IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joseph Litman
July 11, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection done via SpaceX API and webscraping Wikipedia
  - Data wrangling performed to create binary training data and fill in missing data
  - Exploratory data analysis done through visualization and SQL
  - Interactive visual analytics created with Folium and Plotly Dash
  - Predictive analysis done with classification models
- Summary of all results
  - Successful launch rates have increased as time passes.
  - Orbits with the highest success rates: ES-L1 (Lagrange points), GEO (geosynchronous), HEO (highly elliptical), SSO (Sun-synchronous).
  - Decision Tree model produces the most accurate predictions.

Introduction

- Space X provides rocket launches at a significant savings over their competitors due to the ability to reuse the first stage rocket.
- These savings will only be realized if the first stage rocket successfully lands.
- The question to be answered is: “Can we predict if a launch will land successfully?”

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data was collected via the Space X API and web scraping Wikipedia

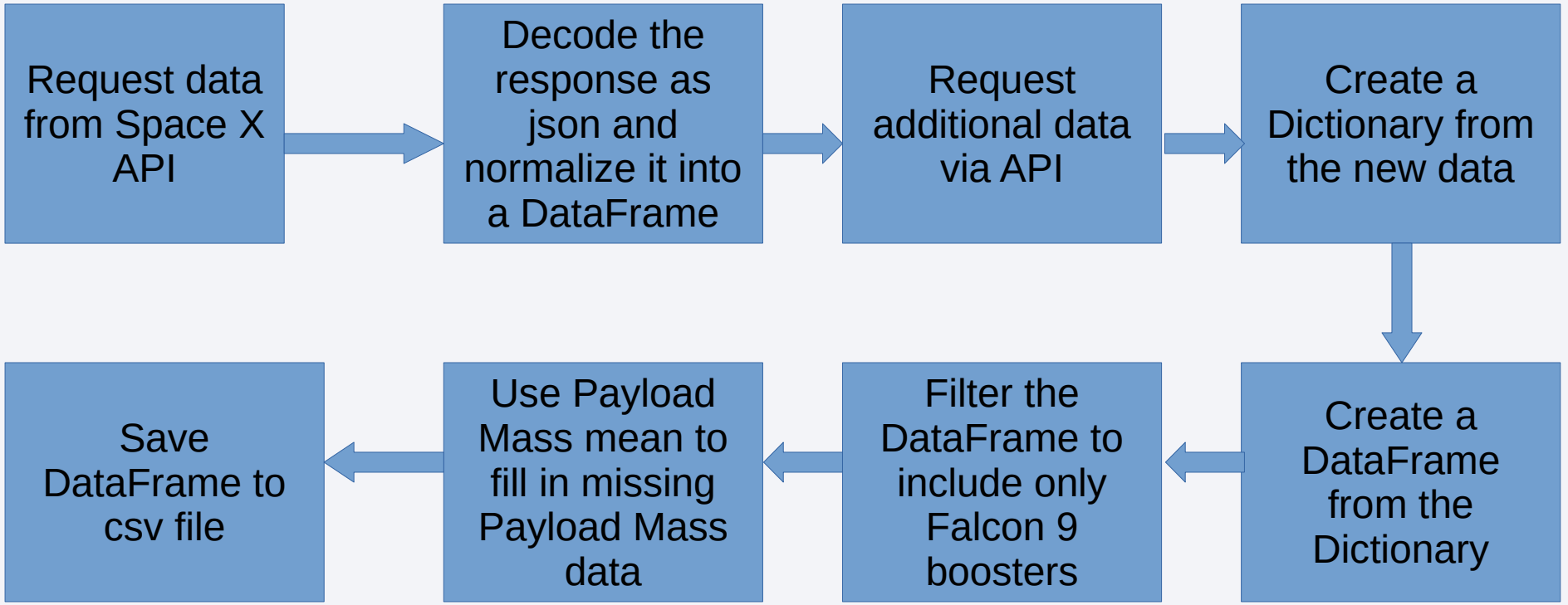
- Perform data wrangling
  - Converted various landing cases into a binary training label and missing data was filled in using the mean of existing data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models used: Logisitic Regression, SVM, Decision Tree, and KNN

Data Collection

Data sets were collected in two ways:

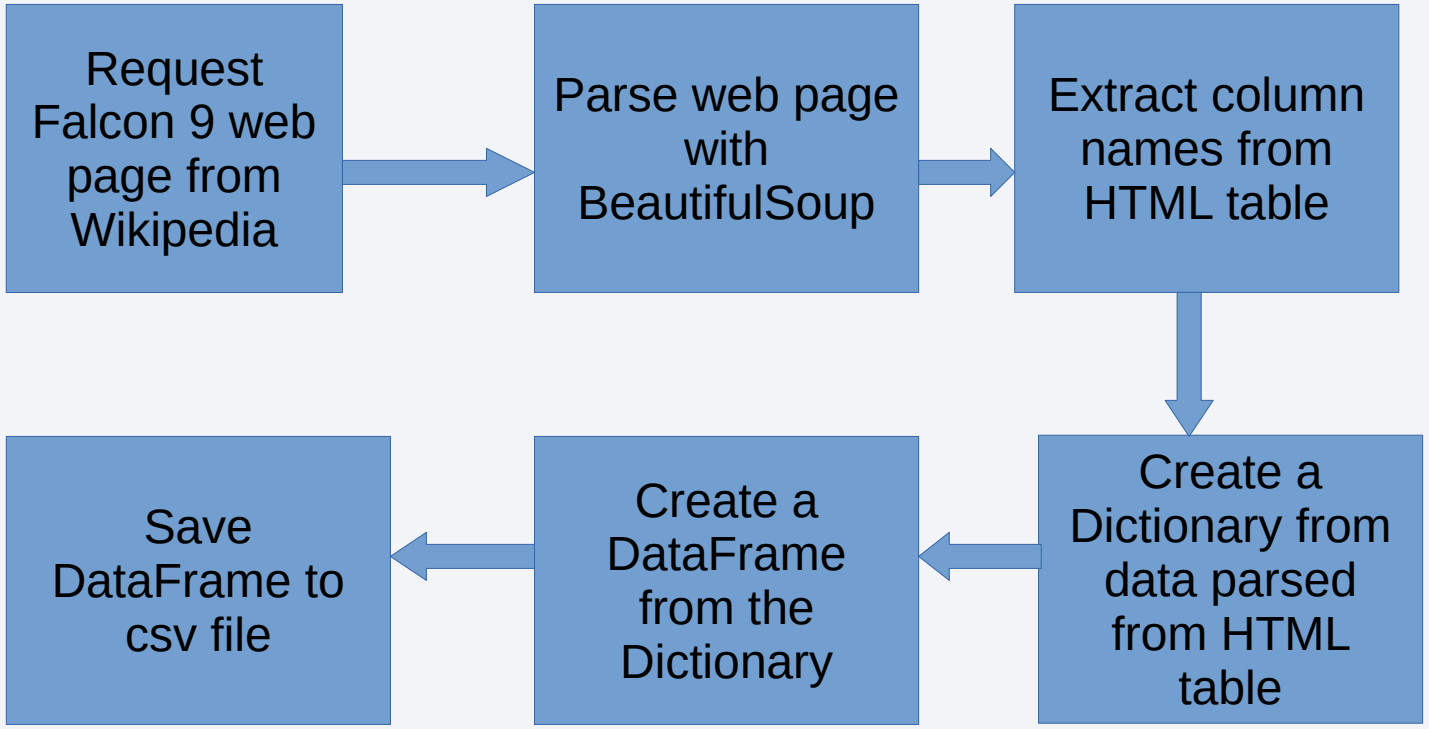
- Space X API
- Web scraping Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API



SpaceX API Data Collection Notebook

Data Collection - Scraping

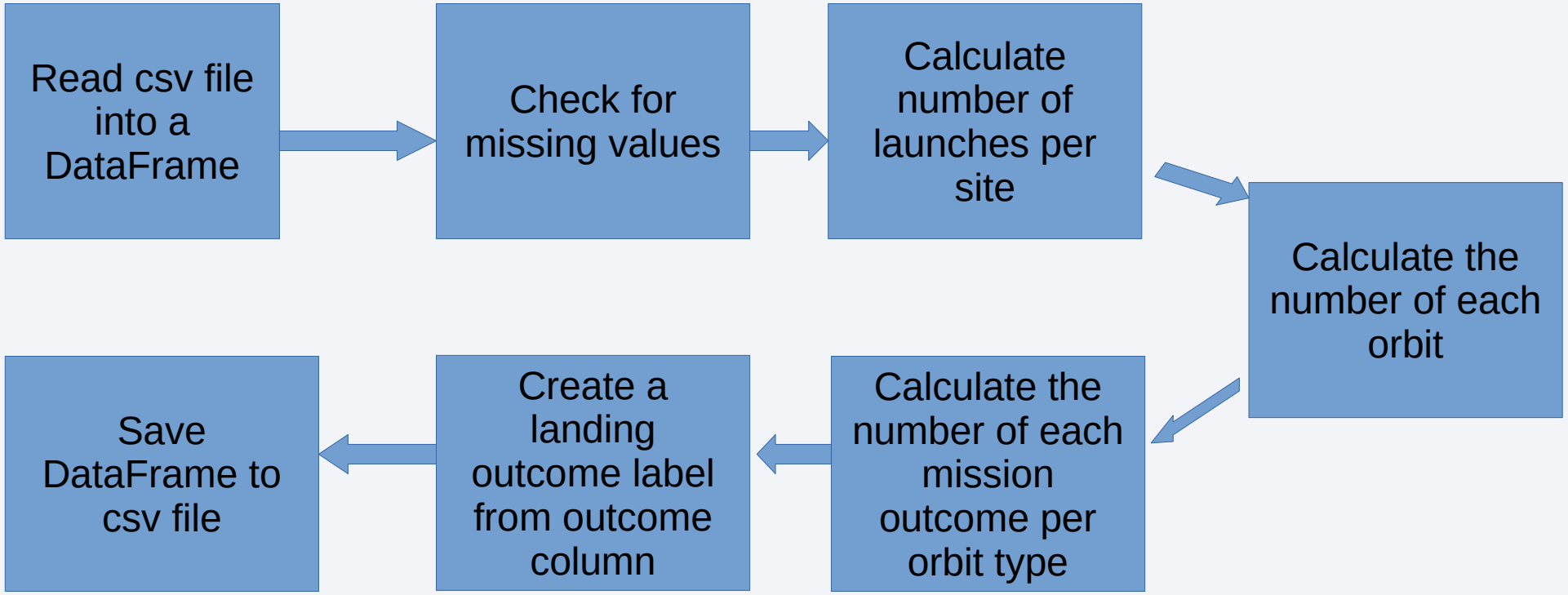


Web Scraping Notebook

Data Wrangling (1)

The original data provides an ‘Outcome’ column which contains mission results depending on where the booster landed and if it landed successfully. For example: a value of ‘True ASDS’ indicates that the booster successfully landed on a drone ship. This column was converted to a ‘Class’ column that indicated success (1) or failure (0).

Data Wrangling (2)



Data Wrangling Notebook

EDA with Data Visualization

Scatter plots were used to show payload mass vs flight number,flight number vs launch site, launch site vs payload mass, flight number vs orbit, and payload mass vs orbit. In all cases the colors of the points indicated success or failure. A scatter plot shows if a relationship exists between two numerical values. This is useful in determining which variables to use during machine learning.

A bar chart was used to show orbit vs success. A bar chart allows comparison of groups of data.

A line chart was used to show year vs average success rate. Line charts are useful when looking at time series data.

EDA with SQL

SQL queries performed:

- Select unique launch sites
- Select 5 records where the launch site begins with ‘CCA’
- Display total payload mass carried by boosters launched by ‘NASA (CRS’)
- Display average payload mass carried by F9 v1.1
- Show the date of the first successful landing on a ground pad
- List the boosters which have successfully landed on a drone ship and have a payload mass between 4000 and 6000 kg
- List total number of successful and failed missions
- List the names of booster versions that carried the maximum payload mass
- List records with month names, failure landings on a drone ship,booster versions, and landing site for the year 2015
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

Folium map provides several map objects for marking a map. The following were used:

- Circles to indicate location of launch sites.
- Marker to add labels to points on the map.
- MarkerCluster to show many markers that have the same coordinates. Different colors were used to indicate if a particular launch was successful.
- Mouse position to show coordinates of the mouse pointer
- Polyline to place a line between launch sites and points of interest (highways, cities, railroads).

Build a Dashboard with Plotly Dash

The plotly dash board has a drop-down list for choosing the launch site (or all launch sites).

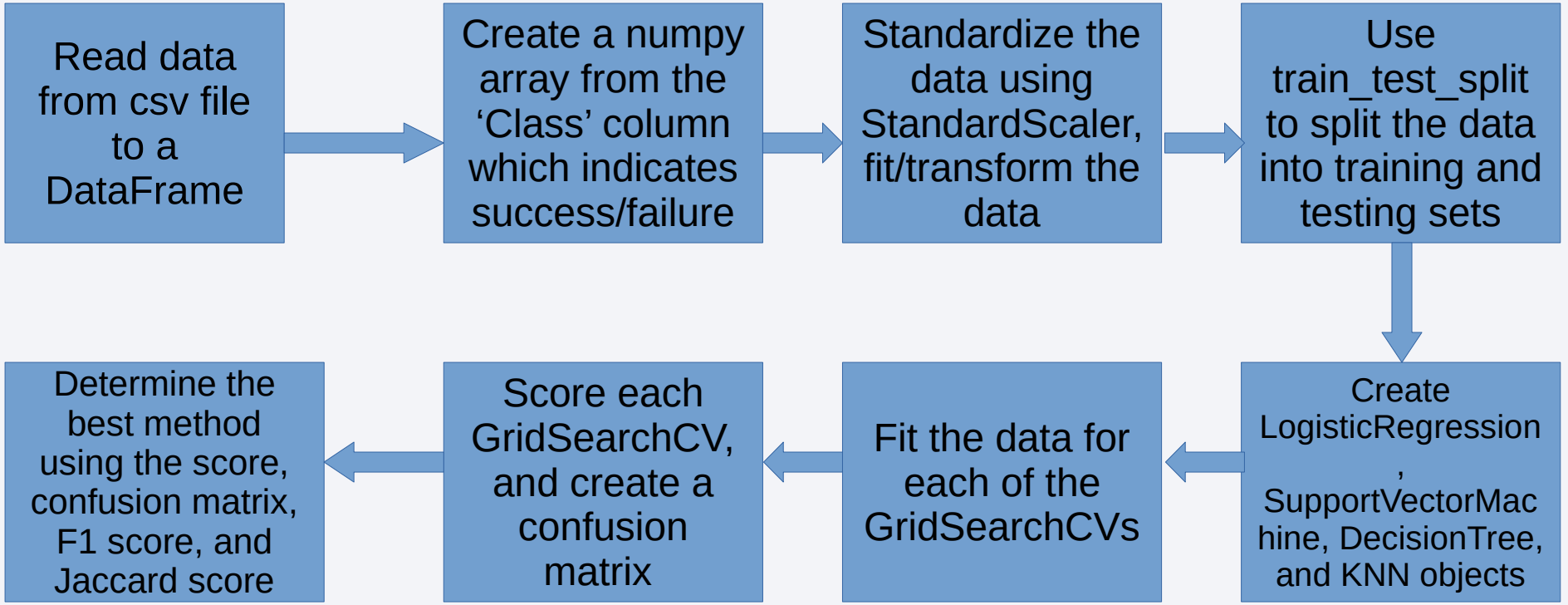
There is a pie chart that displays the percentage of successful launches either at each launch site individually or for all sites.

A slider allows the user to select a payload mass range

Scatter plot shows payload mass versus successful landing. Dots are colored according to booster version

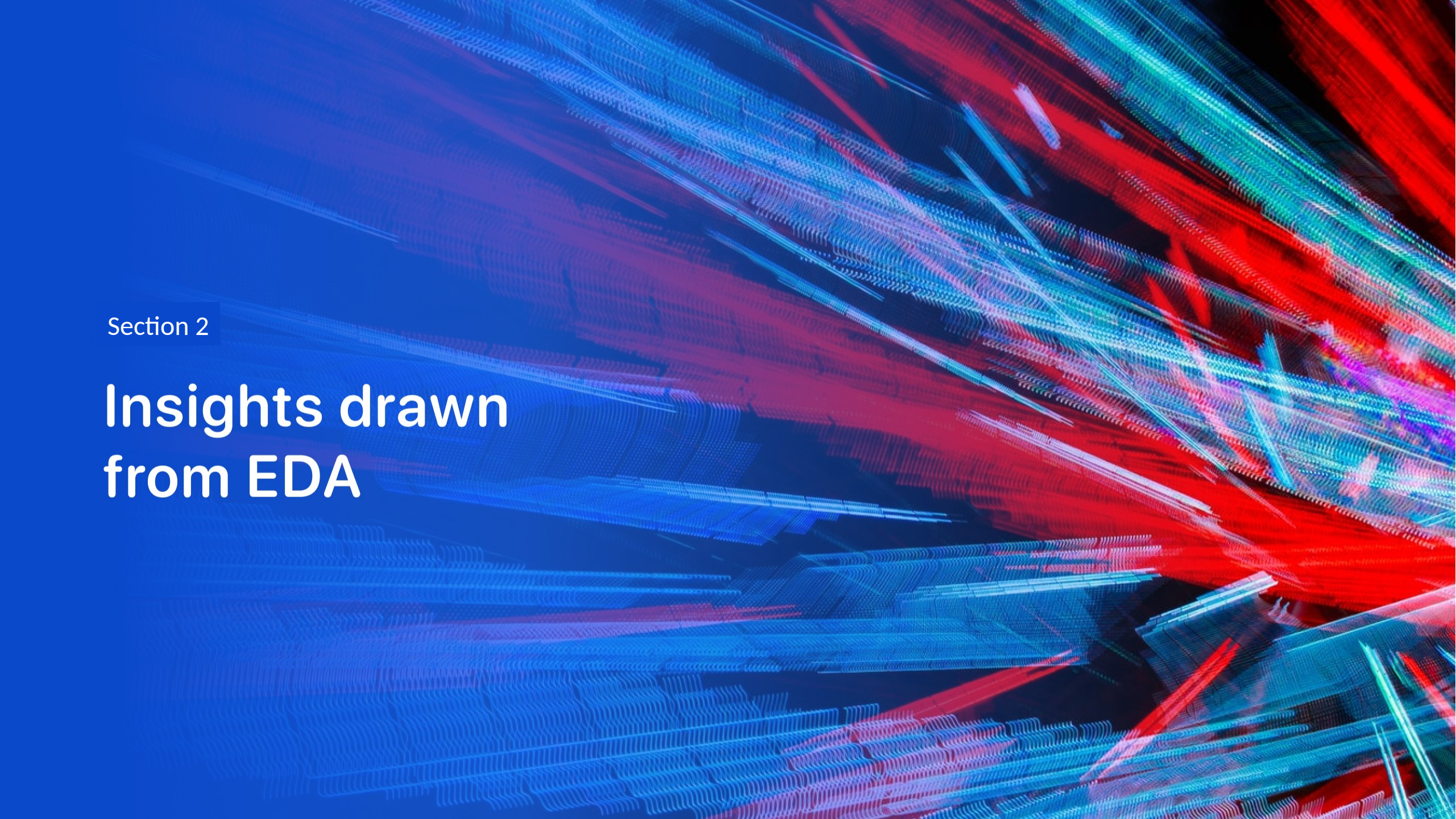
Plottly Dash Python Script

Predictive Analysis (Classification)

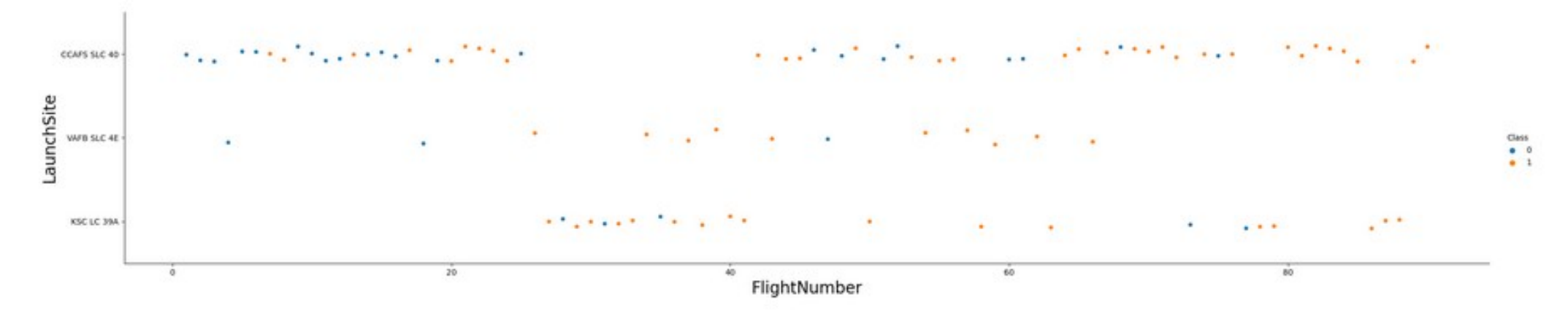


Results

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

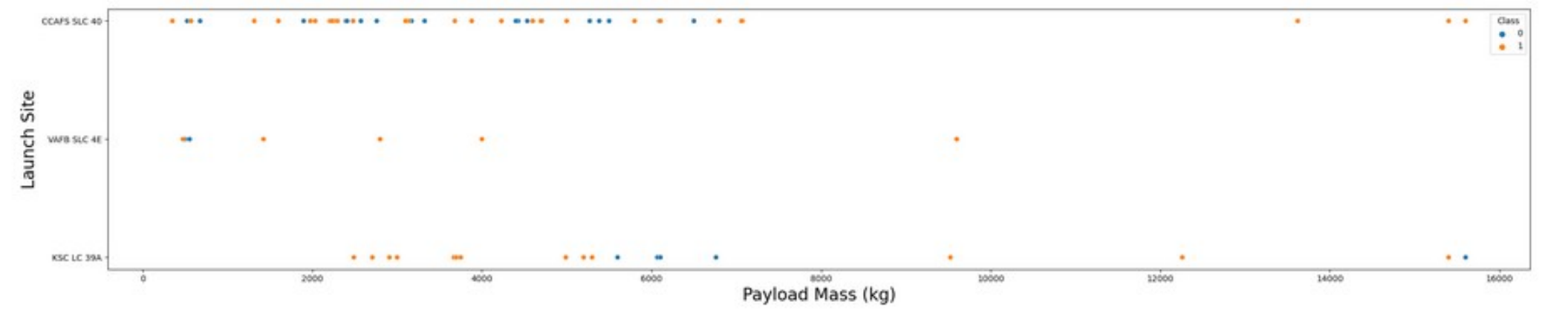

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



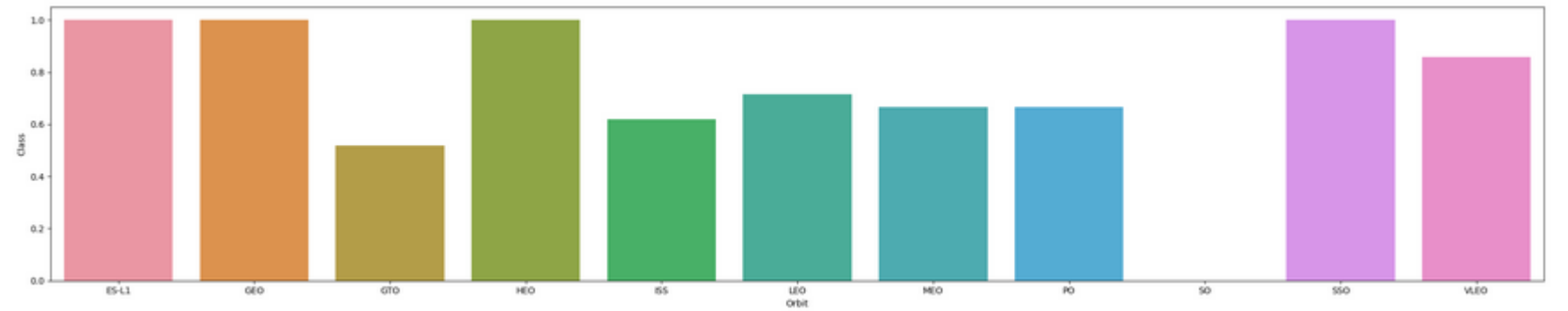
- Most launches have been from CCAFS SLC 40
- Earlier flight numbers were unsuccessful
- All flights after #80 have been successful

Payload vs. Launch Site



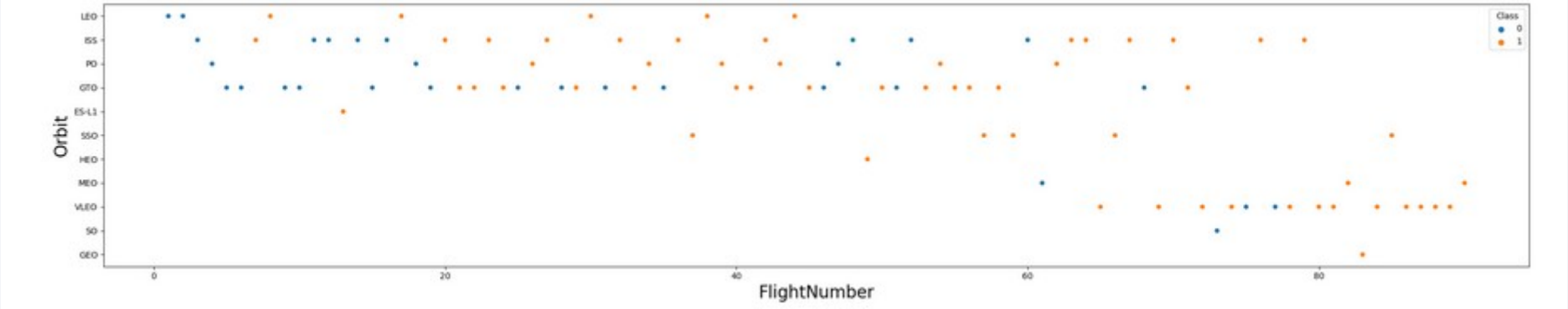
- No rockets heavier than 10000 kg have launched from VAFB SLC.
- KSC LC 39A has not successfully landed a booster when the payload is around 6000 kg.

Success Rate vs. Orbit Type



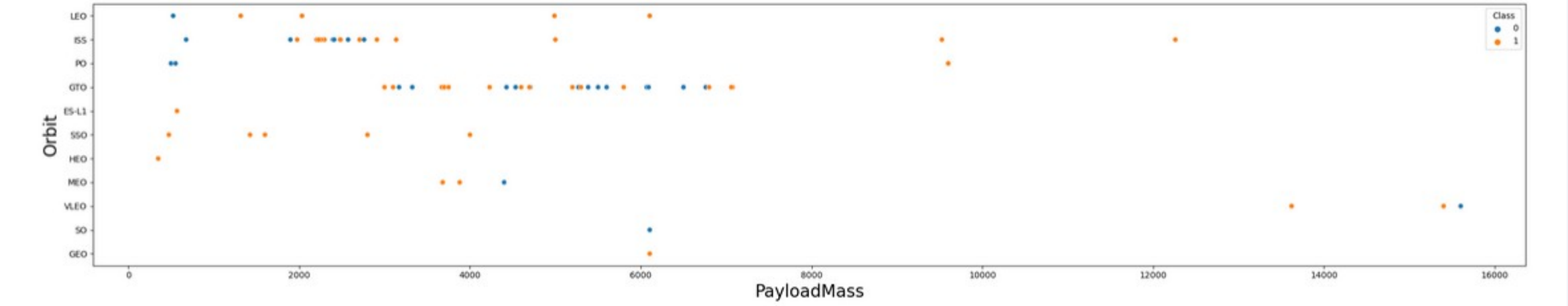
- Orbits ES-L1,GEO,HEO,and SSO have a success rate of 100%
- SO orbit has a 0% success rate

Flight Number vs. Orbit Type



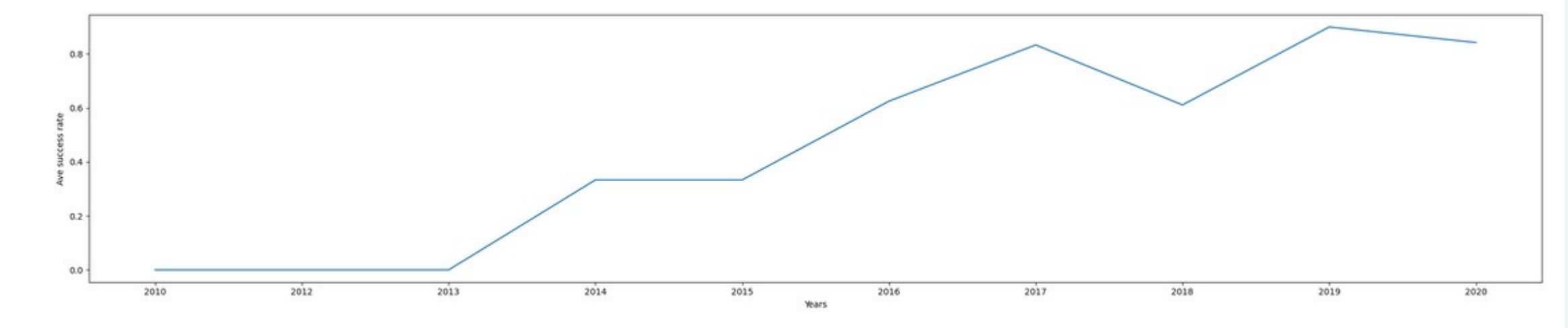
- LEO orbit success shows a positive correlation to flight number
- No relationship is seen in the GTO orbit between success and flight number

Payload vs. Orbit Type



- Success rates with heavy payloads are higher for orbits LEO, ISS, and Polar
- GTO orbit does not show a marked difference for success/failure at different payload masses.

Launch Success Yearly Trend



- Overall the success rate has been increasing since 2013.
- There are marked improvements from 2013-2014, 2015-2017, and 2018-2019.
- There are two decreases in success from 2017-2018 and 2019-2020.

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
None

The SQL statement lists the unique launch sites. The “DISTINCT” keyword ensures that each item returned from the database is unique.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL 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	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

The SQL statement returns 5 rows from the table SPACEXTBL where the Launch_Site begins with CCA.

Total Payload Mass

```
%sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" = "NASA (CRS)"
```

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

SUM("PAYLOAD_MASS__KG_")
45596.0

The SQL statement sums the values in the Payload_Mass_Kg column if the Customer is “NASA (CRS)”.

Average Payload Mass by F9 v1.1

```
%sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" like "F9 v1.1%"
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

The SQL statement calculates the average payload mass when the “Booster_Version” is F9 v1.1

First Successful Ground Landing Date

```
%sql select MIN("Date") from SPACEXTBL where "Landing_Outcome" = "Success (ground pad)"
```

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

MIN("Date")
01/08/2018

This statement selects the minimum Date when there was a successful ground pad landing.

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_" between 4000 AND 6000)
```

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

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

This statement returns the “Booster_Version” when the “Landing_Outcome” was equal to “Success (drone ship)” and the “Payload_Mass_Kg” was between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
success_count = %sql select count(*) from SPACEXTBL where "Mission_Outcome" LIKE "Success%"
fail_count = %sql select count(*) from SPACEXTBL where "Mission_Outcome" LIKE "Failure%"
print("Successful mission outcomes: {}".format(success_count))
print("Failed mission outcomes: {}".format(fail_count))
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
Successful mission outcomes: +----------+
| count(*) |
+----------+
|   100    |
+----------+
Failed mission outcomes: +----------+
| count(*) |
+----------+
|    1     |
+----------+
```

Two statements are used to get the total number of successful and failed missions.

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

A subquery is used to find the maximum payload values which are used in the where clause to return a list of Booster versions.

2015 Launch Records

```
%sql select Substr("Date",4,2) as "Month","Landing_Outcome","Booster_Version","Launch_Site" from SPACEXTBL \
where "Landing_Outcome" = "Failure (drone ship)" and substr("Date",7,4)='2015'
```

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

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

This statement shows launches that failed to land on a drone ship in 2015. Substr is used to retrieve the month and year since SQLite does not support this.

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

Show the count of landing outcomes in descending order for dates between 2010-06-04 and 2017-03-20

```
%sql select "Date", "Landing_Outcome",Count("Landing_Outcome") as "Count" from SPACEXTBL WHERE \
julianday(substr("Date",7)||'-'||substr("Date",4,2)||'-'||substr("Date",1,2)) BETWEEN\
julianday('2010-06-04') AND julianday('2017-03-20') GROUP by "Landing_Outcome" ORDER BY \
Count("Landing_Outcome") DESC
```

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

Date	Landing_Outcome	Count
22/05/2012	No attempt	10
22/12/2015	Success (ground pad)	5
04/08/2016	Success (drone ship)	5
01/10/2015	Failure (drone ship)	5
18/04/2014	Controlled (ocean)	3
29/09/2013	Uncontrolled (ocean)	2
28/06/2015	Precluded (drone ship)	1
12/08/2010	Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

Launch sites are shown in orange. All sites are on the coast and are relatively close to the equator.

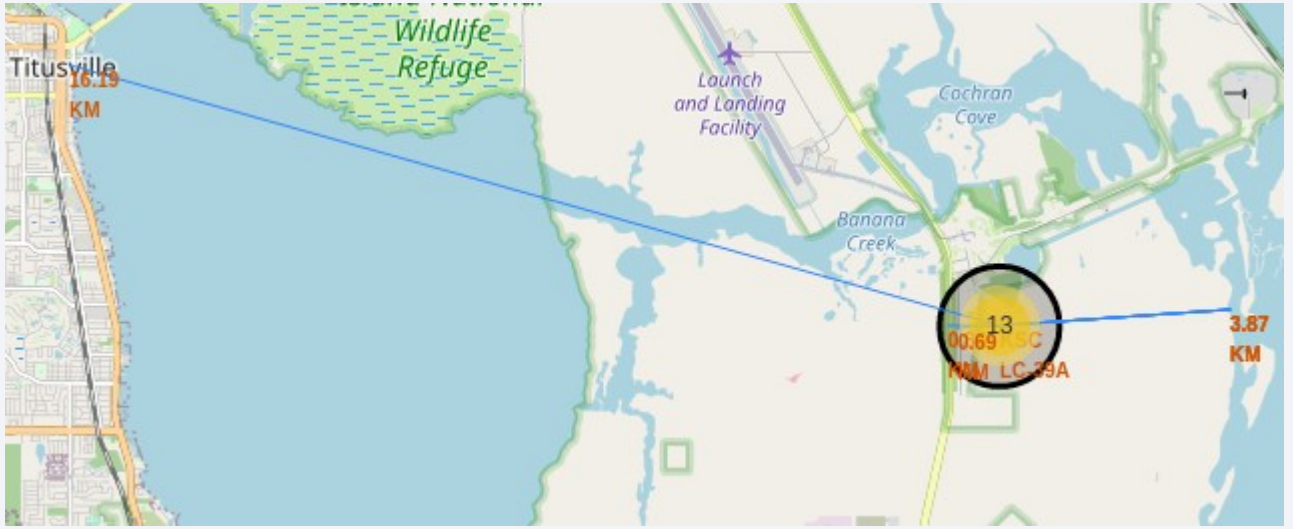
Launch Outcomes for KSC LC-39A

This map shows the launch outcomes for launch site KSC LC-39A.

The green icon indicates a successful landing while the red icon indicates failure. From this it is evident that this launch site has a rate success rate.

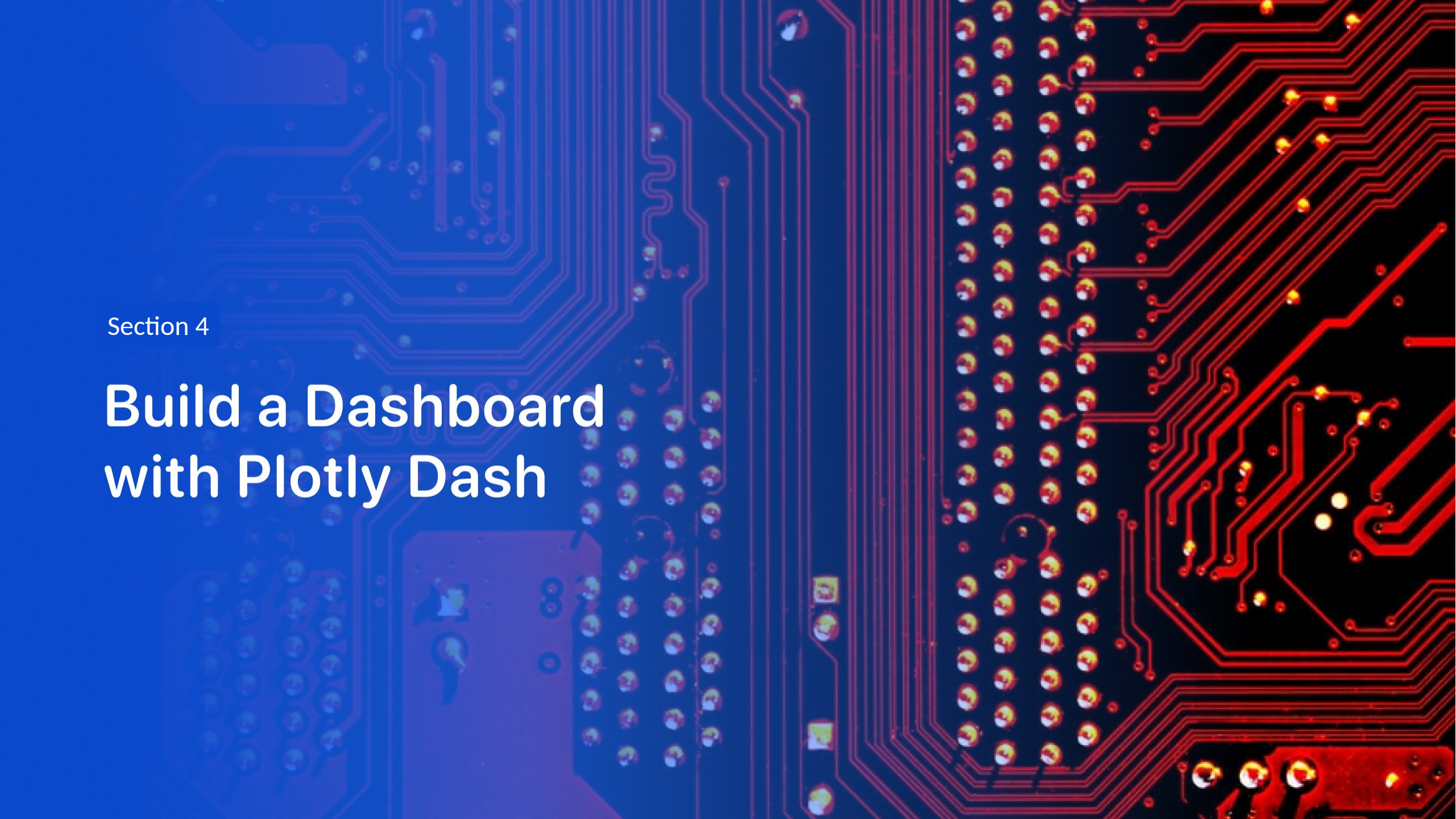
Proximities to KSC LC-39A

The map shows the distance between launch site KSC LC-39A and several locations.

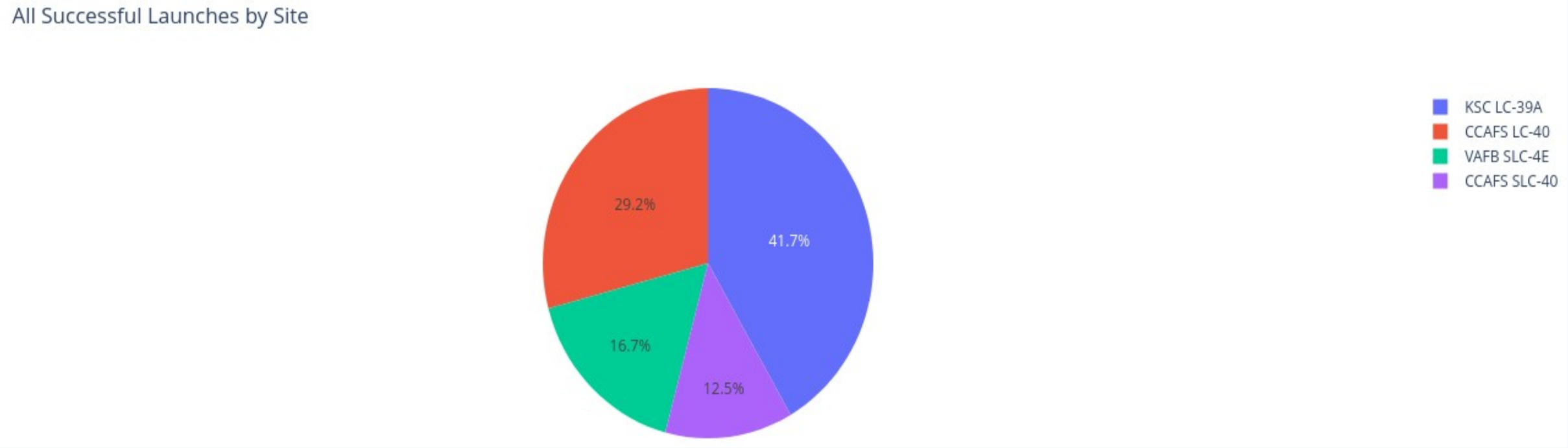


The city of Titusville is a little over 16 km from the launch site.

One coast line is just under 4 km from the launch site.

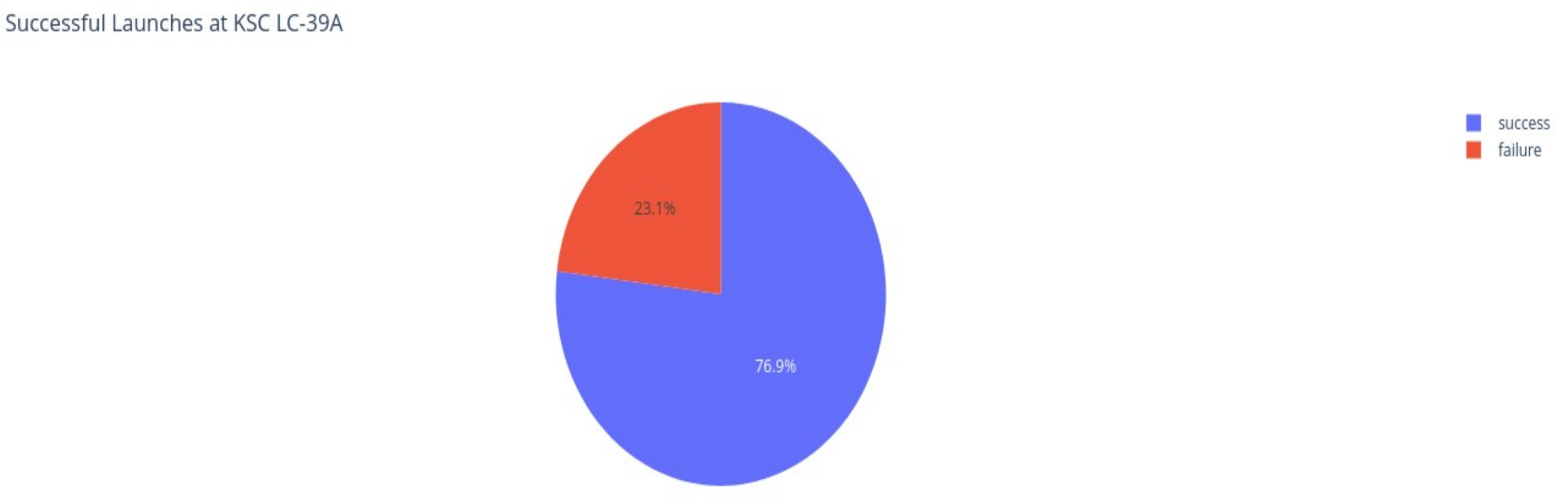

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



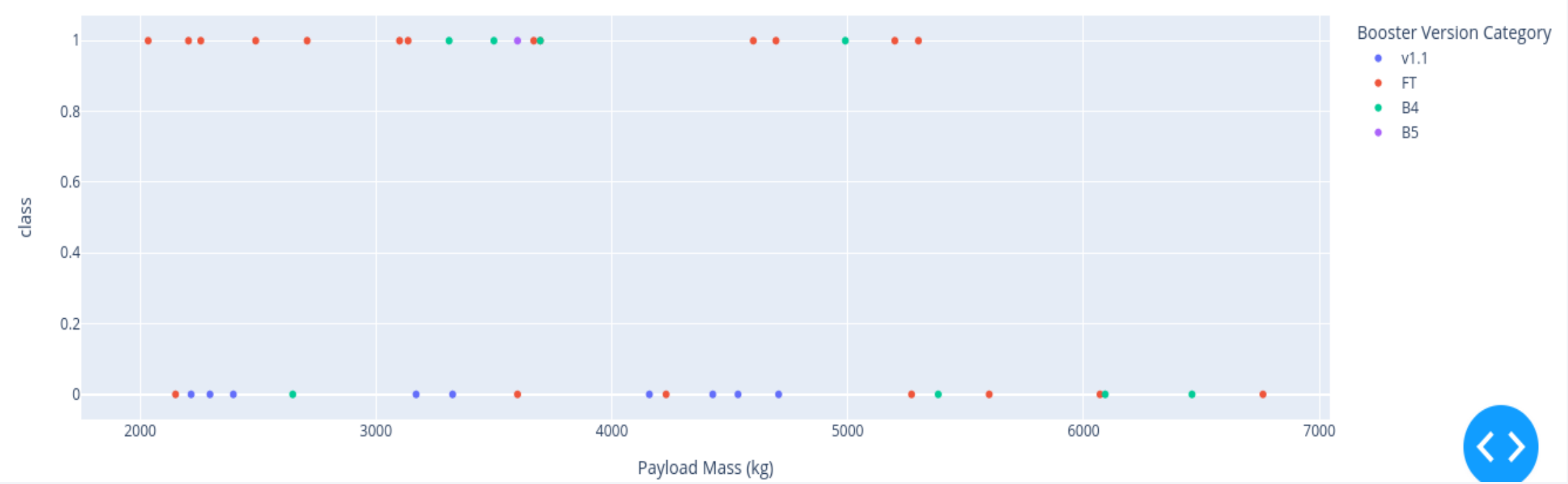
The pie chart shows successful launches by site. KSC LC-39A has the most successful launches while CCAFS SLC-40 has the least.

Ratio of Successful Launches at KSC LC-39A



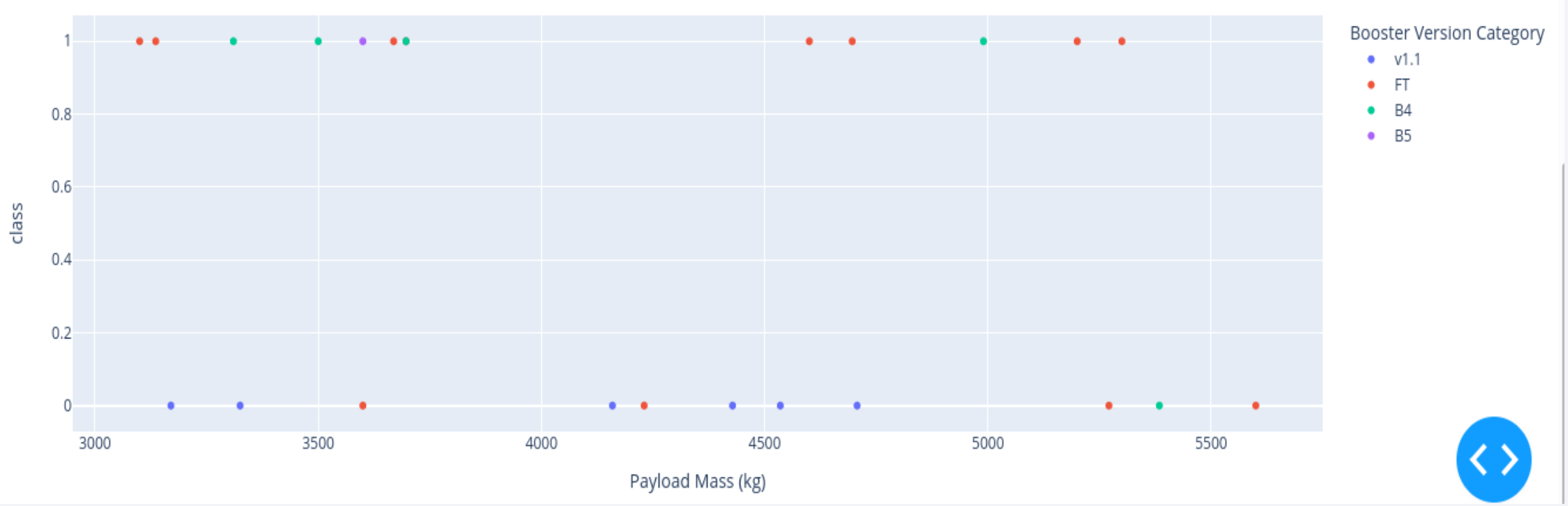
The pie chart shows the ratio of successful launches at KSC LC-39A. KSC LC-39A has the hightest percentage of successful launches at almost 77%.

Successful Launches by Payload Mass - 1



This chart shows payload mass in the range of 1500-7000 kg.

Successful Launches by Payload Mass - 2



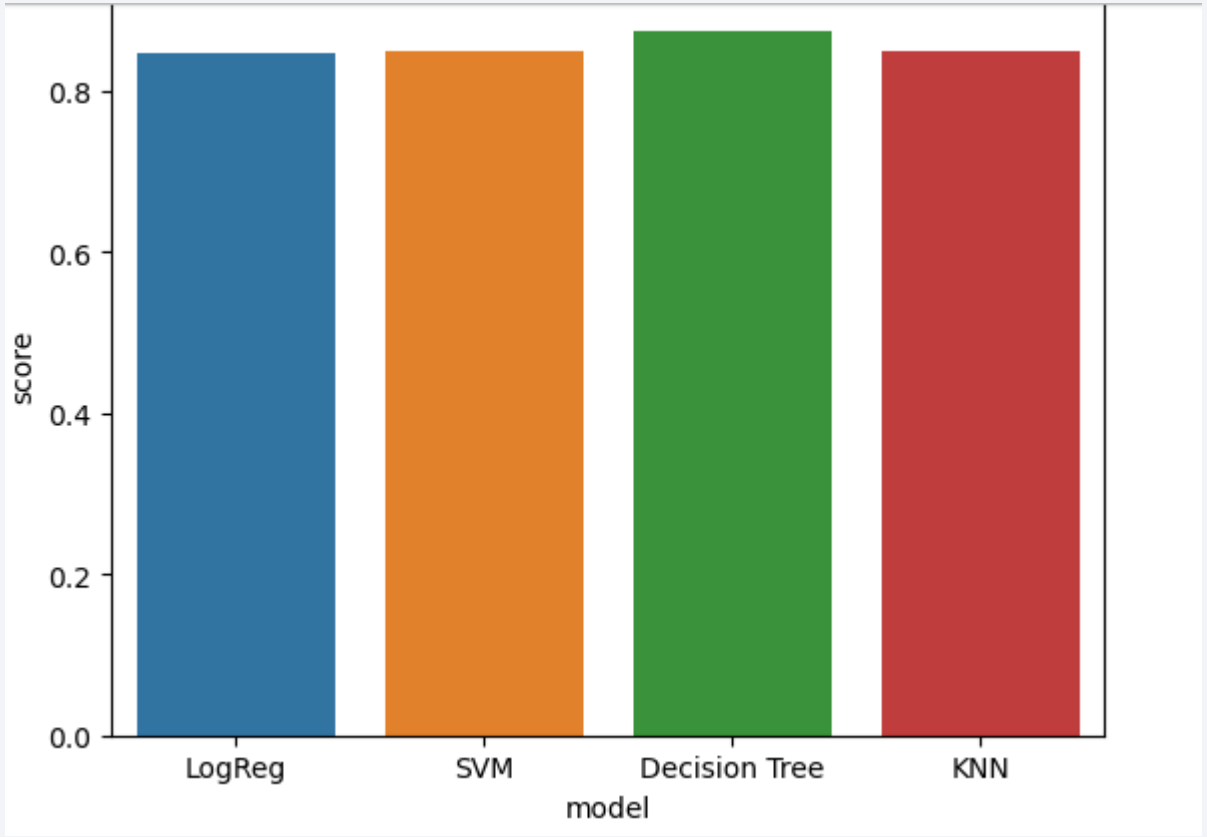
This chart shows payload mass in the range 3000-6000 kg. Payloads in the range of 2000-5000 kg have the most successful launches.

Section 5

Predictive Analysis (Classification)

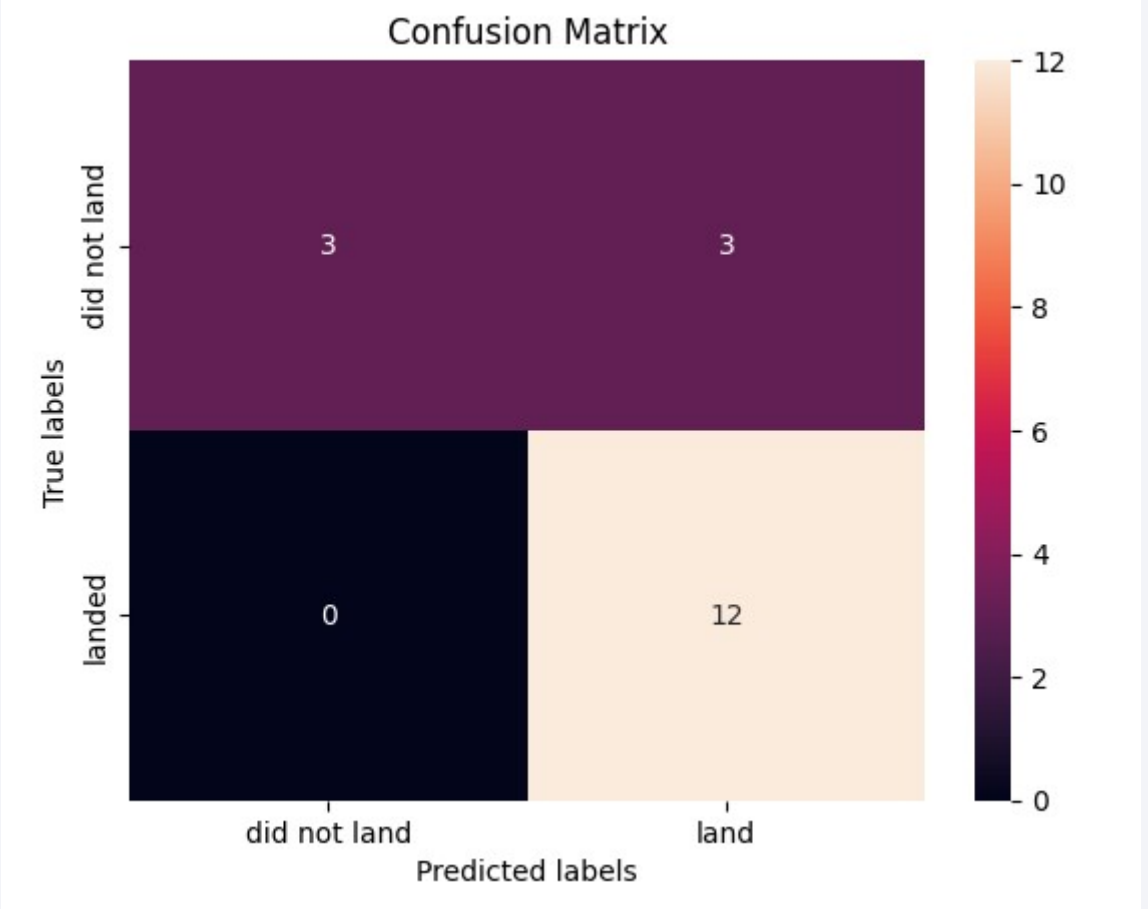
Classification Accuracy

- The Decision Tree model provides the highest classification accuracy



Confusion Matrix for Decision Tree

- The confusion matrix for the Decision Tree model shows:
- 3 true positives
- 3 false positives (Type 1 error)
- 0 false negatives (Type 2 error)
- 12 true negatives
- This shows that only 3 predictions were in error.



Conclusions

- Successful booster landings have increased in the time period from 2013 to 2020.
- The successful landing of a booster rocket is tied to launch location, payload size, and orbit.
- Orbits with the highest success rates are ES-L1, GEO,HEO, and SSO.
- Landing site KSC LC-39A has the highest percentage of successful landings.
- It is possible to predict the successful landing of a booster rocket with a accuracy greater then 80%. All predictive models performed well but the Decision Tree was slightly better than the others.

Appendix

All Python code, SQL queries, charts, and Jupyter Notebooks can be found at:

https://github.com/jlitman/DataScienceCapstone

Thank you!