Winning Space Race with Data Science

Jansen Machado
December 19, 2022

Outline

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

Executive Summary

Summary of methodologies:

> API

> Web Scraping

> Data Wrangling

> Data Analysis and Exploration Using:

- SQL
- Data Visualization
- Folium

> Machine Learning Prediction

Summary of all results:

> Data Analysis and Exploration Results

> Prediction Comparison and Results

Introduction

Project background and context

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 Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. This Capstone Project goes through the entire process and culminates in creating a machine learning pipeline that predicts if the first stage will land, given the data.

Problems you want to find answers

> Where there were successful and unsuccessful landings
> What can impact the success/fail rate and what are tied to them
> What are the patterns, which can be used, and how they influence

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilized API and Web Scraping
- Performed data wrangling
  - Only Falcon 9 data was used; one-hot-encoding for non-numerical categories
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Explored launch sites, orbits, payloads and success/failure rates

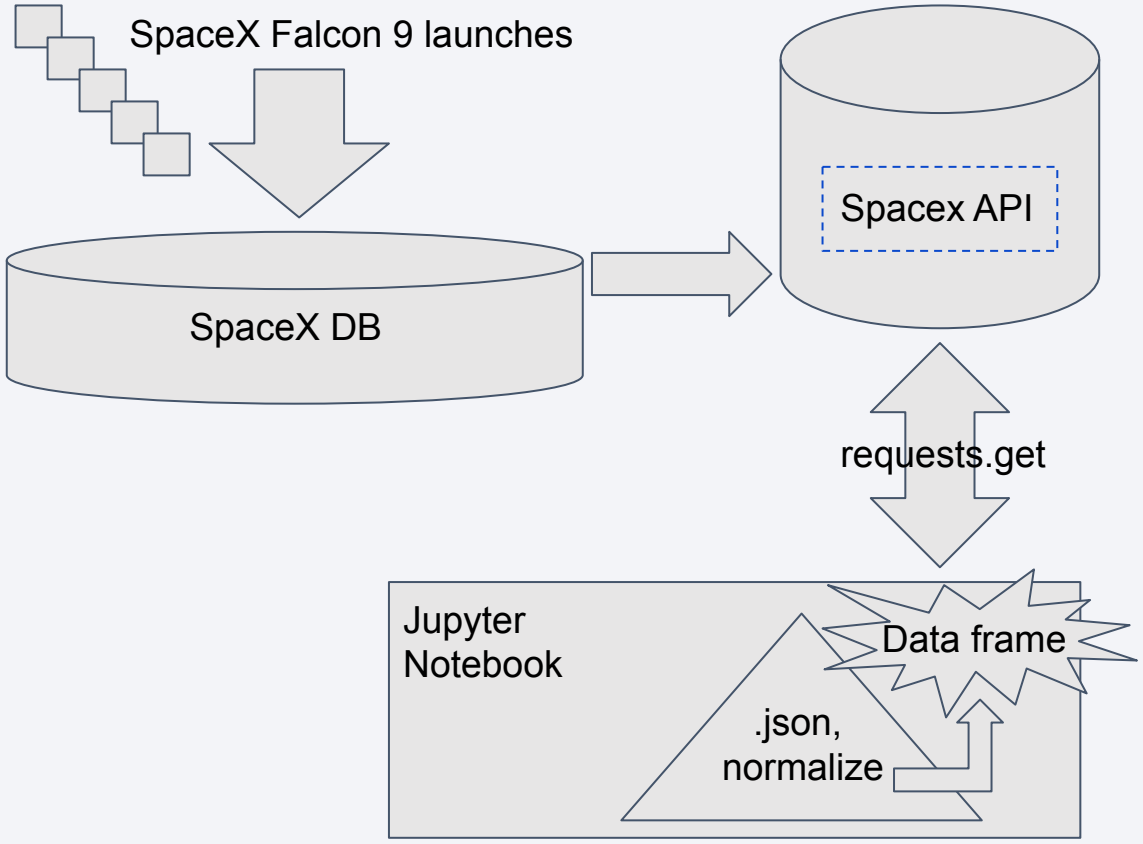
Data Collection

- SpaceX API was used to collect data on:
  - Booster Version
  - Launch Site
  - Payload Mass
  - Core (the reusable booster)
- Web Scraping was used to collect data on Falcon 9 Launch Data on:
  - Flight number, Date and Time
  - Booster Version and Landing
  - Launch Site
  - Payload and Payload Mass
  - Orbit Destination

Data Collection – SpaceX API

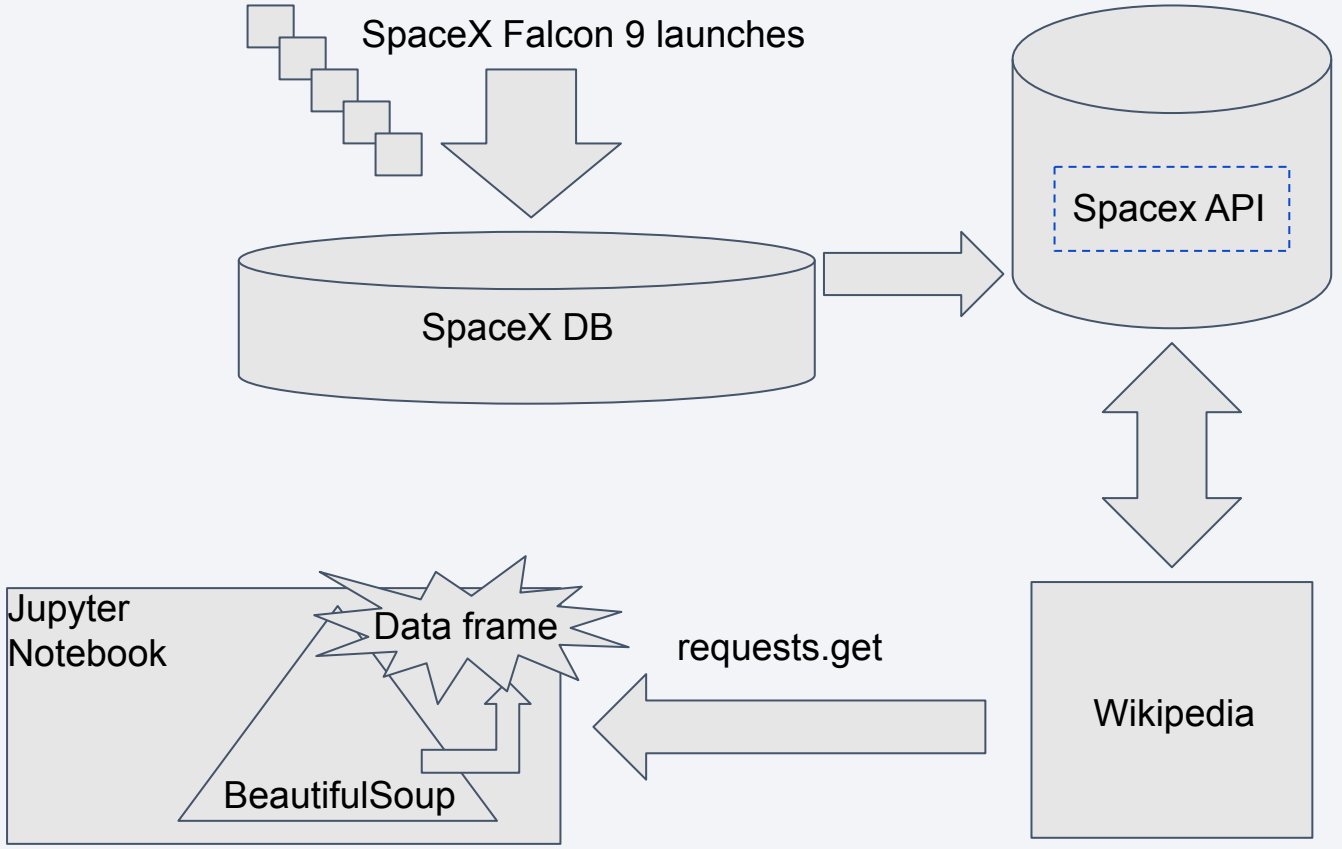
- Data was pulled from SpaceX API using the requests method, normalized, and a specific data frame created with only Falcon 9 launches and certain columns. That included checking for, removing and replacing missing values with the mean, where applicable.

- Data Collection – SpaceX API Github



Data Collection - Scraping

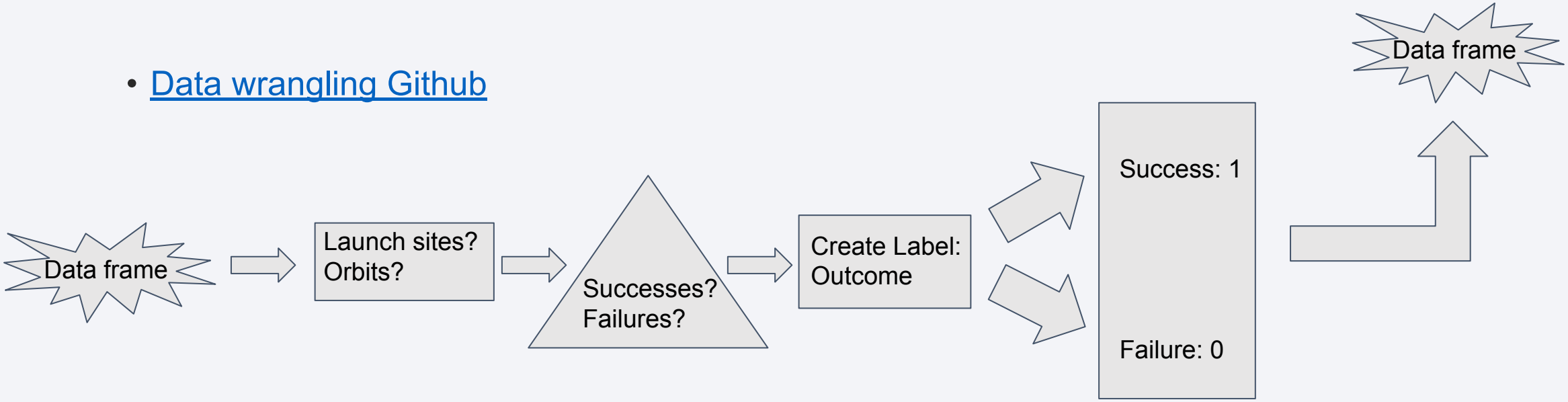
- Data was pulled from a Wikipedia table with Falcon 9 launches using the requests method. A BeautifulSoup object was created, columns identified, a dictionary created in order to parse the information and create a data frame.



- [Data Collection - Scraping Github]

Data Wrangling

- Identified and determined the number of launches on each site and orbits
- Utilized launches' respective success and failure outcomes, and thus, their rates
- Data wrangling Github



EDA with Data Visualization

- Seaborn was used to explore the relationships below:
  - Categorization between Flight number/ Payload Mass, Launch Site
  - Scatter plot for Payload Mass/Launch Site, Flight Number/Orbit and Payload Mass/Orbit
- Matplotlib was used to visualize:
  - Using bars, the relationship between Success Rate and Orbit Type
  - Using a line, the Yearly Trend of launch successes

- EDA with Data Visualization Github

EDA with SQL

- The names of all the unique launch sites in the space mission
- Five records of launch sites beginning with “CCA”
- Total payload mass carried by NASA (CRS) boosters
- Average payload mass carried by F9 v1.1 booster
- Date of the first successful landing on a ground pad
- Boosters that successfully landed in a drone ship with payload between 4000 and 6000 Kg
- Total number of successful and failed mission outcomes
- The name of booster versions which have carried the maximum payload
- Records which show month names, failed landing outcomes in drone ships, booster versions, launch site name for the months in 2015
- Rank the amount of successful landings between 04-06-2010 and 20-03-2017 in descending order

- EDA with SQL Github

Build an Interactive Map with Folium

- Markers and circles were created for all launch sites to visualize their national positioning
- Each site has its own set of marker cluster showing the successful and failed launch locations to deliver a visual of the success rate at each site and their specific positions
- Two lines were drawn from a launch site: one, to a marker delimiting the closest coastline and a second to a major highway intersection. It can be shown how far they are from hazardous activity such as launching rockets a launch site
- Folium Interactive Map Github

Build a Dashboard with Plotly Dash

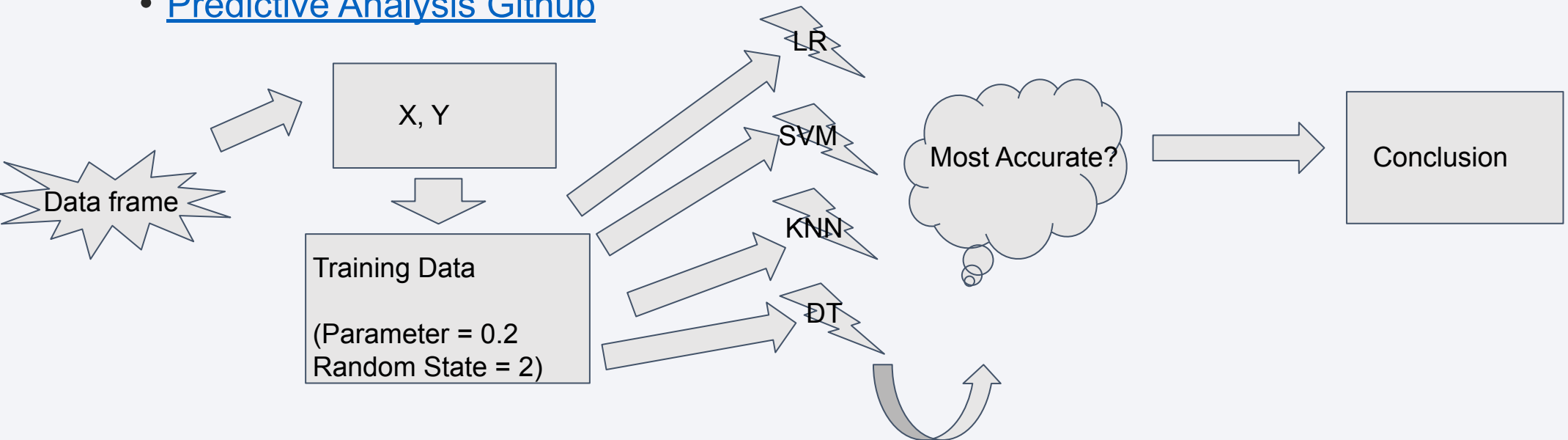
- Interactive Dashboard, so you can choose different launch sites (or all of them combined) and payload masses

- Visuals such as a pie chart to see the amount of launches per site and a scatter plot which changes with the payload mass to observe how mass changes which booster is used

- Plotly Dashboard Github

Predictive Analysis (Classification)

- Started with creating an array with NumPy using the column Class from the data frame to get y, then standardized and transformed to get x
- Obtained training and test data with 0.2 for parameter and random state to 2, ending with 18 test samples
- Proceeded to evaluate using Logistic Regression, Support Vector Machine, KNN and Decision Trees and their respective accuracy and confusion matrices, allowing to pick the best model
- [Predictive Analysis Github]

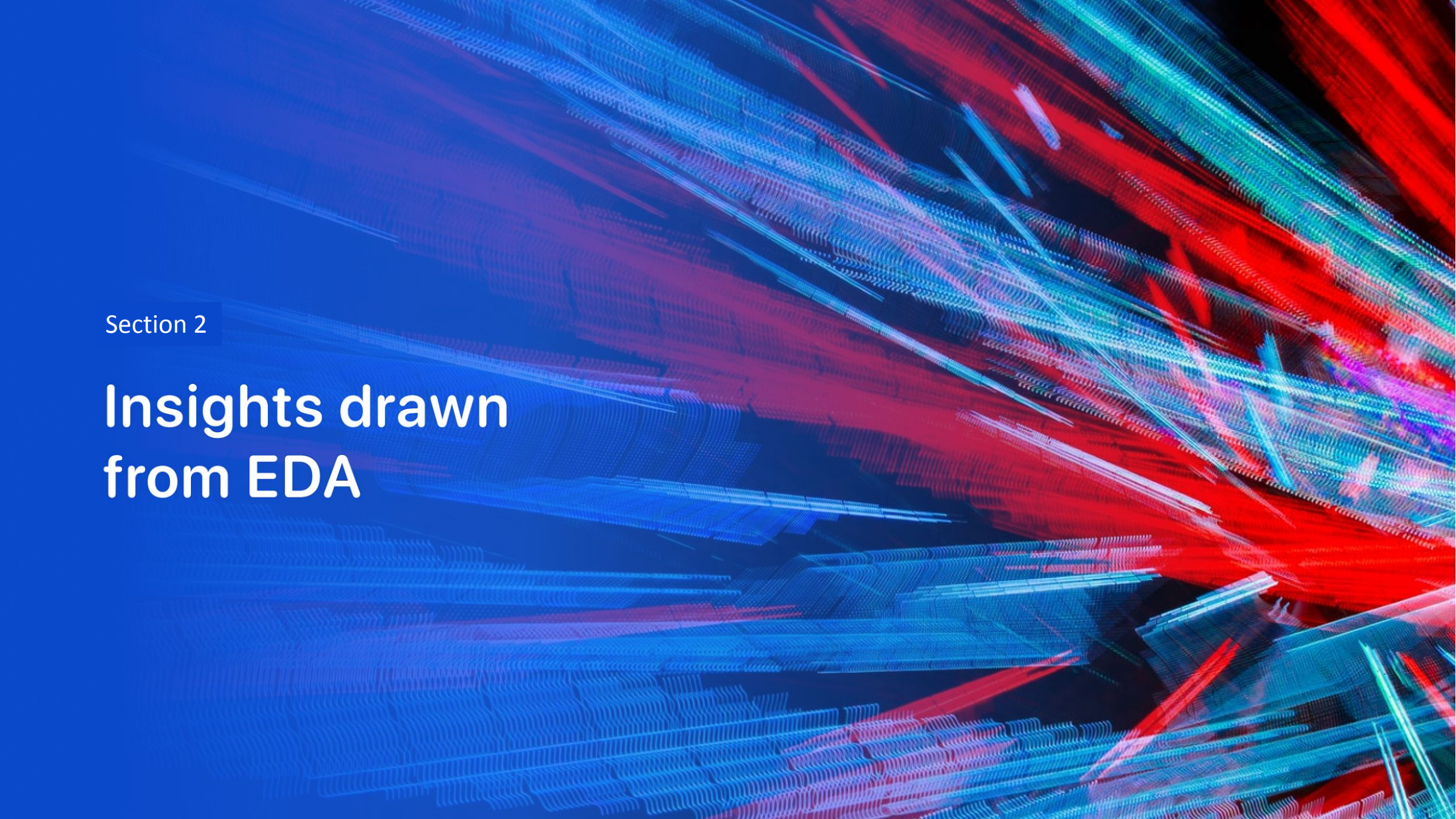


Results

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

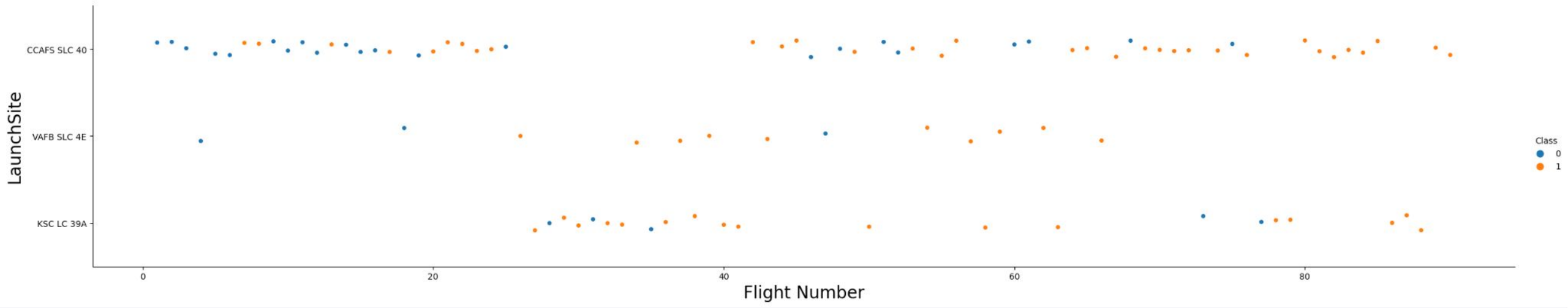
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

With blue dots being failed landings, and orange dots being successful landings, below we see a trend of increasing success rate with higher flight numbers across all launch sites:

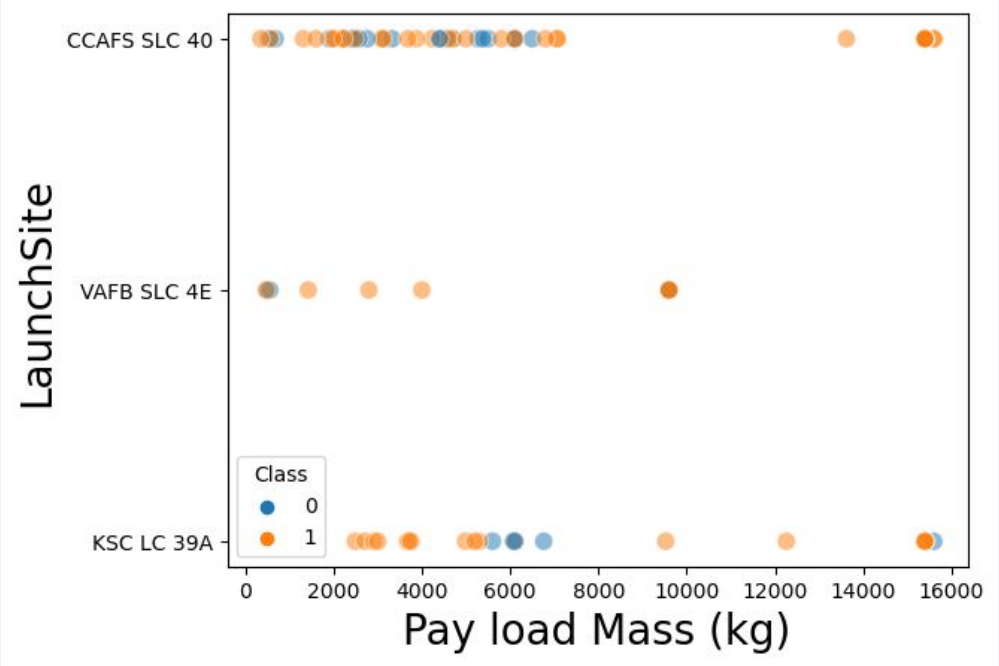


Payload vs. Launch Site

In this scatter plot, which uses the same blue/orange color key as the previous one, CCAFS SLC 40 shows to be used more than any other launch site, regardless of payload mass – however, with a payload gap between 8000 Kg and 14000 Kg with no launches recorded;

Coming in second as far as usage is concerned, the KSL LC 39A also displays a better payload spread;

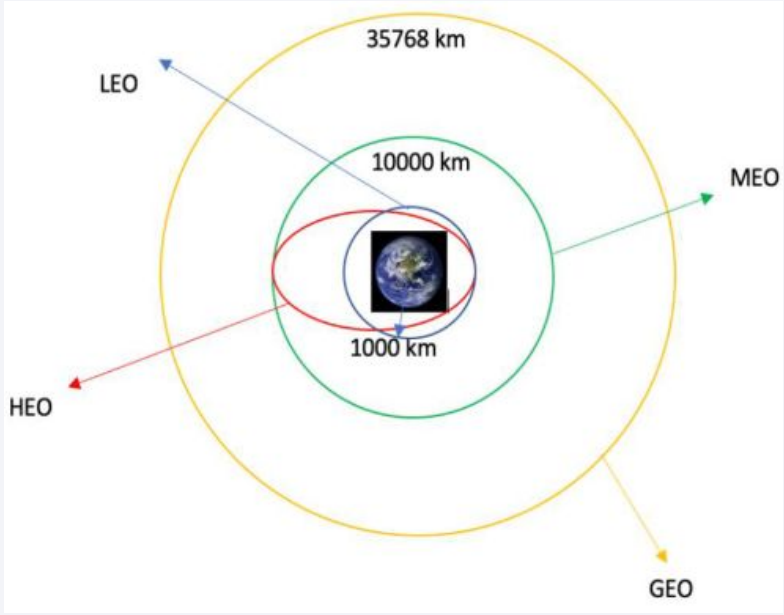
Lastly, VAFB SLC 4E is at last place for usage, with even greater payload gaps between 5000 Kg and 10000 Kg, and over 11000 Kg, approximately.

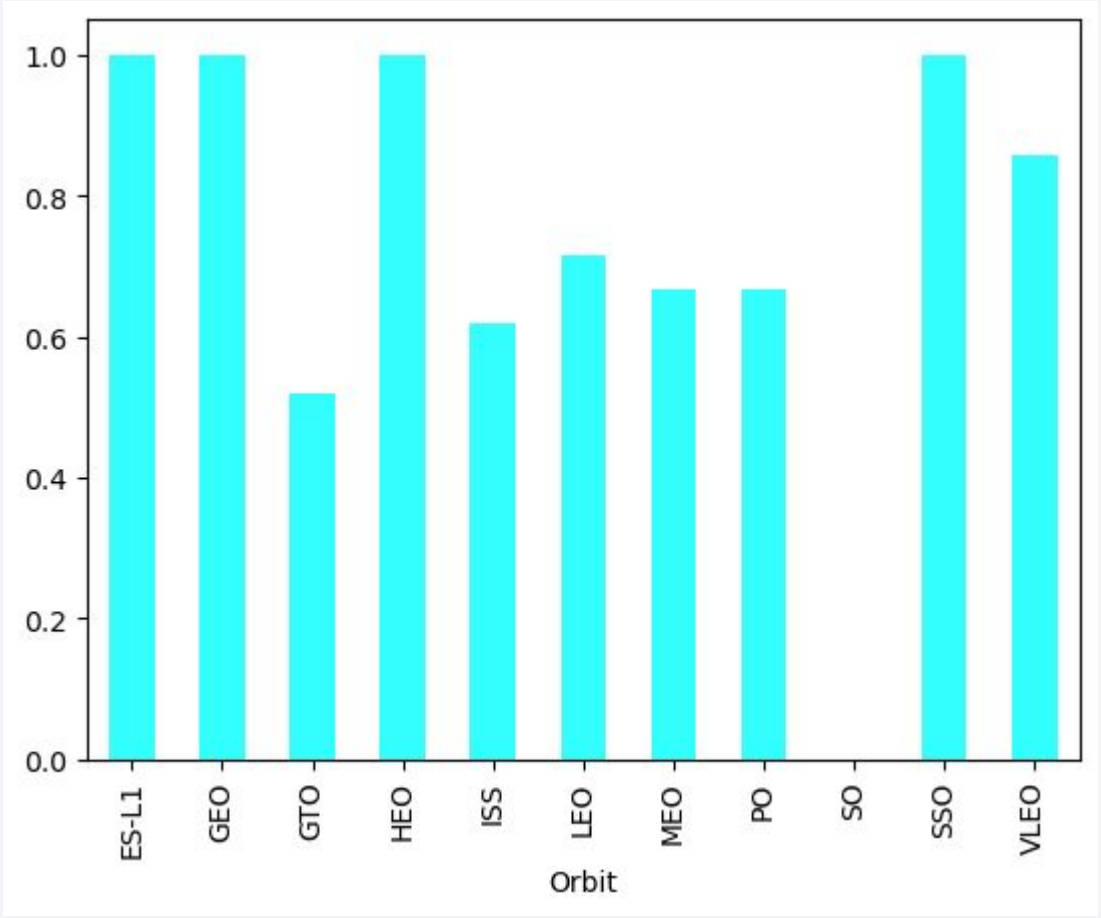


Success Rate vs. Orbit Type

The bar graph to the right shows the success rate per Orbit type (although some soccer fans would argue it's Argentina's World Cup result);

Standing out with the best success rates, from left to right, we have ES-L1, GEO, HEO and SSO (see graph below for a few orbit distances for reference)



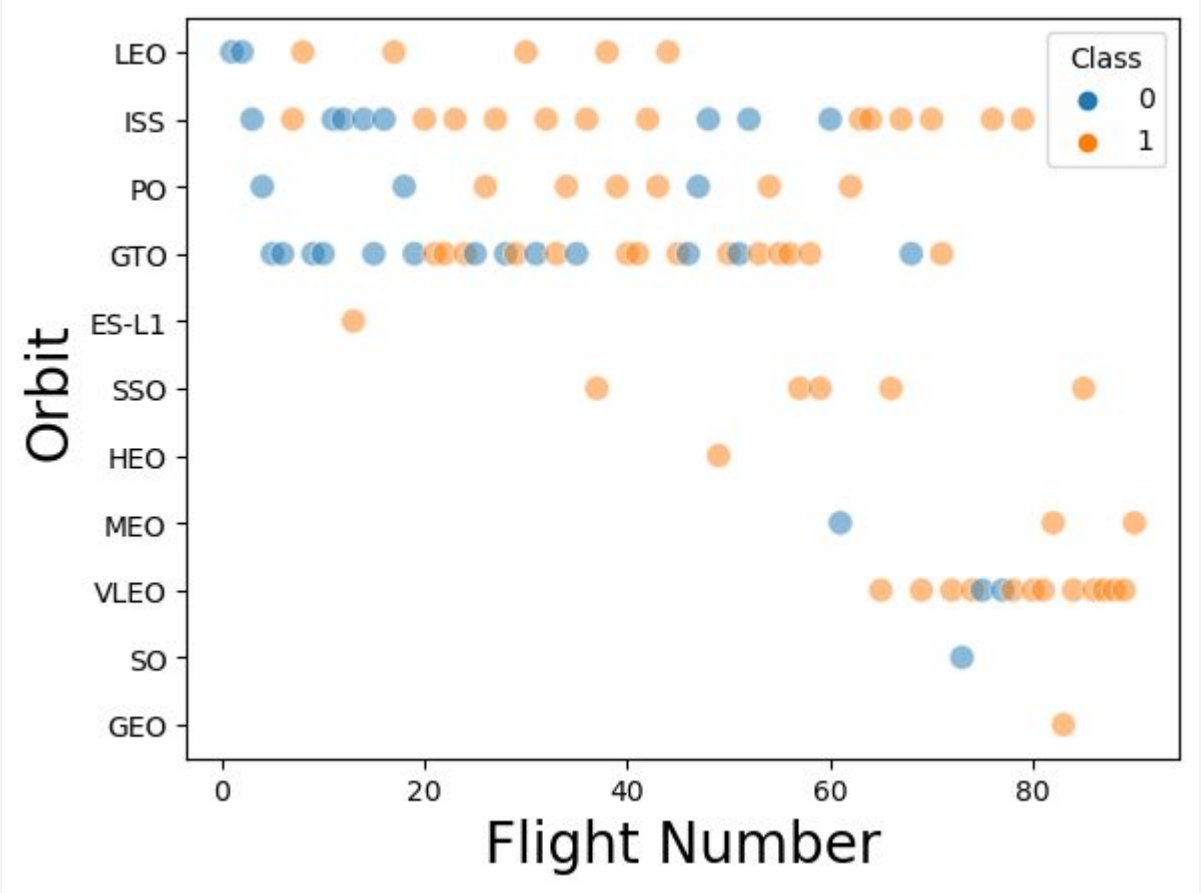


Flight Number vs. Orbit Type

ın this graph, using the same color key as before, we can see the same increase in success rate as the flight number increases;

However, there seems to not be a correlation between orbit types and success rate

Moreover, GTO is a commonly used orbit, since the payload is aligned with the equator



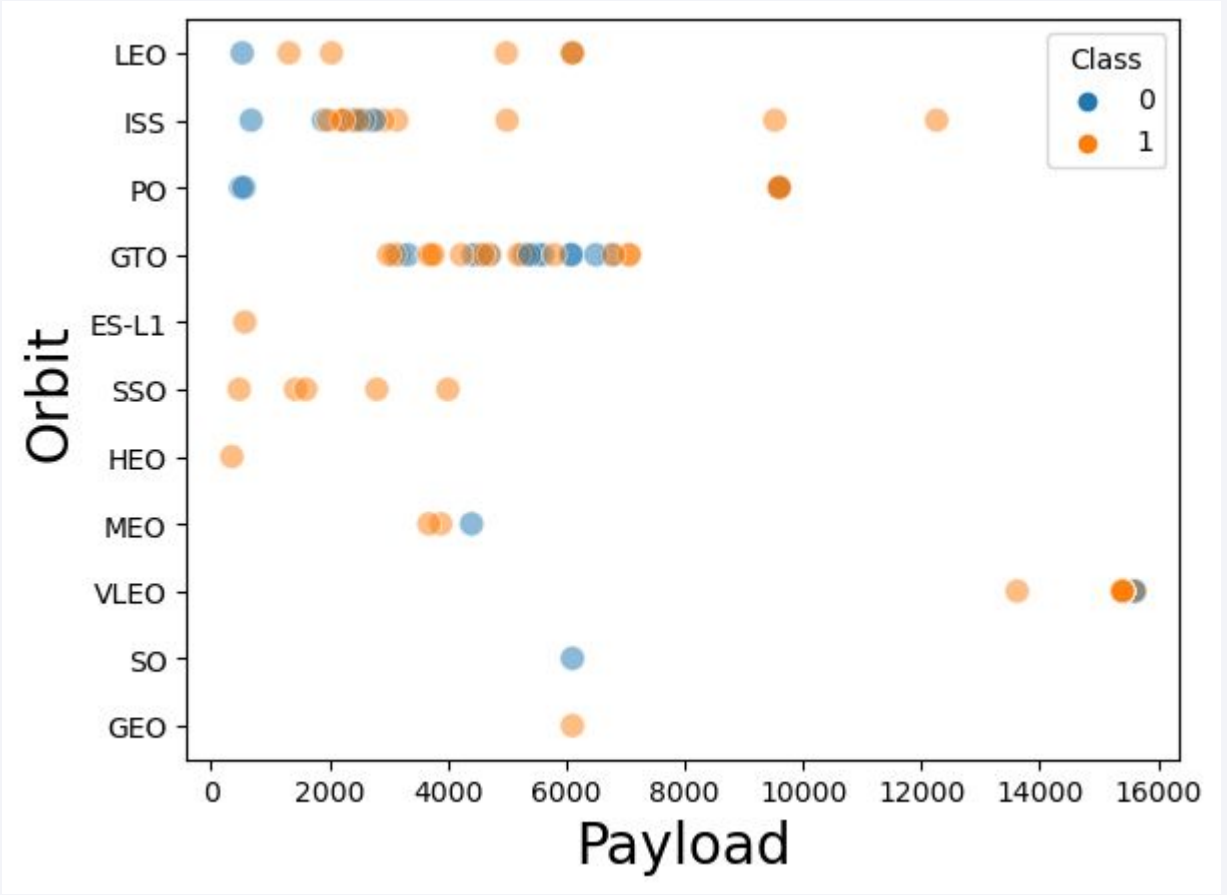
Payload vs. Orbit Type

Again, with the same color key, it is noticeable that GTO has a well-defined payload range between just below 4000 Kg and just below 8000 Kg;

SSO has a few, but all success landings with payloads of less than 6000 Kg;

ISS also displays a good cluster between 2000 and 4000 Kg with mixed outcomes;

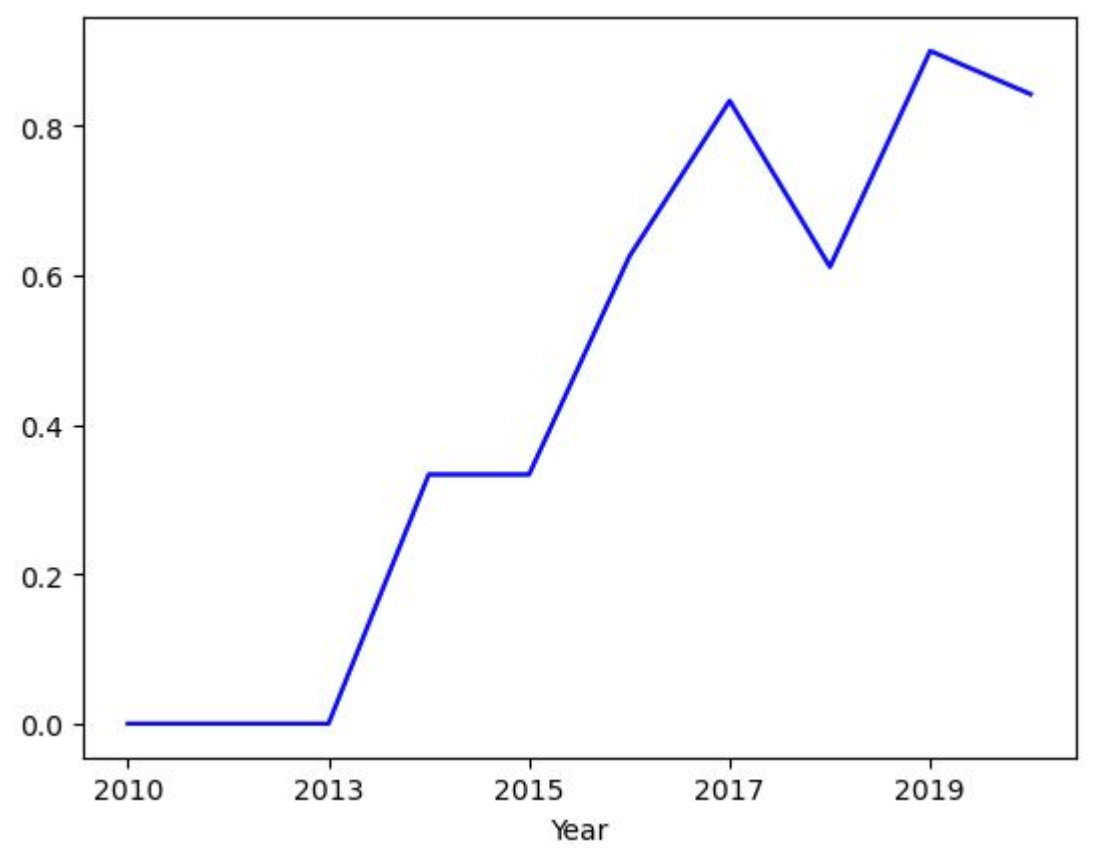
As a whole, we can infer from this graph that heavier payloads are fewer, but with a higher success rate.



Launch Success Yearly Trend

In the line graph to the right, it's clear that the success rate is trading in a positive way.

Considering the payload mass, launch site, weather, other factors, and the fact that they are reusable rockets, one can say it is a successful technology.



All Launch Site Names

- *Task: Find the names of the unique launch sites*

- The DISTINCT command was used to ensure only unique values would be called

```
1 %%sql
2 SELECT DISTINCT "Launch_Site"
3 FROM "SPACEXTBL"
```

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

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

Launch Site Names Begin with 'CCA'

- *Task: Find 5 records where launch sites begin with `CCA`*
- Since there are several rows, the command LIMIT 5 was utilized to only show 5 records

```
%%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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- *Task: Calculate the total payload carried by boosters from NASA*
- Utilized the command SUM() to call the total payload

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total Payload Mass Carried by Boosters by NASA (CRS)'
FROM SPACEXTBL
WHERE "Customer" IS "NASA (CRS)"
```

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

Total Payload Mass Carried by Boosters by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- *Task: Calculate the average payload mass carried by booster version F9 v1.1*
- Utilized the AVG() command to call the average

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS 'Total Payload Mass Carried by F9 v1.1 Boosters'
FROM SPACEXTBL
WHERE "BOOSTER_VERSION" IS "F9 v1.1"
```

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

Total Payload Mass Carried by F9 v1.1 Boosters
2928.4

First Successful Ground Landing Date

- *Task: Find the dates of the first successful landing outcome on ground pad*
- Utilized the MIN() command to call the first successful landing and filtered using “ground pad” to avoid calling the very successful landing (which was not on ground pad)

```
%%sql
SELECT MIN(DATE) AS "FIRST SUCCESSFUL LANDING OUTCOME", *
FROM SPACEXTBL
WHERE "Landing _Outcome" IS "Success (ground pad)"

```

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

FIRST SUCCESSFUL LANDING OUTCOME	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
01-05-2017	01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- *Task: List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 Kg*
- Utilized the command BETWEEN AND to call the needed mass

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION), "Landing _Outcome", "PAYLOAD_MASS__KG_"
FROM SPACEXTBL
WHERE "Landing _Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- *Task: Calculate the total number of successful and failure mission outcomes*
- Utilized the commands WHERE and LIKE to filter only successful missions

```
%%sql
SELECT COUNT("Landing _Outcome") AS SUCCESSFUL_MISSIONS
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Success%'
```

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

SUCCESSFUL_MISSIONS
61

```
%%sql
SELECT COUNT("Landing _Outcome") AS FAILED_MISSIONS
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Failure%'
```

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

FAILED_MISSIONS
10

Boosters Carried Maximum Payload

- *Task: List the names of the booster which have carried the maximum payload mass*
- Utilized subquery method in order to use the command MAX() to get the maximum payload and the booster version
- (SQL part not included to reduce visual noise)

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

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

- Utilized the command substr(Date, n, n) in order to call the month

```
%%sql
SELECT "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE,  substr(Date, 4, 2) AS Month
FROM SPACEXTBL
WHERE "Landing _Outcome" = "Failure (drone ship)" AND substr(Date,7,4) = "2015"
```

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

Landing _Outcome	Booster_Version	Launch_Site	Month
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04

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

- *Task: 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*
- Utilized commands such as COUNT, WHERE, LIKE, BETWEEN, GROUP BY and ORDER BY in order to filter and group to reach the desired results

```
%%sql
SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS COUNT
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Success%' AND "DATE" BETWEEN
GROUP BY "Landing _Outcome"
ORDER BY COUNT DESC
```

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

Landing _Outcome	COUNT
Success	38
Success (drone ship)	14
Success (ground pad)	9


Section 3
Launch Sites
Proximities Analysis

Falcon 9 SpaceX Launch Sites

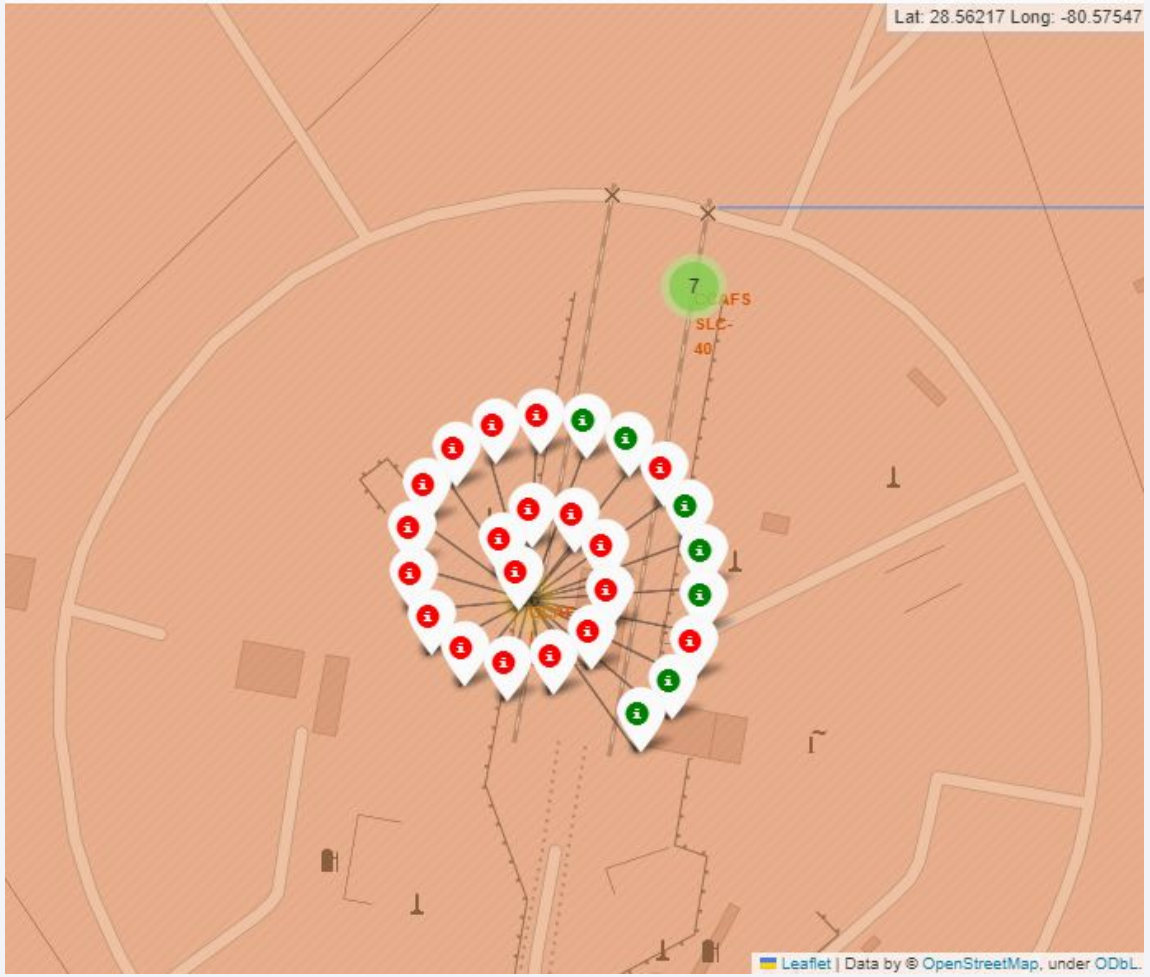
So far, due to physics and cost, SpaceX has 2 main locations for their Falcon 9 launch sites: California and Florida, as shown in red on the map to the right.

Folium color-coded expandable marker clusters

On the right, we have the Florida launch site. At the top, with a green circle and the number 7 inside, is CCAFS SLC-40. If you were to click on that number, it would open up its marker cluster, as shown at the bottom, at CCAFS LC-40.

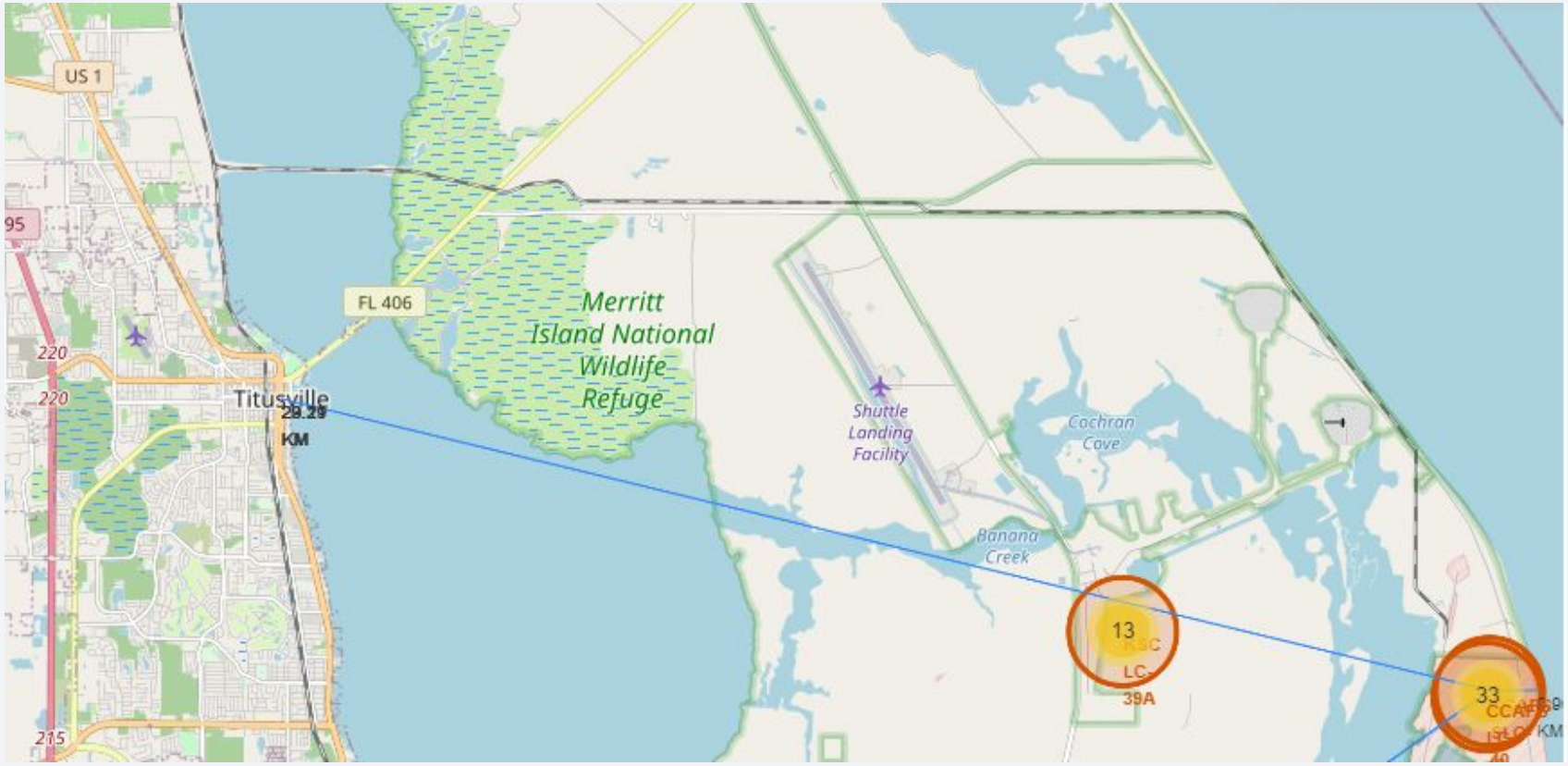
Green marks are for successful landings

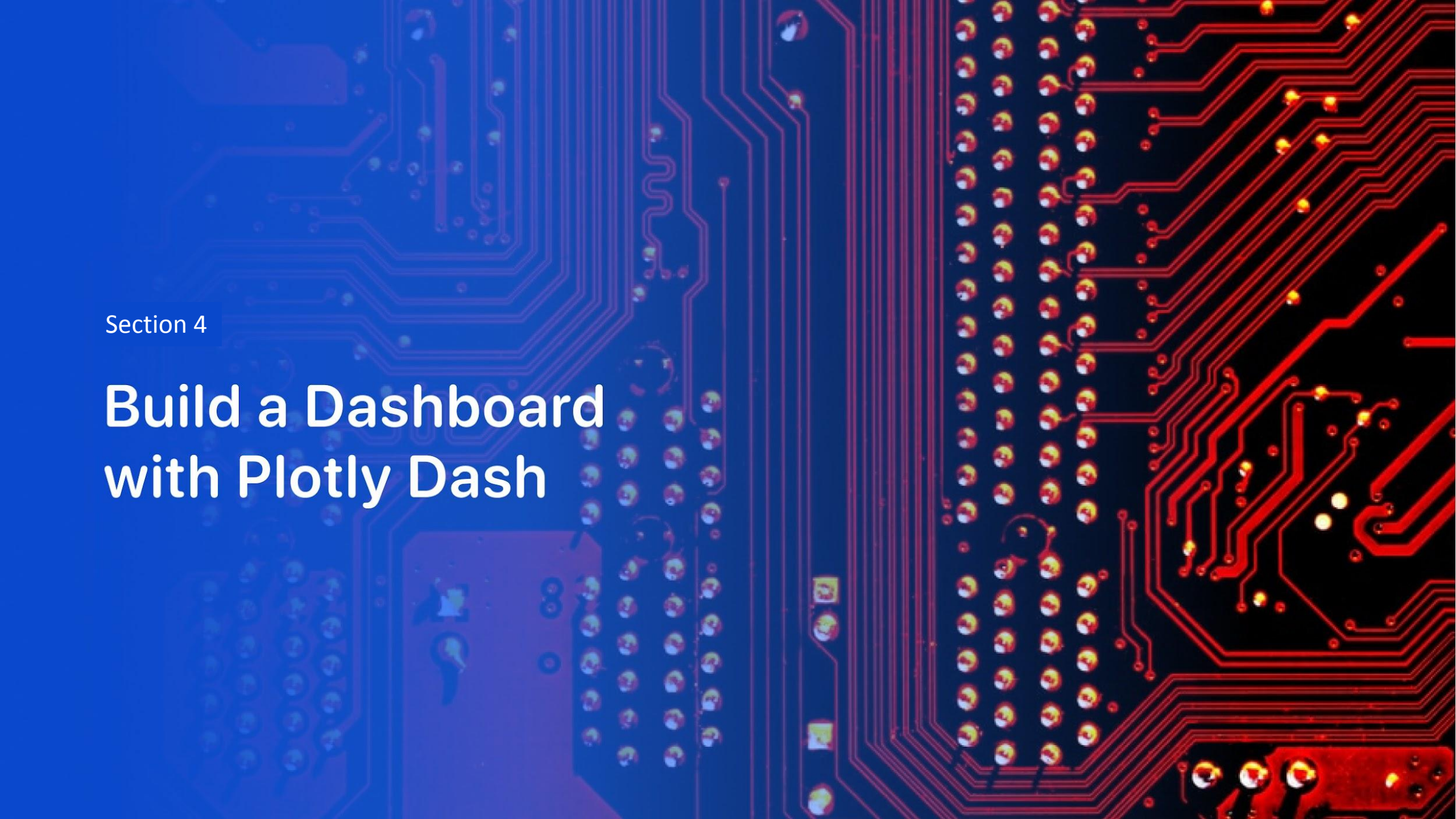
Red marks are for failed landings



Folium Distance Plotting

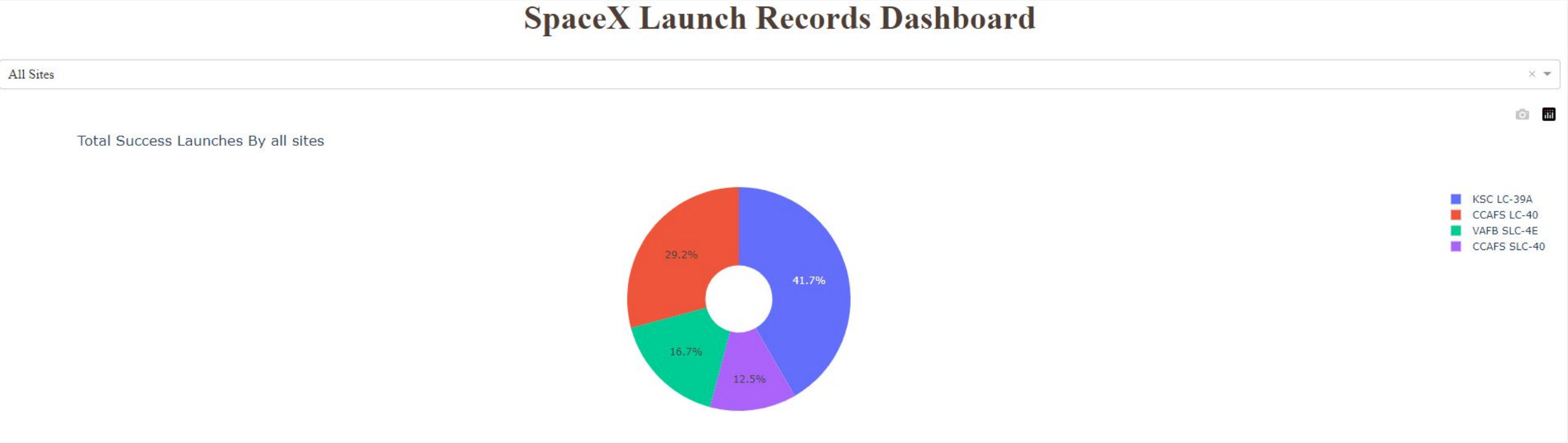
Below is an example of Folium's ability to plot a line using coordinates. In this example, a line from CCAFS, the busiest launch site in our analysis, is connected to its closest city, Titusville. With a population of about 50,000, it is only 29 KM, or 18 miles away from the launch site.




Section 4
Build a Dashboard with Plotly Dash

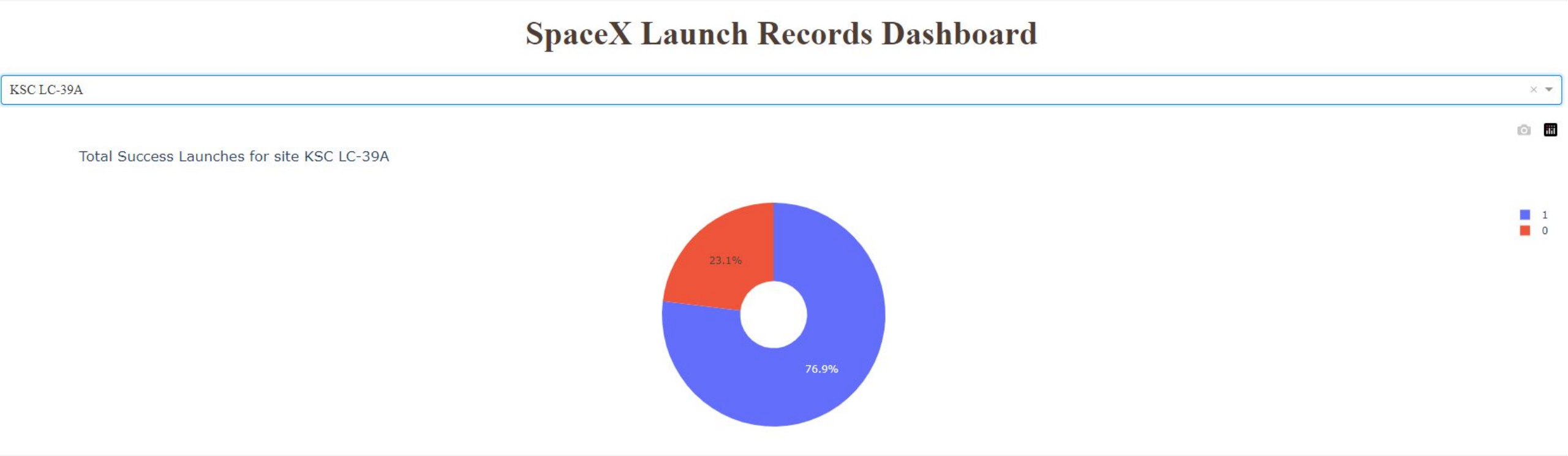
Plotly Dash: Success Count for All Sites

- A pie chart, below, giving a sense of the amount of successful launches by launch site, with KSC LC-39A leads with 41.7% success rate.



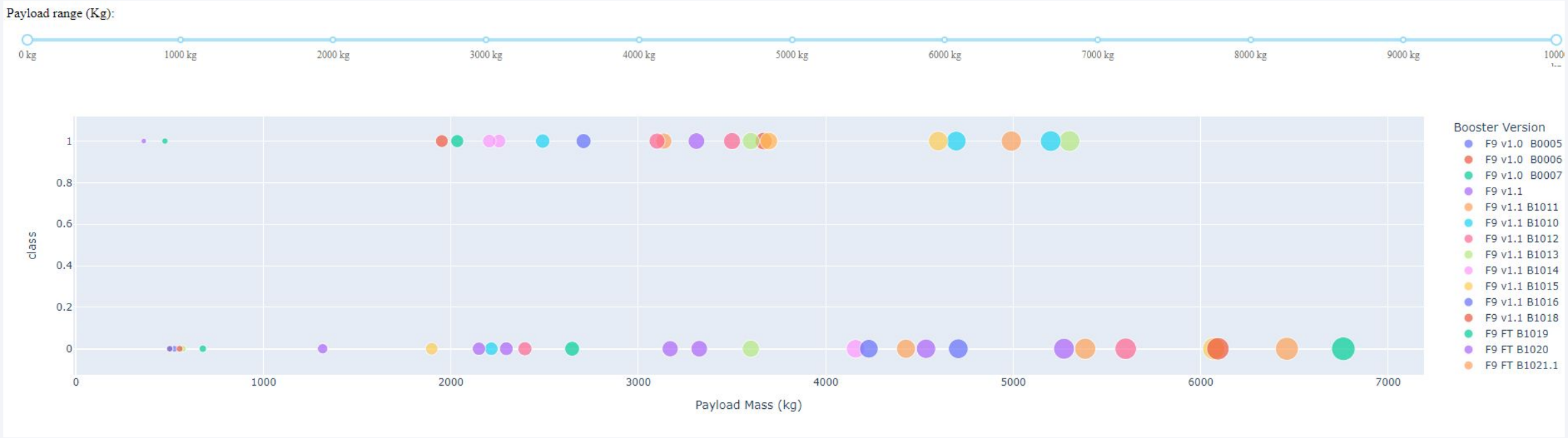
Plotly Dash: Highest Launch Success Ratio

- We learned from the previous slide that KSC LC-39A leads with 41.7% success rate over all other sites. However, by clicking on the interactive menu and choosing this specific site, it reveals its own success ratio of 76.9%:



Plotly Dash: All Sites Payload vs Launch Outcome

Below, at the bottom section of the page, we can use the slider to choose the Payload range in Kg (here, the full spectrum has been selected) and its historical data on success rate, denoted with a “1” and failure as “0”. Note that most the high success rate is in between 2000 and 6000 Kg.

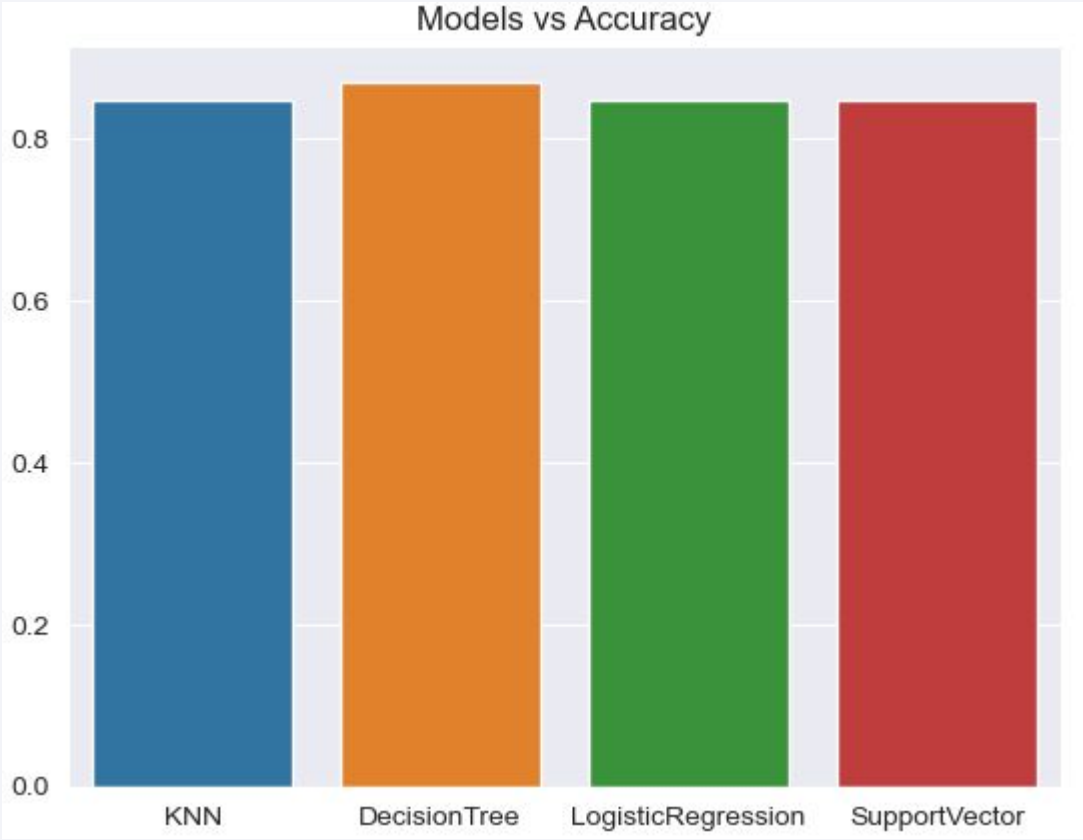


Section 5

Predictive Analysis (Classification)

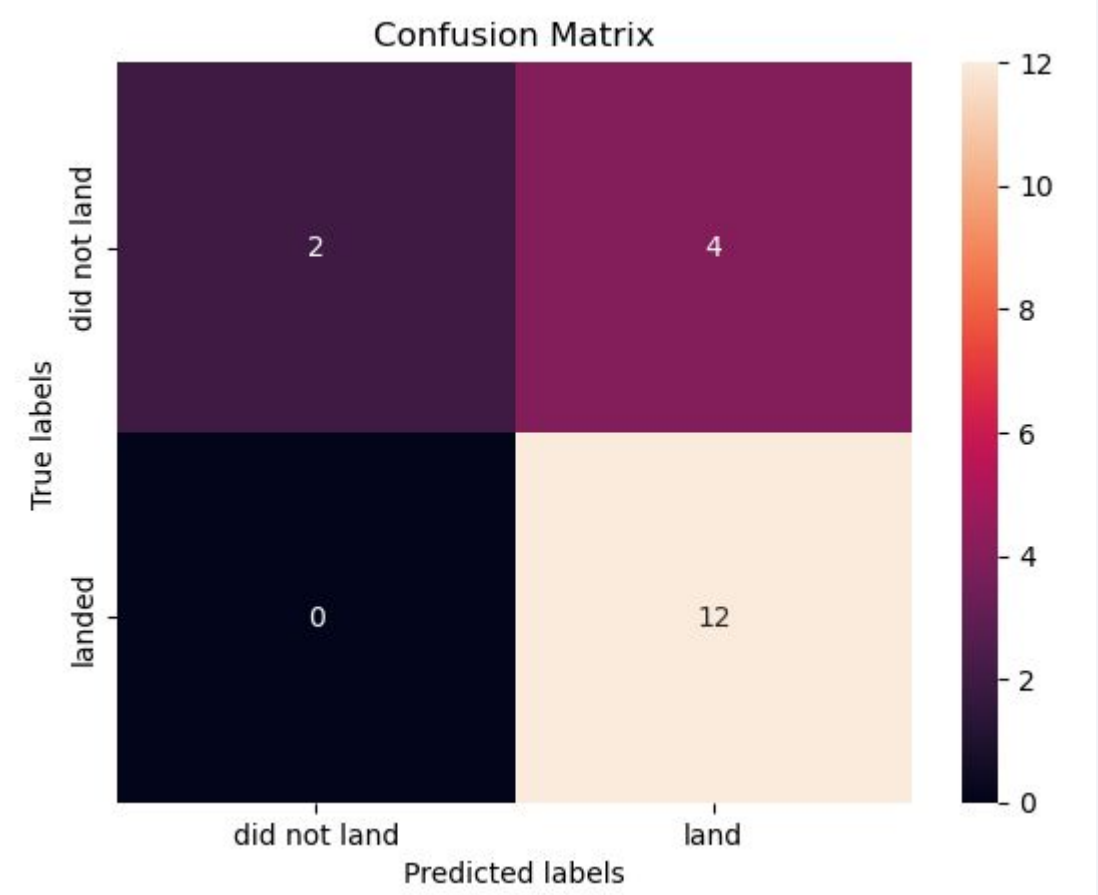
Classification Accuracy

Although all models were very close in accuracy, the *Decision Tree* model came out slightly ahead, as the bar graph on the right shows.



Confusion Matrix

In this Confusion Matrix diagram on the right we can see a pattern among all models: a very high False Positive of 12. However, in all other models, the True Positives at 3, just below the Decision Tree score of 4. To top things off, the True Negative is 2 here, but all others were 3. With all that said, even though the accuracy was slightly higher, the Decision Tree is indisputably the model of choice for this project.



Conclusions

- SpaceX data is readily and openly accessible, with a plethora of details to gather data from;
- Most the high success rate missions had a payload in between 2000 and 6000 Kg;
- All analyzed launch sites are at a good distance from major cities
- The SpaceX Falcon 9 missions, in general, have been a success regardless of payload mass, orbit, or landing site (yes, even drone ships in the middle of the ocean)
- The fact that the reusable rocket has a significant reduced cost compared to competitors, and such high success rate despite the factors mentioned above, it would be difficult to argue against SpaceX's case
- Among K-Nearest Neighbors, Support Vector Machine, Logistic Regression and Decision Tree models, the latter came out ahead with a 0.85 score
- With all points above considered, we can conclude that the Falcon 9 rocket will land. Again.

Thank you!