Applied Data Science Capstone

Pun Yi Jie

April 21, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

1. Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium lab
- Interactive Dashboard with Plotly Dash
- Machine Learning Prediction (Classification)

2. Summary of all results

- There is a correlation between some features of the rocket launches and the outcome of the launches based on the graph.
- It is found that Decision Tree are the best machine learning algorithm in predicting the outcome of the launches.

Introduction

1. **Project background and context**

   - The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes the launch of Falcon 9 rockets for a cost of 62 million dollars; in comparison, other suppliers charge upwards of 165 million dollars each launch; a large portion of the cost reductions are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the launch cost if we can ascertain if the first stage will land. Our prediction will be based on public data and machine learning methods to determine if SpaceX will reuse the first stage.

2. **Problems you want to find answers**

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

Methodology

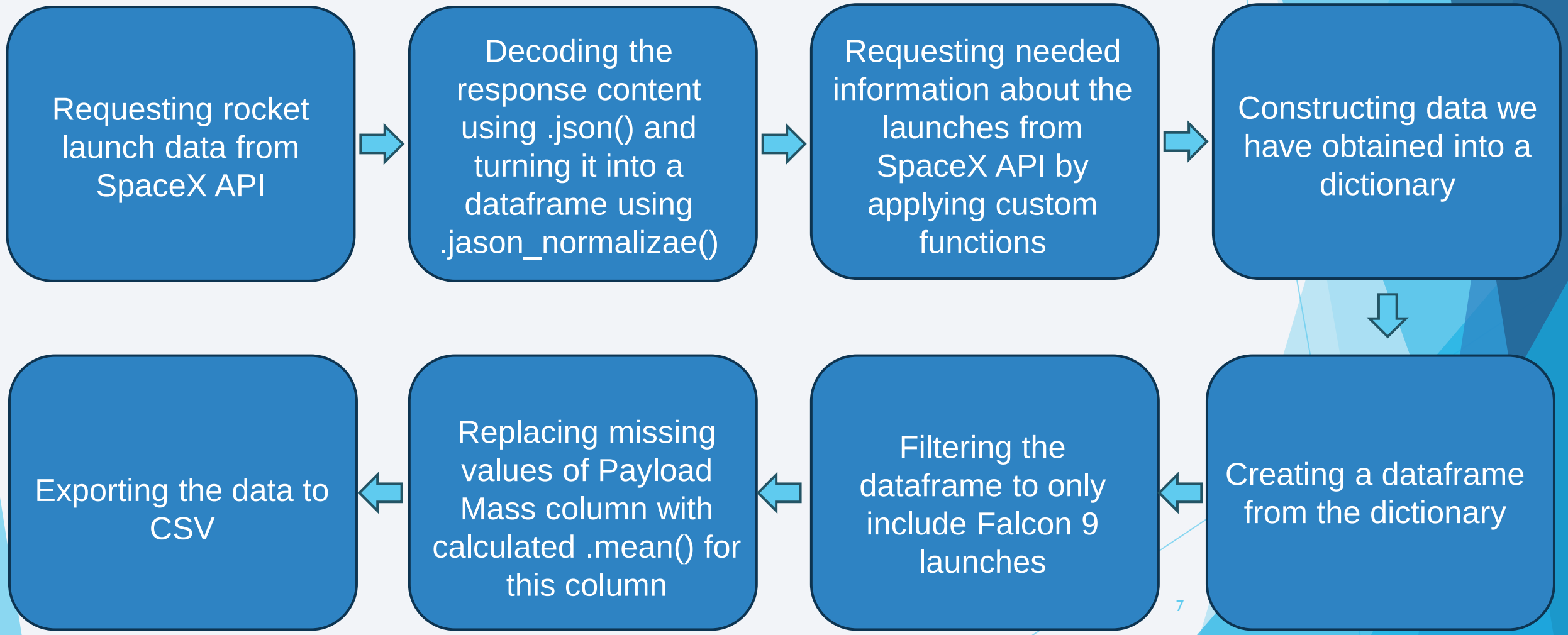
Executive Summary

1. Data collection methodology
   - Using SpaceX Rest API
   - Using Web Scraping
2. Perform data wrangling
   - Filtering the data
   - Dealing with missing values
3. Perform exploratory data analysis (EDA) using visualization and SQL
4. Perform interactive visual analytics using Folium and Plotly Dash
5. Perform predictive analysis using classification models
   - Building, tuning and evaluation of classification models

Data Collection

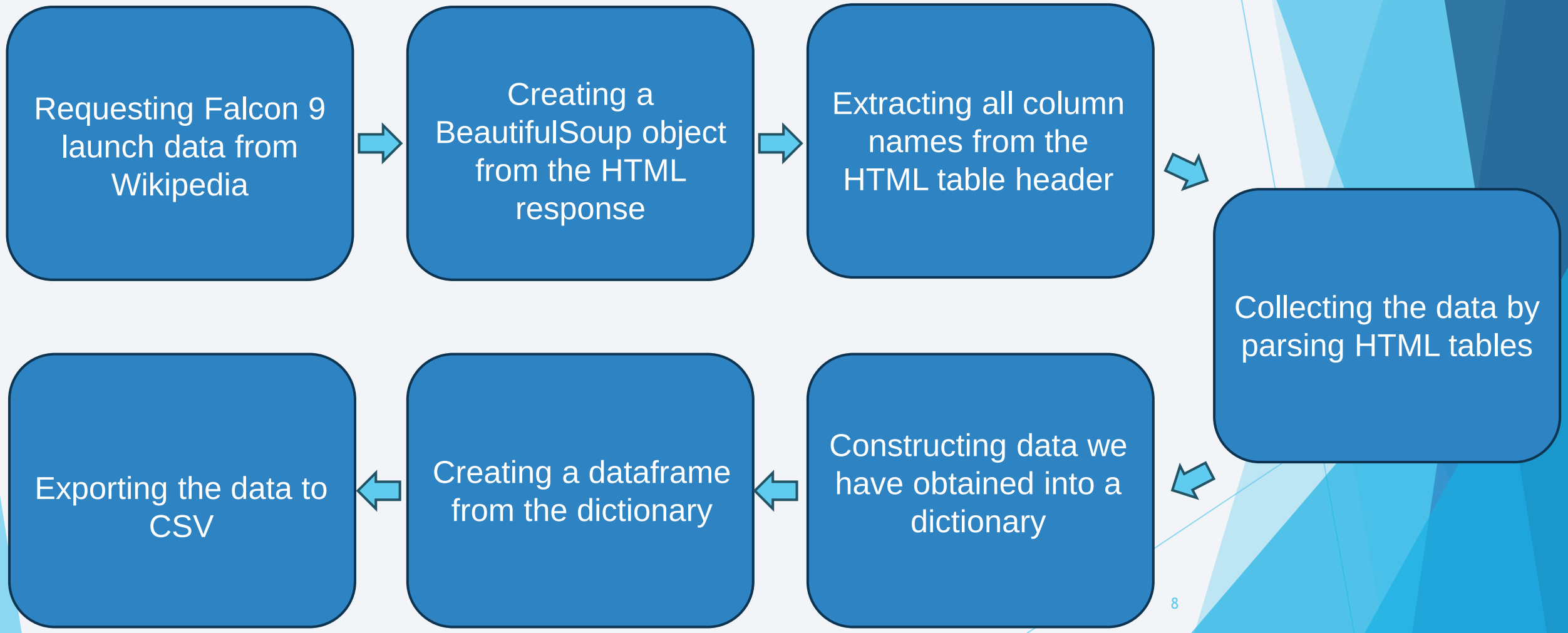
- A mix of web scraping data from a table in SpaceX's Wikipedia article and API queries from the company's REST API were used in the data collecting procedure.
- In order to obtain all the information needed for a more thorough study, we had to employ both of these data gathering techniques for the launches.

Data Collection – SpaceX API



GitHub URL: Data Collection API

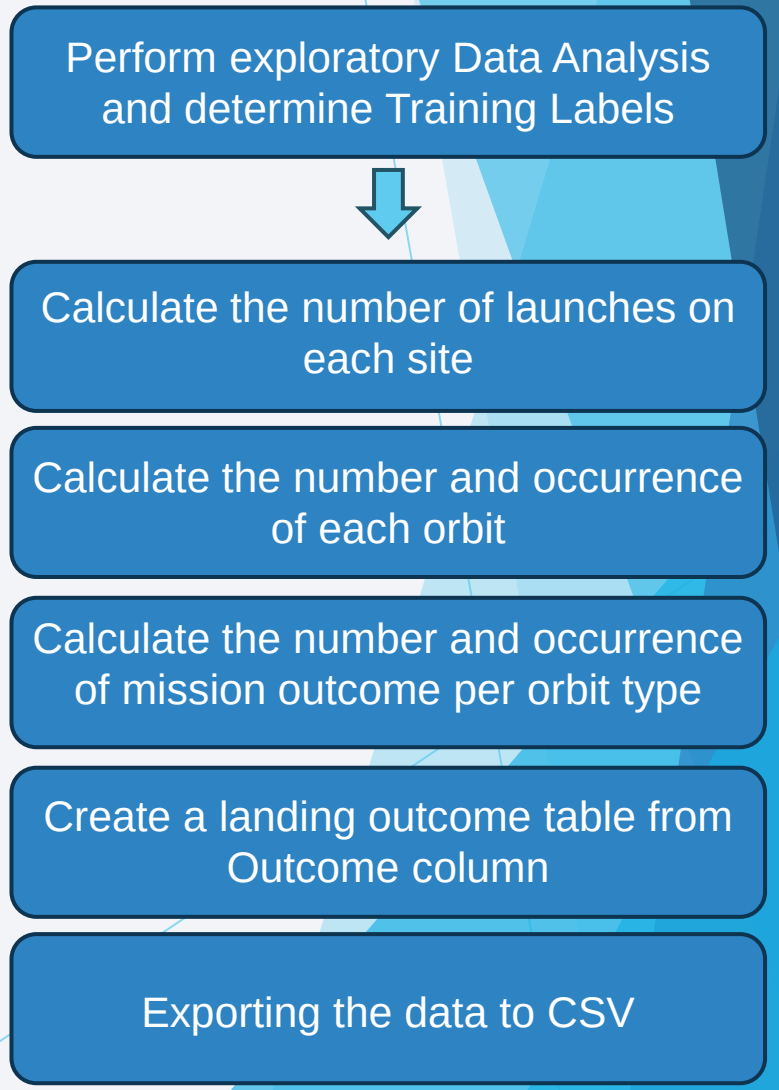
Data Collection – Scraping



GitHub URL: Data Collection With Web Scraping

Data Wrangling

- There are several instances in the data set where the booster failed to land successfully. Sometimes a landing attempt was made but was unsuccessful due to an accident; for instance, a landing that was successful in reaching a particular area of the ocean is known as a True Ocean, whereas a landing that was unsuccessful in reaching a certain area of the ocean is known as a False Ocean. If the mission was successful, the landing on a ground pad is indicated by true RTLS. An failed landing to a ground pad is indicated by a false RTLS.A successful landing of the mission's conclusion on a drone ship is referred to as true ASDS. An failed landing on a drone ship is indicated by a false ASDS.



GitHub URL: Data Wrangling

EDA with Data Visualization

- A few charts were plotted, for example, Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type and etc.
- We can observe the scatter plot to see if there is any relationship between the variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in date over time.

GitHub URL: EDA with Visualization

EDA with SQL

- **Performed SQL queries:**
  - 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 date 2010-06-04 and 2017-03-20 in descending order

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- **Markers of all Launch Sites**
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- **Coloured Markers of the launch outcomes for each Launch Site**
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- **Distances between a Launch Site to its proximities**
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

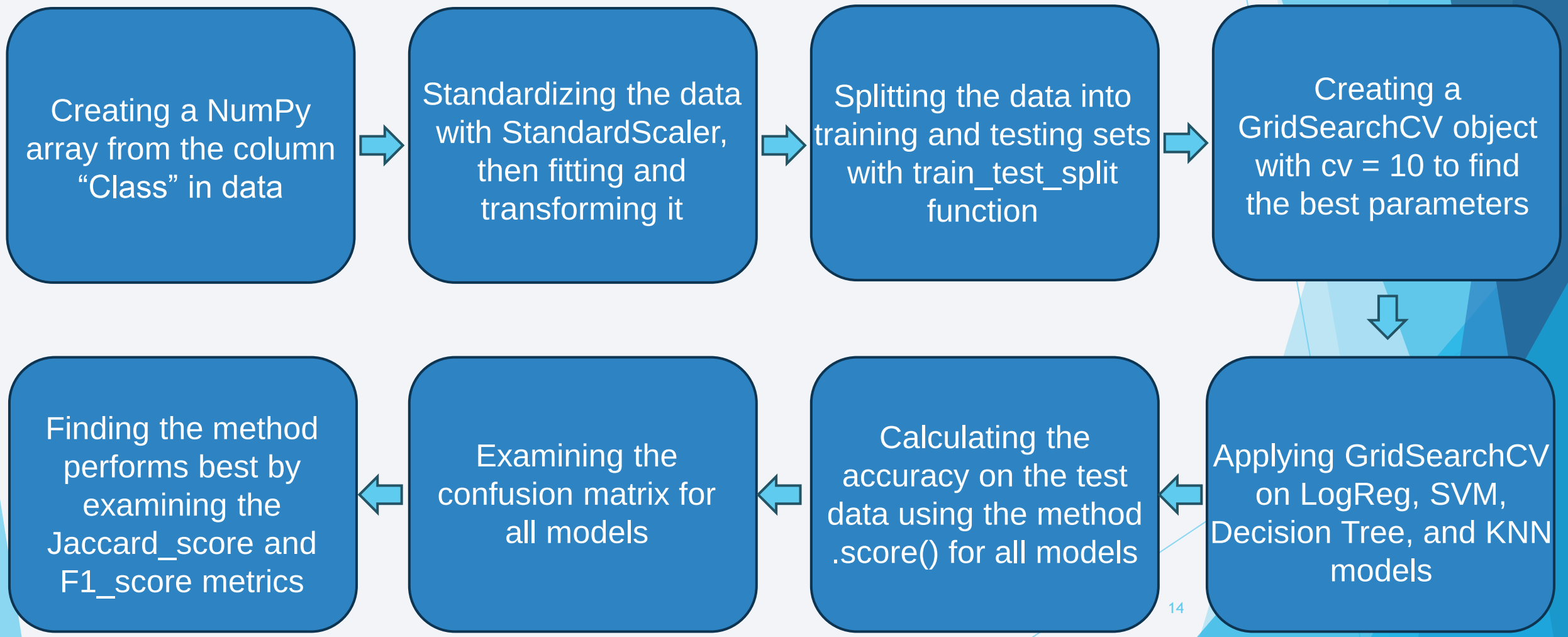
GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List**
  - Added a dropdown list to enable Launch Site selection.
- **Pie Chart showing Success Launches (All Sites/Certain Site)**
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- **Slider of Payload Mass Range**
  - Added a slider to select Payload range.
- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions**
  - Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL: Interactive Dashboard with Plotly Dash

Predictive Analysis (Classification)



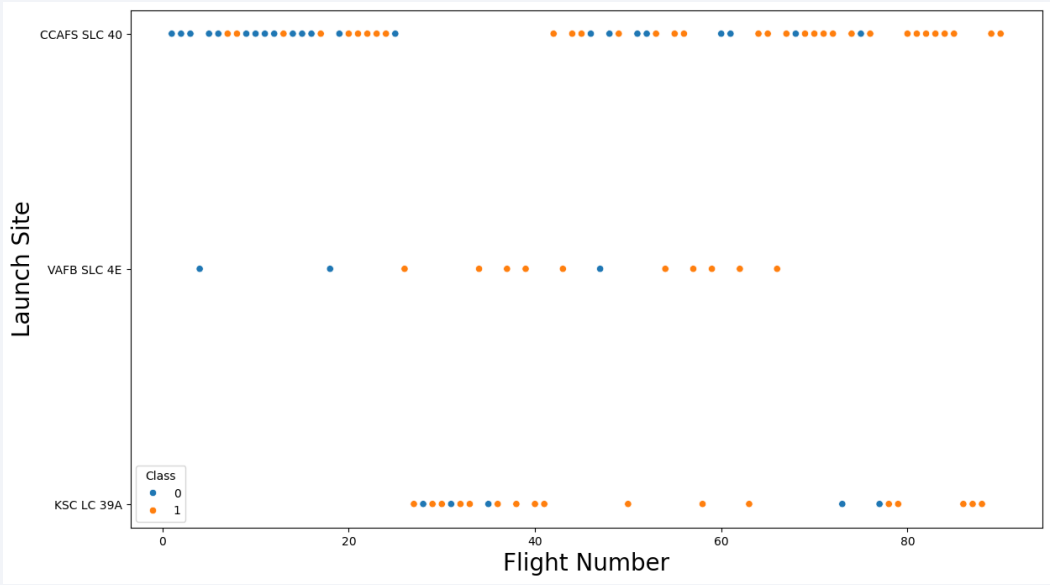
GitHub URL: Machine Learning Prediction

Results

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

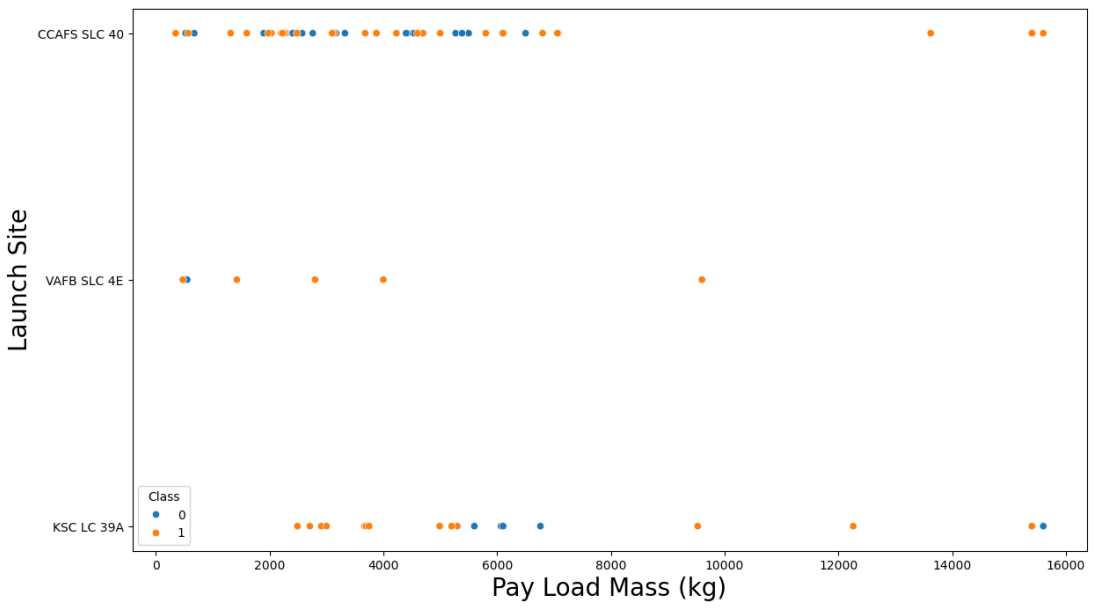

Exploratory
Data Analysis

Flight Number vs. Launch Site



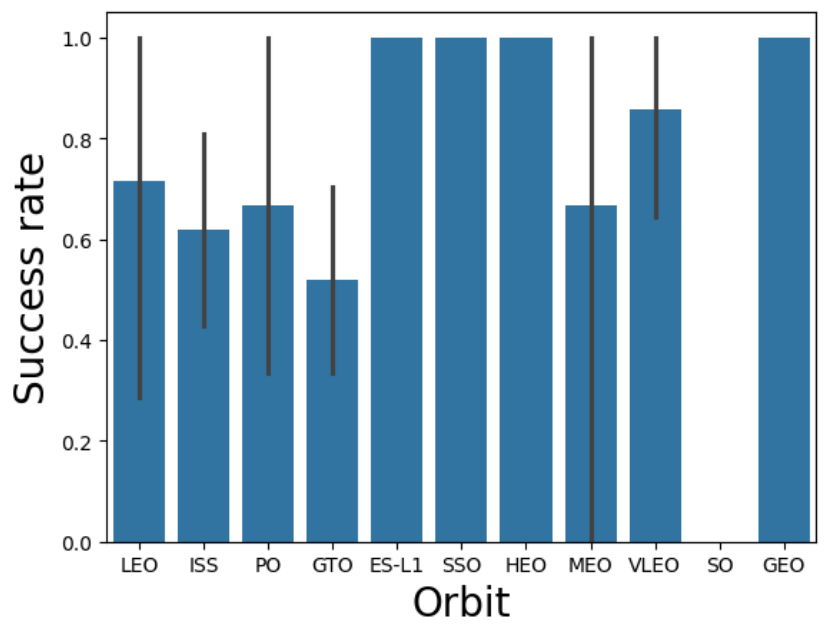
- The majority of the early flights failed while the majority of the latest flights succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



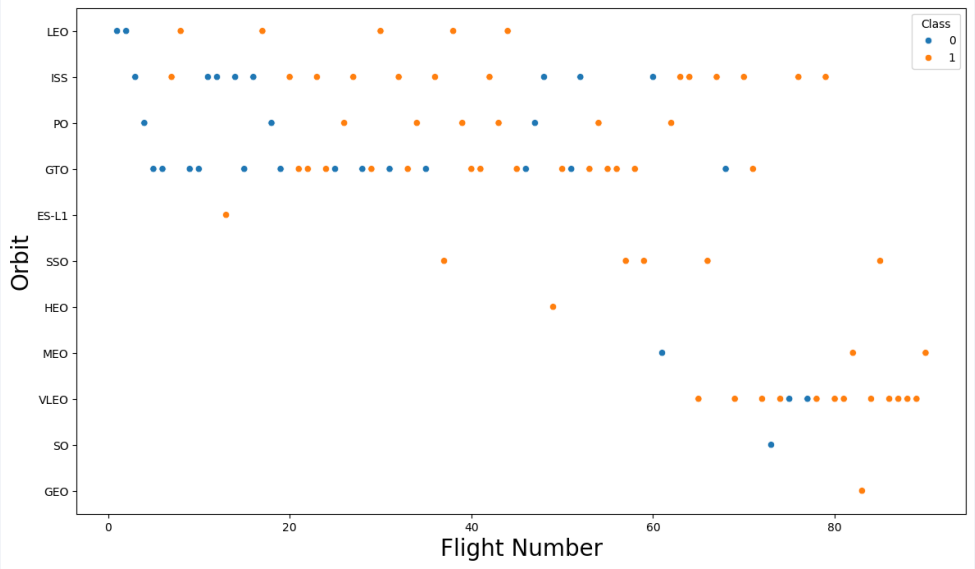
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



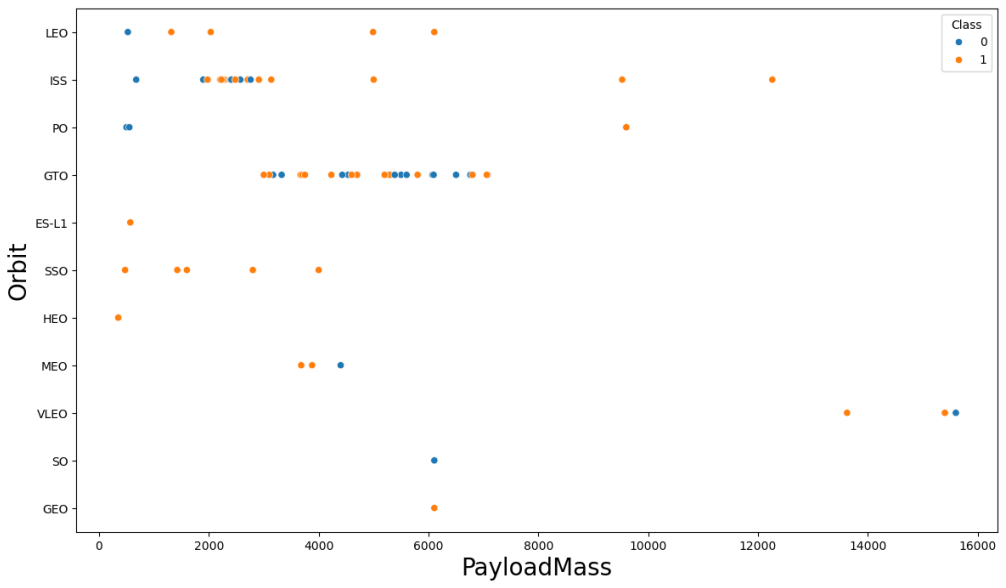
- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



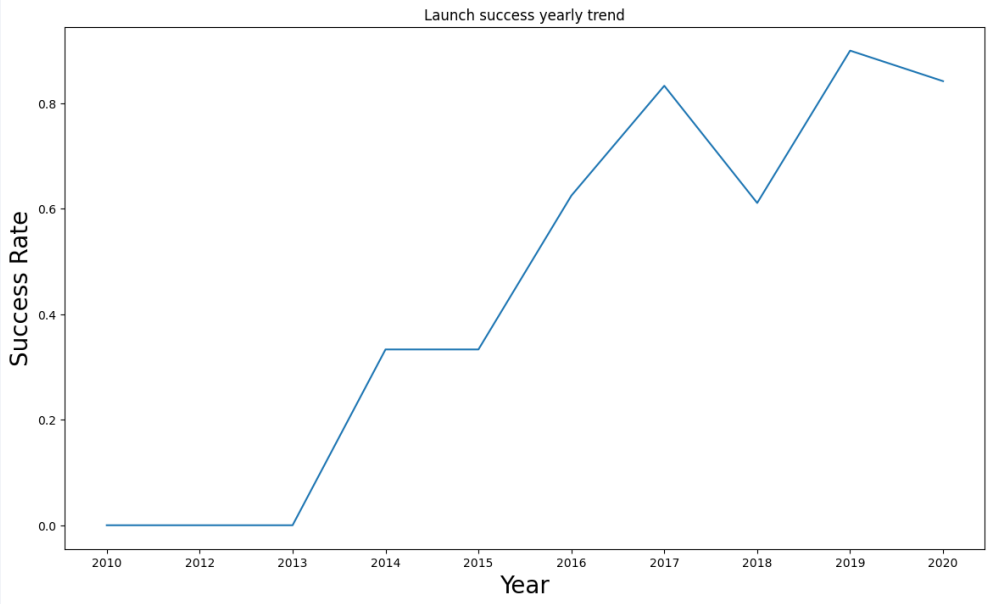
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select distinct launch_site from SPACEXTBL;
```

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

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

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

Total payload mass by NASA (CRS)
45596

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

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION
```

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

Average payload mass by Booster Version F9 v1.1
2928.4

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

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS "Date of first successful landing outcome in ground pad" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Su
```

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

Date of first successful landing outcome in ground pad
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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000
```

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

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM
```

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

number_of_success_outcomes	number_of_failure_outcomes
100	1

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
%sql SELECT substr(Date, 6,2), BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AN
```

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

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

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS Landing_Count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '201
```

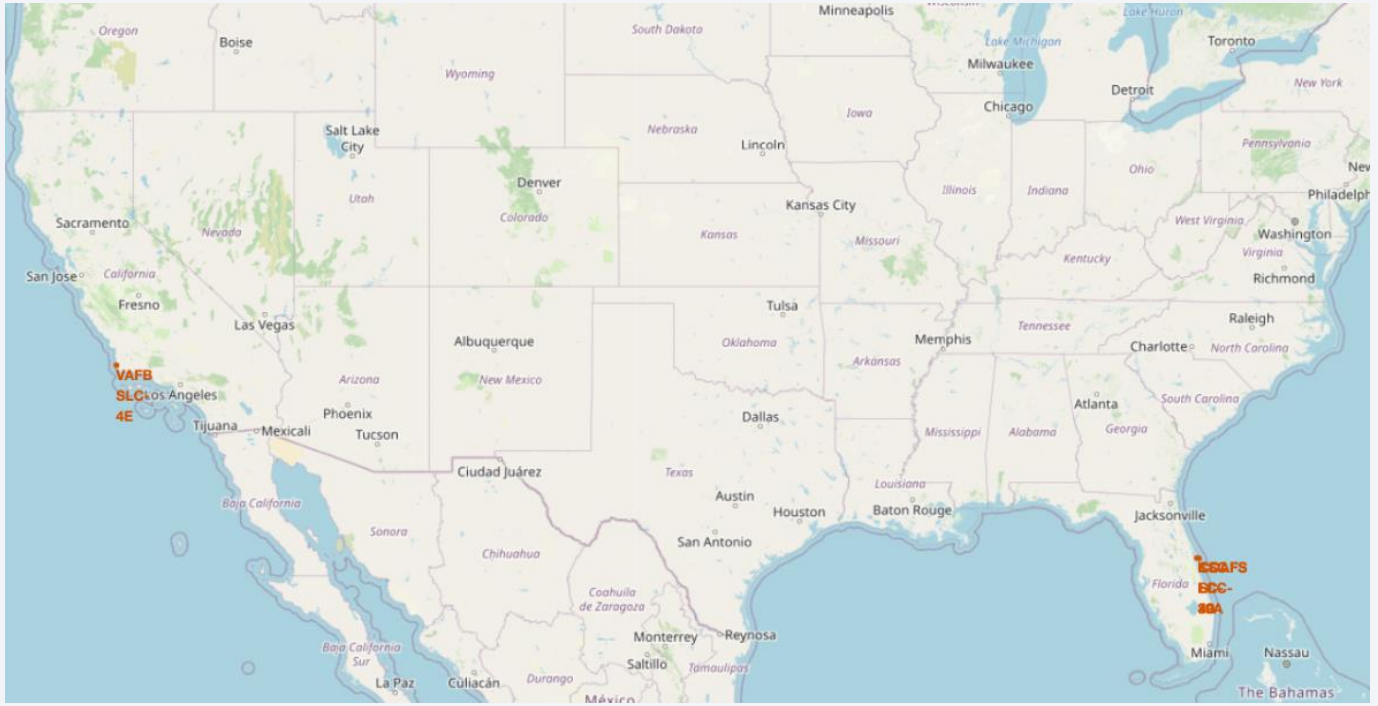
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Landing_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

- 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.

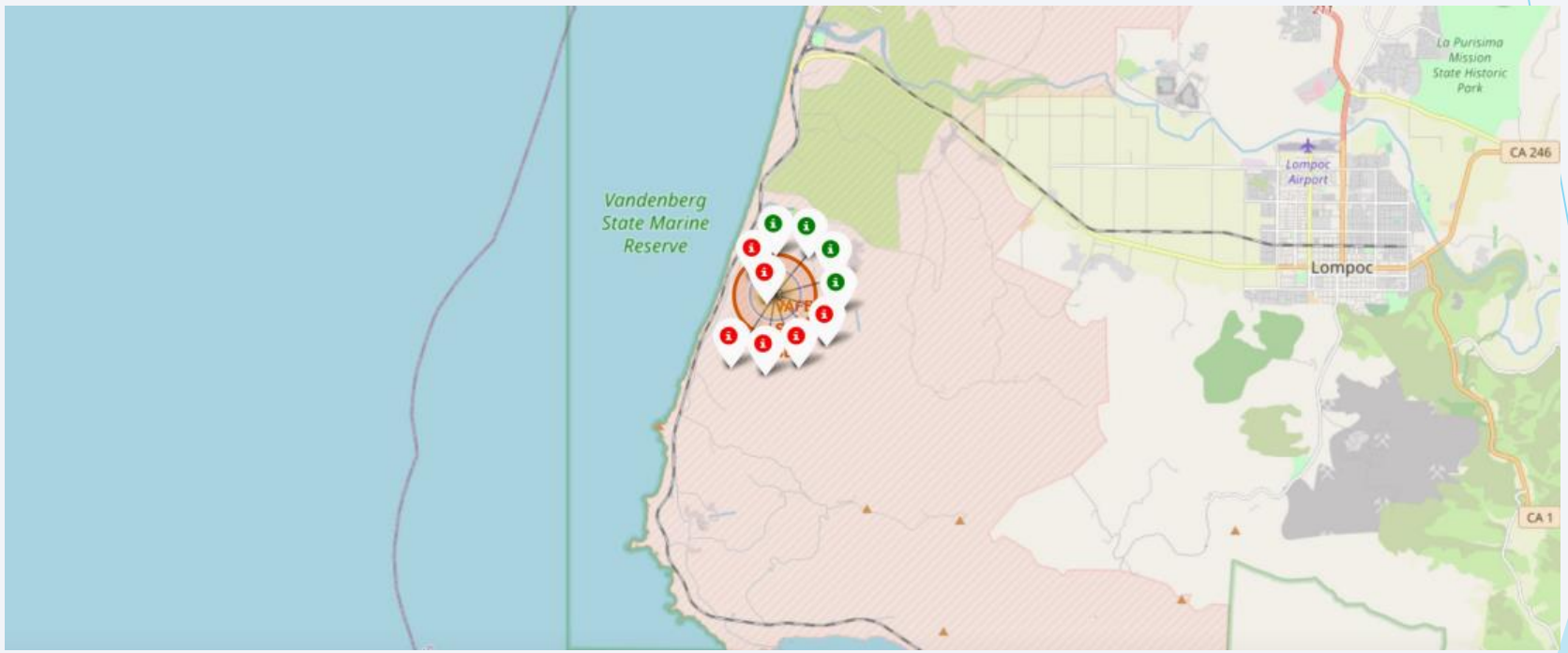

Launch Sites
Proximities Analysis

All launch sites' location markers on a global map



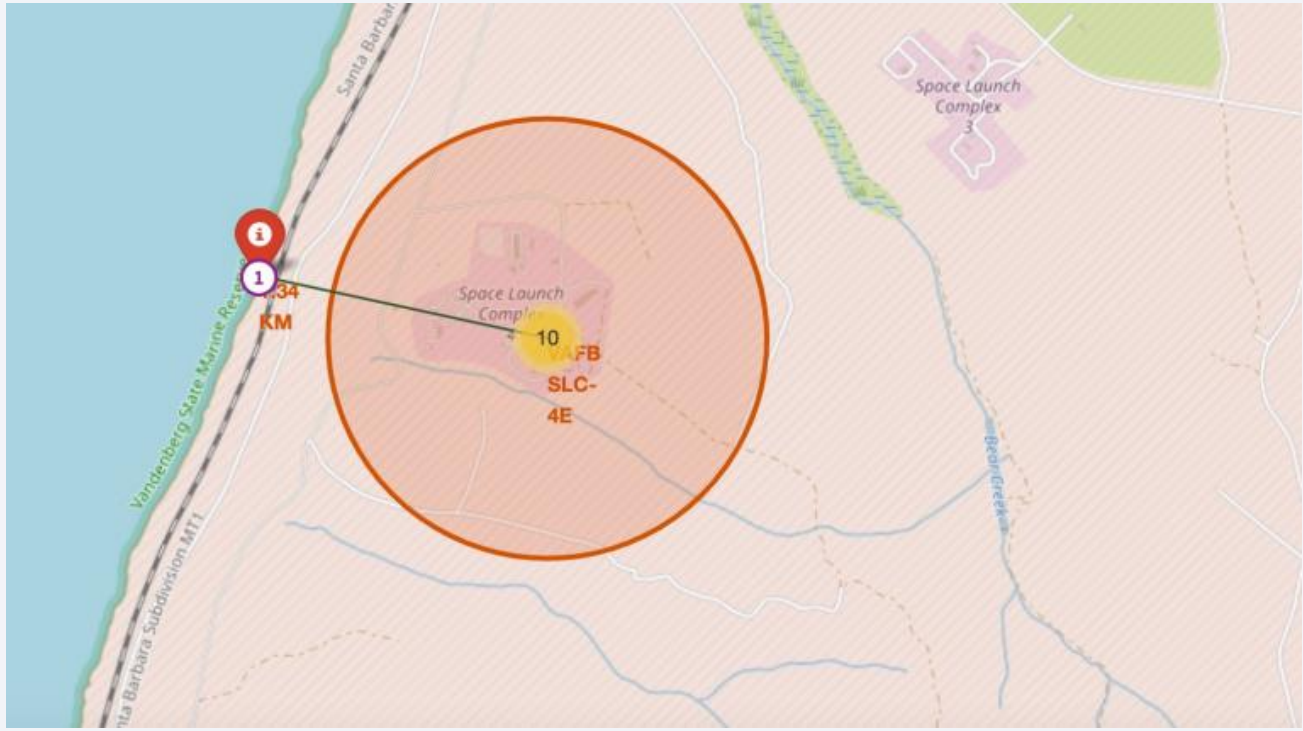
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people

Colour-labeled launch records on the map

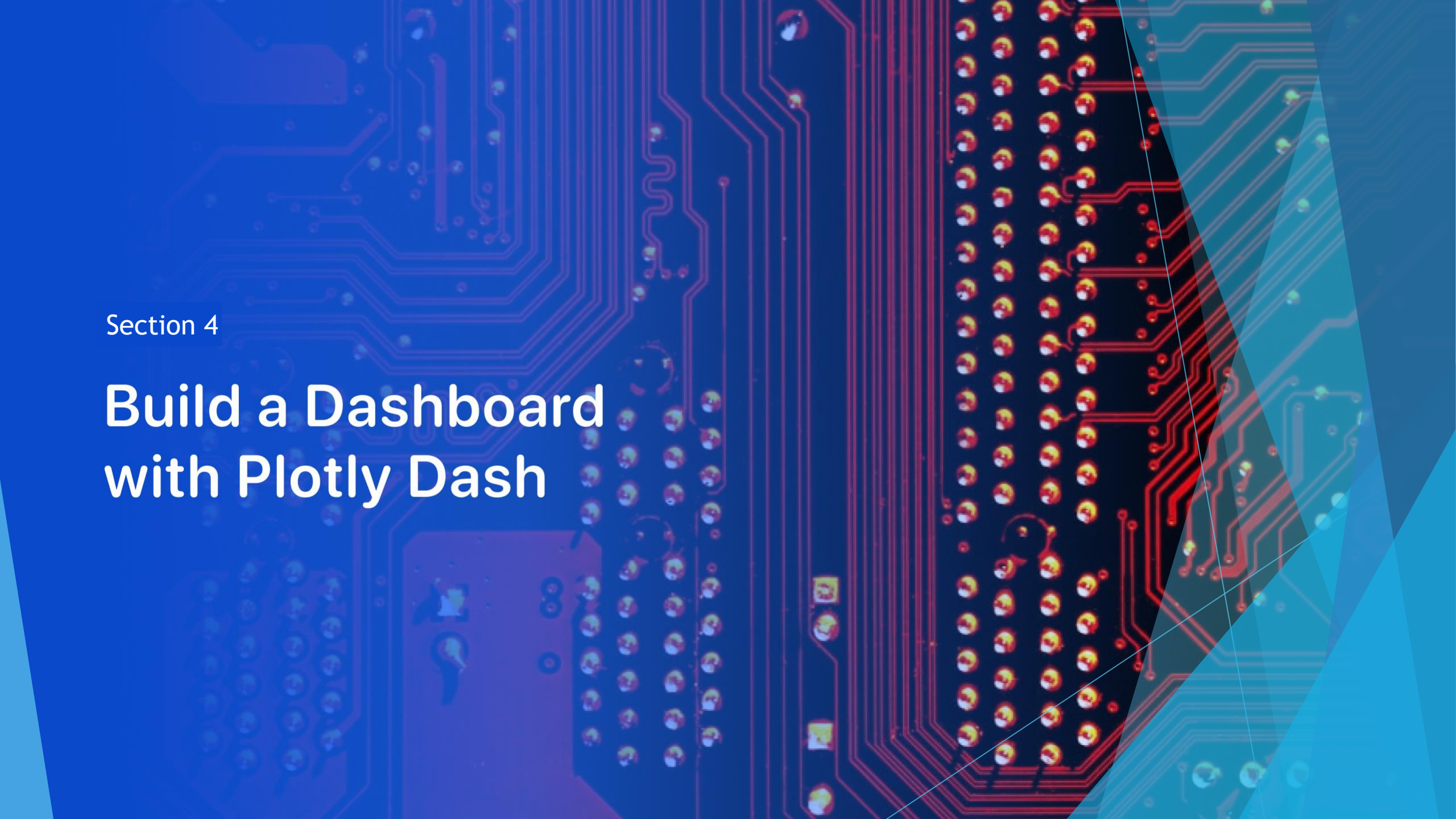


- If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch

Distance from the launch site VAFB SLC-4E to its proximities

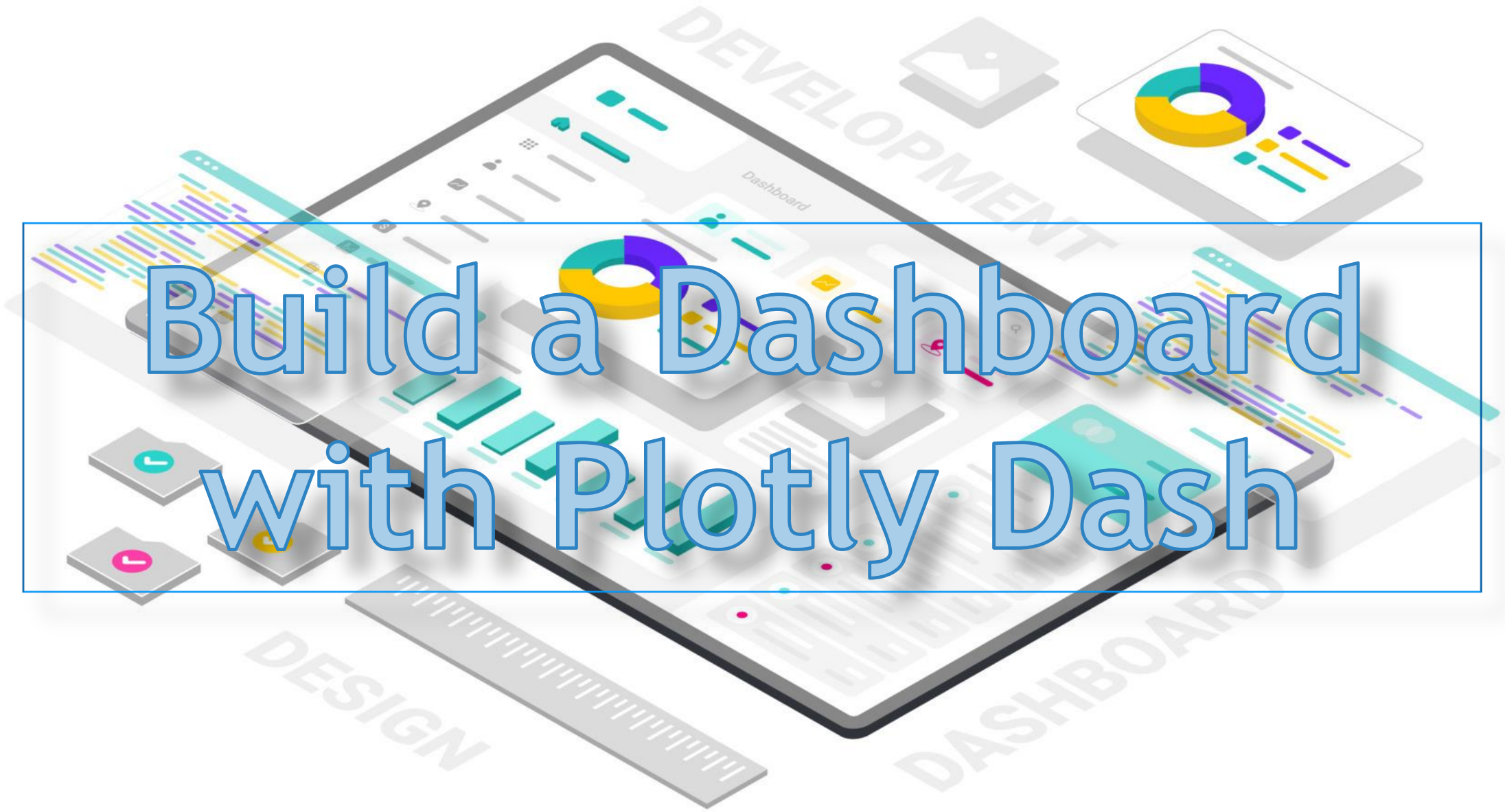


- The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



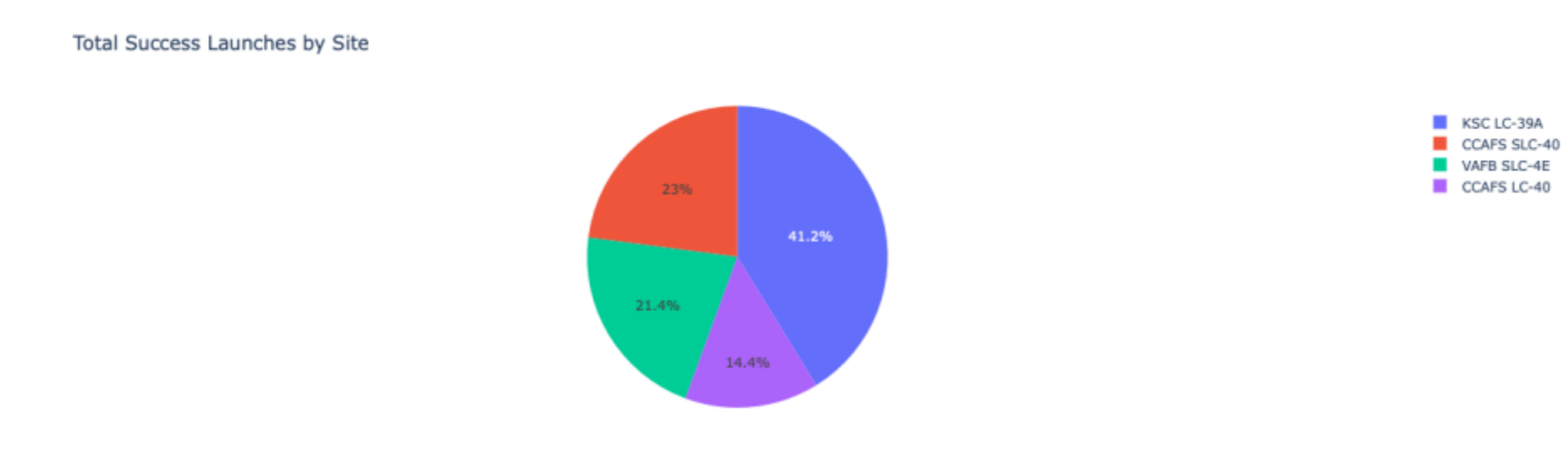
Section 4

Build a Dashboard with Plotly Dash



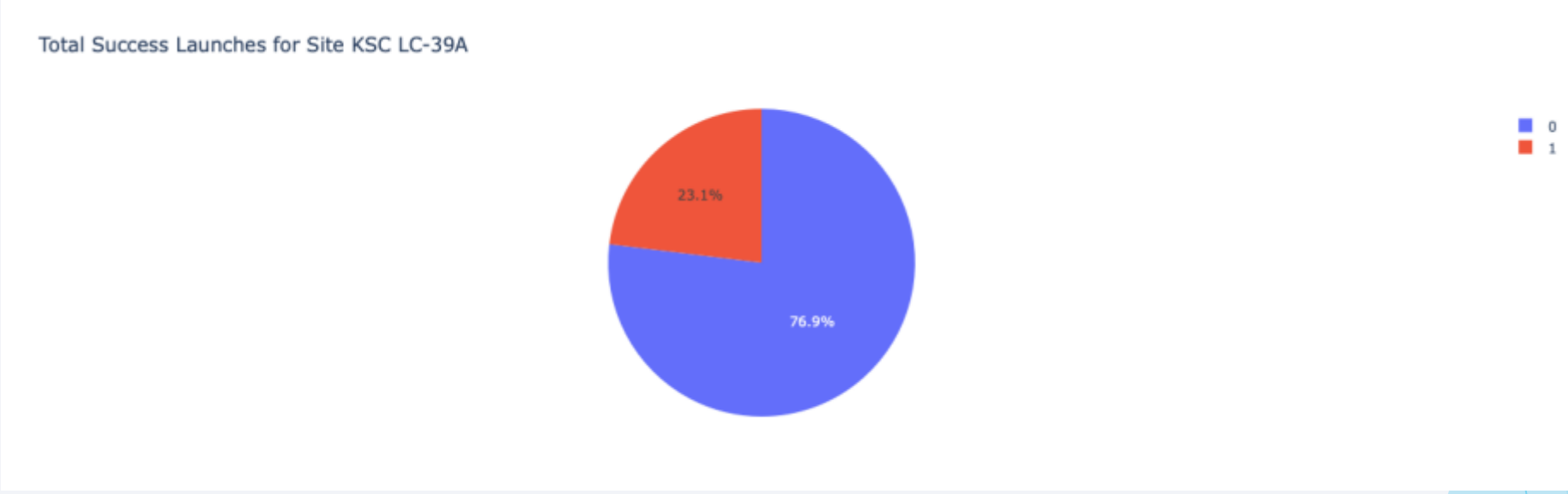
Build a Dashboard with Plotly Dash

Launch success count for all sites



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

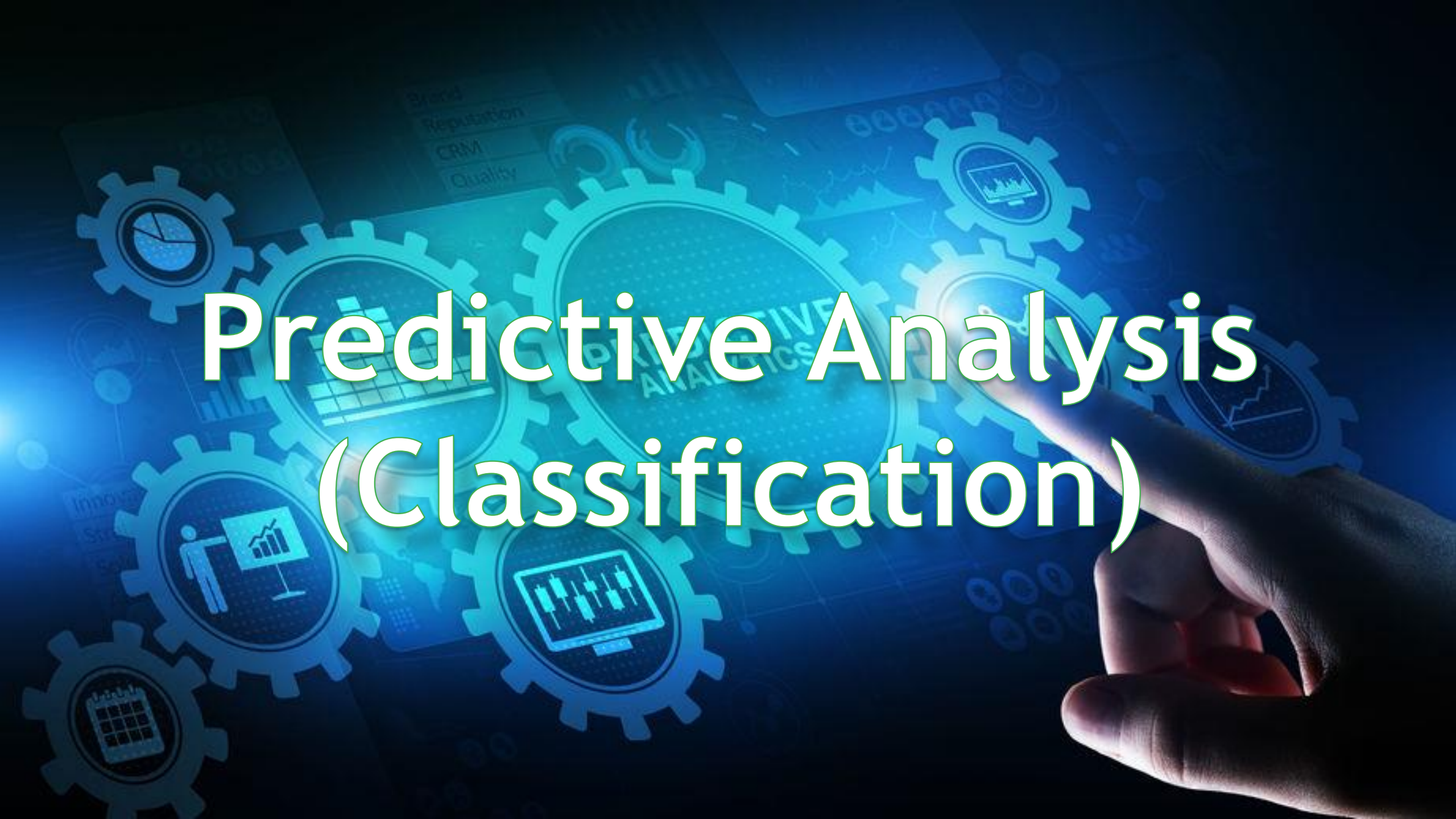
Launch site with highest launch success ratip



- The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline

Payload Mass vs. Launch Outcome for all sites

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

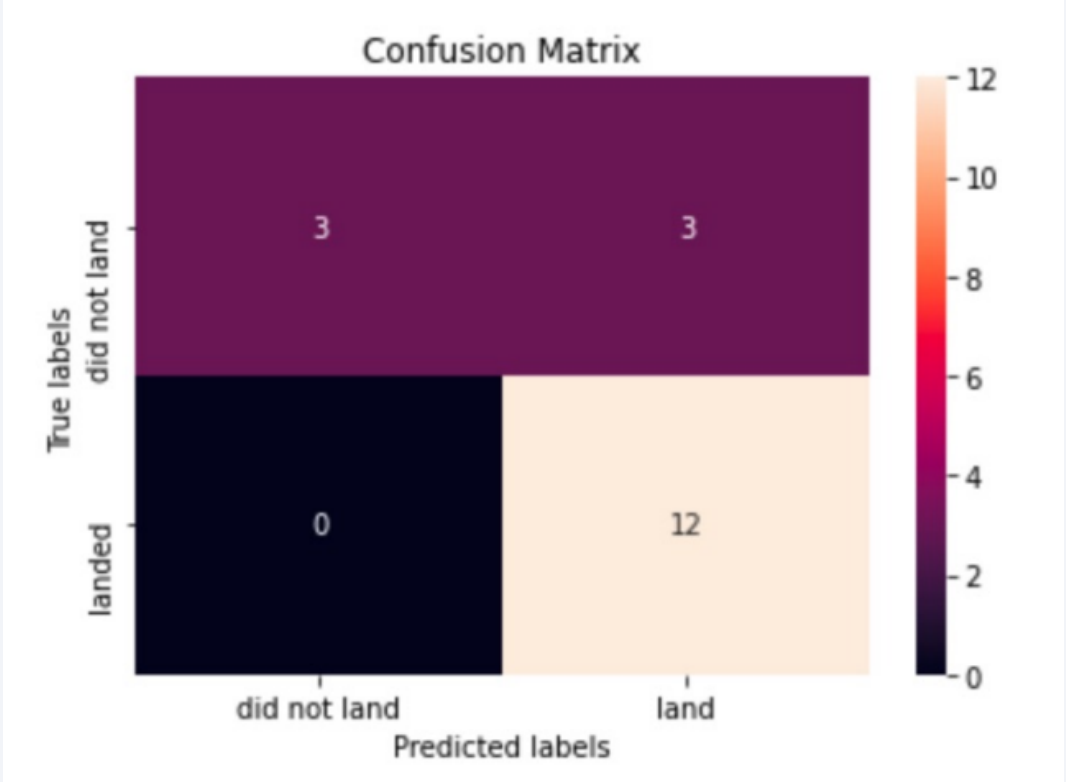

Predictive Analysis
(Classification)

Classification Accuracy

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.889286
KNN	0.848214

- It is found that Decision tree obtained the best score for accuracy.

Confusion Matrix



- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.