IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Auburn Barron-Lutzross
— –, 2024

Outline

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

Executive Summary

- Methodologies
  - Data collected from SpaceX API and Wikipedia
  - Data wrangling for EDA and to determine labels for model training
  - SQL for analysis and insights
  - Data visualization using …
  - Basic statistical analysis
  - Machine Learning Prediction
- Results
  - Falcon 9 launch success increased progressively between 2013 and 2020
  - Orbits ES-L1, GEO, HEO, and SSO are 100% successful

Introduction

- Project background and context
  - The goal of the project is to evaluate the viability of a hypothetical program SpaceY to compete with SpaceX
- Questions to address
  - How do attributes like payload mass, launch site, and orbit affect the success of Falcon 9 first stage landings?
  - Does the success rate of landings increase with time?
  - What is the best algorithm to predict landing success?

Section 1

Methodology

Methodology

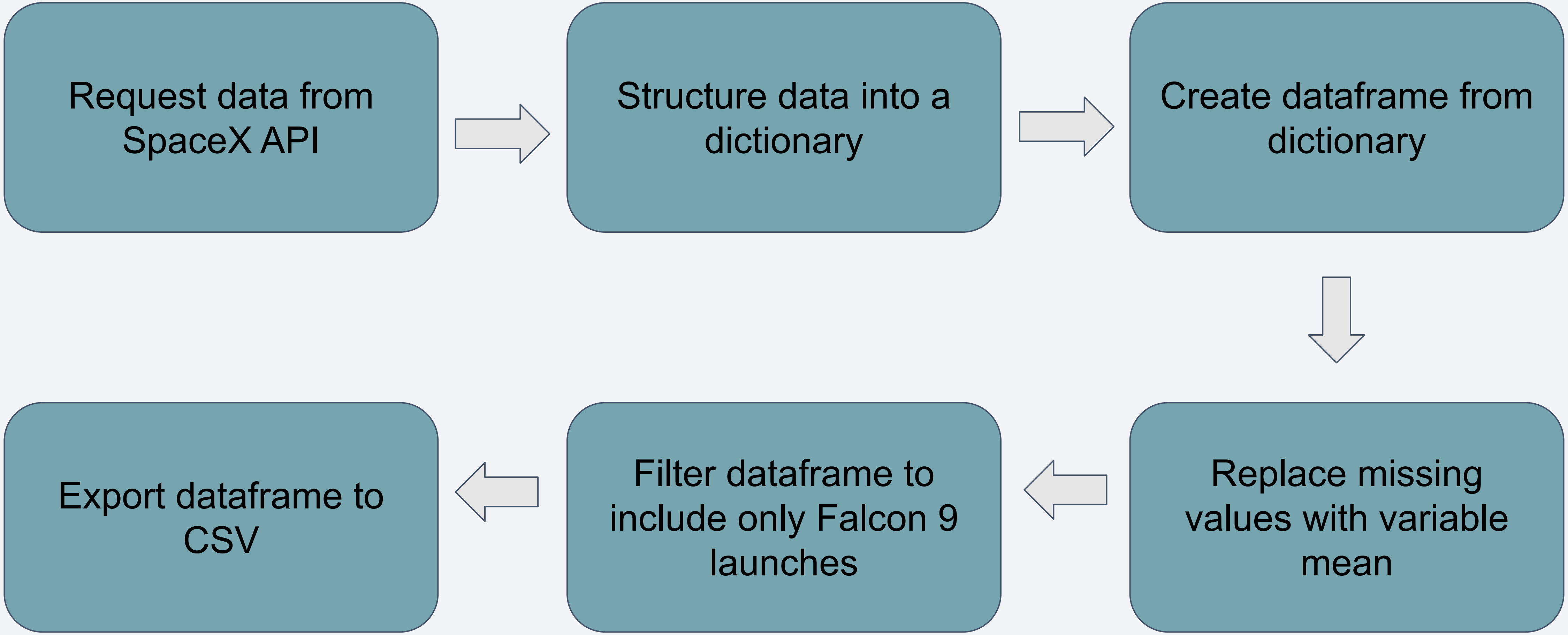
Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Filter data
  - Handle missing values
  - One-Hot Encoding converting categorical data to binary for machine learning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating classification models

Data Collection

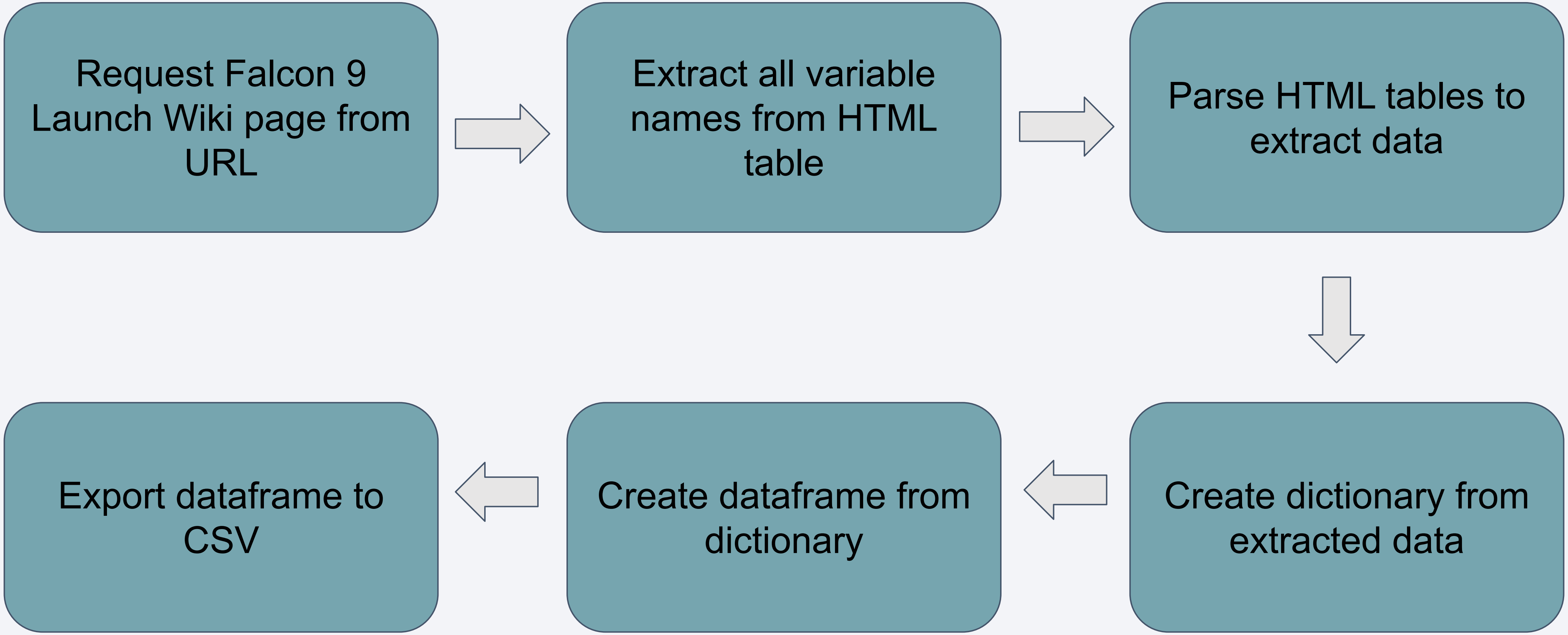
- To get all the data needed for a complete analysis, two sources of data were accessed. These were accessed in different ways for different variables.
- SpaceX REST API
    - Accessed through API requests
    - Flight Number, Launch Date, Booster Version, Payload Mass, Orbit, Launch Site, Launch Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Wikipedia entry on SpaceX
    - Accessed with web scraping
    - Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Launch Date, Launch Time

Data Collection – SpaceX API



https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Collection – Web Scraping



https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

To understand what makes a launch successful, I first performed some Exploratory Data Analysis. This included calculating the number of launches per site, the number of instances of each type of orbit, and mission outcome per orbit type. Mission outcome in the original dataset included the location as well as success. This was converted to a binary 0 or 1 for failed or successful landings.

Exploratory Data Analysis → Launched per site → Instances of orbit type → Instances of orbit type → Mission outcome per orbit type → Create landing outcome variable

https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots, bar charts, and line graphs were used to explore the relationship between the following feature pairs:
    - Payload Mass and Flight Number
    - Launch Site and Flight Number
    - Launch Site and Payload Mass
    - Orbit and Flight Number
    - Payload Mass and Orbit

https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda.ipynb

EDA with SQL

- The following SQL queries were performed
    - Names of unique launch sites in the space mission
    - Top five launch sites with names beginning in ‘CCA’
    - Total payload mass carried by boosters launched by NASA (CRS)
    - Average payload mass carried by booster version F9 v1.1
    - Date of first successful ground pad landing
    - Names of boosters with success in drone ship and having a payload mass between 4000 and 6000 kg
    - Total number of successful and failed missions
    - Names of booster versions carrying the maximum payload mass
    - Booster versions and launch site names for failed landing outcomes in drone ship in the year 2015
    - Landing outcomes between 2010-06-04 and 2017-03-20 ranked by count

https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for each Launch Site
  - Marker with circle, popup label, and text label for NASA Johnson Space Center
  - Markers with circles, popup labels, and text labels for all launch sites
- Color-coding by launch outcome for each Launch Site
  - Identify which sites have the highest success rates
- Distance between a Launch Site and nearest significant proximites
  - Determine whether launch sites are near railroads, highways, and coasts
  - Assess general distance to cities

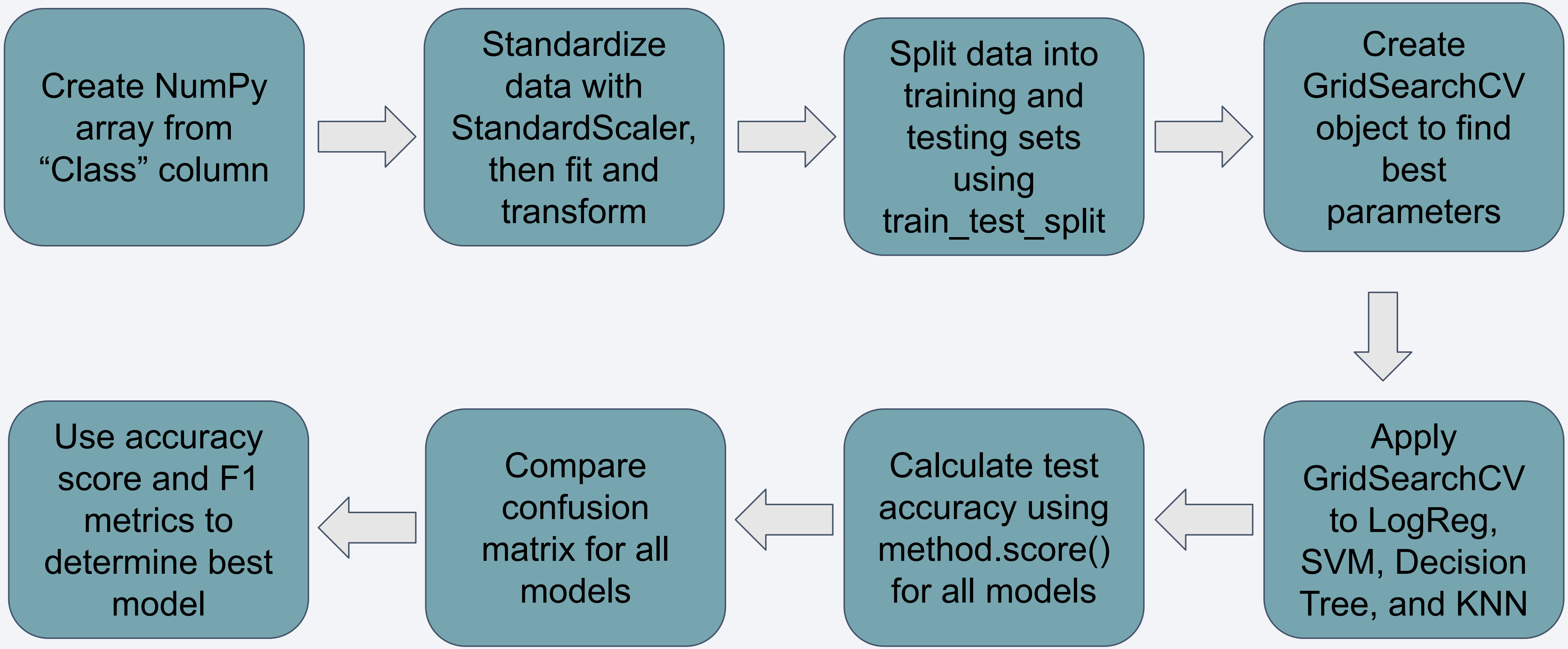
https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Launch Site Dropdown to select individual launch sites or all sites
- Pie Chart to view successful vs. failed launches
- Scatter Plot to demonstrate the relationship between Outcome and Payload Mass for different booster versions

https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/auburnbl/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

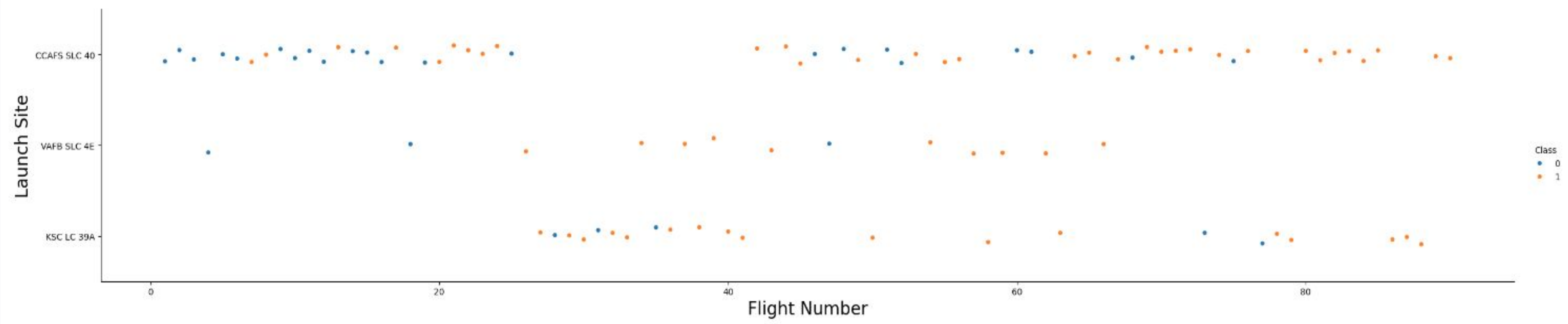
Results

- Exploratory data analysis results
- Launch site evaluation
- Predictive analysis results

Section 2

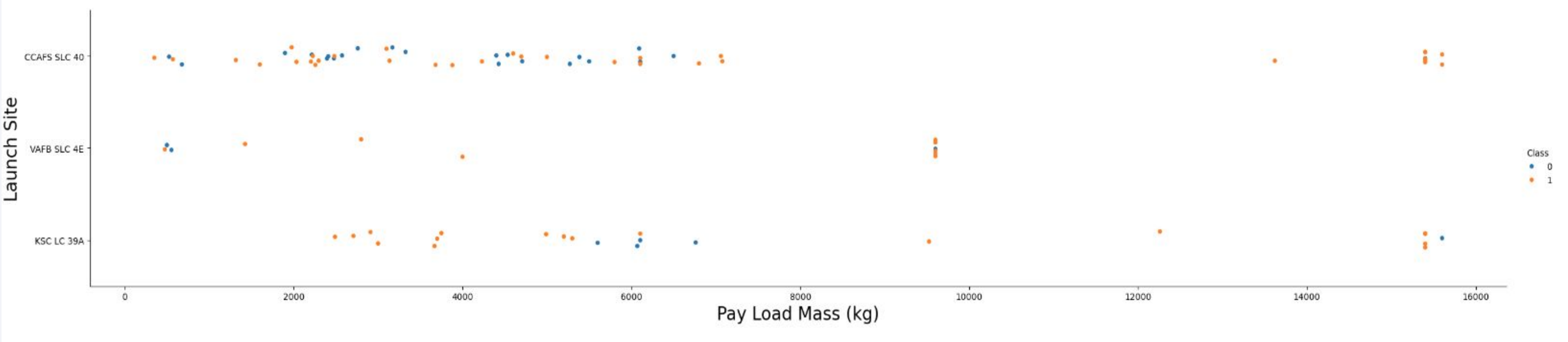
Insights drawn from EDA

Flight Number vs. Launch Site



- The most commonly used launch site is CCAFS SLC 40 with about half of all the SpaceX launches
- The earliest launches across all sites failed (Class 0) while the latest launches all failed. It is clear that success rate has improved over time.
- Sites VAFB SLC 4E and KSC LC 39A have higher success rates.

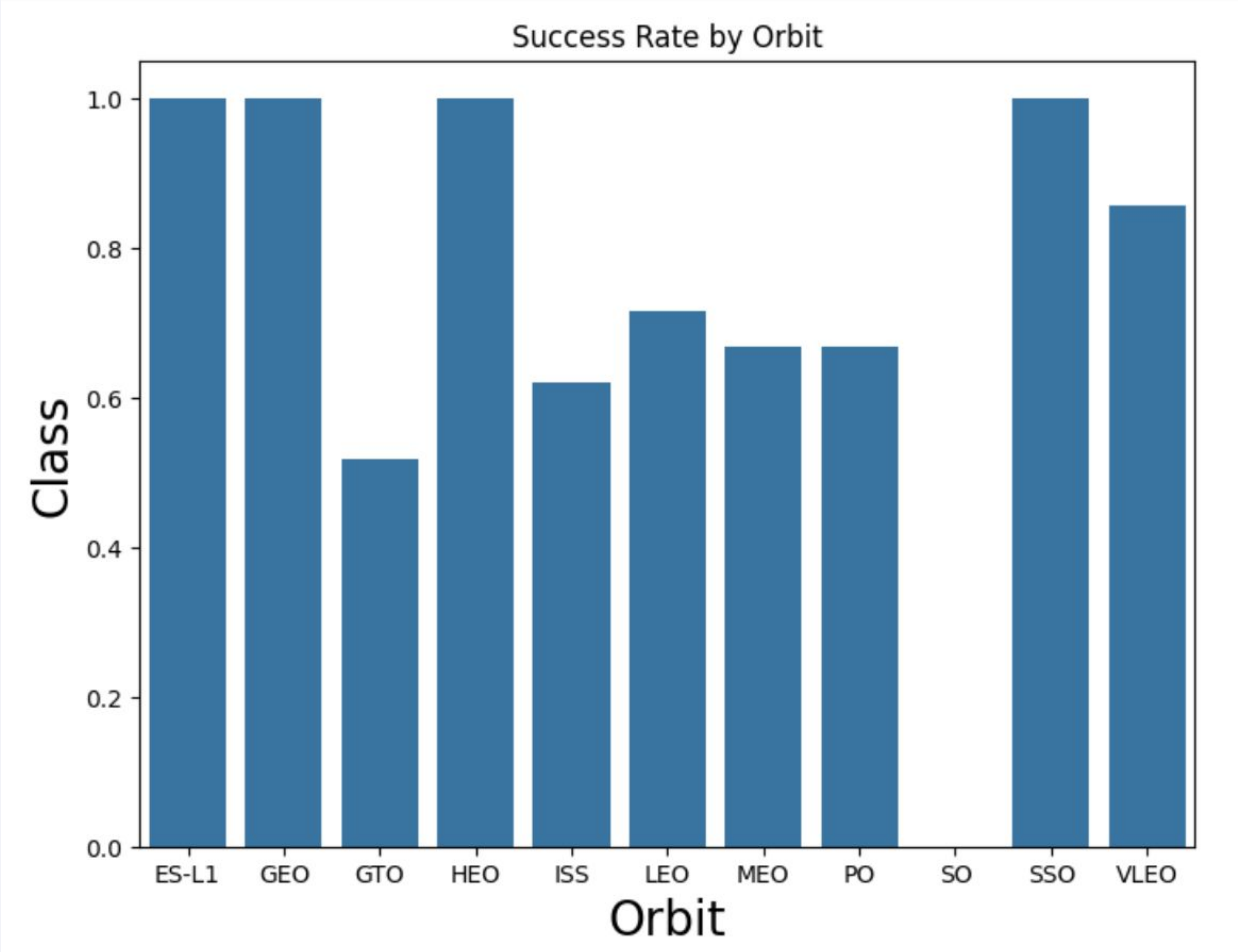
Payload vs. Launch Site



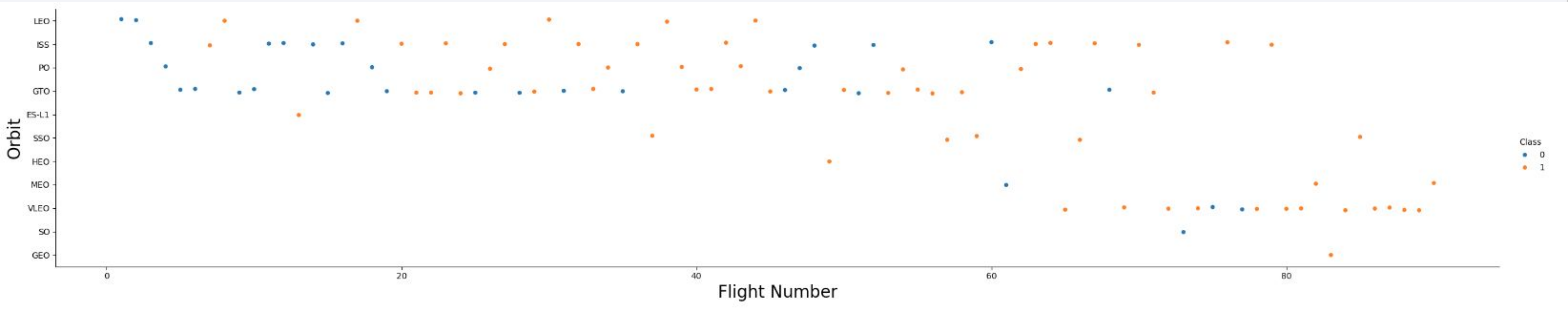
- Success rates improve with payload across launch sites.
- Nearly all of the launches with a payload mass of more than 7,000 kg were successful.
- KSC LC 39A also has a 100% success rate for payloads of less than 5,000 kg.
- VAFB SLC 4E has had no launches over 10,000 kg.

Success Rate vs. Orbit Type

- All but one of the orbits have over 50% success rates.
- Orbits with 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbit VLEO is close behind with a success rate over 80%.
- Orbit SO has a 0% success rate.

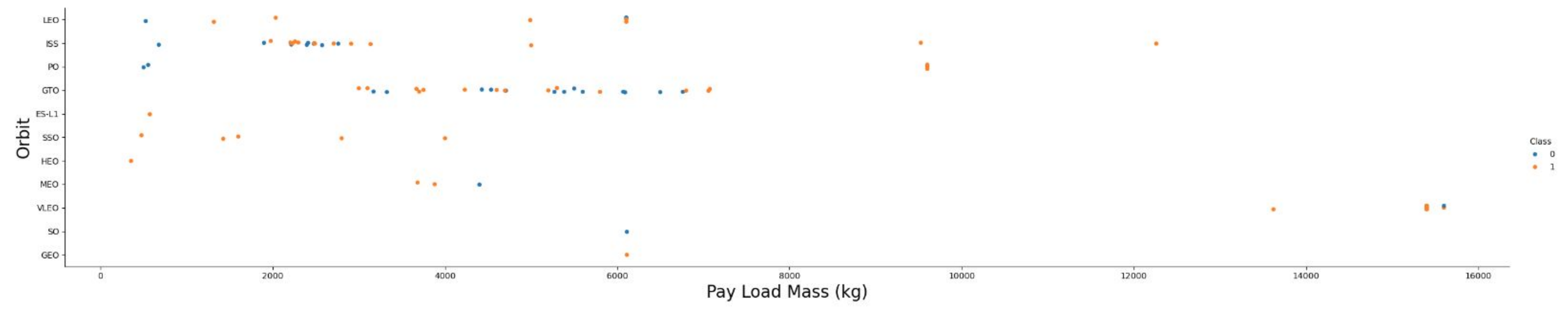


Flight Number vs. Orbit Type



- Success rate improved over time for all orbits.
- Orbits SO and GEO are the most recently developed types, and each has only had one launch.
- Despite becoming successful, the earliest orbit (LEO) has not been attempted in any recent launches.
- The second earliest orbit (ISS) is still attempted regularly.
- VLEO orbit appears to be the most common these days.

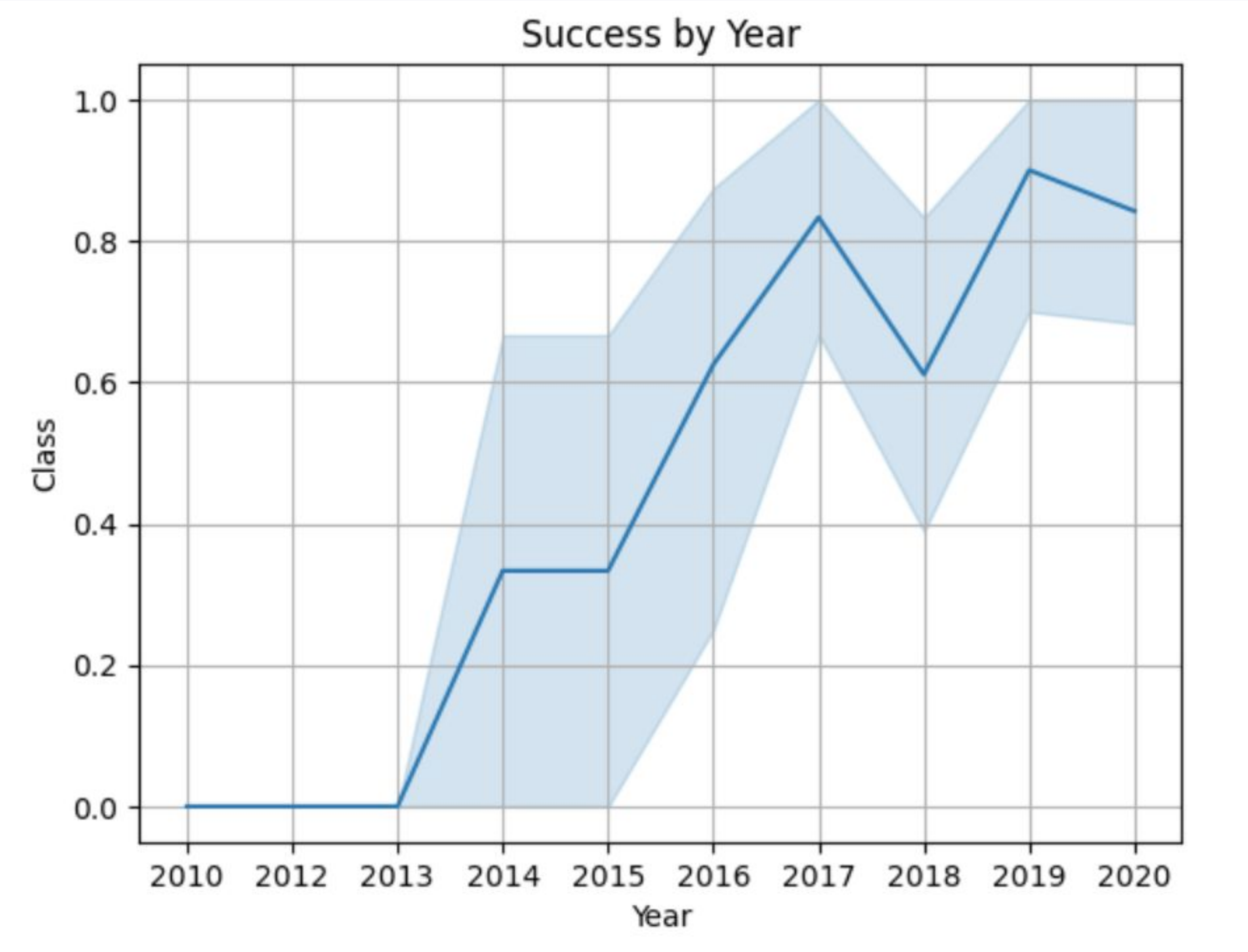
Payload vs. Orbit Type



- Orbit VLEO is the only one with payloads over 13,000 kg.
- Most orbits have payloads under 8,000.
    - ISS, PO, and VLEO are the only exceptions.
- Orbit GEO has no apparent relationship between success rate and payload mass.

Launch Success Yearly Trend

- The first successful launch was in 2013.
- There has been a progressive increase in success, with dips in 2017 and 2019.



All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

Unique launch site names are displayed by selecting distinct occurrences of the variable “launch_site” from the dataset.

Launch Site Names Begin with 'CCA'

```
%%sql

select *
from SPACEXTABLE
where Launch_Site like 'CCA%'
limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_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

By selecting all instances where the name of the launch site starts with ‘CCA’ we can identify all Cape Canaveral launches. The display is limited to five samples.

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

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

total
45596

The total payload for NASA launches is calculated by getting the sum of all payloads where the customer is ‘NASA (CRS)’.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS average
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

average
2534.6666666666665

The average payload mass of booster version F9 v1.1 is calculated as the average of all instances where the Booster_Version variable starts with ‘F9 v1.1’.

First Successful Ground Landing Date

```
%%sql select min(date) as first_successful_landing
from SPACEXTABLE
where landing_outcome = 'Success (ground pad)'
```

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

first_successful_landing
2015-12-22

The date for the first successful ground language was retrieved as the minimum (ie: earliest) date where the landing outcome was ‘Success (ground pad)’.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select Booster_Version
from SPACEXTABLE
where  Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Above are the four booster versions where the landing outcome was a successful drone ship landing, and the payload was between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
Select Mission_Outcome, Count(*) as count
From SPACEXTABLE
Group by Mission_Outcome
```

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

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

This shows the number of launches per mission outcome. It shows a total of 1 failure and 100 successes.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

These are all the boosters which carried the maximum payload mass within the dataset.

2015 Launch Records

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

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

substr(Date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Above is the record for the launches in 2015 showing the date, outcome, booster version, and launch site.

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) as count
FROM SPACEXTABLE
WHERE Date between '2010-06-04' and '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(*) DESC
```

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

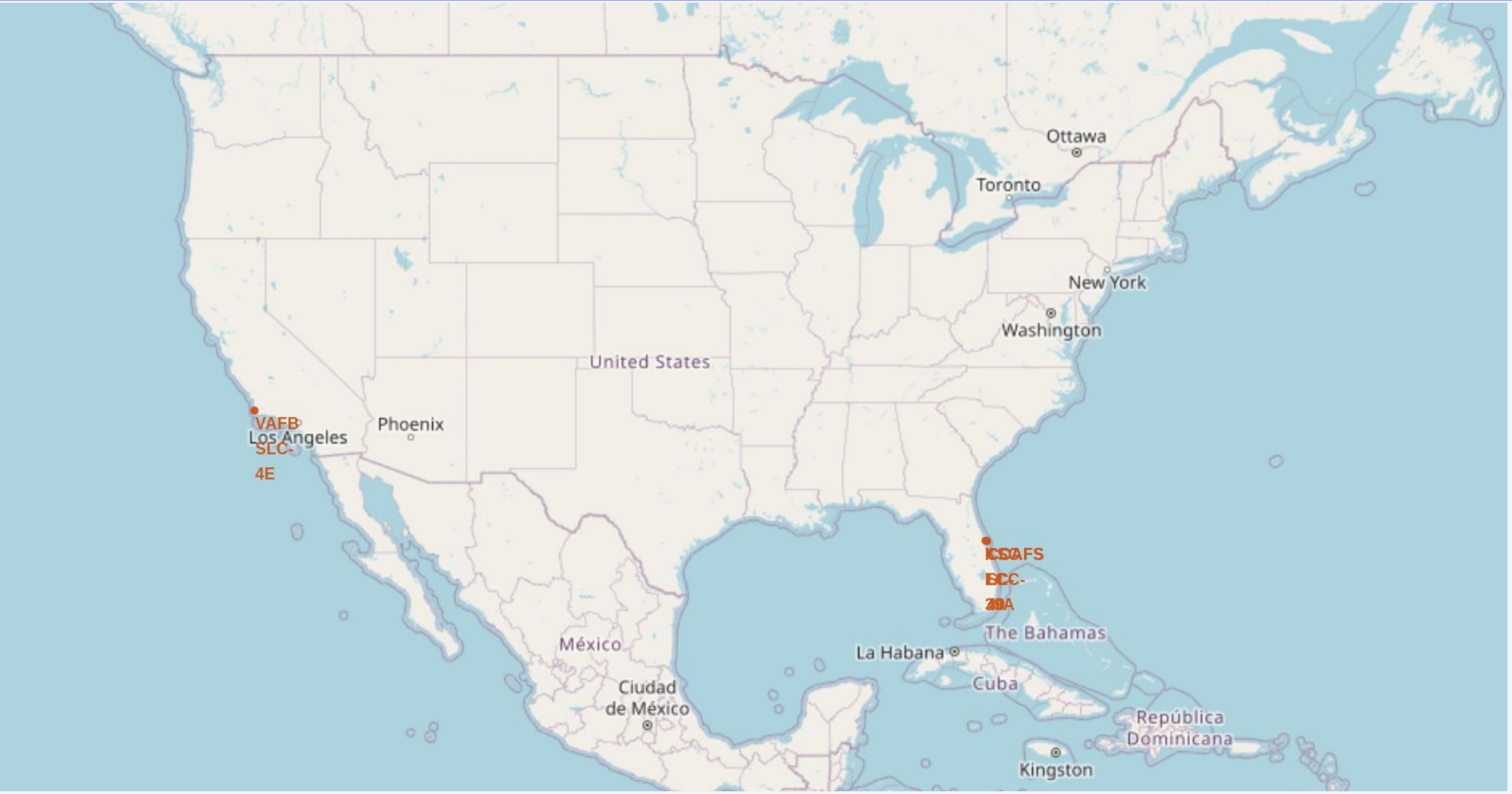
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

A list of the landing outcomes between June 4th, 2010 and March 3rd, 2017, ranked by the number of launches with each outcome.

Section 3

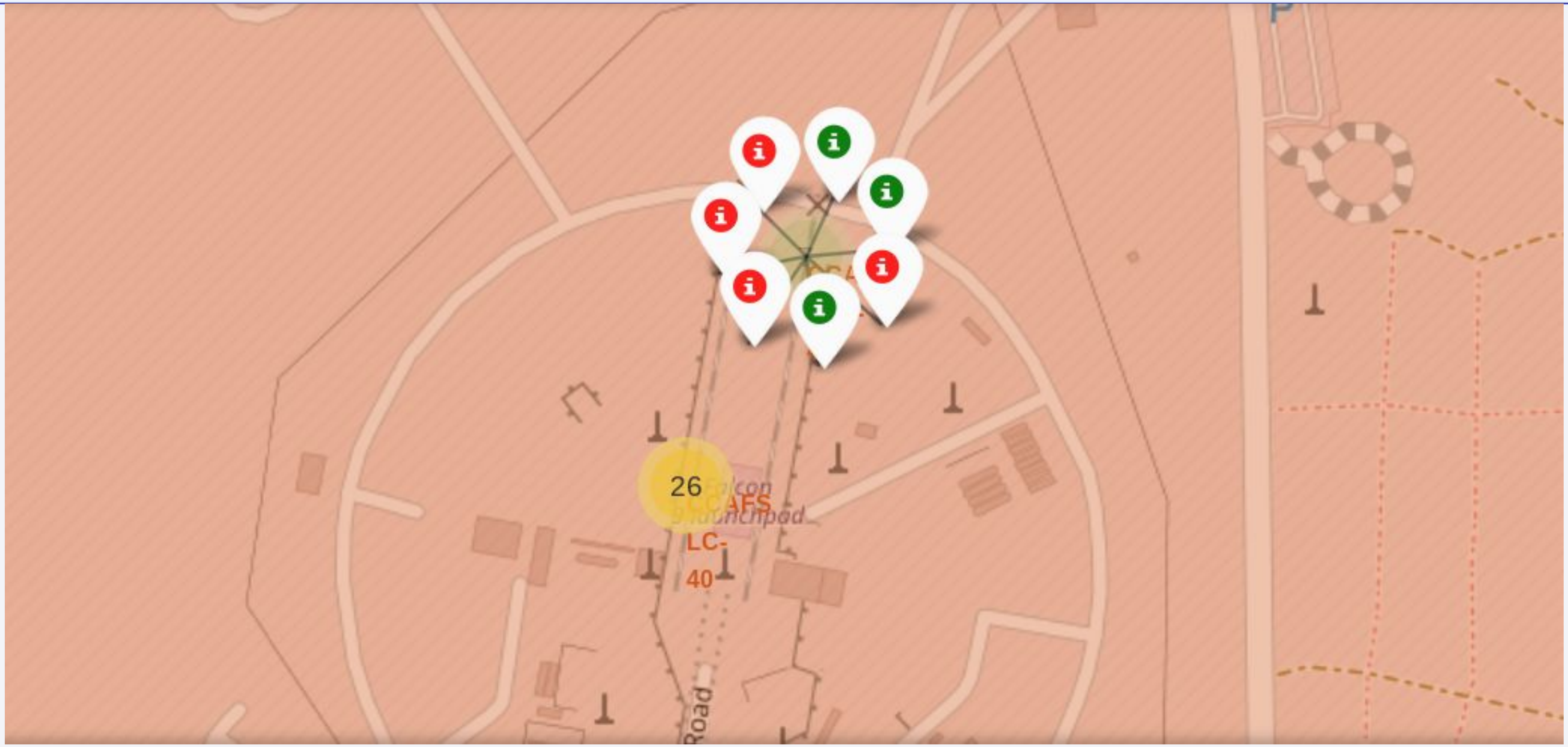
Launch Sites Proximities Analysis

Lunch Sites in the US



- All launch sites are closer to the equator than most of the USA.
- All sites are near oceans for less destructive failed launches, as well as relatively close to roads and railways for access.

Launch Fails and Successes

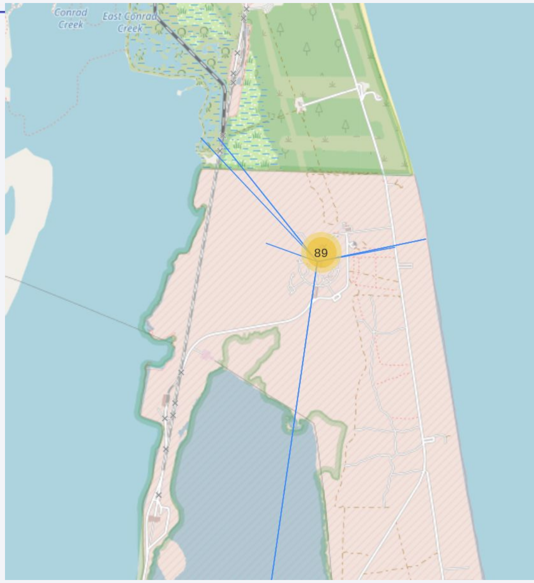


- Using color markers allows for easy identification of successes and fails within a launch site.
- This example, CCAFS LC 40 has a success rate of 57.1%

Logistics and Safety

- Like all sites, CCAFS LC 40 is relatively close to roads, railroads, and coastline, and is positioned far from any major cities for safety.

- This map shows distances between the site and the nearest city, highway, railroad, and coast



Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

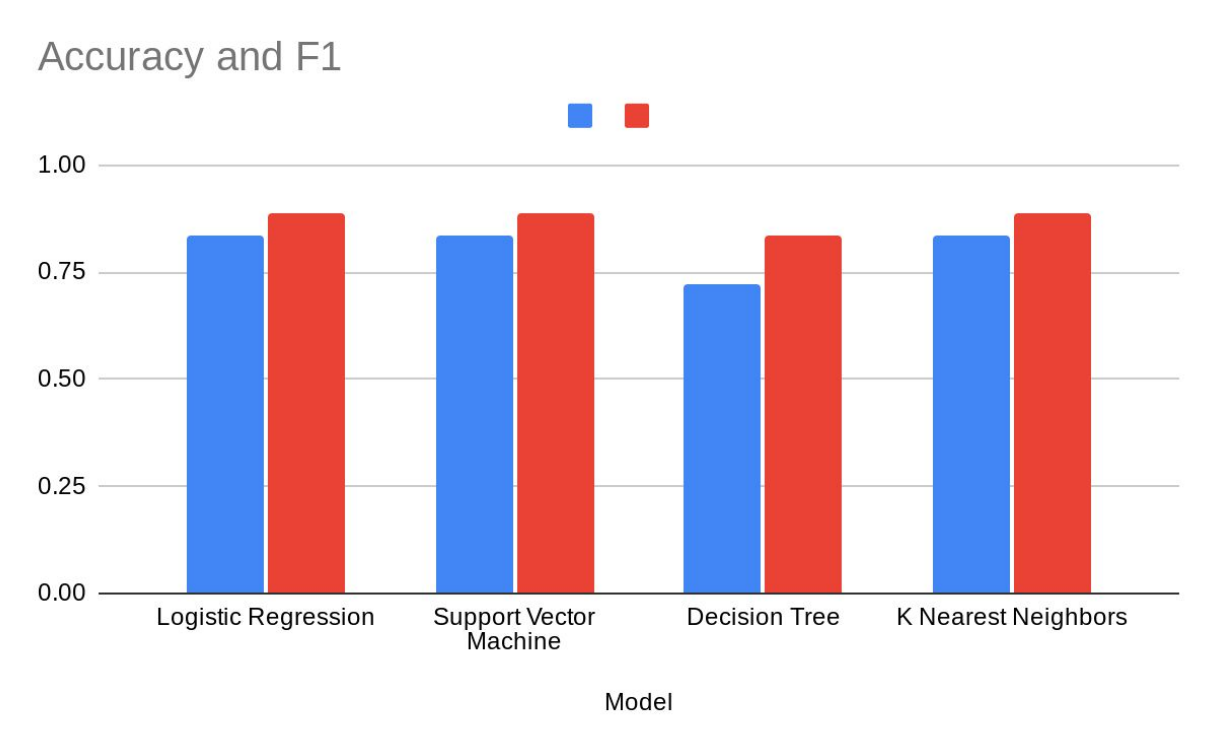
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

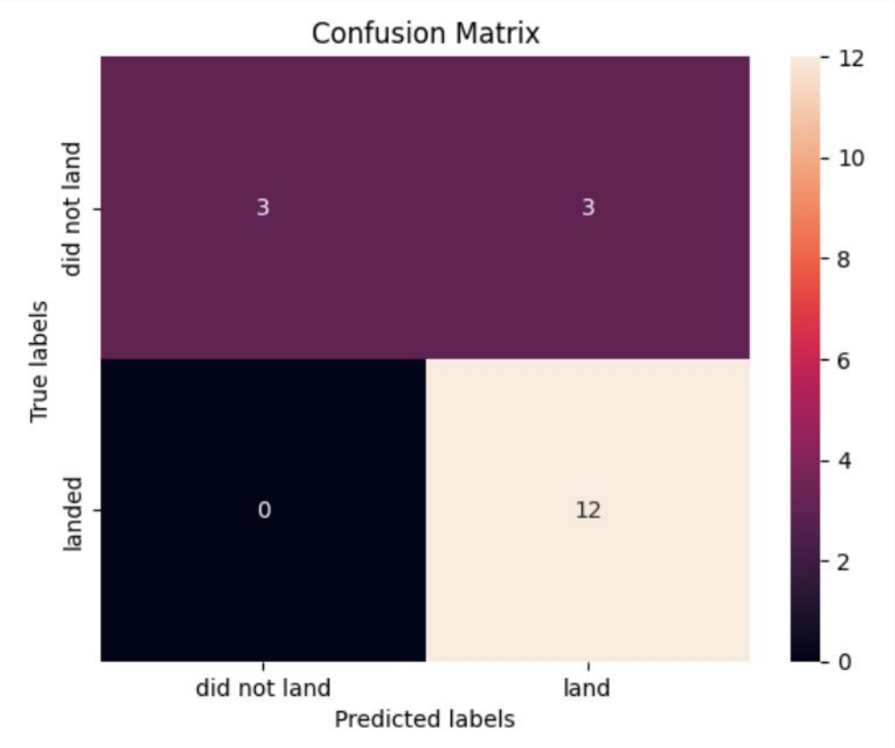
Classification Accuracy

- Bar chart shows Accuracy and F1 scores for:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbors
- Decision Tree model has the lowest score for both Accuracy and F1. All others are equally good.

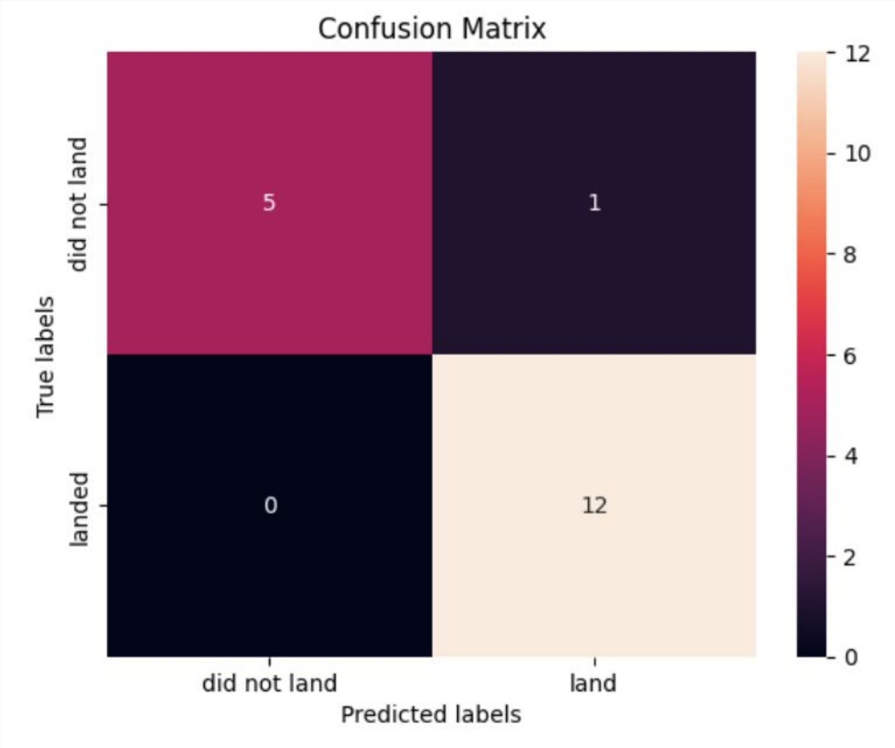


Confusion Matrix

- The confusion matrices show that all models struggled with false positives. The decision tree model was a worse predictor than the other three.



Decision Tree



Logistic Regression (identical to KNN and SVM)

Conclusions

- Predictive analysis shows that all models struggled with false positives, but the Decision Tree model was less accurate than the other three models tested (K-Nearest Neighbor, Logistic Regression, and Support Vector Machine
- Lower payloads have higher likelihood of success
- KSC LC 39A has the highest success rate of all sites
- Launch sites are located relatively close to the equator and are on the coast. They have close proximity to highways and railroads, but are far from major urban areas.
- The success of launches improves over time with advancements in processes and rockets
- Orbits ESL, GEO, HEO, and SSO have 100% success rates.

Appendix

- I was unable to successfully create a dashboard, so I have left these slides out of the presentation.
- Thank you to the instructors, Coursera, IBM, and fellow Coursera members taking this course!

Thank you!