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)

Ans: Inferences from categorical variables plots:

1. 'clear' weathersit has most bookings
2. there is no data for heavy_rain in weathersit
3. most number of bookings happened during 'fall' season
4. there were more bookings in 2019 as compared to 2018

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

Ans: Let us suppose, an independent variable contains n categorical values. We have to remove it and dummy variables corresponding to each n categories. When we add n new variables, as they all are correlated, it will lead to Dummy Variable Trap. It will make difficult to interpret the model and can increase standard errors. Thus to avoid it, we use drop_first=True in get_dummies().

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: As per the pair-plot among numerical variables, temp has 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? (3 marks)

Ans: I have performed following analysis for validating assumptions of Linear Regression:

1. Normality of error terms : Error terms were normally distributed
2. Multicollinearity check : There were insignificant multicollinearity among variables
3. Linear relationship : confirmed by the pair plot graphs
4. Homoscedasticity : there were no visible pattern in residuals

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 contributing towards demand of shared bikes:

1. Temp : coeff = 0.4302
2. Yr : coeff = 0.2347
3. weathersit_light_rain : coeff = - 0.2896

General Subjective Questions

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

Ans: Linear regression is a statistical modeling technique for examining the relationships between a dependent variable and one or more independent variables. Linear regression assumes that there is a linear relationship between the dependent variable and independent variable(s), meaning that the association can be characterized by a linear line.

The equation of a linear regression model is:

$Y = \beta0 + \beta1X1 + \beta2X2 + … + \beta nXn$

where:

Y is the dependent variable

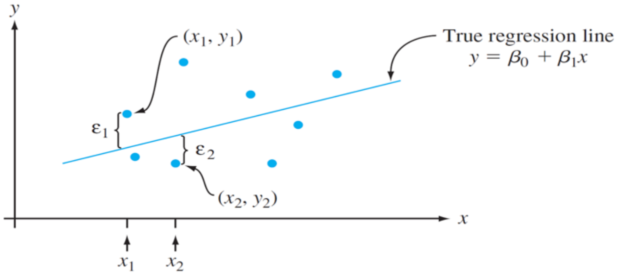
X1, X2, …, Xn are the independent variables

β0 is the intercept

β1, β2, …, βn are the coefficients

Types of linear regression:

1. Simple Linear Regression : It involves only one independent variable and one dependent variable. In case of Simple Linear Regression, X1 is the only independent variable and the equation simplifies to:
   $Y = \beta0 + \beta1X$



2. Multiple Linear Regression : It involves more than one independent variable and one dependent variable.

The goal of linear regression is to find the values of all the constants(β0, β1, ..., βn) that minimize the distance between the predicted values and the actual values. This is often done using a method called Root Mean Square Error.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n \text{ are predicted values}$$
$$y_1, y_2, \ldots, y_n \text{ are observed values}$$
$$n \text{ is the number of observations}$$

Assumptions:

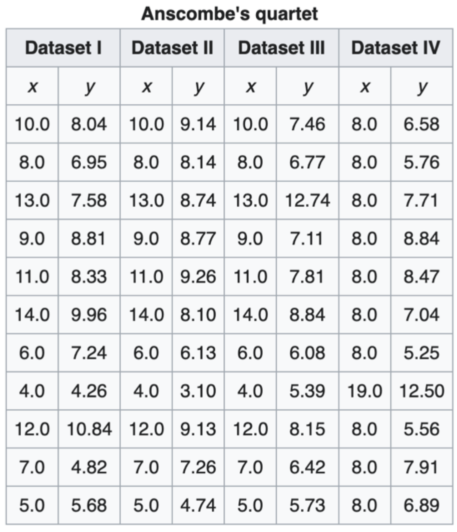
We have to make some assumptions to use Linear regression:

1. Linearity: The dependent and independent variables have a linear relationship.
2. Independence: The observations are independent.
3. Homoscedasticity: The variance of the errors is constant.
4. Normality: The errors must be normally distributed.

If these assumptions are not met, the results of the regression analysis may be faulty.

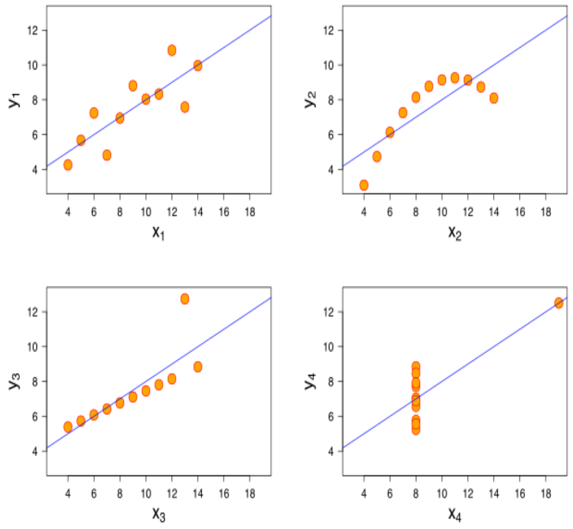
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: In 1973, statistician Francis Anscombe built 4 datasets(Anscombe's quartet) in order to demonstrate the importance of plotting data prior to analyzing it and creating your model. Each of the four datasets all have nearly the same statistical observations, providing you with the same information (in terms of variance and mean) of every x and y point in each dataset. When you plot each dataset, however, they all appear to be very different plots from one another.



Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset 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



Anscombe’s quartet dataset:

1. A standard linear relationship between x and y.
2. A quadratic relationship between x and y.
3. It looks like a tight linear relationship between x and y, except for one large outlier which exerts enough influence to lower the correlation coefficient.
4. It looks like the value of x remains constant, except for one outlier which produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R? (3 marks)

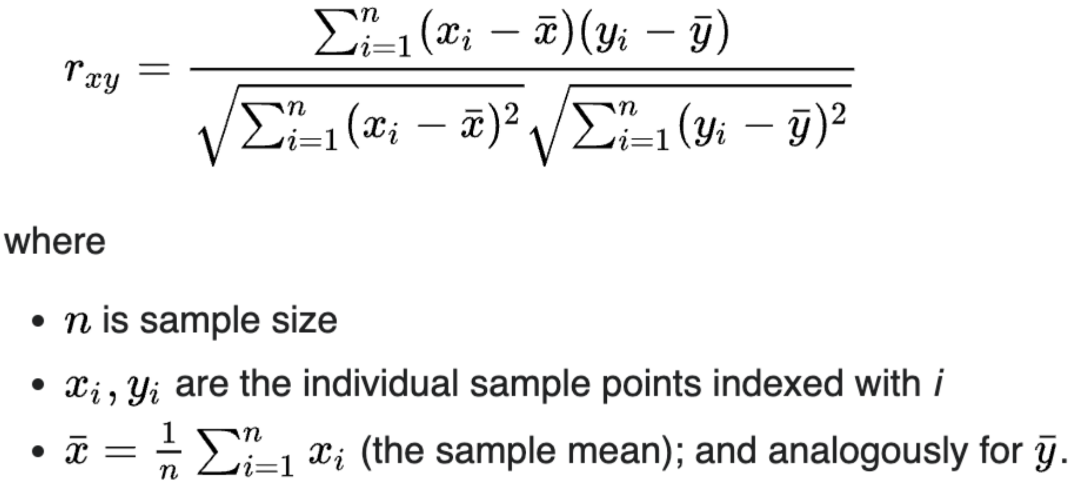
Ans: Pearson's correlation coefficient(r) also known as Pearson's R is an index used to measure the linear relationship between two variables.

The values of r can have values between -1 and 1:

1. r = 1, means there is a perfect positive correlation, meaning the two variables increase or decrease together.
2. r = -1, means there is a perfect negative correlation, meaning one variable will increase and the other will decrease.
3. r = 0, means there is no correlation between the two variables.

Formula:

[PearsonCoefficientFormula.png]



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

[Correlation_coefficient.png]

Direction: The sign of r indicates the direction of the relationship:

Positive r: Both variables increase or decrease together

Negative r: One variable increases as the other decreases

Strength: The absolute value of r indicates the strength of the relationship:
|r| close to 1: Strong relationship
|r| close to 0: Weak relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans: Scaling is the process of transforming data into a common range or scale. In data science/machine learning, this is frequently required to ensure that different features contribute equally to the learning process of the model.

Purpose of Scaling:

1. Equal Contribution: If the features are having different scales (i.e.: one feature in meters, the other thousands), there is a chance it will bias the model. Scaling ensures that all features contribute equally.
2. Convergence Speed: Using scaled features, algorithms like gradient descent will converge more quickly.
3. Regularization: Certain regularization techniques perform better when scaled effectively.

Key difference between Normalized Scaling and Standardized Scaling :

1. Min/Max values are used in Normalized scaling, Mean/StdDev are used for Standardized scaling 2. Scaled value ranges from [0, 1] or [-1, 1] in Normalized scaling, there is no bounded range for Standardized scaling.
2. Outliers affects Normalized scaling, there is no impact of outliers in case of Standardized scaling
3. Standardized scaling is commonly used for data having standard normal distributions, Normalized scaling is used when we dont have knowledge about data distrubtion.
4. Normalized scaling preserves relative differences, Standardized scaling does not preserve it.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: The Variance Inflation Factor (VIF) is a way of examining multicollinearity in regression analysis. It measures the degree to which the variance of a regression coefficient is inflated due to correlations with the other predictors. A VIF that is infinite indicates perfect multicollinearity in your regression dataset. Perfect multicollinearity exists when one of the predictor variables can be perfectly predicted by a linear combination of the other predictor variables. This occurs when the columns of the data matrix are linearly dependent.

Reasons for having infinite VIF:

1. Redundant Variables: When there are two variables that are essentially similar for example, age and "age in years".

2. Linear Combinations: When a variable is measured as a linear combination of other variables for example, "total_sales" = "quantity_sold" * "unit_price".
3. Dummy Variables: When creating dummy variables for categorical variables, you should avoid including all levels. If you include all levels, one of the dummy variables can be perfectly predicted by the others, resulting in infinite VIF.
4. Data Entry Errors: Data entry or data preparation mishaps can lead to perfectly correlated variables.

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

(3 marks)

Ans: A Q-Q plot is a scatter plot developed by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for; the second quantile is the actual distribution you are comparing it to. If two quantiles from the two respective distributions are sampled from the same distribution it is expected that it will fall in line.
The Q-Q plot is a visual tool for comparison, the results can be useful to support or reject the hypotheses.

Use and Importance of Q-Q Plots in Linear Regression:

1. Assessing Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (the differences between the actual and predicted values) are normally distributed. A Q-Q plot can visually check this assumption.
2. Identifying Outliers: Outliers can significantly impact the results of linear regression. Q-Q plots can help identify outliers that might be skewing the distribution.
3. Detecting Skewness and Kurtosis: Q-Q plots can reveal if the data is skewed (asymmetric) or has heavy tails (kurtosis).
4. Choosing Appropriate Transformations: If the residuals are not normally distributed, transforming the data (e.g., log transformation) might help achieve normality. Q-Q plots can guide this decision.

The Use and Importance of Q-Q Plots in Linear Regression:

1. Normality of Residuals: In linear regression, we expect the residuals (the difference between the actual values and the predicted values) to be normally distributed. A Q-Q plot is a good way to check this visual.
2. Detecting Outliers: Q-Q plots can help identify outliers that may affect the distribution.
3. Skewness and Kurtosis: Q-Q plots can demonstrate when the data are skewed (asymmetrical) or has heavy tails (kurtosis)
4. Choosing Transformations: If the residuals are not normally distributed, it may be useful to transform the data (e.g. using log transformation) based on Q-Q plots.