Subjective Questions

Assignment-based Subjective Questions

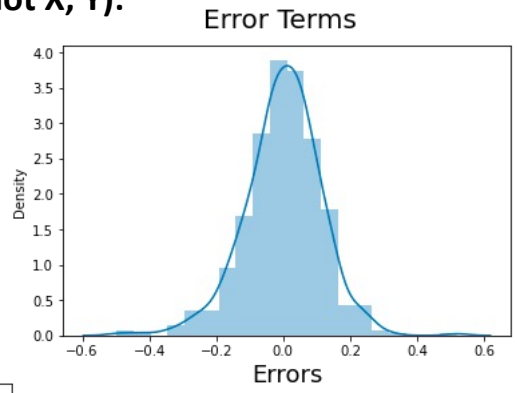
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   - '**season**' is a **strong driver variable** – maximum bikes are rented in the season 3(fall), followed by 2(summer), 4(winter), 1(spring)
   - '**weathersit**' is a **strong driver variable** – more bikes are rented in weathersit 1(Clear, Few clouds, Partly cloudy, Partly cloudy), followed by 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
   - People rented more bikes on an average in the **year 1(2019) than year 0(2018)**
   - People rented more bikes on the **months 7,9,6,8 than the other months**
   - People rented more bikes on **non-holidays than holidays**
   - '**weekday**' has little to no effect on the target variable
   - '**workingday**' has little to no effect on the target variable

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   - When we have a **categorical variable** with say '**n**' levels, the idea of dummy variable creation is to build '**n-1**' variables, indicating the levels
   - We can drop one of the levels as all the **information** can still be **retained**
   - In **pandas**, using "drop_first=True", informs the library to **drop the first level**

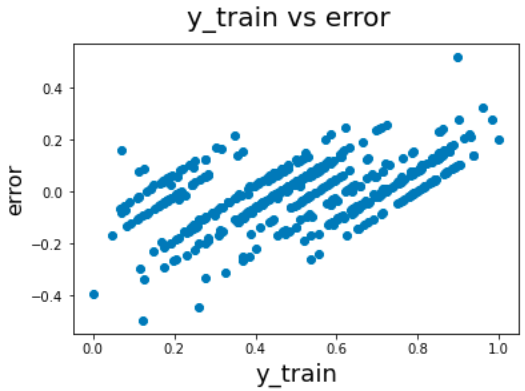
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - **atemp (0.65)**

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

- Assumption 1: **There is a *linear relationship* between X and Y:**
  i. By looking at the correlation values between feature variables, X and target Y
  ii. By making sure the p-values of the features are not significantly high in the linear models
- Assumption 2: **Error terms are *normally distributed* with mean zero(not X, Y):**
  I. By plotting the distribution of the errors



- Assumption 3: **Error terms are *independent* of each other:**
  i. By Plotting y_train vs error



- Assumption 4: **Error terms have *constant variance* (homoscedasticity):**
  - By plotting X vs y with the regression line
  - Not possible since the model is multi-linear and the regression line is a hyperplane
- Assumption 5: No assumptions on the distribution of X or y
  - No assumptions of X or y were considered in model building

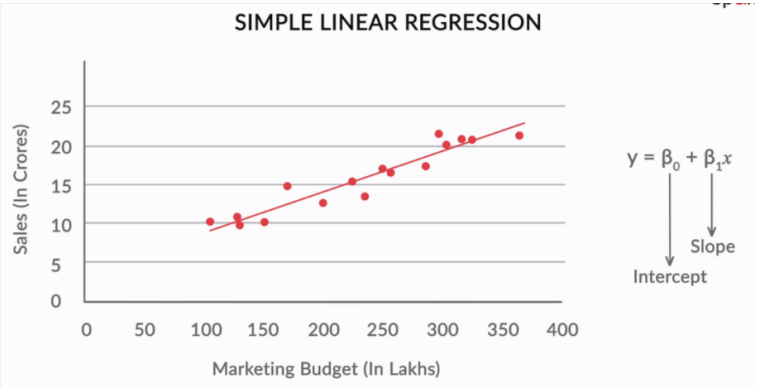
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- **season_3 (fall) - (0.3334)**
- **weathersit_3 (Light snow, …) - (-0.3089)**
- **season_2 (summer) - (0.2580)**

General Subjective Questions

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

- Belongs to Supervised learning technique
- As the name says, the output/target variable is continuous in nature
- A linear regression model attempts to explain the relationship between one or more dependent variables and an independent variable using a regression line
- The regression line or the best-fit line is a straight line in case of Simple Linear Regression and a hyperplane in case of Multiple Linear Regression.
- The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot
- A simple Linear Regression is given by the equation:



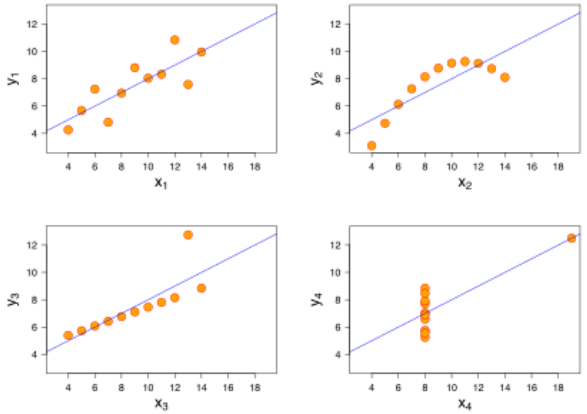
- A Multiple linear Regression is given by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- The strength of the linear regression model can be assessed using 2 metrics:
    1. R^2 or Coefficient of Determination
    2. Residual Standard Error (RSE)
- Linear Regression models can be interpreted using the coefficients
    - Ex: If X increases by a unit of 1, y increases by a unit of B(beta) 1

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

- **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics, yet have very **different distributions** and **appear very different when graphed**.
- Each dataset consists of eleven (*x*,*y*) points.
- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and **the effect of outliers** and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a **simple linear relationship**, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right); while a relationship between the two variables is obvious, it is **not linear**, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is **linear**, but should have a different regression line (a robust regression would have been called for). The calculated regression is **offset by the one outlier** which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when **one high-leverage point is enough to produce a high correlation coefficient**, even though the other data points do not indicate any relationship between the variables.

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

- In statistics, Pearson's *r*, or the correlation coefficient is a **measure of linear correlation** between two sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between **−1 and 1**.
- As with covariance itself, the measure can only reflect a **linear correlation** of variables, and ignores many other types of relationships or correlations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- A **positive value** means the two variables are **directly proportional** and a **negative value** means that they are **inversely proportional** to each other

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

- Scaling is a technique to **standardize** the independent features present in the data to 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.
- Also, in cases like MLR, **without scaling, it is impossible to understand & interpret multiple coefficients**
- Difference: In **normalized scaling**, the feature values are rescaled to **range between 0 and 1**, whereas in **standardized scaling**, the values are rescaled such that they are **centered around the mean with a unit standard deviation**, i.e., mean = 0 and standard deviation = 1.

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

- VIF calculates how well **one independent variable is explained by all the other independent variables combined**, excluding the target variable
- A large value of VIF indicates that there is a **correlation** between the variables (multi-collinearity)
- If there is **perfect correlation**, then VIF = infinity.
- It means the **variances of the feature are perfectly explained** by a combination of other independent features.

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

- In statistics, a **Q–Q plot** (**quantile-quantile plot**) is a probability plot,
- a graphical method for **comparing two probability distributions by plotting their *quantiles* against each other**.
- A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.
- If the two distributions being compared are **similar**, the points in the Q–Q plot will approximately lie on the identity line $y = x$. If the distributions are **linearly related**, the points in the Q–Q plot will approximately **lie on a line**, but not necessarily on the line $y = x$.
- Ex: