitting the best line for the given data involves reducing the residuals, which is the distance between the data and the predicted value from that of a proposed regression line.



Residual Sum of Squares(RSS) =

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$$

The best fit line is the one which has minimum RSS which is the cost function.

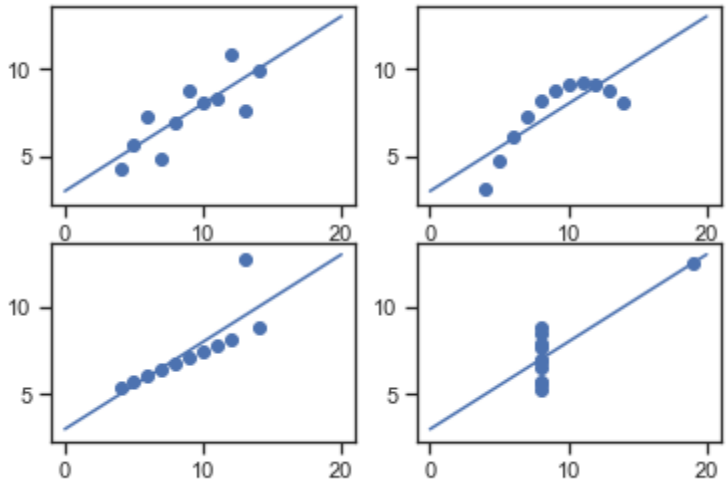
There are 2 methods in which the cost functions can be minimised

1. Closed form selection
2. Iterative method.

Closed form selection involves differentiating the cost function and equating it to zero. The co-efficients that we arrive at after solution is that for the best fit line. The iterative method like gradient descent in which the process starts with some coefficients and the move towards finding better coefficients in an iterative manner.

2. Explain the Anscombe's quartet in detail.

It consists of 4 data sets which are identical in statistical properties. Each of the data set contains 11 data points and was constructed in the year 1973 by statistician Francis Anscombe. This stresses the importance of data visualisation before analysing it and also the effects of the outliers on the statistical properties.

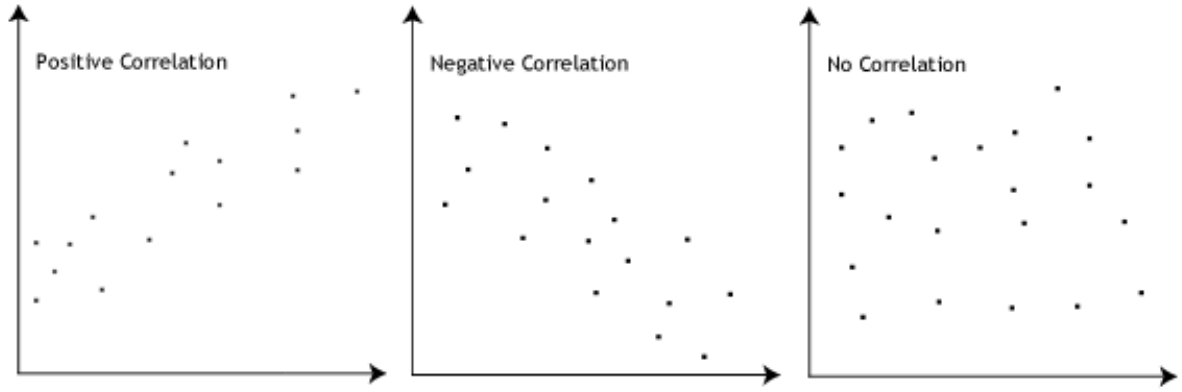


- The data set 1 consists of a set of (x,y) points that represent a linear relationship with some variance.
- The data set 2 shows a curve shape but doesn't show a linear relationship.
- The data set 3 looks like a tight linear relationship between *x* and *y*, except for one **large outlier**.
- The data set 4 looks like the value of *x* remains constant, except for one **outlier** as well.

The data set 3 and 4 is an example of the outliers misleading the statistical characteristics of the data set. It is only with the help of visualisation that we can make out about the data. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

Pearson's R is a measure of the strength of association of 2 variables. It is denoted by **r**. The possible values of **r** is from -1 to +1. An **r** value of > 0 indicates a positive association, that is if one variable increases the other also increases. An **r** value of <0 indicates negative association in which when one variable increases the other decreases. **r** value of 0 indicates that there is no relationship between the two variables. The 3 possible scenarios are shown in the below figure.



Pearson Correlation Coefficient (r) is calculated using the below formula

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

Where x and y are the variables for which we are trying to find the strength of association using **r.**

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

Scaling is a process in the data preparation in which the independent variables are scaled for the purpose of faster computation and also for the ease of interpretation of the coefficients of the linear regression. Sometimes the dataset we have has variables with very different magnitudes, units and range. To bring them all into the same level of magnitude we have to do scaling. It only affects the coefficients and doesn't effects other statistics like t-statistic, F-statistic, p-values, R-squared, etc. There are 2 types of scaling.

1. Normalisation/Min-Max Scaling
2. Standardisation scaling

In Normalisation, all the data are brought on to the range of 0 to 1. Using the below formula.

$$(x - \min(x)) / (\max(x) - \min(x))$$

Standardisation replaces the values with their Z score. It bring all the data onto a standard normal distribution. It is calculated using the below formula.

$$(x - \text{mean}(x)) / \text{sd}(x)$$

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

The value of VIF as infinity indicates that there is multicollinearity problem and there is perfect correlation between the 2 independent variables. If the independent variables are perfectly collinear, then our model becomes singular and it would not be possible to uniquely identify the model coefficients mathematically. In this case we need to drop one of the variables. VIF is defined by the below formula.

$$VIF = 1 / (1 - R^2)$$

R^2 - R-Squared value.

When there is perfect correlation, $R^2 = 1$ and VIF becomes infinity.

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

In data science or in statistics its important to know the kind of distribution. In order to find if a distribution is a normal Uniform distribution, Exponential distribution, Binomial distribution, etc we use Quantile-Quantile plots.

It also helps

1. In finding if 2 populations are of same distribution.
2. Skewness of distribution.
3. In regression in finding if error terms is normally distributed which is one of the assumptions.

In Q-Q plot, the theoretical quantile values are plotted with sample quantiles. If the data sets that are compared are from same type of distribution, we get a rough straight line. The below is an example of normal distribution.