IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nassim Oulhadj
January 21, 2026

Outline

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

Executive Summary

Summary Of Methodologies

1. Data Collection (through SpaceX API)
2. Data Preparation
    a. Cleaned data
    b. Exploratory Data Analysis with SQL
3. Visualizations
    a. Interactive Launch Map with Folium
    b. Interactive Dashboard with Plotly Dash
4. Predictive Analysis
    a. Implemented Logistic Regression, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN)
    b. Evaluation through confusion matrices, test accuracy, test F1 scores and global F1 scores.

Summary Of Results

1. Model Performance
    a. Test Accuracy
        i. Logistic Regression: 0.833
        ii. SVM: 0.833
        iii. Decision Tree: 0.888
        iv. KNN: 0.833
2. Key Findings
    a. Decision Tree performed best
    b. Success can be impacted by launch site and payload mass

Introduction

Project Background And Context

This capstone project focuses on determining whether SpaceX will successfully recover the Falcon 9 first stage after launch. The competitive advantage SpaceX holds in the commercial space industry stems from first-stage reusability, which dramatically reduces launch expenses compared to traditional providers. Through predictive modeling of landing outcomes, we can assess the true cost of each mission and deliver strategic intelligence to competitors seeking to challenge SpaceX's market position.

Problems To Be Answered

1. What factors affect landing success?
    a. Payload mass, orbits, launch site, etc.
2. Can we predict outcomes using machine learning models?
    a. If so which model performs best?

Section 1

Methodology

Methodology

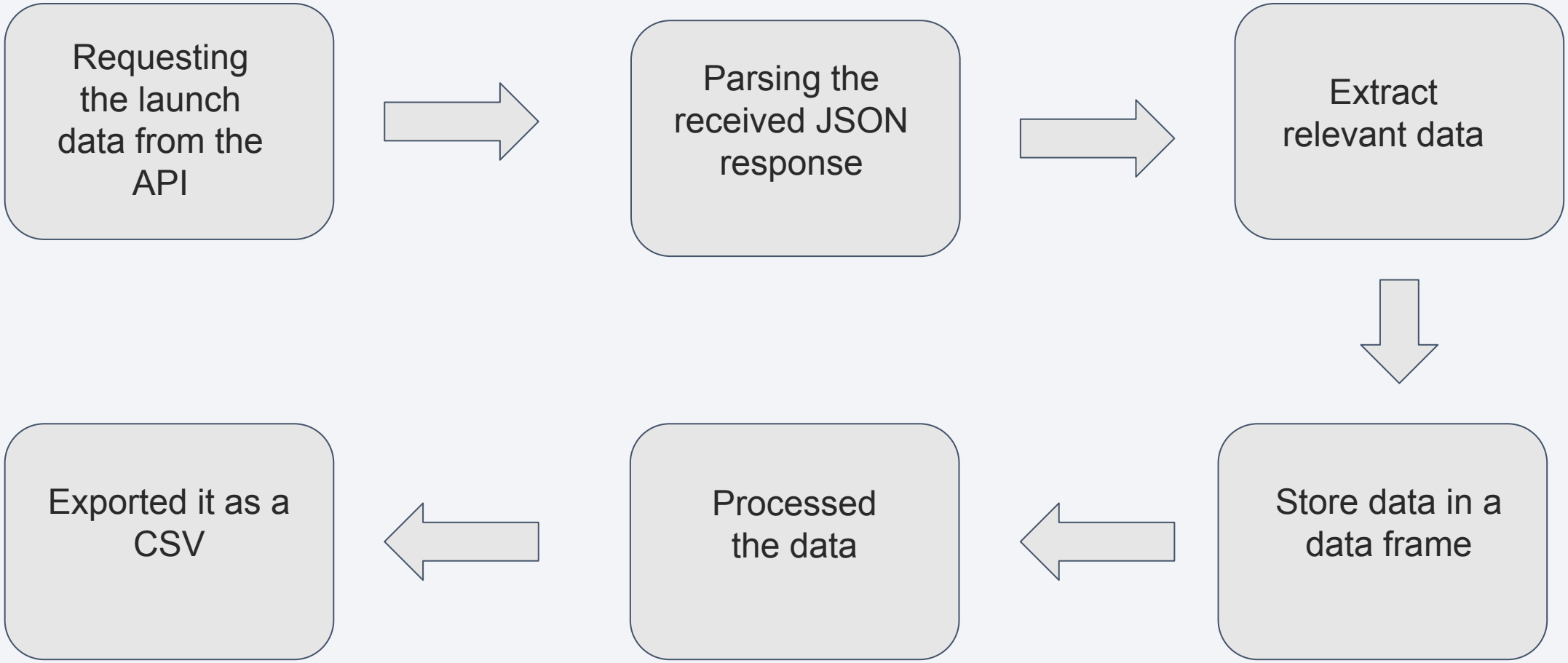
Executive Summary

- Data collection methodology:
  - Data was sourced from SpaceX API as well as web scraping Wikipedia
- Perform data wrangling
  - Data was processed by filtering the data, removing missing values and using One Hot Encoding to standardize it
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Implemented Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees, tuned with GridSearchCV and evaluated based on accuracy

Data Collection

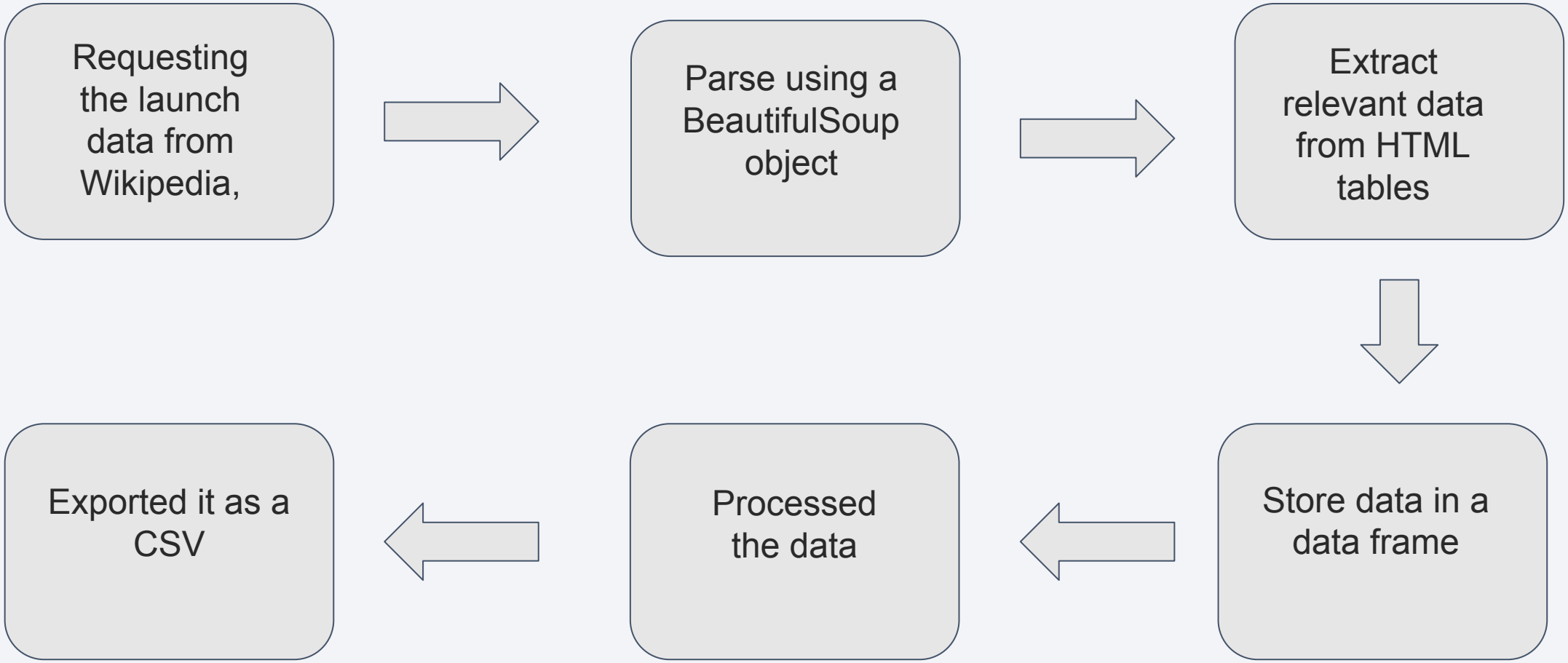
Data collection consisted of both accessing the SpaceX REST API as well as web scraping wikipedia and integrating the data together. The collection process for the API consisted of requesting the launch data from the API, parsing the received JSON response, extract relevant data, store data in a data frame, processed the data, and exported it as a CSV. The collection process for the web scraping procedure consisted of requesting the launch data from Wikipedia, parse using a BeautifulSoup object, extract relevant data from HTML tables, store data in a data frame, processed the data, and exported it as a CSV.

Data Collection – SpaceX API



https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

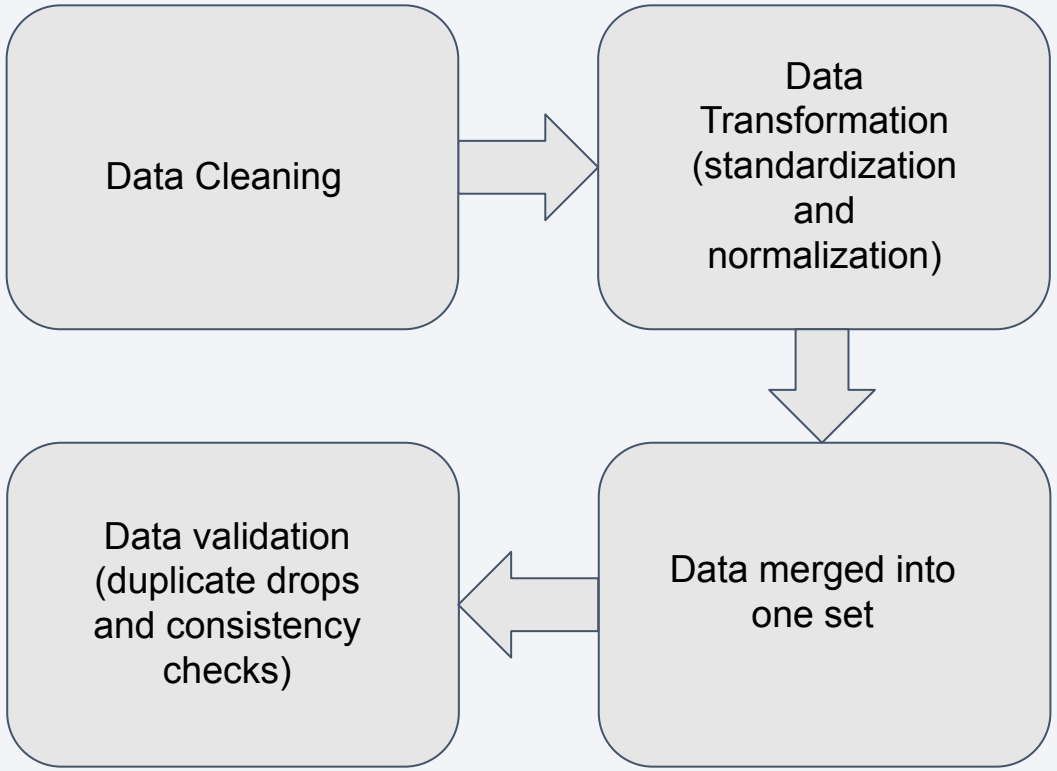
Data Collection - Scraping



https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

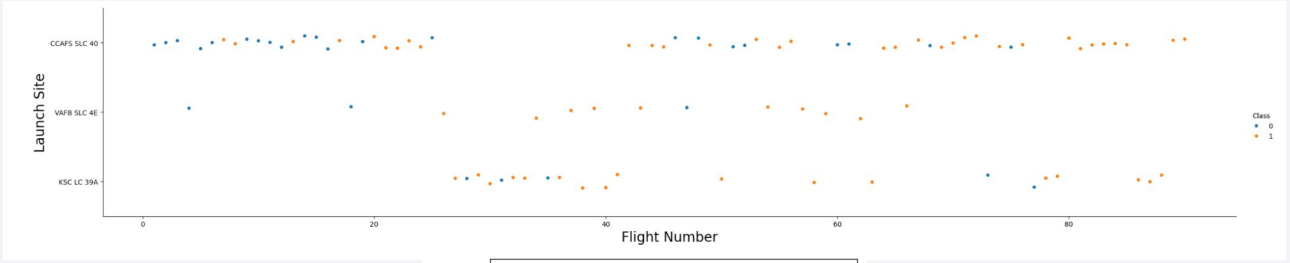
Data wrangling consisted of first cleaning the data by handling missing data by either dropping or filling with replacements. Then the data is transformed by standardizing it into the appropriate formats and normalized it for consistency. Next, the data sets are merged together into one. Lastly, the data is validated by checking consistency and dropping duplicates.

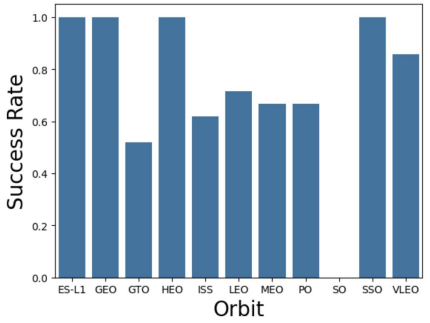


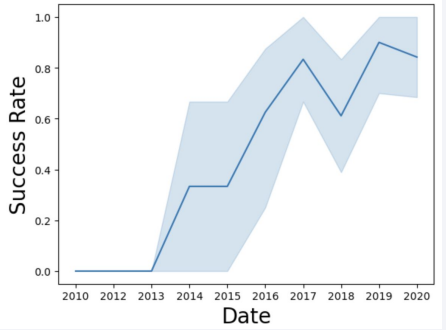
https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- ***Scatterplots***
  - These plots show the relationship between 2 variables which allows us to view correlations between variables and determine if relationships exist for our machine learning models
- ***Bar Charts***
  - These plots allow us to compare categorical variables across different categories which allows us to determine relationships between categorical variables which can not be computed numerically
- ***Line Charts***
  - These plots allow us to see trends in data over time which can help in understanding changes over time







https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/edadataviz.ipynb

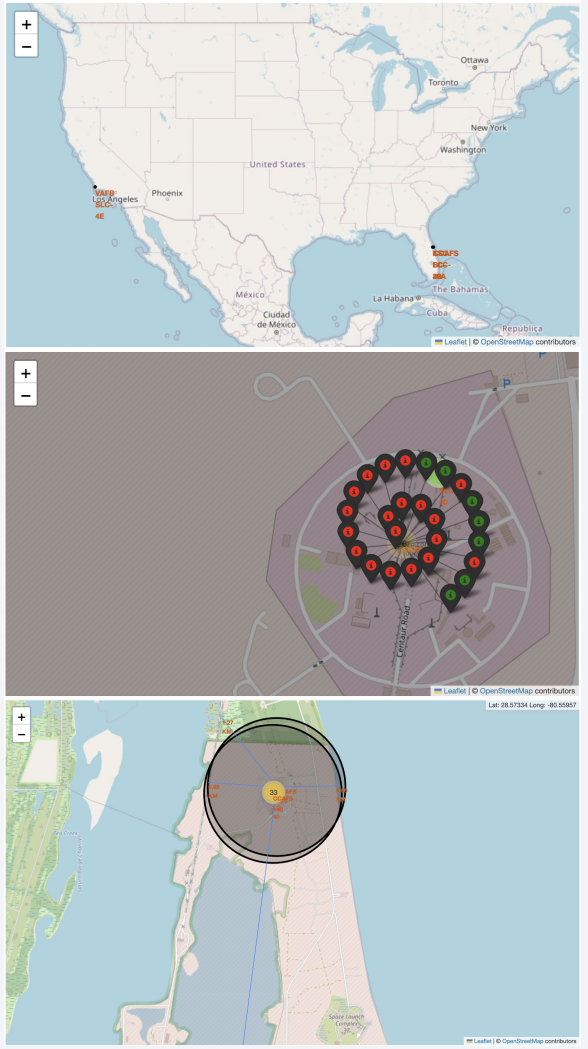
EDA with SQL

- **Display the names of the unique launch sites in the space mission**
  - %sql select distinct(Launch_Site) from SPACEXTBL
- **Display 5 records where launch sites begin with the string 'CCA'**
  - %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'
- **Display average payload mass carried by booster version F9 v1.1**
  - %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1'
- **List the date when the first succesful landing outcome in ground pad was acheived.**
  - %sql select min(Date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
  - %sql select Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000
- **List the total number of successful and failure mission outcomes**
  - %sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome = 'Success' or Mission_Outcome = 'Failure (in flight)'
- **List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.**
  - %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
  - %sql select substr(Date,6,2) as Month, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome= 'Failure (drone ship)' and substr(Date,0,5)='2015'
- **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 * from SPACEXTBL where Landing_Outcome like 'Success%' and (Date between '2010-06-04' and '2017-03-20') order by Date desc

https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- ***Markers***
  - Placed to indicate launch sites
  - Allows users to see where launches occurred as well as providing reference for distance
- ***Colored Markers***
  - Placed to indicate success (green) or failure (red) near launch sites
  - Allows users to determine ideal launch locations
- ***Lines***
  - Placed to indicate distance from coastlines, railroads, highways, and nearest cities
  - Allows users to see how environmental factors can affect launch success



https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Graphs

- **Pie Chart**
  - Shows distribution of successful vs failed launches for a given site
  - Allows users to determine mission success rates in an easily understood way at a glance
- **Scatterplot**
  - Shows the relationship between payload mass and launch success
  - Allows users to see the correlation between the 2 which can allow for planning payload mass accordingly to maximize launch success

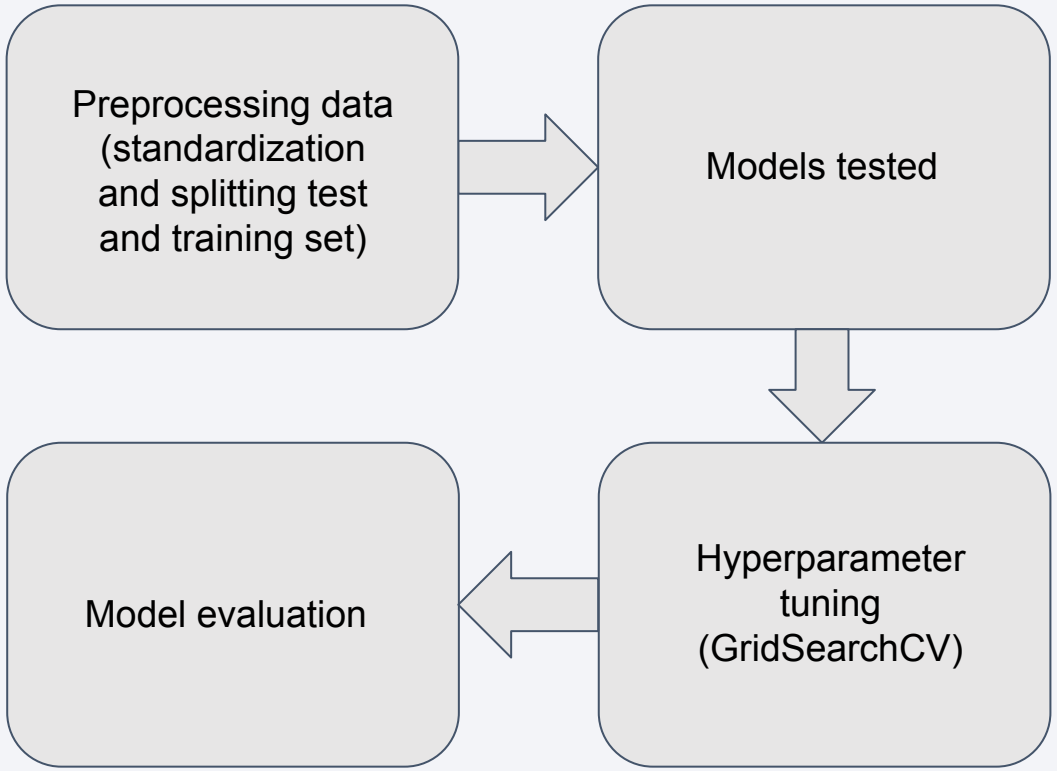
Interactions

- **Dropdown menu**
  - Used to select a specific launch site
  - Allows users to switch between launch sites to view various metrics for each
- **Range Slider**
  - Used to select different payload masses
  - Allows users to explore how payload mass effects mission success in tandem with the scatterplot

https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

Predictive analysis consisted of first preprocessing the data by standardizing it and splitting it into a test and training set. Next, models were tested including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN). Then, hyperparameters were tuned using GridSearchCV to determine the best hyperparameter selection. Lastly, the models are evaluated and the best model is determined based on the confusion matrices, test accuracy, test F1 scores and global F1 scores.



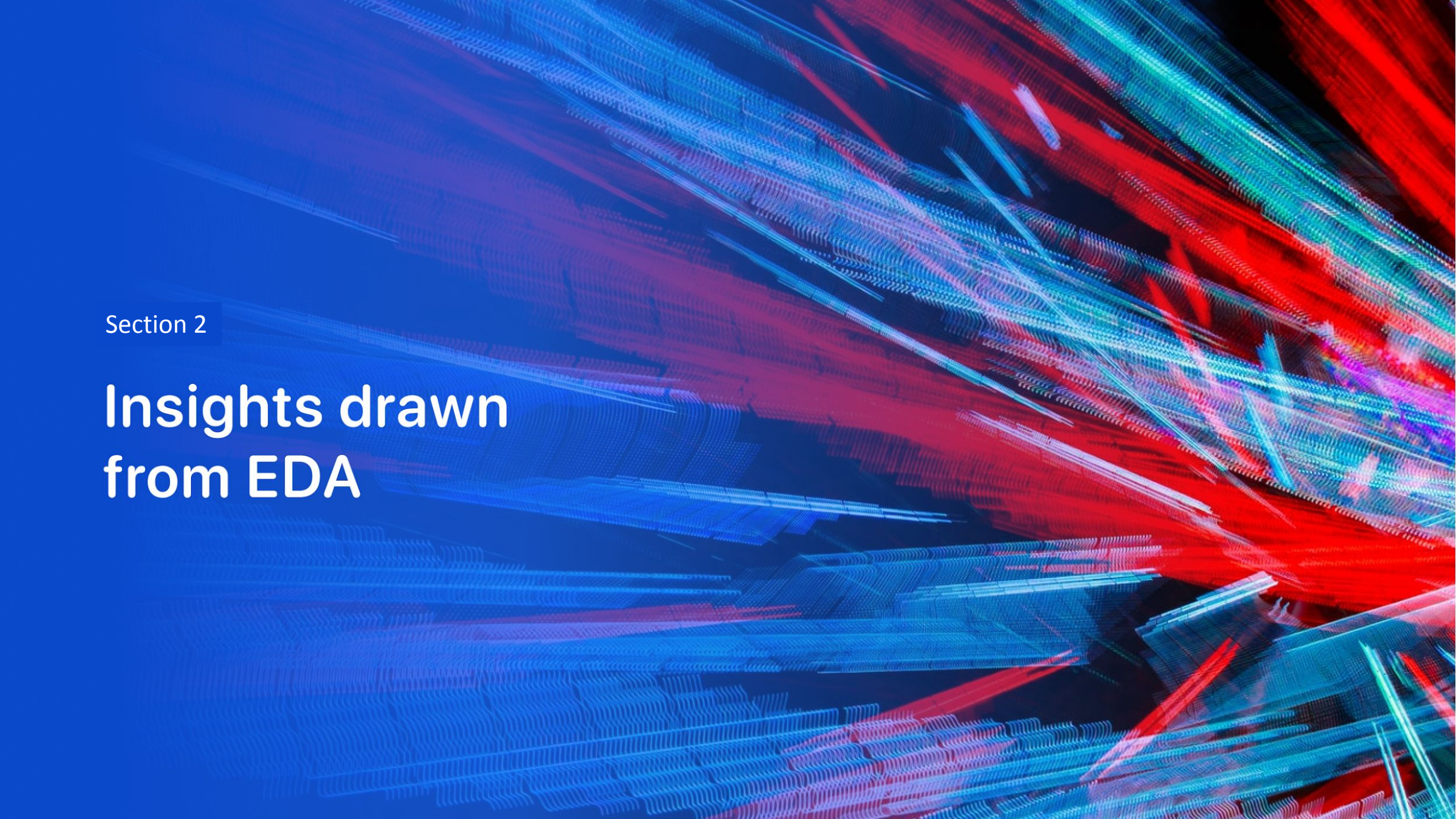
https://github.com/nassim-oulhadj/IBM_Capstone/blob/main/spacex-dash-app.py

Results

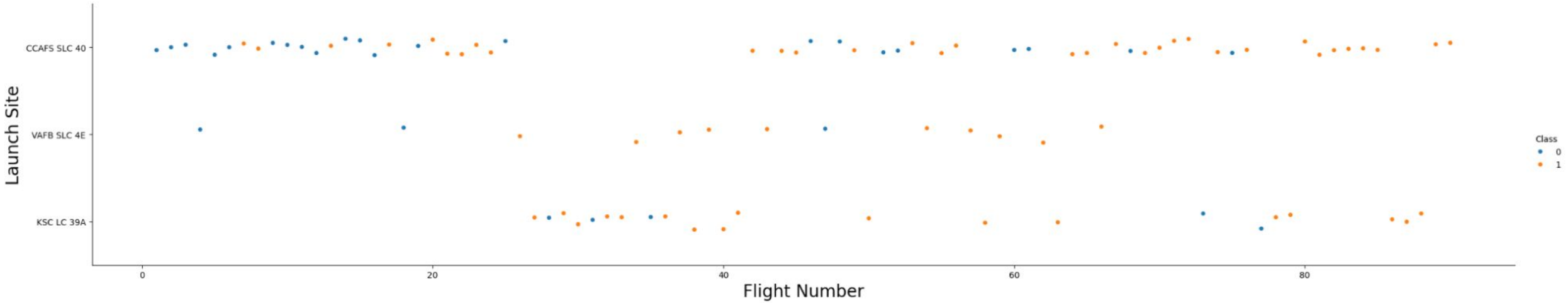
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

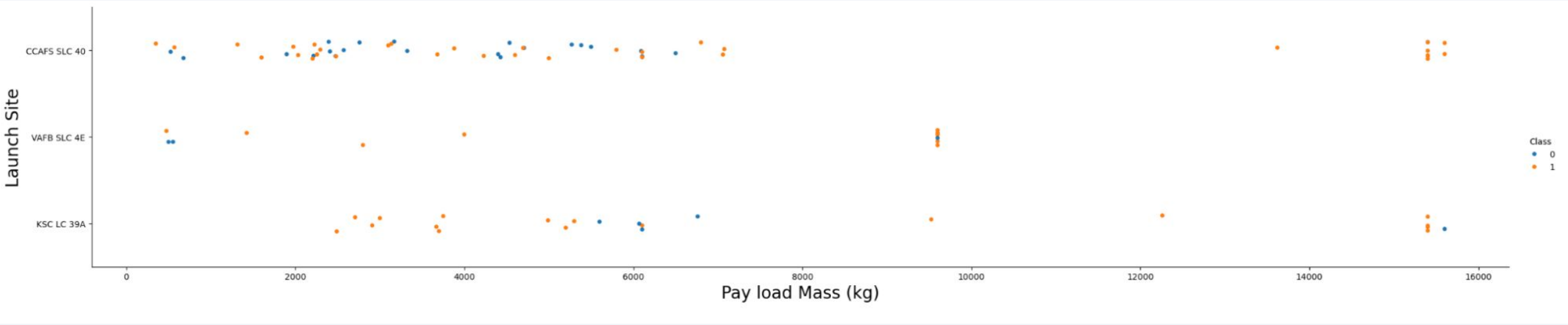


Flight Number vs. Launch Site



- There is a mix of successful and unsuccessful launches for all of them
- CCAFS SLC 40 has the vast majority of launches but also the most failures
- Earlier attempts tended to be more unsuccessful and later attempts tended to be more successful

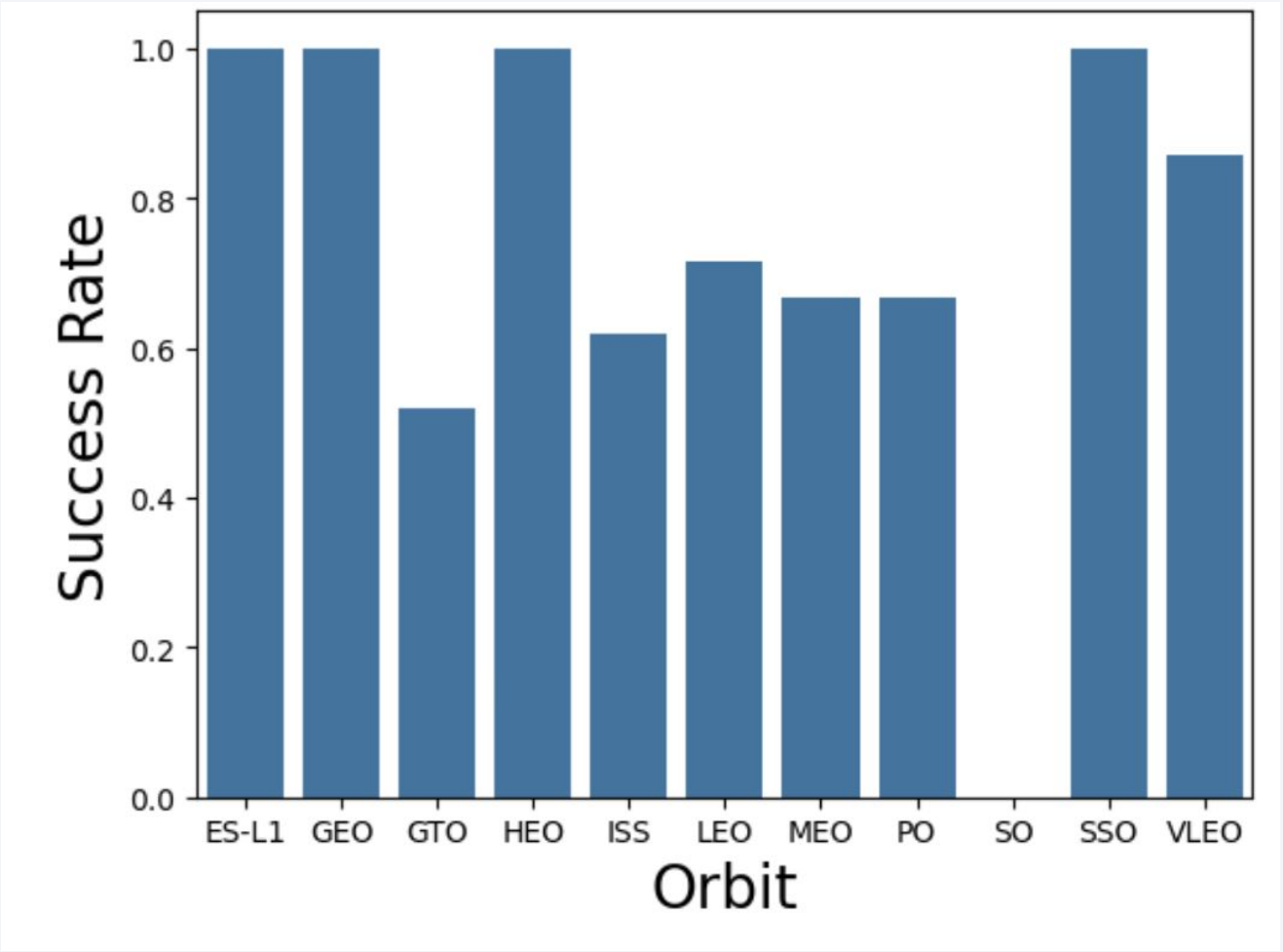
Payload vs. Launch Site



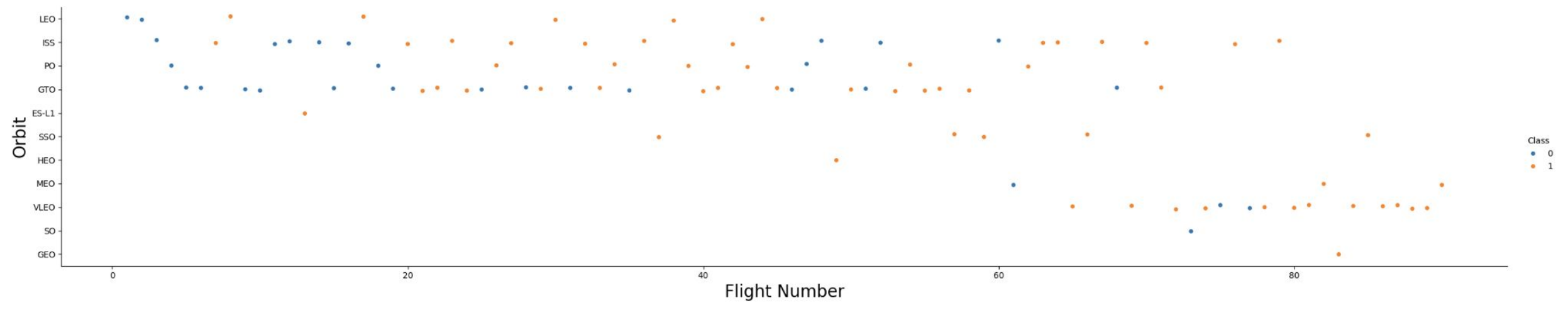
- As the mass gets heavier initially, it increases the rate of unsuccessful launches
- Launches after 9000 kg have a much higher success rate but there are less of them as well
- CCAFS SLC 40 has the majority of heavy successful launches whereas KSC LC 39A appears better suited for lower masses

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have 100% success rates
- GTO, ISS, LEO, MEO, PO and VLEO have success rates between 50% and 80%
- SO is the only orbit with a 0% success rate

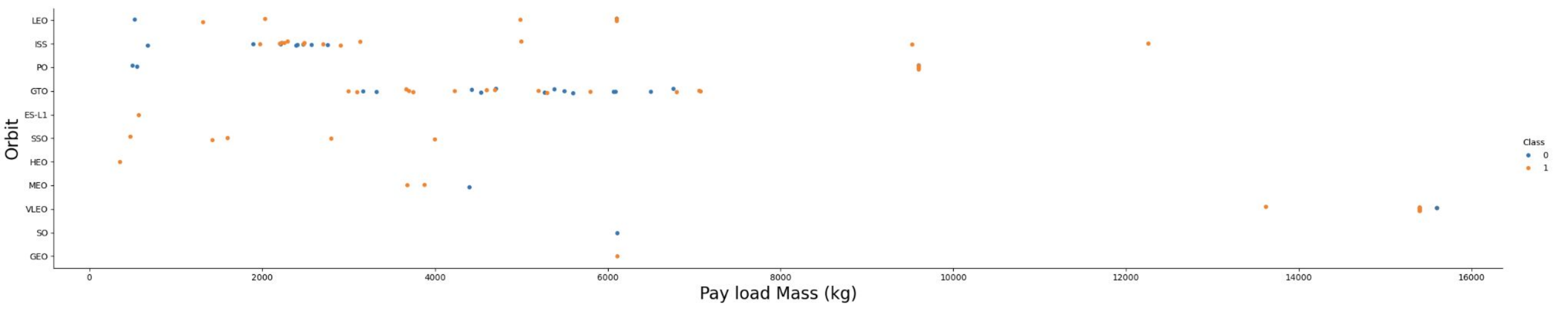


Flight Number vs. Orbit Type



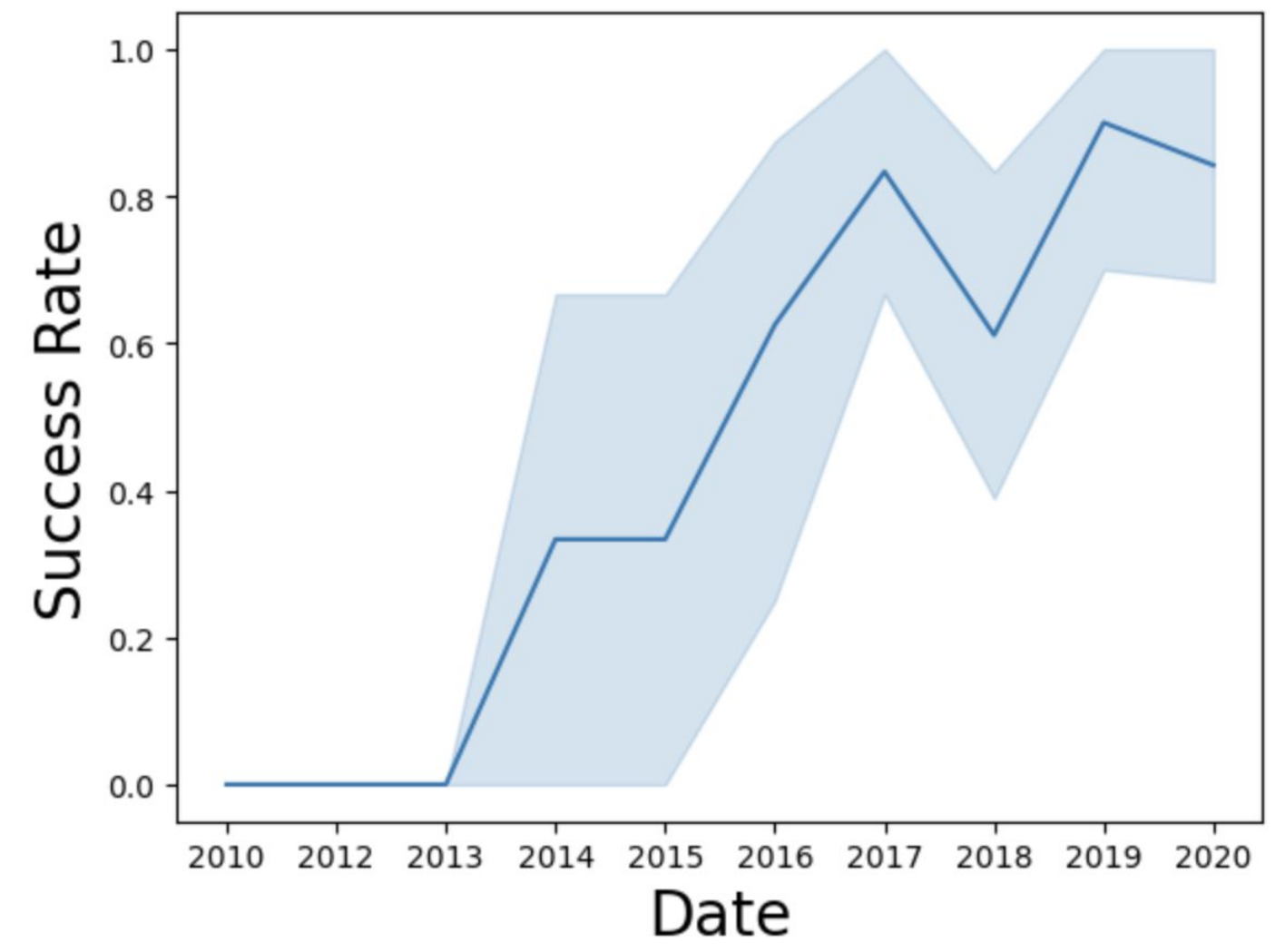
- Most all orbits have better success outcomes for later flight attempts
- GTO appears to have a mix of successful and unsuccessful attempts through its flight attempts

Payload vs. Orbit Type



- Most orbits perform better with lower payload masses especially GTO
- LEO, ISIS and PO actually perform better with higher payload masses

Launch Success Yearly Trend



- There is an 80% increase in success rate from 2013 to 2020
- There was a 20% dip from 2017 to 2018 but it was corrected the next year and appears to be showing consistent growth over time

All Launch Site Names

```
%sql select distinct(Launch_Site) from SPACEXTBL
```

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

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

This query displays all launch site names

Launch Site Names Begin with 'CCA'

```
[34]: %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

[34]:

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

This query displays 5 launch sites with names beginning with CCA

Total Payload Mass

```
[35]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
[35]:
```

sum(PAYLOAD_MASS__KG_)
45596

This query displays the total payload mass launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
[36]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

[36]:

avg(PAYLOAD_MASS__KG_)
2928.4

This query displays the average mass carried by booster F9 v1.1

First Successful Ground Landing Date

```
[37]: %sql select min(Date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'
```

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

[37]:

min(Date)
2015-12-22

This query displays the first successful ground launching date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000
```

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

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

This query displays the boosters with successful drone ship landing with a payload mass between 4000 kg and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) as Outcome_Counter from SPACEXTBL group by Mission_Outcome
```

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

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

This query displays the first successful ground launching date

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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

This query displays the boosters that carried the maximum payload mass

2015 Launch Records

```
%sql select substr(Date,6,2) as Month, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome= 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

This query displays the failed drone ship landing outcomes with the month, booster version and launch site listed

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

```
%%sql select landing_outcome, count(*) as landing_count from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by landing_outcome
    order by landing_count desc
```

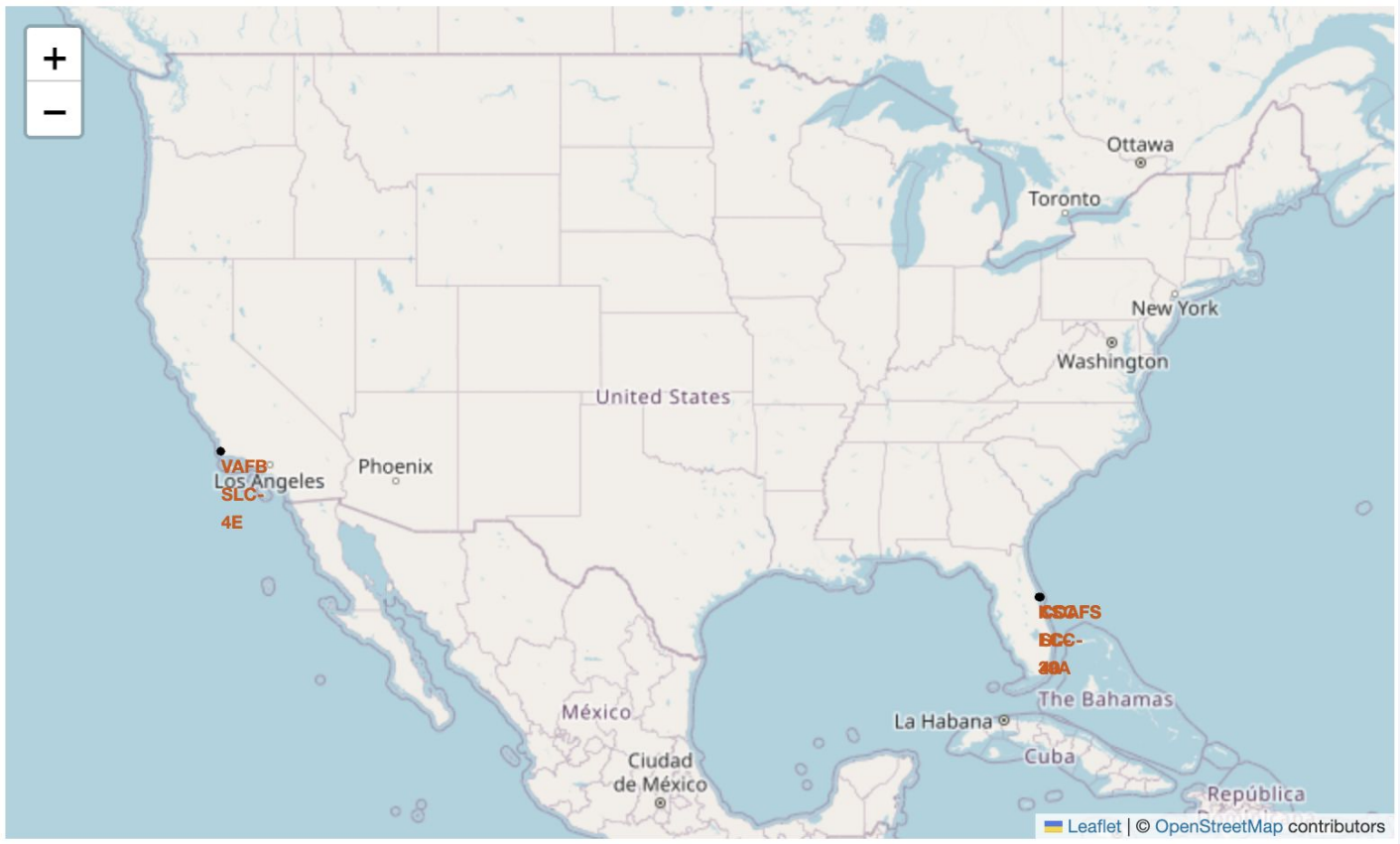
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	landing_count
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

This query displays the count of landing outcomes between 2010-06-04 and 2017-03-20

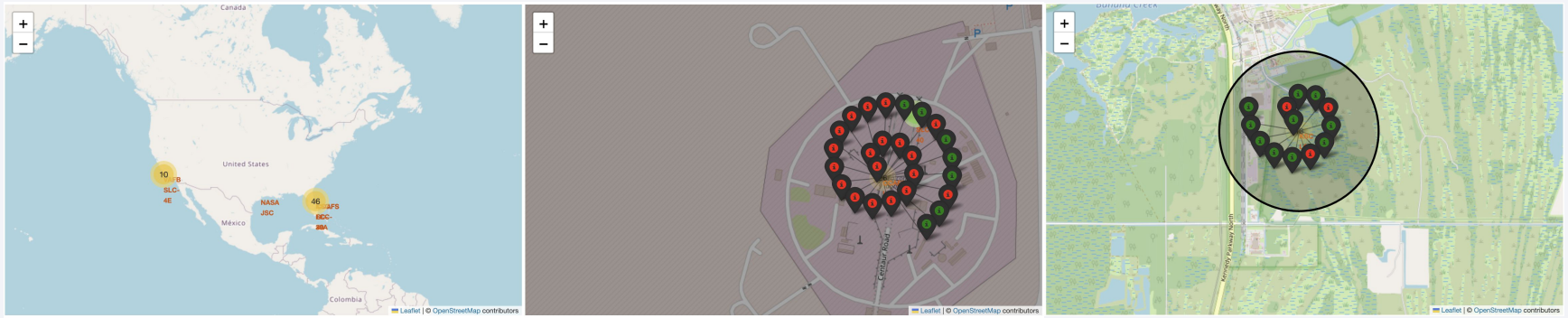

Section 3
Launch Sites
Proximities Analysis

Launch Site Location Markers Global Map



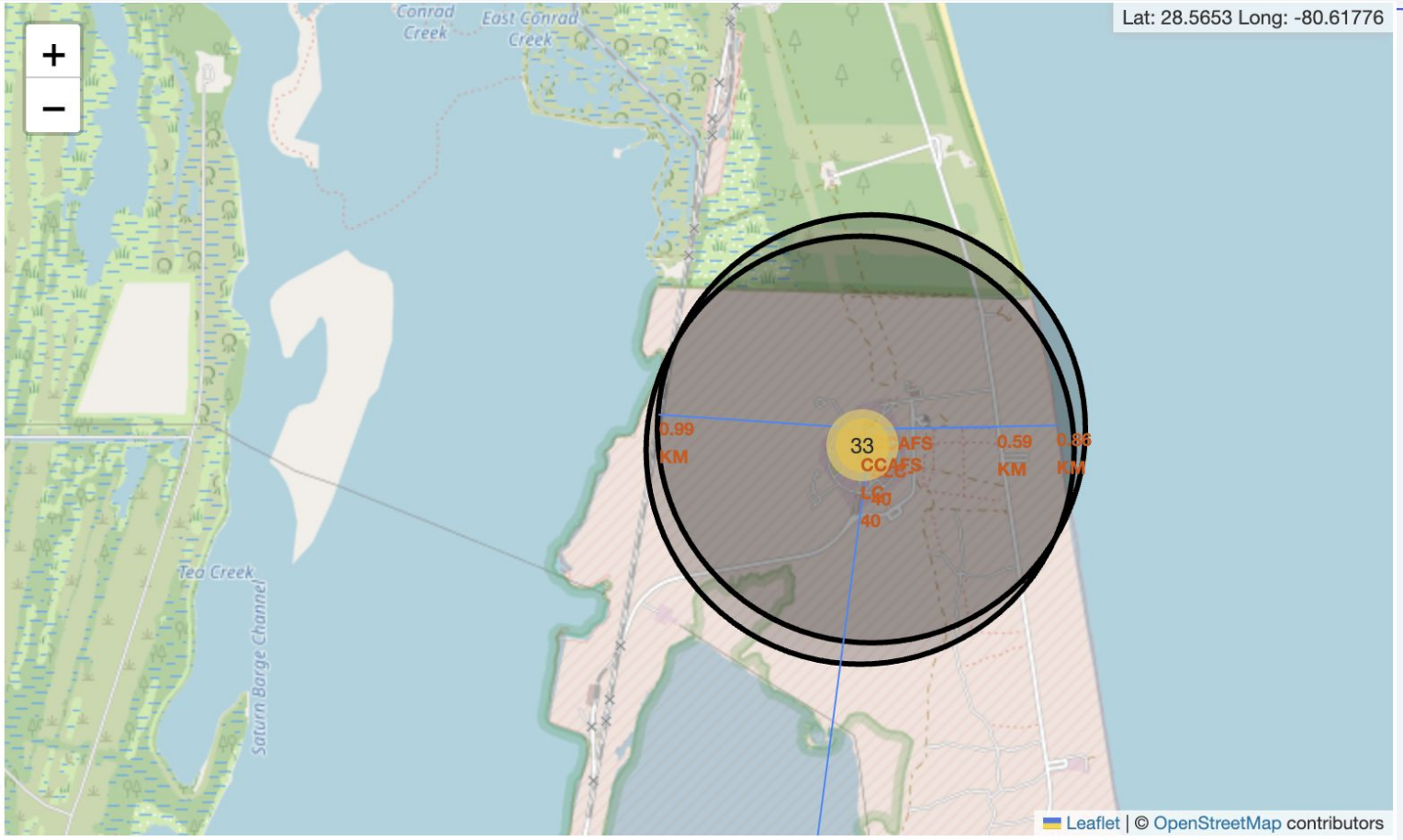
- The launch sites are in areas of the USA that are as south as possible, likely to be close to the equator since that would save on fuel due to the Earth rotating faster there
- The launch sites are also all close to coastlines which is likely due to safety concerns of failed launches being able to just fall in the water
- Most launch sites are in Florida, likely due to Florida offering lucrative tax benefits for development in the state

Color Labeled Launch Outcomes Global Map

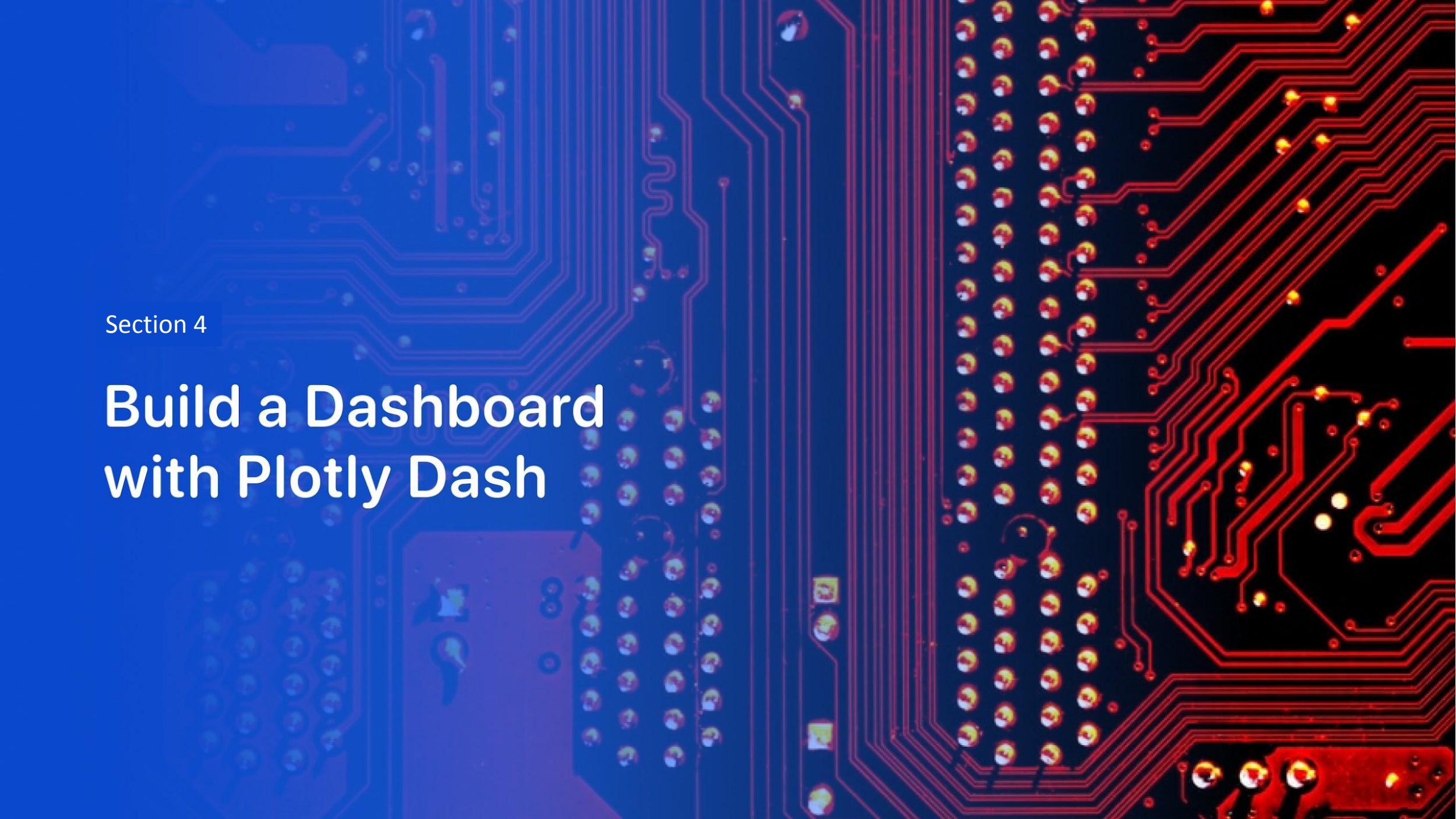


- Green markers represent successful launches whereas red markers are unsuccessful
- CCAFS LC-40 had the worst outcomes with 19 red and 7 green meaning it had a 27% success rate
- KSC LC39-A had the best outcomes with 3 red and 10 green meaning it had a 77% success rate

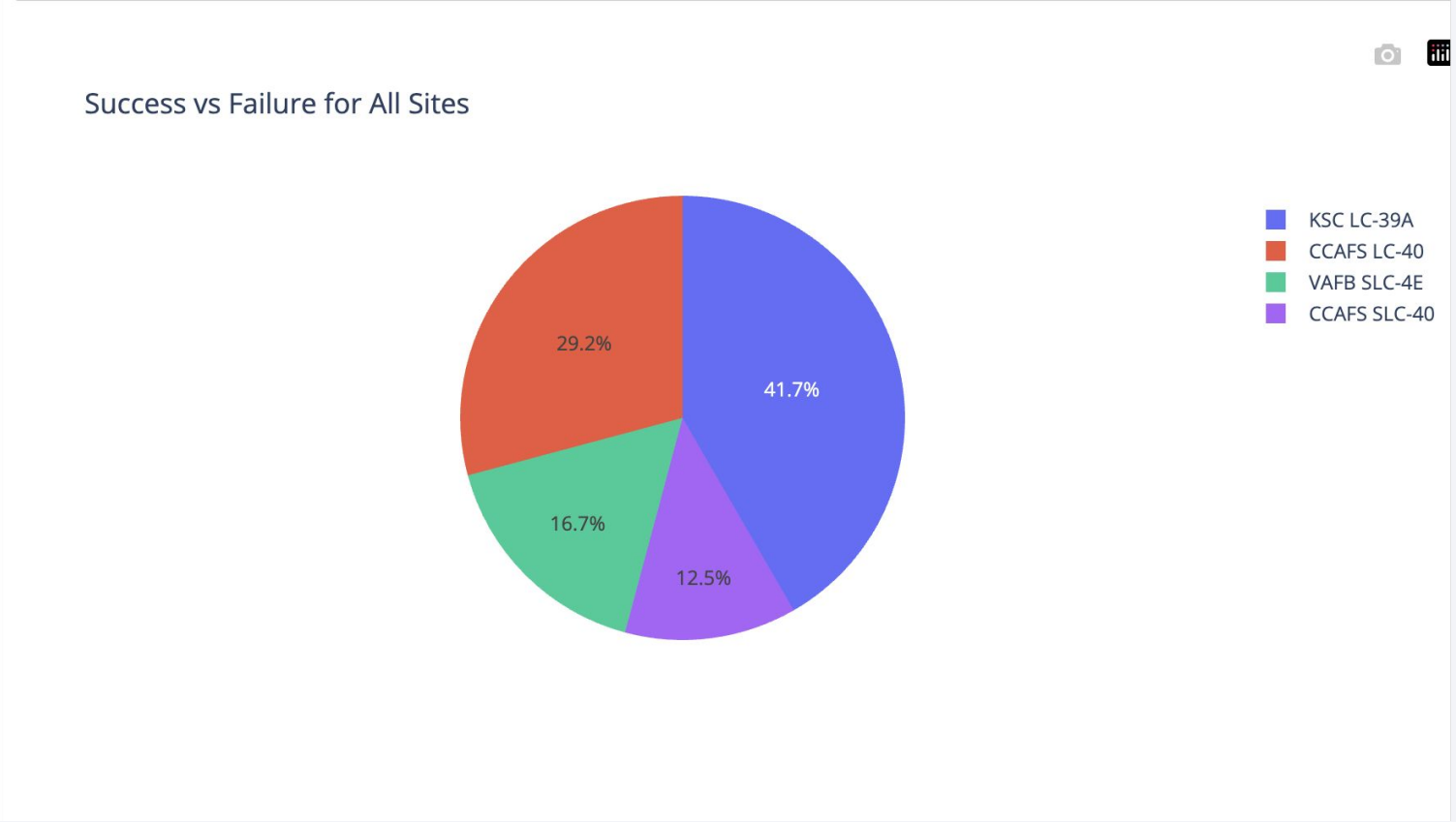
Distance From Launch Sites to Proximities Map



- Launch site CCAFS SLC-40 distance from
  - Coastline: 0.86 km
  - Railroad: 0.99 km
  - Highway: 0.59 km
  - City: 19.66 km
- The close proximity to trains and highways could be a partial cause as to the numerous failures at this site but the distance from the coast and nearest city are ideal

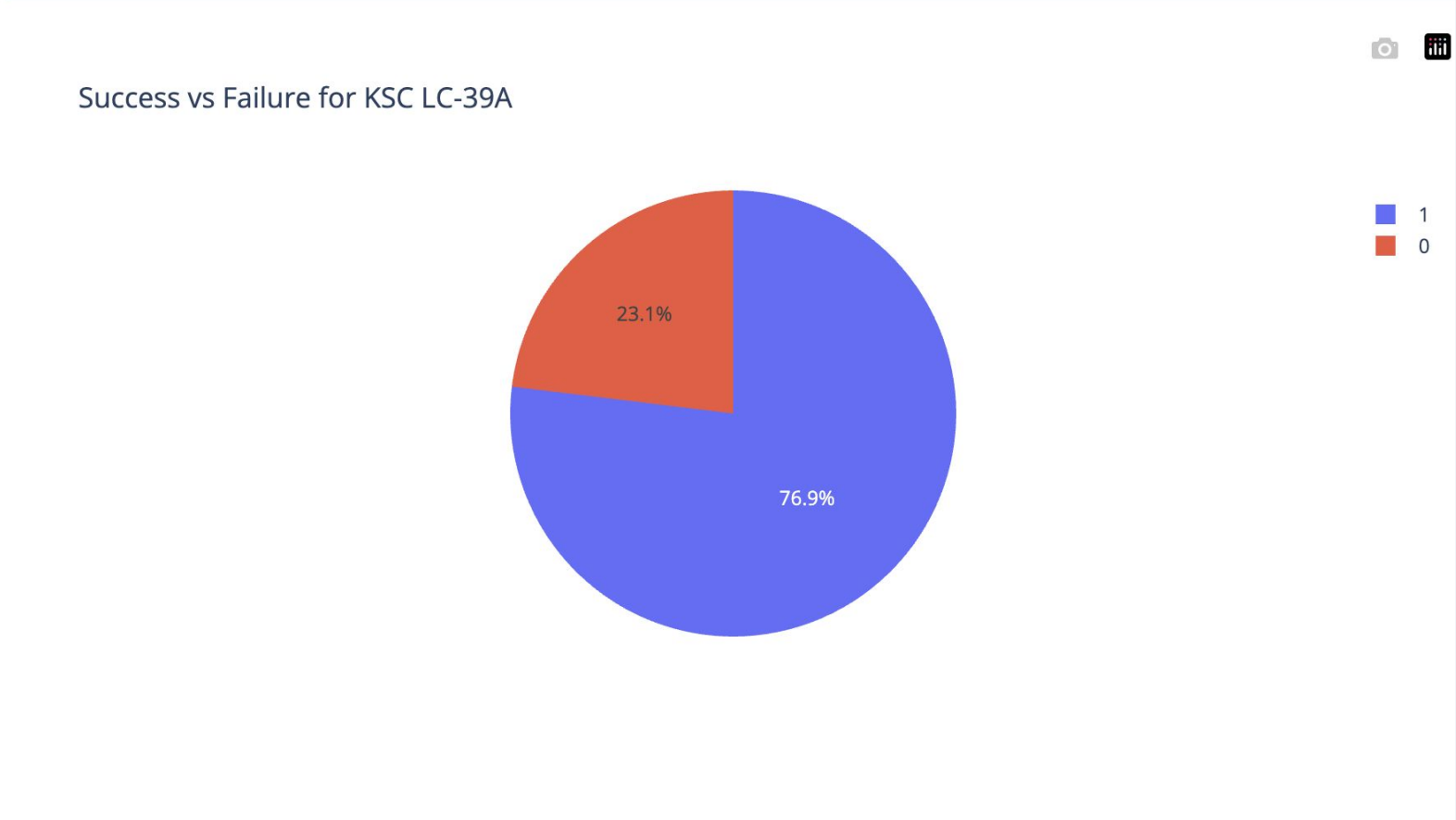

Section 4
Build a Dashboard with Plotly Dash

Pie Chart of All Sites Launch Success



- Percentage of success for all launch sites
  - KSC LC-39A: 41.7%
  - CCAFS LC-40: 29.2%
  - VAFB SLC-4E: 16.7%
  - CCAFS SLC-40: 12.5%
- As seen in the chart, KSC LC-39A has the majority of successful launches making up nearly half of the total

Pie Chart of Most Successful Launch Site



- KSC LC-39A is the most successful launch site
- 76.9% success rate with 10 launches
- 23.1% failure rate with 3 launches

Payload Mass Vs Launch Outcome Scatterplot

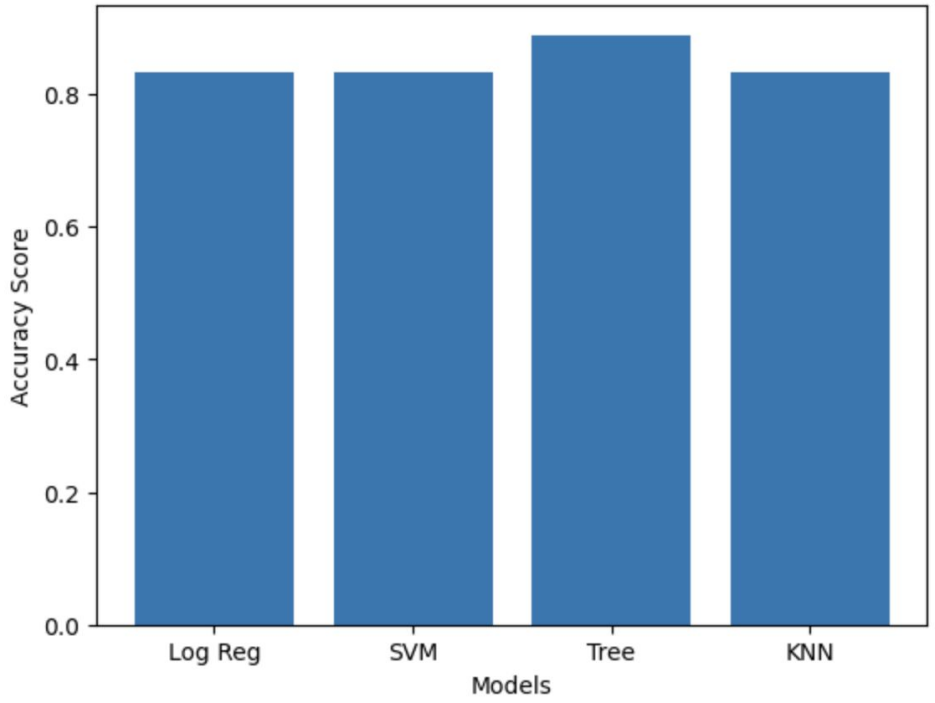
When comparing between a low payload range (0kg-5,000kg) and a high payload range (5,000kg-10,000kg), it is clear that the lower payload range has better success margins with about half of them succeeding compared to only 3 for the high payload range.

Section 5

Predictive Analysis (Classification)

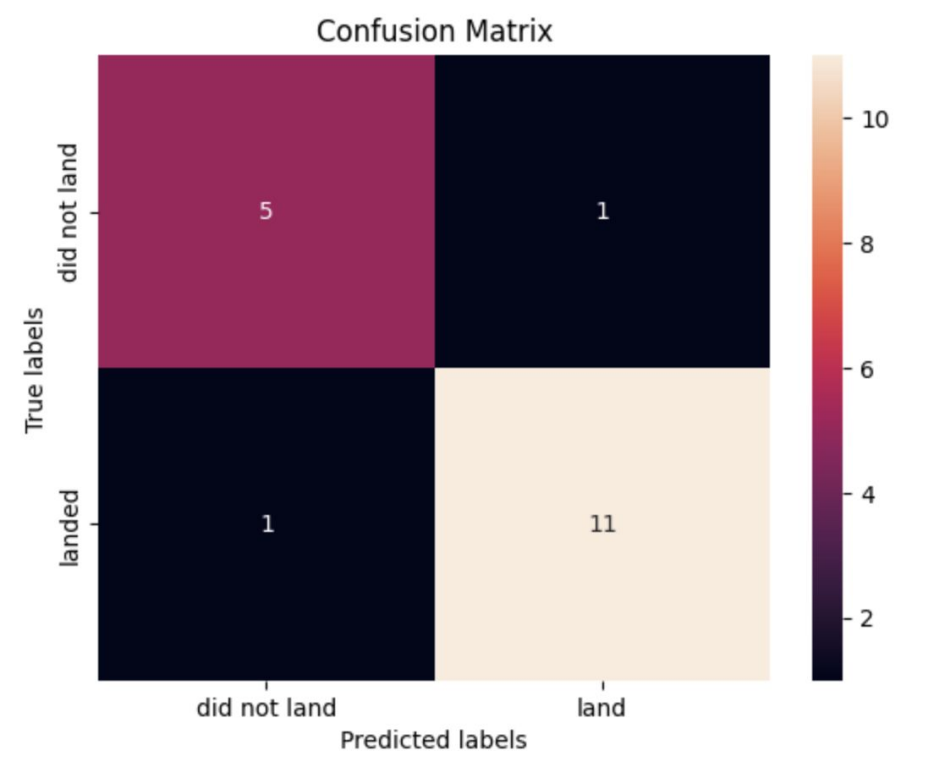
Classification Accuracy

```
Log Reg: 0.8333333333333334
SVM: 0.8333333333333334
Tree: 0.8888888888888888
KNN: 0.8333333333333334
```



The decision tree has the highest accuracy at 0.889 as opposed to all the others that have a tie at 0.833

Confusion Matrix



- This confusion matrix has the least amount of errors
  - Type 1 errors: 1
  - Type 2 errors: 1
  - Total errors: 2
- All other confusion matrices for the models
  - Type 1 errors: 3
  - Type 2 errors: 0
  - Total errors: 3

Conclusions

- Launches with lower payload masses tended to perform better
- Orbits ES-L1, GEO, HEO and SSO have 100% success rates
- There is an 80% increase in success rate from 2013 to 2020
- Launch sites are as close to the equator and coasts as possible for fuel saving and safety benefits
- KSC LC-39A is the most successful launch site with a 76.9% success rate
- The decision tree performed the best with an accuracy of 0.889 as opposed to the other models having an accuracy of 0.833

Appendix

It is interesting to note how all the models when ran were consistent except the decision tree. When running the decision tree it had to be tuned manually because the GridSearchCV object was not returning consistent tuned hyperparameters, at one point being below the others, then the same, then it was above the rest. This effect was not present in any of the other models.

Thank you!