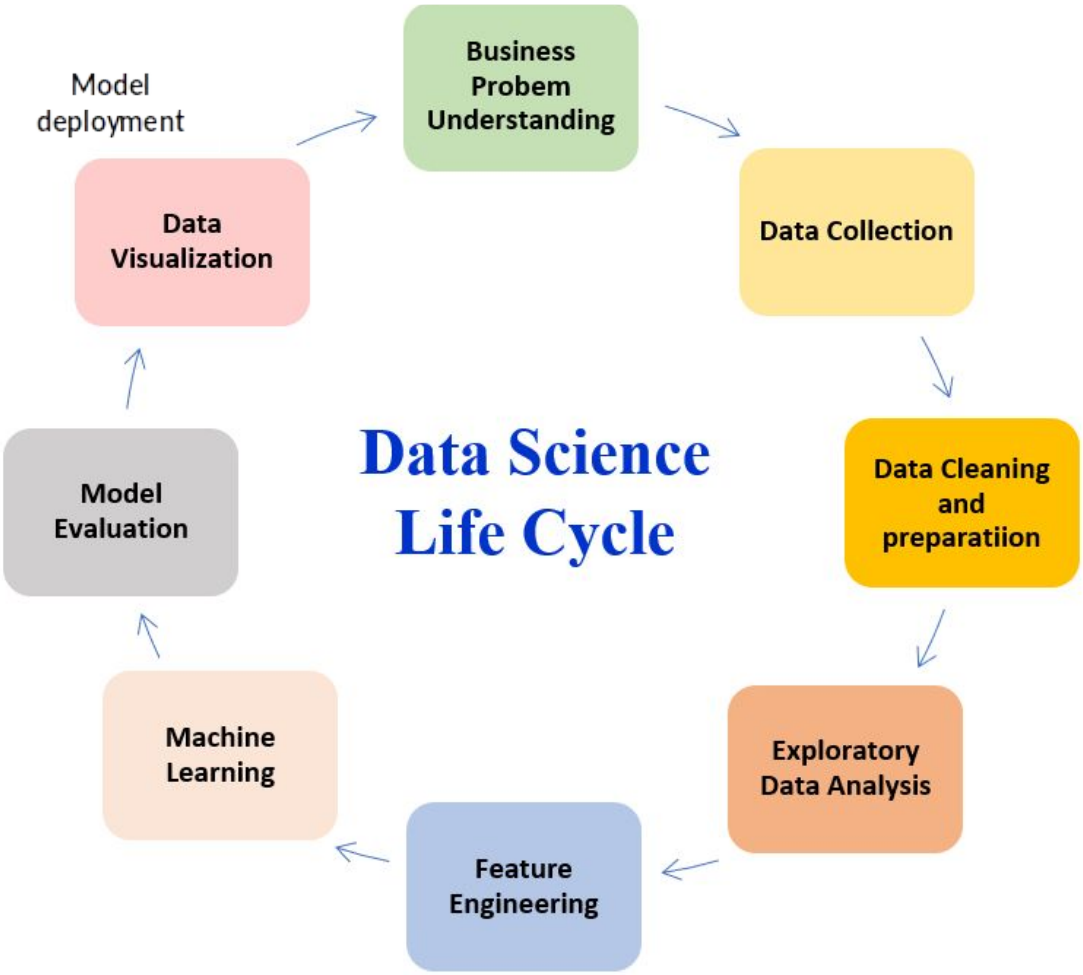
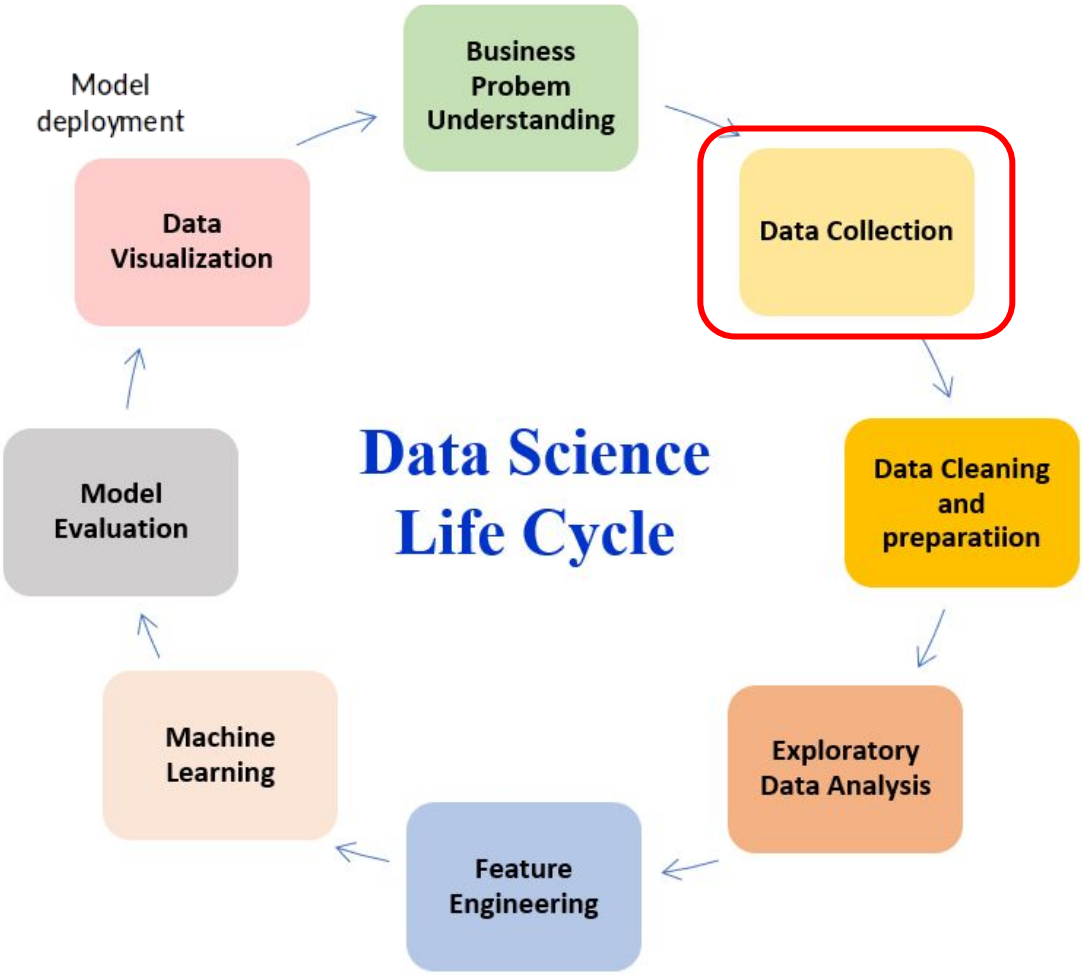
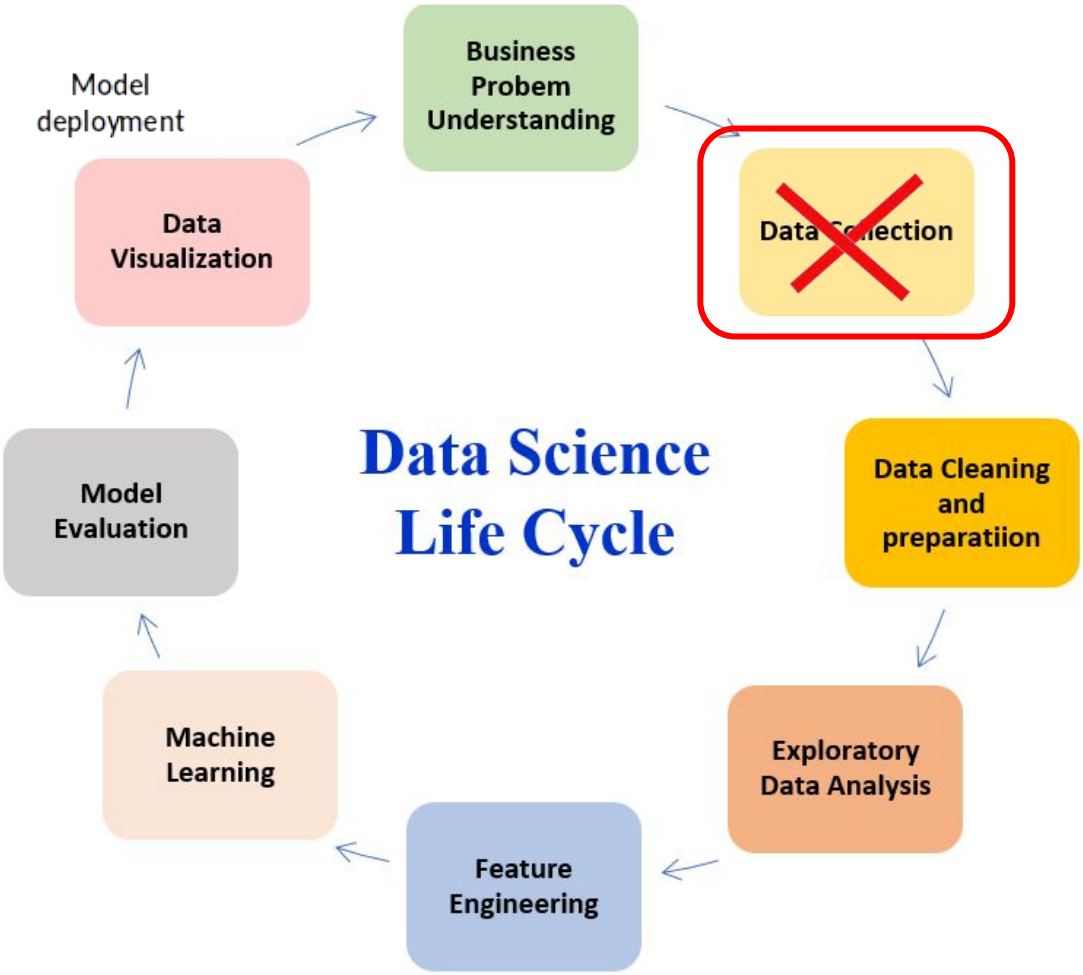
Data Science First Steps

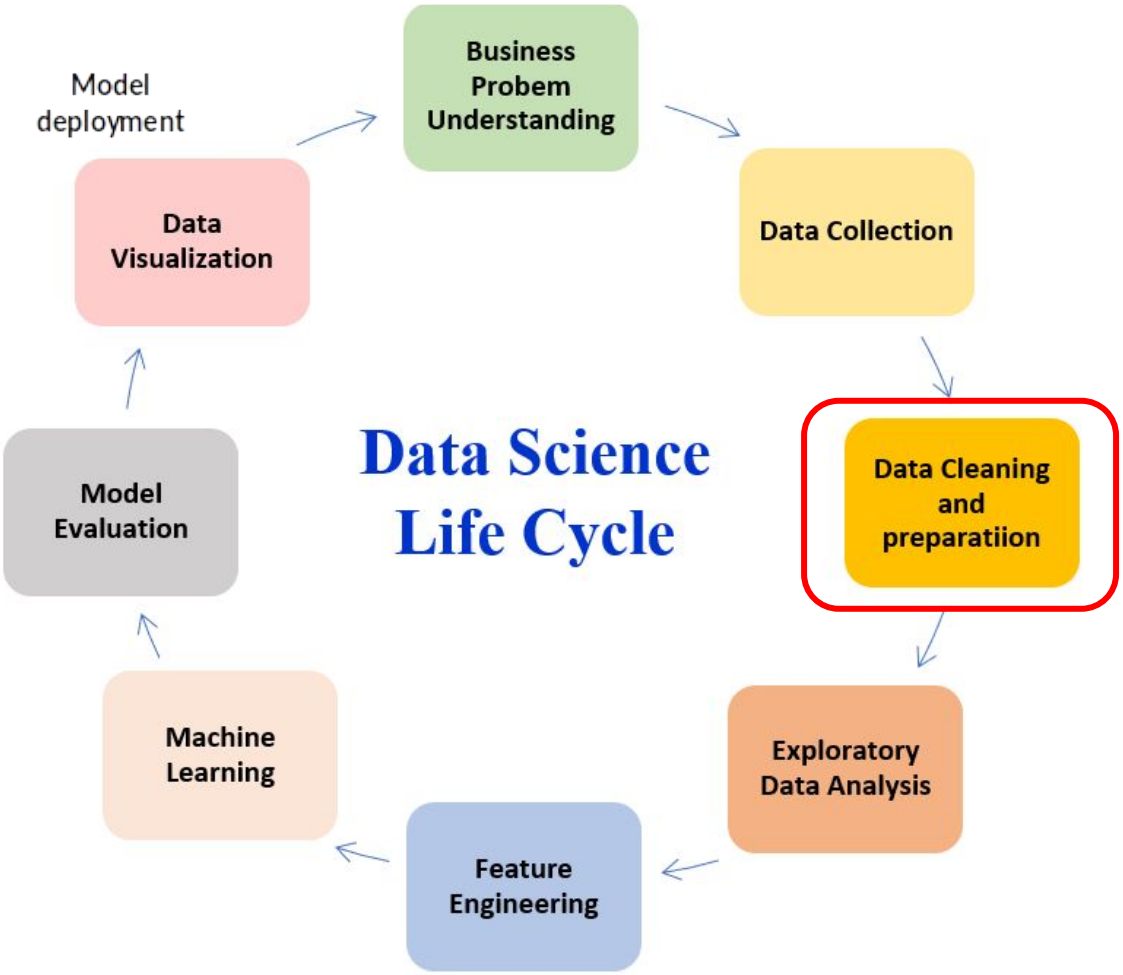

Data Science
Life Cycle
Business
Probem
Understanding
Data Collection
Data Cleaning
and
preparatiion
Exploratory
Data Analysis
Feature
Engineering
Machine
Learning
Model
Evaluation
Data
Visualization
Model
deployment

Business problem understanding

- State clearly the problem to be solved and why.
- Define the potential value of the project.
- Identify the project risks including ethical considerations.
- Develop and communicate a high-level, flexible project plan.


Data Science
Life Cycle
Business
Probem
Understanding
Data Collection
Data Cleaning
and
preparatiion
Exploratory
Data Analysis
Feature
Engineering
Machine
Learning
Model
Evaluation
Data
Visualization
Model
deployment


Model deployment
Business Probem Understanding
Data Collection
Data Cleaning and preparatiion
Exploratory Data Analysis
Feature Engineering
Machine Learning
Model Evaluation
Data Visualization
Data Science Life Cycle


Data Science
Life Cycle
Business
Probem
Understanding
Data Collection
Data Cleaning
and
preparatiion
Exploratory
Data Analysis
Feature
Engineering
Machine
Learning
Model
Evaluation
Data
Visualization
Model
deployment

First Inspection

- Check the dataset dimensions.
- Look at the columns and their values.
- Check column types (int, float, etc.).
- Convert some columns to a specific type (e.g., int, datetime ,etc.).

First Inspection

- Check the dataset dimensions. → **Pandas: .shape**
- Look at the columns and their values. → **Pandas: head, tail, describe, info**
- Check column types (int, float, etc.). → **Pandas: .dtypes, info**
- Convert some columns to a specific type (e.g., int, datetime ,etc.). → **Pandas: astype, to_datetime**

Data cleaning and preparation

- Remove unnecessary features.

Data cleaning and preparation

- Remove unnecessary features.
- Deal with null values:
  - Remove the whole row.
  - Remove whole column.
  - Value imputation: mean, median, mode, apply ML.

Data cleaning and preparation

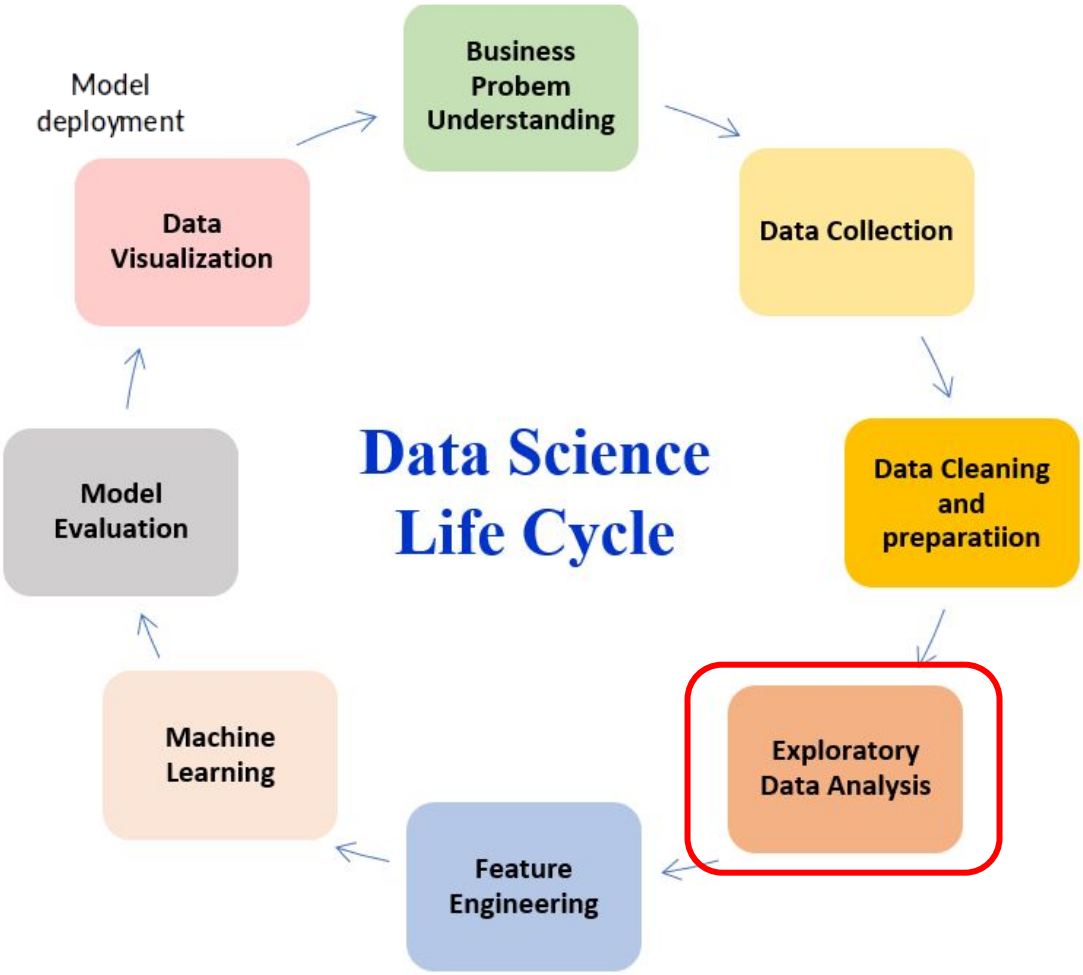
- Remove unnecessary features.
- Deal with null values:
  - Remove the whole row.
  - Remove whole column.
  - Value imputation: mean, median, mode, apply ML.
- Remove duplicates if errors.

Data cleaning and preparation

- Remove unnecessary features.
- Deal with null values:
  - Remove the whole row.
  - Remove whole column.
  - Value imputation: mean, median, mode, apply ML.
- Remove duplicates if errors.
- Outliers:
  - Remove or impute them if errors.
  - Decide what to do if they are not errors.

Data cleaning and preparation

- Remove unnecessary features. → **Pandas: drop**
- Deal with null values: → **Pandas: isnull, isna**
  - Remove whole row. → **Pandas: dropna**
  - Remove whole column. → **Pandas: dropna**
  - Value imputation: mean, median, mode, apply ML. → **Pandas: fillna, mean, median, KNNImputer**
- Remove duplicates if errors. → **Pandas: duplicated, drop_duplicates**
- Outliers:
  - Remove or impute them if errors.
  - Decide what to do if they are not errors.


Data Science
Life Cycle
Business
Probem
Understanding
Data Collection
Data Cleaning
and
preparatiion
Exploratory
Data Analysis
Feature
Engineering
Machine
Learning
Model
Evaluation
Data
Visualization
Model
deployment

Exploratory Data Analysis (EDA)

Summarize main characteristics with visualizations. Some common ideas:

- Distributions of all features: histogram, boxplot, bar plot.

Exploratory Data Analysis (EDA)

Summarize main characteristics with visualizations. Some common ideas:

- Distributions of all features: histogram, boxplot, bar plot.
- Correlation matrix between all pairs of features.

Exploratory Data Analysis (EDA)

Summarize main characteristics with visualizations. Some common ideas:

- Distributions of all features: histogram, boxplot, bar plot.
- Correlation matrix between all pairs of features.
- Pair plot.

Exploratory Data Analysis (EDA)

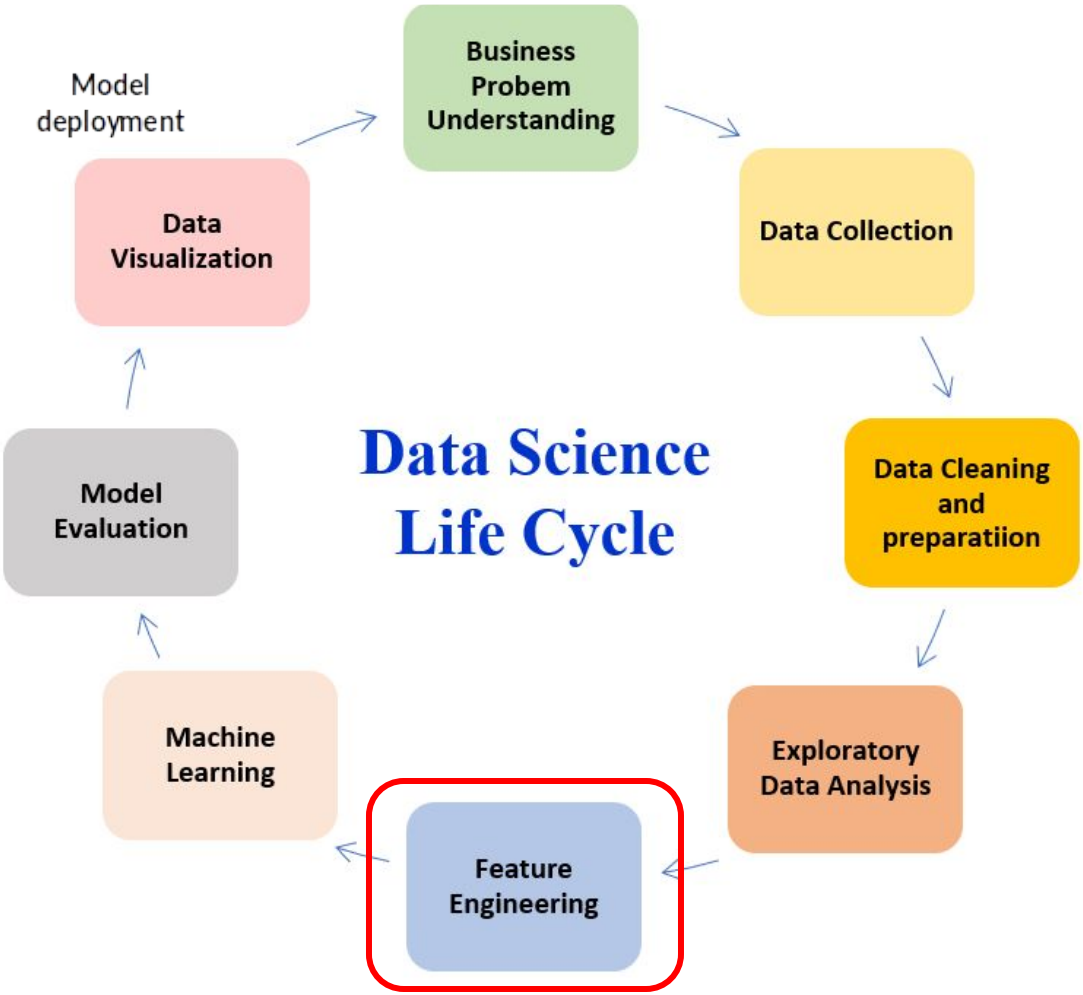
Summarize main characteristics with visualizations. Some common ideas:

- Distributions of all features: histogram, boxplot, bar plot.
- Correlation matrix between all pairs of features.
- Pair plot.
- Correlations of features with target.

Exploratory Data Analysis (EDA)

Summarize main characteristics with visualizations. Some common ideas:

- Distributions of all features: histogram, boxplot, bar plot. → **Matplotlib: hist, boxplot, barplot**
- Correlation matrix between all pairs of features. → **Pandas: corr; Seaborn: heatmap**
- Pair plot. → **Seaborn: pairplot**
- Correlations of features with target. → **Pandas: corr**


Data Science
Life Cycle
Business
Probem
Understanding
Data Collection
Data Cleaning
and
preparatiion
Exploratory
Data Analysis
Feature
Engineering
Machine
Learning
Model
Evaluation
Data
Visualization
Model
deployment

Feature Engineering

- Preprocessing data:
  - Encode categorical features: one-hot encoding, label encoding.
  - Binning.
  - Drop highly correlated features.

One-Hot Encoding

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Ordinal Encoding

SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (NUMERICAL)
None	0
Low	1
Medium	2
High	3
Very-High	4

Feature Engineering

- Preprocessing data:
  - Encode categorical features: one-hot encoding, ordinal encoding.
  - Binning.
  - Drop highly correlated features.

Feature Engineering

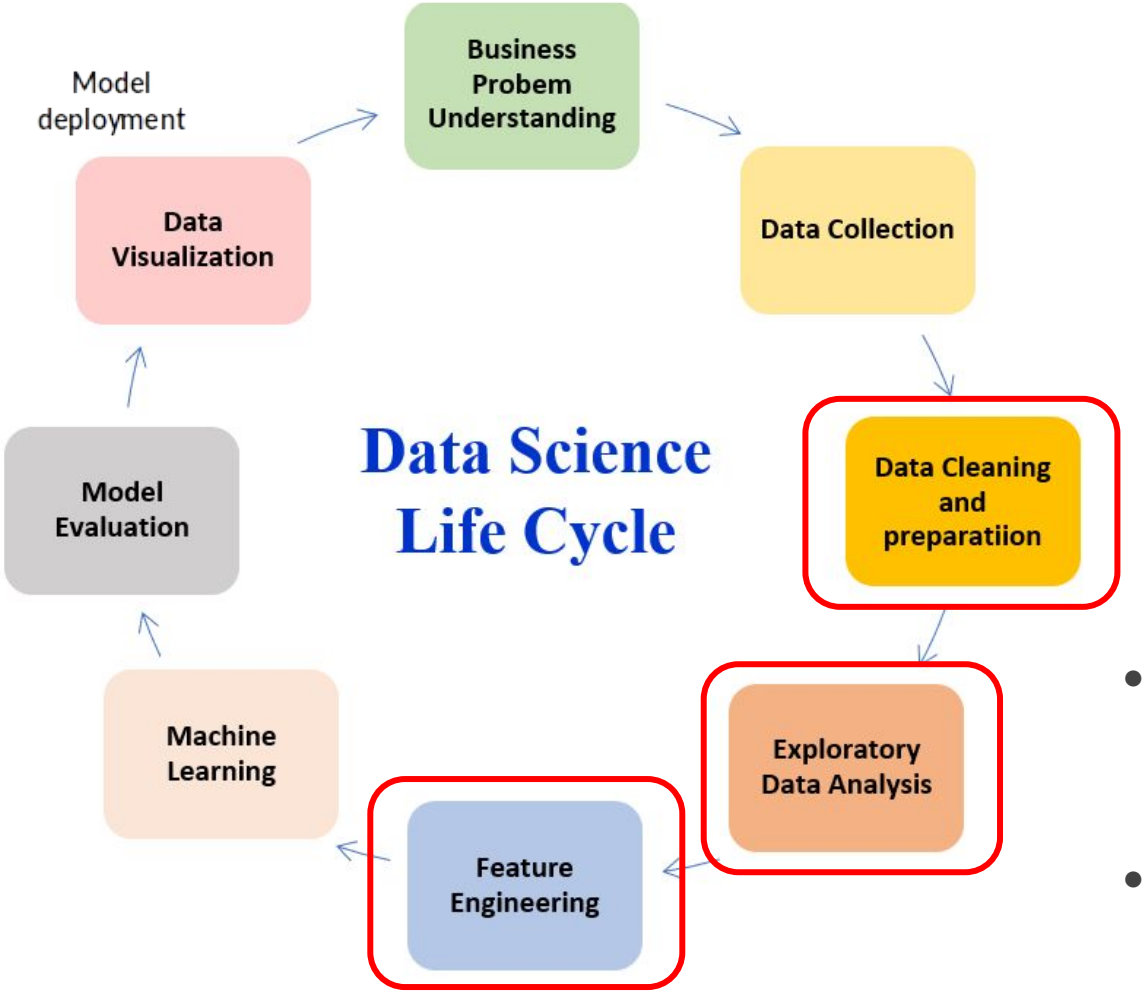
- Preprocessing data:
  - Encode categorical features:
    - One-hot encoding.
    - Ordinal encoding.
  - Binning.
  - Drop highly correlated features.
- Feature selection: RFE, statistical tests, ML models.

Feature Engineering

- Preprocessing data:
  - Encode categorical features:
    - One-hot encoding.
    - Ordinal encoding.
  - Binning.
  - Drop highly correlated features.
- Feature selection: RFE, statistical tests, ML models.
- Add new features: from external datasets or derived from existing features.

Feature Engineering

- Preprocessing data:
  - Encode categorical features:
    - One-hot encoding. → **Pandas: get_dummies**
    - Ordinal encoding. → **Sklearn: OrdinalEncoder**
  - Binning. → **Pandas: cut**
  - Drop highly correlated features. → **Pandas: corr**
- Feature selection: RFE, statistical tests, ML models. → **Sklearn: RFE**
- Add new features: from external datasets or derived from existing features.



- In a single notebook (different sections).

  or...

- In 3 independent notebooks.