IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robbie Carney
13/02/2022

Outline

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

Executive Summary

- Several companies are now providing commercial space travel services, of which one of the most successful is SpaceX. Much of SpaceX's success is explained by their reusing the first stage of projects, which saves them millions of dollars. If we can predict if the first stage of a launch will land, we can predict that launch's cost to the company.
- The aim of this project was to acquire publicly available data on SpaceX launches, process and explore it, and use machine learning classification algorithms to predict if a first stage will land based on relevant features. Data were collected by a combination of API calls and webscraping. Null values were replaced, counts of interest were inspected and the outcome label was converted from an eight-valued to a binary variable.
- During the exploratory data analysis (EDA), the data were explored using SQL queries and visualization. We then performed launch sites proximities analysis using Folium and built an interactive dashboard using Dash to glean further insights.
- Finally, we fitted four classification algorithms to the data -- logistic regression, support vector machine (SVM), decision tree and k-nearest neighbours (KNN). We used grid search to find the best hyperparameters for these models, applied those hyperparameters, and evaluated the models on test data.
- We found that all four models performed equally, and that each allows us to predict whether the first stage of a launch will land with around 83% accuracy. This is a successful result and has considerable implications for the industry, such as if an alternative company wants to bid against SpaceX for a rocket launch.

Introduction

- Several companies are now providing commercial space travel services, of which one of the most successful is SpaceX. Much of SpaceX's success is explained by their reusing the first stage of projects, which saves them millions of dollars. If we can predict if the first stage will land, we can predict a launch's cost.
- The aim of this project was to acquire publicly available data, process and explore it, and use machine learning classification algorithms to predict if a first stage will land based on relevant features.

Section 1

Methodology

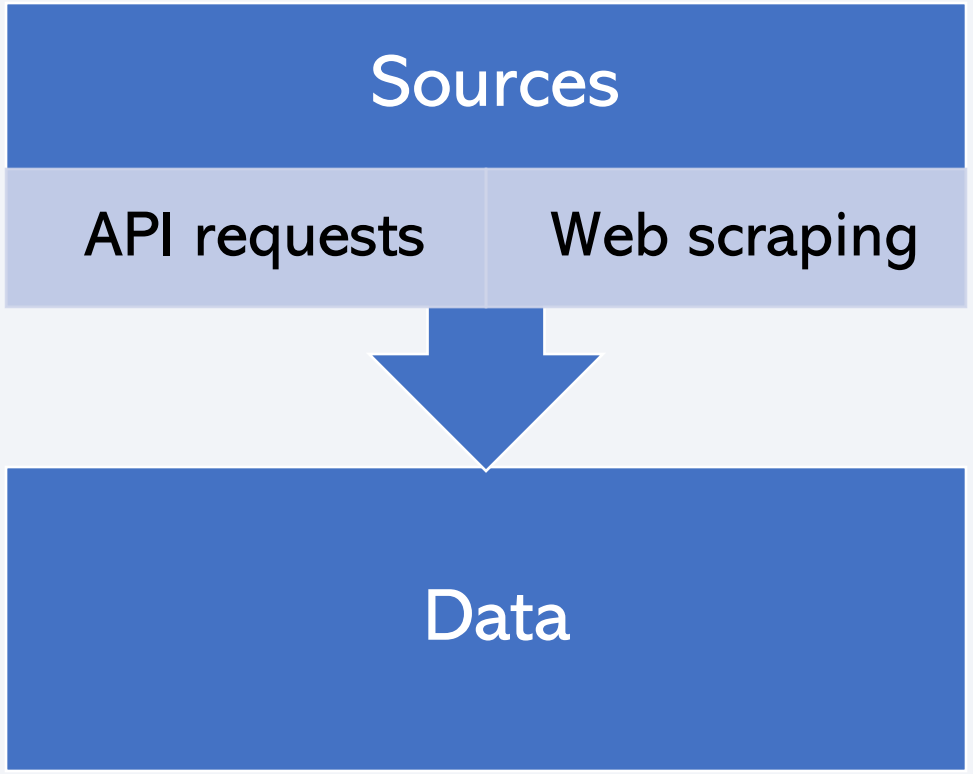
Methodology

Executive Summary

- Data collection methodology:
  - Data were collected by a combination of API calls and webscraping
- Performed data wrangling:
  - Null values were replaced, counts of interest were inspected and the outcome label was converted from an eight-valued to a binary variable.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Four classification models were fitted to the data, tuned using grid search and then evaluated on the test data

Data Collection

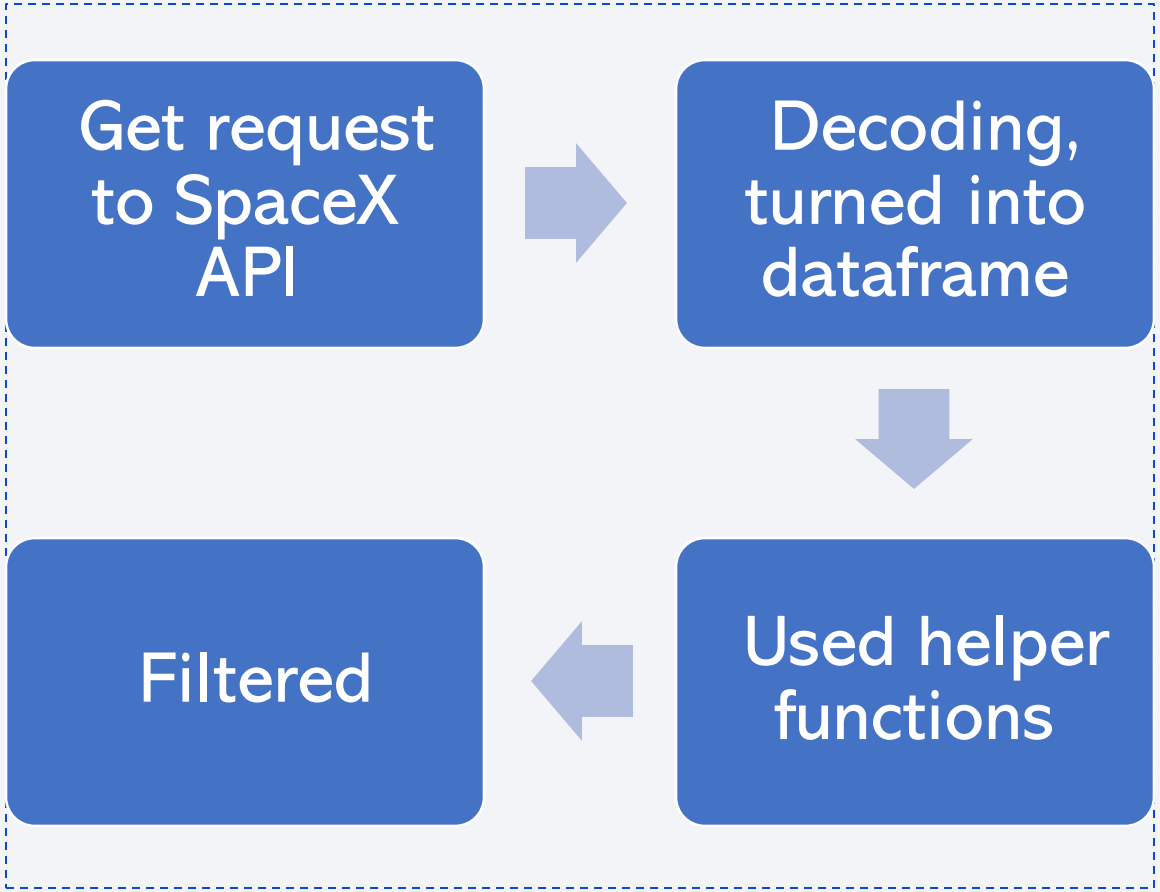
- Data were collected using a combination of requests to the SpaceX API and web scraping on a Wikipedia page:



Data Collection – SpaceX API

- Data was collected by making a get request to the SpaceX API. We defined four helper functions to acquire data of interest, decoded the API as JSON file, turned this into a Pandas dataframe, and used our defined functions to extract some more data of interest. We then filtered the dataframe to include only launches using Falcon 9 boosters.

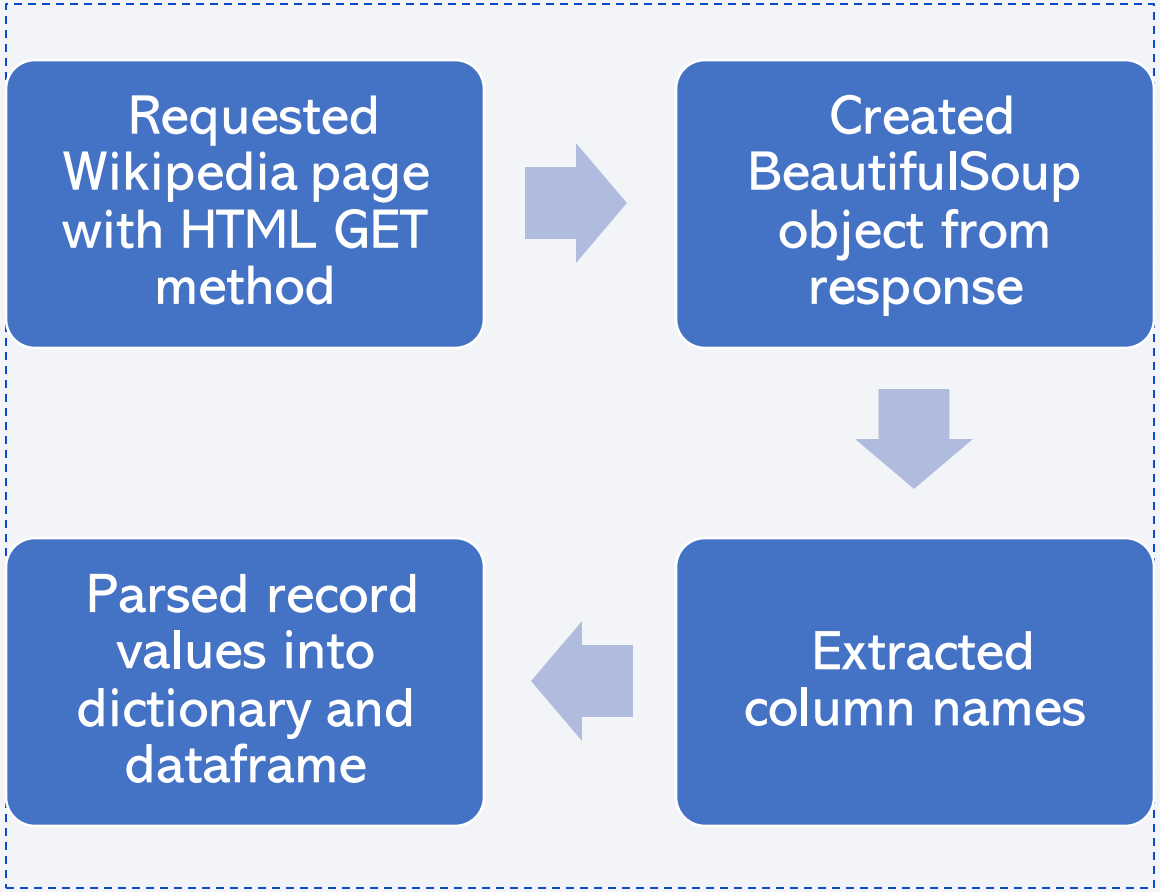
- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/Data%20collection%20API.ipynb



Data Collection – Scraping

- We extracted a Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup. We defined some helper functions, requested the HTML table and created a BeautifulSoup object from the response. We extracted the column names, then parsed the HTML tables into a dataframe.

- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/Data%20collection%20web%20scraping.ipynb



Data Wrangling

- Null values for payload mass in the dataframe drawn from the API were replaced with the mean. We considered the number of launches from each site, the counts of orbit type and the counts of each mission outcome. We then converted the outcome label from an eight-valued variable representing both outcome and landing site to a binary variable representing outcome only (0 for failure, 1 for success).

- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- We began by producing a scatter plot of flight number and payload mass against outcome to see how these two features may affect the label. We then produced scatter plots of flight number and launch site against outcome and payload and launch site against outcome, to see if there is a noticeable difference made by launch site or payload. Next, we produced a bar chart of mean outcome by orbit type, which showed us which orbit types were most correlated with success. We produced scatter plots of flight number and orbit against outcome and payload and orbit against outcome to visualize how these variables interact. Finally, we produced a line chart of average success rate by year from 2010-2020, showing a clear upward trend.
- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/EDA%20visualization.ipynb

EDA with SQL

- We performed SQL queries to display:
  - All unique launch site names
  - 5 records where launch sites begin with 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date when the first successful ground pad landing outcome was achieved
  - The names of boosters which had successful landings on a drone ship, where the payload mass was between 4000 and 6000 kg
  - The total number of successful and unsuccessful mission outcomes
  - The booster versions which have carried the maximum payload mass
  - The failed landing outcomes on drone ships and their booster versions and launch site names for the year 2015
  - The count of landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order

Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/EDA%20SQL.ipynb

Build an Interactive Map with Folium

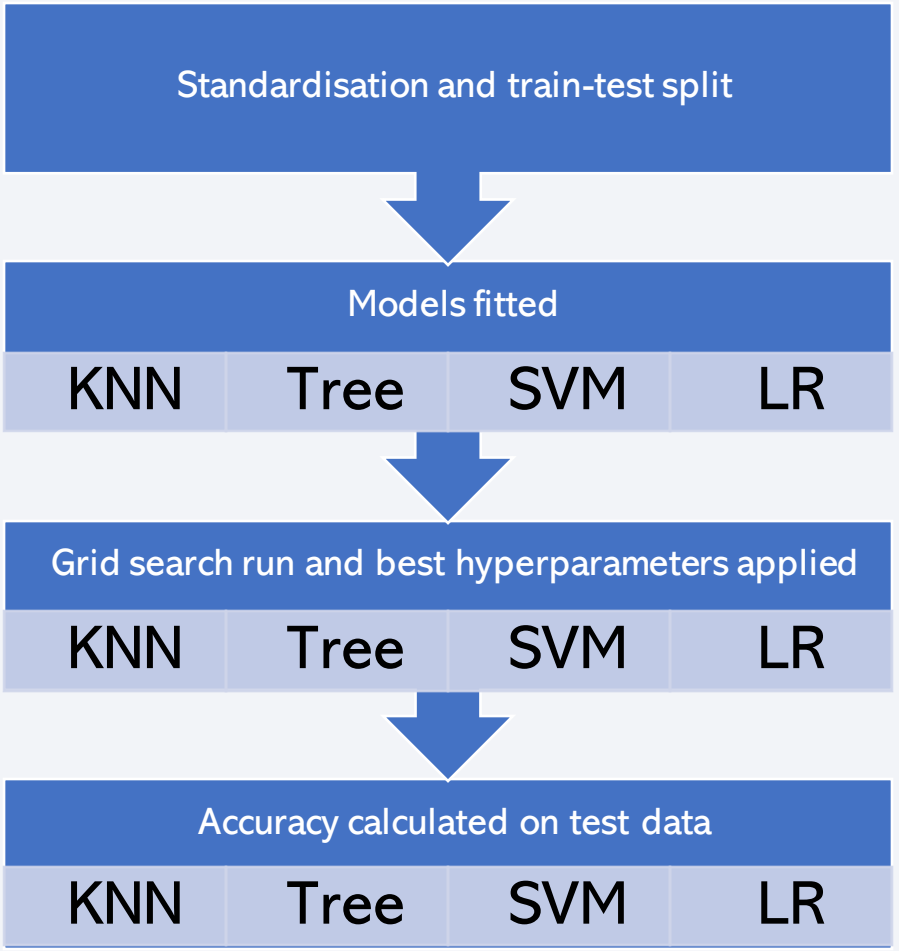
- We created a Folium map object initially centered on the Johnson Space Center in Houston. We then added circle objects with attached text labels at the coordinates of the four launch sites so that each site was highlighted and identifiable by name. From this we could see, for example, that all the launch sites were (a) on the coast and (b) relatively near the equator. We also added markers to the same coordinates, one for each launch that took place at that site. Since this would otherwise result in a cluttered map, we created marker cluster objects to gather these markers together. We also colored the markers red for failed launches and green for successful launches to instantly convey where successful and failed launches occurred. Finally, we considered the distances between launch sites and their proximities such as coastlines, cities, railways and highways using MousePosition. We placed marker objects at these proximities, labelled them with their distance from the launch sites and then drew line objects between them and the launch sites. These distance lines allowed us to gain several insights, including that launch sites are always near the coast, frequently near to roads and railways and frequently at some distance to cities.
- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/Launch%20site%20location.ipynb

Build a Dashboard with Plotly Dash

- We divided the dashboard layout into two sections. The first contains an input object in the form of a dropdown menu of launch sites together with a pie chart output. The default dropdown option is to display data for all sites, resulting in a pie chart of the share of total successful launches originating from each site. By selecting a site, the pie chart updates to show the proportion of successful and unsuccessful launches from that site. Hence, the pie chart shows us which sites were correlated most with success.
- The second part of the layout contains a range slider input and scatter chart output. The slider allows the user to select a range of payload masses; its default is all masses. The scatter chart below is a plot of payload mass against success, filtered by (a) the launch site chosen in the dropdown, if any and (b) the selected payload range, if any. Further, the colour dimension is used on the plot to show booster version. Hence, the scatter chart shows success as a function of launch site, payload mass range, and booster version, allowing us to see how these features affect the label.
- Python file: https://github.com/robbiecarney/DS_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- After standardising the data and splitting it into training and test portions, we fitted four classification models to the data: a logistic regression (LR) model, a support vector machine (SVM) model, a decision tree model, and a k-nearest neighbours (KNN) model. We used grid search to find the best hyperparameters for each of these models and selected those hyperparameters for each model. We then calculated the accuracy of each refined model on the test data.
- Notebook: https://github.com/robbiecarney/DS_capstone/blob/main/ML%20prediction.ipynb

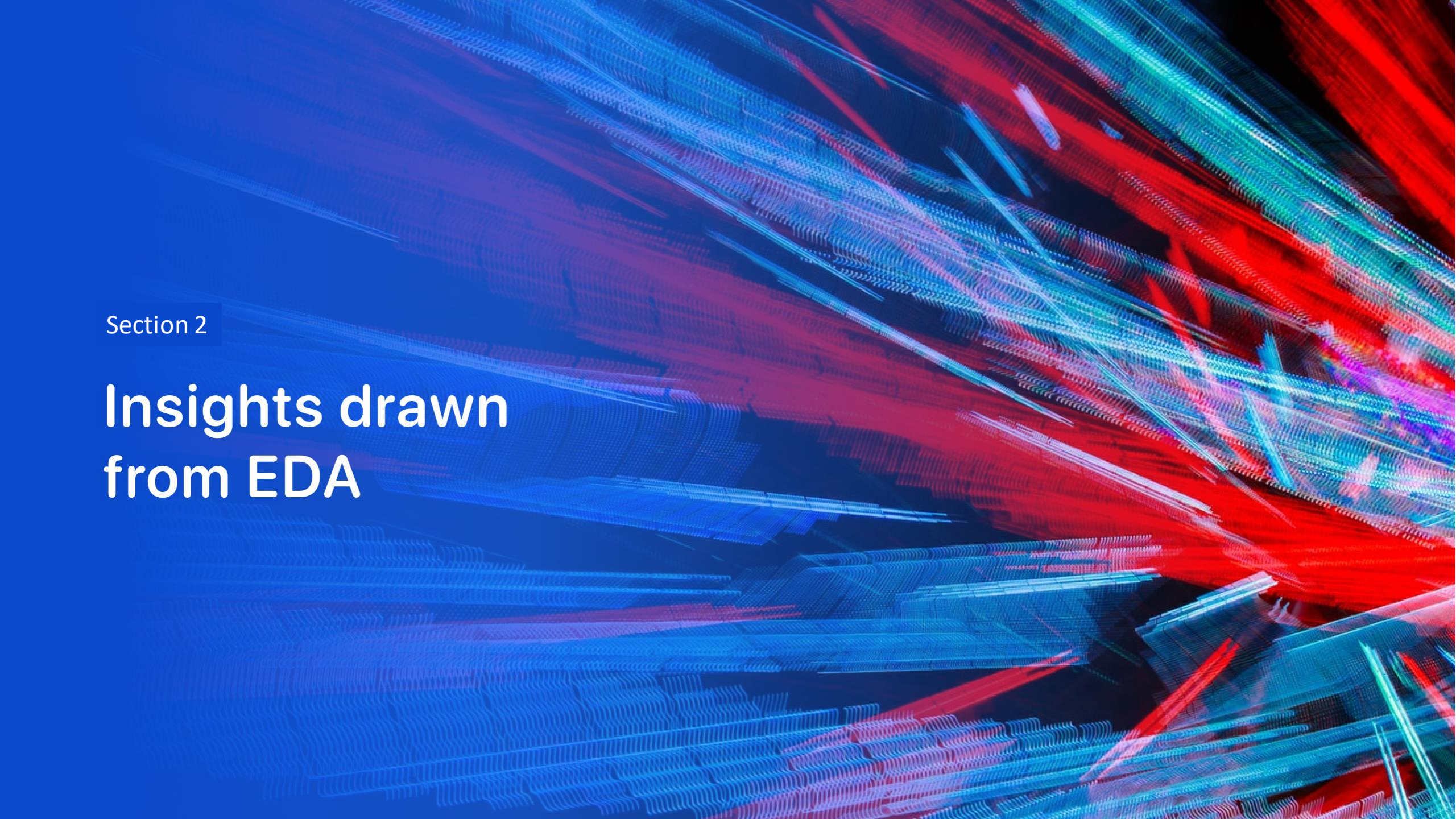


Results

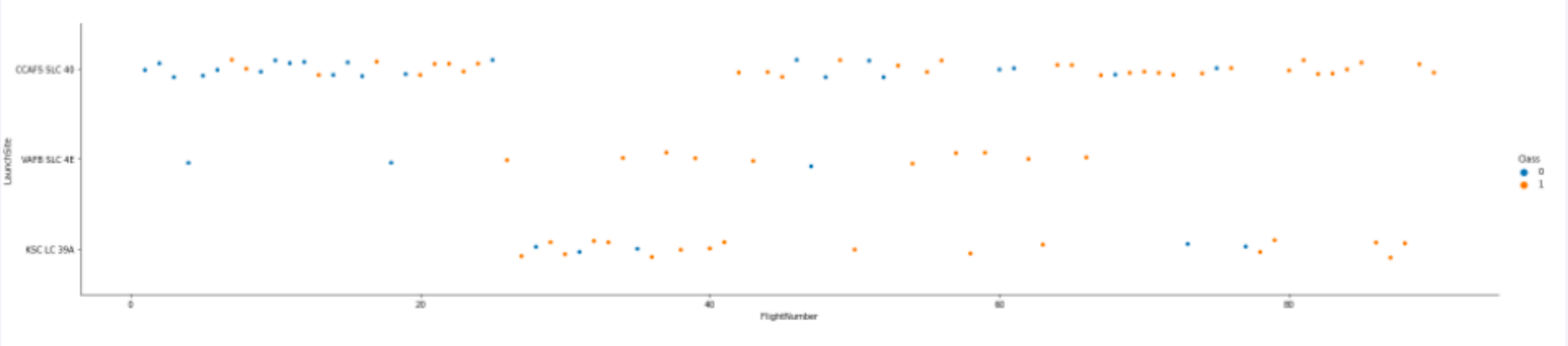
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



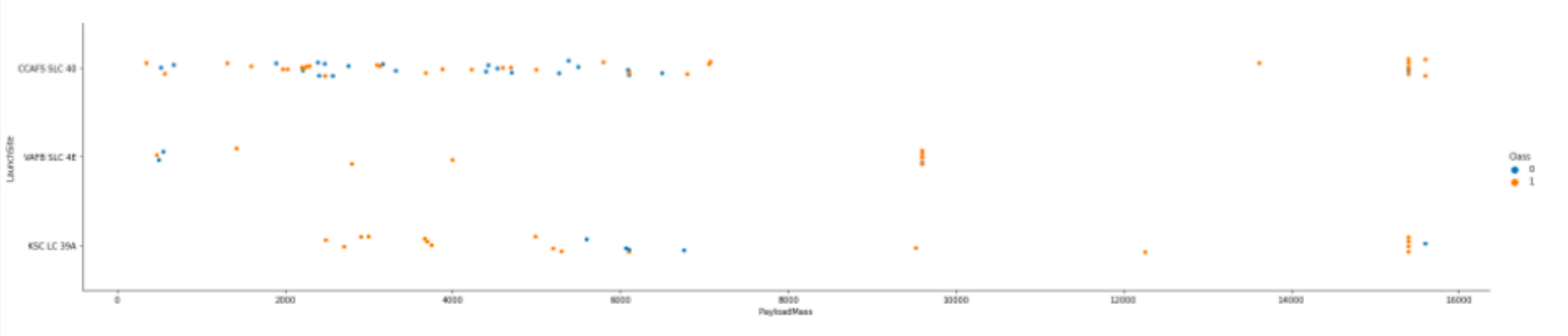
Flight Number vs. Launch Site



Scatter plot of flight number vs launch site, coloured by success.

The plot shows success increasing with flight number, which is unsurprising given that SpaceX likely learns from trial and error over time.

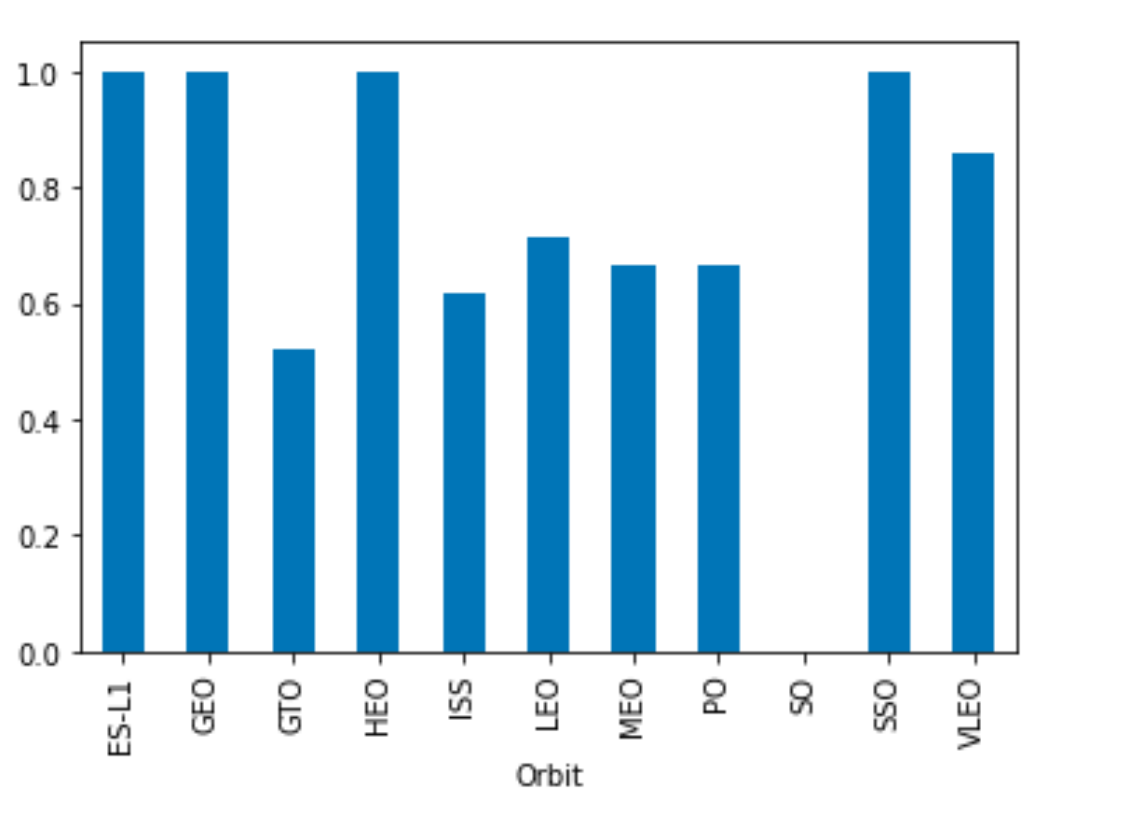
Payload vs. Launch Site



Scatter plot of payload mass vs launch site, coloured by success.

The plot shows success increasing with payload mass. This is more surprising, perhaps explained by heavier payload masses only being attempted on later flights (see first plot in notebook).

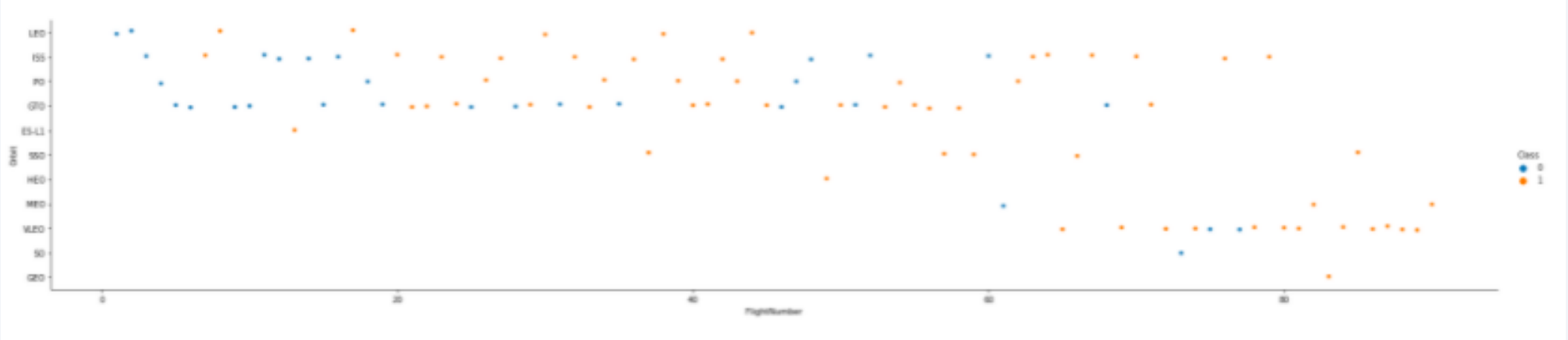
Success rate vs orbit type



Bar chart of success rate for each type of orbit.

The chart shows the most successful orbit types are ES-L1, circular geosynchronous, highly elliptical and sun-synchronous, while the least successful was geosynchronous. (Only one launch is given as orbit type SO, which was unsuccessful, but this is another name for sun-synchronous).

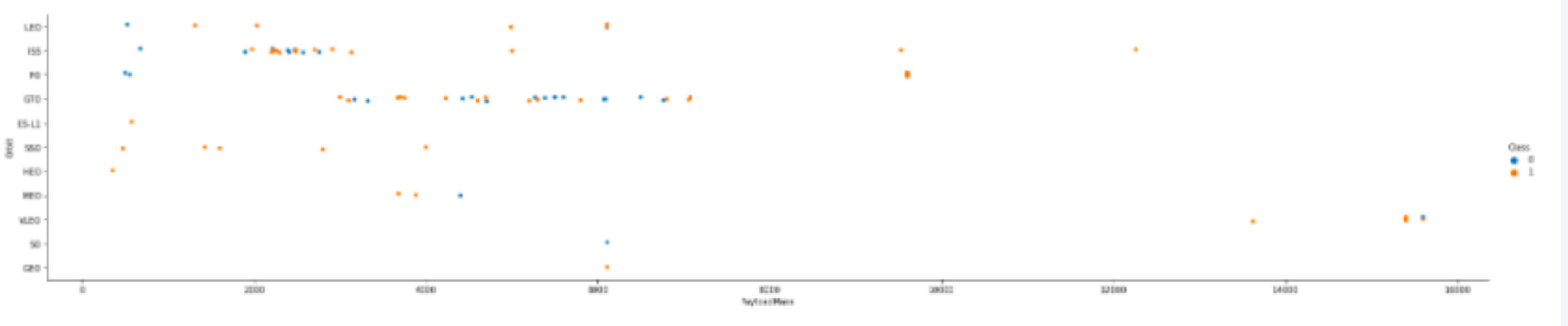
Flight number vs orbit type



Scatter plot of flight number vs orbit type, coloured by success.

The plot shows the last flights for each orbit type are normally successful, but in varying numbers (one final success for geosynchronous vs eight for very low earth orbit). This could imply that SpaceX is able to learn from experience more successfully with certain orbit types than with others.

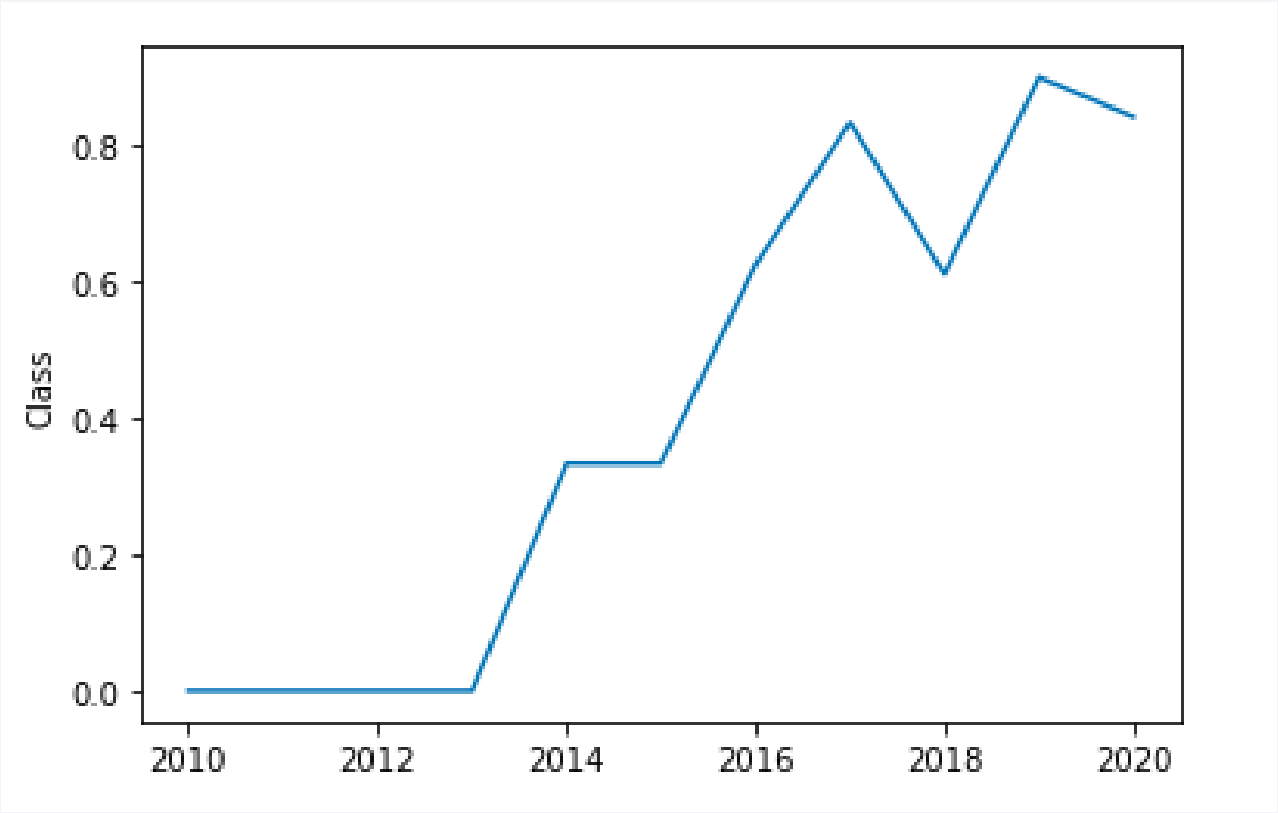
Payload vs orbit type



Scatter plot of payload vs orbit type, coloured by success.

The plot shows the only attempts at payloads over 8,000 kg were for ISS, polar and very low earth orbit orbit types, but most of these were successful.

Launch success yearly trend



Line chart of yearly average success rate.

The plot shows a clear upward trend with a plateau around 2014 and a dip around 2018. The general trend is unsurprising given SpaceX can learn from trial and error over time.

All Launch Site Names

- Unique launch site names were queried as follows. They are displayed in the output.

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET
```

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Launch site names beginning with 'CCA'

- Five launch site names beginning with 'CCA' were queried as follows. They are displayed in the output.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31929/blu
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	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

Total Payload Mass

- Total payload mass for customer 'NASA (CRS)' was queried as follows. The result is displayed in the output.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
    1
45596
```

Average Payload Mass by F9 v1.1

- Below is the query to display average payload mass carried by booster version F9 v1.1, and the result.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/blu
db
Done.
   1
2534
```

First Successful Ground Landing Date

- Query to find first successful ground landing date, with result.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query to list the names of boosters which have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000. Result below.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

booster_version
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1029.2
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1038.1

Total Number of Successful and Failure Mission Outcomes

- Query to display total number of successful and failure mission outcomes, with result. As we can see, missions are almost always considered successful (not to be confused with launch success).

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Boosters Carried Maximum Payload

- Query to list the names of the boosters which have carried the maximum payload mass, with result.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- Query to list failed drone ship landings, their booster versions, and launch site names for the year 2015. Result below.

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

```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

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

- Query to list count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order. Result below. As we can see, the modal landing outcome for this fairly long period was that no landing was attempted.

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

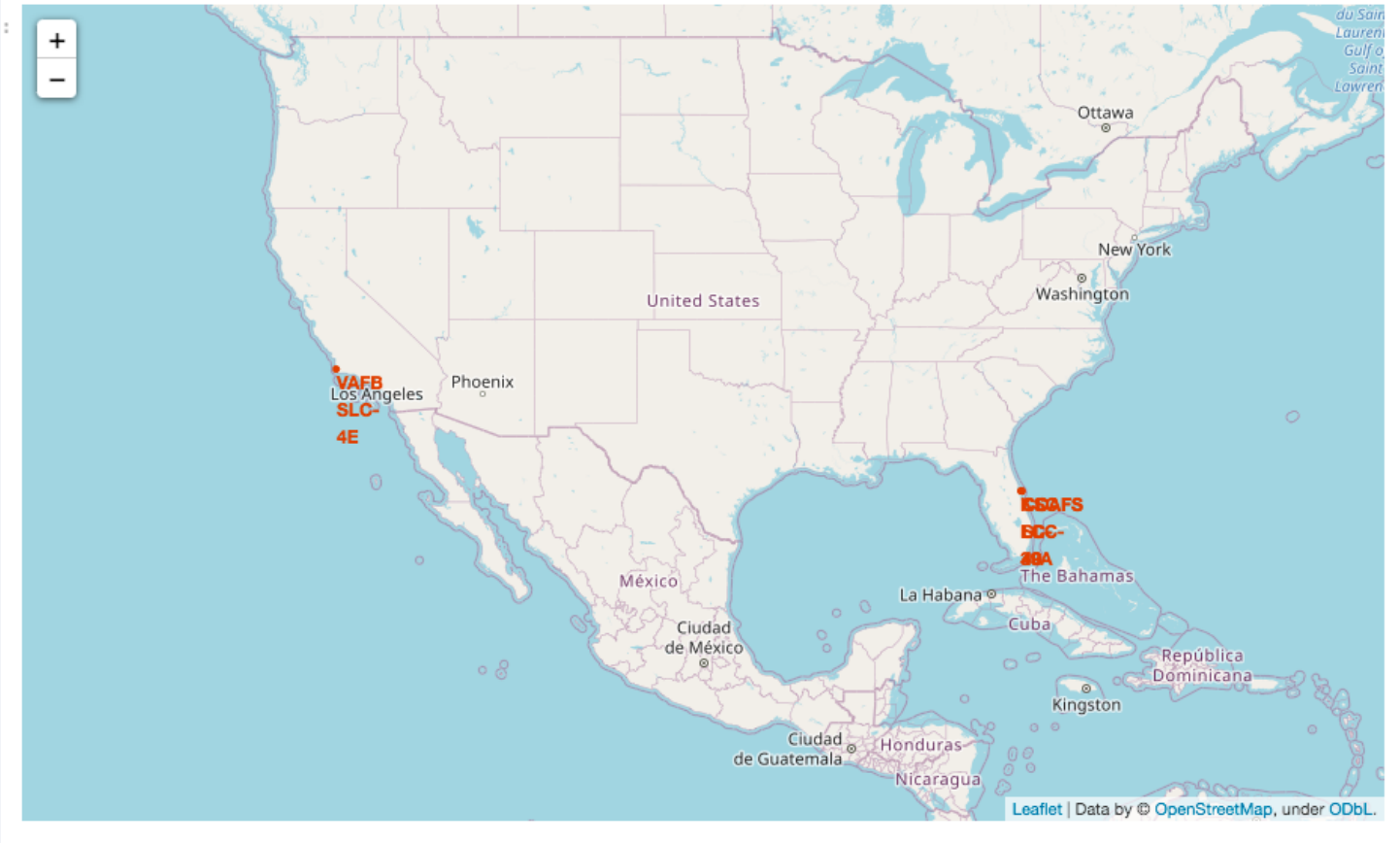
```
 * ibm_db_sa://sps46893:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

landing__outcome	COUNT
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


Section 3
Launch Sites
Proximities Analysis

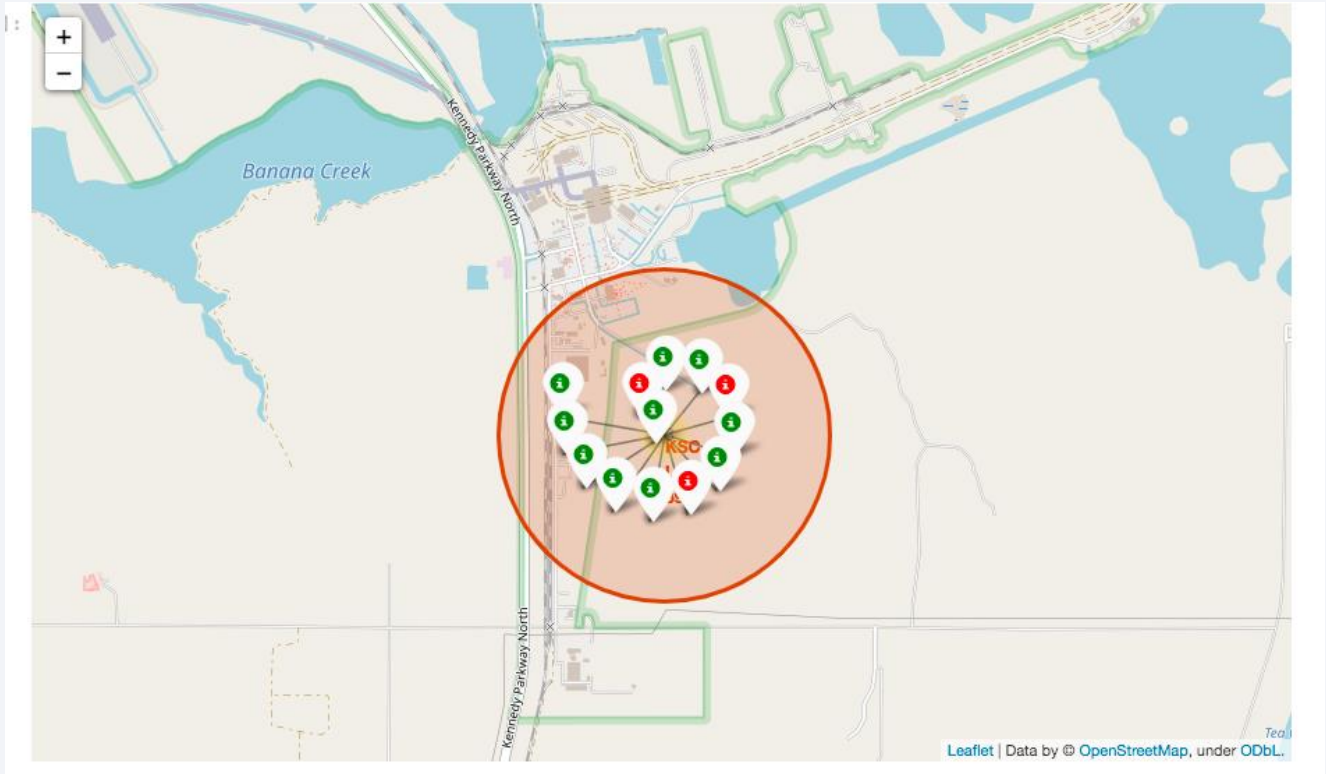
Folium map of all launch sites

- Map of all launch sites. We can see that all launch sites are on the coast and relatively near the equator.



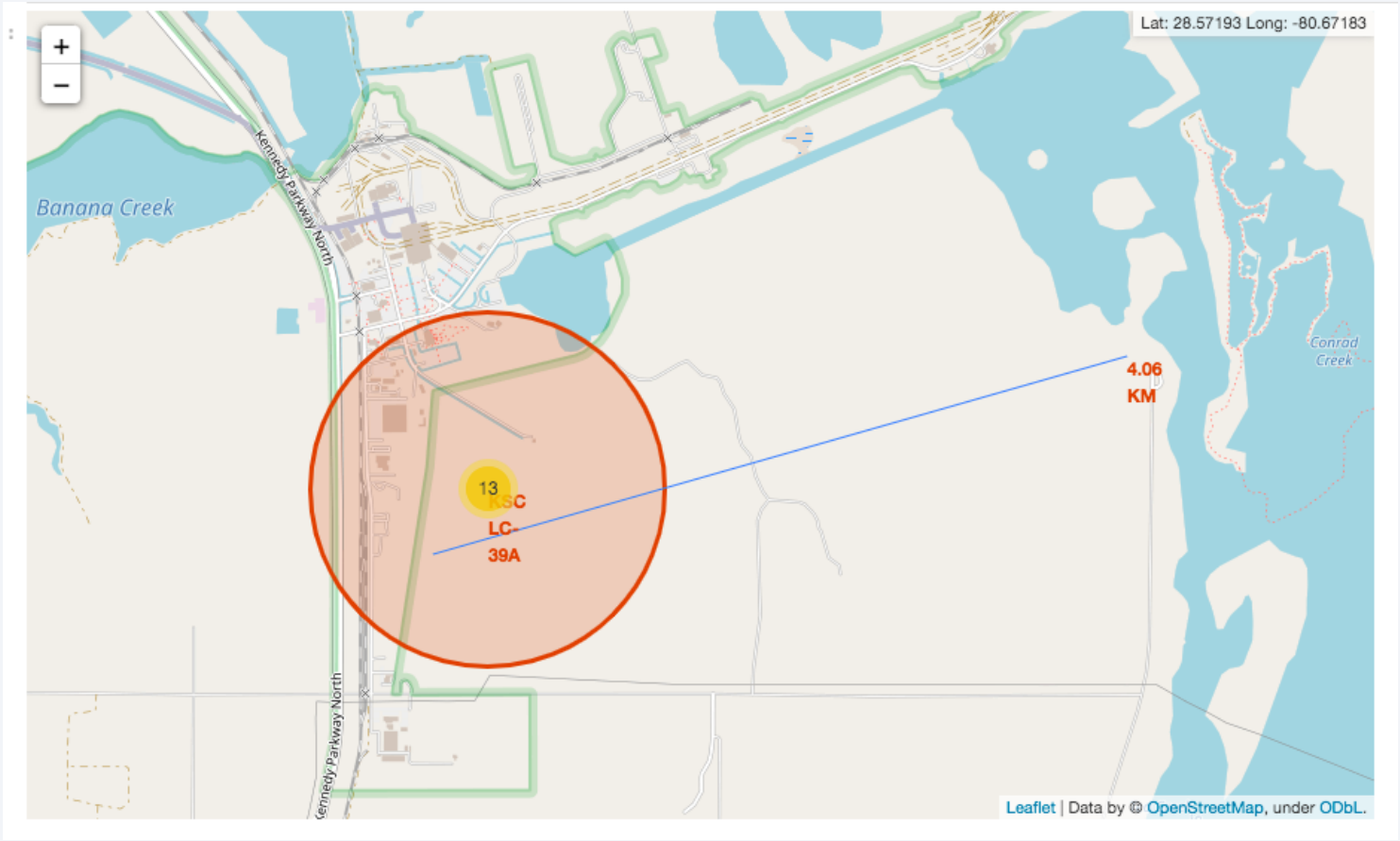
Launch attempt markers

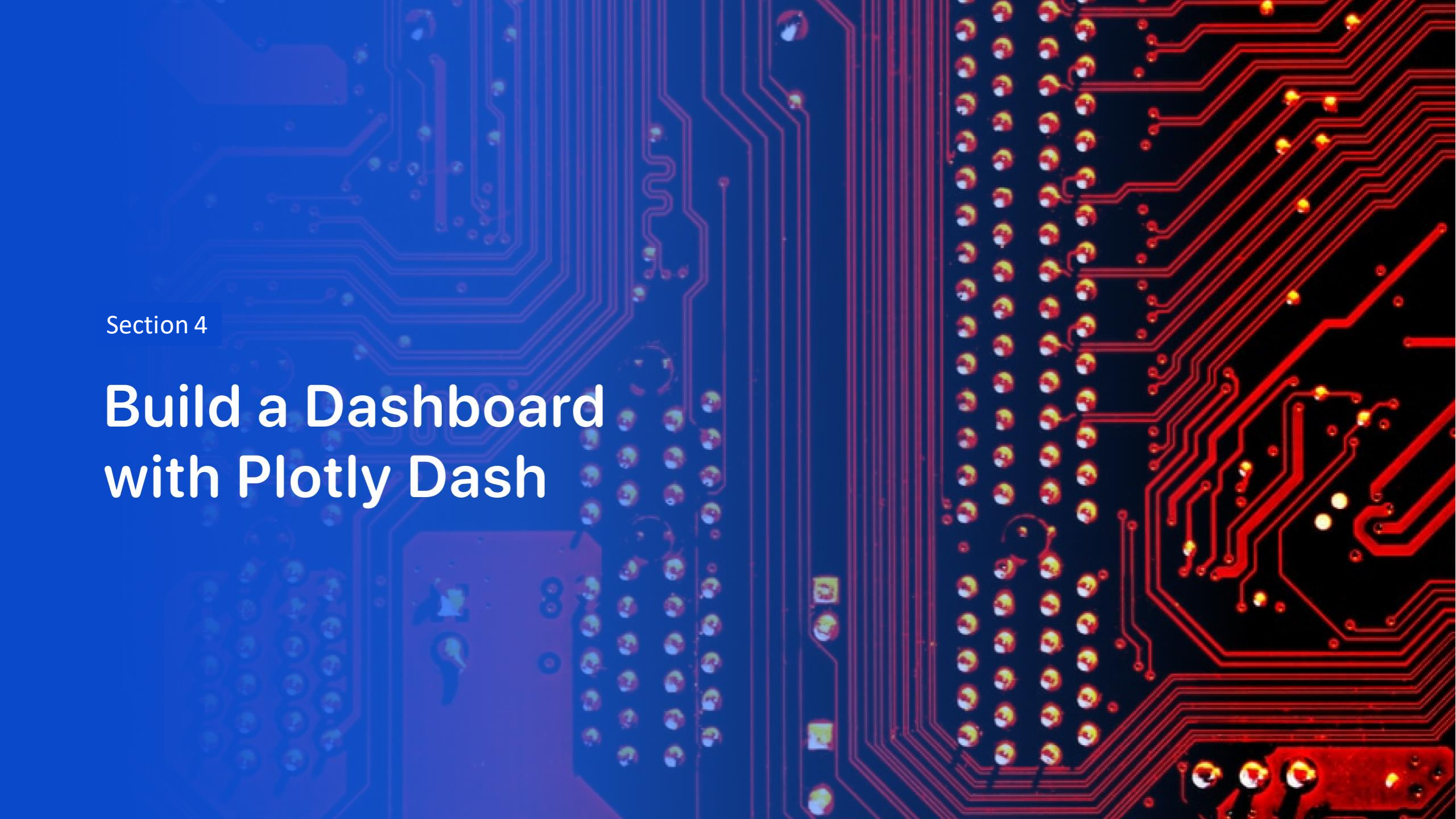
- Colour-coded markers for launch attempts – green for successful, red for failed. We can see the successful and failed attempts from each launch site.



Map of launch site proximities with lines and displayed distance

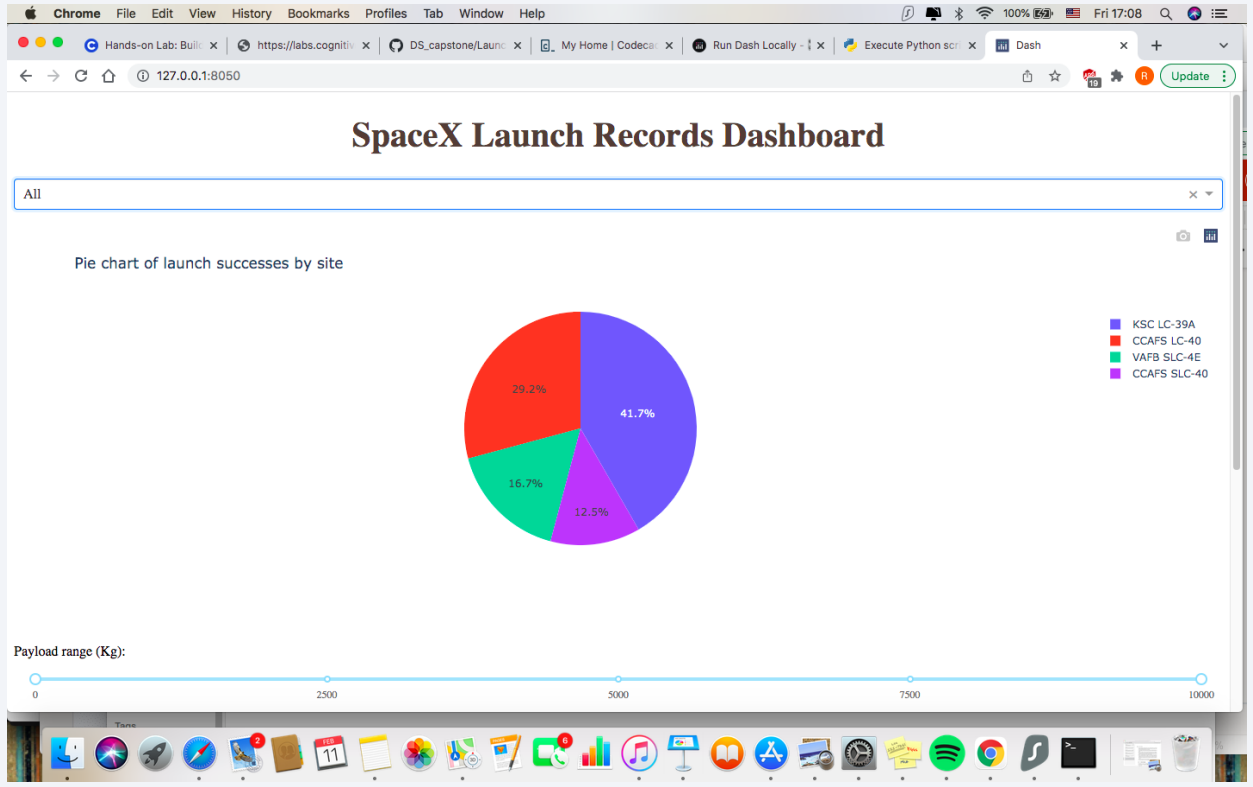
- Here we see the fairly high proximity of the launch site to the coast.




Section 4
Build a Dashboard with Plotly Dash

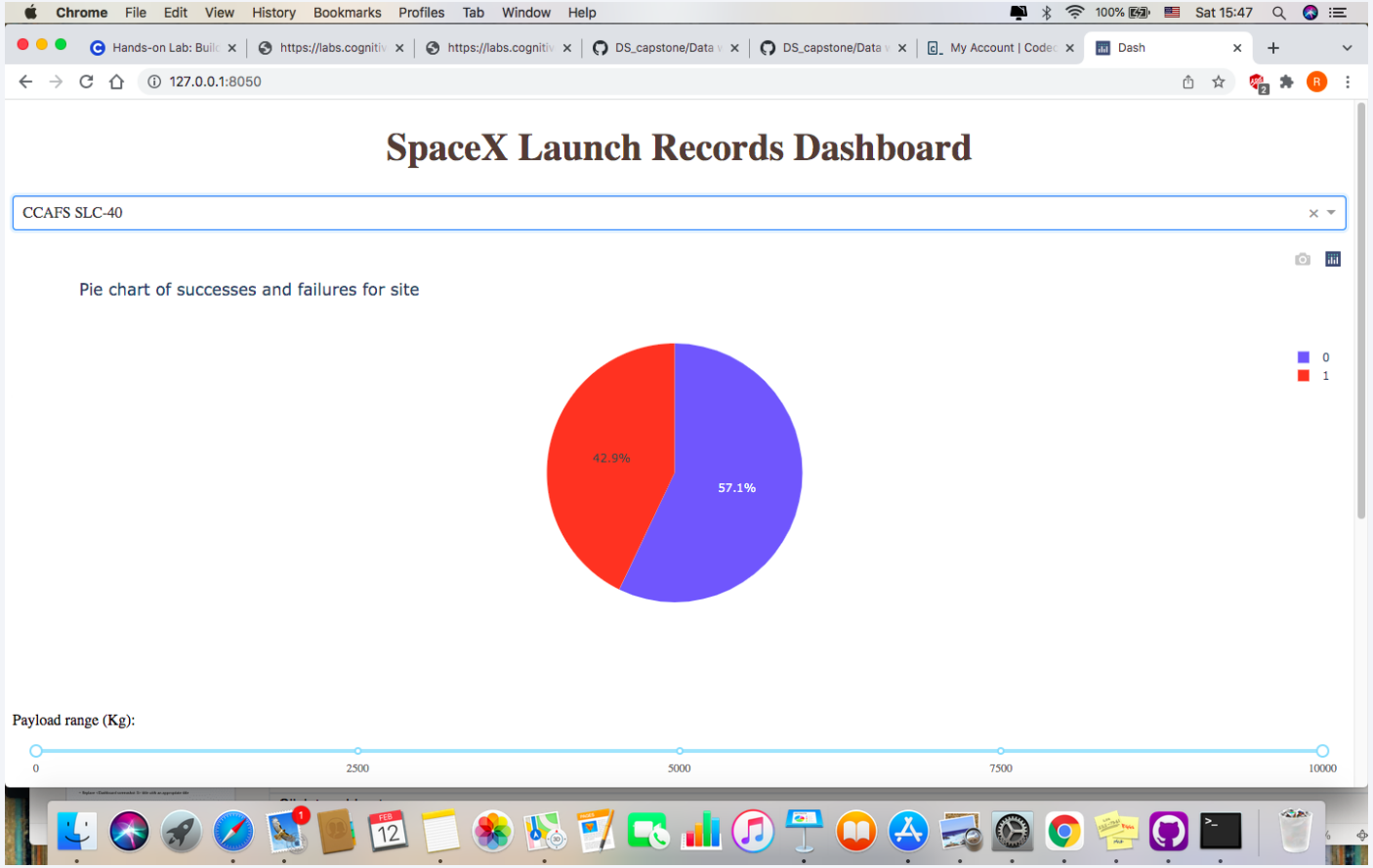
Pie chart - all sites

- Pie chart of share of successes by launch site. We see that the highest number of successful launches originated from the Kennedy Space Center, and the lowest from Cape Canaveral SLC40.



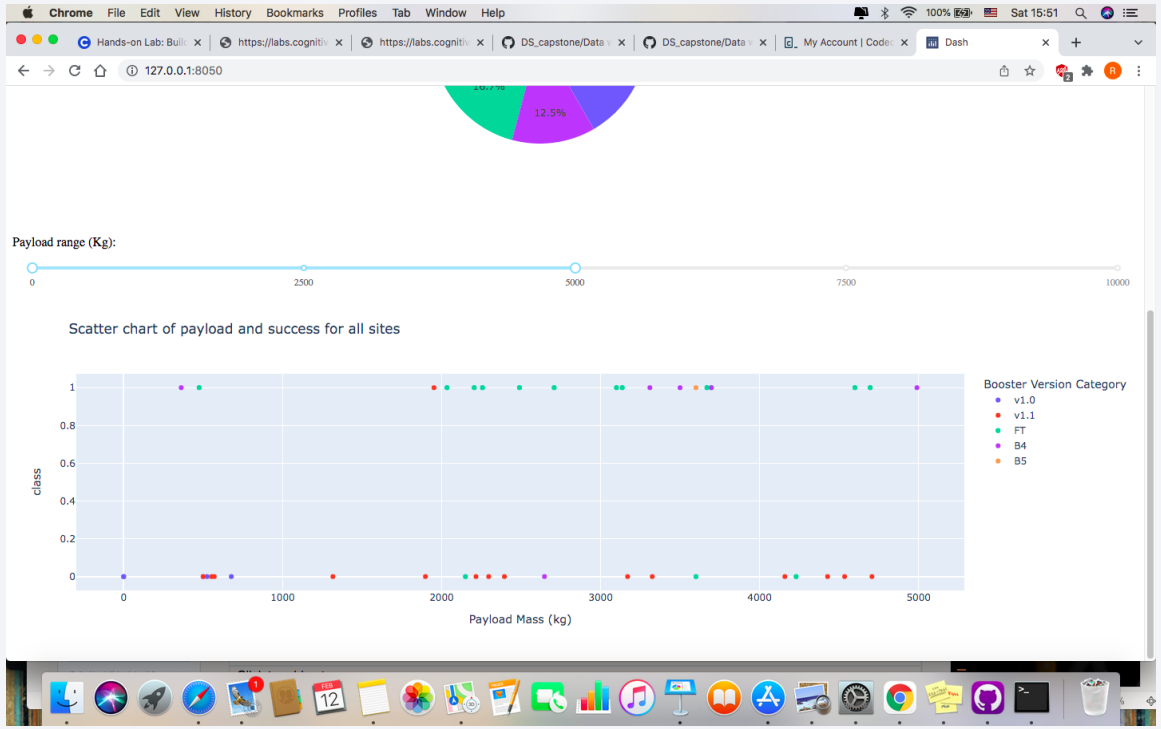
Pie chart – highest success ratio site

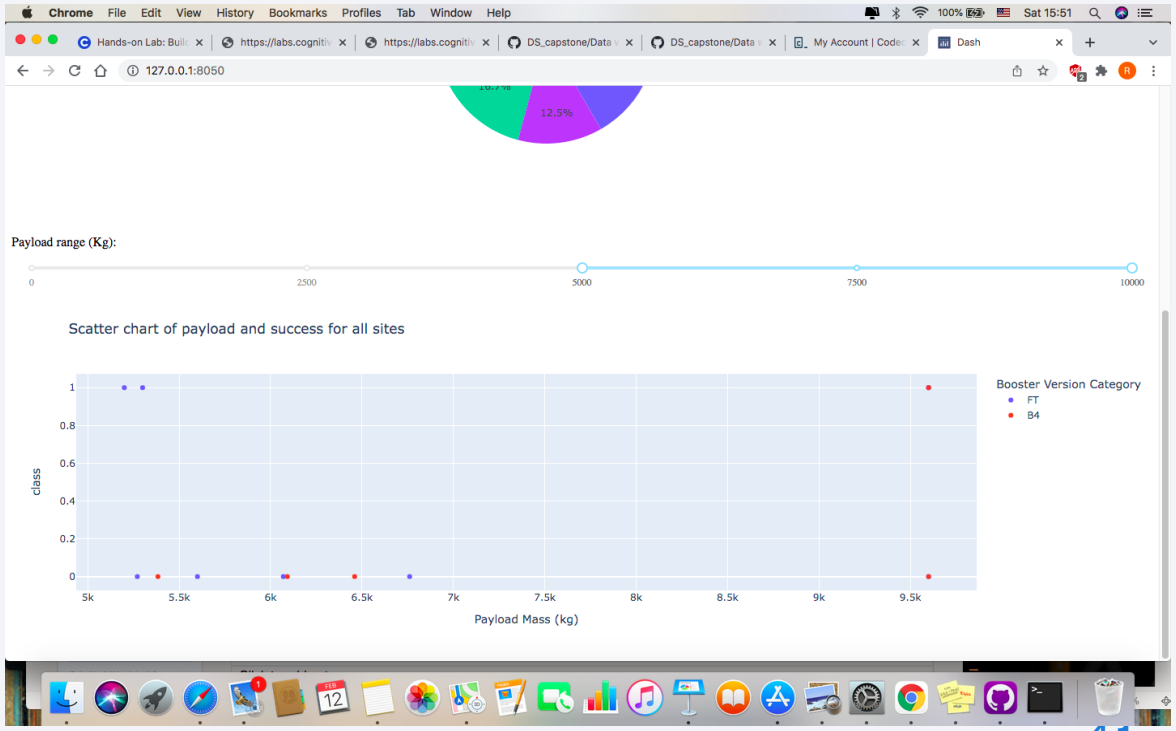
- Pie chart of successes and failures from Cape Canaveral SLC 40. This is the site with the highest ratio of successes to failures.



Scatter plots of payload and success

Scatter plot of payload and success for all sites, coloured by booster version, with payload range 0-5,000kg (left) and 5,000-10,000kg (right). We see that most launches had payloads of 5,000kg or less, that FT boosters have a fairly high success rate at low payload mass, and that v1.1 boosters have a fairly low one.



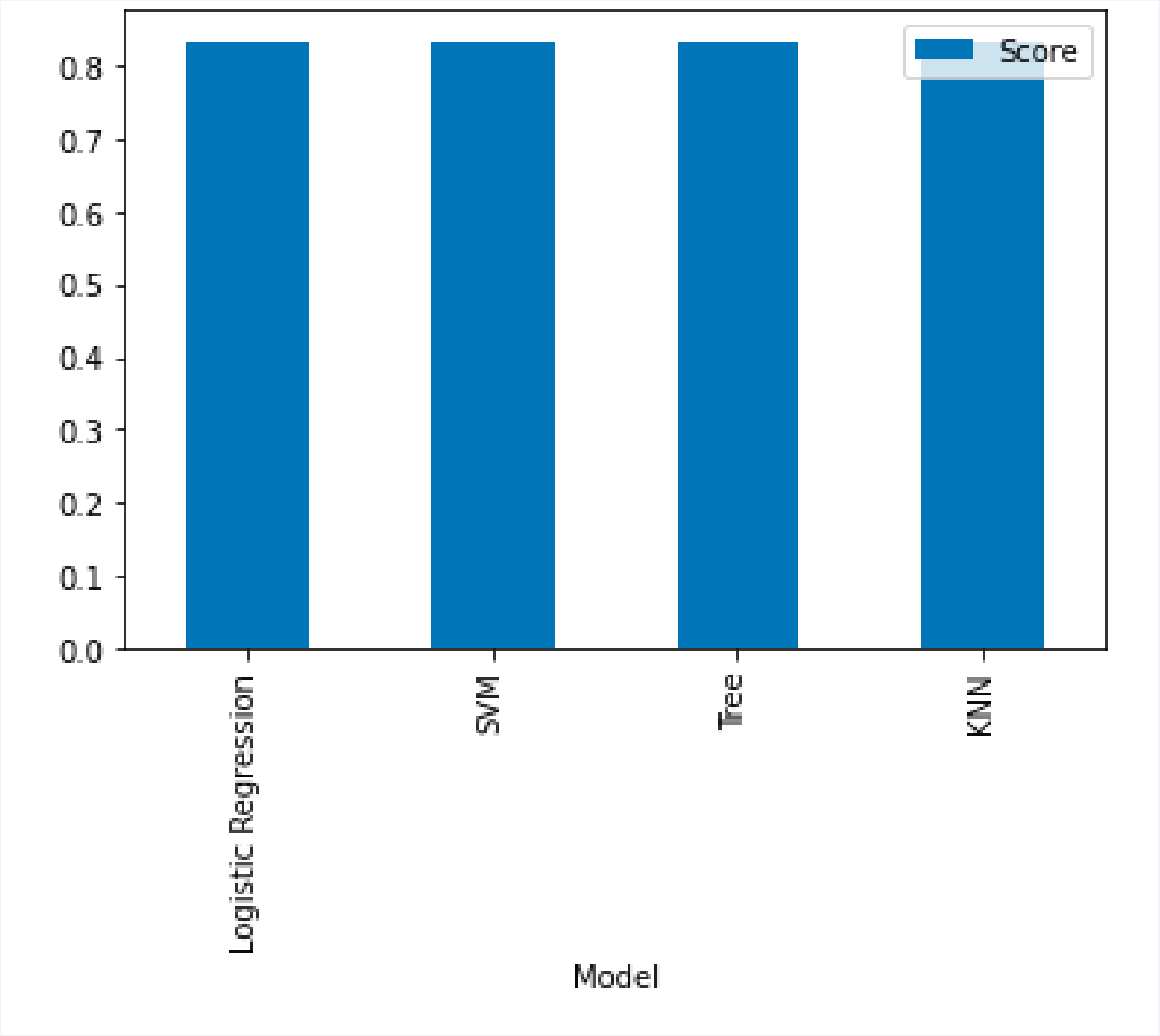


Section 5

Predictive Analysis (Classification)

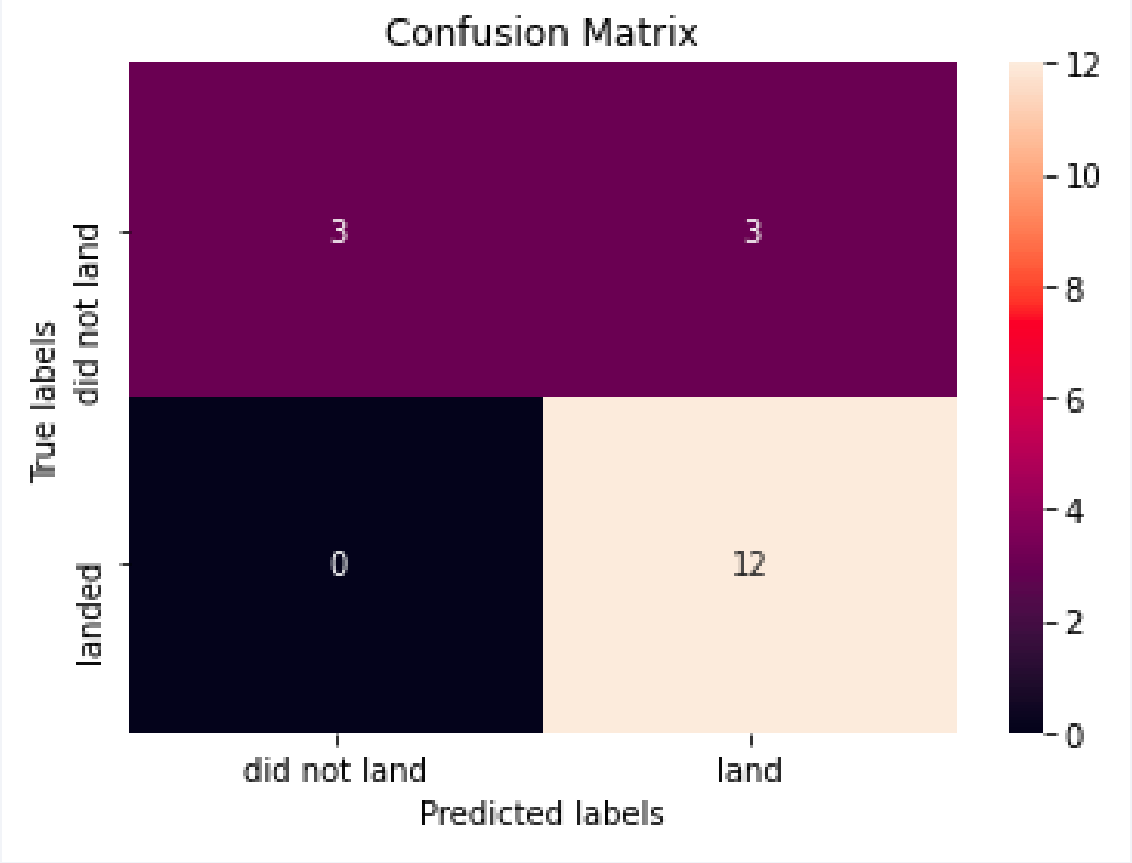
Classification Accuracy

- Once grid search was run and the best parameters applied, all built models had the same accuracy on the test data, as can be seen in the bar chart.



Confusion Matrix

- All four models returned the same confusion matrix, the one shown here. As we can see, the models are fairly accurate, with the main issue being the false positives in the top right quadrant.



Conclusions

- Our four equally-performing classification models allow us to predict whether the first stage will land with around 83% accuracy.
- This is a successful result. It has considerable implications for the industry, allowing us to predict the likely price of a launch, which could be useful if an alternate company wants to bid against SpaceX for a launch.

Appendix

GitHub repository for module: https://github.com/robbiecarney/DS_capstone

Thank you!