IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lewis Joyce
22/06/2023

Outline

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

Executive Summary

- This project aims to determine the factors that lead to the successful landing of SpaceX Falcon 9 rockets. Data on past Falcon 9 launches was obtain using the SpaceX REST API and web scraping from Wikipedia. After cleaning and formatting the data, exploratory data analysis was performed through the use of scatter plots and bar charts. Further insights were obtained using SQL queries and Folium interactive maps. Some of these results were presented as a Plotly dashboard.

- The data was fit to four machine learning models, which predicted whether the landing would be a success or a failure. These were logistic regression, SVM, decision tree and K-nearest neighbours. The logistic regression and SVM models were the most accurate at 83.3%.

- From all of this analysis, it was learned that heavier payloads tend to have a higher success rate, that the largest proportion of successful landings came from the KSC launch site, and that the FT booster version has the highest number of successful landings.

Introduction

- The SpaceX Falcon 9 rocket offers low-cost launches of around $62 million compared to other rocket providers, which can exceed $165 million. The main reason for this is that SpaceX can land and reuse the first stage of the launch, reducing the net cost.

- The aim of this project was to use data corresponding to successful and failed Falcon 9 launches to determine whether the first stage would survive the launch and to assess the factors that affected the success rate, thus being able to infer the impact on the cost.

Section 1

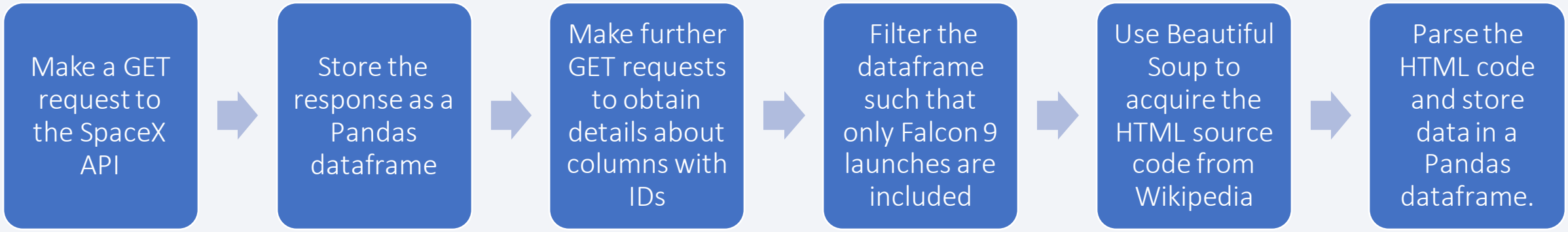
Methodology

Methodology

- Executive Summary
- Data collection methodology.
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- The datasets were obtained by making a number of GET requests to the SpaceX API and scraping data from Wikipedia about historical Falcon 9 launches.
- The details of the process are as follows:



Data Collection – SpaceX API

Make the following get request to the API to obtain past SpaceX launches

```
response = requests.get("https://api.spacexdata.com/v4/launches/past")
```

GET Request

1

Decode the response as a JSON file and convert it into a Pandas dataframe

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle

Raw Data

2

Make further GET requests to obtain information about ID columns, such the booster version of the rockets and the mass of the payloads

```
# Takes the dataset and uses the rocket column to call the API and append the data to the list
def getBoosterVersion(data):
    for x in data['rocket']:
        if x:
            response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
            BoosterVersion.append(response['name'])
```

GET data from ID columns. This is an example of getting booster version from rockets.

3

Keep only the desired columns and reformat the remaining data. This shows a subset of the dataframe .

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721

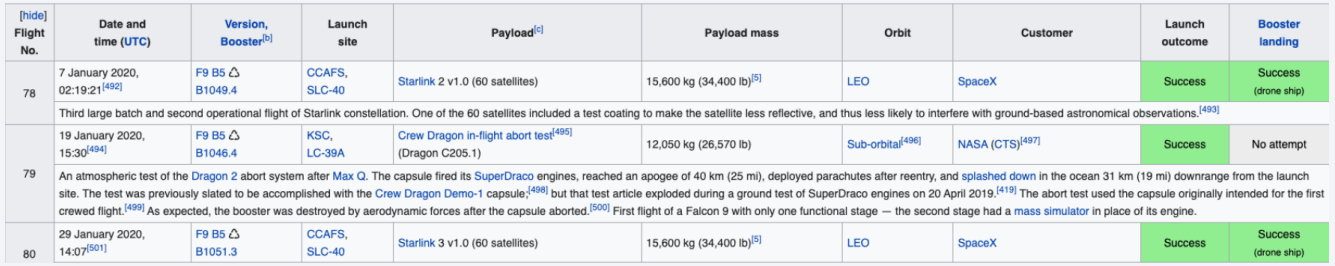
Final data frame before data wrangling

4

- Link : https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%201%20-%20Data%20Collection%20and%20Wrangling.ipynb

Data Collection - Scraping

Request HTML source code from Wikipedia



[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 △ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

1

```
response = requests.get('https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922')
```

Parse the HTML with Beautiful Soup and extract the data from the columns

```
<table class="wikitable plainrowheaders collapsible" style="width: 100%;">
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Version,
br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-Dragon-12">[c]</a></sup>
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
<th scope="col">Customer
</th>
```

2

Store the data as a pandas dataframe

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44

3

- https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%202%20-%20Data%20Collection%20and%20Web%20Scraping.ipynb

Data Wrangling

The data was prepared for the exploration and modelling stages by ensuring all variables were present and in the desired format.

The Payload Mass column contained missing values. These were replaced with the mean.

```
data_falcon9.isnull().sum()
```

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       5
```

```
# Calculate the mean value of PayloadMass column
pl_mean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan,pl_mean,inplace=True)
```

1

Identify all landing outcomes

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

2

Create a set of failure outcomes from this column

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

3

Create a binary categorical outcome column called class (0 for failure, 1 for success)

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class=[]
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

4

- https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%203%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- A number of charts were plotted to gather some initial insights about the relationships between different features of the data and the successful landing of the first stage.
- A scatter plot between the flight number and the payload mass was created with the hue showing the class. This chart provides insight about how the payload mass and the outcome correlate across sequential launches and displays how the mass impacts the landing outcome.
- Similar scatter plots were produced to show the following relationships along with the outcome as the hue:

- Flight Number and Launch Site
- Launch Site and Payload Mass
- Flight Number and Orbit Type
- Payload and Orbit type

- A bar chart was plotted to show the success rate (mean of the class column) of each type of orbit.
- A line plot was created to show how the launch success rate varied yearly.

https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%205%20-%20Data%20Visualisation.ipynb

EDA with SQL

SQL queries were made to the database to obtain various insights about the dataset. The queries made are outlined below.

- The distinct launch site names were selected and five records were displayed from the 'CCFAS' sites.
- The total payload mass launched by NASA and the average mass launched using the F9v1.1 booster were calculated.
- The names of the booster versions that carried the maximum payload mass were listed
- The total number of success and failure outcomes were shown and the types of success outcome were ranked.
- The earliest success outcome was printed and the failures in the year 2015 were obtained.
- The booster versions that successfully landed after carrying a payload mass between 4000kg and 6000kg were listed.

https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%204%20-%20Exploratory%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

An interactive Folium map was created to show the locations and total number of launches at each of the four site. The success of the launch was also displayed through the use of coloured markers.

- Circles of radius 1000m were added at each launch site on the map. Marker objects were included to display the names of each site.
- A marker cluster was assigned to each site where each marker would indicate a launch, with the colour representing the outcome (green for success, red for failure). This allows the viewer to clearly see the number of launches at each location on the map, and they can zoom in and click on the location to view the breakdown of the successes.
- Polylines and markers were included to display the distance between the launch sites and significant geographical features, such as railways, coastlines and cities.

https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%206%20-%20Interactive%20Visual%20Analytics.ipynb

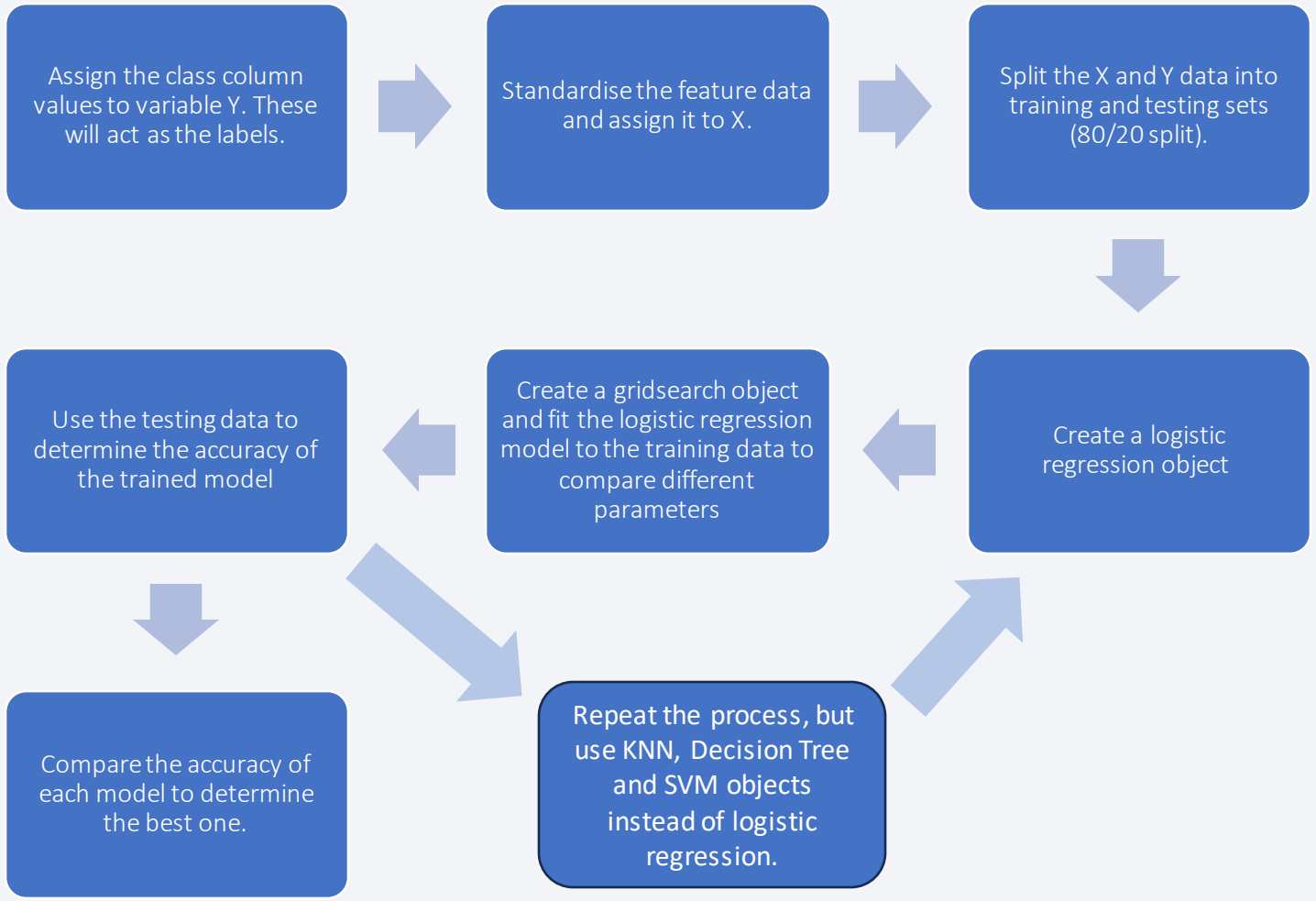
Build a Dashboard with Plotly Dash

A Plotly dashboard was created to display certain results in an elegant and interactive way. The following plots were included in the dashboard.

- A pie chart is displayed that shows the number of successful launches at each site. By default, a chart with all launch sites is shown, but each site can be viewed individually by choosing the desired option from the dropdown box
- A scatter plot displaying the landing outcome class against the payload mass, with the hue corresponding to the booster version. A range slider is included so that the user can focus on the outcomes in a specific range. As with the pie chart, all launch sites are shown by default, but can be focused on individually using the dropdown box.

https://github.com/lewisjjoyce/IBMCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



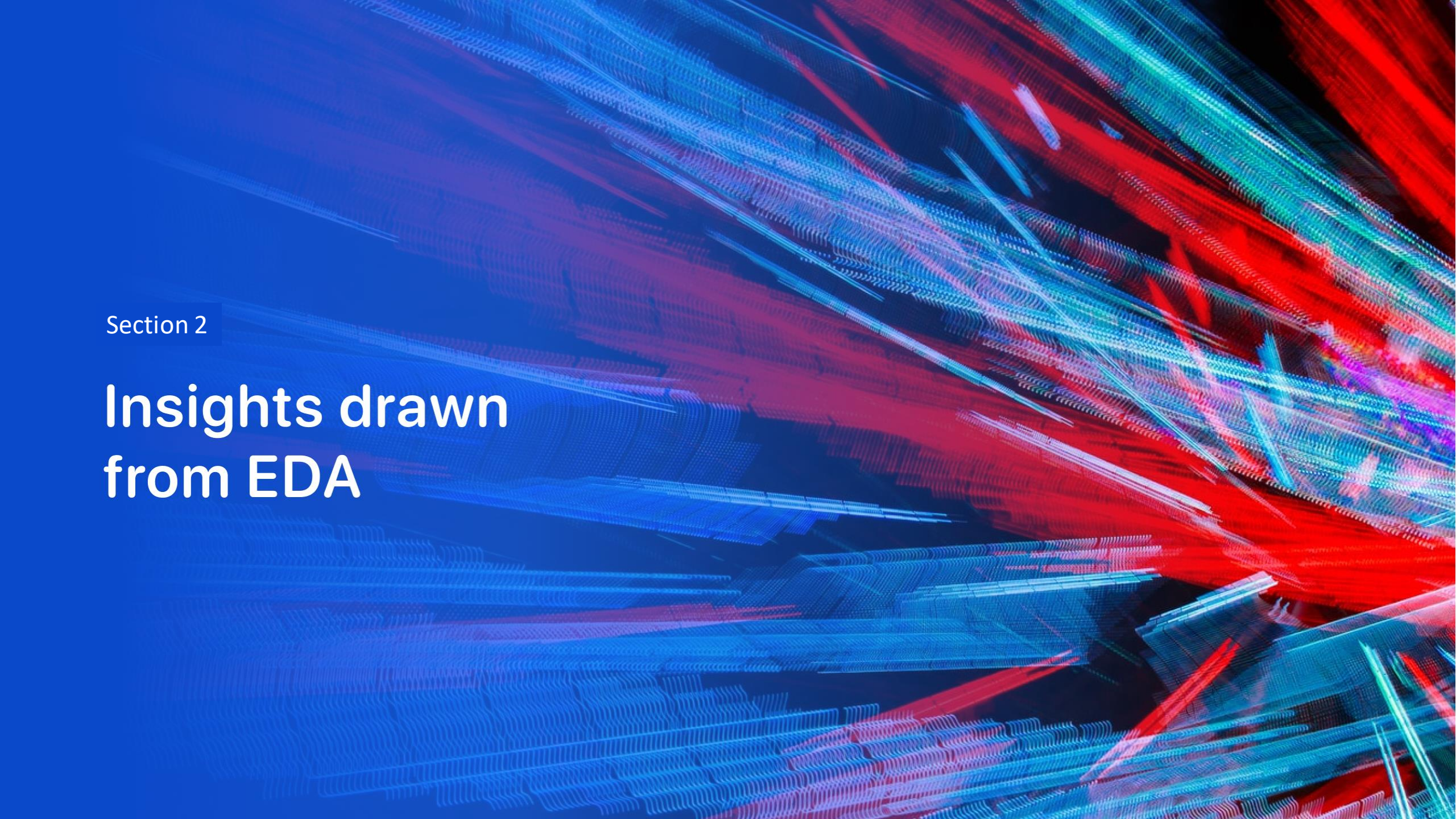
https://github.com/lewisjjoyce/IBMCapstone/blob/main/Part%208%20-%20Machine%20Learning%20Predictions.ipynb

Results

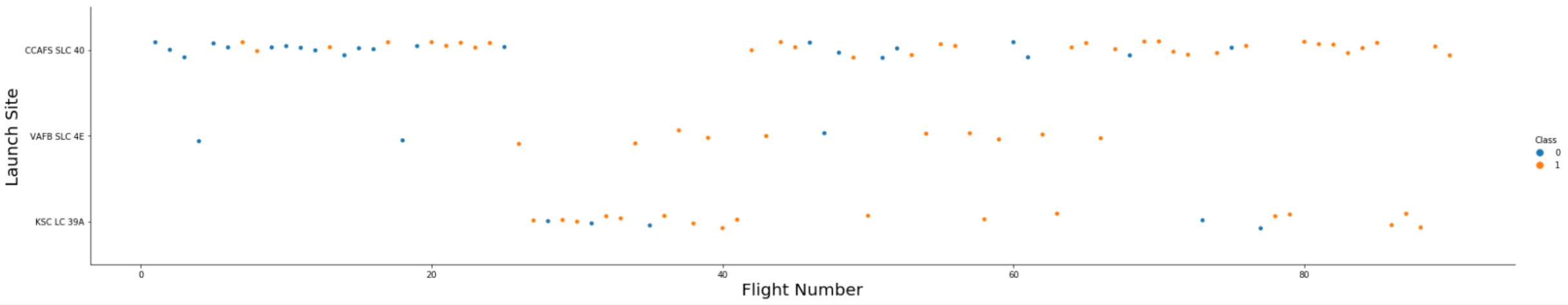
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

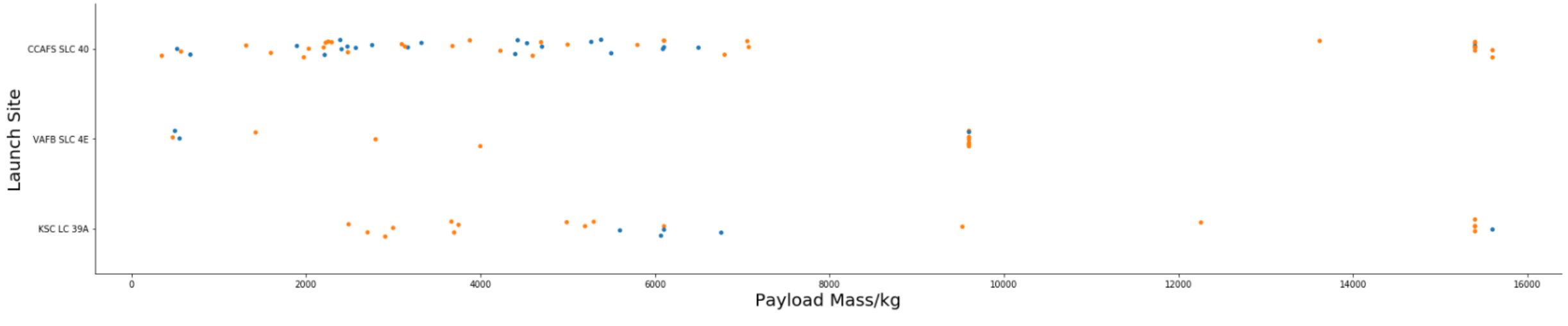


Flight Number vs. Launch Site



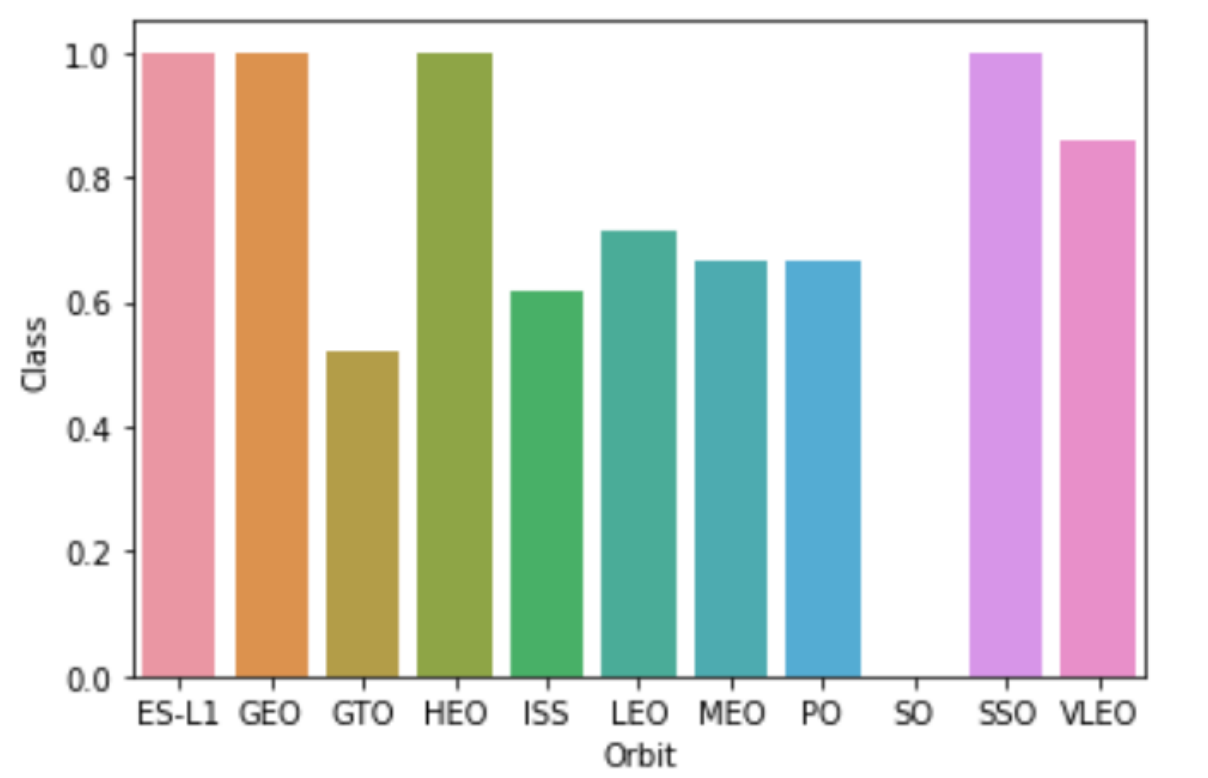
This plot shows that the earlier flights launching from CCAFS SLC 40 had a high failure rate and that over time they have become more successful. It can also be seen that VAFB SLC 4E and KSC LE 39A have higher success rates. The success rate as a whole increases after around 25 flights, which is coincides with the increased use of the other two launch sites. Thus, it is unclear whether the higher success rate at the VAFB and KSC sites can be attributed to the location or simply SpaceX's increased experience at launching Falcon 9 rockets.

Payload vs. Launch Site



From this plot, we can see that heavier payloads have a higher success rate in landing the first stage of Falcon 9 rockets. Payloads exceeding 10000kg have thus far only been launched from the CCAFS and KSC sites.

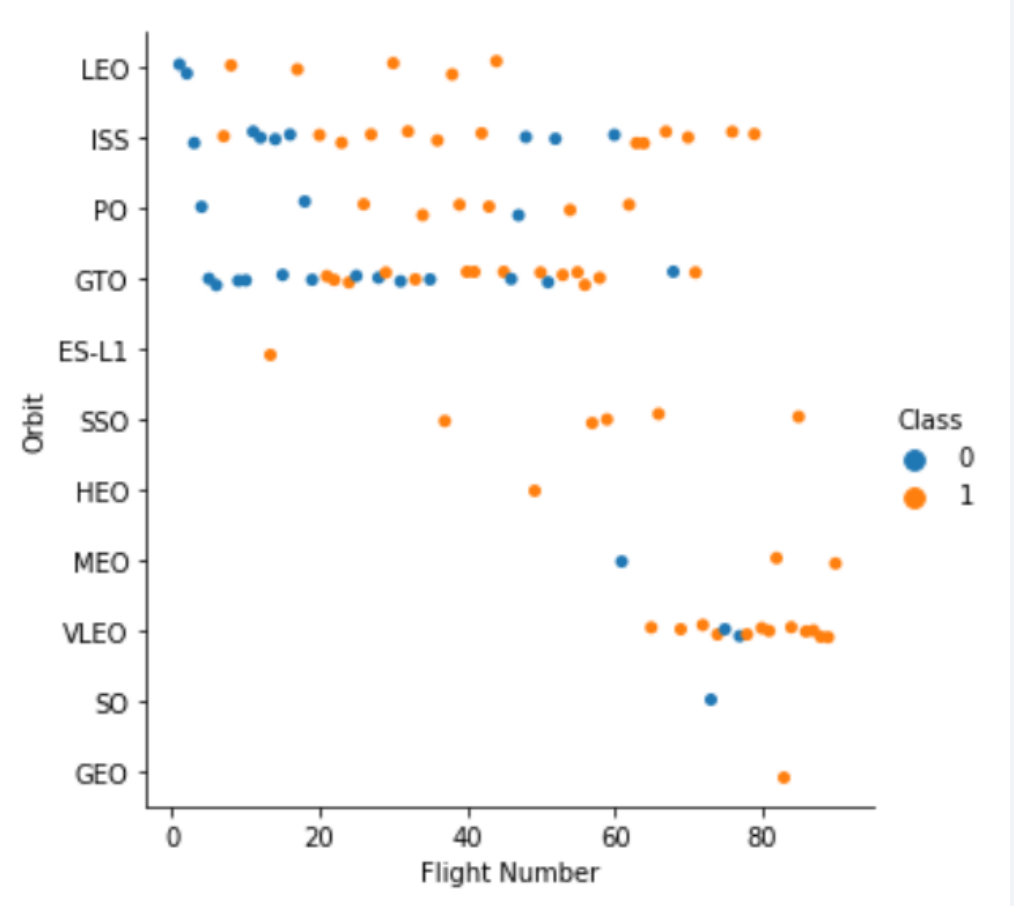
Success Rate vs. Orbit Type



Orbit	Number of Launches
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
SO	1
GEO	1
HEO	1
ES-L1	1

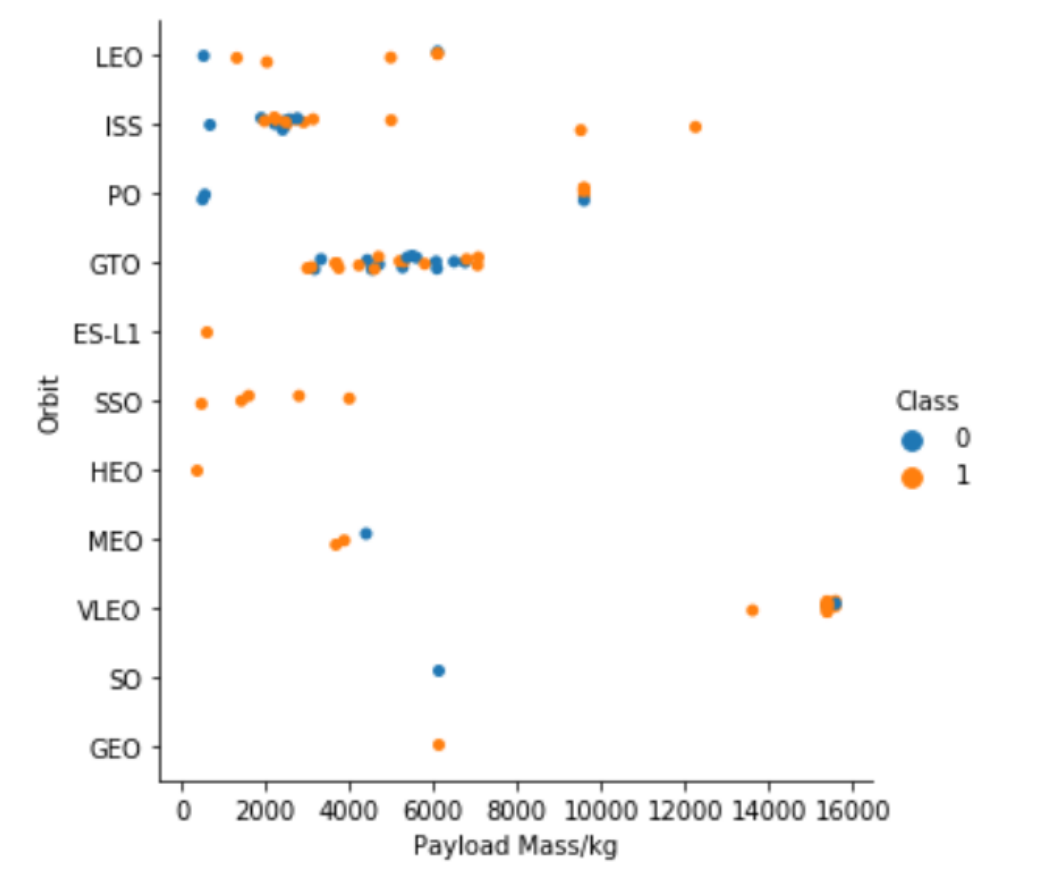
The most successful orbits have been ES-L1, GEO, HEO and SSO with a 100% success rate. However, only a single launch for the first three listed has been recorded and five for SSO, so the sample size is too small to be conclusive. The GTO orbit has the lowest success rate and is also the mode. VLEO has a high success rate, greater than 80%, whilst also having the third highest number of launches, indicating that these orbits seem to correlate to positive outcomes.

Flight Number vs. Orbit Type



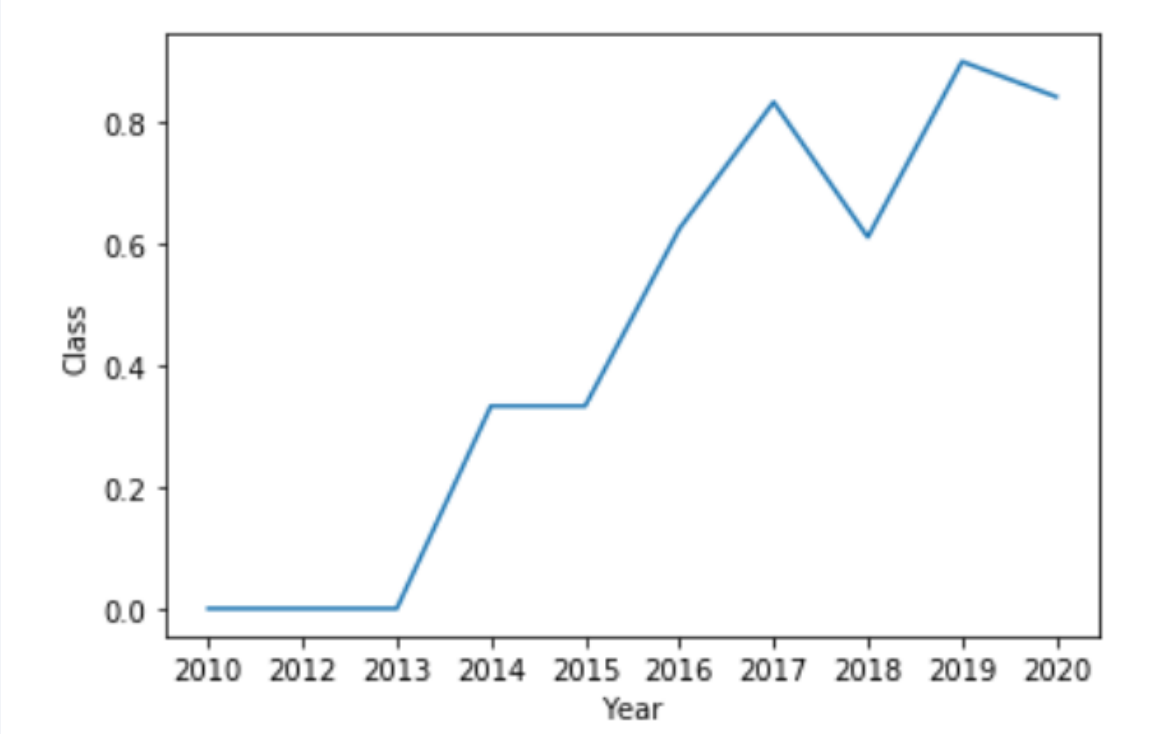
We can see that over time, SpaceX has increased the quantity of VLEO orbits, which have a high success rate. From LEO to GTO orbits, there was initially a high failure rate which has improved over time, however, the GTO launches have retained a higher rate of failure than the others.

Payload vs. Orbit Type



For ISS and PO, heavier launches are more successful. The heaviest launches are reserved for VLEOs. In previous slides, we've observed a correlation between both heavy payloads and VLEOs with higher success rate, which is supported by this plot. GTOs appear to have no correlation between the mass and success.

Launch Success Yearly Trend



This shows that SpaceX Falcon9 rockets have been launched, generally, with increasing success over time, with the first successful launch coming after 4 years in 2014. With the exception of 2018 and 2020, SpaceX has seen a yearly increase in successful launches.

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

This query returns all unique names of launch sites. This shows that there are four sites and that some rows have missing (None) values.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE '%CCA%' LIMIT 5;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

This query returns the first five records for which the launch site contains the characters 'CCA'.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total Payload Mass' FROM (SELECT * FROM SPACEXTBL WHERE Customer = 'NASA (CRS)')
```

Total Payload Mass
45596.0

This query returns the total payload mass for boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average Payload Mass' FROM (SELECT * FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1')
```

Average Payload Mass
2928.4

This query returns the average payload mass launched by booster version F9 v1.1

First Successful Ground Landing Date

```
%%sql SELECT MIN(SUBSTR(DATE,7,4)) AS 'Date of First Successful Landing Outcome' FROM
      (SELECT * FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success%')
```

Date of First Successful Landing Outcome
2015

This query returns the date of the first successful launch outcome.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_Version,PAYLOAD_MASS__KG_,Landing_Outcome FROM SPACEXTBL WHERE
      Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696.0	Success (drone ship)
F9 FT B1026	4600.0	Success (drone ship)
F9 FT B1021.2	5300.0	Success (drone ship)
F9 FT B1031.2	5200.0	Success (drone ship)

This query returns the booster versions that successfully landed on drone ships with a payload between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(*) AS 'Count', Mission_Outcome FROM SPACEXTBL WHERE NOT Mission_Outcome = 'None' GROUP BY Mission_Outcome
```

Count	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

This query returns a count of all of the success and failure outcomes.

Boosters Carried Maximum Payload

```
%%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

This query returns a list of booster versions that carried the heaviest payload.

2015 Launch Records

```
%%sql SELECT SUBSTR(Date,4,2) as 'Month', Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
      WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,7,4)='2015'
```

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query lists the months in 2015 for which there was a failed landing on a drone ship.

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

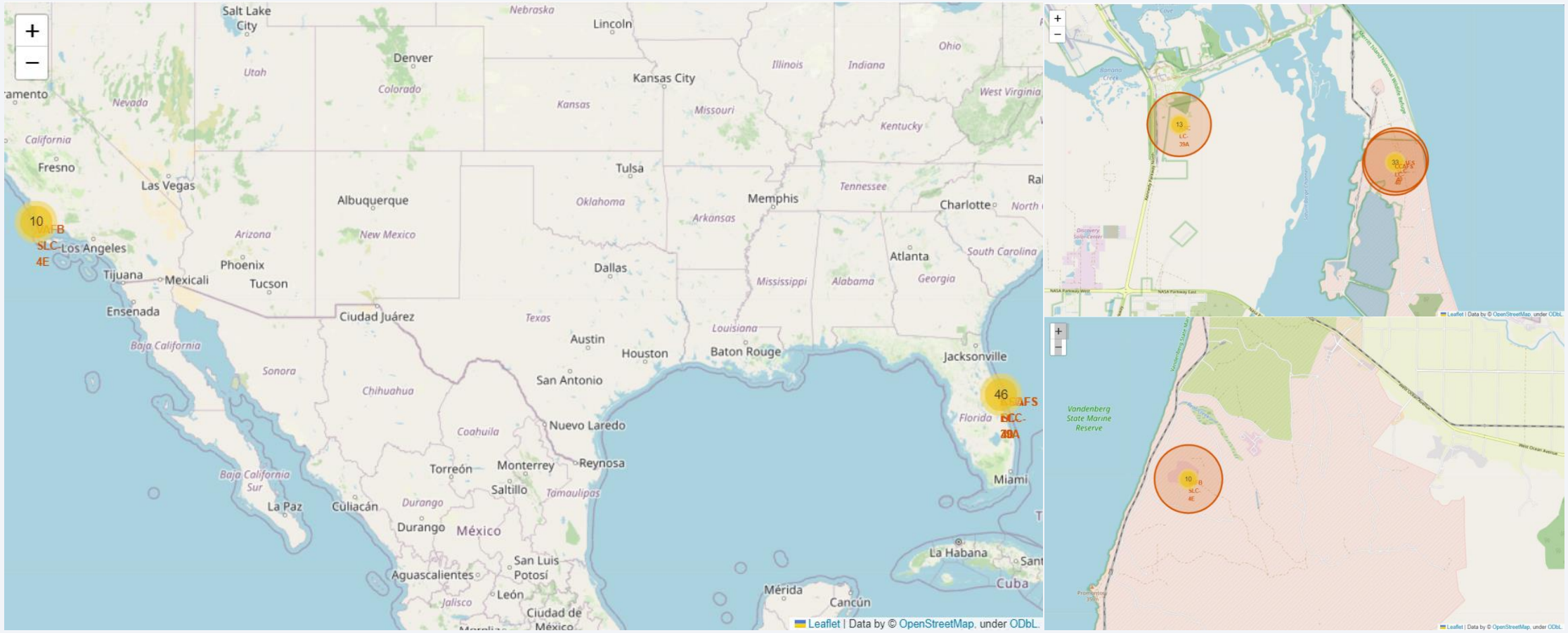
```
SELECT COUNT(*) AS 'Count', Landing_Outcome FROM SPACEXTBL WHERE Date BETWEEN '04/06/2010' AND '20/03/2017' AND Landing_Outcome
LIKE '%Success%' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
```

Count	Landing_Outcome
20	Success
8	Success (drone ship)
7	Success (ground pad)

This query returns an ordered list of the types of success outcome between 4th June 2010 and 20th March 2020.

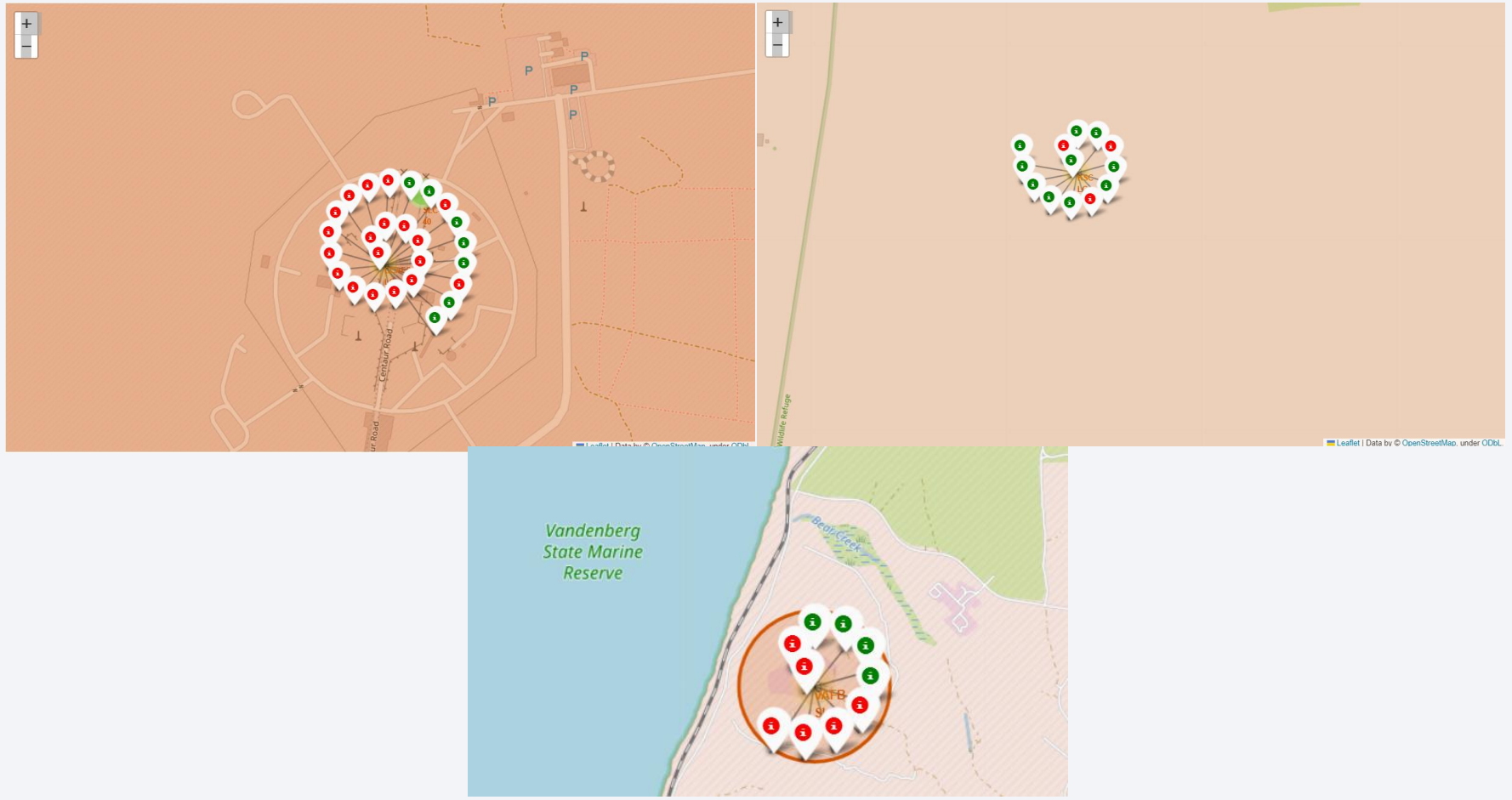

Section 3
Launch Sites
Proximities Analysis

A map of launch sites



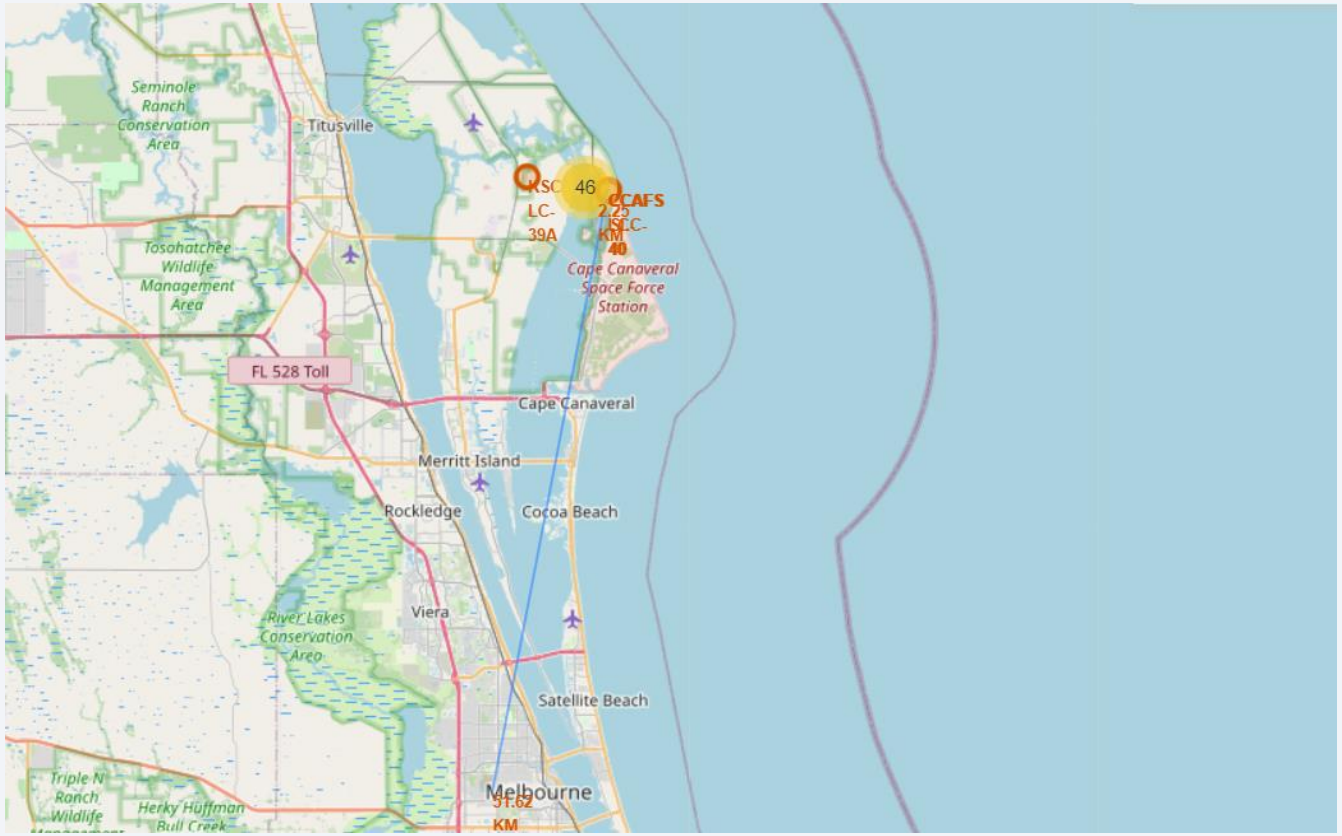
This map geographically displays the number of Falcon 9 launches that took place at each site.

Launch Outcomes

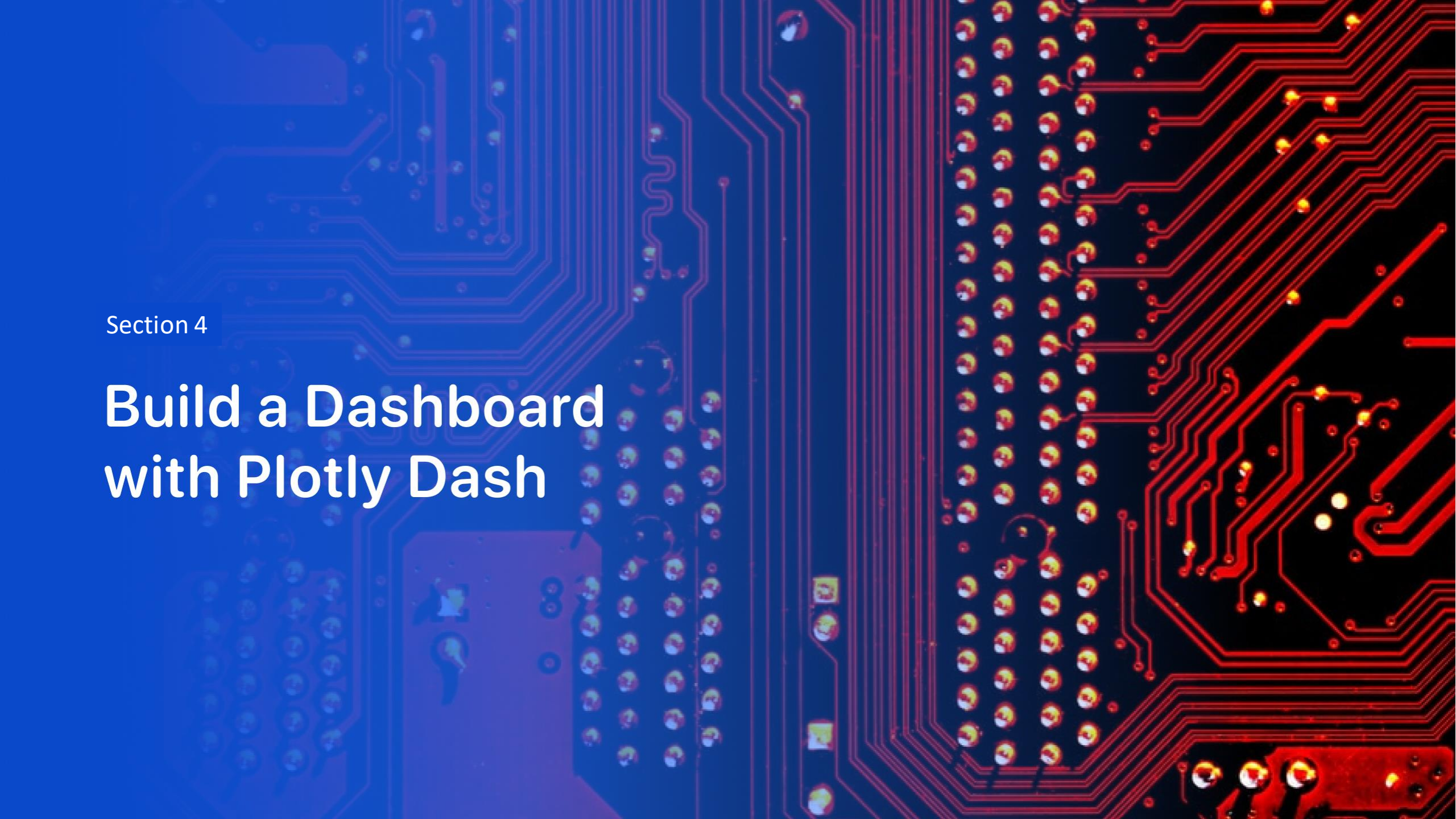


By clicking on the area surrounding the launch site, markers become visible that indicate the outcome of each launch.

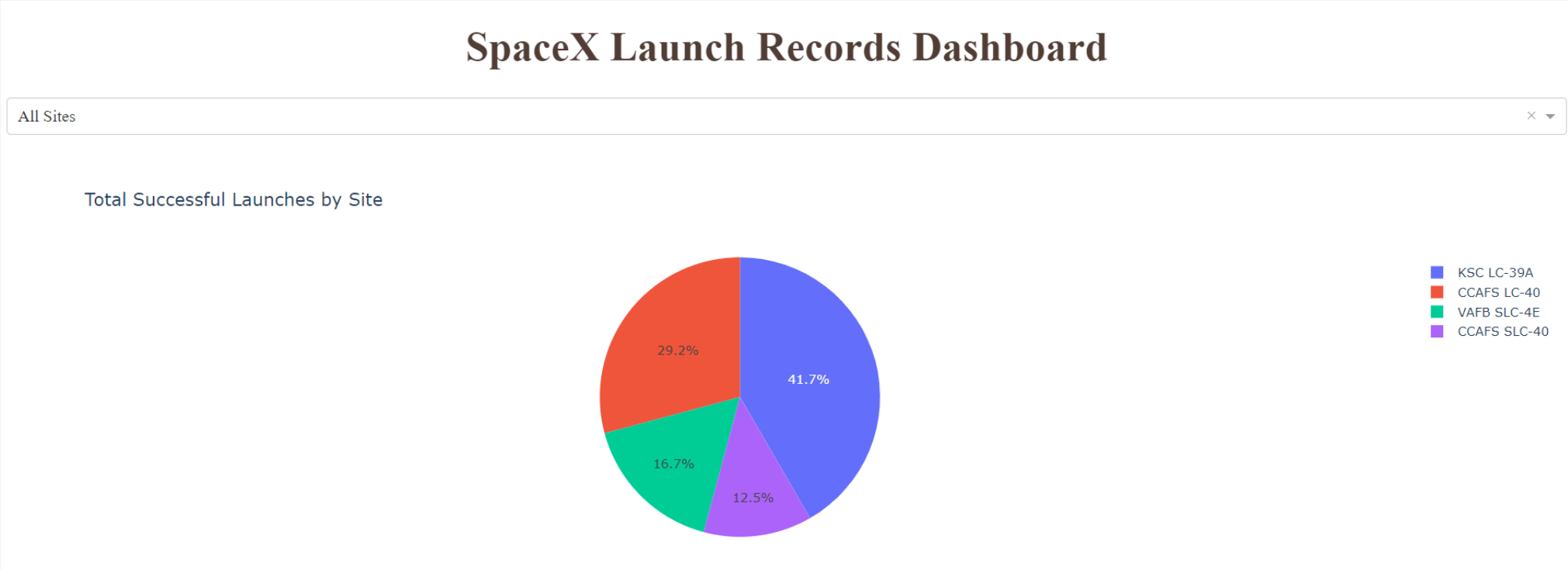
Proximities to geographical features



This map shows the proximity of the CCAFS 40 SLC launch site to the nearest city.

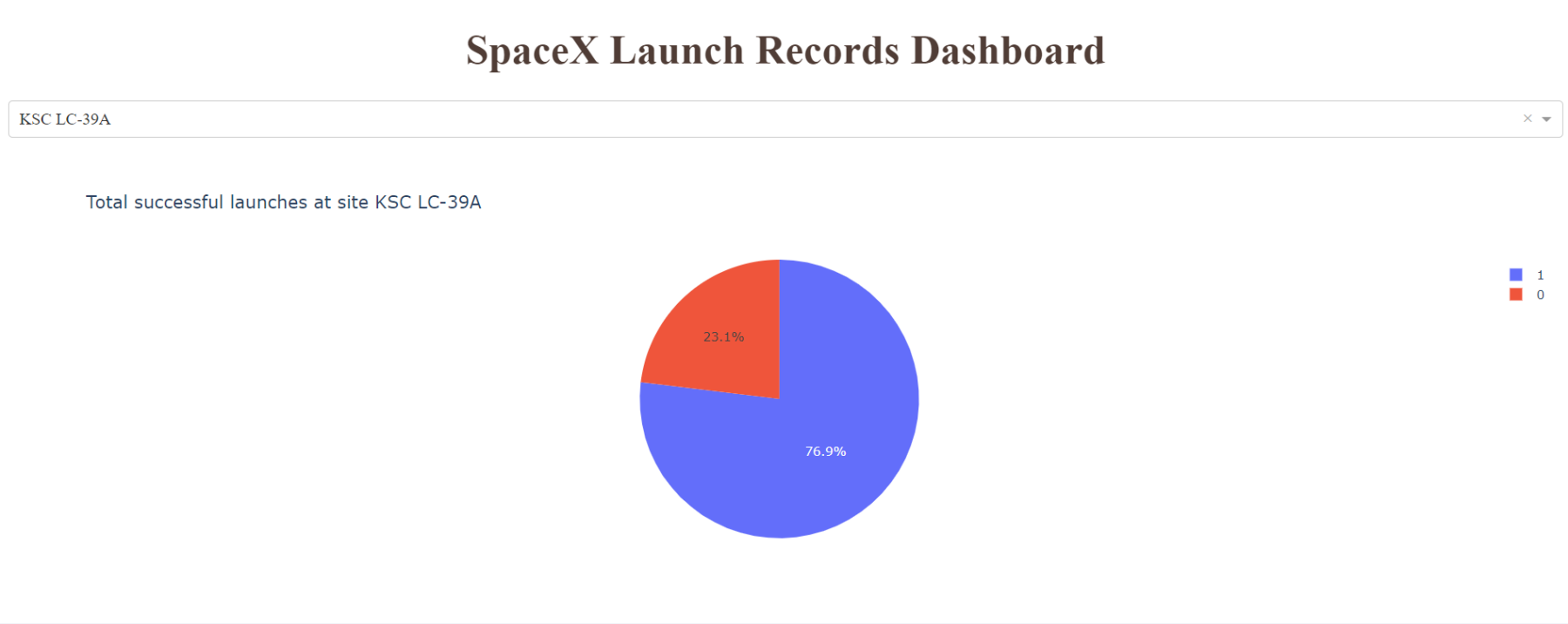

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Success rate for each launch site



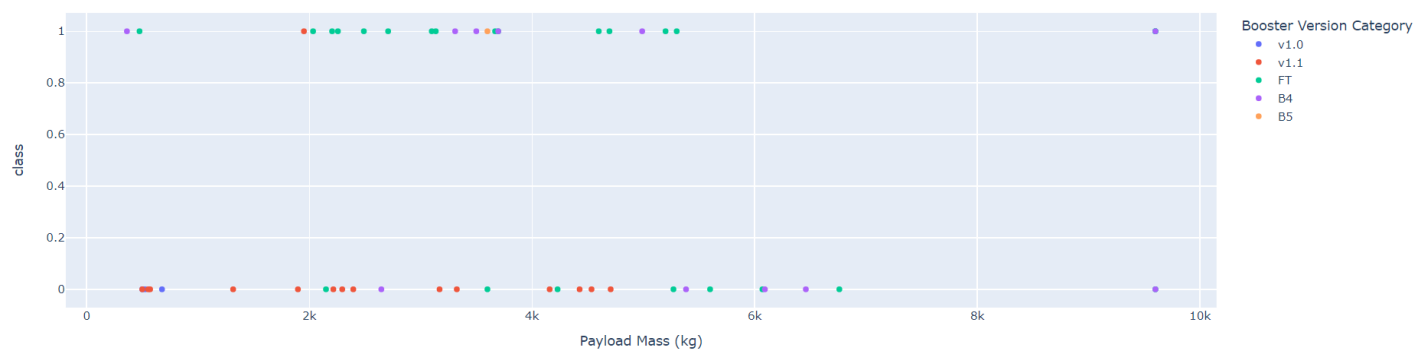
This chart displays the percentage of successful launches that occurred at each site. The KSC site had the highest proportion of successful launches, whilst the CCAFS SLC-40 has the lowest.

Highest Success rate Pie Chart



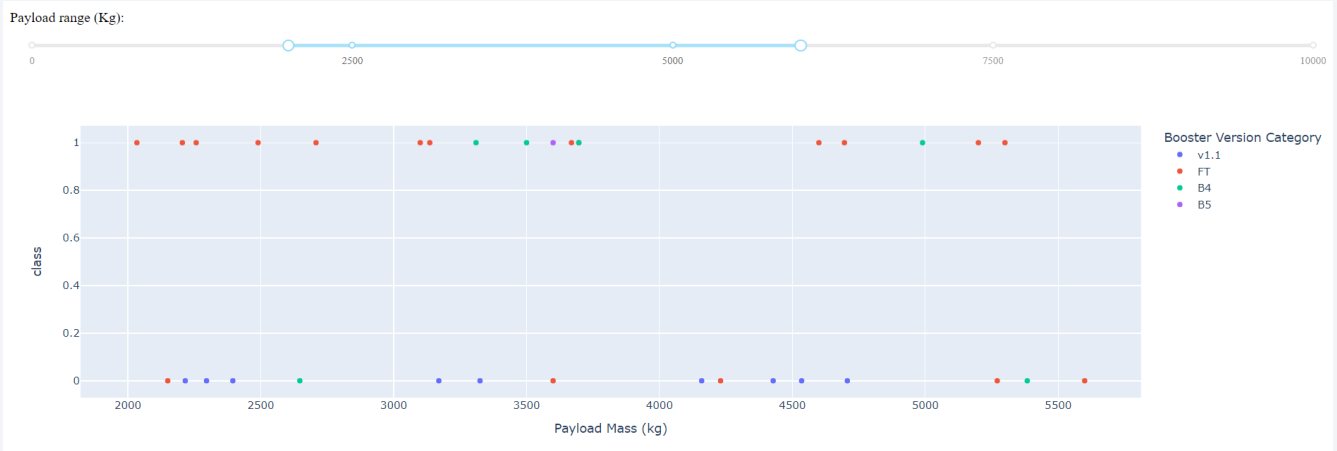
This plot shows the proportion of successes and failures at the best launch site, KSC.

Payload vs Class for different ranges



This plot shows the full range of payloads launched. The majority launched in the 2000kg to 6000kg range. The heavier payloads (greater than 6000kg) were carried by the FT and the B4 booster.

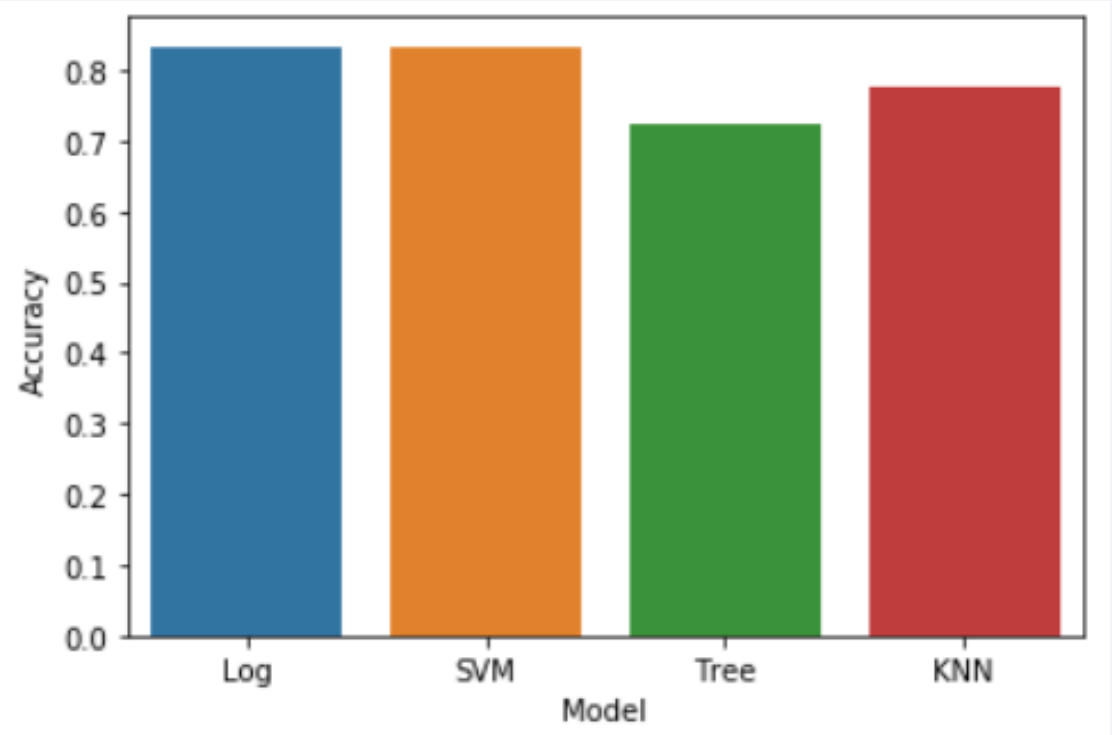
Zooming in to the 2000kg to 6000kg range, we can see that the most successful booster is the FT version, whereas the least successful is v1.1.



Section 5

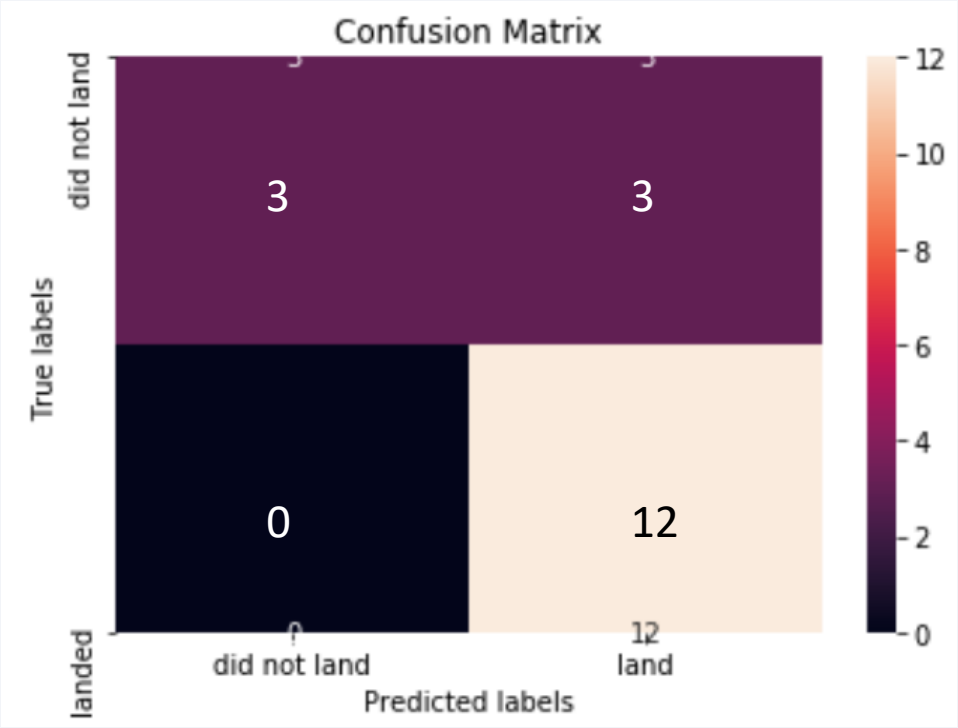
Predictive Analysis (Classification)

Classification Accuracy



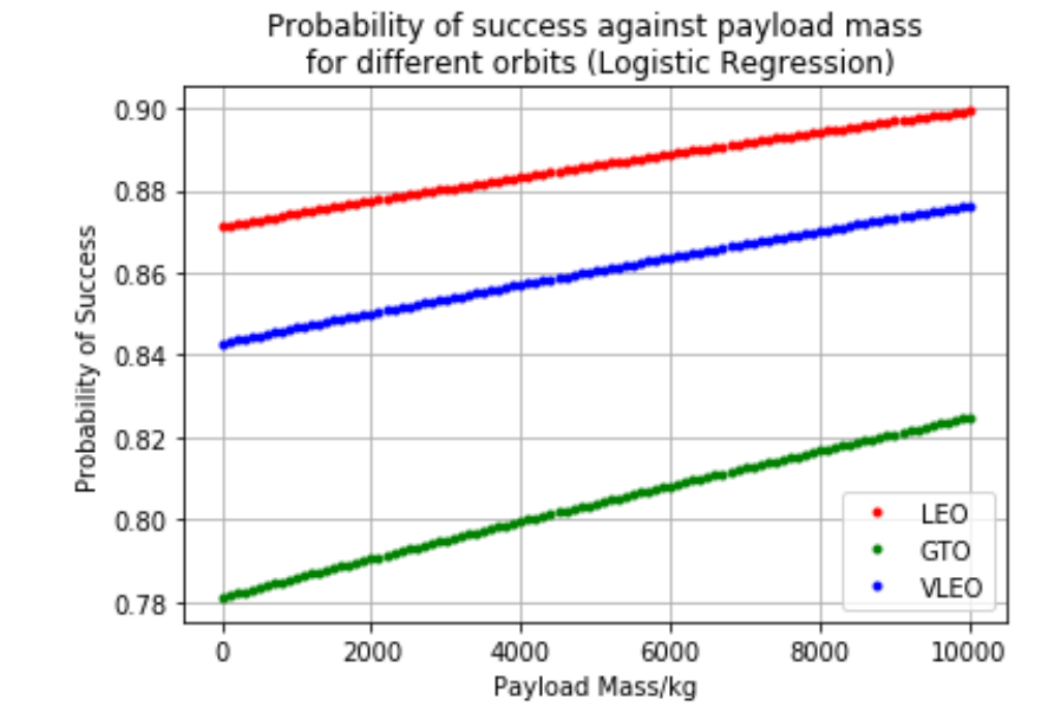
This plot shows the accuracy of each model. The logistic regression and SVM models are most accurate at 83.3%

Confusion Matrix



This is the logistic regression confusion matrix. It shows that it correctly labelled three failed outcomes and twelve successful outcomes. It output three false positives and zero false negatives.

Further Insight – Logistic Regression Payload Predictions



Orbit	PayloadMass	Class
ES-L1	570.000000	1.000000
GEO	6104.959412	1.000000
GTO	5011.994444	0.518519
HEO	350.000000	1.000000
ISS	3279.938095	0.619048
LEO	3882.839748	0.714286
MEO	3987.000000	0.666667
PO	7583.666667	0.666667
SO	6104.959412	0.000000
SSO	2060.000000	1.000000
VLEO	15315.714286	0.857143

Mean Payload Mass (kg) and class from dataset using .groupby() and .mean() methods.

Using the logistic regression model, we can vary the payload mass between 0 kg and 1000 kg, whilst fixing other parameters (see appendix), and observe the model's probability of returning a success outcome. The above plot shows the relationship between the payload and success for the LEO, GTO and VLEO orbits. All three orbits show a positive linear relationship with payload mass, which aligns with earlier analysis. However, when compared to the success rate from the raw data (table on the right), the model appears to overpredict the success of LEO orbits. The confusion matrix of the previous slide showed that the model's inaccuracies tend to be false postives, which appears to also be the case for LEOs.

Conclusions

- Spacex's success rate has generally increased from 2013 to 2020
- Heavier payloads tend to have a higher rate of success
- VLEOs are the most successful type of orbit with a sample size > 5.
- The KSC launch site has seen the most successful landings
- The FT booster version has the highest number of successful landings
- A logistic regression model or SVM provide the most accurate predictions at 83.3%

Appendix

Page 45 – Data for logistic regression predictions – All values except for the payload mass and orbit were fixed.

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount
0	44.0	0.0	1.0	1.0	0.0	1.0	4.0	1.0

	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	Orbit_LEO	Orbit_MEO	Orbit_PO	Orbit_SO	Orbit_SSO	Orbit_VLEO
0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0

	LaunchSite_CCAFS SLC 40	LaunchSite_KSC LC 39A	LaunchSite_VAFB SLC 4E
0	1.0	0.0	0.0

Serial_B1042	Serial_B1043	Serial_B1044	Serial_B1045	Serial_B1046	Serial_B1047
0.0	1.0	0.0	0.0	0.0	0.0

Thank you!