IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SHEKHAR JAYANTHI
05/03/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using Spacex API calls and web scraping.
  - Data wrangling to create training labels for outcome based on successful or unsuccessful mission.
  - Exploratory Data Analysis (EDA) with data visualization using scatter, line, and bar charts.
  - EDA with SQL queries.
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis using classification models.
- Summary of all results
  - EDA highlights the importance of number of flights, orbit type, launch site, and payload mass.
  - Interactive analysis using Folium illustrates the importance of location of the launch site in success ot failure.
  - Dashboard results show the launch site with greatest number of successes, and also the impact of payload mass on launch success for various booster versions.
  - Predictive modeling indicates that launch success can be predicted with an accuracy of 83.33%.

Introduction

- Project background and context
  - Competition is intensifying in the space arena.
  - Cost of launches is determined by first-stage of rocket launches.
  - Successful landing of the first-stage is critical to reducing mission costs.
- Problems you want to find answers
  - What are the key factors that enable successful landing of space launches?
  - Can we predict successful launches using these factors?

Section 1

Methodology

Methodology

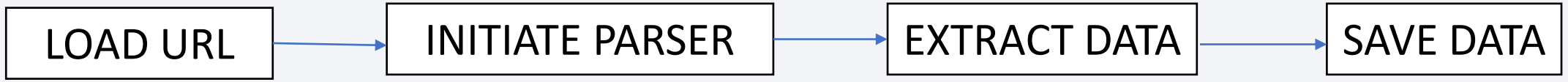
Executive Summary

- Data collection methodology:
  - The primary modes of data collection were SpaceX API calls and web scraping.
- Perform data wrangling
  - We converted mission outcome into training labels by considering successful and unsuccessful booster landings.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using training data, we estimated four classification models – logistic regression, decision tree, Support Vector Machine (SVM), and k Nearest Neighbor search. We used testing data set to calculate accuracy score.

Data Collection

- We collected data using SpaceX REST API calls and web scraping
  - SpaceX API calls were carried out by accessing the site: https://github.com/r-spacex/SpaceX-API
  - Web scraping was carried out using the Falcon 9 and Falcon Heavy Launches Records at .https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches#Past_launches
  - SpaceX API data: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

  - Wikipedia web scraping data: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, TimeDescribe how data sets were collected.



Data Collection – SpaceX API

GitHub URL

- CALL API
  - Request data using SpaceX REST API calls

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

- PARSE RESPONSE
  - Parse the .JSON response file

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

- EXTRACT DATA
  - Use functions to extract data

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

  - Combine columns to create a new data frame

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

- SAVE DATA
  - Filter data frame and save to CSV

```
data_falcon9 = launch_df[launch_df['BoosterVersion'] == 'Falcon 9']
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

GitHub URL

- LOAD URL
  - Perform an HTTP GET method to request the Falcon9 Launch HTML as an HTTP response.

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

- INITIATE PARSER
  - Create and use a BeautifulSoup object.

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

  - Find all tables and assign results to a list.

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

- EXTRACT DATA
  - Extract all column/variable names from the HTML table header.

```
column_names = []
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if(name != None and len(name) > 0):
        column_names.append(name)
```

  - Create an empty dictionary with keys

```
launch_dict= dict.fromkeys(column_names)
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

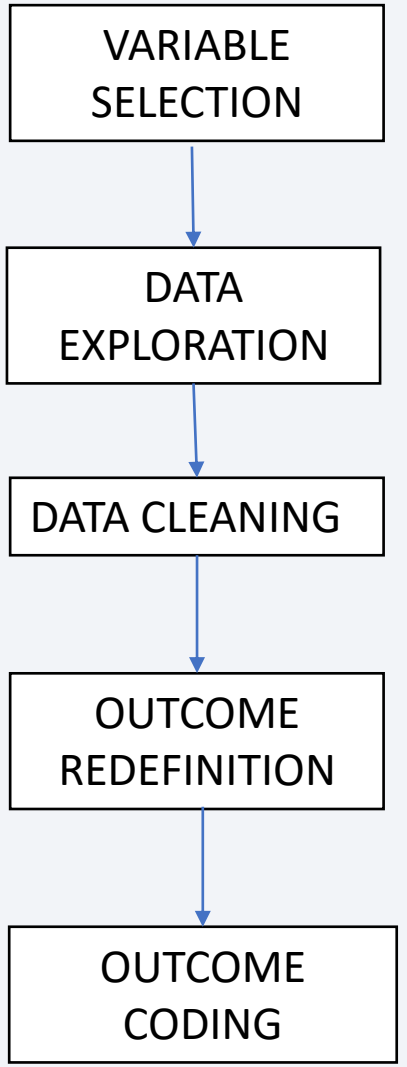
  - Fill dictionary with launch records
- SAVE DATA
  - Create a Pandas data frame by parsing launch HTML tables and export to CSV.

```
df=pd.DataFrame(launch_dict)
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

GitHub URL

- The purpose of data wrangling was to provide outcome labels for training supervised models.
- **Variable selection**: Launch site, orbit type, mission outcome
- **Data exploration**
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit
- **Data cleaning**: Delete observations with missing values
- **Outcome redefinition**:
  - True Ocean or False Ocean (Successful or unsuccessful ocean landing)
  - True RTLS or False RTLS (successful or unsuccessful ground pad landing)
  - True ASDS or False ASDS (successful or unsuccessful drone ship landing)
- **Outcome coding**: Create a label from redefined Outcome column
  - Class=1 (Success) or Class=0 (Failure)



EDA with Data Visualization

GitHub URL

- **Scatter chart**: The purpose of scatter plot was to see if there are any correlations between the variables. Following scatter plots were constructed:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
- **Bar chart**: The purpose of bar chart was to compare target variable across categories of a given variable. Following bar chart was used:
  - Orbit Type vs. Success Rate
- **Line chart**: The purpose of the line chart is to observe trend in the target variable over time. We developed the following line chart:
  - Year vs. Success Rate

EDA with SQL

GitHub URL

- As part f Exploratory Data Analysis (EDA), we performed SQL queries to:
  - Identify the names of unique launch sites.
  - List five records for the launch sites whose names begin with 'CCA.'
  - Evaluate the total payload mass carried by boosters from NASA (CRS).
  - Calculate average payload mass carried by booster version F9 v1.1 rockets.
  - Find the date when successful ground pad landing was achieved.
  - List booster versions with successful drone ship landing that carried payload mass between 4000 and 6000 kg.
  - List successful and failed mission outcomes.
  - List booster versions that carried maximum payload mass.
  - List failed outcomes in drone ship, their booster versions, and launch sites for 2015.
  - Rank the count for landing outcomes between 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

GitHub URL

- We created and added several map objects such as markers, marker clusters, circles to a folium map.
- The purpose of these map objects was to gain insights into the impact of location characteristics of the launch site on landing outcomes.
  - Markers were used to see the geographical location of launch sites.
  - Circles were used to drill down on the launch site to identify various details about the launch site, e.g., ground pad, railway tracks.
  - Color-coded marker clusters were used to visualize successful and failed landing outcomes at a given launch site.
  - Lines were used to visualize the distances of a selected launch site from infrastructure, e.g., railway, highway, city, and coastline.

Interactive Folium Map

Build a Dashboard with Plotly Dash

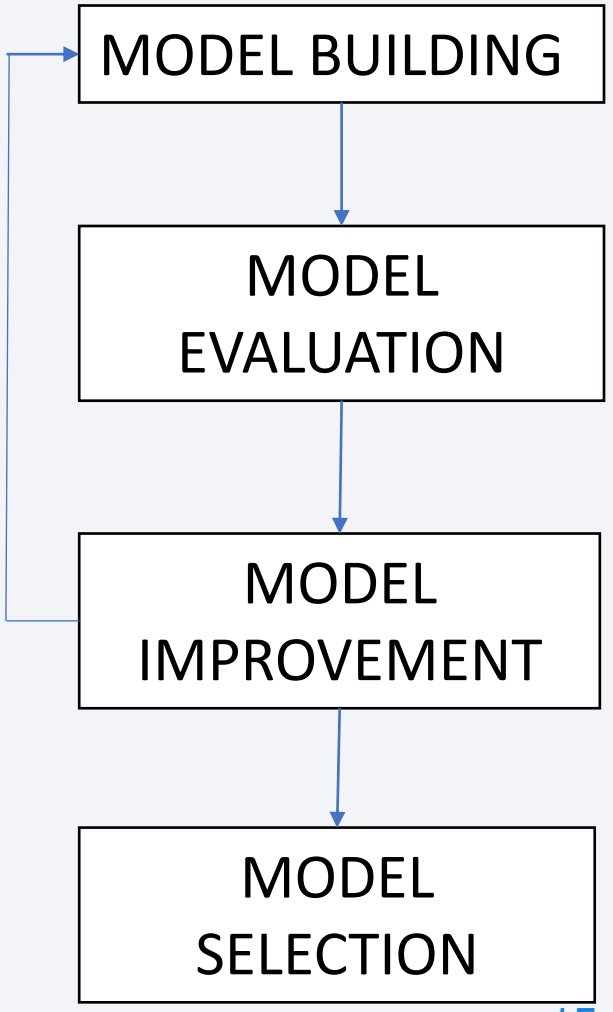
GitHub URL

- The Dashboard consists of two sections – a pie chart and a scatter plot.
- Pie chart
  - The pie chart was built to understand how the launch sites influence landing outcome.
  - The pie chart depicts the percentage successes and failures for different launch sites.
  - The dropdown can be used to change the pie chart for all sites or a particular individual site.
- Scatter plot
  - The scatter plot is designed to understand the impact of various booster versions that carry different payloads on landing outcome at various launch sites.
  - The scatter plot shows the successes and failures for different payload masses carried by rockets with different booster versions.
  - The scatter plot can be changed using a dropdown for all or particular individual sites, and also a slide ruler that enables selection of the payload mass in kg.

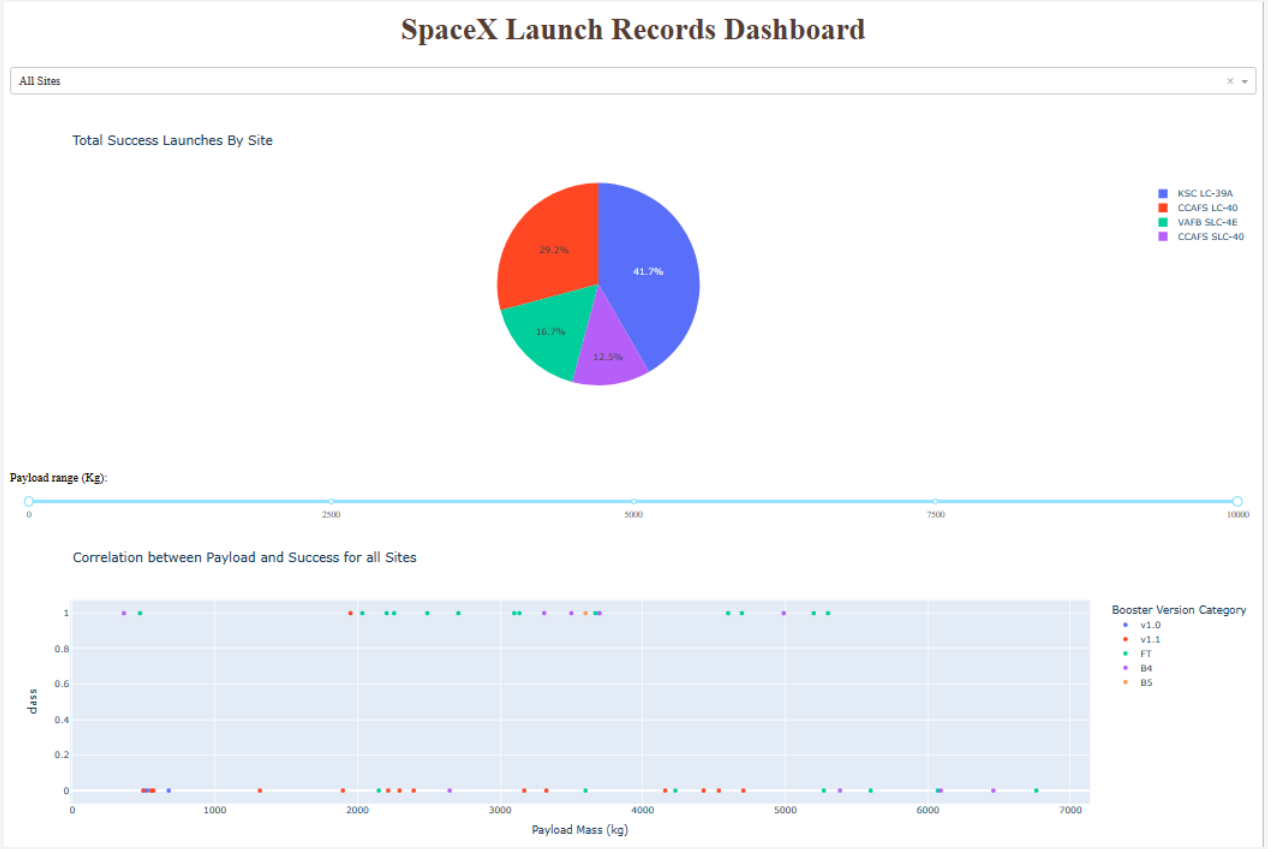
Predictive Analysis (Classification)

GitHub URL

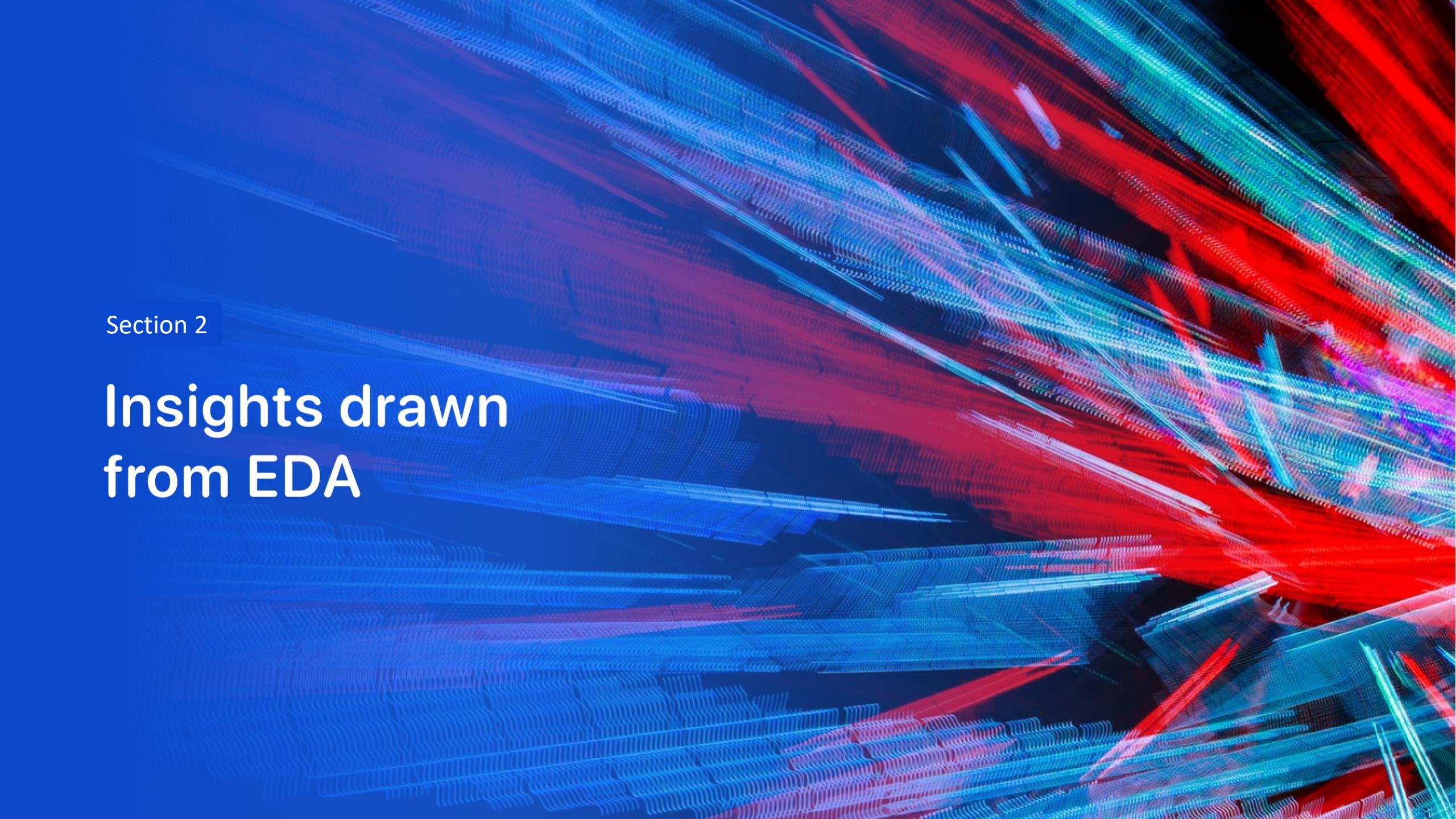
- Model building:
  - Code the dependent variable as success (Class=1) or failure (Class=0).
  - Identify various features found to be important in predicting success rate from exploratory data analysis (EDA).
  - Use these features to create training and testing datasets.
  - Estimate best parameters using Gradient Search on training dataset.
- Model evaluation:
  - Use best parameters to calculate accuracy scores from test dataset.
- Model improvement:
  - Build and evaluate other models using the procedure described in Model building and Model evaluation.
- Model selection:
  - Select the best model by comparing accuracy scores.



Results

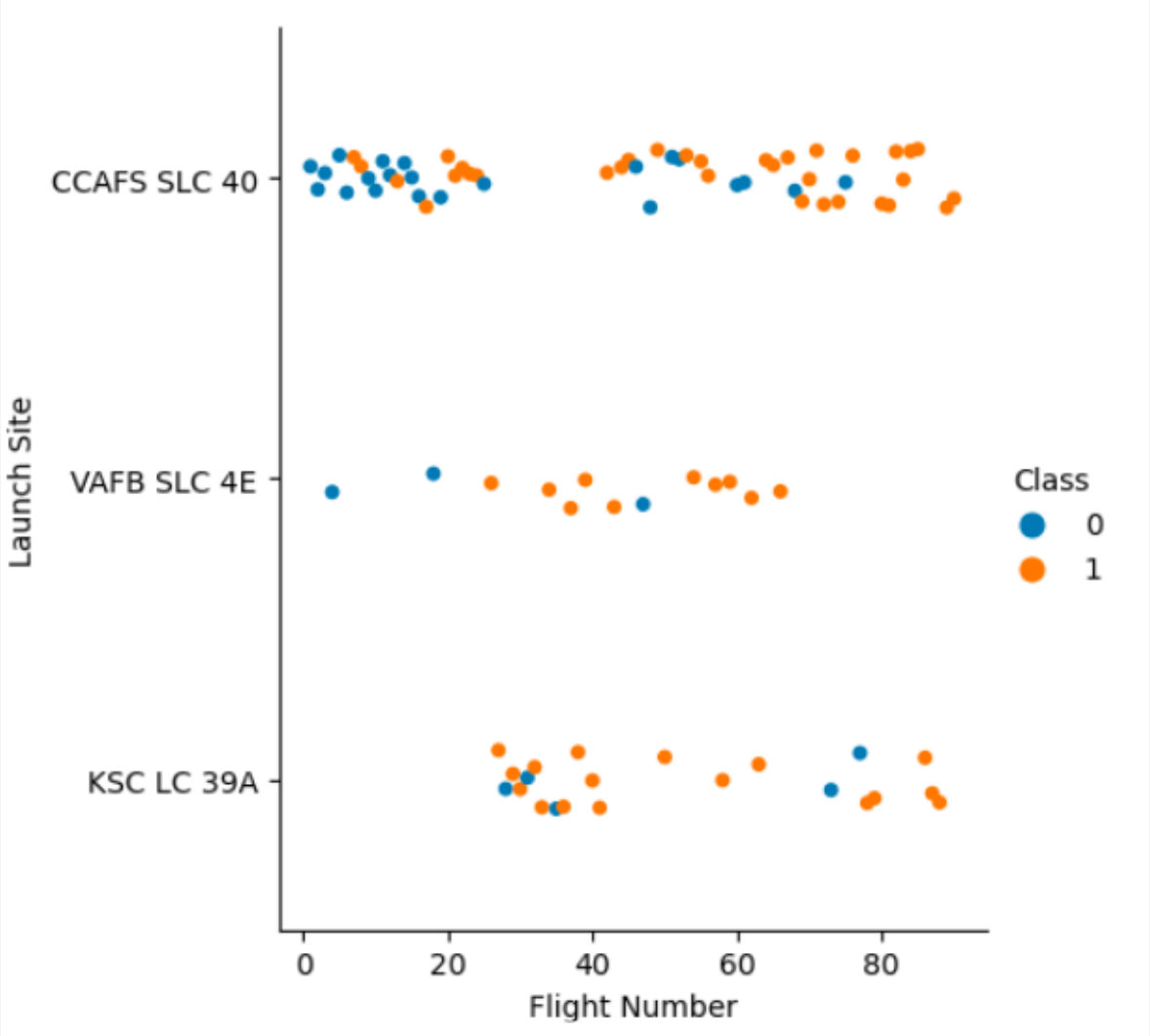


- The exploratory data analysis showed that success rate is dependent on several factors – flight number, orbit type, payload mass, launch site location, and booster version.
- Interactive analytics using a dashboard revealed that KS LC-39A has the highest success rate. Lower payload range fosters a variety of booster versions, while higher payloads favor only versions FT and B4.
- Predictive analysis conducted using four algorithms – logistic regression, decision tree, Support Vector Machine (SVM), and K Nearest Neighbor (KNN) – showed an identical accuracy of 83.33%.


Section 2
Insights drawn from EDA

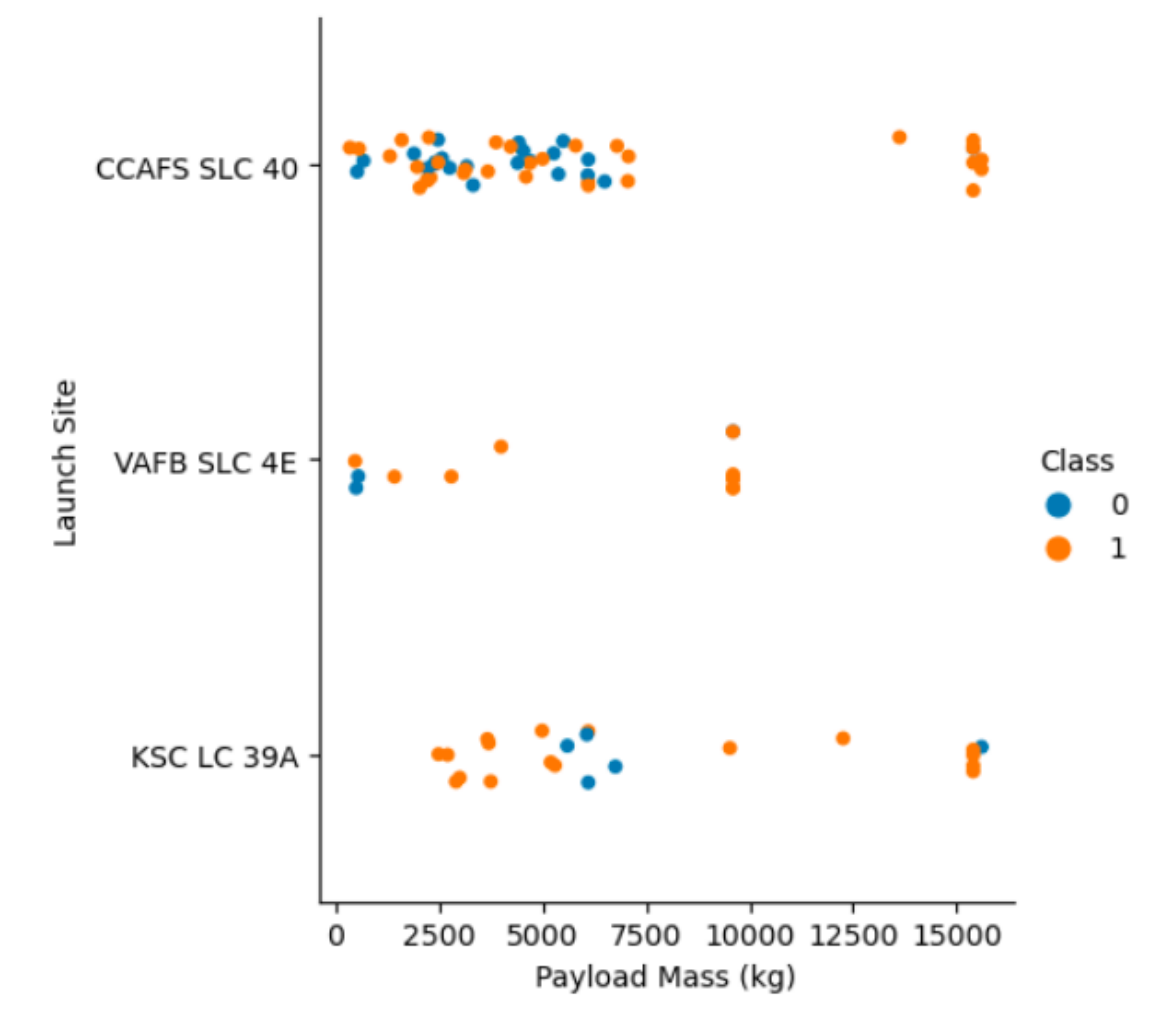
Flight Number vs. Launch Site

- The scatter plot shows the success (Class=1) and failure (Class=0) for each Launch site over flight number.
- In general, increase in flight number improves success of the launch for all the launch sites.
- Success appears to improve dramatically after the 20th flight for all the launch sites.



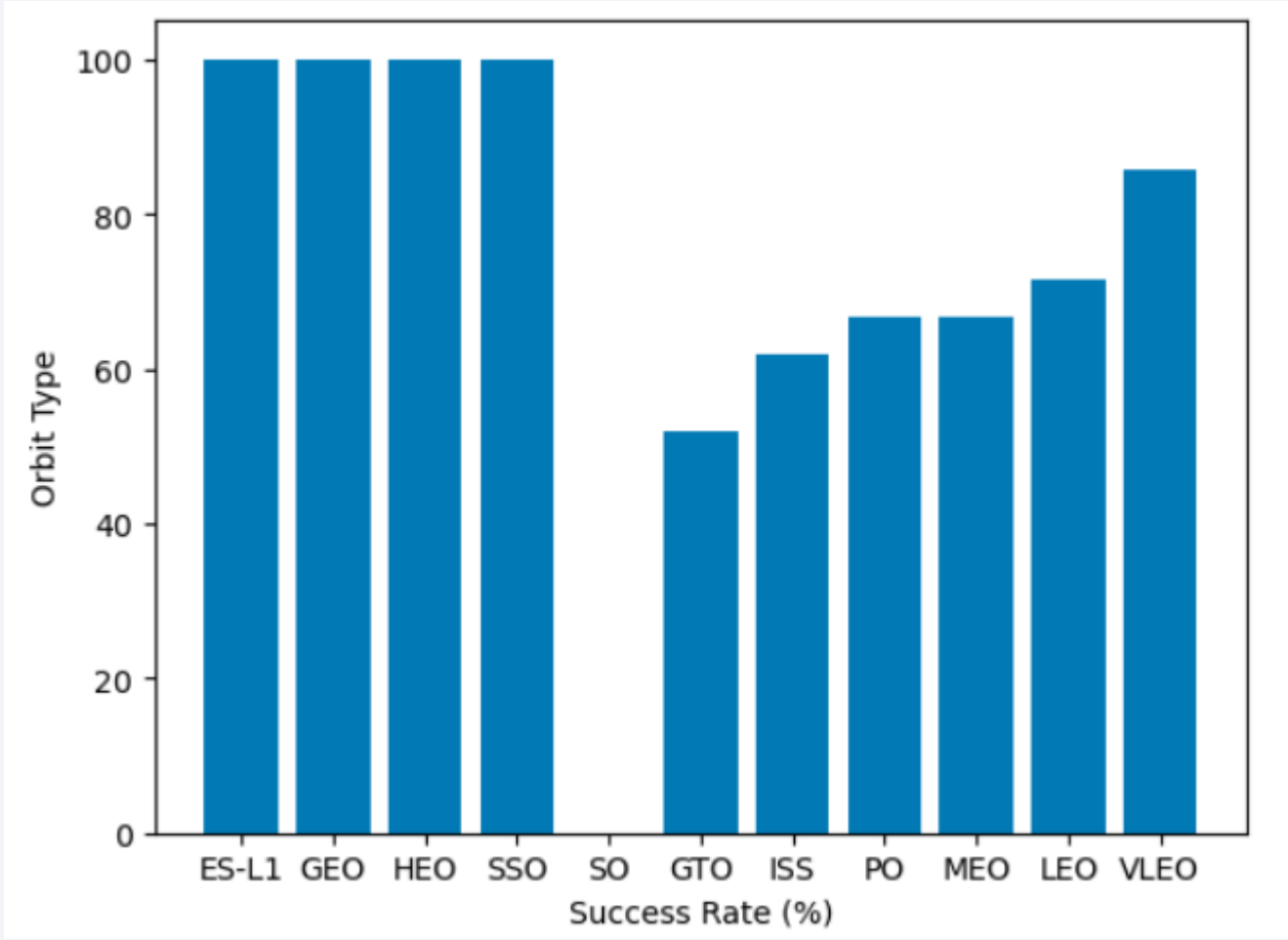
Payload vs. Launch Site

- The scatter plot shows successful (class=1) and failed (class=0) launches in terms of of payload and launch sites.
- The scatter plot shows a more consistent pattern of success for site KSC LC -39A over a wide range of payload mass.
- The success and failures are more concentrated for lighter (0-7500 kg) and heavier (15000 kg) payload for the launch site CCAS SLC -40A.
- For the launch site VAFB SLC -4E, the outcomes are concentrated for either lighter (0-5000 kg) or heavier (10000 kg) payload.



Success Rate vs. Orbit Type

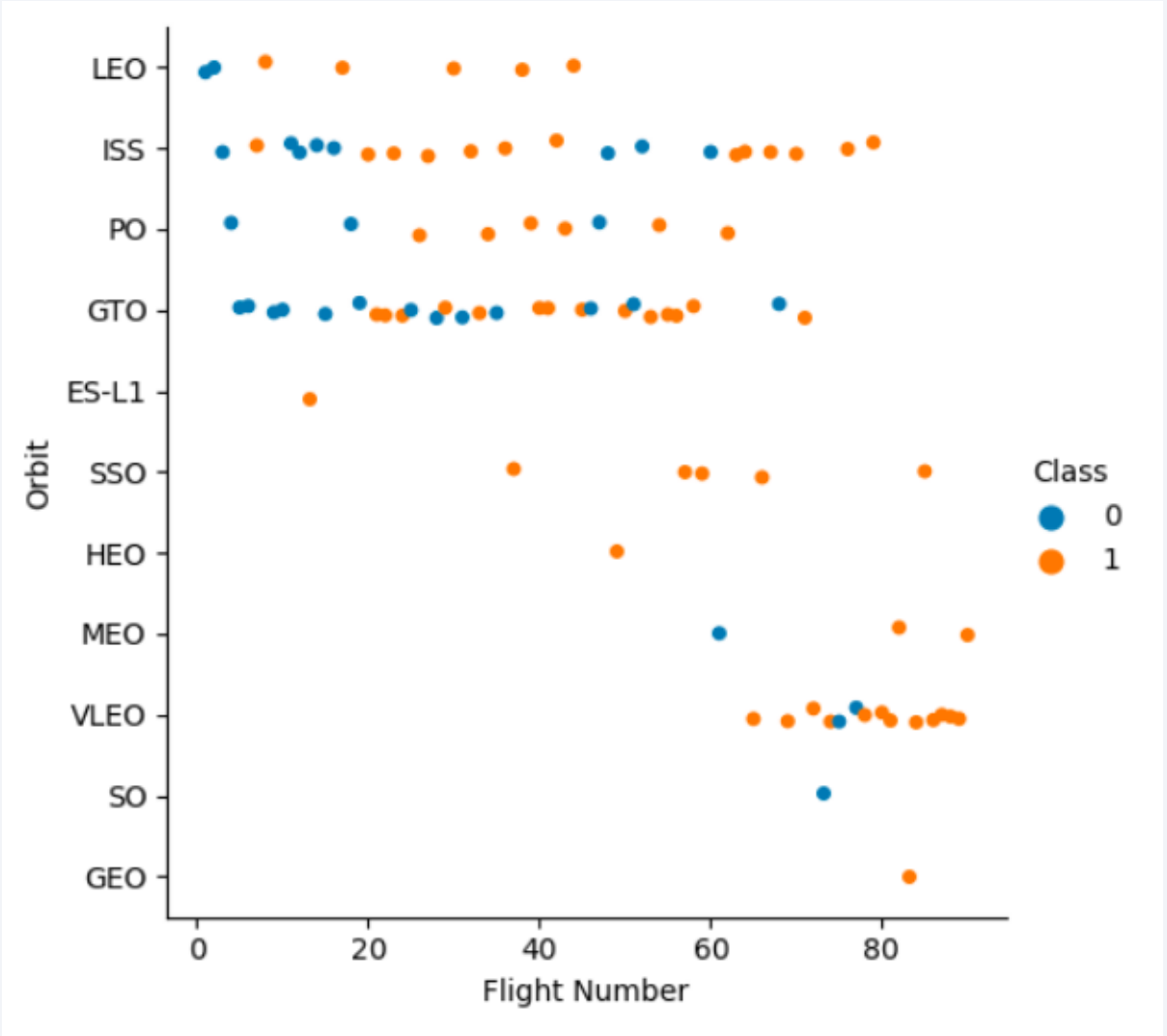
- The bar chart for the success rate of each orbit type shows that orbit types ES-L1, GEO, HEO, and SSO have the highest success rate.
- GTO orbit has the least success rate due to the inclusion of transfer orbits in space launches that may be more complex than a direct orbit, such as GEO.
- MEO, PO, LEO, and VLEO all have success rates in the middle range perhaps due to maneuverability issues.
- Orbit type SO is the only orbit without any success.



Reference: ESA - Types of orbits

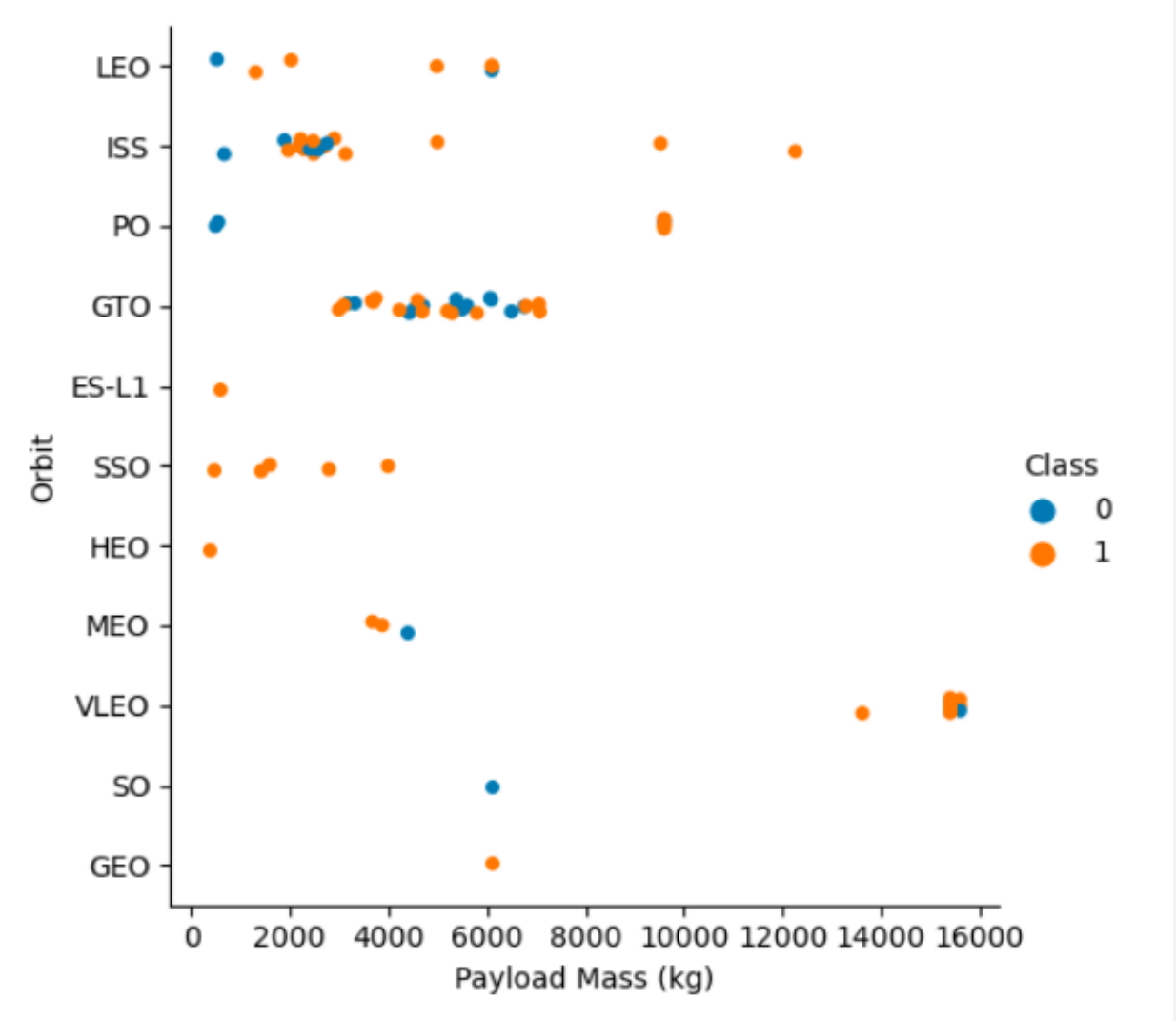
Flight Number vs. Orbit Type

- In general, increase in flight number improves success rate for all the orbit types.
- Initial number of flights are greater under orbit types LEO, ISS, PO, and GTO.
- Later flights seem to build on the experience of the above orbit types to achieve successes with fewer flights for orbits SSO, HEO, MEO, VLEO, and GEO.
- Only orbit SO shows failure (Class=0) and no success for greater number of flights.



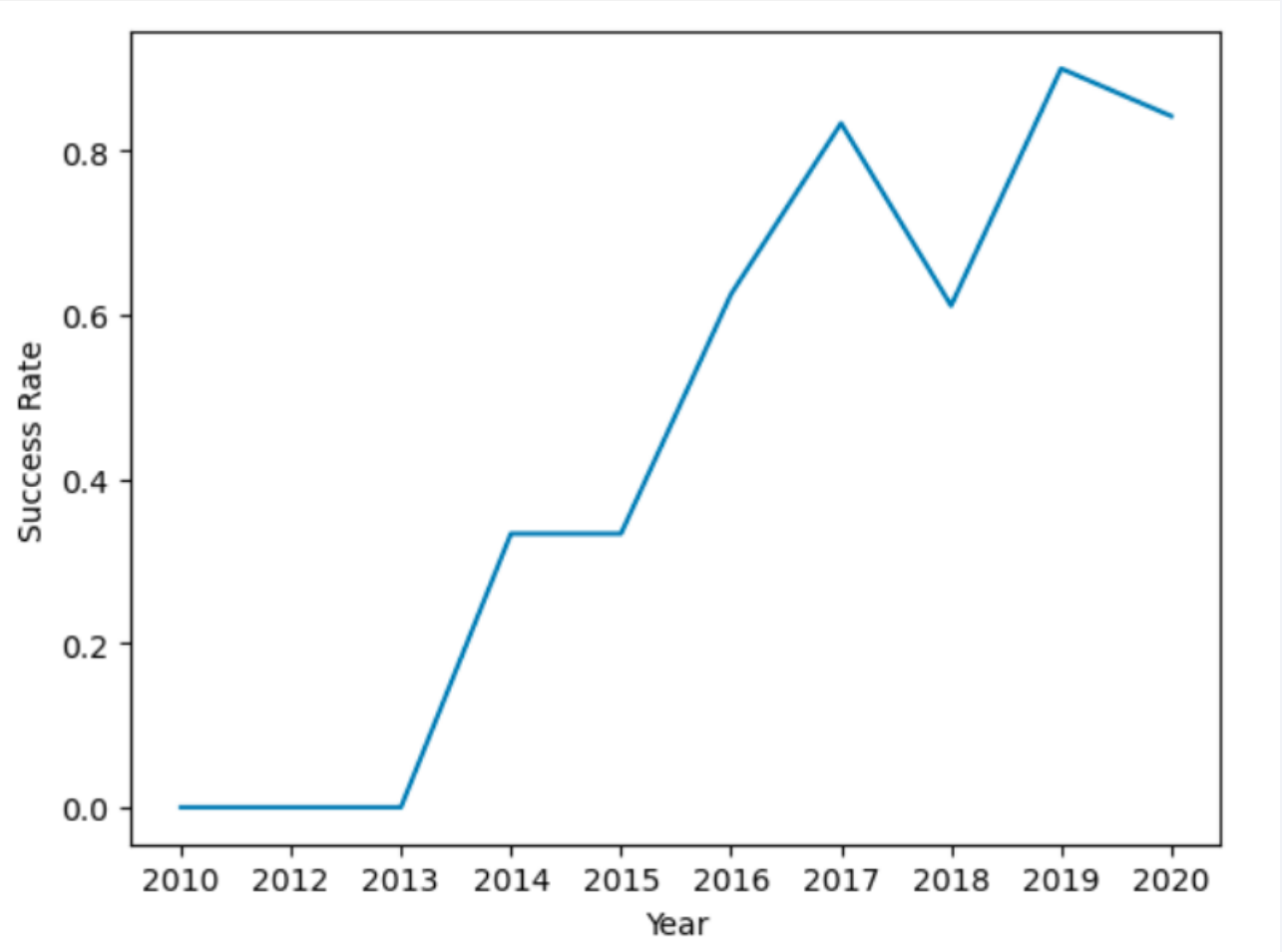
Payload vs. Orbit Type

- The scatter plot shows that greater successes (Class=1) rather than failures (Class=0) are in the range of payload mass from 0 to 6000 kg.
- The lighter payload mass from 0 to 3000 kg is associated with success for all orbit types except SO.
- The orbit type GTO shows a high concentration of success along with failures in the 3000-7000 kg range.
- Orbit types ISS, PO, and VLEO show successes for relatively higher payload masses.



Launch Success Yearly Trend

- In general, the line chart shows that yearly average success rate has significantly improved since 2013.
- However, there are some dips in the yearly average success rate for 2018 and 2020.



All Launch Site Names

QUERY

SELECT DISTINCT Launch_site FROM SPACEX DATA

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

- There are four unique launch sites that SpaceX operates.
- Only one of the sites (VB SLC-4E) is on the West Cost in California.
- The other three sites are on the East Coast in Florida.

Launch Site Names Begin with 'CCA'

QUERY

SELECT * FROM SPACEX DATA

WHERE Launch_site LIKE 'CCA%' LIMIT 5

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

- The query provided a list of five (LIMIT 5) launch sites in the dataset that begin with ‘CCA’.
- The list of these sites on the East Coast exhibit failed or no attempt in retrieving the first-stage rocket.

Total Payload Mass

QUERY

```
SELECT SUM(PAYLOAD_MASS_KG_) AS total_payload_mass_kg
FROM SPACEX DATA
WHERE Customer LIKE  'NASA (CRS)'
```

total_payload_mas_kg
45596

- The query returns the total payload mass SpaceX carried for NASA as 45,596 kg over the years.
- Over a period of approximately 15 years, the payload mass will account to approximately 3000 kg – a payload mass that seems to be the “best” choice for successful landing outcomes.

Average Payload Mass by F9 v1.1

QUERY

```
SELECT AVG(PAYLOAD_MASS_KG_) AS average_payload_mass_kg
FROM SPACEX DATA
WHERE Booster_Version = 'F9 v1.1'
```

average_payload_mass_kg
2928

- The average payload mass carried by Falcon 9 rocket with booster version F9 v1.1 is 2928 kg that is on the lighter side.
- One of the keys to successful launches appears to be using the most appropriate payload mass.

First Successful Ground Landing Date

QUERY

```
SELECT MIN(DATE) AS first_successful_landing_date FROM SPACEX DATA
WHERE Landing_Outcome = 'Success (ground pad)'
```

first_successful_landing_date
2015-12-22

- The first successful landing date on ground pad is 12/22/2015.
- This success shows that there was a considerable period of experimentation before successful landing on the ground.

Successful Drone Ship Landing with Payload between 4000 and 6000

QUERY

```
SELECT Booster_Version FROM SPACEX DATA
WHERE Landing_Outcome = 'Success (drone
ship)' AND (PAYLOAD_MASS_KG_ BETWEEN
4000 AND 6000)
```

booster_version	payload_mass_kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- The query returns boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000 kg.
- The list shows that boosters of the type “FT” alone landed successfully on a drone ship.

Total Number of Successful and Failure Mission Outcomes

QUERY

```
SELECT MISSION_OUTCOME, COUNT(*) AS COUNT
FROM SPACEX DATA
GROUP BY MISSION_OUTCOME
```

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

- The above query returns the number of mission success and failures.
- The accompanying list shows that SpaceX's missions have been highly successful with a success rate of 99%.
- Mission outcomes are different from landing outcomes that are targeted at recovering the first-stage of the rocket.

Boosters Carried Maximum Payload

QUERY

```
SELECT DISTINCT Booster_Version, PAYLOAD_MASS_KG_
FROM SPACEX DATA
WHERE PAYLOAD_MASS_KG_ IN ( SELECT
MAX(PAYLOAD_MASS_KG_)  FROM SPACEX DATA)
```

- The above query finds the maximum payload mass (kg) for various booster versions that are distinct within the dataset.
- The list displayed to the right shows that the booster versions B5 is the ne carrying the maximum payload mass for the Falcon 9 rockets.

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

2015 Launch Records

QUERY

```
SELECT Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEX DATA
WHERE Landing_Outcome = 'Failure (drone ship)' AND
YEAR(DATE) = '2015'
```

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

- The query returns Falcon 9 v1.1 boosters that resulted in failed landing in the drone ship mode in the year 2015.
- Both the instances displayed in the list are located in the launch site (CCAS LC-40) closer to the cast line.

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

QUERY

```
SELECT Landing_Outcome, COUNT(*) AS COUNT
FROM SPACEX DATA
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT DESC
```

landing_outcome	COUNT
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

- The query returns a list of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- The list suggests that success and failure as being equally likely. This could be the result of experimentation during the initial period of rocket launches.

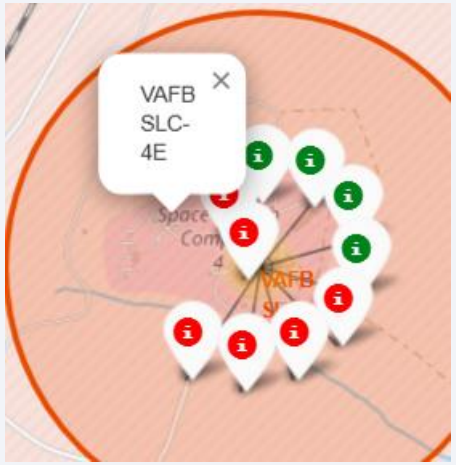
Section 3

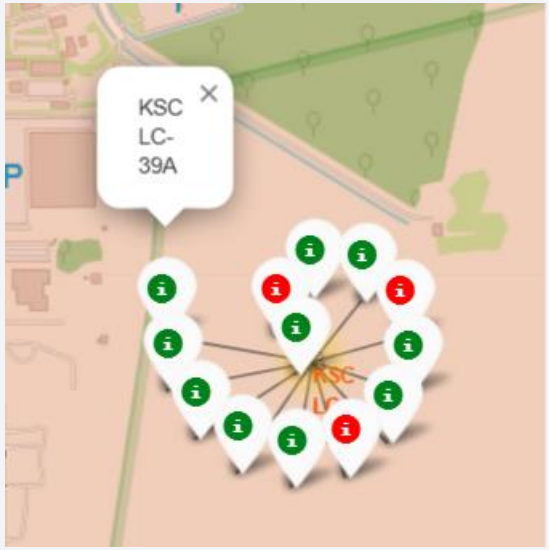
Launch Sites Proximities Analysis

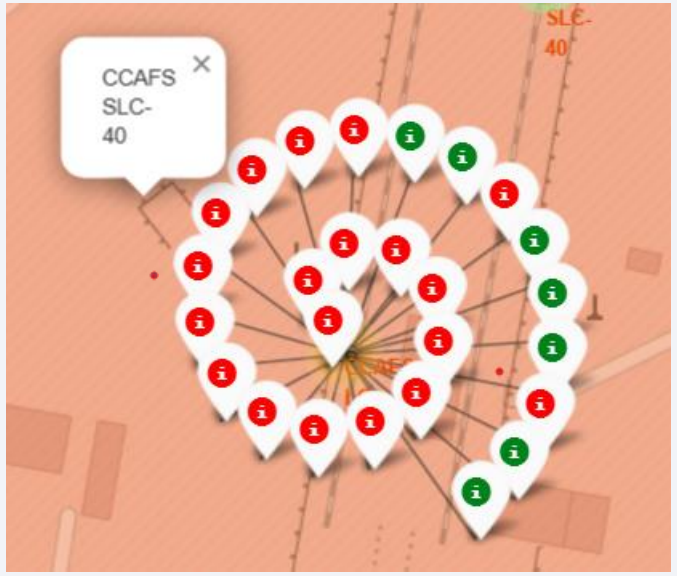
Launch site locations

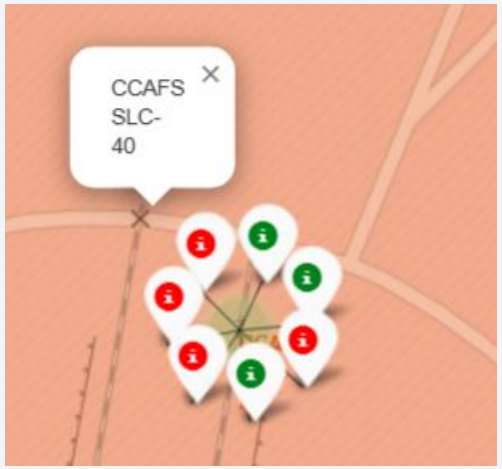
- There are four launch sites within the United States: CCA SLC-40, CCAFS SLC-40, KSLC-39A, VAFB SLC-4E.
- Only one of the launch sites (VAFB SLC-4E) is on the west cast.

Launch site outcomes

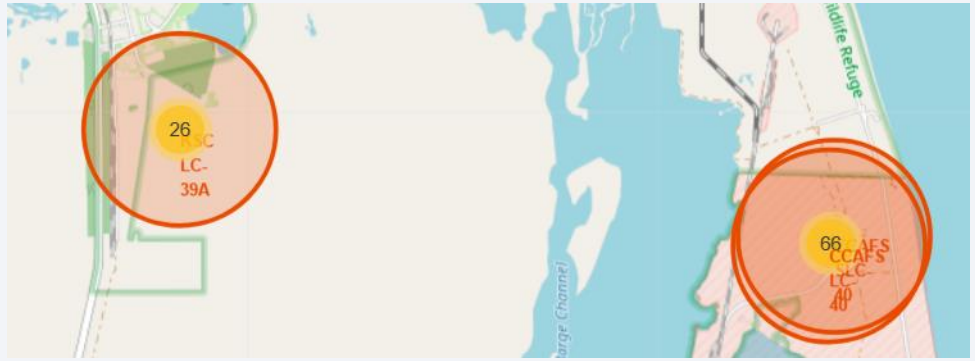




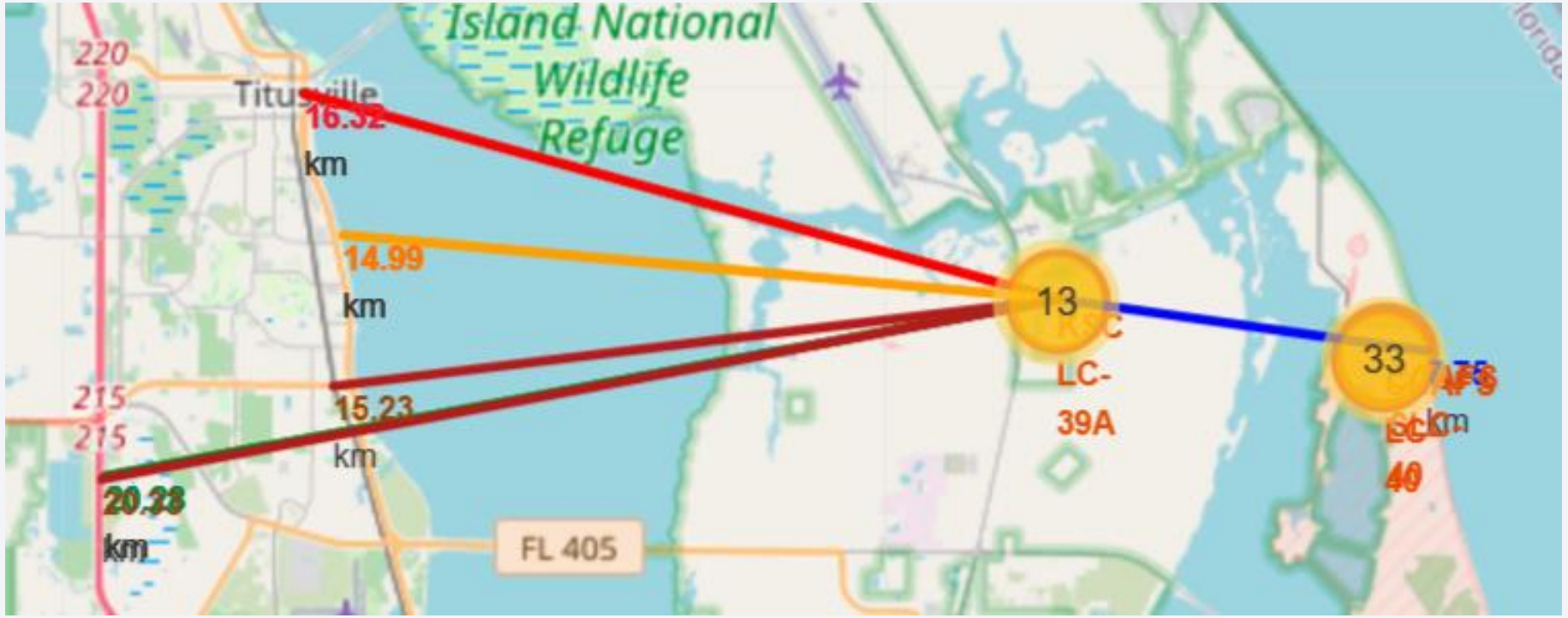




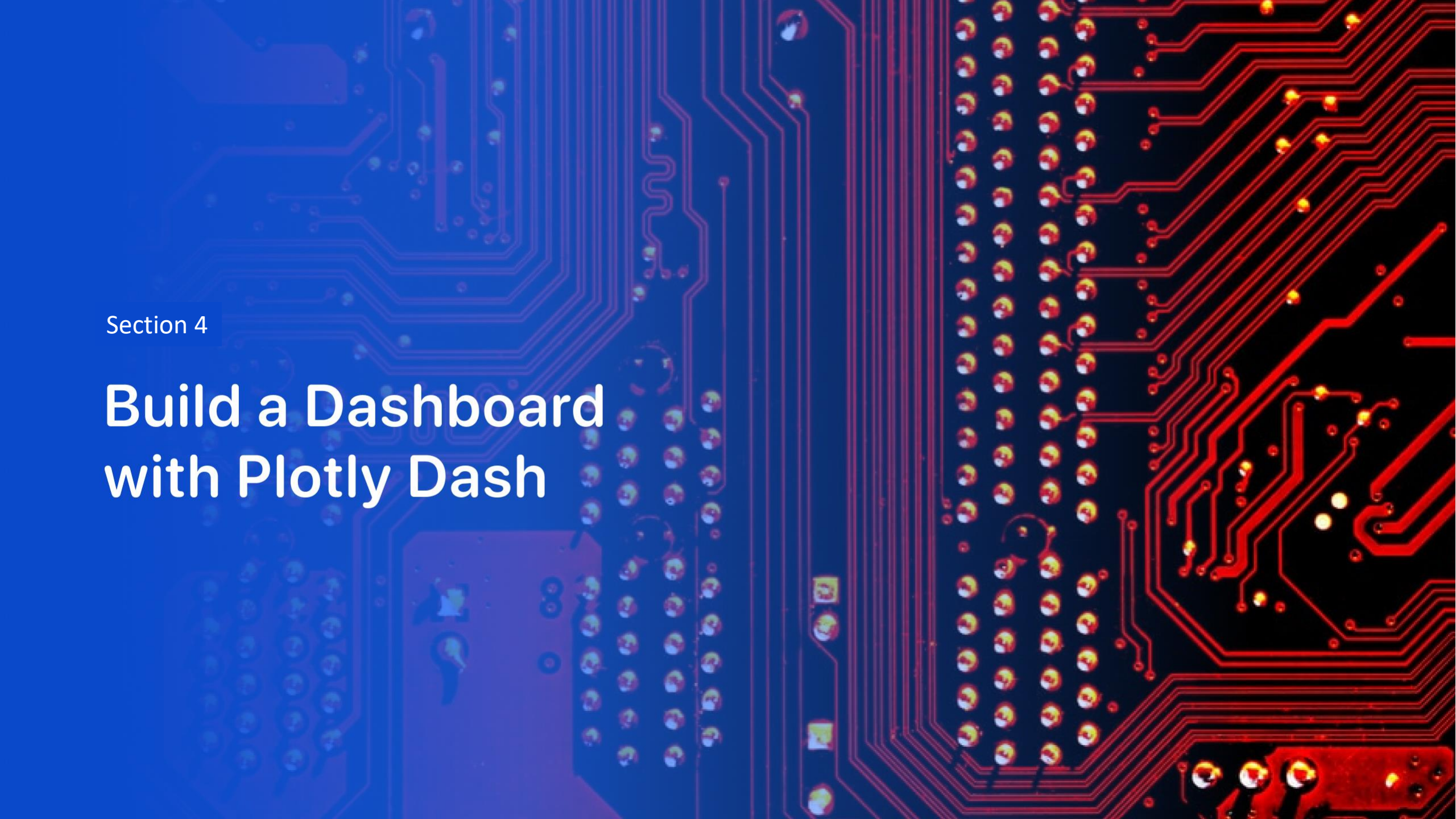
- From the markers, it can be readily seen that the ratio of successful (green labels) as opposed to failed launches (red labels) is the best for KSCLC-39A launch site.



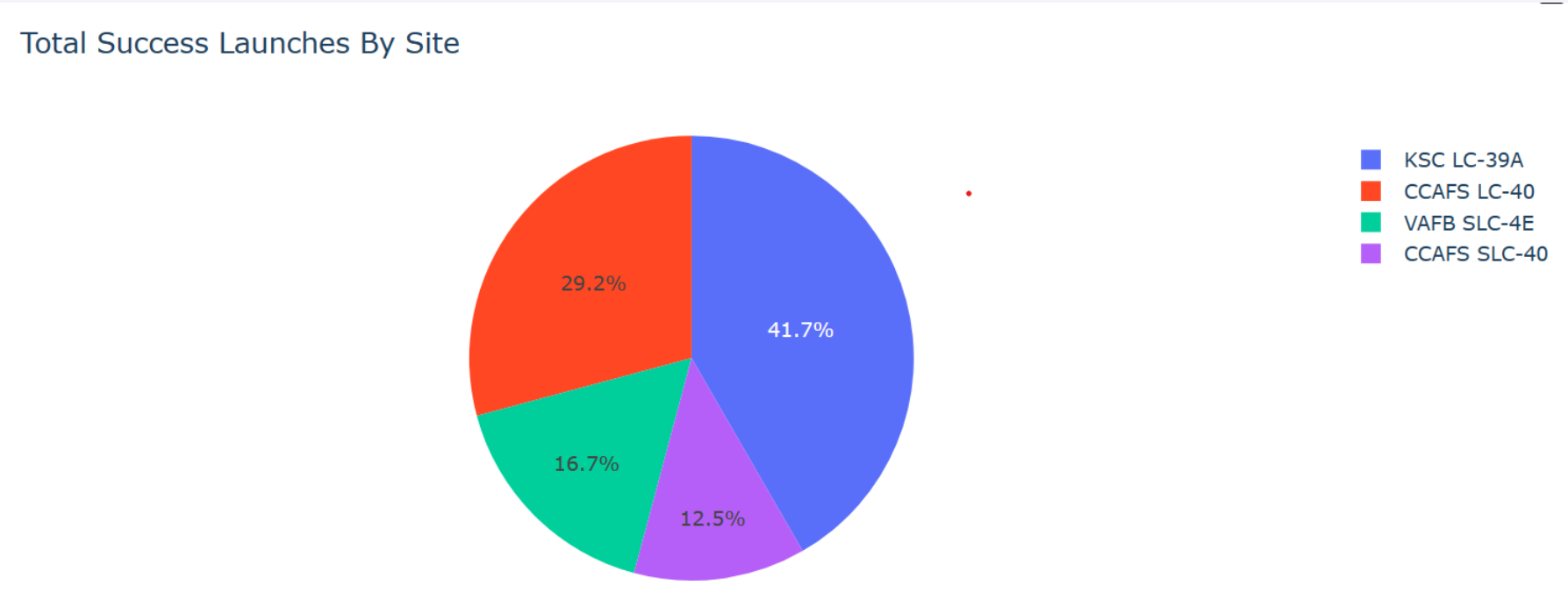
Launch site proximity



- The selected launch site (KSLC-39A) is reasonably far away from nearest city, coastline, railway, and a major highway indicating that failed launches do not influence nearby infrastructure.
- The two launch sites closer to the coast (CCA SLC-40 and CCAFS SLC-40) may facilitate drone ship or ocean landing.

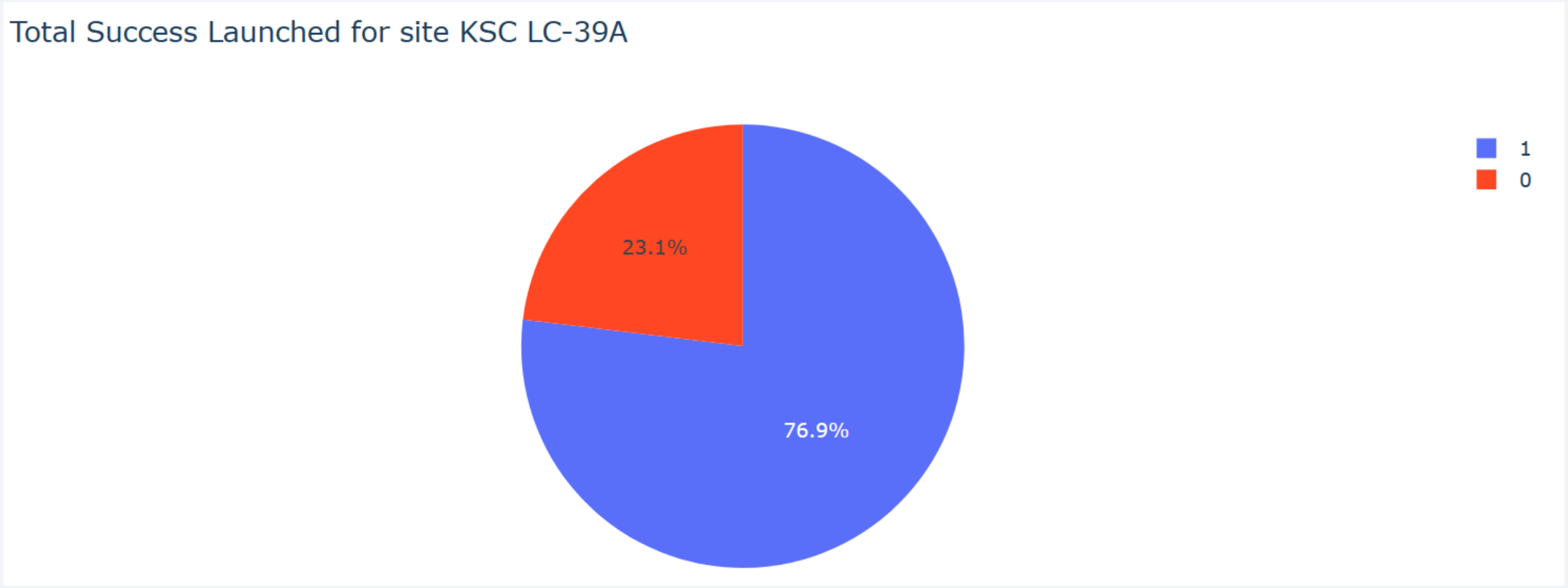

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



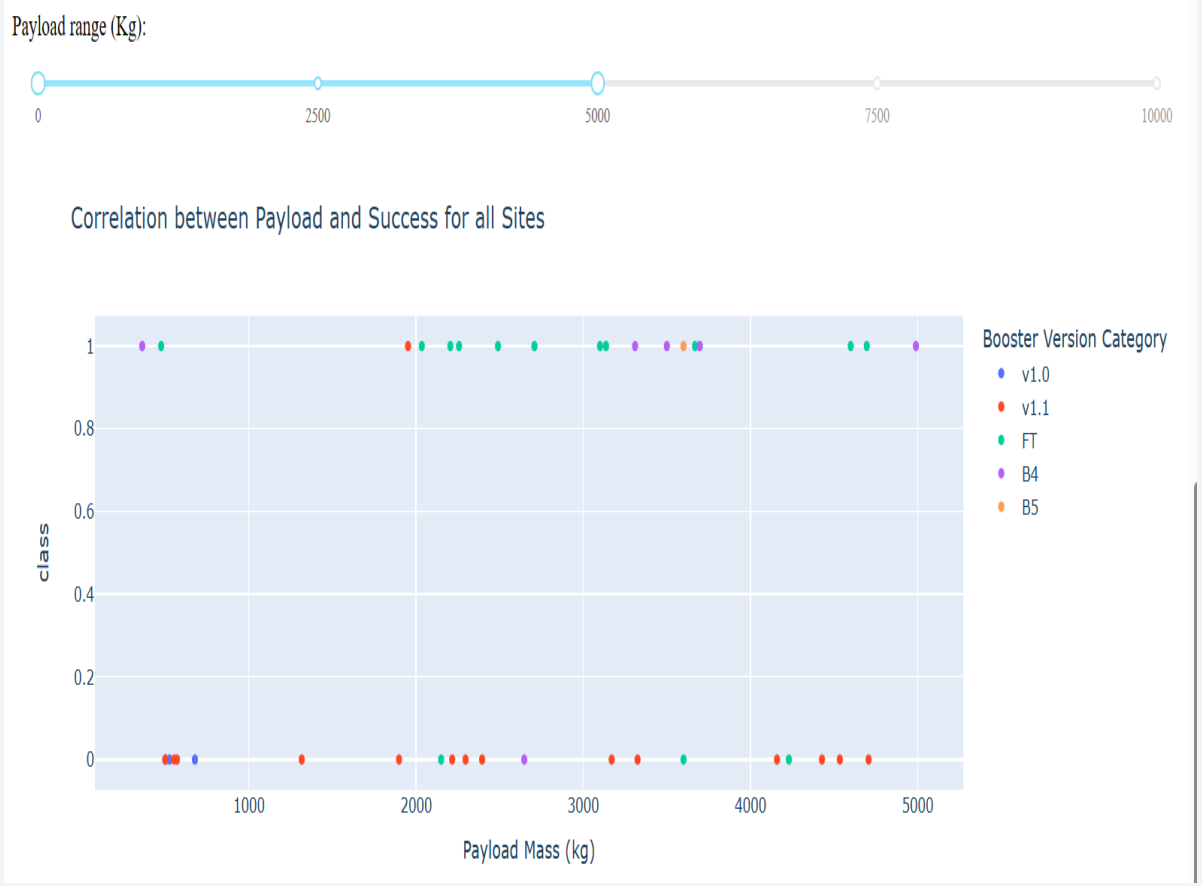
- As per the pie chart, the greatest successes are recorded for the site KSLC -39A, which may be because the launch site becomes operational after SpaceX gained experience from initial flights.
- The success ratio is the least for the launch site CCAFS SLC-40 as this site may be an experimental arena for initial flights launched by SpaceX.

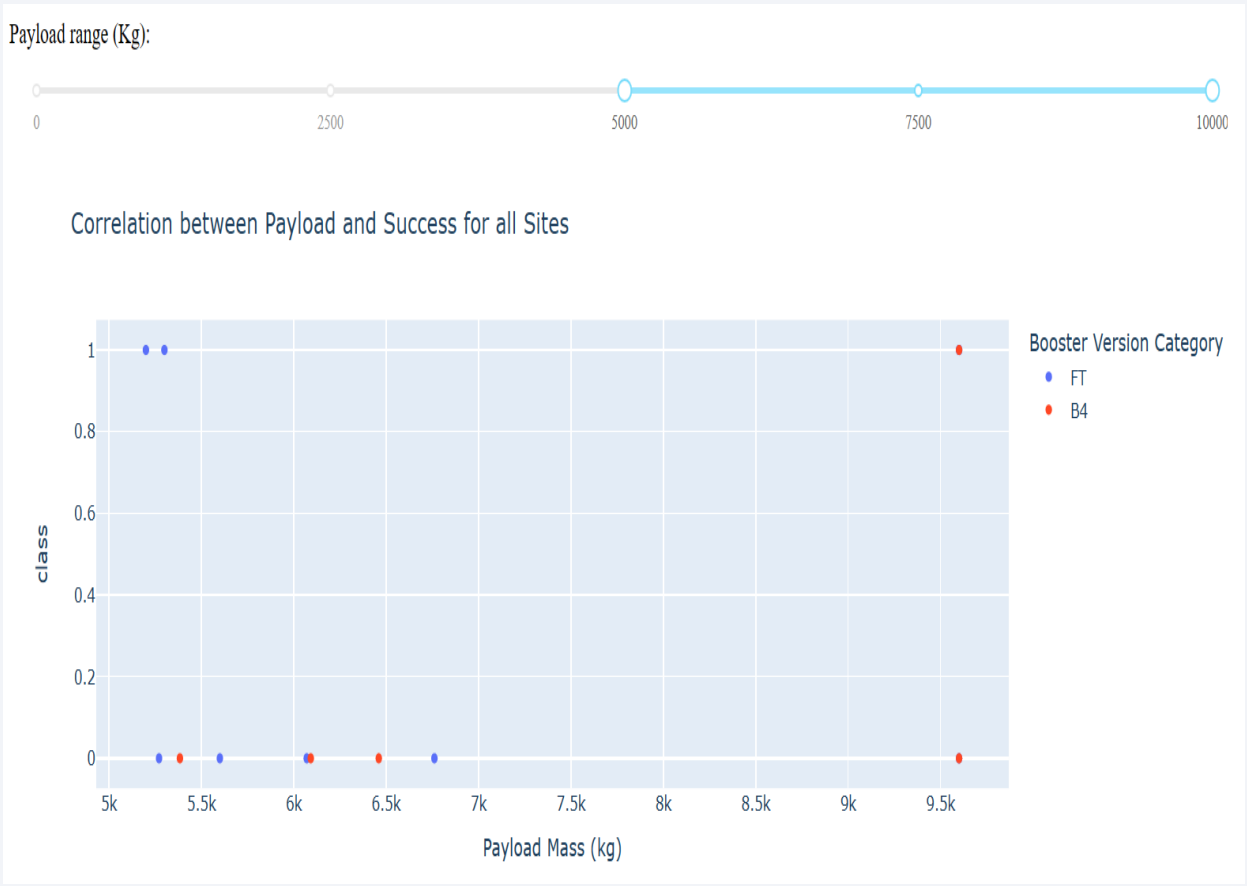
Launch site with highest success



- The success ratio is 76.9% (Class=1) and failure is 23.1% (Class=0) for the launch site KSC LC-39A.

Payload range for all launch sites





- The visualization using slide rule shows that success (Class=1) rate is better for lighter payload range (0-5000 kg) using a variety of booster versions, whereas only booster versions FT and B4 appear in higher payload range (5000 to 10000 kg).

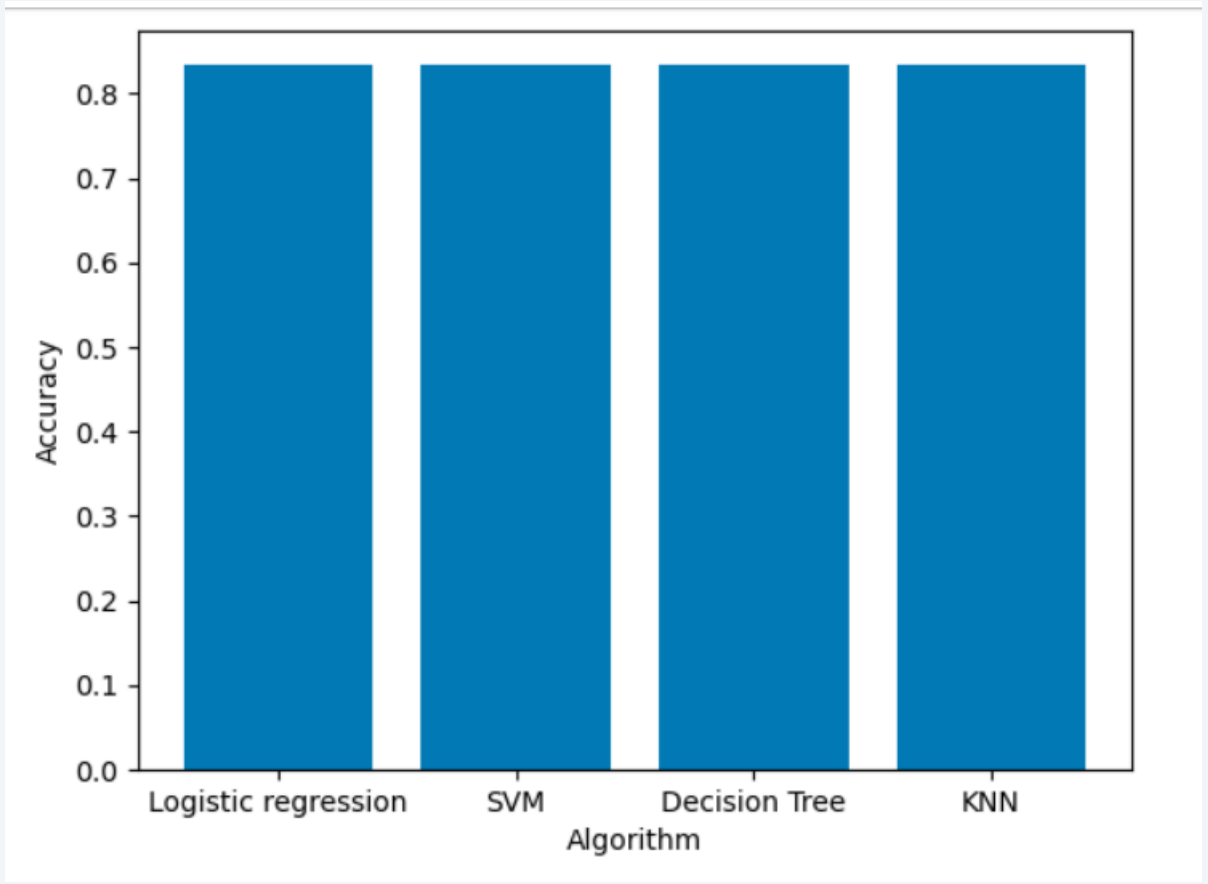
Section 5

Predictive Analysis (Classification)

Classification Accuracy

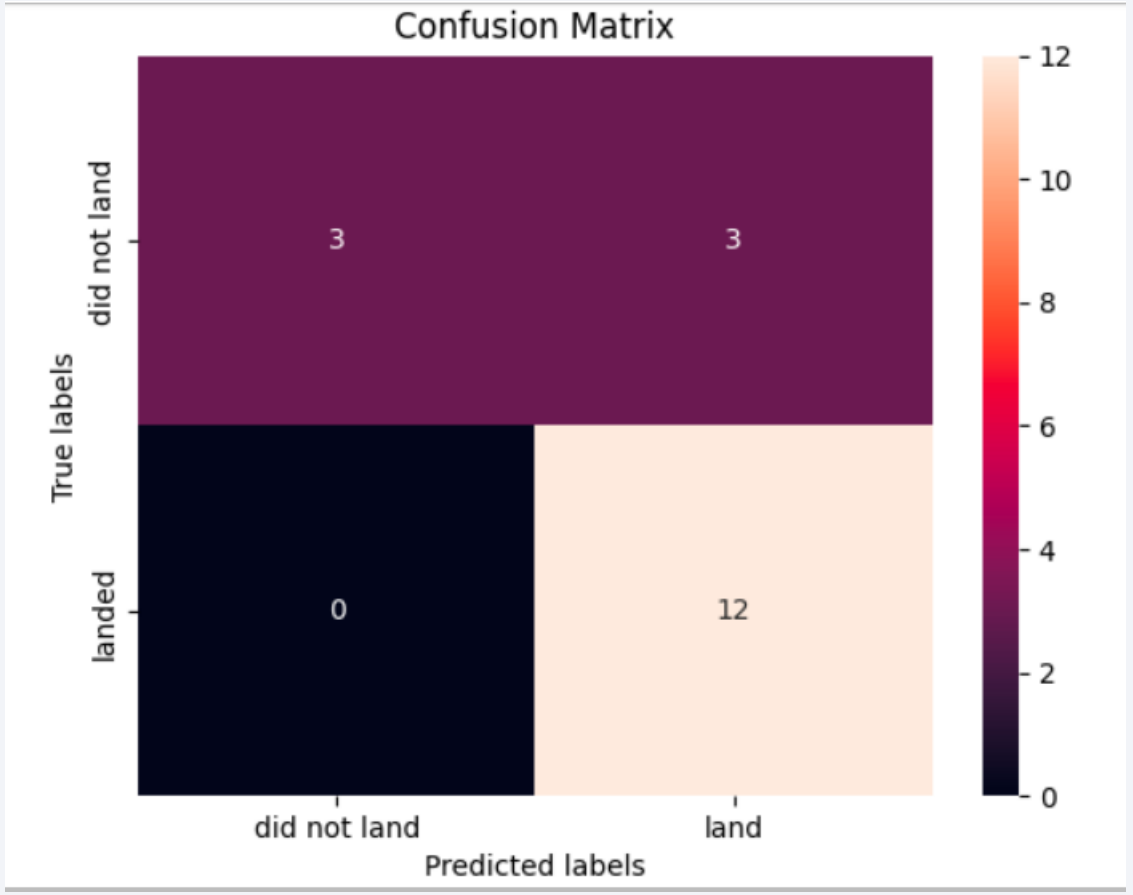
- All models exhibit identical accuracy scores.
- Thus, any of the models can be used for making predictions.

	Accuracy score
Logistic Regression	0.833333
Support Vector Machine	0.833333
Decision Tree	0.833333
k Nearest Neighbor	0.833333



Confusion Matrix

- The confusion matrix for all models is identical.
- The confusion matrix aligns true labels to predicted labels.
- Accuracy of the algorithm can be obtained by dividing the instances of true positives and true negatives by the sum of all instances in the matrix.
- Thus, accuracy for all models is 15/18=0.83.



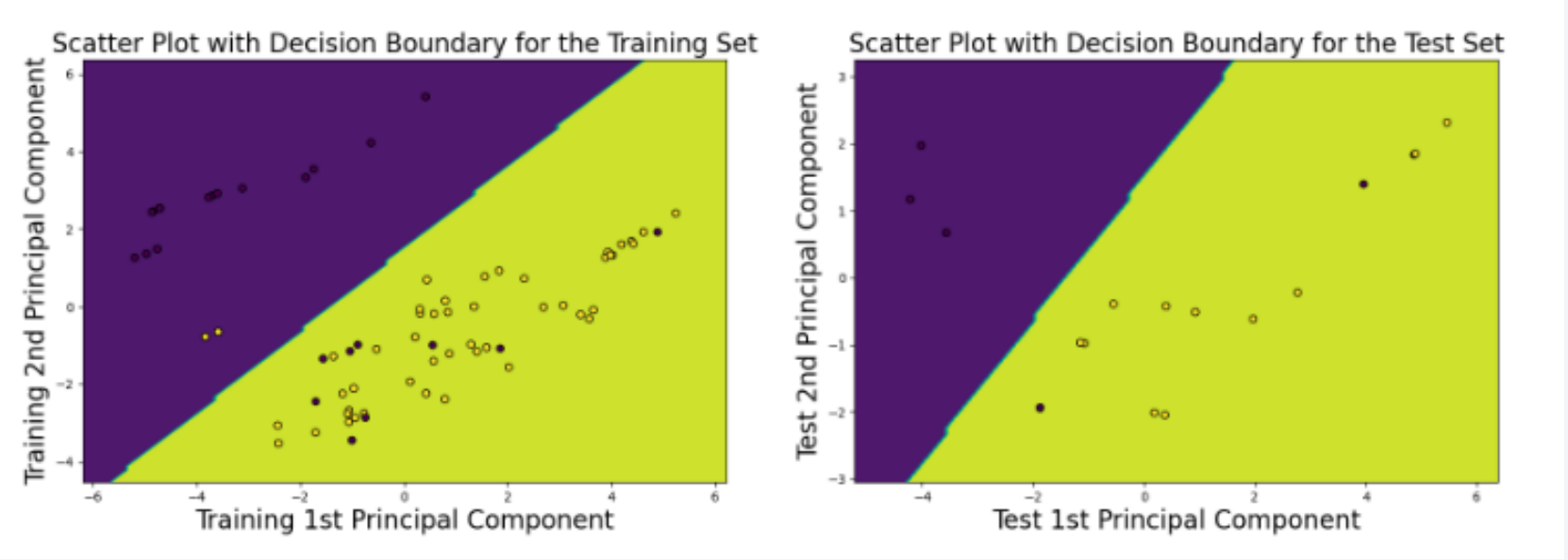
Conclusions

- Success rate improves with experience in terms of number of flights.
- With some hitches, success rate improved consistently after the year 2013.
- Orbit altitude appears to be associated closely with success rate in a U-fashion. SSO, HEO, GEO, and ES-L1 orbits have 100% success rates, while other orbit types show improved success as altitude decreases.
- The launch site KSCLC-39A shows the highest success rate, which can be attributed to experience gained from flights at other launch sites.
- This site (KSCLC-39A) is reasonably far from infrastructure and coastline.
- Lower payload launches have higher success rate than those with larger payloads, especially for higher altitude orbits, and *vice versa*.
- All machine learning models used in the analysis yield an accuracy of 83.33% in predicting launch success.

Appendix

CAPSTONE URL

Decision Boundary



- The decision boundary using Principal Component Analysis based on logistic regression model is consistent with the confusion matrix for the test data.
- There are three failures left of the decision boundary.
- The remaining 15 cases with clear circles appear to the right of decision boundary in green, which include three failures (predicted and true labels do not match) and 12 successes (predicted and true labels agree).

Reference: https://favtutor.com/blogs/decision-boundary-logistic-regression

Thank you!