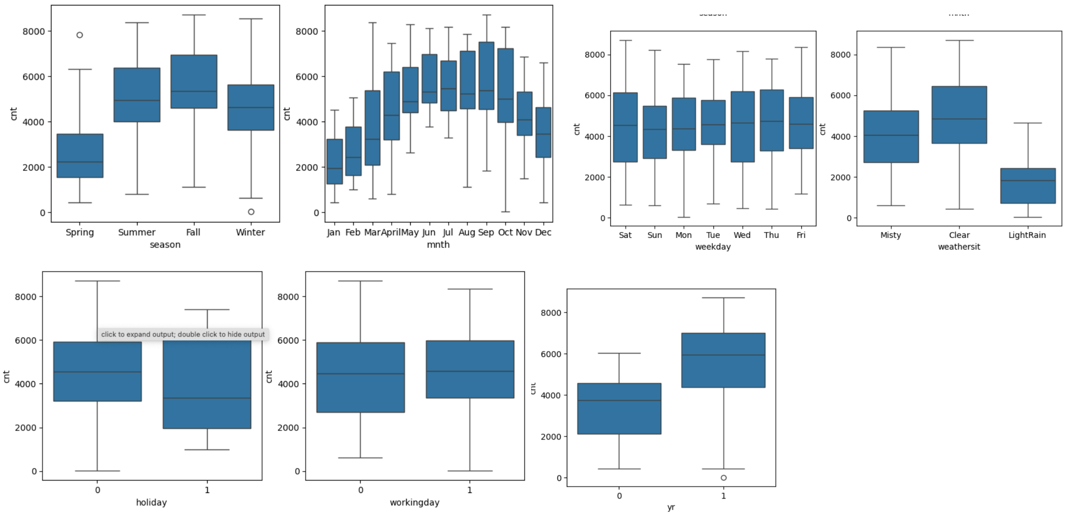
IIITB PGP Program Linear Regression Assignment
Yeshwanth Kumar Bapu Rajaram

**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)**
Here is my assessment of the Categorical variables:

- Season: Ridership / demand is less in Sprint compared to other seasons
- Month: Jun, Jul, Aug, Sep have higher demand compared to other months
- Weekday: Demand is almost similar across all days of the week
- Weather Situation: Naturally as we can expect, clear weather days have high demand
- Holiday: Holidays leads to drop in demand (atleast from median)
- Working Day: Median is almost same for whether it is a working day.
- Year: Year 2019 has seen a sharp rise in demand.



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

The objective of creating dummy variables is to convert categorical values into numerical representations. This is necessary to be able to interpret and infer statistically. So when we have a cat variable with lets say 3 categories, it is sufficient to have 2 dummy variables to represent them. For ex: if categories are A, B, C, we can represent the encoding of this with 2 dummy variables cat_A, cat_B as below.

| CategoryValue | cat_A | cat_B |
|---|---|---|
| A | 1 | 0 |
| B | 0 | 1 |
| C | 0 | 0 |

Thus in general if cat variable has ‘n’ categories, we will only need to generate ‘n-1’ dummy variable. When we don’t care about which category is dropped, we can user drop_first =True to retain n-1 dummy variables.

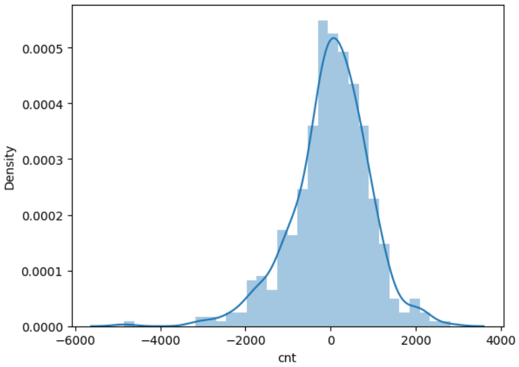
However if we need a specific dummy_variable to be dropped, we can set drop_first=False and manually dropped the desired dumm_variable.

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

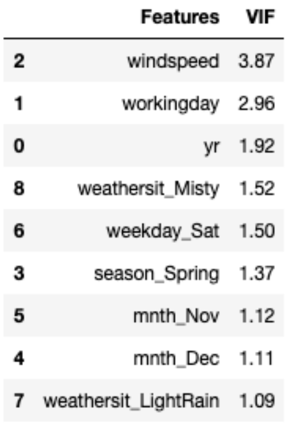
**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

I have validated the assumptions of Linear Regression on my model by the following assessment.

a. **Linearity:** The relationship between independent variables and target variable in linear. This was validated from the results, the residual analysis and visualizations indicating linear dependency between variables and target variable
b. **Normality of Residuals:** I have plotted a distplot of the residuals (y_train - y_train_pred) which indicate how the residuals or error terms are distributed. From the plot it is evident that the residuals are normally distributed and are centered around zero which is one of the assumptions of Linear Regression.

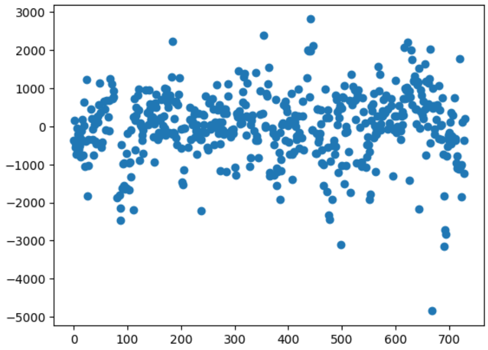


c. **Multicollinearity:** The multicollinearity between independent variables should not be significant. This was validated by checking the Variance Inflation Factor and ensuring that VIF is less than 5 for all independent variables.



| | Features | VIF |
|---|---|---|
| 2 | windspeed | 3.87 |
| 1 | workingday | 2.96 |
| 0 | yr | 1.92 |
| 8 | weathersit_Misty | 1.52 |
| 6 | weekday_Sat | 1.50 |
| 3 | season_Spring | 1.37 |
| 5 | mnth_Nov | 1.12 |
| 4 | mnth_Dec | 1.11 |
| 7 | weathersit_LightRain | 1.09 |

d. **Homoscedasticity:** Residuals or Error terms do not vary much as the value of the predictor variable changes. This was validated by plotting a scatterplot of residuals.



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

1. Spring Season with: -2649 coeff
2. LightRain weather situation: -2422.38 coeff
3. Year – 2053.81 coeff

### General Subjective Questions

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

Linear Regression is a statistical method used for modelling the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). LR algorithm is a supervised learning algorithm which is used when there is a linear relationship between predictors and target variable i.e., when the predictor variable change positively or negatively, the target variable also changes linearly with them.

**Mathematical linear relationship** can be represented with the following equation:

**y = mx + c** where y is target variable and x is predictor variable and c is the constant or intercept indicating the value of y when x =0.

**Postulation: y = linear combination of input x**

Generic model representation for multiple input variables:

Y=β0+β1X1+β2X2+...+βnXn+ϵ

where,

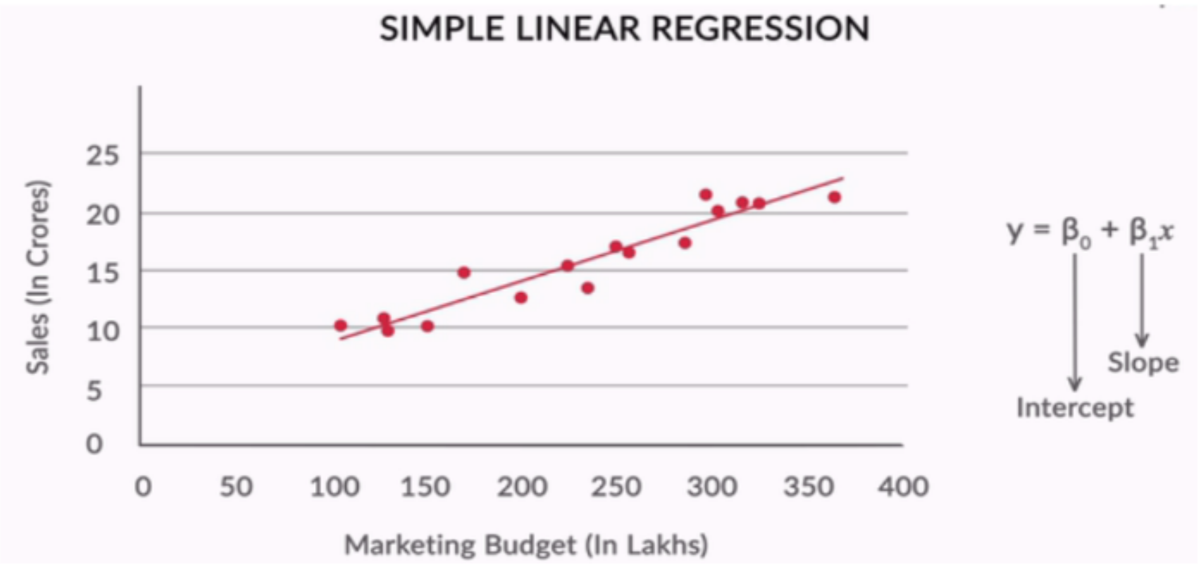
**Y** – target variable

**X1, X2, … Xn** – n predictor variables

**β, β1, β2 … βn** - indicate the coeffecients representing the weightage of each predictor variables

**ϵ -** represents the error terms which is added to account for the difference in observed and predicted values of y.

Sample 2 dimensional linear regression plot is as below.



**Objective of LR model:**

The objective of Linear Regression is to find the values of the coefficients so that minimize the difference between the actual and predicted values by the model.

The coefficients (β0, β1,.. βn) are estimated to find the best fit line.

This is achieved by minimising the error in predicting values.

Mean Square Error (MSE) is given by MSE $= \frac{1}{m}\sum_{1}^{n}(y_i - \hat{y}_i)^2$

To find the optimal fit, we will have to find the minima of this function.

It is evident that MSE is a convex function and to find minima of a convex function 'g' we can equate the derivative of function and equate to zero. When we consider multiple predictor variables it would be a vector and thus we will need to equate the gradient of the function to 0.

Converting to MSE in vector format: MSE $= \frac{1}{m}(Y - X\beta)^T(Y - X\beta)$

$solving\ for\ gradient(MSE) = 0$ , We get $\beta^* = (X^TX)^{-1}\,X^TY$

Where $\beta^*$ is the value of coefficient vector so as to keep the MSE at minima.

This forms the essence of a Linear Regression model.

**Assumptions**: Linear regression makes several assumptions:

- There is a linear relationship between the independent variables and the dependent variable.
- The residuals (the differences between the observed and predicted values) are normally distributed.
- There is homoscedasticity, meaning that the variance of the residuals is constant across all levels of the independent variables.
- There is no multicollinearity, meaning that the independent variables are not highly correlated with each other.

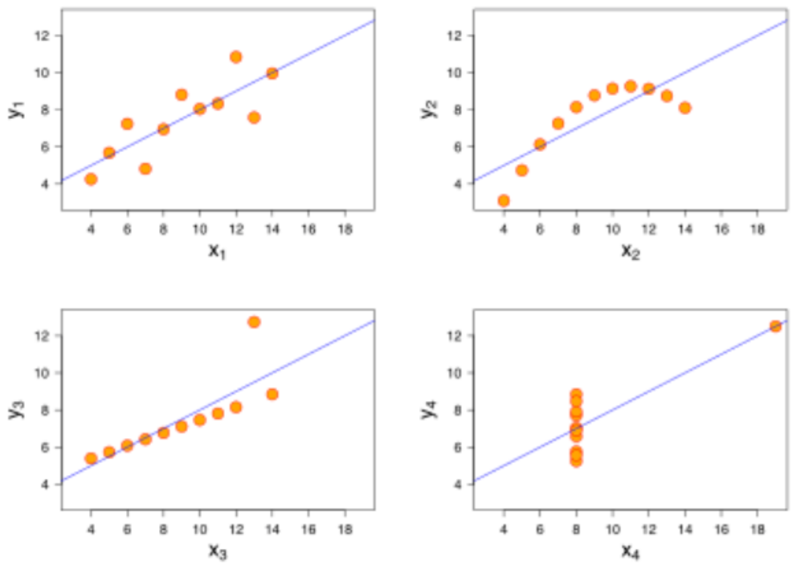
**2. Explain the Anscombe's quartet in detail. (3 marks)**

From my reading, I understand that Anscombe's quartet is a set of four datasets which were created to emphasize the importance of visual exploration of data to gain better insights.

Created by statistician Francis Anscombe in 1973, **Anscombe's quartet** is a set of four datasets that have nearly identical statistics. Each dataset consists of eleven data points (x,y) which have identical statistical properties as described below:

1. Exact mean,
2. Variance for x and y is approximately same
3. Corelation – is same (accurate to 3 decimal places)
4. All for have same linear regression line ($y = 3.00 + 0.500x$)

It is very surprising to see these datasets which tell us to give more emphasis on data visualization and not to solely rely on summary statistics.
In essence, summary statistics can provide useful insights into a dataset, however they may sufficient to understand the dataset and its complexities. Visual exploration through graphs or plots is necessary for understanding patterns, relationships, and outliers in data.



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

Pearson's $r$ (or Pearson correlation coefficient), is a measure of the linear relationship between two variables. It is a statistical measure of the strength and direction of the linear association between two continuous variables.

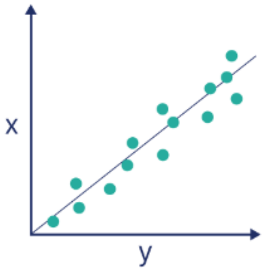
Pearson's $r$ r ranges from -1 to +1:

$r = 1$ indicates perfect positive correlation (as one variable increases, the other variable increases).
$r = -1$ indicates perfect negative correlation (as one variable increases, the other variable decreases).
$r = 0$ indicates No linear correlation (there is no linear relationship between the variables).

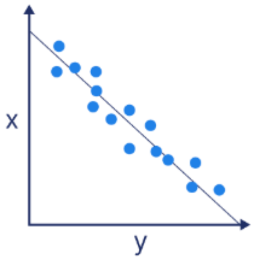
Pearson's $r$ r is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:
$x_i$ *and* $y_i$ are the individual data points.
$\bar{x}$ *and* $\bar{y}$ are the means of the x and y variables, respectively.
$n$ is the number of data points.



Positive corelation (r > 0)



Negative corelation (r < 0)

No corelation (r = 0)

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

**Scaling** is a preprocessing step in data analysis where the values of variables are transformed to fit within a specific range or distribution. It also helps in improving the speed / performance of the algorithms as the algorithms will have to work with numbers in smaller range compared to base data which can be huge number.

**Scaling is performed** to:

- Normalize data – ensure that all variables have the same scale or range. If we do not scale, variables with large range will over influence the predictions and will undermine the other variables with smaller range.

- Standardize data – by transforming the data to have a mean of 0 and standard deviation of 1. This makes the data more interpretable.

**Normalized and Standardized scaling:**
**In normalized scaling,** the data is transformed to fit within a specified range, usually between 0 and 1.

- **Formula-** $\boldsymbol{X_{normalized}} = \frac{\boldsymbol{X - X_{min}}}{\boldsymbol{X_{max} - X_{min}}}$, where $X_{max}$ is the maximum and $X_{min}$ is the minimum value of the variable X

**Standardized scaling** the data is transformed to have a mean of zero and a standard deviation of one.

- **Formula -** $\boldsymbol{X_{std}} = \frac{\boldsymbol{X - \mu}}{\boldsymbol{\sigma}}$ where $\mu$ is mean and $\sigma$ is standard deviation of variable X

One point to note about normalization over standardization is that normalization loses some information in the data, especially about outliers.

Normalization/standardization are designed to achieve a similar goal, which is to create features that have similar ranges to each other.

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

Yes, I have observed this while playing around with the model and variables. High VIF indicates high multicollinearity between variables and an infinite VIF indicates perfect multicollinearity among the independent variables.

Multicollinearity means that one or more of the independent variables can be expressed as a perfect linear combination of the others. In other words, the relationship between the variables is so strong that one variable can be exactly predicted from the others.

VIF is calculated by inversing the correlation matrix among the variables under consideration, when they are perfectly corelated, the matrix is non invertible leading to infinite VIF. Thus an infinite VIF occurs when there is perfect multicollinearity among the independent variables in a regression model, making it impossible to estimate the coefficients accurately.

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

Quantile-Quantile plot or Q-Q plot is a plot used to check whether a dataset follows a common probability distribution. Q-Q plots provide a visual assessment of how well a dataset conforms to a theoretical distribution such as the Normal Distribution.

**Steps -** sort dataset in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The x-axis represents the quantiles (percentiles) of the theoretical distribution. The y-axis represents the quantiles of the observed dataset. A quantile represents a value below which a certain percentage of the data falls.

Q-Q plots are valuable in linear regression for assessing the normality of the residuals (the differences between observed and predicted values). In linear regression, it is assumed that the residuals follow a normal distribution with mean zero. Deviations from this assumption can impact the validity of the regression analysis. By examining the Q-Q plot of the residuals, we can visually inspect whether the residuals are normally distributed.

If the points on the plot follow a straight line, it suggests that the residuals are approximately normally distributed, any other pattern in Q-Q plot indicates the probable presence of outliers, skewness, or other issues which mean that the assumptions of linear regression is violated. Thus Q-Q plot is an important tool to assess a Linear Regression Model.