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)

From the analysis I have conducted analysis on categorical columns using both box plots and bar plots. Here are several key observations drawn from the visualizations:

•Booking activity appears to surge during the fall season, with a significant increase observed in each season from 2018 to 2019.
•The majority of bookings occur between May and October, with a rising trend from the beginning of the year until mid-year, followed by a decline towards the year's end.
•Favorable weather conditions notably attract more bookings, which is unsurprising.
•Thursdays, Fridays, Saturdays, and Sundays show a higher number of bookings compared to the beginning of the week.
•Booking frequency tends to decrease on non-holiday days, which is reasonable as people may prefer spending time at home with family during holidays.
•There is a relatively balanced distribution of bookings between working and non-working days.
•2019 witnessed a notable increase in bookings compared to the previous year, indicating positive business growth.

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

Using drop_first = True is crucial as it aids in minimizing the surplus column generated during the creation of dummy variables, thereby mitigating correlations among these variables.
In terms of syntax, drop_first is a boolean parameter, with a default value of False. When set to True, it indicates whether to derive k-1 dummy variables out of k categorical levels by excluding the first level.
For instance, suppose we have a categorical column with three distinct values, and we intend to generate dummy variables for that column. If one variable is not A and B, it logically implies it's C. Hence, the third variable is redundant for identifying C, rendering it unnecessary.

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

'temp' variable has the highest correlation with the target variable.

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

I have verified the assumptions underlying the Linear Regression Model through the following five criteria:
•Checking for Normality of Error Terms: Ensuring that error terms exhibit a normal distribution.
•Assessing Multicollinearity: Verifying that there is no significant multicollinearity among the variables.
•Linearity: Verifying that the predictor variables in the regression have a straight-line relationship with the outcome variable.
•Homoscedasticity: The variance of the dependent variable is the same for all the data.
•Confirming Independence of Residuals: Ensuring absence of autocorrelation within the 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)

- Temperature
- Weather
- Year

General Subjective Questions

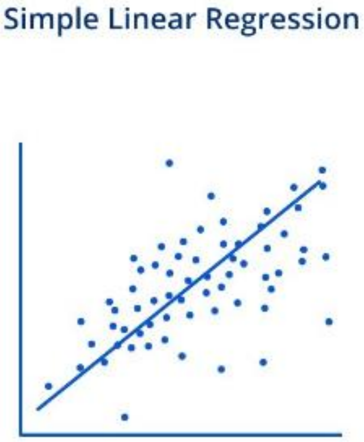
1. **Explain the linear regression algorithm in detail. (4 marks)**

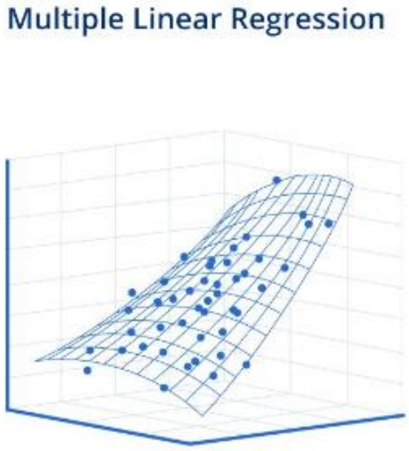
Linear regression is a statistical model used to analyze the linear association between a dependent variable and a given set of independent variables. This implies that changes in the independent variables result in corresponding changes (either increases or decreases) in the dependent variable.

This relationship is mathematically represented by the equation Y = mX + c, where:
•Y represents the dependent variable being predicted.
•X represents the independent variable used for predictions.
•m is the slope of the regression line, indicating the impact of X on Y.
•c is a constant known as the Y-intercept. When X = 0, Y equals c.

Additionally, the linear relationship can be categorized as either positive or negative:
•Positive Linear Relationship: Both independent and dependent variables increase.
•Negative Linear Relationship: The independent variable increases while the dependent variable decreases.

Linear regression can be classified into two types: Simple Linear Regression and Multiple Linear Regression.



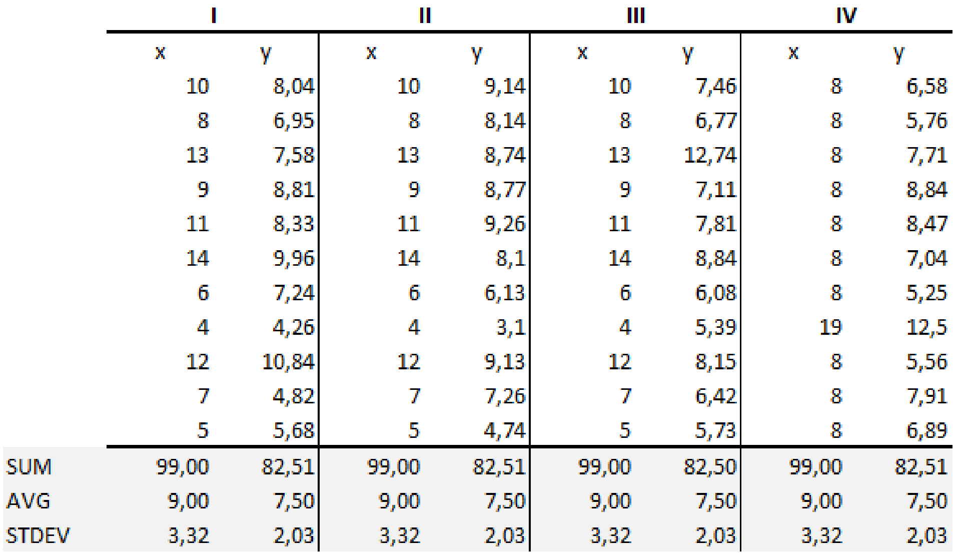


When employing linear regression, our main goal is to identify the optimal line of fit, ensuring that the discrepancy between predicted and actual values is minimized. The best-fit line is characterized by the lowest possible error. The aim of linear regression is to determine the coefficients of a linear equation that most accurately fits the training data. This is achieved by adjusting the coefficients in the direction of the negative gradient of the Mean Squared Error with respect to the coefficients.

Assumptions about the dataset made by the Linear Regression model include:
•Multi-collinearity: The model assumes minimal or no multi-collinearity in the data, which occurs when independent variables or features are interdependent.
•Auto-correlation: Linear regression assumes minimal or no auto-correlation in the data, where residual errors exhibit dependency.
•Relationship between variables: The model assumes a linear relationship between response and feature variables.
•Normality of error terms: Error terms are expected to be normally distributed.
•Homoscedasticity: There should be no discernible pattern in residual values.

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

Anscombe’s Quartet, devised by statistician Francis Anscombe, consists of four datasets, each comprising eleven (x, y) pairs which have identical descriptive statistics. However, when visualized through graphs, each graph presents a distinct story.



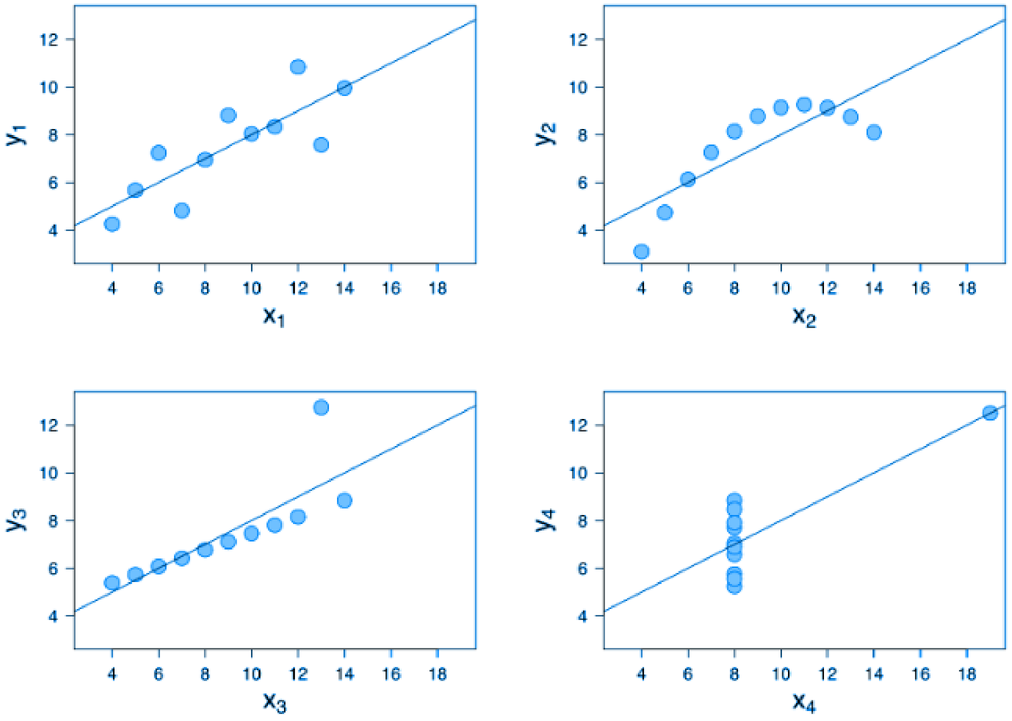
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics indicate that the means and variances are consistent for both x and y across the datasets:

•The mean of x is 9, and the mean of y is 7.50 for each dataset.

•Likewise, the variance of x is 11, and the variance of y is 4.13 for each dataset.

•The correlation coefficient (indicating the strength of the relationship between two variables) between x and y is 0.816 for each dataset.



Upon plotting these datasets on an x/y coordinate plane, it becomes evident that they exhibit the same regression lines. However, each dataset conveys a unique narrative:

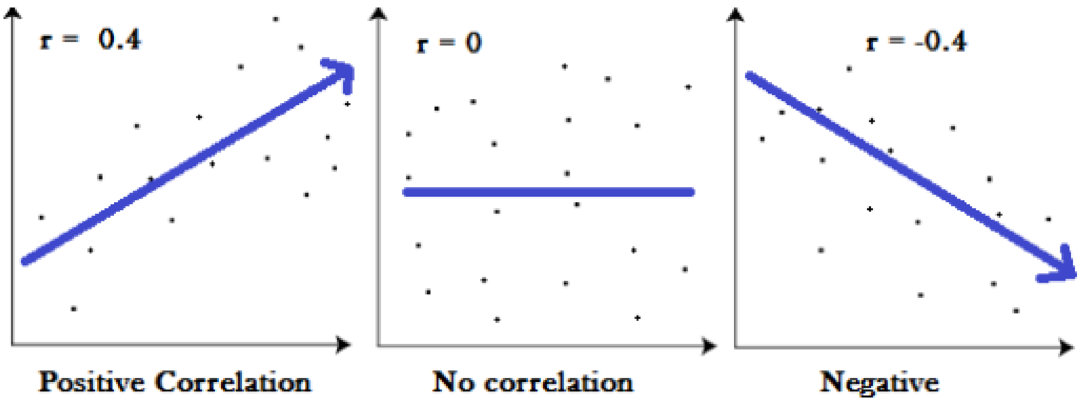
•Dataset I showcases clean and well-fitting linear models.

•Dataset II does not follow a normal distribution.
•In Dataset III, the distribution is linear, but the calculated regression is distorted by an outlier.
•Dataset IV demonstrates how a single outlier can inflate the correlation coefficient.

This quartet underscores the significance of visualization in Data Analysis. Examining the data visually reveals the underlying structure and provides a clearer understanding of the dataset.

3. What is Pearson’s R? (3 marks)

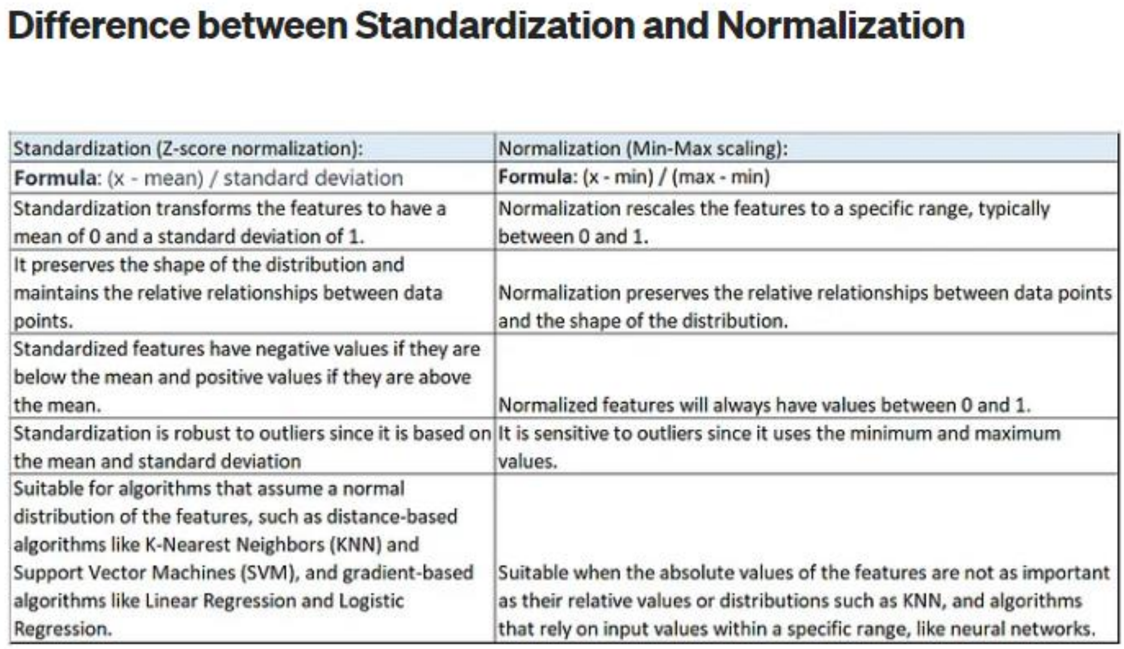
Pearson's correlation coefficient, denoted as "r," provides a numerical measure of the strength of the linear relationship between variables. When variables tend to move in the same direction, the correlation coefficient is positive. Conversely, when variables move in opposite directions, with low values of one variable corresponding to high values of the other, the correlation coefficient is negative. Pearson's r can range from +1 to -1. A value of 0 signifies no association between the variables. A positive value indicates a positive association, meaning that as one variable increases, the other variable also tends to increase. Conversely, a negative value indicates a negative association, where an increase in one variable is associated with a decrease in the other variable.



Graphs showing a correlation of -1, 0 and +1

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling is a method used to normalize the independent features within the dataset to a consistent range. This process is typically carried out during data pre-processing to address disparities in magnitudes, values, or units. Without feature scaling, machine learning algorithms may assign greater importance to larger values and perceive smaller values as less significant, irrespective of their unit of measurement.



Difference between Standardization and Normalization

Standardization (Z-score normalization):	Normalization (Min-Max scaling):
Formula: (x - mean) / standard deviation	**Formula**: (x - min) / (max - min)
Standardization transforms the features to have a mean of 0 and a standard deviation of 1.	Normalization rescales the features to a specific range, typically between 0 and 1.
It preserves the shape of the distribution and maintains the relative relationships between data points.	Normalization preserves the relative relationships between data points and the shape of the distribution.
Standardized features have negative values if they are below the mean and positive values if they are above the mean.	Normalized features will always have values between 0 and 1.
Standardization is robust to outliers since it is based on the mean and standard deviation	It is sensitive to outliers since it uses the minimum and maximum values.
Suitable for algorithms that assume a normal distribution of the features, such as distance-based algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), and gradient-based algorithms like Linear Regression and Logistic Regression.	Suitable when the absolute values of the features are not as important as their relative values or distributions such as KNN, and algorithms that rely on input values within a specific range, like neural networks.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

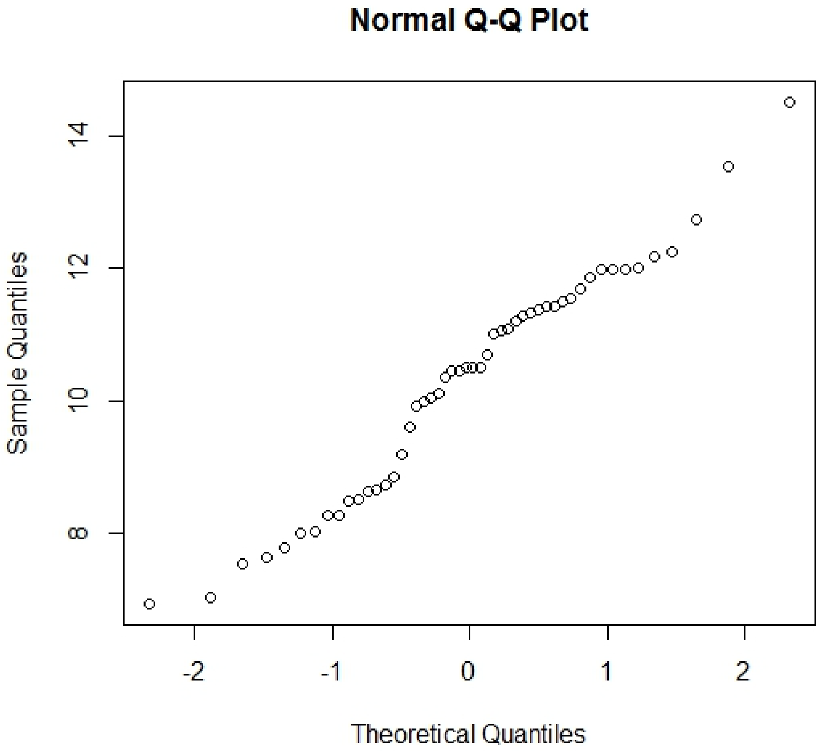
If there's perfect correlation, the Variance Inflation Factor (VIF) becomes infinite. A high VIF suggests correlation between variables. For instance, a VIF of 4 indicates that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity.
When VIF reaches infinity, it indicates perfect correlation between two independent variables. In such cases, R-squared (R2) equals 1, resulting in 1/ (1-R2) becoming infinity. To address this, it's necessary to remove one of the variables from the dataset that is causing this perfect multicollinearity.

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

(3 marks)

The quantile-quantile (Q-Q) plot serves as a graphical method to assess whether two datasets originate from populations with a shared distribution.
Function of Q-Q plot: A Q-Q plot compares the quantiles of one dataset with those of another. A quantile represents the fraction (or percentage) of data points below a given value. For instance, the 0.3 (or 30%) quantile denotes the point where 30% of the data lies below and 70% lies above that value. Additionally, a 45-degree reference line is included. If the datasets stem from populations with identical distributions, the points should generally align along this reference line. The further the points deviate from this line, the stronger the indication that the datasets arise from distributions that differ.

Significance of Q-Q plot: A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically. When analysing two data samples, it's often crucial to ascertain whether the assumption of a shared distribution holds true. If confirmed, estimators of location and scale can combine both datasets to derive estimates of the common location and scale. Conversely, if the datasets exhibit disparities, understanding the nature of these differences becomes valuable. In this context, the Q-Q plot offers deeper insights into the disparities compared to analytical methods like the chi-square and Kolmogorov-Smirnov 2-sample tests.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles