IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sujeet Ghorpade
July 21, 2024

Outline

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

Executive Summary

Summary of methodologies

- Summary of methodologies
- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

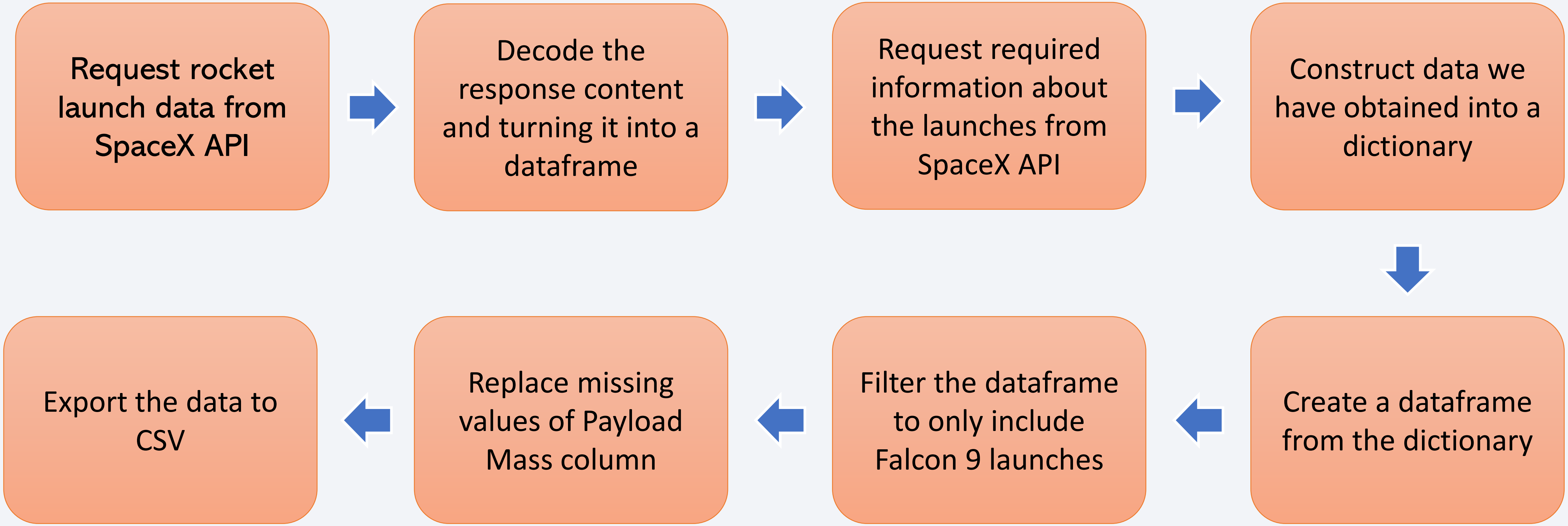
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- 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 evaluation of classification models to ensure the best results

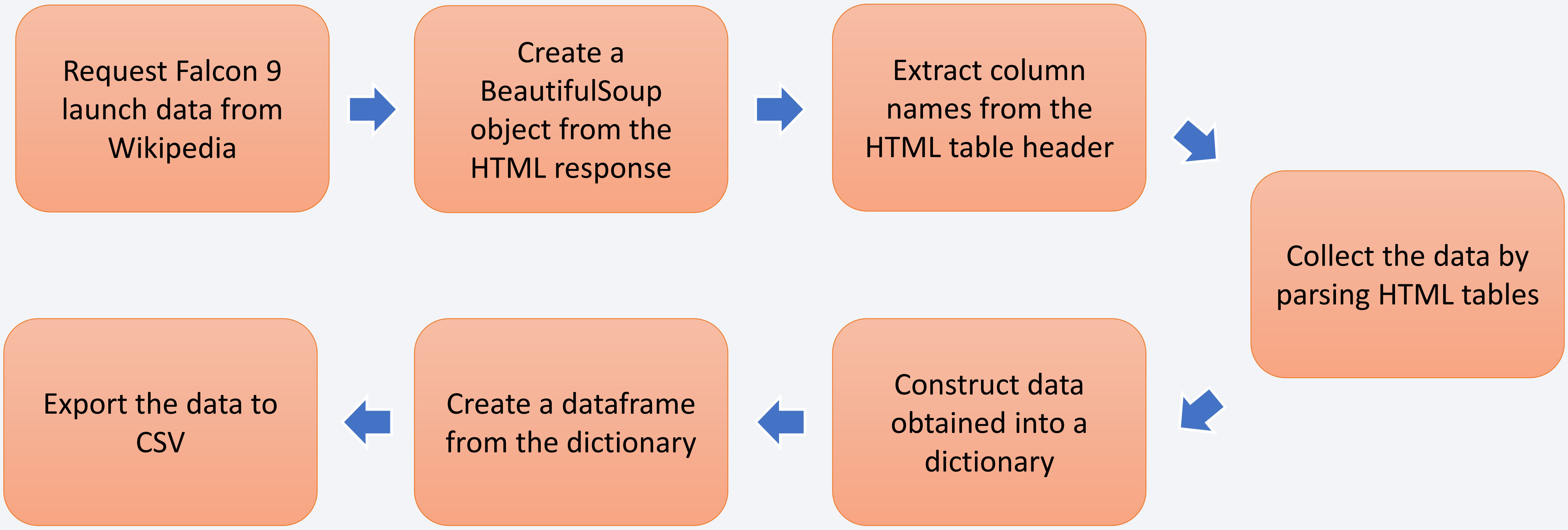
Data Collection

- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- Data obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

Several occurrences in the data set show unsuccessful booster landings. In other cases, landing attempts were unsuccessful owing to accidents, such as True.

The term "Ocean" refers to a successfully landed mission in a certain ocean zone, whereas "False Ocean" refers to a failed landing. True RTLS indicates that the mission successfully landed on a ground pad. False RTLS indicates a failed landing on a ground pad.True ASDS indicates that the task was successfully completed on a drone ship. False ASDS indicates that the mission outcome was unsuccessfully landed on the drone ship.

We translate results into Training Labels, where "1" indicates a successful landing and "0" indicates a failure.

EDA with Data Visualization

This work involved plotting charts for flight number, payload mass, launch site, orbit type, success rate, and payload mass. Yearly Trend

Scatter plots depict the relationships between variables. If a relationship exists, it can be utilized in a machine learning model.

Bar charts display comparisons between distinct categories. The purpose is to demonstrate the correlation between the compared categories and a measured value.

Line charts depict patterns in data across time.

EDA with SQL

In this work following SQL queries were performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

- Markers for all launch sites have been updated, including a circle, popup label, and text label for NASA Johnson Space Center, which uses latitude and longitude coordinates as its starting position.
- Markers with Circle, Popup Label, and Text Label for all Launch Sites utilizing latitude and longitude coordinates to indicate their geographical location and closeness to the equator and coastlines.
- Colored markers indicate success (Green) and failure (Red) for each launch point using Marker Cluster to find places with high success rates.
- Distances between a launch site and its surroundings:
- Colored lines indicate distances between Launch Site KSC LC-39A and nearby locations, such as railways, highways, coastlines, and cities.

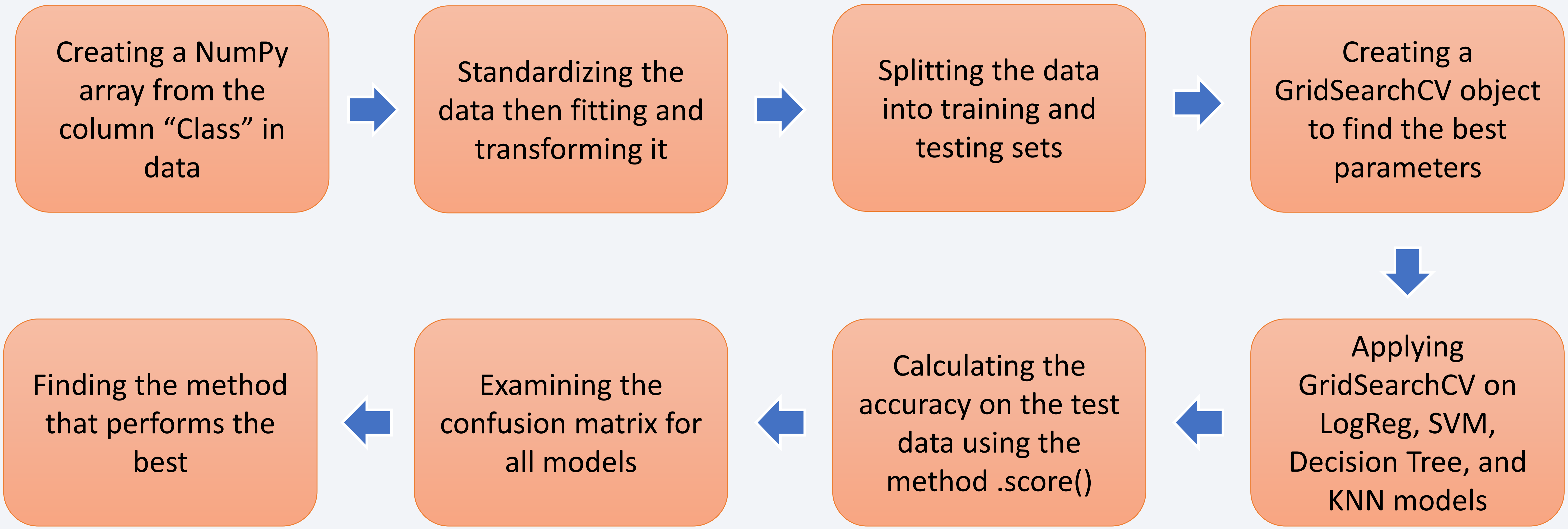
Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List: A dropdown list has been added to allow users to pick their Launch Site.

Pie Chart of Successful Launches (All Sites/Specific Site):

- A pie chart now displays the overall number of successful launches across all sites, as well as the success/failure ratio for a single launch site.
- Added a slider to choose the Payload Mass Range.
- Added a scatter graphic comparing payload mass and success rate for different booster versions.

Predictive Analysis (Classification)



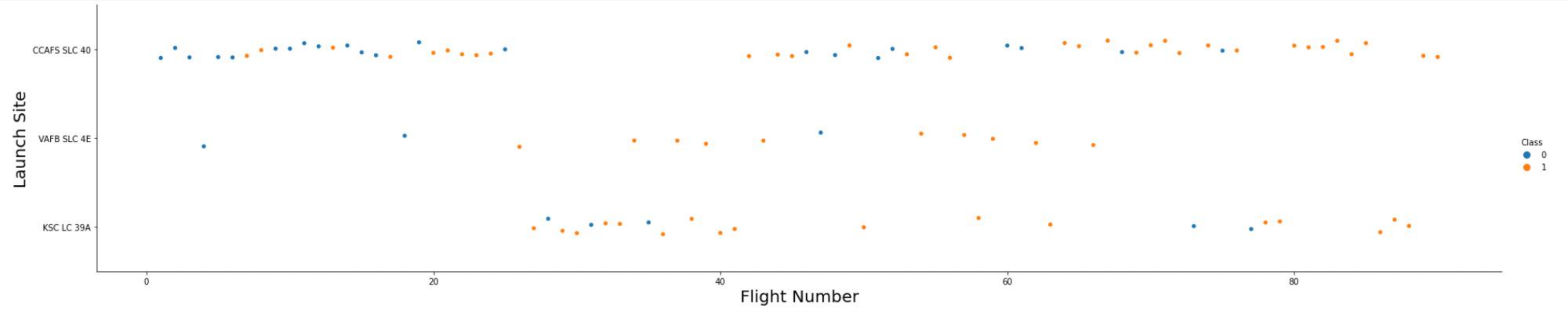
Results

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

Section 2

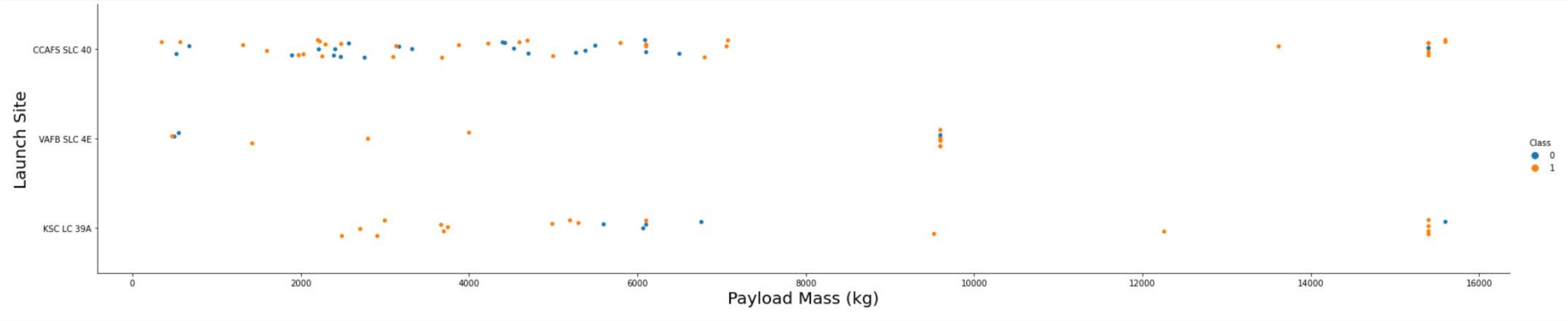
Insights drawn from EDA

Flight Number vs. Launch Site



- The CCAFS SLC 40 launch site accounts for almost half of all launches, although VAFB SLC 4E and KSC LC 39A have greater success rates. All previous flights were unsuccessful.
- It may be expected that each new launch has a larger chance of success.

Payload vs. Launch Site



- Better payload mass correlates with better success rates at each launch location. The majority of missions weighing above 7000 kg were successful.
- KSC LC 39A has a 100% success rate with payload masses under 5500 kg.

Success Rate vs. Orbit Type

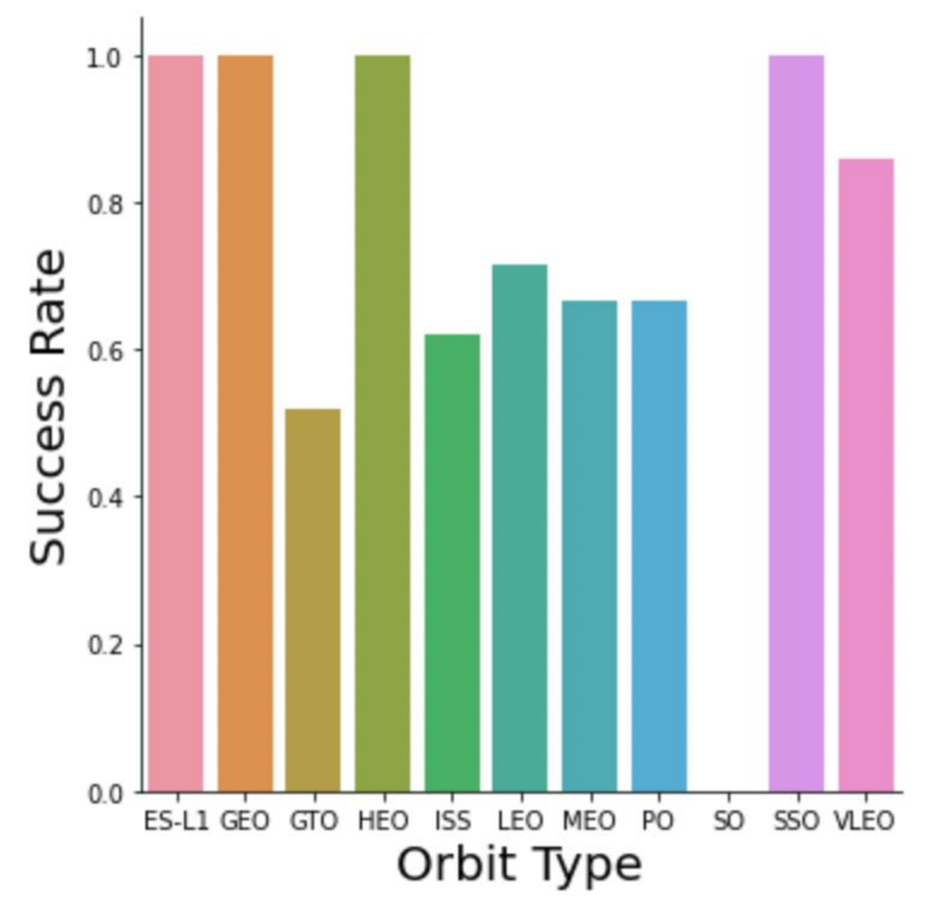
- Orbits with 100% success rate:

ES-L1, GEO, HEO, SSO

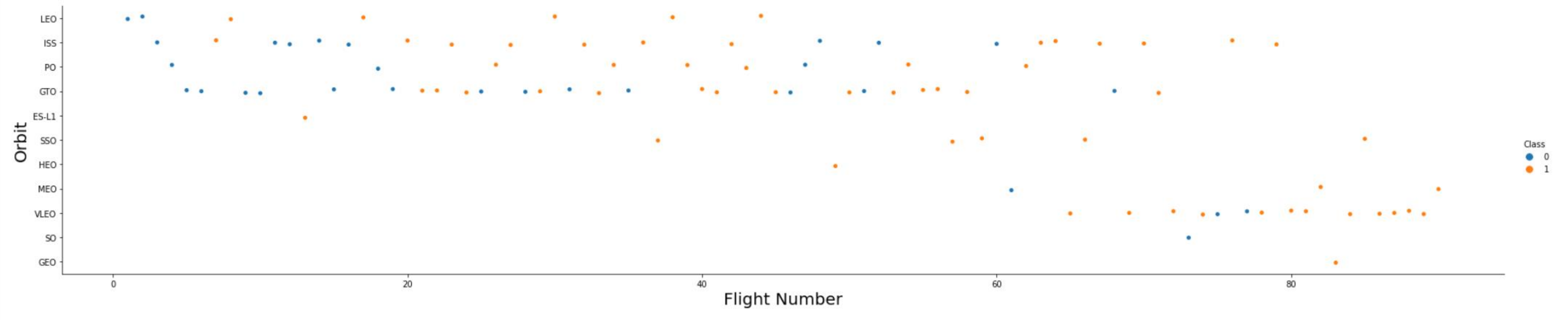
- Orbits with 0% success rate: SO

- Orbits with success rate between 50% and 85%:

 GTO, ISS, LEO, MEO, PO

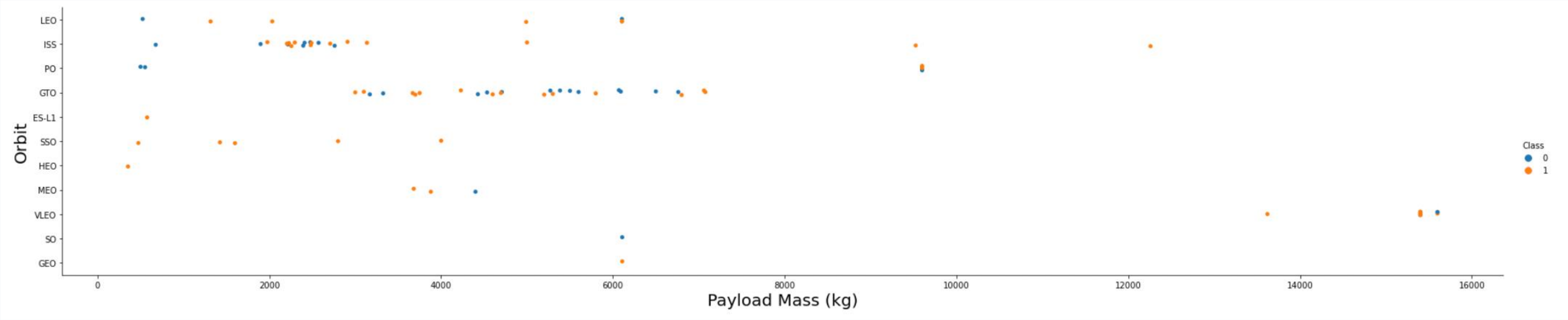


Flight Number vs. Orbit Type



- In LEO orbit, success is linked to the number of flights. However, in GTO orbit, there appears to be no correlation.

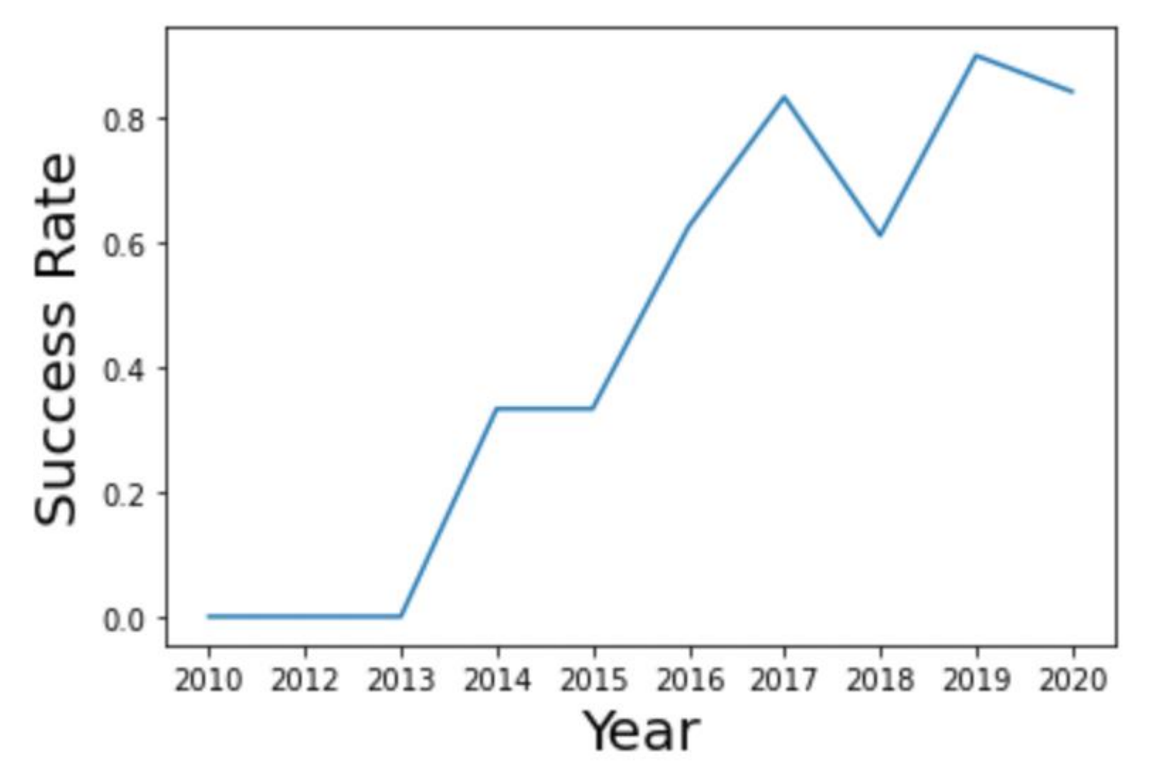
Payload vs. Orbit Type



- Heavy payloads negatively impact GTO orbits but positively affect Polar LEO (ISS) orbits.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020, with a stagnant period between 2014-2015.
- The success rate briefly plumets in 2018 but quickly raises back up in 2019.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

- Displaying the names of the space mission's distinct launch locations.

Launch Site Names Begin with 'CCA'

In [5]:
```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

- Displaying 5 entries where the launch site begins with the string 'CCA'.

Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

- Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

- Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[9]:
```

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

- List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[10]:
```

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

- Displaying the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

- Displaying the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

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

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

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

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

landing__outcome	count_outcomes
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

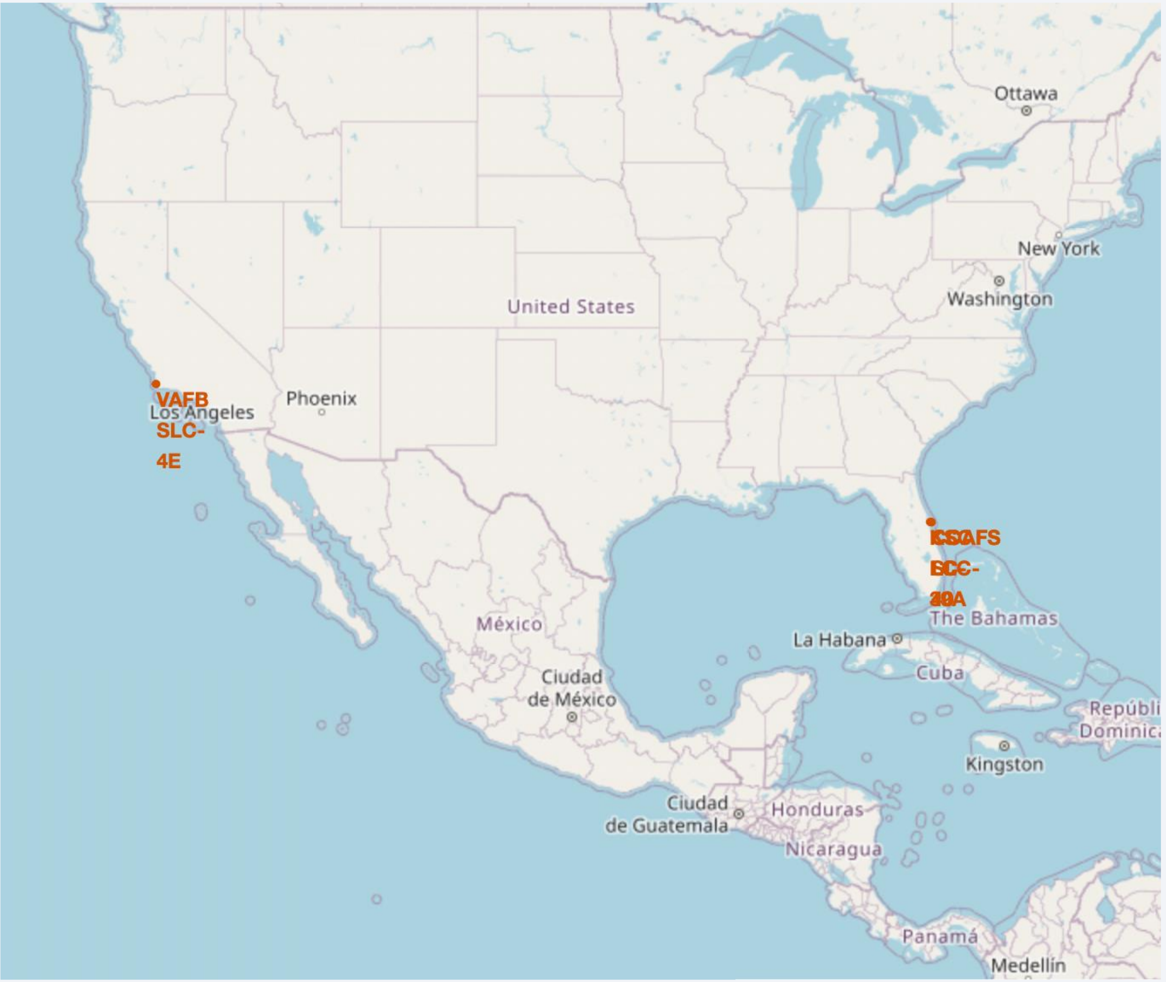
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

Launch Site Location Markers

- Most launch locations are along the equator line. At the equator, land moves faster than anywhere else on Earth's surface. The Earth's surface at the equator moves at 1670 km/hour. When a ship is launched from the equator, it travels into space and continues to move around the Earth at its original speed. This is due to inertia.
- This speed enables the spacecraft to maintain a stable orbit.
- Launch locations are close to the shore, reducing the possibility of debris dropping or exploding nearby.



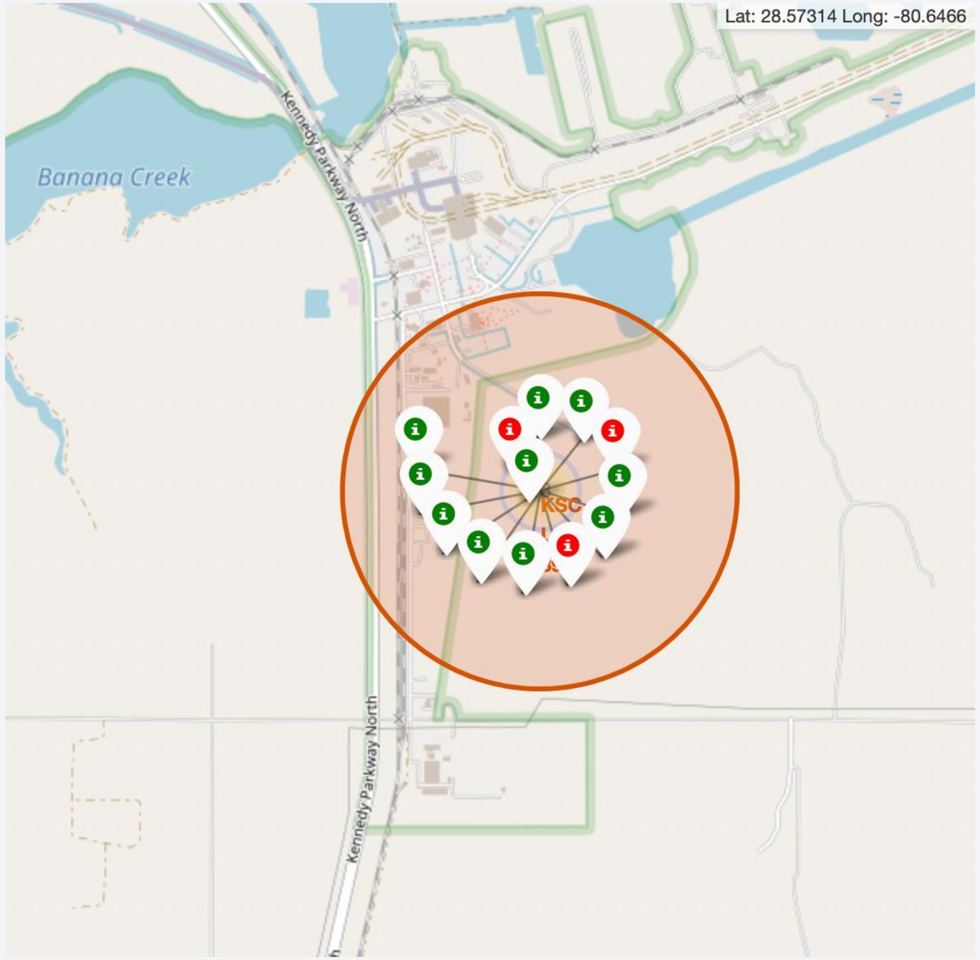
Color-Coded Launch Record Markers

- Color-coded markings help identify launch sites with high success rates.

-> Green Marker indicates successful launch.

-> Red marker indicates failed launch.

- Launch Site KSC LC-39A has a high success rate.



Distance from the launch site KSC LC-39A to its proximities

- Visual study of launch location KSC LC-39A reveals that it is:

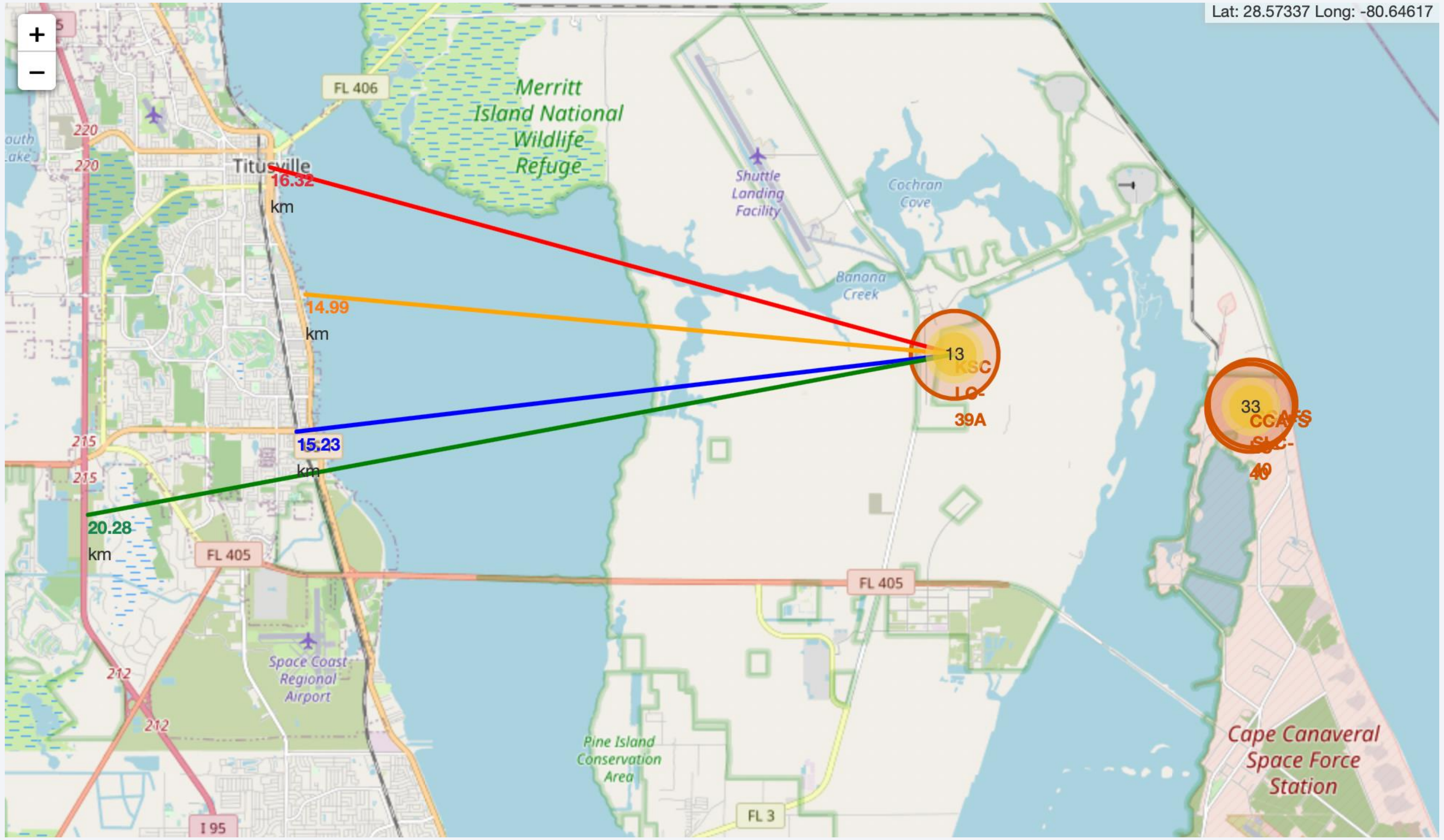
1. relative proximity to the railway (15.23 kilometers)

2. comparatively near to the highway (20.28 kilometers)

3. quite near to the shore (14.99 km)

• The KSC LC-39A launch site is adjacent to the nearest city, Titusville (16.32 miles).

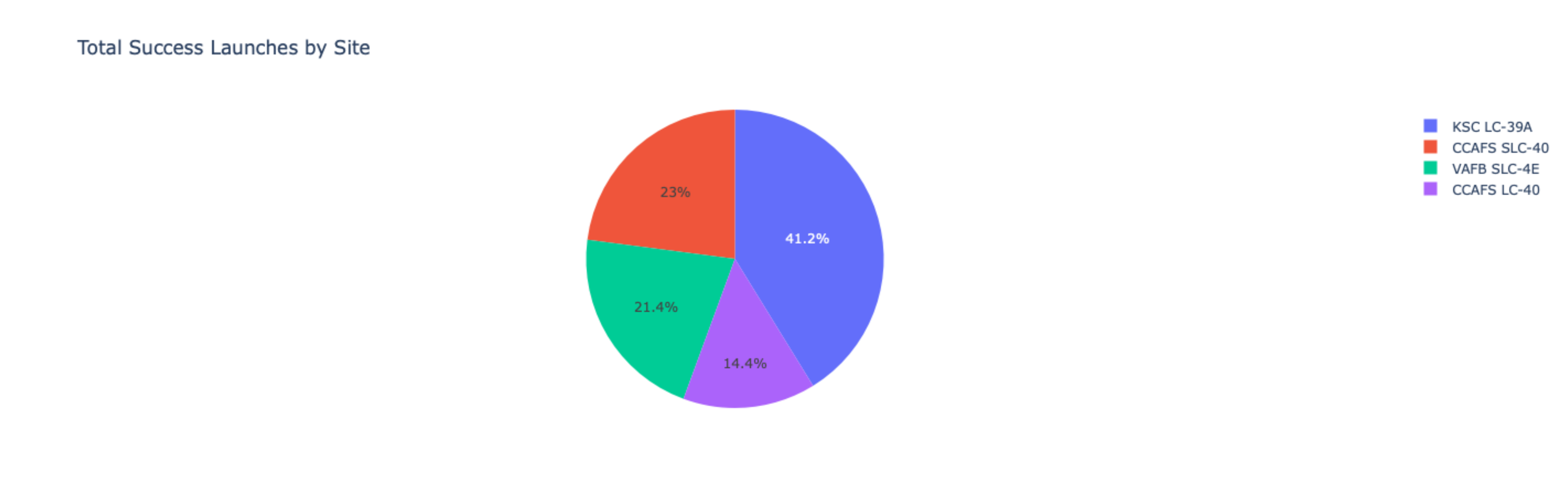
• A failed rocket's tremendous speed allows it to travel up to 15-20 kilometers in seconds. It may pose a risk to inhabited regions.



Section 4

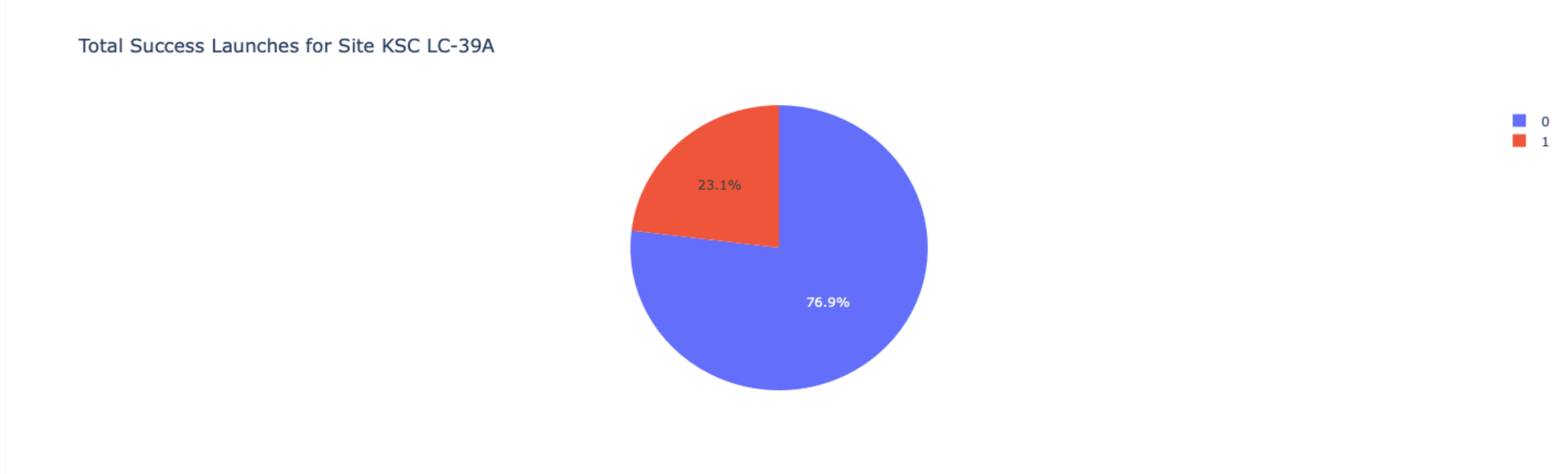
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The figure indicates that KSC LC-39A has the most number of successful launches among all locations.

Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%), with 10 successful landings and only 3 failed attempts.

Payload Mass vs Launch Outcome

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

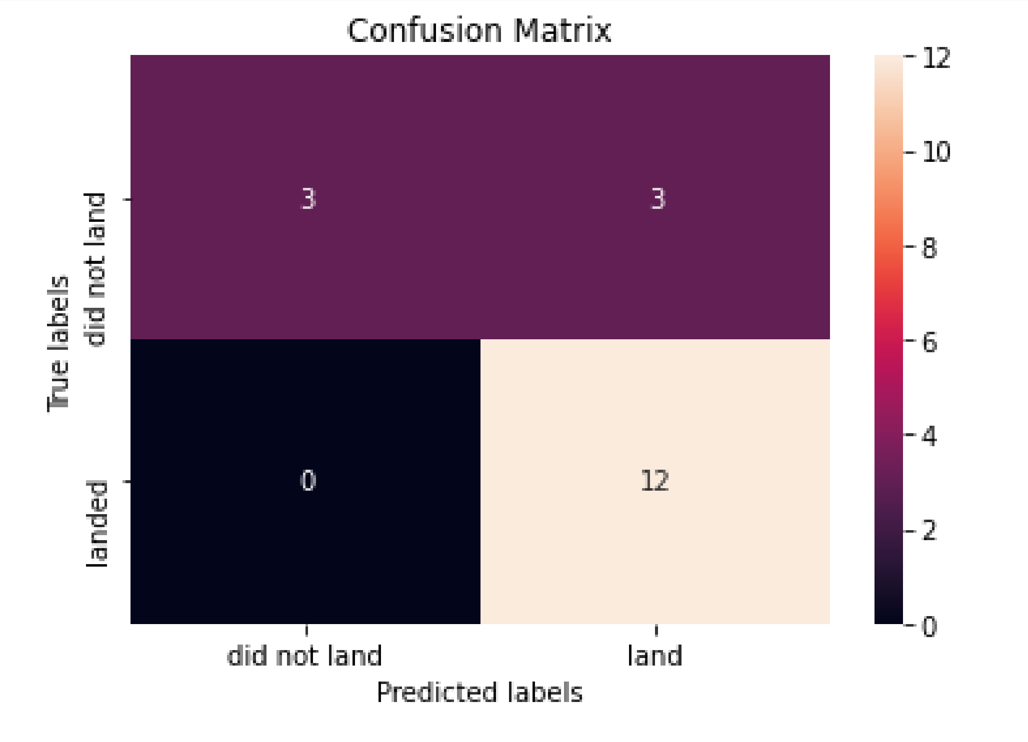
```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

- From our findings we see that the decision tree classifier is the model with the highest classification accuracy

Confusion Matrix

- The confusion matrix demonstrates the decision tree classifier's ability to discriminate between classes.

- The biggest issue is false positives.The classifier considers failed landings as successful ones.



Conclusions

We can deduce that:

- A launch site's success rate increases as the number of flights increases.
- From 2013 to 2020, the launch success rate increased.
- Orbits ES L1, GEO, HEO, SSO, and VLEO achieved the highest success rate.
- KSC LC 39A was the most successful launch of all locations.
- The Decision tree classifier is the most effective machine learning method for this problem.

Thank you!