Winning Space Race with Data Science

Marianne Campbell

4 March 2023

Outline

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

Executive Summary

- Successful commercial spaceflight requires accurate prediction of launch success
- Launch success can be predicted through
  - Using a REST API, webscraping relevant Wikipedia pages, data wrangling, data visualization, using SQL for querying, using Folium for mapping, using Plotly Dash for obtaining insights interactively, and using machine learning to build classification models
- The methodologies showed that
  - Successful launch probability increased over time
  - Orbit type, launch site, and payload mass were associated with different rates of success
  - Launch sites are far from cities and close to coastlines
  - All models used predicted success similarly

Introduction, Part I

- Government space agencies initiated space travel to explore worlds beyond Earth and identify alternative spaces for human habitation
- Launch costs were budgeted from taxes collected from citizens
- With the rise in aerospace technologies, private companies have entered the arena and contributed new perspectives to decades-old problems
- To commercialize air travel, it must be affordable and reliable
- Selling tickets for recreational air travel requires business-level accounting accuracy for all costs of a spaceflight for paying passengers
- If we can determine if the first stage of a rocket will launch, we can determine the cost of a launch

Introduction, Part II

- SpaceX's Falcon 9 can recover the first stage of the rocket, which significantly reduces costs.
- We are forming SpaceY to compete with SpaceX.
- Goal: determine price of each launch
- Method: gather information about SpaceX and make dashboards for the team
- We will figure out the conditions under which SpaceX reuses the first stage
- To achieve this, we will train a machine learning model and use public information to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

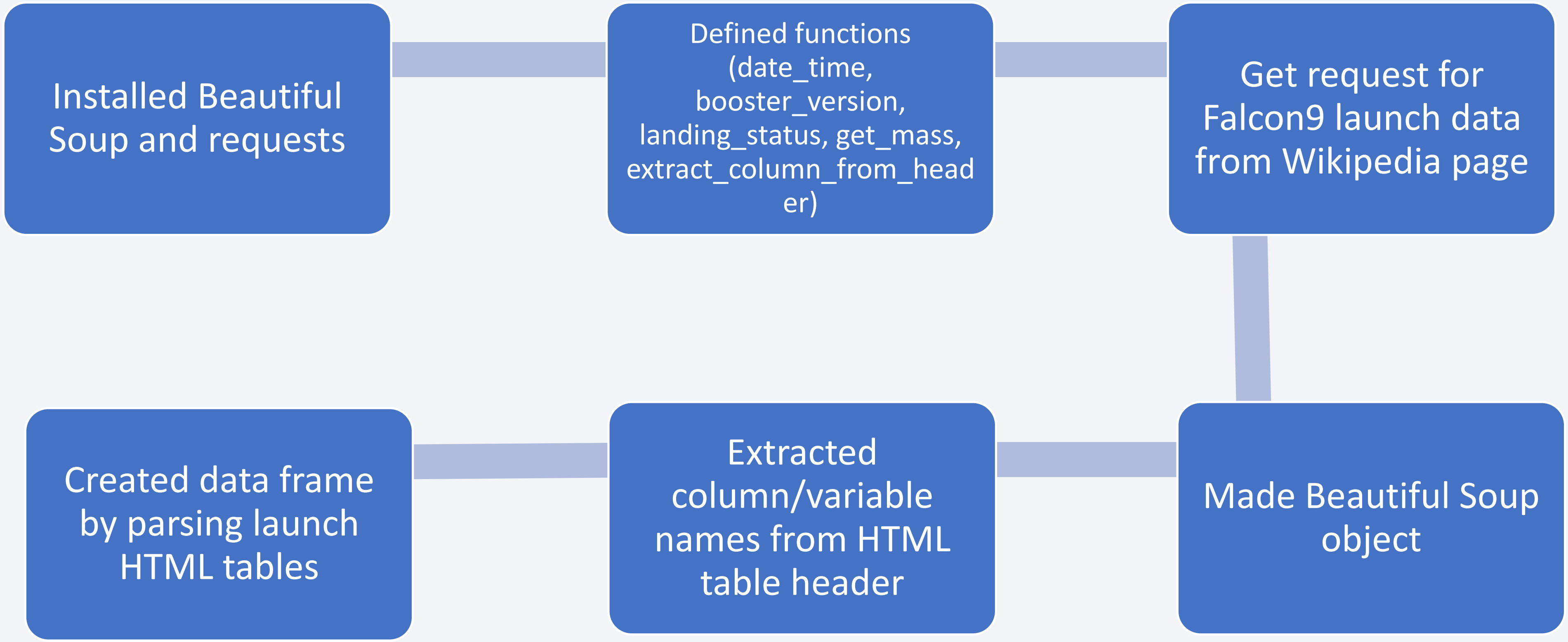
- Data collection methodology:
  - Data were gathered using a SpaceX REST API and webscraping related Wiki pages
- Perform data wrangling
  - Data was processed using an API, filtering for Falcon 9, and removing nulls (NaN)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A machine learning pipeline was built using train/test/split to perform and evaluate the results of logistic regression, support vector machines, decision tree classification, and K-nearest neighbors, outputting confusion matrices

Data Collection – SpaceX API

1. Imported requests library
2. Defined functions (get BoosterVersion, get LaunchSite, get PayloadData, get CoreData
3. Requests.get
4. Normalized
5. Restricted by data and number of cores
6. Called functions
7. Combined columns into dictionary
8. Filtered for Falcon 9 launches

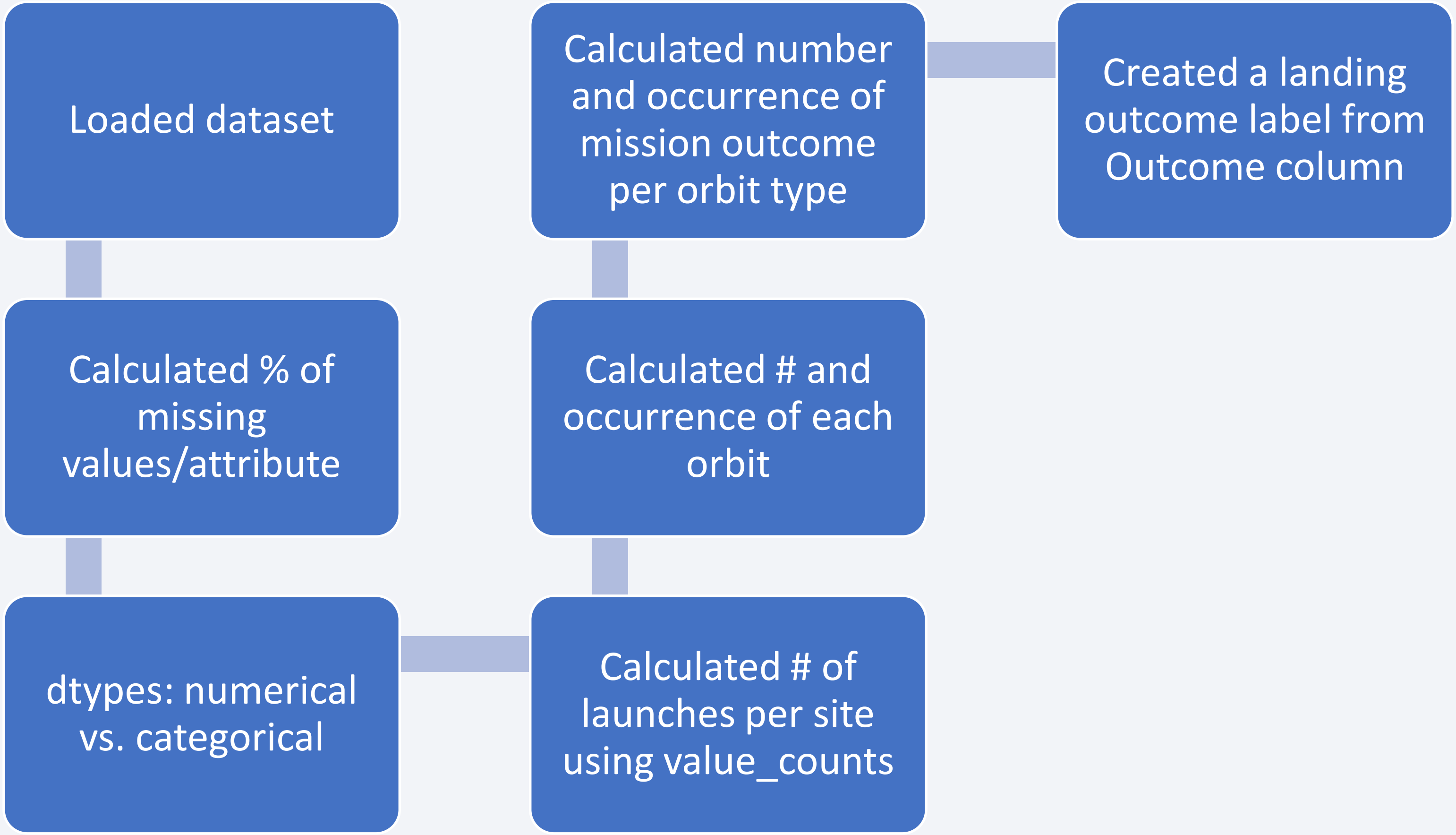
- GitHub link to URL

Data Collection - Scraping



[GitHub URL]

Data Wrangling



- GitHub URL

EDA with Data Visualization

GitHub URL

Question	Chart type	Reason for selecting chart type
How does FlightNumber (continuous launch attempts) and Payload variables affect launch outcome?	Scatter plot	X and Y variables are numerical, distinguishable dots using class variable show chronological trend
How does LaunchSite affect outcome?	Scatter plot	
Is there any relationship between launch site and payload mass that affects outcome?	Scatter plot	
Is there any relationship between success rate and orbit type?	Bar graph	Orbit type is categorical while success rate is numerical
Does orbit type affect the success likelihood over time?	Scatter plot	One variable is numerical and other is categorical, dots show chronological trend
Are payload and orbit type related?	Scatter plot	
What is the trend for successful launches over time?	Line graph	Time is a continuous variable

EDA with SQL: used SQL commands to accomplish the following

[GitHub URL]

- Displayed names of unique launch sites
- Displayed 5 records where launch sites began with the string 'CCA’
- Displayed total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Found date when first successful landing outcome on ground pad was achieved
- Found names of the boosters which had success landing on drone ship and had payload mass of 4000-6000 kg
- Identified total number of successful and failed mission outcomes
- Identified booster versions which carried the maximum payload mass
- Listed failed landing outcomes on drone ship, their booster versions, and launch site names for 2015
- Ranked 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]

Launch success rates might depend on proximity to specific land features

Object added (type)	Reason added
Launch sites (circles)	Show proximity to natural (coast) and man-made land features
Successful and failed launches (marker cluster)	Identify patterns for different launch sites
Latitude and longitude identifier (mouse position)	Identify coordinates of any location
Line (polyLine)	Show distance between markers (ex. launch site to coast, railroad tracks, highway

Build a Dashboard with Plotly Dash

GitHub URL

Feature	Purpose
Dropdown list	Enable launch site selection
Pie chart	Show the total successful launches count for all sites/each site
Slider bar	Select payload range
Scatter plot	Show the correlation between payload and launch success
Callback function #1	\`Site-dropdown\` as input, \`success-pie-chart\` as output
Callback function #2	\`Site-dropdown\` and \`payload-slider\` as inputs, \`success-payload-scatter-chart\` as output

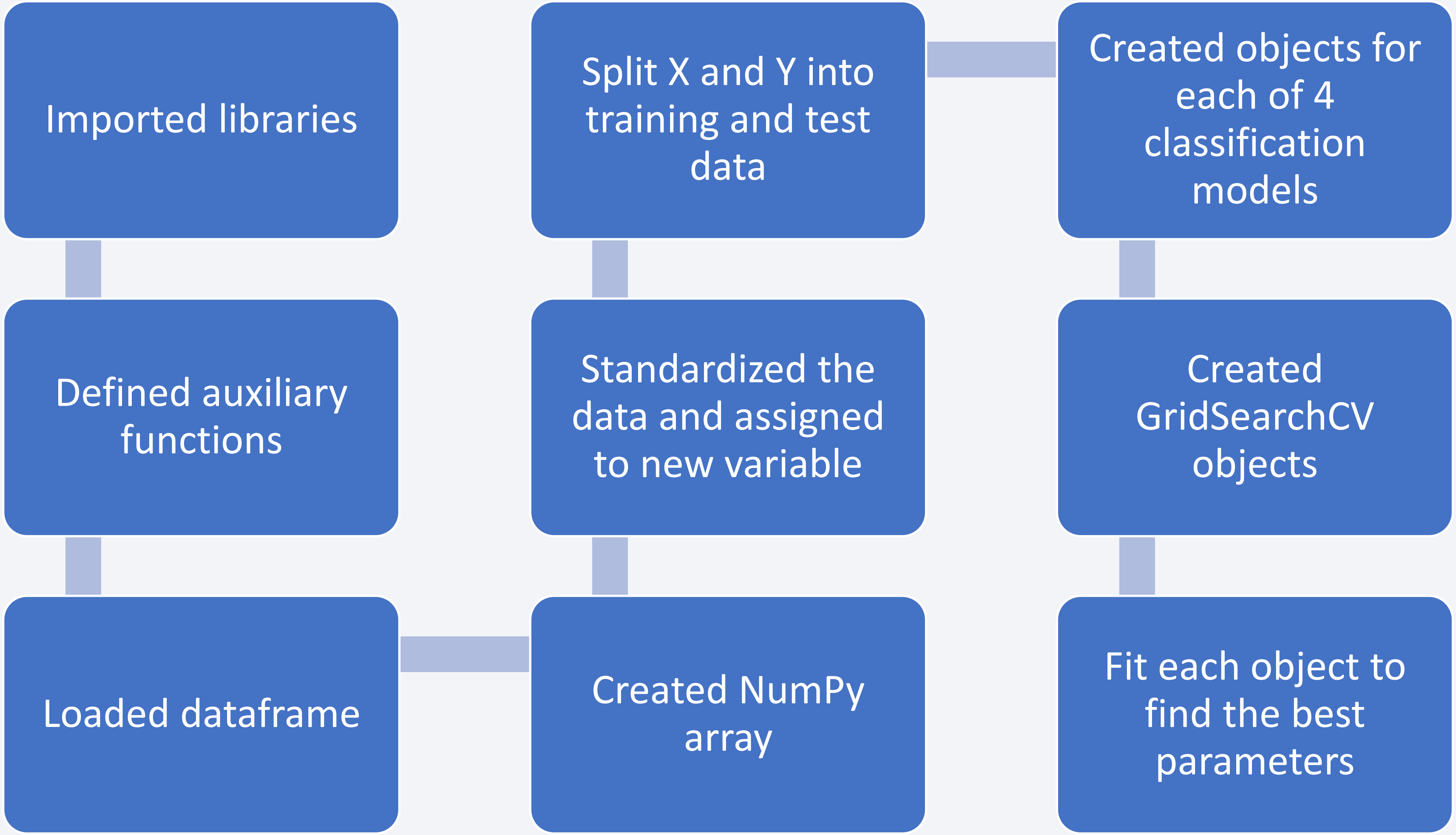
Predictive Analysis (Classification, Part I)

GitHub URL

- Building models: imported libraries, imported dataframe, created NumPy array, standardized the data, deployed train_test_split, created 4 types of objects
  - Logistic regression, support vector machine, decision tree classifier, and K nearest neighbors
- Evaluation: For each model, calculated accuracy using score, assessed accuracy using confusion matrix
- Improvement: Used GridSearchCV to find the best parameter values to achieve the greatest accuracy
- Compared accuracy values to determine the best-performing model

Predictive Analysis (Classification, Part II)

GitHub URL

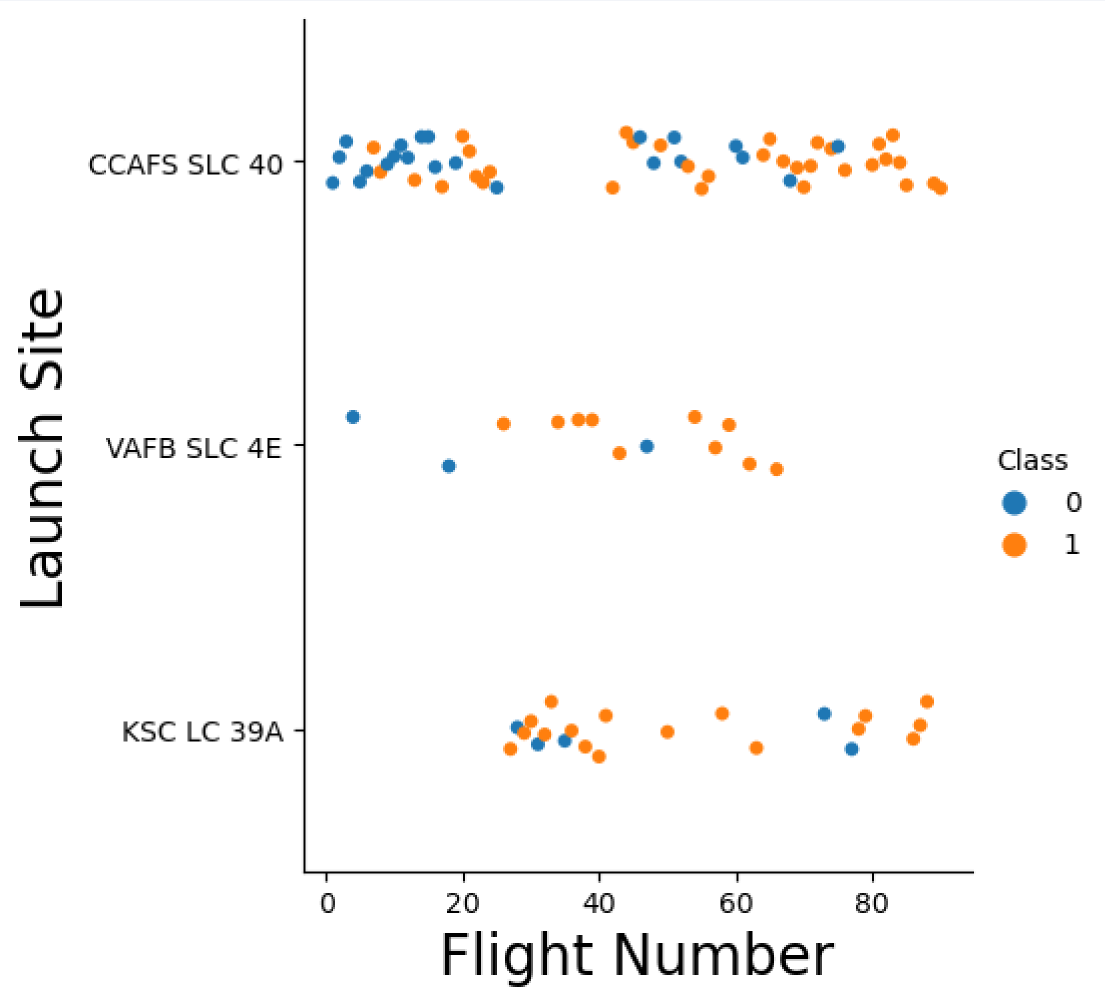


Section 2

Insights drawn from EDA

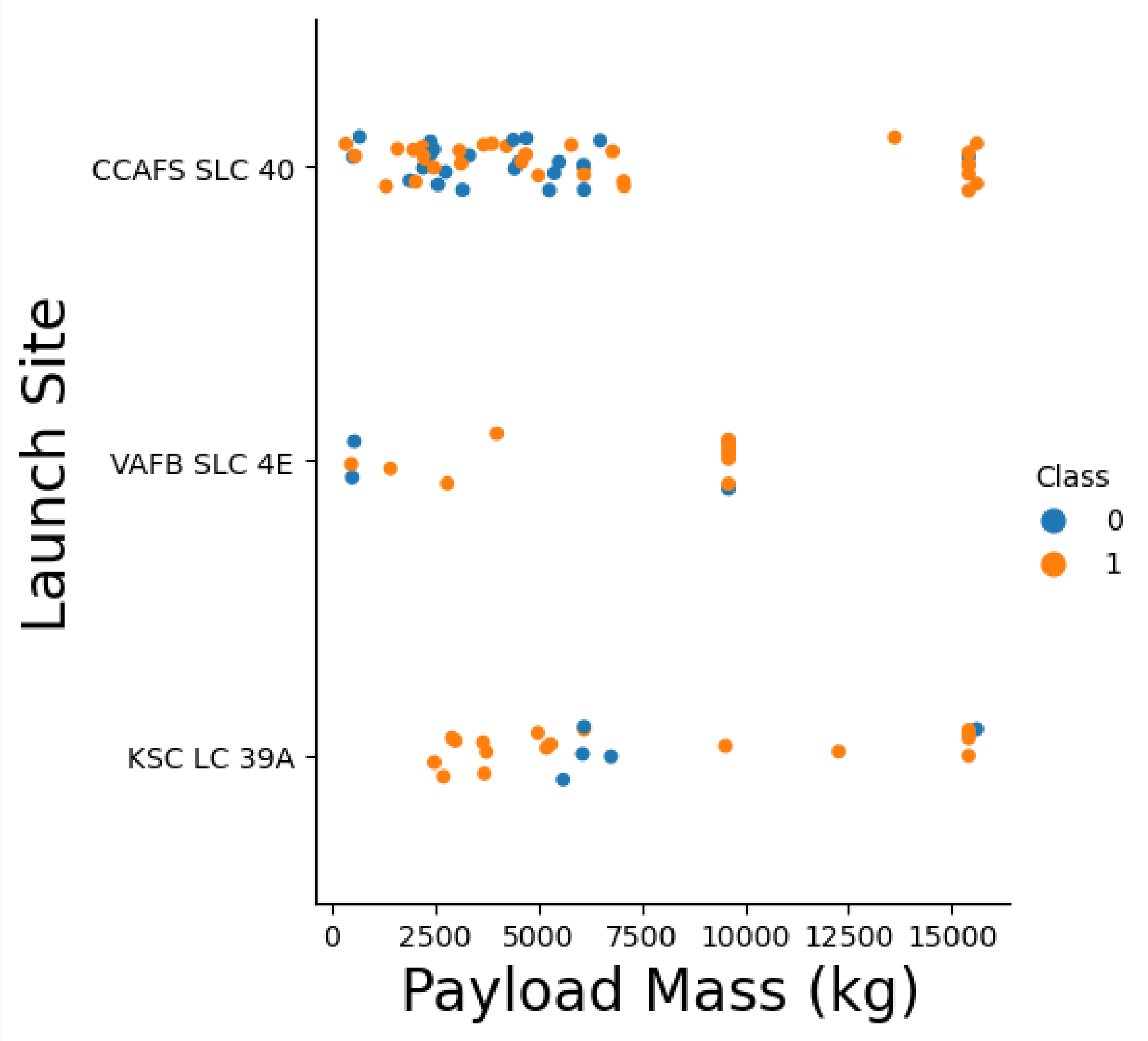
Flight Number vs. Launch Site

- For all launch sites, as flight number increases, so does the rate of success
- Cape Canaveral launched the greatest number of flights
- Kennedy Space Center and Vandenberg Air Force Base had higher success rates than Cape Canaveral



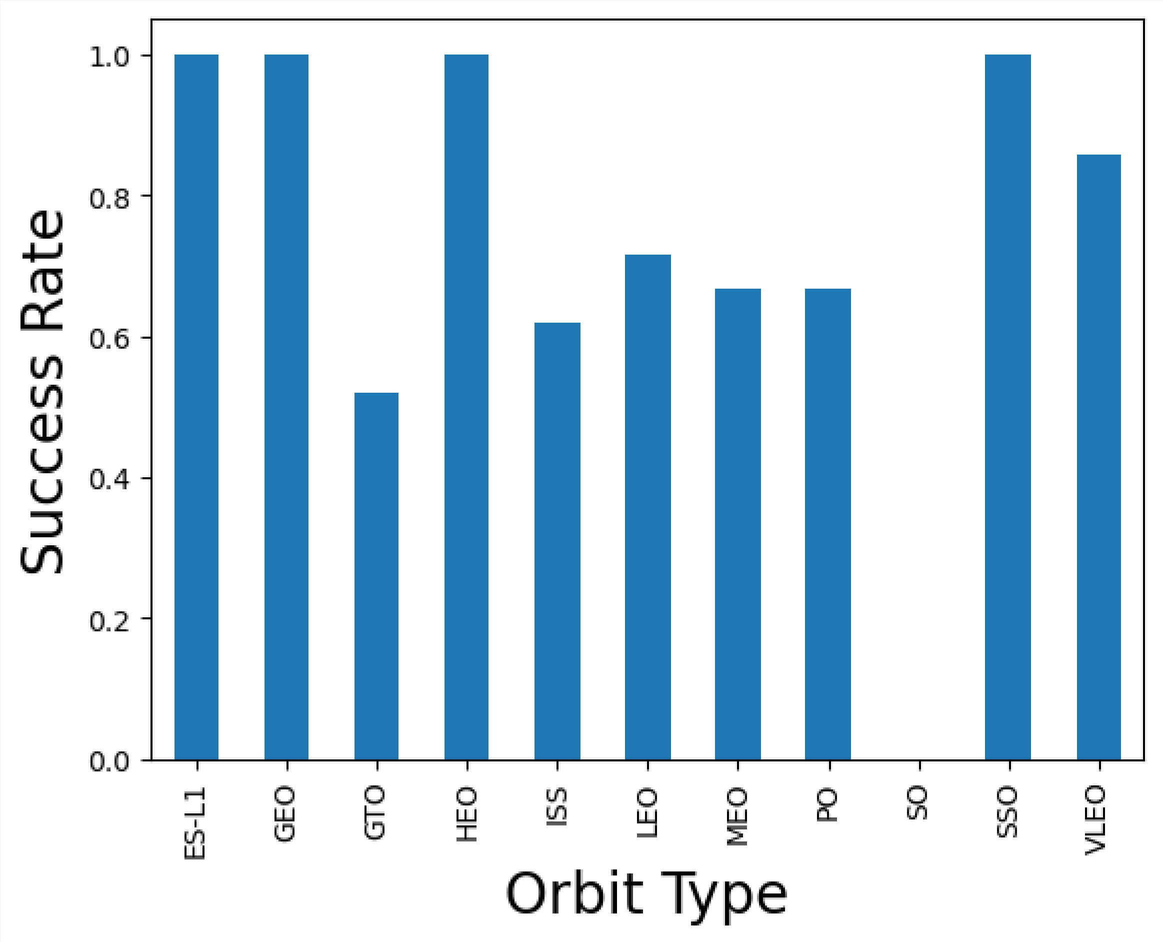
Payload vs. Launch Site

- Cape Canaveral mainly launched low payload craft, but all its high payload launches were successful
- For Vandenberg, there are no rockets launched for heavy payload mass (greater than 10000)
- Payloads of 5000-7500 kg were least successful at Kennedy Space Center



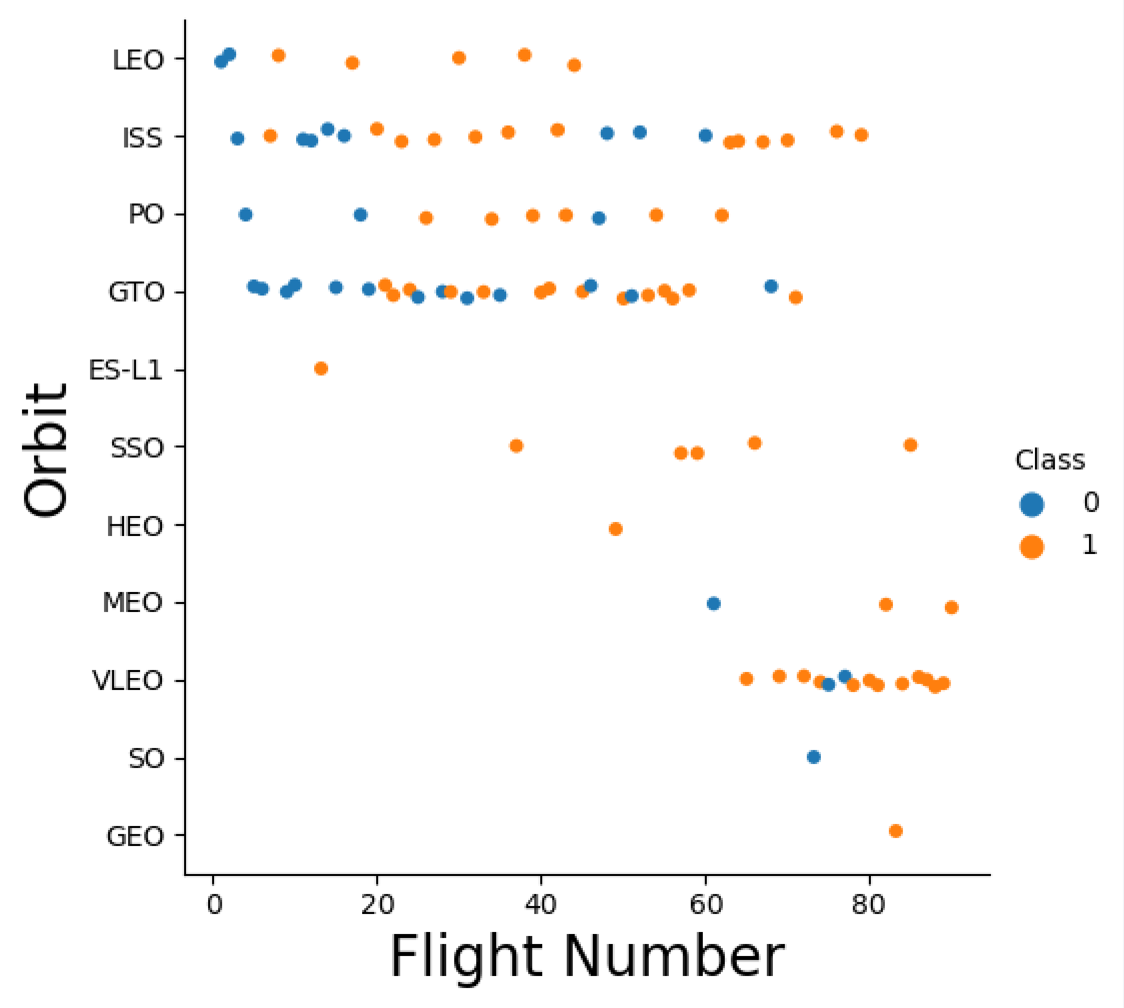
Success Rate vs. Orbit Type

- Orbit types with the greatest success rates include ESL-1, GEO, HEO, and SSO
- Orbit types GTO and SO should be avoided due to low success rates
- A confounding variable is the dates of each of the orbit types: if some orbit types were favored earlier or later in the process, launch success may not be related to orbit type



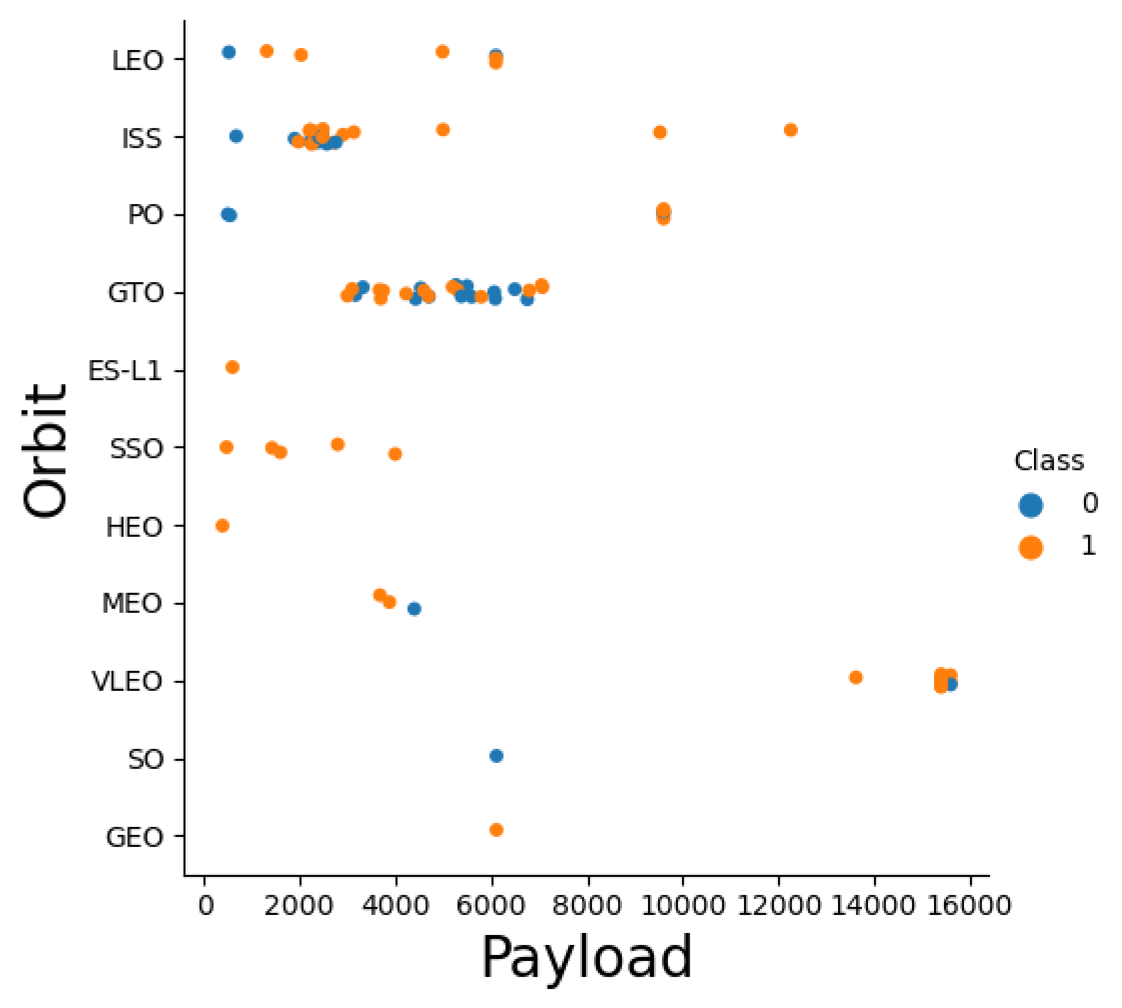
Flight Number vs. Orbit Type

- In the LEO orbit the success appears to be related to the number of flights
- There seems to be no relationship between flight number and success when in GTO orbit
- Do GTO orbits present technical challenges unresolved by repeated flights?
- SSO orbits have been uniformly successful despite few attempts
- SSO and VLEO orbits have the highest overall success rates



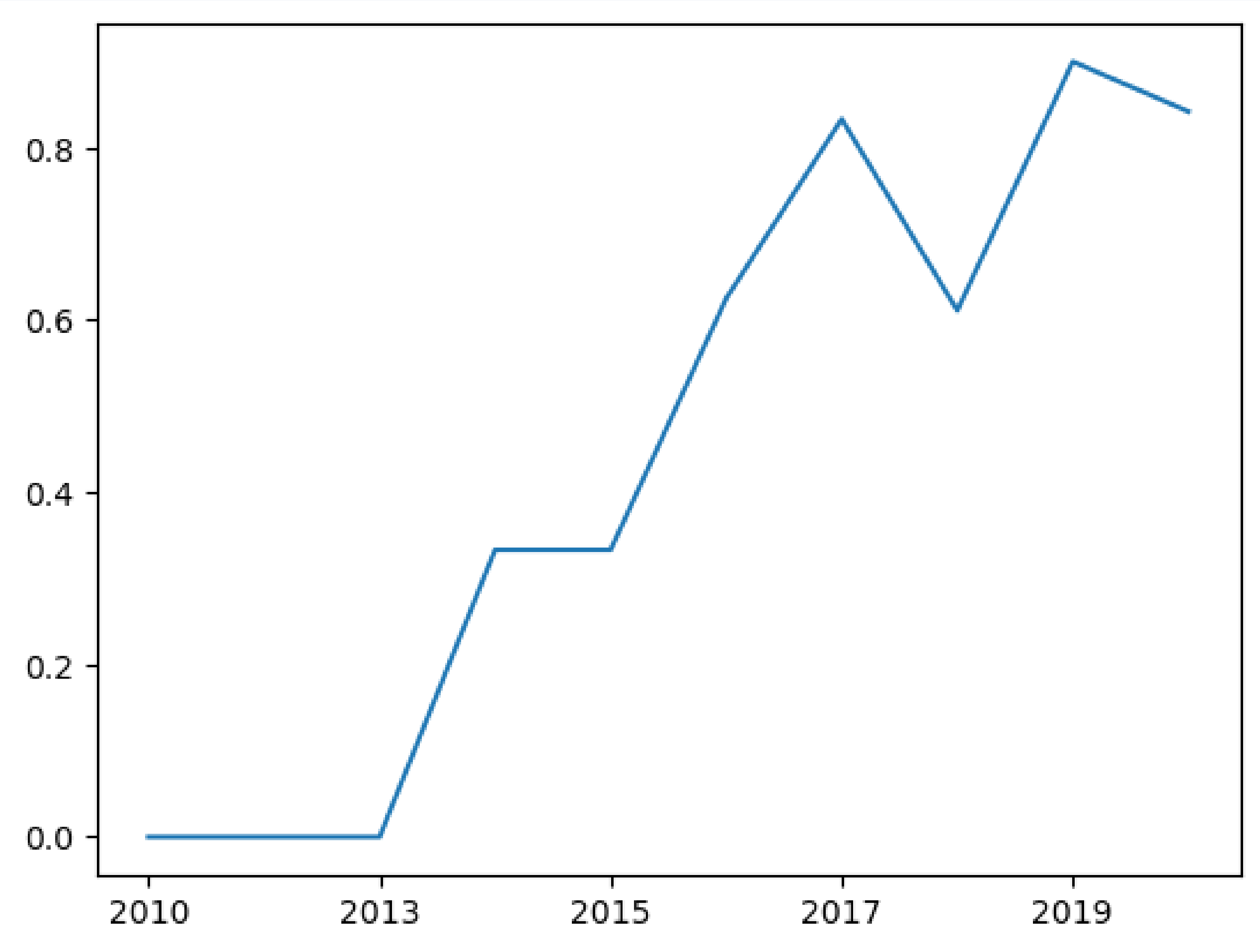
Payload vs. Orbit Type

- GTO orbits have the smallest range of payloads and lowest rate of success
- SSO orbits are suitable for light payloads only
- Polar, LEO and ISS orbits result in higher success rates with heavy payloads
- ES-L1, HEO, VLEO, SO, and GEO orbits have not been attempted enough to analyze trends



Launch Success Yearly Trend

- Launch success increases between 2013 and 2020
- The dip in 2018 may be attributed to a greater number of launches with less preparation, small sample size, random error, or other factors



All Launch Site Names

Find the names of the unique launch sites- used SELECT DISTINCT

Display the names of the unique launch sites in the space mission

```
In [7]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX

 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[7]:

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

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`: used SELECT with a WHERE clause

Display 5 records where launch sites begin with the string 'CCA'

In [8]:
```
%sql SELECT launch_site FROM spacex WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[8]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Calculate the total payload carried by boosters from NASA: used SELECT and sum with like statement

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [9]: %sql SELECT sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS__KG_ from spacex where customer like 'NASA (CRS)'

 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[9]:

total_payload_mass__kg_
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1: used SELECT and avg with like statement and where clause

Task 4

Display average payload mass carried by booster version F9 v1.1

In [10]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) from spacex WHERE Booster_Version LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[10]:

1
2928

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad: Used SELECT with where clause containing an equality

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [11]:
```
%sql SELECT min(Date) from SPACEX WHERE "Landing__Outcome" = 'Success (ground pad)';
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[11]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000: Used SELECT with equality statement and multiple conditions

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [14]:
```
%sql SELECT Booster_Version from spacex WHERE "Landing__Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4001 AND
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes: Used SELECT with where clauses, group by clause and % wildcard

List the total number of successful and failure mission outcomes

In [35]:
```
_Outcome, COUNT(*) FROM spacex WHERE (Landing__Outcome like 'Succ%' OR Landing__Outcome like 'Fail%') GROUP BY Landing__Outcome;
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[35]:

landing__outcome	2
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass: Used SELECT and subquery with equality and max function

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [40]:
```
#%sql SELECT Booster_Version FROM spacex WHERE (SELECT max(PAYLOAD_MASS__KG_))

%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM spacex WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacex);
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[40]:

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

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015: Used SELECT, where clause, equality, and multiple conditions with AND

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [42]:
```
%sql SELECT Booster_Version, Landing__Outcome, launch_site FROM spacex WHERE (landing__outcome ='Failure (drone ship)' AND year(d
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[42]:

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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order: Used SELECT, where clause, BETWEEN, AND, LIKE, and group by

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

In [43]:
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__C
```

```
 * ibm_db_sa://lwz86339:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[43]:

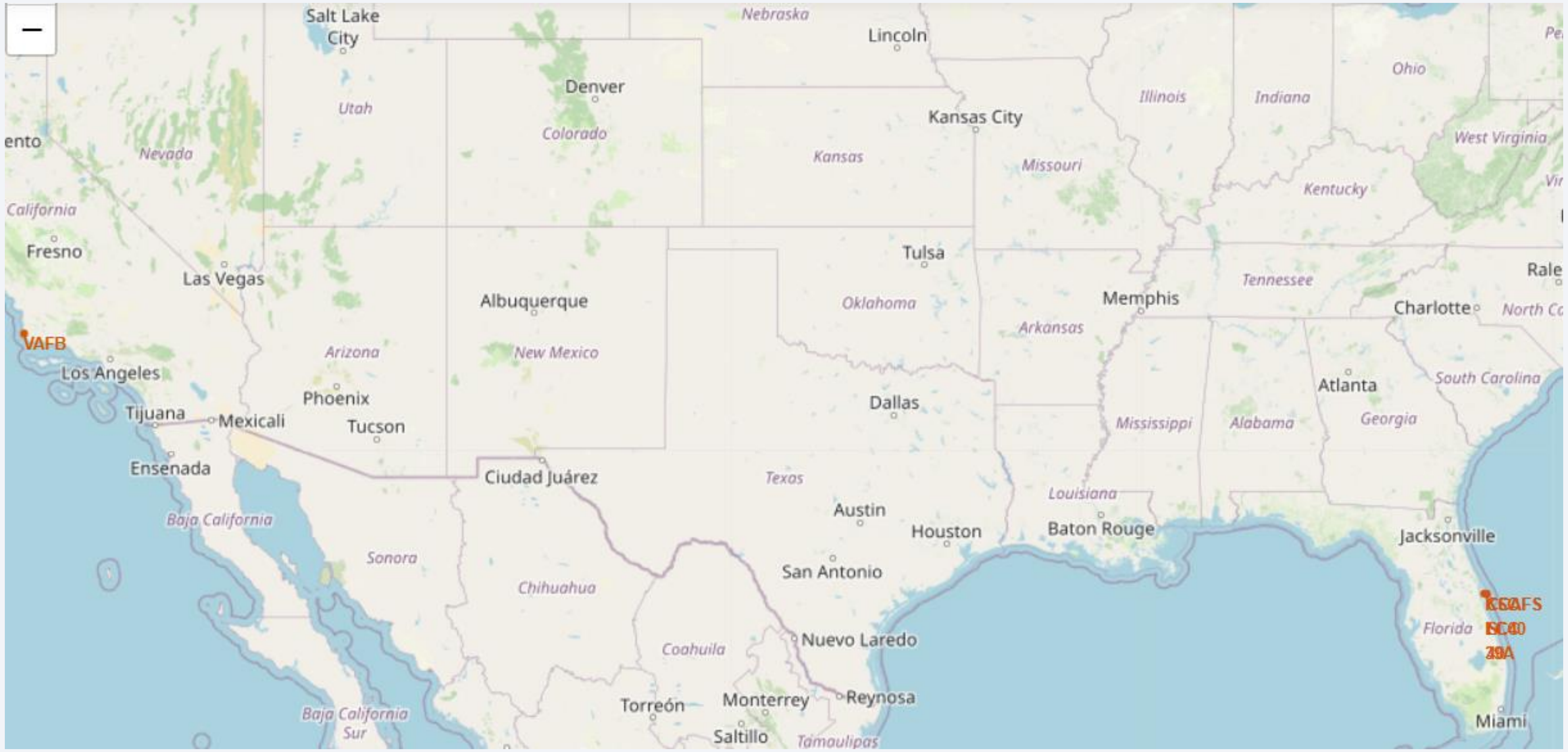
landing__outcome	2
Success (drone ship)	5
Success (ground pad)	3

Section 3

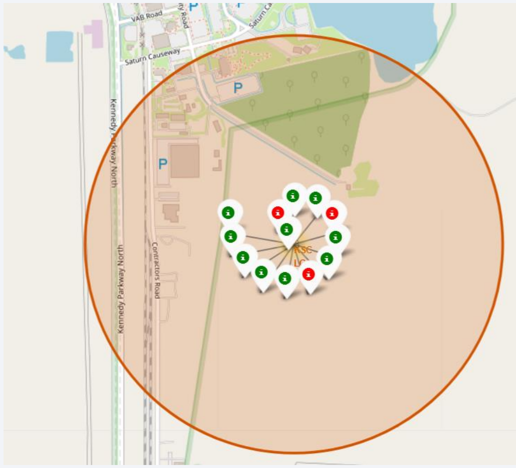
Launch Sites Proximities Analysis

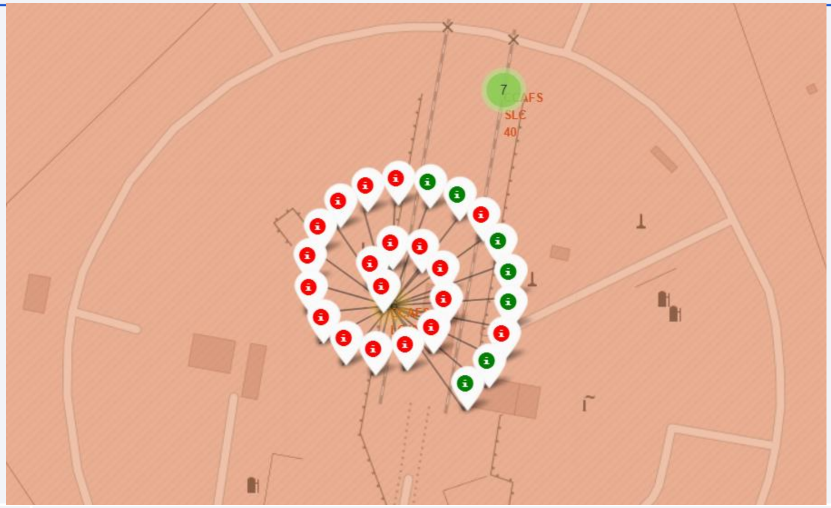
All Launch Sites Mapped

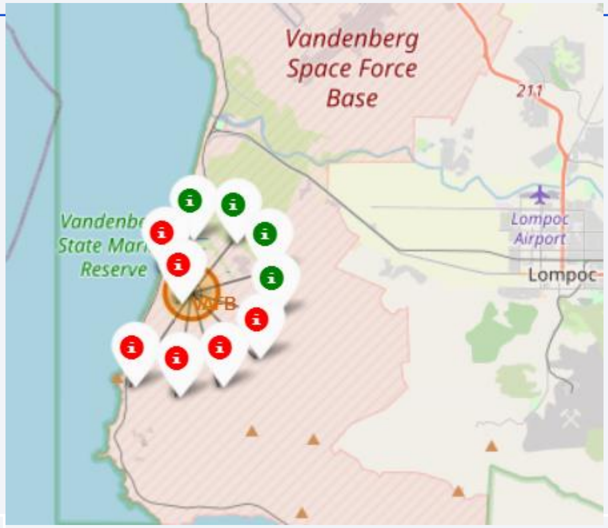
- All launch sites are located near coastlines, presumably to mitigate casualties in the event of malfunctions
- All launch sites were previously used for space or air flight, presumably to maximize existing infrastructure
- All launch sites are situated in areas that were sparsely populated at inception



Success and Failure of Launch Sites can be Visualized







Kennedy SC	Cape Canaveral AFS – both sites	Vandenberg AFB

- Kennedy Space Center showed the highest rate of launch successes
- Cape Canaveral showed the lowest rate of launch successes
- Land features (natural and man-made) may not be related to launch success
- Causation and correlation must not be confused

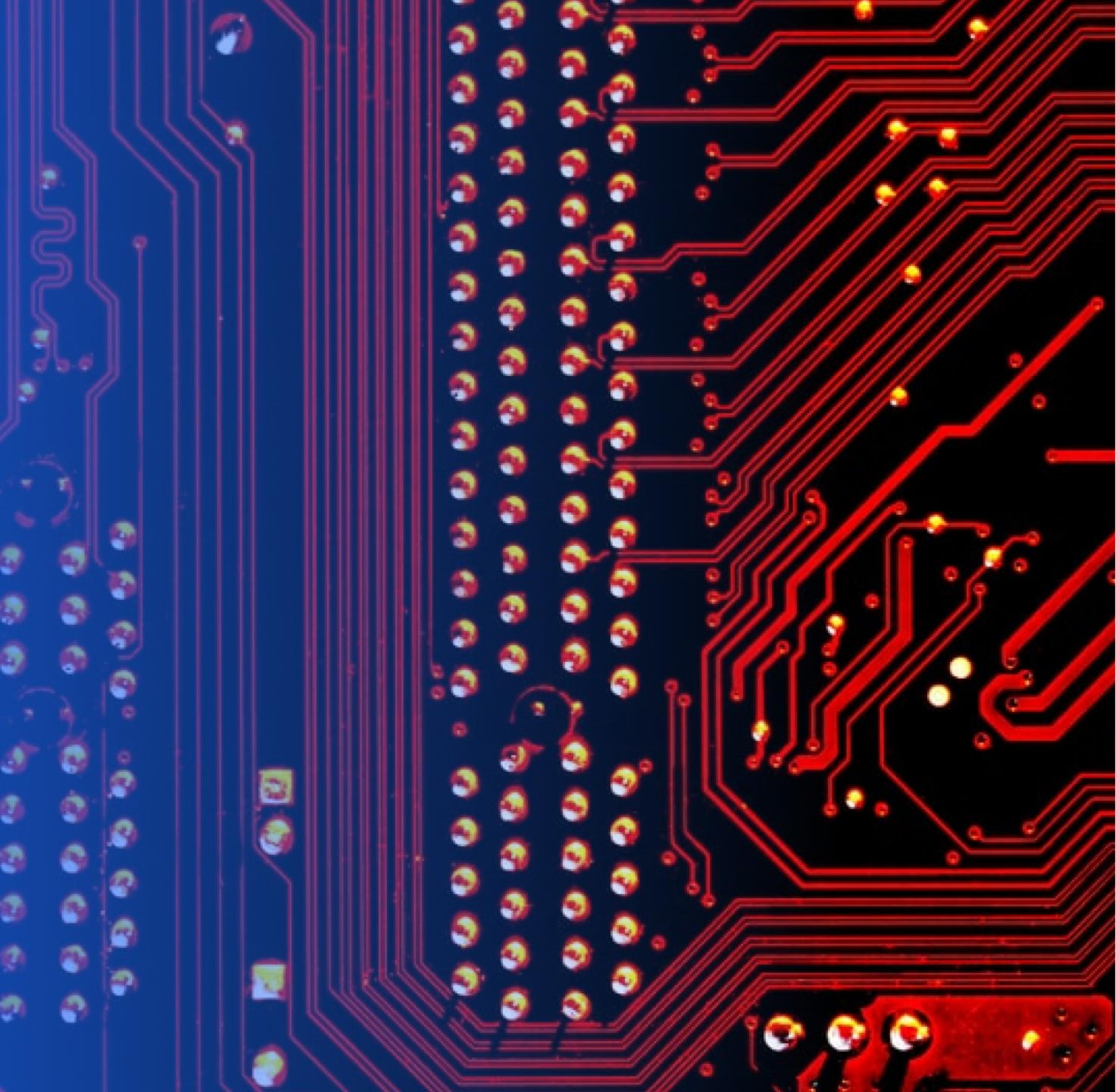
Distance to Important Features can be Calculated

- CCAFS SLC 40 is 0.86 kilometers from the coastline

- In general, launch sites are far from cities, close to coastlines, and near purpose-built airstrips

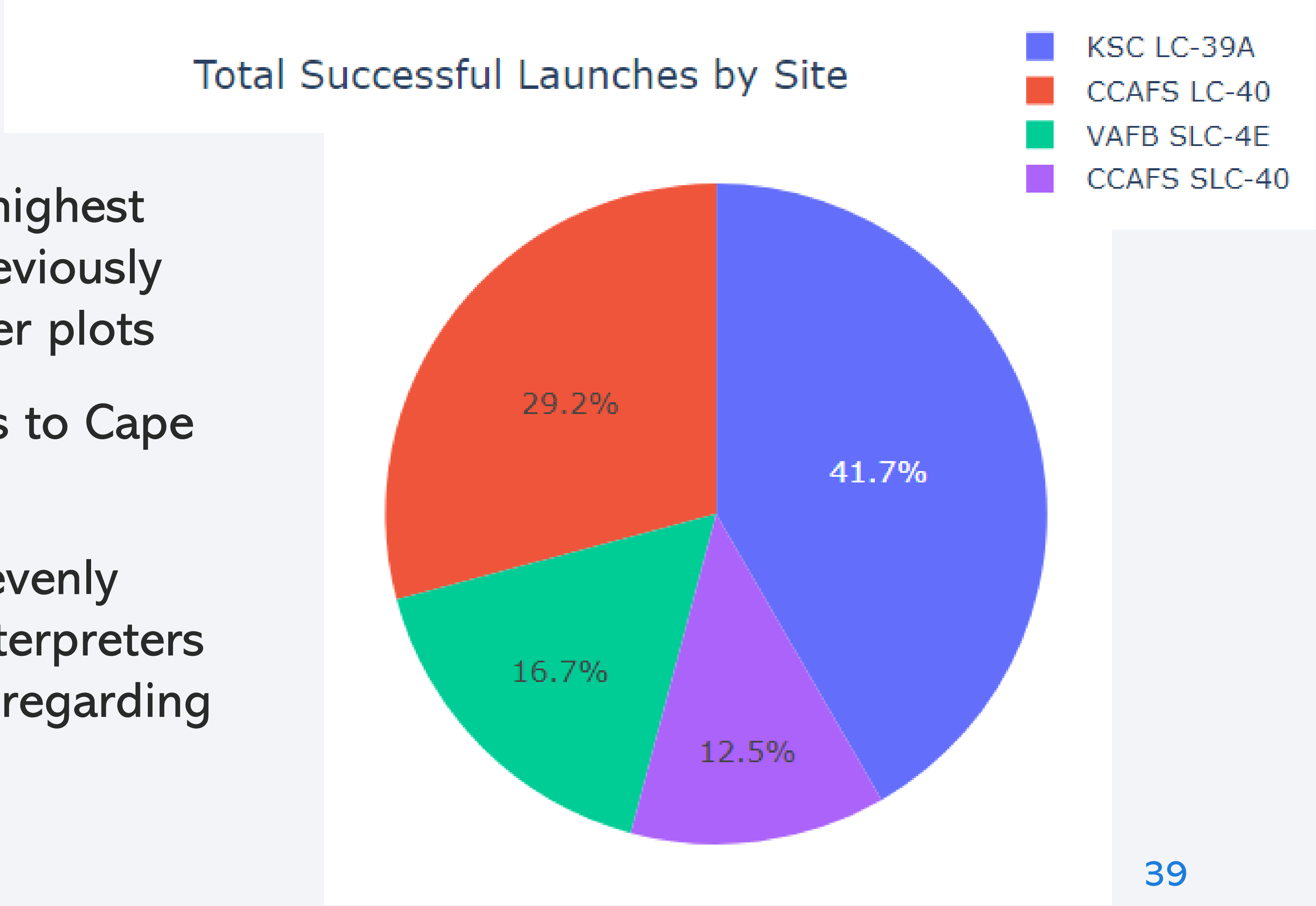
Section 4

Build a Dashboard with Plotly Dash



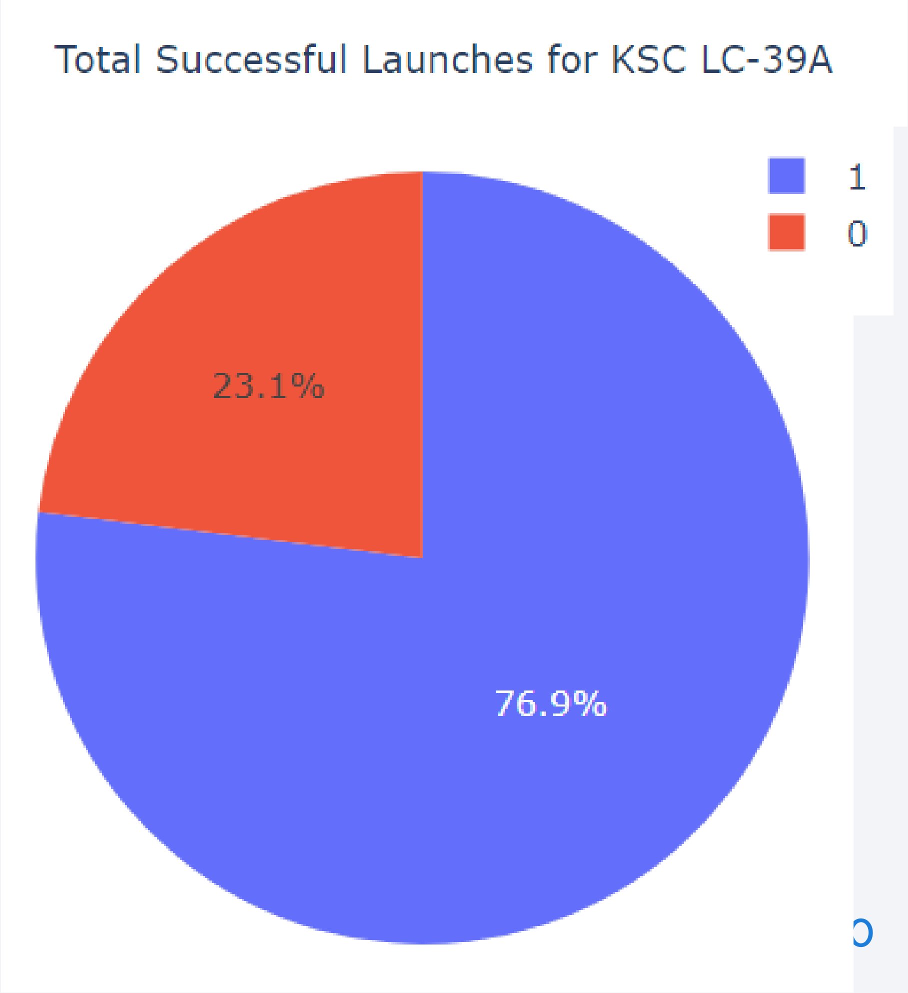
Launch success

- Kennedy Space Center has the highest success rate for launches, as previously shown in Folium data and scatter plots
- The lowest success rate belongs to Cape Canaveral's SLC-40 site
- Launch attempts are likely not evenly divided by date among sites; interpreters should avoid inferring causality regarding site and success

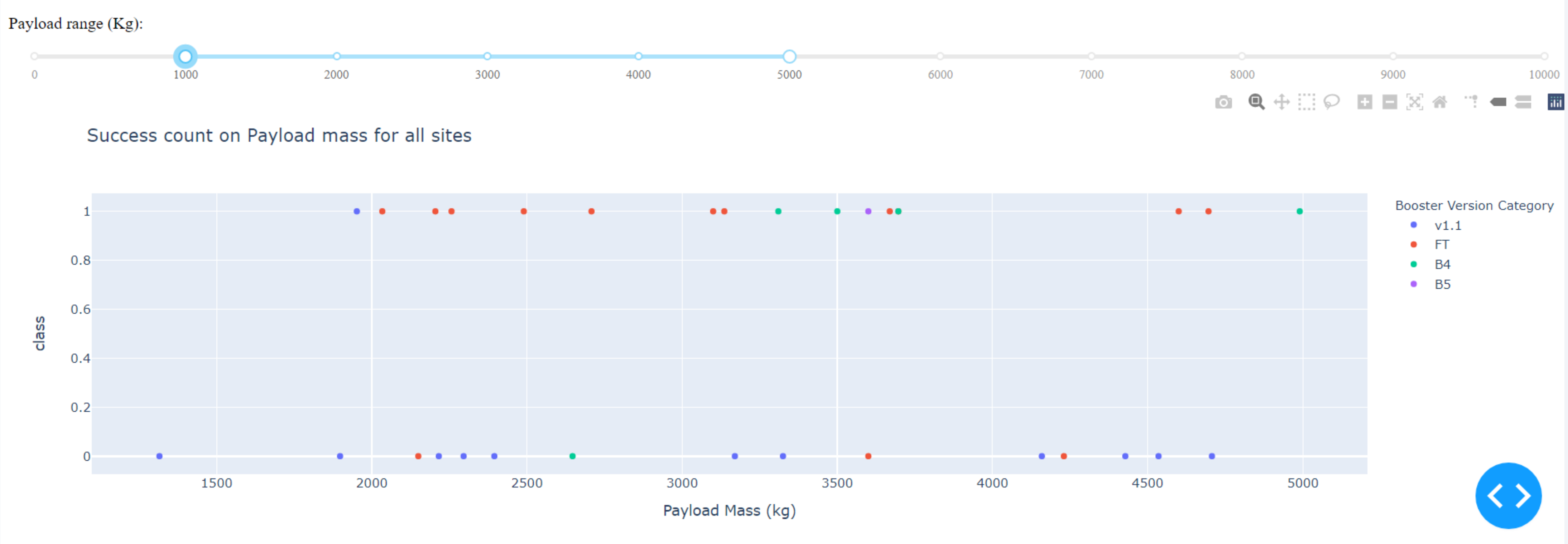


Kennedy Space Center has the Highest Launch Success Rate

- Over 75% of launches attempted at the Kennedy Space Center were successful
- Further study: which of the following factors contributed to the success rate compared to other sites?
  - Personnel
  - Weather
  - Natural features of the site
  - Launch dates (season and stage of technology development)
  - Rocket components
  - Regulatory procedures
  - Other variables

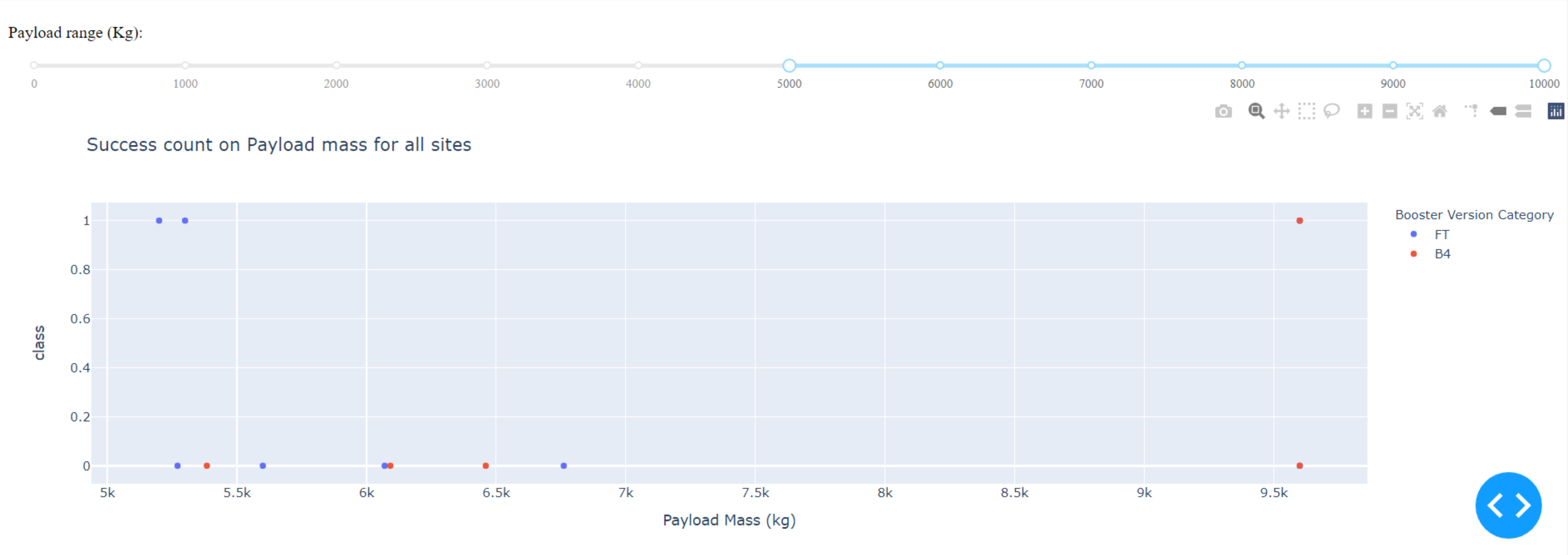


Light Payload vs. Outcome for All Sites



- For light payloads (1000-5000 kg),
  - v1.1 boosters have a low success rate
  - B4 and FT have high success rates

Heavy Payload vs. Outcome for All Sites



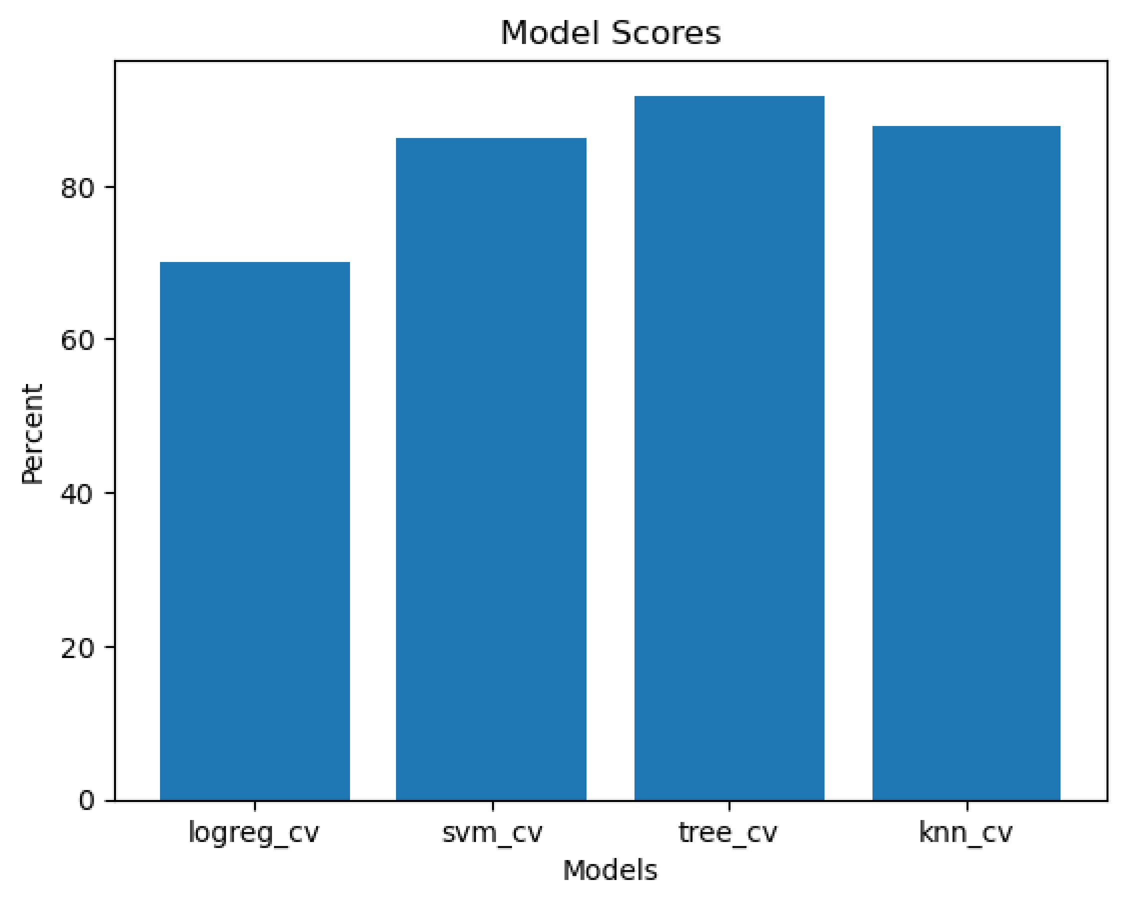
- For heavy payloads (5000-10000 kg),
  - Only FT and B4 boosters were used
  - FT boosters are successful at lower masses and B4 boosters are successful at higher masses

Section 5

Predictive Analysis (Classification)

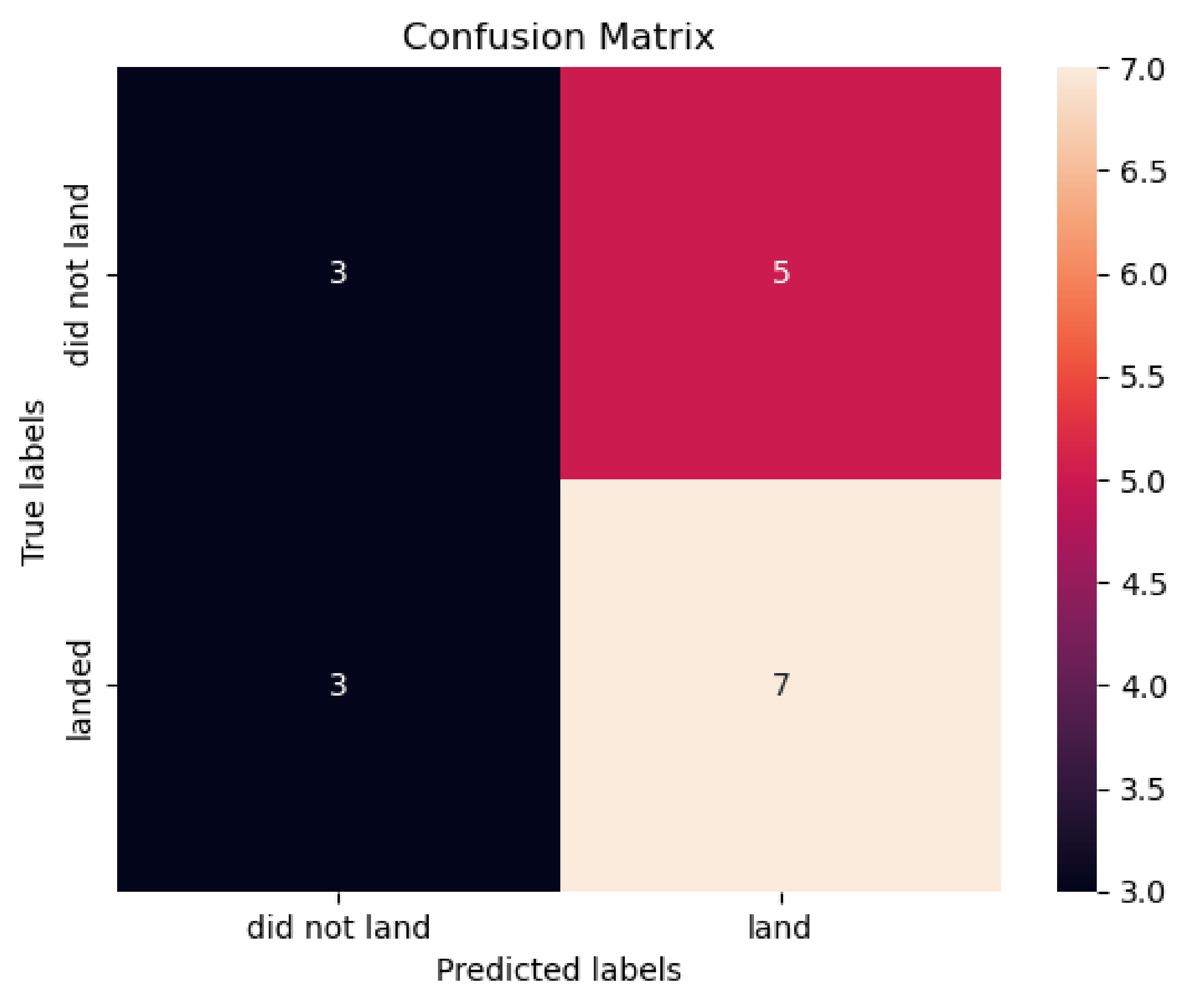
Classification Accuracy

- Built model accuracy varies by classification type
- Accuracy is similar for all models
- Accuracy is highest for decision trees and lowest for logistic regression
- Results of other researchers may be different due to stochastic results of iterations



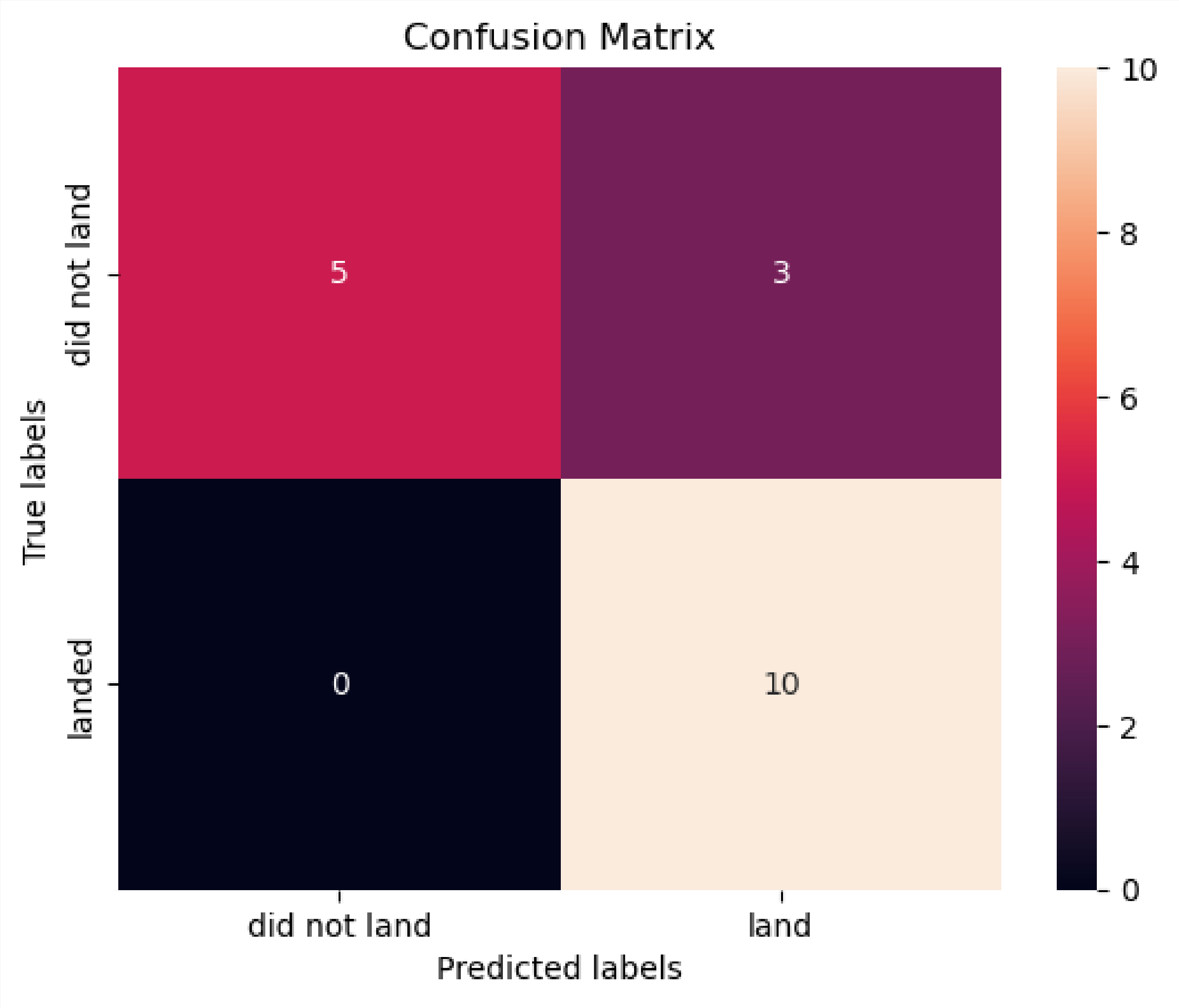
Evaluation of the Decision Tree Model

- The decision tree model showed the best performance with regard to accuracy
- Five false positives were predicted
- Three false negatives were shown



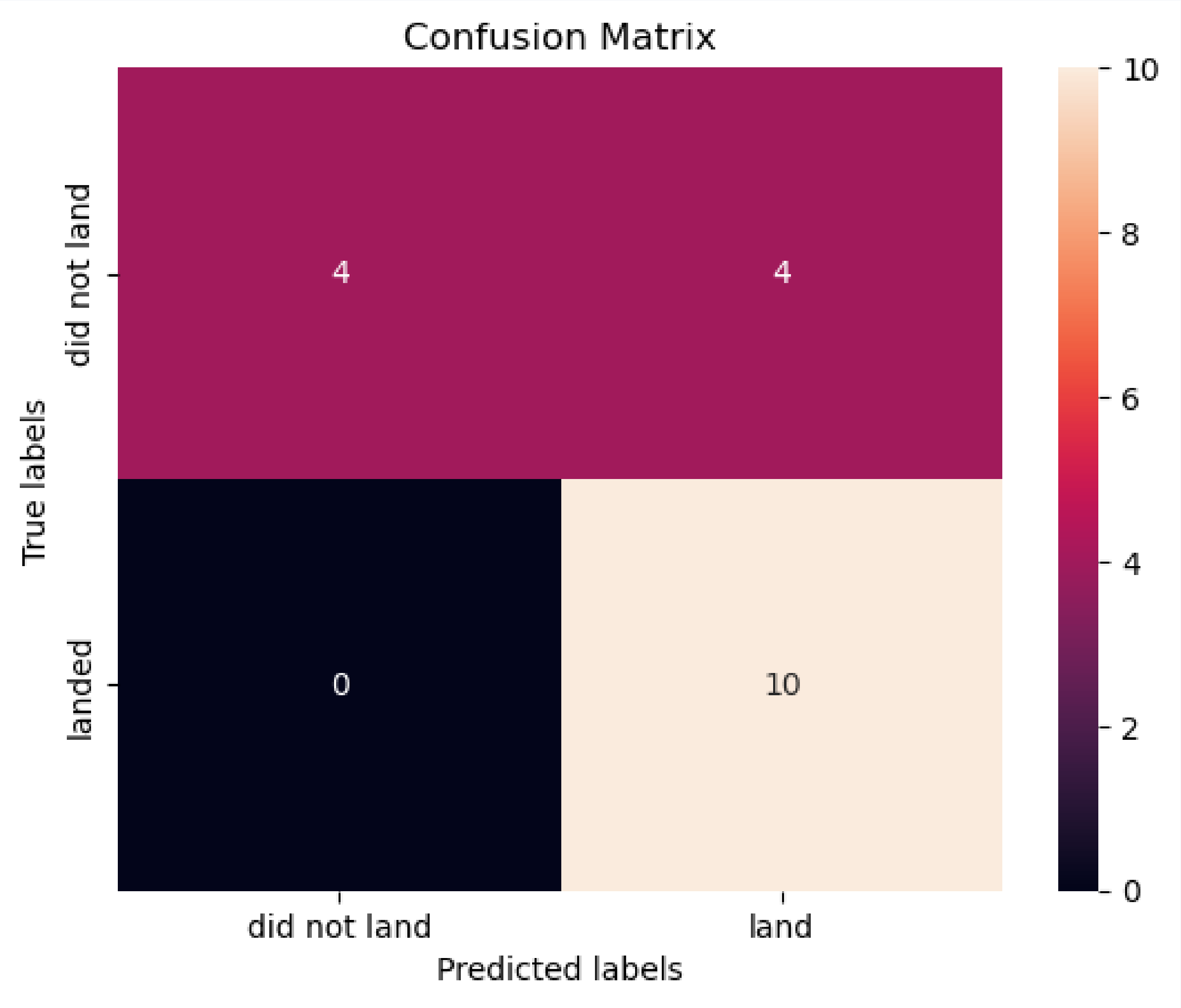
Evaluation of the Logistic Regression Model

- The logistic regression mode performed worse than the decision tree model
- There were three false positives
- Zero false negatives existed



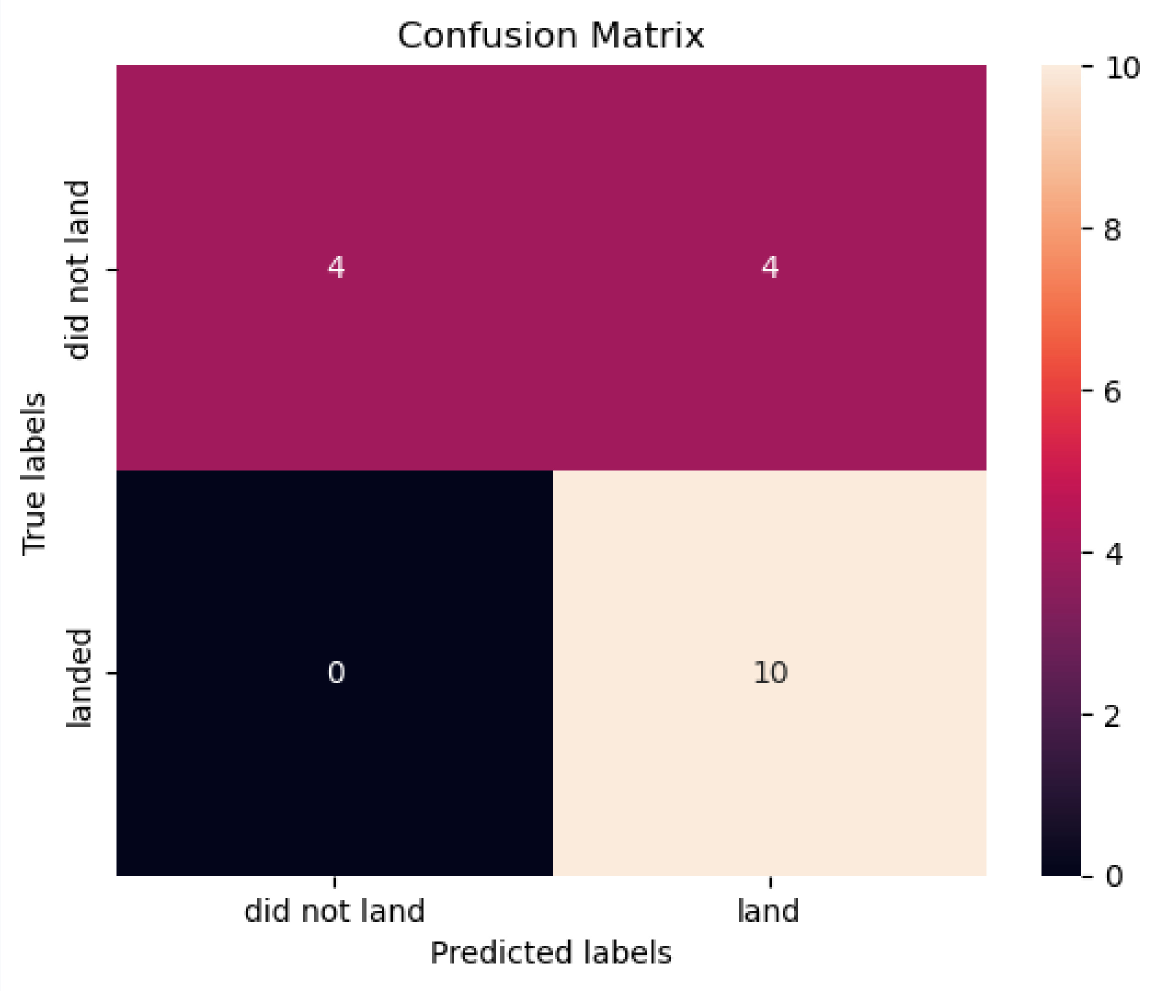
Evaluation of the SVM Model

- The SVM model was inferior in performance to the decision tree model
- Four false positives were predicted
- Zero false negatives were predicted



Evaluation of the K Nearest Neighbors Model

- The KNN model was inferior to the decision tree model with regard to accuracy
- False negatives matched those of the SVM model
- False positives matched those of the SVM model



Conclusions

- Probability of successful launches increased between 2013 and 2020
- Kennedy Space Center showed the highest rate of launch successes
- SSO and VLEO orbits have the highest overall success rates
- Different orbits are associated with different payload ranges when successful
- The number of successful landings on drone ships exceed those of ground pads
- In general, launch sites are far from cities, close to coastlines, and near purpose-built airstrips
- B4 boosters have high rates of success for a wide range of payloads
- The decision tree model is most useful at predicting launch success

Appendix

All assets relating to this project may be accessed at
https://github.com/mcselkirk/testrepo

Thank you!