Winning Space Race with Data Science

Sean K. LIANG
30 March 2024

Outline

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

Executive Summary

- Methodology
  - **Data collection** using SpaceX REST API and web scraping
  - **Data wrangling** by filtering data, eliminate missing values & one-hot encoding
  - **Data exploring** by exploratory data analysis (EDA) using visualization and SQL
  - **Data visualization** with interactive visual analytics using Folium and Plotly Dash
  - **Predictive analysis** by building, tuning, and evaluating classification model
- Results
  - Higher payload mass has higher success rate for Polar, LEO and ISS orbits.
  - Launch success rate increases over time (from 2013 to 2020).
  - Orbits ES-L1, GEO, HEO and SSO have 100% success rate
  - KSC LC-39A has the highest success rate among launch sites.
  - The models performed similarly with the decision tree model slightly outperforming
  - All launch sites are closed to coastal area and equator.

Introduction

- Background

  SpaceX, a pioneering force in the space sector, aims to democratize space travel by ensuring affordability. Notable achievements include manned space missions and internet-providing satellite constellation. Central to its cost-effectiveness is the reuse of Falcon 9 rocket's first stage, reducing launch expenses to $62 million, whereas other providers charge over $165 million per launch. By leveraging public data and machine learning models, we aim to predict first stage successful rate, crucial for estimating launch costs and facilitating competitive bidding in the aerospace industry.

- Problems to resolve
  - What factors determine the success of the first-stage landing?
  - What is the pattern of the rate of successful landings over time?
  - What are the best operating conditions for a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection** using SpaceX REST API and web scraping
- **Data wrangling** by filtering data, eliminate missing values & one-hot encoding
- **Data exploring** by exploratory data analysis (EDA) using visualization and SQL
- **Data visualization** with interactive visual analytics using Folium and Plotly Dash
- **Predictive analysis** by building, tuning, and evaluating classification models

Data Collection – SpaceX API

- Dataset collected from SpaceX API
- Objective: Request to the SpaceX API and clean the requested data

Completed deliverables: Link

Data Collection – Scraping

- Web data collected from Wikipedia
- Objective: Web scrap Falcon 9 launch records with BeautifulSoup

Completed deliverables: Link

Data Wrangling

- Objective: Perform Exploratory Data Analysis and determine Training Labels (Features Engineering)
- During the Exploratory Data Analysis, the following calculation have been made to provide insights
  - # of launches on each site
  - # & occurrence of orbit
  - # & occurrence of mission outcome of orbits

Completed deliverables: Link

EDA with SQL

Performed SQL queries

- Names of the unique launch sites
- Top 5 records where launch sites begins with 'CCA’
- Total pay load mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on ground pad
- Names of the boosters which had success landing on drone ship and have payload mass between 4000 & 6000 kg
- Total number of successful and failure missions
- Names of booster versions which have carried the max payload
- Failed landing out comes on drone ship, their booster versions, and launch site names for the months in 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc order)

Completed deliverables: Link

EDA with Data Visualization

- Charts
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Success Rate of Orbits
  - Flight Number vs. Orbit
  - Payload Mass vs. Orbit
  - Launch Success Yearly Trend
- Analysis
  - Provide a clear view to show relationships between variables. Scatter plots, bar charts and line charts have been used for different type of variables.
  - Identify possible relationships between variables for further use in ML predictive analysis.

Completed deliverables: Link

Build an Interactive Map with Folium

- Markers with Circles for Launch Sites & NASA Johnson Space Center
  - Circles with popup labels added to show names and coordinates
- Marker clusters for Launch Outcomes in each Launch Sites
  - To show Successful (green) and unsuccessful (red) launches in each launch site
- Lines for Distances Between a Launch Site to Proximities
  - To show distances between launch site and nearest coastline, railway, highway, and city

Completed deliverables: Link

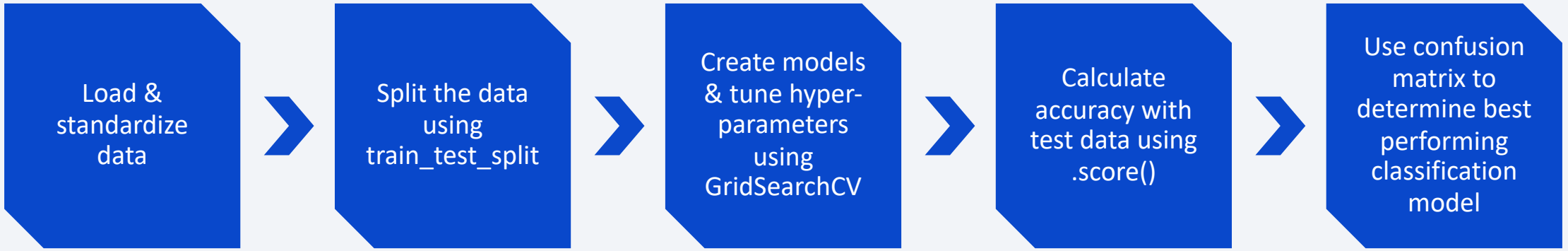
Build a Dashboard with Plotly Dash

- Dropdown List for Launch Sites Selection
  - User can select all launch sites or a specific launch site
- Pie Charts Showing Successful Launches
  - User can see successful launches from all sites, or launch outcomes from a specific launch site
- Slider of Payload Mass Range
  - User can filter a specific payload mass range
- Scatter Chart Showing Payload Mass vs. Outcome by Booster Version
  - User can see the correlation between Payload Mass and Outcome

Completed deliverables: Link

Predictive Analysis (Classification)

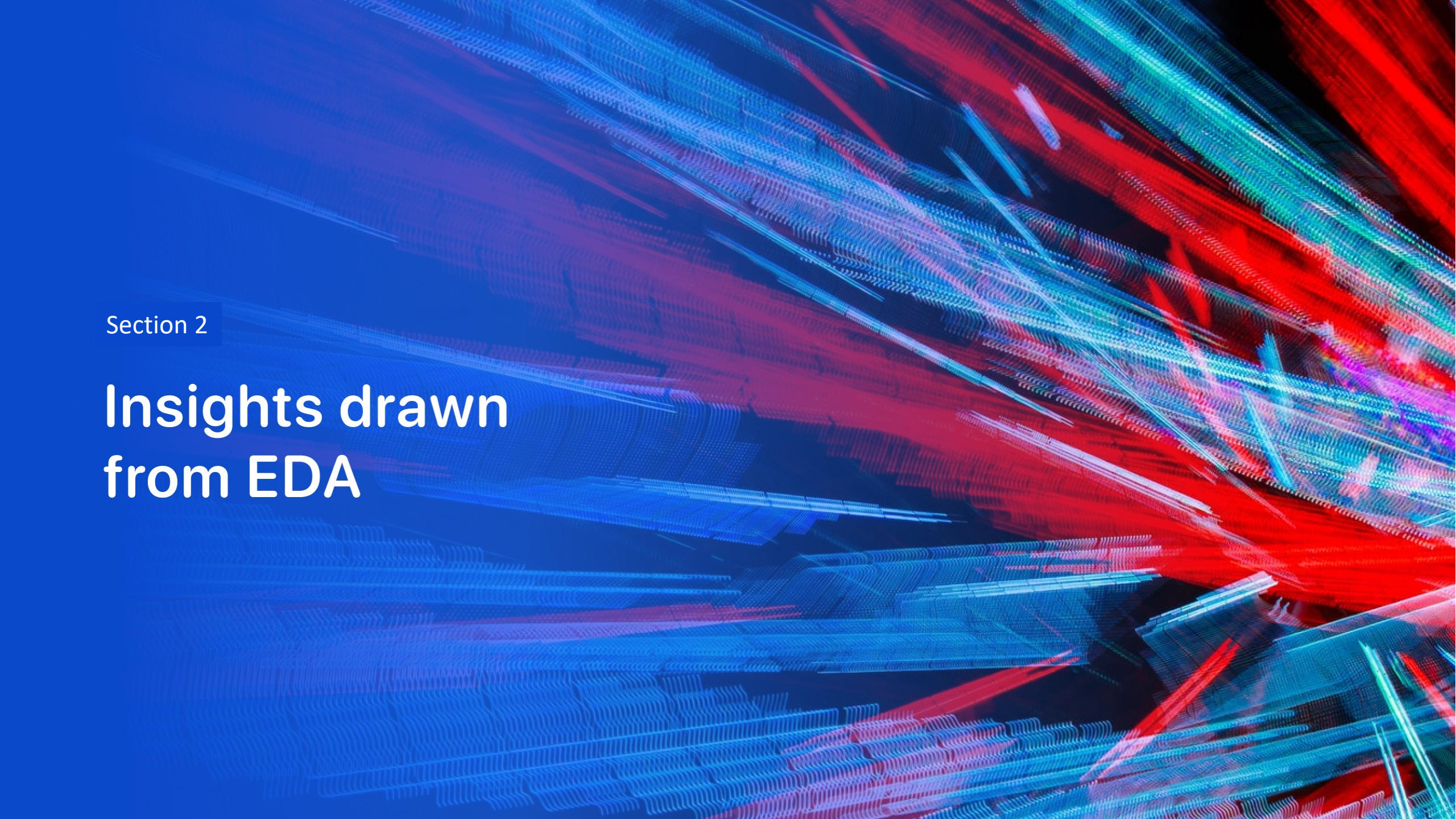
- Methods used: Support Vector Machine, Logistic Regression, K Nearest Neighbor, Decision Tree
- Objective: Perform exploratory data analysis and determine training labels, and test to find the method that performs best.



Completed deliverables: Link

Results Summary

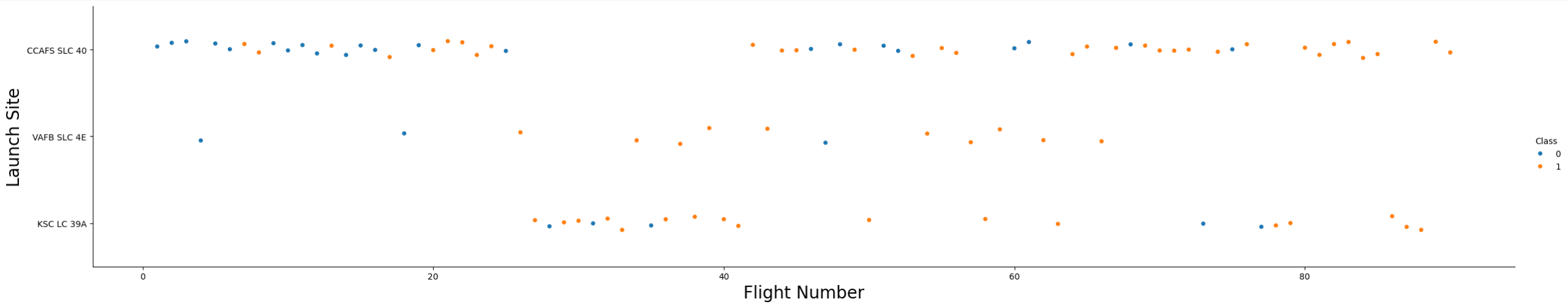
- Exploratory data analysis
  - Higher payload mass has higher success rate for Polar, LEO and ISS orbits
  - Launch success has improved over time
  - Launch site KSC LC-39A has the highest success rate
  - Launch site CCAPS SLC 40 has the most launches
  - Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Interactive analytics
  - Launch sites are close to the coast and near the equator
  - Launch sites are accessible for logistics needs but far from civil infrastructures to prevent damages
- Predictive analysis
  - The decision tree model provides highest classification accuracy


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Key findings

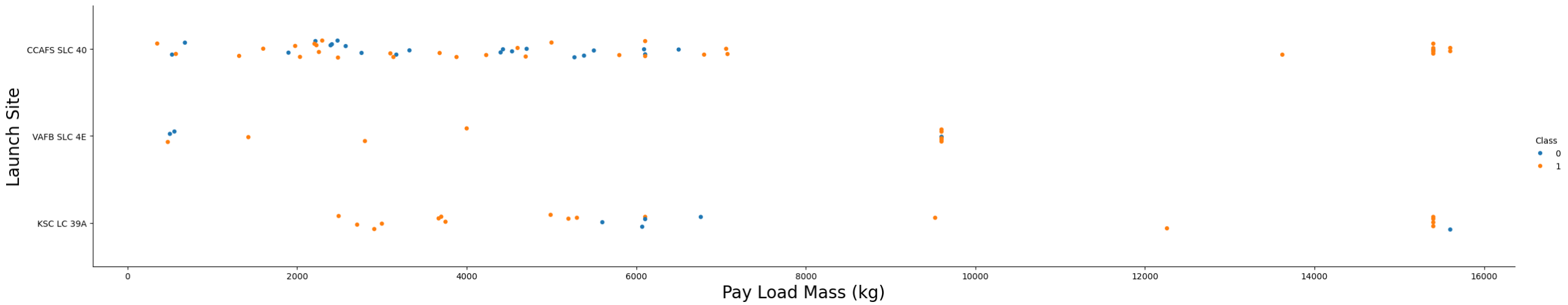
- CCAFS SLC 40 performed most(>50%) of launches
- KSC LC 39A followed by VAFB SLC 4E have higher success rates
- New launches have higher success rates



Payload vs. Launch Site

Key findings

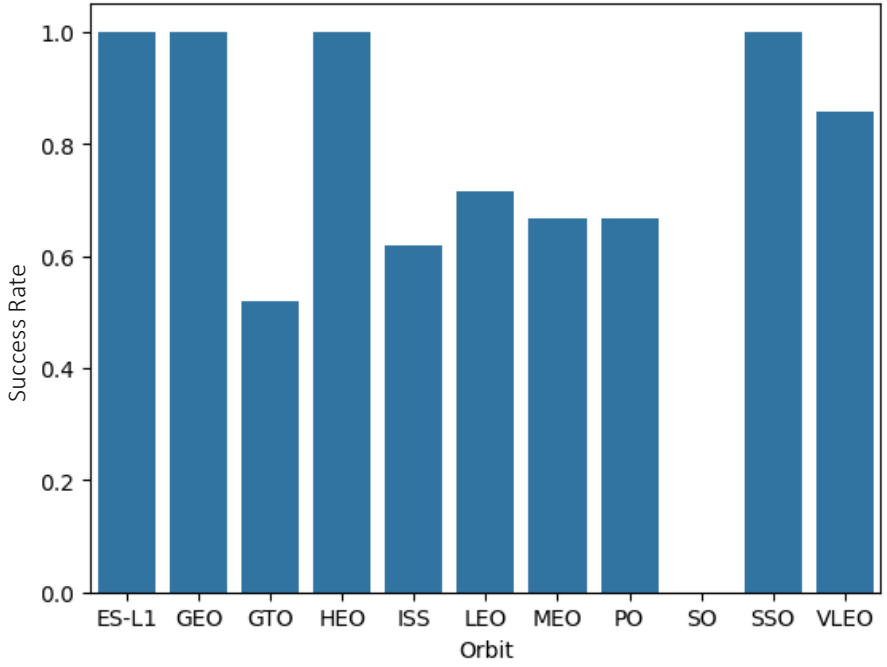
- The heavier the payload, the higher the success rate for CCAFS SLC 40
- VAFB SLC 4E has only launched <10,000 kg payload
- KSC LC 39A has succeeded in all launches <5,000 kg



Success Rate vs. Orbit Type

Key findings

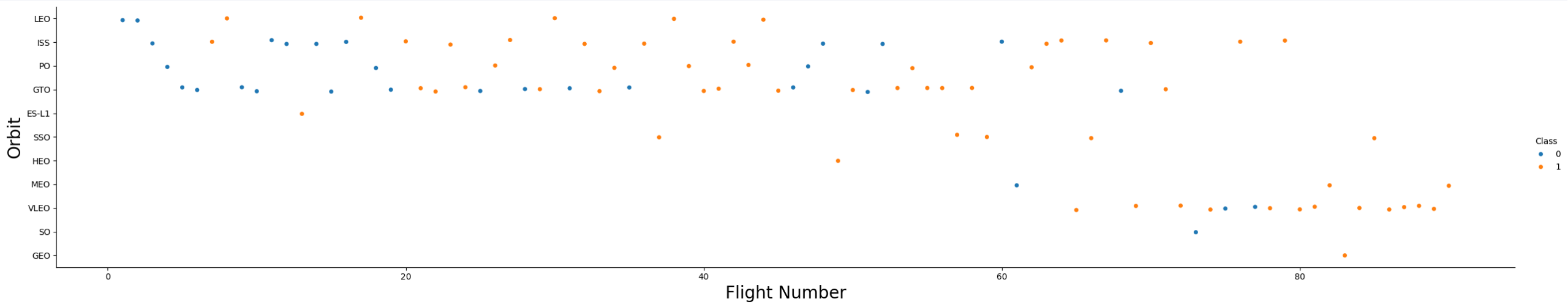
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Orbit SO has 0% success rate



Flight Number vs. Orbit Type

Key findings

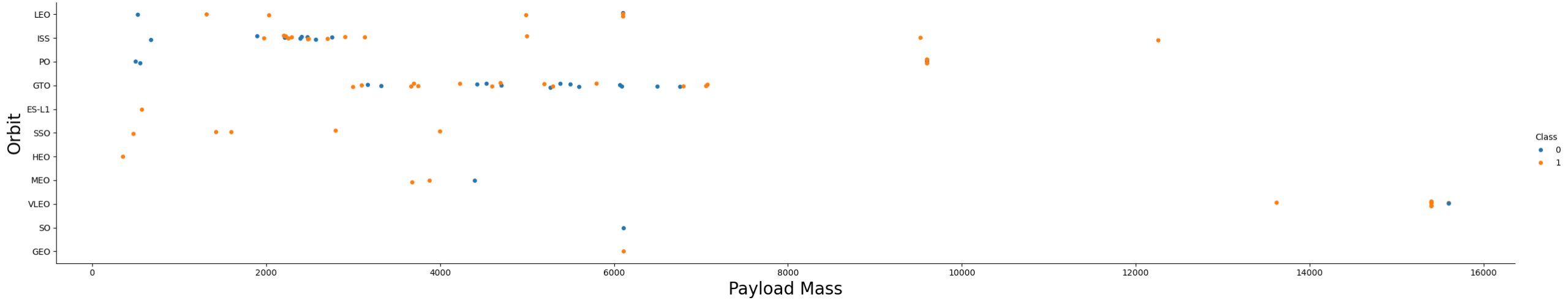
- The success rate is generally increasing in most of the orbits
- Orbit LEO is highly matched with the relationship stated above
- As of Orbit GTO, however, the relationship is not significant



Payload vs. Orbit Type

Key findings

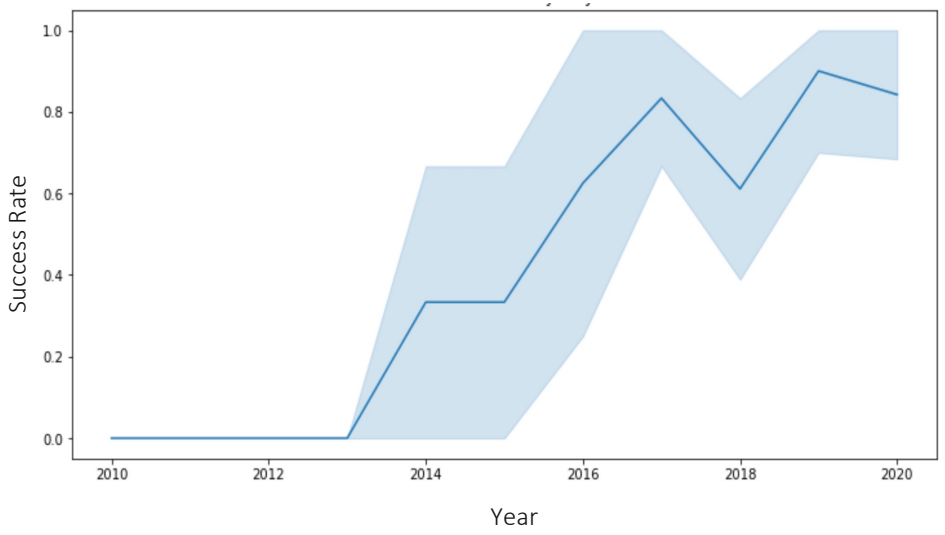
- With heavy payloads, the landings are more likely to succeed for PO, ISS and LEO orbits.



Launch Success Yearly Trend

Key findings

- Overall, the success rate has improved from 2013 to 2020



Launch Site Info Queries

All Launch Site Names

- DISTINCT is used to show unique values

```
%sql SELECT distinct(launch_site) FROM SPACEXTABLE
```

```
 * sqlite:///04_SpaceX_EDA_SQL_data.db
   sqlite:///db
Done.
```

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

Launch Site Names Begin with 'CCA’

- WHERE / LIKE is used to show only launch sites begin with 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	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Launch Site Info Queries

Total Payload Mass

- 45,596 kg in total carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_")
AS "Total payload mass carried by boosters launched by NASA (CRS)"
FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)'
```

Total payload mass carried by boosters launched by NASA (CRS)
45596

Average Payload Mass

- 2,928 kg in average carried by booster version F9 v1.1

```
%sql SELECT ROUND(AVG("PAYLOAD_MASS__KG_"),2)
AS "Average payload mass carried by booster version F9 v1.1"
FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

Average payload mass carried by booster version F9 v1.1
2534.67

Landing & Mission Outcomes

First Successful Ground Landing Date

- 22 December 2015
- WHERE is used to filter to “Success (ground pad)”

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (ground pad)'
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B1022 / F9 FT B1026 / F9 FT B1021.2 / F9 FT B1031.2
- WHERE & BETWEEN are used to filter the result

```
%sql SELECT Booster_Version FROM SPACEXTABLE
WHERE Landing_Outcome == 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight, 99 Success, 1 Success (payload status unclear)
- COUNT is used to provide total numbers of success/failure events

```
%sql SELECT Mission_Outcome, COUNT(*) AS "Total" FROM SPACEXTABLE GROUP BY Mission_Outcome
#%sql SELECT Mission_Outcome, COUNT(*) AS "SUCCESS COUNT" FROM SPACEXTABLE WHERE Mission_Ou
#%sql SELECT Mission_Outcome, COUNT(*) AS "FAILURE COUNT" FROM SPACEXTABLE WHERE Mission_Ou
```

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

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

Boosters Carried Maximum Payload

Key findings

- Result:

  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 in the WHERE clause and the MAX() function are used to determine the booster versions with maximum payload

```
%sql SELECT Booster_Version AS "Booster_versions carried Max payload mass" \
FROM SPACEXTABLE \
WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

Booster_versions carried Max payload mass
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

Key findings

- WHERE clause, along with LIKE, AND, and BETWEEN conditions are used to filter for failed landing outcomes in drone ship, their booster versions, and launch sites for the year 2015

```
%sql SELECT substr(Date, 6,2) AS "Month", Landing_Outcome, Booster_Version, Launch_Site \
    FROM SPACEXTABLE\
    WHERE substr(Date,0,5)='2015' AND Landing_Outcome == "Failure (drone ship)"
```

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

Key findings

- COUNT, WHERE, GROUP BY, ORDER BY are used to filter the dates and provide descending order of counts of landing outcomes

```
%sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE\
     WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" \
     GROUP BY Landing_Outcome\
     ORDER BY COUNT (*) desc
```

Landing_Outcome	COUNT(*)
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

All launch sites

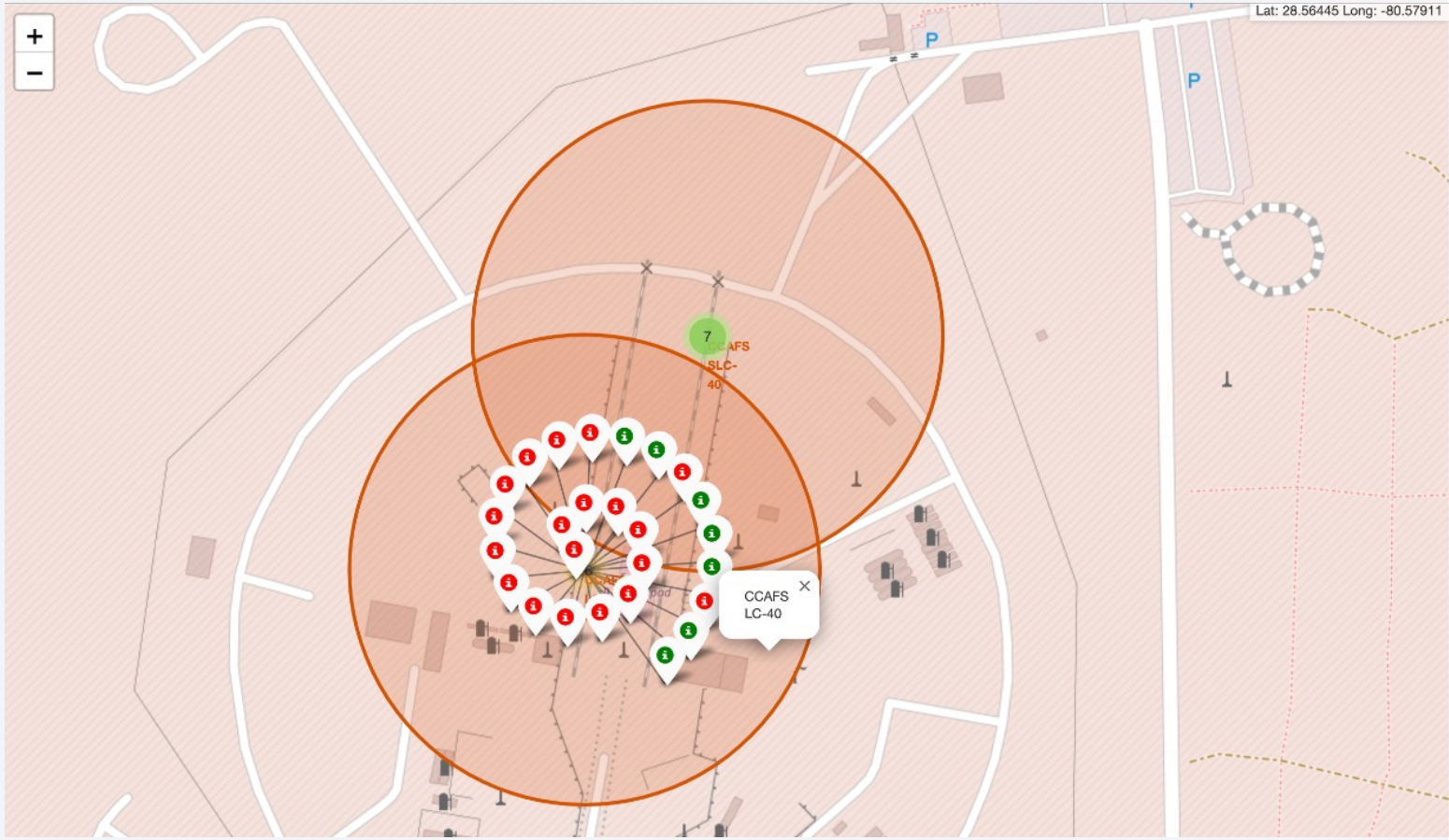
Key findings

- Launch sites are all close to Equator.
- Launch sites are all in the coastal areas.

Launch Outcomes

Key findings

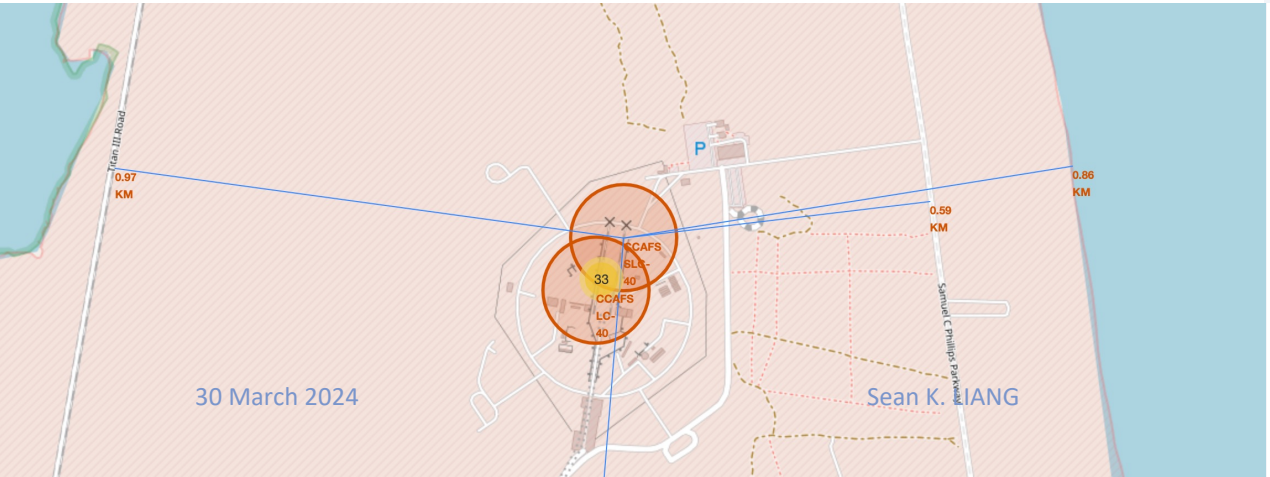
- Green/red markers makes it easy to show the launch outcomes of a certain site.
- The CCAFS LC-40 has 7 successful launches and 19 unsuccessful launches. (26.9% success rate)



Launch Outcomes

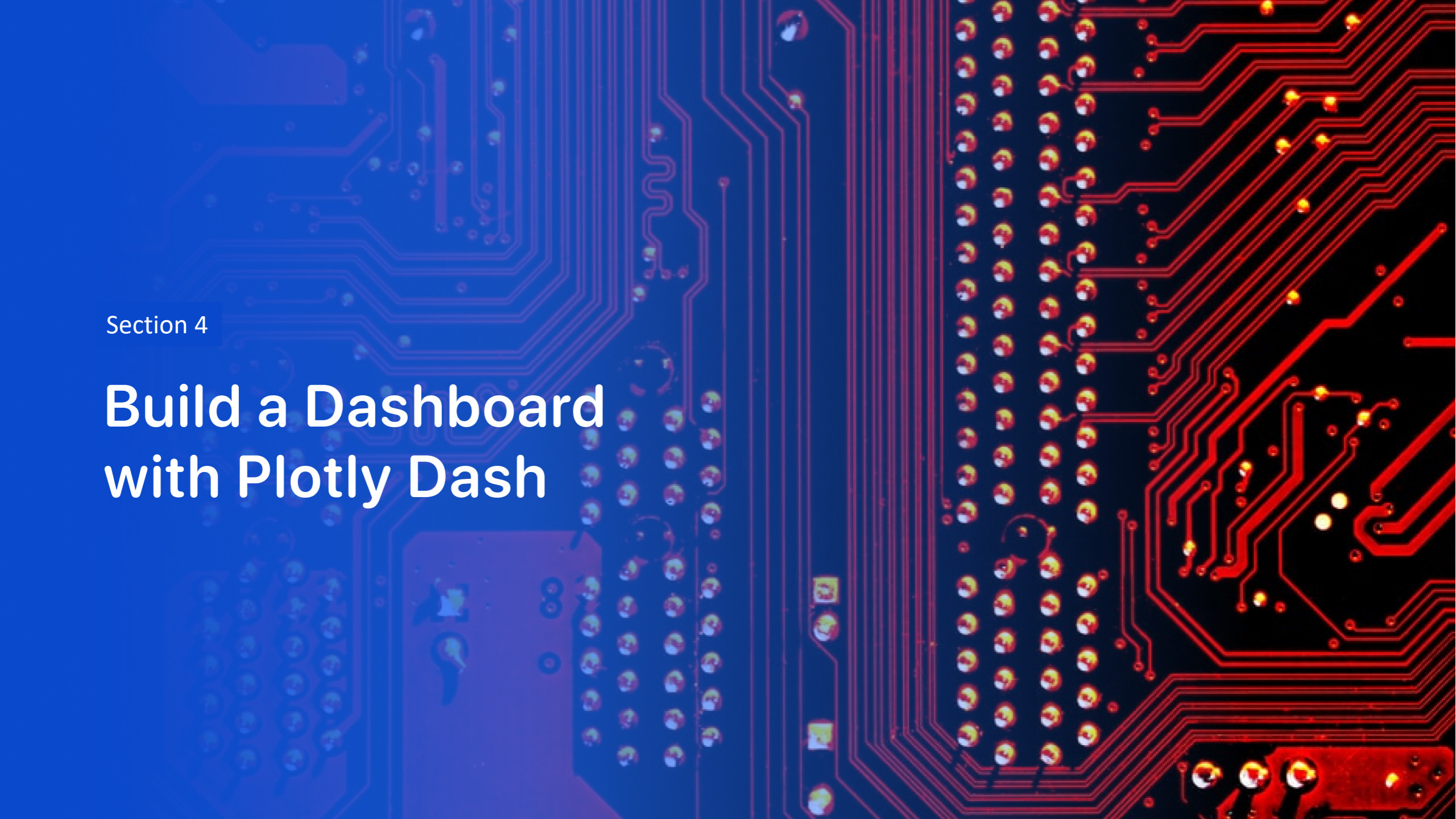
Key findings

- The launch site has a good accessibility regarding railway and highway connection in order to fulfill logistic needs.
- The launch site is closed to the coast to ensure the launch path is far from inhabited areas.



Key findings

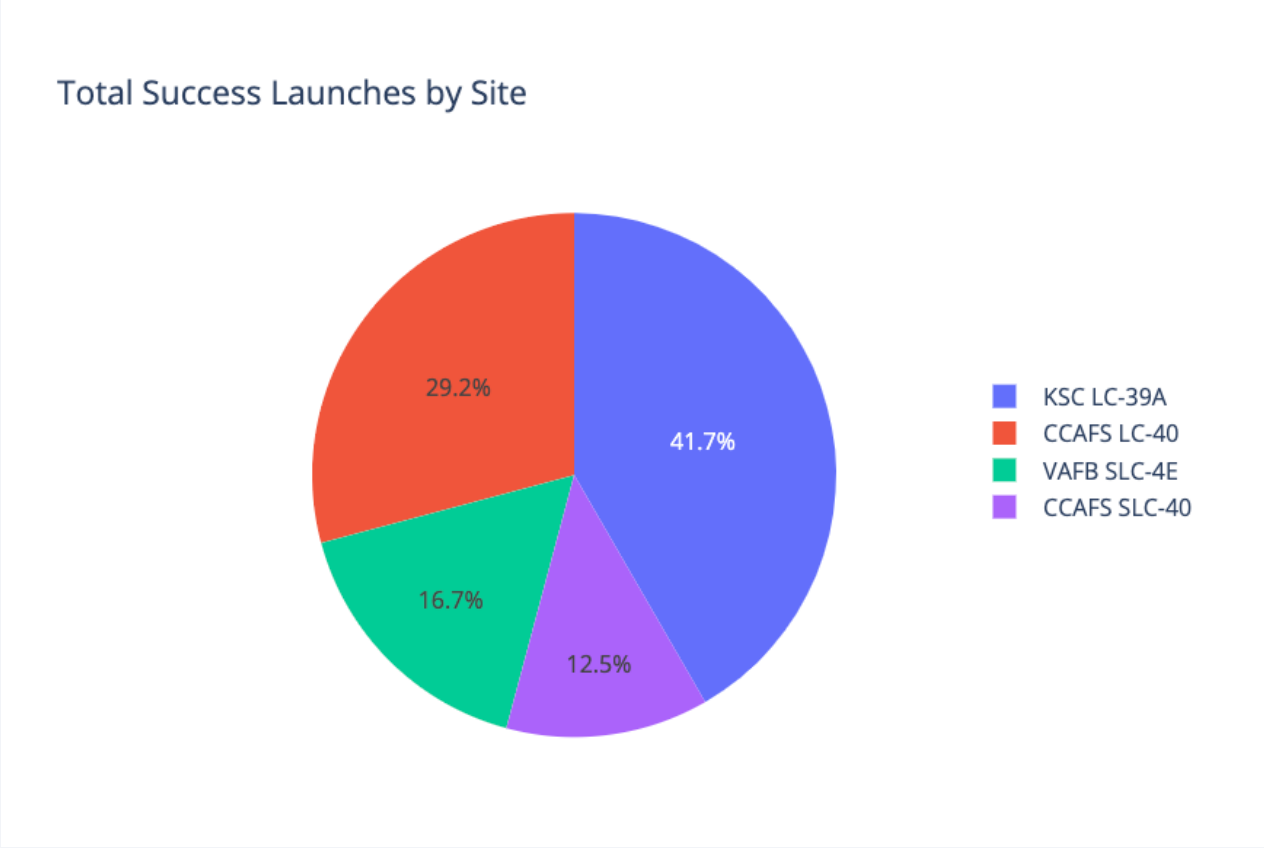
- The launch site is far from cities in order to prevent damages.


Section 4
Build a Dashboard with Plotly Dash

Success Launches by Launch Site

Key findings

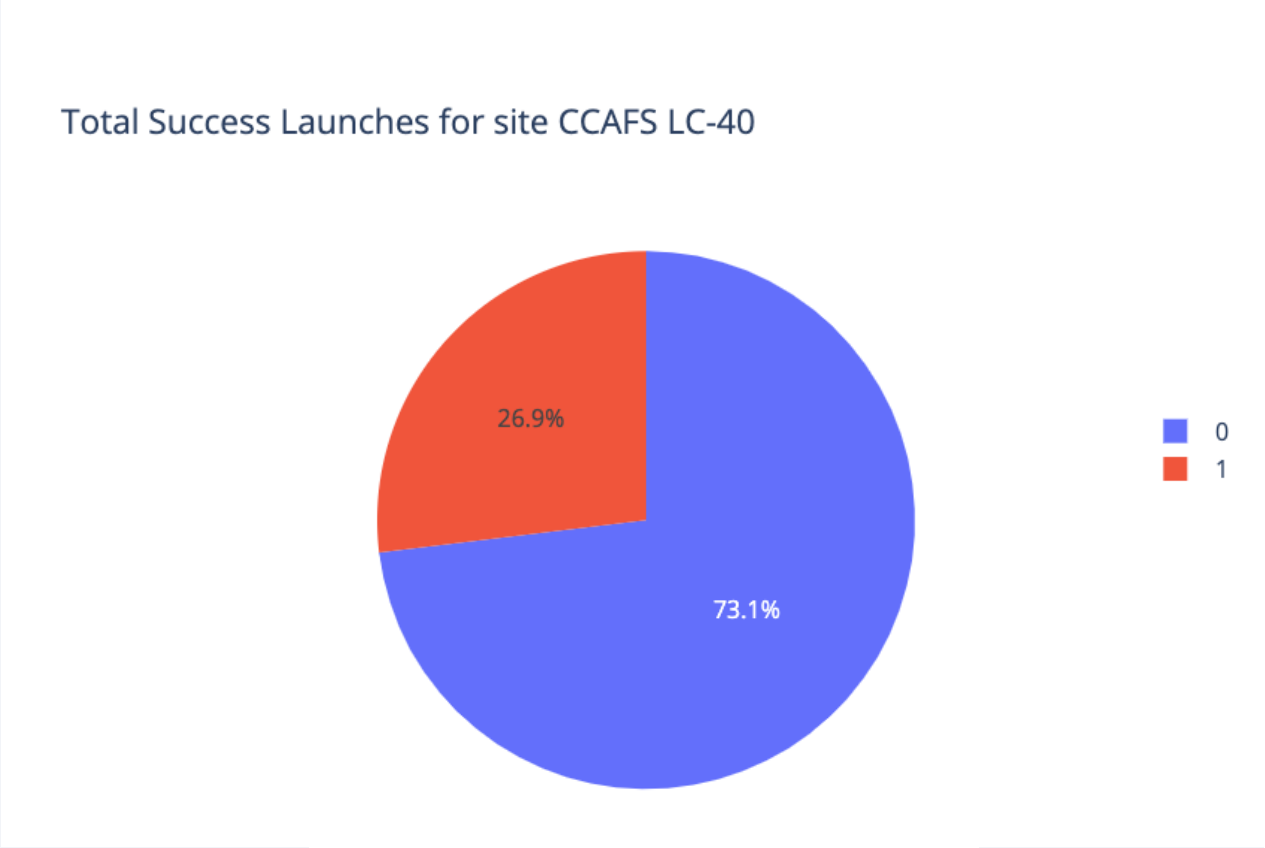
- KSC LC-39A has highest(41.7%) success rate among all sites.



Success Launches for CCAFS LC-40

Key findings

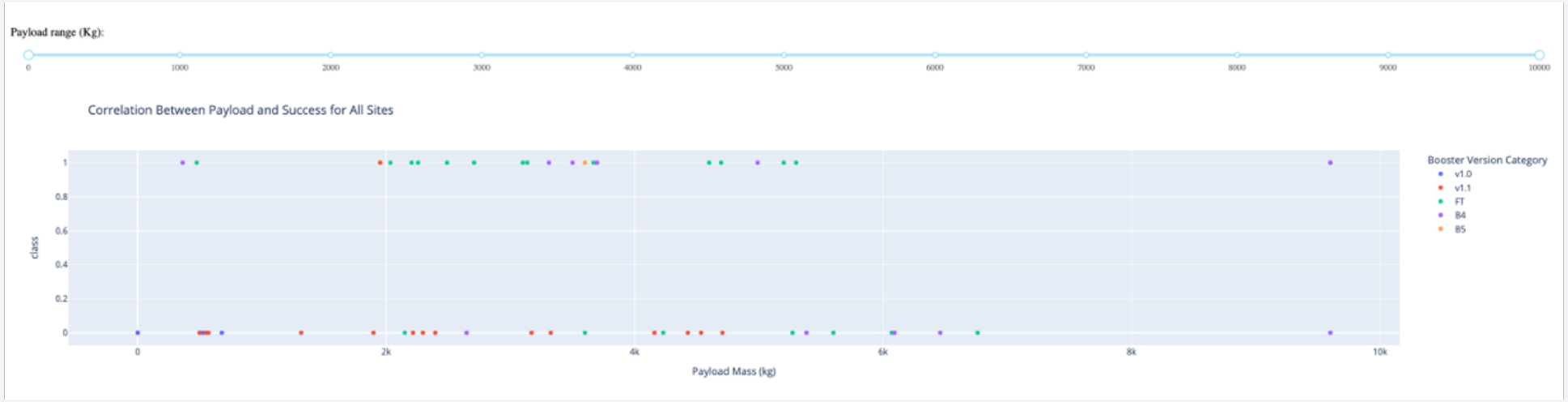
- CCAFS LC-40 has 26.9% successful launches against 73.1% unsuccessful launches.



Success Launches with Payloads

Key findings

- Success launches are highly concentrated with payloads between 2,000 kg and 5,000 kg



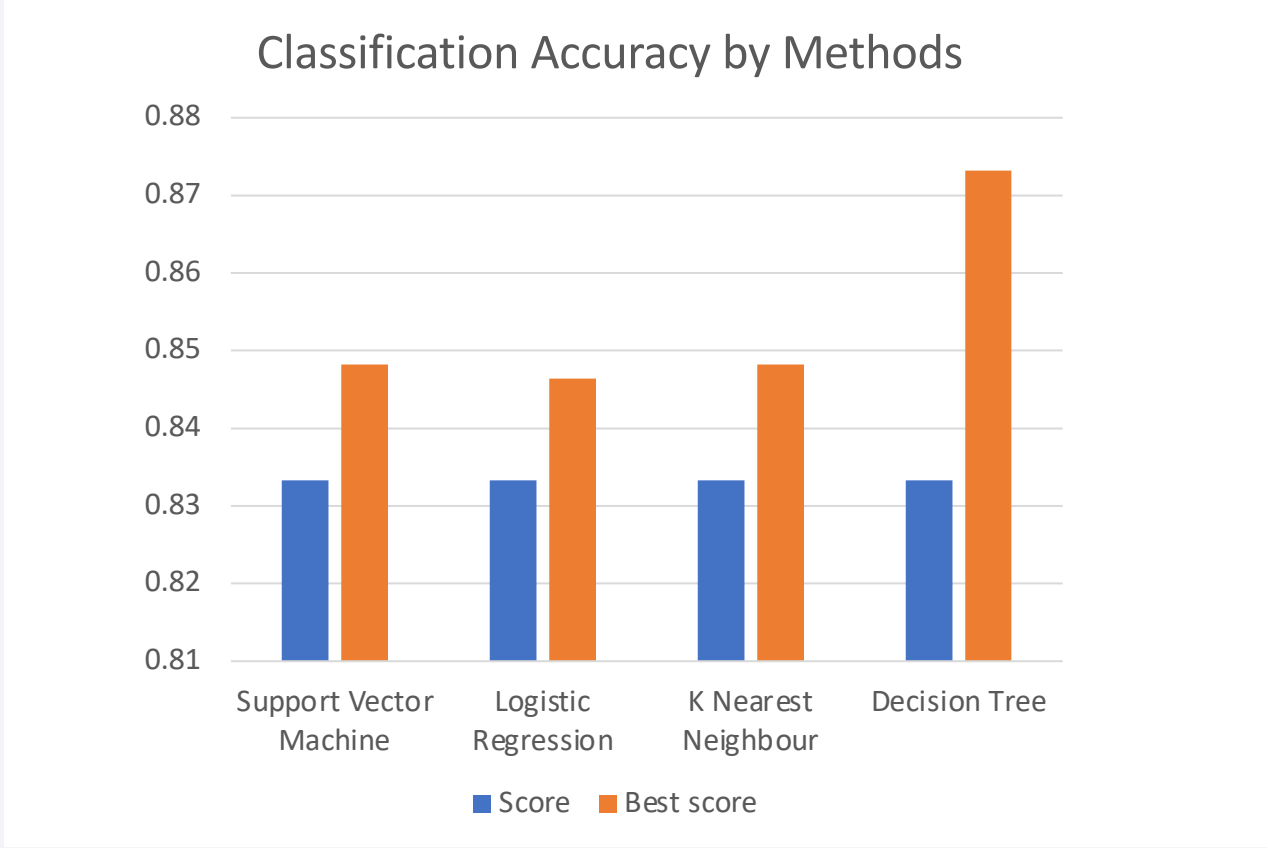
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Key findings

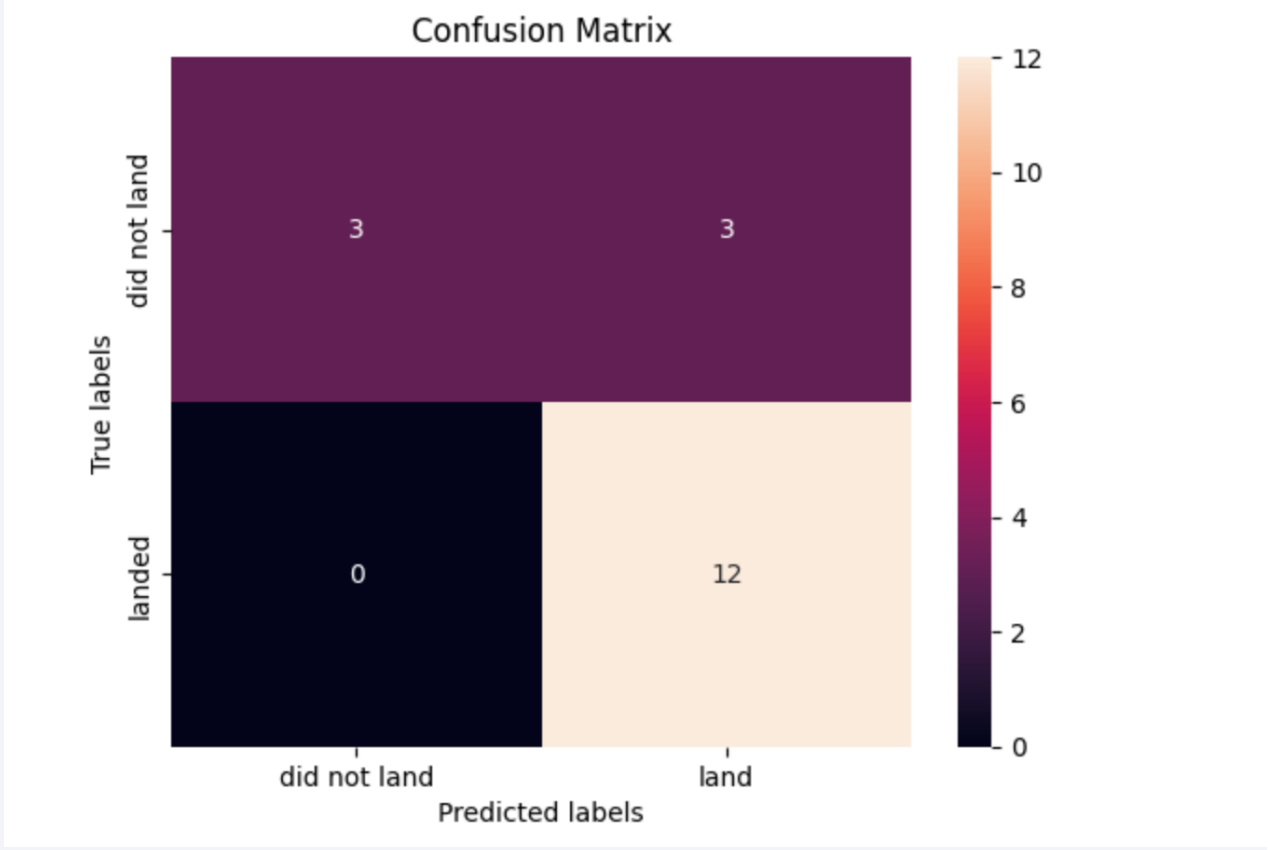
- All the methods performed have the same scores and accuracy.
- However, the Decision Tree model slightly outperformed the rest when looking at their best scores.



Confusion Matrix

Key findings

- All the confusion matrices are the same
- The false positives (Type 1 error) exists, which should be further focused on and lower down



Conclusion

- Higher payload mass has higher success rate for Polar, LEO and ISS orbits.
- Launch success rate increases over time (from 2013 to 2020).
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- KSC LC-39A has the highest success rate among launch sites.
- The models performed similarly with the decision tree model slightly outperforming
- All launch sites are closed to coastal area and equator.

Thank you!