IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Michael Asrat

08/01/2022

Outline

- Data Collection through API and WebScraping
- Data Wrangling
- Exploratory Data Analysis with SQL and with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Executive Summary

SpaceX’s Falcon 9 v1.0, was designed as an expandable launch system and attempts to recover it, adding lightweight thermal protection system capability and using a parachute recovery, failed. The v1.1 was designed to allow powered re-entry and one first stage was recovered, but real breakthrough came with the Full Thrust version in 2015. Data shows that a first stage has been reused up to 6 times, and the yearly success rate has achieved 75%~80%.

Building a model, on top of the payload mass and orbit, features of the rocket including its number of flights/re-use count, block version, landing pad coordinates, presence of gridfins and legs (all absent in the earliest version of Falcon 9), help yield good predictions of the chance of success of a landing, with an accuracy of 90%.

Introduction

- Project background and context

SpaceX's Falcon9 rocket launch costs 62 million dollars. Similar rocket launches from other providers costs 165 million dollars. These huge difference in cost comes as SpaceX was able to recover (reuse) part of the launched rocket (Stage 1) by successfully landing it back.

- This study will address the following questions:

SpaceY wants to predict the successful landing of it's rocket by developing a ML model.

- What segment of the market does SpaceX address, in terms of payload mass and orbits?
- Which launch sites do they rely on? What is their launch frequency?
- Is there an observable learning curve? Which parameters can we play on to make the learning curve steeper?

Section 1

Methodology

Methodology

Summary

- Data collection methodology:
  - Combined data from SpaceX API and SpaceX Wikipedia page
- Perform data wrangling
  - Applying categorical features: true landing as successful, otherwise 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
  - Tuned models using GridSearchCV

Data Collection

Space X API Data Columns:

Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

Wiki Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

Request (SpaceX APIs)

JSON file + Lists(Launch Site, Booster Version, Payload Data)

Normalize the data from JSON format

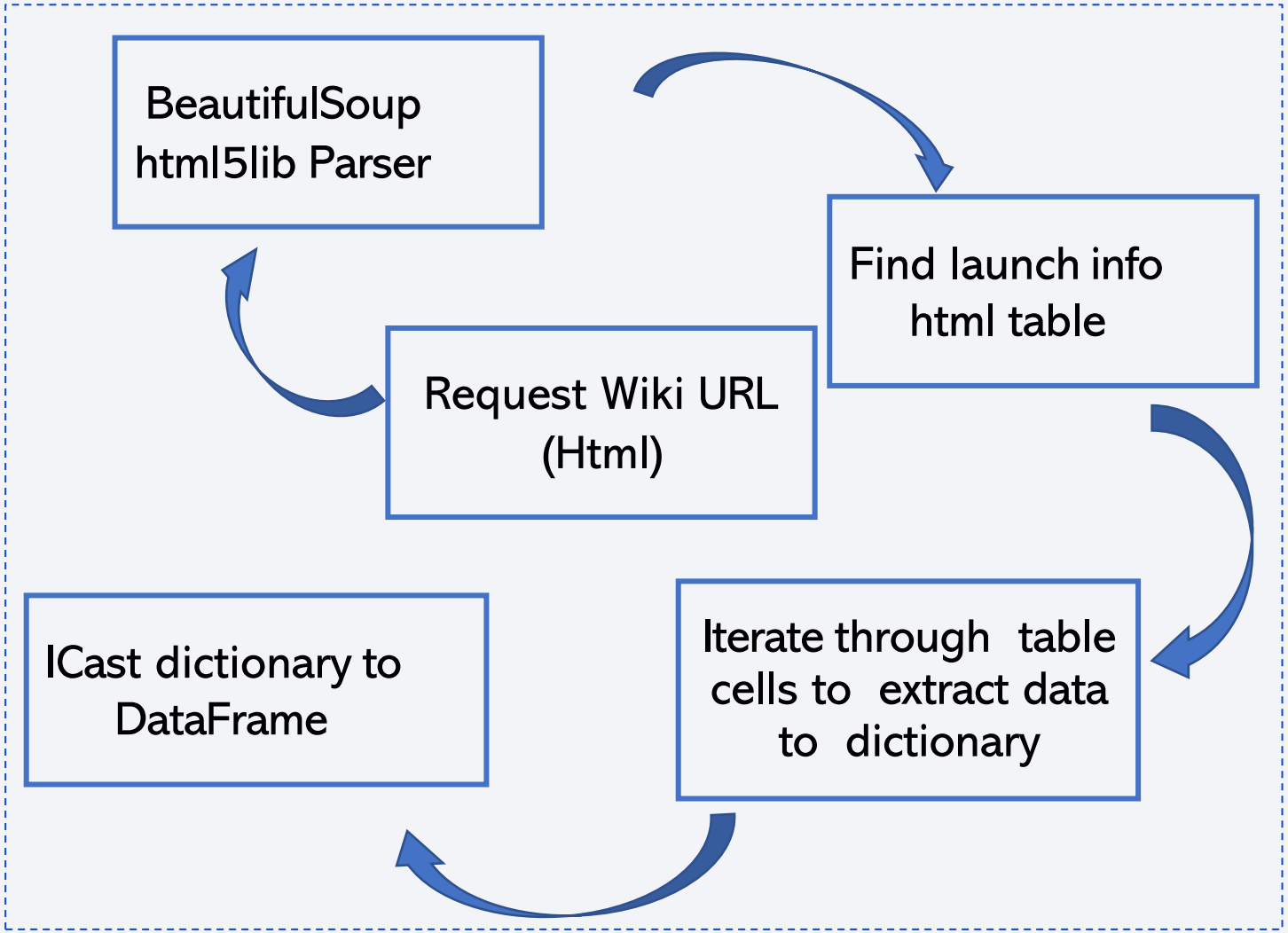
Filter data to only include the Falcon 9 launches

Impute missing PayloadMass values with mean

https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Collecting%20the%20Data.ipynb

Data Collection - Scraping

- Web scrapping applied on Falcon 9 launch records with BeautifulSoup
- The table was parsed and converted into a pandas dataframe.



- https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Web%20Scraping%20Falcon%209.ipynb

Data Wrangling

1. Load the data set

Load Space X dataset, from last section.

```
df=pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv")
df.head(10)
```

3. Landing Outcome

```
# Landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

5. Determine the success

We can use the following line of code to determine the success rate:

```
df["Class"].mean()
```

```
0.6666666666666666
```

2. Calculate the number of launches on each site

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

4. Find the bad outcome

We create a set of outcomes where the second stage did not land successfully:

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

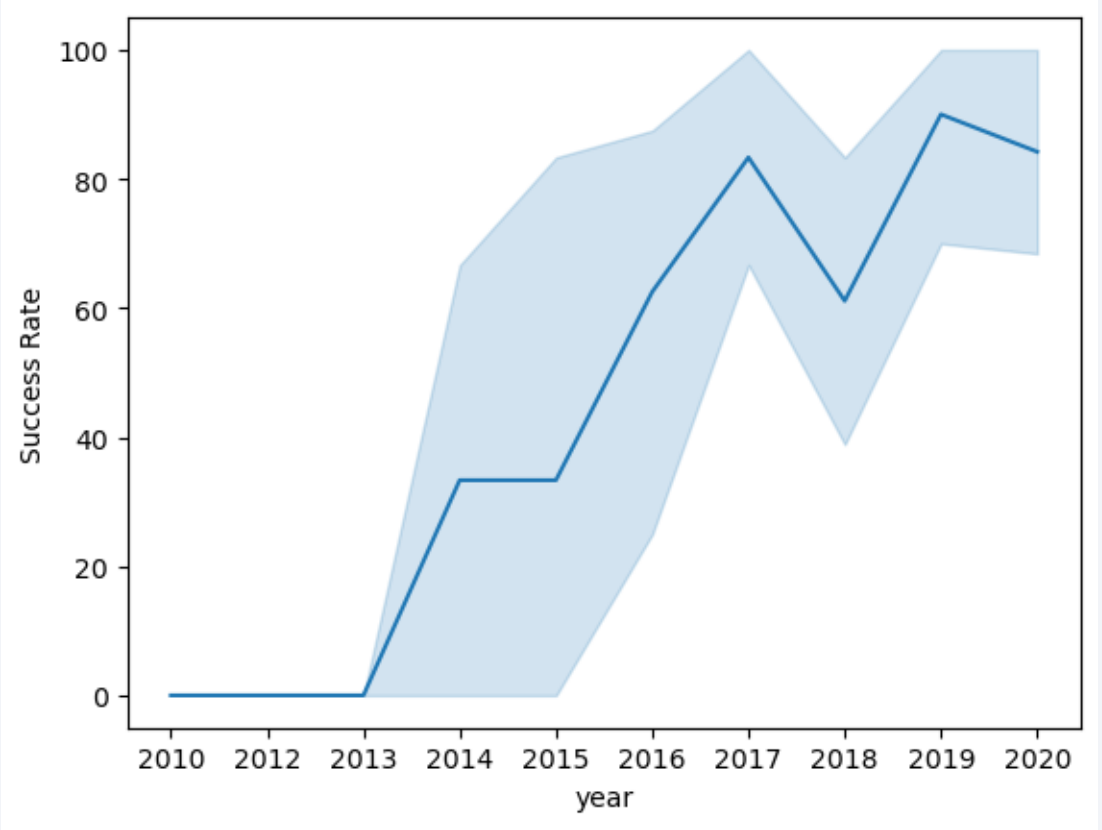
https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling%20Falcon%209.ipynb

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



https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Queried using SQL Python integration using PostgreSQL database.
- Queried information about:
  - The names of unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The total payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

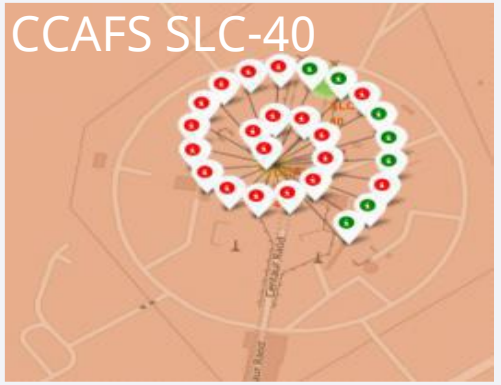
SpaceX launches rockets from 4 sites

- East coast: CAFS SLC-40, CCAFS SLC-40, KSC LC-39A
- West coast: VAFB SLC-4E

Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.

All 4 sites share the same features:

- Proximity to coastline
- Distance from cities, highways and railways



VAFB SLC-4E

KSC LC-39A

Success/failed launches for each site

https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart showing the total launches by a certain sites and used to visualize launch site success rate.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Plot%20Dash

Predictive Analysis (Classification)

1. Building the model

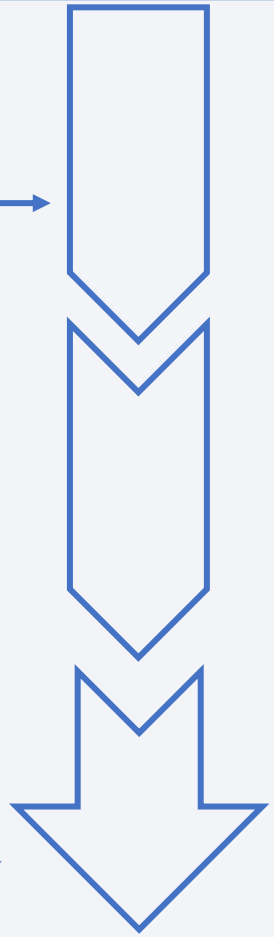
- Create column for the class
- Standardize the data
- Split the data into train and test sets
- Build GridSearch CV model and fit the data

2. Evaluating the model

- Calculating the accuracies
- Calculating the confusion matrix
- Plot the results

3. Finding the optimal model

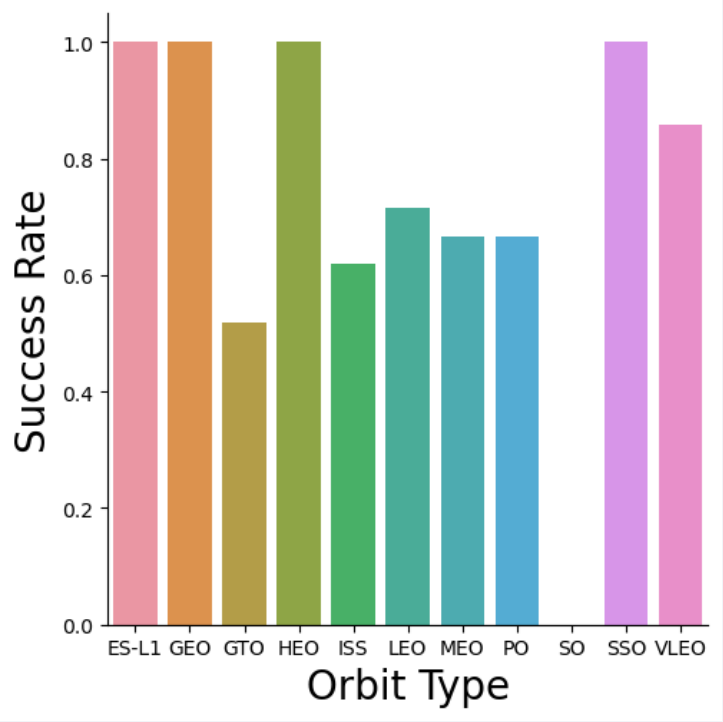
- Find the best hyperparameters for the models
- Find the best model with highest accuracy
- Confirm the optimal model



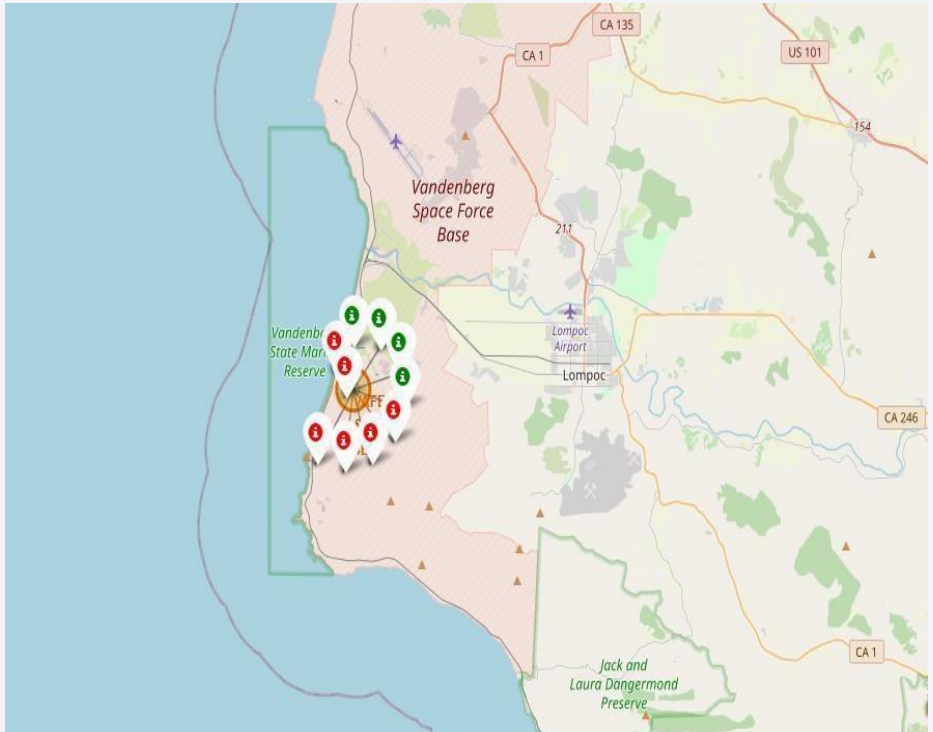
https://github.com/t1-michael/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction%20(1).ipynb

Results

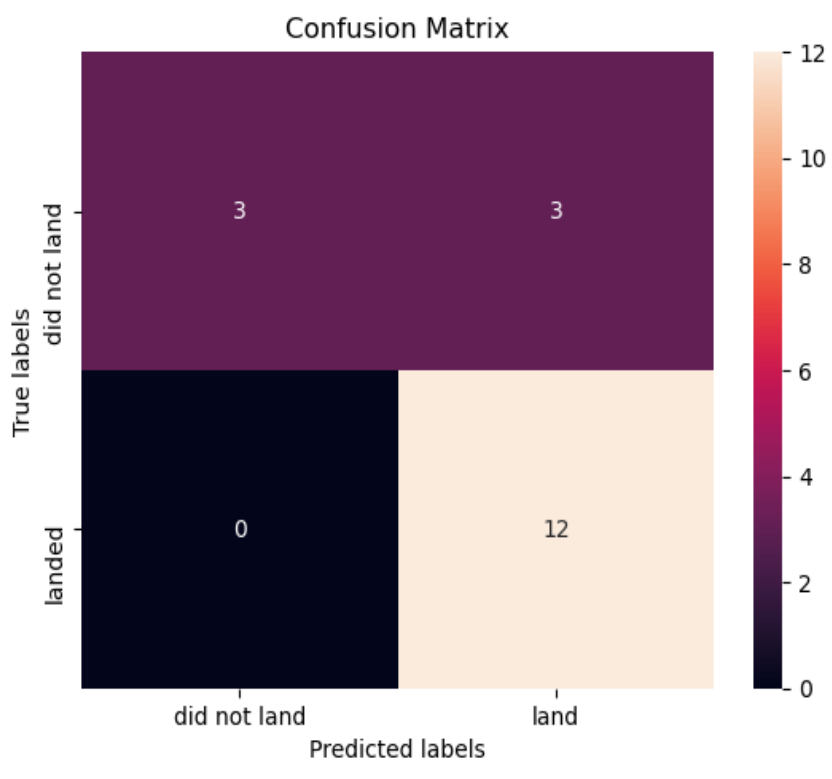
Exploratory data analysis results

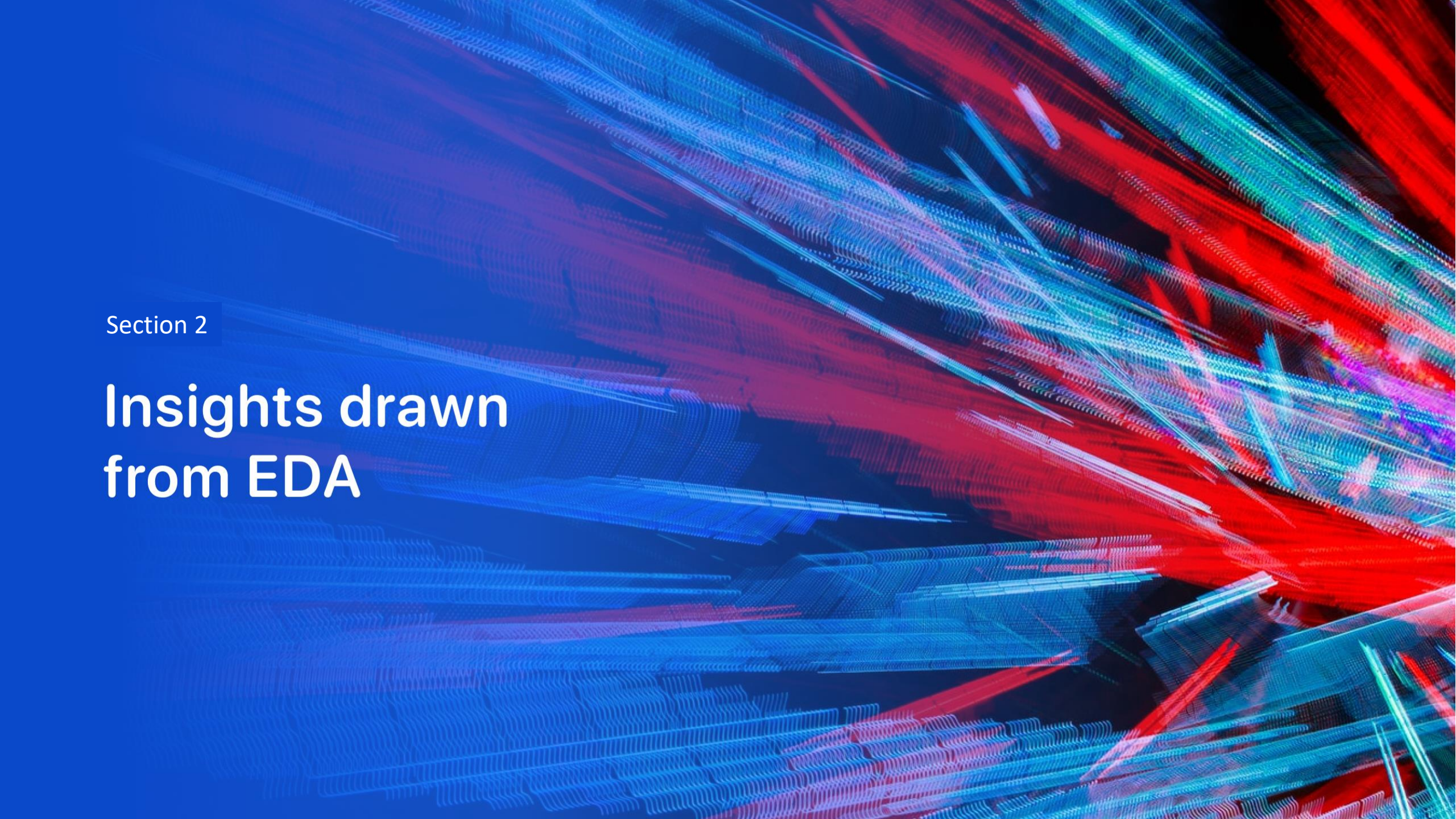


Interactive analytics demo in screenshots

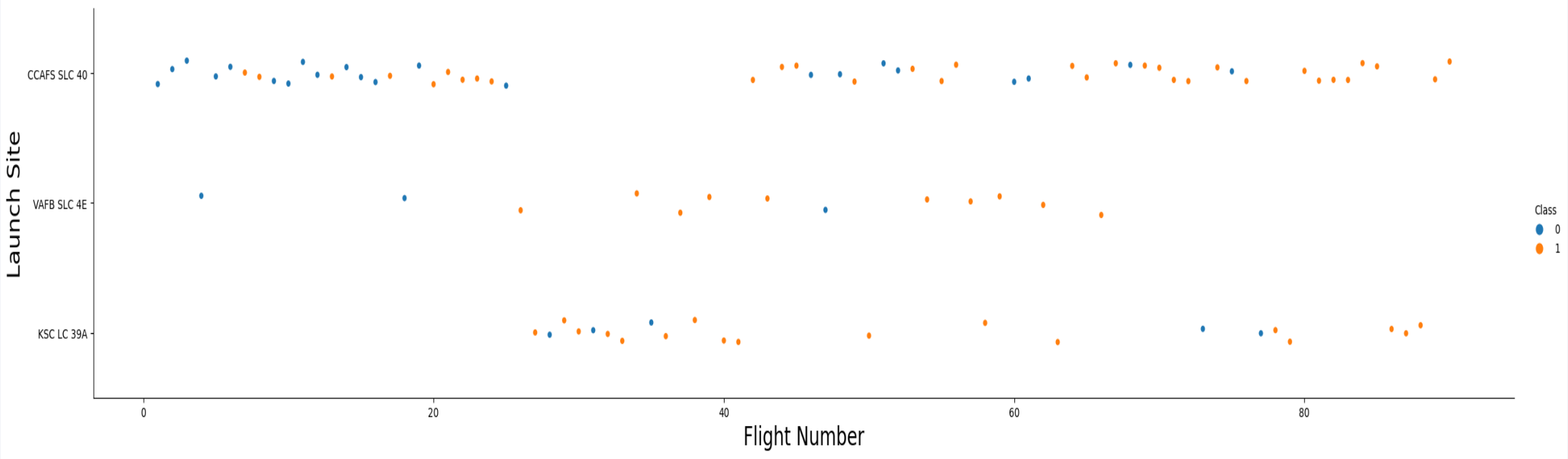


Predictive analysis results



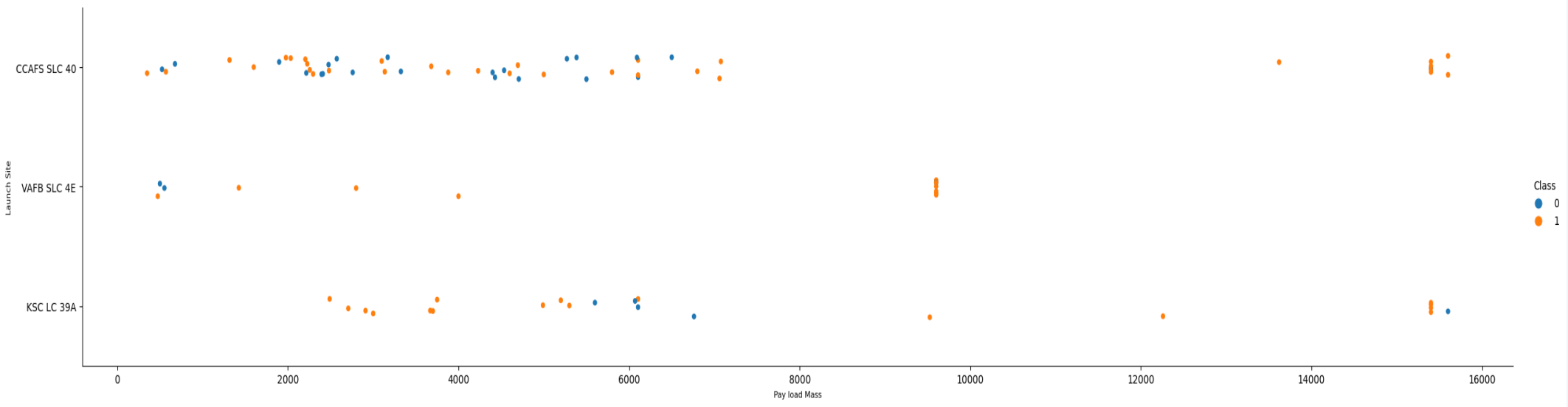

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- ✓ Graphic suggests an increase in success rate over time (indicated in Flight Number).
- ✓ Likely a big breakthrough around flight 20 which significantly increased success rate.
- ✓ CCAFS appears to be the main launch site as it has the most volume.

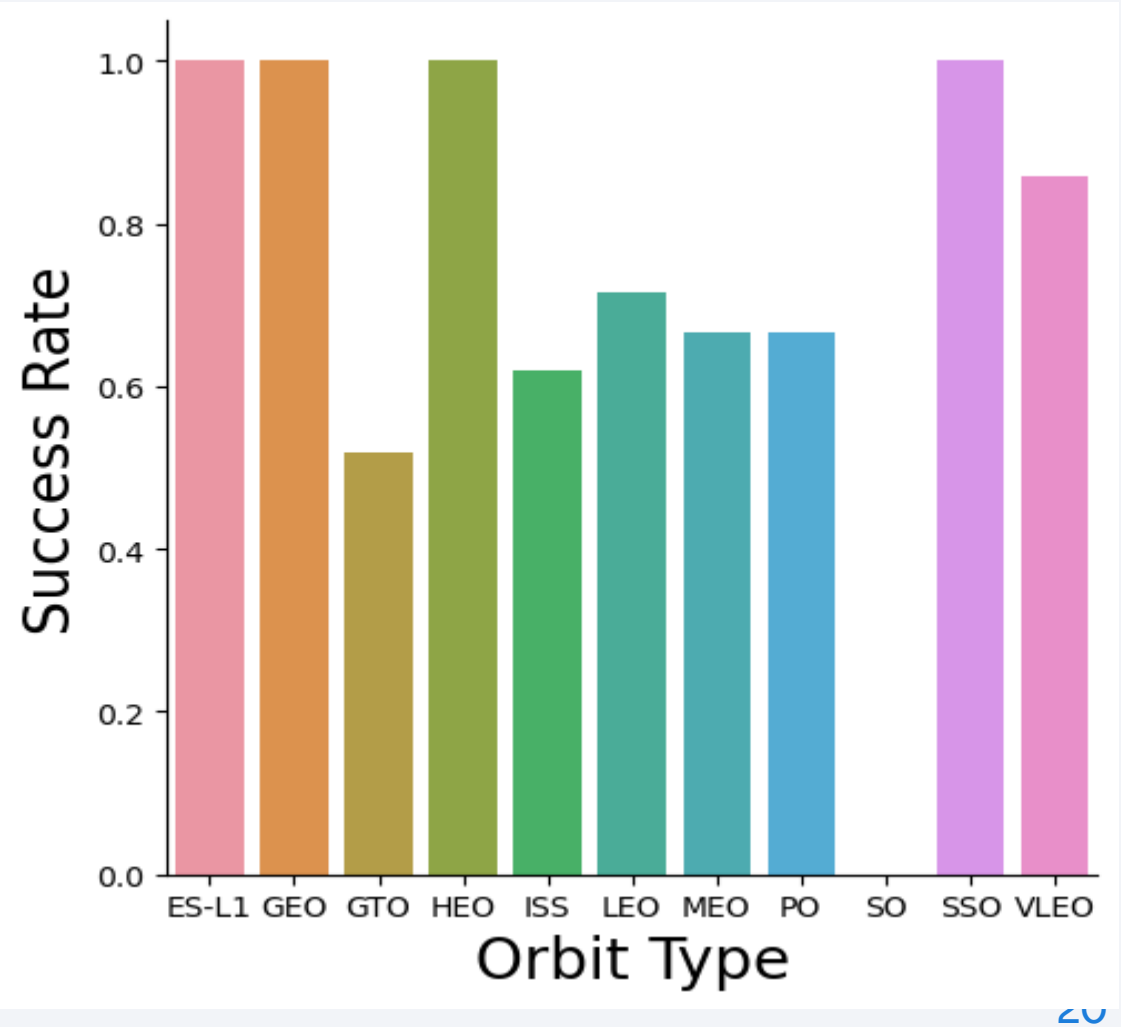
Payload vs. Launch Site



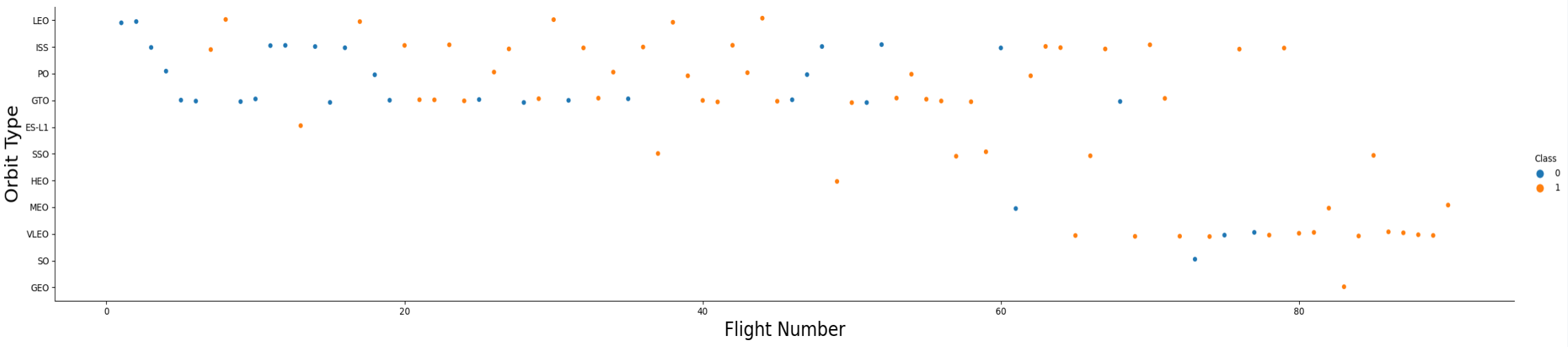
- Payload mass appears to fall mostly between 0-6000 kg.
- With heavy payloads the successful landing or positive landing rates are more.

Success Rate vs. Orbit Type

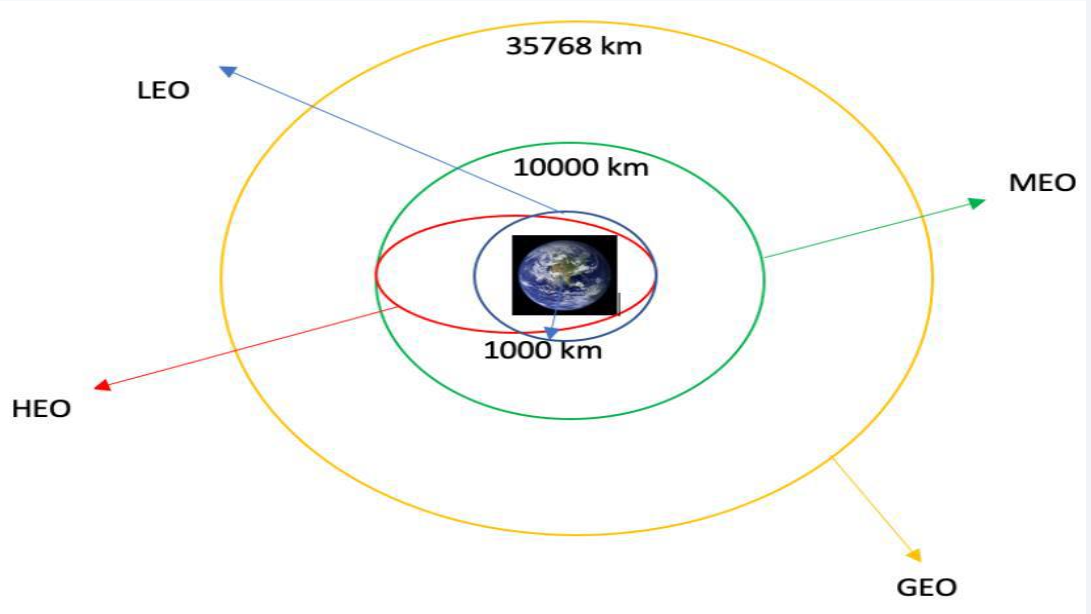
- In terms of orbits, ES-L1, GEO, HEO and SSO have recorded no failed landings. VLEO still manages to exceed 80% success rate, whilst the other orbits have recorded 25% to 50% of failures.
- GTO shows by far the highest occurrence, followed with ISS.



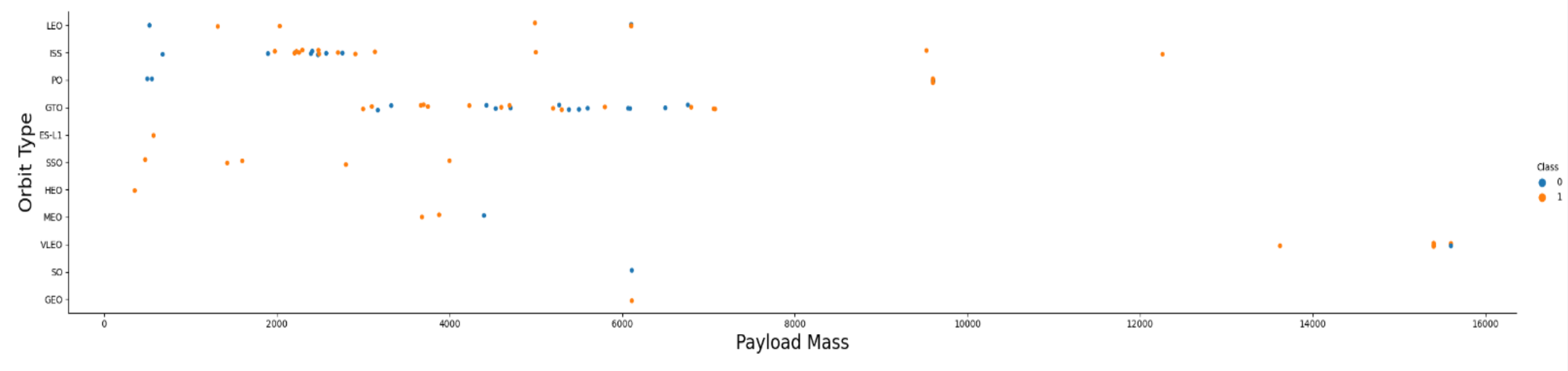
Flight Number vs. Orbit Type



- The higher the Flight number on each orbit, the greater the success rate (Except for GTO orbit).
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.



Payload vs. Orbit Type

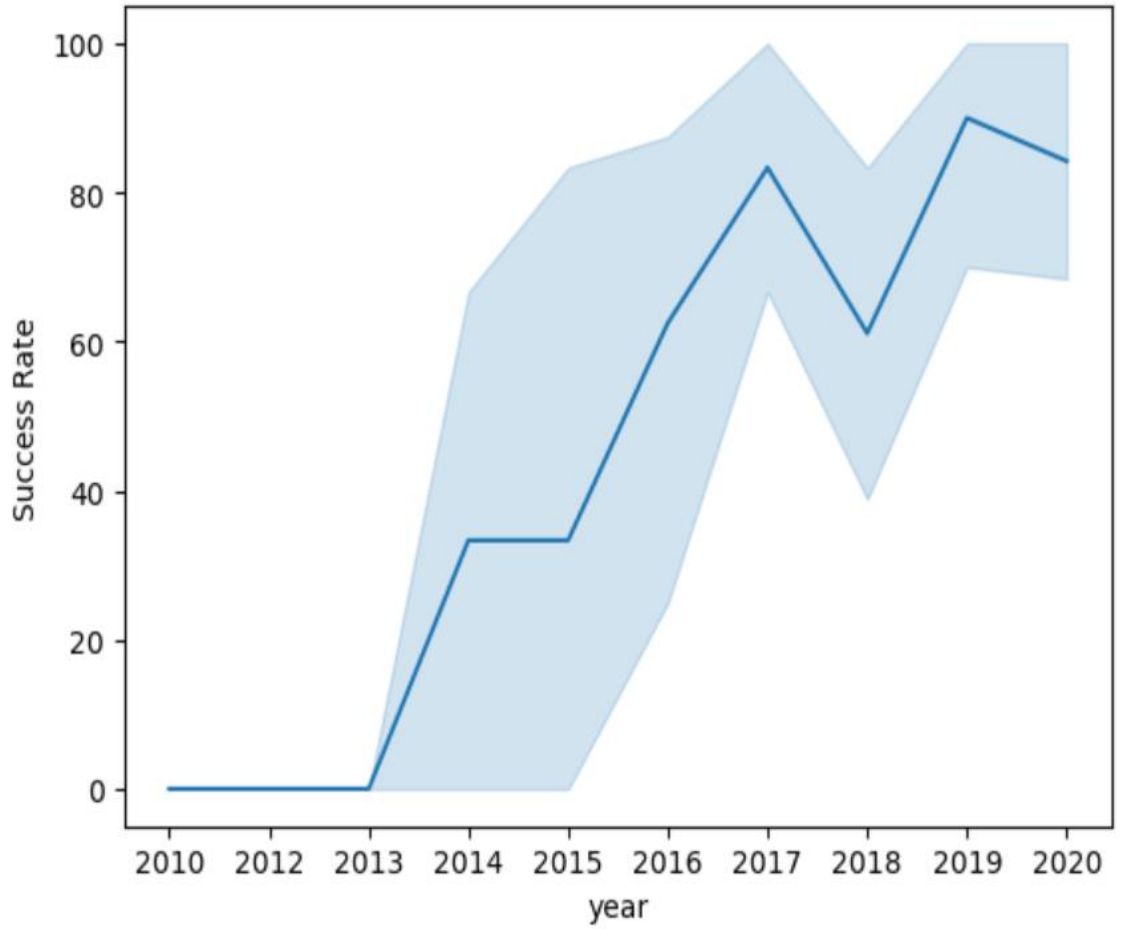


- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

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

- Overall, 95% confidence interval (light blue shading)



EDA with SQL

All Launch Site Names

- The key word DISTINCT is used to specify the unique launch sites from the SpaceX data.

```
%sql select DISTINCT LAUNCH_SITE from SPACEDATASETX
```

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

Launch Site Names Begin with 'CCA'

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

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

The total payload carried by boosters launched from NASA (CRS)

```
%sql select sum(payload_mass__kg_) as sum from SPACEDATASETX where customer like 'NASA (CRS)'
```

```
Done.
```

SUM
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) as Average from SPACEDATASETX where booster_version like 'F9 v1.1%'
```

Done.

average
2534

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad

```
%sql select min(date) as Date from SPACEDATASETX where mission_outcome like 'Success'
```

Done.

DATE
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

List of boosters which have successfully landed on drone ship which had payload mass greater than 4000 but less than 6000

```
%sql select BOOSTER_VERSION from SPACEDATASETX where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_
```

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

The total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEDATASETX GROUP by mission_outcome ORDER BY mission_outcome
```

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

Boosters Carried Maximum Payload

List the names of the boosters which have carried the maximum payload mass using subquery.

```
%sql select BOOSTER_VERSION from SPACEDATASETX where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEDATASETX)
```

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

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEDATASETX WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

Done.

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

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

Rank of 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

```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEDATASETX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

Done.

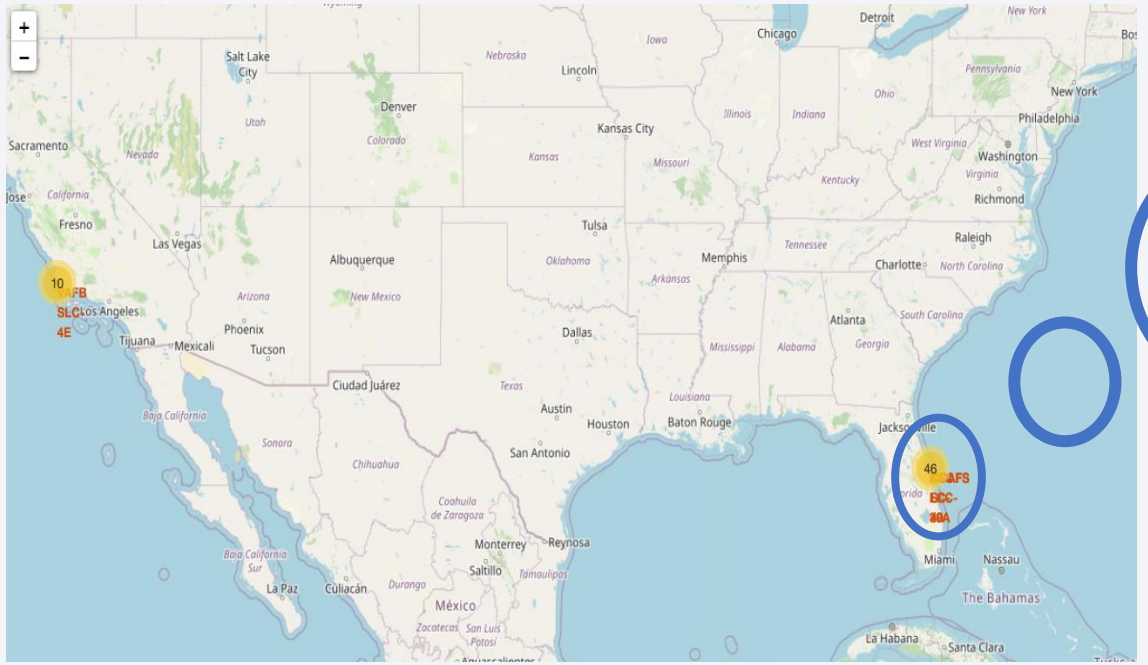
Landing Outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

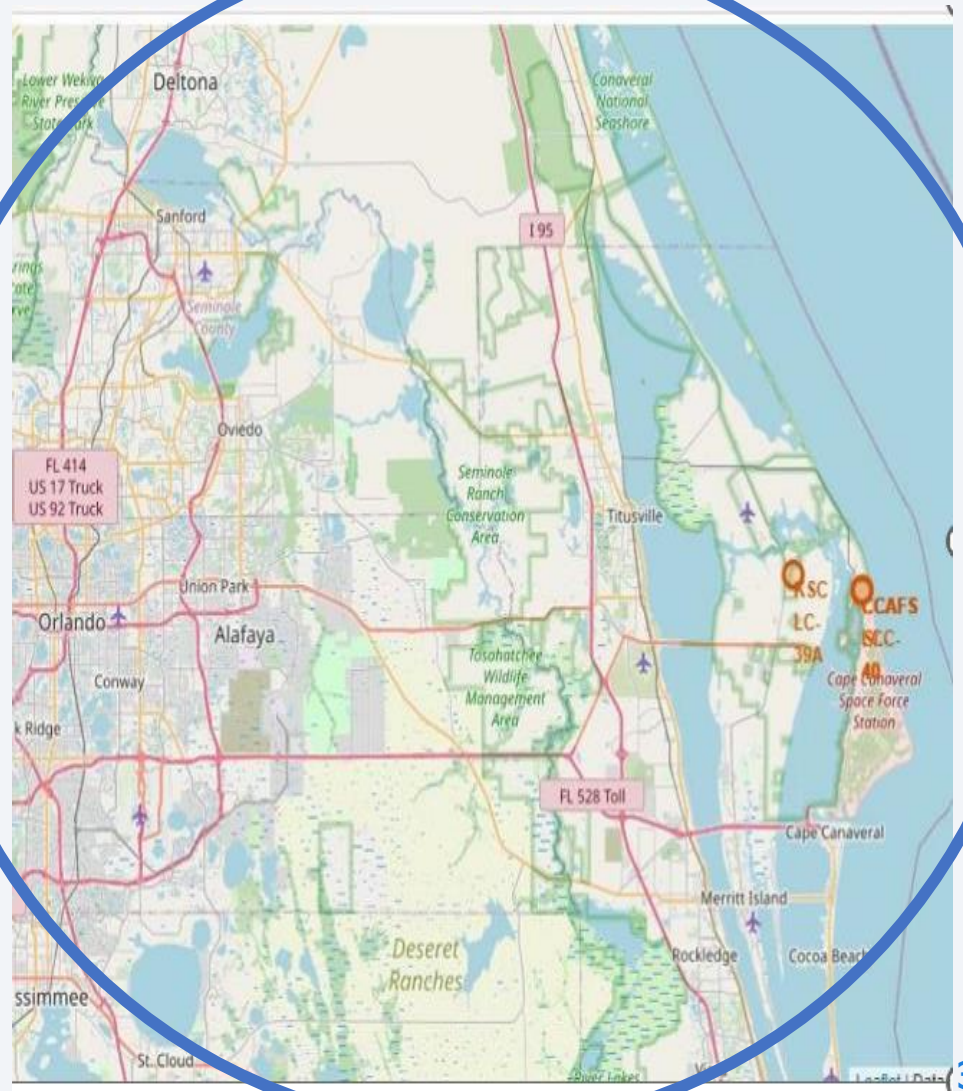
Section 3

Launch Sites Proximities Analysis

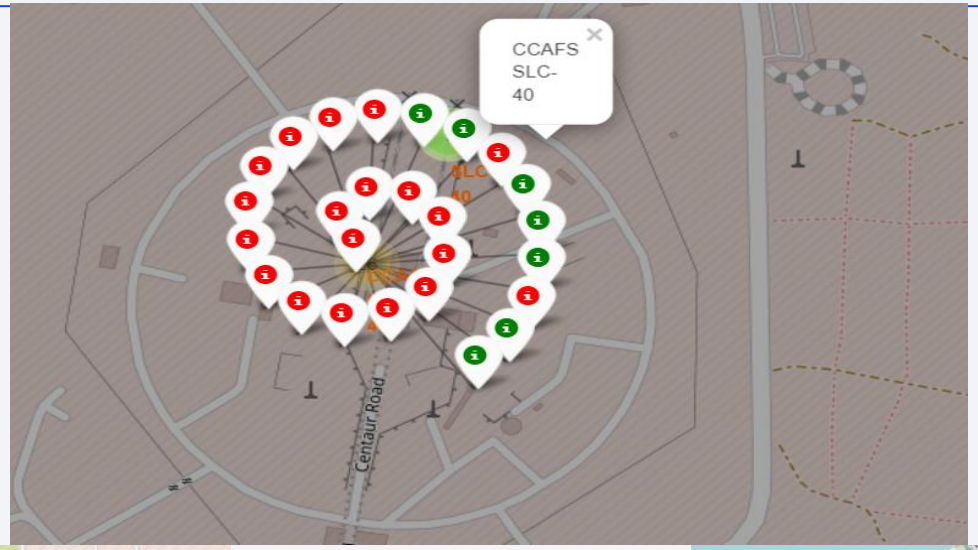
Launch Site Locations

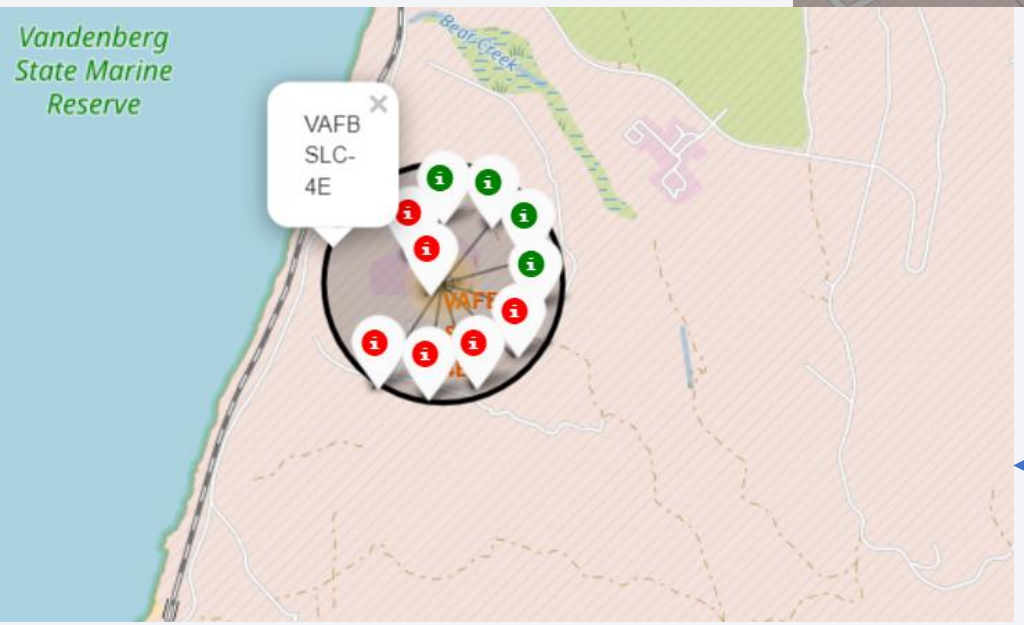
- The left bottom shows the Florida and California launch sites
- In Florida, there are two launch sites (right).
- All sites are located at coastline.



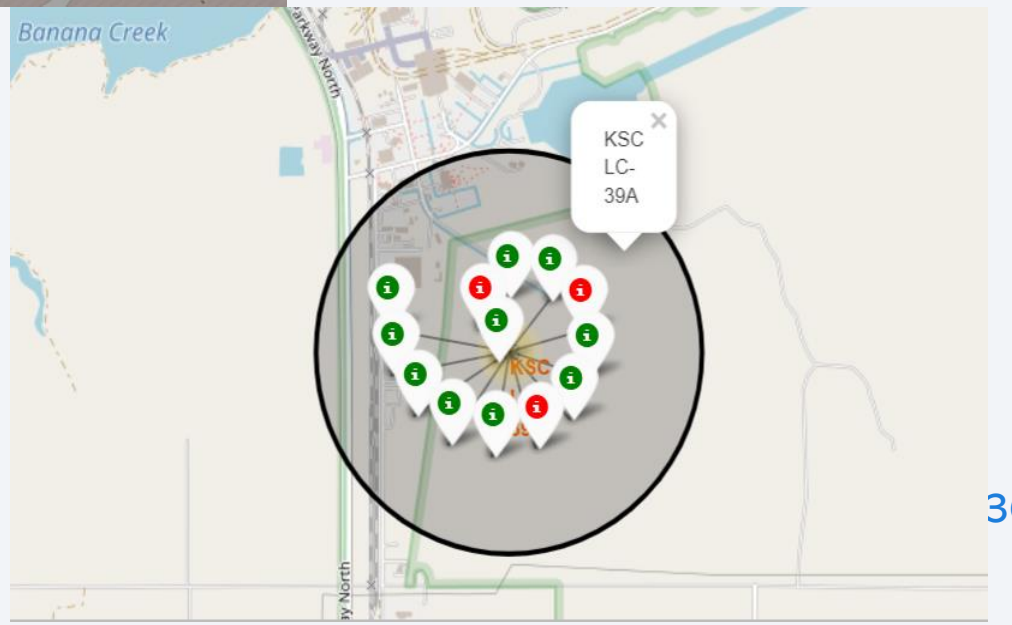


Color Coded Launch Markers

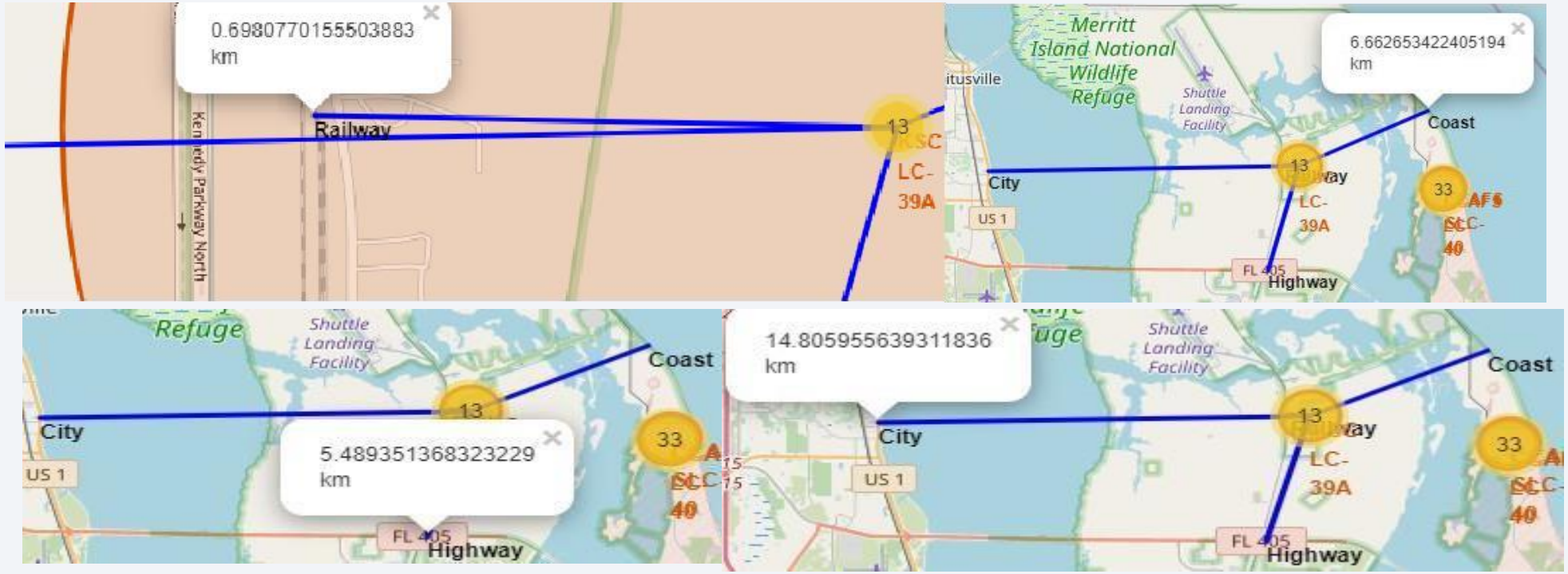




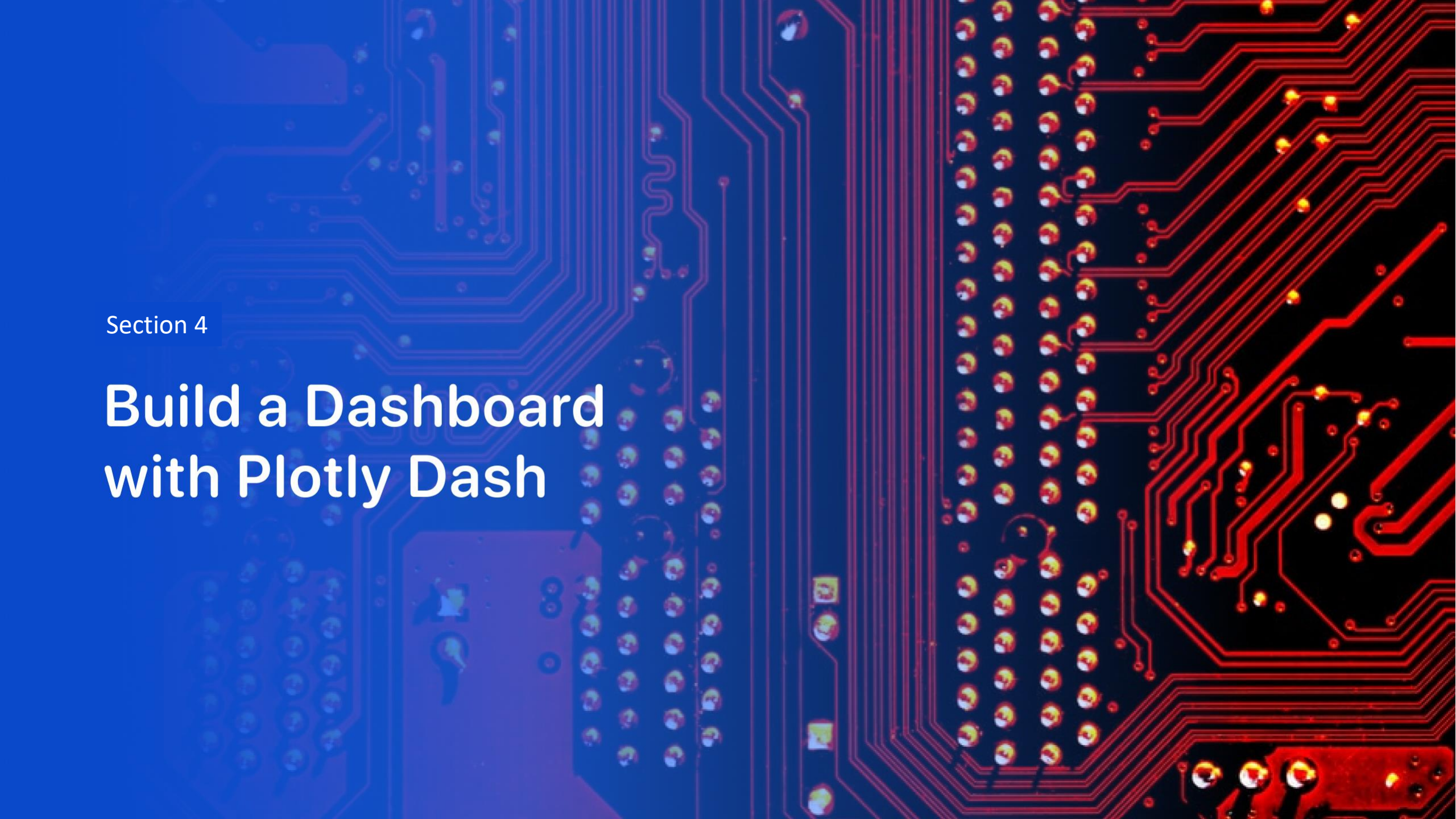
- Colored launch markers display each successful landing (green icon) and failed landing (red icon).
- In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



Launch Site Distances to its Proximities



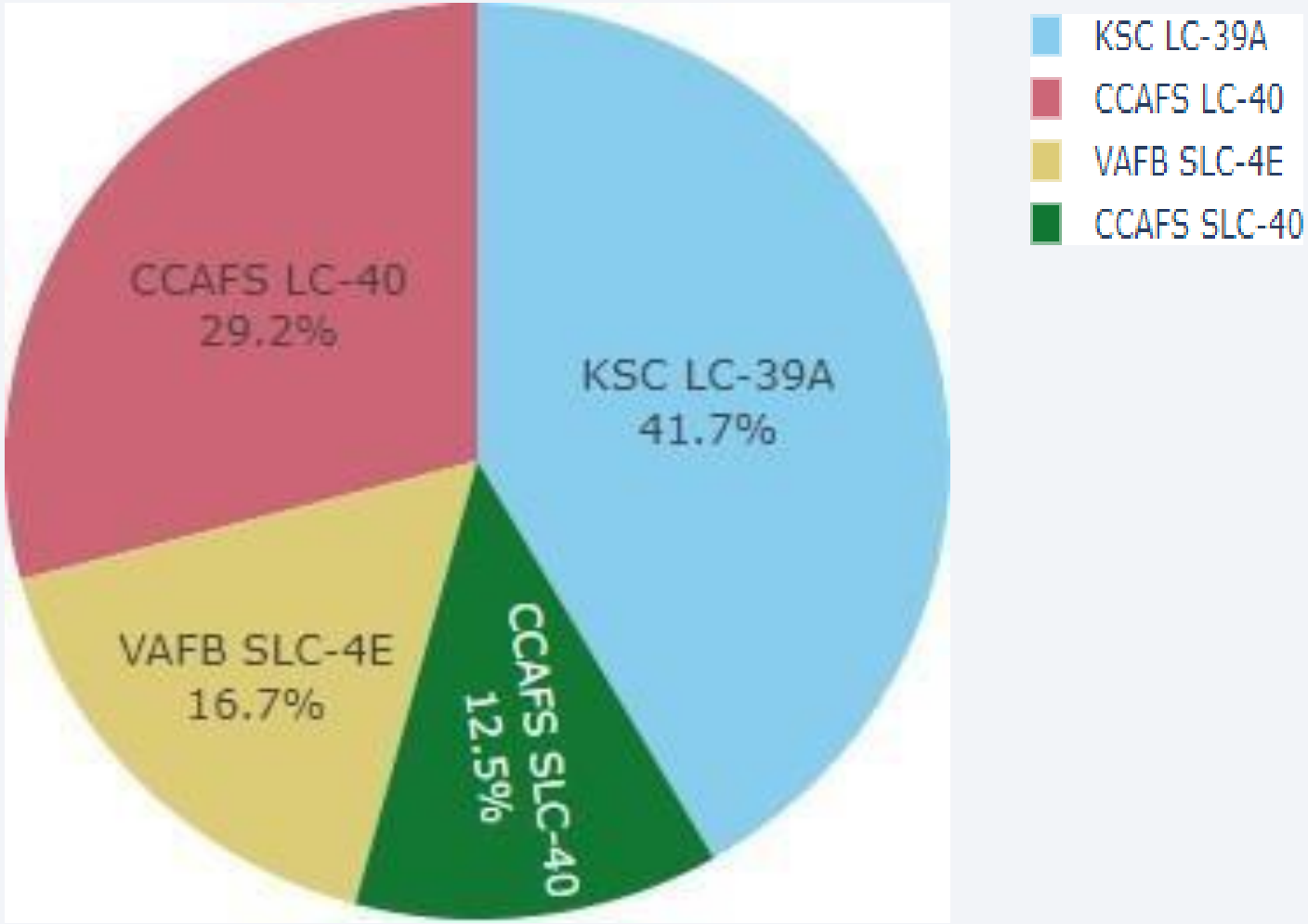
- Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.


Section 4
Build a Dashboard with Plotly Dash

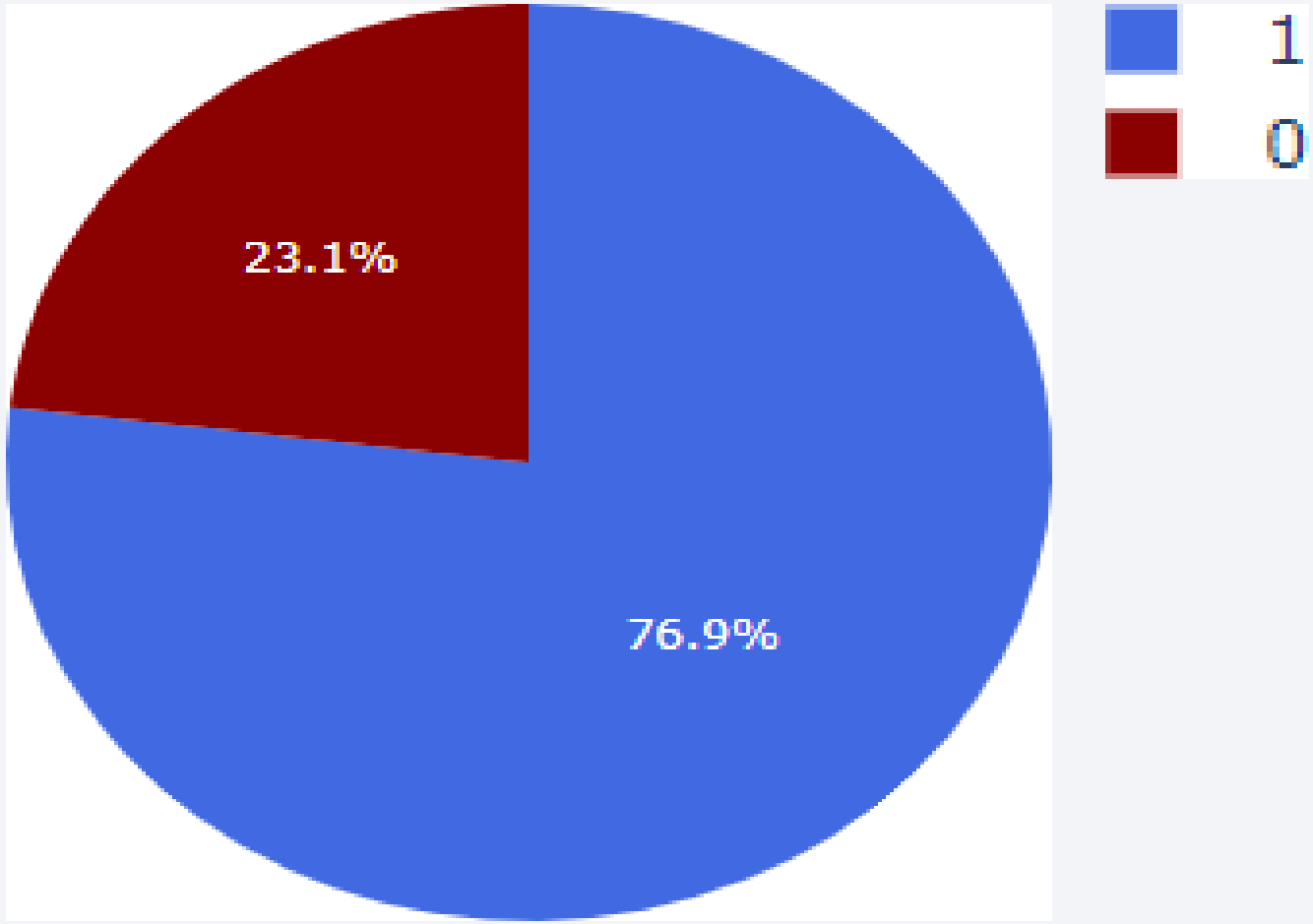
Successful Launches at Each Launch Site

- KSC LC shows the highest share of successful landing.

- VAFB has the smallest share of successful landings. The attributes might be the smaller number of launches and the west coast location.

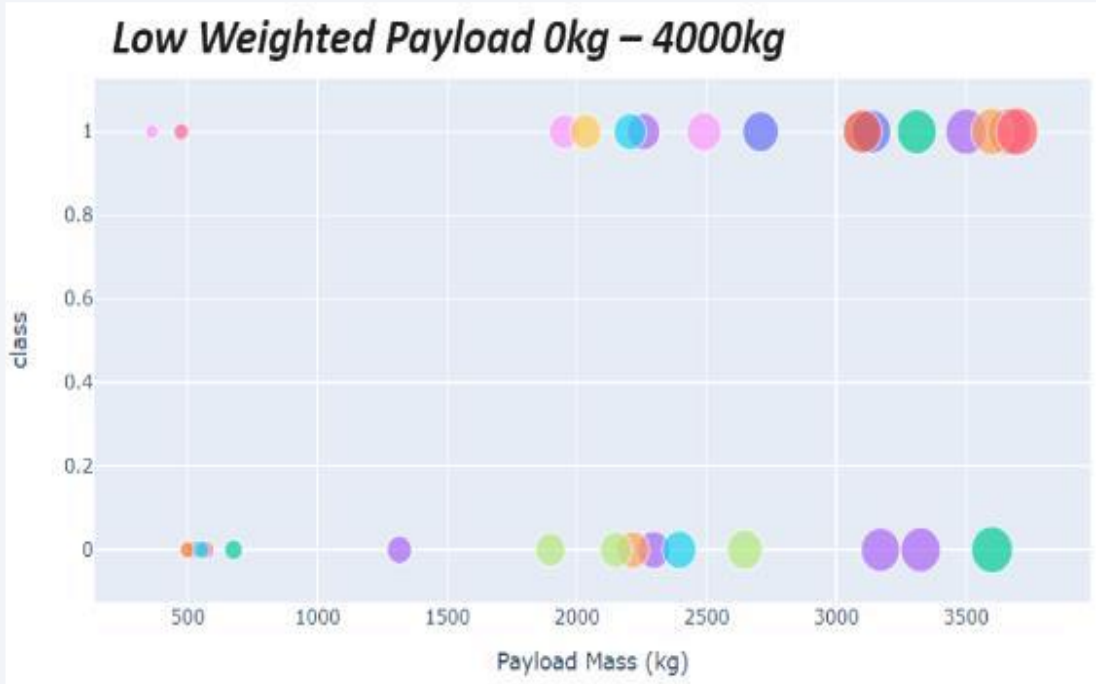


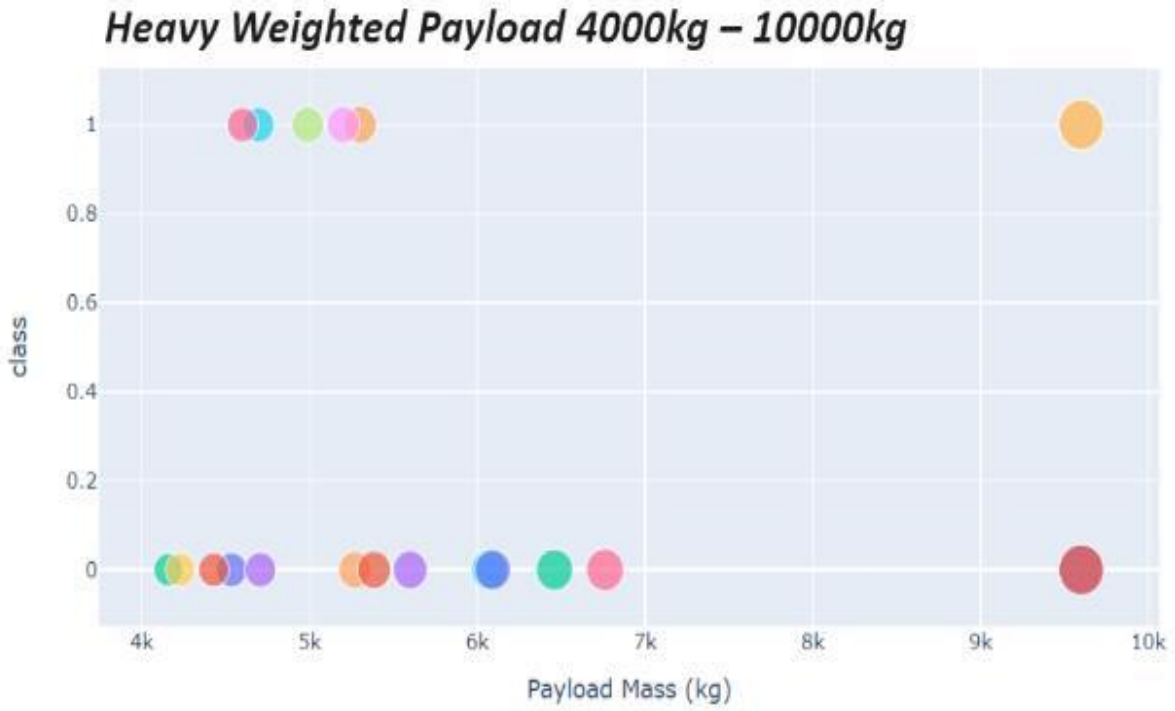
The Highest Success Rate Launch Site



- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Payload VS Launch Outcome





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Find the method performs best:

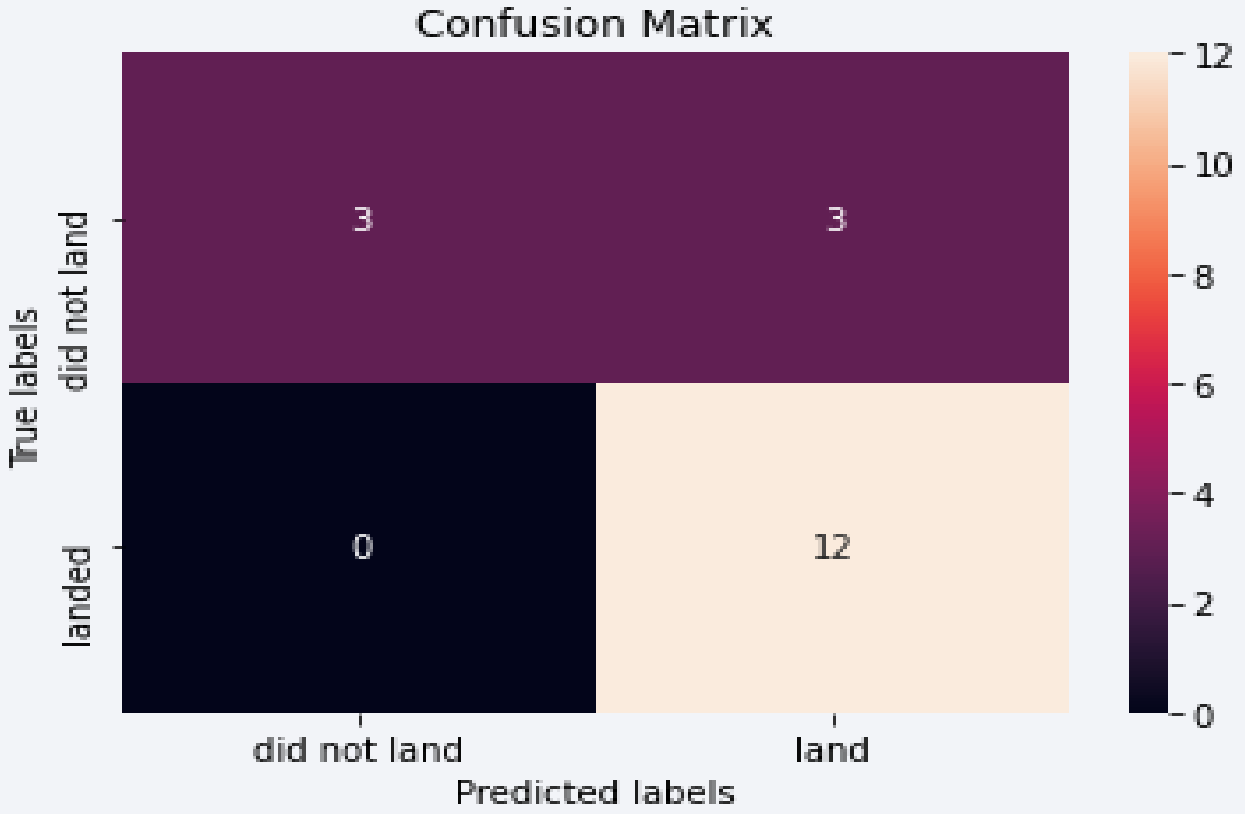
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9027777777777778
Best params is : {'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}
```

- Best model DecisionTree with a score of 0.90

Confusion Matrix



- Correct predictions are on a diagonal from top left to bottom right.

Conclusions

- The Tree Classifier Algorithm is the best Machine Learning approach for this dataset.
- The low weighted payloads (which define as 4000kg and below) performed better than the heavy weighted payloads.
- Starting from the year 2013, the success rate for SpaceX launches is increased, directly proportional time in years to 2020, which it will eventually perfect the launches in the future.
- KSC LC-39A have the most successful launches of any sites; 76.9%
- SSO orbit have the most success rate; 100% and more than 1 occurrence.

Appendix

- Programing Language: Python
- Database: SQL, DB2
- IDE: Jupyter Notebook, Pycharm
- Version Control Tools: Github
- Github Repo URL: https://github.com/t1-michael/Applied-Data-Science-Capstone
- Course URL: https://www.coursera.org/professional-certificates/ibm-data-science

Thank you!