IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nicolas SAINTIER
November 04, 2021

Outline

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

Executive Summary

- Identifying key factors in a successfull landing is critical to lower cost and become competitive.
- To address this issue, data were collected from SpaceX (and complemented from some webpage) and then cleansed.
- We examined graphically the impact of some factors (like Launch site, Payload Mass, booster version ….) on success landing. We found that the success rate generally increases year after year and is heavily influenced by orbit type, launch site, payload mass and booster version.
- Eventually we tested some supervised predictive model to identify the most suitable for our problem. We found that Tree model performs slightly better.
- All notebooks are publicly available at https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website 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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- We want

1. to identify the factors that most impact on the success of a landing, and
2. train a supervised model to predict the success of future landings.

Section 1

Methodology

Methodology

Executive Summary

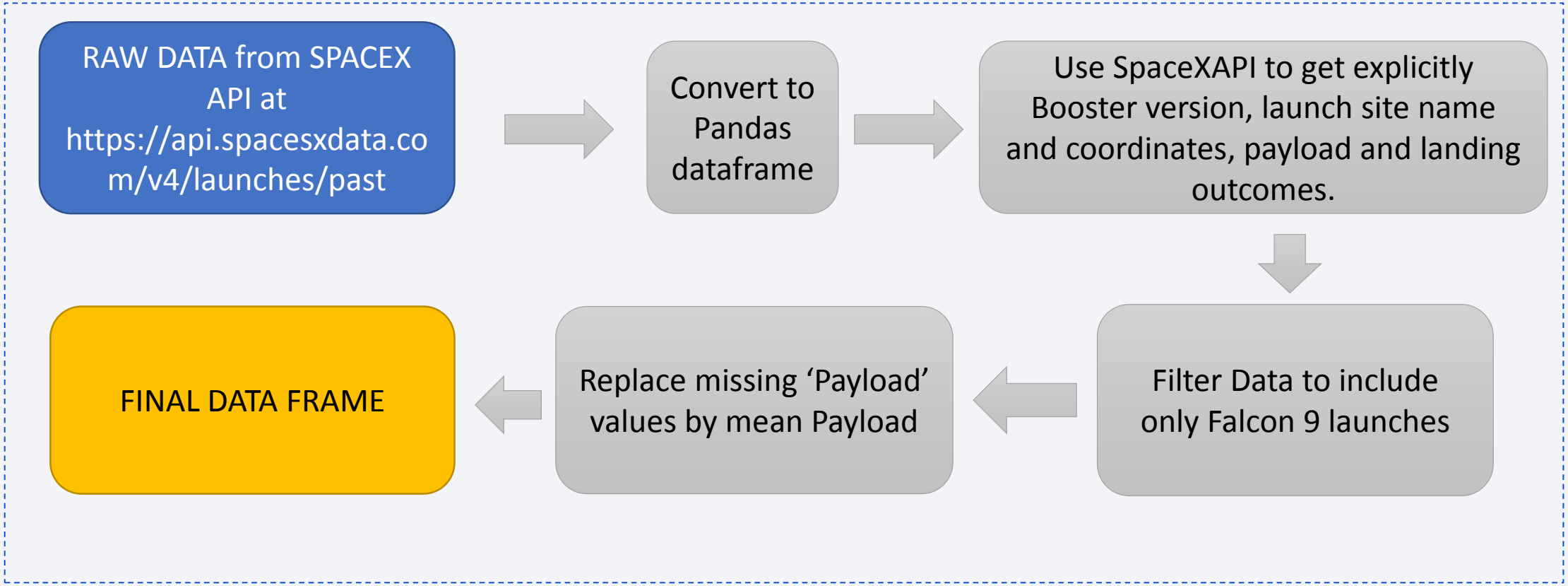
- Data collection methodology:
  - Data were collected from 2 sources: Space X (through its API) and Wikipedia webpage to get info about Falcon 9 launcher.
- Perform data wrangling
  - To prepare predictive analysis, we icreated a new column ‘Class’ containing 0 or 1 to classify outcomes as ‘bad’ and ‘good’.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After preprocessing and splitting the data into train and test set, we studied 4 supervised model (logistic regression, SVM, Tree, KNN) and studied their performance. Parameters for each models were selected through a Gridsearch.

Data Collection

Data were collected from 2 sources:

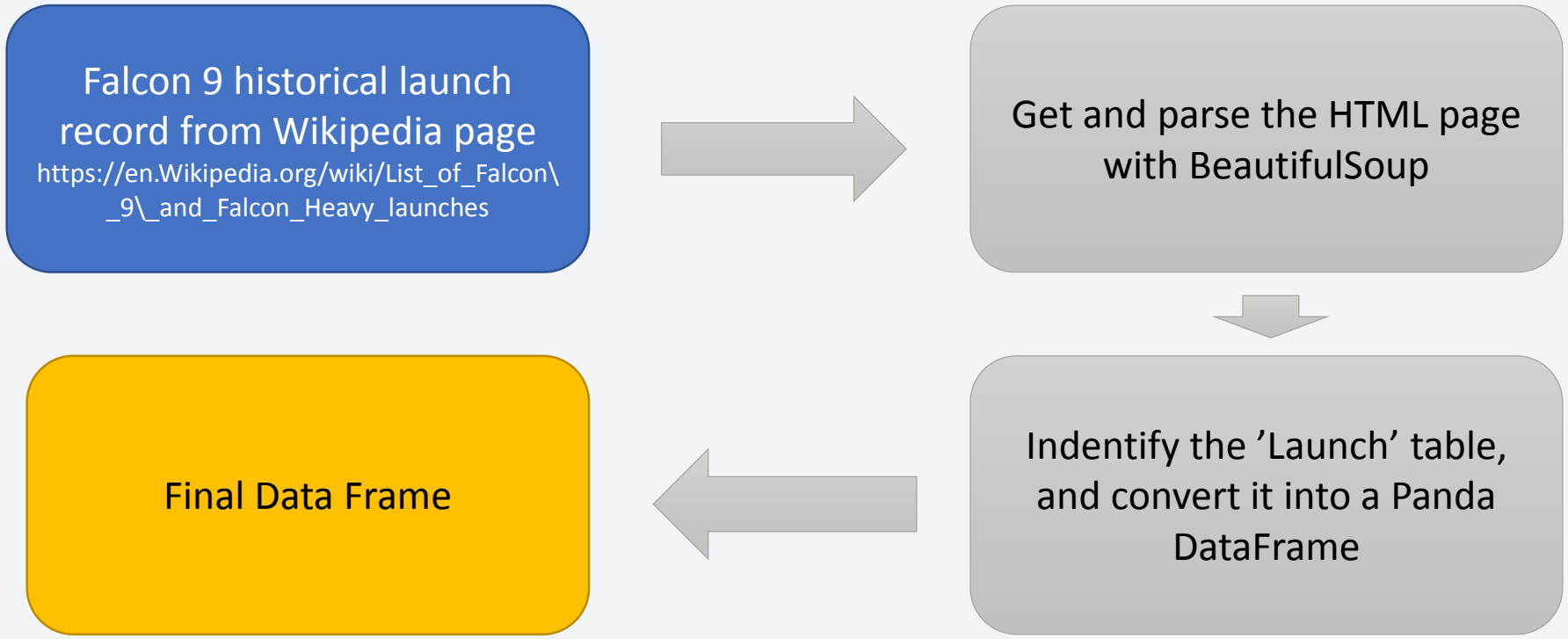
1) SpaceX API
2) web scraping of a Wikipedia pages about Falcon 9 launcher.

Data Collection – SpaceX API



- GitHub:: https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

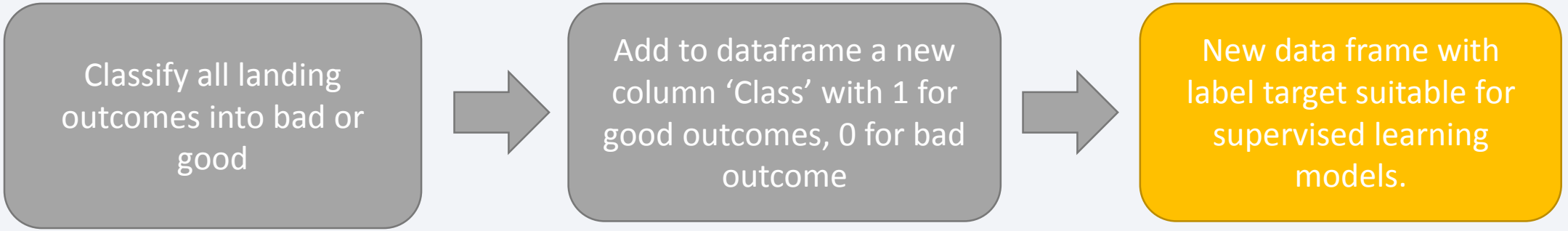
Data Collection - Scraping



GitHub URL: https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

Data Wrangling

- Purpose classify landing outcomes for future supervised learning models.



GitHub : https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

EDA with Data Visualization

We examine graphically the interplay between pairs of variable and launch outcome.

1. To assess the **interplay between PayloadMass, launch site and the outcome of the launch**, we plotted *scatter plots* of

- FlighNumber vs PayloadMass,
- FlighNumber vs LaunchSite,
- PayloadMass vs LaunchSite,

and overlay the outcome of the launch.

EDA with Data Visualization

2. To assess the **interplay between PayloadMass, Orbit type and the outcome of the launch**, we plotted

- *Bar plot* of the success rate by Orbit type,
- *Scatter plot* Orbit vs FlighNumber and Orbit vs PayloadMass (and overlay the outcome of the launch for both plots)

3. We also investigate the **yearly trend of success rate** plotting

GitHub: : https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

EDA with SQL

We performed the following queries in SQL:

1. *Display the names of the unique launch sites in the space mission*
2. *Display 5 records where launch sites begin with the string 'CCA'*
3. *Display the total payload mass carried by boosters launched by NASA (CRS)*
4. *Display average payload mass carried by booster version F9 v1.1*
5. *List the date when the first successful landing outcome in ground pad was acheived.*
6. *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*

EDA with SQL

7. List the total number of successful and failure mission outcomes

8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery

9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

GitHub: : https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

Build an Interactive Map with Folium

- In order to **assess the impact of the launch sites location on the success launch rate**, we created an interactive map of EEUU with Folium to which we added:

i. Circles at the 4 launch sites with pop-up markers indicating the launch site name,

ii. For each launch site, markers for all the launches with color green/red to indicate success/failure of the launch,

iii. Lines from each launch sites to the closest city, transport red (highway or railroad), sea coast.

- This allows to **quickly evaluate the success rate of each launch site and the possible relations with geographic** (proximity of the sea) **and infrastructure elements** (nearby city, road/railroad).

GitHub:: https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

Build a Dashboard with Plotly Dash

- We created an **interactive dashboard** (using Dash and Plotly) where the user can select the launch site and see

1. The proportion of success and failed launches with a *pie chart*,
2. A *scatter plot* displaying the success/failed launches VS payload mass (payload mass range selected by the user) and the influence of the booster version.

- This allows to **quickly see the success rate of each launch site** and appreciate **its relation with payload mass** and **the booster version**.

GitHub: https://github.com/NicolasSaintier/IBM-Certificate-Coursera-capstone-project-SpaceX

Predictive Analysis (Classification)

Data preprocessing (standardization)
Data splitting into train and test set)

We considered **4 supervised models:**
1. logistic regression,
2. SVM,
3. Tree,
4. KNN.

For each model:
1. we performed a GridSearch to find the best parameter set on the train set.
2. Compute the accuracy on test set

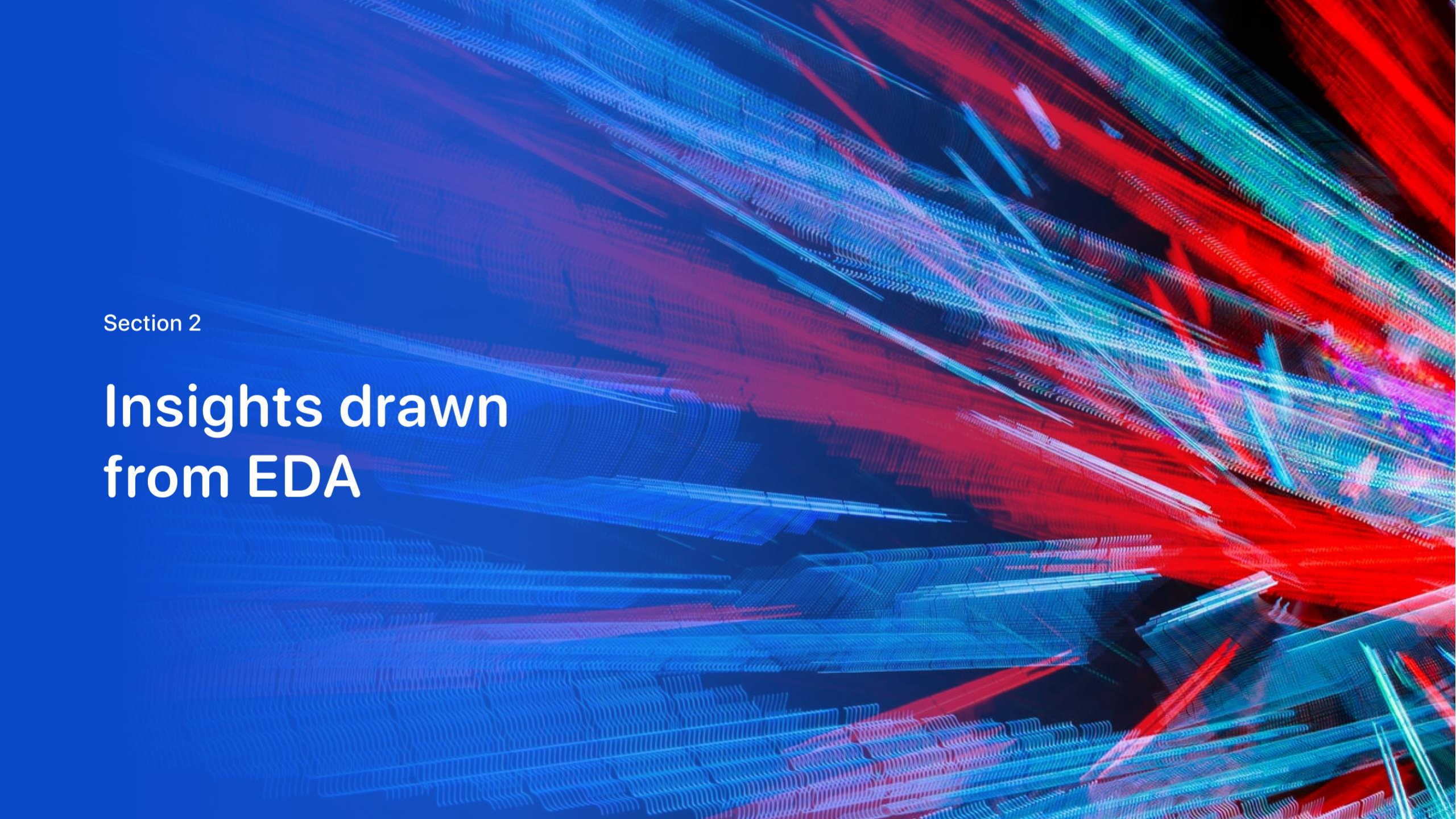
Graphical visualization:
1. Bar plot of accuracy and score
2. Plot of Confusion matrix on test set.

Results

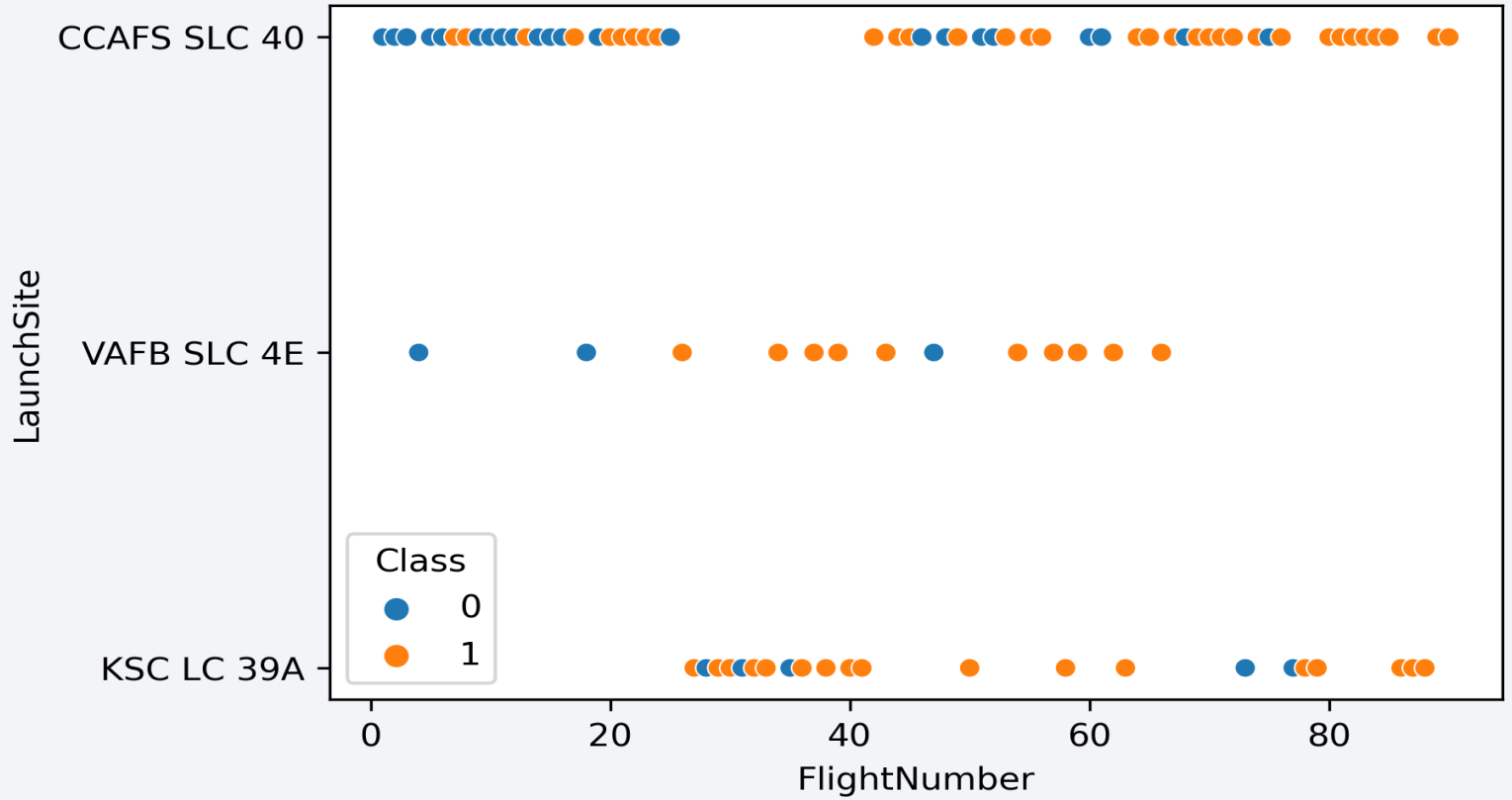
- **The success rate generally increases year after year.**
- Influence of orbit type: orbits **ES-L1, GEO, HEO, SSO** have the highest success rate.
- Influence of launch site: launch site KSC-LC-39A has the highest success rate,
- Influence of payload mass: most successful payload range is **3000-4000 kg** (**70% success rate)**.
- **Influence of booster version: B4** is the most successful Booster version (**45.5% success**)
- Predictive analysis results: We tested 4 predictive models: Logistic regression, SVM, Tree, and KNN. **Tree models seems to perform slightly better than others**.

Section 2

Insights drawn from EDA

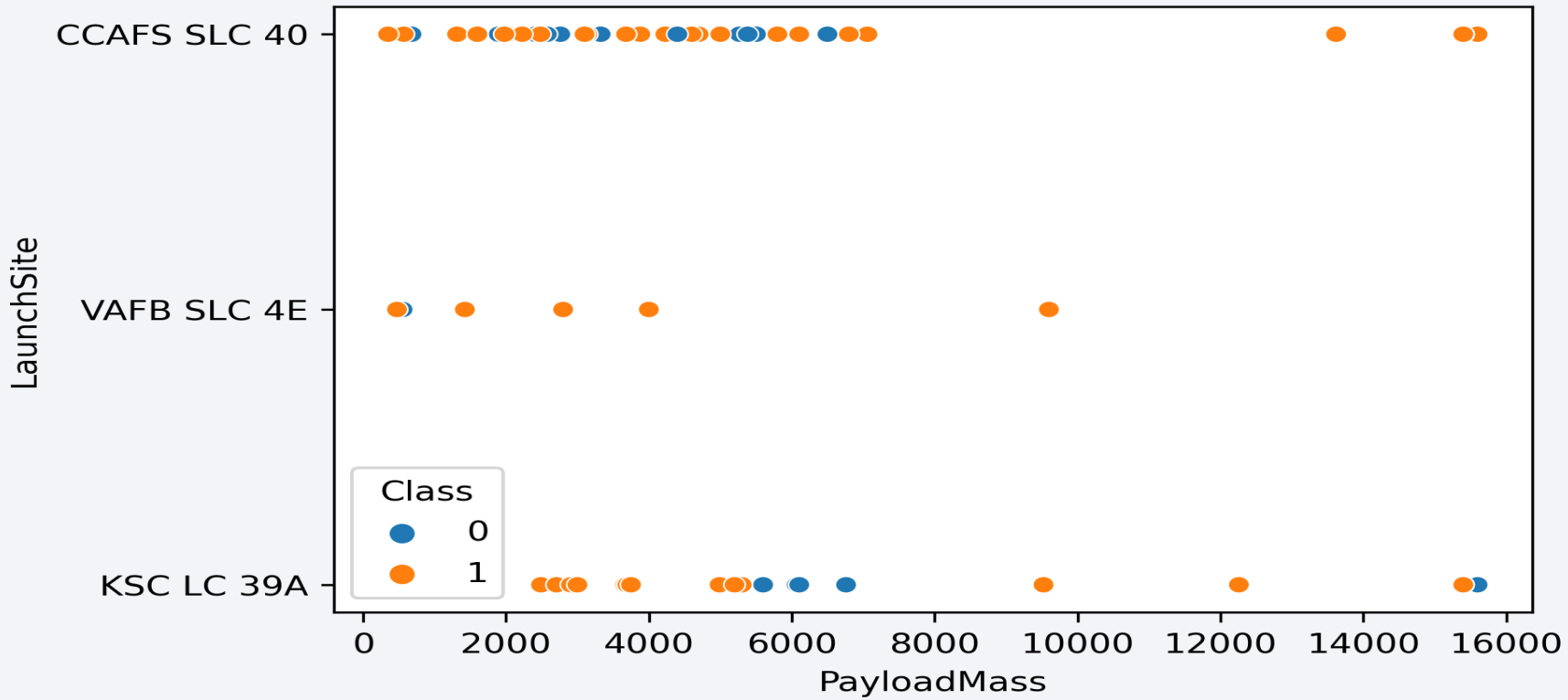


Flight Number vs. Launch Site



- Success rate: 60% for CCAFS LC-40, 77% for KSC LC-39A and VAFB SLC 4E
- Success rate seems to increase with time

Payload vs. Launch Site

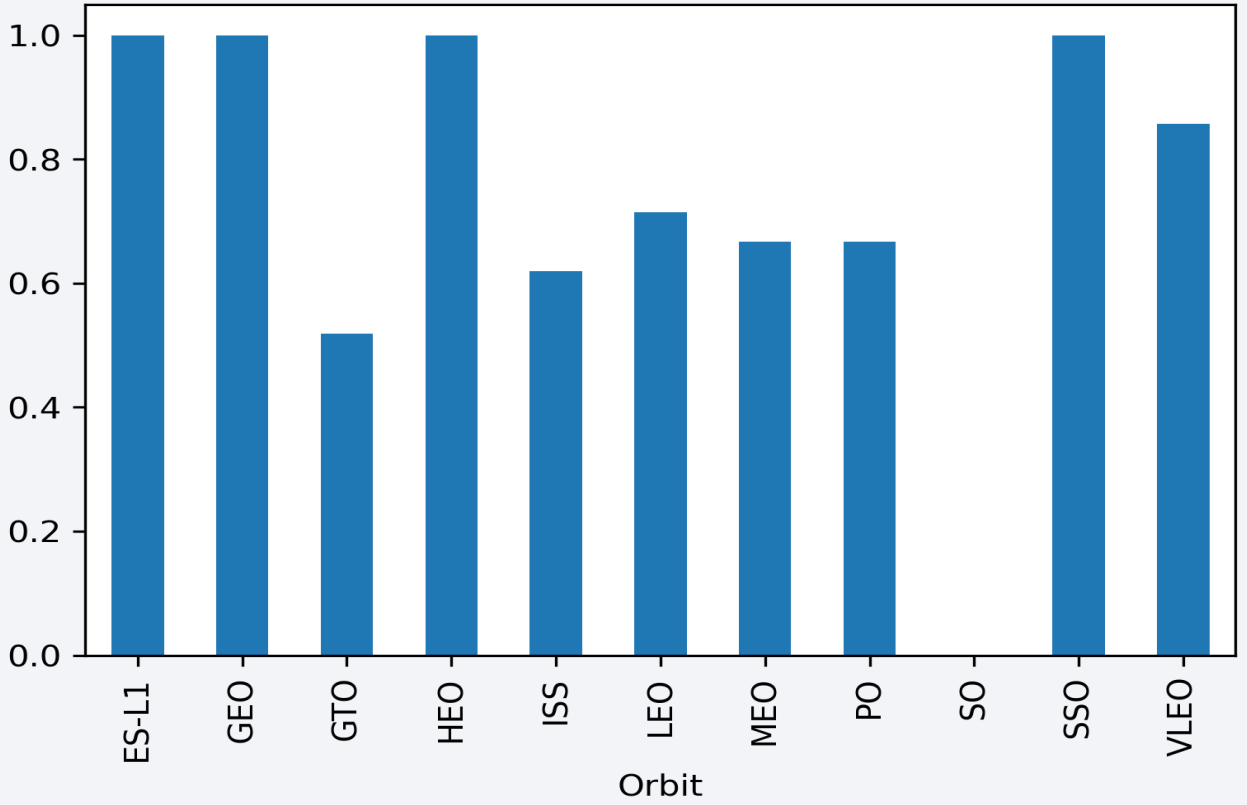


- Failure are concentrated in the mid-range 3000-7000 due to CCAFS SLC-40 and KSC LC 39A.
- Low (<3000) and high (>7000) shows only success for all 3 sites.

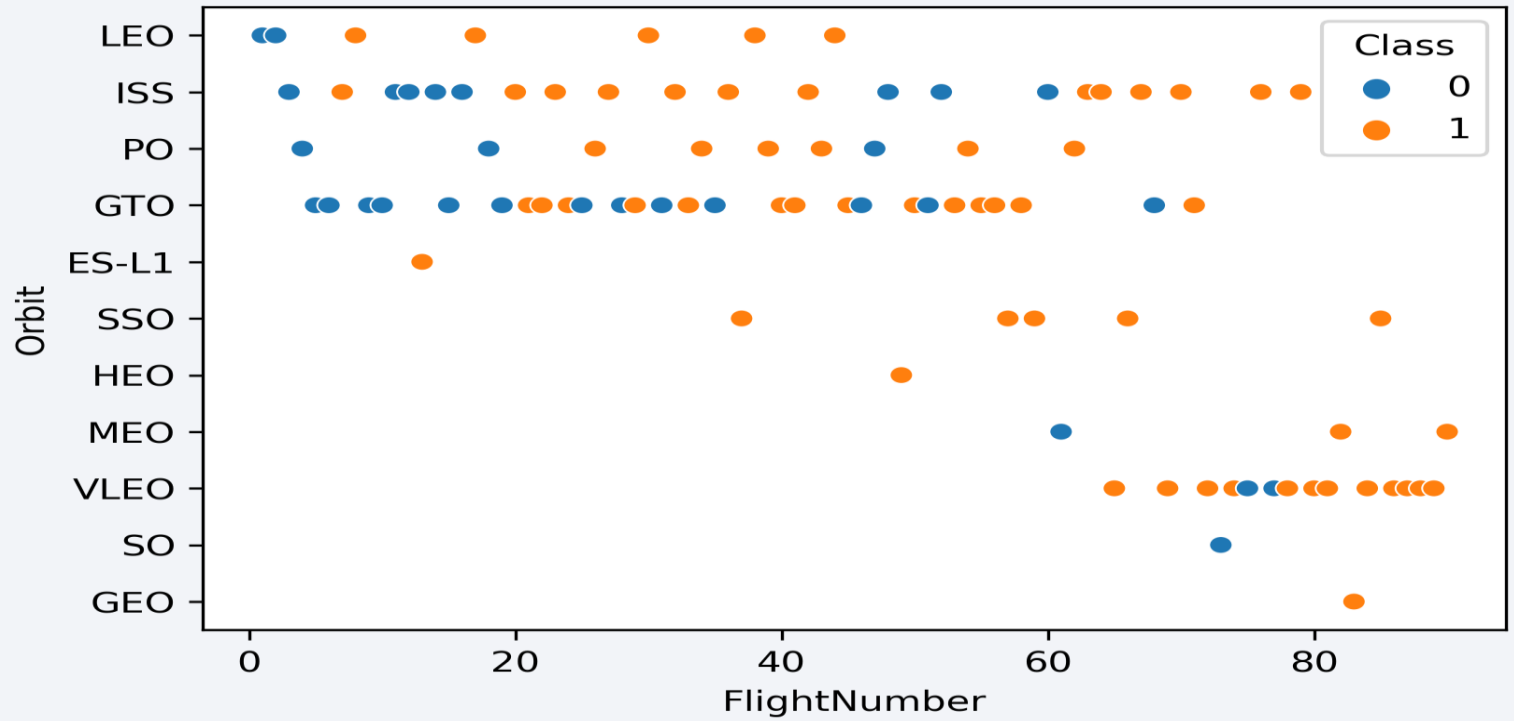
Success Rate vs. Orbit Type

Orbits can be split into **3 groups**:

- 100% success: ES-L1, GEO, HEO, SSO
- 80% success: VLEO
- <70%: GTO, ISS, LEO, MEO, PO

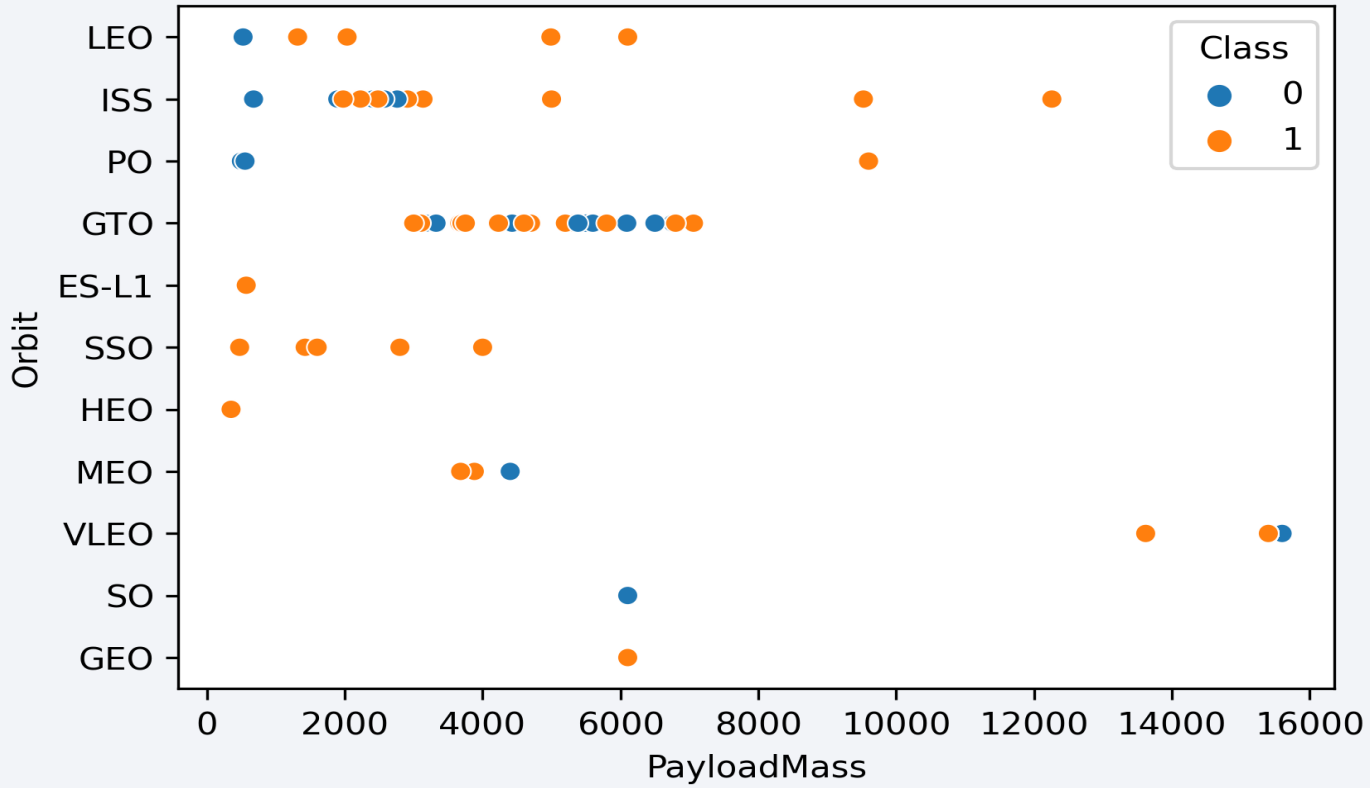


Flight Number vs. Orbit Type



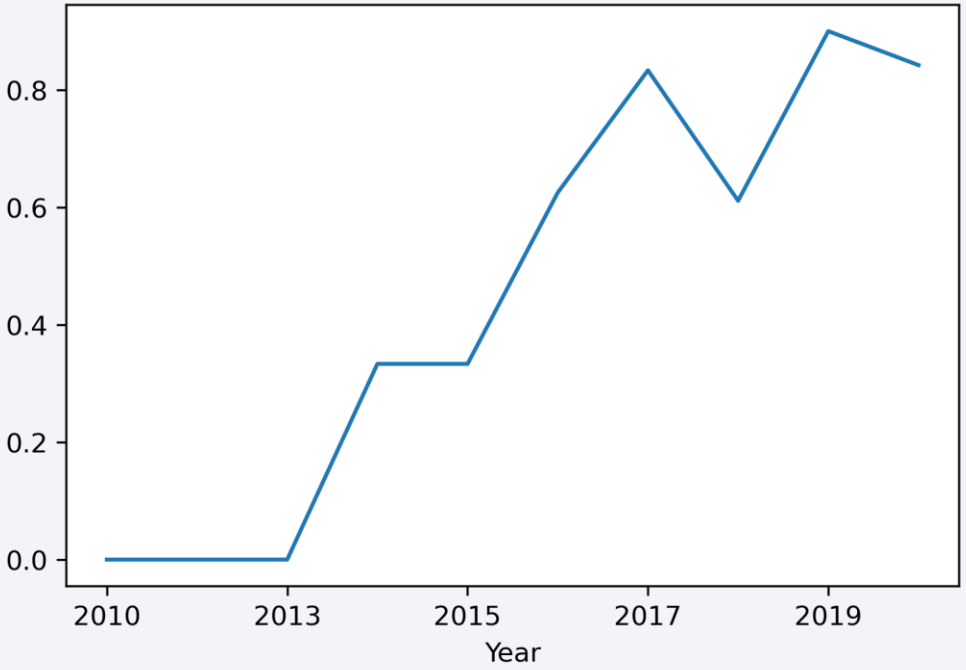
- LEO orbit success appears related to the number of flights
- there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Heavy payloads have a negative influence on GTO, VLO orbits
- positive on PO, LEO, ISS orbits.

Launch Success Yearly Trend



- Since 2013, **the success rate increases** (except in 2017 and 2019).

All Launch Site Names

Task 1

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

In [5]:
```
%%sql
select distinct LAUNCH_SITE from SPACEXDATASET;
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[5]:

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

- The launch sites are **CCAFS-LC-40, CCAFS SLC-40, KSC LC 39A**, and **VAFB SLC 4E**

5 Launch Site Names Beginning with 'CCA'

Task 2

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

In [6]:
```
%%sql
select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5;
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[6]:

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 carried by boosters from NASA

Task 3

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

In [7]:
```
%%sql
select PAYLOAD_MASS__KG_ from SPACEXDATASET where CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[7]:

payload_mass__kg_
500
677
2296
2216
2395
1898
1952
3136
2257
2490
2708
3310
2205
2647
2697
2500
2495
2268
1977
2972

Average Payload Mass by F9 v1.1

Task 4

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

In [8]:
```
%%sql
select SUM(PAYLOAD_MASS__KG_)/count(*) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

1
2928

- The average payload mass by booster version F9 v1.1 is **2928** kg.

Dates of the First Successful Ground Landing

Task 5

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

Hint:Use min function

In [9]:
```
%%sql
select DATE,LANDING__OUTCOME  from SPACEXDATASET where LANDING__OUTCOME like 'Success%' ;
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

DATE	landing__outcome
2015-12-22	Success (ground pad)
2016-04-08	Success (drone ship)
2016-05-06	Success (drone ship)
2016-05-27	Success (drone ship)
2016-07-18	Success (ground pad)
2016-08-14	Success (drone ship)
2017-01-14	Success (drone ship)

The first successful landing outcome on ground pad occurred in 2015-12-22, 2016-04-08, …..

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [10]:
```
%%sql
select BOOSTER_VERSION from SPACEXDATASET where (LANDING__OUTCOME like 'Success%ship%') and (PAYLOAD_MASS__KG_ between 4000 and
6000);
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[10]:

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

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are: F9FT B1022, F9FT B1026, F9FT B1021.2, F9FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [11]:
```
%%sql
select MISSION_OUTCOME, count(*) as total from SPACEXDATASET group by MISSION_OUTCOME;
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

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

- There was 1 failure in flight, 99 success (+1 with payload status unclear)

Boosters Carried Maximum Payload

Task 8

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

In [12]:
```
%%sql
select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXDATASET
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET);
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[12]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

Task 9

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

In [13]:
```
%%sql
select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXDATASET
where YEAR(DATE)=2015 and LANDING__OUTCOME like 'Failure%';
```

```
 * ibm_db_sa://zps87196:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

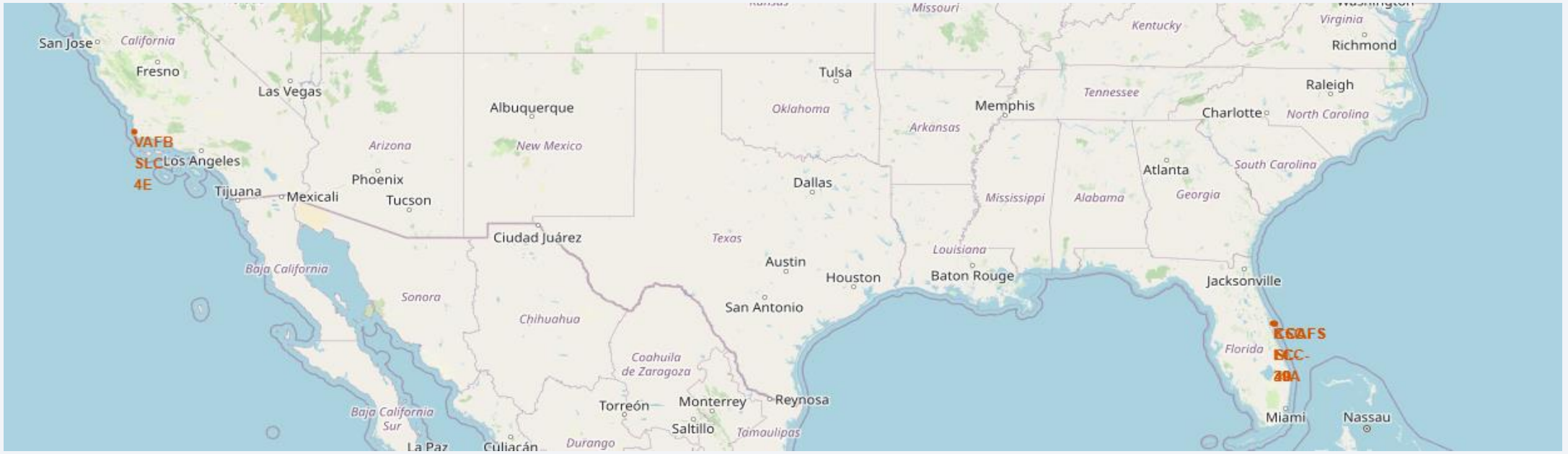
Out[13]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Section 4

Launch Sites Proximities Analysis

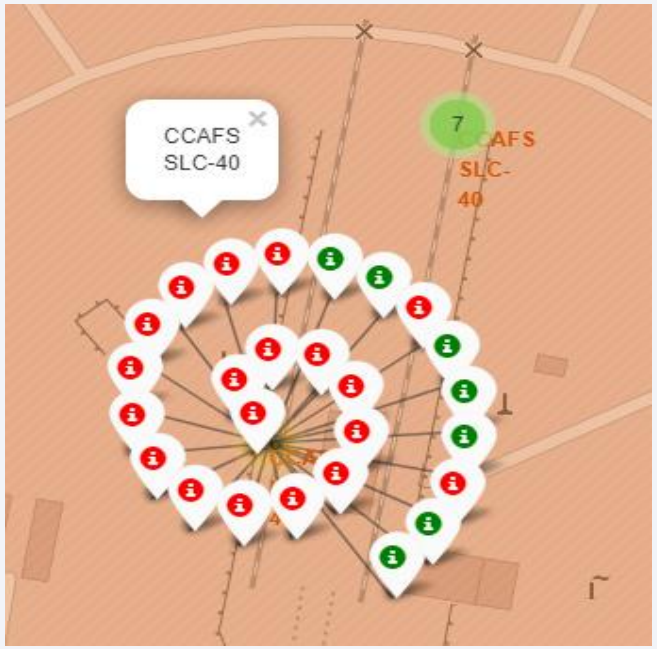
Launch site location

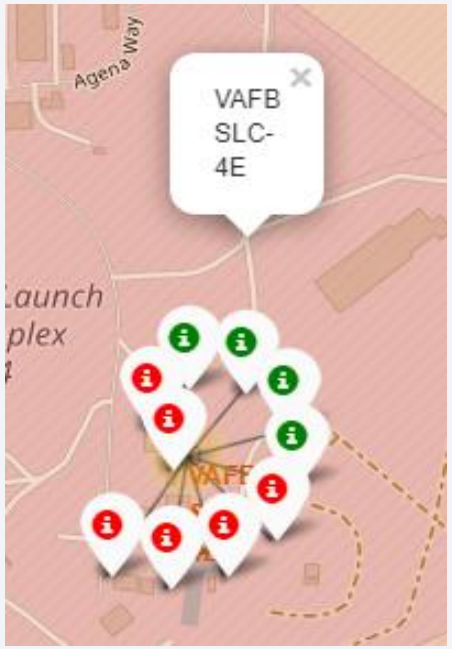


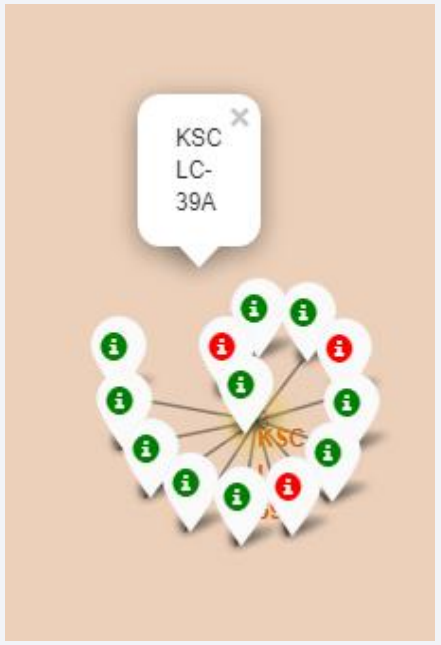
3 sites on the west coast (Florida), 1 site on the East coast (California).

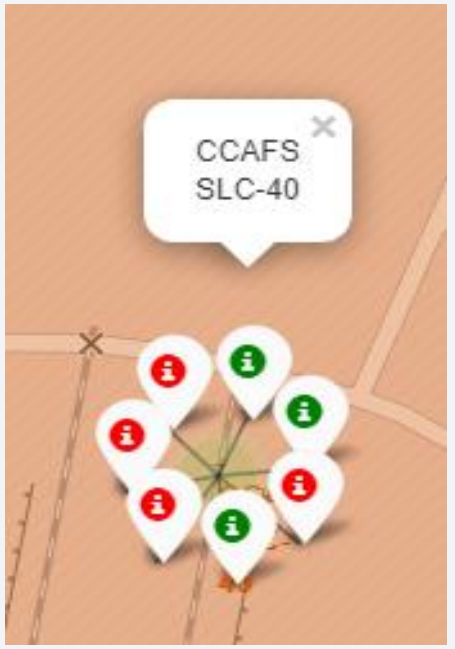
All the sites are very close to the coast.

Success/failed launches for each site



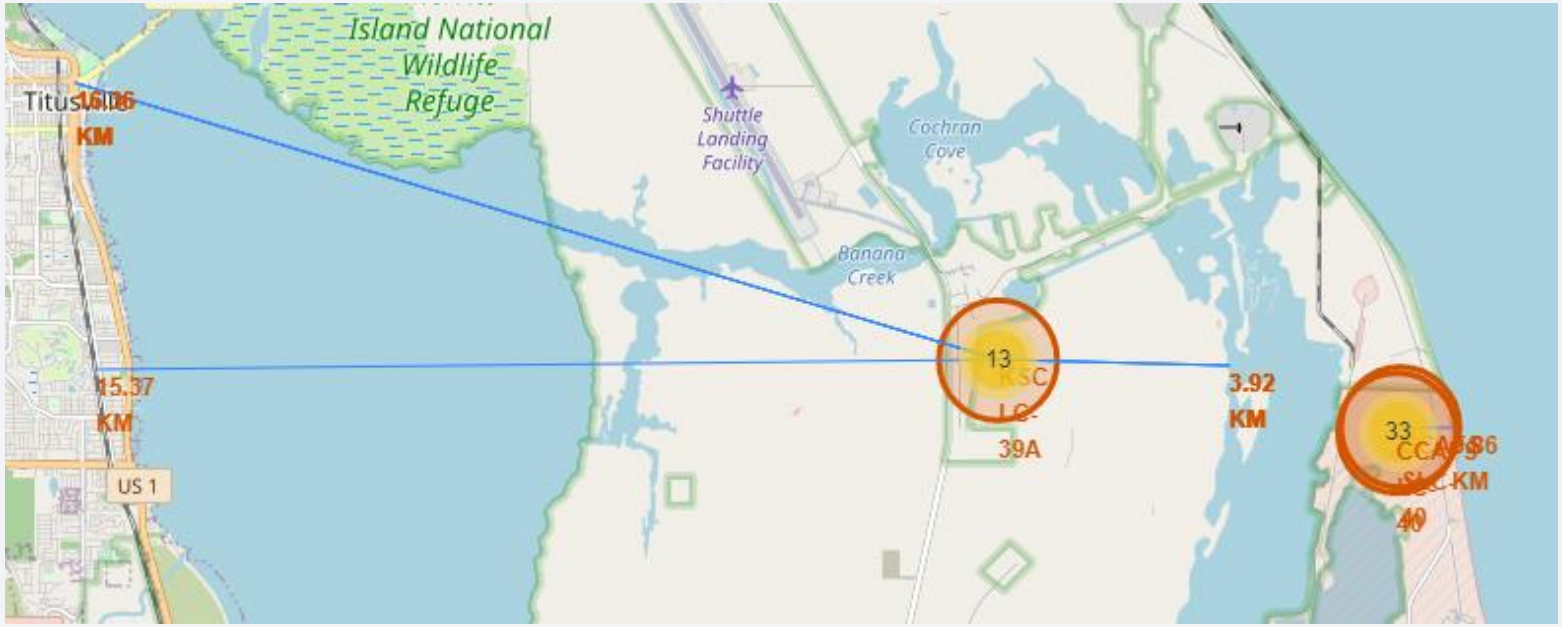






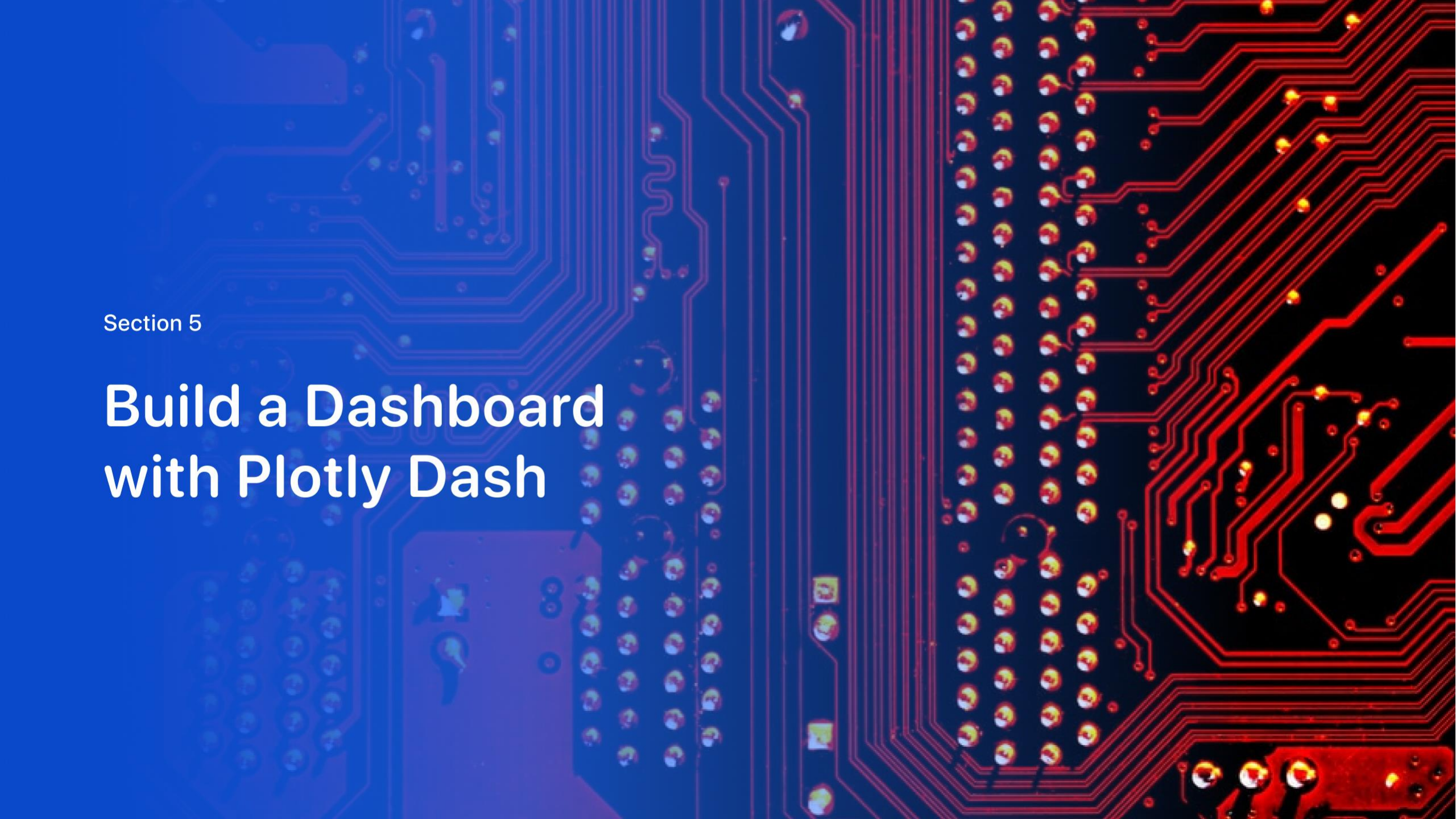
Relations between launch site and surrounding

- Launch sites are far from cities (~15km to nearest city), and close to the coast (1 to 3 km). Distance to highway/railroad vary.

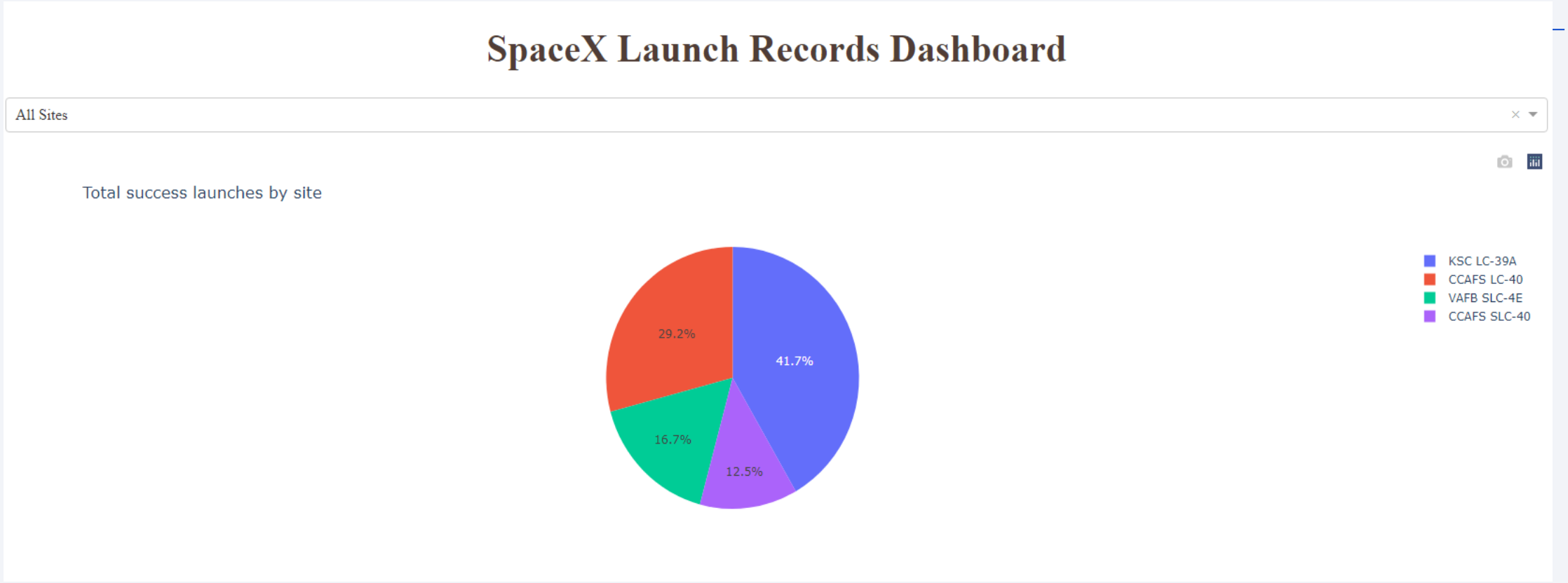


Section 5

Build a Dashboard with Plotly Dash



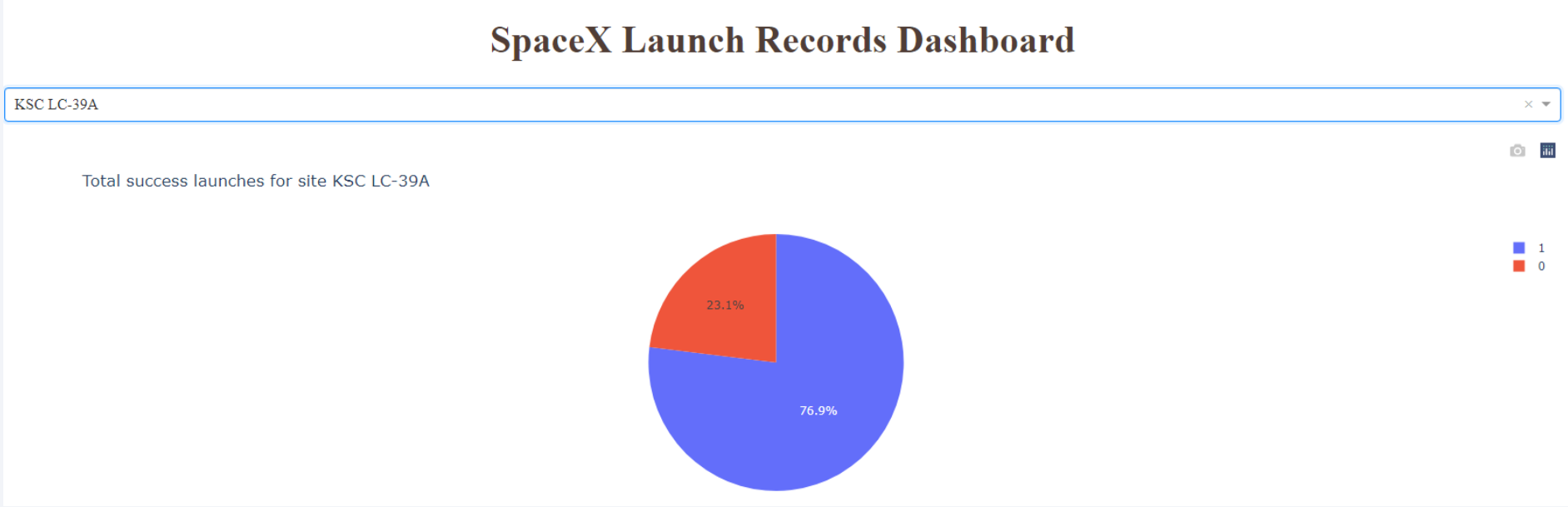
Success rate of the distinct launch sites



- Launch site ***KSC-LC-39A*** **has by far the highest *number* of successful launches w.r.t. total number of launches** (~42% - the 2nd one has ~29%).

Detailed analysis of the most successful launch site.

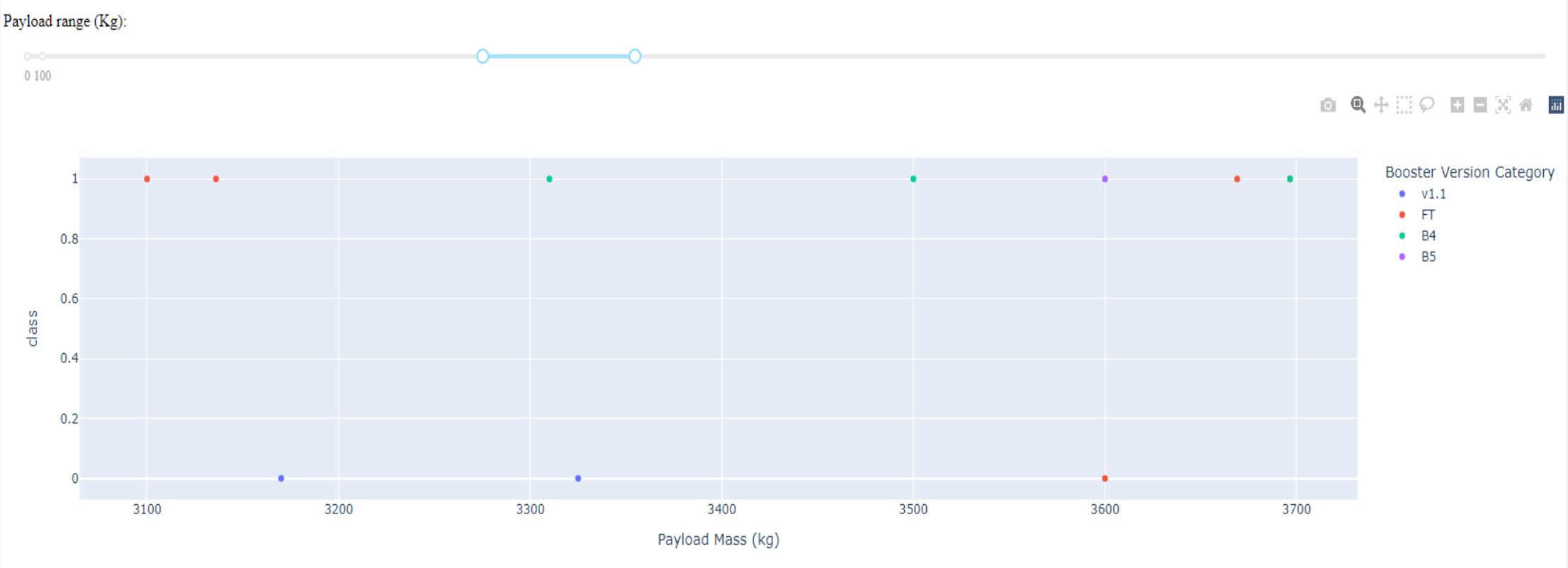
- We saw KSC-LC-39A has by far the highest total number of success launches.



- This is due to (i) it has **the highest success rate**, and

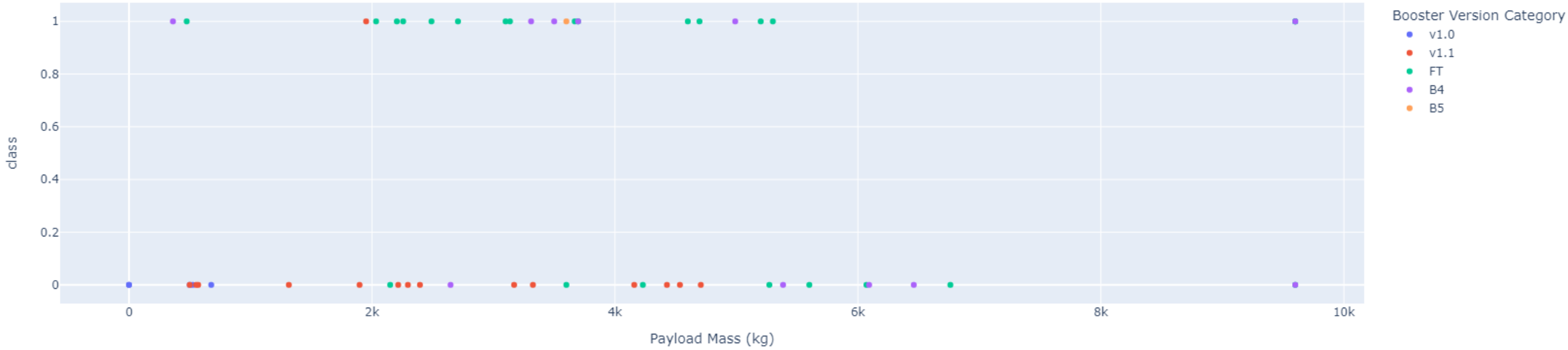
  (ii) it is the **2nd most used launch site** (after CCAFS LC 40).

Launch outcomes VS payload mass



- The **most successful payload range is 3000-4000 kg**: **70% success rate**

Launch outcomes VS Booster version

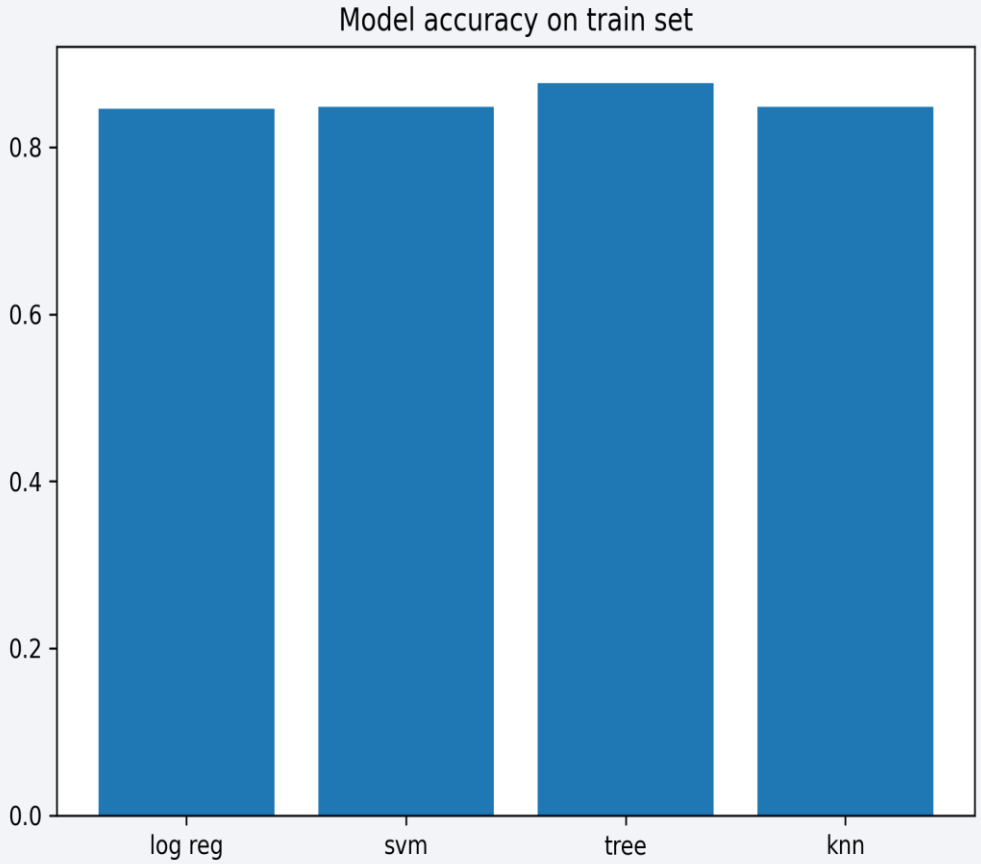


- The **most successful Booster version is B4** (45.5% success)

Section 6

Predictive Analysis (Classification)

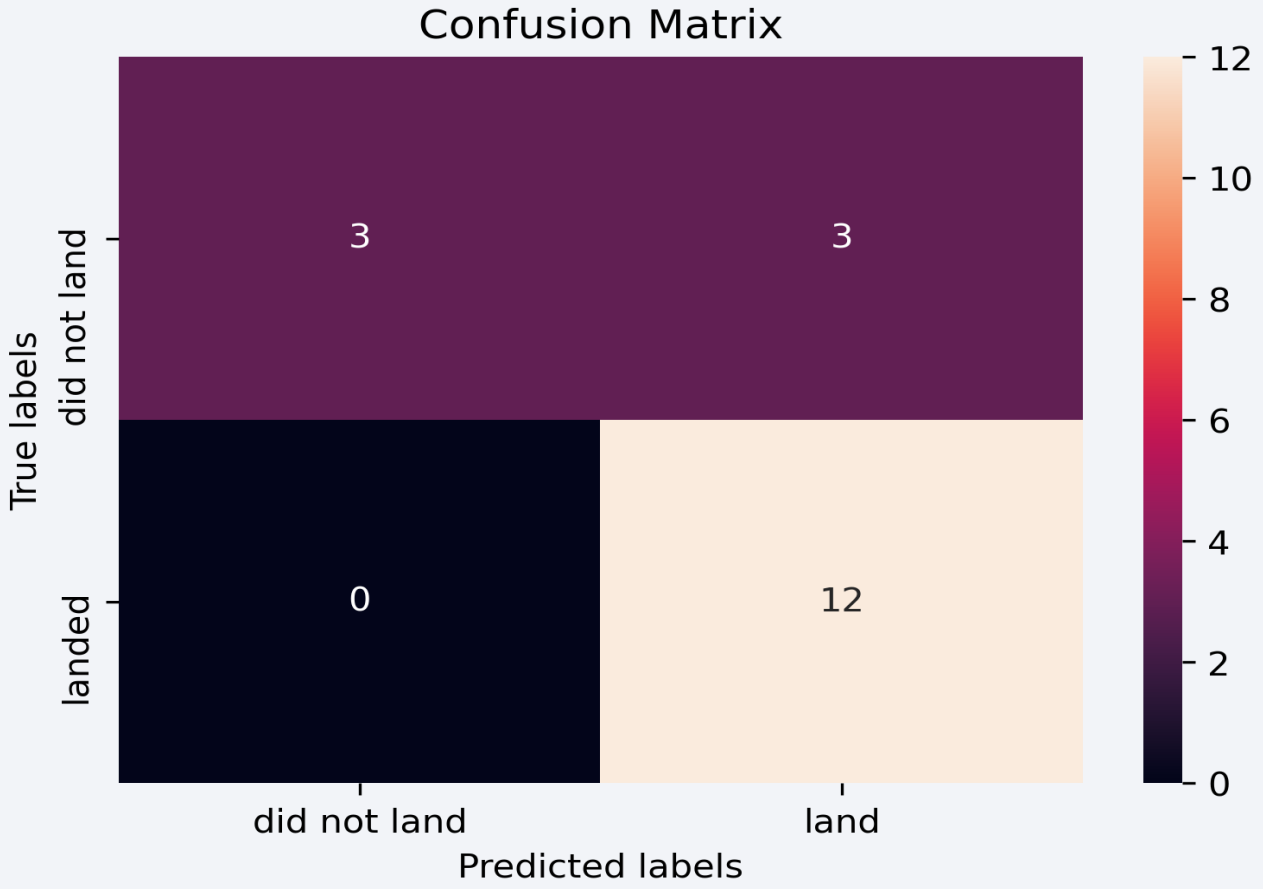
Classification Accuracy



- All 4 models performs equally on the test set. Tree performs slightly better on train set. **I recommend considering Tree model.**

Confusion Matrix

All 4 considered models have the same confusion matrix:



Conclusions

- We considered 4 models: logistic regression, SVM, Tree model, KNN.
- All 4 models performed identically on test set.
- Tree performed slightly better on train set.
- All 4 models present the same confusion matrix exhibiting problems with false positive.
- Conclusion: **I recommend using Tree model.**

Thank you!