https://ds3.ucsd.edu/ | ds3@ucsd.edu

data science
student society

Life Cycle of a Data Science Project

Consulting Subcommittee
Qiaoxuan (Josh) Wang
Joyce Lu

Table of Contents

1) Background
   - What’s a Data Science Project?
   - What sets them apart?
   - Why Call it a Lifecycle?
2) Life Cycle Stages
3) Engagement Activity
4) Q&A

data science
student society

[01]
Background

1.Background

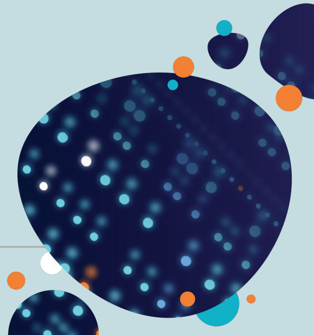
2. Life Cycle Stages

3. Engagement Activity

4. Q&A

What is a Data Science Project?

- A practical application of technical skills & domain knowledge
- Makes use of data to extract actionable insights and solve real-world problems
- Combines principles and practices from a variety of fields
- Common applications of data science in business include…
  - Quantifying Performance
  - Detecting Anomalies
  - Streamlining Operations
  - Informing Next steps



Source: [What Is Data Science? 5 Applications in Business]

1.Background

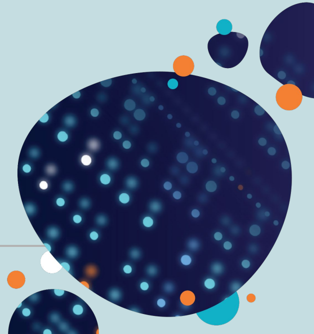
2. Life Cycle Stages

3. Engagement Activity

4. Q&A

What Sets Data Science Apart?

- “Data Science enhances the traditional and more conservative world of Statistics with advanced algorithms to enable us to make sense out of soaring amounts of data”
- Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. These insights can be used to guide decision making and strategic planning.



Source: What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

data science student society

Why Call it a Lifecycle?

- All data science projects are not built the same, so their life cycle varies as well. Still, we can picture a general lifecycle that includes some of the most common data science steps.
- The term “lifecycle” refers to the iterative steps taken to build, deliver and maintain any data science product
- The work is never truly finished – there is always something to refine!

Source: [What Is Data Science Life Cycle? Steps Explained]

[02]
Life Cycle Stages

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

Ideation

- The first step is to produce a clear definition and understanding of the problem or business case and then translate that into a data science problem with actionable steps and goals
- Starts with “why”
  - State clearly the problem to be solved and why
  - Define the potential value of the forthcoming project
  - Identify the key stakeholders
  - Identify the type of problem being solved

Source: [What is a Data Science Life Cycle? - Data Science Process]()

Data Acquisition

- Collecting data from relevant sources either in structured or unstructured form
- Some common ways to get data:
  - Kaggle
  - Public websites (i.e. government websites)
  - APIs
  - Web scraping
  - Gather your own data (i.e. surveys, questionnaire, etc.)

Source: Data Science life Cycle | Towards Data Science

1.Background

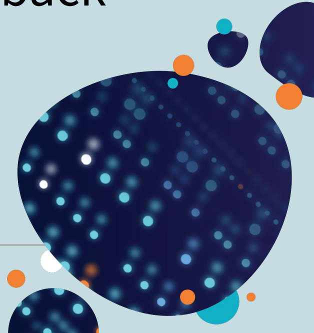
2. Life Cycle Stages

3. Engagement Activity

4. Q&A

Preparation

- Once you have the data, start exploring it. You can conduct activities such as:
  - Document the data quality
  - Clean the data (i.e. should we omit, replace, or fill in missing data?)
  - Combine various data sets to create new views
  - Load the data
  - Present initial findings to stakeholders and solicit feedback



Source: [Data Science life Cycle | Towards Data Science]

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

data science
student society

Research & Development

Conduct Exploratory Data Analysis (EDA)

- Use data visualizations to quickly understand features and their interactions
- Conduct hypothesis testing or check any assumptions that you may have
- What tools can we use?
  - Clustering and dimension reduction techniques
  - Univariate visualization
  - Bivariate visualizations and summary statistics
  - Multivariate visualizations
  - K-means Clustering
  - Linear regressions

Source: [Data Science life Cycle | Towards Data Science]

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

Research & Development (cont.)

data science student society

Model Building

- Data modeling: the process of converting raw data into a form that can be transverse into other applications as well
    - Mostly performed in using statistical tools and databases
- Steps:
    - Developing datasets for training and testing
    - Choose the appropriate model that best fits the data and the question
    - Learn whether the problem is a classification, regression, or clustering problem
    - After analyzing the model family, choose the algorithms to implement them
    - Run the model

Source: [Data Science life Cycle | Towards Data Science]

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

data science student society

Validation & Ethics Audit

Model evaluation

- Using different evaluation metrics to understand a model’s performance
- Continue to refine and reiterate the previous steps, tuning model parameters, until we reach a satisfactory performance
- Understand model strengths and weaknesses
- Interpret results

Ethics

- Are you being fair, accountable, and transparent?
- Are our findings reproducible? Do our findings/models add value?
- Do they run the risk of harming anybody?
- How can we prevent our findings from being misinterpreted?

Source: Chapter 5. Choosing and evaluating models

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

data science
student society

Present, Deploy, and Monitor

- The final stage of the Data Science Lifecycle is to communicate your findings to your stakeholders and deploy the model in production in the desired format and preferred channel
- The model you’ve built can be used on something as simple as an output on a Tableau dashboard or as complex as scaling it to the cloud for millions of users
- Once your model has been deployed, you must continually monitor & maintain its performance to ensure that it’s meeting expectations, making enhancements as needed

Source: What is a Data Science Life Cycle?

[03]
Engagement Activity

[Kaggle Dataset]

[Workshop GitHub Repo]

1.Background

2. Life Cycle Stages

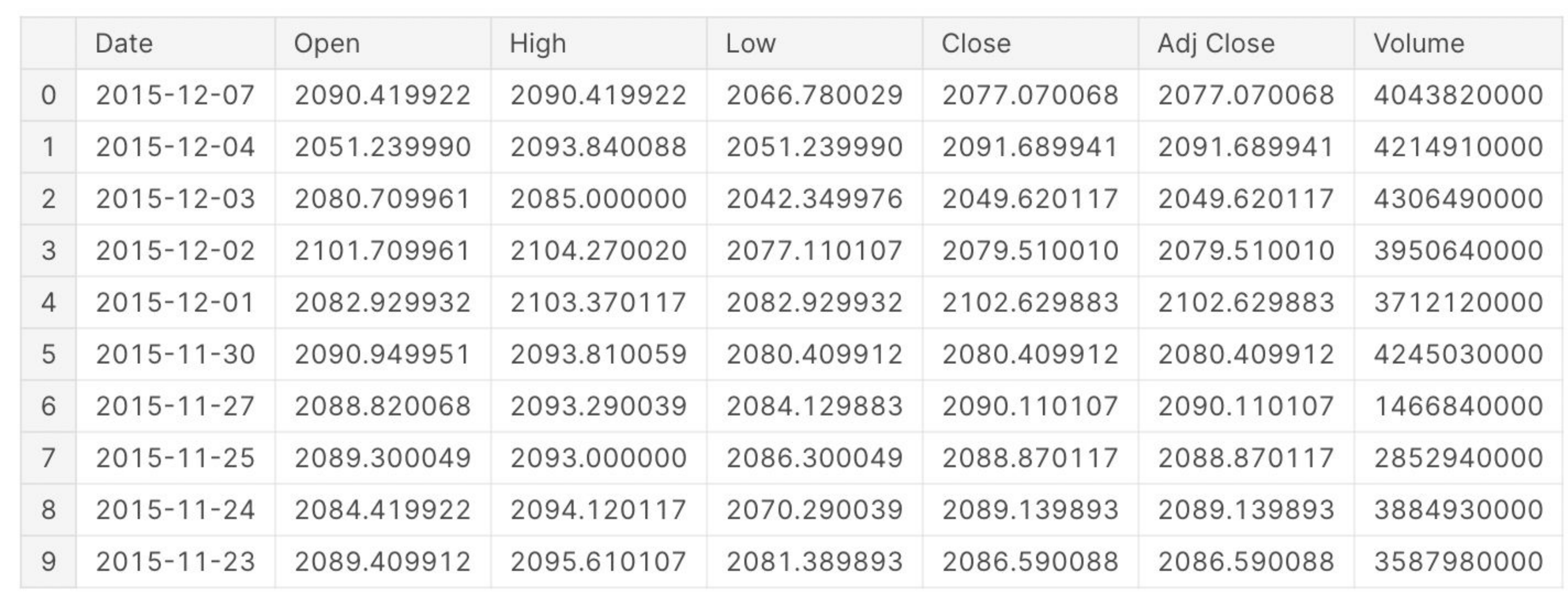
3. Engagement Activity

4. Q&A

data science student society

Yahoo Stock Price Dataset

First 10 rows



	Date	Open	High	Low	Close	Adj Close	Volume
0	2015-12-07	2090.419922	2090.419922	2066.780029	2077.070068	2077.070068	4043820000
1	2015-12-04	2051.239990	2093.840088	2051.239990	2091.689941	2091.689941	4214910000
2	2015-12-03	2080.709961	2085.000000	2042.349976	2049.620117	2049.620117	4306490000
3	2015-12-02	2101.709961	2104.270020	2077.110107	2079.510010	2079.510010	3950640000
4	2015-12-01	2082.929932	2103.370117	2082.929932	2102.629883	2102.629883	3712120000
5	2015-11-30	2090.949951	2093.810059	2080.409912	2080.409912	2080.409912	4245030000
6	2015-11-27	2088.820068	2093.290039	2084.129883	2090.110107	2090.110107	1466840000
7	2015-11-25	2089.300049	2093.000000	2086.300049	2088.870117	2088.870117	2852940000
8	2015-11-24	2084.419922	2094.120117	2070.290039	2089.139893	2089.139893	3884930000
9	2015-11-23	2089.409912	2095.610107	2081.389893	2086.590088	2086.590088	3587980000

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

1. Ideation

data science student society

- First, we need to define and understand the problem we want to solve:
- **What is the problem that we are trying to solve?**
- In this example, the author asked: What will be the value of Yahoo's assets in the near future?

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

2. Data Acquisition

- Next, we want to collect the data we need
- We have our data already!
- **What if today the data isn’t provided? What are some ways we can collect the data we need? What data do we want to collect?**

3. Data Preparation / Exploration

- **What are some tasks we can perform in this step?**
  - Data Cleaning
  - Combining it with other datasets
  - Visualizing the data
  - Loading the data into a target location
  - …

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

1.Background

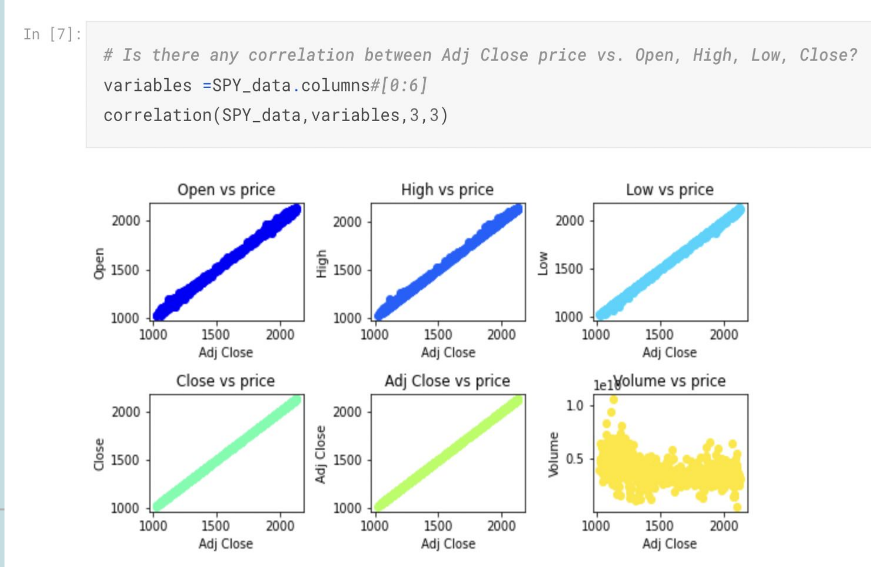
2. Life Cycle Stages

3. Engagement Activity

4. Q&A

4. Research & Development

- Conducting Exploratory Data Analysis (EDA)
  - **Do we have any assumptions or hypotheses about the data?**
  - **What can you do / what tools can you use to quickly test your assumptions and understand the relationships between various features?**



```
In [7]:
# Is there any correlation between Adj Close price vs. Open, High, Low, Close?
variables =SPY_data.columns#[0:6]
correlation(SPY_data,variables,3,3)
```

1.Background

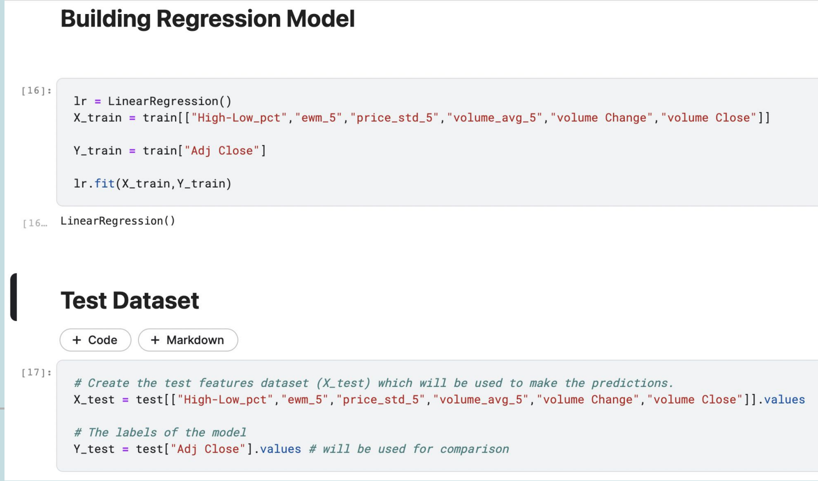
2. Life Cycle Stages

3. Engagement Activity

4. Q&A

4. Research & Development

- Model Building
  - **What approach can we take?**
  - **What model is best fit for the data?**
- In this example, the author uses a regression model
- The dataset is split into train & test sets



Building Regression Model

```
[16]:
lr = LinearRegression()
X_train = train[["High-Low_pct","ewm_5","price_std_5","volume_avg_5","volume Change","volume Close"]]

Y_train = train["Adj Close"]

lr.fit(X_train,Y_train)
```

```
[16…  LinearRegression()
```

Test Dataset

```
[17]:
# Create the test features dataset (X_test) which will be used to make the predictions.
X_test = test[["High-Low_pct","ewm_5","price_std_5","volume_avg_5","volume Change","volume Close"]].values

# The labels of the model
Y_test = test["Adj Close"].values # will be used for comparison
```

1.Background

2. Life Cycle Stages

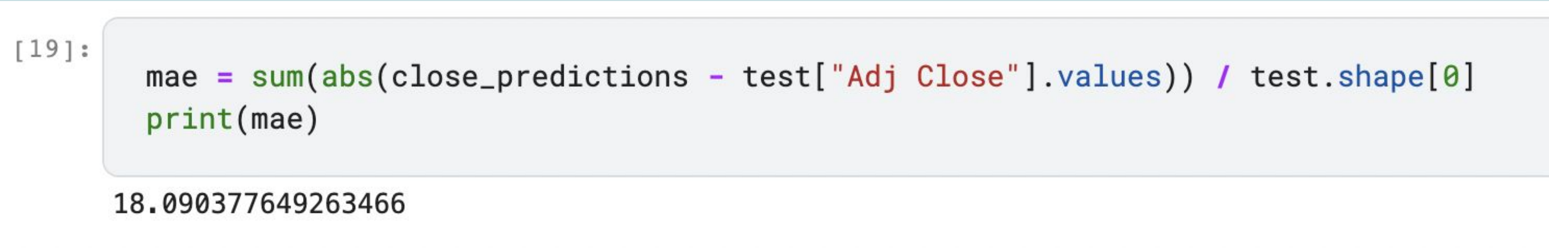
3. Engagement Activity

4. Q&A

data science
student society

5. Validation & Ethics

- Validation
- In this case study, the author uses Mean Absolute Error (MAE) for model evaluation
- **What are some ways we can improve our model?**



```
[19]:  mae = sum(abs(close_predictions - test["Adj Close"].values)) / test.shape[0]
       print(mae)
```

```
18.090377649263466
```

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

5. Validation & Ethics

- Ethical Considerations
- **Are we being fair, accountable and transparent?**
- **Do they run the risk of harming anybody? If so, how?**
- **How can we prevent our findings from being misinterpreted?**
- **Are our findings reproducible?**

1.Background

2. Life Cycle Stages

3. Engagement Activity

4. Q&A

6. Present, Deploy, and Monitor

- We are at the final stage of the lifecycle of data science
- If we are deploying a product, the model is exposed to real time data flowing into the system and outputs are being generated
- It is important to maintain and keep updating the model and the data

[04]
Q&A

Find us on social media!

@DS3UCSD

@DS3UCSD

@DS3.UCSD

https://tinyurl.com/discordds3

ds3.ucsd.edu

ds3@ucsd.edu

Thank you!