IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<HTET AUNG LYNN>
<Date: 18 – Sep – 2024>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies – analysis steps
  - Data collection (SpaceX API + Space X Wikipedia)
  - Data wrangling (modify correct format)
  - Exploratory data analysis (investigate basic stats)
  - Data visualization (visual analysis)
  - Predictive analysis (data modelling + evaluation)
- Summary of all results – finding best model using GridSearchCV
  - Logistic Regression (84.6% accurate)
  - Support Vector Machine (84.8% accurate)
  - K-Nearest Neighbors (84.8% accurate)
  - Decision Tree (87.5% accurate) best

Introduction

- Background
  - Space X advertised Falcon 9 rocket launches with a cost of $62 million,.
  - Other providers use $165 million.
  - Much cost saving for Space X is recovering the first stage (stage 1).
  - The Space Y (our project) want to compete with Space X
- Problems you want to find answers
  - Gather information about Space X
  - Clean data into correct format
  - Create dashboard for our team
  - Predict the first stage reusage by machine learning model

Section 1

Methodology

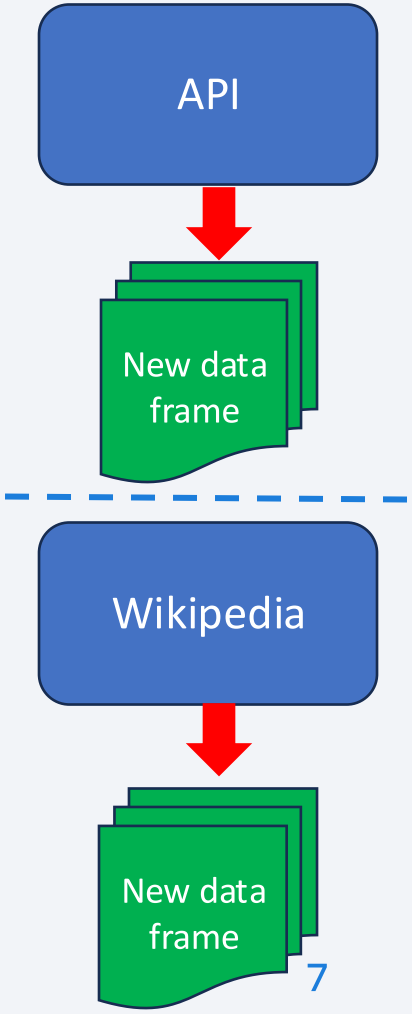
Methodology

Executive Summary

- Data collection methodology:
  - Collect data from Space X (Public API + Wikipedia)
- Perform data wrangling
  - Investigate missing values and classify landing classes (1-Success/0-Fail)
- Perform exploratory data analysis (EDA) using data visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build models (Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors)
  - Tune model using GridSearchCV
  - Evaluate for best model based on accuracy scores

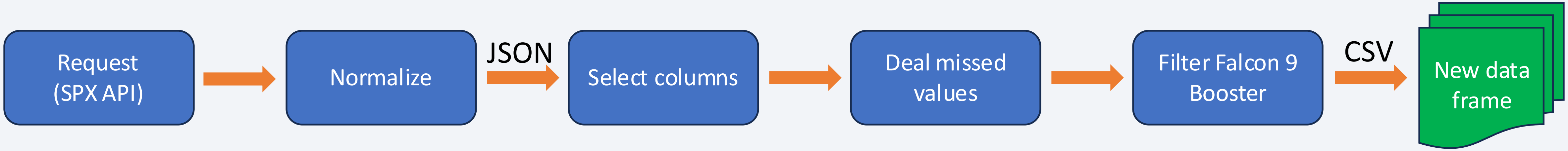
Data Collection

- Data were collected from SpaceX public API and Wikipedia
  - SpaceX Public API data is collected into JSON format using Requests | LINK
    - Get rocket launch data in JSON format
    - Normalize the JSON data
    - Filter required column
    - Create a new data frame from dictionary
  - Falcon 9 historical launch record from Wikipedia using BeautifulSoup | LINK
    - Get Falcon 9 rocket launch data response
    - Transform response data into HTML table using BeautifulSoup
    - Create empty dictionary and add data with correct format for each column
    - Create a new data frame from dictionary



Data Collection – Space X API

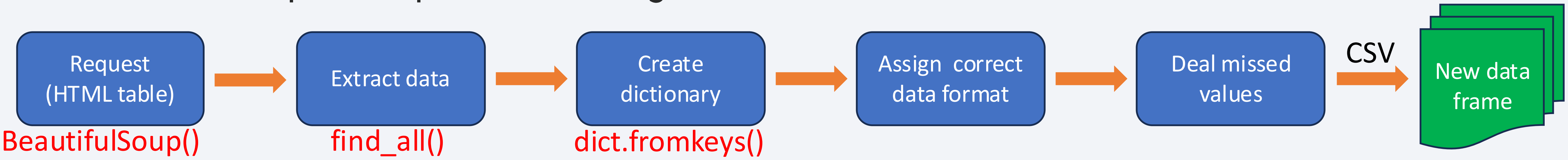
- Total 43 columns include in JSON format data
- Select 6 columns ('rocket', 'payloads', 'launchpad', 'cores', 'flight_number','date_utc')
- Created empty dictionary and append related data into it from associated Space X API path
- The most complicated part is extracting 'cores' data from the cores column



SpaceX API call notebook | GitHub

Data Collection - Scraping

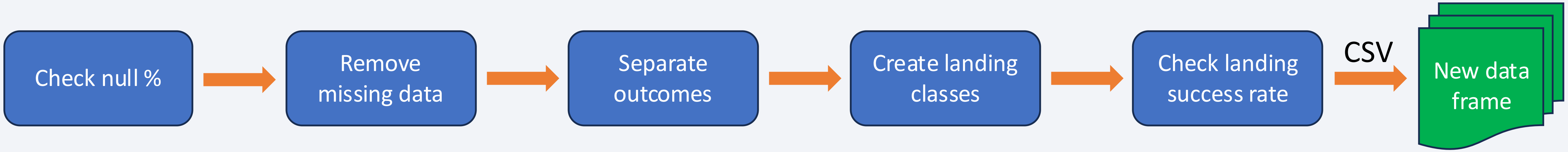
- 26 HTML tables are included BeautifulSoup object
- Second table is chosen to manipulate.
- 75 columns include in the table
- Created an empty dictionary and transform into a new data frame with 11 columns.
- The most complicated part is extracting data from the HTML table



Web-scraping notebook | GitHub

Data Wrangling

- Initially, check null rate and determine to remove or replace.
- Remove missing data that is not useful
- Separate bad landing outcomes from landing outcomes
- Create 'Class' column for landing outcomes (1 for success | 0 for bad)



Mapping:

True ASDS', 'True RTLS' and 'True Ocean' are set to 1

'None None', 'False ASDS', 'False Ocean', 'None ASDS' and 'False RTLS' are set to 0

Data wrangling notebook | GitHub

EDA with Data Visualization

- Scatter plots are used to know the relationship between the following columns:
    - Flight Number vs. Launch Site vs. Outcomes
    - Launch Site vs. Payload vs. Outcomes
    - Orbit Type vs. Success Rate vs. Outcomes
    - Flight Number vs. Orbit Type vs. Outcomes
    - Payload Mass vs. Orbit Type vs. Outcomes

- Line chart is used to show the time series relationship:
    - Success Rate over the years 2010 to 2020

EDA with Visualization notebook | GitHub

EDA with SQL

Exploratory Data Analysis (EDA) process includes the following tasks after creation of an empty database:

1. Display unique launch site names
2. Display 5 records where launch site name begin with 'CCA'
3. Display total payload mass carried by NASA (CRS) booster launch
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List booster names having success in drone ship and have payload mass between 4000 but and 6000 kg.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster_versions which have carried the maximum payload mass.
9. List the month names, failure landing outcomes in drone ship, booster versions, launch_site for the months in year 2015.
10. Rank the landing outcomes count (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

EDA with SQL notebook | GitHub

Build an Interactive Map with Folium

The interactive map allows easy exploration of location-based data. Using Folium, the following elements are displayed:

- **Circles**: represent launch sites, making it easy to identify their exact locations on the map.
- **Markers**: display landing outcomes with color-coded icons, allowing quick recognition of success or failure at a glance.
- **Lines**: illustrate the proximity to important areas, helping assess whether the distance is
  - safe from city limits and
  - close enough to railways or highways for supply and transportation needs.

Interactive map notebook | GitHub

Build a Dashboard with Plotly Dash

The dashboard provides dynamic data visualization using real-time information. It includes the following features and charts:

- **Dropdown menu:** Select options from a predefined list.
  - **All:** Shows success rates by total count across all launch sites.
  - **Individual launch sites:** Displays both success and failure rates (with counts) for each site.
- **Range slider:** Adjust the payload mass range from 0 to 9600 kg.
- **Scatter plot:** Visualizes data based on dropdown and payload selections.
  - **All:** Displays the relationship between Payload Mass and Class, categorized by Booster Version.
  - **Individual launch sites:** Shows the correlation between Payload Mass and Class, grouped by Booster Version for each launch site.

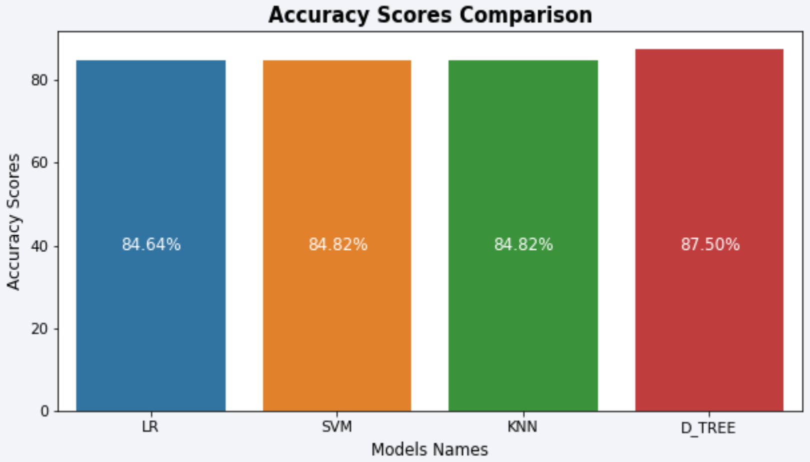
This setup allows for flexible and detailed analysis of launch data.

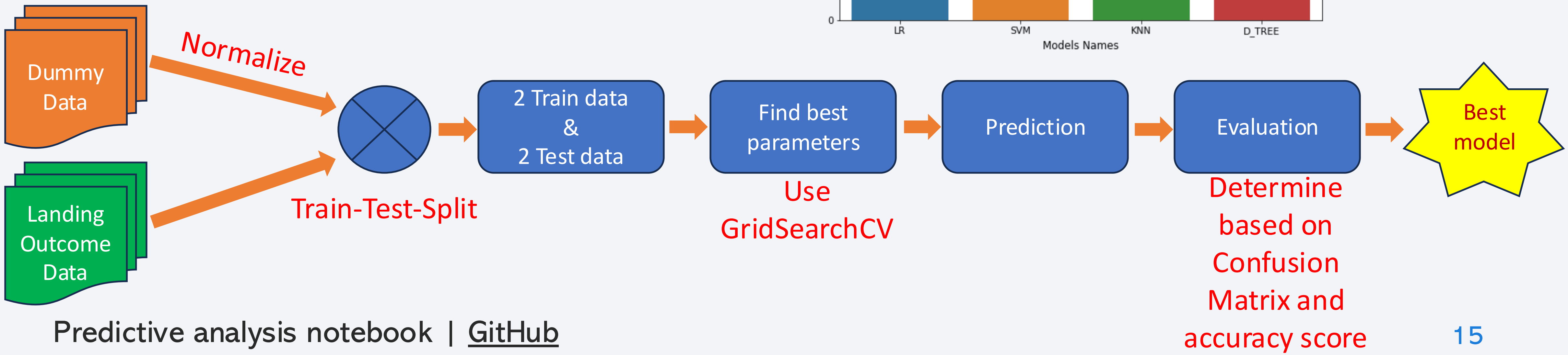
Interactive dashboard notebook | GitHub

Predictive Analysis (Classification)

All of the machine learning classification models are trained, evaluated and compared accuracy scores to choose the best model for predictive analysis. There are total 4 classification models

1. Logistic Regression
2. Support Vector Machine
3. K-Nearest Neighbor
4. Decision Tree (best model)





Predictive analysis notebook | GitHub

Results

- Figure 1: EDA for Landing Success Rate based on launch sites and landing outcomes.
- Figure 2: Interactive map to display information landing outcomes.
- Figure 3: ML model comparison to choose the best model based on accuracy score.

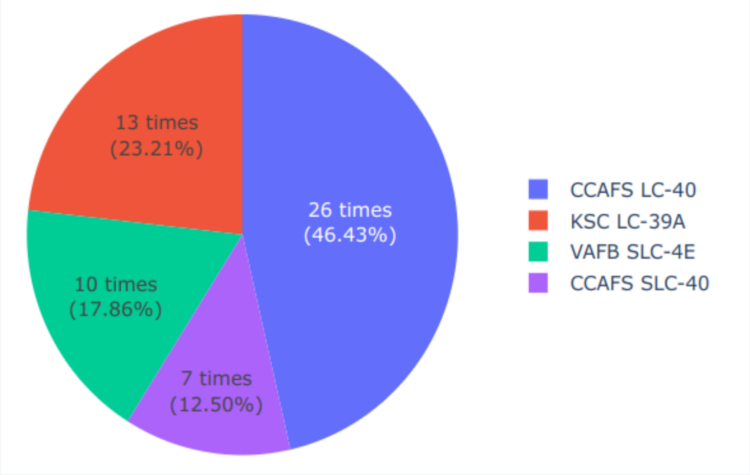


Figure 1

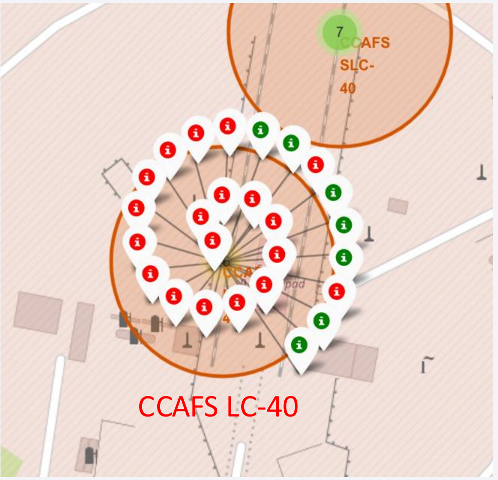


Figure 2

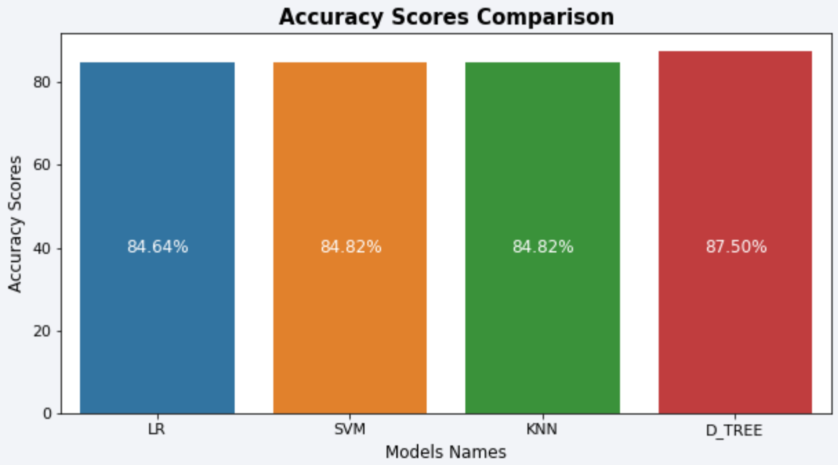


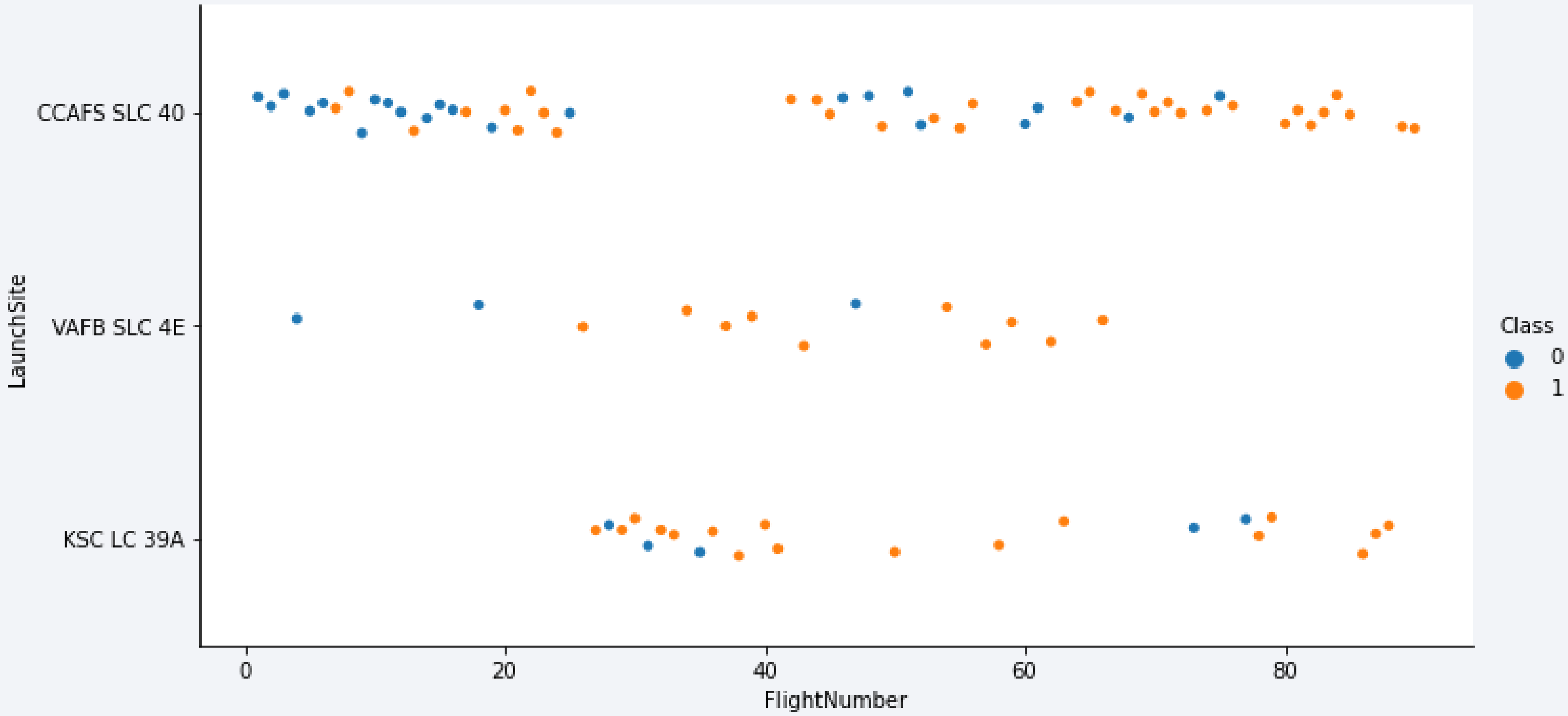
Figure 3

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

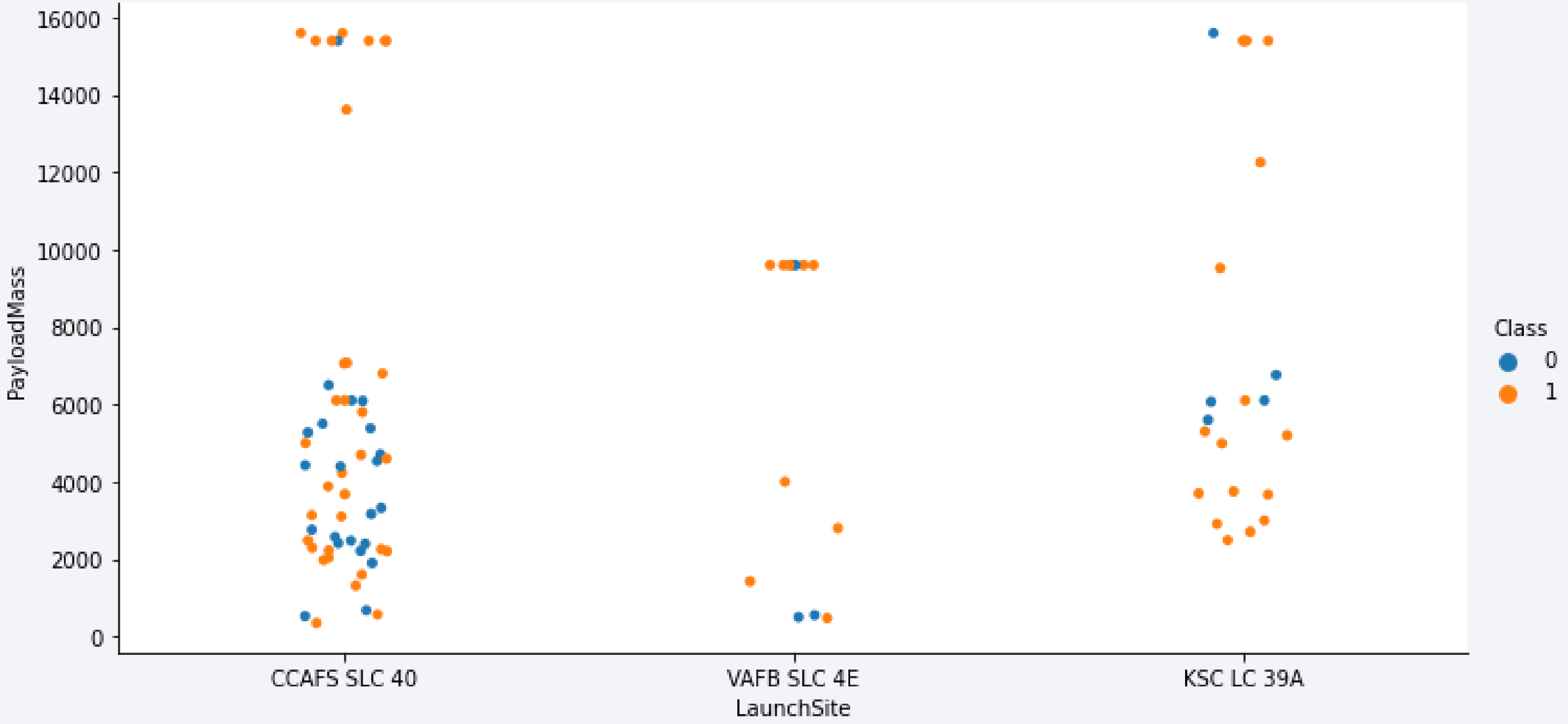
- All launch sites are mostly successful land in higher flight number.



EDA with visualization note book | GitHub

Payload vs. Launch Site

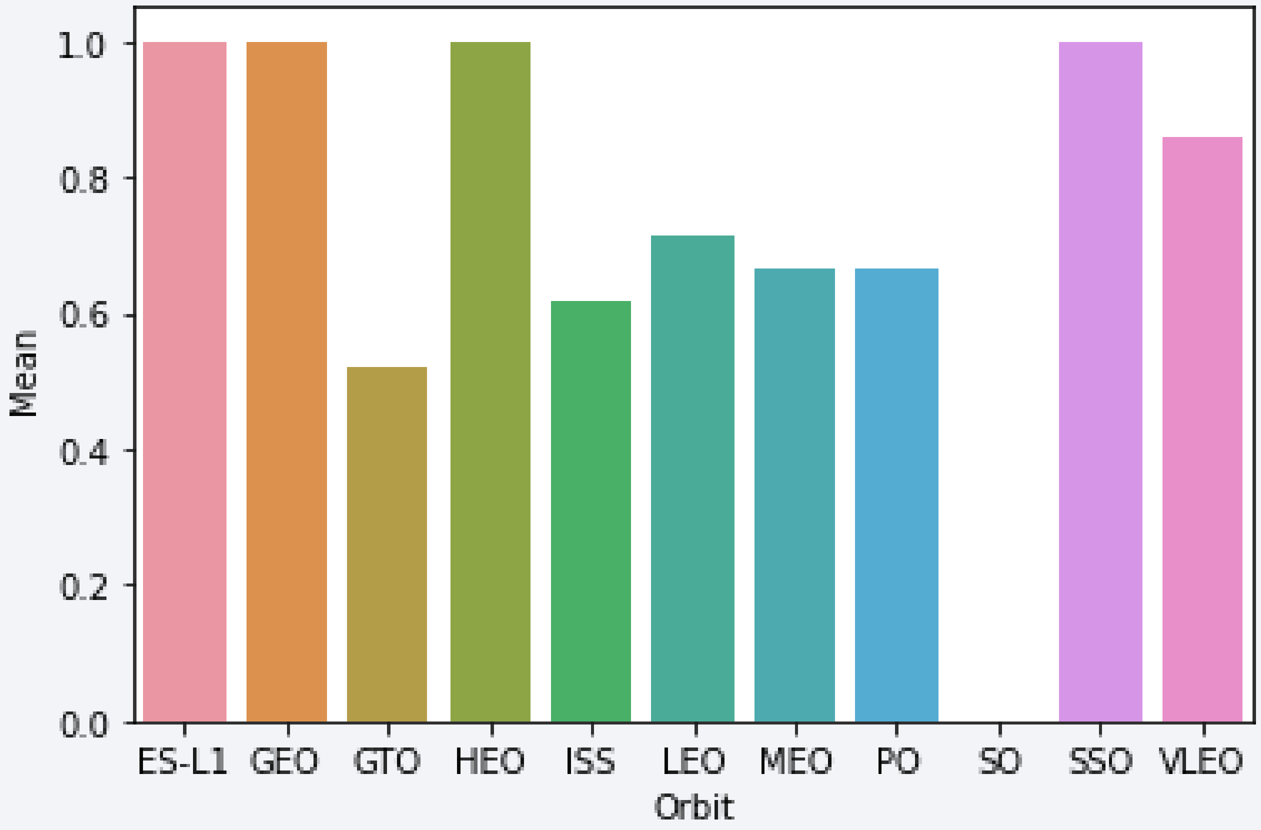
- No rocket launch is greater than 10000 kg in VAFB SLC 4E.
- More success rate with heavy payload in CCAFS SLC-40, while with light payload in KSC LC-39A.



EDA with visualization note book | GitHub

Success Rate vs. Orbit Type

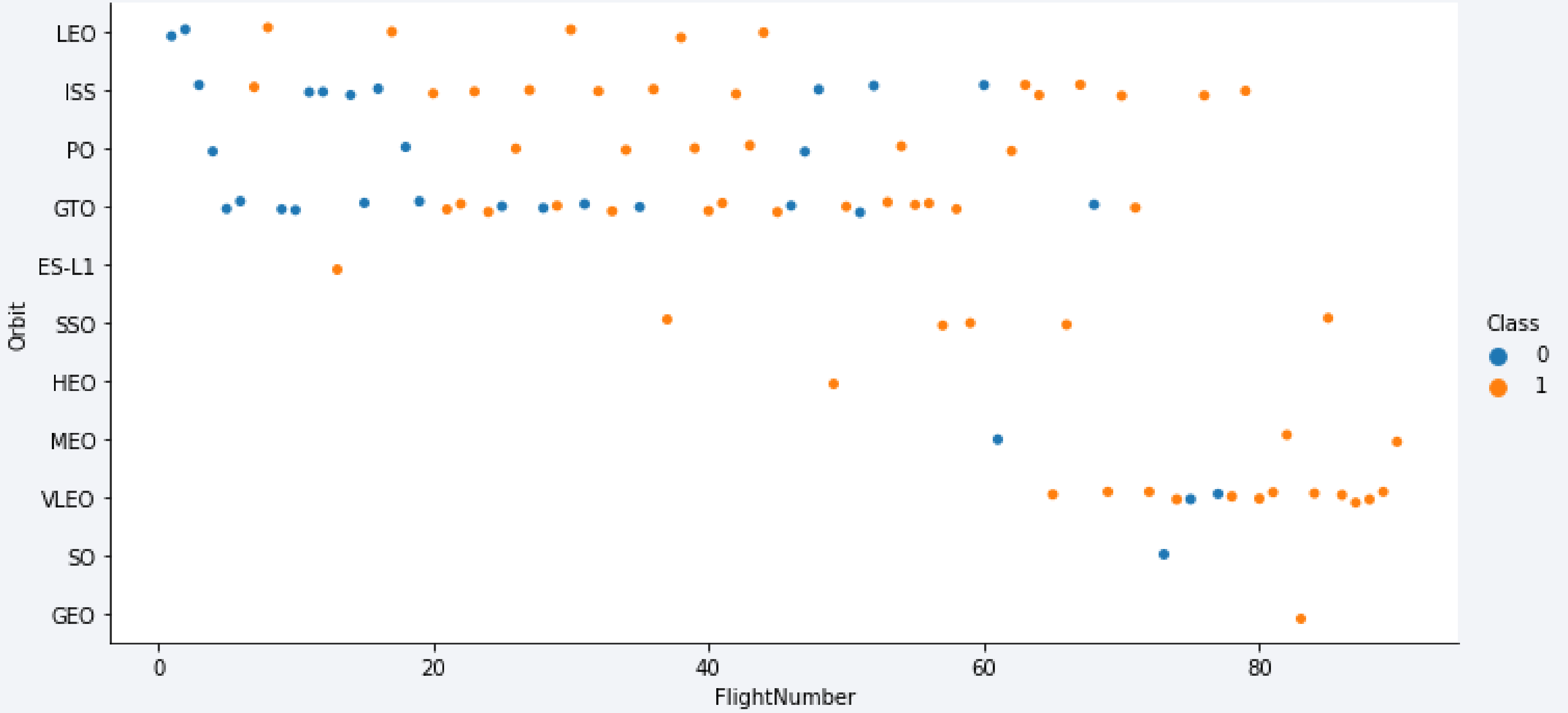
- ES-L1, GEO, HEO and SSO have the highest success rate.



EDA with visualization note book | GitHub

Flight Number vs. Orbit Type

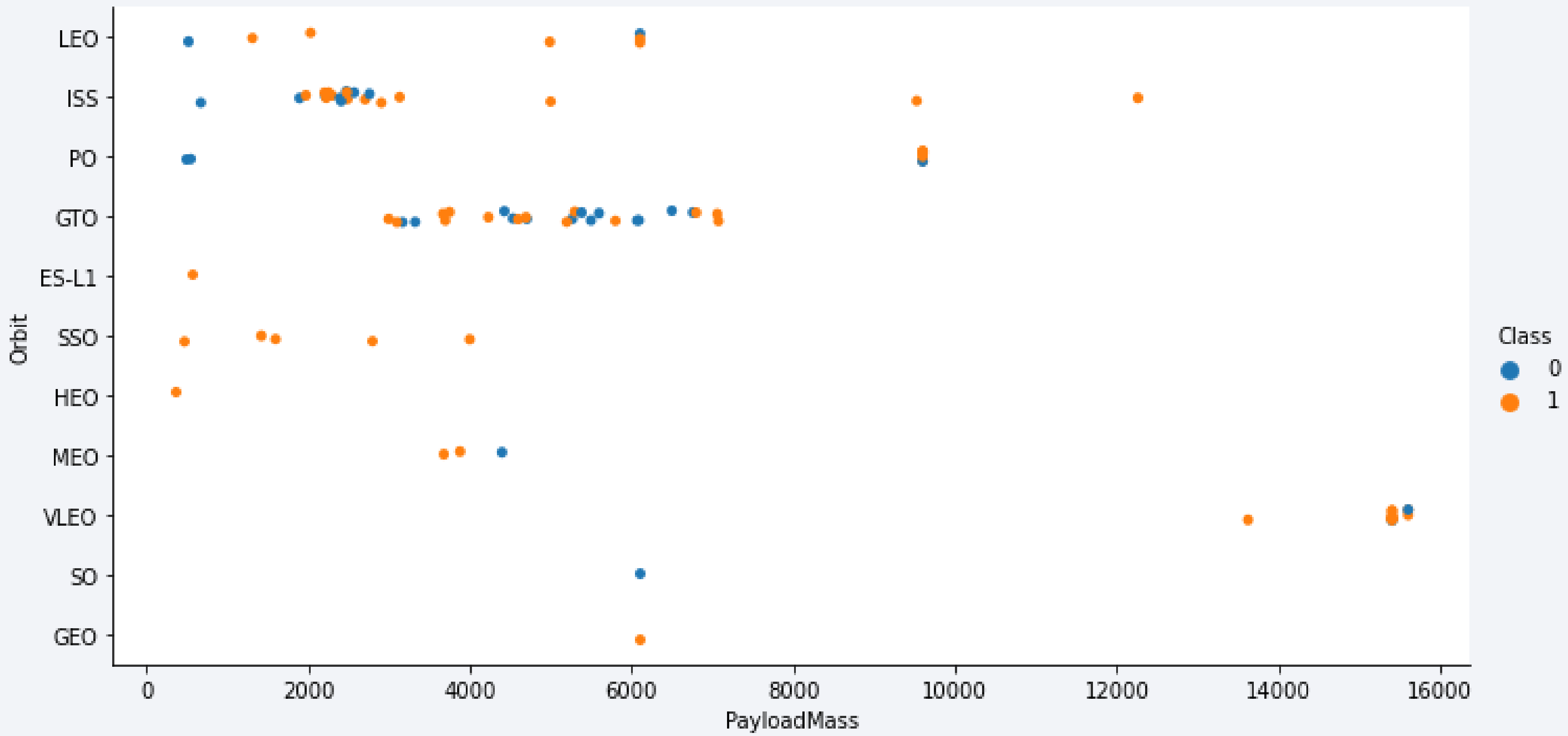
- ES-L1, GEO, HEO and SSO have the highest success rate.
- VLEO orbit has higher success in higher flight number.



EDA with visualization note book | GitHub

Payload vs. Orbit Type

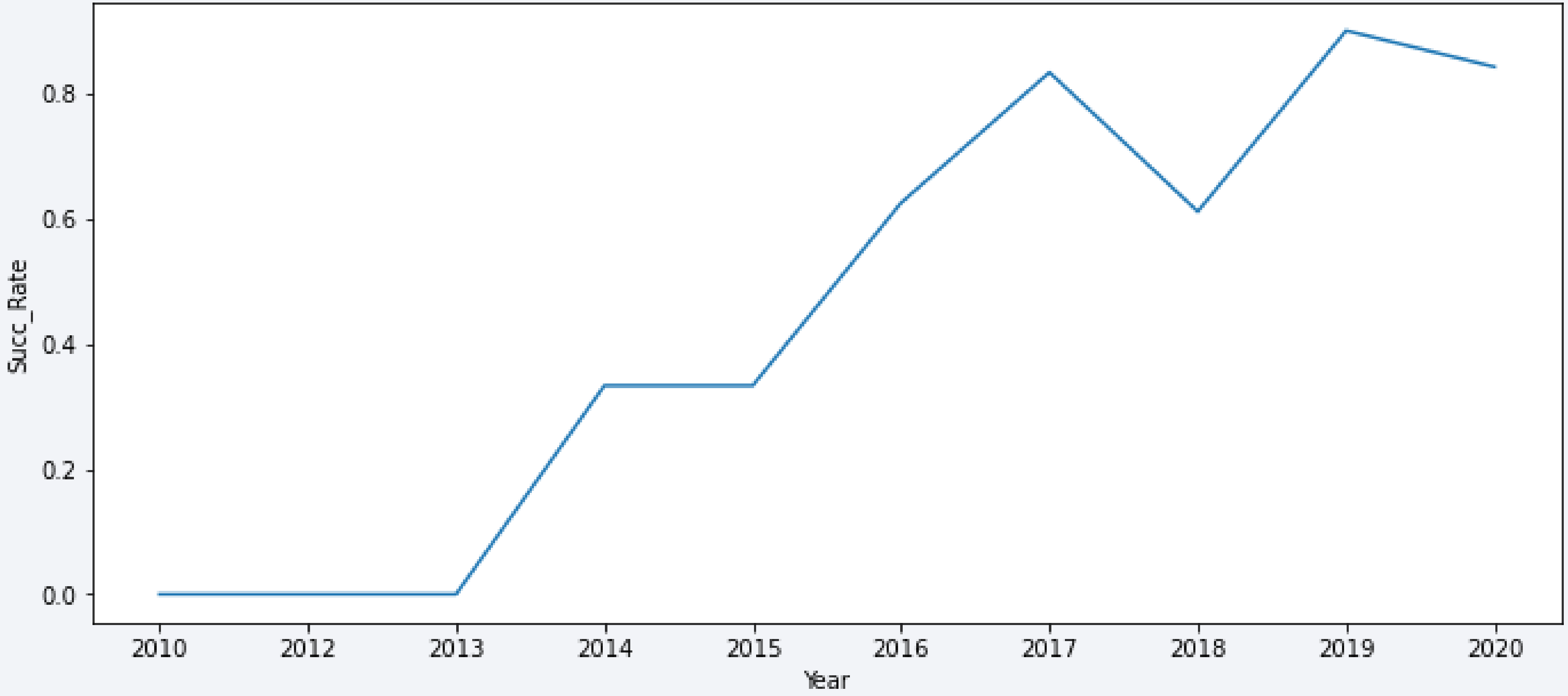
- ISS, PO and VLEO have more success landing over heavy payload.



EDA with visualization note book | GitHub

Launch Success Yearly Trend

- The launch site success rate is increased from 2014 until 2020.



EDA with visualization note book | GitHub

All Launch Site Names

- 4 unit launch sites are received from the following query.

```
%%sql
select distinct(Launch_Site)
from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Total 60 results are found, but shown only 5 results due to the limited space.

```
%%sql
SELECT * from SPACEXTABLE
where Launch_Site like 'CCA%'
limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The result of "Total Payload Mass" is 45,596 kg.

```
%%sql
select SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass"
from SPACEXTABLE
where Customer LIKE 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass by F9 v1.1 is ~2,535 kg.

```
%%sql
select AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass"
from SPACEXTABLE
where Booster_Version like 'F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
```

Average Payload Mass
2534.6666666666665

First Successful Ground Landing Date

- The first successful ground landing date is 22 December 2015.

```
%%sql
select min(Date) as first_succ_lo, Landing_Outcome
from SPACEXTABLE
where Landing_Outcome
LIKE 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

first_succ_lo	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Three customers (Sky Perfect JSAT Group and SES and SES EchoStar) are found.

```
%%sql
select Customer, PAYLOAD_MASS__KG_ from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ > 4000
and PAYLOAD_MASS__KG_ < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Customer	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	4696
SKY Perfect JSAT Group	4600
SES	5300
SES EchoStar	5200

Total Number of Successful and Failure Mission Outcomes

- The success outcome is 100 and failure outcome is 1 as per query result.
  (Remark mission outcome "Success" is separately show due to typing error.)

```
%%sql
select Mission_Outcome, count(*) as Count
from SPACEXTABLE
group by Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The booster version name that can carry the maximum payload is F9 B5 B1048.4.

```
%%sql
select Booster_Version, max(PAYLOAD_MASS__KG_)
from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	max(PAYLOAD_MASS__KG_)
F9 B5 B1048.4	15600

2015 Launch Records

- Boosters were launched in a single launch site in January and April 2015.

```
%%sql
select substr(Date,6,2) as Month,
Landing_Outcome as Failure_Landing_Outcome,
Booster_Version, Launch_Site,
substr(Date,0,5) as Year from SPACEXTABLE
where substr(Date, 0, 5) = '2015'
and Landing_Outcome like 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Failure_Landing_Outcome	Booster_Version	Launch_Site	Year
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Total 8 landing outcomes are found.

```
%%sql
select Date, Landing_Outcome, count(*) as Outcome_Count
from SPACEXTABLE
where (Date between '2010-06-04' and '2017-03-20')
and (Landing_Outcome = 'Failure (drone ship)'
     or Landing_Outcome = 'Success (ground pad)')
group by Landing_Outcome
order by Outcome_Count desc;
```

```
 * sqlite:///my_data1.db
Done.
```

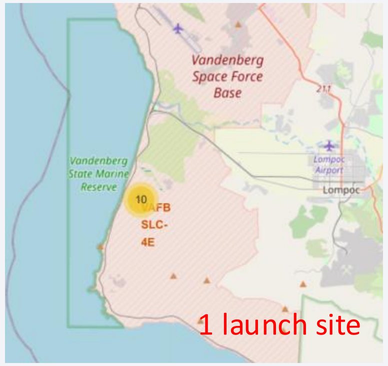
Date	Landing_Outcome	Outcome_Count
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3

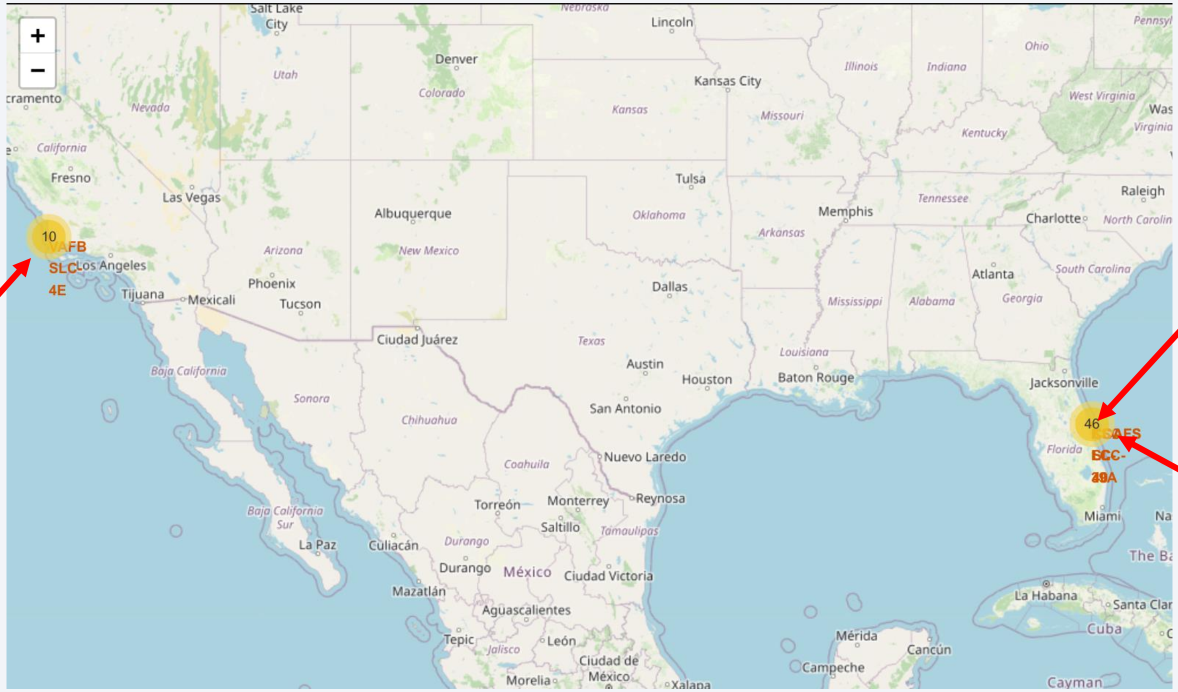
Section 3

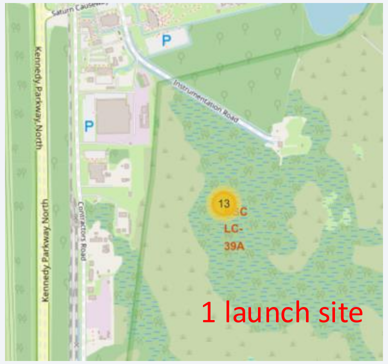
Launch Sites Proximities Analysis

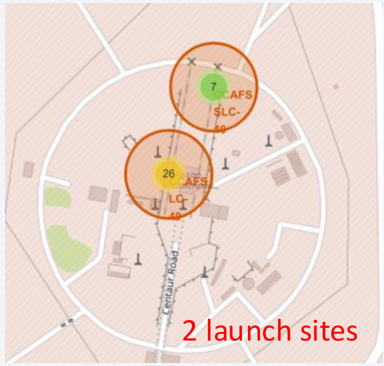
Launch Sites Locations

- There are total 4 launch site locations on the map.
- All of the launch sites are located near coastline.





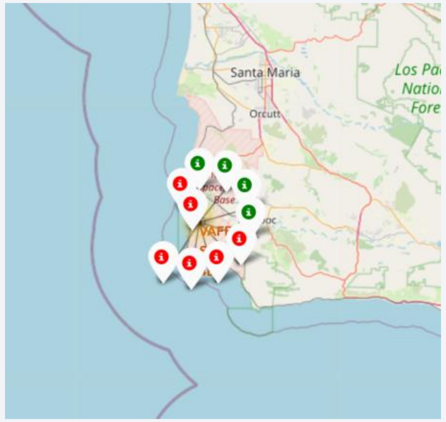




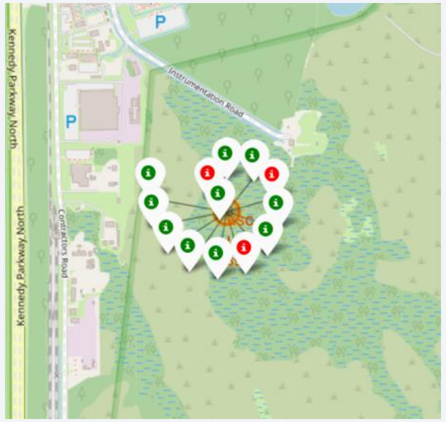
Interactive map notebook | [GitHub]

Landing Outcomes

- Landing outcomes can be seen on the map in order to markers' colors
- RFC LC-39A has more successful outcomes.
- On the other hand, CCAFS LC-40 is used for landing most of the time.

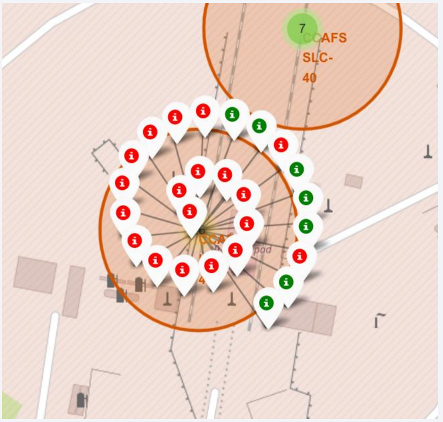


VAFB SLC-4E



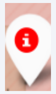
RFC LC-39A

CCAFS SLC-40



CCAFS LC-40

Successful landing

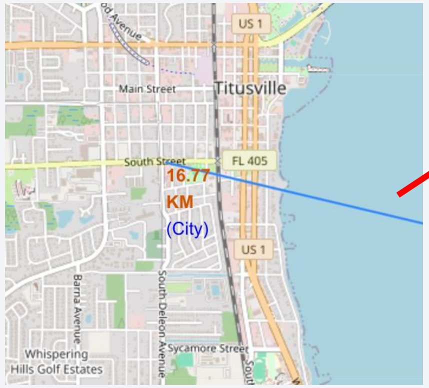


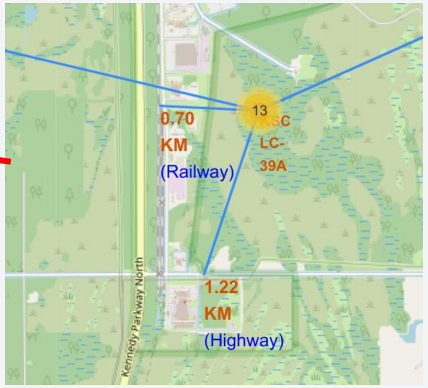
Faile landing

Interactive map notebook | GitHub

Proximities Check on RFC LC-39A

- The launch site area is located 16.77 KM from the nearest city (safe enough).
- It is 6.7 KM from the coast line, 0.7 KM from railway and 1.22 KM from high way.
- It is safe enough for people for landing failure and good enough for supply transportation





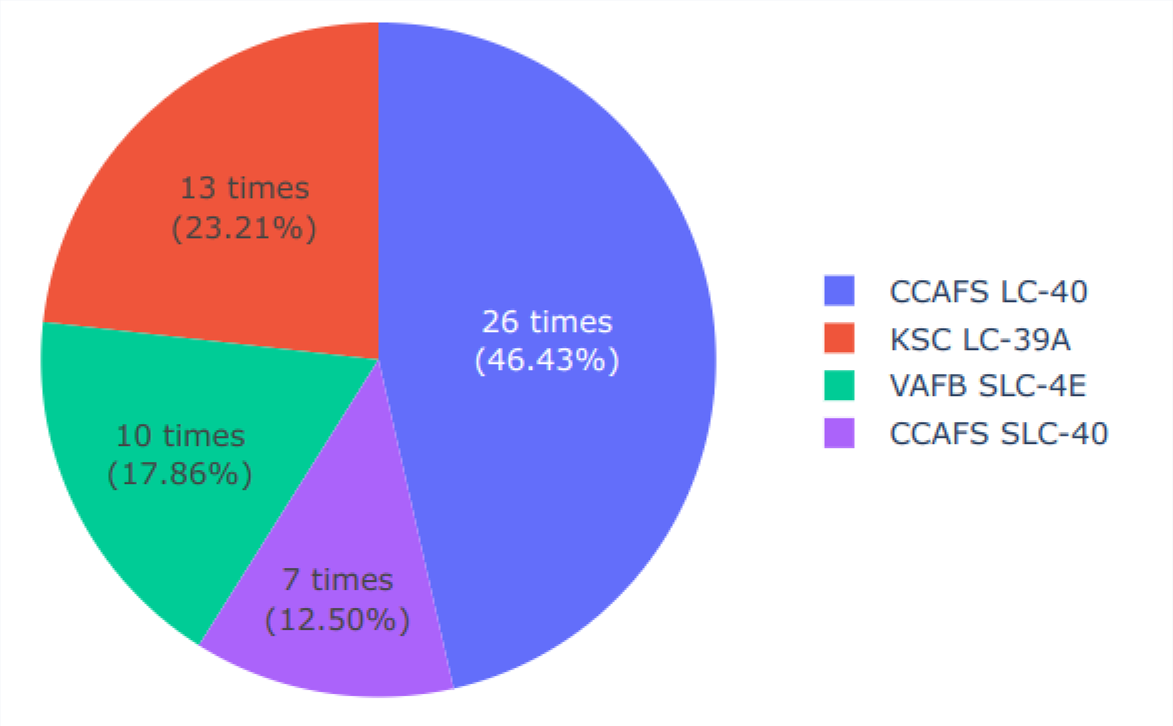
Interactive map notebook | GitHub

Section 4

Build a Dashboard with Plotly Dash

Launch Success Counts for All Sites

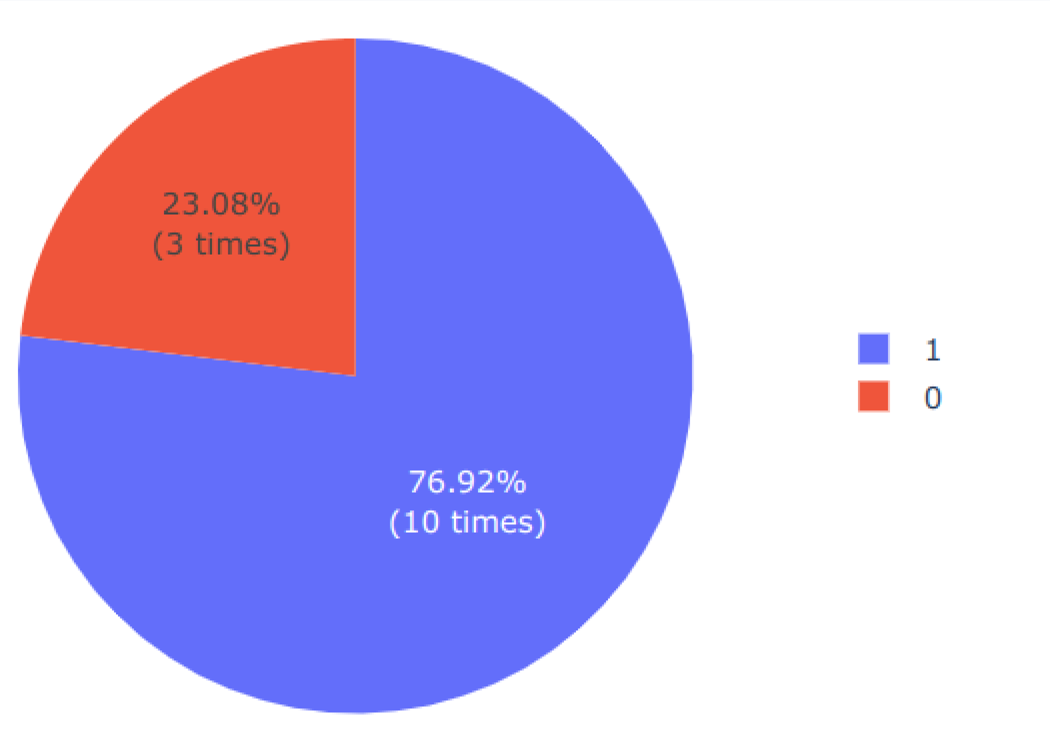
- CCAFS LC-40 has the highest launch success count with 46.43%.
- CCADS SLC-40 has the lowest launch success count with 12.5%.



Interactive map notebook | GitHub

Highest Launch Success Ratio

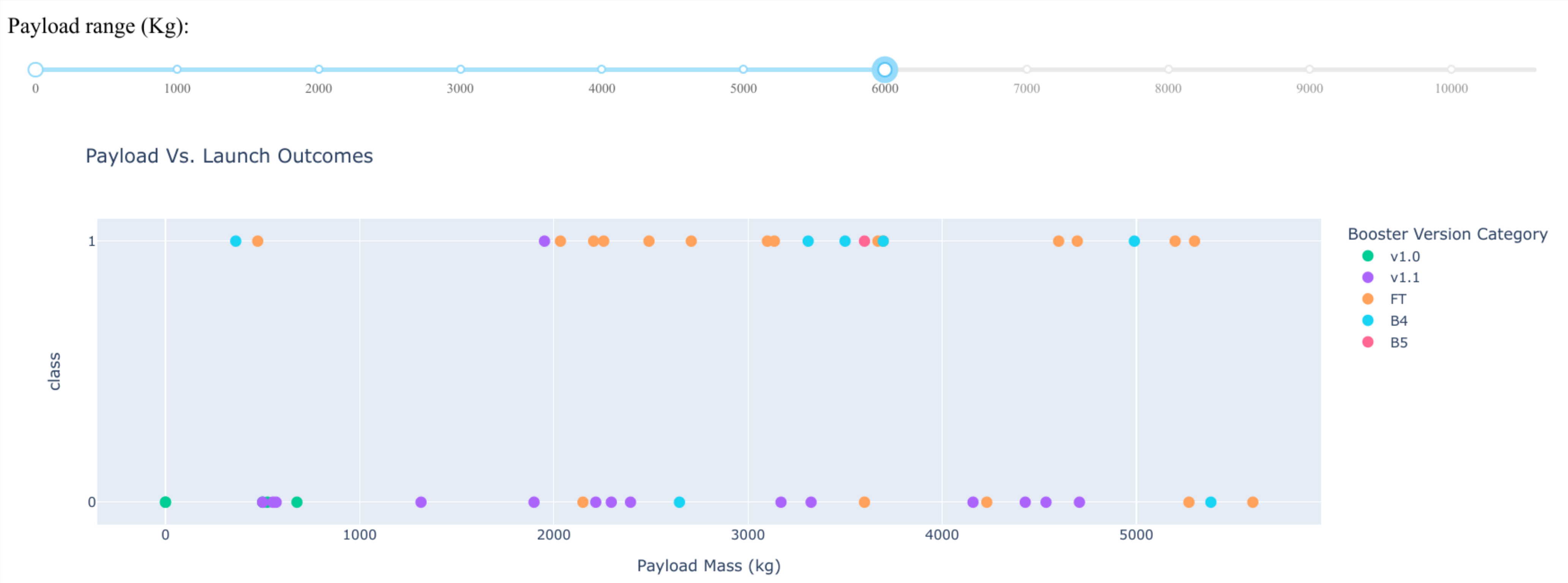
- TSKC LC-39A has the highest success ratio by 76.92%.
- The total launch count is 13 times, which can lead to higher percent values.



Interactive map notebook | GitHub

Payload Mass Vs. Success Counts

- The available max payload is 9600 kg.
- The payload range 0 – 6000 kg and FT booster version have the largest success counts compared to other payload range.



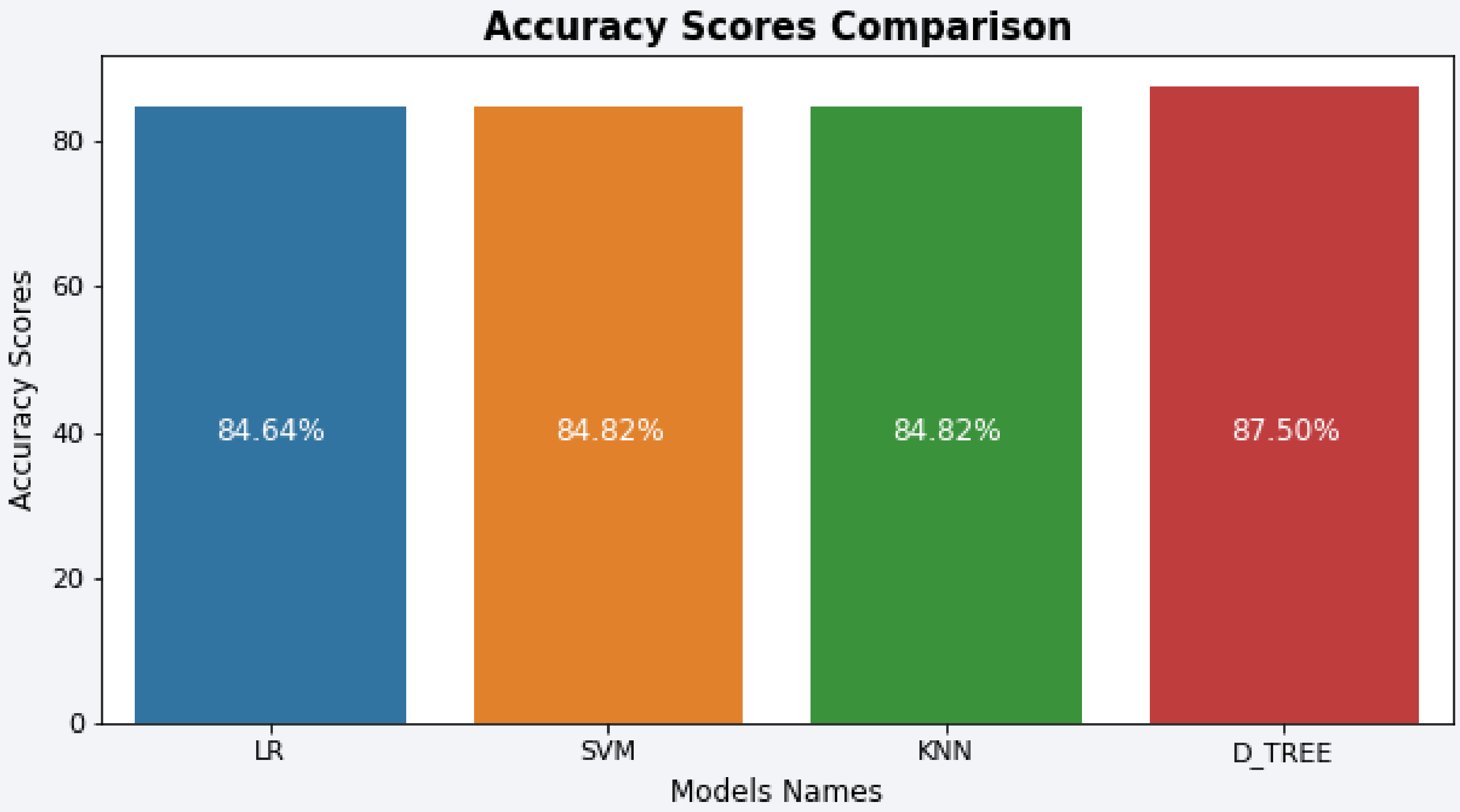
Interactive map notebook | GitHub

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- In order to my analysis, Decision Tree model has the highest classification accuracy.



LR = Logistic Regression

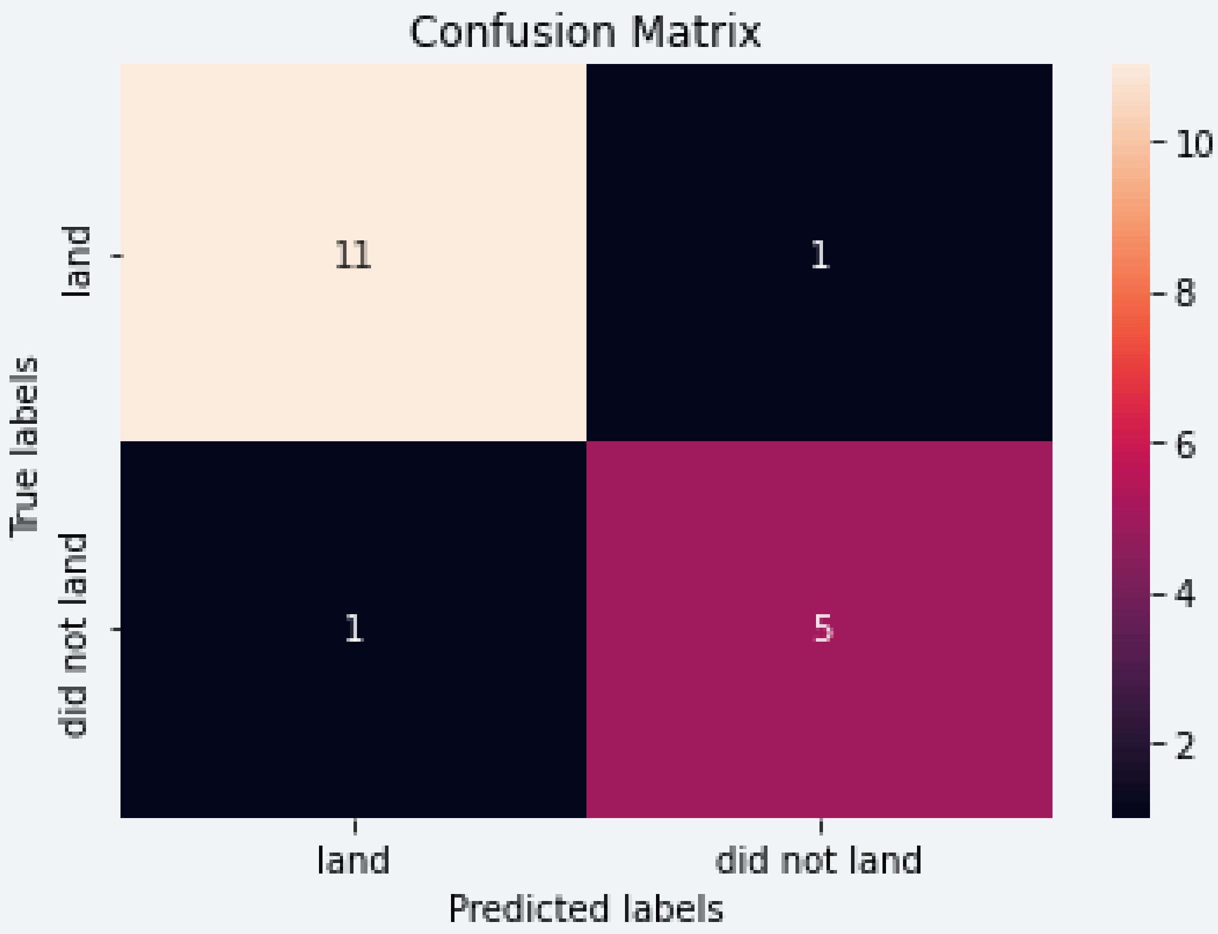
SVM = Support Vector Machine

KNN = K-Nearest Neighbor

D_TREE = Decision Tree

Confusion Matrix

- The best performing model is Decision Tree.



Explanation

- This model can predict values closer to the actual value.
- Correctly predicted 5 "**did not land**" values compared with other models
- Accurately predicted 11 "**land**" values compared with other models.
- In contrast, it can give the least fault prediction results (***False Negative*** and ***False Positive***).

Conclusions

The Space Y project aims to develop a predictive model to compete with SpaceX, focusing on accurately predicting rocket landings and saving over $100 million.

Objective: Build a model to predict successful Stage 1 rocket landings back to earth.

Data Collection: Sourced from SpaceX API and Wikipedia using web scraping.

Data Preparation: Cleaned, labeled, and organized into a structured format.

Analysis: Conducted exploratory analysis using SQL and visualizations to uncover insights.

Dashboard: Created an interactive tool for real-time data updates.

Modeling: Developed and tested machine learning models, selecting the most accurate one.

Future Work: Model accuracy will improve with more data.

Appendix

GitHub Repository URL | LINK

Instructors

- Yan Luo (Ph.D., Data Scientist and Developer at IBM)
- Joseph Santarcangelo (Ph.D., Data Scientist at IBM)

Special thanks to all instructors…

Thank you!