IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MANSI TAK
07-06-2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

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

Introduction

Project background and context:

We predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems that needed solving:

- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX API and web scraping from Wikipedia Pages
- Perform data wrangling
  - Data was collected and converted into Pandas data frame for visualization and analysis
- 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 collected were normalized, divided into training and test data sets, and evaluated by four different classification models. The accuracy of each model was assessed using various combinations of parameters

Data Collection

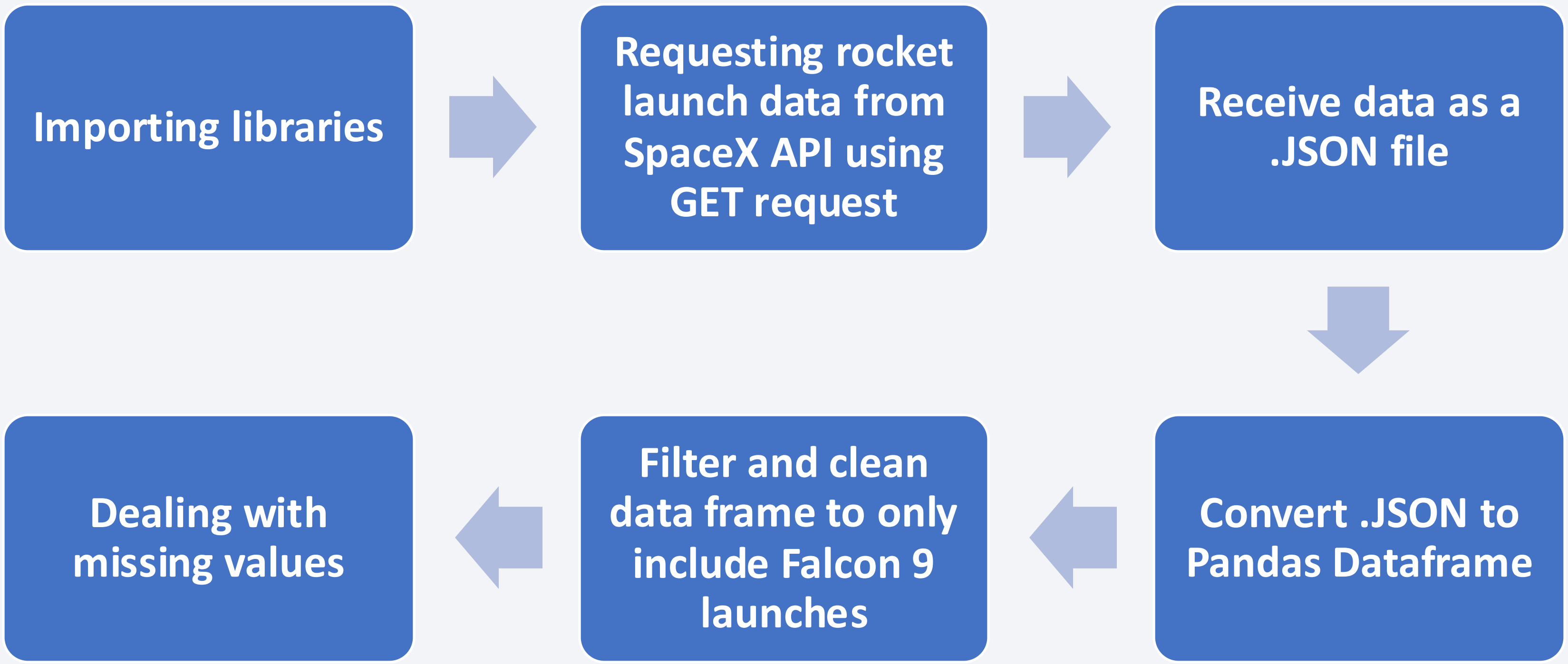
- Data collection of rocket launch data from SPACEX API using GET request.

1. Requesting rocket launch data from SpaceX API using GET request
2. Receive data as a .JSON file
3. Convert .JSON to Pandas Dataframe
4. Filter and clean data in dataframe
5. Dealing with missing values

1. Web Scraping Falcon 9 launch records
2. Use BeautifulSoup to web scrape HTML Tables
3. Extract HTML Tables
4. Parse data from tables
5. Convert table into Pandas data frame

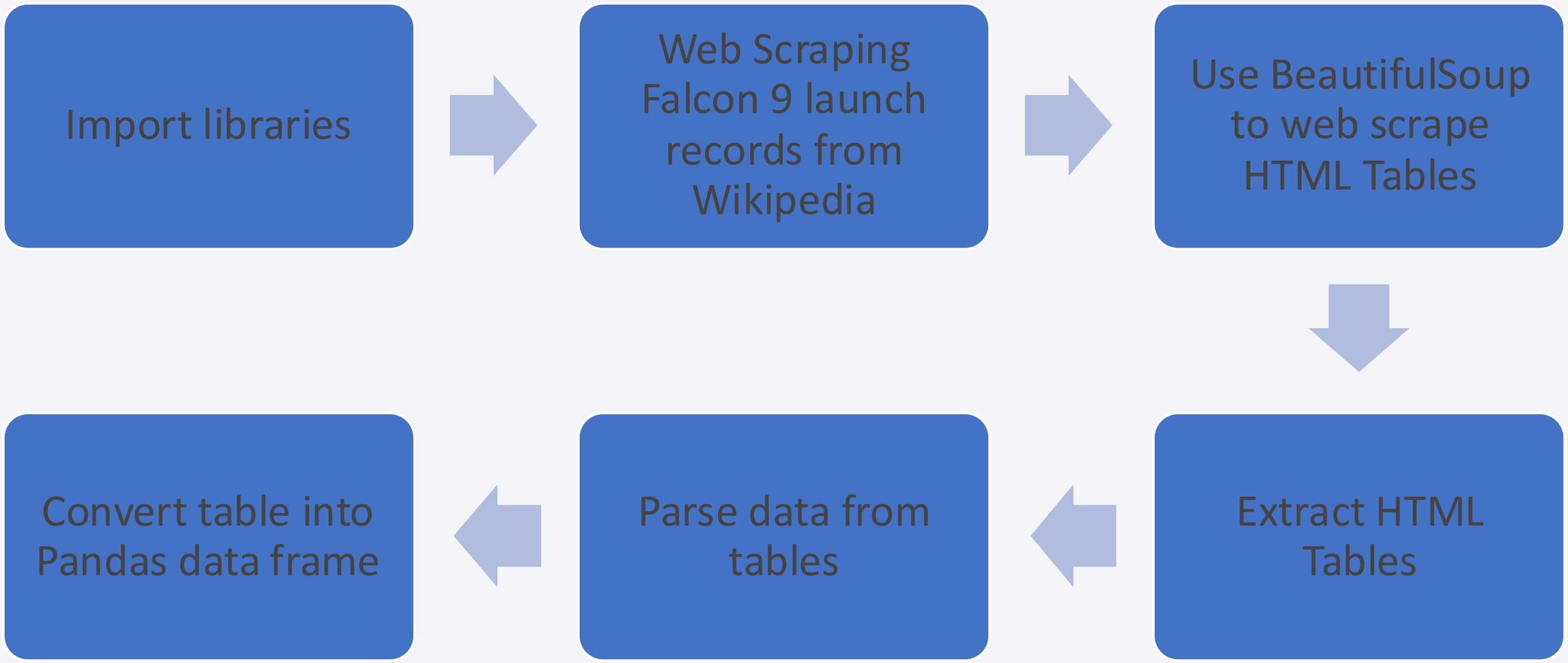
Data Collection – SpaceX API

- Data Collection, cleaning, removing misising values.
- [Data-Collection-API-Notebook]



Data Collection - Scraping

- Data Scarping from Wikipedia using BeautifulSoup.
- [Web-Scraping-Notebook]



Data Wrangling

- Data Wrangling using Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- [Data-Wrangling-Notebook]

Load SpaceX dataset as Pandas data frame → Identify missing values and which columns are numerical and categorical → Calculate the number of launches on each site → Calculate the number and occurrence of each orbit → Calculate the number and occurence of mission outcome of the orbits → Create a landing outcome label from Outcome column

EDA with Data Visualization

Exploratory Data Analysis using Data Visualization

What charts were plotted and why:

1. Catplot(scatter plot) : to help visualize the relationship between a numerical value and a categorical value
   - FlightNumber vs. PayloadMass
   - FlightNumber vs. LaunchSite
   - PayloadMass vs. LaunchSite
   - FlightNumber vs. Orbit
   - PayloadMass vs. Orbit
2. Bar Chart: to visually check if there are any relationship between success rate and orbit type
   - Success Rate vs. Orbit
3. Line Chart
   - Launch success yearly trend

EDA-Data-Visualization-Notebook

EDA with SQL

- EDA was applied with SQL to help obtain information about the data.
- Launch sites starting with “CCA” were pulled out
- Total payload mass carried by boosters/boosters F9 v1.1were determined
- Successful landing(ground pad)/(drone ship) outcomes determined
- Total number of success and failures of missions
- The failed landing outcomes in drone ship, their booster version and launch site names
- EDA-with-SQL-Notebook

Build an Interactive Map with Folium

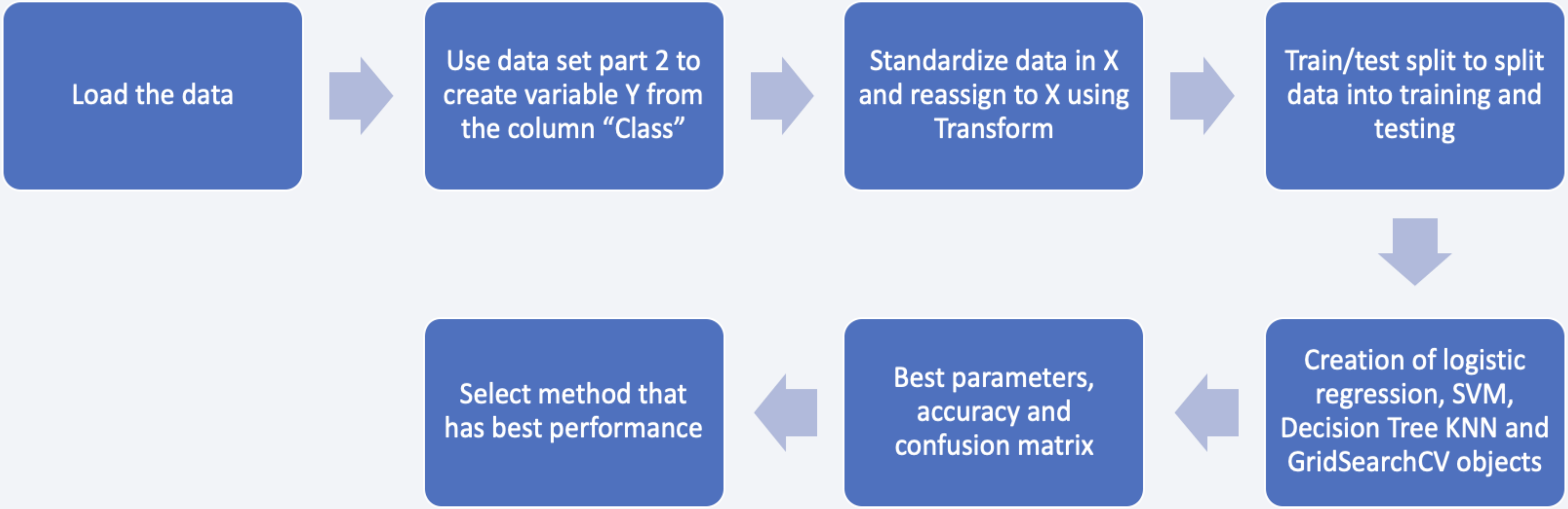
- All launch sites were marked and map objects (markers, circles, and lines) were added to the map to mark the success or failures of each launch. Additional features were added to help address questions as well.
- A highlighted circle area with a text label on a specific coordinate for each launch site on the site map was added
- Cluster objects were added to simplify the map that had many markers having the same coordinate
- Mouse position was added to have a pointer show up when hovering over the map
- Line were drawn between each launch site and its closest city, railway and highway.
- Launch-Site-Locations-with-Folium

Build a Dashboard with Plotly Dash

- This dashboard was created to preform interactive analysis of SpaceX launch data in real time.
- This dashboard contains various inputs
- Launch site drop down
- Callback function – this renders a success pie chart based on the selected site drop down.
- A range slider used to select payload
- [Interactive-Plotly-Dashboard]

Predictive Analysis (Classification)

- Through the machine learning process the best hyperparameters for logistic regression, SVM, Decision Trees and KNN classifiers were able to be determined
- SpaceX-Machine-Learning-Prediction-Notebook



Results

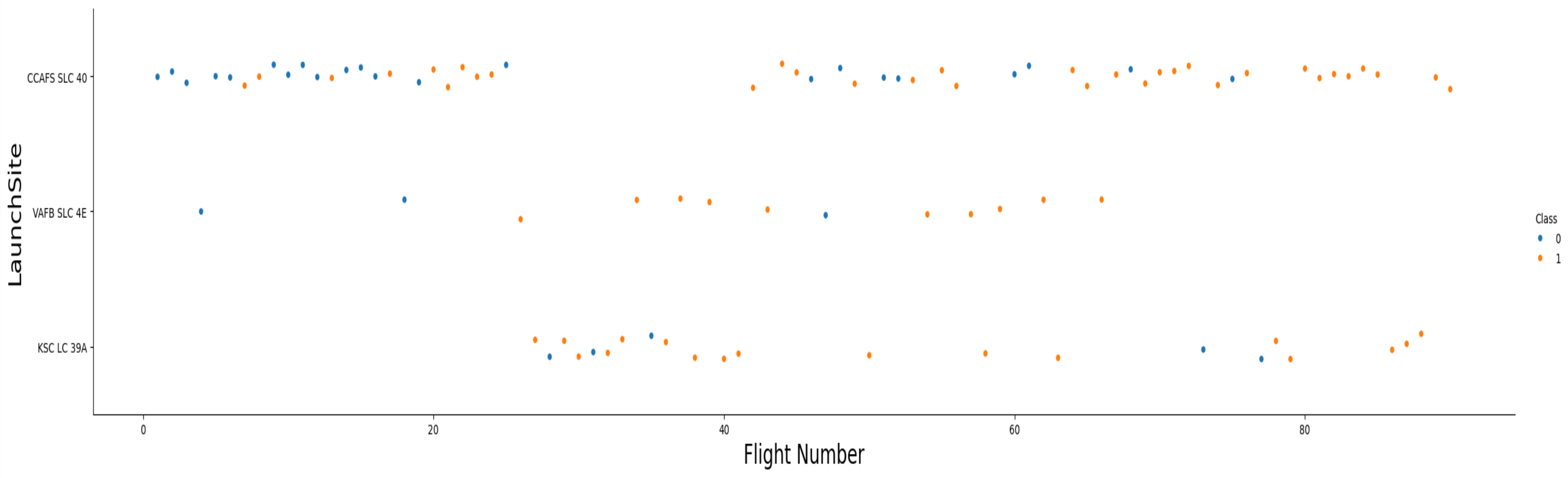
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

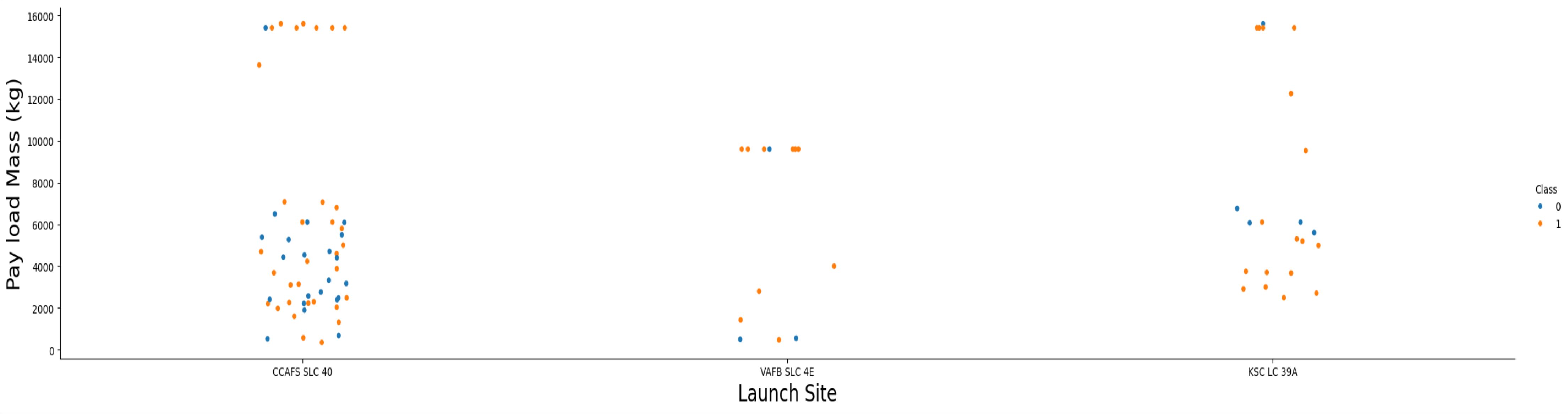
Flight Number vs. Launch Site

- Over time the success of the launch rates have increased at every launch site. Most notably at CCAFS SLC-40 where most launches take place
- VAFB and KSC represent 1/3 of the total launches



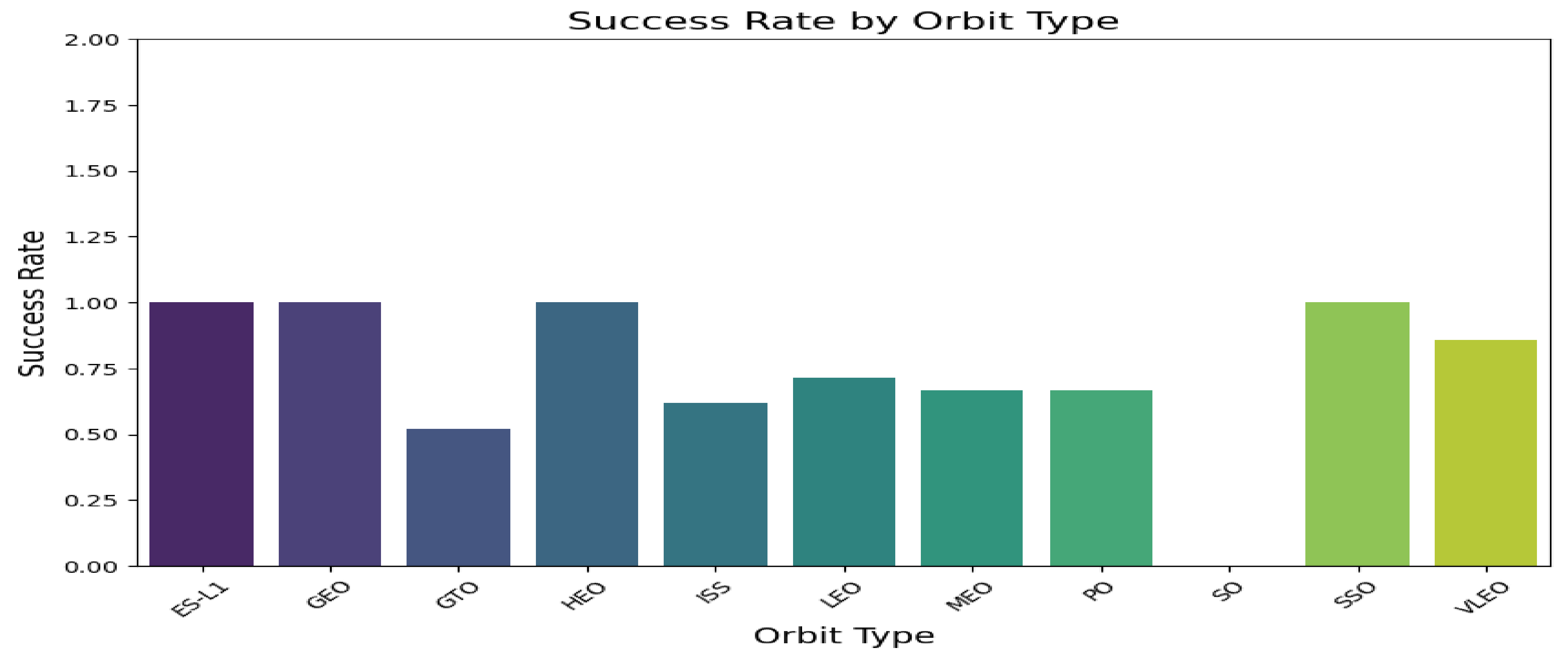
Payload vs. Launch Site

- VAFB does not launch rockets with a heavy payload mass (greater than 10000 kg)
- KSC does not launch rockets with a lower payload mass (less than 2500 kg)
- CCAFS has launched rockets less than 7500kg and more than 13000kg but does not launch rockets in-between that range



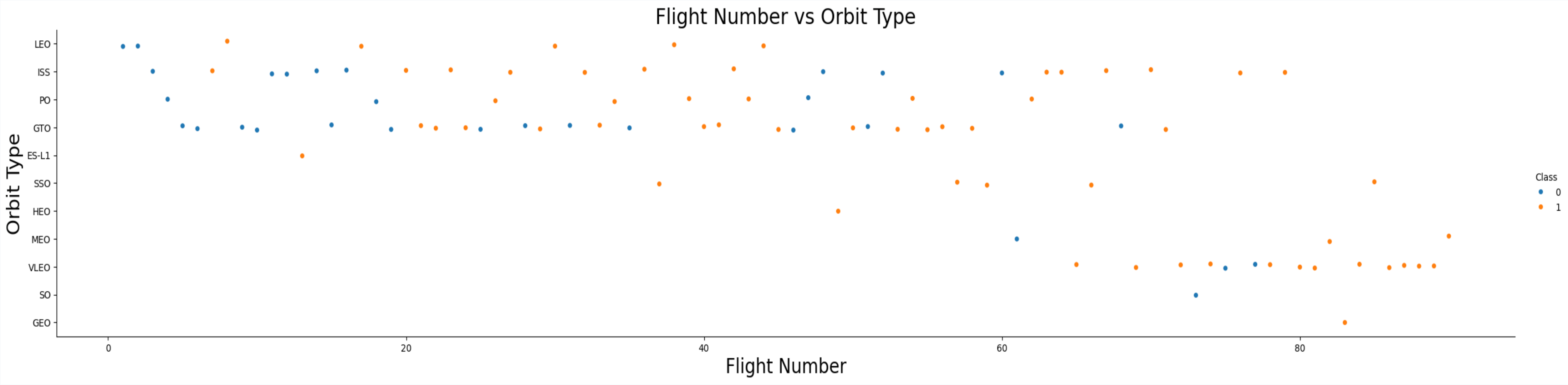
Success Rate vs. Orbit Type

- The first four (SSO, HEO, GEO, ES-L1) have the best success rate



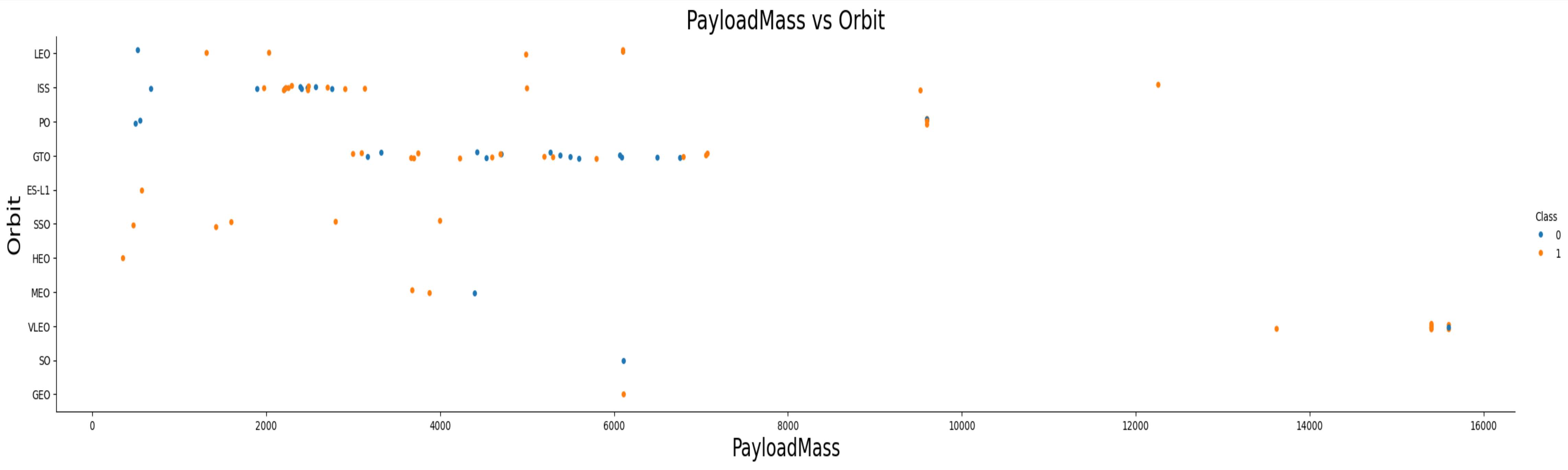
Flight Number vs. Orbit Type

- There are more failures at the beginning of launch series but as more progress the ratio improvs by the number of unsuccessful landings reducing
- GTO and ISS orbits have the highest concentration of launches with the lowest ratio of successful landings
- The orbits with a higher success rate have an overall fewer launches made



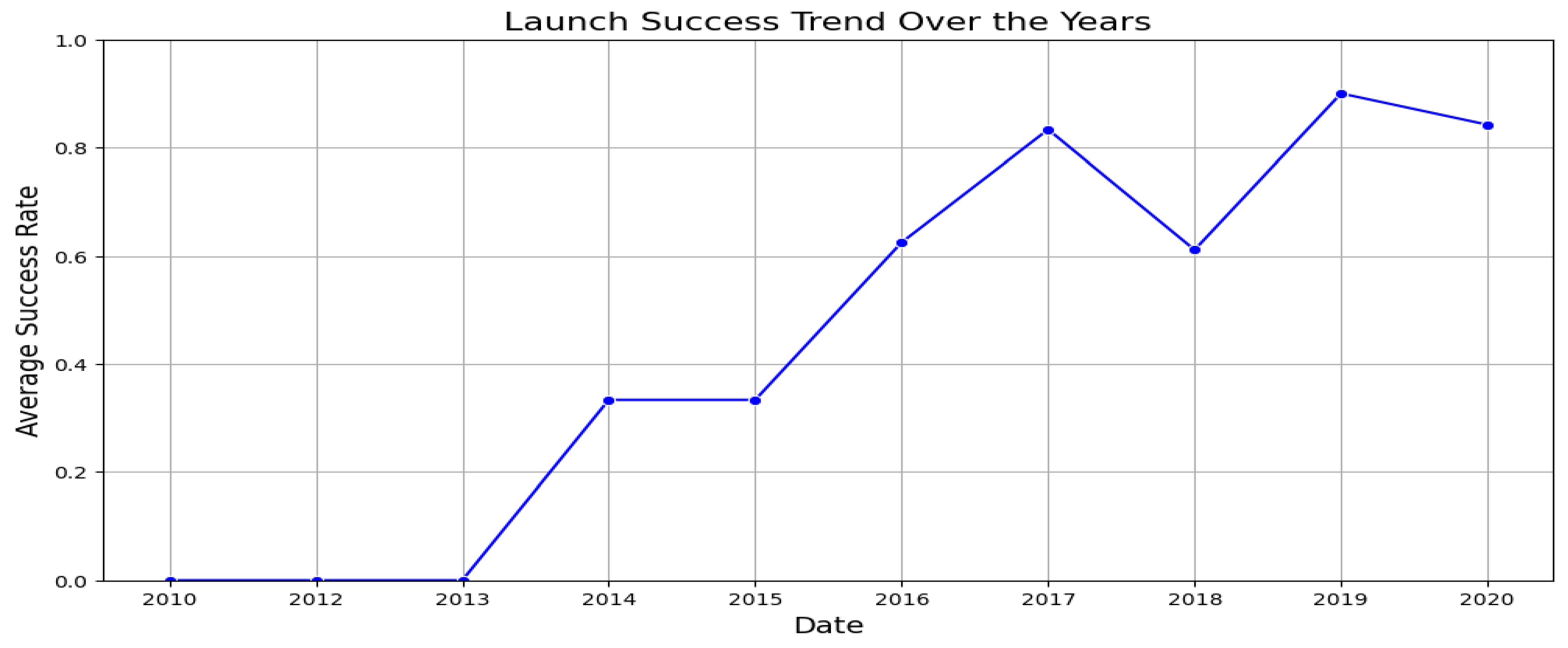
Payload vs. Orbit Type

- For heavy payloads PO, LEO and ISS have a higher landing success rate
- GTO has both positive and negative landing rates making it hard to determine overall success rate



Launch Success Yearly Trend

- The success rate continued to increase from 2013 - 2020



All Launch Site Names

- Four unique launch sites are – CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- "DISTINCT" in SQL finds all unique values
- SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * 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'

- The query used SELECT * FROM SPACEXTABLE where "Launch_Site" like 'CCA%' LIMIT 5

```
%sql SELECT * FROM SPACEXTABLE 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_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Suc
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	Suc
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Suc
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Suc
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Suc

Total Payload Mass

- Using SUM and WHERE the total payload mass carried by booster launched by NASA (CRS) was calculated
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer ='NASA (CRS)';

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Using AVG() function the average payload mass carried by booster version F9 v1.1 was calculated
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version ='F9 v1.1';

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- This is the data from the first successful ground pad landing outcome. This was achieved using MIN
- SELECT MIN("Date") AS "First_Landing_Date" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)";

```
%sql SELECT MIN("Date") AS "First_Landing_Date" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success
```

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

First_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. This is found using WHERE and AND clause together
- SELECT * FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000

```
%sql SELECT * FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AN
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes. This uses COUNT with GROUP BY
- SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE "Mission_Outcome" = "Success"
- SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "%Failure%"
-

```
Total_Success = %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE "Mission_Outcome" = "Success"
Total_Failure = %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "%Failure%"

print(Total_Success)
print(Total_Failure)
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
+-------------------------+
| COUNT(Mission_Outcome)  |
+-------------------------+
|           98            |
+-------------------------+
+-------------------------+
| COUNT(Mission_Outcome)  |
+-------------------------+
|           1             |
+-------------------------+
```

Boosters Carried Maximum Payload

- These are the names of the booster versions which have carried the max payload mass, this was found using subquery
- SELECT Booster_Version FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

```
MPM = %sql SELECT Booster_Version FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") F

print(MPM)
```

```
 * 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

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- SELECT substr("Date", 6, 2) AS "Month", Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Failure (drone ship)%" AND substr("Date", 1, 4) = '2015';

```
%sql SELECT substr("Date", 6, 2) AS "Month", Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE
```

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

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

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order. Query uses COUNT, WHERE, BETWEEN and GROUP BY
- SELECT Landing_Outcome, COUNT("Landing_Outcome") AS "Count" FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY "Count" DESC;

```
%sql SELECT Landing_Outcome, COUNT("Landing_Outcome") AS "Count" FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-0
```

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

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

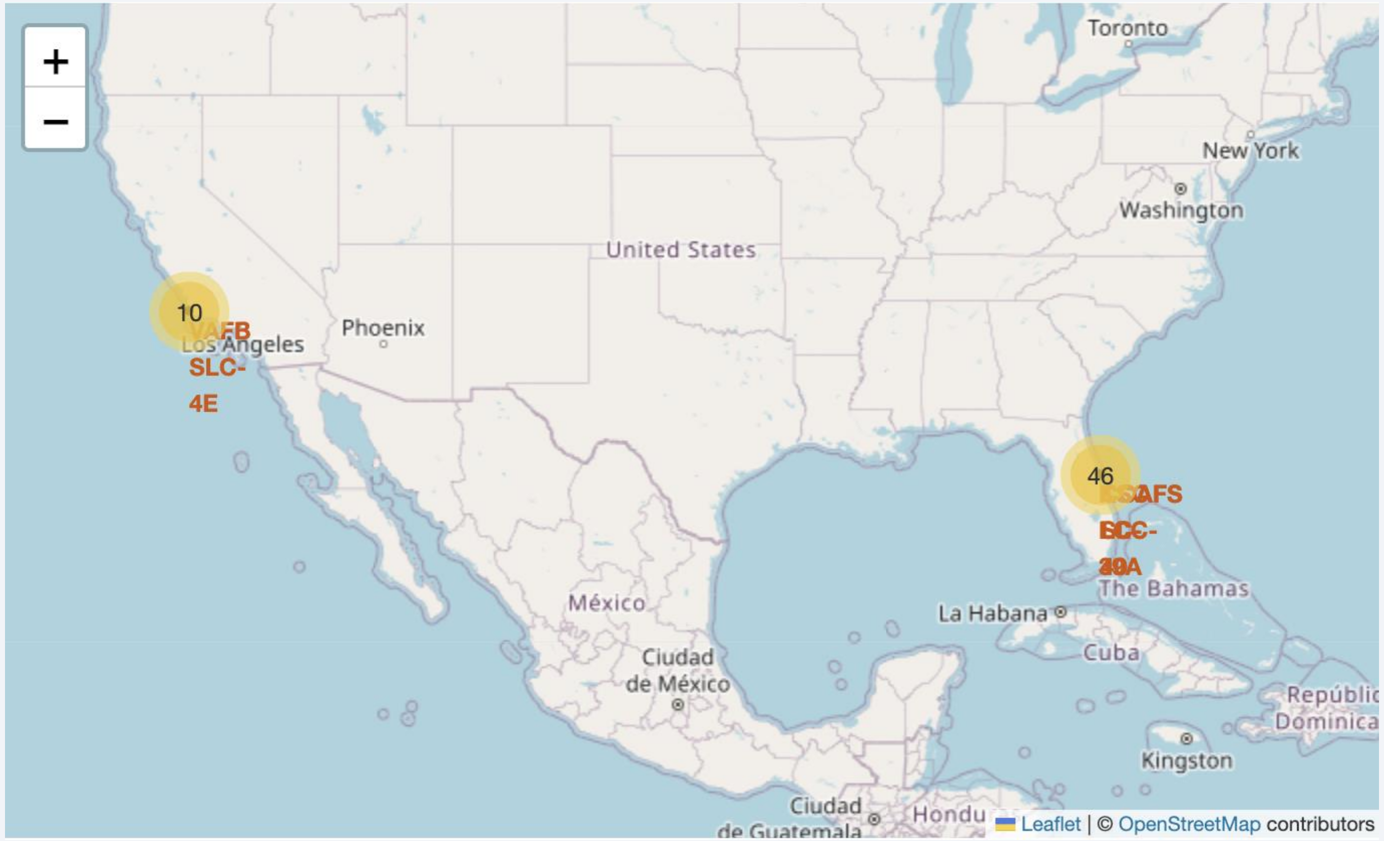
Launch Sites Proximities Analysis

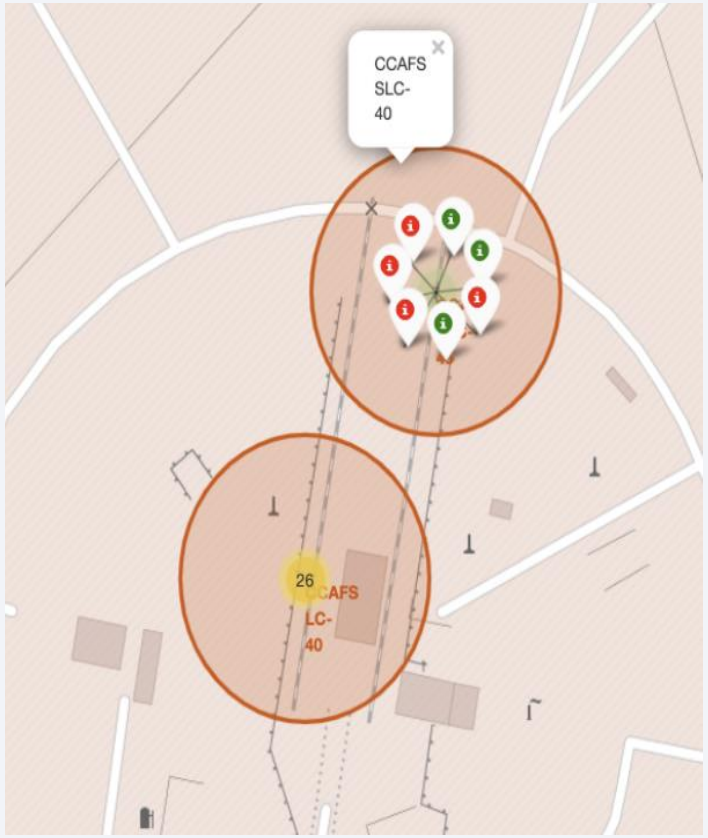
All Launch Sites

- Launch sites are close to the coast and within the US California and Florida

Success/Failed Launches per Site

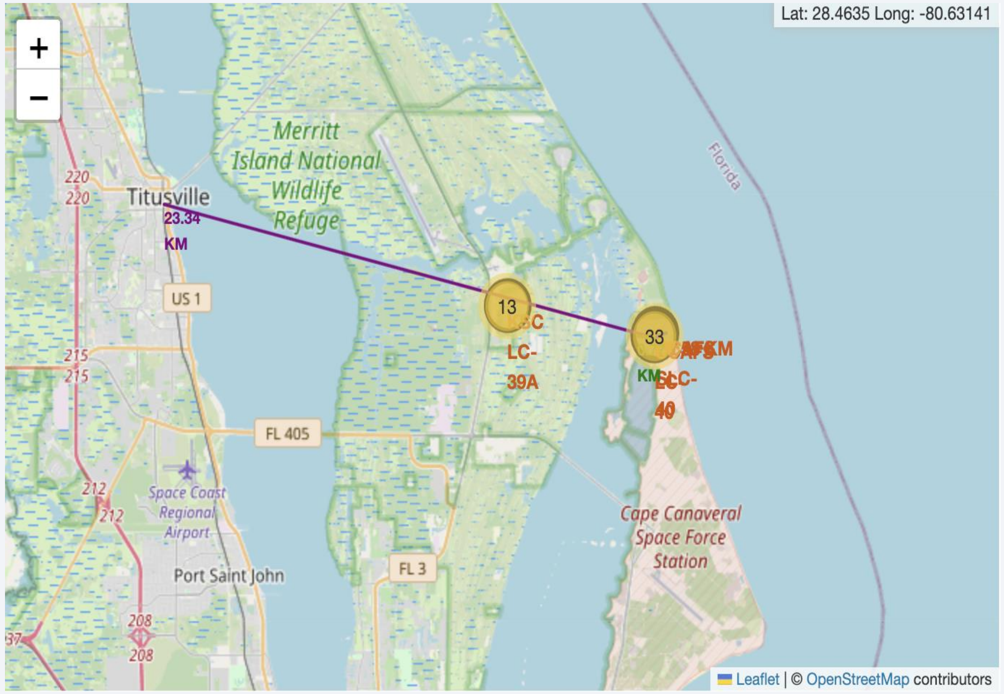
- The first map shows clusters for each launch site. The second map shows green markers if the launch was successful and red markers if it was not

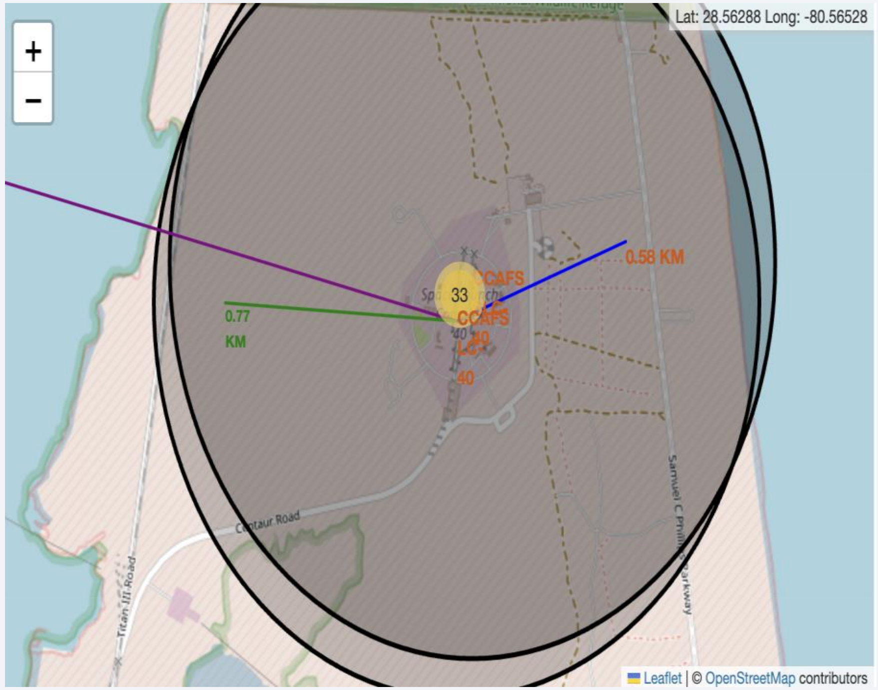




Proximities to Launch Site

- This shows how close the launch sites are to railways, roads, coastlines, etc.
- Here, nearest coastline is 0.58km, railways is 0.77 km and roads is 23.34 km



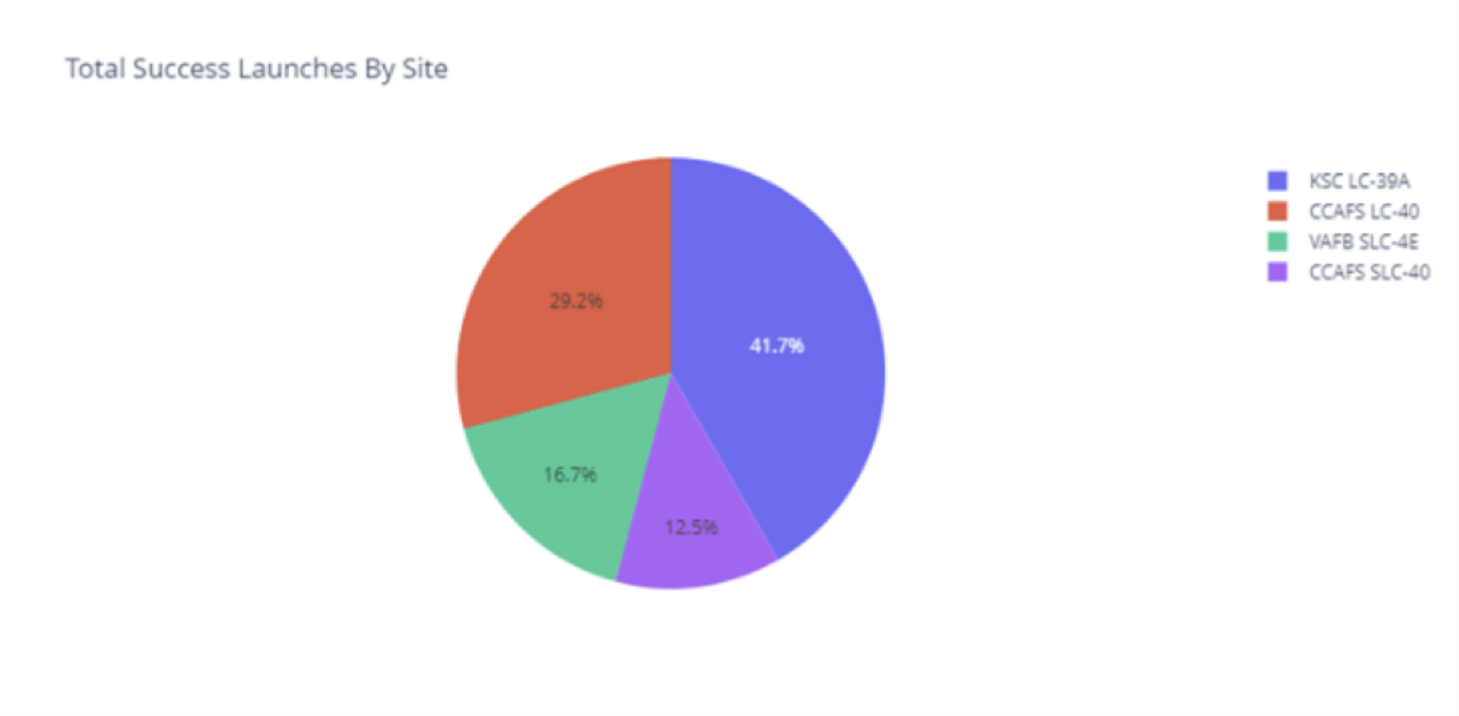


Section 4

Build a Dashboard with Plotly Dash

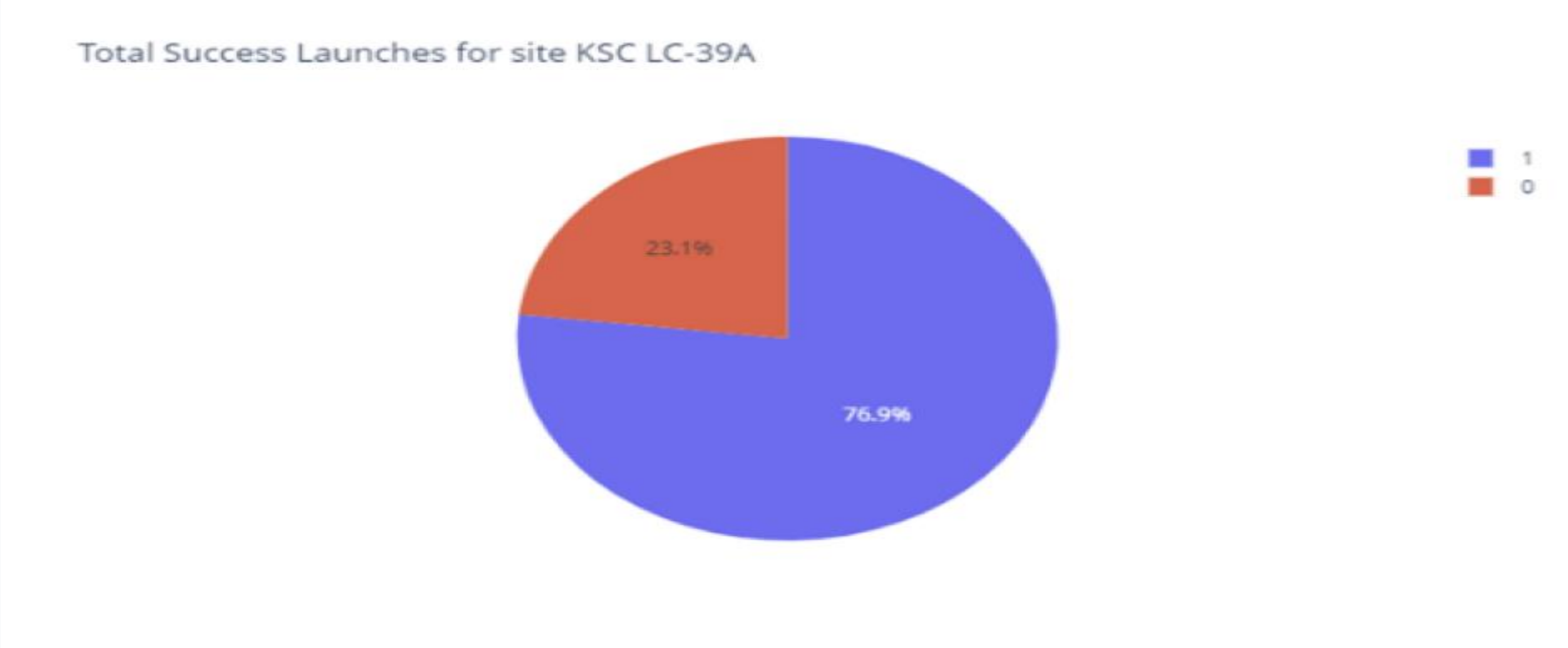
Total Success Launches by All Site

- KSC has the highest successful launches followed by CCAFS



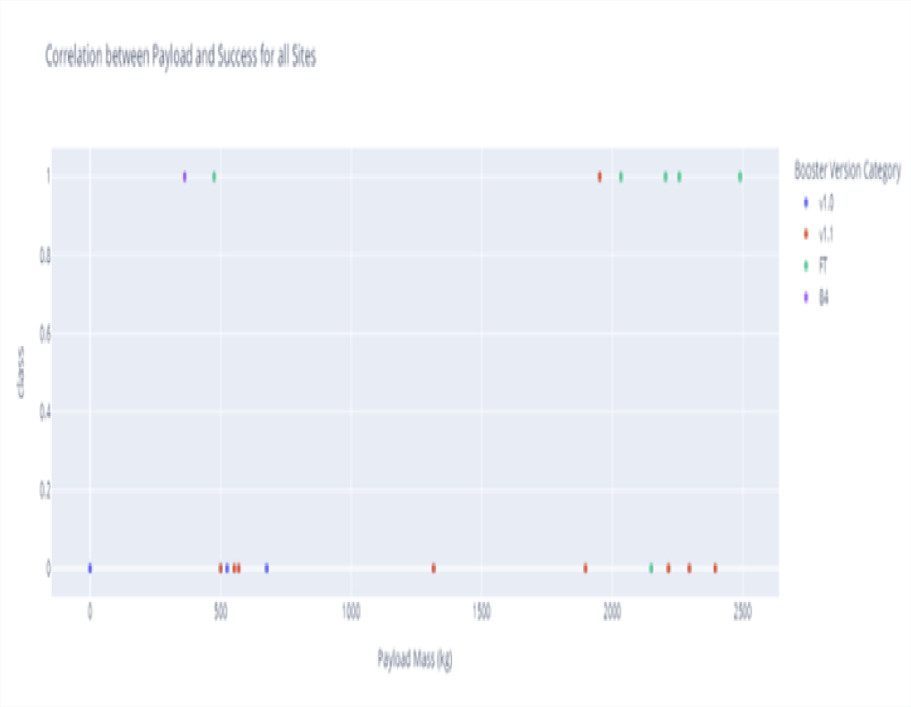
Highest Launch Success Ratio

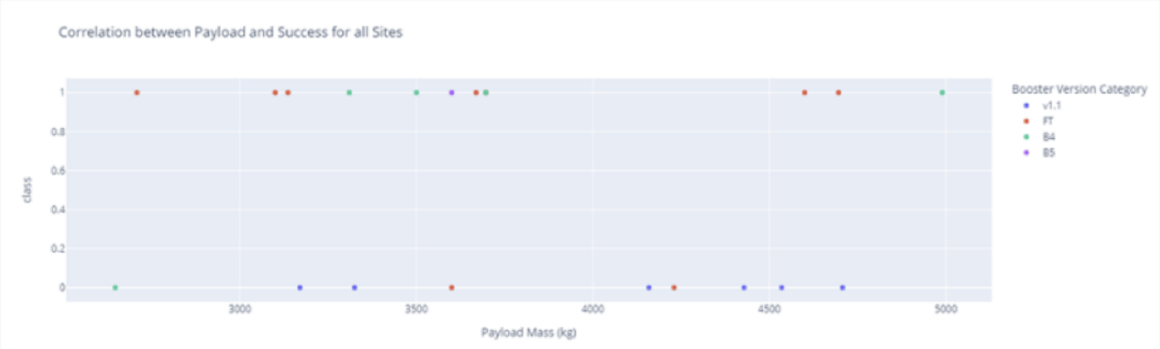
- This pie chart sows KSC as the site which the highest launch success ratio

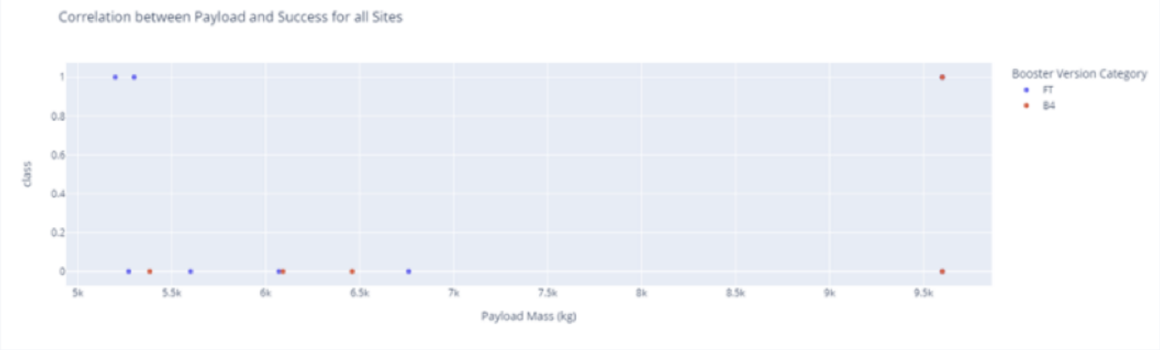


Payload VS Launch Outcome

- Scatter plot for all sites (2500kg, 5000kg and 10000kg) payload ranges
- Payload range 2500-5000kg have most of the successful launches





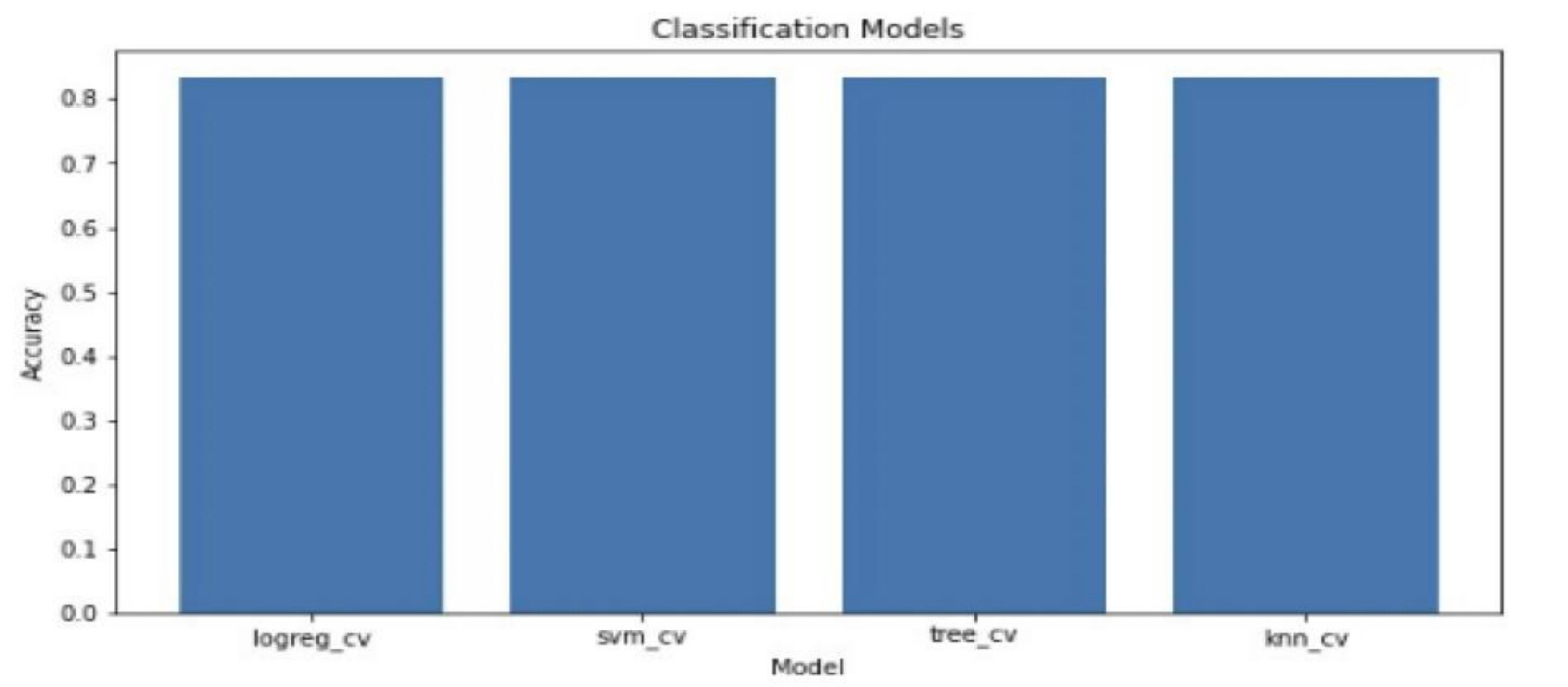


Section 5

Predictive Analysis (Classification)

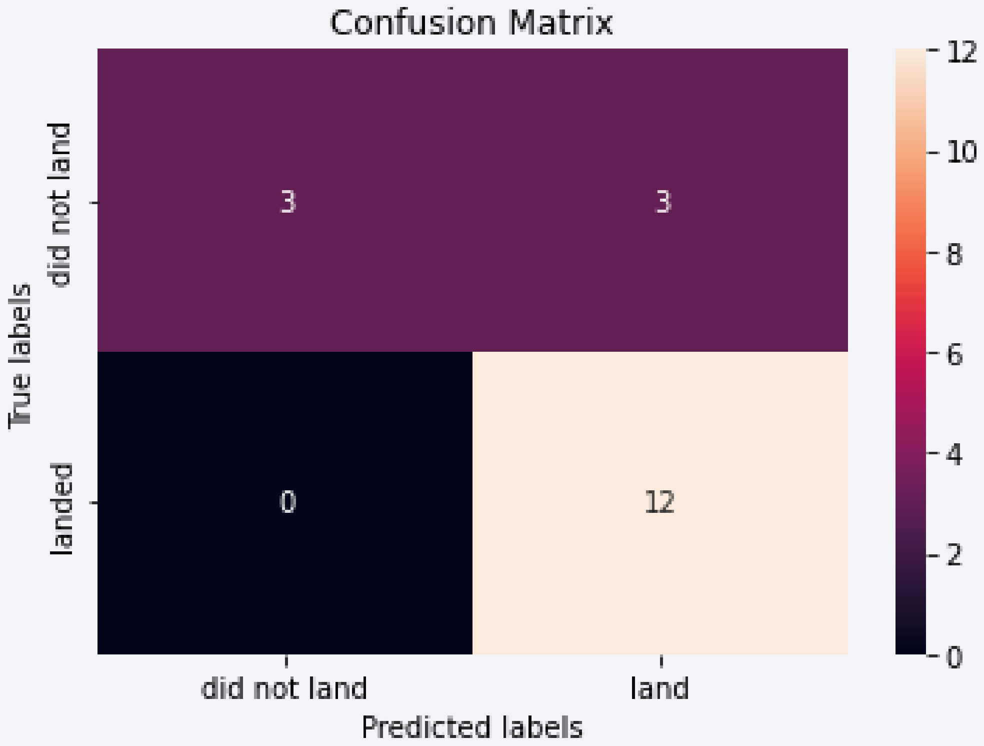
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside
- The classifier model with the highest accuracy is the Decision Tree



Confusion Matrix

- The Decision Tree confusion matrix shows the largest number of true positive and true negative compared to other models.



Conclusions

In conclusion:

- The more flights that take place at a launch location the higher the chances of are of a successful launch
- Successful launches increased from 2013-2020
- KSC LC-39A had the most successful launches of any site
- All machine learning models seem to have the same amount of accuracy, but the Decision Tree model gave the most accurate results.

Appendix

- For additional information / raw files of the above please go to this GitHub Repository.
- Git-Hub-Repository

Thank you!