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:

Season, Weather Situation, holiday, month, working day and weekday are the identified categorical variables.

   a. **Season**: spring season has the lowest bike rent count whereas fall season has the maximum bike rents
   b. **working day**: Bike rents were low during the holidays
   c. **weather Si**t: Rents were high during the Clear weather and low during the 'Light Snow'
   d. **month**: September month has high rentals** while October has the lowest
   e. **weekday**: Weekends have significant increase in rentals

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax - drop first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

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

   Ans: temp and atemp has the highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans:
   1. Distribution of residual values should be normal and centred around 0
   2. Test the assumption by producing the distplot/histplot and check if they follow the normal distribution
   3. The residuals are also scattered around the mean 0.

Standard 5 assumptions needs to be evaluated:

1. Normality of error terms : Error terms should be normally distributed
2. Multicollinearity check : There should be insignificant multicollinearity among variables.
3. Linear relationship validation : Linearity should be visible among variables
4. Homoscedasticity : There should be no visible pattern in residual values.

5. Independence of residuals o No auto-correlation

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

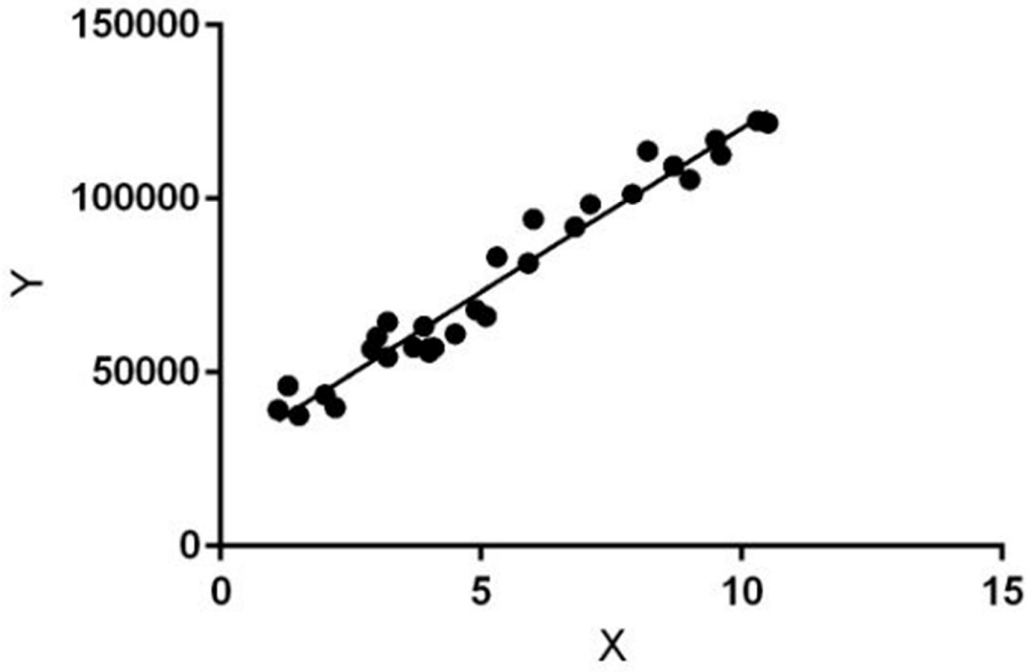
Ans:

1. Temperature (0.3999) With a unit increase in temperature it increases 0.3999 bike rentals
2. Mist (-0.3647) With a unit increase it decrease 0.3647 bike rentals
3. Season (spring) (-0.6842) With increase in unit value it decreases bike rentals by 0.6842

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning methodology. Basically, it performs a regression task. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input and y(output). Hence the name is linear regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Linear Regression may further divide into

1. Simple Linear Regression/ Univariate Linear regression
2. Multivariate Linear Regression

Simple Linear Regression/ Univariate Linear Regression: When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression. The mathematical equation can be given as: $Y = \beta0 + \beta1*x$ Where

- Y is the response or the target variable
- x is the independent feature
- β1 is the coefficient of x
- β0 is the intercept

β0 and β1 are the model coefficients (or weights). To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable such as Sales!

Multivariate Linear Regression: Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable. Multiple Linear Regression Formula

$$yi = \beta0 + \beta1xi1 + \beta2xi2 + \ldots + \beta pxip + \epsilon$$

Where:

- yi is the dependent or predicted variable
- β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
- β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
- βp is the slope coefficient for each independent variable
- ϵ is the model's random error (residual) term.

Assumptions:

Multi – collinearity: Linear regression model assumes that there is very little or no-multi-collinearity in the data. It occurs when the independent variables or features have dependency in them.

Auto-Correlation: Very little or no auto-correlation in the data, the auto-correlation happens when there is dependency between residual errors.

Relationship between variables : relationship between response and feature variable must be linear

Normality of error iterms: Error terms should be normally distributed

Homoscedasticity: There should no visible pattern in residual values.

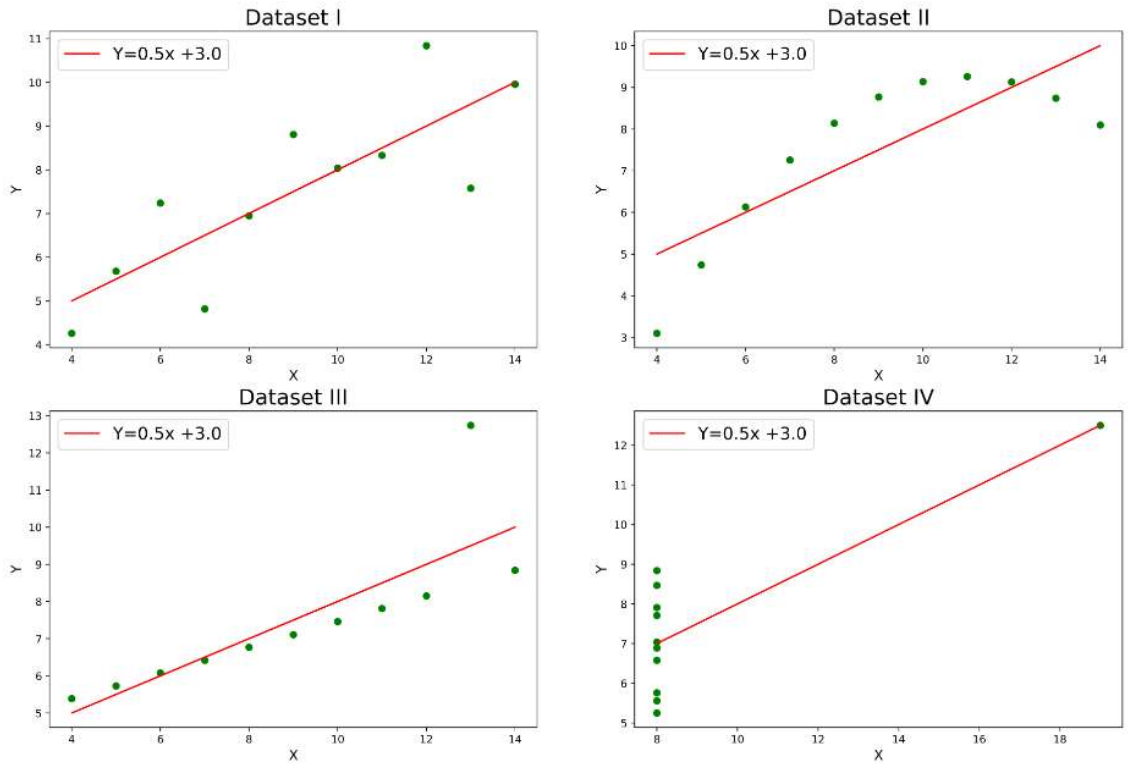
2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The four datasets of **Anscombe's quartet.**

I		II		III		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 Plot

Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this output:

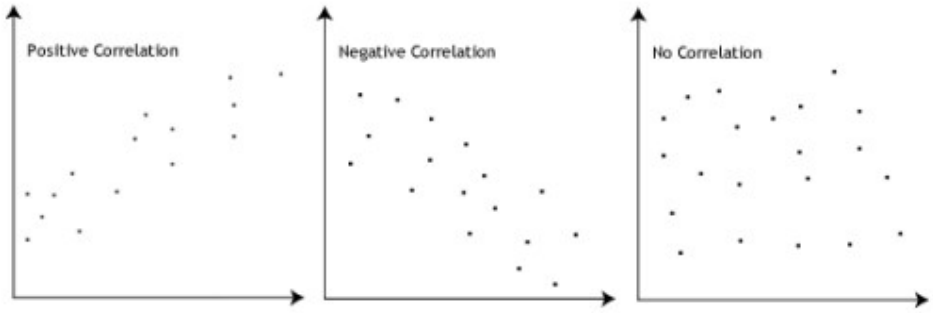
- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3. What is Pearson’s R? (4 marks) (3 marks) (3 marks)

Ans

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below: 4. What is scaling? Why is scaling performed? What is the difference b



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

Ans:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which 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)

Ans:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.