IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Igor Osmolovskii
07.07.2023

Outline

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

Executive Summary

- Summary of methodologies

The open sourced data was used during the data collection: SpaceX API (called via Python requests) and Wikipedia (webscrapping).
The technics of data collection, exploratory analysis, visualization, locating on a map, dashboard and machine learning were applied.

- Summary of all results

We get the understanding of which launch sites and launch parameters are more likely to have the successful outcome which is presented via the maps, dashboards and charts. Besides we're able to make the prediction of the successfulness of a launch.

Introduction

- Project background and context

As a private space company we are analyzing the SpaseX launches data to learn their efficiency and make predictions on the success outcome of a launch. This allows us to avoid already known mistakes and achieve our company's goals efficiently.

- Problems you want to find answers

1. What are the essential parameters for a successful launch?
2. What are the optimal parameters?
3. What is the optimal launch site location? Where a new location should be built?
4. What is the optimal payload for a chosen orbit?
5. What are the most efficient booster parameters?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Official SpaceX API
  - Wikipedia pages
- Perform data wrangling
  - Data cleaning to remove the non-consistent values.
  - Adding new features needed for analysis
  - Transformation and encoding for machine learning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

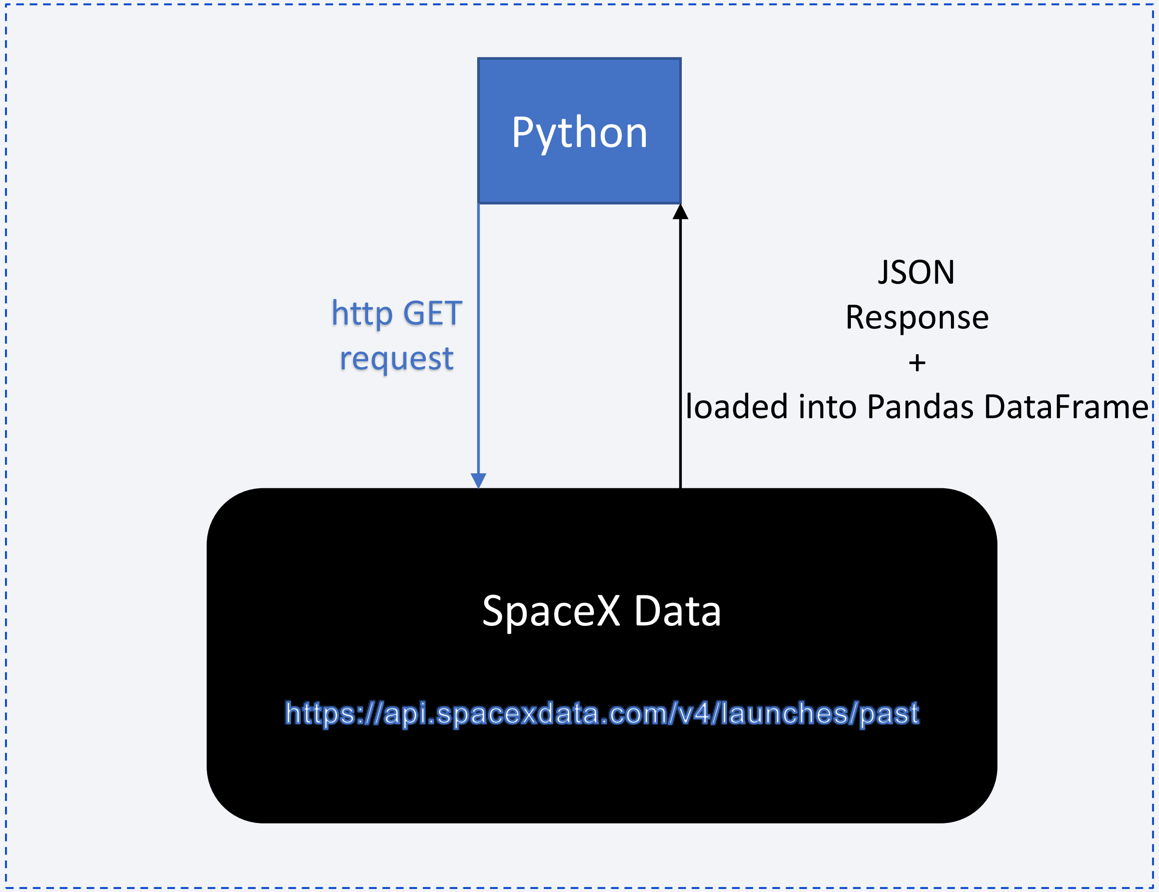
Data Collection

The data was collected with the next approaches:

- Official SpaceX API (called via Python requests library)
- Wikipedia pages (webscrapping performed with Python BeautifulSoup library)

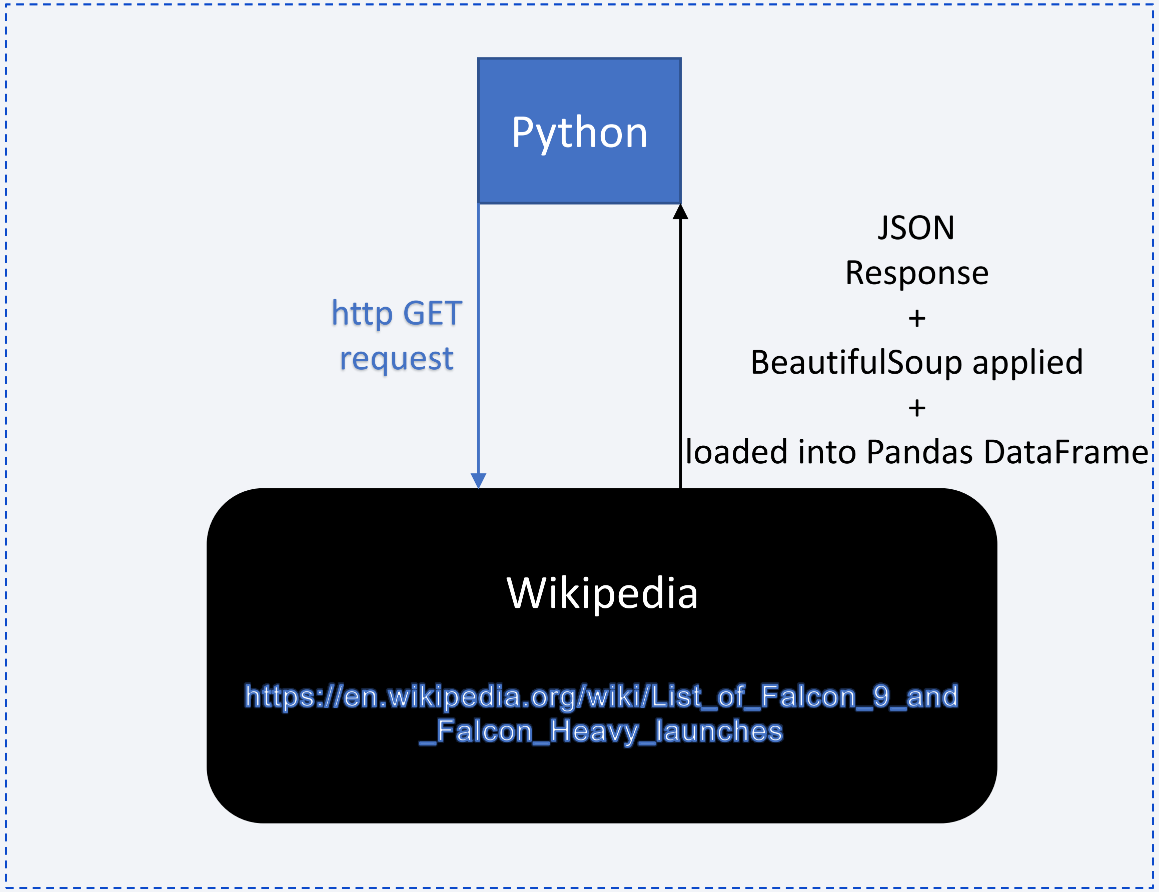
Data Collection – SpaceX API

- Python requests library for API calls
- URL = "https://api.spacexdata.com/v4/launches/past"
- The GitHub URL of the completed SpaceX API calls notebook


Python
http GET request
JSON Response + loaded into Pandas DataFrame
SpaceX Data
https://api.spacexdata.com/v4/launches/past

Data Collection - Scraping

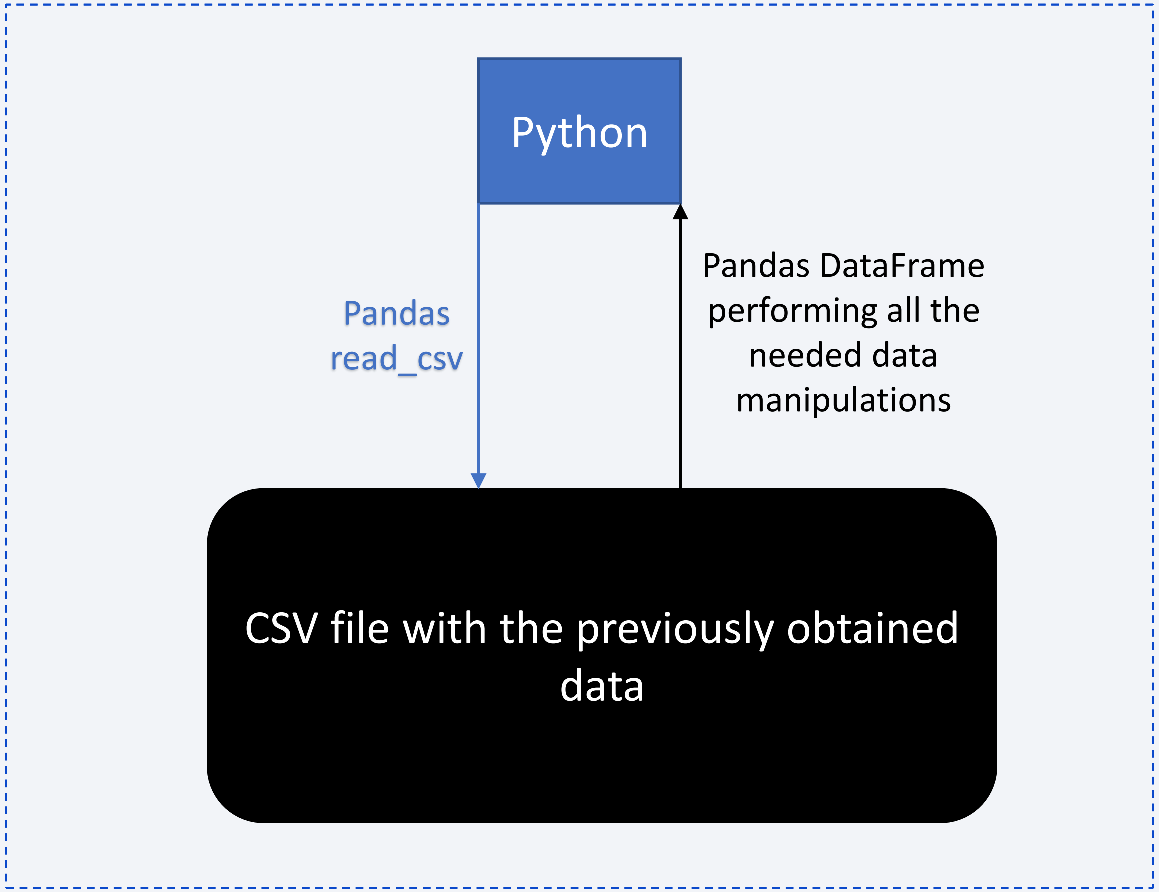
- Python BeautifulSoup library scraps a webpage
- URL = https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- [The GitHub URL of the completed web scraping notebook]()


Python
http GET request
JSON Response + BeautifulSoup applied + loaded into Pandas DataFrame
Wikipedia
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Wrangling

- How data were processed
  - Calculation of the number of launches on each site
  - Calculation of the number and occurrence of each orbit
  - Calculation of the number and occurence of mission outcome per orbit type
  - Creation of a landing outcome label from Outcome column

- [The GitHub URL of the completed data wrangling related notebook]()


Python
Pandas read_csv
Pandas DataFrame performing all the needed data manipulations
CSV file with the previously obtained data

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Scatterplot Flight Number VS Payload Mass with a success color marker to see if the efficiency of carrying the loads was increasing along he timeline.
  - Scatterplot Flight Number VS Launch Site with a success color marker to see if the efficiency of launches was increasing along he timeline
  - Scatterplot Payload Mass VS Launch Site with a success color marker to see which launch sites performed better with different loading
  - Bar chart Orbit VS Success to find out if there is a dependency on which orbit we're launching
  - Scatter plot Flight Number VS Orbit with a success color marker to see if there was a progress along the time in sending a rocket to a specific orbit
  - Scatter plot Payload VS Orbit with a success color marker to see if there was a progress along the time in sending a rocket with specific payload to a specific orbit
  - Launch success yearly trend to see the overall progress in launchings
- The GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

- The SQL queries performed:
  - the names of the unique launch sites
  - display 5 records where launch sites begin with the string 'CCA‘
  - the total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - the date when the first succesful landing outcome in ground pad was achieved
  - the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - the names of the booster_versions which have carried the maximum payload mass
  - the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - 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.
- The GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
    - All launch sites were added to check the geography of those
    - All launches with the success color marker were added to investigated the launch site ranks
    - The distance to a nearest town was added to validate how far they can be located

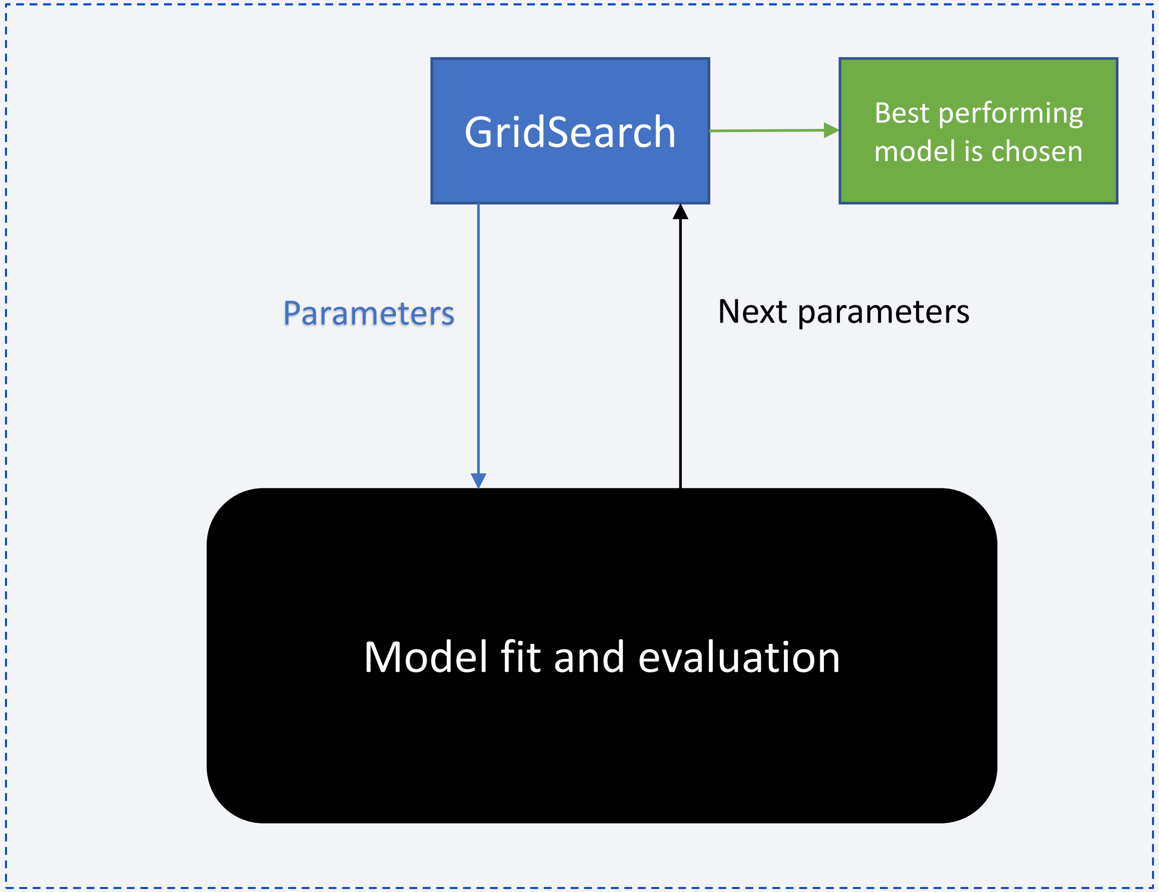
- The GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Successfulness launch rate per launch site
  - Successfulness of a launch per Booster and payload
- Explain why you added those plots and interactions
  - We want the statistics of a successful outcome per site to better predict for next launches
  - We want the statistics how payload impacts a launch per site per booster to understand their impact on the outcome
- The GitHub URL of the completed Plotly Dash lab

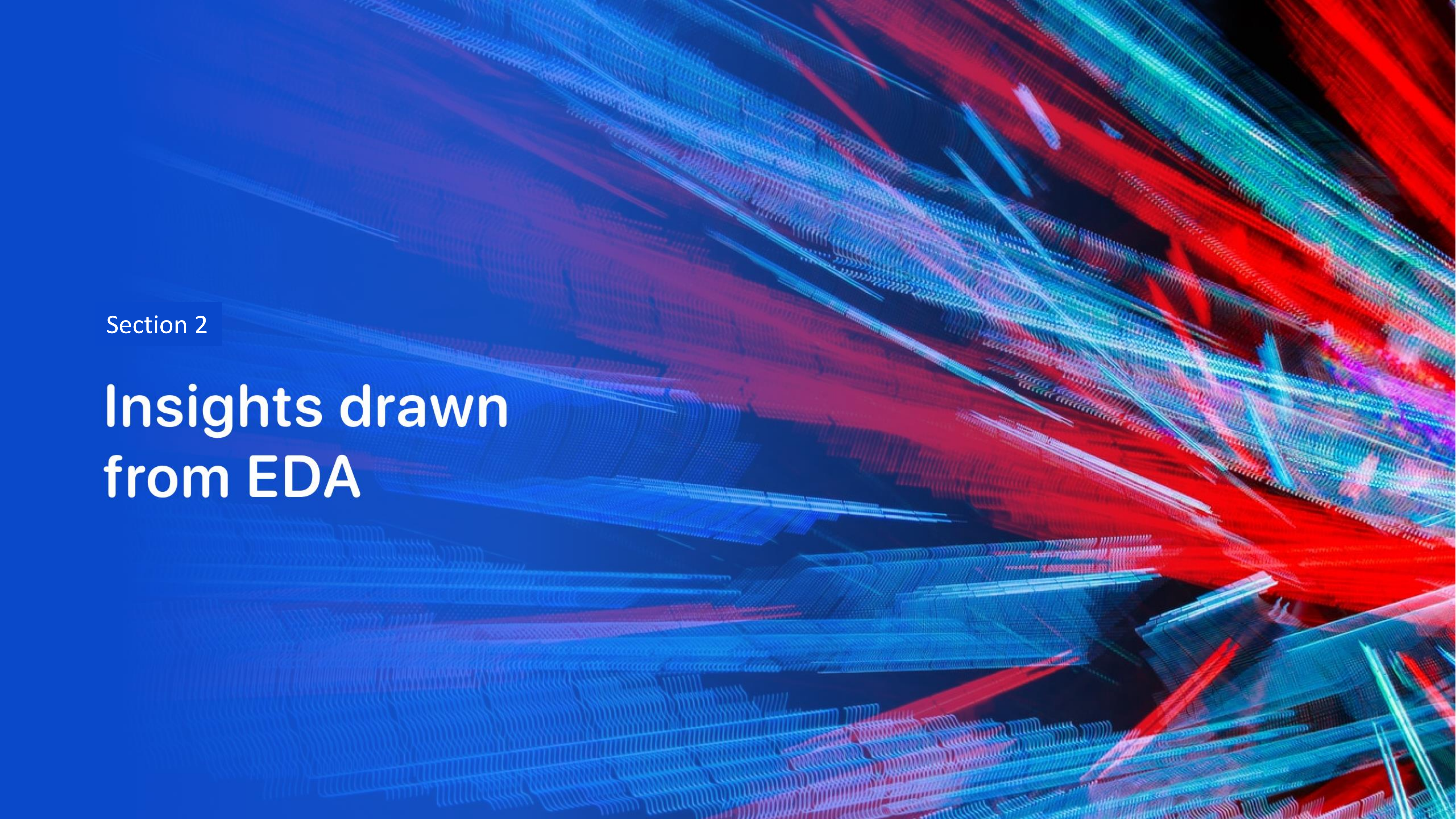
Predictive Analysis (Classification)

- Summary on how the best performing classification model built, evaluated, improved, and found.
  - Loaded the data into a DataFrame.
  - Created a result variable array from a Class (success) column.
  - Standrardized the data and splitted it into the train and test sets.
  - Applied GridSearch for a number of parameters to different algorithms.
  - Evaluated model performances with their best parameters checking the accuracy and confusion matrix.
- The GitHub URL of your completed predictive analysis lab



Results

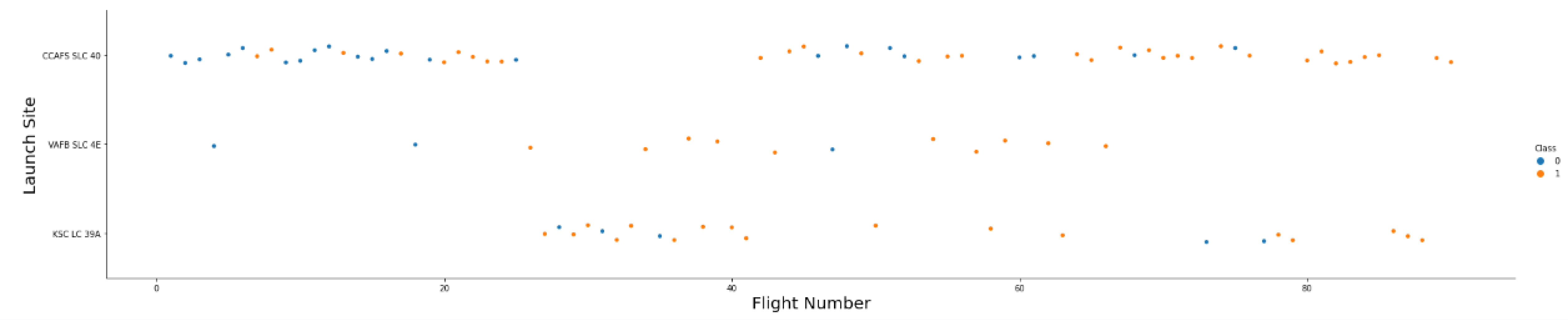
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

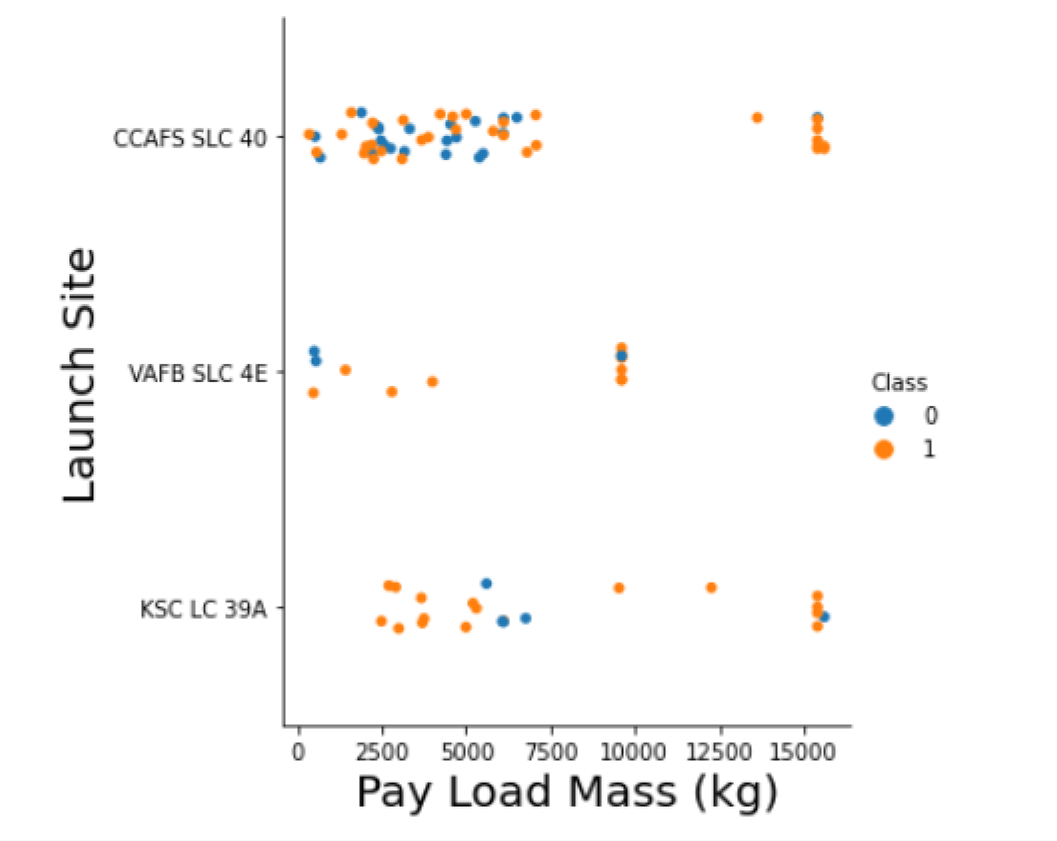
Insights drawn from EDA

Flight Number vs. Launch Site



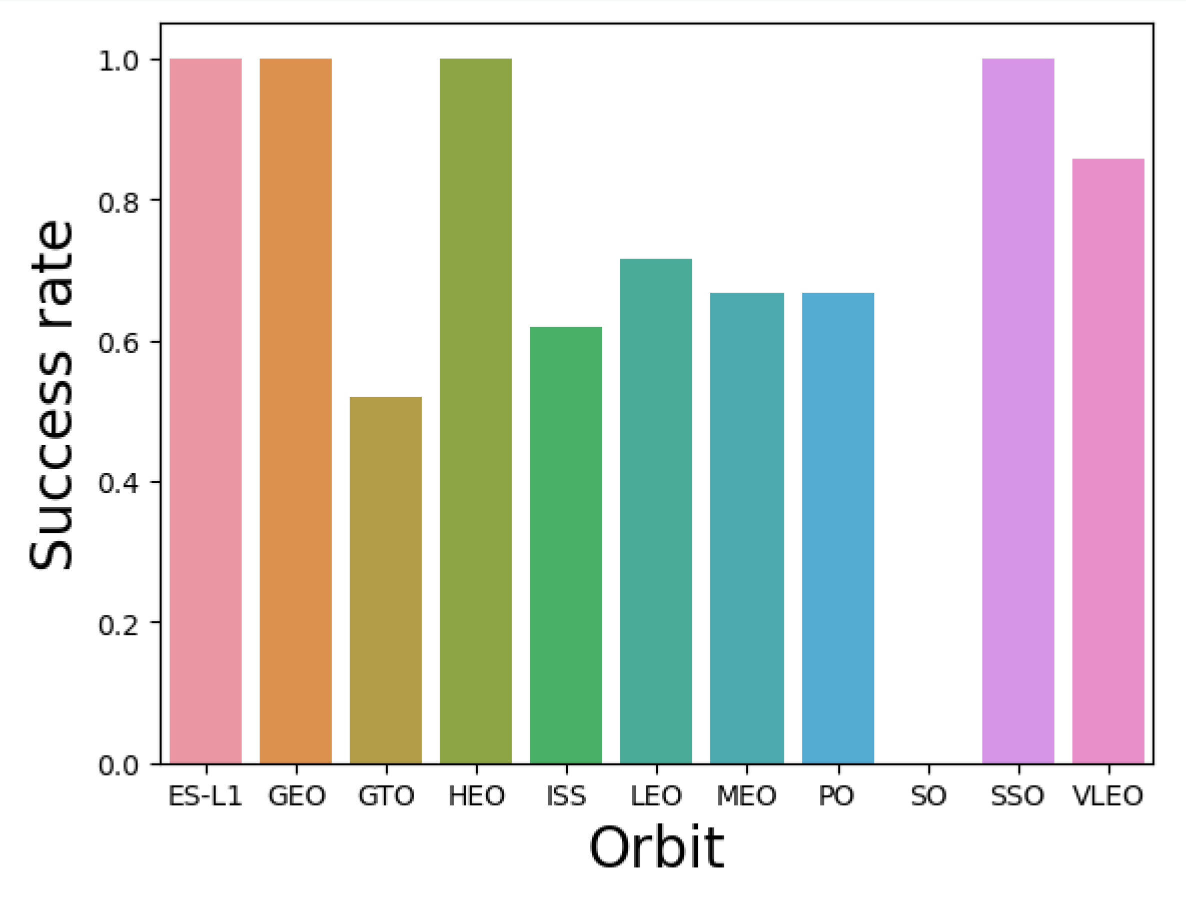
We see that successful launch number is increasing with the flight number in all launch sites

Payload vs. Launch Site



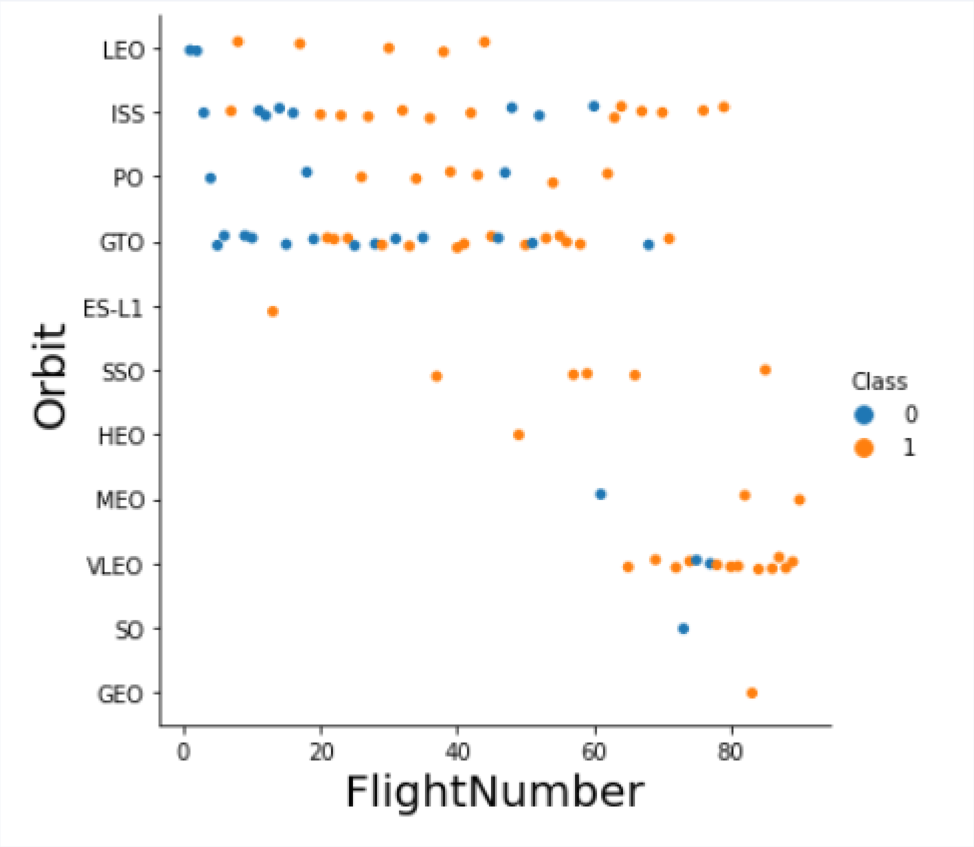
CCAFS SLC-40 and KSCLC 39A have more successful launches with the higher payload whereas for VAFB SLC 4E there is almost no such dependency.

Success Rate vs. Orbit Type



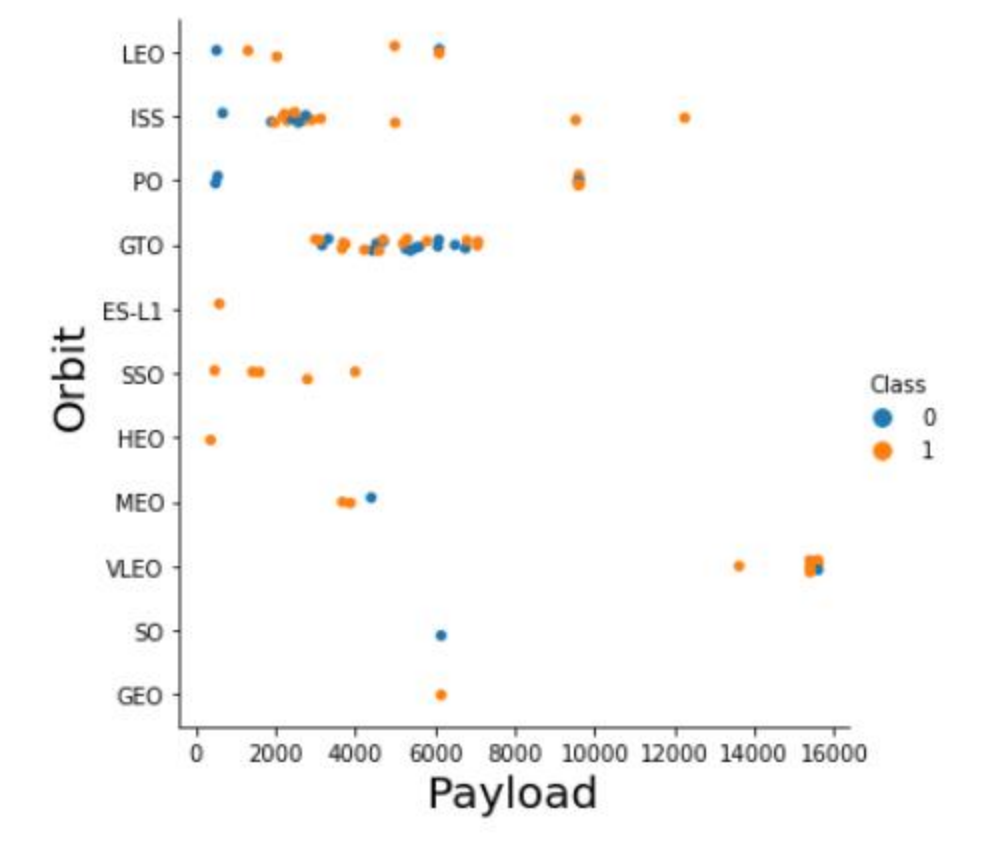
ES-L1, GEO, HEO, SSO orbits have the highest success rate.

Flight Number vs. Orbit Type



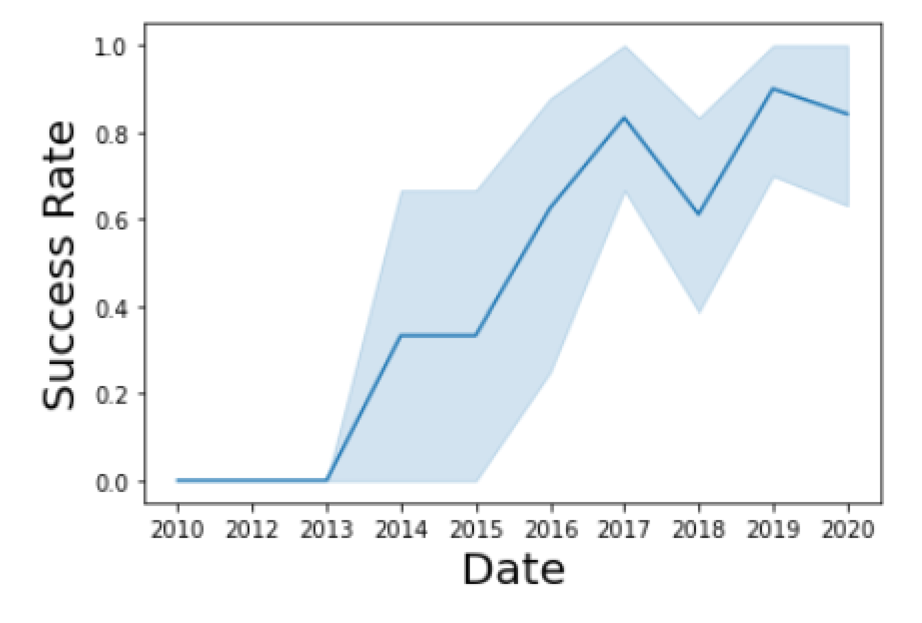
In all the orbits the success rank increases with a flight number

Payload vs. Orbit Type



Higher payload works better for ISS. SSO is just perfect. Other orbits successes are not stable in regards to the payload.

Launch Success Yearly Trend



Yearly trend is very positive starting from 2013

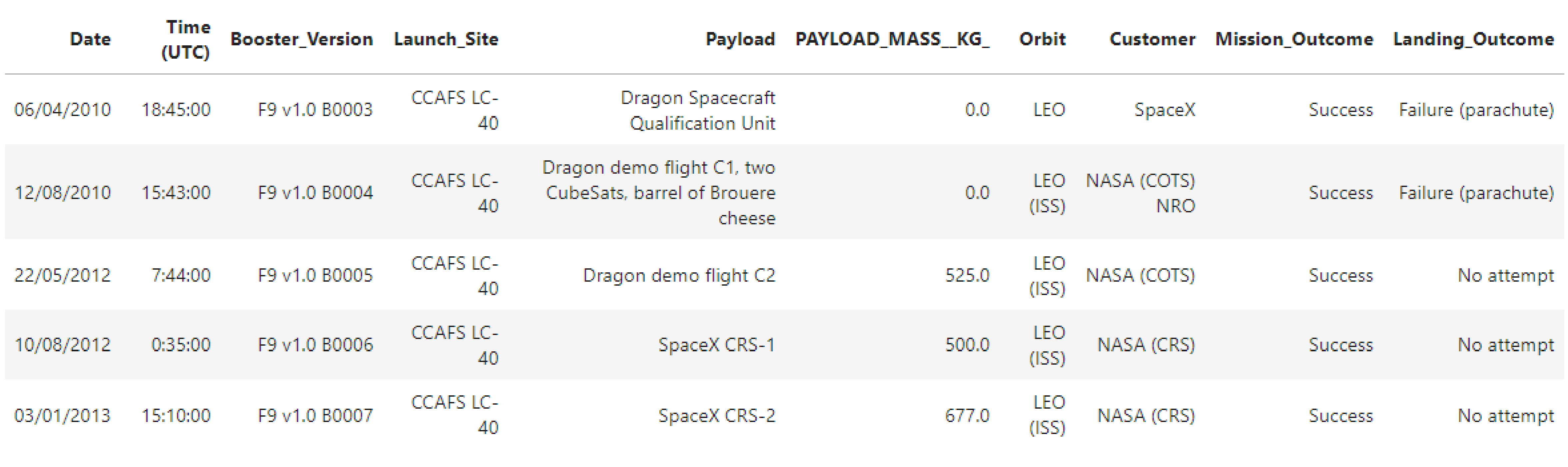
All Launch Site Names

The next sites are presented in the data. Some data lines are missing the values.

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

Launch Site Names Begin with 'CCA'

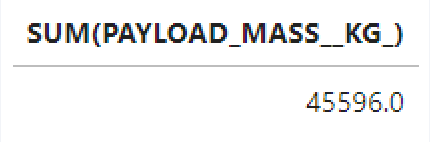
Here are 5 records showing the launch data from the CCAFS LC-40 site



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

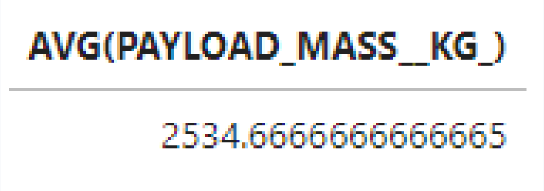
Total NASA launches payload throughout 2013-2020 is 45596 kg



SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

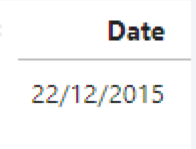
The average F9 v1.1 payload mass is 2534.67 kg



AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

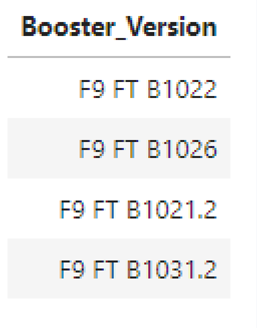
The first successful Ground Landing Date is 22 December 2015



Date
22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

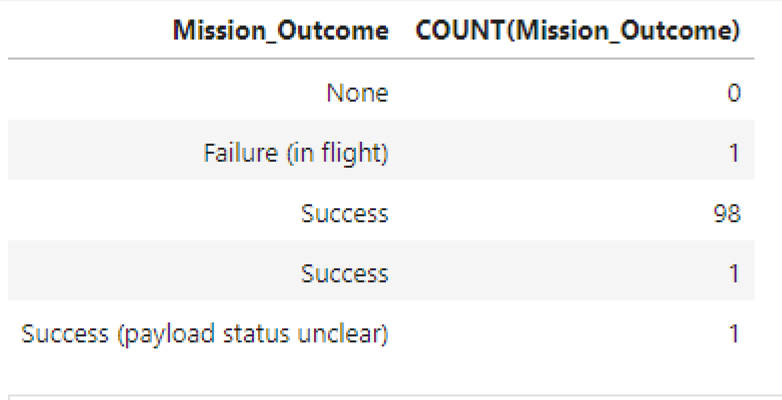
The next Boosters successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



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

Total Number of Successful and Failure Mission Outcomes

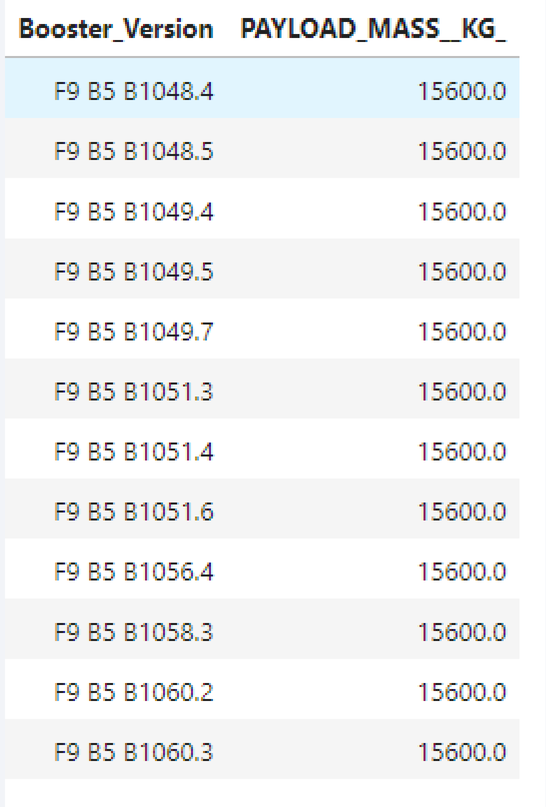
- There were 100 successful launches and 1 failure



Mission_Outcome	COUNT(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

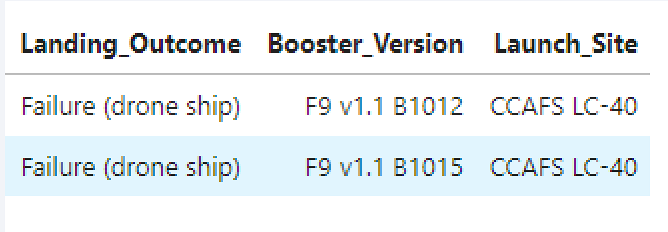
- The next boosters carried the maximum payload mass of 15600 kg



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1049.7	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1060.3	15600.0

2015 Launch Records

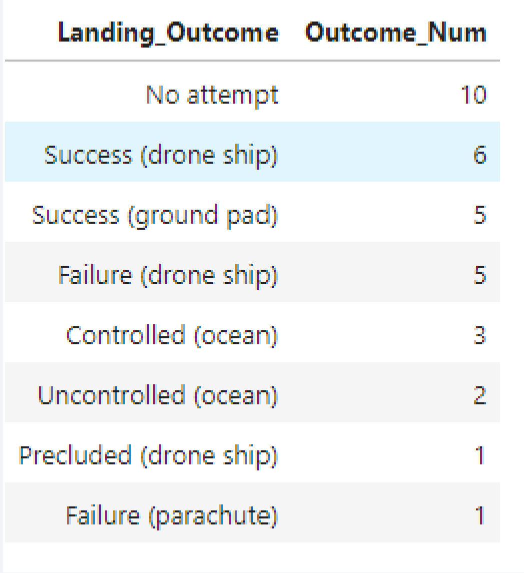
- List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

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

- The landing outcome ranks between 2010-06-04 and 2017-03-20, in descending order

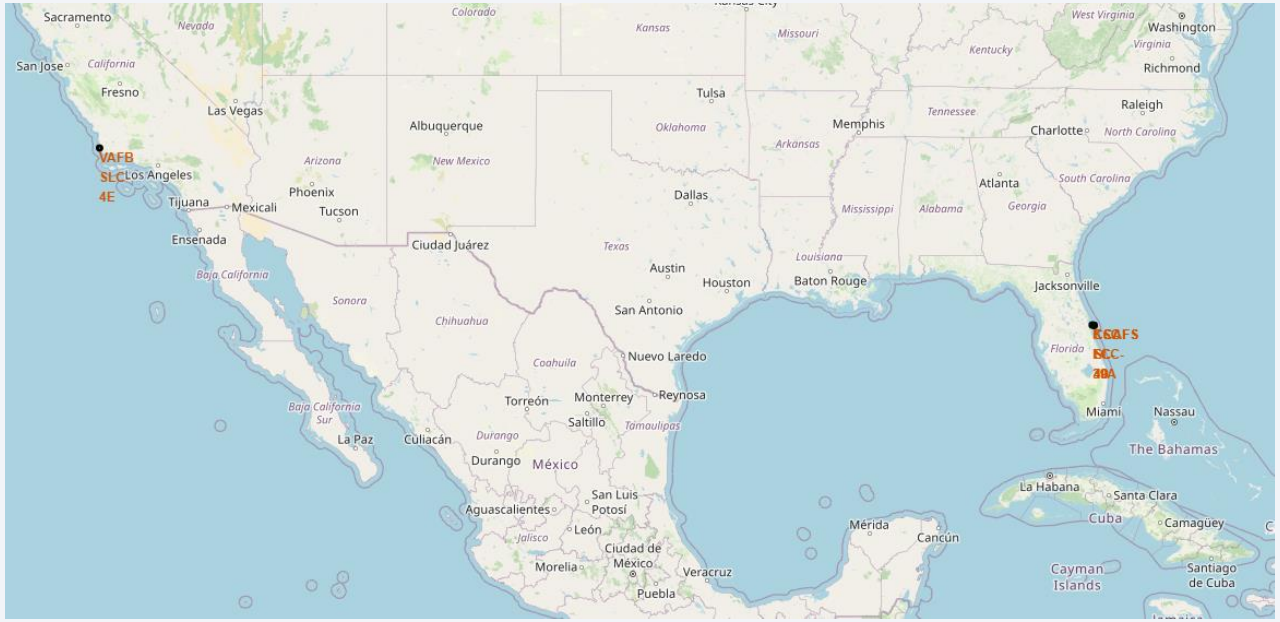


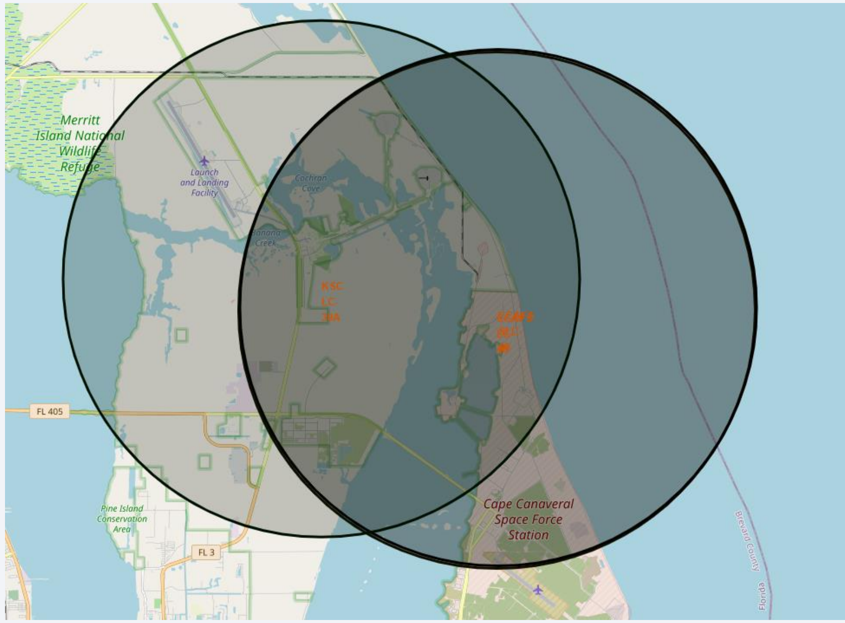
Landing_Outcome	Outcome_Num
No attempt	10
Success (drone ship)	6
Success (ground pad)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch site locations

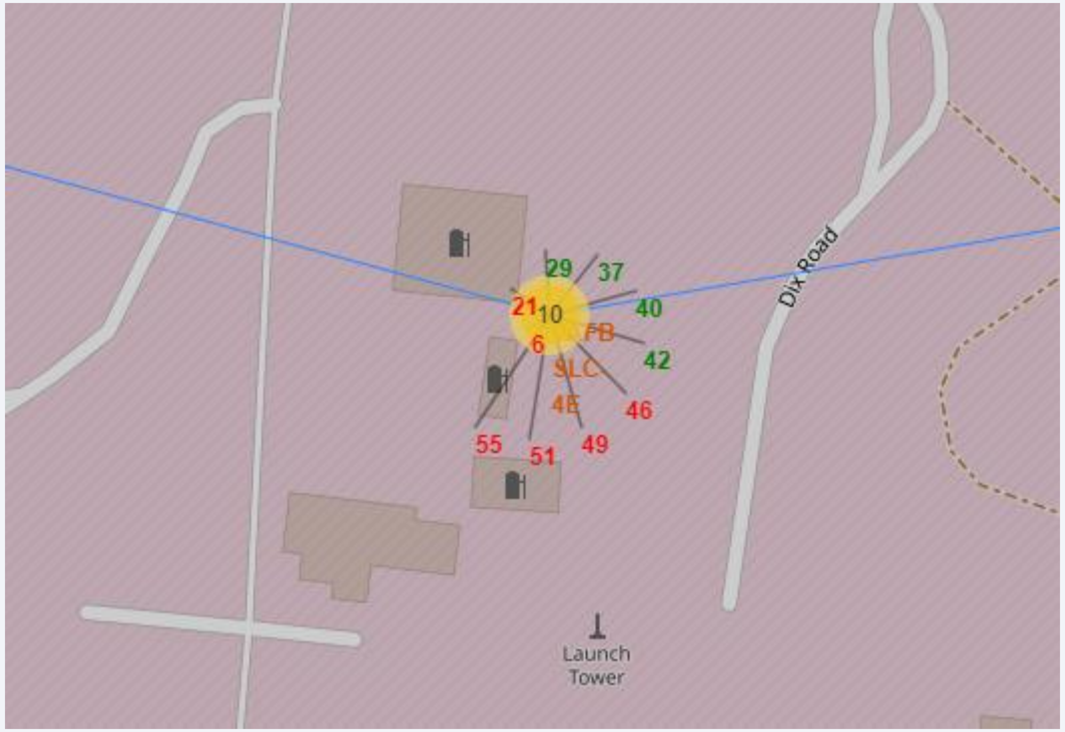



Merritt Island National Wildlife Refuge
Launch and Landing Facility
FL 405
FL 3
Pine Island Conservation Area
Cape Canaveral Space Force Station

There are 4 launch sites, all are in US: 3 in Florida and 1 in California.

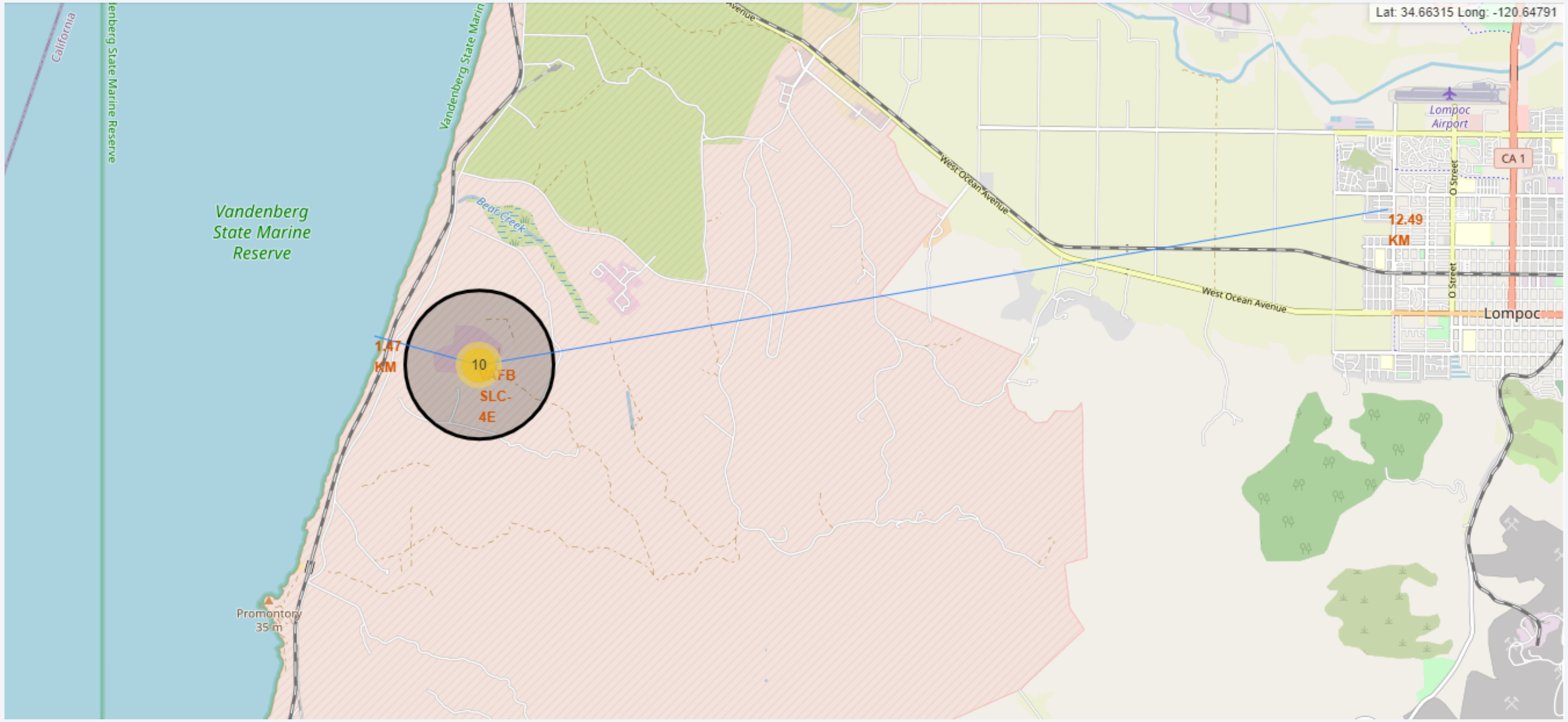
3 in Florida are really close to each other and can be better distinguished with zoom in in the right picture.

Launch outcomes for the VAFB SLC-4E site

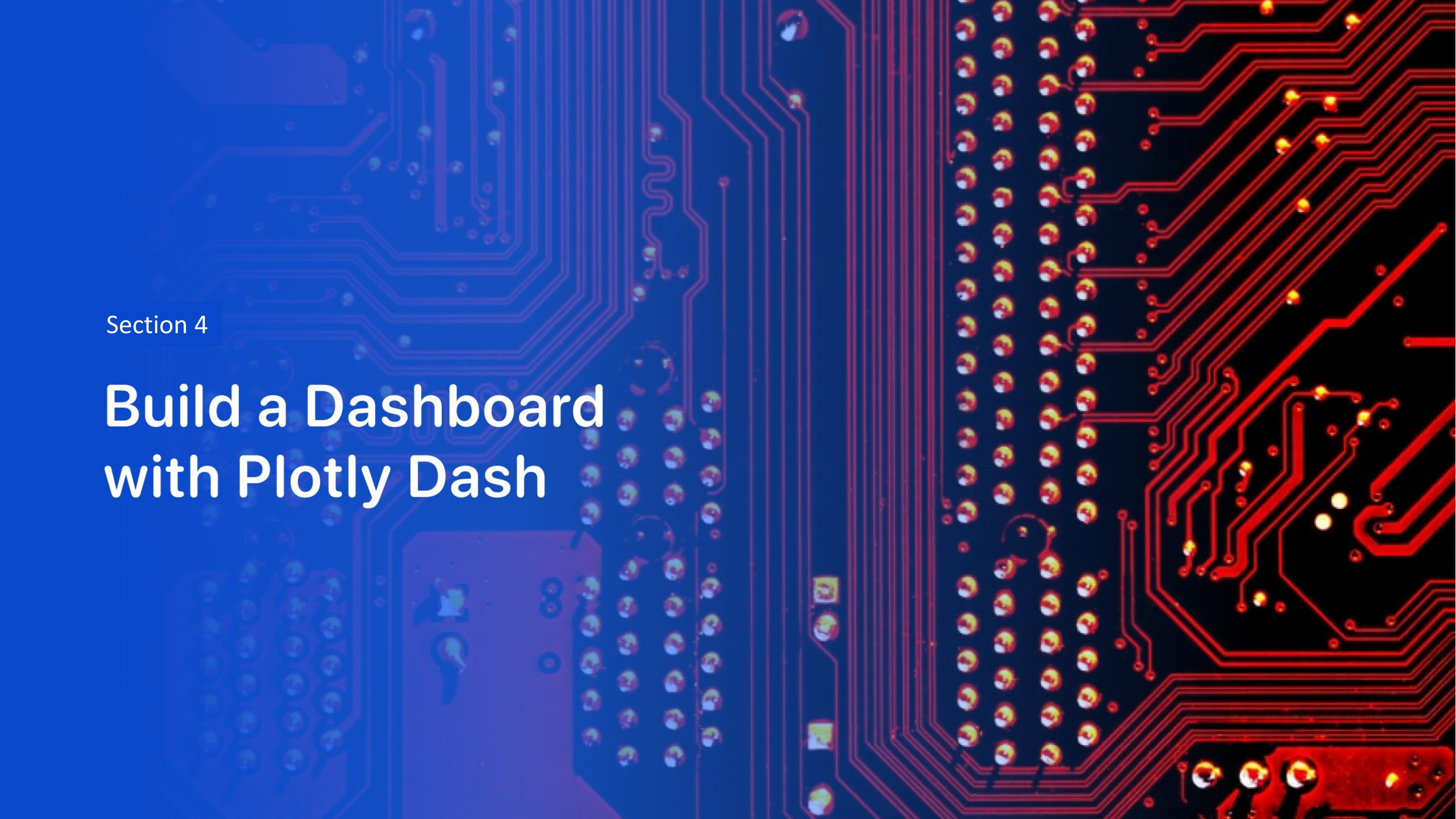


Green = successful, Red = failure. Numbers represent a flight number. We can see that even later flights were not successful and there is no successfulness trend.

VAFB SLC-4E site proximities



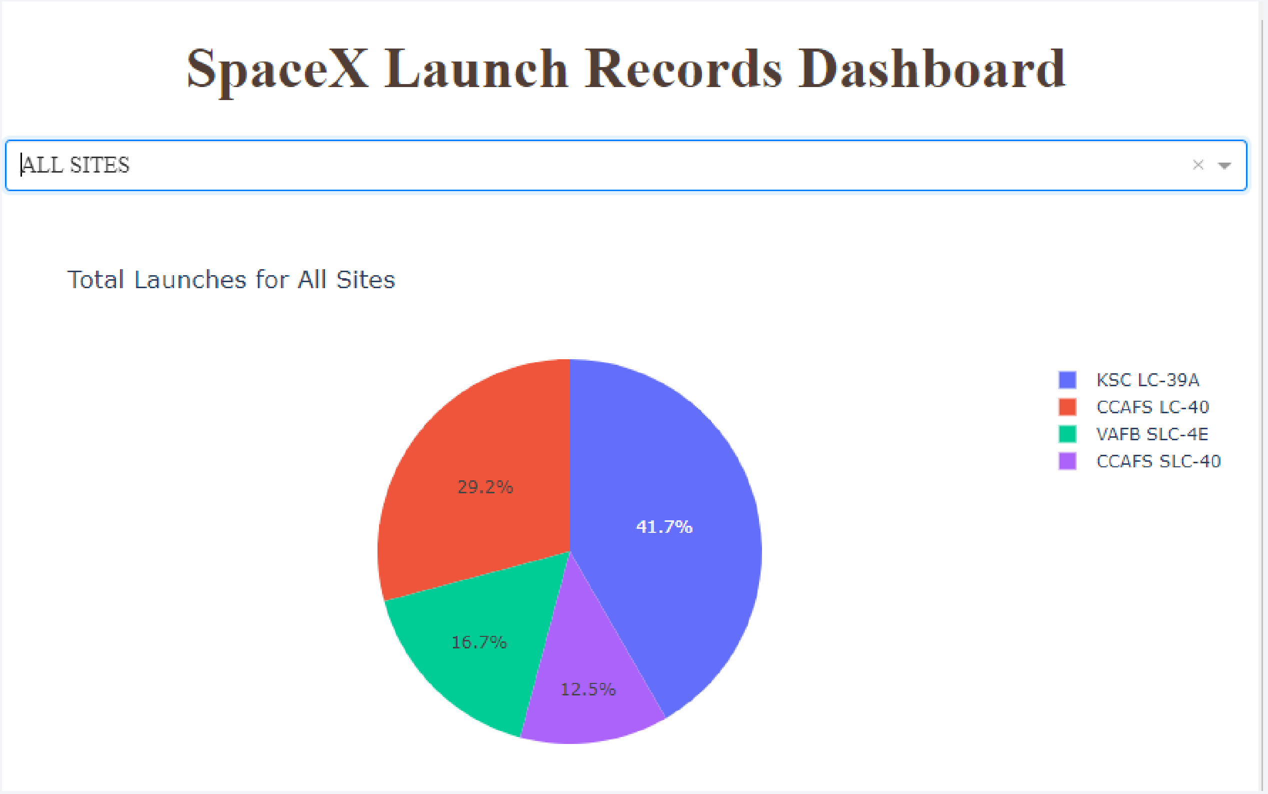
It is usual the launch sites are locate close to the coast and farther from the residential areas. Moreover the closer to the equator the better.



Section 4

Build a Dashboard with Plotly Dash

Success launches across sites



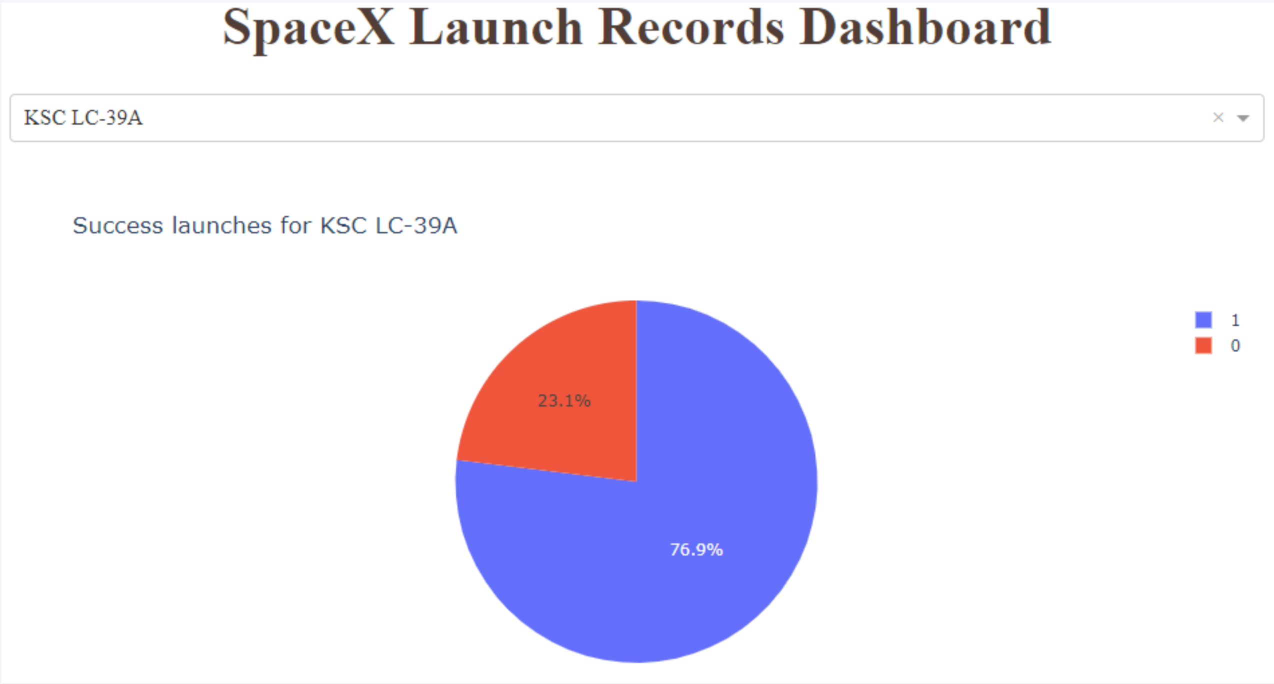
SpaceX Launch Records Dashboard

ALL SITES

The maximum number of successful launches is from KSC LC-39A site.

The minimum is from CCAFS SLC-40.

KSC LC-39A site launch statistics



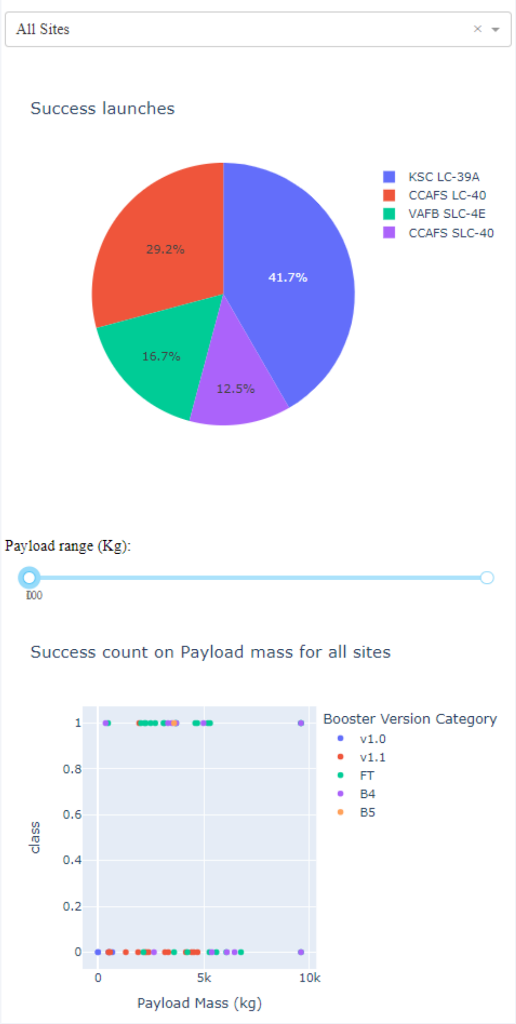
SpaceX Launch Records Dashboard

KSC LC-39A

Success launches for KSC LC-39A

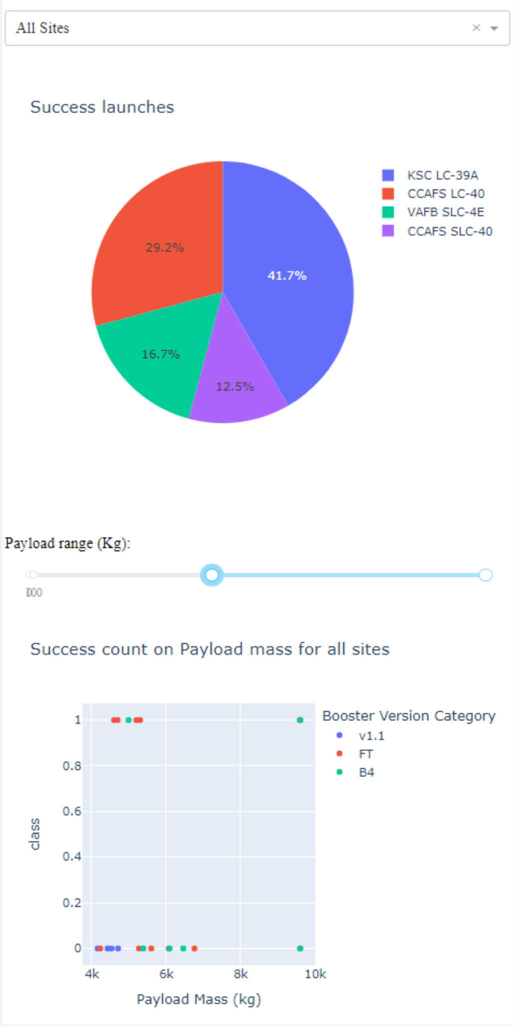
This is the statistics of successful launches (=1) of the best rank site.

Payload vs. Launch Outcome across all sites



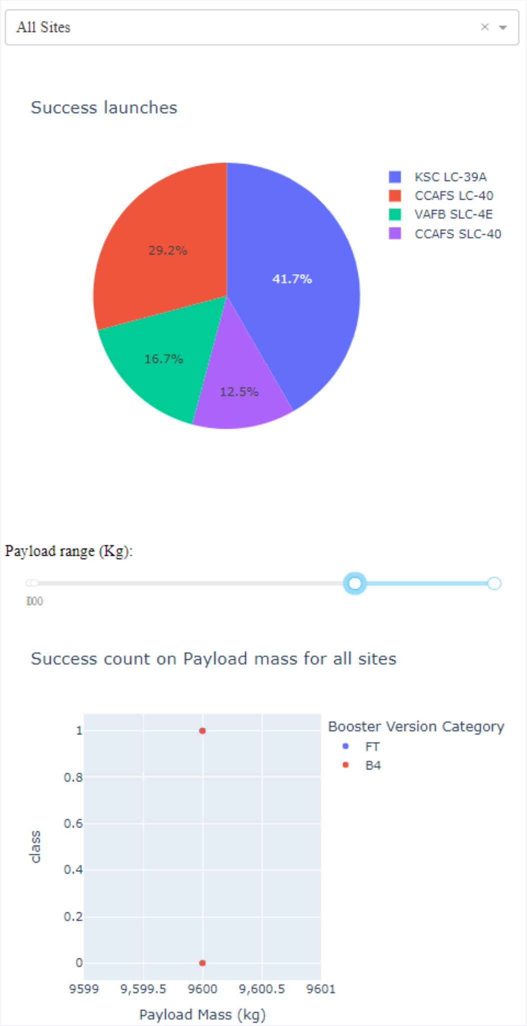
All Sites

Payload range (Kg):



All Sites

Payload range (Kg):



All Sites

Payload range (Kg):

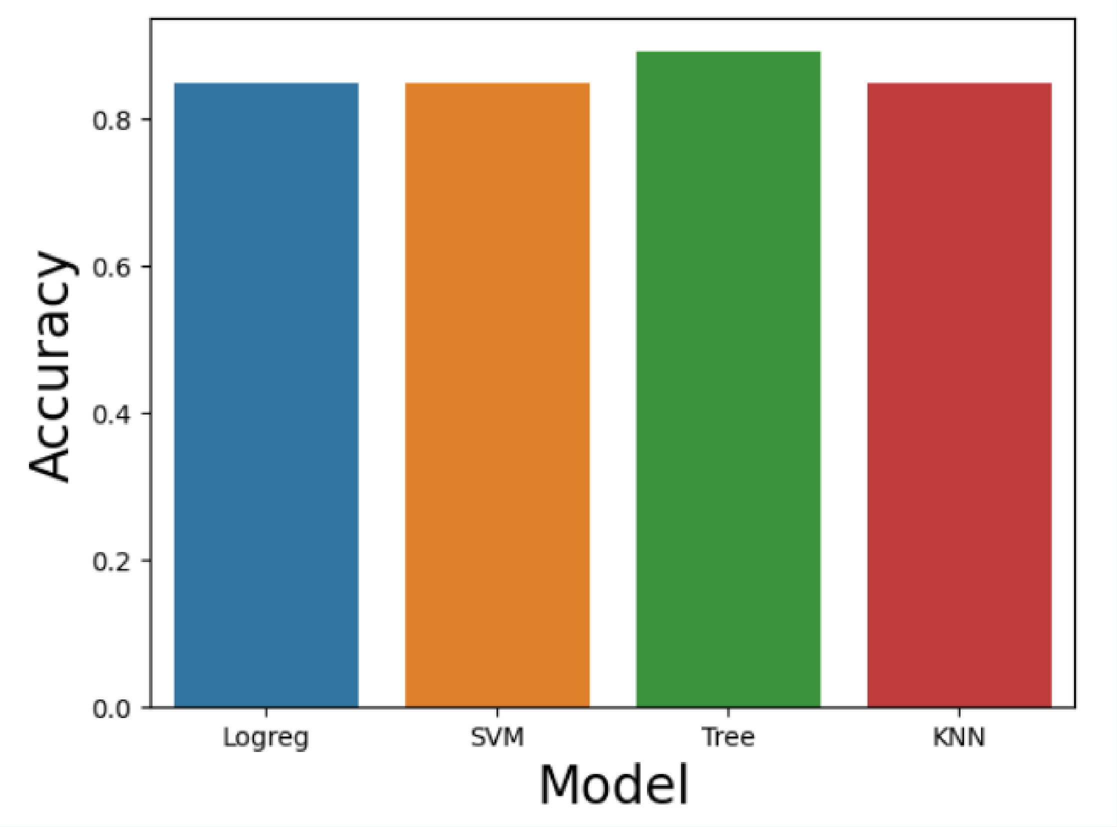
As we see the most successful launches were performed for the payload not more than 5.5k kg.

FT booster is the top one.

Section 5

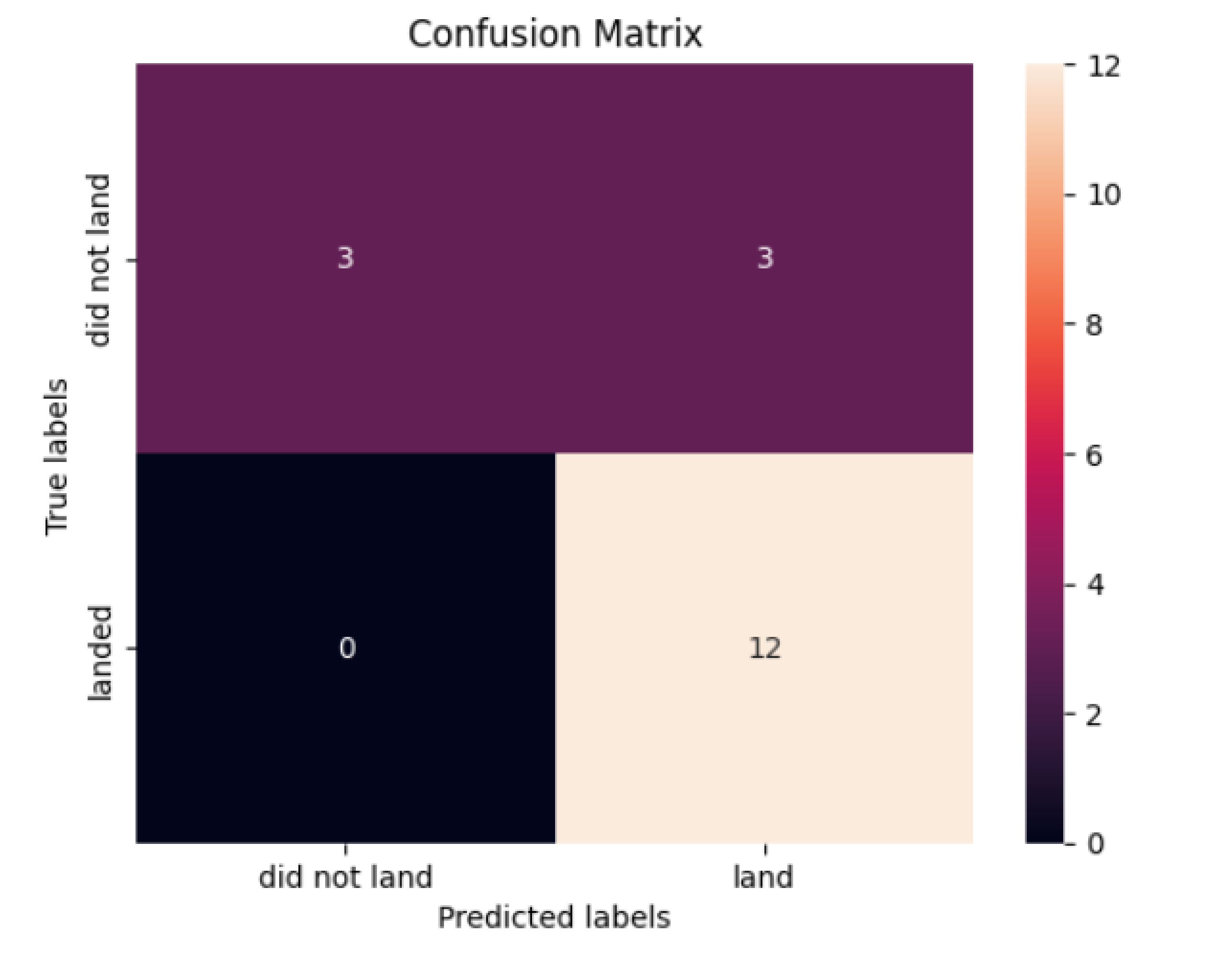
Predictive Analysis (Classification)

Classification Accuracy



The best performing model (even though they are quite close) is the Decision Tree Classifier with the accuracy level = 0.891

Confusion Matrix for the Decision Tree Classifier



We can see that the main problem of the model is False Positives for the “did not land” case.

All “landed” cases were predicted perfectly.

Conclusions

- There are at least 2 data sources: SpaceX API and Wikipedia
- The data should be preprocessed: cleaning, wrangling
- There’s enough data to perform the exploratory data analysis to make findings like:
    - There’s a yearly trend of increasing launch success rate
    - Higher payload tend to be more successful for launching
    - For all orbits the success rate increases
    - All kind of statistics, e.g. average payload, first successful landing, NASA total payload, landing outcomes etc
- All launch sites are located on the coast and as close to the equator as possible
- Residential areas are located in some secure distance from the launch sites
- The highest success rate is at the KSC LC-39A launch site in Florida with 76.9% rate
- The best prediction model trained on the data is the Decision Tree Classification

Appendix

All the code created can be found in the public GitHub repository under the link

https://github.com/igosm/Data-Science-Capstone

Acknowledgements

I would like to express my sincere gratitude to all the authors of the Data Science course that provided me with the deeper understanding of the subject.

Additional thanks for the Capstone course which I find extremely useful as it discovers something you could miss or misunderstand and calibrates the knowledge acquired.

Looking forward to seeing and taking more captivating courses!

Thank you!