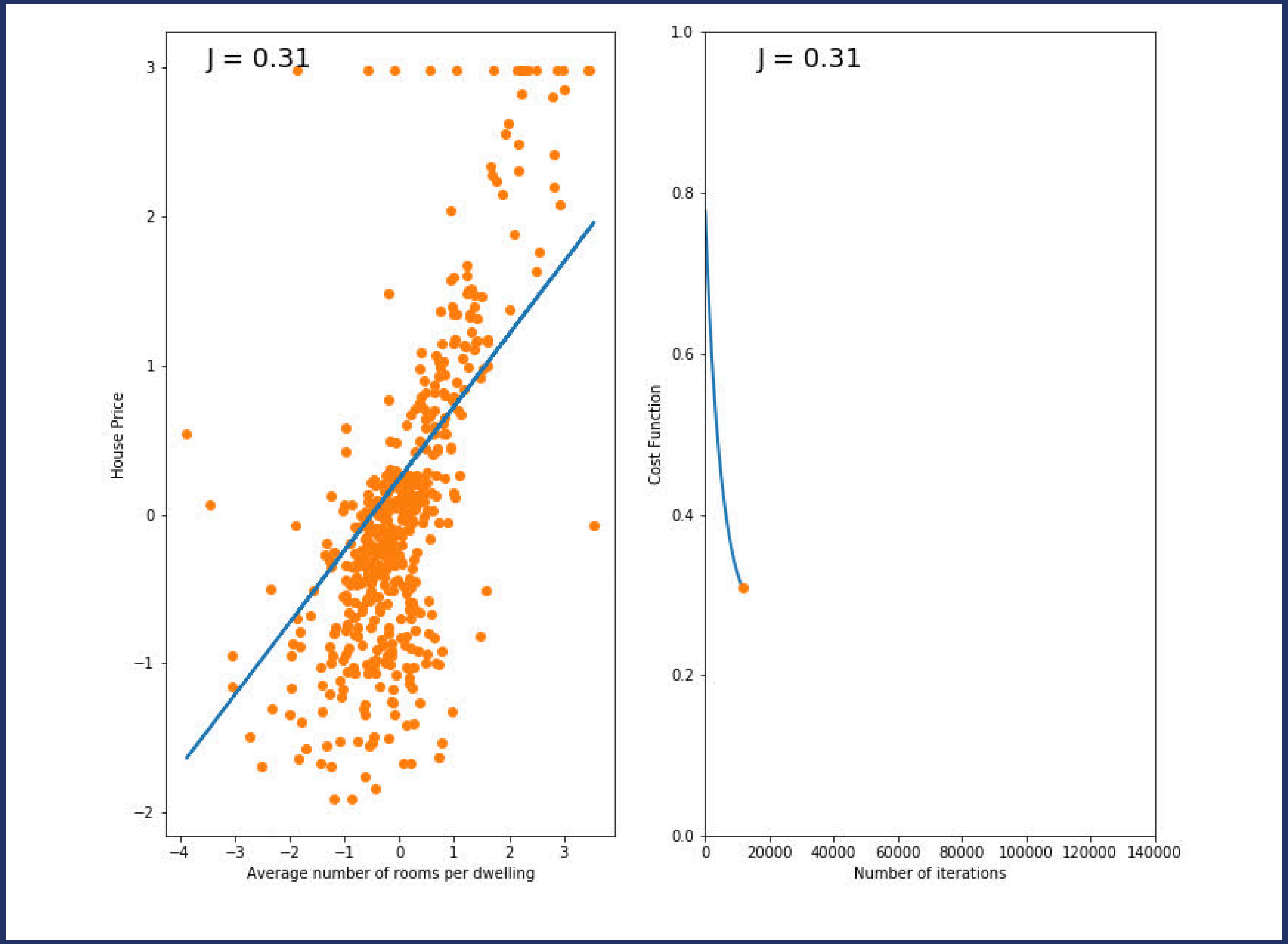
Statistics, Math Models and Algorithms

Fahad Faruqi, Ali Mohamed, Md Omit,
Evan Perez, Jennifer Saeteros, Tak Kit Yeung

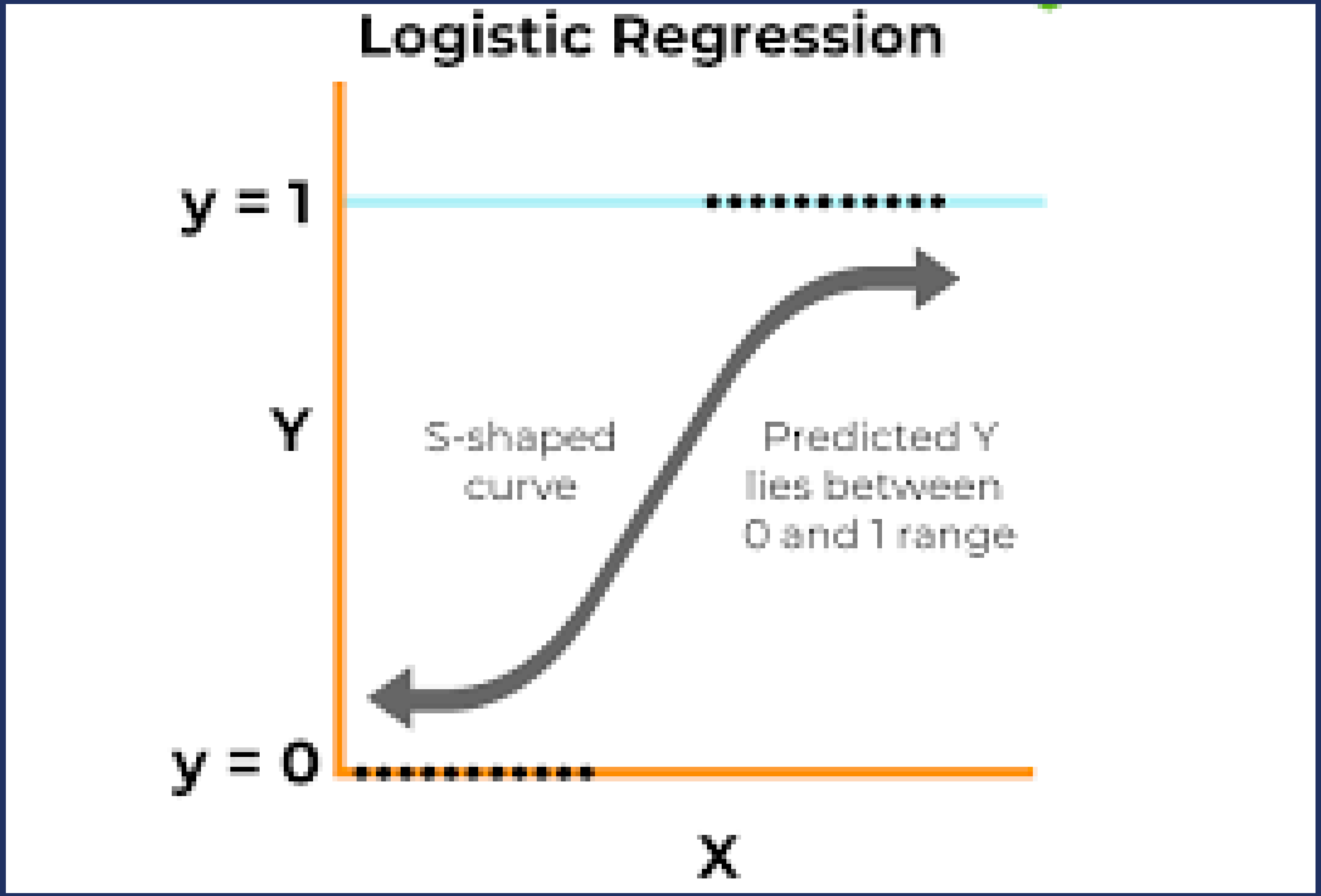
4219 Industries LLC

Linear Regression

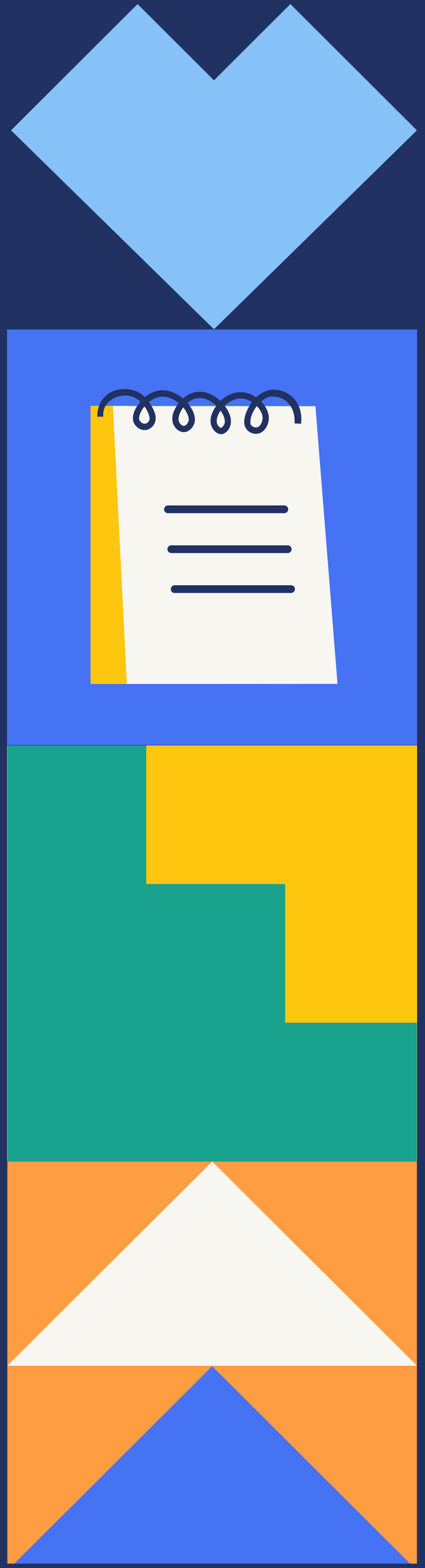


Finding a linear relationship between variables.

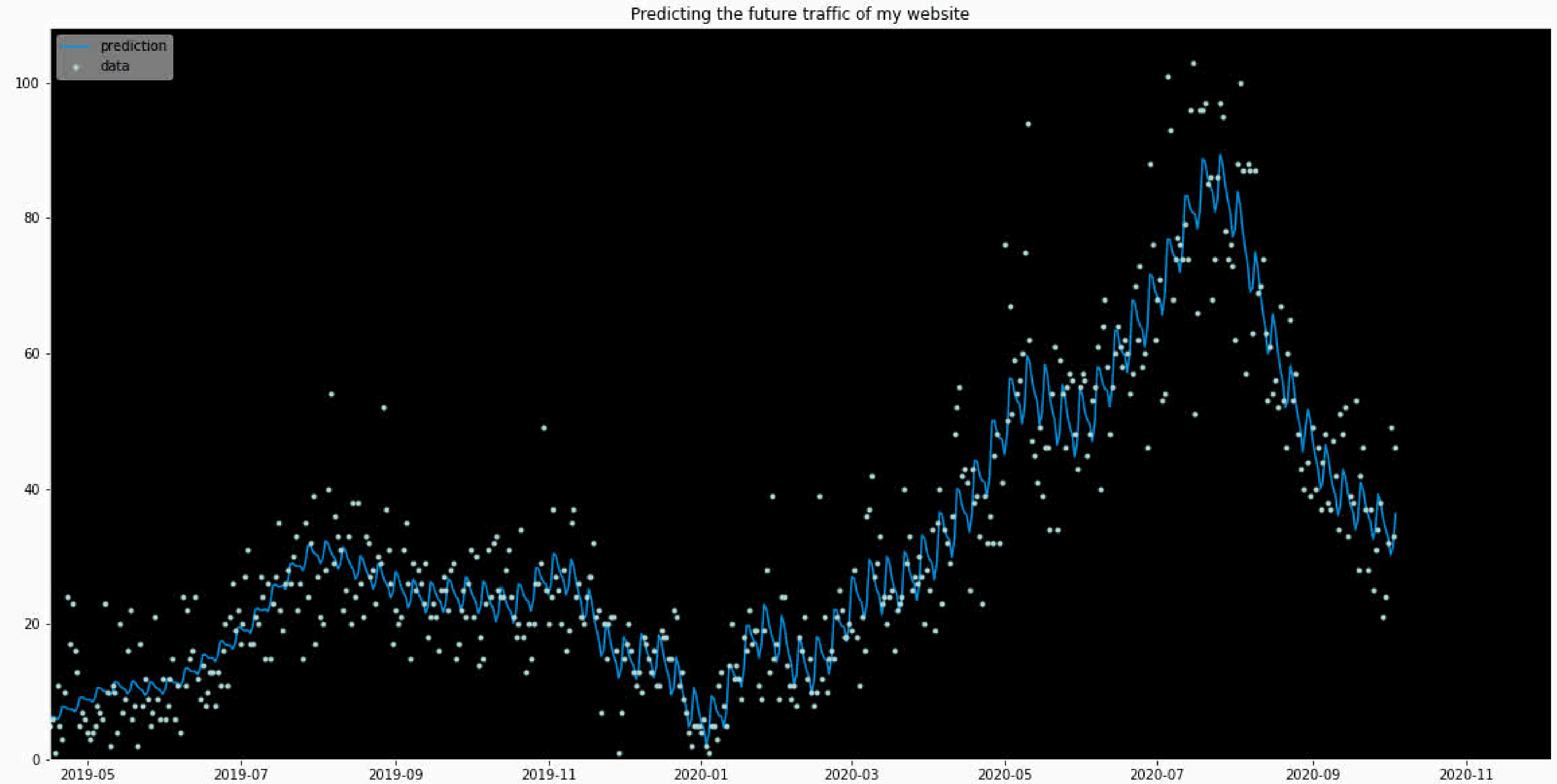
Logistic Regression



Determines probability that an event will occur.

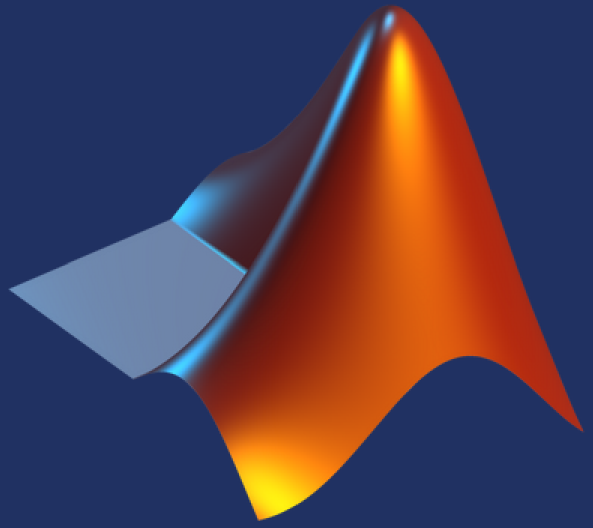


Time Series



Find patterns and forecast future values based on historical data.

Common Statistical Softwares



Matlab

PowerBI

JMP

Median

Set A: 8, 3, 4, 9, 6

3, 4, $\boxed{6}$, 8, 9

↓

Median

Set B: 11, 17, 3, 14, 19, 7

3, 7, $\boxed{11, 14}$, 17, 19

$$\text{Median} = \frac{11 + 14}{2} = \boxed{12.5}$$

Mean

Set A: 8, 3, 4, 9, 6

6

↓

Mean

Set B: 11, 17, 3, 14, 19, 7

$$\text{Mean} = \frac{11 + 17 + 3 + 14 + 19 + 7}{6} = 11.8$$

Common Statistical Metric

R^2

correlation between two datasets
high R^2== high correlation

RMSE

differences between true or predicted values

Recall, Precision, Accuracy

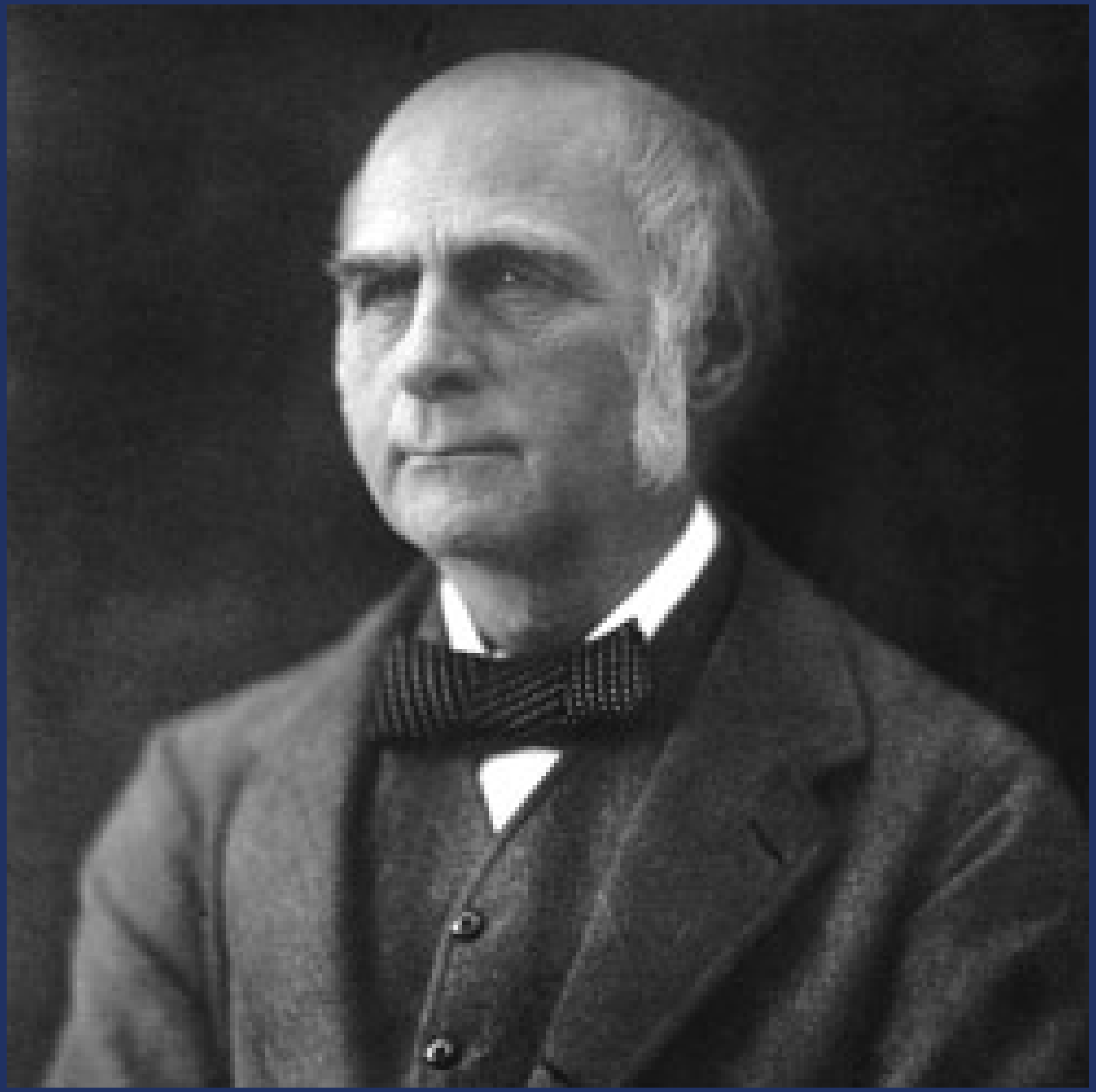
true positive rate, how closely the predictions group together, percentage correct prediction

History

Linear Regression Model

Introduced in 1894.

Derived by Sir Francis Galton after conceptualizing regression towards the mean.



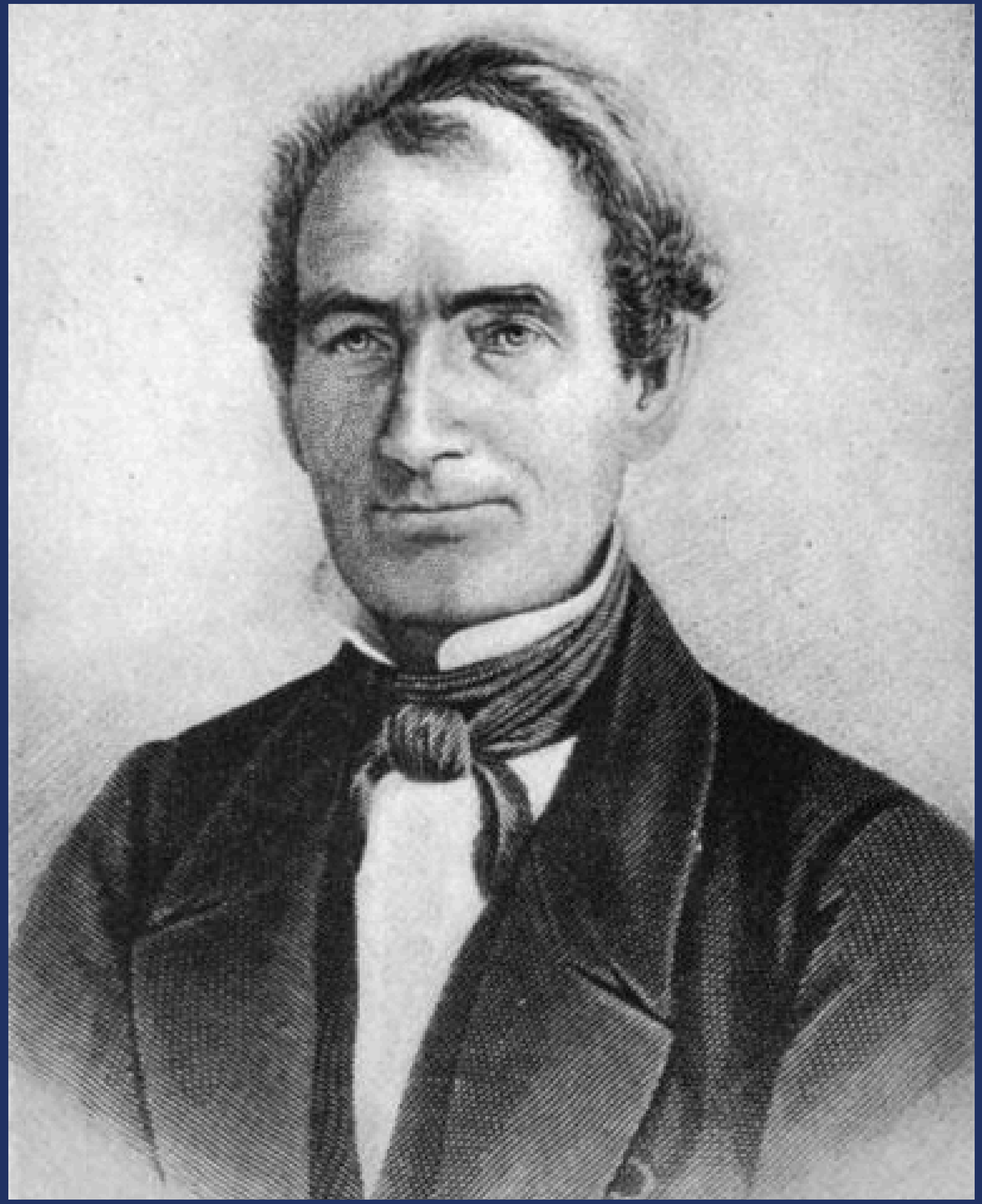
Used to determine relationships between variables and predict the value of one variable based on independent variables.

History

Logistic Regression Model

Found by Pierre-Francois Verhulst in 1838.

Rediscovered in 1920 as a model for population growth.



Predicts the probability of an event occurring, classifies data into different categories.

History

Time series Model

Data dates as far back as 800 BC China.

Analysis was created by Udny Yule in 1927.

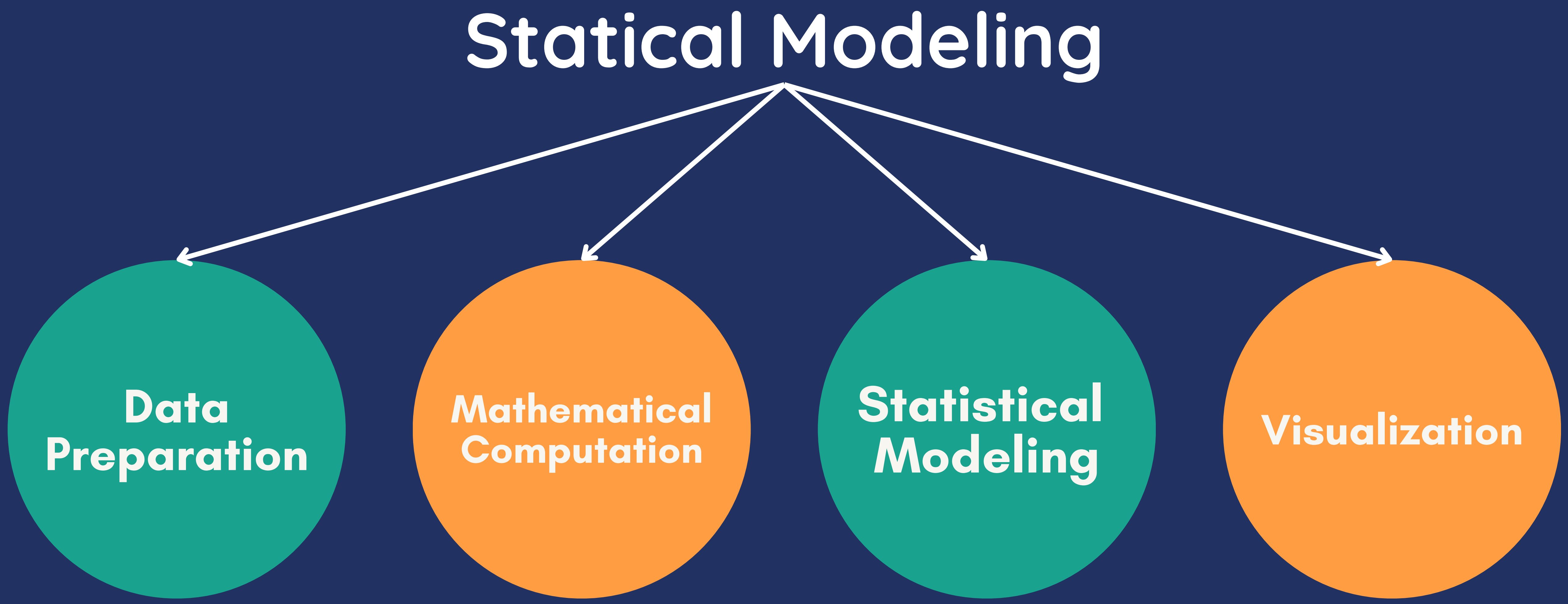
Understand past performance and predict future outcomes in a relevant and actionable way.

Usecases

Popularity and Community Support

Time Series	Logistic Regression	Linear Regression
Forecasting Capability	Binary Classification	Statistical Foundation
Widely Applicable	Widely Applicable	Widely Applicable
Temporal Dependency		Simplicity

Popularity and Community Support



Statical Modeling

Data Preparation

Mathematical Computation

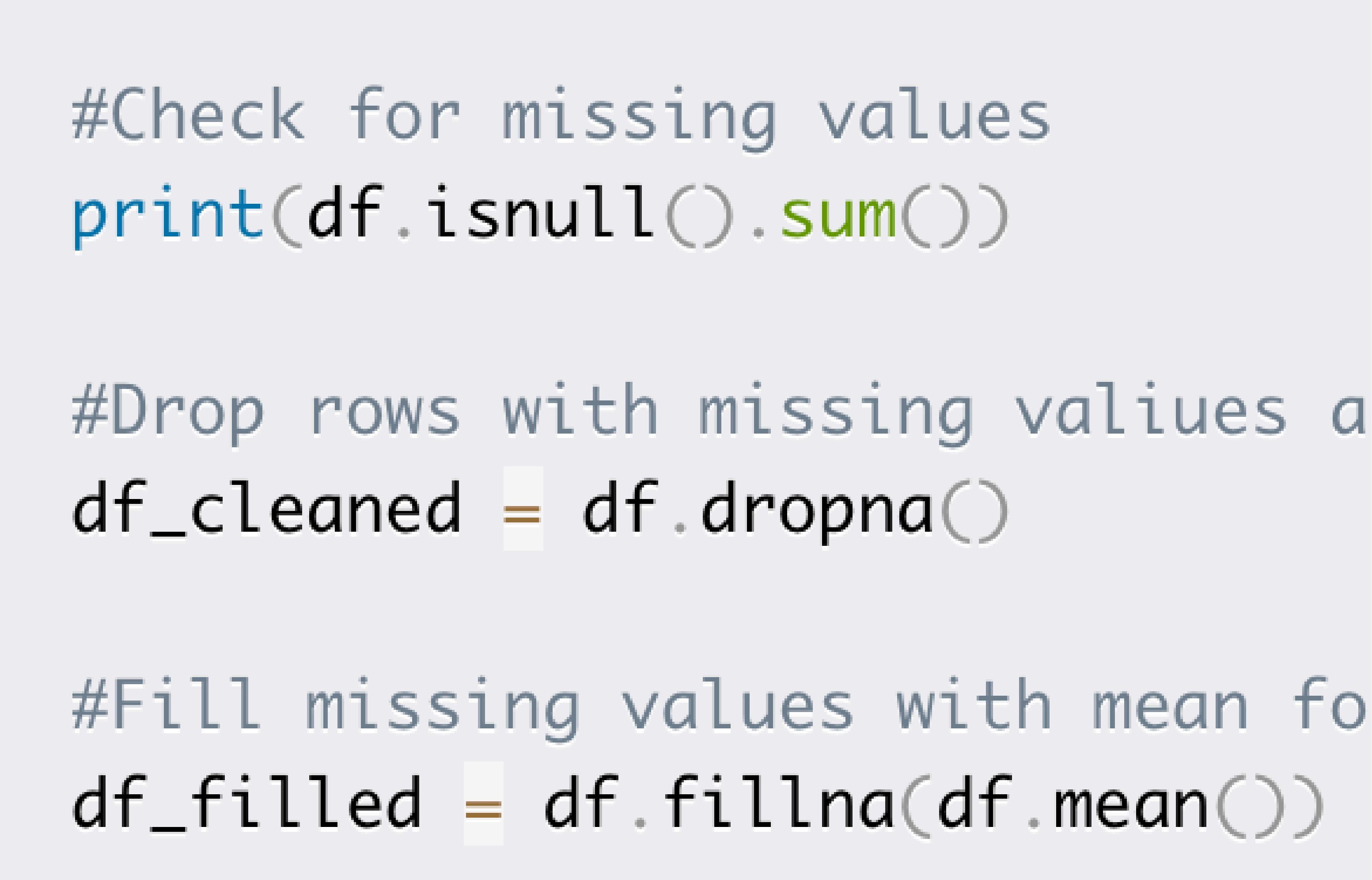
Statistical Modeling

Visualization

Preprocessing

Data Cleaning and Prepartion

Pandas

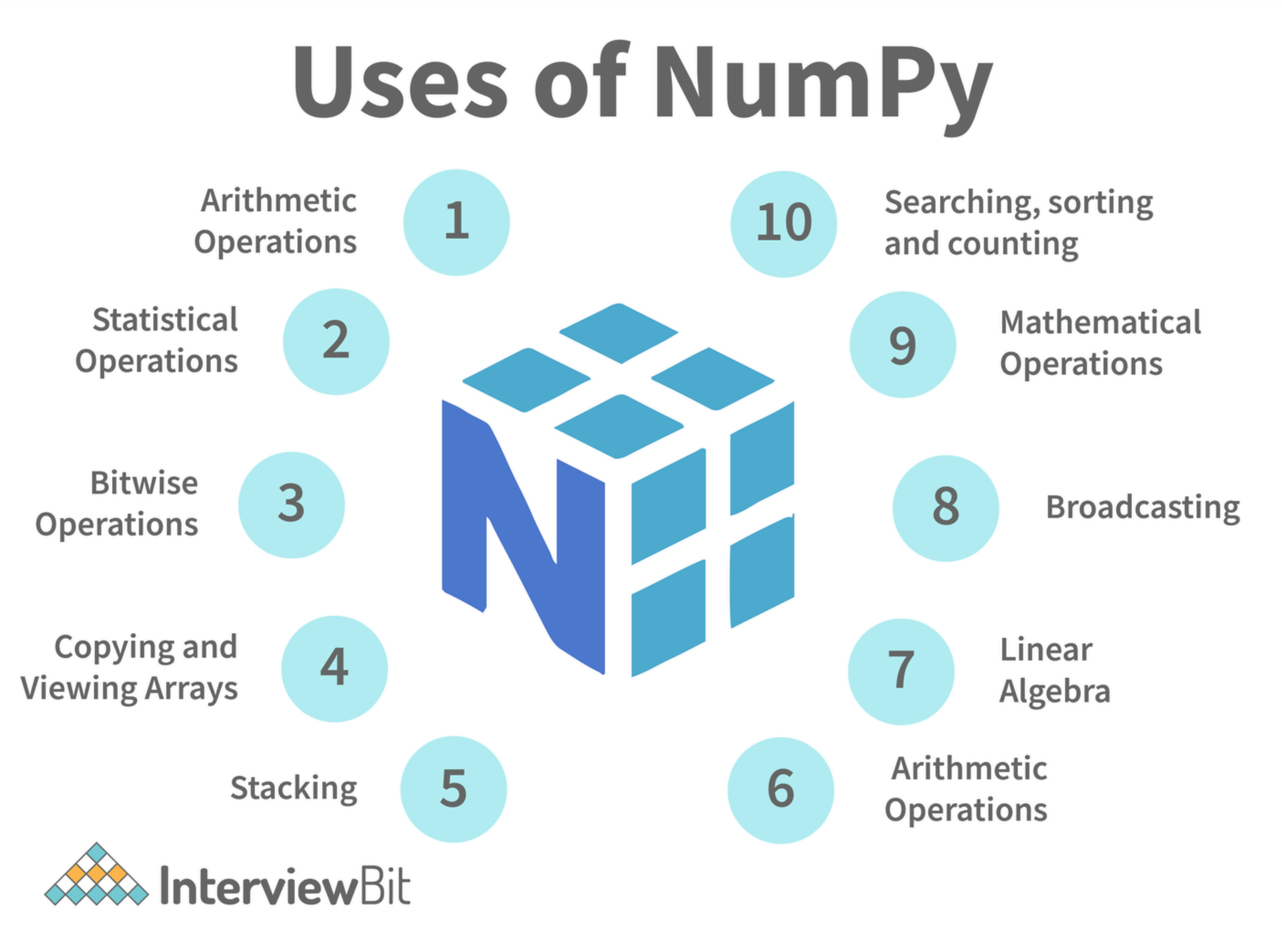


```
#Check for missing values
print(df.isnull().sum())

#Drop rows with missing valiues a
df_cleaned = df.dropna()

#Fill missing values with mean fo
df_filled = df.fillna(df.mean())
```

Doing the math

Statical Computation

NumPy
SciPy



Uses of NumPy

1. Arithmetic Operations
2. Statistical Operations
3. Bitwise Operations
4. Copying and Viewing Arrays
5. Stacking
6. Arithmetic Operations
7. Linear Algebra
8. Broadcasting
9. Mathematical Operations
10. Searching, sorting and counting

InterviewBit

Statistical Modeling

statsmodels

scikit learn

machine learning in Python

R

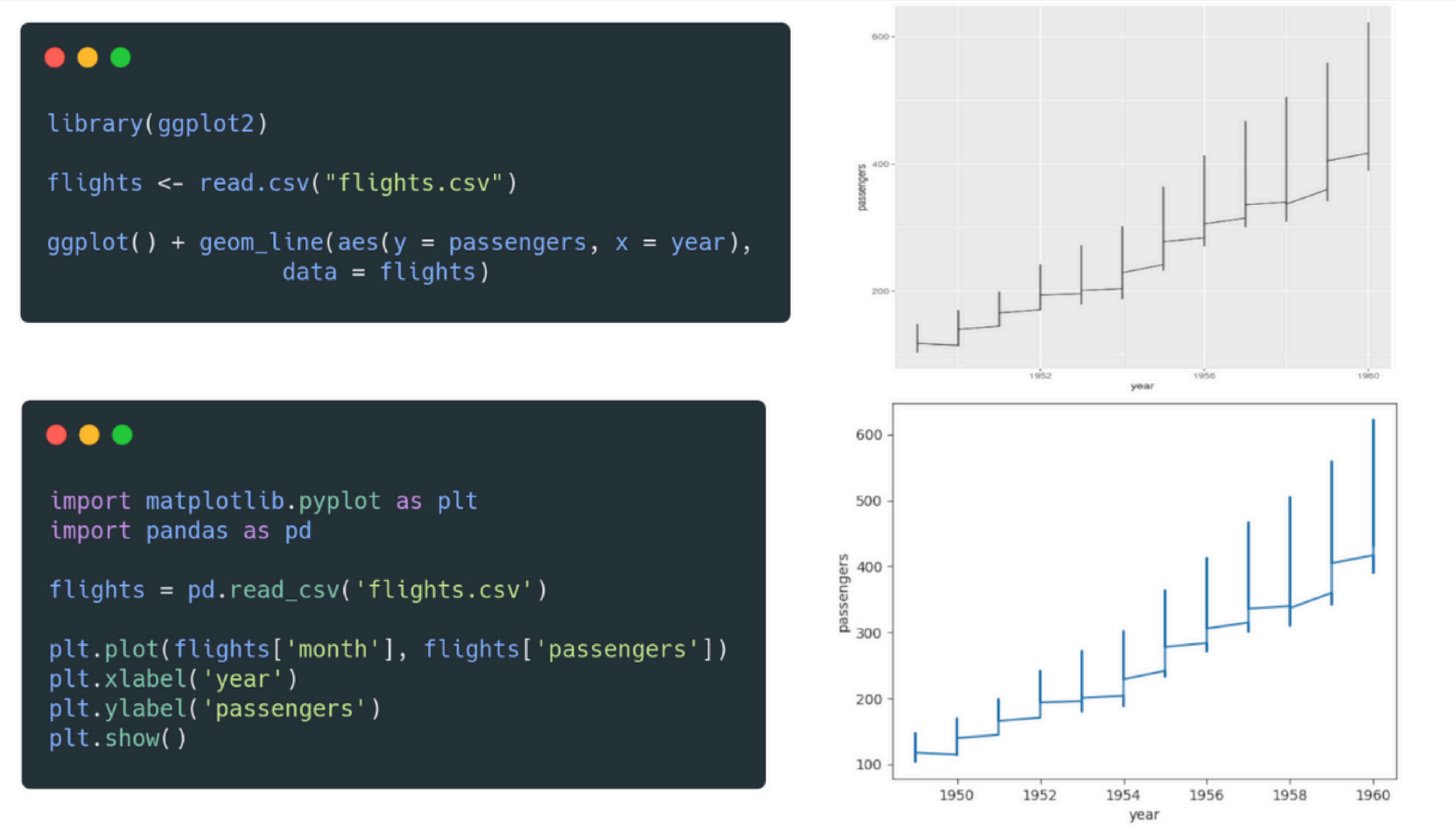
Statsmodels

Scikit-learn

R

Matplotlib
Seaborn
Plotly
ggplot2

Visualization



```r
library(ggplot2)

flights <- read.csv("flights.csv")

ggplot() + geom_line(aes(y = passengers, x = year),
                     data = flights)
```

```python
import matplotlib.pyplot as plt
import pandas as pd

flights = pd.read_csv('flights.csv')

plt.plot(flights['month'], flights['passengers'])
plt.xlabel('year')
plt.ylabel('passengers')
plt.show()
```

Tabular Comparison

	Predictive Analysis	Probability Analysis	Requires lots of data
Linear Regression	✓		✓
Logistic Regression		✓	✓
Time Series	✓	✓	✓

Linear Regression

Pros

best model for linear relationships
fast & simple
low training times

Cons

succeptible to overfitting
cannot determine complex relationships
needs high sample size

Logistic Regression

Pros

can extend to multiple classes (labels)

fast & simple

good accuracy for linear datasets

Cons

assumes linear boundaries

easily outperformed in determining complex relationships

assumes few outliers

Time Series

Pros

identifies historical trends
identifies outliers in data
shows effect over time

Cons

needs comprehensive data
data must have linear relationship
requires human interpretation

Industry and Academic Relevance

Linear Regression Model

Academic

Biological, behavioral, environmental, social sciences

Industry

Business, insurance

Use Case

Continuous variable based on another variable

Industry and Academic Relevance

Logistic Regression Model

Academic

Machine learning

Industry

Marketing, healthcare

Use Case

Categorical dependent variable

Industry and Academic Relevance

Time Series Model

Academic

Economics,
education

Industry

Finance,
retail

Use Case

Data that are
constantly
fluctuating
over time

Recommended Analysis Technique:

Time Series

- **Forecasting**
    - predict future values
- **Classification**
    - identify and assign categories
- **Descriptive**
    - find trends, patterns, cycles
- **Curve fitting**
    - study relationship between variables
- Obtaining lots of data can be easier due to periodic data collection

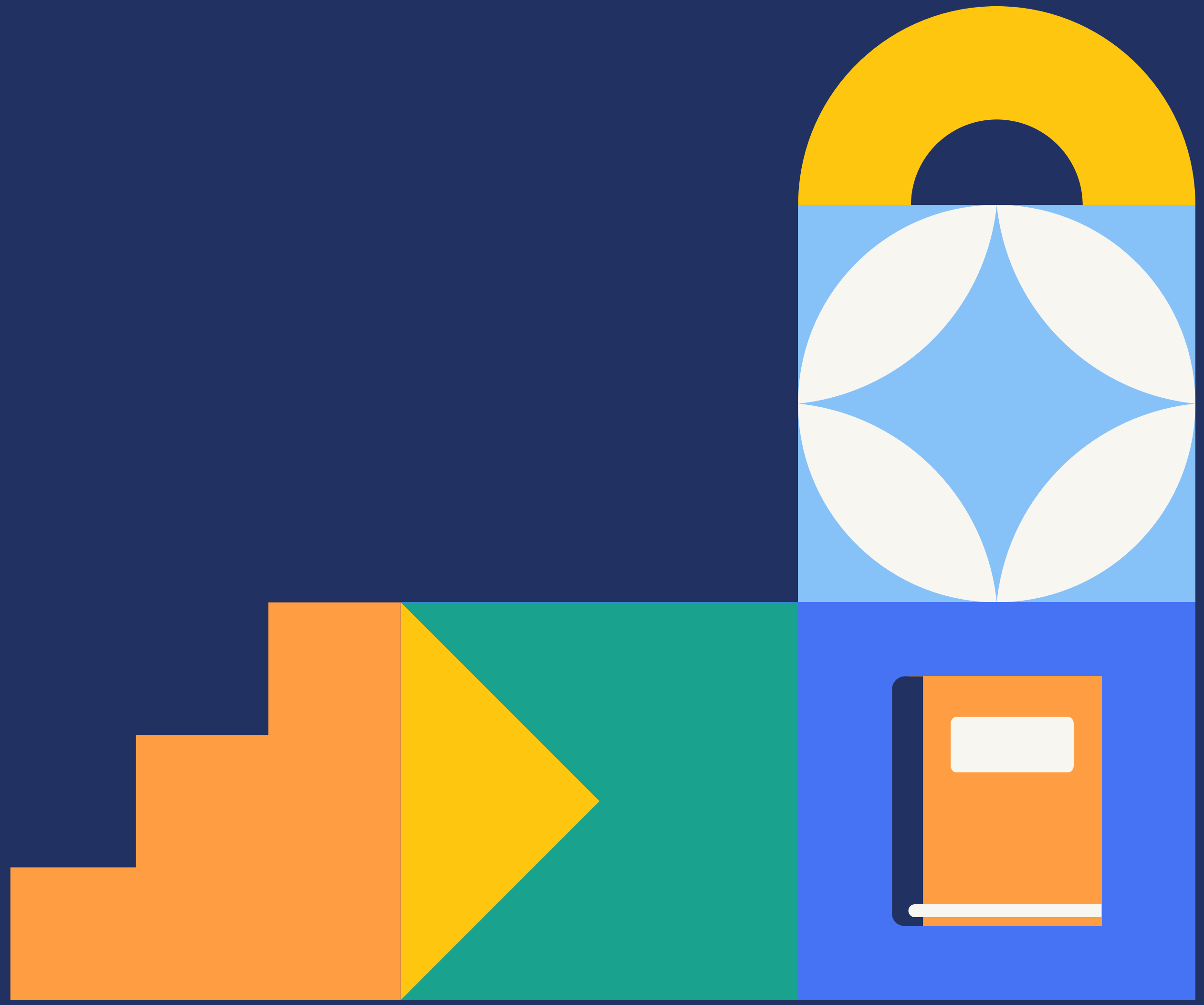
Conclusion

- Different problems require different techniques, especially if the problem is fine-grained
- Statistical analysis methods should be chosen based on:
    - **Data availability**
    - **Data quality**
    - **Focus of the problem at hand**

Thank You For Listening!

References

- Mathspace. "Statistics."
- Institute of Mathematical Statistics (IMS)
- R Consortium:ocuses on supporting the R programming language and community, but it’s also highly relevant for statisticians who use R for statistical models
- Cross Validated (StackExchange): A popular Q&A platform for statisticians, data scientists, and researchers



Q&A