EXPLORATORY DATA ANALYSIS

By:- SHOBHIT TYAGI

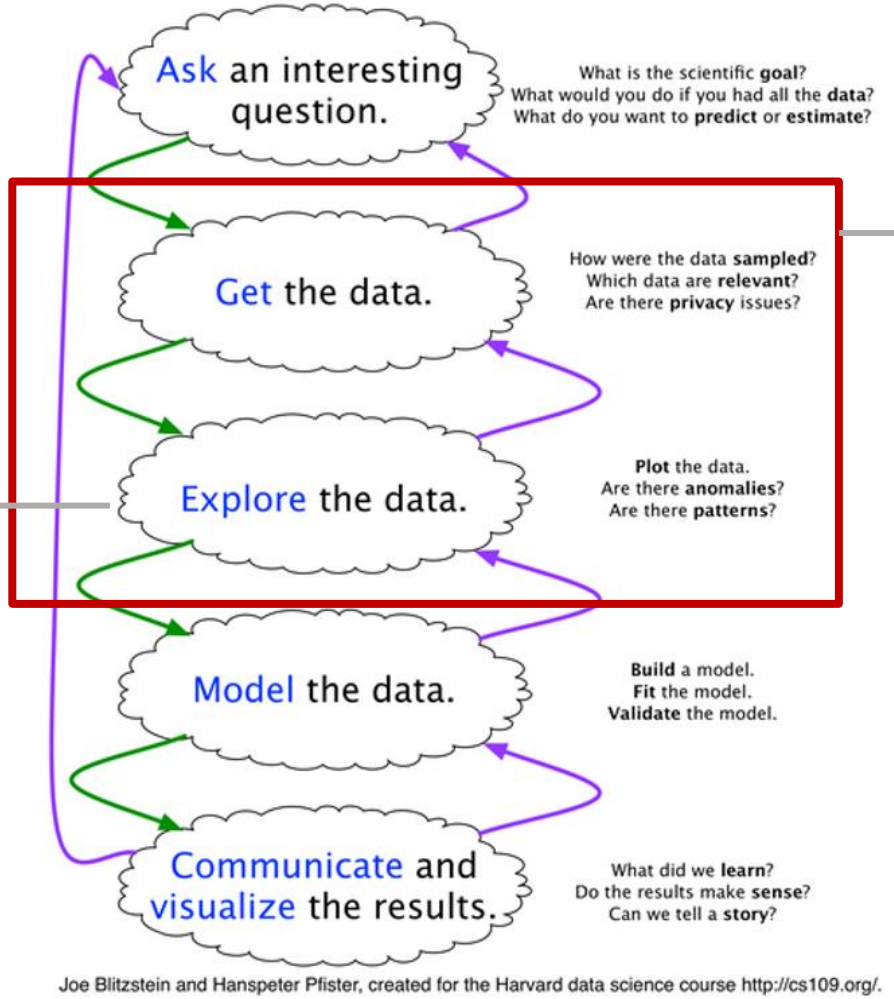
Contents

- Data Types
- Data Analysis Steps
    1) Variable Identification
    2) Univariate Analysis
    3) Bi-variate Analysis
    4) Missing value treatment
    5) Outlier Treatment
    6) Variable transformation
    7) Encoding

Data Preparation/Exploration

Data Exploration

- Descriptive Statistics
- Graphs



www.quora.com

Data Preparation

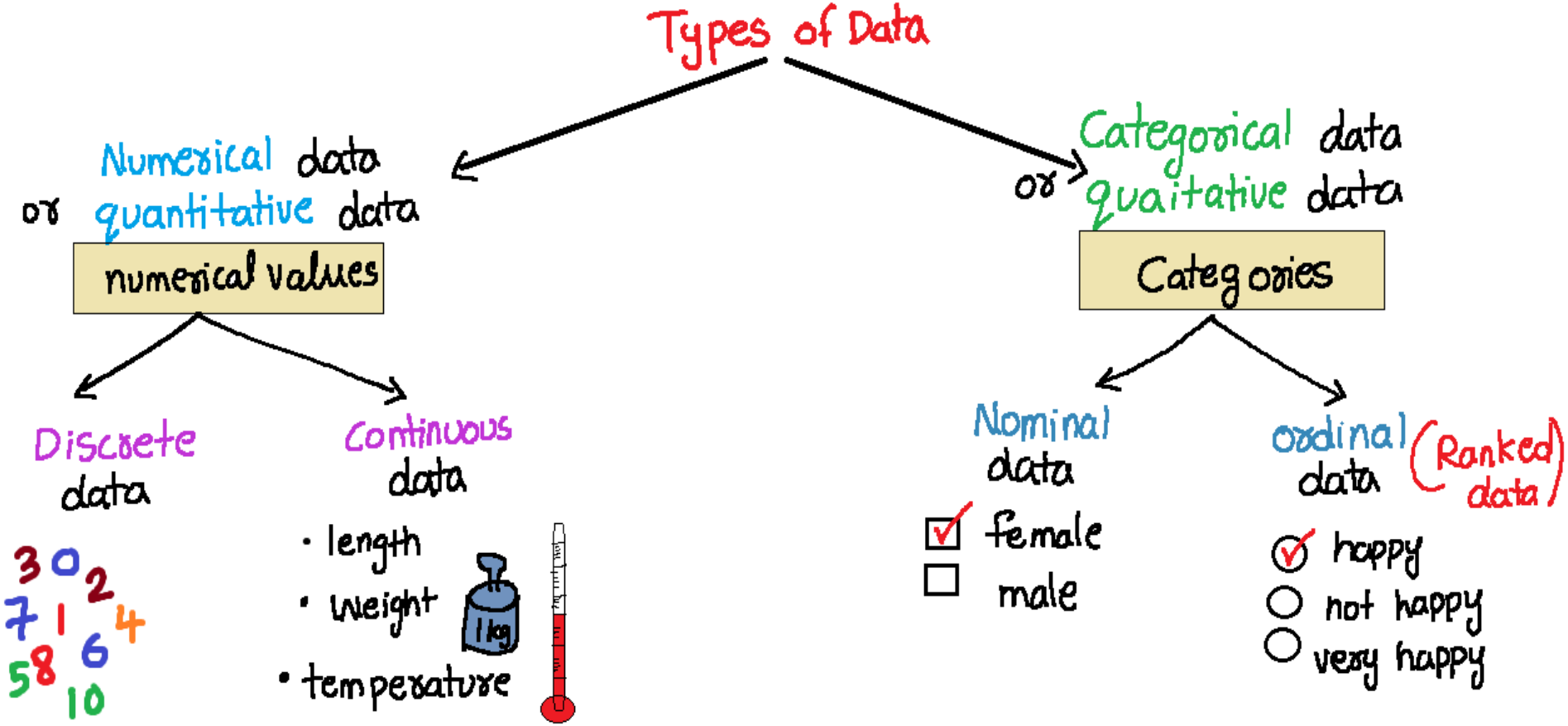
- Ensure Data Quality
- Data Representation

Steps Involved:

1) Variable Identification
2) Univariate Analysis
3) Bi-variate Analysis
4) Missing value treatment
5) Outlier Treatment
6) Variable transformation
7) Variable creation

Data Types

Data Types

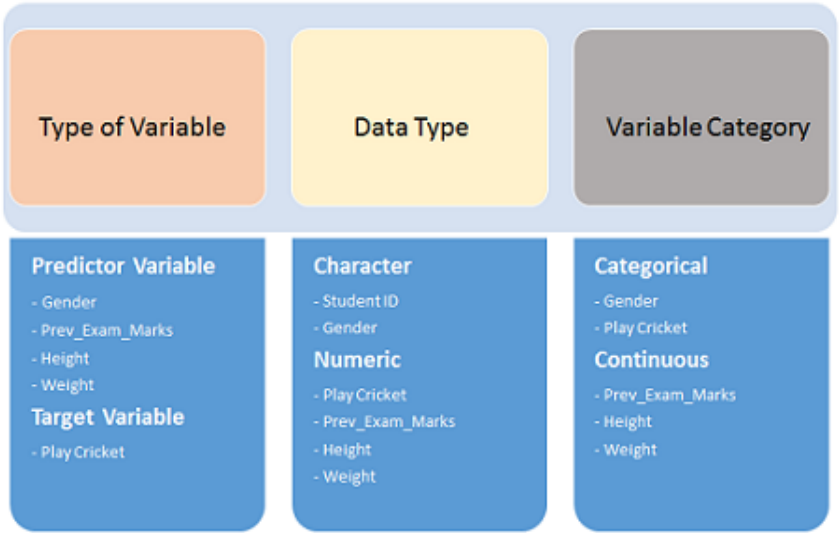

Types of Data
Numerical data
or quantitative data
numerical values
Discrete data
3 0 2
7 1 4
5 8 6
10
Continuous data
• length
• weight
1kg
• temperature
Categorical data
or quaitative data
Categories
Nominal data
female
male
ordinal data
(Ranked data)
happy
not happy
very happy

Steps1: Variable Identification

- First, identify **Predictor (Input)** and **Target (output)** variables. Next, identify the data type and category of the variables.

- **Example:-** Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables.

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



Steps2 (Univariate Analysis)- Step3 (Bivariate Analysis)

- **Univariate Analysis:**
  - We explore variable **one by one**.
  - Univariate analysis is also used to highlight missing and outlier values.

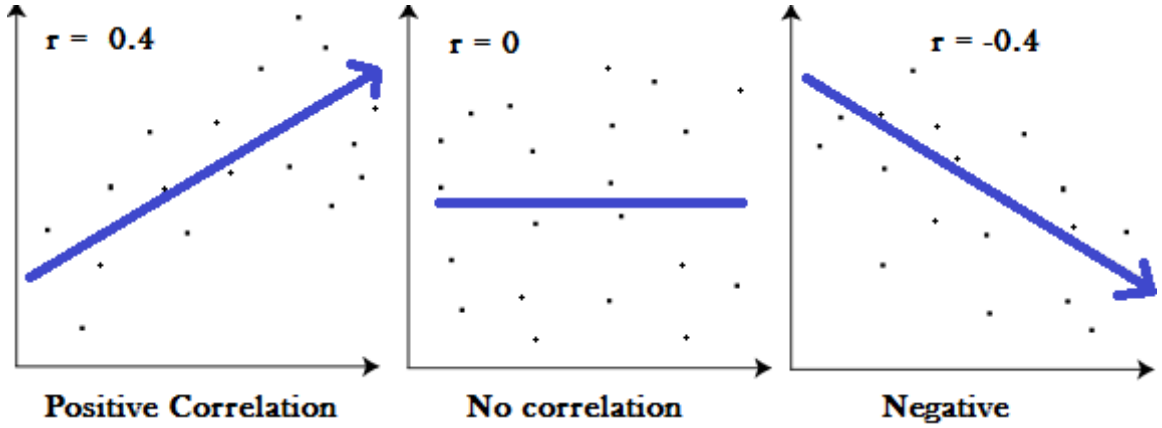
- **Bivariate Analysis:**
  - It finds out the relationship between **two variable**.
  - Different methods are used like Scatter Plot, Correlation values, two-way tables etc.

- **Note: The above methods help us to identify the relationship between variables.**

Bi-variate Analysis : Correlation

- Statistical technique that shows how strongly pairs of variables are related.

- E.g. Height and Weight where taller people tend to be heavier than shorter people.



Steps4: Missing Value Treatment (Ensure Data Quality)

- **Why missing value treatment is required?**
  - Missing data in the training data set can reduce the power / fit of a model.
  - It can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly.
  - It can lead to wrong prediction or classification.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Not Treated Missing Values

The **inference** from this data set is that the chances of playing cricket by males is higher than females.

After Missing Value Treatment

After treatment of missing value(based on gender) we can see that female have higher chances of playing cricket compared to males

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Steps4: Missing Value Treatment(Contd.)

- **Which are the methods to treat missing values?**

1) Deletion:

1) We delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

2) Mean/Mode/Median Imputation:

I. Imputation is a method to fill in the missing values with estimated ones.

II. The objective is to employ known relationships that can be identified in the val ɔ assist in estimating the missing values.

III. It can be of two types:

I. **Generalized Imputation:** In this case, we calculate the mean or n values of that variable then replace missing value with mean or median.

II. **Similar case Imputation:** In this case, we calculate average for gender "Male" (29.75) and "Female" (25) individually of non missing values then replace the missing value based on gender.

Steps4: Missing Value Treatment(Contd.)

- **Which are the methods to treat missing values?**

3) Prediction Model:

I) We create a predictive model to estimate values that will substitute the missing data. The value can be determined using regression, decision tree etc.

4) KNN Imputation:

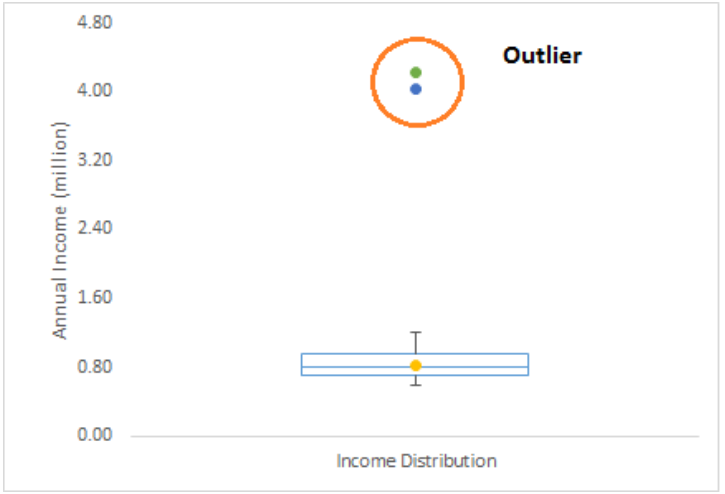
I. The process of replacing attributed values from the available data is known as Imputation.

II. The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing.

III. The similarity of two attributes is determined using a distance function.

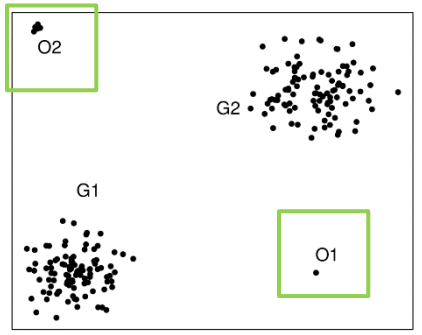
Steps5: Outlier Treatment

- **What is an Outlier?**
  - Outlier is an observation that appears far away and diverges from an overall pattern in a sample.
  - In other words, It is an observation that lies an abnormal distance from other values in a random sample from a population.
  - **Example:** we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population.

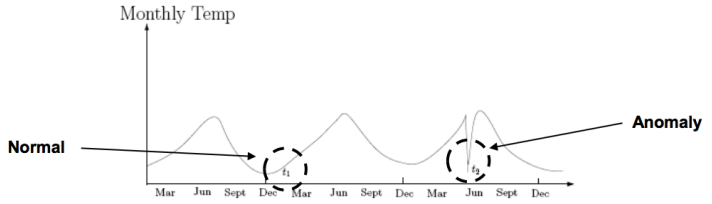


Steps5: Outlier Treatment(Contd.)

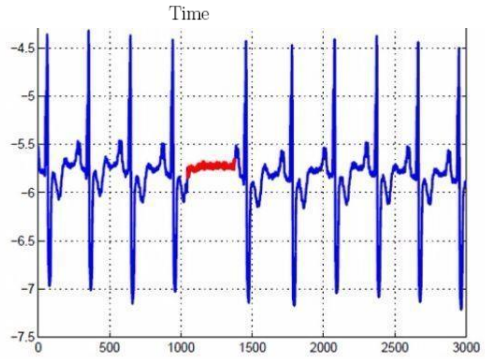
- **Types of Outlier?**
  - There are three kinds of Outliers.
    - **a. Global outlier (Point Anomaly)**
      - i. If an object significantly deviates from the rest of the data
      - **ii. Example**: Figure(1)



    - **b. Contextual Outlier**
      - i. If an individual data instance is anomalous in a specific context (but not otherwise)
      - **i. Example:** 80 F in Urbana – outlier? (depending upon summer or winter



    - **c. Collective Outlier**
      - i. A subset of data objects collectively deviates significantly from the whole data set.
      - ii. The individual data points inside the collective outlier may not be outliers by themse occurrence together as a collection is anomalous
      - iii. Example:



Steps5: Outlier Treatment(Contd.)

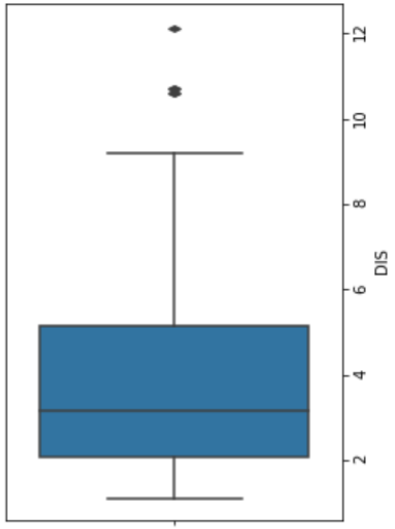
- **What is impact of Outlier?**
  - Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:
  - If the outliers are non-randomly distributed, they can decrease normality
  - They can bias or influence estimates that may be of substantive interest

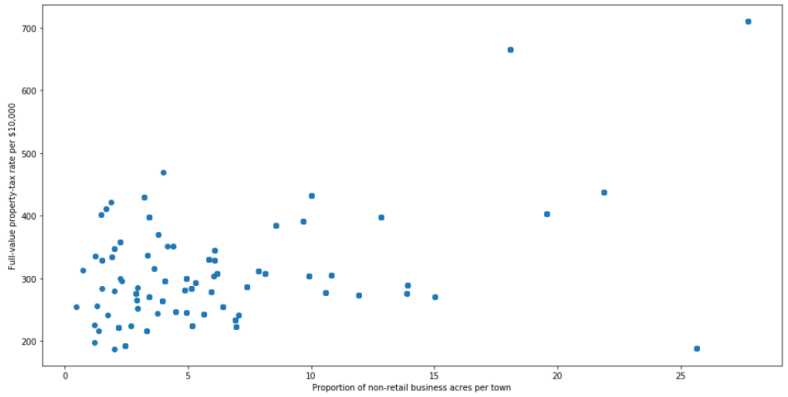
Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

As you can see, data set with outliers has significantly different mean and standard deviation. In the first scenario, we will say that average is 5.45. But with the outlier, average score to 30.
This would change the estimate completely.

Steps5: Outlier Treatment(Contd.)

- **How to detect Outlier?**
  - Most commonly method used is visualization. Like **Box-plot, Histogram, Scatter plot**.
  - Some analysts use some thumb rule to detect outliers.





Steps5: Outlier Treatment(Contd.)

- **How to remove outliers?**

1) Deletion:

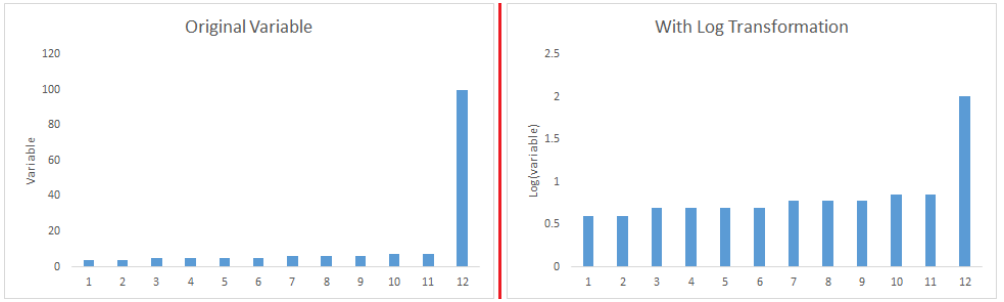
I. We delete outliers value if it is due to data entry error, data processing error or outlier observations are very small in numbers.

II. We can also use trimming at both ends to remove outliers.

III. But it is not acceptable to drop observations because there can be legitimate observations & sometime interesting too.

2) Transforming:

I. Transforming variables can also eliminate outliers**. Natural log** of a value reduces the variation caused by extreme values.



Step6: Variable Transformation - Scaling

- Usually numerical features don't have certain range and differ from each other.
  - E.g. Age and income can never be in same range.

- Scaling helps to bring all features to same range. It can be divided in two ways-
  - **Normalization**: Scales all values in a fixed range between 0 and 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

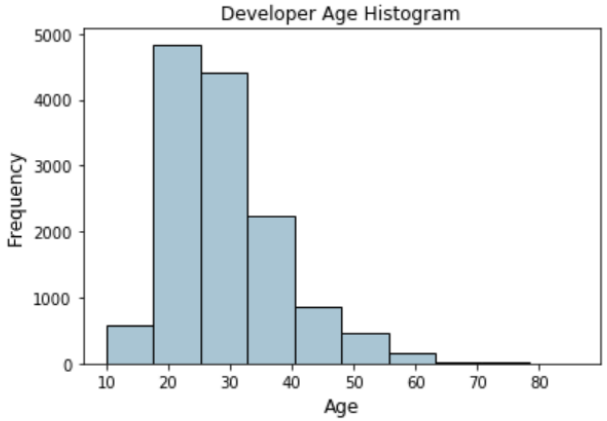
  - **Standardization**: Scales the values while taking into account standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

Step6: Variable Transformation - Binning

- Transforming continuous numeric features into discrete ones
- Makes the model robust and prevent overfitting.

Numerical Data



Histogram depicting developer age distribution

```
Age Range : Bin
---------------
 0 -  15  : 1
16 -  30  : 2
31 -  45  : 3
46 -  60  : 4
61 -  75  : 5
75 - 100  : 6
```

Categorical Data

```
#Categorical Binning Example

Value      Bin
Spain  ->  Europe
Italy  ->  Europe
Chile  ->  South America
Brazil ->  South America
```

Label Encoding & One Hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Thank you