IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chris North
18th February 2022

Outline

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

Executive Summary

Summary of methodologies:

- Data Collection - API
- Data Collection - Web Scraping
- Data Wrangling
- Exploratory Data Analysis - SQL
- Exploratory Data Analysis - Visualization
- Interactive Visual Analytics – Folium
- Interactive Dashboard – Plotly Dash
- Machine Learning Prediction

Summary of all results:

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

Introduction

Project background and context:

In this project, we will predict 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 you want to find answers:

1. Determine when / whether Space X re-uses the 1st stage and what factors & features affect this
2. Determine the cost of each launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API & web scraping Wikipedia
- Perform data wrangling
  - Data was processed to create a landing outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data Collection - API

- Data was collected using get request to the Space X API
- The response content was decoded as a Json and turned into a Pandas data frame
- Columns combined into a Dictionary
- Data filtered to only include Falcon 9 launches
- Missing values identified and replaced where appropriate

Data Collection - Web Scraping

- Data collected from Wikipedia page titled: List of Falcon 9 & Falcon Heavy Launches
- Data extracted using BeautifulSoup
- Falcon 9 launch records HTML table extracted
- Table parsed and converted into Pandas data frame

Data Collection – SpaceX API

- Data was collected using get request to the Space X API
- The response content was decoded as a Json and turned into a Pandas data frame
- Columns combined into a Dictionary

- [GitHub URL of the completed SpaceX API calls notebook]

Request and parse the SpaceX launch data using the GET request

↓

Filter the data frame to only include Falcon 9 launches

↓

Missing values identified and replaced where appropriate

Data Collection - Scraping

- Data collected from Wikipedia page titled: List of Falcon 9 & Falcon Heavy Launches
- Data extracted using BeautifulSoup
- Falcon 9 launch records HTML table extracted
- Table parsed and converted into Pandas data frame

- [GitHub URL of the completed web scraping notebook]()

Request the Falcon9 Launch Wiki page from its URL

↓

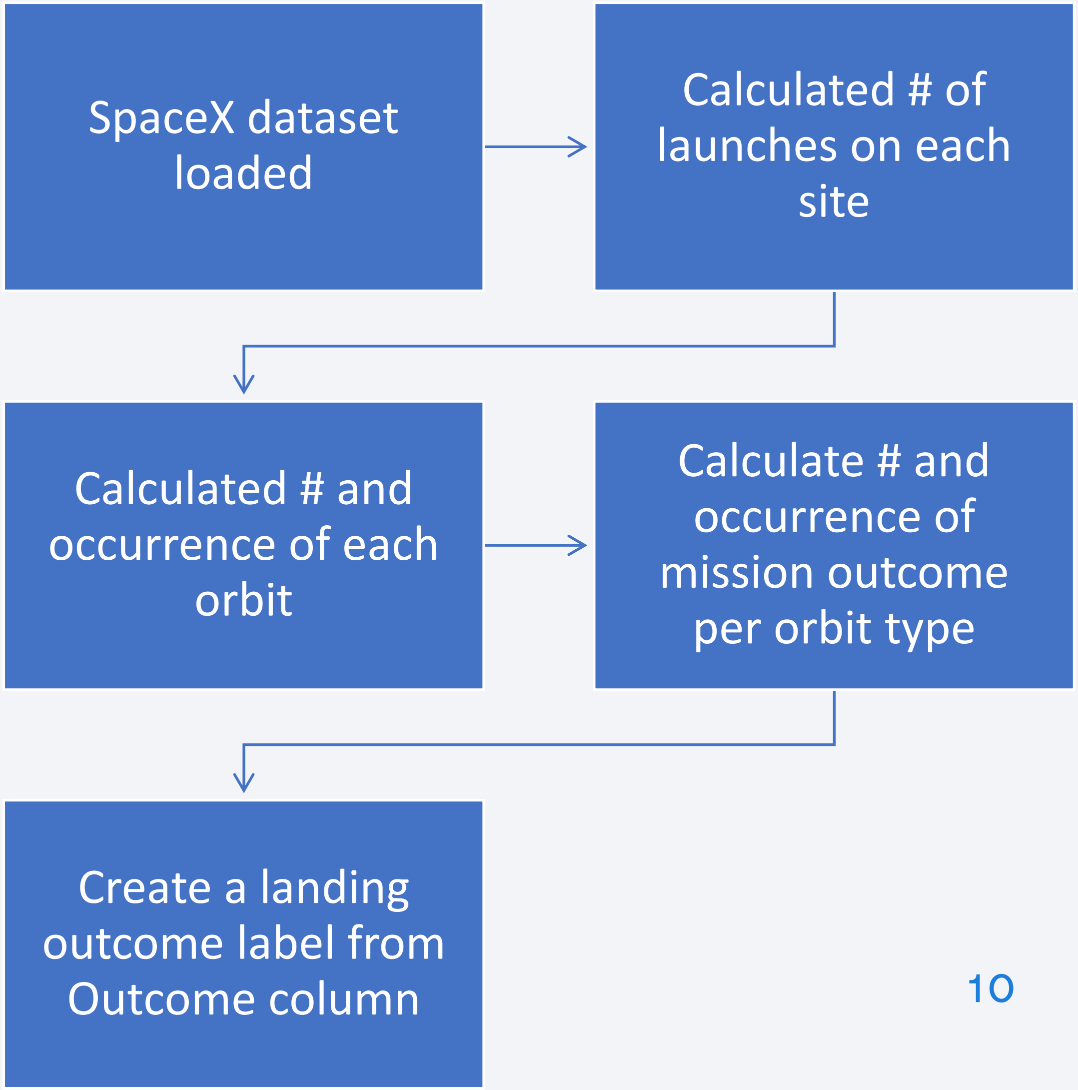
Extract all column/variable names from the HTML table header

↓

Create a data frame by parsing the launch HTML tables

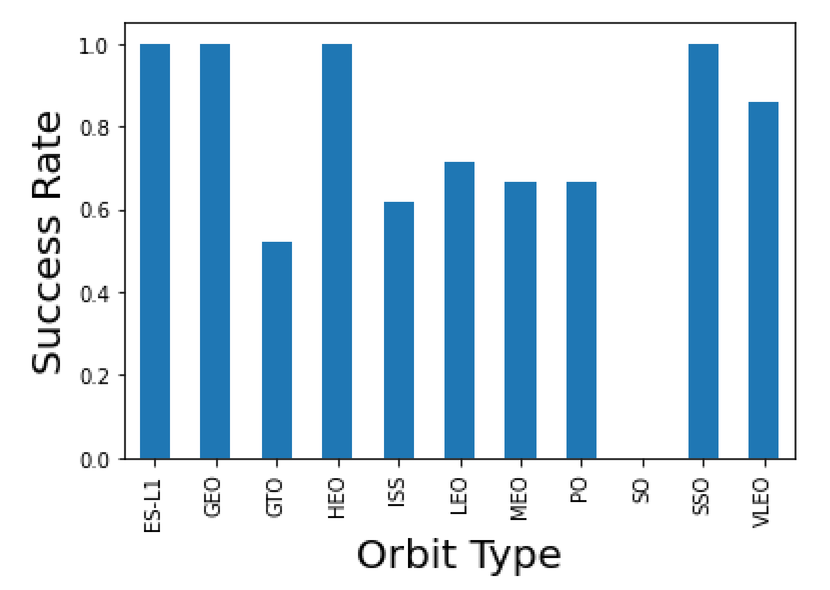
Data Wrangling

- Exploratory data analysis performed and determined training labels (landing outcome)
- [GitHub URL of the completed data wrangling notebooks]()



EDA with Data Visualization

- The following charts were plotted:
  - Flight Number vs Pay Load Mass (scatter)
  - Flight Number vs Launch Site (scatter)
  - Pay Load Mass vs Launch Site (scatter)
  - Success Rate by Orbit Type (bar → example)
  - Flight Number vs Orbit Type (scatter)
  - Pay Load Mass vs Orbit Type (scatter)
  - Yearly Success Rate (line)



- All these charts were plotted to help visualize the relationship between different variables and get a better understanding of the data
- [GitHub URL of completed EDA with data visualization notebook]()

EDA with SQL

- The following SQL queries were performed:
  1. *Displayed the names of the unique launch sites in the space mission*
  2. *Displayed 5 records where launch sites begin with the string 'KSC'*
  3. *Displayed the total payload mass carried by boosters launched by NASA (CRS)*
  4. *Displayed average payload mass carried by booster version F9 v1.1*
  5. *Listed the date where the first successful landing outcome in drone ship was achieved*
  6. *Listed the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000*
  7. *Listed the total number of successful and failure mission outcomes*
  8. *Listed the names of the booster versions which have carried the maximum payload mass using a subquery*
  9. *Listed the records which display the month names, successful landing outcomes in ground pad, booster versions & launch site for the months in year 2017*
  10. *Ranked the count of successful landing outcomes between the dates 2010-06-04 & 2017-03-20 in descending order*
- GitHub URL of completed EDA with SQL notebook

Build an Interactive Map with Folium

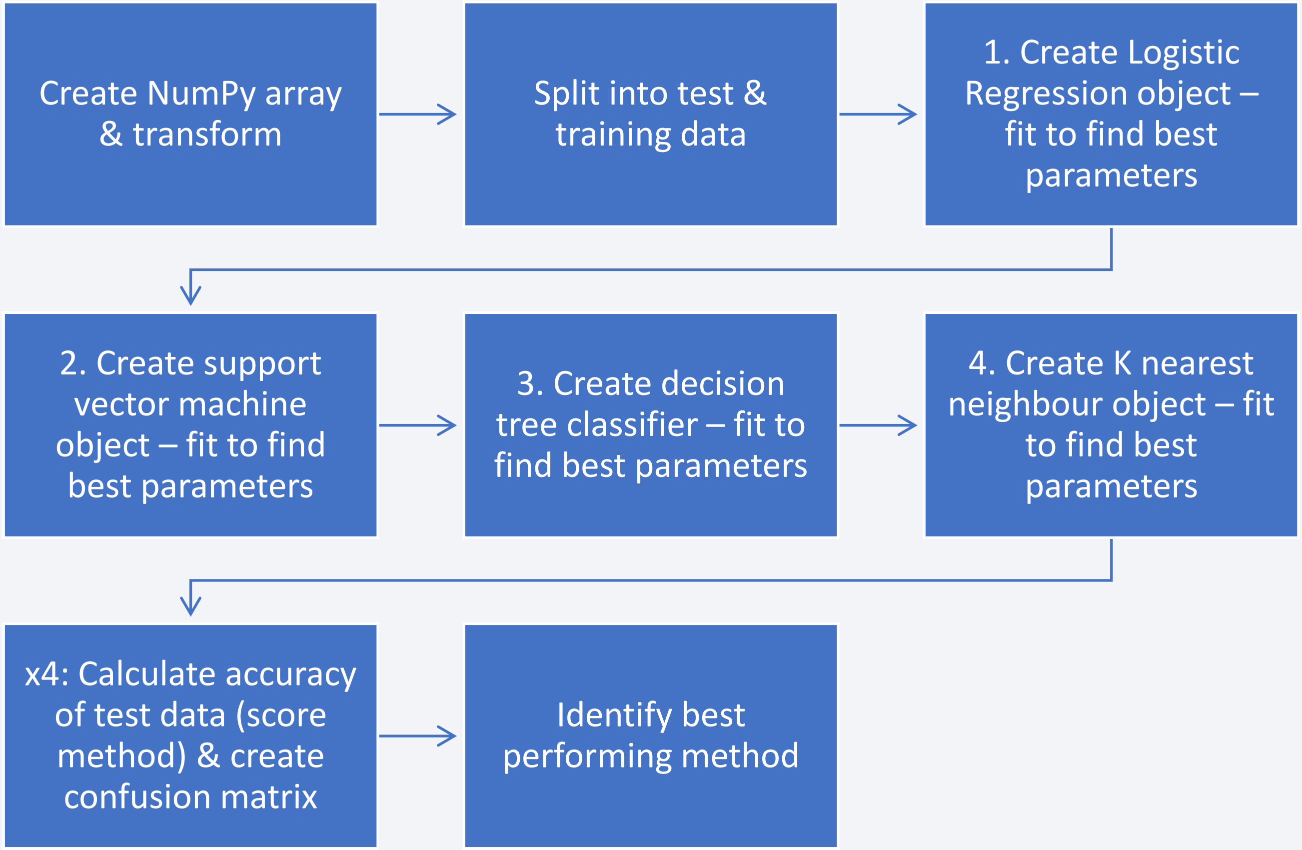
- Map objects created and added as follows:
  - All launch sites using markers
  - Success / failure for each launch site using colors clusters
  - Distance between launch sites to its proximities using PolyLine
- These objects were added in order to identify geographical patterns about launch sites, for example, proximity to coastlines
- GitHub URL of completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- The Dashboard is comprised of the following:
  - Launch site drop-down input component
  - Callback function to render pie chart based on selected site drop-down
  - Range slider to select pay load mass
  - Callback function to render payload mass scatter plot based on range slider
- These plots and interactions enable various visual observations to improve our understanding of the data:
  - Identify which sites have the largest success counts and detailed success rate by site
  - How payload may be correlated with mission outcomes for selected site(s)
  - Visualization of mission outcomes with different boosters
- [GitHub URL of the completed Plotly Dash code]() (unable to save lab environment to GitHub so provided completed code instead)

Predictive Analysis (Classification)

- Performed exploratory data analysis
- Build different machine learning models and tune different parameters using GridSearchCV
- Evaluated using score method and confusion matrices
- [GitHub URL of completed predictive analysis lab]



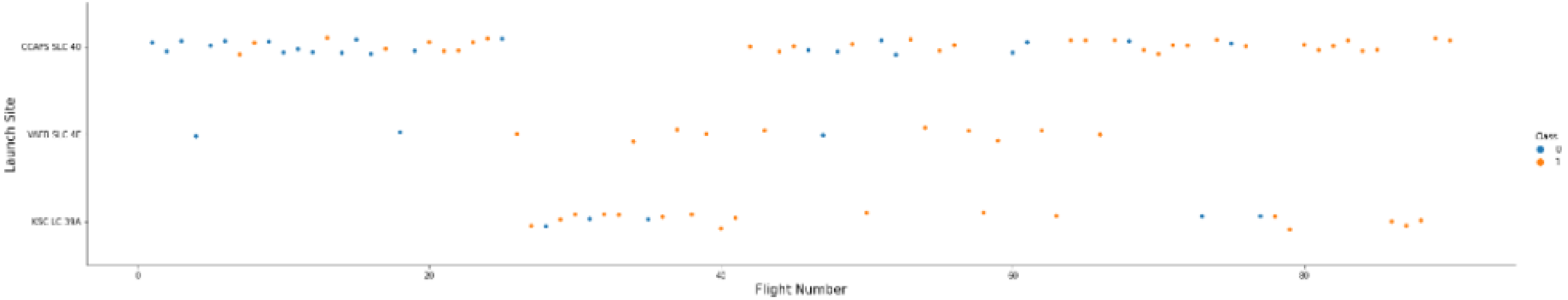
Results

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

Section 2

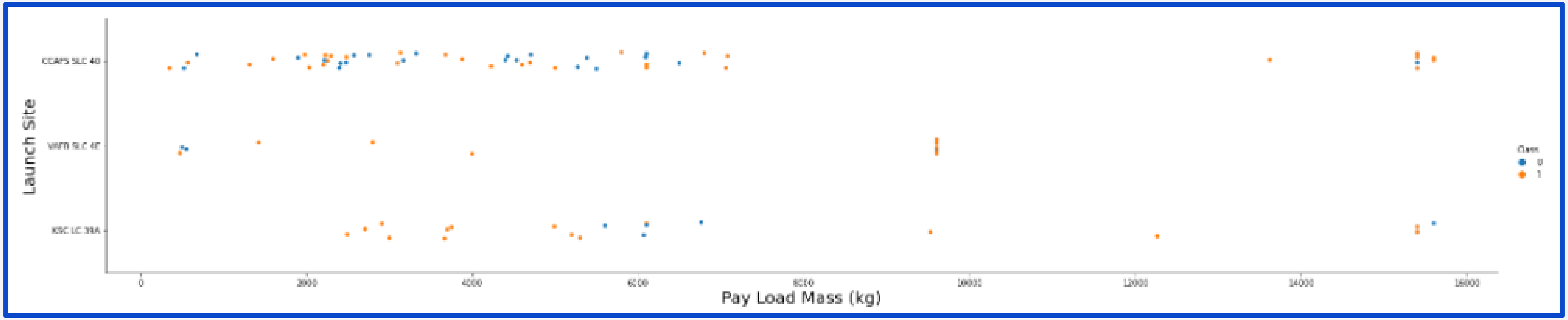
Insights drawn from EDA

Flight Number vs. Launch Site



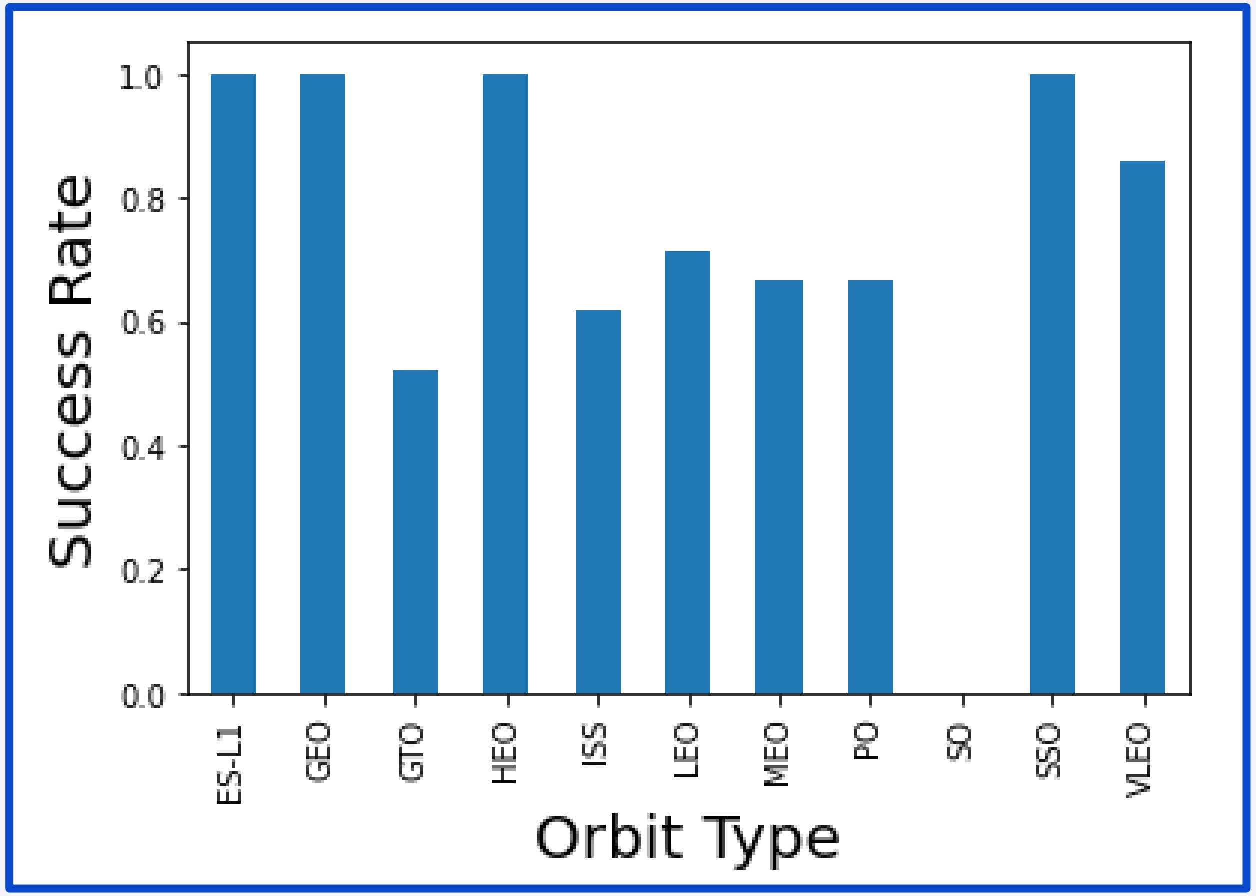
General trend shows success increasing through flight attempts

Payload vs. Launch Site



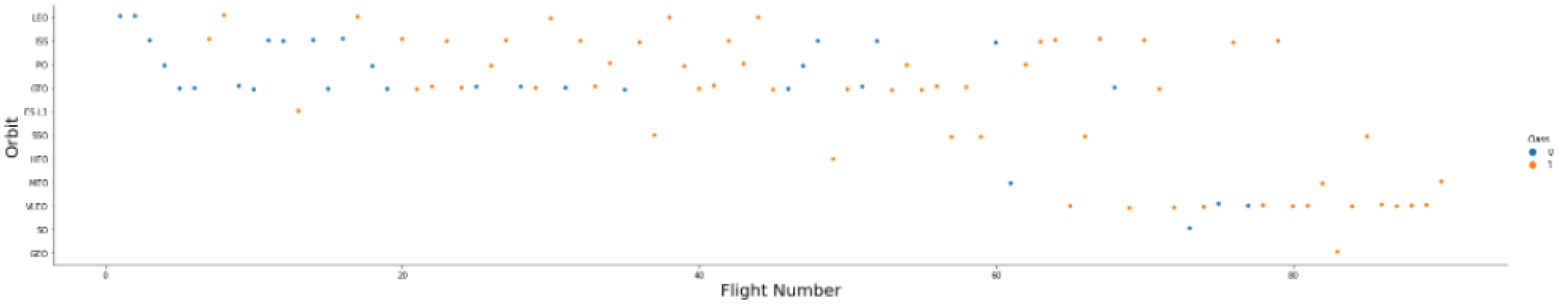
No Payload Launches >10,000KG @ VAFB-SLC and general trend also below at other sites

Success Rate vs. Orbit Type



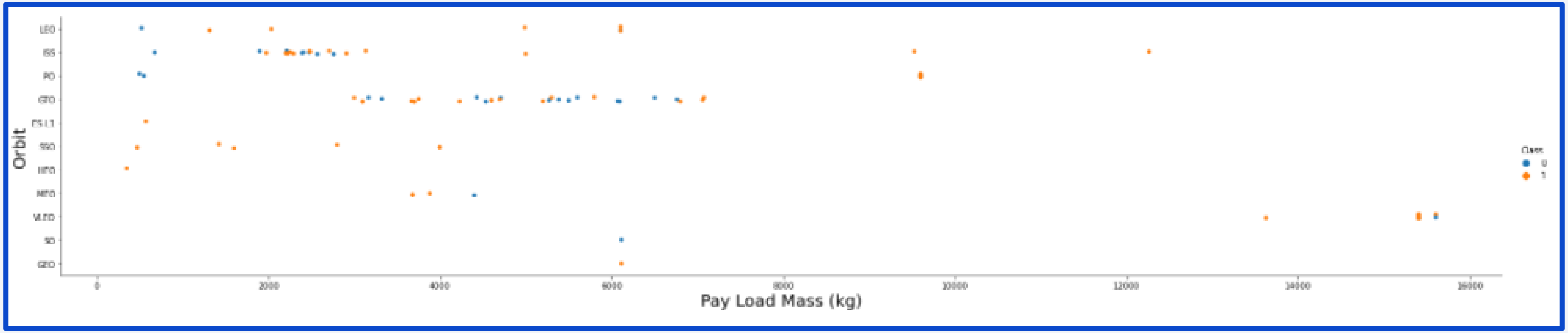
Bar Chart shows us that Orbit Types ES-L1, GEO, HEO & SSO have the best success rate

Flight Number vs. Orbit Type



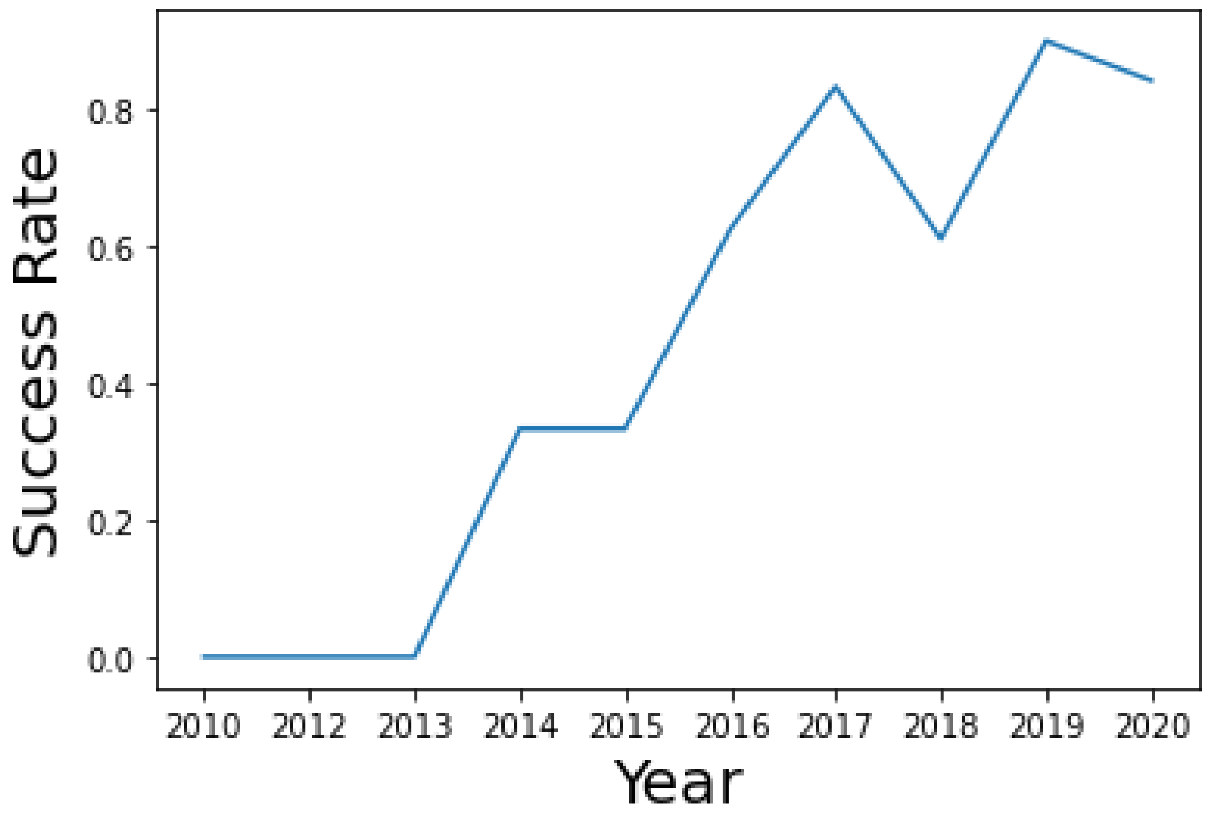
LEO orbit Success appears related to the number of flights; whereas there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



With heavy payloads the successful landing rate is better for PO, LEO and ISS

Launch Success Yearly Trend



After 3 years of initial failure, the success rate has been improving since 2013

All Launch Site Names

In [16]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

```
 * ibm_db_sa://lhp68004:***@2
Done.
```

Out[16]:

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

SQL query showing 4 DISTINCT launch sites

Launch Site Names Begin with 'KSC'

Task 2

Display 5 records where launch sites begin with the string 'KSC'

In [10]:
```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'KSC%'
LIMIT 5;
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

Out[10]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

SQL query showing 5 launch site records beginning with ‘KSC’

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [18]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2
Done.
```

Out[18]:

1
45596

SQL query showing 45,596kg total payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

Task 4

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

In [9]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791
Done.
```

Out[9]:

1
2928

SQL query showing 2928kg average payload mass carried by booster version F9 v1.1

First Successful Drone Ship Landing Date

Task 5

List the date where the first succesful landing outcome in drone ship was acheived.

Hint:Use min function

In [20]:
```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (drone ship)';
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io
Done.
```

Out[20]:

1
2016-04-08

SQL query showing first successful landing outcome in drone ship achieved on 2016-04-08

Note: Updated slide to show Drone Ship landing (not Ground Pad Landing) per Lab question

Successful Ground Pad Landing with Payload between 4000 and 6000

Task 6

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

In [11]:
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cl
Done.
```

Out[11]:

booster_version
F9 FT B1019
F9 FT B1025.1
F9 FT B1031.1
F9 FT B1035.1
F9 B4 B1039.1
F9 FT B1035.2

SQL query showing names of booster versions with success in ground pad with payload mass > 4000kg, but < 6000kg

Note: Updated slide to show Ground Pad landing (not Drone Ship Landing) per Lab question

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [22]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d02186
Done.
```

Out[22]:

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

SQL query showing 100 successful missions and 1 failure outcome

Boosters Carried Maximum Payload

Task 8

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

In [23]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databas
Done.
```

Out[23]:

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

SQL query showing the boosters which have carried the maximum payload mass

2017 Launch Records

Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

In [19]:

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, MONTHNAME(DATE) AS MONTH_NAME
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)'
    AND YEAR(DATE) = 2017;
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

Out[19]:

landing__outcome	booster_version	launch_site	month_name
Success (ground pad)	F9 FT B1031.1	KSC LC-39A	February
Success (ground pad)	F9 FT B1032.1	KSC LC-39A	May
Success (ground pad)	F9 FT B1035.1	KSC LC-39A	June
Success (ground pad)	F9 B4 B1039.1	KSC LC-39A	August
Success (ground pad)	F9 B4 B1040.1	KSC LC-39A	September
Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40	December

SQL query showing month names, successful landing outcomes in ground pad, boosters versions and launch sites for the year 2017

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

Task 10

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

In [36]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://lhp68004:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.ap
Done.
```

Out[36]:

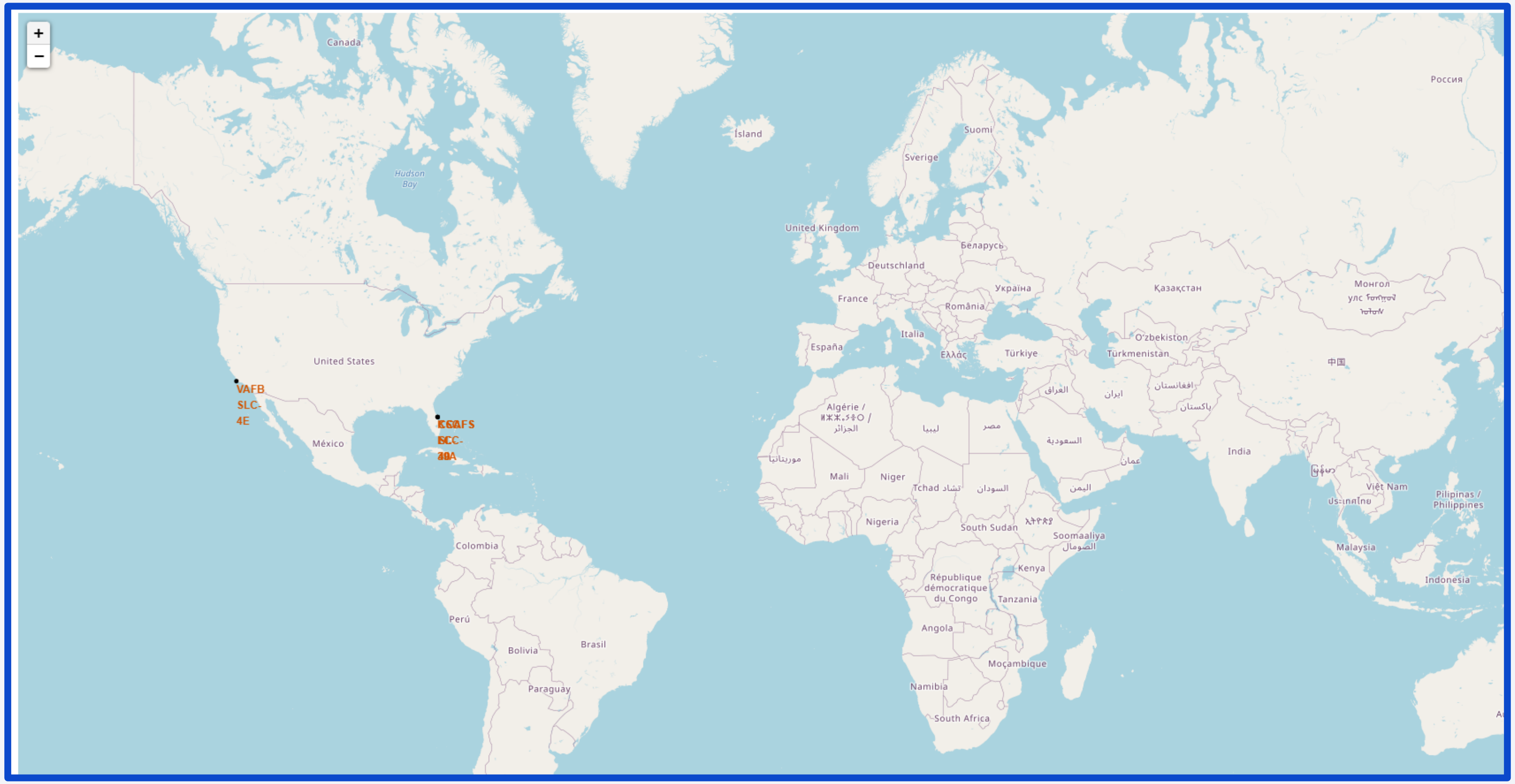
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

SQL query ranking the count of successful landing outcomes between 2010-06-04 & 2017-03-20 in descending order

Section 3

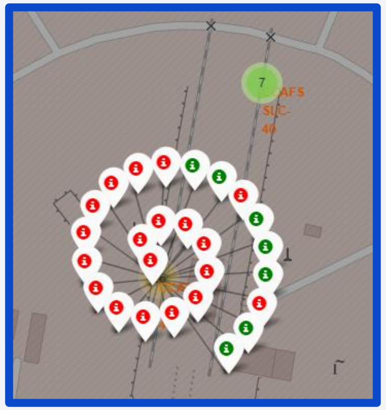
Launch Sites Proximities Analysis

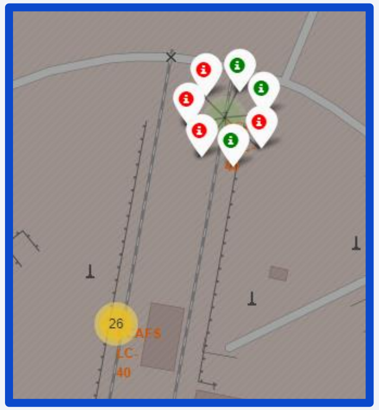
All Launch Sites marked using Folium

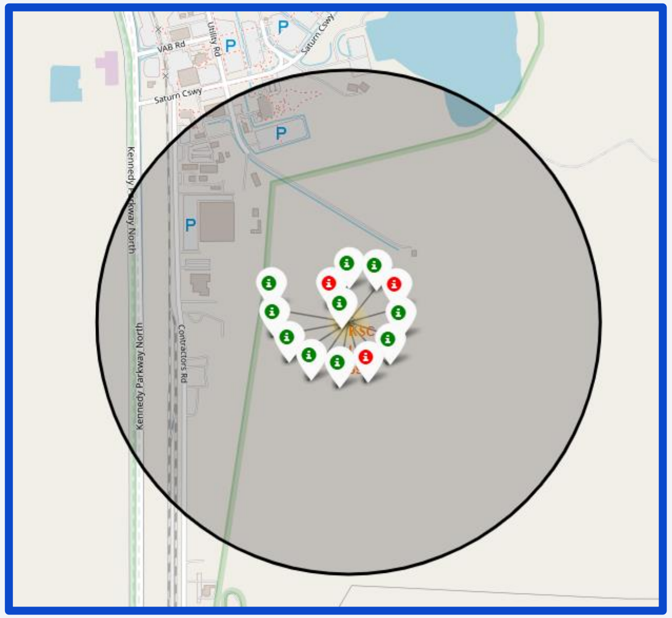


Global launch sites located on coastlines in the USA in the states of Florida & California

Success / failure marked for each site with color labels

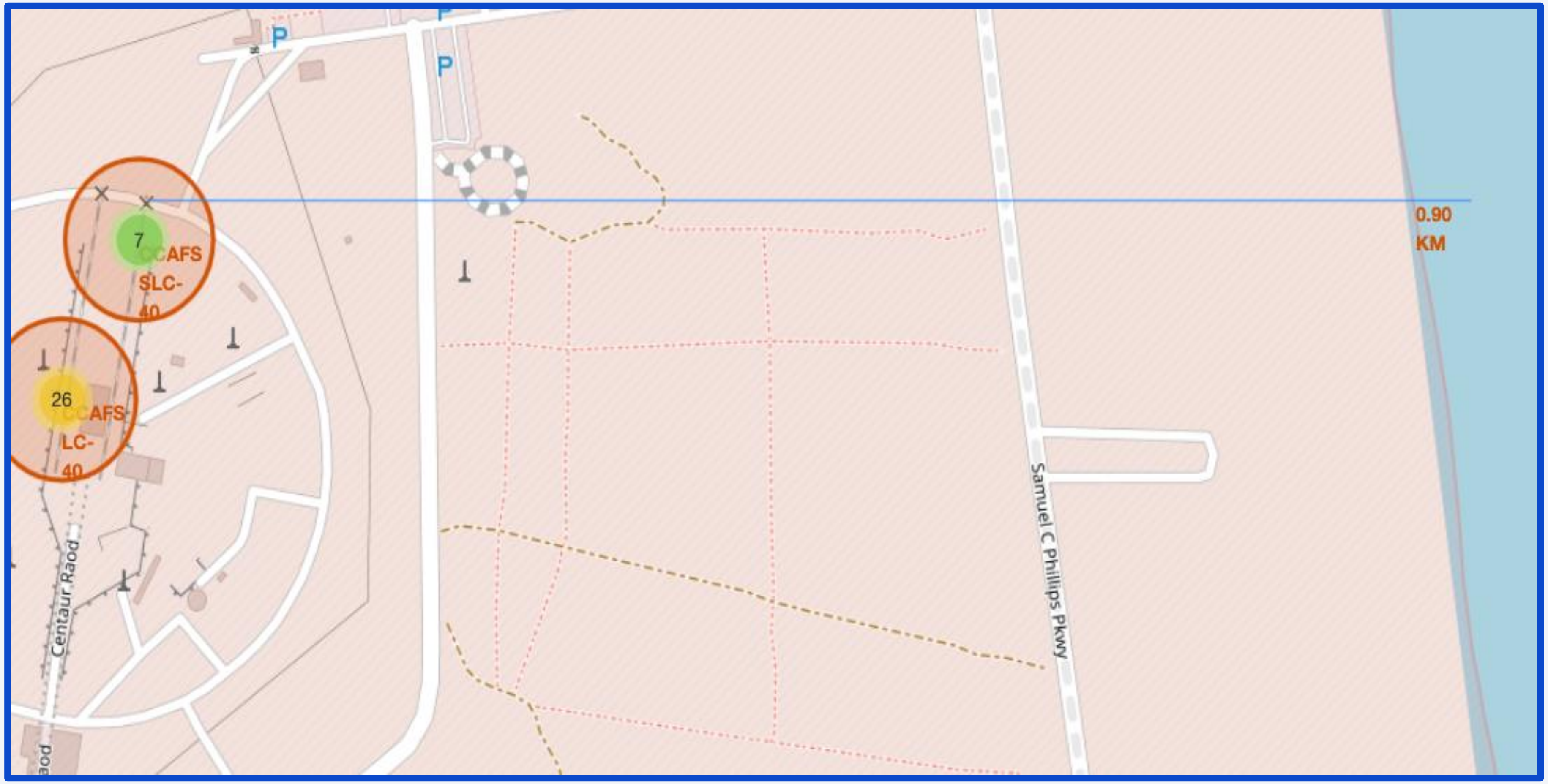






Green markers show successful launches, red markers show failures

Distances between Launch Sites & Proximities

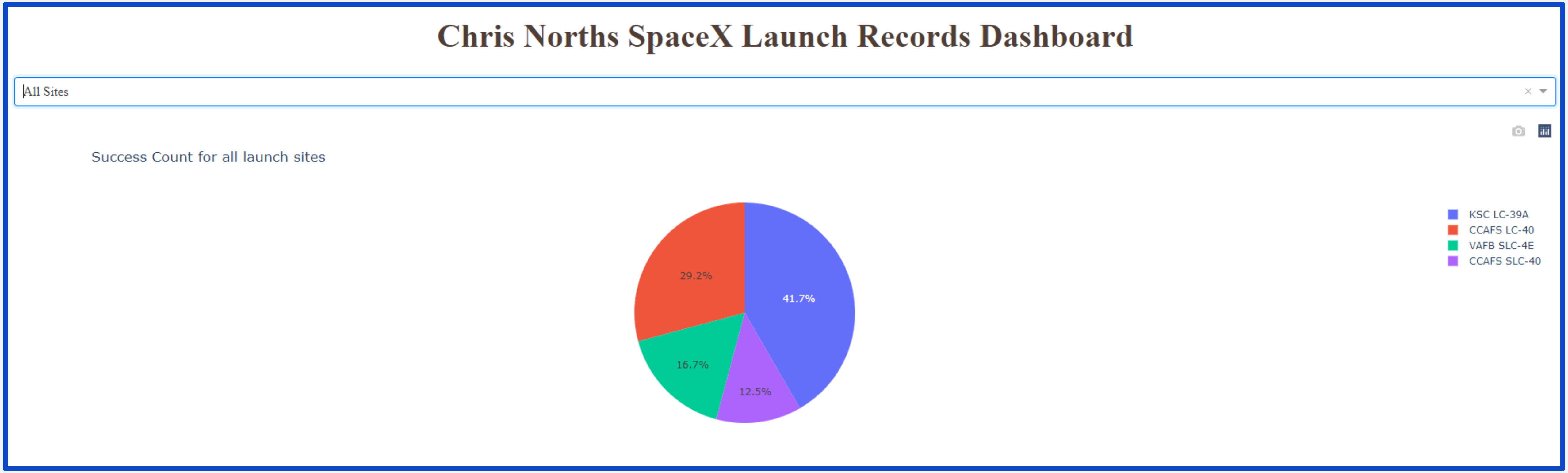


Proximity to coast line shown. Adversely, distance to railways, highways and cities is greater due to safety

Section 4

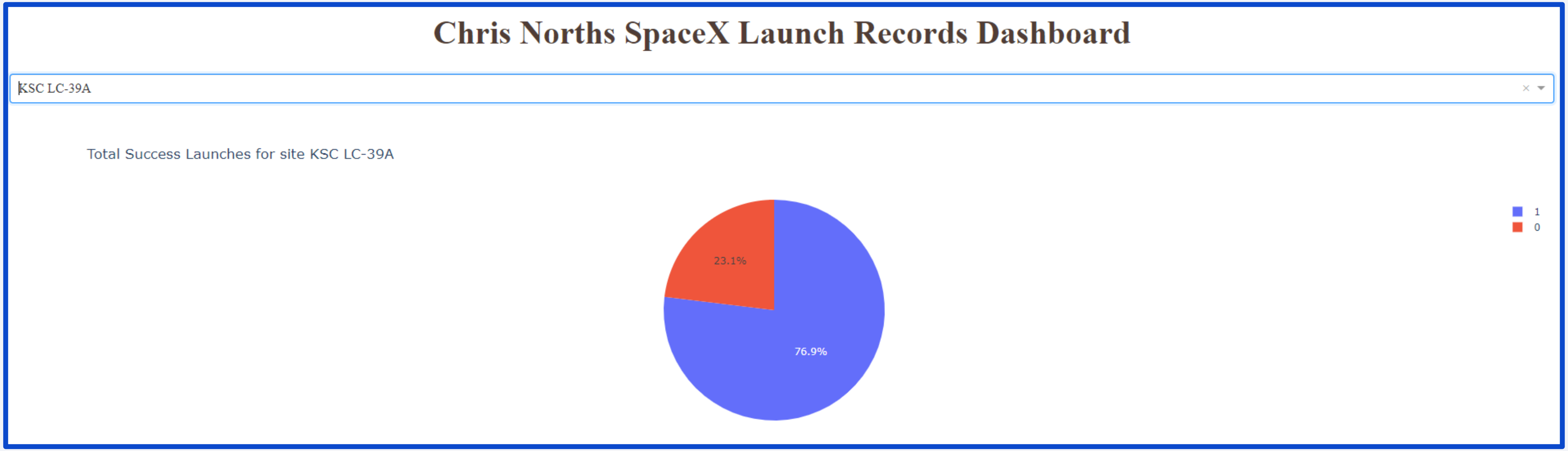
Build a Dashboard with Plotly Dash

Dashboard showing launch site success count for all sites



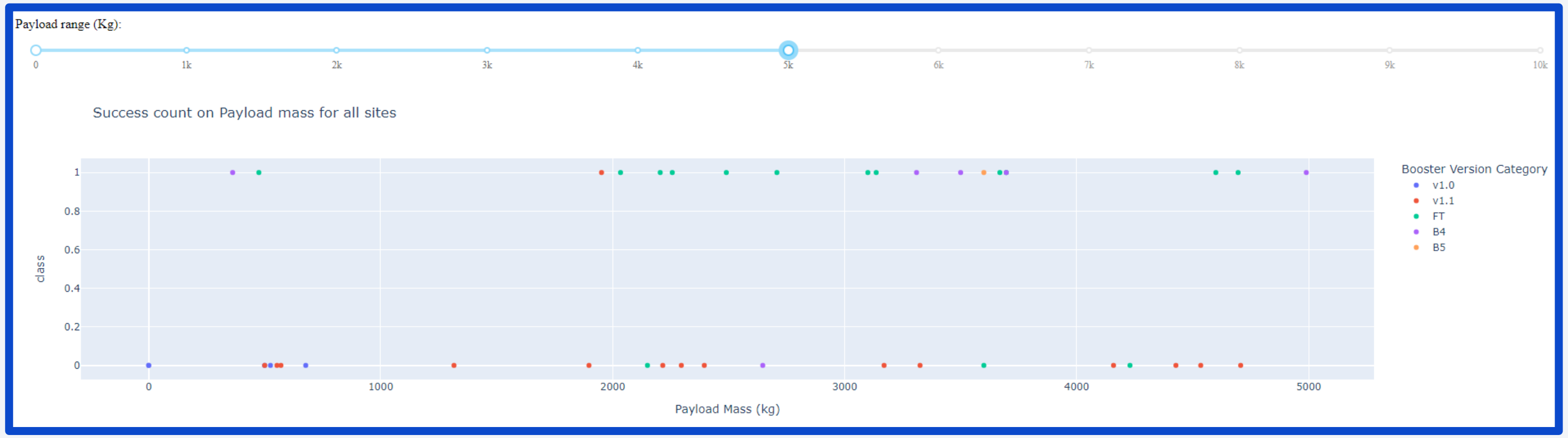
KSC LC-39A has the highest success count, CCAFS SLC-40 the lowest

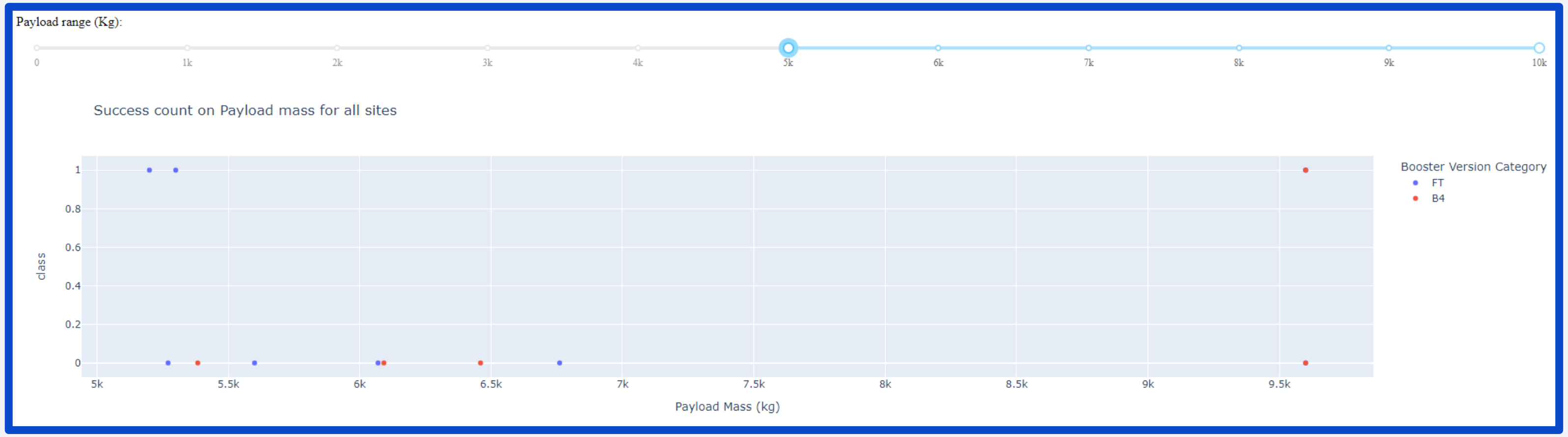
Dashboard showing launch site with the highest launch success ratio



KSC LC-39A has the highest success ratio of 76.9%

Payload v Launch Outcome scatter for all sites



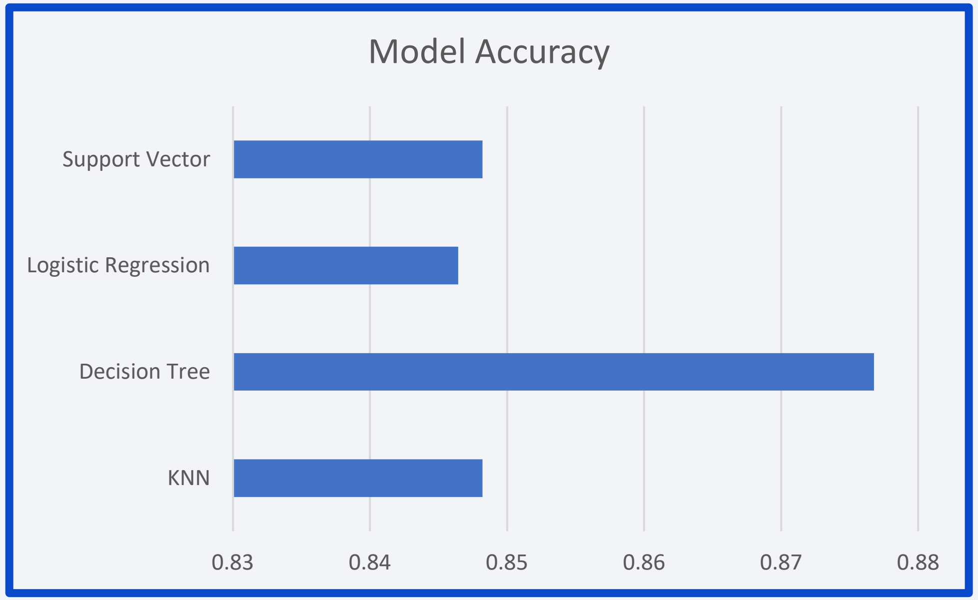


Success rate is higher in the 0-5000kg range vs the 5000-10000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Decision Tree model has the highest classification accuracy

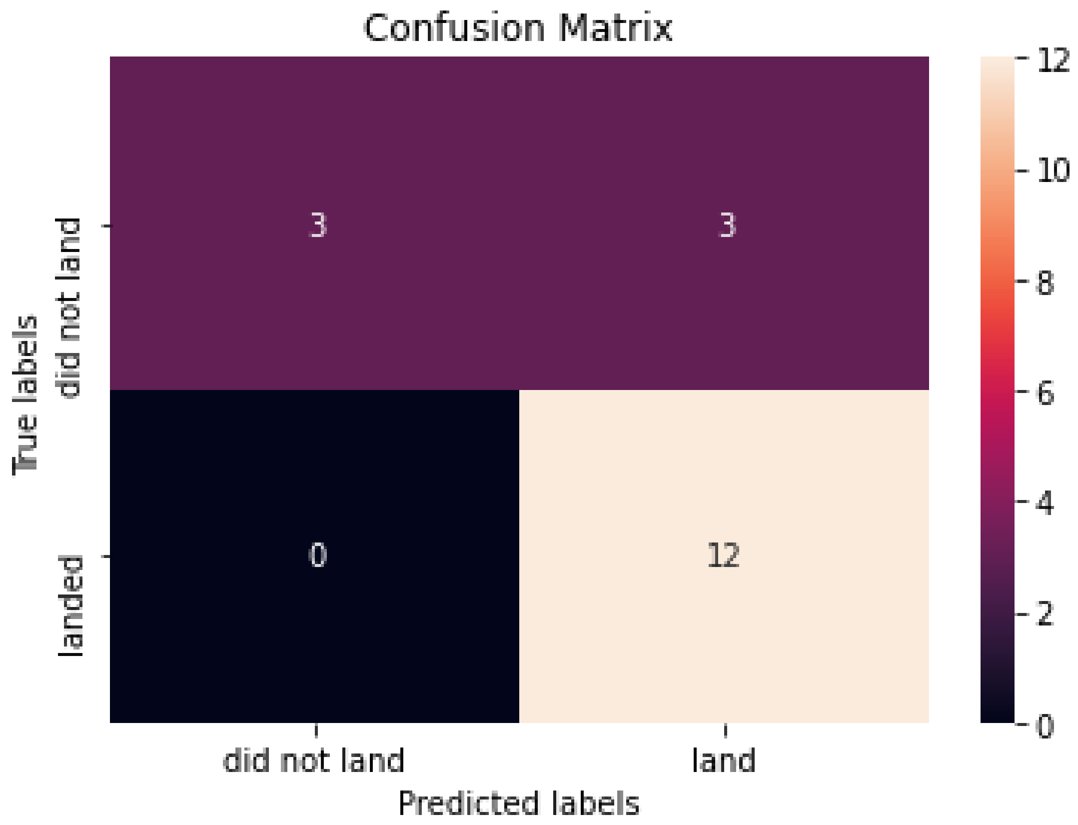
Find the method performs best:

In [32]:
```python
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 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5,
'splitter': 'best'}
```

Confusion Matrix – Decision Tree Model



The Decision Tree Model Confusion Matrix shows a high volume of True Negatives (12) correct predictions and no false negatives

Conclusions

- Orbit Types ES-L1, GEO, HEO & SSO have the best success rate
- With heavy payloads the successful landing rate is better for PO, LEO and ISS orbit types
- After 3 years of initial failure, the launch success rate has been improving since 2013
- Proximity to the coastline, and distance to railway, cities and highways is a key safety factor in launch site selection
- Launch site KSC LC-39A has the highest success count, CCAFS SLC-40 the lowest
- The Decision Tree model has the highest classification accuracy

Thank you!