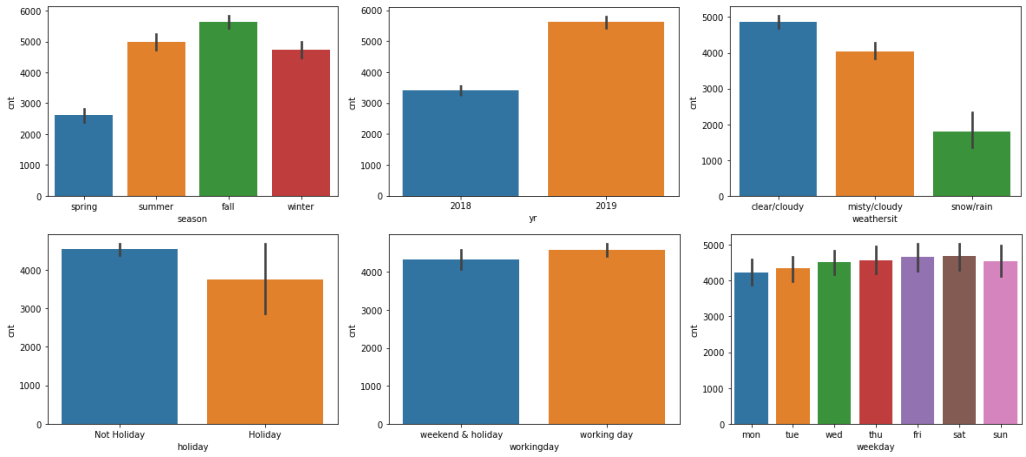
Assignment-based Subjective Questions

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

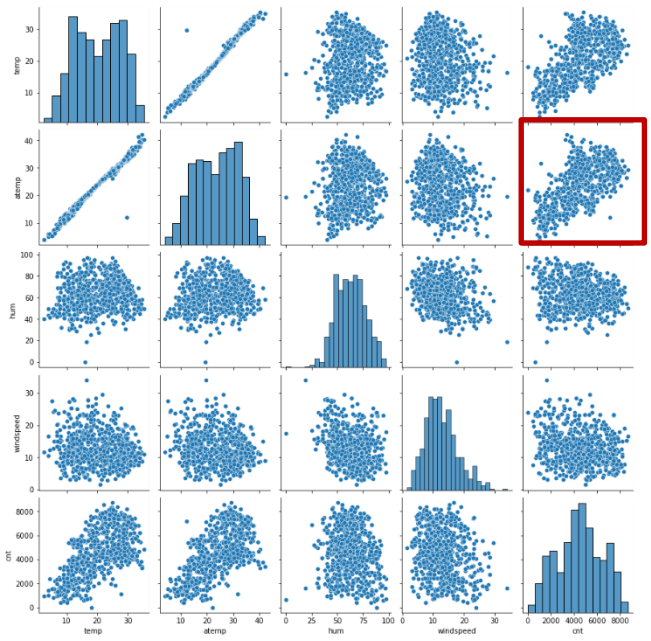
-There is no significant effect of weekday or working day on the count of customers sharing the bikes

2. 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.

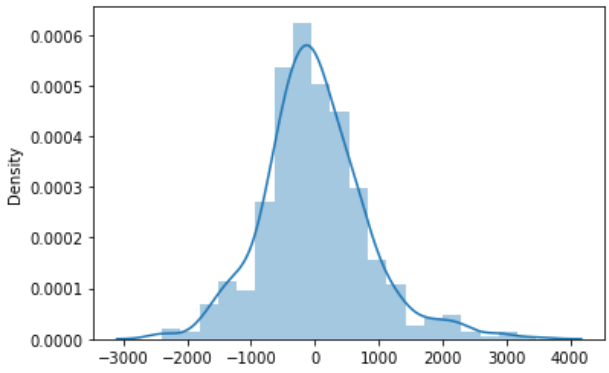
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

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.



4. 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



5. 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:

The equation of the best fit line is : `cnt = 835.9 + 5015.1 * atemp + 2007.0 x yr + 1027.1* winter + 525.9 * summer + 75.4 * weekday - 2090.3 * light rain/snow - 812.6 * windspeed - 599.2 * holiday - 595.4 * misty`

General Subjective Questions

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

Answer: Linear regression 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's represented as below:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

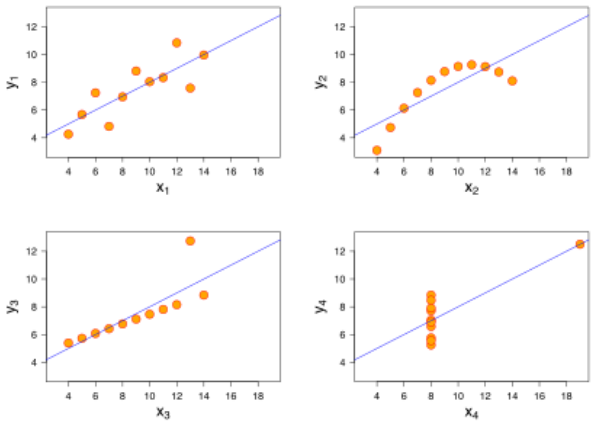
Where beta are the coefficients and X are the independent variables used to predict dependent variable Y

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.

2. 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 of Image: Wikipedia

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

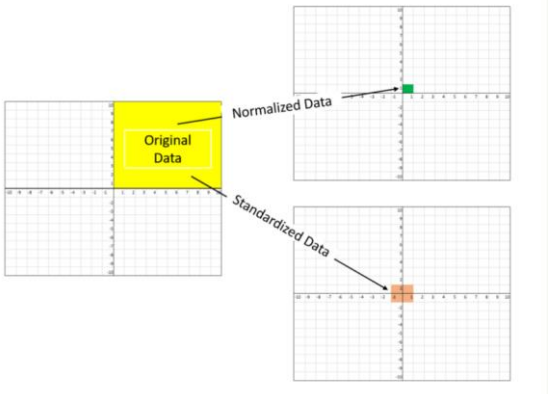
Answer: Pearson's R is a type of correlation coefficient. It is used to measure the relationship between two variables. It varies between -1 and +1, -1 indicates that there is a strong negative relation between the variables, meaning when one variable increases the other variable decreases, 0 correlation indicates the one variable doesn't affect the other variable and +1 indicates the strong positive relation between the variables. It can also be visualized as the slope of the line representing the relation between two variables on the scatter plot. It also represents the tightness of the relation between the two variables. It is calculated by the formula:
Correlation Coefficient= Covariance (x, y)/Sx*Sy where Sx & Sy are standard deviations of the respective variables. The expanded formula is given as follows:

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

Source of Image: https://byjus.com/correlation-coefficient-formula/

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

Answer: Scaling is a technique used to help the model learn your data faster. The machine learning algorithms works on reducing the distance between the data points to optimize the model. It is essential that we try to bring the data points as close as possible so that the differences become lower and it is easier to minimize the distances. It is performed to reduce the computation time of your optimization algorithm. Also, during some real-life problems, you will have different features with different units, for example, if you have 2 features, 10g weight and 10crore rupees, your model will treat both as same values, to make the two features comparable, we need some scaling technique which identifies the distribution in that feature and tries to assign a value which is comparable with other features. If it is not performed features with higher values get more importance in the algorithm which is not preferred. Normalized Scaling tries to squeeze your data within a range typically between [-1, 1]. It is done as follows for a variable x: $X_{normalized} = (x - x_{min}) / (x_{max} - x_{min})$. On the other hand, standardized scaling assumes the distribution and tries to fit a normal distribution of the data with mean 0 and standard deviation as 1. It is done as follows for variable x with mean u and std deviation s: $x_{standardized} = (x-u)/s$
The difference between normalized and standard scaling can be seen below:



Source of Image: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

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

Answer: As we can see from the formula that VIF is calculated as follows:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

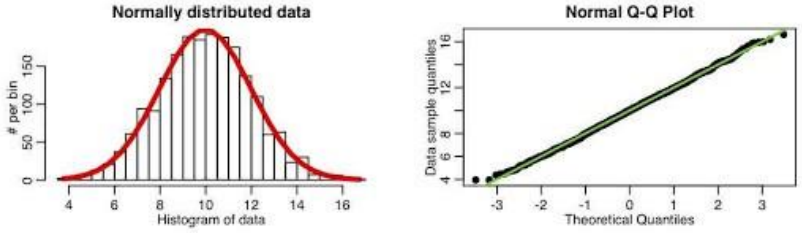
Source of Image: Wikipedia

Where R_i^2 is the coefficient of determination of the linear regression fit for i^{th} feature

The value of VIF for a feature can be infinite if the R_i^2 is equal to 1. This indicates that there is multicollinearity and the feature has perfect correlation with other variables and is linearly related with the other features in the model. It also means the feature is a linear combination of other features and having that feature for the linear regression model would be insignificant, because in linear regression the features have to be independent and only the target variable should be dependent. It can also happen if one feature is scaled up value of another variable, for example weight in grams and weight in kilograms are two features it will make the VIF to infinite

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

Answer: It's known as Quantile-Quantile plot and is used to analyze and compare two probability distributions by plotting their quantiles against each other. If the 2 distributions are exactly equal, then all the points will lie on a straight line with slope as 1. It is also powerful enough to depict the type of distribution whether it is Gaussian, Uniform, Exponential or Pareto distribution. It is mostly used to identify whether a distribution is normal or not, because many events in nature are normally distributed, so if a perfect straight line is formed, we can say it's a normal distribution. Also, skewness (a measure of asymmetry) can be checked with this plot, if its curved and tilted out at top, its right skewed or if it's tilted in towards bottom, its left skewed. The theoretical quantiles or normal distribution quantiles are plotted on x-axis and on y-axis. The below example shows how normality can be checked using a Q-Q plot:



Source: https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

In linear regression, we generally check on our model, whether the assumptions of the linear regression are satisfied or not, one of the assumptions was that the residuals which are the error terms between actual and predicted data are normally distributed or not, to understand it quantitatively, we can check how good is the fit of the points on the Q-Q plot with the trend line, if most of the points lie near the line, we can say the residuals are normally distributed thereby satisfying the assumption of linear regression.