IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sager Dave Sircar
29/Nov/2022

Outline

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

Image Source: SpaceX

Executive Summary

Methodology

- Data source: public SpaceX API & Wikipedia's list of SpaceX launches
- Creation of a 'class' label for successful / failed landings
- Data exploration using visualization, SQL, Folium maps & Plotly dashboard
- Predictive model built using cross sampling (GridSearchCV) using
  - logistic regression
  - Support Vector Machine
  - Decision Tree
  - K-nearest neighbor

Results

- All models: Accuracy 83.333 %
- Predictions identical (no exception)

Image Source: Bruno Kelzer

Introduction

Project background and context

- Space X has best pricing ($62 million vs. $165 million US-$)
- Largely due to ability to recover parts of their spacecrafts (Stage 1)
- Space Y wants to compete with Space X

Task given by SpaceY:

- Prediction model for successful stage 1 recovery

Section 1

Methodology

Methodology

Data collection through SpacX's web API & Wikipedia's list of past launches

Data wrangling

- Filtered for Falcon 9 launches
- Missing payloads are replaced by mean values & multiple payloads are ignored

Exploratory data analysis (EDA) using visualization and SQL

Interactive visual analytics using Folium and Plotly Dash

Predictive analysis using classification models

- Build with Logistic Regression, SVM, Decision Tree & k-NN
- Tuned using GridSearch CV
- Evaluate by accuracy and Confusion Matrix

Data Collection Sources

Online collection through SpacX's web API (https://api.spacexdata.com/v4/launches/past)

Web scraping Wikipedia's list of past launches (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

- Collection with SpaceX REST calls using requests.get(URL)
- Data Columns: 'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'

GitHub URL:

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/01-%20Data%20Collection%20API.ipynb

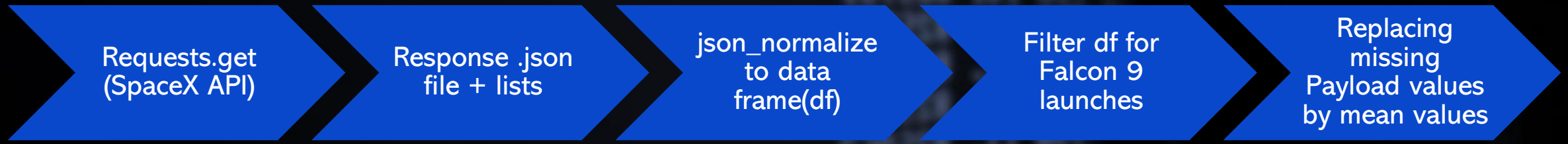


Image Source: Vitali Vlasov

Data Collection – Scraping

- Creation of a BeautifulSoup object after requests.get(URL)
- Data Columns: 'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Version Booster', 'Booster landing', 'Date', 'Time'

GitHub URL:

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/02-%20Data%20Collection%20WebScraping.ipynb

Image Source: Vitali Vlasov

Data Wrangling

Created a new label named ‘class’

- Successful landings = 1 / failures and no landing attempts = 0
- Generated from label ‘Outcome’
  - Successful landing tags (aka 1): True ASDS, True RTLS, & True Ocean
  - Other landing tags (aka 0): None None, False ASDS, None ASDS, False Ocean, False RTLS

Github URL:

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/03-%20Data%20Wrangling%20Lab.ipynb

Image Source: SpaceX

EDA with Data Visualization

Scatter plots:

- Flight Number / Payload Mass / class
- Flight Number / Launch Site / class
- Payload / Launch Site / class
- Flight No. / Orbit / class
- Payload / Orbit / class

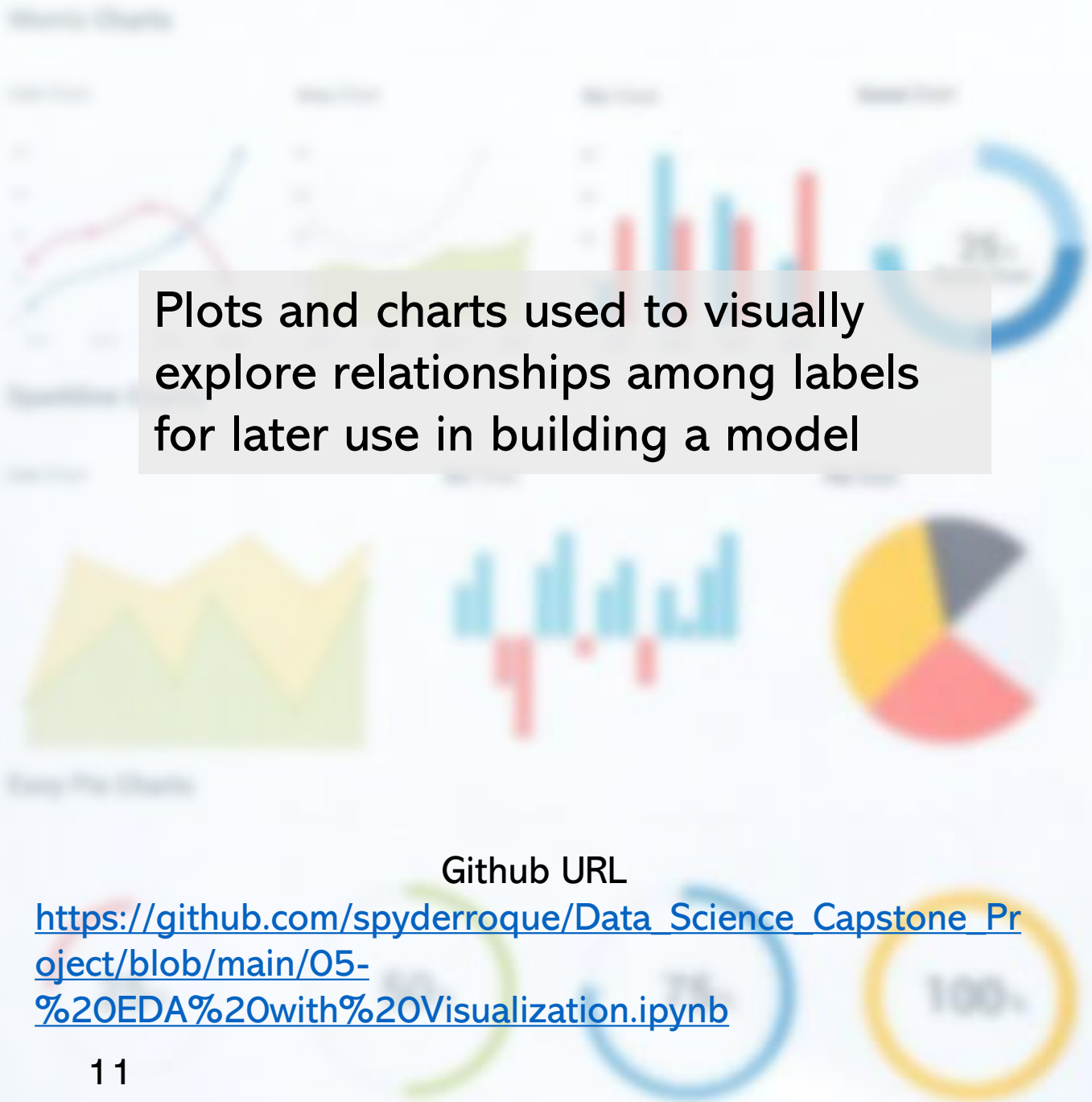
Line chart

- Year / Success rate

Bar plots:

- Payload orbit to Success rate

Plots and charts used to visually explore relationships among labels for later use in building a model



Github URL

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/05-%20EDA%20with%20Visualization.ipynb

EDA with SQL

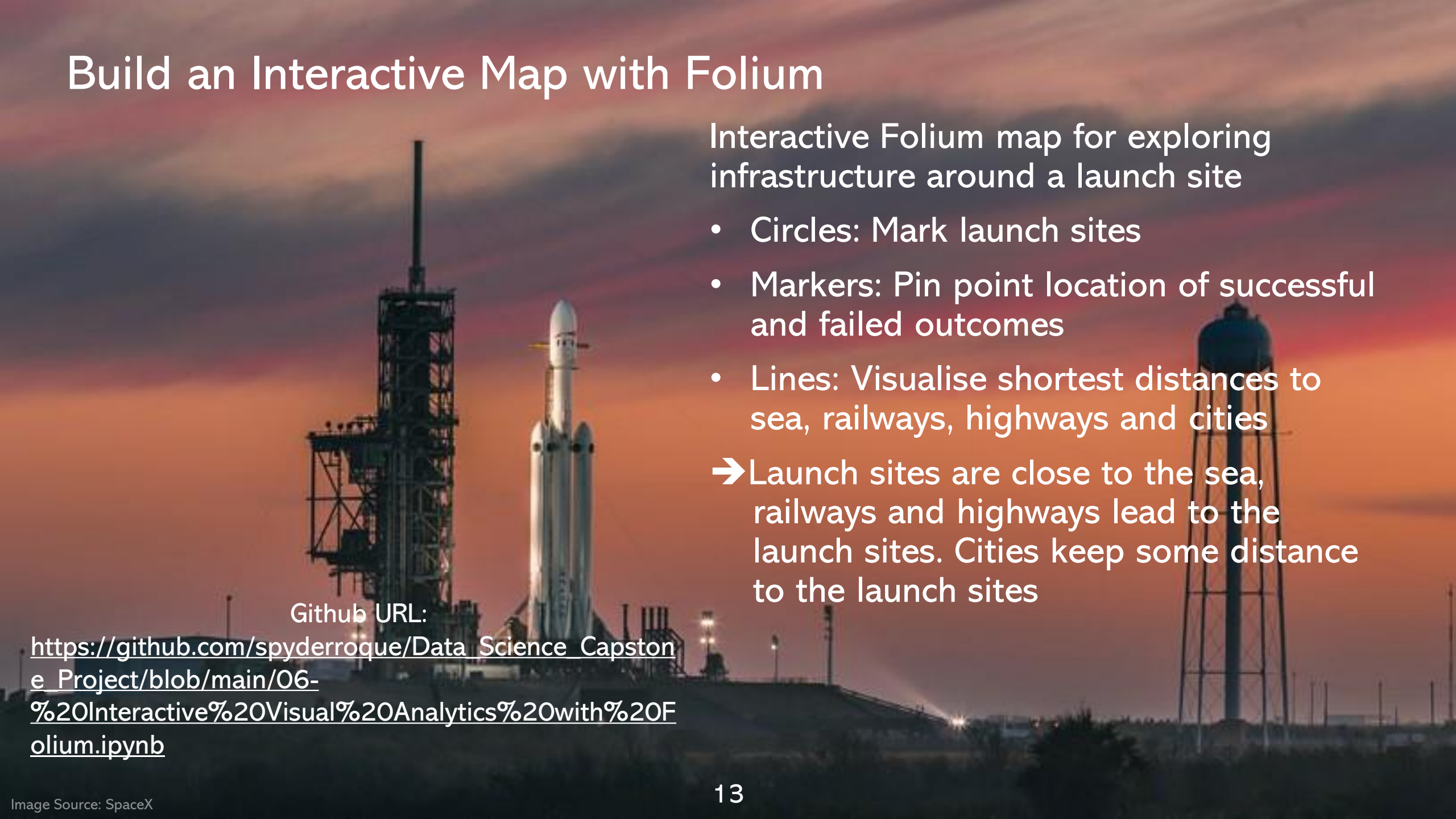
SQL query summary:

- (Upload csv file to DB2 & correct data type)
- Display names of launch sites
- Display total payload carried for NASA
- Average payload by booster version ‘F9 v1.1’
- First successful landing on a ground pad
- Booster versions successful landed on drone ships
- Total no. failed / successful mission outcomes
- Name of booster versions, which carried max. payload
- Failed landings / booster version / launch sites since 2015
- Count of landing outcomes

Github URL

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/04-%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium



Interactive Folium map for exploring infrastructure around a launch site

- Circles: Mark launch sites
- Markers: Pin point location of successful and failed outcomes
- Lines: Visualise shortest distances to sea, railways, highways and cities

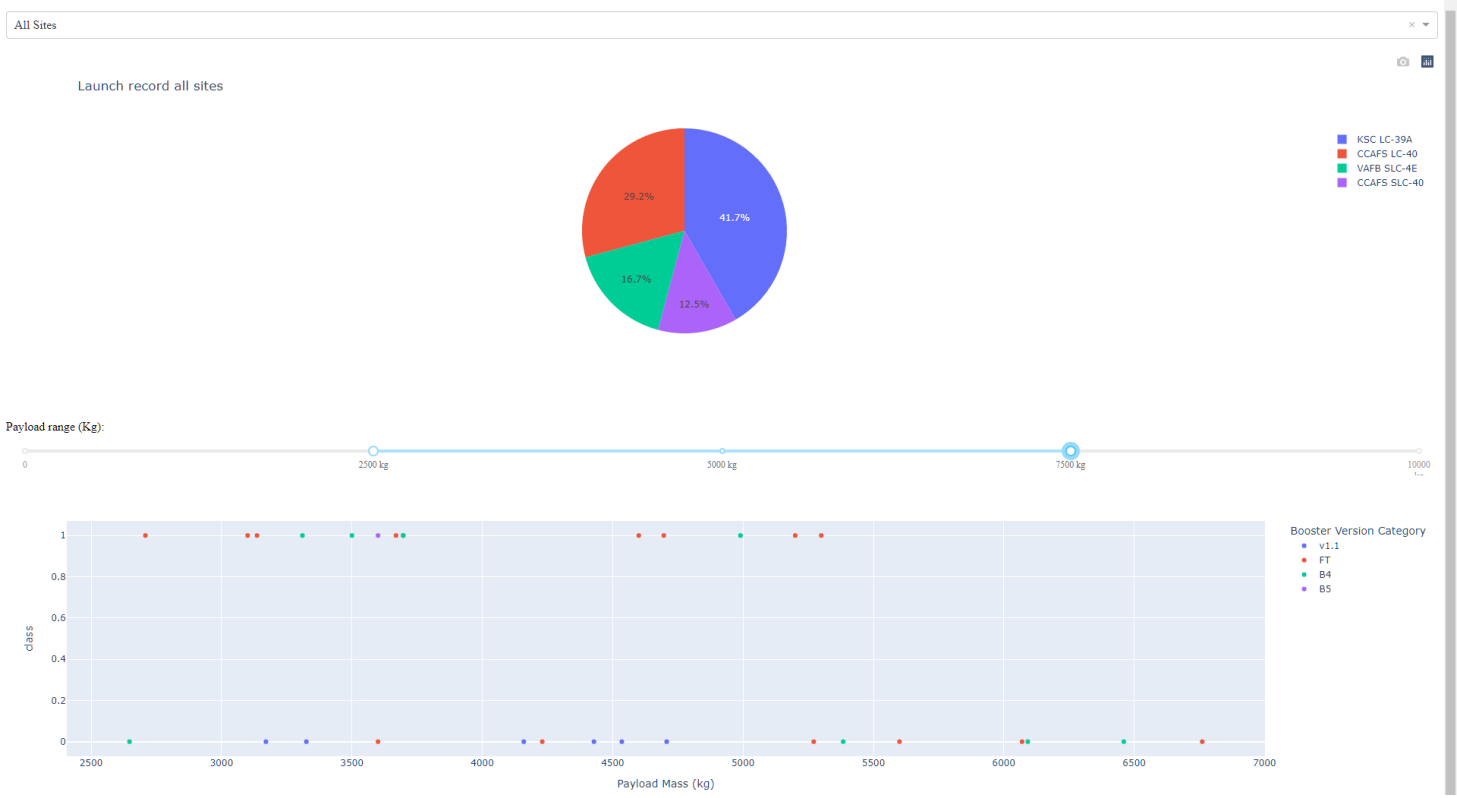
➔Launch sites are close to the sea, railways and highways lead to the launch sites. Cities keep some distance to the launch sites

Github URL: https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/06-%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Image Source: SpaceX

Build a Dashboard with Plotly Dash

Github url
https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/07-%20spacex_dash_app.py



Elements:

- Pie chart: Successfull landings rate all / specific launch sites
- Scatterplot: Payload (x-Axis), Success (y-Axis), Booster Version (colour)
- Selection tools: Drop down: Choose launch sites; Slider: Constrict payload mass

Motivation:

- Help explore data more convienently and show progress to customer

Predictive Analysis (Classification)

GitHub URL:

https://github.com/spyderroque/Data_Science_Capstone_Project/blob/main/08-%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Image Source: Vitali Vlasov

Results

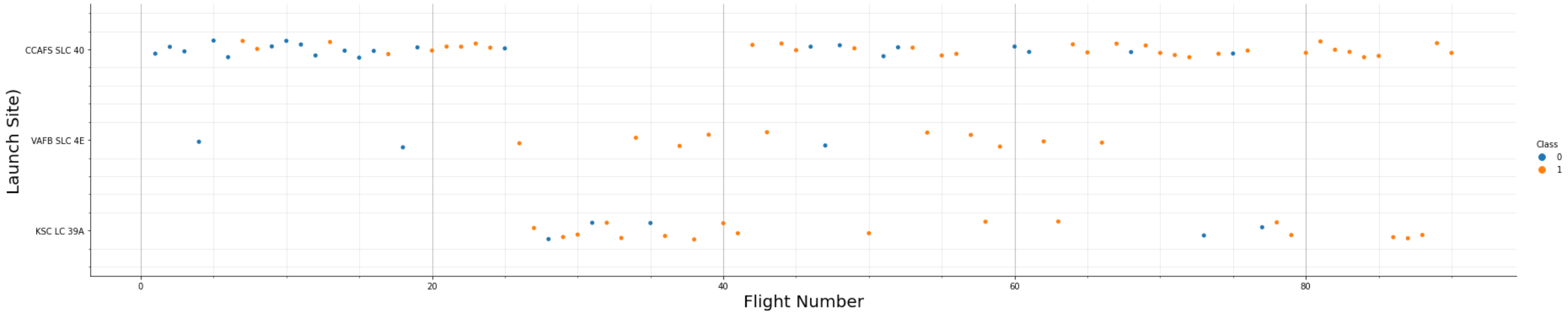
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

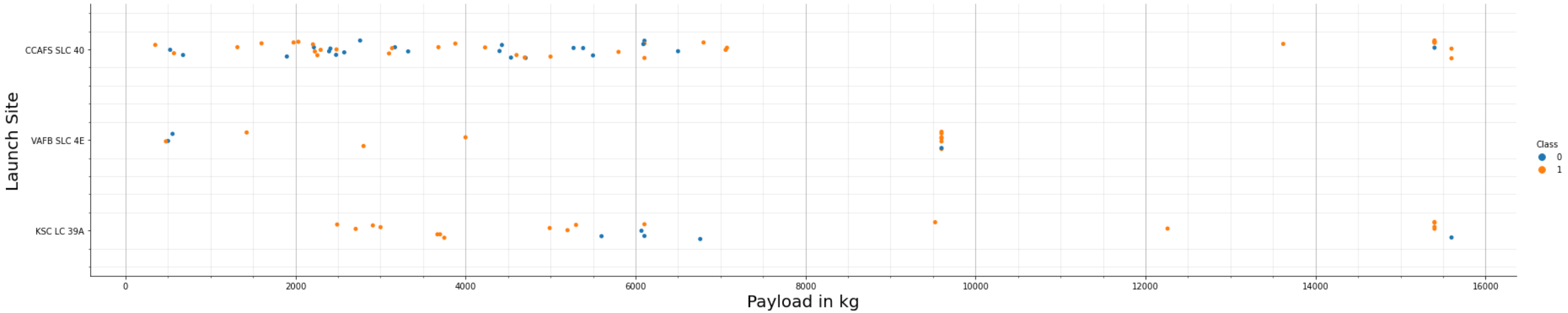
Flight Number vs. Launch Site

- Increase in success rate over time (indicated in Flight Number)
- Possible breakthrough at flight 20 increased success rate+
- CCAFS preferred launch site



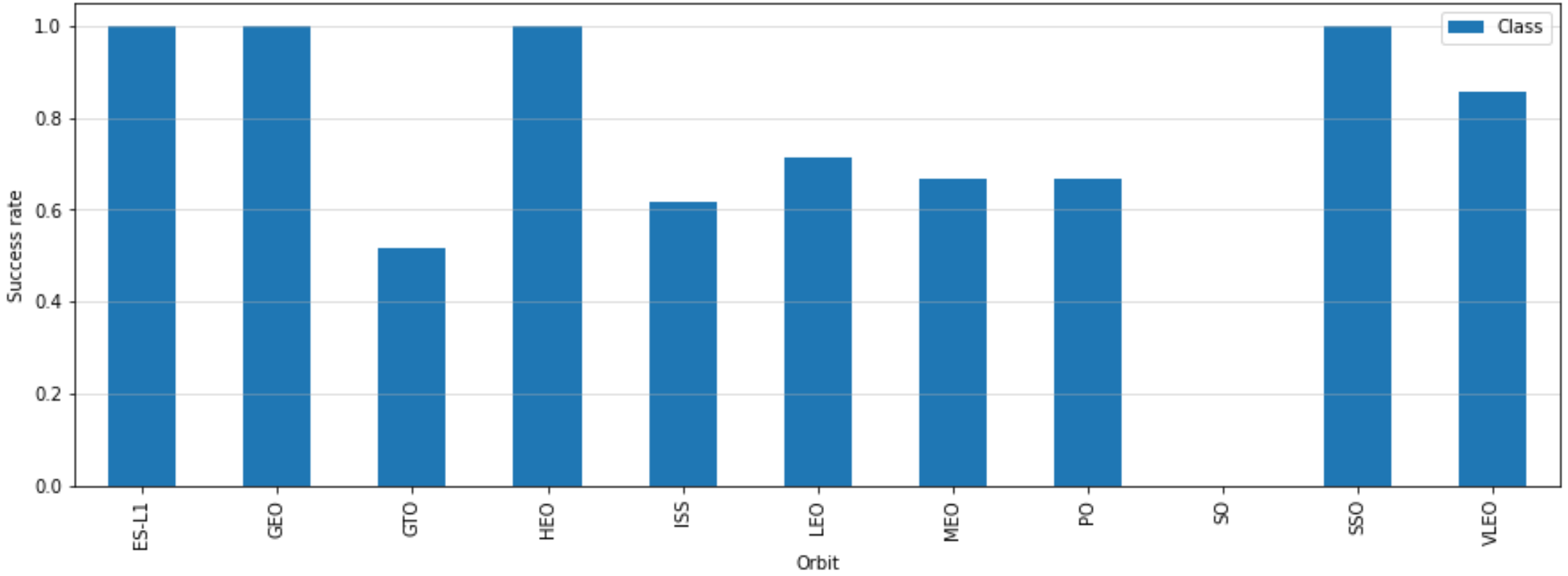
Payload vs. Launch Site

- Payloads > 9000 seem to correlate with launch sites
- Payloads above 7000 kg rare
- CCAFS & KSC LC 39A preferred launch site till 7000 kg



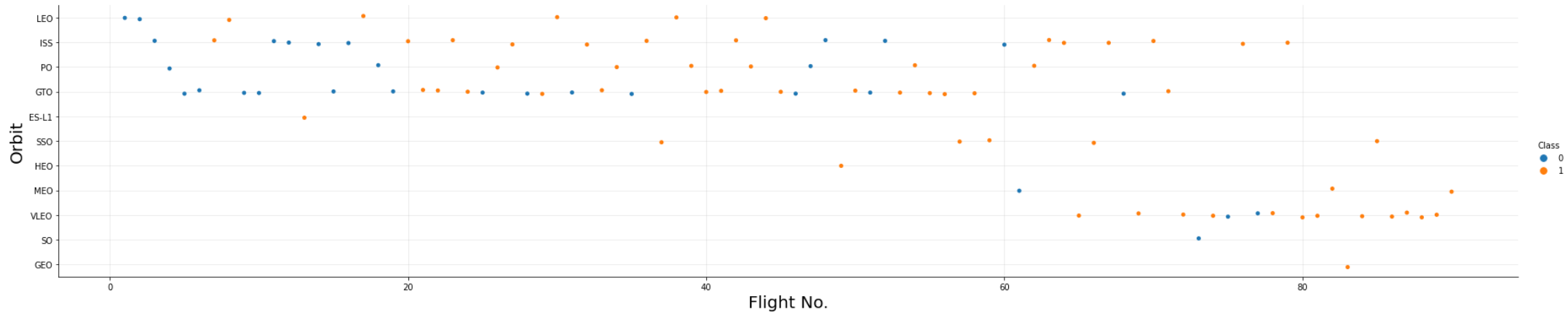
Success Rate vs. Orbit Type

- L1, GEO, HEO and SOO have very high success rates



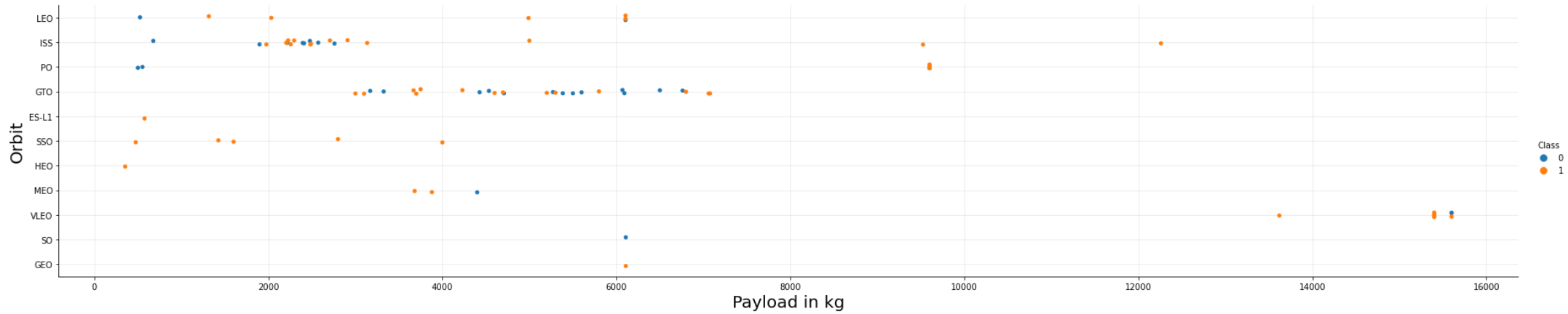
Flight Number vs. Orbit Type

- Success rate of LEO orbit related to the number of flights
- No relationship between flight number and GTO orbit discernable



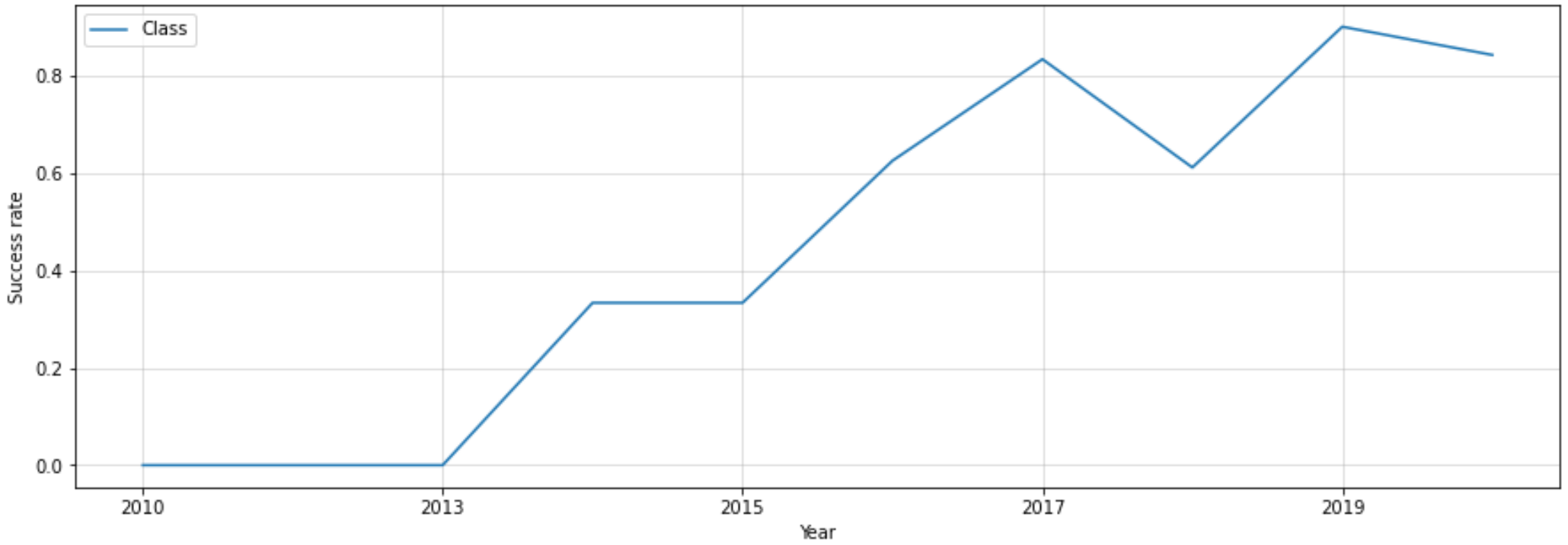
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- For GTO we cannot distinguish this well, as both positive landing rate and unsuccessful missions are both there



Launch Success Yearly Trend

- Sucess rate since 2013 kept increasing till 2020



All Launch Site Names

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

Query for unique launch site names

- CCAFS LC and SLC probably same location (new and old name)

→ 3 launch sites available

Launch Site Names Begin with 'CCA'

DATE	time__utc_	booster_ version	launch_ site	payload	payload _mass_ _kg_	orbit	customer	mission_ outcome	landing__outco me
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

First 5 entries for the launch sites starting with ‚CCA‘

Total Payload Mass

```
%sql select customer, sum(PAYLOAD_MASS__KG_) as Total_Payload_kg from spacextbl group by customer having customer = 'NASA (CRS)';
```

customer	total_payload_kg
NASA (CRS)	45596

This is total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql select BOOSTER_VERSION, avg(PAYLOAD_MASS__KG_) as Avg_Payload_kg from SPACEXTBL group by BOOSTER_VERSION having BOOSTER_VERSION = 'F9 v1.1';
```

booster_version	avg_payload_kg
F9 v1.1	2928

This is average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

Is 22nd / Dec / 2015

```
%sql select min (DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)';
```

Ground landing was not the first successful landing, apparently landing on drone ships is safer and less challenging

Successful Drone Ship Landing with Payload between 4t and 6t

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

```
%sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXTBL where LANDING__OUTCOME =
'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000 ;
```

This query reuturns the version of the booster for the 4 flights in this range.

No additonal inisghts gained here.

Total Number of Successful and Failure Mission Outcomes

mission_outcome	total_no
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Total_No from SPACEXTBL group by MISSION_OUTCOME;
```

SpaceX has a 98 % success outcome for it‘s missions. This is a good rate.

Boosters Carried Maximum Payload

booster_version

F9 B5 B1048.4

F9 B5 B1049.4

F9 B5 B1051.3

F9 B5 B1056.4

F9 B5 B1048.5

F9 B5 B1051.4

F9 B5 B1049.5

F9 B5 B1060.2

F9 B5 B1058.3

F9 B5 B1051.6

F9 B5 B1060.3

F9 B5 B1049.7

```
%sql select booster_version from Spacextbl where payload_mass__kg_=(select
MAX(payload_mass__kg_) from Spacextbl);
```

These booster versions carried the max payload of 15,600 kg. Apparently the Booster type F9 B5 has enough safety margin to carry this load safely.

```
%sql select MAX(payload_mass__kg_) from Spacextbl;
returns 15600
```

2015 Launch Records

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

```
%sql select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, Date from SPACEXTBL where year(DATE) = 2015
and LANDING__OUTCOME like '%Failure (drone ship)%';
```

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site of 2015 launches where stage 1 failed to land on a drone ship.

→ 2 failures.

→ At a success rate of 33% this means: 1 success

Ranked Landing Outcomes 2010-06-04 and 2017-03-20

landing__outcome	total
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

```
%sql select LANDING__OUTCOME, count(LANDING__OUTCOME) as Total from
SPACEXTBL where DATE between '2010-06-04'and'2017-03-20' group by
(LANDING__OUTCOME) order by Total desc;
```

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site of 2015 launches where stage 1 failed to land on a drone ship.

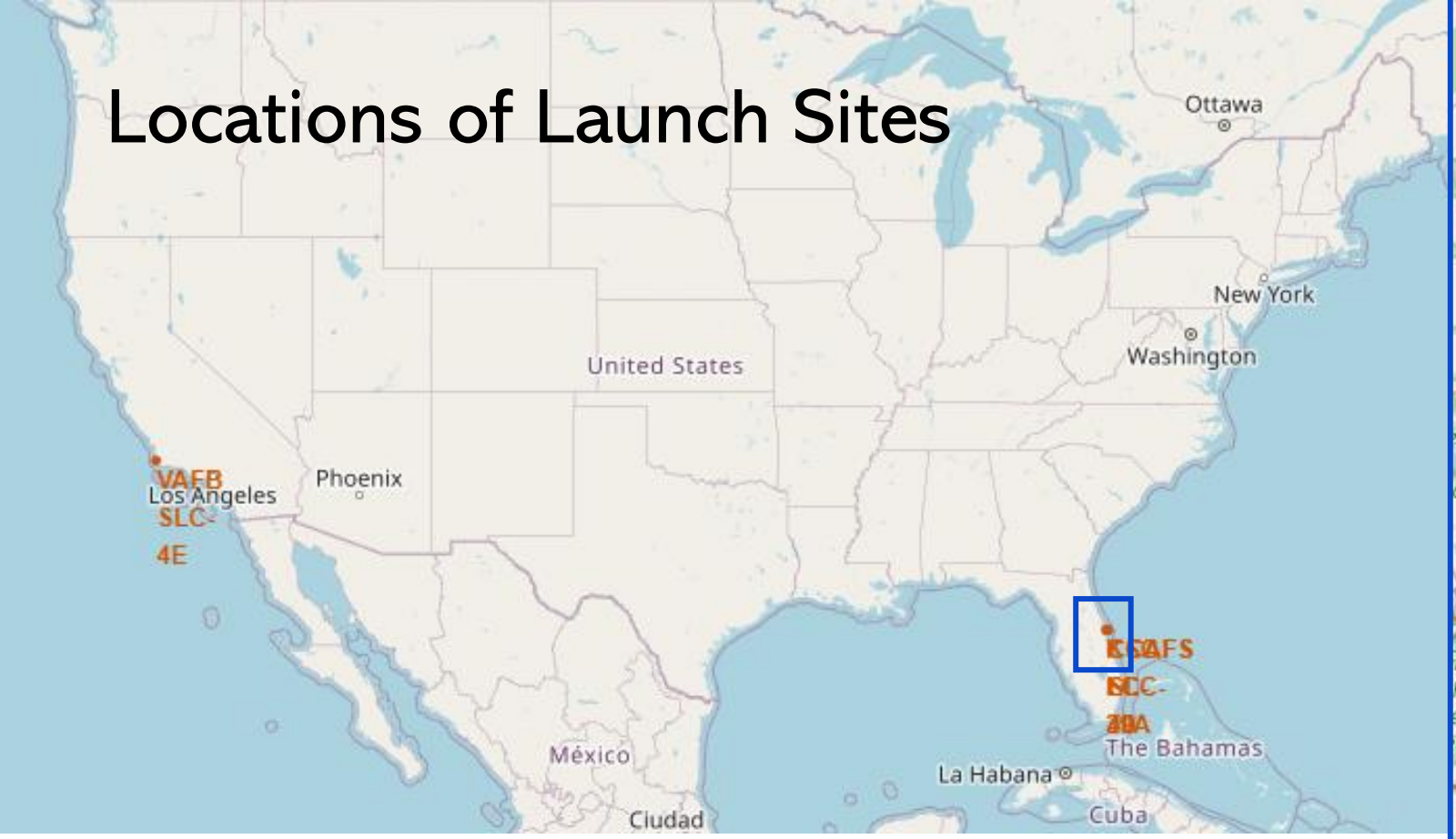
→ 2 failures.

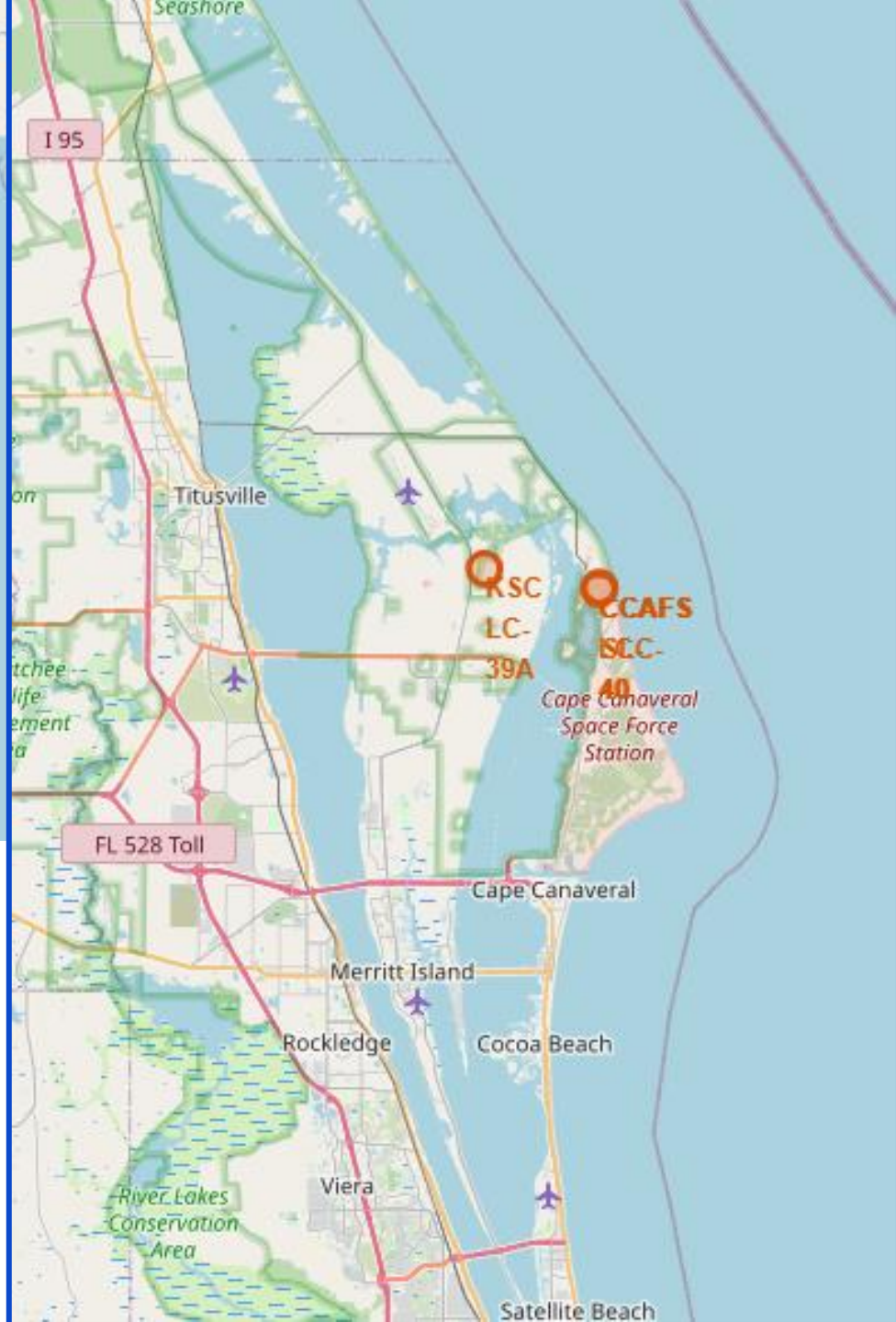
→ At a success rate of 33% this means: 1 success

Section 3

Launch Sites Proximities Analysis

Locations of Launch Sites



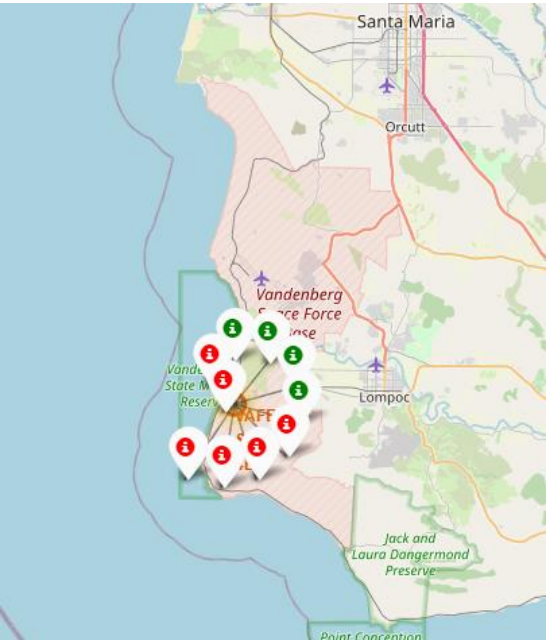


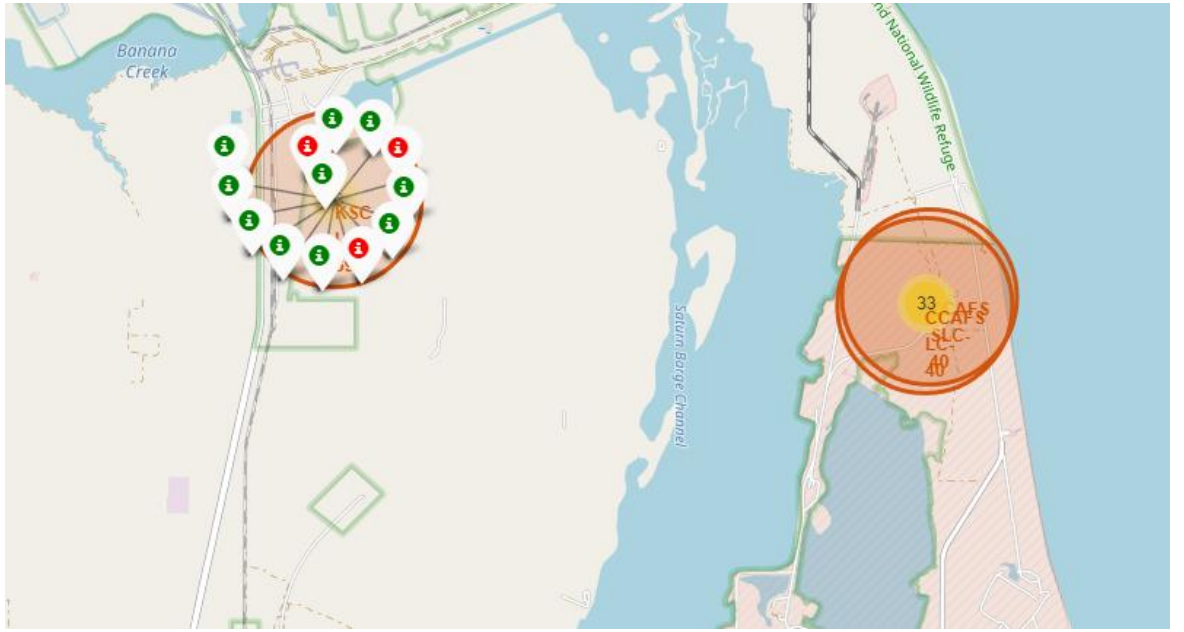
Above map: Location of all 3 sites

Right map: Locations of KSC and CCAFS

→ Launch sites are close to the sea

Interactive Map Elements





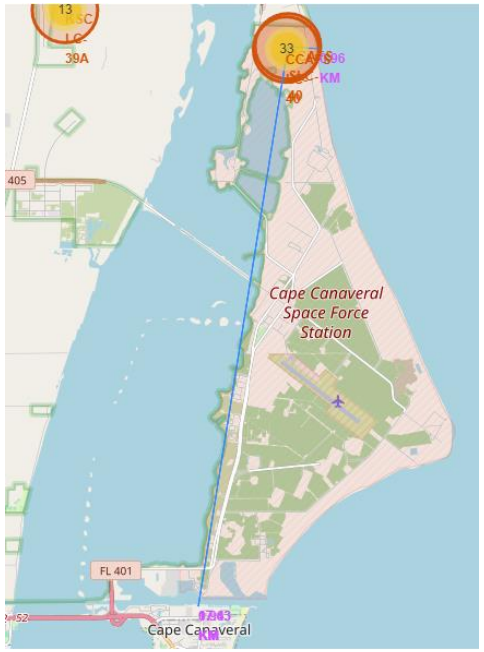
Markers:

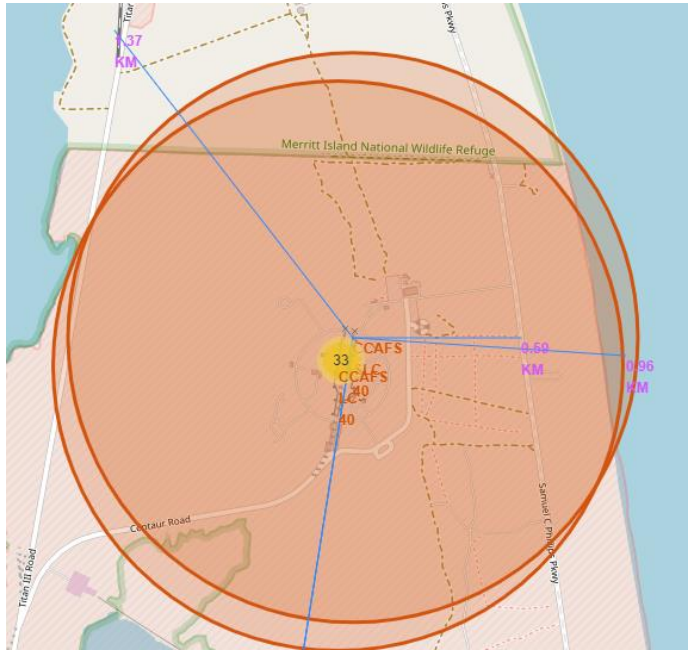
Red: Failed outcomes

Green: Successful outcomes

Yellow Circles: not selected agglomeration of outcomes

Distance to infrastructure





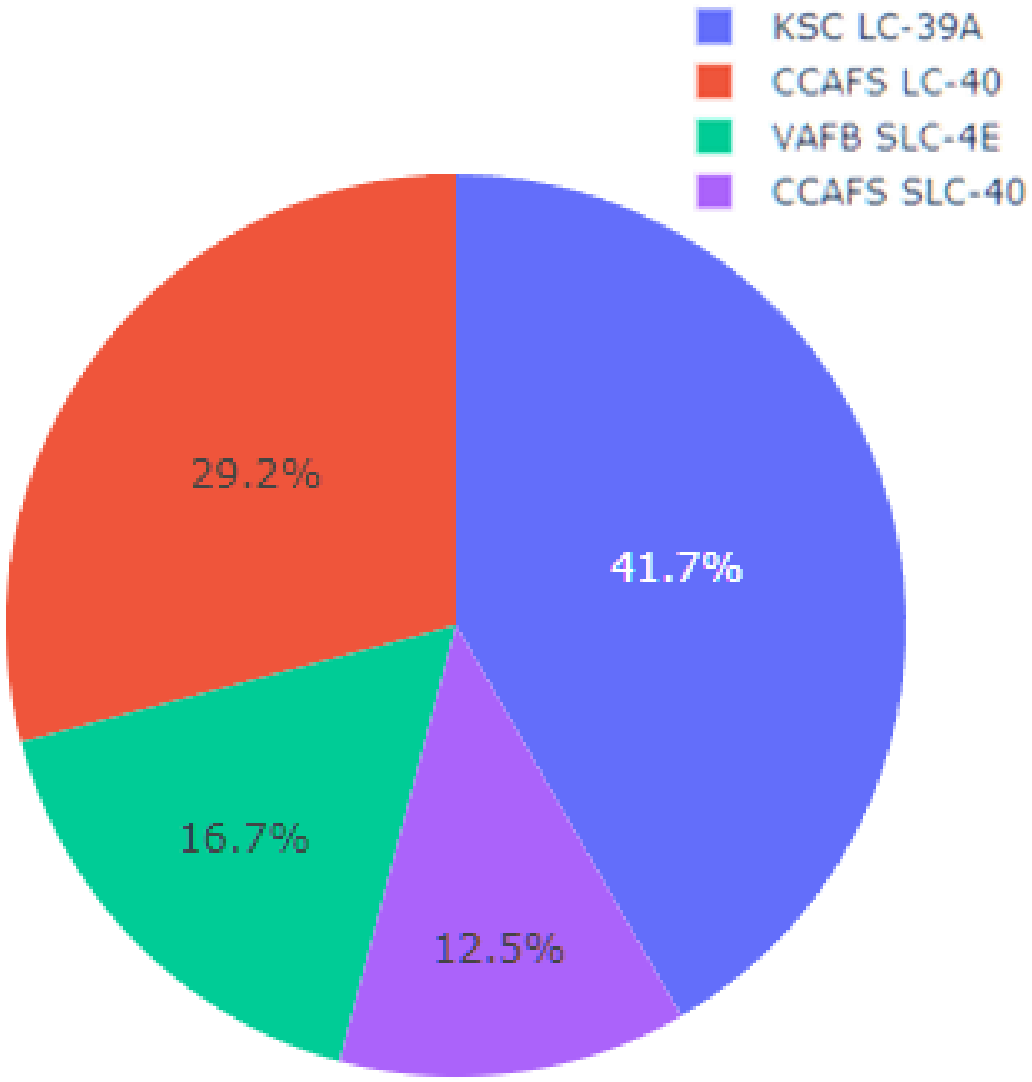
Launch sites are close to the sea, railways and highways lead to the launch sites. (exemplary for CCAFS)

Cities are obviously in some distance to the launch sites

Section 4

Build a Dashboard with Plotly Dash

Launch record on all sites



Note: CCAFS LC-40 and CCAFS SLC-40 are the same location

At CCAFS and KSC each 41.7% of all launches
Both sites are in close proximity

→ For most orbits Florida is preferred over US Westcoast

VAFB SLC-4E – Launch site with highest success rate

The Westcoast site has the highest success rate, however.

All sites: Booster Version and Success

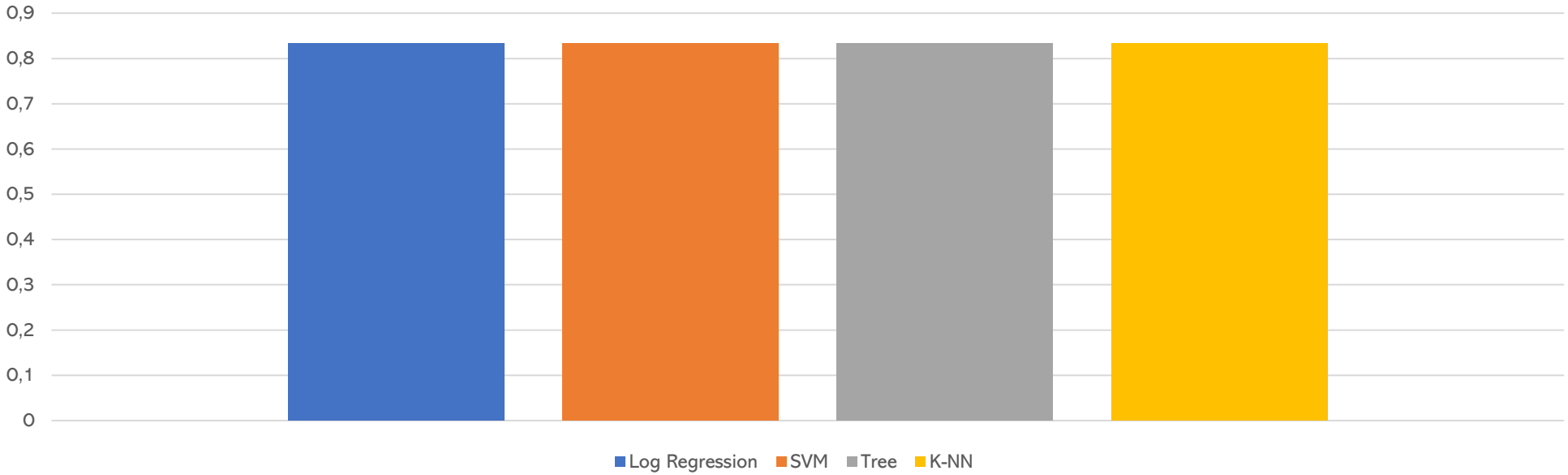
In the range >2.5t and < 7t there have been more failures than successes. Within the group of successful missions there is no Booster v1.1

Most successful missions were carried out by version FT & B4, B5 is probably newer and an improved version of B4. The only data point in this range shows success

Section 5

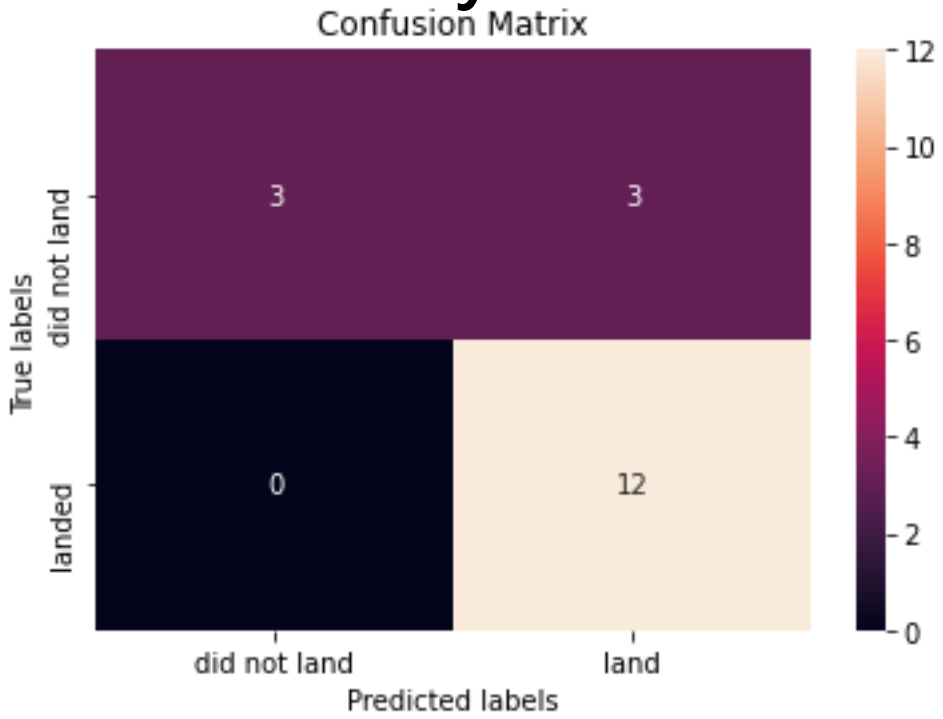
Predictive Analysis (Classification)

Classification Accuracy



Accuracy of each model is practically the same (83.333 %)

Classification Accuracy



The confusion matrix is identical for all classification models

Conclusion

- Given Task: Develop machine learning model for SpaceY for competing against SpaceX
- Model Goal: Prediction of successful stage 1 recovery
- Used Data: SpaceX public API and web scraping of SpaceX's wiki page
- Created data labels and stored data into a SQL database
- Visualized data
- Created a dashboard
- Generated 4 machine learning models with identical accuracy of 83.333%
- Model ca be used by SpaceY for predicting whether a launch will have a successful Stage 1 landing / recovery prior launching

Image Source: Ivan Diaz (unsplash)

Appendix

Github path:

https://github.com/spyderroque/Data_Science_Capstone_Project

Image Source: Planet Volums (unsplash)

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!