IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

TO TRONG KIEN

28 Jan, 2022

Outline

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

Executive Summary

- Collected data with SpaceX REST API and SpaceX Wikipedia web page, data wrangling with API, dealing with Nulls, created labels with ‘class’ column which classifies successful landings. Explored data using SQL, visualization charts, folium maps and interactive dashboards. Transformed and normalized data to find best parameters for machine learning modelss using Grid Search CV. Visualized accuracy score of all machine learning methods.

- Four machine learning methods included Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors. All methods provide similar performances with accuracy score about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

- Background and context:
  - SpaceX advertises Falcon9 rocket launches with a cost of $62 million dollars (the best pricing).
  - Largely due to ability to recover part of rocket (Stage 1)
  - SpaceY wants to compete with SpaceX

- Problems:
  - SpaceY want to train a machine learning model and use public information to predict successful landinings if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data from SpaceX REST API and SpaceX Wikipedia web page
- Perform data wrangling
  - Dealing with Nulls, classify true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data to training set and test set, transform and normalize data with one-hot encoding
  - Tuning model performances using GridSearchCV and evaluate using accuracy score

Data Collection

- Data collection process involved a combination of API requests from SpaceX API and Web scraping data from a table in SpaceX Wikipedia entry.

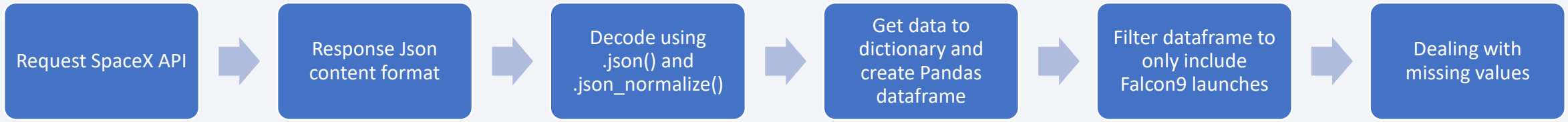
- **SpaceX API data columns include:** *FlightNumber, Data, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*

- **Wikipedia Data Columns include:** *Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Lauch outcome, Version Booster, Booster Landing, Date, Time*

Data Collection – SpaceX API

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%201:%20Data%20Collection%20with%20Web%20Scraping/W1_HOL_Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

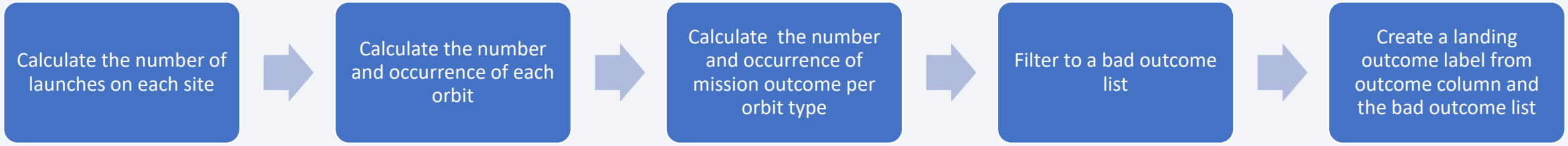
- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%201:%20Data%20Collection%20with%20Web%20Scraping/W1_Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%201:%20Data%20Collection%20with%20Web%20Scraping/W1_HOL_Data%20Wrangling.ipynb



EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%202%20:%20Exploratory%20Data%20Analysis%20Overview/W2_HOL_The%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- Loaded dataset into IBM DB2 Database
- Queried using SQL Python intergration
- Queries were made to get a better understanding of the dataset
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions and landing outcomes

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%202%20:%20Exploratory%20Data%20Analysis%20Overview/W2_HOL_The%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Coastline, Railway, Highway, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%203:%20Interactive%20Visual%20Analytics%20and%20Dashboards/W3_HOL_The%20Data%20Visualization%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

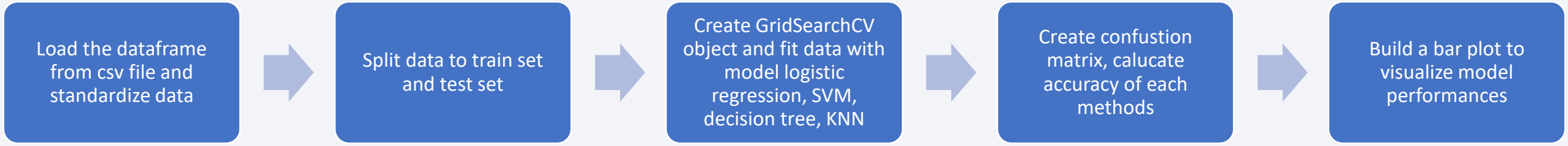
- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%203:%20Interactive%20Visual%20Analytics%20and%20Dashboards/W3_HOL_spacex_dash_app.ipynb

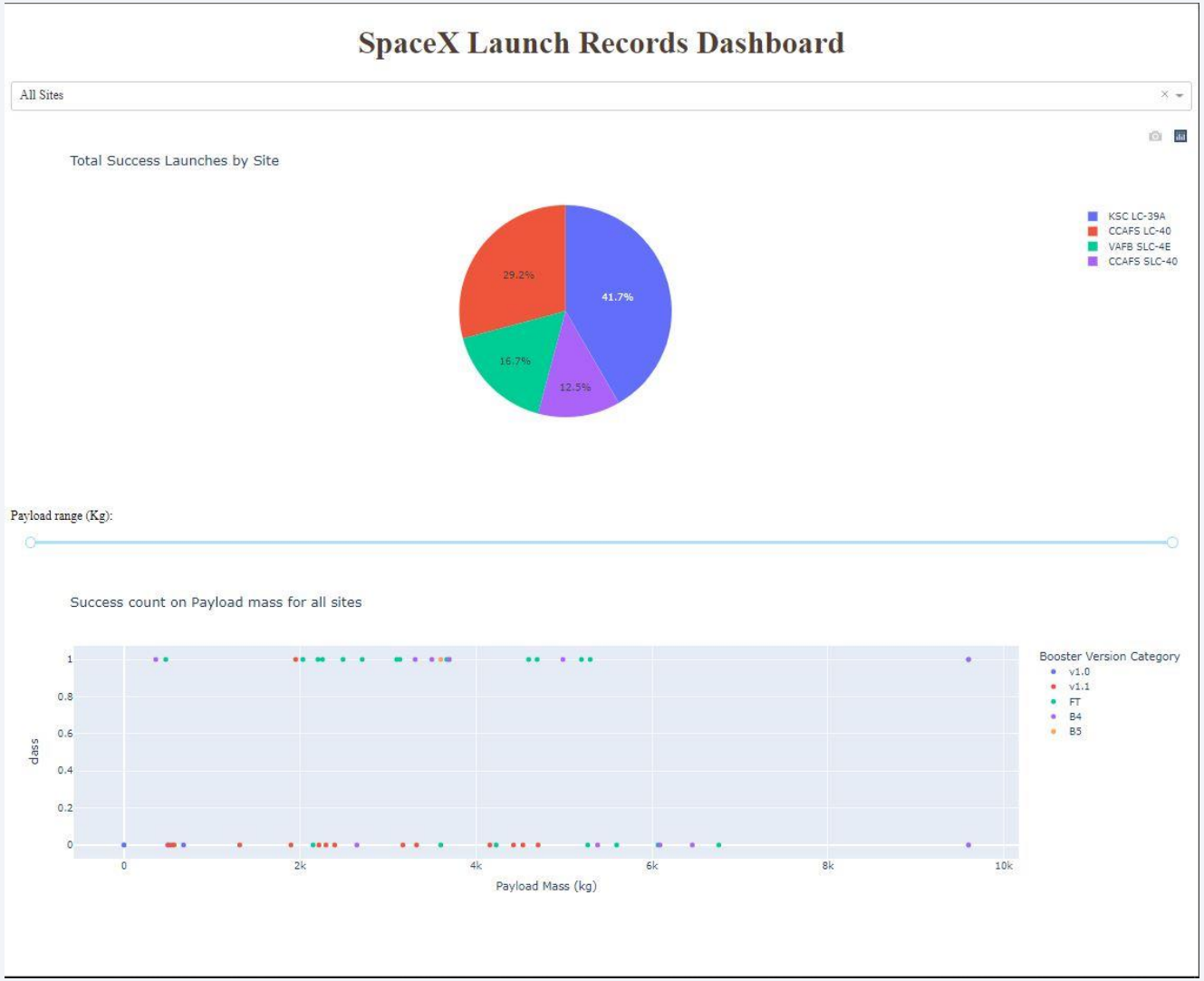
Predictive Analysis (Classification)

- GitHub URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/blob/da5994509d4eed37ba3948f3f717749755246d8e/Week%204:%20Predictive%20Analysis%20Overview/W4_HOL_the%20Machine%20Learning%20Predictive%20Analysis.ipynb



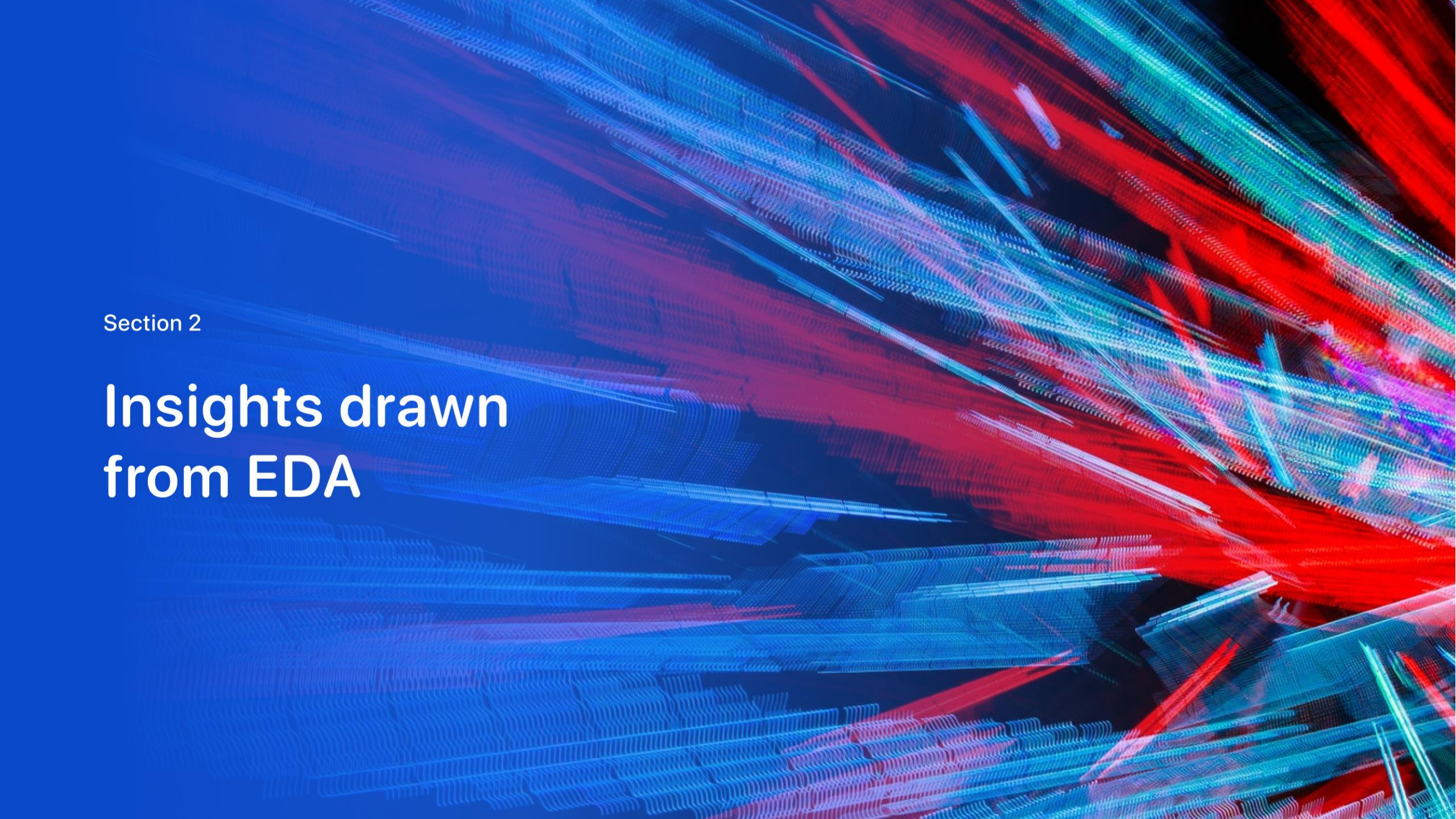
Results



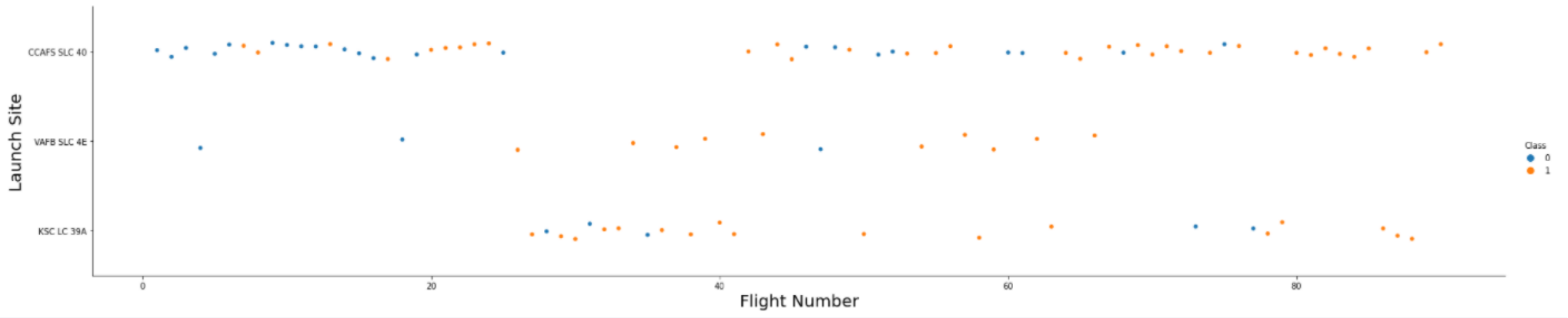
This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

Section 2

Insights drawn from EDA

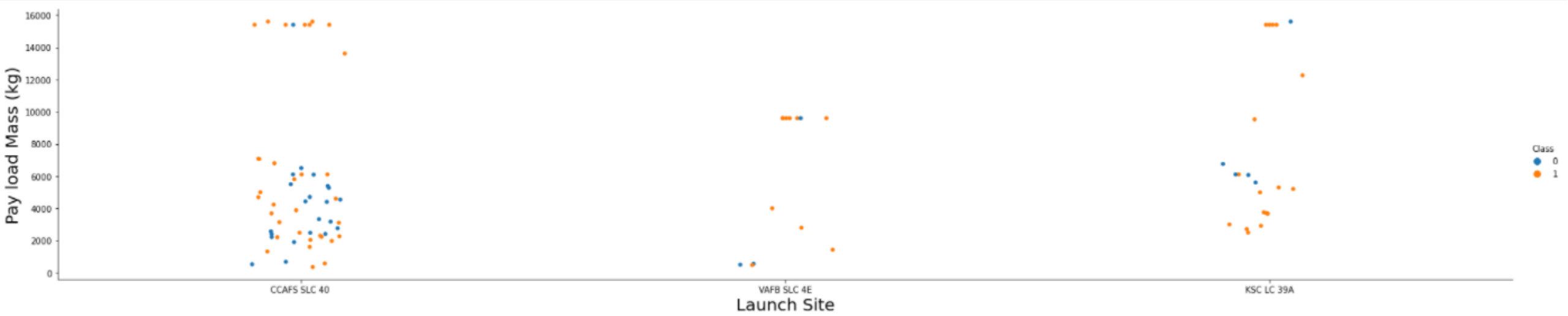


Flight Number vs. Launch Site



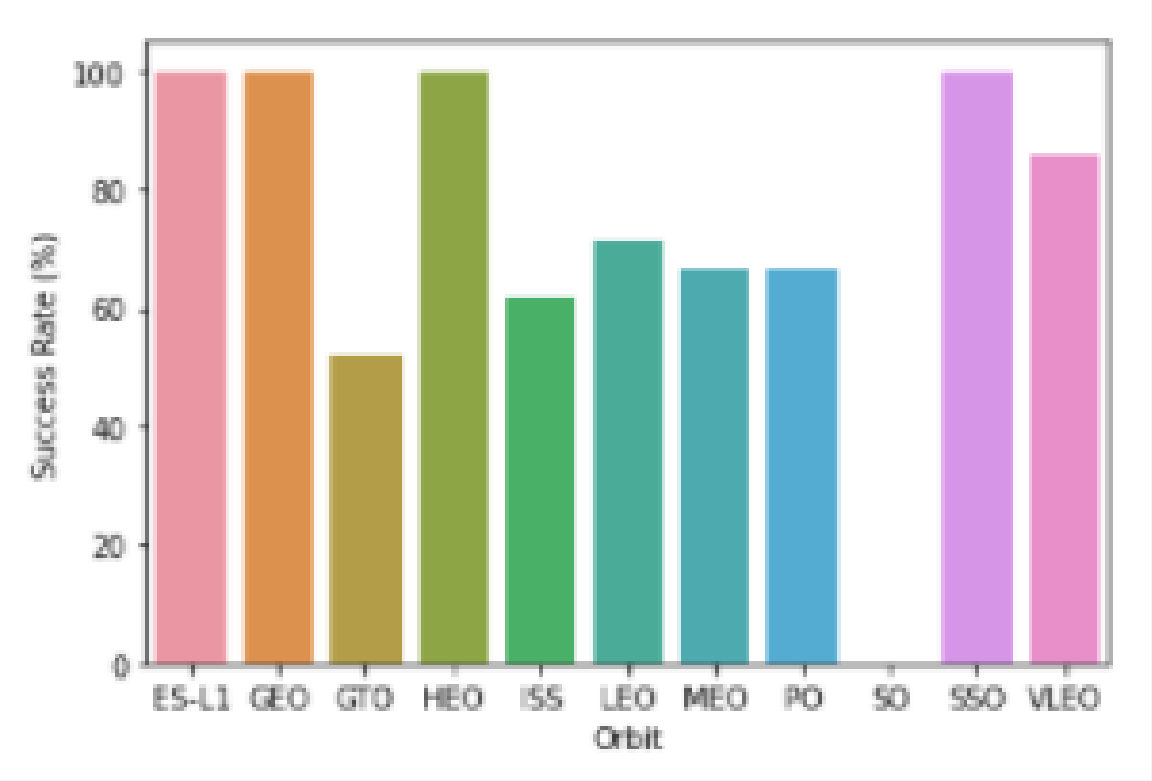
- Orange indicates successful launchs, Blue indicates unsuccessful launchs
- Success rate increase over time (indicated in Flight Number).
- CCAFS had the most volume of flight number

Payload vs. Launch Site



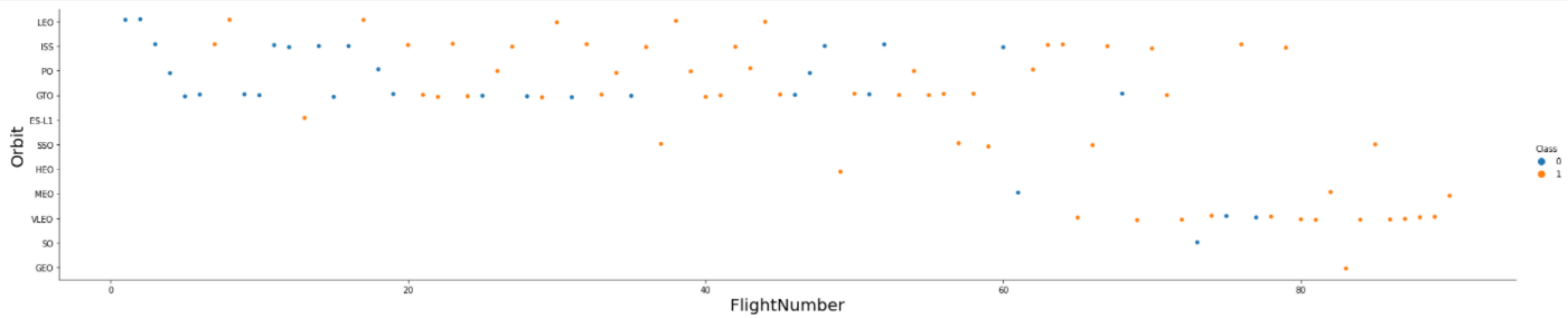
- Orange indicates successful launchs, Blue indicates unsuccessful launchs
- Payload Mass is between 0 – 16000 kg
- Uncessfull launches is mostly between 0-6000 kg

Success Rate vs. Orbit Type



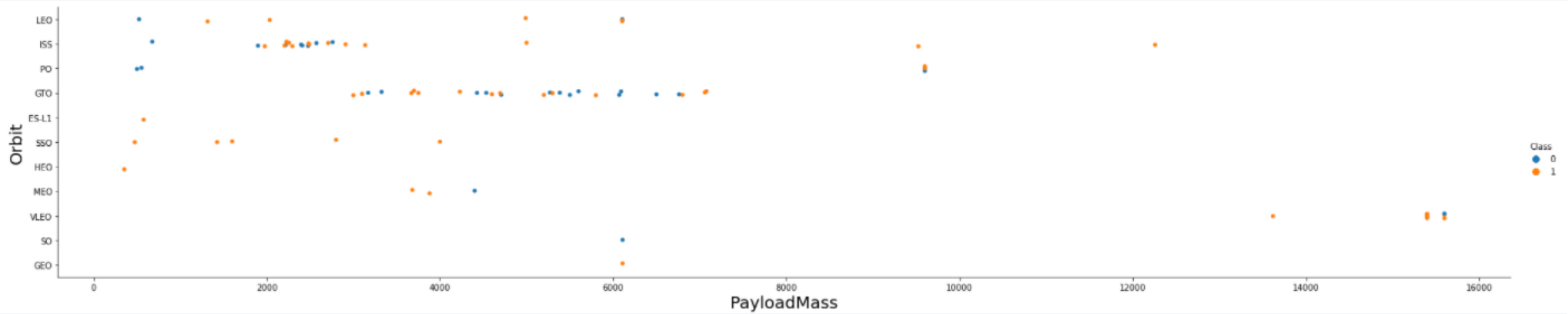
- ES-L1, GEO, HEO, SSO have 100% success rate
- VEO has around 90% success rate
- LEO has around 70% success rate
- MEO and PO have the same around 65% success rate
- ISS has 60% success rate
- GTO has around 55% success rate
- SO have no information about success rate

Flight Number vs. Orbit Type



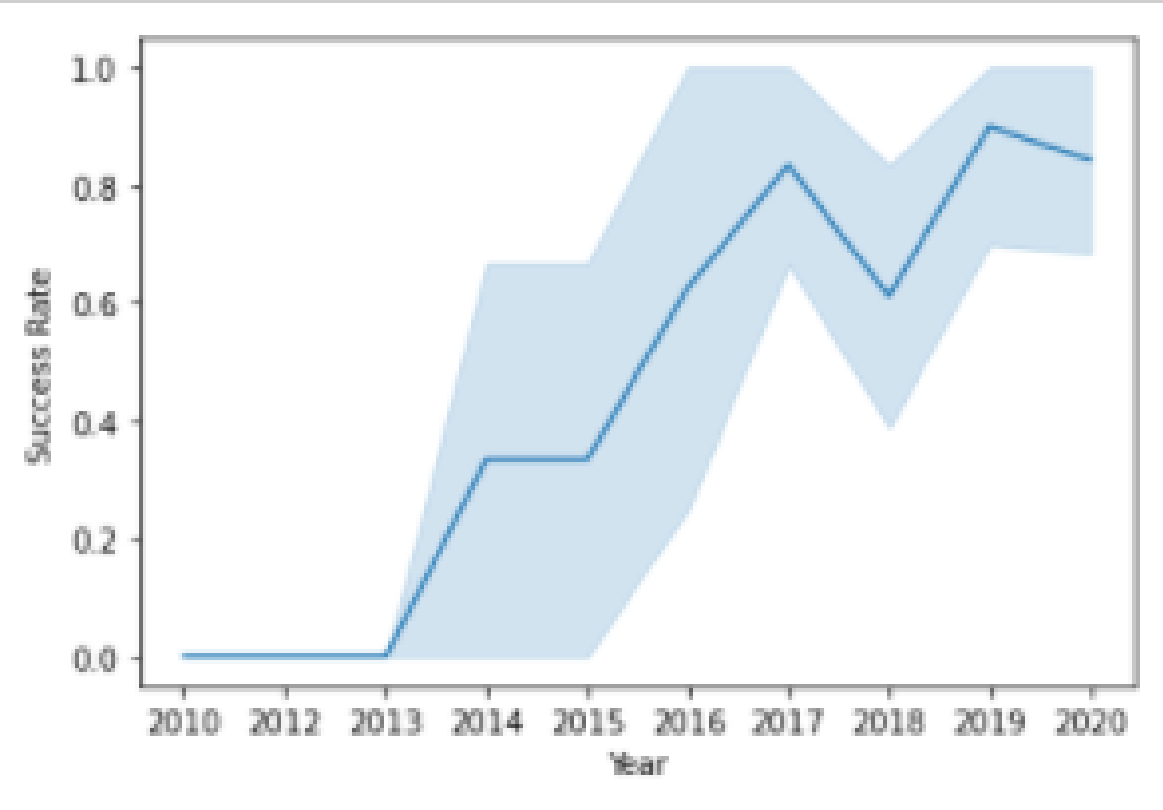
- Orange indicates successful launchs, Blue indicates unsuccessful launchs
- Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



- Orange indicates successful launchs, Blue indicates unsuccessful launchs
- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



95% confidence interval (light blue shading)

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%

All Launch Site Names

```
In [8]: %%sql

        select distinct launch_site
        from SPACEXTBL

         * ibm_db_sa://gkn98086:***@125f9f6
        Done.
```

Out[8]:

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

Query unique launch site names from database.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

In [9]:
```
%%sql

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

```
 * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[9]:

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

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

```
In [8]: %%sql

        select sum(payload_mass__kg_) as total_payload_mass
        from SPACEXTBL
        where customer = 'NASA (CRS)'

         * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-939
        Done.

Out[8]: total_payload_mass
                    45596
```

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
In [9]: %%sql

        select avg(payload_mass__kg_) as total_payload_mass
        from SPACEXTBL
        where booster_version = 'F9 v1.1'

         * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-939
        Done.

Out[9]: total_payload_mass
                    2928
```

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Landing Date

```
In [16]: %%sql

         select min(DATE) as DATE
         from SPACEXTBL
         where landing__outcome = 'Success (ground pad)'

          * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9
         Done.

Out[16]:
```

DATE
2015-12-22

This query returns the first successful ground pad landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [11]:
```
%%sql

select distinct booster_version
from SPACEXTBL
where landing__outcome = 'Success (drone ship)'
    and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

```
 * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-9399-c8177b21803
Done.
```

Out[11]:

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

In [18]:
```
%%sql

select
    mission_outcome,
    count(*) as total_number
from SPACEXTBL
group by mission_outcome
```

```
 * ibm_db_sa://gkn98086:***@125f9f61-9715-
Done.
```

Out[18]:

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

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload

```
In [13]: %%sql

         select distinct booster_version
         from SPACEXTBL
         where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)

          * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tg
         Done.
```

Out[13]:

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

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

2015 Launch Records

In [14]:
```
%%sql

select booster_version, launch_site
from SPACEXTBL
where landing__outcome = 'Failure (drone ship)'
    and YEAR(DATE) = 2015
limit 100
```

```
 * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-
Done.
```

Out[14]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

This query returns Booster Version and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

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

```
In [15]: %%sql

         select landing__outcome, count(*) as rank
         from SPACEXTBL
         where DATE between '2010-06-04' and '2017-03-20'
             and (landing__outcome = 'Failure (drone ship)'
                  or landing__outcome = 'Success (ground pad)')
         group by landing__outcome
         order by rank desc

          * ibm_db_sa://gkn98086:***@125f9f61-9715-46f9-9399-c817
         Done.
```

Out[15]:

landing__outcome	RANK
Failure (drone ship)	5
Success (ground pad)	3

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

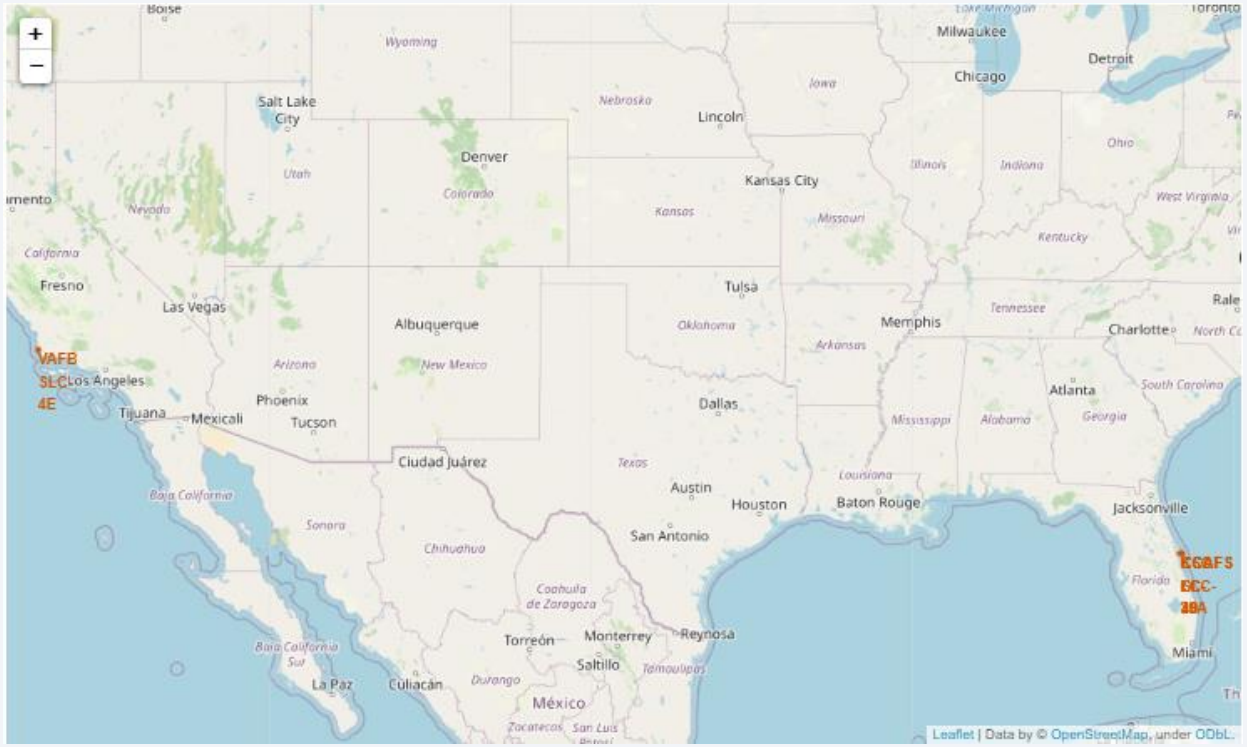
There are two types of successful landing outcomes: drone ship and ground pad landings.

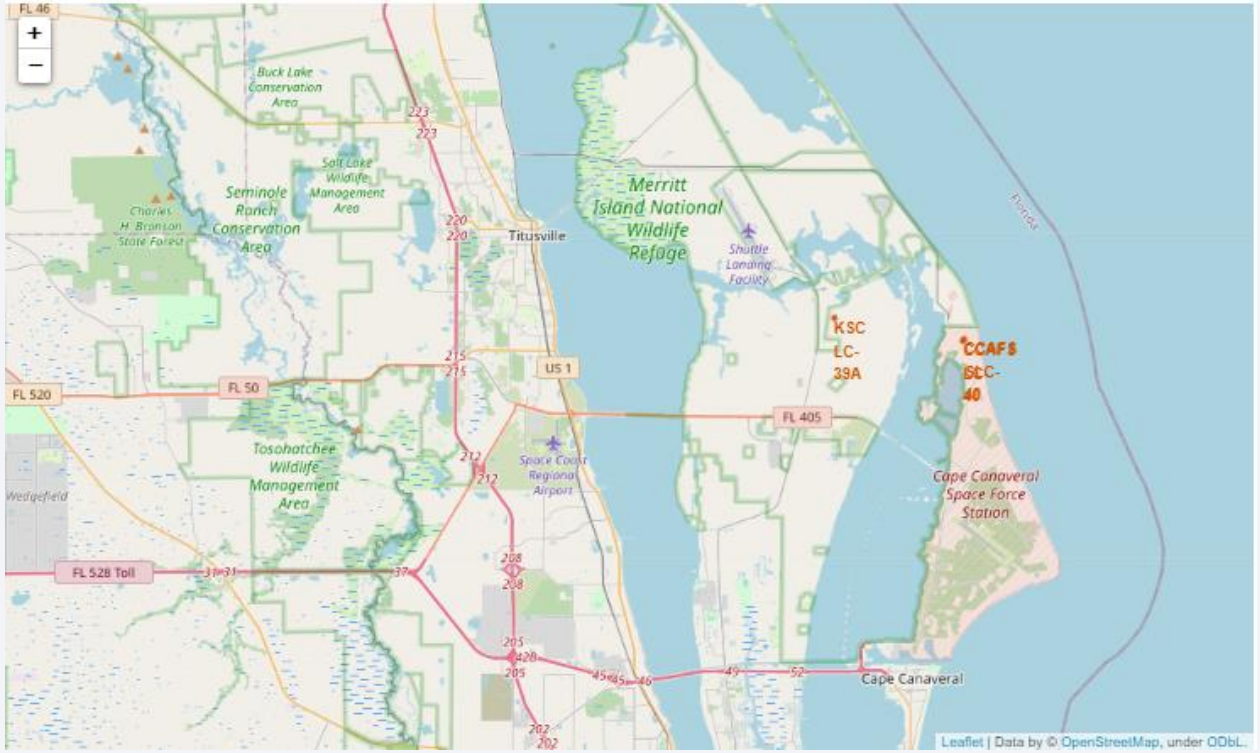
There were 8 successful landings in total during this time period

Section 4

Launch Sites Proximities Analysis

Launch Site Locations

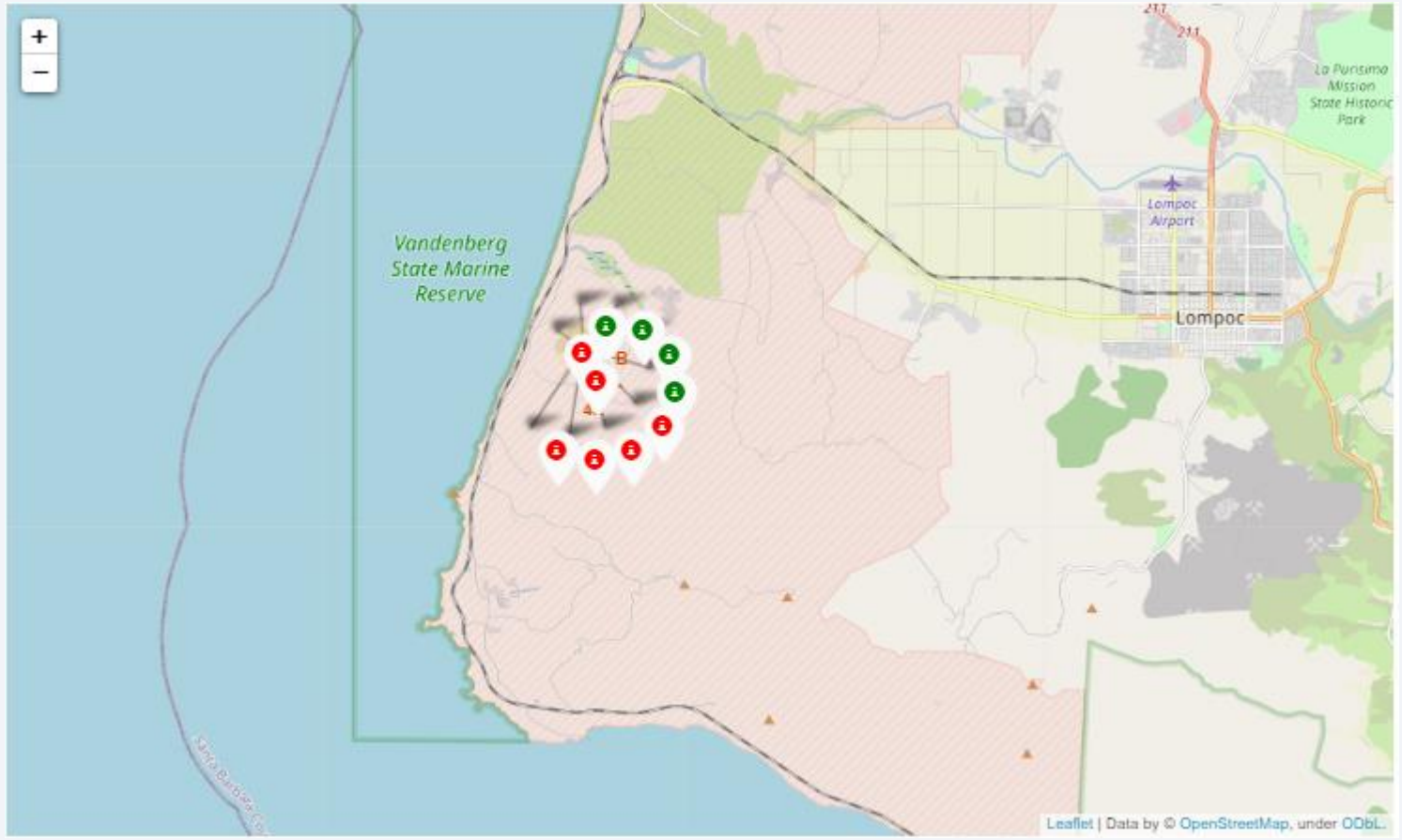




The left map shows all launch sites relative US map.

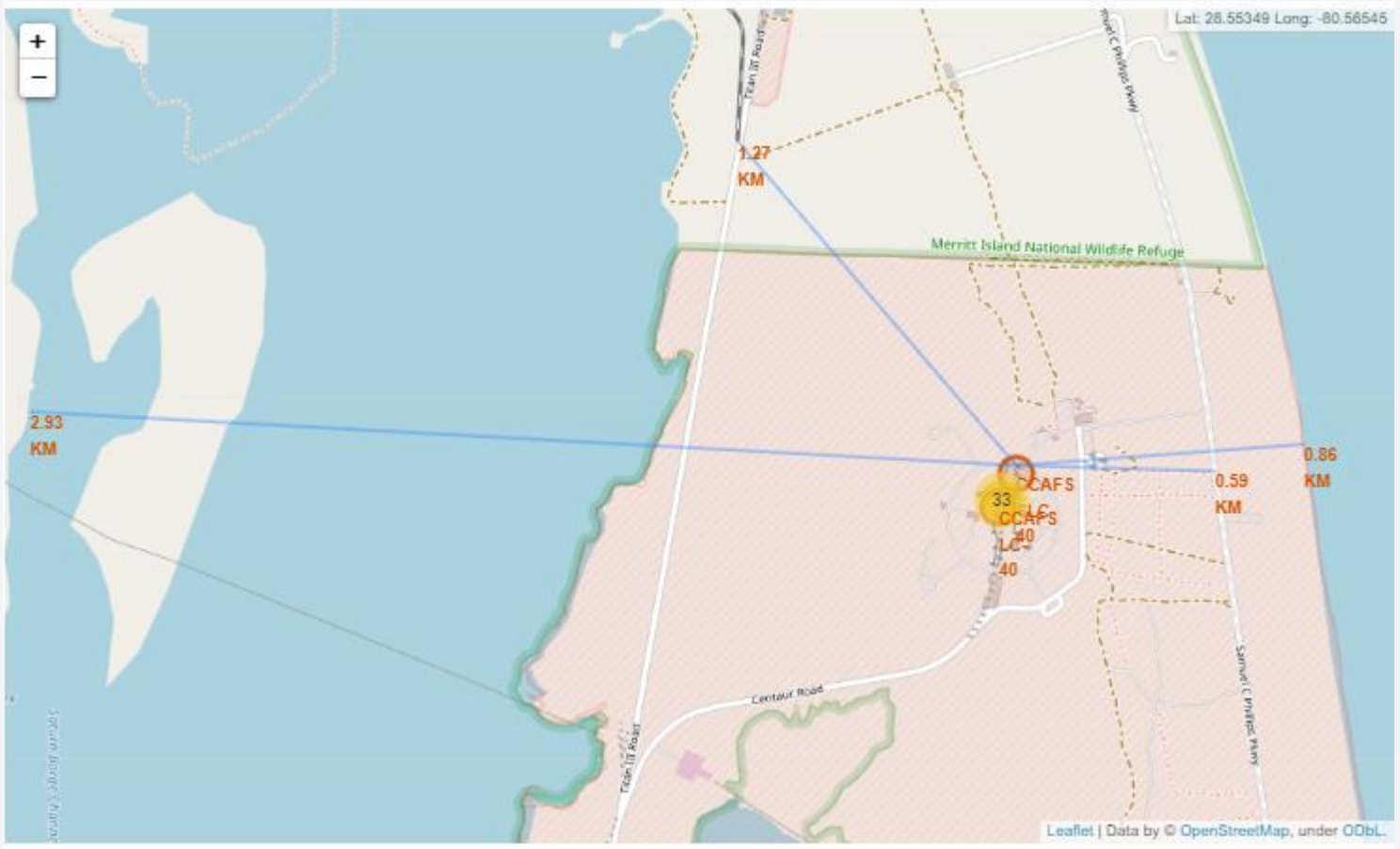
The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

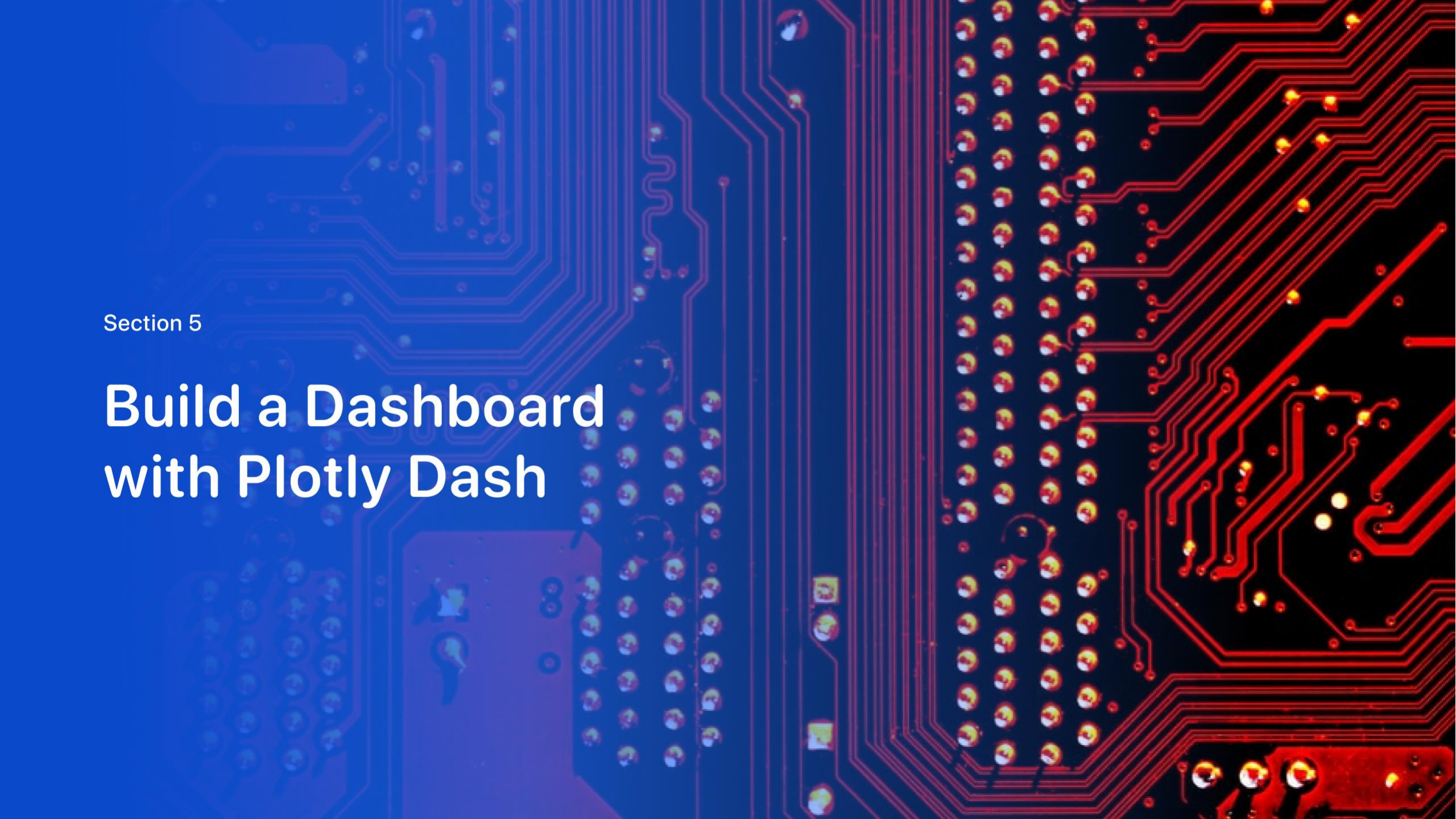
Key Location Proximities



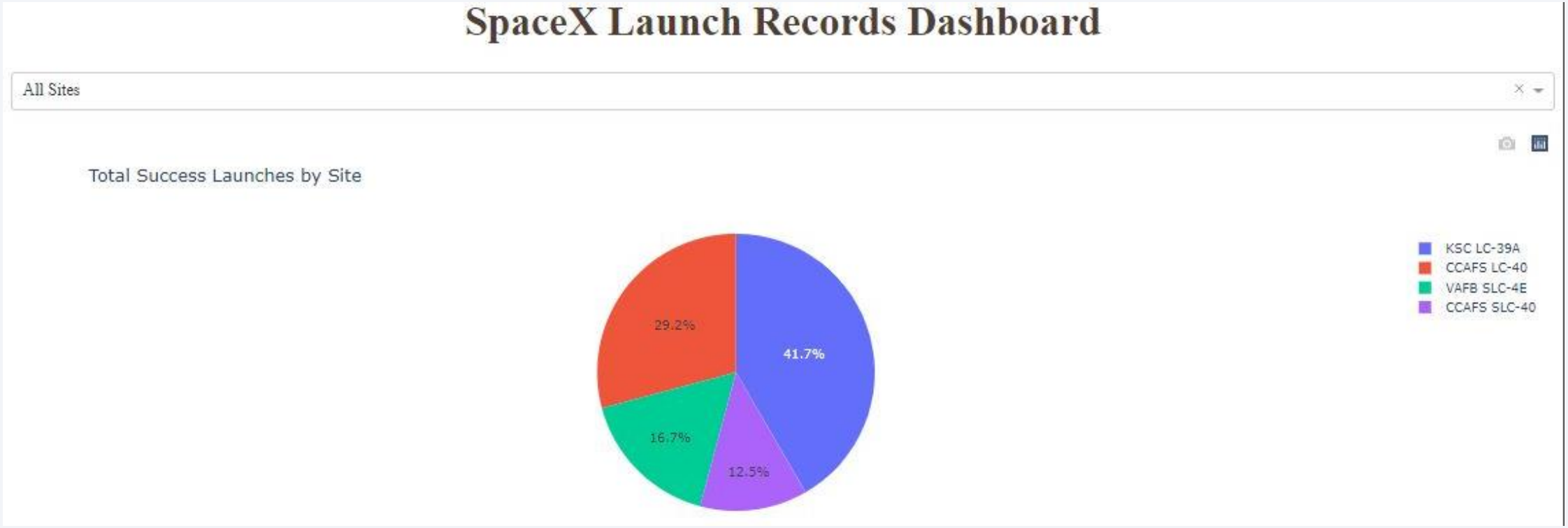
Choose a launch site and calculate distance between it and the closest railway, highway, coastline and city

Section 5

Build a Dashboard with Plotly Dash

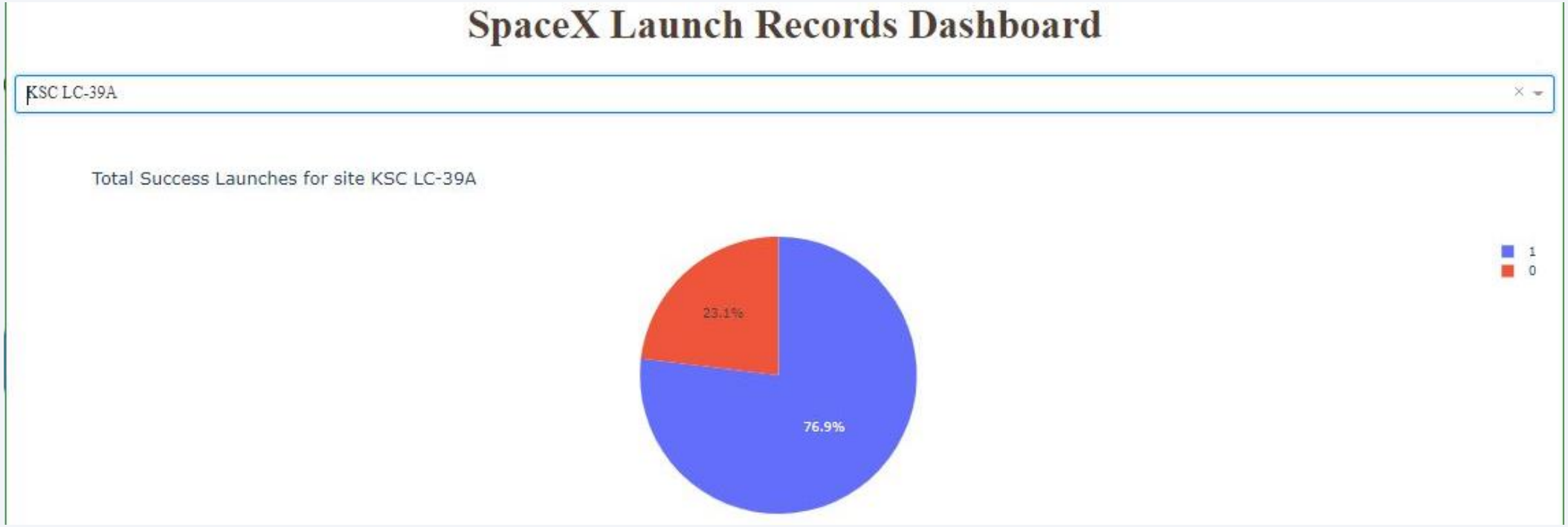


Total Success Launches by Site



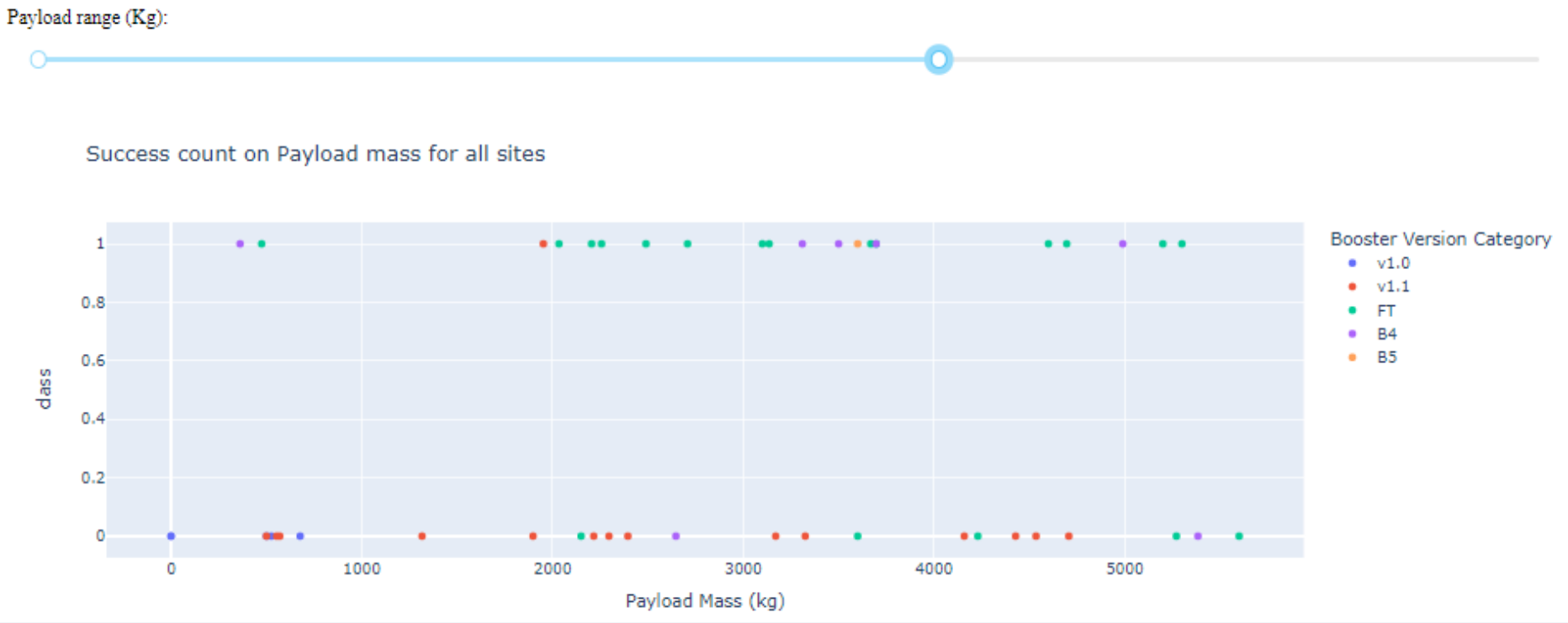
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings.

Total Success Launches for site KSC LC-39A



KSC LC-39A has the highest success rate with 10 successful landings (76.9%) and 3 failed landings (23.1%).

Payload Mass vs. Landing Outcome vs.Booster Version Category

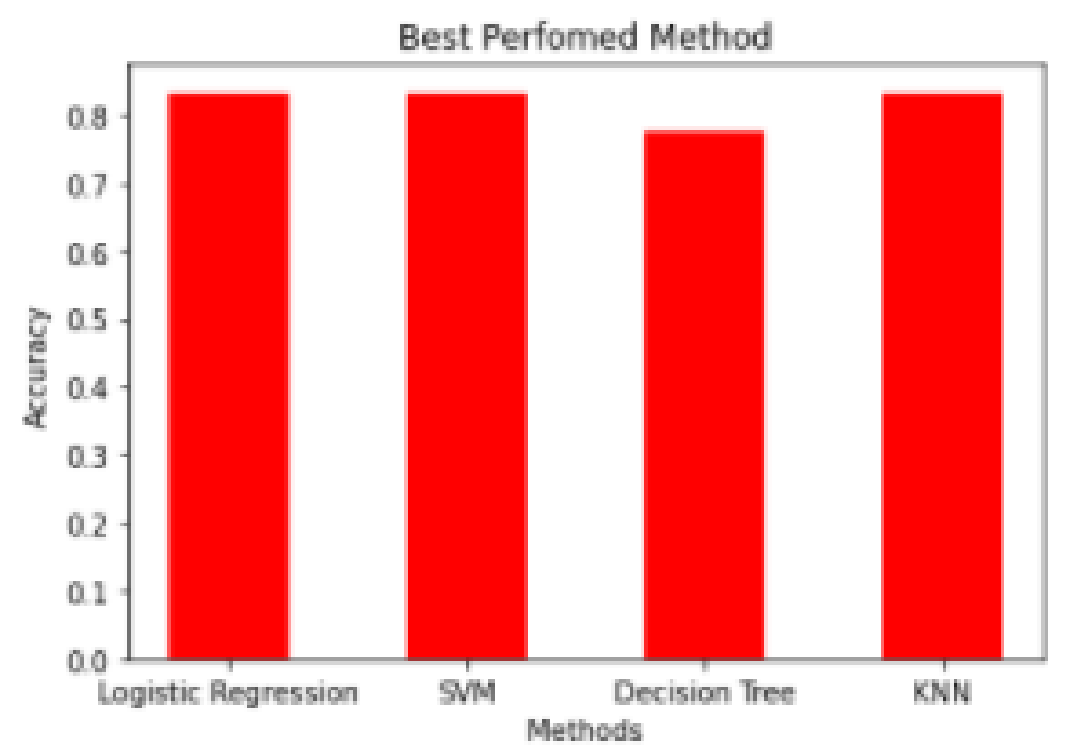


Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

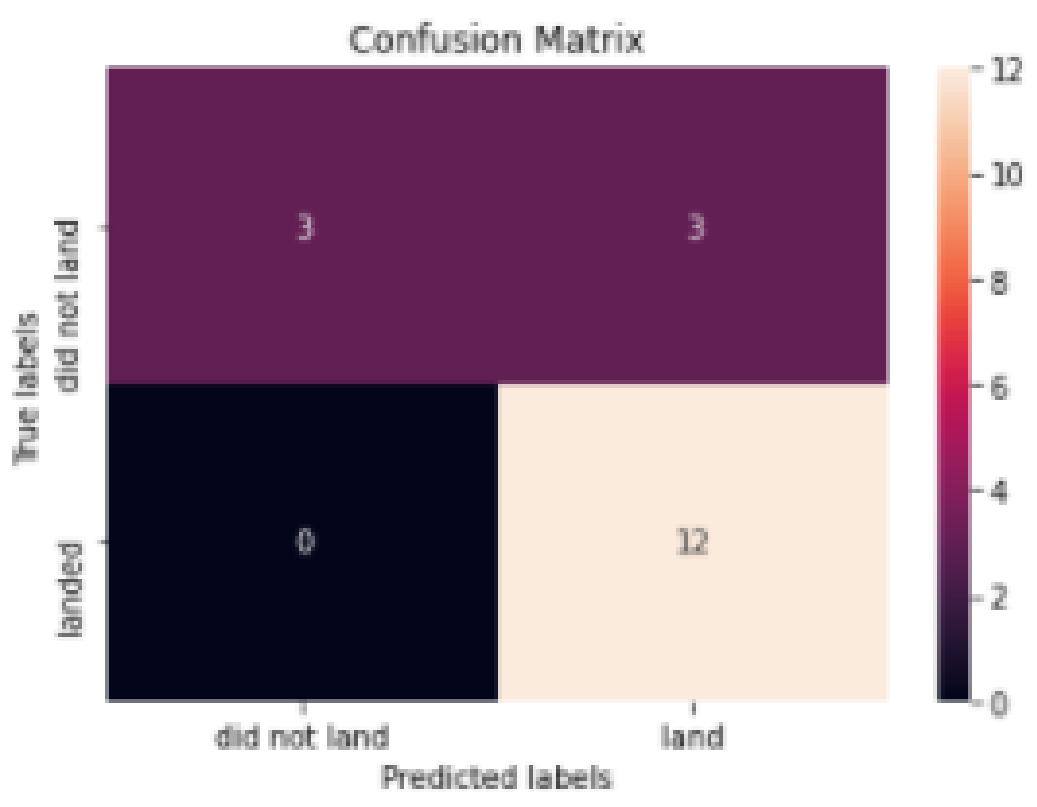


LogisticRegression, SV and KNN models had virtually the same accuracy on the test set at 83.33% accuracy.

It should be noted that test size is small at only sample size of 18.

We likely need more data to determine the best model.

Confusion Matrix



Since all models performed the same for the test set, the confusion matrix is the same across all models.

The models predicted 12 successful landings when the true label was successful landing (true positives).

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing (true negatives).

The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

- GitHub Repository URL:

https://github.com/tokien1998/IBM-DS-Applied-Data-Science-Capstone/tree/master

Thank you!