IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dario Sangrigoli
21 September 2021

Outline

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

Executive Summary

- Objectives: determine the price of each SpaceX's launch and first stage reusage
- Methodologies used:
  - Data Collection & Data Wrangling
  - Exploratory analysis with Python & SQL
  - Interactive analytics with Plotly and Folium
  - Predictive analysis and Machine learning

Executive Summary

- The following were found to be the most decisive factors in terms of increasing the success rate of a launch:
  - Number of flights
  - Payload mass (depending on the launch site)
  - Orbits GEO, HEO, SSO, ES-L1
  - Time
- These factors have an impact often dependent on specific elements, as per subsequent analysis

Introduction

- Context:
  - SpaceY wants to enter the commercial space market and needs to collect data in regards to costs and methodologies.
  - SpaceX is the most successful competitor, which SpaceY intends to analyze to make its own project for market penetration.
- Questions:
  - What are the factors most affecting the success rate of a landing?
  - How are these factors influenced by other variables (ie, launch site, payload etc.)?
  - What are the best settings we need to plan in order to achieve the best success launch rate?

Section 1

Methodology

Methodology

Executive Summary

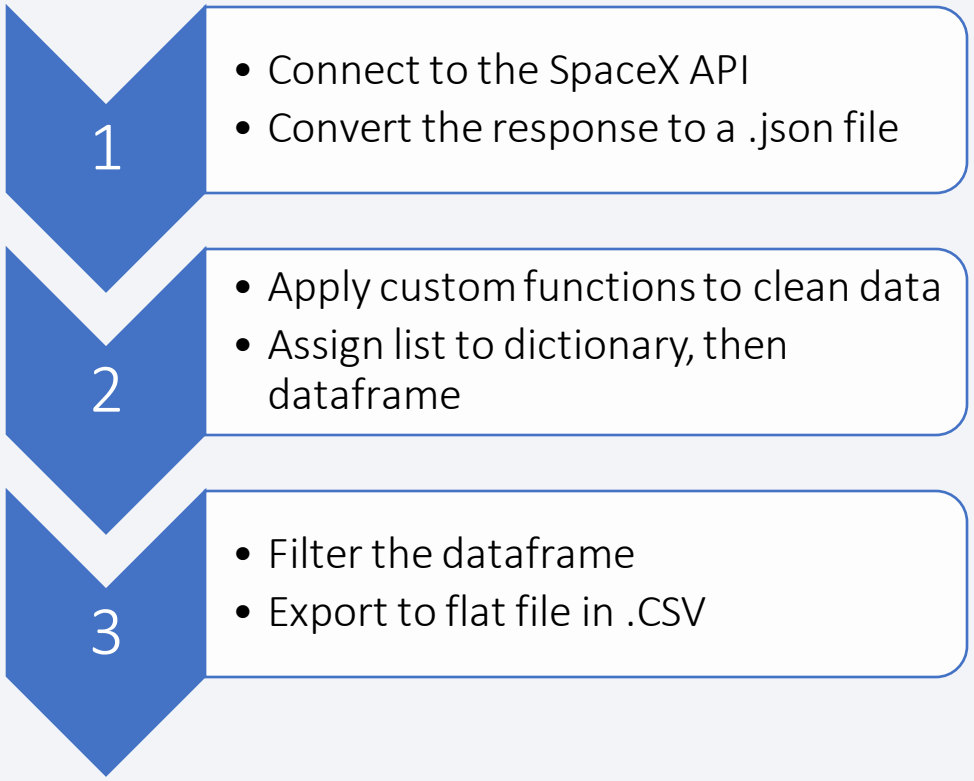
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data were collected using the following:
  - SpaceX REST API. This provides a comprehensive list of data about launches including the type of rocket, the payload, the specifications, the landing specifications and outcome.
  - Wikipedia WebScraping. The Wikipedia page contains the data in a similar format, and provides us with further tools and insights to process the analysis.

Data Collection – SpaceX API

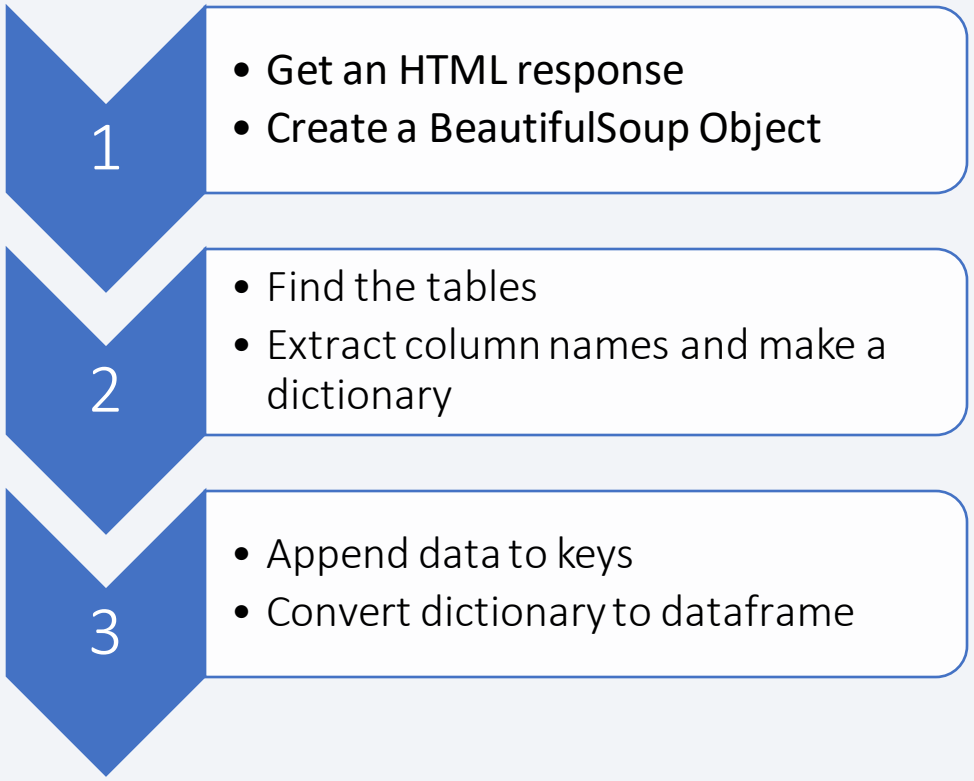
- The data collection from the SpaceX API was done by connecting to it and acquiring a .JSON file as a response. The data were then normalized into flat data in a .CSV format.



- Notebook URL

Data Collection - Scraping

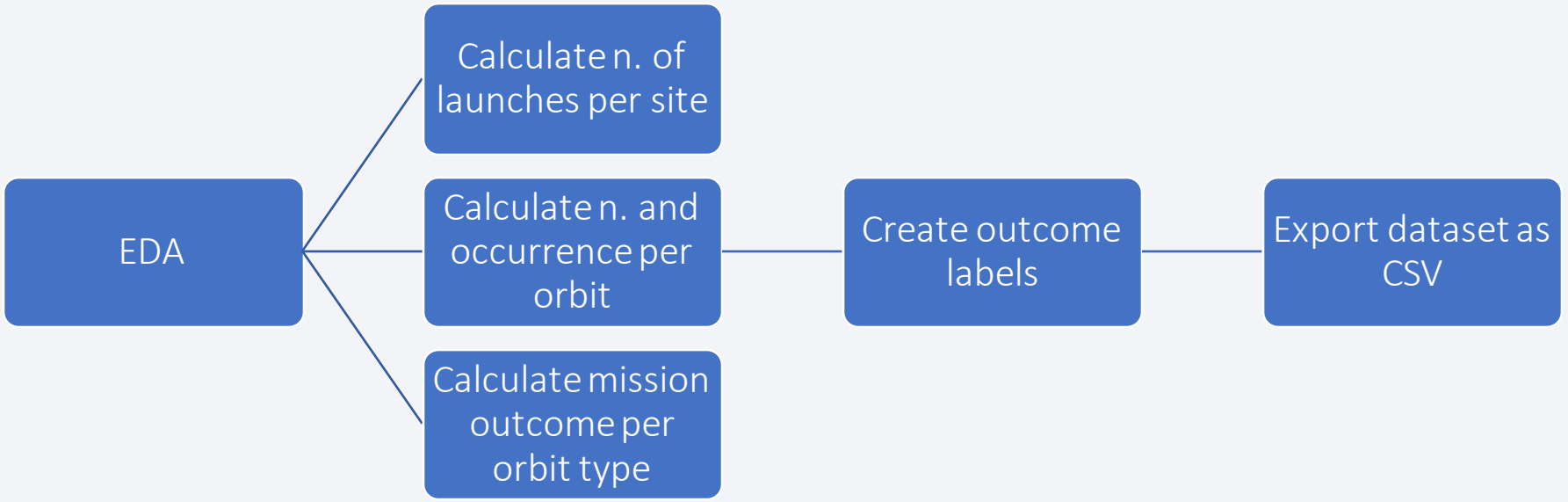
- The Web Scraping process involves connecting and extracting the HTML page as text with BeautifulSoup and then parsing the HTML into a dictionary, then make a dataframe out of it.



- Notebook URL

Data Wrangling

- The data were preliminary processed by performing Exploratory Data Analysis (EDA) and determining training labels for the next analysis steps.



- Notebook URL

EDA with Data Visualization

- I utilize and visualize the below charts. The aim is to uncover the relationships between the different variables and hint patterns, as illustrated in the next sections:
  - FlightNumber vs PayloadMass to SuccessRate scatterplot.
  - FlightNumber vs LaunchSite scatterplot
  - PayloadMass vs LaunchSite scatterplot
  - Orbit vs SuccessRate barchart
  - FlightNumber vs Orbit scatterplot
  - PayloadMass vs Orbit
  - Success rate time series.

- Notebook URL

EDA with SQL

- The SQL EDA was performed with the following steps:
  - Load the SQL extension
  - Connect to the IBM DB2
  - Display the names of unique launch sites
  - Display 5 records where launch sites begin with 'CCA'
  - Display total payload mass carried by NASA's boosters
  - Display average payload mass carried by boosters F9 v1.1
  - List the date when first successful landing outcome in ground pad was achieved
  - List names of boosters successful in drone ship with payload mass > 4000 and <6000
  - List total number of successful and failure mission outcomes
  - List names of the booster version which carried the max payload mass
  - List failed landing outcomes in drone ship, their booster versions and launch site names in 2015
  - Rank the count of landing outcomes in a date range

- Notebook URL

Build an Interactive Map with Folium

- To build our map, I performed the following:
  - Identified the sites' coordinates
  - Created a Folium object centered to NASA Johnson Space Center in Huston, adding a highlighted circle area with text label
  - Added a circle and a marker for each site, to identified them and ascertain their relative position
  - Marked the success-failed launches of each site, to identify patterns and clusters
  - Calculated the distances beween a launch site to its proximities, identifying points of interest, to acquire further insights on the specific charachteristics of the environment of each launch site

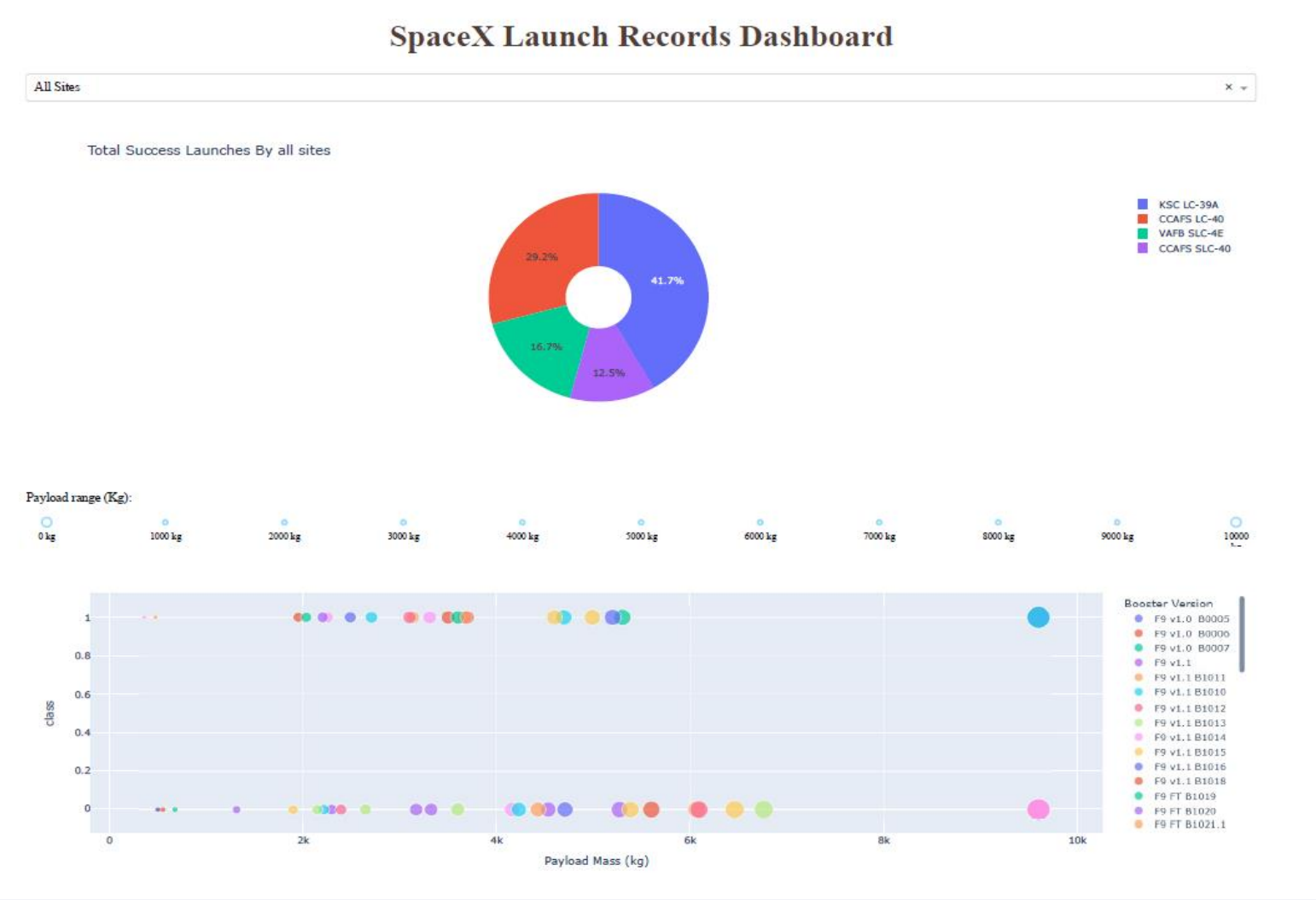
- Notebook URL

Build a Dashboard with Plotly Dash

- The dashboard was built with Dash framework, and hosts the following content:
  - A pie chart showing the total launches per site or of all sites. It displays the relative proportion of multiple classes of data.
  - A scatter graph, showing the relationship of two variables per different booster version: Outcome and Payload Mass. It shows in a clear and straightforward presentation the non-linear pattern in the data.

- CODE URL - please refer to the next slide to see the resulting Dashboard

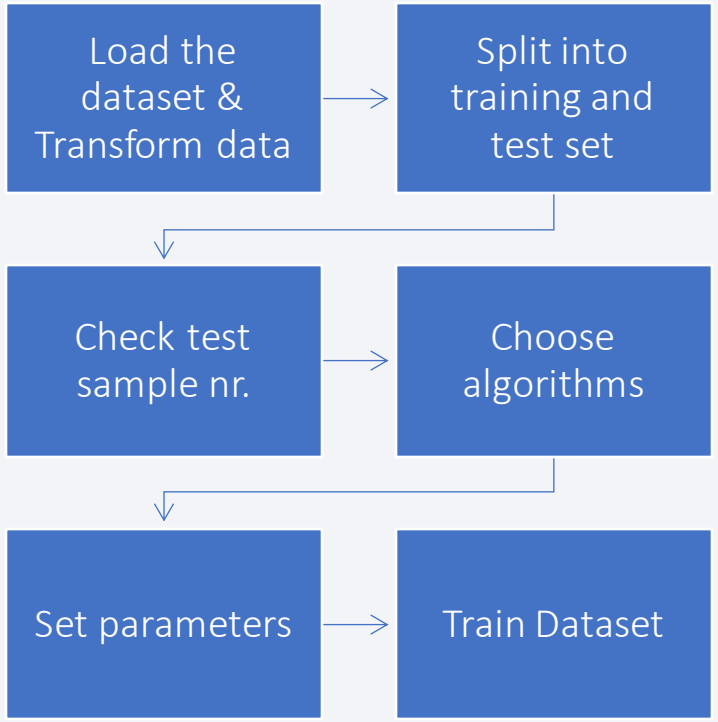
Build a Dashboard with Plotly Dash



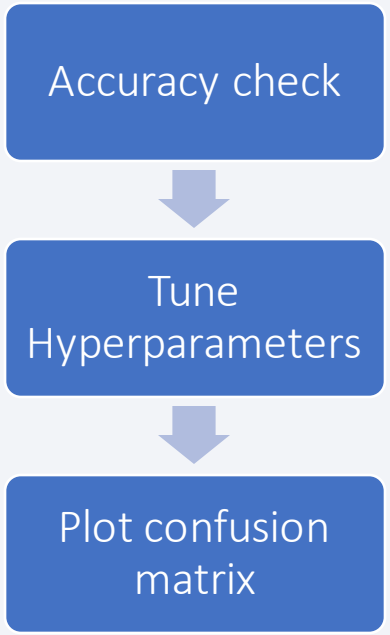
Predictive Analysis (Classification)

- Notebook URL

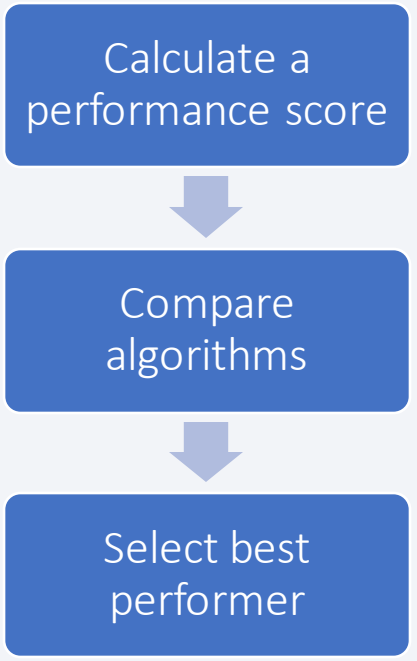
Model Building



Model Evaluation

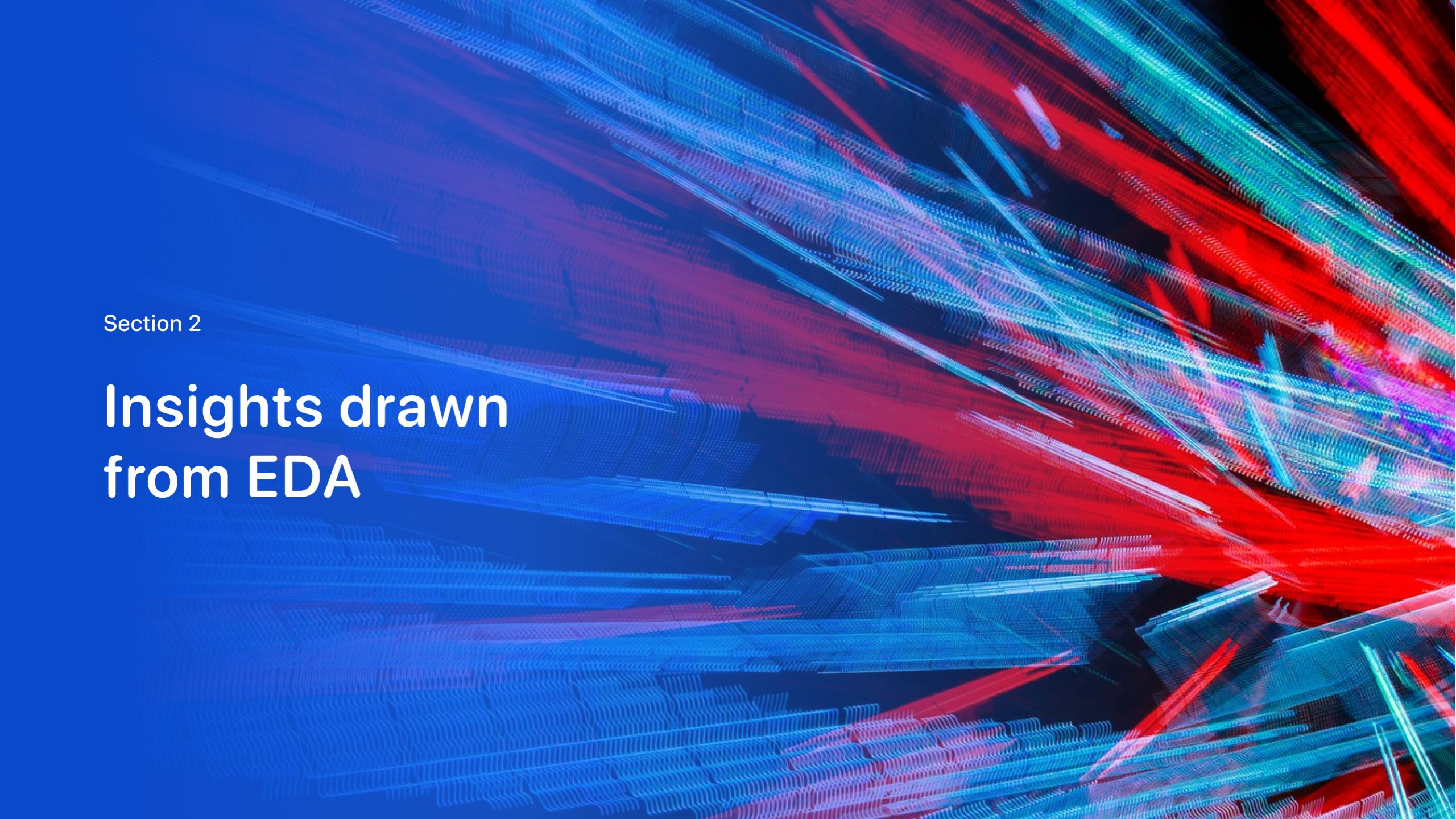


Model Performance



Results

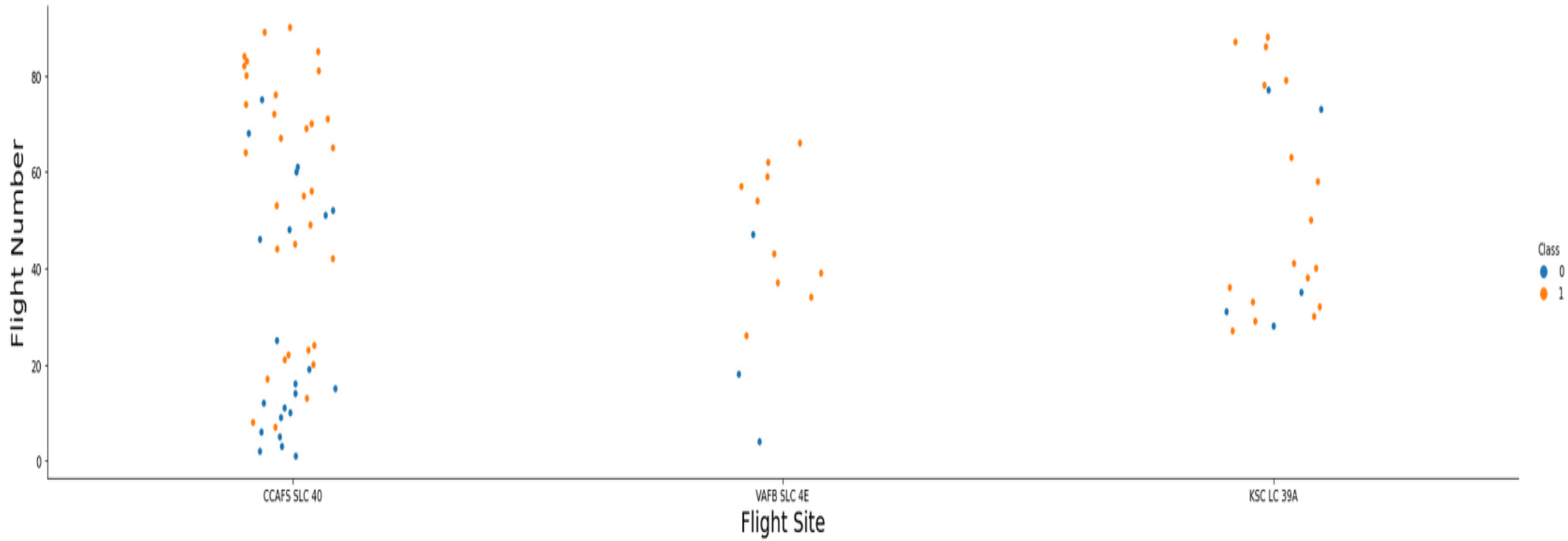
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

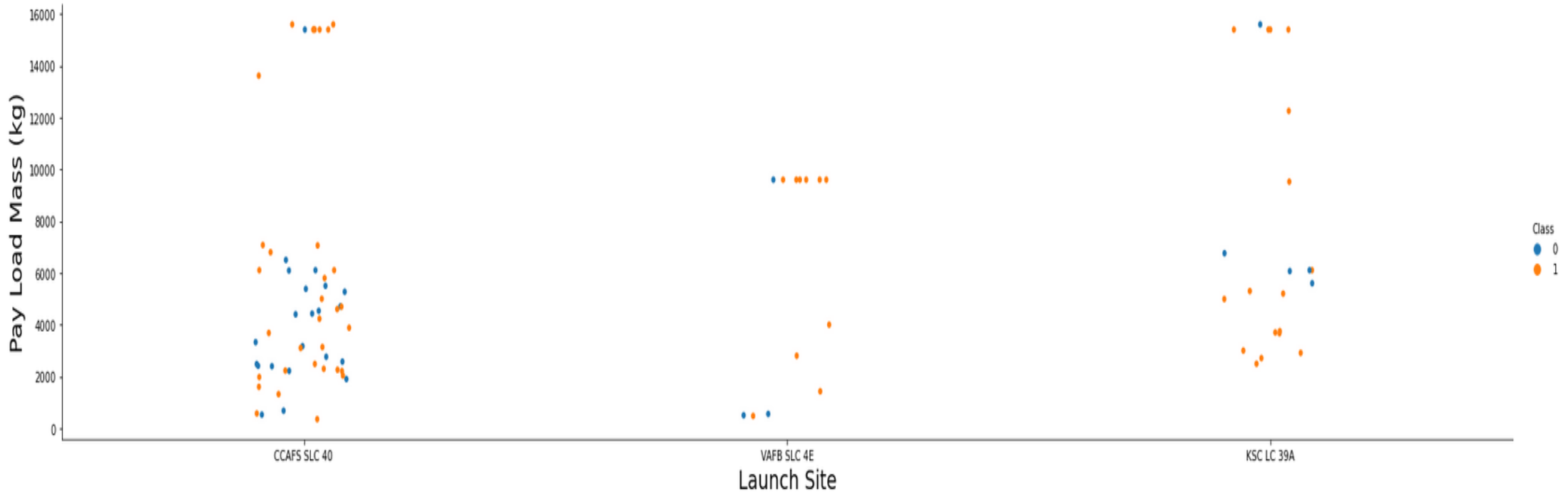
Insights drawn from EDA

Flight Number vs. Launch Site



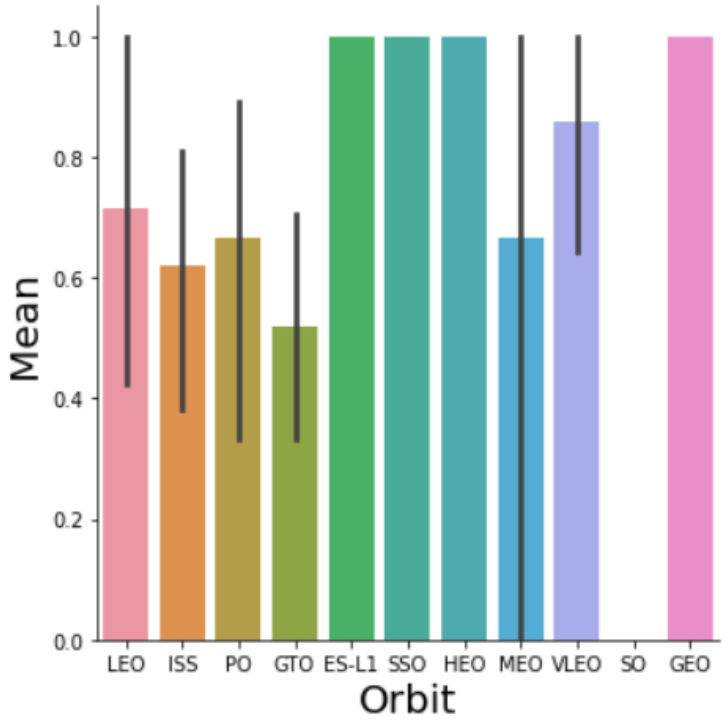
The relationship between these two parameters is overall unclear. there seems to be a slight positive correlation of the success rate with the number of flights, especially for CCAFS SLC 40 where there seem to be concentrations of success/failure hotspots, but it is overall not conclusive.

Payload vs. Launch Site



Different launch sites seem to host launches of different Pay Load Mass. CCAFS SLC 40 and KSC LC 39A seem to host mostly launches with high or low Pay Load Mass, while VAFB SLC 4E focuses more on medim Pay Load Mass launches. Failures seem concentrated more when the Pay Load Mass is low, while the rate of failure is low when considering medium and high Pay Load Mass launches.

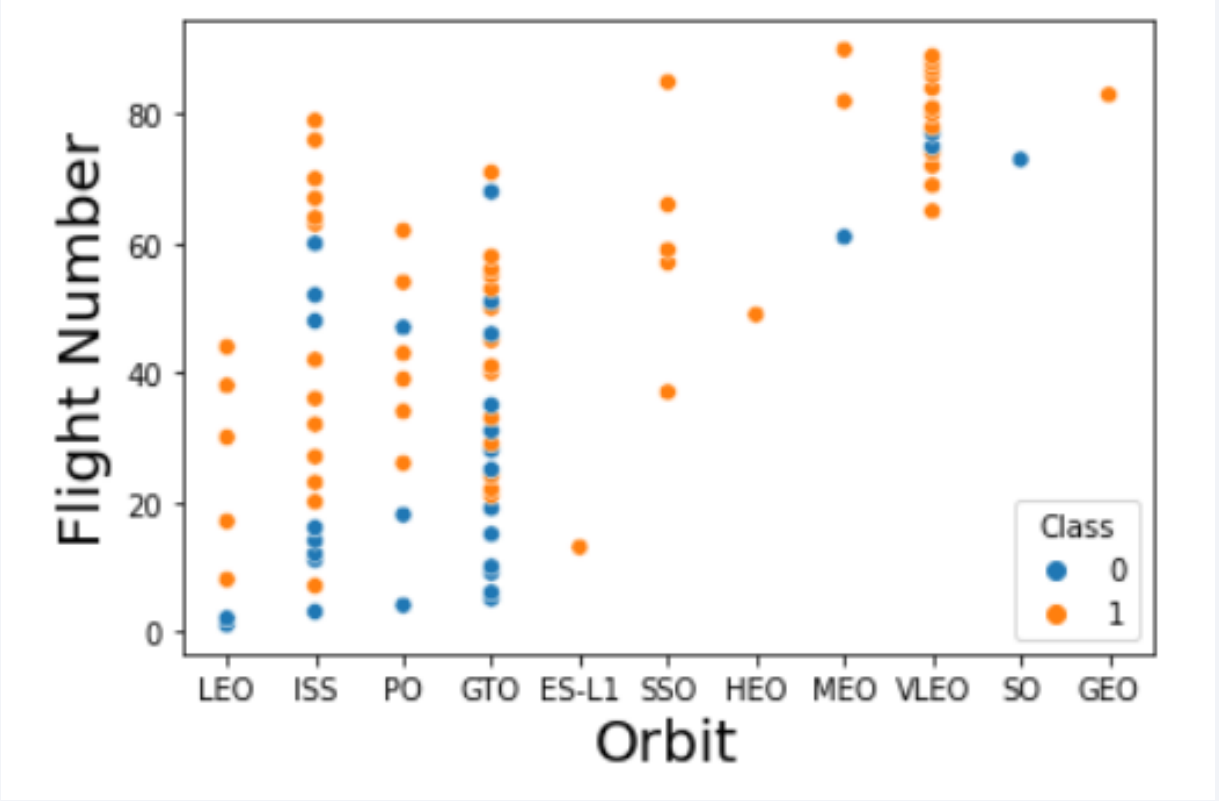
Success Rate vs. Orbit Type



THe orbits with the best success rate are GEO, HEO, SSO, and ES-L1.

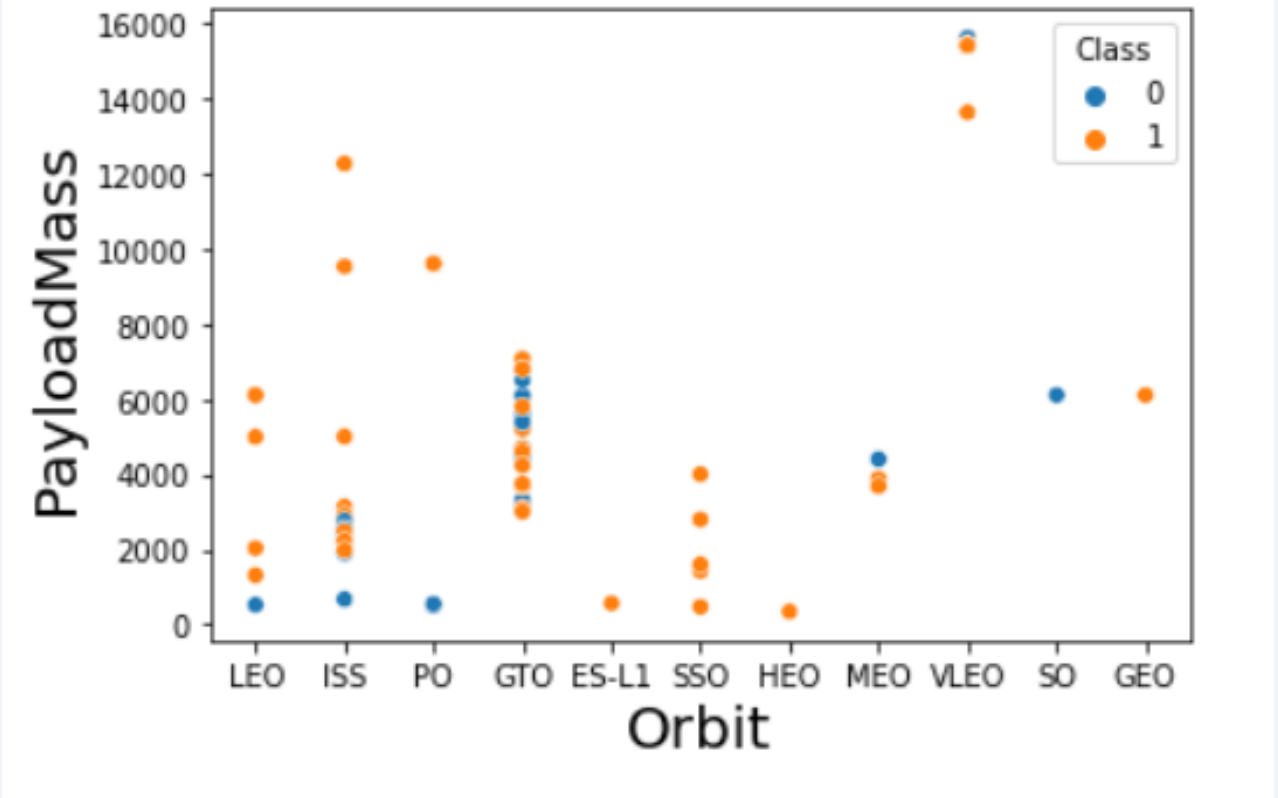
Flight Number vs. Orbit Type

You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



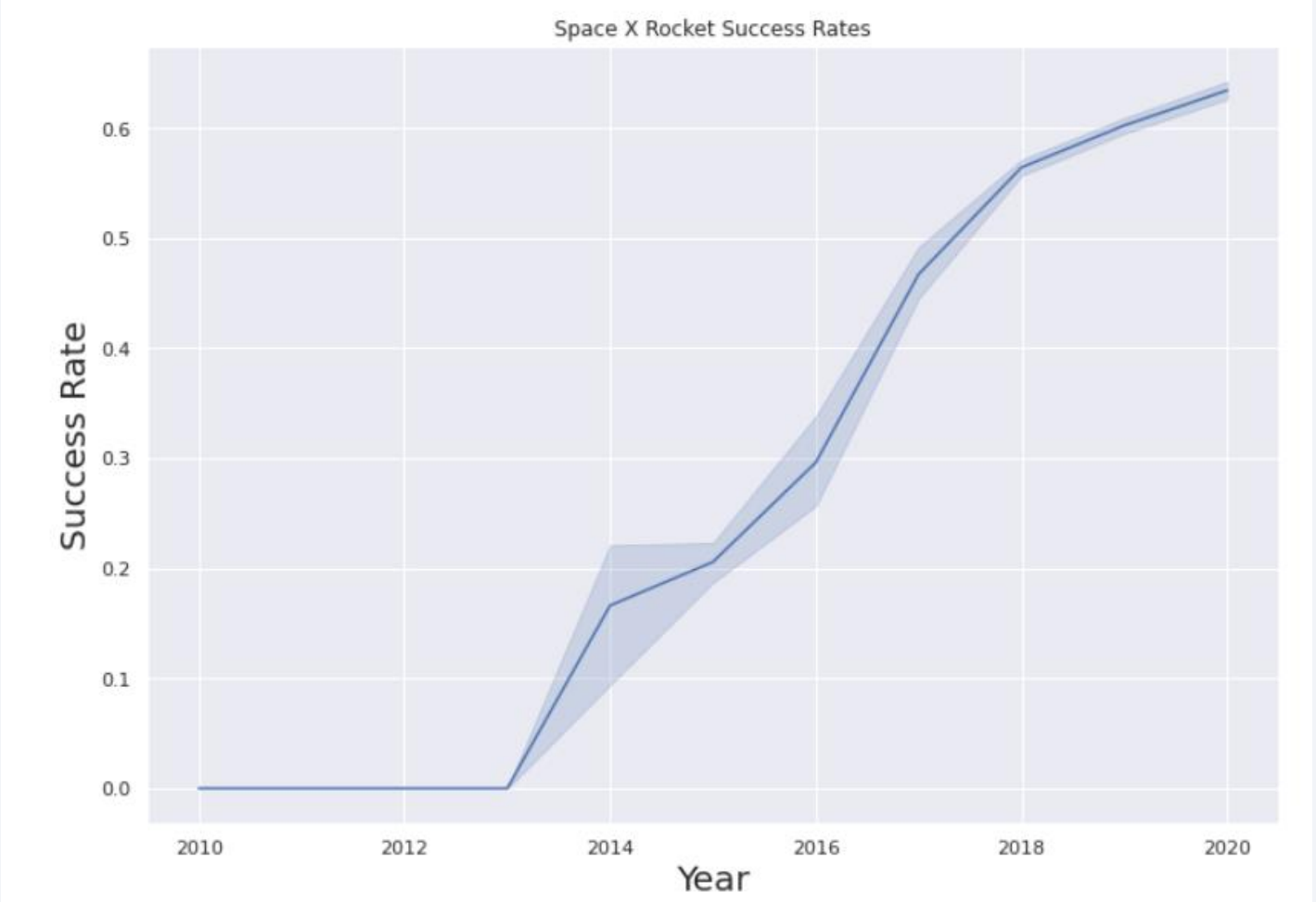
Payload vs. Orbit Type

- You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- You can observe that the sucess rate since 2013 kept increasing till 2020. This is most probably due to a learning curve process and the advancement of technology over time.



All Launch Site Names

- The query to retrieve the launch site names has been structured as a simple SELECT / DISTINCT statement.

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

Launch Site Names Begin with 'CCA'

- The list below was retrive by selecting all dataset, using a wildcard to restrict the result to those beginning with 'CCA' and adding a Limit 5 condition.

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

- `%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer = 'NASA (CRS)'`

- The above code retrieves the Payload mass with a Select statement restricted to NASA (CRS) customers.
- This returns a result of 45596.

Average Payload Mass by F9 v1.1

- `%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version = 'F9 v1.1';`
-
- The above queries selects the average payload mass restricting the selection to the booster version F9 v1.1.
- The result obtained is 2928.

First Successful Ground Landing Date

- `%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE mission_outcome = 'Success';`

- The above query finds the 'minimum' date where the outcome of the mission was a success.
- The resulting date is 2010-06-04.

Successful Drone Ship Landing with Payload between 4000 and 6000

- `%sql SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;`
-

- The above query selects the distinct booster version in a restricted range.

booster_version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

- `%sql SELECT mission_outcome, COUNT(*) AS Total FROM SPACEXDATASET GROUP BY mission_outcome;`
- The above query counts the distinct mission outcomes grouped by outcome.

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

Boosters Carried Maximum Payload

- %sql SELECT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET)

- The above query uses a subquery to retrieve the boosters that carried the 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

2015 Launch Records

- `%sql SELECT booster_version, launch_site, landing__outcome, DATE FROM SPACEXDATASET WHERE landing__outcome LIKE 'Failure%' AND DATE LIKE '2015%'`
-
- This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

- %sql SELECT landing__outcome, COUNT(landing__outcome) FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY COUNT(landing__outcome) DESC
-
- The query ranks 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

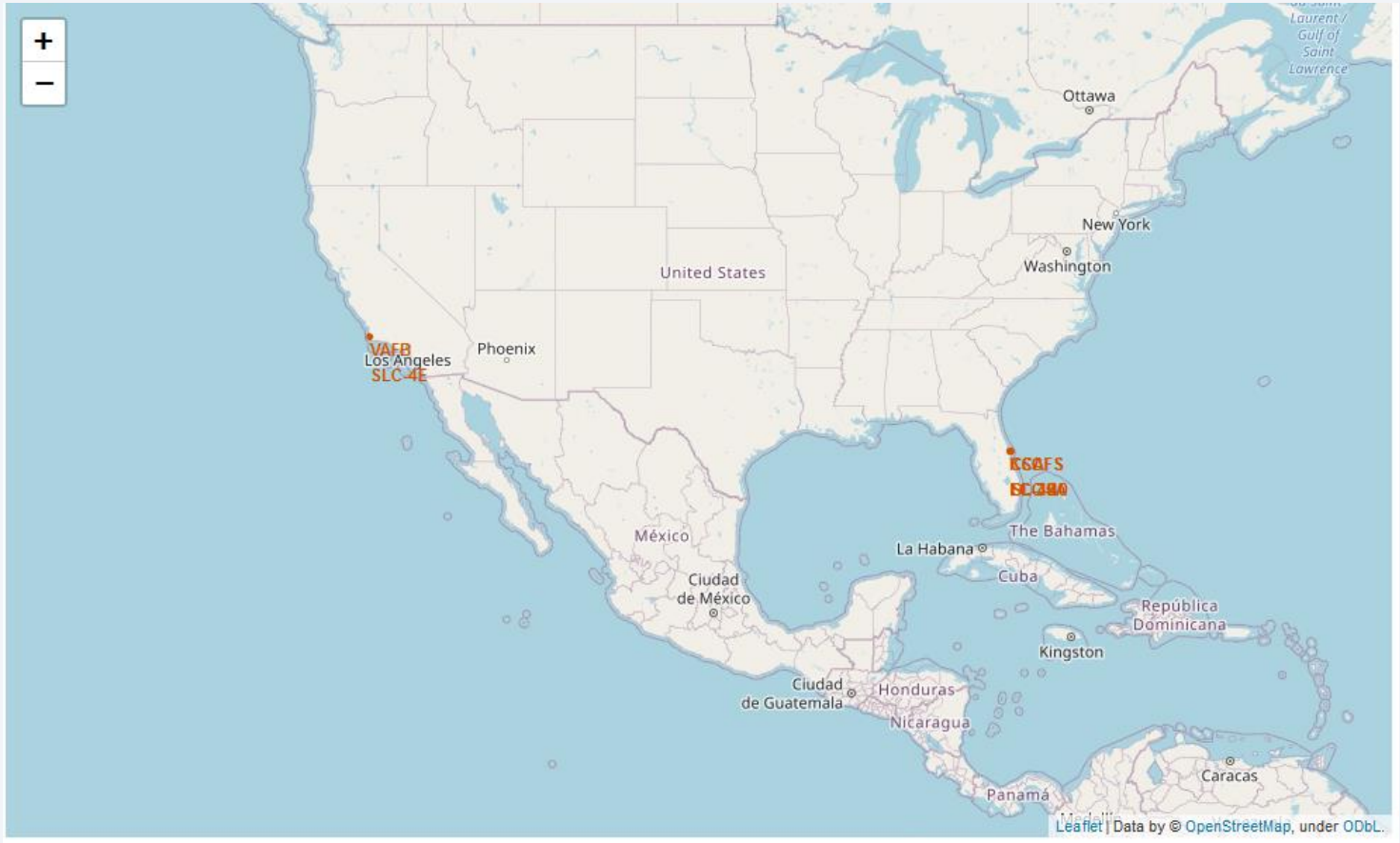
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

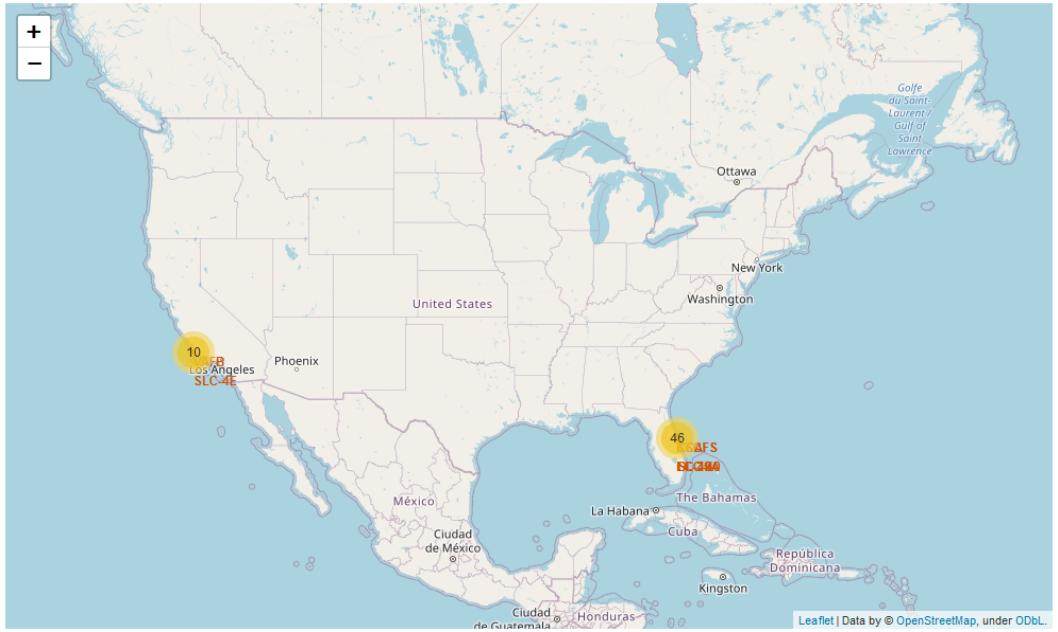
Launch sites location

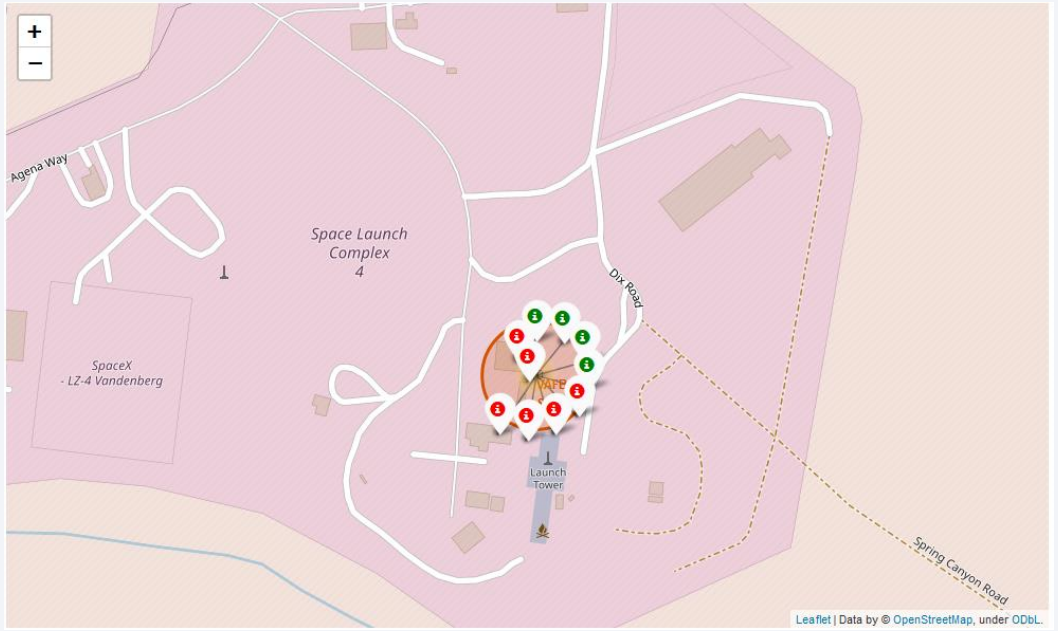
- This map shows the exact location of the launch sites, distributed in a few close-distance areas of the east and west coasts of US.



Launch outcomes map

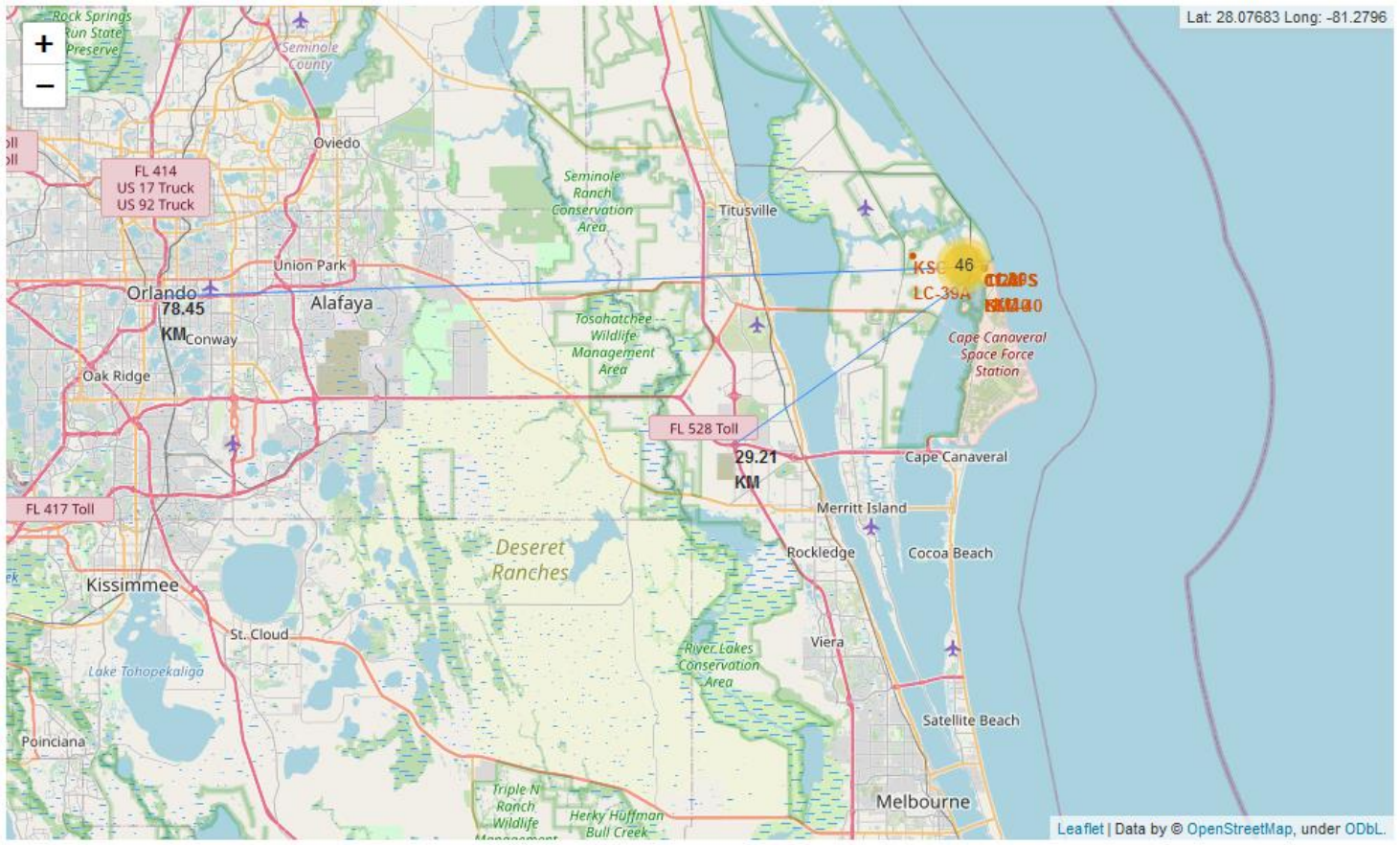
- The left map shows, in absolute numbers, the success launches per each location. If we drill down as per map on the right, we can ascertain exactly the number of success and fail launches per location.

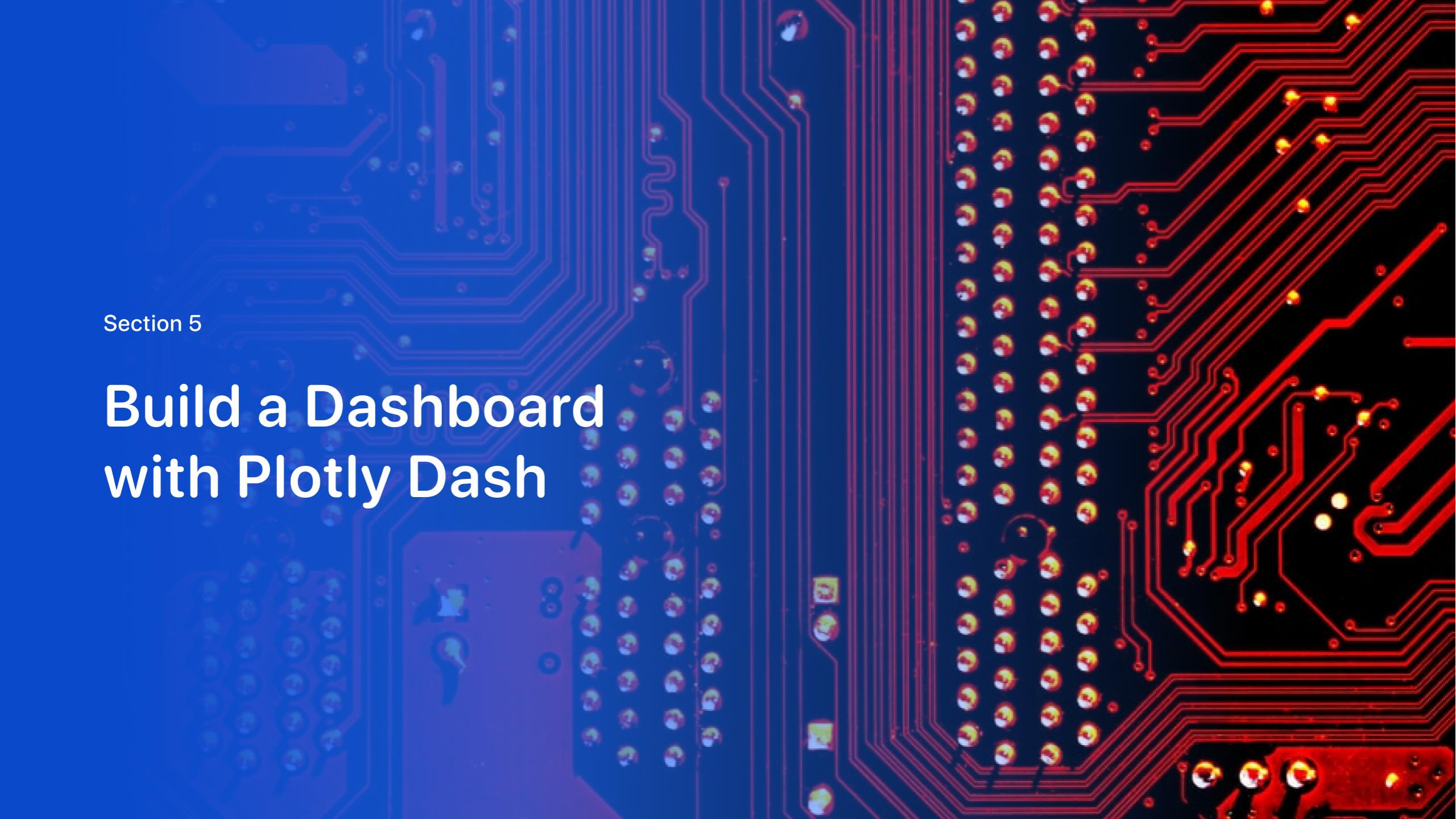




Proximity to points of interest

- The map shows with blue lines the proximity of a location to a number of points of interest. In this case, the location is close to highways and the coastline, but not to a railway or Florida City.



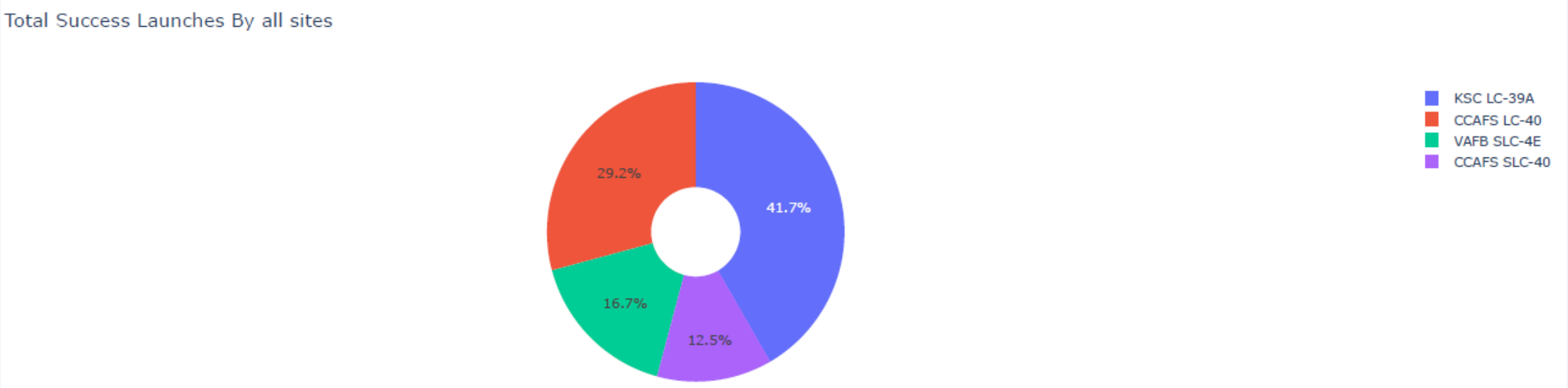


Section 5

Build a Dashboard with Plotly Dash

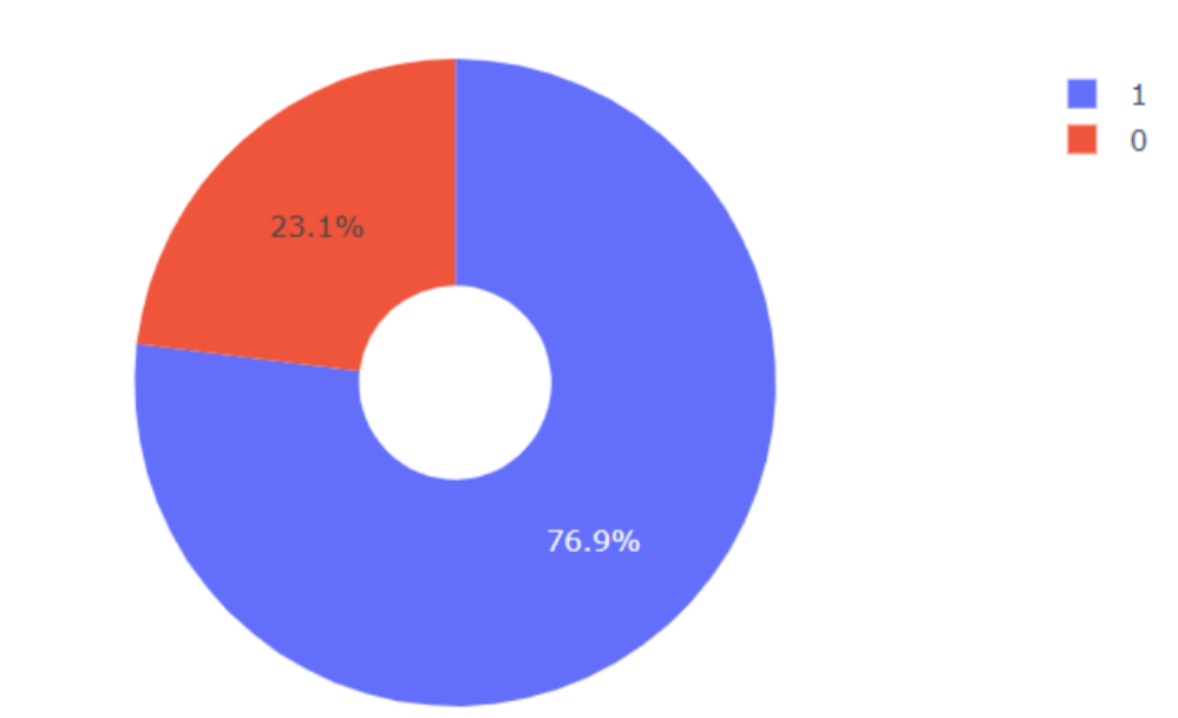
Success launches per site

- The chart shows the proportion of success launches for all sites. KSC LC-39A has the highest proportion with respect to all other launch sites.



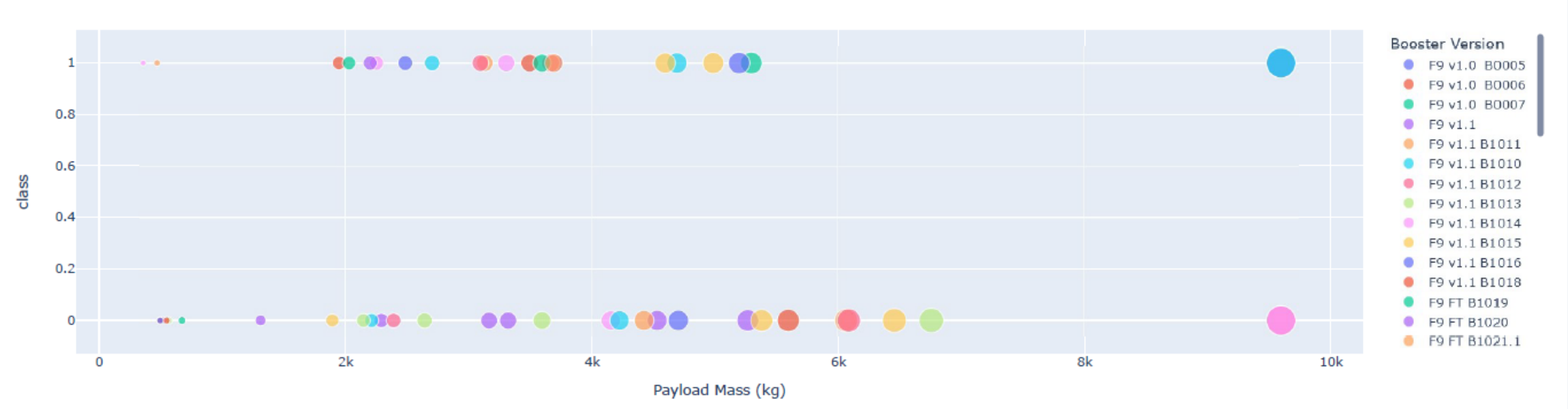
Highest launch success ratio

- The highest launch success ratio was achieved by KSC LC 39-A.



Payload vs Launch outcome

- The highest success rate was achieved with the F9 v1.1B1016 and F9 v1.1B1016 boosters.

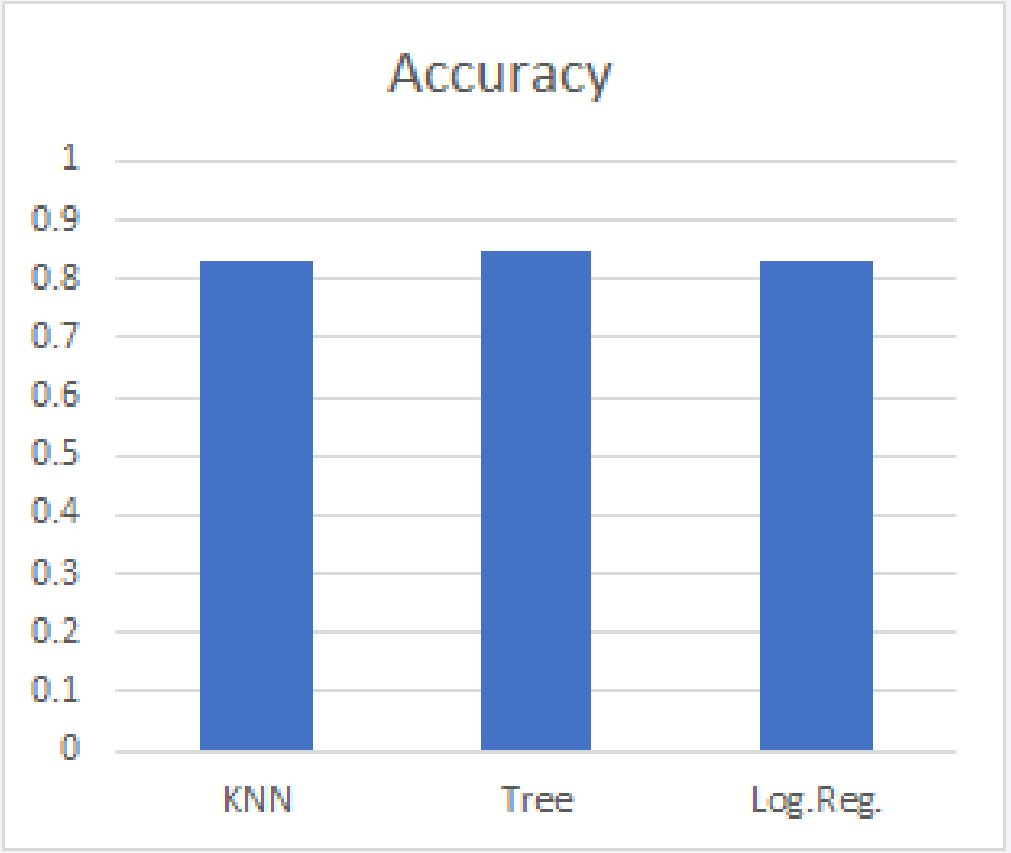


Section 6

Predictive Analysis (Classification)

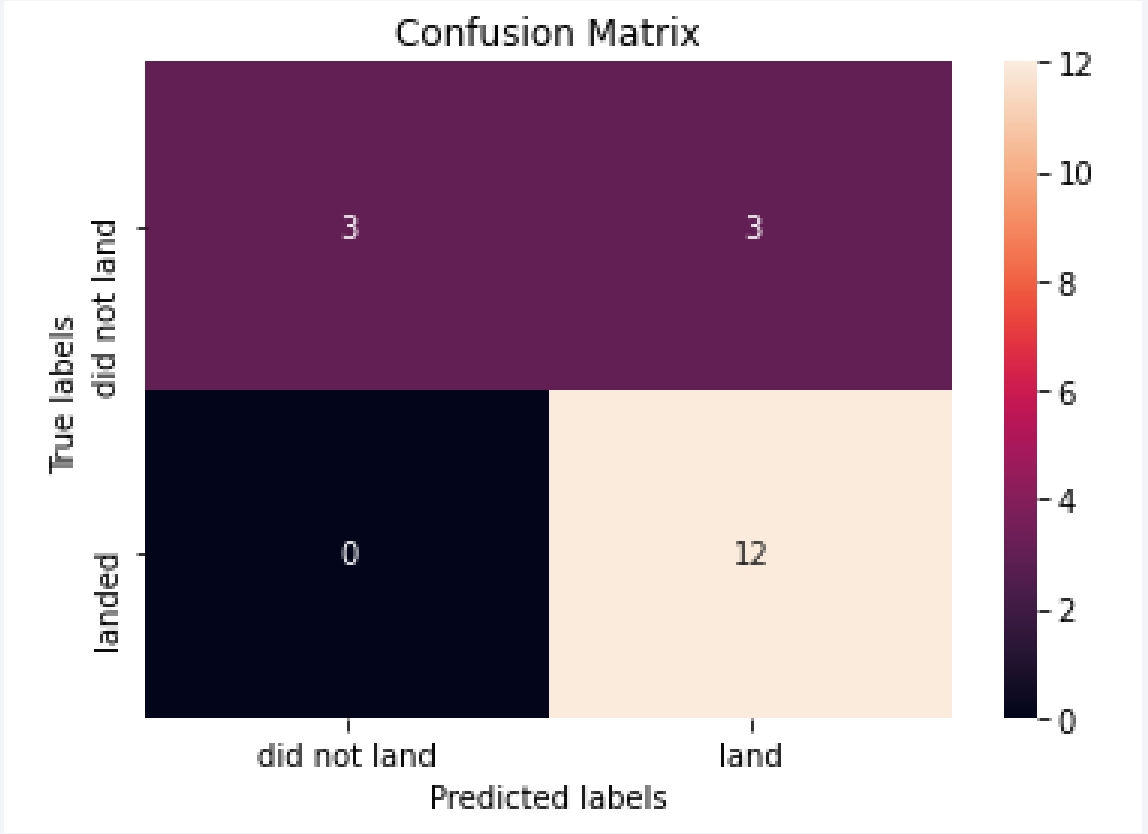
Classification Accuracy

- The highest accuracy score was obtained by the decision tree, with a precision of 84.82%.



Confusion Matrix

- The Confusion Matrix for the Decision Tree model performed relatively poorly with respect to false positive, while the performance for false negatives was perfect.



Conclusions

- The success rate is highly time-dependent. The learning curve and the technology enhancements are probably highly correlated with improvement of performance.
- The level of payload has shown to be a fairly good predictor of the launch success. Weaker results were obtained when considering the number of flights per site.
- The launch sites are influenced in partially different ways when different variables are compared. Overall, the best performing site seems to be KSC LC 39-A.
- The best predictor of launch result seems to be obtained applying a decision tree analysis. The confusion matrix analysis corroborates this conclusion.

Appendix

- All assets, code, graphs and results are stored in the notebooks linked in the first slides. Please make reference to them and any note you will find beside them.

Thank you!