IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>Rion Sato
<Date>December 14th, 2023

Outline

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

Executive Summary

Summary of methodologies

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Summary of all results

Exploratory Data Analysis:
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:
- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:
- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Project background and context

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. SpaceX's accomplishments include sending spacecraft to the International Space Station, launching satellite constellations that provide Internet access, and sending manned missions into space. SpaceX is able to do this because of the relatively low cost of rocket launches ($62 million per launch) due to its novel approach of reusing the first stage of Falcon 9 rockets. Other providers that cannot reuse the first stage cost more than $165 million per launch. The launch price can be determined by determining whether the first stage will land. To do so, public data and machine learning models can be used to predict whether SpaceX, or a competing company, can reuse the first stage.

Problems

➢ What factors contribute to a higher landing success rate for the first stage?

Things to explore to solve the above problems include:

- Effect of payload mass, launch location, number of flights, and orbit on first stage landing success
- Landing success rate over time
- Best Prediction Model for Landing Success (Binary Classification)

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology**

using SpaceX REST API and web scraping techniques

- **Perform data wrangling**

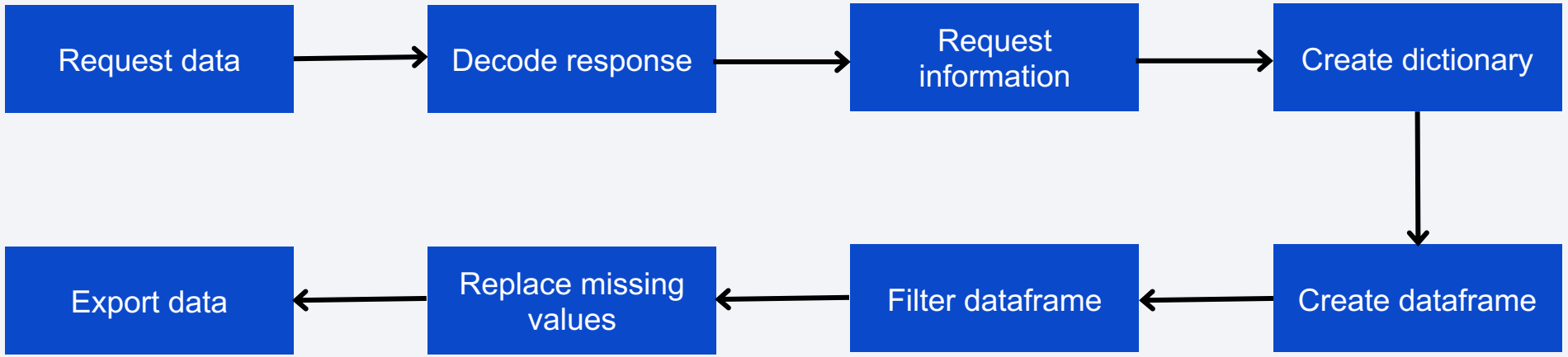
by filtering the data, handling missing values and applying one hot encoding to prepare the data for analysis and modeling

- **Perform EDA using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**

using classification models and tune and evaluate models to find best model and parameters

Data Collection – SpaceX API

Flowcharts

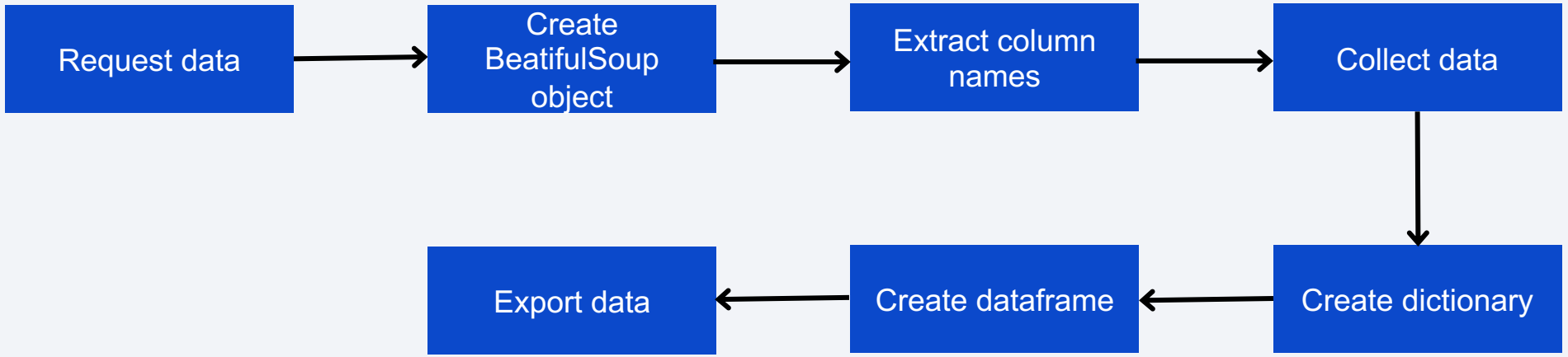


GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Flowcharts

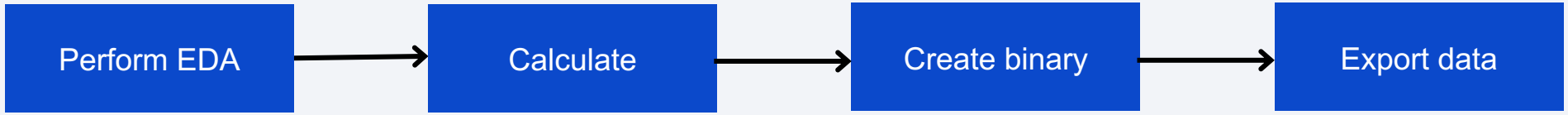


GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week1/jupyter-labs-webscraping.ipynb

Data Wrangling

Flowcharts



Note(How data were processed):
When we look at Landing Outcome, we find that Landing was not always Successful and Landing Outcome has different types of outcomes.
To make these data easier to handle and understand, I converted outcomes into 1 for a successful landing and 0 for an unsuccessful landing.

GitHub URL
https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week1/labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

Charts

- Flight Number vs Payload Mass (kg) (scatter plot)
- Flight Number vs Launch Site (scatter plot)
- Payload Mass (kg) vs Launch Site (scatter plot)
- Success Rate vs Orbit Type (bar chart)
- Flight Number vs Orbit Type (scatter plot)
- Payload Mass (kg) vs Orbit type (scatter plot)
- Year vs Success Rate (line chart)

Analysis(The reason why I used those charts)

- **View relationship** by using **scatter plots**. The variables could be useful for machine learning if a relationship exists.
- **Show comparisons** among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value.
- **Line charts** show changes over time.

GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week2/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Queries

Display:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added blue **circle** at **NASA Johnson Space Center's coordinate** with a **popup label** showing its name using its latitude and longitude coordinates
- Added red **circles** at **all launch sites coordinates** with a **popup label** showing its name using its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

• Added **colored markers** of **successful** (green) and **unsuccessful** (red) **launches** at each launch site to show which launch sites have high success Rates

Distances Between a Launch Site to Proximities

• Added **colored lines** to **show distance between** launch site **CCAFS SLC- 40** and its proximity to the **nearest coastline, railway, highway, and city**

GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site Slider of Payload Mass Range
- Allow user to select payload mass range

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

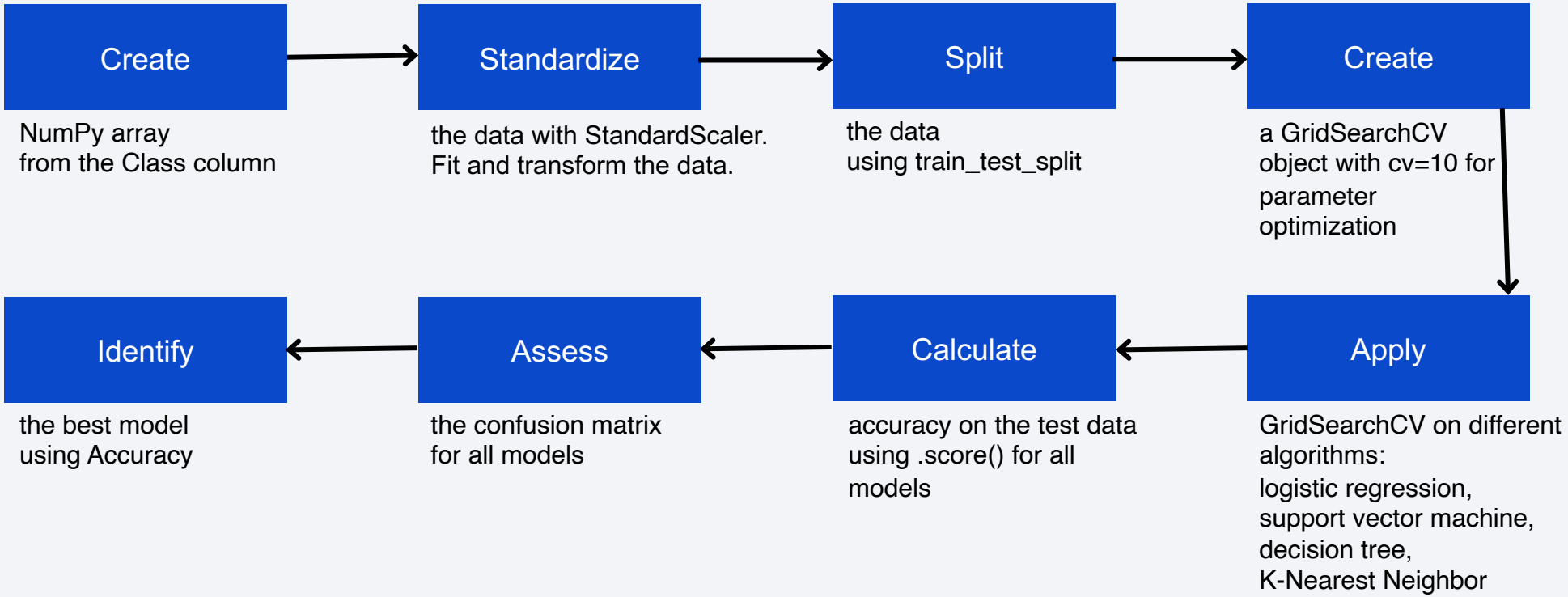
- Allow user to see the correlation between Payload and Launch Success

GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week3/spacex_dash_app.py

Predictive Analysis (Classification)

Flowcharts&explain



GitHub URL

https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate/blob/main/Week4/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

Exploratory Data Analysis

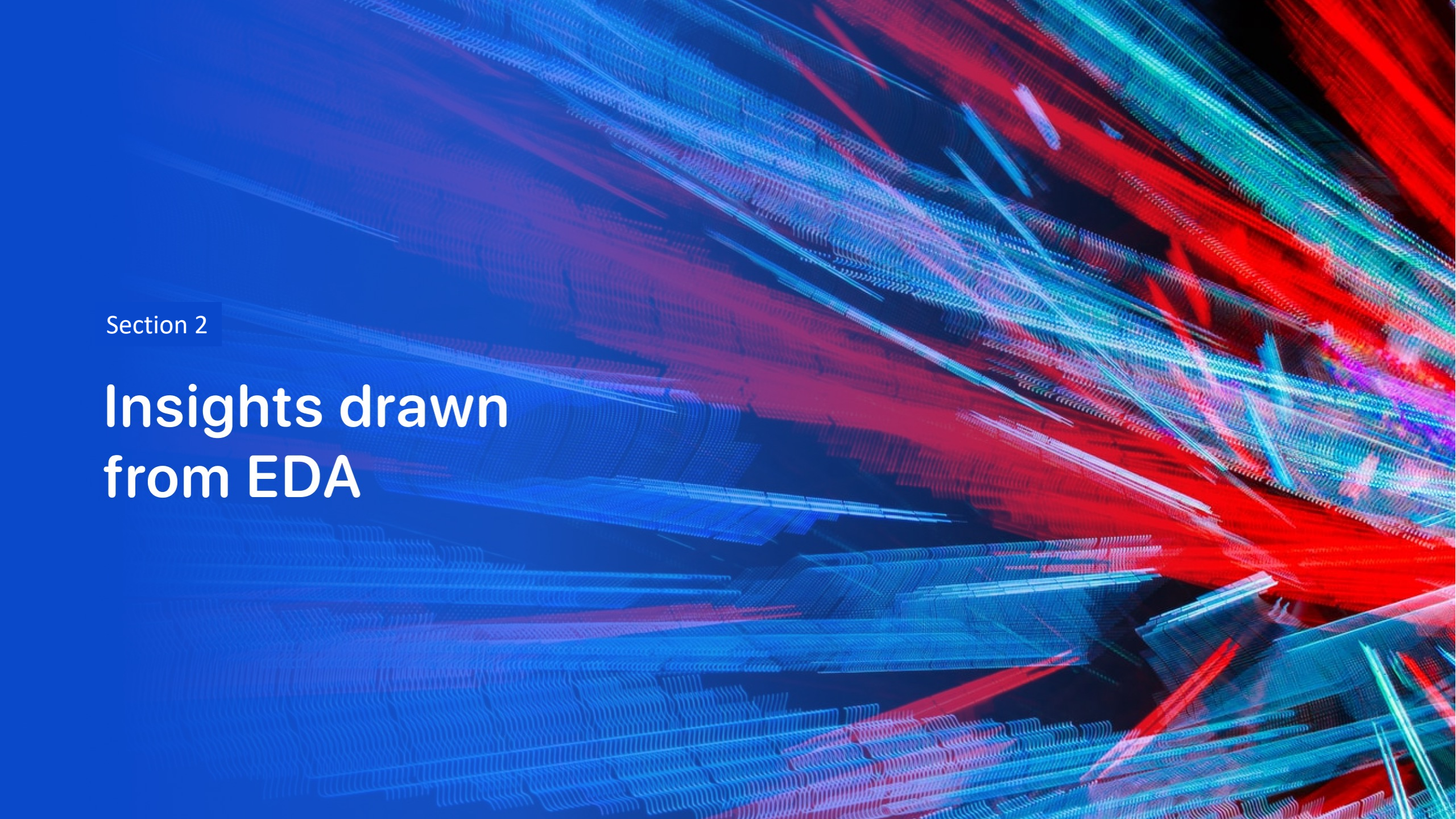
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

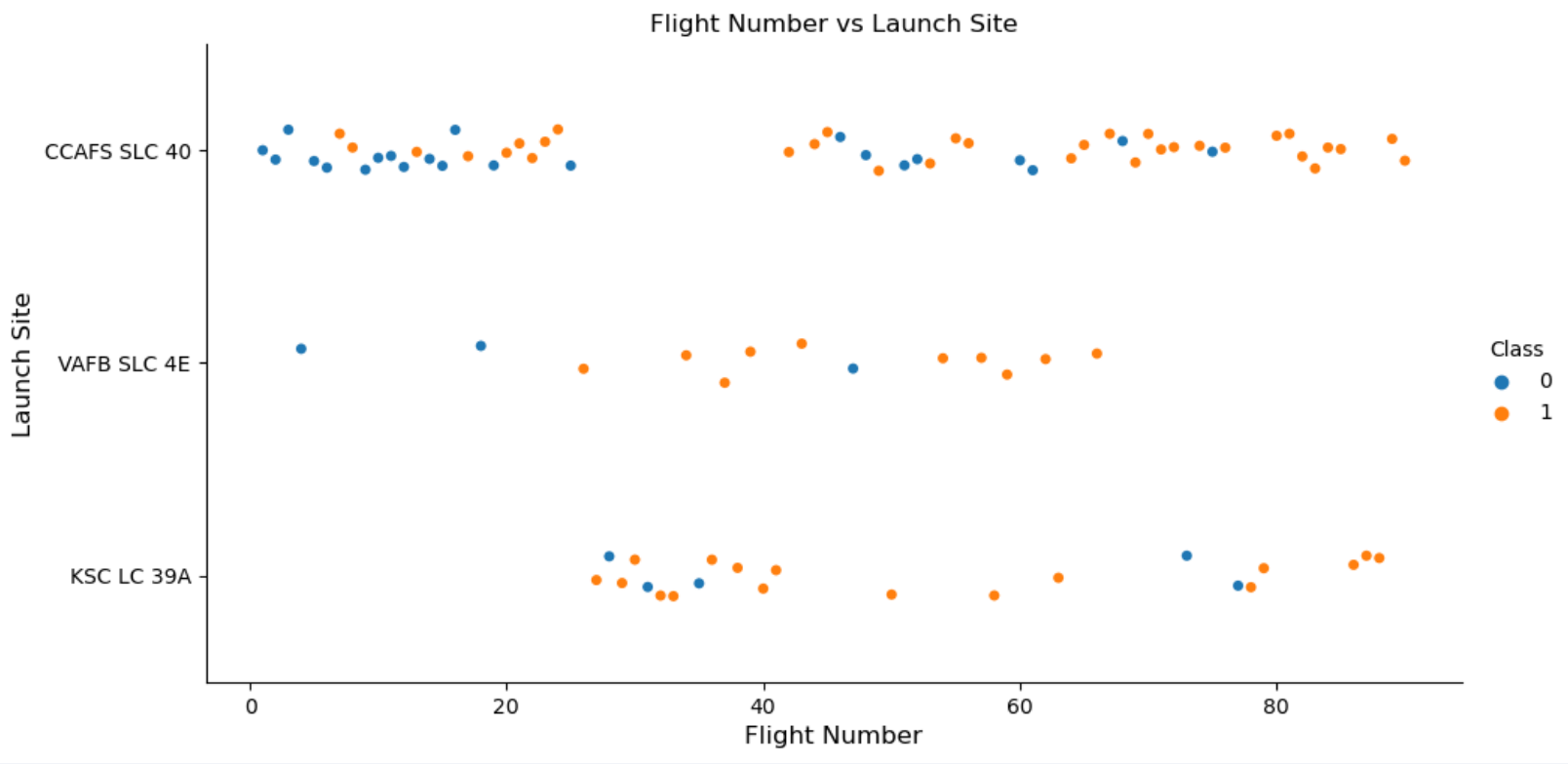
Predictive Analytics

- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

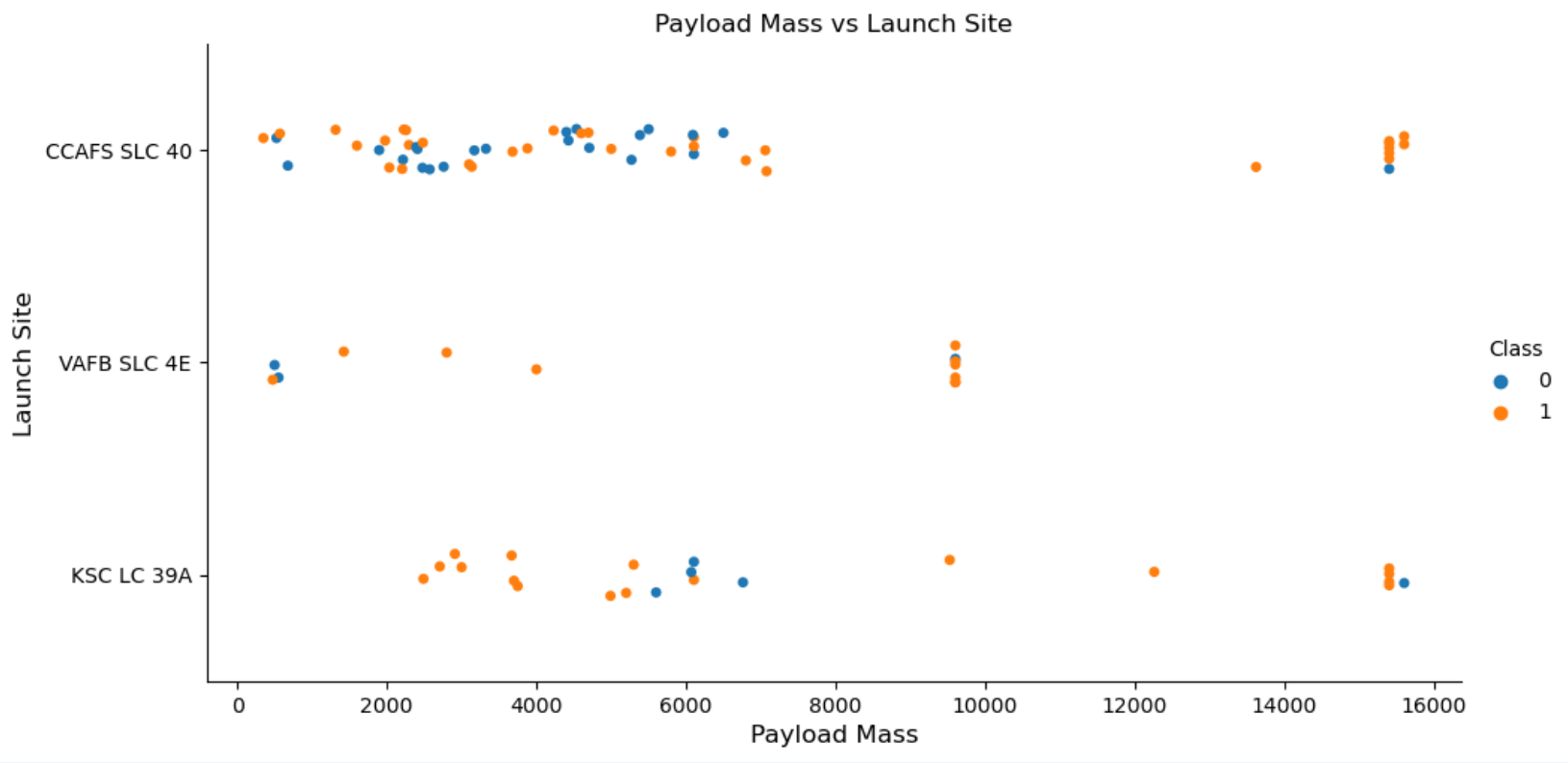
Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



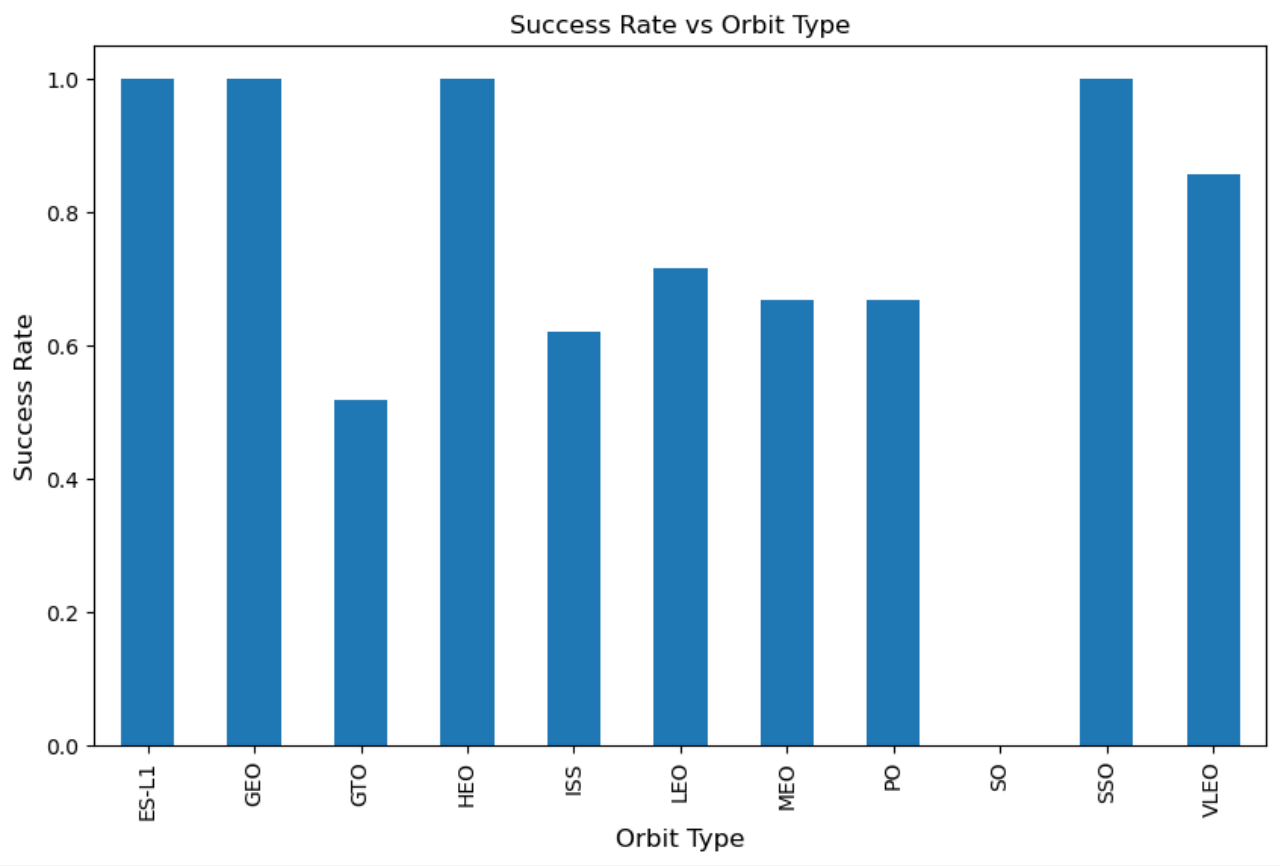
Payload vs. Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



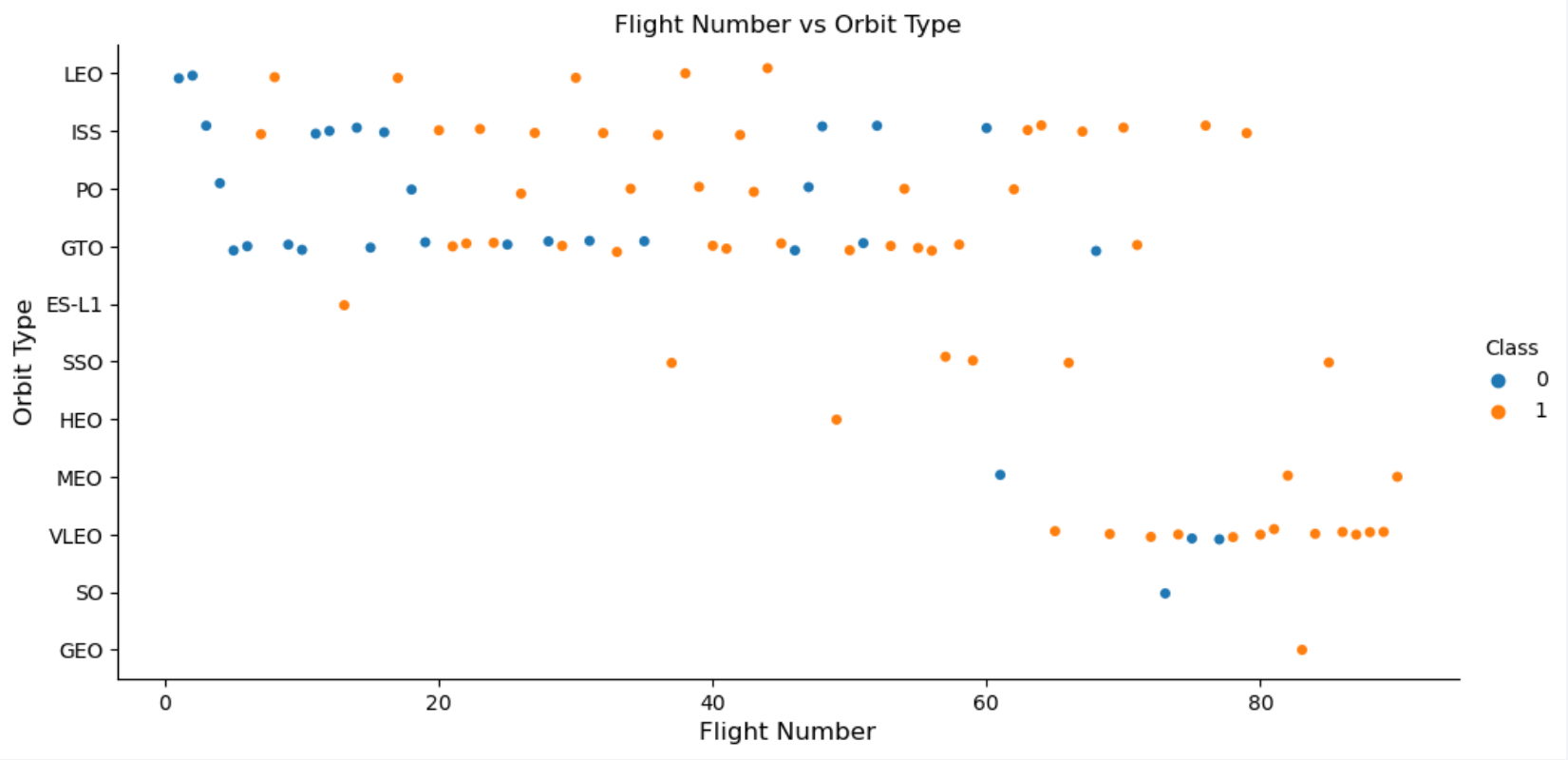
Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



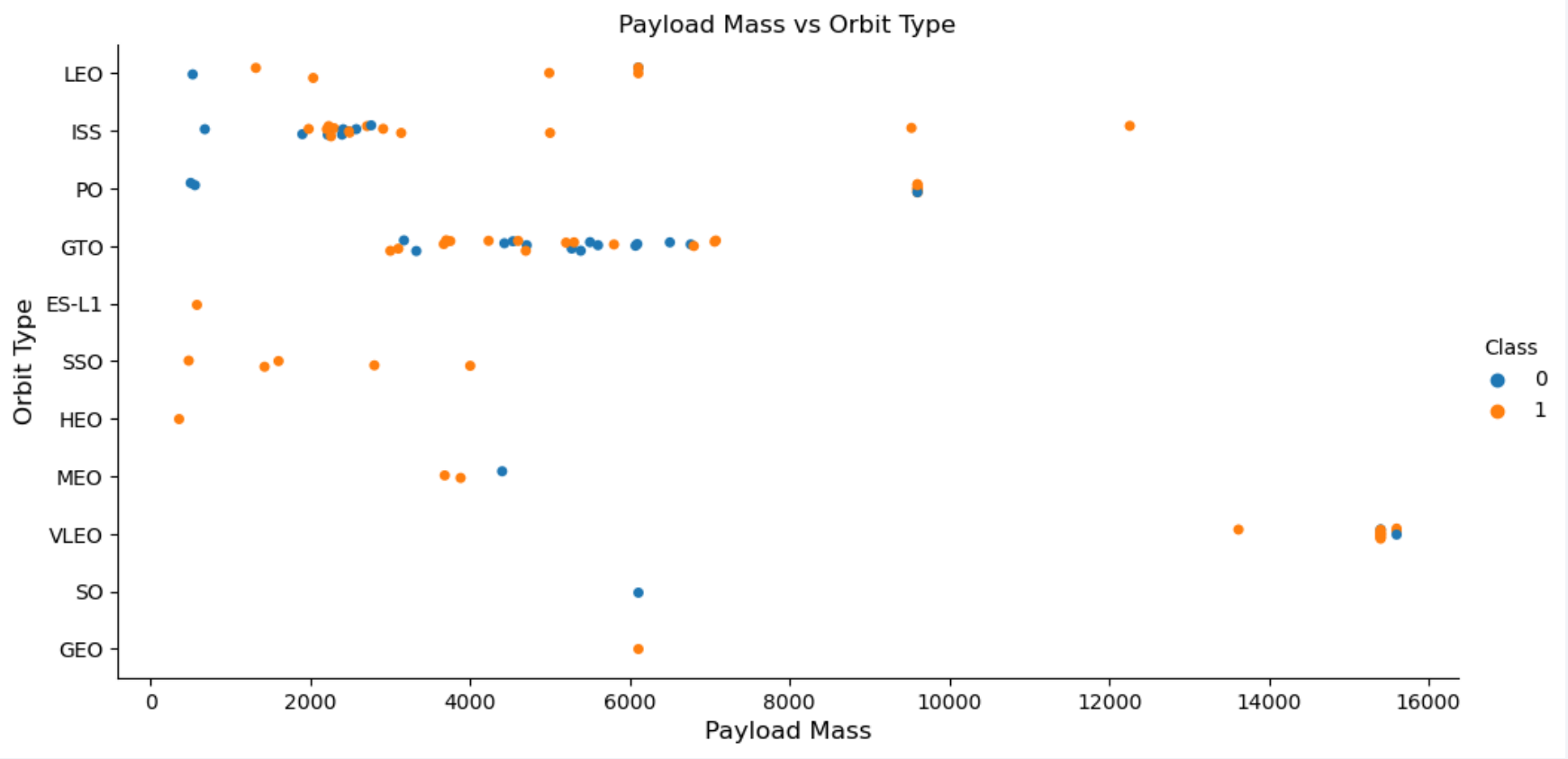
Flight Number vs. Orbit Type

- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



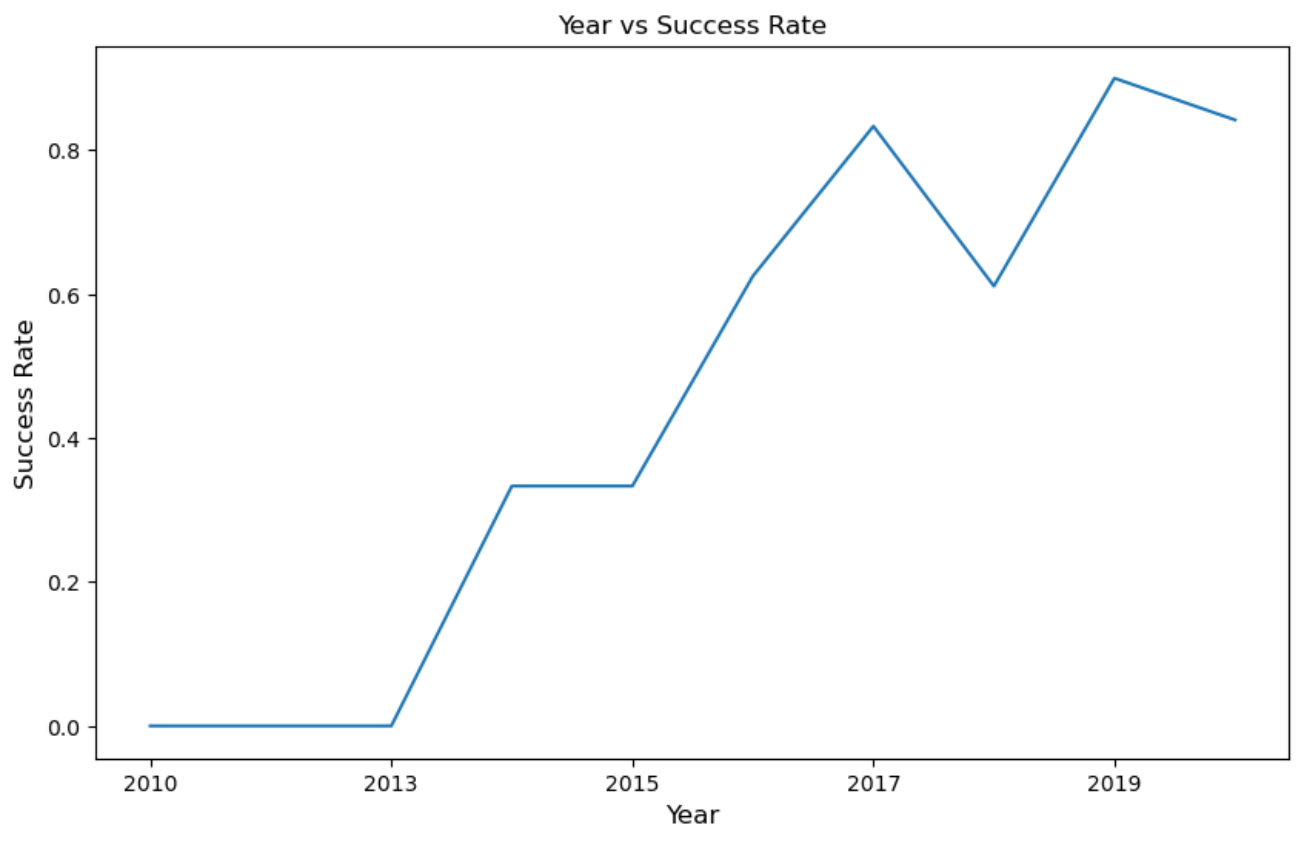
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

Launch Site Names

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

```
%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

Launch Site Names Begin with 'CCA'

Records with Launch Site Starting with CCA

- Displaying 5 records below

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

[14]:

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

Total Payload Mass

- 107,010 kg (total) carried by boosters launched by NASA

```
[15]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like "%NASA%";
 * sqlite:///my_data1.db
Done.
[15]: sum(PAYLOAD_MASS__KG_)
107010
```

Average Payload Mass by F9 v1.1

Average Payload Mass

- 2,534 kg (average) carried by booster version F9 v1.1

```
[16]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "%F9 v1.1%";
 * sqlite:///my_data1.db
Done.
[16]: avg(PAYLOAD_MASS__KG_)
      2534.6666666666665
```

First Successful Ground Landing Date

1st Successful Landing in Ground Pad

- 12/22/2015

```
[18]: %sql select min(Date) from SPACEXTBL where Landing_Outcome = "Success (ground pad)";
 * sqlite:///my_data1.db
Done.
[18]: min(Date)
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000:
  JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

```
[21]: %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.
```

[21]:

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

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

Carrying Max Payload

- 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

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL) group by Booster_Version;
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

In 2015

- Showing month, date, booster version, launch site and landing outcome

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

```
 * 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

Ranked Descending

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%%sql
select Landing_Outcome, count(Landing_Outcome)
from SPACEXTBL
where "Date" between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(Landing_Outcome) desc;
```

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

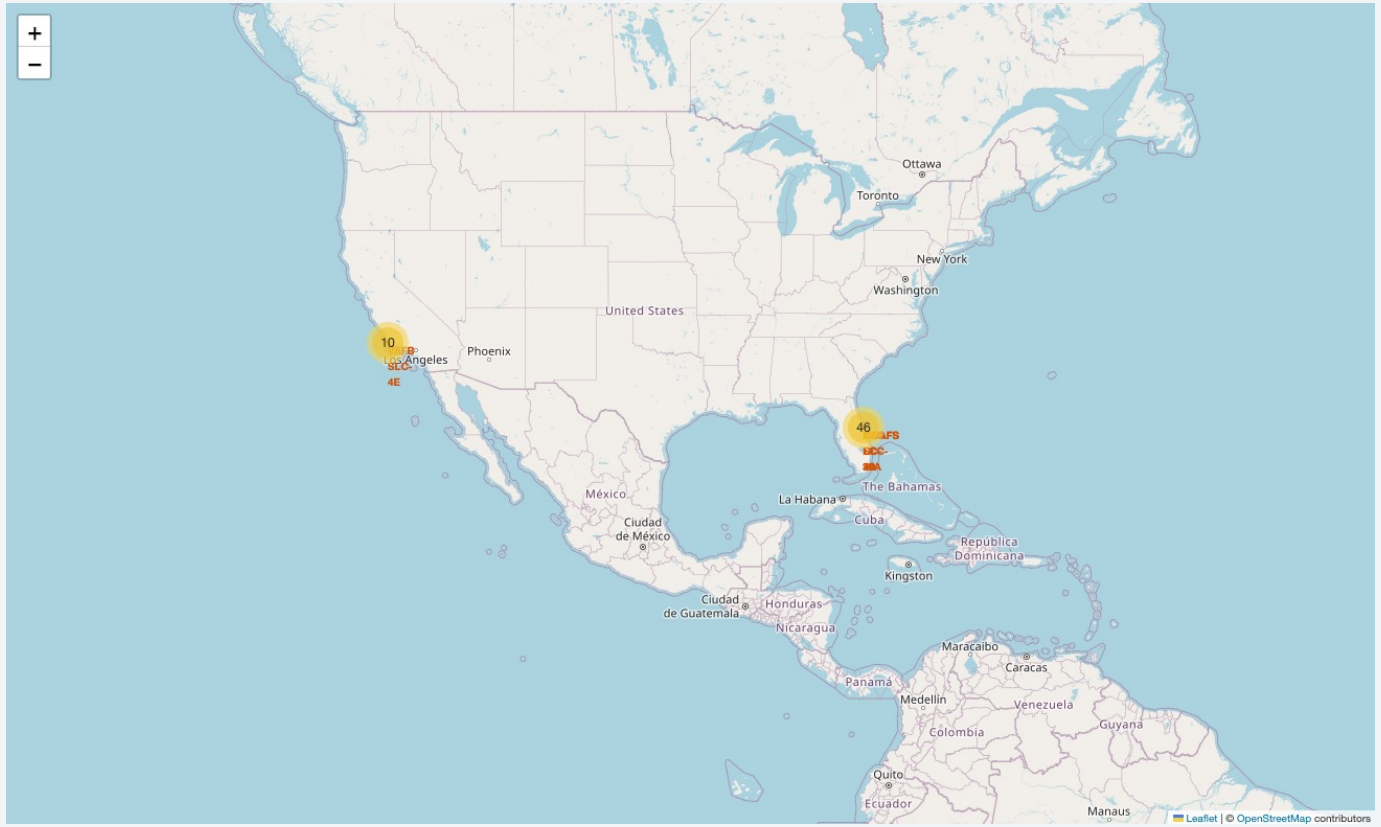
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites

With Markers

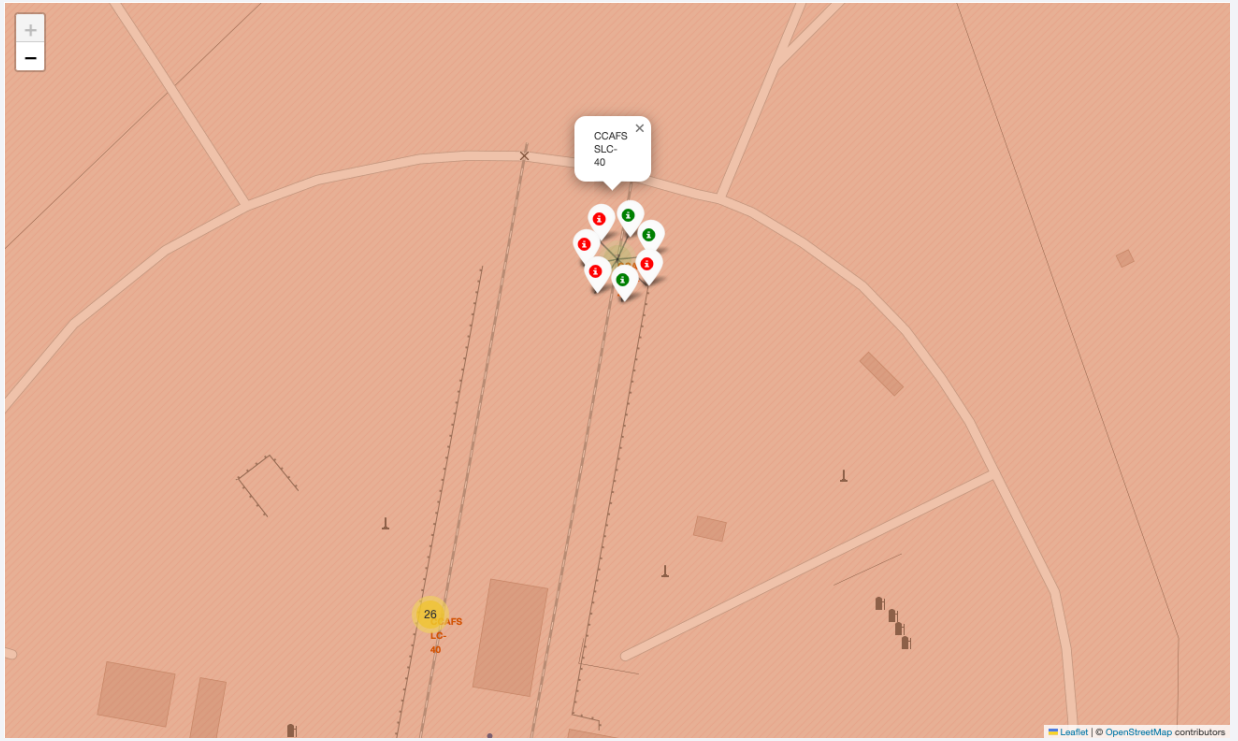
- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost due to the rotational speed of earth that helps save the cost of putting in extra fuel and boosters.



Launch Outcomes

At Each Launch Site

- Outcomes:
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

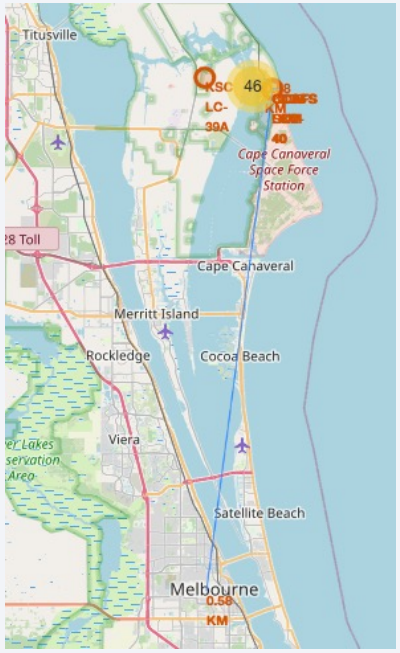


Distance from CCAFS SLC-40 (one of the Launch Sites)

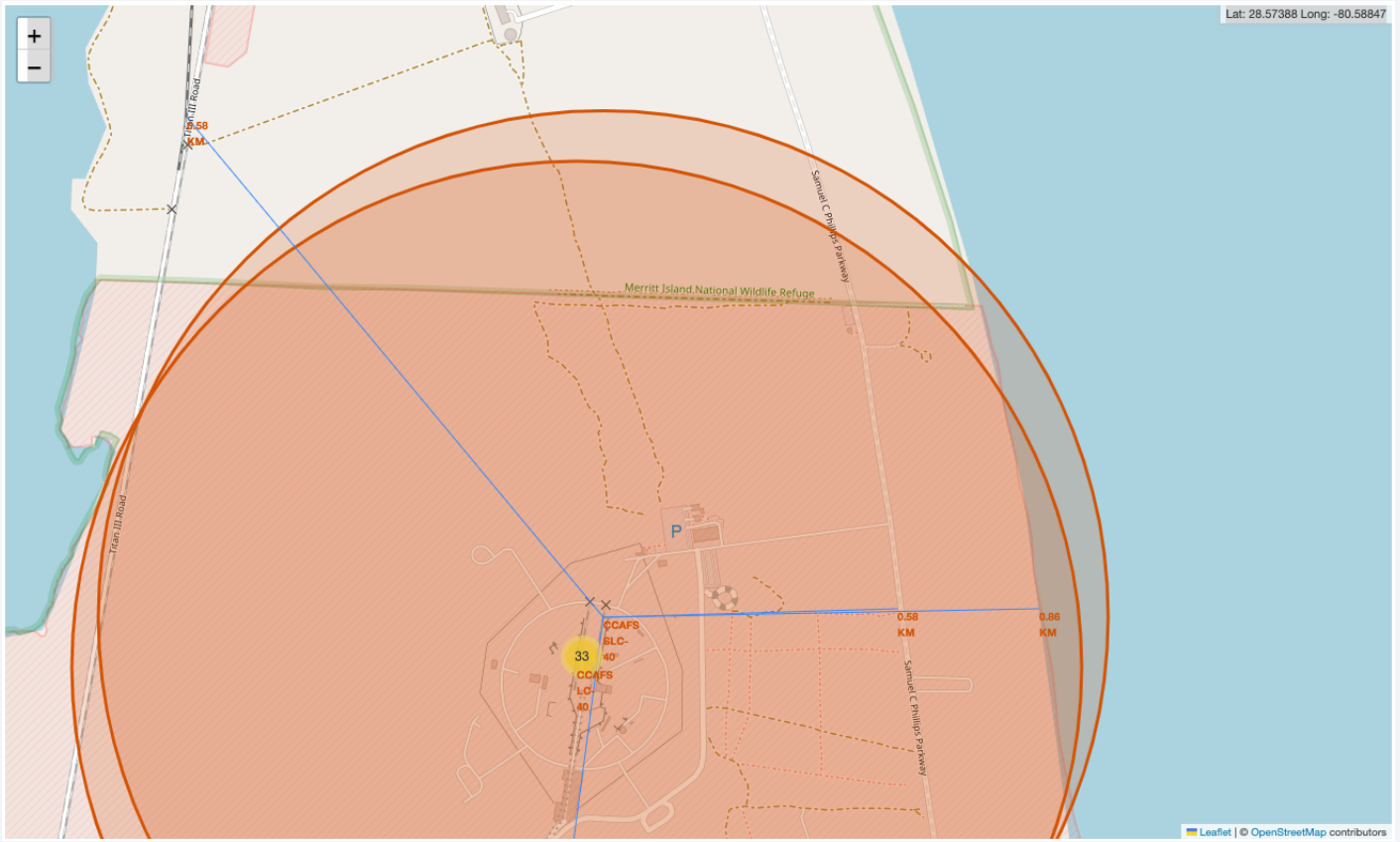
CCAFS SLC-40

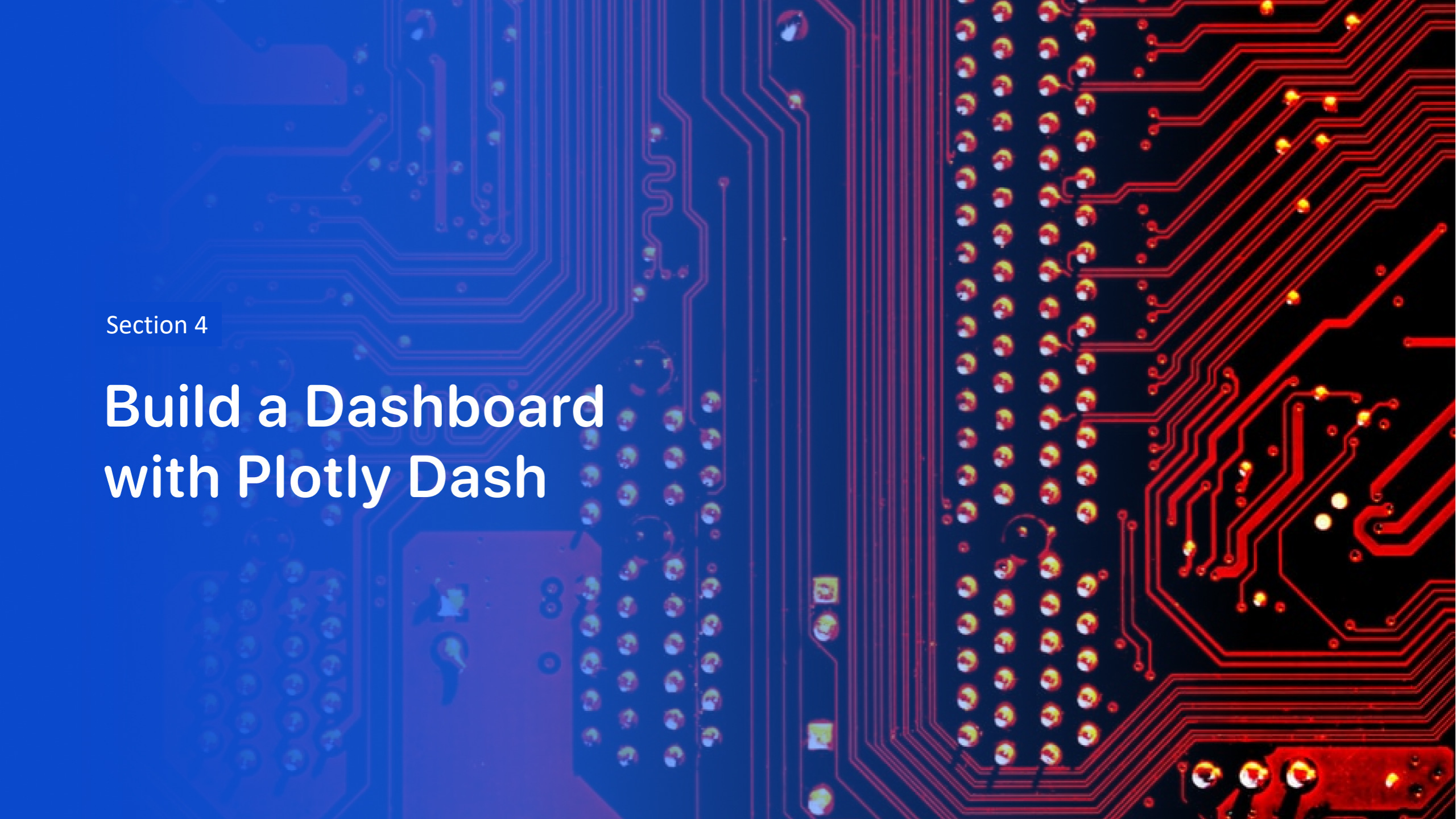
- 0.86 km from nearest coastline
- 0.58km from nearest railway
- 0.58 km from nearest city
- 0.58 km from nearest highway

city



coastline, railway, highway

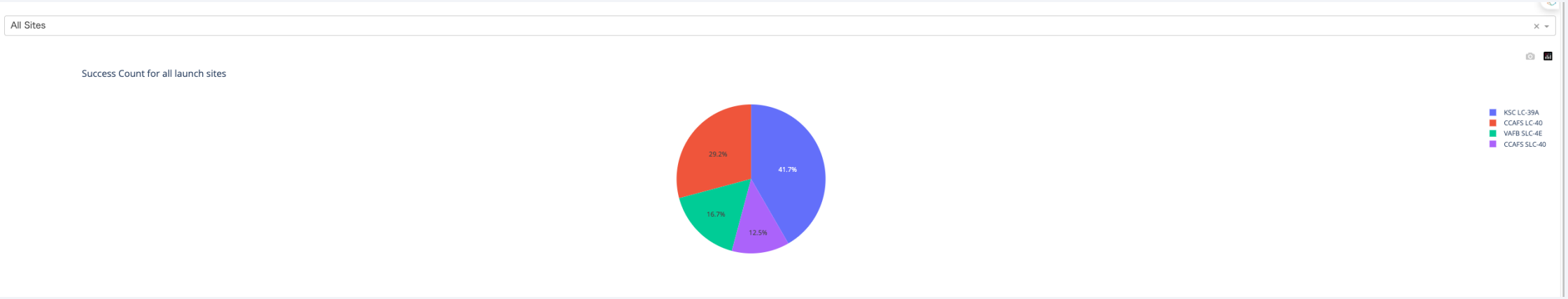



Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site

Success as Percent of Total

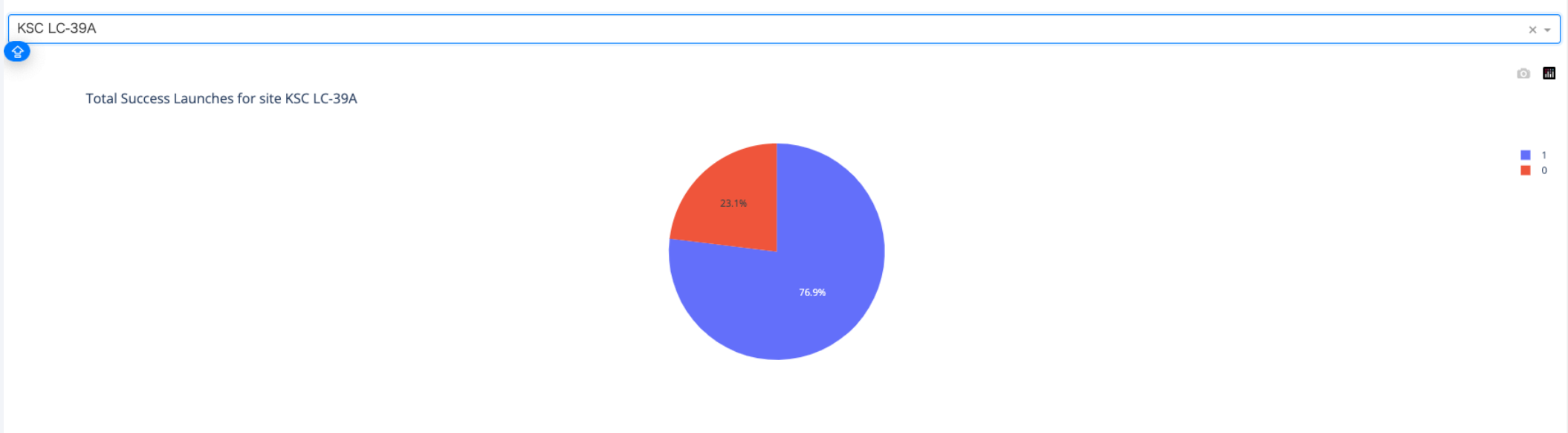
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-39A)

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)



Payload Mass and Success

By Booster Version

- Payloads between 4,000 kg and 6,000 kg have the highest success rate
- Payloads of more than 6000 kg almost failed

Section 5

Predictive Analysis (Classification)

Classification Accuracy

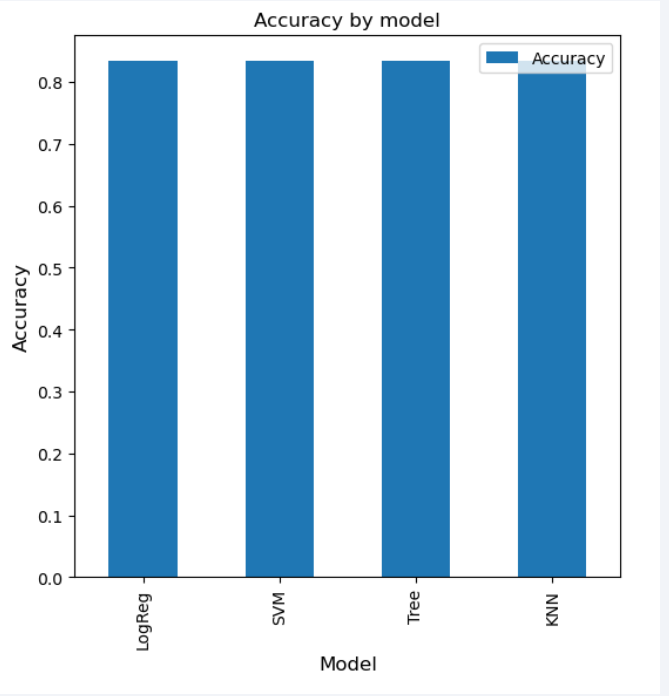
Accuracy

- All the models performed at about the same level and had the same accuracy.

Accuracy table

	LogReg	SVM	Tree	KNN
Accuracy	0.833333	0.833333	0.833333	0.833333

Accuracy bar chart

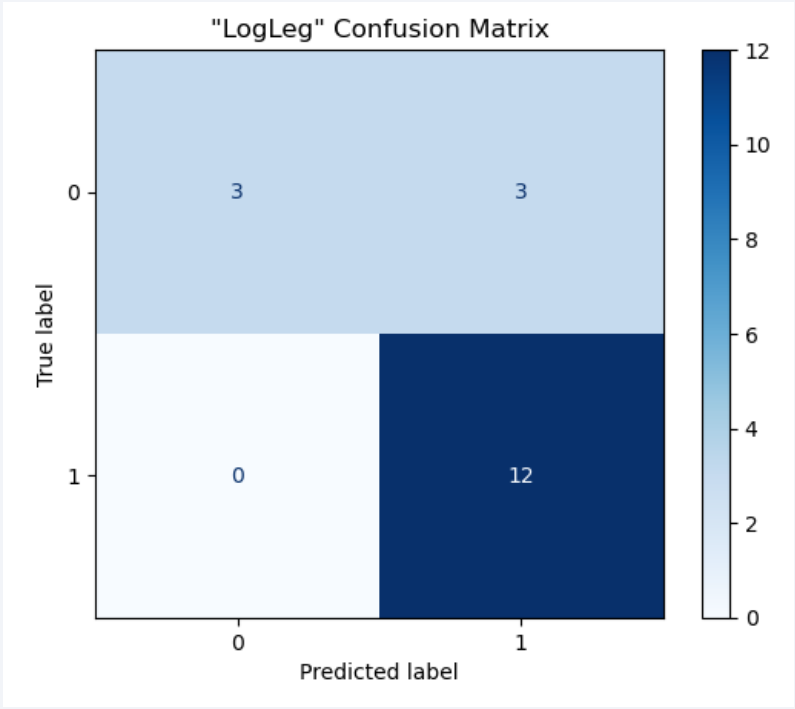


Confusion Matrix

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative

Confusion Matrix (ex. LogReg)



Conclusions

Research

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost due to the rotational speed of earth which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

My GitHub repository for this course:
https://github.com/Rion-Sato/My_Final_Project_of_IBM_Data_Science_Professional_Certificate

Thank you!