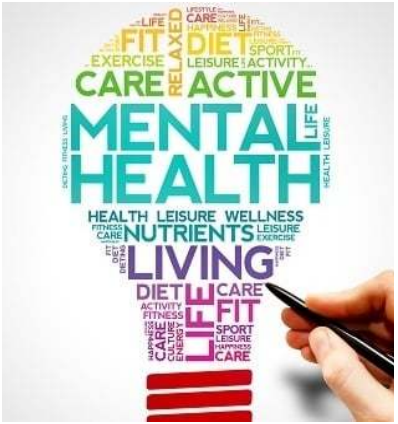
Medical Health Expenses

Understanding Problem Statement

- Predict the future medical expenses of patients based on certain features.
- Factors affecting the medical expenses of the patients:-
  - **Age**
  - **Gender**
  - **Body Mass Index**
  - **Region**
  - **Smoking Behaviour**

Business Implications of the Project

- Health is the center of everyone's life.
- Every part of our life relies on good health.
- Health is the extent of an individual's continuing physical, emotional, mental, and social ability to cope with the environment.



Univariate Analysis

- It involves only one variable.
- It is used to understand the distribution of the variables present in the dataset and derive meaningful insights from them.
- It can be used to check the distribution of both Numerical as well as Categorical variables.

Bivariate Analysis

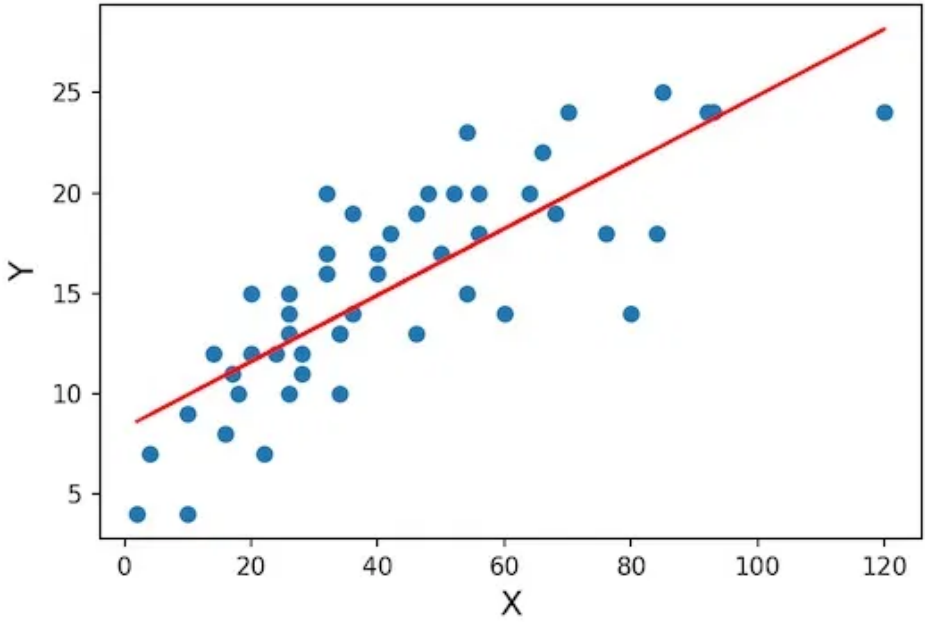
- Bivariate Analysis is one of the simplest forms of quantitative analysis.
- It involves analysis of two variables for determining the empirical relationship between them.
- It can be helpful in testing simple hypothesis of association.

Feature Scaling

- Feature Scaling is used to normalize the range of independent variables or features of data.
- Feature Scaling is also known as Data Normalization.
- It helps to normalize the data within a particular range.
- It also helps in speeding up the calculations in an algorithm.

Linear Regression

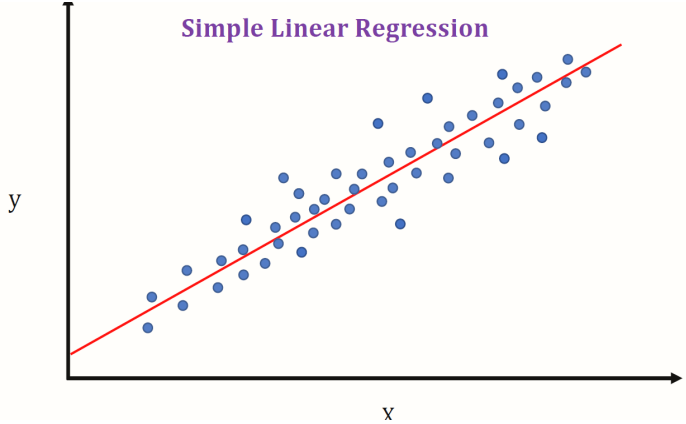
- Linear Regression is a linear approach of modelling the relationship between a dependent and one or more independent variables.



data is good

Assumptions of Linear Regression

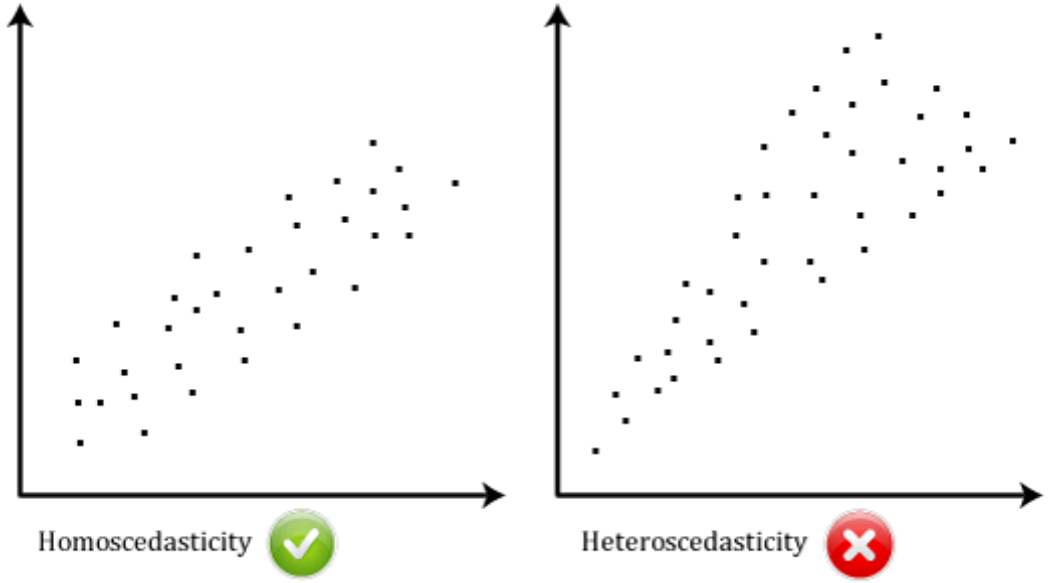
- **Linearity**
  - The relationship between the dependent and the independent variables must be linear.
  - Trend lines between two variables must be either in increasing or decreasing pattern.



data is good

Assumptions of Linear Regression

- **Homoscedasticity**
  - In statistical terms, the variance of all the variables must be same.



Assumptions of Linear Regression

- **Independence**
  - All the observations must be independent of each other.

- **Normality**
  - All the variables must follow a normal distribution.

Evaluation Metrics

- **R2 Score**
  - It is generally used to determine the strength of correlation between the independent features and the target column.
- **Root Mean Squared Error(RMSE)**
  - It is the square root of the mean of the differences between actual and the predicted values.

Random Forest

- Random forest is an ensemble learning method for classification and regression by constructing multiple number of decision trees at training time.
- And it outputs the average prediction of the individual trees in case of regression and mode of the classes in case of classification.

Gradient Boosting Model

- Gradient boosting works sequentially adding the previous predictors under fitted predictions to the ensemble.
- In case of Gradient boosting, ensembling happens sequentially.
- The model is built until and unless the errors are optimised in the best way.

Cross Validation

- Cross Validation is a resampling procedure which is used to evaluate the machine learning models on limited data samples.
- The goal of cross validation is to test the model's ability to predict new data.
- It has a single parameter called **k**.
- **k** indicates the number of groups the data would be split into.

More things to try

- We can give different labels to each Region. It might give better results.
- Keep 4 and 5 number of children in our analysis instead of capping them and see how the results vary.
- We can try some more predictive models and compare the results.
- Try converting Expense column to a normal distribution using log or square root transformation.

Major Takeaways from the project

- Understood the importance of Data Analysis and Data Visualization for determining the association between features.
- How to build intuition by building insights from Data Visualization.
- How to deal with categorical variables.
- Learnt about different Assumptions to be satisfied before using a Linear Regression Model.

Major Takeaways from the project

- Learned different predictive models such as Linear Regression, Random Forest and Gradient Boosting model.
- Learned how to ensemble different models.
- Understood the importance of cross validation on the data.
- Learned the importance of comparing different predictive models.