IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

David Sekkes
May 10, 2025

Outline

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

Executive Summary

Summary of all methodologies

- Data collection using the SpaceX API and Web scrapping the Space X Wiki page :
- Perform data wrangling cleaning up missing values an creating an Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis by building and evaluated four different classification models identifying the most accurate

Summary of all results

By examining the parameters of an individual launch, we can predict with 83.3% accuracy if the landing will be successful.

- Heavier payloads are most successful
- Payloads sent to low Earth orbit are most successful
- Payloads launched from Kennedy Space Center are most successful

Introduction

Falcon 9 rocket launches cost $62M, over $100M less than competitors in large part because of Space X's ability to reuse the first stage of the rocket by landing it intact.

By examining open-source rocket launch data we wanted to predict the outcome of a launch.

Section 1

Methodology

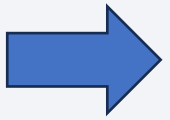
Methodology

Executive Summary

- Data collection methodology:
  - GET request to the SpaceX API and Web scraping from the SpaceX Wikipedia page
- Perform data wrangling
  - Removed columns with null values
  - Created a column called Outcome to store successful launch status
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built and evaluated four different classification models identifying the most accurate

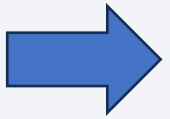
Data Collection

The **SpaceX API** provided launch data that included the booster version, payload, orbit, launch site, outcome, and several other details for each launch



This data was extracted to a Pandas data frame using the built in in JSON reader

Launch records were found on the SpaceX **Wikipedia** page in a table.



This data was scraped using BeautifulSoup into a Pandas data frame for analysis.

Data Collection – SpaceX API

1. Request and parse the SpaceX launch data using the GET request
2. Filter the dataframe to only include Falcon 9 launches
3. Replaced Missing Values with the mean of the payload values
4. Exported the datadrame as a CSV

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod1A-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Using BeautifulSoup: Extracted the Falcon 9 launch records HTML table from Wikipedia. Then, the table was parsed and converted it into a Pandas data frame

1. Requested the Falcon9 Launch Wiki page from its URL
2. Extracted all column/variable names from the HTML table header
3. Created a data frame by parsing the launch HTML tables
4. Exported the dataframe to a CSV

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod1B-jupyter-labs-webscraping.ipynb

Data Wrangling

- Loaded the data and examined is data types and structure
- Identify any missing values, only LandingPad had null values
- Calculated the number of launches at each unique site
- Calculated the number of occurrences and outcome for each orbit type
- Created an Outcome column to show if the landing was successful or not

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod1C-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot of Flight Number vs Payload Mass- We see that as the flight number increases, the first stage is more likely to land successfully
- Scatter plot of Flight Number vs Launch Site- Successes increased in frequency across all sites as the flight number increased. VAFG SLC4E was the least used site, while CCAFS SLC40 was the most used
- Scatter plot of Payload Mass vs Launch Site – VAFB had no launches with payloads exceeding 10,000 kg
- Bar Chart showing the mean success rate for each orbit type – ES-L2, GEO, and SSO all had the completely successful attempts while the remaining orbit types had about a 60% success rate
- Scatter plot of the Flight Number vs Orbit type – This shows that GEO and SSO orbit attempts were not made in the early development stages
- Scatter plot of Payload Mass vs Orbit – this shows that most attempts were below 8,000 kg, but higher mass payloads were almost exclusively got VLEO

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod2B-edadataviz.ipynb

EDA with SQL

- Displayed all the unique launch sites
- Examined the launch sites with ‘CCA’ in their name
- Calculated the total payload mass launched by NASA
- Calculated the average payload carried by booster version F9 v1.1
- Extract the date od the first successful outcome using a ground pad
- List the names of the boosters that had successful drone ship landings with a payload mass between 4,000 and 6,000 pounds
- Counted the number of successful missions
- Listed all the booster versions that launched with the maximum payload
- Listed the landing outcomes with the count of occurrences

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod2A-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In the interactive map, a marker cluster was added for each launch site that shows the successful launches (green) and failures (red).
- Lines were drawn from the Cape Canaveral Space Force Station (CCSFS) launch site to the nearest coastline, city, highway, and railroad to examine the proximity to these site
- The CCSFS is right along the coast and any launches to the east will be far enough away from the nearest city of Titusville to be safe.
- Nor should affect any critical transportation such as US 1 or railroad lines.

- https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod3A-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The Dashboard includes:
  - A pie chart showing the success for a launch site
  - A scatter plot showing the payload mass for the selected launch site
  - A slider to filter the payload mass to a specific range
- These interactive graphs can be used to explore the relationship between the success rate and payload mass for each (or all) launch sites

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod3B-Capstone_Dashboard_DAS.py

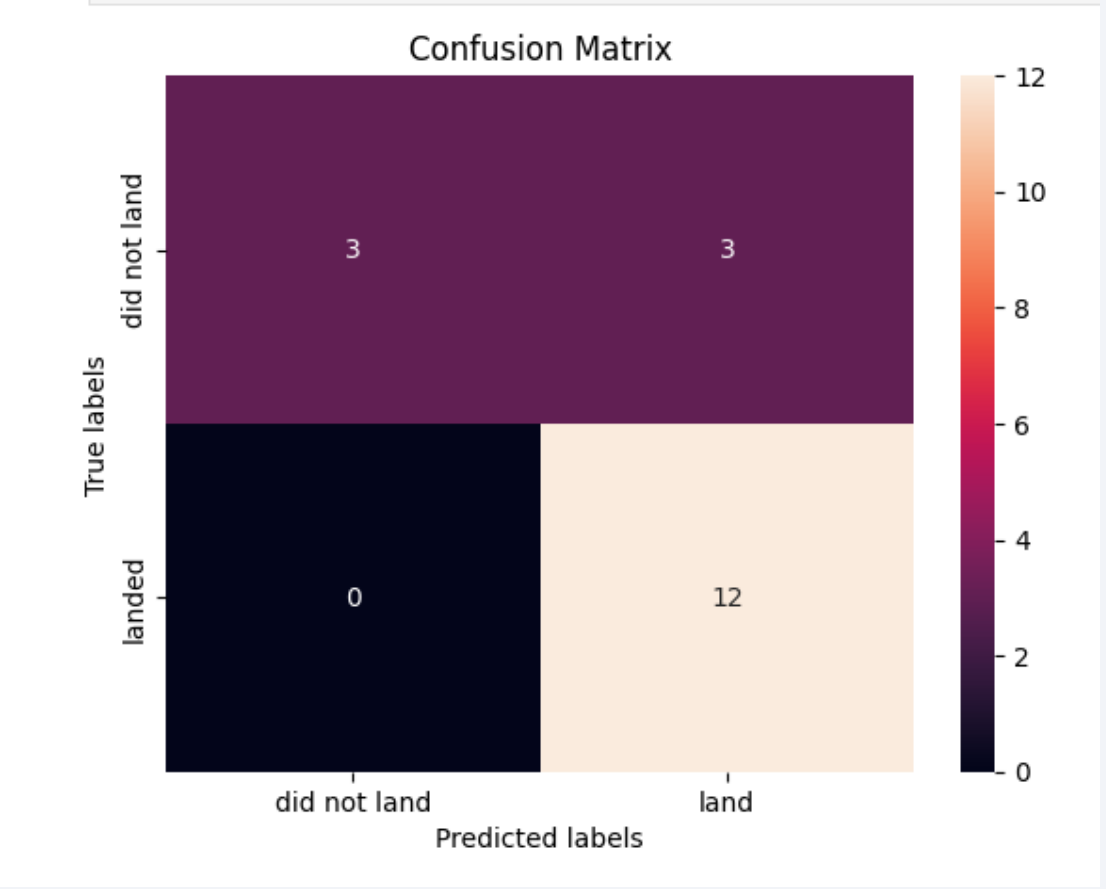
Predictive Analysis (Classification)

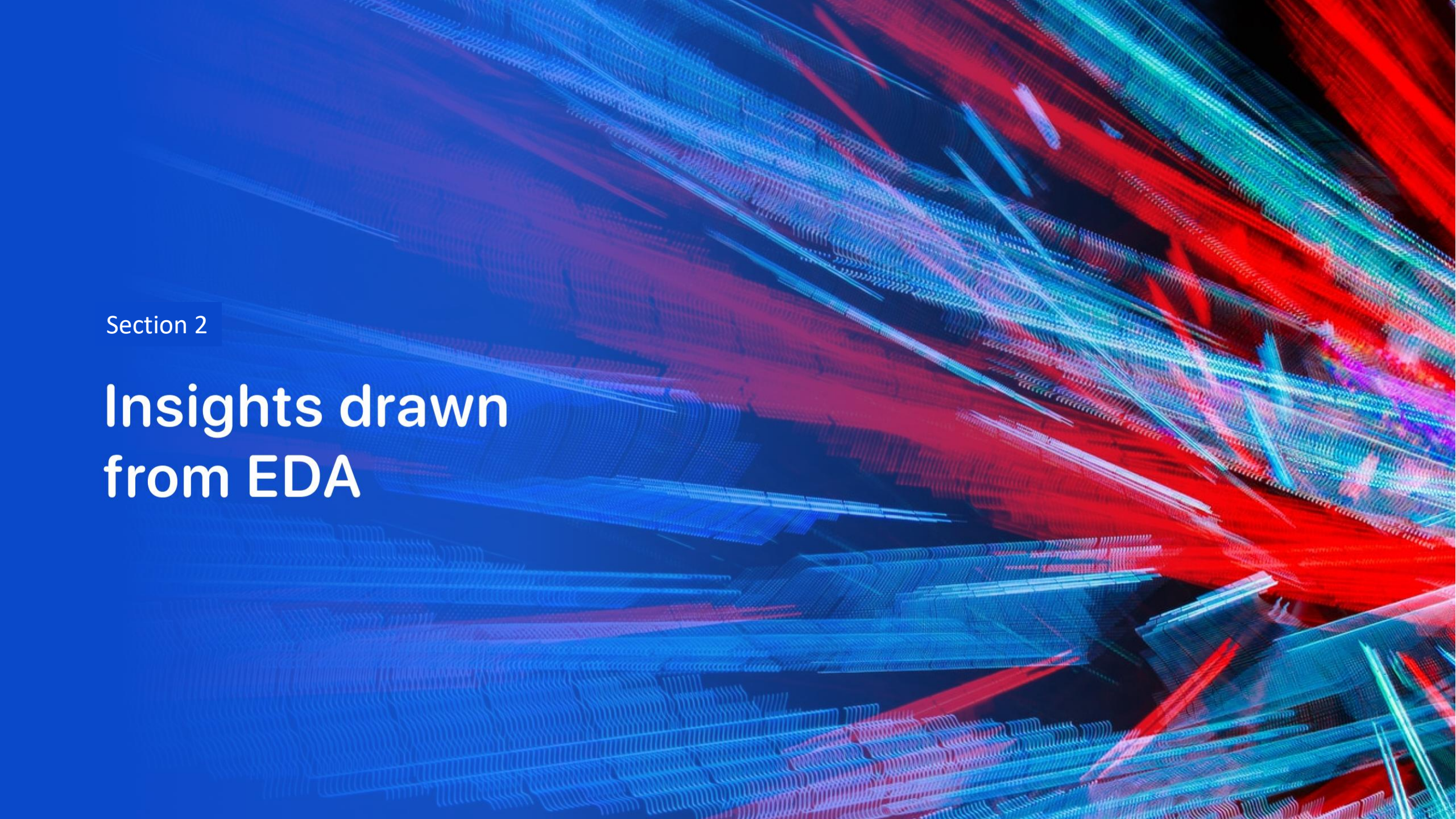
- Created a numpy array of the Class column which stores the success of the launch
- Standardized the data using the Standard Scaler
- Split the data into test and train subsets using a 20% test size
- Using the training data, a Grid Search was created to test the accuracy several different estimators
- The confusion matrix was generated, as well as an accuracy score calculated to find the best estimator

https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS/blob/main/Mod4-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

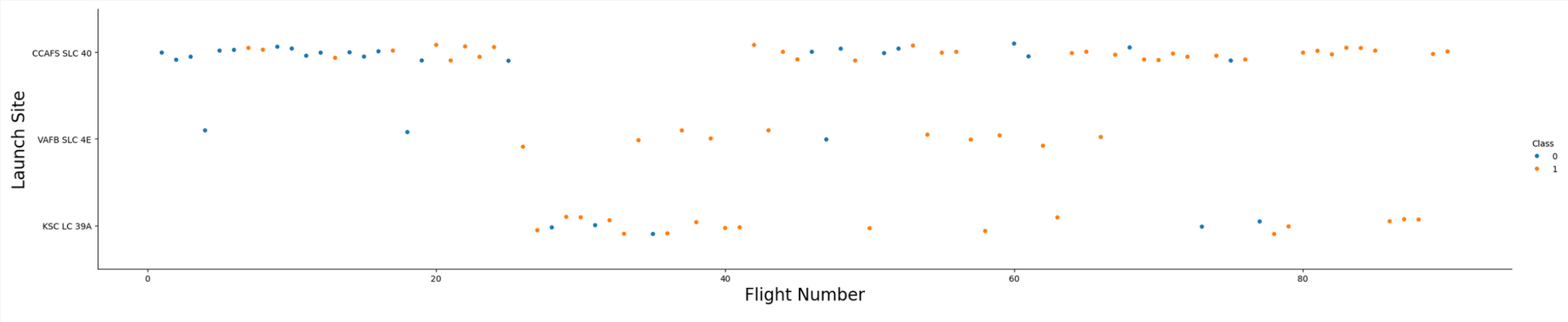
Results

- Exploratory data analysis accuracy results
  - Logistic Regression (84.64%)
  - Support Vector Machines (84.82%)
  - Decision Tree (88.75%)
  - K Nearest Neighbors (84.82%)



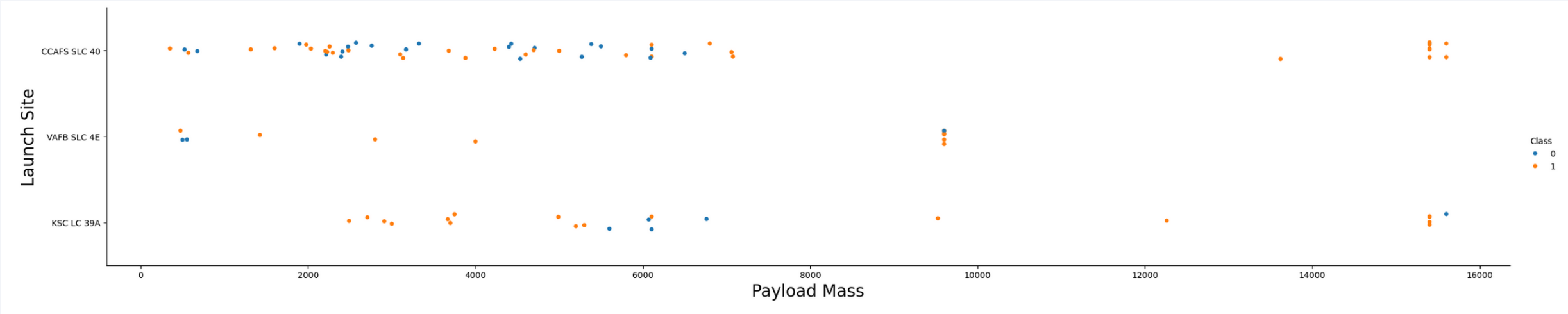

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The first 20 flights had a high number of failures (blue).
- The number of successful launches (orange) increased dramatically following flight 20.

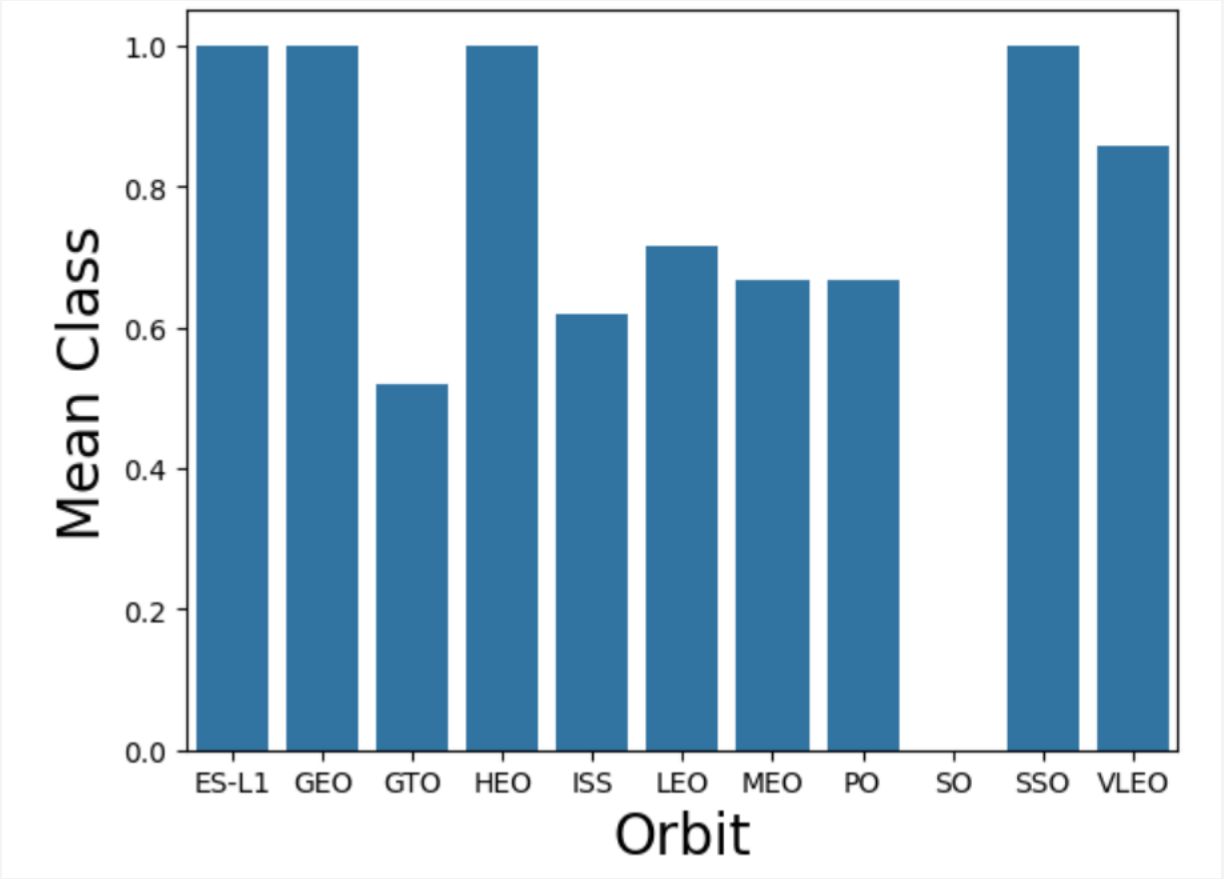
Payload vs. Launch Site



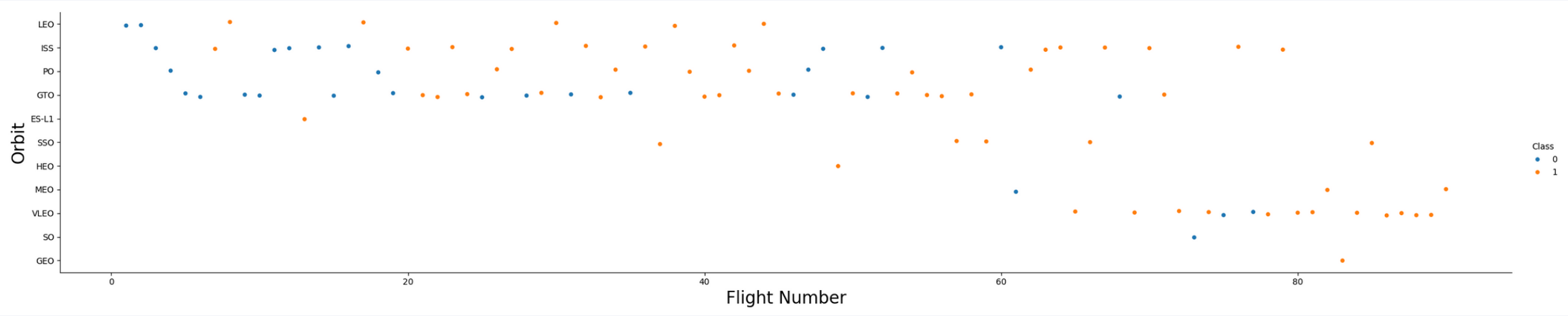
- All launches at VAFB launch site carried a payload mass of under 10,000 kg
- Max payload masses from CCAFS SLC-40 were all successful

Success Rate vs. Orbit Type

- There were 4 orbits that had a 100% success rate
  - ES-L1
  - GEO
  - HEO
  - SSO
- SO had no successful launches

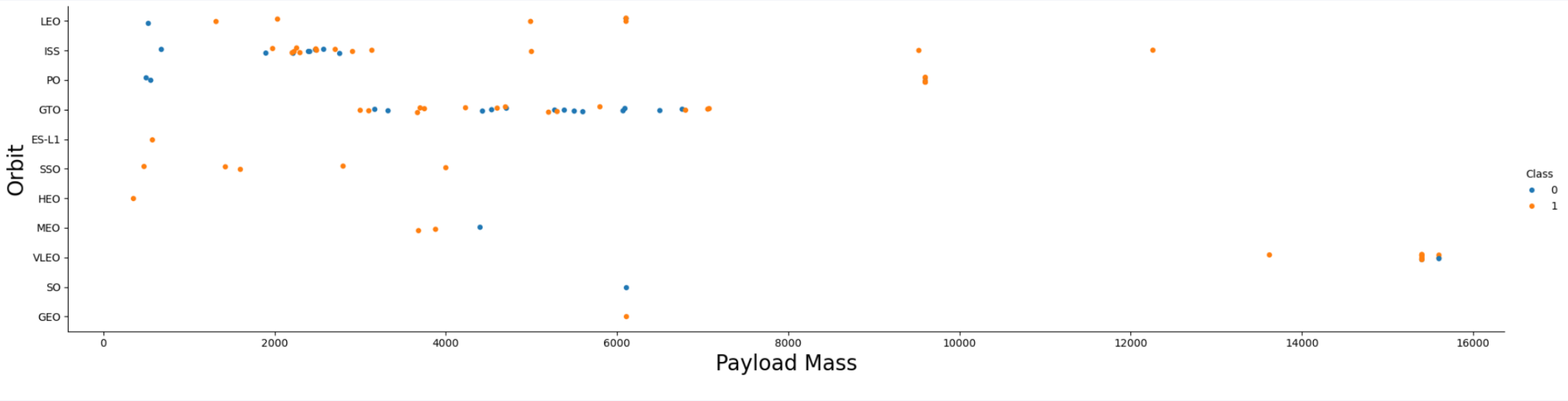


Flight Number vs. Orbit Type



- Success rate of the orbits overall got better over time
- LEO stopped being attempted around 50 flights for more concertation on VLEO and ISS

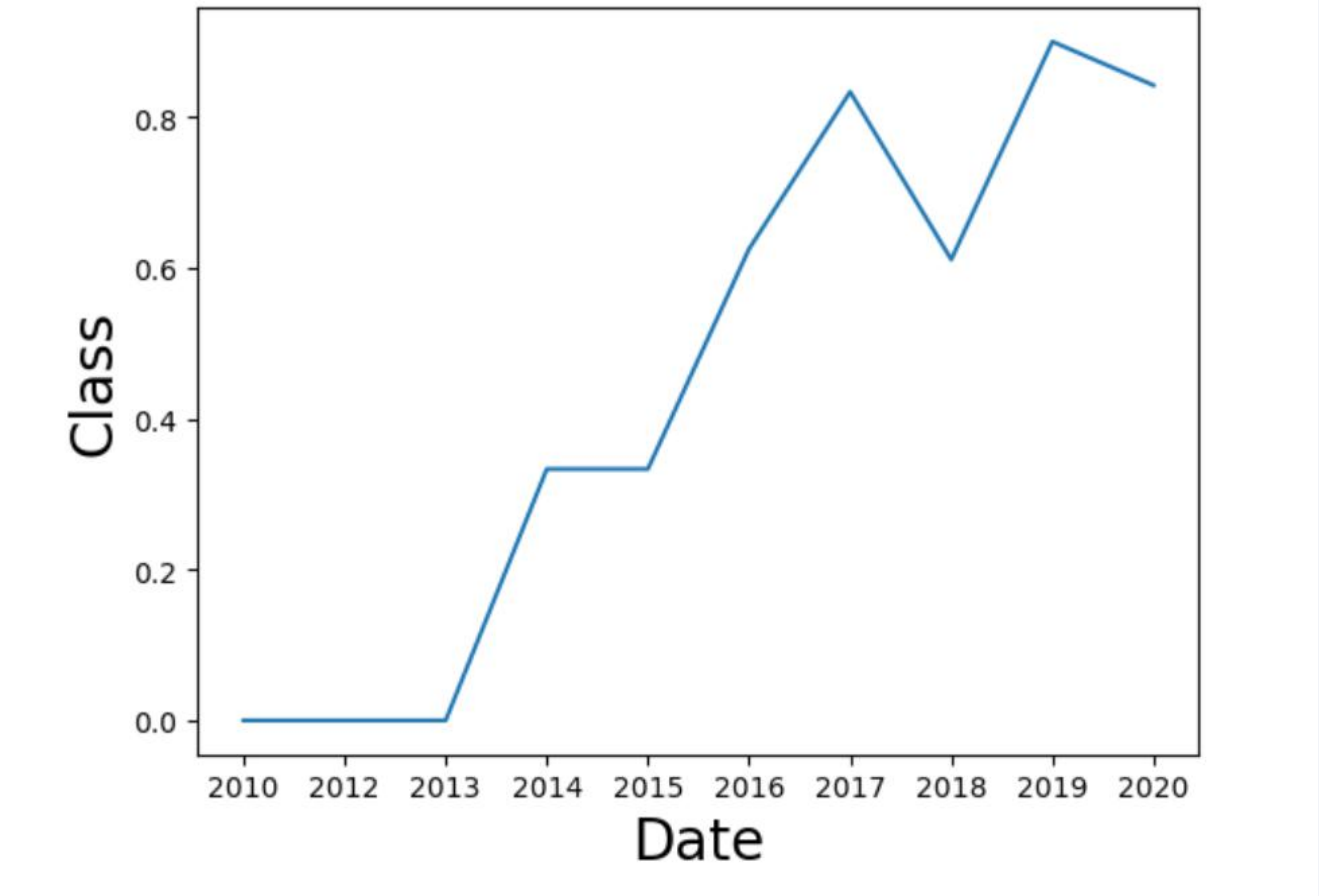
Payload vs. Orbit Type



- Almost all launches have a payload mass under 8,000 km
- Some orbit types such as the ISS and GTO are more concentrated in their payload mass
- The highest payload masses are all for a VLEO orbit

Launch Success Yearly Trend

The average success rate trended upwards for every year between 2013 and 2020 except for a dip in 2018



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

There are four launch sites in the data:

- CCAFS-LC-40
- VAFB SLC-4E
- KSC LC-40
- CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
%sql SELECT "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Launch_Site", 1, 3) =  'CCA' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

This is a sample of 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_")  FROM SPACEXTABLE WHERE "CUSTOMER" = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
SUM("PAYLOAD_MASS__KG_")
                   45596
```

Summing all the data in the PAYLOAD_MASS__KG_ column gives a value of: 45,596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_")  FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
AVG("PAYLOAD_MASS__KG_")
2928.4
```

- The average payload mass was used using the AVG function
- The total payload mass carried by the F9 v1.1 is 2,928.4 kg

First Successful Ground Landing Date

```
%sql SELECT min("DATE") FROM SPACEXTABLE WHERE "Mission_Outcome" = 'Success'
```

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

min("DATE")
2010-06-04

- The date of the first successful landing outcome on ground pad was found using the MIN function.
- The date is June 4, 2010

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)'  AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000
 * sqlite:///my_data1.db
Done.
```

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

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were found using the DISTINCT function and setting the where clause to the appropriate values

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Count("Mission_Outcome") FROM SPACEXTABLE WHERE "Mission_Outcome" = 'Success'
```

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

Count("Mission_Outcome")
98

```
%sql SELECT Count("Mission_Outcome") FROM SPACEXTABLE WHERE "Mission_Outcome" != 'Success'
```

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

Count("Mission_Outcome")
3

The total number of successful and failure mission outcomes were found using the COUNT function using the where clause to only count successful or unsuccessful launches.

- Successful launches = 98
- Unsuccessful launches = 3

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT "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

- The names of the booster which have carried the maximum payload mass are found using the DISTINCT and MAX functions in a where clause
- The results can be found in the image to the left

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) as "Month", "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.
```

Month	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

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 were found using the SELECT a substr on the Date column and setting the appropriate WHERE clause to ensure the date was from 2015 and the outcome was a ‘Failure (drone ship)’

- F9 v1.1 B1012
- F9 v1.1 B1015

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

```
%sql SELECT  "Landing_Outcome"  AS "Outcome", COUNT(*) AS "Count" FROM SPACEXTABLE WHERE "Date" >= '2010-06-04' AND "Date" <= '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY "Count" DESC
 * sqlite:///my_data1.db
Done.
```

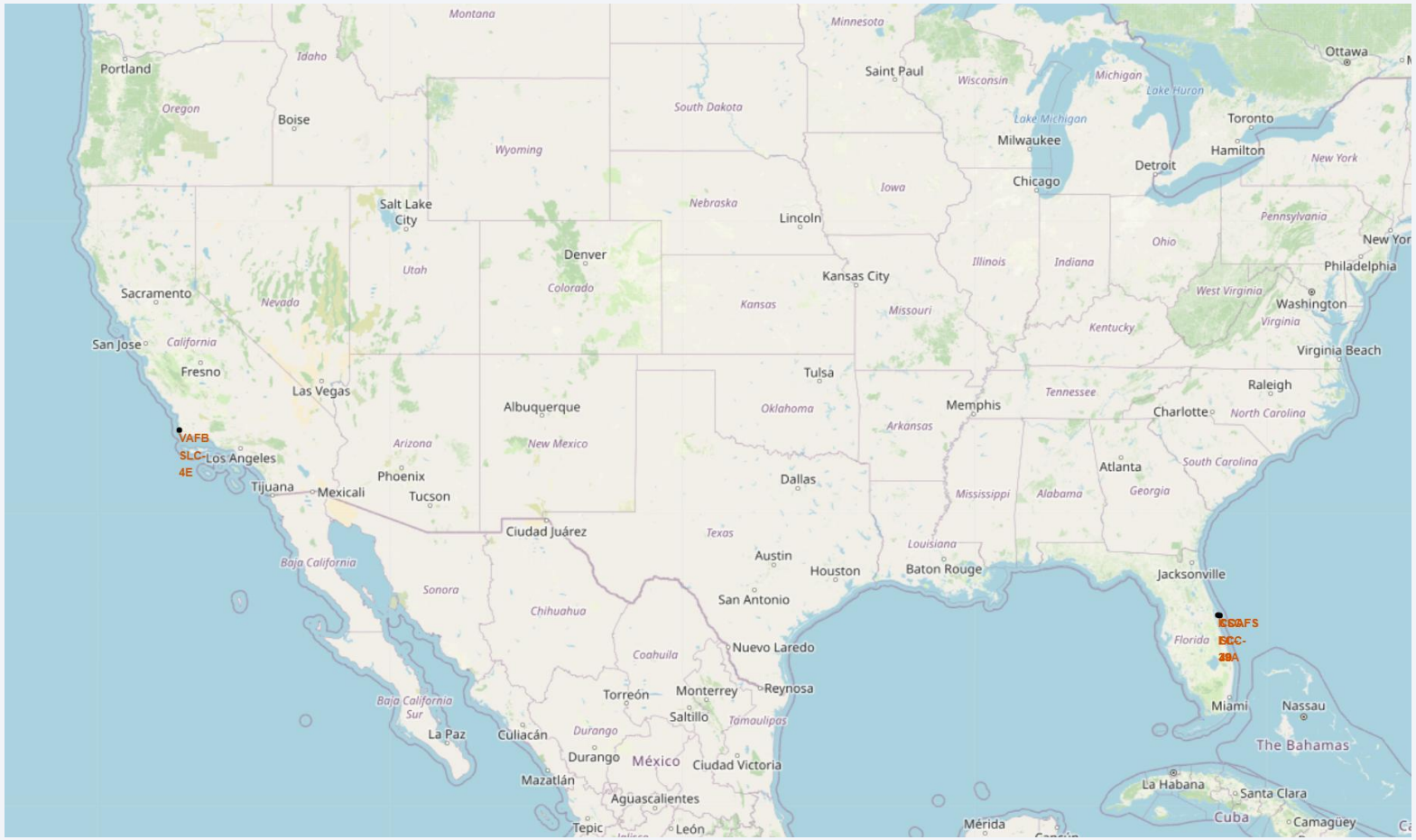
Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 are ranked using a SELECT statement and then GROUP BY and ORDER BY to present in the descending order

Section 3

Launch Sites Proximities Analysis

Folium Map - Launch Sites



- Three launch sites aee in proximity to each other in Florida on the east coast.
- The fourth site is in California in the west coast

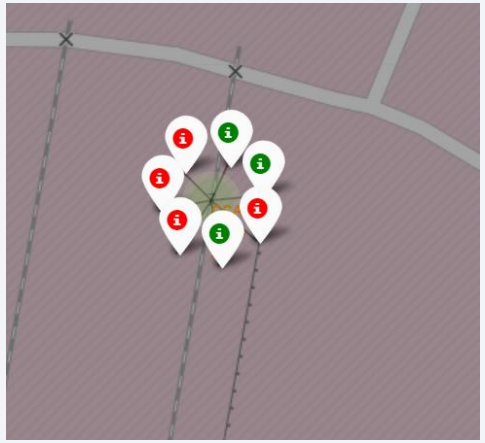
Folium Map – Launch Site Outcomes

West Coast Launch Sites

Vandenberg SLC

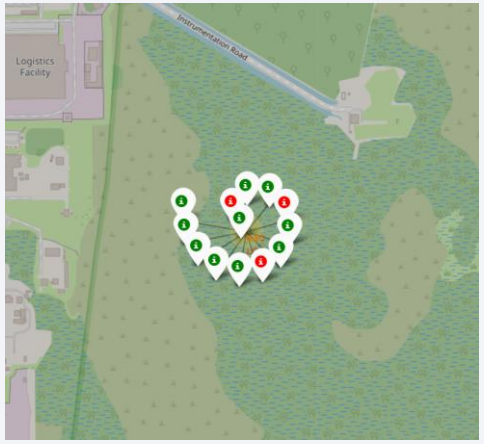
- These maps of the launch sites show the outcomes of each launch at the four sites.
- Red = Failure
- Green = Success
- CCAFA SLC40 (lower right) is the most successful launch site.

East Coast Launch Sites



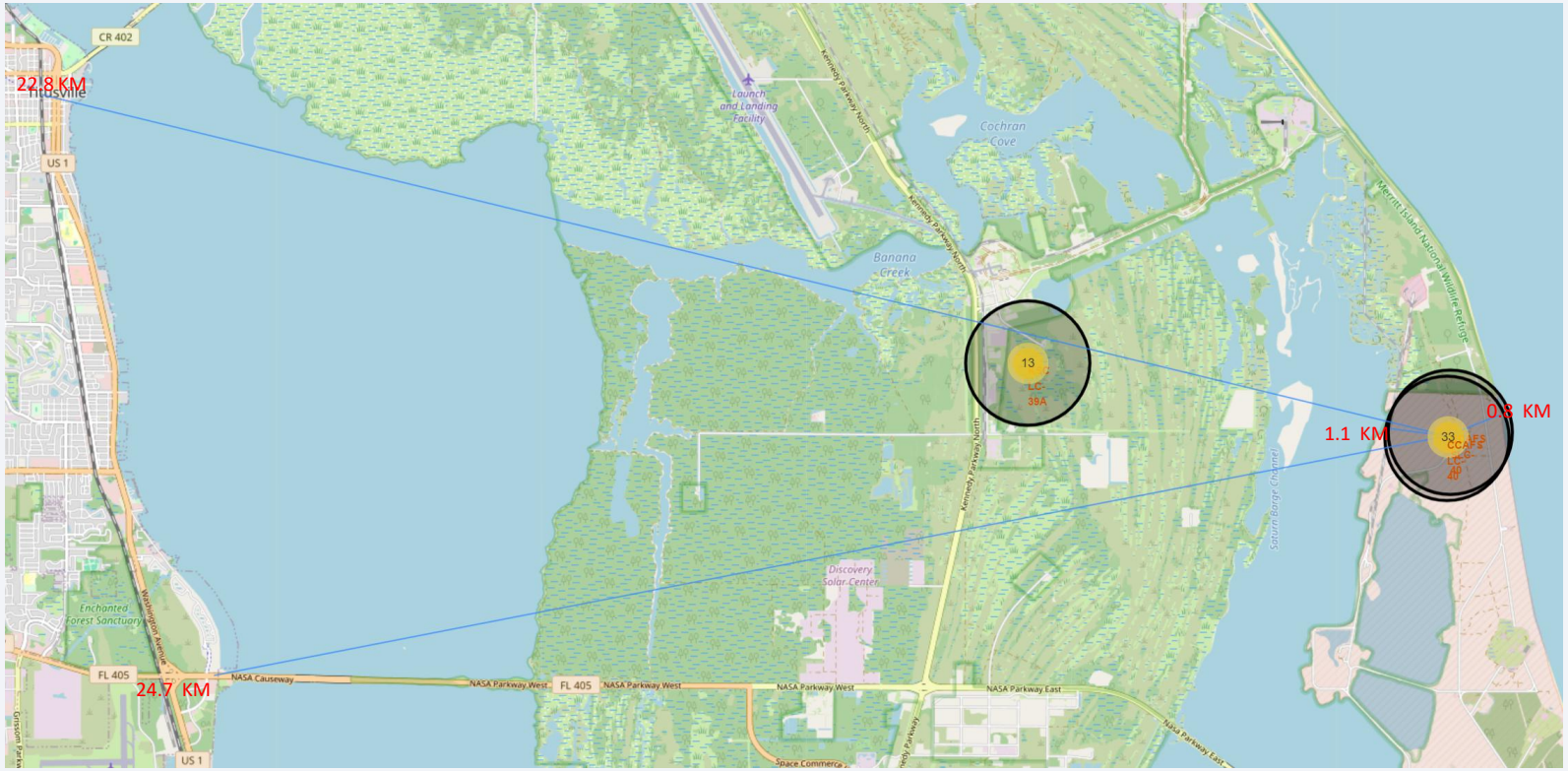
Kennedy Space Center LC-39A

Cape Canaveral AFS LC40

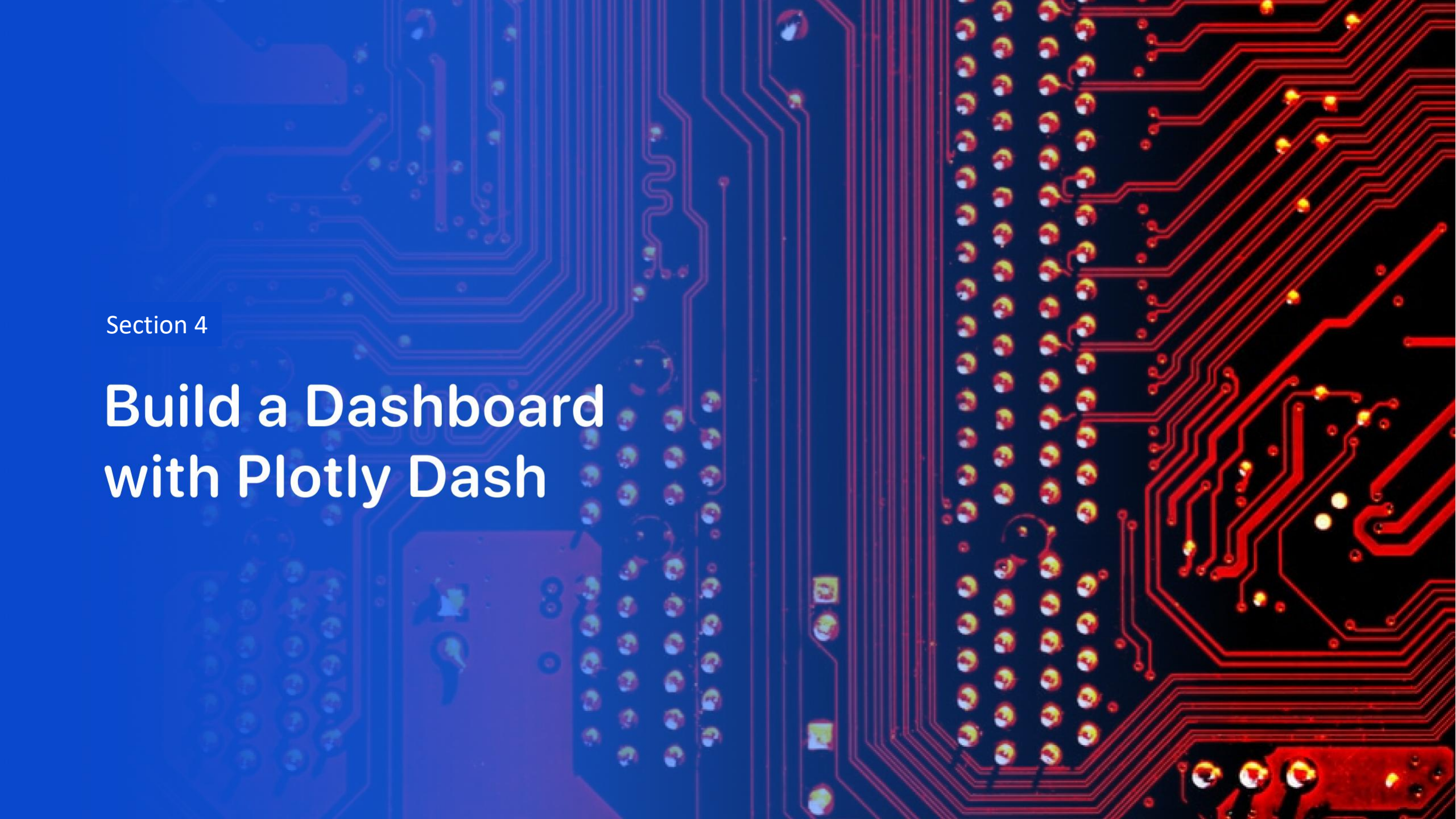


Cape Canaveral AFS SLC40

Folium Map - Launch Site Proximity

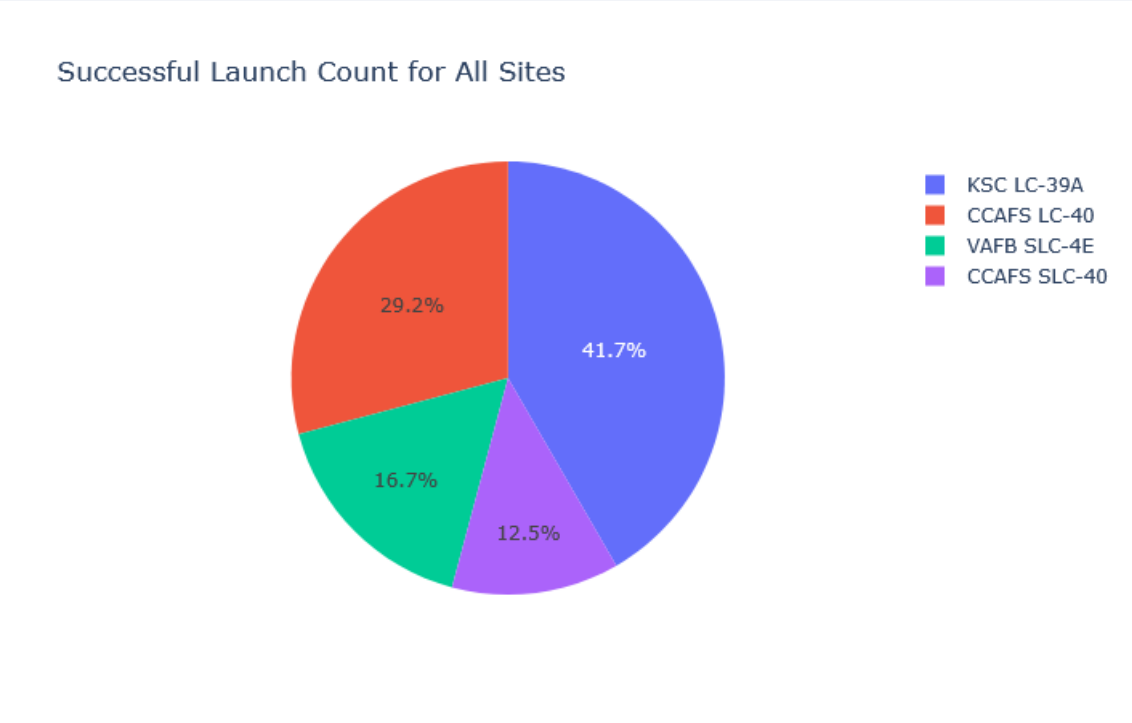


- Locations of the nearest:
  - City 22.8 km
  - Highway 24.7 km
  - Railroad 1.0 km
  - Coastline 0.8 km
- All the critical facilities are at a safe distance from the launch sites


Section 4
Build a Dashboard with Plotly Dash

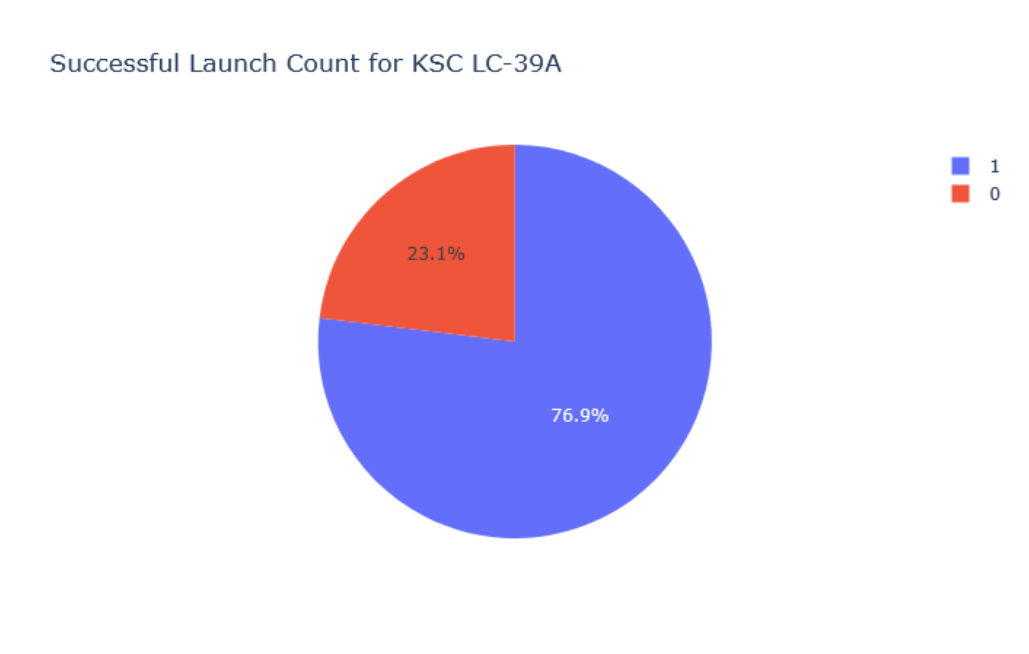
Successful Launches From All Sites

- KSC LC-39A was the most successful launch site
- It accounted for 41% of all successful launches

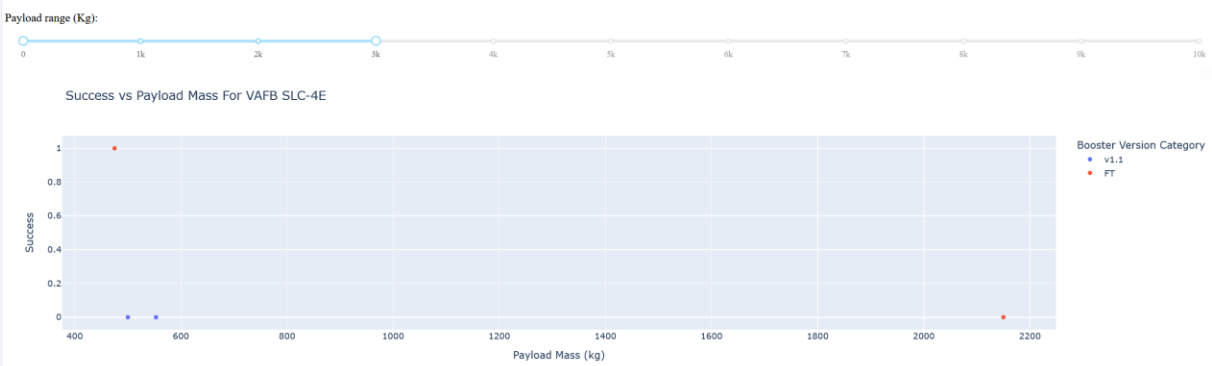


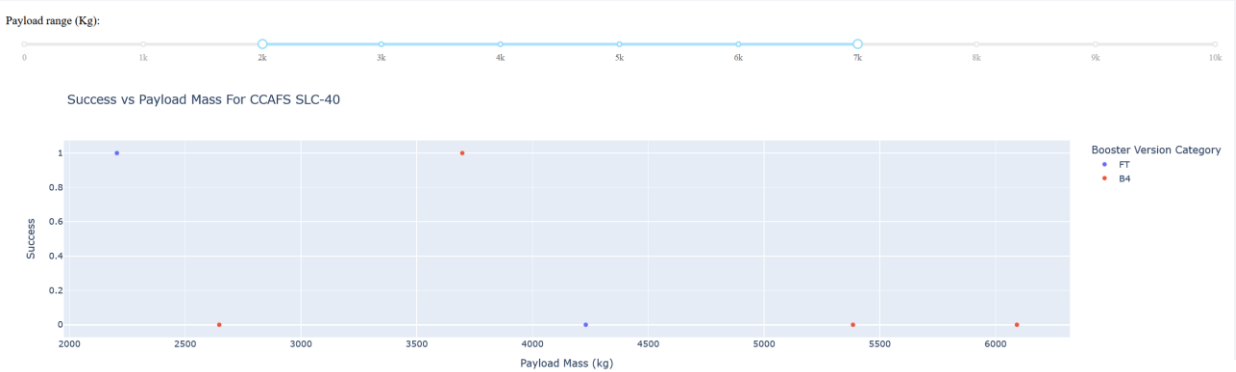
Most Successful Launch Site: KSC LC-39A

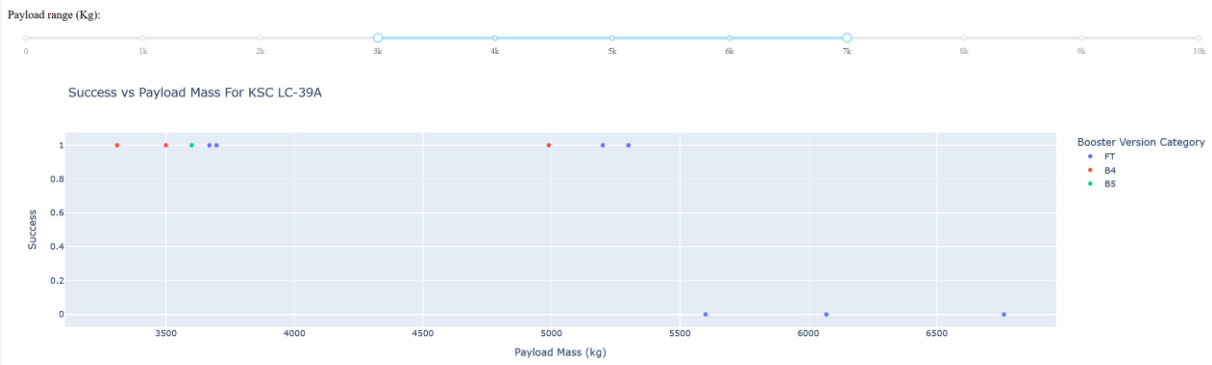
- KSC LC-39A was the most successful launch site.
- It landed 76.9% successfully

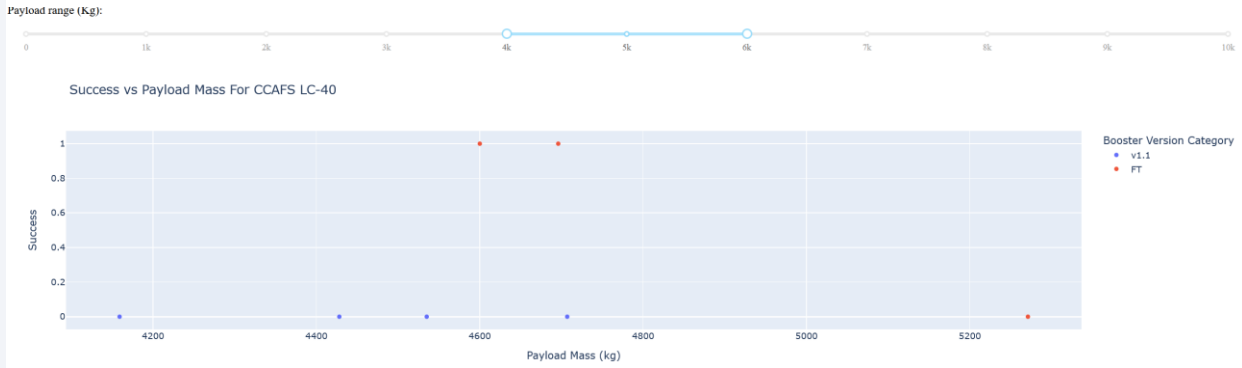


Payload vs Launch Outcome









- There is a clear trend of a higher success rate for payload masses between 2,500 and 5,000 kg at KSC

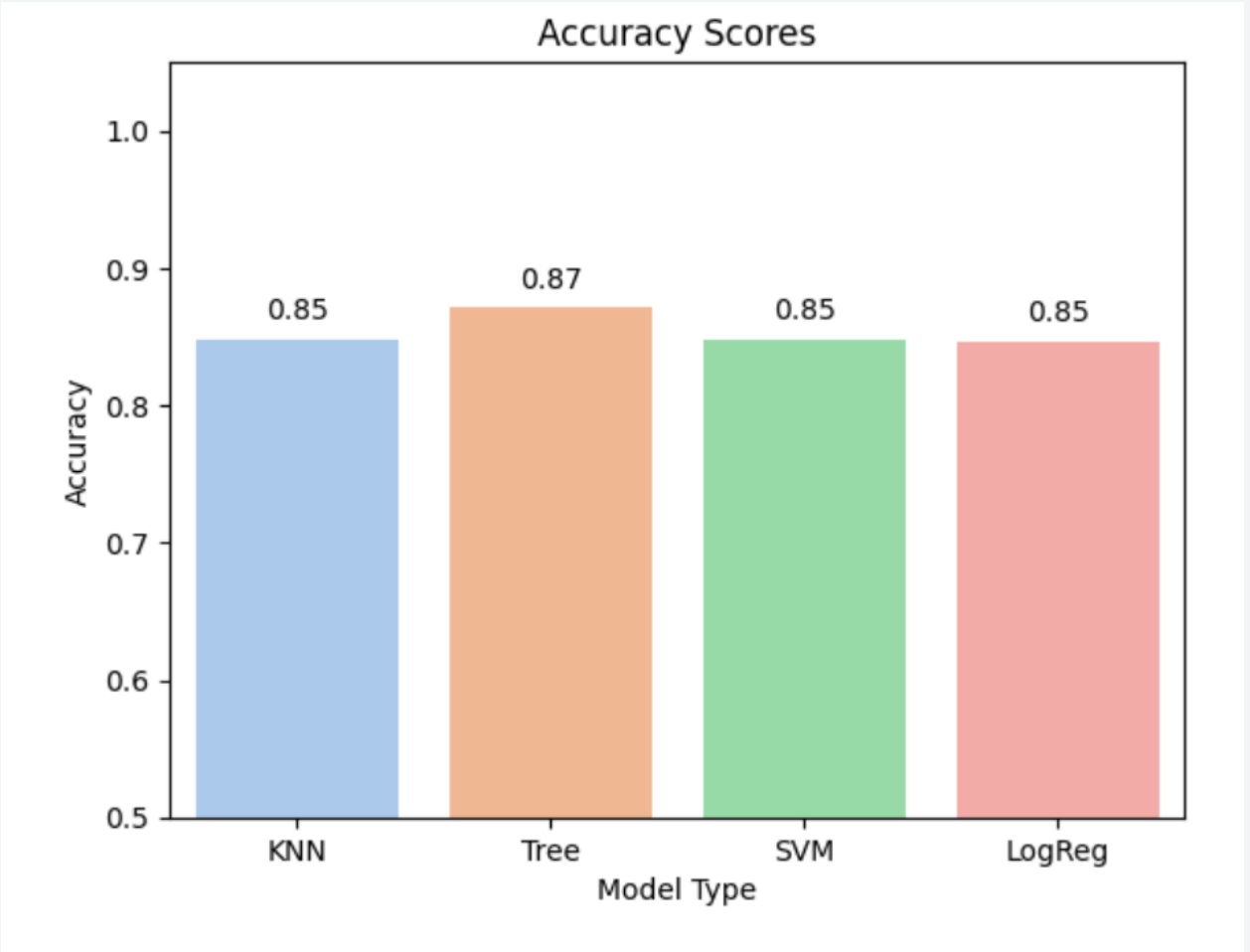
Section 5

Predictive Analysis (Classification)

Classification Accuracy

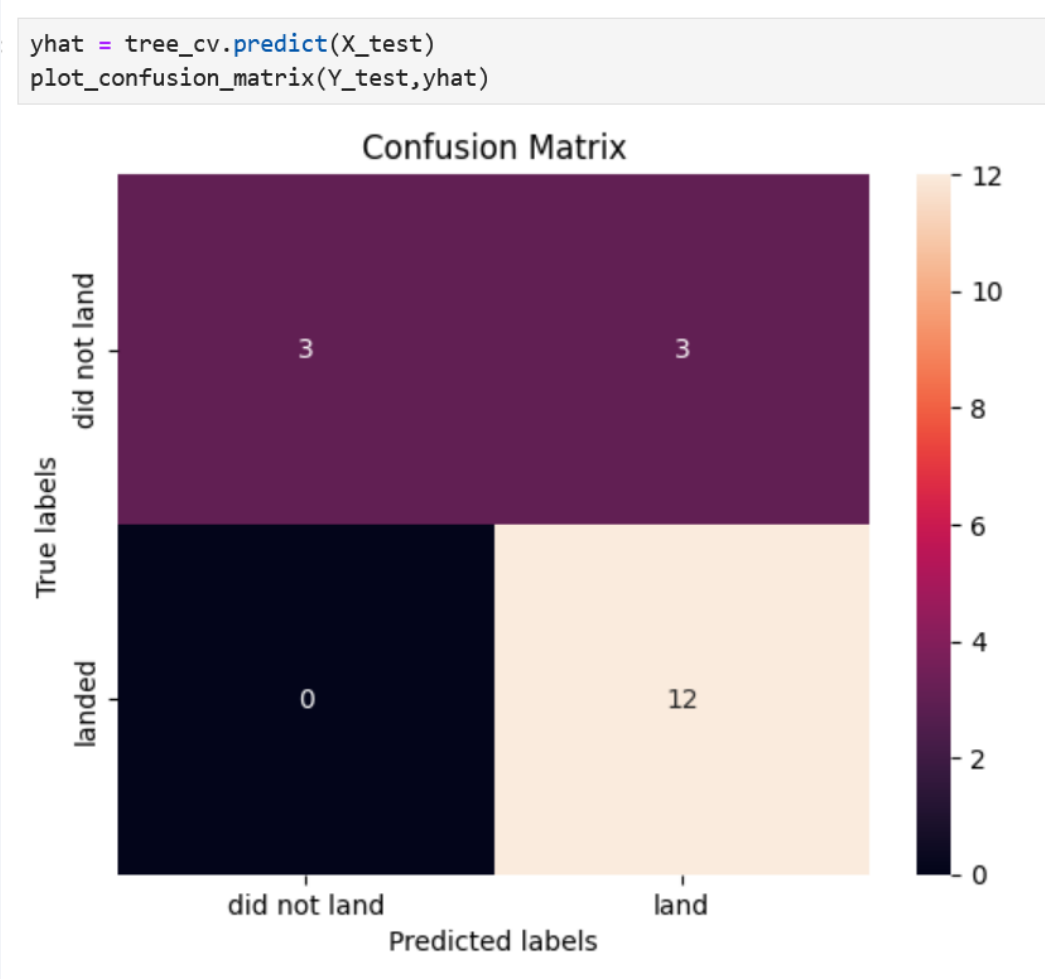
- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy



Confusion Matrix

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



- Show the confusion matrix of the best performing model with an explanation

Conclusions

- The average success rate trended upwards for every year between 2013 and 2020 except for a dip in 2018
- KSC LC 39A accounted for 41% of all successful launches. It landed successfully 76.9% of the time
- There is a clear trend of a higher success rate for payload masses between 2,500 and 5,000 kg at KSC
- Publicly available data can be used to predict the success of a launch using a Decision Tree model with an accuracy of 88.75%

Appendix

- All relevant assets including Python code, SQL queries, charts, Notebook outputs, or data sets that you may created during this project can be found on my GitHub Page

- https://github.com/dsekkes/IBM-Data-Science-Capstone-Project---DAS

Thank you!