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?

   Snow/Rain seems to have a large negative effect on rentals, while misty or cloudy weather also has a negative effect, but much smaller in size
   Year number has a large positive effect, indicating that demand will grow with time
   Summer and fall have the highest rentals among seasons, followed by winter (spring has the lowest)
   Weekends vs weekdays do not have much effect on demand, but holidays do have slightly lower demand

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

   If there are N distinct levels for a categorical column, you only need N-1 dummy variables to represent it, as one of the levels can be represented by a vector of N-1 zeros (the remaining levels will be one-hot encoded). The above command drop_first=True drops one of the dummy columns to ensure N-1 output columns instead of N. This is an efficient representation that avoids redundant columns that would make our dataset unnecessarily larger.

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

   Temp and atemp have the highest correlation with the target variable (temp and atemp have a correlation of >0.99 so it is hard to choose between the 2)

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

   Plotted a histogram of residuals to ensure they are centered around zero and look approximately normal

QQ plot of residuals also used to test normality (points should all be on the line, where they would theoretically lie on a normal distribution)

Plotted residuals vs predicted values to see if there is any pattern in the residuals and whether they are homoscedastic (variance is roughly the same throughout)

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

   Temp (higher temp corresponds to more rentals)
   Yr (higher rentals can be anticipated in the 2nd year due to growth over time)
   Rain/Snow weather (has a negative impact on rentals)

General Subjective Questions

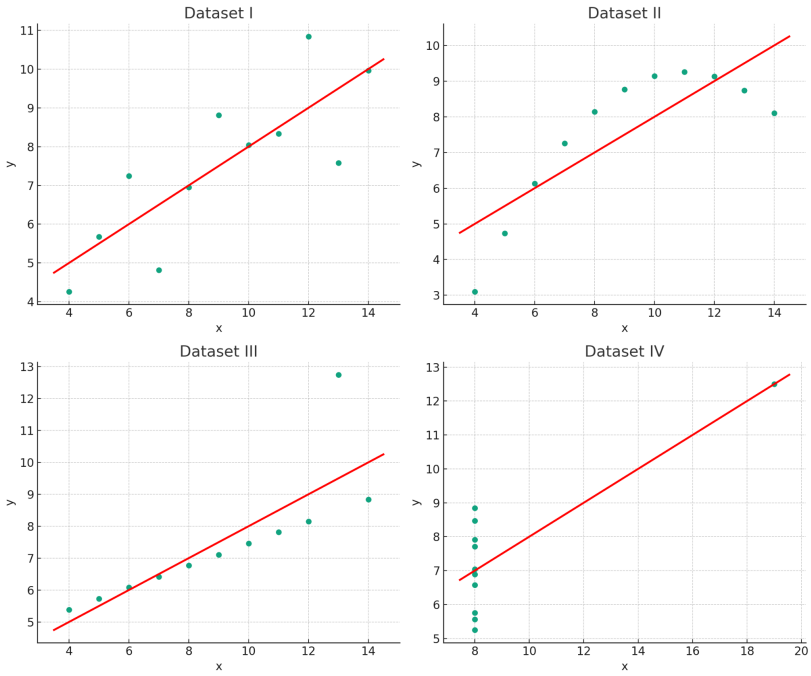
1. Explain the linear regression algorithm in detail.

   Linear regression assumes a linear relationship between independent variables (Xs) and a target variable y. The equation takes the form $y=a+b_1X_1+b_2X_2+b_3X_3+...b_nX_n + \varepsilon$ where ε is the error term (or residuals), which is any component not explained by the X variables.
   The values of the coefficients (a, b_1,b_2, etc.) are determined by minimizing a cost function, namely the Mean Squared Errors (MSE), which is the mean of the squared difference between actual and predicted values.
   This cost function can be minimized using Ordinary Least Squares(OLS), or in case of large datasets with many variables, Gradient Descent can be used to arrive at an approximate solution.
   The model places several assumptions on the data, namely linearity (linear relationship between X and y), independence (observations should be independent of each other), normality (residuals should be normally distributed with mean 0) and homoscedasticity (residuals should have constant variance).

Once the model has been trained and the best fit line has been calculated, these assumptions need to be tested. The model can be evaluated using several metrics such as R-squared, adj. R-squared, AIC and BIC. The results should also be validated against a test dataset (unseen samples that were not used on the training) to avoid overfitting to the training data.

2. Explain the Anscombe's quartet in detail.

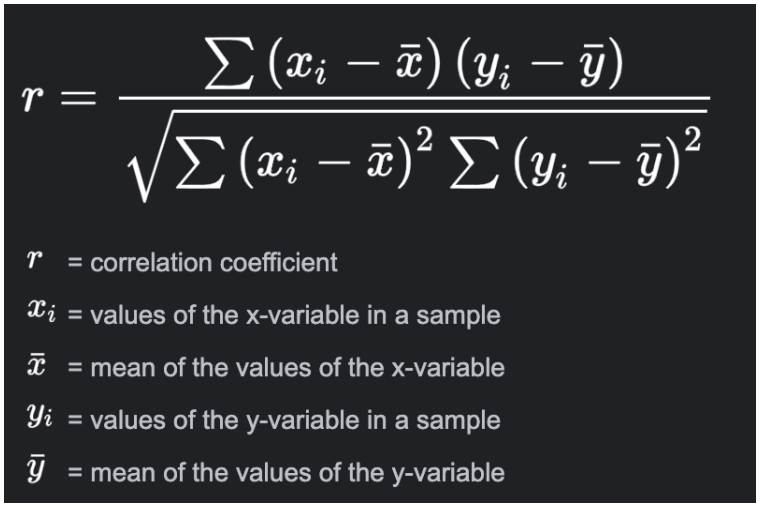
Anscombe's quartet was constructed by statistician Francis Anscombe to illustrate the importance of visualizing data before conducting statistical analysis or model building.
It consists of 4 datasets, each of which have the same descriptive statistics, such as mean and variance. If one naively applied a linear regression model, all 4 datasets would output the same best fit regression line. However, when we plot the data we see 4 very different patterns.



As one can see plotting the data,
Dataset I fits the linear model well
Dataset II is not a linear relationship
Dataset III is linear, but has an outlier which drastically alters the best fit line
Dataset IV has all x values the same, apart from an outlier - clearly not a linear relationship

3. What is Pearson's R?

   It is a measure of linear correlation between 2 variables (say x and y). It takes values between +1 and -1, denoting the strength and direction of the linear relationship (+1 being perfect positive correlation, -1 being perfect negative correlation).
   Generally a high positive correlation means when x increases, y also usually increases and vice versa. A high negative correlation means when x increases, y usually decreases and vice versa. Correlation of low magnitude means when x increasing or decreasing has little to no bearing on the direction y moves and vice versa.
   The formula is given below -



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

   Scaling is the process for transforming a variable in a way that brings it within a range of values. In the context of linear regression, it is performed so that the coefficients can be interpreted for effect size.
   For example, let us say we are predicting house value based on Area in square feet and the floor number. Floor number will take on very low values (say 1-15), compared to area (which will be in thousands). Due to this, the coefficient size is not comparable as the coefficient of floor

number would naturally be higher even if it had a smaller effect. By bringing both variables to a similar scale, the coefficients are comparable.
Normalization involves restricting the values a variable can take within a specific range. For mixmax scaling, this range is [0,1].
Standardization uses mean and variance of a distribution to scale it to having a mean 0 and variance 1, but places no upper or lower bound on the values. A massive outlier could still take on very high values.
Normalization is more suitable when your data does not have a Gaussian distribution.

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

   VIF is a measure of collinearity between predictor variables. The formula is 1/(1-Rsquared) where Rsquared refers to the model of the other predictor against the variable in question. If the variable in question is almost completely explained by the other predictors, it will have an Rsquared close to 1, which means the denominator will be close to zero, leading to an infinite VIF score.

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

   A Quantile-Quantile plot is a plot that is used to test if a dataset follows a particular theoretical distribution. It plots the quantiles of the actual data against the theoretical quantiles that would emerge from that particular distribution. If the data actually follows that distribution, the points should fall on the 45 degree line.
   In linear regression, the model assumes that the residuals are normally distributed. This assumption can be tested by plotting a Q-Q plot of the residuals, with a normal distribution for the theoretical quantiles. If we see that many points do not fall on the 45 degree line, this points to the fact that this assumption has been violated and our model is not valid. We may need to change our predictor variables or apply some transformations to the dependent variable and rebuild the model.