IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

Executive Summary

- After collecting raw data from SpaceX API, it was preprocessed before visualized and modelled using various ML methods.
- This powerpoint presentation serves to illustrate my findings.

Introduction

- As space exploration grows in popularity, society must determine its importance and weight in comparison to other areas of research. Specifically, how much should governments invest in this field (in terms of expertise and monetary value)?
- One of the main players, namely SpaceX, is attempting to land a used rocket from outer space onto Earth's surface — an idea widely criticized and condoned by experts. This, however, does not stop Elon Musk from pursuing this idea.
- Thus, we seek to find: will the first stage of Falcon 9 land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: using SpaceX's API for extensive data on rocket ships
- Perform data wrangling: reduce raw data into useful data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models (e.x. SVM, Decision Tree, etc)

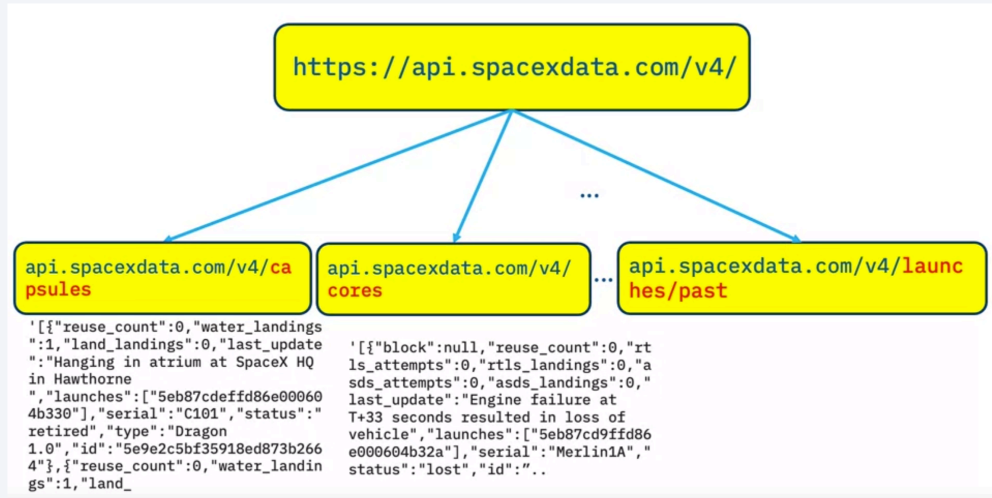
Data Collection

- Datasets collected: Rockets, Launchpads, Payloads, Cores
  - Retrieved using requests with SpaceXData's RESTful API
- Flowcharts and diagrams in following slides better illustrate this methodology.

Data Collection – SpaceX API

1. Sent get request to SpaceX API
2. Retrieved data in JSON format
3. Filtered dataframe

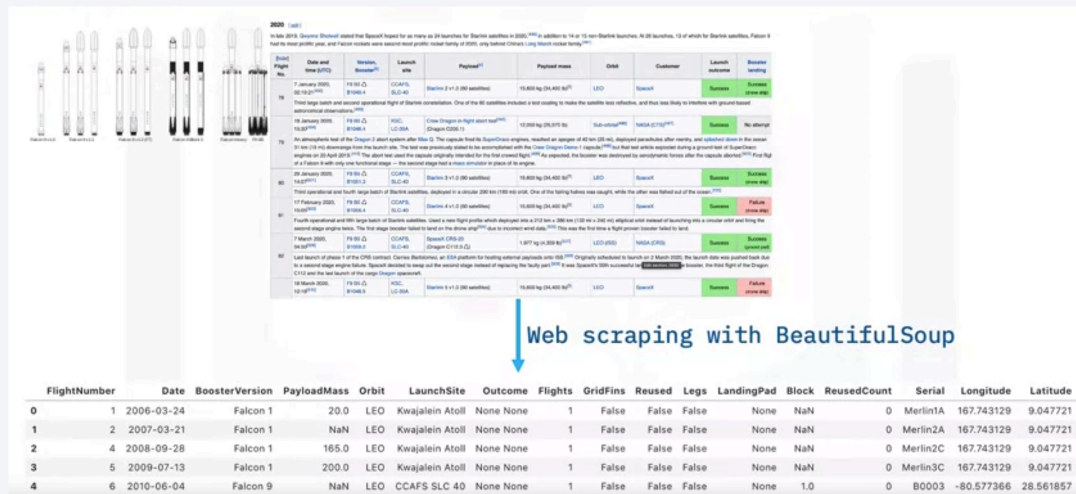
GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/spacex-data-collection-api.ipynb



Data Collection - Scraping

1. Extracted Falcon 9 Launch Records from Wikipedia
2. Parsed table using BeautifulSoup
3. Converted table into Pandas data frame

- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/spacex-webscraping.ipynb



	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
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Wrangling

1. Perform exploratory data analysis
2. Determine training labels

- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

Calculate number of launches on each site

↓

Calculate number of occurrence of each orbit

↓

Calculate number and occurrence of mission outcome per orbit type

↓

Create landing outcome label from Outcome column

EDA with Data Visualization

- Plotted charts: Cat plots, line plots, bar charts
- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL Queries:
  - Information regarding launch sites, payload mass, successful landing dates and names of boosters, successful/failure mission outcomes.
- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/eda-sql.ipynb

Build an Interactive Map with Folium

Launch success is determined by several factors including location and proximities of a launch site. This is what this portion is used for. Steps include:

1. Marked all launch sites on an interactive map
2. Marked the success/failed launches for each site
3. Calculated distances between launch sites and proximities

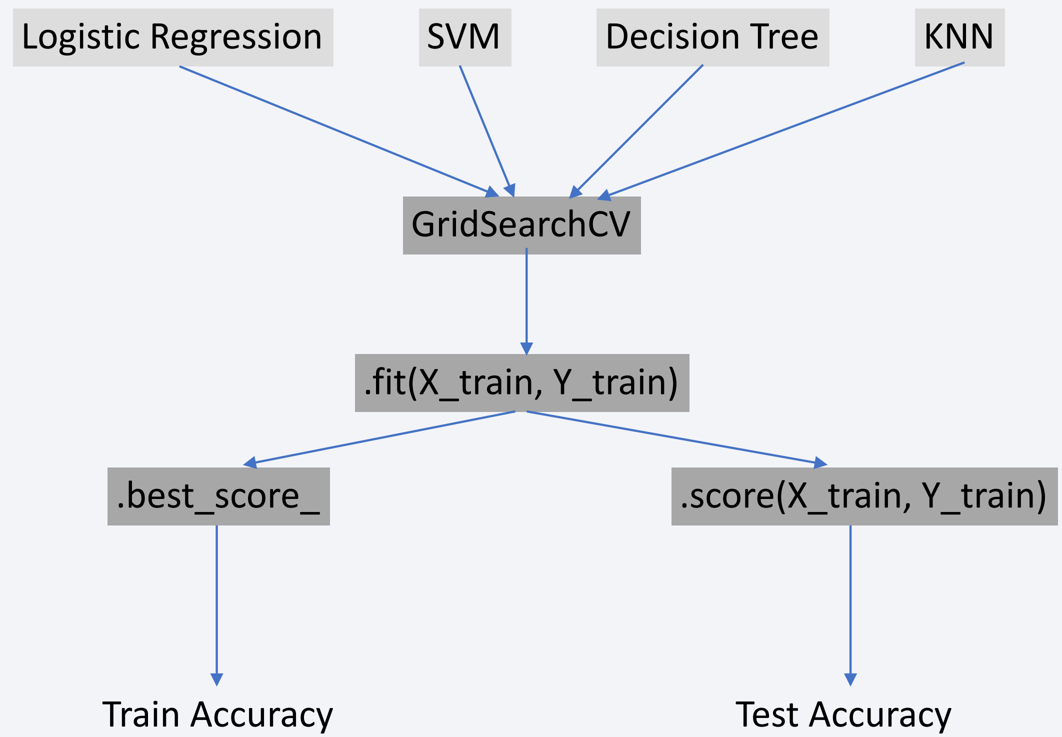
- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plots added: pie charts, scatter plots
- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Created 4 objects: logistic regression, support vector machine (SVM), decision tree classifier, and k-nearest neighbours (KNN).
- GitHub: https://github.com/R1chard04/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction.jupyterlite.ipynb

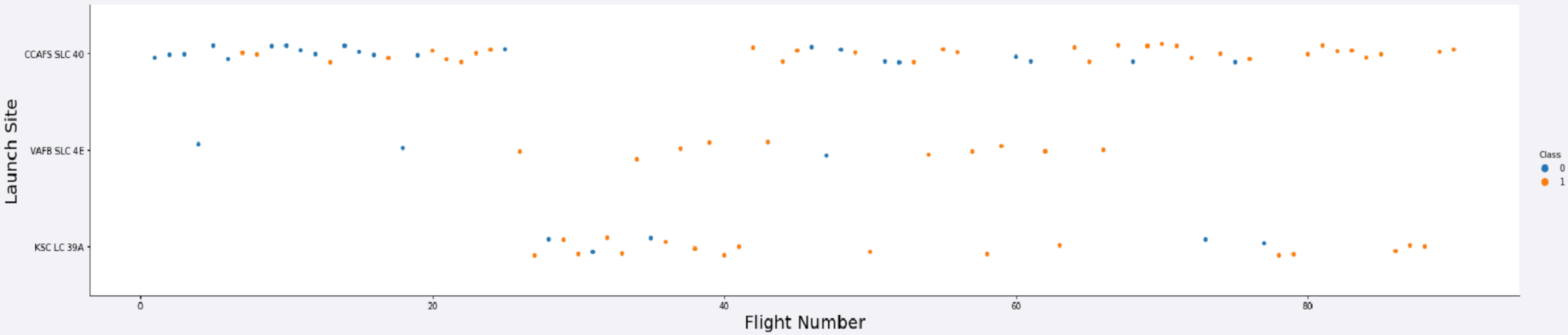


Section 2

Insights drawn from EDA

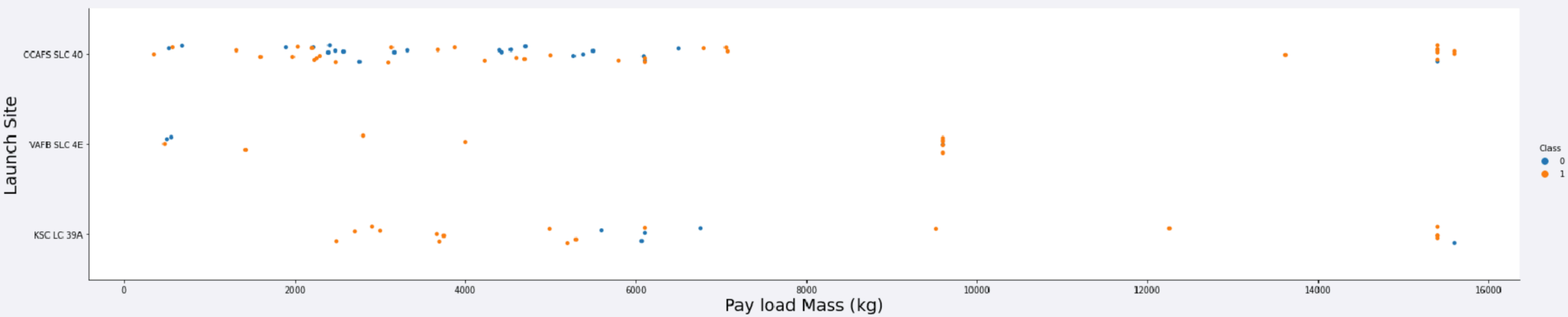
Flight Number vs. Launch Site

For each site, flight number and successful landings is proportional.



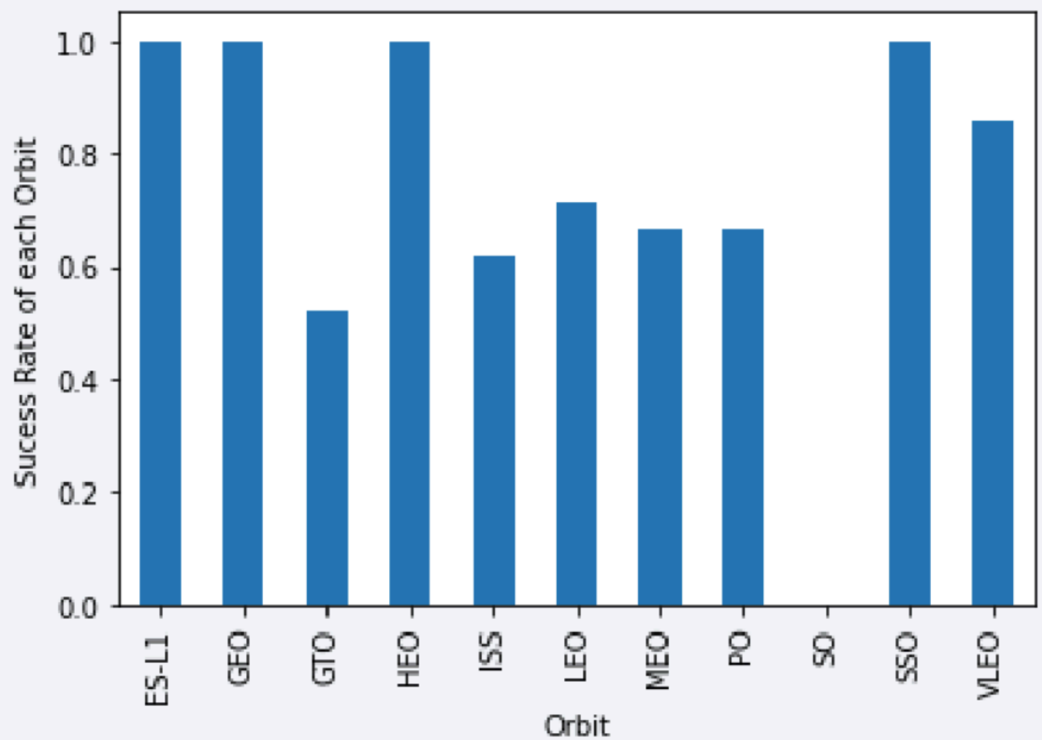
Payload vs. Launch Site

At approximately 7000 kg payload, the number of landing successes decreases substantially.



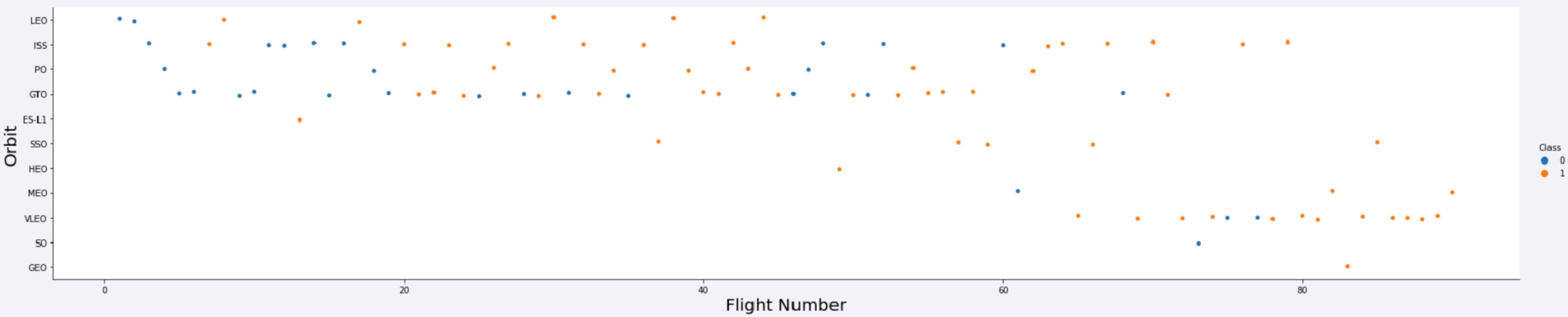
Success Rate vs. Orbit Type

The highest success rates result from orbits ESL1, GEO, HEO and SSO.



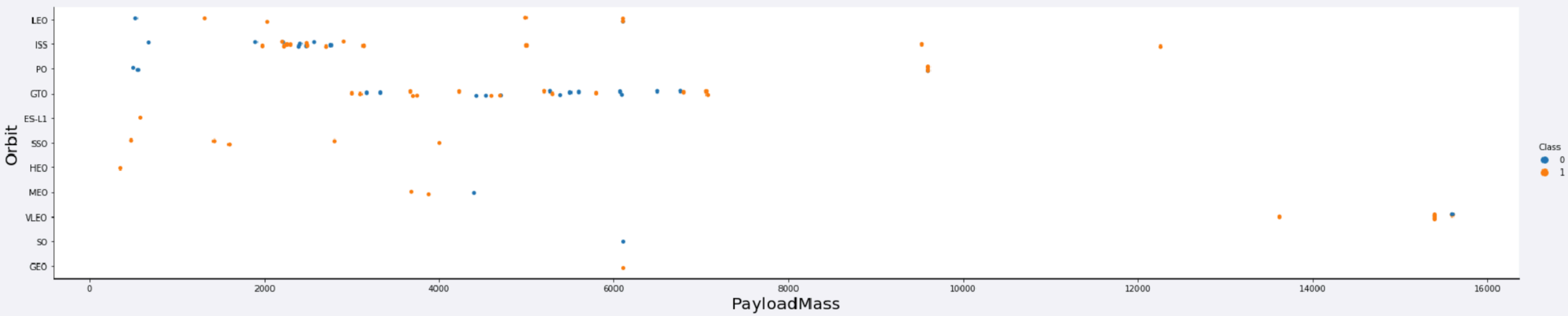
Flight Number vs. Orbit Type

Generally, for each orbit type, flight number and the frequency of successful landings is proportional



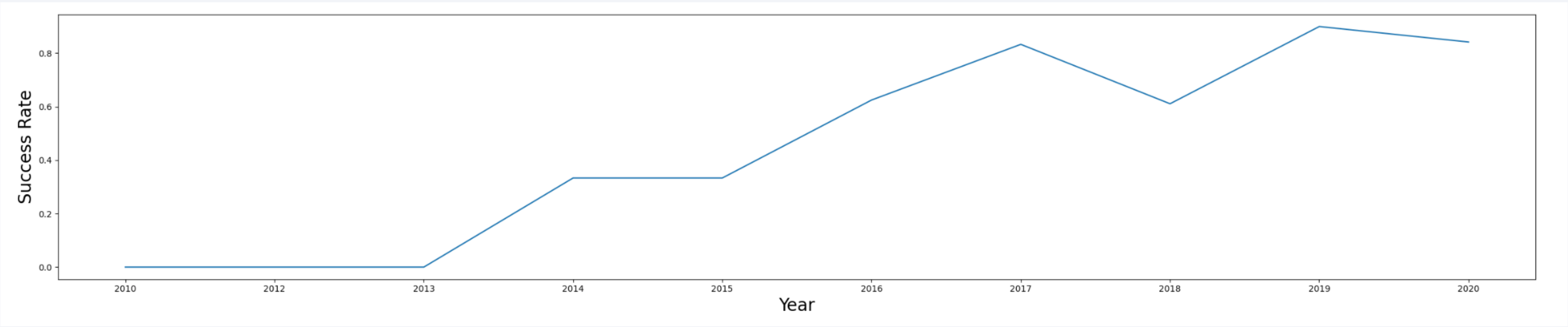
Payload vs. Orbit Type

Different payload masses are successful in different orbits. However, this trend is significantly more evident at lower weights (i.e. <7000 kg).



Launch Success Yearly Trend

As the year increases, the success rate increases as well.



All Launch Site Names

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

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

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

Launch Site Names Begin with 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
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

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

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

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

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%';
```

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

AVERAGE_PAYLOAD_MASS
2534.6666666666665

First Successful Ground Landing Date

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" LIKE '%Success%'
```

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

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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 "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAI
```

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, (SELECT COUNT("MISSION_(
```

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

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%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

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT substr(DATE, 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Failure (drone ship)
```

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

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "LANDING _OUTCOME", 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("LANDING _OUTCOME") DESC;
```

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

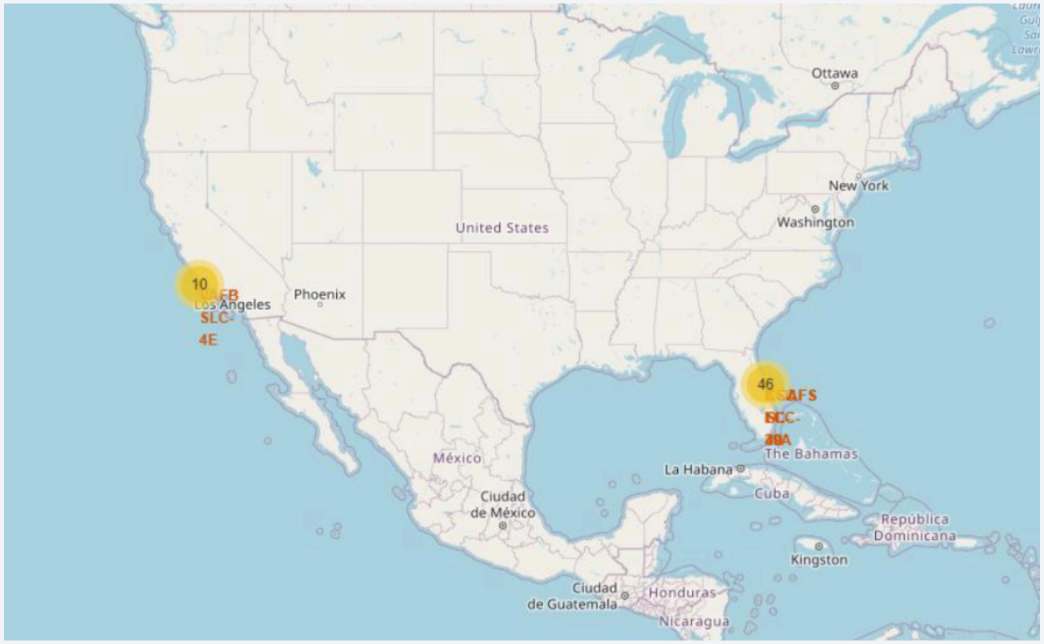
Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

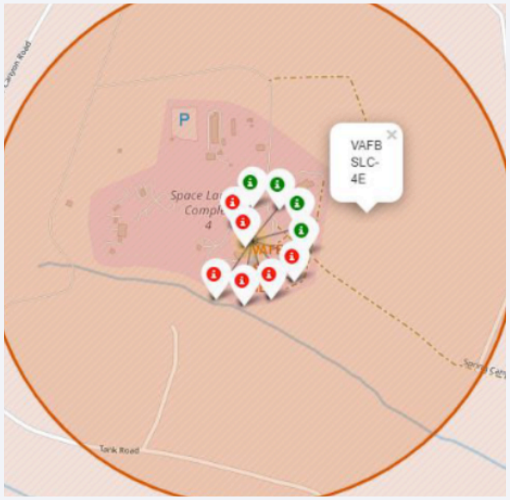
Folium Map: Launch Site Locations

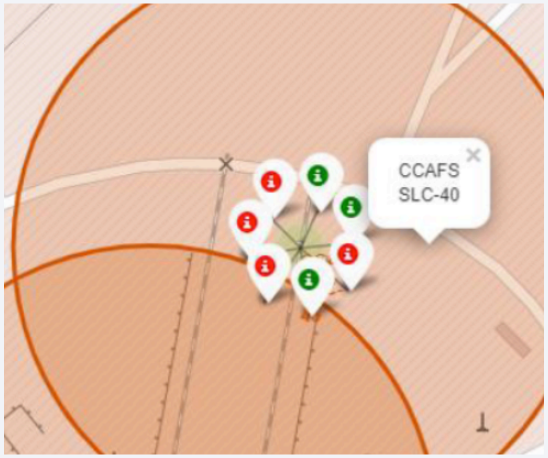
Evidently, SpaceX launch sites are located in California and Florida



Folium Map - Successful/Unsuccessful Launch Locations

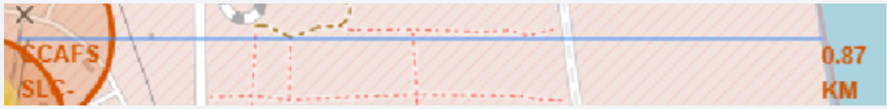
Green (information symbol) markers represent successful launches; red (information symbol) markers represent unsuccessful launches).



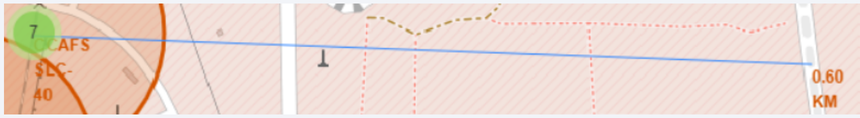


Folium Map - CCAFS SLC-40 Proximities

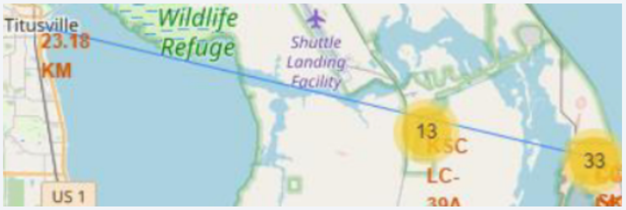
Coastline



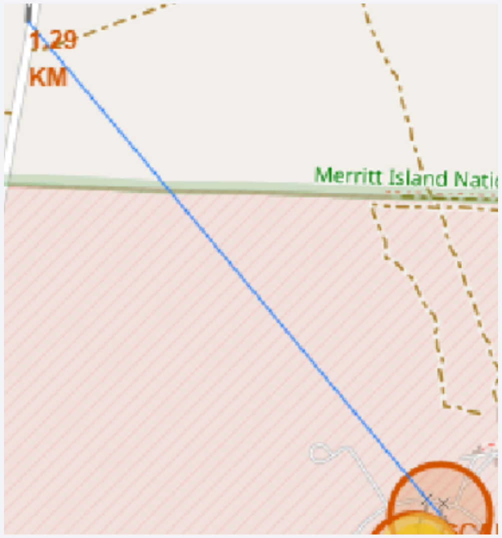
Railway



Highway

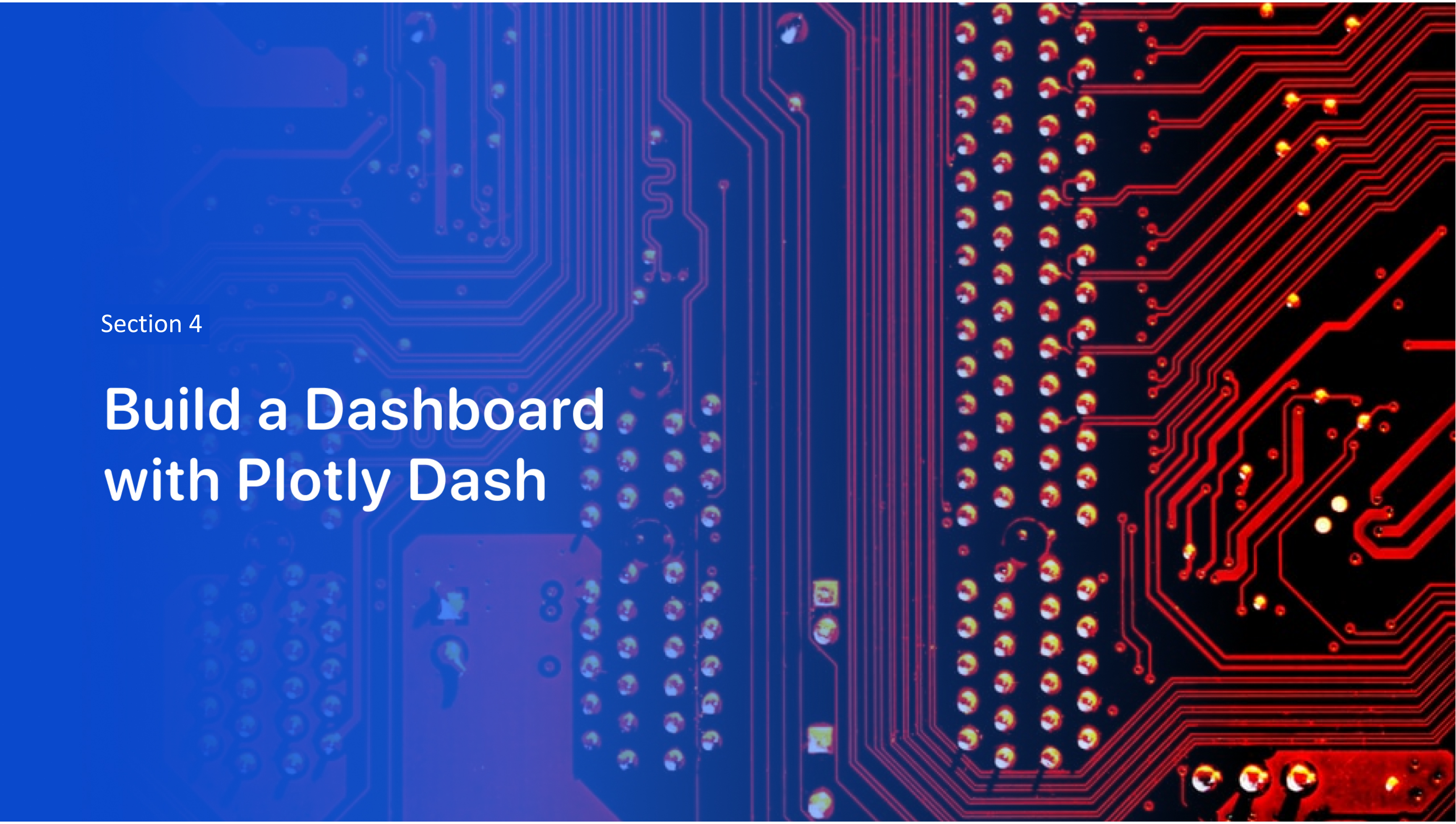


Cities

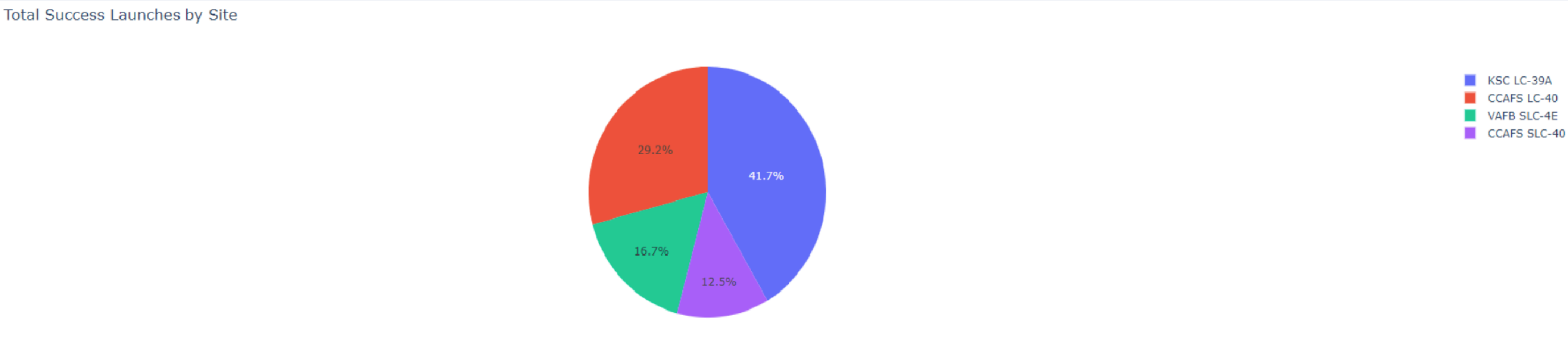


Section 4

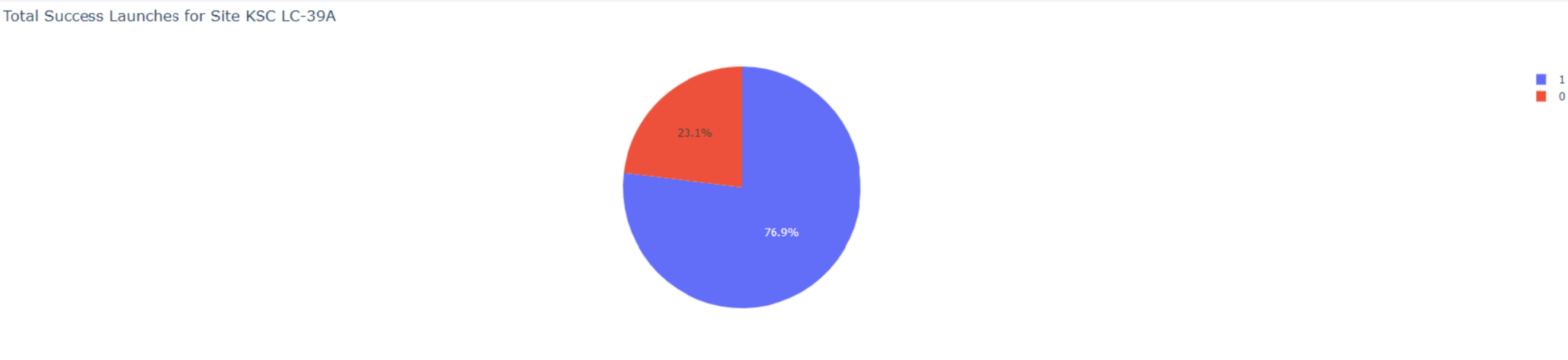
Build a Dashboard with Plotly Dash



Dashboard - Total Success by Site

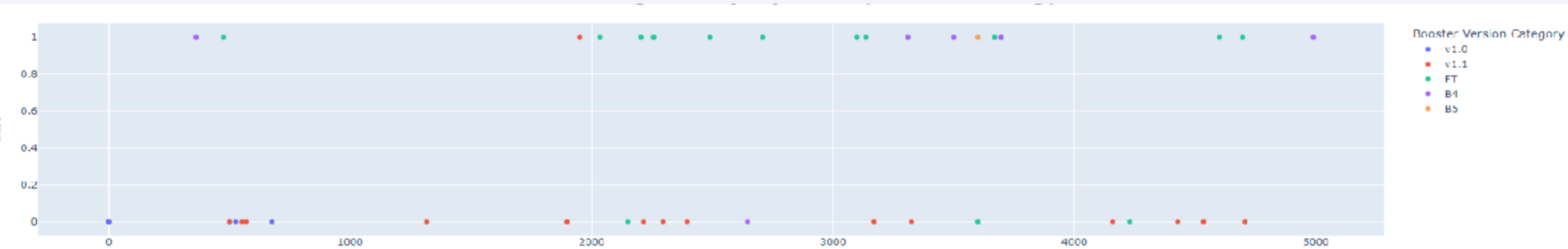


Dashboard - Total Success Launches for Site KSC LC-39A

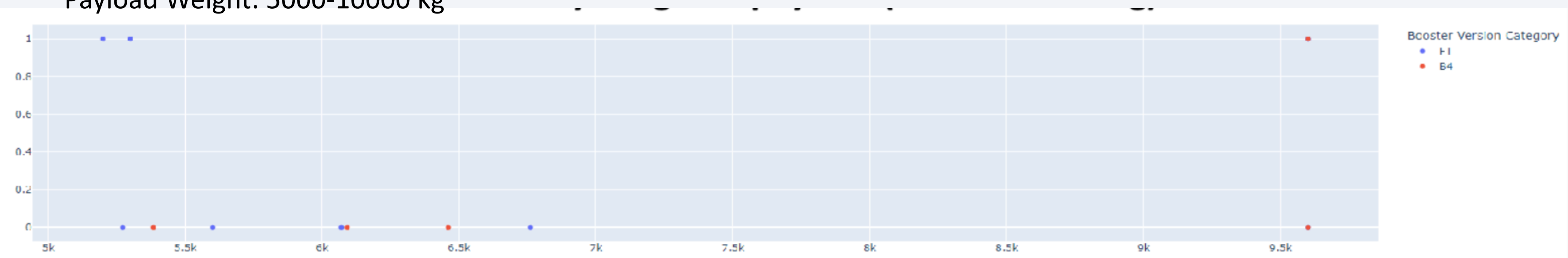


<Dashboard Screenshot 3>

Payload Weight: 0-5000 kg



Payload Weight: 5000-10000 kg



Section 5

Predictive Analysis (Classification)

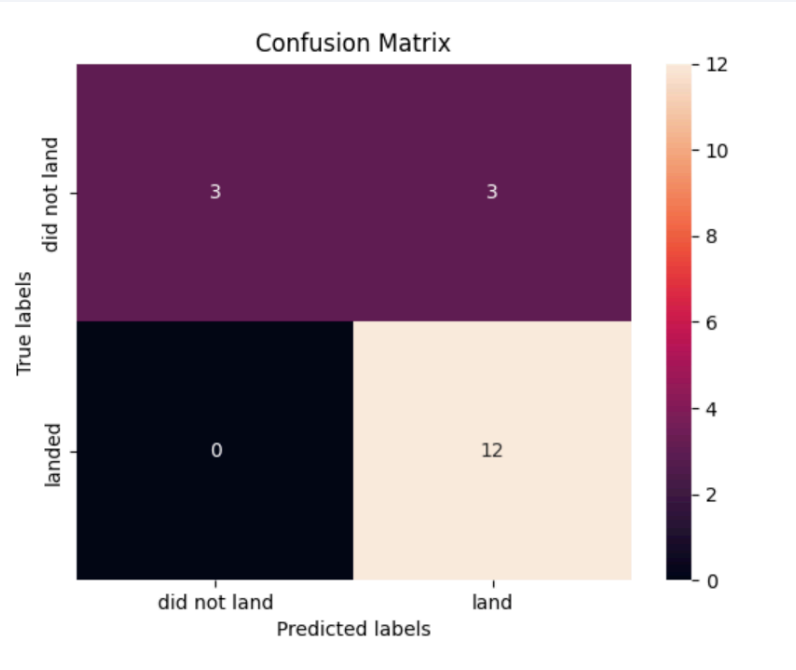
Classification Accuracy

All methods perform similar; hence, it's difficult to state for certain that one outperforms another with high certainty. That being said, the decision tree method has the highest training accuracy and the support vector machine (SVM) has the highest test accuracy.

	Train Accuracy	Test Accuracy
LogReg	0.846429	0.875000
SVM	0.848214	0.888889
Tree	0.876786	0.861111
KNN	0.848214	0.861111

Confusion Matrix

This is the confusion matrix that yielded the best accuracy. As shown, the number of false positives and false negatives is minimized



Conclusions

- Orbits with most success landings: GEO, HEO, SSO and ES-L1
- Payload mass is a high consideration when it comes to landing successes, especially weight below and above 7000 kg.
- Decision tree classification and SVM methods proved to be (slightly) superior models when compared to KNN and logistic regression models.


Thank you!