IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marc Adlam
August 31, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection by calling SpaceX REST API and by web scraping Wikipedia.org
  - Data wrangling using Pandas and Numpy
  - Exploratory Data Analysis (EDA) using Sqlite and Python
  - Visual EDA using Matplotlib and Seaborn libraries in Python
  - Mapping of launch sites using Folium
  - Dashboard creation using Dash and Plotly
  - Predictive analytics using a varied set of machine learning classifier models
- Summary of key results
  - Decision Tree classifier model scored best with this data set
  - Launch success rates increase over time, with low-payload launches outperforming larger ones
  - Specific rocket models and orbit types enjoy greater success, with launches occurring near the equator and in proximity to shoreline

Introduction

- Project background and context
  - Our startup Space Y is looking to potentially enter the market to compete with market leader SpaceX
  - Our job is to learn from historical data and make predictions to guide business toward the best decisions on whether and how to enter the market
- Questions needing to be answered
  - What types of rockets work best?
  - What orbit types should we consider?
  - Where should we consider choosing for our launch sites?
  - What machine learning models should we consider applying as we seek to make predictions?

Section 1

Methodology

Methodology

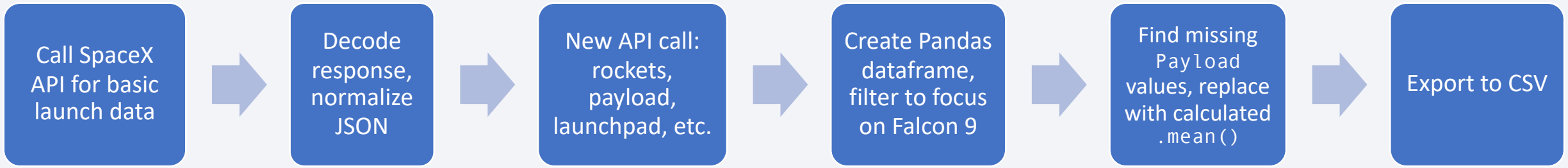
Executive Summary

- Data collection methodology:
  - Data obtained from SpaceX public API and from Wikipedia.org
- Perform data wrangling
  - Used Pandas to identify numerical vs categorical data, calculated landing Outcomes
- Command-line exploratory data analysis (EDA) using SQL;
  Visualization using EDA with Seaborn & Matplotlib
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classifier models
  - Details provided below on building, tuning, and evaluating various classification models

Data Collection

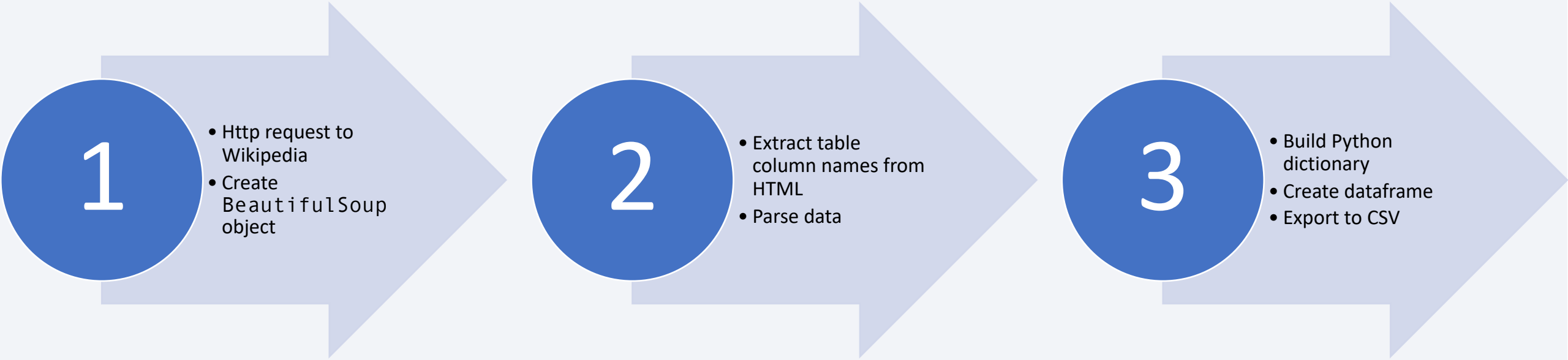
- Data sets collected using SpaceX public API and Wikipedia.org
  - https://api.spacexdata.com/v4/launches/past
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data from SpaceX public REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Supplemental data scraped from Wikipedia:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



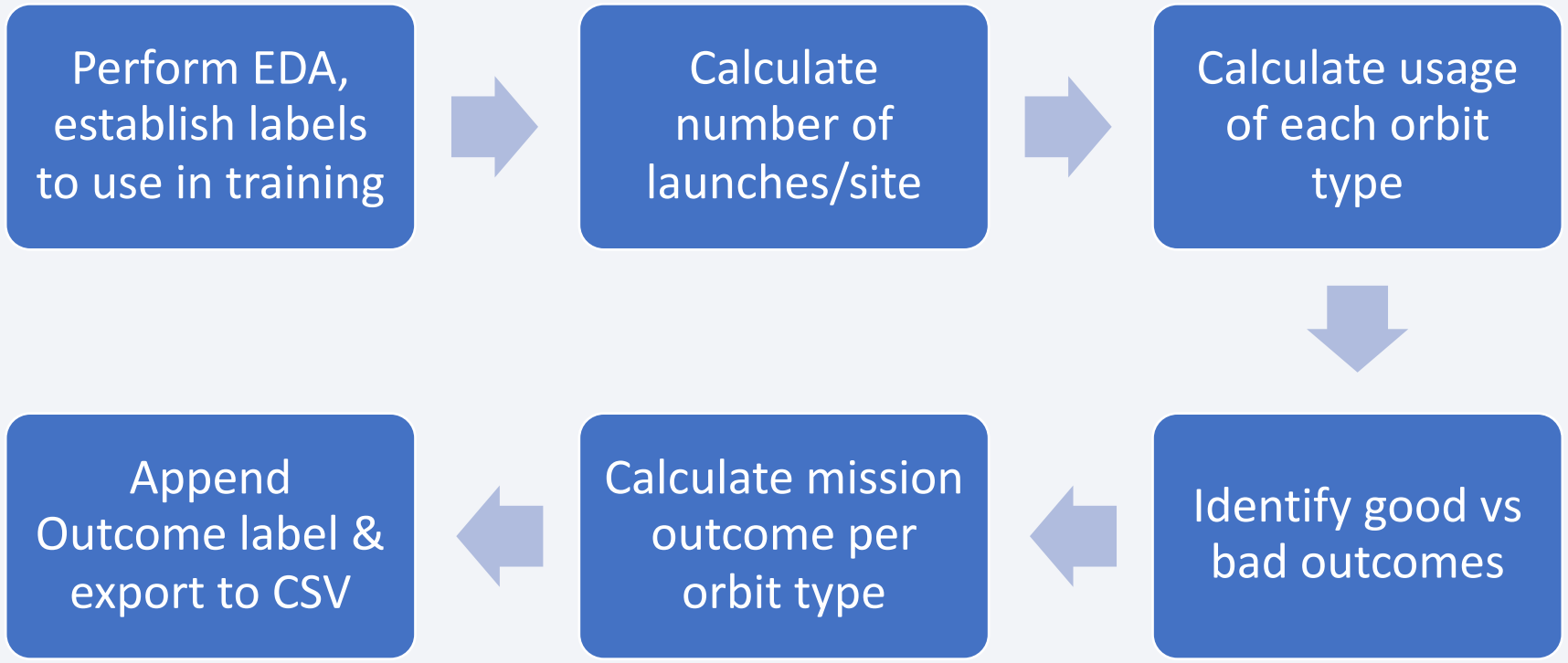
https://github.com/madlam/ibm-data-science/blob/main/Notebooks/01-DataCollection.ipynb

Data Collection - Scraping



https://github.com/madlam/ibm-data-science/blob/main/Notebooks/02-WebScraping.ipynb

Data Wrangling



https://github.com/madlam/ibm-data-science/blob/main/Notebooks/03-DataWrangling.ipynb

EDA with SQL

- Find names of the unique launch sites in the space mission
- Display total payload mass carried by boosters launched by NASA (CRS)
- Calculate average payload mass carried by booster version F9 v1.1
- Determine date when the first successful landing outcome on a ground pad was achieved
- List the names of boosters that have had success in drone ship *and* have payload mass greater than 4000 but less than 6000
- Count the total number of successful and failure mission outcomes
- Find the names of booster versions which have carried the maximum payload mass
- Show the records with month names, failure landing outcomes in drone ship, booster versions, and launch site for any month in 2015.
- Rank the total landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

https://github.com/madlam/ibm-data-science/blob/main/Notebooks/04-ExploratorySQL.ipynb

EDA with Data Visualization

- Charts that were plotted include:
  - 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
  - Yearly Trend of Success Rate Trend

Chart types and applicability:

Scatter plots *were used to show the relationship between variables.*

- *If a relationship exists, they could be used in machine learning model.*

Bar charts (or histograms) *were used to show comparisons among categorical data.*

- *Their purpose is to show magnitude and compare relationship between various categories of data.*

Line chart *was used for success rate to show progress over a period of time.*

https://github.com/madlam/ibm-data-science/blob/main/Notebooks/05-ExploratoryVisual.ipynb

Interactive Map with Folium

Created markers of all launch sites utilized by SpaceX:

- Included Marker with Circle, Popup Label, and Text Label of NASA Johnson Space Center using
- its geographic coordinates. Set this as a start location.
- More Markers added with Circle, Popup Label, and Text Label of all Launch Sites using their coordinates to show their locations and proximity to both the equator and nearby coastlines.

Produced colored Markers showing the launch outcomes for each launch site:

- These Markers indicate launch success (Green) and failure (Red).
- Also utilized Marker Cluster to visually identify launch sites with higher success rates.

Mapped distances between launch sites and nearby locations:

- Included lines of various colors to show distance to nearby locations such as railways, highways, coastline, and cities.

https://github.com/madlam/ibm-data-science/blob/main/Notebooks/06-MapLaunchSites.ipynb

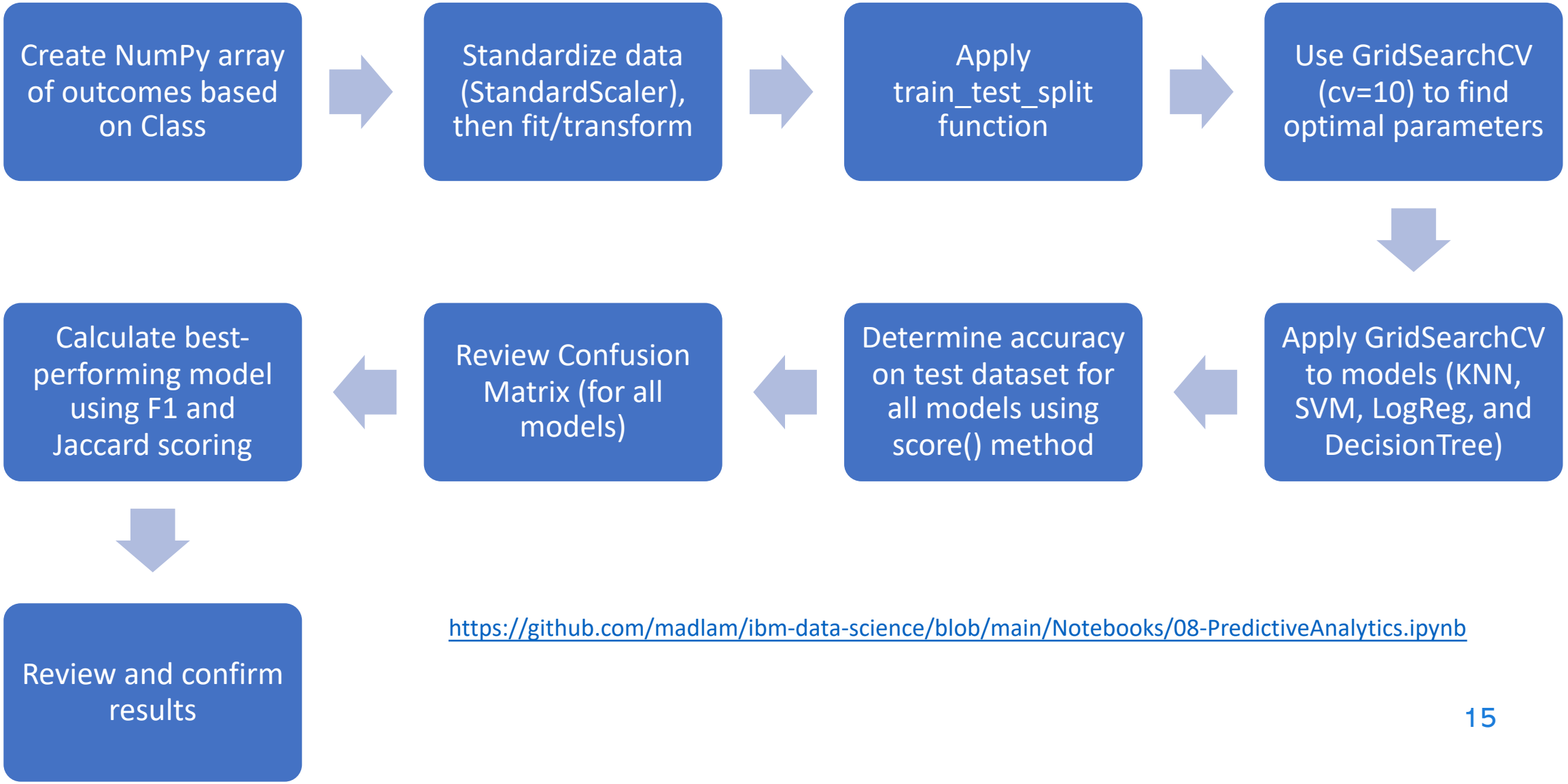
Interactive Dashboard with Plotly Dash

Dashboard features:

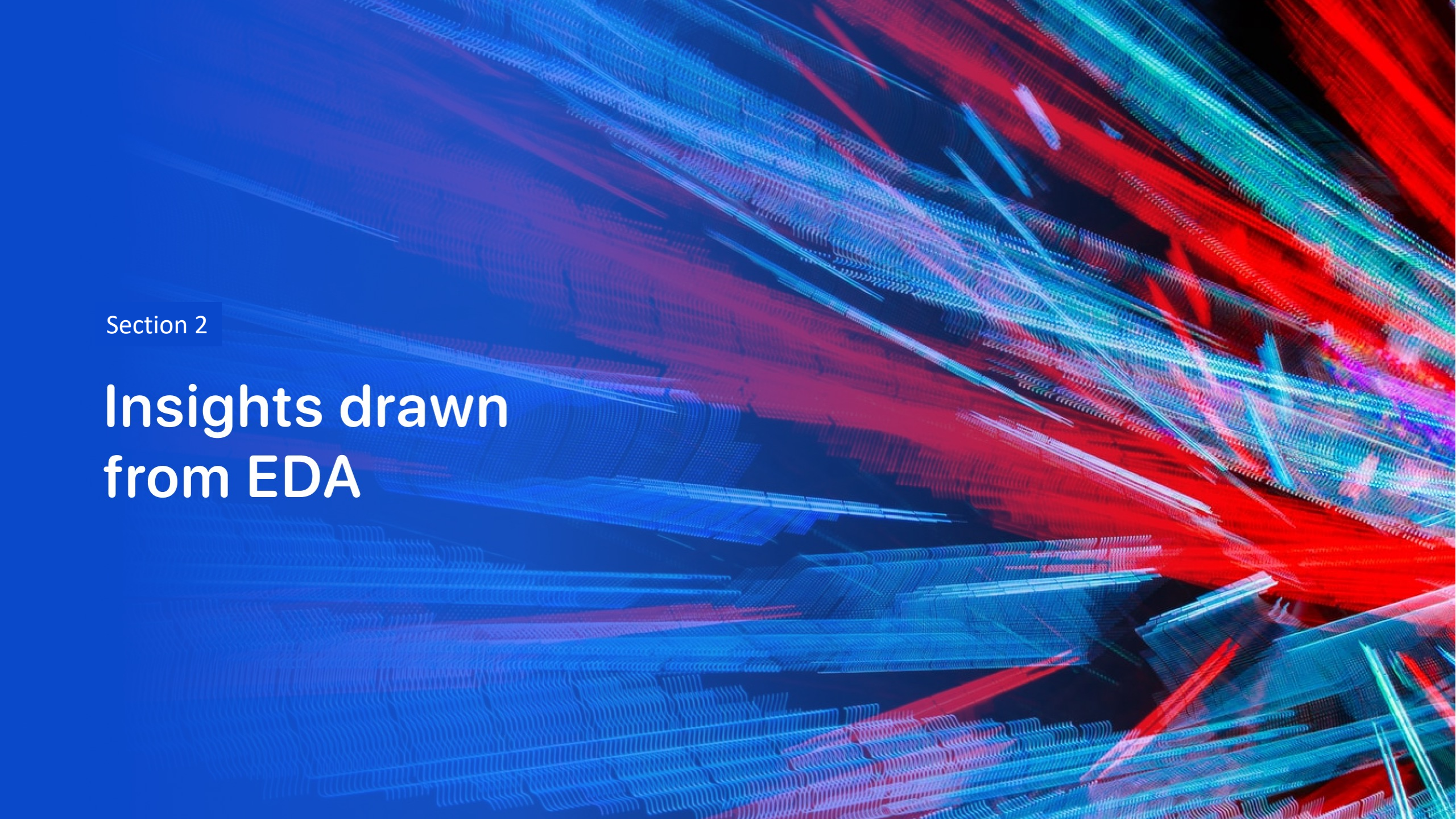
- Dropdown List
  - Enables Launch Site selection.
- Pie Chart
  - Shows total successful launches for all sites
  - Also can show total successes or failures for a given site (if only one site selected)
- Slider
  - Allows user to select Payload range.
- Scatter Chart
  - Displays relationship between Payload Mass and Launch Success.

https://github.com/madlam/ibm-data-science/blob/main/Code/7-spacex_dash_app.py

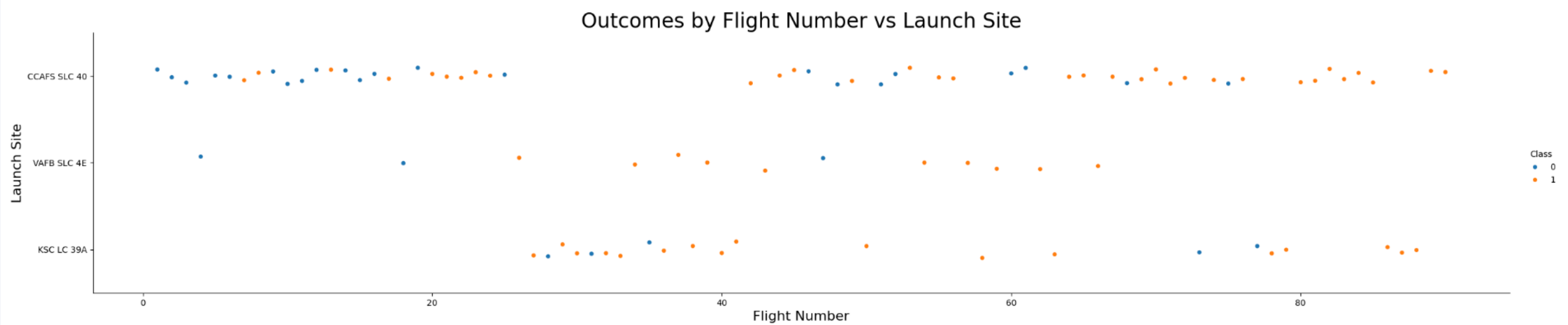
Predictive Analysis (Classification)



https://github.com/madlam/ibm-data-science/blob/main/Notebooks/08-PredictiveAnalytics.ipynb


Section 2
Insights drawn from EDA

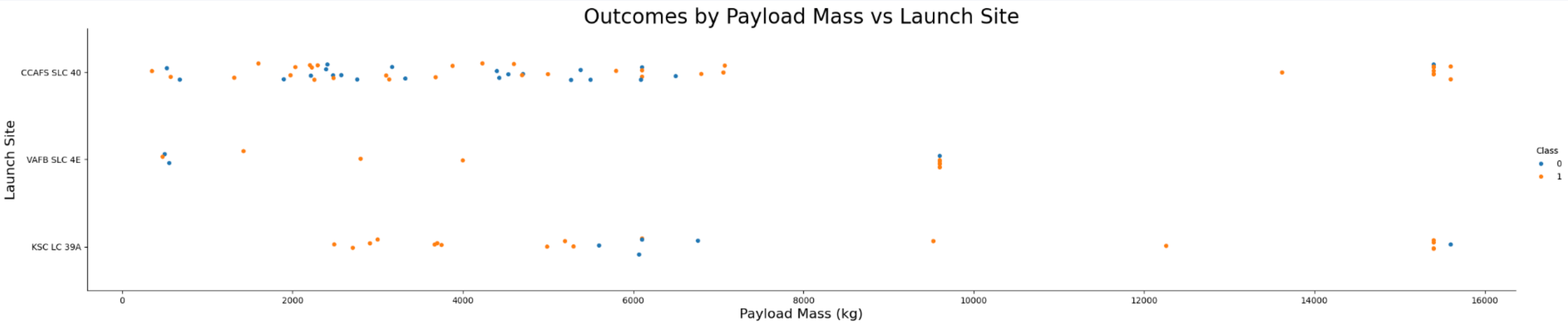
Flight Number vs. Launch Site



Patterns observed:

- The first 6 flights all failed.
- The most recent 13 flights all succeeded.
- The CCAFS SLC 40 launch site has hosted the majority of launches, but also has the most failed launches.
- The VAFB SLC 4E and KSC LC 39A sites have higher success rates, although VAFB is used less frequently.

Payload vs. Launch Site



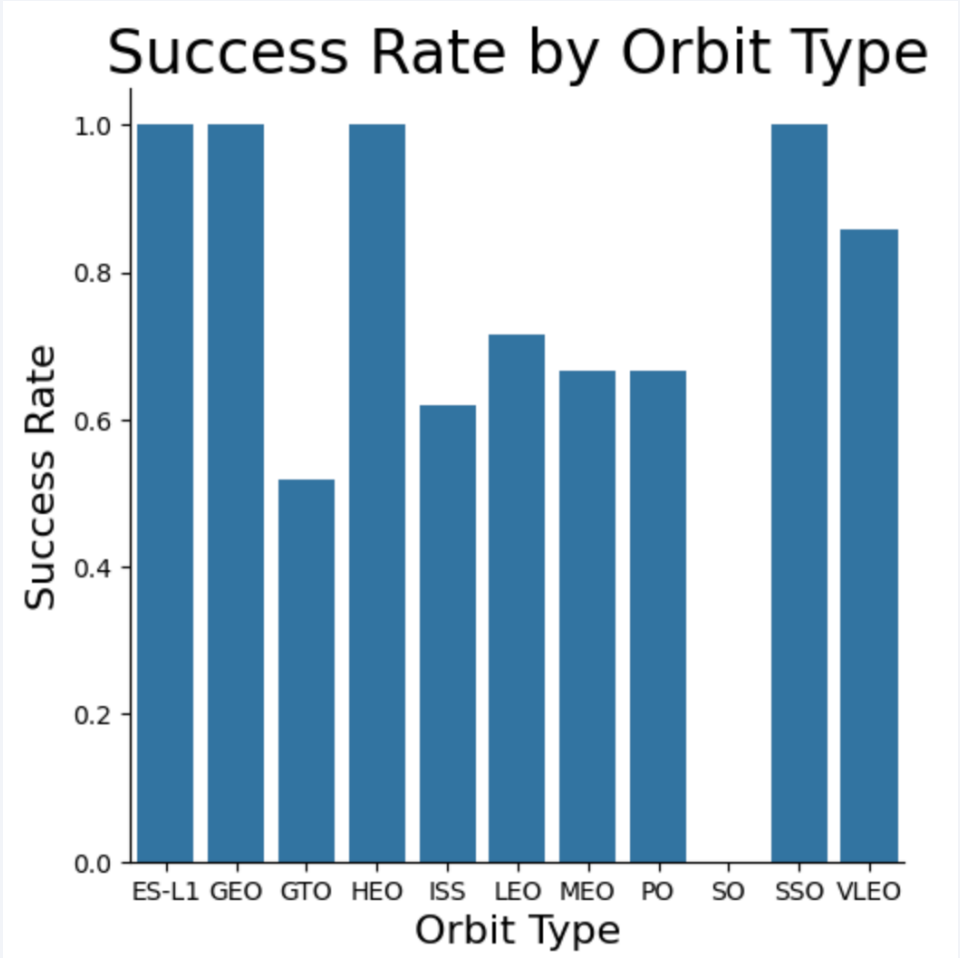
Observations:

- KSC LC 39A has a perfect track record for payloads less than 5,500kg
- Looking at VAFB-SC, we see no heavy launches (greater than10,000kg)
- The heavier payloads have seen greater success (i.e., payloads > 7,000kg)

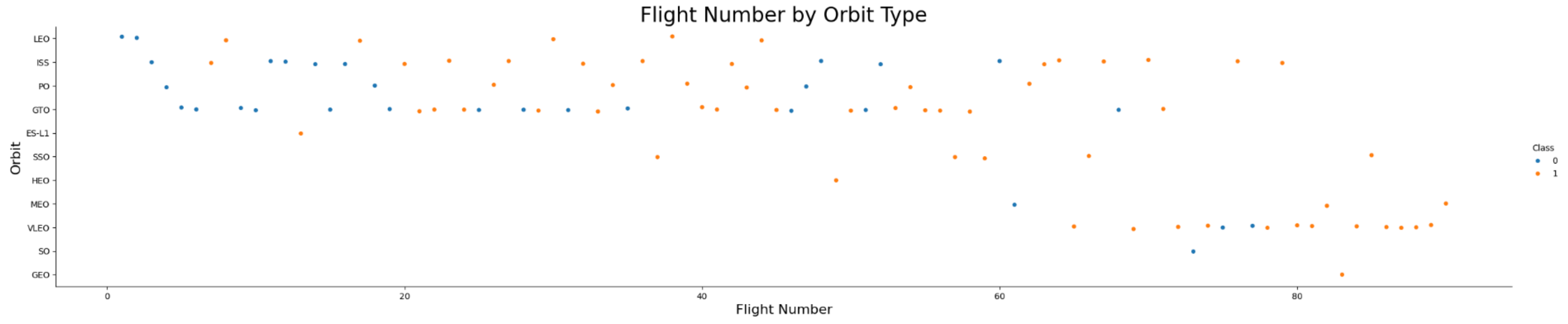
Success Rate vs. Orbit Type

Conclusions drawn:

- The following orbit types enjoy a 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- We can rank all other orbit types in descending order of success:
  - VLEO
  - LEO
  - MEO and PO (tied)
  - ISS
  - GTO
- Of all these, SO is the only orbit type that has never been successful.



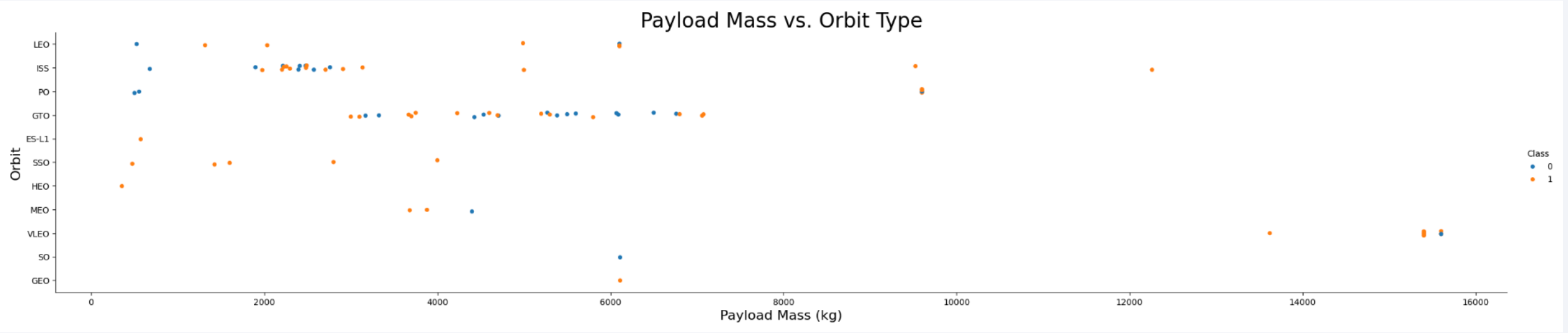
Flight Number vs. Orbit Type



Observations:

- With LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



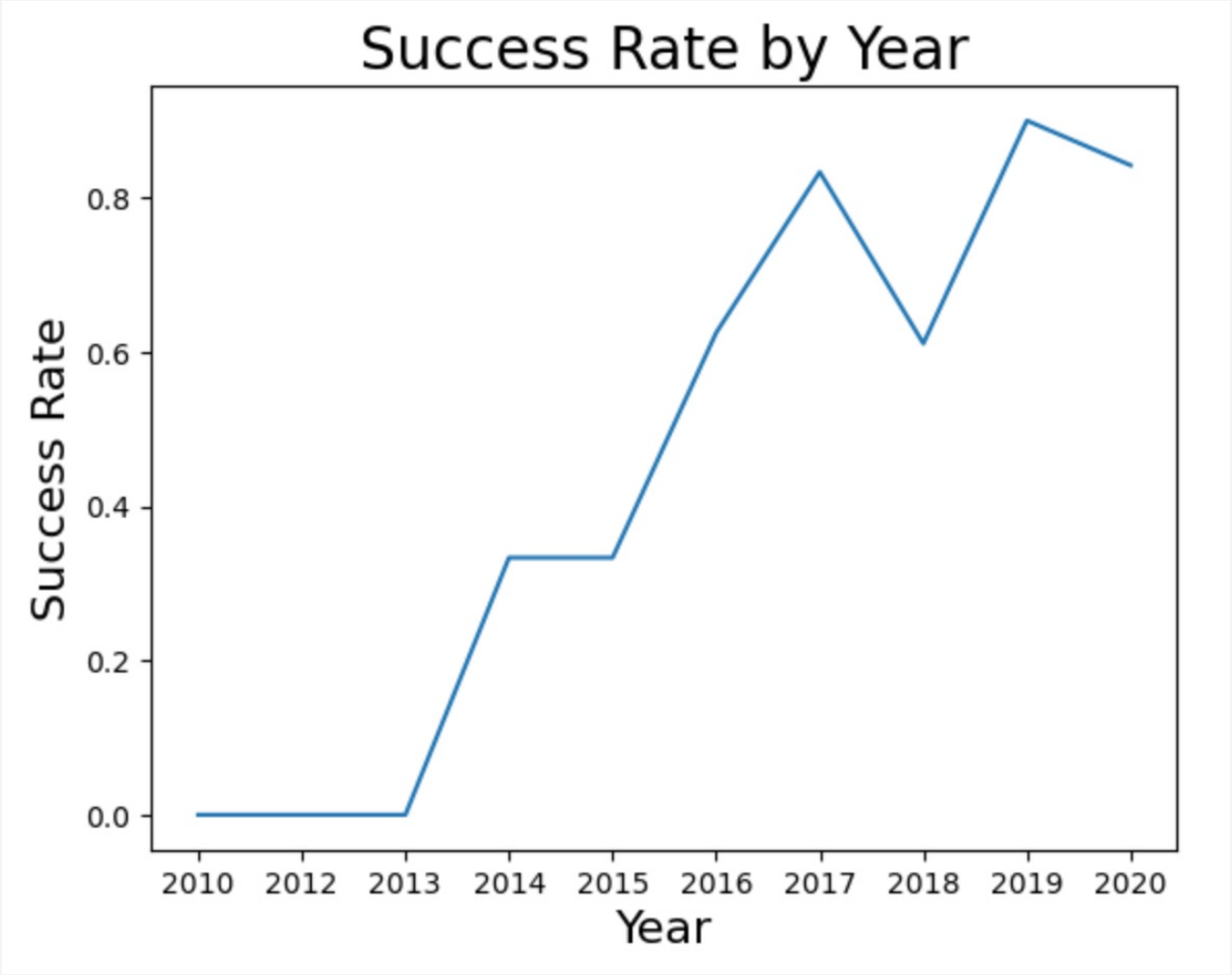
Observations:

- Looking at heavy payloads, we see more successful landing rate for Polar, LEO and ISS orbit types.
- Considering GTO orbit, it's challenging to distinguish between successful and unsuccessful landing, because we observe both types of outcomes in the data.

Launch Success Yearly Trend

Observations:

- SpaceX saw zero success from 2010 through 2013
- First success came in 2014, and remained flat through 2015
- Success rates rose dramatically through 2017, dropped sharply in 2018, then rose to new heights in 2019
- 2020 led to slightly lower success than 2019, but it was still the 2nd best year in company history



All Launch Site Names

- We used SQL to find the names of the unique launch sites
- This list was current at a point in time, but it may have changed since then.

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

```
%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'

- We used SQL to find 5 records where launch sites begin with `CCA`
- This simply demonstrates facility with the SQL language for filtering results.

Display 5 records where launch sites begin with the string '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__KG_	Orbit	Customer	Mission_Outcome	L
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- We use the aggregate function `SUM( )` to produce this result

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

```
%sql select sum(PAYLOAD_MASS__KG_) from spacextable;
```

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

sum(PAYLOAD_MASS__KG_)
619967

Average Payload Mass by F9 v1.1

- We use SQL to calculate the average payload mass carried by booster version F9 v1.1
  - Leveraging the aggregate function AVG( ) along with a predicate to filter on booster version type

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";
```

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

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

- Using SQL to find the dates of the first successful landing outcome on ground pad
- The `MIN( )` function applied to a `DATE` datatype achieves this result

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) from spacextable where mission_outcome = "Success";
```

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

min(date)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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 distinct(booster_version) from spacextable \
    where mission_outcome = "Success" \
    and payload_mass__kg_ between 4000 and 6000;
```

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

- Results column can be seen at the right
  - Notice the use of keyword `BETWEEN` in the SQL query to filter the desired results

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Using SQL to calculate the total number of successful and failure mission outcomes
- This query combines the aggregate function `SUM()` together with the `CASE WHEN` keywords to identify different sets of outcomes and total them

List the total number of successful and failure mission outcomes

```
%sql select sum(case when mission_outcome like 'Success%' then 1 end) as successful_missions, sum(case when mission_outcome like 'Failure%' then 1 end) as failed_missions from spacextable;
```

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

successful_missions	failed_missions
100	1

Boosters Carried Maximum Payload

- Using SQL to list the names of booster versions that have carried the maximum payload mass
  - The use of a subquery lets use find max payload and use it as a predicate

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery ¶

```
%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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Using `SUBSTR()` function lets us pull the month from a larger `DATE` datatype

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. ¶

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(date,6,2) as month, substr(date,0,5) as year, landing_outcome, booster_version, launch_site from spacextable where year = '2015' and landing_outcome like('Failure (drone%');
```

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

month	year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Using SQL to rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
  - Query uses `RANK()` and `ORDER BY count(*) DESC` along with `GROUP BY` to provide the necessary results

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 landing_outcome, count(*) as outcome_tally, rank() over (order by count(*) desc) as ranking from spacextable where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by ranking;
```

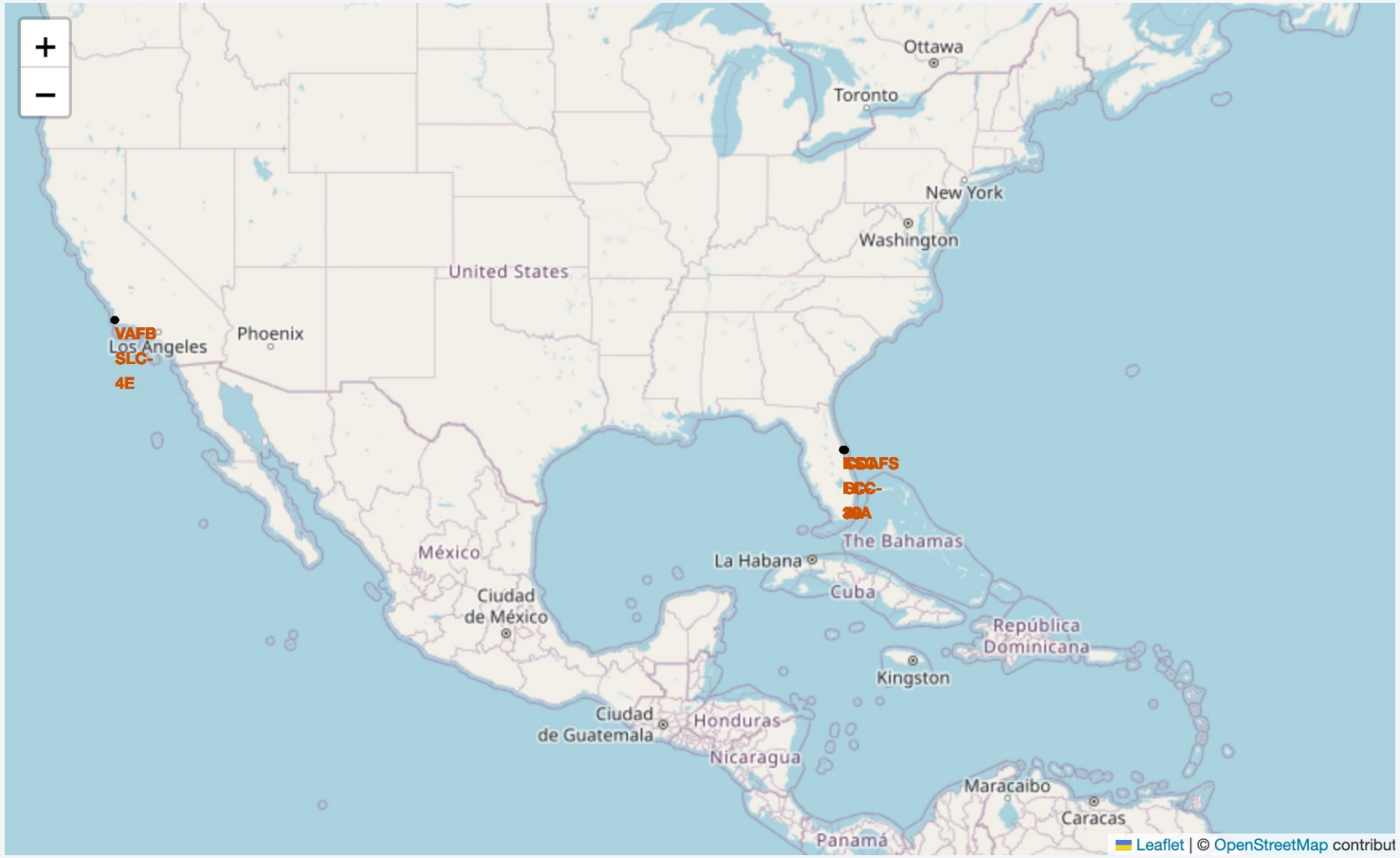
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	outcome_tally	ranking
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

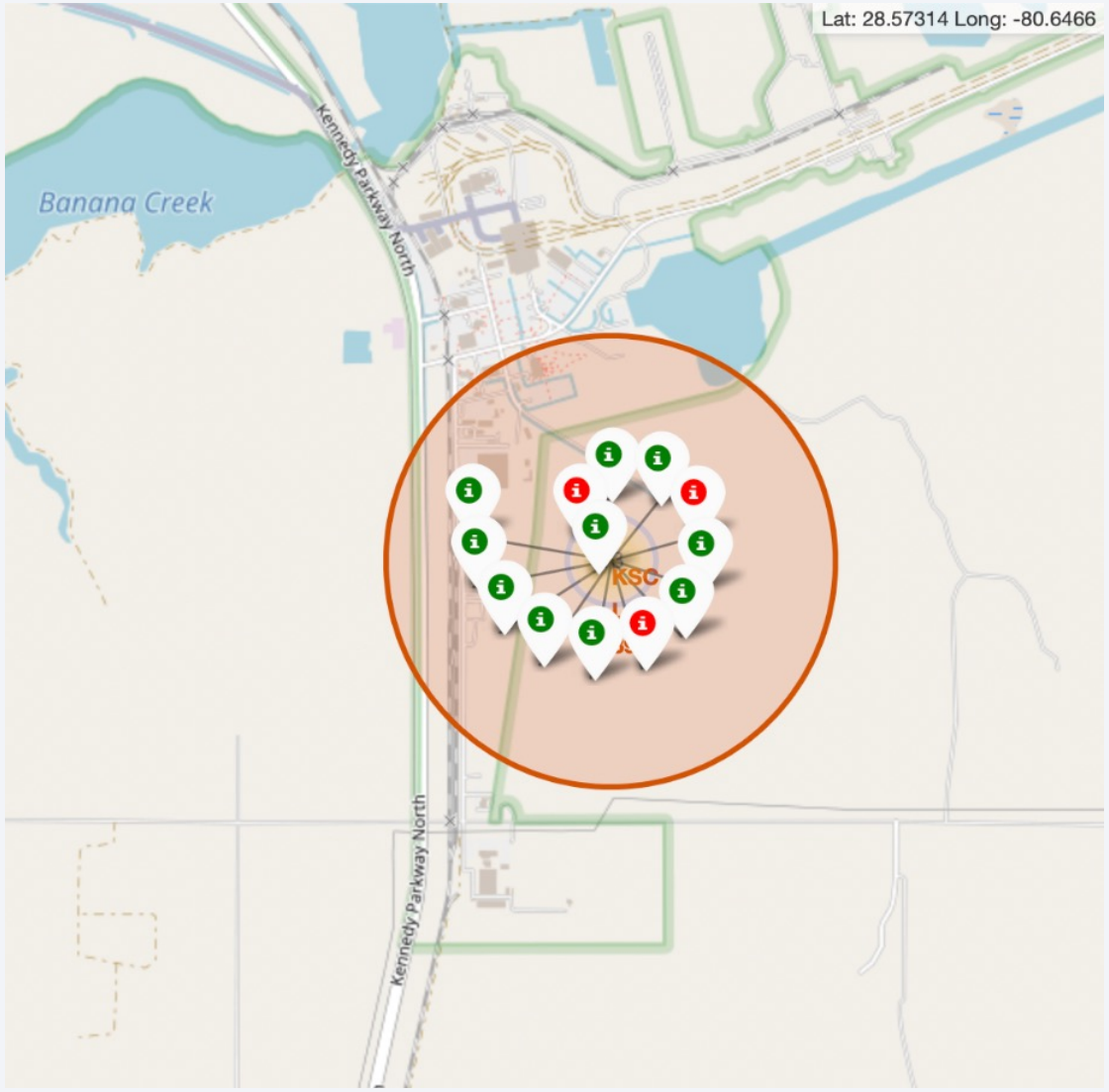

Section 3
Launch Sites
Proximities Analysis

Viewing Launch Site Locations across USA

- Launch sites all located in southern latitudes of the US
- Earth’s rotational velocity at equator exceeds 1000 mph
- Velocity is transferred to launched rockets due to inertia, helping achieve & maintain orbital speed
- All sites also situated near coastlines, enabling launch over oceans which reduces risks to populated areas



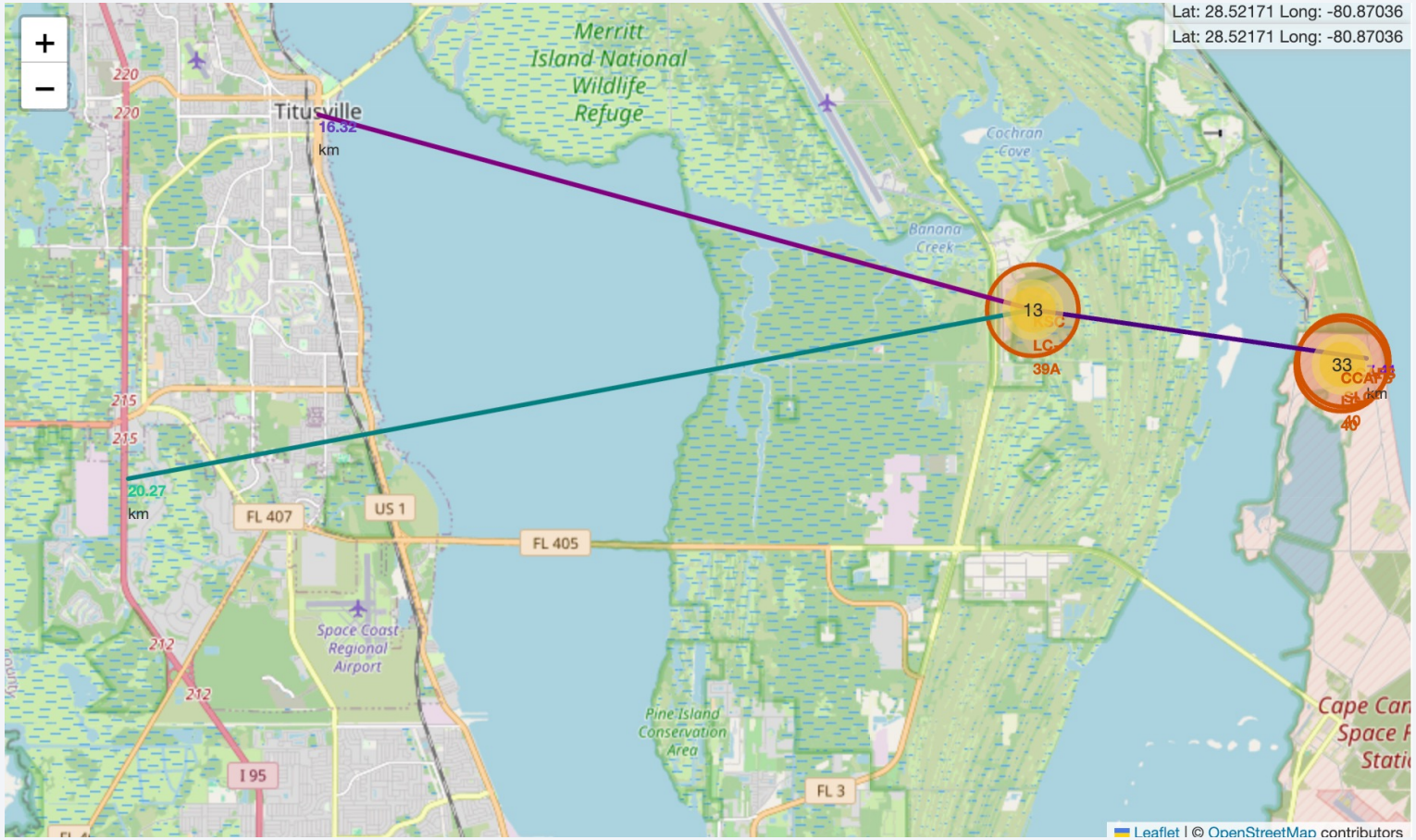
Using Colored Launch Markers to Visualize Performance

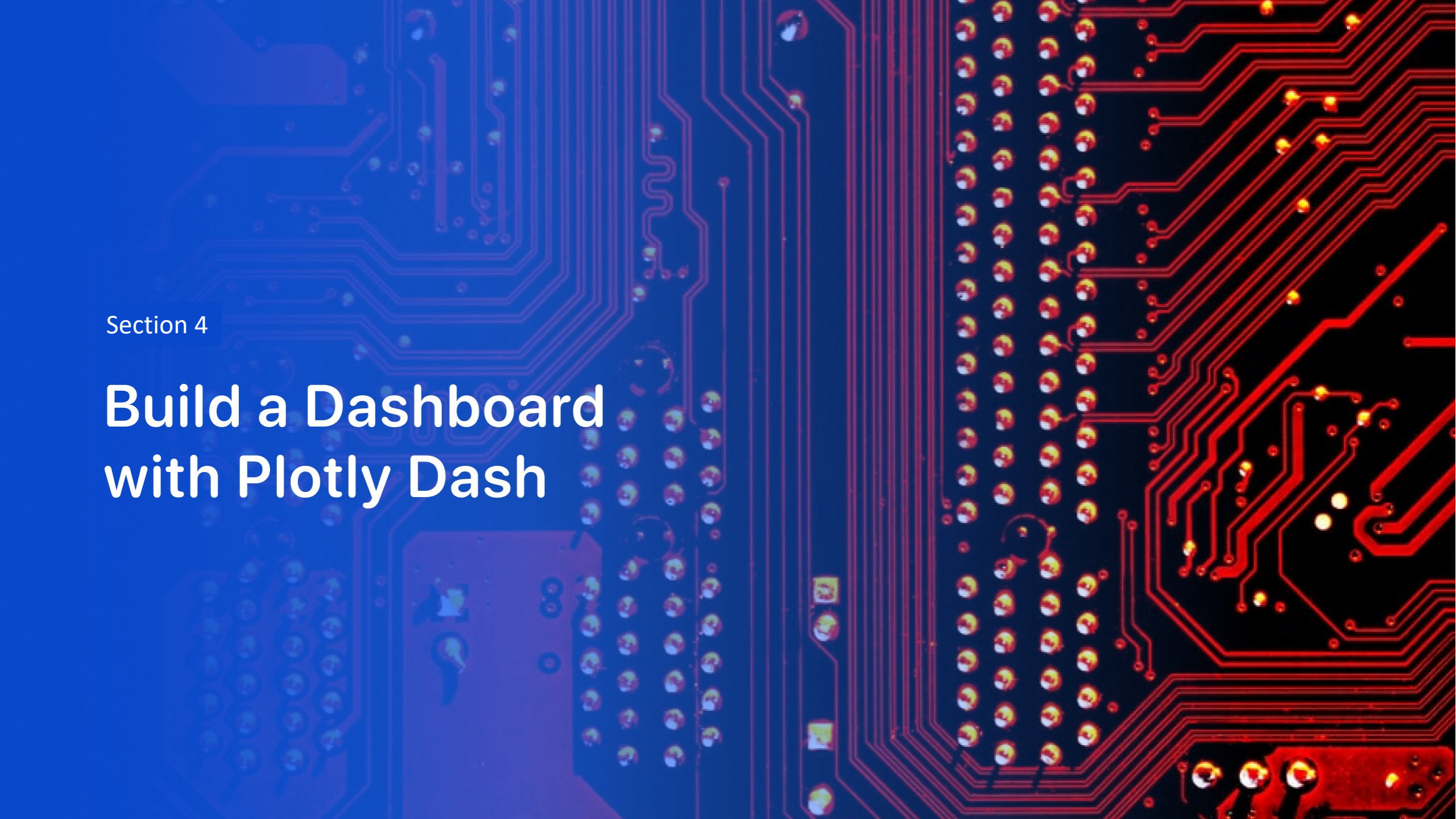


- The use of color on our Folium Map helps us to quickly visualize the success and failure at various launch sites
- As you might suspect:
  - Red markers indicate failure
  - Green markers indicate success
- Here we are looking down at Kennedy Space Center's Launch Complex 39A (KSC LC-39A)
- We can easily see the relatively high success rate at this particular site

Within Proximity of Kennedy Space Center

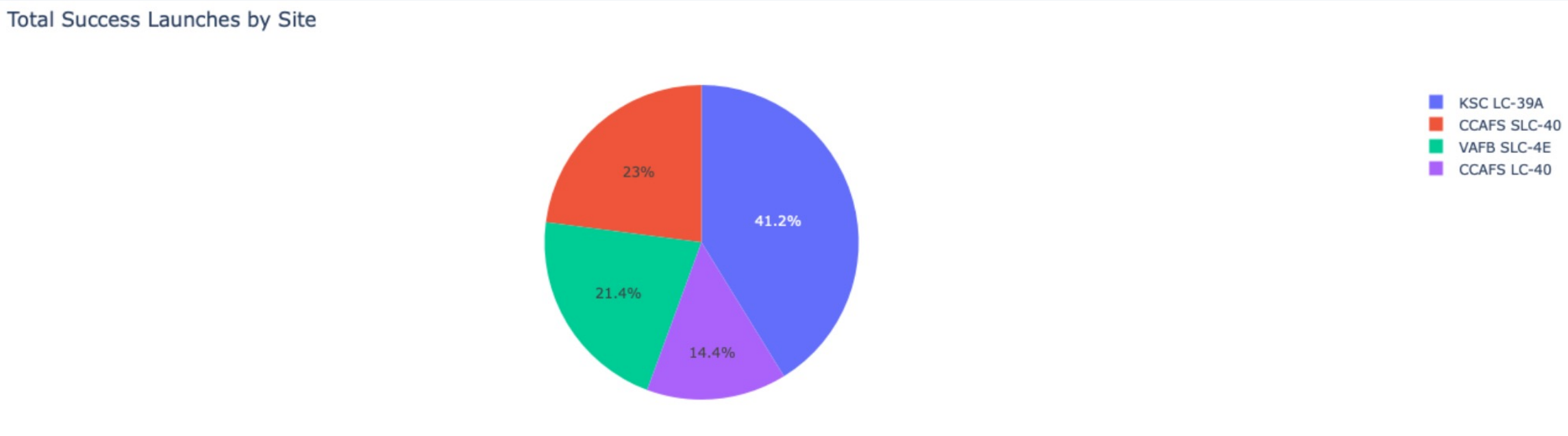
- Using Folium we can drop in a PolyLine (or 3 in this case) to mark straight-line distance to various locations.
- Here we can see how close KSC is (~16km) to the nearest city, Titusville.
- Also marked are distances to a nearby railway line and Cape Canaveral launch site, which itself is located directly on the coast




Section 4
Build a Dashboard with Plotly Dash

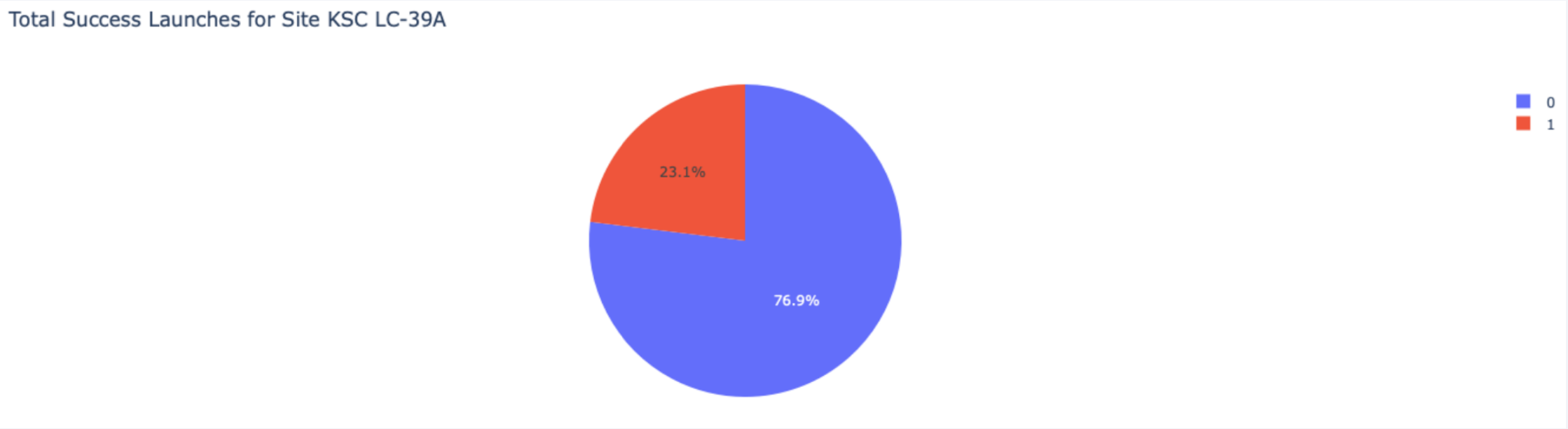
Assessing Success by Launch Site

- This pie chart shows the superior success rate at Kennedy Space Center's Launch Complex 39A (KSC LC-39A)
- Note how this aligns well with the Folium Map data we observed on slide 37



Drilling Down into Kennedy Space Center's Results

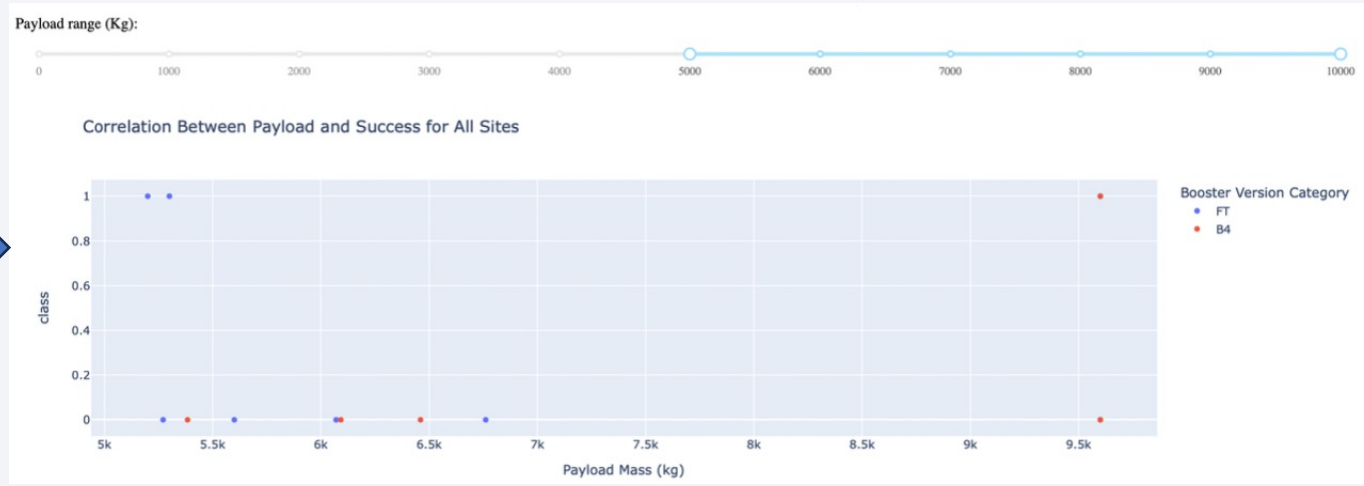
- Although KSC LC-39A is the most successful launch site, it is not perfect.
- This site has still seen a 23.1% failure rate



The Impact of Payload Mass on Success

- Using the Slider control, we can see the impact that payloads 5000 kg or smaller have on mission success
- Note the Slider filters results to show only rocket boosters which have performed missions of that size.

- Moving the Slider out to capture larger payloads, we see different boosters involved, with differing outcomes
- Conclusion: the "sweet spot" for payload mass lies somewhere between ~1900 and ~5500 kg



Section 5

Predictive Analysis (Classification)

Classification Accuracy

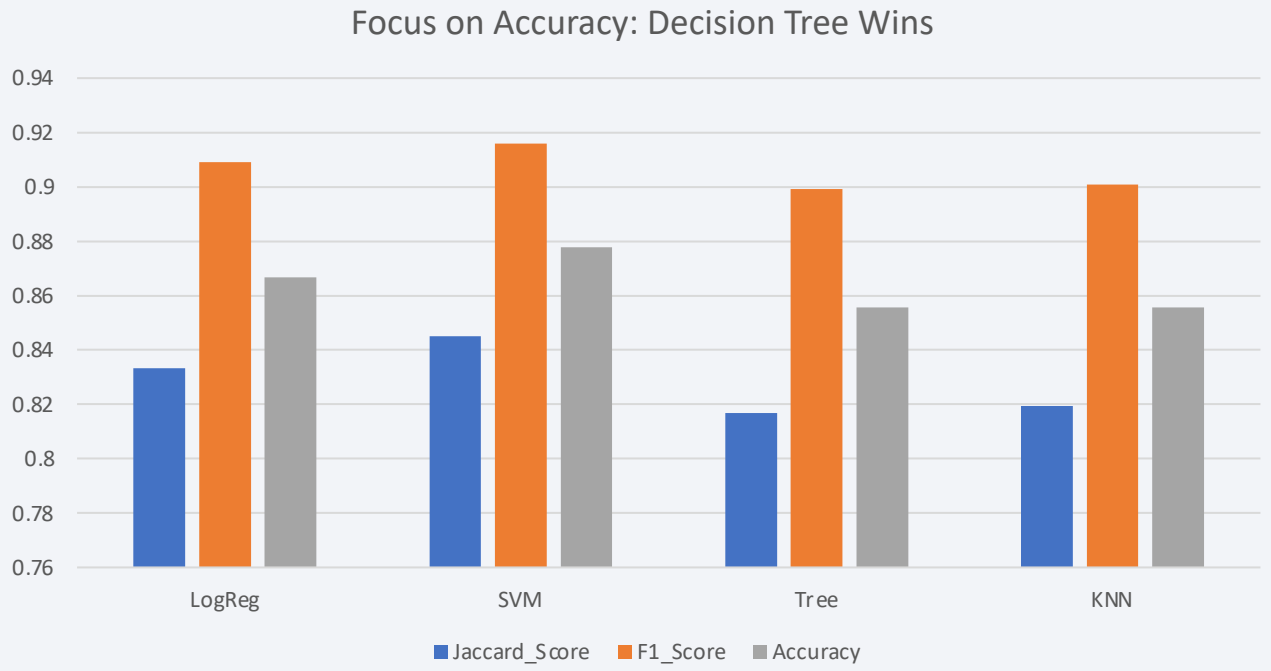
- Conclusion: Decision Tree classifier has the greatest accuracy.
  - The small number of sample in the Test dataset is not providing us enough coverage, but when we look at the entire data set, then we see the Decision Tree has a slight edge on Accuracy over the other models.

Accuracy on Test Set Only

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Accuracy on Full Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.816901	0.819444
F1_Score	0.909091	0.916031	0.899225	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556



Confusion Matrix for Decision Tree Classifier

- The confusion matrix of the Decision Tree Classifier can be seen at right.
  - This is the best performing model, able to distinguish between different classes.
  - With the small Test dataset (only 18 samples), it had no False Positives but 3 False Negatives.

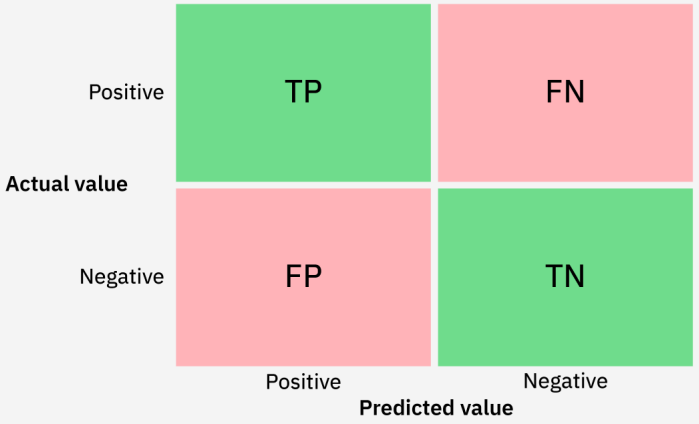
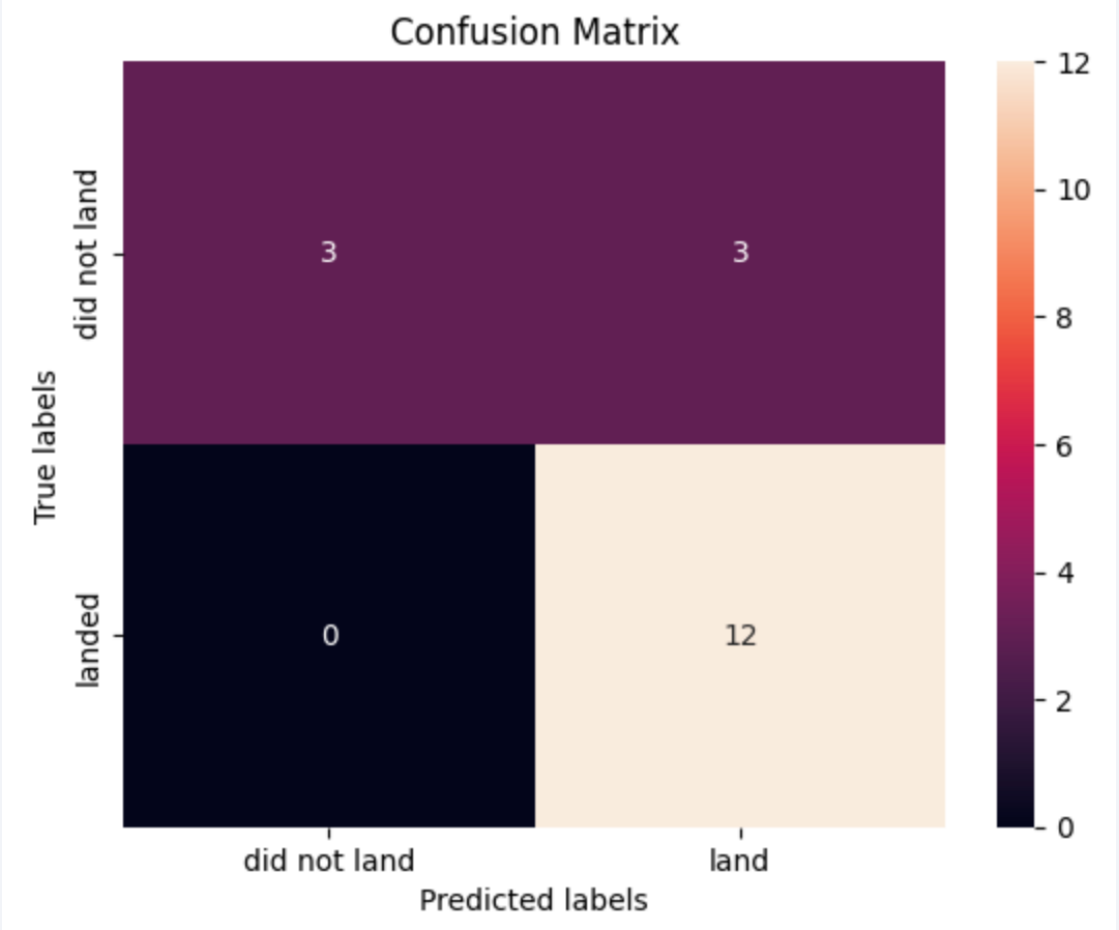


Image credit: IBM Developer
https://developer.ibm.com/tutorials/awb-confusion-matrix-r/



Conclusions

- Choose launch sites near equator and coasts
  - orbital velocity advantage and risk mitigation
- Planning factors such as site, mass, and orbit can predict launch success
  - lower payload mass launches have higher success rates
  - KSC LC-39A has highest launch success rate among all sites
  - 100% success for ES-L1, GEO, HEO, and SSO orbital trajectories
- Success rates increasing over years due to technological advancements
- Decision Tree Model best algorithm for analyzing launch data

Thank you!