IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Giovanni BRICCONI
Dec 27 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Public data and web scraping / data cleaning pandas and SQL / data visualization and online dashboards with Folium and Dash / data visualization with matplotlib / how to predict successful returns using different algorithms
- Publicly available data allows to understand key points on SpaceX business
  - Where launch sites should be located
  - How launch reliability has improved during these years
  - Kind of payload launched and rocket recovery success rate

Introduction

SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars; other providers cost upward of 165 million dollars each

Much of the savings is because SpaceX can reuse the first stage.

Cans SpaceY understand when first stage reusage is possible and gain precious insights on how to orient its business?

Which are the most effective launch sites?

Which kind of payloads and orbit offer the best opportunities?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Calls to web services endpoints. Scraping of Wikipedia pages with BeautifulSoup
- Perform data wrangling
  - NaN values handling, outcomes normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train test data splitting, hyperparameters tuning, data normalization, benchmark of LogisticRegression, SupportVectorMachines, DecisonTrees and Kneighbors Classifier.

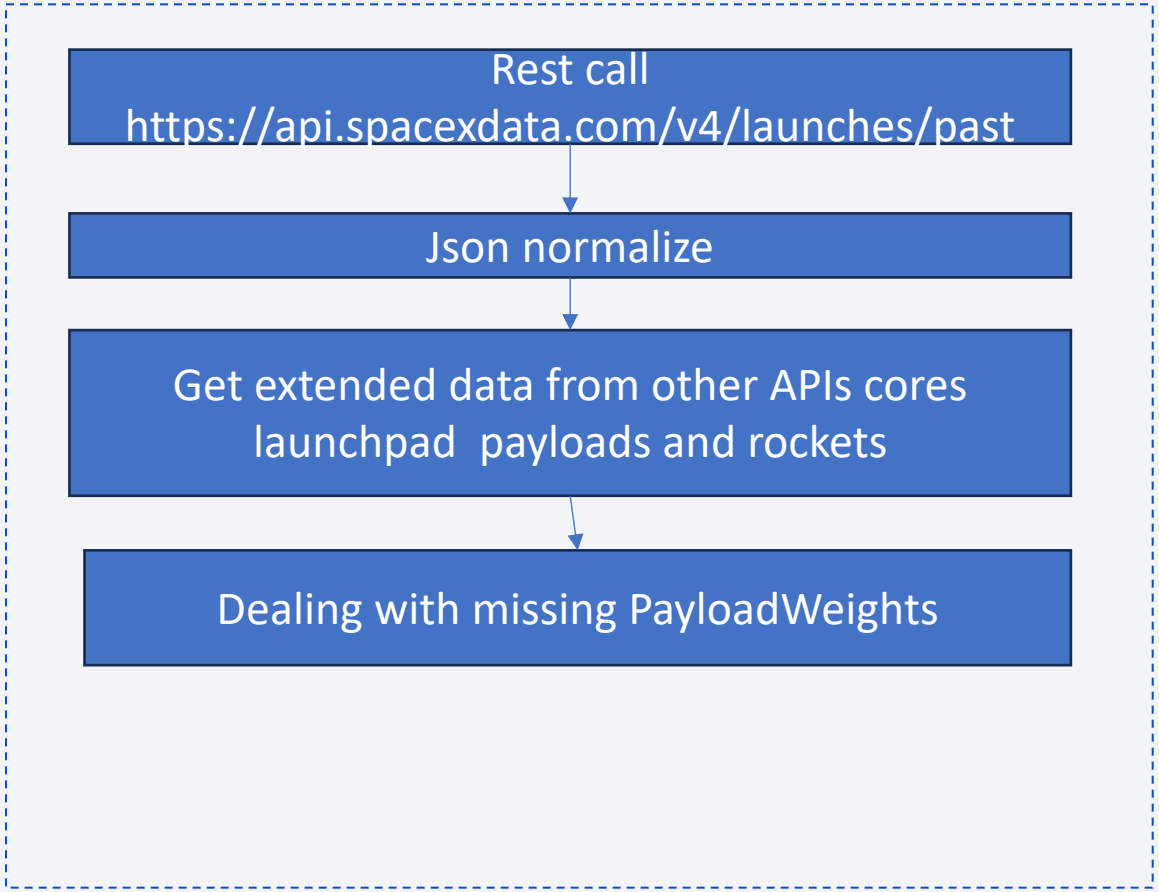
Data Collection

- SpaceX public API used to get launch information (`pandas` and `requests` python libraries)
- Filtering on Falcon 9 launches
- Mean values used to substitute NaN values
- Wikipedia scraping with BeautifulSoup lib

Data Collection – SpaceX API

- First collect data from launches.
- Collect extendend information from other APIs
- Detect missing values and use average payload weights as replacement

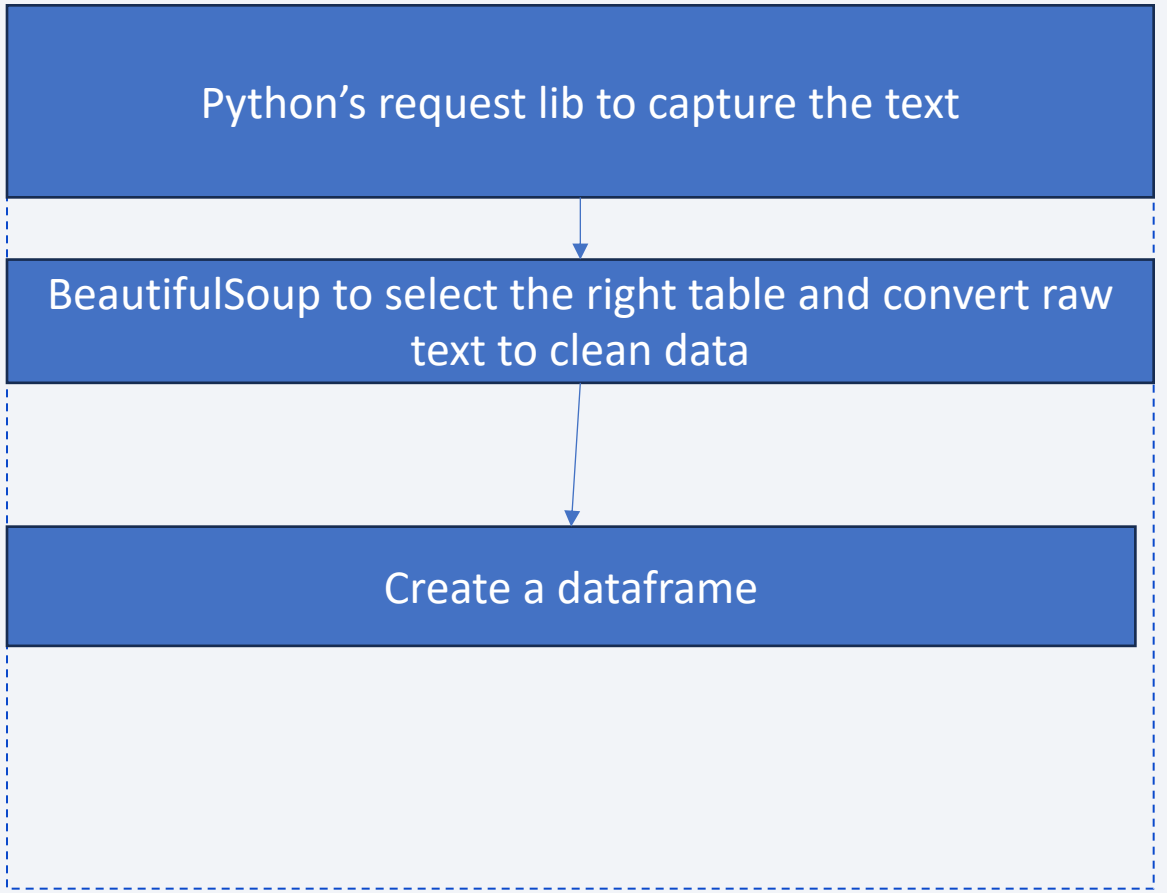
- GitHub code



Data Collection - Scraping

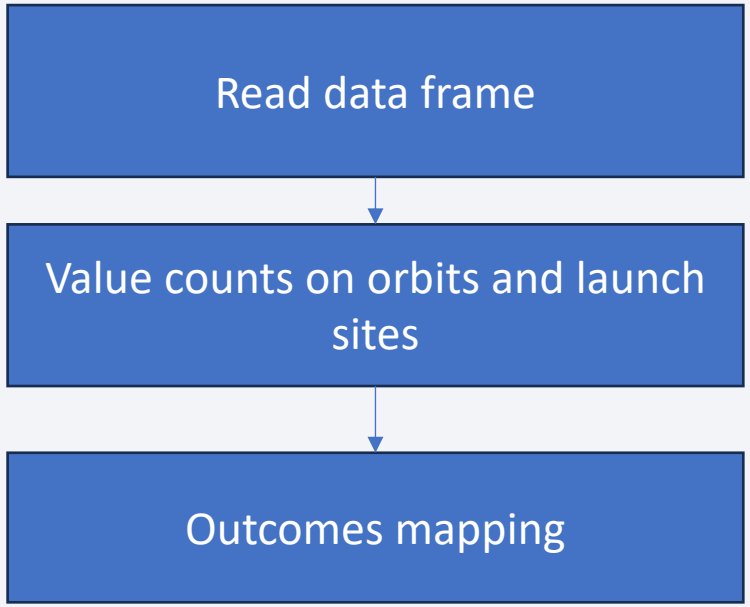
- Wikipedia contains complementary information, but in HTML format.
- Using BeautifulSoup to convert text data into a pandas dataframe

- GitHub code



Data Wrangling

- Using Pandas to
  - Identify launch sites
  - analyze orbits kinds and frequencies
  - Analyze outcomes kinds -> we want to identify first stage recovery
- GitHub link



EDA with Data Visualization

- Payload mass evolition wrt flight number
- Lunches by launch sites
- Payload vs Launch site
- Success rate by orbit type
- Orbit / payload relation
- Success rate evolution during these years
- GitHub code

EDA with SQL

- View launces from CCA
- Total payload lanched from NASA CRS
- F9 v1.1 average payload
- Discover first ground pad successful outcome
- Date processing with sqlite
- Could just take pictures of the notebook, save option was not working. See the Screenshot pictures in https://github.com/gibri/data-science-training/tree/master/spacex

Build an Interactive Map with Folium

- Locate launch sites and launches on the maps using circles and explanatory markers. Draw distance lines from a specific point. Launch sites are near rail stations and far from cities.
- GitHub URL

Build a Dashboard with Plotly Dash

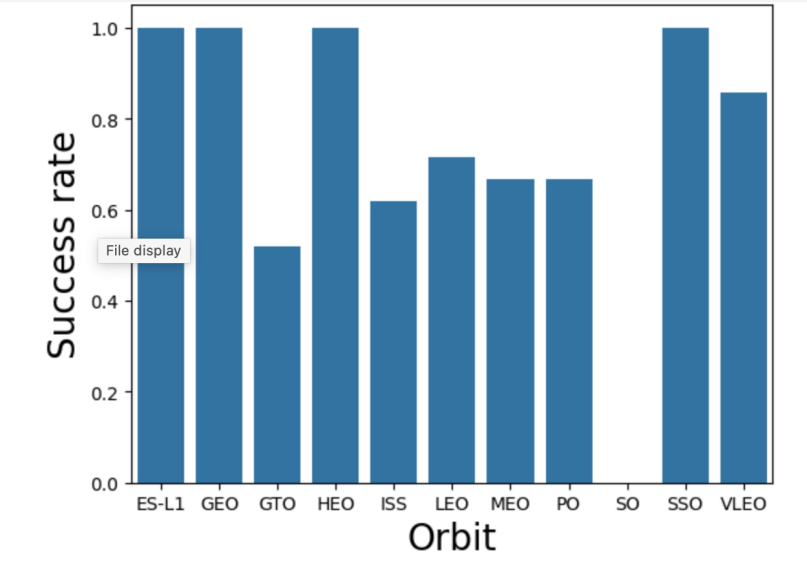
- Analyze success rate by launch site with pie chart
- Filter on payload weight
- Show scatter plot representing launches outcome
- Which site has the largest successful launches? KSK lc-39A
- Which site has the highest launch success rate? KSK lc-39A
- Which payload range(s) has the highest launch success rate? 2/4k
- Which payload range(s) has the lowest launch success rate? 6/10k
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? FT

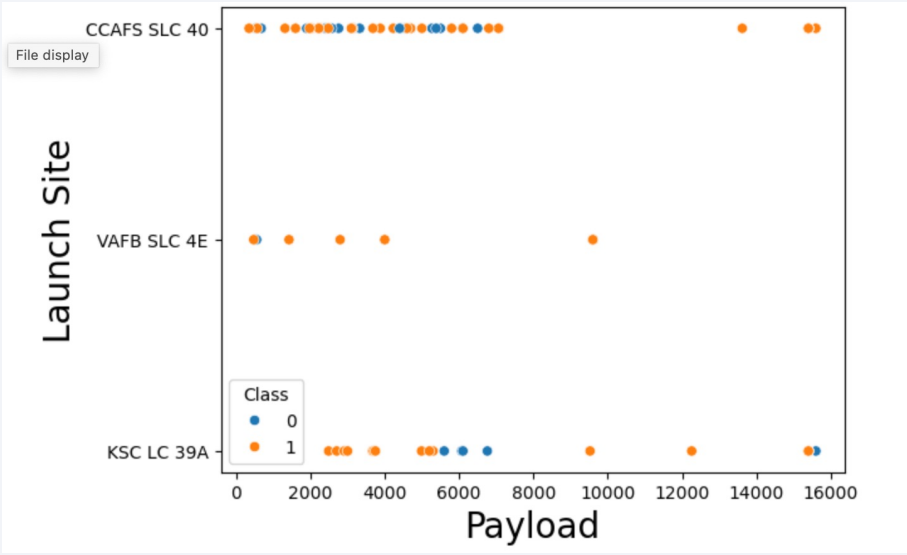
- Explain why you added those plots and interactions
- GitHub URL

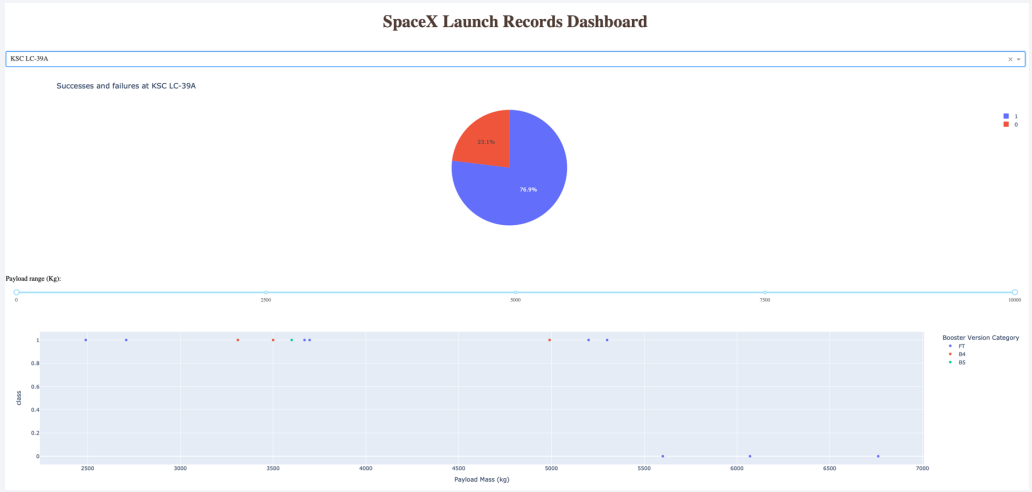
Predictive Analysis (Classification)

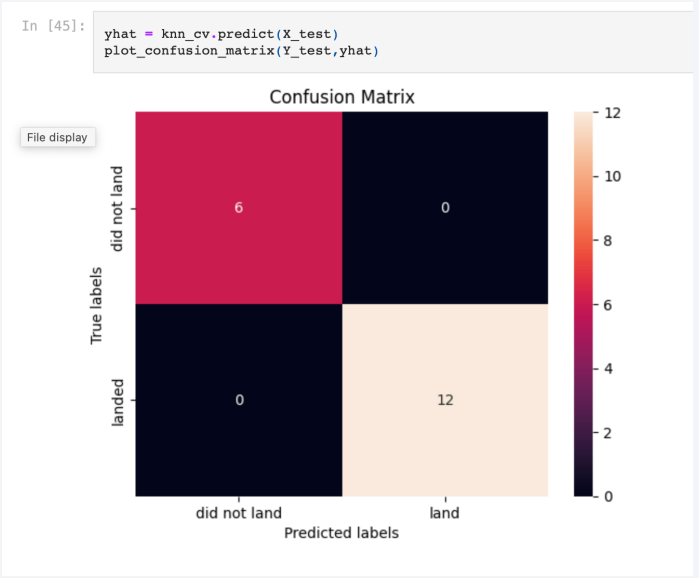
- Input data has been normalized and divided into test and train sets (20%-80%)
- Four models have been compared: SVM, Decision Trees, logistic regression and Knearest neighbors
- Hyperparameter tuning cross validation 10 folds
- Performances compared using confusion matrixes, and correct scores on test data
- KNN is the best model
- https://github.com/gibri/data-science-training/blob/master/spacex/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

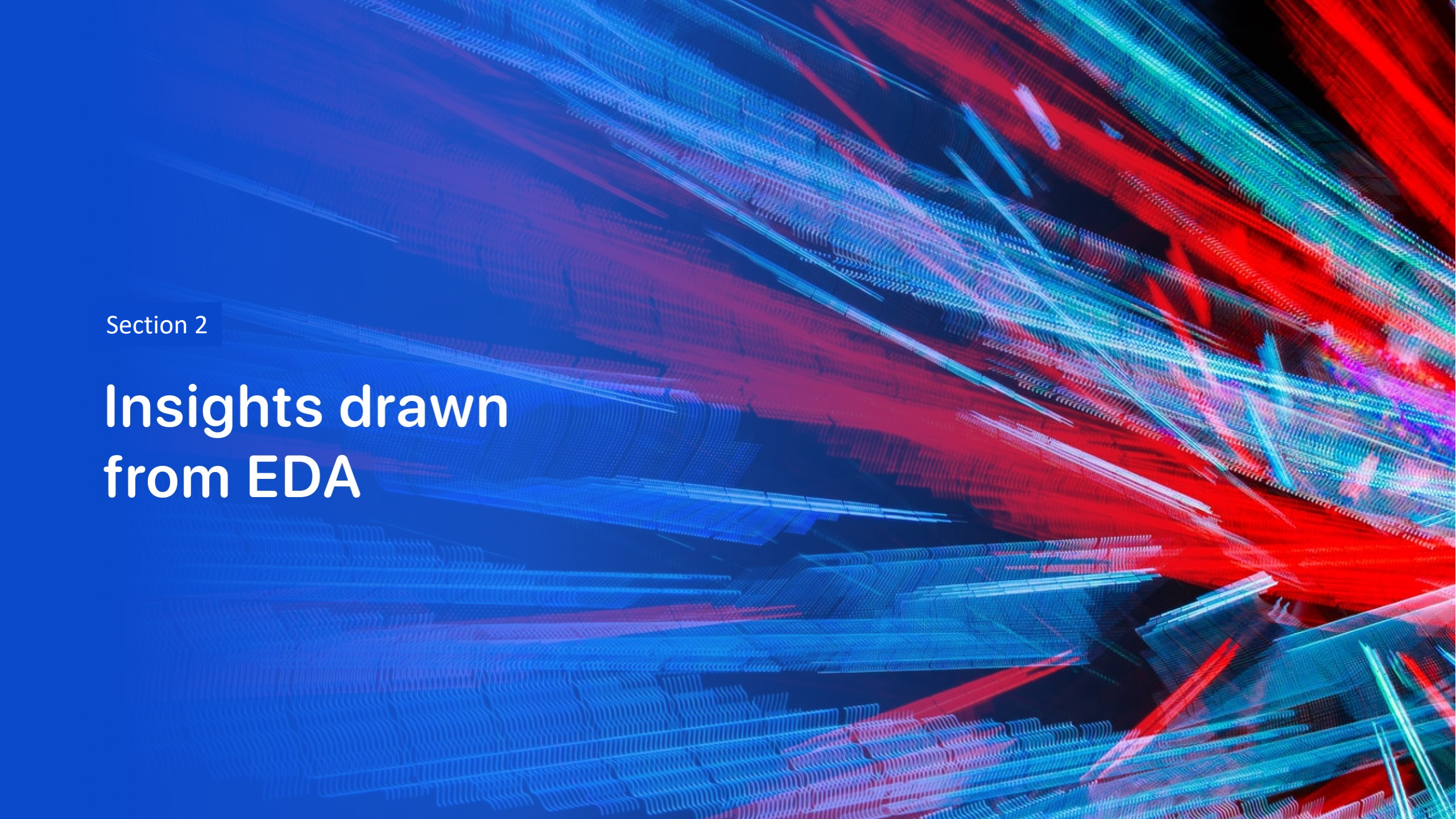
Results knn is the best model



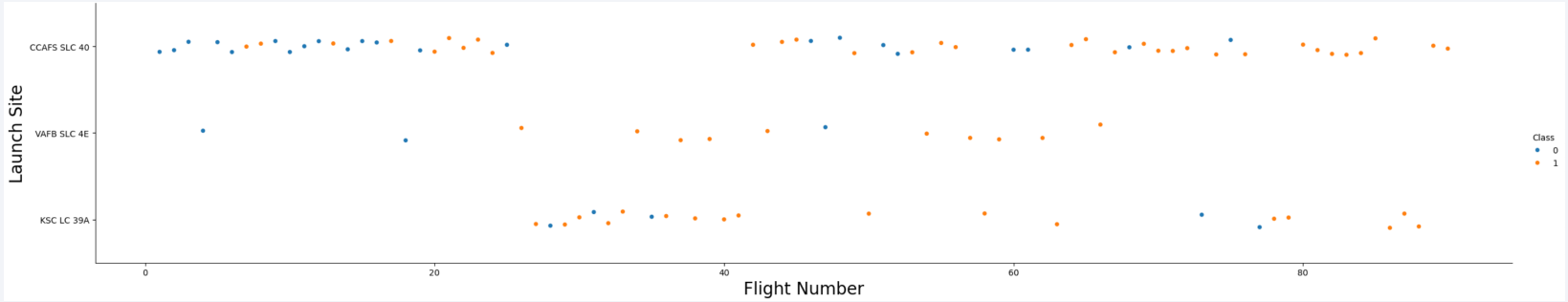







Section 2
Insights drawn from EDA

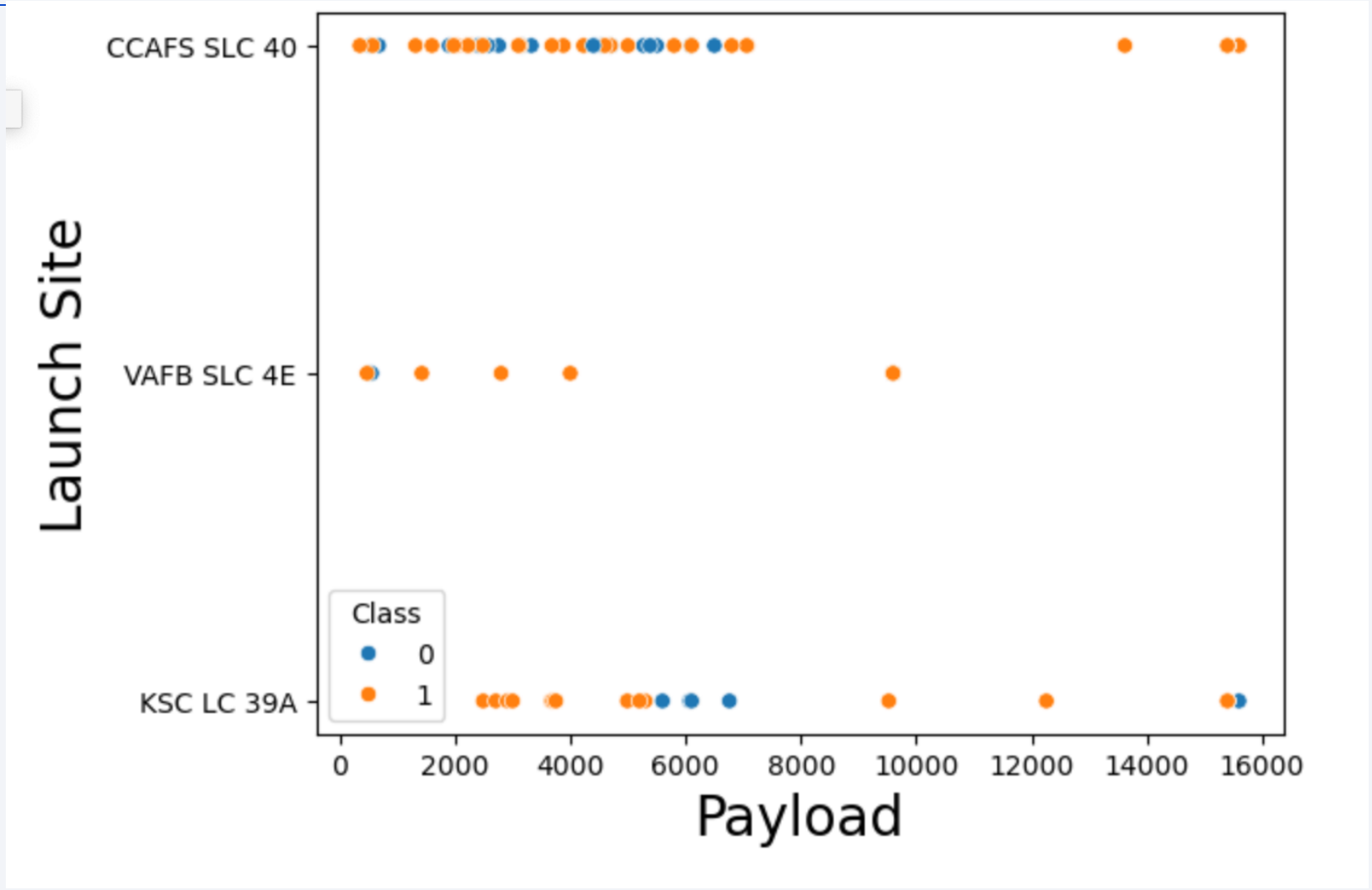
Flight Number vs. Launch Site



- Flight Number vs. Launch Site. SpaceX operates mainly from 2 sites, but a third one was used too. Orange dots represent successful return of the 1st stage. New launches are usually successful.

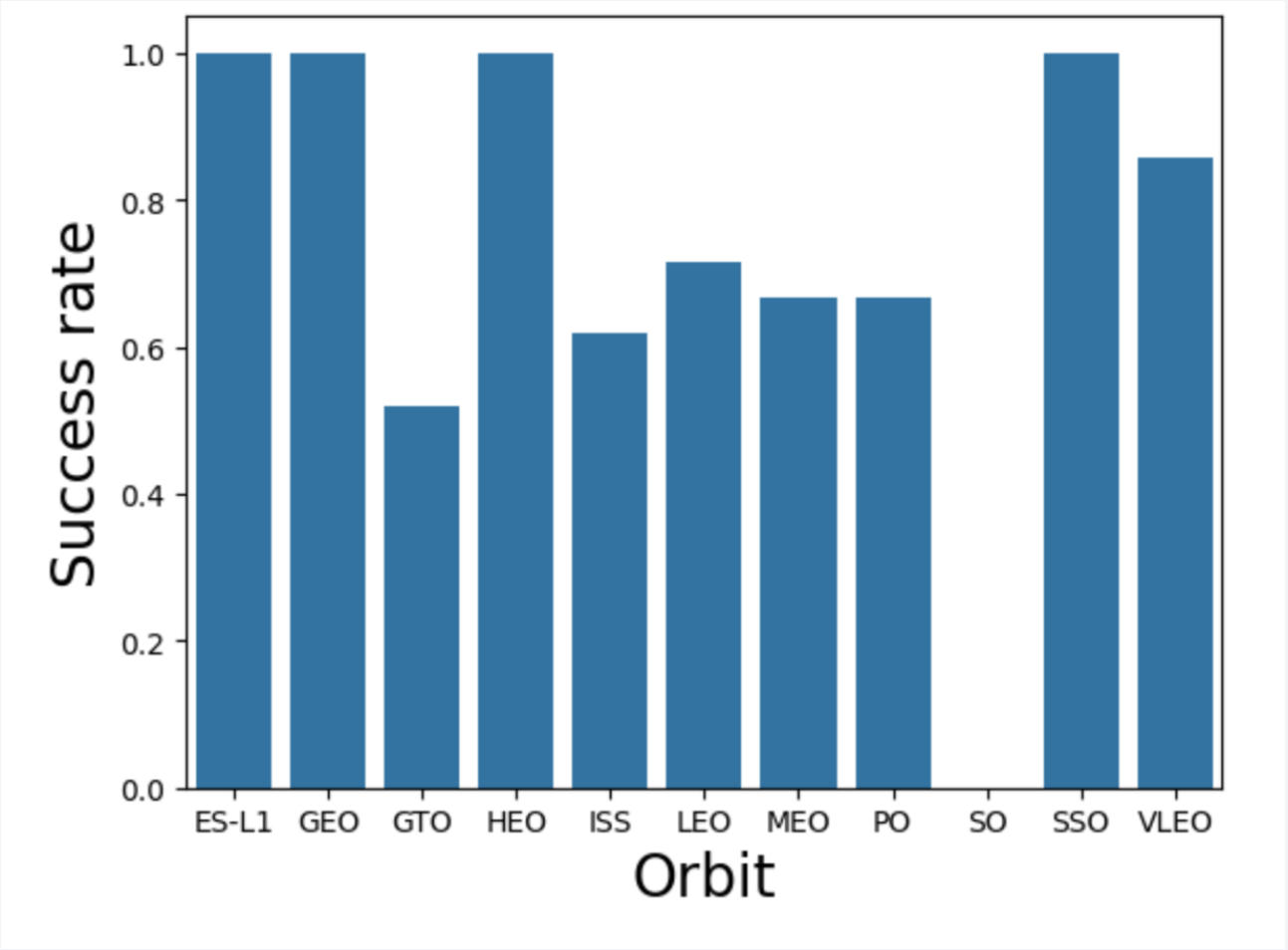
Payload vs. Launch Site

- VAFB site launch only light payload missiles.
- The other 2 sites are more versatile



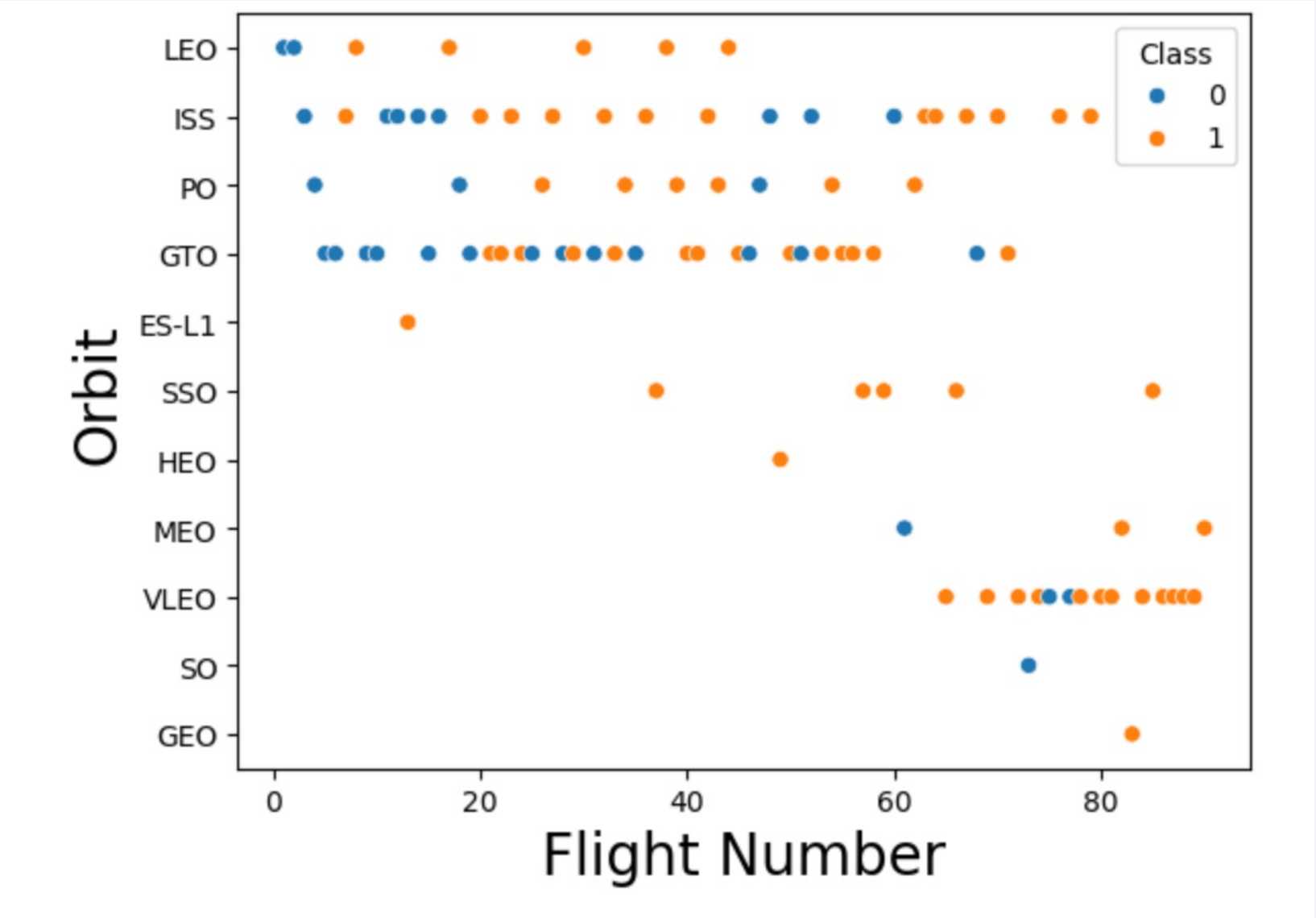
Success Rate vs. Orbit Type

- It's possible to launch satellites with 11 different orbit types.
- The success rate depends on the orbit type



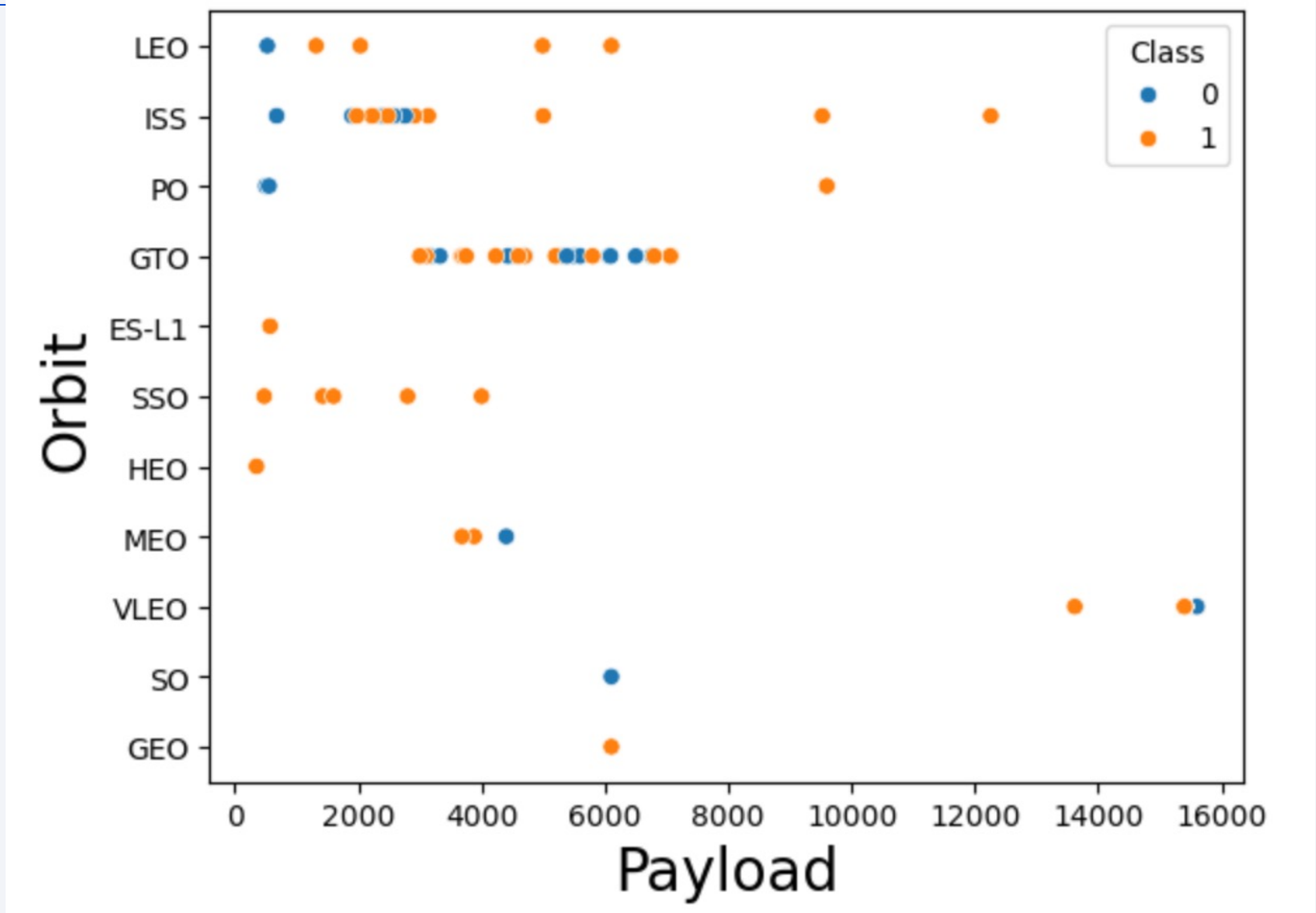
Flight Number vs. Orbit Type

- Recently VLEO orbit stated to be very used for new launches.
- In the LEO orbit the Success appears related to the number of flights



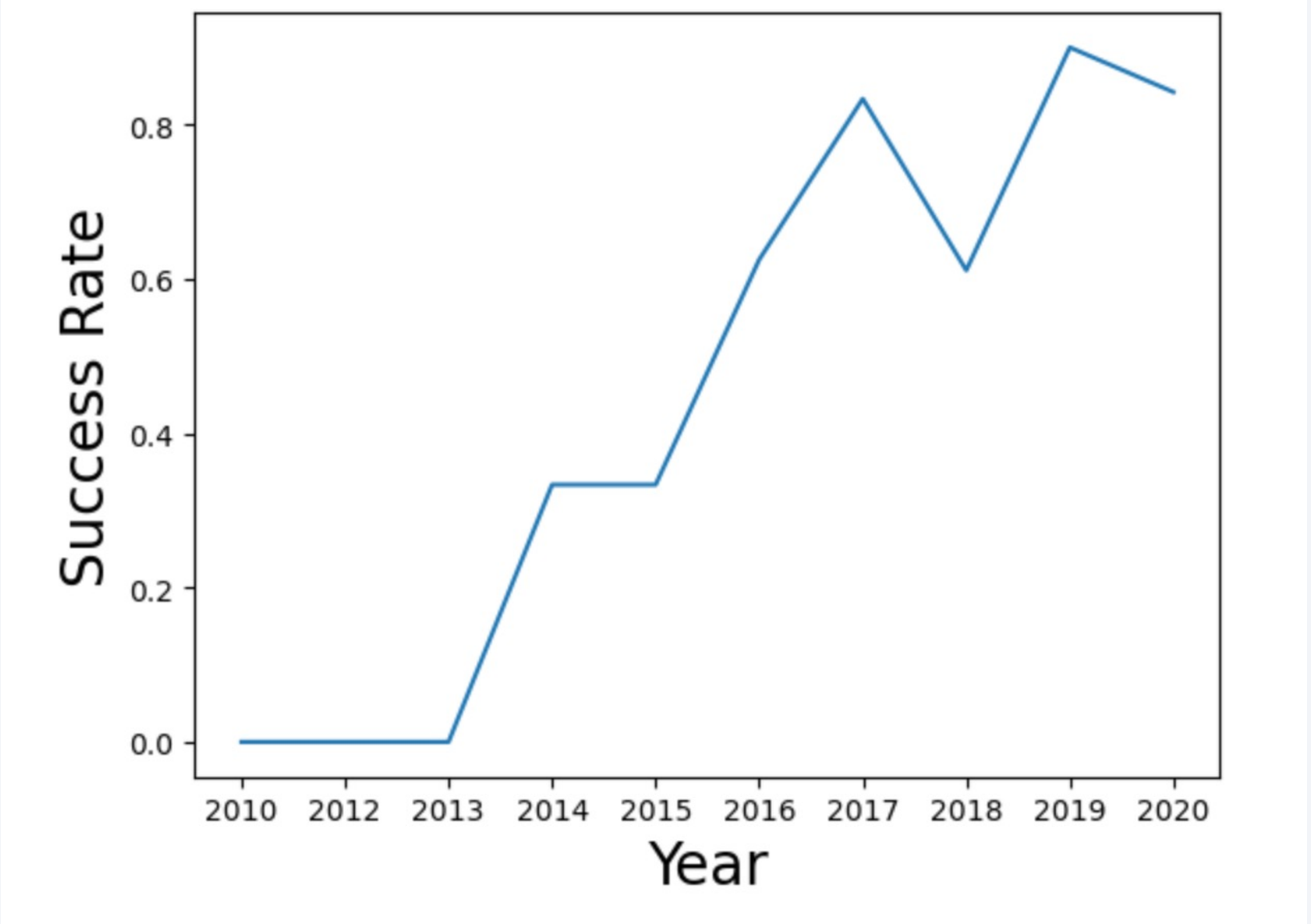
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- Since 2013 the success rate started to increase



All Launch Site Names

- We see that actually rockets are launched from 4 different sites, although 2 are in the same space center.

```
[20]: %sql select distinct "Launch_Site" from spacextable
 * sqlite:///my_data1.db
Done.
```

[20]:

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

Launch Site Names Begin with 'CCA'

- Some examples of launces from the `CCA` space center

```
[21]: %sql select * from spacextable where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

[21]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA is about 46,000 Kg

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

```
[27]: %sql select Customer, sum(payload_mass__kg_) from spacextable where Customer like 'NASA (CRS)' group by Customer
```

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

[27]:

Customer	sum(payload_mass__kg_)
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,500 Kg

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

```
%sql select 'F9 v1.1%' booster_version, avg(payload_mass__kg_) from spacextable where booster_version like 'F9 v1.1%'
```

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

booster_version	avg(payload_mass__kg_)
F9 v1.1%	2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened in December 2015

```
[35]: %sql select min(Date) from spacextable where landing_outcome='Success (ground pad)' order by date asc
 * sqlite:///my_data1.db
Done.
[35]: min(Date)
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Rockets can land on a drone ship. Successfully landed on drone ship with payload mass greater than 4000 but less than 6000

```
[37]: %sql select * from spacextable where landing_outcome='Success (drone ship)' and payload_mass__kg_>=4000 and payload_mass__kg_<6000
 * sqlite:///my_data1.db
Done.
```

[37]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
[38]: %sql select mission_outcome, count(*) from spacextable group by mission_outcome
```

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

[38]:

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

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass

```
[48]: %%sql select booster_version from spacextable where payload_mass__kg_ = (select max(payload_mass__kg_)
      from spacextable )
```

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

[48]:

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

2015 Launch Records

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

[50]:
```
%%sql select substr(Date, 6,2) as month, landing_outcome, booster_version, launch_site
from spacextable
where substr(Date,0,5)='2015'
and landing_outcome='Failure (drone ship)'
```

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

[50]:

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

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

- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

[53]:
```
%%sql select landing_outcome, count(*) as total
from spacextable
where date between '2010-06-24' and '2017-03-20'
group by landing_outcome
order by count(*) desc
```

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

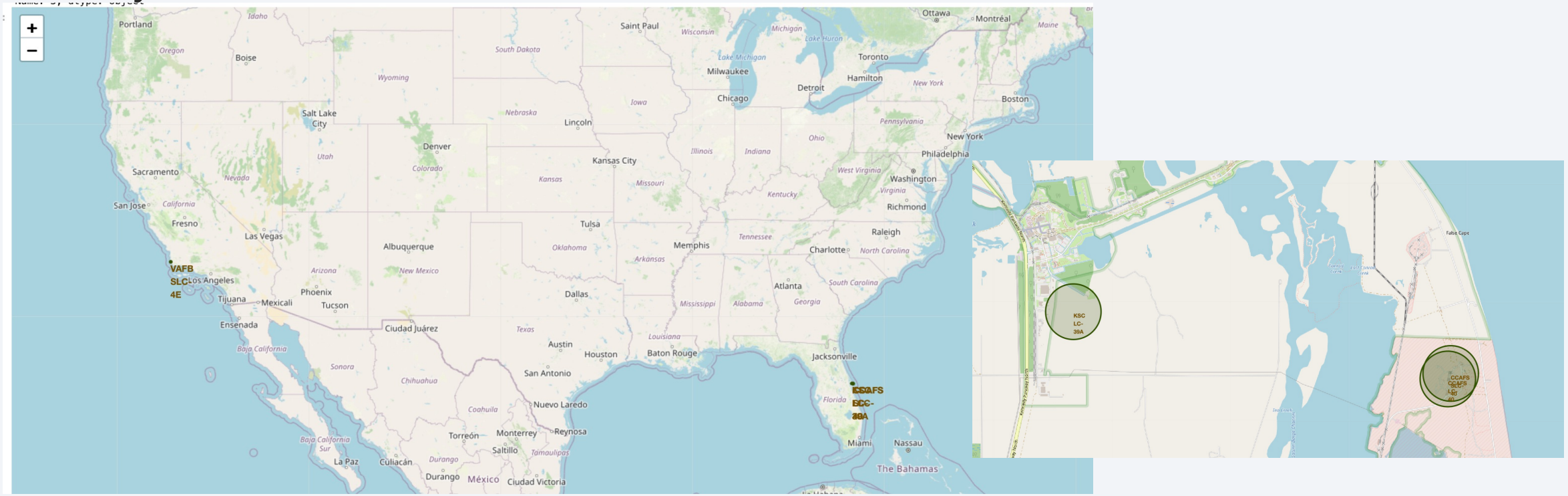
[53]:

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


Section 3
Launch Sites
Proximities Analysis

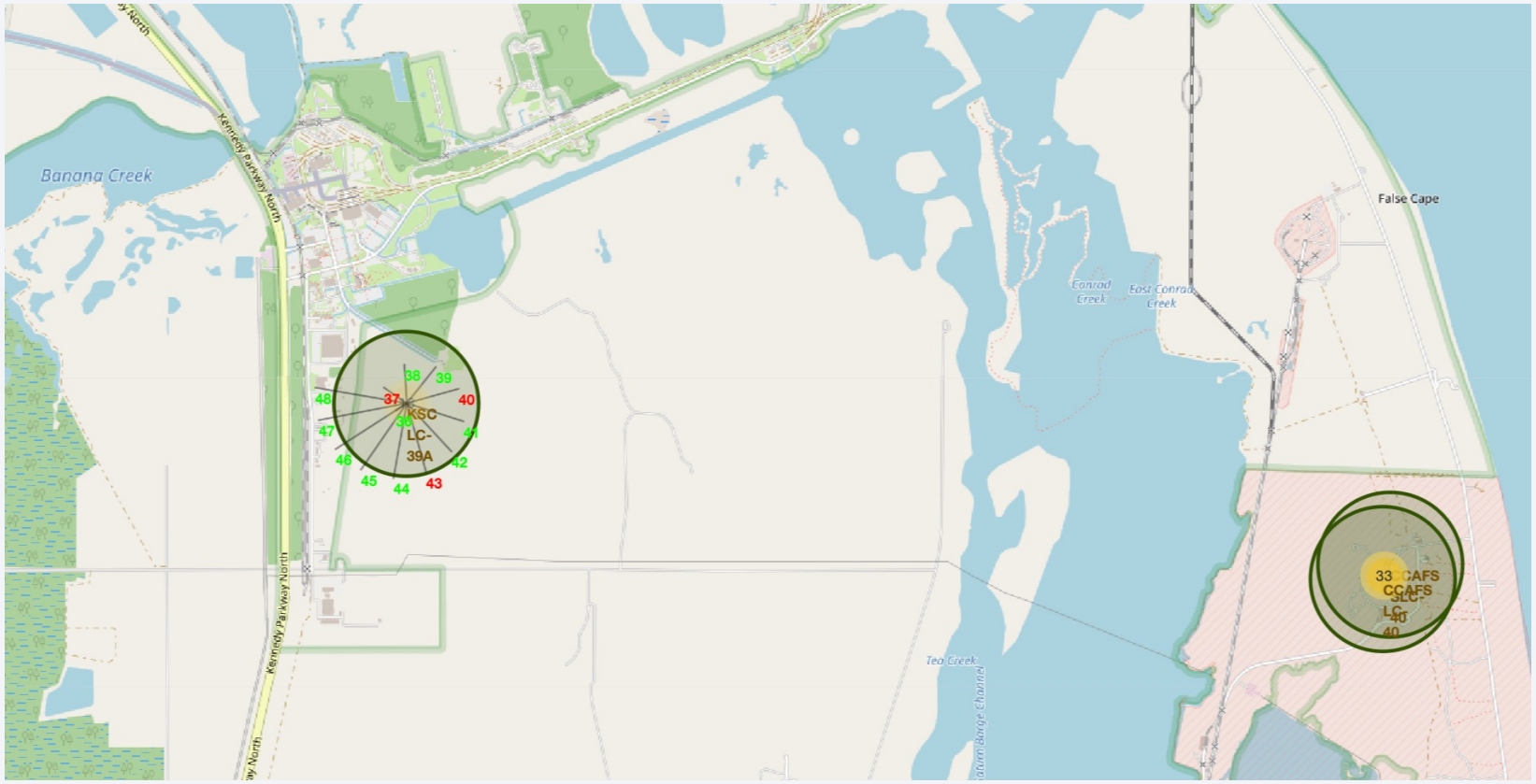
Launch site locations

- SpaceX launches from the East and the West coast. Sites on the east coast are very close



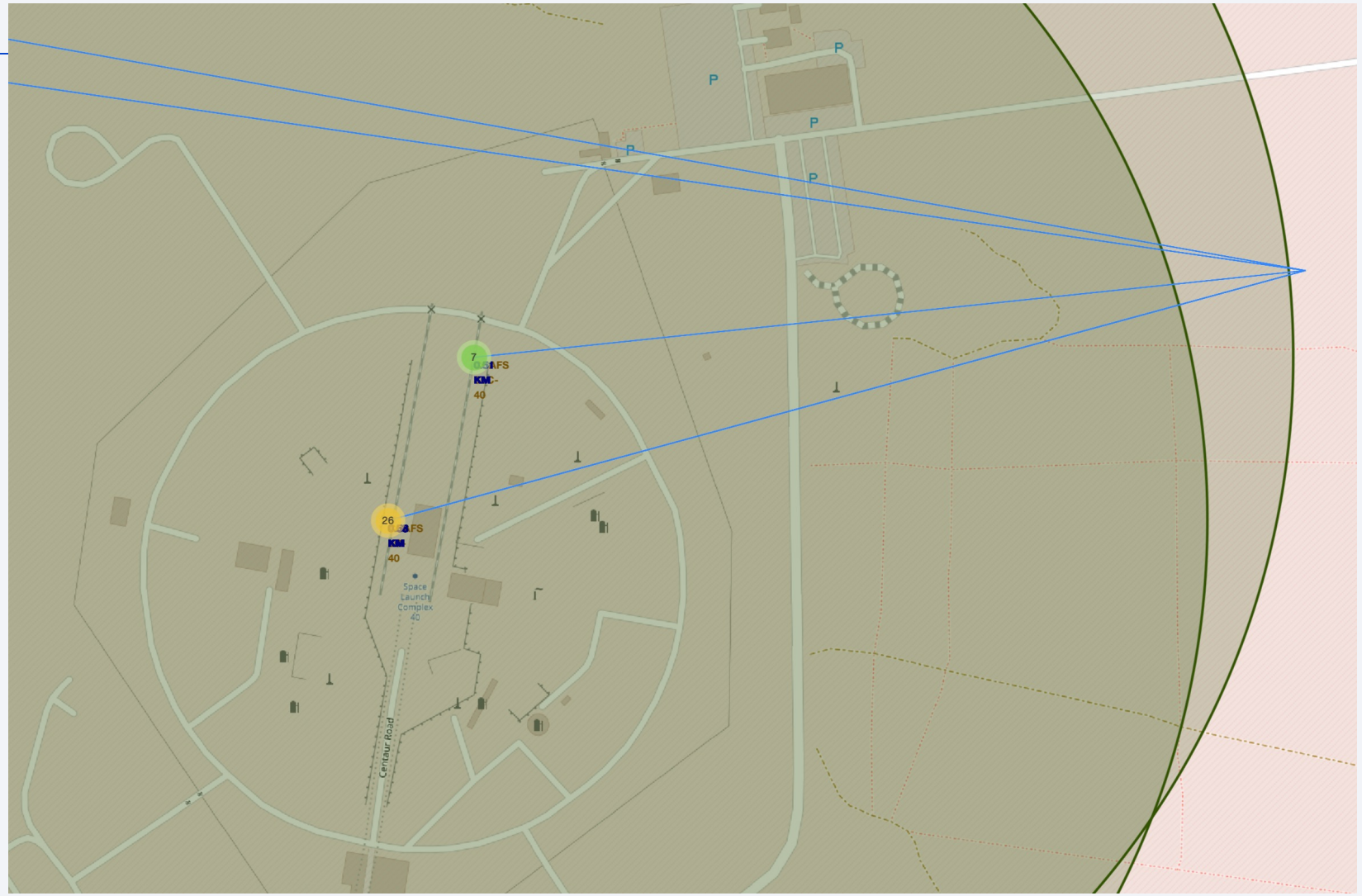
Successful launches by site

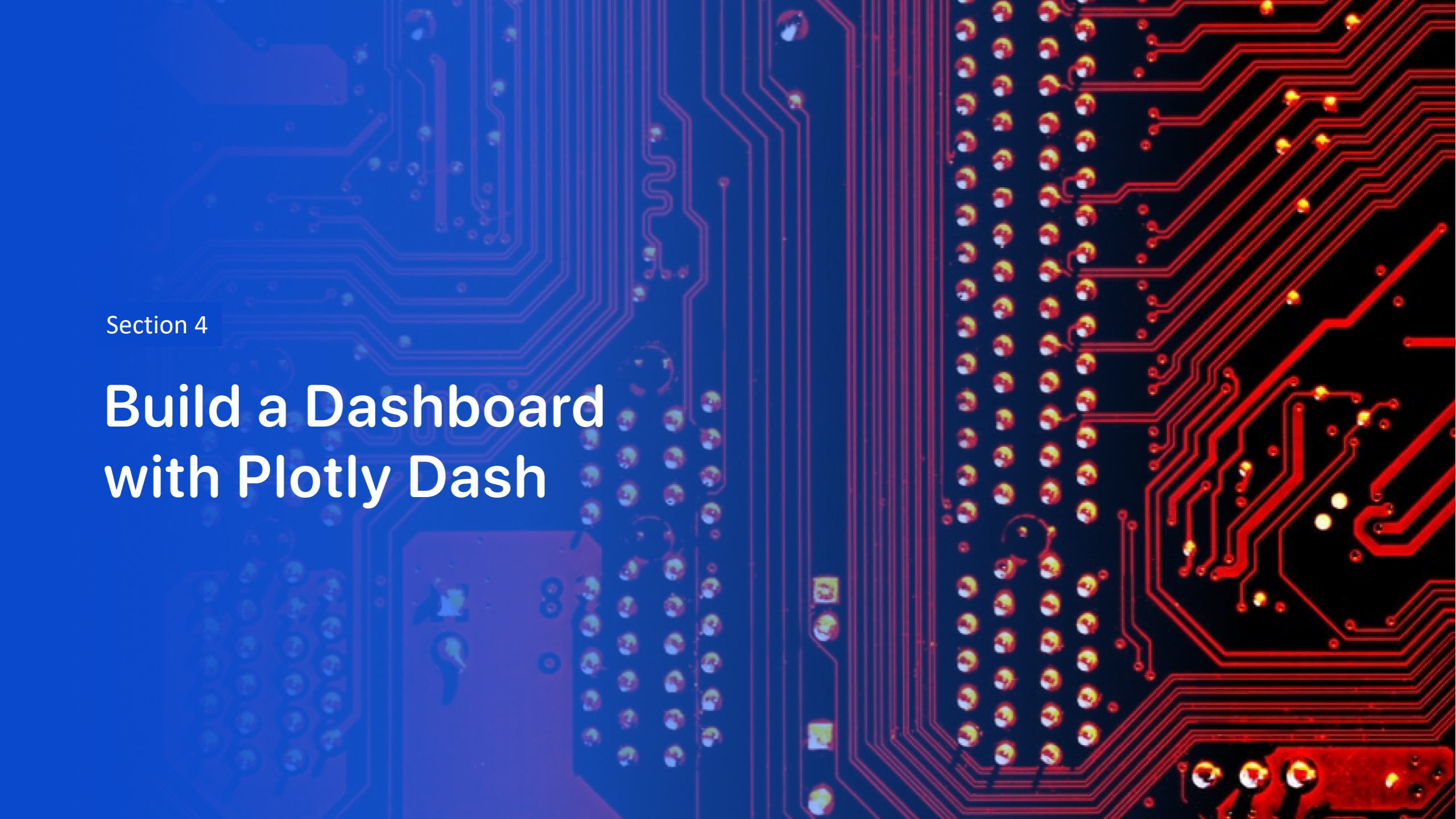
- We can explore successful launches by site



Launch sites are close to railroads

- Highway and coastline are close, and cities are far away.

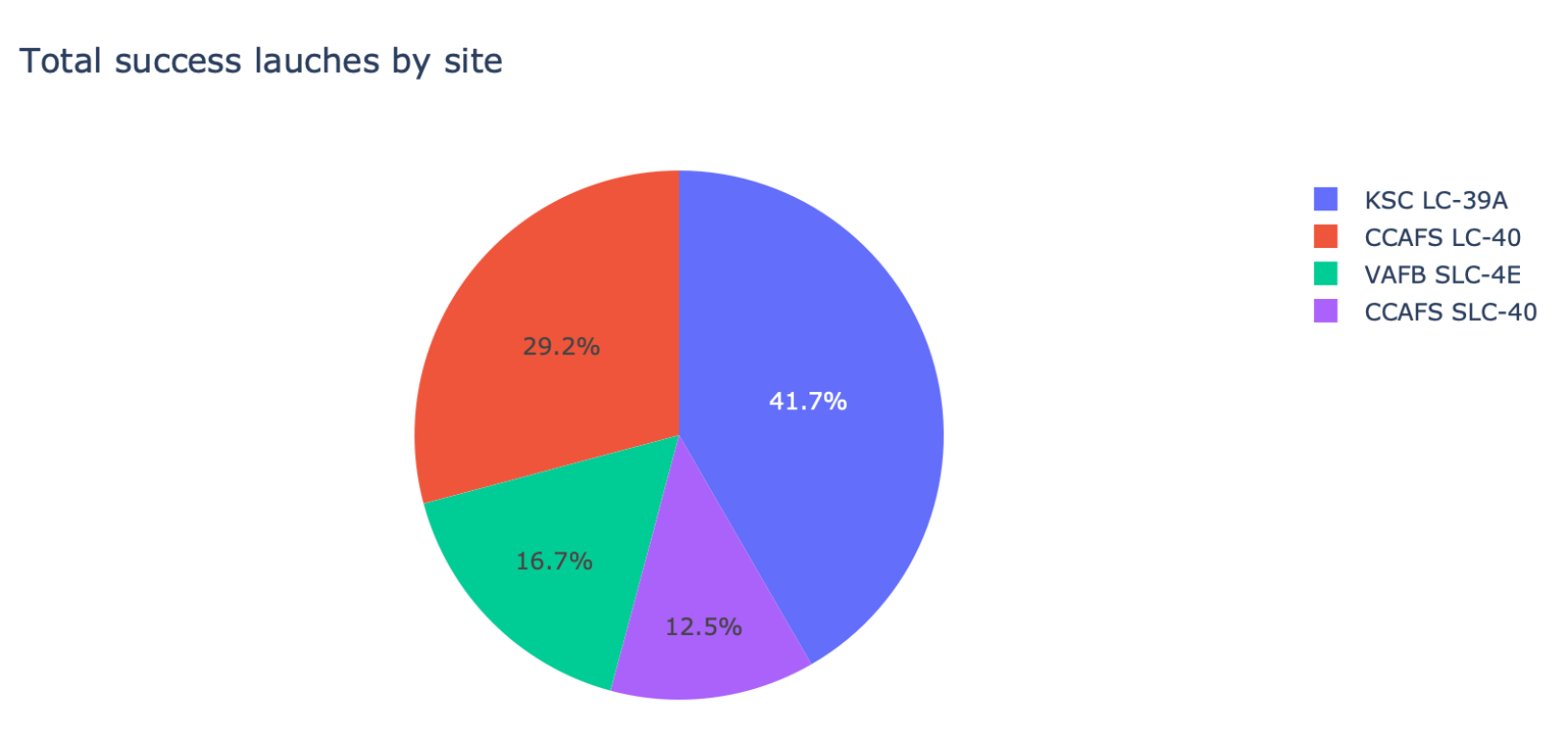




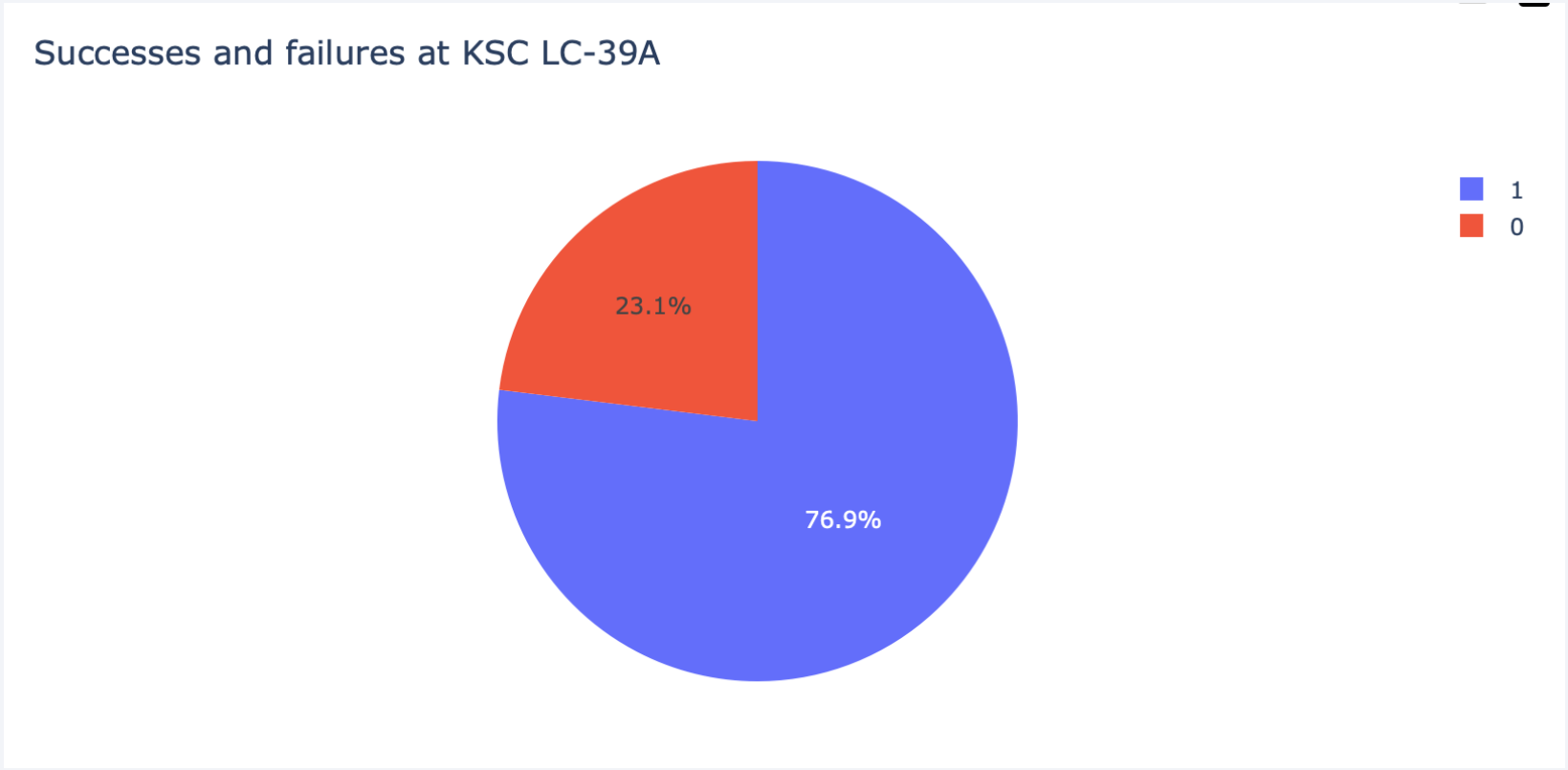
Section 4

Build a Dashboard with Plotly Dash

KSC is the most successful launch site

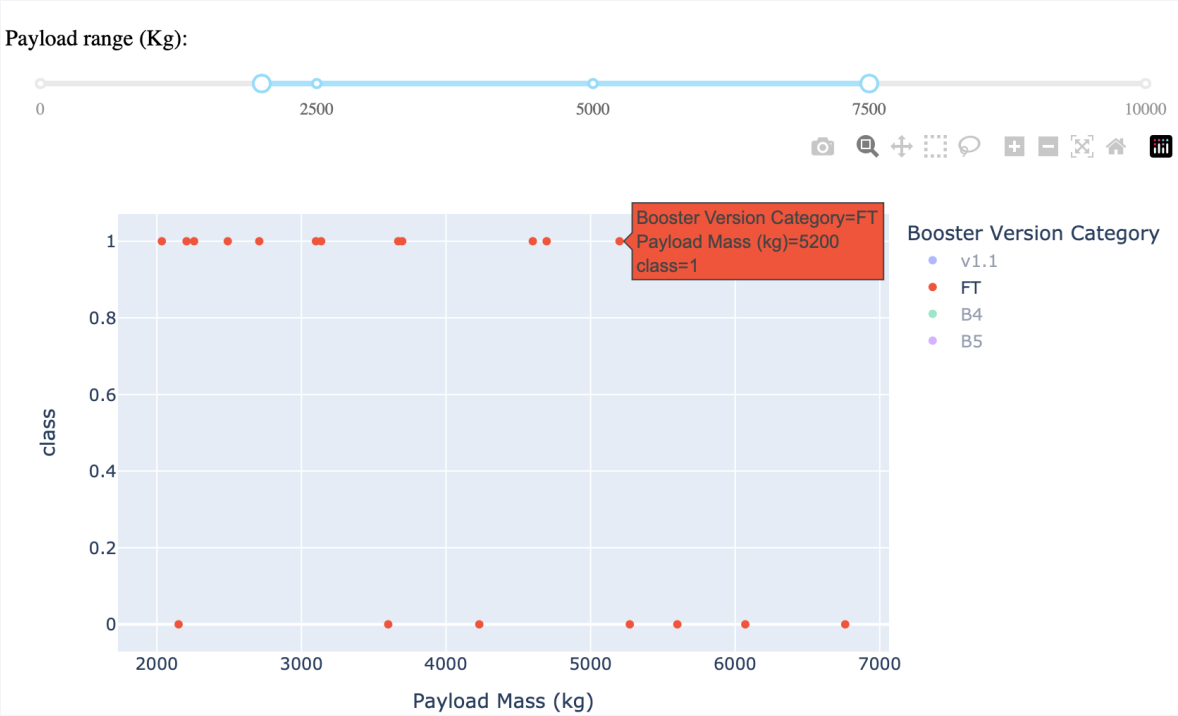


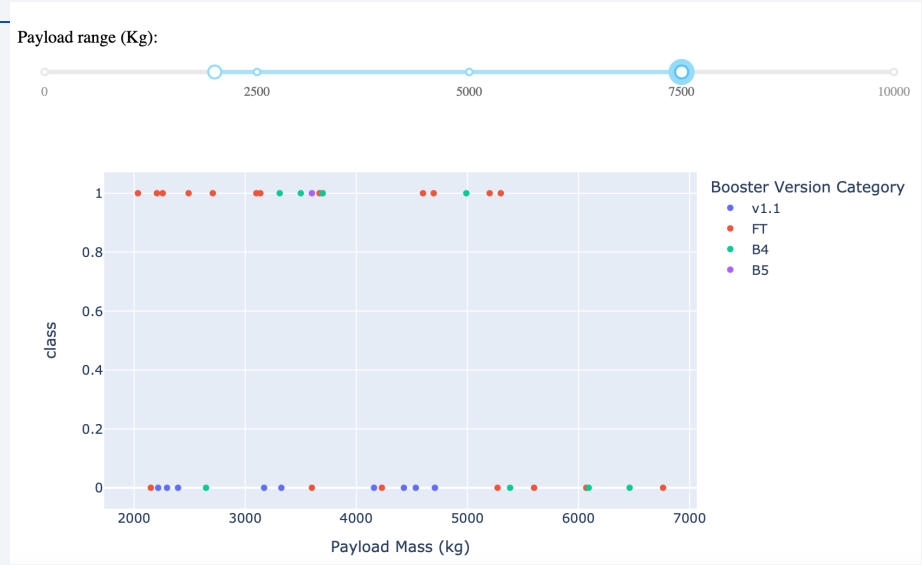
76% of launches are successful on KSC



Success and failure on a payload range

- FT boosters are the most effectives on this common payload range

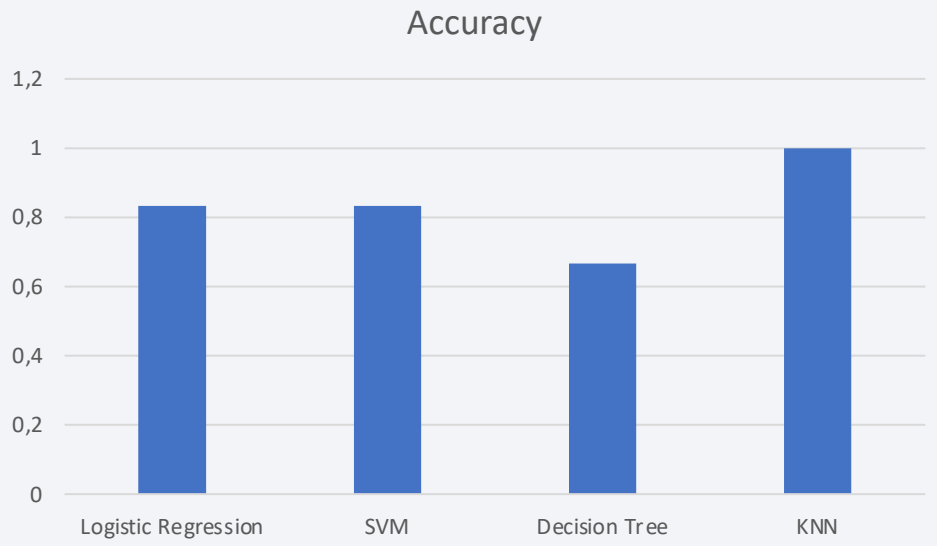




Section 5

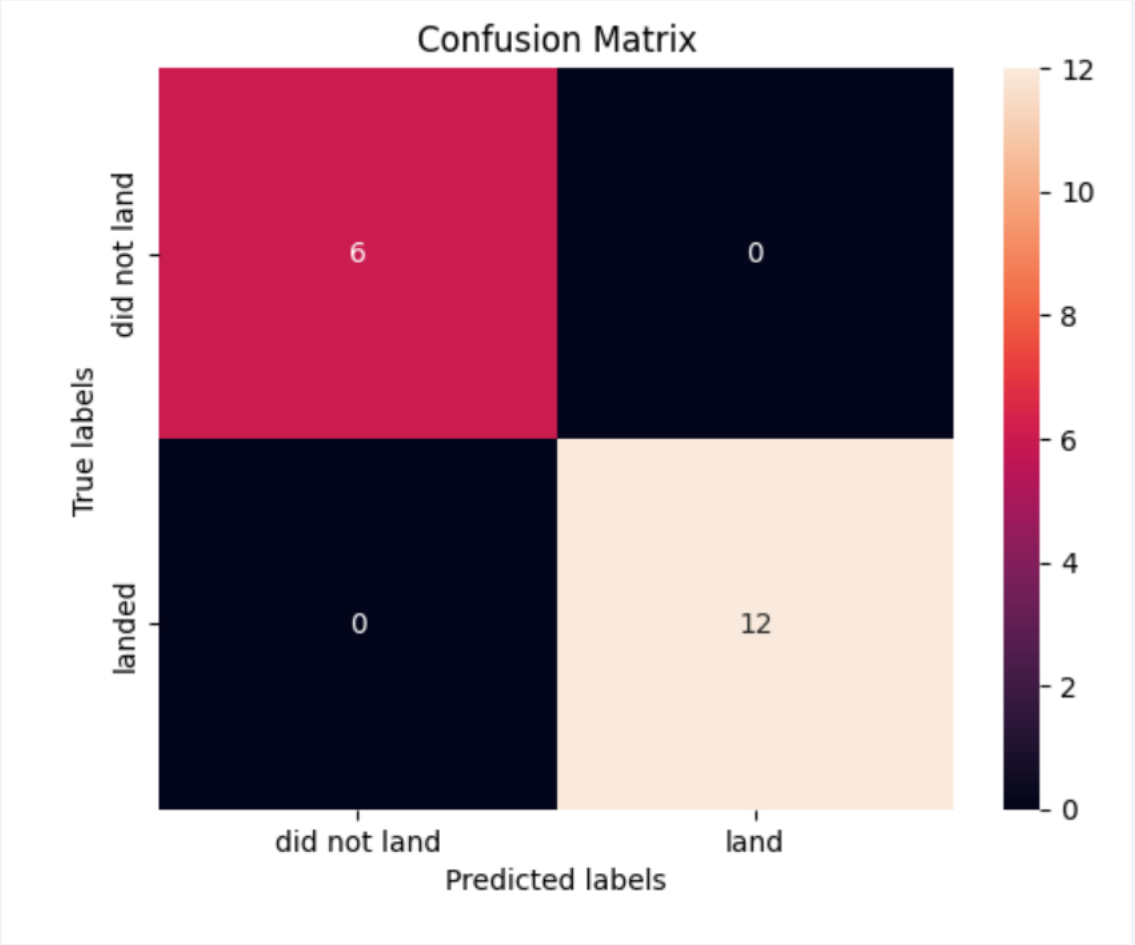
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- KNN classifies always correctly the records in the test set.



Conclusions

- Launch sites should be close to railways
- It is possible to shoot satellites on many different types of orbits
- We can predict a launch outcome using information on similar launches (knn).

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!