IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

John Bergmann
22/01/2023

Outline

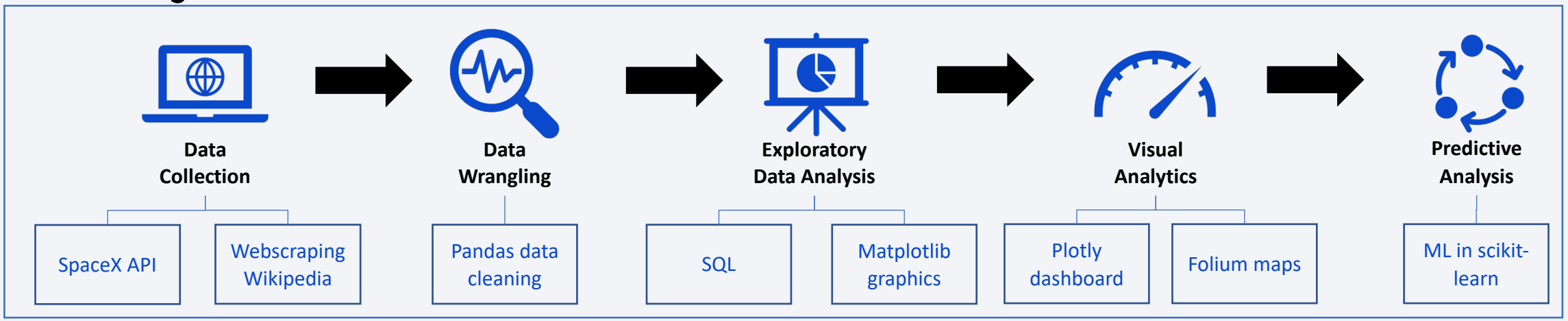
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Problem:

Launching rockets without the re-use of the booster's (SpaceX) creates costs more than twofold those of traditional space companies. SpaceY aims to predict when these "first-stages" are successful and therefore save costs.

Methodologies



Results:

A trained decision tree was able to predict if a rocket's boosters could be re-used with an **accuracy of 87.5%**, saving SpaceY costs of **1.236 billion dollars** compared to traditional launch techniques. Furthermore, the launch site "CCAFS SLC-40" has the highest launch success ratio and should therefore be prioritized. Lastly, the "ES-L1", "GEO", "HEO", and "SSO" orbit's have a booster landing rate of 100%, making these first-stage success favorite orbits which should be prioritised (when possible).

Introduction

Put in the shoes of a Space-Company named “SpaceY”, a new company aimed at being a direct competitor to SpaceX, we want to predict the cost of launching rockets into space. Two options exist:

- The traditional use case where rockets are not re-used.
- SpaceX’s “trademark” approach of re-using boosters (the “first stage” lands).

Re-using rockets cuts costs from 165 (or more) million dollars by more than half; to 65 million dollars.

Therefore, if our machine learning model is able to predict if the first stage will land, large quantities of money can be saved (costs can be avoided).

Section 1

Methodology

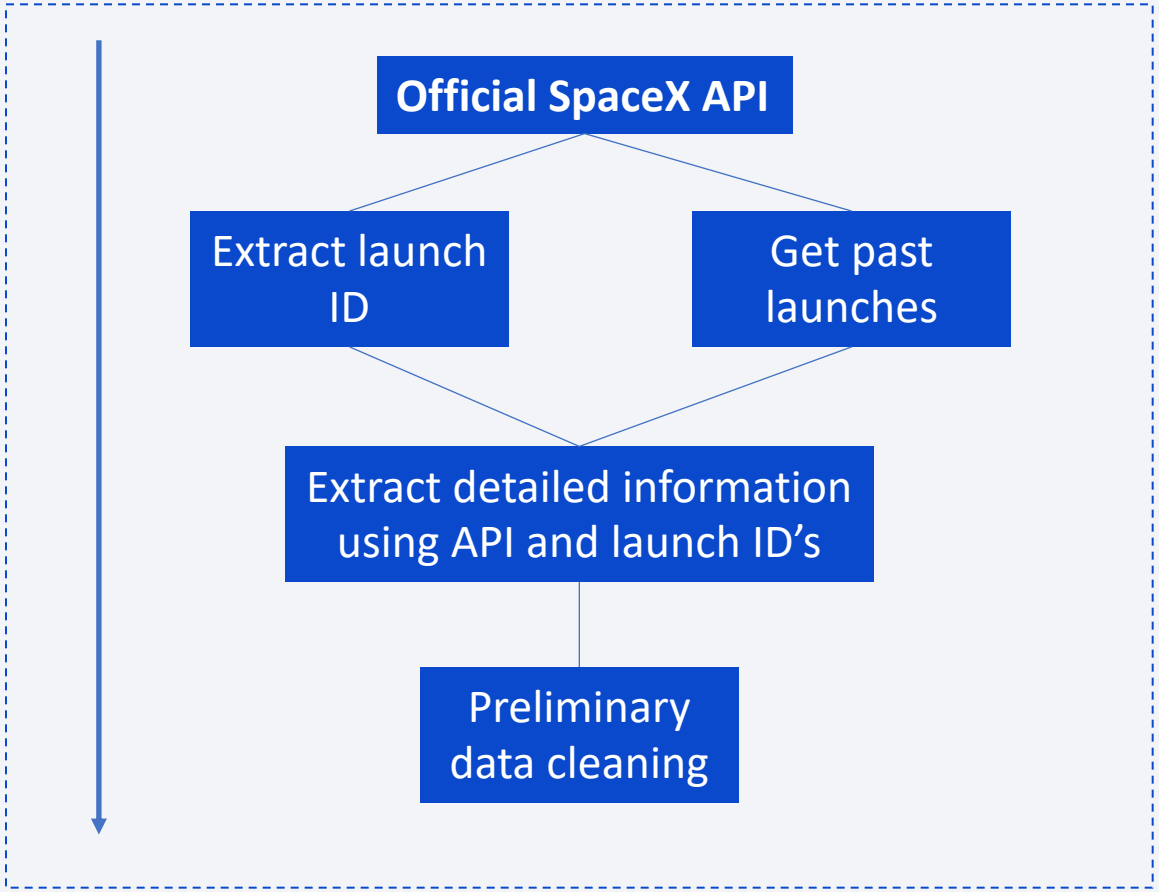
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX's official API and Wikipedia.
- Perform data wrangling
  - Data was processed using python. More specifically using the pandas library.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were created, trained, and evaluated in python using the scikit-learn library.

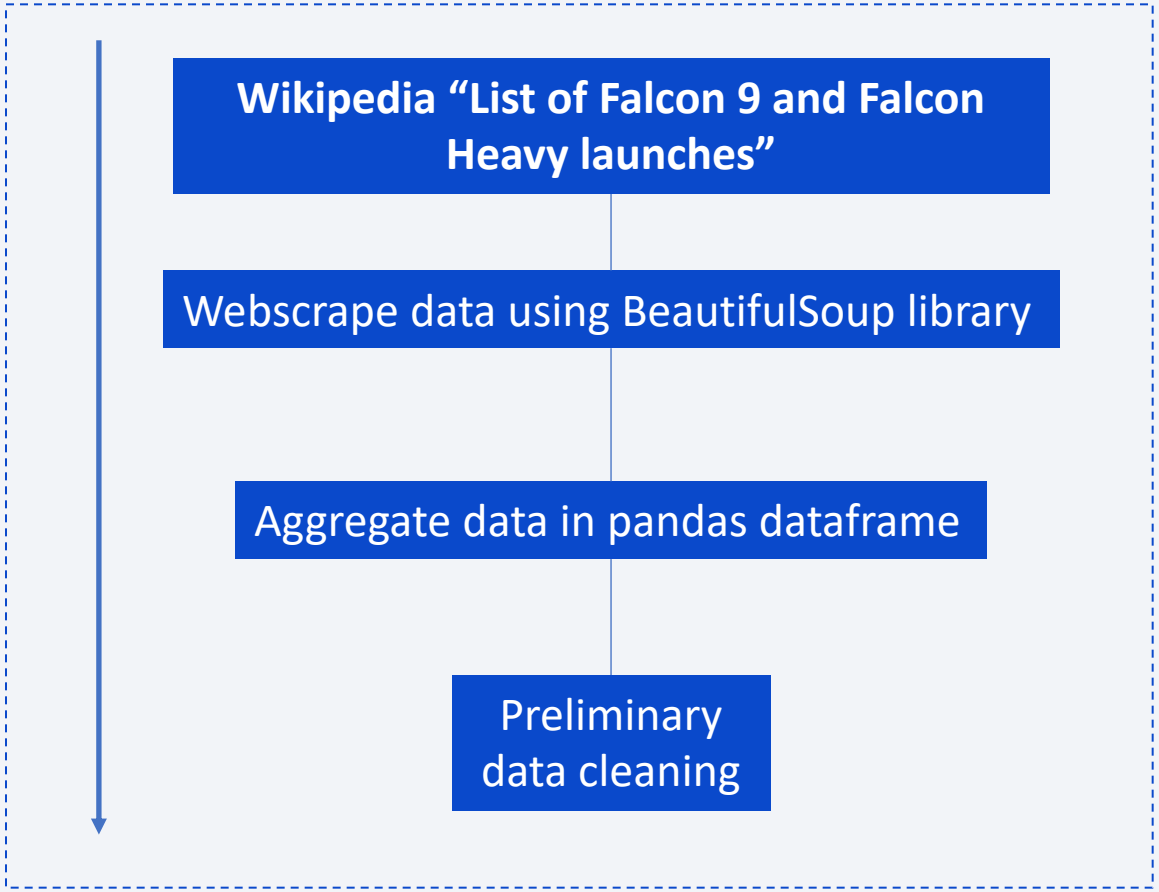
Data Collection – SpaceX API

- The SpaceX API is called in order to obtain starting information and launch ID's. Based on these ID's, more detailed historical launch information is extracted from the same API using a different endpoint.
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/1_Data_Collection/API_Data_Collection.ipynb



Data Collection – Scraping

- Additional information on launches was extracted from Wikipedia using the python web scraping library “BeautifulSoup”.
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/1_Data_Collection/Web%20Scraping.ipynb

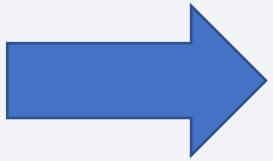


Data Wrangling

- Data Wrangling mainly had the goal of creating an independent variable (the "Launch Outcome") which will later be used to test our prediction model. This variable is ultimately a simplified/aggregated version of the "Outcome" feature.
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/2_Data%20_Wrangling/Data%20Wrangling.ipynb

```
landing_outcomes = df["Outcome"].value_counts()
landing_outcomes
```

```
True ASDS     41
None None     19
True RTLS     14
False ASDS     6
True Ocean     5
False Ocean    2
None ASDS      2
False RTLS     1
Name: Outcome, dtype: int64
```



```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

EDA with Data Visualization

- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/3_Exploratory%20Data%20Analysis/3.2_Pandas_Matplotlib/EDA%20Pandas.ipynb
- Chart overview:

Chart	Reason/Explanation
Flight Number vs. Payload Mass	Identify if more launches also lead to more sucesful outcomes and understand if heavier or lighter rockets are more successful.
Flight Number vs. Launch Site	See if rockets departing from certain launch sites are statistically more successful than others.
Payload Mass vs. Launch Site	See if launch sites are used depending on payload mass.
Orbit vs. Success Rate	Identify if booster landings are more successful from certain orbits compared to others.
Flight Number vs. Orbit	Look for a change in orbit based on flight “experience”.
Payload Mass vs. Orbit	Identify a relationship between Orbit and payload mass.
Success Rate vs. Year	See how booster landing success has changed over time.

EDA with SQL

- SQL queries:
  - Select unique launch sites
  - Show 5 launch sites that start with the letters ‘CCA’
  - Display the aggregated payload mass carried by boosters launched for ‘NASA (CRS)’
  - Display the average payload mass carried by boosters of version ‘F9 v1.1’
  - Show the date when the first successful ground pad landing was done
  - List of successful booster with a payload mass between 4000 and 6000 kg
  - List of the total number of successful and failed missions
  - List of booster versions that carried the maximum payload mass
  - List of failed drone ship landings in 2015
  - Descending list of landing outcomes between 2010-06-04 and 2017-03-20
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/3_Exploratory%20Data%20Analysis/3.1_SQL/DB2/EDA%20SQL.ipynb

Build an Interactive Map with Folium

- Map features and functionalities:
  - Markers of launch sites
  - Markers showing the launch outcomes of each site (and clustering them together for better visibility, showing the aggregated number of launches for each site)
  - Added distance line to nearest shore and railway as reference points (coordinates can be derived from mouse cursor coordinate functionality)
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/4_Visual_Analytics_and_Dashboard/Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard features and functionalities:
  - Dashboard contains 2 filters: One for launch site and another for payload mass (a slider)
  - By default, a pie chart displaying the sum of successful launches for each site is shown
    - Filtering by launch site shows the number of successful and failed launches for that site
  - By default, a scatter plot showing the payload mass vs. success of launch
    - Filtering by launch site shows these metrics only for that specific site
    - Adjusting the payload mass slider only affects the scatter plot, increasing or decreasing the number of instances shown
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/4_Visual_Analytics_and_Dashboard/spacex_dash_app.py

Predictive Analysis (Classification)

- Building Models:
  - Predictive Analysis was done in Python using the scikit-learn library.
  - The models tested were Logistic Regression, Support Vector Classification, Decision Tree, and K-Nearest-Neighbors.
- Evaluation:
  - Models were evaluated using the standard train/test split of 80/20.
  - Furthermore, cross-validation has been applied using GridSearch
- Improvement:
  - Using GridSearch, optimal hyperparameters were chosen for each model, optimizing its performance
- Best Model:
  - The best model turned out to be a decision tree
  - The deciding metric used in this case was model accuracy (the total amount of correctly classified instances)
- Github: https://github.com/JPBergmann/IBM_DS_Capstone/blob/master/5_Predictive_Analysis_(Classification)/Predictive%20Analysis.ipynb

Results

Exploratory Data Analysis

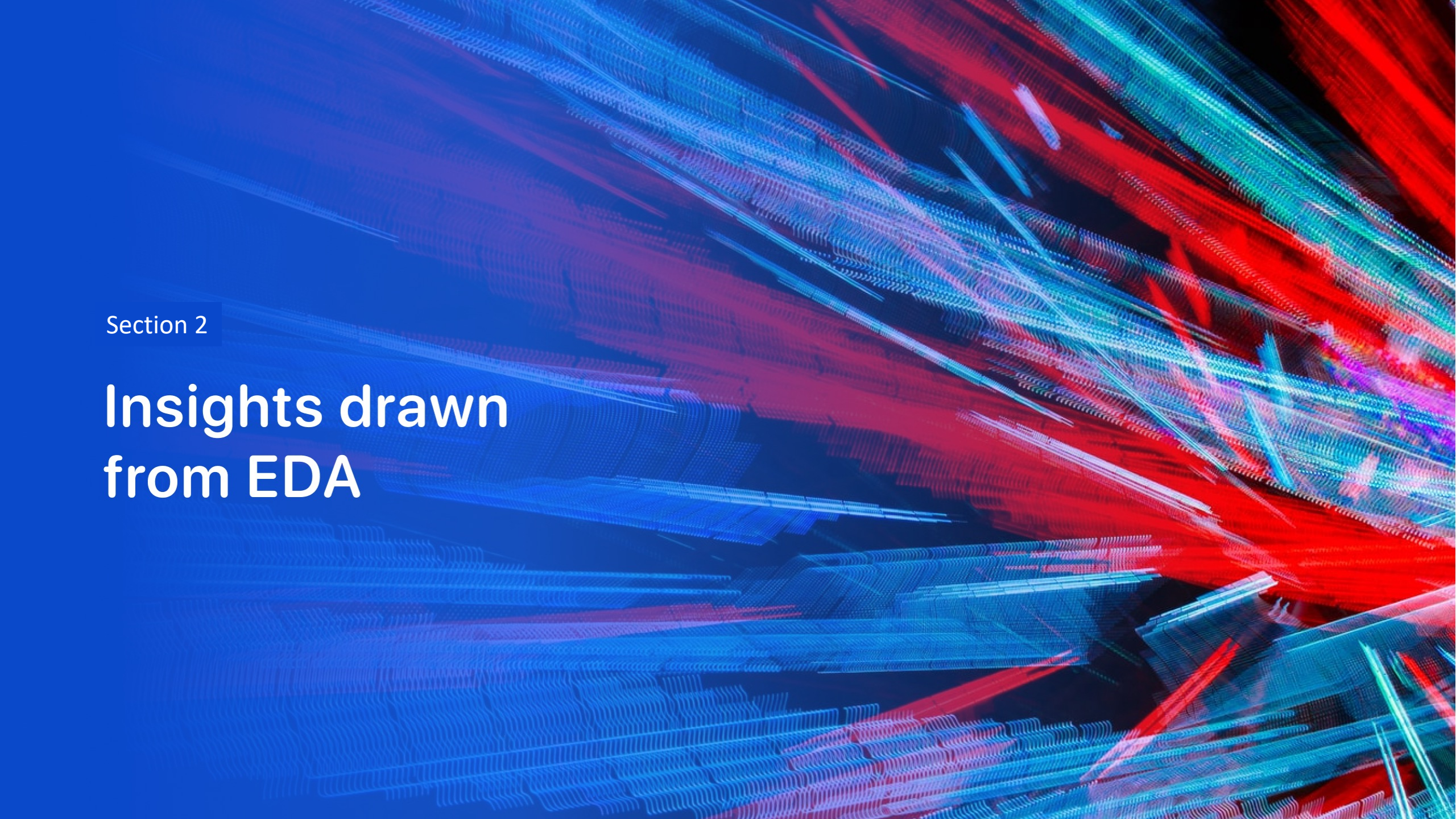
- VAFB-SLC launch site has no rocket launches with a payload mass above 10,000 kg.
- Relationship between number of flights and success in 'LEO' Orbit.
- LEO, ISS, and Polar orbit associated with positive landing, especially with heavier payloads.
- Over time, success rate has increased until 2020, after which a decrease can be observed.

Interactice Analytics

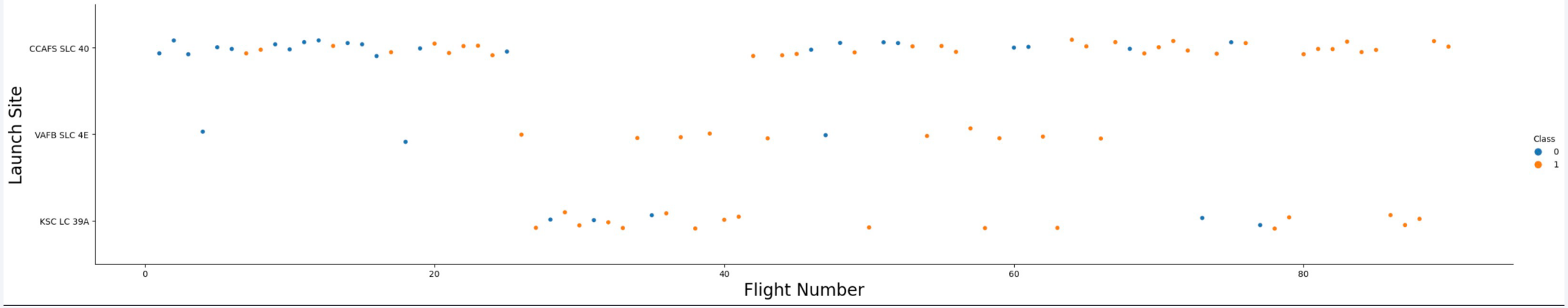
- KSC LC-39A has the most successful launches out of all inspected launch sites.
- CCAFS SLC-40 has the highest ratio of successful to non-successful launches.
- Rockets using boosters of type "FT" and "B4" are strongly correlated with successful launches, among a broad variety of payloads.

Predictive Analysis

- The best model to predict if a launch was successful or not is a decision tree.
- The evaluation metric used was model accuracy, achieving a score of 87.5%.
- The most important aspect to (potentially) keep improving are False Positives (predicting a rocket will land, but it doesn't)

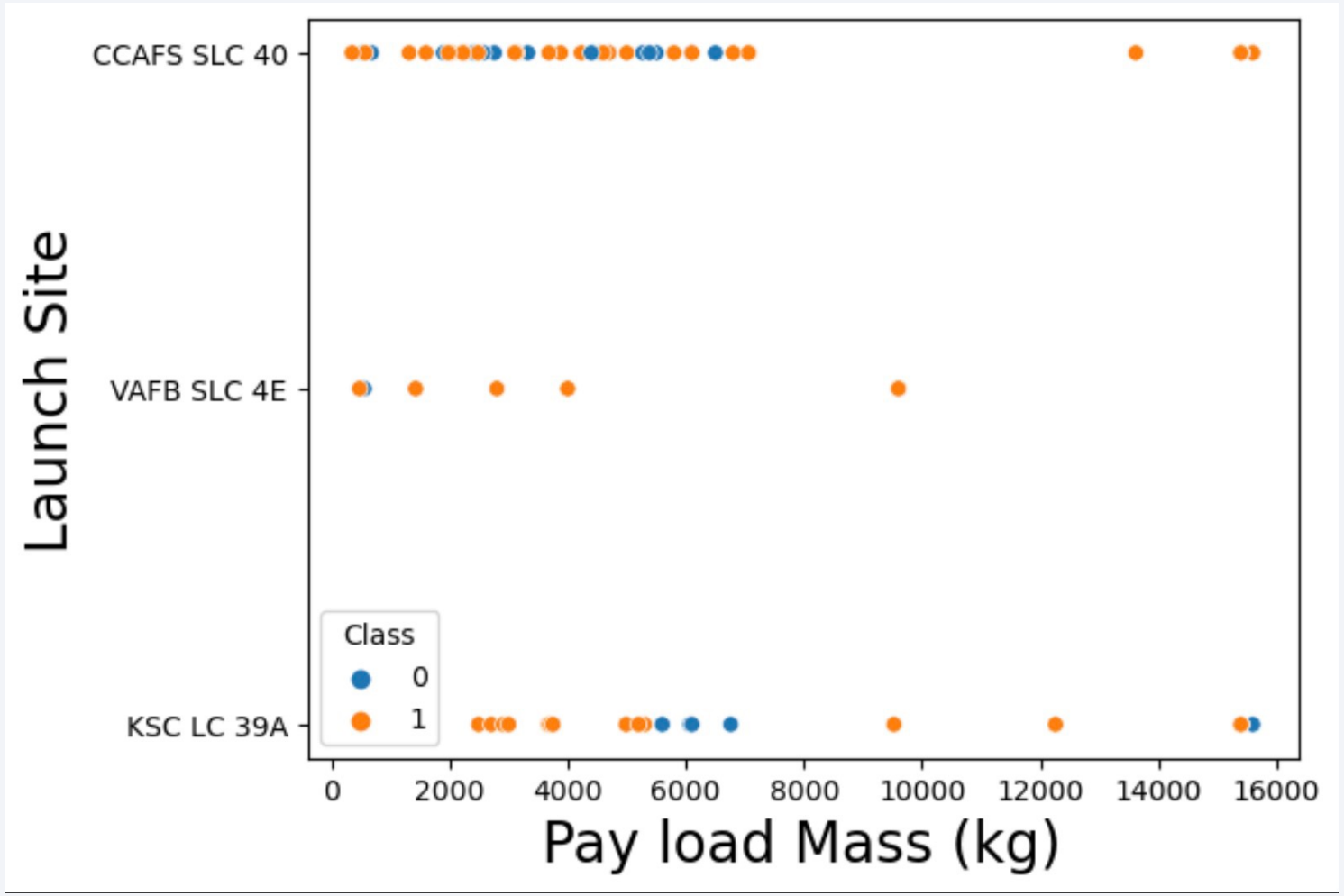

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



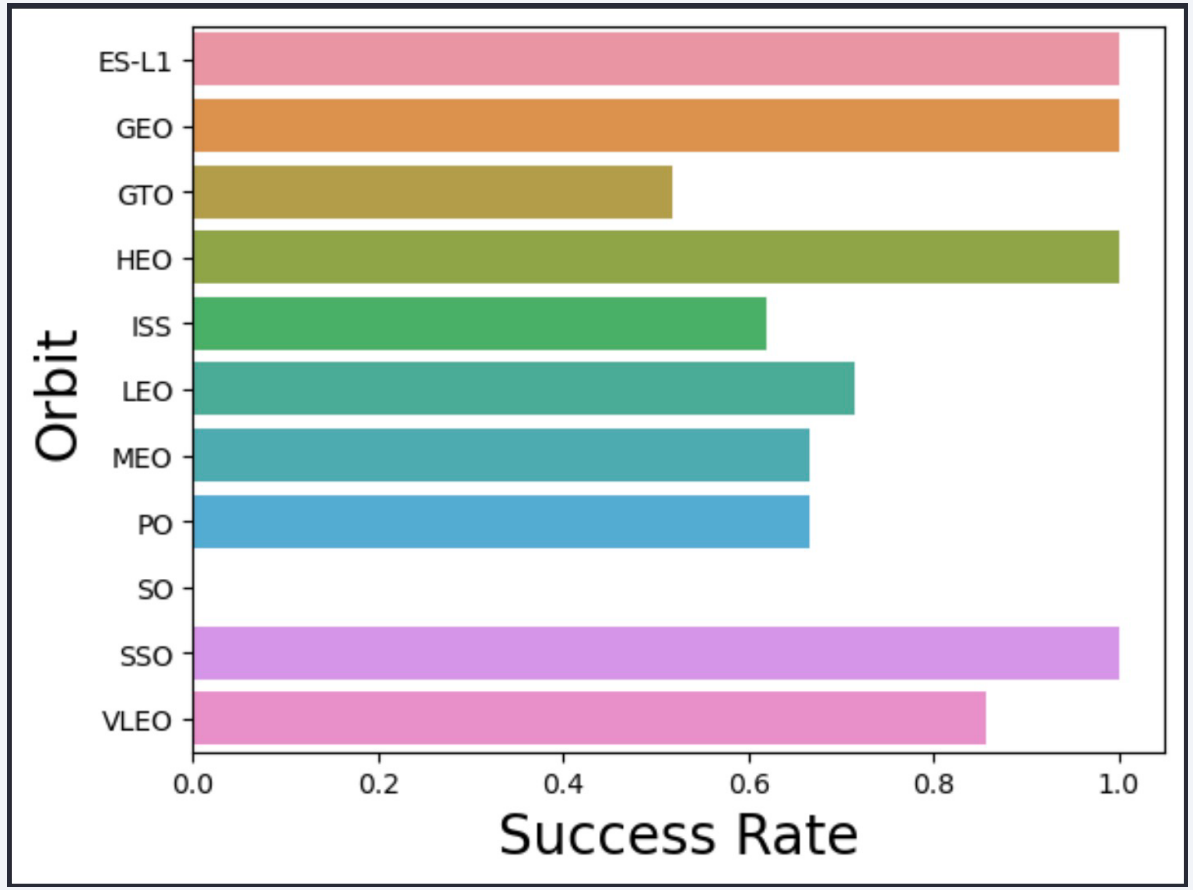
- Launch site VAFB SLC 4E has not been used after around the 70th launch.
- Furthermore, launch site KSC LC 39A has been used more frequently between flight 30 and 40.
- Ultimately, launch site CCAFS SLC 40 is being used most often in the present day (at least until our data allows, approx. 2020/2021) with no unsuccessful launches after flight 80.

Payload vs. Launch Site



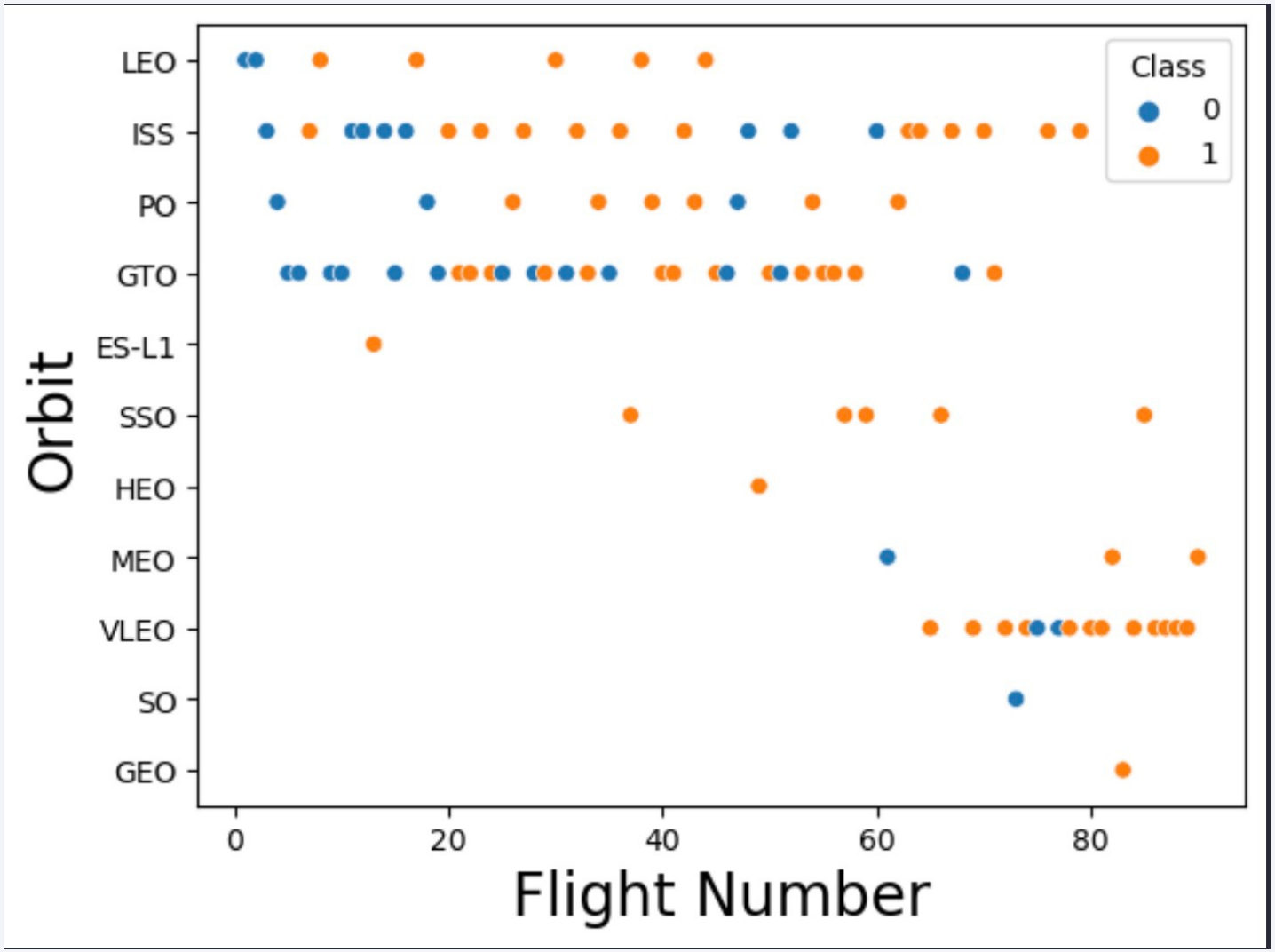
- Launch site VAFB SLC 4E has not been used for payloads over 10,000 kg.
- While launch sites CCAFS SLC 40 and KSC LC 39A have both carried heavier payloads (> 10,000 kg), only the former has done so without any failed booster landings.

Success Rate vs. Orbit Type



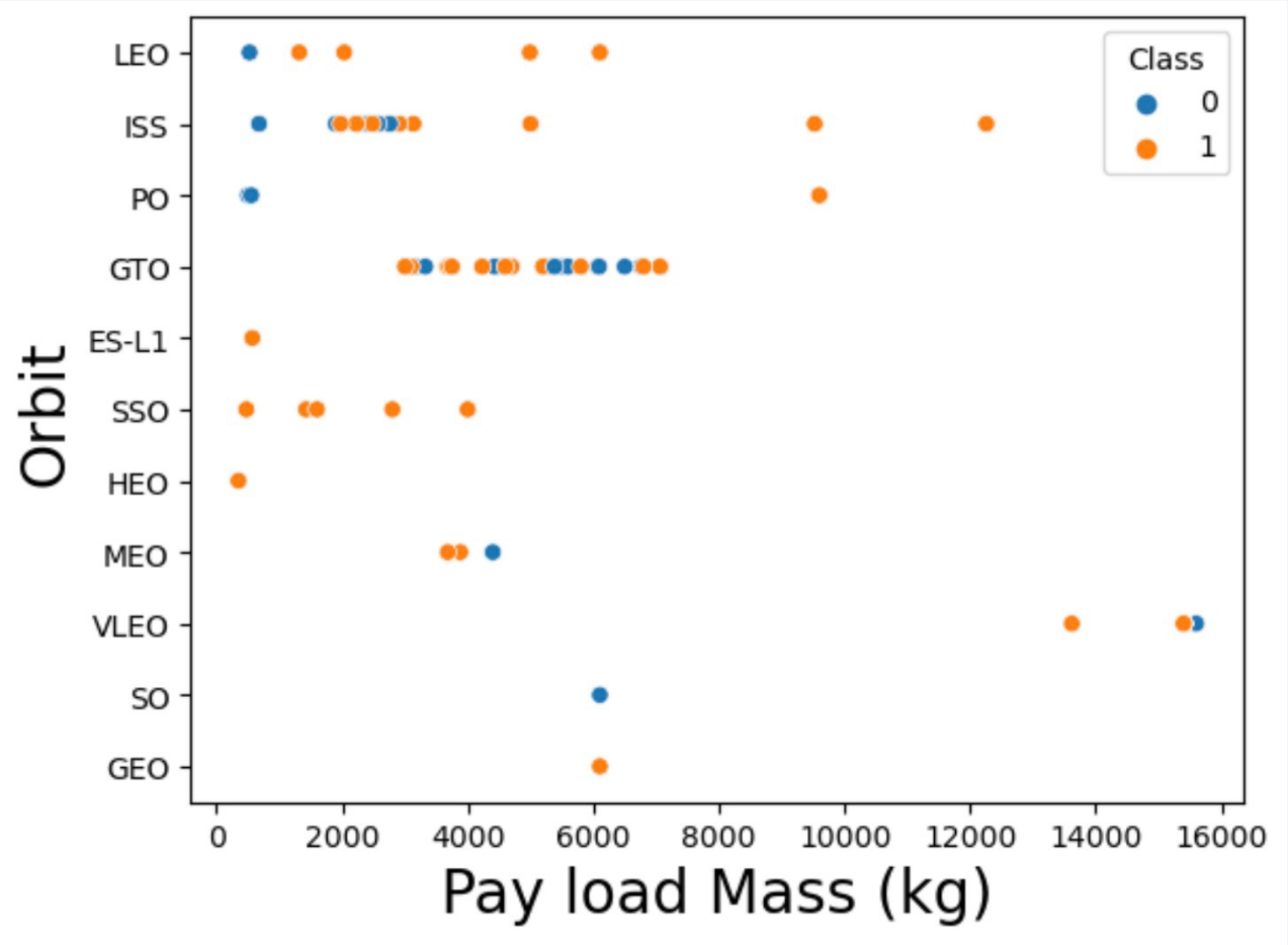
- The Orbits with no unsuccessful launch outcomes are ES-L1, GEO, HEO, and SSO.
- These are closely followed by the VLEO Orbit, having a success rate of approximately 85%.

Flight Number vs. Orbit Type



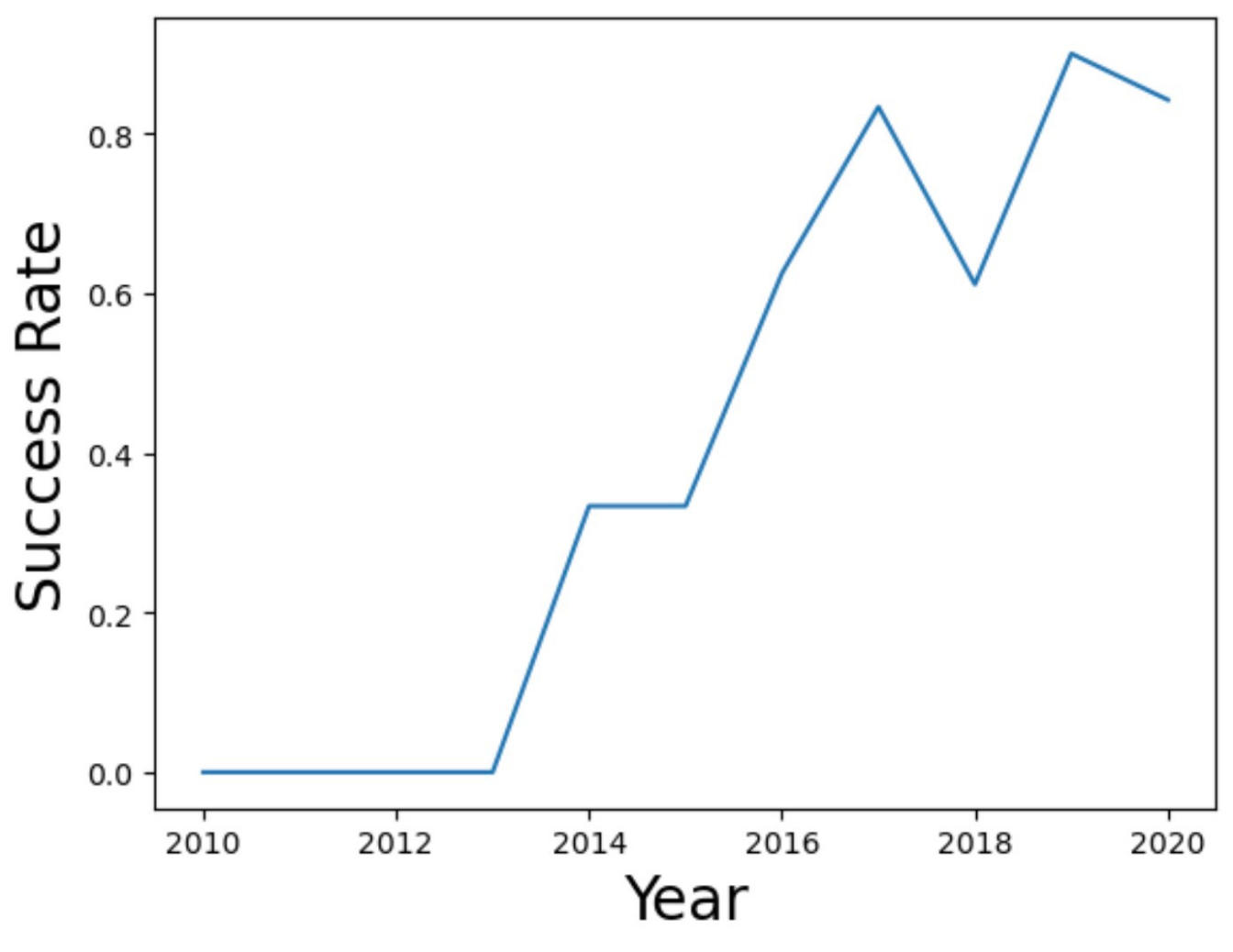
- Across all Orbit types, the success rate of booster landings has increased with increasing number of launches (flight number).

Payload vs. Orbit Type



- Rockets launched into the SO and GEO Orbit exclusively carry payload with a mass of 6,000 kg.
- Any rocket with a payload over approximately 13,500 kg is sent to the VLEO Orbit.
- Rockets launched to the ISS Orbit carry a broad variety of payloads successfully, ranging from 2,000 to 12,200 kg.

Launch Success Yearly Trend



- The success rate of booster landings has steadily increased until 2017.
- After a small decline in 2017 and subsequent recovery in 2018, success rate has begun to slowly decline from 2019 to 2020.

All Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEX;
```

```
* ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4
Done.
```

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

- The available launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.
- These were extracted from the database using the DISTINCT function.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Python

```
* ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Using the LIKE function, queries on strings can be done in a specific manner, such as only finding words that begin with ‘CCA’.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databas
Done.
```

1
45596

- Using the SUM clause in combination with the WHERE function allows to extract the total payload mass for NASA (CRS) launch site only.
- This equates to a total payload mass of 45,596 kg.

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.app
Done.
```

1
2928

- Using the AVG clause in combination with the WHERE function allows to extract the average payload mass for rockets that used the ‘F9 v1. 1’ booster.
- This equates to an average payload mass of 2,928 kg.

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdo
Done.
```

1
2015-12-22

- The earliest date out of a timeline can be found using the MIN operator. Combined with a WHERE clause, this allows for the extraction of the first successful ground pad landing from the dataset.
- This was the 22 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_, LANDING__OUTCOME FROM SPACEX \
WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_  > 4000 AND PAYLOAD_MASS__KG_ < 6000);
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Using the AND clause, two conditions inside of a WHERE statement can be tested together. In this case, to find successful drone ship landings with a payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS OUTCOME_COUNT FROM SPACEX GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.
```

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

- GROUP BY allows to perform operations such as COUNT on subsets of data.
- This allows for the identification of successful and failed mission outcomes.

Boosters Carried Maximum Payload

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

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

- Subqueries allow for more complex extractions and comparisons between aggregation functions such as MAX. In this case showcasing all booster versions which carried the maximum payload mass.

2015 Launch Records

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

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE FROM SPACEX \
WHERE (YEAR(DATE) = 2015) AND (LANDING__OUTCOME = 'Failure (drone ship)');
```

```
 * ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- Using the YEAR function in combination with an AND statement, only records in 2015 are extracted where the landing outcome was “Failure (drone ship)”.
- 2 records have been identified.

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

```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS AMOUNT_LANDINGS FROM SPACEX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC;
```

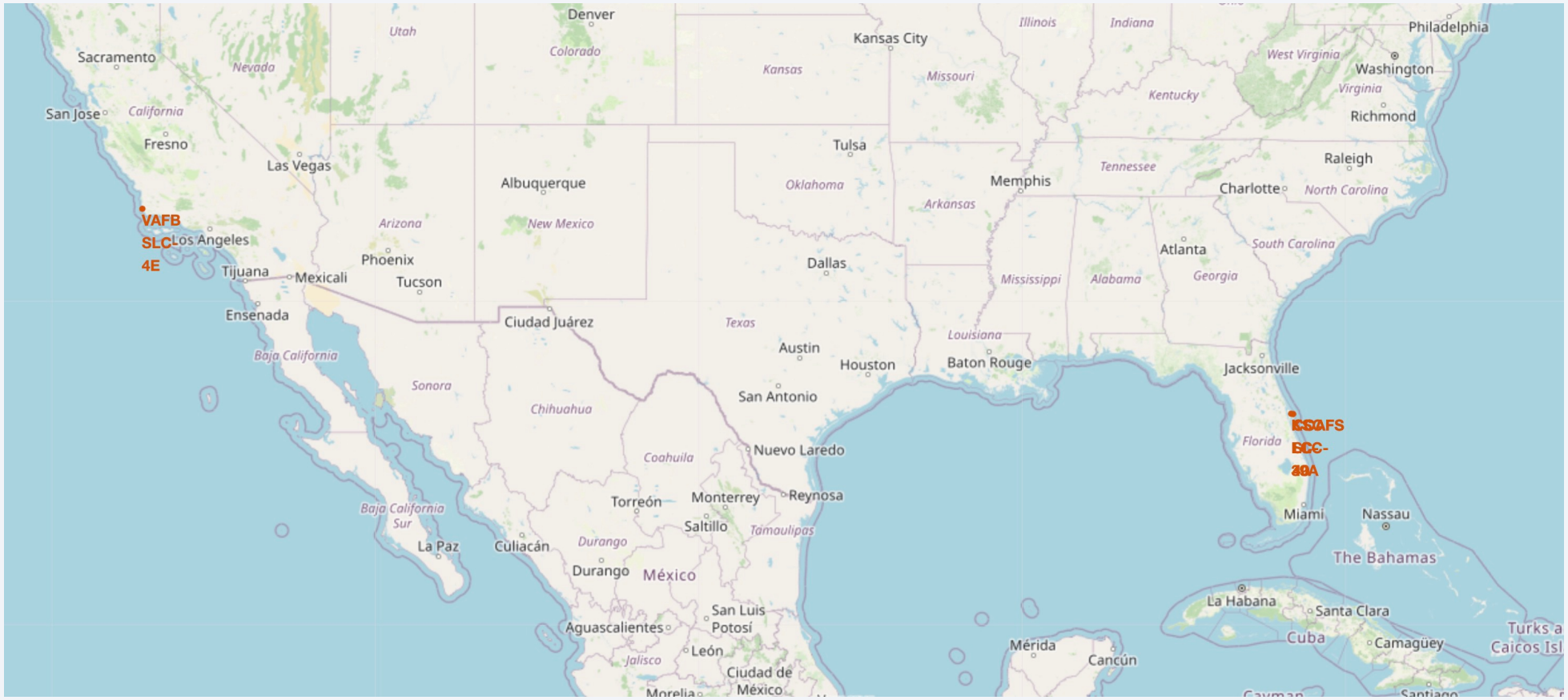
```
* ibm_db_sa://gjd47046:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

- Using GROUP BY and ORDER BY, operations on landing outcomes can be performed and then sorted in a desired way.
- Here, in the specified date range, landings are counted, grouped by their outcome.

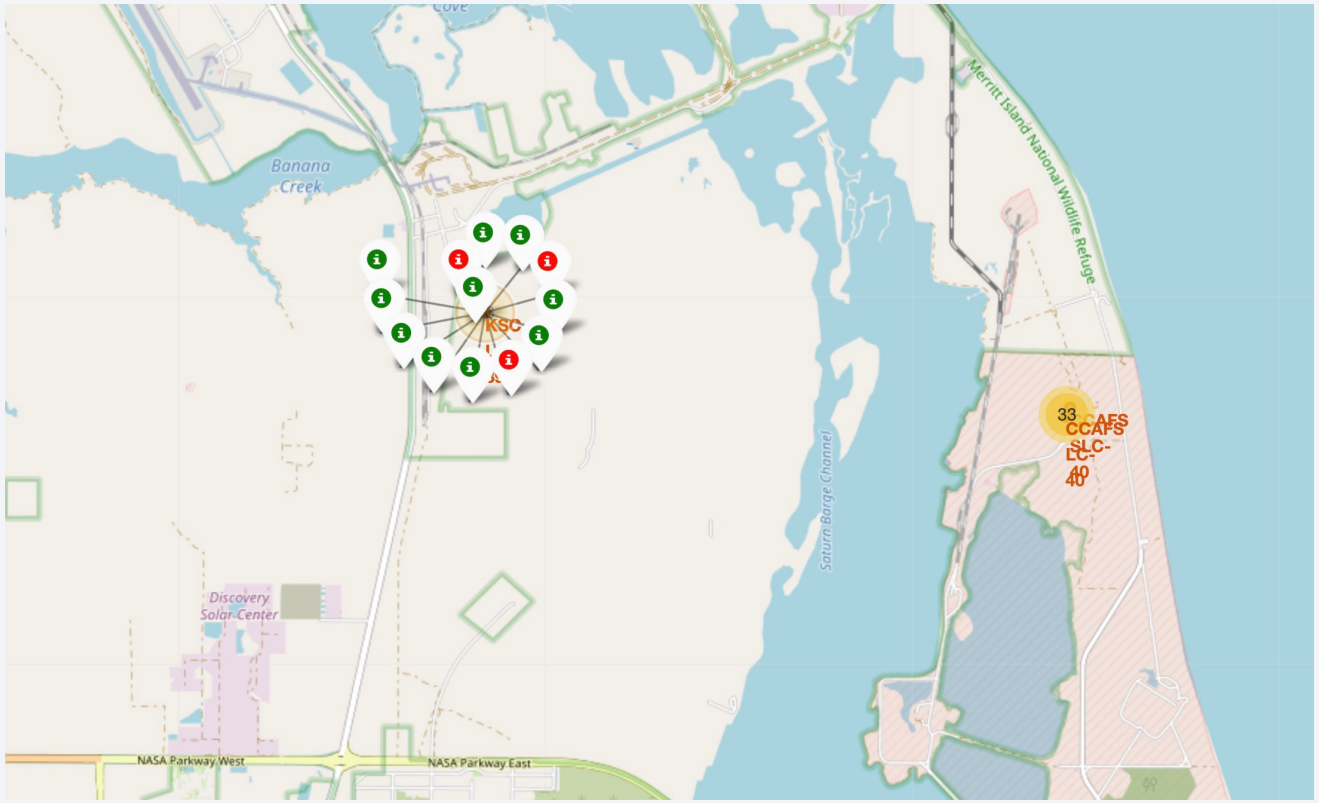

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations



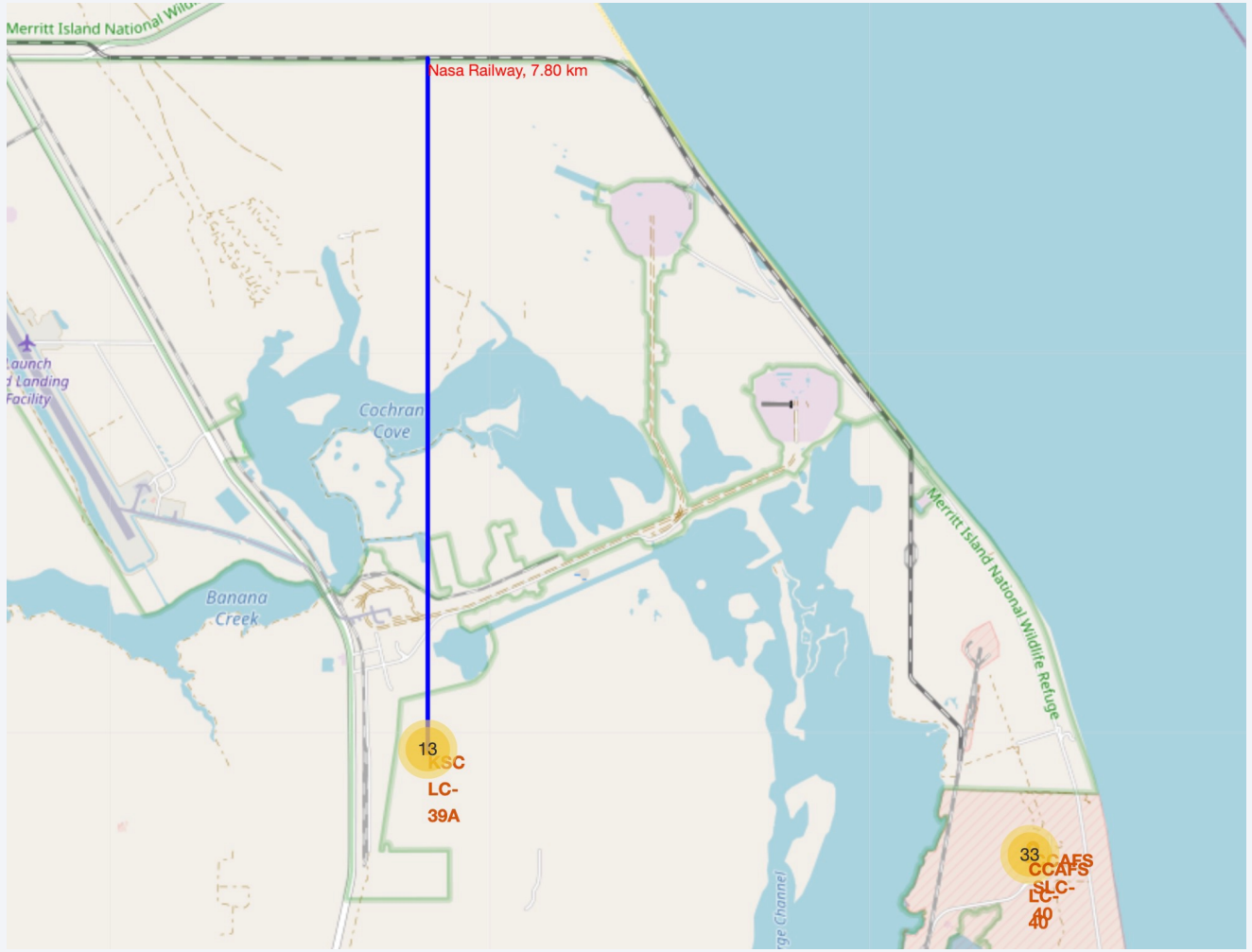
- All launch sites are near a water source.
- Furthermore, most launch sites are located on the east-coast.

Launch Site Specific Map Display of Landing Outcomes

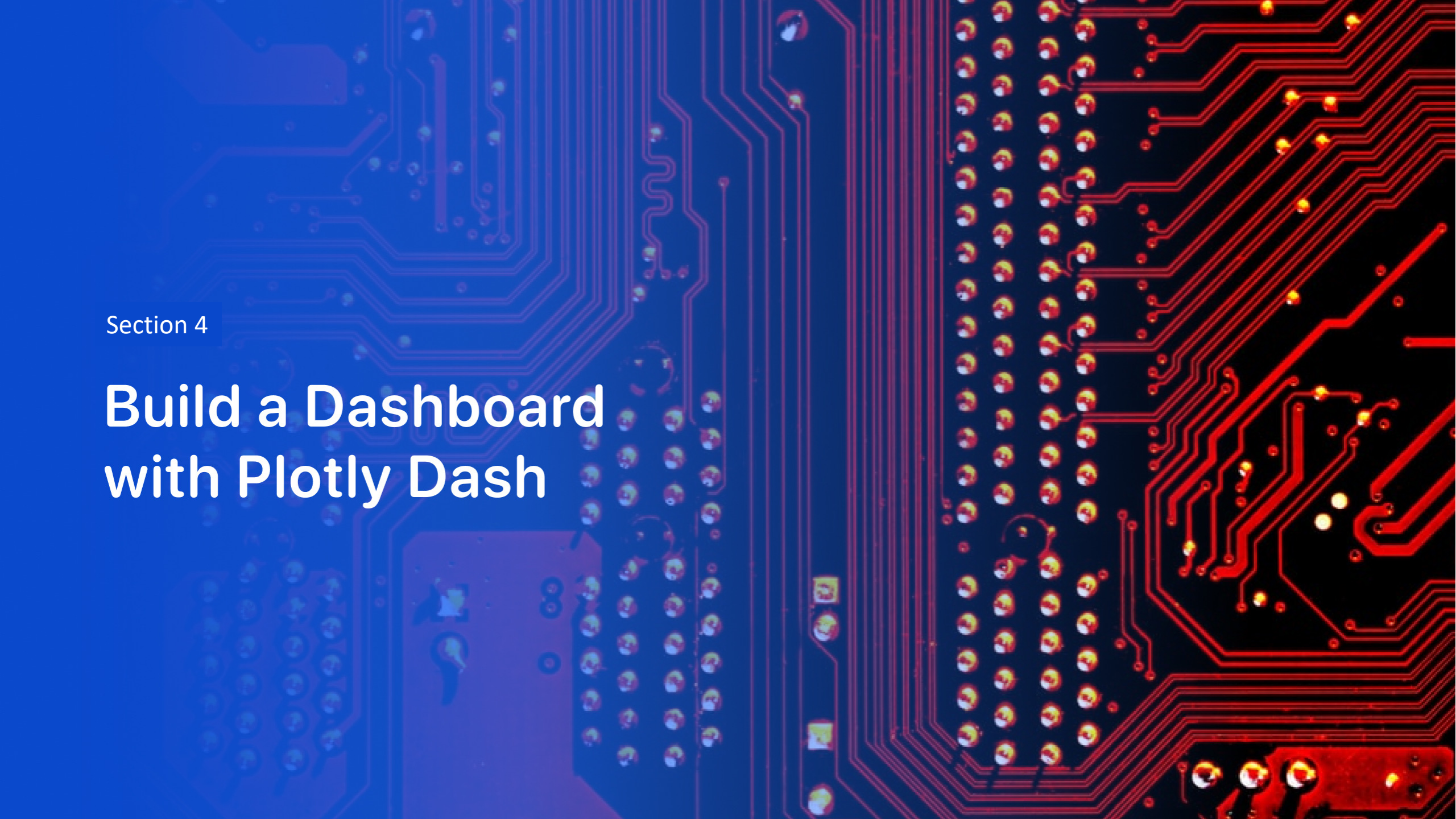


- At Kennedy Space Center, most landings were classified as being successful.
- However, its neighboring launch sites have more launch outcomes (13 vs 33).

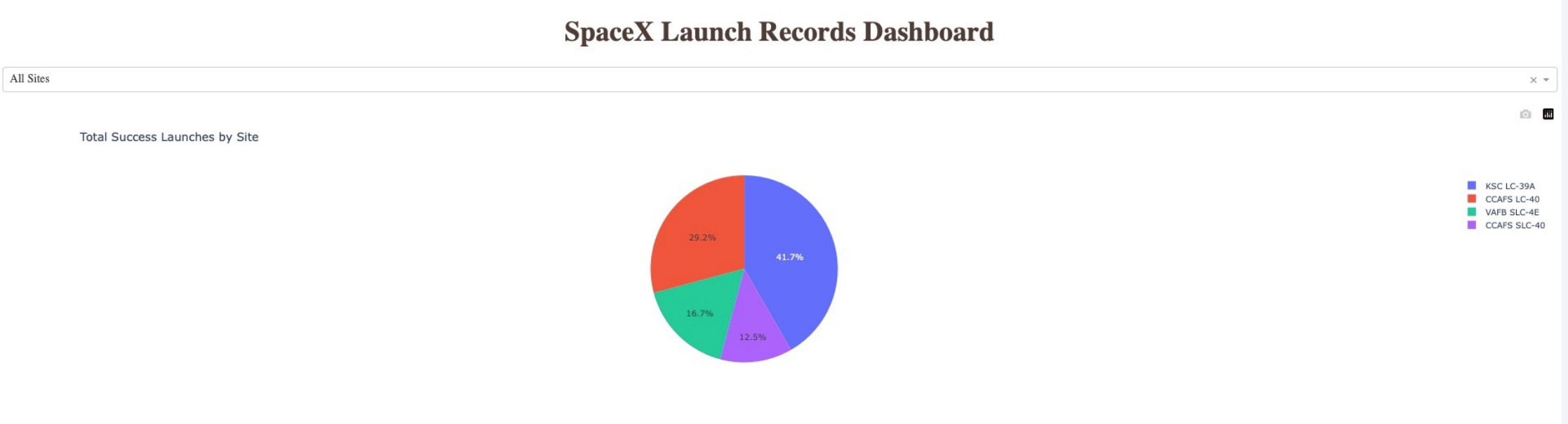
Launch Site Proximity to Railway



- The KSC launch site lies approximately 7.8 km from a railway which can be used to transport rockets to the launch site.

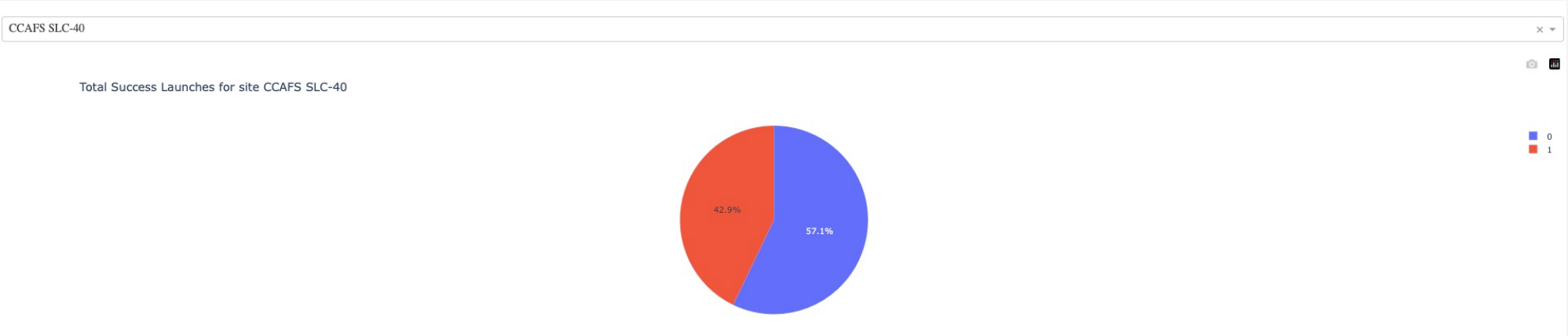

Section 4
Build a Dashboard with Plotly Dash

Dashboard - Launch Site Success Rate



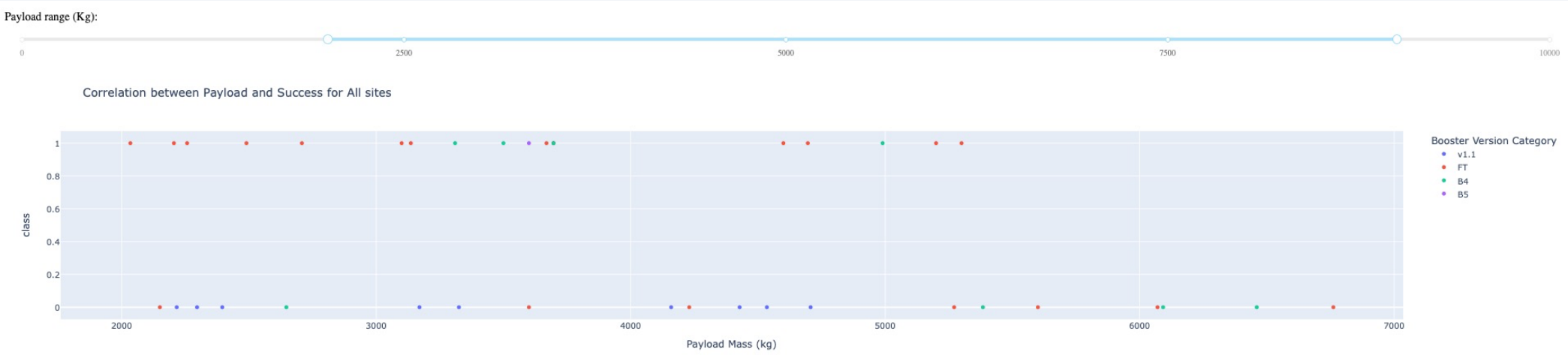
- The largest share of successful launches came from the KSC LC-39A site, making up 41.7% of all successful rocket missions.
- This ratio is found by: Successful launches (specific launch site) / all successful launches.

Dashboard – Detailed Launch Site Outcomes



- Launch site CCAFS SLC-40 has the highest launch-success ratio.
- This ratio is found by: Successful launches (specific launch site) / failed launches (specific launch site).

Dashboard – Correlation Between Payload and Launch Success



- The following dashboard allows one to view the relationship between successful booster landings and the rocket's payload mass.
- Furthermore, it is affected by prior filtering of the launch site and payload mass using the provided slider.
- In the above, it can be observed that rockets using boosters of type "FT" and "B4" are strongly correlated with successful launches, among a broad variety of payloads.

Section 5

Predictive Analysis (Classification)

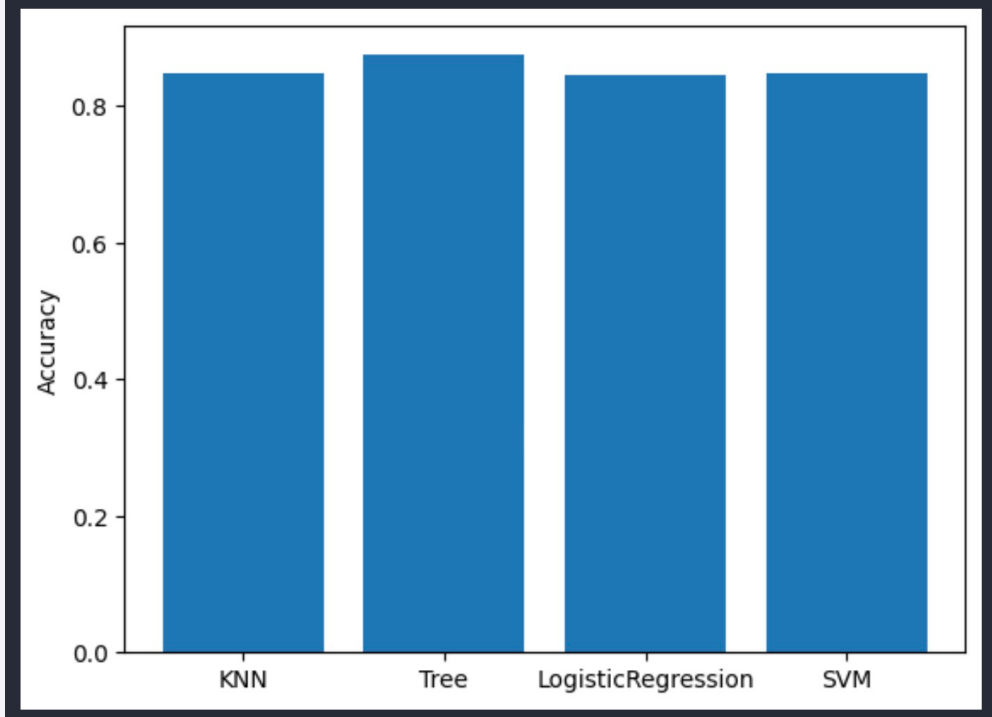
Classification Accuracy

```
# Bar Chart model acc

plt.bar(range(len(algorithms)), list(algorithms.values()), align = 'center')
plt.xticks(range(len(algorithms)), list(algorithms.keys()))
#plt.xlabel('Model')
plt.ylabel('Accuracy')
```

✓ 0.1s

Text(0, 0.5, 'Accuracy')

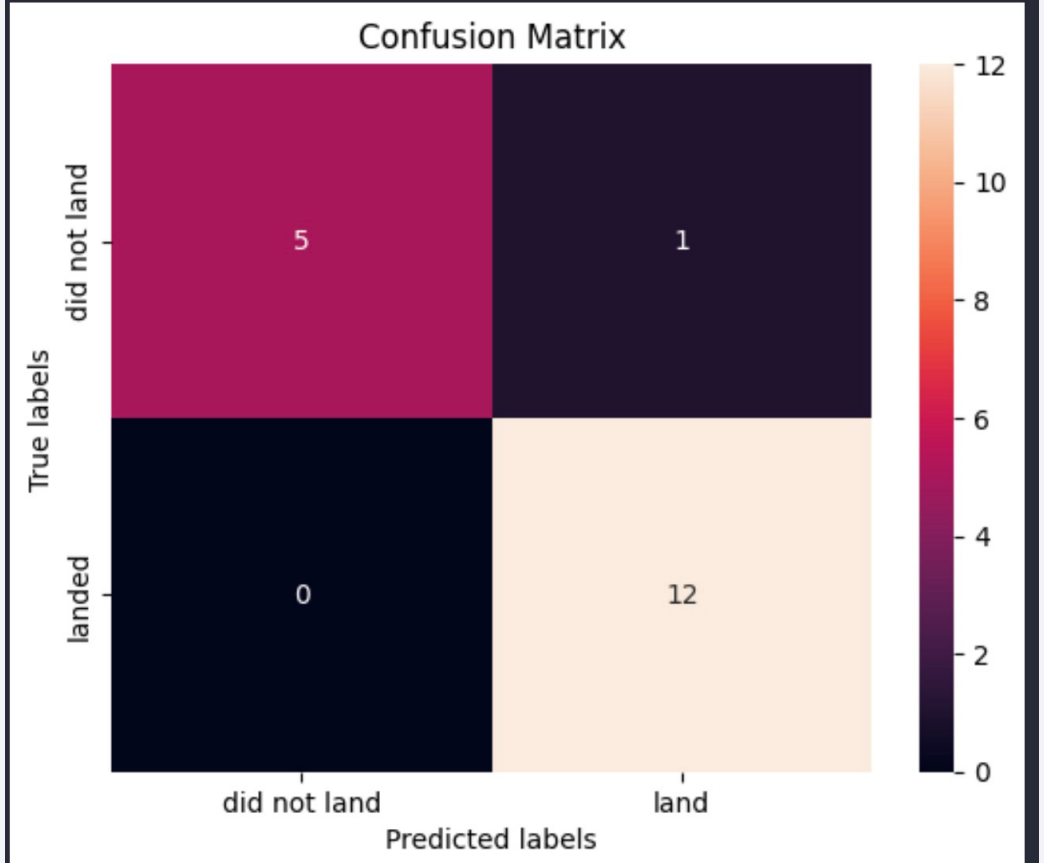


- The trained decision tree provides the highest model accuracy with 87.5%.
- It is hereby important to not focus on the accuracy metric alone since it can be misleading (known as the “accuracy paradox”).

Confusion Matrix

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

✓ 0.1s

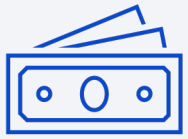


- The confusion matrix shows that the model predicted in the following manner:
  - False Positives: 1
    - The model predicted the rocket to land, but it did not.
  - False Negatives: 0
    - The model predicted the rocket to crash, but it landed successfully.
  - True Positives: 12
    - The model predicted the rocket will land correctly
  - True Negatives: 5
    - The model predicted the rocket will crash correctly

Conclusions

The CCAFS SLC-40 launch site has the highest launch success ratio, making it a valuable candidate to choose as a primary launch location.

Rockets sent to the ES-L1, GEO, HEO, or SSO orbit have a booster landing rate of 100%, furthermore making these orbit's very attractive for SpaceY.



- If a rocket's boosters can land successfully, the launch cost equates to 62 million dollars. If the rocket boosters fail or traditional vendors are used, the cost for one launch is 165 million dollars.
- Launching all rockets using traditional vendors/technologies, costs of 165 x 18 = **2.970 billion** dollars arise.
- Using the trained model, launch costs equate to (62 x 12) + (6 x 165) = 744 + 990 = **1.734 billion** dollars.
- A difference of 2.970 – 1.734 = **1.236 billion dollars is being saved using the developed machine learning model.**

Thank you!