Winning Space Race with Data Science

UMAR SHEHZAD
October 10Th, 2021

Outline

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

Executive Summary

- Summary of methodologies

Data collection

Data wrangling

Exploratory Data Analysis

Interactive Visual Analytics and Dashboard

Predictive Analysis (Classification)

- Summary of all results

Exploratory Data Analysis results

Interactive Analytics demo in screenshots

Predictive Analysis results

Introduction

- Project background and context

Unlike other commercial spaceflight providers, SpaceX provides cheaper spaceflights by using rockets that can reuse the first stage. As the first stage sometimes fails to land, *the cost of a launch can be determined by whether the first stage will land successfully*. Being able to predict this will be useful for a company competing against SpaceX as a commercial spaceflight provider.

- Problems you want to find answers

Using public information, can we predict whether SpaceX will reuse the first stage?

Can we predict how different variables would affect the launch success rate?

Section 1

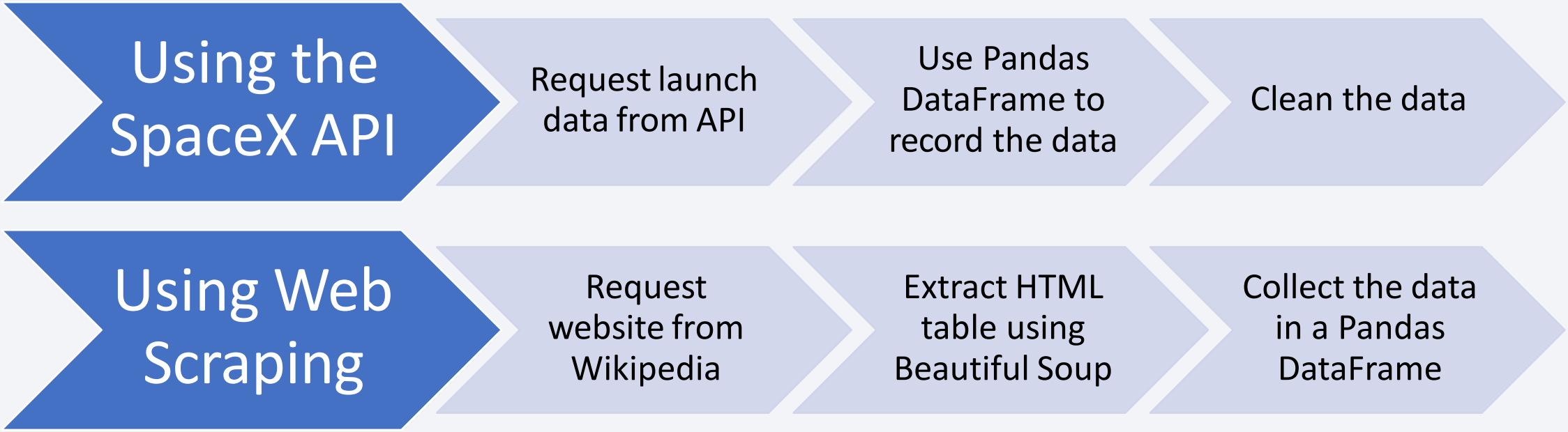
Methodology

Methodology

Executive Summary

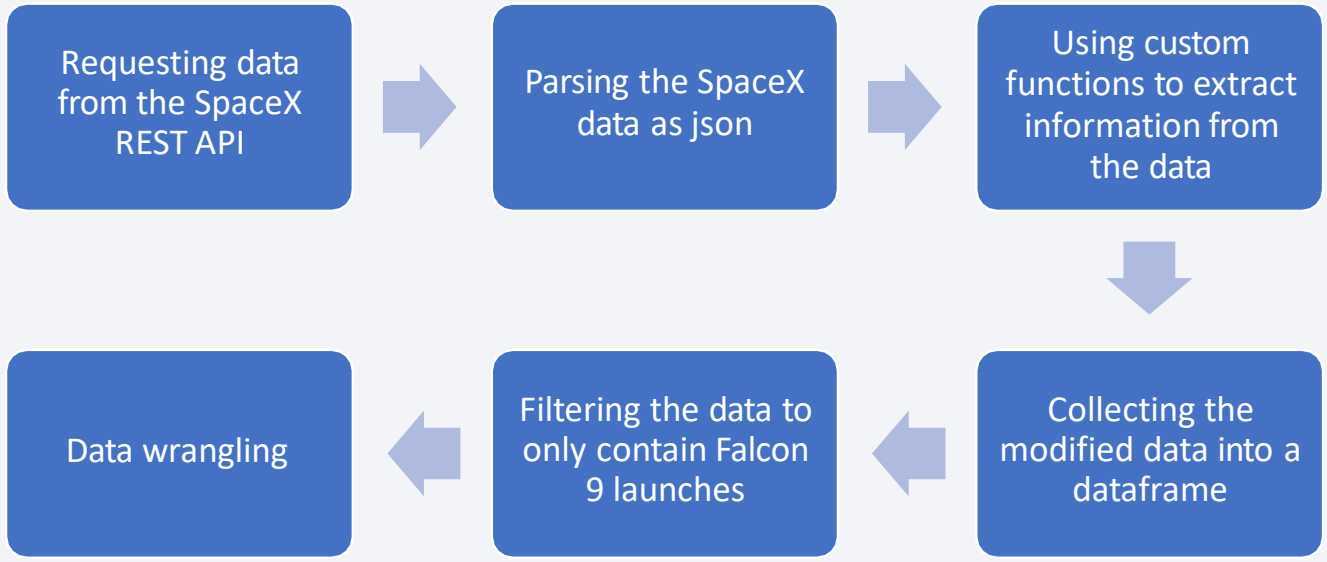
- Data collection methodology:
  - Using the SpaceX REST API to collect public launch data from SpaceX
  - Using web scraping to collect data from Wikipedia
- Perform data wrangling
  - Processing missing values and One-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Training logistic regression, SVM, decision tree, and KNN models to predict the outcome of each launch based on publicly available data

Data Collection



Data Collection - SpaceX API

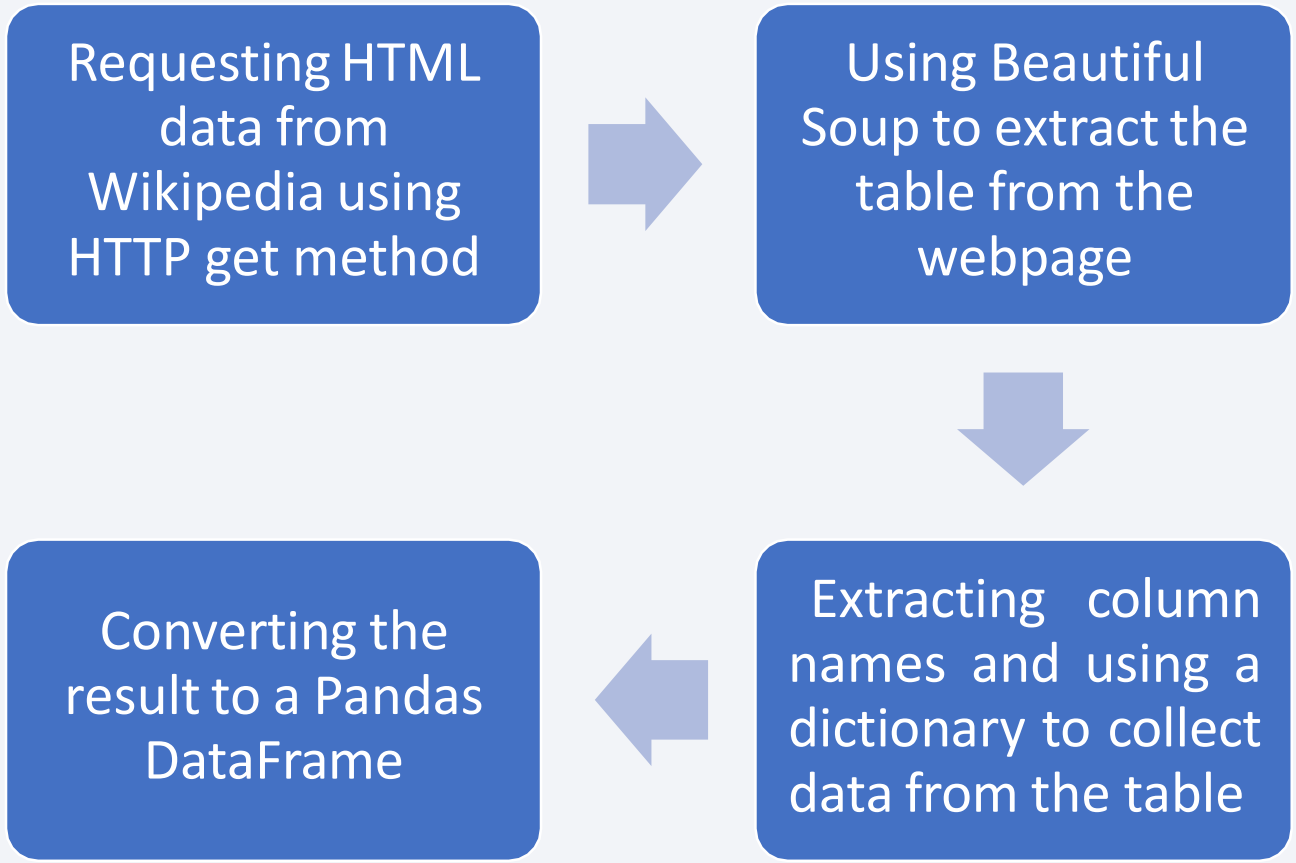
- Data was requested from the SpaceX REST API in the form of json
- Functions were used to extract useful information from the data
- GitHub URL: https://github.com/OmerShehzad/Predict-if-Spacex-Falcon-9-first-stage-will-land-successfully.git



Data Collection - Scraping

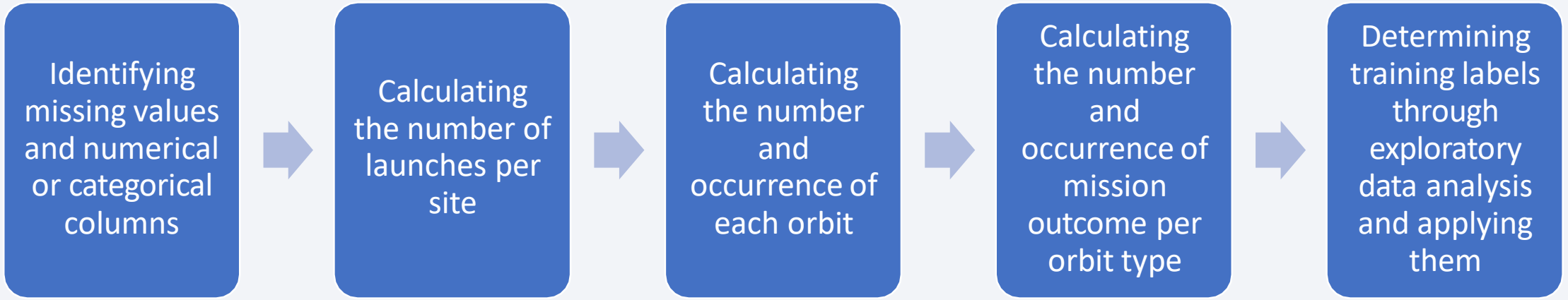
- SpaceX launch data was requested from a Wikipedia page using HTTP
- The table in the webpage was used as our data
- GitHub URL:

https://github.com/OmerShehzad/Predict-if-Spacex-Falcon-9-first-stage-will-land-successfully.git



Data Wrangling

- The data was labeled according to whether it was a success(training labels), and further insight about the data was gained through exploratory data analysis.



EDA with Data Visualization

Graph	Purpose
Scatter plot, flight number vs payload mass	To check the effect flight number and payload mass has on the success of each launch
Scatter plot, flight number vs launch site	To see if there are patterns related to the launch site
Scatter plot, payload mass vs launch site	To check for relationships between launch site and payload
Bar graph, orbit type vs success rate	To check the launch success rate for each type of orbit
Scatter plot, flight number vs orbit type	To check for relationships between flight number and orbit
Scatter plot, payload mass vs orbit type	To see what effect payload mass has on launch success rate for each type of orbit
Line graph, year vs success rate	To observe the yearly launch success trend

The above graphs were plotted to explore the data.

EDA with SQL

- The following SQL queries were performed to explore the data.

1. Display the unique names of each launch site in the data
2. Display 5 records where the launch site name begins with 'CCA'
3. Display the total payload mass carried by the customer 'NASA (CRS)'
4. Display the average payload mass carried by the booster F9 v1.1
5. List the date when the first successful landing on ground pad was achieved
6. List the names of boosters which had success landing on a drone ship with a payload mass between 4000 and 6000 kilograms
7. List the total number of successful and failed mission outcomes
8. List the names of boosters which have carried the maximum payload mass
9. List the failed landings on a drone ship, their booster versions, and launch site names for launches in the year 2015
10. Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Build an Interactive Map with Folium

Map object	Purpose
folium.Circle with text label	To mark the location of NASA's Johnson Space Center
folium.Circle with text label	To mark the locations of each launch site
Marker cluster with color based on success	To mark the launch outcomes for each launch onto each site
MousePosition object	To get the coordinates for any point on the map
folium.Marker for distance	To mark the distance between each site and the nearest railway point
PolyLine object	To mark the distance between each site and either the nearest railway point, city, coastline, or highway

The above map objects were added to a folium map to explore the data.

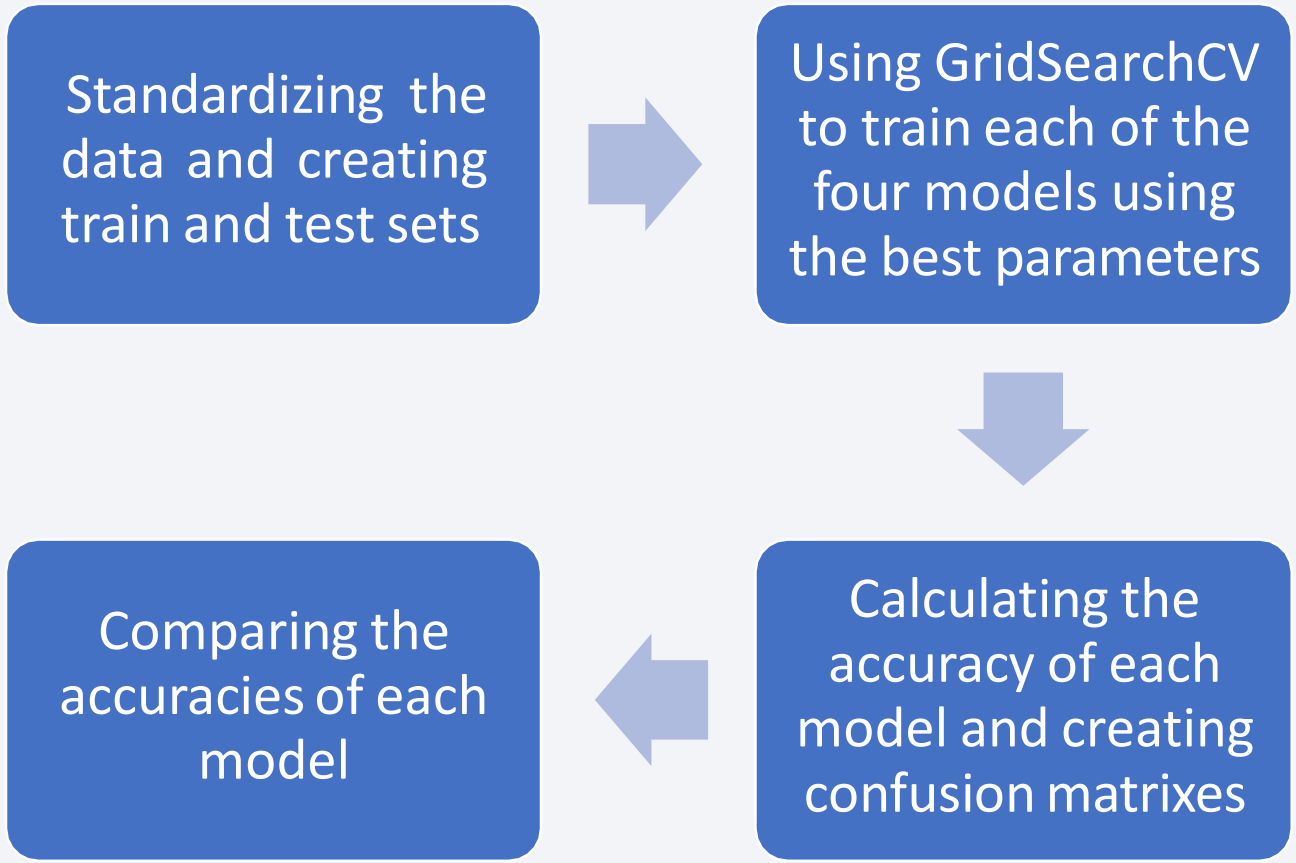
Build a Dashboard with Plotly Dash

Dashboard element	Purpose
Dropdown component	To be able to select a launch site we want to view the data for, or all launch sites
Pie chart with callback function	To display either the amount of successful launches per site when all sites are selected or the amount of successful and failed launches at the site when a single site is selected
Range slider	To be able to select a range of payload masses we want to view the data for
Scatter plot with callback function	To display the mission outcome of each launch based on payload, and to also see the booster versions of each launch

The above elements were added to a dashboard to explore the data.

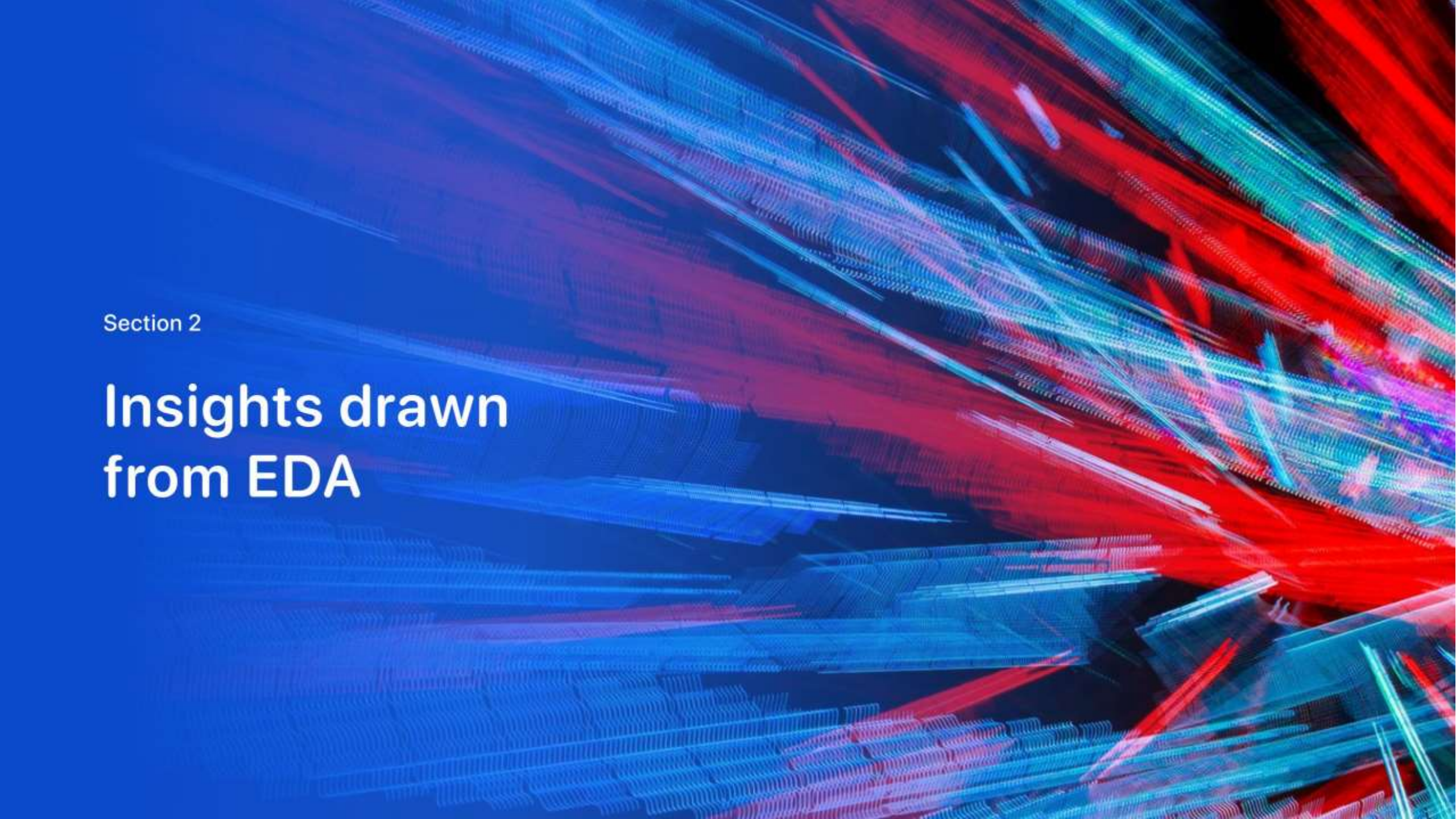
Predictive Analysis (Classification)

- Logistic regression, SVM, Decision tree, and KNN models were trained on the same data and evaluated for the best results.

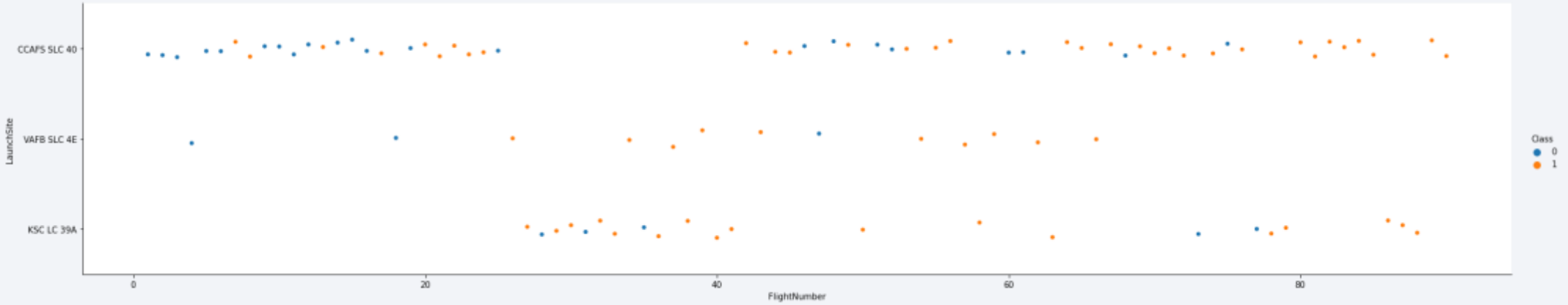


Section 2

Insights drawn from EDA

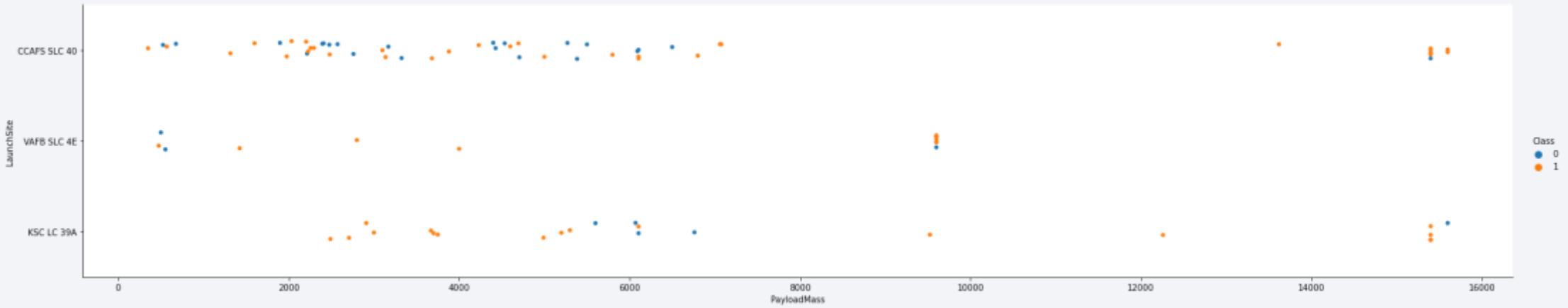


Flight Number vs. Launch Site



- Different sites were mainly used for SpaceX launches over time.
- As time goes on, the success rate of launches increases on all launch sites.

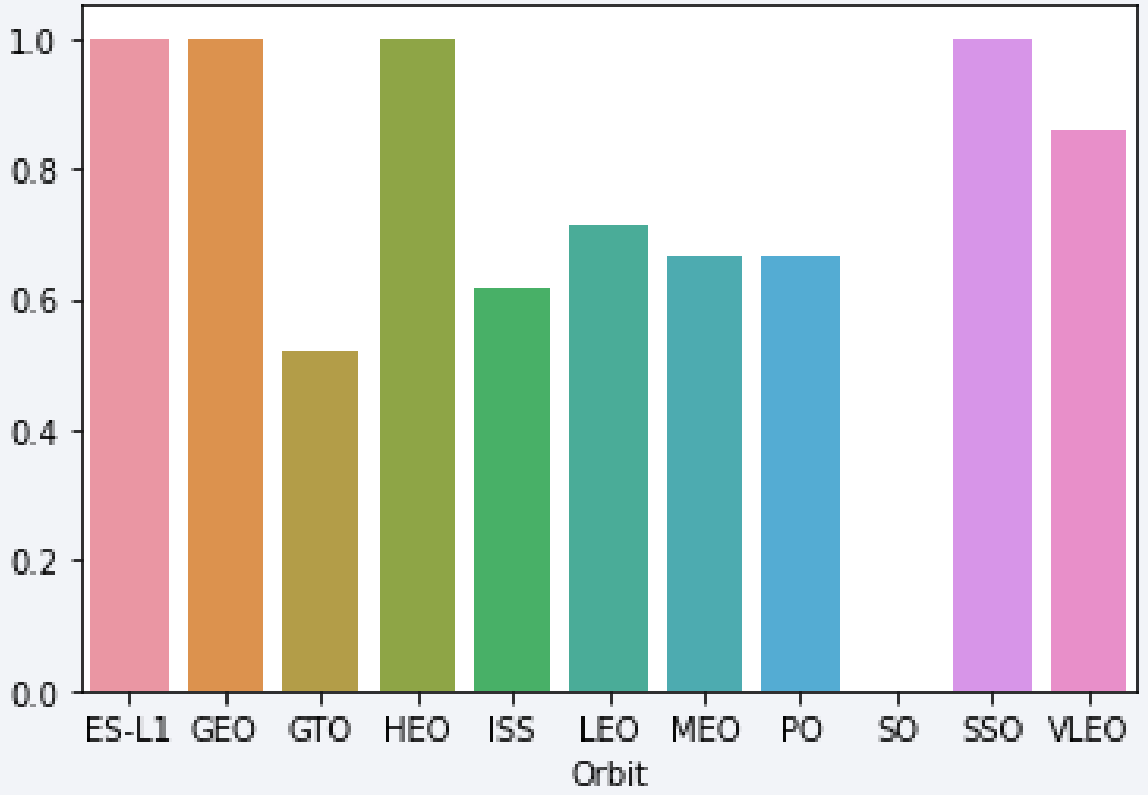
Payload vs. Launch Site



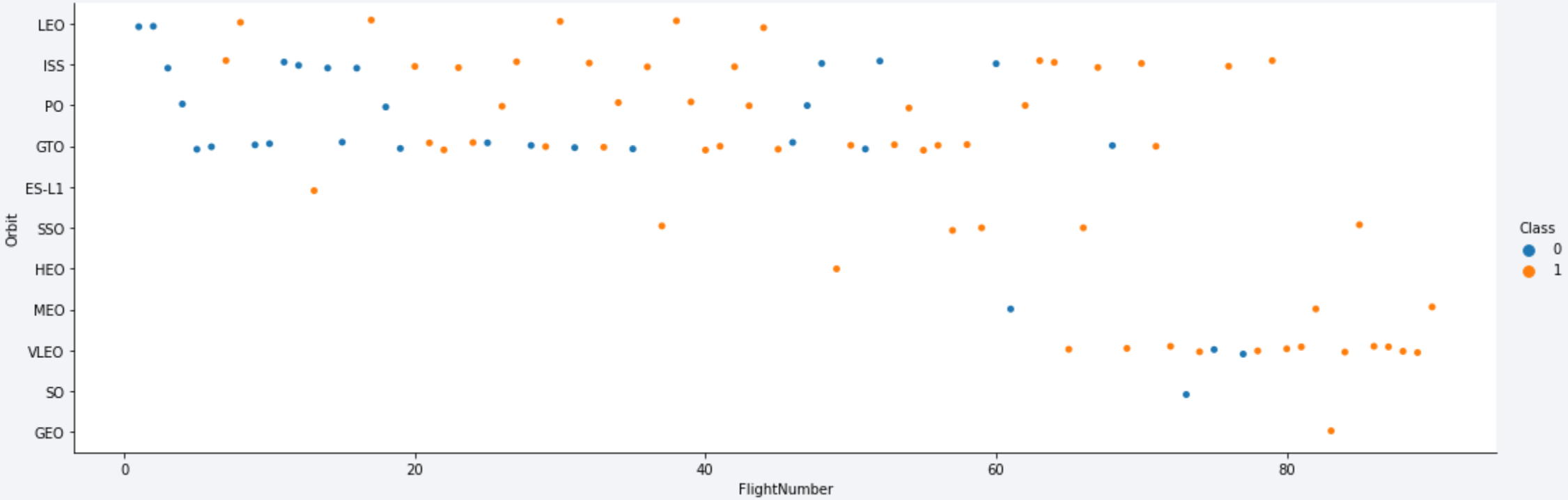
- The launch site VAFB SLC 4E was used for launches with lighter payloads.
- For launch site CCAFS SLC 40, larger payloads have a larger success rate.
- The majority of launches with payloads *heavier than 7000kg* were successful.

Success Rate vs. Orbit Type

- All launches to ES-L1, GEO, HEO, and SSO orbits were successful, and no launches to the SO orbit was successful.
- Out of the rest, launches to VLEO had the highest success rate, and launches to GTO had the lowest.

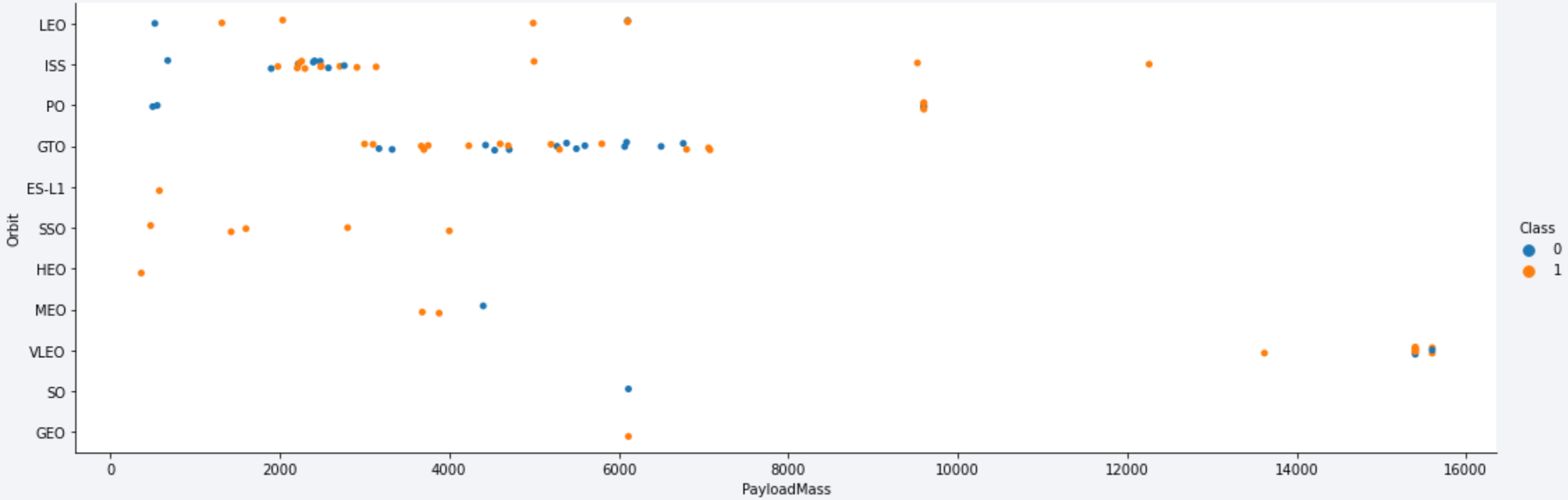


Flight Number vs. Orbit Type



- The success rate increases as time goes on. This is most visible in launches to LEO, while less visible in launches to GTO.

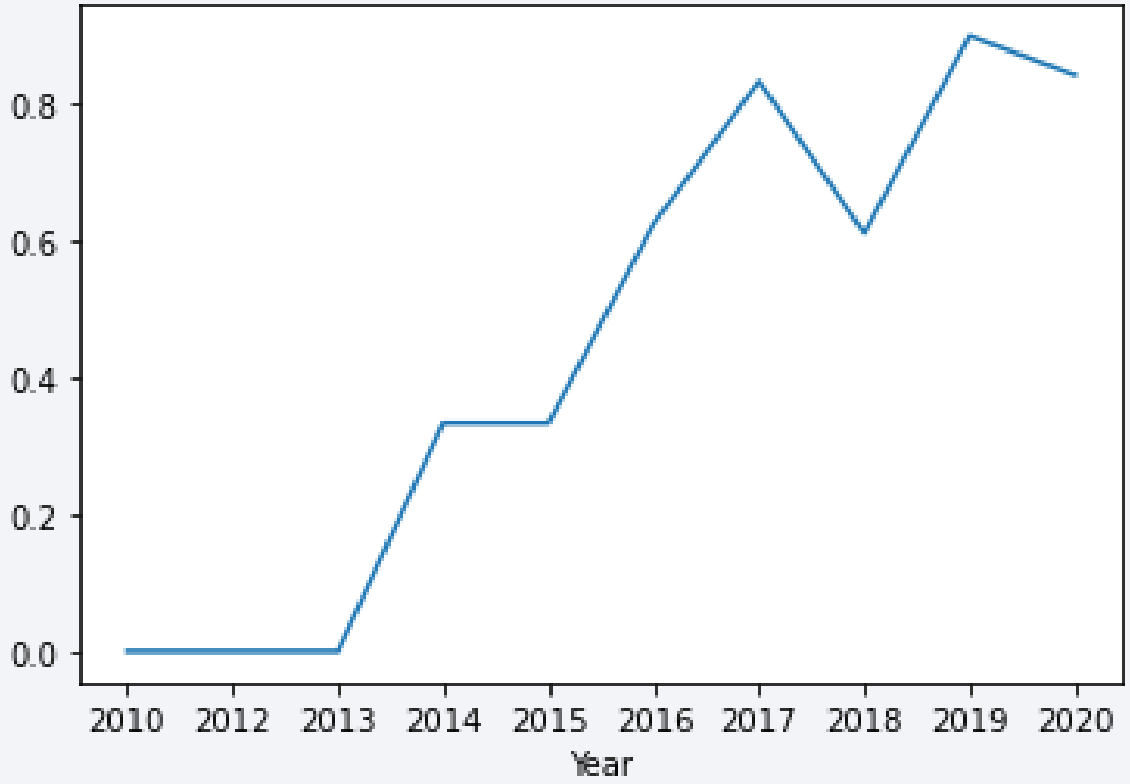
Payload vs. Orbit Type



- A heavier payload has a negative influence on launches to GTO and MEO, while it appears to have a positive influence on launches to LEO, ISS, and PO orbits.

Launch Success Yearly Trend

- The average success rate was zero from 2010 to 2013 but began to increase to up to over 80% in 2017.
- From 2010 to 2020, the change in the success rate was **generally an upwards trend**.



All Launch Site Names

- “select distinct” was used to select unique values from the launch site column.

```
%sql select distinct launch_site from spacexdataset
```

	LAUNCH_SITE
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- “limit” was used to limit the number of results to 5.

```
%sql select * from spacexdataset where launch_site like 'CCA%' limit 5
```

Python

	DATE	TIME_UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS_KG_	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING_OUTCOME
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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

- “sum()” was used to calculate the sum of the values in the payload mass column from the selected rows.

```
%sql select sum(payload_mass__kg_) from spacexdataset where customer like 'NASA (CRS)'
```

	1
0	45596

Average Payload Mass by F9 v1.1

- “avg()” was used to calculate the average of values in the payload mass column in the selected rows of data.

```
%sql select avg(payload_mass__kg_) from spacexdataset where booster_version like 'F9 v1.1%'
```

	1
0	2534

First Successful Ground Landing Date

- “min()” was used to get the minimum number, or in this case the first date, in the date column from the selected rows.

```
%sql select min(date) from spacexdataset where landing__outcome like 'Success (ground pad)'
```

	1
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- “and” was used to query using multiple conditions.

```
%sql select distinct booster_version from spacexdataset where landing__outcome like 'Success (drone ship)' and payload_mass__kg_<6000 and payload_mass__kg_>4000
```

	BOOSTER_VERSION
0	F9 FT B1021.2
1	F9 FT B1031.2
2	F9 FT B1022
3	F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- “count(*)” was used to count the number of rows that were grouped into each value of the mission outcome column.

```
%sql select mission_outcome, count(*) from spacexdataset group by mission_outcome
```

	MISSION_OUTCOME	2
0	Failure (in flight)	1
1	Success	99
2	Success (payload status unclear)	1

Boosters Carried Maximum Payload

- A subquery was used to calculate the maximum payload, the result of which was then used to select the booster versions associated with the value.

```
%sql select booster_version from spacexdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset)
```

	BOOSTER_VERSION
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
...	...
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

12 rows × 1 columns

2015 Launch Records

- “date like ‘2015%’” was used to filter the data for dates in 2015.

```
%sql select landing__outcome, booster_version, launch_site from spacexdataset where landing__outcome like 'Failure (drone ship)' and date like '2015%'
```

	LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Dates can be compared using <, >, <=, and >=.

```
%sql select landing__outcome, count(*) from spacexdataset where date>='2010-06-04' and date<='2017-03-20' group by landing__outcome order by count(*) desc
```

	LANDING_OUTCOME	2
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Failure (parachute)	2
6	Uncontrolled (ocean)	2
7	Precluded (drone ship)	1

Section 4

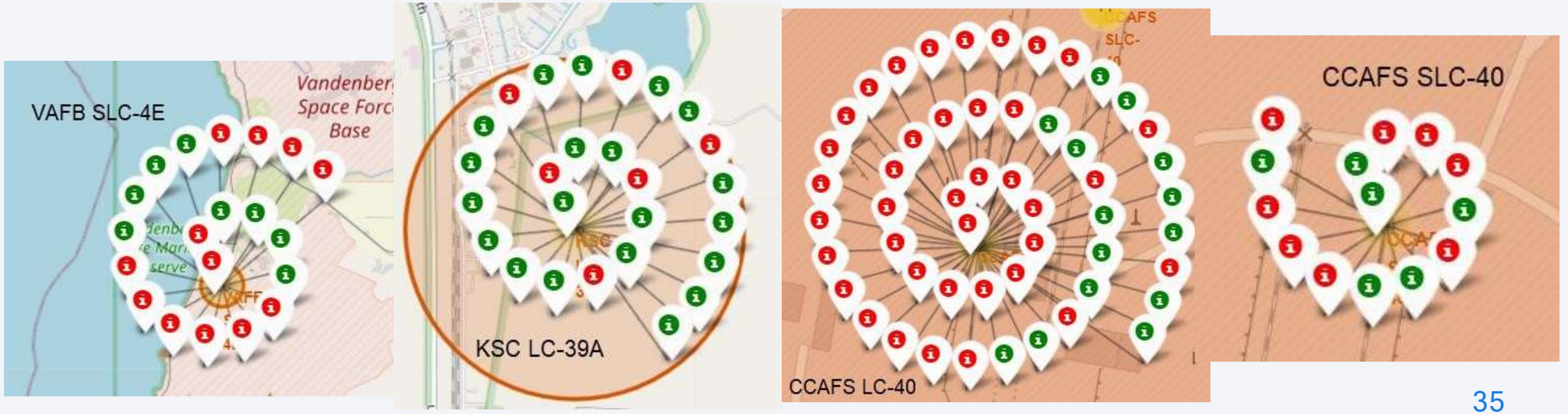
Launch Sites Proximities Analysis

Locations of Launch Sites

- The launch sites were located on the **west or east coast** of the United States, and towards the **southern parts**.

Color-labeled Launch Markers

- Green and red markers each show successful and failed launches respectively.
- Launches from CCAFS LC-40 had the lowest success rate, while launches from KSC LC-39A had the highest.

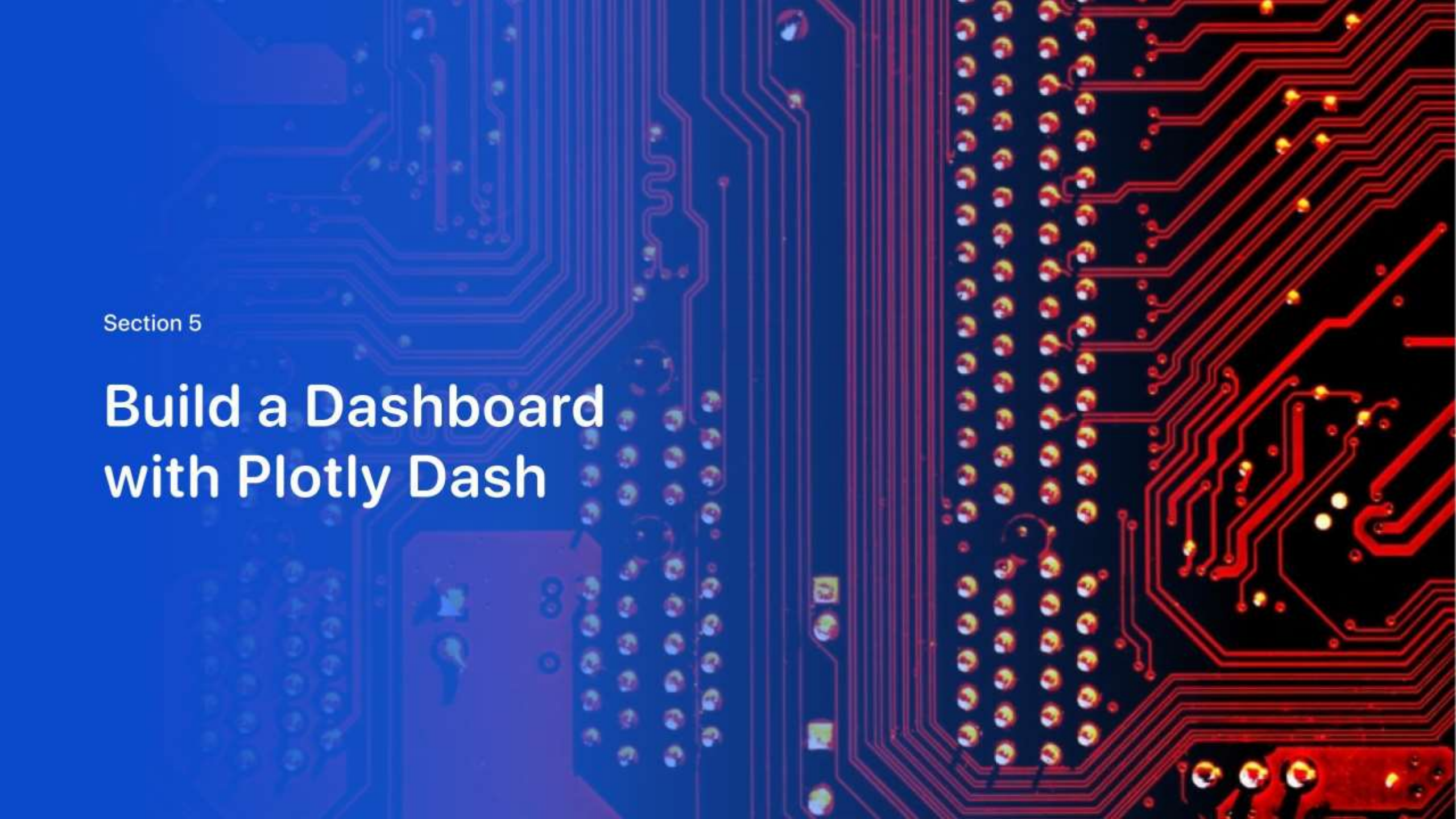


Distances from each Launch Site

- The distance from CCAFS SLC-40 to the nearest railway is about 1.29km.

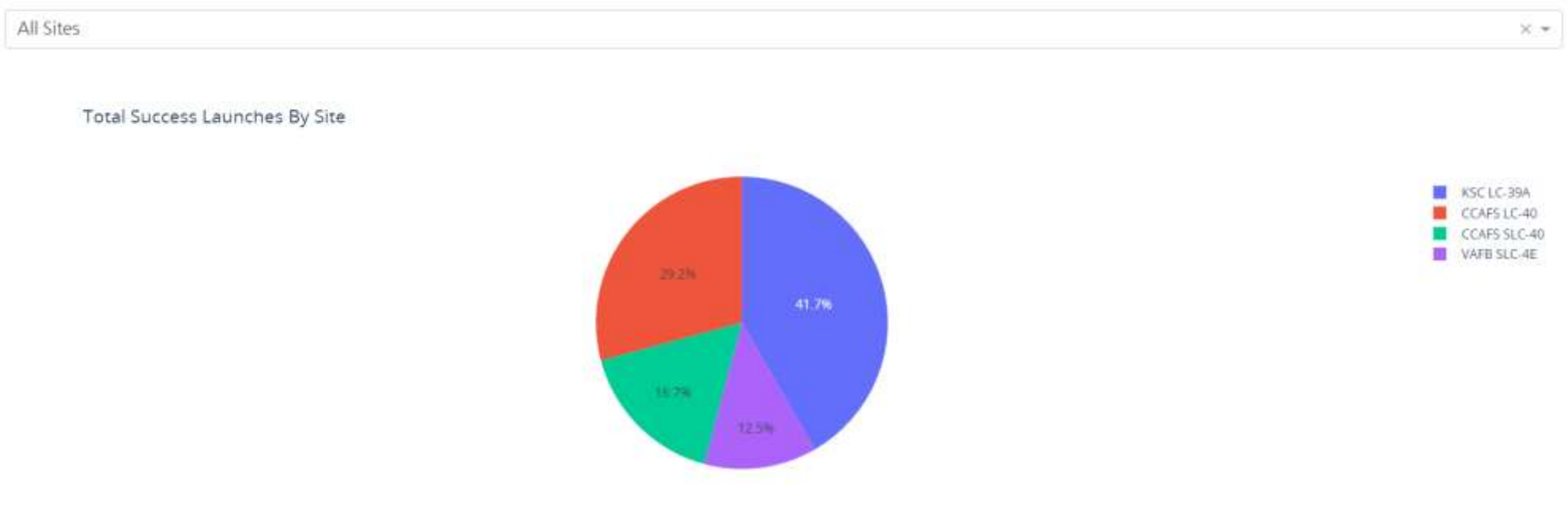
Section 5

Build a Dashboard with Plotly Dash



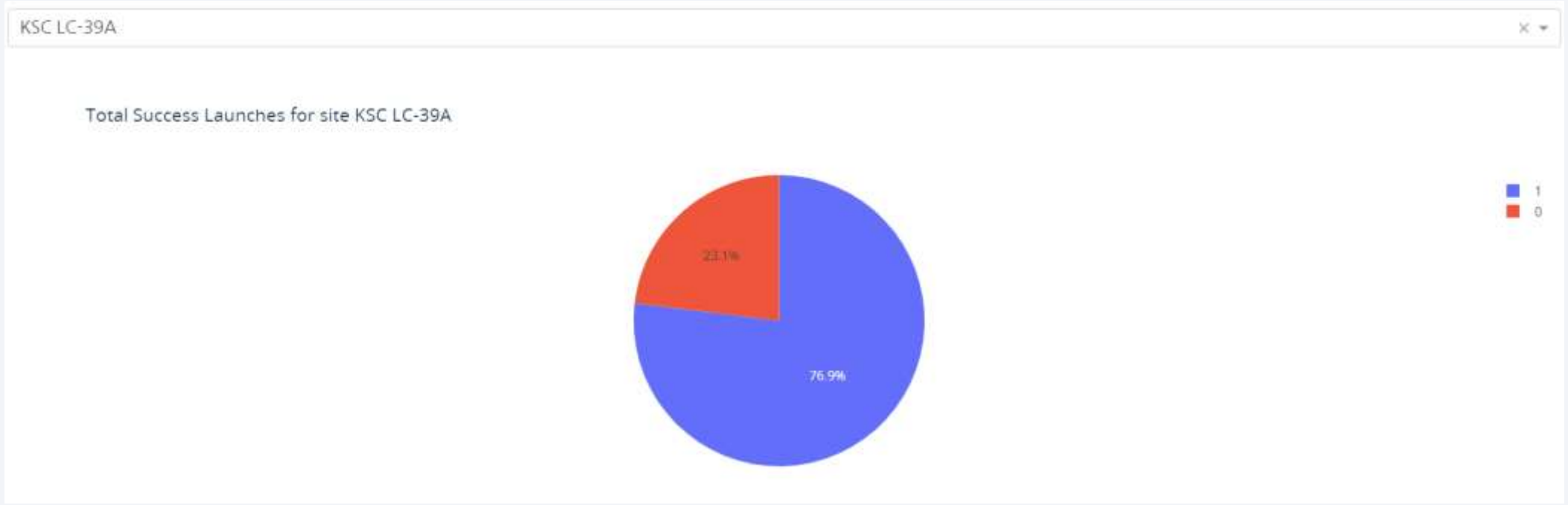
Number of Successful Launches for all Sites

- **KSC-LC-39A** had the most successful launches, and CCAFS SLC-40 had the least successful launches.



Launch Site with highest Launch success ratio

- **KSC LC-39A** had the highest launch success rate of 76.9%.



Payload Mass and Booster Version vs. outcome

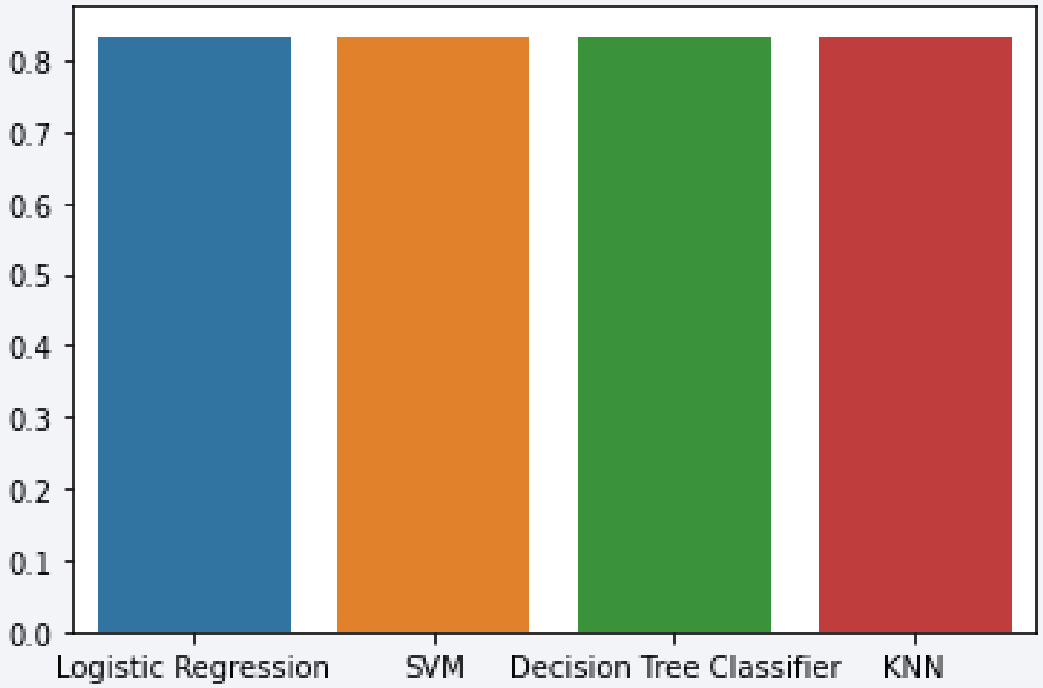
- Launches with a payload range *between 2000 and 4000 kilograms* appear to have a much higher success rate than the rest.
- The booster version categories of FT has a high success rate, while v1.0 and v1.1 boosters have a low success rate.

Section 6

Predictive Analysis (Classification)

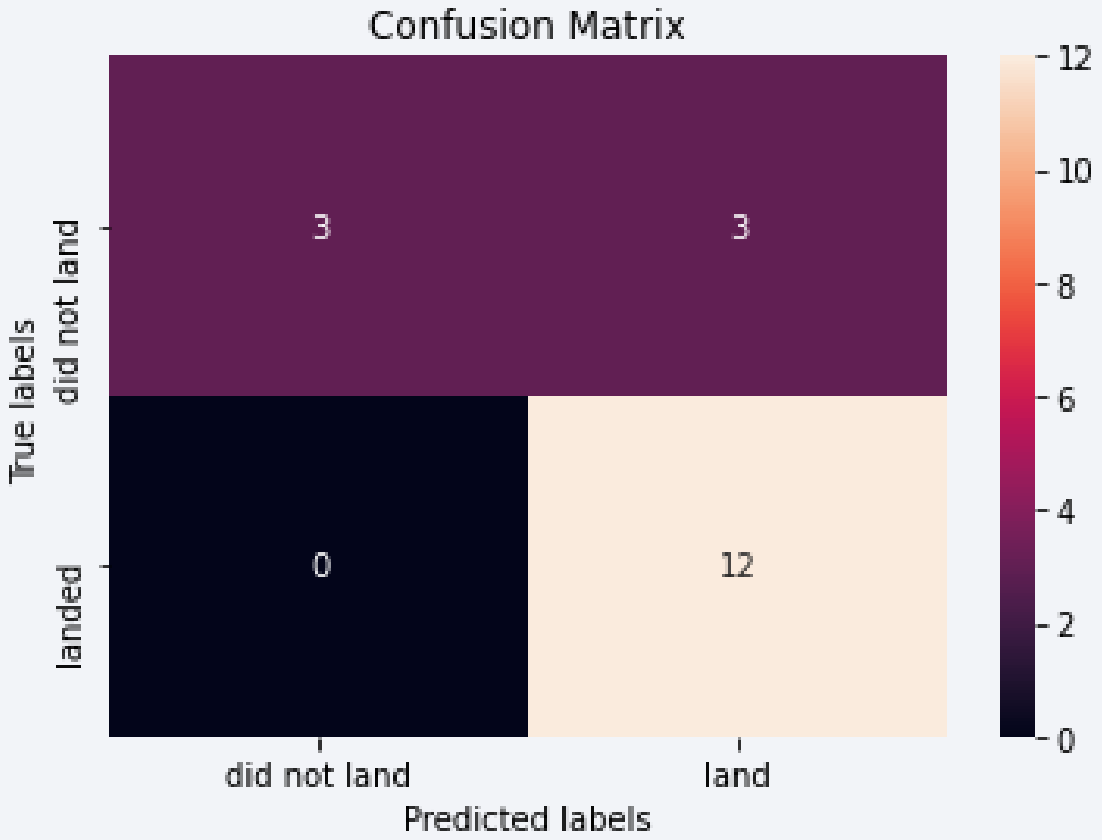
Classification Accuracy

- All algorithms have the *same accuracy of 0.833*, classifying 15 cases correctly out of 18.
- We cannot use this information to determine the best performing algorithm.



Confusion Matrix

- All algorithms classified the test set in the same way, yielding identical confusion matrixes.
- All algorithms misclassified three of the failed launches as successes, and none of the successful launches as failures.
- The major problem is *false positives*.



Conclusions

- From EDA with Data Visualization, we were able to conclude that the *success rate of launches has increased over time*, and we were able to see the effect certain variables have on the success rate for each launch site and intended orbit.
- From EDA with SQL, we were able to gain insight about the individual launches through queries on the dataset.
- From interactive visual analytics, we were able to gain insight about the launch sites and the launches performed in each of them. We were also able to see the success rate for certain payload mass ranges and booster versions.
- Using four different algorithms, we were able to train models that can *classify the outcome of a launch using publicly available data* with 83.3% accuracy.

Thank you!