IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rodrigo Calaboni
14/08/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection and wrangling
  - Exploratory data analysis using data visualization and SQL
  - Creating an interactive map with Folium and Dashboard with Plotly Dash
  - Predictive analysis testing different machine learning models
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX is a commercial space responsible for expressive price changes in space travel. Falcon 9 launches advertised on its website costs 65 mi dollars, while other providers costs more than 165 mi dollars. The main reason for Falcon 9 being affordable is due to the fact that they can reuse the first stage of it's launch.
- Problems you want to find answers
  - Is there a way to have a better understanding and predicting the outcomes of a successful launch, capable of reusing the first stage?
  - How does each variable available affect the success rate?
  - How does the success rate evolve with the passing years?

Section 1

Methodology

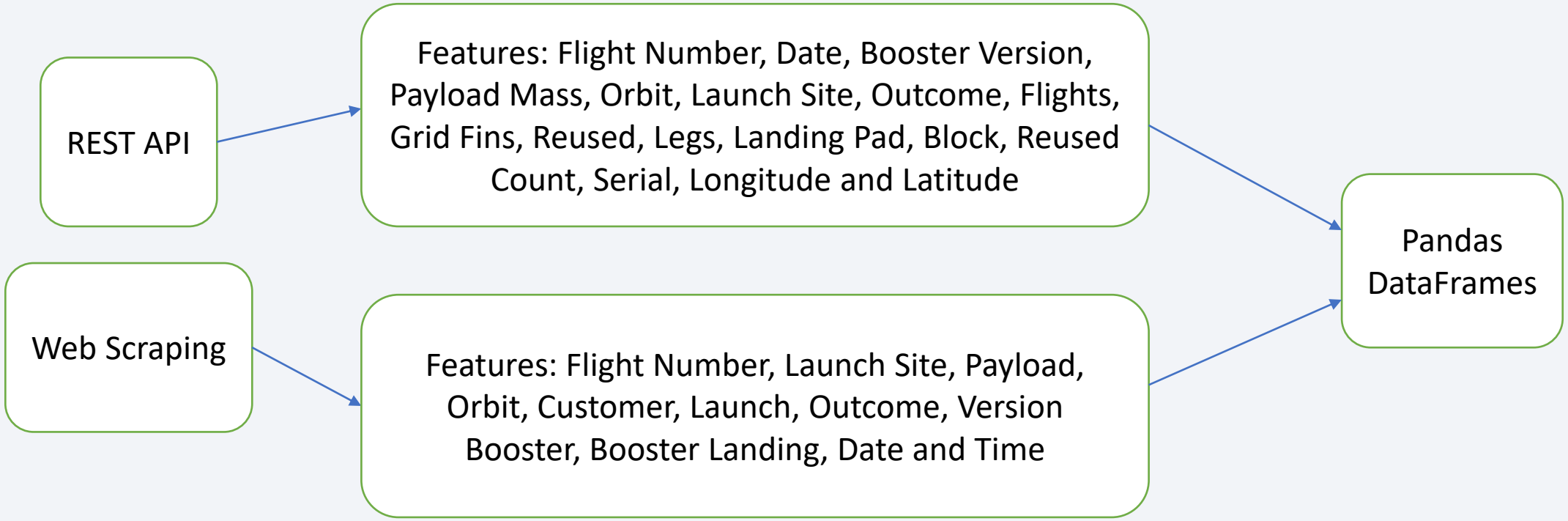
Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API and Web Scrapping from Wikipedia page
- Perform data wrangling
  - In this step, data was filtered, cleaned from missing values and categorical variables were prepared for binary approaches using One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different models were built, tuned and evaluated for finding the best results among them

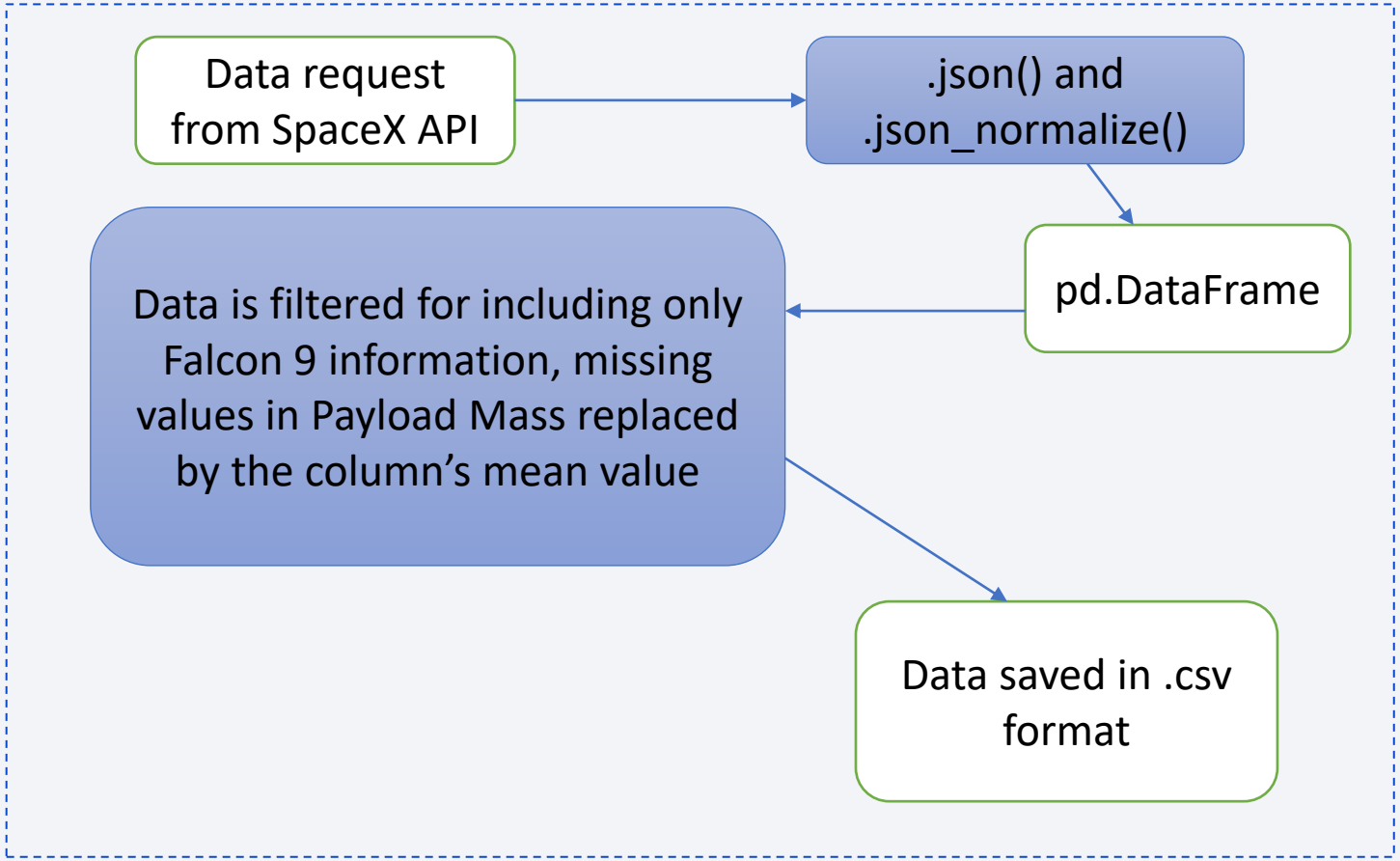
Data Collection

- The project started with the collection of Data from SpaceX REST API and Web Scraping data from SpaceX Wikipedia page.



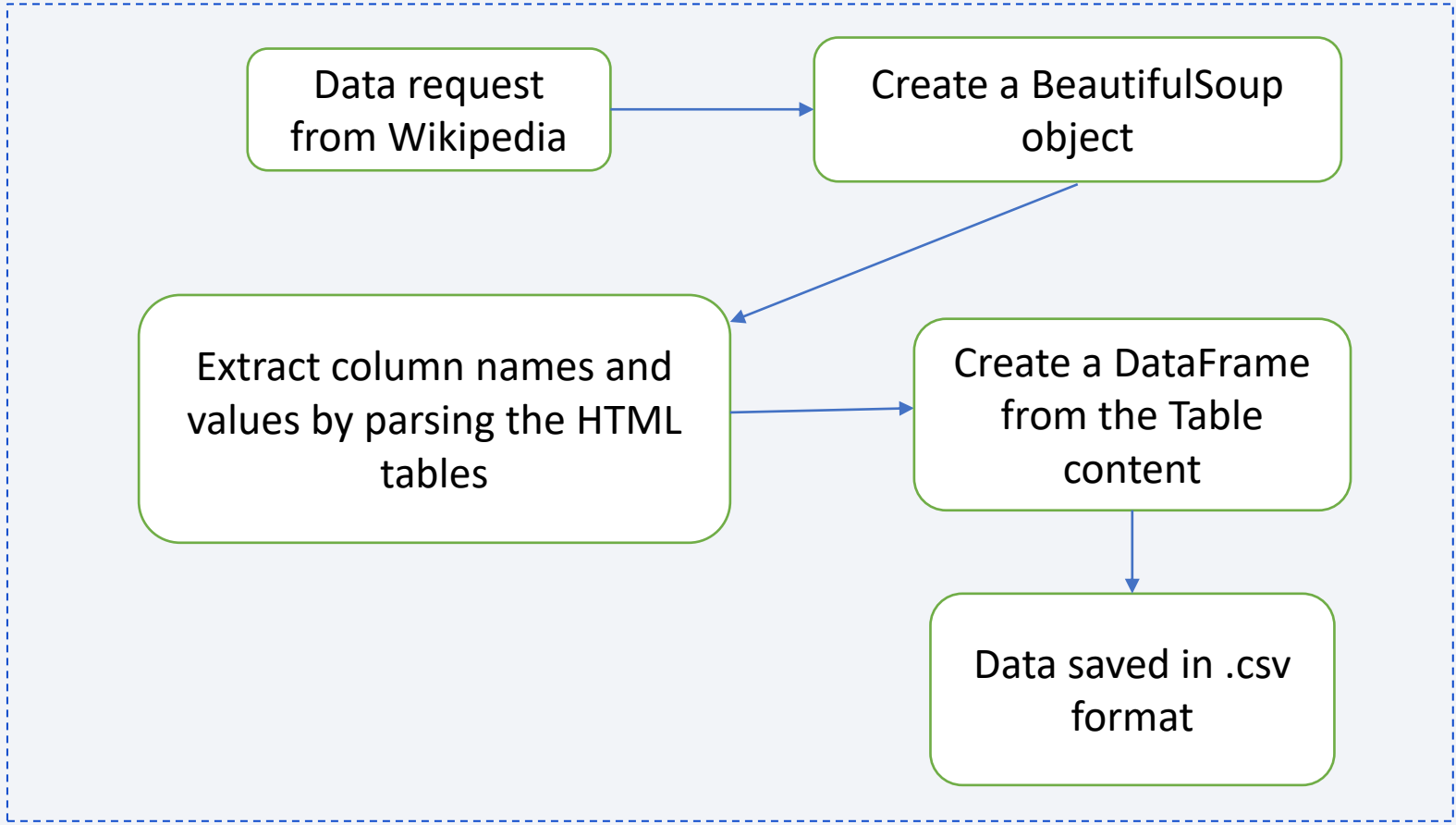
Data Collection – SpaceX API

GitHub URL for the Data Collection REST API Code



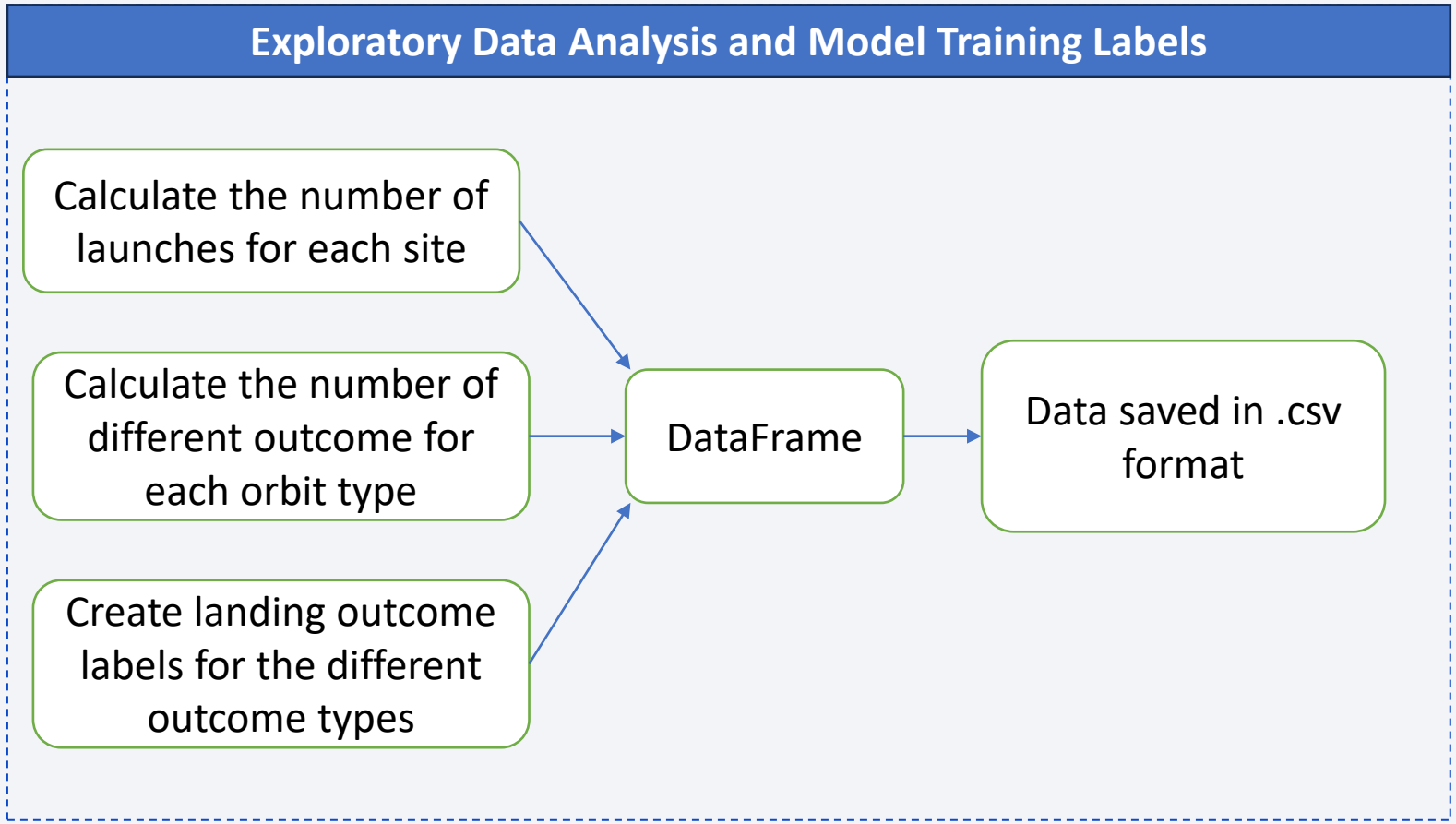
Data Collection - Scraping

GitHub URL for the Data Collection Scraping Code



Data Wrangling

GitHub URL for the Data Wrangling Code



EDA with Data Visualization

GitHub URL for the Data Visualization Code

Charts plotted contained the following information:

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit Type vs Success Rate
- Flight Number vs Orbit Type
- Payload Mass vs Orbit Type
- Success Rate vs Year

Scatter plots helps visualizing correlation between two variables. If there is any correlation, it can be calculated through machine learning models.

Bar charts helps visualizing the relationship between different categorical data types. The distribution of certain categories between succeeded and unsucceeded launches can lead to identifying clusters.

Line charts show the evolution of the value over time. If there is a chronological trend in Success Rate over the years, it can be easily visualized.

EDA with SQL

GitHub URL for the SQL EDA Code

SQL Queries for displaying the following information from the Database:

- Names of each launch site in the database
- Five records Where launch sites begin with CCA text
- Total payload mass carried by boosters launched by NASA's site
- Average payload mass carried by booster version F9 1.1
- The date of the first successful landing in ground pad
- The names of boosters successfully landed in drone ship with payloads between 4000 and 6000 kg
- Total number of successful and unsuccessful landing outcomes
- Names of booster versions with the maximum payload mass
- The failed landing outcomes for drone ships, their booster version and launch distributed along the months for 2015
- Rank of the value Count of landing outcomes between 04/06/10 and 20/03/17

Build an Interactive Map with Folium

GitHub URL for the Folium Map Code

- Markers pointing to each launch site using Latitude and Longitude information
- Coloured marker clusters showing successful and unsuccessful launches for each site
- Distance between a Launch Site to the nearest railway, highway and coastline
- The map shows that launch sites are installed in low latitudes, close to coastal lines and have safe distances to population dense areas

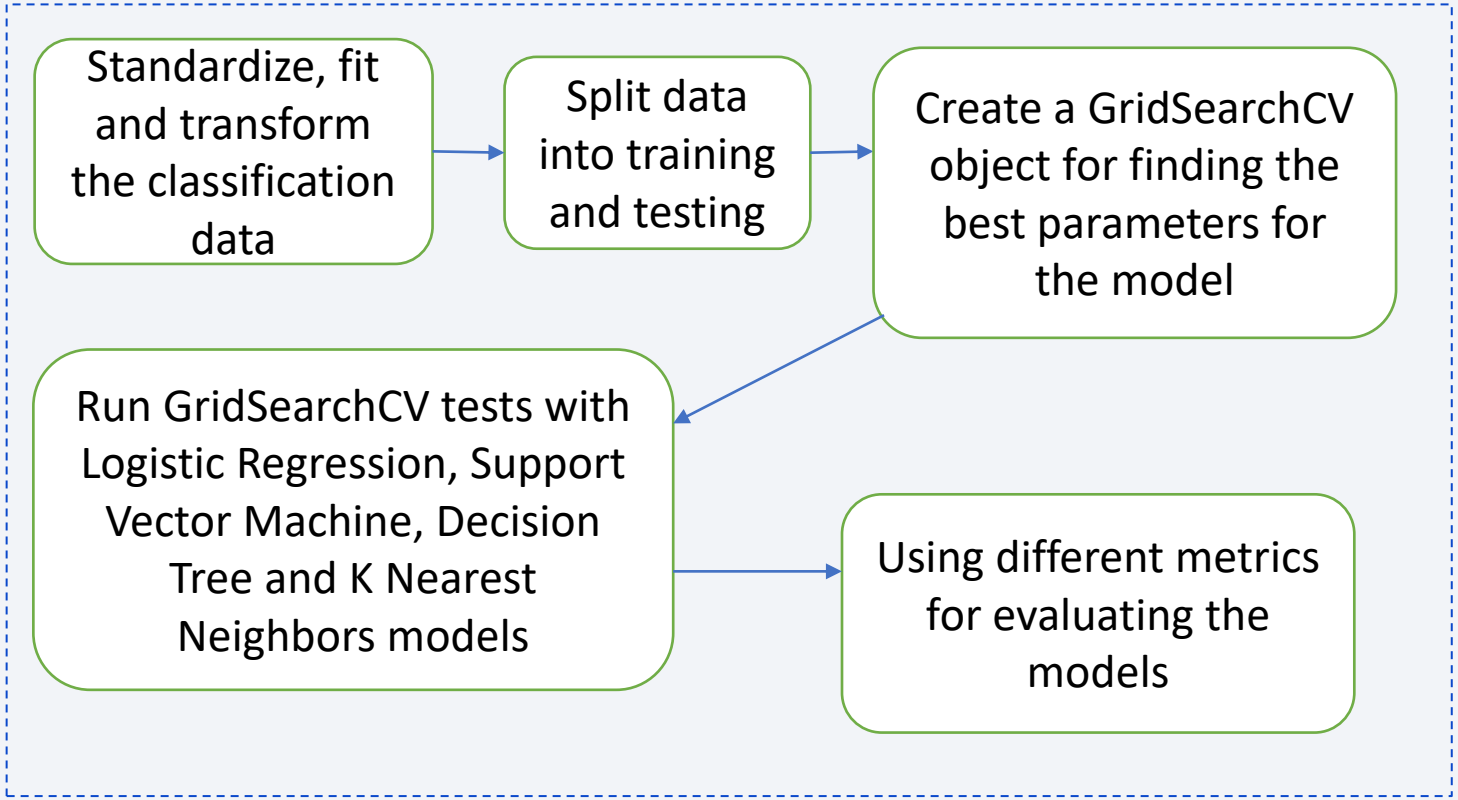
Build a Dashboard with Plotly Dash

GitHub URL for the Dashboard Code

- Dropdown list for Launch Site selection
- Pie Chart showing Success Rate for selected Launch Sites
- Slider for specifying minimum and maximum Payload Mass
- Scatter Plot of Payload Mass vs Success Rate for each Booster Version

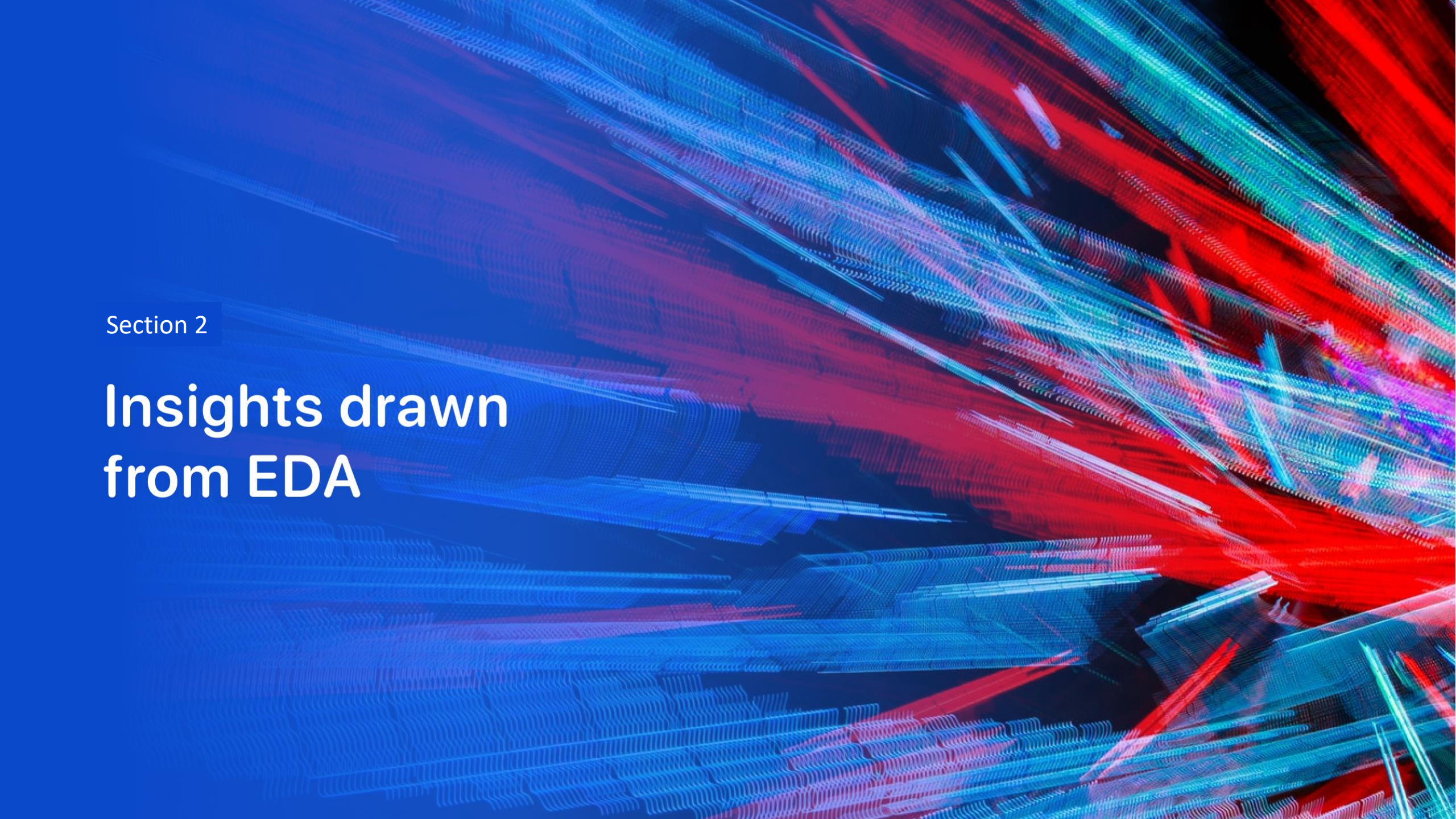
Predictive Analysis (Classification)

GitHub URL for the Classification Code

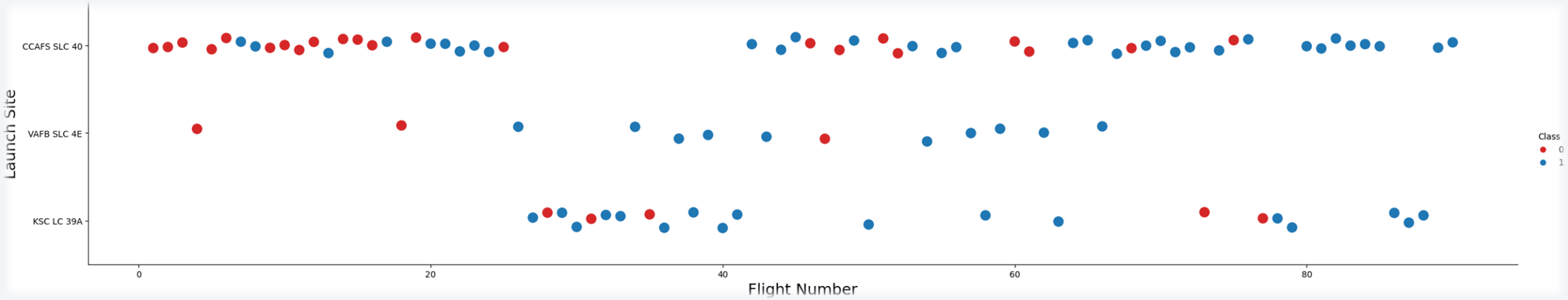


Results

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

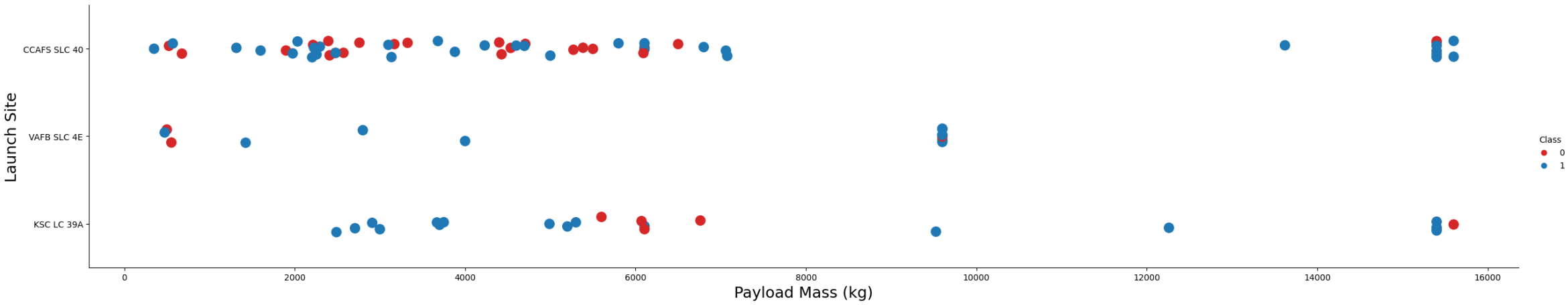

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The initial 20 launches, executed in CCAFS SLC 40 and VAFB SLC 4E sites have lower sauccess rate. The following launches had better success for all Launch Sites.
- This result shows that the number of launches executed has a higher influence to success rate than the Launch Site location.

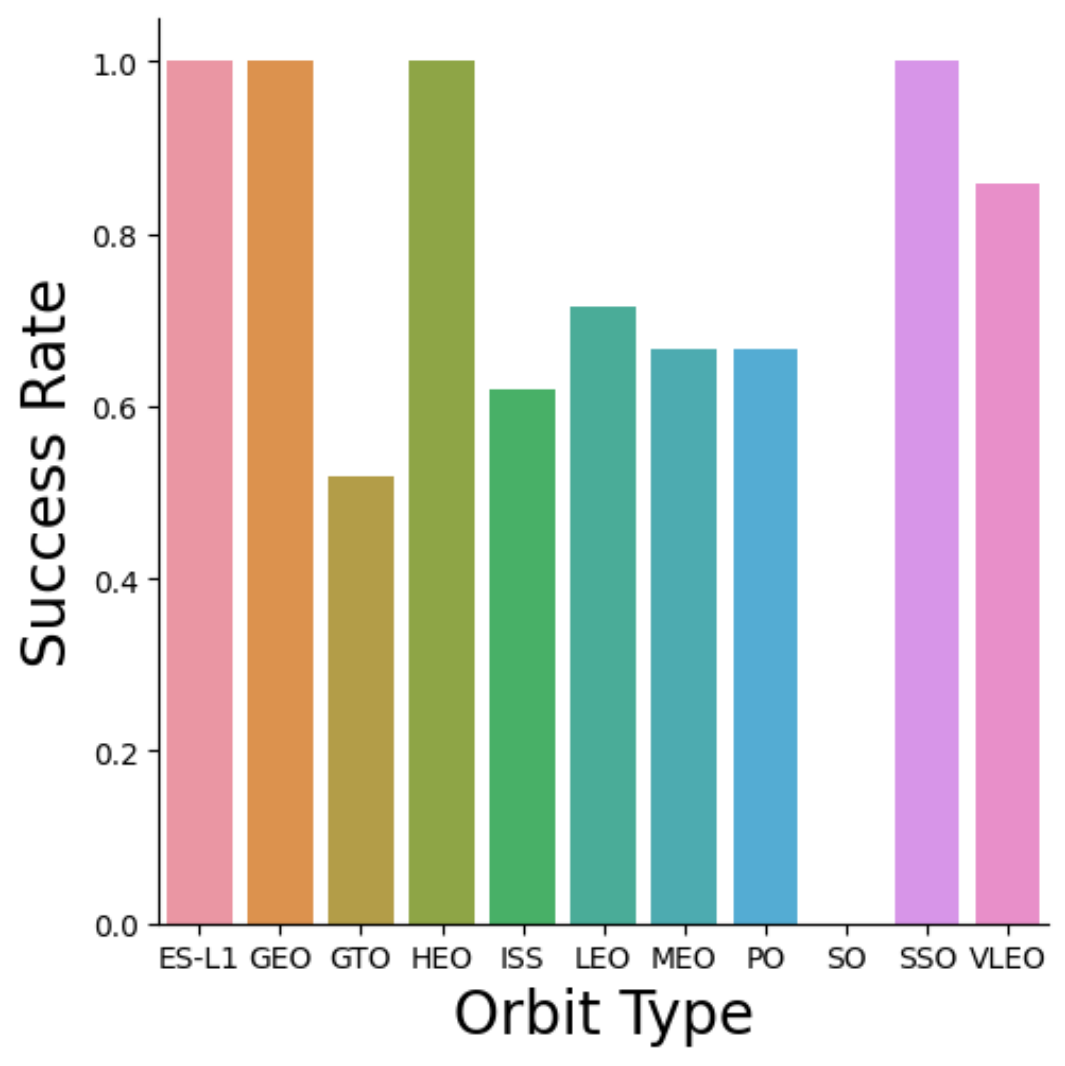
Payload vs. Launch Site



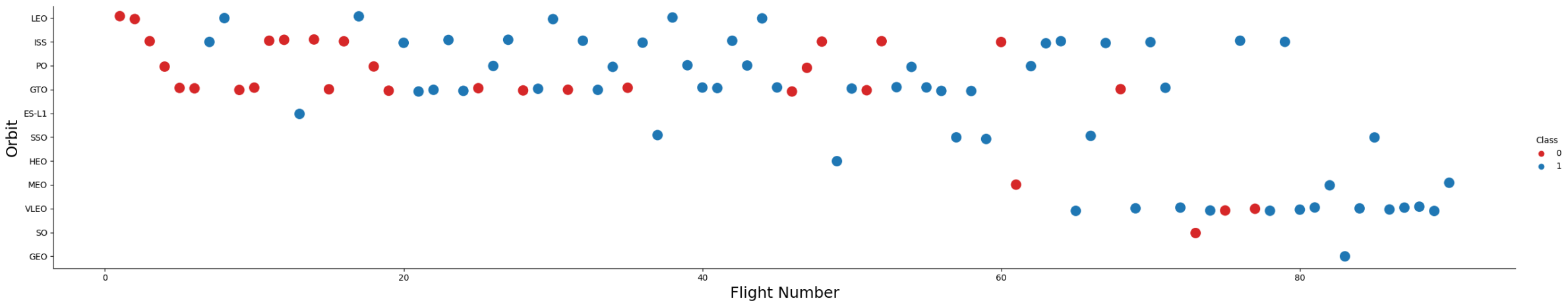
- Payloads seems to have better success rate with mass greater than 8000 kg

Success Rate vs. Orbit Type

- Orbits with the best rate were ES-L1, GEO, HEO and SSO, with100% success rate, and VLEO with more than 80%
- SO Orbit had no succeded landing
- The other Orbit Types had success rate between 50 and 70% success rate

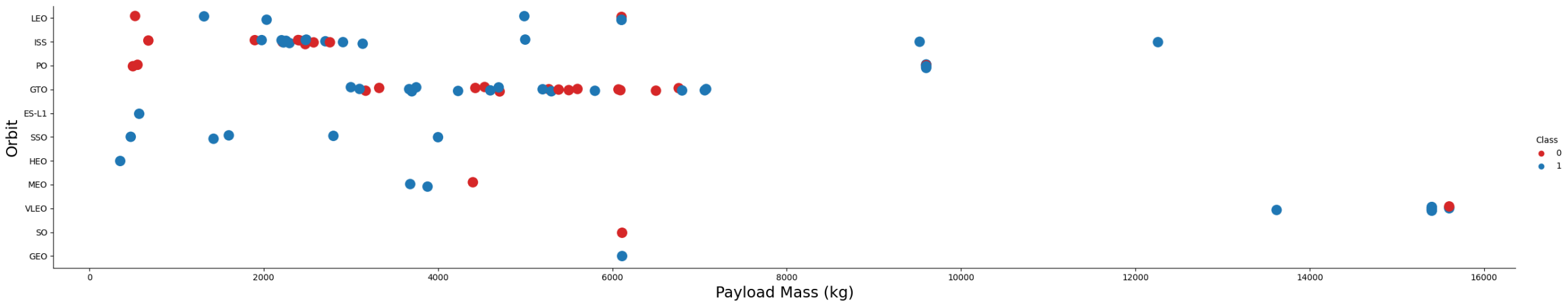


Flight Number vs. Orbit Type



- In general, the orbit types seem to have higher success rates with more launch attempt numbers.
- The GTO orbit is an exception, showing no clear correlation with between flight number and success rate

Payload vs. Orbit Type

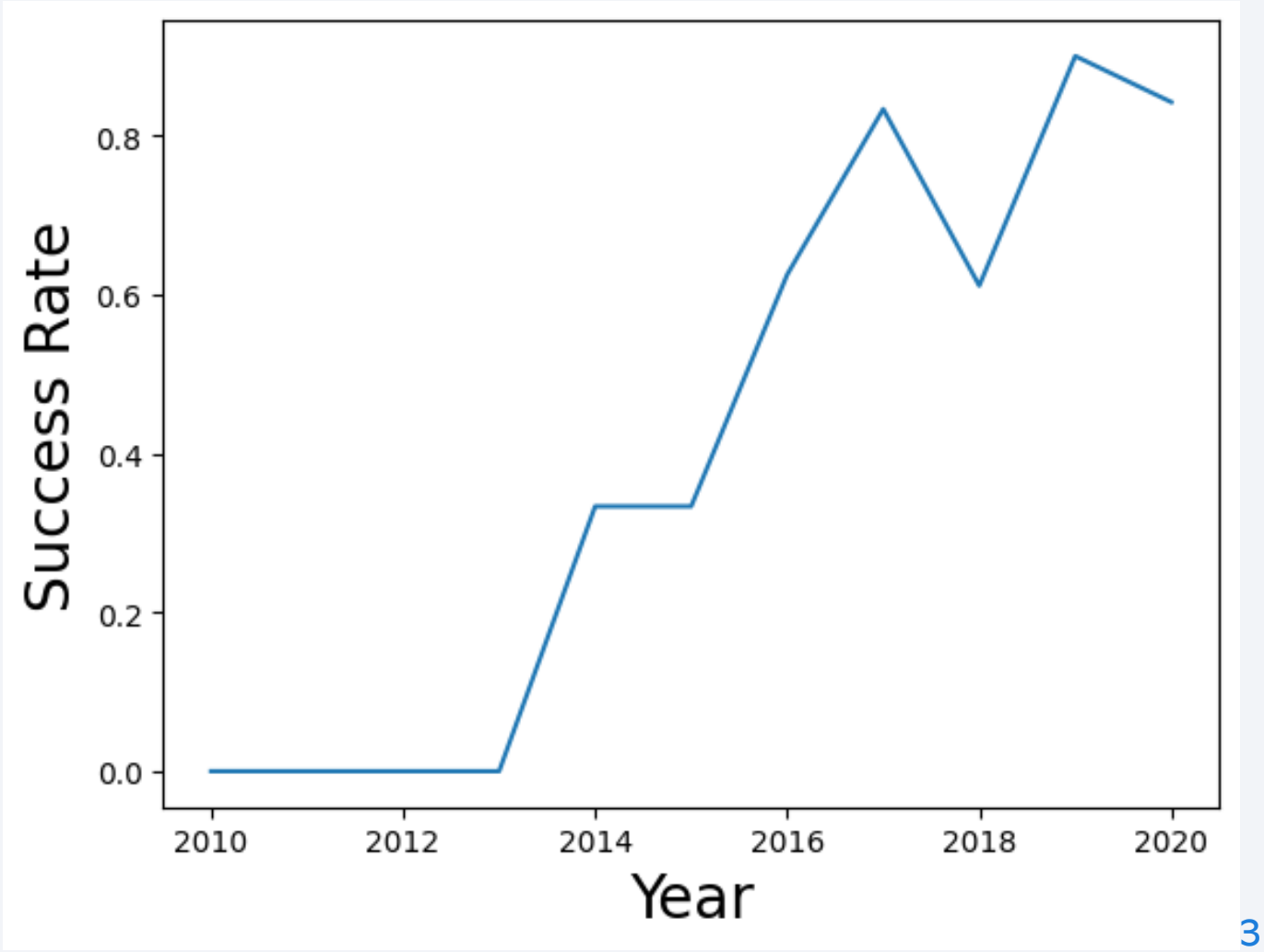


- In general, the orbit types seem to have higher success rates with heavier payloads.
- As well as it was seem for the Flight Number, GTO orbit has no clear correlation between Payload Mass and success rate.

Launch Success Yearly Trend

- The yearly success rate trend shows better results as time passes.

- This plot corroborates with the flight number assumption that success rate has positive correlation to the number of launch attempts



All Launch Site Names

- SQL Query for displaying each Launch Site name on the specified column

```
%sql select distinct launch_site from SPACEXTABLE;
```

✓ 0.0s

```
 * 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'

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS...	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraf...	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flig...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flig...	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- SQL Query for showing the first five entries with Launch Sites with names beginning with ‘CCA’

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

total_payload_...
45596

- SQL Query for showing the sum of all payload mass carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

average_payload...
2534.66666666666...

- SQL Query for showing the average payload mass carried by F9 rockets with Booster Version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

first_successful_...
2015-12-22

- SQL Query for showing the date of the first successful landing attempt on a ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

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

- SQL Query for showing the names of boosters which have successfully landed on a drone ship and had payload mass between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload s...	1

- SQL Query for showing the sum of each mission outcome type

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * 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

- SQL Query for listing the booster versions of launches carrying the maximum payload mass

2015 Launch Records

```
%%sql select "Landing_Outcome", substr(Date,1,4) as "year", substr(Date,6,2) as "month",  "Booster_Version", "Launch_Site" from SPACEXTABLE
    where "Landing_Outcome" = "Failure (drone ship)" and substr(Date,1,4)="2015";
```

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

Landing_Outcome	year	month	Booster_Version	Launch_Site
Failure (drone ship)	2015	10	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	2015	04	F9 v1.1 B1015	CCAFS LC-40

- SQL Query for showing the boster versions, launch sites and months for landing outcome failed in drone ships in 2015

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

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

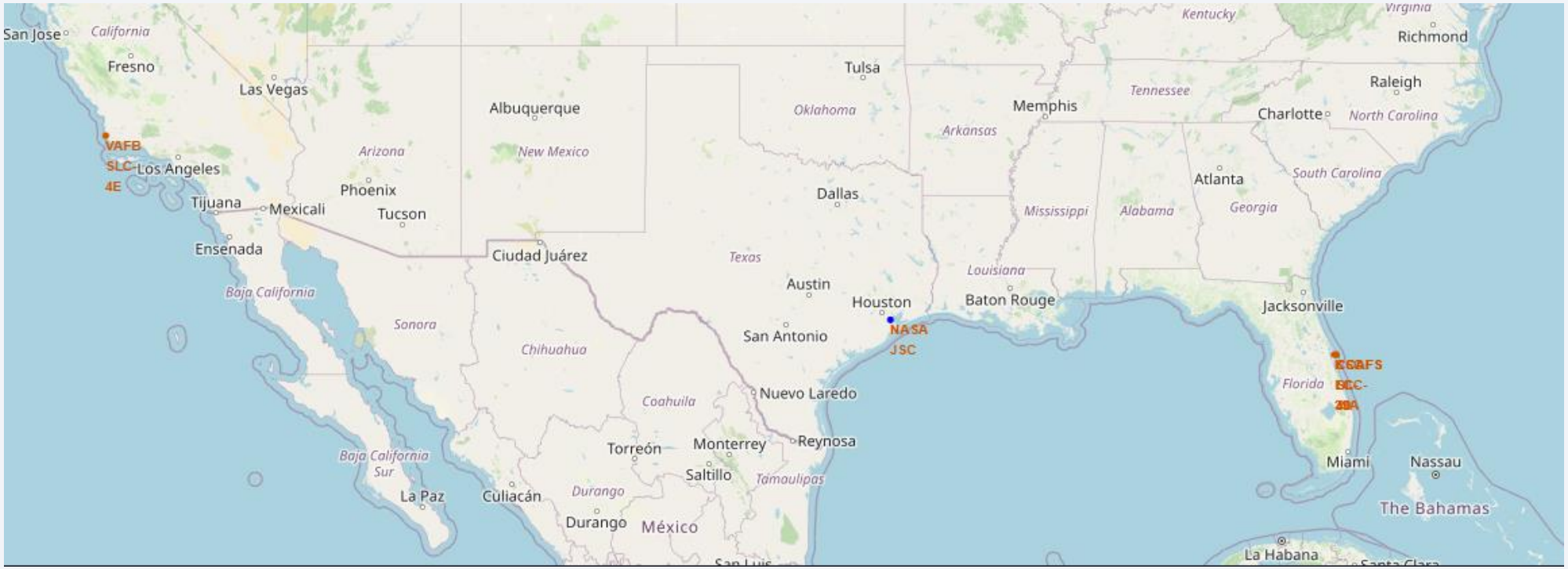
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
No attempt	10
Success (ground p...	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocea...	2
Precluded (drone s...	1
Failure (parachute)	1

- SQL Query for showing a descending ranking of the total landing outcomes between 2010/06/04 and 2017/03/20


Section 3
Launch Sites
Proximities Analysis

Launch Sites displayed on the global map



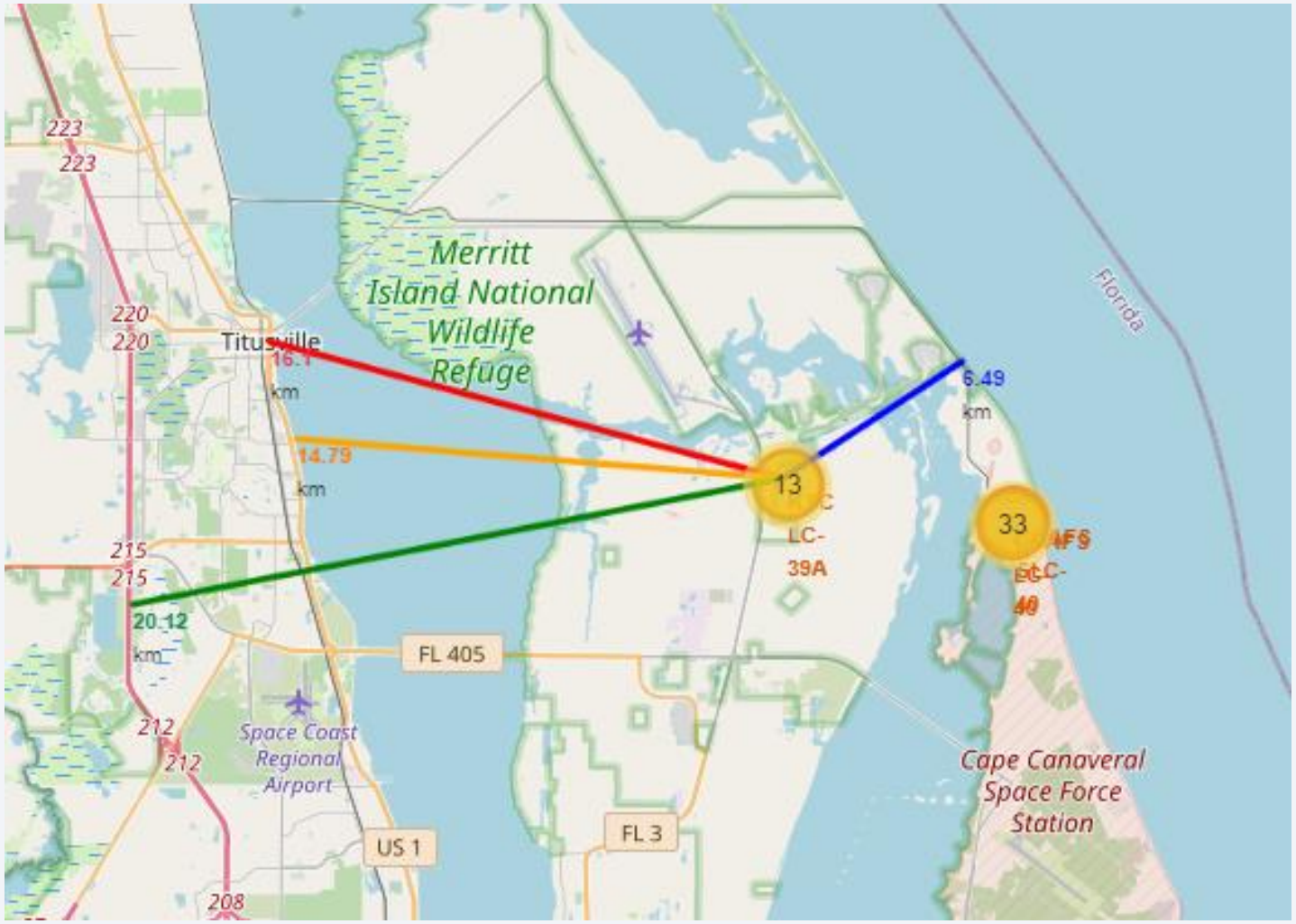
- The sites are close to the south limits of USA, wich are the closest parts to equator line. The closer to equator line, higher is the momentum from land surface, making it easier to maintain speed for staying in orbit.
- All sites are positioned in coastal areas, reducing the risk of debris falling in populated areas, if an accident occurs.

Launch records categorized by outcome

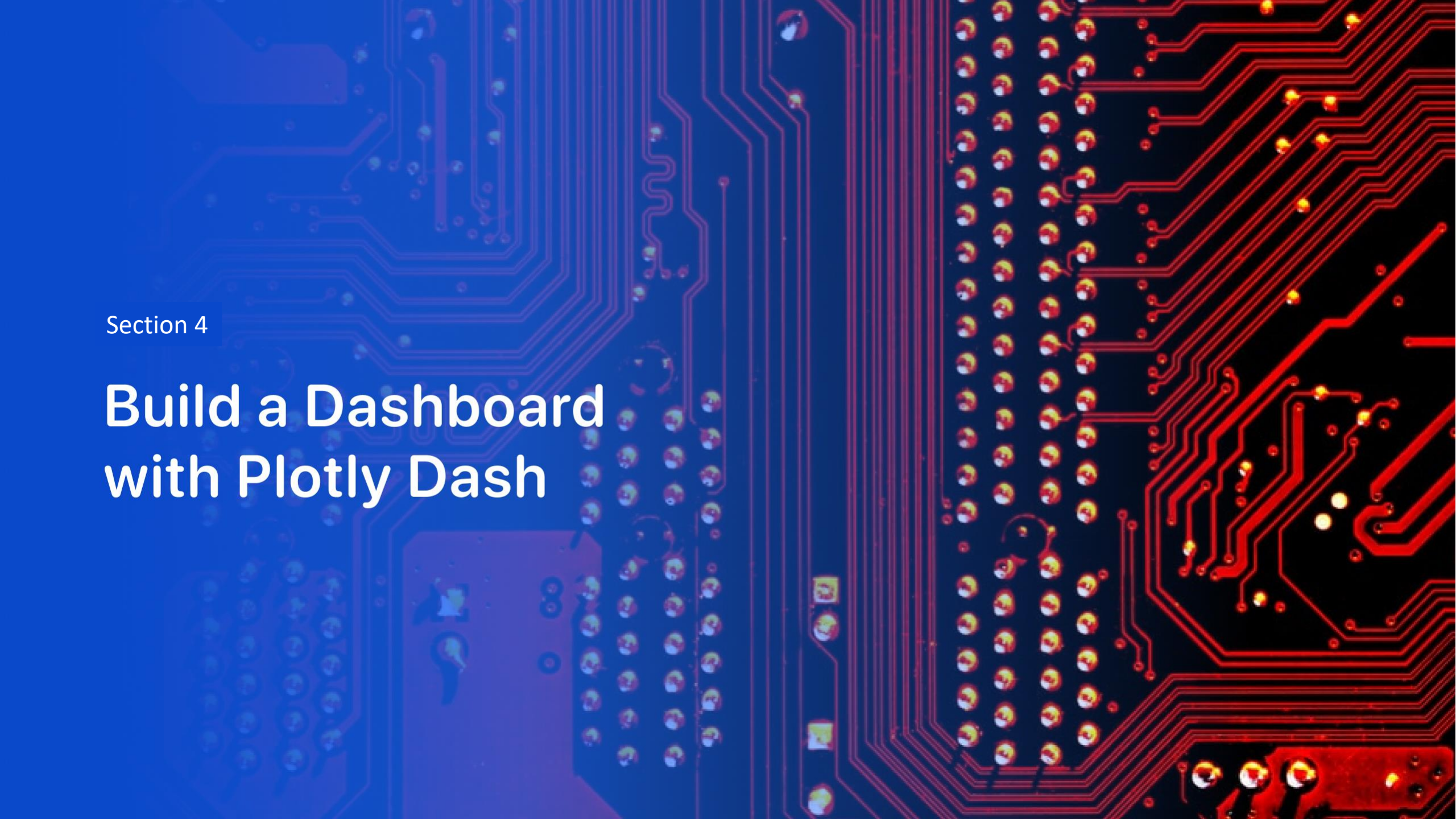
- The markers labeled in green and red helps visualizing the success rate from each highlighted launch sites.

- The example shows a success rate of 40% for the VAFB SLC-4E Site, with 4 successful launches out of 10

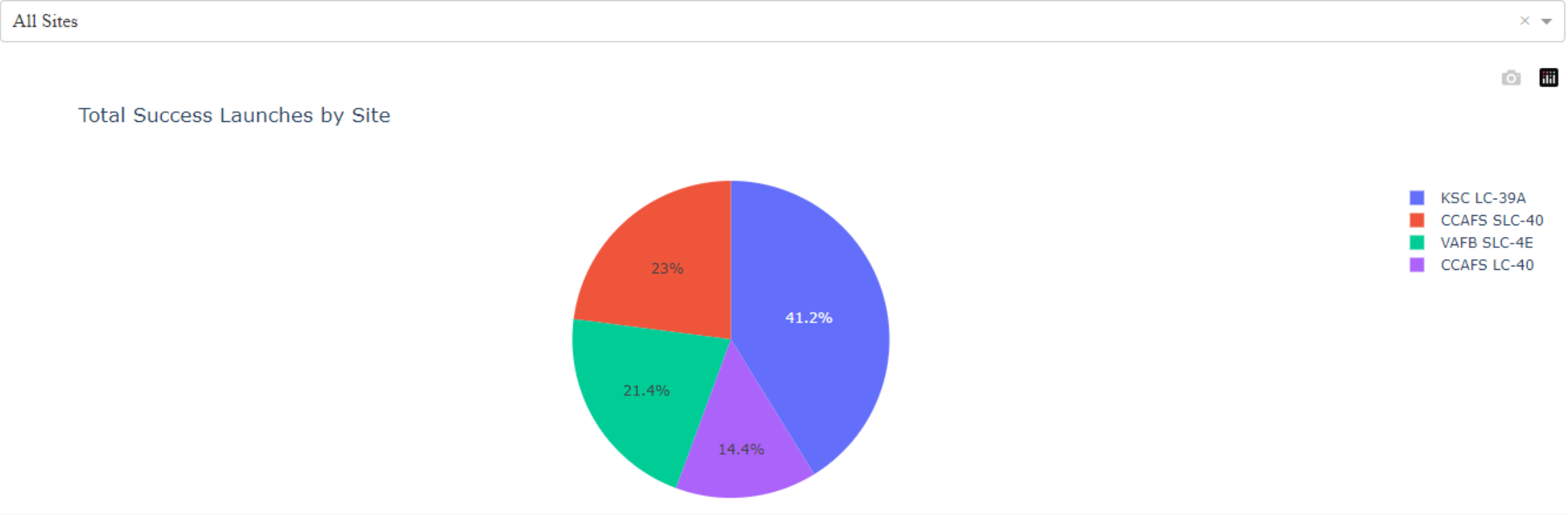
KSC LC-38A distance to important features



- The map shows that launch site is relatively close to areas that are accessible to people, with closest railway, highway and city distances between 14.79 and 20.12 km. This proximity to population should be a concern for the debris fall risks.
- The site is closer to the coast line (6.49 km), so launch missions should focus on east-northeast directions for avoiding risks.

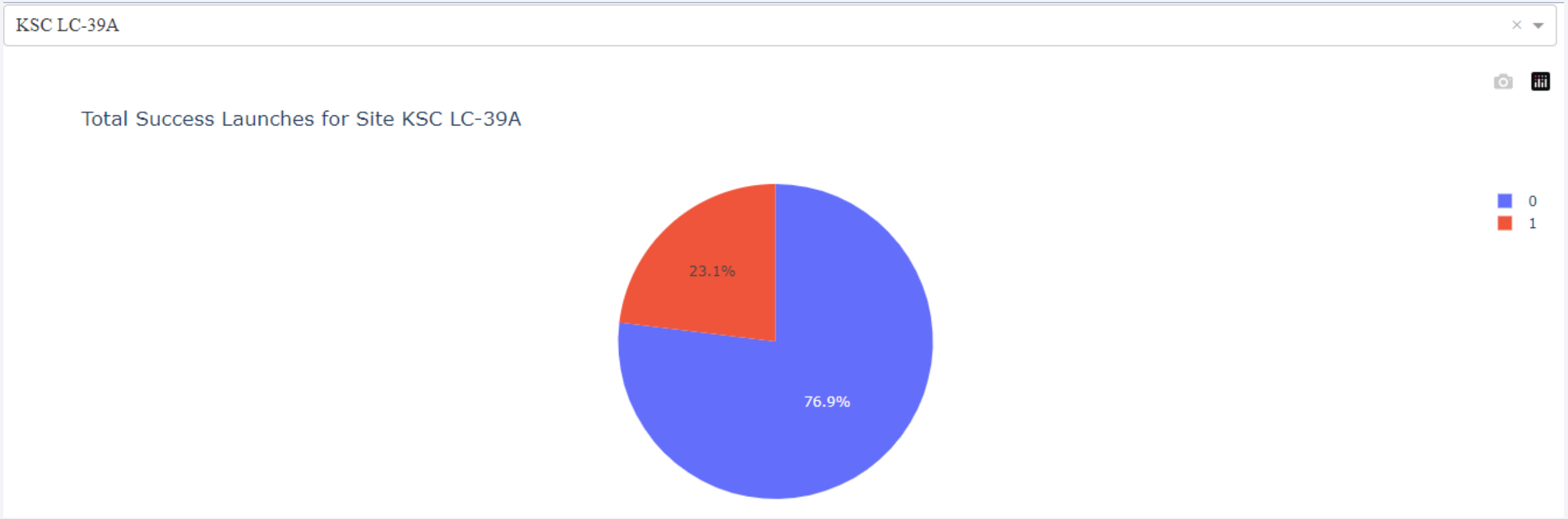

Section 4
Build a Dashboard with Plotly Dash

Success Launches count by site



- The pie chart shows that KSC LC-39A has most of the success launches, if compared to the other sites.
- The count for more success can be a false high success rate if that site has failed attempts counts even higher than other sites. This can occur when the dataset have different sizes for each feature analyzed.

Success Rate Pie Chart for the Best Score



- Going through the pie charts for each launch site, the higher success rate for the KSC LC-39A was confirmed, with positive outcome for 76.9% of the attempts

Payload Mass vs Launch Outcome

- The scatter plot for all sites and athe whole payload mass range shows that payloads between 2000 and 4000 kg have the higher positive outcomes.
- The second chart highlights that range of payload mass.

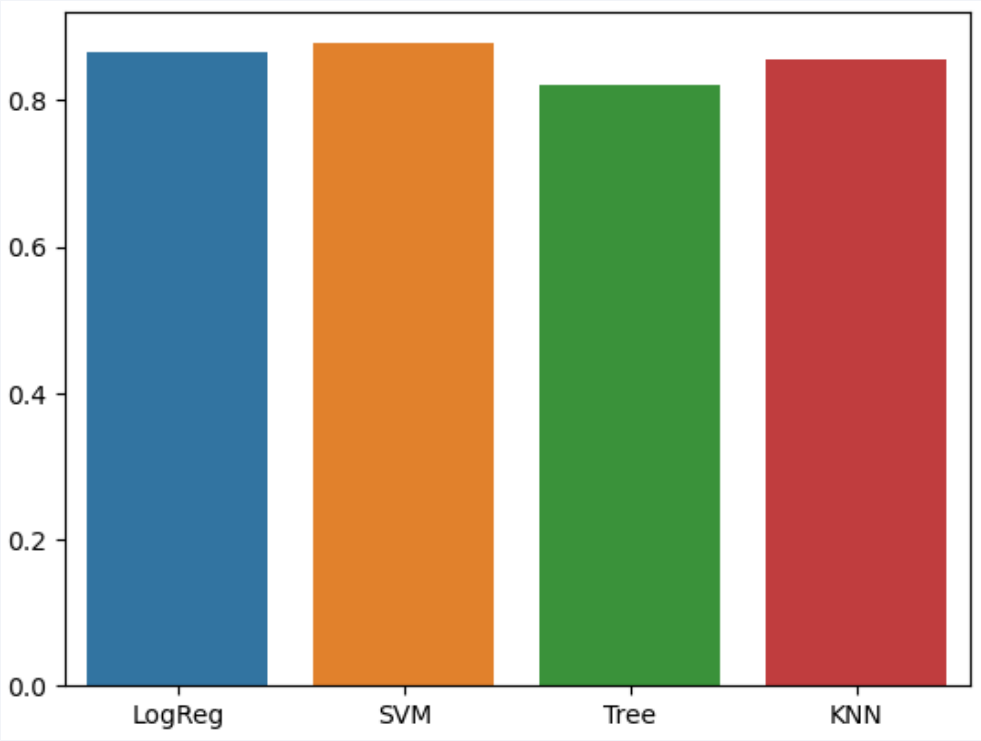
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All models showed relatively high accuracies, with more than 80%

- For this task, the Support Vector Machine model seems to have the best performance

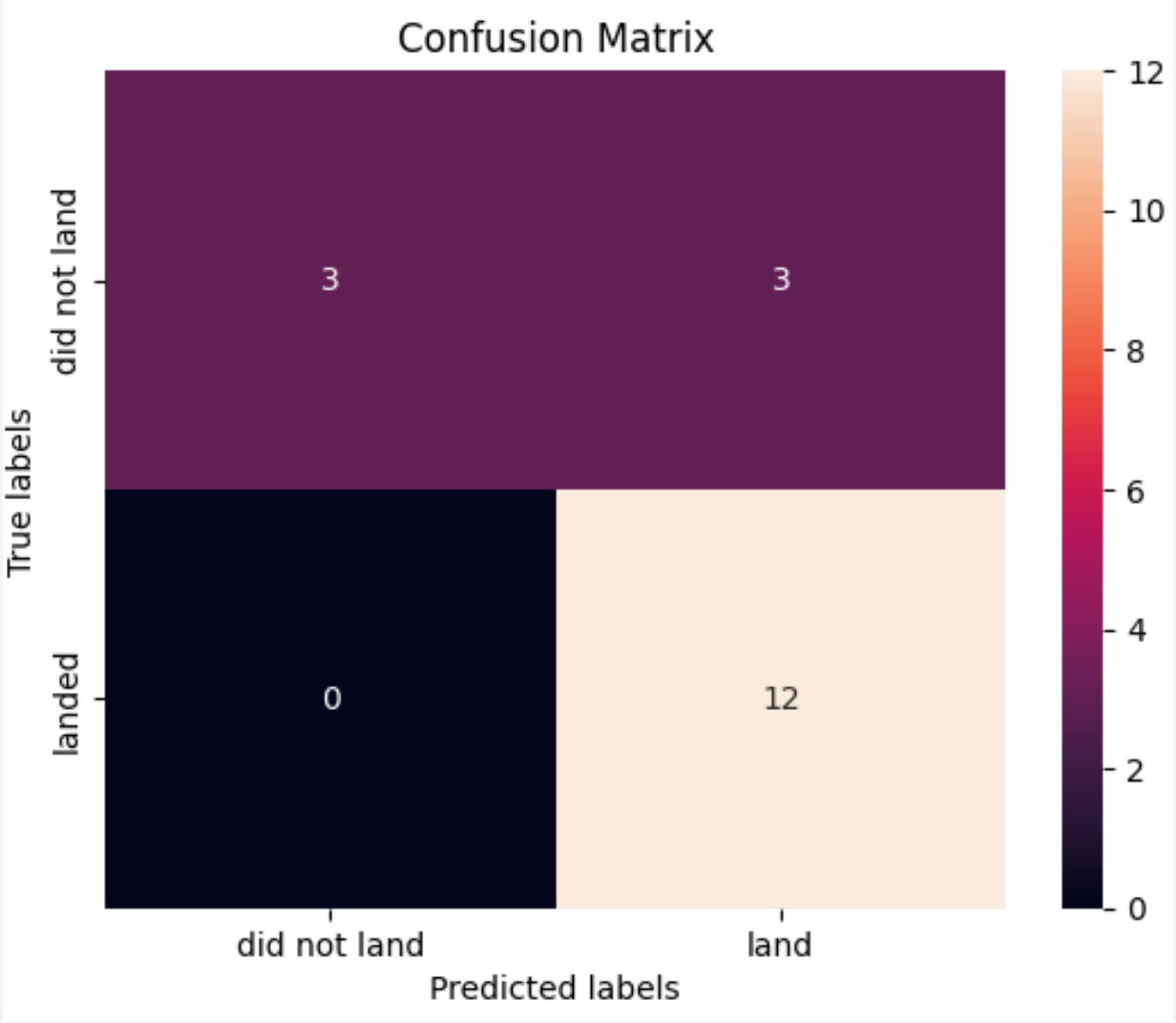


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.777778	0.819444
F1_Score	0.909091	0.916031	0.875000	0.900763
Accuracy	0.866667	0.877778	0.822222	0.855556

Confusion Matrix

- A Confusion Matrix shows the true classes in the vertical axis and the predicted labels in the horizontal axis. The numbers in the main diagonal (3 and 12) are right predictions. Numbers out of the main diagonal are wrong predictions.

- The confusion matrix for the SVM model shows that, for the part of the dataset separated for testing, there were 12 positive landing outcomes and 6 negative. The model was able to predict all the 12 posivite outcomes correctle, but predicted only half of the negative ones.

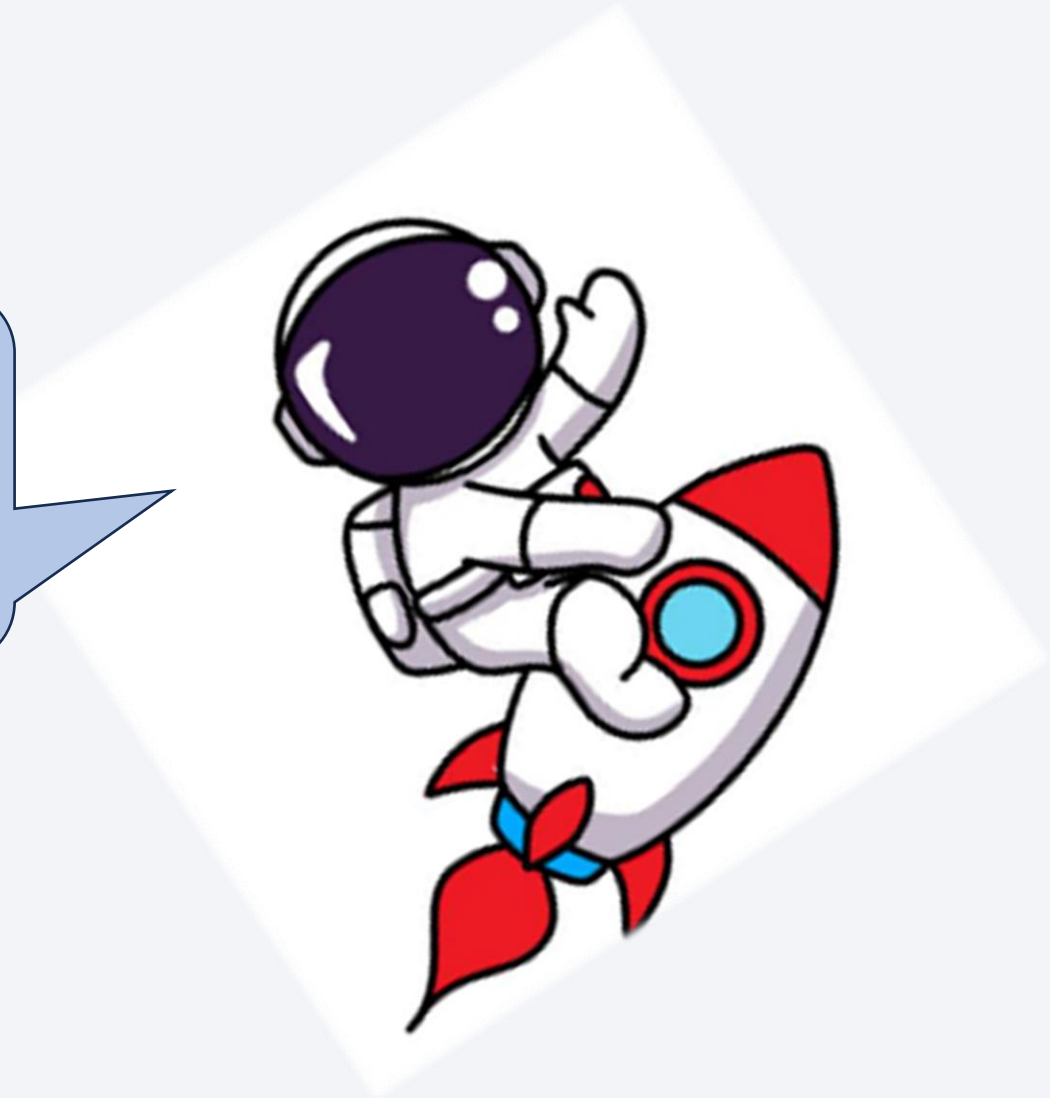


Conclusions

- Although SVM had the best performance, all prediction models had good results and should be fine tuned and tested for the future data.
- The confusion matrix for the SVM showed that it predicted 15 positive outcomes and 3 negative. 3 of the positive predictions were false positives, meaning that 3 unsuccessful landings could not be predicted.
- The data used for training the models had 48 positive outcomes of a total of 72, resulting in 66.7% of the data used showing positive outcomes. The unbalanced dataset used for training can make the model “get used” to predict more positive outcomes, resulting in the false positives seen on the confusion matrix.
- From the exploratory analyzis executed, the dataset shows that the landing outcomes have correlations to payload mass, the number of launches executed and the types of orbits.

Appendix

Special thanks to Instructor from IBM and Coursera and my colleagues for the Peer Reviews!



Thank you!