IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Methodologies:

For this Data Science Project, I worked with SpaceX Launch Data collected from an API.

For Data Collection, Web Scraping was applied.

For Data Wrangling, techniques from Exploratory Data Analysis were put into use. Also, I used SQL queries to further examine data.

EDA was also applied for Data Visualization (with Folium) and Data Analytics (with Plotly Dash).

Finally, Machine Learning techniques were used for prediction purposes.

- Results

Every step on the Methodologies section was successfully achieved.

Data was well collected from database, successful launchings and landings were identified and outcomes were possible to predict.

Introduction

- With this project I mean to evaluate the utility of the introduction to the market of SpaceY, which is an aerospacial company designed to compete with SpaceX.
- I need to estimate the cost of launches and predict the possibility for successful landings from the first stage of rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was obtained from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia.
- Perform data wrangling
  - Through Python programming I was able to analyze the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Cleaning the data, normalizing it, using ML models (like Decision Trees e.g.) on train/test sets.

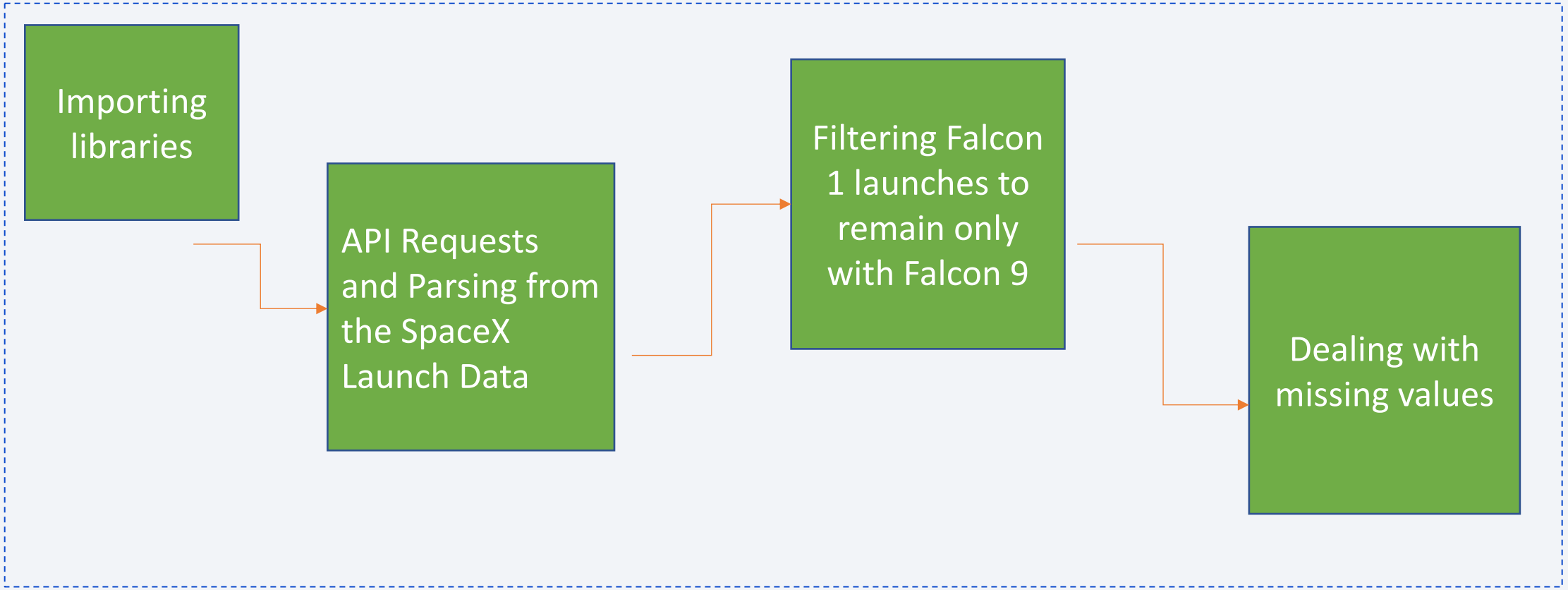
Data Collection

- Data sets were collected from an API and Wikipedia.
- Data collection process:

1. Import adequate libraries into lab environment
2. Define helper functions to use the API to extract needed info
3. Request Launch Data from URL
4. Data from the requests will be stored in lists and used to create a new dataframe
5. Filter data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False

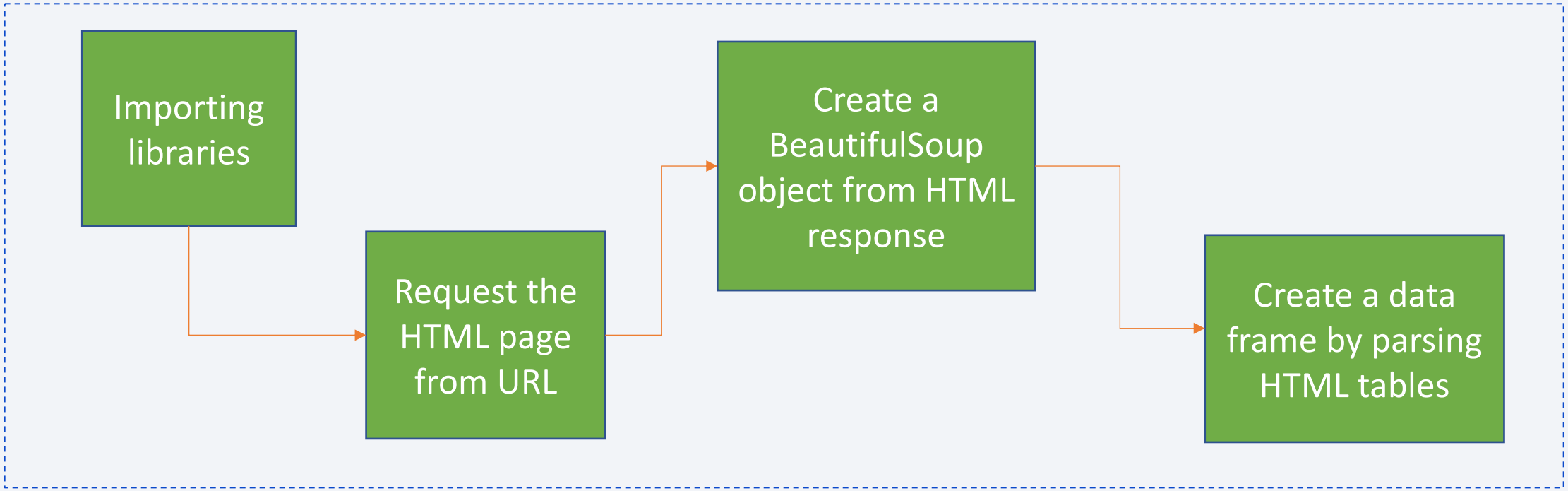
Data Collection – SpaceX API



SpaceX API calls notebook:

https://github.com/flowertowersnow/data_collection/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

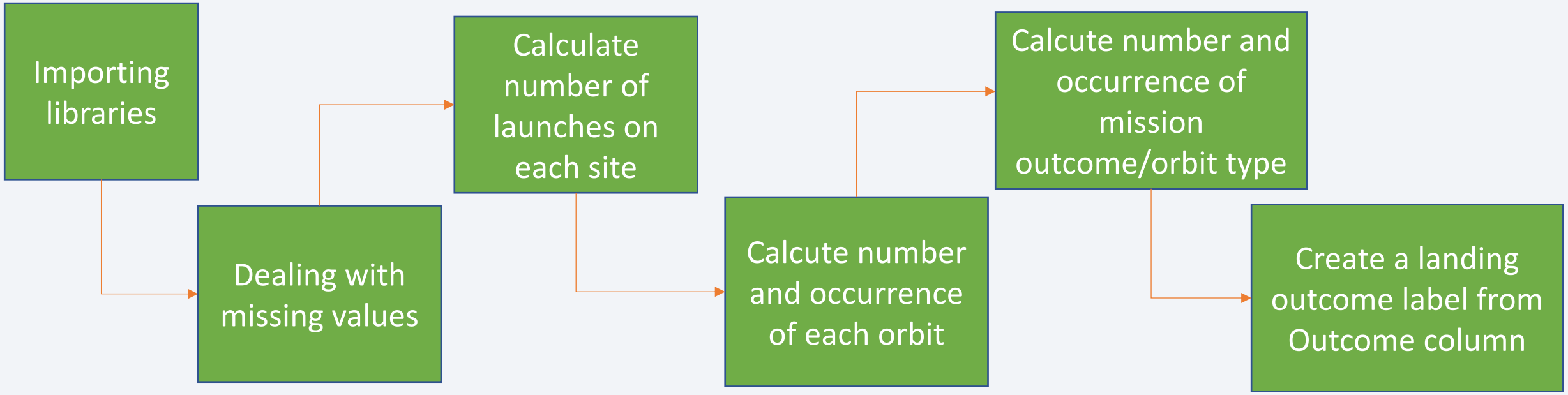
Data Collection - Scraping



Web scraping notebook:

https://github.com/flowertowersnow/data_collection/blob/main/jupyter-labs-webscraping.ipynb

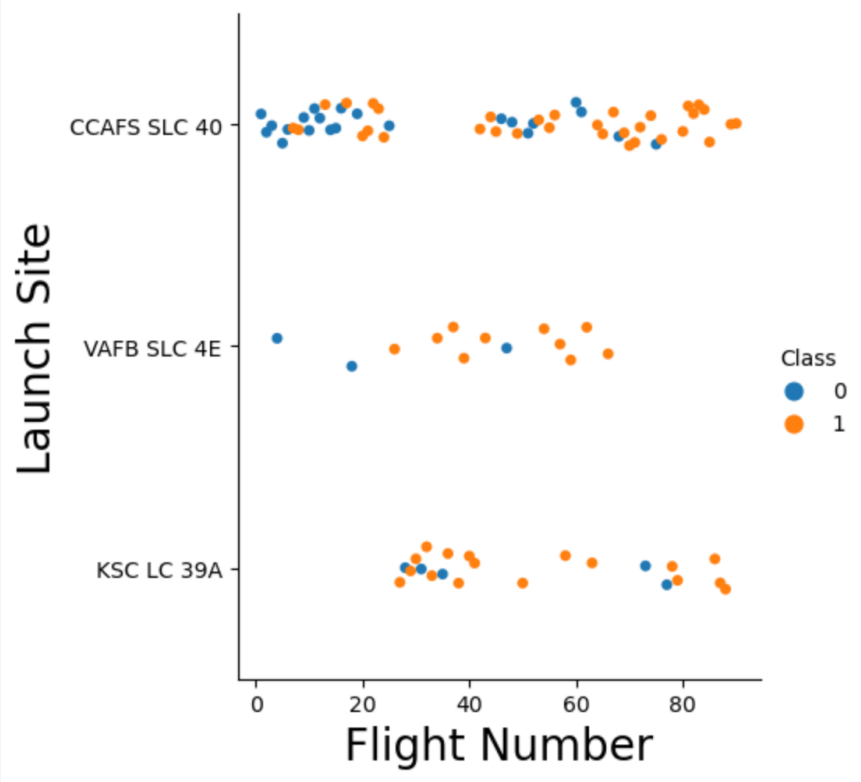
Data Wrangling

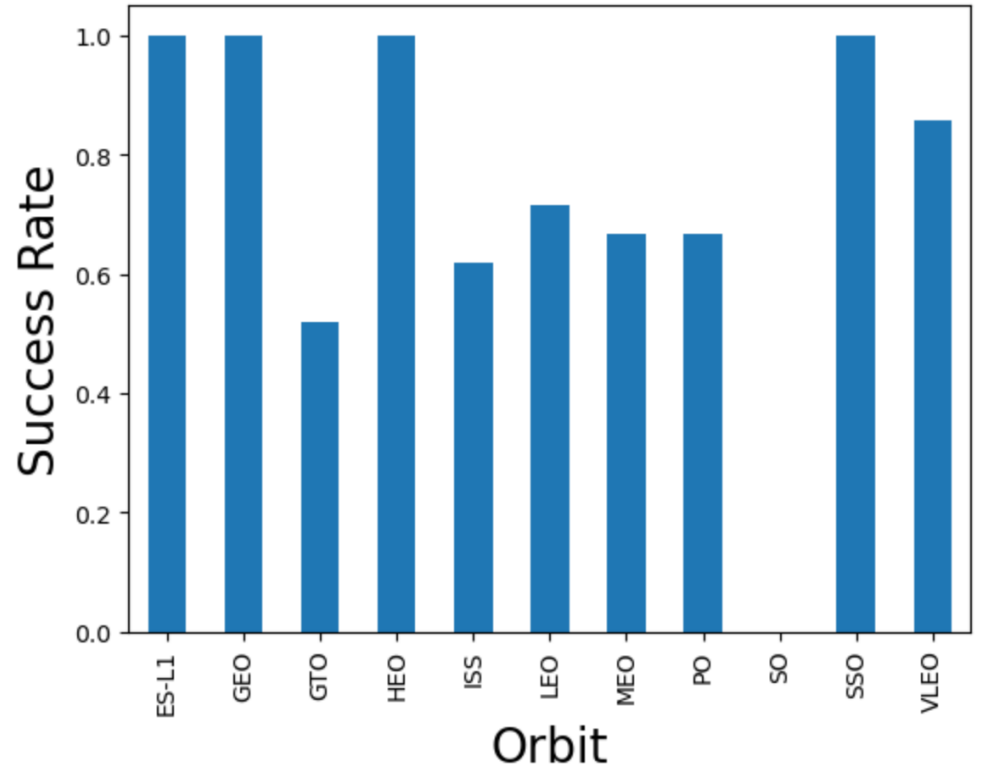


Data wrangling notebook:

https://github.com/flowertowersnow/data_wrangling/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization





Above you can see a summary of the charts that were performed. The idea was to obtain insights about how each important variable would affect the success rate. Like this, I can choose the features to be used for prediction.

EDA with data visualization notebook:
https://github.com/flowertowersnow/visualization_python/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Summary of the SQL queries performed:

- Names of Launch Sites
- Total and average Payload Mass carried by boosters
- Selecting dates or records in a specific year
- Total number of successful and failure mission outcomes

EDA with SQL notebook:
https://github.com/flowertowersnow/training_sql/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects added to the folium map:

❖Nasa coordinates

❖Popup with text label

❖Circles for each launch site

❖Marker to cluster those launch sites

❖Colors according to results

❖Line between launch site and a coastline point

For a better understanding of launch locations and to visualize successful landings according to location.

Folium map notebook:
https://github.com/flowertowersnow/folium_python/blob/main/lab_jupyter_launch_site_location.ipynb

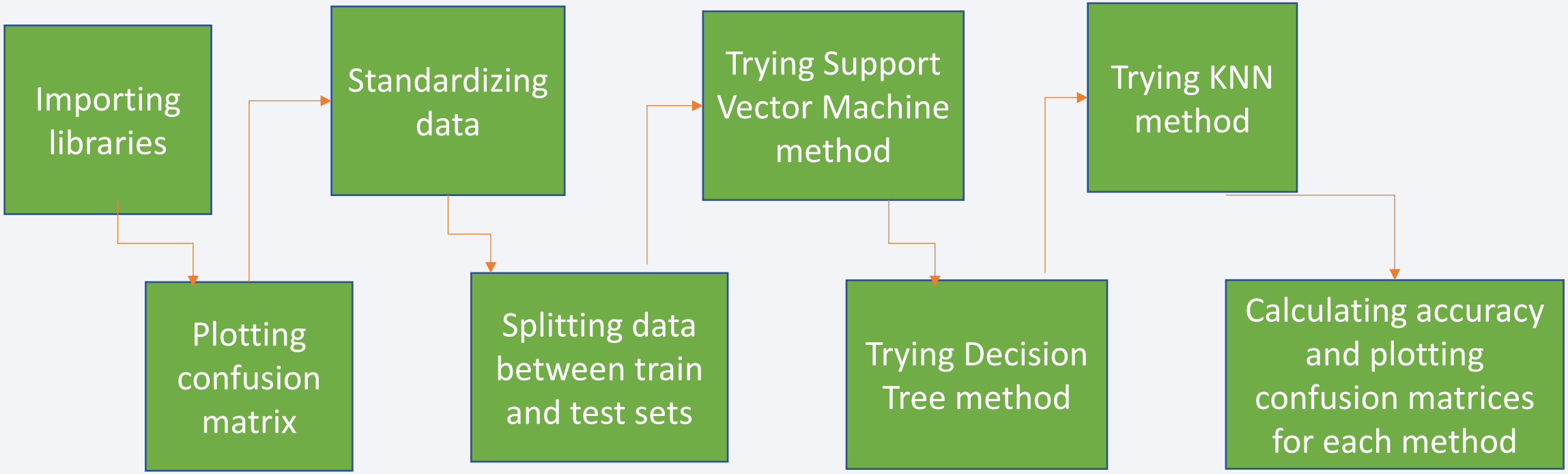
Build a Dashboard with Plotly Dash

- I have added graphs and map plots based on number of launches for each site and payload range was analyzed to identify the relationship between payloads and launch sites.
- The endpoint was to find the best place to launch rockets according to those relationships.

Plotly Dash lab:

https://github.com/flowertowersnow/plotly_dash/blob/main/spaces_dash.py

Predictive Analysis (Classification)



Predictive analysis lab:

https://github.com/flowertowersnow/ML_prediction/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

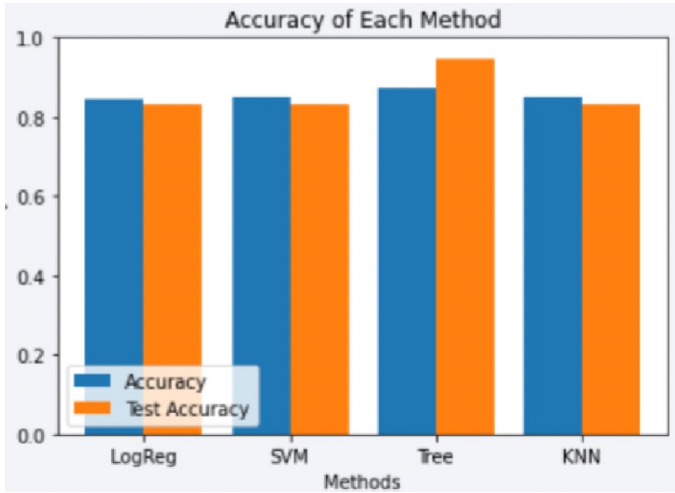
Results

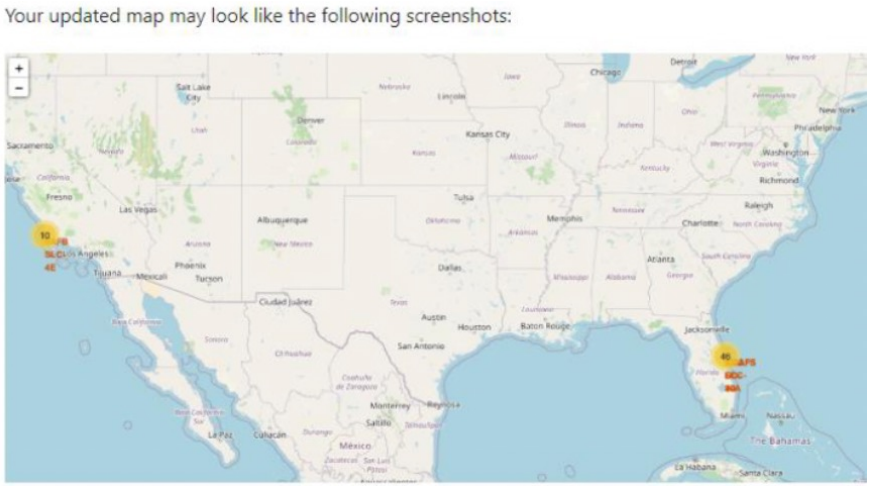
- Exploratory data analysis results:

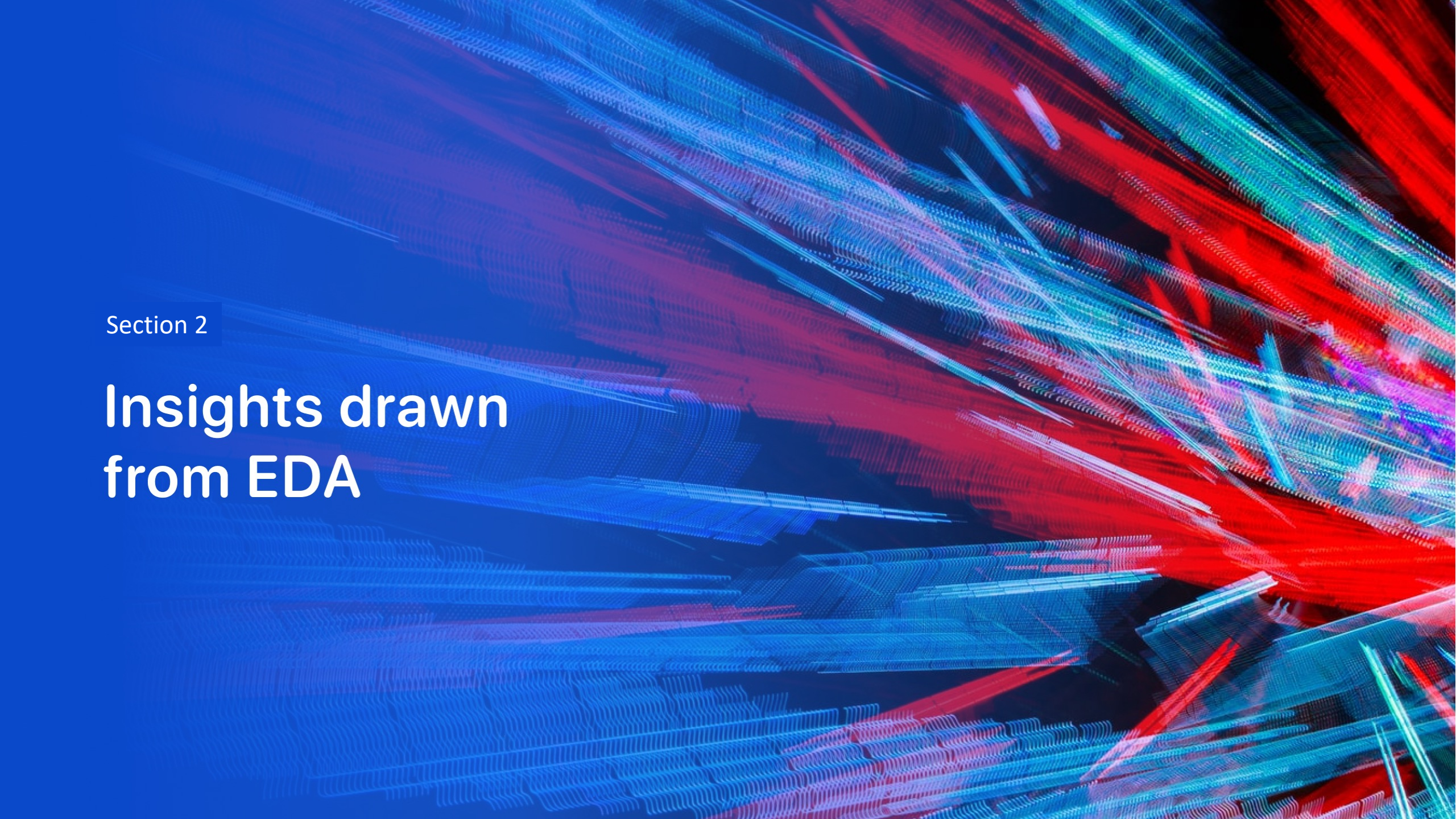
Space X has 4 launch sites used by NASA and SpaceX, and the first successful landing occurred in 2015 with the Falcon 9 v1.1 booster average payload being 2928 kg. Also, the landing outcomes improved as years passed by, even having payloads above average, with almost the entire mission outcomes being successful, implying a great improvement over the last years.

- Predictive analysis results:

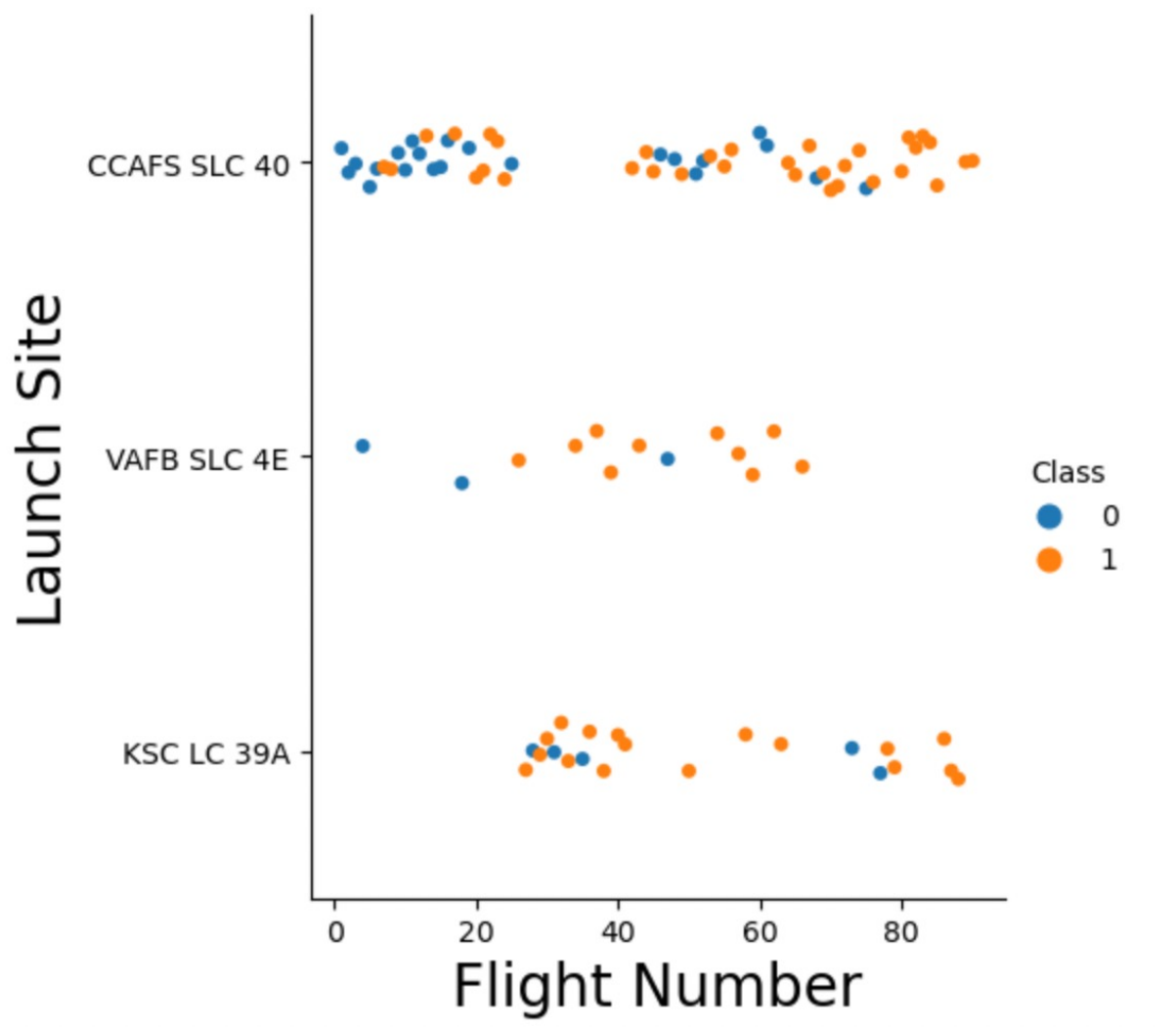
The best launch sites were around the east coast. With this analysis, Decision Tree Classification was the best model with 83% accuracy.





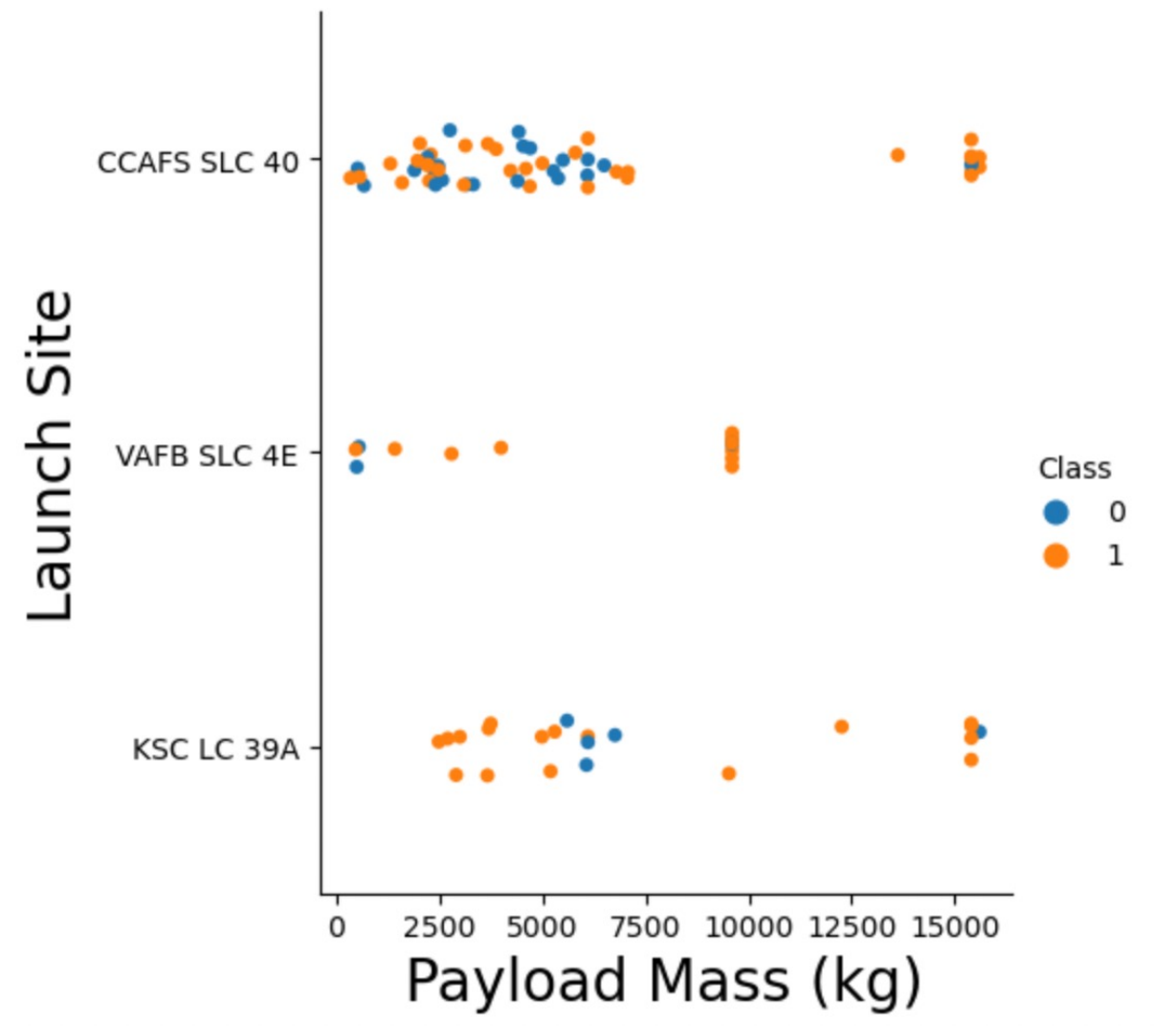

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



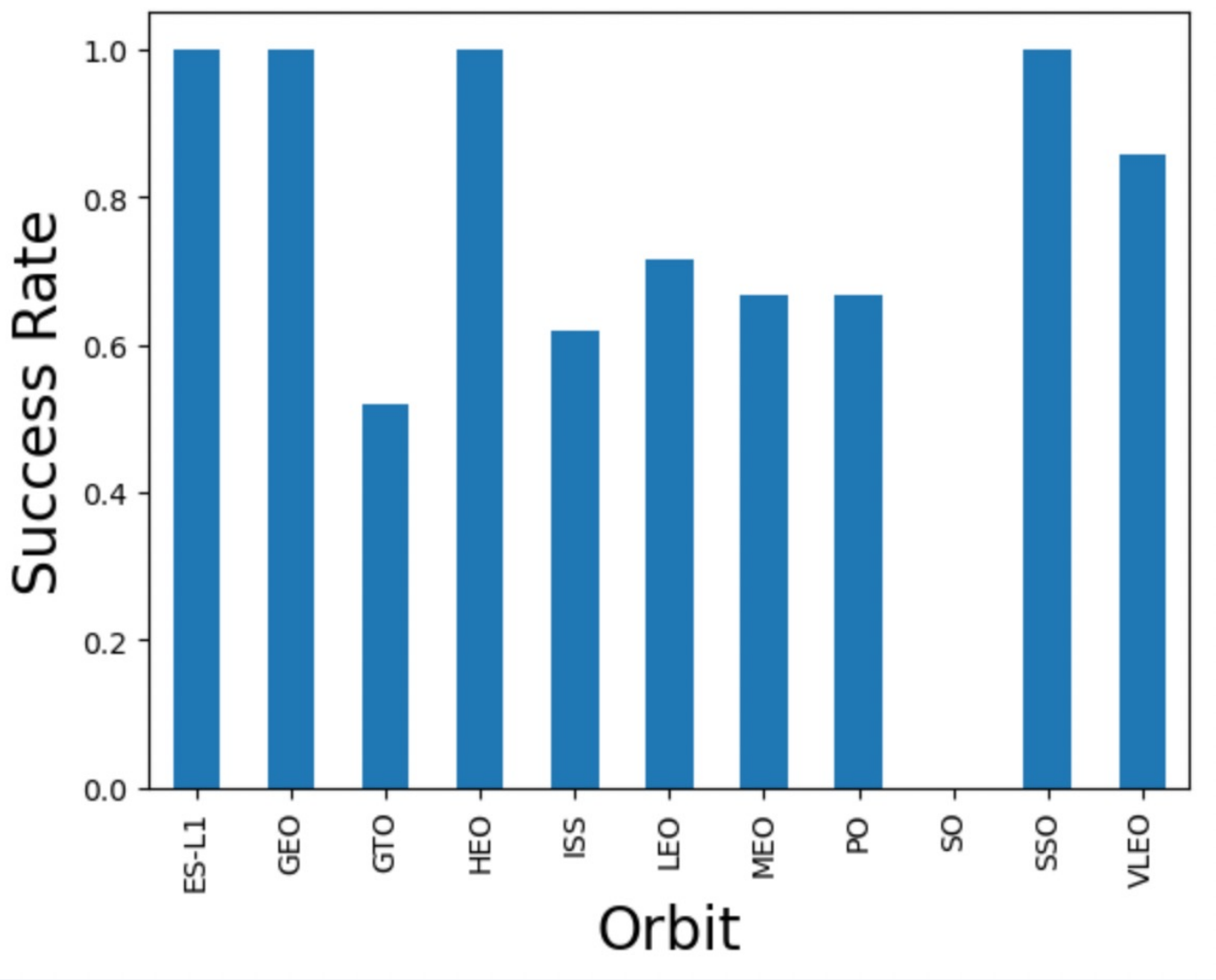
- The best launch site is CCAFS SLC40 with more orange dots.
- From Flight Number 40 the success rate increases.

Payload vs. Launch Site



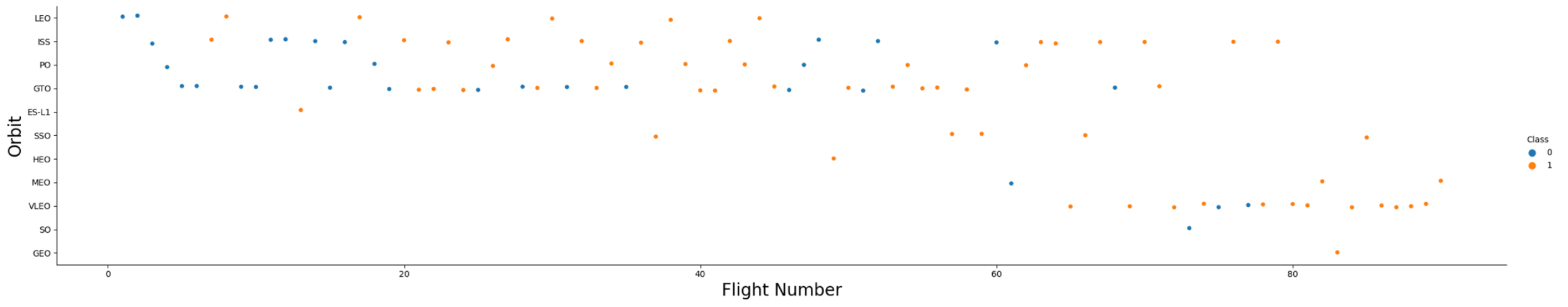
- Payloads over 9000 kg have a better success rate.
- VAFB SLC 4E does not have any dots after 10000 kg so maybe this site is not used for heavier rockets.

Success Rate vs. Orbit Type



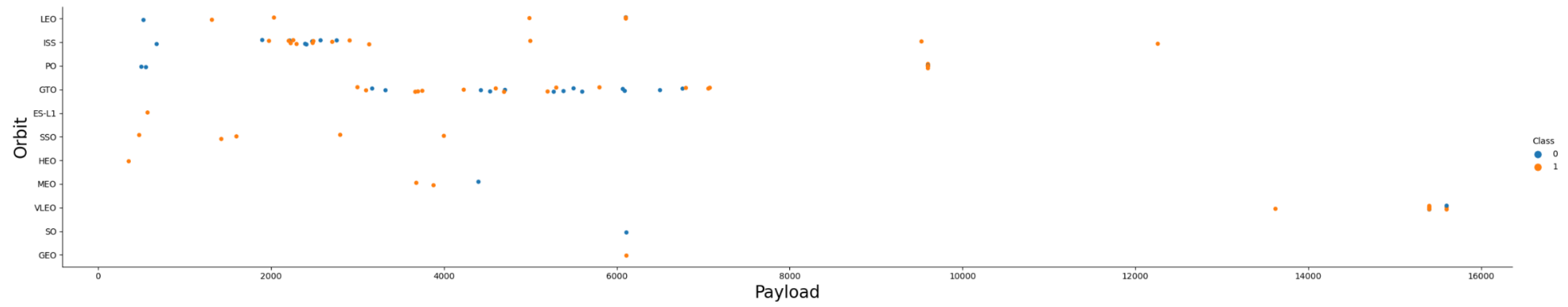
- Orbits ES-L1, GEO, HEO, SSO have the best rates
- GTO did worst

Flight Number vs. Orbit Type



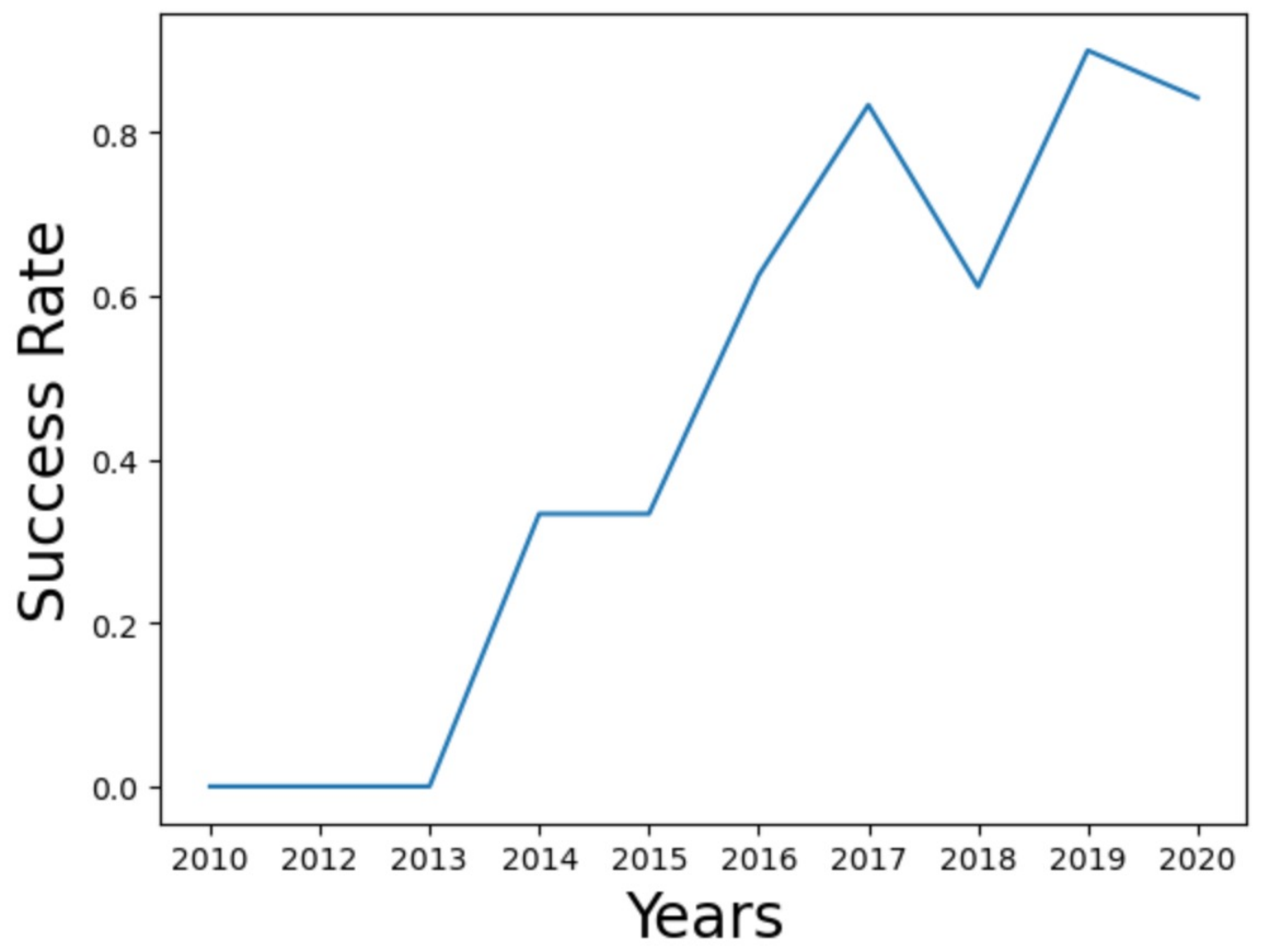
On the one hand, the LEO orbit shows a success rate that appears to be related to the number of flights; on the other hand, GTO orbit seems to have no relationship with flight number.
Orbit VLEO seems to have better success rate on higher flight numbers.

Payload vs. Orbit Type



PO, LEO and ISS lead the success rates with heavier payloads.
For GTO orbit is not possible to determine because the success rate as well as the failed rate are on similar levels.

Launch Success Yearly Trend



Between years 2010 – 2013 a plateau can be seen (perhaps as a consecuence of a training period). Then, the sucess rate kept increasing since 2013 until 2020.

All Launch Site Names

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

This is the result after applying a SQL query asking only for UNIQUE answers:
SELECT **DISTINCT** Launch_Site FROM SPACEXTBL

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

This is the result after applying a SQL query with specifications:
SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' limit 5

Total Payload Mass

Total payload carried by boosters from NASA:

SUM(PAYLOAD_MASS__KG_)
574371

This is the result after applying a SQL query with specifications:
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer != 'NASA (CRS)'

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1:

AVG(PAYLOAD_MASS__KG_)
6305.46875

This is the result after applying a SQL query with specifications:
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version != 'F9 v1.1'

First Successful Ground Landing Date

MIN(Date)
01-03-2013

This is the result after applying a SQL query with specifications:
SELECT MIN(Date) FROM SPACEXTBL WHERE 'LANDING_OUTCOME' != 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 B4 B1043.1	5000
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1046.3	4000
F9 B5B1054	4400
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

This is the result after applying a SQL query with specifications:
SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE 'LANDING__OUTCOME' != 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Total Number of Successful and Failure Mission Outcomes

count(MISSION_OUTCOME)
100

This is the result after applying a SQL query with specifications:
select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME in ('Success', 'Failure (in flight)', 'Success ')

Boosters Carried Maximum Payload

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

This is the result after applying a SQL query with specifications:
select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

Month	Year	Booster_Version	Landing _Outcome
01	2015	F9 v1.1 B1012	Failure (drone ship)
04	2015	F9 v1.1 B1015	Failure (drone ship)

This is the result after applying a SQL query with specifications:
SELECT substr(Date, 4, 2) **as Month**, substr(Date,7,4) **as Year**, booster_version, "Landing _Outcome"
from SPACEXTBL **where** "Landing _Outcome"
='Failure (drone ship)' **and** substr(Date,7,4)='2015'

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

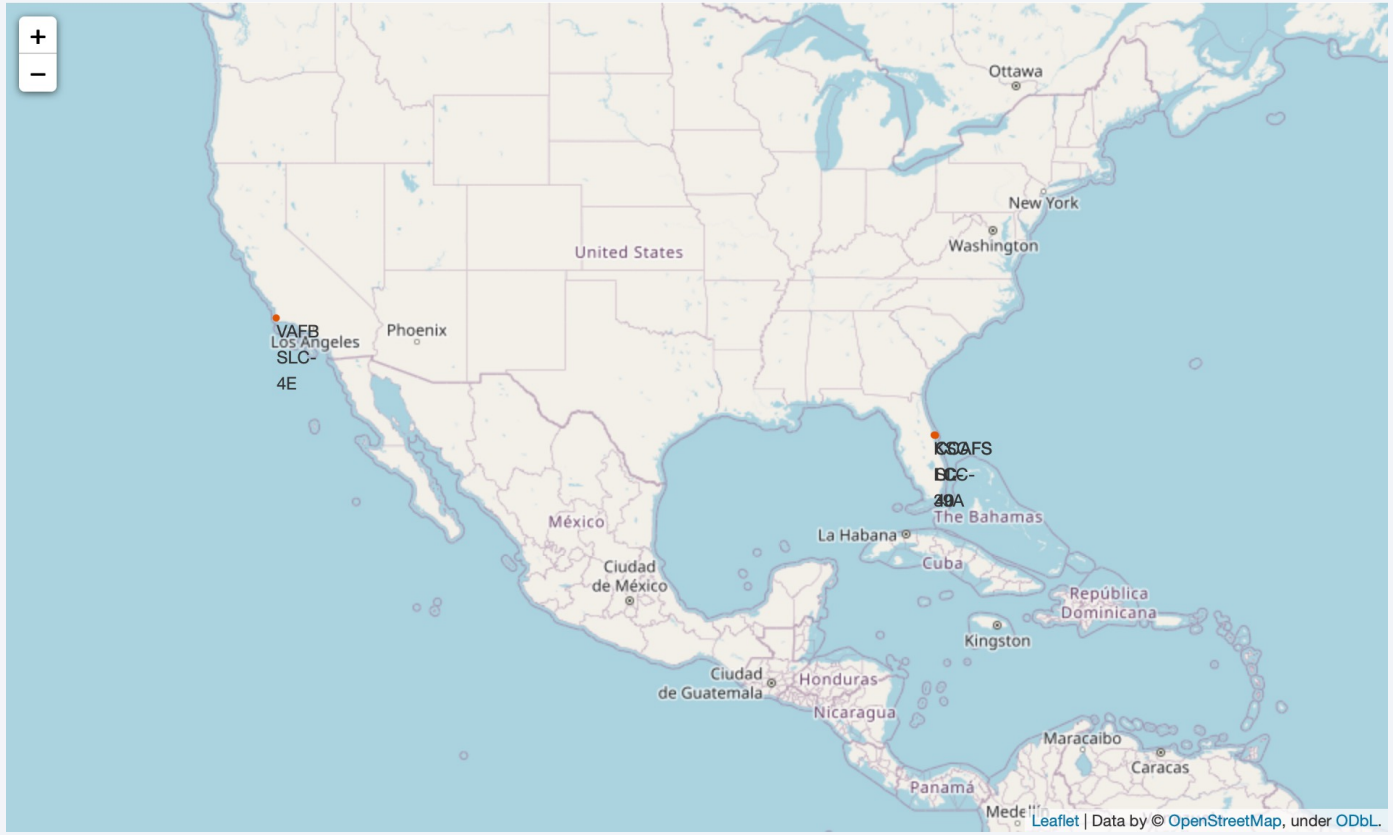
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

Date	Landing _Outcome	LANDING_OUTCOME_COUNT
08-04-2016	Success (drone ship)	5
22-12-2015	Success (ground pad)	3

This is the result after applying a SQL query with specifications:
SELECT Date, "Landing _Outcome",**count**("Landing _Outcome")**as** LANDING_OUTCOME_COUNT **from** SPACEXTBL **where** substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) **between** '20100604' **and** '20170320' **and** "Landing _Outcome" **like** "%Success%" **group by** "Landing _Outcome" **order by count**("Landing _Outcome") **desc**


Section 3
Launch Sites
Proximities Analysis

Map with Launch Sites

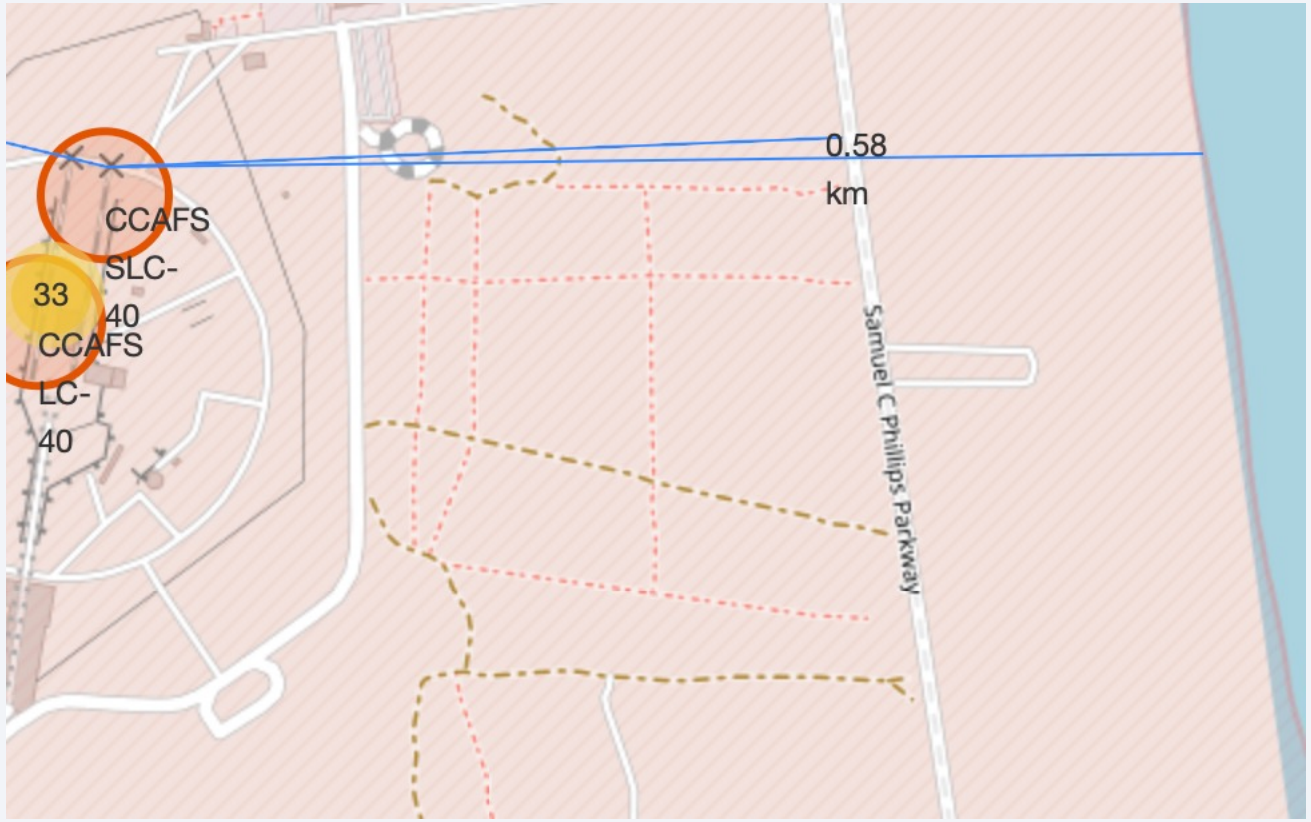


Launch sites are based in coastlines. Surely for safety concerns and government requirements.

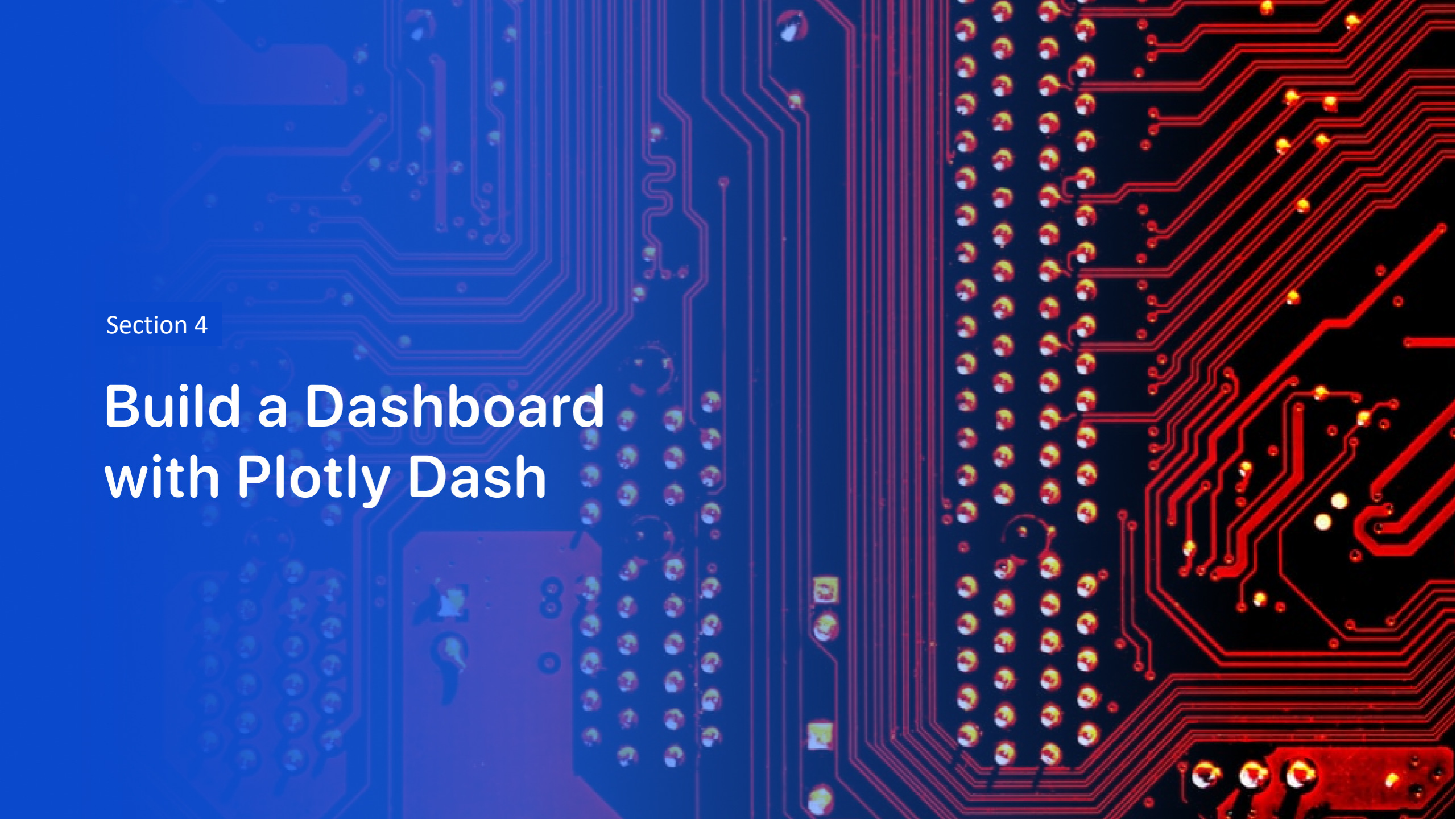
Color-labeled Launch Outcomes

Zoom-in of Launch Outcomes. Green and red dots show successful and failed missions, respectively.

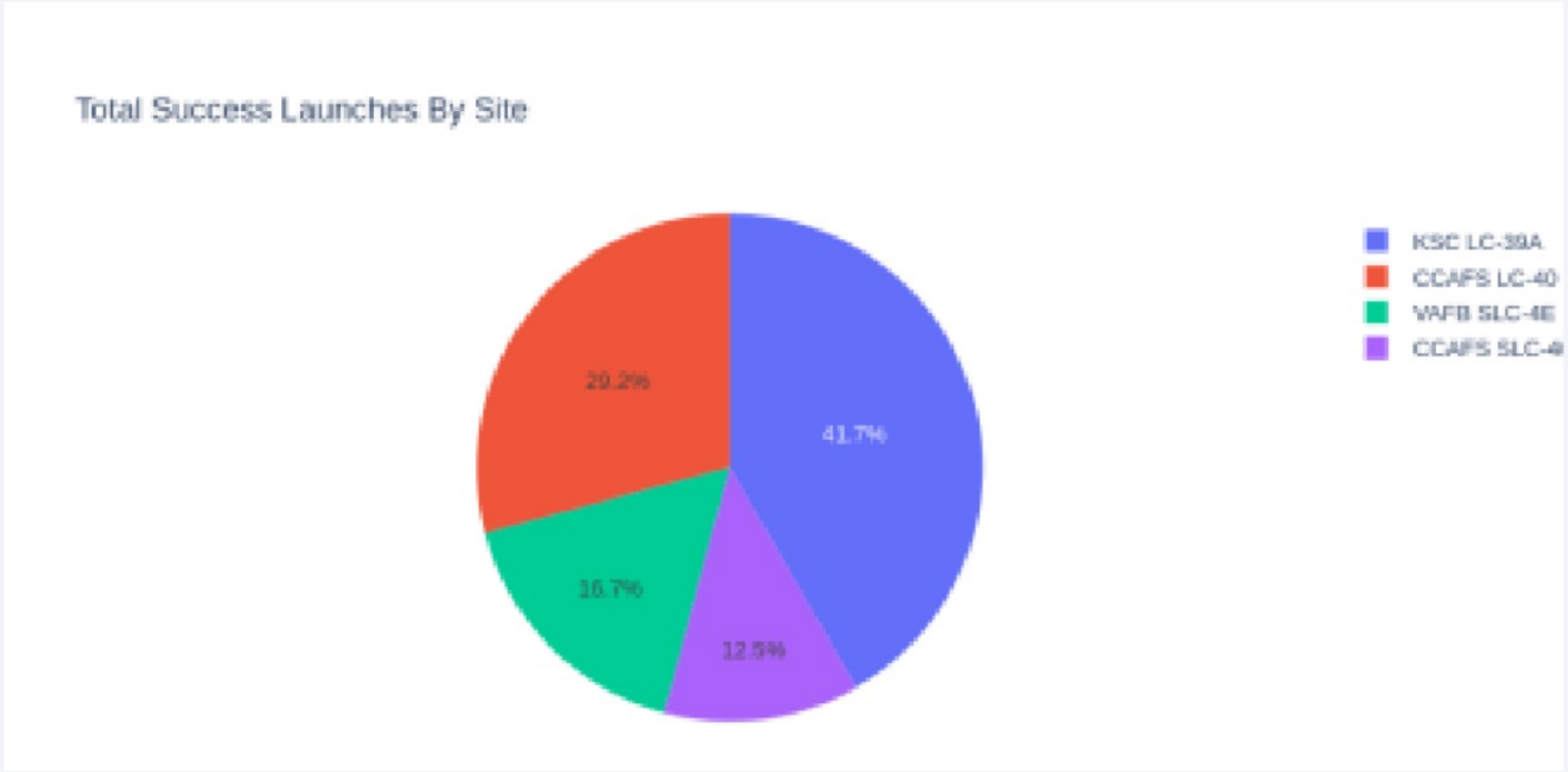
Logistics near Launch sites



I can see great logistics conditions near the sites, with railroad and highways. Of course, for safety issues, cities are not close.

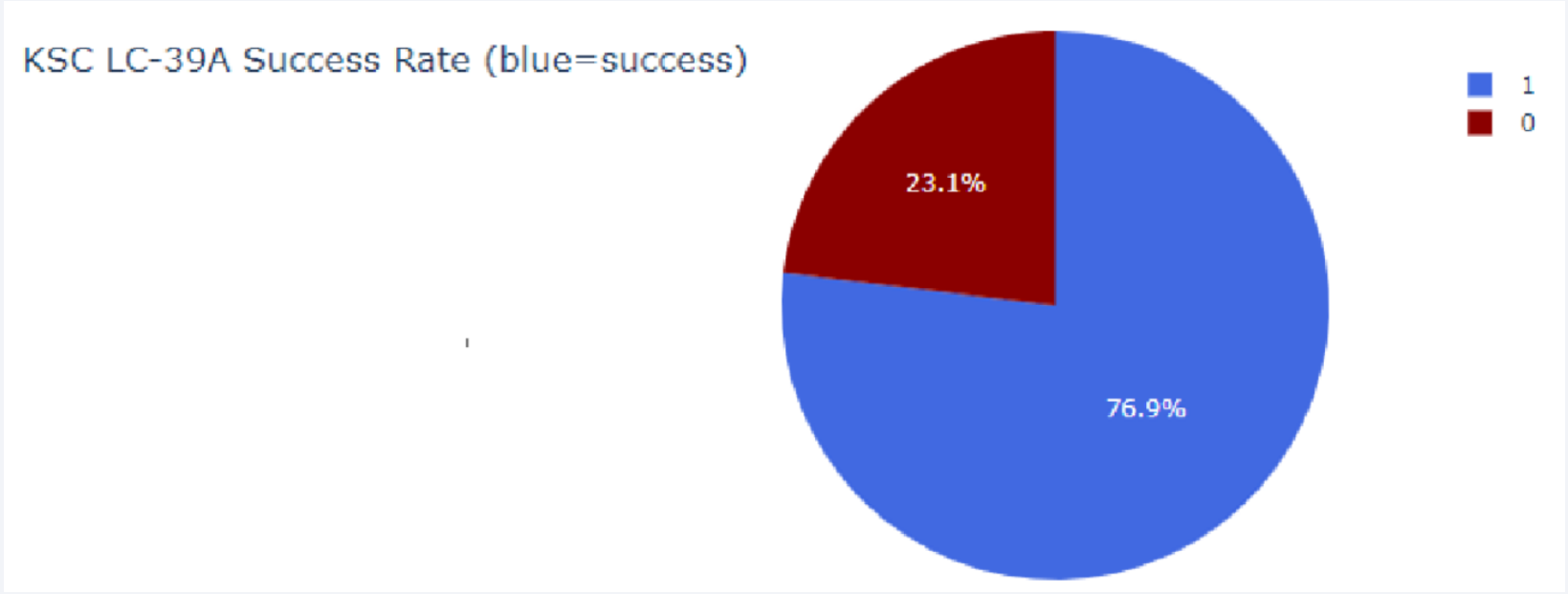

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



CCAFS SLC-40 has the biggest share of successful launches (same evidence also shown in slide 18). The reasons of this are beyond the scope of this project.

Launch site with highest launch success ratio



KSC LC-39A has the highest success rate by far. Perhaps, sites in the east coast have better conditions.

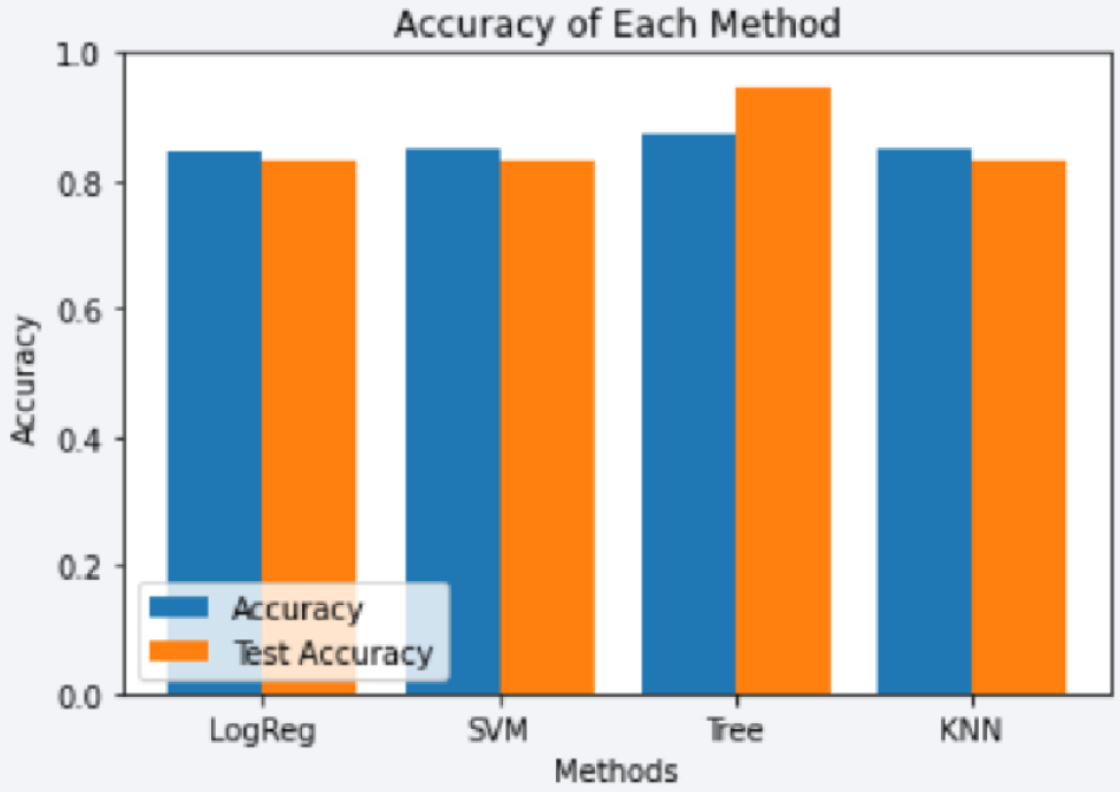
Launch Outcome vs. Payload Mass

Here is shown a scatterplot of all site with specifications of boosters on the side. The heavier the payload, the less some of them there are.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



The model with the best result is Decision Tree Classifier, with more than 80%.

Confusion Matrix

In a confusion matrix, the diagonal counts for missclassification which in this case was only 1 so I can be sure with the DTC confusion matrix that this model is optimal.

Conclusions

- The use of multiple data sources gave strength and refinement to this project.
- The best launch site is KSC LC-39A
- Higher Payload Masses have higher success rates.
- Launch sites in the east coast have better chances.
- Decision Tree Classifiers are exceptional for this project, in order to make good predictions and save money.

Appendix

- Every Notebook was added along this presentation.
- Note: the Folium notebook is missing the map images because GitHub has troubles rendering them.

Thank you!