IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kevin Nguyen
February 29, 2024

Outline

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

Executive Summary

The Falcon 9 rocket launches for SpaceX are quoted to cost significantly less than competing space programs due to reuse of the first stage. In this investigation, launch data available from the SpaceX API was studied to predict whether the first stage will successfully land, thus reducing the cost of the rocket launch. Exploratory data analysis was performed using Matplotlib and Dash Plotly for graphical visualization, Folium for geographical exploration, and SQL for targeting specific statistics and data groupings. The important features contributing to the success of a Falcon 9 launch were identified to be Launch Site location, Payload Mass, and Orbit type. A comparison of four predictive classification algorithms showed that Decision Trees achieved the greatest accuracy of 94.4% compared to 83.3% achieved by Logistic Regression, Support Vector Machines, and K-Nearest Neighbors.

Introduction

- The Falcon 9 rocket launches for SpaceX is advertised to cost $US 62 million, while other space programs have costs of up to $US 165 million. The savings come from the fact that SpaceX can reuse the first stage.
- The goal of this investigation is to predict if the Falcon 9 first stage will land successfully, and if so, the cost of the launch can be determined.
- If any alternate company wants to bid against SpaceX for a rocket launch, this data can be used for comparison.

Section 1

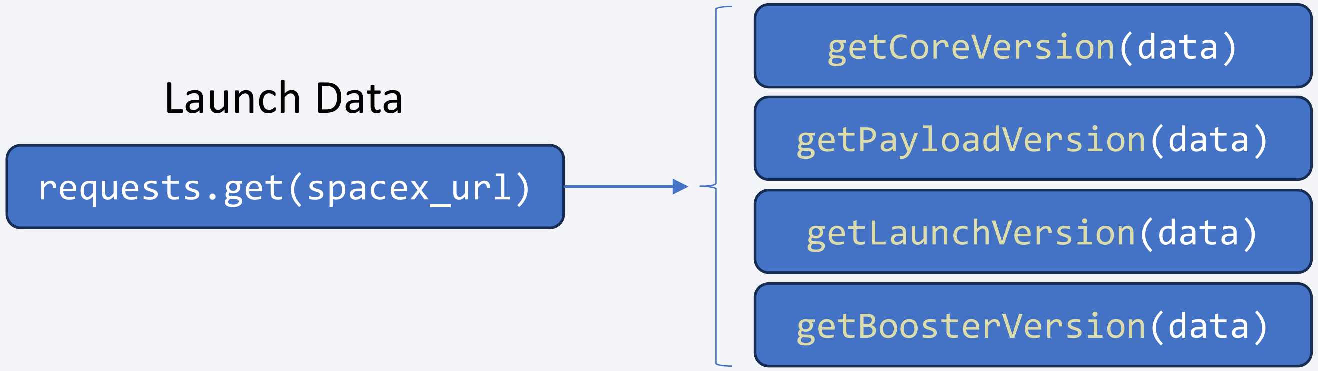
Methodology

Methodology

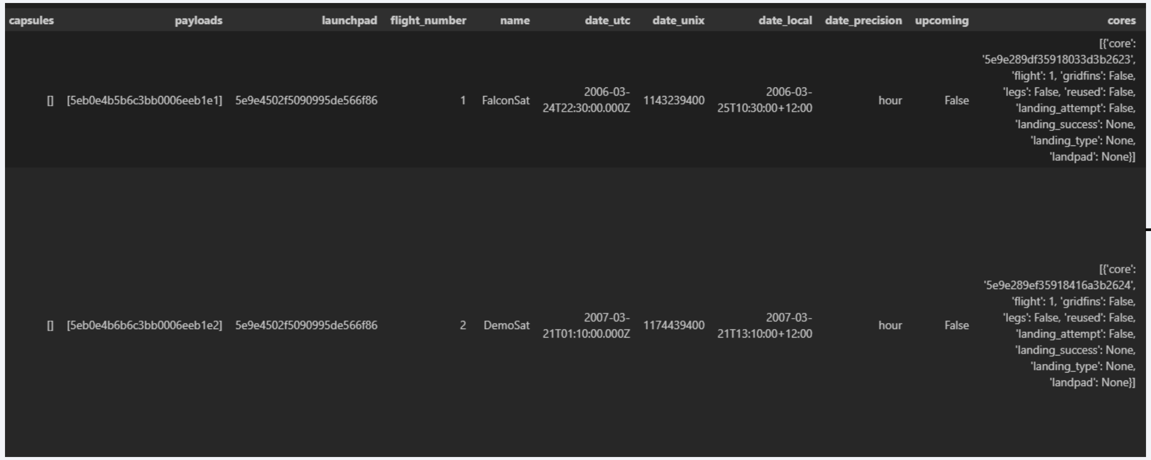
Executive Summary

- Data collection methodology:
  - Data collected using the SpaceX API, with alternative method being webscraping
- Perform data wrangling
  - Data was processed by distinguishing between good and bad landing outcomes and classifying each launch in as a binary variable. A success rate was determined from this.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare Logistic Regression, SVM, Decision Trees, and K-Nearest Neighbors

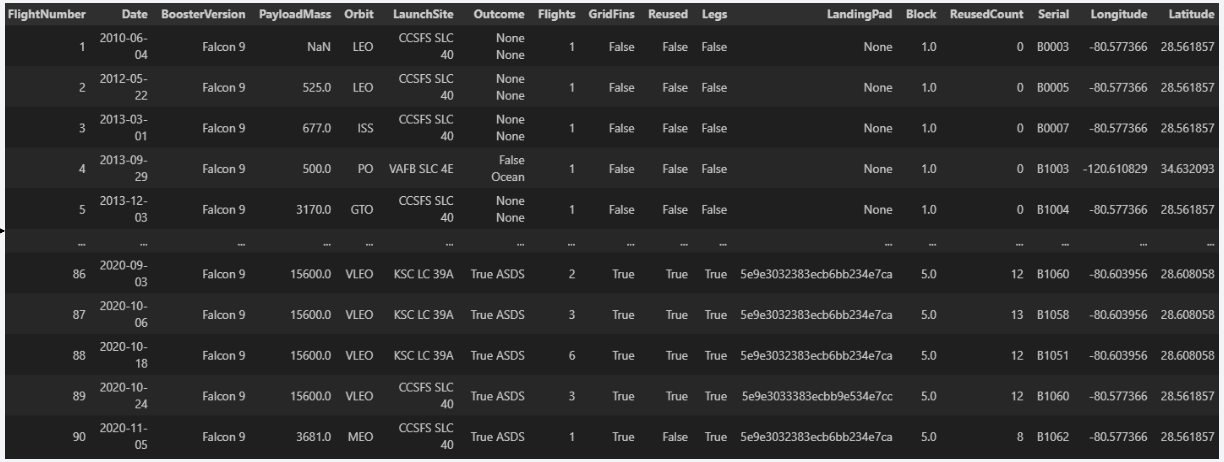
Data Collection – SpaceX API



- Past launch data is readily available from the SpaceX API.
- The launch data contains ID numbers. The IDs are used to extracting the measured data from the API.
- Each get function collects data from API and generates lists to be combines into a dataframe.



capsules	payloads	launchpad	flight_number	name	date_utc	date_unix	date_local	date_precision	upcoming	cores
[]	[5eb0e4b5b6c3bb0006eeb1e1]	5e9e4502f5090995de566f86	1	FalconSat	2006-03-24T22:30:00.000Z	1143239400	2006-03-25T10:30:00+12:00	hour	False	[{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}]
[]	[5eb0e4b6b6c3bb0006eeb1e2]	5e9e4502f5090995de566f86	2	DemoSat	2007-03-21T01:10:00.000Z	1174439400	2007-03-21T13:10:00+12:00	hour	False	[{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}]

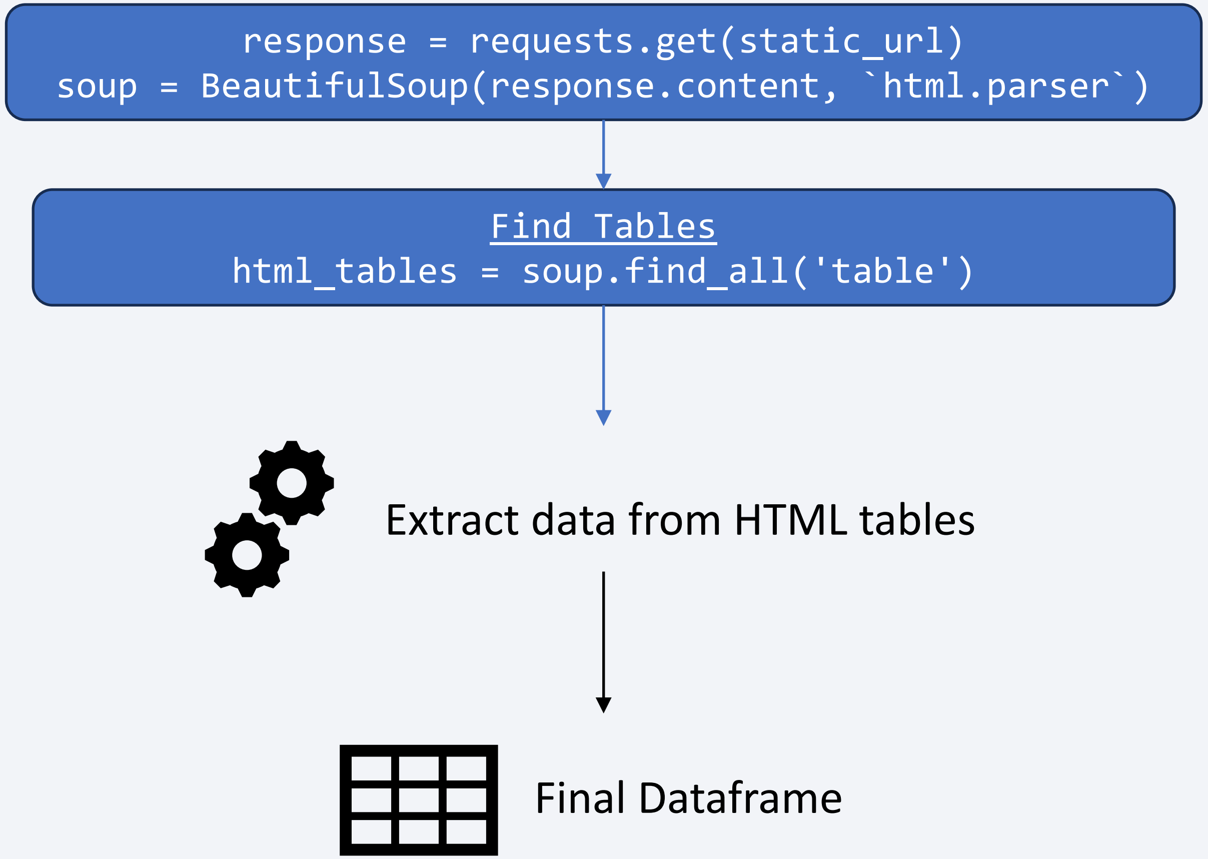


FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

Data Collection - Scraping

- Beautiful Soup package for Python webscraping

- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/jupyter-labs-webscraping.ipynb



```
response = requests.get(static_url)
soup = BeautifulSoup(response.content, `html.parser`)
```

Find Tables

```
html_tables = soup.find_all('table')
```

Extract data from HTML tables

Final Dataframe

Data Wrangling

- Separate the landing outcomes into those that failed and those that passed.
- Use the outcomes to categorize each launch into a pass or fail class
  - This will be the variable to be predicted using supervised machine learning
- Determined launch success rate from mean of class = 66.67%
- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Landing Outcome	Good/Bad Launch?	Class
True ASDS	Good	1
None None	Bad	0
True RTLS	Good	1
False ASDS	Bad	0
True Ocean	Good	1
False Ocean	Bad	0
None ASDS	Bad	0
False RTLS	Bad	0

EDA with Data Visualization

- Scatterplots between categorical and continuous variables are used between important variables.
  - Variables: Payload Mass, Orbit Type, Launch Site, Flight Number
- Each point is color-coded by class, making it easy to visual any noticeable clusters of passes and fails.
- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Queries:

- Unique Launch Sites in the space mission
- Launch Sites beginning with the sequence `CCA`
- Total Payload Mass carried by Boosters launched by `NASA (CRS)`
- Average Payload Mass carried by Booster Version `F9 v1.1`
- Date of achieving first successful landing on ground pad
- Names of boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
- Total number of successful and failed mission outcomes
- Names of boosters that have carried the maximum payload mass
- Records displaying month names, failed landing outcomes in drone ship, booster version, and launch site in 2015
- Ranking the count of different landing outcomes between 2010-06-04 and 2017-03-20 in descending order

GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

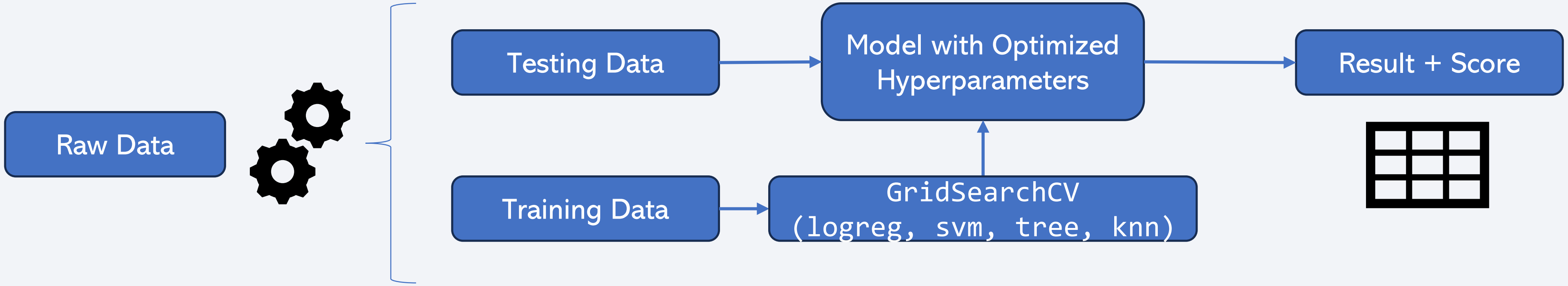
- A global map was generated to identify any commonalities in location between the launch sites.
  - Each site was labeled with a marker and pop-up label
  - Sites are clustered into Marker Cluster objects for a more appealing visualization
  - Launch Outcome Disposition was color-coded for each site
- Success
- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart for summarizing dispositions:
  - Proportion of successful launches between all Launch Sites
  - Success and fail rate for each site
- Scatter Plot to identify any trends or success indicators:
  - Launch Outcome Disposition by Payload Mass, Booster Version, and Launch Site
- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Compared 4 different machine learning classification algorithms
  - Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors
- Data is preprocessed using one-hot encoding and standardization
- Data was split into training and testing sets using sklearn train_test_split
- Models were trained and hyperparameters were optimized using sklearn GridSearchCV
- GitHub: https://github.com/ktnguyen10/ibm_ds_capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb


Raw Data
Testing Data
Training Data
Model with Optimized Hyperparameters
GridSearchCV (logreg, svm, tree, knn)
Result + Score

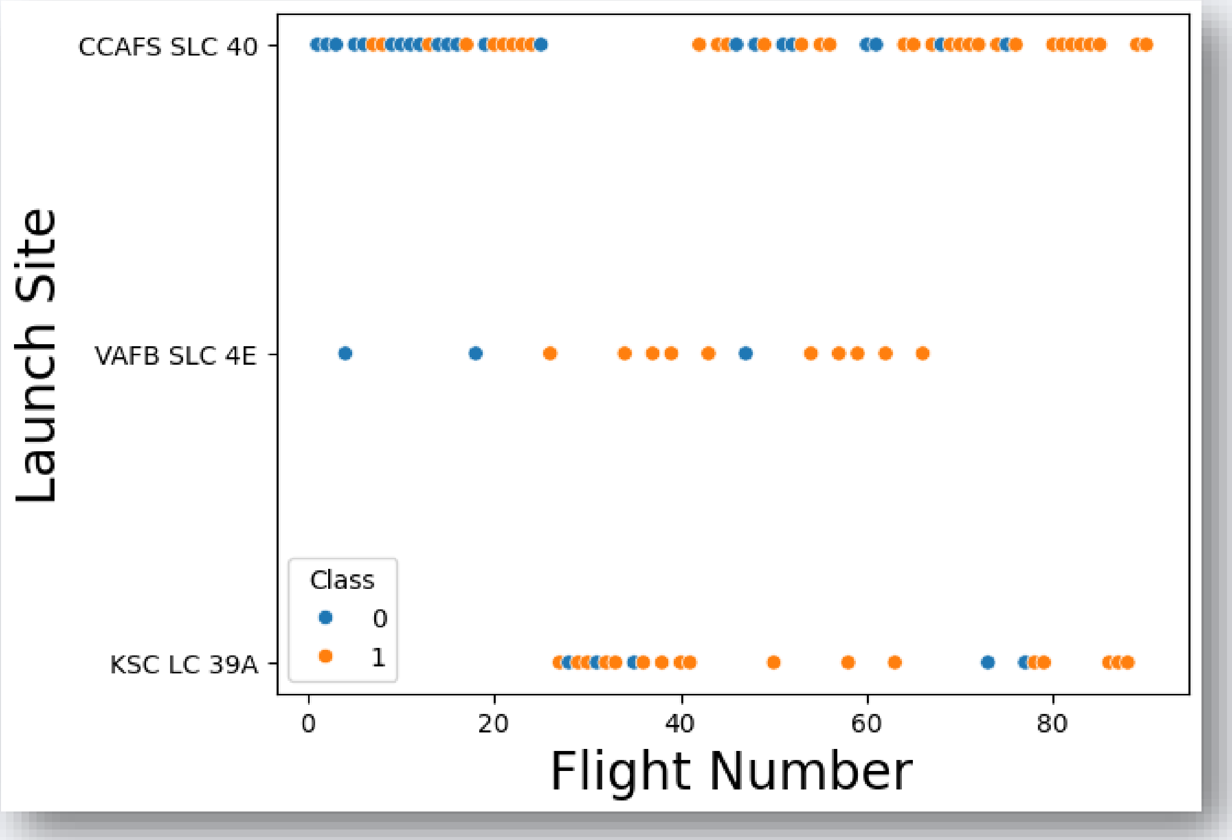
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

<u>Observations:</u>

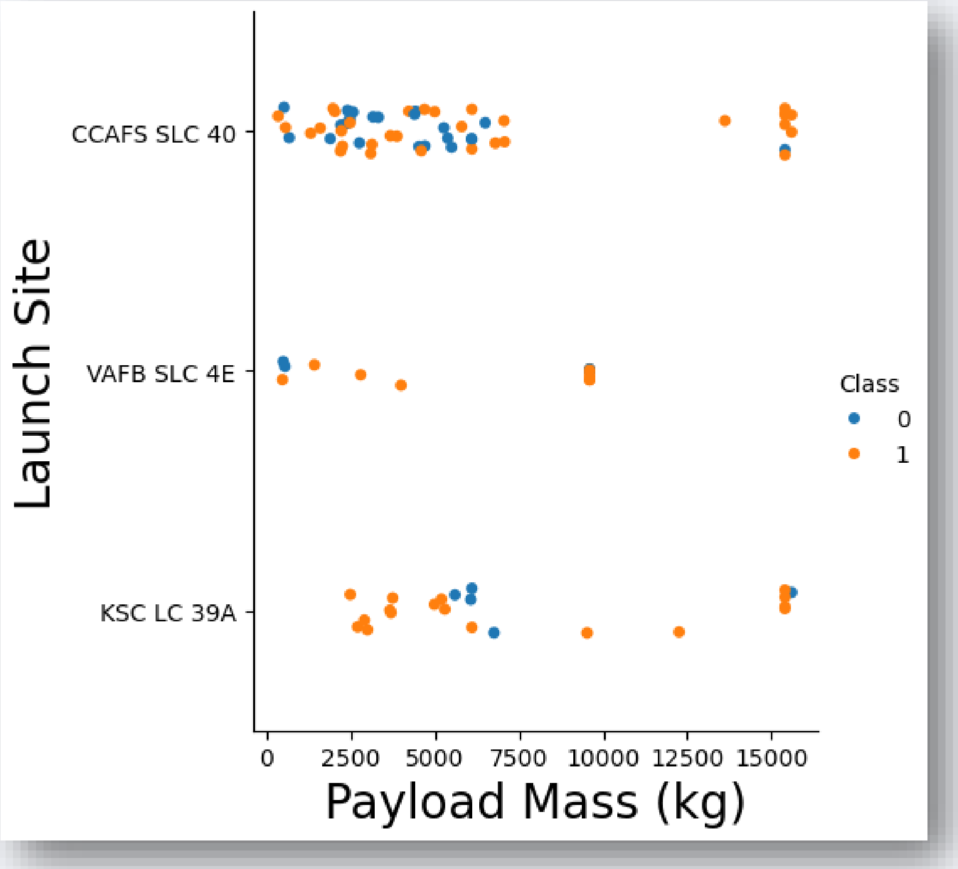
- Earlier flights have higher rate of failure
- Earlier flights are mostly at CCAFS SLC 40.



Payload vs. Launch Site

<u>Observations:</u>

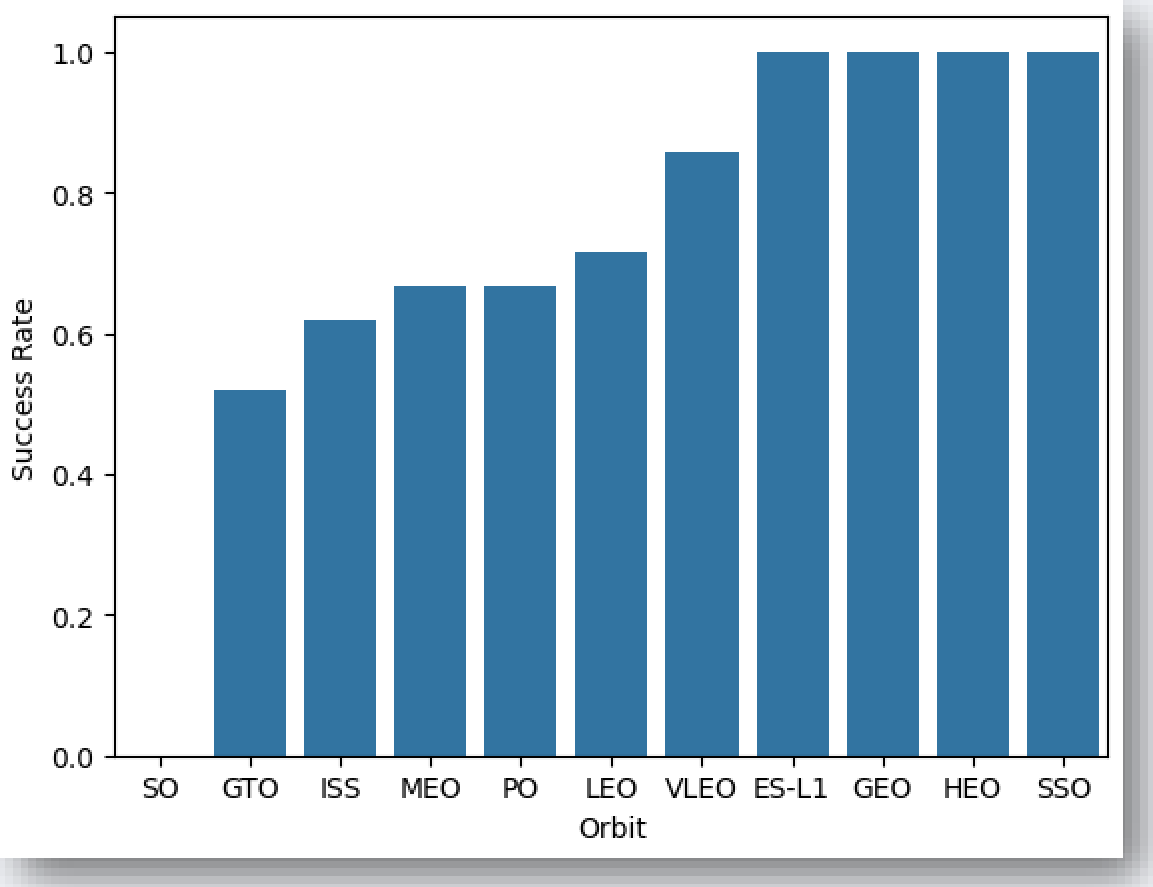
- High payload mass >7500 has a high success rate
- There is a boundary that can be inferred at 6000 kg for launch site KSC LC 39A.



Success Rate vs. Orbit Type

<u>Observations:</u>

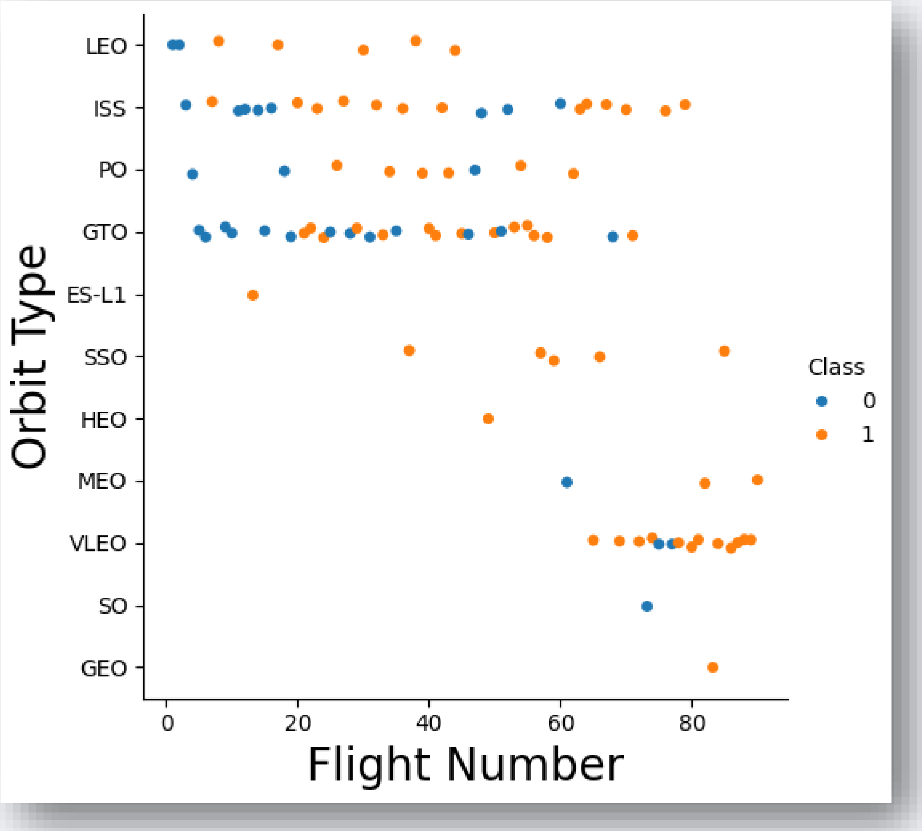
- ES-L1, GEO, HEO, and SSO Orbit types have 100% landing success rate.
- SO has a 0% success rate.



Flight Number vs. Orbit Type

Observations:

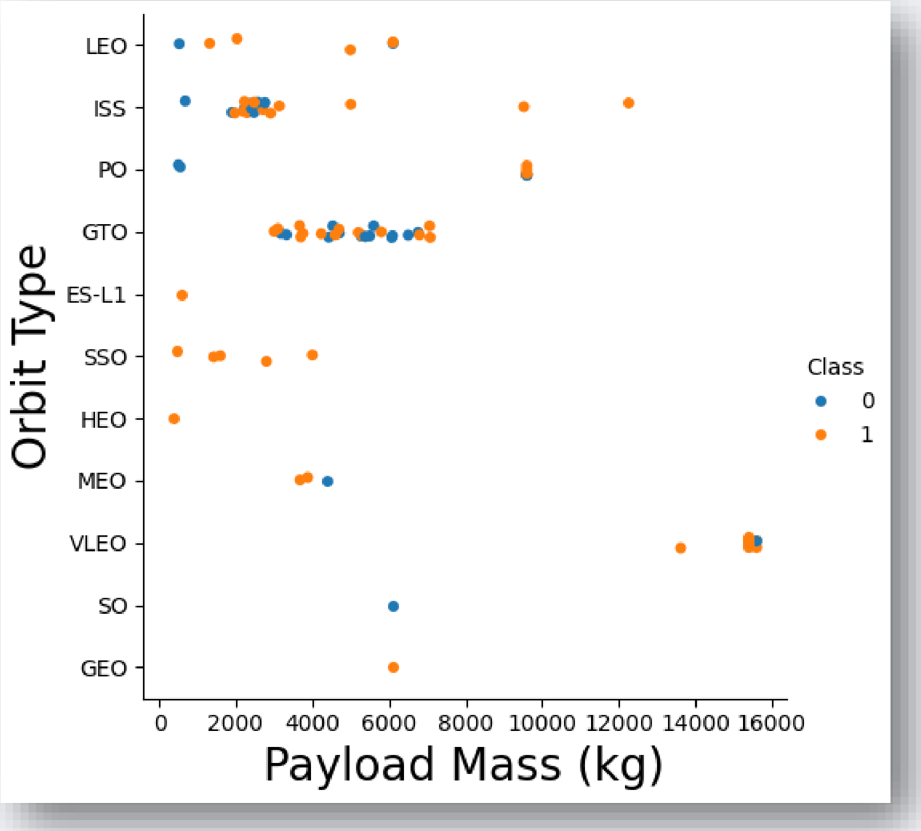
- ES-L1, GEO, HEO, and SSO Orbit types have 100% landing success rate, **however** the launch sample size is small.
- The Orbit types with the most frequent launches are ISS, GTO, and VLEO.
- VLEO has the highest success rate after excluding orbits with 100% and 0% success rates.



Payload vs. Orbit Type

Observations:

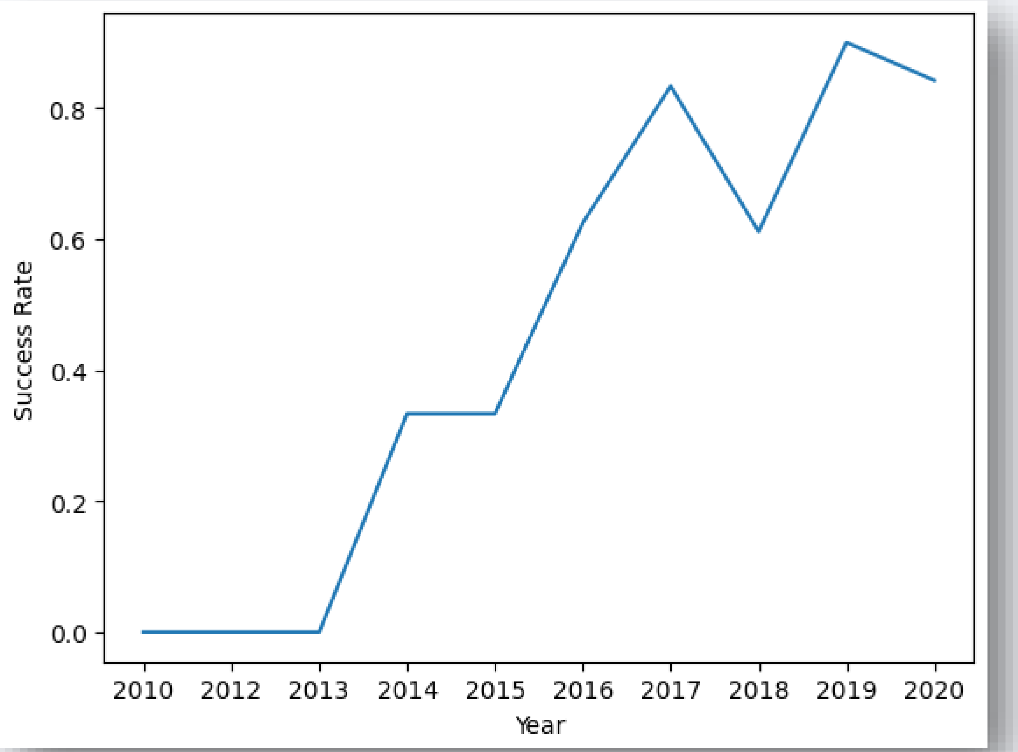
- Certain orbit types have specific payload mass ranges



Launch Success Yearly Trend

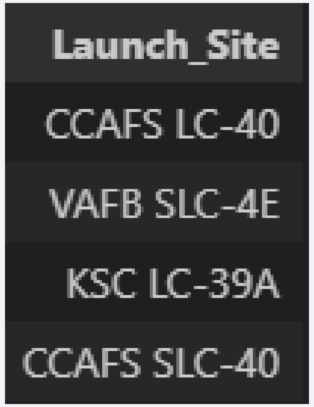
<u>Observations:</u>

- Launches improve over time, as illustrated by the launch success increase over the years.



All Launch Site Names

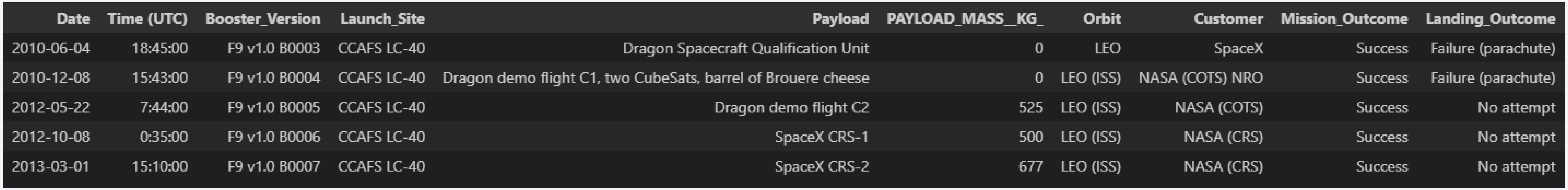
- Unique launch sites can be found by applying the DISTINCT function on the launch site column.
- The different launch sites are CCSFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40



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

Launch Site Names Begin with 'CCA'

- Displaying 5 full records where launch site names start with `CCA`.

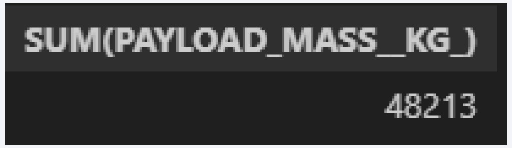


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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 total payload mass carried by boosters from NASA (CRS) is 48,213 kg.

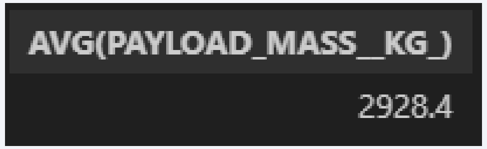
- The mass of all payloads where the customer name contains NASA and CRS was summed together.



SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

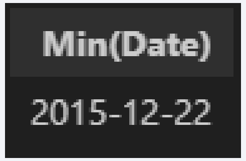
- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.
- The average of all payload masses where the Booster Version contains F9 and v1.1 was calculated.



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

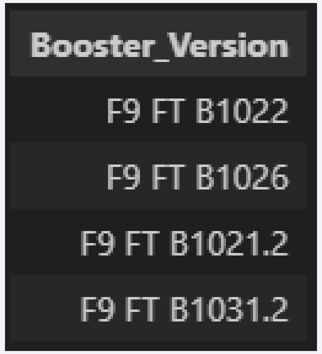
- The minimum date, 2015-12-22, was selected from records where the landing outcome contains `success` and `ground pad`



Min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

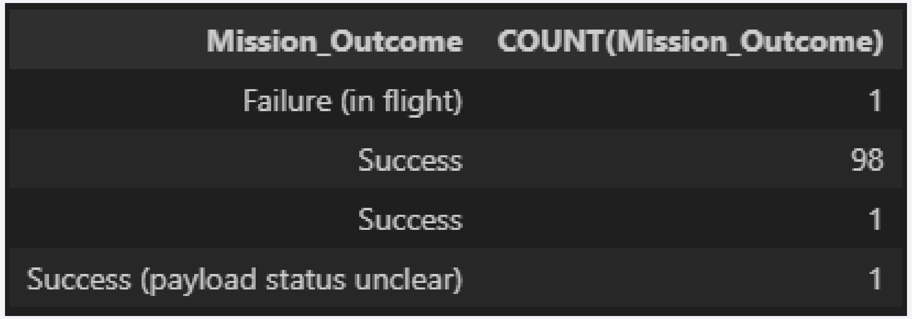


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The unique booster versions that have a successful landing outcome on a drone ship with payload mass between 4000 and 6000 are F9 FT B1022, F9 FT B1026, F9 FtB1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

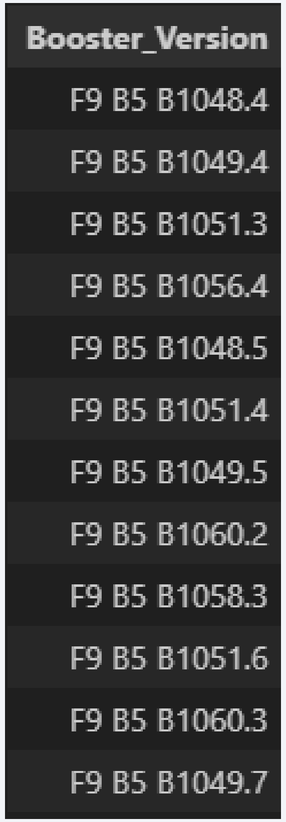
- Count all mission outcomes, grouping by the mission outcome.



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

Boosters Carried Maximum Payload

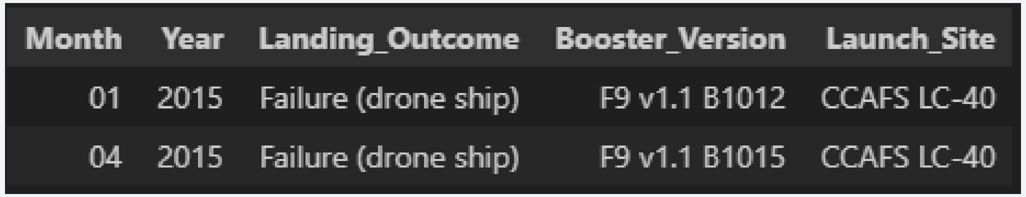
- The boosters versions that have carried the maximum payload mass are shown in the table.



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

2015 Launch Records

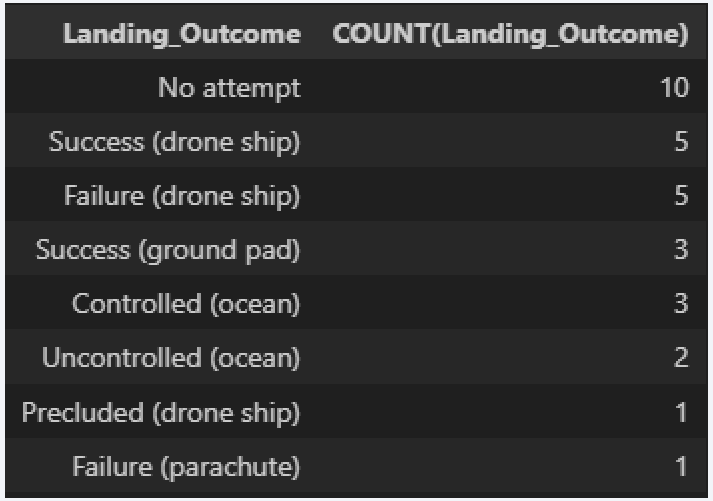
- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:



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

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

- Rank of 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:

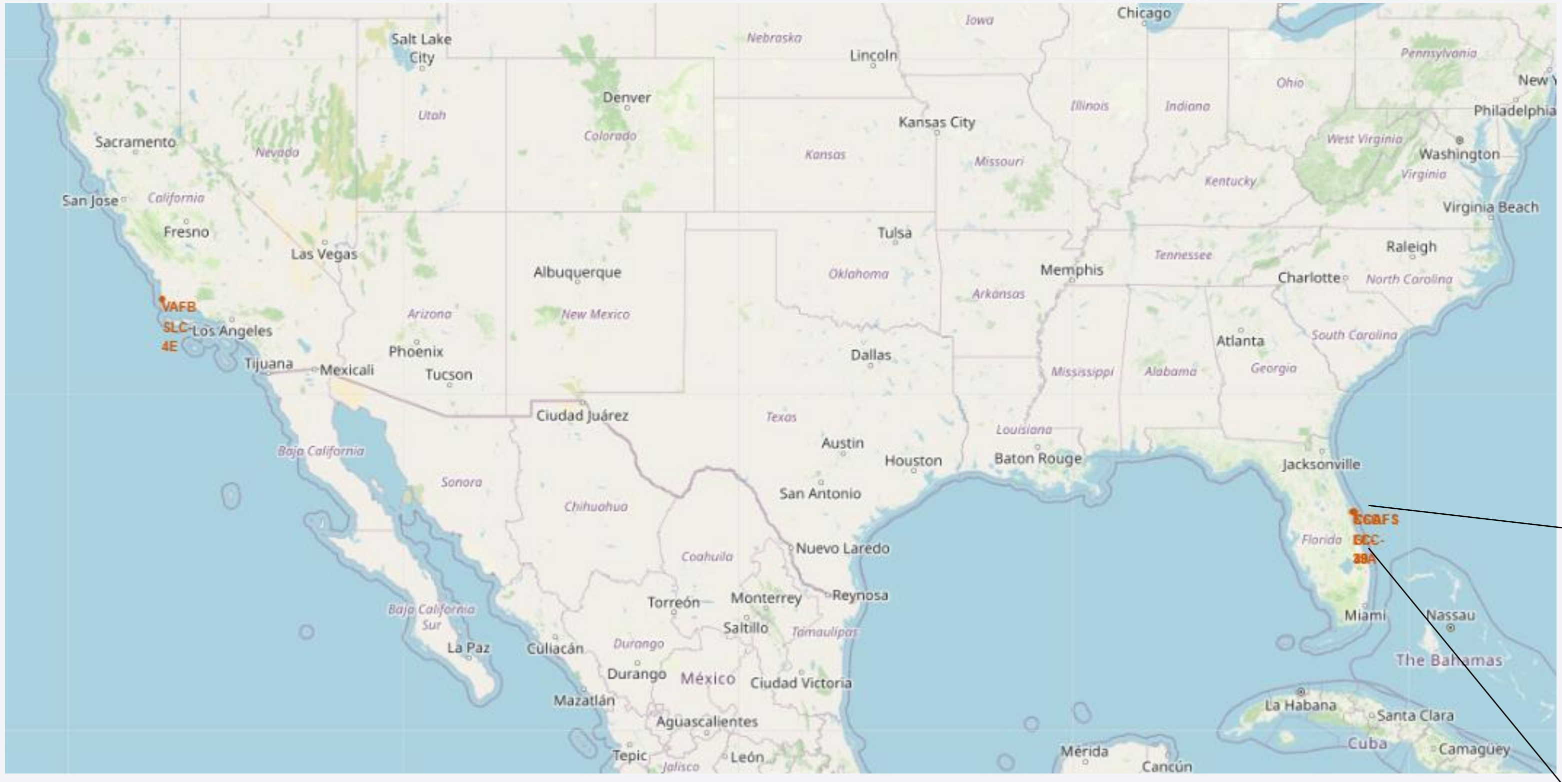


Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

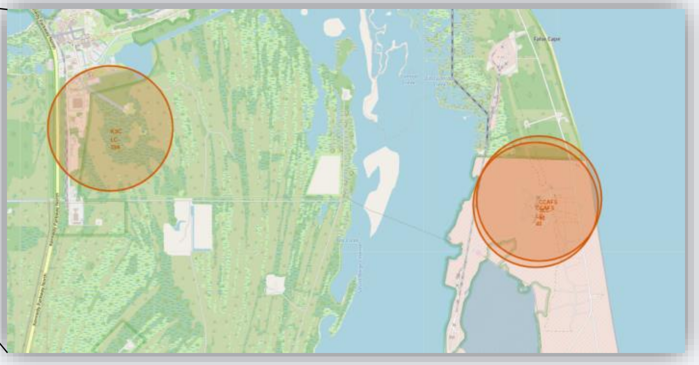
Section 3

Launch Sites Proximities Analysis

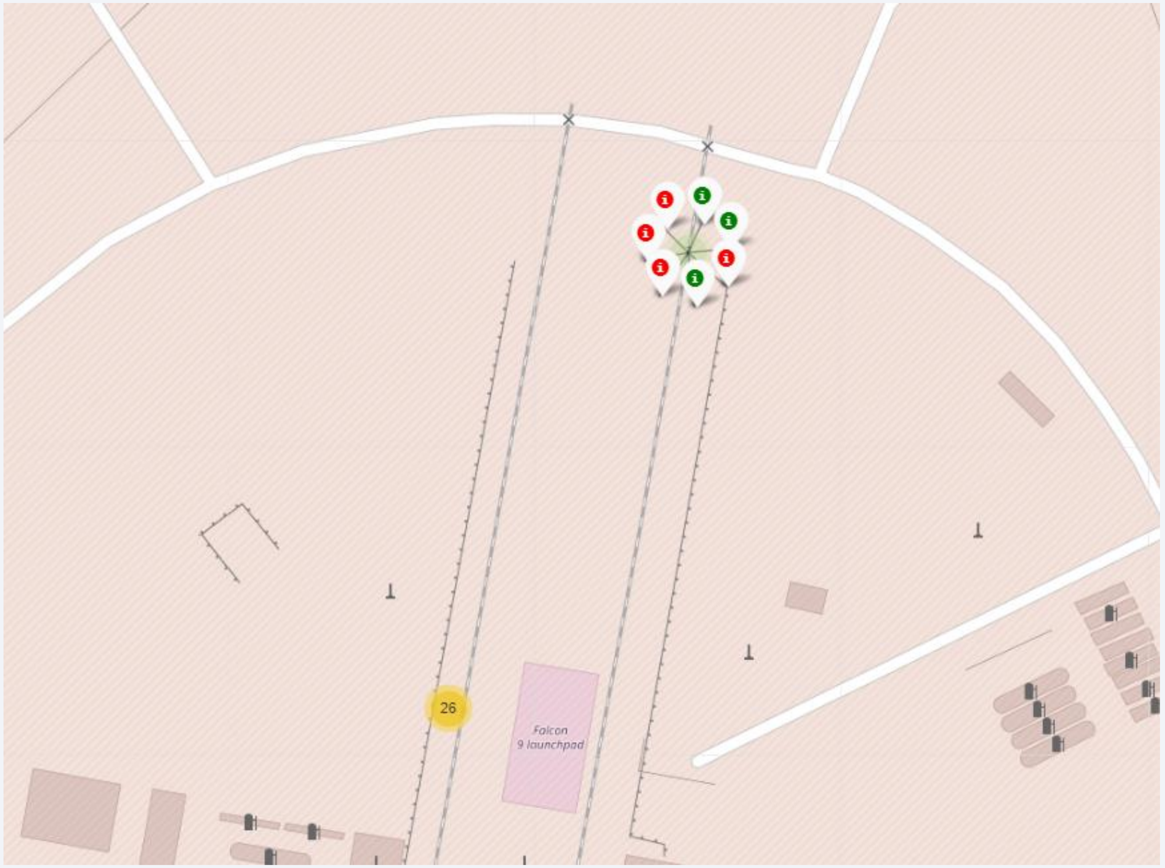
SpaceX Launch Site Locations



- 3 launch sites are in close proximity to each other in Florida
- Last site is on the West Coast in Southern California

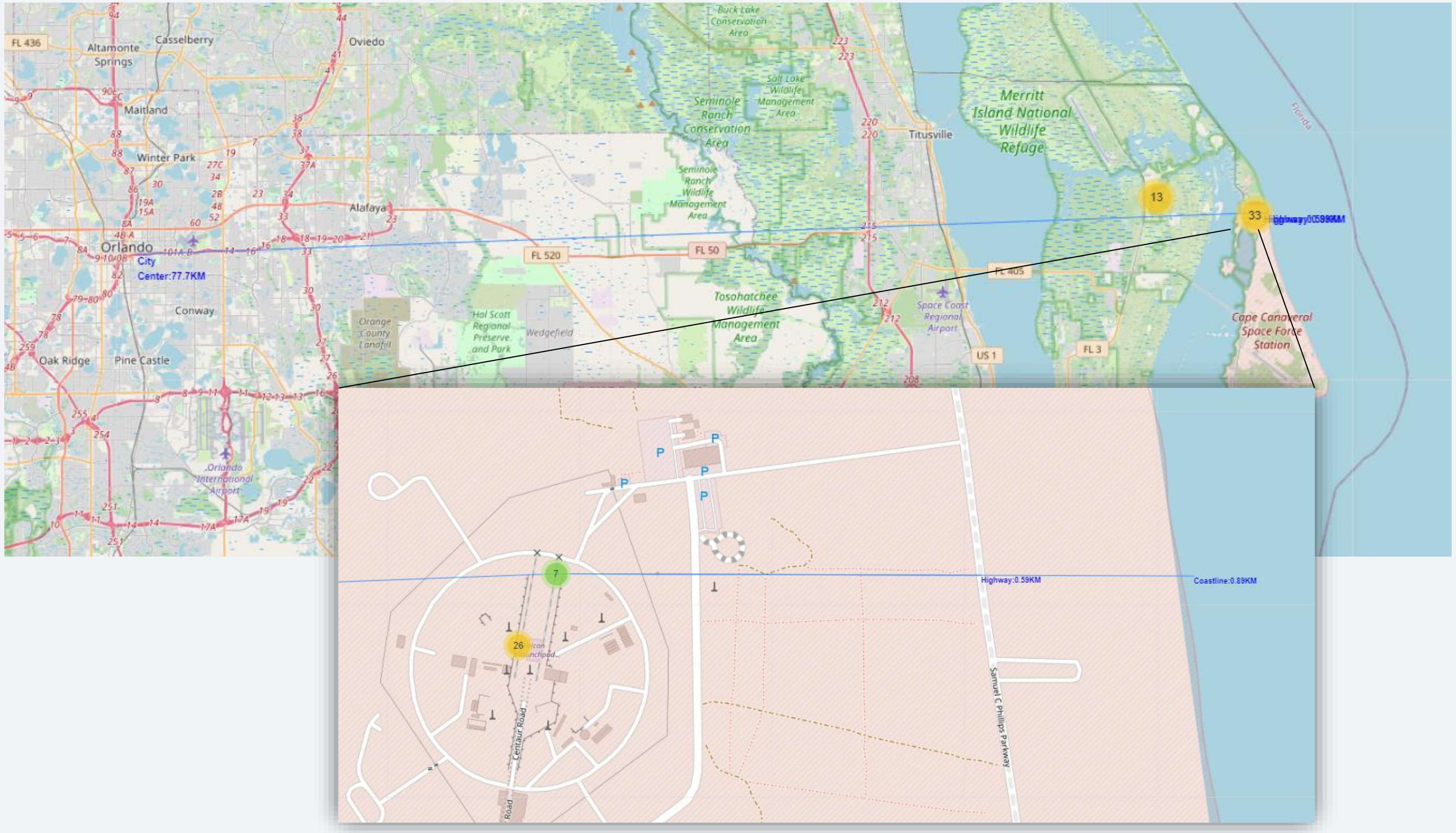


Launch Outcomes by Site



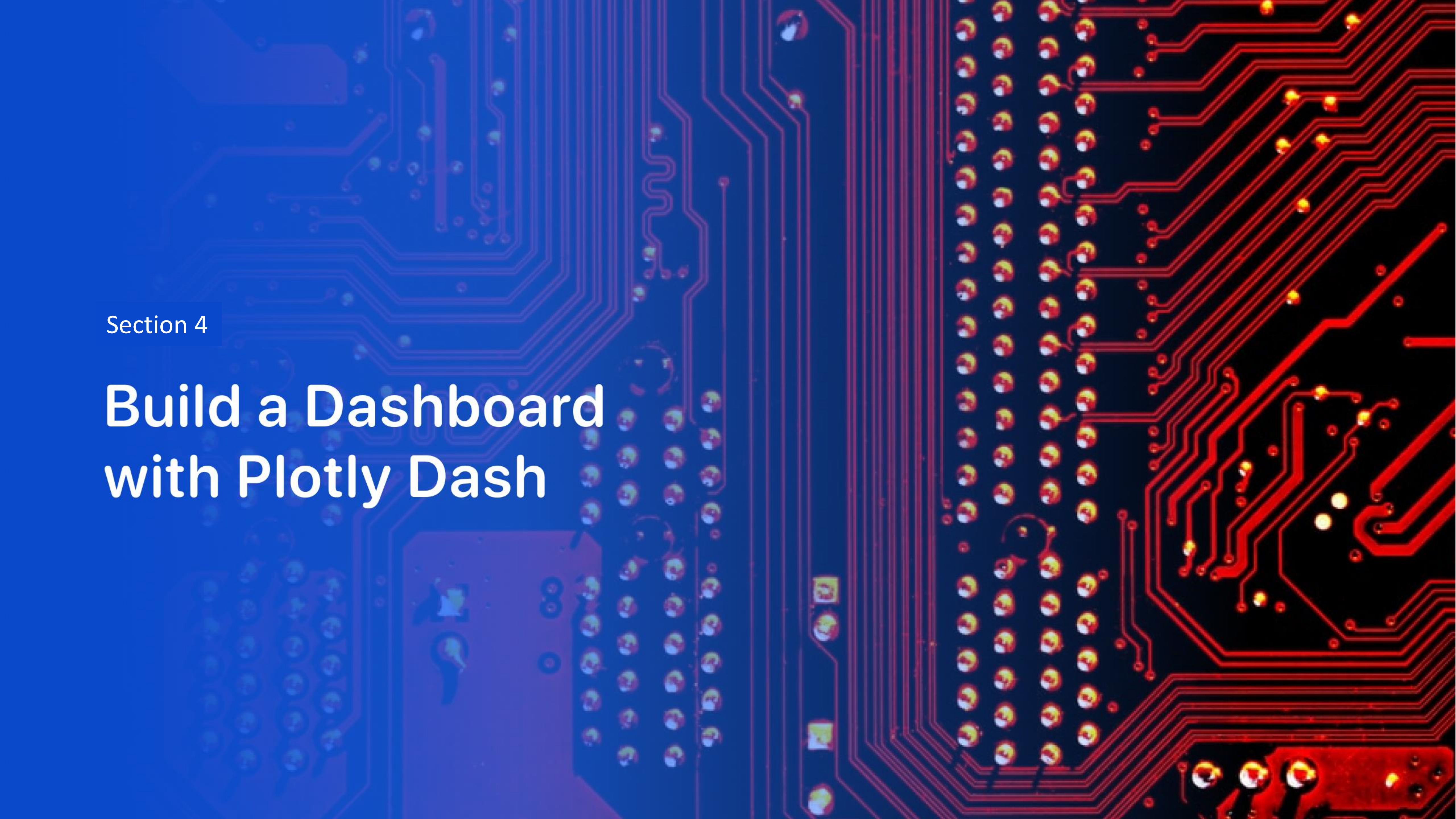
- Showing launch outcomes for launch site CCAFS SLC-40
- Green indicates success, red indicates fail

Proximity Analysis


Orlando
City Center:77.7KM
Titusville
Merritt Island National Wildlife Refuge
Cape Canaveral Space Force Station
Space Coast Regional Airport
Orlando International Airport
Oviedo
Altamonte Springs
Casselberry
Maitland
Winter Park
Alafaya
Conway
Oak Ridge
Pine Castle
Wedgefield
Tosohatchee Wildlife Management Area
Seminole Ranch Conservation Area
FL 520
FL 50
FL 405
FL 3
US 1
13
33
7
26
Highway:0.59KM
Coastline:0.89KM
Centaur Road
Samuel C Phillips Parkway

- Showing proximity to nearest coast, highway, and city for launch site CCAFS SLC-40

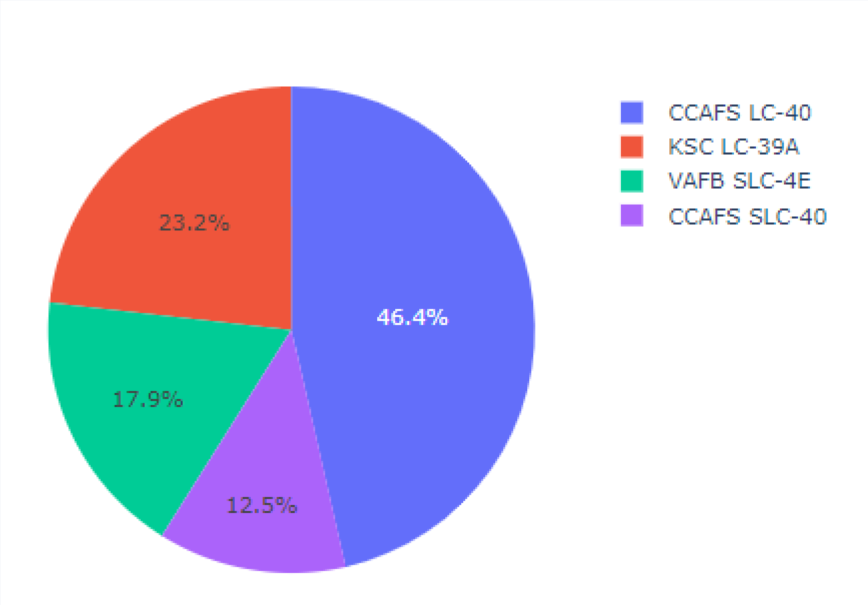
- Launch sites are built near to the coast and a highway, located far from highly populated cities.



Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

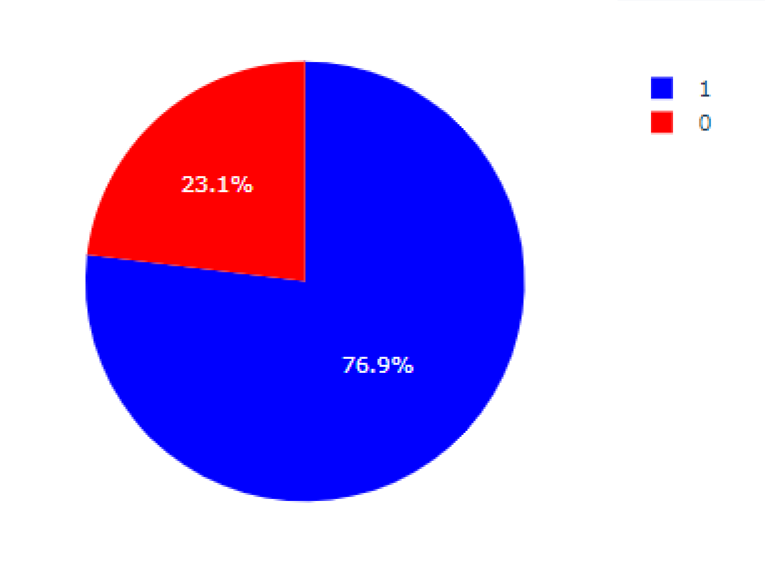


The percentage breakdown of successful launches by site is illustrated by the pie chart.

CCAFS LC-40 has the highest proportion out of all successful launches, followed by KSC LC-39A, VAFB SLC-4E, and CCAFS SLC-40

Launch Success Rate by Site

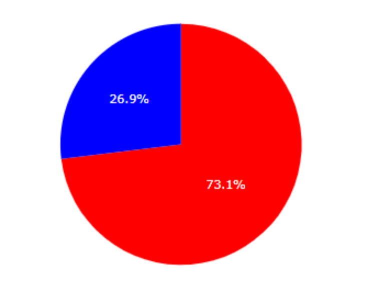
KSC LC-39A



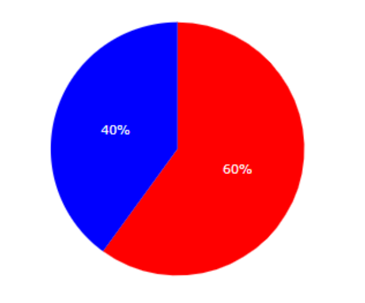
The site with the highest success rate is KSC LC-39A, with 76.9% of their launches being successful.

This is over 30% greater than the next highest success rate of 42.9%.

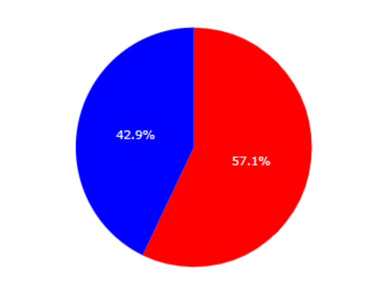
CCAFS LC-40



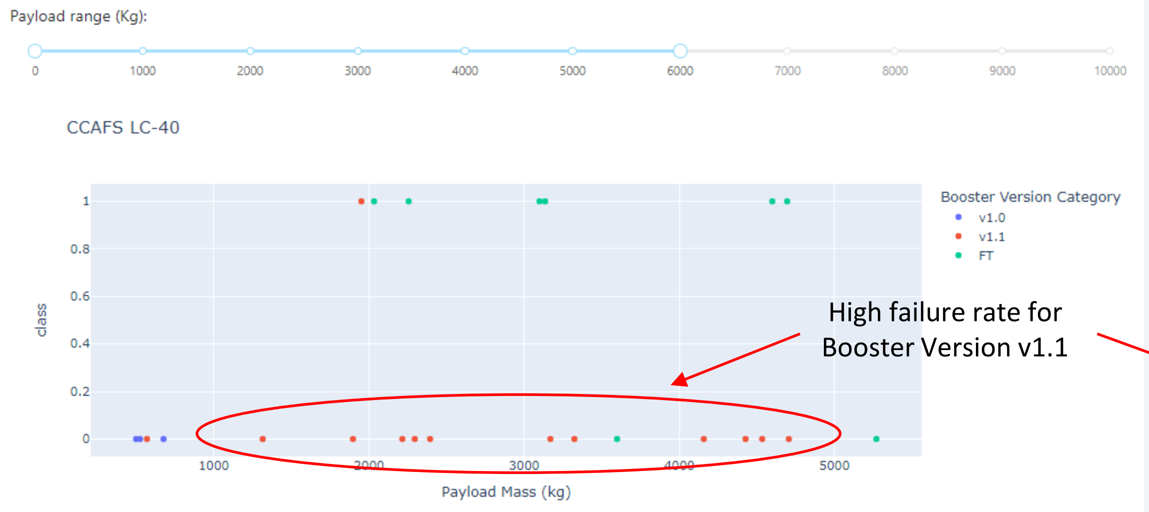
VAFB SLC-4E



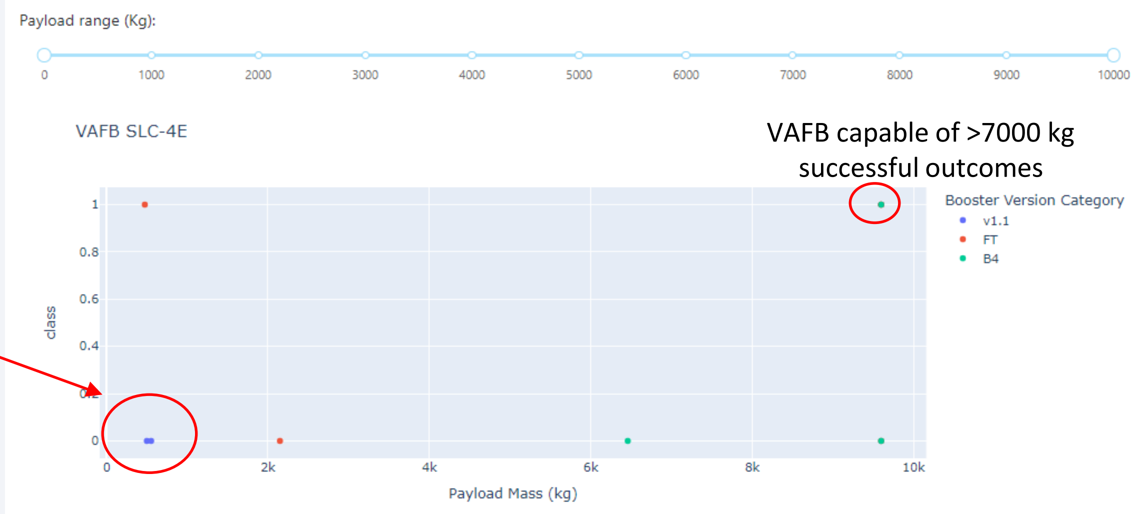
CCAFS SLC-40



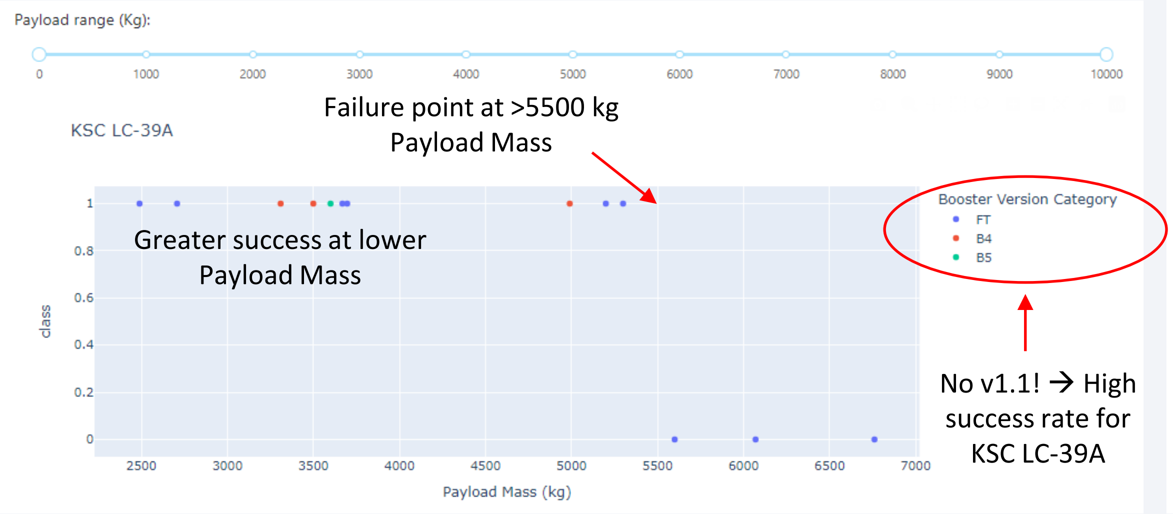
Launch Outcome by Payload Mass, Booster Version, and Site



High failure rate for Booster Version v1.1



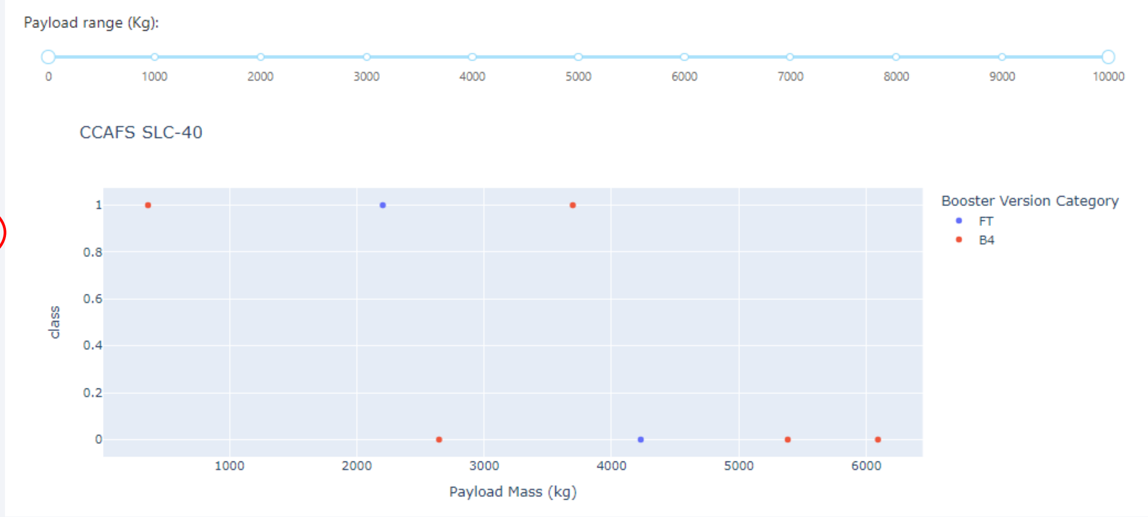
VAFB capable of >7000 kg successful outcomes



Failure point at >5500 kg Payload Mass

Greater success at lower Payload Mass

No v1.1! → High success rate for KSC LC-39A



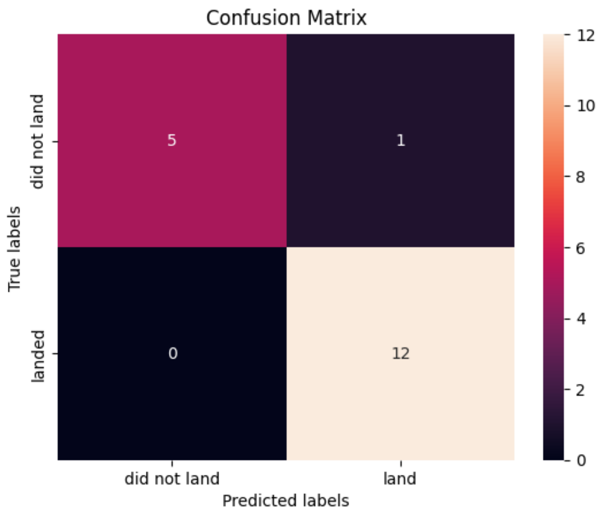
Section 5

Predictive Analysis (Classification)

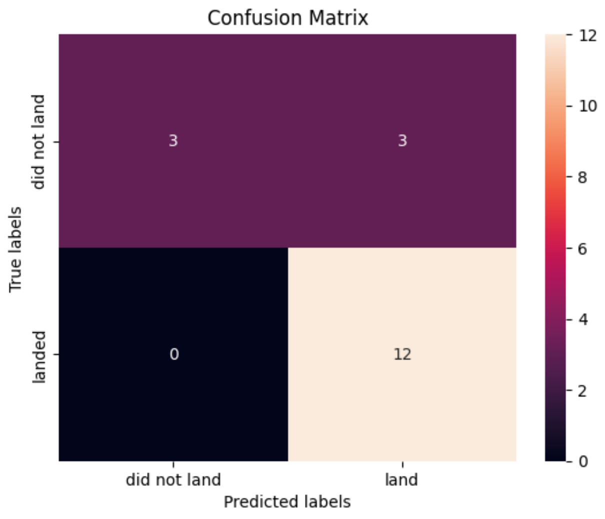
Classification Accuracy and Confusion Matrix

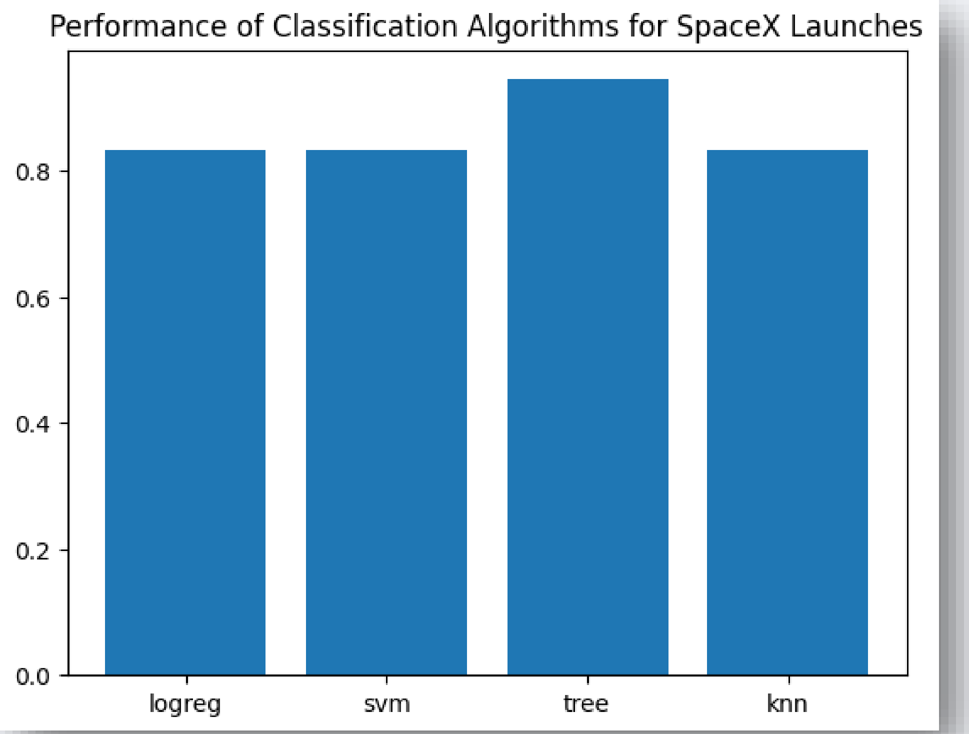
- Tree algorithm provides the best accuracy with the given dataset
- From confusion matrix, only 1 landing was misclassified as landed when it failed
- Other algorithms misclassified 3 failed landings as successes

Decision Tree



Other Algorithms





Conclusions

- Launch Sites are typically located by a highway and ocean coast. Strategically, this allows for quicker supply and transport, and serve as gateways to Asia (Pacific) and Europe (Atlantic)
- The success rate at KSC LC-39A launch site is >30% better than the next best site. However, this may be due to not having the low-success-rate Booster Version v1.1 launching from there.
- Success rate increases year-over-year, except for 2018.
- Decision Tree classification correctly classifies 94.4% of the training data launches, making it the strongest performing prediction model for our data.

Thank you!