IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohsin Khan
4 November 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive analytics Demo in Screenshots
  - Predictive Analysis Results

Introduction

Project background and context

- The era of commercial space has arrived
- Several companies are making space travel affordable for everyone
- SpaceX is the most successful of them

- their rocket launch is relatively inexpensive
- SpaceX advertises Falcon 9 rocket launches, with a cost of 62 million dollars.
- It's because SpaceX can reuse the first stage
- Therefore, we will predict if the Falcon 9 first stage will land successfully

Problems you want to find answers

- Correlations between each rocket variables and successful landing rate
- Conditions to get the best results and ensure the best successful landing rate

Section 1

Methodology

Methodology

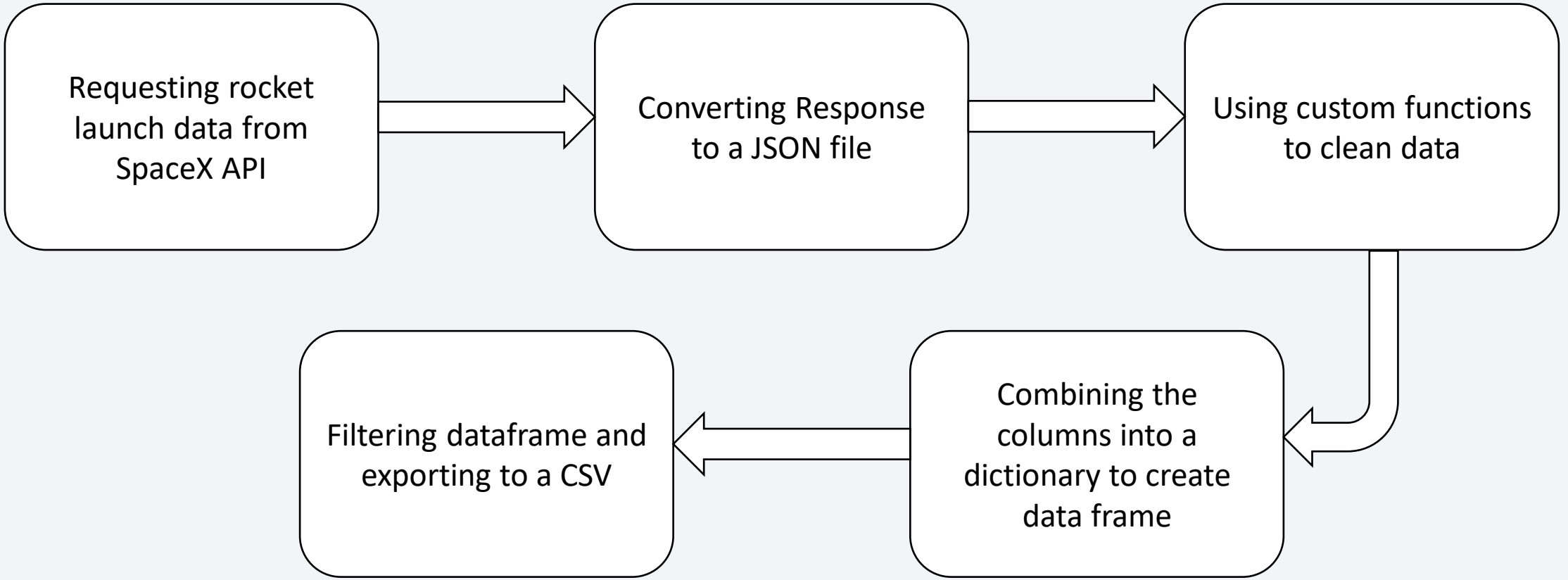
Executive Summary

- Data collection methodology:
  - SpaceX API & Web Scraping SpaceX Wikipedia Page
- 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
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

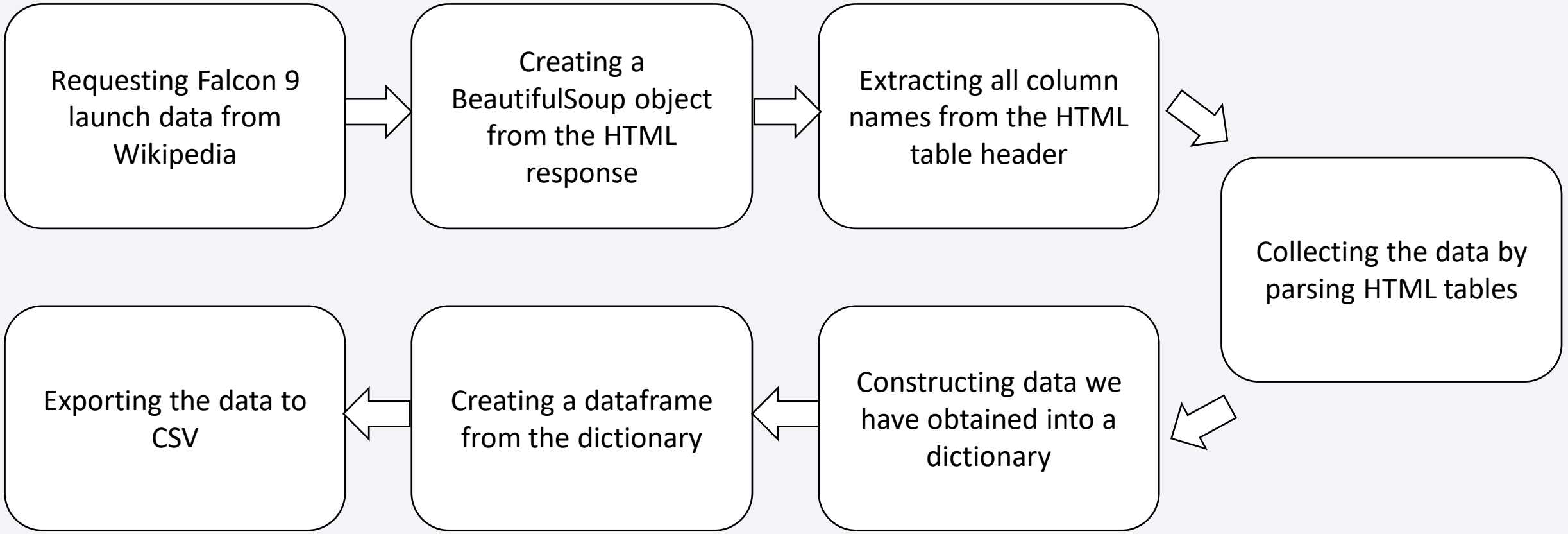
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia.
- Data Columns are obtained by using SpaceX REST API:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL

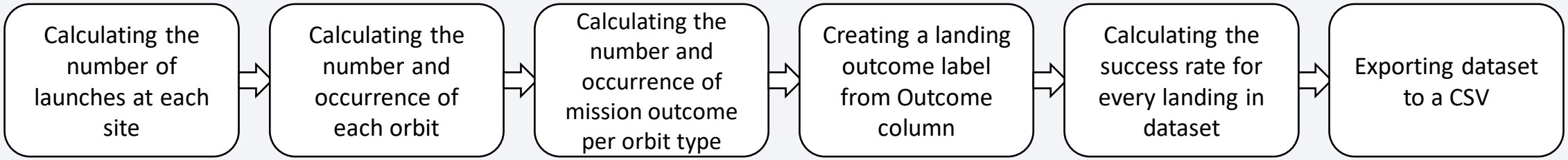
Data Collection - Scraping



GitHub URL

Data Wrangling

- There are several cases in which the booster failed to successfully land on the dataset, and sometimes it attempted to land but failed because of accident.
- Converting the results into training labels:
  - 1 = successful / 0 = failure



GitHub URL

EDA with Data Visualization

Scatter chart:

- Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, Payload vs. Orbit Type.
- A scatter plot shows how much one variable is affected by another. The relationship between two variables is called a correlation. This plot is generally composed of large data bodies.

Bar chart:

- Orbit Type vs. Success Rate.
- A Bar chart makes it easy to compare datasets between multiple groups at a glance. One axis represents a category and the other axis represents a discrete value. The purpose of this chart is to indicate the relationship between the two axes

Line chart:

- Year vs. Success Rate
- A Line chart shows data variables and trends very clearly and helps predict the results of data that has not yet been recorded

GitHub URL

EDA with SQL

- Performed SQL queries:
  - 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
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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 URL

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest Cit

GitHub URL

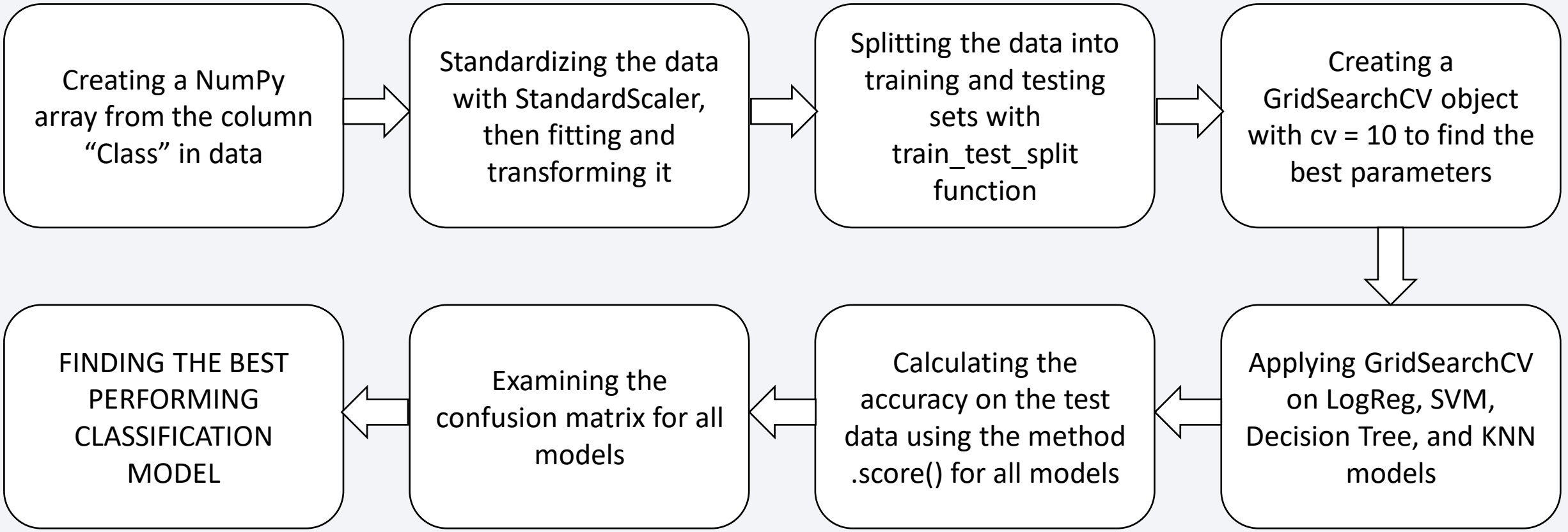
Build a Dashboard with Plotly Dash

The dashboard application contains a pie chart and a scatter point chart.

- Pie chart
  - For be selected to indicate a successful landing distribution across all launch sites or to indshowing total success launches by sites
  - This chart can icate the success rate of individual launch sites.
- Scatter chart
  - For showing the relationship between Outcomes and Payload mass(Kg) by different boosters
  - Has 2 inputs: All sites/individual site & Payload mass on a slider between 0 and 10000 kg
  - This chart helps determine how success depends on the launch point, payload mass, and booster version categories

GitHub URL

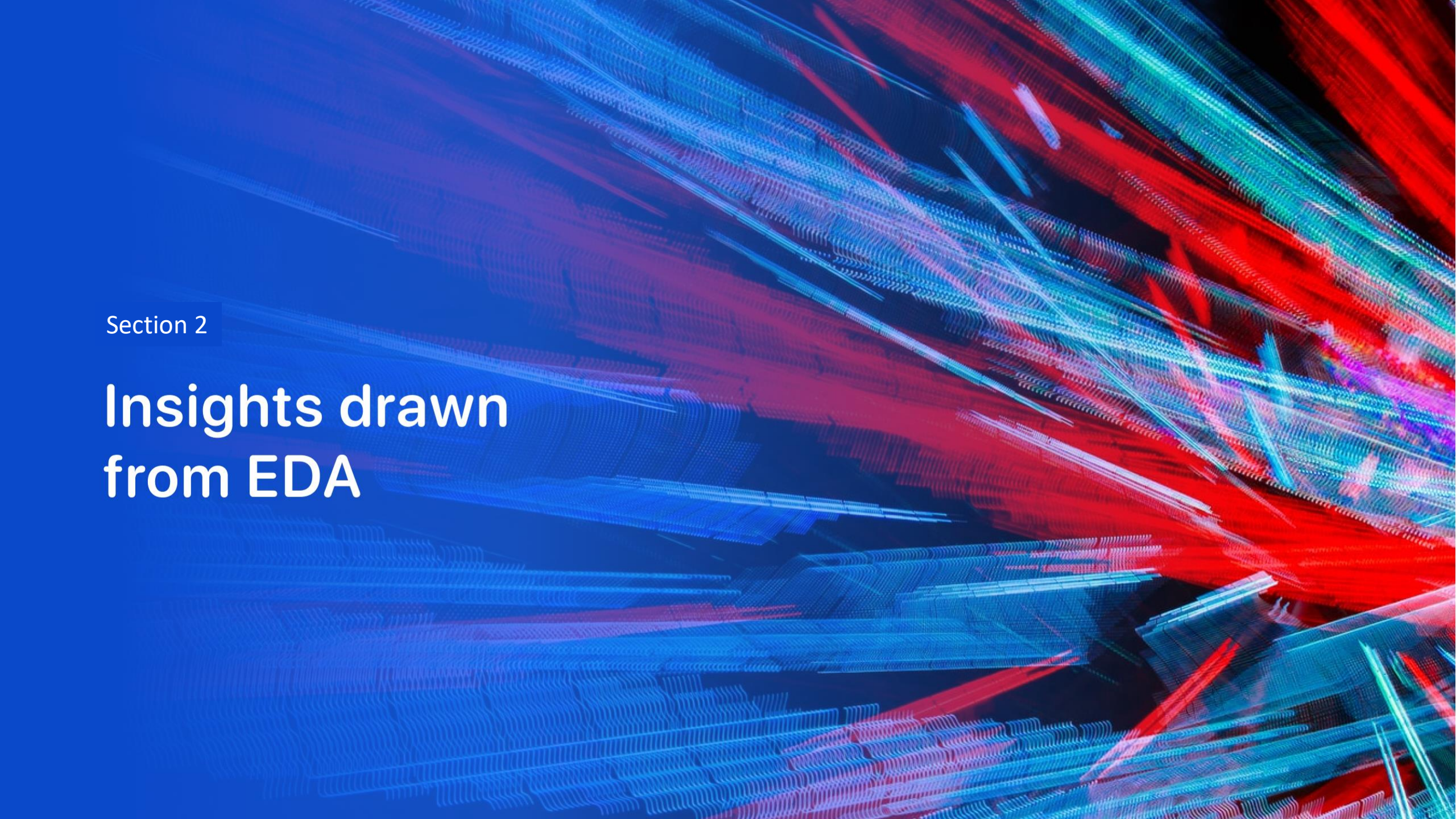
Predictive Analysis (Classification)



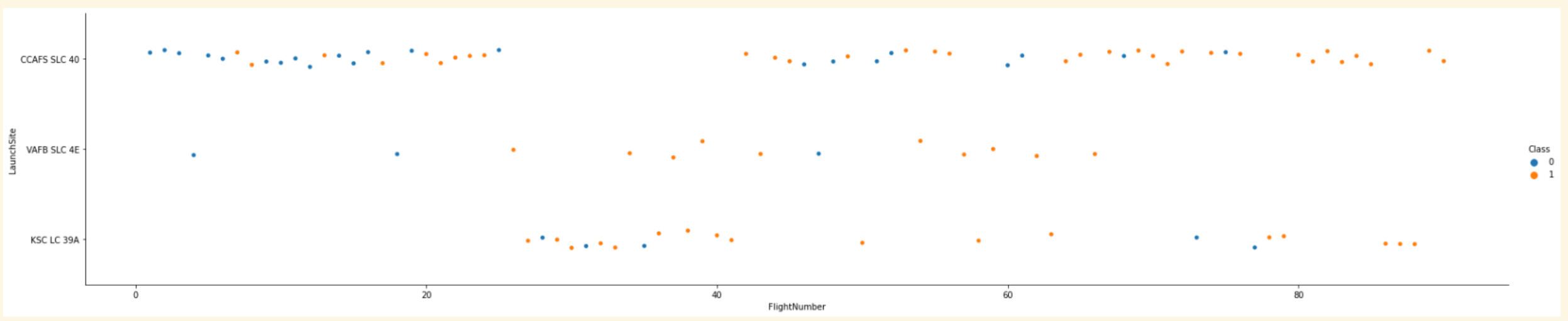
GitHub URL

Results

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

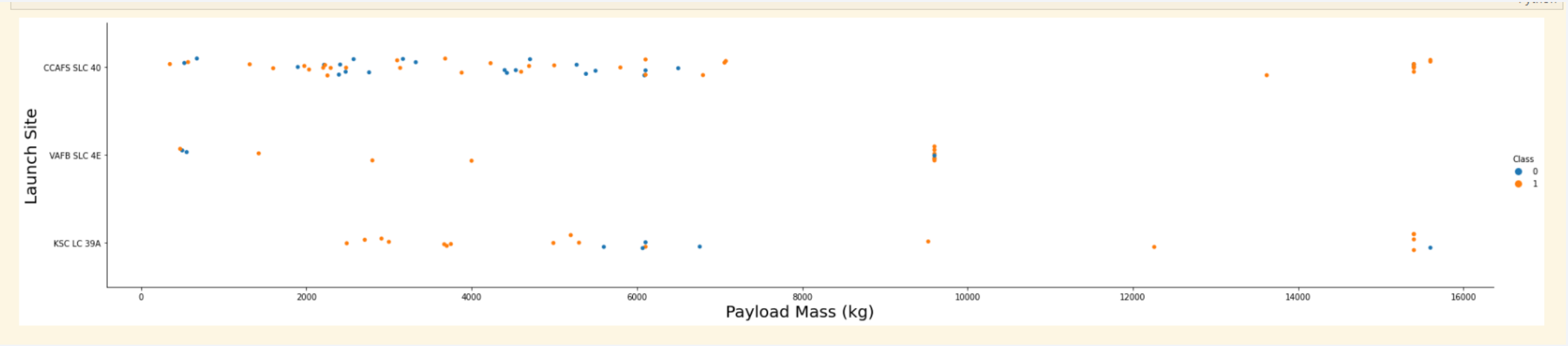

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



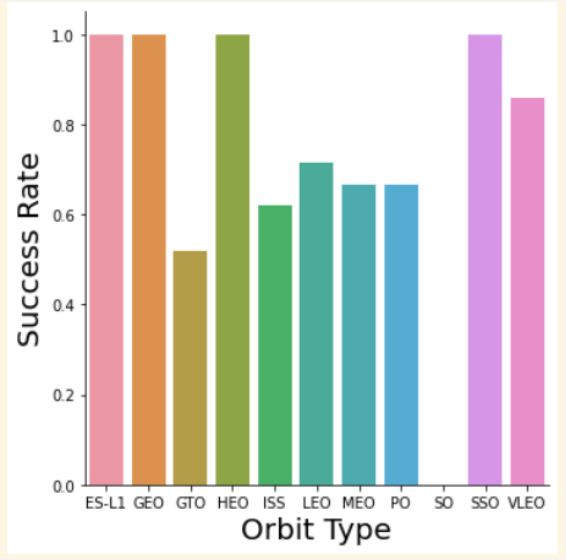
Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



For every launch site the higher the payload mass, the higher the success rate. Most of the launches with payload mass over 7000 kg were successful. KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

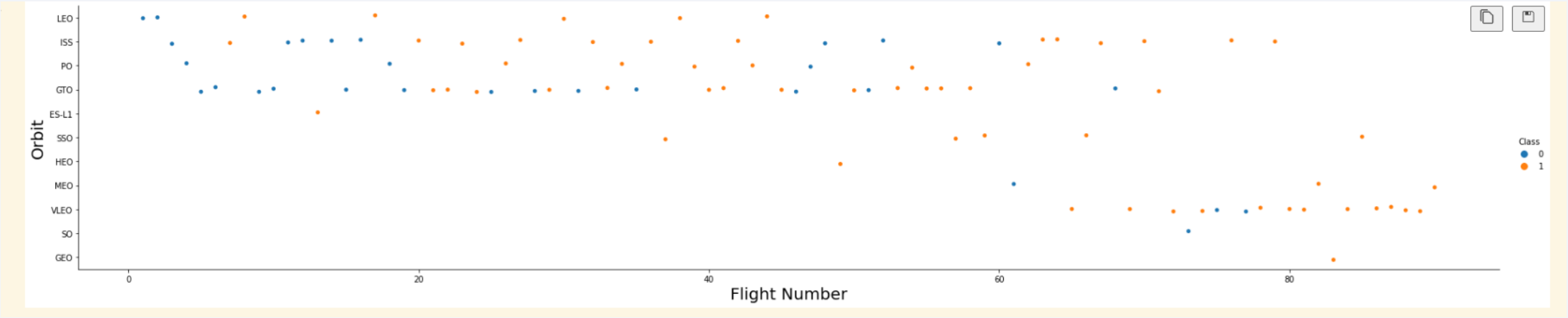
Success Rate vs. Orbit Type



Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%).

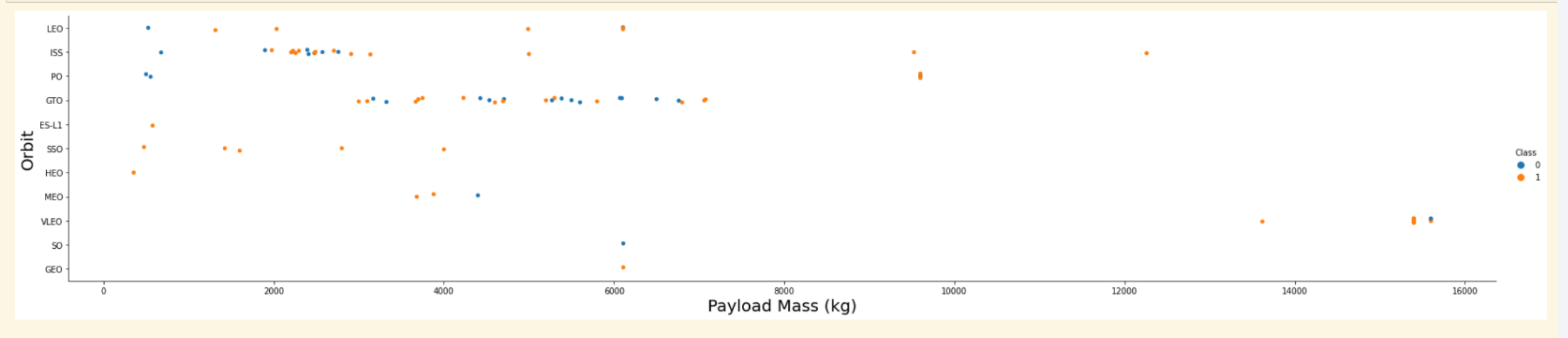
On the other hand, the success rate of orbit type GTO is only 50%, and it is the lowest except for type SO, which recorded failure in a single attempt

Flight Number vs. Orbit Type



Launch Orbit preferences changed over Flight Number.
Launch Outcome seems to correlate with this preference.
SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

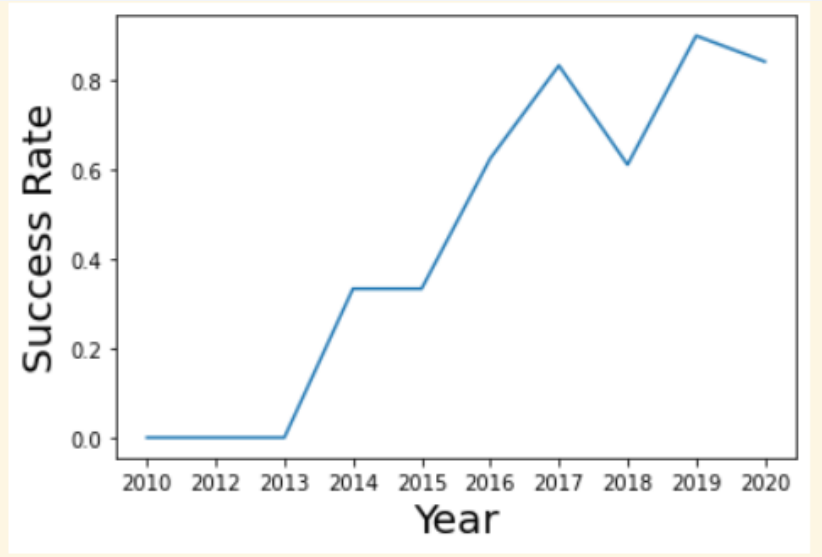


Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



Since 2013, the success rate has continued to increase until 2017.
The rate decreased slightly in 2018.
In 2020, it has shown a success rate of about 80%

Section 3

EDA with SQL

All Launch Site Names

```
%sql select distinct(LAUNCH_SITE) from SPACEXDATASET
```

[3]

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9
Done.
```

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

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

When the SQL DISTINCT clause is used in the query, only unique values are displayed in the Launch_Site column from the SpaceX table

Launch Site Names Begin with 'CCA'

```
[4]  %sql select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/BLUDB
Done.
```

DATE	TIME	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

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

Using the LIKE operator and the percent sign (%) together, the Launch_Site name starting with CAA could be called

Total Payload Mass

```
[6]  %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)'

...  * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.

</>  total_payload_mass
     45596
```

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

In the WHERE clause, filter the dataset to perform calculations only if Customer is NASA (CRS)

Average Payload Mass by F9 v1.1

```
[12]  %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/BL
Done.
```

average_payload_mass
2534

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

In the WHERE clause, filter the dataset to perform calculations only if Booster_version is F9 v1.1

First Successful Ground Landing Date

```
[13]  %sql select min(DATE) from SPACEXDATASET where landing_outcome = 'Success (ground pad)'

...   * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.datab
      Done.

</>           1
      2015-12-22
```

Listing the date when the first successful landing outcome in ground pad was achieved

In the WHERE clause, filter the dataset to perform a search only if Landing__outcome is Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[14]  %%sql select booster_version
      from SPACEXDATASET
      where landing_outcome = 'Success (drone ship)'
      and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-982
Done.
```

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

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

Using the AND operator to display a record if additional condition PAYLOAD_MASS__KG_ is between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
[15]  %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdom
Done.
```

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

Listing the total number of successful and failure mission outcomes

Using the GROUP BY statement, groups rows that have the same values into summary rows to find the total number in each Mission_outcome

Boosters Carried Maximum Payload

```
[16] %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/BLUDB
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

Listing the names of the booster versions which have carried the maximum payload mass

Using a subquery, first, find the maximum value of the payload by using MAX() function, and second, filter the dataset to perform a search if PAYLOAD_MASS__KG_ is the maximum value of the payload

2015 Launch Records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing_outcome from SPACEXDATASET
where landing_outcome = 'Failure (drone ship)' and year(date)=2015
```

[20]

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32
Done.
```

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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

Using the AND operator to display a record if additional condition YEAR is 2015

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

```
[21] %%sql select landing_outcome, count(*) as count_outcomes from SPACEXDATASET
        where date between '2010-06-04' and '2017-03-20'
        group by landing_outcome
        order by count_outcomes desc;
```

```
 * ibm_db_sa://zgq47707:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0l
Done.
```

landing_outcome	count_outcomes
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

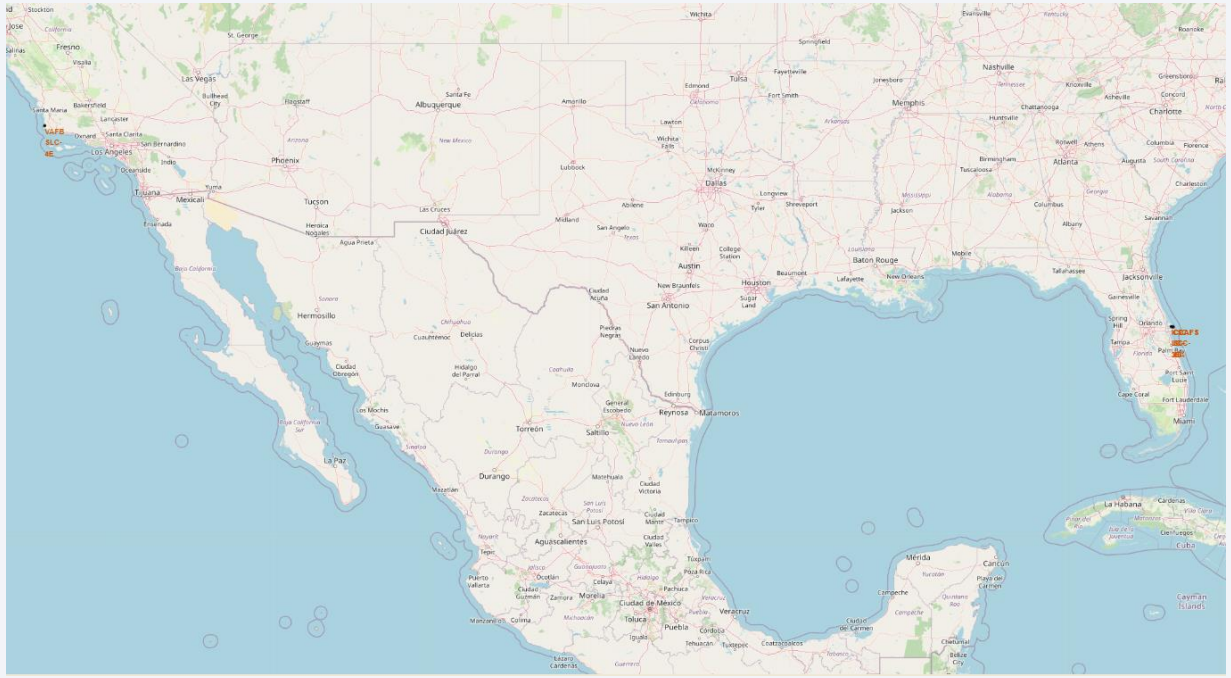
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

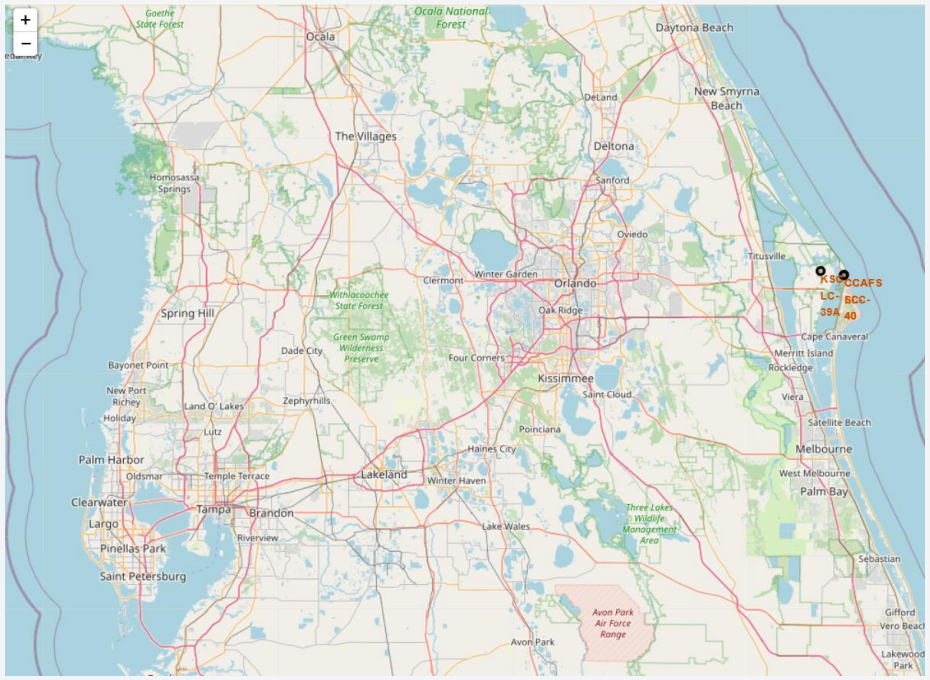
In the WHERE clause, filter the dataset to perform a search if the date is between 2010-06-04 and 2017-03-20.

Using the ORDER BY keyword to sort the records by total number of landing, and using DESC keyword to sort the records in descending order


Section 4
Launch Sites
Proximities Analysis

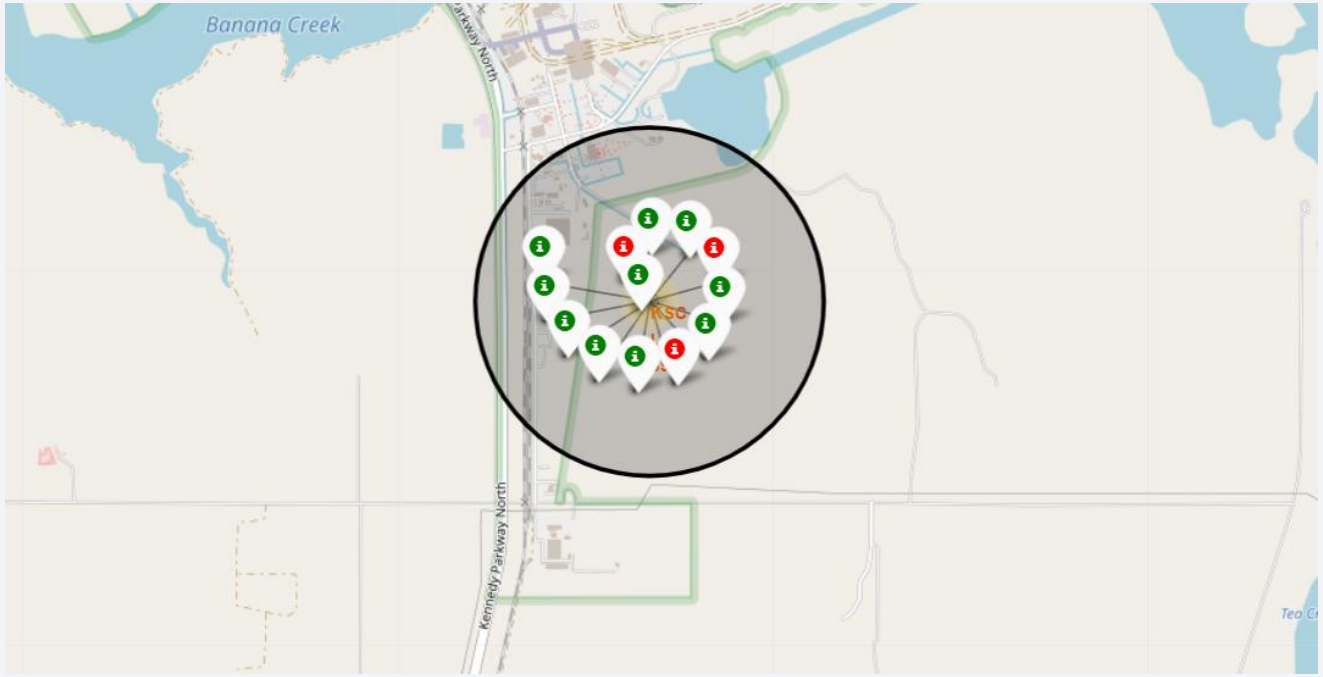
All launch sites' location markers on a global map





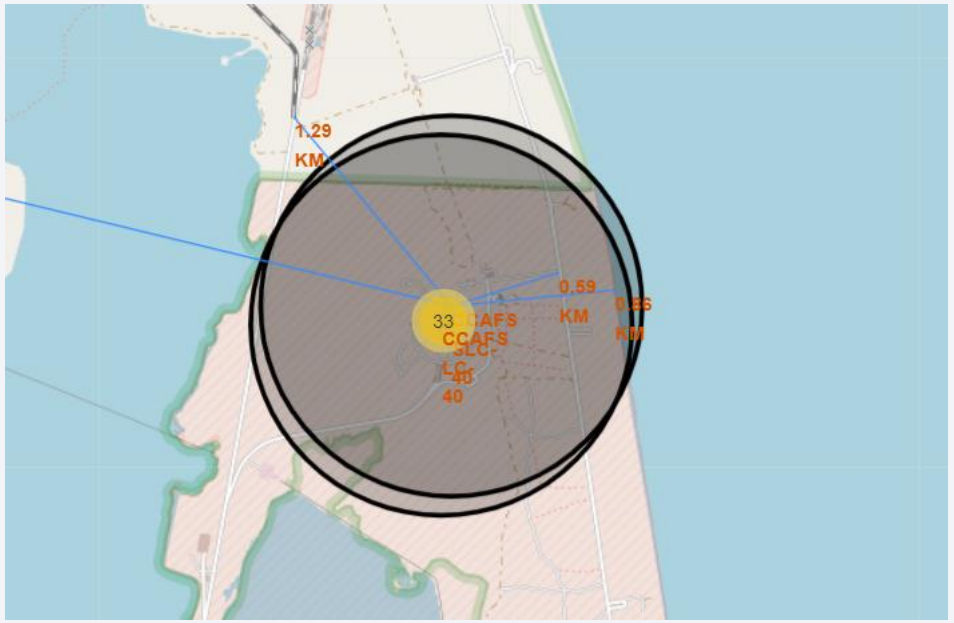
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean

Colour-labeled launch records on the map

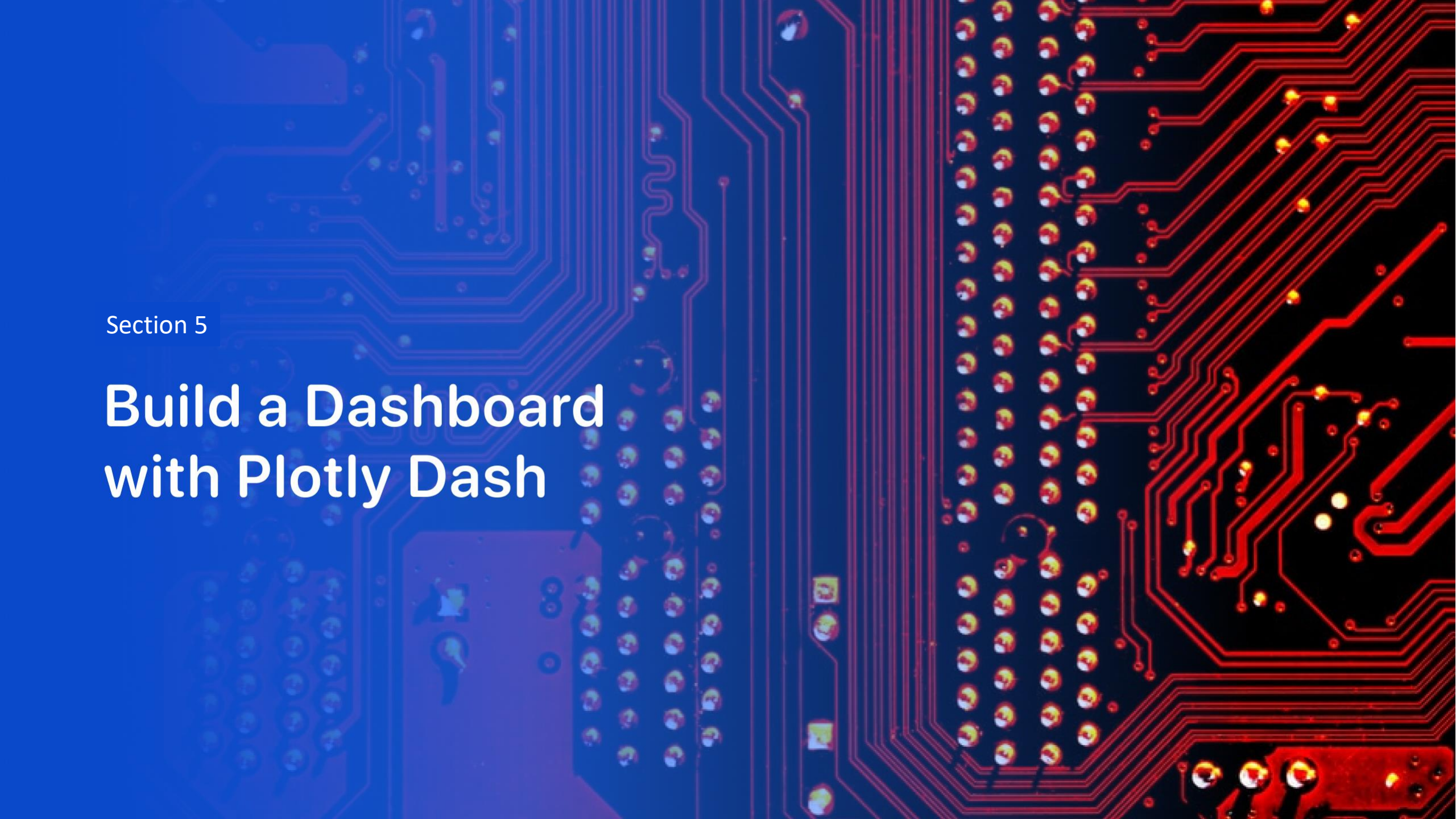


Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

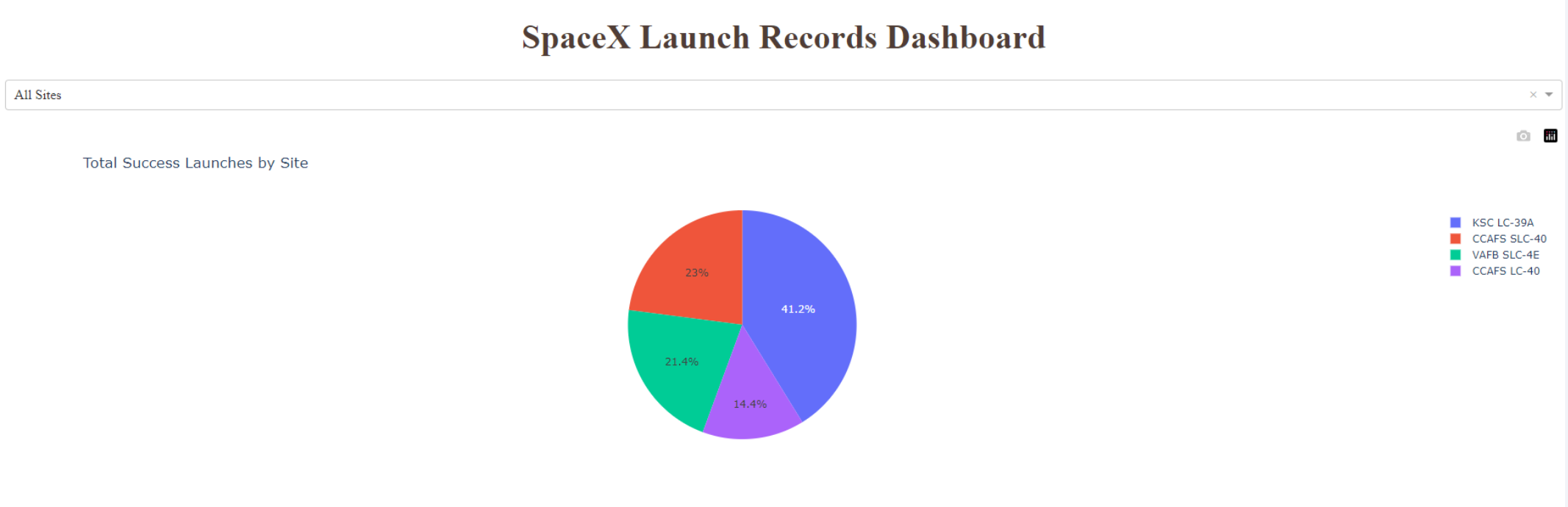
Proximities of Launch Sites



It can be found that the launch site is close to railways and highways for transportation of equipment or personnel, and is also close to coastline and relatively far from the cities so that launch failure does not pose a threat

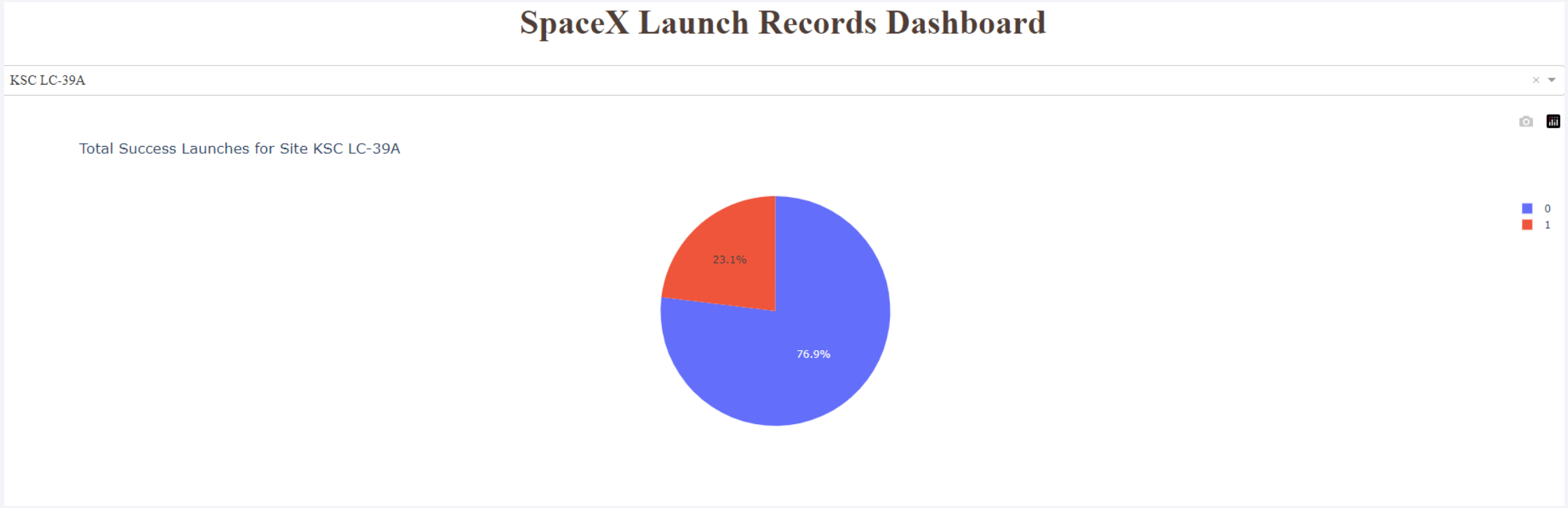

Section 5
Build a Dashboard with Plotly Dash

Launch success count for all sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

Launch site with highest launch success ratio



KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%)

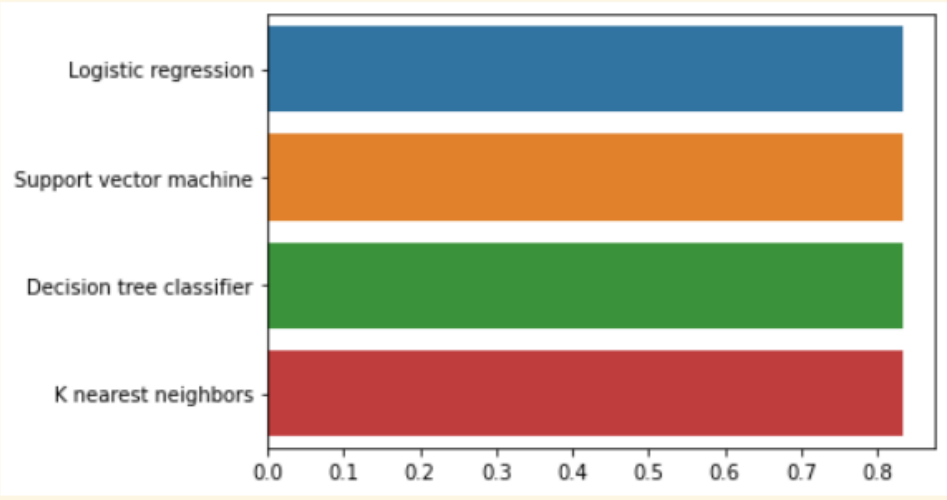
<Dashboard Screenshot 3>

These figures show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is higher than that of heavy weighted payloads(5000-10000 kg)

Section 6

Predictive Analysis (Classification)

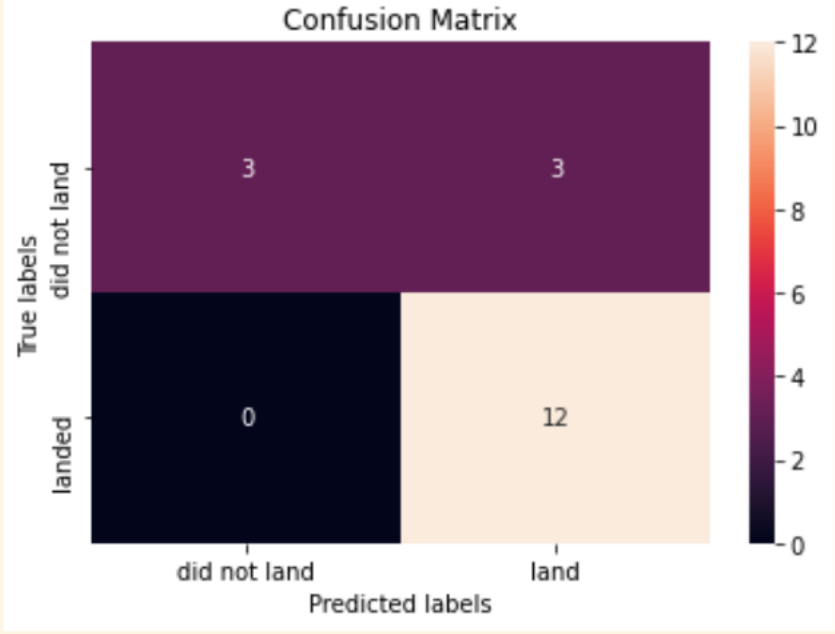
Classification Accuracy



	Method	Accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.833333
3	K nearest neighbors	0.833333

- In the test set, the accuracy of all models was virtually the same at 83.33%.
- It should be noted that the test size was small at 18.
- Therefore, more data is needed to determine the optimal model

Confusion Matrix



- The confusion matrix is the same for all models because all models performed the same for the test set.
- The models predicted :
  - 12 successful landings when the true label was successful
  - 3 failed landings when the true label was failure.
  - 3 predictions that said successful landings when the true label was failure (false positive).
- Overall, these models predict successful landings

Conclusions

- As the number of flights increased, the success rate increased, and recently it has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- The launch site is close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.

Appendix

- GitHub URL
- Coursera Applied Data Science Capstone URL

Thank you!