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)

I have conducted an analysis on categorical columns using boxplots and bar plots. Here are some key observations drawn from the visualizations:

- Bookings are notably higher during clear weather conditions, which aligns with expectations.
- Booking frequencies appear relatively consistent between working days and non-working days.
- Thursday, Friday, Saturday, and Sunday show higher booking numbers compared to the early days of the week.
- Non-holiday periods seem to have fewer bookings, which is reasonable as people may prefer to stay at home and enjoy time with family during holidays.
- The majority of bookings occurred in May, June, July, August, September, and October. There is a rising trend from the beginning of the year until the middle, followed by a decline towards the year-end.
- The fall season appears to have garnered more bookings, with a significant increase in booking counts across all seasons from 2018 to 2019.
- 2019 exhibited a higher number of bookings compared to the previous year, indicating positive progress in terms of business.

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

Using **drop_first = True** is crucial because it aids in minimizing the creation of an additional column during the generation of dummy variables. This, in turn, mitigates the correlations that may arise among these dummy variables.

Suppose we have a categorical column with three distinct values, and we intend to generate dummy variables for this column. If a variable is not A or B, it automatically represents C. In this scenario, having a third variable to identify C becomes unnecessary.

Syntax:
drop_first: bool,
default: False,
which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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 assessed the assumptions of the Linear Regression Model, focusing on the following five aspects:

- **Multicollinearity Check:**
  - There should be no significant multicollinearity among variables.
- **Homoscedasticity:**
  - There should be no discernible pattern in the residual values.
- **Independence of Residuals:**
  - There should be no auto-correlation in the residuals.
- **Normality of Error Terms:**
  - The error terms should exhibit a normal distribution.
- **Linear Relationship Validation:**
  - A visible linear relationship should exist among variables.

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

- Winter
- Temp
- Sep

General Subjective Questions

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

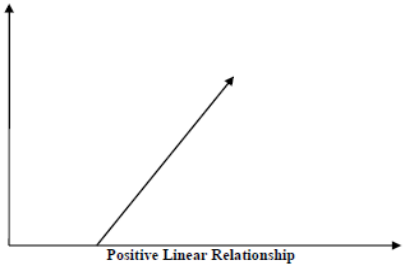
Linear regression is a statistical model that examines the linear association between a dependent variable and a given set of independent variables. In the context of this model, a linear relationship denotes that as the value of one or more independent variables changes (either increasing or decreasing), the corresponding value of the dependent variable changes proportionally (either increasing or decreasing).

This relationship can be expressed mathematically through the following equation:
Y=mX+c where

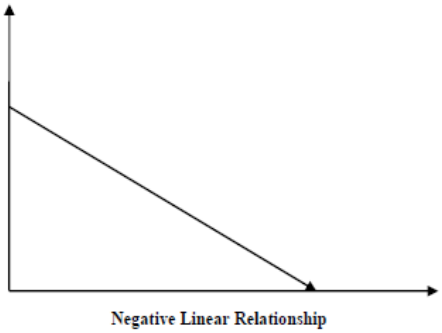
- Y *r*epresents the dependent variable under consideration.
- m signifies the slope of the regression line, indicating the impact of X on Y
- X denotes the independent variable used for making predictions.
- c is a constant known as the Y-intercept. If X=0, then Y would be equal to c.

Moreover, the linear relationship can be characterized as either positive or negative, elucidated as follows:

- **Positive Linear Relationship:**
    - A linear relationship is termed positive when both the independent and dependent variables increase simultaneously. This concept is illustrated through the following graph:



- **Negative Linear Relationship:**
    - A linear relationship is considered negative when the independent variable increases while the dependent variable decreases. This concept is exemplified through the following graph:



Linear regression is of two types. They are as follows

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

Assumptions about dataset that is made by Linear Regression model are as follows:

- **Auto-correlation**
  - The underlying assumption of a linear regression model is that there is minimal or no auto-correlation in the data. Auto-correlation arises when there is a dependence between residual errors.
- **Normality of error terms**
  - Error terms have to be distributed normally
- **Multicollinearity**
  - The linear regression model assumes the presence of minimal or no multi-collinearity in the data. Multi-collinearity arises when there is dependency among the independent variables or features.
- **Relationship between variables**
  - Linear Regression model assumes that response and feature variables relationship must be linear.
- **Homoscedasticity**
  - No visible values in residual patterns should exist

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

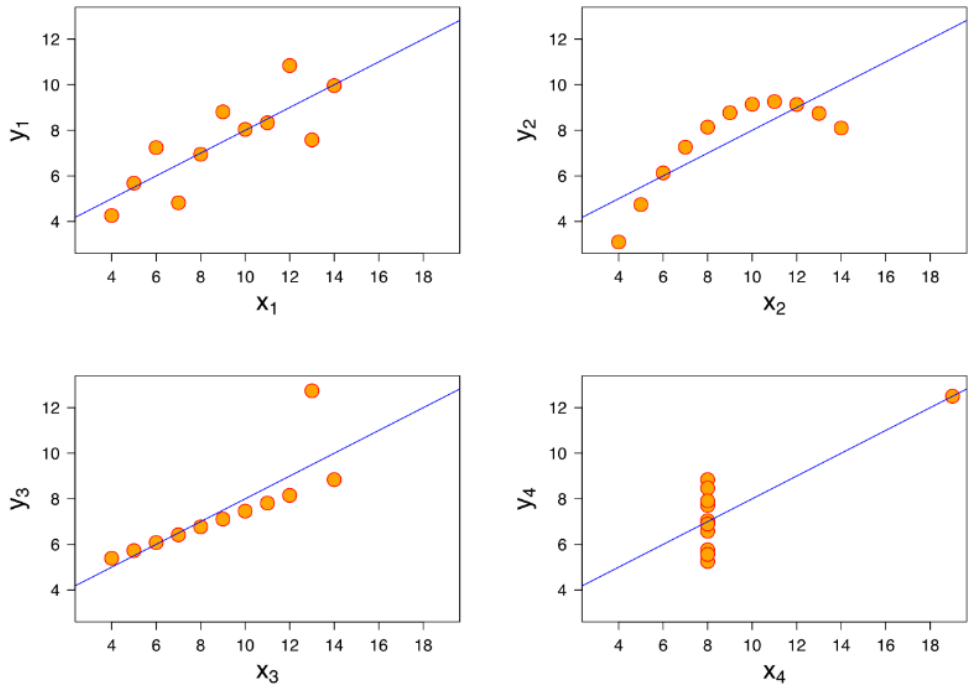
The Anscombe's Quartet, formulated by the statistician Francis Anscombe, consists of four datasets, each comprising eleven (x, y) pairs. Notably, these datasets share identical descriptive statistics. However, a crucial point to emphasize is that when visualized graphically, each dataset tells a distinctly different story, despite their similar summary statistics.

	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 reveal uniform means and variances for both x and y within each dataset:

- The mean of x is consistently 9, and the mean of y is consistently 7.50 for each dataset.
- Likewise, the variance of x is consistently 11, and the variance of y is consistently 4.13 for each dataset.
- The correlation coefficient, indicating the strength of the relationship between x and y, is consistently 0.816 across all datasets.

Despite these consistent summary statistics, when these four datasets are graphed on an x/y coordinate plane, it becomes evident that they share the same regression lines. However, intriguingly, each dataset conveys a distinct narrative.



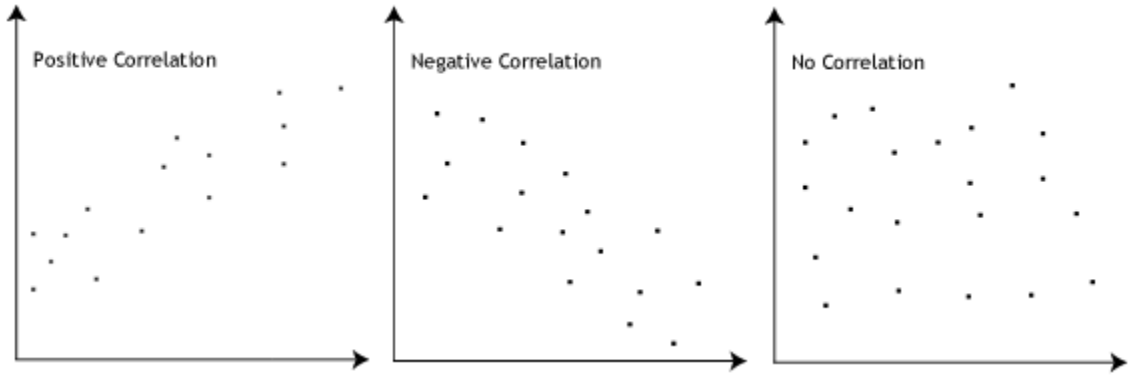
- Dataset I exhibits clean and well-fitted linear models.
- Dataset II does not display a normal distribution
- In Dataset III, the distribution appears linear, but the calculated regression is disrupted by an outlier.
- Dataset IV highlights that a single outlier can lead to a notably high correlation coefficient.

This quartet underscores the significance of visualization in data analysis. Examining the data visually unveils the underlying structure and provides a clear picture of the dataset.

3. What is Pearson’s R? (3 marks)

Pearson's R serves as a quantitative measure of the intensity of the linear relationship between variables. A positive correlation coefficient indicates that the variables tend to move in the same direction—increasing or decreasing together. Conversely, a negative correlation coefficient suggests that the variables move in opposite directions, with low values of one variable corresponding to high values of the other.

The Pearson correlation coefficient, denoted as r, spans a range from +1 to -1. A value of 0 signifies no discernible association between the two variables. A positive value indicates a positive association, wherein an increase in the value of one variable corresponds to an increase in the value of the other variable. Conversely, a negative value signifies a negative association, where an increase in one variable is associated with a decrease in the other variable. This relationship is illustrated in the diagram below:



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 a dataset to a consistent range. This process is integral during data pre-processing, especially when dealing with variables that exhibit significant differences in magnitudes, values, or units. Failure to implement feature scaling may lead a machine learning algorithm to disproportionately emphasize higher values, treating them as more influential, while downplaying smaller values, irrespective of their actual unit significance.

Example: Without employing feature scaling, an algorithm might erroneously perceive a value like 3000 meters as greater than 5 kilometers, which is not accurate. In such instances, the algorithm is prone to making incorrect predictions. Feature scaling is employed to standardize all

values to the same magnitudes, addressing this disparity and ensuring accurate comparisons between variables.

S.No	Normalized Scaling	Standardized Scaling
1	Scaling uses minimum and maximum values of features	Scaling uses mean and standard deviation
2	Used when features are of distinct scales	Used to ensure zero mean and unit standard deviation
3	Values of scales are between [0,1] or [-1,1]	Bounded by no certain range
4	Much affected by the outliers	Less affected by the outliers
5	SciKit-Learn provides MinMaxScaler transformer for normalization	SciKit-Learn provides StandardScaler transformer for normalization

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

In the presence of perfect correlation, the Variance Inflation Factor (VIF) becomes infinite. A high VIF value indicates correlation among variables. For example, a VIF of 4 implies that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity.

When VIF reaches infinity, it signifies a perfect correlation between two independent variables. In instances of perfect correlation, the R-squared (R2) equals 1, resulting in 1/(1-R2) equating to infinity. To address this issue, it becomes necessary to eliminate one of the variables from the dataset 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.

Usage of Q-Q Plot:
A Q-Q plot juxtaposes the quantiles of the first dataset against those of the second dataset. Here, a quantile represents the fraction or percentage of points below a given value. For instance, the 0.3 (or 30%) quantile signifies the point where 30% of the data falls below, and 70% falls above that value. A 45-degree reference line is included. In cases where both sets stem from populations with the same distribution, the points should align approximately along this

reference line. The deviation from this line indicates the strength of evidence supporting the notion that the two datasets have distinct distributions.

Significance of Q-Q Plot:
When dealing with two data samples, it is often crucial to assess whether the assumption of a common distribution holds. If it does, location and scale estimators can amalgamate both datasets to derive estimates for the common location and scale. Conversely, if the two samples differ, the Q-Q plot provides valuable insights into the nature of these differences, surpassing the capabilities of analytical methods such as chi-square and Kolmogorov-Smirnov 2-sample tests.