Winning Space Race with Data Science

Khang Nguyen Do
6th Nov 2022

Outline

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

Executive Summary

- Methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Results
  - Exploratory Data Analysis results
  - Interactive Analytics in screenshots
  - Predictive Analytics results

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.
- Therefore, the listed presentation below will demonstrate a machine learning pipeline to predict if the first stage will land given the data from the preceding labs to give information regarding cost advantage with space X.

Section 1

Methodology

Methodology

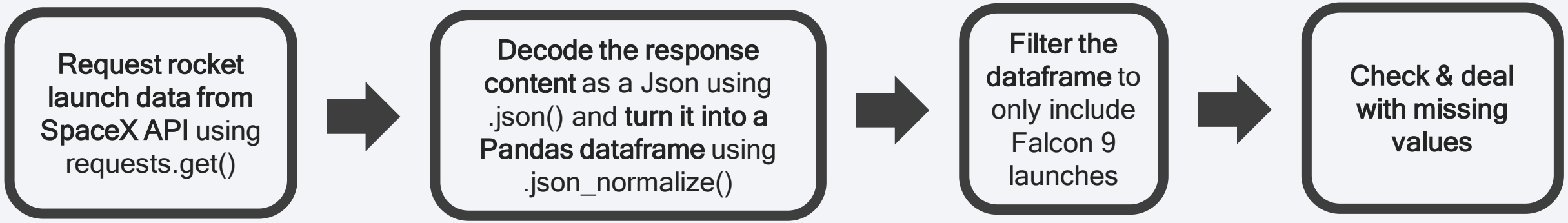
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

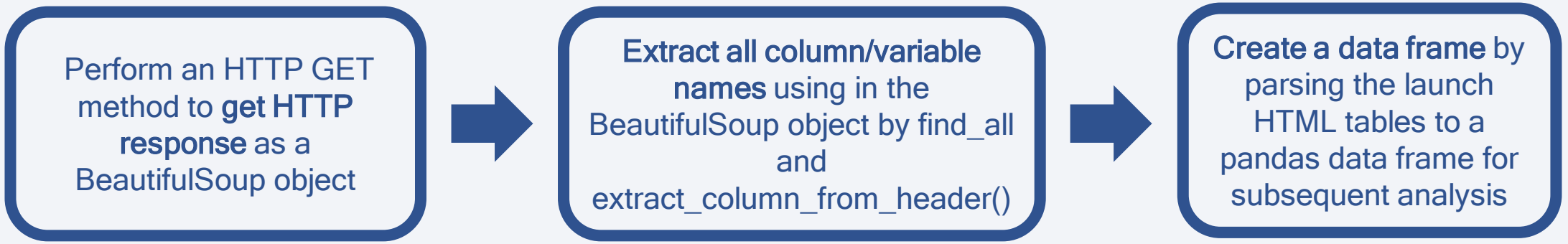
Data Collection

- The data collection process can be presented with 2 methods as below:

A. SpaceX API - Data wrangling

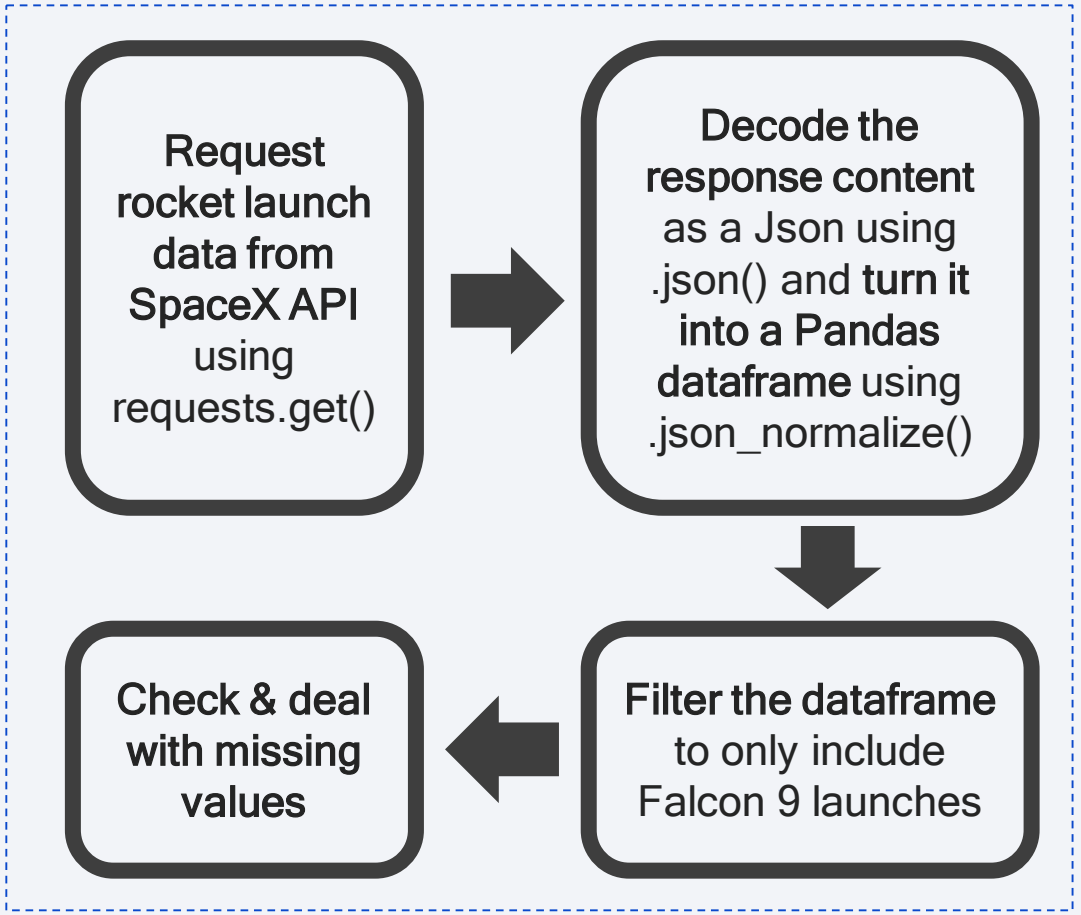


B. Web Scrapping



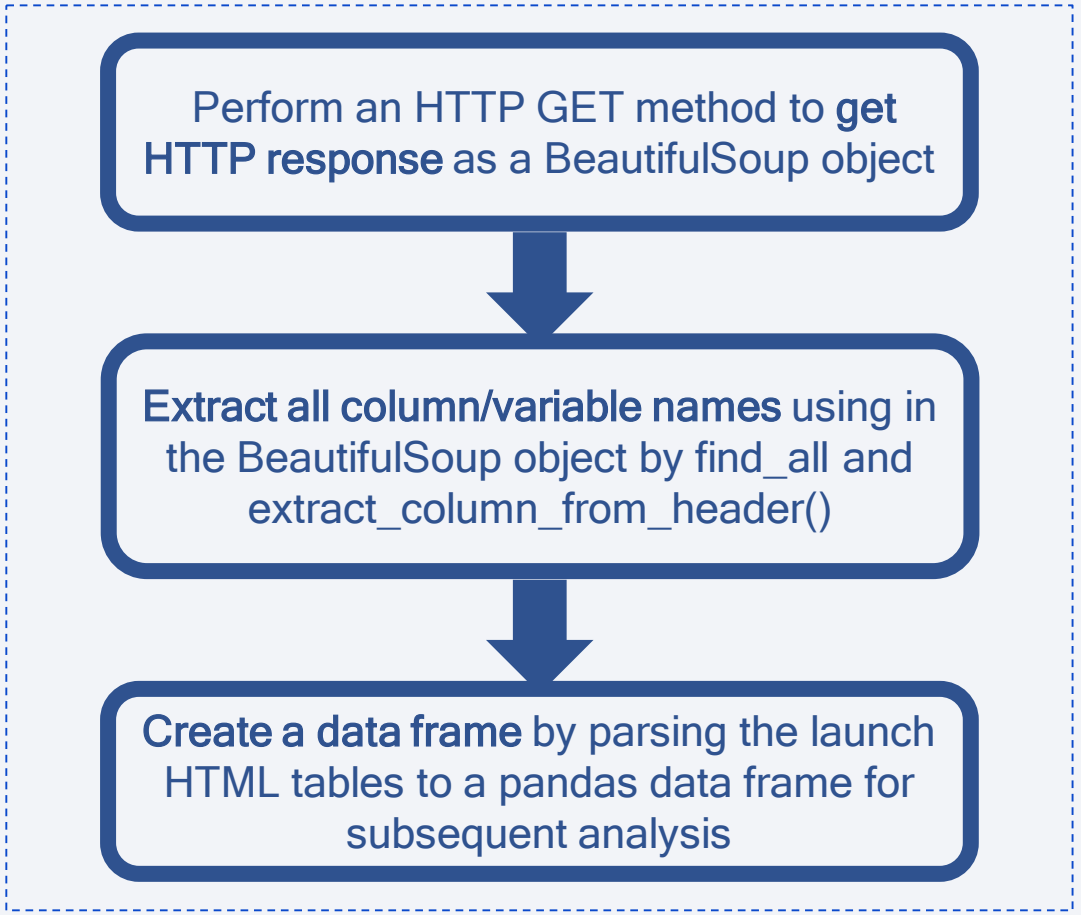
Data Collection - SpaceX API

- The data was collected from SpaceX API, then was cleaned and finally formatted as requested.
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- Link: https://github.com/khangxinh/coursera/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



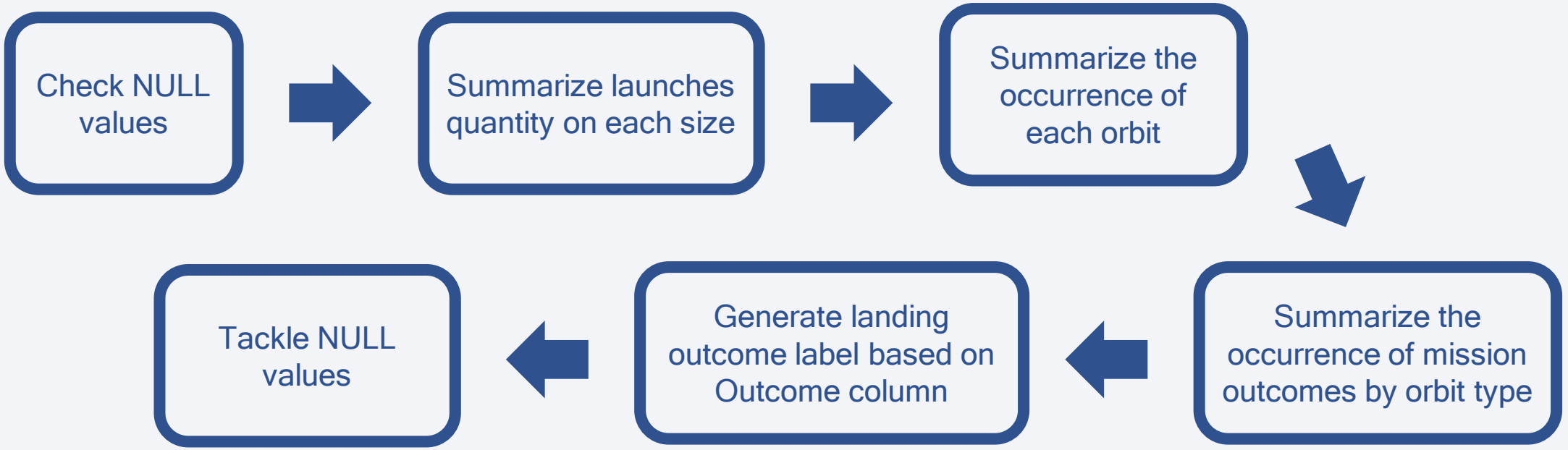
Data Collection - SpaceX API

- Data was requested from Falcon9 Launch Wiki page's URL, then all column/variable names was extracted and finally the required data frame was created by parsing the launch HTML tables
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose
- Link: https://github.com/khangxinh/coursera/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

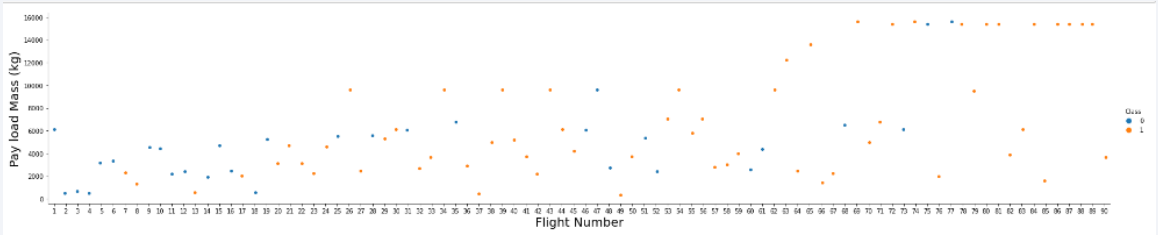
- Data was processed as the flowcharts below:



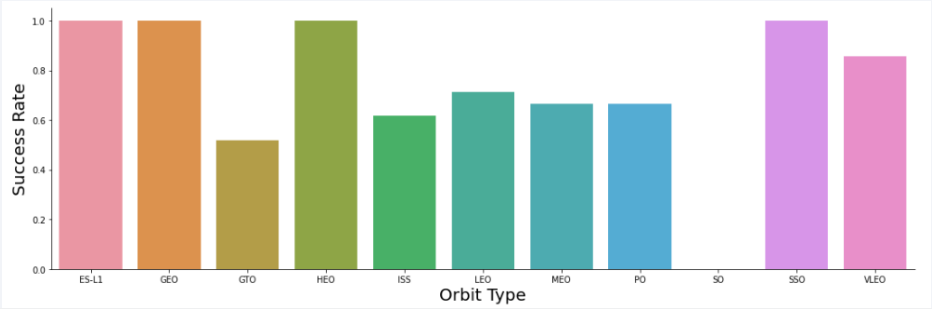
- Link: https://github.com/khangxinh/coursera/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

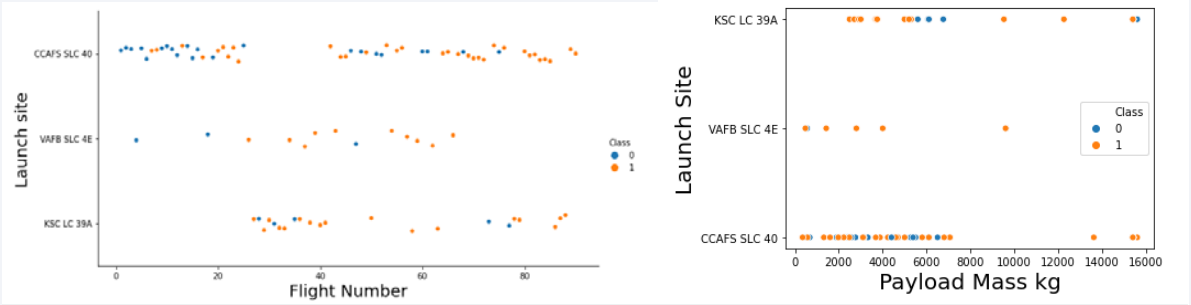
- Link: https://github.com/khangxinh/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb



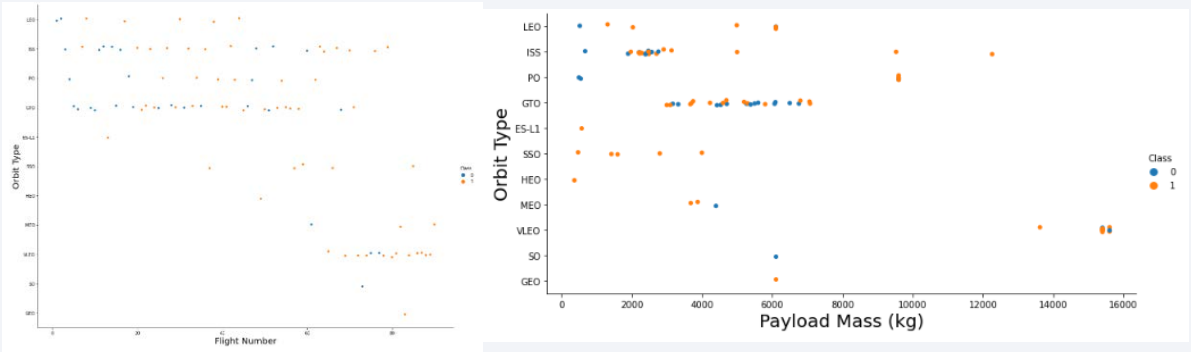
Examine FlightNumber vs. PayloadMass correlation

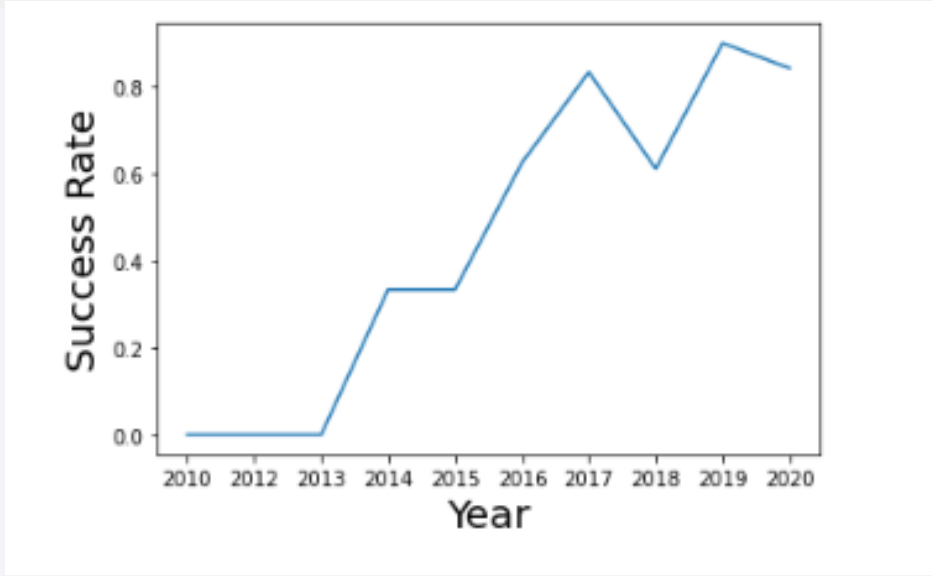


Find which orbits have high sucess rate



Examine 2-variable relationship



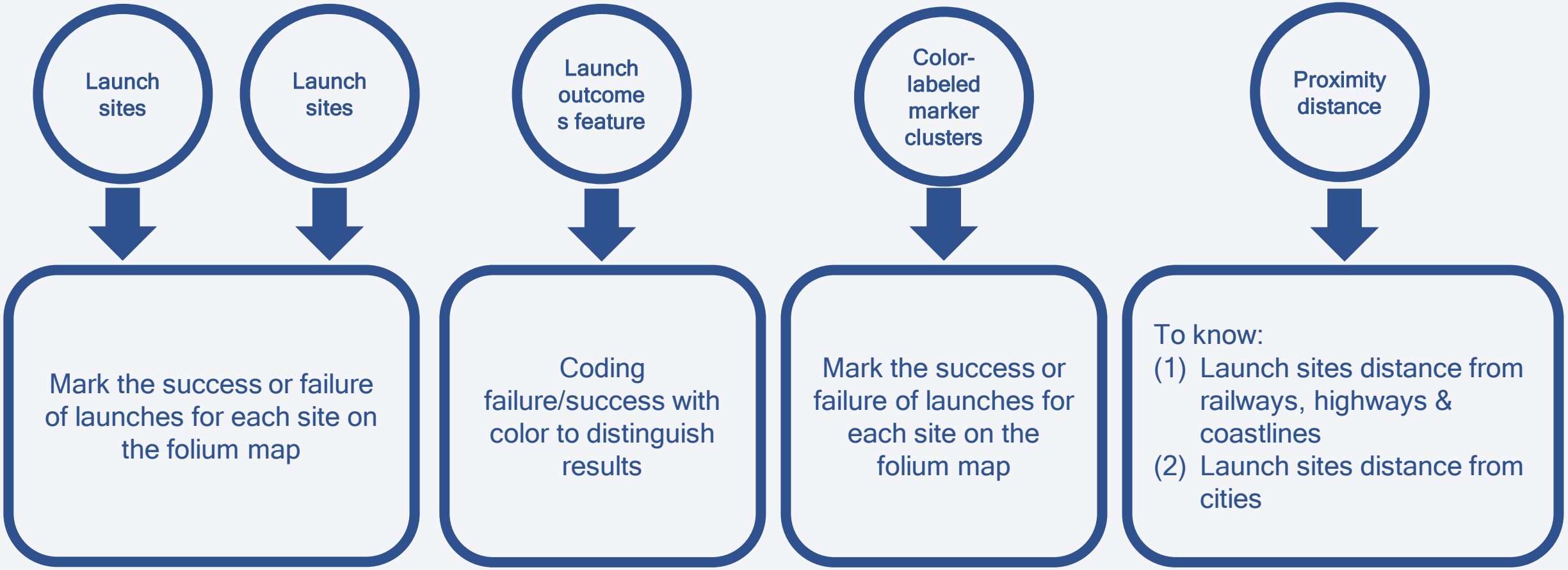


Success rate by time series

EDA with SQL

- SQL script
  - ***Display the names of the unique launch sites in the space mission***
  - ***Display 5 records where launch sites begin with the string 'CCA'***
  - ***Display the total payload mass carried by boosters launched by NASA (CRS)***
  - ***Display average payload mass carried by booster version F9 v1.1***
  - ***List the date when the first succesful landing outcome in ground pad was acheived.***
  - ***List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000***
  - ***List the total number of successful and failure mission outcomes***
  - ***List the names of the booster_versions which have carried the maximum payload mass. Use a subquery***
  - ***List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.***
  - ***Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.***
- Link: https://github.com/khangxinh/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20-%20del.ipynb

Build an Interactive Map with Folium



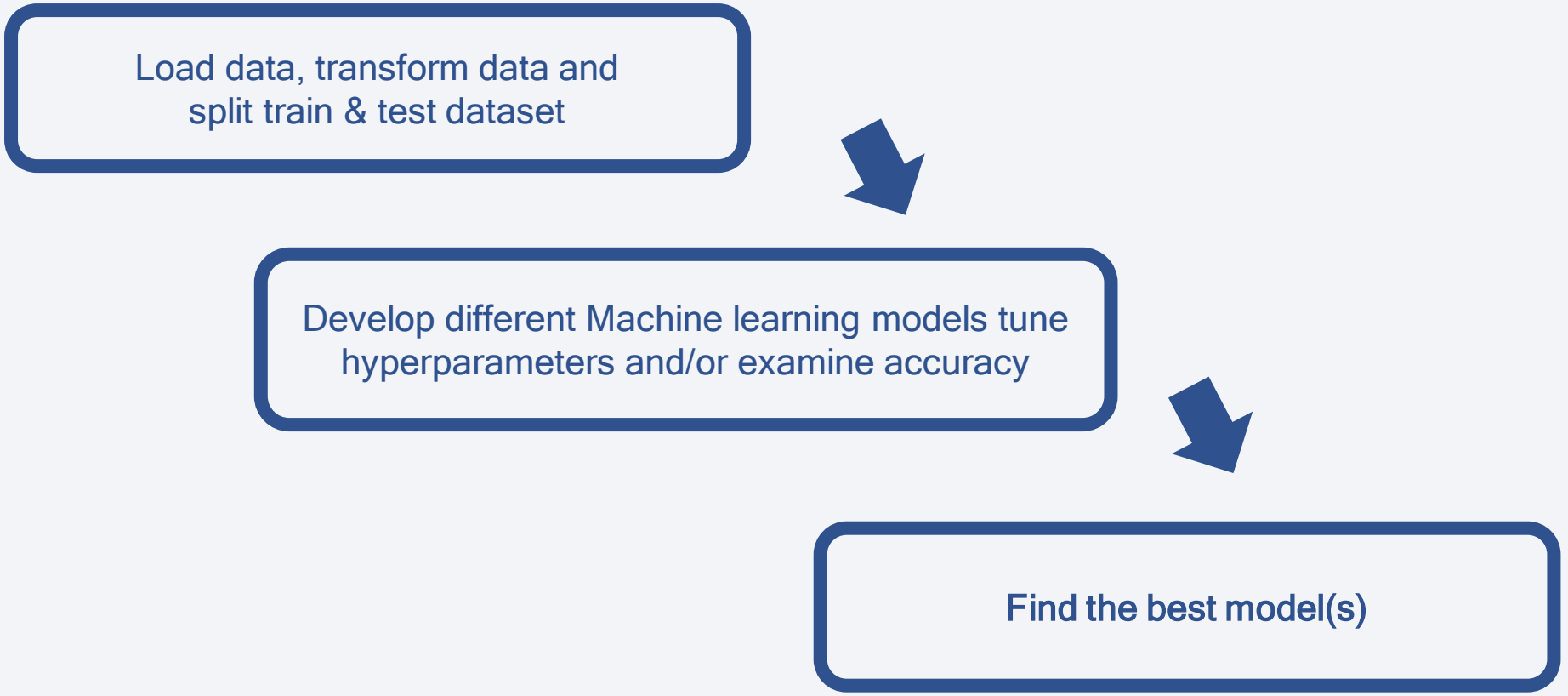
- Link: https://github.com/khangxinh/coursera/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Interactive dashboard with Plotly dash includes

- Link: https://github.com/khangxinh/coursera/blob/main/jupyter-labs-plotly-spacex.ipynb

Predictive Analysis (Classification)



- Link: https://github.com/khangxinh/coursera/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - KSC LC 49A site has the best performance;
  - Orbit GEO, HEO, SSO, ES L1 perform well regarding success rate;
  - The lower the payloads the better the performance;
  - The success rate perform gradually better through the recorded time series.

- Interactive analytics demo in screenshots:

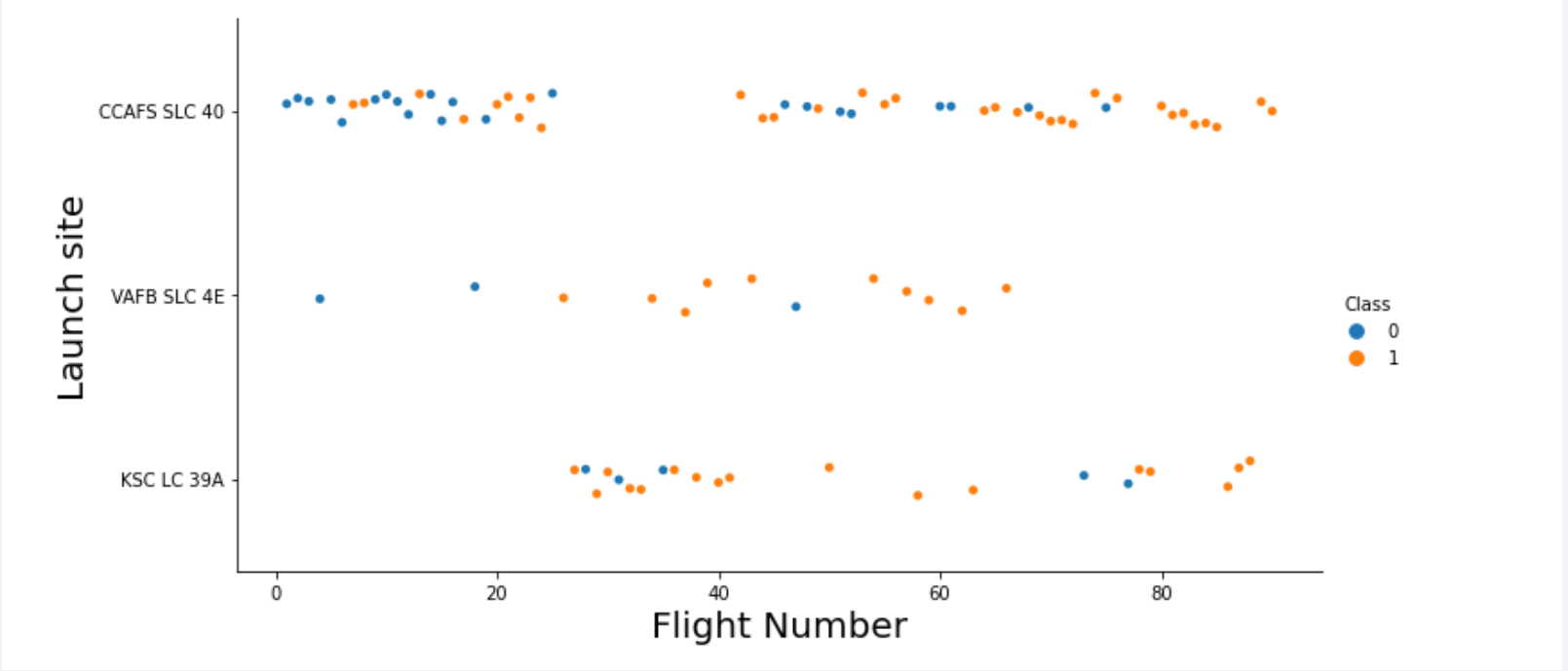
- Predictive analysis results:

  The SVM, KNN, and Log Reg are considered to be the best ML models for success rate prediction in terms of accuracy.

Section 2

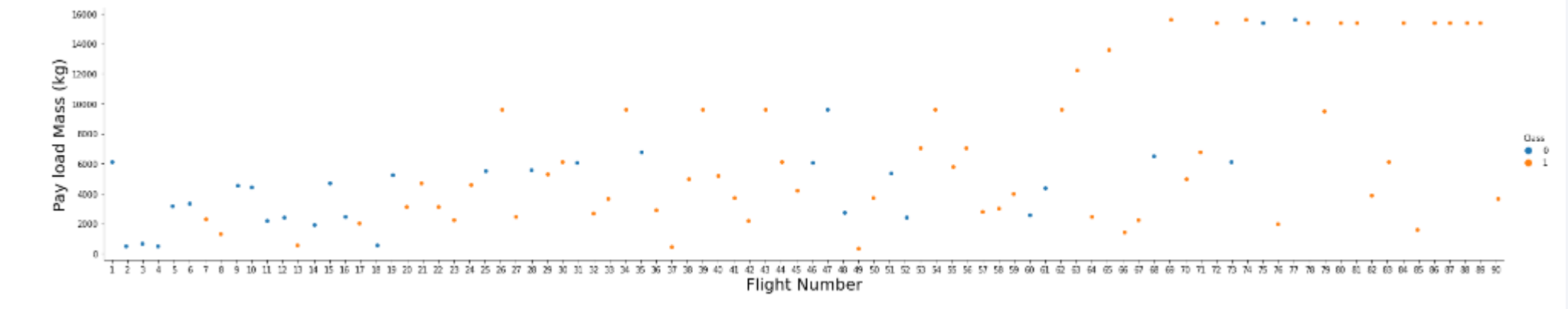
Insights drawn from EDA

Flight Number vs. Launch Site



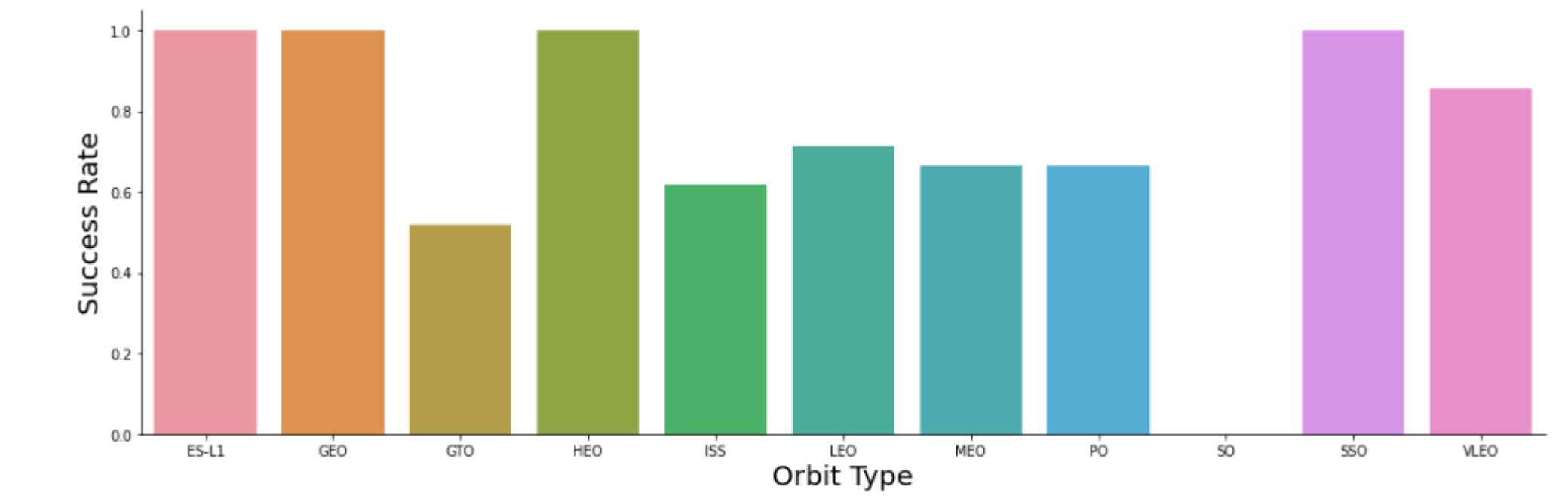
- We see that as the flight number increases, the first stage is more likely to land successfully.

Payload vs. Launch Site



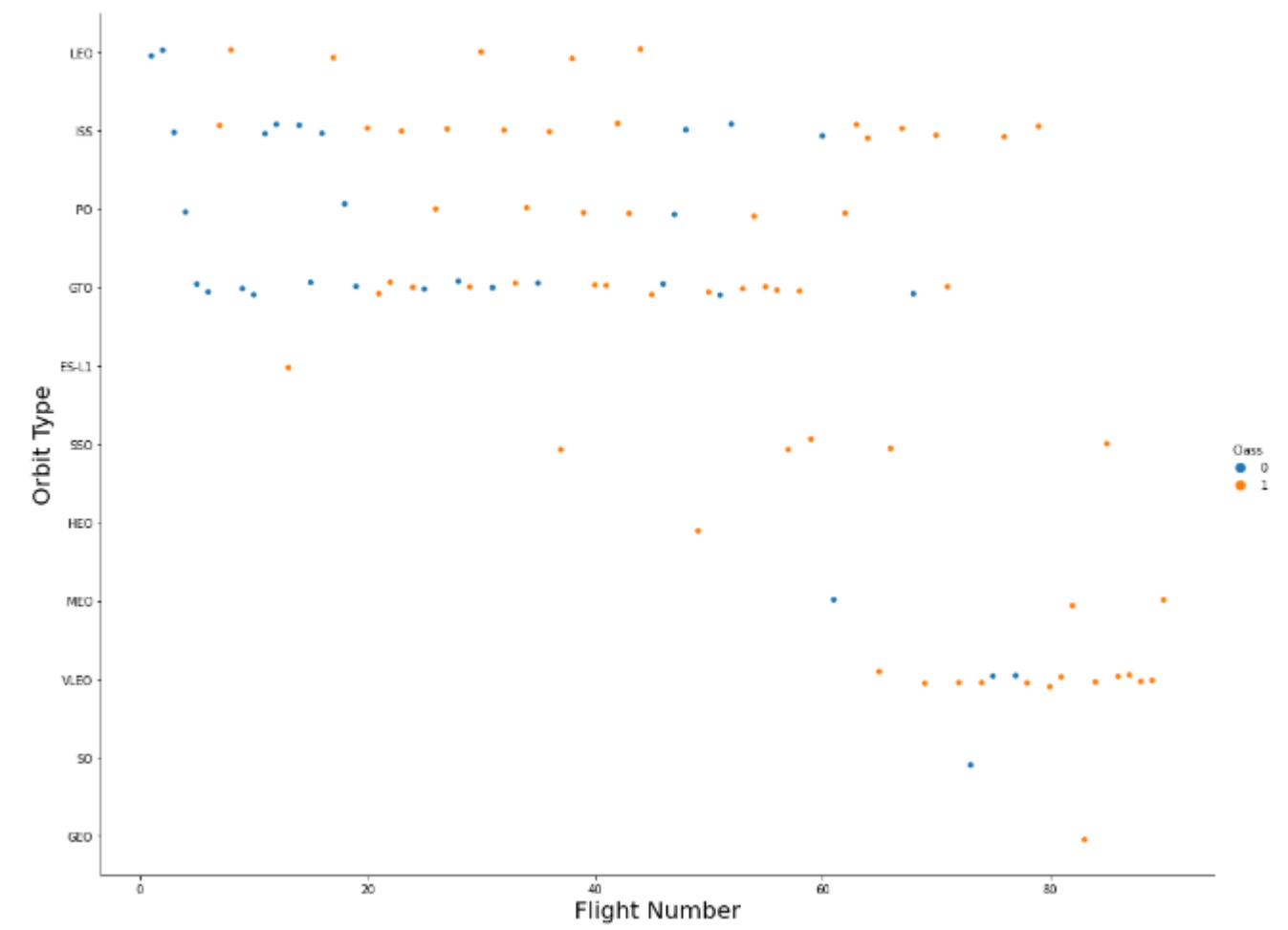
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

Success Rate vs. Orbit Type



- From the plot, we can see that launches with ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

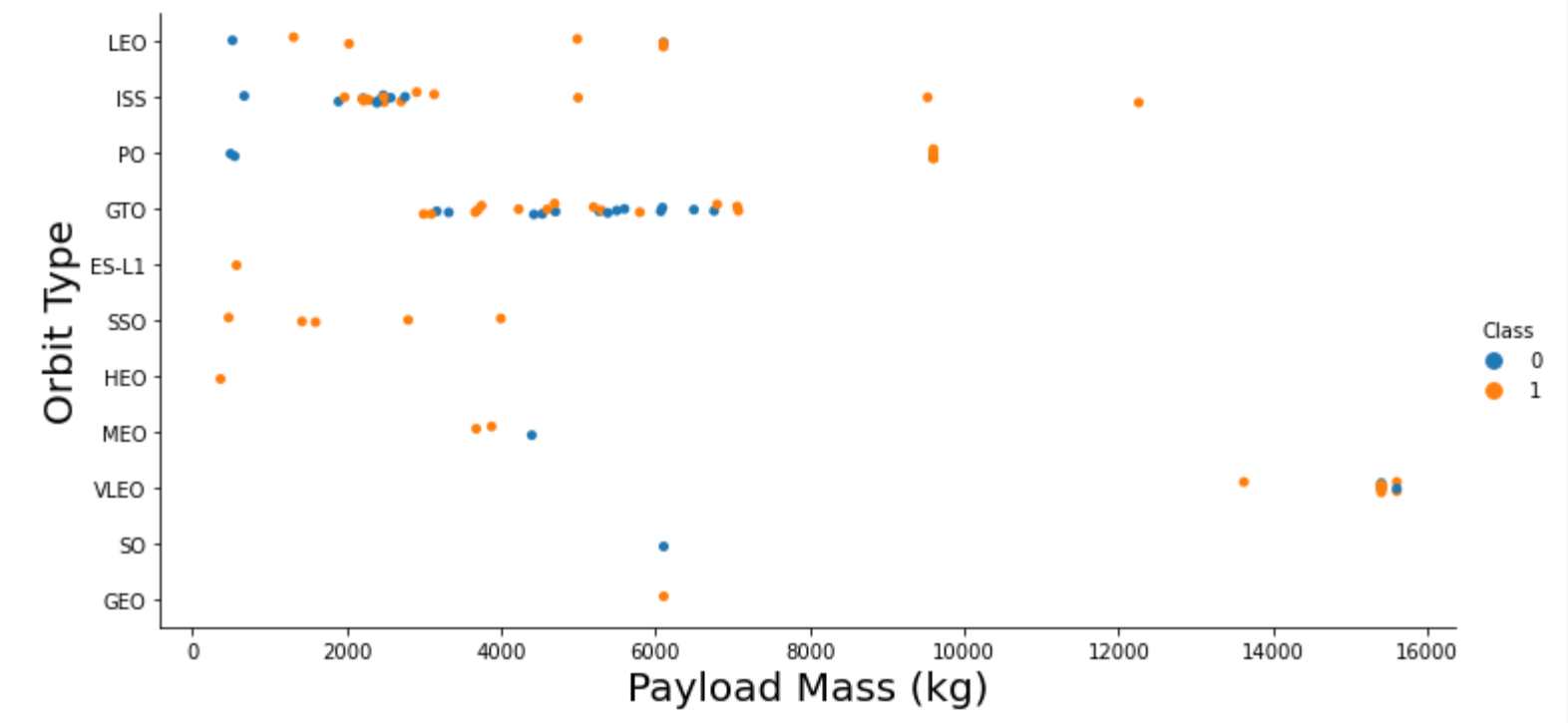
Flight Number vs. Orbit Type



- VLEO seems to be chosen more for the latter time period.

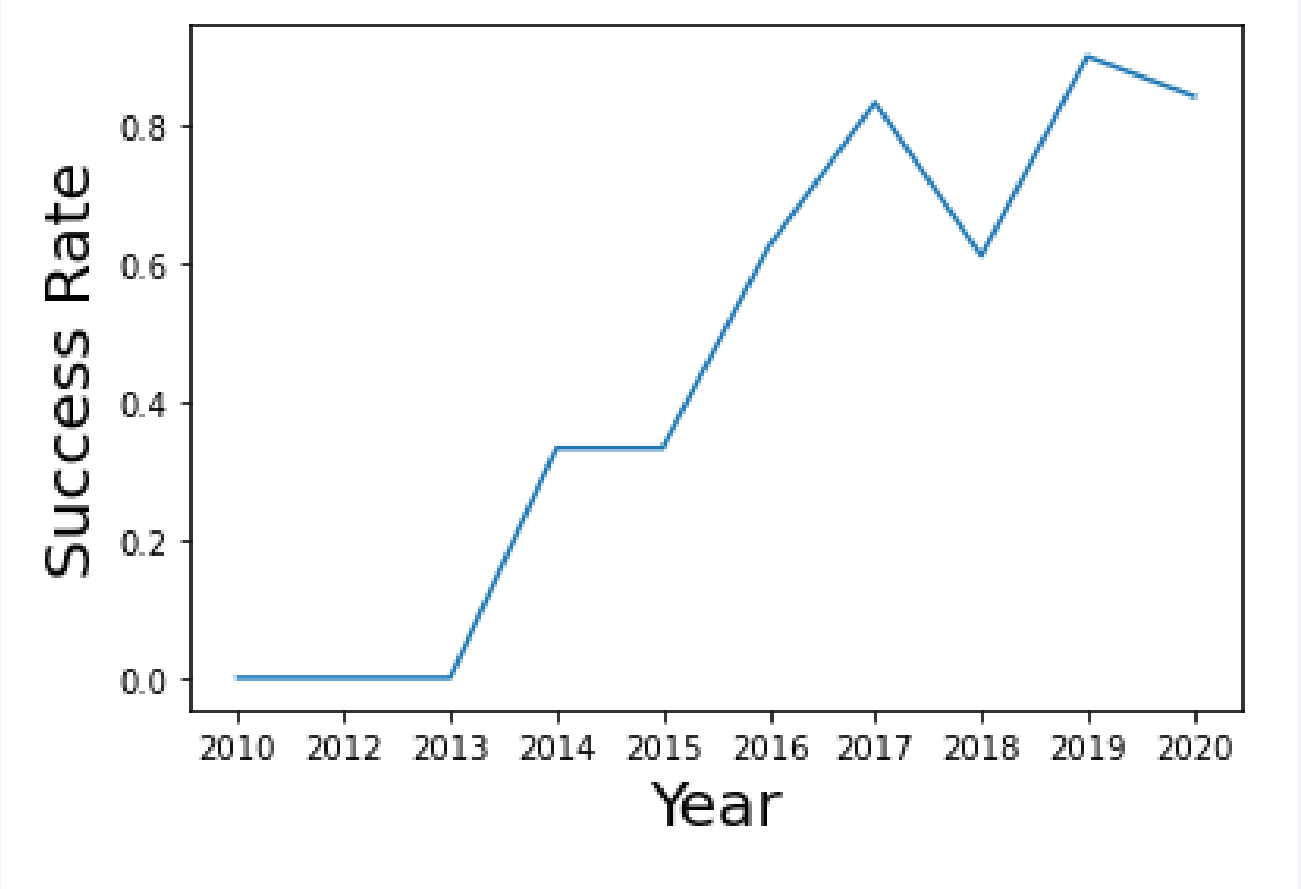
Payload vs. Orbit Type

- Strong correlation between payload mass and success rate, regardless of orbit types



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM METIBMVL;
```

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * from METIBMVL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass
FROM METIBMVL WHERE Customer = 'NASA (CRS)';
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) average_payload_mass
FROM METIBMVL WHERE Booster_Version LIKE 'F9 v1.0%';
```

average_payload_mass
340

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) first_succesful_landing_date
FROM METIBMVL WHERE Landing__Outcome = 'Success (ground pad)';
```

first_succesful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM METIBMVL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM METIBMVL GROUP BY MISSION_OUTCOME;
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM METIBMVL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)FROM METIBMVL);
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM METIBMVL WHERE Landing__Outcome = 'Failure (drone ship)'
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM METIBMVL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Launch site on US map

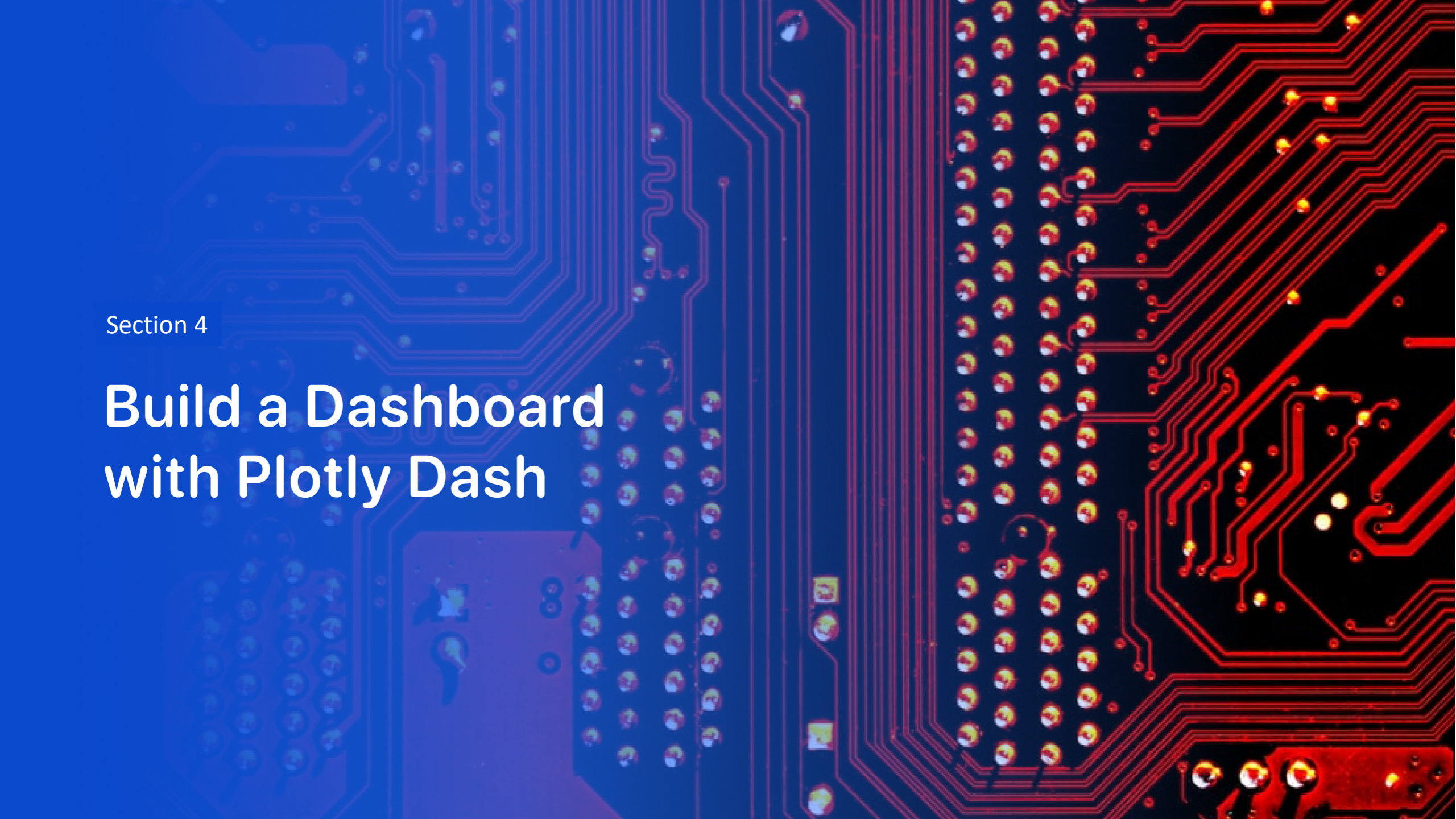
- Launch site demonstrated on the map of the USA

Launch sites and launches labels

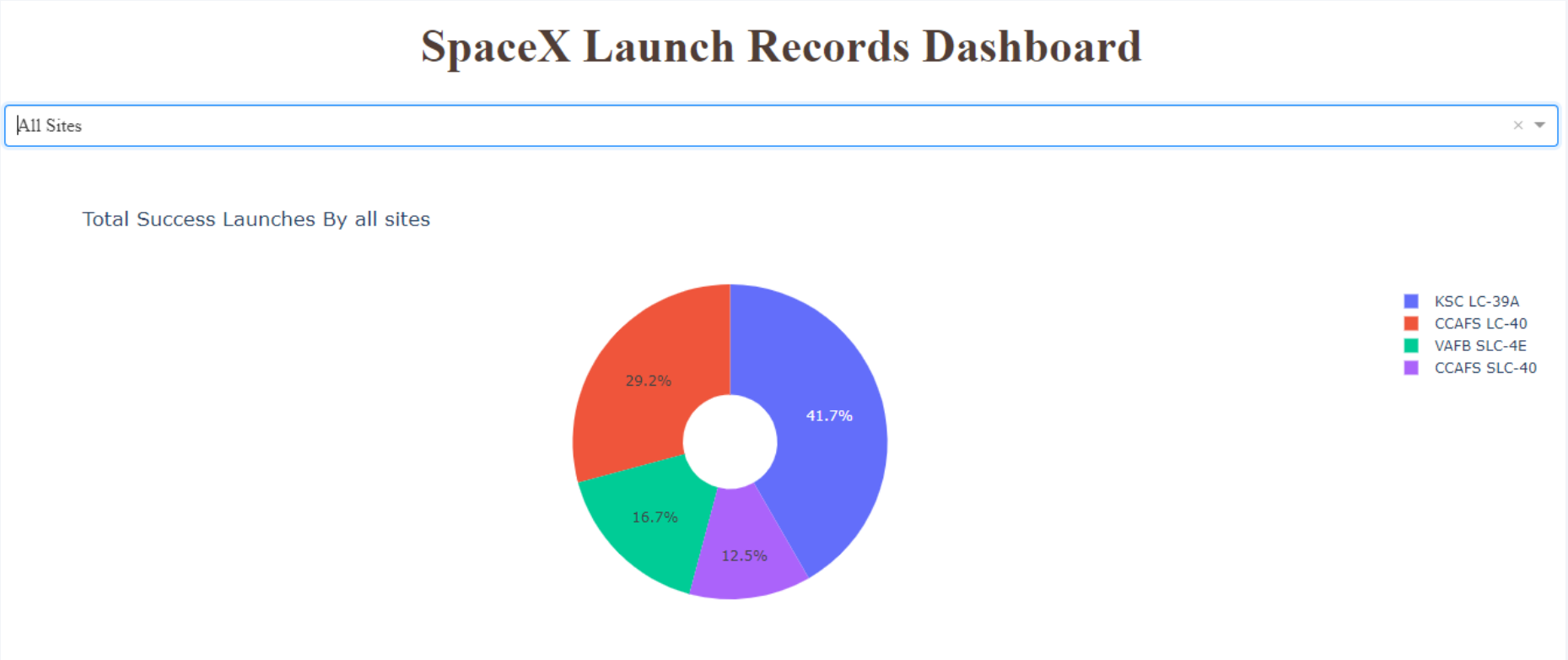
- Florida and California Launch sites, Green label shows success launch and Red label show failed launch

Launch Site distance to points

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

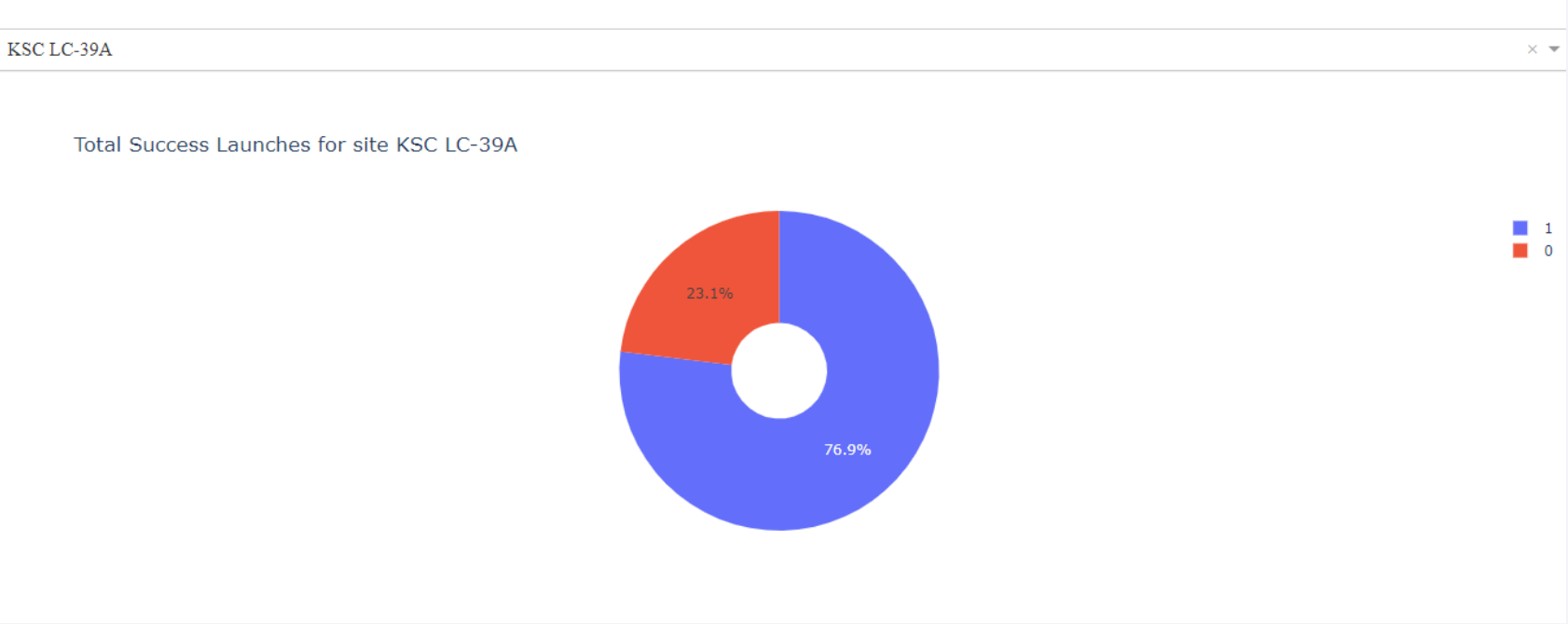

Section 4
Build a Dashboard with Plotly Dash

Success launches share by sites



- KSC LC-39A has the most successful launches

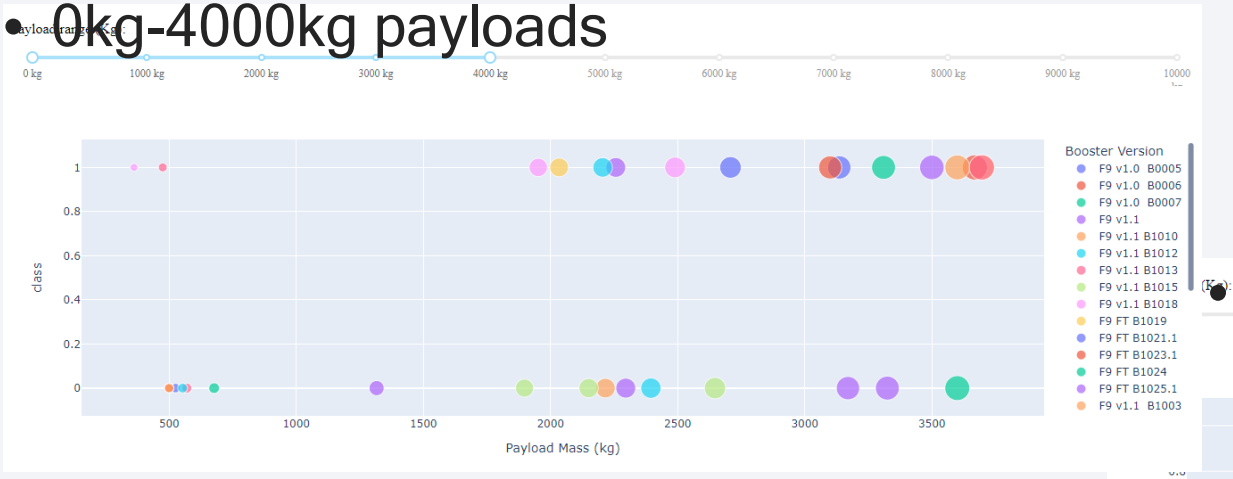
Success rate of KSC LC-39A



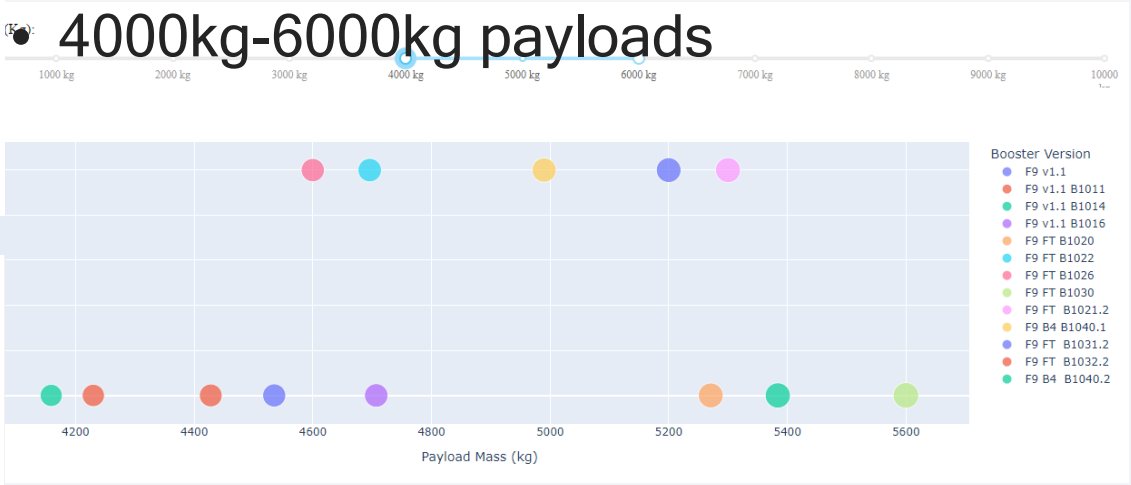
- KSC LC-39A success rate is 76.9%

Payloads & Success Scatterplot

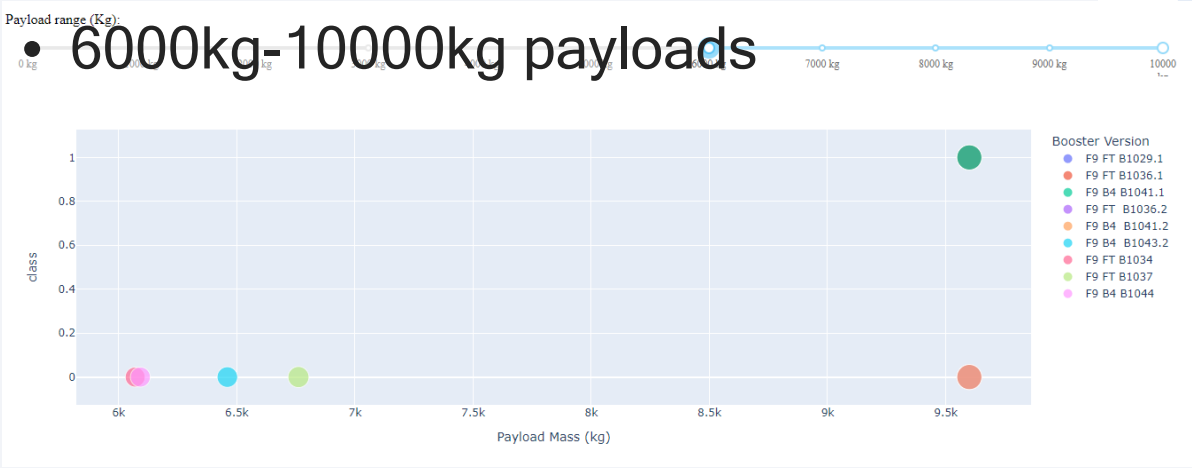
- 0kg-4000kg payloads



- 4000kg-6000kg payloads



- 6000kg-10000kg payloads

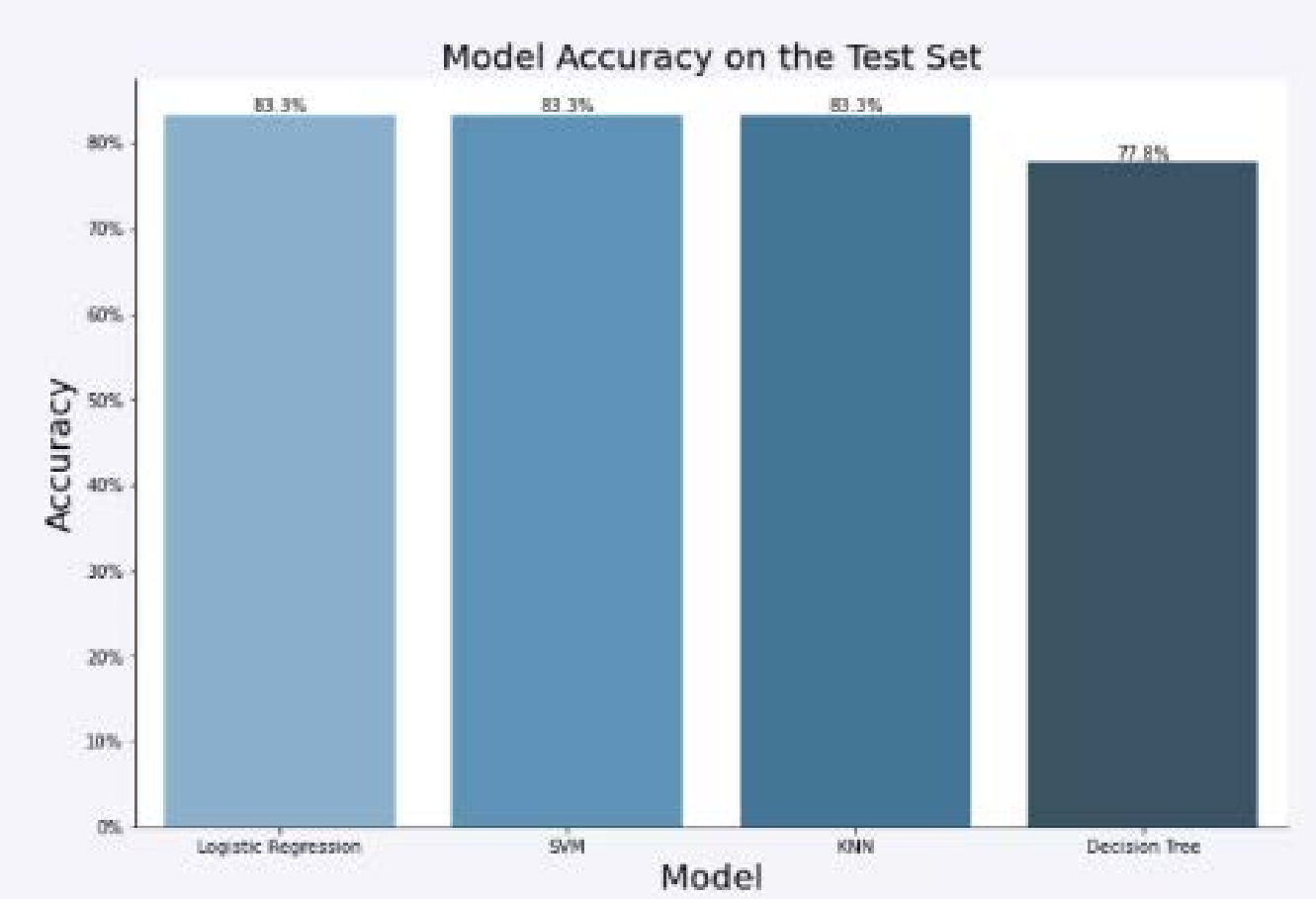


- Payloads correlates with success rate with a negative trends

Section 5

Predictive Analysis (Classification)

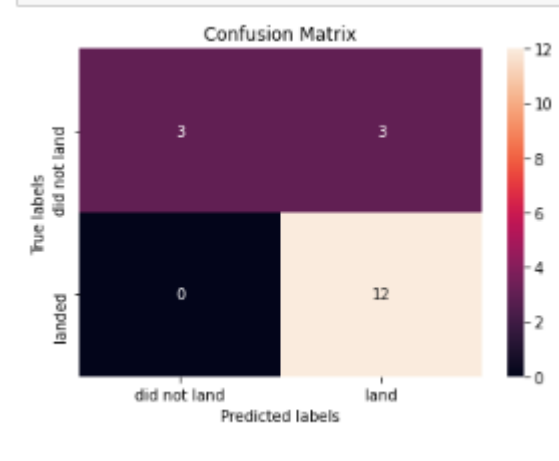
Classification Accuracy



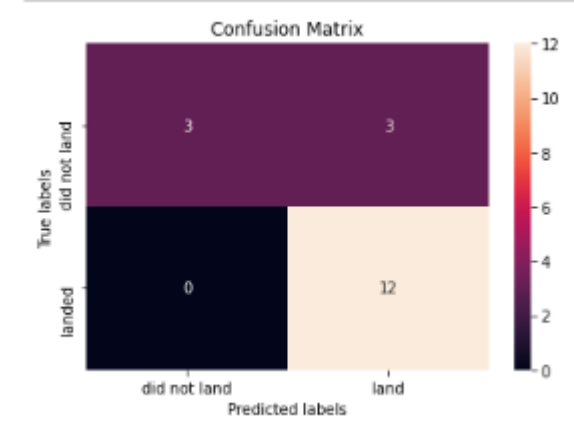

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7222222222222222
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix

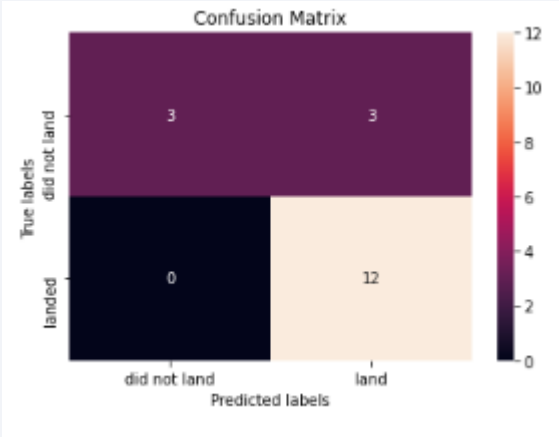
Tree



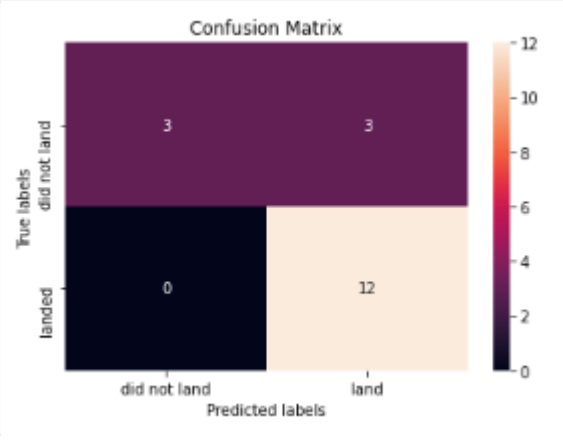
Log Reg



SVM



KNN



Conclusions

- The SVM, KNN & Log Reg models are the best models for predictions
- Success rate would be higher:
  - Lower payloads
  - By time series
  - Orbits ES-L1, GEO, HEO, SSO, VLEO
  - Site KSC LC-39A

Thank you!