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)

   The categorical variables in the dataset are:

   - Weather: Clear conditions correlate with increased bike rentals.
   - Season: Higher rentals in summer and fall.
   - Weekday: Similar counts on all weekdays.
   - Holiday: Lower rentals on holidays.
   - Year: More rentals in 2019.
   - Month: Peak rentals in September.

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

   Using drop_first=True in dummy variable creation is crucial as it minimizes redundancy and mitigates multicollinearity, preventing the introduction of noisy data that could negatively impact the model-building process.

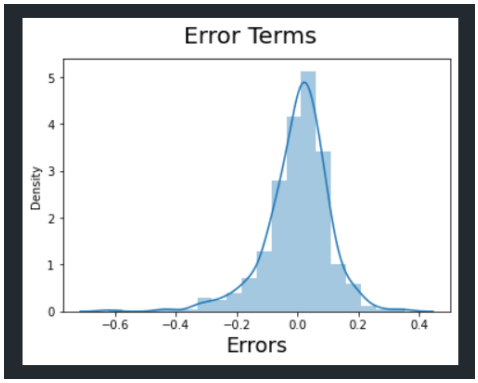
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   We can see that **temperature** has the highest correlation with target variable **count**.

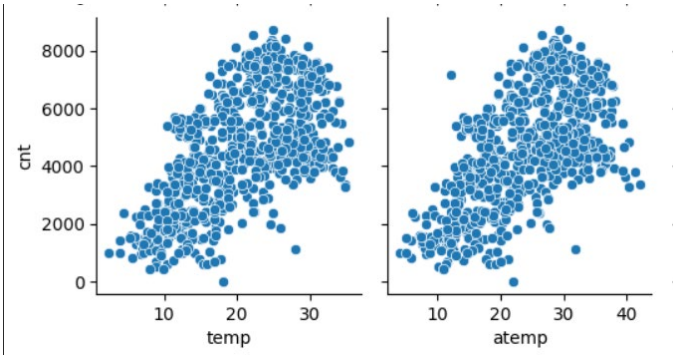
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   The assumptions of linear regression is that a linear relationship exists between the independent variables (X) and the dependent variable (y), where y represents the dependent variable Count ('cnt') in our case. We validate these assumptions by:

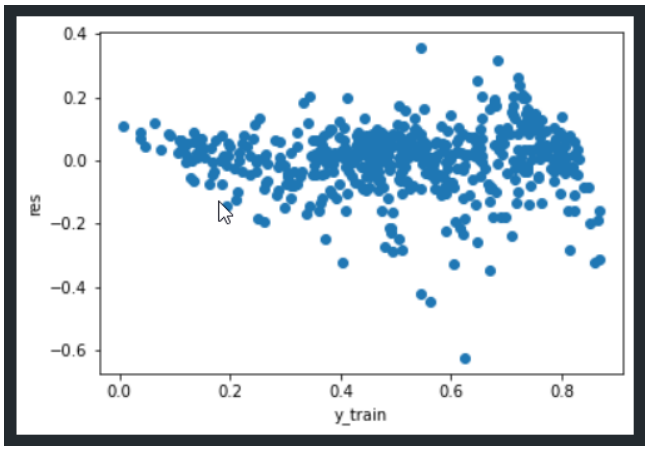
   - Ensuring a normal distribution of normal errors, by making sure error residuals are normally distributed:



- By checking the scatter plots we can tell whether independent variable is linearly related to the dependent variable:



- Homoscedasicty is another important way to justify linear regression assumptions by plotting residuals against y_test data:



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

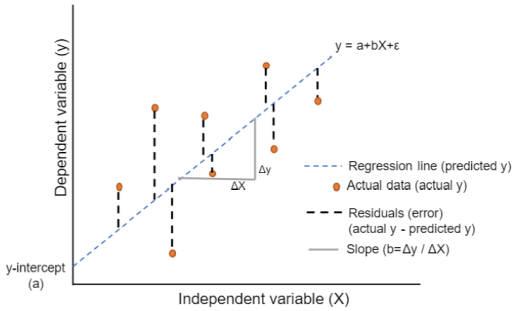
The primary factors significantly influencing the rise in shared bike demand are:

- Temperature (temp)
- Year (yr)
- Winter season (winter)

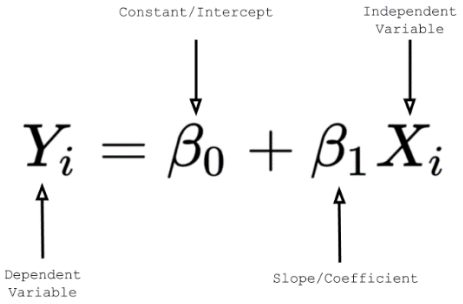
General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables). The algorithm assumes a linear relationship between the predictors and the outcome.



With Simple Linear Regression we can form a equation for the dependent variable Y(i) such as :



Here our aim is to predict the value of a dependent variable based on one or more independent variables. Assuming that there is a linear relationship between independent and dependent variables. For Multiple Linear Regression we put in i = 1 to n for the multiple predictors we have.

We then proceed to form a Objective Function (Cost Function) who's aim is to:

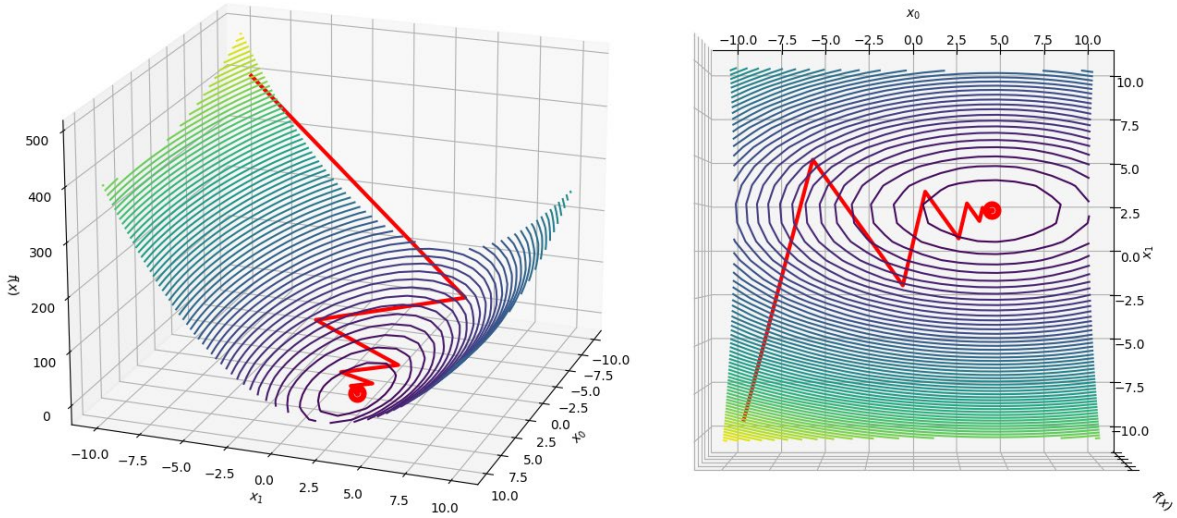
- Minimize the difference between predicted and actual values.
- The commonly used objective function is the Mean Squared Error (MSE), which calculates the average squared difference between predicted and actual values.

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2$$

Replace $y_{i\,pred}$ with $mx_i + c$

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

We then perform a Gradient Descent where we optimize the coefficients by iteratively updating them to minimize the cost function. The algorithm adjusts the coefficients in the direction of steepest descent.



Training the Model consists of splitting the dataset into training and testing sets. The model is trained on the training set to learn the coefficients. Prediction is made by using the learned coefficients to predict the outcome variable for new or unseen data.

Evaluation can be done to assess the model's performance using metrics like R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

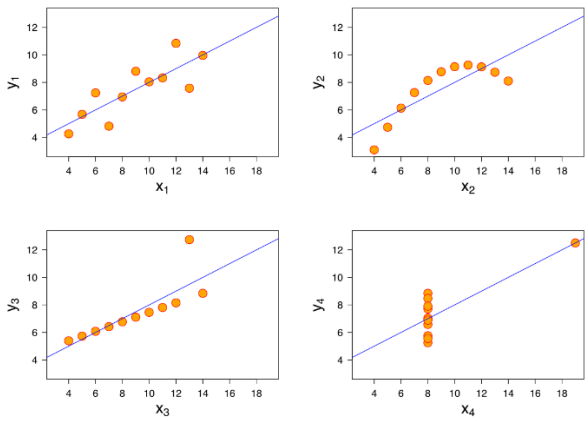
Linear Regression is commonly used in fields like economics, finance, biology, and social sciences for predicting numerical outcomes. Linear regression provides a simple and interpretable model, making it a foundational algorithm in statistical modeling and machine learning.

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

Anscombe's quartet consists of four datasets, each with eleven (x, y) points, meticulously crafted by statistician Francis Anscombe in 1973.

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

Despite their nearly identical basic statistical properties, the datasets exhibit striking graphical differences.

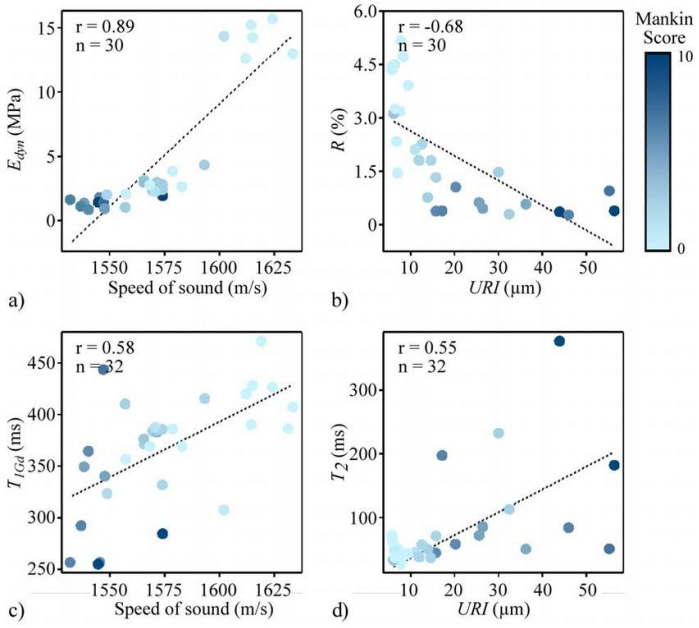


Anscombe designed these quartets to underscore the significance of visualizing data before analysis and to highlight the impact of outliers on statistical properties. This quartet emphasizes the critical need for plotting data features, aiding in the identification of anomalies like outliers, data distribution, and linear separability. It underscores that linear regression is suitable only for datasets with linear relationships, illustrating the importance of thorough data exploration.

3. What is Pearson's R? (3 marks)

Pearson's R, or the correlation coefficient, quantifies the degree of a linear association between two variables. A positive correlation (R > 0) implies variables moving together, while a negative correlation (R < 0) suggests opposite directions. R ranges from -1 to +1, where -1 signifies a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and 0 denotes no linear association. Scatter plots visually convey the strength and direction of the correlation, with points aligning more closely along a line as R increases.

For example look at the following scatter plots for a visual understanding :



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

Scaling is a crucial data preprocessing step which is aimed at normalizing data to a specific range for independent variables. When datasets have varying units and value ranges, algorithms might misinterpret the results, as they focus on values rather than units. Scaling ensures uniform magnitudes across all variables, preventing biases in the analysis.

Scaling is performed to mitigate issues arising from disparate units and ranges within a dataset. Algorithms can be sensitive to these variations, leading to unreliable results. By scaling, we can bring all variables to a consistent magnitude, enabling algorithms to use the data more accurately and preventing skewed outcomes.

Difference between Normalized Scaling and Standardized Scaling are :

Normalized Scaling (Min-Max Scaling):

- Uses the minimum and maximum values of features for scaling.
- Scales values between 0 and 1 (or -1 and 1).
- Highly influenced by outliers.
- Utilizes MinMaxScaler from sklearn for normalization.

Standardized Scaling:

- Uses mean and standard deviation for scaling.
- Aims for zero mean and 1 standard deviation.
- Less affected by outliers.
- Employs StandardScaler from sklearn for normalization.

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

An infinite VIF occurs when two independent variables have a perfect correlation, resulting in an R-squared value of 1. The VIF is calculated as 1/(1-R2), leading to infinity in this scenario. This shows a severe issue of multicollinearity, indicating that one of the correlated variables should be removed to establish a reliable regression model.

In the given bike dataset, three variables—cnt, casual, and registered—result in an infinite VIF. This occurs because the count (cnt) is a linear combination of the casual and registered variables, expressed as cnt = registered + casual. The infinite VIF reflects the strong correlation and linear relationship among these three variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical method which can be used to assess whether a given dataset follows a particular theoretical distribution, such as the normal distribution. When performing linear regression, Q-Q plots are valuable for checking the assumption of normality of residuals. Residuals are the differences between observed and predicted values in the regression analysis.

Here's how Q-Q plots are utilized in linear regression:

1. Q-Q plots compare the quantiles of the observed residuals against the quantiles of the normal distribution.

2. By examining the plot, we can identify deviations from the expected linear pattern. Departures from a straight line may indicate non-normality in the residuals.

3. A roughly straight Q-Q plot suggests that residuals are approximately normally distributed.

4. If residuals are normally distributed the validity of statistical tests and confidence intervals derived can be confirmed.

Therefore Q-Q plots are an important tool in linear regression, helping analysts check the assumption of normality in residuals, which is essential for drawing valid conclusions.