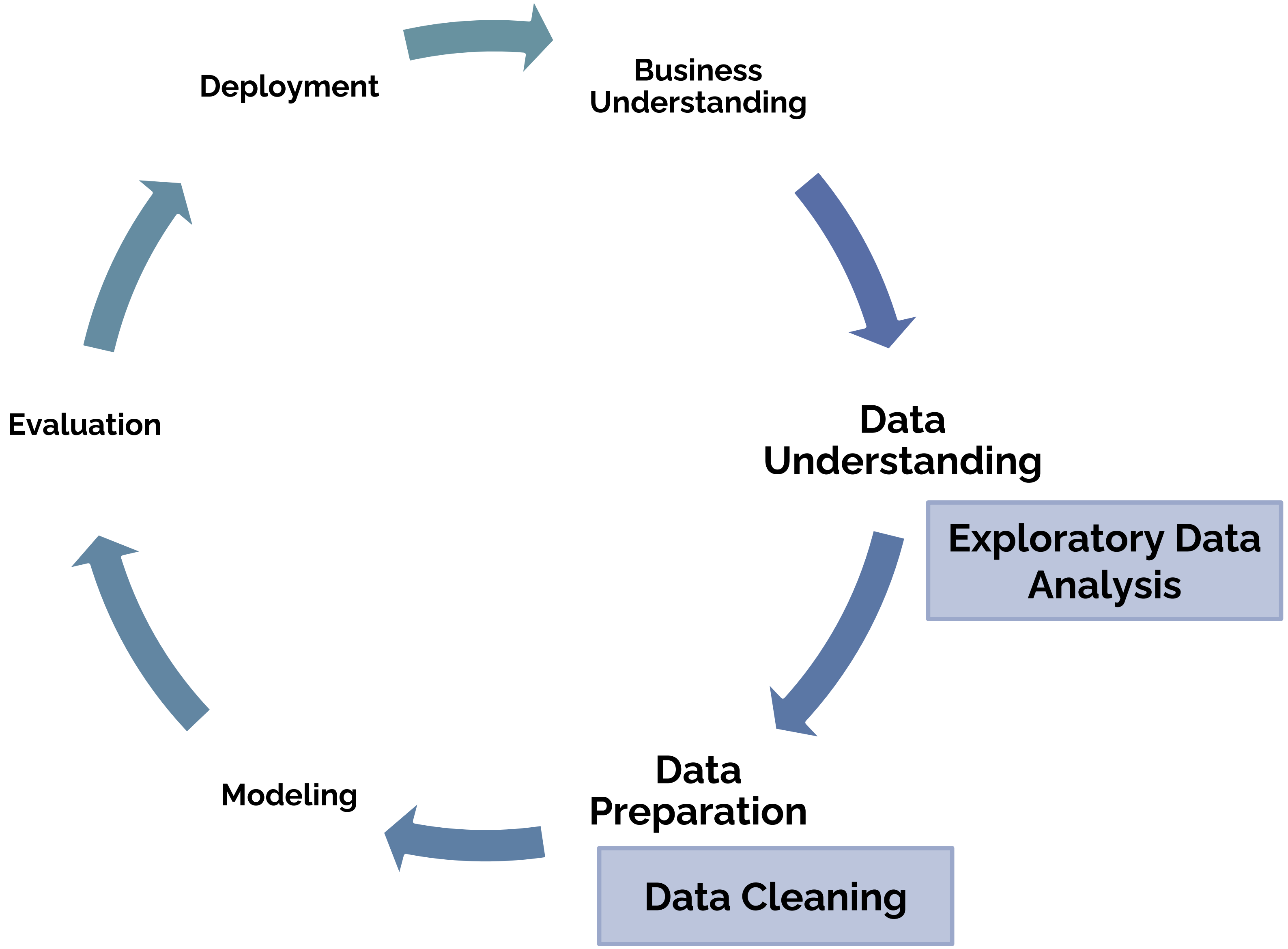
Welcome…

Feature Engineering

CS 797Q
Fall 2024

09/30/2024

REVIEW



Deployment

Business Understanding

Data Understanding

Exploratory Data Analysis

Data Preparation

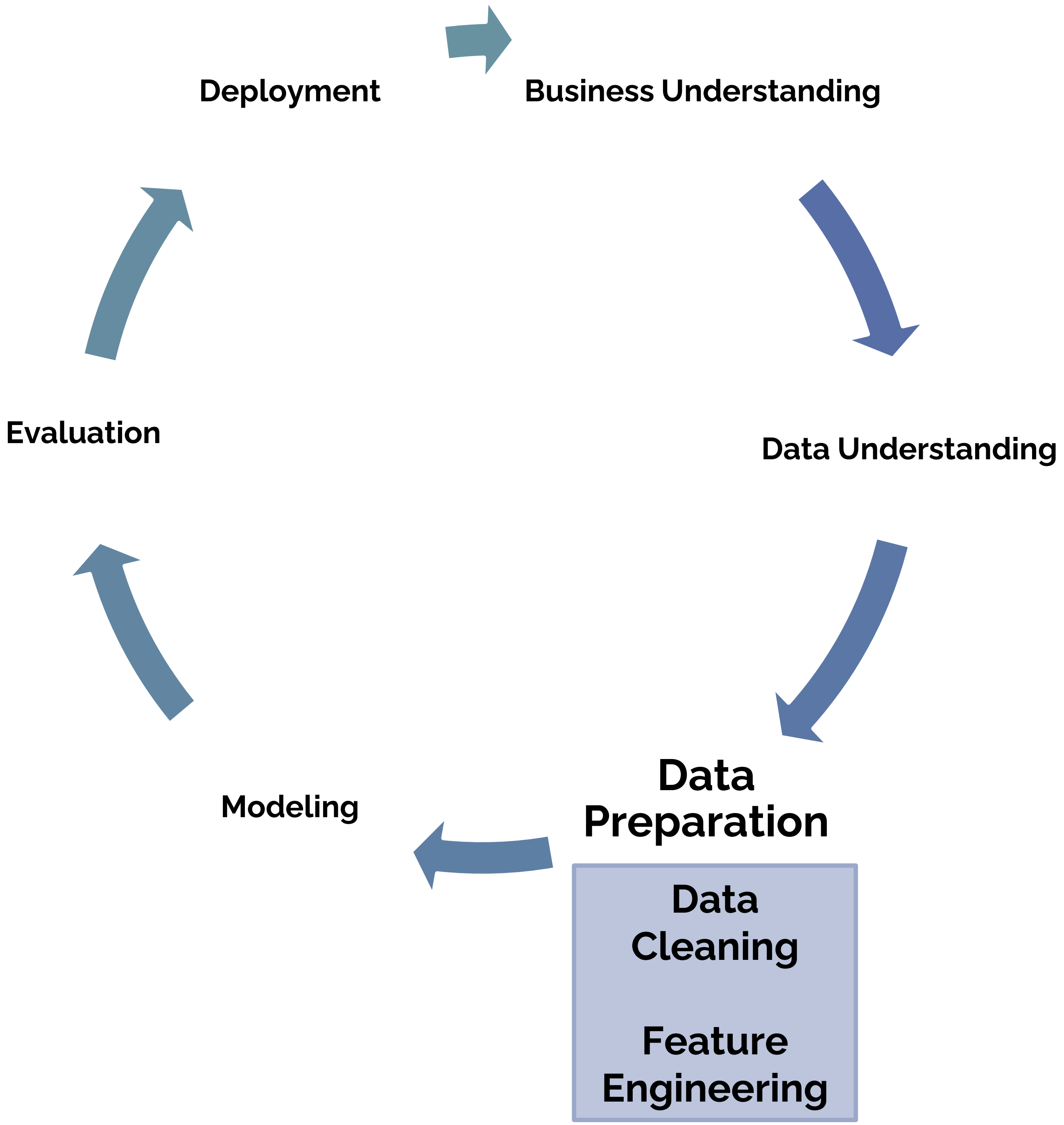
Data Cleaning

Modeling

Evaluation

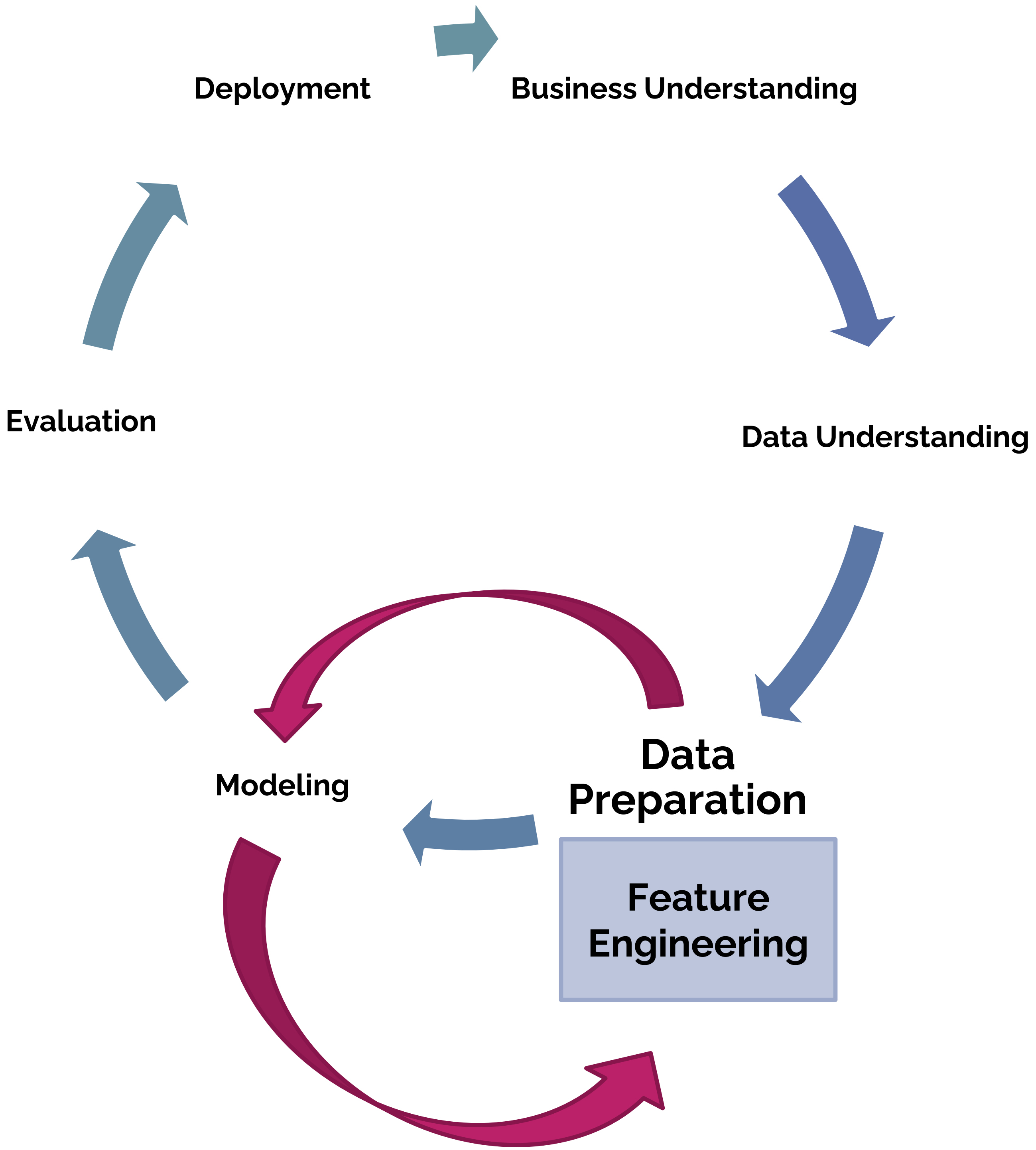
BEST PRACTICES

- Goals of Feature Engineering
  - Create, select, manipulate, and transform features for machine learning models
- Purpose of Feature Engineering
  - Meet algorithm requirements
  - Improve the performance of machine learning models
  - Better interpretability of relationships



FEATURE ENGINEERING

- Feature Engineering can be iterative with modeling
- Create and test new features to improve model performance

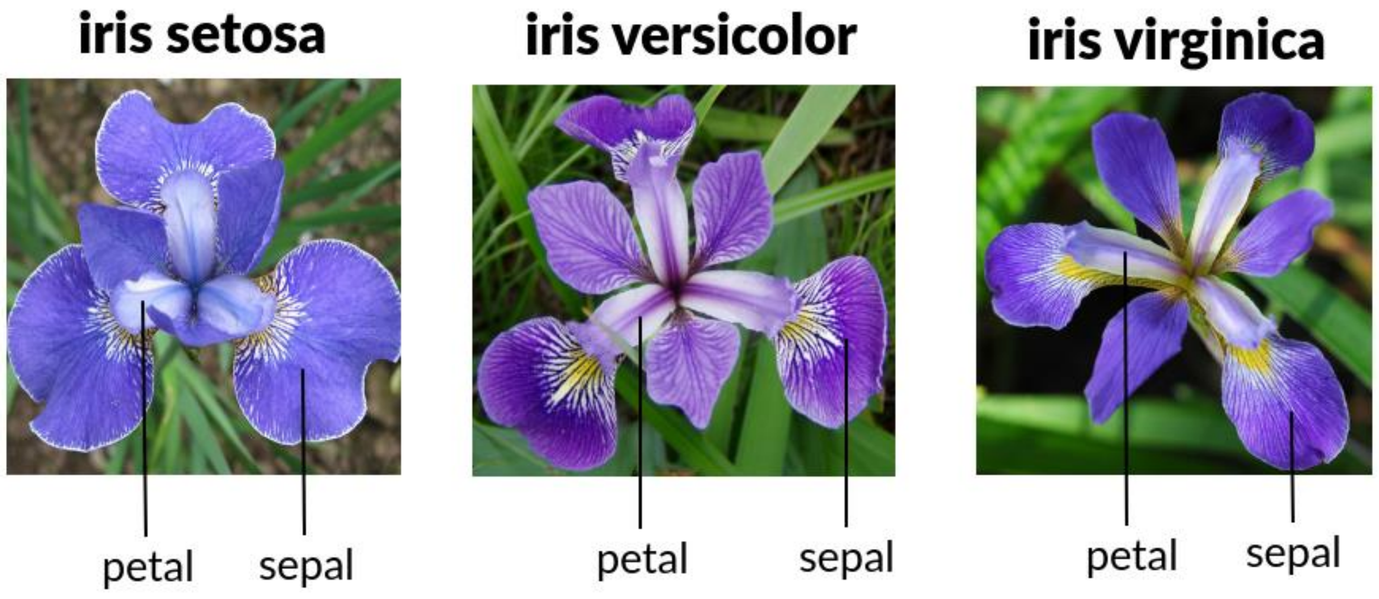


FEATURE ENGINEERING

Object	**Price Per Pound**	**Calories Per Pound**	**Is fruit**	**Is vegetable**	**Calories Per Dollar**
Broccoli	2.78	154	0	1	55.4
Banana	1.58	404	1	0	255.7
Mango	1.82	271	1	0	148.9
Cabbage	0.78	118	0	1	151.3

FEATURE ENGINEERING

Species	Sepal Length	Sepal Width	Petal Length	Petal Width	Setosa	Virginica	Versicolor	Sepal Size	Petal Size
Iris Setosa	5.1	3.5	1.4	0.2	1	0	0	17.85	0.28
Iris Virginica	6.3	3.3	6.0	2.5	0	1	0	20.79	15
Iris Versicolor	7.0	3.2	4.7	1.4	0	0	1	22.4	6.58

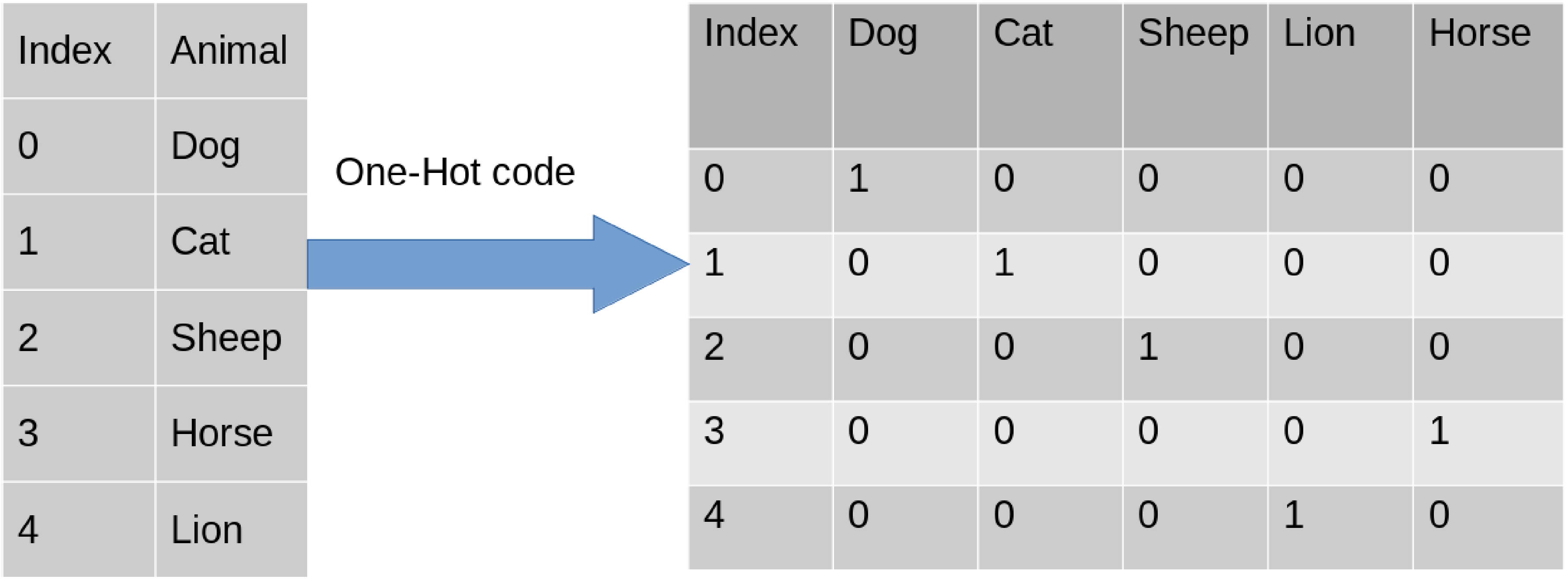


FEATURE ENGINEERING TECHNIQUES

- Encoding
- Binning
- Grouping
- Feature Splitting
- Extracting Date

ENCODING

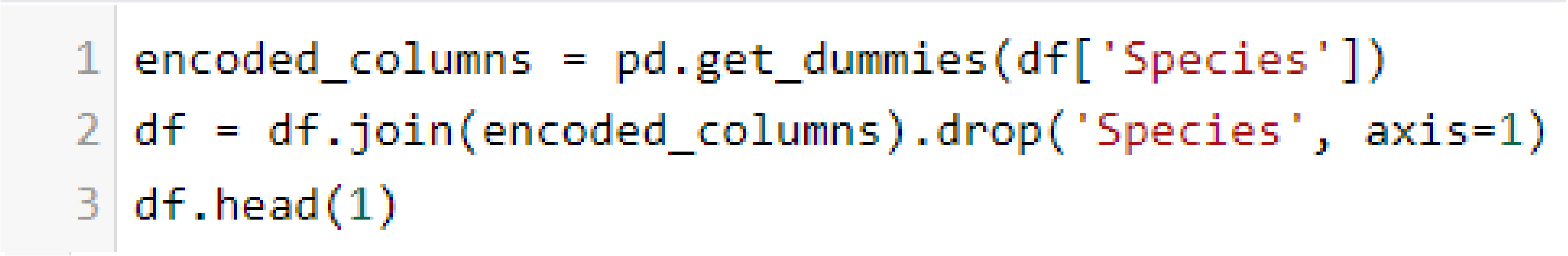
- Used on
  - Categorical features
- Purpose
  - Converting categorical features to numeric features
- Used frequently



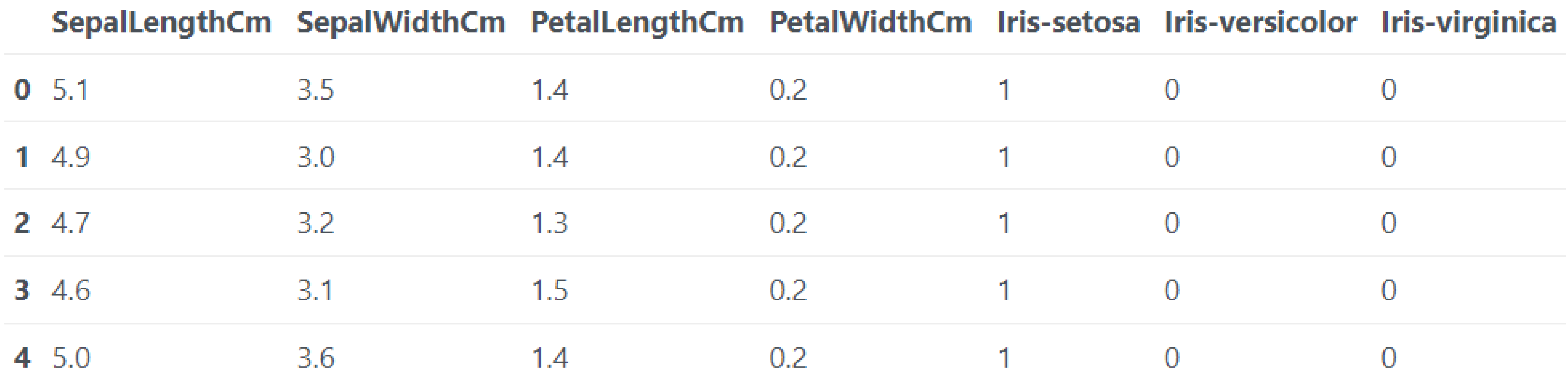
Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0



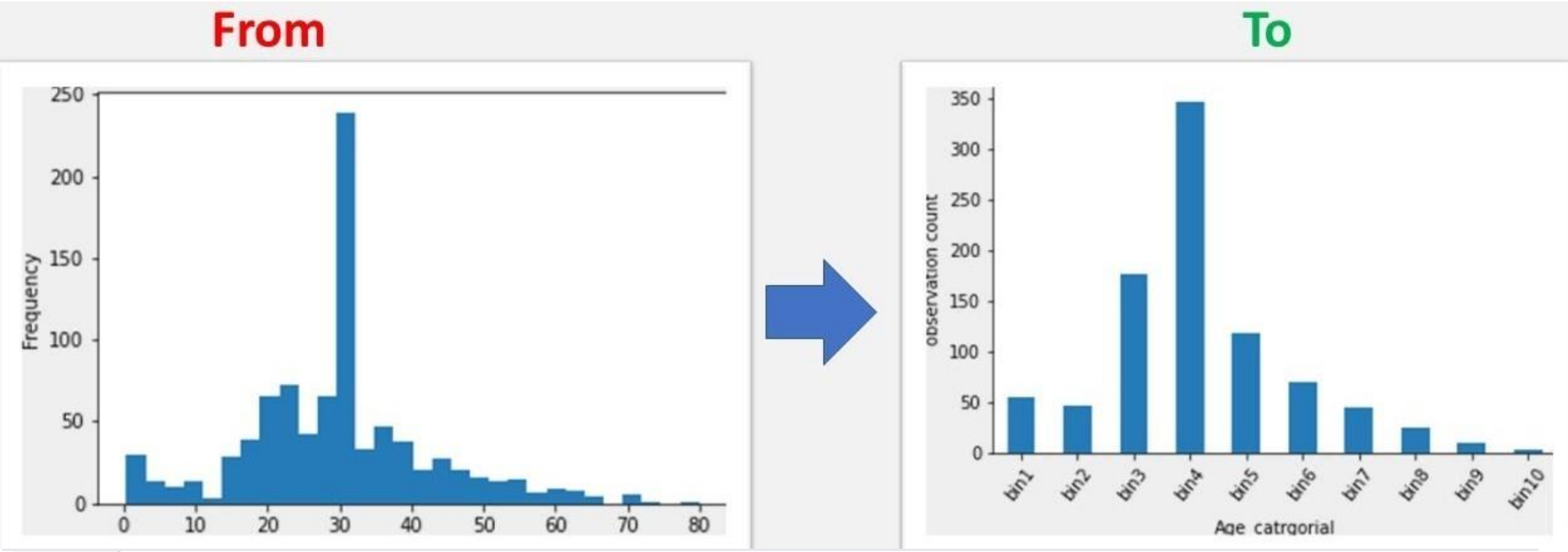
```
encoded_columns = pd.get_dummies(df['Species'])
df = df.join(encoded_columns).drop('Species', axis=1)
df.head(1)
```

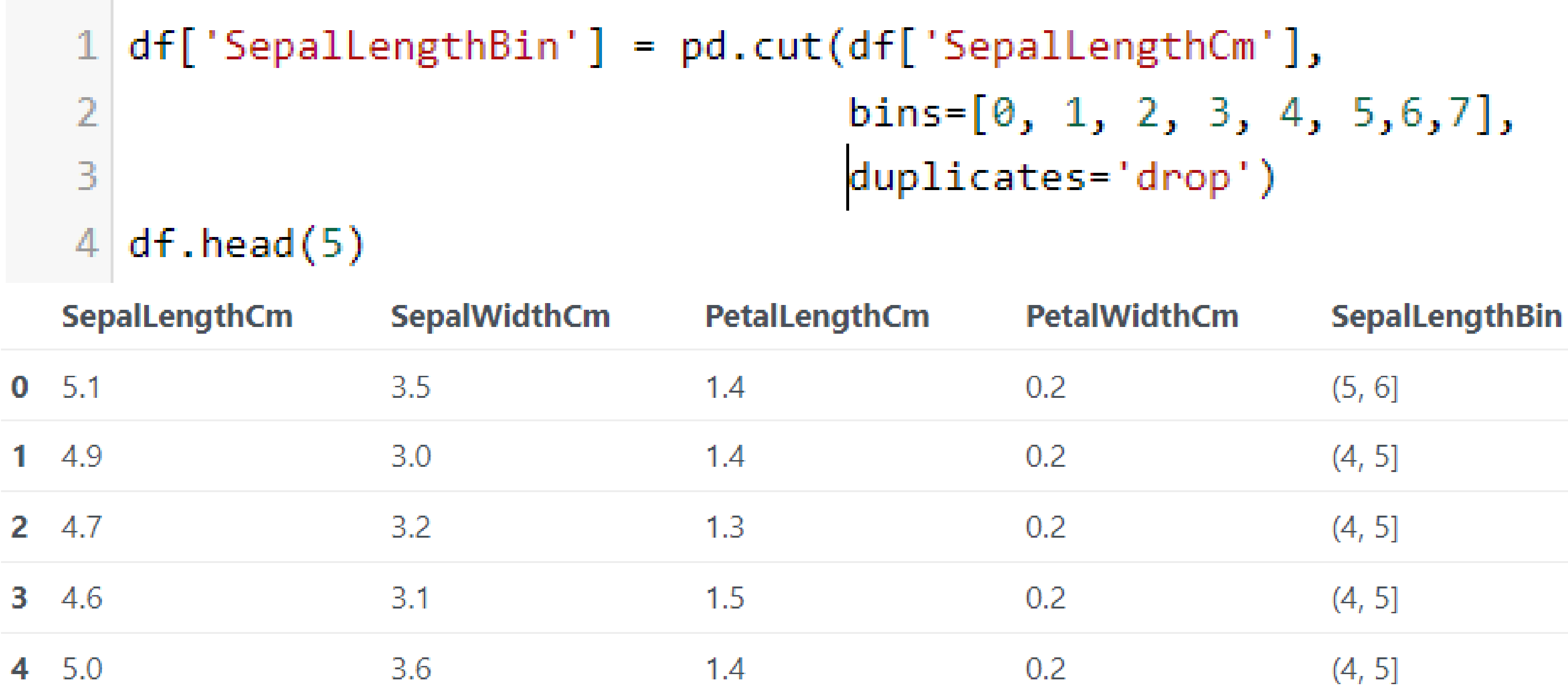


	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Iris-setosa	Iris-versicolor	Iris-virginica
0	5.1	3.5	1.4	0.2	1	0	0
1	4.9	3.0	1.4	0.2	1	0	0
2	4.7	3.2	1.3	0.2	1	0	0
3	4.6	3.1	1.5	0.2	1	0	0
4	5.0	3.6	1.4	0.2	1	0	0

BINNING

- Used on
  - Categorical and numeric data
- Purpose
  - Create a more robust model and prevent overfitting
- Binning to fewer categories causes data loss
- More appropriate for categorical feature labels that occur infrequently



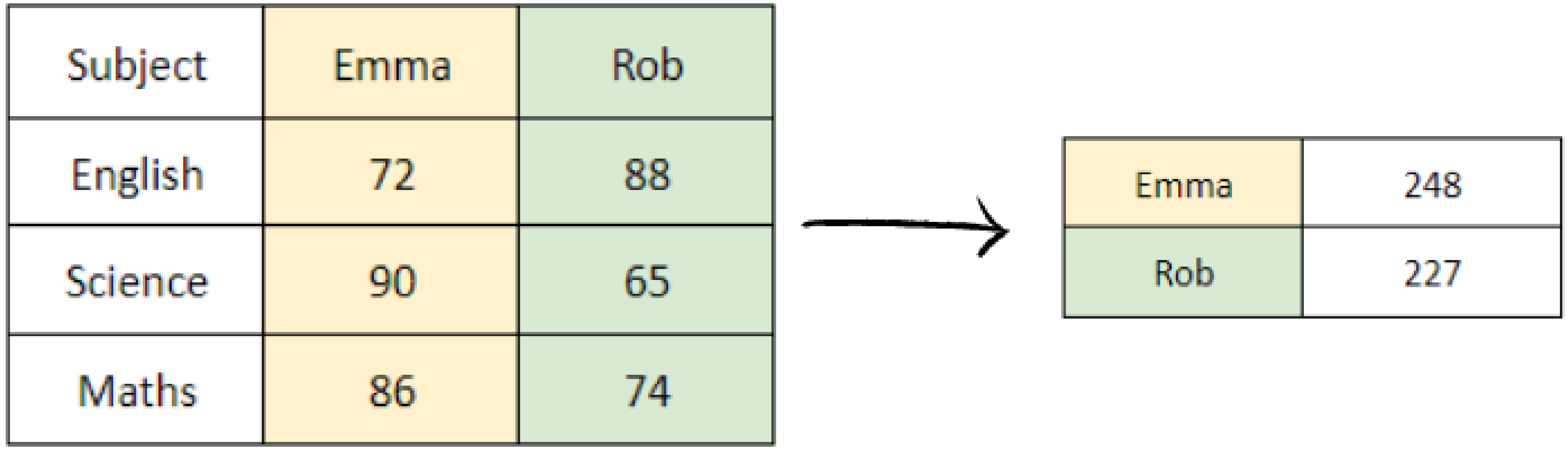


```
df['SepalLengthBin'] = pd.cut(df['SepalLengthCm'],
                              bins=[0, 1, 2, 3, 4, 5,6,7],
                              duplicates='drop')
df.head(5)
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	SepalLengthBin
0	5.1	3.5	1.4	0.2	(5, 6]
1	4.9	3.0	1.4	0.2	(4, 5]
2	4.7	3.2	1.3	0.2	(4, 5]
3	4.6	3.1	1.5	0.2	(4, 5]
4	5.0	3.6	1.4	0.2	(4, 5]

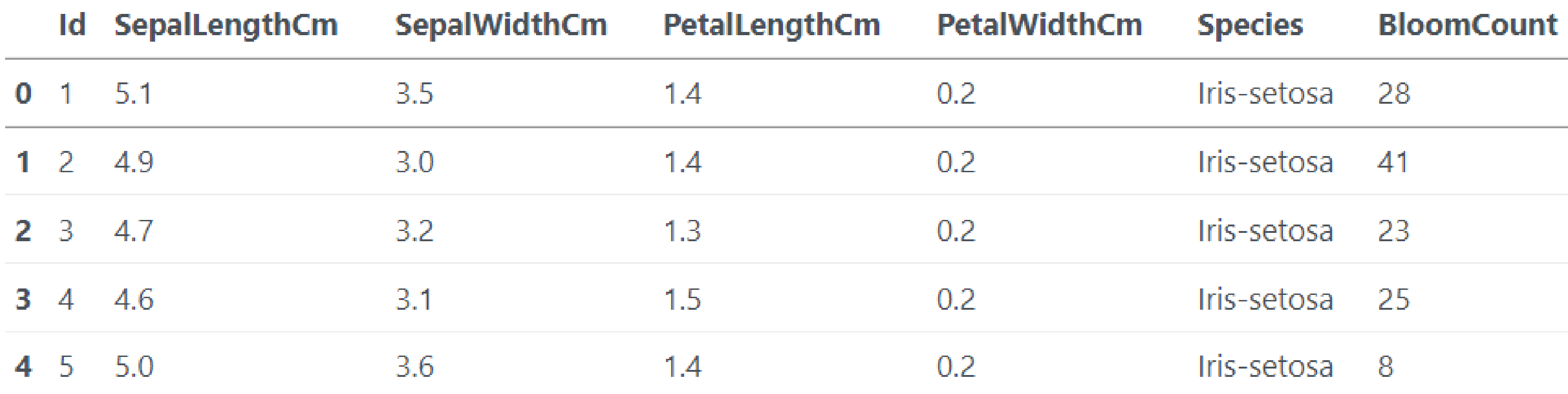
GROUPING

- Used on
  - Categorical and numeric data
- Purpose
  - Produce tidy datasets and create more robust features for modeling
- Can be an alternative to binning

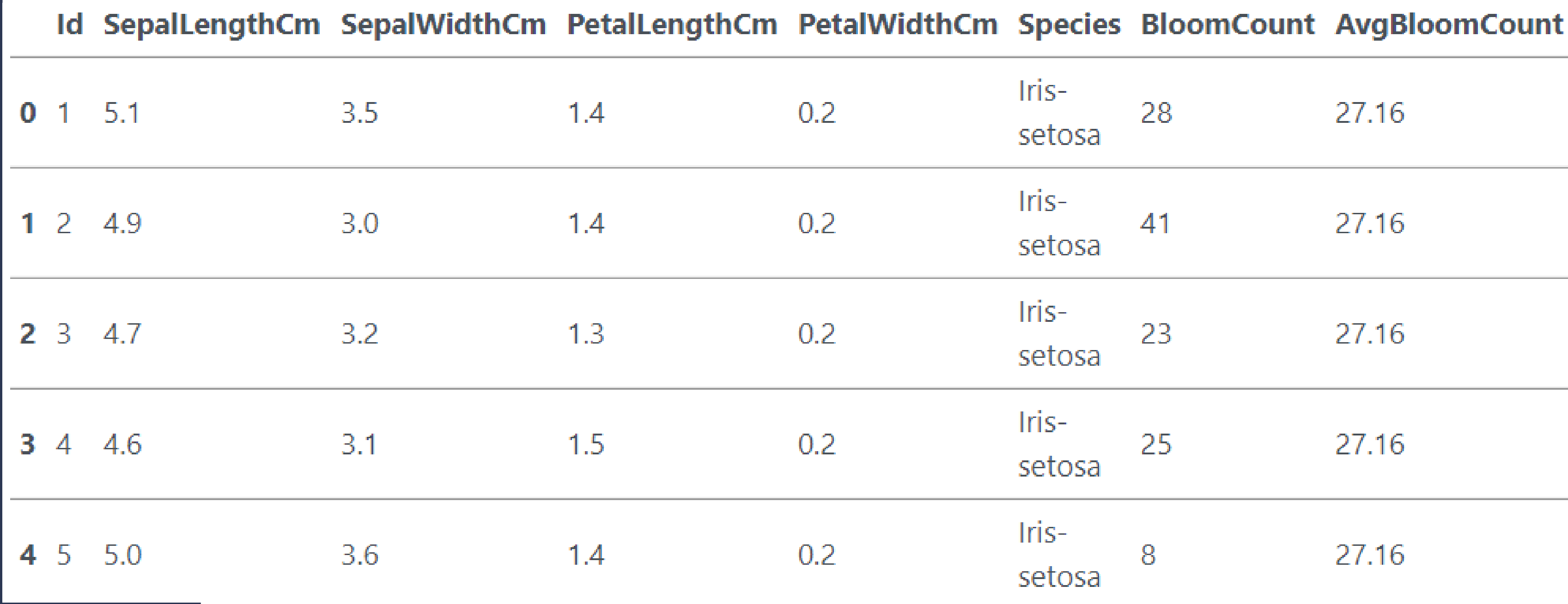


Subject	Emma	Rob
English	72	88
Science	90	65
Maths	86	74

Emma	248
Rob	227



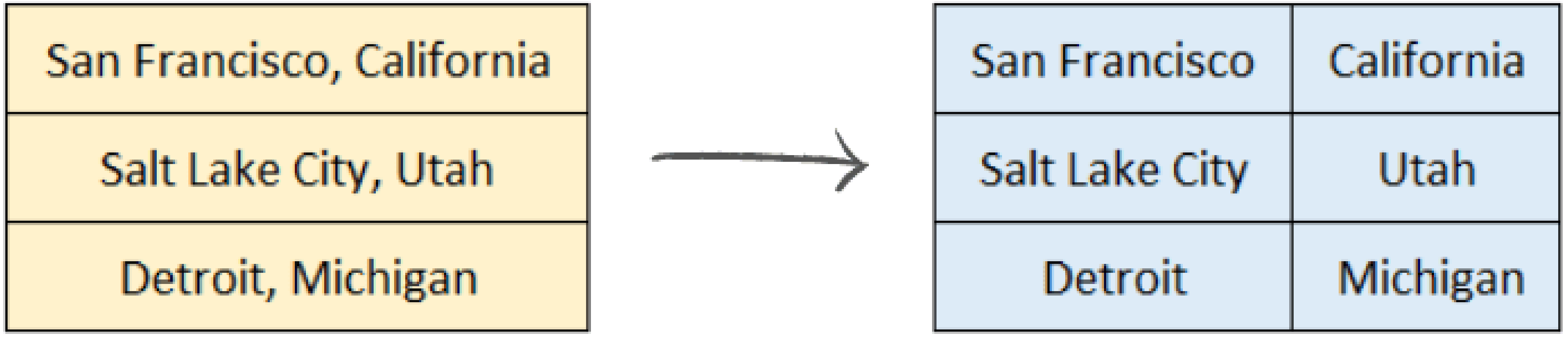
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	BloomCount
0	1	5.1	3.5	1.4	0.2	Iris-setosa	28
1	2	4.9	3.0	1.4	0.2	Iris-setosa	41
2	3	4.7	3.2	1.3	0.2	Iris-setosa	23
3	4	4.6	3.1	1.5	0.2	Iris-setosa	25
4	5	5.0	3.6	1.4	0.2	Iris-setosa	8



	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	BloomCount	AvgBloomCount
0	1	5.1	3.5	1.4	0.2	Iris-setosa	28	27.16
1	2	4.9	3.0	1.4	0.2	Iris-setosa	41	27.16
2	3	4.7	3.2	1.3	0.2	Iris-setosa	23	27.16
3	4	4.6	3.1	1.5	0.2	Iris-setosa	25	27.16
4	5	5.0	3.6	1.4	0.2	Iris-setosa	8	27.16

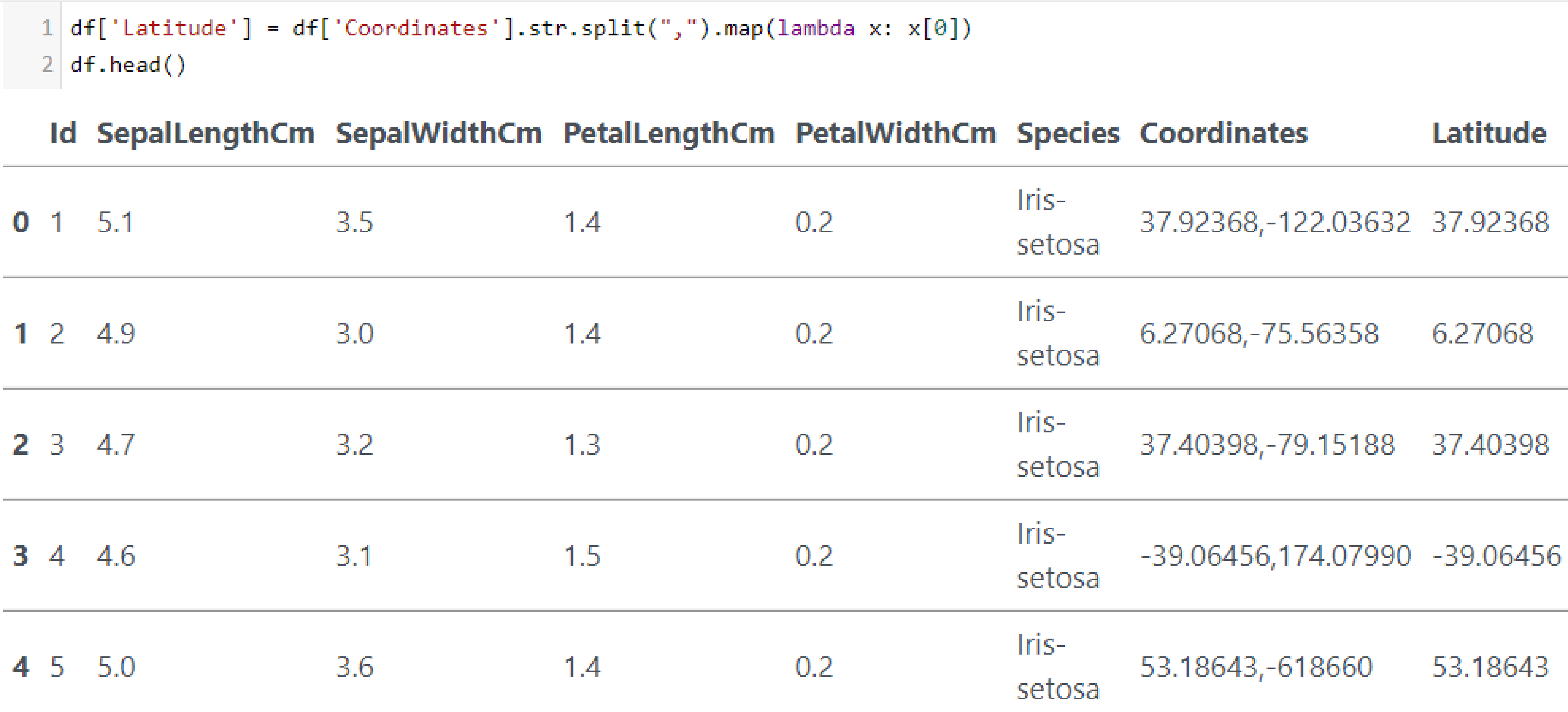
FEATURE SPLITTING

- Used on
  - Categorical and nominal data
- Purpose
  - Create features for ML algorithms
  - Enables binning and grouping
  - Improve model performance
- Flexible and used often



San Francisco, California
Salt Lake City, Utah
Detroit, Michigan

San Francisco	California
Salt Lake City	Utah
Detroit	Michigan

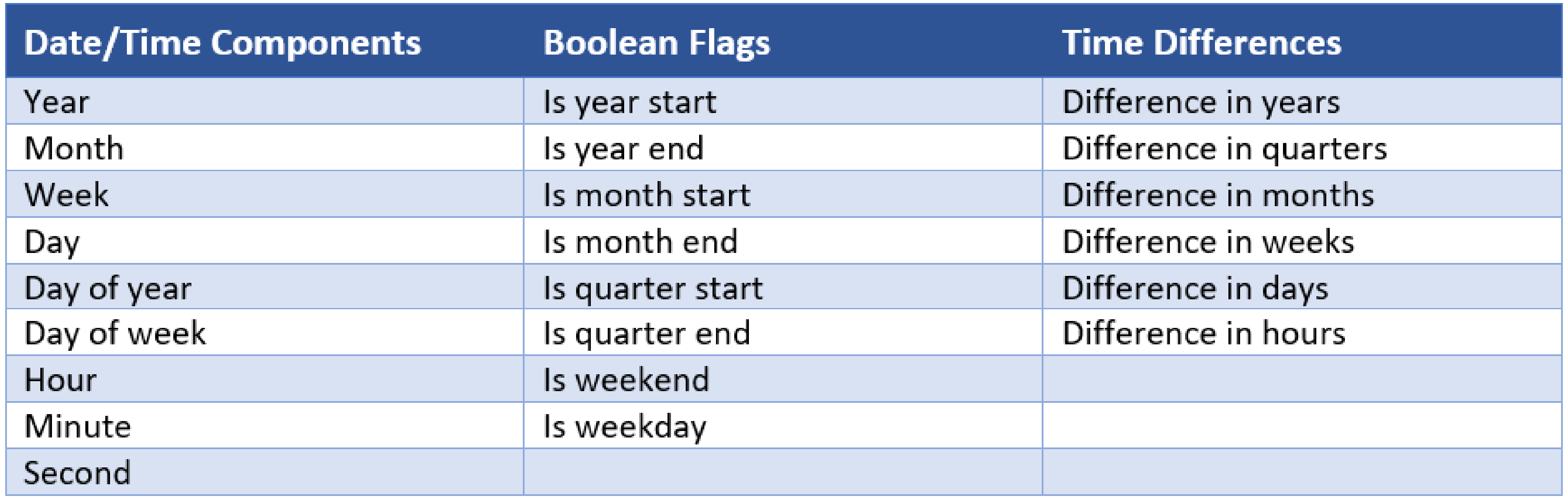


```
df['Latitude'] = df['Coordinates'].str.split(",").map(lambda x: x[0])
df.head()
```

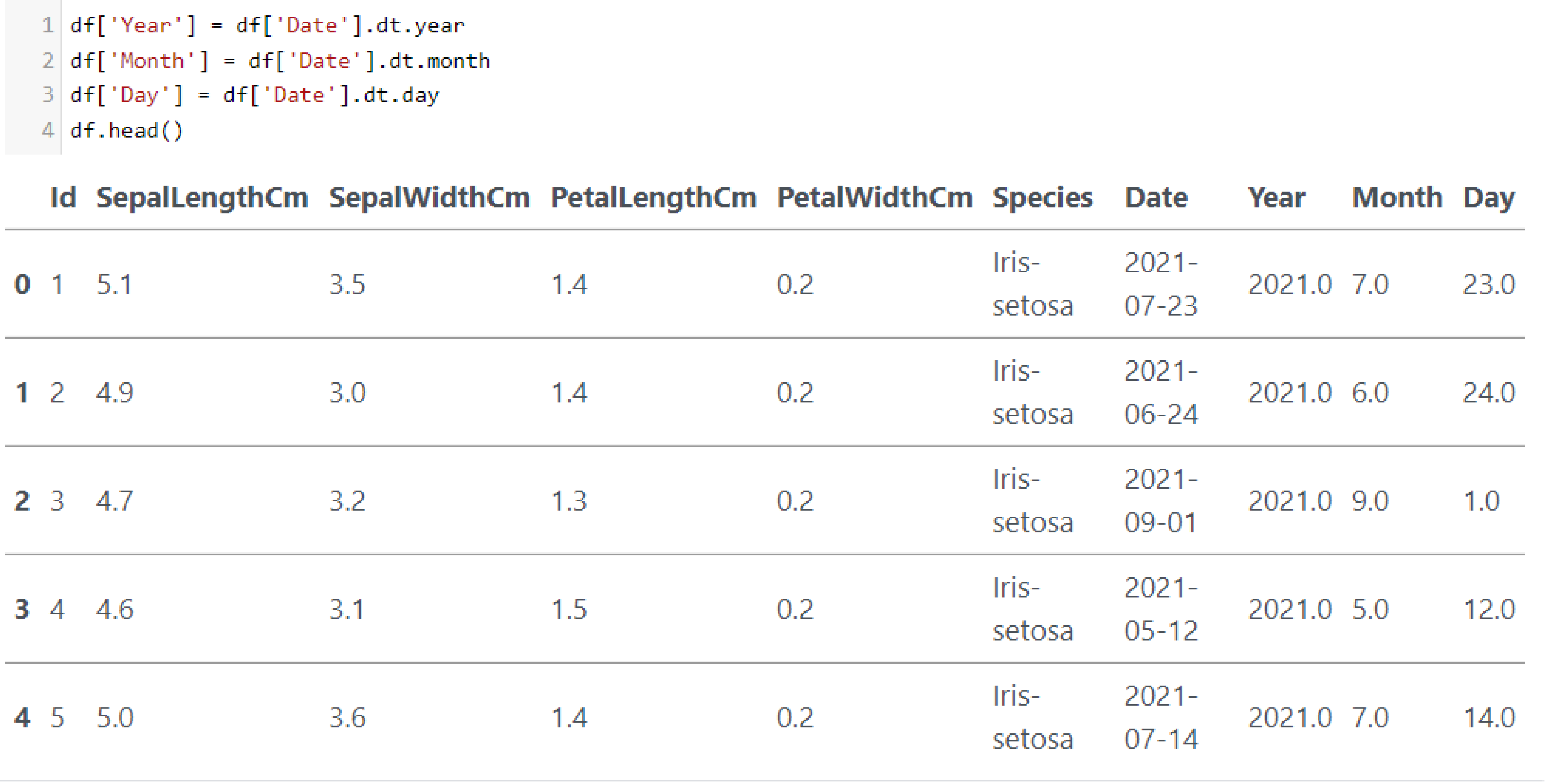
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Coordinates	Latitude
0	1	5.1	3.5	1.4	0.2	Iris-setosa	37.92368,-122.03632	37.92368
1	2	4.9	3.0	1.4	0.2	Iris-setosa	6.27068,-75.56358	6.27068
2	3	4.7	3.2	1.3	0.2	Iris-setosa	37.40398,-79.15188	37.40398
3	4	4.6	3.1	1.5	0.2	Iris-setosa	-39.06456,174.07990	-39.06456
4	5	5.0	3.6	1.4	0.2	Iris-setosa	53.18643,-618660	53.18643

EXTRACTING DATE FEATURES

- Used on
  - Date/timestamp data
- Purpose
  - Create features for ML algorithms
  - Enables binning and grouping
  - Improve model performance
- Flexible and used often



Date/Time Components	Boolean Flags	Time Differences
Year	Is year start	Difference in years
Month	Is year end	Difference in quarters
Week	Is month start	Difference in months
Day	Is month end	Difference in weeks
Day of year	Is quarter start	Difference in days
Day of week	Is quarter end	Difference in hours
Hour	Is weekend	
Minute	Is weekday	
Second		

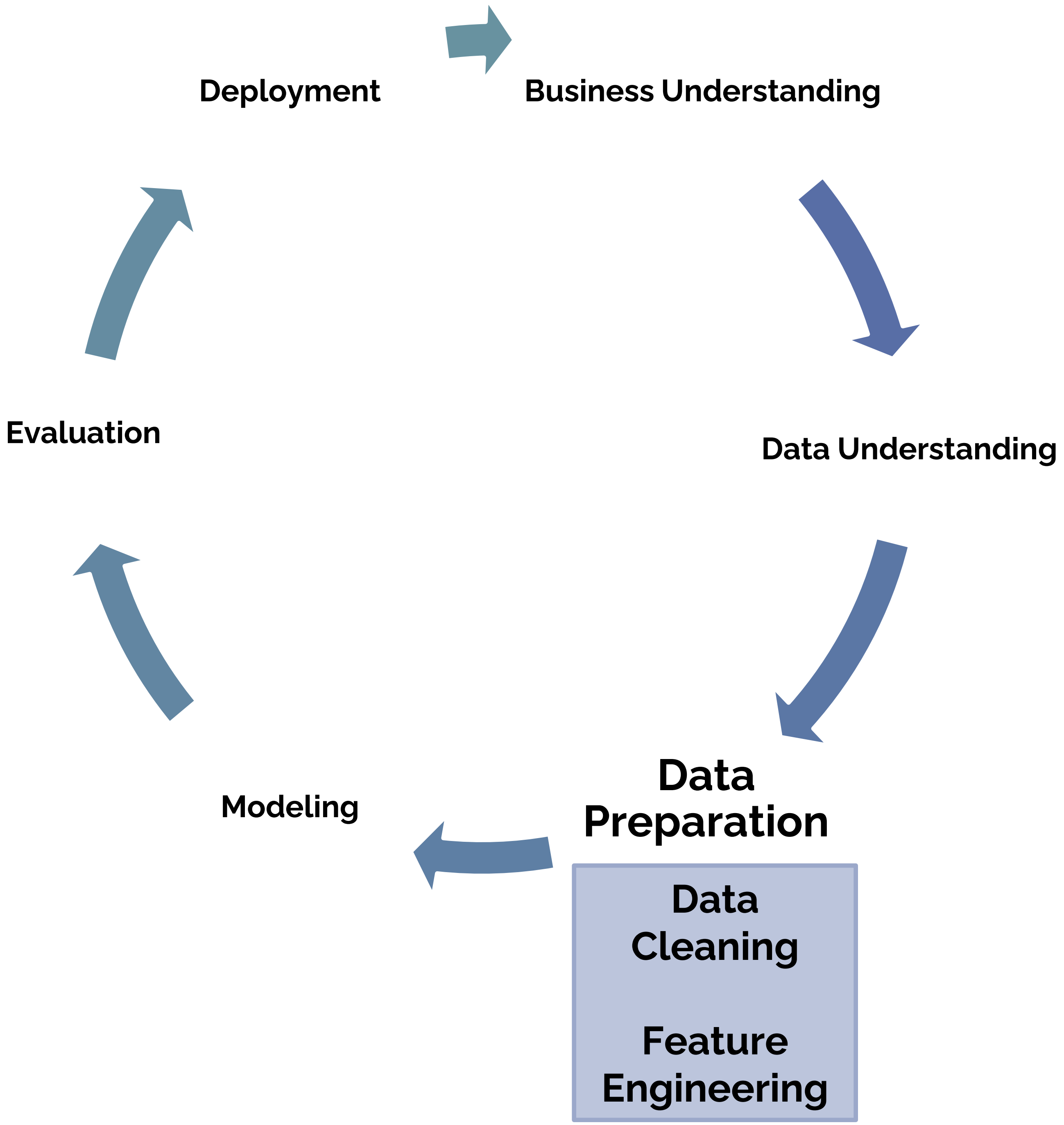


```
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
df.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Date	Year	Month	Day
0	1	5.1	3.5	1.4	0.2	Iris-setosa	2021-07-23	2021.0	7.0	23.0
1	2	4.9	3.0	1.4	0.2	Iris-setosa	2021-06-24	2021.0	6.0	24.0
2	3	4.7	3.2	1.3	0.2	Iris-setosa	2021-09-01	2021.0	9.0	1.0
3	4	4.6	3.1	1.5	0.2	Iris-setosa	2021-05-12	2021.0	5.0	12.0
4	5	5.0	3.6	1.4	0.2	Iris-setosa	2021-07-14	2021.0	7.0	14.0

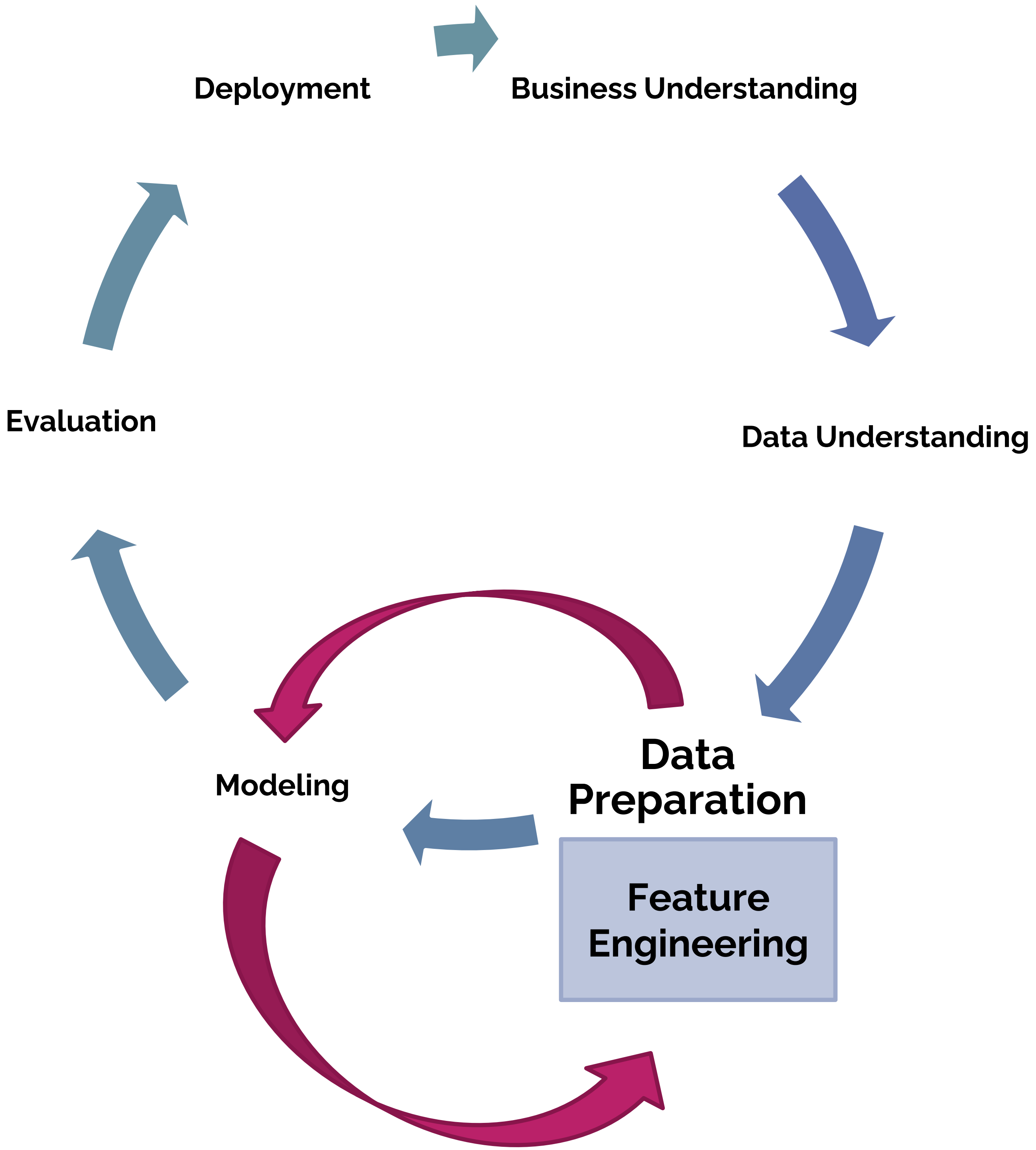
BEST PRACTICES

- Goals of Feature Engineering
  - Create, select, manipulate, and transform features for machine learning models
- Purpose of Feature Engineering
  - Meet algorithm requirements
  - Improve the performance of machine learning models
  - Better interpretability of relationships



FEATURE ENGINEERING

- Feature Engineering can be iterative with modeling
- Create and test new features to improve model performance



FEATURE ENGINEERING TECHNIQUES

- Encoding
- Binning
- Grouping
- Feature Splitting
- Extracting Date