IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alex Belenky
December 23, 2022

Outline

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

From Wikipedia

Executive Summary

- Summary of methodologies
  - - data collection via API (SpaceX API) and web scraping (Wikipedia SpaceX page)
  - - Exploratory Data Analysis, Data visualization, Feature engineering
  - - Machine Learning Prediction (utilizing 4 different algorithms)
- Summary of all results
  - - Decision Tree algorithm predicted with the highest accuracy
  - - Prediction accuracy is highly dependent on random state
  - - public data source are good enough data sources for success prediction

Introduction

- Background:
- As a research for a new company SpaceY, successful rocket launch parameters and mission success rates are highly important in determining the price of a launch.
-
- The main idea in the present work is to determine:
- 1. Launch characteristics
- 2. Success landing rates

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Data collection via API (SpaceX API)

- Web scraping with beautifulsoup4 from Wikipedia SpaceX page

Data wrangling:

- Collected data was enriched by creating a landing outcome

label based on outcome data, empty cells were treated

Exploratory data analysis (EDA) using visualization and SQL

- - Perform interactive visual analytics using Folium
- and Plotly Dash

Methodology (cont.)

Executive Summary

Predictive analysis using classification models

1. The data was scaled using standard scaler from scikit-learn package
2. The data was split into training and test sets
3. Four differnet classification algorithms were used
4. Accuracy of the algorithms was checked

Data Collection

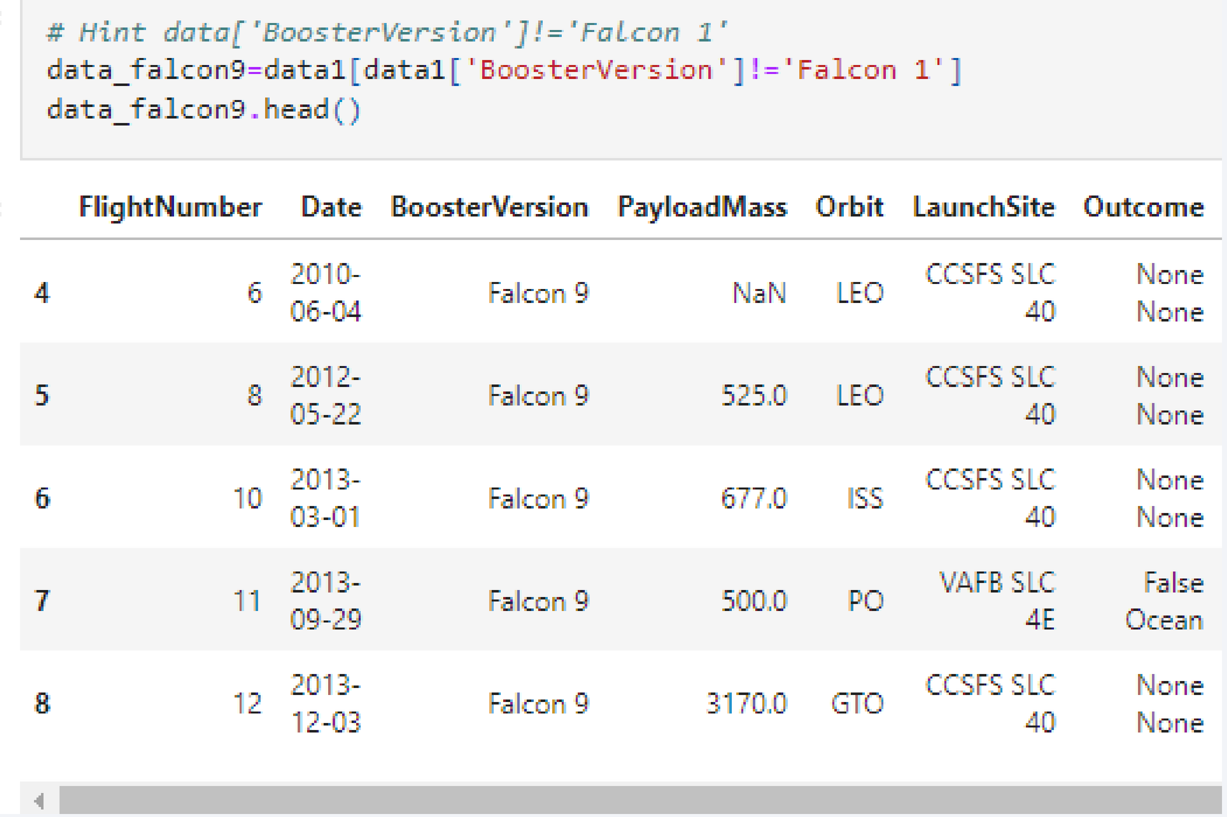
1. Data collection via API (SpaceX API):

https://api.spacexdata.com/v4/

2. Web scraping with beautifulsoup4 from Wikipedia SpaceX page:

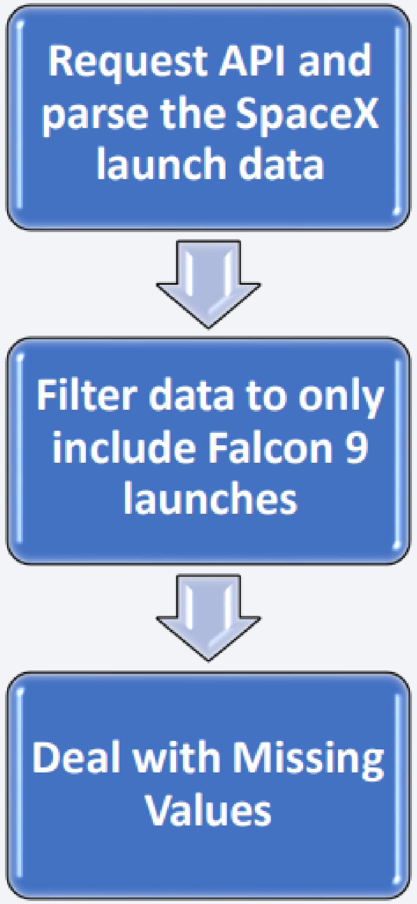
Wikipedia SpaceX page

Data Collection – SpaceX API



```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9=data1[data1['BoosterVersion']!='Falcon 1']
data_falcon9.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None



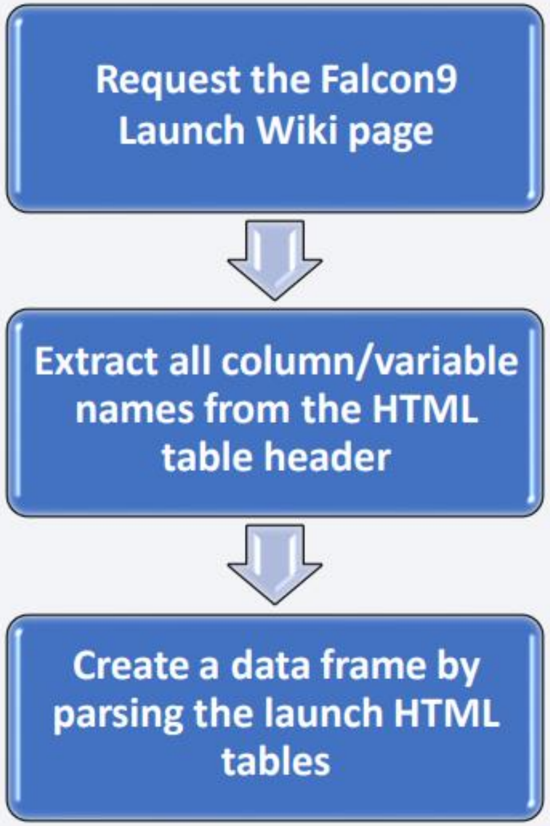
Request API and parse the SpaceX launch data

Filter data to only include Falcon 9 launches

Deal with Missing Values

GitHub URL: https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/week_1_Data_Collection_API_Lab.ipynb

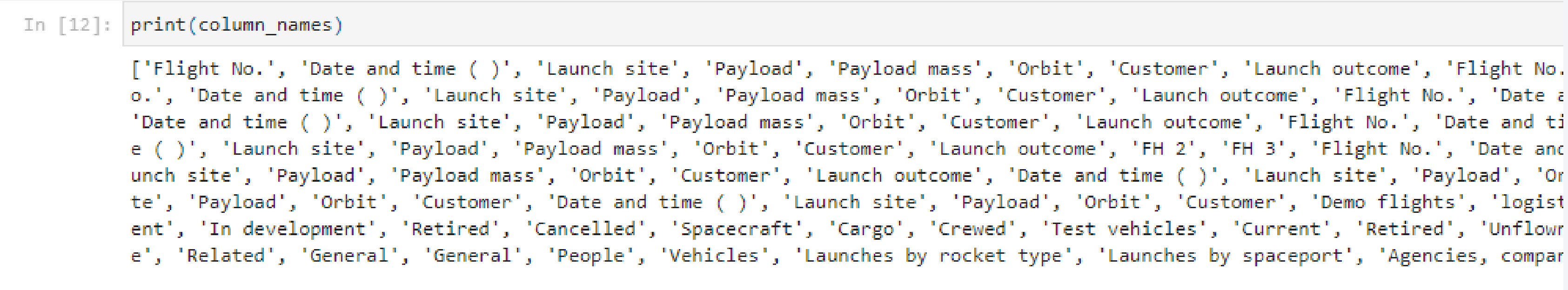
Data Collection - Scraping



Request the Falcon9 Launch Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

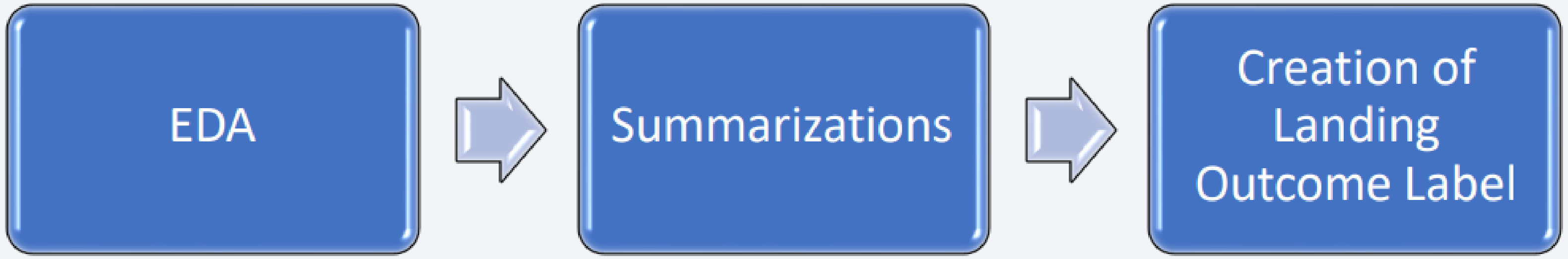


```
In [12]: print(column_names)

['Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Flight No.
o.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Flight No.', 'Date a
'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Flight No.', 'Date and ti
e ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'FH 2', 'FH 3', 'Flight No.', 'Date and
unch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Date and time ( )', 'Launch site', 'Payload', 'Or
te', 'Payload', 'Orbit', 'Customer', 'Date and time ( )', 'Launch site', 'Payload', 'Orbit', 'Customer', 'Demo flights', 'logist
ent', 'In development', 'Retired', 'Cancelled', 'Spacecraft', 'Cargo', 'Crewed', 'Test vehicles', 'Current', 'Retired', 'Unflown
e', 'Related', 'General', 'General', 'People', 'Vehicles', 'Launches by rocket type', 'Launches by spaceport', 'Agencies, compan
```

GitHub URL:
[https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/week 1 - Data Collection with Web Scraping lab.ipynb](https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/week%201%20-%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb)

Data Wrangling



EDA

Summarizations

Creation of Landing Outcome Label

- Exploratory Data Analysis to find patterns in the data was performed

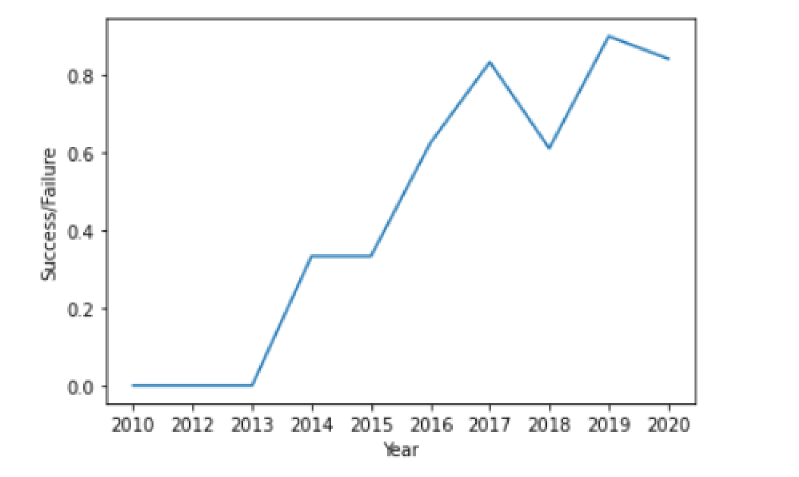
- Converting launch outcomes into Training Labels

- Translating outcomes into different column

GitHub URL:

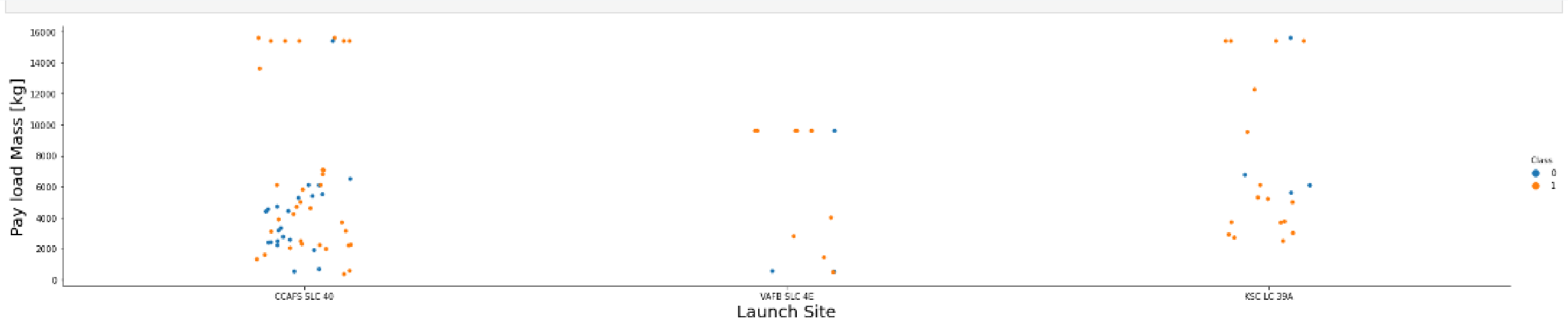
https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/week1%20-%20labs-jupyter-spacex-Data-wrangling.ipynb

EDA with Data Visualization



This chart shows that in later years there were more successful launches

Pay load Mass vs. Launch site chart shows that from VAFB SLC 4E site no heavy launches were made (above 10000 kg)



GitHub URL: https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission were obtainded
- Top 5 launch sites whose name begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA(CRS): 45596 [kg]
- Average payload mass carried by booster version F9 v1.1: 2928 [kg]
- Date when the first successful landing outcome in ground pad was achieved: 2010-06-04
- List the names of the boosters which have success in drone ship and have payload mass 4000 between 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20

GitHub URL:
https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/eda-sql-coursera-nopass.ipynb

Build an Interactive Map with Folium

Summary of map objects that were created:

- add folium.Circle and folium.Marker for each launch site on the site map

- Mark the success/failed launches for each site on the map

- Calculating the distances between a launch site to its proximities (adding lines to the map)

GitHub URL:
https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/Launch_Sites_Locations_Analysis_with_Folium%20(1).ipynb

https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/inter_map_with_Folium.pdf

Build a Dashboard with Plotly Dash

The next charts were created:

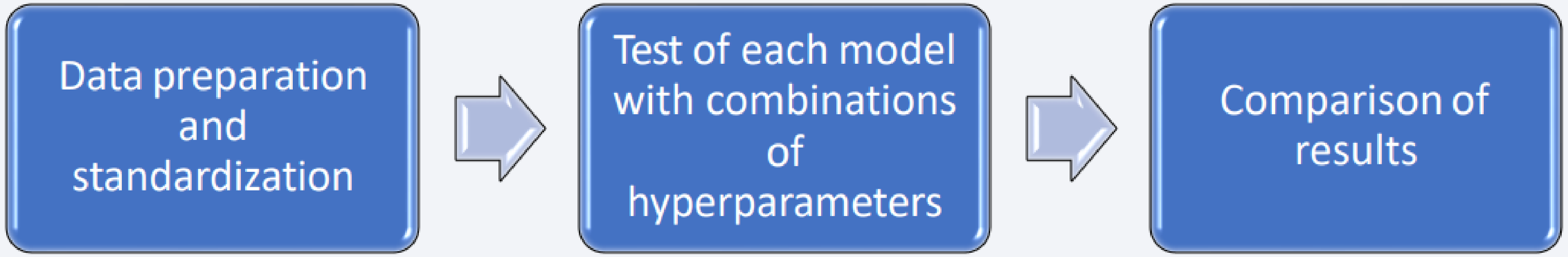
- success pie chart based on selected site: allows quickly understand the successful launches per site

- success payload scatter chart: allows to check success versus payload for different buster types

GitHub URL:
https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/Dash_app_week3_course10.ipynb

Predictive Analysis (Classification)

1. The data was scaled using standard scaler from scikit-learn package
2. The data was split into training and test sets
3. Four differnet classification algorithms were used
4. Accuracy of the algorithms was checked


Data preparation and standardization
Test of each model with combinations of hyperparameters
Comparison of results

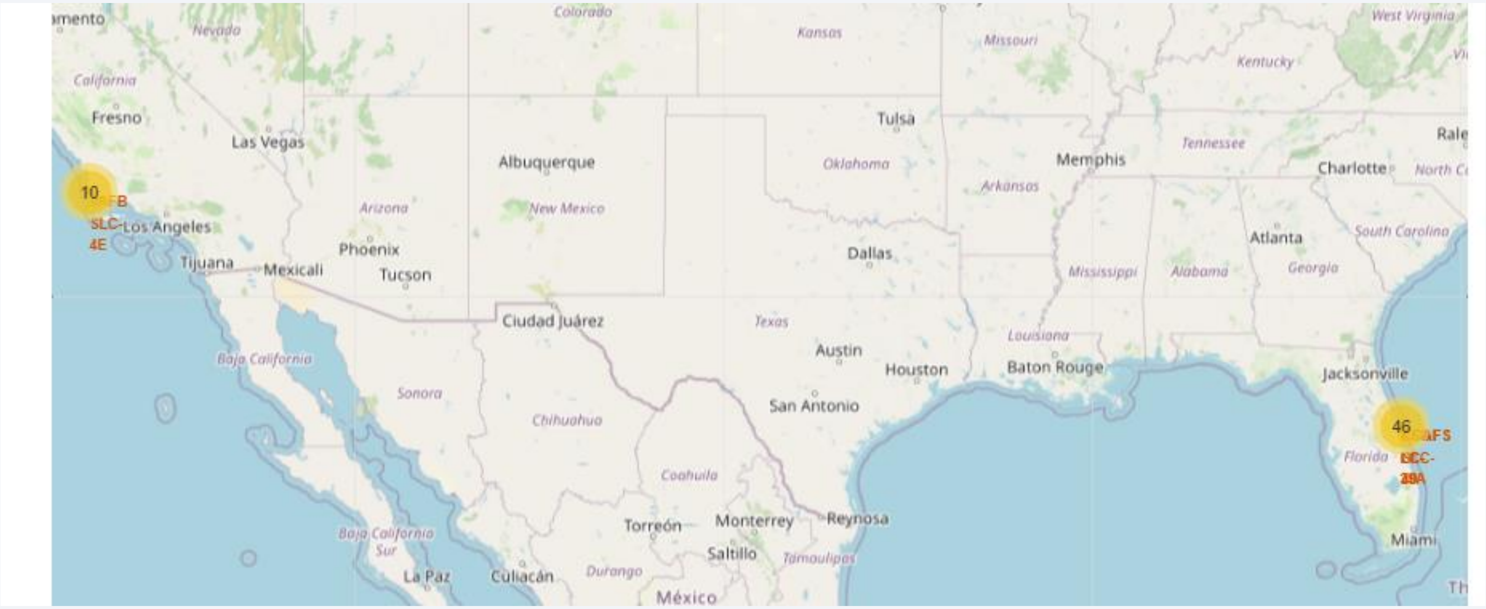
GitHub URL:
https://github.com/sbel12/Capstone_Project_DS_AB/blob/master/Labs_module_4_SpaceX_Machine_Learning_Prediction_v1.ipynb

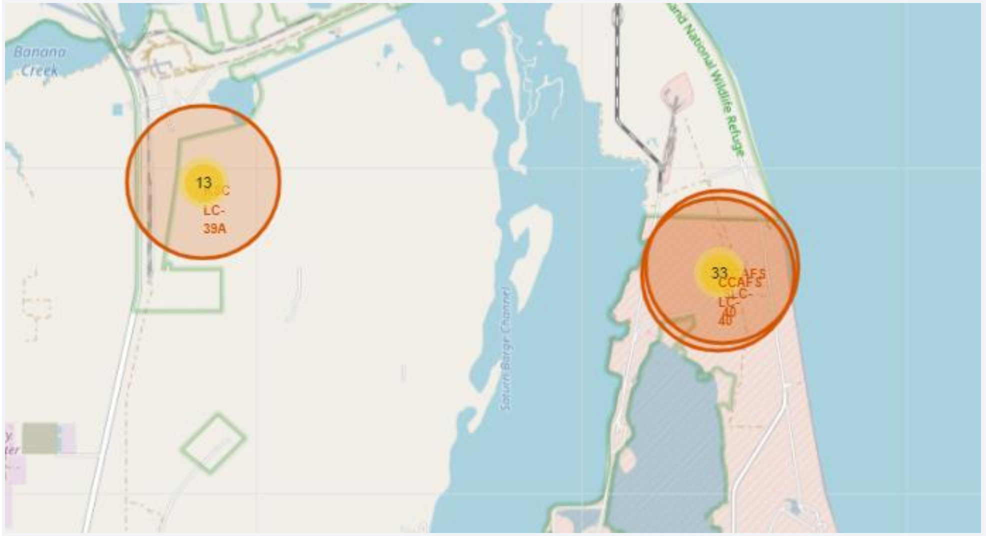
Results - EDA

Exploratory data analysis results:

- SpaceX launches from 4 different launch sites
- The first launches were performed from SpaceX and NASA sites
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 fiver year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average

Results – Interactive analyses





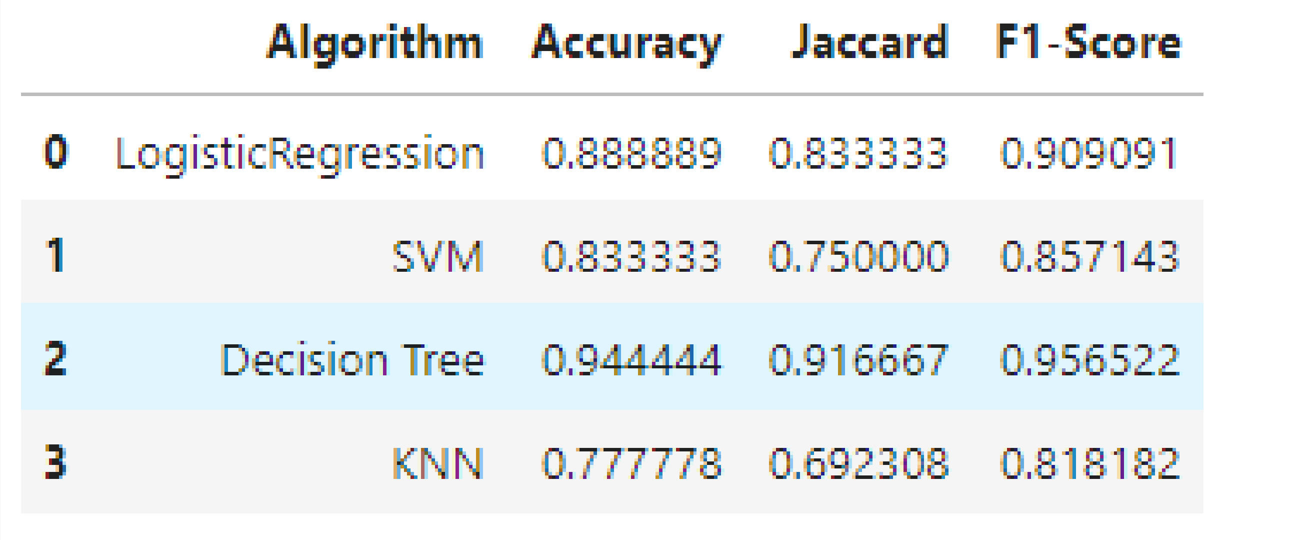
- all launch sites are near oceans, have a good safety distances, good infrastructure around them

- most launches happened at the east cost

Results – Predictive Analysis

- from 4 different algorithms, Decision Tree Classifier has the best model predicition

- this prediction are highly dependent on random state in data splitting

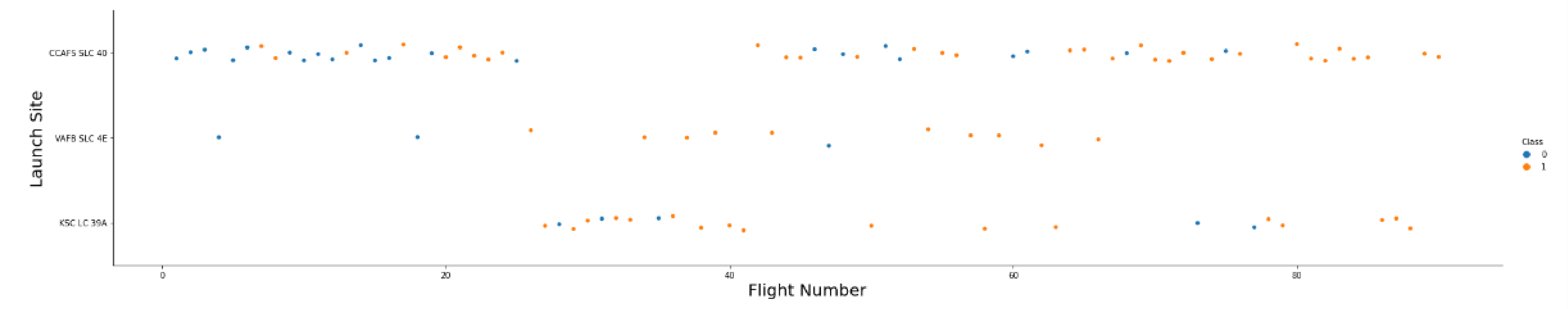


	Algorithm	Accuracy	Jaccard	F1-Score
0	LogisticRegression	0.888889	0.833333	0.909091
1	SVM	0.833333	0.750000	0.857143
2	Decision Tree	0.944444	0.916667	0.956522
3	KNN	0.777778	0.692308	0.818182

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

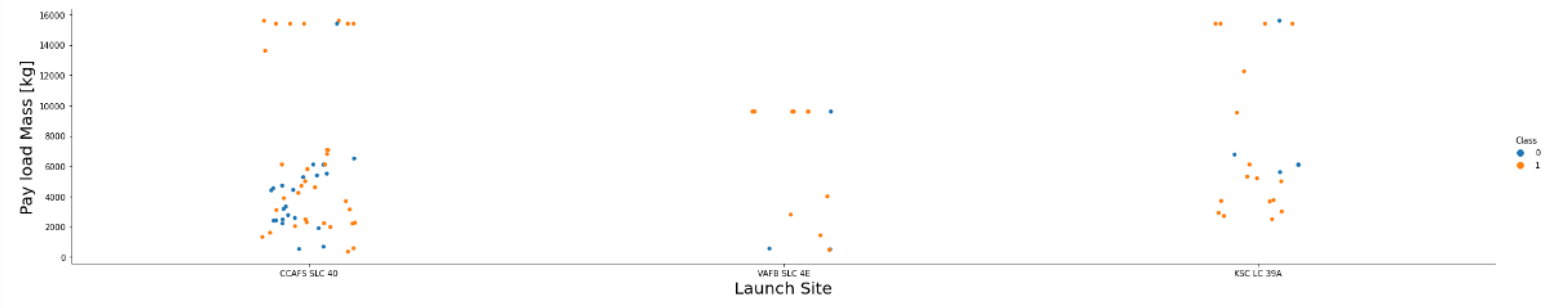


- the best launch site is CCAF5 SLC 40, where most of recent launches were successful

- CCAF5 SLC 40 launch site is most used

- latest launches have more success rate

Payload vs. Launch Site



- from VAFB SLC 4E site no heavy launches were made (above 12000 kg)

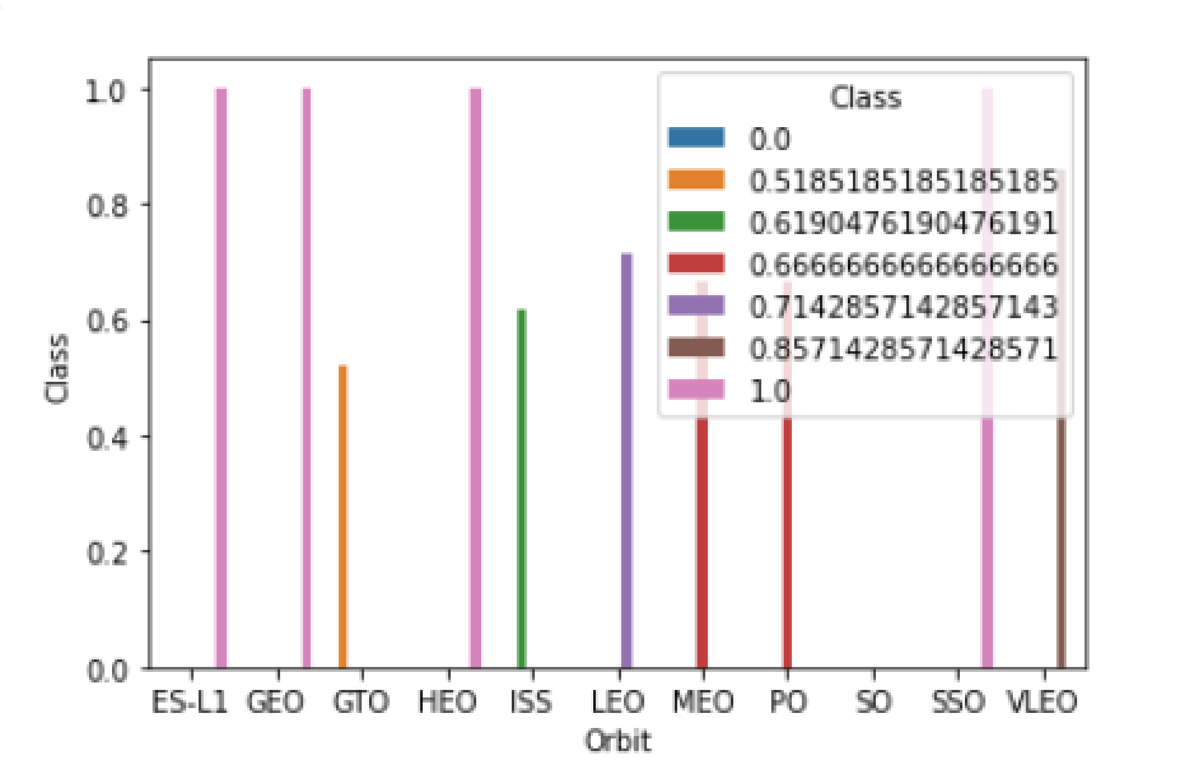
- only two of above 10000 kg launches were unsuccessful

Success Rate vs. Orbit Type

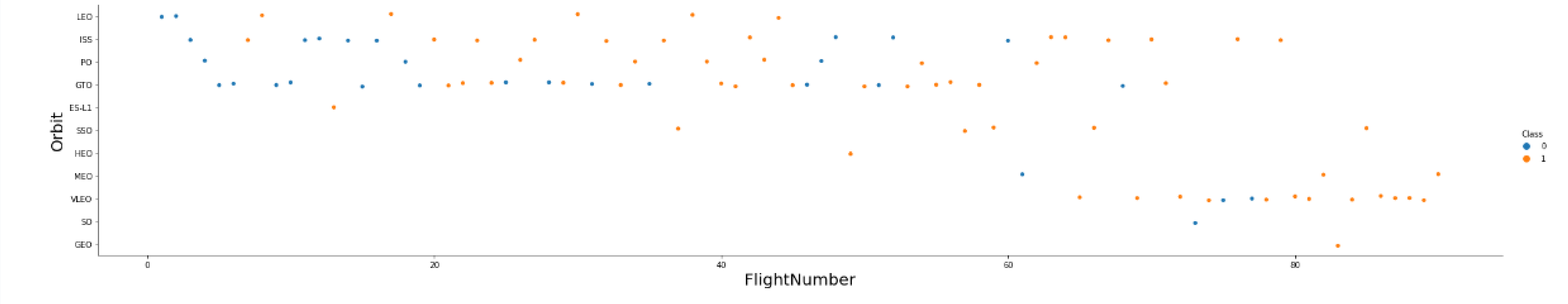
- Most successful orbits: ES-L1, GEO, HEO and SSO

- VLEO with ~86% of success

- Leo with ~71% of success

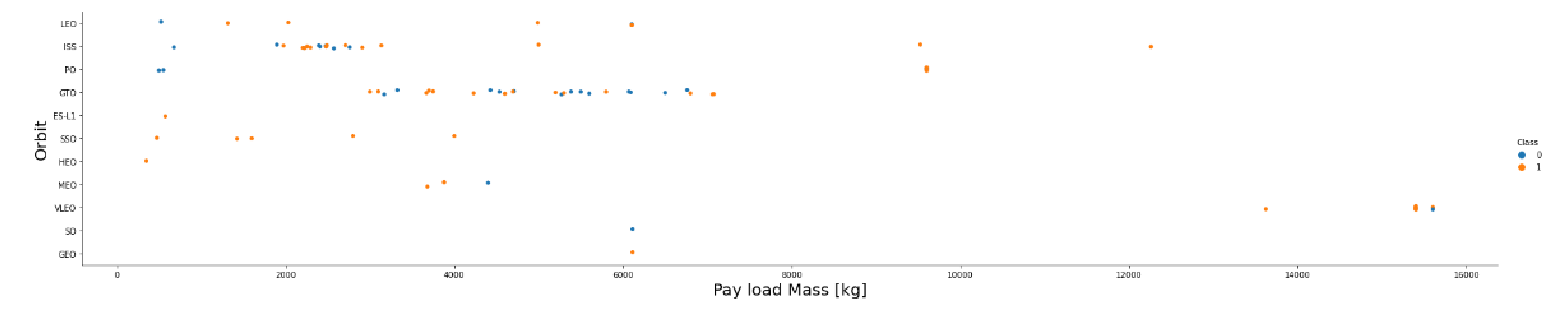


Flight Number vs. Orbit Type



- Latest launches are more successful
- VLEO orbit is mostly used in latest launches

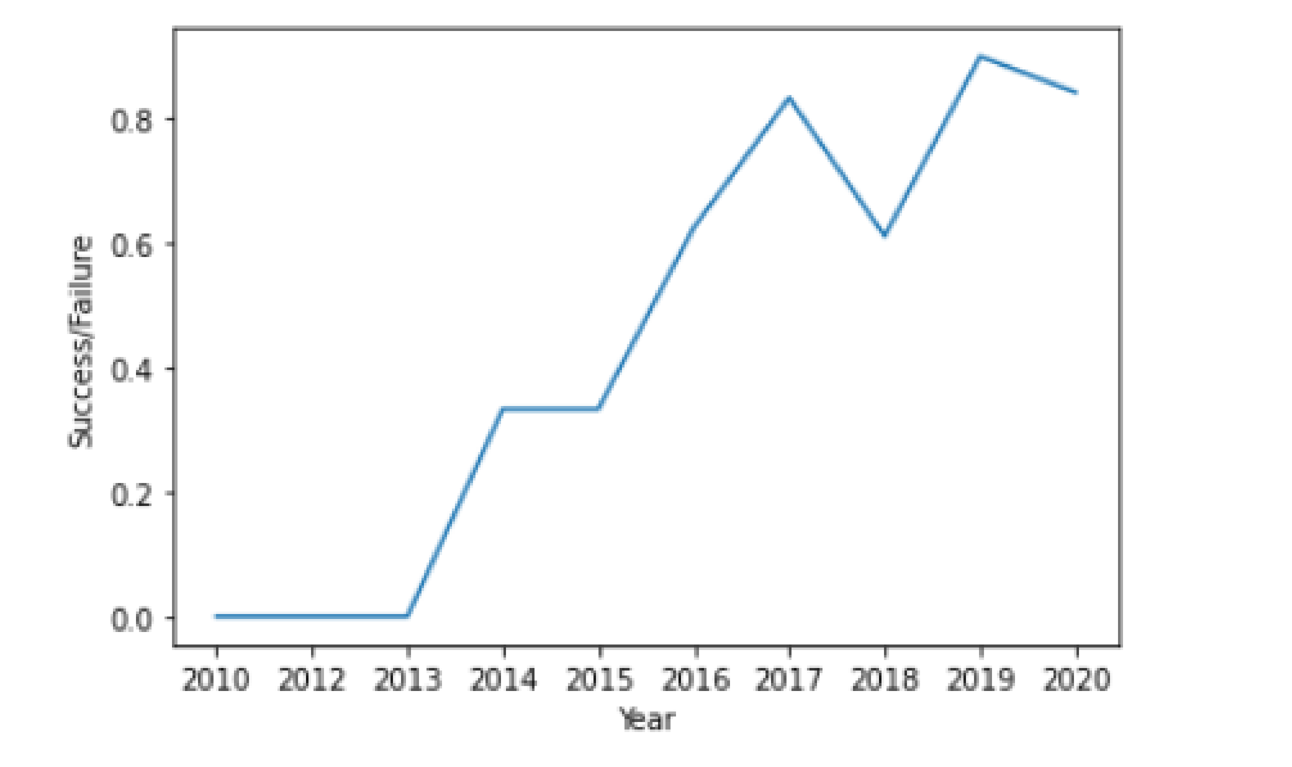
Payload vs. Orbit Type



- Only three orbits were used for heavy payloads (above 10000 kg)
- So and GEO orbits are less frequently used
- GTO and ISS orbits are very frequently used

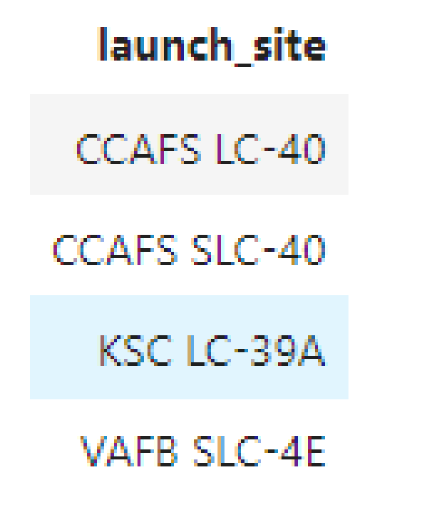
Launch Success Yearly Trend

- success rate constantly increases from 2013
- First three reported years were unsuccessful
- Latest years from 2015 are more successful



All Launch Site Names

According to the data, these are the launch sites that were used:

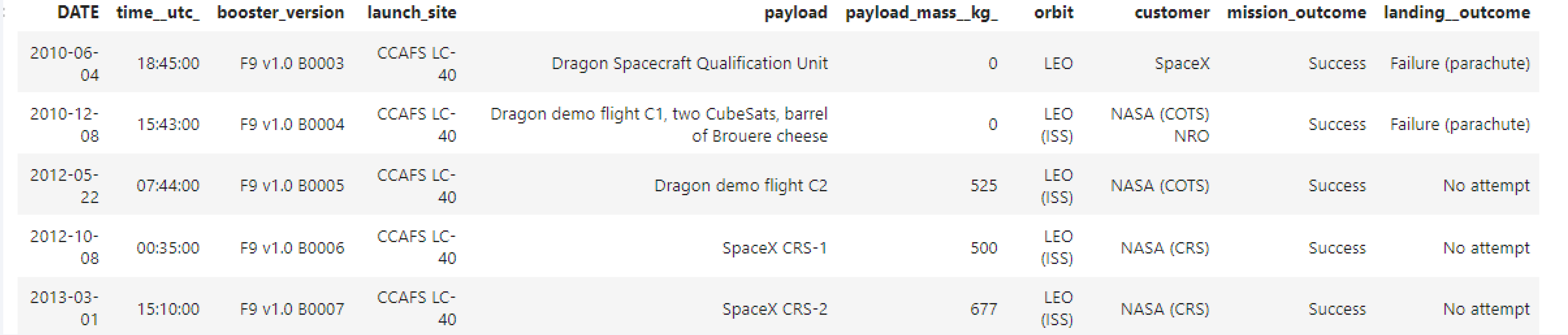


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

This results was obtained applying DISTINCT on LAUNCH_SITE column

Launch Site Names Begin with 'CCA'

These are 5 records where launch sites begin 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	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

This results was obtained using “like 'CCA%' LIMIT(5)”

Total Payload Mass

The total payload carried by boosters from NASA is: 45596 [kg]

This calculation was don by applying “SUM”

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928 [kg]

This calculation was don by applying “AVG “ command

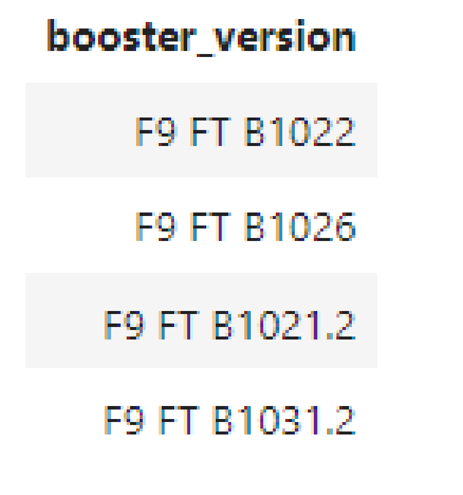
First Successful Ground Landing Date

The first successful landing outcome on ground pad was on 2010-06-04

This query was done by using “DATE” and limiting it to “ MISSION_OUTCOME='Success' order by Date asc LIMIT(1) “

Successful Drone Ship Landing with Payload between 4000 and 6000

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

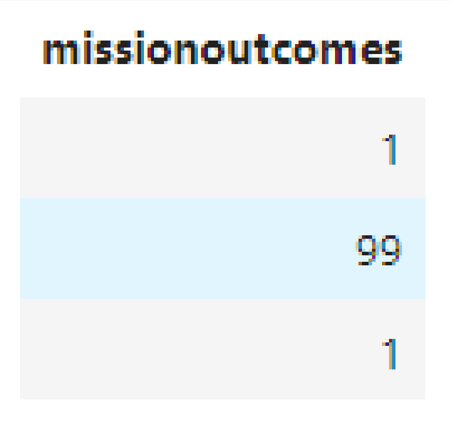


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

This query was done by limiting “PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000”

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes:

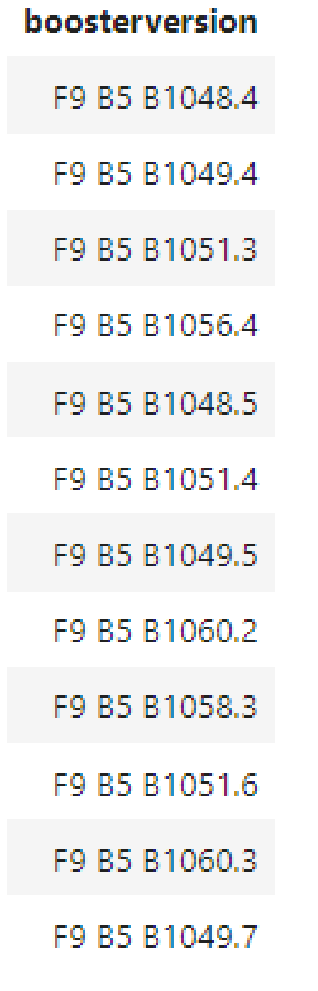


	missionoutcomes
Success (payload status unclear)	1
Success	99
Failure (in flight)	1

This query was done by using “GROUP BY”

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass:

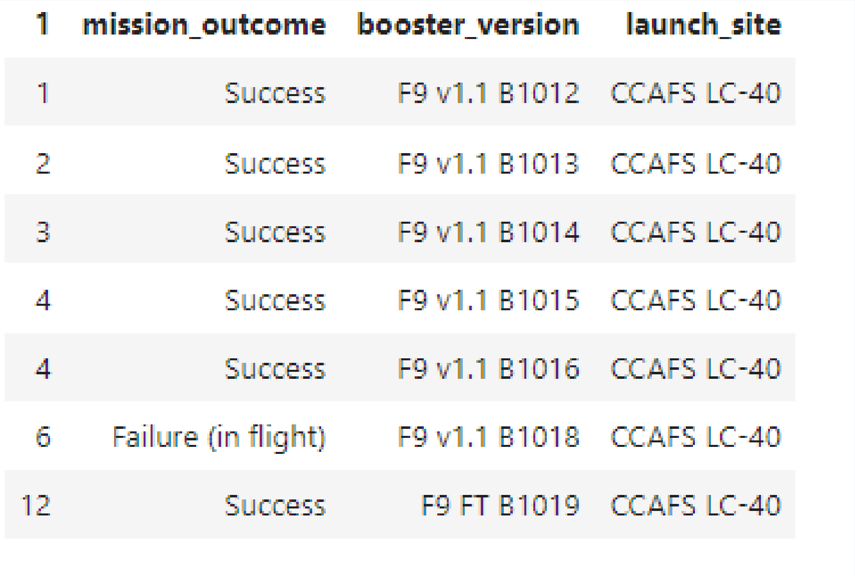


boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

This was done by using subquery

2015 Launch Records

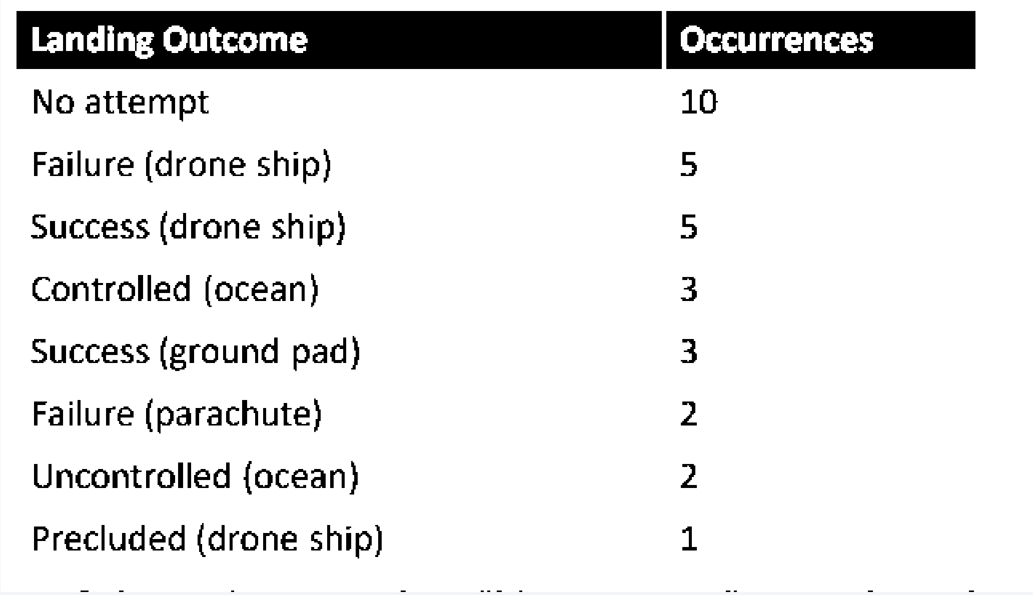
The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:



1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

Here is the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order



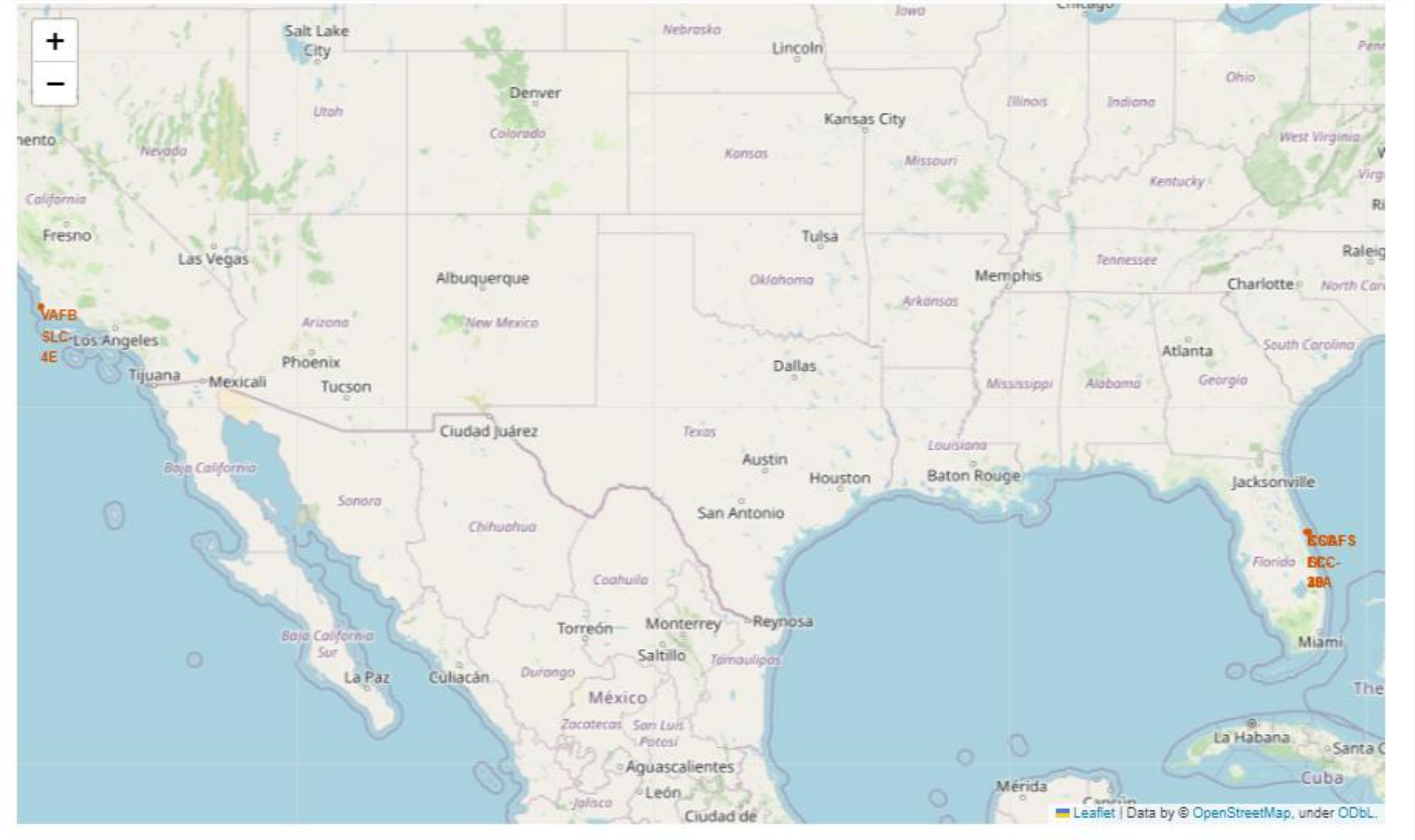
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

There is also “No attempt” occurences

Section 3

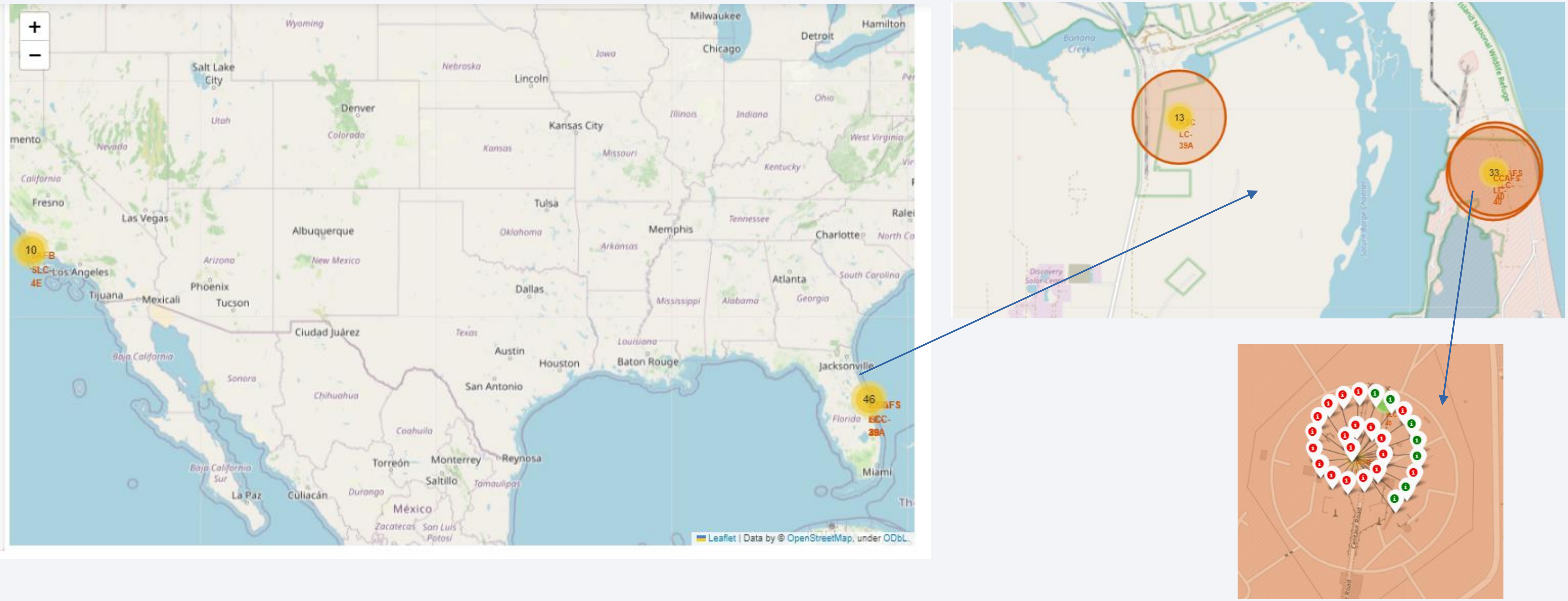
Launch Sites Proximities Analysis

All Launch Sites



All launch sites are near the oceans

Launch Outcomes per site



Green markers are successful launches, red are failures

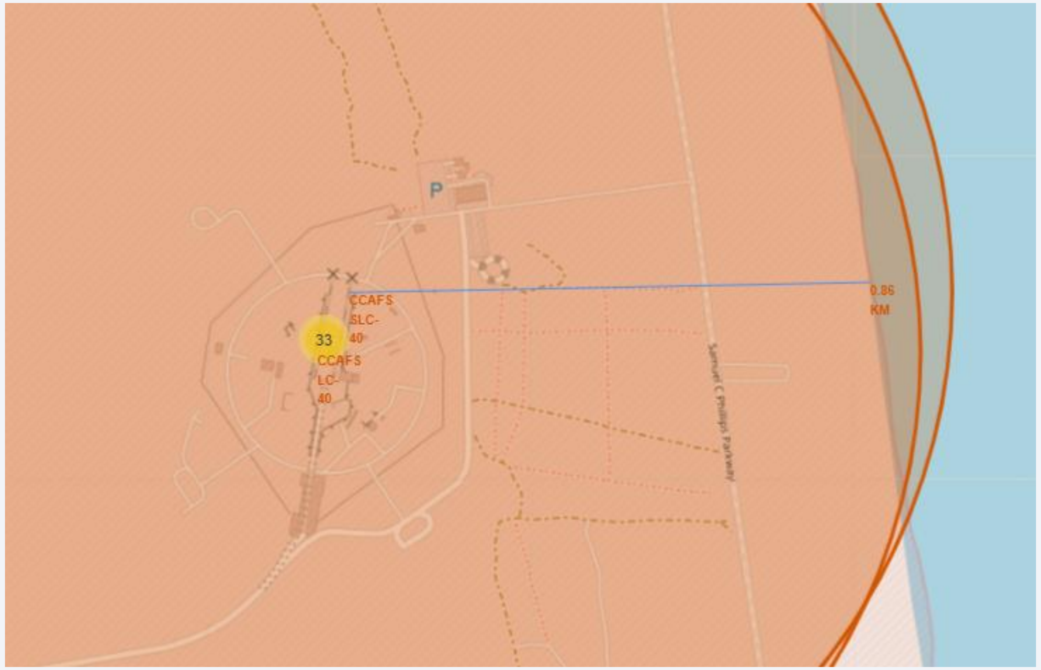
Measuring distance to different objects

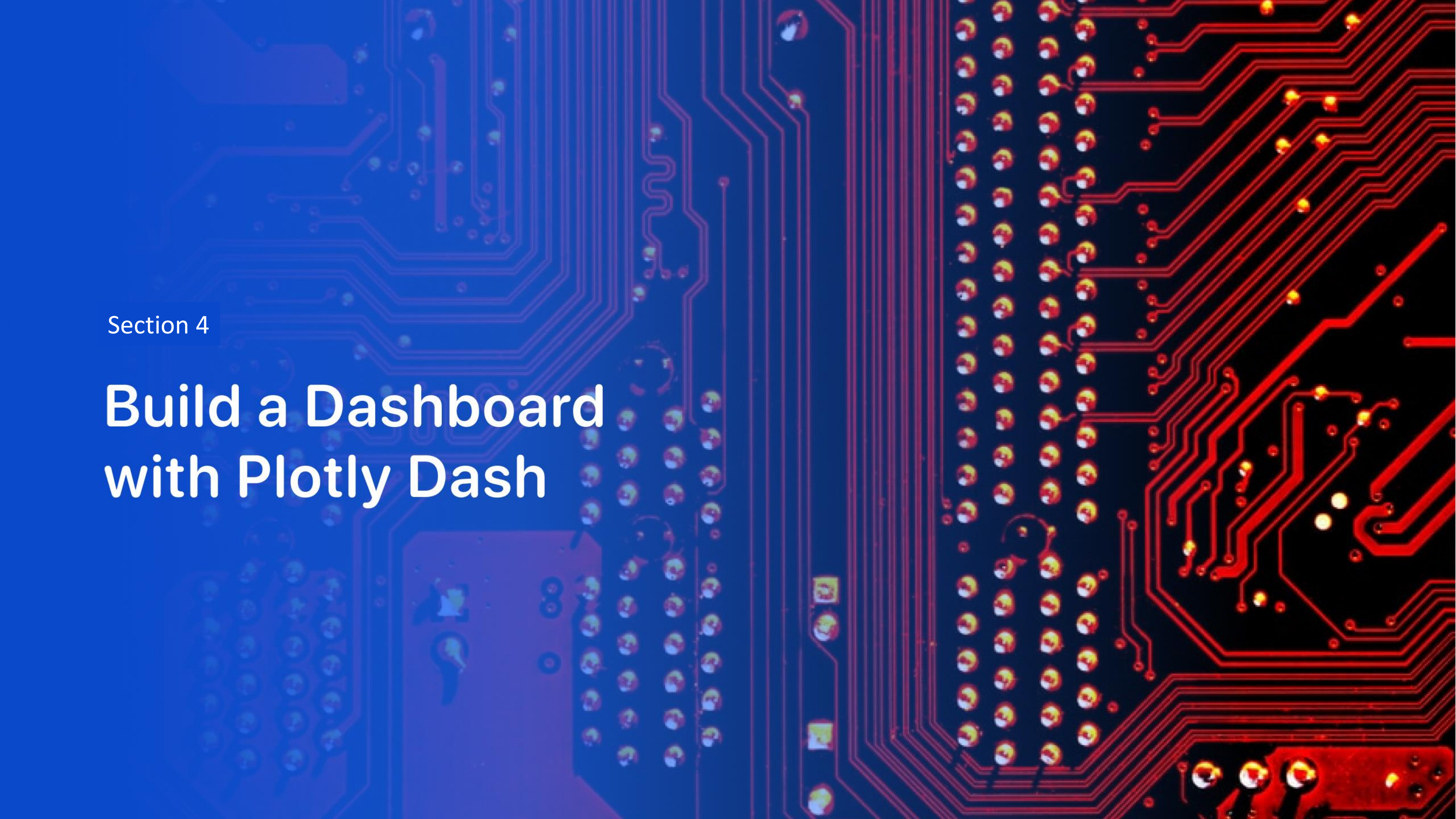
Distance to the coastline: 0.86 [km]

Distance to highway = 0.58 [km]

Distance to railroad = 1.28 [km]

Distance to nearest city = 51.4 [km]

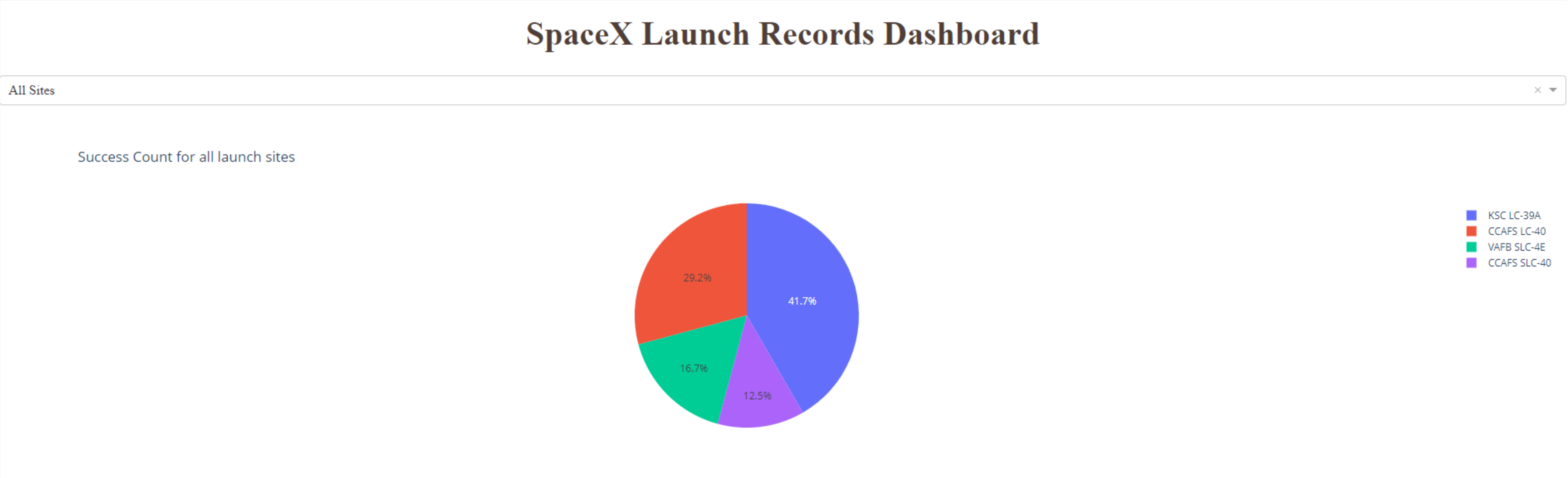




Section 4

Build a Dashboard with Plotly Dash

Success Launches Counts for all sites

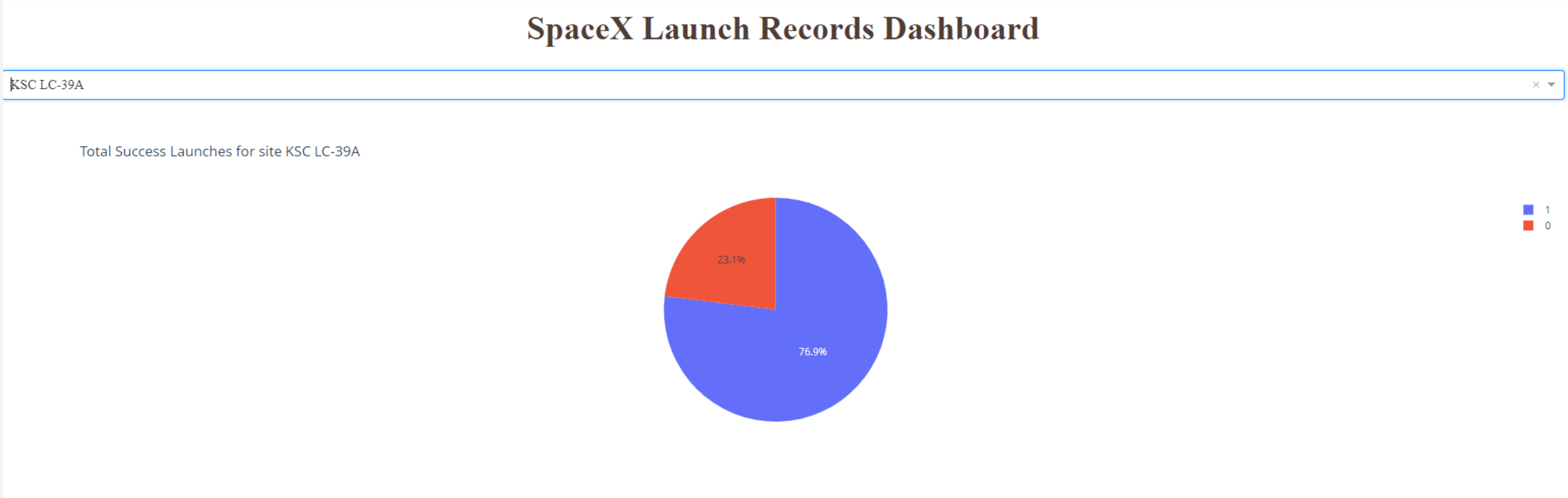


SpaceX Launch Records Dashboard

All Sites

41.7% of the successful launches were made from KSC LC-39A

Launches success ratio for KSC LC-39A



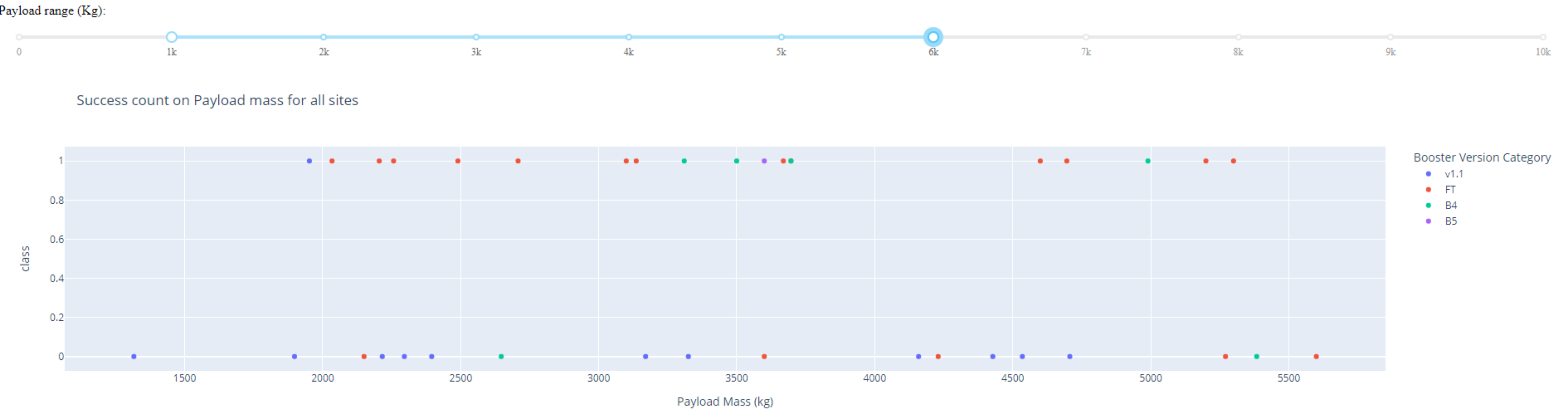
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Almost 77% are successful launches from this site

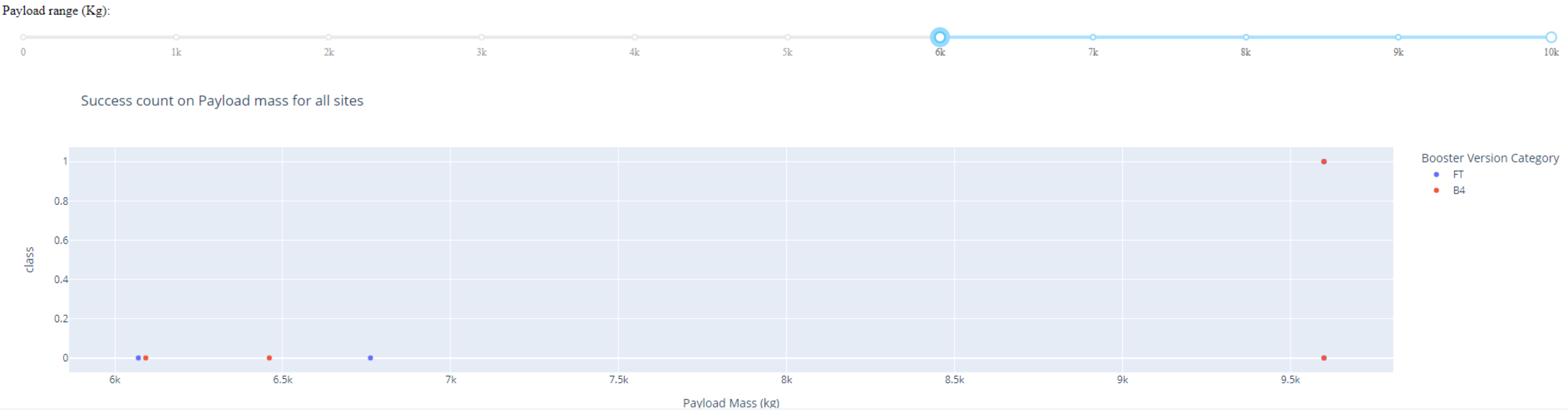
Payloads vs. Launch Outcome plots



Payload range (Kg):

FT booster for Payloads less then 6000 [kg] is the most successful

Payloads vs. Launch Outcome plots (cont.)



Above 6000 [kg] of payload there was only one successful launch

Section 5

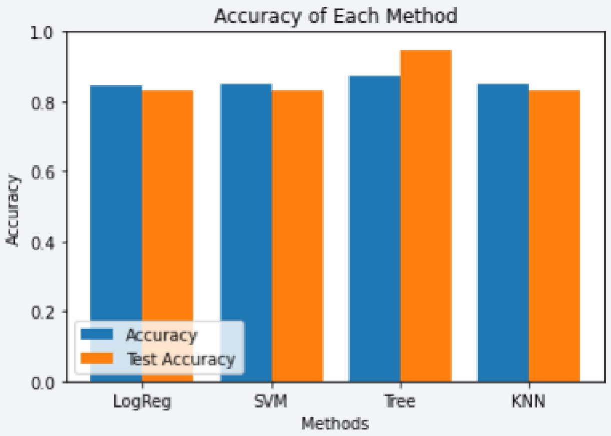
Predictive Analysis (Classification)

Classification Accuracy

- Four different classification algorithms were applied on the dataset

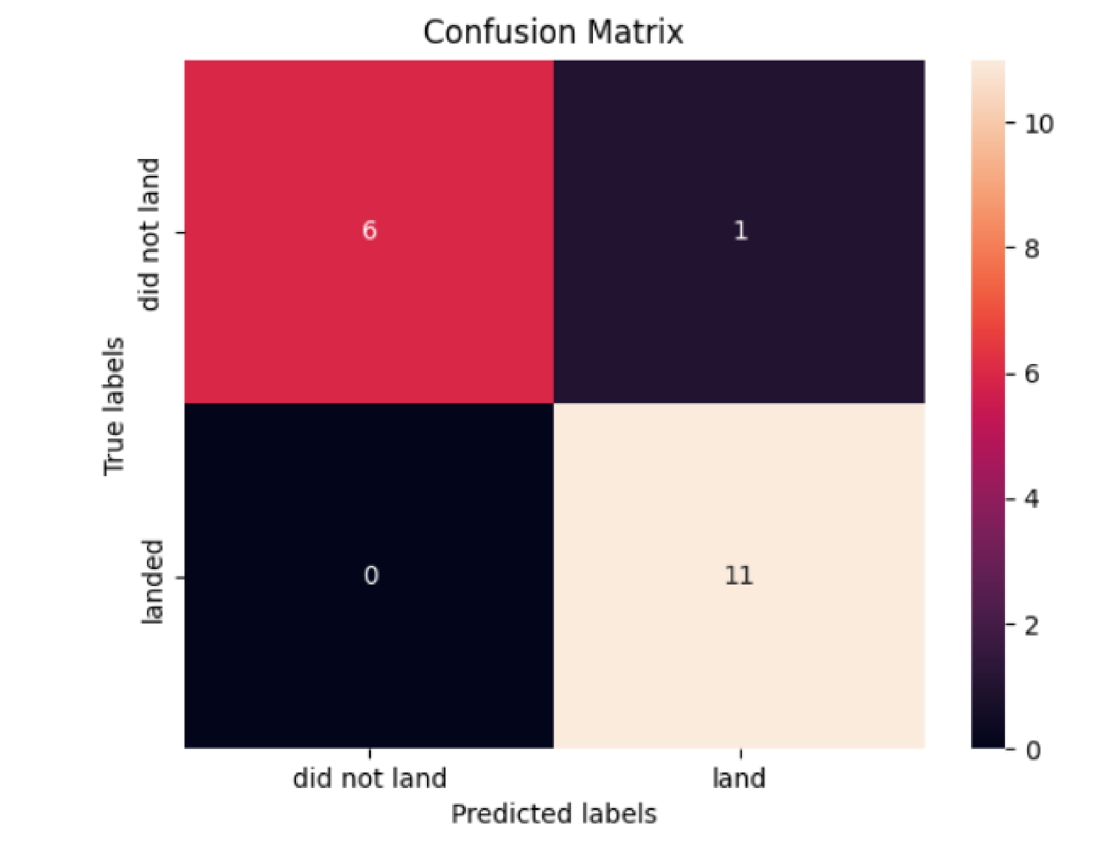
- Decision Tree Classifier has the highest accuracy of 94.4%

- The results are highly dependent on random state in splitting data stages



Confusion Matrix for Decision Tree Algorithm

- Only one instance was classified as False Negative by Decision Tree algorithm
- All others were classified correctly



Conclusions

- Success rate constantly increases from 2013

- All launch sites are near the oceans

- Most successful orbits: ES-L1, GEO, HEO and SSO

- Decision Tree Classifier has the highest accuracy of 94.4%

- The results are highly dependent on random state in splitting data stages

- More points are available throughuot the previous pages

Appendix

- Folium maps are not shoun properly on GitHub, so there is an extra pdf file of a notebook with maps shown explicitly

Thank you!