Winning Space Race with Data Science

Arya Katebian
2024/09/07

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
    - API
    - Web Scraping
  - Exploratory Data Analysis (EDA)
    - SQL
    - Visualization
  - Interactive Visual Analytics and Dashboards
    - Folium
    - Plotly Dash
  - Predictive Analysis (Classification)

- Summary of all results
- SpaceX has gotten better at launching successfully over time, with the overwhelming number of failures happening early in their process
- SpaceX launches primarily from California and Florida
- KSC LC-39A has the highest success rate for launches at nearly 77%
- All predictive models had the same level of classification accuracy

Introduction

- Project background and context
  - Our firm, SpaceY, will be bidding against our main competitor, SpaceX, to procure government contracts for the launching of rockets into space
  - SpaceX dominates this field by producing substantially cheaper launches with a cost of $62 million while competitors produce costs of $165 million per launch
  - SpaceX is able to produce cheaper launches by its ability to reuse rockets in the first stage of launching, eliminating the need to produce new rockets each time
  - We will use data science to examine the Falcon 9's launch success rate to determine the cost of a launch
- Problems you want to find answers
  - How has SpaceX improved with launches over time?
  - What is the most successful launch site?
  - How can we predict the success rate of future launches and from which site should they be launched from?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Data was collected using SpaceX's API and web scraping from SpaceX's Wikipedia
- Data Wrangling
  - A dataset for each launch and relevant variables was processed, cleaned, and prepared primarily examining the success and fail rates of rocket launches
- Exploratory data analysis (EDA) using visualization and SQL
  - The SpaceX dataset was loaded into a table and SQL was used to develop the analysis
  - Categorical variables were converted using one-hot encoding to prepare the data for a machine learning model predicting successful launches in the first stage
  - Pandas and Matplotlib were used to visualize the successful and failed launch data

Methodology

Executive Summary

- Interactive visual analytics using Folium and Plotly Dash
  - Folium was used to analyze launch site geo data
  - A Plotly dashboard was developed to show visual patterns on the success rate of each type of launch
- Predictive analysis using classification models
  - A machine learning pipeline was used to predict the landing success rate during the Falcon 9's first stage

Data Collection

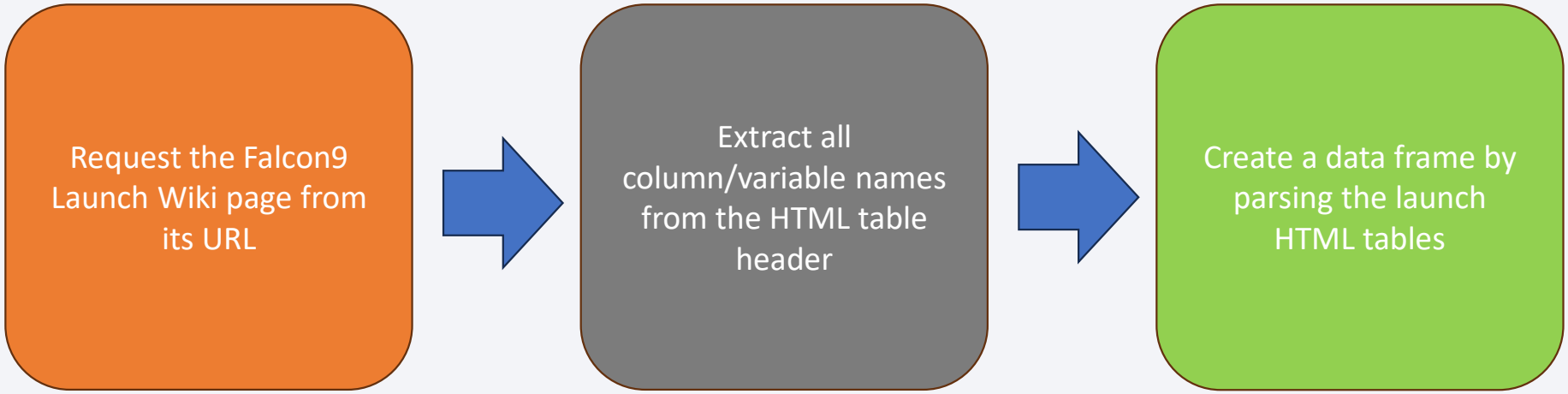
- Data was collected using SpaceX's REST API and webscraping SpaceX's Wikipedia and generated into CSV files

Data Collection – SpaceX API

- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

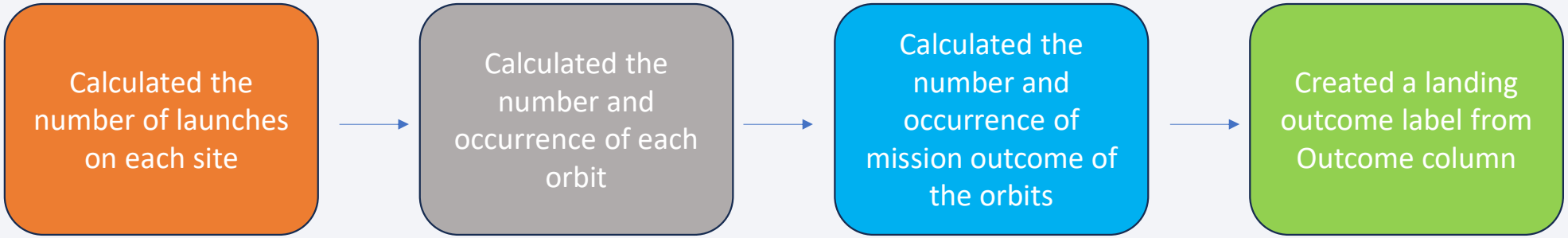
Data Collection - Scraping

- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- A dataset for each launch and relevant variables was processed, cleaned, and prepared primarily examining the success and fail rates of rocket launches
- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plots, bar chats, and line charts were developed to show the relationship between the following variables:
  - The Flight Number and Payload Mass
  - The Flight Number and Launch Site
  - Payload Mass and Launch Site
  - The success rate of each Orbit type
  - Flight Number and Orbit Type
  - Payload Mass and Orbit Type
  - The launch success over time
- This data was visualized to show SpaceX's launch success rate over time and the launch success behind the decisions they make to use specific sites based on their total payload
- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Displayed the names of the unique launch sites in the space mission
  - Displayed 5 records where launch sites began with the string 'CCA'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date when the first successful landing outcome in ground pad was achieved.
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes
  - Listed the names of the booster versions which have carried the maximum payload mass
  - Listed the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
  - Ranked 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: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

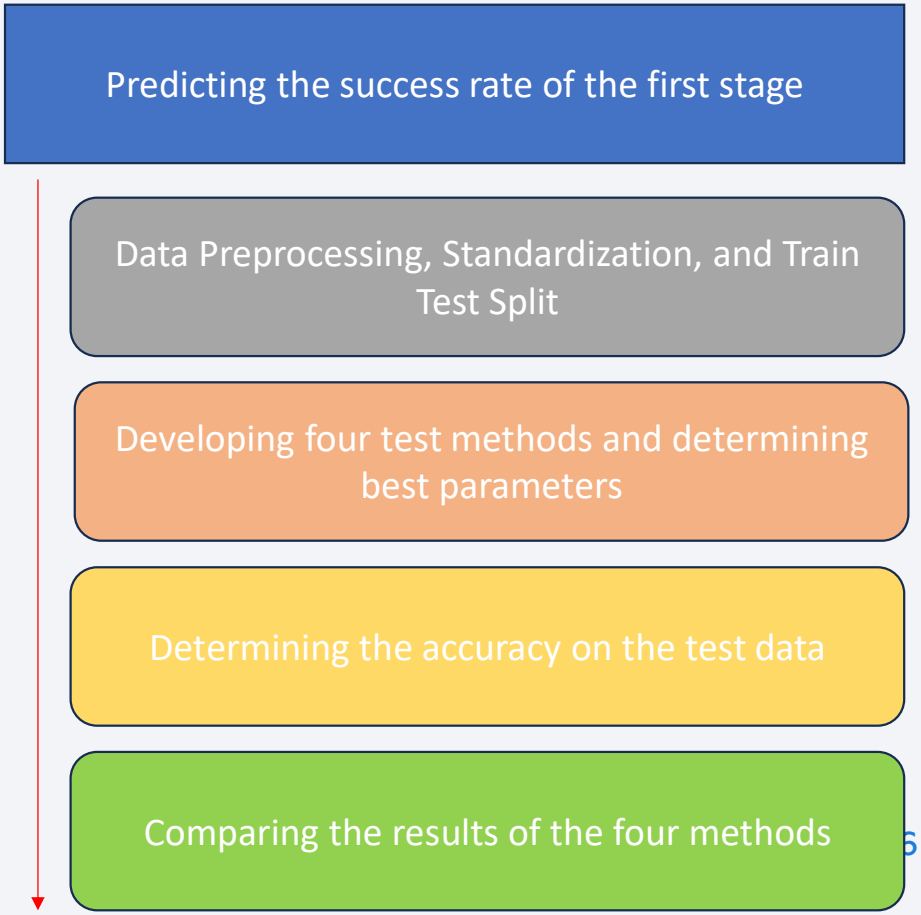
- Markers were developed showing the locations of the launch sites on an interactive map, including a text box with names of the launch sites when clicked
- Successful launches were tagged in Green, failed launches in Red
- Distances from a launch site to the nearest coastline and city were drawn with a blue line with the associated value displayed in km.
- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A Pie chart was developed to show the share of launches for each site
- Pie charts were developed to show the success rate for each launch site
- A scatter chart was developed for each launch site to show the correlation between payload and success for each booster version, with a sliding scale for the payload range
- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/spacex_dash_app.py

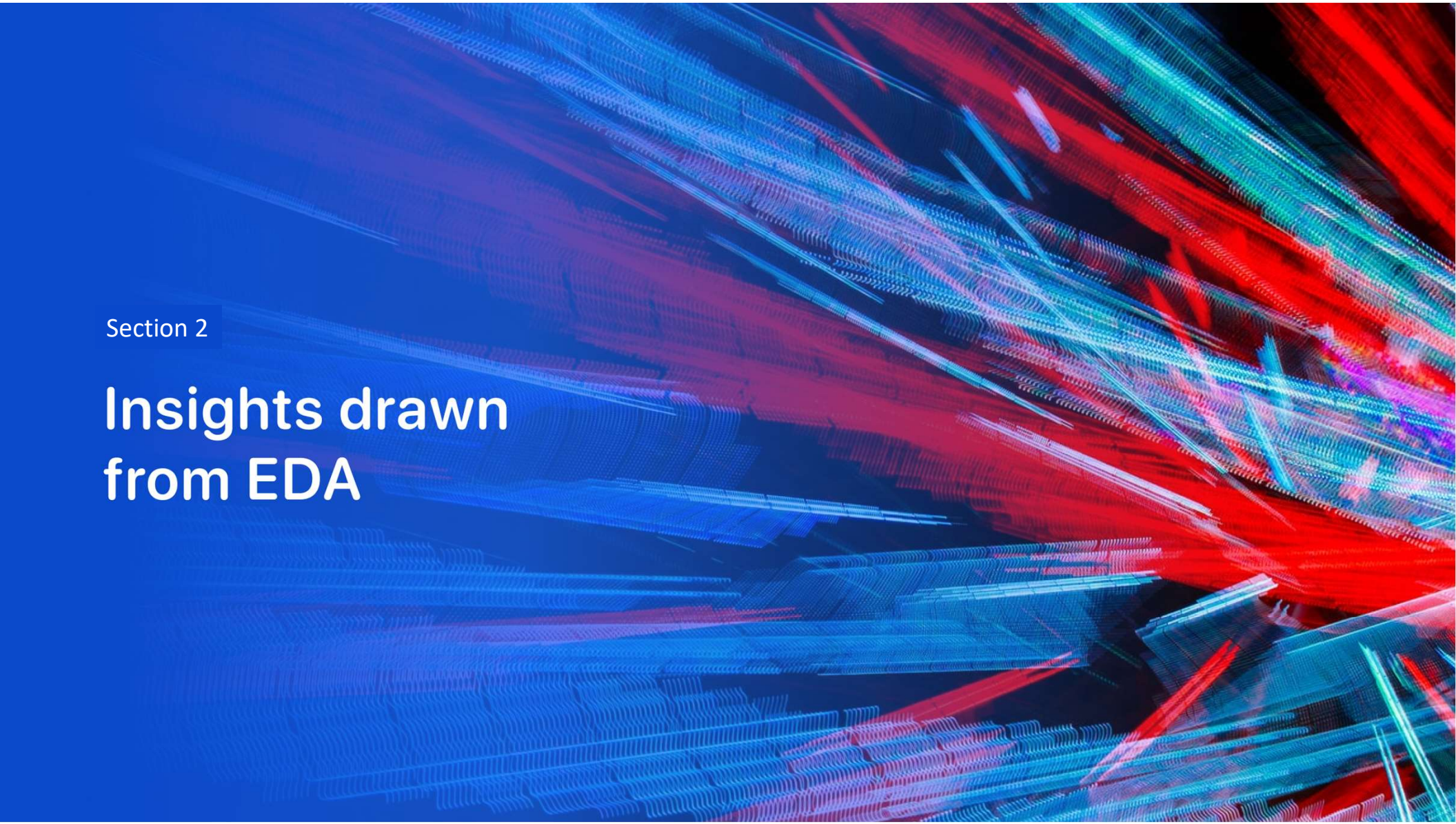
Predictive Analysis (Classification)

- Using test data, the following four methods were used to determine which performed the best classification
  - Logistics Regression
  - Support Vector Machines (SVM)
  - Decision Tree
  - K-Nearest Neighbors
- GitHub URL: https://github.com/bsunak/AppliedDataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

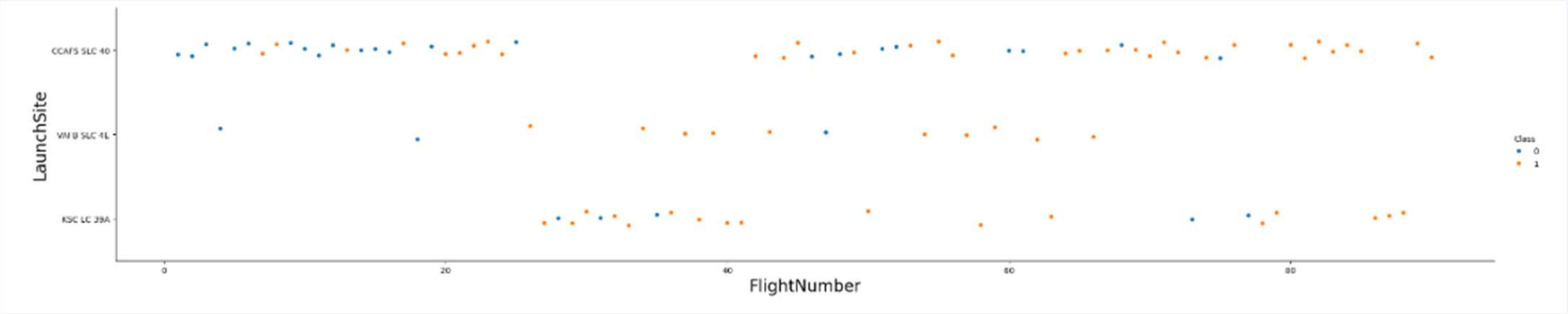
- Exploratory data analysis results
  - As the number of flights increased over time, the success rate went up
- Interactive analytics demo in screenshots
  - SpaceX primarily launches from California and Florida, with Florida the most successful location
- Predictive analysis results
  - All models had the same level of classification accuracy



Section 2

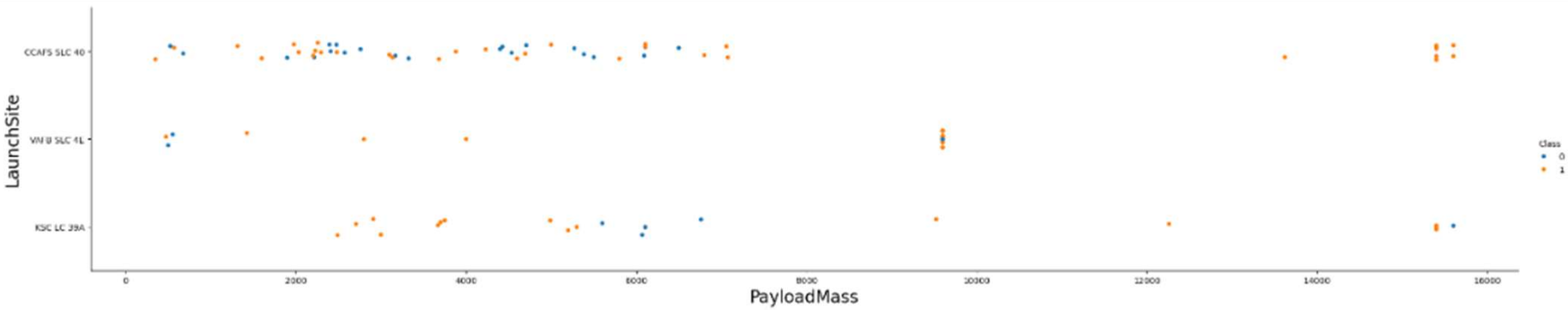
Insights drawn from EDA

Flight Number vs. Launch Site



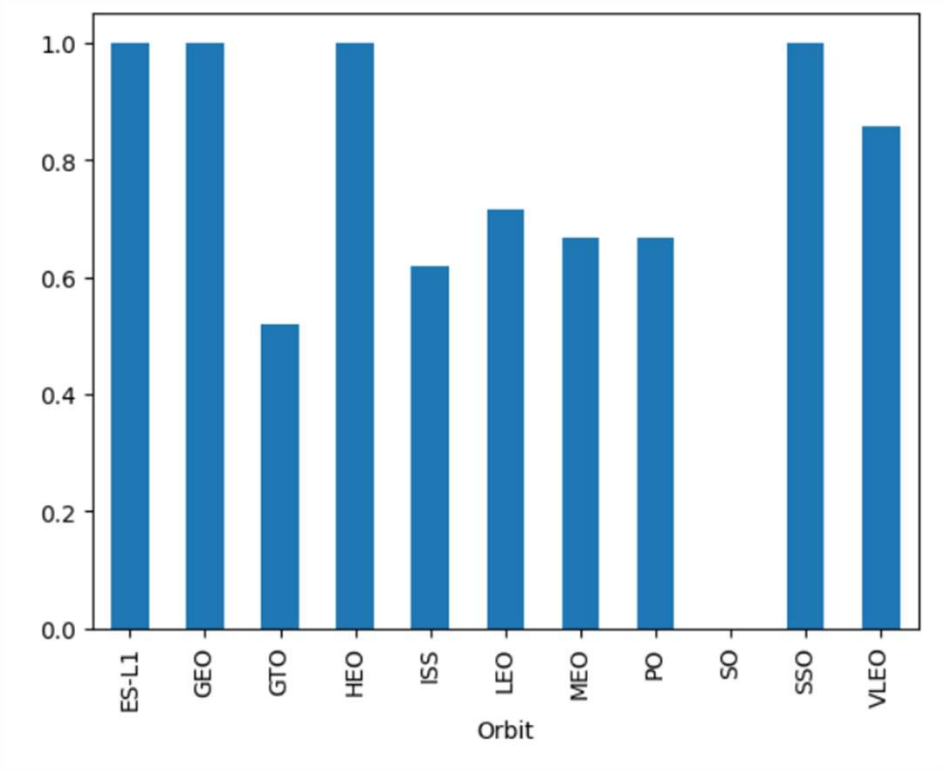
- As the number of flights increased, the success rate for launches went up in general, indicating that SpaceX learned from past mistakes to most recently launch overwhelmingly successful launches
- Early data for CCAFS SLC 40 showed a lack of confidence in the site determining the need to test alternative sites. However, as SpaceX became more successful, the launch site was reused as the primary launch site

Payload vs. Launch Site



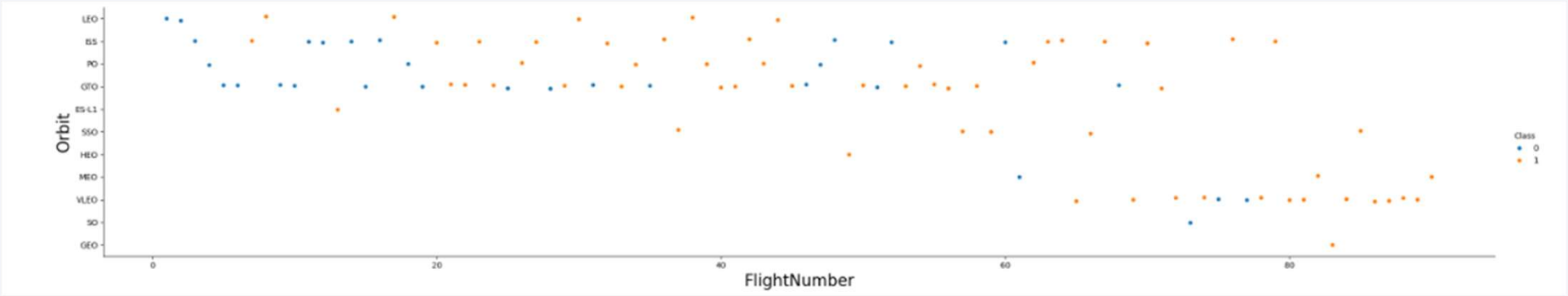
- As Payload Mass increases, the success rate increases as well
- Considering in the prior slide that early launches were failures, it's possible the payloads were intentionally lower to test the feasibility of the rocket launches

Success Rate vs. Orbit Type



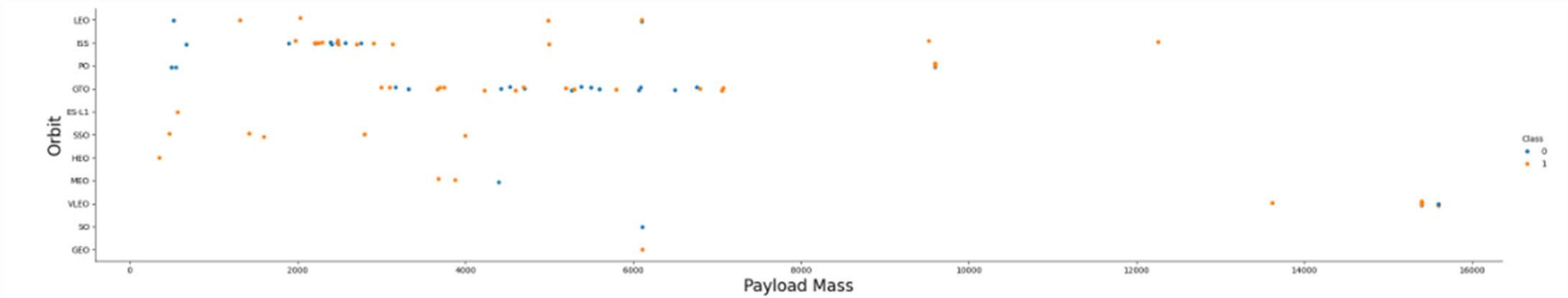
- ES-L1, GEO, HEO, and SSO have perfect success rates
- SO has a 0% success rate
- All other orbits have a success rate of 50% or greater

Flight Number vs. Orbit Type



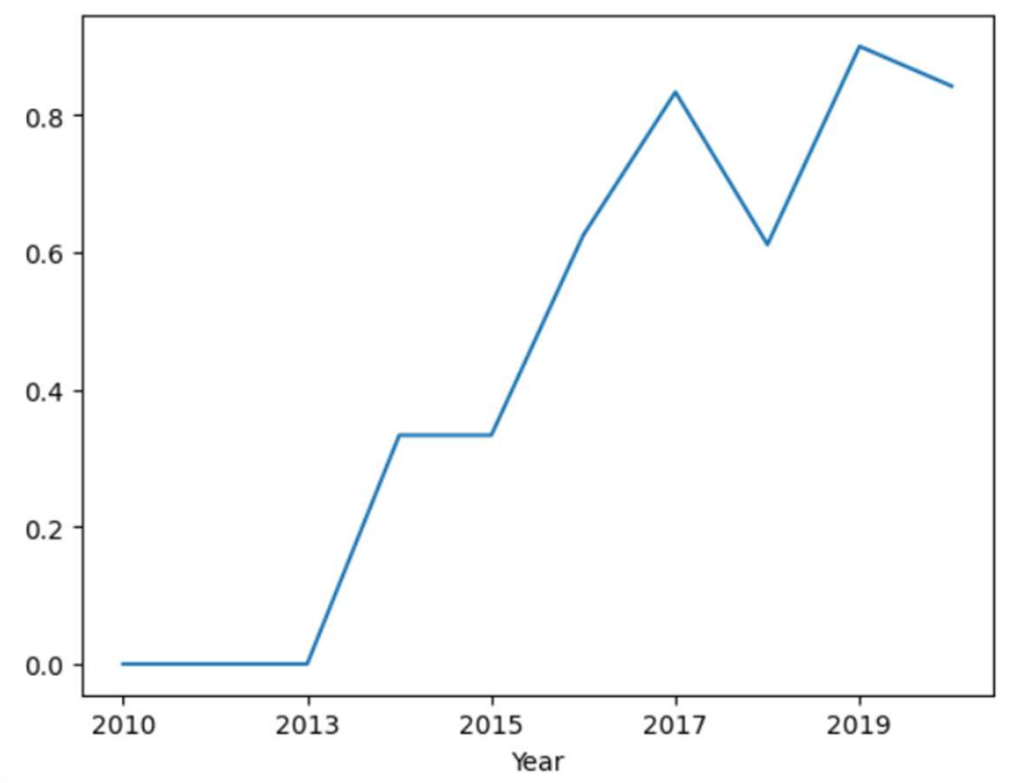
- Again we see the success rate increase over time as the flight number increases
- SO's previously identified success rate of 0% only has a sample size of 1
- Most recently SpaceX primarily launches in the VLEO orbit

Payload vs. Orbit Type



- LEO is successful for most payload masses
- SSO has proven successful for smaller payload masses without information on how heavier masses would fare
- Otherwise, there does not seem to be much correlation between payload mass and orbit type for other sites

Launch Success Yearly Trend



- Demonstrated again here, the success rate has substantially increased over time
- The most recent dip does not seem significant enough to establish a negative trend
- The first successful launch did not occur until after 2013

All Launch Site Names

A query was used to show the unique launch sites from the database

```
In [11]: %sql select distinct launch_site from SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

- A query was used to show 5 launches from a launch site beginning with “CCA”

```
In [12]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

```
In [16]: %sql select sum(payload_mass__kg_) as Total_Payload_Mass from SPACEXTABLE where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[16]:
```

Total_Payload_Mass
45596

- A sum was calculated of the total payload carried by boosters from NASA, finding it to be a total of 45,596 kg

Average Payload Mass by F9 v1.1

```
In [18]: %sql select avg(payload_mass__kg_) as Average_Payload_Mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
 * sqlite:///my_data1.db
Done.
Out[18]:
```

Average_Payload_Mass
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 was calculated and found to be 2,534.67 kg

First Successful Ground Landing Date

In [20]:
```
%sql select min(date) as First_Successful_Landing_Outcome from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

Out[20]:

First_Successful_Landing_Outcome
2015-12-22

- The first successful landing date was found to be December 22, 2015

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
```

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

Out[22]:

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

- Four boosters have successfully landed on the drone ship and had payload mass greater than 4000 but less than 6000
- * The entire line of code in Jupyter Notebook would not fit so a screenshot was taken of the file from VSCode. The output is a screenshot from the Jupyter Notebook file (click to verify)

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT 'Success' AS Outcome, SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS Count FROM SPACEXTABLE GROUP BY Outcome UNION SELECT 'Failure' AS Outcome, S
```

```
, SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS Count FROM SPACEXTABLE GROUP BY Outcome;
```

Python

Done.

Out[27]:

Outcome	Count
Failure	1
Success	100

- There have been 100 successful outcomes and 1 failure outcome
- * The entire line of code in Jupyter Notebook would not fit so a screenshot was taken of the file from VSCode. The output is a screenshot from the Jupyter Notebook file (click to verify)

Boosters Carried Maximum Payload

```
In [28]: %sql SELECT booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

Out[28]:

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

- A list of names of the booster which have carried the maximum payload mass was queried from the database

2015 Launch Records

```
%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 is 'Failure (drone ship)';
```

Python

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

Out[29]:

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

- Two failed landing outcomes in drone ship happened in January and April of 2015
- * The entire line of code in Jupyter Notebook would not fit so a screenshot was taken of the file from VSCode. The output is a screenshot from the Jupyter Notebook file (click to verify)

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

In [32]:
```
%%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.
```

Out[32]:

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

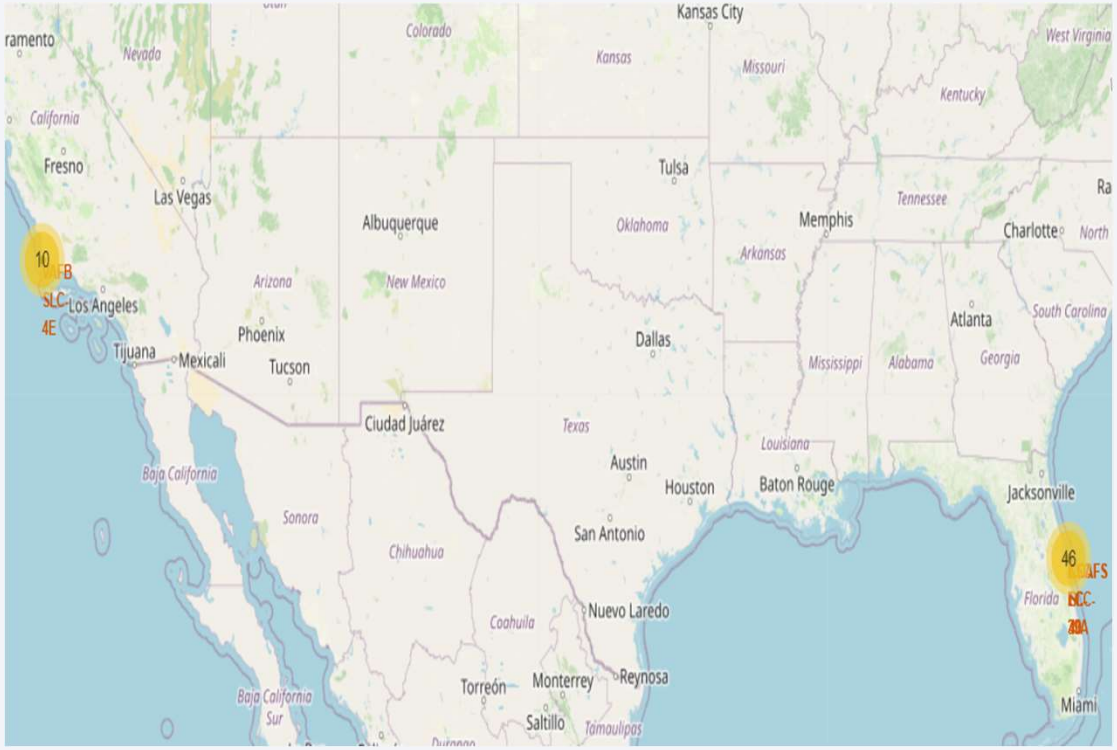
- Every type of successful and failed landing outcome was queried between 2010-06-04 and 2017-03-20, in descending order

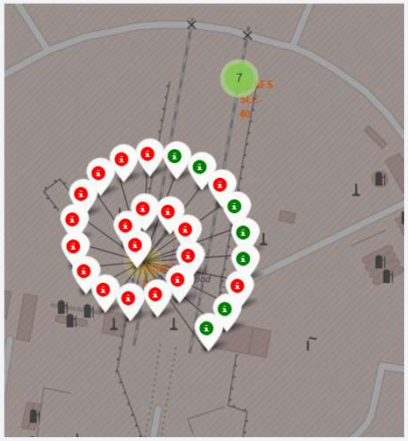
Section 3

Launch Sites Proximities Analysis

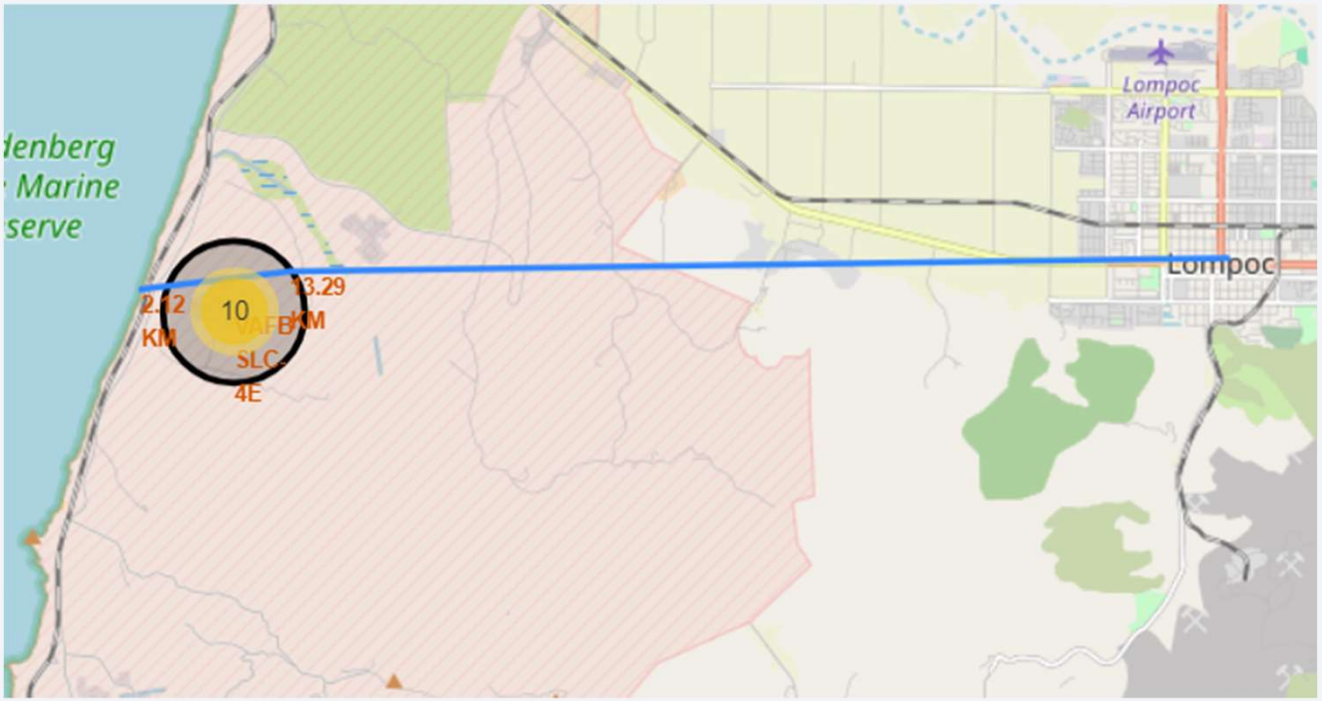
SpaceX primarily launches from California and Florida

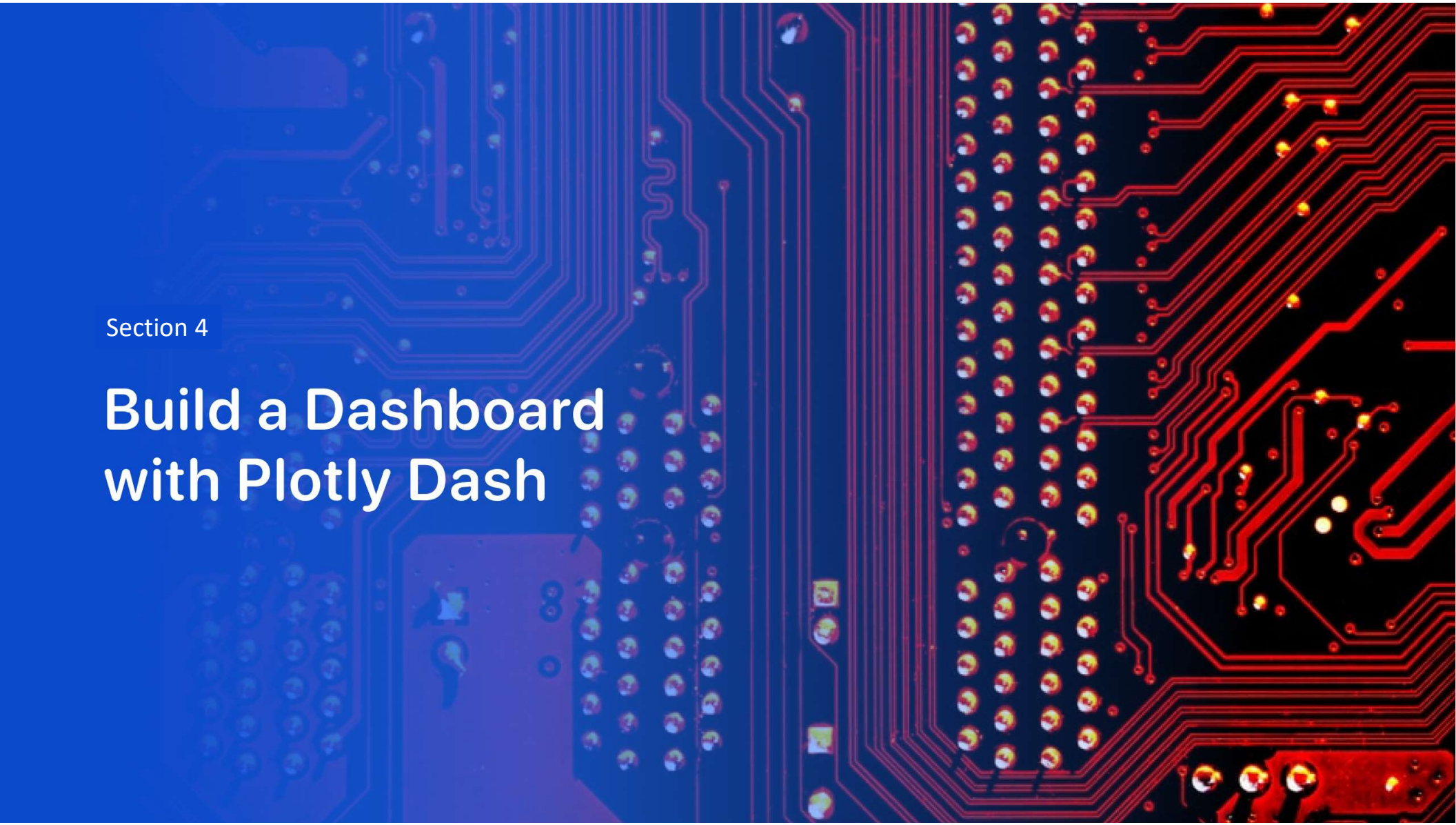
Most launches (46) were in Florida, with the remaining (10) in California. The most successful launch site was KSC LC-39A region while CCAFS SLC-40 region overwhelmingly failed





The California launch site is near the coast but distant from the nearest city

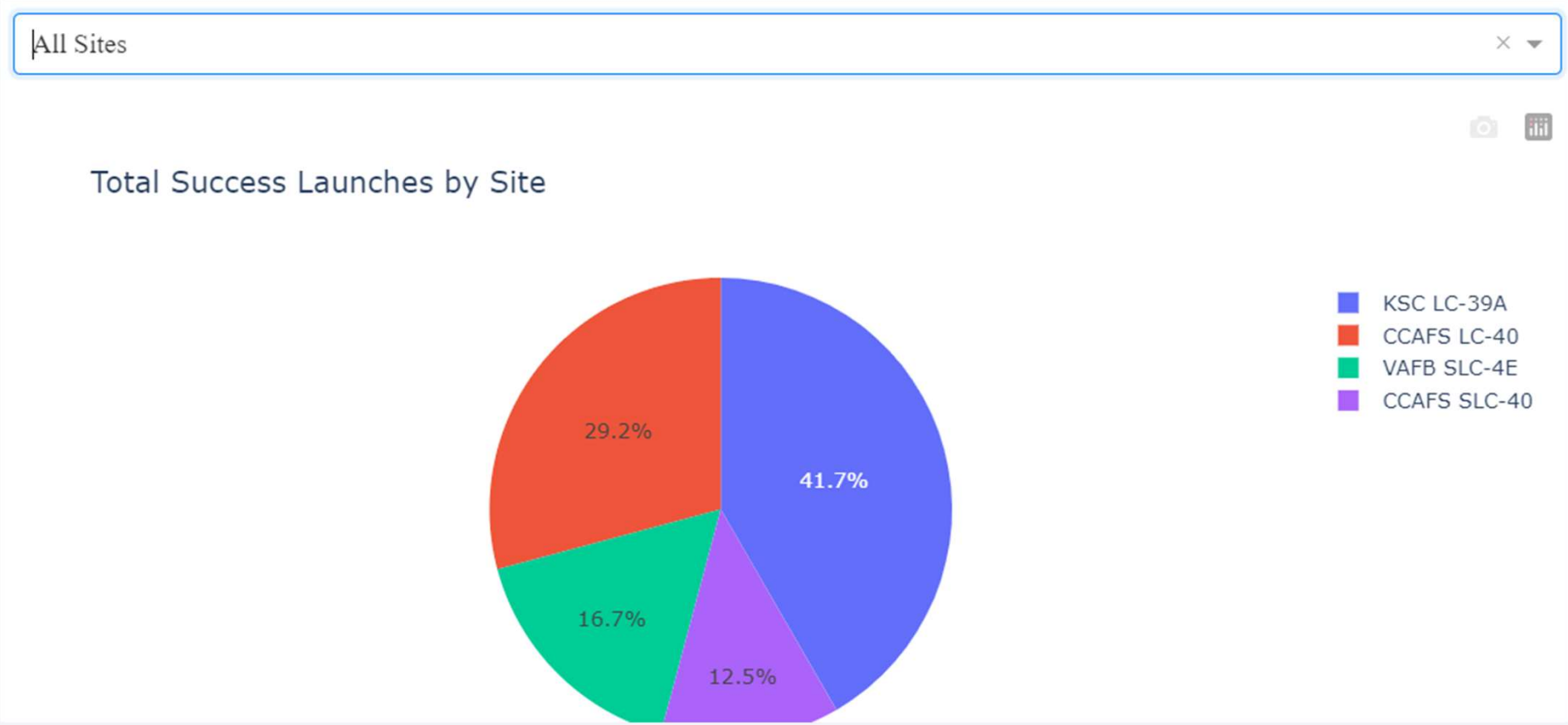




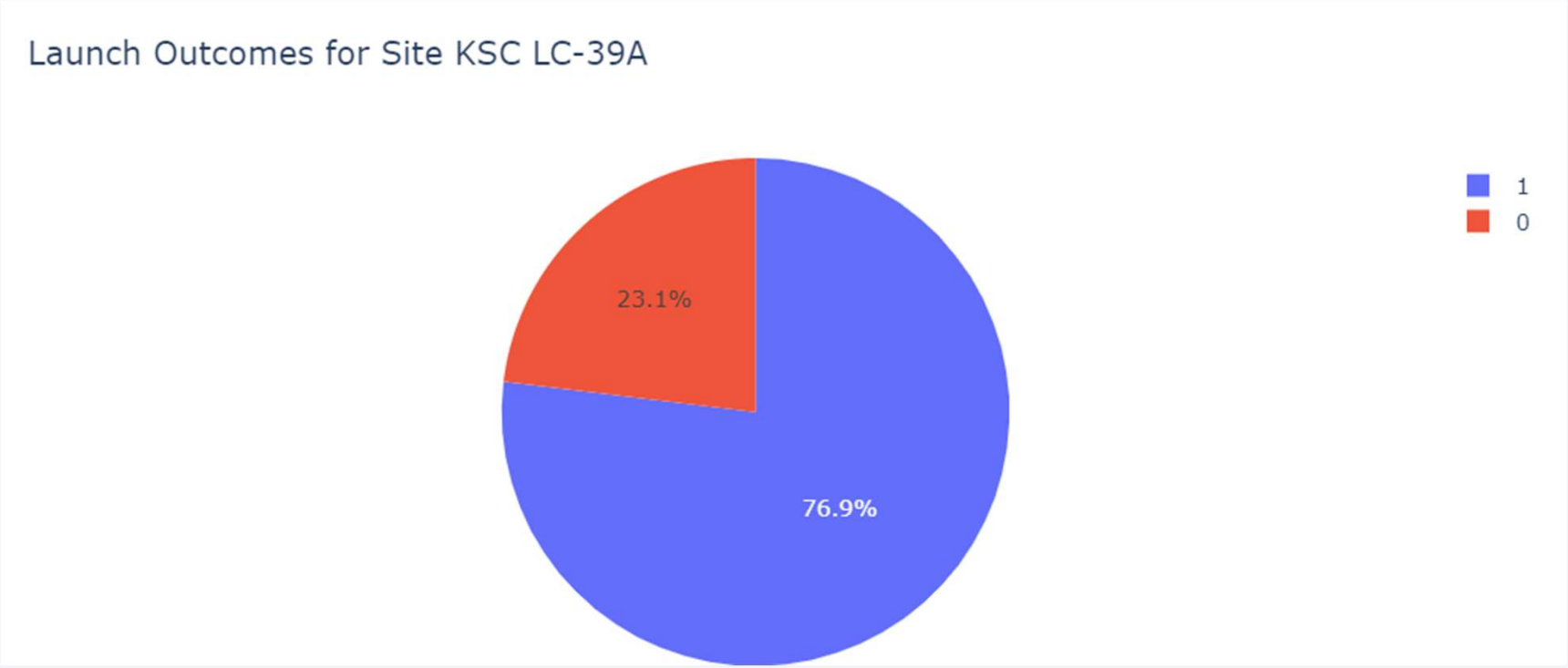
Section 4

Build a Dashboard with Plotly Dash

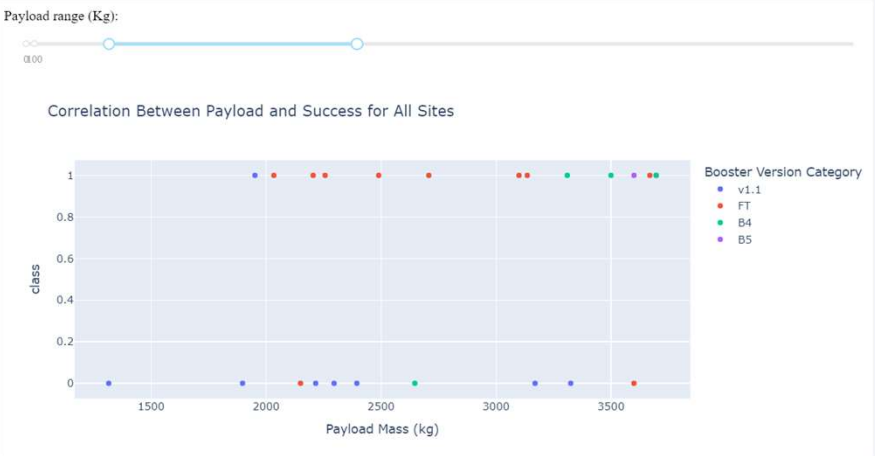
KSC LC-39A is the most successful launch site, with CCAFS SLC-40 the least successful



KSC LC-39A has the highest success rate for launches at nearly 77%



v.1.1 has no correlation between payload and success for all sites



Section 5

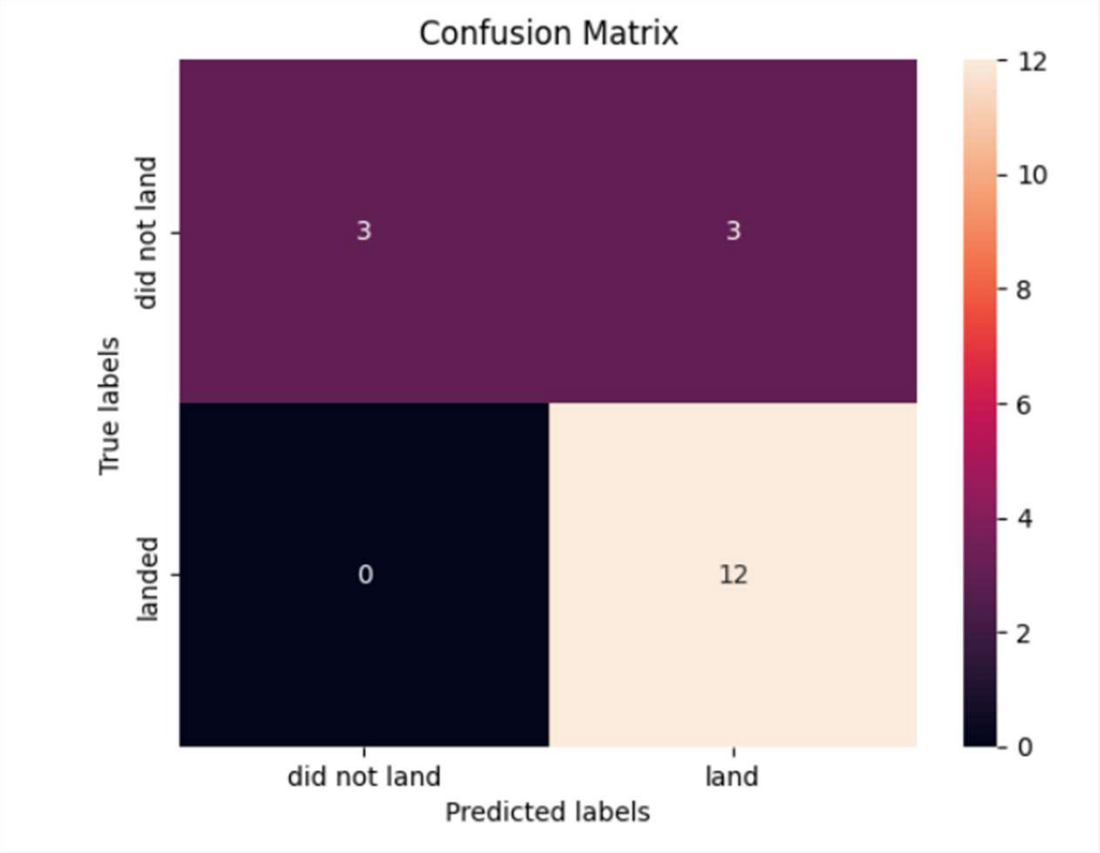
Predictive Analysis (Classification)

Classification Accuracy

- All tests had the same level of classification accuracy at 83.33%

Confusion Matrix

- All models performed exactly the same and had identical confusion matrices
- Provided is an example of K-Nearest Neighbors



Conclusions

- SpaceX has gotten better at launching successfully over time, with the overwhelming number of failures happening early in their process
- SpaceX launches primarily from California and Florida
- KSC LC-39A has the highest success rate for launches at nearly 77%
- All predictive models had the same level of classification accuracy

Thank you!