TMA Innovation Center

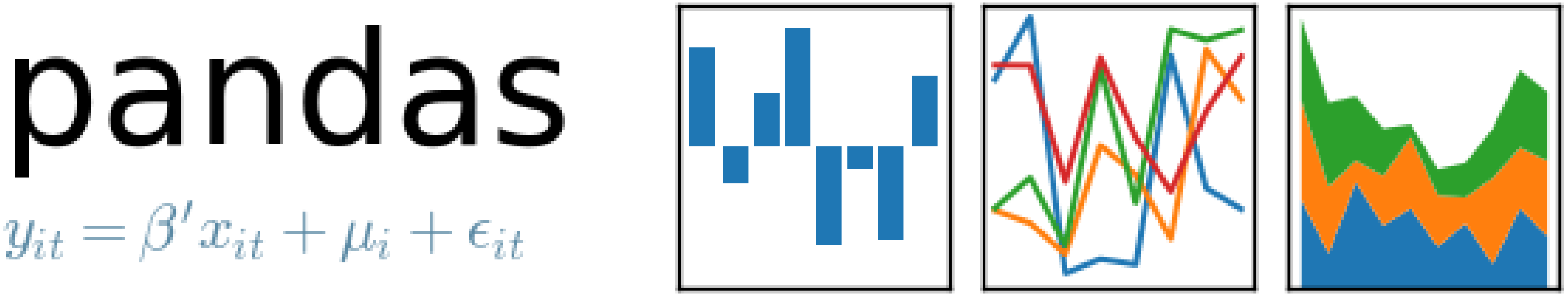
Data Science Level 1

-- Session 2--

Data manipulation with pandas

Hung Nguyen

TIC Data Team Lead



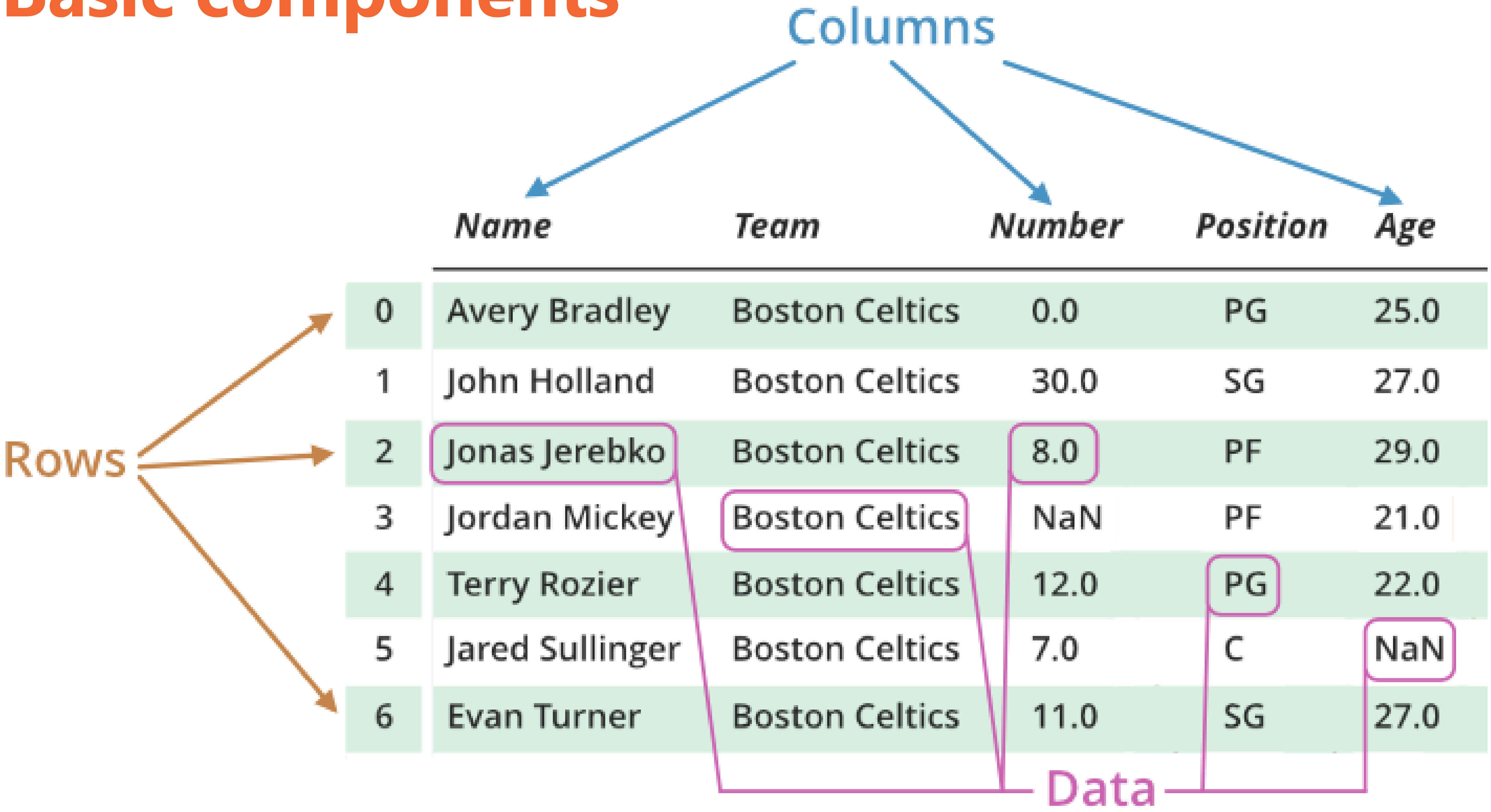
- high-performance
- easy-to-use
- data structures and analysis tools

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

- Easier to clean & wrangle Data.
- Features of Pandas make it a great choice for Data Science and Analysis.
- More useful with Matplotlib & Numpy

Basic components



	Name	*Team*	*Number*	*Position*	*Age*
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Create data frame

pd.DataFrame()

- List of rows
- List of columns
- Matrix
- …

pd.read_*()

- **csv** – comma/tab separated value
- xls – Excel file
- sql - query
- …

Reshape and select

Reshape

- **add**: pd.concat()/df.append()
- **delete**: df.drop()
- **join**: df.merge()

Select

- column:

  df['column name']/df.column_name

- row:

  df.loc()/df.iloc

  df.head()/df.tail()

- cell/range of cell
- filter:

  df[logical expression]

Summarize a data frame

Objective	Command
Get shape	df.shape()
Basic information	df.info()
Descriptive analysis	df.describe()/mean()/sum()/max()/min()…
Counting missing value	df/df[column] .isnull().sum()
Correlation matrix	df.corr()
Duplicate rows	df.duplicated()
Unique values	df[column].unique()/value_counts()

Basic transformations for preprocessing

Objective	Function
Sorting	df. sort_values()
Drop duplicates	df. drop_duplicates()

Dealing with missing value
- drop
- fill
	
	df.dropna()
	df.fillna()

Changing a column - map value to value - apply function - normalization - standardization	df[column] = df[column].map() df[column].apply()
Creating new column	df[new column] =
One-hot encoding	pd.get_dummies()

TMA Innovation Center

Data Science Level 1

-- Session 2--

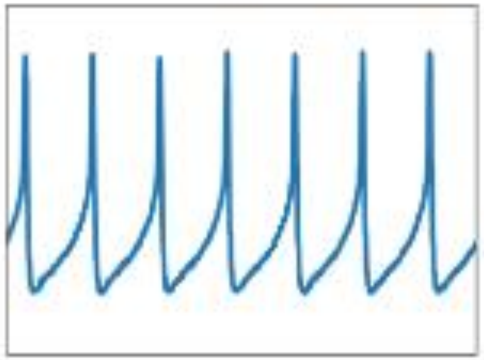
Data visualization with matplotlib

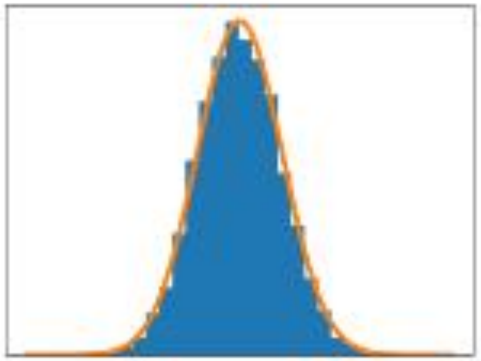
Hung Nguyen

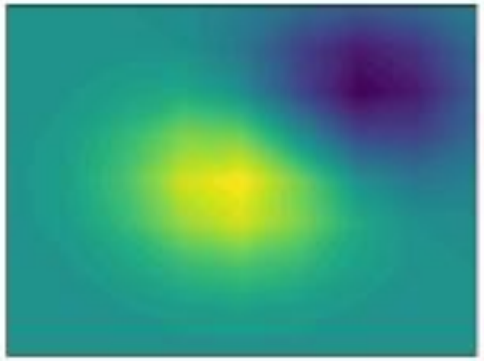
TIC Data Team Lead

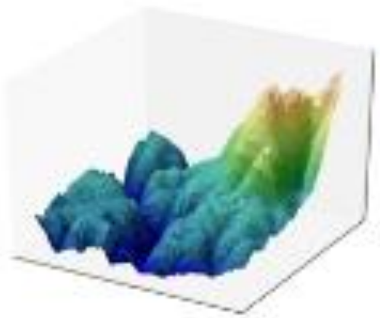
Data visualization

- Human brain processes graph faster than texts and tables
- Help to:
  - ✓ Convey concepts in a universal manner
  - ✓ Identify areas that need attention or improvement.
  - ✓ Clarify which factors influence customer behavior.
  - ✓ Help you understand which products to place where.
  - ✓ Predict sales volumes.



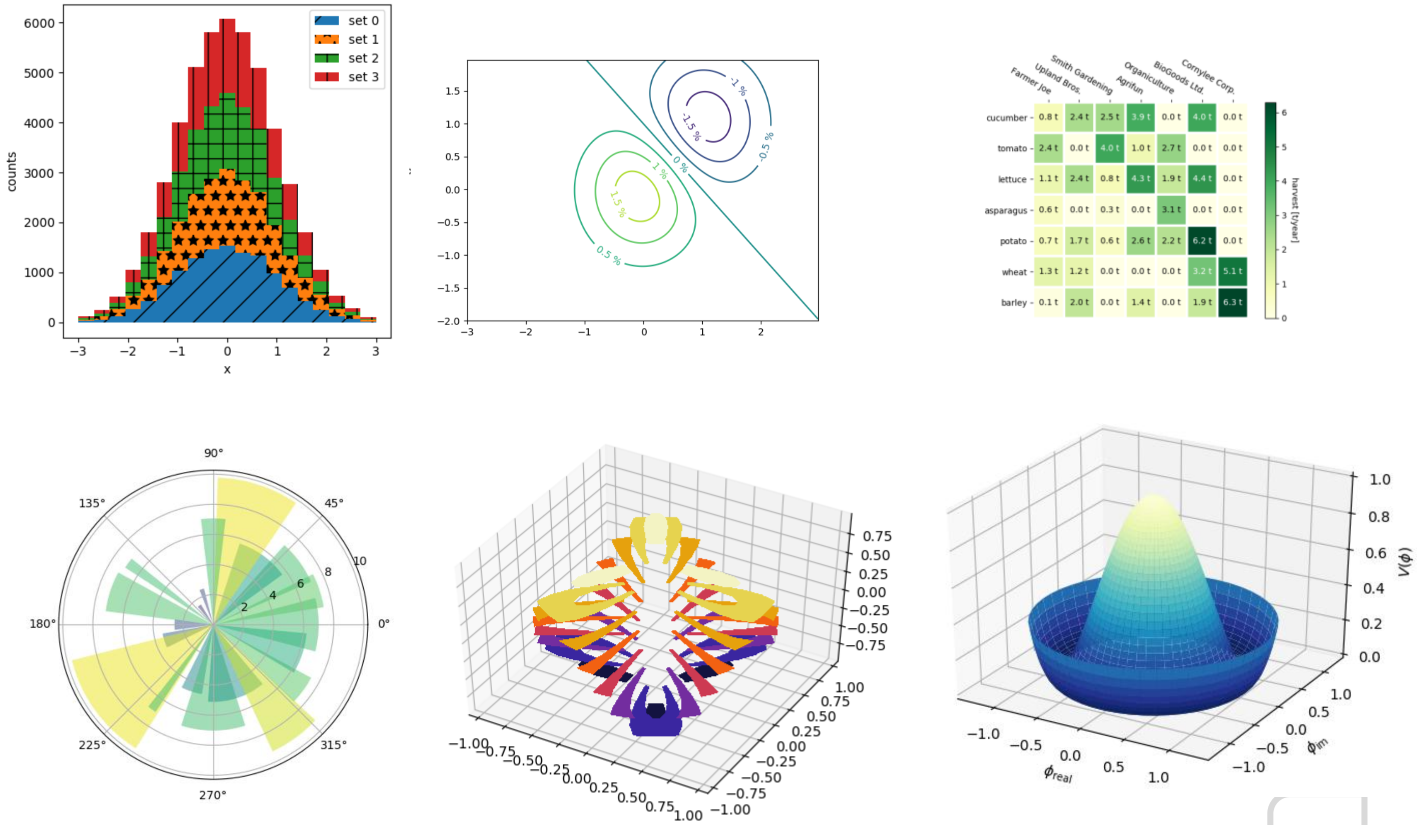






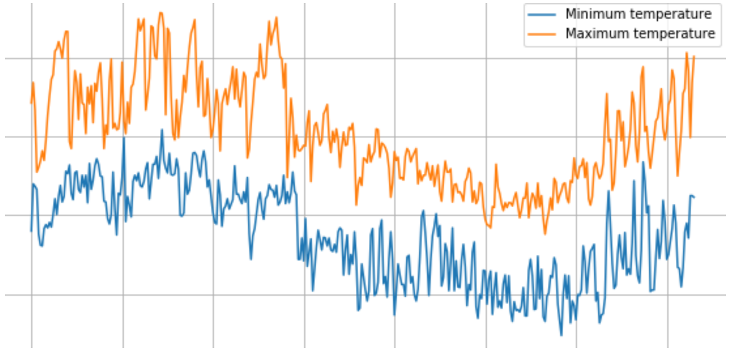
matplotlib

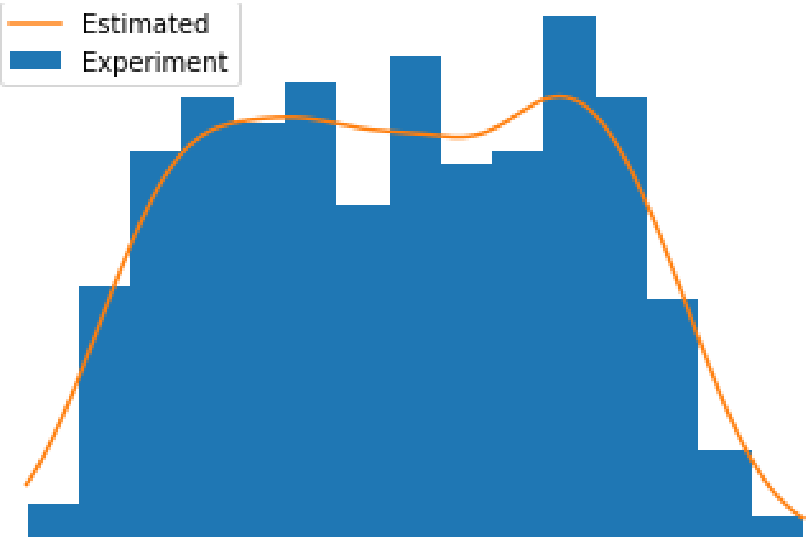
- ➢ used for plotting the beautiful and attractive Graphs
- ➢ Uses the numpy to handle large arrays of data sets
- ➢ Intergrates with pandas

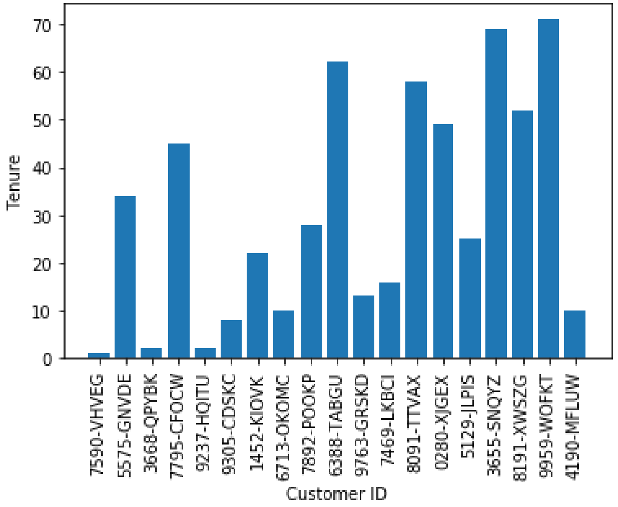


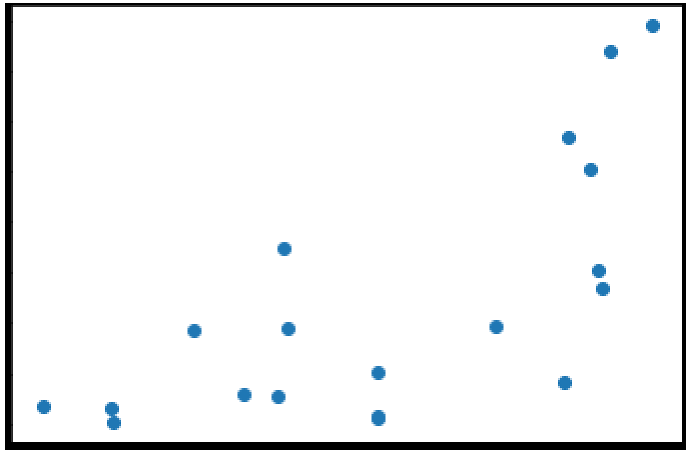
	Farmer Joe	Upland Bros.	Smith Gardening	Agrifun	Organiculture	BioGoods Ltd.	Cornylee Corp.
cucumber	0.8 t	2.4 t	2.5 t	3.9 t	0.0 t	4.0 t	0.0 t
tomato	2.4 t	0.0 t	4.0 t	1.0 t	2.7 t	0.0 t	0.0 t
lettuce	1.1 t	2.4 t	0.8 t	4.3 t	1.9 t	4.4 t	0.0 t
asparagus	0.6 t	0.0 t	0.3 t	0.0 t	3.1 t	0.0 t	0.0 t
potato	0.7 t	1.7 t	0.6 t	2.6 t	2.2 t	6.2 t	0.0 t
wheat	1.3 t	1.2 t	0.0 t	0.0 t	0.0 t	3.2 t	5.1 t
barley	0.1 t	2.0 t	0.0 t	1.4 t	0.0 t	1.9 t	6.3 t

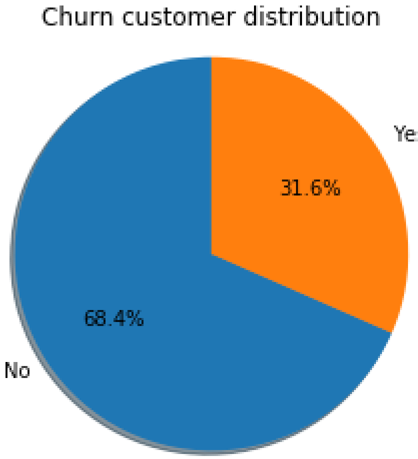
Basic graph types











THANK YOU!

Vietnam:	**84-2839-951-059**
North America:	**+1 844 224 4188**
Australia:	**+61 414 734 277**
Japan:	**+81 364 324 994**
Website:	**http://tma-innovation.center**
Email:	**innovation@tma.com.vn**