IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dylan van Wyk
March 2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone.
- Virgin Galactic is providing suborbital spaceflights.
- Rocket Lab is a small satellite provider.
- Blue Origin manufactures sub-orbital and orbital reusable rockets.
- SpaceX's is sending spacecraft to the International Space Station, has launched Starlink, a satellite internet constellation providing satellite Internet access and sending manned missions to Space.
- SpaceX is the most successful seeing as their rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- The purpose of this capstone is to take on the role of a data scientist working for a new rocket company: Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk.
- The question we would like to answer is to determine the price of each launch. We will do this by
- gathering information about Space X and creating dashboards for the team.
- training a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Methodology

Executive Summary

- Data collection methodology:
    - Used the SpaceX API
    - Extracted the Falcon 9 launch records HTML table from Wikipedia
- Perform data wrangling
    - Converted outcomes into training labels with 1 means the booster successfully landed 0 means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Created a machine learning pipeline to predict if the first stage will land given the data
    - Tuned and evaluated different classification models

Data Collection

- **SpaceX REST API:** The SpaceX Api , is an unofficial open source REST Api for SpaceX launch, rocket, core, capsule, starlink, launchpad, and landing pad data.
- **Web scraping:** collecting Falcon 9 historical launch records from a Wikipedia using BeautifulSoup. Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML.

Data Collection – SpaceX API

- We used a *get* request to the SpaceX API get a *json* file of the data. This *json* file was then converted into a dataframe.
- Github: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Request and parse the SpaceX launch data using the GET request

↓

Filter the dataframe to only include Falcon 9 launches

↓

Replaced missing values with the mean value

Data Collection - Scraping

- Web scraped Falcon 9 launch records with BeautifulSoup.
- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Request the Falcon9 Launch Wiki page from its URL

↓

Extract all column/variable names from the HTML table header

↓

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots, barplots and line charts were used to visualize the relationship between pair of features:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success rate and Orbit type
  - Flight Number and Orbit type
  - Payload and Orbit type
  - Launch success by year

- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

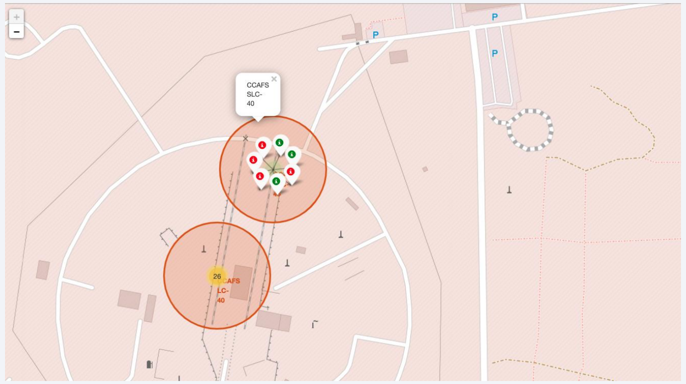
EDA with SQL

- The following SQL queries were performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass
  - 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.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The following makets were added on the map:
  - All launch sites : visualizing locations by pinning them on a map.
  - success/failed launches for each site on the map: see which sites have high success rate.
  - Calculate the distances between a launch site and:
    - Railways
    - Cities
    - Coastlines
    - Highways

  for safety purposes.



- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- GitHub: https://github.com/DylanVW313/Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four different models were used on the data:
  - logistic regression
  - support vector machine
  - decision tree
  - K nearest neighbors
- GitHub https://github.com/DylanVW313/Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

1. **Load the dataframe**
2. Standardize the data
3. Split the data into training and test data
4. Created a model
5. Fitted the model
6. Tested the Accuracy of the model

Results

Exploratory data analysis results:

- Space X uses 4 different launch sites;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes improved over time.
- Launches took place mainly on the east coast of the US.
- Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

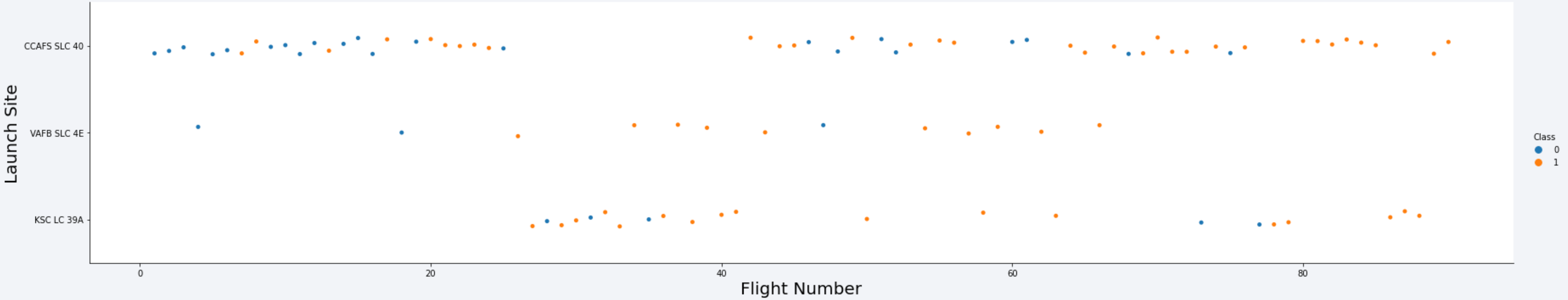
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

According to the plot below:

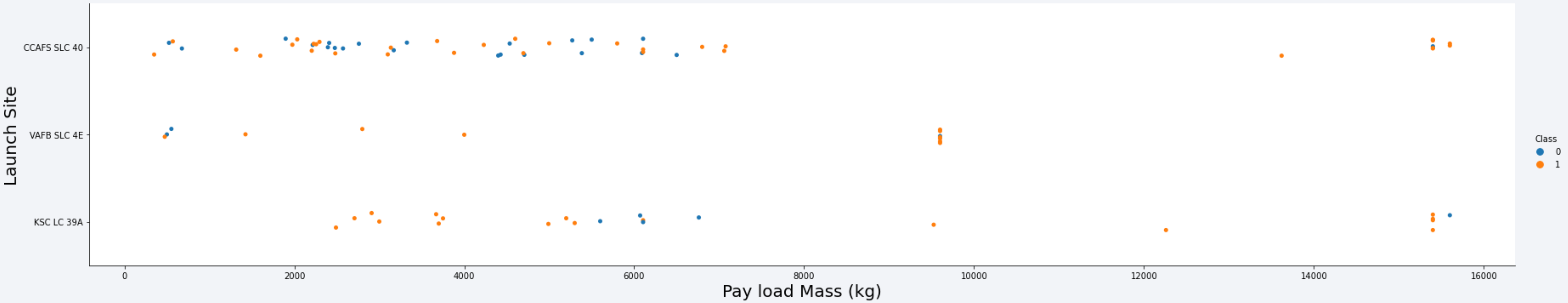
- CCAF5 SLC 40 has been the longest used launch site
- CCAF5 SLC 40 has been used the most recently
- VAFB SLC 4E has been used less frequently and has not been used recently
- KSC LC 39A is the newest addition and the last 5 launches have been successful.



Payload vs. Launch Site

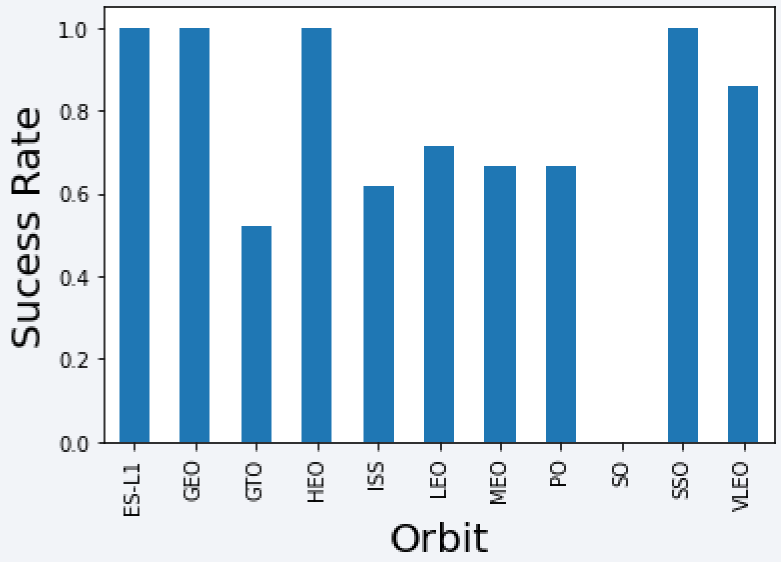
According to the plot below:

- CCAF5 SLC 40 has been used to test smaller payload sizes, with mixed results.
- VAFB SLC 4E has a maximum payload size of 9000kg
- KSC LC 39A has done well with smaller and larger payload sizes, but has failed more often in the 5000 – 7000kg range
- Heavier payloads have been more successful overall.



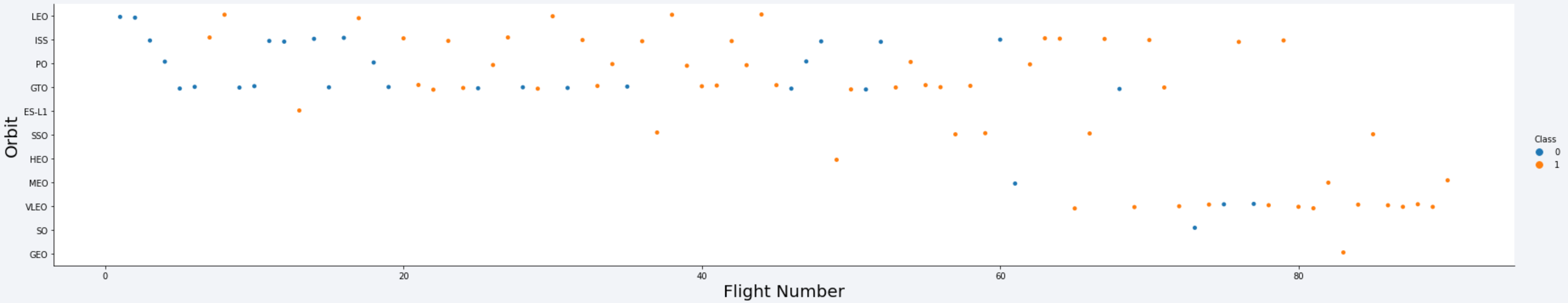
Success Rate vs. Orbit Type

- The following orbits have a 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- The following orbits have a highsuccess rate (>70%):
  - VLEO
  - LFO



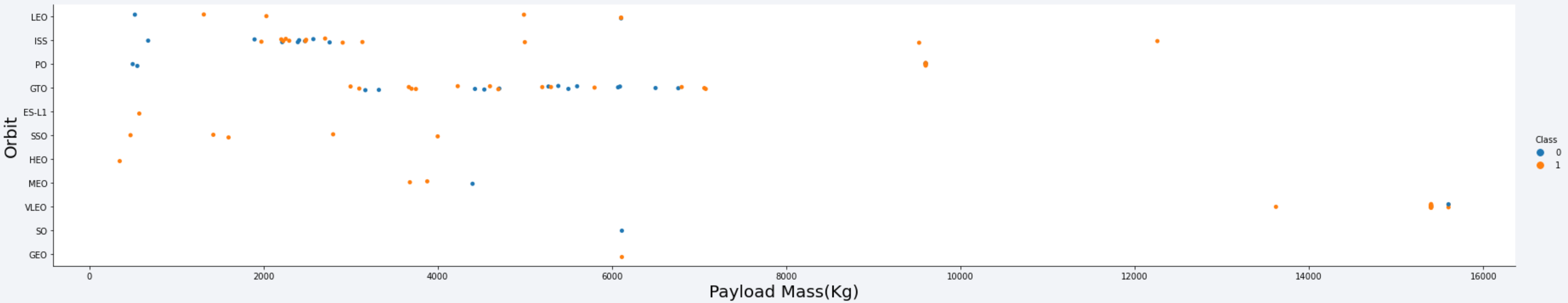
Flight Number vs. Orbit Type

- LEO, ISS, PO and GTO were the first orbit types to be used, with low success rates initially
- More recently WEO has been the most used orbit, with reasonably high success rate
- GTO and ISS seem to have the worst success rates
- VLEO seems to have the best success rate



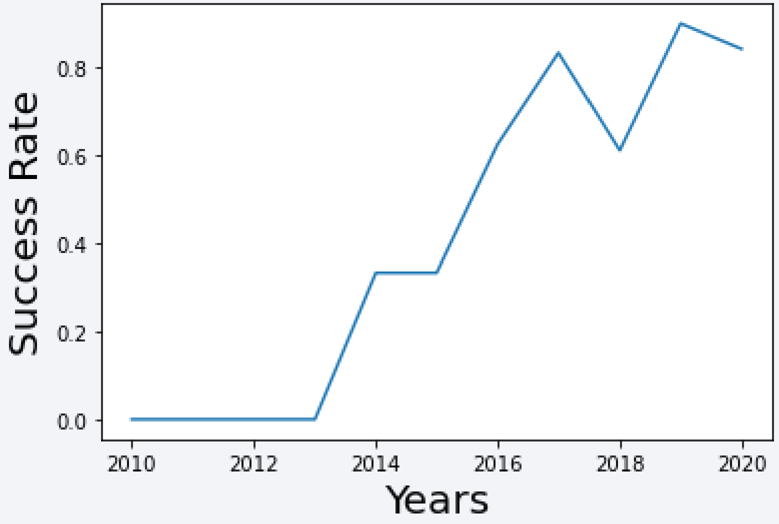
Payload vs. Orbit Type

- VLEO is used exclusively for large payloads
- GTO caters for payloads between 3000 and 7000kg
- ISS has the widest range of payload weights.



Launch Success Yearly Trend

- Success rates have been steadily improving
- There were no perfect years
- 2010 -2013 had zero success rate – could be considered a testing phase
- 2018 was lower than trend
- Future launches should have high success rates if trend continues



All Launch Site Names

- Use distinct() to find unique values

In [11]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL;
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

- Use LIKE '%String%' to search records which contain a certain string

```
%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
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

- Use SUM() to find the total of a column

```
[13]: %sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE CUSTOMER LIKE '%CRS%';
```

[13]:

SUM(payload_mass__kg_)
48213

Average Payload Mass by F9 v1.1

- Use AVG() to find the average of a column

```
In [14]: %sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version LIKE '%F9 v1.1%';
```

Out[14]:

AVG(payload_mass__kg_)
2534.6666666666665

First Successful Ground Landing Date

- Use MIN() to find the smallest value in a column

In [15]:

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success (ground pad)%';
```

Out[15]:

MIN(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use BETWEEN xxx and xxx to find values in a certain range

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

In [17]:

```
%sql select BOOSTER_VERSION from SPACEXTBL where "Landing_Outcome"='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

Out[17]:

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

Total Number of Successful and Failure Mission Outcomes

- Use COUNT() to count the number of occurrences in a column
- Use GROUP BY by to arange identical data into groups

List the total number of successful and failure mission outcomes

In [20]:
```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Mission_Outcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

Out[20]:

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

Boosters Carried Maximum Payload

- Use a subquery first use MAX() to find largest value
- Then SELECT the boosters with the max value

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

In [21]:
```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

Out[21]:

boosterversion
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

- Use LIKE '%2015%' to find dates which have 2015 in them

In [25]:
```
%sql SELECT DATE, "Landing_Outcome", booster_version, launch_site FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Failure (drone ship)%' AND DATE LIKE '%
```

Out[25]:

Date	Landing_Outcome	Booster_Version	Launch_Site
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Use ORDER BY Field DESC to arrange outcomes by a field

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

In [32]:

```
%sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017
```

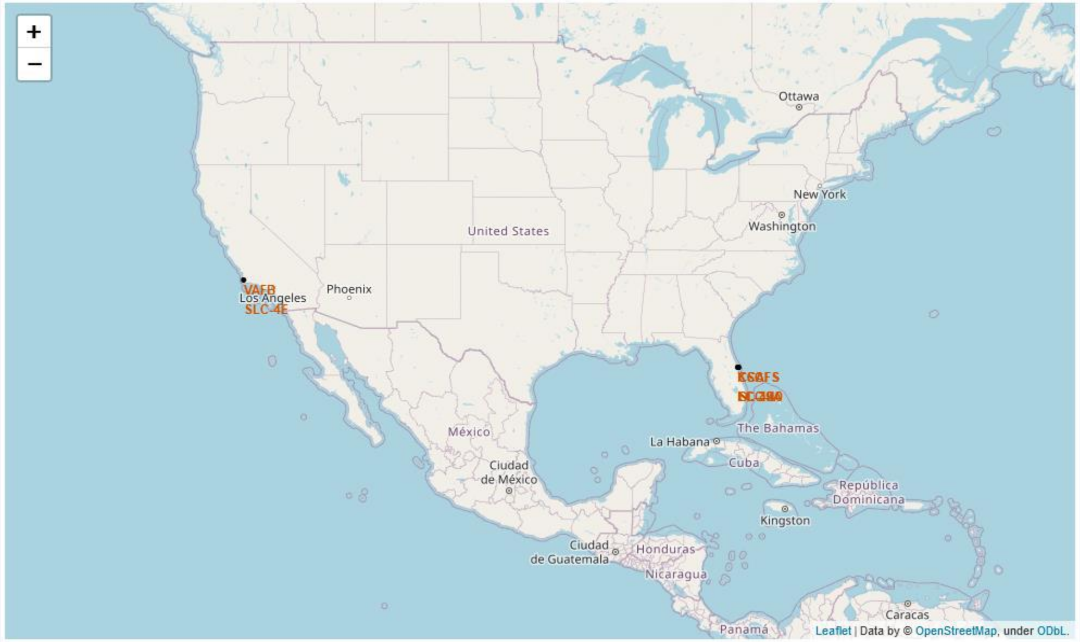
Out[32]:

Landing Outcome	Occurrences
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

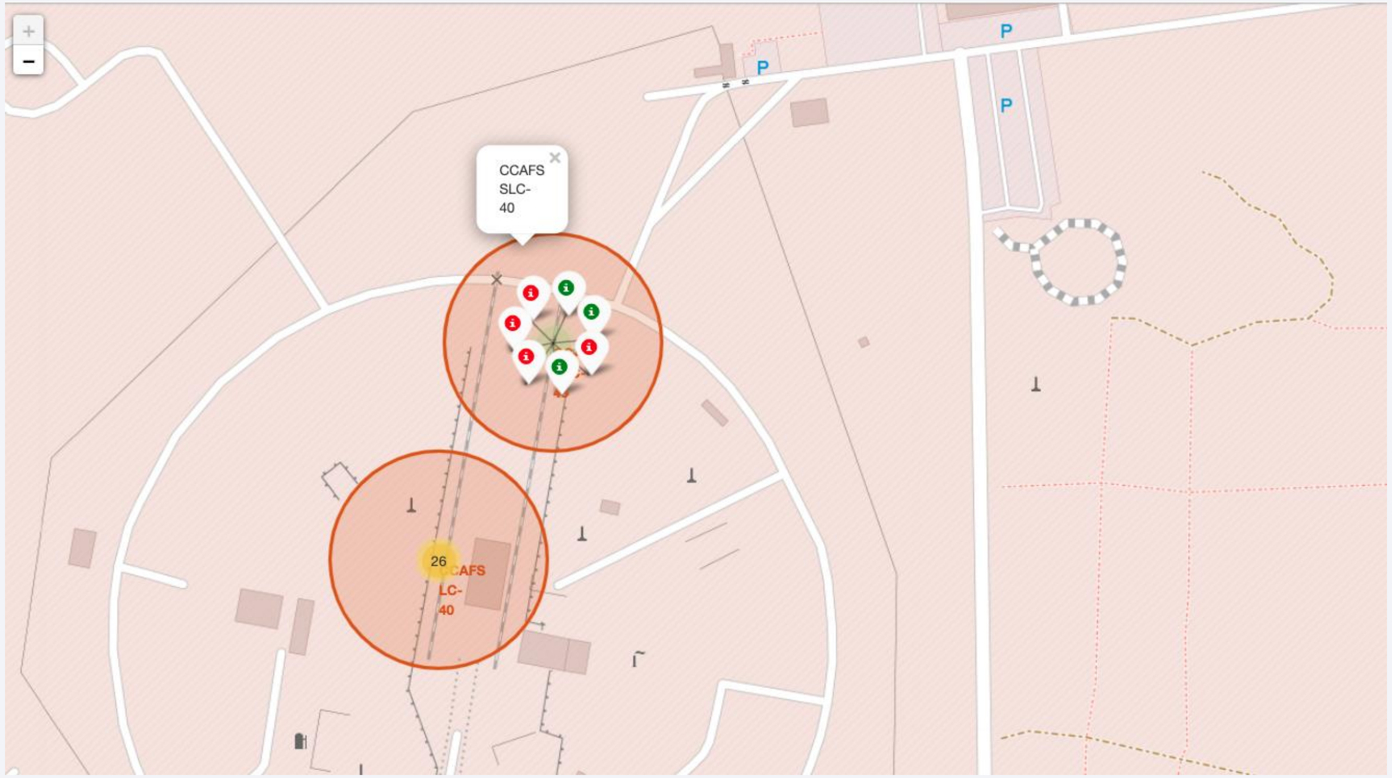
Launch Sites Proximities Analysis

All Launch Sites



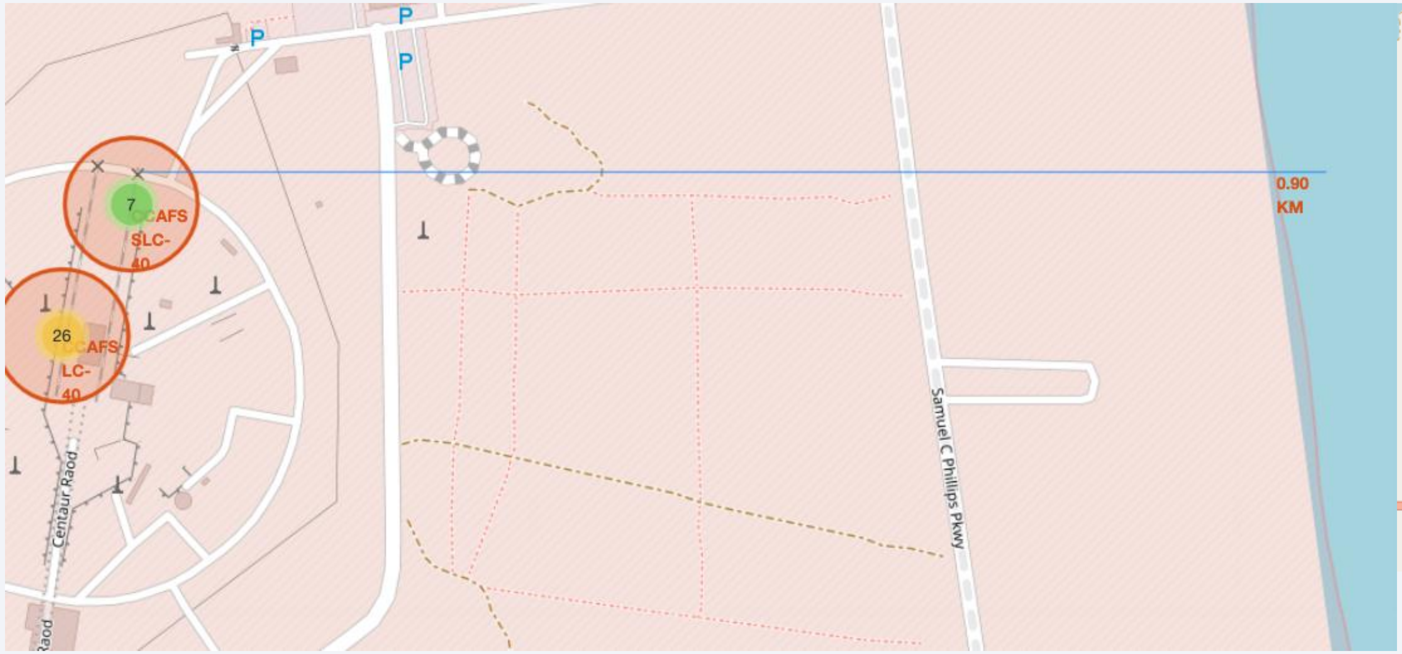
- Launch sites are close to the coast in the USA

Successful and failed launches



- Red indicates failed launces, green indicates successful launces

Launch site proximity

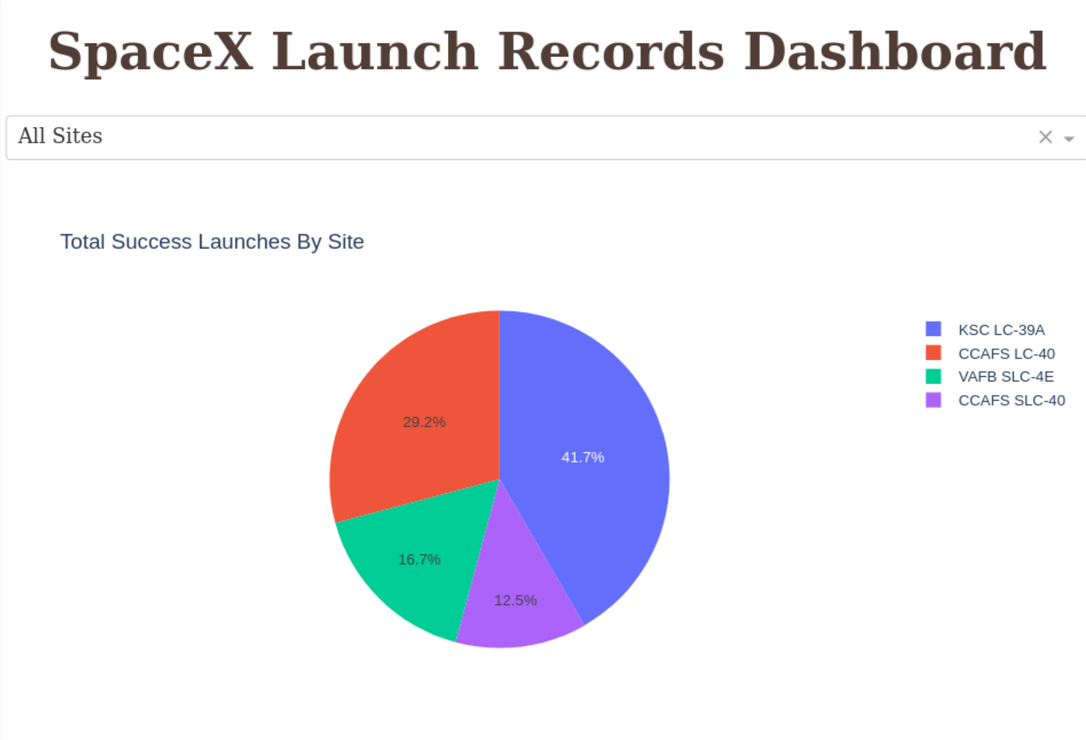


- Launch sites are close to the ocean as well as railway lines
- They are generally far away from populated areas

Section 4

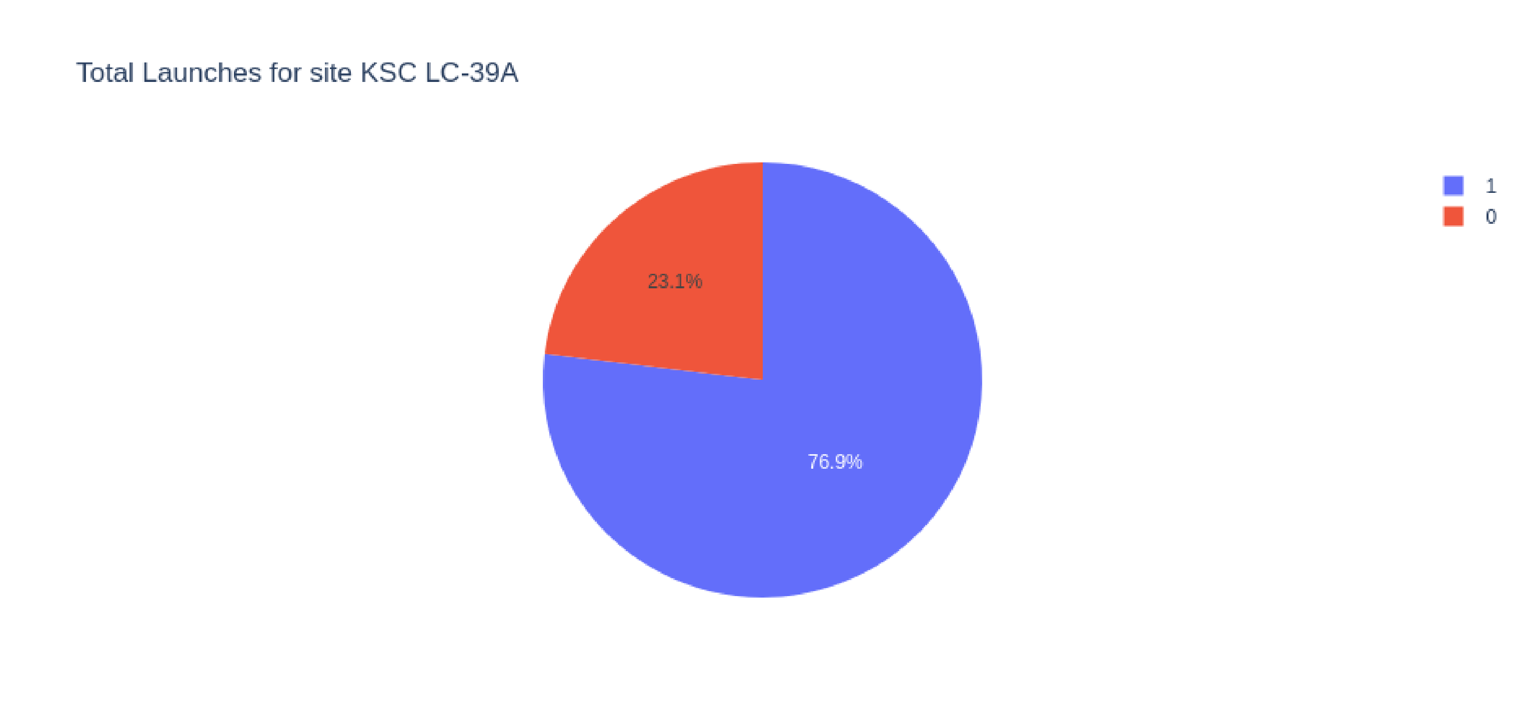
Build a Dashboard with Plotly Dash

Successful Launches by Site



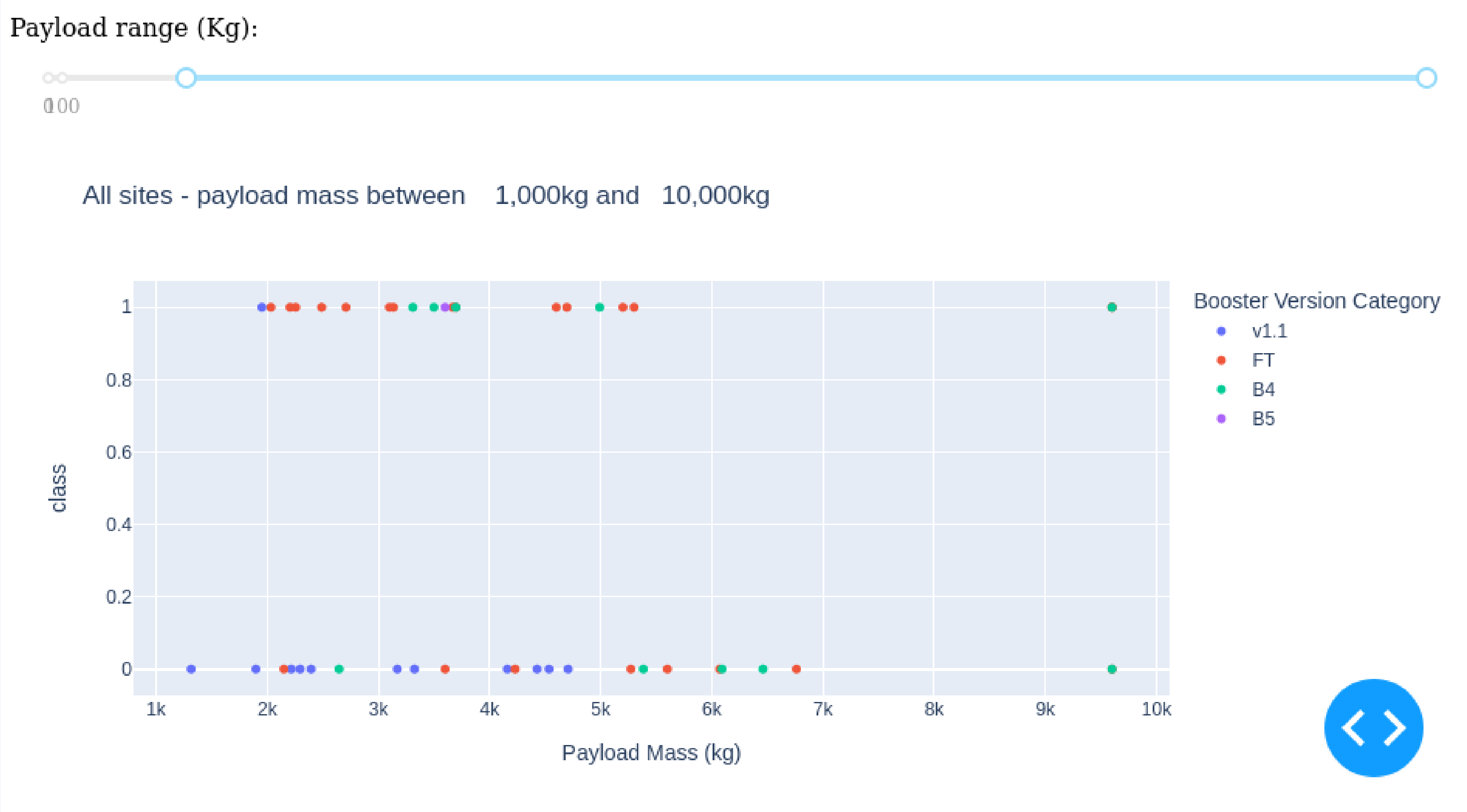
- Launch sites seem to be a key factor in success rate

Launch Success Ratio for KSC LC 39A



- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



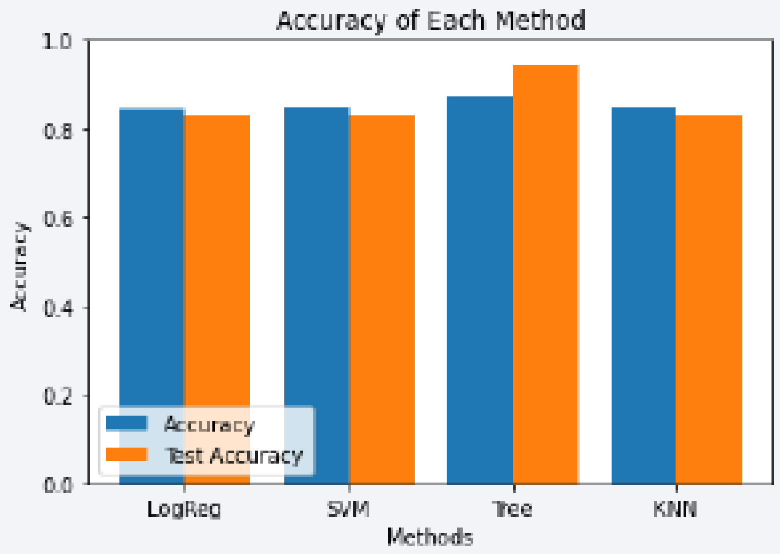
- Payloads under 6,000kg and FT boosters are the most successful combination.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- Descion tree delivered the best result with no false negatives and only 3 false positives

Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC 39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!