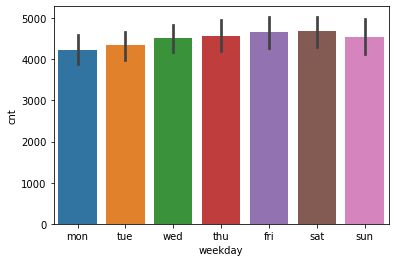
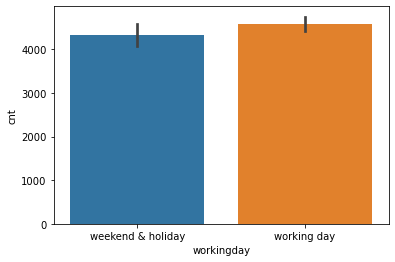
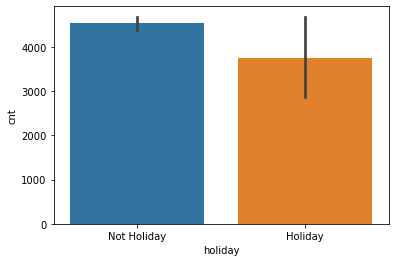
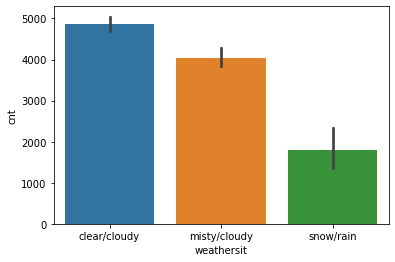
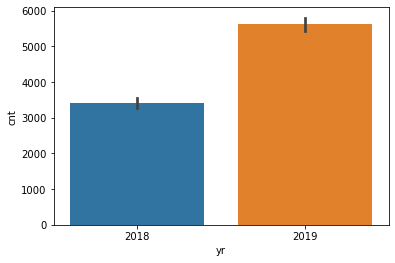
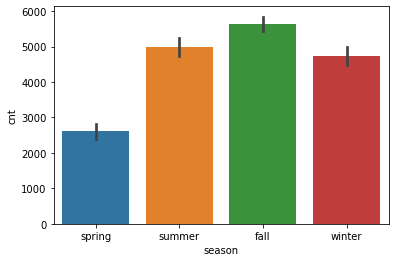
# Assignment-based Subjective Questions

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



Answer: From the exploratory data analysis on categorical variables, a bar plot was plotted taking into account the various categories in x axis and the mean of dependent variable on y axis. The results are as follows:

-There are low customers sharing bikes during spring season

-There are higher customers (nearly doubled) sharing bikes in 2019 as compared to 2018

-People share bikes lesser in holidays as compared to non-holidays

-People have taken more bikes in clear and cloudy weather as compared to other weather conditions

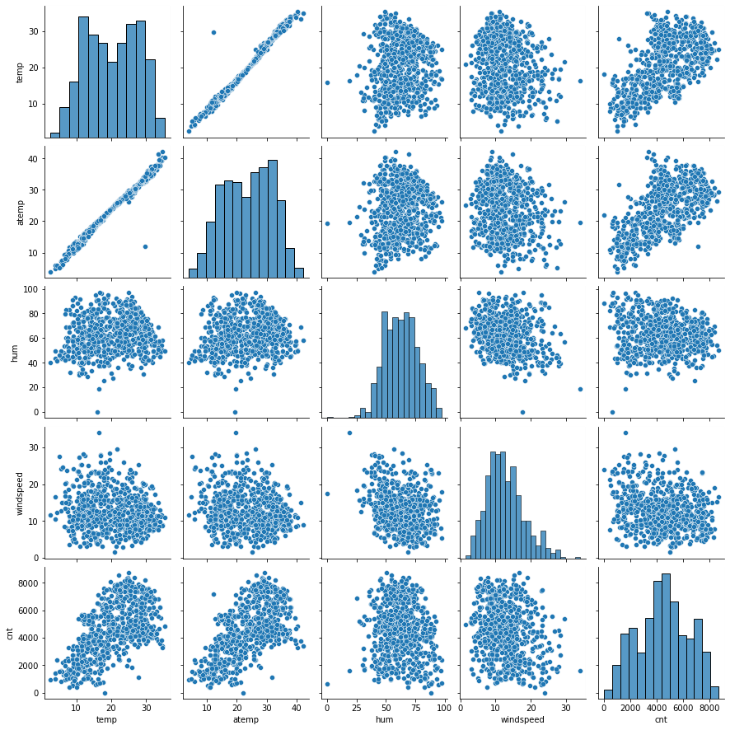
-There is no significant effect of weekday or working day on the count of customers sharing the bikes

1. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark)

Answer: This is because we can depict the categorical variable with ‘n’ categories into (n-1) dummy variables. We can drop the first or last dummy variable, all the information can be depicted in n-1 features. The first column will be redundant and this step helps us reduce correlation amongst dummy variables.

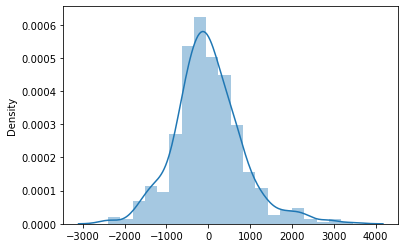
1. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Looking at the pair-plot amongst the numerical variables, actual temperature has the highest correlation with target variable, but it is closely followed by temperature, however as temperature and actual temperature are highly correlated, we can say actual temperature has the highest significant correlation with target variable.



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

Answer: In the multi linear regression model which was built on the training data, we used the model to predict the target variable for training features, we find the residuals which is the difference between actual target variable and predicted target variable and plot a distribution plot of the residuals which shows a normal distribution with mean at 0



1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

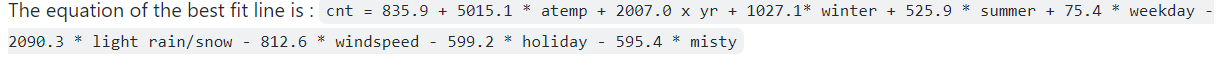
Answer: The equation of the best fit line shows the coefficients of the independent variables which explains its effect on the target variable which is the demand of the shared bikes. The top 3 features contributing significantly are:

1. Actual Temperature (Positive Effect)

2. Weather condition: light rain/ snow (Negative Effect)

3. Year (Positive Effect)

This is derived from:



# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

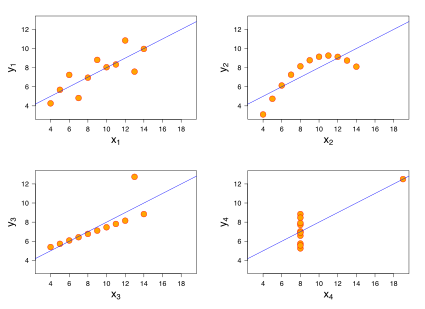
Answer: Linear regression algorithm is an algorithm which tries to find a relation between independent input variables and a dependent output variable. It is used to predict output on numeric variables. The linear regression finds how the value of the dependent variable is changing according to the value of the independent variable. It tries to fit a straight line which best represents the output variable in terms of input features.

It tries to minimize the mean square error which is the error between the actual data points and the predictor line, the best fit line is represented as the line which has the lowest mean squared error and highest R-squared value which is a metric used to measure the fit of the line. The linear regression assumptions are that the output variables should be linearly dependent on each input variable.

The model which is fit should have random errors and the residuals should be normally distributed with a mean of 0, they should be homoscedasticity (equal variance). There are two types of linear regression, a simple linear regression which has one predictor while multiple linear regression (MLR) has multiple predictors and in MLR, the predictors should not be multicollinear which can be evaluated using metrics like variance inflation factor.

1. Explain the Anscombe’s quartet in detail. (3 marks)

Answer: The Anscombe’s quartet is a special set of 4 datasets(x,y) which is used to indicate the significance of using graphs to understand data distributions. Anscombe had prepared four datasets which have different distributions but they all had the same summary statistics, all the four sets of (x,y) data had same mean, median and standard deviation but when plotted on a scatter plot, they all showed completely different story. This shows that we should not only use summary statistics to compare datasets but also understand their graphical distribution to make a valid comparison. Surprisingly, they all when fit on a linear regression line show the same R-squared which is 0.67 and have the same correlation of 0.816



Source: Wikipedia

1. What is Pearson’s R? (3 marks)

Answer:

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
2. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

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

(3 marks)