IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Laura Roberts
23 Nov 2024

https://github.com/laurahc55/dataScienceCapstone/tree/main

Outline

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

Executive Summary

The data was collected from SpaceX REST API and the Wiki pages about the Falcon 9 launches. We analyze 90 landings in total with 60 successful landings and 30 unsuccessful ones.

Visualization to see the relationships between the different variables were created and some interesting questions were answered by summarizing some of the data using SQL. Maps of the launch sites are included with the number of launches from each site and the outcome, to quickly see the performance of each launch site.

A Decision Tree with an 89% of accuracy on the test data was the best model to predict the outcome of a landing mission. Using this model, we will predict if the Falcon 9 stage will land successfully and determine the cost of a launch to be able to bid against SpaceX for a rocket launch.

Introduction

SpaceX advertises Falcon 9 rocket cost on its website, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. We will use this information to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

Collected data from SpaceX REST API to analyze the Falcon 9 launches and information from the Wiki pages about the Falcon 9 launches was gathered using web scrapping techniques. We included launches from before13th Nov 2020 and with a single core and payload. Missing *payloadMass* values were replaced by the mean. We analyze 90 landings in total with 60 successful landings and 30 unsuccessful ones.

Visualization of the variables to be included in the model were created to understand the relationships between the different variables and some questions were answered by analyzing the data using SQL. The launch sites were included in a map with the number of launches from each site and the outcome to quickly see where the launches happened and the success rate of each site.

The models trained to predict the outcome of a landing were: Logistic regression, Support Vector Machine, Decision Tree and K Nearest Neighbor clustering. We trained thousands of models to find the best parameters for each models and then tested the accuracy using data that those models on the testing data.

Data Collection

Launch data from the SpaceX REST API, which contains:

- rocket used
- payload delivered
- launch specifications
- landing specifications
- landing outcome

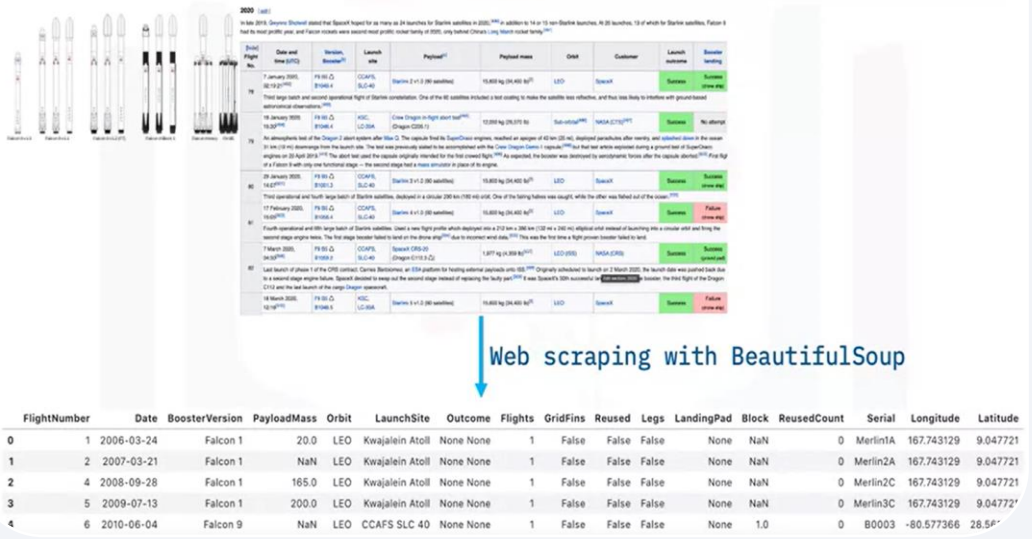
```
url="https://api.spacexdata.com/v4/launches/past"

response =requests.get(url)

response.json()
```

```
response.json()
```


```
[{'fairings': {'reused': False,
  'recovery_attempt': False,
  'recovered': False,
  'ships': []},
 'links': {'patch': {'small': 'https://images2.imgbox.com/3c/0e/T8iJcSN3_o.png',
   'large': 'https://images2.imgbox.com/40/e3/GypSkayF_o.png'},
  'reddit': {'campaign': None,
   'launch': None,
   'media': None,
   'recovery': None},
  'flickr': {'small': [], 'original': []},
  'presskit': None,
  'webcast': 'https://www.youtube.com/watch?v=0a_00nJ_Y88',
  'youtube_id': '0a_00nJ_Y88',
  'article': 'https://www.space.com/2196-spacex-inaugural-falcon-1-rocket-lost-launch.html',
  'wikipedia': 'https://en.wikipedia.org/wiki/DemoSat'},
 'static_fire_date_utc': '2006-03-17T00:00:00.000Z',
 'static_fire_date_unix': 1142553600,
 'tbd': False,
 'net': False,
 'window': 0,
 'rocket': '5e9d0d95eda69955f709d1eb',
 'success': False,
 'details': 'Engine failure at 33 seconds and loss of vehicle',
 'crew': [],
 'ships': [],
 'capsules': [],
 'payloads': ['5eb0e4b5b6c3bb0006eeb1e1'],
 'launchpad': '5e9e4502f5090995de566f86',
 'auto_update': True,
 'failures': [{'time': 33,
   'altitude': None,
   'reason': 'merlin engine failure'}],
 'flight_number': 1,
 'name': 'FalconSat',
 'date_utc': '2006-03-24T22:30:00.000Z',
 'date_unix': 1143239400,
 'date_local': '2006-03-25T10:30:00+12:00',
 'date_precision': 'hour',
 'upcoming': False,
 'cores': [{'core': '5e9e289df35918033d3b2623',
   'flight': 1,
   'gridfins': False,
   'legs': False,
   'reused': False,
   'landing_attempt': False,
   'landing_success': None,
   'landing_type': None,
   'landpad': None}],
 'id': '5eb87cd9ffd86e000604b32a'},
```


Wiki pages about the Falcon 9 launches

Web scraping Falcon 9 Launch records



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.04772
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.56

Data Collection – SpaceX API

Set up
- Import required libraries
- Define functions to collect the data

Import data
- Request rocket launch data from SpaceX API
- Convert the Json result into a dataframe

Clean data
- Only keep the colunms: rocket, payloads, launchpad, cores, flight numbers and date
- Remove rows with multiple cores because those are falcon rockets with 2 extra rocket boosters and rows that have multiple payloads in a single rocket
- Extract single value of payloads and cores
- Convert the date_utc to a datetime and restrict the dates of the launches to be before Nov 13, 2020

Filter data
- Only include Falcon 9 launches
- Reset the flight number column to only account for selected rows

Handle missing values
- Input the mean PayloadMass into the rows with missing values in the PayloadMass column

https://github.com/laurahc55/dataScienceCapstone/blob/main/1-Spacex-Data_collection_API.ipynb

Data Collection - Scraping

Set up

- Import required libraries
- Define functions to process web scraped HTML table

Import data

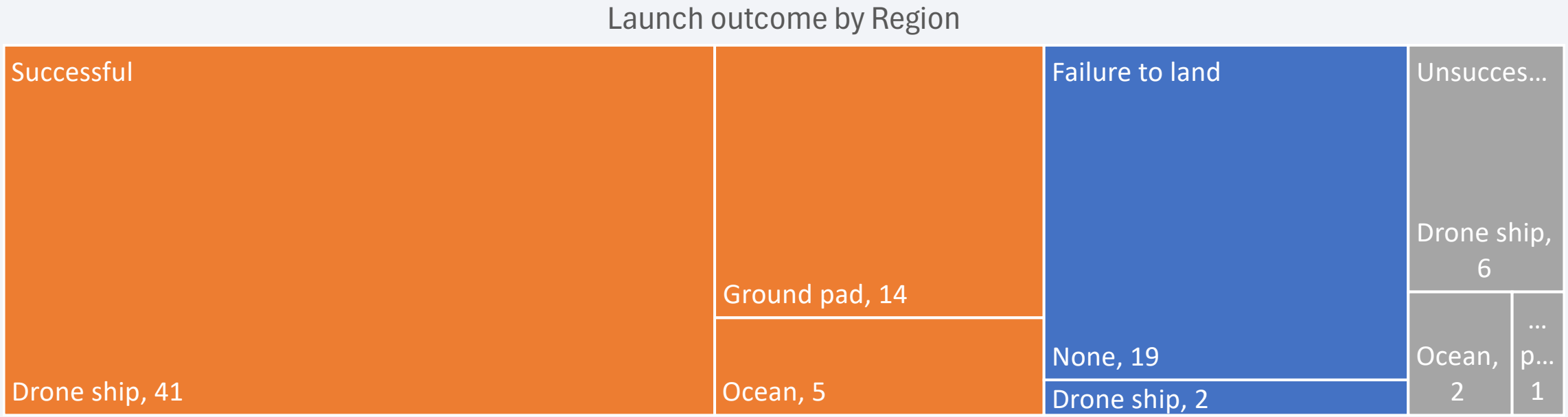
- Scrape the data from a snapshot of the Wikipage updated on 9th June 2021
- Perform an HTTP GET method to request the Falcon9 Launch HTML page
- Create a object from the HTML
- Create a BeautifulSoup object from a response text content

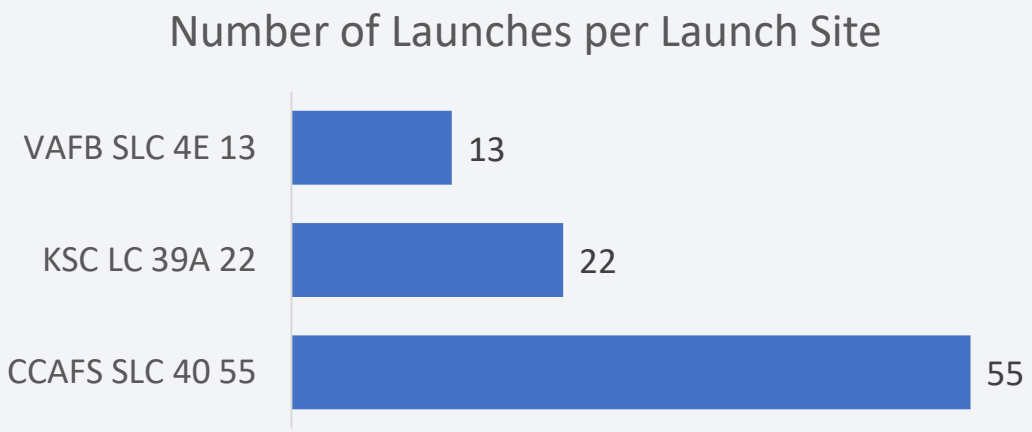
Extract data

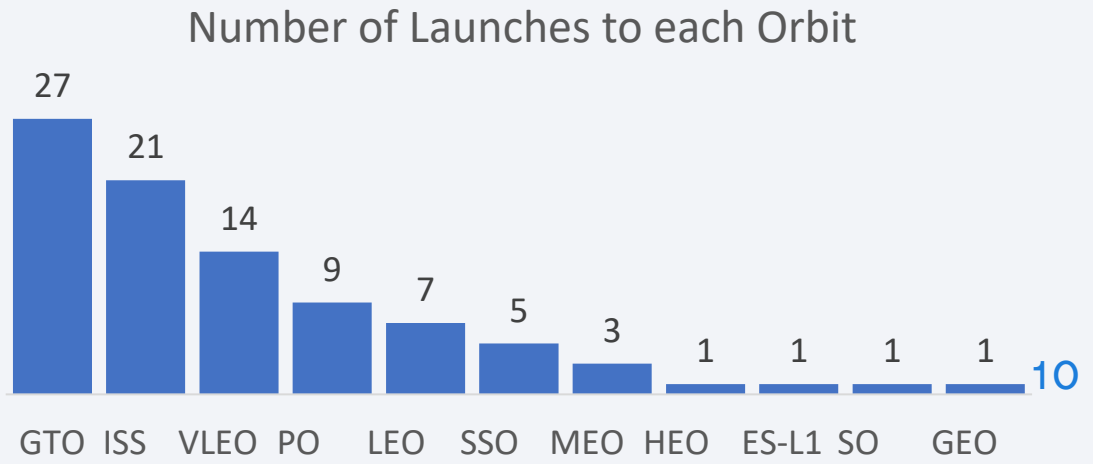
- Find column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

https://github.com/laurahc55/dataScienceCapstone/blob/main/2-Webscraping.ipynb

Data Wrangling







https://github.com/laurahc55/dataScienceCapstone/blob/main/3-Spacex-Data_wrangling.ipynb

EDA with Data Visualization

We created various charts to quickly understand the success rate by, payload mass, orbit, launch site and year and noticed some patterns that might be able to help predict the outcome of a mission. These graphs showed some interesting findings that were highlighted in those specific sections.

https://github.com/laurahc55/dataScienceCapstone/blob/main/4-Exploratory_data_analysis-Data_visualization.ipynb

EDA with SQL, we found:

- The names of the unique launch sites in the space mission
- 5 records from launch sites that begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved
- The boosters that have successfully landed a rocket with a payload mass between 4000 and 6000 in a drone ship
- The total number of successful and failure mission outcomes
- The booster versions which have carried rockets with the maximum payload mass
- The month, booster versions and, launch site of the missions that fail to land in a drone ship in 2015
- The frequency of each lading outcome from 2010-06-04 and 2017-03-20, listing the most frequent first

https://github.com/laurahc55/dataScienceCapstone/blob/main/5-Exploratory_data_analysis-with_SQLeda-sql-coursera_sqllite.ipynb

Interactive Map with Folium

Maps with the location of each launch site, the number of launches from each site and the outcome of the mission represented by the color (green meaning successful and red meaning unsuccessful). Lines from each launch site to the closest city, railroad, highway and coastline including the distance in km.

https://github.com/laurahc55/dataScienceCapstone/blob/main/6-Visual_Analytics_with_Folium.ipynb

Dashboard with Plotly Dash

A dashboard was created to visualize and compare the successful launches from all the launch sites and the proportion of successful launches vs unsuccessful launches for each site.

We also created a dynamic graph that allows us to see the relationship between payload and the outcome of the mission and what booster version was used for each one. Including a slider to select the range of the payload mass we want to see, and we can also filter by launch site.

https://github.com/laurahc55/dataScienceCapstone/blob/main/7-Dashboard.py

Predictive Analysis (Classification)

First, we divided our data into 72 observations for training and 18 observation for testing our results.

We trained and selected the best model for each technique out of:

- 30 Logistic Regression
- 1,250 Support Vector Machine
- 3,240 Decision Trees
- 800 KNN clustering

To evaluate the accuracy of the four best models we used our test data set and selected the model with the highest accuracy on the test data.

https://github.com/laurahc55/dataScienceCapstone/blob/main/8-SpaceX_Machine_Learning_Prediction.ipynb

Results

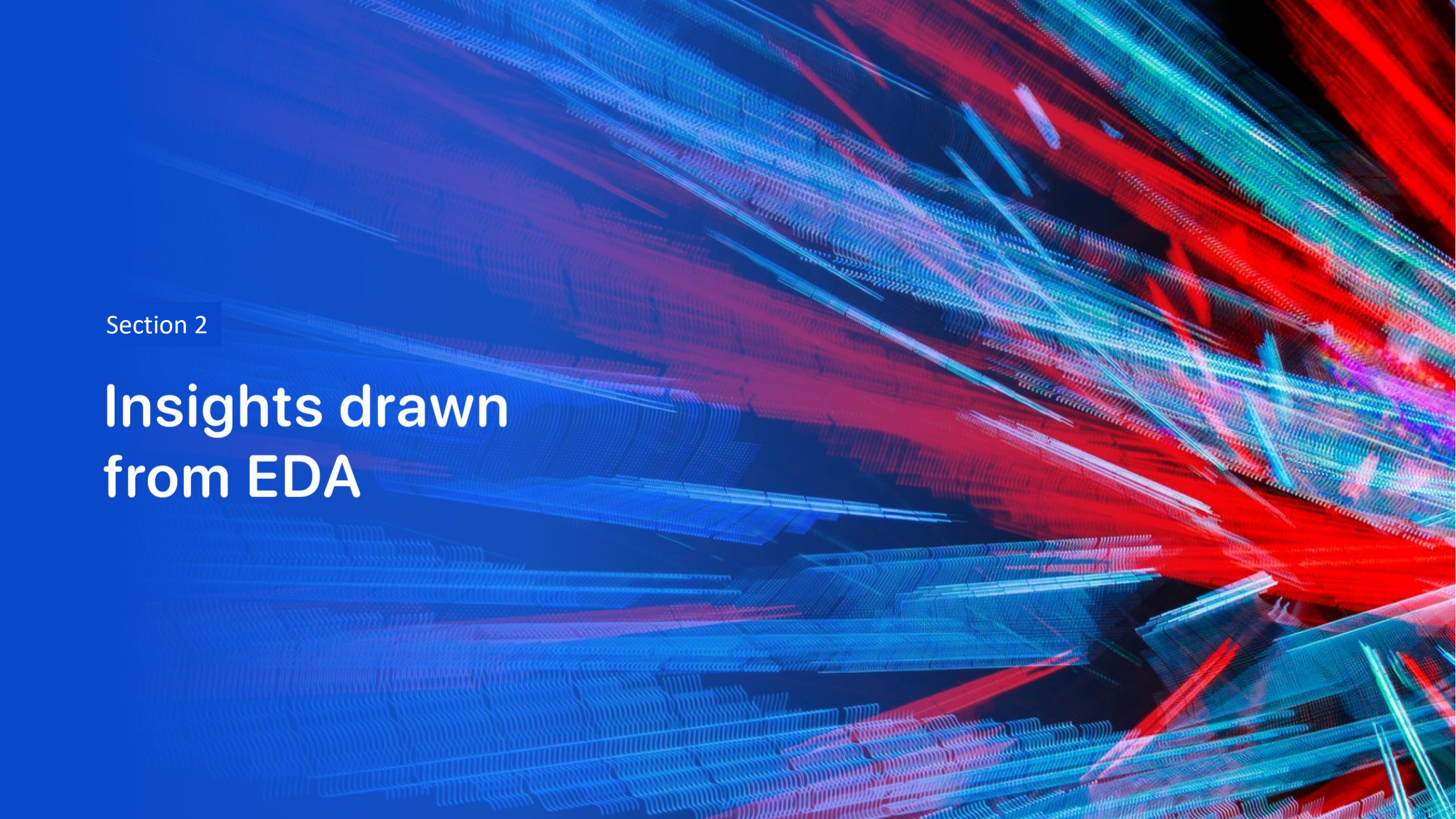
The selected model was a decision tree with an 89% accuracy, correctly predicting the landing outcome of 16 out of the 18 test cases.

The variables used in the model are:

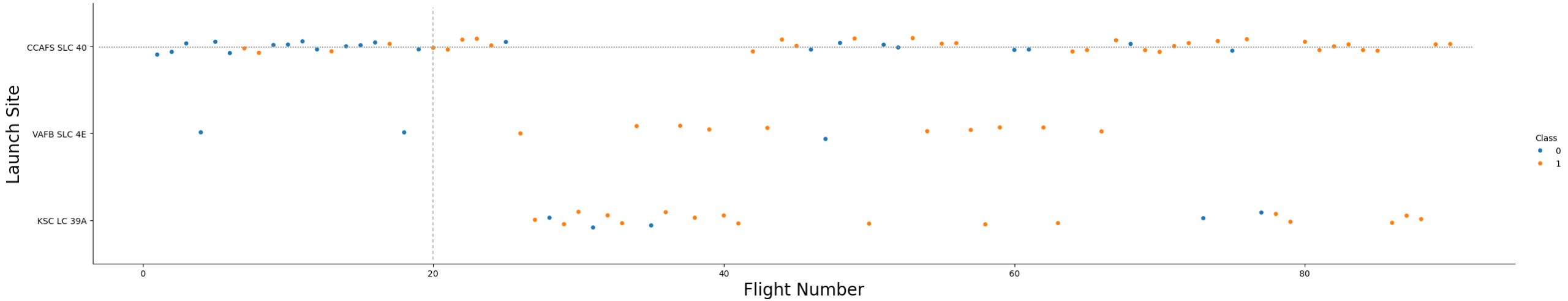
> Payload Mass, Orbit, Flights, Blocks, Reused, Legs, Grid fins, Booster type, Launch Site, Serial and , Landing Pad

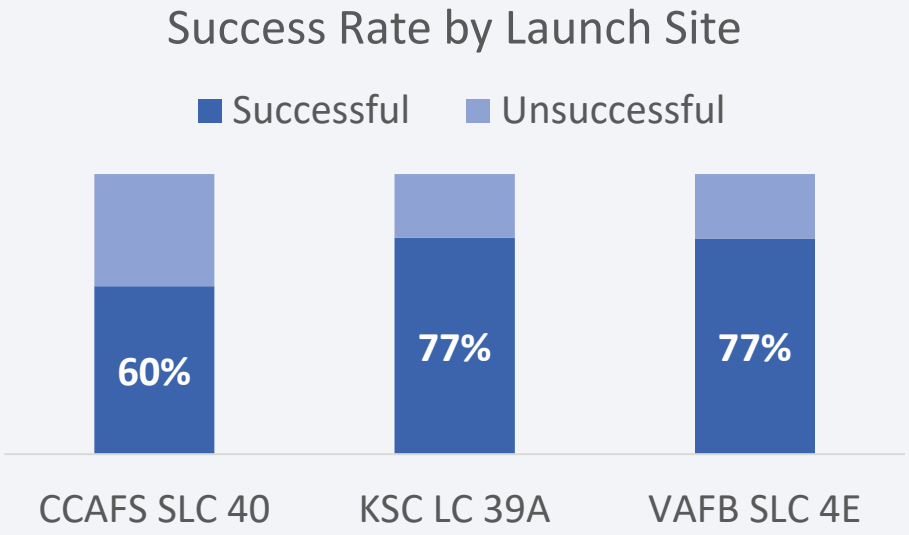
The success rate of the last 70 is more than 3 times higher than the success rate of the first 20 launches. Which also means the success rate of a mission has been much higher in the recent years.

The success rate to the orbit VLEO (86%) is the highest of the three orbits with the most frequent launches.


Section 2
Insights drawn from EDA

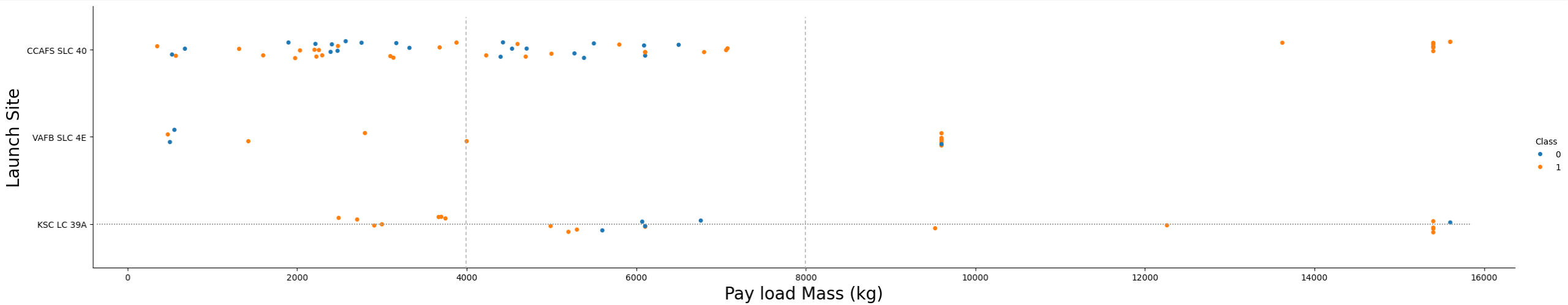
Flight Number vs. Launch Site

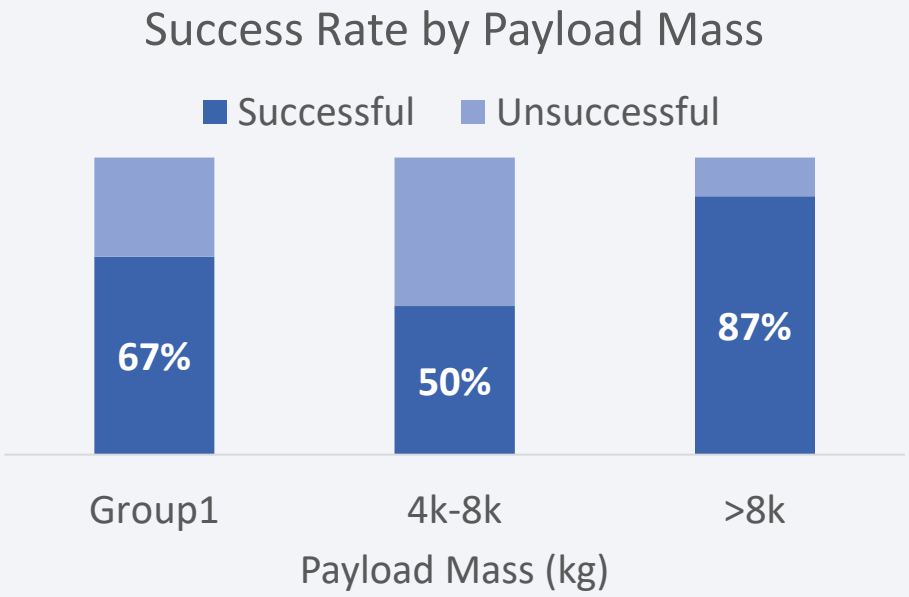




- The success rate for the first 20 launches was only 25% compared to 80% for the last 70 launches
- The CCAFS SLC 40 station has the lowest success rate at about 60% and like the previous finding the success rate of the first 18 missions from this site only 28% compared to a 76% success rate on the last 37 missions

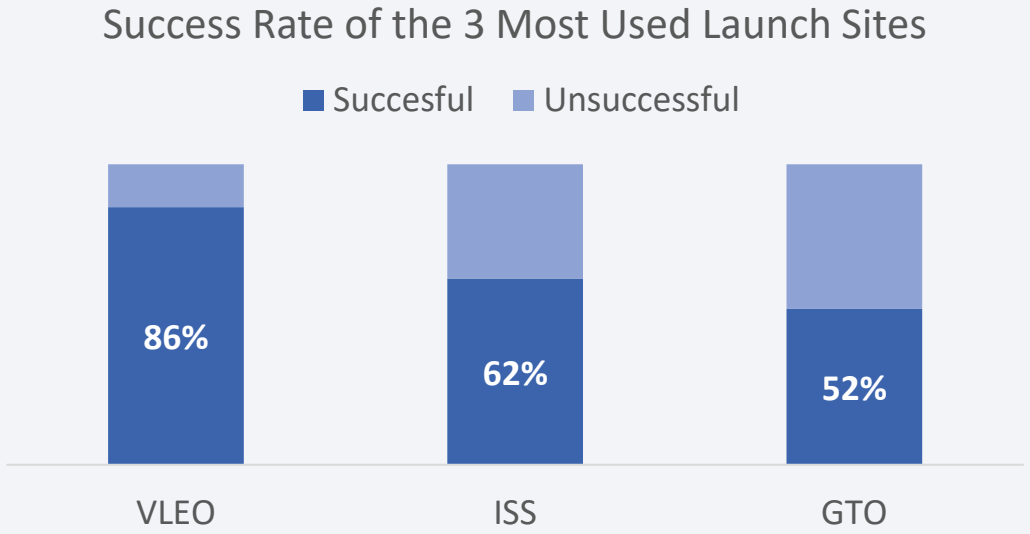
Payload vs. Launch Site

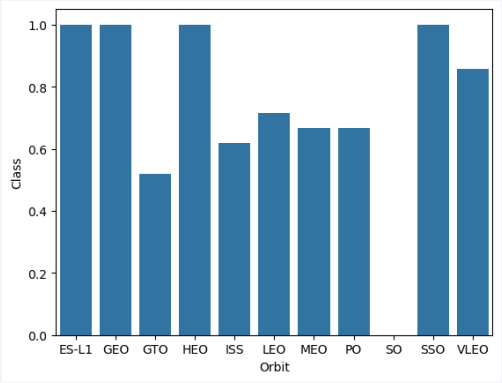


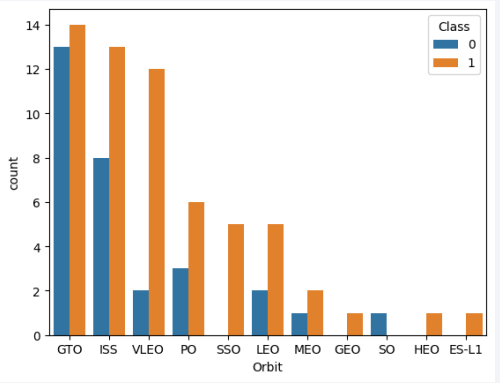


- The heavier (>8k kg) payloads have a higher success rate of 87% than the medium (4k-8k kg) with only 50%.
- The launch station KSC LC 39A have 4 out of 5 unsuccessful launches from rockets with medium payload mass
- The launch station VAFB SLC 4E have only launched light (<=4k kg) and heavy rockets (>8k kg)

Success Rate vs. Orbit Type

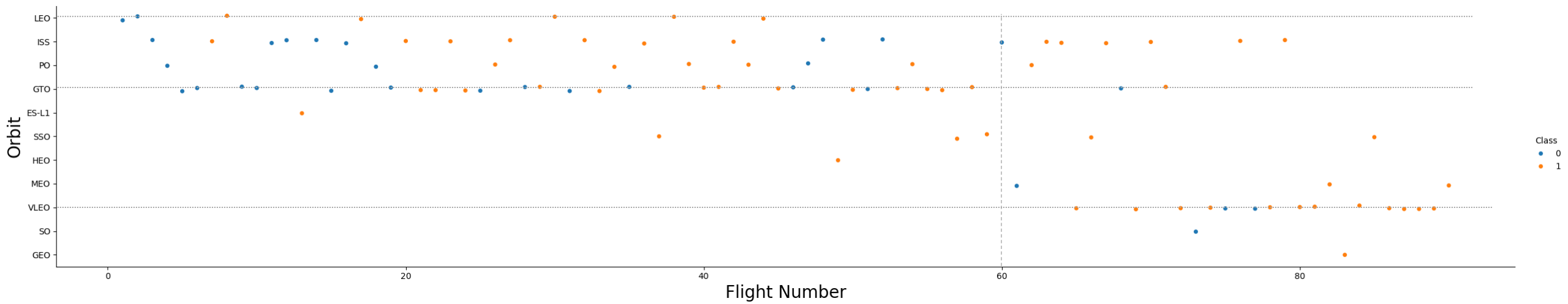






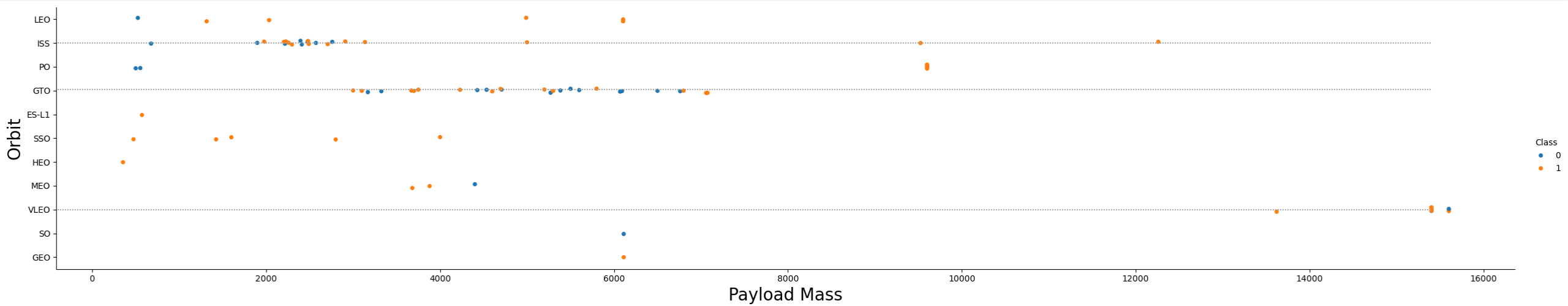
- Of the three launch stations that account for ~70% (62) of the launches, VLEO has the highest success rate of 86% followed by ISS with 62% and GTO has the lowest with 52%
- All the launch to ES-L1 (1), GEO (1), HEO (1) and SSO (5) have being successful but those only account for ~9%. launches
- The only launch to the SO was unsuccessful

Flight Number vs. Orbit



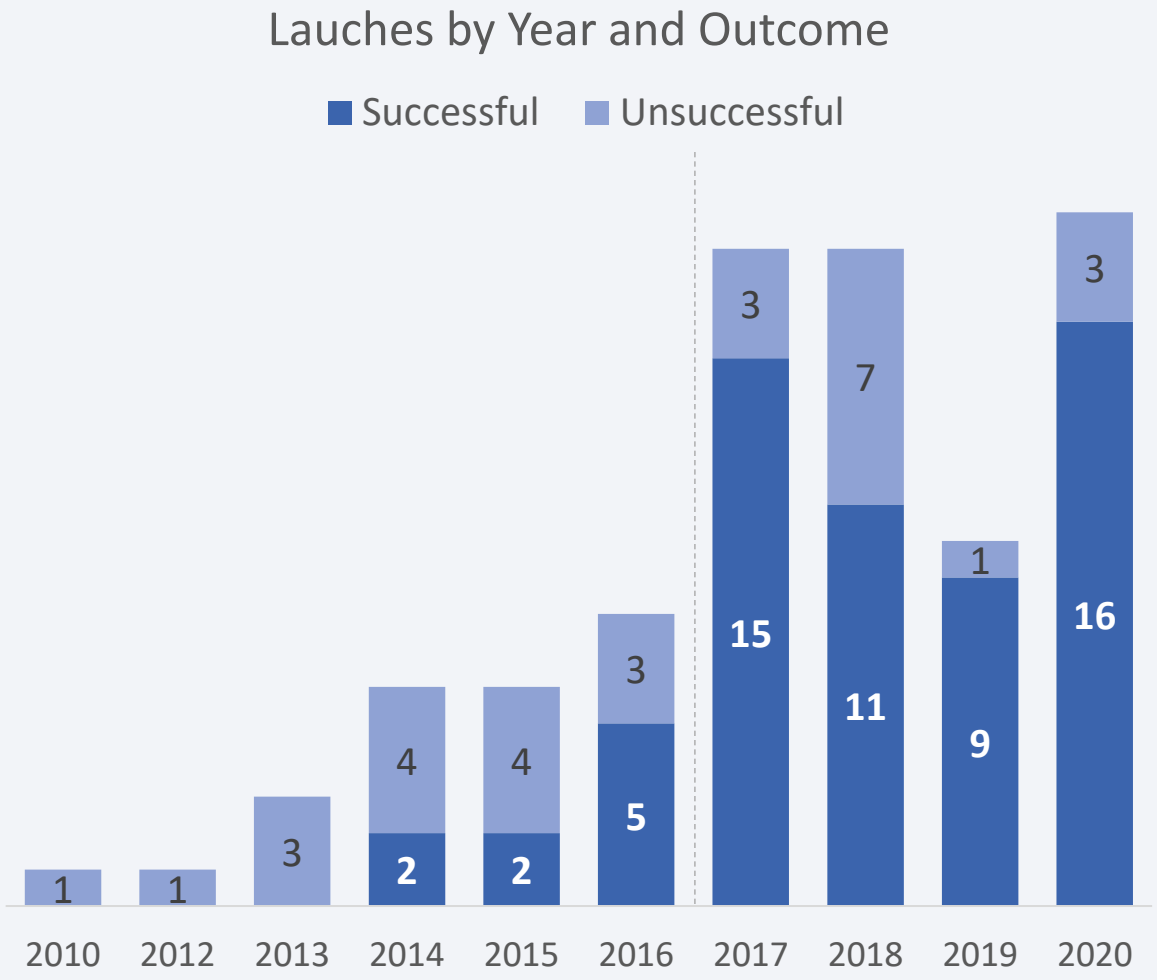
- There hasn't been a launch to the LEO orbit since the 44th flight
- The last 6 launches to the ISS orbit have been successful
- The launches to the GTO orbit went from 25 out of the first 60 (~40%) to only 2 out of the last 30 (~7%)
- Almost half of the last 30 launches were to the VLEO orbit

Payload vs. Orbit Type

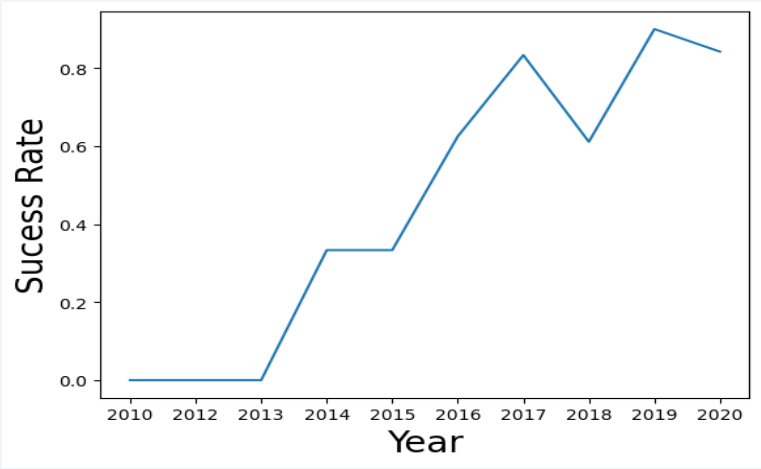


- ISS orbit: Most (~60%) of the unsuccessful launches of the rockets with light payload mass (4k kg) were to this orbit
- GTO orbit: ~80% of the unsuccessful launches of the rockets with medium payload mass (4-8k kg) were to this orbit vs only ~15% of the unsuccessful launches of the rockets with light payload mass (4k kg)
- VLEO orbit: all the rockets launch to this orbit have the heaviest payload mass (>8k kg) and the success rate of ~85%

Launch Success Yearly Trend



- The success rate since 2017 have being >60%
- In 2019 we saw the highest success rate of 90%
- ~72% of launches happened from 2017 to 2020



All Launch Site Names

The four launch sites for the Falcon 9 missions are:

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

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select distinct "Launch_Site" from SPACEXTBL
```

```
 * 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'

Here is a sample of some of the launches made from sites the begin with CCA

Sites like:

- CCAFS LC 40
- CCAFS SLC 40

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL 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

NASA (CRS) has launched a total of 45,596 kg of payload mass.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(PAYLOAD_MASS__KG_)  from SPACEXTBL where "Customer" = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The booster version F9 v1.1 carries 2,928.4 kg of payload mass on average.

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_)  from SPACEXTBL where "Booster_Version" = "F9 v1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The first successful ground landing was on the 22nd of December 2015

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min("Date") from SPACEXTBL where "Landing_Outcome" like "%ground%"
```

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

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The booster version that have successful landed drone ship landing with a payload between 4k-6k are:

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

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select distinct "Booster_Version" from SPACEXTBL where "Landing_Outcome" like "%Success (drone ship)%" and PAYLOAD_MAS
```

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

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

Total Number of Successful and Failure Mission Outcomes

- 61 of the 101 missions had a successful outcome
- 30 of the 101 missions did not land
- 10 of the 101 missions had an unsuccessful outcome

Task 7

List the total number of successful and failure mission outcomes

```
%sql select case when "Landing_Outcome" like "%Success%" then "Success" when "Landing_Outcome" like "%Failure%" then "Failu
```

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

Outcome	count(*)
None	30
Failure	10
Success	61

Boosters Carried Maximum Payload

The booster version that have carried the maximum payload mass are:

- 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

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select "Booster_Version" from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

In 2015 there were two launches, both failed to land on a drone ship to the CCAFS LC-40 launch site; one in January, booster version F9 v1.1 B1012 and one in April, booster version F9 v1.1 B1015.

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select case when substr(Date, 6,2) = "01" then 'January' when substr(Date, 6,2) = "04" then 'April'end  as month_name,
```

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

month_name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Half of the rockets from Jun 4th,2010 to Mar 20th ,2017 did not attempt a landing

Task 10

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.

```
%sql select "Landing_Outcome", count(*) from SPACEXTBL where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_
```

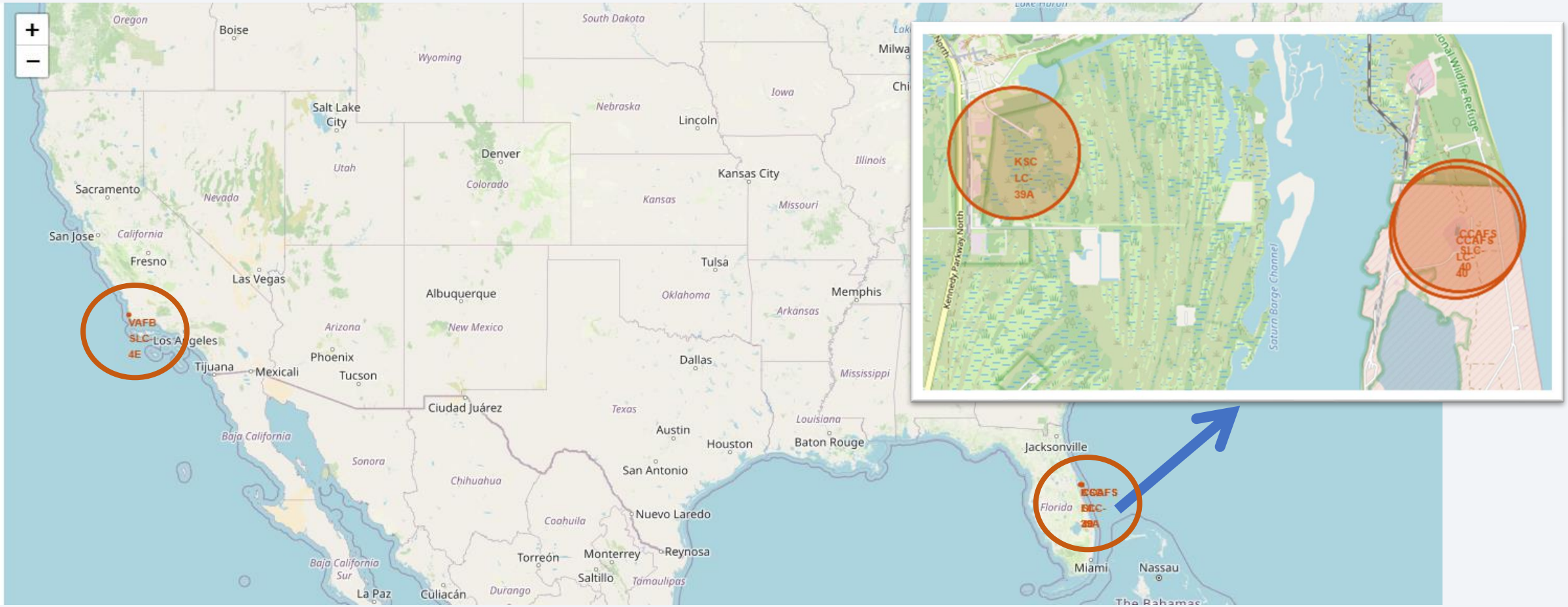
```
* sqlite:///my_data1.db
Done.
```

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

Section 3

Launch Sites Proximities Analysis

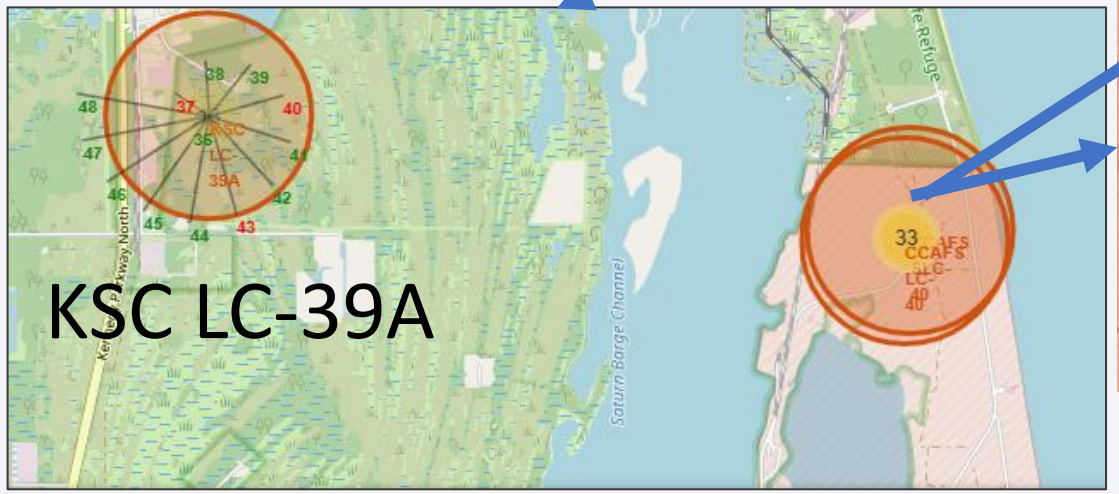
Launch site location



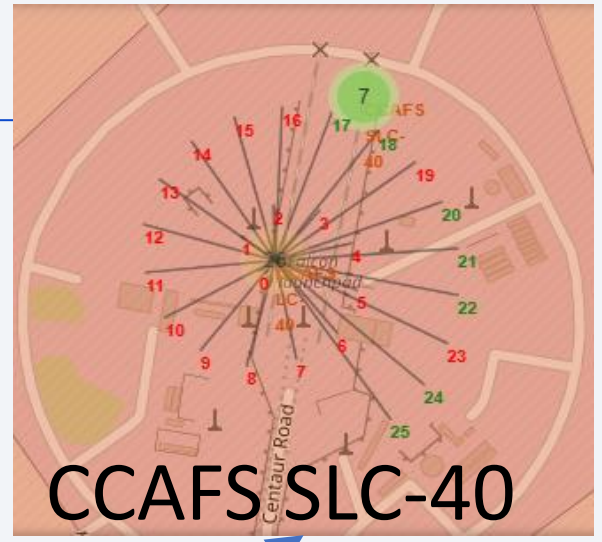
Only one of the four launch sites is in California, the rest of the launch sites are in Florida. With the CCAFS launch sites being within a mile from each other

Outcome of the launch per landing site

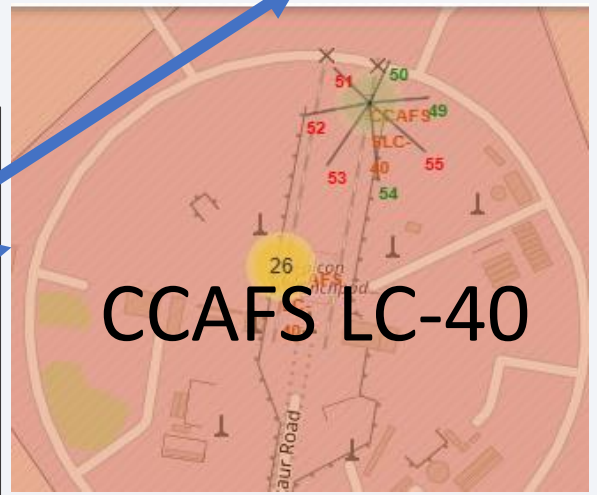
VAFB SLC-4E



KSC LC-39A



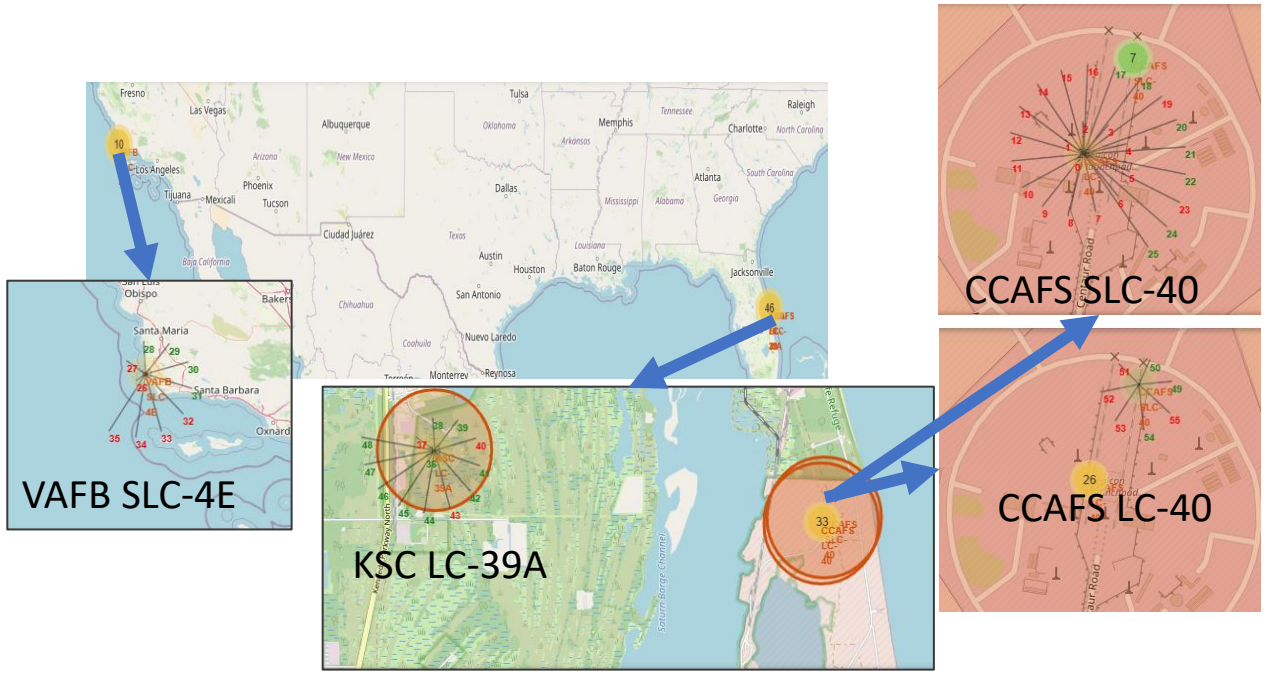
CCAFS SLC-40



CCAFS LC-40

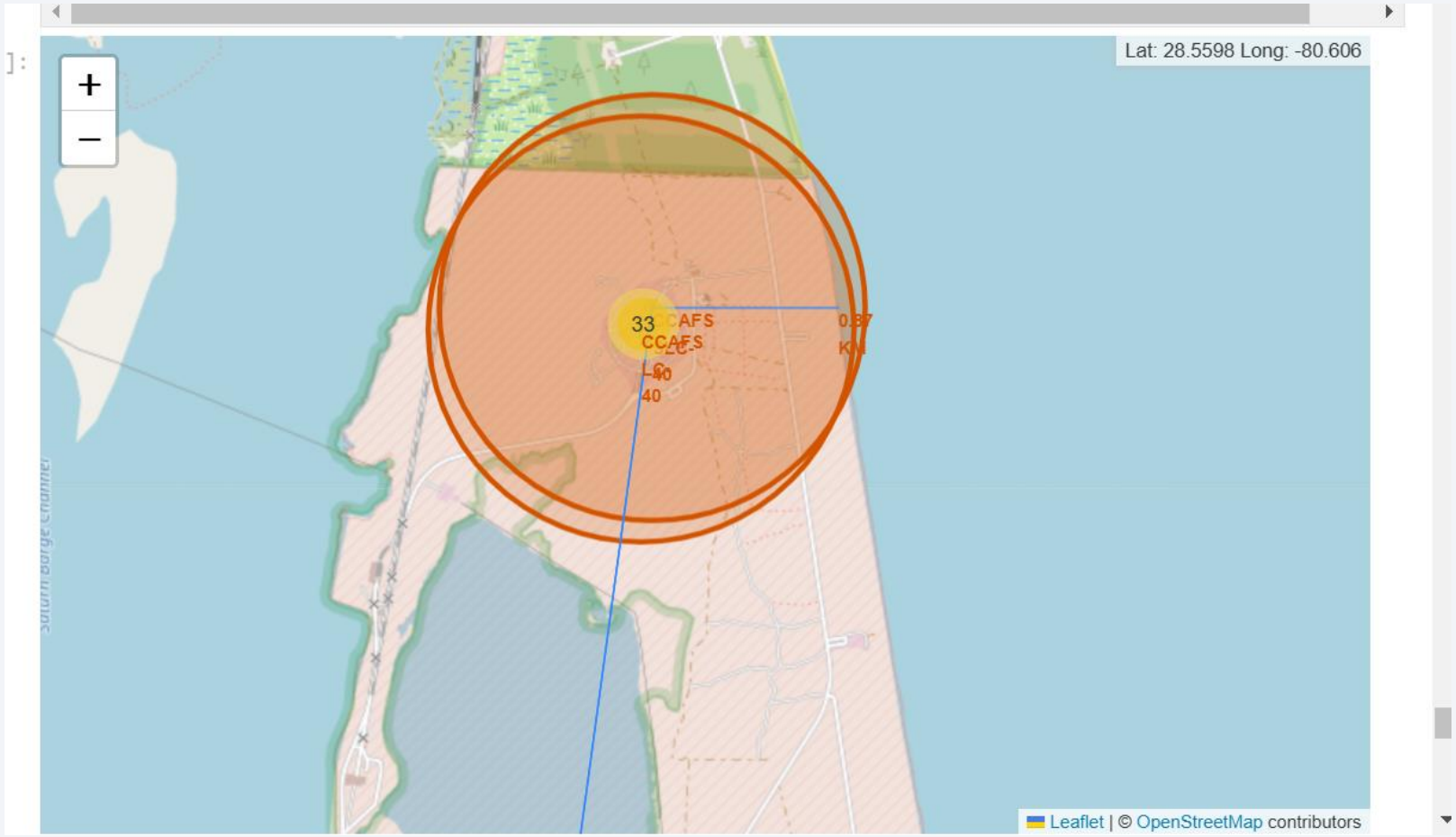
Outcome of the launch per landing site

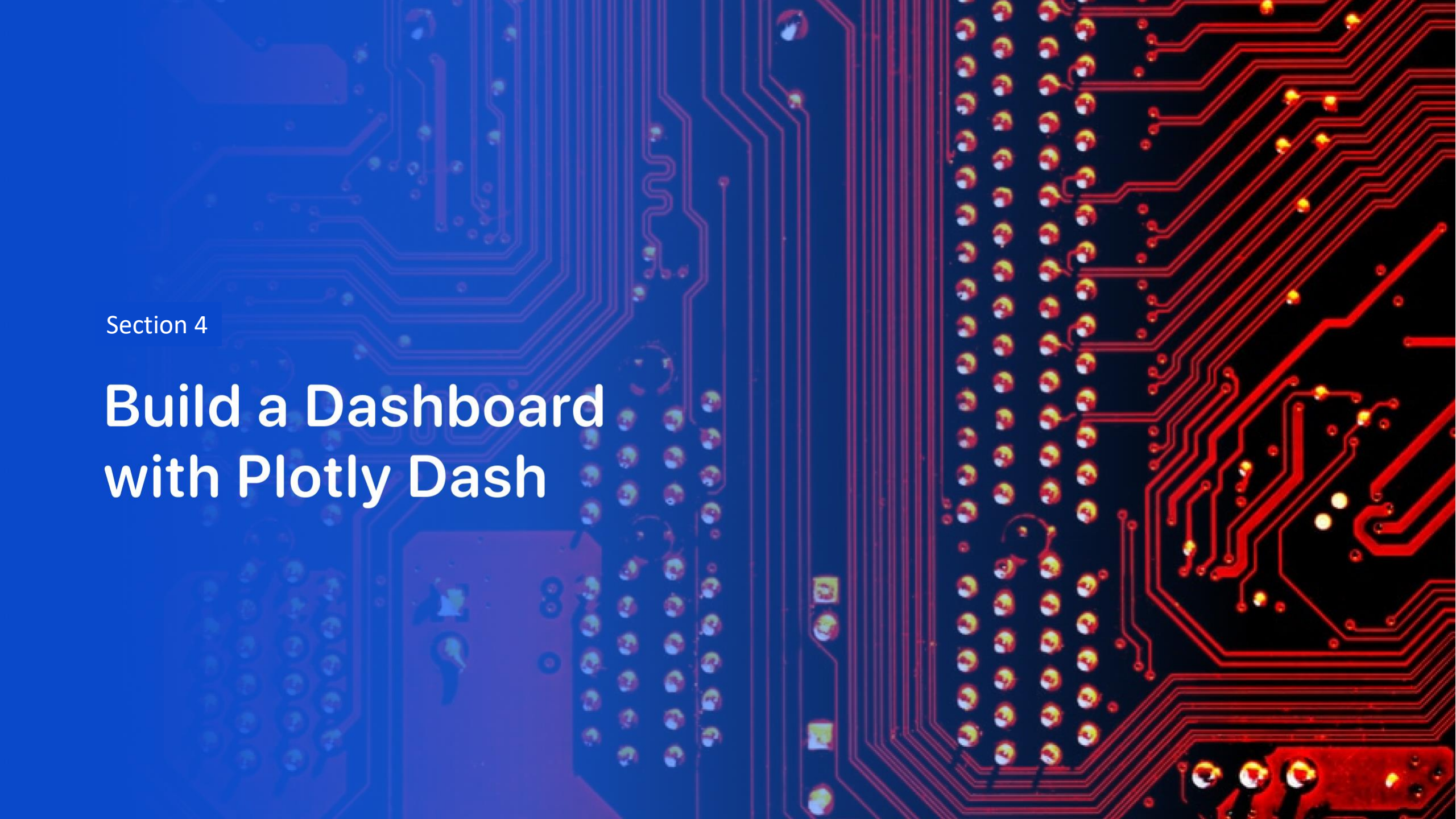
- KSC LC-39A has the highest proportion of successful landings
- CCAFS LC-40 only have 7 landings with only 3 successful ones
- 4 out of the 10 landings in the landing site VAFB SLC-4E were successful
- CCAFS SLC-40 is the landing site with the most landing but only 7 have being successful



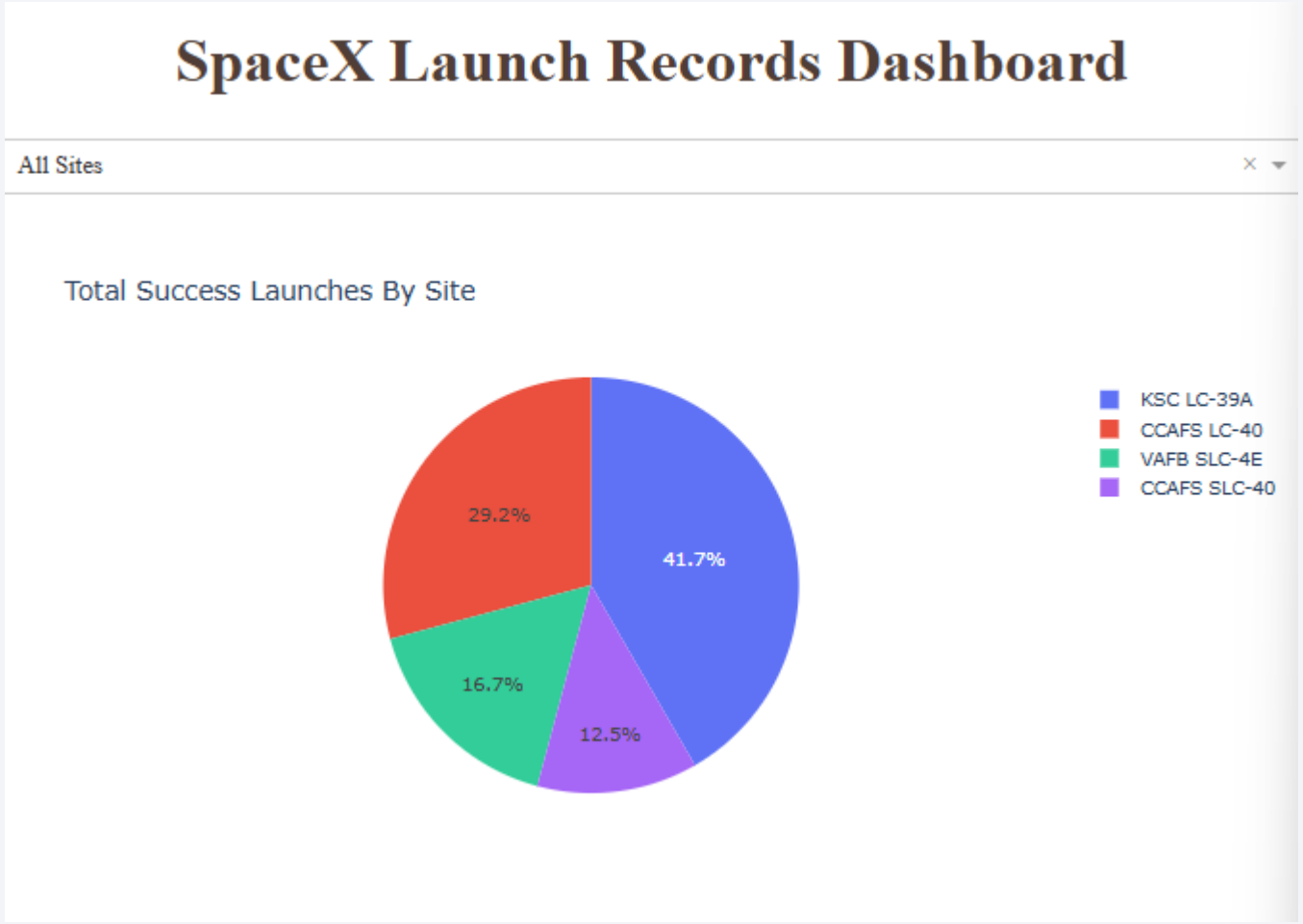
Launch site proximities to railway, highway and, coastline

Launch sites tend to be close to a railway, highway and the coastline but not too close to cities



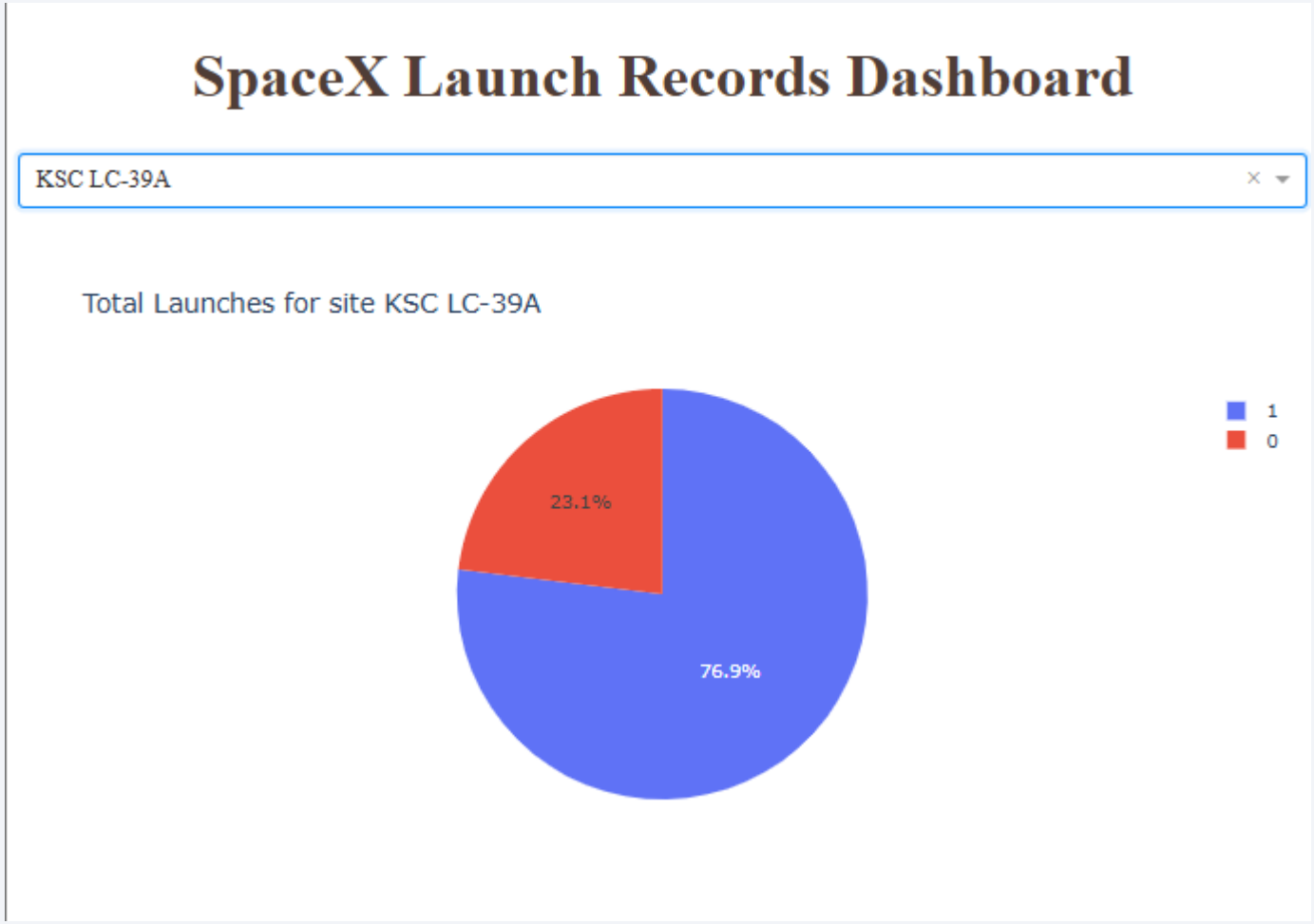

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



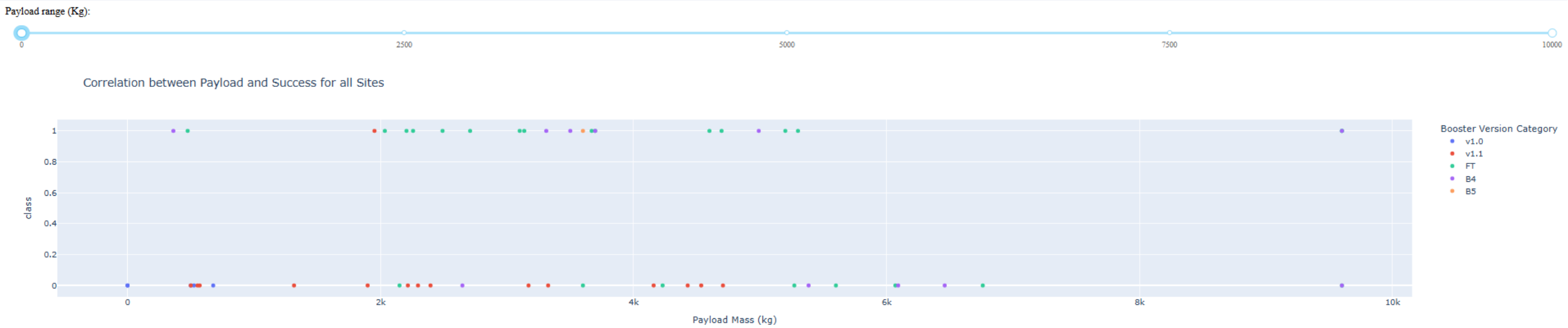
- 41.7% of the successful launches are from the launch site KSC LC-39A
- 29.2% of the successful launches are from the launch site CCAFS LC-40

Launch site with highest launch success ratio



- The launch site KSC LC-39A has the highest launch success ratio with 76.9% of launches being successful

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

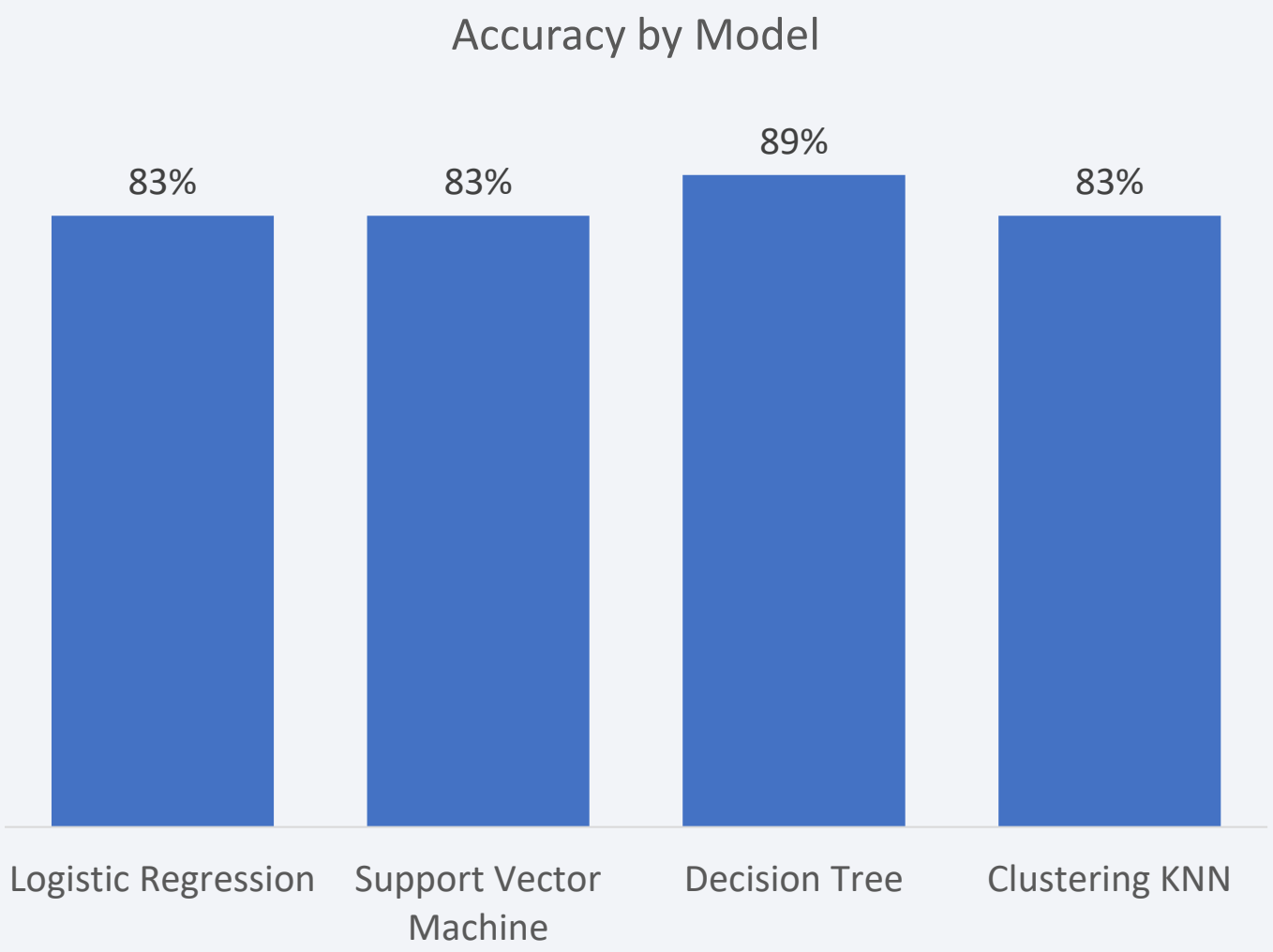


- Most launches have a payload mass of 2k-6k kg.
- The category booster version of most of the rockets with the heaviest payload is B4
- The rockets with the booster version category v1.1 tend to have an unsuccessful mission outcome

Section 5

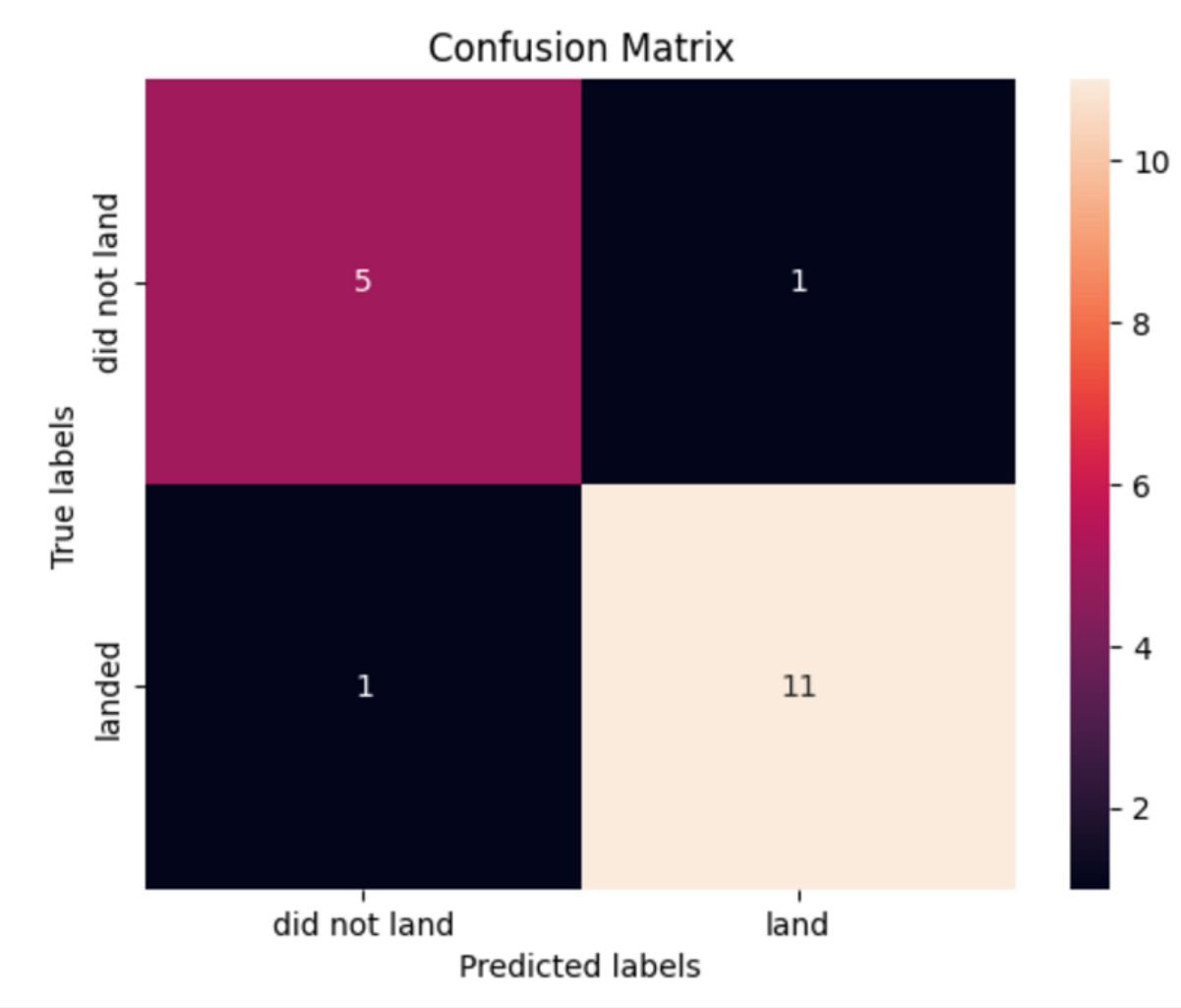
Predictive Analysis (Classification)

Classification Accuracy



The model with the highest accuracy was the decision tree with an accuracy of 89%

Confusion Matrix



The decision tree correctly classify 16 out of the 18 test cases

Conclusions

- We can predict with high accuracy the landing outcome of the different launches which we can in turn use for predict the cost of the next launch
- The variables that impact the landing outcome of a launch are:
  - Payload Mass, Orbit, Flights, Blocks, Reused, Legs, Grid fins, Booster type, Launch Site, Serial and , Landing Pad
- Most of the landings in the year 2017-2020 were successful
- The rockets with the heaviest payload mass tend to have more successful landings

Thank you!