IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nam Nguyen
16/03/2023

Outline

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

Executive Summary

Summary of Methodologies:
Data was gathered online using the SpaceX API and web-scraping. The data was then explored using SQL and matplotlib/seaborn/folium. The methods shows that the successful landing of a falcon 9 rocket stage 1 depends on a lot of factors such as Payload Mass, Orbits, Flight Numbers and different classification models can be used to accurately predict the outcome of future launches

Summary of Results:
There were 4 SpaceX Launch Sites, all of which were located near coastlines on either side of the US. The most successful launch sites have been CCAFS SLC-40 with 42.9% success rate. Over the years, SpaceX has massively improved the success rate of their missions, from 0% during the first 2 years, increasing to around 80% in 2020. This shows great promise of consistent rocket landing and reuse.

Introduction

- Project background

Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch, which could be beneficial information to a SpaceX competitor company.

- Project objectives

The aim of this project is therefore to analyze records of past launches by SpaceX and their outcome, in order to predict outcome of future launches based on different features such as launch site, payload mass, orbit, etc..

Section 1

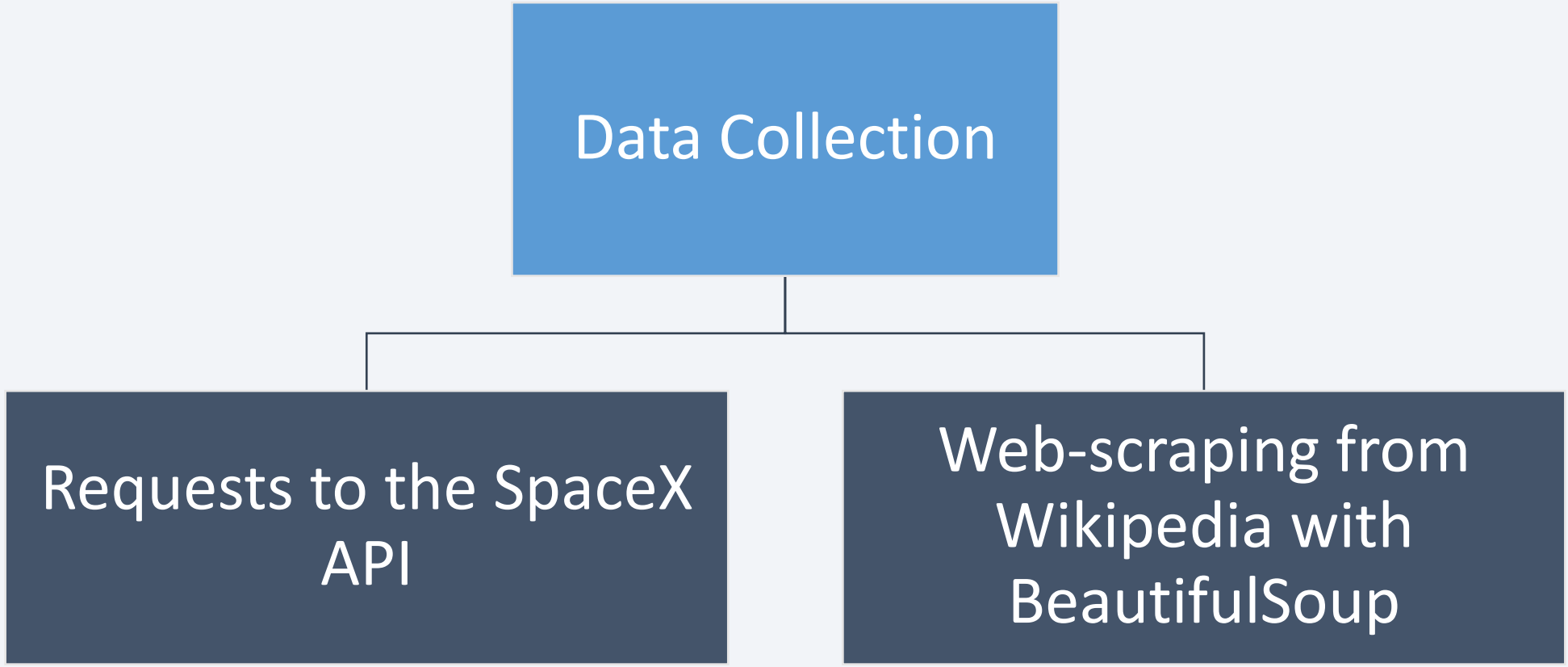
Methodology

Methodology Overview

- Data collection:
  - Data was collected by using requests to the SpaceX API, along with webscraping from Wikipedia with BeautifulSoup
- Data wrangling:
  - Missing data replacement, one-hot encoding, etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models and parameters were tested for best prediction performance.

Data Collection

There are two main methods of collecting data:



Data Collection – SpaceX API

IDs dataset acquired using the requests library and pd.json_normalize() function

Apply pre-defined functions to call on the SpaceX API for every acquired ID value

Dataset with actual values are constructed from the returned values with the given ID, i.e., actual Booster Versions, Launch Site names etc.

Notebook permalink for reference:

https://github.com/andersonpac/ibm_capstoneproject/blob/e2144174d08d4c924c0ed61e0b9623f7d08bb8e2/jupyter-labs1-spacex-data-collection-api.ipynb

Data Collection - Scraping

Get html text of Falcon9 launch Wiki page by using requests.get() and parsing with BeautifulSoup to a soup object

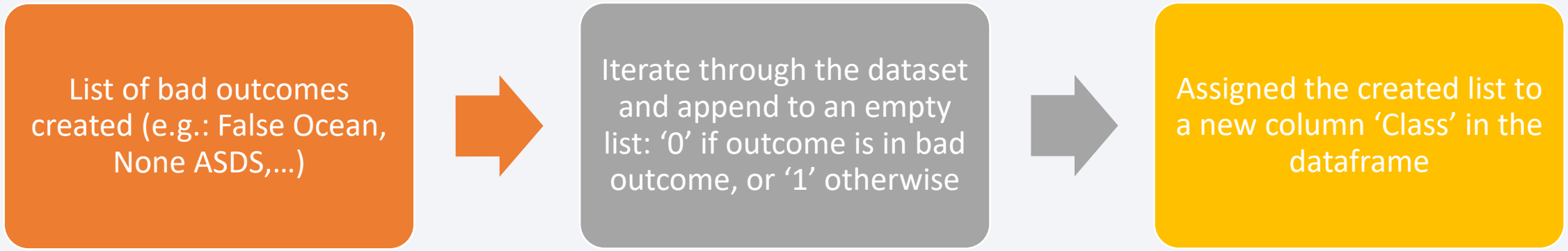
Extract all column names using soup.find_all(), applied to the tag ‘th’

Iterate through the table and add value to the empty columns to create the desired dataframe

Notebook permalink for reference:
https://github.com/andersonpac/ibm_capstoneproject/blob/e2144174d08d4c924c0ed61e0b9623f7d08bb8e2/jupyter-labs2-webscraping.ipynb

Data Wrangling

- Functions such as ‘df.isnull()’ and ‘df.dtypes’ were used on the dataset to determine missing values and value type of each column
- A new column to categorize good vs bad outcome from each mission is then created based on the data from the original landing outcome column:



Notebook permalink:
https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/jupyter-labs3-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

List of plots:

- Scatter plots: to determine if there are any relationship between two variables and the launch outcome (each datapoint is colored based on mission success/failure):
  - Payload Mass vs Flight Number
  - Launch Site vs Flight Number
  - Launch Site vs Payload Mass
  - Orbit vs Flight Number

- Bar graph: Class vs Orbit: to visualize the average mission success rate for each type of orbit.
- Line graph: Class vs Year: to track the yearly mission success rate over a period of time.

Notebook permalink:
https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/jupyter-labs5-eda-dataviz.ipynb

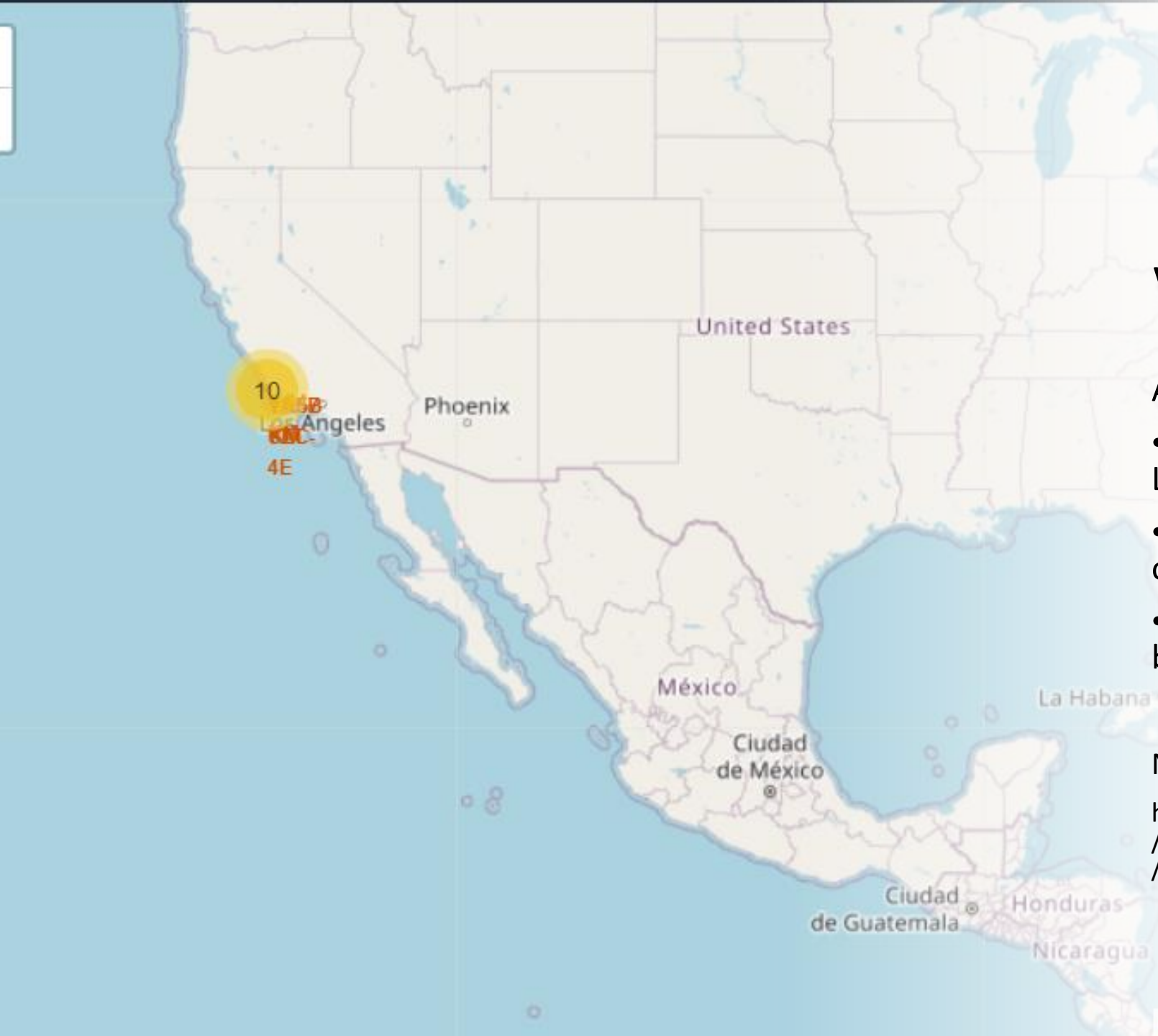
EDA with SQL

Summary of SQL queries performed:

- Determined unique launch sites
- Displayed 5 records where the launch site code name started with ‘CCA’
- Displayed the total payload mass carried by NASA (CRS)
- Determined average payload mass carried by booster version F9 V1.1
- Determined the date when successful landing on a ground pad was first achieved
- Showed Boosters with payload mass between 4000 and 6000kg which landed successfully on a drone ship
- Listed the total number of successful and unsuccessful mission outcome
- Showed booster versions that have carried the maximum payload
- Show month record for failed drone ship missions in 2015
- Ranked the count of different successful outcomes from 2010 to 2017

Notebook permalink:

https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/jupyter-labs4-eda-sql-coursera_sqllite.ipynb



Build an Interactive Map with Folium

Added objects to the map include:

- Circle/Text Marker displaying each Launch Site
- Marker Clusters displaying mission outcome on each Site
- Polylines/Text Marker displaying distance between launch site and nearby utilities

Notebook permalink:

https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/jupyter-labs6-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

Main elements of the Dash application:

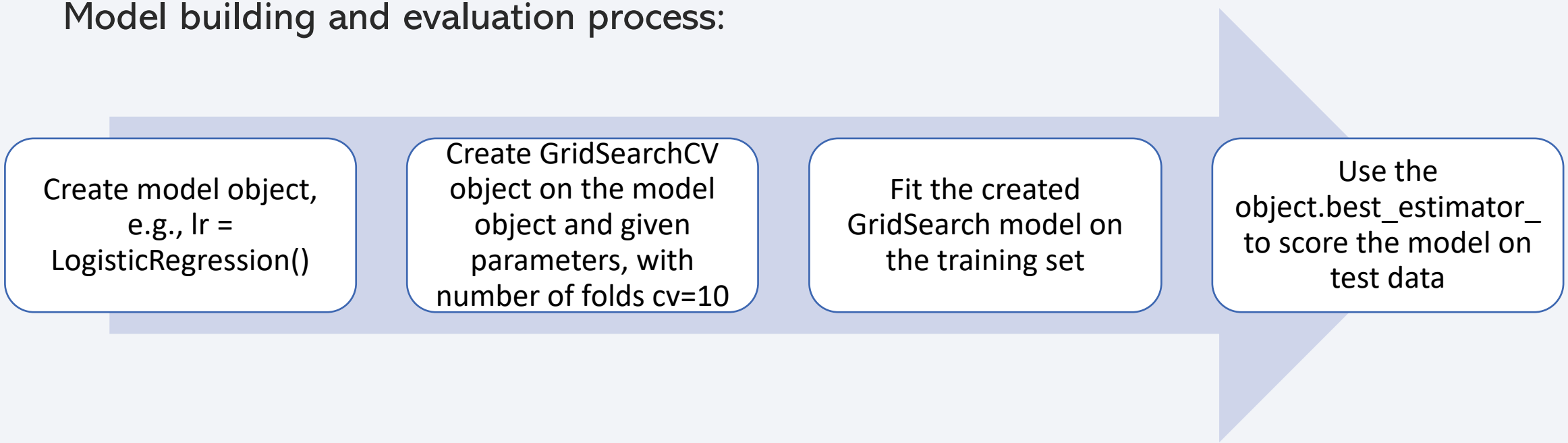
- Launch Site Dropdown: this was added because it is helpful to be able to choose which Launch Site to display data for.
- Success rate pie chart: added to display success rate for all or individual launch sites
- Payload Mass Range Slider: added to choose the range of payload mass over which the user would like to see success data for
- Class vs Payload Mass scatter plot, color-coded with booster version name: displays the Successful (class 1) or Failure (class 0) outcomes of each mission along the chosen range of payload mass.

Application file permalink:

https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/spacex_dash_app.py

Predictive Analysis (Classification)

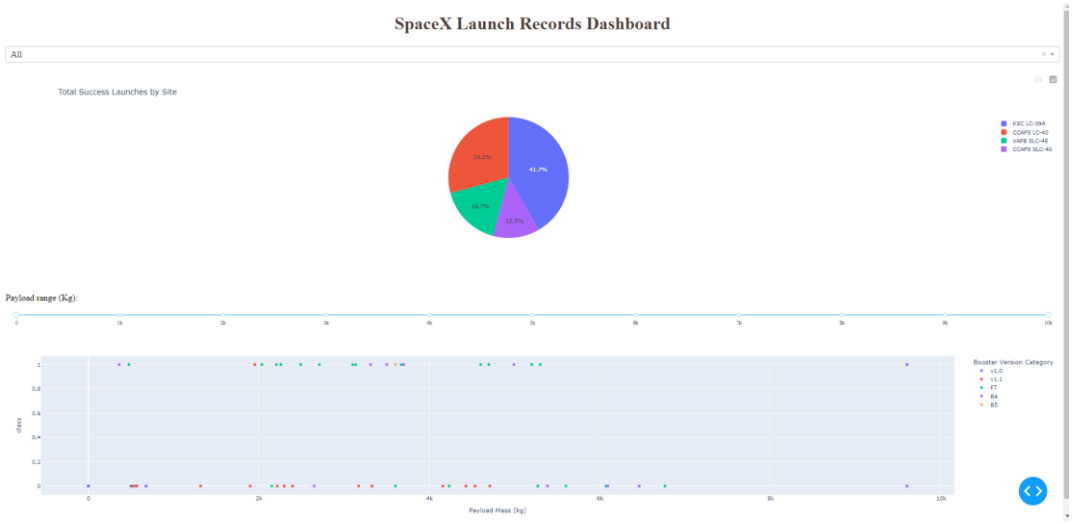
Model building and evaluation process:

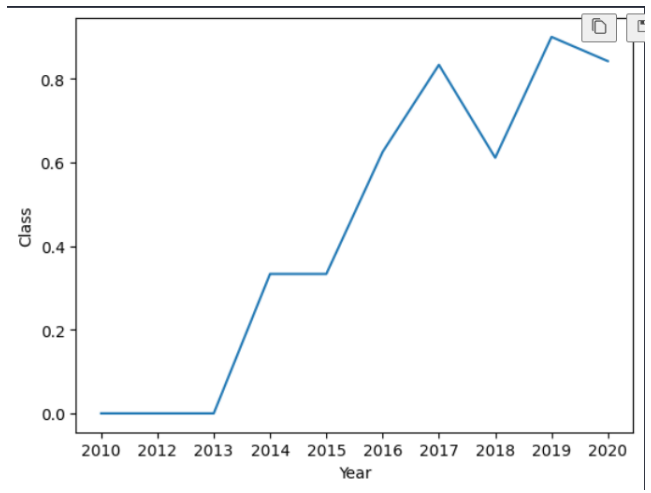


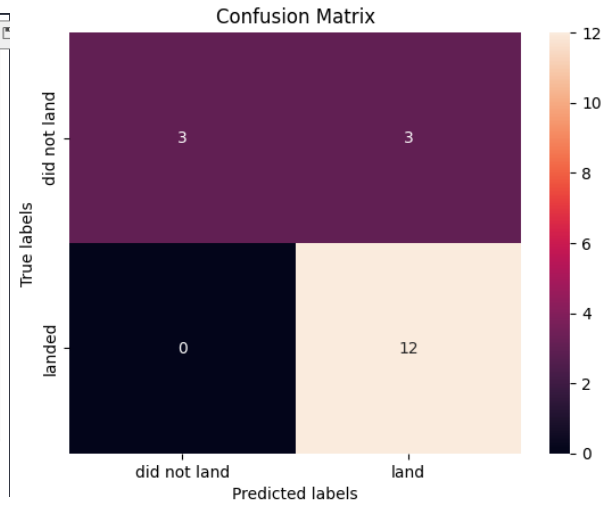
This process was applied to several classification models to determine which one performed the best, for example Logistic Regression, K-nearest Neighbors, etc. For code file please refer to this link:

https://github.com/andersonpac/ibm_capstoneproject/blob/e848a7f77059851a59298974c3f4948d85527099/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

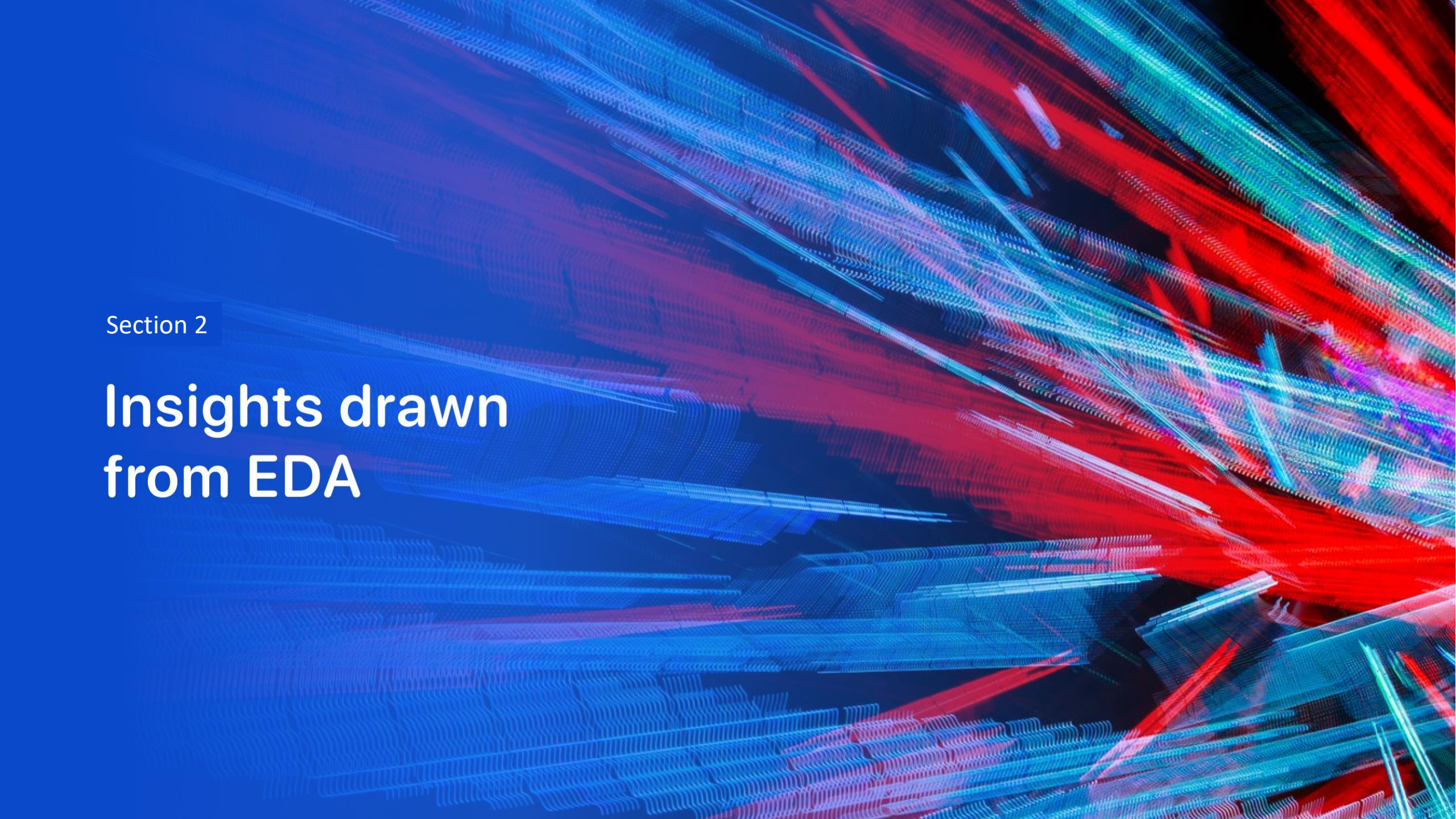
Results







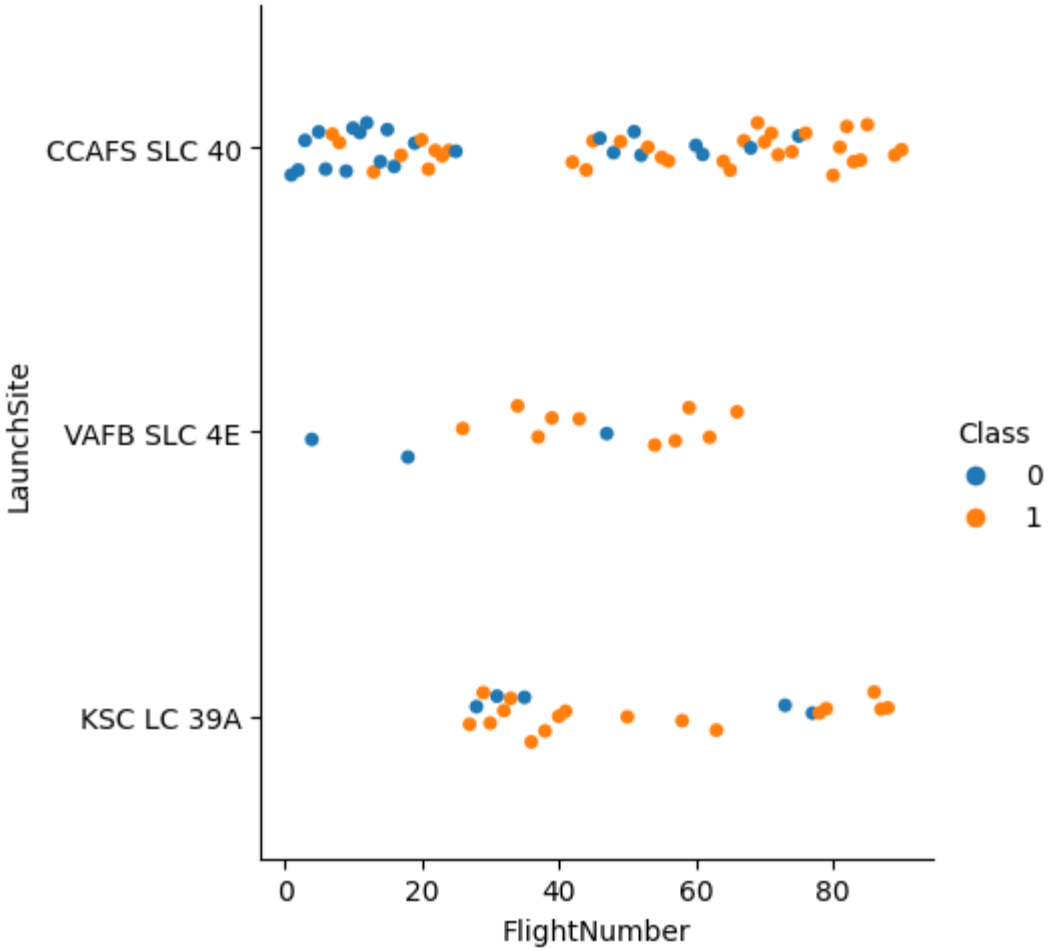
From left to right: Dash Application, Launch Success rate over the years, and confusion matrix by predictive analysis


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

This figure displays launch site vs flight number, color coded with the mission outcome (0: unsuccessful and 1: successful). Some key observations:

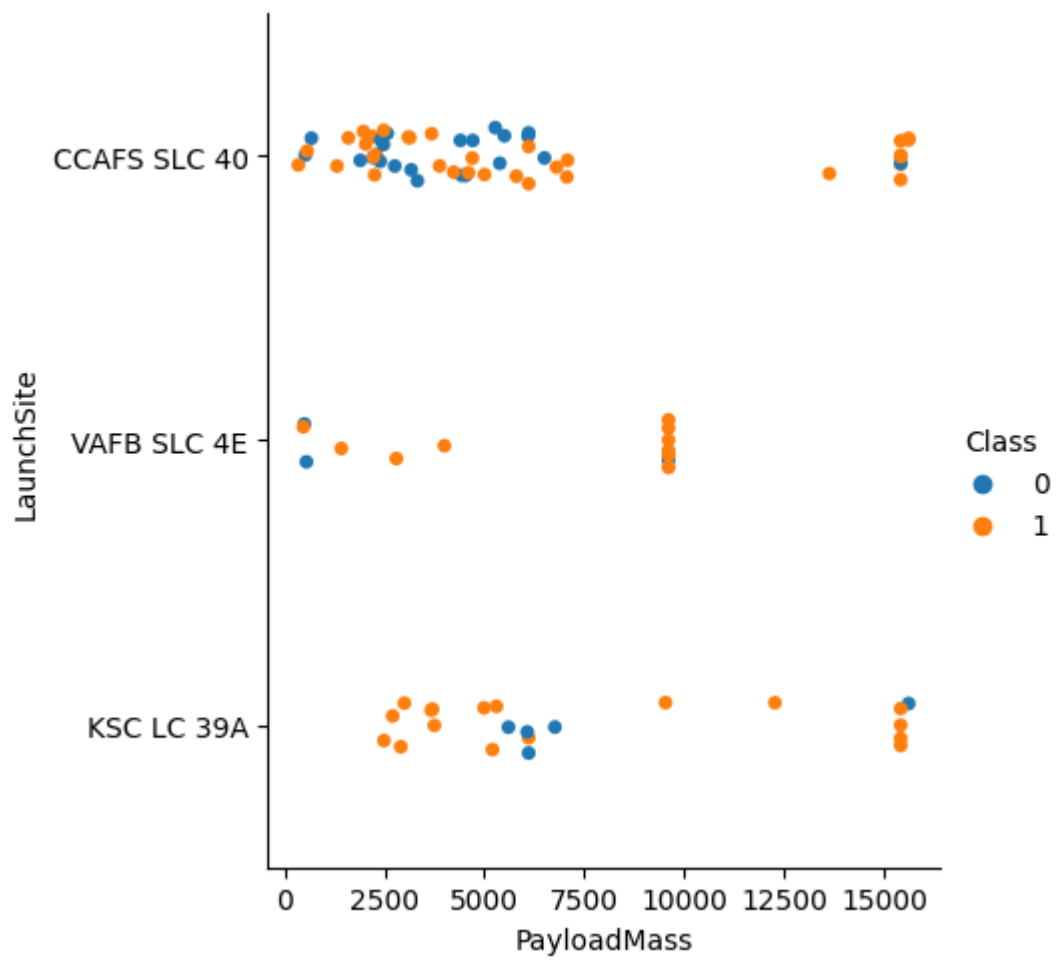
- Most earlier launches (flight no. < 25) took place at CCAFS SLC 40
- Larger flight numbers are more successful



Payload vs. Launch Site

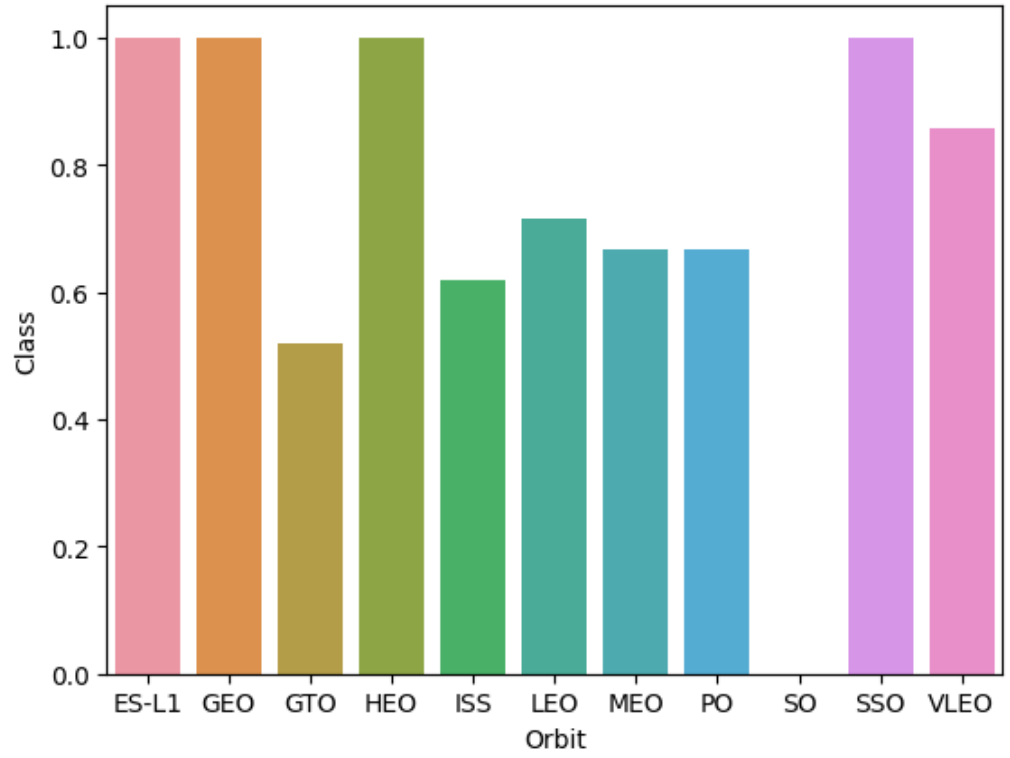
This figure displays Launch Site vs Payload Mass, color coded with the mission outcome (0: unsuccessful and 1: successful). Some key observations:

- All sites have a cap on payload mass e.g., 10000 kg for VAFB SLC 4E



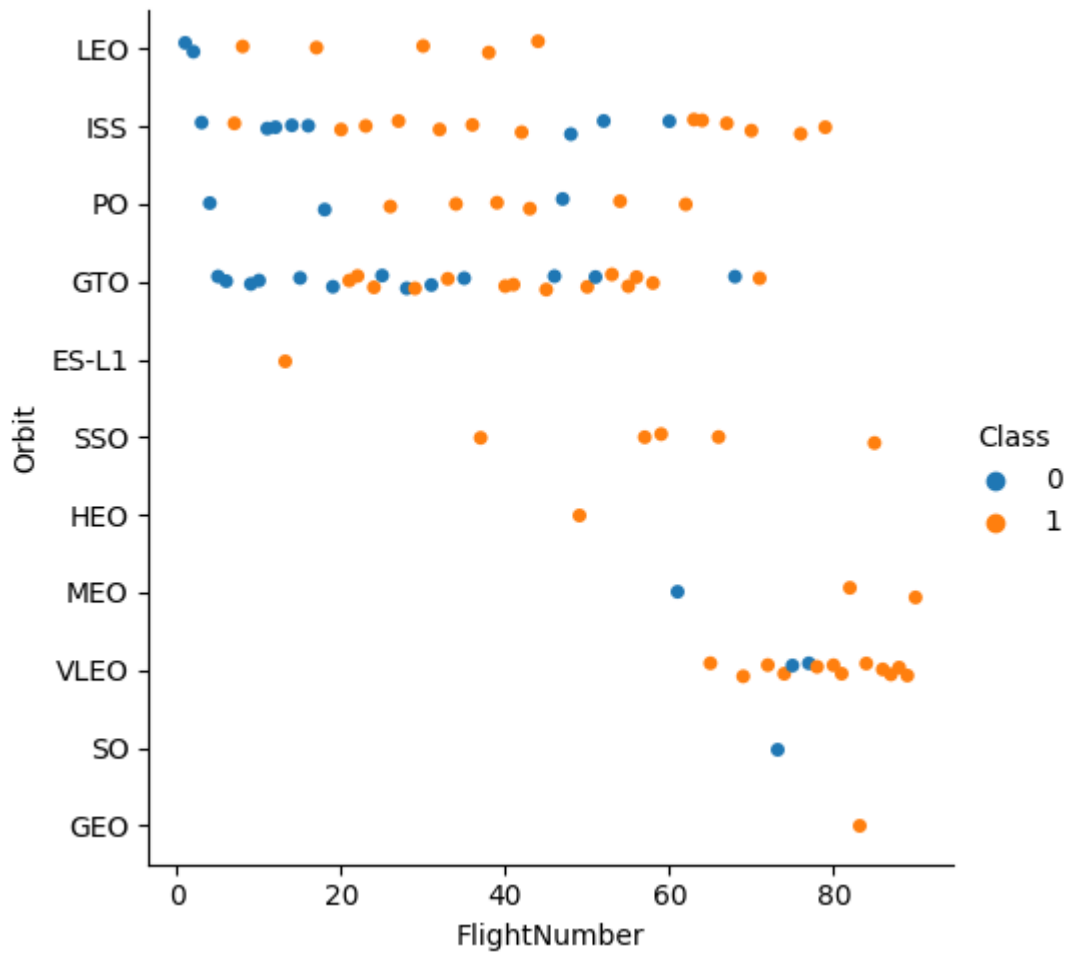
Success Rate vs. Orbit Type

This chart shows the average success rate for each of the orbit type that the launch will follow in the mission. As can be seen, missions on ES-L1, GEO, HEO and SSO orbits are totally successful while missions on SO orbits have never worked out.



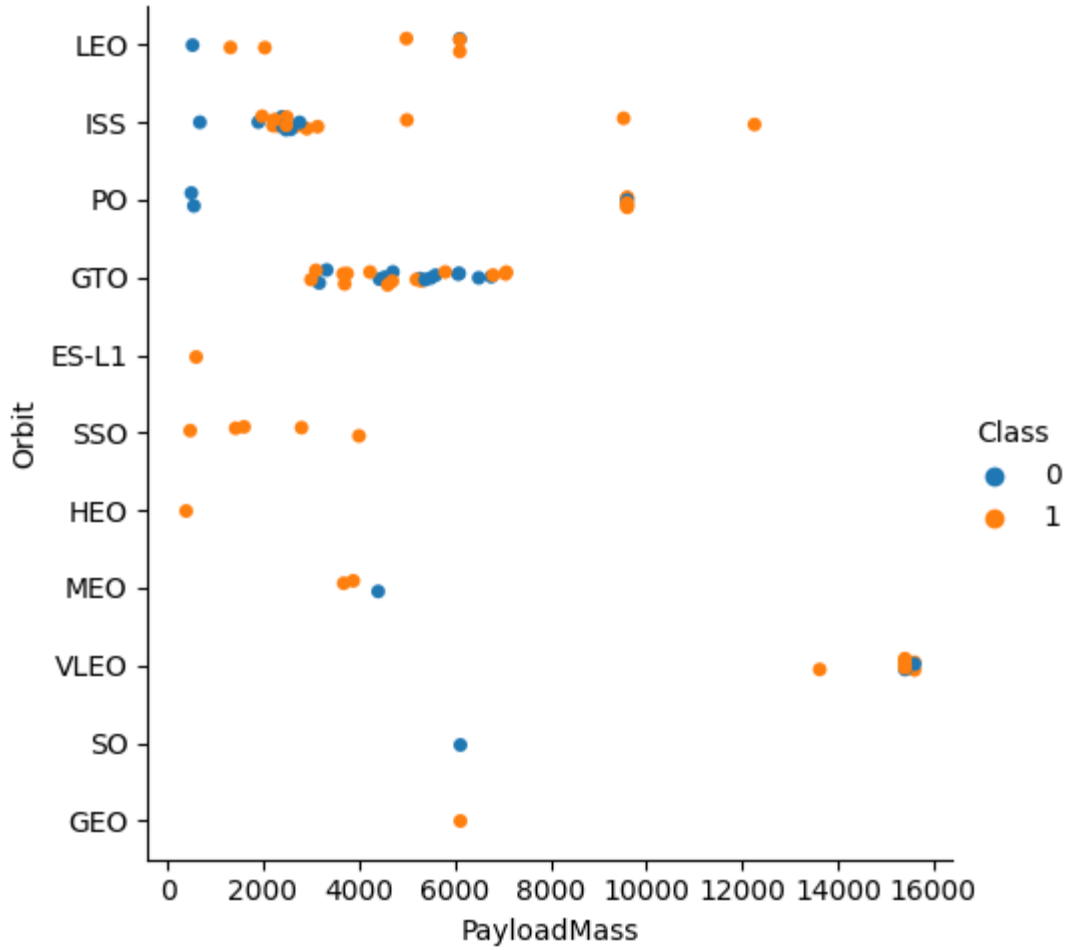
Flight Number vs. Orbit Type

This plot shows the relationship between the mission's Orbit and Flight Number. As we can see, most recent flight numbers are on the VLEO and ISS orbits.



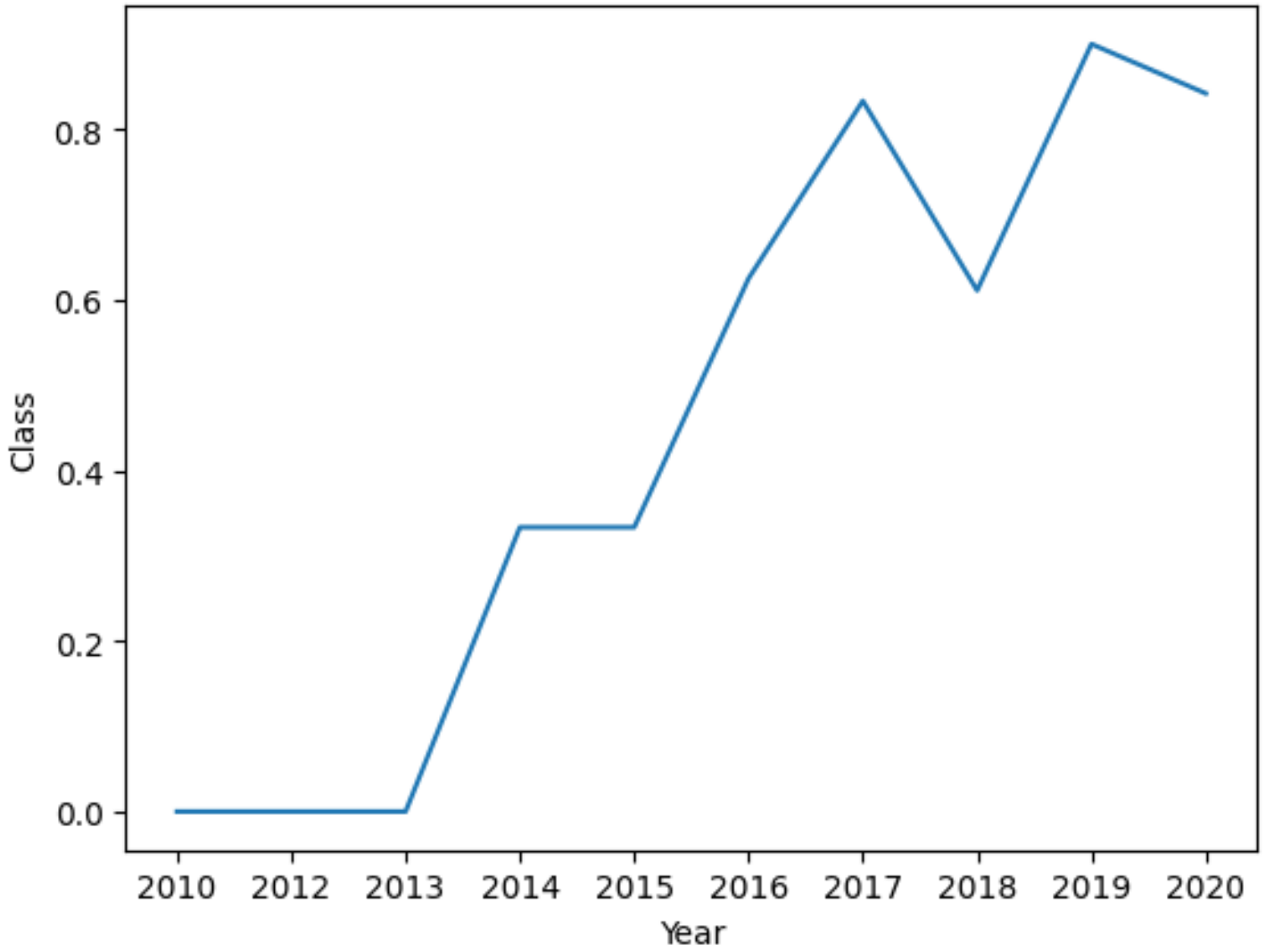
Payload vs. Orbit Type

This figure plots the Orbit against Payload Mass. We can see that payloads larger than around 13000 kg are only ever launched into the VLEO orbit. The ISS has the widest range of payload mass



Launch Success Yearly Trend

This line chart shows the yearly success rate of Falcon 9 launches from 2010 up to 2020. The general trend is evidently an increase in success, with the first few years being complete failure, but then steadily increases to around 0.8 in 2020



All Launch Site Names

- 4 Launch Sites were used throughout all launches
- The names of the sites recorded in the database were code names

```
%%sql
SELECT DISTINCT Launch_site FROM spacextbl;
```

[6]

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

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

Launch Site Names Begin with 'CCA'

The query returned the first five records of launches from site CCAFS LC-40 since the site name contained ‘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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by NASA (CRS) was found to be 45596kg

```
%%sql
SELECT SUM(payload_mass__kg_) as Total_Payload_kg, customer FROM spacextbl
WHERE customer = 'NASA (CRS)'
```

[8]

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

Total_Payload_kg	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 were calculated to be around 2534.67 kg across all past launch records of the booster version

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

```
%%sql
SELECT booster_version, AVG(payload_mass__kg_) as AVG_Payload FROM spacextbl
WHERE booster_version LIKE '%F9 v1.1%'
```

[9]

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

Booster_Version	AVG_Payload
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

- The query returned 22 December 2015 as the first date when a successful ground pad landing was achieved

```
%%sql

SELECT Date, `Landing _Outcome` FROM spacextbl
WHERE `Landing _Outcome` = 'Success (ground pad)'
ORDER BY substr(Date,7,4), substr(Date,4,2), substr(Date,1,2) DESC LIMIT 1
```

[10]

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

Date	Landing _Outcome
22-12-2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query returned 4 different booster versions which satisfied the constraints

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_, `Landing _Outcome` FROM spacextbl
WHERE (`Landing _Outcome` = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

[11]

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

```
%%sql

SELECT COUNT(), `Landing _Outcome` FROM spacextbl
GROUP BY `Landing _Outcome`
```

[12]

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

COUNT()	Landing _Outcome
5	Controlled (ocean)
3	Failure
5	Failure (drone ship)
2	Failure (parachute)
21	No attempt
1	No attempt
1	Precluded (drone ship)
38	Success
14	Success (drone ship)
9	Success (ground pad)
2	Uncontrolled (ocean)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes. A lot of missions were successful as shown.

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_ FROM spacextbl
WHERE PAYLOAD_MASS__KG_ = (SELECT Max(PAYLOAD_MASS__KG_ ) FROM spacextbl)
```

[13]

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

Boosters Carried Maximum Payload

- The query lists the names of the booster which have carried the maximum payload mass, which was found to be 15600kg

2015 Launch Records

The query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, along with the month the mission was carried out. There were only 2 failed drone ship landings in 2015.

```
%%sql
SELECT substr(Date,4,2) AS Month, `Landing _Outcome`, booster_version, launch_site FROM spacextbl
WHERE substr(Date,7,4)='2015' AND `Landing _Outcome` = 'Failure (drone ship)'
```

[14]

```
* 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

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

- This query ranks the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- The most successful landing outcome was found to be drone ship, leading with 14 counts

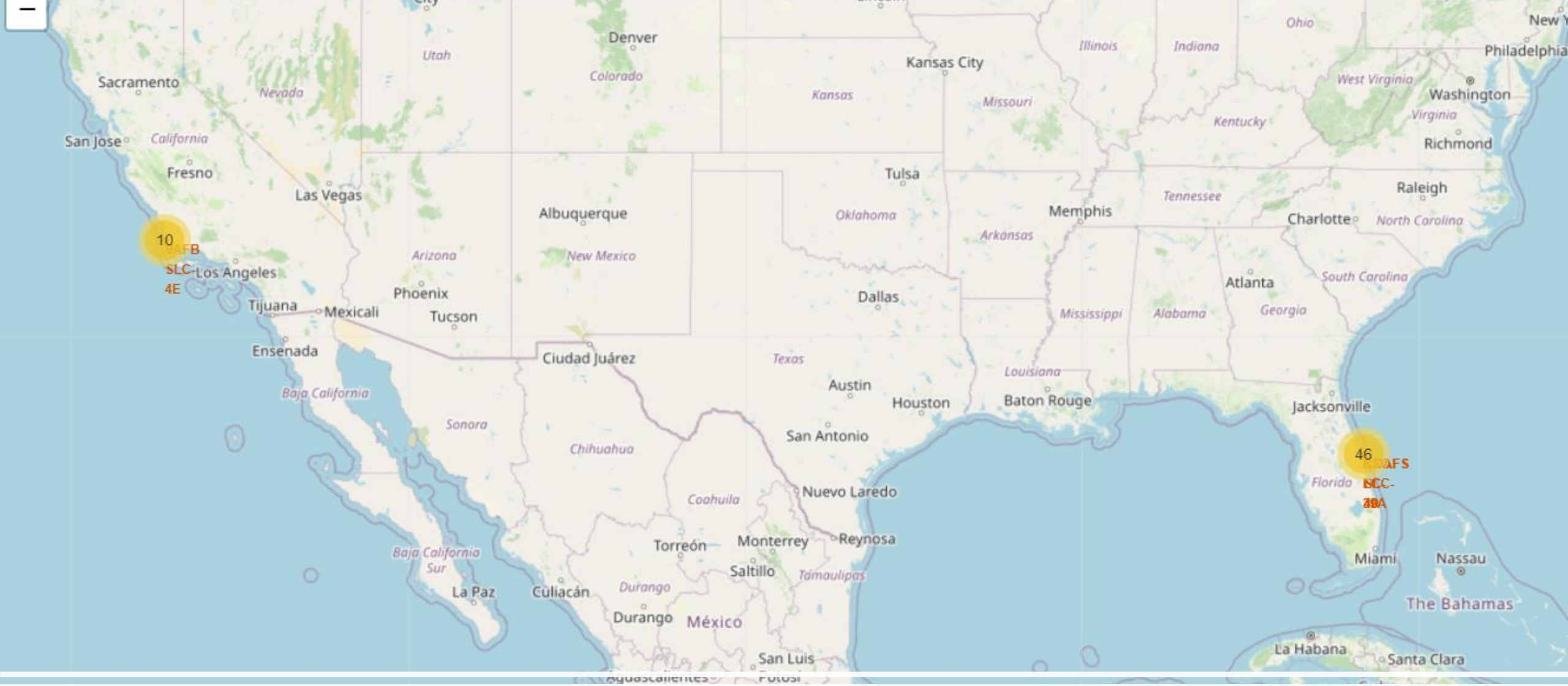
```
%%sql
SELECT `Landing _Outcome`, COUNT(`Landing _Outcome`) AS Count FROM spacextbl
GROUP BY `Landing _Outcome`
HAVING CAST(substr(Date,7,4) AS INT)<2017
    AND ((`Landing _Outcome` LIKE '%Success%')
        OR `Landing _Outcome` LIKE 'Controlled%')
ORDER BY COUNT(`Landing _Outcome`) DESC
```

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

Landing _Outcome	Count
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5

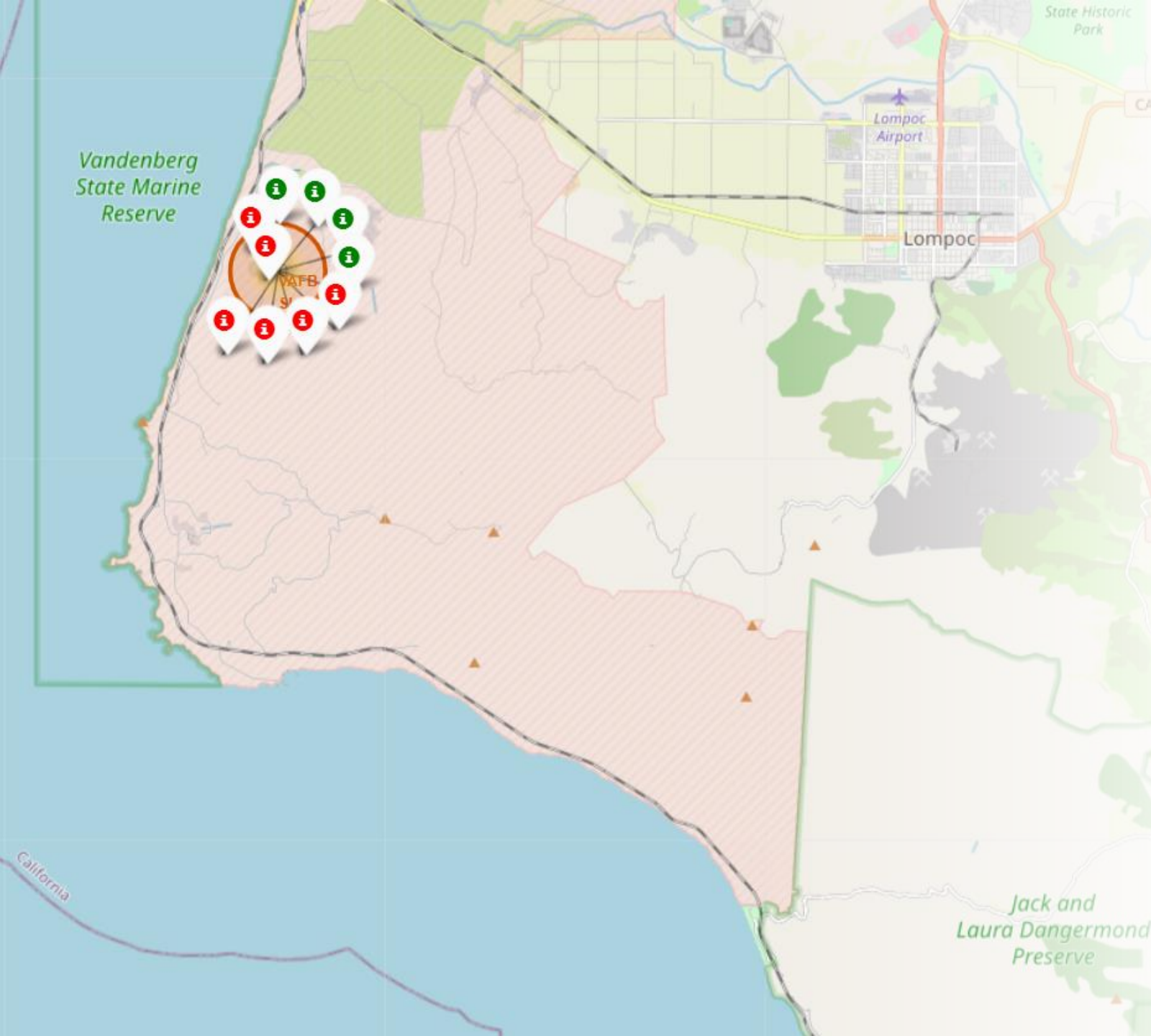
Section 3

Launch Sites Proximities Analysis



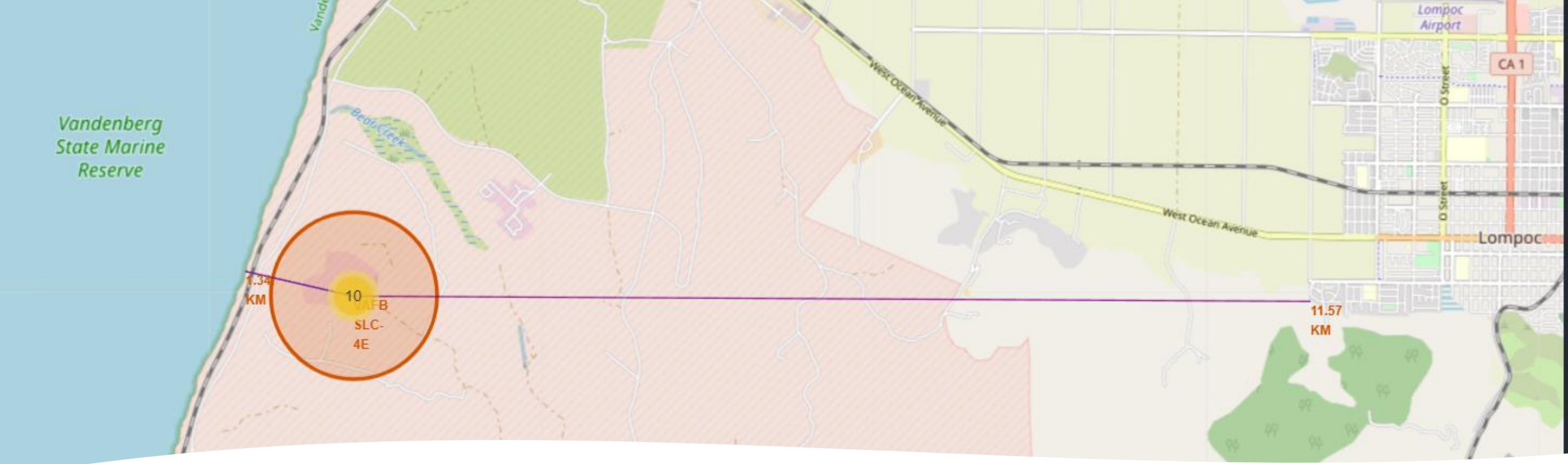
SpaceX Launch Site Locations

SpaceX has 2 main launch locations, one in California and one in Florida, both very close to the coastline.



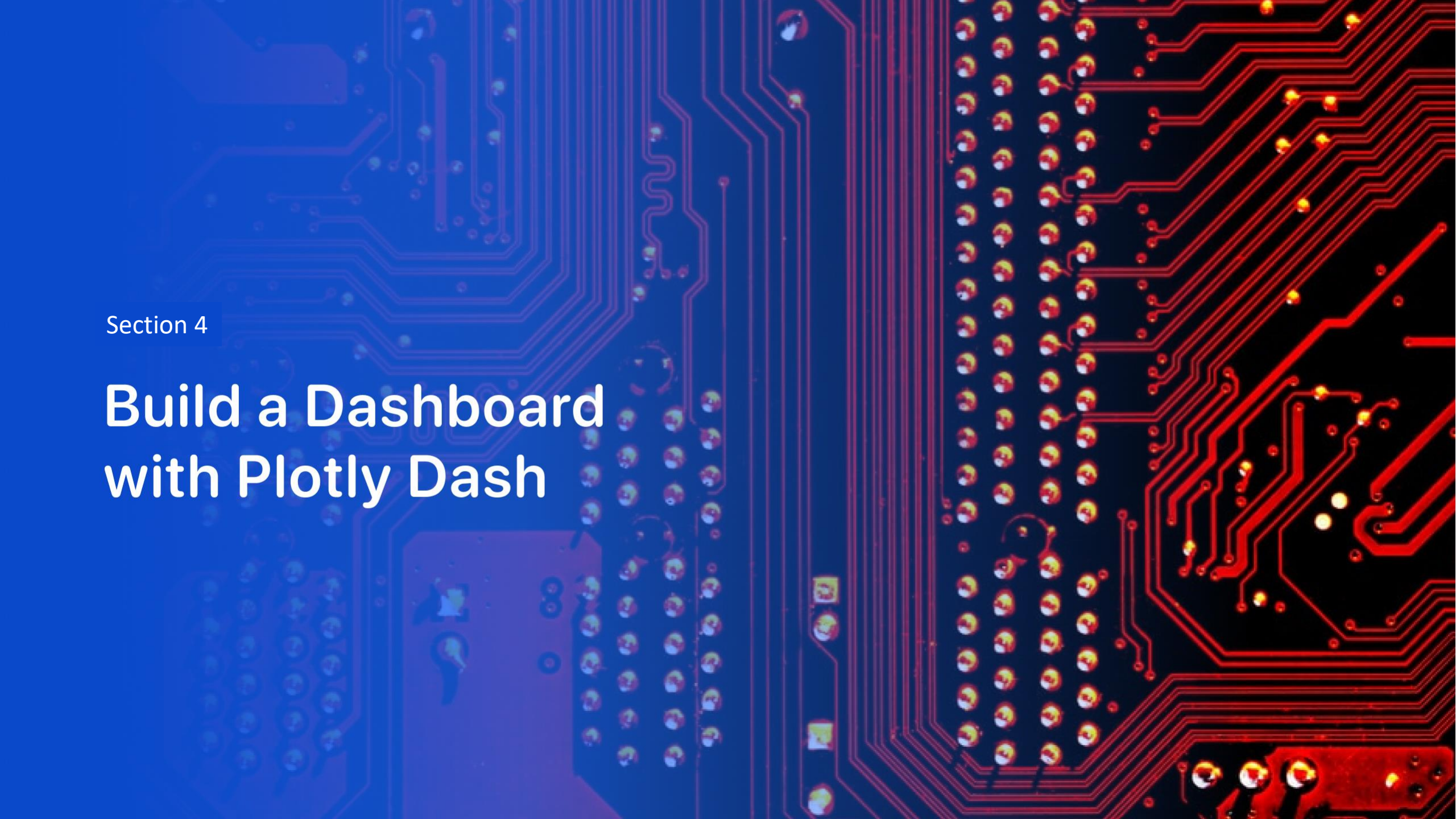
Vandenberg Site with colored-label mission outcomes

- This is the only SpaceX launch site on the west coast of the US
- Only 4 missions from the site have been successful so far

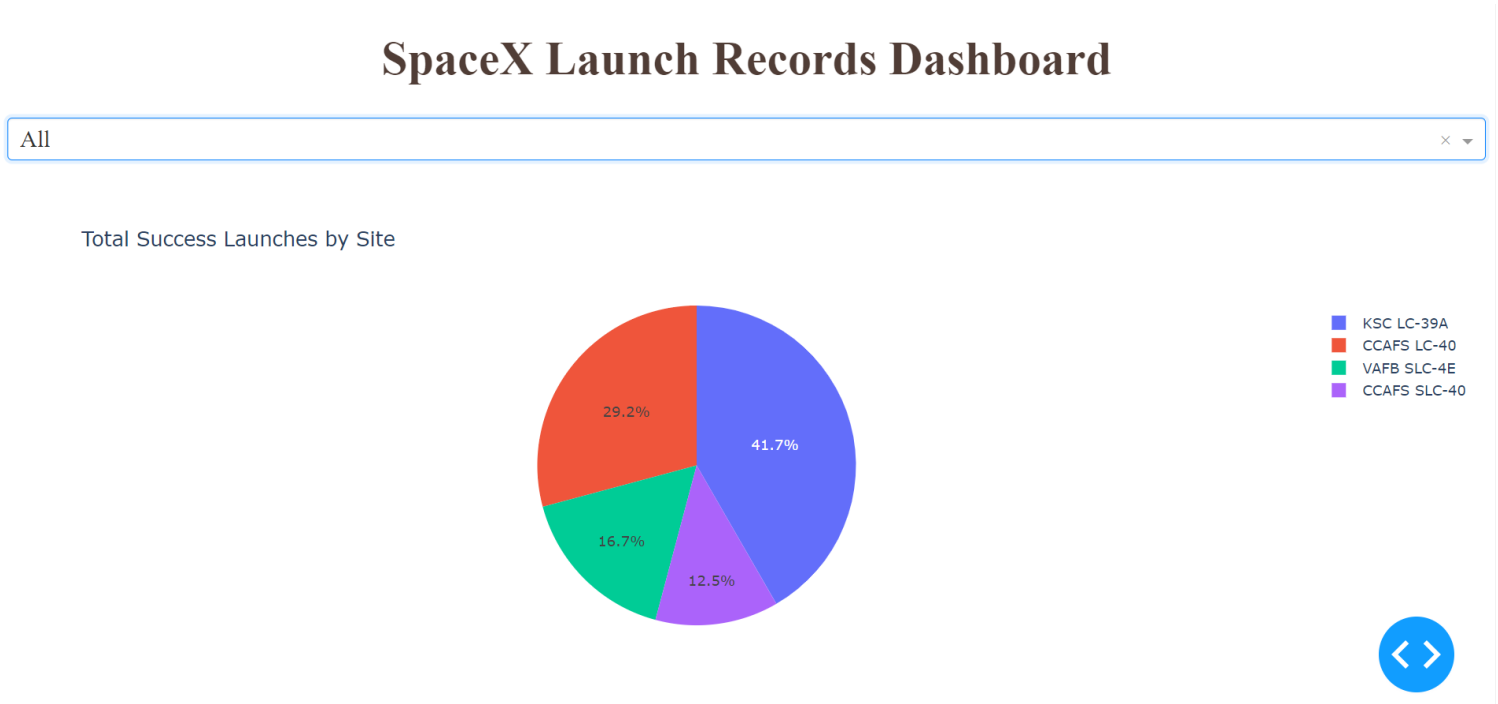


Vandenberg site distance to the nearest coastline and city

- The launch site is located extremely close to the waters (1.34 km) and other utilities such as railways and highways. However, it is very remote from metropolitan areas such as cities – with the closest city being Lompoc – almost 12km away. This is sensible since launching a rocket requires a lot of resources, which could be provided via train or trucks, but it would not be desirable to have the launch take place near a city since it can be extremely dangerous and disruptive.


Section 4
Build a Dashboard with Plotly Dash

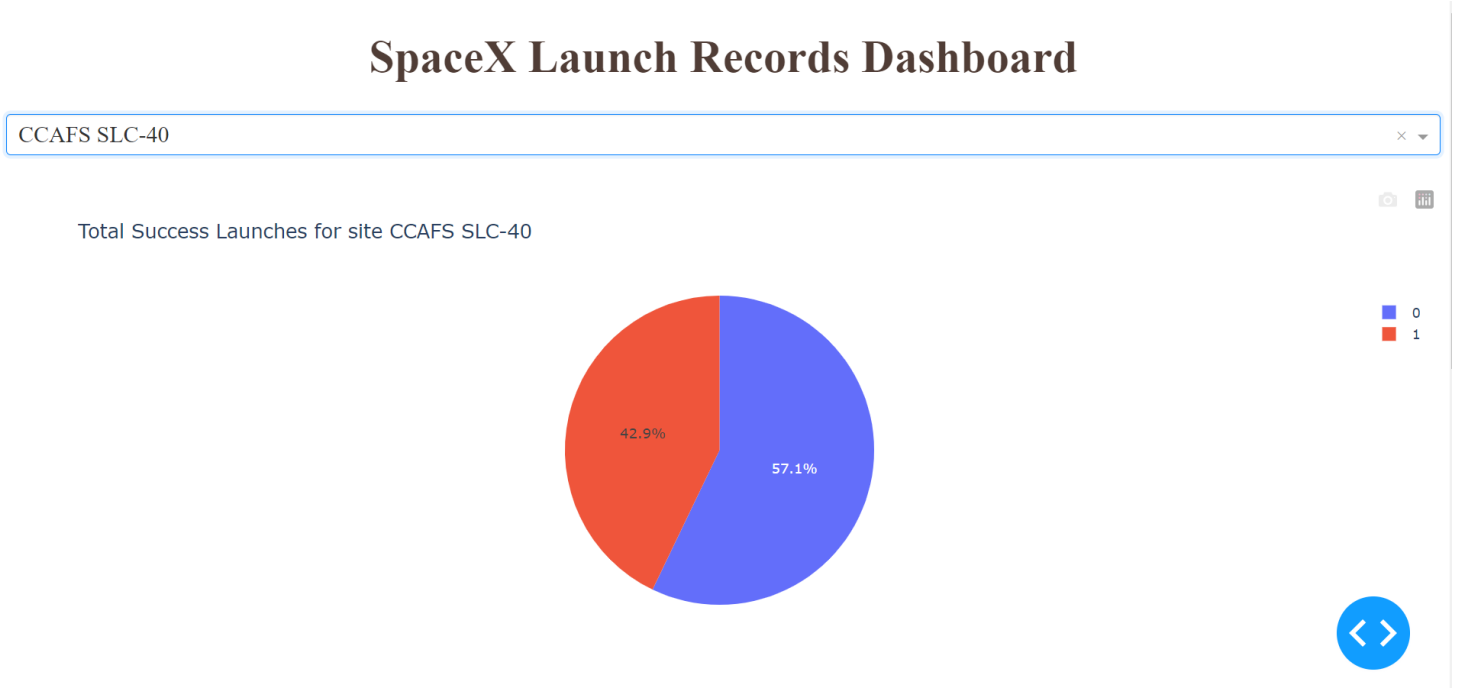
Pie chart showing total successful launches by all site

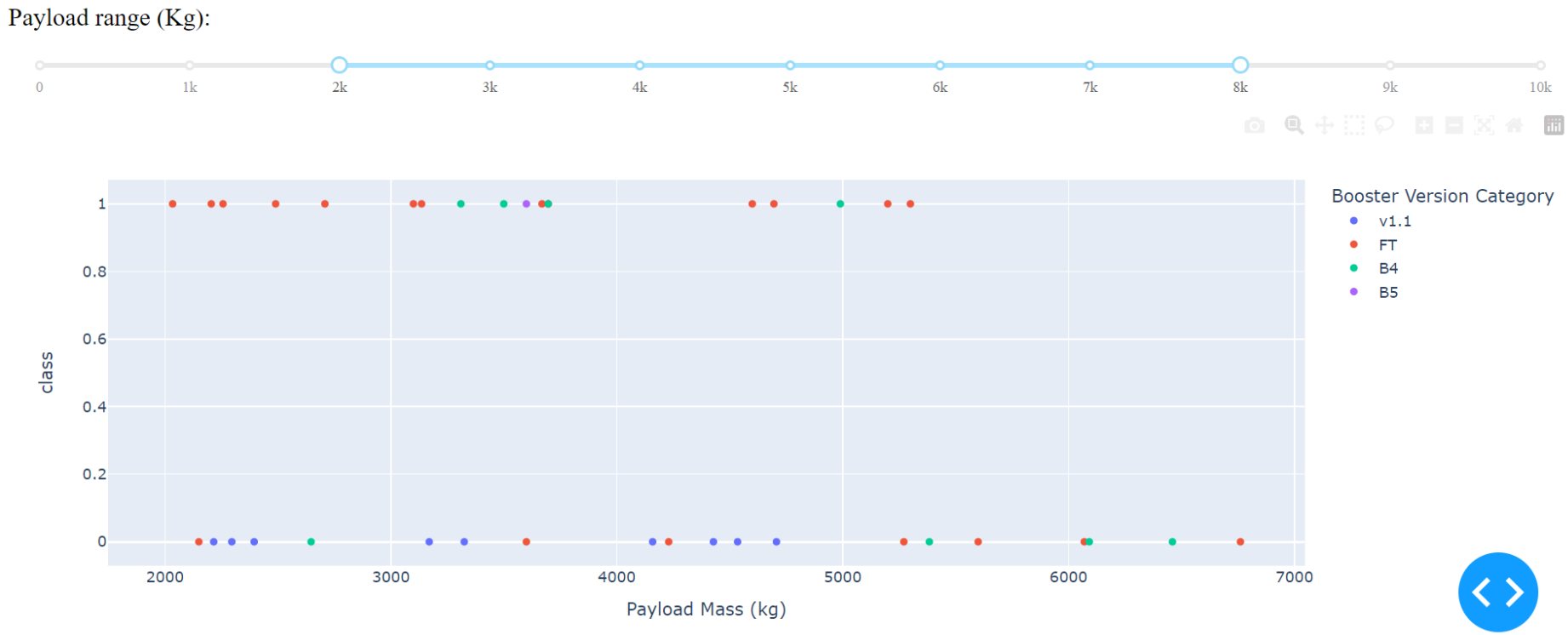


The site that makes up the greatest number of successful launches was KSC LC-39A at nearly 42%, followed by CCAFS LC-40 at around 30%

Pie chart showing success/failed launches for CCAFS SLC-40

- This is the site with the highest ratio of success/fail mission outcome, as can be seen almost half of all missions ever carried out here have been successful.





Outcome vs Payload Mass, color-coded with Booster Version category

- The most successful booster version category for a payload mass between 2k and 8k kg was evidently the FT type, while most of the v1.1 type's outcome were negative.

Section 5

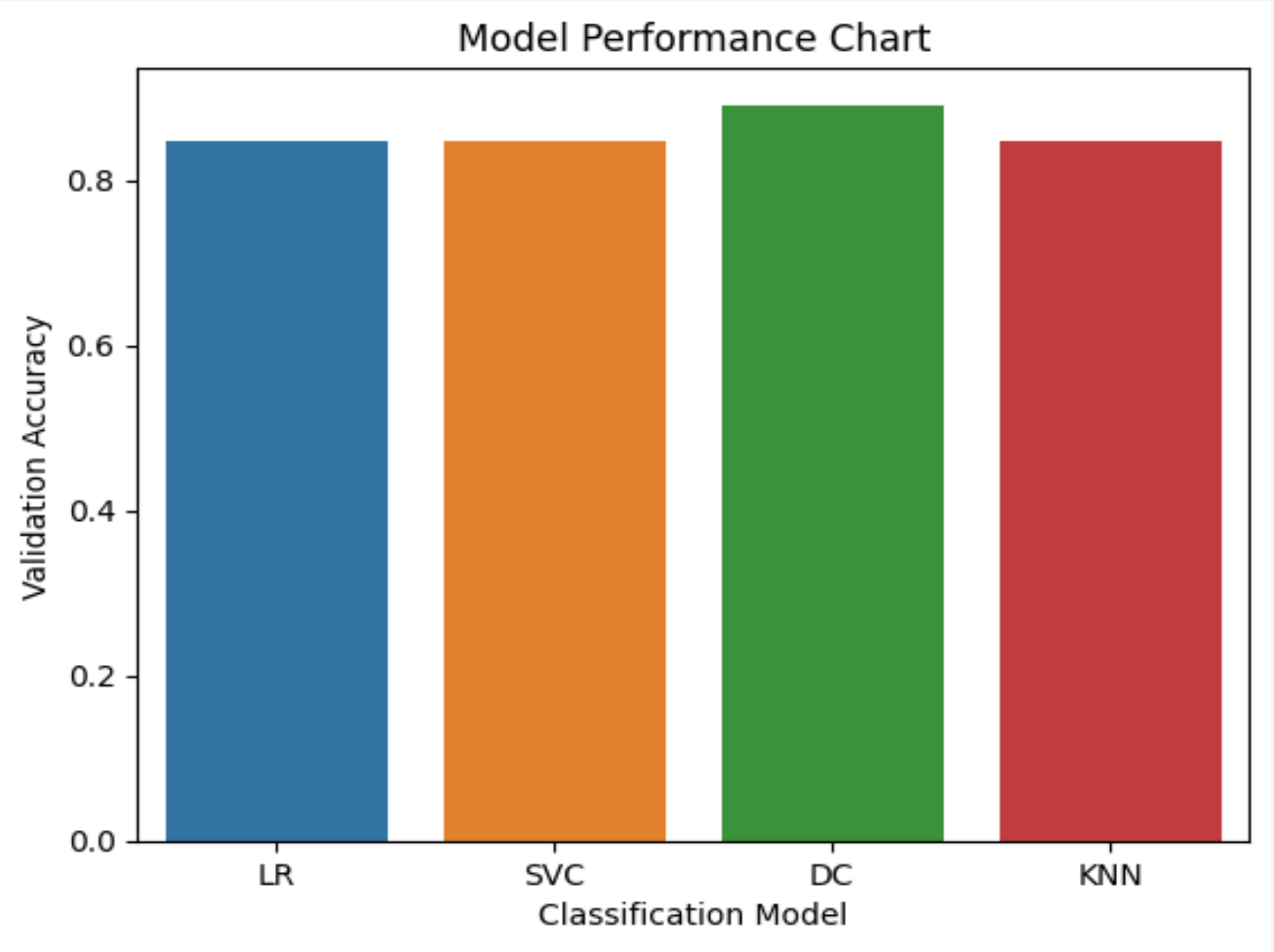
Predictive Analysis (Classification)

Classification Accuracy

The four proposed models were:

- LR: Logistic Regression
- SVC: Support Vector Machine
- DC: Decision Tree
- KNN: K-nearest neighbors

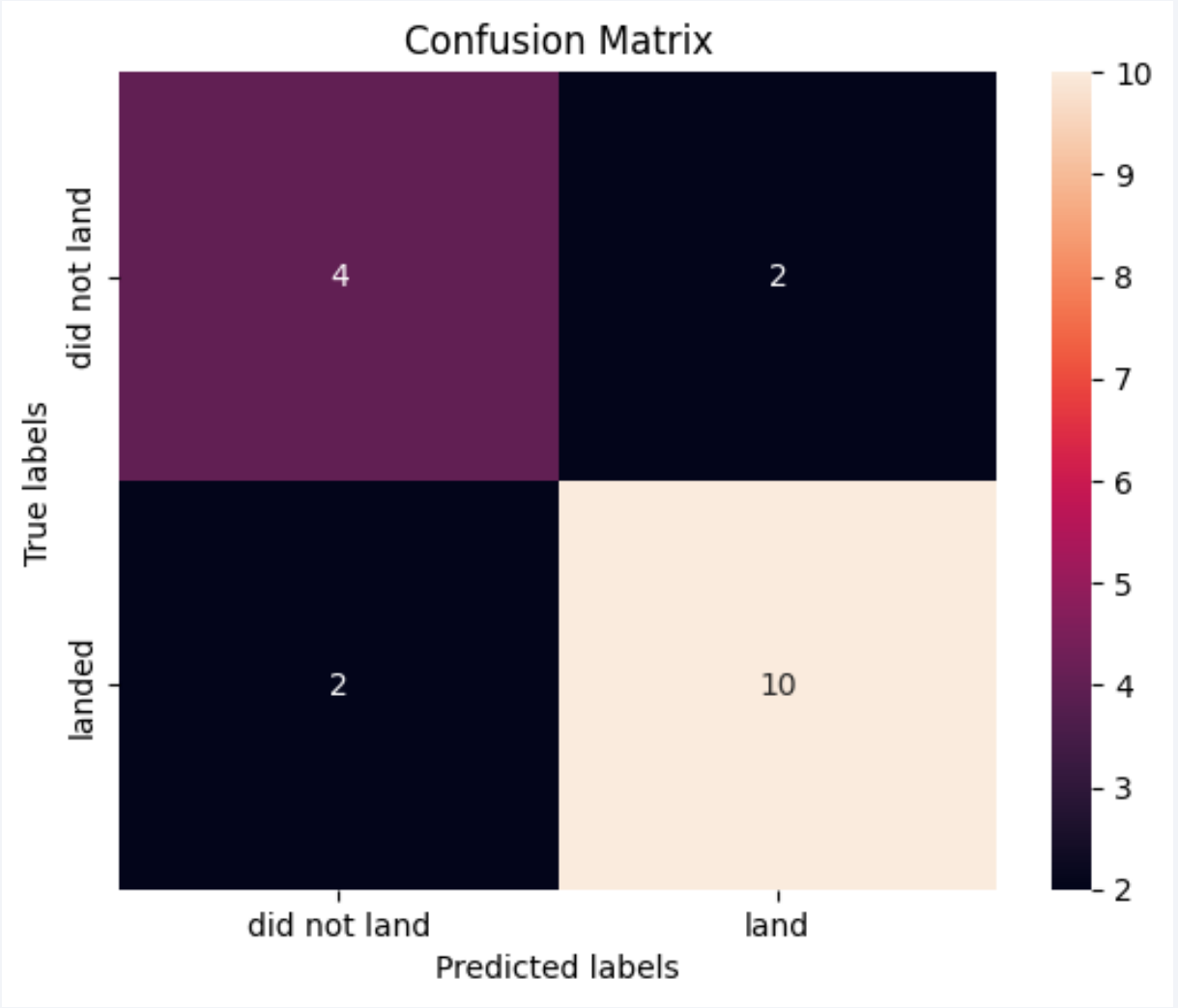
The best performing model on validation was DC



Confusion Matrix

On the right is the confusion matrix for the decision tree model, which shows the outcome of prediction on the test dataset. The model:

- Correctly predicted 10 successful landing out of 12
- Correctly predicted 4 failed landings out of 6.



Conclusions

A lot of insights were drawn from the work done in this project, and can be summarized as follows:

- SpaceX uses 2 main locations for launching their rockets, with 4 distinct Launch Sites – 3 on the East and 1 on the West Coast
- The outcome of landing and reusing the stage 1 depends on a lot of factors such as Payload Mass, Launch Site, Orbit, etc.
- A Decision Tree classification model can be used to predict the landing outcome of future missions with accuracy up to ~89%, based on the aforementioned features.

Appendix

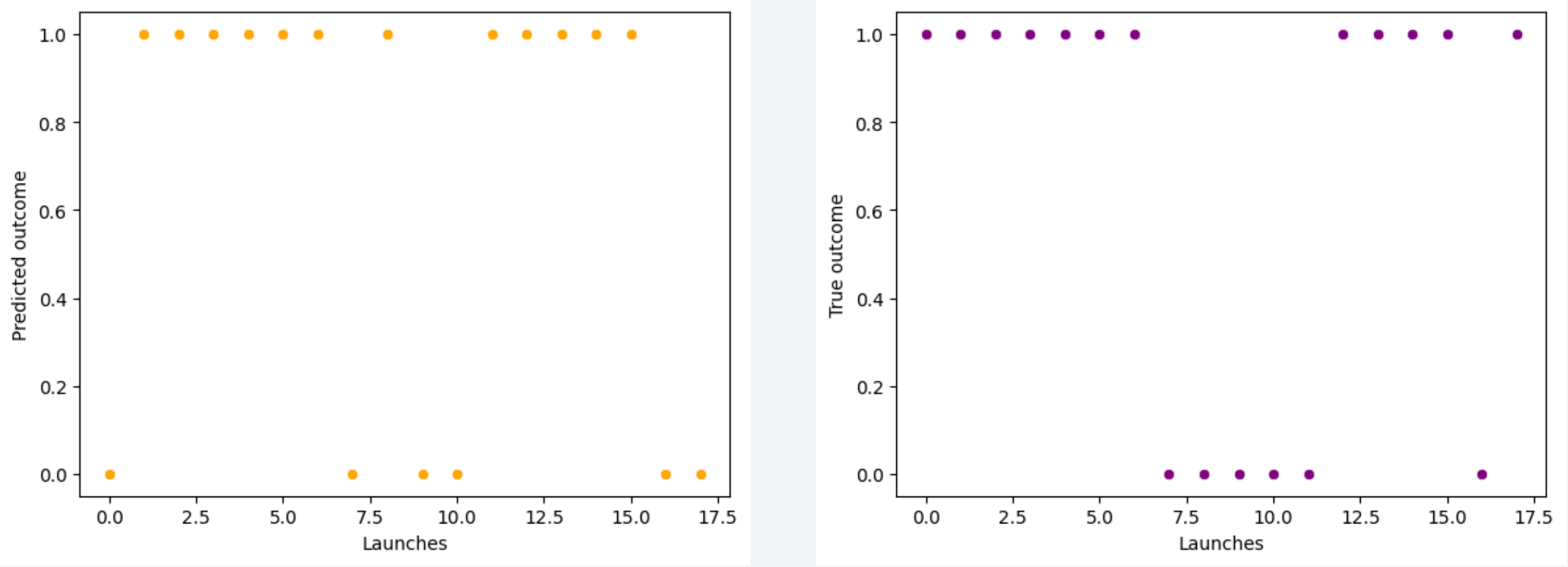


Figure: Outcome prediction by the Decision Tree classification model compared to the True values of the test data. As can be seen, they are quite similar graphically and shows that the model performs relatively well with out-of-sample data.

Thank you!