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?**

Ans:

I have performed analaysis and visulaization for categorical variables for bike sharing data using box plaot and bar chart for year 2018 and 2019 and based on that following can be summarized:

- ✓ Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- ✓ Most of the bookings has been done during the month of may, june, july, aug, sep and oct.Trend increased starting of the year tillmid of the year and then it started decreasing as we approached the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
- ✓ Clear weather attracted more booking which seems obvious. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.
- ✓ Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- ✓ When its not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- ✓ Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.

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

Ans:

In linear regression, using drop_first=True when creating dummy variables helps to avoid the problem of multicollinearity. Multicollinearity occurs when two or more predictor variables are highly correlated, leading to unstable and unreliable coefficients.

For example,

- ◆ consider a dataset with a categorical variable "Color" with three categories: Red, Green, and Blue.
- ◆ To include this variable in a linear regression model, you need to create 3 dummy variables: "Color_Red", "Color_Green", and "Color_Blue".
- ◆ If you don't use drop_first=True, all 3 dummy variables will be included in the model.
- ◆ However, if you use drop_first=True, only two dummy variables, "Color_Green" and "Color_Blue", will be included, preventing multicollinearity and improving the stability and interpretability of the coefficients.

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

Ans:

temp and atemp variables are highly correlated with target variable cnt

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

Ans:

I have validated the assumption of Linear Regression Model based on below **5** assumptions

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

✓ **Independence of residuals**: 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?

Ans:

Demand of bikes depend on following top 3 features:
-> atemp
-> sep
-> winter.

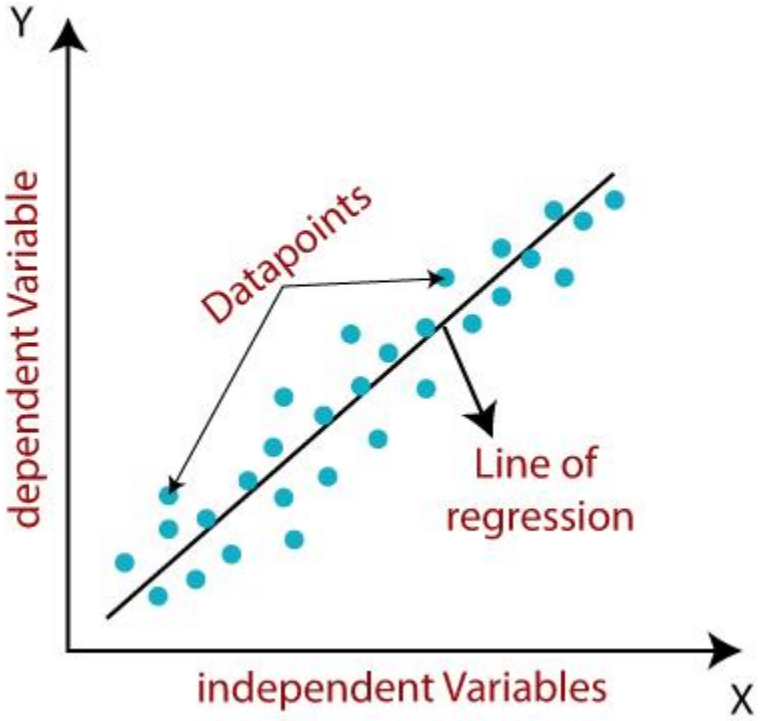
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans:

Linear Regression is a supervised learning algorithm that is used to predict a continuous outcome variable (also called dependent variable) based on one or more predictor variables (also called independent variables). It assumes a linear relationship between the outcome and the predictor variables.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Here's the algorithm in detail:

1. Collect the data: This involves gathering a dataset that contains the predictor variables and the outcome variable.
2. Determine the type of regression: Simple linear regression is used when there is one predictor variable and multiple linear regression is used when there are multiple predictor variables.
3. Prepare the data: This involves cleaning and transforming the data, if necessary, to ensure that it is ready for analysis.

4. Choose a model: Linear regression models can be fit using various algorithms such as ordinary least squares, gradient descent, and others.
5. Estimate the model parameters: The objective of this step is to determine the coefficients (weights) for each predictor variable that best predict the outcome variable. This is done by minimizing the sum of the squared differences between the predicted and actual values of the outcome variable.
6. Validate the model: This step involves evaluating the accuracy of the model by using metrics such as mean squared error, R-squared, and others.
7. Use the model: If the model is deemed accurate, it can be used to make predictions on new data by using the estimated coefficients.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = /fracn/sumxy - (/sumx)(/sumy)n/sumx^2 - (/sumx)^2$$

$$a(intercept) = /fracn/sumy - b(/sumx)n$$

Here, x and y are two variables on the regression line.

b = Slope of the line
a = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset

The following are some assumptions about dataset that is made by Linear Regression model
1. **Multi-collinearity**

✓ Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2. **Auto-correlation**

✓ Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
3. **Relationship between variables**

✓ Linear regression model assumes that the relationship between response and feature variables must be linear.
4. **Normality of error terms**

✓ Error terms should be normally distributed
5. **Homoscedasticity**

✓ There should be no visible pattern in residual values.

The goodness of fit of the linear regression model can be evaluated by several metrics, including R-squared, mean squared error (MSE), and adjusted R-squared.

- ✓ **R-squared** is a measure of the proportion of variation in the outcome variable that is explained by the predictor variables. A higher R-squared value indicates a better fit of the model to the data.
- ✓ **MSE** is a measure of the average squared difference between the predicted and actual values of the outcome variable. A lower MSE value indicates a better fit of the model to the data.
- ✓ **Adjusted R-squared** takes into account the number of predictor variables in the model and adjusts the R-squared value accordingly. It provides a more accurate measure of the fit of the model when there are multiple predictor variables.

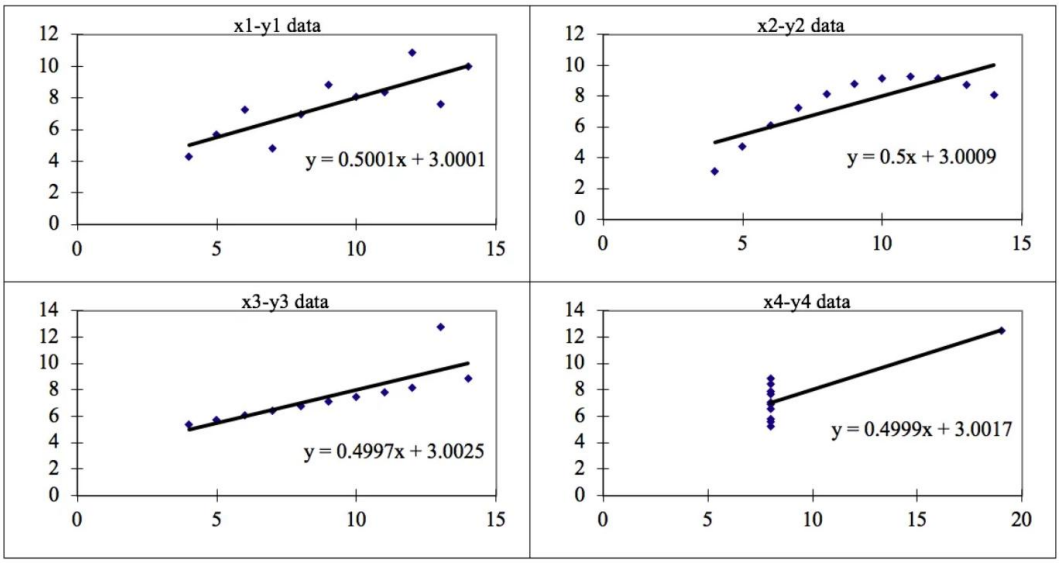
2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a set of four datasets that were created by Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. The four datasets have nearly identical descriptive statistics, such as the mean and variance of the x and y variables, yet they have vastly different distributions and relationships between the variables.

Each of the four datasets consists of 11 pairs of (x, y) data points. When plotted on a scatter plot, the datasets look very different from one another despite having similar summary statistics.

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression mode

In conclusion, Anscombe's quartet serves as a reminder of the importance of visualizing data before analyzing it. Descriptive statistics can provide a useful summary of the data, but they do not always accurately reflect the relationships and patterns within the data. Therefore, it is crucial to create

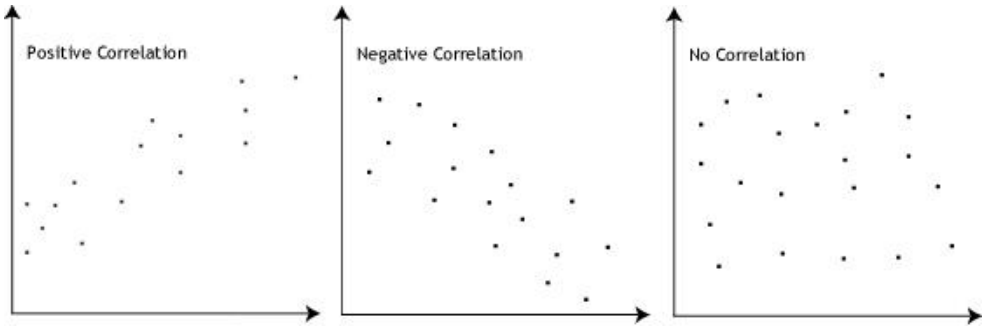
visualizations such as scatter plots, histograms, and box plots to better understand the structure and properties of the data.

3. What is Pearson's R?

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 variableassociated 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 thediagram below:



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

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?

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.

Ans:

A Q-Q (Quantile-Quantile) plot is a graphical representation of the comparison between two sets of quantiles, usually between the observed values and expected values of a dataset. In the context of linear regression, a Q-Q plot is often used to assess the normality of residuals, which are the differences between the observed and predicted values of the response variable.

The purpose of a Q-Q plot is to check whether the residuals are approximately normally distributed, which is an assumption of many statistical methods, including linear regression. If the residuals are normally distributed, the points in the Q-Q plot should lie approximately along a straight line. Deviations from normality, such as skewness or heavy tails, can be identified as deviations from the straight line in the Q-Q plot.

The importance of checking for normality in the residuals is that many statistical methods, including linear regression, assume that the residuals are normally distributed. If this assumption is not met, the results of the regression analysis can be unreliable, and alternative methods should be considered. A Q-Q plot is an easy and effective way to check for normality in the residuals, and it is an important step in the process of performing a regression analysis.

In summary, a Q-Q plot is a graphical representation of the comparison between the observed and expected values of a dataset, used to assess the normality of the data. In the context of linear regression, it is used to assess the normality of the residuals, which is an important assumption in many statistical methods