IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Erdi Köse
January 11, 2023

Outline

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

Executive Summary

Summary of methodologies

- Collecting public SpaceX data via a RESTful API and web scraping from wiki page
- Data wrangling, including filtering, one hot encoding and missing value imputation
- Exploratory data analysis (EDA) using visualization techniques and SQL queries to gather insights
- Performing interactive visual analysis with folium for interactive geospatial analysis and Plotly Dash for interactive dashboard
- Building predictive models (classification) namely Logistic Regression, Support Vector Machines, Classification Trees and k-Nearest Neighbors.

Summary of all results

- Even though their training performances vary, all models make the same predictions with an 83.33% accuracy (15 out of 18).
- This project will also demonstrate some meaningful insights from the outputs of EDA, interactive maps and dashboard.

Introduction

Project Background

- At the dawn of the commercial space age, companies are making space tourism more affordable. SpaceX, probably the most successful of them, launches Falcon 9 with a cost of $62 million; other rocket providers cost upwards of $165 million each.
- Much of the savings is because SpaceX, unlike others, can reuse the first stage. Therefore, if we can predict whether the first stage will successfully land or not, we can determine the cost of a launch.

Problems

- Predicting if the Falcon 9 first stage will land successfully
- Determining the price of each launch
- Determining if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

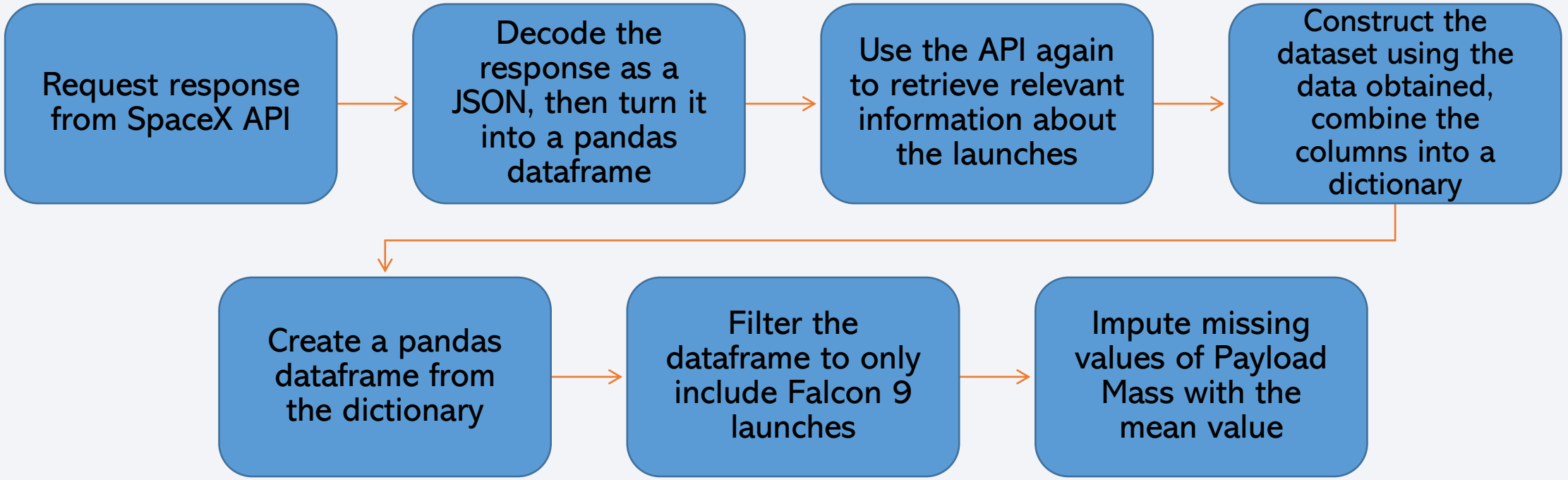
- Data collection methodology
  - Data was gathered by using RESTful API and web scraping.
- Perform data wrangling
  - Filtering Falcon 9 launches, imputing missing values of ‘Payload Mass’ with mean value, converting landing outcomes into a binary variable (numeric)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize the data with Standard Scaler, then split it into training and test data sets.
  - Train four different classification models tuned by using GridSearchCV and compare them according to their accuracy.

Data Collection

SpaceX launch data is collected from SpaceX REST API and web scrape the List of Falcon 9 and Falcon Heavy launches Wikipedia page.

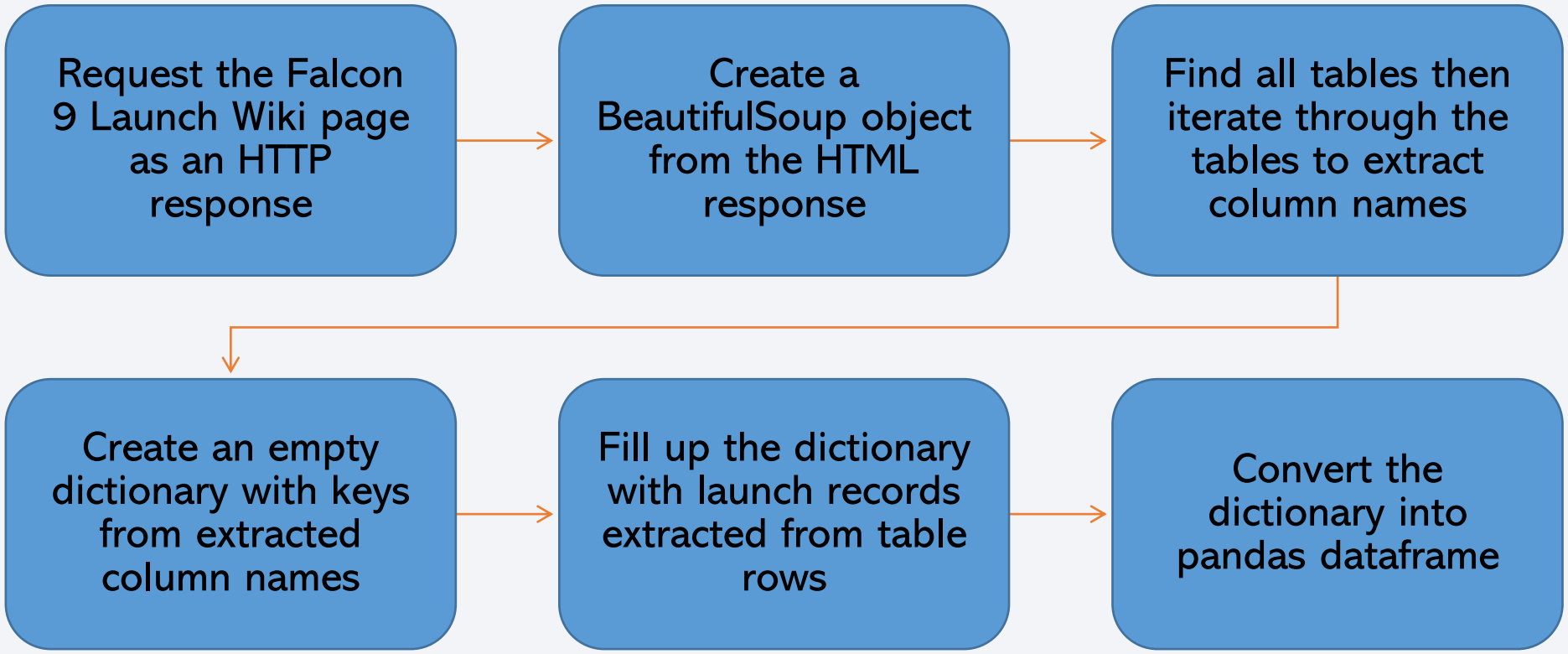
- SpaceX REST API
  - Using SpaceX API to get data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome.
- Web scraping
  - Using the Python BeautifulSoup package to web scape some HTML tables that contain Falcon 9 launch records.

Data Collection – SpaceX API



GitHub Link

Data Collection - Scraping



GitHub Link

Data Wrangling

- Identify and calculate the percentage of missing values
- Determine the number of launches on each site
- Determine the number and occurence of each orbit
- Create 'Landing Class' label from 'outcome' column:
  - There are 8 types of outcome in the outcome column
  - e.g: 'True Ocean' means mission outcome was successfully landed (True) to a specific region of the ocean (Ocean) while False RLTS means the mission outcome was unsuccessfully landed (False) to a ground pad (RLTS)
  - The outcomes False ASDS, False Ocean, False RTLS, None ASDS and None None represent unsuccessful first stage and are labeled **0**.
  - Similarly, True ASDS, True Ocean and True RTLS represent successful missions and are labeled **1**.

GitHub Link

EDA with Data Visualization

Scatter plots are made to study the relationship between two numeric variables with the option of showing subgroups.

- Flight Number vs Launch Site
- Payload vs Launch Site
- Flight Number vs Orbit Type
- Payload vs Orbit Type

Bar charts show the relationship between one numeric and one categorical variable. An ordered bar chart is a good choice to provide additional information.

- Success Rate by Orbit Type

Line charts display the evolution of one or several numeric variables and are often used to visualize time trend.

- Success Rate Yearly Trend

GitHub Link

EDA with SQL

To have a better understanding of the dataset, the following SQL queries are performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the ‘CCA’
- Display the total payload mass carried by boosters launched by NASA
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- 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 outcome in drone ship, booster versions and the launch sites for the months in year 2015
- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20

GitHub Link

Build an Interactive Map with Folium

Mark all launch sites on a map

- Add a marker and a circle for each launch site

Mark the success/failed launches for each site.

- Cluster them since there are many markers having same coordinates
- Color the markers (green for successful and red for failed launches) so that we can easily identify which launch sites have relatively high success rate

Calculate the distances between a launch site to its proximities

- Thus, we can have a general knowledge regarding the location of a launch site.

GitHub Link

Build a Dashboard with Plotly Dash

The interactive dashboard consists of two visualization methods:

Pie chart of the total successful landings per launch site

- Shows the percentage of successful landings by sites. (default option)
- By filtering launch sites, we can discover and compare launch success rates and counts of both successful and unsuccessful landings for different sites.

Scatter plot of landing outcome vs. payload mass for all sites grouped by booster versions

- Can be modified by filtering booster versions and/or payload mass range to extract booster version and payload mass specific insights.

GitHub Link

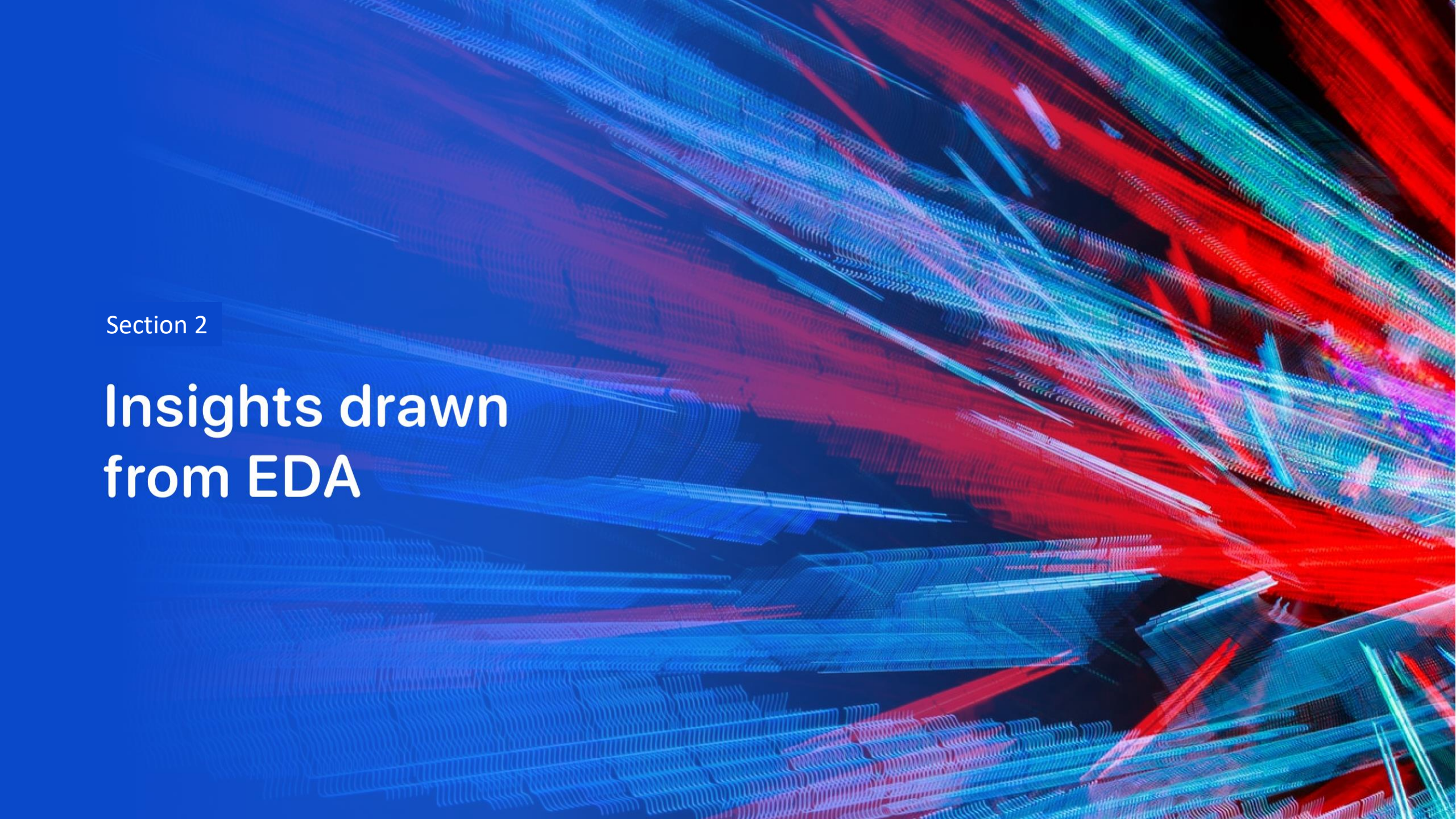
Predictive Analysis (Classification)

- Create a ‘Class’ column.
- Standardize the data with Standard Scaler.
- Split the data into train and test data.
- Train four classification models: Logistic Regression, SVM, Decision Tree Classifier and kNN. Find best hyperparameters with grid search, for each model.
- Calculate the accuracy on the test data and create a confusion matrix for each tuned model.
- Compare the models based on their accuracy on the test data to determine best performing model.

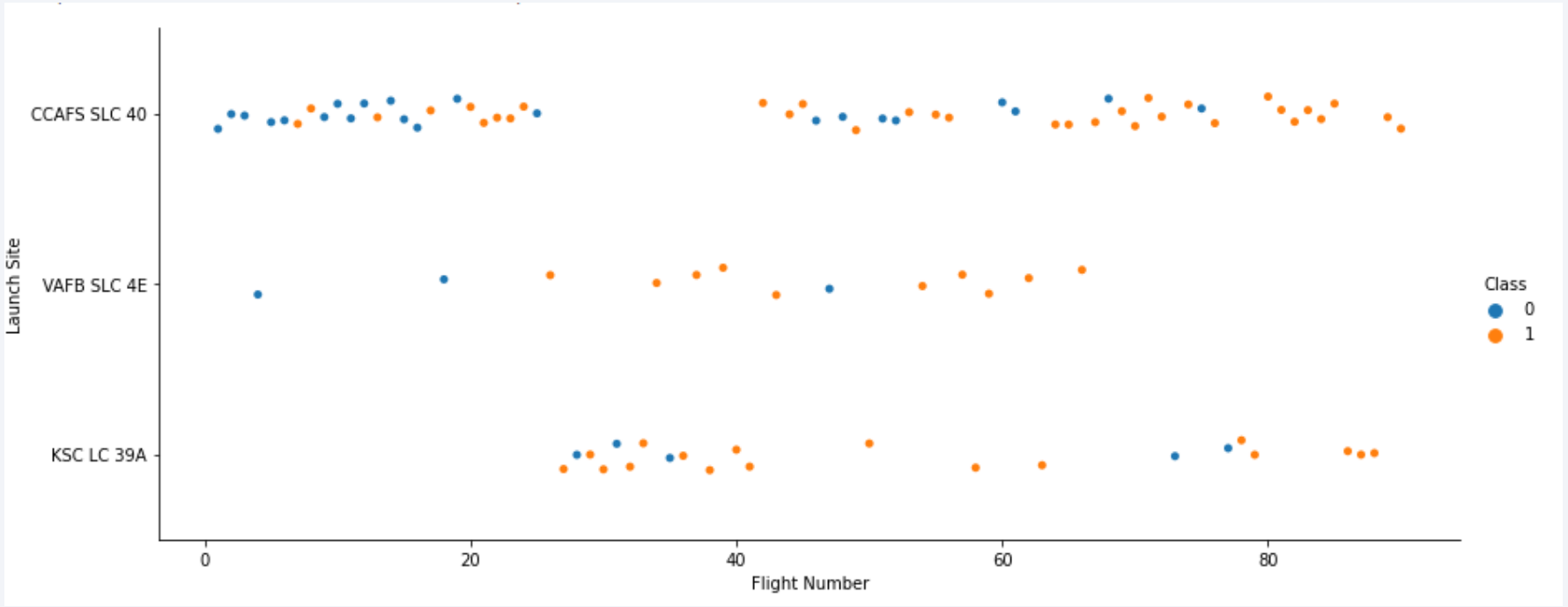
GitHub Link

Results

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


Section 2
Insights drawn from EDA

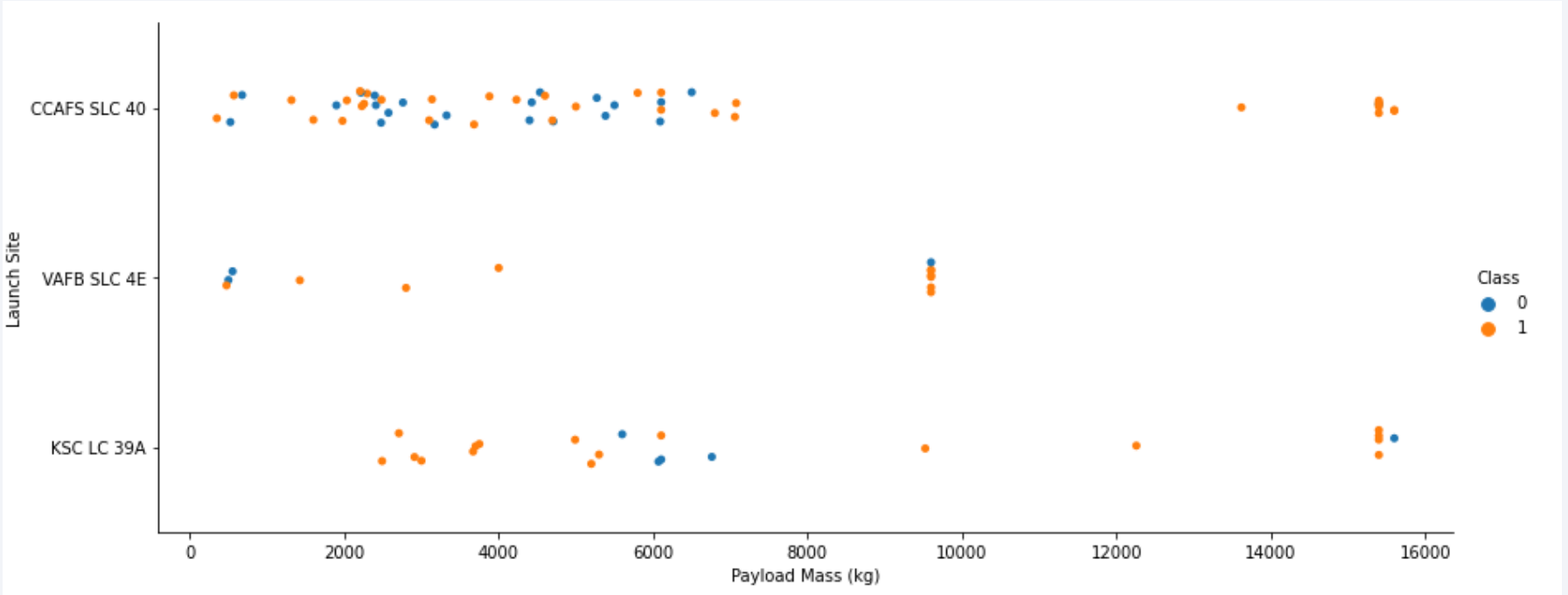
Flight Number vs. Launch Site



- CCAFS SLC 40 hosted more launches than other launch sites but there is no launches from here after flight #25 up until flight #40.

- Early launches were mostly unsuccessful. As time progresses, the success rate increases for all three launch sites.
- Excluding early launches of CCAFS SLC 40, all three launch sites have similar success rates.

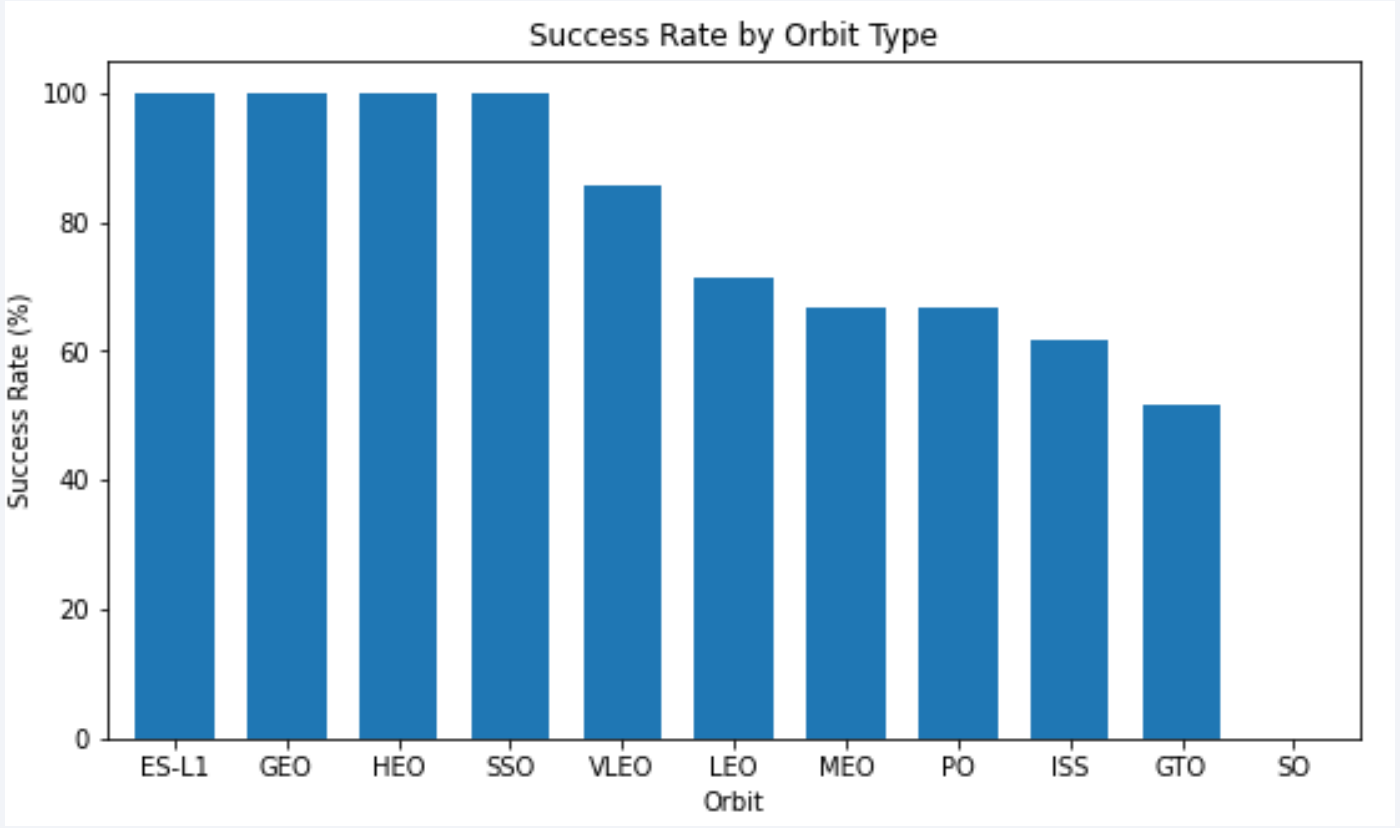
Payload vs. Launch Site



- Every launch from VAFB SLC 4E has a payload mass of less than 10,000 kilograms.

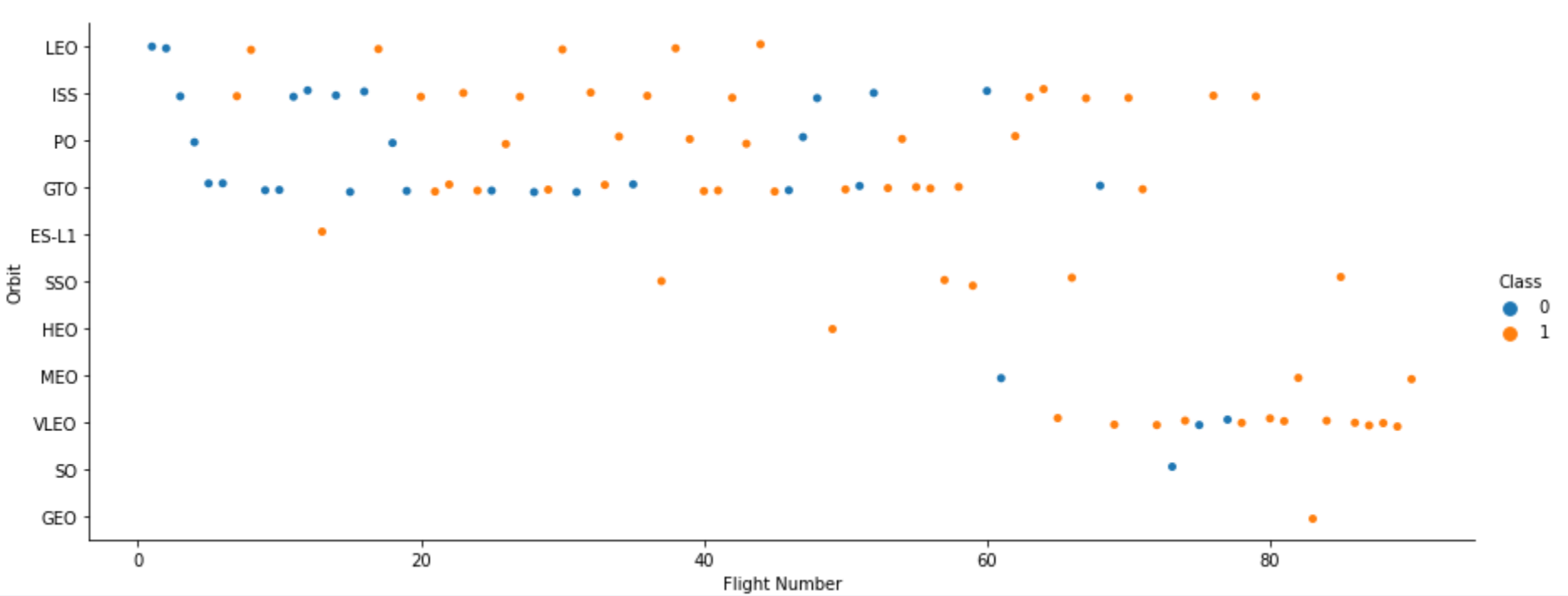
- Most launches have a payload mass less than 7,000 kilograms but above this level of payload mass, landing success rate is higher.
- However, there is no linear correlation between payload mass and landing success rate observed for a given launch site.

Success Rate vs. Orbit Type



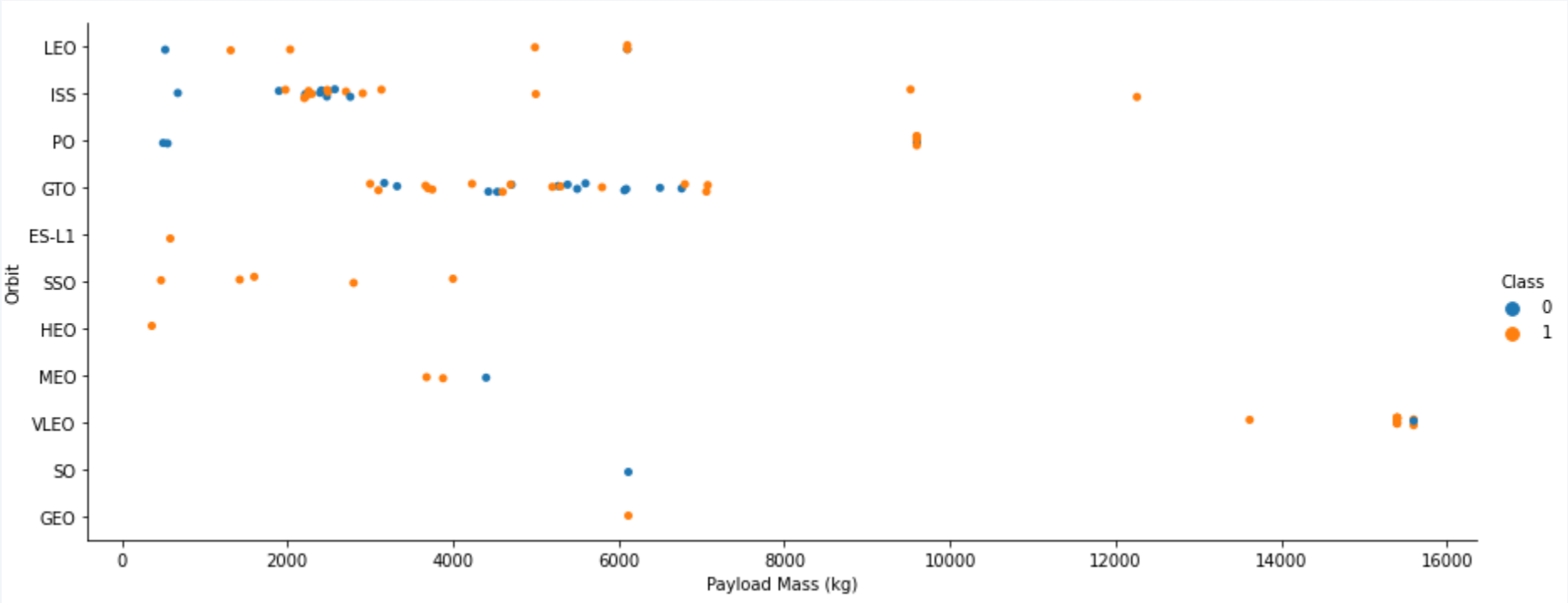
- ES-L1, GEO, HEO, SSO have 100% success rate, but only SSO has multiple attempts (5).
- SO has the lowest success rate with 0% (0 for 1).
- 9 out of 11 orbit types have at least 60% success rate.

Flight Number vs. Orbit Type



- By the time, orbit preference shifted to Very Low Earth Orbit (VLEO). Low Earth Orbit (LEO) and Polar Orbit (PO) were not used since VLEO appeared.
- ISS and GTO are most common orbit types; however, after Flight #80 they are not seen anymore.

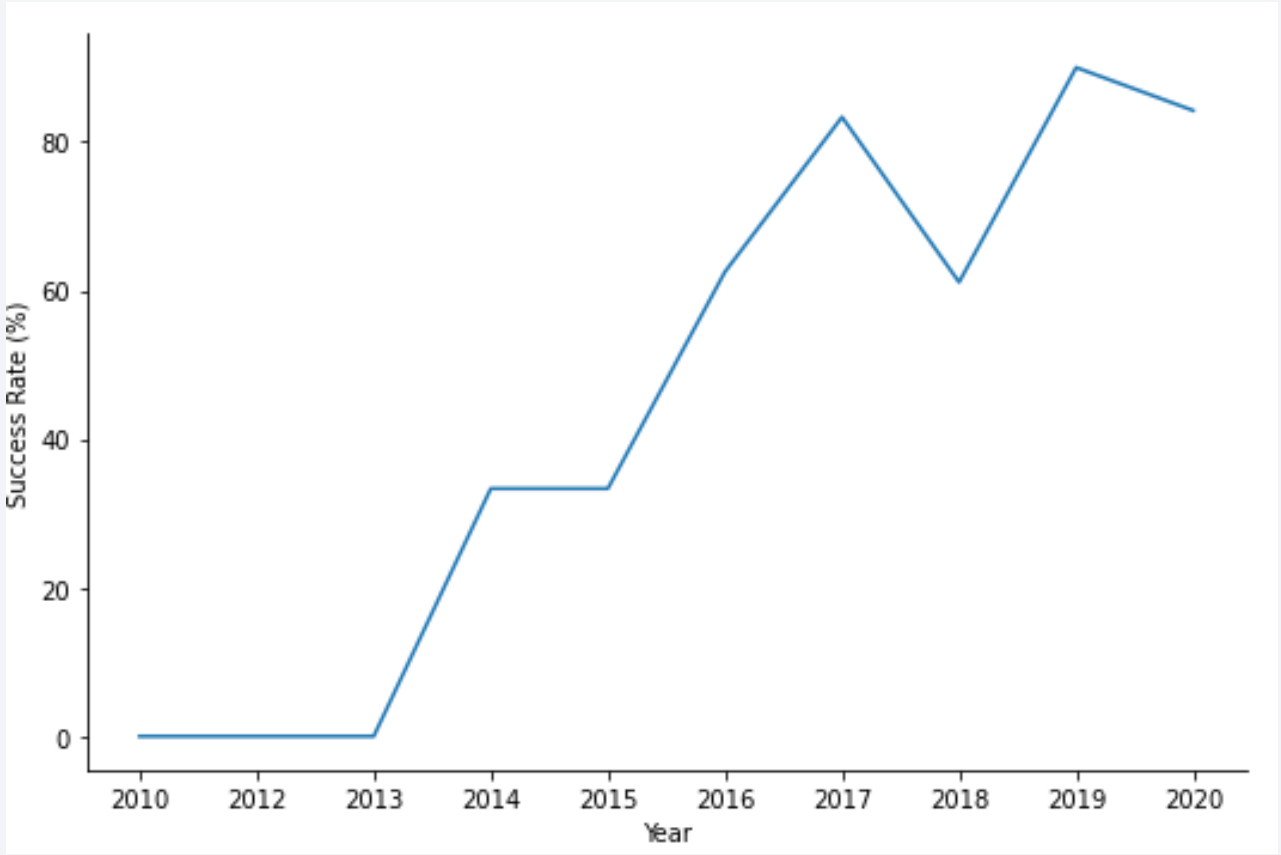
Payload vs. Orbit Type



- VLEO is preferred for the heaviest payloads.
- GTO has a compact and well-balanced payload mass distribution.

Launch Success Yearly Trend

- The success rate tends to increase over time.
- From 2010 to 2013, all launches were unsuccessful. Also note that, there were no launches in year 2011.
- For the years 2019 and 2020, success rate was above 80%.



All Launch Site Names

There are 4 unique launch sites in the space mission.

```
%%sql
select distinct LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

The first 5 records where launch sites begin with `CCA` are as follows:

```
%%sql
select * from SPACEXTBL
where LAUNCH_SITE like 'CCA%'
limit 5
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload carried by boosters from NASA

```
%%sql
select SUM(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTBL
where Customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 is slightly below 3000 kg.

```
%%sql
select AVG(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTBL
where Booster_Version = 'F9 v1.1'
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

average_payload_mass
2928

First Successful Ground Landing Date

The first successful landing outcome on ground pad was on December 22, 2015.

```
%%sql
select MIN(Date) as "Date" from SPACEXTBL
where Landing__Outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as follows:

```
%%sql
select Booster_Version from SPACEXTBL
where Landing__Outcome = 'Success (drone ship)' and
      (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
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

There are 99 successful mission outcomes (not to be confused with landing outcome) out of 101 launches.

```
%%sql
select Mission_Outcome, count(*) as frequency from SPACEXTBL
group by Mission_Outcome
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

mission_outcome	frequency
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

```
%%sql
select distinct Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

These are the records which will display the month names, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015.

```
%%sql
select monthname(Date) as "Month", Landing__Outcome, Booster_Version, Launch_Site from SPACEXTBL
where Landing__Outcome = 'Failure (drone ship)' and substr(Date, 1, 4) = '2015'
```

```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

Month	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

The count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
select Landing__Outcome, COUNT(Landing__Outcome) as frequency from SPACEXTBL
where (Date between '2010-06-04' and '2017-03-20') and Landing__Outcome like 'Success%'
group by Landing__Outcome
order by frequency DESC
```

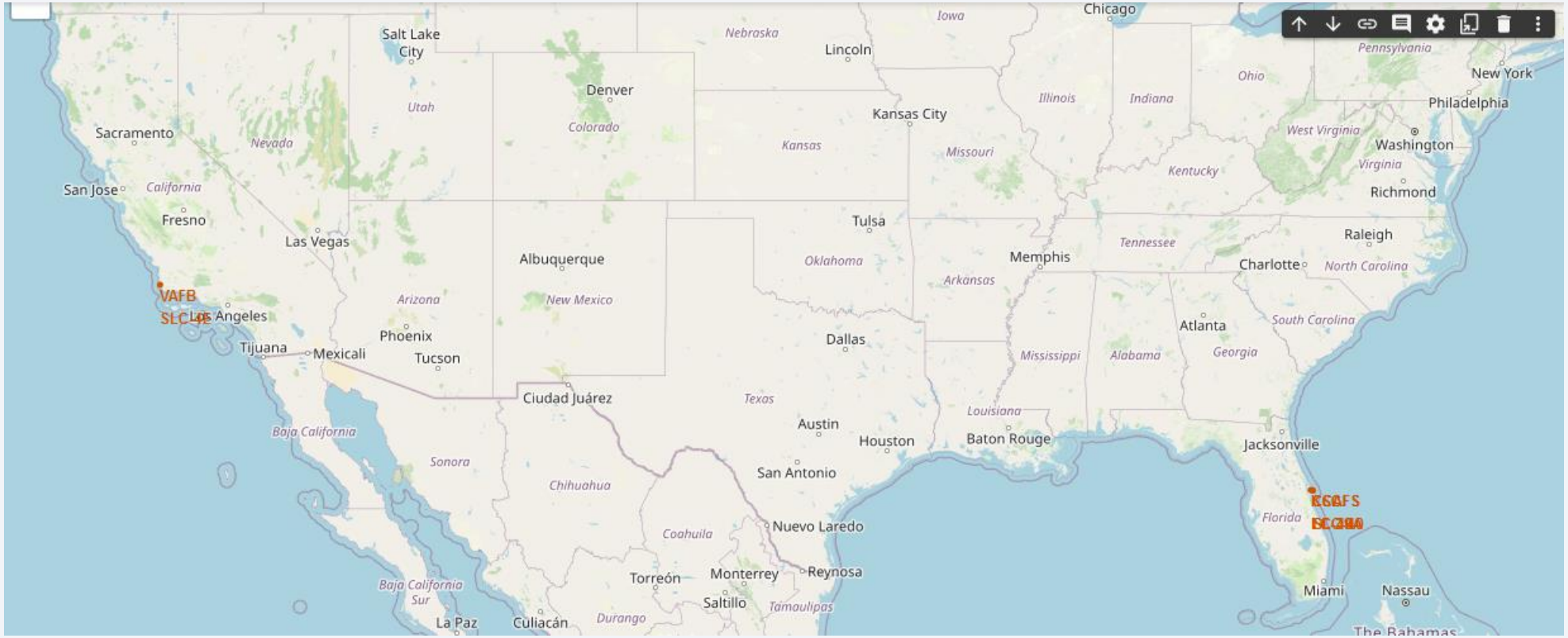
```
 * ibm_db_sa://gws33216:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/BLUDB
Done.
```

landing__outcome	frequency
Success (drone ship)	5
Success (ground pad)	3

Section 3

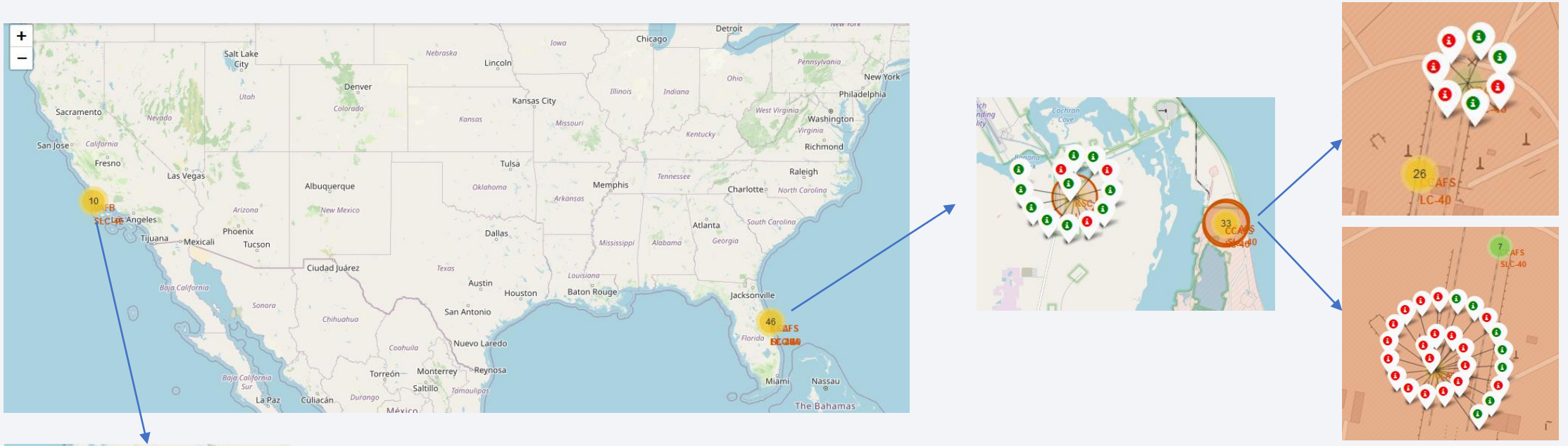
Launch Sites Proximities Analysis

All Launch Site Locations



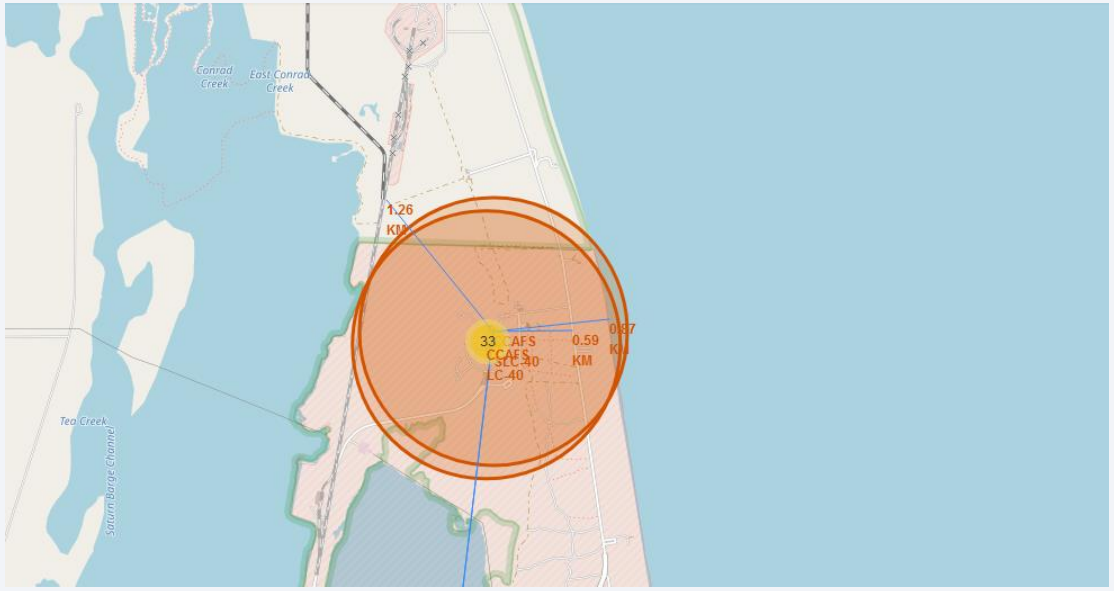
All launch sites are on coasts of California and Florida, two of the largest states in the U.S. (One site in California and three sites in Florida)

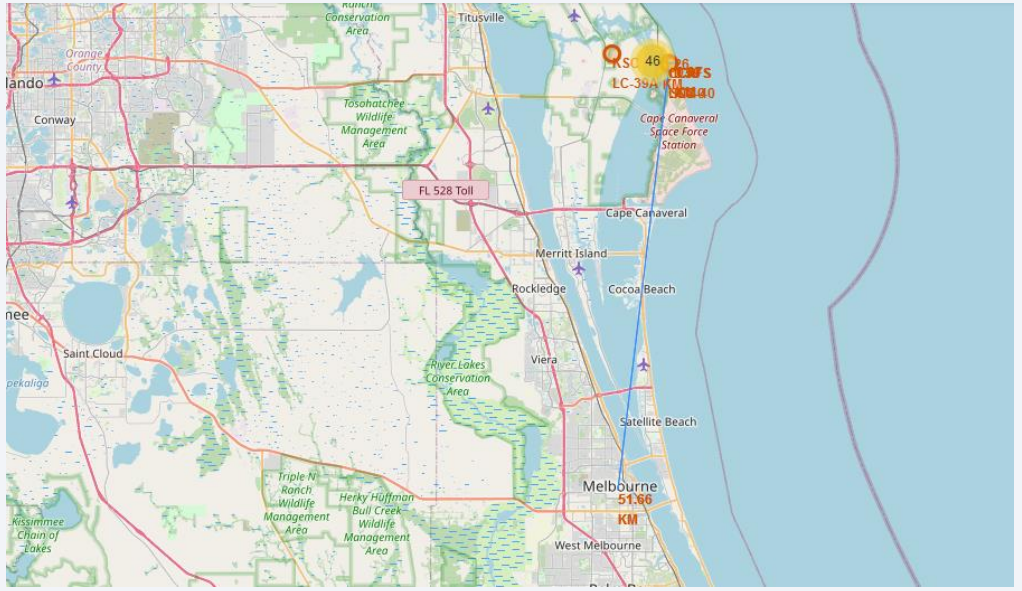
Successful and Failed Launches with Color-Labeled Markers



There were 10 launches in California (VAFB SLC-4E). 4 of them were successful and 6 of them were failed. KSC LC-39A is, by far, the most successful site among them with 10 successful launch out of 13 (both by percentage and count). CCAFS LC-40 has most launches with 26 but 19 of them were failed which makes it the least successful launch site.

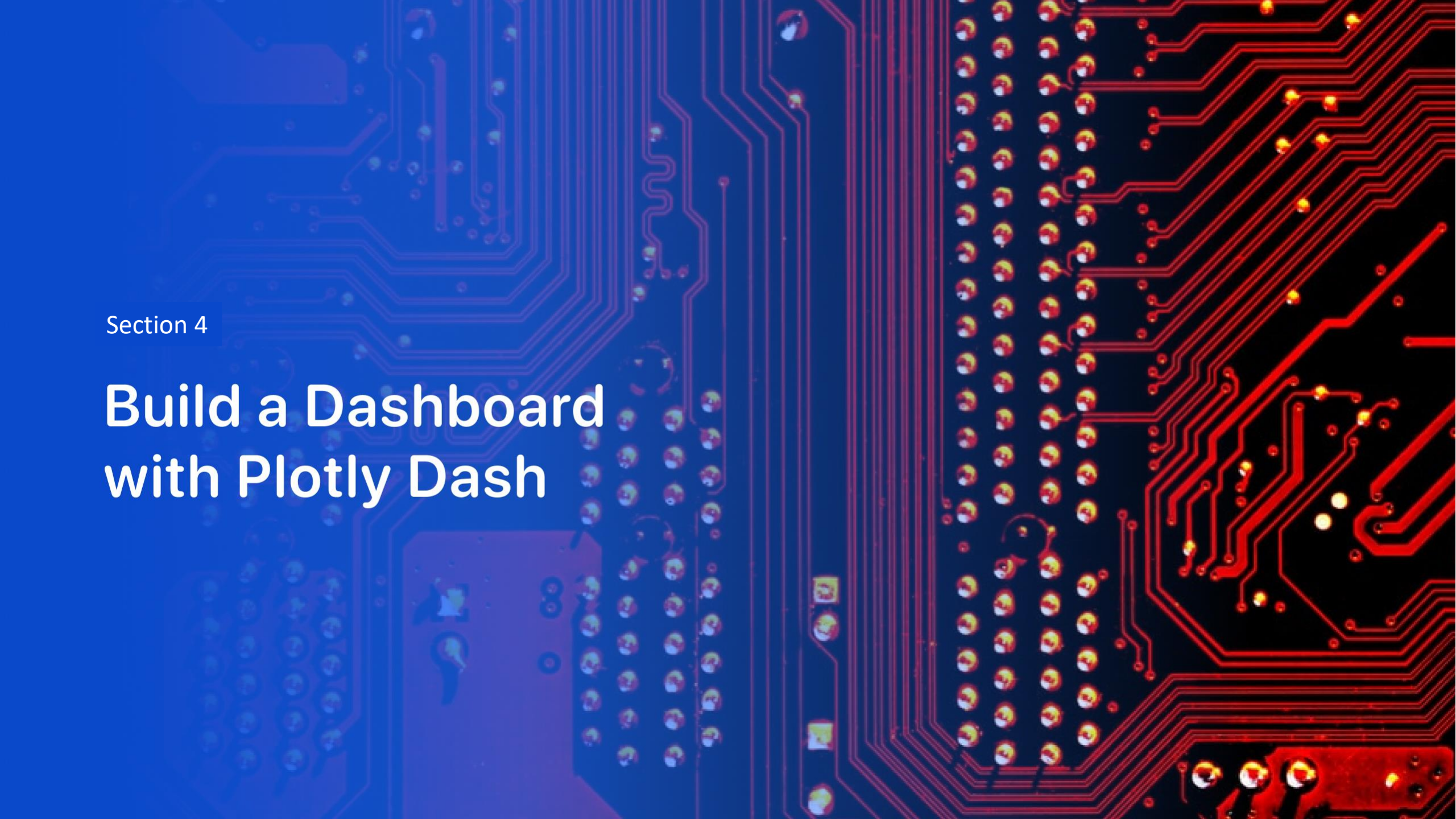
Launch Site Proximities



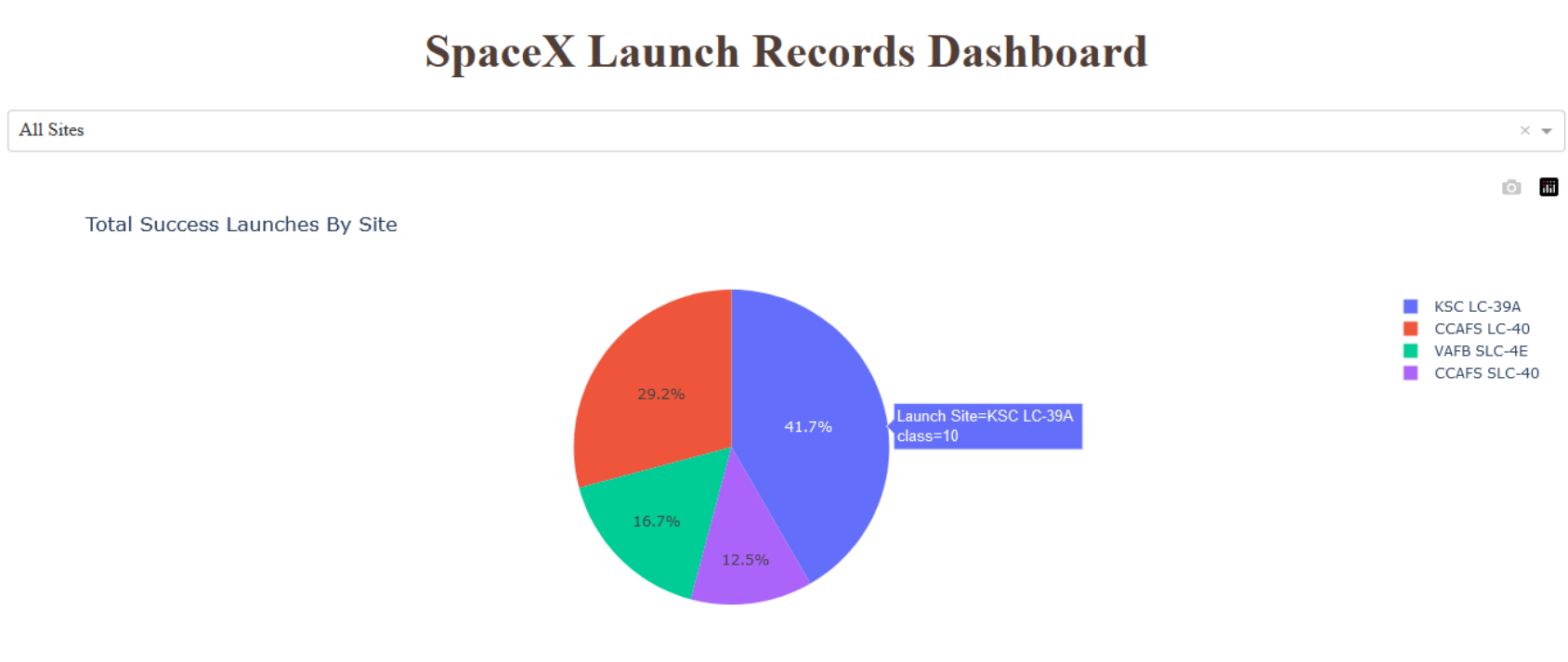


Launch Sites: CCAFS SLC-40 and CCAFS LC-40 (Florida)

The launch sites are near the coastline, highway and railway within a radius of 1.3 km and the nearest city is far more than 50 km. This may be expected due to safety concerns and ease of carriage.

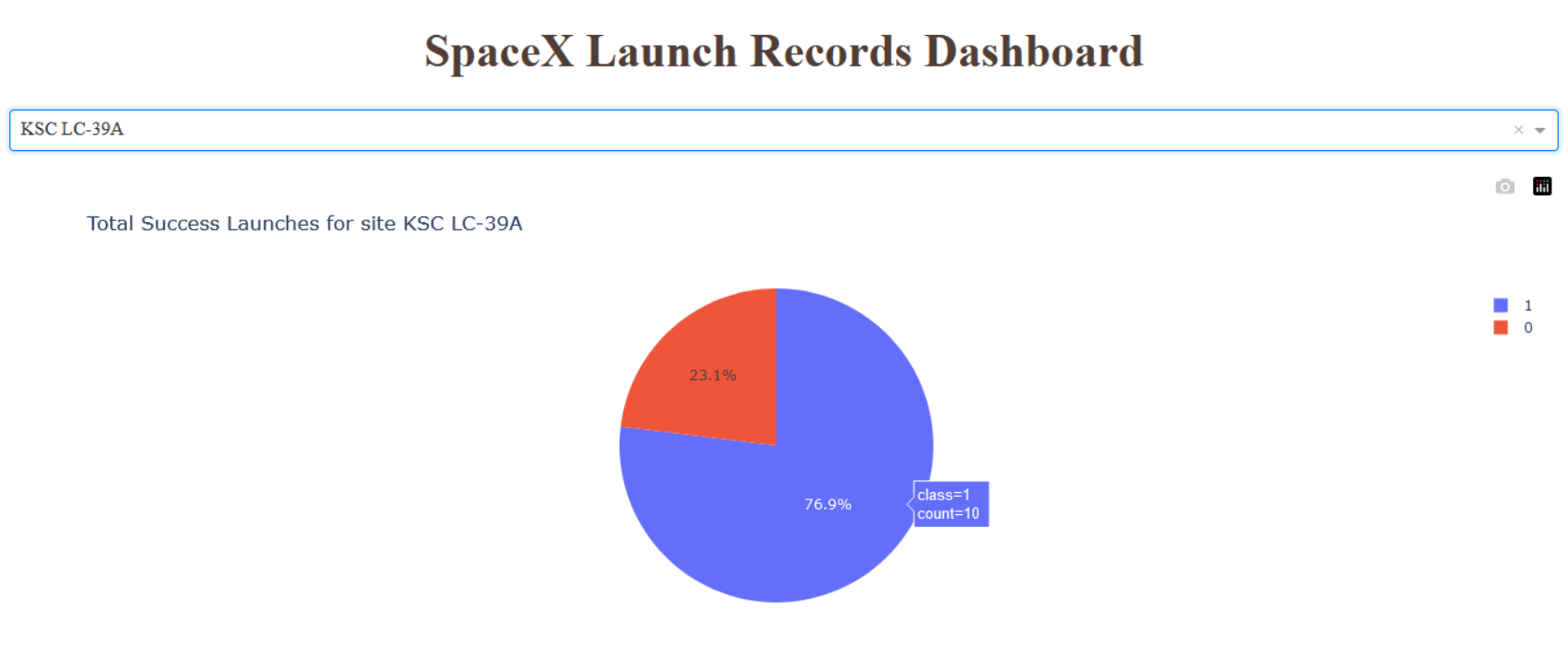

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



KSC LC-39A has the most successful launches with 10, which constitutes %41.7 of total successful launches.

Highest Launch Success Rate Among Launch Sites



KSC LC-39A has also the highest launch success rate among all launch site with 76.9% success rate (10 out of 13).

Payload vs. Launch Outcome

- After the level of 5,000 kg payload mass, there are only a few successful launches.
- For v1.0 and v1.1 boosters, all launches have a payload mass of less than 5,000 kg and there is only one successful landing out of 19 launches.

Payload vs. Launch Outcome (Cont'd)

- Having excluded boosters v1.0 and v1.1, which have exceptionally low success rates (around 5% combined), the launch success rate for payload mass of less than 5,000 kilograms is significantly higher.
- More specifically, launches with payload mass 2,000-5,000 kg are more likely to be successful when all booster versions considered.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All models have the same accuracy of 83.33% (15/18).
- Moreover, their predictions are all the same.
- But keep in mind that the accuracy of some models such as Decision Tree depends on randomness or random state chosen which is set as 2 in this project.

	Train Score	Test Score
Decision Tree	0.876786	0.833333
kNN	0.848214	0.833333
SVM	0.848214	0.833333
Logistic Reg.	0.846429	0.833333

- One possible reason why all models have the same accuracy might be the limited size of the testing dataset since the models' training scores differ.

Confusion Matrix

- The models can distinguish between the different classes.
- The models have 100% (3/3) sensitivity and 80% specificity (12/15).
- Therefore, the major problem here is false positives (upper right).

Conclusions

- The launch success rate tends to increase over time. Moreover, in 2019 and 2020, the success rate exceeded 80%.
- Recently, orbit preference shifted to Very Low Earth Orbit (VLEO), especially for heavier payloads. Also, VLEO has a success rate of over 80%.
- KSC LC-39A (Florida) is the most successful launch site, by far.
- In general, launch sites are close to the ocean, railroads, and highway for safety and transportation purposes.
- The early booster versions v1.0 and v1.1 have extremely low success rates; launches with a payload of less than 5,000 kg are more likely to be successful.
- The classification models have the same test accuracy of 83.33%. However, Decision Tree has better train accuracy. Finally, false positives are a major problem for the models.

Thank you!