IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anastasiia Filchenko
January 21, 2022

Outline

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

Executive Summary

Summary of methodologies

- Tools: Python and SQL
- Data sources: API and web page
- Visualizations: scatter plots, folium maps, dashboard
- Supervised machine learning: classification

Summary of all results

- SpaceX is quite successful in recovering the first stage of its rockets, continually improving since 2013
- Payload mass, destination (orbit) and booster version to a large extent define how successful the landing outcome will be
- SpaceX fails continuously to recover the first stage after launching rockets to ISS or GTO orbits; SpaceY can investigate this opportunity further

Introduction

Project background and context

SpaceY wants to enter the aerospace manufacturing business and looks up to a potential rival company SpaceX. What makes SpaceX attractive to the clients is the lower cost of the rockets which is achieved by reusing the most expensive part of the rocket – it's first stage. The main task of this project is to analyze public SpaceX launch data (Falcon 9) and predict whether the first stage will be recovered (successful landing outcome) and eventually determine the price of the rocket manufacturing for SpaceY.

Questions

- How successful is SpaceX in recovering the first stage of the rockets?
- What drives un/successful landing outcomes?
- Should SpaceY bid against SpaceX?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and web scraping of Wikipedia page
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Convert all categorical variables to dummy variables (0/1) and normalize
  - Split dataset into training and test sets (80/20)
  - Train classification models with selected hyperparameters and apply on test sets
  - Compare confusion matrices and accuracy score to define the best performing model

Data Collection

Falcon 9 data was collected using:

- SpaceX API with Requests library
  - https://api.spacexdata.com/v4/launches/past
- Web scraping of Wikipedia page with Requests and BeautifulSoup libraries
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

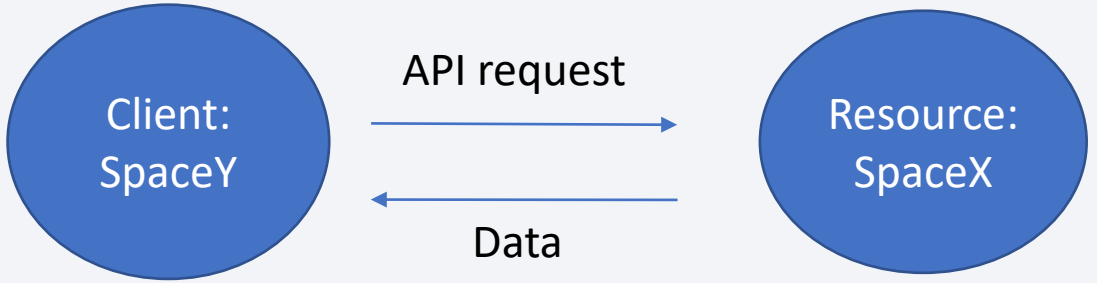
Data extracted:

```
FlightNumber
Date
BoosterVersion
PayloadMass
Orbit
LaunchSite
Outcome
Flights
GridFins
Reused
Legs
LandingPad
Block
ReusedCount
Serial
Longitude
Latitude
```

Data Collection – SpaceX API

Collecting data thru API:

- Get data using URL and *get* method from Requests library
- Decode response content as JSON using *json* method and turn it to dataframe using Pandas *json_normalize* method



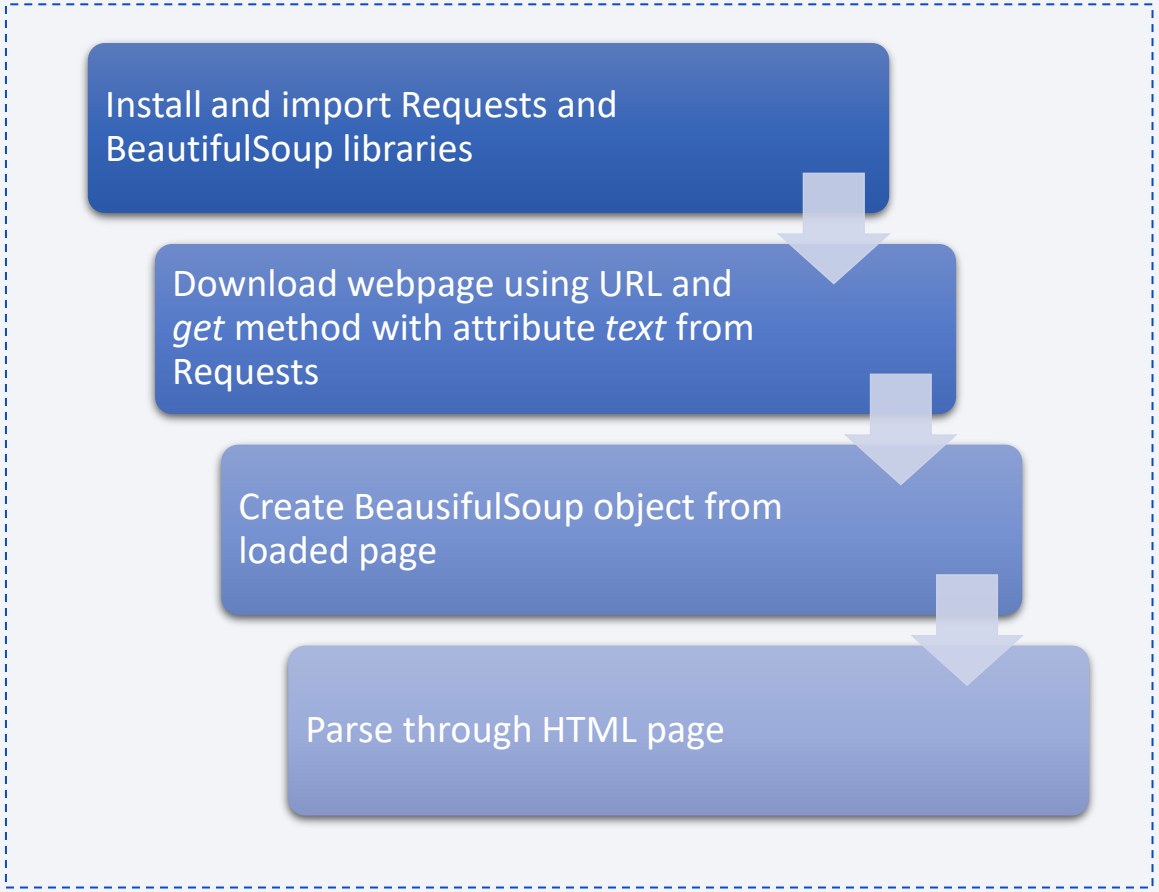
API request in form of HTTP message was sent from Client (SpaceY) to Resource (Space X). Response in form of JSON file containing requested data was sent back to Client.

https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/data-collection-api.ipynb

Data Collection - Scraping

Scraping HTML page:

- Find table header tags ('th') and table rows ('td') using method *find_all* from BeautifulSoup
- Iterate through elements of each row to extract and save data to dictionary
- Convert dictionary to Pandas dataframe



https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/data-collection-webscraping.ipynb

Data Wrangling

- Data types were checked, numerical and categorical variables identified
- Dataset was treated for missing values
- Ad hoc EA was performed on launch sites and orbit types data using pandas *value_counts* method
- Based on categorical variable Outcome a new variable Class was introduced: 0/1 – un/successful landing of the1st stage

https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/data-wrangling.ipynb

EDA with Data Visualization

To get a better insight into data, i.e. what might explain launch outcomes, the following relationships between flight number (continuous flight attempts), payload mass, launch site and orbit type were examined:

- FlightNumber vs PayloadMass
- FlightNumber vs LaunchSite
- PayloadMass vs LaunchSite
- FlightNumber vs OrbitType
- PayloadMass vs OrbitType

https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/EDA-visualization.ipynb

EDA with SQL

SQL queries performed:

- Landing outcome: the first successful landing, landing types, list of boosters for specific outcomes, and ranking the count of outcomes for certain periods
- Payload mass: total payload mass per customer
- Boosters: explore boosters for heavy payload mass rockets

https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/EDA-sql.ipynb

Build an Interactive Map with Folium

To get a better overview over rocket launch locations, the following objects were added to the interactive map:

- Launch site areas
- Launch outcomes per site
- Launch site proximity to highways, railroads, towns and coast

https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/EDA-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

Data suggests that payload mass of the rocket and its booster version have a strong affect on landing outcome. To visualize these relationships, two charts were added to the dashboard:

- Pie chart: Success rate (per site), %; and Total successful landings by site, %
- Scatter plot: Launch outcomes by site within selected payload mass range, giving insight into which booster versions were used

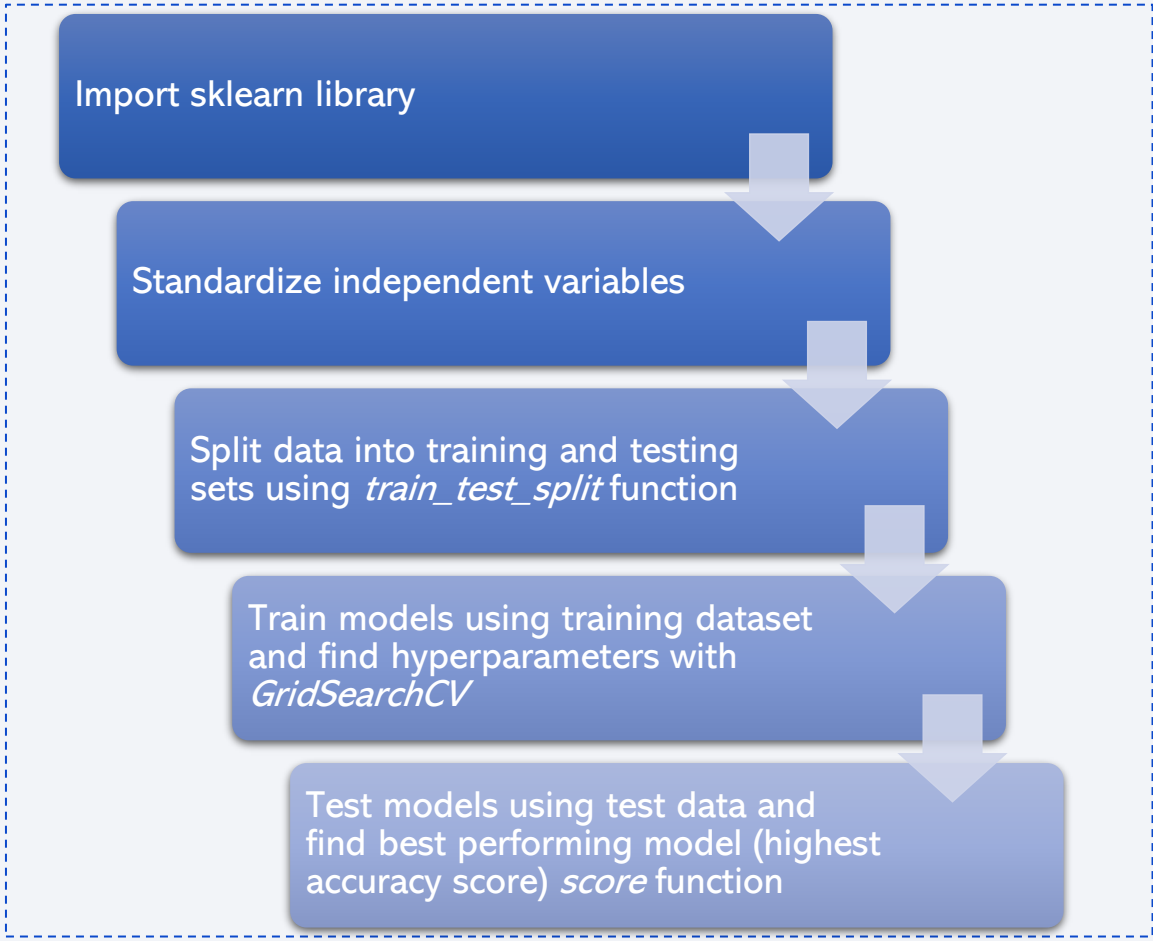
https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/EDA-dashboard.py

Predictive Analysis (Classification)

Following the process steps described on the right, four classification models were tested:

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K-Nearest Neighbors

To find the best performing model, confusion matrix was plotted, and accuracy score (score method) was calculated for each model.



https://github.com/anastasiiaf/ibm_data_science/blob/master/Capstone/ML-prediction.ipynb

Results

Exploratory data analysis results for Falcon 9

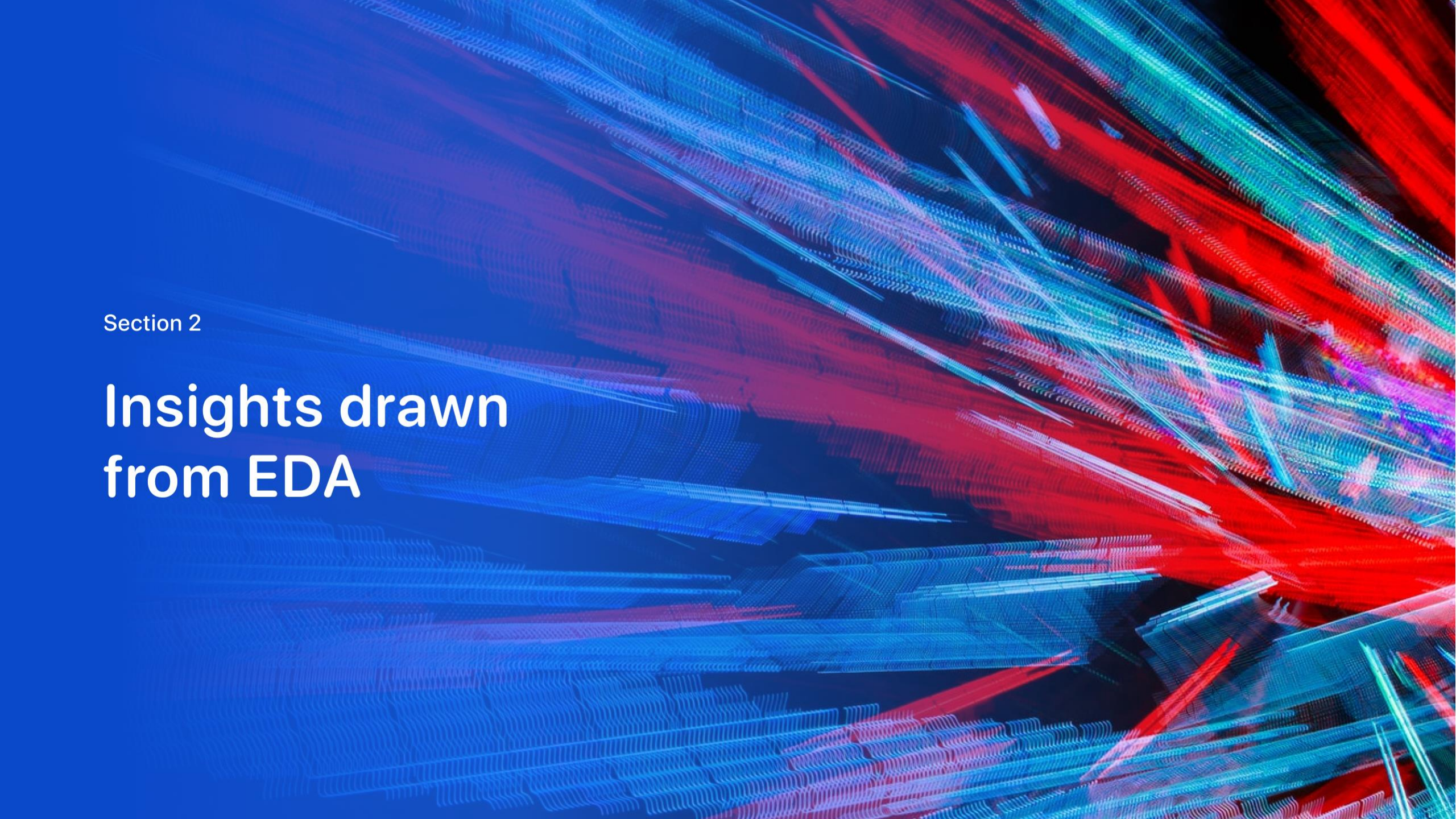
- The average success rate of retrieving the first stage is around 67% and around 80% in 2020
- Depending on destination (orbit) and payload mass, the launch outcomes variate:
  - Very successful with launching heavy payload mass rockets
  - Very successful with launching to orbits VLEO, LEO and PO
  - Numerous but least successful with launching to orbits ISS and GTO
  - In majority of successful landings booster versions FT and B4 are used

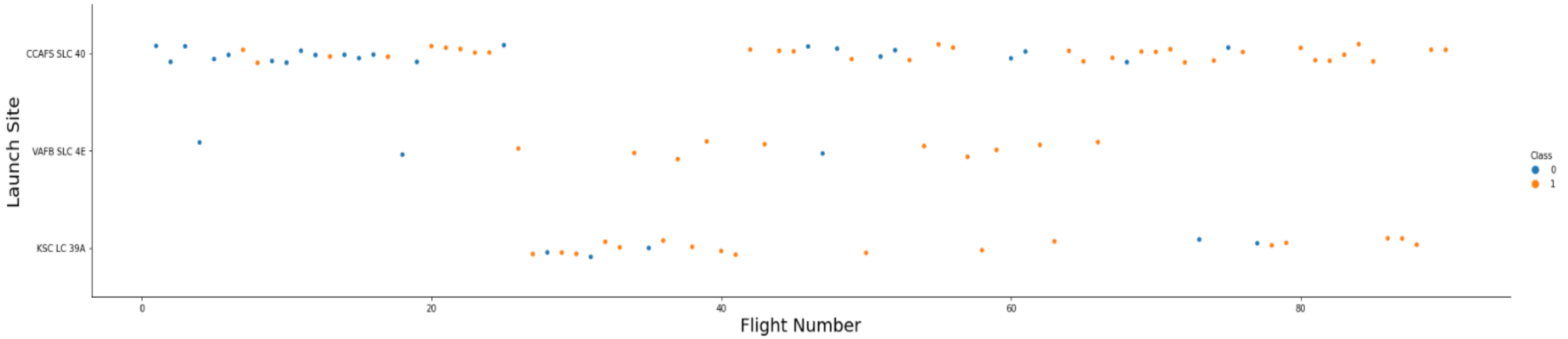
Predictive analysis results

- Due to small data sample, all four classification models tested yield appx the same results, meaning every model can be used for predicting the launch outcome, and eventually price of the rocket manufacturing.

Section 2

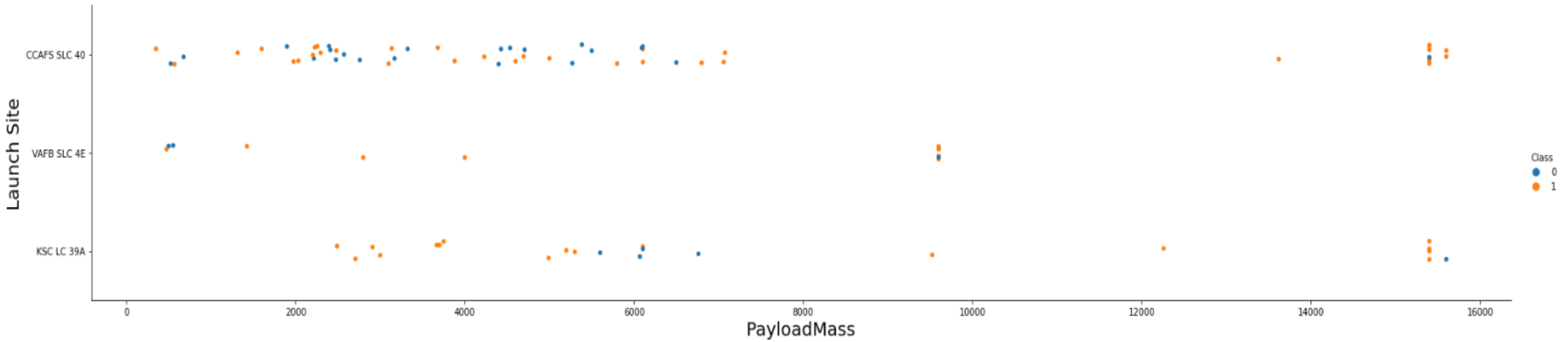
Insights drawn from EDA





Flight Number vs. Launch Site

- Majority launches took place at CCAFS SLC 40, although success rates are higher at two other sites
- More launches (continuous launch attempts) increase the likelihood of getting successful landing

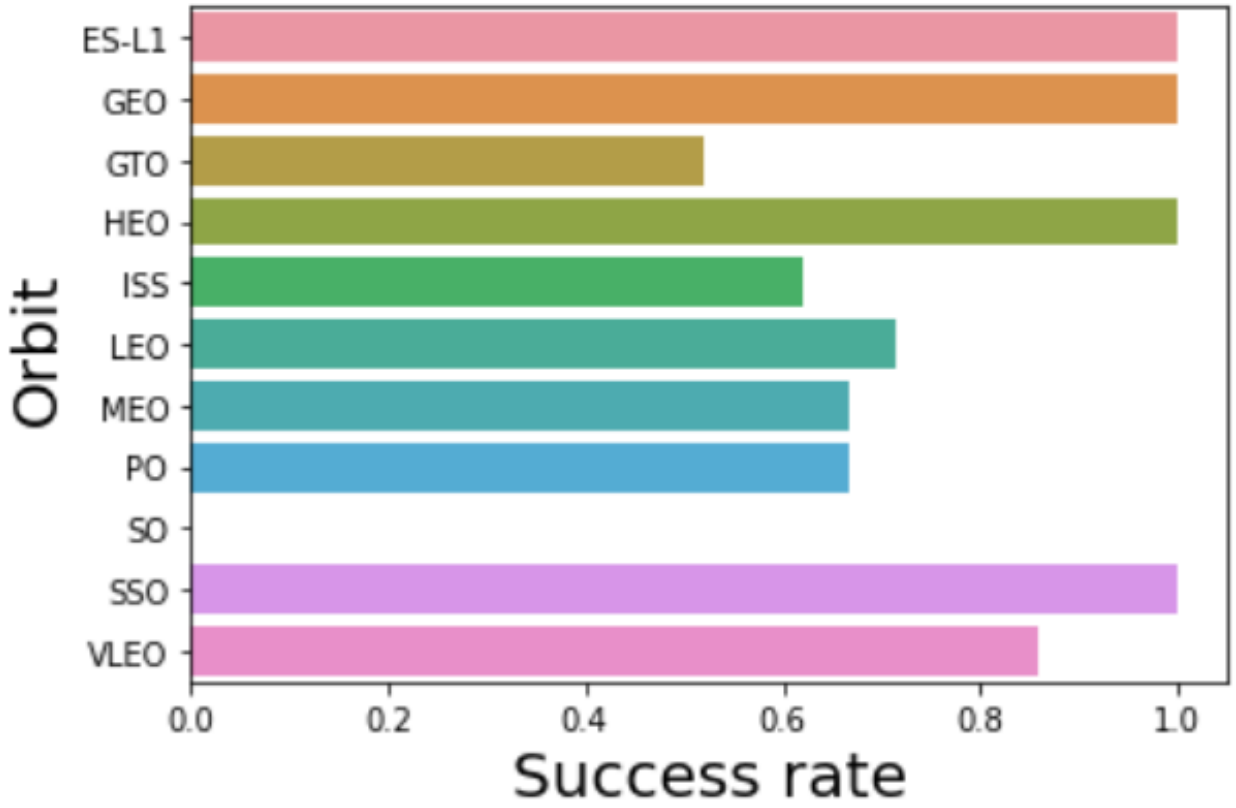


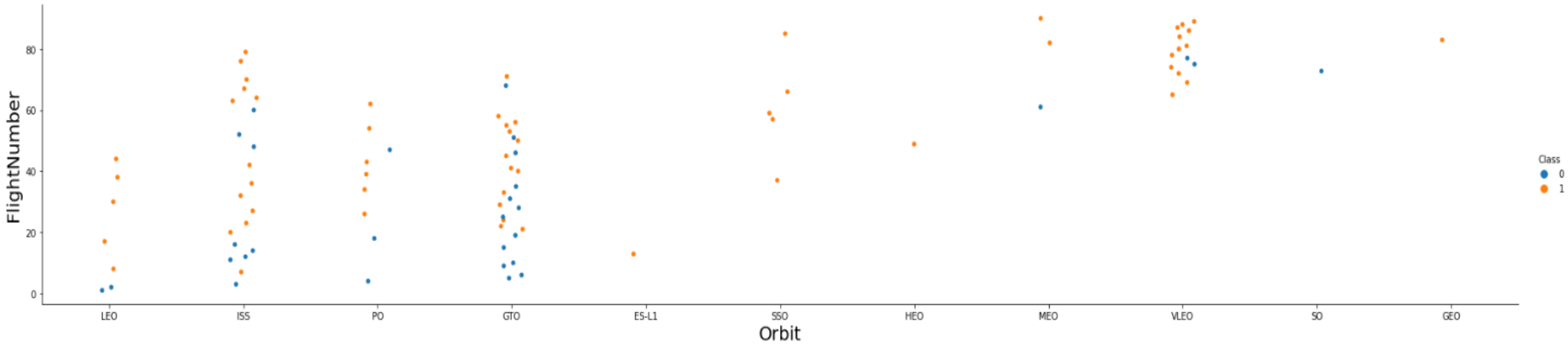
Payload vs. Launch Site

- Majority of launched rockets have payload mass between 500 kg to 7000 kg
- VAFB SLC 4E site has no rockets with heavy payload mass

Success Rate vs. Orbit Type

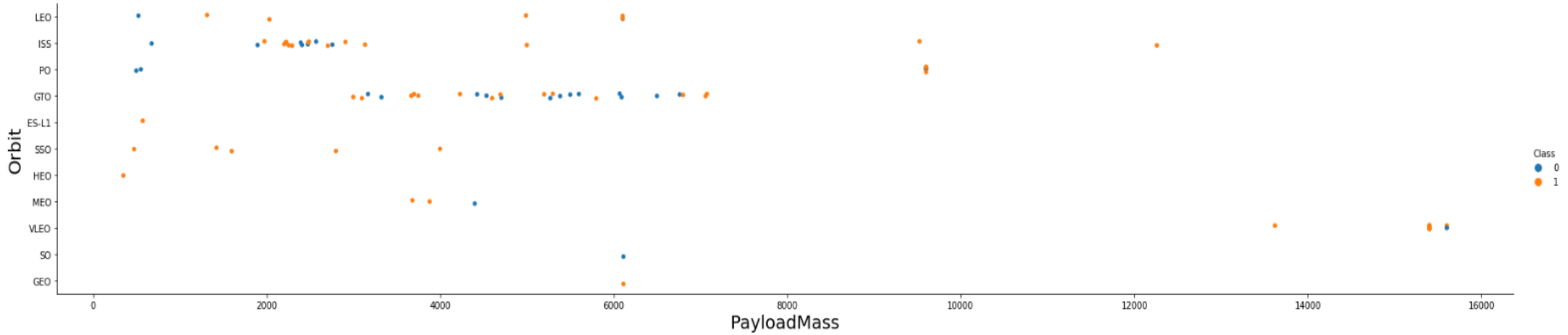
- Launches to orbits GTO, ISS, MEO and PO were least successful
- Majority of launches happened to orbits GTO, ISS, VLEO, PO and LEO
- There are few (1 to 3) launches to orbits ES-L1, GEO and HEO





Flight Number vs. Orbit Type

- Successful outcome in LEO orbit is related to the number of flights (improvement with every next launch), which is not the case for orbits ISS and GTO

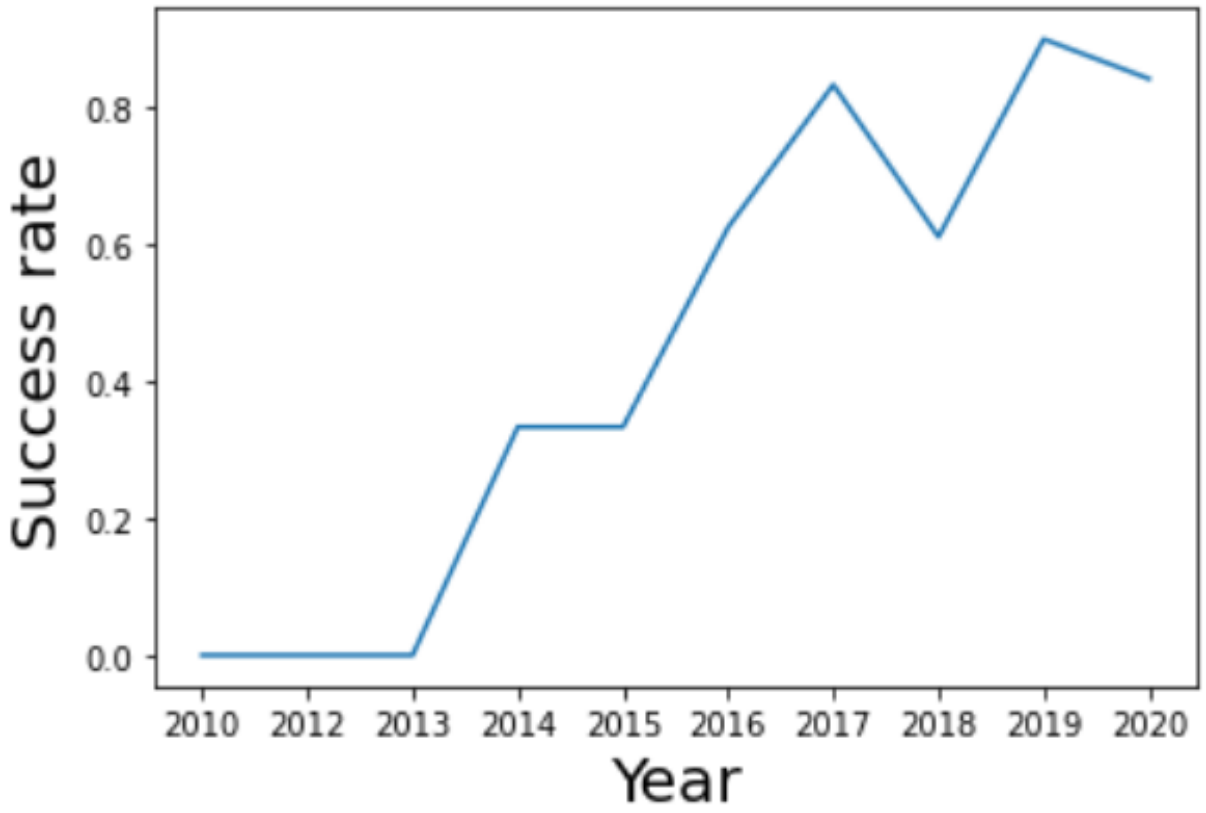


Payload vs. Orbit Type

- Heavy payload mass rockets have successful outcome in orbits ISS, PO and VLEO

Launch Success Yearly Trend

- From 2013 until 2020 success rate is increasing



All Launch Site Names

- Query result shows the names of 4 unique launch sites
- SQL: keyword DISTINCT is used in SELECT statement

```
%sql SELECT distinct launch_site FROM SPACEX
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85
Done.
```

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

Launch Site Names Begin with 'CCA'

- Query result shows 5 records where launch sites begin with `CCA`
- SQL:
  - keyword LIKE is used to retrieve entries which contain ‘CCA’ in launch_site column
  - Keyword LIMIT 5 restricts the retrieved dataset to 5 entries

```
%sql SELECT * FROM SPACEX WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Query result shows the total payload carried by boosters from NASA
- SQL: function SUM is used to get the total sum

```
%sql SELECT sum(payload_mass__kg_) FROM SPACEX WHERE customer=UCASE('NASA (CRS)')
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u9
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Query result shows the average payload mass carried by booster version F9 v1.1
- SQL: function AVG is used to get an average value

```
%sql SELECT avg(payload_mass__kg_) FROM SPACEX WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g
in.cloud:32286/bludb
Done.
```

1
2534

First Successful Ground Landing Date

- Query result shows the dates of the first successful landing outcome on ground pad
- SQL: function MIN is used to get a minimum value in a column; can be applied to date type as well

```
%sql SELECT min(date) FROM SPACEX WHERE landing__outcome LIKE 'Success%'
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0
in.cloud:32286/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- SQL: keywords BETWEEN and AND are used to retrieve data based on several conditions in one statement

```
%%sql SELECT booster_version FROM SPACEX
WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.datab
in.cloud:32286/bludb
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

- Query result shows the total number of successful and failed mission outcomes
- SQL: function COUNT is used to calculate the number of mission outcomes

```
%sql SELECT distinct mission_outcome FROM SPACEX
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-
in.cloud:32286/bludb
Done.
```

mission_outcome
Failure (in flight)
Success
Success (payload status unclear)

```
%sql SELECT count(mission_outcome) FROM SPACEX
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-
in.cloud:32286/bludb
Done.
```

1
101

Boosters Carried Maximum Payload

- Query result shows the names of the boosters which have carried the maximum payload mass
- SQL: nested SELECT statement in WHERE clause

```
%sql SELECT booster_version FROM SPACEX WHERE payload_mass__kg_=(SELECT max(payload_mass__kg_) FROM SPACEX)
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
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

- Query result shows the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

- SQL: function YEAR is used to retrieve year from the date

```
%%sql SELECT date, booster_version, launch_site, landing__outcome FROM SPACEX
WHERE landing__outcome='Failure (drone ship)' and year(date)=2015
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk
oud:32286/bludb
Done.
```

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Query result shows the ranked 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: ranking is done with GROUP BY and ORDER BY clauses

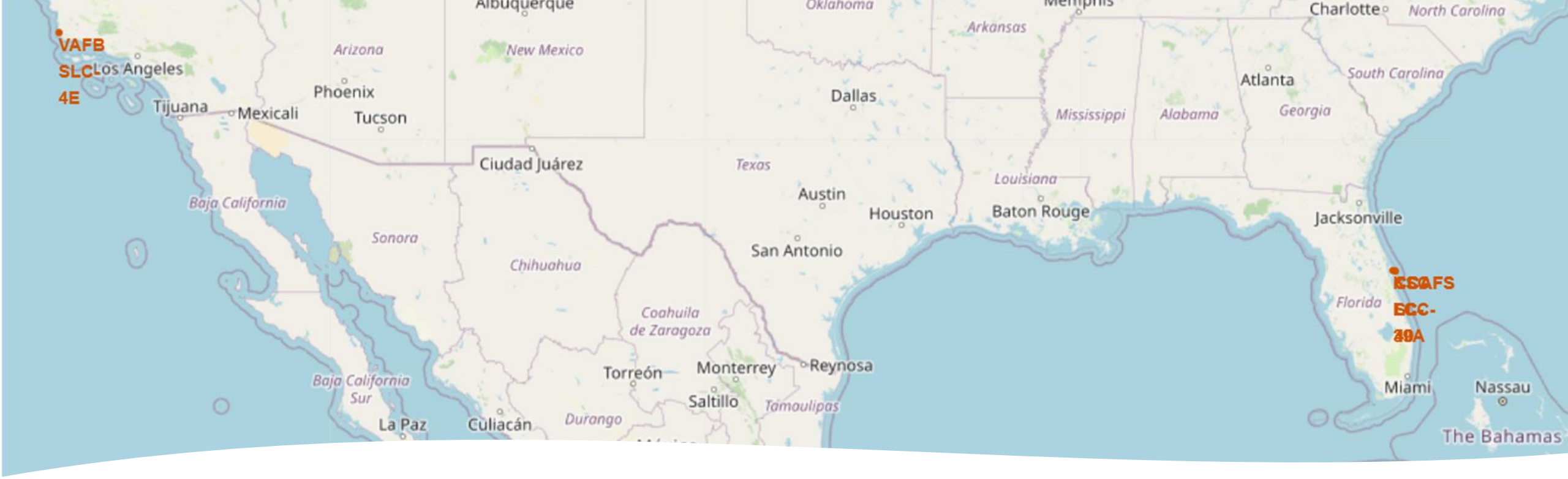
```
%%sql SELECT landing__outcome, count(landing__outcome) FROM SPACEX
WHERE date >= '2010-06-04' and date <'2017-03-21' GROUP BY landing__outcome ORDER BY 2 DESC
```

```
 * ibm_db_sa://qhz81330:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databa
oud:32286/bludb
Done.
```

landing__outcome	2
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

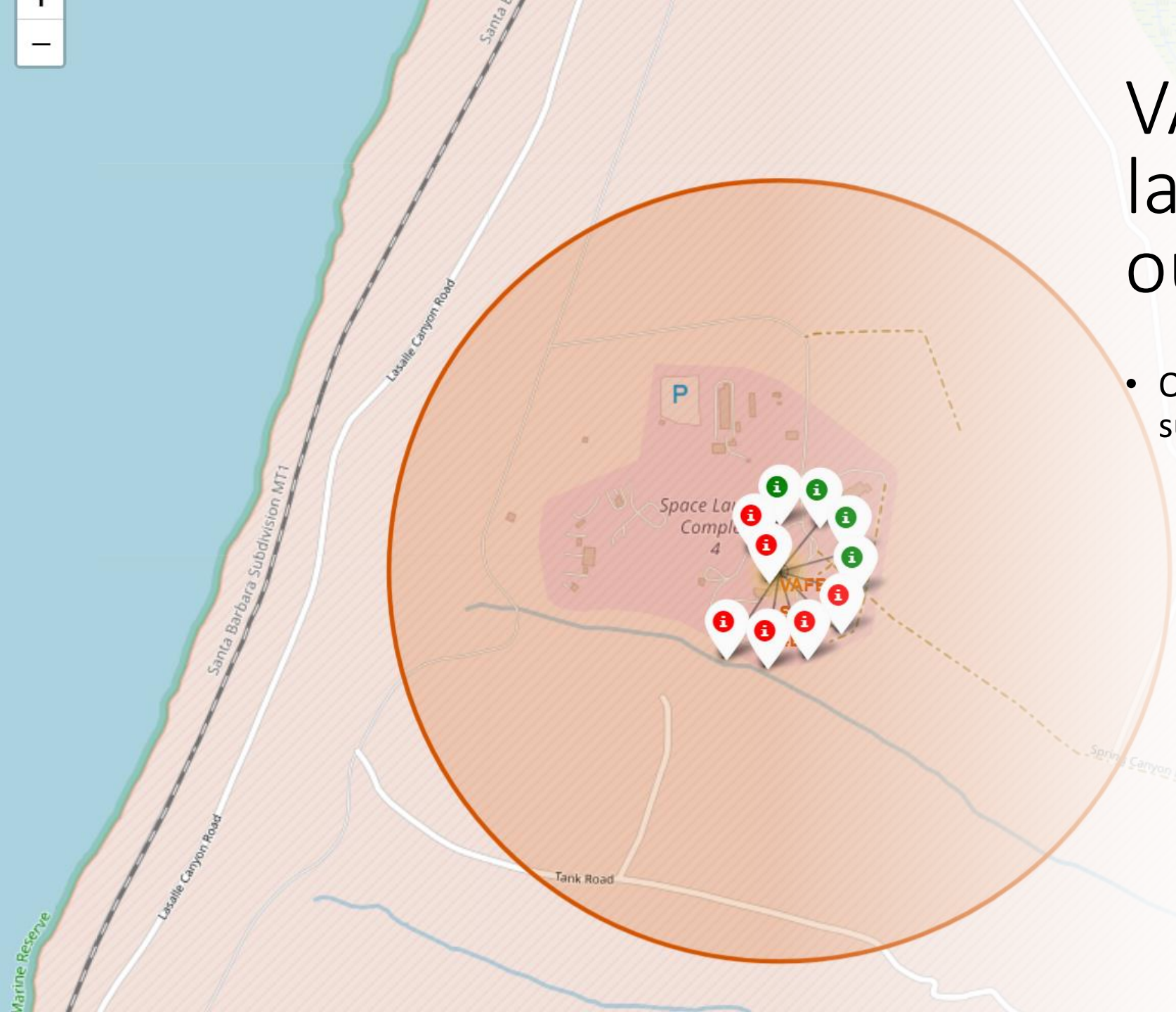
Section 4

Launch Sites Proximities Analysis



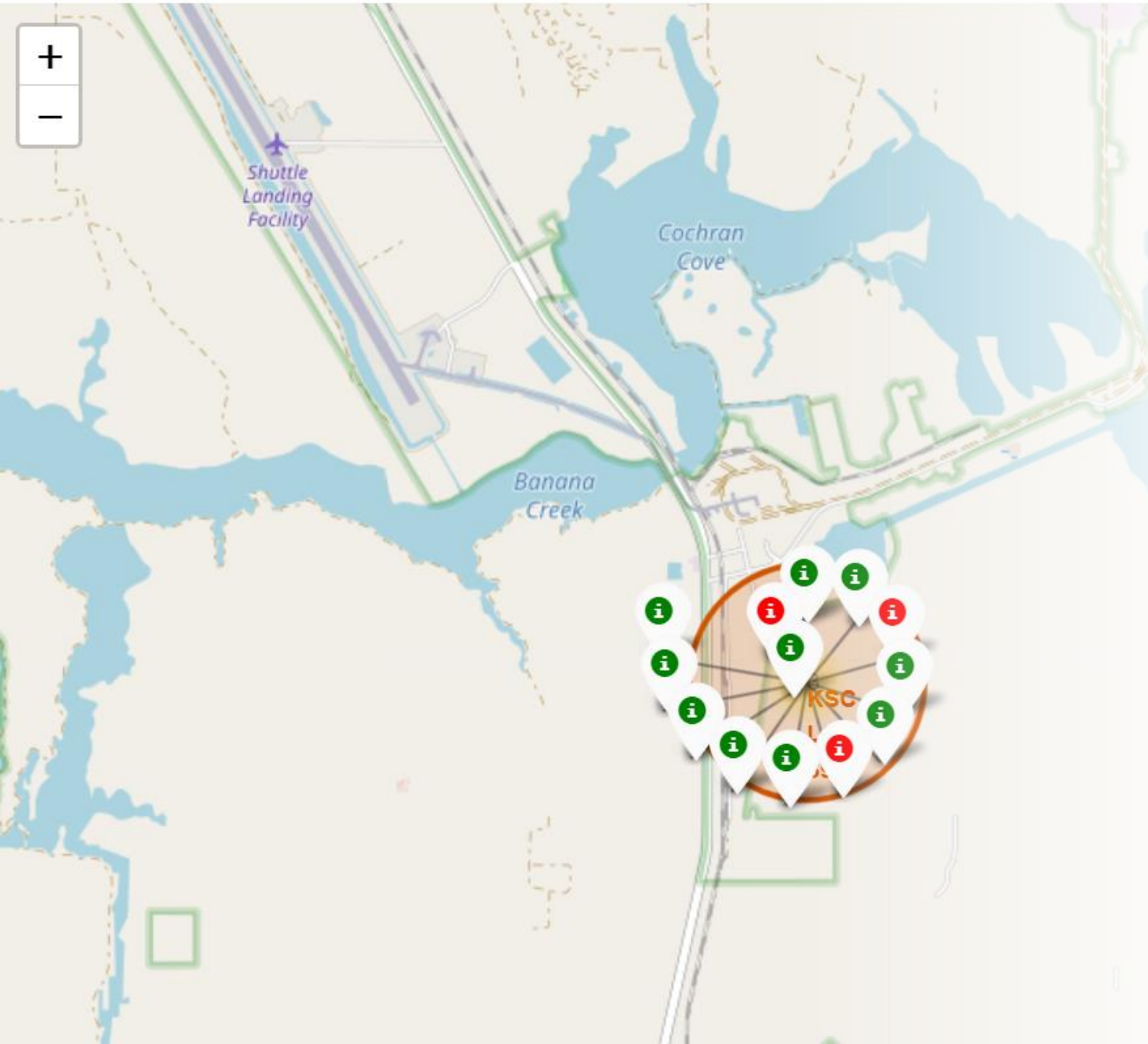
Launch sites

- Launching sites are located very close to the coastal area: VAFB SLC-4E site on west coast (California, Pacific ocean) and others on east coast (Florida, Atlantic ocean)
- Sites CCAFS S/LC-40 and KSC LC-39A are closer to the equator



VAFB SLC-4E: launch outcomes

- Out of 10 launches only 4 were successful

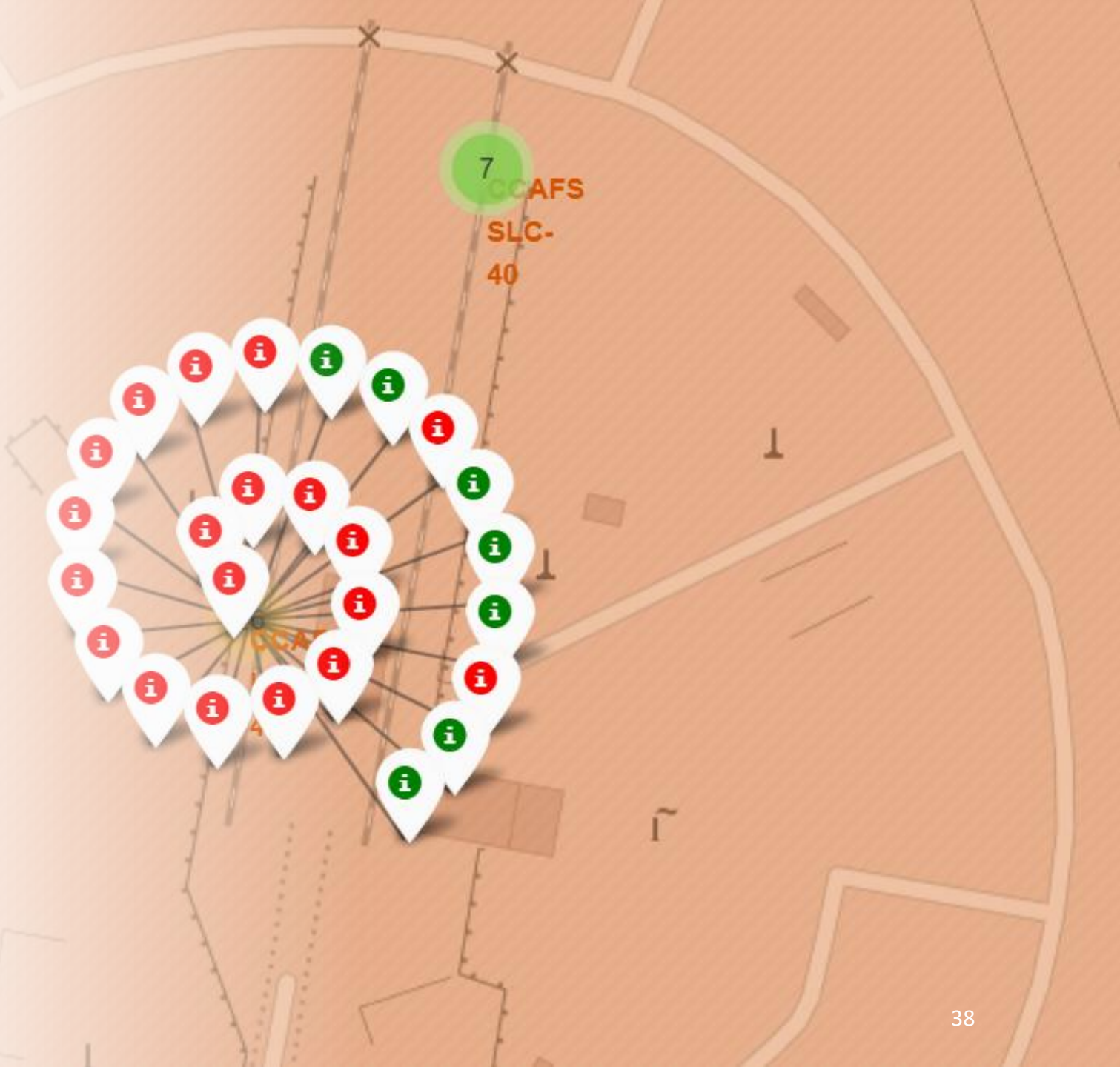


KSC LC – 39A: launch outcomes

- Out of 13 launces 10 were successful

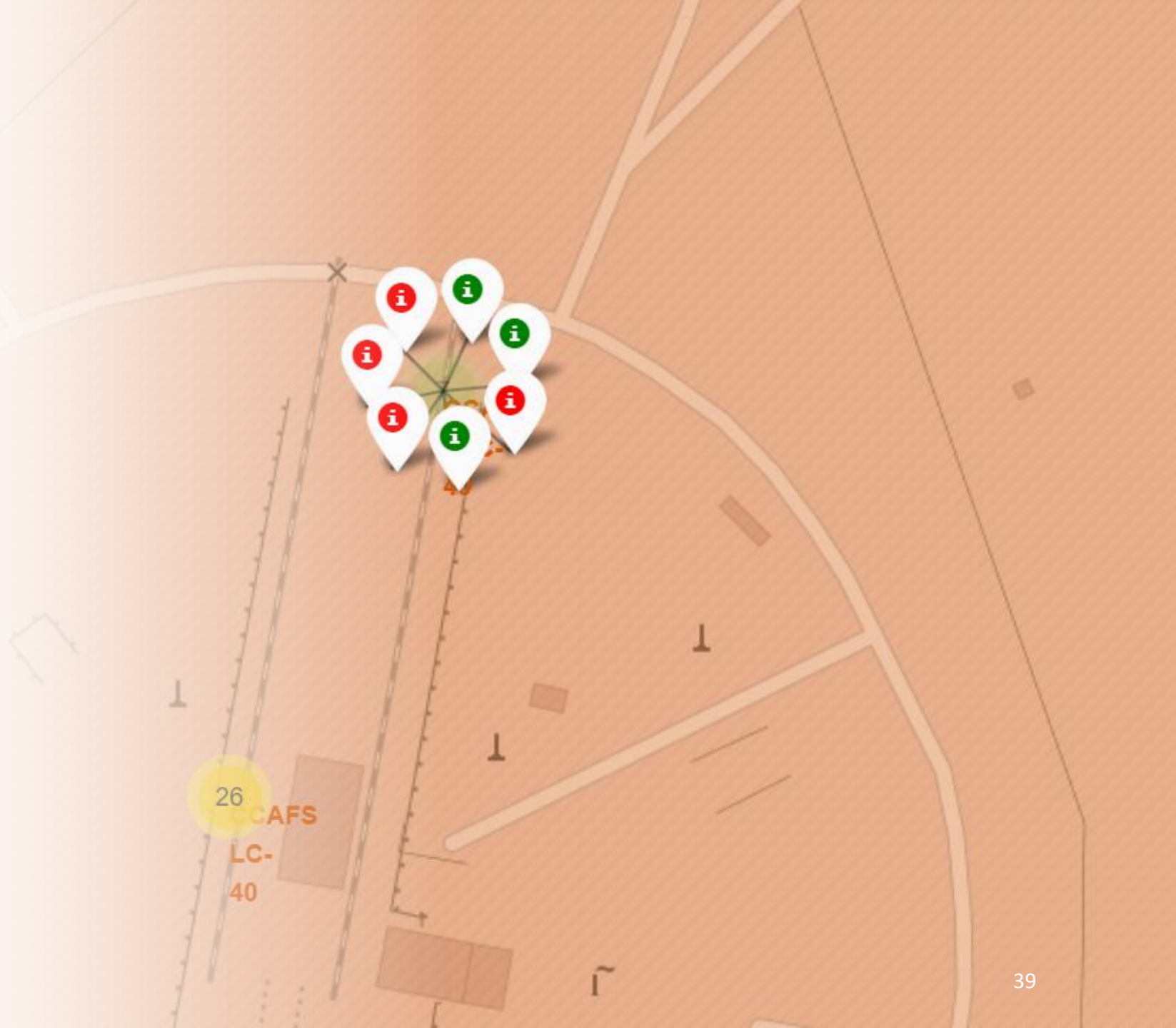
CCAFS LC-40: launch outcomes

- Out of 26 launches only 7 were successful



CCAFS SLC-40: launch outcomes

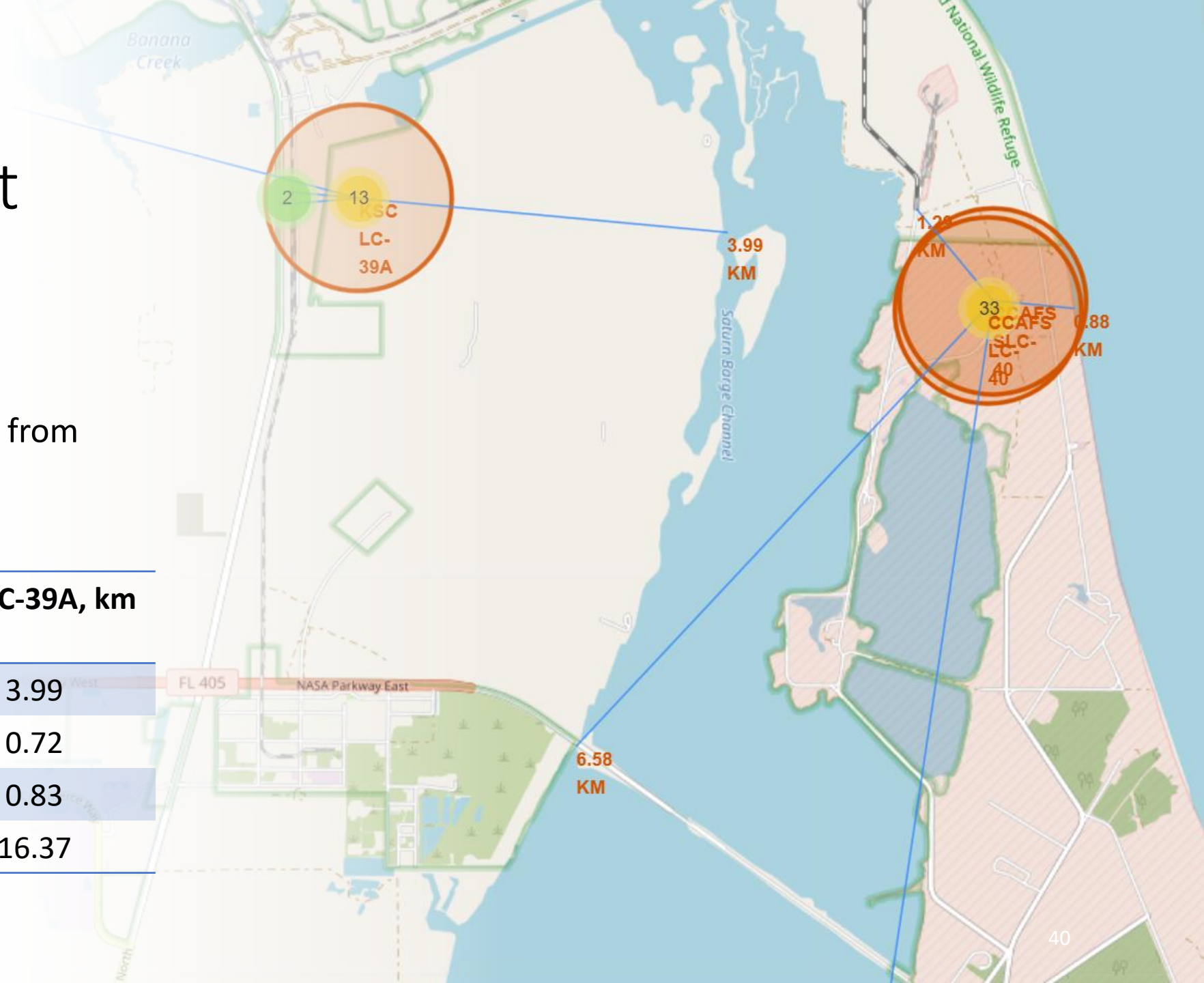
- Out of 7 launches only 3 were successful

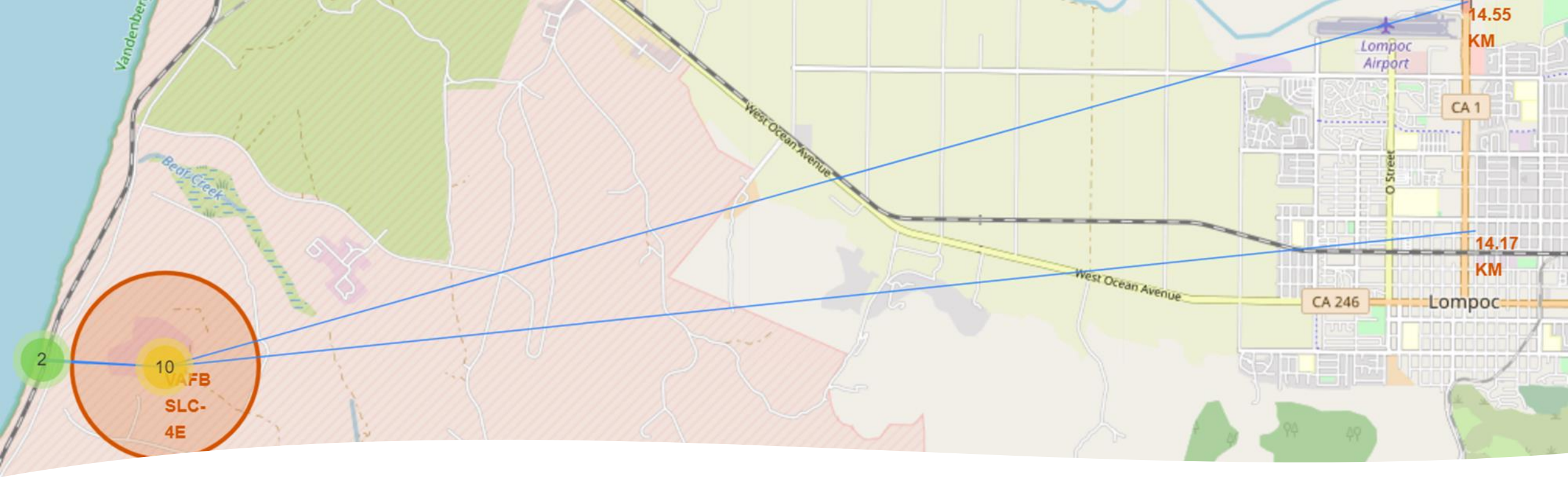


Launch sites proximity: East coast

- Sites are located closer to railroads and further away from towns

Object / Site	CCAFS S/LC 40, km	KSC LC-39A, km
Coast	0.88	3.99
Railroad	1.29	0.72
Highway	6.58	0.83
Town	18.16	16.37





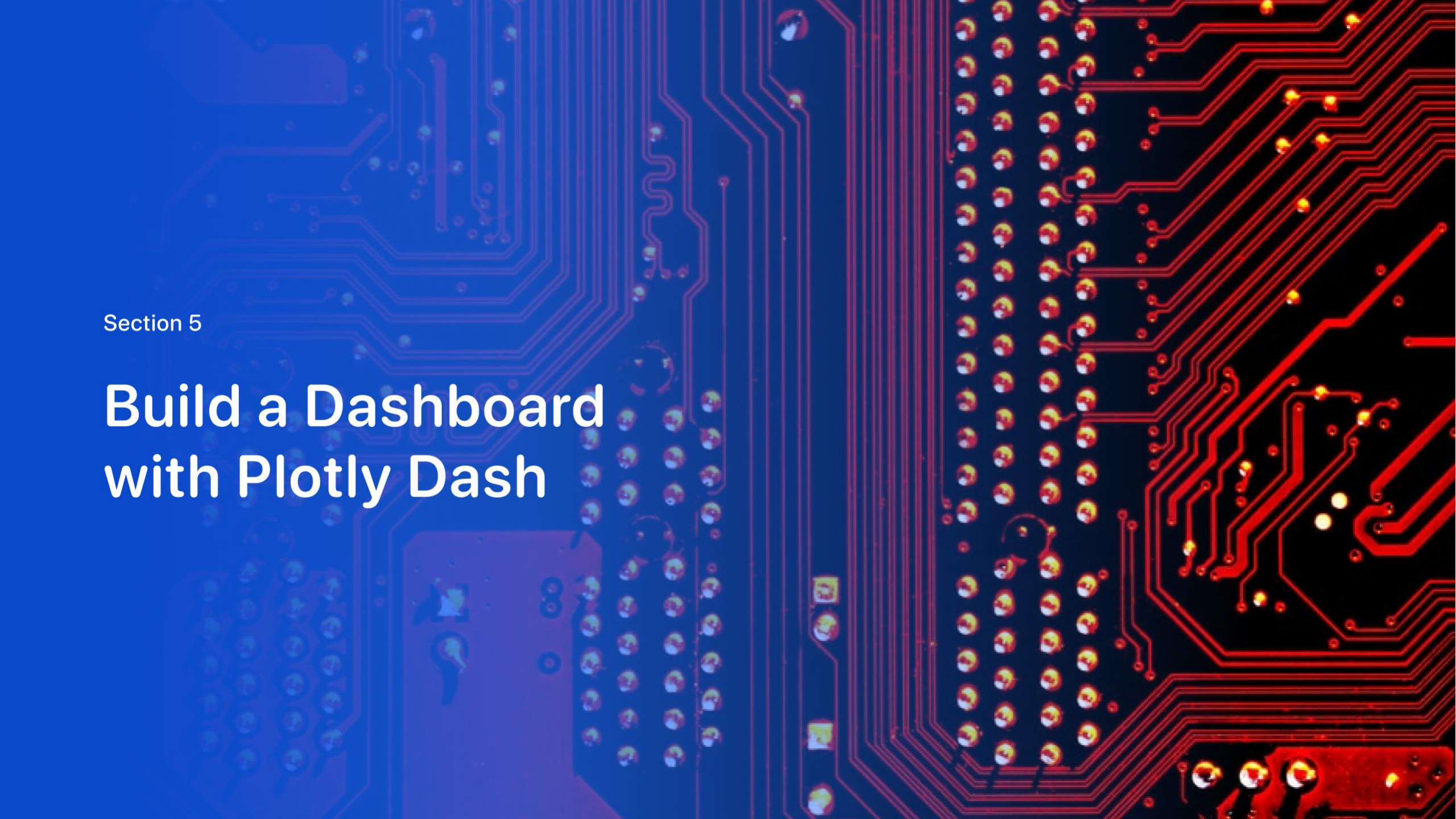
Launch sites proximity: West coast

- West coast site is located nearby coast and railroad, and equally away from the closest highway and town

Object / Site	VAFB SLC-4E, km
Coast	1.38
Railroad	1.27
Highway	14.55
Town	14.17

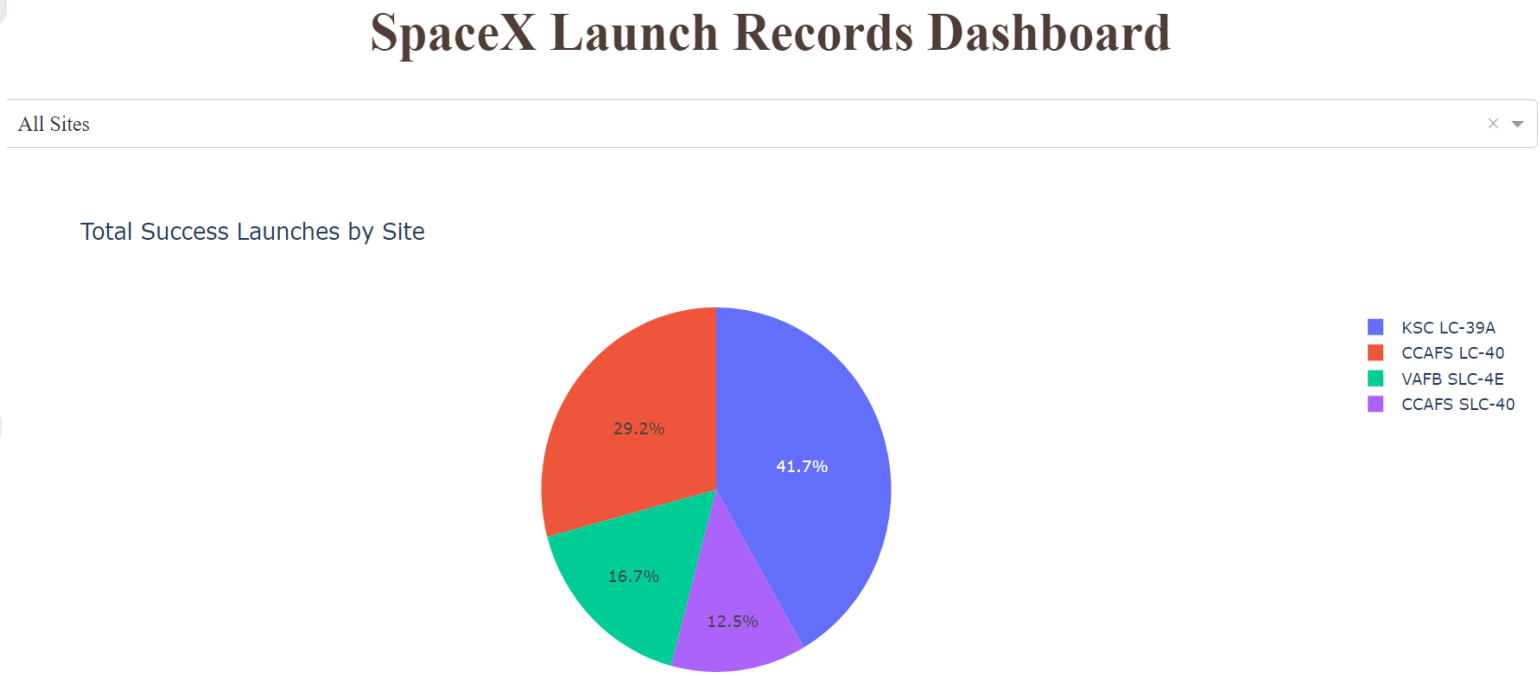
Section 5

Build a Dashboard with Plotly Dash



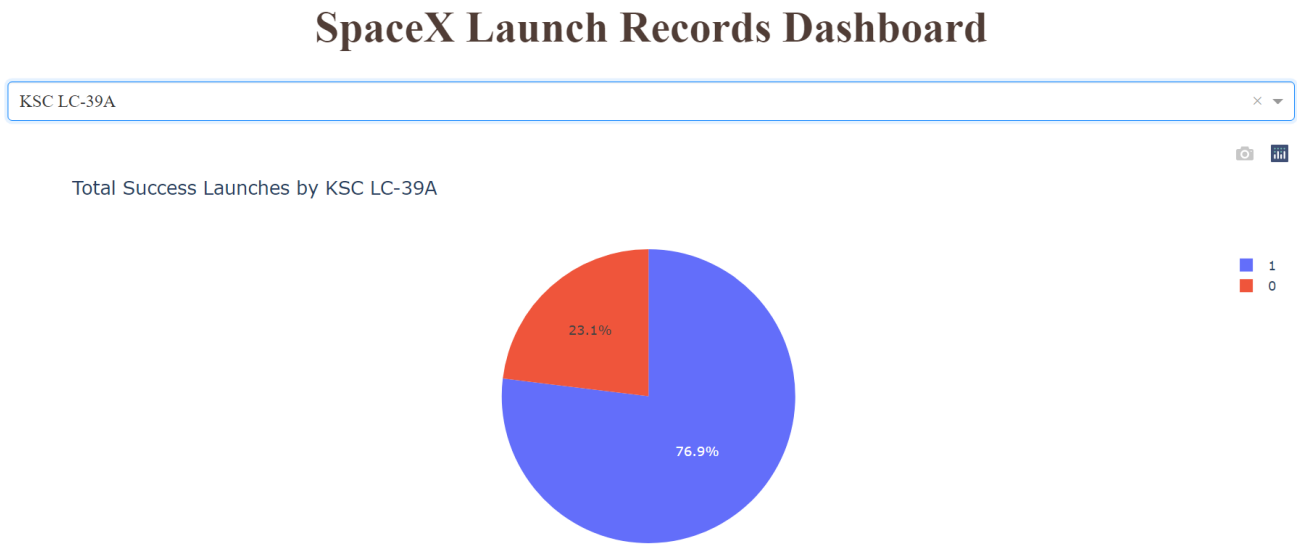
Launch sites: Success share

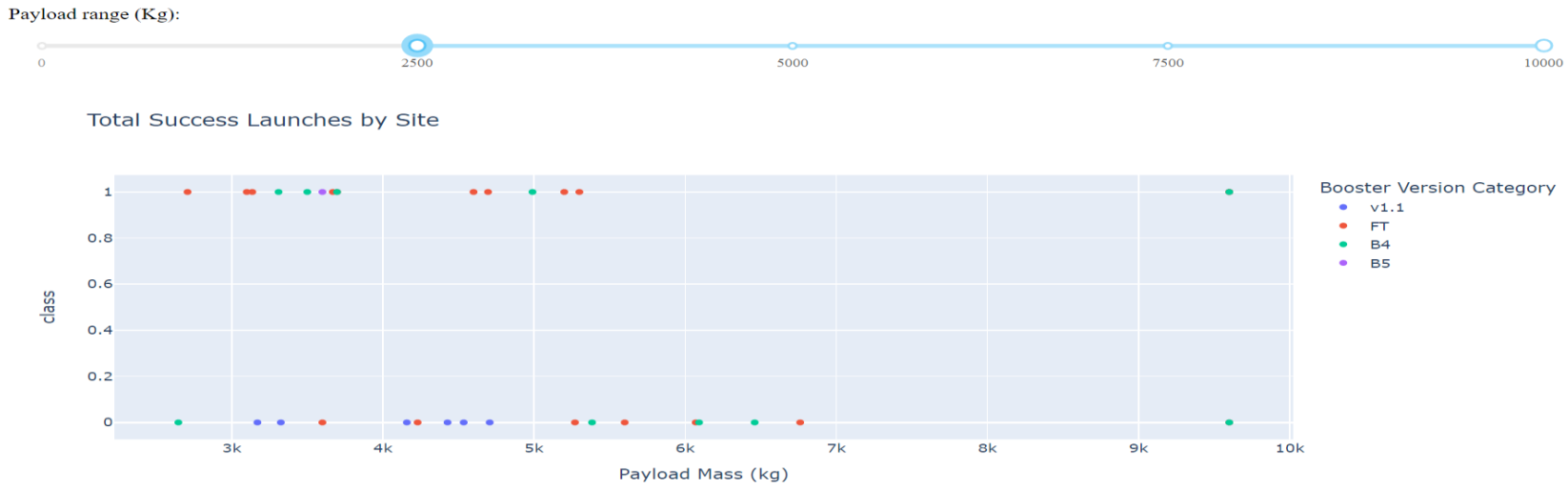
According to the data, around 42% of all successful rocket launch outcomes happened on site KSC LC-39A located in Florida



Highest launch success ratio

Among 13 launches 10 were marked as successful, yielding success ratio of 77%





Payload mass and booster

Majority of successful launch outcomes occurred where:

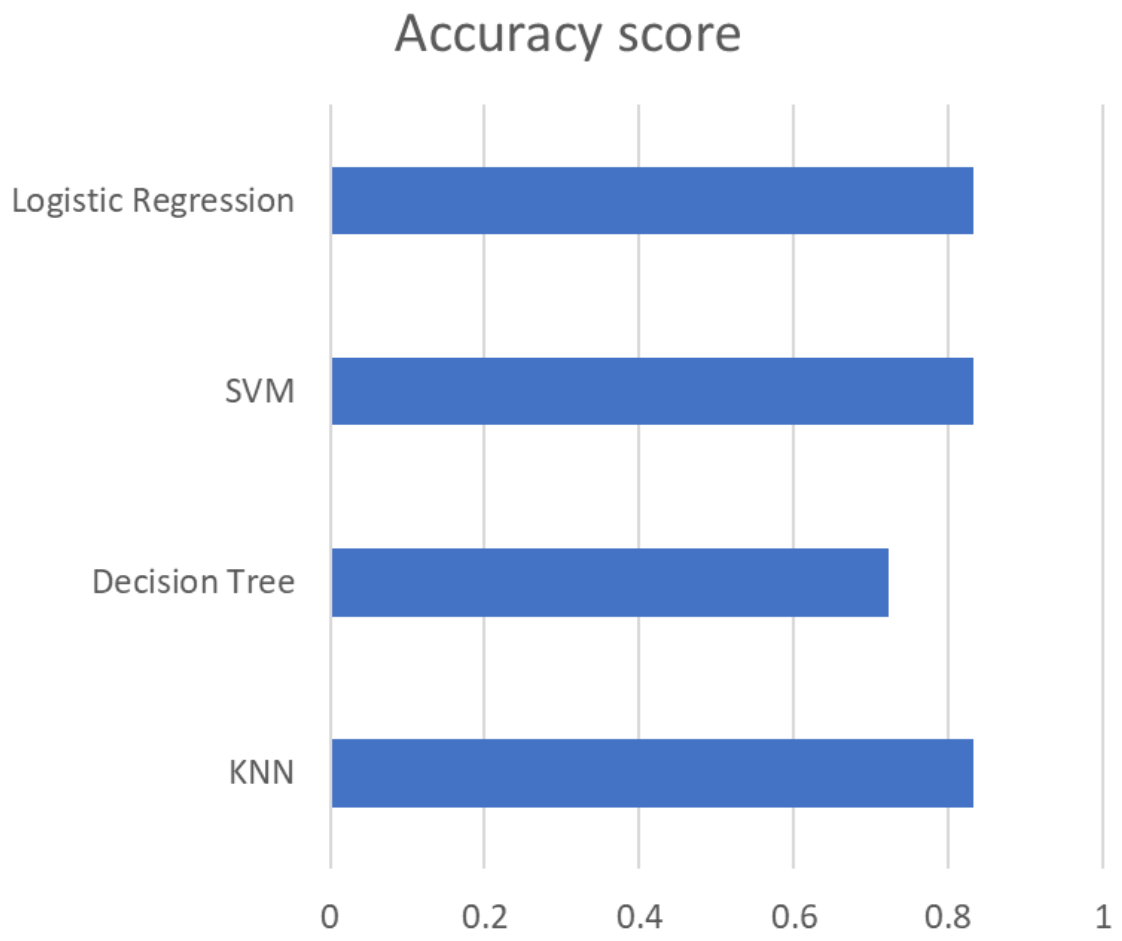
- Payload mass was in range from 2500 kg to 5300 kg
- Booster version used was FT, followed by B4

Section 6

Predictive Analysis (Classification)

Classification Accuracy

All classification models yielded approximately the same accuracy score on test dataset, meaning that all models perform relatively well.

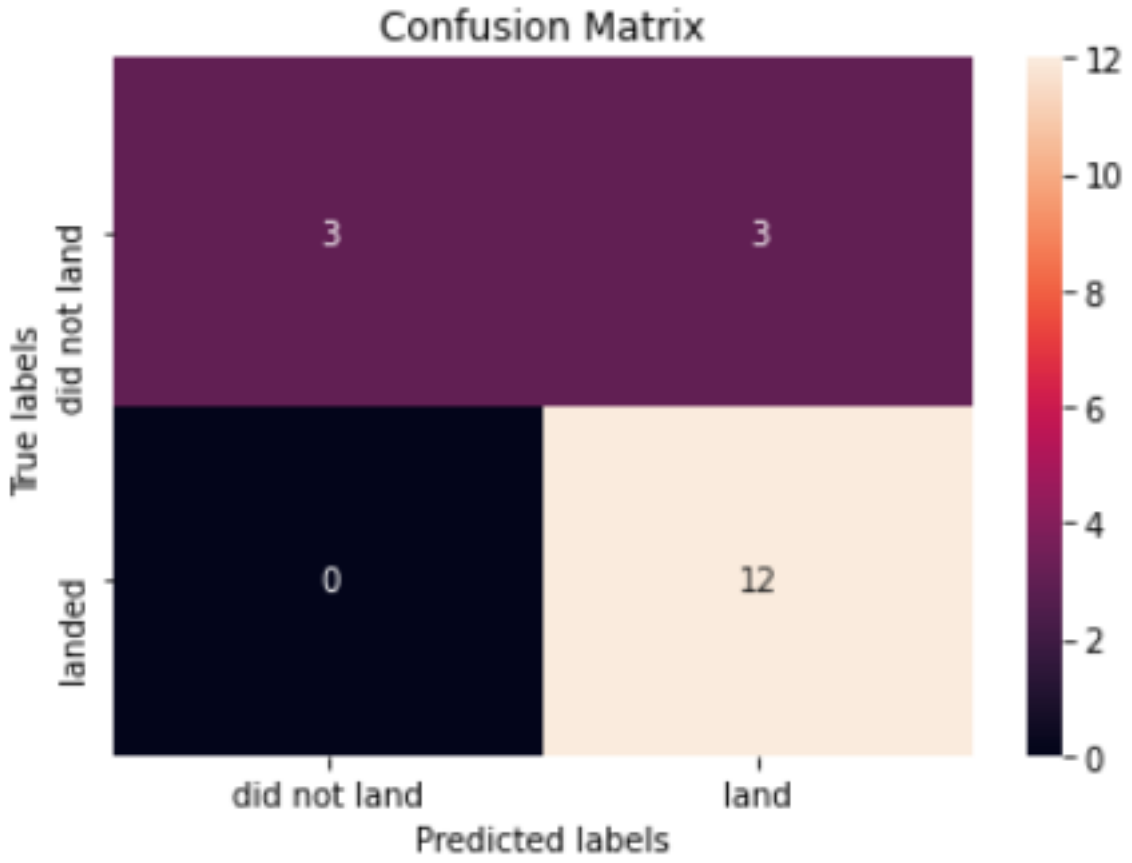


Algorithm	Jaccard	F1-score	Accuracy
KNN	0.700	0.815	0.833
Decision tree	0.569	0.715	0.722
SVM	0.700	0.815	0.833
LogisticRegression	0.700	0.815	0.833

Confusion Matrix

All classification models had the same results when confusion matrices were plotted:

- models were able to distinguish different classes (land / did not land)
- 3 landing outcomes were not classified properly (false positives)



Conclusions

- With a collected data consisting of 90 launch entries, all tested classification models perform well and can be used to predict landing outcome and eventually price of the rocket manufacturing
- To improve further model performance, collect more data on Falcon 9

Thank you!