IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ryan Mascarenhas
28 Sep 2024

Outline

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

Executive Summary

SpaceX advertises rocket launches for its Falcon 9 rocket for $62M – a discount of more than $100M (62%) compared to its competitors.

The goal of this report is to use a disciplined approach to collect, wrangle and explore the data through various approaches – SQL queries, visualizations.

Armed with this initial analyses, we are able visualize launch related data on maps and interactive web dashboards to demonstrate findings and build some initial hypotheses around relevant features

Finally, we build and choose a classification model from a variety of models using GridSearch with Cross Validation to tune hyper parameters.

By using this model, we could imagine that a competitor could selectively bid on launches with a higher success rate and thus a higher re-use of the rockets.

For reference, Github Repo: https://github.com/ryanmas13/SpaceY_FinalProject.git

Introduction

- SpaceX advertises rocket launches for its Falcon 9 rocket for $62M – a discount of more than $100M (62%) compared to its competitors!
- The company claims it can do that because it can re-use the first stage.
- Our goal is to analyze previous launch data and identify a good model to predict if the first stage will land successfully
- The information would be useful for a competitor to close the pricing gap

Section 1

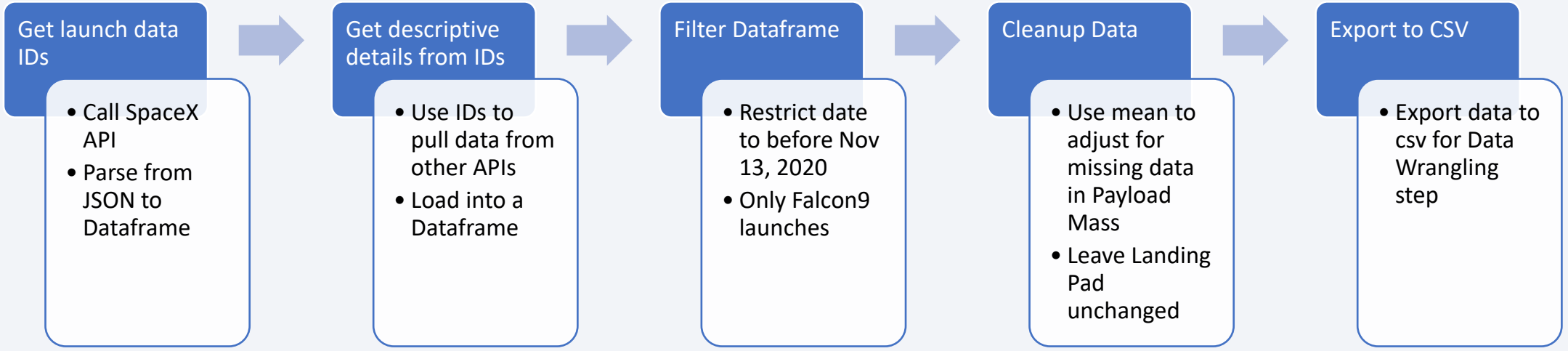
Methodology

Methodology

Executive Summary

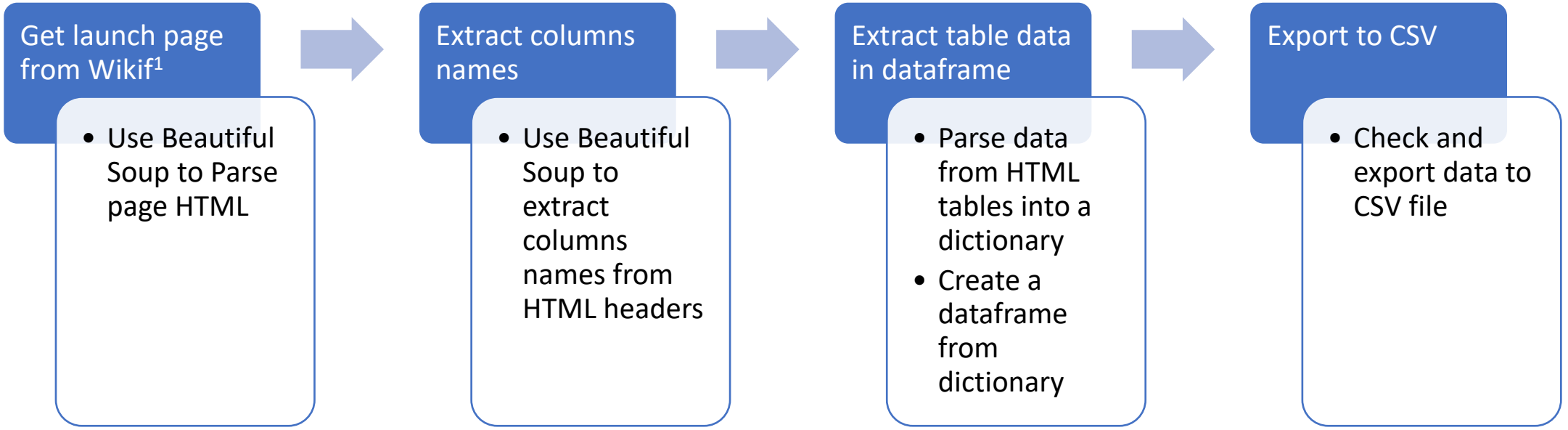
- Collect launch data: Falcon 9 only
  - Selectively parse and filter data from a Spacex REST API and webscraped Wiki page
- Perform data wrangling:
  - Analyze the data structure, then, clean it and define a target variable for launch success
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use a variety of classification models with GridSearch and then compare their accuracy to pick a good model

Data Collection – SpaceX API



Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/1_Data_collection_SpacexAPI.ipynb

Data Collection – Webscraping



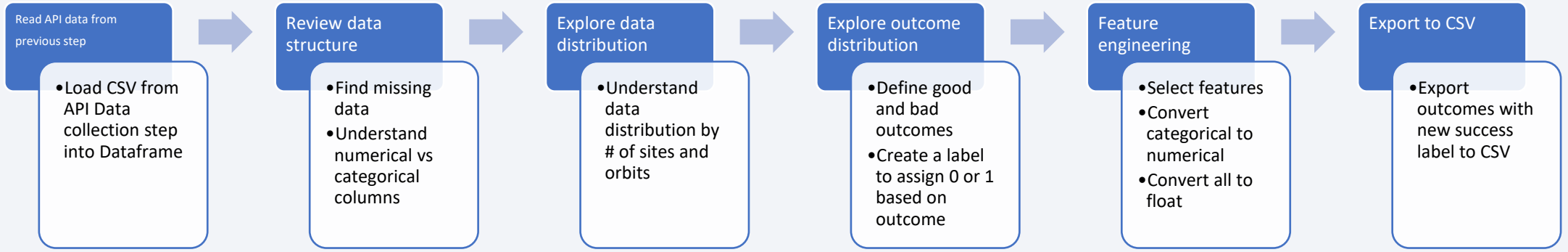
Github URL to notebook:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/2_Data_collection_WebscrapingWiki.ipynb

1 Wiki link to data
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Wrangling – Understand and assign label for training data

Pg.7

Output of Data Collection – SpaceX API



Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/3_Data_wrangling.ipynb

EDA with SQL – High level summary

1. Create new table to transfer launch data from csv
2. Query unique launch sites
3. Query loads for specific customer and specific boosters
4. Query to find the first successful ground pad landing date
5. Query boosters that were successful with range of payloads
6. Query mission success distribution
7. Query boosters that have carried the max payload mass
8. Query landing outcome distribution

Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/4_EDA_SQlite.ipynb

EDA with Data Visualization

- Post Data Wrangling, the following charts were plotted to understand the relationship between various features and outcomes and thus identify the features of interest
  - The relationship with Flight number vs Payload on outcome (referenced as “Class”) shows success improving with experience and mass influencing success
  - The relationship between Launch Site and Flight number shows that Launch Site has influenced success and failure
  - The relationship between Launch Site and Payload shows that one site didn’t launch heavy payloads
  - There is a relationship between the type of orbit and successful outcomes as well as types of orbit and flight number and payload

Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/5_EDA_DataViz.ipynb

Build an Interactive Map with Folium

- In the previous step (EDA) there was a relationship between launch sites and successful outcomes.
- Objective was to dig in further to see if this was related to geographic features.

1. Marked out all launch sites on the map
2. Added markers differentiating launch success and failures by site
3. Measured distance to proximities like coastlines, railways, cities, highways to look for patterns

Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/6_Visualization_Folium_Launch_Sites.ipynb

Build a Dashboard with Plotly Dash

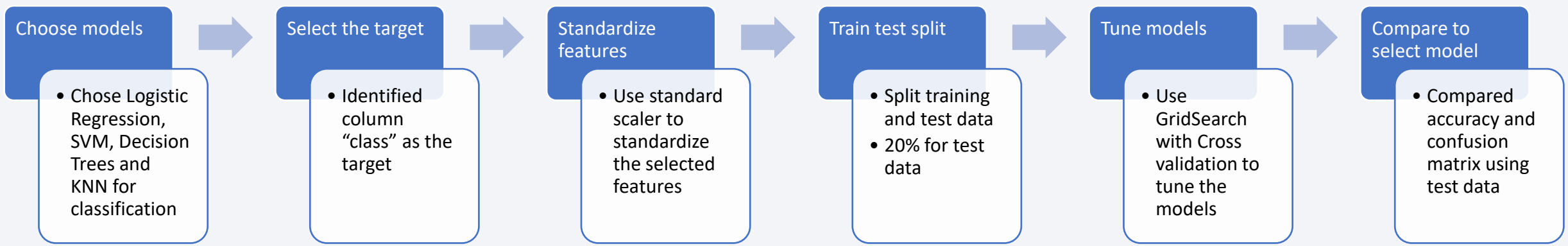
- Created a dynamic dashboard to explore the interaction between sites and launch success rates using a pie chart
- Additionally, added a scatter plot to observe the effect of payload on success
- Further, added ability to drill into these details for different launch sites using a simple drop down for site selection

Github URL:

https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/6_Visualization_Folium_Launch_Sites.ipynb

Predictive Analysis (Classification)

Pg.9 *Output of Data Wrangling – Encoded features and numeric target*

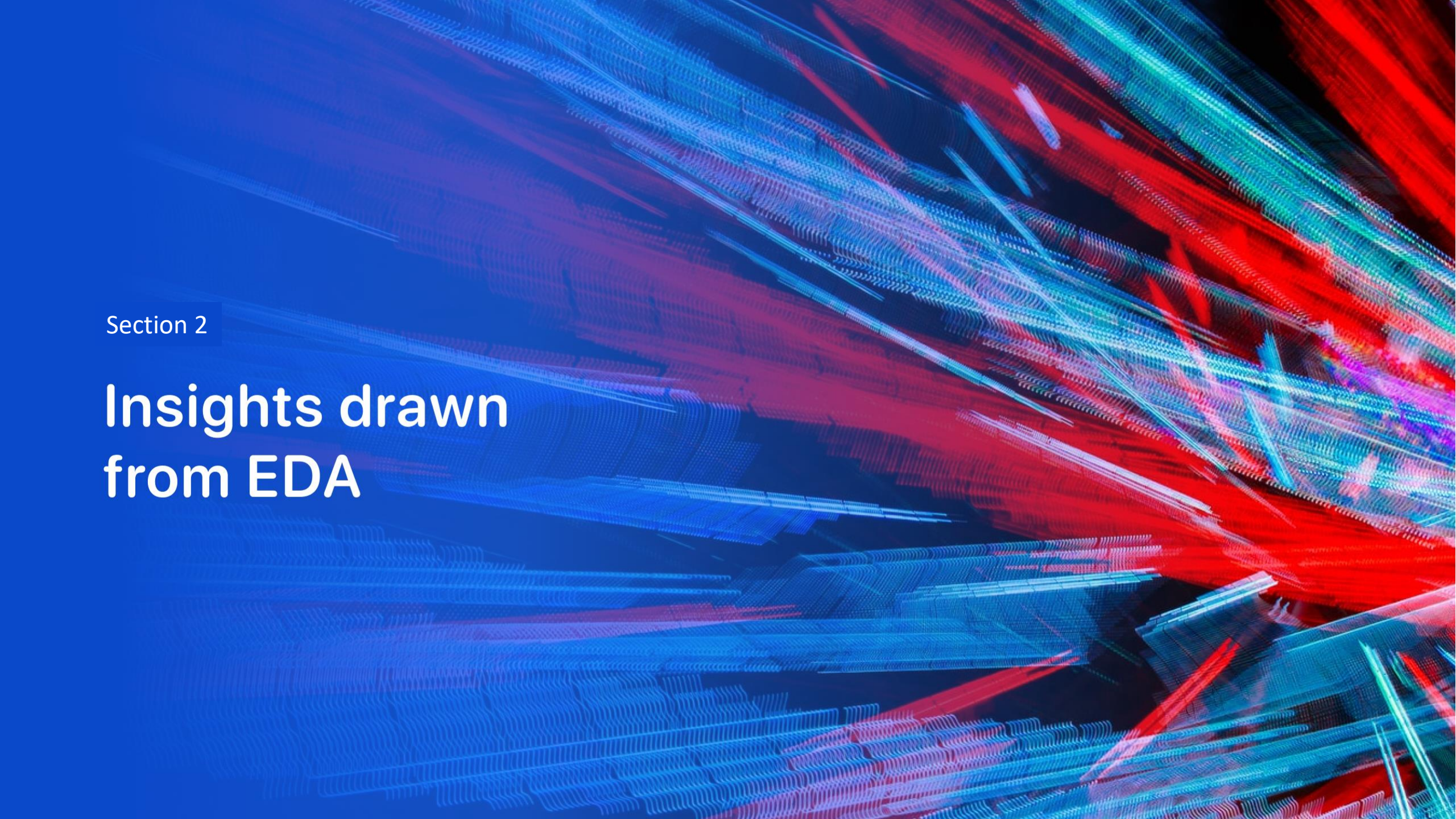


Github URL:
https://github.com/ryanmas13/SpaceY_FinalProject/blob/main/8_ML_Prediction.ipynb

Results

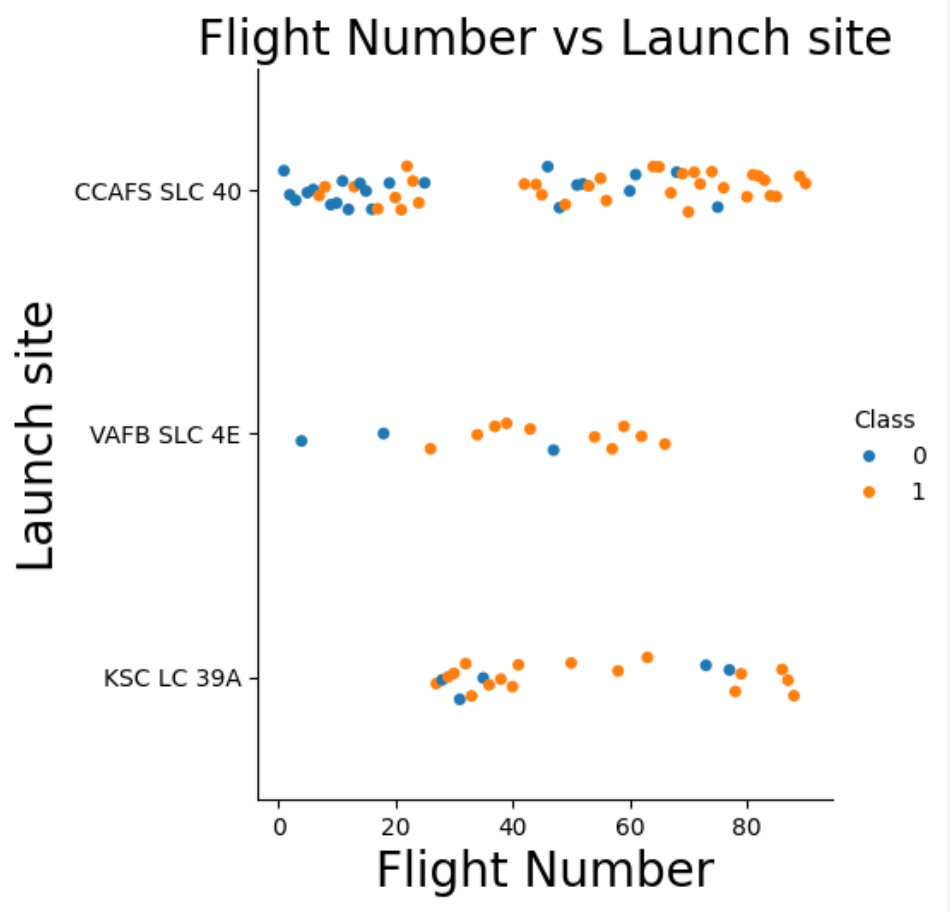
In the next section, we will go through:

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


Section 2
Insights drawn from EDA

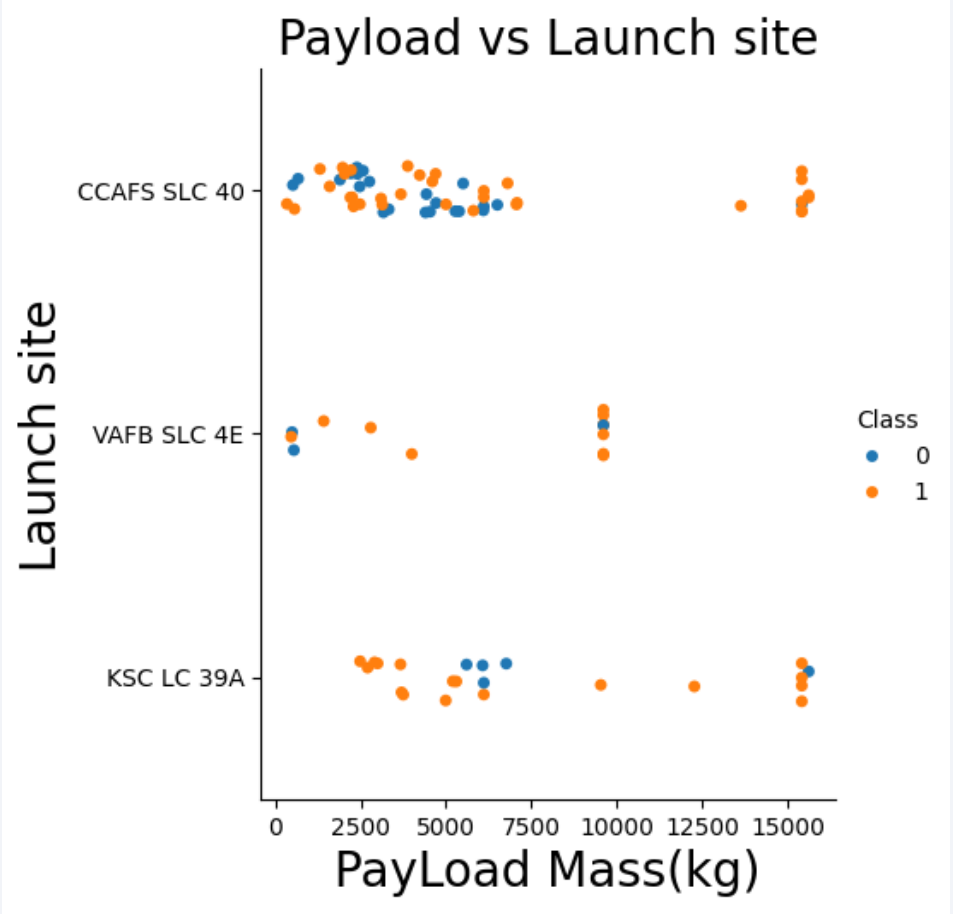
Flight Number vs. Launch Site

- The success rate (“Class”, 0 indicates failure) seems better with KSC and VAFB
- However more than 50% were from CCAFS as well as most of the initial flights so this could be part of the learning curve.
- Also, overall success has improved with experience from all sites.



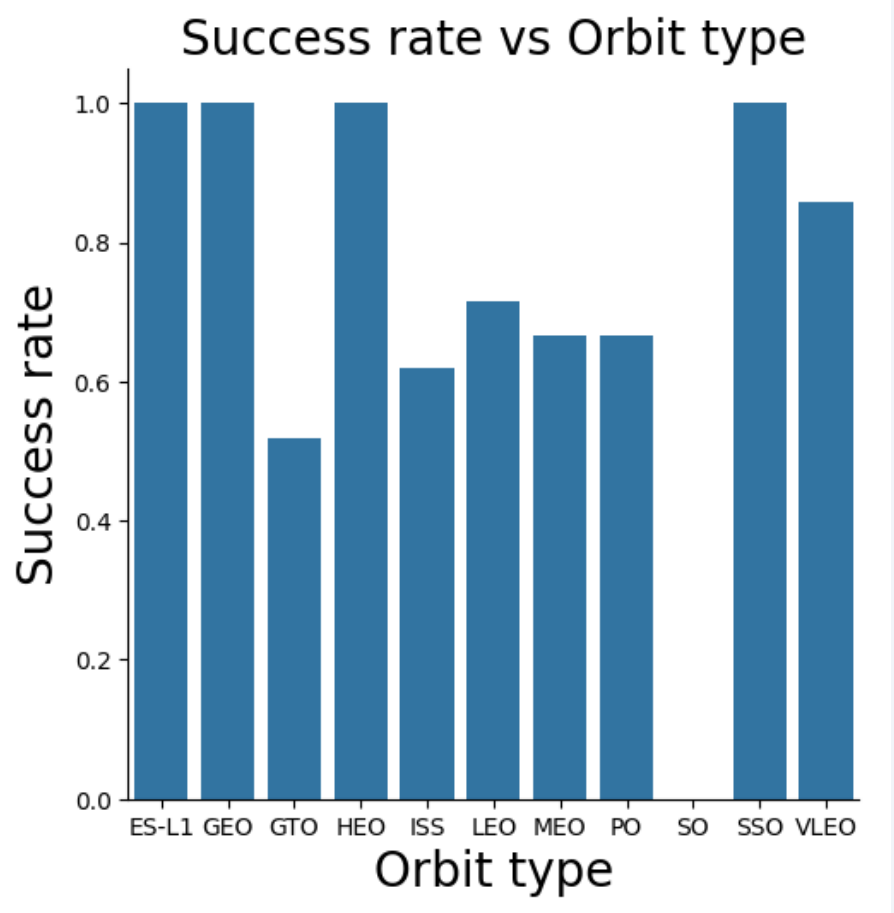
Payload vs. Launch Site

- Generally, most payloads are under 7.5 tons
- Here, KSC and VAFB have been more successful
- Heavy payloads (more than 10 tons) have only been launched from CCAFS and KSC of which CCAFS has had more success



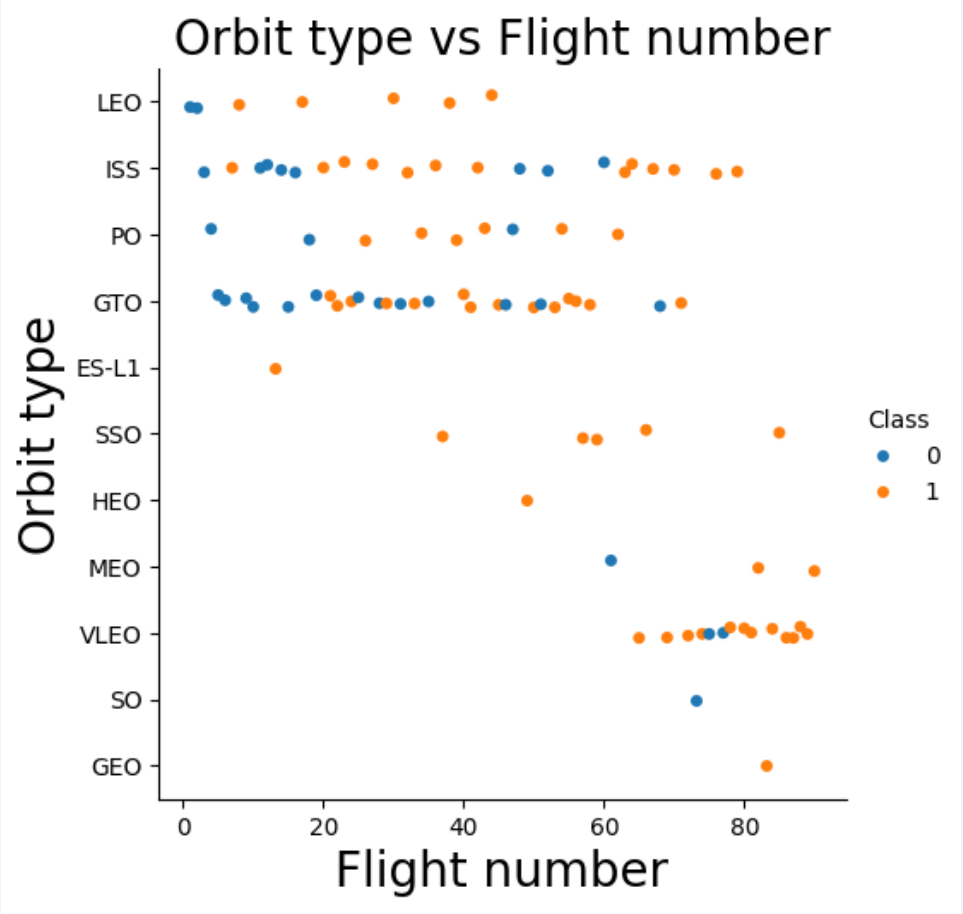
Success Rate vs. Orbit Type

- Out of 11 orbit types, 4 orbits have 100% success
- 6 others are generally mixed but above 50% successful
- Just one, SO, has had no success
- We need to look at flight number distribution by orbit to understand the context of successful performance



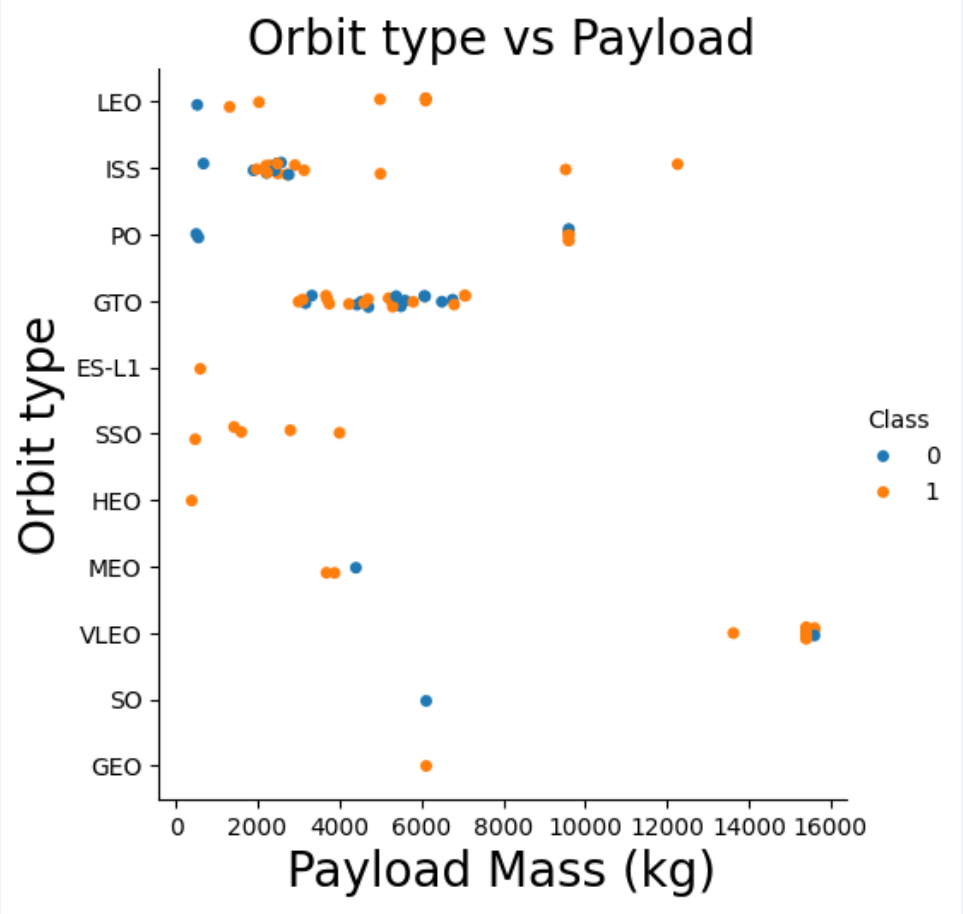
Flight Number vs. Orbit Type

- Orbit types are not evenly distributed over the flight numbers
- Of the 4 successful orbit types from earlier, just SSO has had more than 1 flight, showing consistency
- Later flight numbers have been to other orbits like SSO, VLEO
- While LEO, ISS, GTO are most typical type of orbit used with Falcon 9, spread out across flight numbers
- SO has had only 1 flight and it was a failure, providing context for 100% failure there.



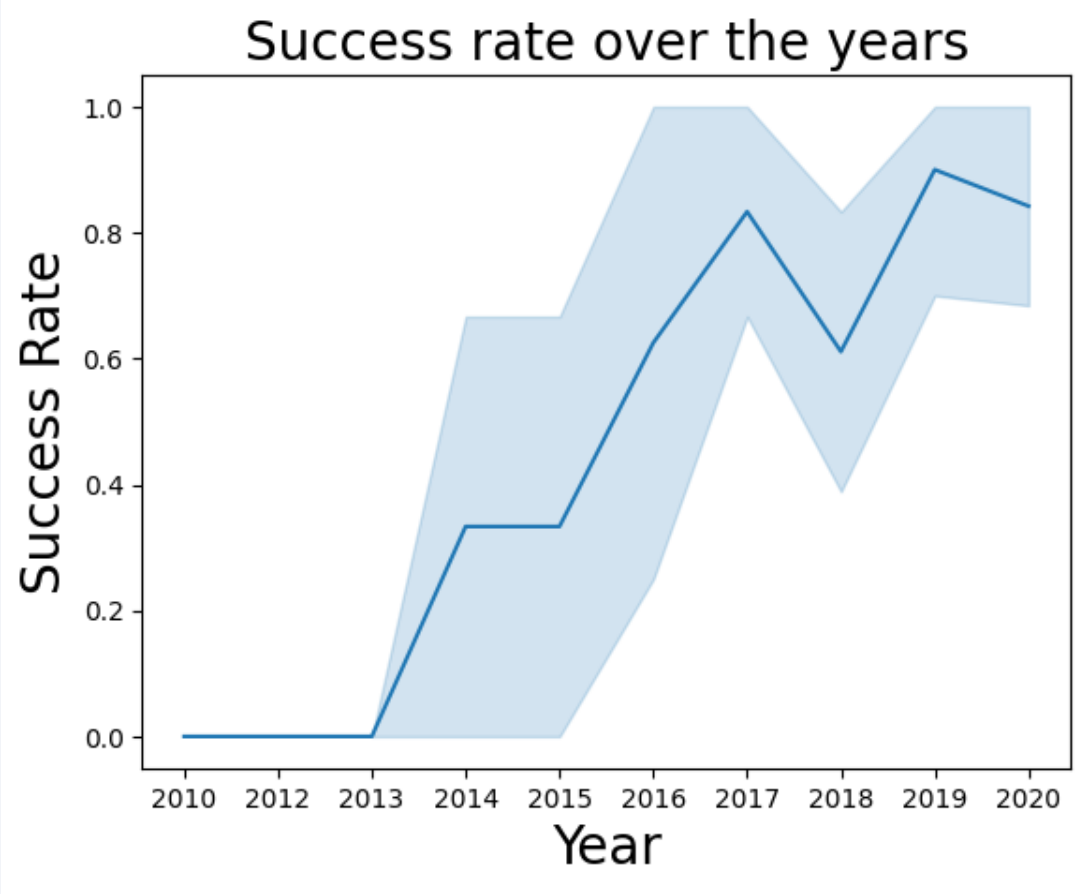
Payload vs. Orbit Type

- With heavier payloads, successful landings have been from ISS, VLEO and PO orbits
- SSO on the other hand has only been with lower payloads
- GTO is mixed with a tighter distribution of payloads



Launch Success Yearly Trend

- The success rate has progressively increased over the years with a plateau between 2014 and 2015 (stable) and a meaningful dip in 2018



All Launch Site Names

- Use DISTINCT, we can query the data table to find the unique list of launch sites

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

```
[32] %sql select DISTINCT(Launch_Site) from SPACEXTABLE;
```

```
 * 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'

- Using “LIKE” with a regular expression we can find sites beginning with CCA
- We can also limit results using “LIMIT”

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

```
%%sql SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

[33] Python

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Using SUM function to query the total payload for a specific customer, “NASA”

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer='NASA (CRS)';
```

[34]

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Used AVG function to calculate the average payload mass carried by booster version F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version='F9 v1.1';
```

[35]

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Use the MIN function to find the first successful landing outcome on a ground pad

```
%%sql
SELECT MIN(Date) FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE 'SUCCESS%GROUND%'
```

[36]

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success%drone%ship%'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__kg_ < 6000;
```

[37]

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*), Mission_Outcome FROM SPACEXTABLE GROUP BY(Mission_Outcome)
```

[38]

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

COUNT(*)	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%%sql
select Booster_Version from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
);
```

[39]

```
 * 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

2015 Launch Records

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

```
%%sql
select
substr(Date,6,2) as Month,
Landing_Outcome,
Booster_Version,
Launch_Site
from SPACEXTABLE
where substr(Date,0,5) = '2015'
AND Landing_Outcome LIKE 'Failure%'
```

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

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

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

- 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

```
%%sql
Select count(Landing_outcome) as Count, Landing_Outcome from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Count DESC
```

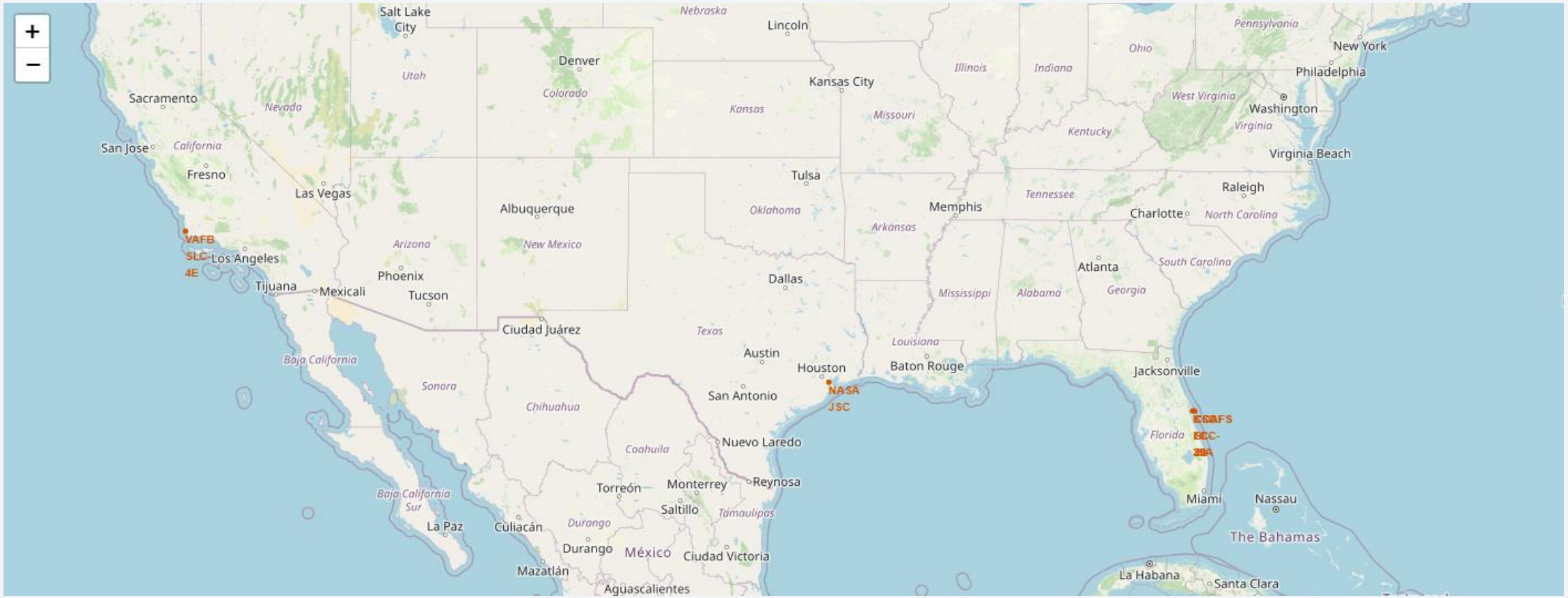
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

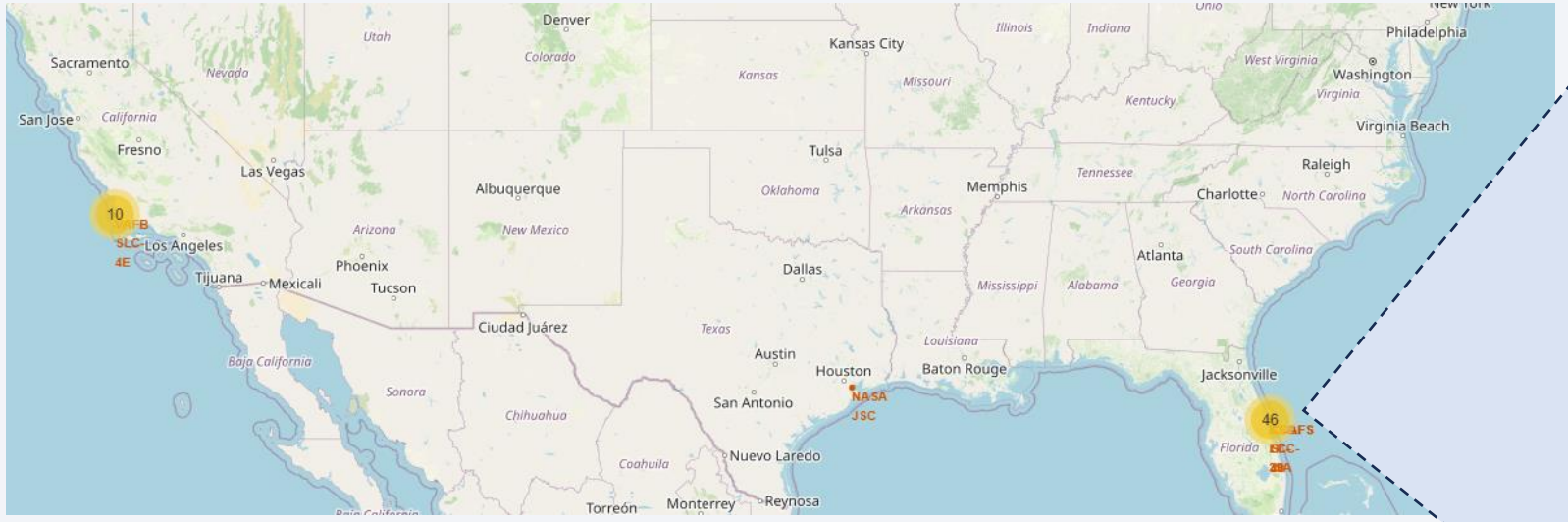
Launch Sites Proximities Analysis

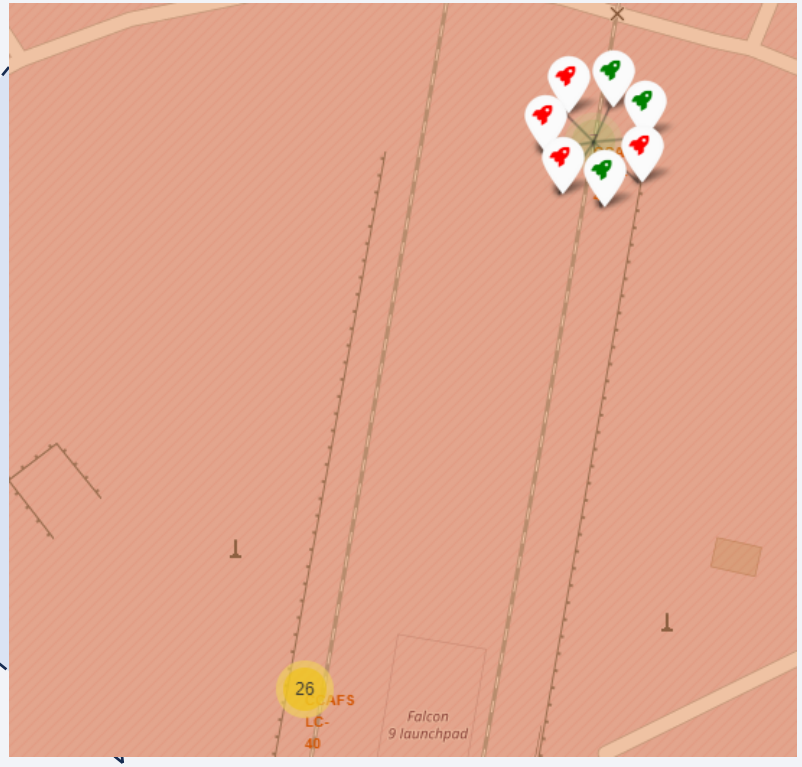
Launch sites



Launch sites are closer to the equator (south) and closer to the coastline

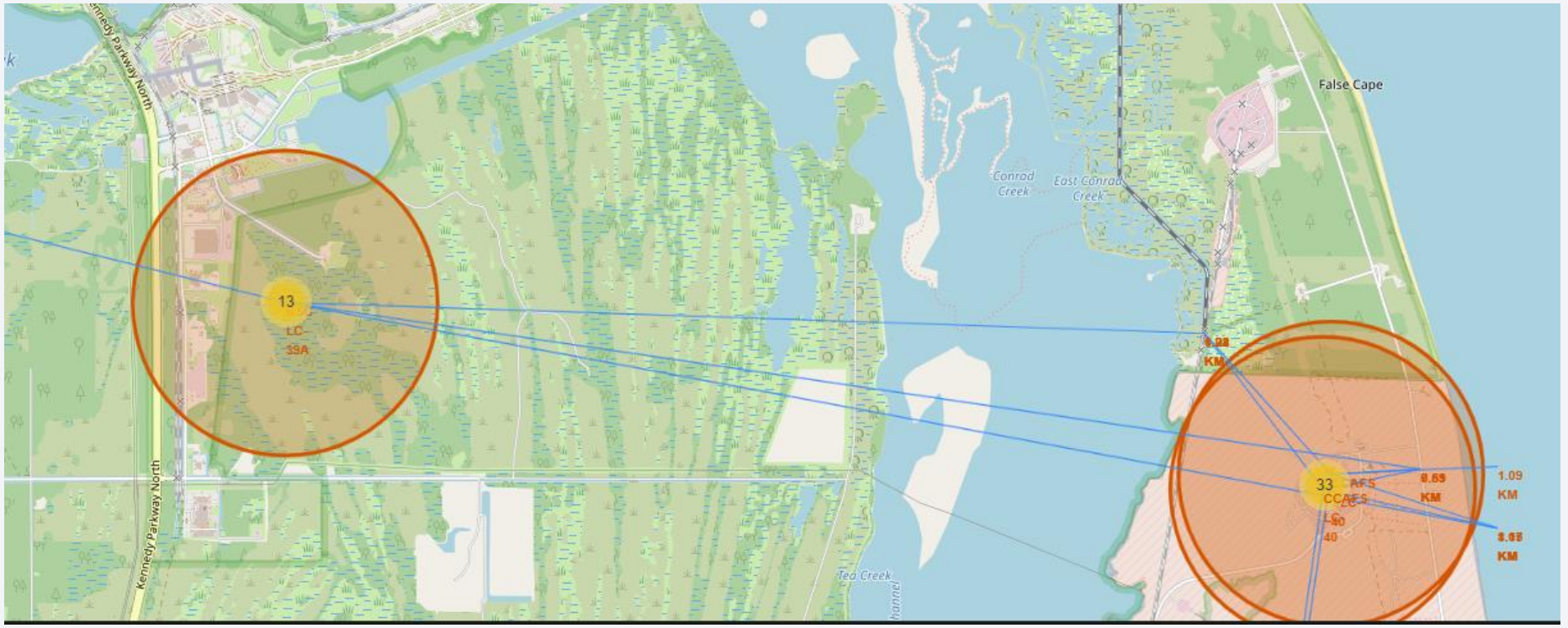
Launch outcomes plotted to launch sites



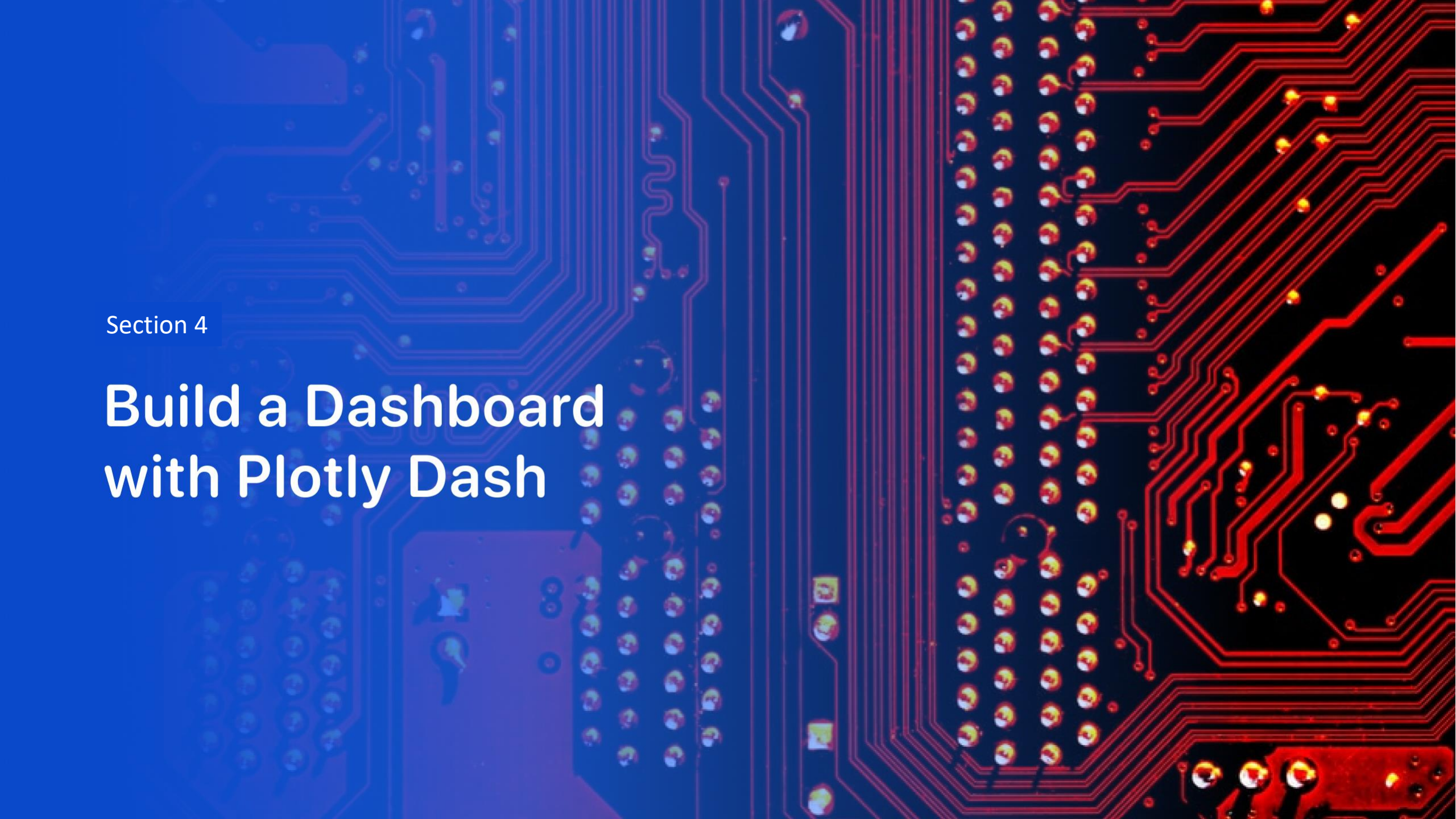


- The launch outcomes are mapped to each site with a marker cluster object.
- This allows us to see the success rate of a launch site by zooming into the map

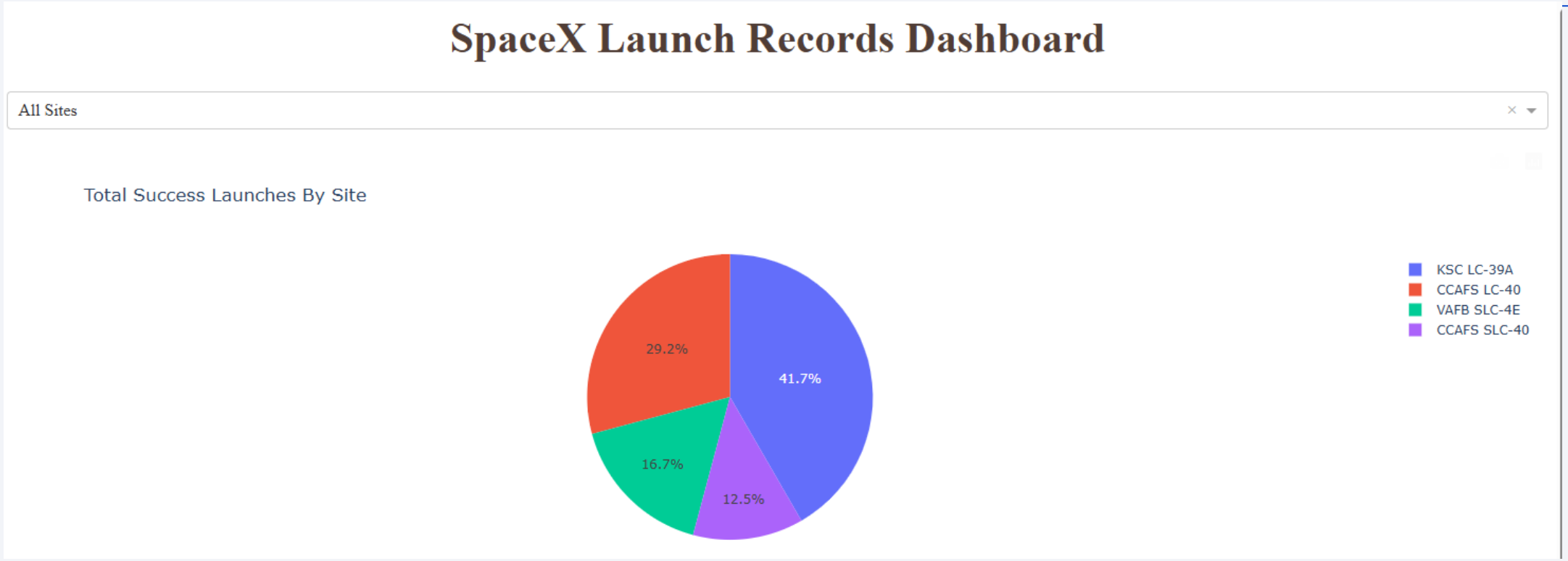
Analyzing proximities to launch sites



- We can explore the generated folium map around Cape Canaveral and Kennedy Space Center to see its proximity to railway lines, highways and coastlines
- Looking across sites, they have similar elements of proximity to coastlines as well as supply routes through rail and road but further away from major cities.

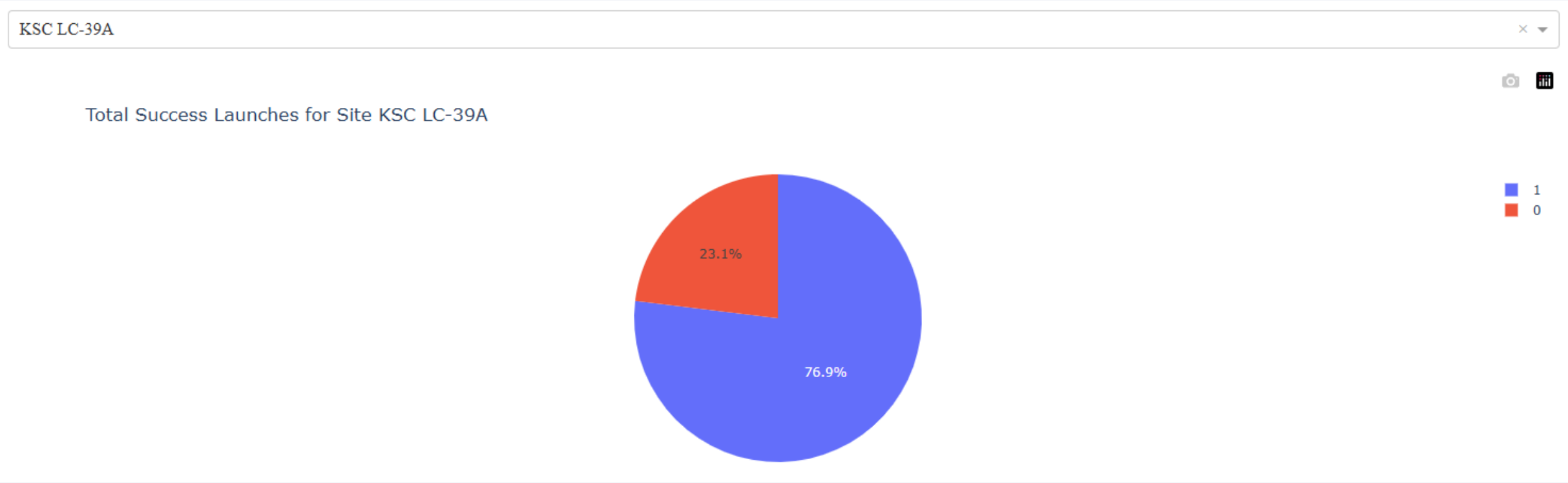

Section 4
Build a Dashboard with Plotly Dash

Launch success by Site



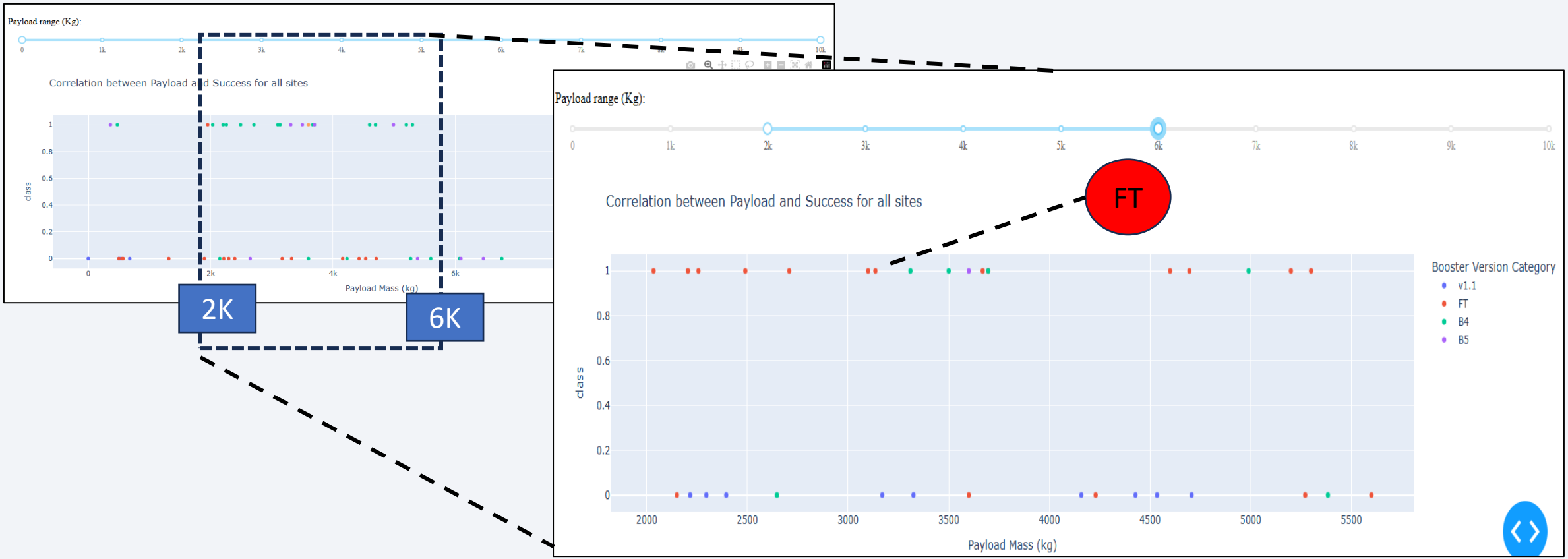
For the default, All Sites option, the pie chart shows the proportion of success by site.

Kennedy Space Center has the highest success ratio



Selecting Kennedy Space from the dropdown shows the breakdown of success vs failure of that site

Payload vs Launch Outcome (class)



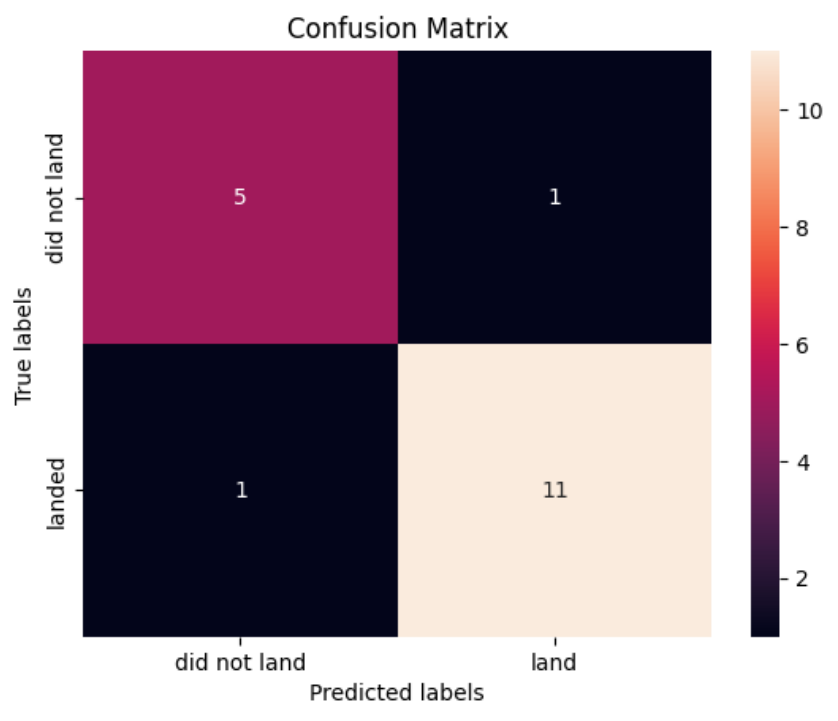
The 2 ton to 6 ton range of payload shows the highest concentration of successes with the FT version of the booster having the most successes

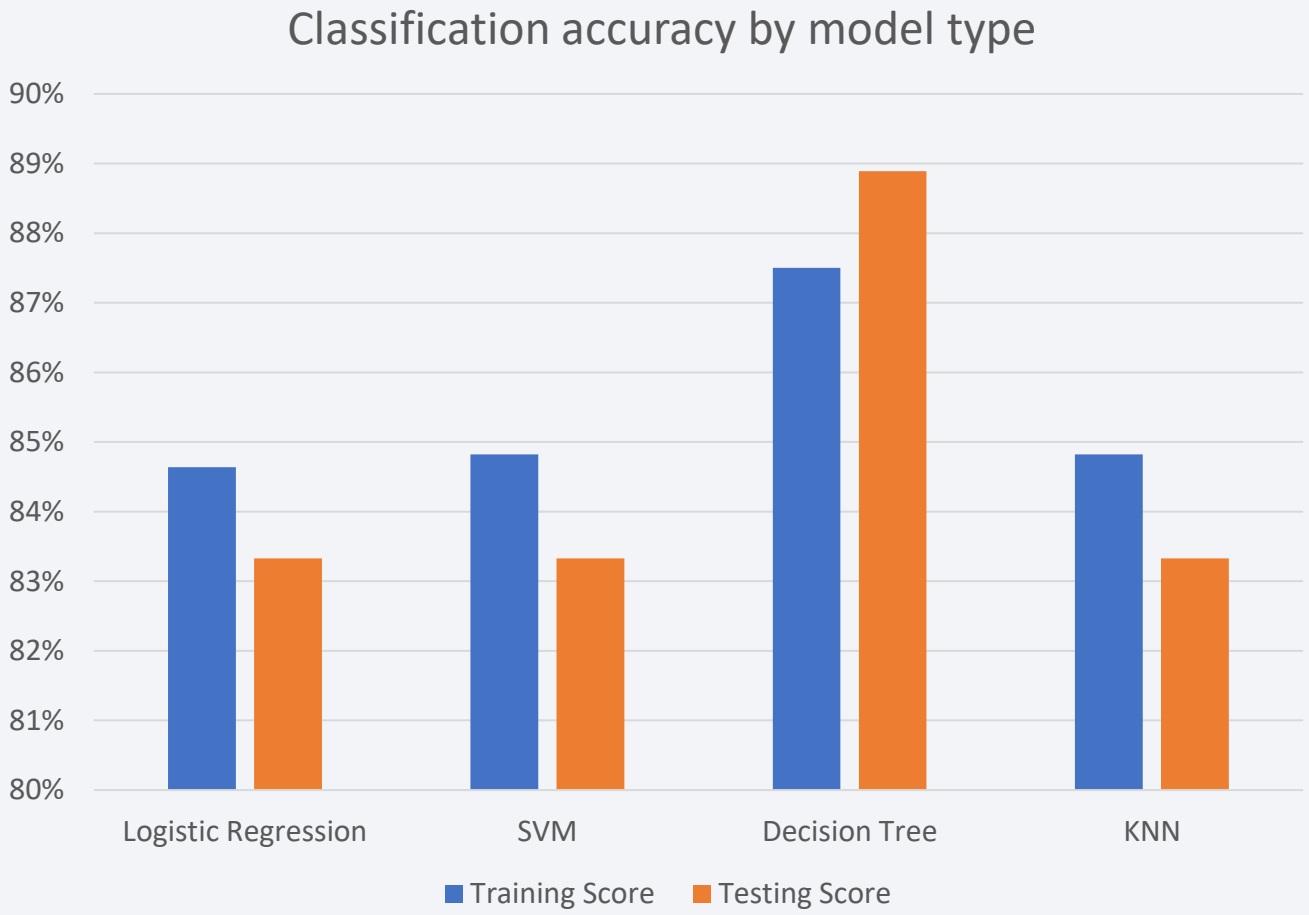
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We find that in this case, Decision Tree had the best score in predicting outcomes
- Confusion Matrix for Decision tree below





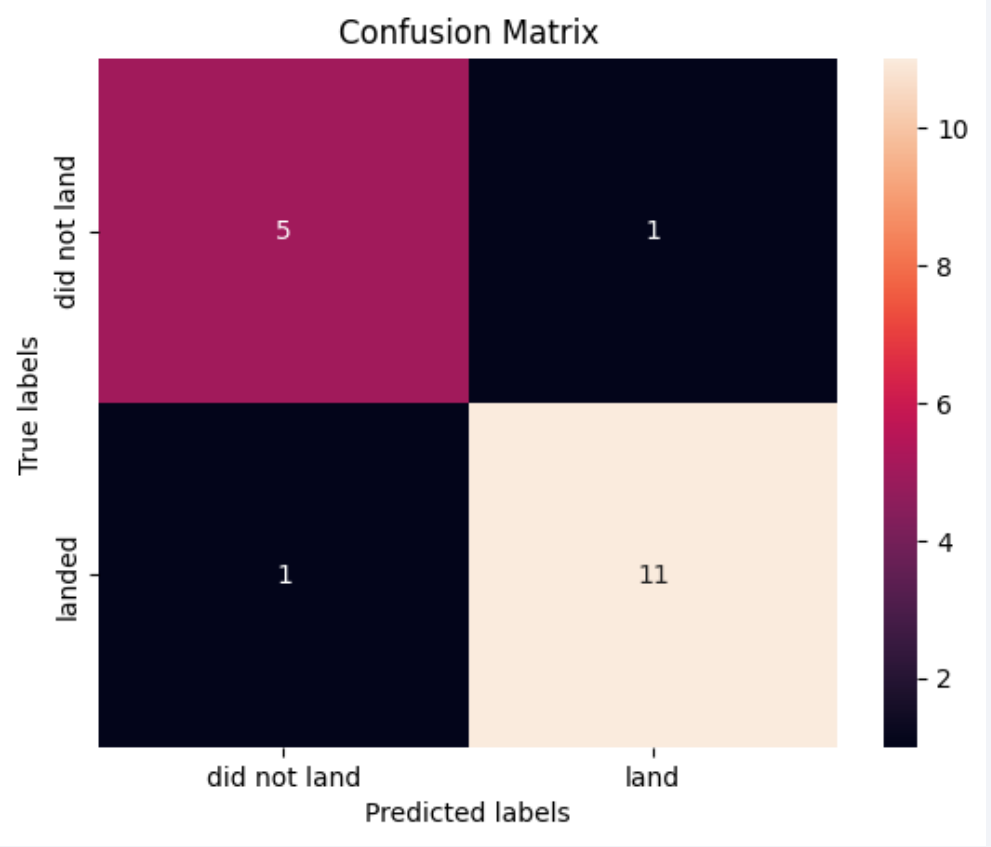
Confusion Matrix

Decision Tree was the best performing model.

We can see that there is only 1 false positive and one false negative implying high precision and recall

Classification Report for Decision Tree

	precision	recall	f1-score	support
0	0.83	0.83	0.83	6
1	0.92	0.92	0.92	12
accuracy			0.89	18
macro avg	0.88	0.88	0.88	18
weighted avg	0.89	0.89	0.89	18



Conclusions

- We observe that SpaceX launches have improved in their success rate over time indicating re-use of Falcon 9 rockets
- Analyzing over various features like payload, launch sites and orbit types, shows us that these features play an important role in whether the launch in a success
- E.g. Most successes from 2 ton to 6 ton range using FT booster type
- Moreover, we observe that launch sites tend to be near coastlines and certain orbit types have a higher chance of success
- Further, we can build a prediction model using Decision Tree to predict with good accuracy the outcome of future launches
- This information would be helpful to a competitor to selectively bid on launches with a high chance of success

Appendix

- Public data link for webscraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!