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?

Some of the categorical variables have a significant influence on the dependent variable. Some of them influence positively while others influence negatively. Even though few categories like holiday, fall season were identified as highly influencing while doing the EDA, once the model was built and ran RFE or checked VIF, few of them were eliminated due to multicollinearity between other independent variables. In case of year variable, visualization showed a greater demand for the year 2019 than 2018 and it remained valid as per the model as year became the variable with highest positive coefficient.

There was a total of 11 columns that were analysed and among them four were continuous variable. Windspeed was the only continuous variable that became significant at the end. Most of the variance in dependent variable is explained by 9 categorical columns namely,

- 'yr'
- 'weathersit_clear'
- 'mnth_September'
- 'mnth_November'
- 'mnth_February'
- 'mnth_December'
- 'mnth_January'
- 'season_spring'
- 'weathersit_light rain/snow'

As per the model year, clear weather condition and September month positively affect (positive coefficients) demand for bike sharing. Whereas weather condition as light rain/snow, spring season and Nov-Dec-Jan-Feb months negatively affects (negative coefficients) the demand. Variable 'yr' has the highest positive coefficient and light rain/snow weather condition has the highest negative coefficient (in magnitude) .

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

When drop_first = True is mentioned while creating dummy variable, it will drop the reference variable and thus eliminates the variable (keep k-1 dummies out of k categorical levels). It is important to use drop_first = True during dummy variable creation as it will help to reduce the multicollinearity created amongst the dummy variables.

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

Both temperature and feeling temperature have the highest correlation with the target variable, 'cnt'

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

Below given are the assumptions and how it was validated on training set.

- There should be a linear relationship between dependent and independent variables.
  Above assumption was checked by using visualization techniques like pair plot, bar plot, correlation matrix and heatmap.
- Error terms must be normally distributed with mean zero.
  Above assumption was checked by plotting histogram(seaborn.distplot) of error terms (y_actual – y_predicted). Error terms are normally distributed with mean nearly at zero.
- There should not be multicollinearity between independent variables.
  This assumption was validated by finding correlation matrix and VIF(Variance Inflation Factor) values. Variables with VIF greater than 4 were dropped and model was rebuilt in iterations.
- Error terms are independent of each other.
  This was checked by plotting a scatterplot between y actual and residuals (error terms). There was no observable pattern in the scatter plot.
- Error terms have constant variance (homoscedasticity):
  This was checked by plotting y_predicted values and error terms. Even though the error terms seemed to have a constant variance, towards the higher values of predicted variable, error seemed to be increasing. As it was only for few data points, it was ignored.

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

The top 3 features and their coefficients which significantly contribute towards explaining the demand of the shared bikes are given as below.

yr	2140.656301
weathersit_light rain/snow	-1894.157790
windspeed	-1819.452046

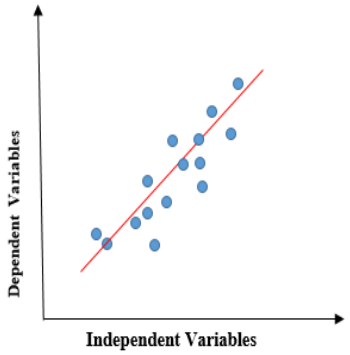
Here year positively affects the demand and weather condition of light rain or snow, windspeed negatively affects the demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm that tries to find out the best relation between dependent and independent variables. For applying linear regression, as name suggests, the target and predictor variables must be linearly related.
There are two types of linear regression. One is simple linear regression and the other is multiple linear regression.
Simple Linear Regression:
Here there will be only one dependent and one independent variable. For example, regression model created for sales and money invested in marketing can represent a simple linear regression model. The linear regression algorithm gives a best fitted sloped line describing the relationship

between variables.



In above figure blue dots represent data points and red line is the best fitted line. To arrive at the best fitted line linear regression uses below equation.

$$y = \beta_0 + \beta_1 x$$

Here y is the target variable
x is the predictor
β_0 is the intercept
β_1 is the slope

The equation can be interpreted as, for a unit increase in x, there will be β_1 times increase in y. If x is equal to zero, y is equal to β_0. In case of sales-marketing example, finding the best fitted line will help to predict sales for a particular marketing money.

Finding the best fitted line is by finding model coefficients (β_0 and β_1). This is calculated using OLS (Ordinary Least Square) method. OLS starts with a basic fitted model and then the residuals or error terms (y_actual - y_predicted) are calculated. As the idea is to achieve least residual sum of squares (RSS), optimisation algorithms are used to reduce cost function in iterations until optimum values of model coefficients are reached. In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values.
By simple linear equation y=mx+b we can calculate MSE as. Let's assume y = actual values, y_i = predicted values

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

The line with optimum coefficients will be the best fitted line.

Multiple Linear Regression:
It is an extension of simple linear regression. In case of multiple linear regression there will be more than one predictor variables and one target variable. Here instead of best fitted line, the regression algorithm finds out best possible hyperplane that fits the data points. Equation is as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

where y is the target variable
x_1 , x_2, x_3 are predictor variables

Here also optimum coefficients are obtained using least square criteria. An example will be the effect of various marketing channels like TV, Radio and Newspaper on the target variable 'Sales'.

In conclusion, linear regression is a powerful statistical method to find relationship between variables. It can be used for both predication and forecasting.

2. Explain the Anscombe's quartet in detail.

To signify the importance of both visualization of data before analysing it and effect of outliers on statistical properties, in 1973, the statistician Francis Anscombe has created four data sets that have identical descriptive statistics (mean, standard deviation, and correlation), yet having very different distributions and appearance when they are graphed. This model is known as Anscombe's quartet.

Given below is an image of the four datasets. Each datasets contains 11 datapoints for two variables x and y.

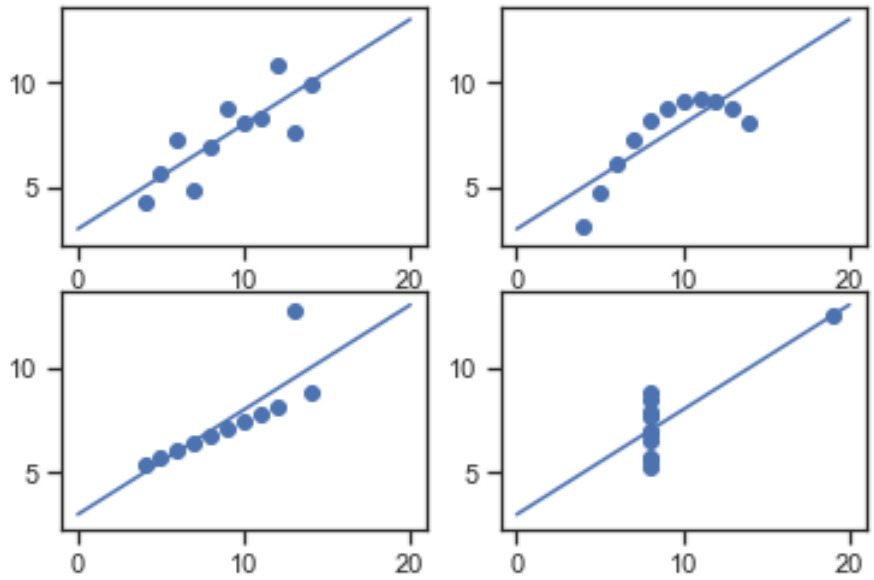
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The resulting descriptive statistics for each datasets are given below,

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

As per the descriptive statistics, mean, standard deviation and correlation for variables is exactly same for all four datasets. Now let's see the visualizations for each datasets.



Observations:

1. Dataset I (Top left) : As per the scatterplot x and y has a linear relationship.
2. Dataset II (Top right) : The graph shows a curve, so x and y has a non-linear relationship.
3. Dataset III (Bottom left) : Here x and y has a perfect linear relationship except for an outlier.
4. Dataset IV (Bottom right) : For this set x value remains constant except for 1 extreme point. Here because of one extreme point this dataset gives out a high correlation between x and y.

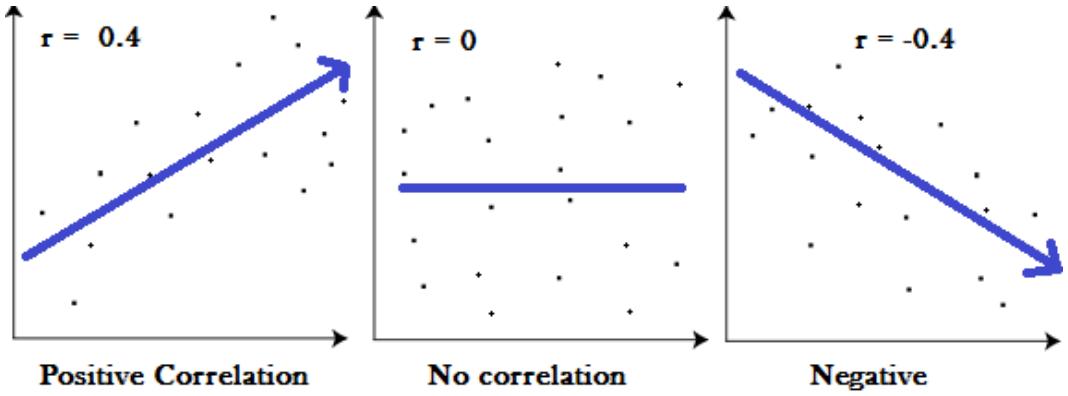
As we have seen above, even though the numerical calculations were exact, they all showed different patterns when they were graphed. So Anscombe's quartet clearly demonstrates the importance of plotting variables while doing exploratory data analysis, because by only looking at the descriptive statistics can mislead the conclusions we make on the dataset. Visualizations help us to reveal some of the underlying patters which we wouldn't have thought about in the first place.

3. What is Pearson's R?

Pearson's R or Pearson's correlation coefficient R is a measure of how strong is the linear relationship between two variables. It is the ratio between covariance of two variables and product of their standard deviation. Correlation coefficient always has a value between -1 and 1, where:

- 1 indicates a strong positive relationship
- -1 indicates a strong negative relationship
- 0 indicates no relationship at all

Given below are graphs representing various correlations and their is Pearson's R value.



A positive correlation indicates a direct relationship between variables. For example, as age increases the height of the child also increases proportionally. Whereas a negative correlation indicates an inverse relationship. For example when temperature decreases, more heaters are purchased. Correlation of zero says that change in one variable does not affect or result in a linear change in another variable. For example, there is no relation between the number of road accidents and amount of tea drunk. But one thing that need to be remembered here is that correlation does not mean causation. Meaning, even if there exists a correlation between two variables, it does not mean that change in one has caused the change in other.

Correlation coefficient = covariance(X,Y) /(stand.deviation of X)*(stand.deviation of Y)
Here are the assumptions for calculating Pearson's R:

- Variables must be normally distributed
- There should not be any significant outliers in the data
- Variables must have a linear relationship
- Variables must be continuous

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

In machine learning, a step under data pre-processing for brining all independent variables to a normalized range is known as feature scaling. As the machine learning algorithms only considers numbers, if feature scaling is not done, it tends to give more importance to higher figures and less importance to smaller values regardless of the unit of the values.
Scaling is done for easier interpretation and faster convergence for algorithms that use gradient descent as optimisation method. Imagine a dataset consisting of various independent variables such as 'Length' in meters (ranging from 1-20), 'Area' in square feet (ranging from 1000 to 5000) and 'Price' in lakhs (ranging from 10 to 50) . When the machine learning algorithm learns this data, it obviously have an inclination towards the variable having larger magnitudes. Here in this case the algorithm will be biased towards the variable area and it will think that area is more important than length and price. So it is important to make all independent variables to one scale.
Two prominent scaling techniques are normalization and standardization.

Normalization: Normalized scaler shifts and transforms the data into a range between 0 and 1. It is also called as Min-Max Scaling.

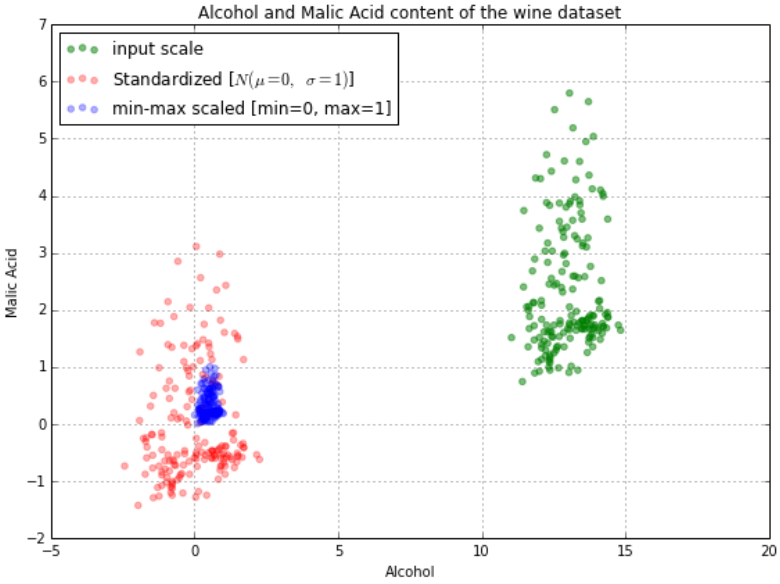
Formula is $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where Xmax and Xmin and maximum and minimum values of X.

Standardization: Standardized scaler transforms the data in such a way that the resulting distribution has a mean of 0 and a standard deviation of 1.

Formula is $$X' = \frac{X - \mu}{\sigma}$$

Where mu is the mean and sigma is the standard deviation of X. Below given graph shows the difference between normalized data and standardized data.



As we can see the normalized data is now between zero and one. Because of this normalization is can be used when there are outliers in the data but some of the information about feature will be lost here. Whereas standardization retains the information about the data.

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

In multiple linear regression there can be chances that two or more variables are highly correlated with one another. This is known as multicollinearity. Multicollinearity makes it difficult to uniquely identify model coefficients and model becomes unreliable. One way to detect multicollinearity in the model is by using VIF.

VIF -Variance Inflation Factor- is an index that measure how much variance in one independent variable is explained by the other independent variables. We fit regression model between independent variables to determine VIF. Formula is,

$$VIF = \frac{1}{1-R^2}$$

R-squared value is calculated to determine how well an independent variable is described by the other independent variables. So a high value of R-squared indicates that the predictor variable is highly correlated with other predictor variables.

When R-squared value becomes closer to one, VIF value will become infinite which indicates that almost 100 percent variance in this particular independent variable is explained by other independent variables and it has a high multicollinearity. In this case, the variable must be dropped to deal multicollinearity.
In general below given table helps to identify the level of multicollinearity,

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

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

A Q-Q plot helps to identify whether a set of data possibly come from a theoretical distribution such as a Normal, exponential or Uniform distribution. It is also used to check if two datasets come from populations of similar distribution. It basically plots the quantiles of first dataset with the second dataset.
Interpretations:

- If two datasets have similar distribution then Q-Q plot follows a 45° line.
- If all points lie away from the 45° line, then datasets have different distributions.

As one of the assumptions of a linear regression is the error terms to be normally distributed, we can use Q-Q plot to identify whether it follows a normal distribution or not. Usually we check this criteria using histogram. However it has some limitations. Histogram becomes tattered when there is less number of data points. Also the fit to a continuous probability density function depends on the intervals we choose for histogram. Overall, histogram cannot give us a guarantee for confirming normal distribution of error terms. Instead we could use Q-Q plot to check the distribution. Here quantiles of standardized residuals is plotted against quantiles of theoretical normal distribution. If the plot comes close to a 45° straight line, we can confirm that the error terms are normally distributed.