Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- SpaceY aims to compete with SpaceX as a commercial rocket launch provider.
- SpaceX offers launch services starting at $62 million, assuming the first-stage booster is recovered and reused.
- The estimated cost of building a Falcon 9 first-stage booster (excluding R&D and profit) is around $15 million.
- A predictive model was developed to estimate first-stage landing success with 83.3% accuracy, based on payload mass and target orbit.
- This allows SpaceY to make more informed and competitive bids by using landing predictions as a proxy for launch cost.

Introduction

- This report has been prepared as part of the IBM Data Science learning
- In the Applied Data Science Capstone course the role of a data scientist working for a new rocket company called has been trusted to me
- With the help of the data science findings and models in this report, SpaceY will be able to make more informed bids against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

The following data science methodologies have been used:

- Data collection
- Data wrangling
- Exploratory data analysis
- Data visualization
- Model development
- Reporting results to stakeholders

Data Collection

Past launches

2010 to 2013

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb) [19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	

- The data has been acquired from 2 publicly available sources
  - Wikipedia – of html formt
  - Opensoure SpaceX API – of JSON format
- Both required preprocessing by using dedicated programming libraries
- Falcon 9 data was a subject for further analysis

Data Collection – SpaceX API

- Steps to data processing from API:
  - Request the JSON data from the link https://api.spacexdata.com/v4/launches/past
  - Filter the data include Falcon 9 launches only
  - Missing data was handled by using mean value

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40

Example of the data prepared

** Results of the data analysis concerning API processing is [available here]*

Data Collection - Scraping

- Steps to data processing from wikipedia:
  - Request the html data from the link https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Parse the tables in order to create data frame
  - Final dataset contained 121 records regarding Falcon 9 first stage boosters

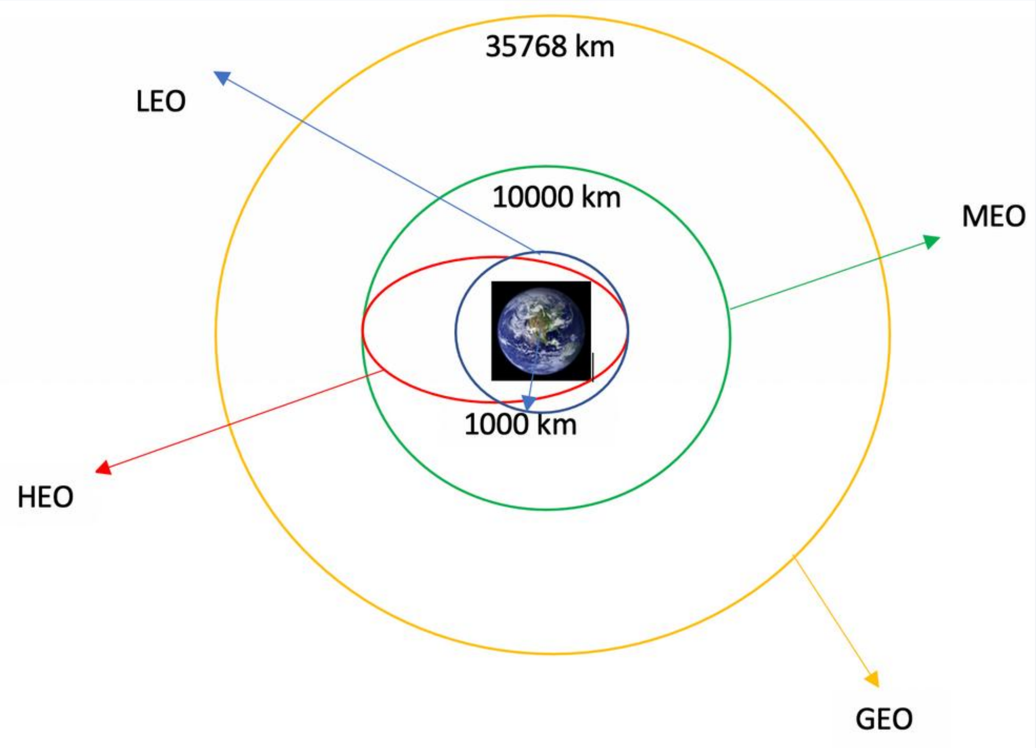
** Results of the data analysis concerning Webscrapping processing is [available here]()*

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)
...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak

Example of the data prepared

Data Wrangling

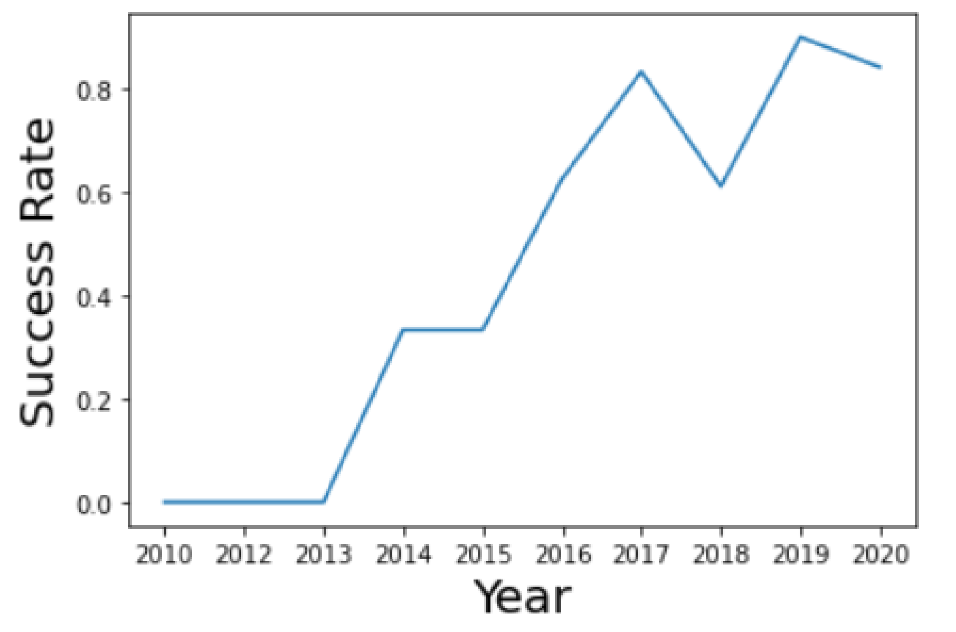
- Data wrangling consisted of the following activities:
  - Calculation the number of launches on each site
  - Calculation the number and occurrence of each orbit
  - Calculation the number and occurence of mission outcome per orbit type
  - Calculation of the landing outcome label from Outcome column (0.66)



** Results of the data analysis concerning Data Wrangling processing is [available here]*

EDA with Data Visualization

- The following vizualizations were produced as the result of EDA:
  - Relationship between Flight Number and Launch Site
  - Relationship between Payload Mass and Launch Site
  - Relationship between success rate of each orbit type
  - Relationship between FlightNumber and Orbit type
  - Relationship between FlightNumber and Orbit type
  - Launch success yearly trend (see the graphics right)



** Results of the data analysis concerning EDA processing is [available here]()*

EDA with SQL 1 / 2

- The following SQL command queries were run on sqlite database (SPACEXTAB)

Operation	SQL query
Display the names of the unique launch sites in the space mission	*SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL*
Display 5 records where launch sites begin with the string 'CCA'	*SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5*
Display the total payload mass carried by boosters launched by NASA (CRS)	SELECT SUM(PAYLOAD_MASS__KG_)FROM SPACEXTBLWHERE Customer = 'NASA (CRS)'
Display average payload mass carried by booster version F9 v1.1	SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'
List the date when the first succesful landing outcome in ground pad was acheived	SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
Display the names of the unique launch sites in the space mission	*SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL*

* *Results of the data analysis concerning EDA processing is [available here]*

EDA with SQL 2 / 2

Operation	SQL query
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000	SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND 4000 < PAYLOAD_MASS__KG_ < 6000
List the total number of successful and failure mission outcomes	SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME
List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function	*SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)*
List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015	*SELECT Date, Booster_Version, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015'*

* *Results of the data analysis concerning EDA processing is [available here]()*

Build an Interactive Map with Folium

In **Folium**, both **Marker** and **Circle** are used to highlight specific geographic locations on a map, but they serve slightly different visual and functional purposes:

- **Marker** marks a specific point on the map
- **Circle** highlights a region around a point using a radius (in meters)

Line in Folium is used to draw a connected sequence of points on a map - essentially a path or route

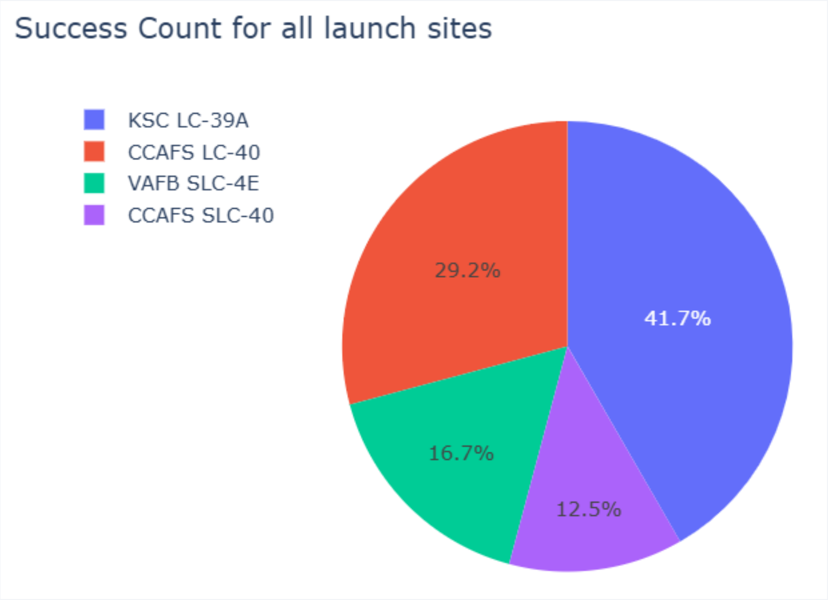
In the **Lab Markers** were used to pin the location of the NASA Johnson Space Center's coordinate and separately launch sites: Vanderberg and Cape Canaveral. Along with markers **Circle** objects were used together with Markers

Line objects was used to visualize the distance to a closest city, railway and highway from Cape Canaveral

* *See the results regarding visualization of [launch sites here](#)*

Build a Dashboard with Plotly Dash

- Plotly Dash was used to visualize analytics with regards to launch sites and success rate
- It offers the possibility to choose the site and select (narrow) the payload of the rocket.
- Based on the selected criteria a pie chart (success rate) gets displayed in real time along with scatter plot (success against the payload).



** See the results regarding visualization with Plotly visualization here*

Predictive Analysis (Classification)

- The following steps were taken in order to come up with model cevelopment
  - Data load along followed by its standarization and split into training and test
  - Fit the training data to various models
    - Logistic regression
    - Support Vector Machine
    - Decission Tree Classifier
    - Nearest Neighbour Classifier
  - Cross-validated grid-search over a variety of hyperparameters to select the best ones for each model was used
  - Evaluated accuracy of each model using test data to select the best model was evaluated

** See the predictive analysis [results here](#).*

*** Please note: it was not possible to complete the tasks fron No. 6 onward in the lab envirnment*

Results

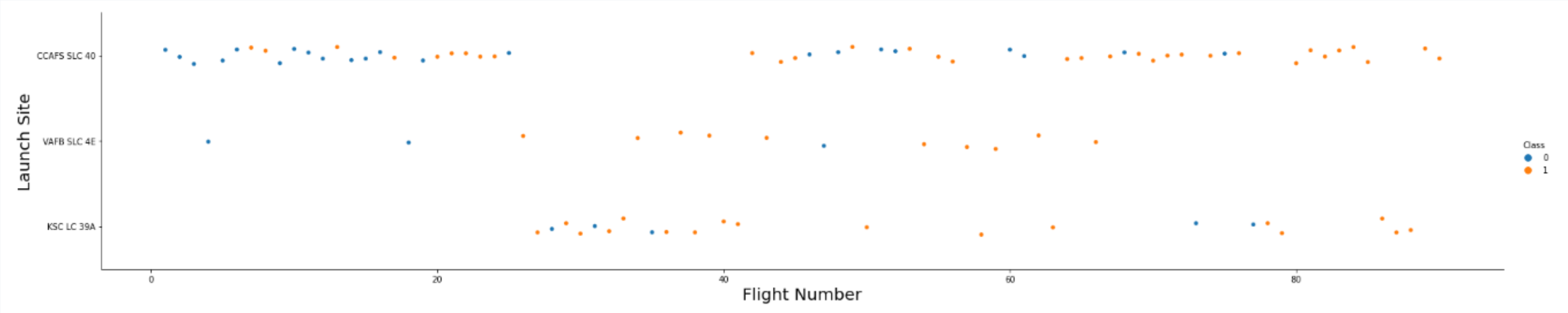
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

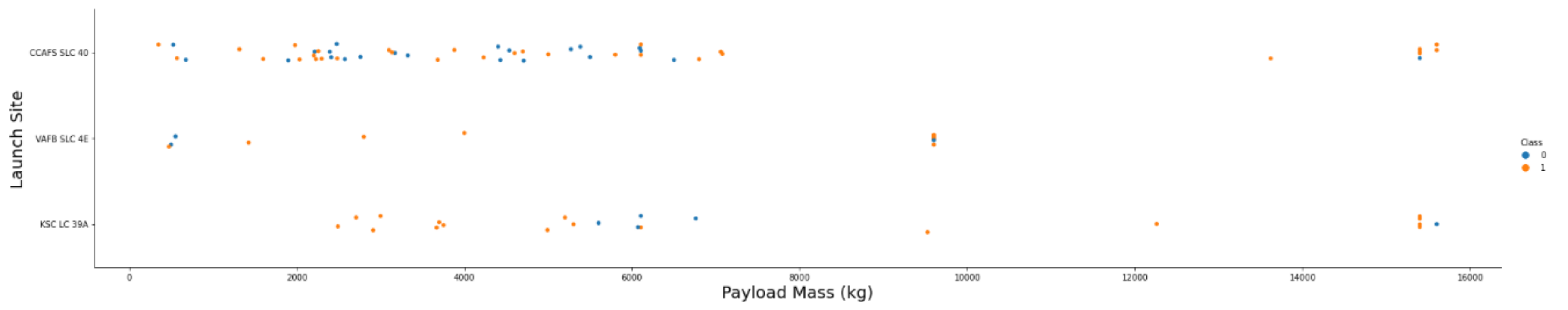
Flight Number vs. Launch Site

- The following graphics shows scatter plot of the rocket launches from different sites
- Successful and non-successful missions have been differentiated by colour



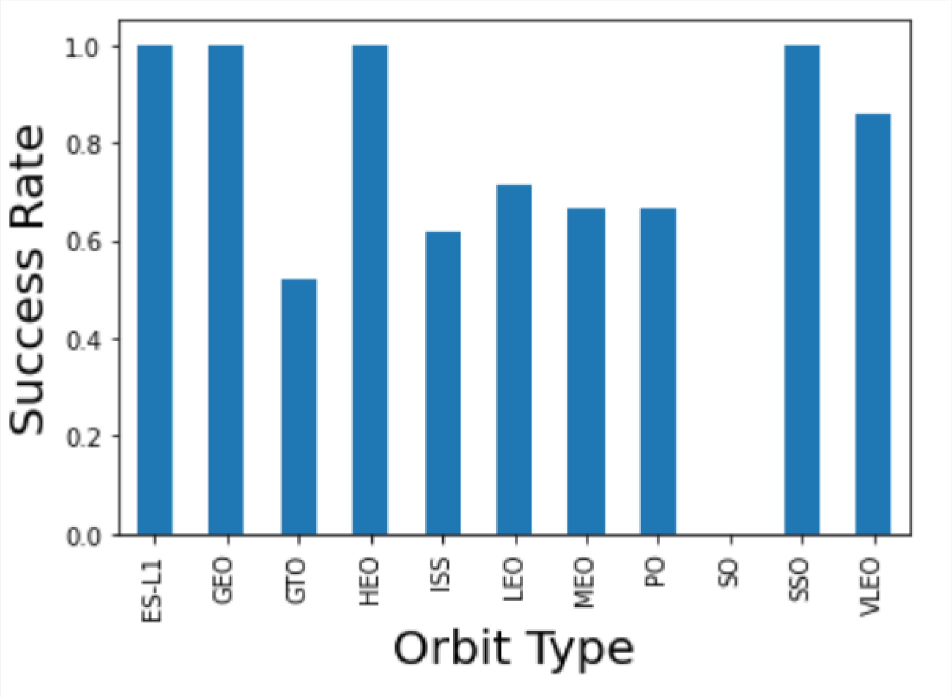
Payload vs. Launch Site

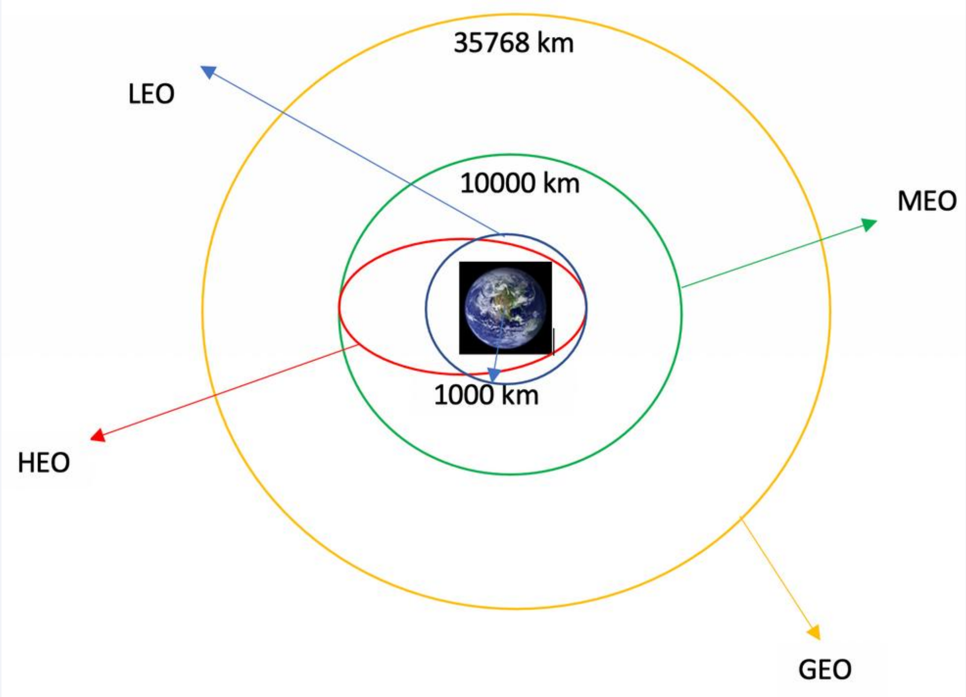
- The following graphics shows scatter plot of the rocket launches from different sites
- Successful and non-successful missions have been differentiated by colour



Success Rate vs. Orbit Type

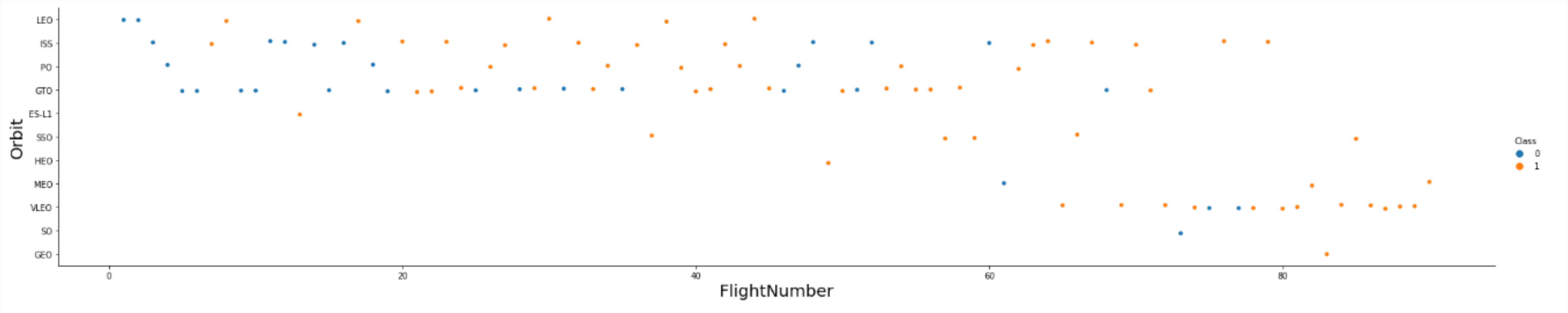
- The following graphics shows box plot of the rocket launches to different orbits
- The mission to geostationary orbit is most difficult one





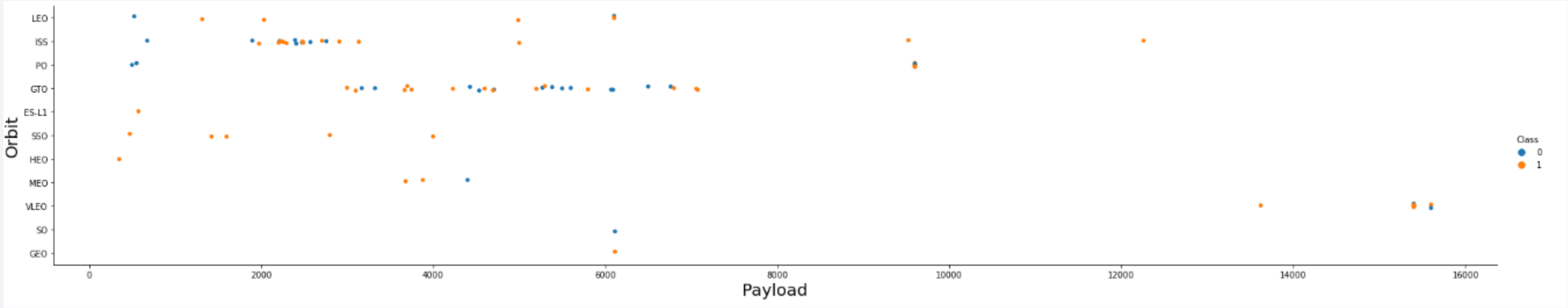
Flight Number vs. Orbit Type

- The following graphics shows scatter plot of the rocket launches to specific orbits
- Successful and non-successful missions have been differentiated by colour



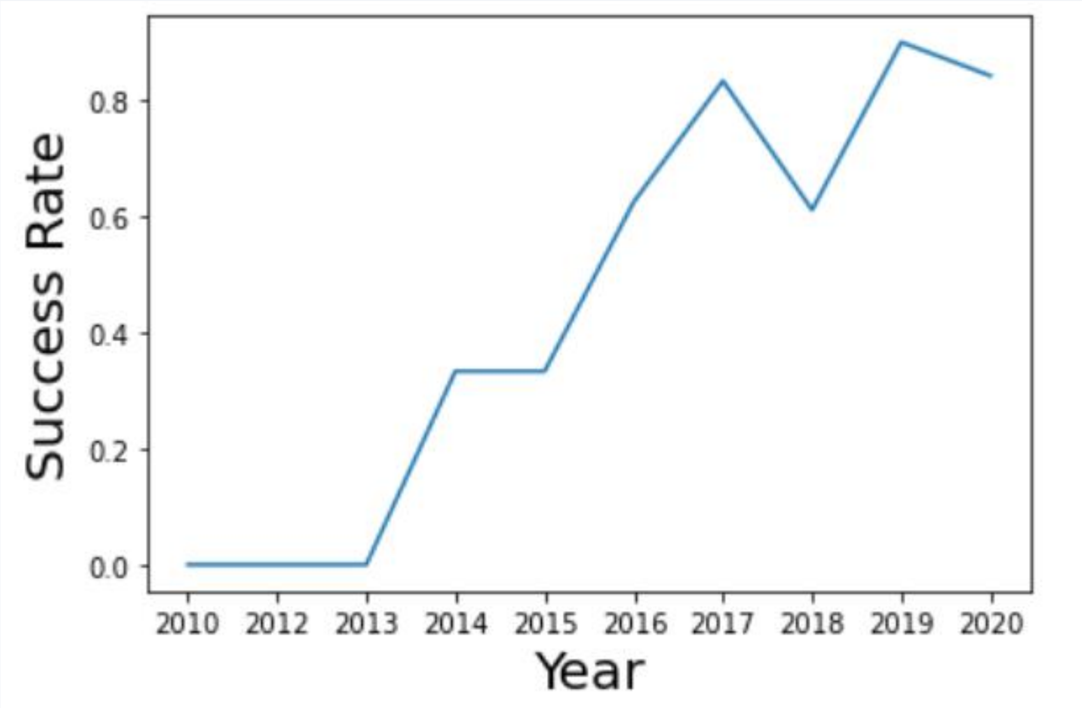
Payload vs. Orbit Type

- The following graphics shows scatter plot of the rocket launches with payload to specific orbit type
- Successful and non-successful missions have been differentiated by colour



Launch Success Yearly Trend

- The following graphics shows the trend of the launch success from 2013 onward



All Launch Site Names

- The list of the Launch sites can be found using e.g. the following query in EDA with SQL Lab:

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- Below the list of last five records of the launch sites where the name starts with ‘CCA’ (Task 2 from SQL with EDA Lab)

Task 2

Display 5 records where launch sites begin with the string 'CCA'

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

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Below the result of the Task 3 within SQL with EDA Lab

Task 3

Display the **total** payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Below the average load of a launch as the result of Task no 4 from SQL with EDA Lab

Task 4

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Below the first successful launch as the result of Task no 5 from SQL with EDA Lab

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below the query to list of boosters with payload between 4000 and 6000 successfully launched (Task no. 6 SQL with EDA Lab)
- On the right the result of the query executed

Task 6

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

```
%%sql
SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Below the number of successful and non successful as the result of Task no 7 from SQL with EDA Lab

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- Below the query to list boosters carried maximal payload (Task no. 8 SQL with EDA Lab)
- On the right the result of the query executed

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%%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

2015 Launch Records

- Below the list of missions with failed outcome as the result of Task no 9 from SQL with EDA Lab

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
SELECT Date, Booster_Version, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND
```

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

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

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

- Below the ranking of landing outcomes between 2010 and 2017 as the result of Task no 10 from SQL with EDA Lab

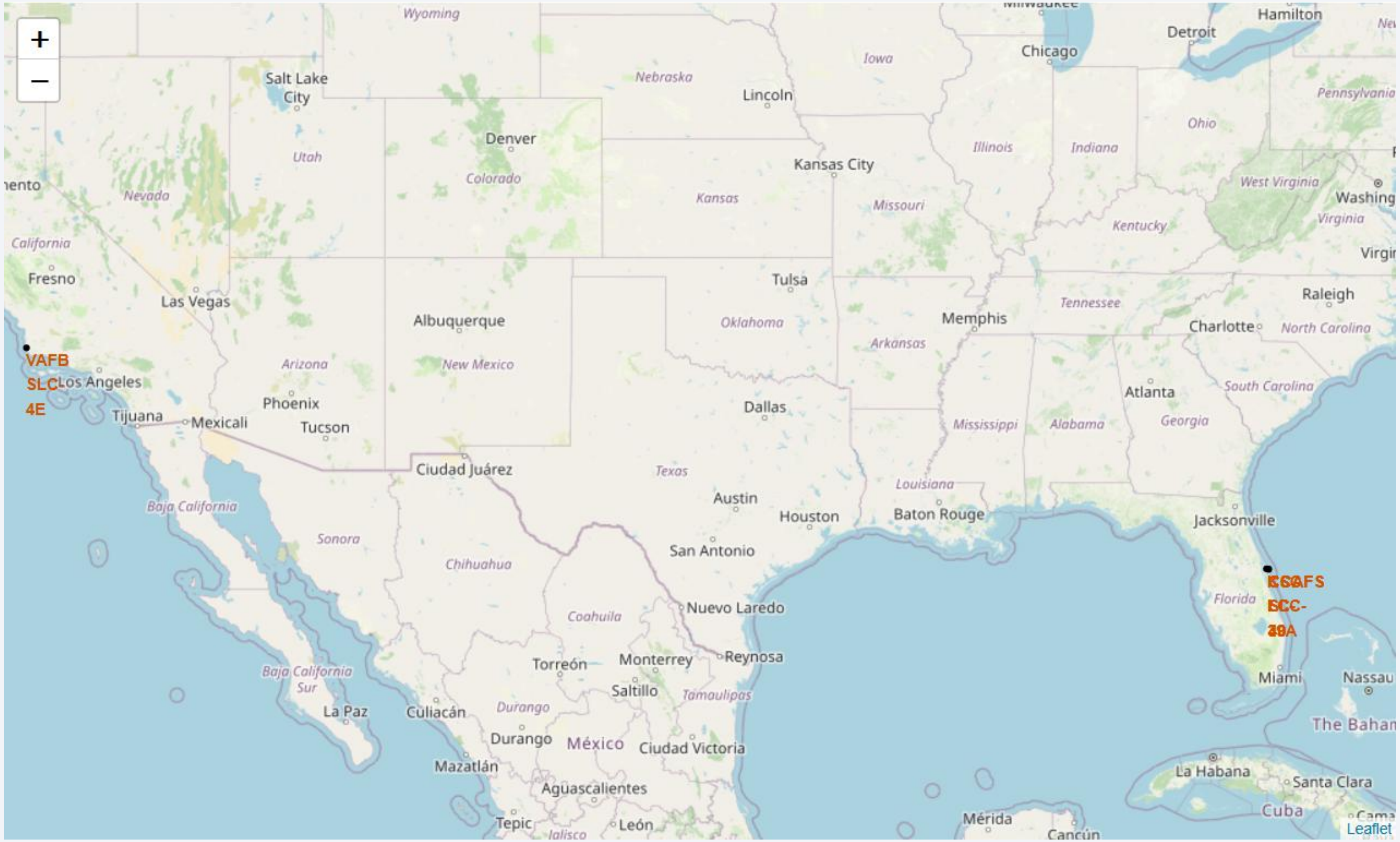
```sql
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

landing__outcome	total_number
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

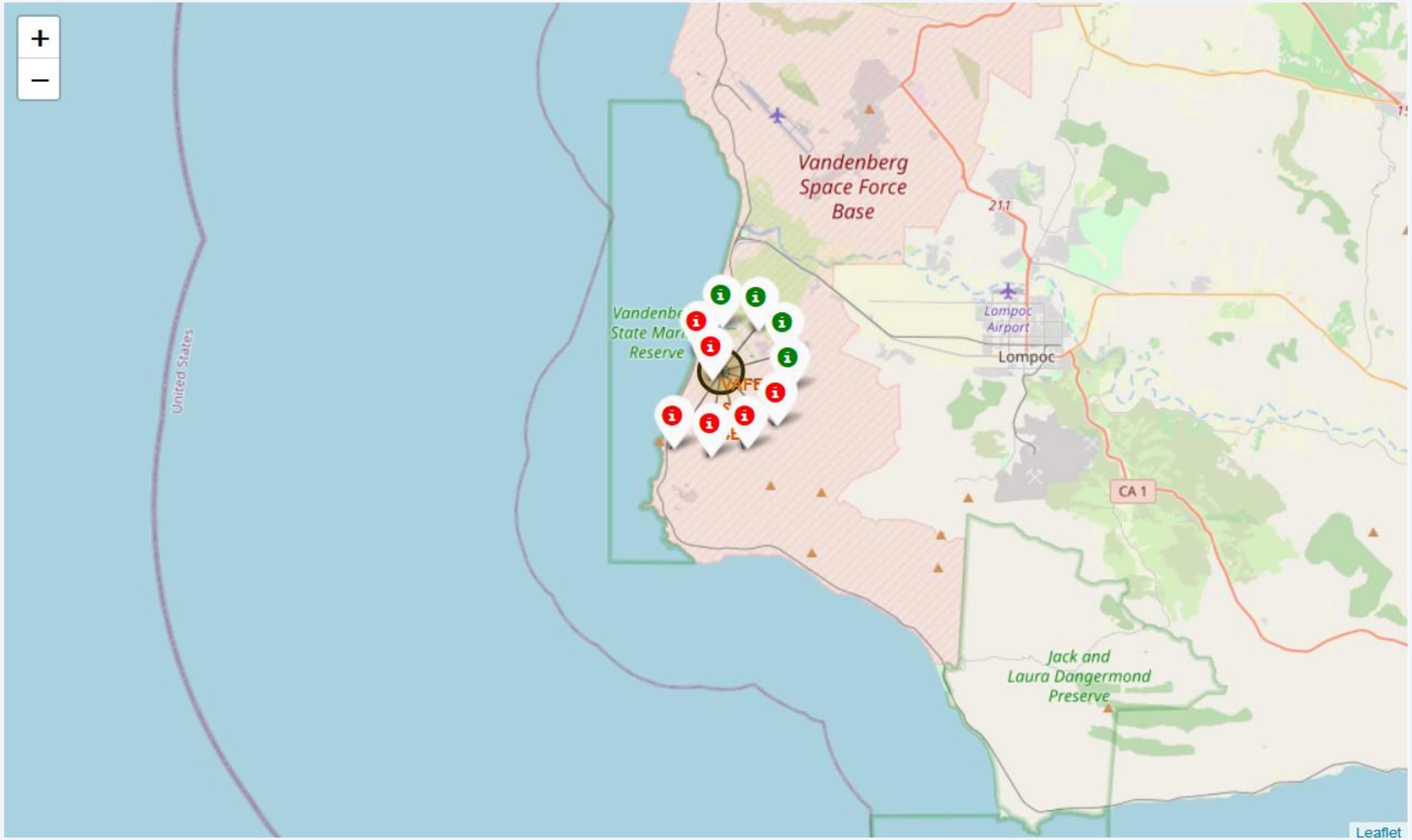
Section 3

Launch Sites Proximities Analysis

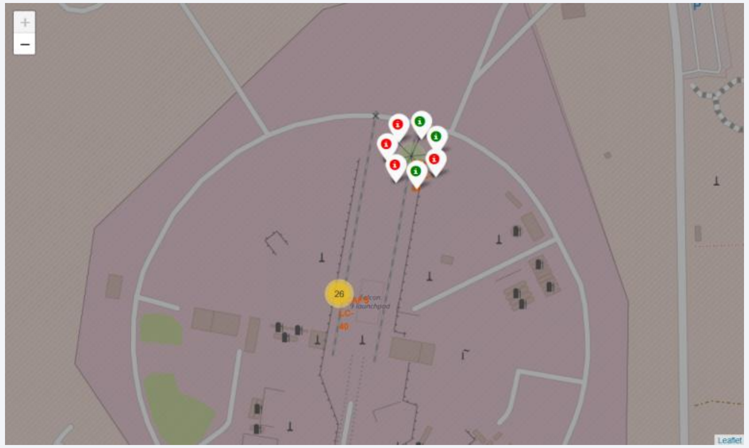
Launch sites on global map

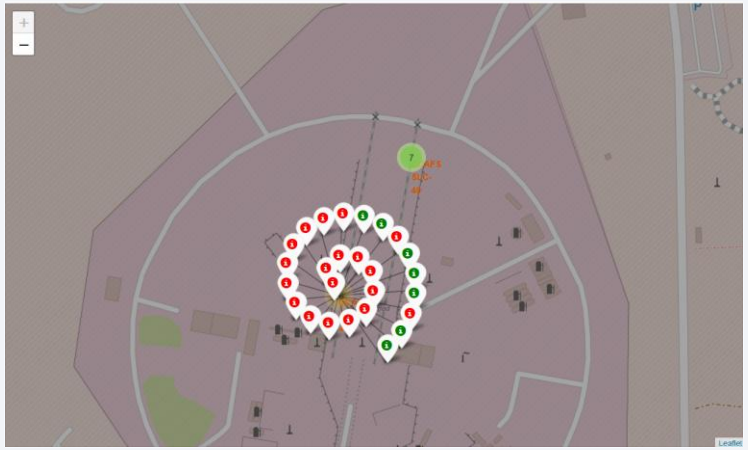


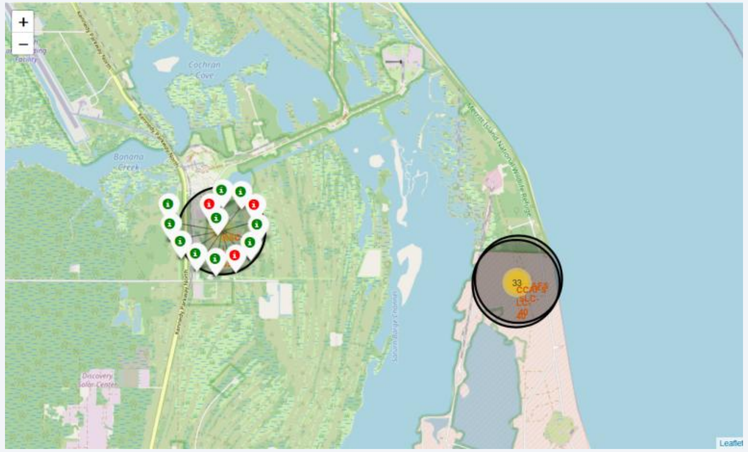
Launch outcomes of Vanderberg site



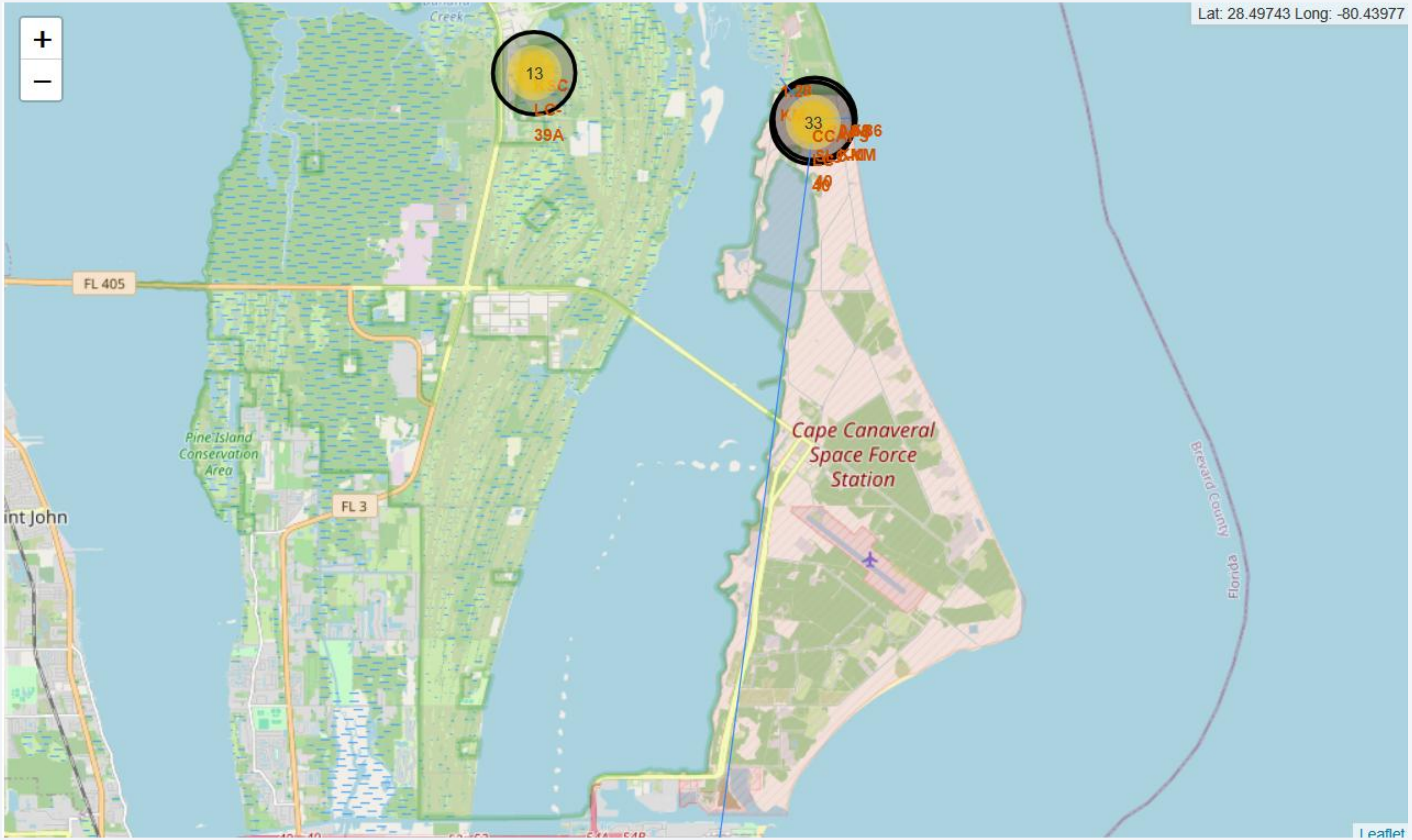
Launch outcomes of Cape Canaveral and Kannedy Space Center





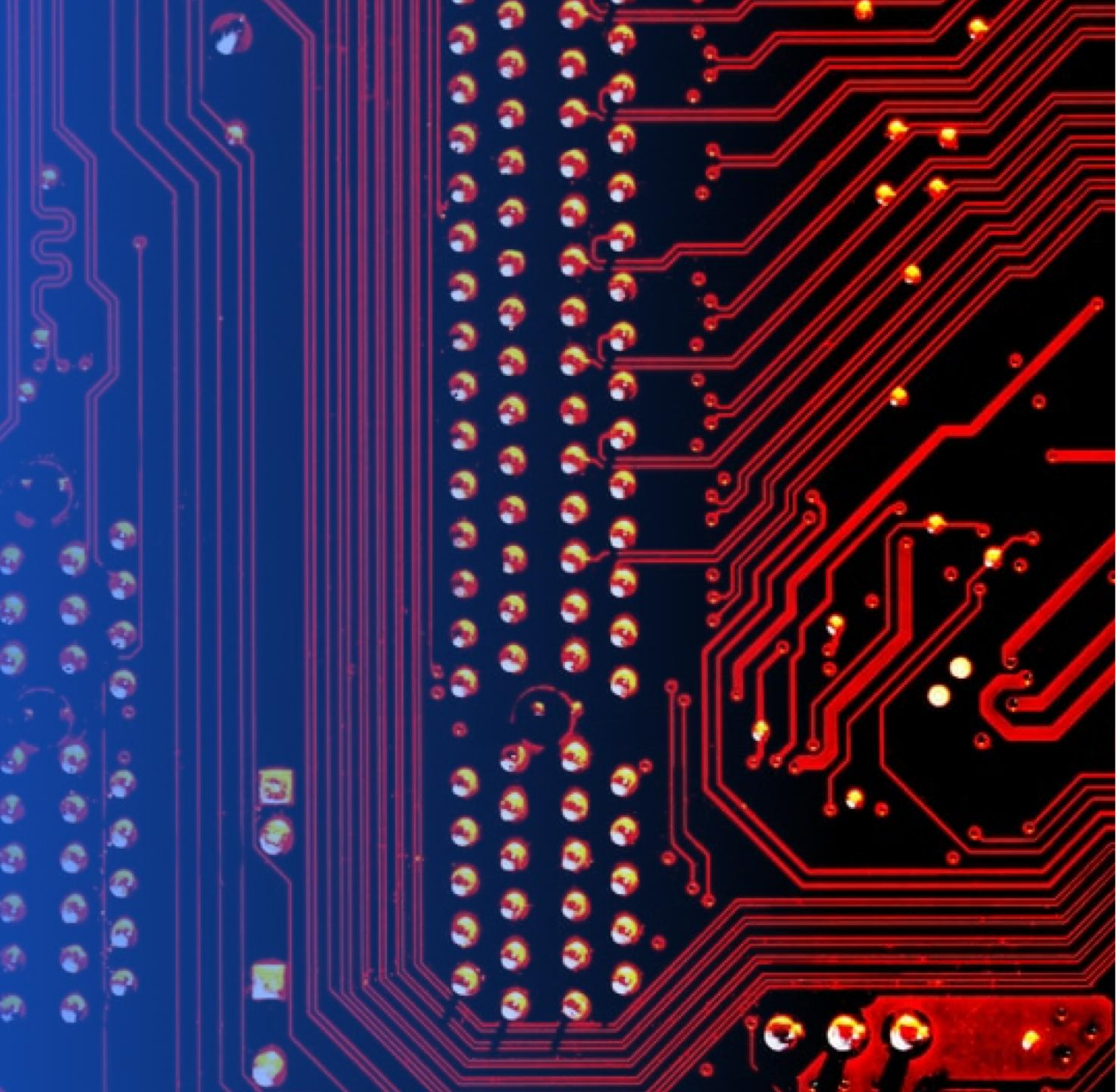


Proximities from Cape Canaveral

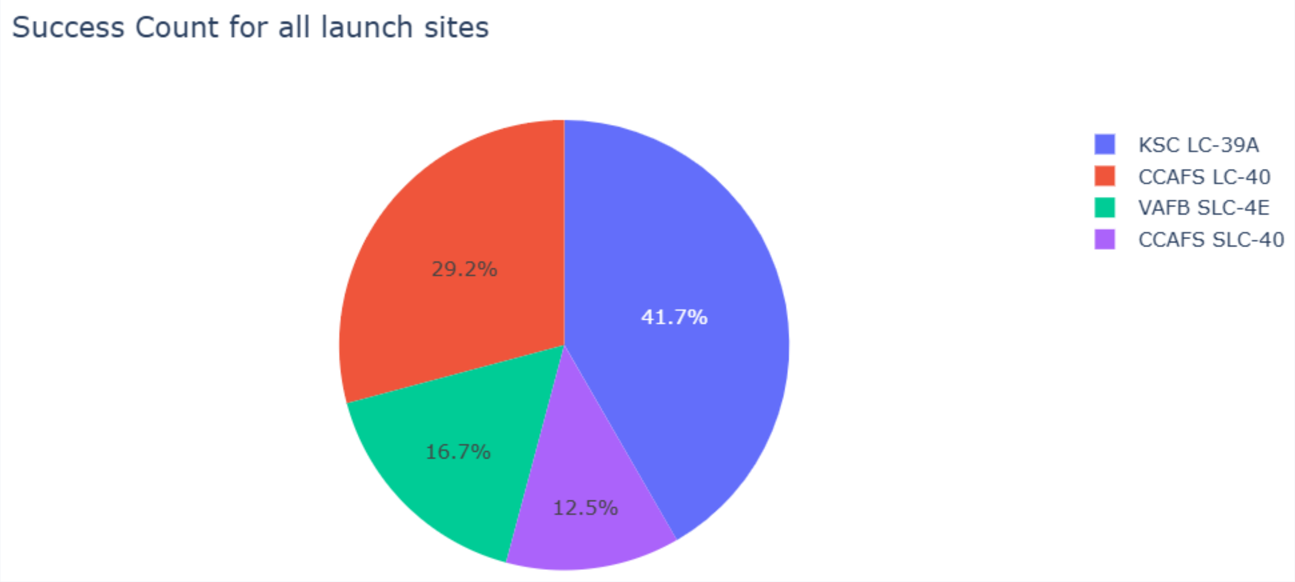


Section 4

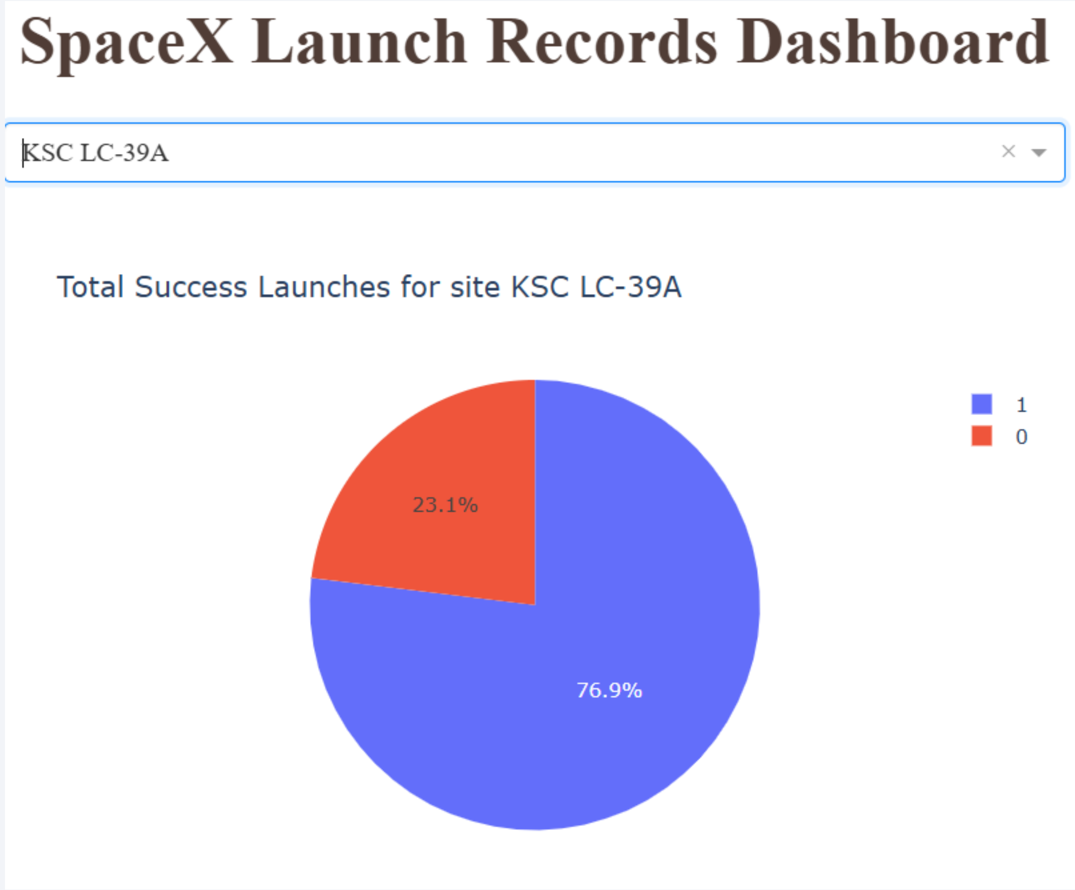
Build a Dashboard with Plotly Dash



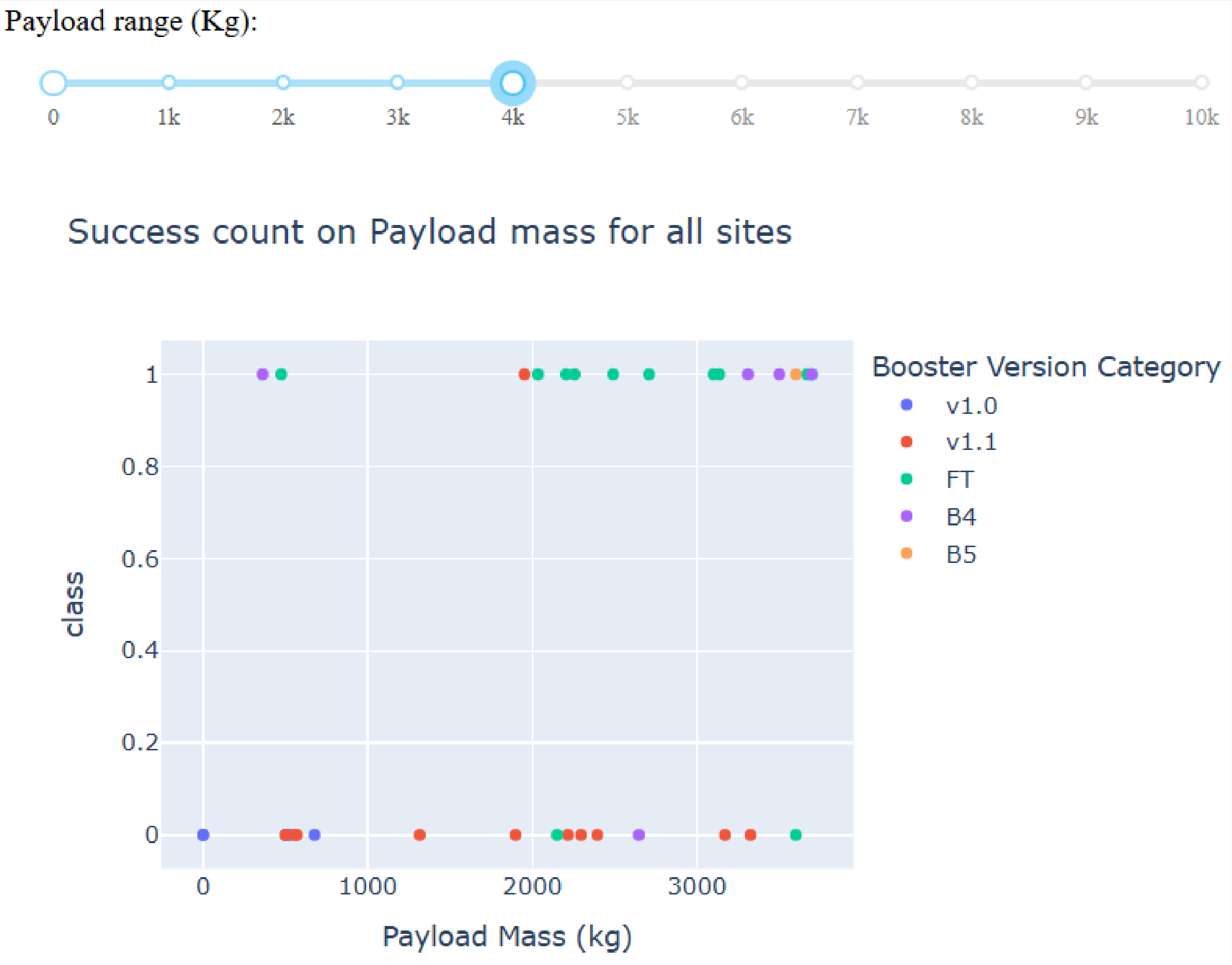
Success rate from all launch sites



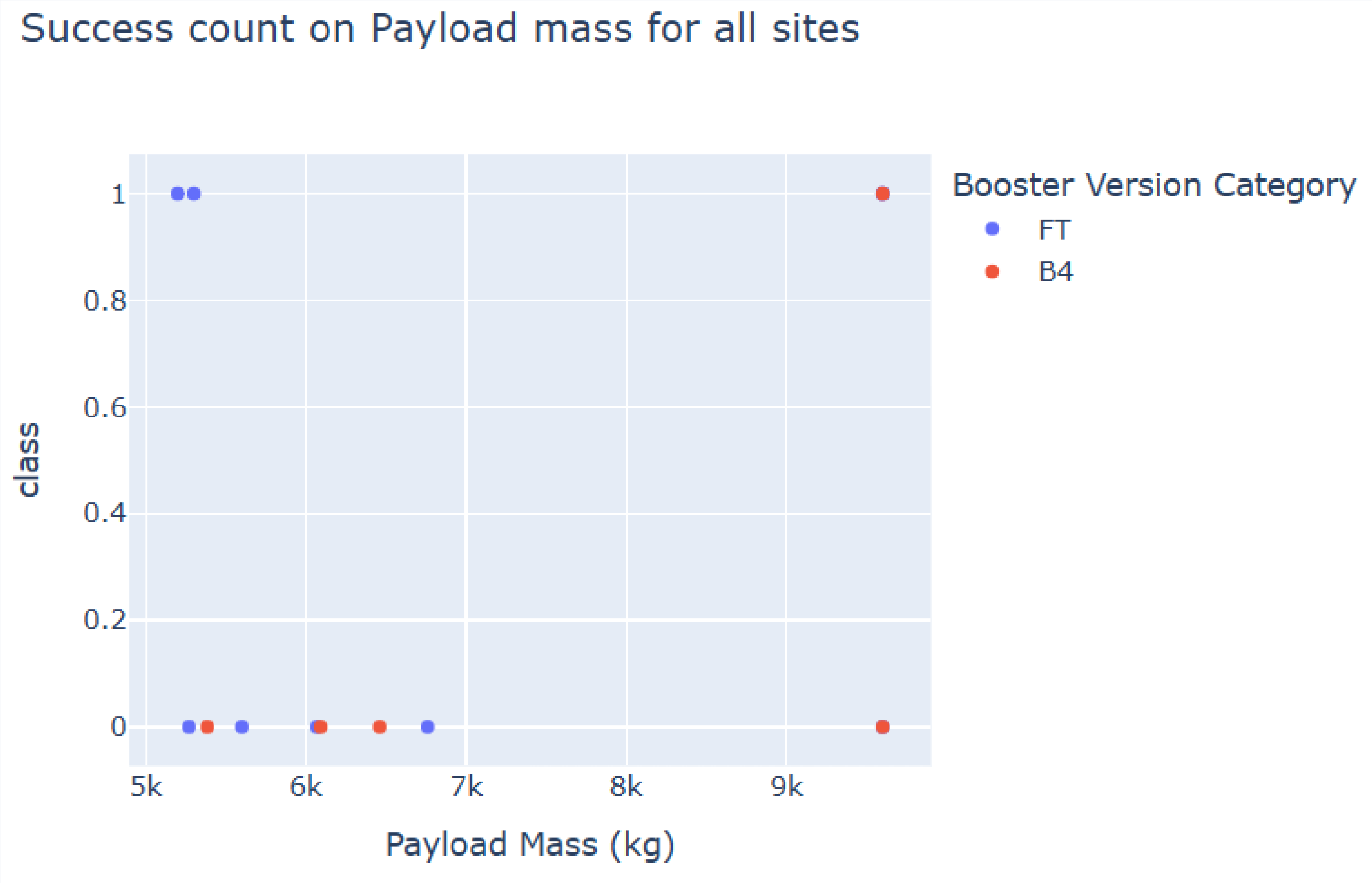
Launch outcomes from Kennedy Space Center



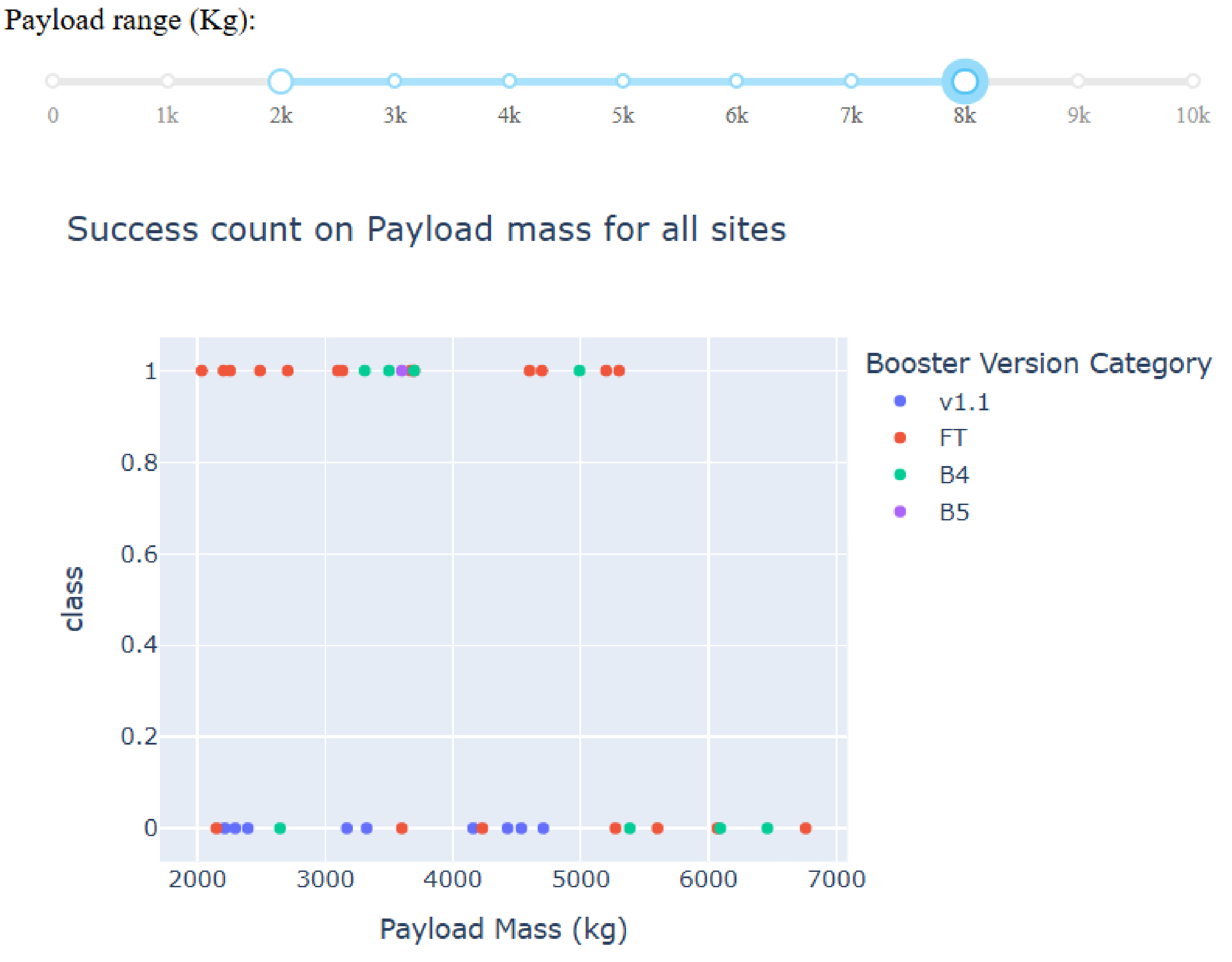
Launch outome against payload with different range 1/3



Launch outome against payload with different range 2/3



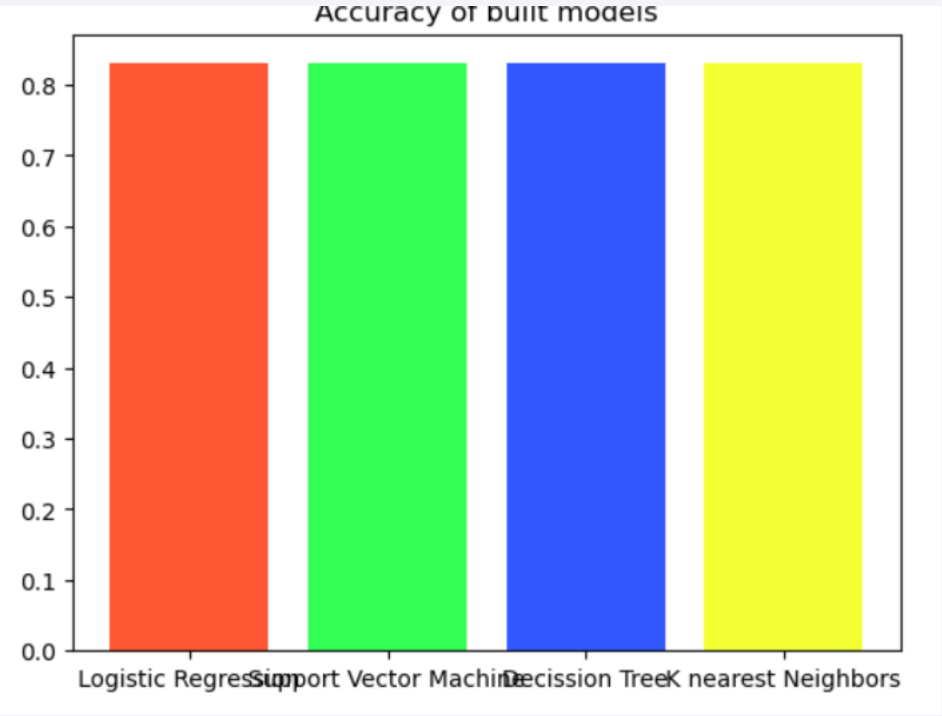
Launch outome against payload with different range 3/3



Section 5

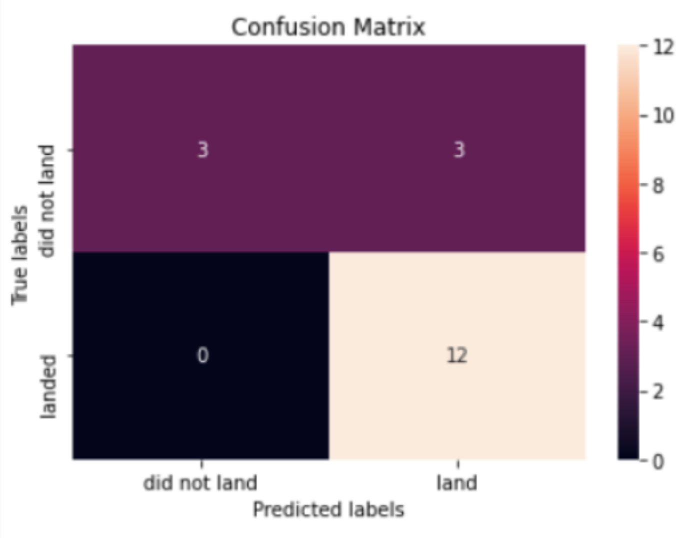
Predictive Analysis (Classification)

Classification Accuracy



Each of the four models built came back with the same accuracy score, 83.33%

Confusion Matrix



- The confusion matrices of the best performing models are the same
- Prediction of 1st stage boosters landing is however inproperly categorized (false positive)

Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!