Winning Space Race with Data Science

Stefano Lanza

07/06/2023

Outline

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

Executive Summary

Summary of methodologies:

- Data collection
- Data wrangling
- Exploratory Data Analysis
- Generating (interactive) data visualizations
- Predictive Analysis via Machine Learning techniques

Summary of all results:

- Exploratory Data Analysis results, supported by Data Visualizations
- Interactive map built with Folium, and an interactive Dashboard
- Predictions obtained employing four Machine Learning techniques

Introduction

Project background:

- Commercial space flights provide an alternative to public agencies-operated flights
- Allegedly, SpaceX is the cheapest, with costs around 62 million dollars per flight (against an average of 165 million dollars of other companies)
- SpaceX reduces the costs because the rocket's first stage can be recovered…
- …but the landing of the first stage is not always successful.

Questions answered:

- Is the success rate increasing over the years?
- What are the variables that influence the most the success of the landing of the first stage?
- How can we model the success rate, in terms of the variables?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - via SpaceX Rest API and web scraping from the SpaceX Wikipedia page
- Perform data wrangling
  - Filtering and cleaning the data, with one-hot encoding for categorical data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - building, tuning, evaluating the classification models

Data Collection

In order to get a more complete picture, data about SpaceZ's rocket launches were collected from two different sources: the SpaceX REST API and SpaceX's Wikipedia page.

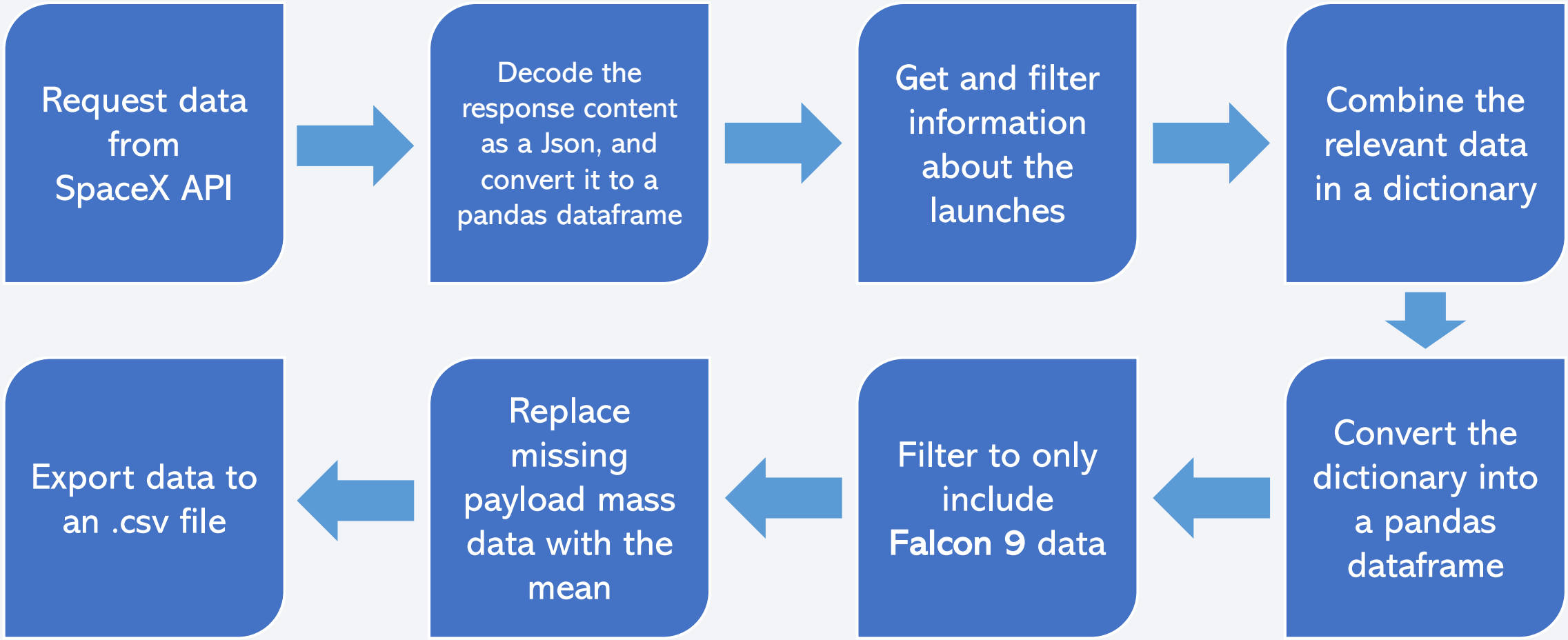
Data in SpaceX REST API:

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

Data in SpaceX's Wikipedia page:

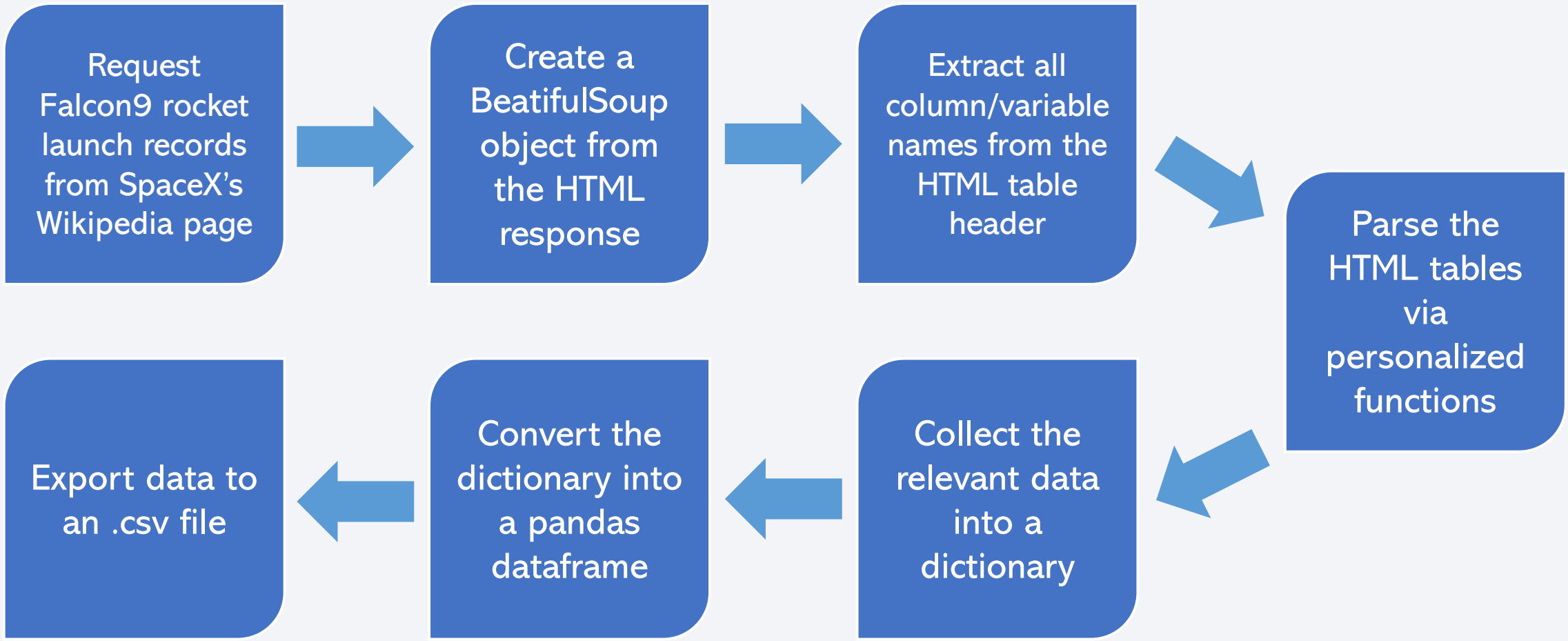
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub Link: Data Collection REST API

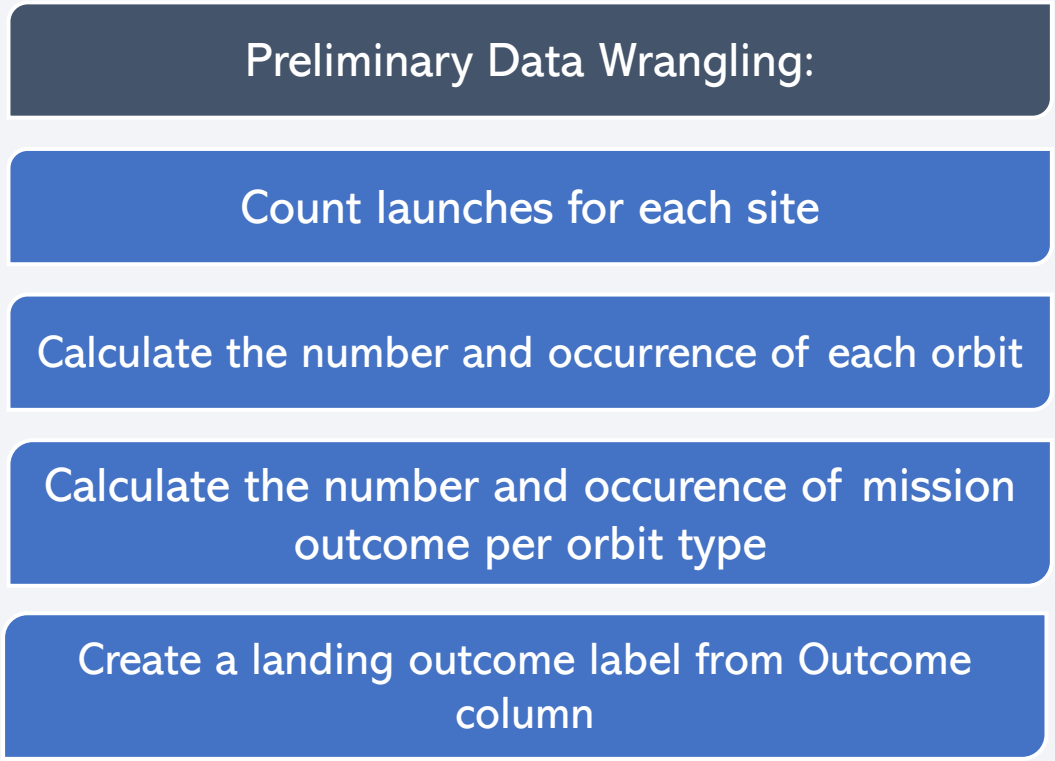
Data Collection – Web Scraping



GitHub Link: Data Collection Web Scraping

Data Wrangling

- The data previously collected from Web Scraping and REST API were imported into a pandas dataframe.
- We performed a preliminary data wrangling, helpful for the subsequent steps of the analysis. The transformed data considered is indicated in the flowchart.
- In particular, we created a landing outcome label, that is 1 if the first stage landed successfully, and 0 otherwise.



GitHub Link: Data Wrangling

EDA with Data Visualization

Charts plotted:

- Flight Number vs. Payload Mass (scatter plot)
- Flight Number vs. Launch Site (scatter plot)
- Payload Mass vs. Launch Site (scatter plot)
- Orbit Type vs. Success Rate (bar chart)
- Flight Number vs. Orbit Type (scatter plot)
- Payload Mass vs Orbit Type (scatter plot)
- Success Rate Yearly Trend (line chart)

GitHub Link: EDA with Data Visualization

EDA with SQL

The SQL queries we performed are:

- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster versions which have carried the maximum payload mass, using a subquery
- Listing the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub Link: EDA with SQL

Build an Interactive Map with Folium

Employing Folium, we created interactive maps, with:

- Markers for all SpaceX Falcon 9 launch sites
  (with additional markers of NASA launch sites for reference)
- Markers for every Falcon 9 launch, colored in green if the first stage landing was successful, and red otherwise.
- Distances to relevant proximities (e.g. coastline, railway, cities) highlighted.

GitHub Link: Maps with Folium

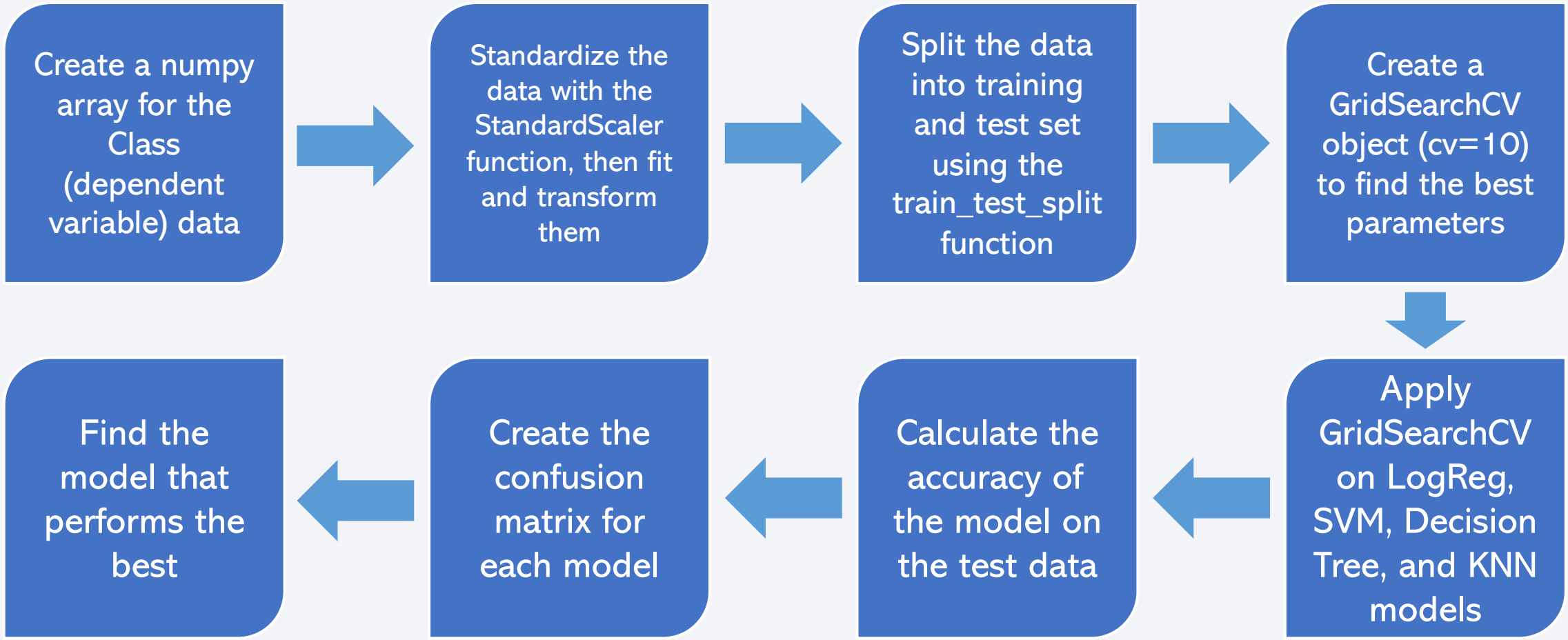
Build a Dashboard with Plotly Dash

With Plotly Dash, we created an interactive dashboard containing:

- a Launch Site Drop-down Input Component;
- a callback function to render success-pie-chart based on selected site dropdown;
- a Range Slider to Select Payload;
- a callback function to render the success-payload-scatter-chart scatter plot.

GitHub Link: Interactive Dashboard

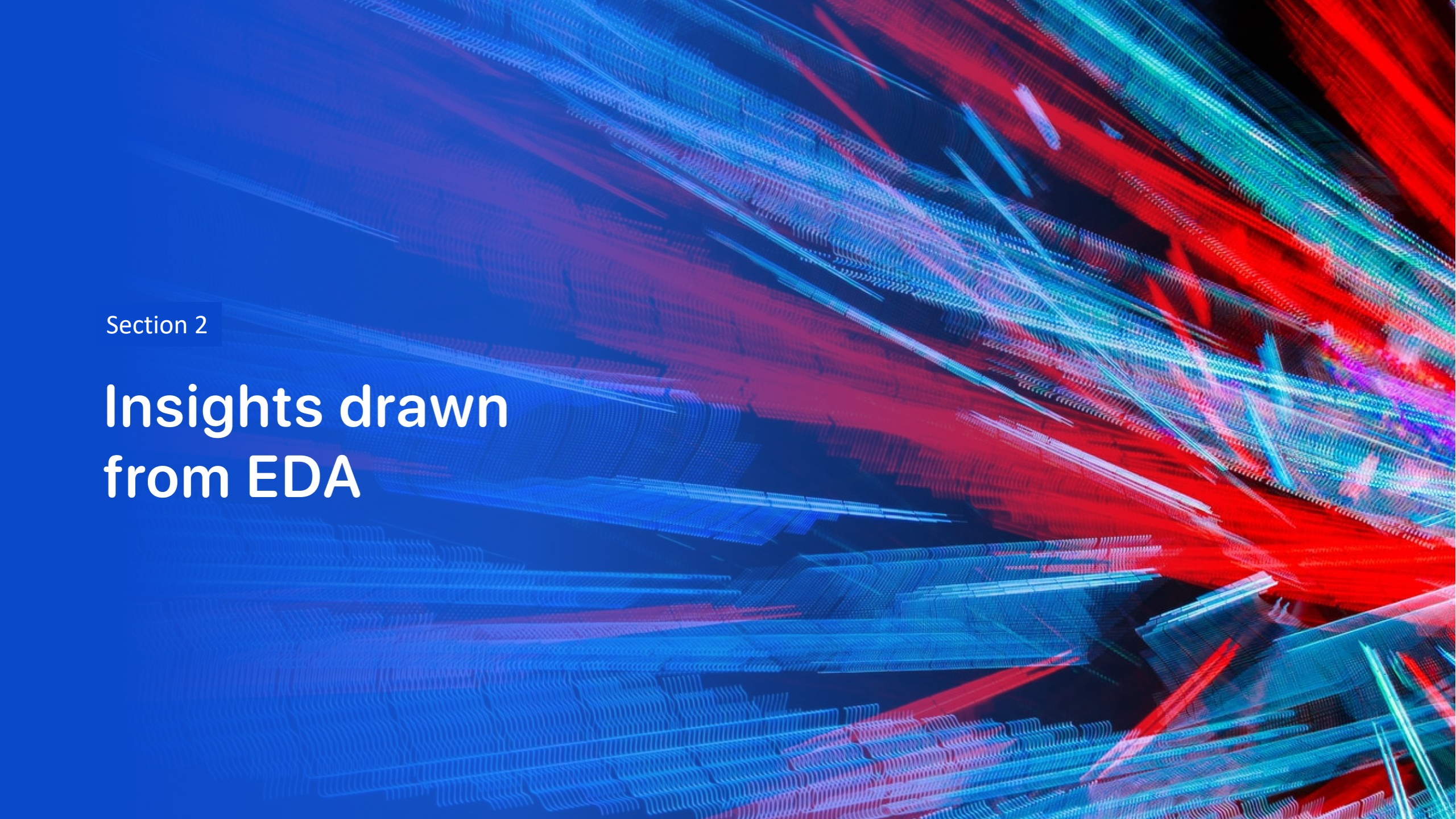
Predictive Analysis (Classification)



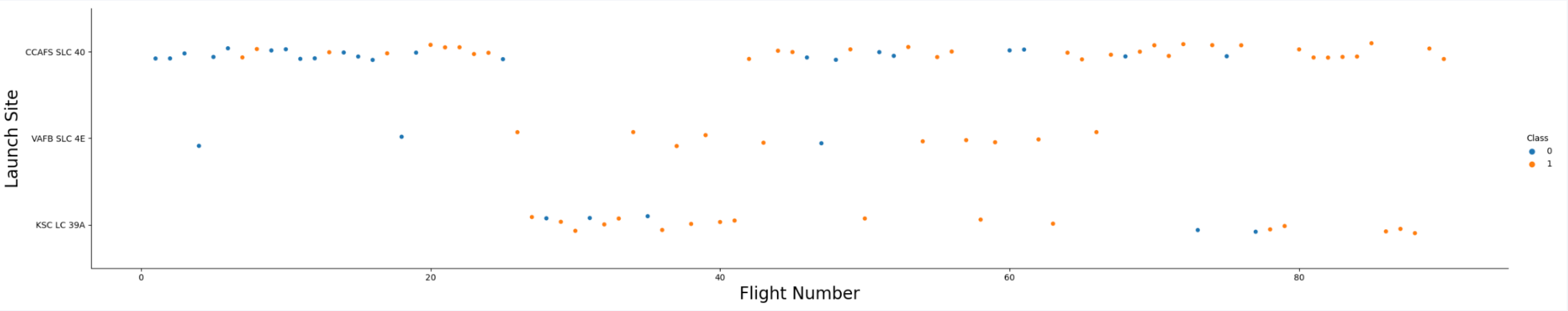
GitHub Link: Machine Learning Predictive Analysis

Results

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

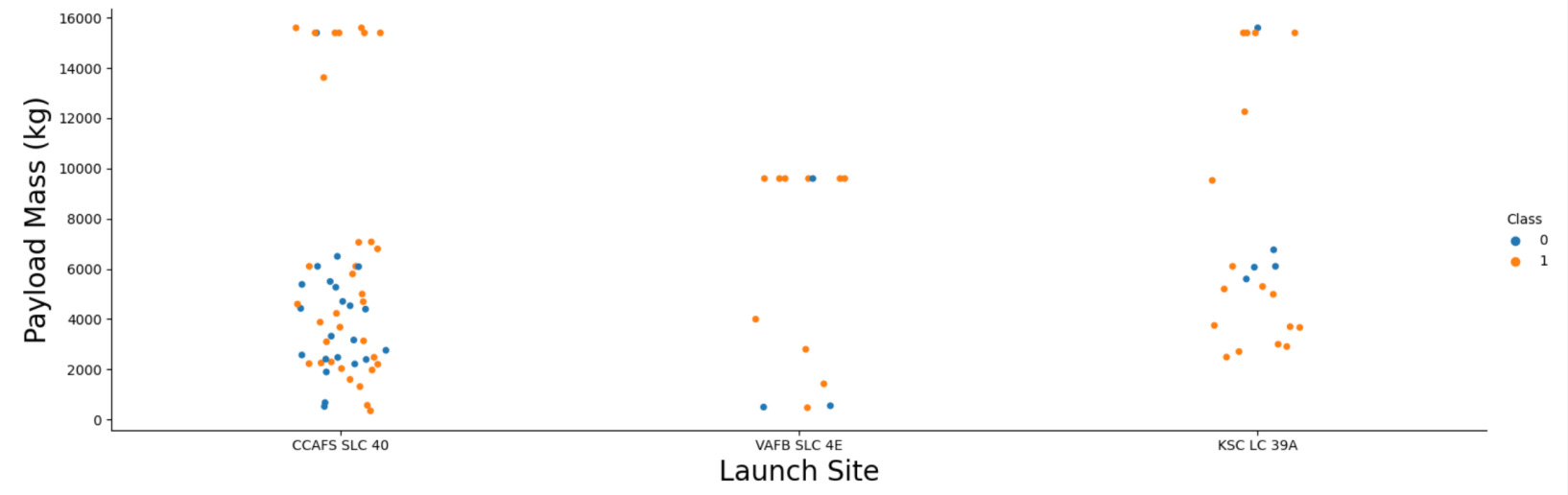

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



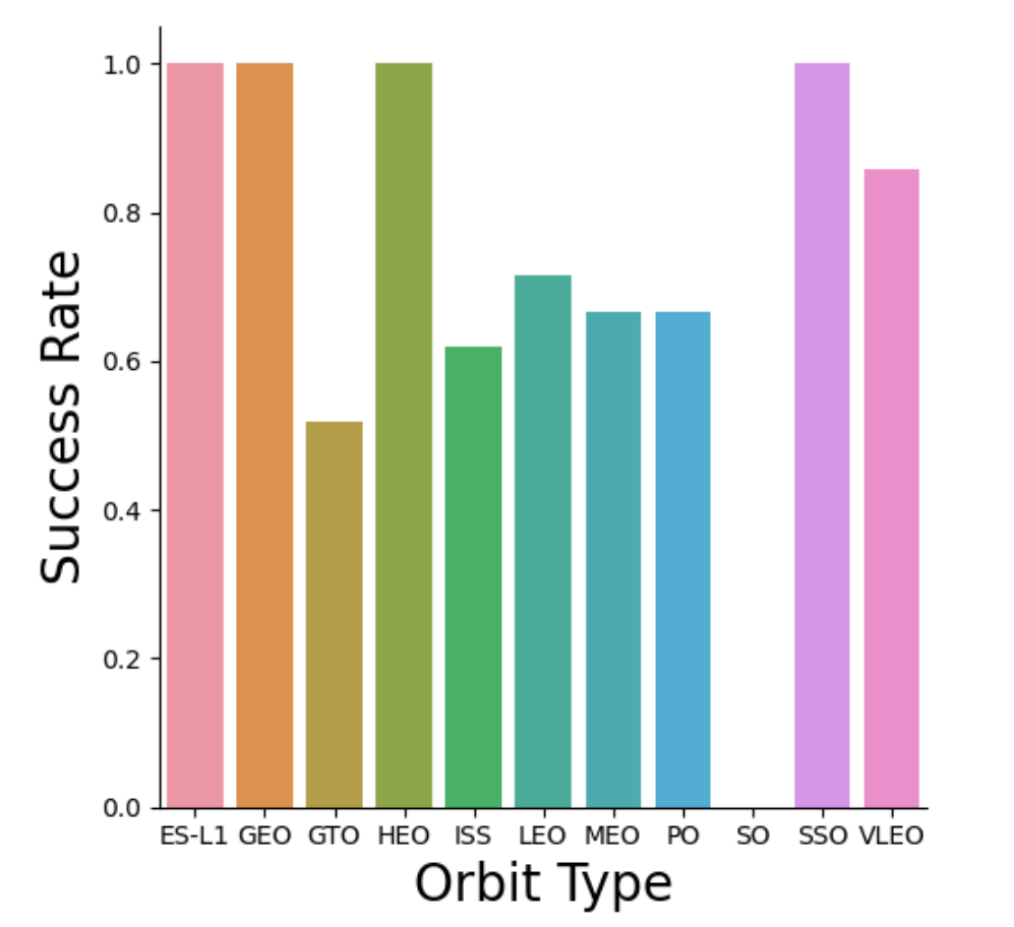
- The earliest launches (from the CCAFS SLC 40 launch site) were mostly unsuccessful.
- The latest launches (from the CCAFS SLC 40 and the KSC LC 39A sites) were successful.
- Only a few launches used the VAFB SLC 4E site.
- Most of the launches came from the CCAFS SLC 40 site.

Payload vs. Launch Site



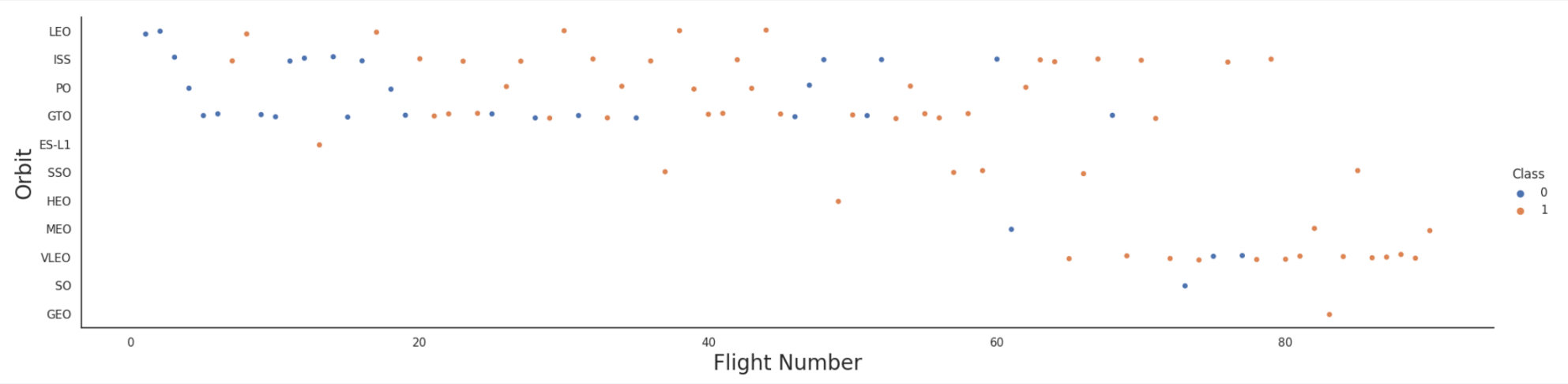
- The CCAFS SLC 40 and the KSC LC 39A launch site hosted the launches with high payload.
- Most launches with high payload (greater than 8000 kg) were successful.
- The launch site KSC LC 39A was mostly successful also for smaller payloads.

Success Rate vs. Orbit Type



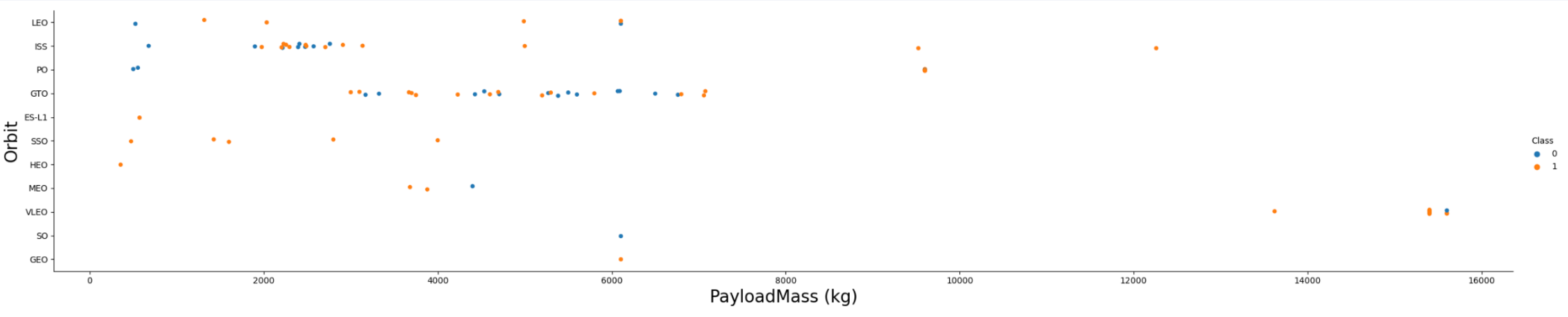
- The orbits ES-L1, GEO, HEO and SSO exhibit 100% success rate.
- The SO orbit has no success rate.
- However, other orbits (GTO, ISS, LEO, MEO, PO, VLEO) have a success rate greater than 50%.

Flight Number vs. Orbit Type



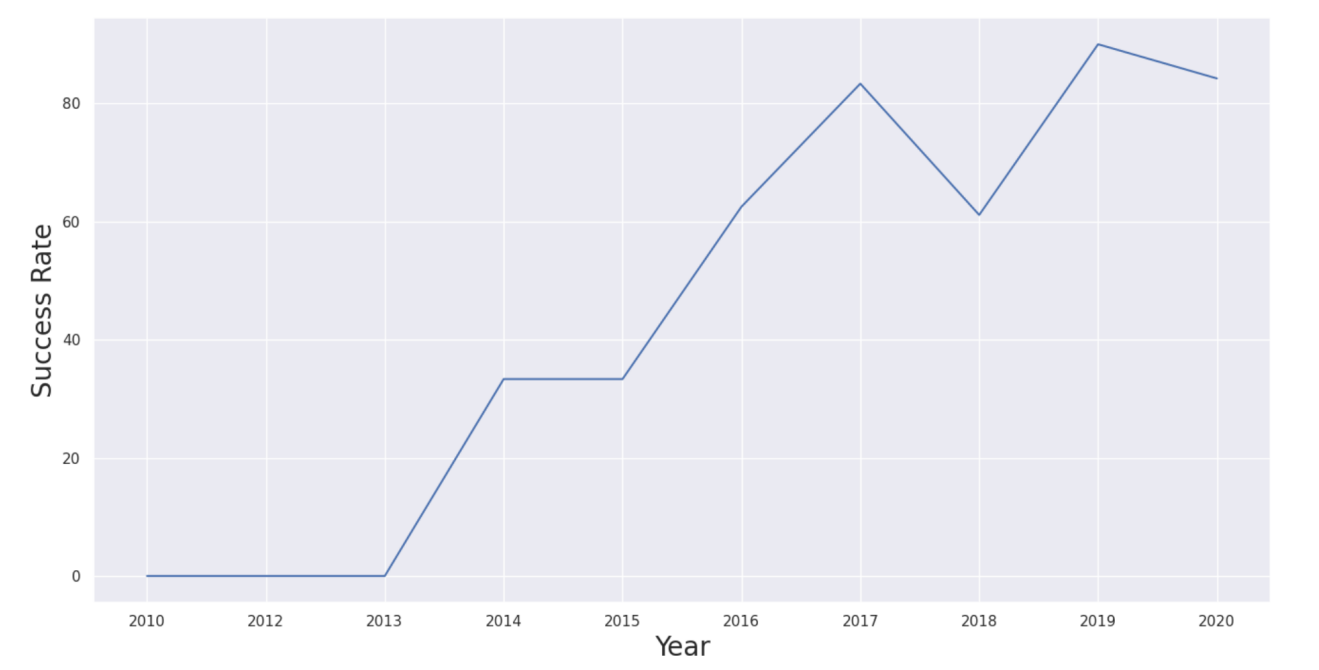
- This plot shows the trends observed in the previous slide in terms of the flight numbers.
- Only a single, unsuccessful flight was to do the SO orbit.
- The GEO, ES-L1, HEO are associated with single, successful flights

Payload vs. Orbit Type



- For GTO orbits, high payload typically lead to an unsuccessful landing.
- For other orbits – SO excluded – an high payload does not seem to influence the landing success.

Launch Success Yearly Trend



- The success rate of the first stage landing has greatly improved across the years.
- In particular, it rose sharply from 2016 onwards.
- The increasing in the success rate was only interrupted in 2018.

All Launch Site Names

❖ Displaying the names of the unique launch sites in the space mission:

```
%%sql

select distinct "Launch_Site"
from SPACEXTBL
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

❖ Displaying 5 records where launch sites begin with the string 'CCA':

```
%%sql

select *
from SPACEXTBL
where ("Launch_Site" LIKE "CCA%")
limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

❖ Displaying the total payload mass carried by boosters launched by NASA (CRS):

```
%%sql

select sum("Payload_Mass__KG_") as "Total NASA Payload Mass"
from SPACEXTBL
where ("Customer" LIKE "NASA (CRS)")
```

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

Total NASA Payload Mass
45596.0

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1:

```
%%sql

select round(avg("Payload_Mass__KG_"),1) as "Average Payload Mass Booster Version F9 v1.1"
from SPACEXTBL
where ("Booster_Version" LIKE "F9 v1.1%")
```

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

Average Payload Mass Booster Version F9 v1.1
2534.7

First Successful Ground Landing Date

❖ Listing the date when the first successful landing outcome in ground pad was achieved (format: yyyymmdd):

```
%%sql

select min(substr(Date, 7, 4)||substr(Date, 4, 2)||substr(Date, 1, 2)) as "First successful landing date in a ground pad"
from SPACEXTBL
where "Landing_Outcome" == "Success (ground pad)"
```

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

First successful landing date in a ground pad
20151222

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:

```
%%sql

select distinct "Booster_Version"
from SPACEXTBL
where ("Landing_Outcome" == "Success (drone ship)" and "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000)
```

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

❖ Listing the total number of successful and failure mission outcomes:

```
%%sql

select "Mission_Outcome", count(*)
from SPACEXTBL
group by "Mission_Outcome"
order by "Mission_Outcome"
```

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

Mission_Outcome	count(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass, using a subquery:

```
%%sql

select "Booster_Version"
from SPACEXTBL
where PAYLOAD_MASS__KG_ == (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

- Listing the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015:

```
%%sql

select substr(Date, 4, 2) as Month, "Landing_Outcome", "Booster_version", "Launch_Site"
from SPACEXTBL
where "Landing_Outcome" == "Failure (drone ship)" and substr(Date,7,4) =='2015'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	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

- Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order:

```
%%sql

select "Landing_Outcome", count(*) as "count_succ"
from SPACEXTBL
where (substr(Date, 7, 2)||substr(Date, 4, 2)||substr(Date, 1, 2)
      between '20100406' and '20170320')  and "Landing_Outcome" like "Success%"
group by "Landing_Outcome"
order by "count_succ" desc
```

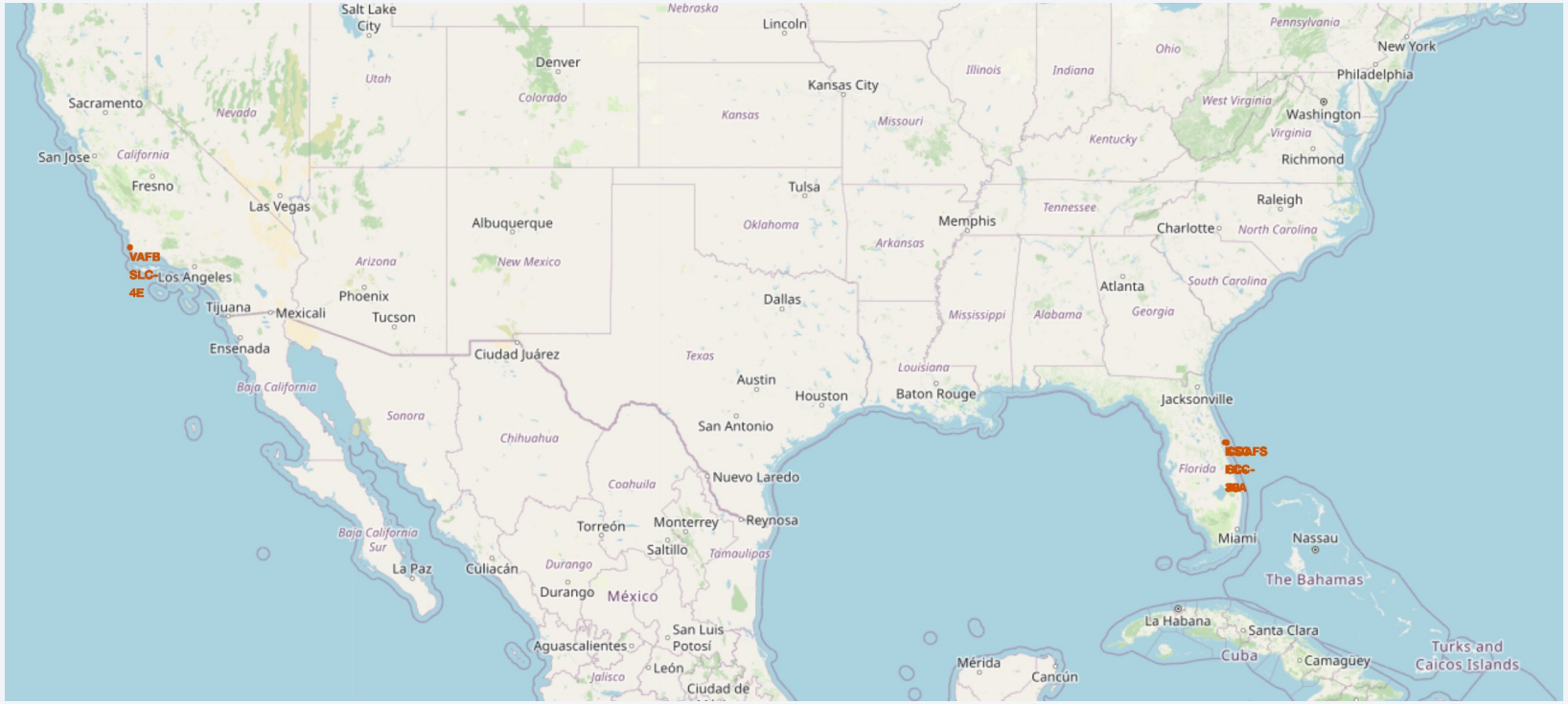
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_succ
Success	11
Success (drone ship)	3
Success (ground pad)	2


Section 3
Launch Sites
Proximities Analysis

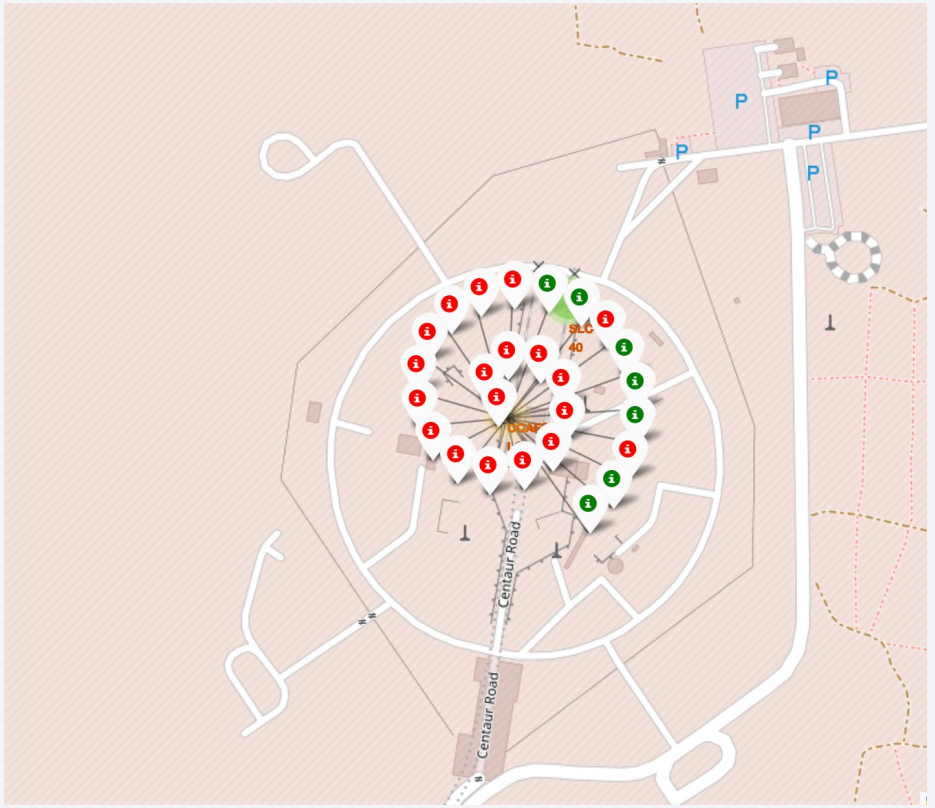
Launch sites' locations

- Below are highlighted the positions of Falcon 9 flights' launch sites.
- The sites are close to the equator line, where the escape velocity may be reached more easily.



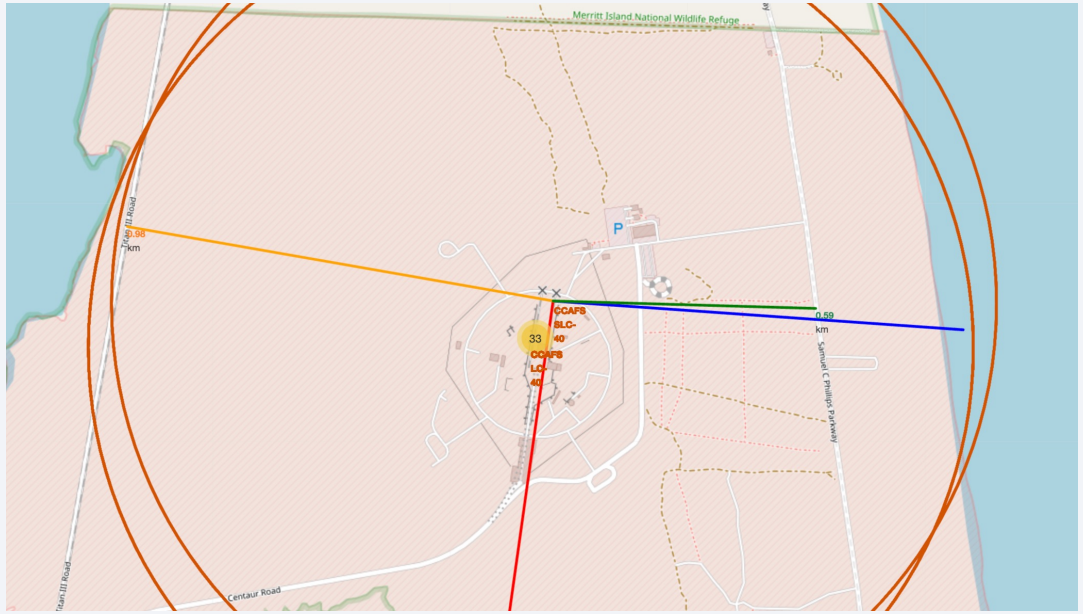
Landing success from given launch site

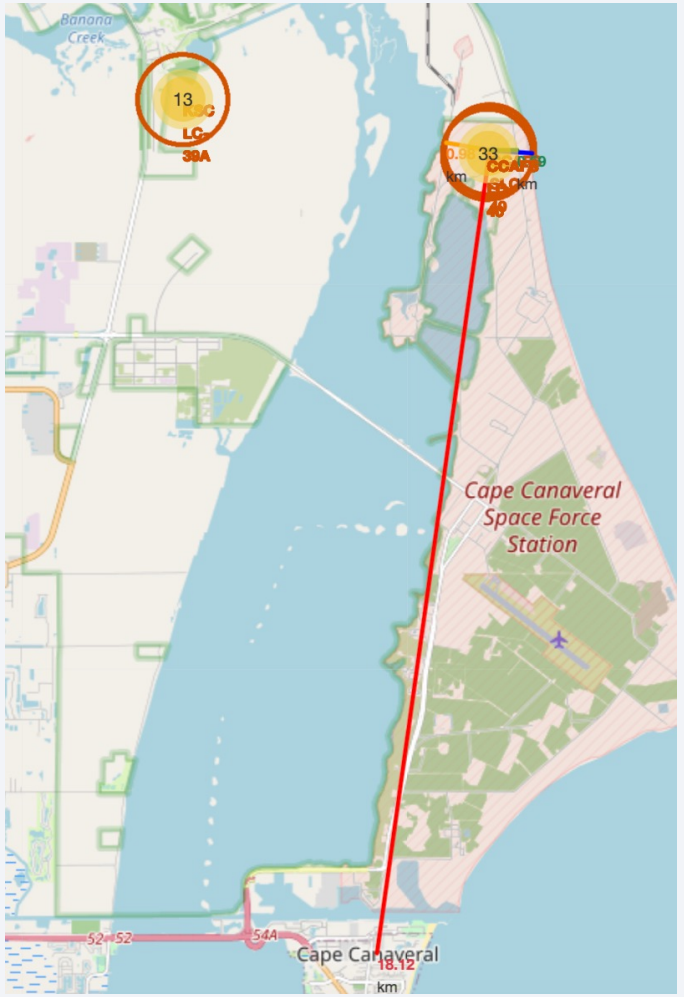
- We introduced a mark for each launch. Clearly, they cluster around each launch site.
- Each mark is colored in green if the first stage landing was successful, and red otherwise.
- Beside is an example for the launch site CCAFS LC-40.

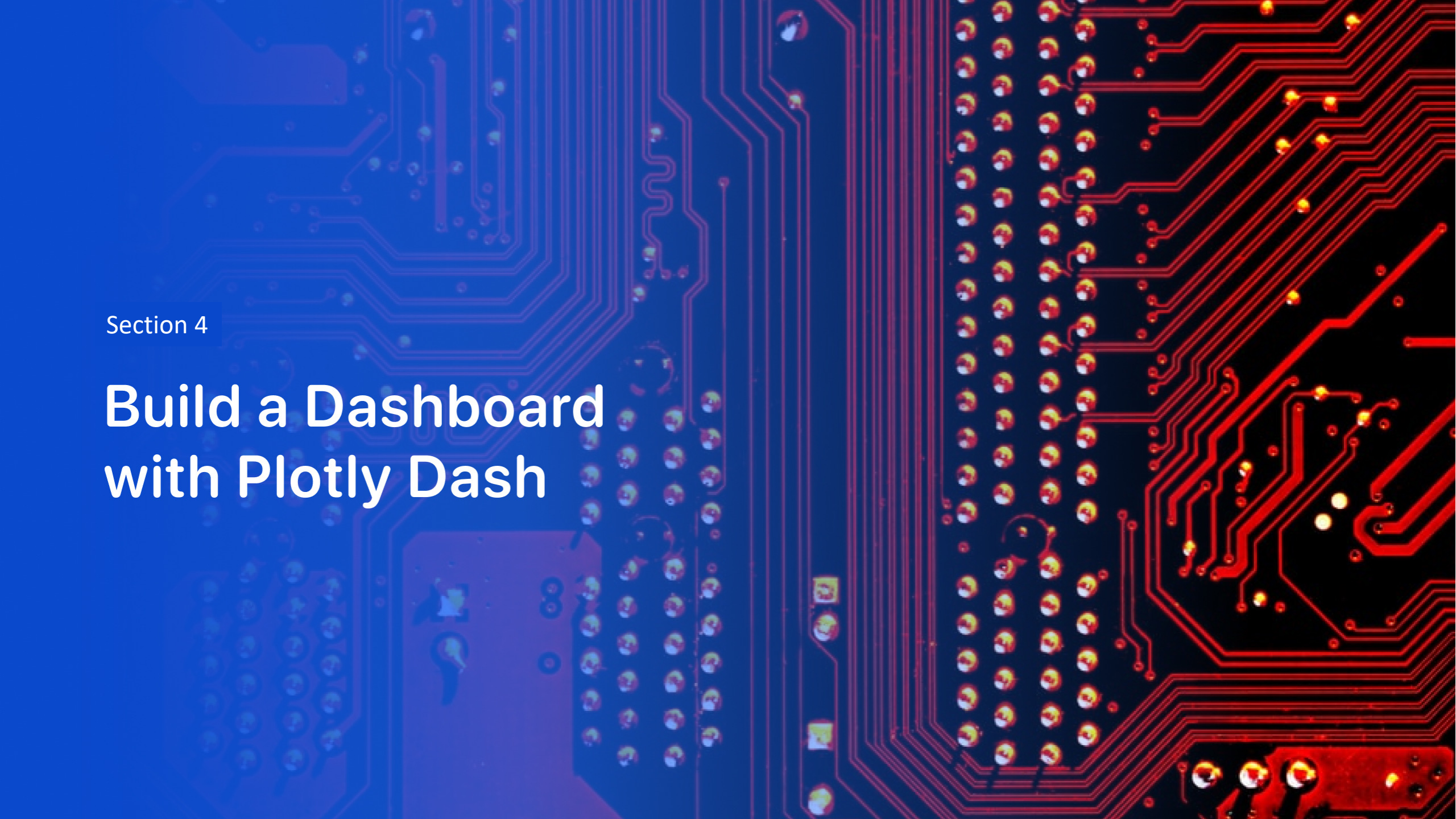


Distances of the launch site CCAFS LC-40 from proximities

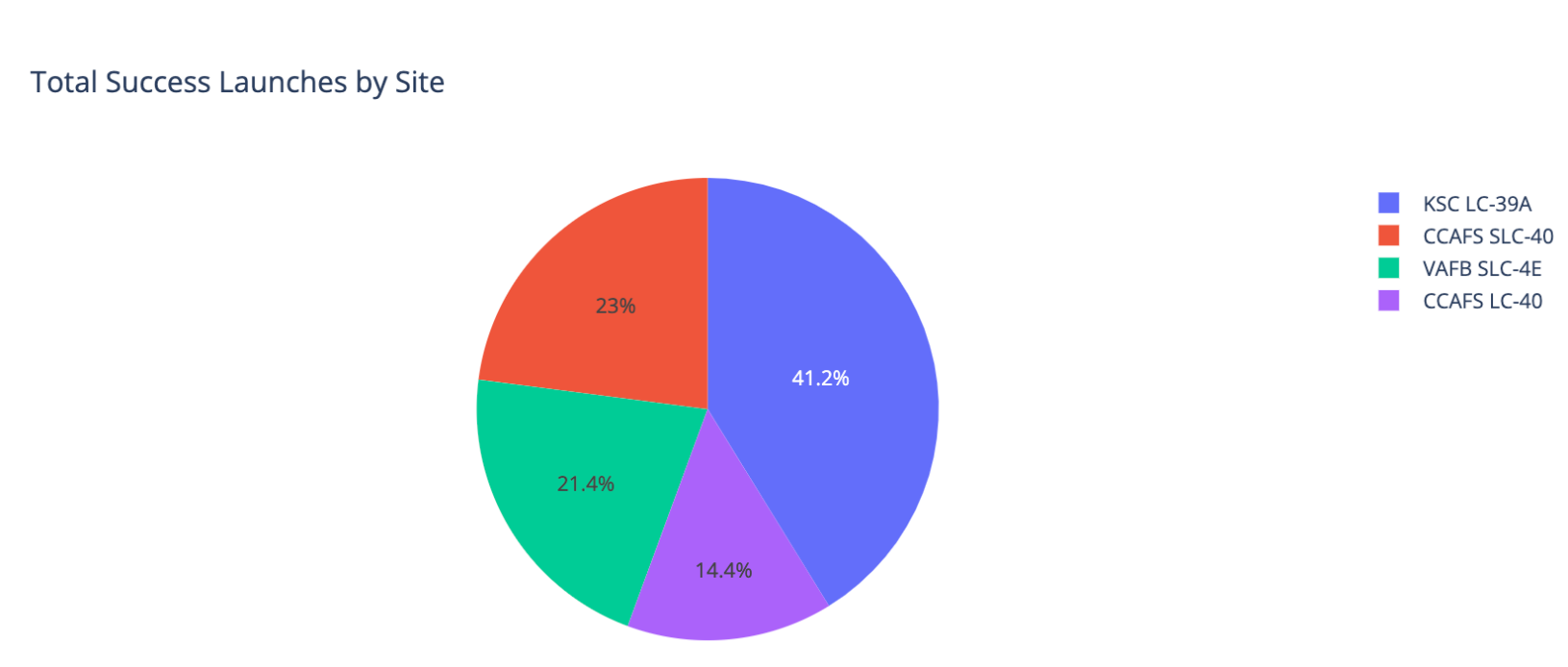
Launch sites are very close to railways (orange line), highways (green), coastline (blue) – see picture below – but far from city centers (red) – see figure on the right.





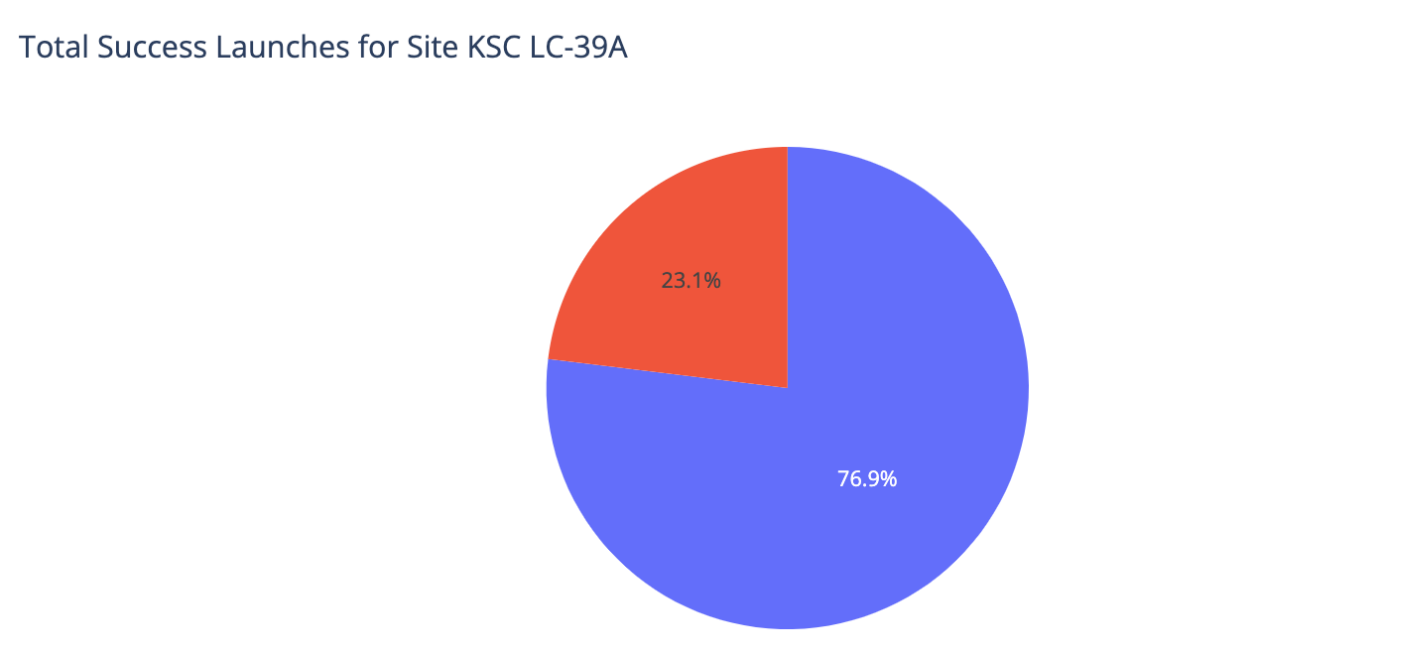

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



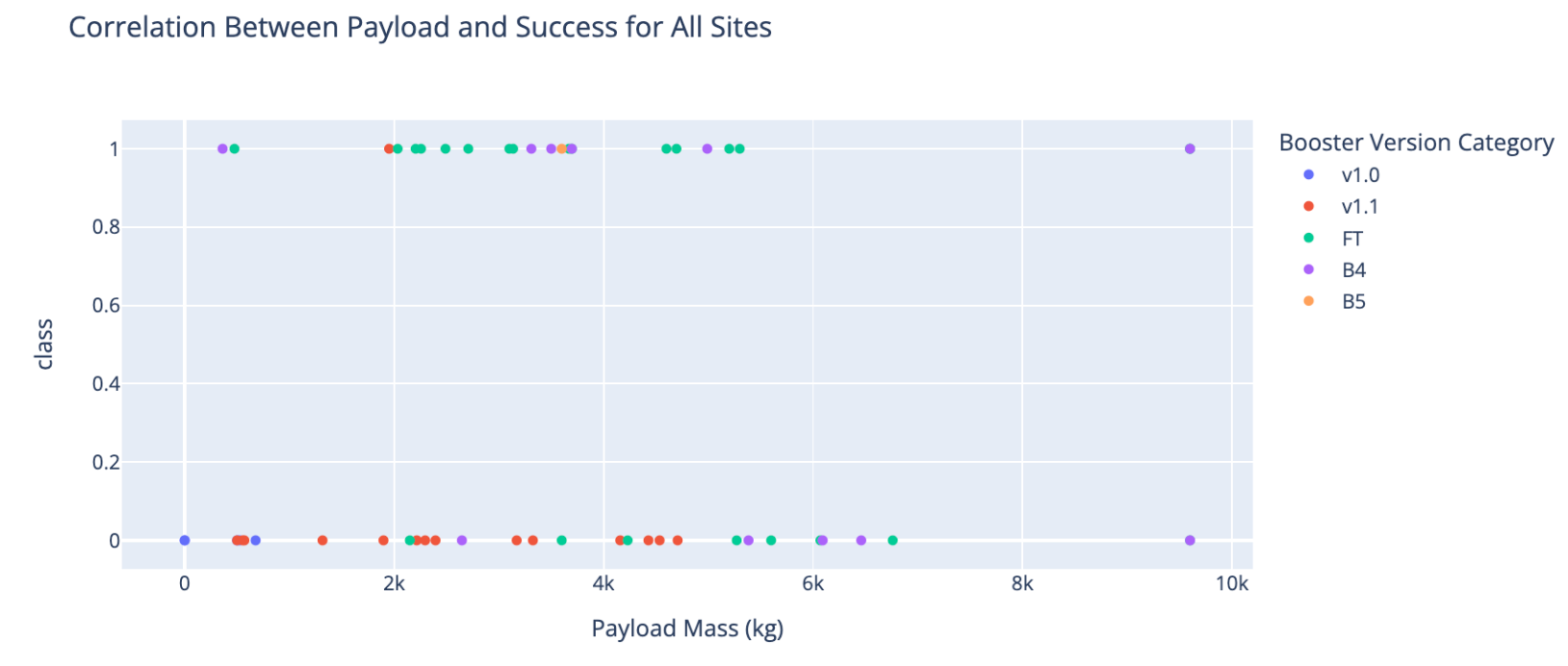
- Most of the successful launches were performed from the KSC LC-39A launch site.

Piechart for the launch site with highest launch success ratio



- The most successful launch site, KSC LC-39A, has a high success rate (76.9%)

Payload vs. Launch Outcome scatter plot for all sites



- For small payloads, there are more successful landings.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

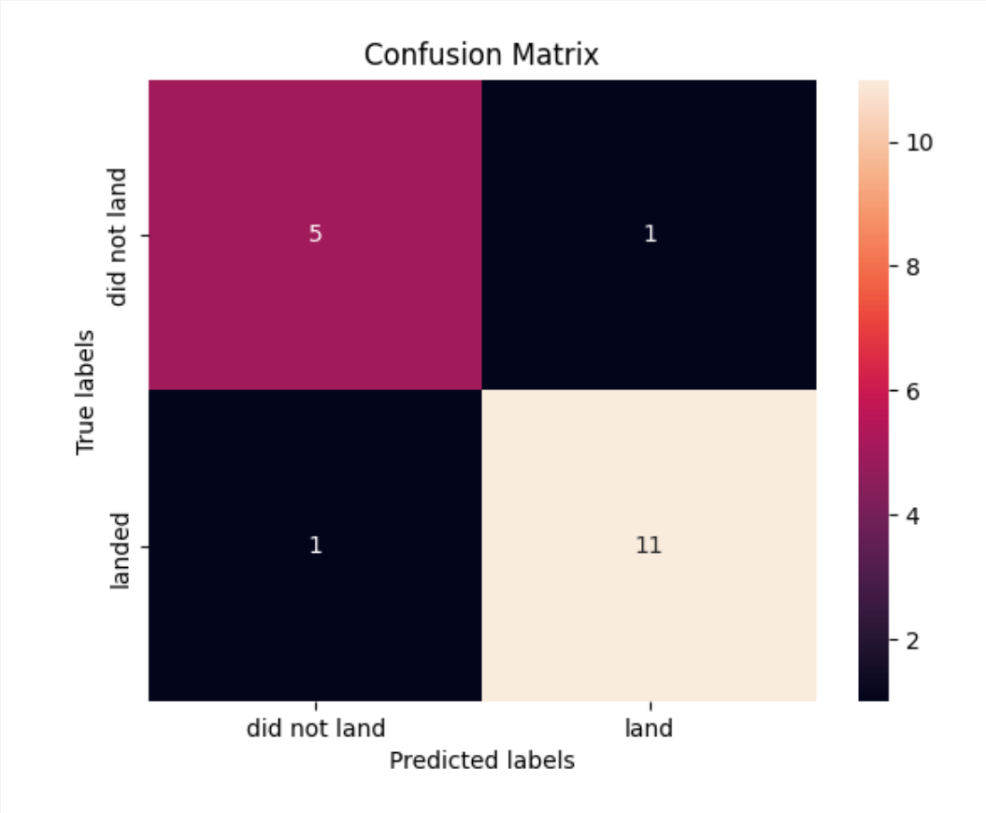
- We tested different models: LogReg, SVM, Decision Tree, and KNN models
- Their performance is encoded in the following metrics (measured on the test set):

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

- The Decision Tree model is the one that performs the best.

Confusion Matrix

- The best performing model is the decision tree classifier, with an accuracy of 0.89
- Only two launches are not classified properly.



Conclusions

- Location of sites: launch sites are close to the Equator, distant from cities but close to relevant highways and railways
- Success over the years: the success rate of the launches is improving greatly over the last years
- Payload relevancy: launches with lighter payload seem to perform the best
- Orbit relevancy: some orbits have 100% success rate, although only few launches have been performed to reach these orbits
- Modeling: the model with the best performance at predicting the launch success seems to be the decision tree classifier.

Appendix

The Jupyter notebooks containing the codes employed for the analysis are all available at the following GitHub page of the author:

GitHub Link

Thank you!