IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Boyang Chen
01/11/2021

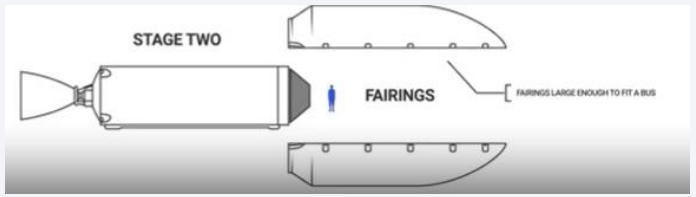
Outline

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

Executive Summary

- Summary of methodologies
  - Data collection – API request and webscraping
  - Data wrangling – data cleaning, transformation using Python
  - Exploratory data analysis – Matplotlib to create scatter, bar and line plots. Use SQL to query to gain insights on the data
  - Interactive data analysis – Plotly Dashboard and Folium maps to visualize launch site data.
  - Machine Learning – using classification models to predict the landing success of the first stage boosters.
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction



- Falcon 9 is a partially reusable two-stage-to-orbit medium-lift launch vehicle designed and manufactured by SpaceX.
- The first stage of the Falcon 9 is capable of re-entering the atmosphere and landing vertically This feat was achieved for the first time on flight 20 in December 2015.
- SpaceX Falcon 9 rocket launches cost around 62 million dollars. This is relatively inexpensive compared to the costs from other rocket companies (upto $165 million).
- Much of this difference is because SpaceX can reuse the first stage, with individual first stages flying as many as ten times.

- The data science problem:
- Predict the cost of a rocket launch.
- Predict the landing success rate of the first stage using machine learning model.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting data from the SpaceX API using get.requests.
  - Webscraping Falcon 9 launch data from Wikipedia page using BeautifulSoup.
- Perform data wrangling
  - Data was converted from JSON to data frame format
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

HTTP Request

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

Client

SpaceX API URL

Result

```
response.status_code
200
```

Response.Json()

Convert to data frame

```
data = pd.json_normalize(response.json())
```

Repeat this for:
BoosterVersion data
LaunchSite data
Payload data
Core data

Click here: Github notebook

Data Collection - Scraping

HTTP Request

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
response = requests.get(static_url).text
```

Client

Falcon9 Launch HTML

Create BeautifulSoup object

```
soup = BeautifulSoup(response, 'html5lib')
```

```
html_tables= soup.find_all('table')
```

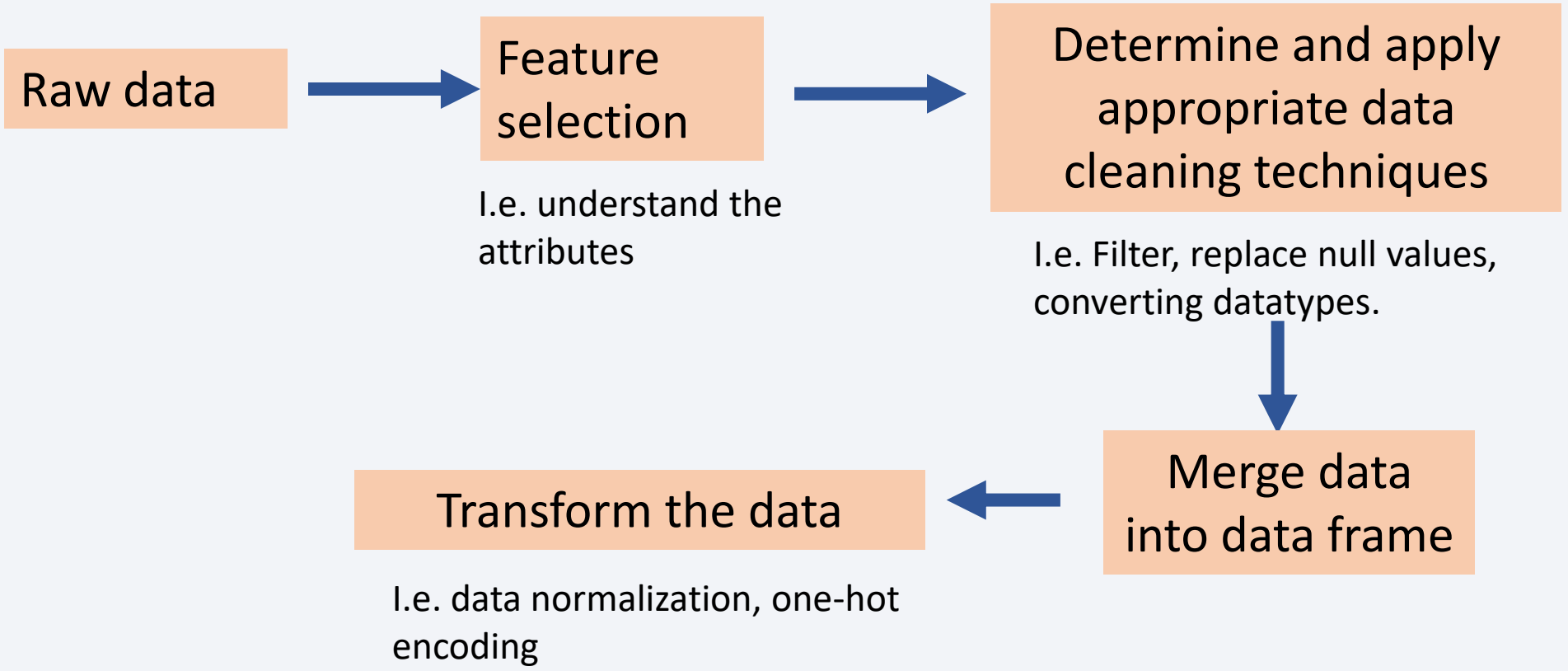
Extract column names

```
for th in first_launch_table.find_all('th'):
    column = extract_column_from_header(th)
    if  column is not None and len(column) > 0:
        column_names.append(column)
```

Parse through the html tables and create data frame

Click here: Github link

Data Wrangling



Click here: Github link

EDA with Data Visualization

- Scatter plots – show the relationship between two variables.
- Bar charts – compare categorical data
- Line plots – show trends over time.

Click here: Github link

EDA with SQL

SQL queries performed:

Click here: Github link

- Unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

Rail

- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Total number of successful and failure mission outcomes
- names of the booster_versions which have carried the maximum payload mass.
- failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- landing outcomes between the date 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

- Map objects
  - Circles – highlight each launch site
  - Markers – colored markers to display launch outcomes.
  - Lines – to indicate the distance between launch sites and proximities.

Click here: Github link

Build a Dashboard with Plotly Dash

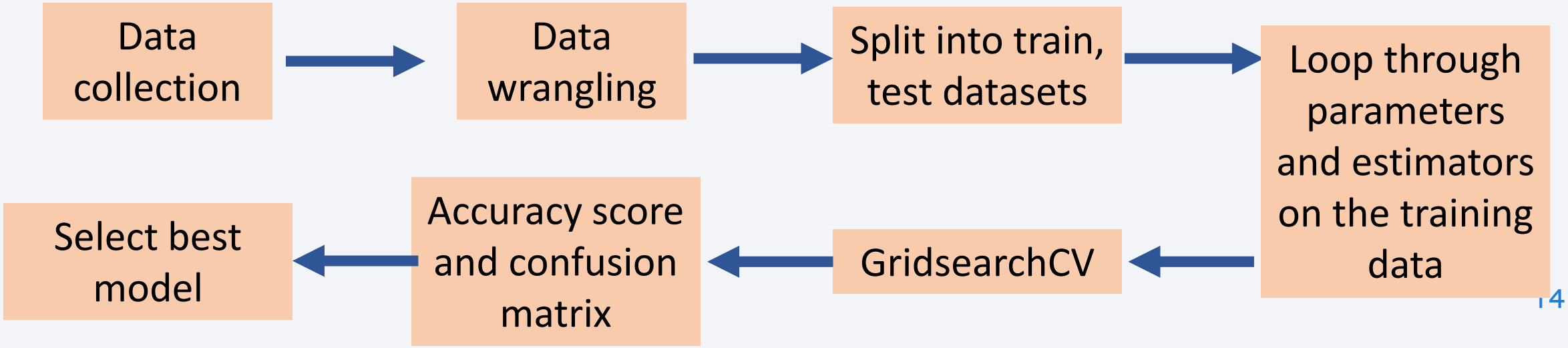
- Dashboard features
  - Drop down menu for each launch site
    - Allows the user to select the launch sites of interest
  - Pie chart showing the total successful launches for each launch site
    - Allows the user to visualize the proportion of successful and unsuccessful launches at each site
  - Range slider for payload mass
    - Allows the user to select specific payload mass ranges for comparison.
  - Scatter plot showing the correlation between payload mass and success rates.
    - Allows the user to visualize the relationship between payload mass and success rates at specific launch sites.

Click here: Github link

Predictive Analysis (Classification)

- Data was converted into a dataframe, features and class were assigned X and Y variable, respectively.
- X data was normalized using standardScaler
- Data was split into training and test datasets.
- Used GridSearchCV to find the best parameters for each model.
- Evaluate model using accuracy score and confusion matrix.

Click here: Github link

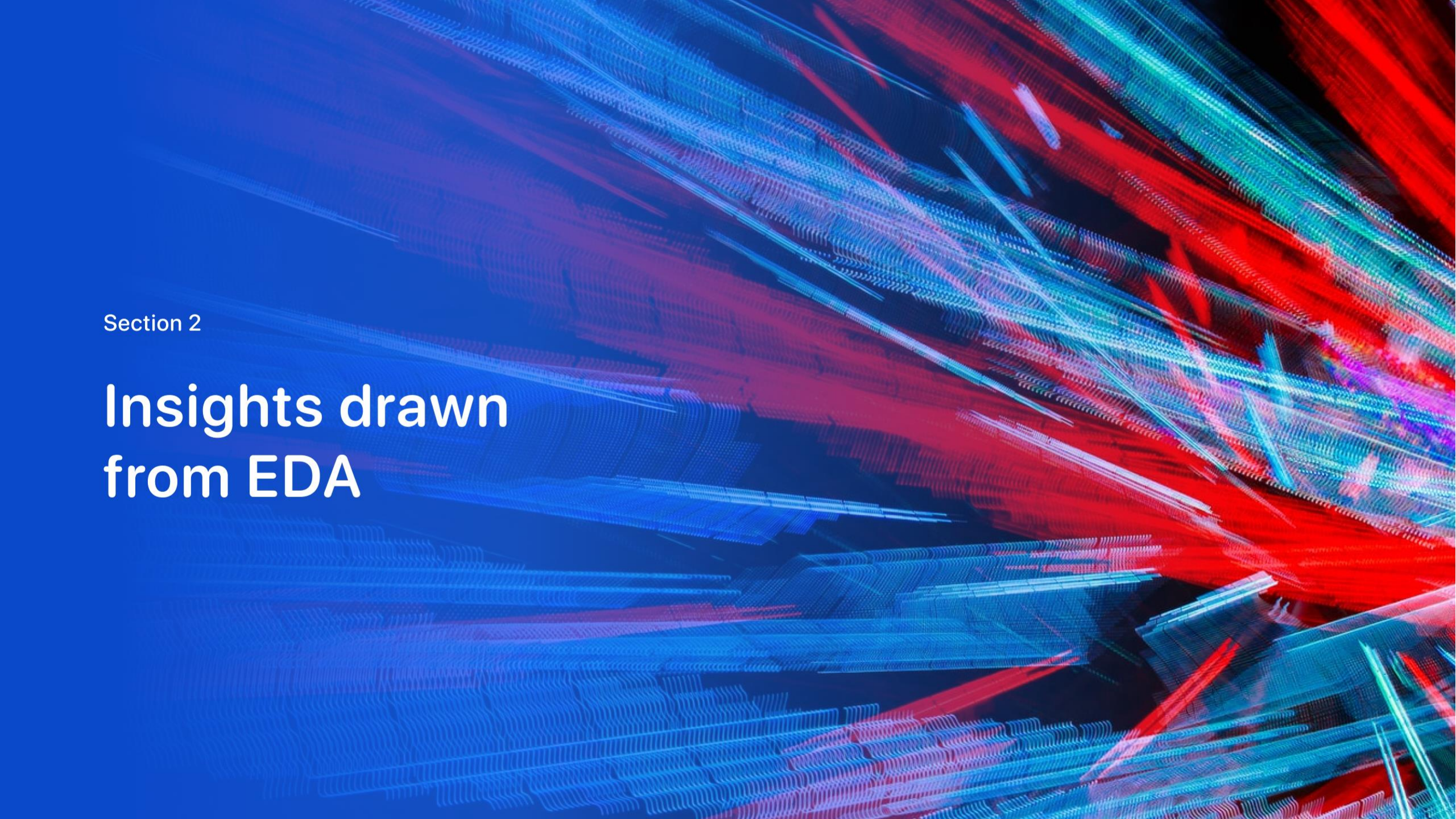


Results

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

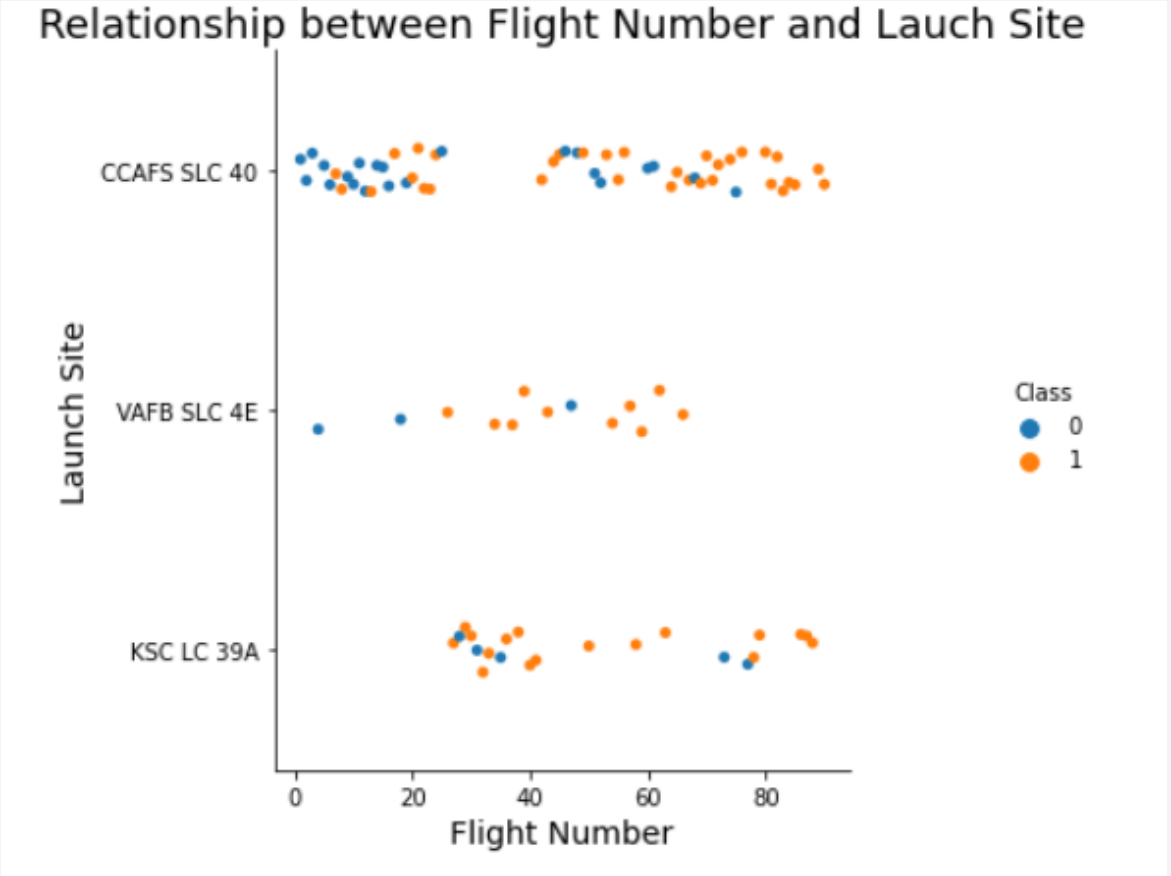
Section 2

Insights drawn from EDA



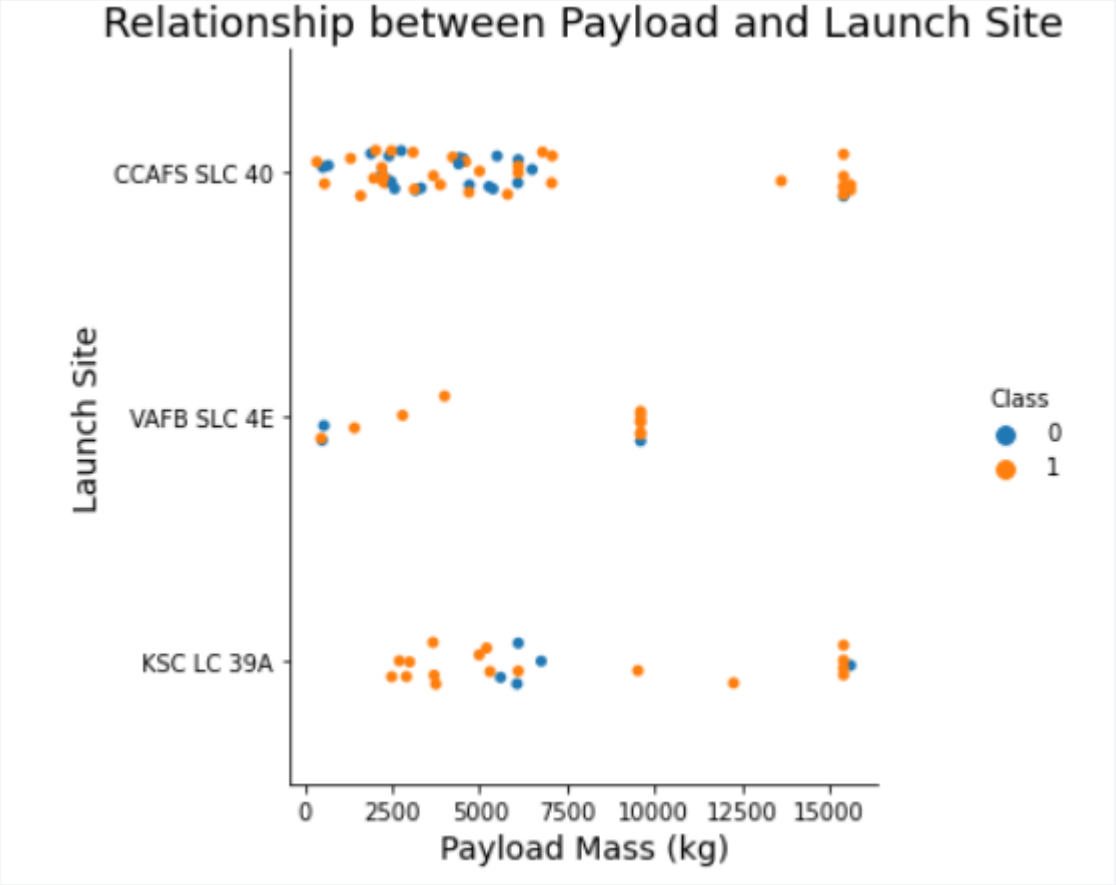
Flight Number vs. Launch Site

- Launch site CCAFS SLC-40 has the highest number of launches.
- CCAFS SLC-40 has been in operation longer than the other two sites.
- Number of failed launches decreased with the increase in flight number.



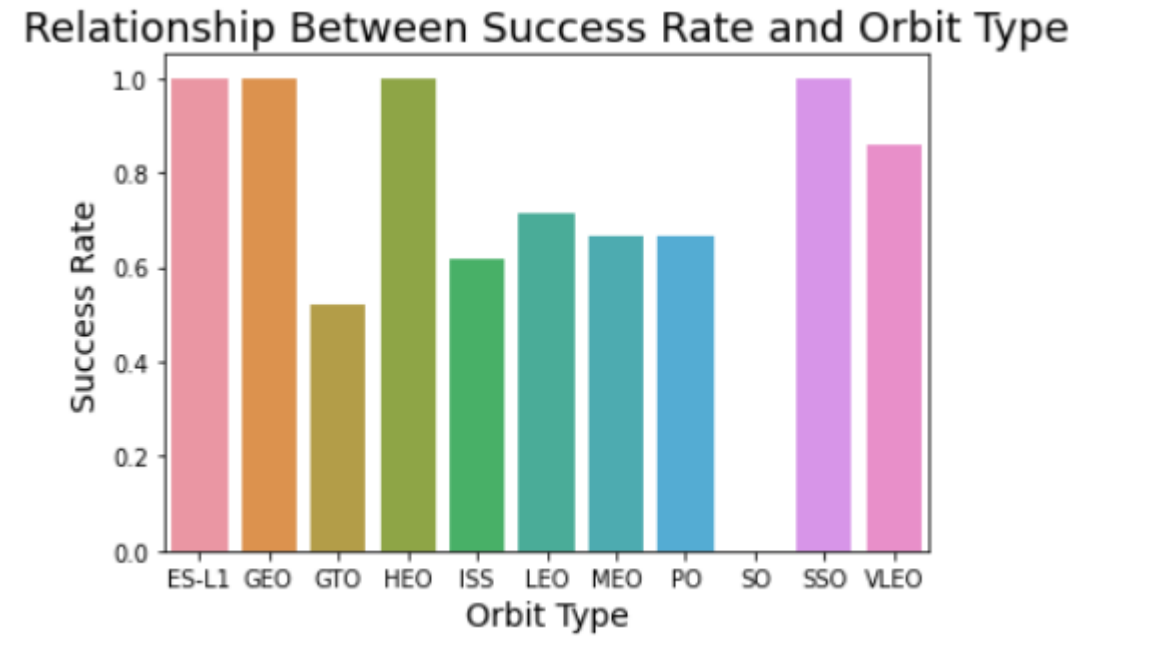
Payload vs. Launch Site

- CCAFS has a large launches under 7500 kg of payload.
- CCAFS and KSC have launched 15000 kg payloads.
- Maximum payload mass at VAFB is 10000 kg.
- More unsuccessful launches observed at lower payload masses.



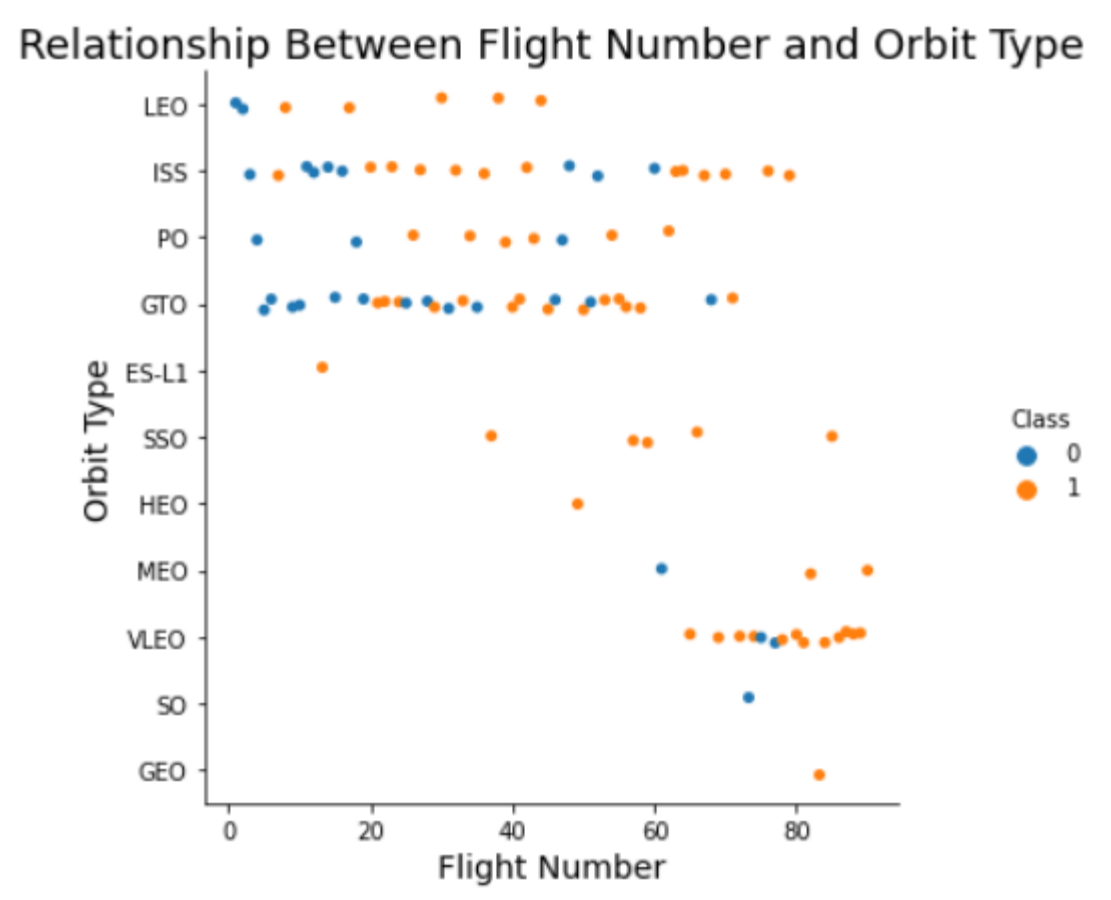
Success Rate vs. Orbit Type

- Orbit types ES-L1, Geostationary orbit (GEO) , High earth orbit (HEO) and Polar orbit and Sun-synchronous orbit (SSO) all have 100% success rate.
- Orbit type SO have 0% success rate (only 1 sample size)



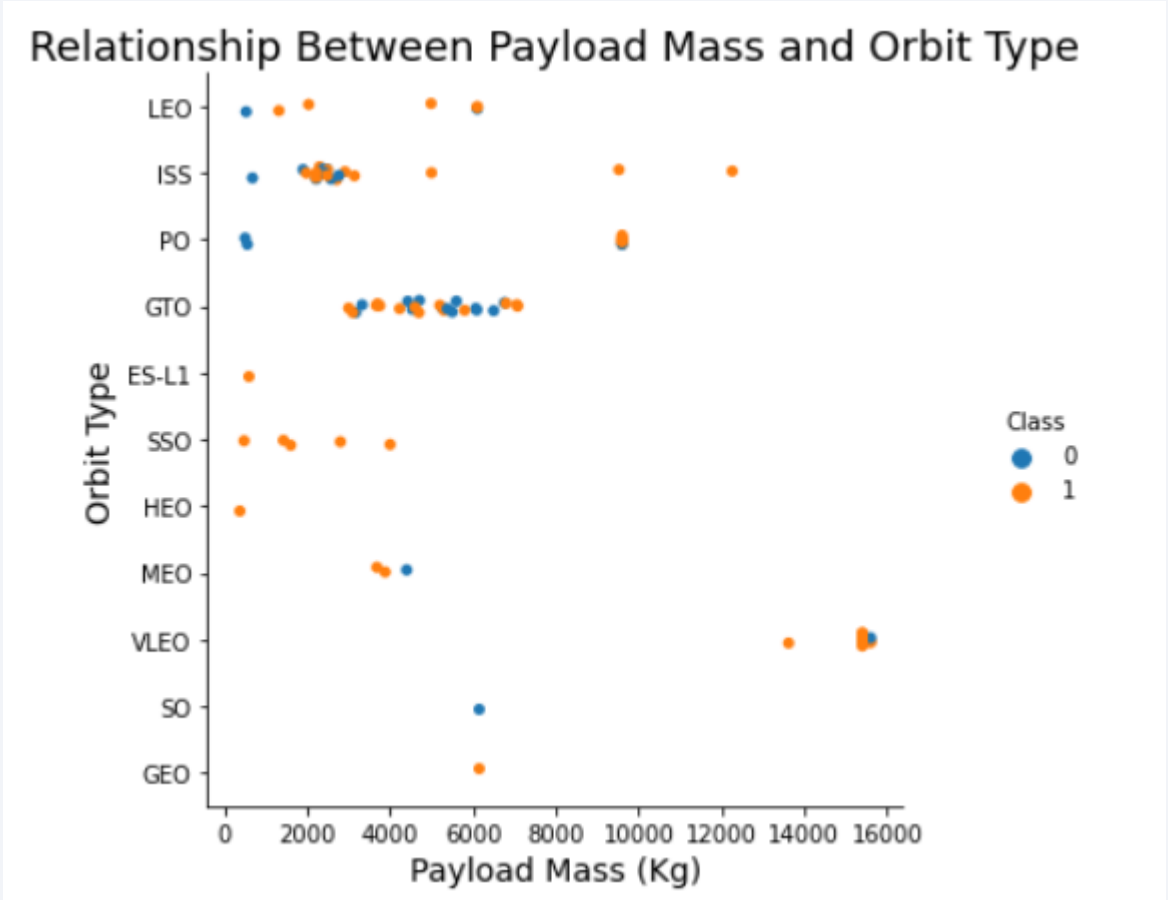
Flight Number vs. Orbit Type

- Earlier on the most popular orbit types were LEO, ISS, PO and GTO.
- Success rate increased with flight number for all orbit types.
- Orbit types SO, GEO, ES-L1 and HEO have only 1 data point.



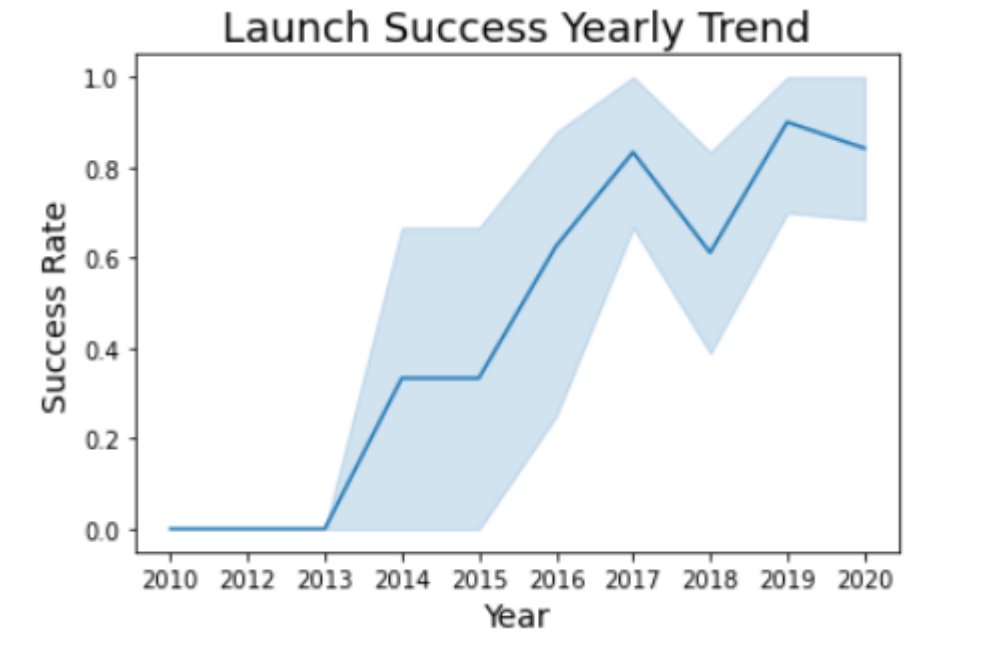
Payload vs. Orbit Type

- For the majority of orbit types, the payload mass is lower than 8000 kg.
- At a payload mass of 16000 kg, VLEO orbit is recommended.



Launch Success Yearly Trend

- Average success rate has been trending upwards from 2013 to 2020
- There is a dip in 2018.



All Launch Site Names

- There are four unique launch sites
- select DISTINCT launch_site
- FROM SPACEXDATASET

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

Launch Site Names Begin with 'CCA'

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

- This table shows the five record where launch site names beginning with ‘CCA’.
- SELECT *
- FROM SPACEXDATASET
- WHERE launch_site LIKE 'CCA%’
- LIMIT 5

Total Payload Mass

total_payload_mass
45596

- The total payload mass (kg) carried by boosters launched by NASA (CRS).
- SELECT SUM(payload_mass__kg_)
- FROM SPACEXDATASET
- WHERE customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

avg_payload_mass
2928

- SELECT AVG(payload_mass__kg_)
- FROM SPACEXDATASET
- WHERE booster_version = 'F9 v1.1'

First Successful Ground Landing Date

first_successful_landing
2015-12-22

- SELECT MIN(DATE) AS First_Successful_Landing
- FROM SPACEXDATASET
- WHERE landing__outcome = 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Four boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- SELECT booster_version
- WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000

Total Number of Successful and Failure Mission Outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SELECT mission_outcome, COUNT(mission_outcome) AS total_number
- FROM SPACEXDATASET
- GROUP BY mission_outcome;

Boosters Carried Maximum Payload

- These boosters have carried the maximum payload.
- SELECT booster_version
- FROM SPACEXDATASET
- WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXDATASET)

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

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

- These two boosters failed in 2015.
- SELECT booster_version, landing__outcome, launch_site
- FROM SPACEXDATASET
- WHERE landing__outcome ='Failure (drone ship)' AND YEAR(DATE) = 2015

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

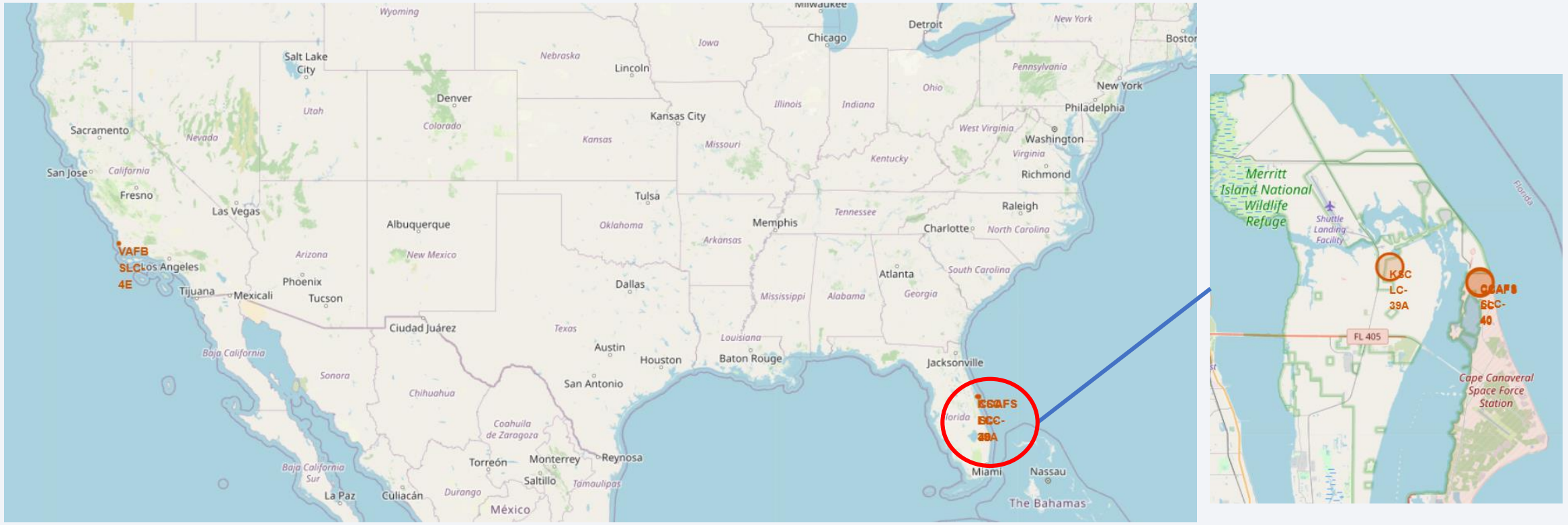
- Of the 11 drone ship landing attempts, 5 of the drone ship landings were successful.
- 3 successful ground pad landings
- 2 failed parachute landings.

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

Section 4

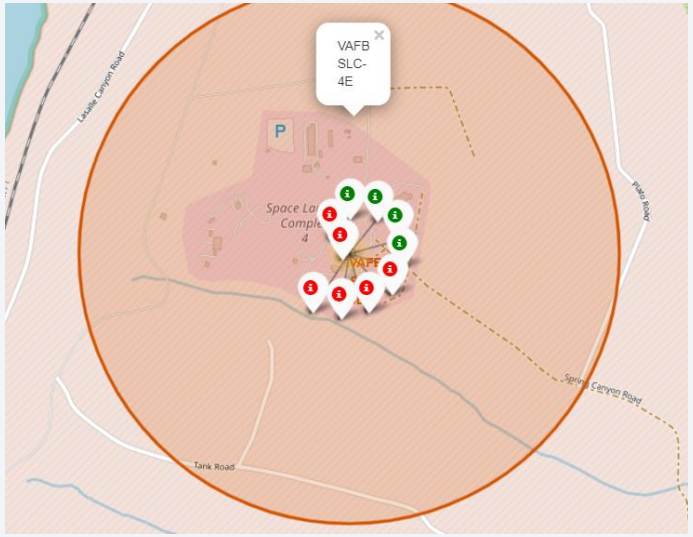
Launch Sites Proximities Analysis

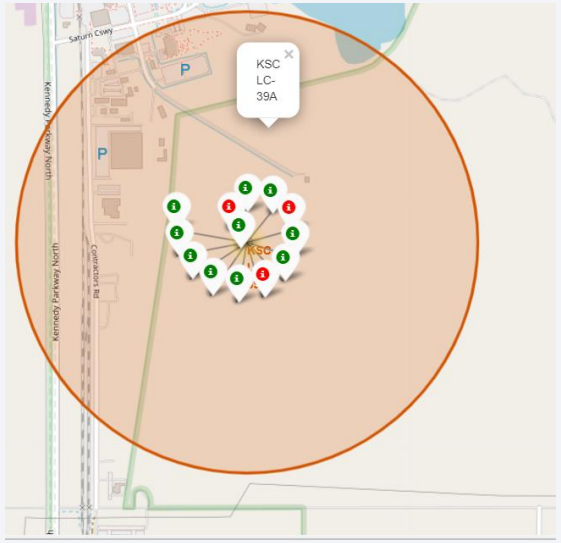
Locations of all launch sites

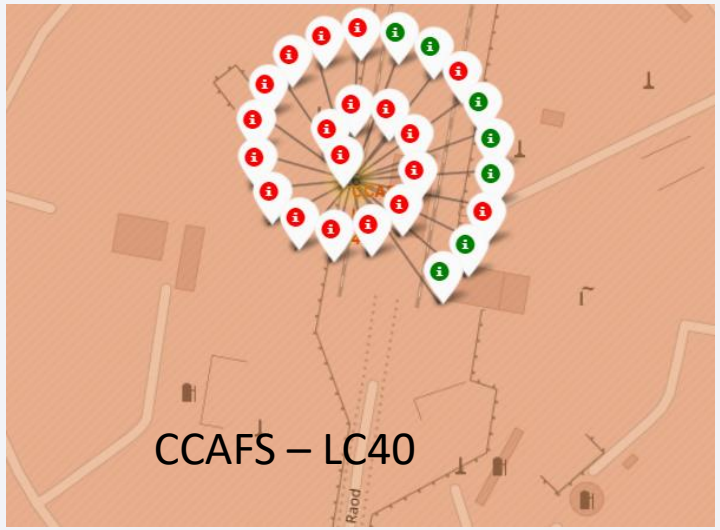


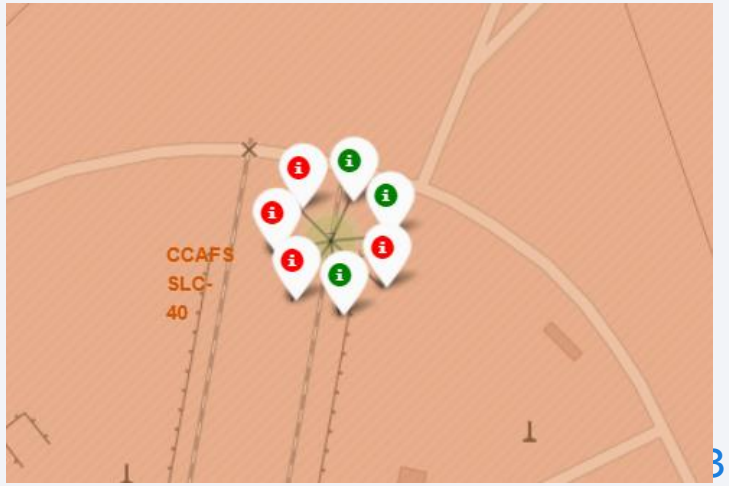
Launch site VAFB SLC-4E is located on the west coast, near LA.
Launch sites CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A is located on the east coast, near Orlando.

Launch outcomes at each launch site









Launch site KSC LC-39A has a high launch success rate (76%).
Launch site CCAFS-LC40 has the lowest success rate (27%).
Launch sites VAFB SLC-4E and CCAFS SLC-40 has 40% and 43% success rates respectively.

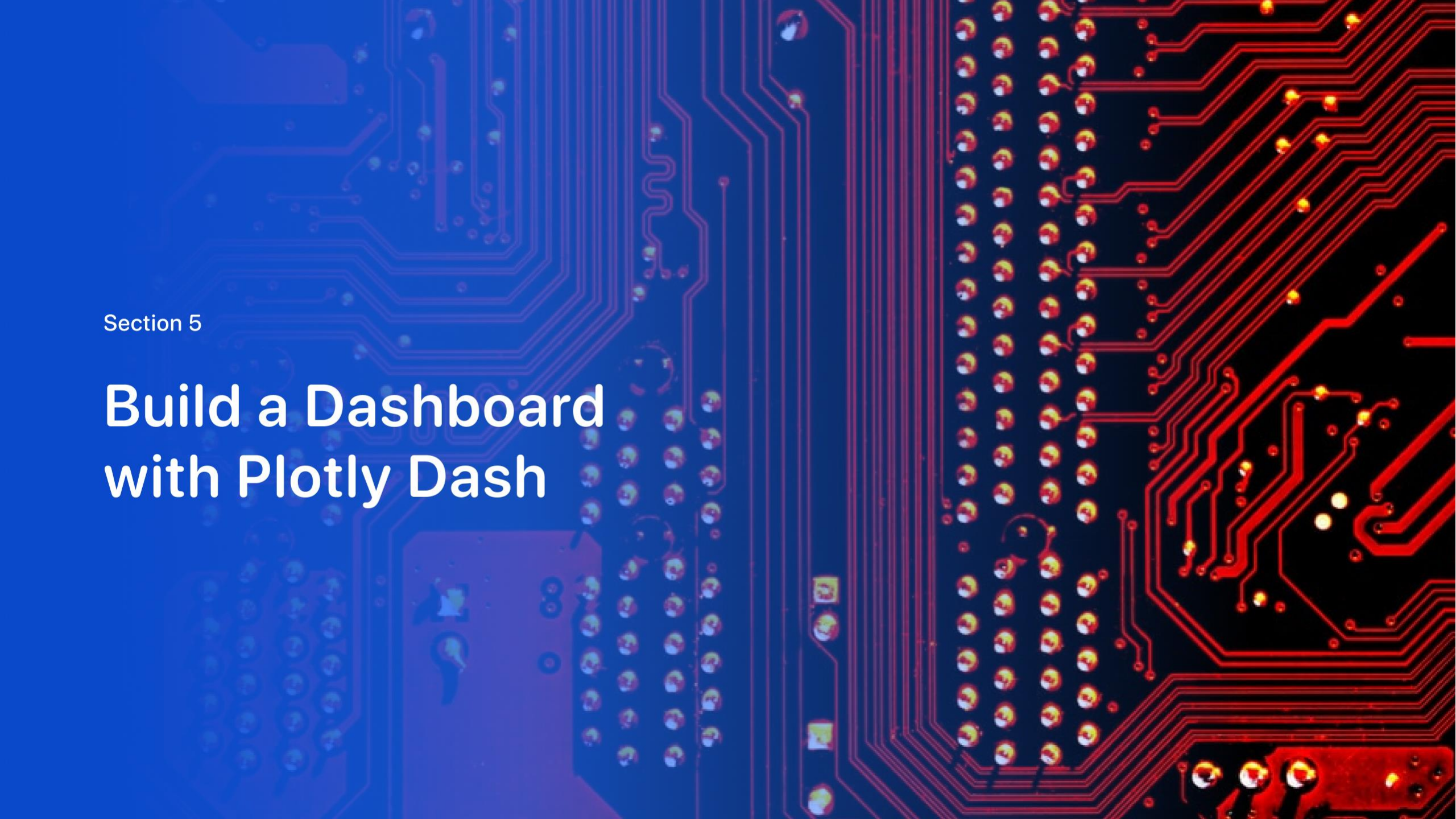
Launch site proximities

Launch sites are positioned close to transportation infrastructure such as railways, highways and coastlines. This makes it easier to bring people and materials to the site.

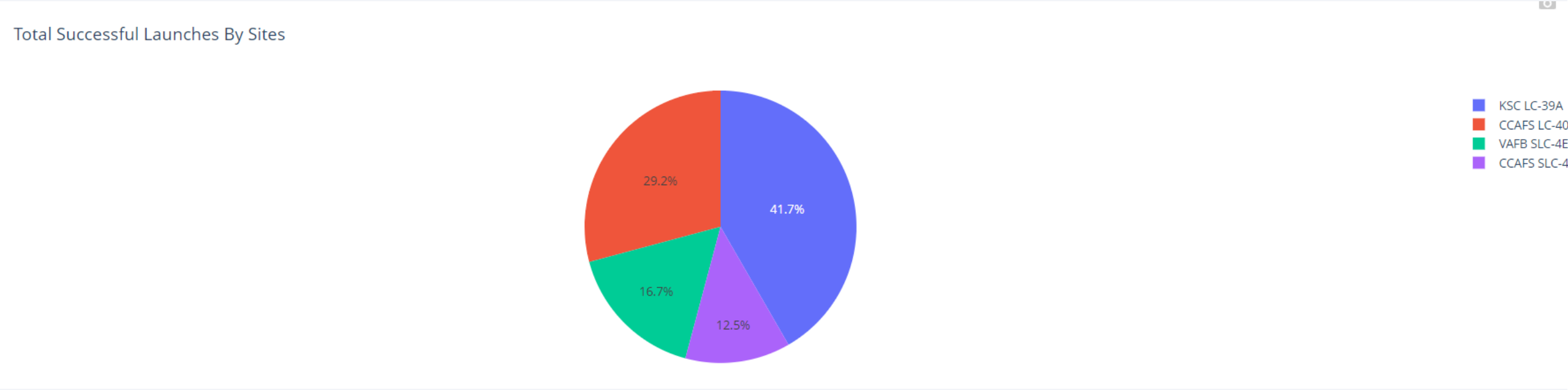
Launch sites are positioned far away from cities and residential areas. During a launch, stages are dropped along the launch path and must be kept away from people or properties.

Section 5

Build a Dashboard with Plotly Dash

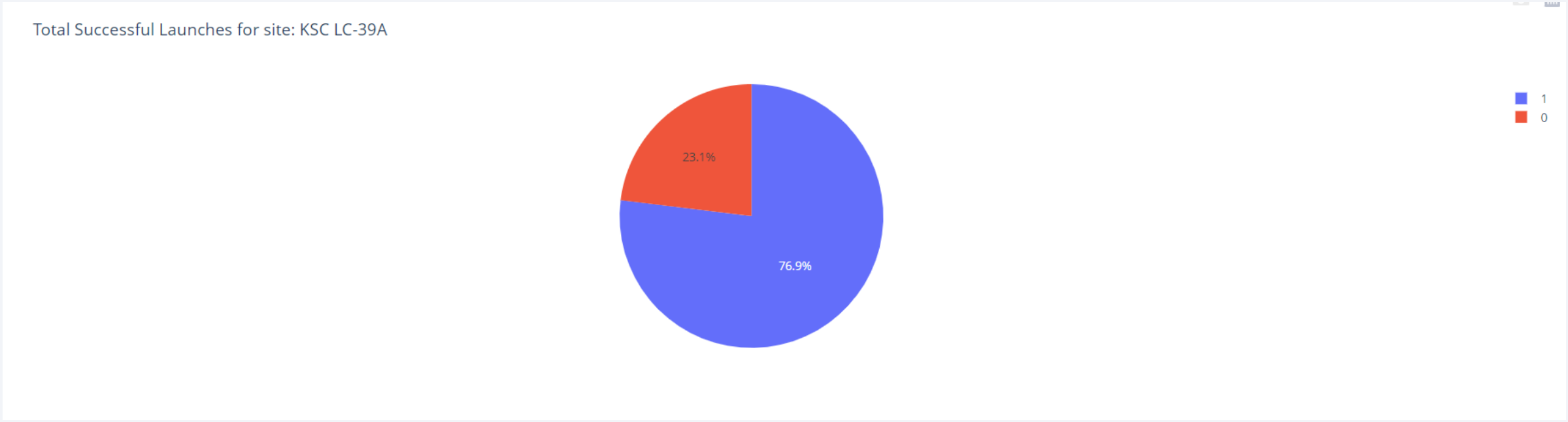


Launch Success count for all sites



- Launch site KSC LC-39A had the highest number of successful launches

Launch Site with Highest Success Ratio



- KSC LC-39A has the highest success ratio at 76.9%

Relationship between Payload and Launch Outcome

Higher launch success rates at lower payload masses

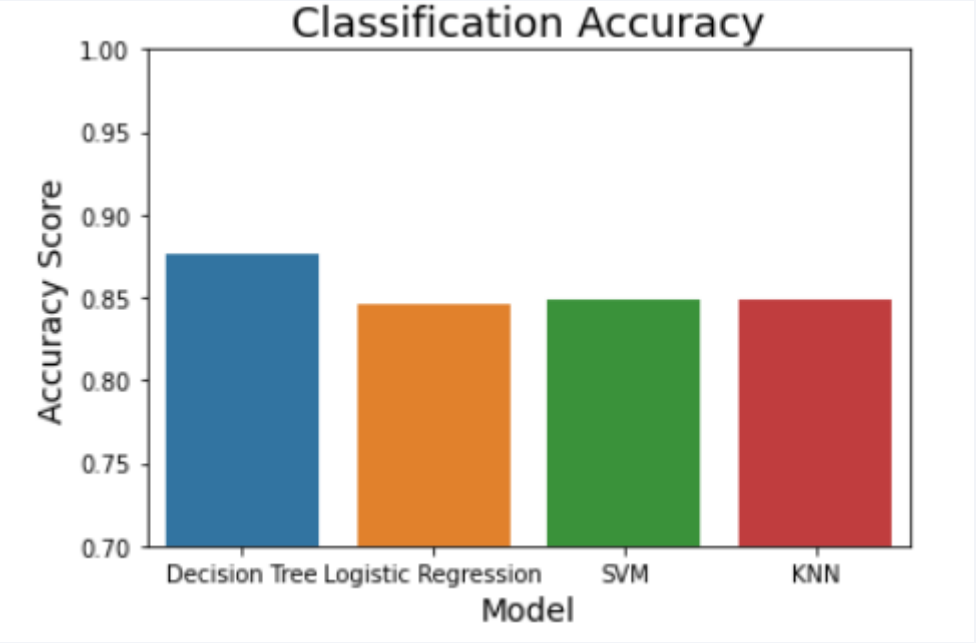
FT booster has the highest success rates

V1.0 and v1.1 has the lowest success rates

Section 6

Predictive Analysis (Classification)

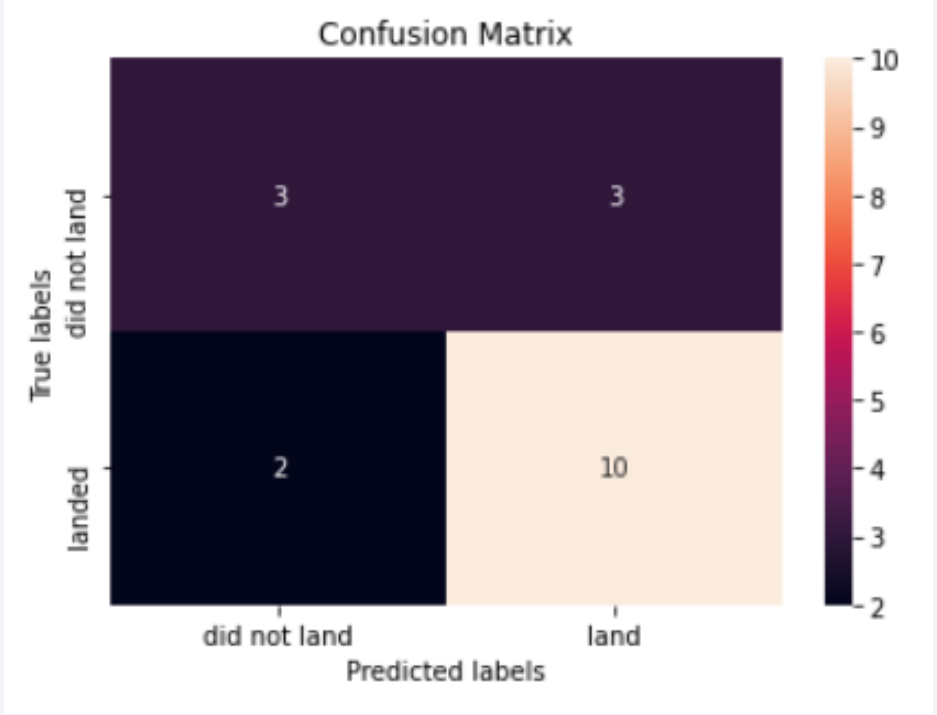
Classification Accuracy



- Decision tree model has the highest accuracy with a score of 0.8768.

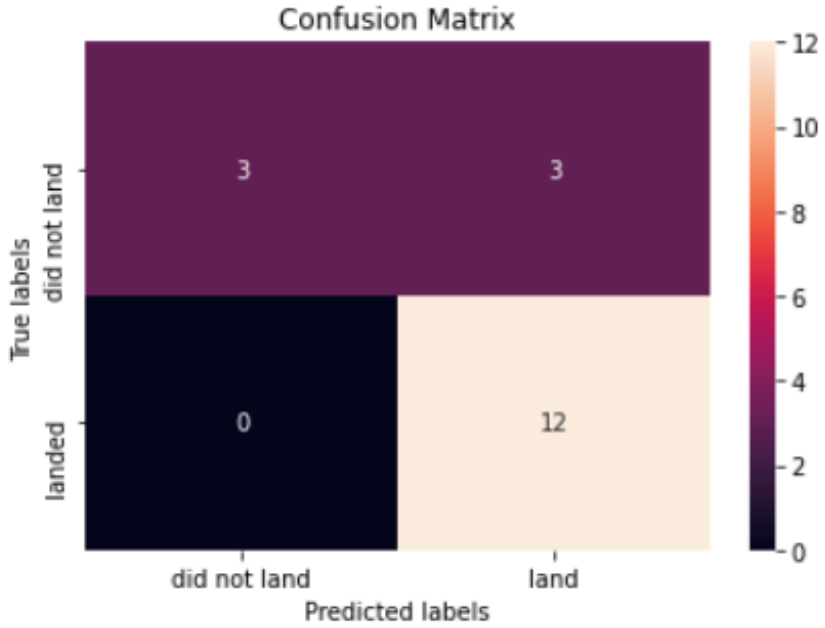
Confusion Matrix – Decision Tree

- Confusion matrix contains 2 false negatives and 3 false positives.
- The decision tree has the highest accuracy score with a score of 0. 8768
- However when tested on examples it hasn't seen, the accuracy score is reduced to 0.7222.
- This is likely a result of overfitting, since the decision model tends to grow to maximum depth by default.



Confusion Matrix - SVM

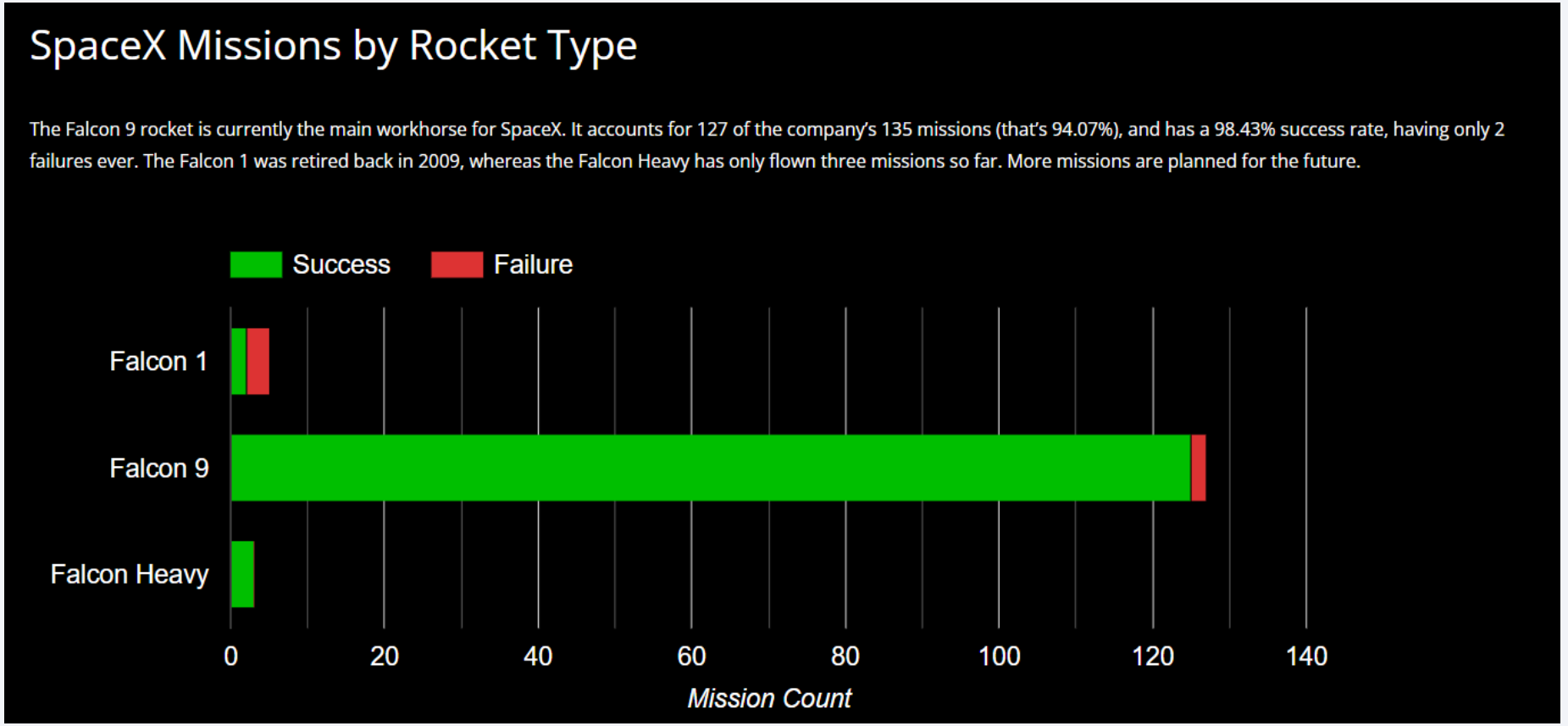
- A better model would the SVM model which has a training accuracy of 0.8482 and a test accuracy of 0.8333.
- The confusion matrix shows no false negatives.



Conclusions

- Consider using the B5 booster for large payloads.
- Ground pad landings has higher success rate than drone ship landings. However majority of landings were on drone ships.
- Launch sites should be positioned close to the equator as it is easier to reach equatorial orbit.
- Launch sites location should be close to coastlines, highways and railways, but far away from cities.
- Use SVM model to predict whether the first stage of the Falcon 9 will land successfully.

Appendix



Taken from [SpaceX website]

Thank you!