Winning Space Race with Data Science

Anar Seyf
November 2021

Executive Summary

The goal of this project is to provide analysis on SpaceX launches, with focus on stage reuse and landings, to provide our company, SpaceY, competitive information on cost of launches and prediction methods on whether SpaceX will reuse a given stage.

This report details launch and landing data collection and exploration methods, followed by exploratory and visual analysis results, and finally the results of training classification models on our data to predict landing success (and thus stage reuse, with implications on cost).

Introduction

The report presents analysis results on SpaceX launch and landing data, including:

- Data collection and preparation methods (API calls, web scraping)
- Exploratory analysis (SQL, Plotly charts)
- Interactive visualization (Folium maps, Dash)
- Prediction using classification models.

A detailed Methodology section dives into each element of the analysis, followed by results of exploratory analysis, launch site mapping, launch success visualization, and finally results of predictive analysis.

Section 1

Methodology

Methodology

Executive Summary

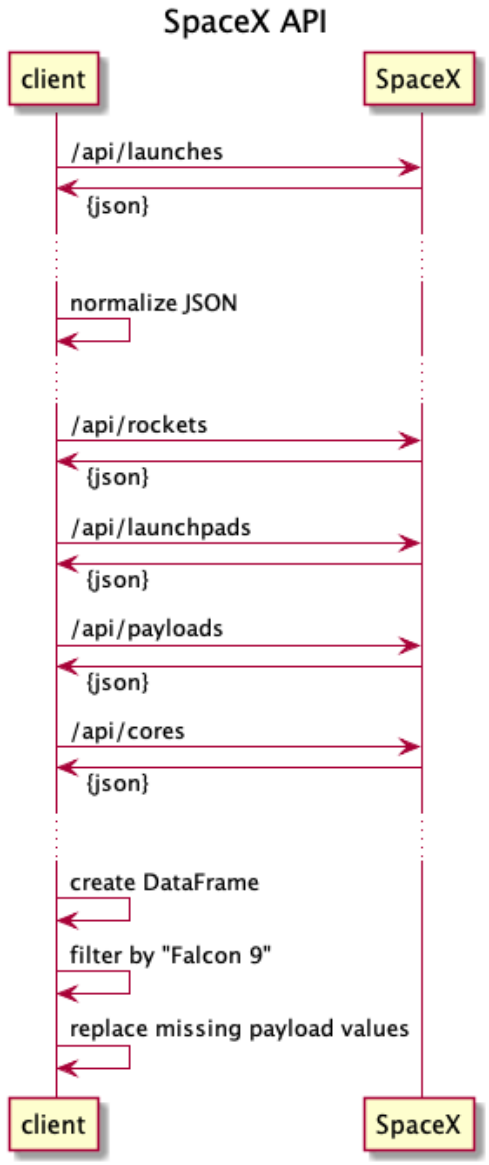
- **Data collection**:
  - SpaceX API: Past landing data, plus info on boosters, cores, payloads;
  - Wikipedia scrape: Parse tables of past launches (including outcomes — success/failure) into a DataFrame for next stage.
- **Data wrangling**:
  - Outcomes (“True ASDS”, “True Ocean”, etc.) grouped into two classes — success (1) and failure (0);
  - Column ‘class’ added to DataFrame, to be used for model building and prediction.
- **Exploratory data analysis** (EDA) using Visualization and SQL:
  - Data loaded into IBM Cloud DB2 instance;
  - SQL queries run from Jupyter notebook;
  - Plotly and Seaborn; categorical plots (scatter, bar, line);
  - One-hot encoding of features.
- **Interactive visual analytics** using Folium and Plotly Dash:
  - Folium maps + marker clusters: show launch sites;
  - Dashboard: show successes for all launch sites, success percentage per site, and payload per launch.
- **Predictive analysis** using classification models:
  - Split data between training and testing;
  - Train and test 4 models: LogReg (logistic regression), SVM (support vector machine), Tree (decision tree), KNN (K nearest neighbors);
  - Grid Search with a set of hyperparams; check model scores (R^2); build confusion matrix for each; —> select best model.

Data Collection

- Fetch info on **past launches** from SpaceX API; join with data on rockets, launchpads, payloads, cores (also from the API);
- Normalize JSON;
- Convert JSON to a pandas DataFrame;
- Filter entries to “Falcon 9” only;
- Replace missing payload values with the mean payload.

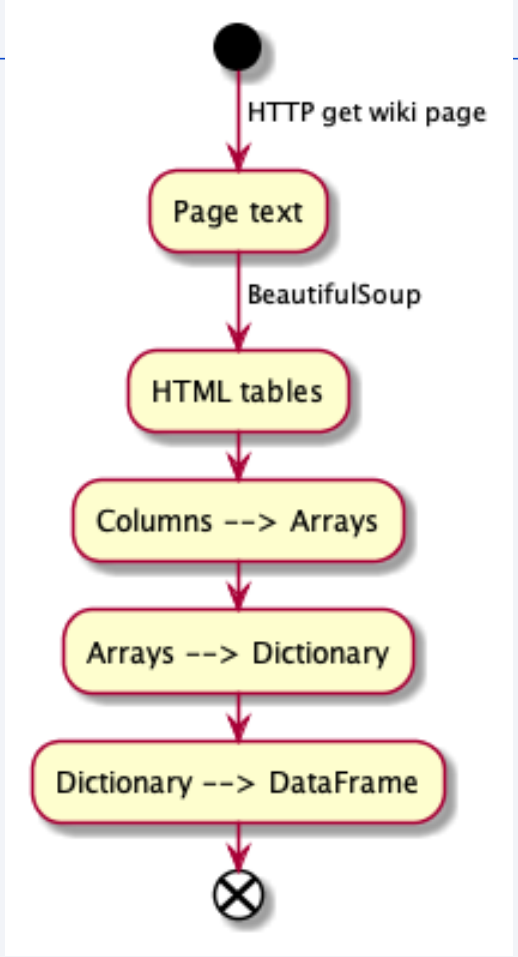
Data Collection – SpaceX API

- Data on past launches is retrieved from the API as JSON, normalized into a DataFrame format, then enriched with data on rockets, launchpads, payloads, and cores (also from the API).
- The DataFrame is filtered to Falcon 9 entries only.
- Missing payload values are replaced with the mean payload.
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/1-jupyter-labs-spacex-data-collection-api.ipynb



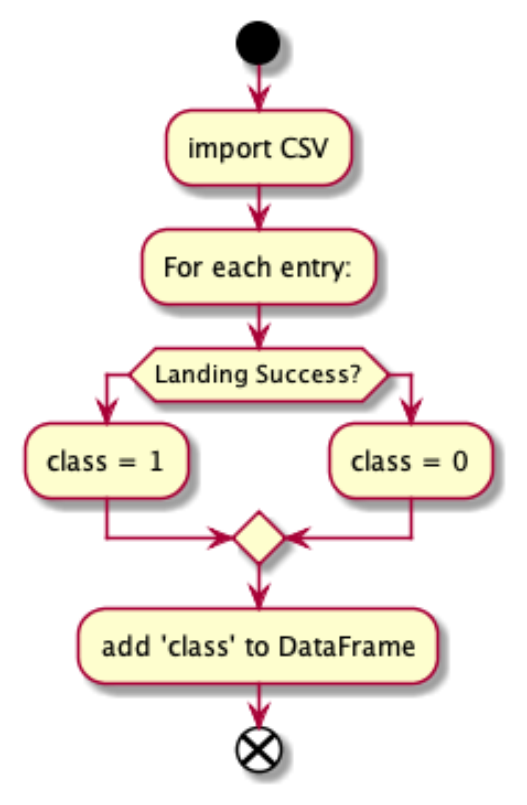
Data Collection - Scraping

- The text of the wikipedia entry “List of Falcon 9 and Falcon Heavy launches” is retrieved and parsed into a searchable HTML structure.
- Columns from each table listing past launches are converted into lists of values, joined into a dictionary, and converted to a DataFrame.
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/2-jupyter-labs-webscraping.ipynb



Data Wrangling

- The source data contains 8 landing outcome values, 5 of which are considered failures (such as “False Ocean”).
- We convert these values to a success (1) or failure (0) class, and add a “class” column to the DataFrame.
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/3-jupyter-labs-spacex-data-wrangling.ipynb



EDA with SQL

These SQL queries were run:

- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/4-jupyter-labs-eda-sql-coursera.ipynb

- List distinct launch sites;
- Find records with launch site names starting with ‘CCA’;
- Calculate total payload carried by boosters launched by NASA;
- Calculate average payload carried by a specific booster version;
- Find the date of the first successful ground pad landing;
- List boosters with payloads in a specific range that have successfully landed on a drone ship;
- Show the total number of mission successes and failures;
- List the booster versions that have carried the maximum payload;
- List failed drone ship landings, their booster versions, and launch sites in 2015;
- Show each landing outcome in a specific date range, ranked by total count.

EDA with Data Visualization

Charts plotted:

- Flight # vs Payload
  - Show payload trend over time (with success trends)
- Flight # vs Launch site
  - Show launches at each site (with success trends)
- Payload vs Launch site
  - Show payload distribution at each site (with success trends)
- Success rate per orbit (bar)
  - Look for success trends across orbit types
- Flight # vs Orbit
  - Look for trends in which orbits are “busiest” at a given time
- Success rate by year (line)
  - Show high-level trend of launch “quality” over time

- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/5-jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

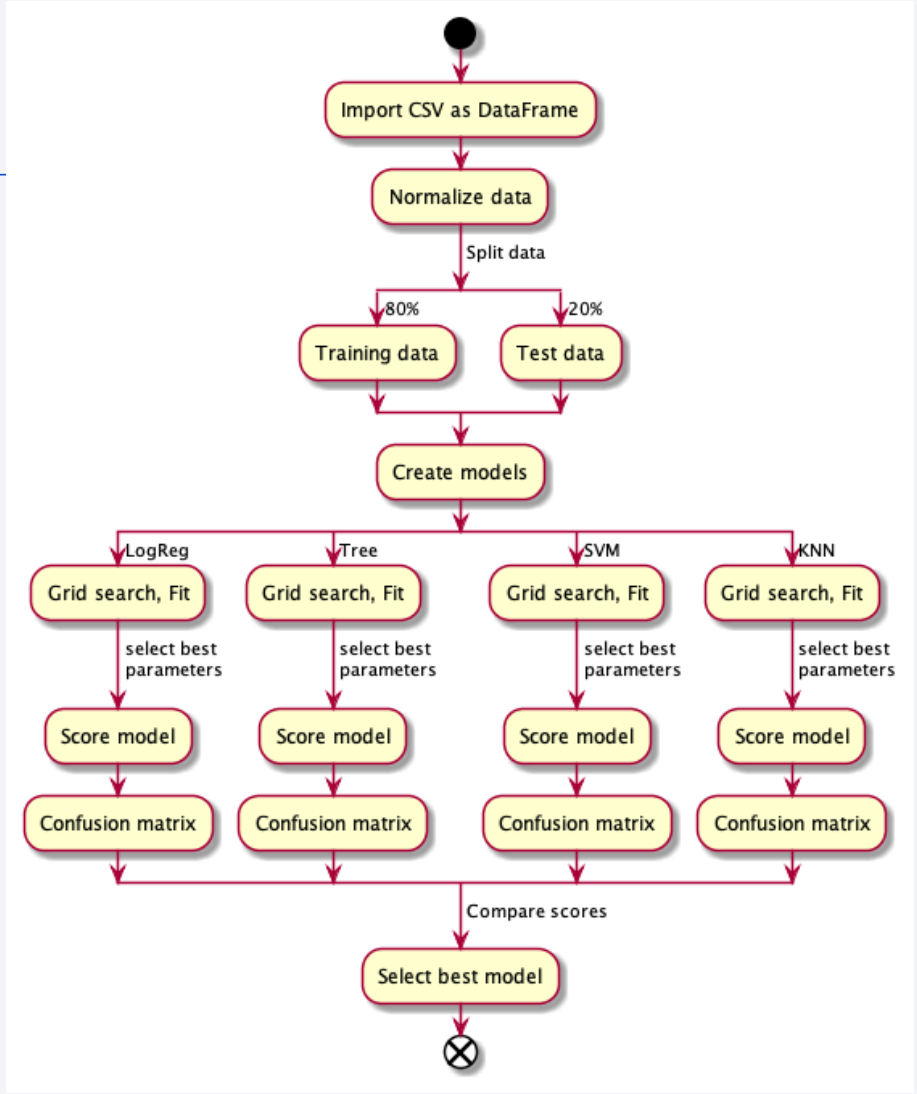
- Objects added to interactive map:
  - Circle — show location of each launch site
  - Marker — label each launch site's name
  - Popup — show more info on mouse click (for interactivitiveness)
  - Marker clusters — to show all launch successes/failures at each site
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/6-jupyter-labs-launch_site_location.ipynb

Build a Dashboard with Plotly Dash

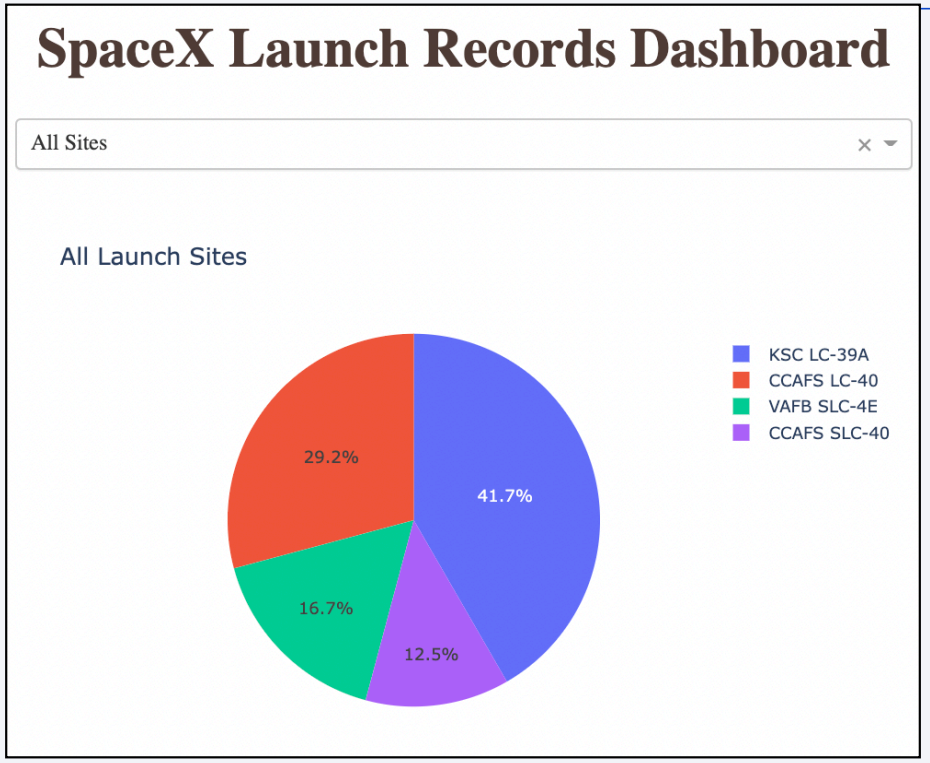
- The Dashboard shows:
  - Interactions: selectors for launch sites and payload limits;
  - Plots:
    - Successful Launches — all and per site (pie chart)
    - Payloads vs Success/Failure (scatterplot; color indicates success or failure).
- These enable a quick overview of success rates at a given launch site and for a given payload range.
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/dash/spacex_dash_app.py
- See **screenshots** in the following slides.

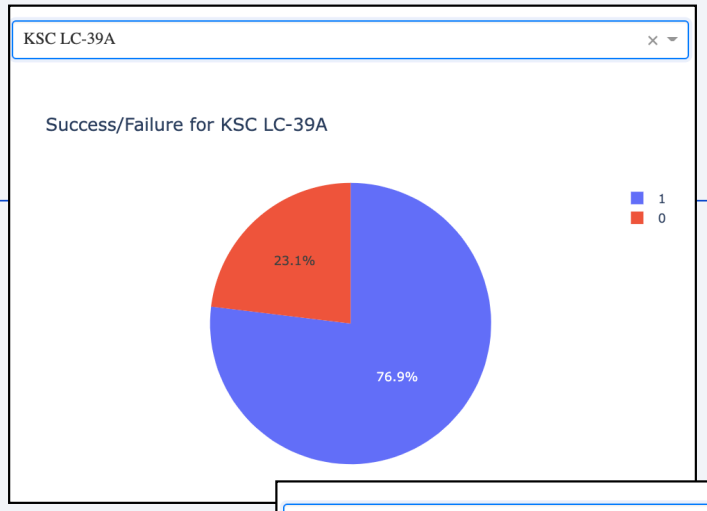
Predictive Analysis (Classification)

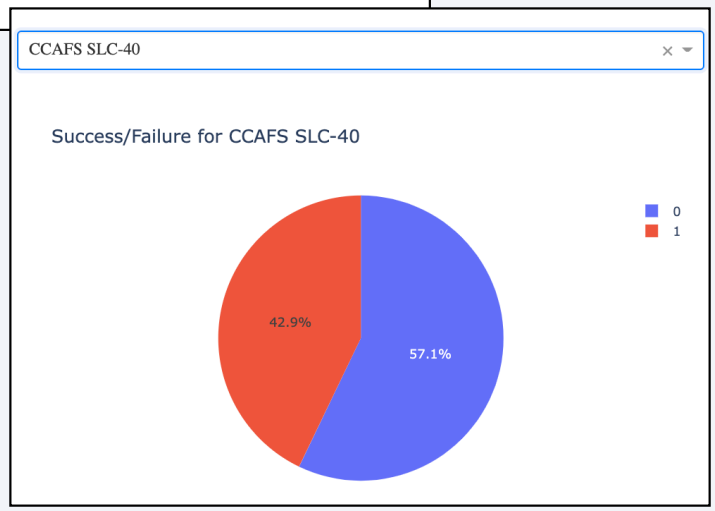
- Four models were trained and tested using the process in the chart. The models were then compared on training data accuracy as well as test data scores (R^2).
- https://github.com/anarseyf/ibm-ds-spacex/blob/main/notebooks/7-jupyter-labs-ml-prediction.ipynb



Dashboard: Success rates



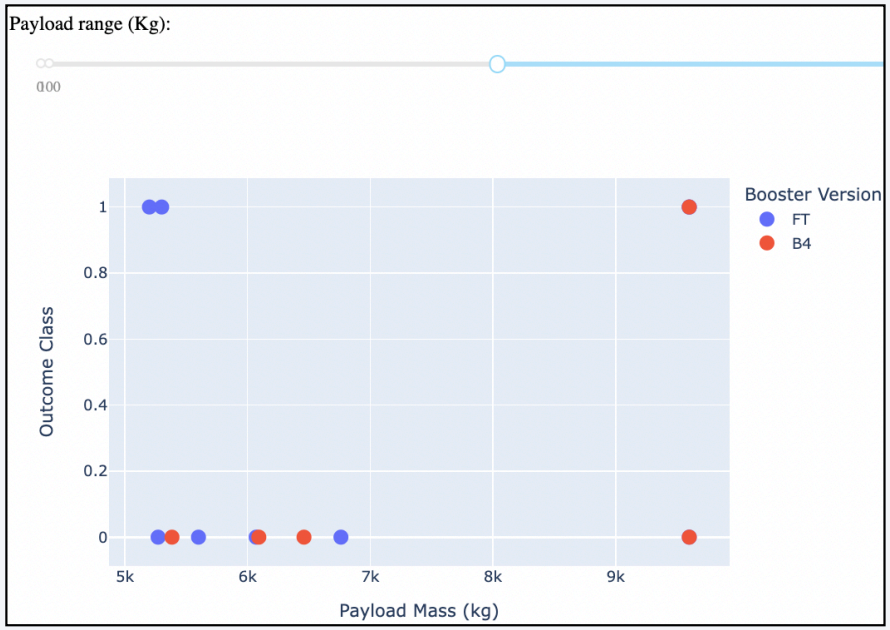




- LC-39A has the most successful launches (by count, not by rate)
- SLC-40 has the least

- However, LC-39A has the lowest rate of successes (23%)…
- …while SLC-40 has the highest (43%)

Dashboard: Payloads



- More lighter loads (up to 5000 kg)
- Higher success percentage
- Most successes on FT and B4 boosters

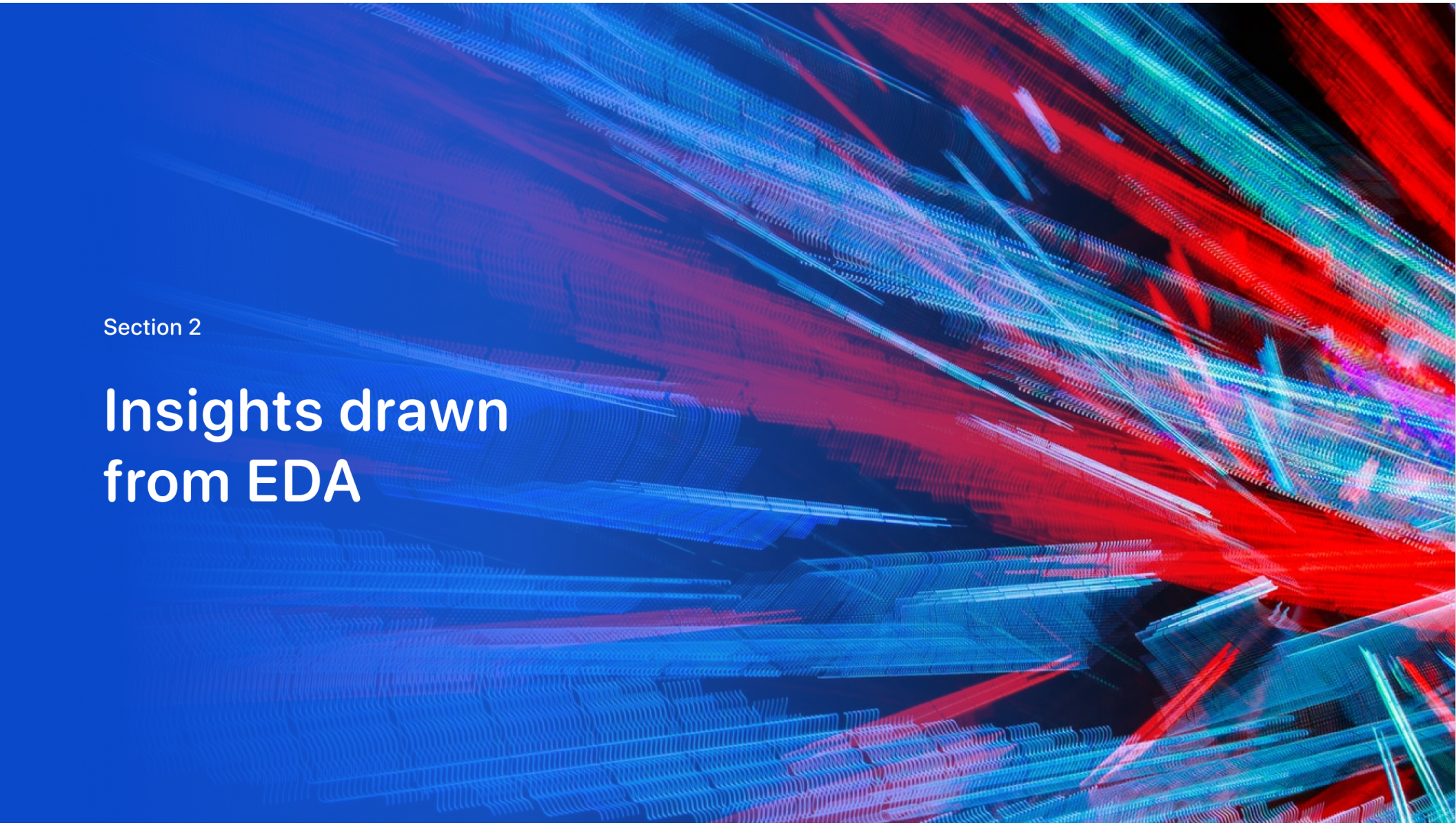
- Few heavy payloads (over 5000kg)
- Far more failures than successes
- Only two booster types seem capable of carrying the higher loads (FT and B4)

Results: Exploratory analysis

- SQL:
  - Majority of loads are to **Geosynchronous orbit** — communications, weather satellites
  - Second most common is Low-Earth orbit (including ISS)
  - Majority of landings are on a **drone ship** (ASDS)
  - Overall **landing success rate is 2/3** (67%)
  - Out of 101 launches in 2010-2020, **only 1 was deemed a failure** in terms of payload delivery.
    (Another one, the secret Zuma satellite launched on Jan 8, 2018 may have been lost, but its status is classified.)
- Visualization:
  - Most initial launches were failures, but the success rate dramatically improved in the last few years.
  - The vast majority of launches are performed in Florida, with only a handful in California (VAFB).
  - Payload capacity of Falcon 9 has increased, reaching a **maximum of 15,400kg** starting in 2020.
  - There are two distinct clusters of payload range: for ISS the range is 2,000-3,000 kg, and for GTO (geosynchronous) it is higher, at 3,000-7,000 kg.
  - All launch sites are located in close proximity to the ocean: this minimizes risk to populated areas in case of rocket failure.
  - All launch sites are also situated in the southern latitudes of the US: The centrifugal force close to the equator provides some assistance with the launch.

Results: Predictive analysis

- Four models were trained and tested on 90 rows of data:
  - LogReg (logistic regression)
  - SVM (support vector machine)
  - Tree (decision tree classifier)
  - KNN (K nearest neighbors)
- Accuracy on test data was the **highest for SVM** at **88%.** The other models were at **85%**.
- The Tree model's R^2 score on test data was lowest: **0.61**, vs **0.83** for the others.
- **Confusion matrix was identical** for the three higher-performing models, with 3 false positives and no false negatives. The Tree model had 5 false positives and 2 false negatives.
- A larger training dataset is needed to distinguish more clearly between the models.

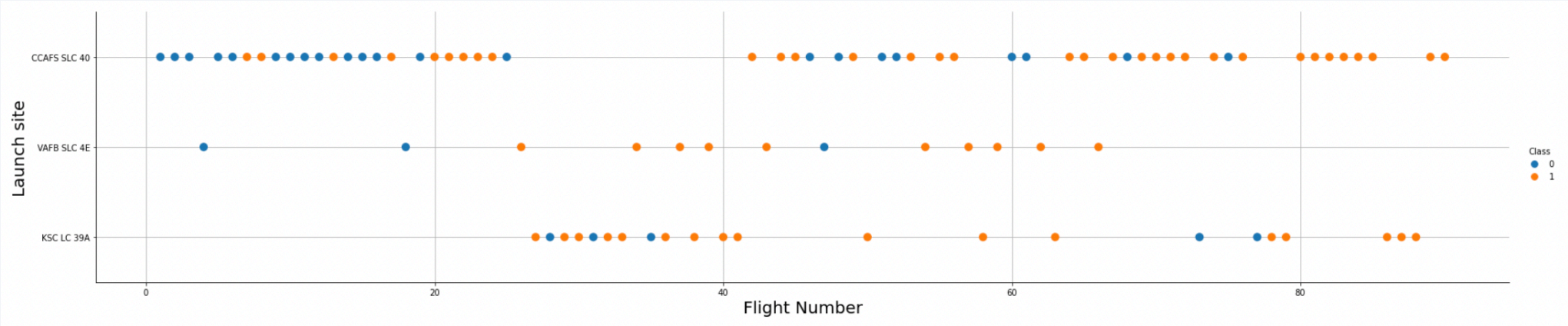


Section 2

Insights drawn from EDA

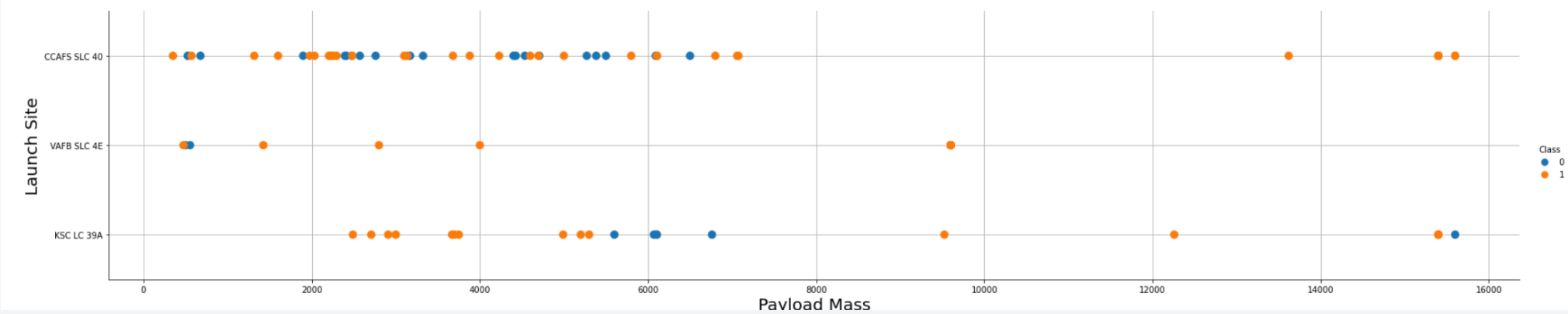
Flight Number vs. Launch Site

- Flight Number vs. Launch Site
- Color indicates success vs failure
- Y axis shows launch sites. The middle one (VAFB) is in California, the others are close together in Florida (Kennedy Space Center)
- KSC sites perform the vast majority of launches.



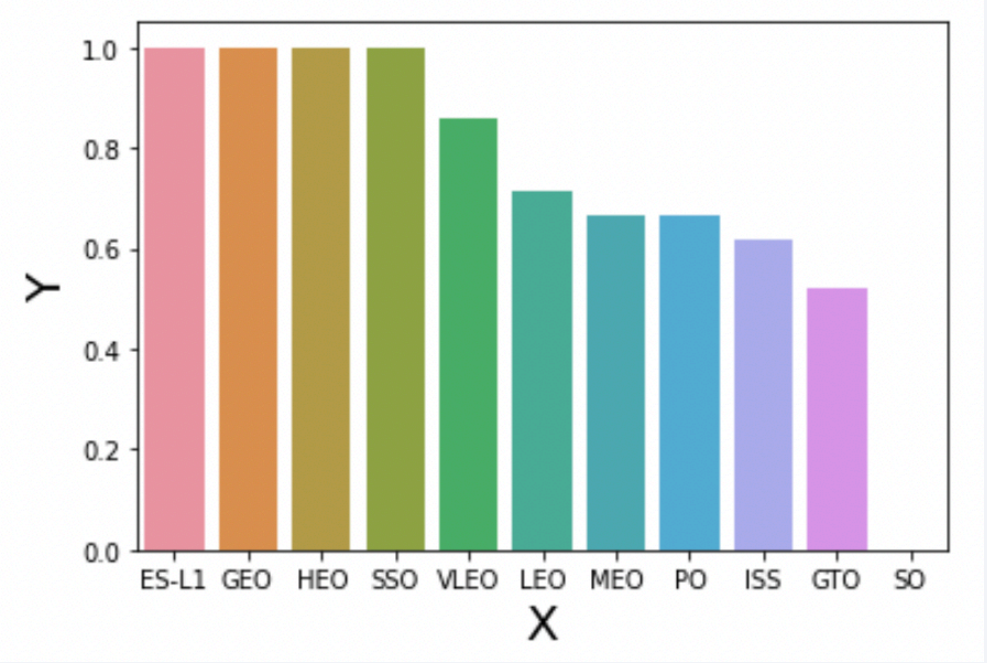
Payload vs. Launch Site

- Payload (kg) by Launch site
- Color indicates success vs failure
- Majority of payloads are below 6,000 kg (Falcon 9 has only reached a maximum of 15,400 kg starting in 2020)
- Similar to previous slide, Y axis represents launch sites. The middle one is VAFB in California, the rest at Kennedy Space Center in Florida.



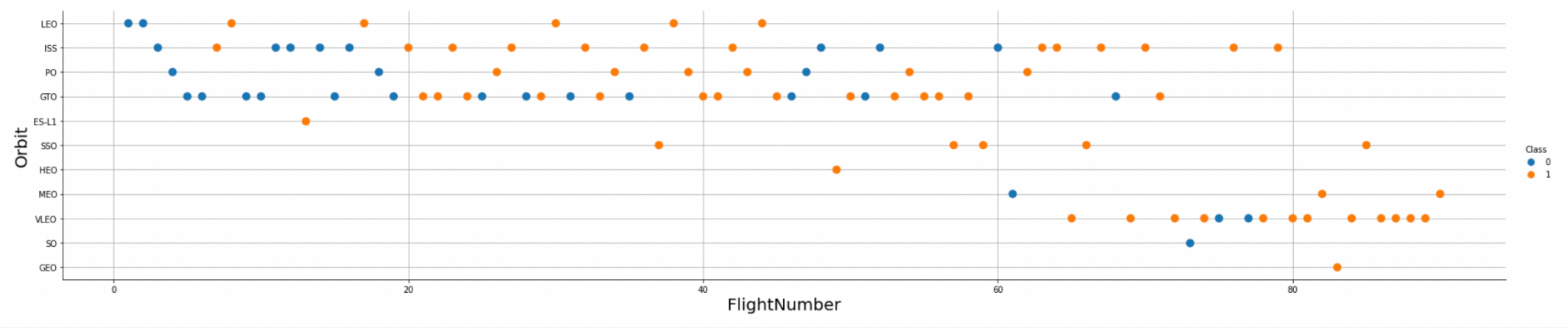
Success Rate vs. Orbit Type

- Success rate for each orbit type
- The low-earth orbits (VLEO, LEO, ISS) have lower success rates. GTO (geosynchronous) is also low. These types are also the majority of total launches (see the Exploratory Analysis slide).
- SO (Sun-syncronous) shows a success rate of 0, but there has only been one launch of that type.



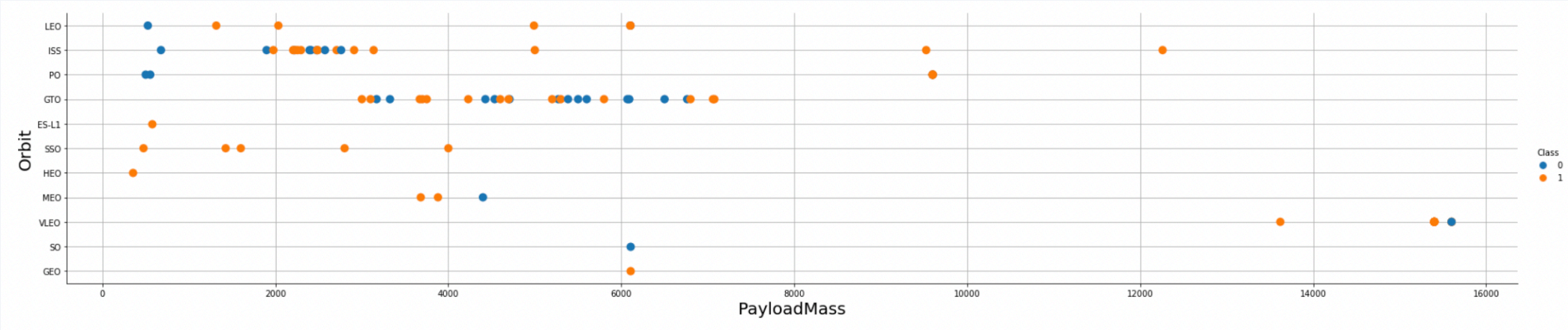
Flight Number vs. Orbit Type

- Y axis shows the orbit type
- Most early launches have been to LEO/ISS orbit, as well as to GTO (geosynchronous)
- More recently the VLEO orbit has been “busy”. This corresponds to SpaceX’s launches of Starlink satellite batches in 2019-2020.



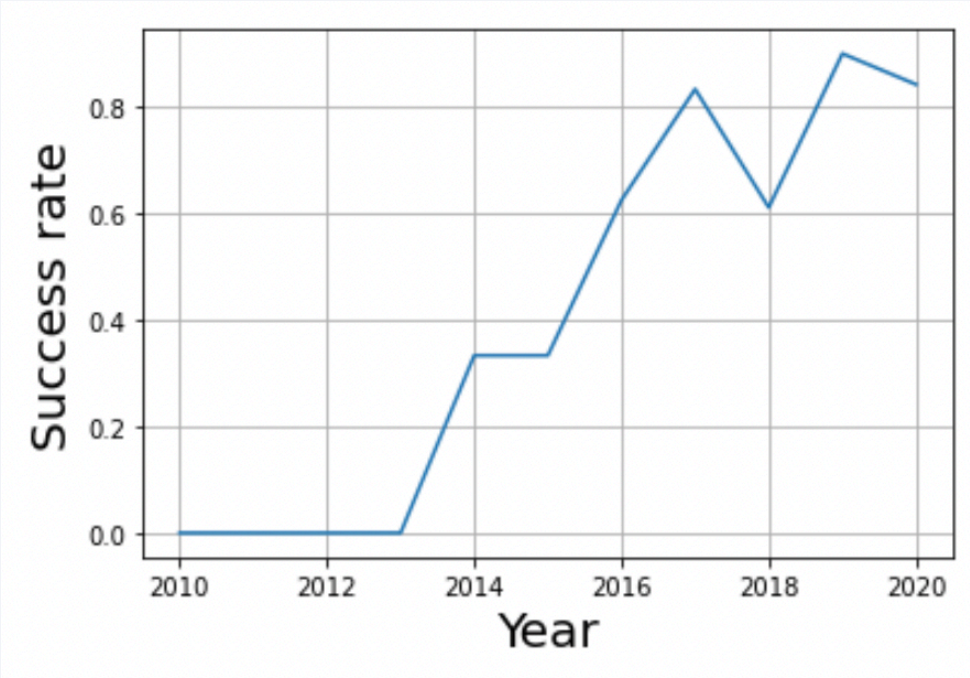
Payload vs. Orbit Type

- Payload vs. orbit type
- As noted in Exploratory Analysis results, two distinct clusters:
  - 2,000-3,000 kg ISS payloads
  - 3,000-7,000 kg Geosynchronous payloads (comms, weather sats, etc.)



Launch Success Yearly Trend

- Average success rate across all launches.
- The trend confirms our observations in previous slides: most early launches were failures (blue dots), and most recent ones were successes (orange dots).



All Launch Site Names

- Four distinct launch sites:
  - Three in Florida (all clustered close together at Kennedy Space Center)
  - One in California (VAFB)

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Not too much to say here…

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 carried by boosters from NASA: **45,596 kg**

```
%sql select sum(payload_mass__kg_) from spacextbl
where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b
027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.
appdomain.cloud:30875/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: **2,534 kg**

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

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b
027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.
appdomain.cloud:30875/bludb
Done.
```

1
2534

First Successful Ground Landing Date

- First successful landing on ground pad: **December 22, 2015**

```
%sql select date from spacextbl \
where landing__outcome = 'Success (ground pad)' \
order by date limit 1
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b
027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.
appdomain.cloud:30875/bludb
Done.
```

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had 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_ between 4000 and 6000
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b027
-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdom
ain.cloud:30875/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- 100 successes, 1 failure
- One of the “successes” is the secretive Zuma satellite launch in 2018, with “payload status unclear”. This is the reason for the subquery.

```
%sql select count(*) as total, outcome from \
(select \
     case mission_outcome like 'Success%' \
     when TRUE then 'Success' else 'Failure' \
     end outcome \
from spacextbl) \
group by outcome
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

total	outcome
1	Failure
100	Success

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass.
- (The maximum is 15,400 kg.)

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

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd
0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
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

```
%sql select booster_version, launch_site, landing__outcome from spacextbl \
where landing__outcome = 'Failure (drone ship)' and extract(year from date) = 2015
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databas
es.appdomain.cloud:30875/bludb
Done.
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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 count(*) as total, landing__outcome from spacextbl \
where date between '2010-06-04' and '2017-03-20' \
group by landing__outcome \
order by total desc
```

```
 * ibm_db_sa://jpx92421:***@98538591-7217-4024-b027-8baa776ffad1.c3
es.appdomain.cloud:30875/bludb
Done.
```

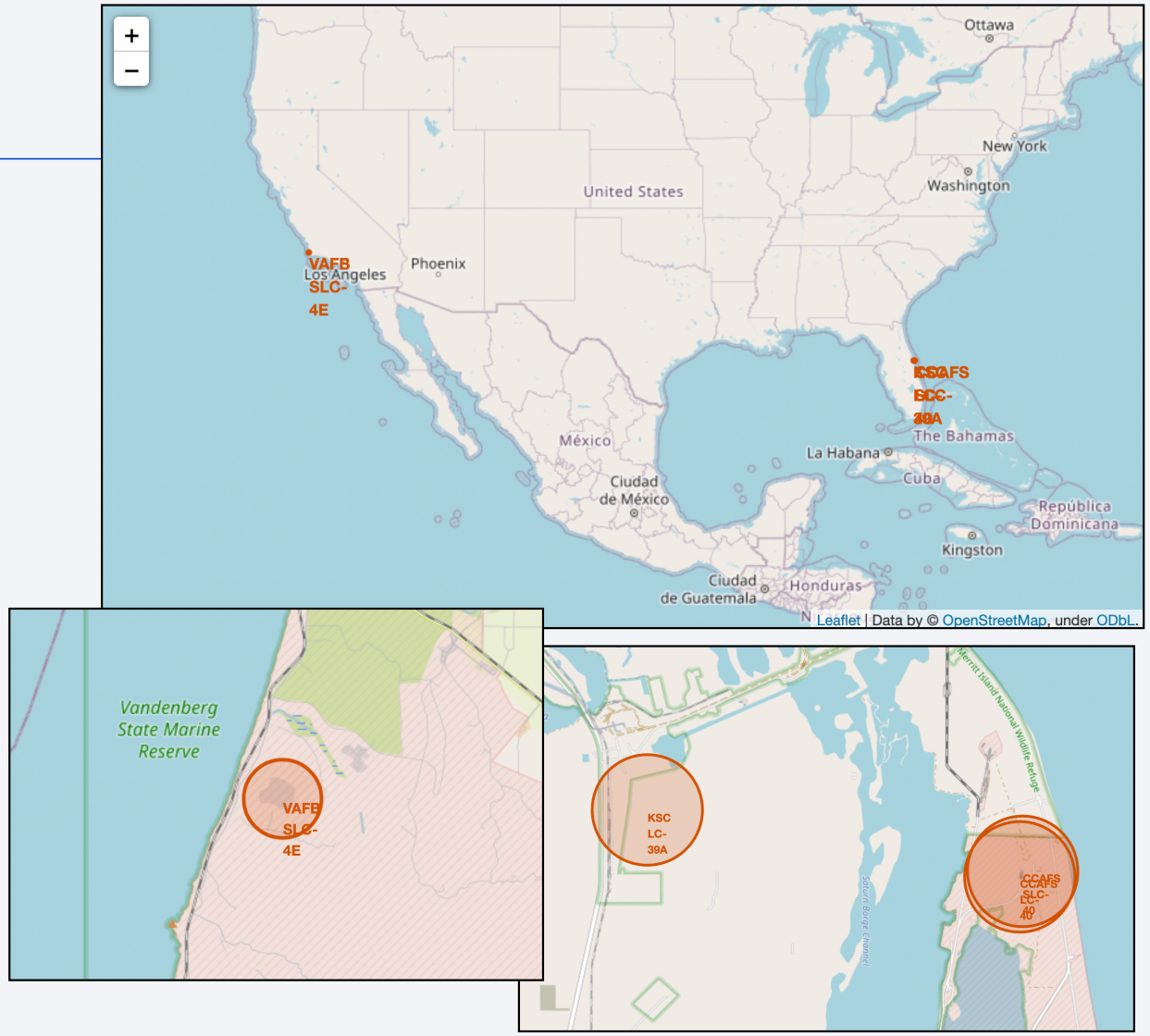
total	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

Launch Sites Map

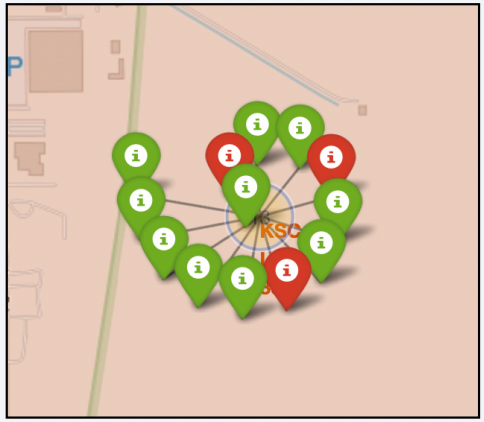
- The main map shows all 4 launch sites, and the insets zoom in on each site.
- All sites are located close to coastline (to reduce risk to populated areas).
- All sites are located south (to take advantage of Earth's centrifugal force close to the equator).



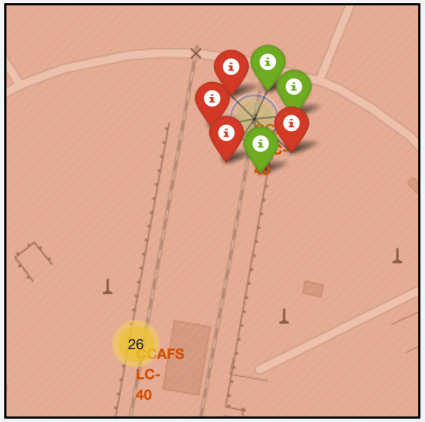
Successes vs Failures at each Launch site

- These inset zoom in on each of the launch sites.
- A green dot indicates a successful launch, a red dot a failure.
- VAFB (California) is in lower-left.
- At KSC most launches have been successful.
- At CCAFS LC-40, most early launches have been failures, but recent ones (outermost part of the spiral) were successes.

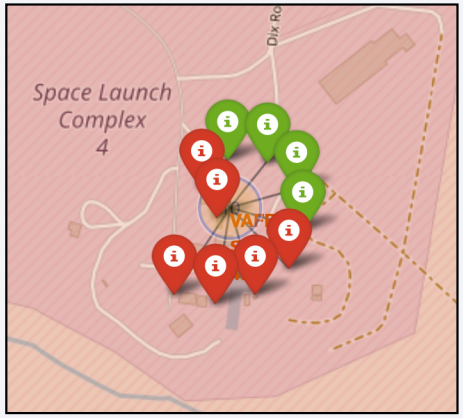
KSC LC-39A



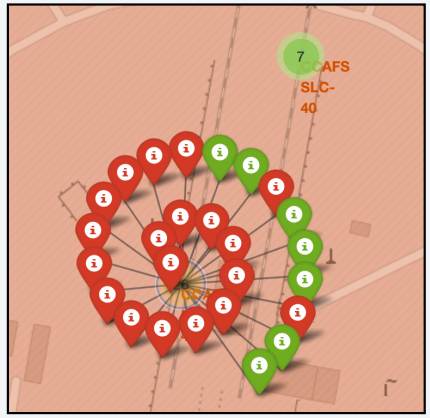
CCAFS SLC-40

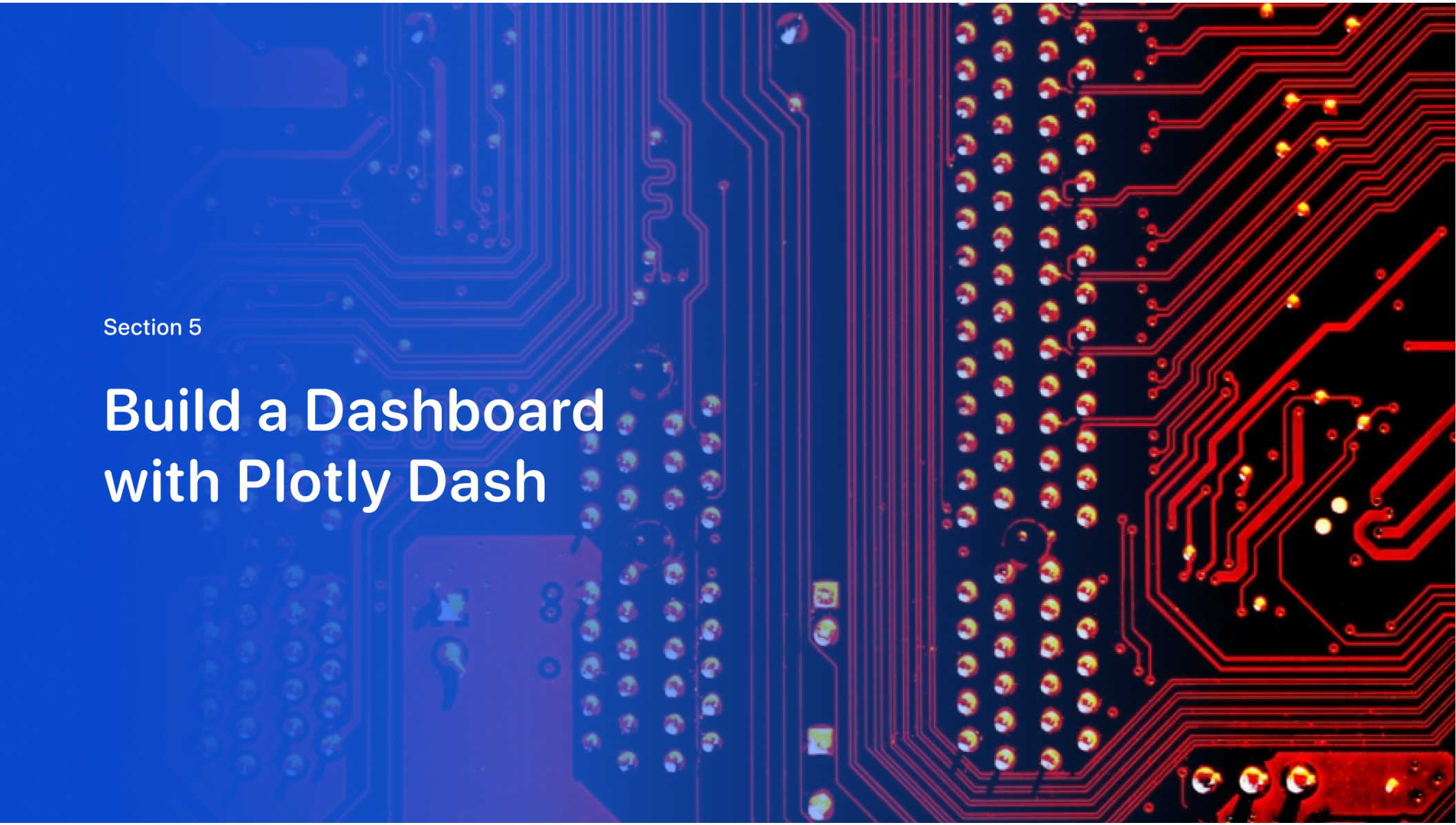


VAFB SLC-4E



CCAFS LC-40



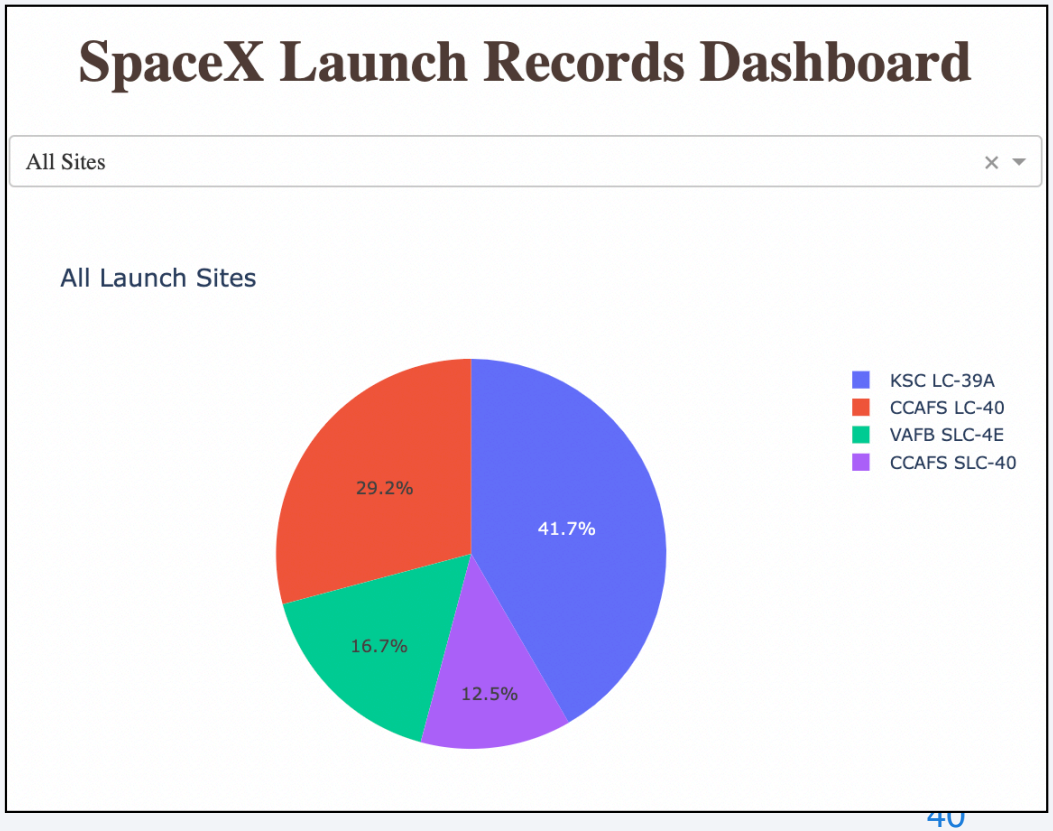


Section 5

Build a Dashboard with Plotly Dash

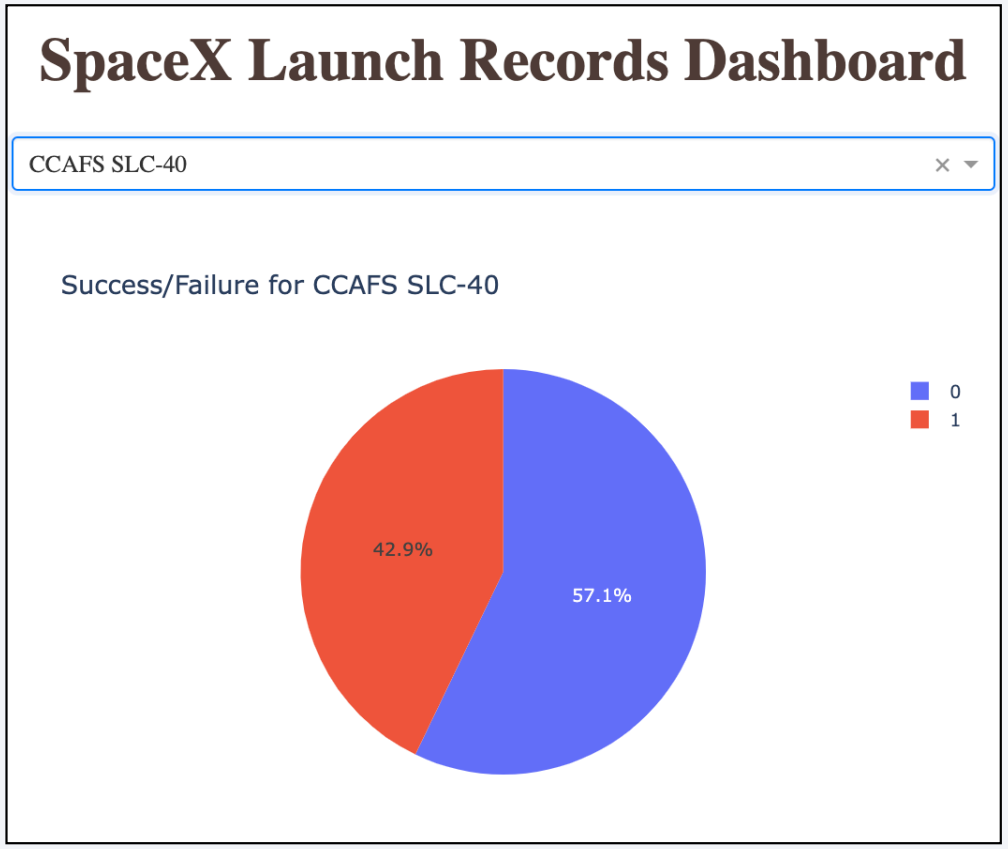
Dashboard: Launch Successes — All sites

- Share of successful launches by site (all 4 sites)
- KSC has highest share at **41.7%**, and CCAFS SLC-40 has the lowest at **12.5%**
- Note that this is not success **rate** — that is shown on the next slide



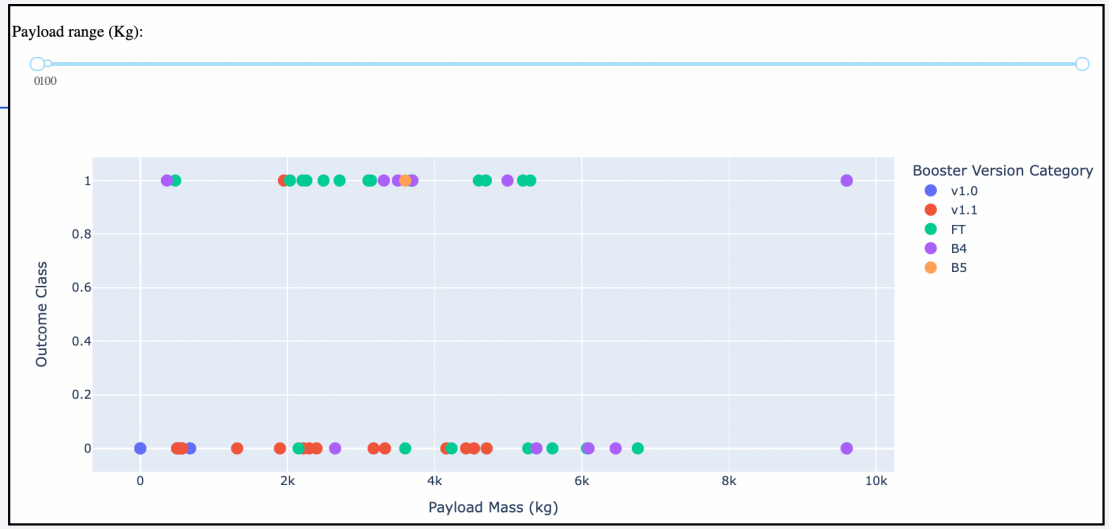
Dashboard: Launch Successes by Site

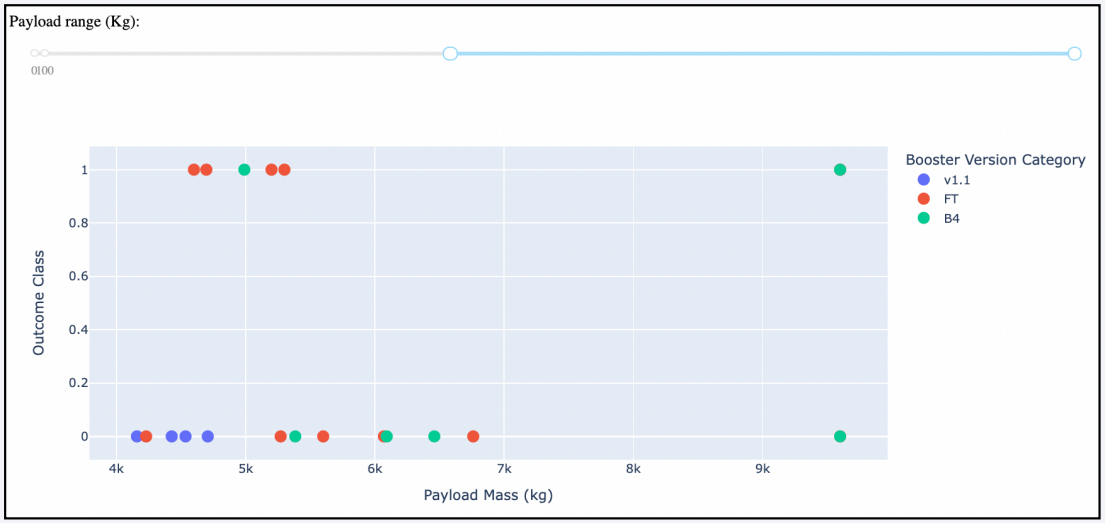
- CCAFS SLC-40 has the highest success ratio at **42.9%**.



Dashboard: Payload range

- X axis: Payload (kg)
- Y axis: Success (1) vs Failure (0)
- First screenshot: all payloads
- Second screenshot: only the heavy payloads (4,000 kg+)
- The FT and B4 boosters have the highest overall success rate
- FT and B4 are also the only ones carrying the highest payloads



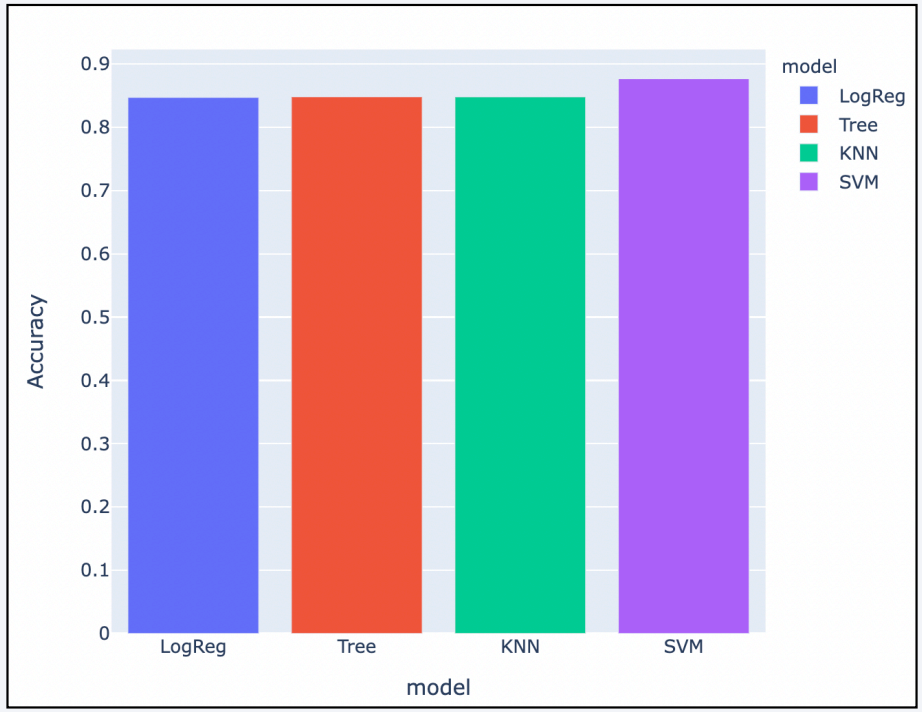


Section 6

Predictive Analysis (Classification)

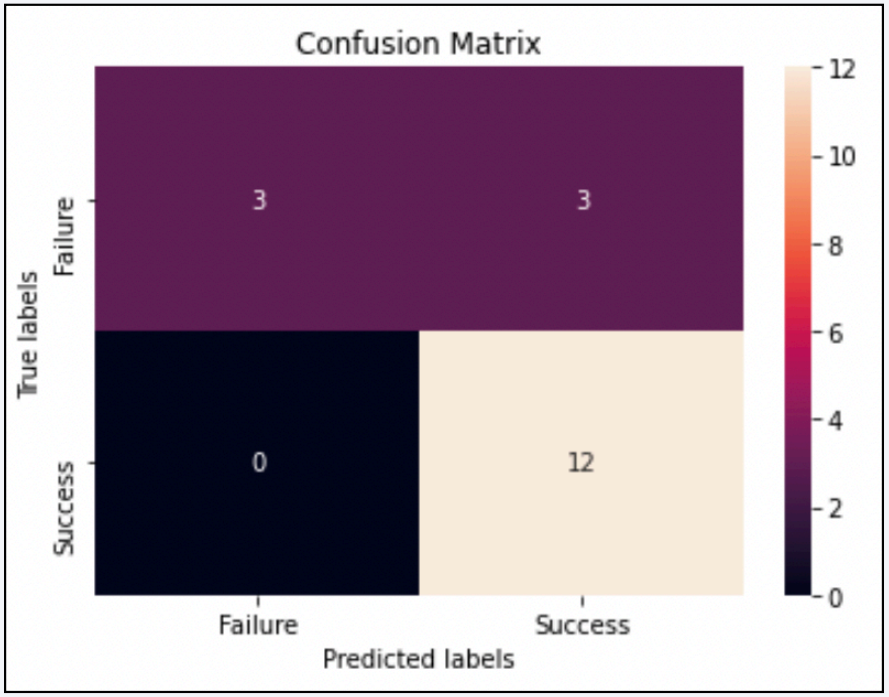
Classification Accuracy

- Model accuracy for all 4 models.
- The **SVM** (support vector machine) model had the best accuracy at **0.88 on training data**, with the rest identical at 0.85.
- On test data LogReg, SVM and KNN all performed similarly, with an R^2 score of 0.83. The Tree model had a lower score of 0.62-0.77 (varied depending on the run).



Confusion Matrix

- The three models with the higher R^2 score produced identical confusion matrices, with 3 false positives and no false negatives:



Conclusions

- Out of 4 prediction models (LogReg, SVM, Tree, KNN), **SVM had the best accuracy** on training data, and among the best at scores on test data.
- The Tree model appeared the weakest.
- Test vs Training data accuracy are not the same.
- The dataset is too small (only 90 observations) to distinguish any further between these models.