IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sue Ann Oh
19 May 2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Summary of all results

- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Introduction

Project background and context

- A classification model was built to predict if Falcon 9 first-stage will land successfully, based on certain input features.
- Falcon 9 rocket launches from SpaceX cost more than 2.5 times less than those of other providers, primarily due to first-stage reuse.
- Predicting the likelihood of a successful landing can help estimate launch costs—valuable information for potential competitors.

Questions to be answered

- The model is trained on historical SpaceX launch data.
- Before training, the raw dataset must be cleaned, filtered, and transformed into a usable format. Key questions include:

i) Is the target variable (i.e. landing outcome) already in a format suitable for modeling?

ii) What are the features (i.e. conditions/factors) & their interactions which affect landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API request to the SpaceX API
  - Web scraping from Wikipedia page
- Perform data wrangling
  - Convert landing outcome entries from categorical to numerical (0: failure, 1: success)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  For a selection of classifiers (Logistic Regression, Support Vector machines, Decision Tree Classifier & K-nearest neighbors)

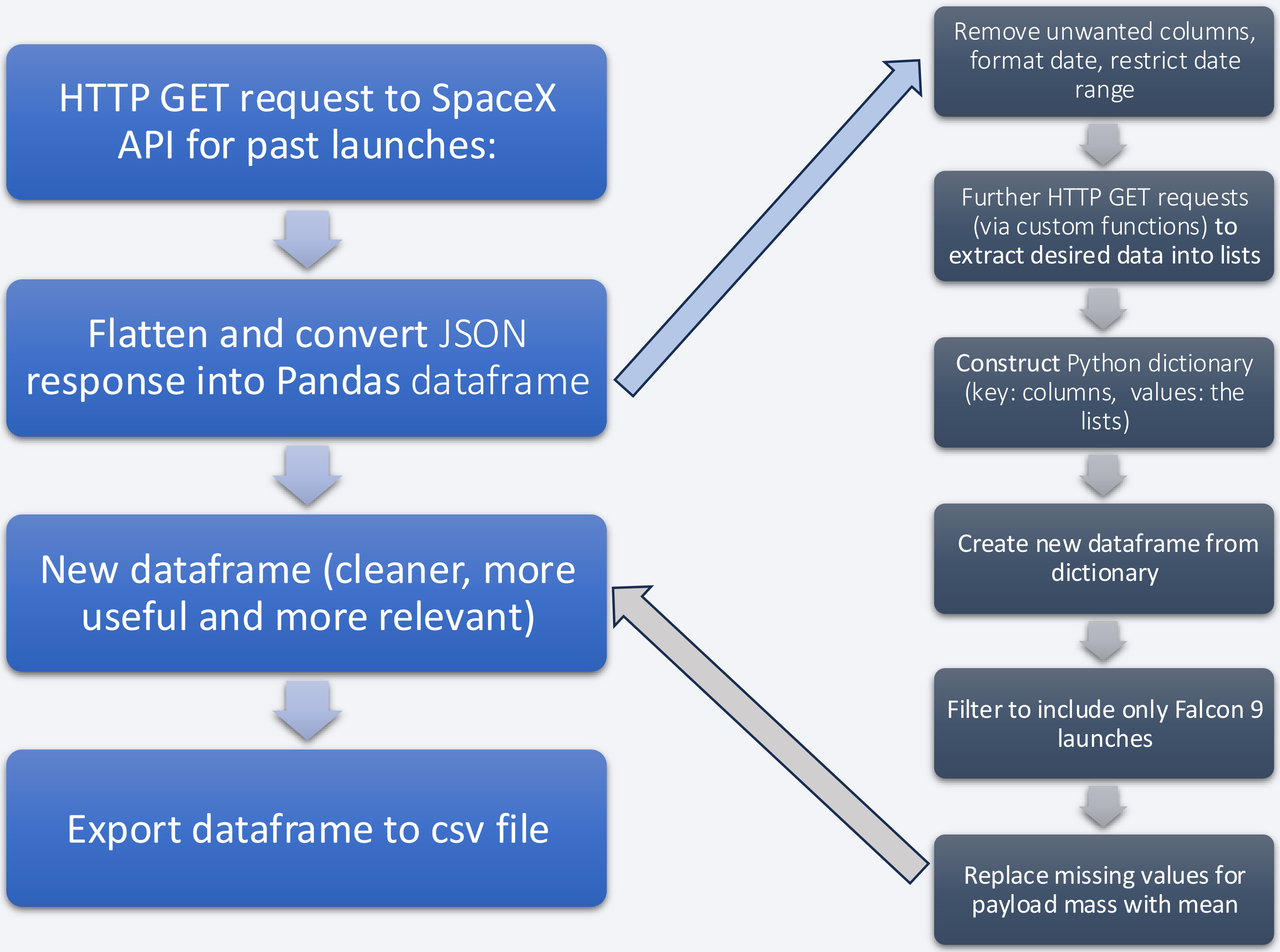
  - Standardize and split data into training and test set. Train model
  - Grid search to find best hyperparameters
  - Determine best model by evaluating accuracy on test data. Inspect confusion matrix for insight

Data Collection

- Two methods were used here - API and web scraping.
- Both begin with an HTTP GET request to retrieve raw data. For API, the response is in JSON format, whereas web scraping returns HTML content that must be parsed to extract relevant information.
- For the API approach, data was first obtained from the "Past launches" API v4 endpoint to form the underlying dataframe. Then the IDs in certain columns were used to reference further API endpoints (rockets, launchpads, payloads, cores) and fetch the corresponding data.
- For webscraping, the data was scraped from a snapshot of the "List of Falcon 9 and Falcon Heavy launches" Wikipage updated on 9th June 2021

Data Collection – SpaceX API

GitHub URL



```
response = requests.get(spacex_url)
```

HTTP GET request to SpaceX API for past launches:

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

Flatten and convert JSON response into Pandas dataframe

```
data_falcon9['PayloadMass'] =
data_falcon9['PayloadMass'].replace(np.nan,
mean_payload)
```

New dataframe (cleaner, more useful and more relevant)

```
data_falcon9.to_csv('dataset_part_1.csv',
index=False)
```

Export dataframe to csv file

Clean, preprocess & wrangle data

Remove unwanted columns, format date, restrict date range

Further HTTP GET requests (via custom functions) to extract desired data into lists

Construct Python dictionary (key: columns, values: the lists)

Create new dataframe from dictionary

Filter to include only Falcon 9 launches

Replace missing values for payload mass with mean

Dataframe does not yet contain desired columns. But existing columns provide "reference" to desired data on other SpaceX APIs.

Data Collection - Scraping

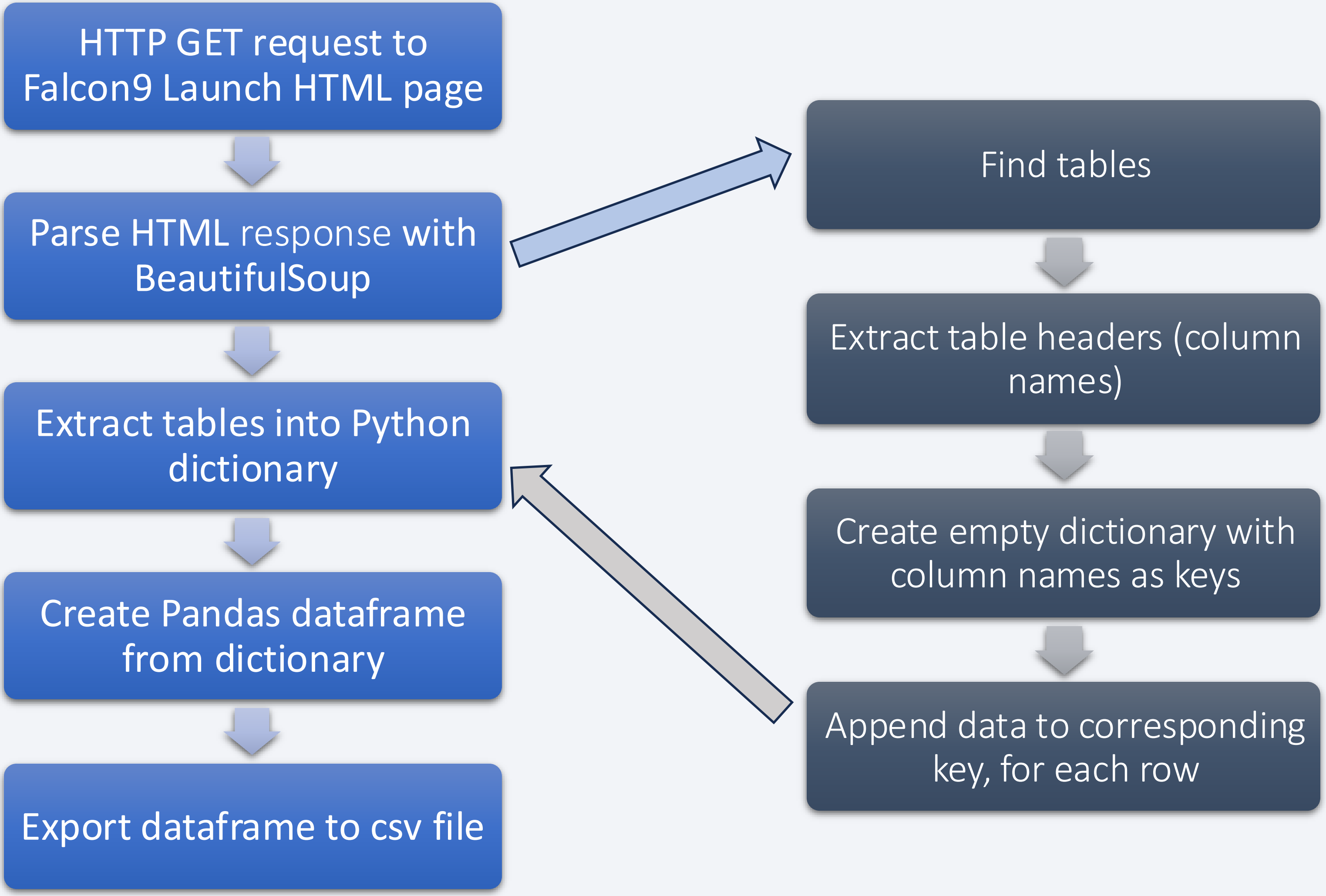
GitHub URL

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

```
soup
= BeautifulSoup(response.text, 'html.parser')
```

```
df=
pd.DataFrame( { key:pd.Series(value) for key, value in
launch_dict.items() } )
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```



HTTP GET request to Falcon9 Launch HTML page

Parse HTML response with BeautifulSoup

Extract tables into Python dictionary

Create Pandas dataframe from dictionary

Export dataframe to csv file

Find tables

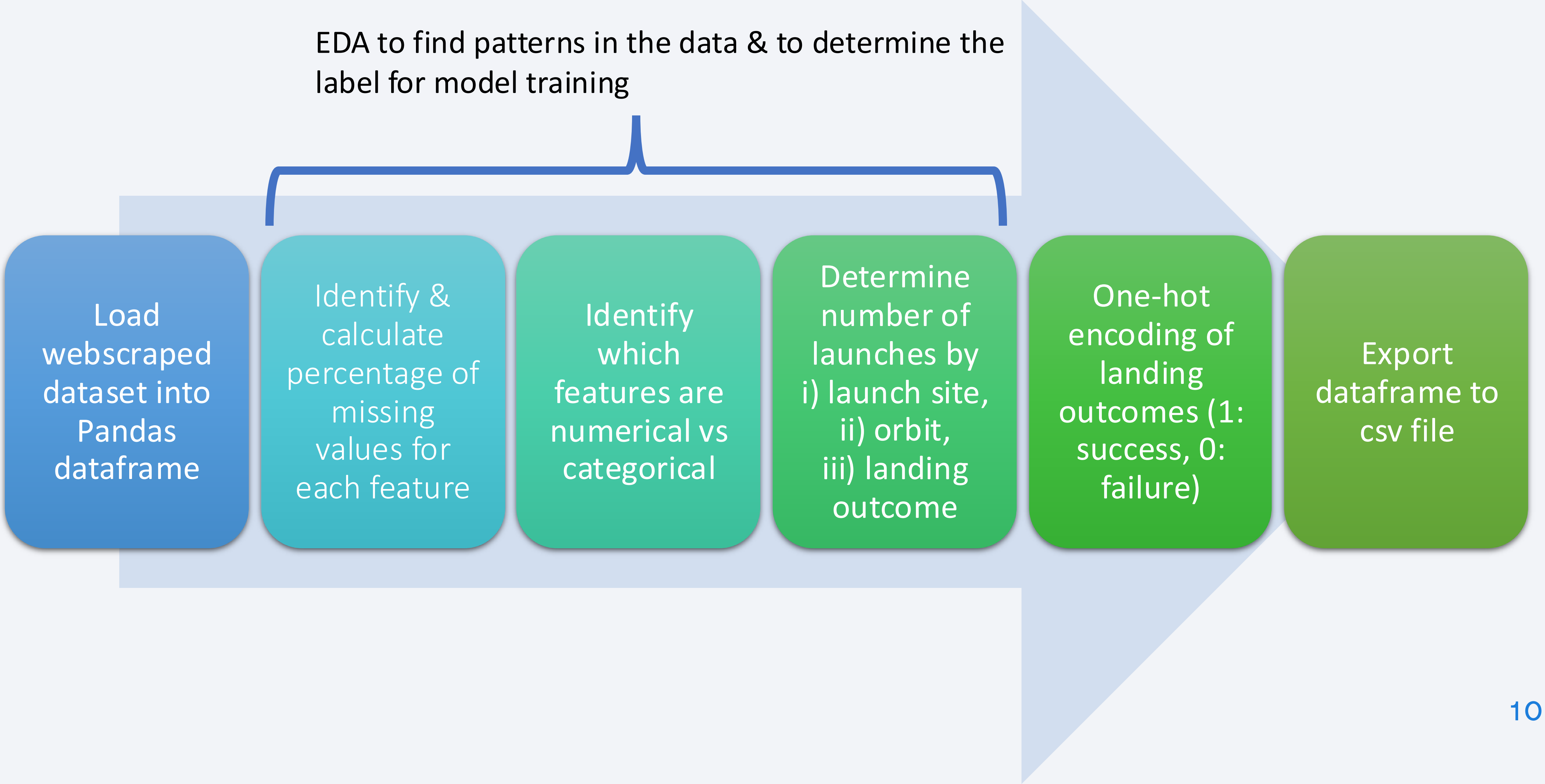
Extract table headers (column names)

Create empty dictionary with column names as keys

Append data to corresponding key, for each row

Data Wrangling

GitHub URL



EDA to find patterns in the data & to determine the label for model training

1. Load webscraped dataset into Pandas dataframe
2. Identify & calculate percentage of missing values for each feature
3. Identify which features are numerical vs categorical
4. Determine number of launches by i) launch site, ii) orbit, iii) landing outcome
5. One-hot encoding of landing outcomes (1: success, 0: failure)
6. Export dataframe to csv file

EDA with Data Visualization

GitHub URL

Features investigated: **Flight number** (corresponds to **time**), **Payload mass**, **Launch site**, **Orbit**
Label: **Landing success/ success rate**

Scatter plot:

FlightNumber vs. **PayloadMass**

FlightNumber vs. **Launch site**

FlightNumber vs. **Orbit**

Examine how mission characteristics (payload mass, launch site, orbit) evolve over time, in terms of landing success rate, frequency and distribution.

Payload mass vs. **Launch site**

Payload mass vs. **Orbit**

Examine how landing success rate relate to payload mass for different launch sites and orbits.

Bar plot: **Success rate** vs. **Orbit**

- To visualize which orbit types had the highest or lowest landing success rates.

Line plot: **Year** vs **Success rate**

- To track how the overall success rate improved each year.

EDA with SQL

GitHub URL

EDA was also carried out by executing SQL queries to:

- **Display the names of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA'**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first succesful landing outcome in ground pad was achieved.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List all the booster_versions that have carried the maximum payload mass.**
- **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
- **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.**

Build an Interactive Map with Folium

GitHub URL

The following map objects and plugins were created and added to the folium map:

- Circle: folium.Circle()

  Marks the location of each launch site with a circle; Popup label (folium.Popup()) shows the names of the launch sites when circle is clicked.

- Markers: folium.Marker()

  Marker displayed as pin if takes argument icon=folium.icon(); as text if icon=folium.DivIcon()

- Lines: folium.PolyLine()

  Create a line between launch site and points of interest

- MousePosition plugin: MousePosition()

  Shows current latitude and longitude of mouse cursor as you hover over the map. Needed to obtain the coordinates of points of interest.

- Marker clusters: MarkerCluster()

  Multiple launch records are associated with each launch site --> overlapping markers; Less messy when organized into clusters.

Build a Dashboard with Plotly Dash

GitHub URL

For interactive visualization, a dashboard containing the following interactive components was created:

- **Dropdown list:**

  Enables either selection of All Sites or a specific launch site

- **Pie chart:**

  If All Sites is selected - shows composition of successful launches by launch site.

  If a specific launch site was selected - shows the composition of successful vs. failed launches for the particular site.

- **Scatter chart:**

  Plots payload mass vs outcome class (0: failure, 1: success) by booster version category; shows correlation between payload mass and launch success

- **Slider:**

  Allows interactive adjustment of the payload mass from the scatter chart

Predictive Analysis (Classification)

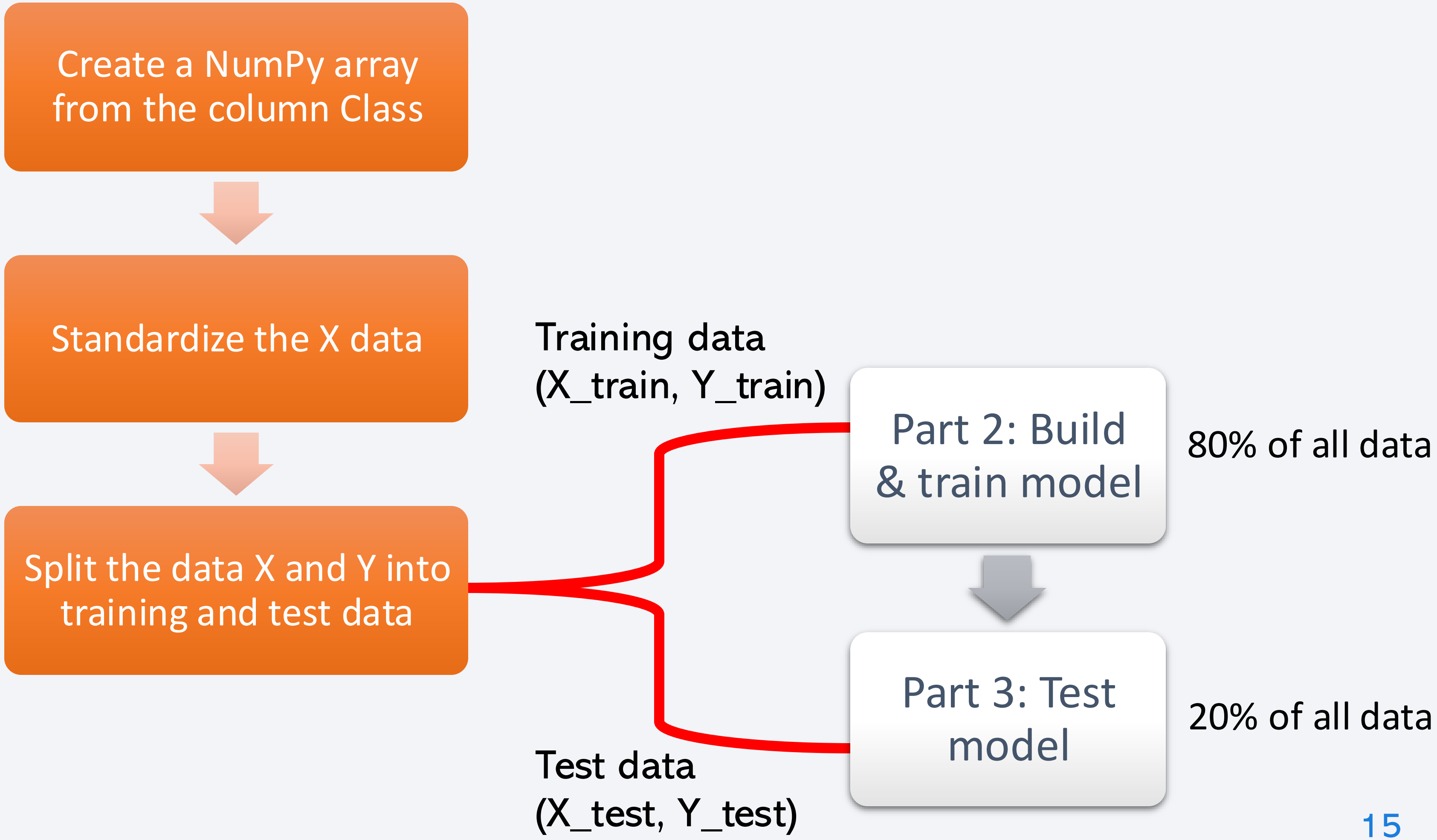
GitHub URL

Part 1: Preparing the data

```
Y = data['Class'].to_numpy()
```

```
transform =
preprocessing.StandardScaler()
X = transform.fit_transform(X)
```

```
X_train, X_test, Y_train, Y_test
= train_test_split(X, Y, test_size=0.2, random_state=2)
```



Create a NumPy array from the column Class

Standardize the X data

Split the data X and Y into training and test data

Training data (X_train, Y_train)

Test data (X_test, Y_test)

Part 2: Build & train model

80% of all data

Part 3: Test model

20% of all data

Predictive Analysis (Classification)

GitHub URL

Part 2: Build & train model

e.g. for SVM:

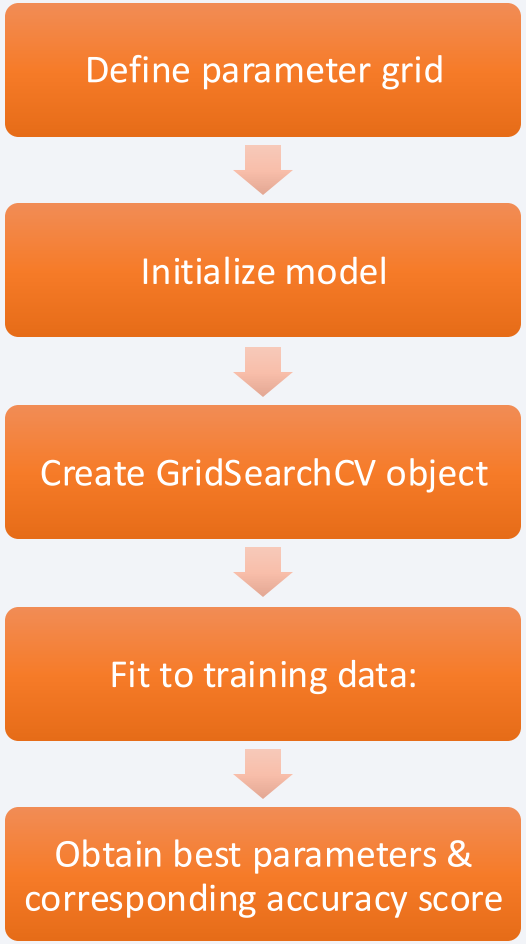
```
parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
              'C': np.logspace(-3, 3, 5), 'gamma':np.logspace(-3, 3, 5)}
```

```
svm = SVC()
```

```
svm_cv = GridSearchCV(svm, parameters, cv=10, scoring='accuracy')
```

```
svm_cv.fit(X_train, Y_train)
```

```
svm_cv.best_params_
svm_cv.best_score_
```



Define parameter grid

Initialize model

Create GridSearchCV object

Fit to training data:

Obtain best parameters & corresponding accuracy score

Predictive Analysis (Classification)

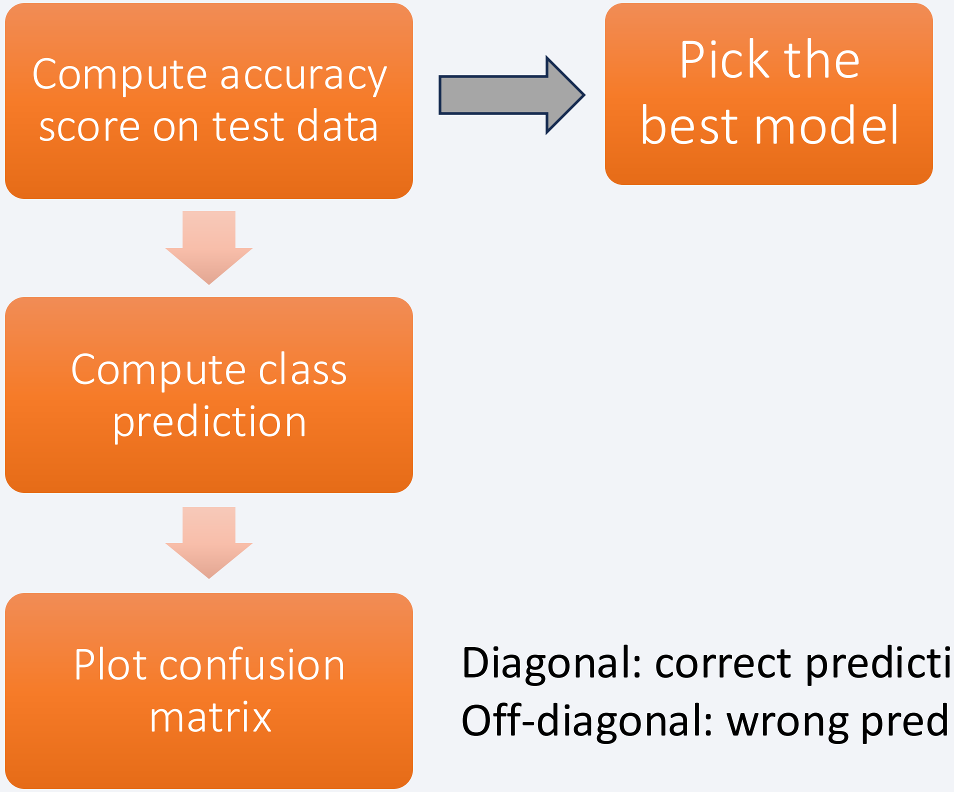
GitHub URL

Part 3: Test model on test (unseen) data

```
acc_svm = svm_cv.score(X_test,Y_test)
```

```
yhat=svm_cv.predict(X_test)
```

```
plot_confusion_matrix(Y_test,yhat)
```



Compute accuracy score on test data

Pick the best model

Compute class prediction

Plot confusion matrix

Diagonal: correct prediction,
Off-diagonal: wrong prediction

Results

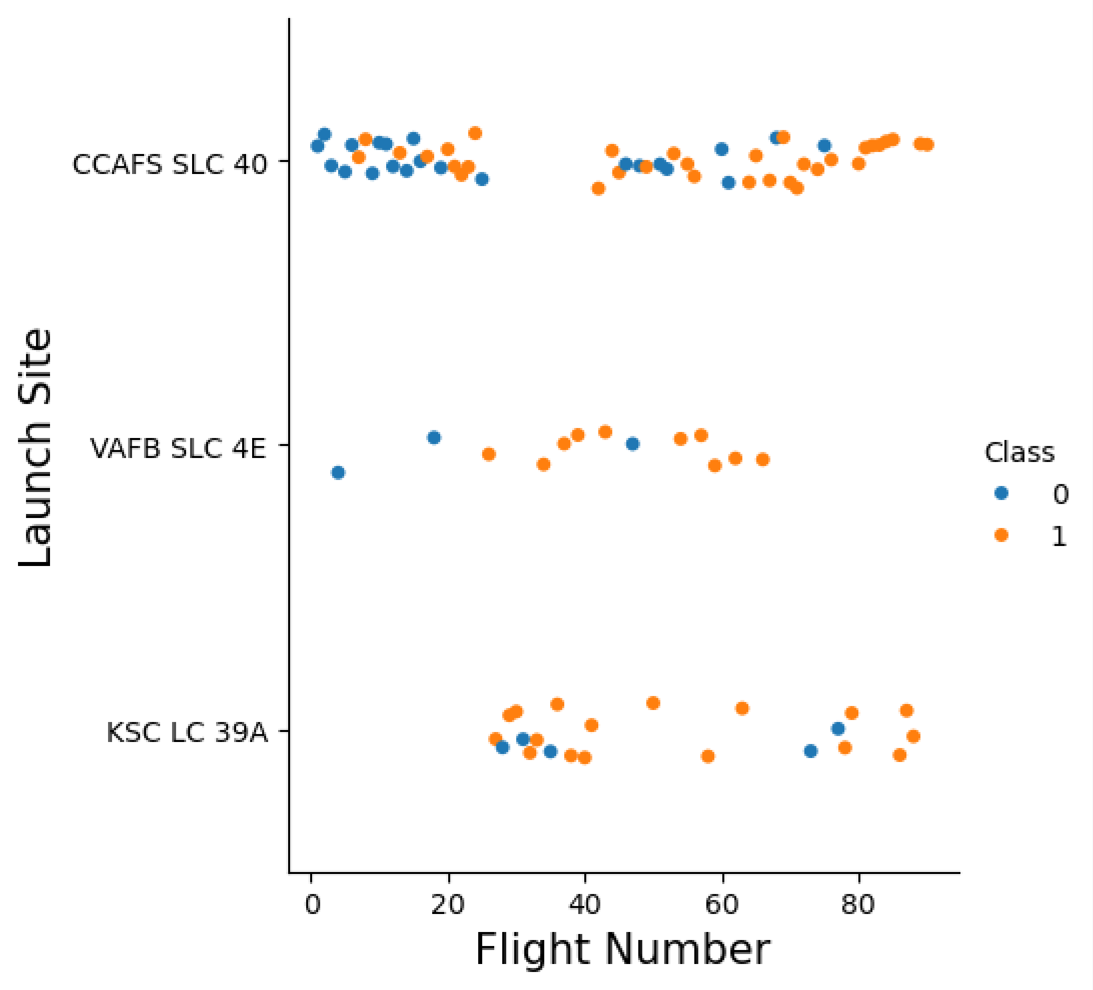
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

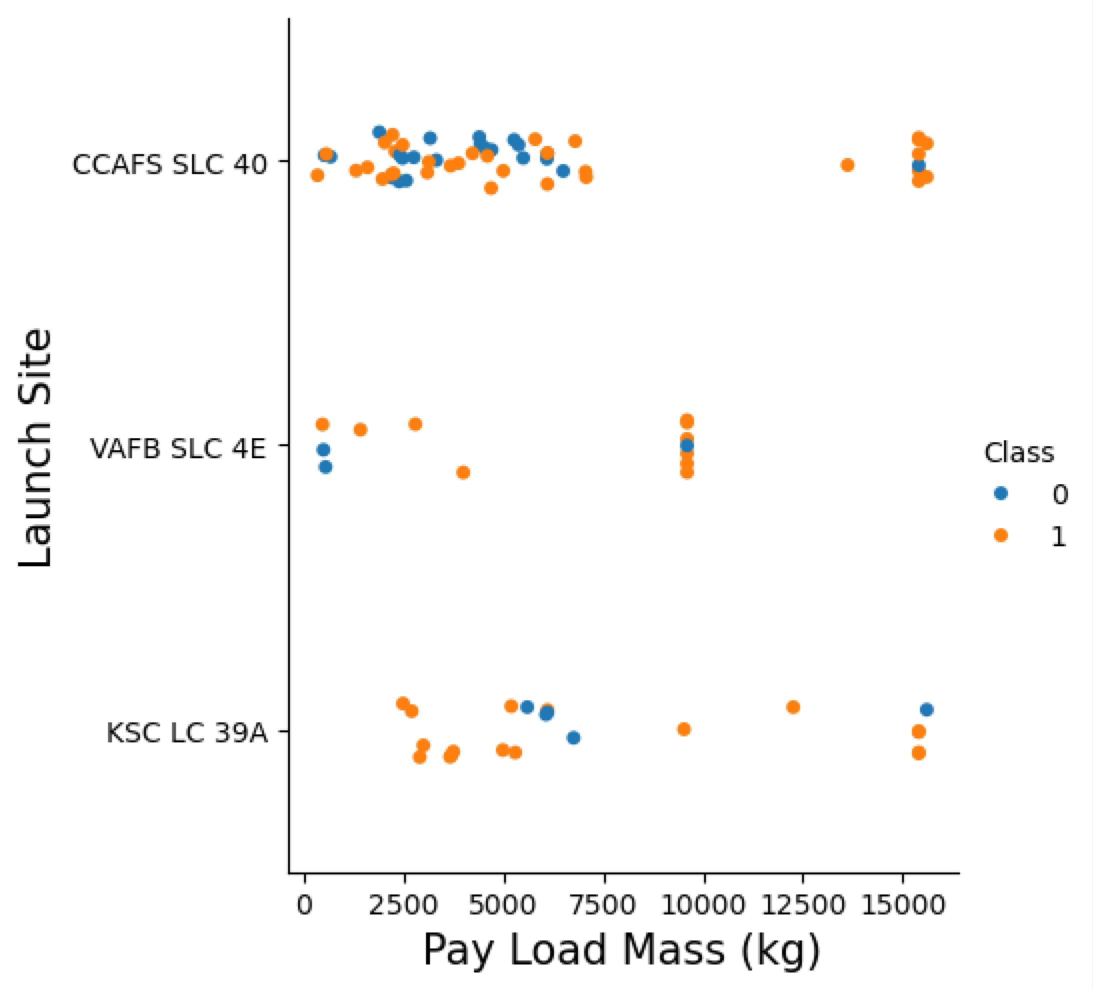
Flight Number vs. Launch Site

- Success rate improves over time (as reflected by the flight number) for all sites. This is likely due to improved technology/processes and increased experience.
- Usage differ between launch sites:
  - ➢ CCAFS SLC 40: Used consistently throughout flight history
  - ➢ VAFB SLC 4E: Fewer total launches here.
  - ➢ KSC LC 39A: Used more in later phases.



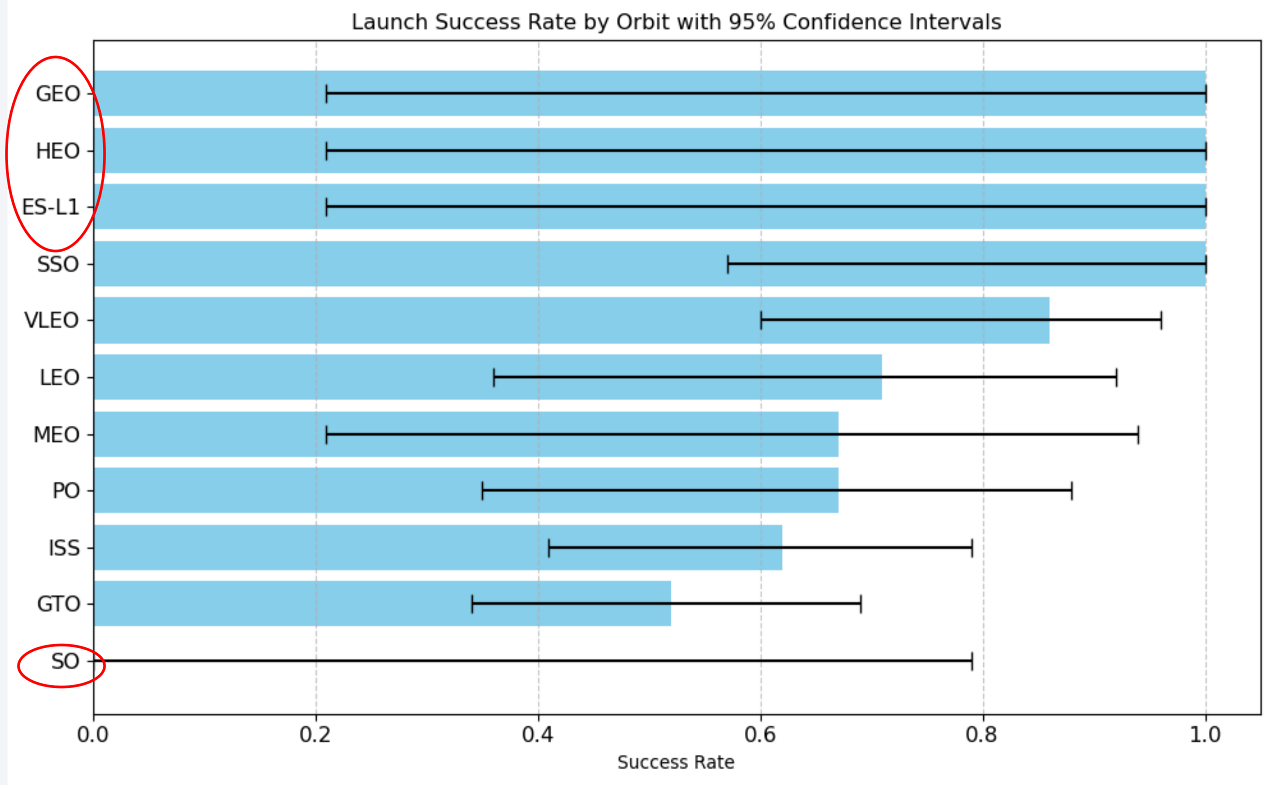
Payload vs. Launch Site

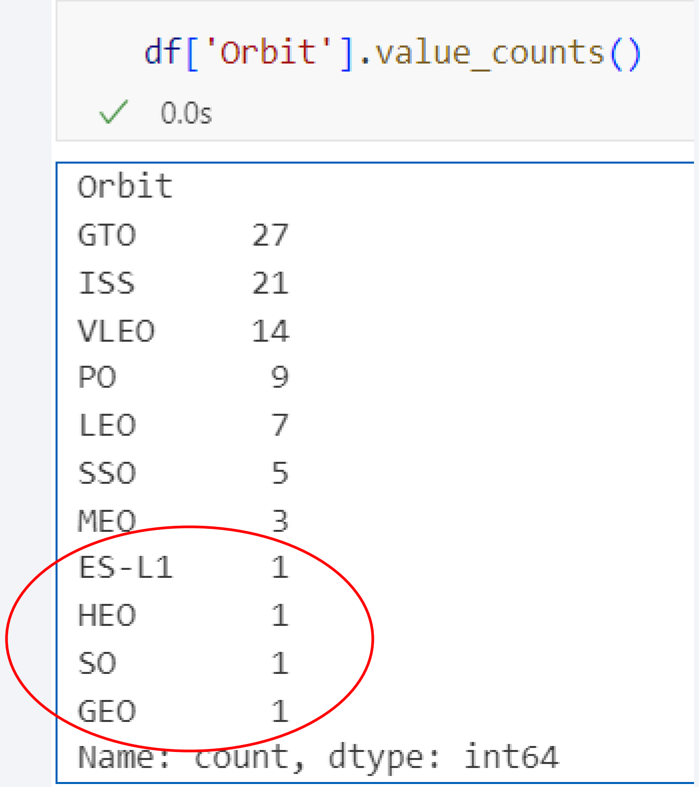
- No strong correlation between payload mass and launch success.
- However, heavier payloads tend to succeed more often. This is perhaps because heavier payloads tend to be associated with more mature rockets and careful execution.
- VAFB-SLC 4E is used mostly for lower payloads, with no rockets launched for mass greater than 10000kg.
- CCAFS SLC 40 is the most active launch site but has the most mixed outcome particularly for lower to moderate pay load mass.



Success Rate vs. Orbit Type

- Note: Sample size differences must be accounted for via confidence intervals (here Wilson Score Intervals were used), as it can affect how reliable our inference is, e.g. four of the orbits have a sample size of only 1, which is very statistically unreliable. Taking this into account:
- SSO, followed by VLEO are the orbits which give the most (statistically meaningful) success rates
- The GTO orbit has the worst success rate





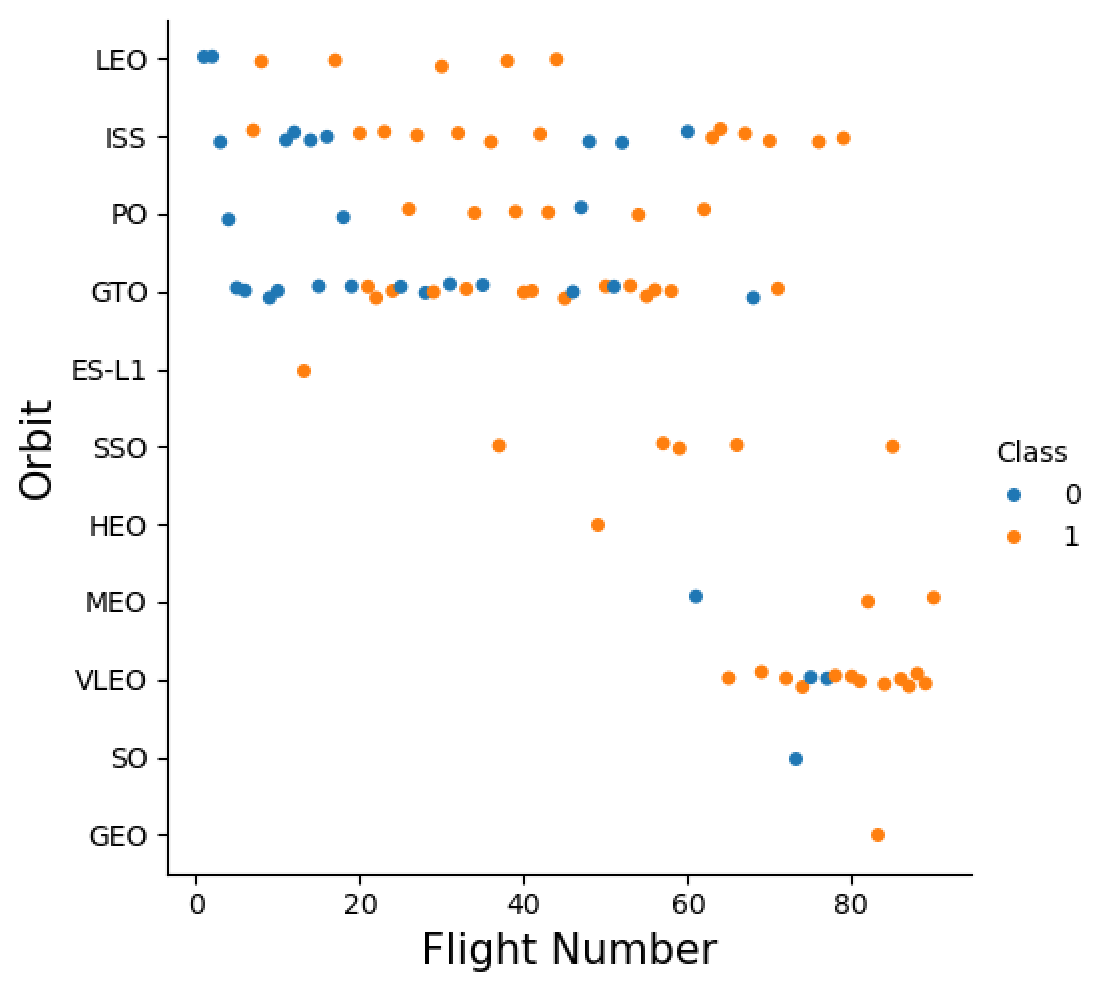
```
df['Orbit'].value_counts()
```

✓ 0.0s

```
Orbit
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: count, dtype: int64
```

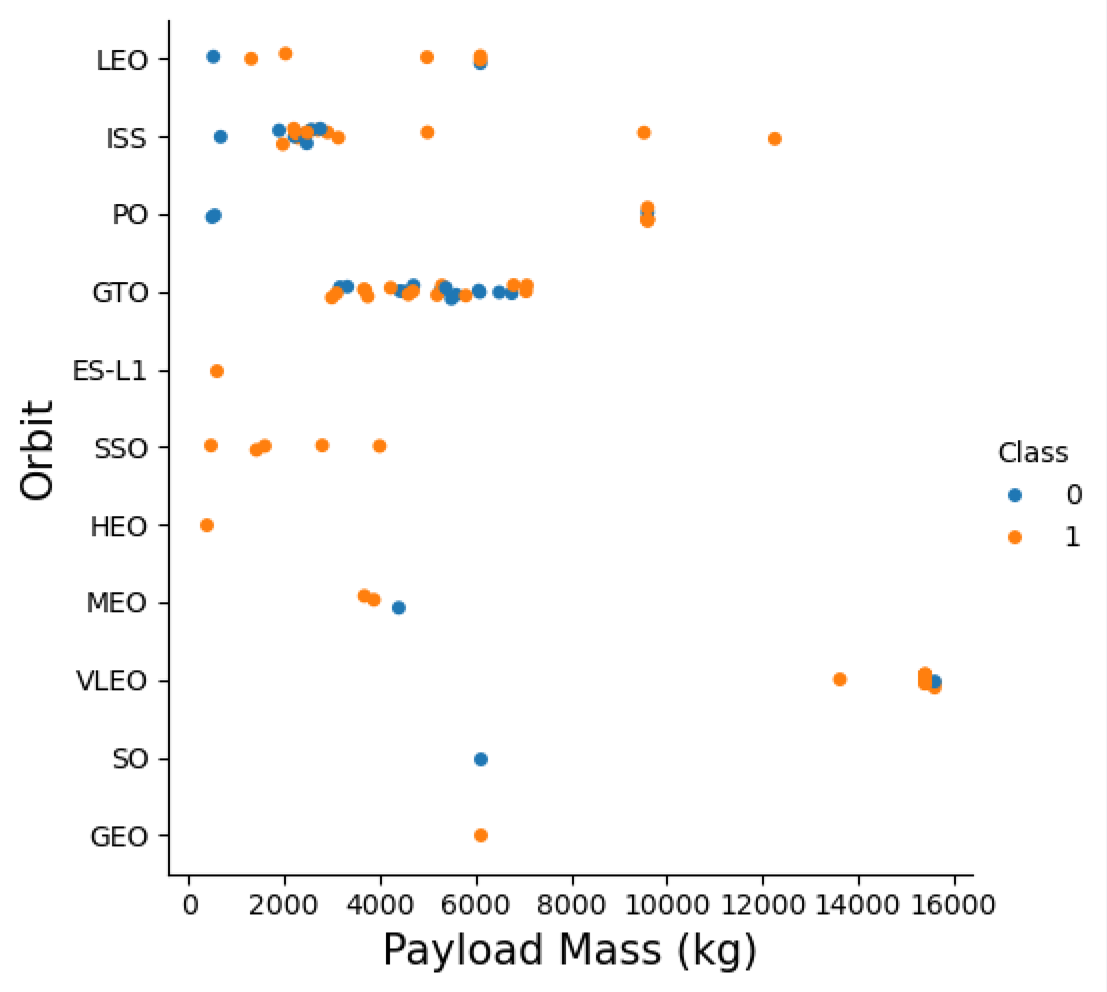
Flight Number vs. Orbit Type

- LEO, ISS and PO: success rate increased with time (flight number)
- GTO orbit: no clear relationship between success rate and time.
- VLEO: landings have been consistently successful
- SSO, SO, HEO, MEO, GEO, ES-L1: the number of data points are too small to make any conclusions.
- Not all orbits began being launched at the same time.



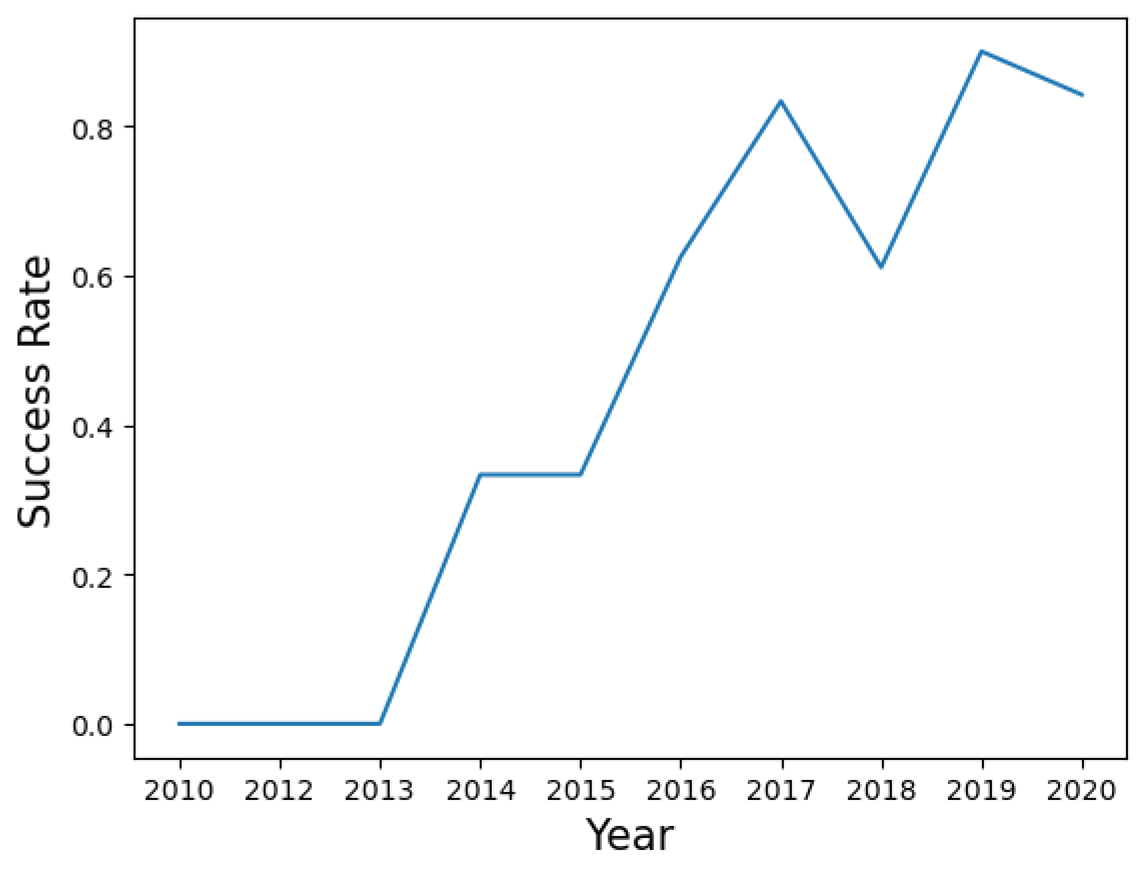
Payload vs. Orbit Type

- The improvement in landing success rate with payload mass seems to be the most evident for POLAR, LEO and ISS.
- For GTO, there is no clear relationship between payload mass and landing success (balanced mix between success and failure)
- This similarity in trend with the previous chart (Flight number vs. Orbit type) suggests that improvement in success rate with payload mass may be due to improvement over time rather than the increased payload mass itself.
- Heavier payloads tend to be launched later with improved technology, careful planning/testing etc.
- Payload mass distribution varies by orbit.



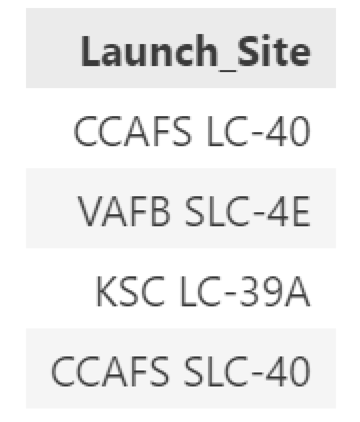
Launch Success Yearly Trend

- 2010–2013: all landings failed
- From 2013, the success rate has been generally increasing (except for a plateau around 2014, a dip in 2018 and a slight decrease around 2019 – 2020)
- Overall positive trend in landing success over time.



All Launch Site Names

- In this slide and the following next 9 slides, results of SQL queries will be presented.
- There are four launch sites in the space mission, as listed below:

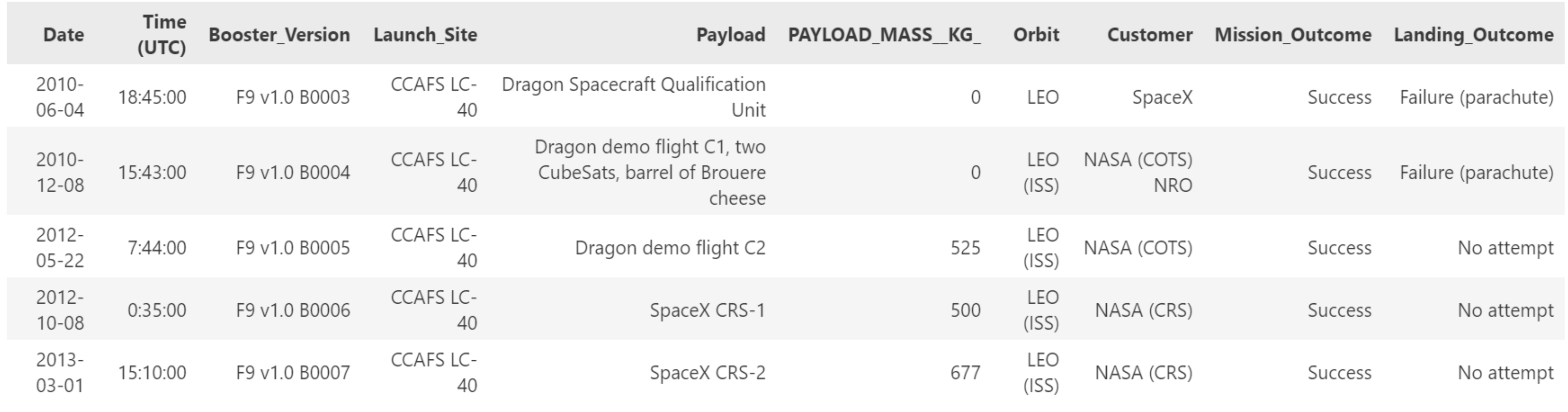


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- CCAFS LC-40: Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40: Cape Canaveral Air Force Station Space Launch Complex 40
- KSC LC-39A: Kennedy Space Center Launch Complex 39A
- VAFB SLC-4E: Vandenberg Space Force Base Space Launch Complex 4E

Launch Site Names Begin with 'CCA'

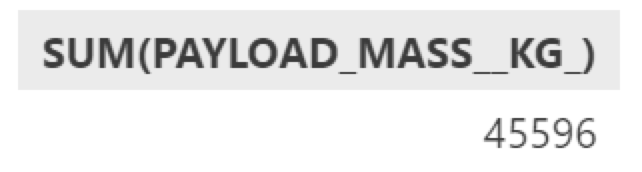
- The first five records where launch sites begin with "CCA" are shown below:



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

Total Payload Mass

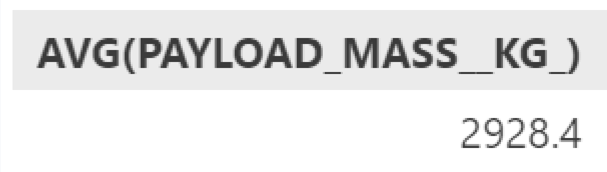
- The total payload carried by boosters from NASA is 45596 kg:



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

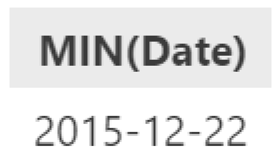
- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg:



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

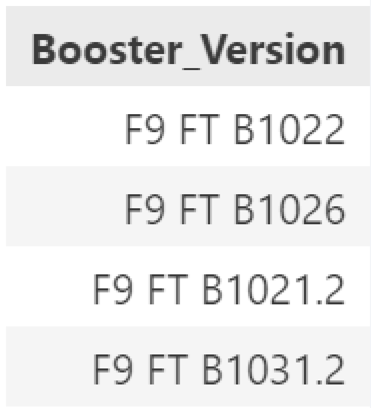
- The first successful landing outcome on a ground pad took place on 22 December 2015:



MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg

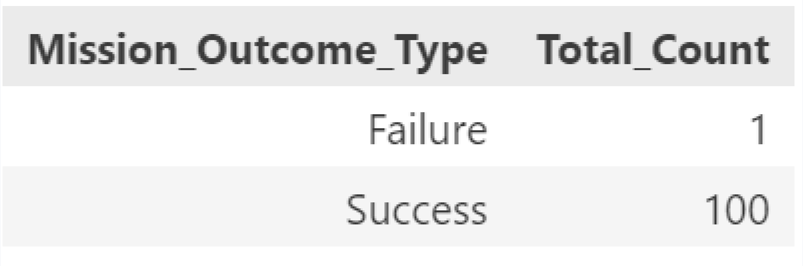
- The names of boosters which have successfully landed on drone a ship and where 4000 kg < payload mass < 6000 kg are given as follows:



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

Total Number of Successful and Failure Mission Outcomes

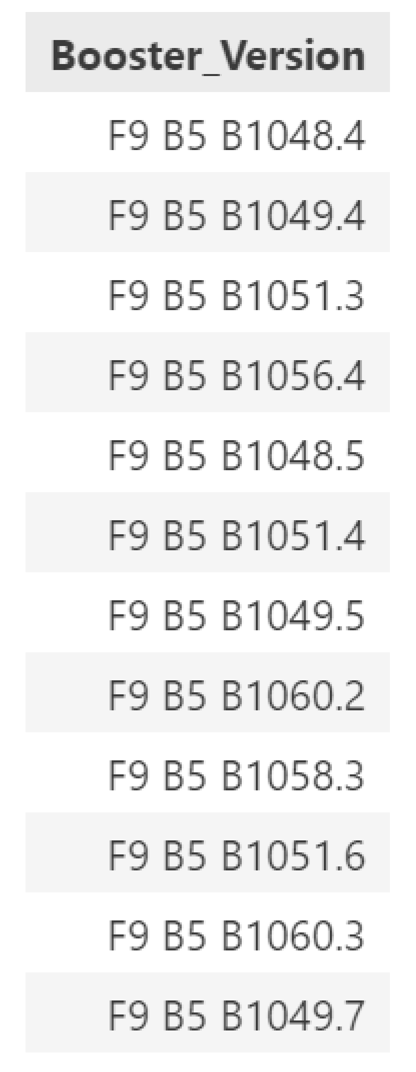
- The total number of successful and failed mission outcomes is displayed below, showing only one failed mission out of a total of 101 missions.



Mission_Outcome_Type	Total_Count
Failure	1
Success	100

Boosters Carried Maximum Payload

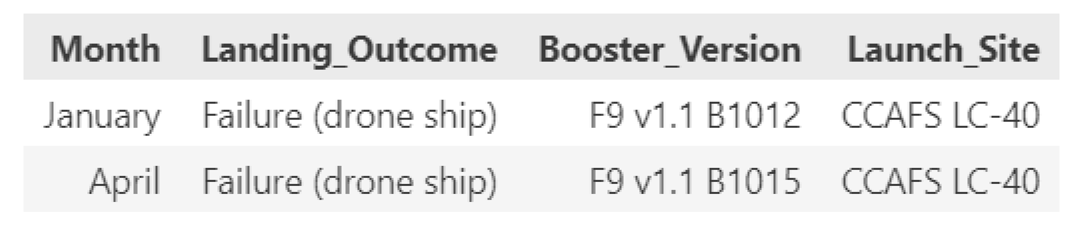
- The names of the boosters which have carried the maximum payload mass are listed here:



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

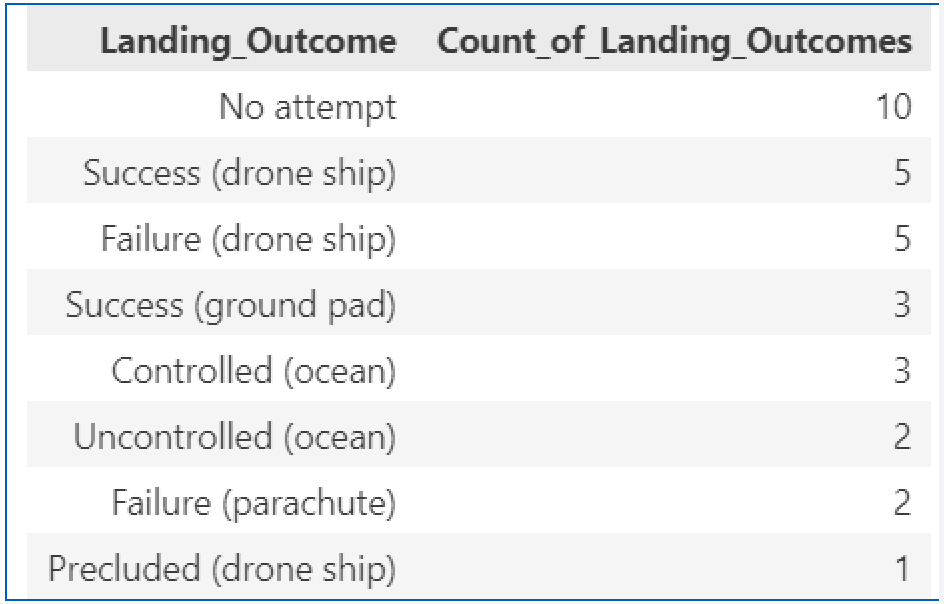
- In 2015, there were two failed landing outcomes on a drone ship - the first occurred in January and the second in April. Both were associated with the CCAFS LC-40 launch site:



Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The landing outcomes in the period from 2010-06-04 to 2017-03-20, arranged in descending order in terms of number of occurrences are listed on the right:
- We can further group the data into separate outcomes where the first stage was reusable/not reusable. It appears the split is quite even:
  - ➢ Reusable (sum of successes): 8
  - ➢ Non-reusable (sum of failures + Uncontrolled): 9
  - ➢ N/A (No attempt + Precluded): 14

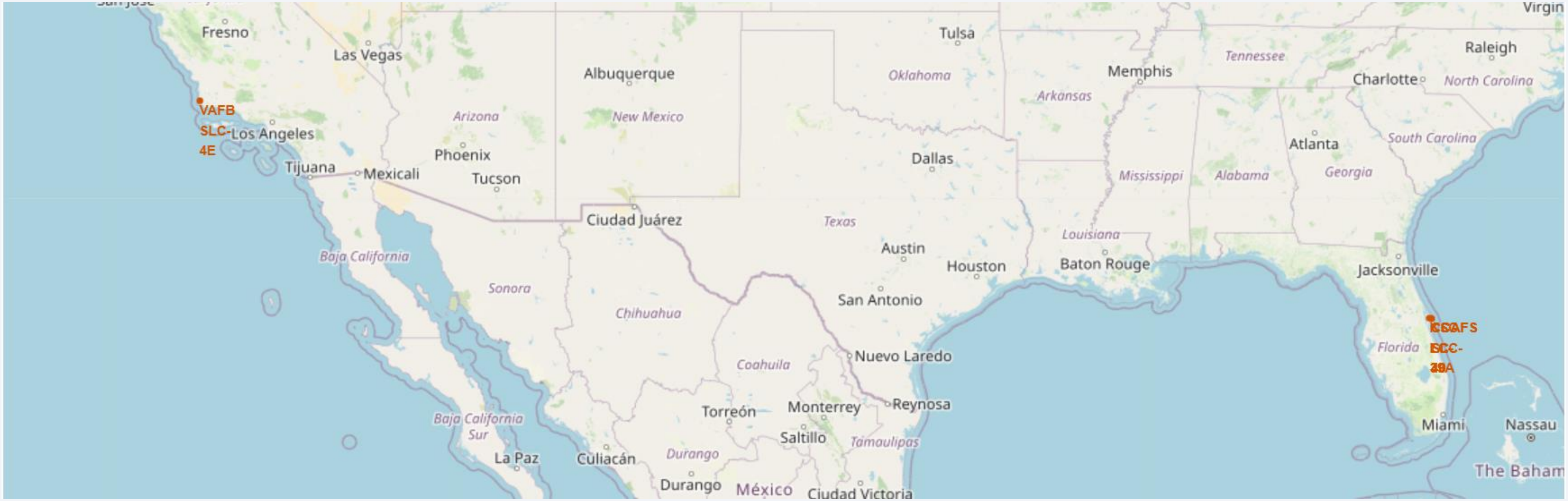


Landing_Outcome	Count_of_Landing_Outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

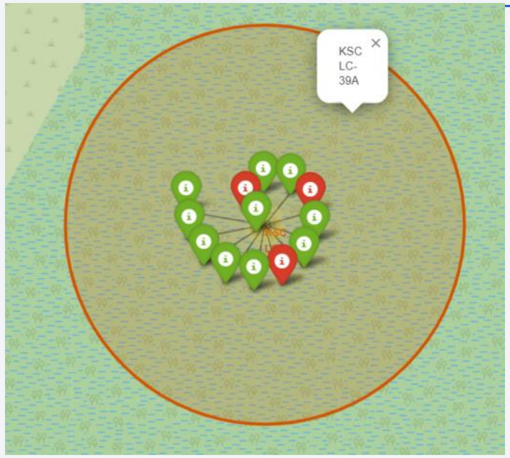
Launch Sites Proximities Analysis

Launch site locations

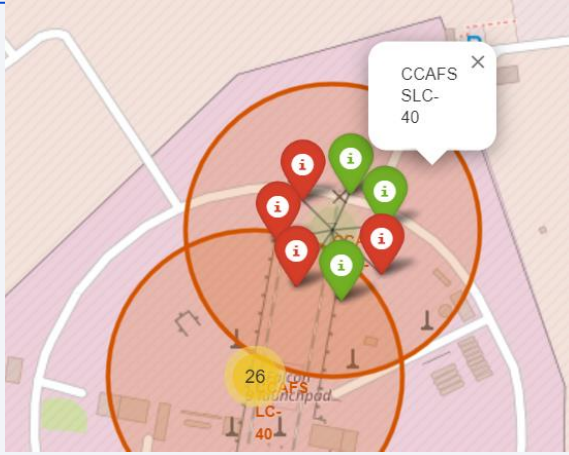


- All the launch sites lie in very close proximity to the coast, since the rocket launches take place over the ocean.
- Because the earth's rotation is fastest at the equator, rocket launches close to the equator get extra boost. However this benefit applies only to eastward launches due to the eastward rotation of the earth. Thus only the launch sites on the eastern coast are built close to the equator.

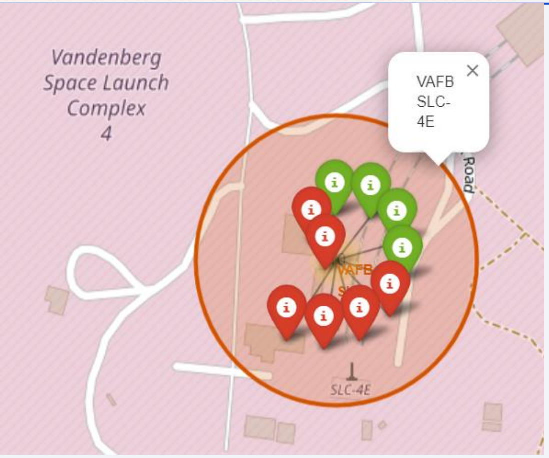
Success rates by launch site



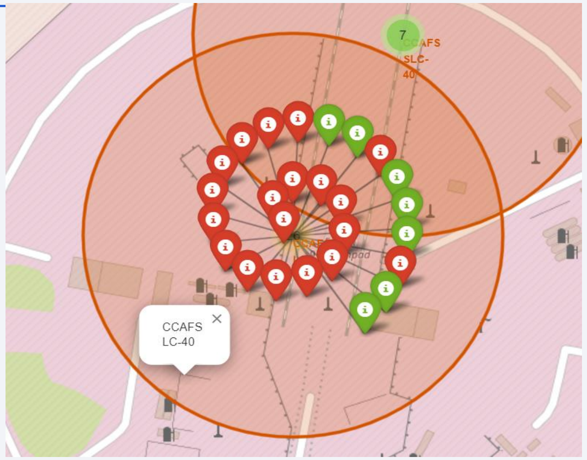
Best: KSC LC-39A (76.92%)



CCAFS SLC-40 (42.86%)



VAFB SLC-4E (40%)



Worst: CCAFS LC-40 (26.92%)

- Launch outcome success rates differ by launch site:

KSC LC-39A has the best launch outcome success rate of almost 77%. However all the other launch sites have success rates of less than 50%, with CCAFS LC-40 having a success rate of only about 27%.

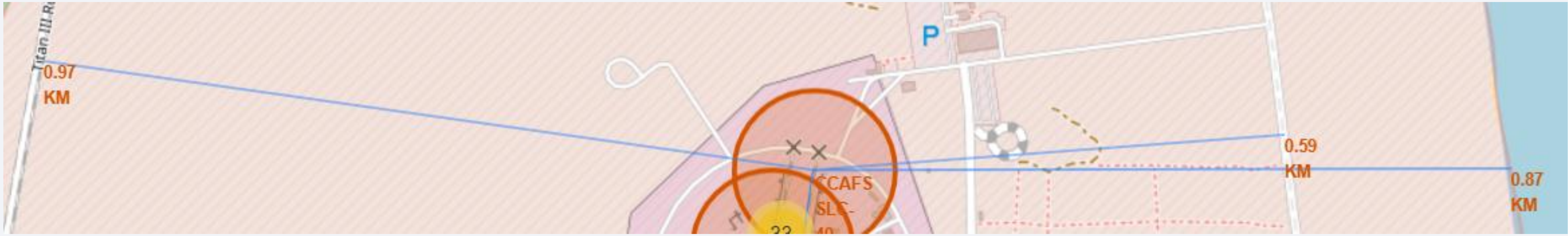
Proximities of Launch Sites

- Launch sites tend to be in close proximity to railways, highways and coastlines. Example: < 1 km to the nearest railway, highway and coastline in the case of CCAFS SLC-40
- Explanation:

Railways: Railways are used to transport heavy payloads, rocket components, and fuel

Highways: Facilitates access for logistics as well as for personnel, supply trucks, and emergency services

Coastlines: Rockets are launched over the ocean to minimize risk to civilians, easier trajectory planning.



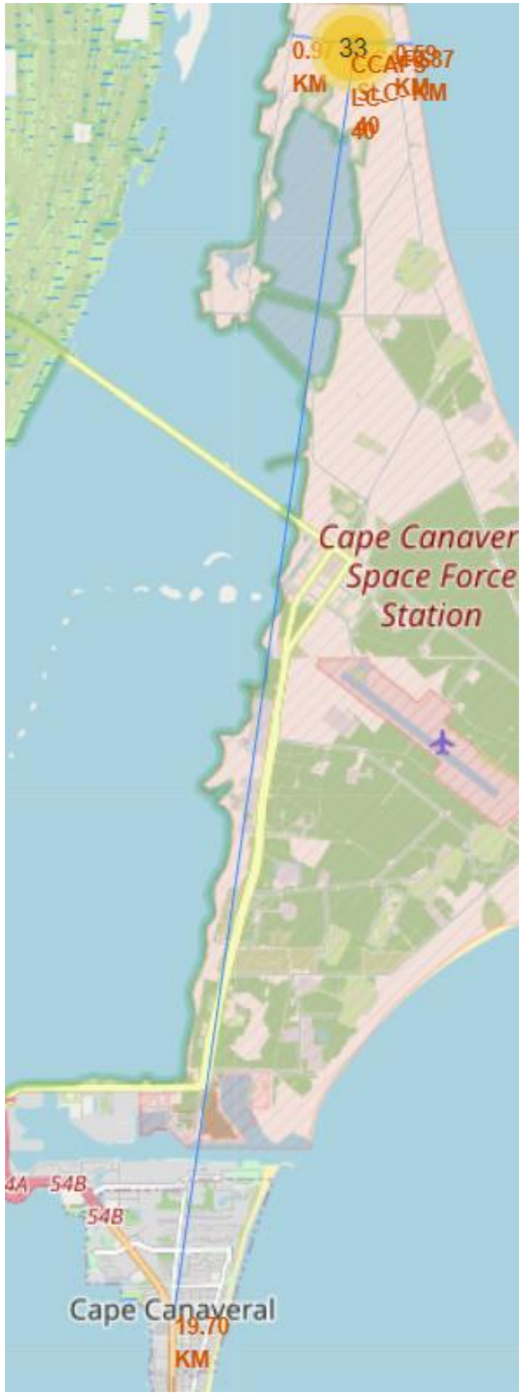
In contrast…

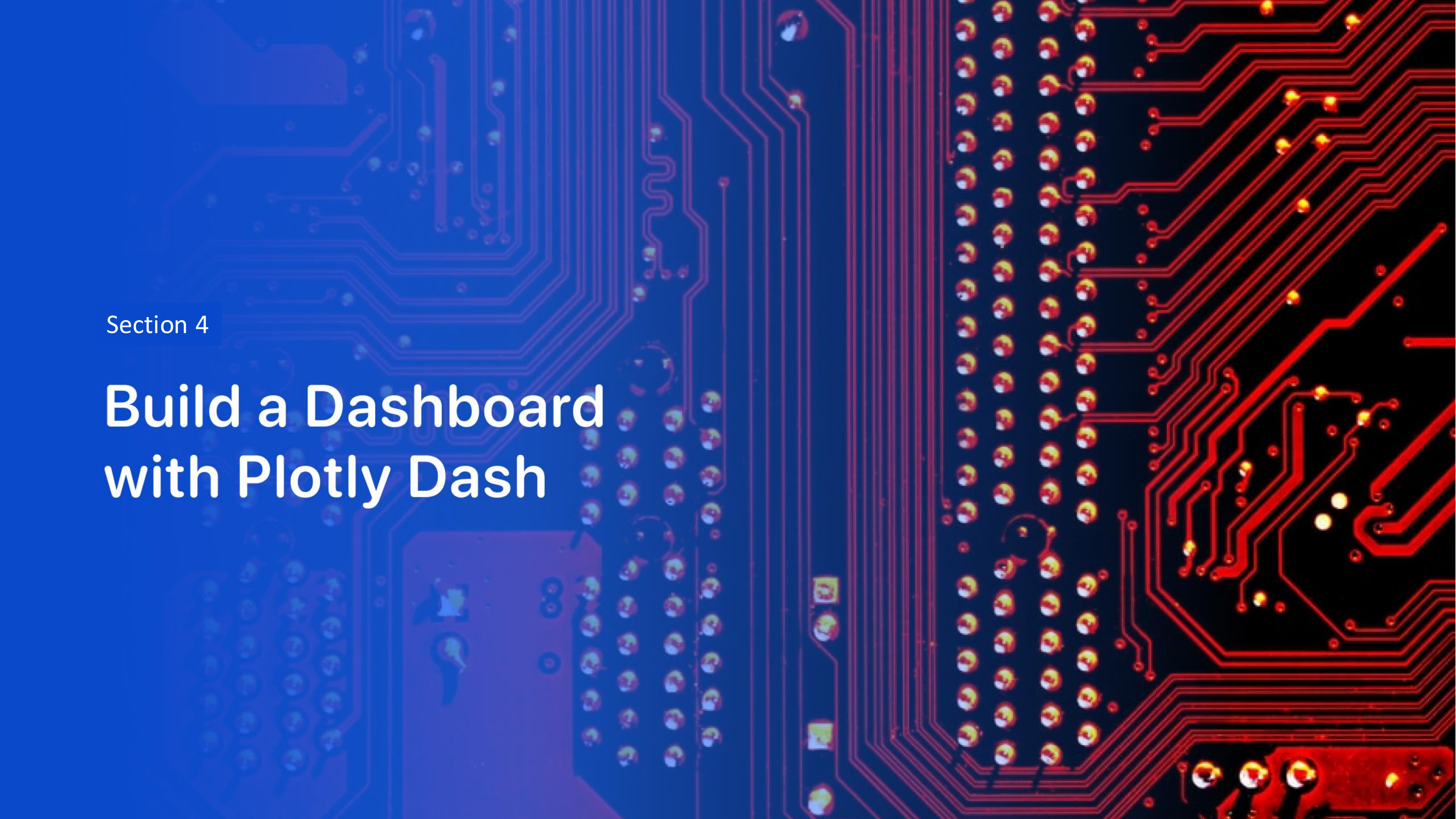
- Launch sites tend to be located far away from cities. Example (CCAFS SLC-40): Distance to the nearest city is > 20 times the distance to the nearest railway, highway and coastline
- Explanation:

Safety: launches involve massive energy and risk of explosions or debris.

Reduces noise pollution and shockwave impact on civilians.

Allows for restricted airspace and land zones without interfering with daily life.



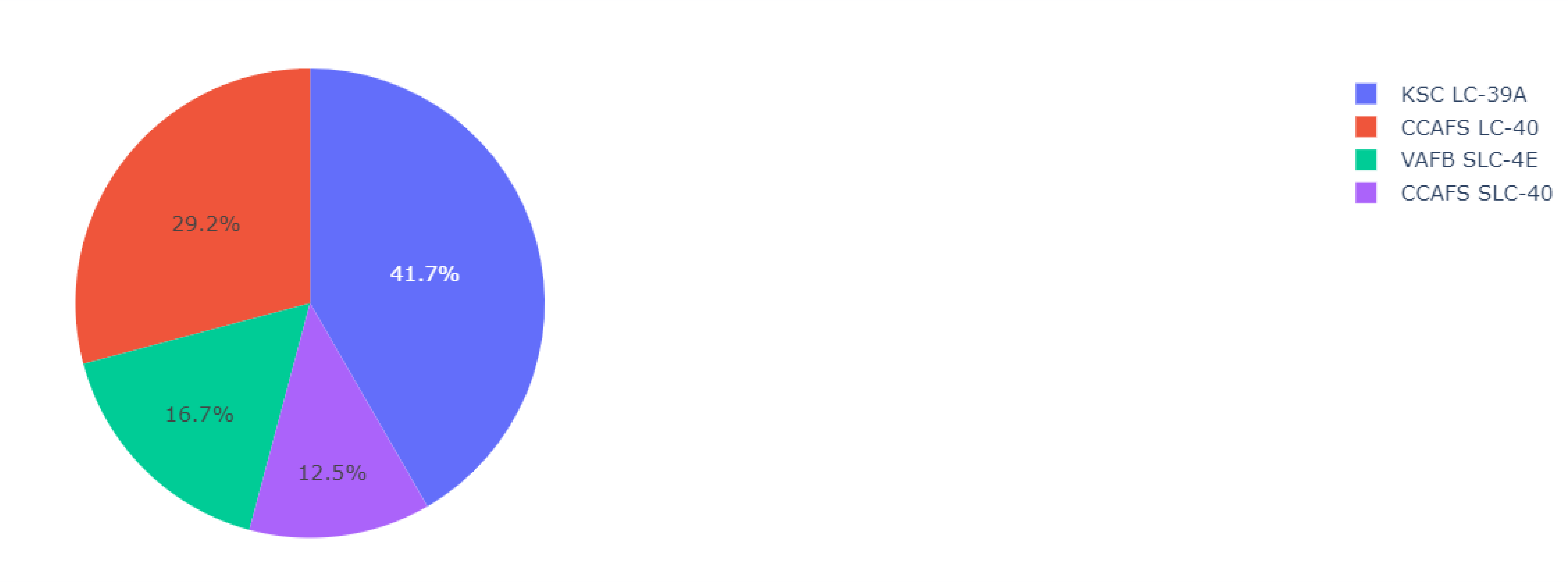


Section 4

Build a Dashboard with Plotly Dash

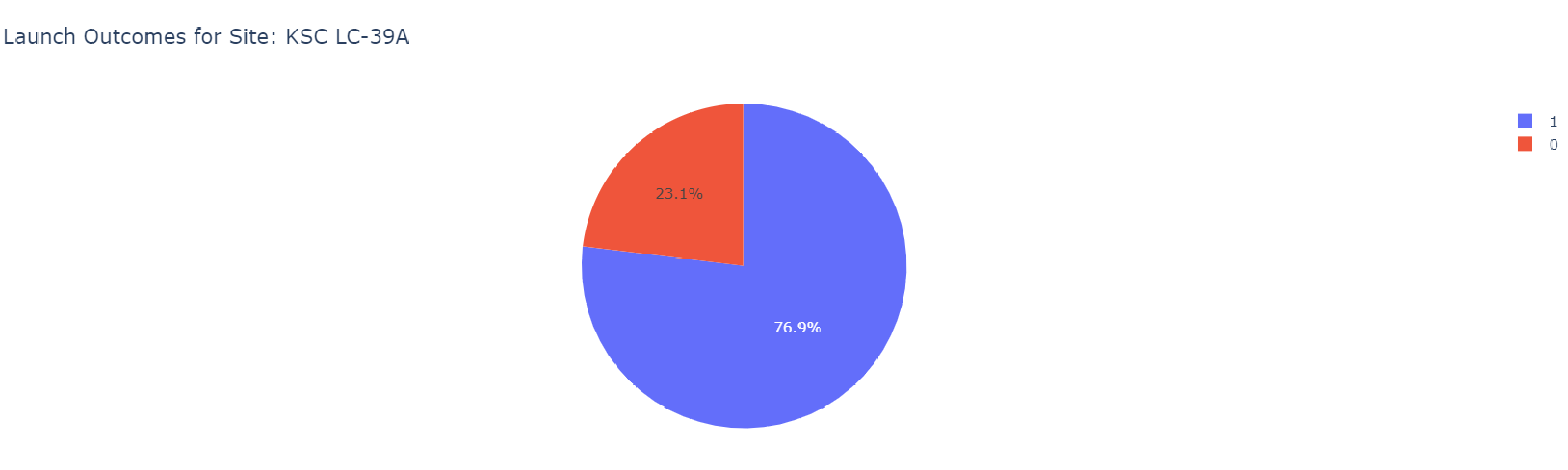
Percentage of Successful Launches by Site

- The pie chart shows the contribution to successful launches from each launch site.
- The largest share of successful launches come from KSC LC-39A.
- CCAFS SLC-40 has the fewest successful launches.



Launch Outcomes of the Most Successful site

- The following pie chart shows that 76.9% of the launches at the most successful site, KSC LC-39A, were successful (in agreement with the success rate from the Folium section).



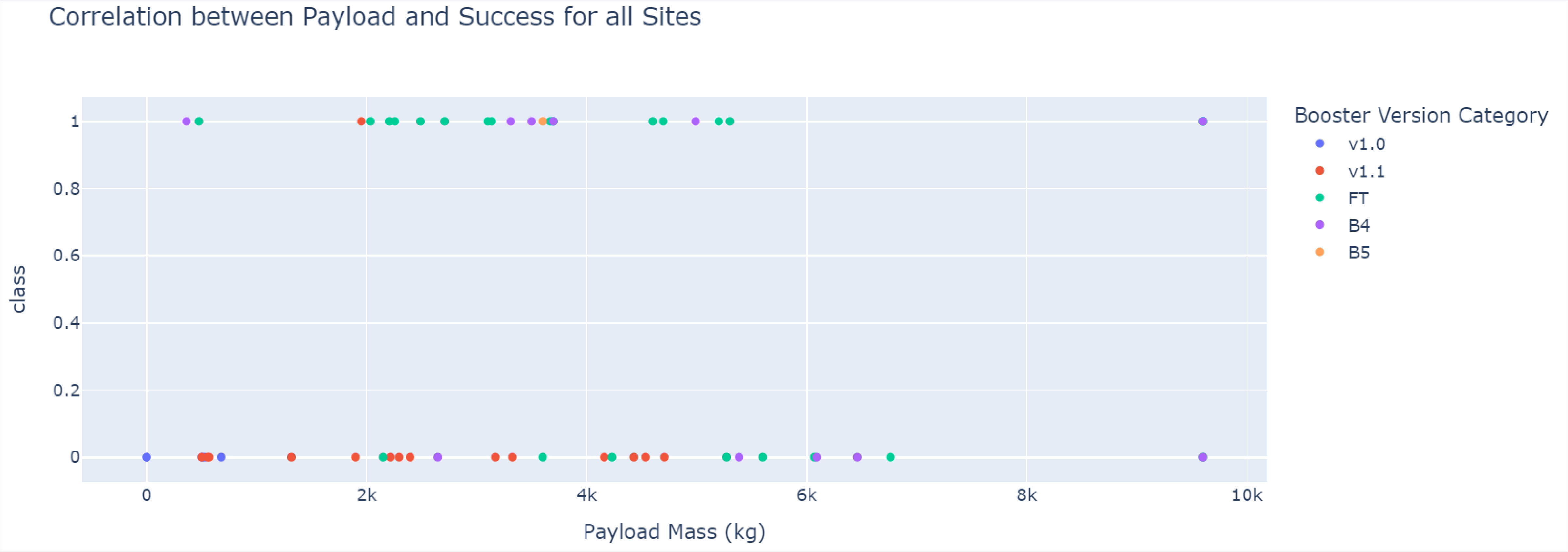
Relationship between Payload and Launch Outcome (all launch sites)

FT - most successful booster version; v1.0 (all launches failed) & v1.1 (all but one failed) - worst booster versions;

B4 - nearly equal number of failures and successes;

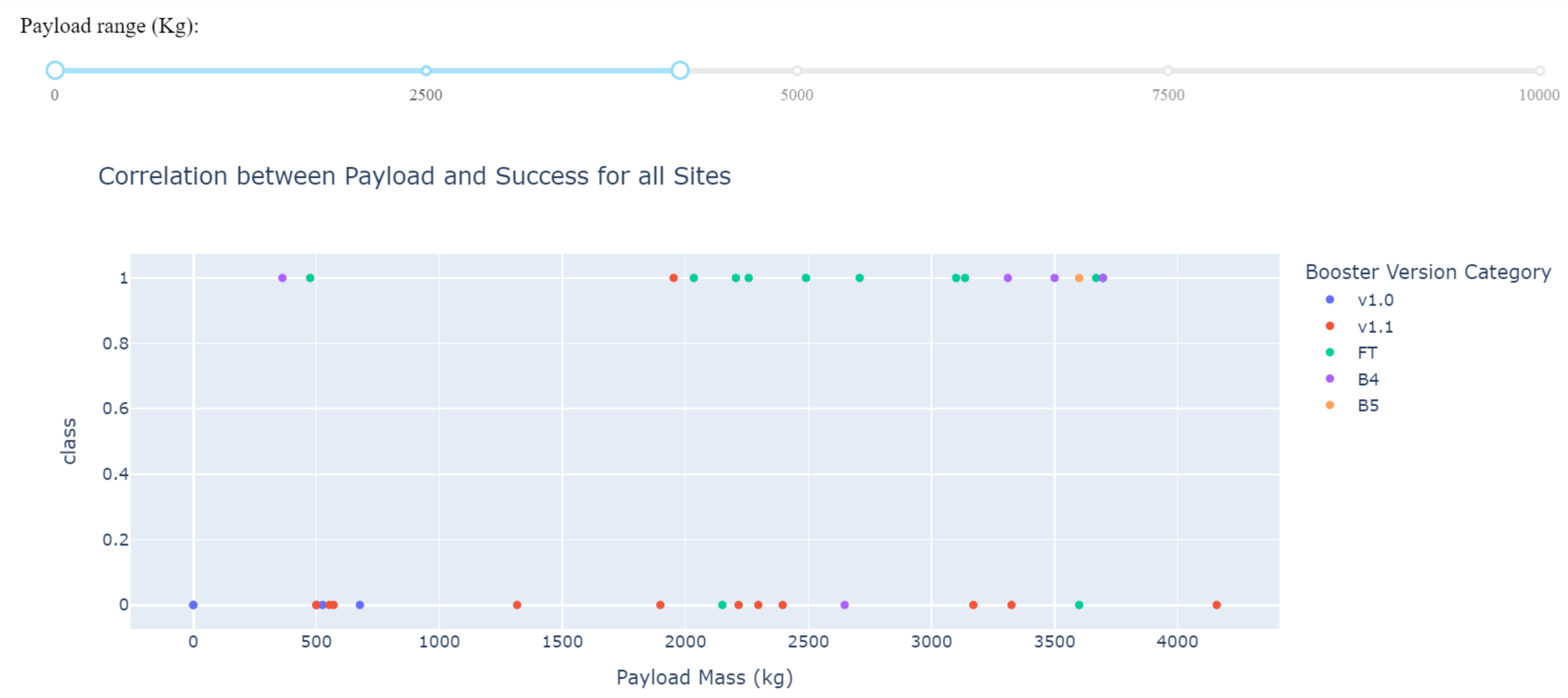
B5 – only one data point

66% of the launches are concentrated within the medium payload mass range (2000 – 6000 kg), with very few launches in the low and high end of the mass spectrum.



Analysis of Payload Mass Range of 0 – 4000 kg

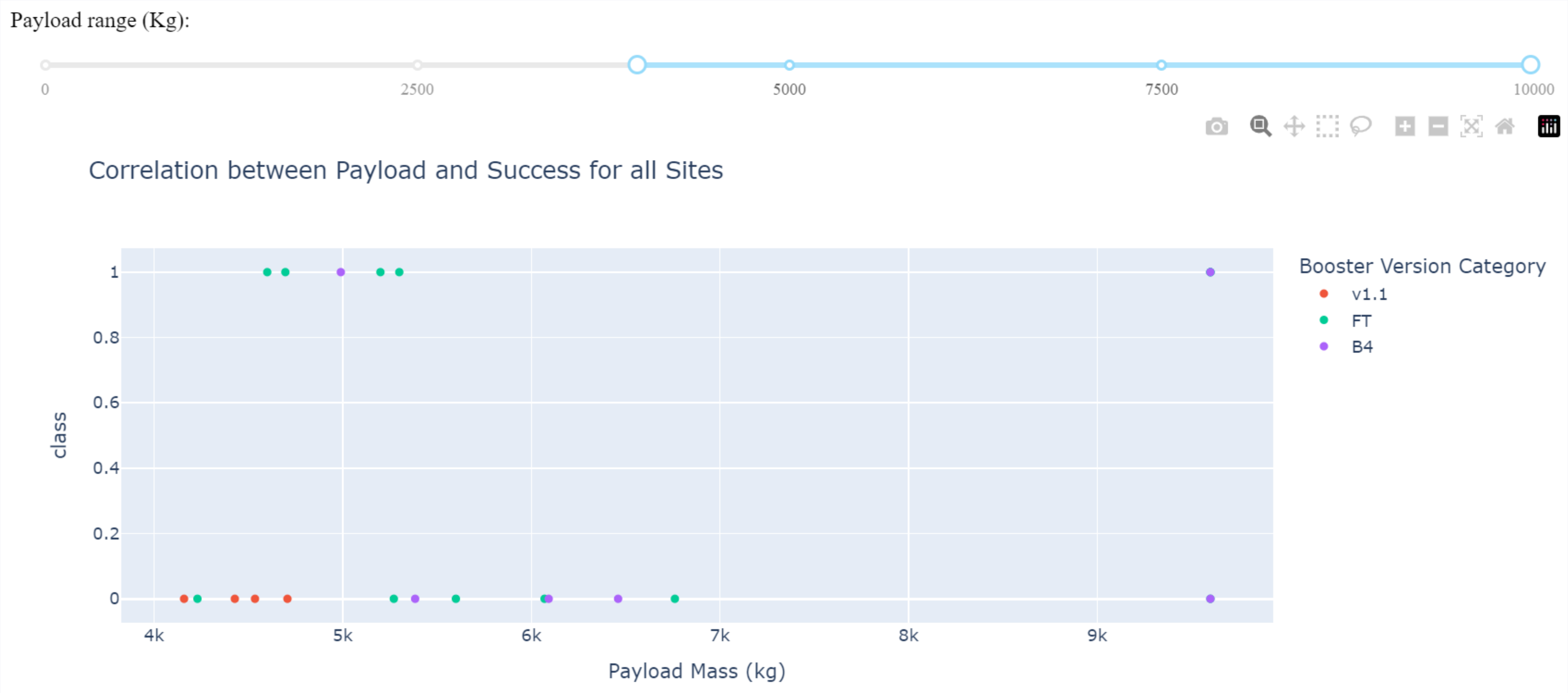
- Success rate in the entire range - FT: 82%, B4: 80%
- Below 2000 kg: 100% success rate for FT & B4. Also the only successful launch from v1.1 falls in this range. But the small sample size may not be reliable
- The 2000 – 4000 kg range is the most active range, with good success rates – FT: 80%; B4: 75%; B5: 100% (but only one data point)



Analysis of Payload Mass Range > 4000 kg

- Success rate: FT: 44%, B4: 33%

- From 5600 kg onwards: only failures from FT & all failures but one for B4. FT and B4 are the only booster versions used in this range.



Relationship between Payload Mass and Success

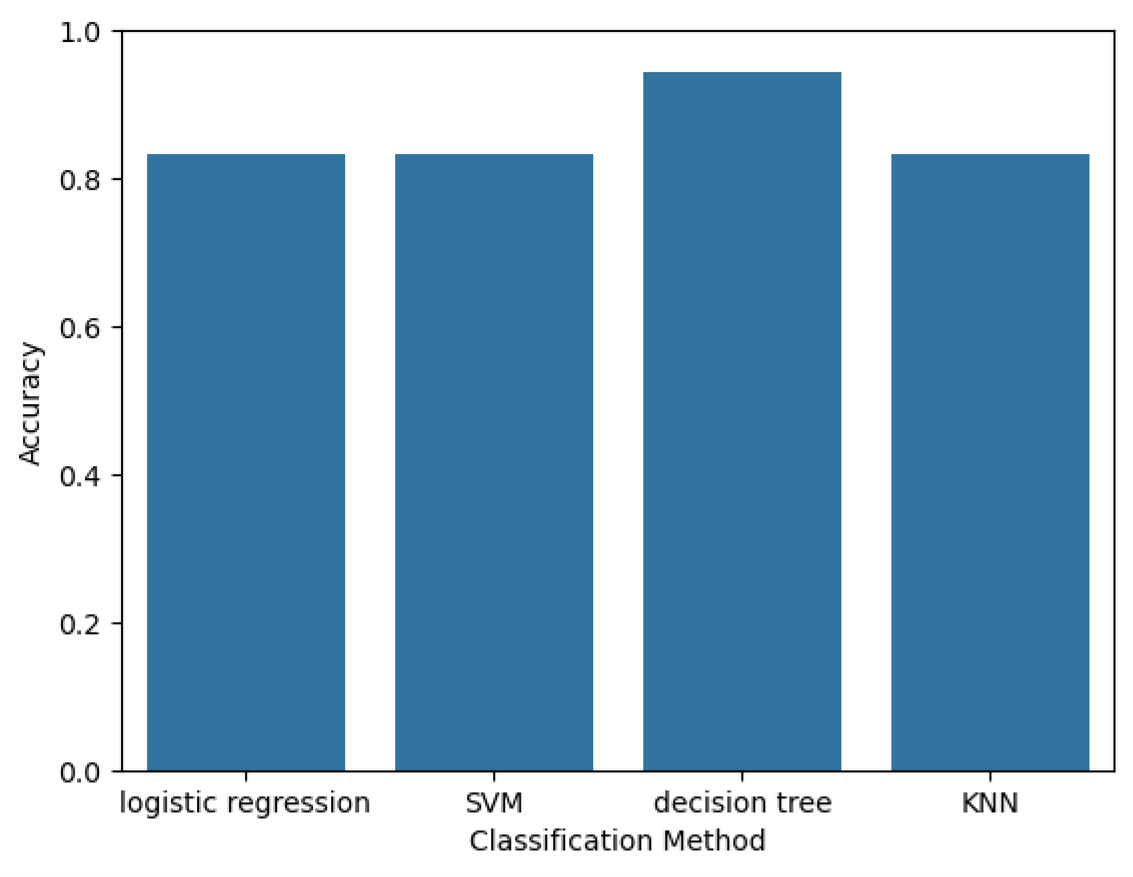
- The results show that lighter payloads tend to correlate with better success rates. This is unsurprising since lighter payloads require less fuel.
- The best success rates occur at payload masses below 4000 kg. Accounting for both the most frequent payload mass range used as well as success rate, the 2000 – 4000 kg payload mass range is the most successful.
- The worst success rates occur at the higher end of the payload mass scale. There are also very few launches in this range, and only involving FT and B4 boosters.
- Unlike in the previous section ("Insights drawn from EDA"), the dataset here spans a relatively short timeframe/flight number range. So the confounding effect of tech advancement over time is reduced – left with “true” relationship between payload mass and success.
- The earlier boosters, v1.0 (flight number range: 1 – 5) and V1.1 (flight number range: 6 – 21) performed poorly regardless of the payload mass

Section 5

Predictive Analysis (Classification)

Classification Accuracy

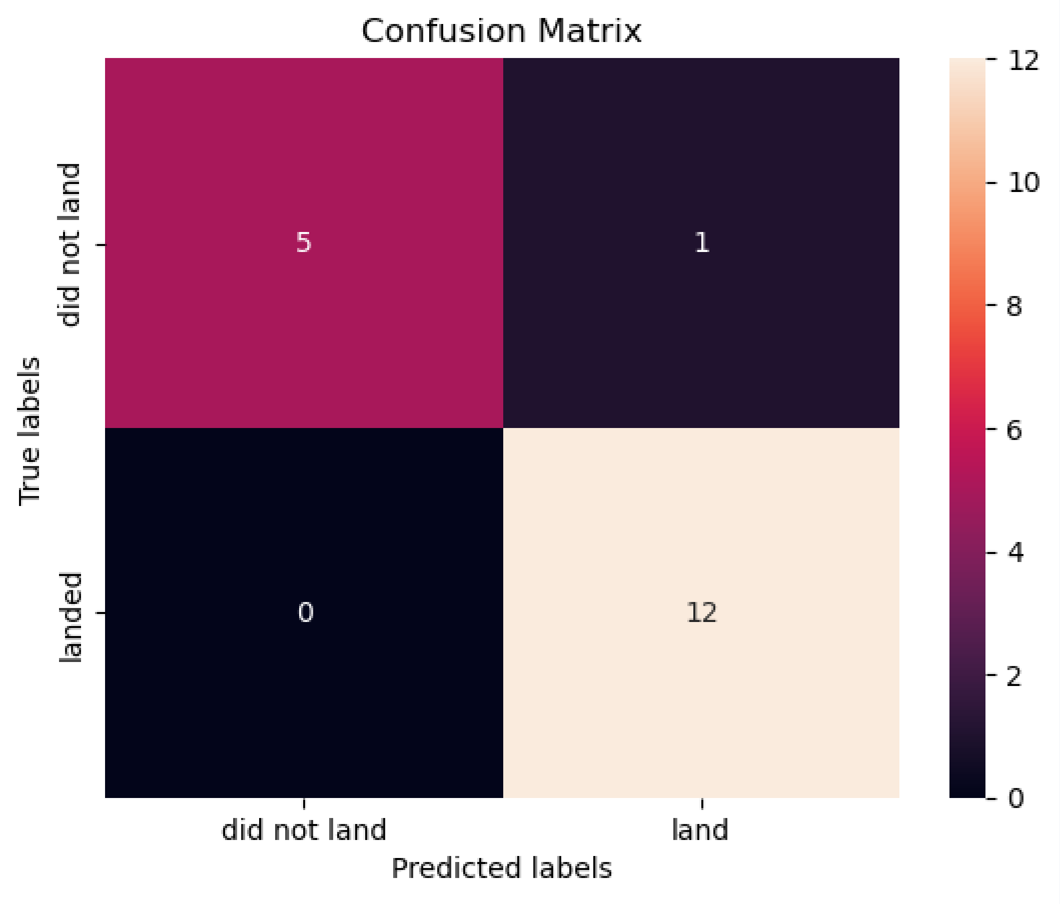
- The model with the highest classification accuracy is the decision tree classifier, with an accuracy score of 0.94



Confusion Matrix

The high prediction accuracy of the decision tree classifier model is also reflected in its confusion matrix:

- All predictions were correct except for only one false positive



Conclusions

- KSC LC-39A is the most successful launch site, with the highest success rate and the largest contribution of successful launches.
- SSO, followed by VLEO are the orbits which give the most (statistically meaningful) success rates.
- Launch sites are always located far from cities, but close to railways, highways and coastlines.
- Unsurprisingly, success rates improved with time/ flight number.
- Lower payload masses are associated with better success rates. Nevertheless, we may sometimes observe dataset that suggests the opposite, but this is spurious. Heavier payloads are more common in later flights, so what we are really seeing is the relationship between success rate and time/ flight number.
- The majority of the launches are in the medium payload mass range. When considering the combination of frequency of use and success rate, the 2000–4000 kg payload mass range is the most notable.
- The most successful booster version is FT.
- The model with the highest classification accuracy is the decision tree classifier.
- Some of the sample sizes were too small to obtain a statistically reliable analysis. For future analysis, a larger dataset could improve the accuracy.

Appendix: Data Collection

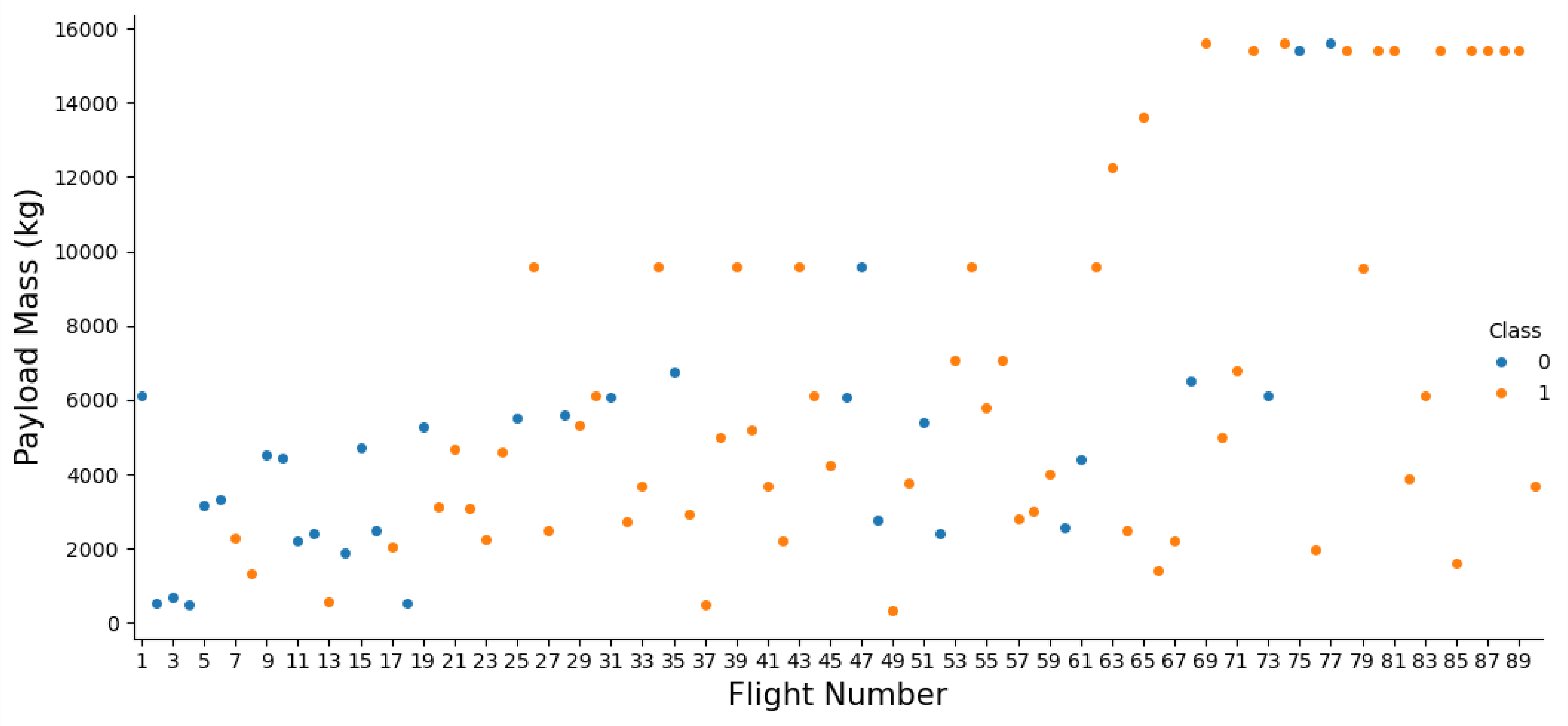
- API endpoints used for data collection:
- https://api.spacexdata.com/v4/launches/past
- https://api.spacexdata.com/v4/rockets/
- https://api.spacexdata.com/v4/launchpads/
- https://api.spacexdata.com/v4/payloads/
- https://api.spacexdata.com/v4/cores/

URL of wikipage for webscraping:

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

Appendix: Relationship between Flight no./Time with Payload Mass

Payload mass (and its corresponding success rate) is positively correlated with flight number: In earlier flights, only lighter payloads were used and there were more failures. As time progressed, the range of payload mass and success rates increased. Also, the more massive the payload, the less likely the first stage will return. Eventually however, after many flights, all launches were successful regardless of the payload mass.



Thank you!