IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dr Bekezela Bobbie Khabo
16 August 2023

Outline

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

Executive Summary- Methods

- **Business Understanding**: a binary Classification challenge- predicting whether Falcon9 rockets can land successfully or not
- **Data Understanding**: revealed need for multiple data sources, therefore compiled a dataset from,
    - SpaceX Rest API
    - Webscraping the Falcon rocket Wikipedia page
- **Data Preparation**: used several Python data manipulation and visualisation libraries, along with SQL to manipulate the data and ensure a clean meaningful dataset stored in a database
- **Exploratory Data Analysis**: carried out using Pandas and Matplotlib, and Seaborn to identify feature variables
- **Key outputs**:
    - Developed an interactive leaflet map with Folium, and an interactive dashboard with Plotly Dash
    - Predictive model for Falcon 9 Landings using K-Nearest neighbour
- **Outstanding:**
    - Evaluation with stakeholders
    - Model deployment
    - Further refinement

Introduction

- **<u>Business Understanding:</u>**
- Space travel is an expensive exercise, with each launch costing USD62m by SpaceX, and almost three times as much from other vendors.
- Predicting whether a SpaceX Falcon9 rocket will land is crucial to deciding a vendor for a rocket launch
- A successful prediction would result in savings of up to tens of millions of dollars if the decision is correct
- Question: Can publicly available data of historic SpaceX launches be used to develop a Machine Learning model to predict whether Falcon9 launches land successfully or not?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Compiling a dataset from,
    - SpaceX Rest API
    - Complemented with data from webscraping SpaceX Wikipedia page
      - Using BeautifulSoup method
- Perform data wrangling
  - Payload mass had five missing values: replaced with the mean
- 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

- SpaceX REST API used to gather some of the data
  - Gathered data includes Booster version, longitude and latitudes of launch sites, payload, cores, light number, and date
  - Filtered out other Falcon versions to leave only Falcon9
- Webscraping of the Falcon9 Wikipedia page (link here) used to extract critical tables for the final dataframe
  - Data points under this included: Flight number, Date, Time, Booster Version, Launch site, Payload, Payload mass, Orbit, Customer, Launch Outcome, and Booster Landing
  - Methods used: Get requests, BeautifulSoup, and html.parser

Data Collection – SpaceX API

- After steps in flowchart on right, then filtered out other Falcon versions to leave only Falcon9
- The GitHub URL of the completed SpaceX API calls notebook is here

1st API call to rockets: For Booster Version names

2nd API Call to launchpads: For launch sites along with their longitudes & latitudes

3rd API Call to Payloads: for payload masses and orbits

4th API Call to Cores: for landing outcome, info on cores, gridfins, legs, and landing pad

Data parsed in a Dictionary with column names as keys, and data as values

Data Collection - Scraping

Sent a Get Request to Falcon9 Wiki page

Created a BeautifulSoup Object

Used findall() to search for all tables on page

Iterated through column names

Appended empty table with contents, and then cast it into a dataframe

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

Data Wrangling

Checked for missing data under each variable and expressed as a %

Checked data types of each variable using method dtypes to prevent issues later in analysis and modeling

Conducted value_counts() on key variables

Used "Outcome" variable to add new column "Class" depending on whether successful landing or not

- The GitHub URL of the completed data wrangling related notebook is [here]

EDA with Data Visualization

- Used Pandas, Matplotlib and Seaborn to plot
  - Scatterplot of Flight number, Payload mass, and Class
    - To assess whether progressive flights succeed or not, and effect of payload mass on safe return
  - Scatterplot of Flight number, Launch site, and Class
    - To check for striking data points on launch site, and flight number and their effect on class
  - Bar chart showing success rate of each type of orbit
    - To see if any orbits have outstanding outcomes compared to others
  - Scatterplot of Orbit, Flight number, and Class
    - To see if any orbits have striking relationships with flight number and class
  - Scatterplot of Orbit, Payload mass, and Class
    - Assess relationship between Orbit, payload mass and success rate
  - Line plot of mean annual success rate
    - To show whether the program has been improving with time

GitHub URL- here

EDA with SQL

- A query to identify unique site used for launch by Spacex
- Showed five records from launch sites with names containing string “CCA”
- Displayed total payload mass carried by boosters launched by NASA (CRS)
- Showed the mean payload mass carried by booster version F9 v1.1
- Listed 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 greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- Used a subquery to list the names of the Booster versions which have carried the maximum payload mass
- Used Substr() to List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- 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.

GitHub URL here

Build an Interactive Map with Folium

- Created circles and markers for each of the four launch sites
- Added additional markers to each launch site in which
  - Green markers for successful launches
  - Red markers for unsuccessful launches
- Put a marker on nearest coastline, and calculated distance to a launch site using the geometry of the earth(assuming the earth's radius is 6373km)
- Drew a Polyline between the coastline above, and the launch site
- Put markers on, and then calculated distances to the following objects,
  - Highway (very close, for logistical ease)
  - Railroad (very close for logistical ease)
  - City (further away, for public safety)

- The GitHub URL of completed interactive map with Folium map is here

Build a Dashboard with Plotly Dash

- Added,
  - a “Launch Site” Drop-down Input Component
  - a callback function to render ”Success Pie-chart” based on selected site dropdown
  - a Range Slider to select “Payload”
  - a callback function to render the ”Success Payload Scatter-chart” scatter plot
- These will assist the business colleagues to make informed decisions using an interactive platform to gain insights

The GitHub URL of the completed Plotly Dash lab is here

Predictive Analysis (Classification)

Created a Numpy array for the target Y from "Class" variable

Used StandardScaler to standardize the data of the Features (X)

Did Train-Test split using 80:20

For each approach,

1. Identified best parameters

2. Assessed them accuracy using test data

3. Plotted Confusion Matrix

Identified best-performing approach

Approaches used,

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree Classifier
4. K-Nearest Neighbour

- The GitHub URL of the completed predictive analysis lab is here

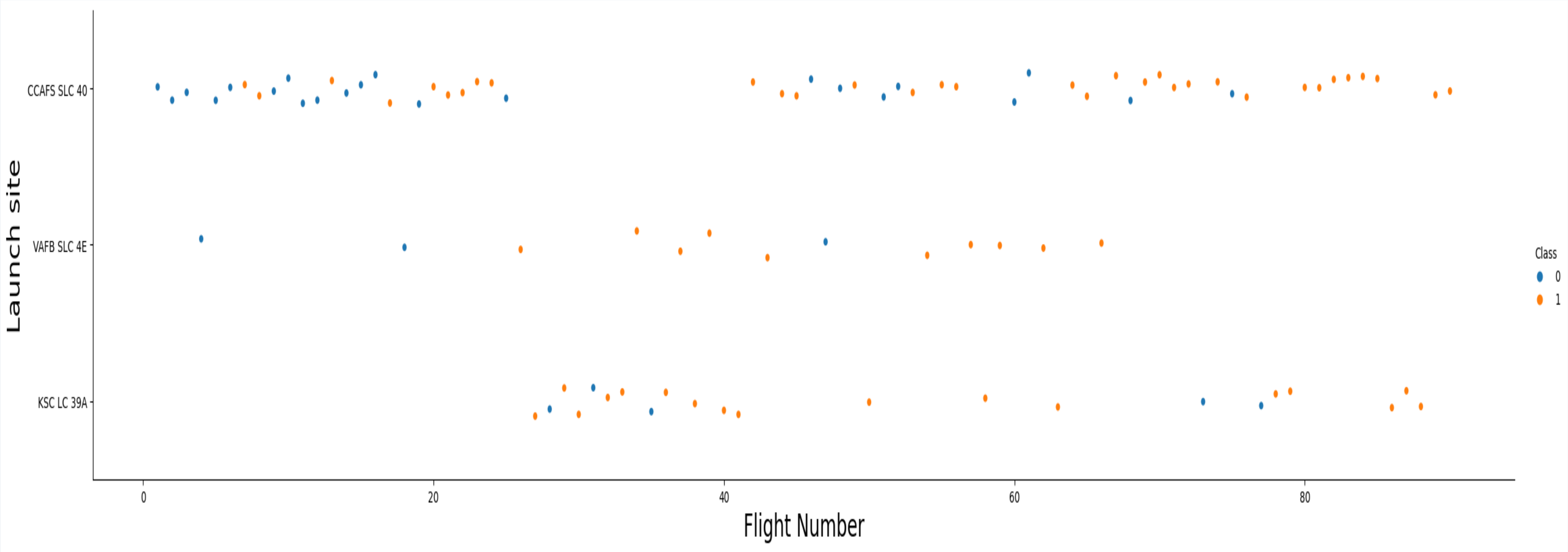
Results

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

Section 2

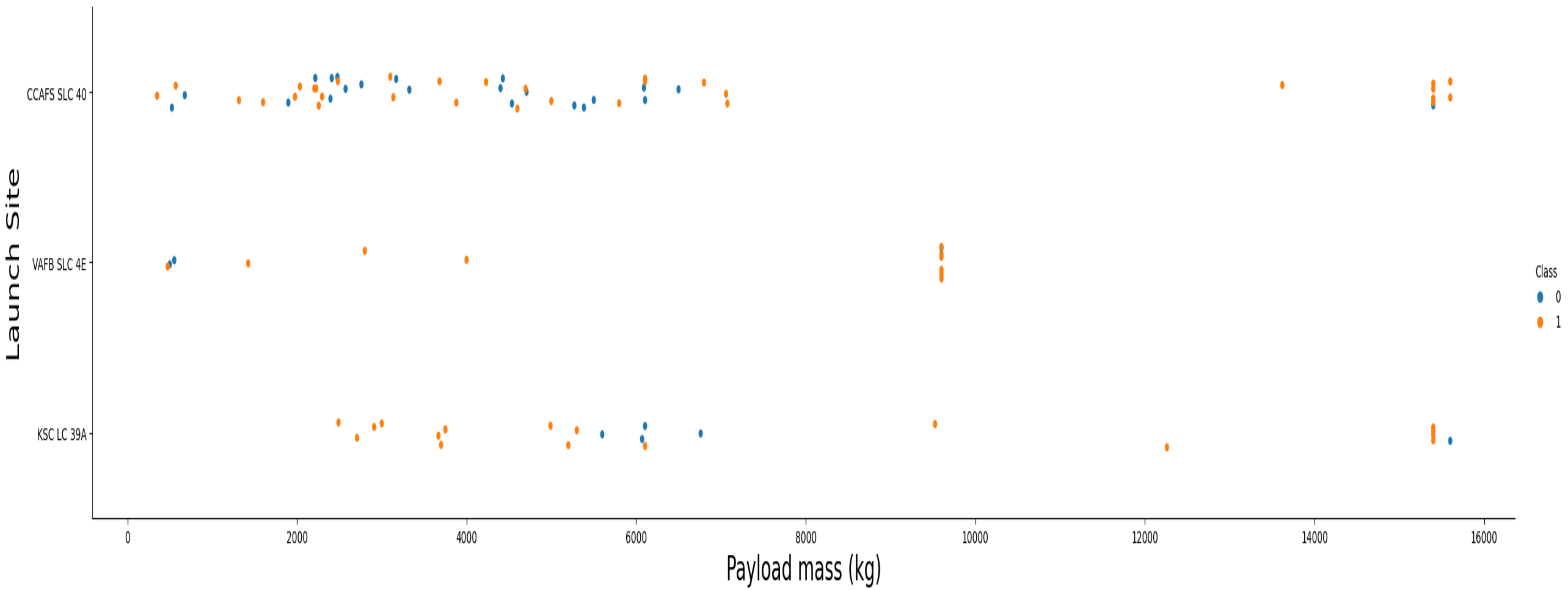
Insights drawn from EDA

Flight Number vs. Launch Site



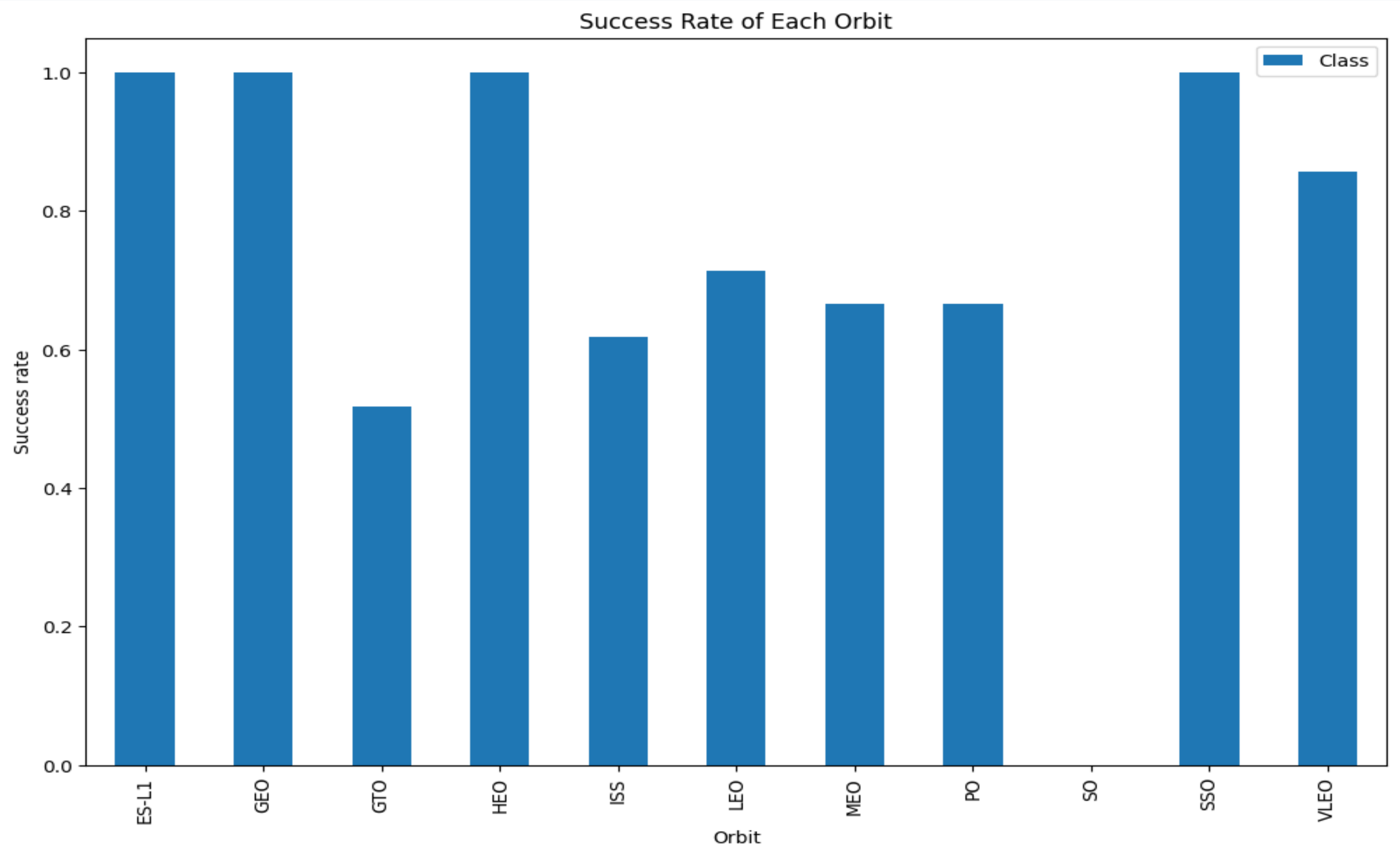
As flights progress, chances of successful landing also increase. CCAFS SLC 40 is the most frequently used site for launches

Payload vs. Launch Site



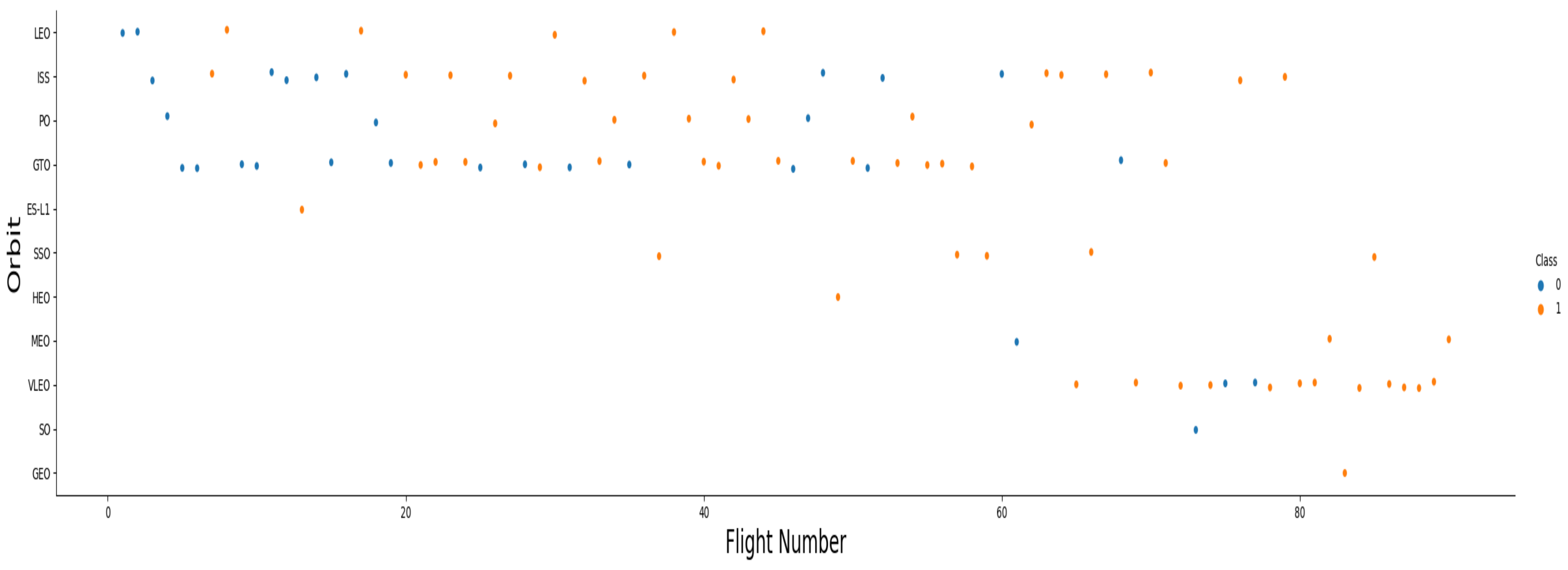
Payloads above 8,000kg have a very successful landing streak compared to smaller payloads

Success Rate vs. Orbit Type



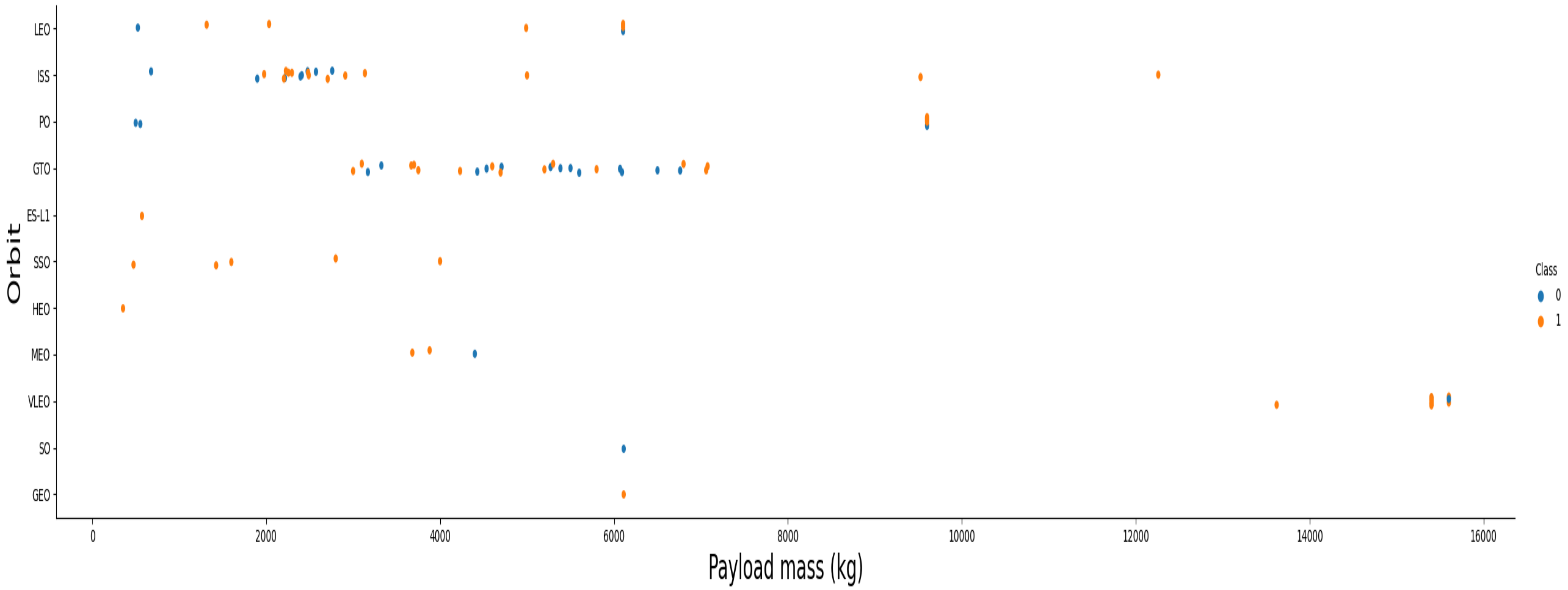
- ES-L1, GEO, HEO, and SSO have the highest success rates for successful landing

Flight Number vs. Orbit Type



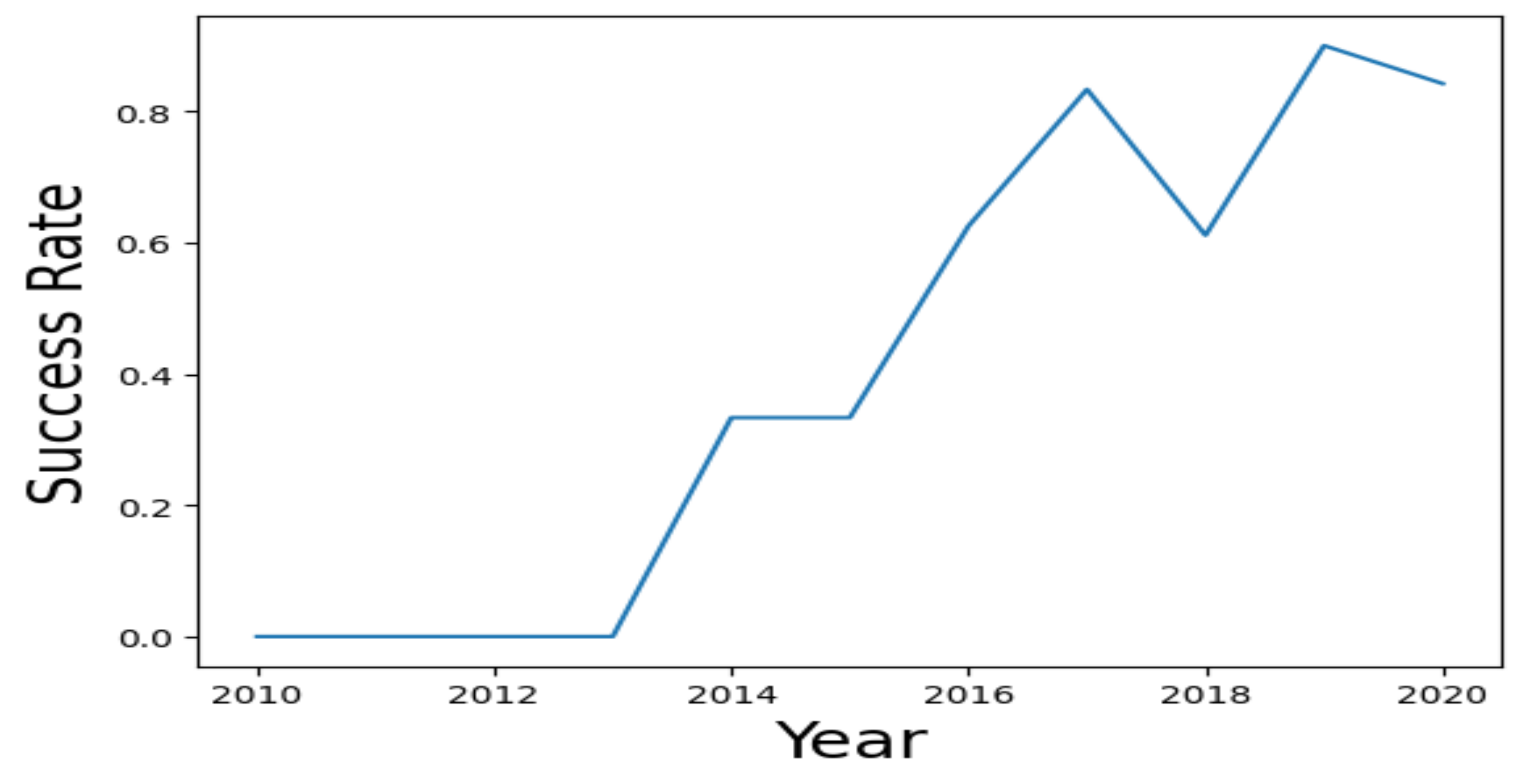
- LEO, and to a lesser extent VLEO, have had great success with landings

Payload vs. Orbit Type



- Large payloads are usually to VLEO and the ISS

Launch Success Yearly Trend



- The yearly average success rate, on average, is on an upward trend, with notable dip in 2018

All Launch Site Names

```
In [16]:  1  %sql select DISTINCT ("Launch_Site") from SPACEXTABLE

 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

- As expected, the database had four launch sites

Launch Site Names Begin with 'CCA'

In [24]:
```
%sql Select * from SPACEXTABLE where "Launch_Site" like "%CCA%" limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[24]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Five records from database where launch sites begin with string `CCA`

Total Payload Mass

In [38]:
```
%sql Select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer"="NASA (CRS)";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[38]:

SUM("PAYLOAD_MASS__KG_")
45596

- Query showing the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
In [39]:  1  %sql select AVG ("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version"="F9 v1.1"

 * sqlite:///my_data1.db
Done.

Out[39]:
```

AVG ("PAYLOAD_MASS__KG_")
2928.4

- Query showing the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
In [50]:  1 %sql Select * from SPACEXTABLE where "Landing_Outcome" like "%success%" order by Date asc limit 1

 * sqlite:///my_data1.db
Done.
```

Out[50]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- Query showing the dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

In [18]:

```
ion") from SPACEXTABLE where "Landing_Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_" between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

Out[18]:

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

- Query of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
In [52]:  1 %sql Select "Mission_Outcome", Count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome"

 * sqlite:///my_data1.db
Done.
```

Out[52]:

Mission_Outcome	Count("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Calculation of the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
In [63]:  1 %sql Select DISTINCT "Booster_Version","PAYLOAD_MASS__KG_" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (Select MAX

 * sqlite:///my_data1.db
Done.
```

Out[63]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Query of the names of the booster which have carried the maximum payload mass

2015 Launch Records

In [72]:
```
%sql Select "Booster_Version", "Launch_Site" FROM SPACEXTABLE where "Landing_Outcome" = "Failure (drone ship)" AND
```

```
 * sqlite:///my_data1.db
Done.
```

Out[72]:

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
In [65]: %sql Select "Landing_Outcome", count(*) from SPACEXTABLE where Date between '2011-06-04' and '2017-03-20' group by

 * sqlite:///my_data1.db
Done.
```

Out[65]:

Landing_Outcome	count(*)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1

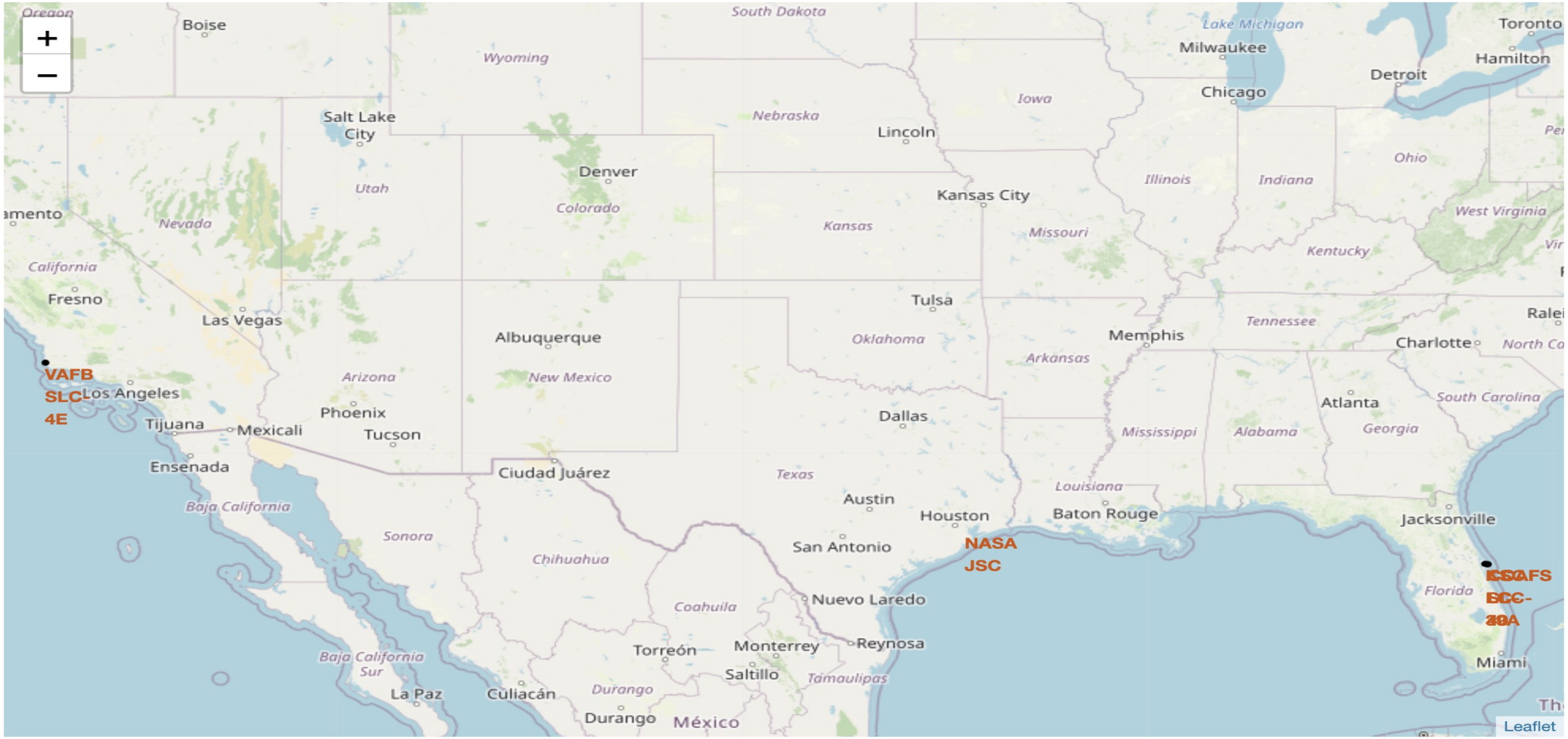
- Ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

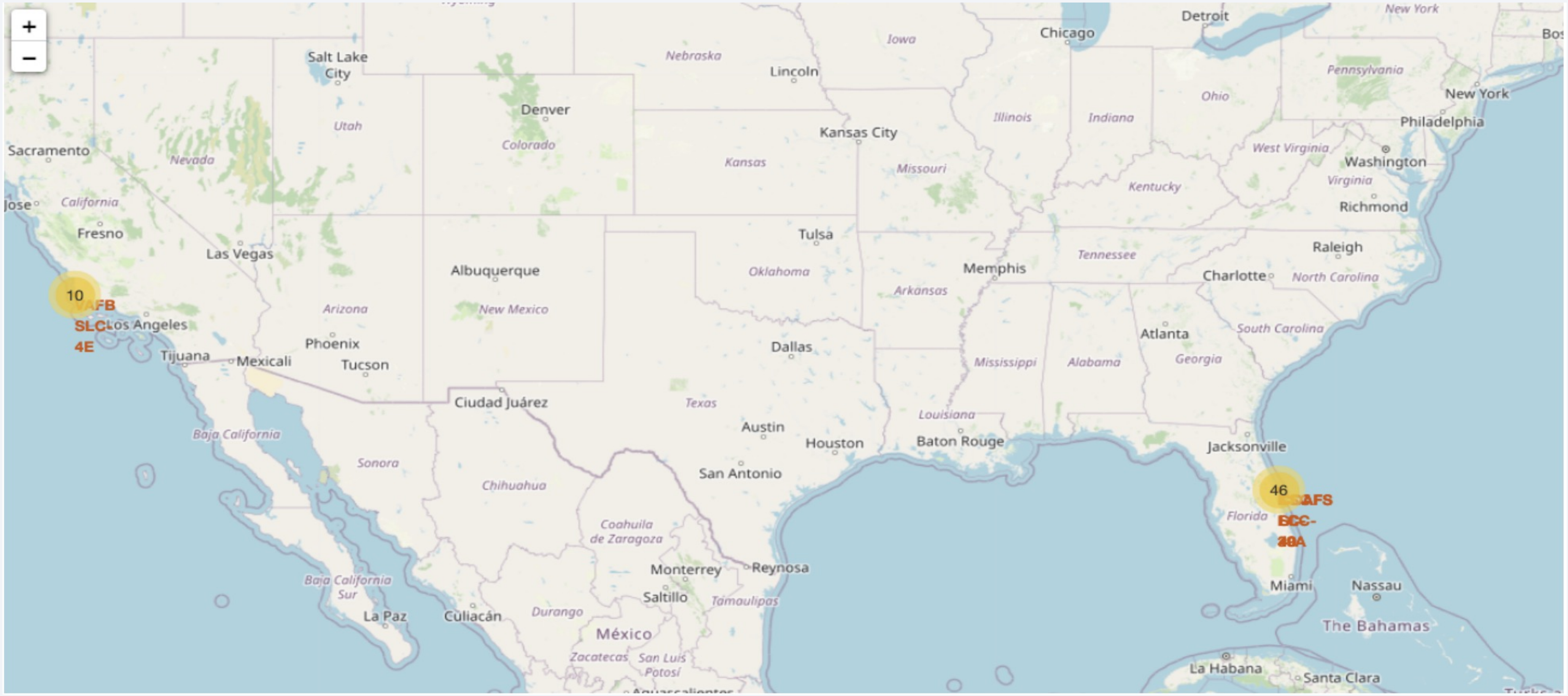
Folium Leaflet Map of SpaceX launch sites

Out[13]:



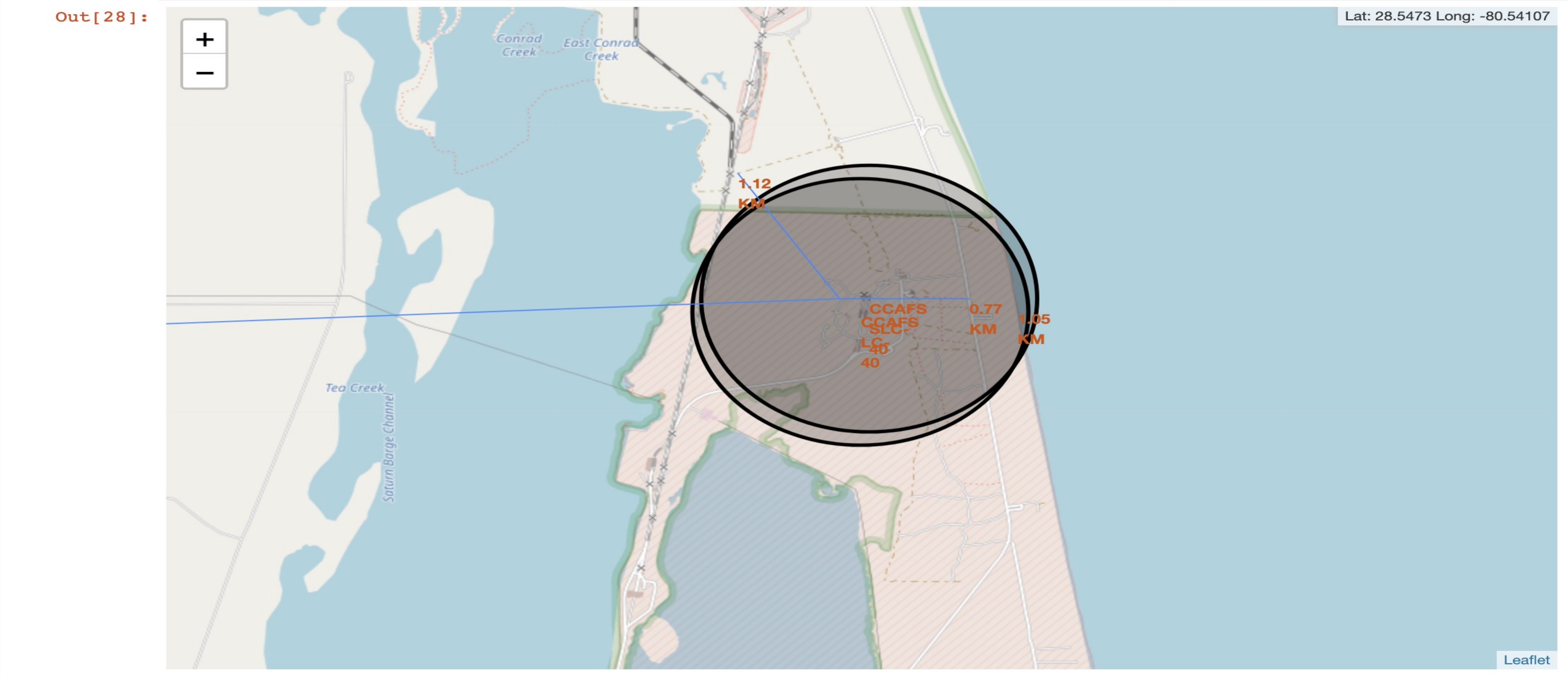
- Leaflet map showing the four SpaceX launch sites, two slightly superimposed in Florida

Folium Map showing Success and Failure of launches



- Success and failure of launches on leaflet map using Folium

Folium Map with Polylines to three landmarks

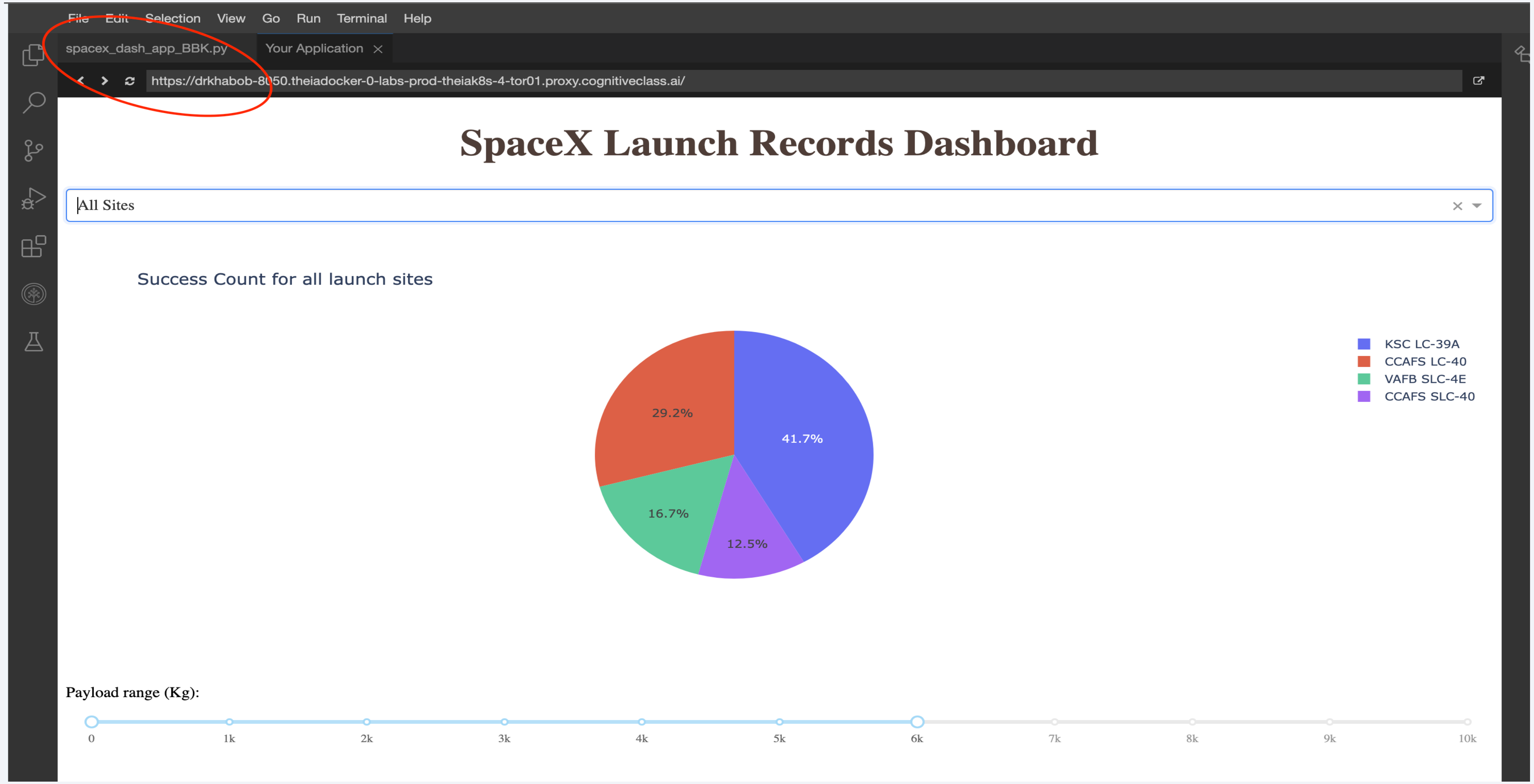


Folium map showing the proximity of a launch site to nearest railway(due northwest line), highway(due easterly), and city (due westerly) with distance calculated and displayed

Section 4

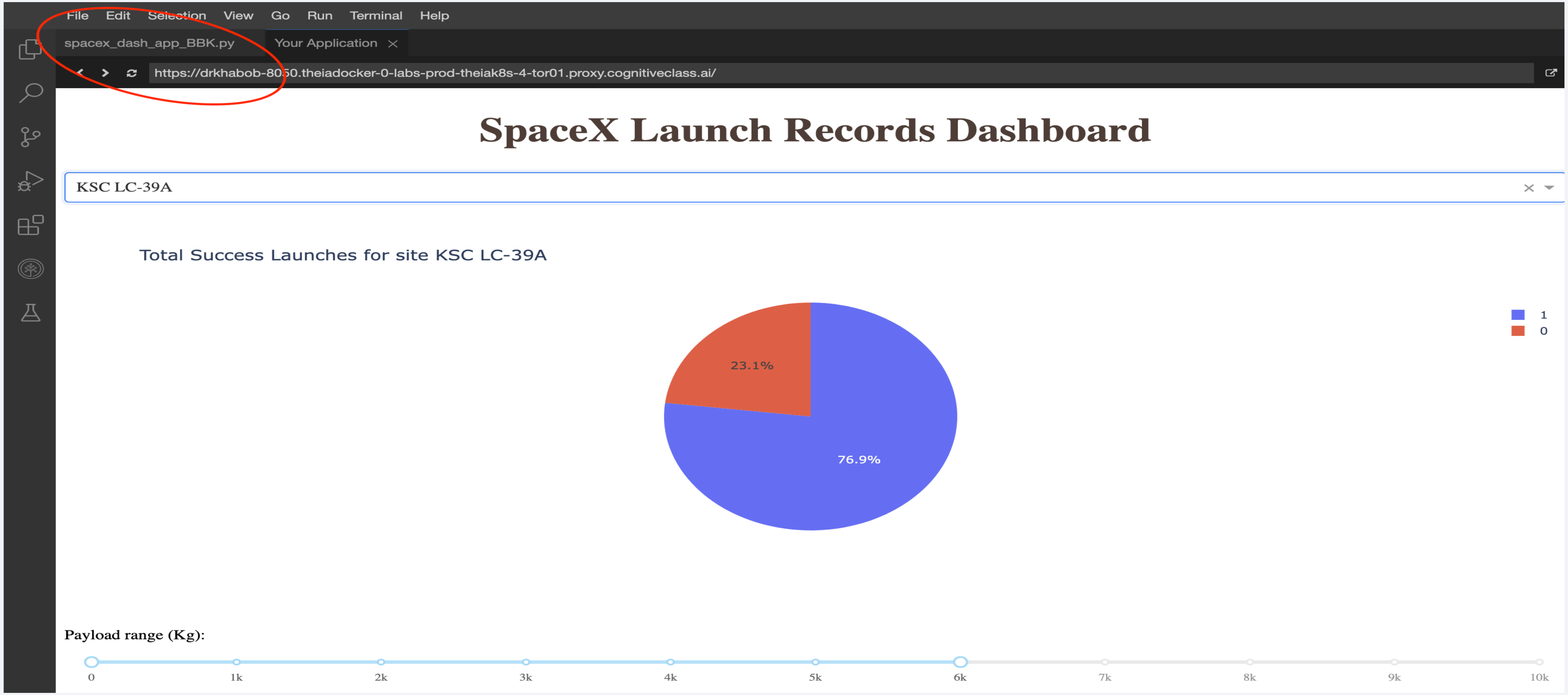
Build a Dashboard with Plotly Dash

Dashboard: SpaceX launch successes by launch site



Contribution of launch sites to successful launches, with KSC LC-39A contributing the most

Dashboard: Success rate of the most prolific launch site



- Pie-chart for the launch site with highest launch success ratio (76.9%)

Dashboard: Payload Vs Launch Outcome (0-9000kg)

- Payload vs. Launch Outcome scatter plot for all sites in the range 0-9000kg

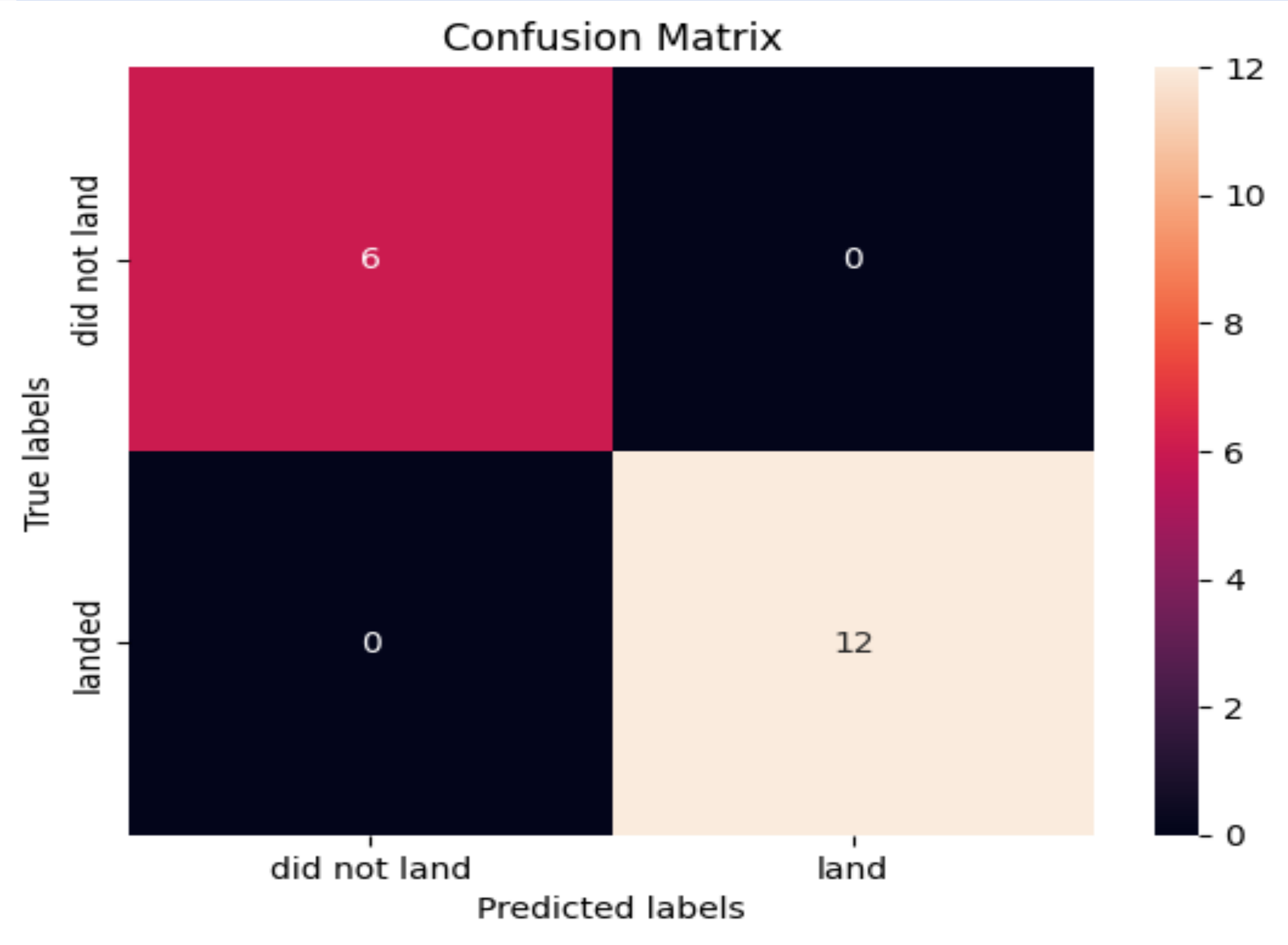
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Three models had high out-of-sample accuracy of ~0.83, but K-Nearest Neighbour had the highest classification accuracy of 1.0 (perfect score).

K-Nearest Neighbour Confusion Matrix



- The Confusion Matrix for K-Nearest Neighbour showing that it classified perfectly under the conditions, with neither false positives, nor false negatives.

Conclusions

- Data science can be used to predict Falcon9 landings,
    - With near-perfect accuracy of out-of-sample prediction by some models i.e. KNN (and high accuracy by the other three models)
    - This can potentially save tens of millions in USD per launch
- Data mining of high-quality, clean and labelled data can be an asset when developing predictive models
- High-quality interactive dashboards and leaflet maps for day-to-day operations of business is demonstrably feasible
- Model evaluation methods to improve model selection based on meritocratic performance measures of out-of-sample data is also feasible

Appendix

- Full GitHub repository: https://github.com/drkhabo/datasci-capstone
- Reference: An Introduction to Statistical Learning (Python): https://hastie.su.domains/ISLP/ISLP_website.pdf
- SpaceX API- https://api.spacexdata.com
- Falcon9 and Falcon Heavy Wikipedia page- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- *For any questions, email: drkhabo.b@gmail.com*

Thank you!