IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

JP Berard
July 22, 2022

Outline

Executive Summary

Summary of methodologies

- Data was collected
- Data was explored, cleaned and visualized with static graphs and interactive visualizations

Summary of all results

- Once data was cleaned, multiple distinct factors were analyzed to determine which could help in predicting a successful first stage landing
- Multiple machine learning algorithms were applied and tested for its most effective in predicting a successful landing of the first stage.

Introduction

- Space X 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 Space X can reuse the first stage
- If we can predict if the first stage of the launch will land, we can predict how much money the launches cost.
- With that information, an alternative company can decide if they could put up a competitive bid against Space X

Section 1

Methodology

Methodology

Executive Summary

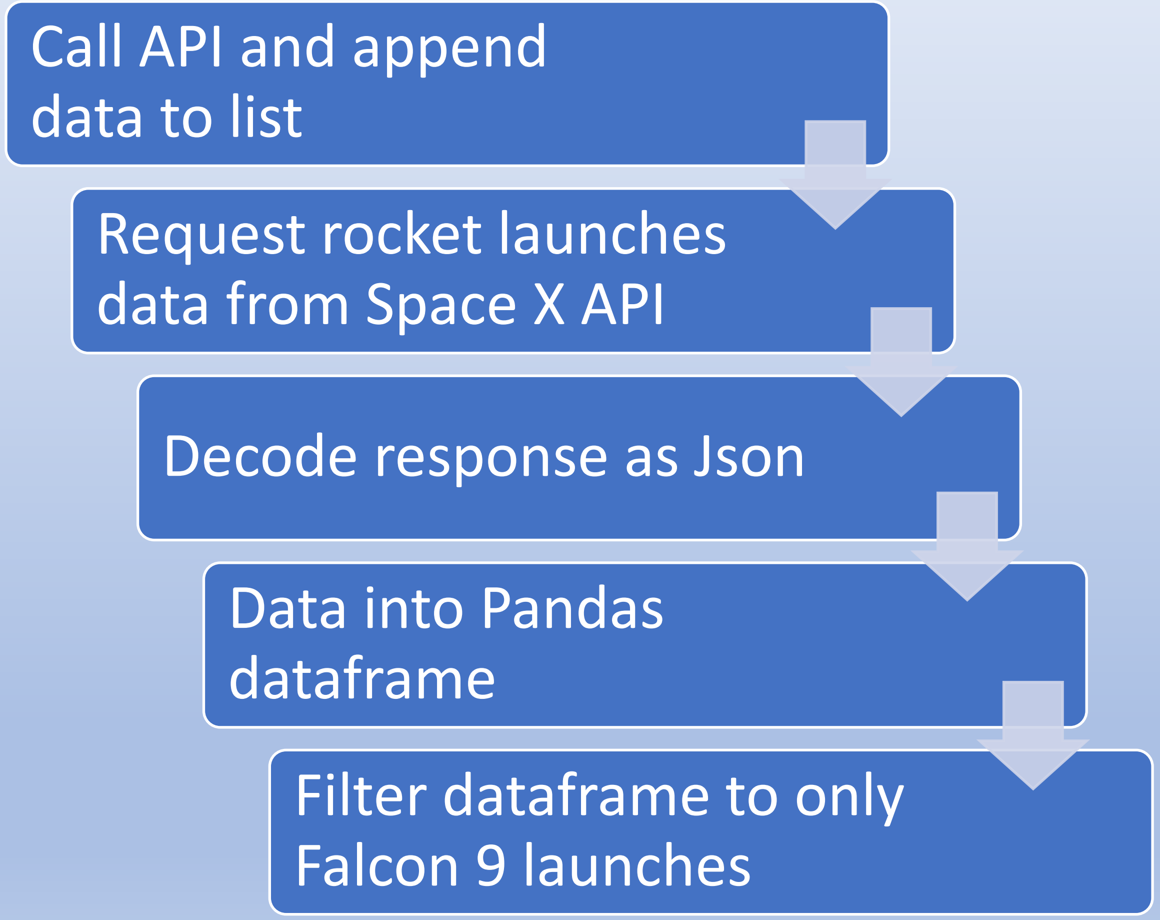
- Data collection methodology:
  - Data was collected through the Space X API and web scraping through Wikipedia
- Perform data wrangling
  - Data wrangling was performed to identify and missing or erroneous values and correct them
- Perform exploratory data analysis (EDA) using visualization and SQL
- 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 was collected using two methods
  - Data was acquired through the Space X API using Pandas, Numpy and Datetime.
    - Data was acquired and a Pandas Dataframe was created
  - Data was extracted from Wikipedia using web scraping
    - Created a BeautifulSoup and extracted all columns/variables from website

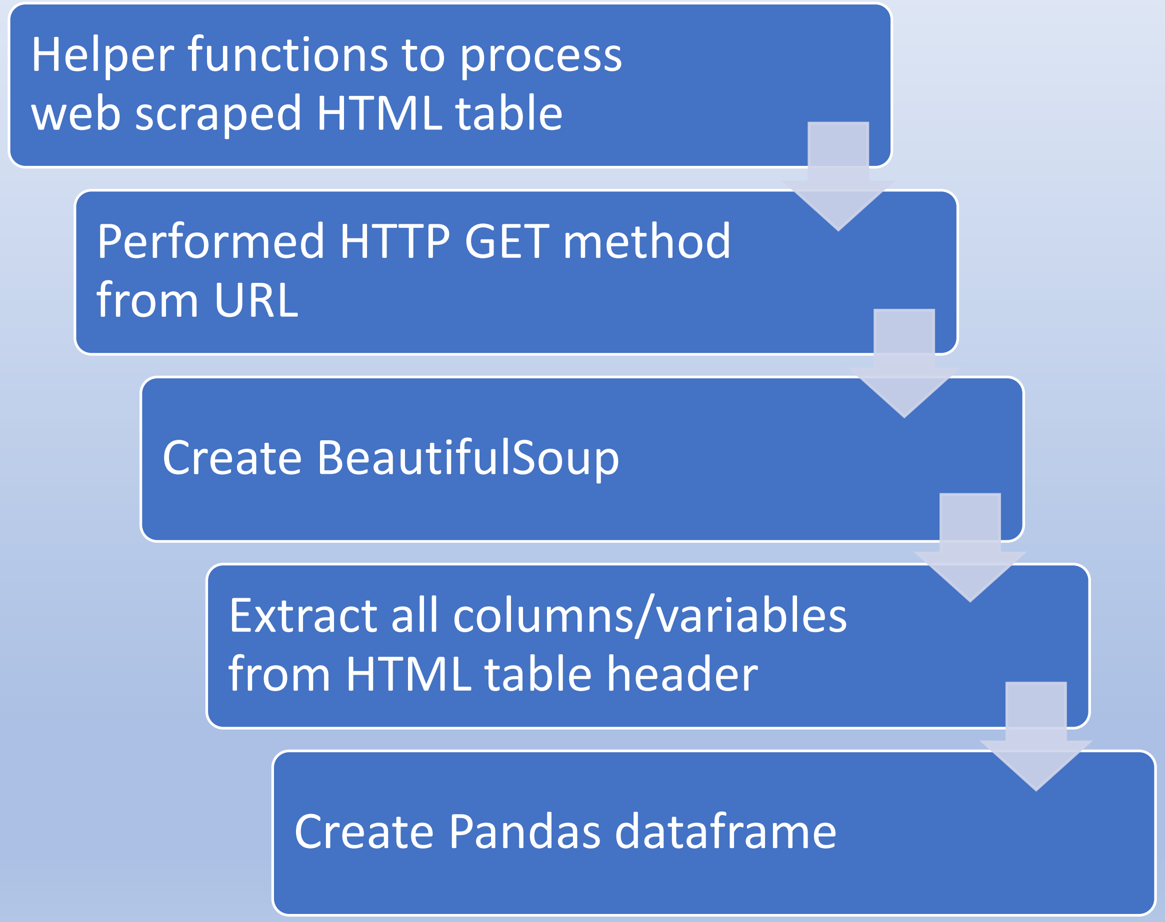
Data Collection – SpaceX API

- The following steps were taken to extract data from the Space X API
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/Capstone%20v1.ipynb



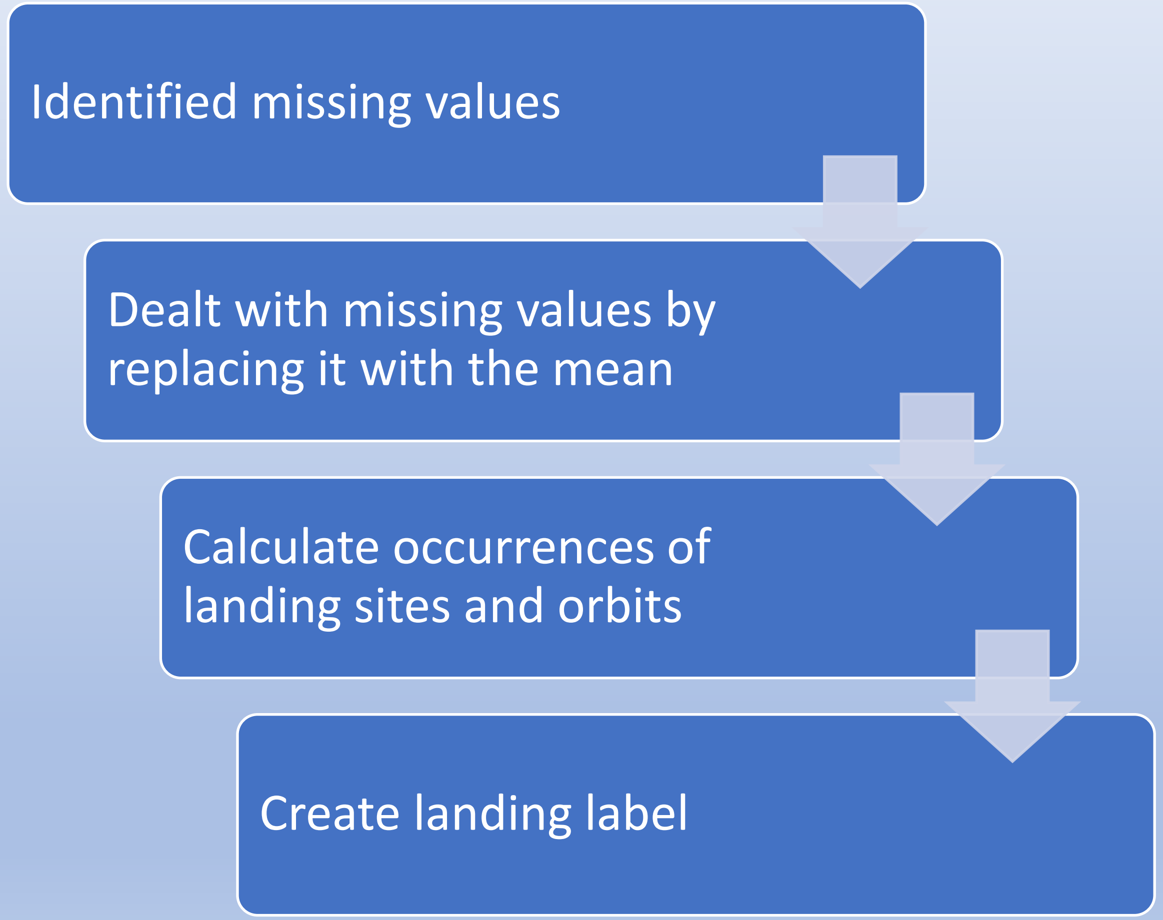
Data Collection - Scraping

- Data was obtained from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 and data was scraped
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/webscraping%20v1.ipynb



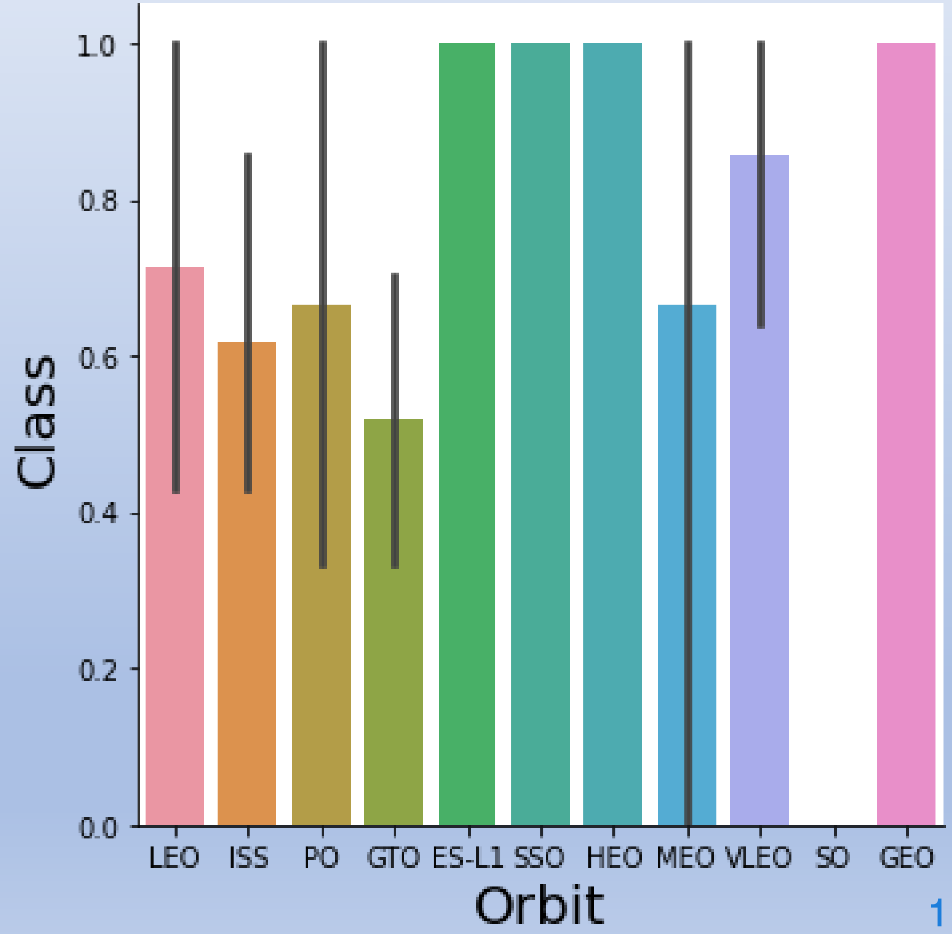
Data Wrangling

- Data wrangling was performed to transform data into a useable state where it could be explored and analyzed
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/Data%20wrangling%20v1.ipynb



EDA with Data Visualization

- Scatter plots, bar charts and line charts were used to visualize the data and analyze trends between different variables
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/EDA%20v2.ipynb



EDA with SQL

- The following queries were made using SQL:
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/EDA.ipynb

Selected distinct launch sites

Selected launch sites that began with ‘CCA’

Calculated total payload mass by NASA (CRS)

Calculated average payload mass carried by booster F9

Obtained date of first successful landing outcome on ground

List boosters that landed on drone ship between 4000-6000kg

Total number of successful and failure mission outcomes

List names of boosters that carried maximum capacity

List boosters that have had failed landings on drone ships in 2015

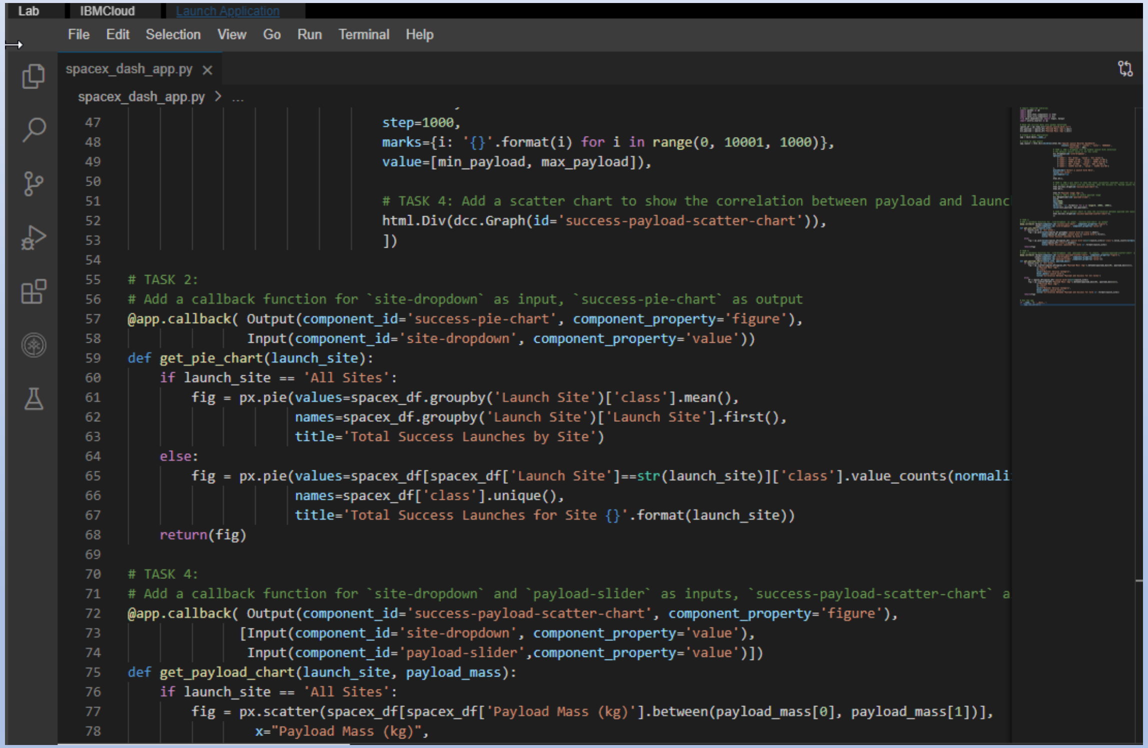
Rank landing outcomes between 2010-2017

Build an Interactive Map with Folium

- Map objects such as circles, markers and lines were used on a Folium Map
- Those objects were used to draw attention and reveal information related to the data such as sites and distances between two points
- Click on the following work to check out the Jupyter notebook:
- https://github.com/jpberard85/Master/blob/main/Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash

- A scatter plot and pie chart were created to display the data in an interactive form. To control these plots, a drop down menu as well as a range slider was used to control parameters
- These plots were created to better visualize the data and let the user control the parameters to infer their own analysis
- ***The external tool would not allow me to save link to GitHub



Predictive Analysis (Classification)

- 4 different models were used to compare for the most accurate
  - Logistcal regression
  - Support vector machine
  - Decision Tree
  - K nearest neighbor
- Click on the following work to check out the Jupyter notebook
- https://github.com/jpberard85/Master/blob/main/Machine%20Learning%20Prediction.ipynb

Create Numpy array, standardize and split into training/testing data

Create logistical regression, SVM, decision tree and KNN models

Calculate accuracy of each model

Find method that performs the best

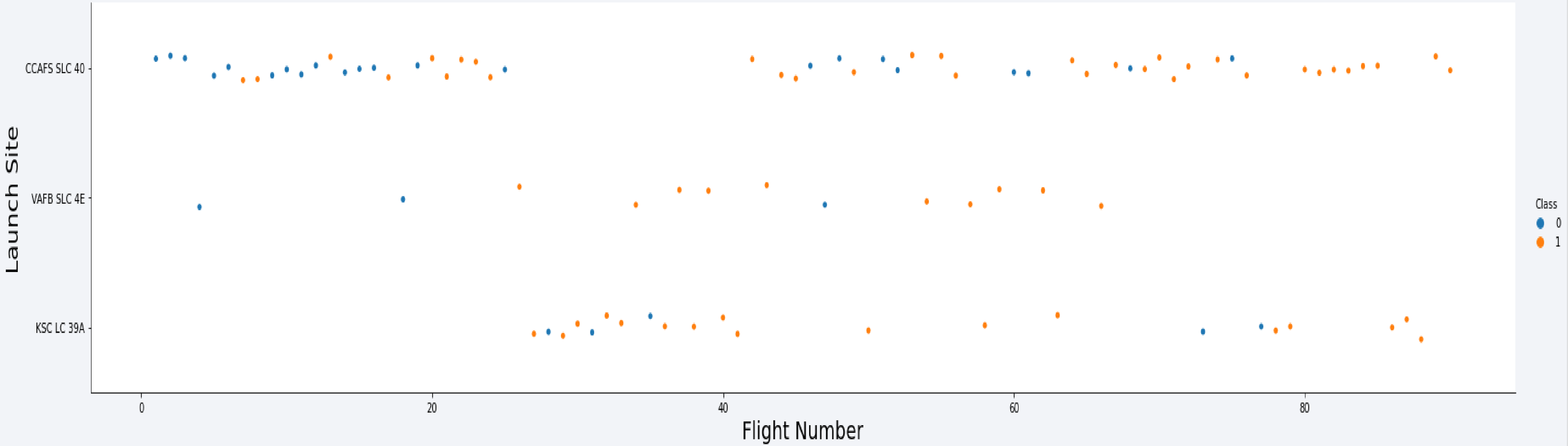
Results

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

Section 2

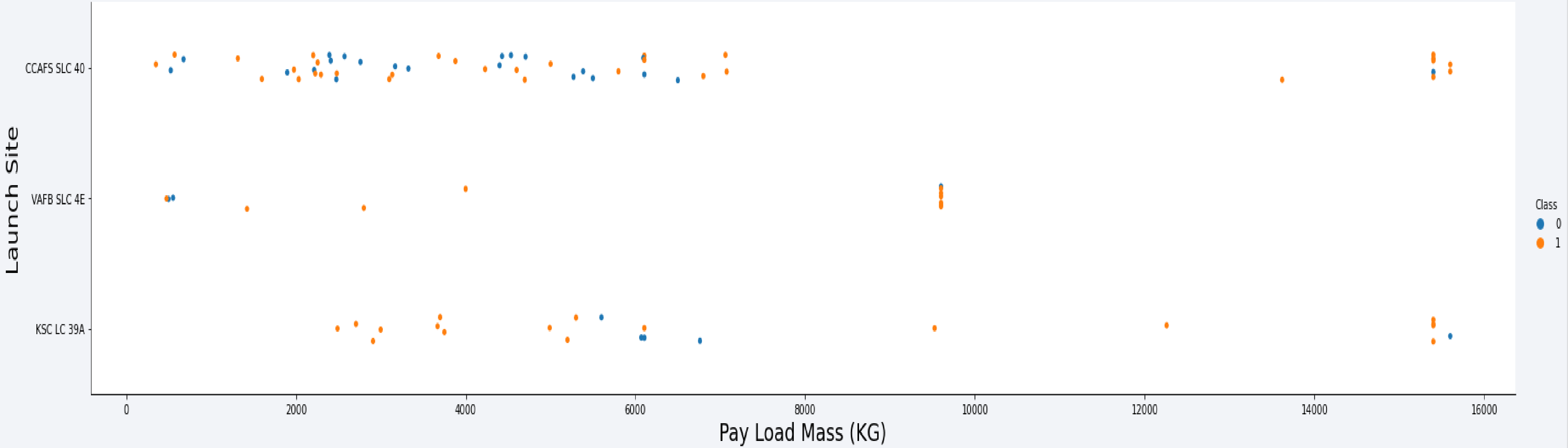
Insights drawn from EDA

Flight Number vs. Launch Site



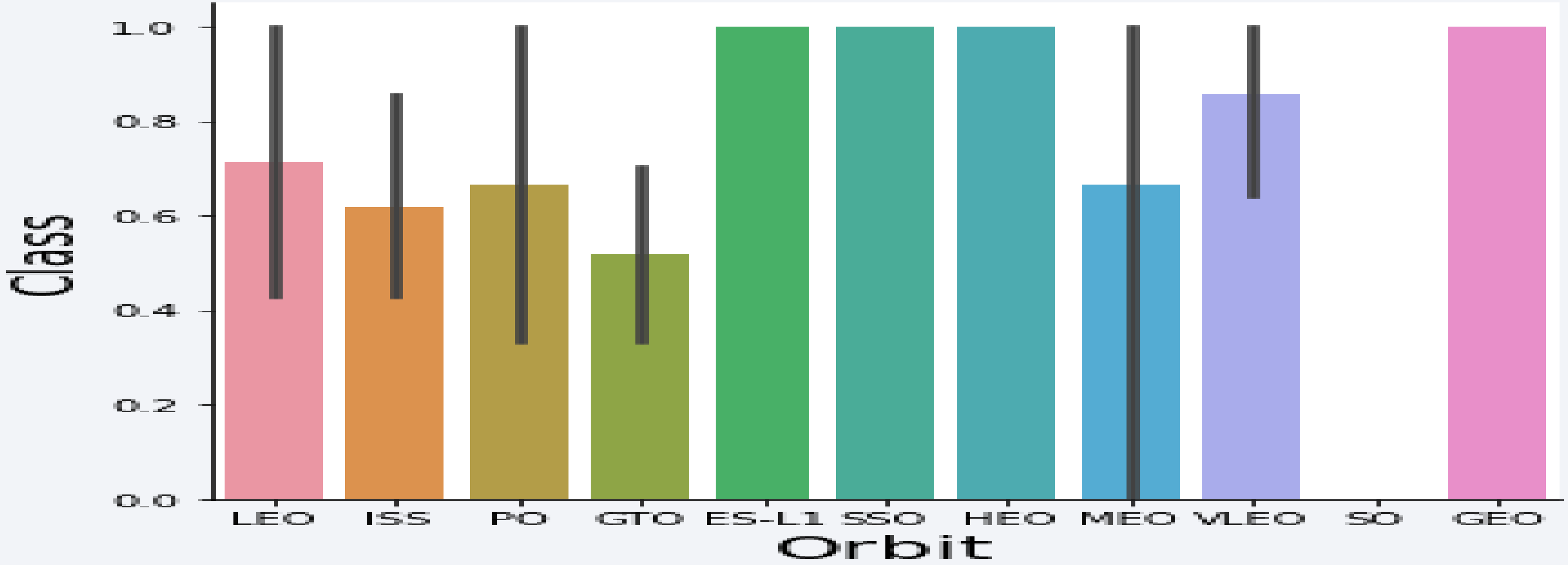
- Increasing flight numbers is correlated to an increase in successful missions.
- We can also appreciate which sites are the most frequently used (CCAFS SLC 40)

Payload vs. Launch Site



- Missions with heavier payloads appear to have a higher success rate
- Launch site VAFB SLC 4E does not have any payloads greater than 10000kg

Success Rate vs. Orbit Type



- We can see that ES-L1, SSO, HEO and GEO have the highest success rate

Flight Number vs. Orbit Type

In [7]:
```
# HINT use groupby method on Orbit column and get the mean of Class column
mean=df.groupby(['Orbit']).mean()
mean
```

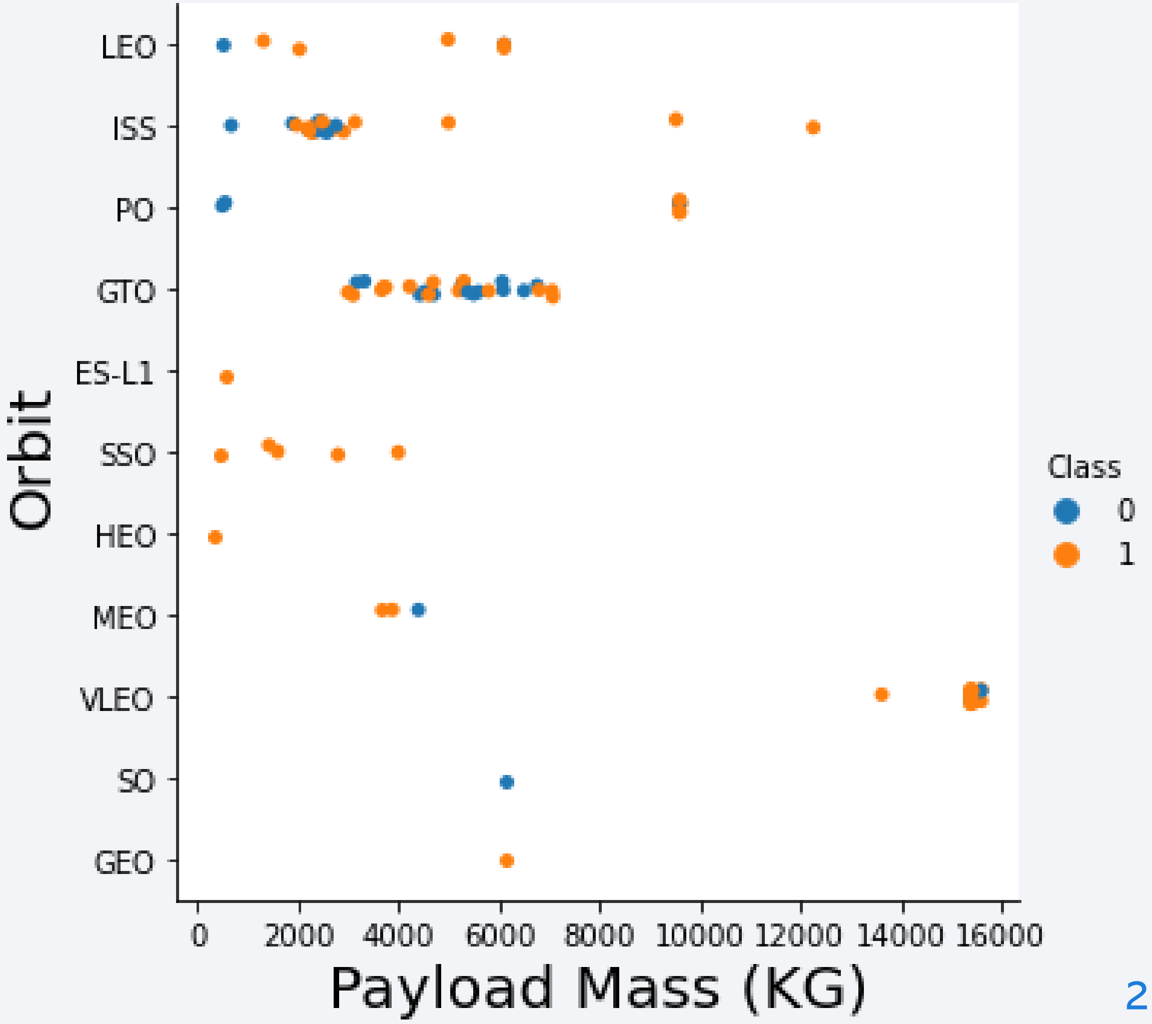
Out[7]:

Orbit	FlightNumber	P[No Title]ss	Flights	GridFins	Reused	Legs	Block	ReusedCount	Longitude	Latitude	Class
ES-L1	13.000000	570.000000	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	-80.577366	28.561857	1.000000
GEO	83.000000	6104.959412	2.000000	1.000000	1.000000	1.000000	5.000000	2.000000	-80.577366	28.561857	1.000000
GTO	35.037037	5011.994444	1.407407	0.629630	0.333333	0.629630	3.037037	0.962963	-80.586229	28.577258	0.518519
HEO	49.000000	350.000000	1.000000	1.000000	0.000000	1.000000	4.000000	1.000000	-80.577366	28.561857	1.000000
ISS	39.142857	3279.938095	1.238095	0.809524	0.238095	0.857143	3.142857	1.285714	-80.583697	28.572857	0.619048
LEO	20.000000	3882.839748	1.000000	0.571429	0.000000	0.714286	2.142857	0.428571	-80.584963	28.575058	0.714286
MEO	77.666667	3987.000000	1.000000	0.666667	0.000000	0.666667	5.000000	0.666667	-80.577366	28.561857	0.666667
PO	36.333333	7583.666667	1.333333	0.888889	0.333333	0.777778	3.222222	1.555556	-120.610829	34.632093	0.666667
SO	73.000000	6104.959412	4.000000	0.000000	1.000000	0.000000	5.000000	3.000000	-80.603956	28.608058	0.000000
SSO	60.800000	2060.000000	2.400000	1.000000	0.800000	1.000000	4.600000	3.200000	-112.604136	33.418046	1.000000
VLEO	78.928571	15315.714286	3.928571	1.000000	1.000000	1.000000	5.000000	3.928571	-80.586862	28.578358	0.857143

- Above is a portion of the data obtained for the bar graph. Please see appendix.

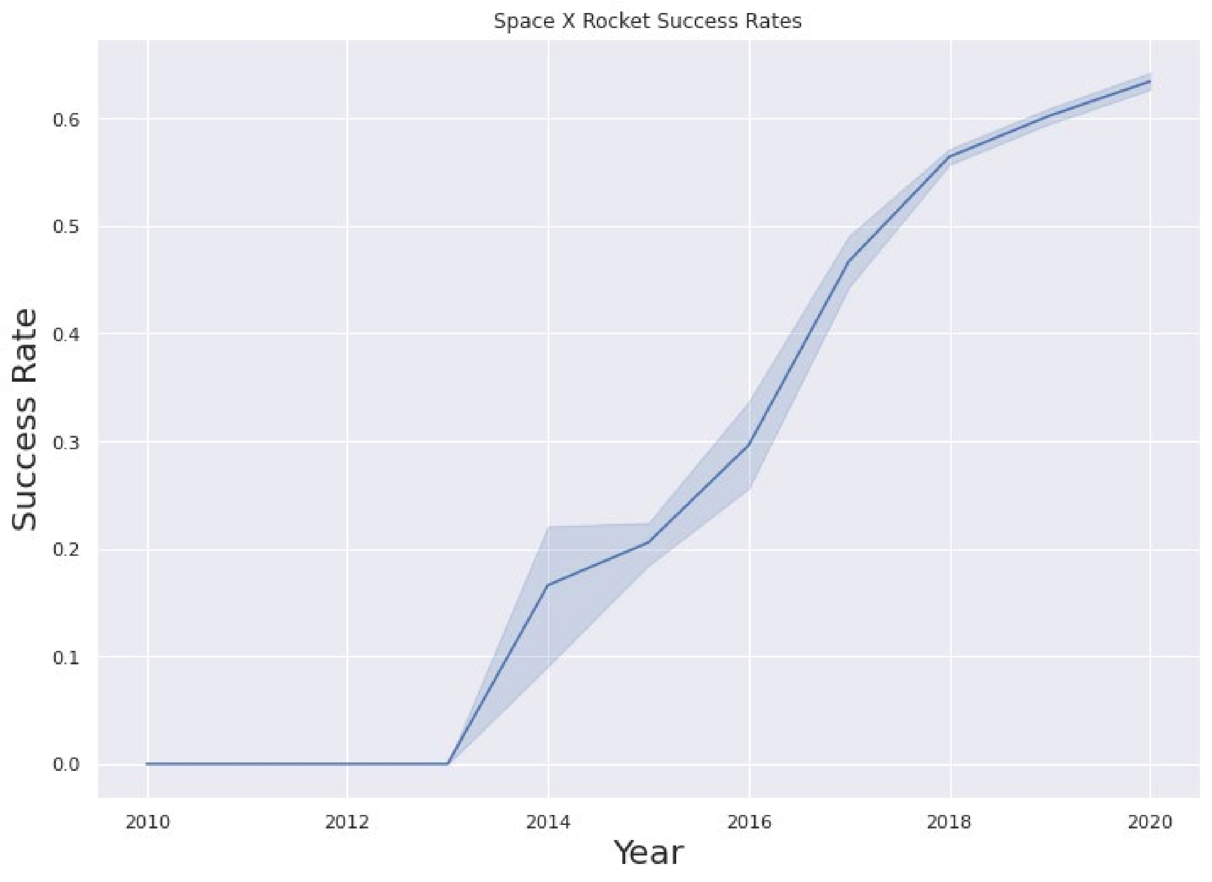
Payload vs. Orbit Type

- Some orbit types are more frequently used such as GTO and ISS
- There is a varying degree of success for the most common types of orbit types used



Launch Success Yearly Trend

- Since 2013, the success rate has continued to increase over time
- 2018 shows an inflection point that may suggest a plateauing of success rate



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [14]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[14]:

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

- Space X uses 4 launch sites. This information was obtained through a SQL query.

Launch Site Names Begin with 'CCA'

Task 2

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

In [15]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[15]:

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

- Above are 5 records with a launch site that begins with “CCA” obtained by a SQL query

Total Payload Mass

Task 3

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

In [19]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[19]:

Total payload mass by NASA (CRS)
45596

- Above is the total payload mass by NASA (CRS) obtained by a SQL query

Average Payload Mass by F9 v1.1

Task 4

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

In [20]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[20]:

Average payload mass by Booster Version F9 v1.1
2928

- Above is the result of a SQL query returning the average payload mass by Booster Version F9 v1.1

First Successful Ground Landing Date

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [21]:
```
%sql SELECT MIN(DATE) AS "Date of first successful landing outcome in ground pad" FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[21]:

Date of first successful landing outcome in ground pad
2015-12-22

- 2015-12-22 was the date of the first successful landing by Space X
- This was obtained using a SQL query

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

In [ ]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AN
D 6000;
```

- This query would yield the number of successful and failures of missions. It appears the query was not run. See Appendix

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [22]:
```
%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM SPAC
EXTBL WHERE MISSION_OUTCOME LIKE 'Success%'
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
(ibm_db_dbi.ProgrammingError) ibm_db_dbi::ProgrammingError: SQLNumResultCols failed: [IBM][CLI Driver][DB2/LINUXX8664] SQL0104N
An unexpected token "END-OF-STATEMENT" was found following "COME LIKE \'Success%\'".  Expected tokens may include:  "ESCAPE <es
cape_char>".  SQLSTATE=42601 SQLCODE=-104
[SQL: SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM SP
ACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%']
(Background on this error at: http://sqlalche.me/e/f405)
```

- This query returned an error.

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://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
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

- Above is the list of booster versions that have carried the maximum payload mass

2015 Launch Records

Task 9

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

In [24]:

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE year(DATE) = '2015' AND LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[24]:

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

- Only 2 results returned for the above SQL query determining how many failed landing outcomes in 2015 with their associated details

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

Task 10

Rank 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

In [25]:

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

```
 * ibm_db_sa://mnq88798:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[25]:

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

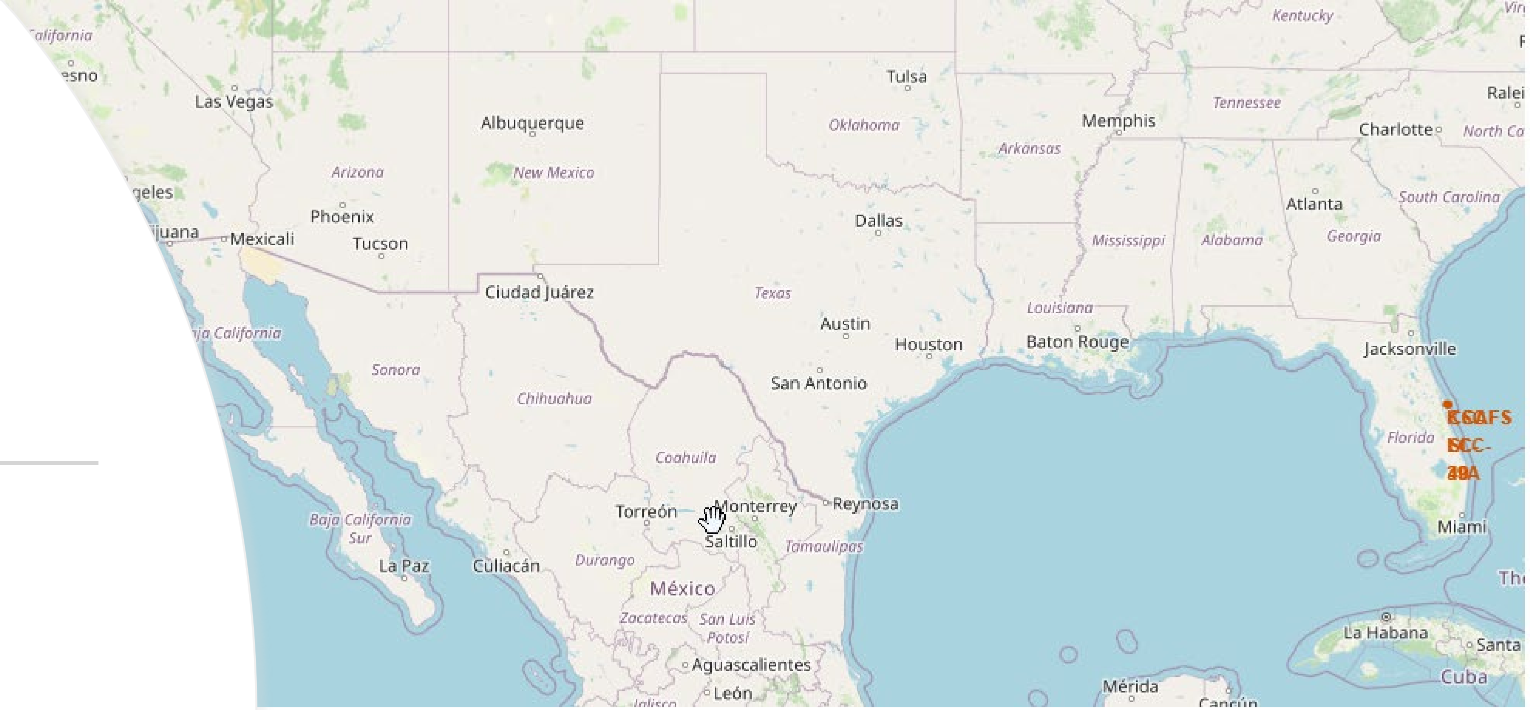
- In ascending order is the count of landing outcomes between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

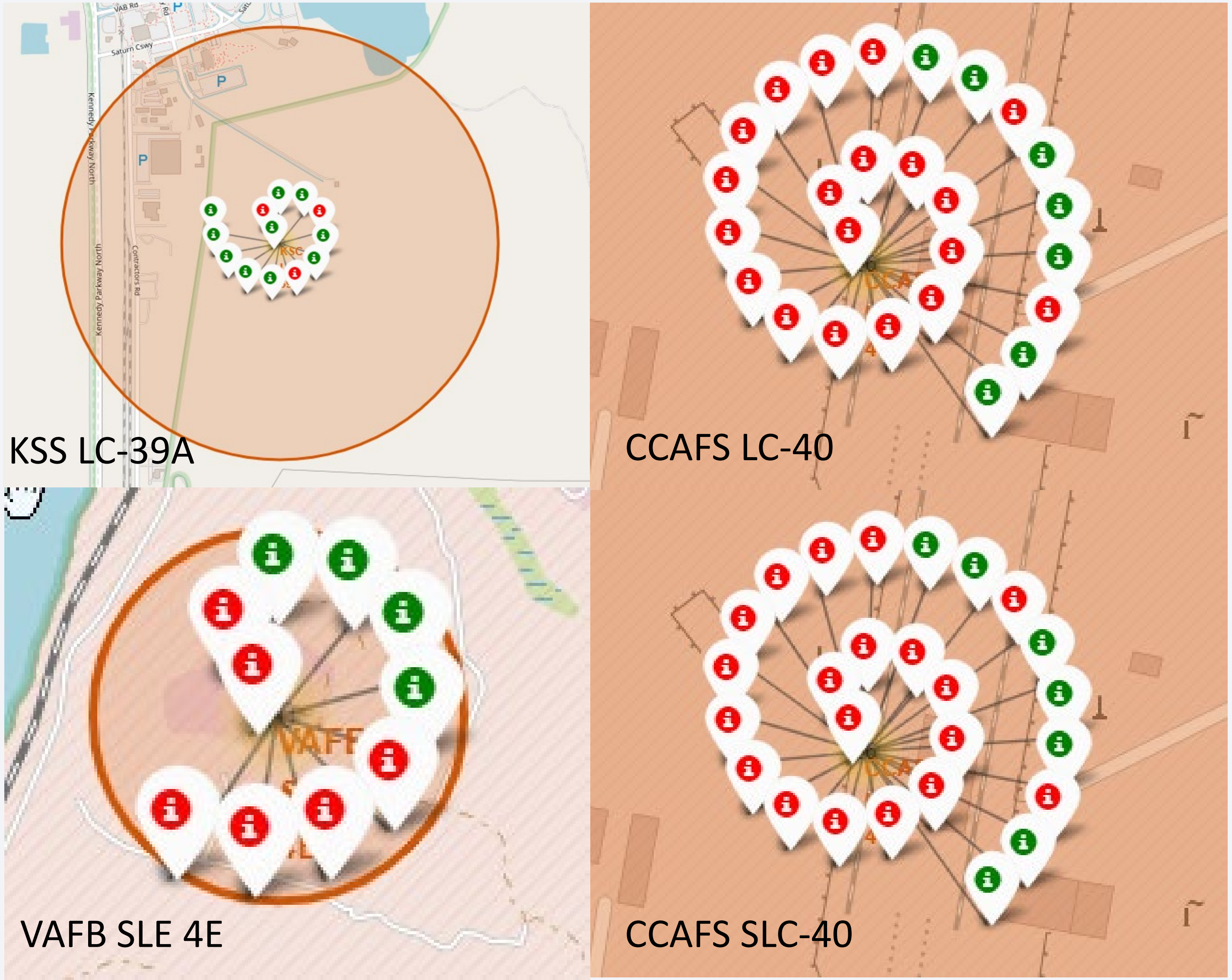
Space X Launch Sites

- The Space X Launch Sites are located in California and Florida and are on the coastal regions of those states
- This Map allows you zoom in to better appreciate the detail of the site

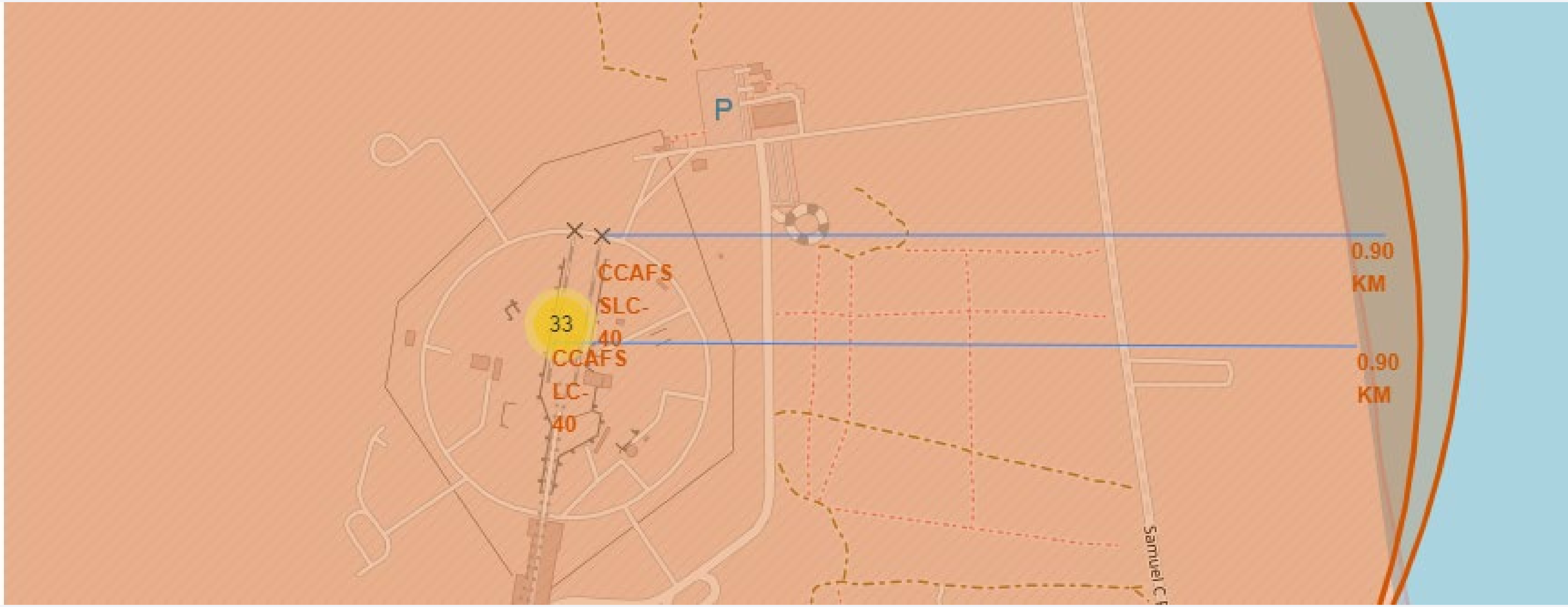


Successful Launches at each site

- All 4 sites are demonstrated at varying levels of zoom on the interactive map
- Green icons show a successful landing while Red icons demonstrate and unsuccessful landing
- Clicking on the site shows the entire marker cluster



Launch site to surrounding landmark

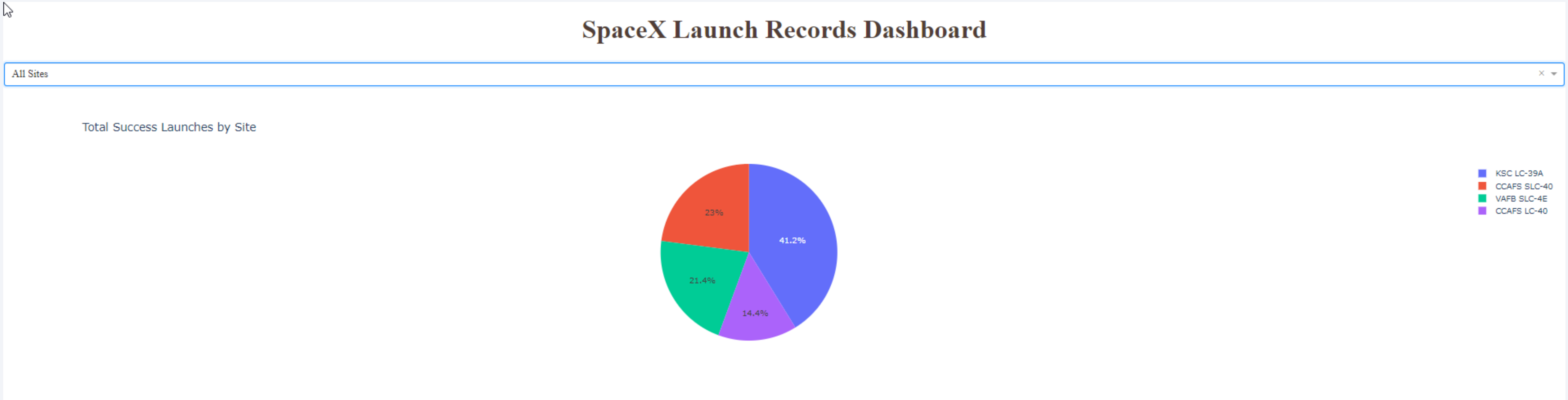


- The distance from CCAFS-LC 40 to the ocean is 0.9km.
- The proximity to the coast seems important for launches to mitigate aborted and failed launches

Section 4

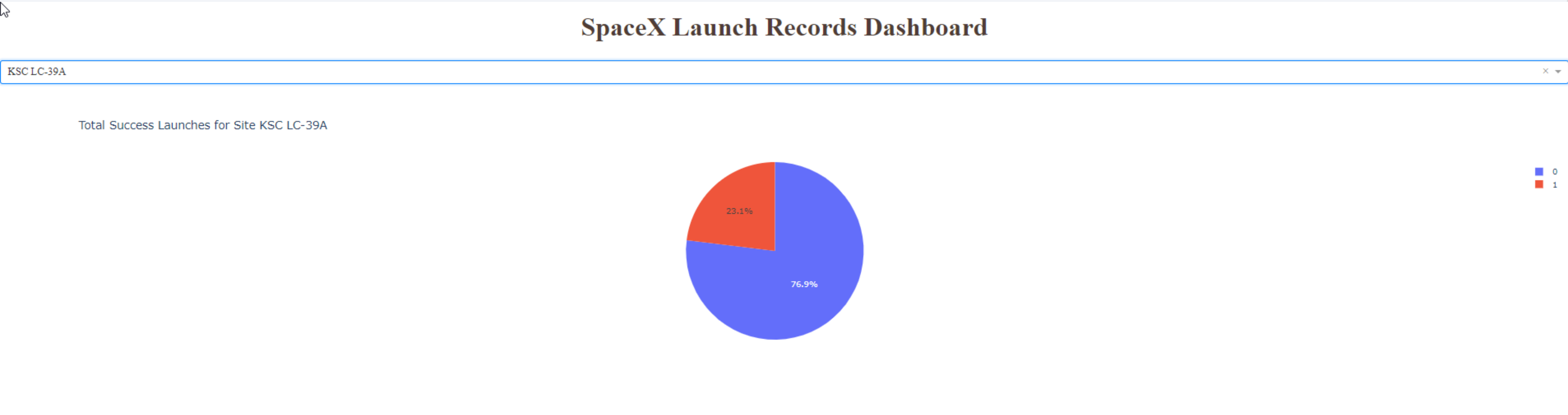
Build a Dashboard with Plotly Dash

Successful Launches per Site



- This chart shows that KSC LC-39A has the most successful launches

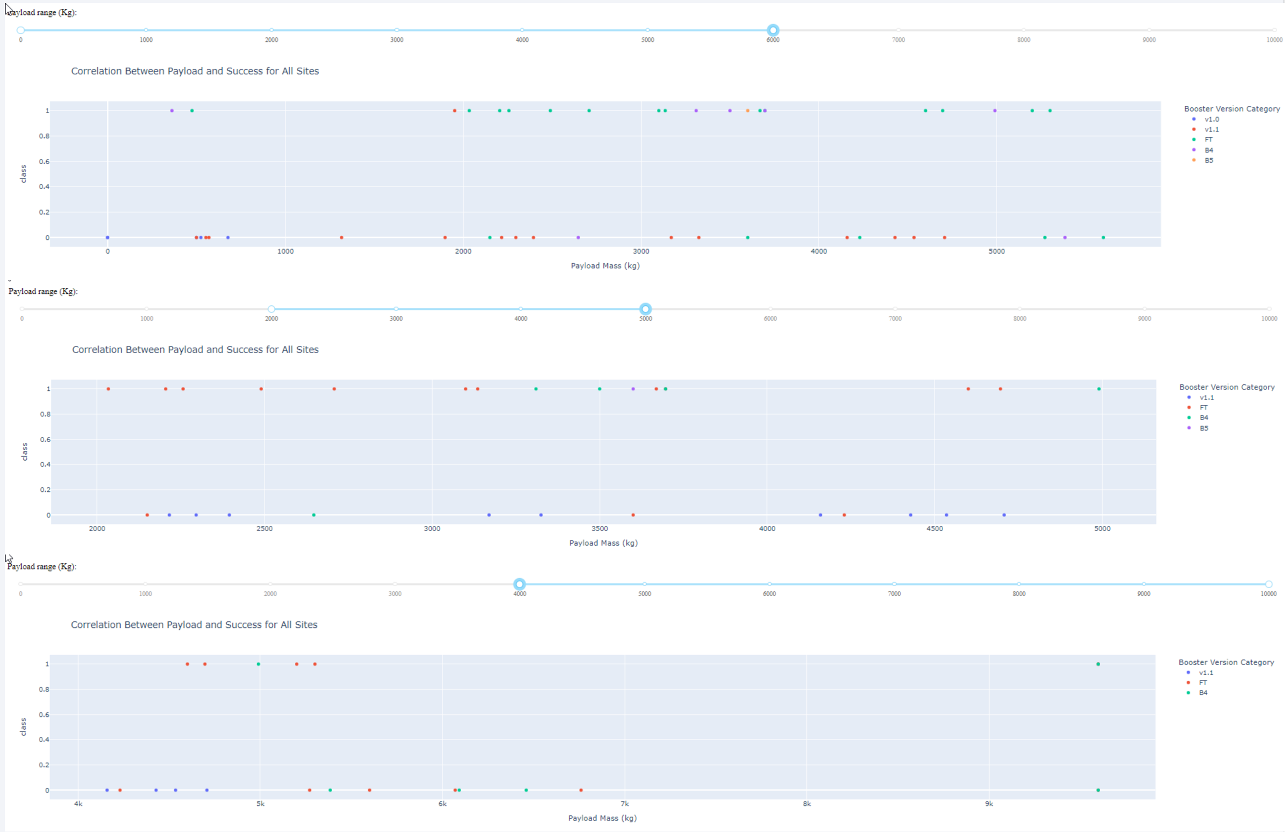
Most Successful Site



- KSC LC-39A is the most successful launch site with a 76.9% success rate.

Payload and Success Rates

- 3 screen shots from the dashboard are attached
- Booster version FT appears to be the most successful.
- The range that seem to have the most success is between 2000kg and 6000kg. It does have its fair share of unsuccessful landings as well suggesting it is the most common payload launched



Section 5

Predictive Analysis (Classification)

Classification Accuracy

TASK 12

Find the method performs best:

In [66]:
```
#All the accuracies are the same. The scores were different however
ml_scores = {'Logistic regression': [logreg_cv.best_score_], 'SVM': [svm_cv.best_score_], 'Decision tree': [tree_cv.best_score
_], 'K Nearest Neighbors': [knn_cv.best_score_]}
df1 = pd.DataFrame.from_dict(ml_scores, orient='index', columns=['Best scores'])
df1
```

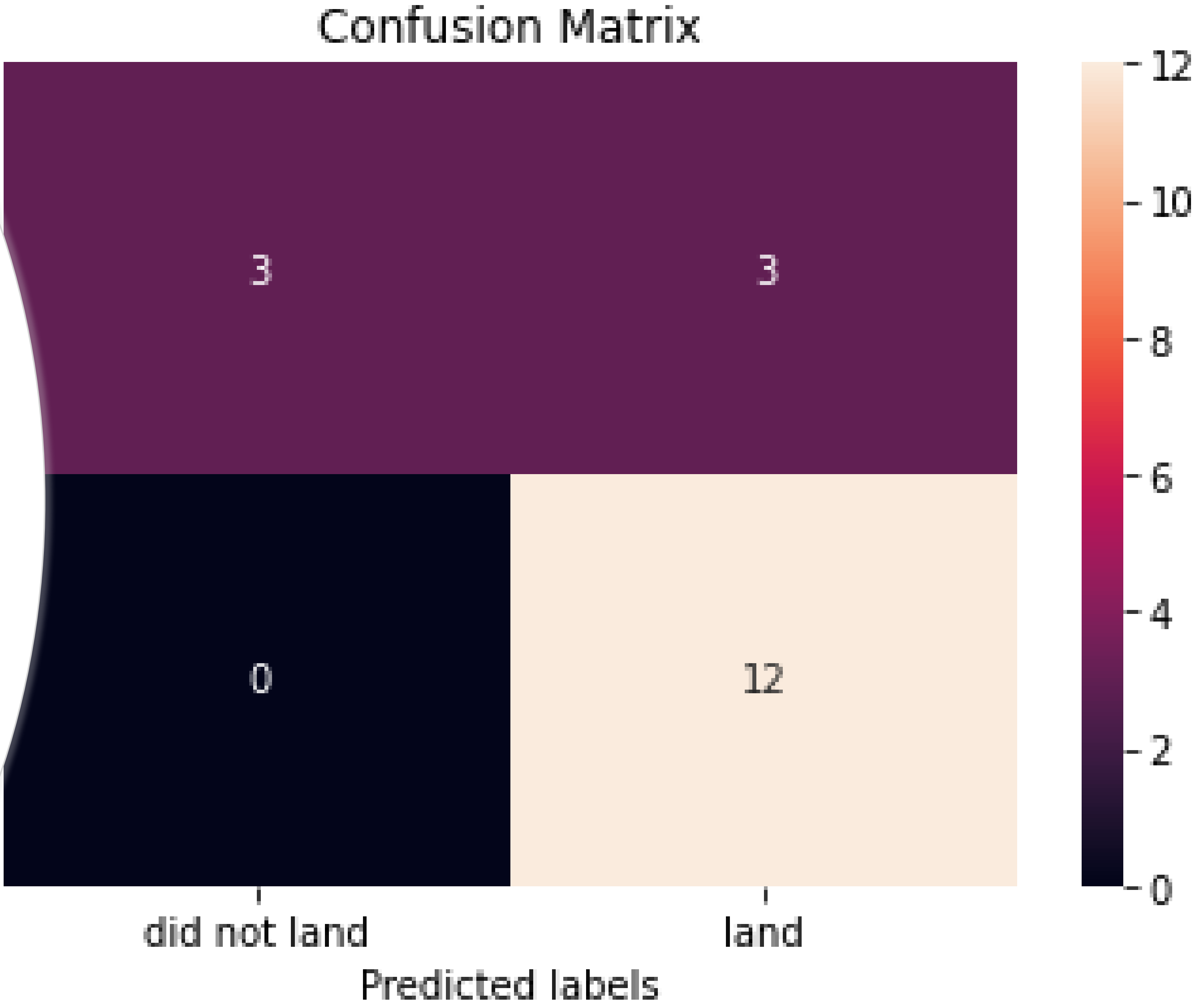
Out[66]:

	Best scores
Logistic regression	0.846429
SVM	0.848214
Decision tree	0.889286
K Nearest Neighbors	0.848214

- All model accuracies were the same however the Decision Tree did score the best compared to the other 3 models tested
- Please see appendix

Confusion Matrix

- This is the confusion matrix of the Decision Tree
- It produced:
  - 12 correct positives
  - 3 false positives
  - 3 correct negatives
  - 0 false negatives



Conclusions

- Over time, Space X has been more successful and are doing so at an accelerated rate but potentially plateauing
- KSC LC-39A is the site that yields the most successful launches
- Boosters with payloads between 2000kg-6000kg are the most frequent and most successful
- The Decision Tree algorithm has the best results for prediction of successful launches
- 4 orbit types have the most successful landings: ES-L1, SSO, HEO and GEO

Appendix

- Flight Number vs. Orbit Type: Only realized that I only completed a portion of the work. I have run out of free computing to correct this
- Classification Accuracy Slide: Was unaware that a bar graph needed to be produced. The scores were produced however I have run out of free computing to correct this
- Successful Drone Ship Landing with Payload between 4000 and 6000: It appears I forgot to run the code. I have fun out of free computing to correct this

Thank you!