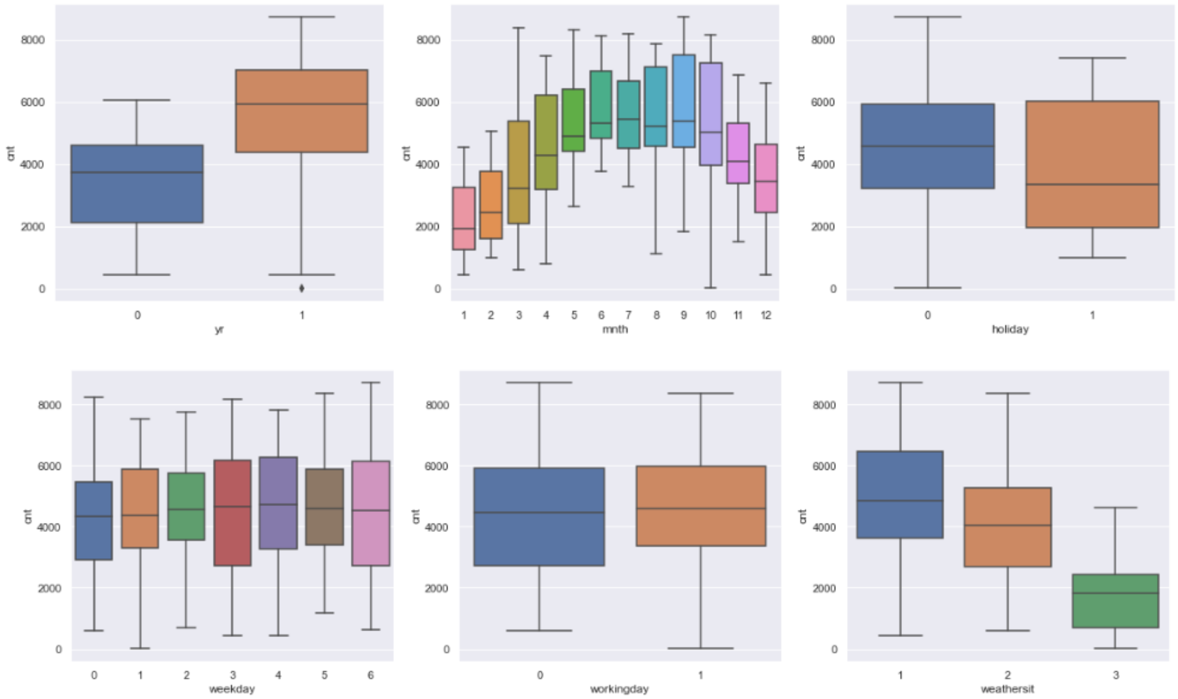
Assignment-based Subjective Questions

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



Following were the observations from the categorical variables. These were derived by looking at the box plots drawn between the categorical variables and the target variable cnt.

- The bikes rental seems to grow in 2019, as compared to 2018
- The bike rental seems to be higher in months May till October
- The bike rentals seems to be higher on non-holidays
- Weekday doesnt seem to have much impact on the demand of bikes
- Working day doesnt seem to have much impact on the demand of bikes
- Weather situation seems to impact the demand of bikes, with less demand in category 3 and 4

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

Lets say we have 4 types of values in a categorical variable column e.g. Summer, Winter, Rain, Fall season. Now to create dummy variable for this variable, we can have 4 columns and the respective column will have a zero or one to depict a certain season. E.g. for Summer season, Summer column will have 1 and rest other columns (Winter, Rain, Fall) will have zeros. However, if 3 columns are all zeroes, then it can be construed that the 4th column is 1. So for a categorical variable with 4 values, we can create dummy variables with 3 columns only.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

While creating dummy variables using pandas, drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Variables temp and atemp have the highest correleation with target variable cnt. But both these variables are highly correlated to each other, hence we have dropped atemp after looking at VIF. Thus temp is the highest correlated variable with target variable cnt.

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

   - The dataset had no missing values and hence there was no need to impute data. This was validated by looking at the dataframe using bikeDF.info().
   - The continuous numerical variable are positively correlated with target variable cnt. Hence Multiple Linear Regression could be applied on this dataset. This was validated by looking at the pairplot of the continuous numerical variables.
   - There was correlation between categorical values and target variable cnt as well. This was validated by looking at the boxplots between categorical variables and target variable.
   - There were no variables that contained yes/ no values, hence it was not needed to convert them to 1/0. This was validated by looking at the value_counts() for each of the categorical variables.
   - There is no Multicollinearity between independent variables. This has been validated by looking at the VIF scores. When the model was rendered the first time, there were 7 independent variables that had VIF of Inf meaning that they were perfectly collinear with some other independent variables. However, after looking at combination of p-values and VIF scores, some variables were dropped and model was re-rendered. This way, the collinearity between independent variables was removed.
   - Residuals were normally distributed. Residuals were calculated and a distplot was created to observe that residuals are normally distributed.
   - The model is accurate on both training and test data set. This was verified by looking at the R squared values on both data sets. We can see that the R Squared on the training set was 0.785 (78.5 %), whereas R squared on test set is 0.772 (77.2 %), which means that 77.2% of the variation on test data set can be explained by this linear regression model

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

   The final model is as given below:

   cnt = 0.3333 + (0.2285 x yr) - (0.0692 x holiday) + (0.0229 x workingday) + (0.5093 x temp) - (0.2582 x hum) - (0.2490 x windspeed) + (0.0722 x "2") - (0.0135 x "1") - (0.0076 x "2") - (0.0294 x "2") - (0.0505 x "2") + (0.0300 x "8") + (0.1171 x "9") + (0.1308 x "10")

Based on this model, the top 3 features contributing significantly towards explaining the demand of shared bikes are temperature, humidity and year.

General Subjective Questions

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

Linear regression provides a means to explain the relationship between magnitude of one variable and that of a second e.g. if X increases, does Y also increase or decrease simultaneously. While the same can be explained via Correlation, the difference is that linear regression helps quantify the strength of association between 2 variables.

Linear regression can be either Simple Linear Regression or Multiple Linear Regression.
Simple Linear Regression estimates how much Y will change when X changes by a certain amount. A simple linear regression can be represented as Y = mX + c.
In this equation, “c” is known as the intercept and “m” signifies the slope.
Multiple Linear Regression estimates how much Y will change when there are multiple predictors. A Multiple linear regression can be represented as Y = c + m1X1 + m2X2 + ….+ mnXn.

Linear regression has many practical uses:

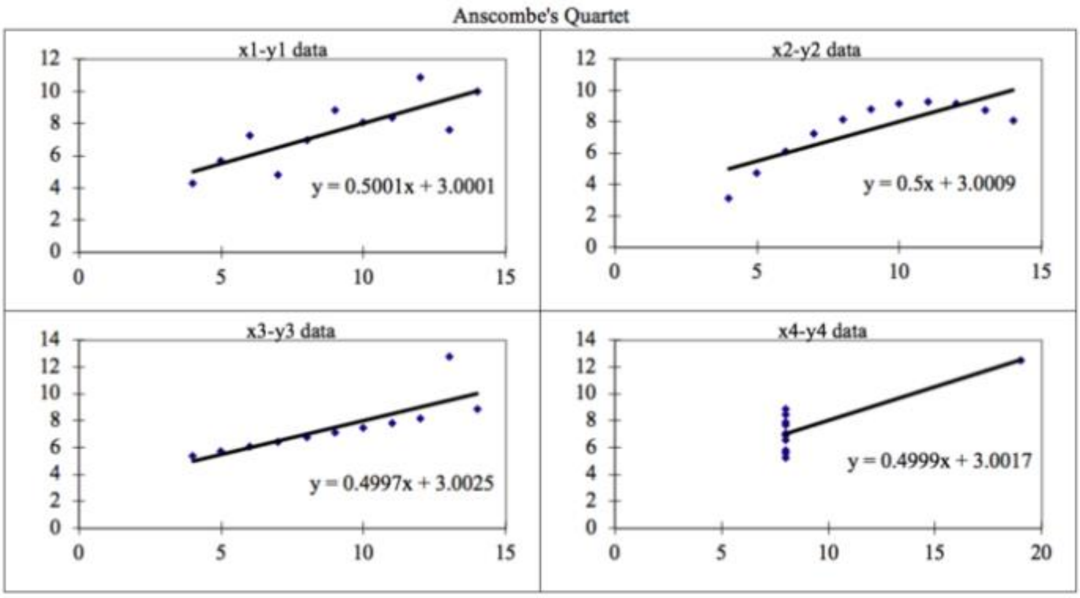
- Predicting
- Forecasting
- Error reduction

Advantage of Linear regression is that estimation procedure is simple. However, the main limitation of linear regression is the assumption of linearity between the dependent variables and the independent variables.

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

Anscombe's quartet was propounded in 1973 by Francis Anscombe to demonstrate the importance of analysing the data using graphs. It takes into consideration four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business. But there’s a danger in relying only on summary statistics and ignoring the overall distribution. This tells us about the importance of visualizing the data before building a model.



If we were to consider 4 datasets, they may have similar averages, variances, correlation between X and Y and even a Linear Regression equation that explains the dataset. However, when these 4 datasets are looked at visually, 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 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 outlier. Dataset IV looks like x remains constant, except for one outlier as well.

Hence summary statistics shouldn’t be used in isolation but in conjunction with visualization.

3. What is Pearson’s R? (3 marks)

Pearson’s R is also known as the Pearson correlation coefficient (PCC), or the Pearson product-moment correlation coefficient (PPMCC), is a measure of linear correlation between two sets of data.

It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

It is denoted by the formula as given below and it varies between -1 and +1:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Since Pearson's r is a bivariate statistical model that analyzes two variables, it should not be used to test an attributive research hypothesis because an attributive research hypothesis only includes one variable. Another issue with Pearson’s r is that it is not able to tell the difference between dependent variables and independent variables.

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

Scaling is a step carried out during data pre-processing on independent variables to normalize the data within a particular range (typically between 0 and 1 for the optimization to become faster). Data for numerical variables can be on different scales ranging from small values to extremely high values. On the other hand, categorical variables will have values as only 0 or 1. Because of the variables being on different scales, their coefficients will also be very different thus incorrectly suggesting that one variable is a stronger predictor than the other variable.

Hence it becomes imperative to have all the variables on the same scale, so that the model is easily interpretable. When we are doing Simple Linear Regression, scaling is not very important. However, its extremely important in Multiple Linear Regression.

There are 2 common ways of scaling:

- Min-Max scaling (also known as Normalization scaling)
- Standardization scaling

<u>Min-Max (Normalization) Scaling:</u>
This technique compresses all the data between 0 and 1. It is calculated by the formula: (x – xmin)/ (xmax – xmin). The maximum value is mapped to 1 and minimum value is mapped to zero. This technique takes care of the outliers (maximum value is mapped as 1).

<u>Standardization Scaling:</u>
This technique compresses the data in such a way that the mean is zero and standard deviation is 1. It is calculated by the formula:
(x – mu)/ sigma. The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier).

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

If the VIF of a certain variable is infinite, then it means that it is perfectly correlated with another variable. For the perfect correlation between two independent variables, the R2 =1, which leads to 1/(1-R2) infinity. We drop one such variable to take care of the multi-collinearity.

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

A standard normal deviation is one in which the units on the x-axis are expressed in terms of standard deviation away from the mean. A QQ Plot is used to visually determine how close a sample is to a specified distribution.

The QQ plot orders the z-scores from low to high and plots each value's z-scores on the y-axis.

The Q-Q plot (quantile-quantile plot) is used to help assess if a sample comes from a known distribution such as a normal distribution. For regression, when checking if the data in this sample is normally distributed, we can use a Normal Q-Q plot to test that assumption.