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?

Ans:

- When compared to other seasons, the demad of bicycles is lower in the month of spring.
- During the week, the demand for bicycles is nearly same.
- When compared to the previous year, the demand for bicycles grew in 2019.
- The months of June through September are the busiest for bike sales. The month of January has the lowest demand.
- There is no discernible difference in bike demand between working and nonworking days
- The demand for bicycles is lower over the holidays than when it is not.

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

Ans: drop first= True is necessary to utilise since it reduces the additional column formed during the construction of dummy variables. As a result, the correlations between dummy variables are reduced.

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

Ans: Temperature has significantly high correlation with Target variable(Cnt)

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

Ans: 1) checking linear relationship between x and y.

2) Error terms are normally distributed with mean zero(not x, y).

3) Error terms are independent of each other.

4) Error terms have constant varaiance(homoscedaiscity)

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

Ans: Top 3 features on final model

1) `mnth_sept`
2) temp
3) season_summer

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a supervised learning ML method. Linear regression predicts a dependent variable (goal) based on the provided independent variable (s). As a result, this regression approach determines a linear connection between a dependent variable and the other independent variables offered.

Equation: y=mx+c

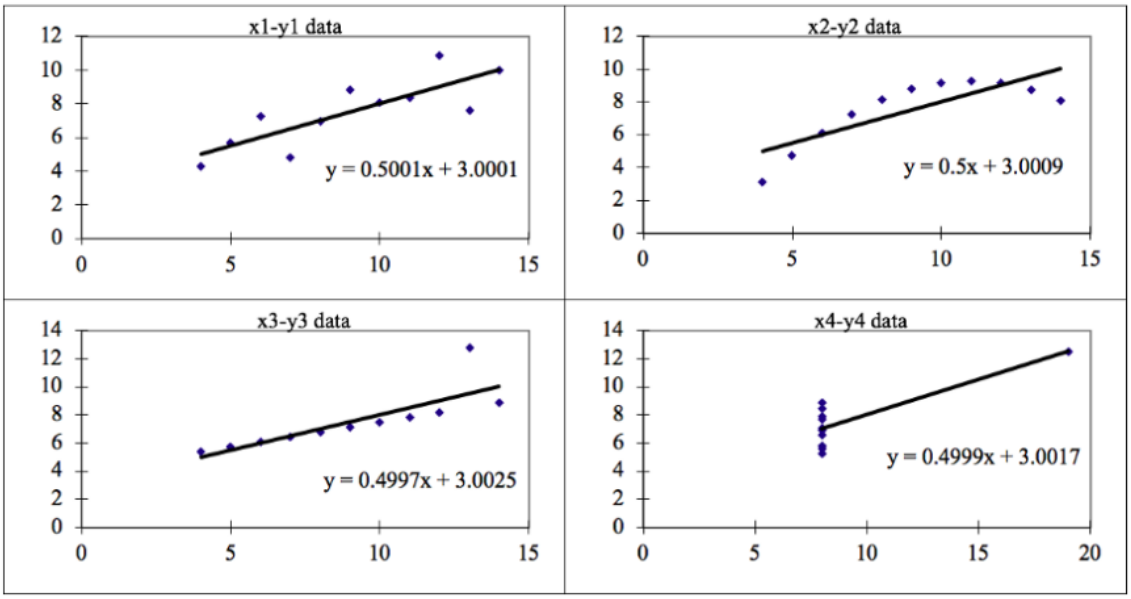
We get the best fit line once we discover the best m and c values. So, when we use our model to forecast, it will predict the value of y for the input value of x.

Cost Function

The model seeks to predict y value such that the error difference between projected value and real value is as little as possible by attaining the best-fit regression line. As a result, it is critical to update the m and c values in order to get the optimal value that minimises the error between the predicted y value (pred) and the real y value (y).

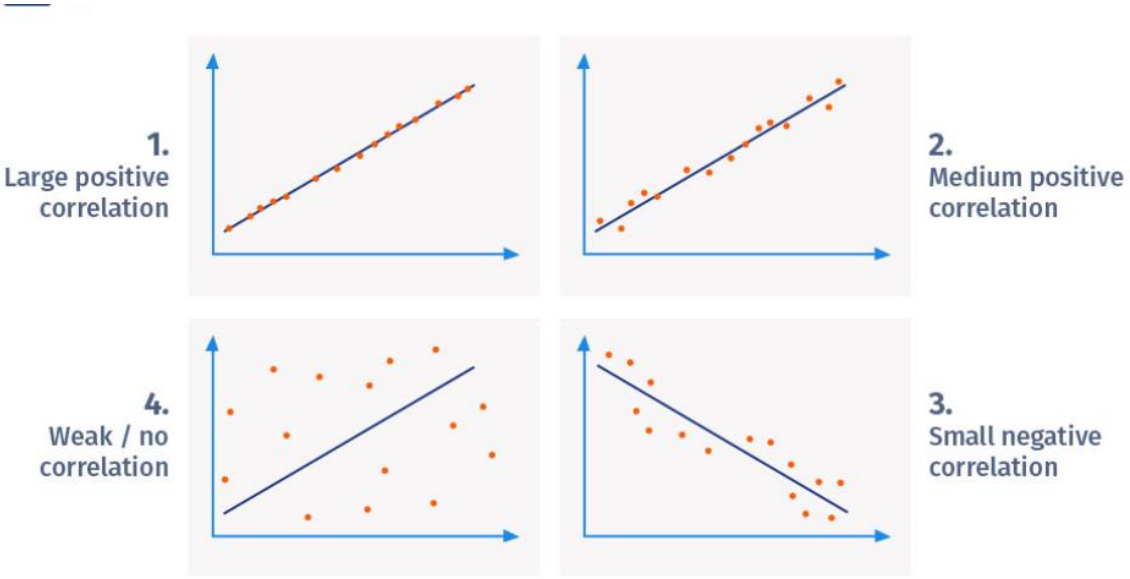
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet is a collection of four data sets that are virtually identical in simple descriptive statistics, but have certain idiosyncrasies that deceive the regression model if formed. They have highly diverse distributions and show very differently on scatter plots.



3. What is Pearson's R?

Ans: Pearson coefficient correlation is statistically significant. It investigates the connection between two variables. It attempts to draw a line between the data of two variables in order to demonstrate their link. The Pearson correlation coefficient calculator is used to calculate the connection between the variables. Positive or negative linear relationships can exist.



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

In Simple Linear Regression, scaling doesn't impact your model. Here we can see that except for area, all the columns have small integer values. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

As you know, there are two common ways of rescaling:

1. Min-Max scaling (normalisation): data between 0 and 1 (x-xmin)/(xmax-xmin), it will take out the outliers

2. Standardisation (mean-0, sigma-1): (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). Now, let's rescale and fit the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: VIF = infinite if there is perfect correlation. This demonstrates an exact correlation between two independent variables. In the event of perfect correlation, R2 = 1, resulting in 1/(1-R2) infinite. To resolve this issue, we must remove one of the variables from the dataset that is producing the perfect multicollinearity. An infinite VIF value suggests that a linear combination of other variables may represent the relevant variable perfectly (which show an infinite VIF as well).

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

Ans: Q-Q Plots (Quantile-Quantile Plots) are comparisons of two quantiles. A quantile is a fraction of values that fall below that quantile. For example, the median is a quantile where 50% of the data falls below it and 50% fall above it. The goal of Q Q plots is to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45-degree angle is drawn; if the two data sets are from the same distribution, the dots will fall on that reference line.

The points in the Q–Q plot will roughly lie on the line y = x if the two distributions being compared are comparable. The points in the Q–Q plot will roughly lie on a line if the distributions are linearly connected, but not necessarily on the line y = x. Q–Q plots may also be used to estimate parameters in a location-scale family of distributions graphically.

A Q–Q plot is used to compare the morphologies of two distributions, showing how features like location, size, and skewness are similar or different in the two.