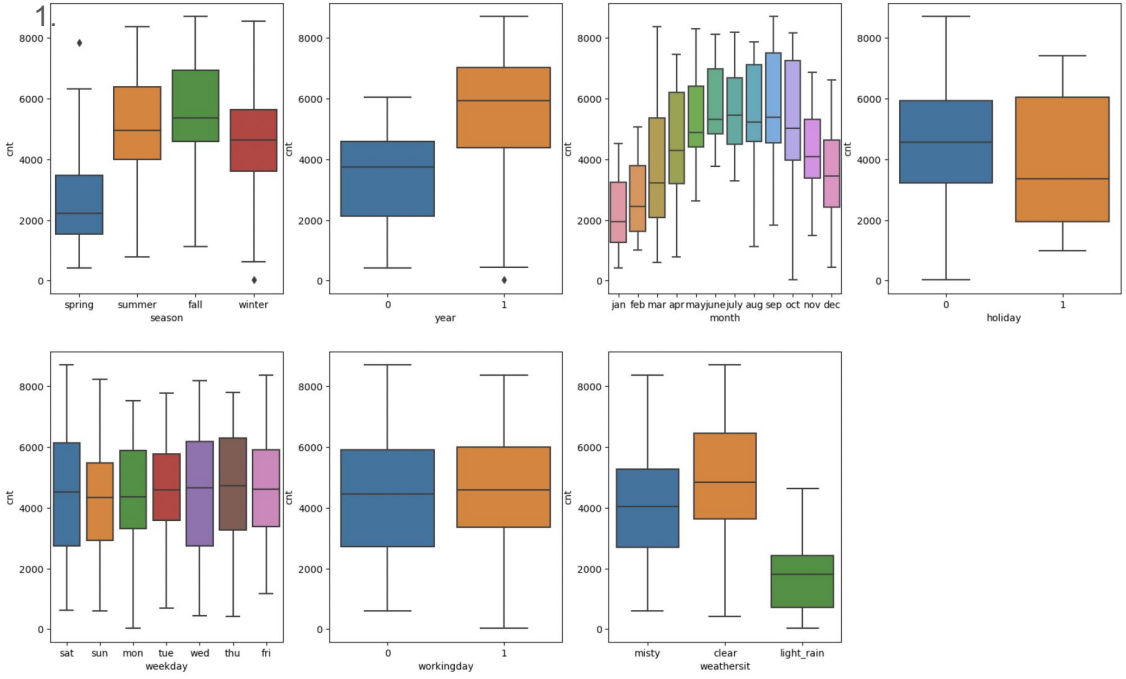
Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

I used box-plot along with subplots to visualize and analyse the categorical variables and how they affect the dependant variable (‘cnt’).



I can infer the following effects of categorical variables on the depend variable from my analysis.

1. Weather has a clear effect on the demand for shared bikes. Summer and Fall seasons seem to attract more demand for shared bikes, with fall attracting the most demand.
2. I can clearly see an increase in shared bikes' demand in 2019 as compared to the previous year 2018. This clearly shows that the business for the shared bikes is gaining more popularity and is a good sign for the business.
3. There is more demand in the months from may through nov. These months are also the seasons of summer and fall in america and thus confirms the first inference noted above.
4. There seems to be more bookings in the holidays than in the non-holidays, which we can attribute to people wanting to spend more time outside during holidays
5. Compared to Tuesday, Wednesday and Thursday, there are good bookings on Friday, Saturday, Sunday and Monday. We can infer that there are more bookings on the the weekend and a day before and after weekend.
6. Weekday also has an affect, there is more bookings on the weekend, which also aligns with the above inference.
7. Clear weather attracts more bike sharing demands

Q2. Why is it important to use drop_first=True during dummy variable creation?

The get_dummies() from pandas library converts categorical variables into dummy variables. Each categorical variable will be converted to as many dummy variables as the number of values in the categorical variable, with each dummy variable containing value 0 or 1. When we have n dummy variables of 0/1, we can live with n-1 variables, as the nth variable can be represented by a all 0 combination of all other n-1 variables. Having lesser variable is desired in model building.

This is where drop_First comes into picture. drop_First helps in removing this additional dummy variable. Setting drop_First = True when using get_dummies() will remove the first dummy variable.

Syntax: drop_First: bool

Example usage from my assignment: *df_season=pd.get_dummies(df_bike.season,drop_first=True)*

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variable ‘temp’ has the highest correlation with the target variable. One more note here is that the variable ‘atemp’ also has similar correlation with the target variable and from the heatmap we can see that they both seem to have same correlation coefficient(0.63) with target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Residuals (error terms) follow a normal distribution
2. The mean of the residuals is zero
3. The residuals have a common variance (Homoscedasticity)

I validated all these by calculating the residuals and plotting its distribution.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Based on the coefficients of my final model, the top 3 features contributing significantly towards explaining the demand of shared bikes are:

1. Year
2. Season
3. Weathersit

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

In layman terms, Linear Regression is a machine learning technique to learn from historical data to establish a linear relationship between input and output so that it can be represented by a formula (y = mx + c) which can be used to predict future outputs.

In advance terms, Linear Regression is a supervised machine learning algorithm which helps us establish a linear relationship between the independent variables (also called predictor variables) and the dependant variable (also called the target variable) so that we can build a model with that relationship and use it to predict the outcome of the future events.

Linear Regression is used when the target variable is known and is a continuous variable.

Simple Linear Regression:

When there is only one independent variable, it is called a Simple Linear Regression. It can be described by a formula

Y = mX + C

Where X is the independent variable and Y is the dependant variable.

C is the intercept and m is the slope of the line which also represents how much Y changes with a X changes a unit.

There are two kinds of linear relationships.

- Positive Linear Relationship
  - When m is positive, it means that Y is having a positive relationship with X; when X increases Y increases and when X decreases Y also decreases.
- Negative Linear Relationship
  - When m is negative, it means that Y is having a negative relationship with X; in other words, when X increases Y decreases and when X decreases Y increases.

Multiple Linear Regression:

When there are multiple independent variables, it is called a Multiple Linear Regression. The formula for multiple linear regression is similar to the simple linear regression with a small change that instead of having beta (slope) for only one variable, you will now have betas for all variables. The formula can be simply given as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Where Y is the dependent variable and Y is the dependant variable.

C is the intercept and m is the slope of the line which also represents how much Y changes with a X changes a unit.

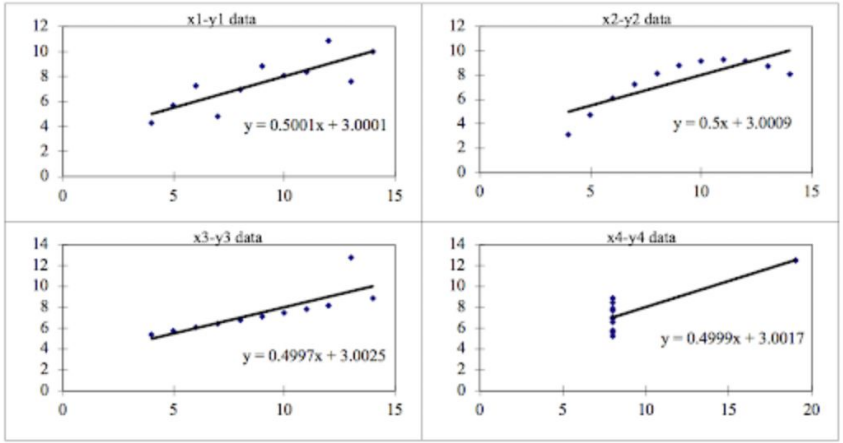
We can build a linear regression model using python libraries sklearn or statsmodels.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. It is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. These four data sets have nearly the same statistical observations, which provide the same statistical 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.
The below is the 4 data sets, we can see that the statistical information are approximately similar for all 4 data sets.

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:

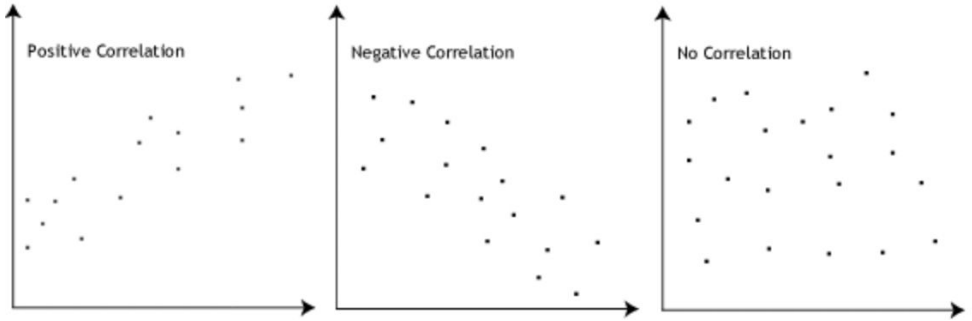


We can describe the four data sets as:
Data Set 1 seems to fit the linear regression model pretty well.
Data Set 2 cannot fit the linear regression model because the data is non-linear.
Data Set 3 shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4 shows the outliers involved in the data set, which also cannot be handled by the linear regression model.
Anscombe's quartet helps us to understand the importance of data visualization before attempting to interpret and model the data or implement a machine learning algorithm.

Q3. What is Pearson’s R?

Pearson's r (also referred as Pearson’s correlation coefficient) is a numerical measure of the strength of the linear association between two variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What is feature scaling?

When you have a lot of independent variables in a model, some of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. This is where feature scaling comes into picture.

Feature Scaling is a technique to standardise the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Feature scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

Why is it performed?

1. It guarantees that all features are on a comparable scale and have comparable ranges. This makes sure that each feature contributes equally to the learning process by scaling the features.
2. It improves algorithm performance improvement. When the features are scaled, machine learning methods generally perform better or converge more quickly.
3. It ensures that each feature is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions.

Normalised vs Standardised Scaling:

In Normalised scaling (also called MinMax Scaling), the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

Formula: $$x = \frac{x - min(x)}{max(x) - min(x)}$$

In Standardised scaling, the variables are scaled in such a way that their mean is zero and standard deviation is one.

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

<u>Differences:</u>

Normalised scaling rescales the values between 0 and 1, whereas standardised rescaling centers the data around the mean and scales to a standard deviation of 1

Normalised scaling is sensitive to outliers whereas standardised rescaling is less sensitive to outliers.

Normalised scaling retains the shape of the original distribution whereas standardised rescaling changes the shape of the original distribution.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The formula for VIF is 1/(1-Ri-squared)
VIF becomes infinite when Ri-squared = 1, i.e, 1/(1-1) = 1/0 = infinity
VIF infinite (or Ri-squared = 1) indicates that there is a perfect (100%) correlation between the variables.
To solve this, we need to drop one of the variables which are highly correlated from the model building.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. It is used to determine if both the two sets of quantiles come from the same distribution. If they came from the same distribution, we should see the points forming a line that's roughly straight.
Quantiles are are points in your data below which a certain proportion of your data fall. For example, a 0.3 quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

Importance of Q-Q plot:
When there are two data samples, often we want to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.