IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

August 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with Web Scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with Pandas and Matplotlib
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Predictive Analysis with Machine Learning Model
- Summary of all results
  - Exploratory Data Analysis
  - Interactive Visual and Dashboard Analytics
  - Predictive Analytics with Machine Learning

Introduction

- Project background and context
  - SpaceX is the most revolutionary rocket launch company on the earth. The company has developed Falcon 9, which first-stage rocket can be reused, to reduce the cost of rocket launch to 62 million USD, compared to its competitors at 165 million.
  - The goal of this project is to predict the landing outcome of the first-stage rocket utilizing data science skills, identifying the market chance of a new company Space Y in the rocket launch industry.
- Problems you want to find answers
  - Figure out the factors that affect the outcome of landing of the first-stage rocket.
  - How to increase the chance of success in future launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia static html page
- Perform data wrangling
  - Calculate the number of launches on each site and the occurrence of each orbit
  - Calculate the number and occurence of mission outcome of the orbits
  - Create a landing outcome label from Outcome column

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Normalized the processed data
    - Divided into training and test sets
    - Evaluated accuracy with different machine learning models

Data Collection

- Describe how data sets were collected.
  - Collected data from SpaceX REST API (https://github.com/r-spacex/SpaceX-API)
  - Scraped the data from a snapshot of the List of Falcon 9 and Falcon Heavy launches Wikipage updated on 9th June 2021 (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- You need to present your data collection process use key phrases and flowcharts

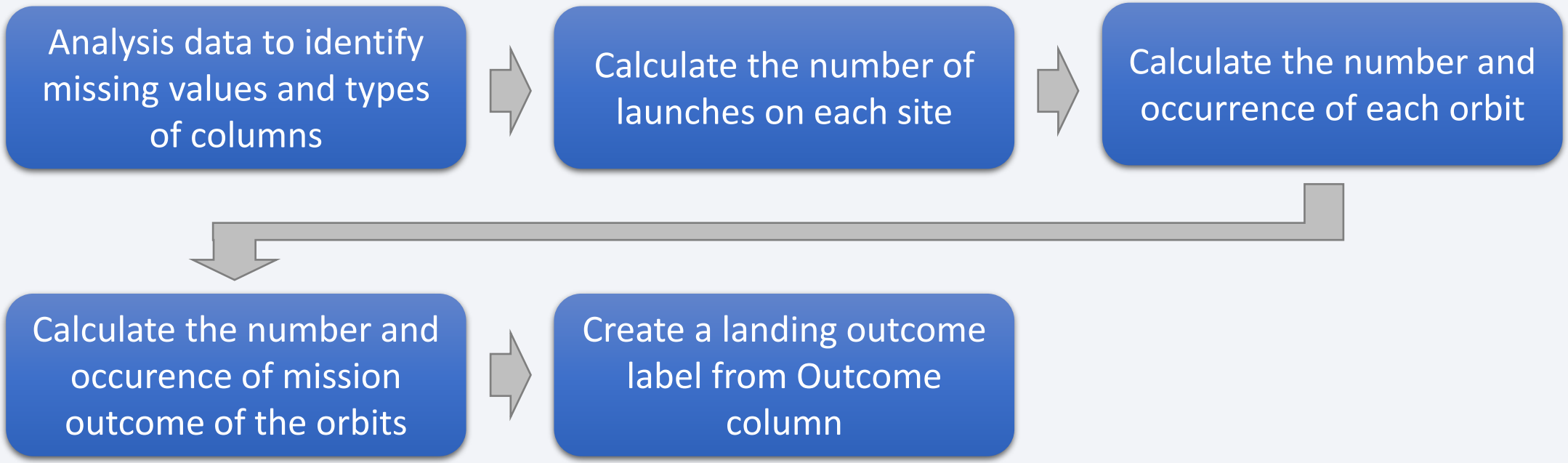
Data Collection – SpaceX API

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling



source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship with scatter plot of Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit type, and Payload and Orbit type.
- Visualize the relationship between success rate of each orbit type with bar chart.
- Visualize the launch success yearly trend with line chart.
- Create dummy variables to categorical columns and cast all numeric columns to float64.

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- 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
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Use markers, circles, lines and marker clusters to:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Create an interactive dashboard with:

- A dropdown list to select distinct Launch Site
- A pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site
- A slider to select payload range
- A scatter chart to show the correlation between payload and launch success

source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

Perform exploratory Data Analysis and determine Training Labels

- create a column for the class
- Standardize the data
- Split into training data and test data

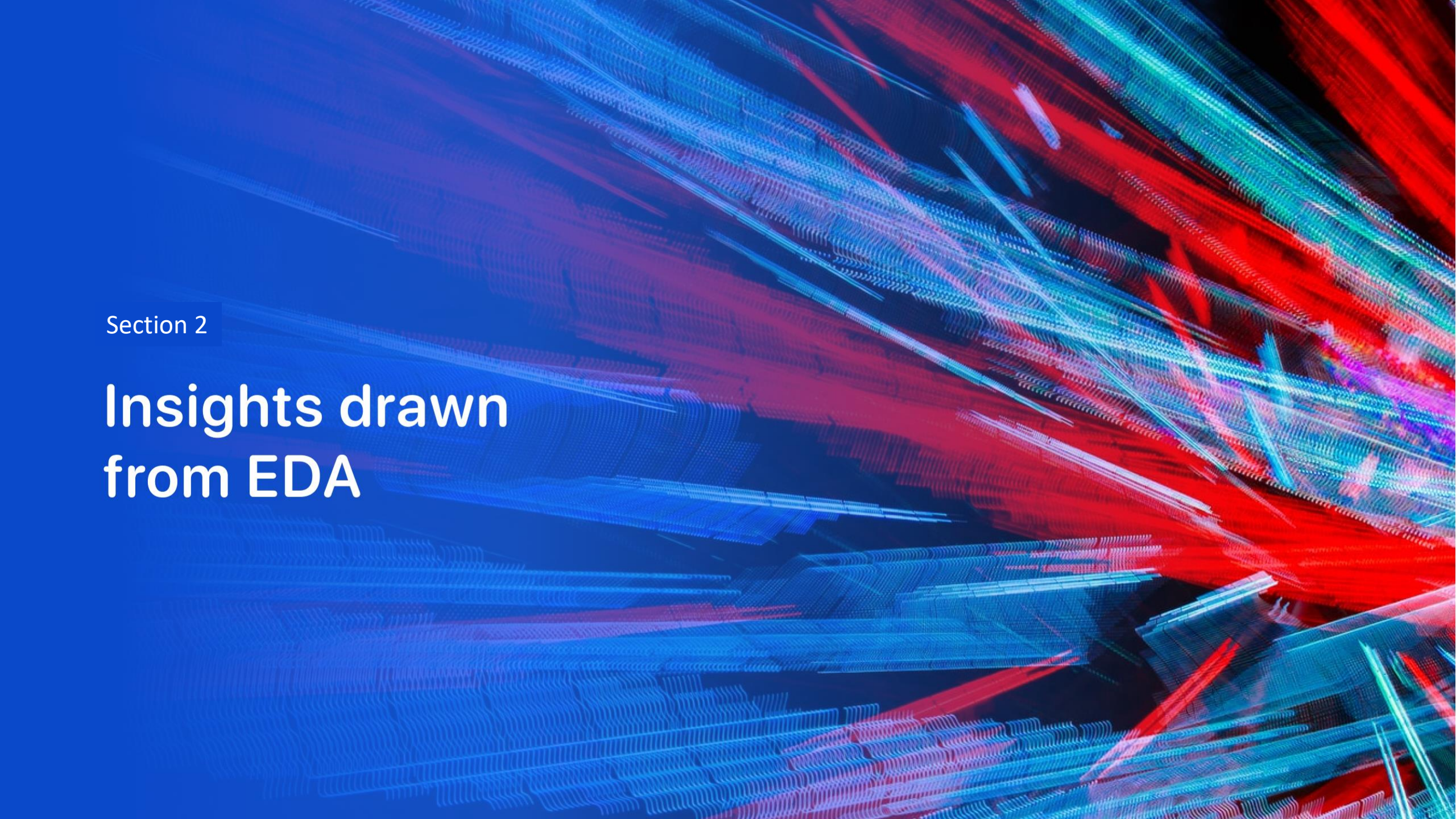
Find best Hyperparameter for Logistic Regression, SVM, Decision Trees and KNN

- Find the method performs best using test data

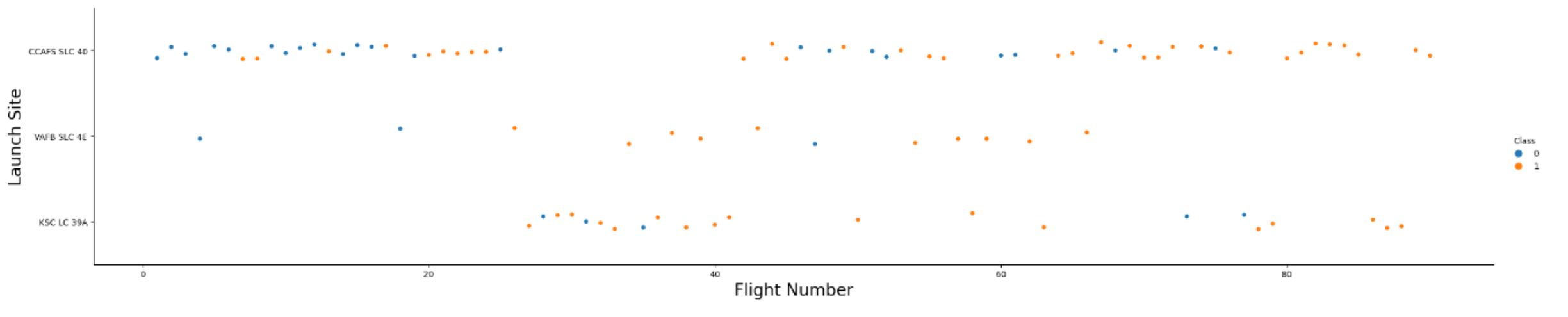
source:
https://github.com/zachho9/IBM-Applied-Data-Science-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

Results

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

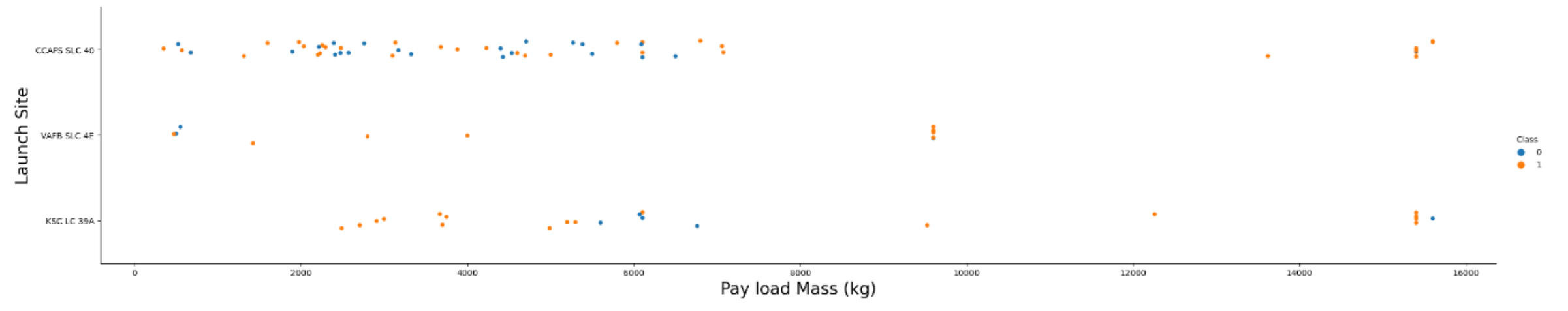

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



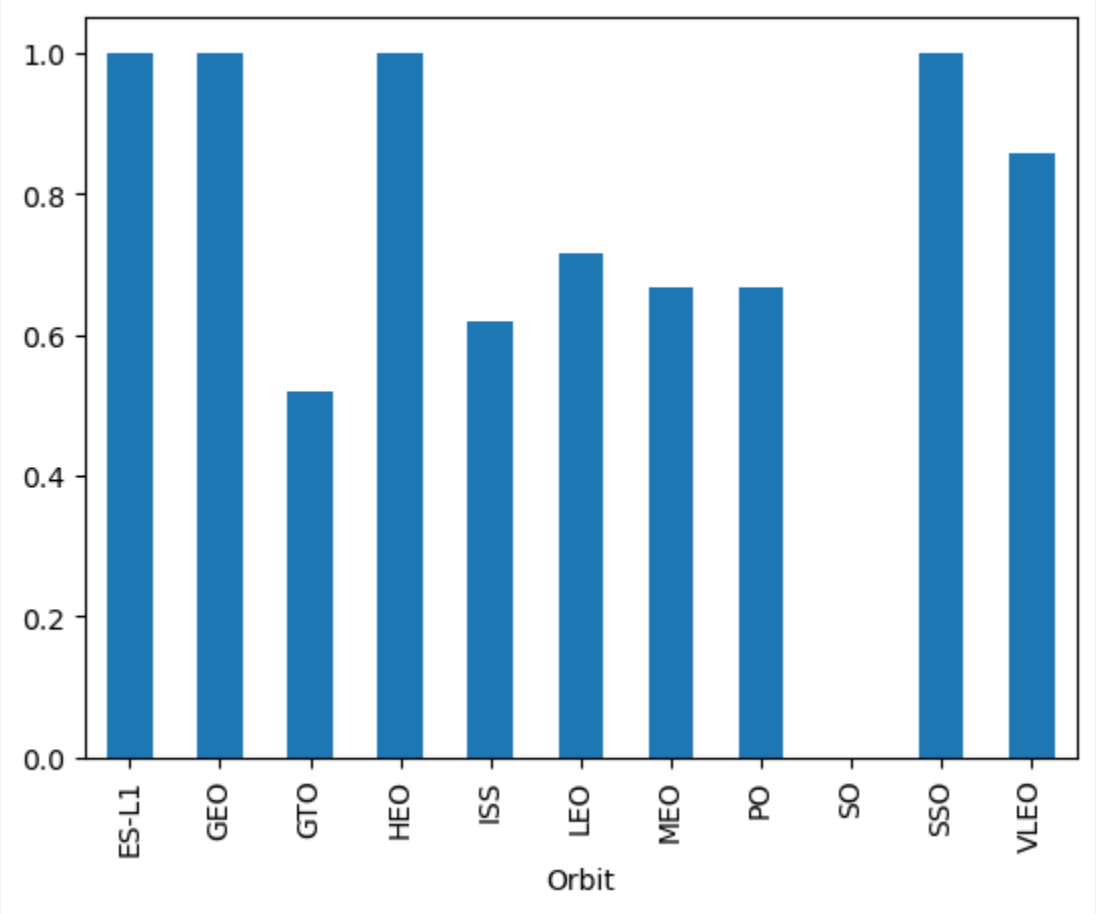
- CCAF5 SLC 40 is the most successful launch site, followed by VAFB SLC 4E and KSC LC 39A.
- CCAF5 SLC 40 is the most frequent used launch site.
- The overall success rate increases in the three sites over time.

Payload vs. Launch Site



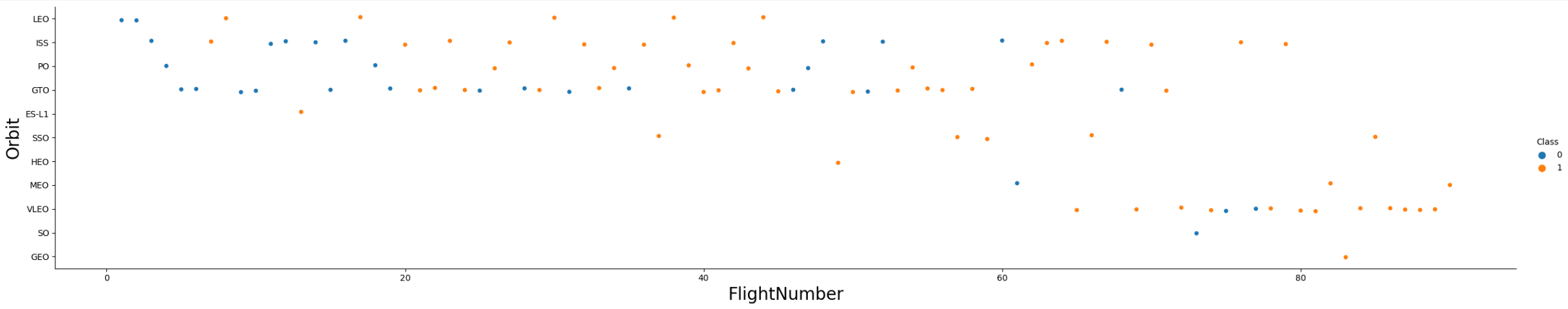
- The majority of payloads are below 7500kg, but nearly half of the launches are failed.
- Payloads above 8500kg have significant success rate.
- CCAFS SLC 40 and KSC LC 39A are mainly used for payloads above 12000kg.

Success Rate vs. Orbit Type



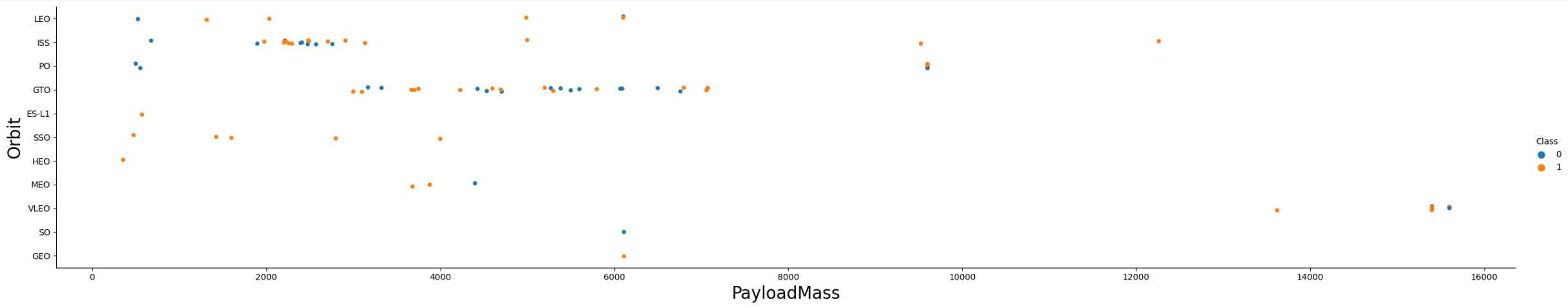
- ES-L1, GEO, HEO, and SSO are the most successful orbit type with success rate at 1.0.
- VLEO ranks 2nd at around 0.8.
- LEO ranks 3rd at around 0.7.
- GTO is the least successful orbit, at nearly 0.5.

Flight Number vs. Orbit Type



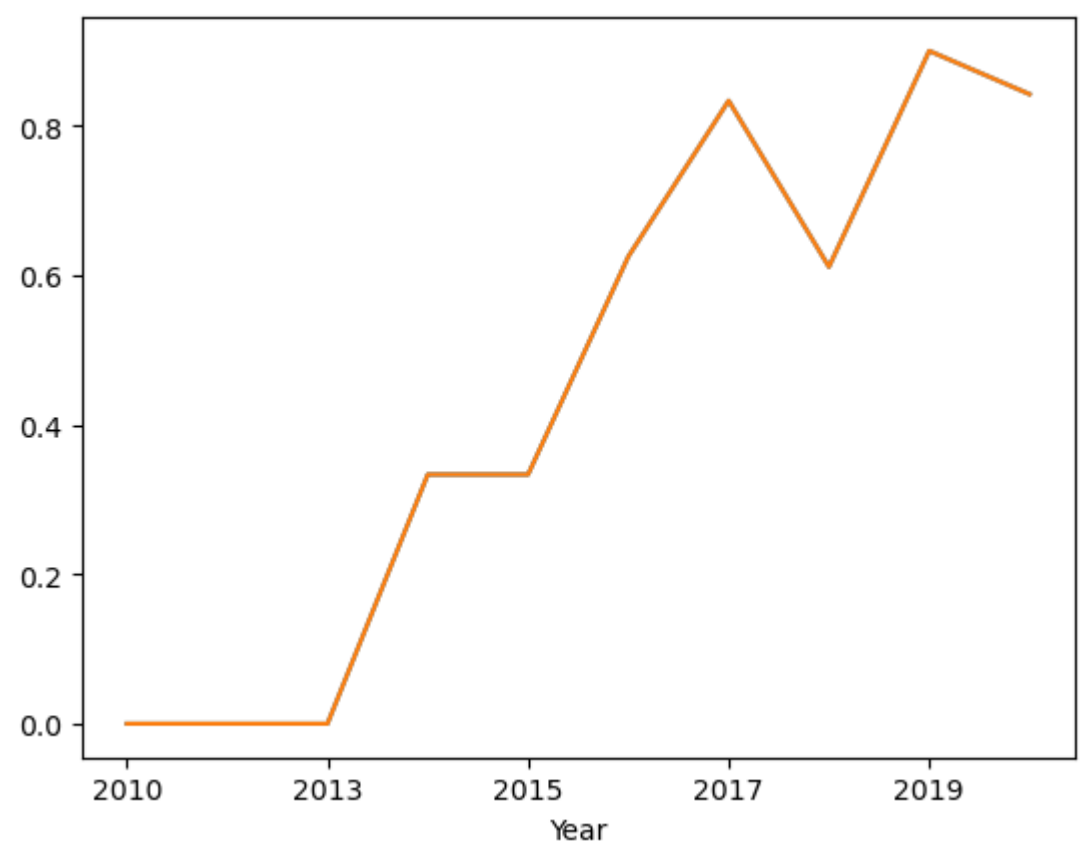
- LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

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

```
In [5]: %sql select distinct launch_site from spacex
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[5]:

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

- Use SQL query to select unique launch site values from the spacex dataset.

Launch Site Names Begin with 'CCA'

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

In [6]:
```
%sql select * from spacex where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[6]:

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

- Use SQL query to select 5 records with launch site name begin with cca from the spacex dataset.

Total Payload Mass

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

```
In [7]: %sql select sum(PAYLOAD_MASS__KG_) as Total from spacex where CUSTOMER='NASA (CRS)'
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[7]:

total
45596

- Use SQL query with sun() function to calculate total amount of payload mass with the customer column nasa(crs) from the spacex dataset.

Average Payload Mass by F9 v1.1

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

In [8]:
```
%sql select avg(PAYLOAD_MASS__KG_) as avg from spacex where BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[8]:

AVG
2928

- Use SQL query with avg() function to calculate average amount of payload mass carried by F9 v1.1 booster version from the spacex dataset.

First Successful Ground Landing Date

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

Hint:Use min function

In [9]:
```
%sql select min(DATE) as first_date from spacex where LANDING_OUTCOME like '%ground pad%'
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[9]:

first_date
2015-12-22

- Use SQL query with min() function to extract the earliest date of successful ground landing from the spacex dataset.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [10]:
```
%sql select BOOSTER_VERSION from spacex where LANDING_OUTCOME='Success (drone ship)' and (PAYLOAD_MASS__KG_ >= 4000 and PAYLOAD_MASS__KG_ <= 6000)
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[10]:

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

- Use SQL query to extract the payload between 4000 and 6000, as well as the landing outcome equals to success (drone ship) from the spacex dataset.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [11]:
```
%sql select MISSION_OUTCOME, count(*) as total from spacex group by MISSION_OUTCOME
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[11]:

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

- Use SQL query with count() function to calculate the total number different mission outcomes from the spacex dataset.

Boosters Carried Maximum Payload

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

```
In [12]: %sql select BOOSTER_VERSION from spacex where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacex) order by BOOSTER_VERSION
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[12]:

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

- Use SQL query with subquery to extract the maximum payload for each unique booster version from the spacex dataset.

2015 Launch Records

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

In [13]:
```
%sql select DATE, LAUNCH_SITE, BOOSTER_VERSION, LANDING_OUTCOME from spacex where LANDING_OUTCOME='Failure (drone ship)' and year(date)=2015
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[13]:

DATE	launch_site	booster_version	landing_outcome
2015-10-01	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
2015-04-14	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

- Use SQL query to extract the failed landing record in year 2015 from the spacex dataset.

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

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 [14]:
```
%sql select LANDING_OUTCOME, count(LANDING_OUTCOME) as total from spacex where DATE between '2010-06-04' and '2017-03-20' group by LANDING_OUTCOME order by
```

```
 * ibm_db_sa://wml38749:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[14]:

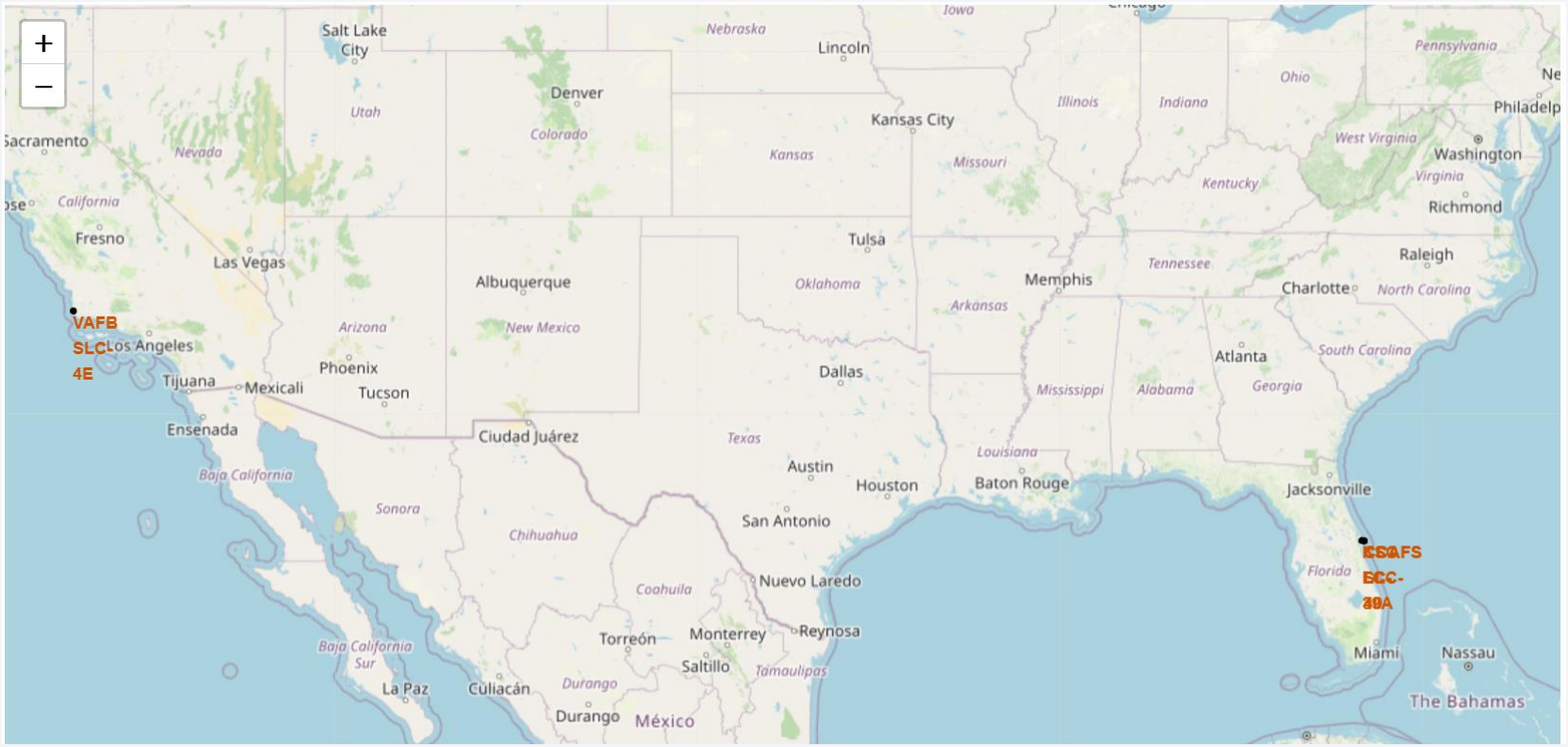
landing_outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

- Use SQL query with count() function to extract the landing outcome within the given period of time, and ordered by total number from the spacex dataset.

Section 3

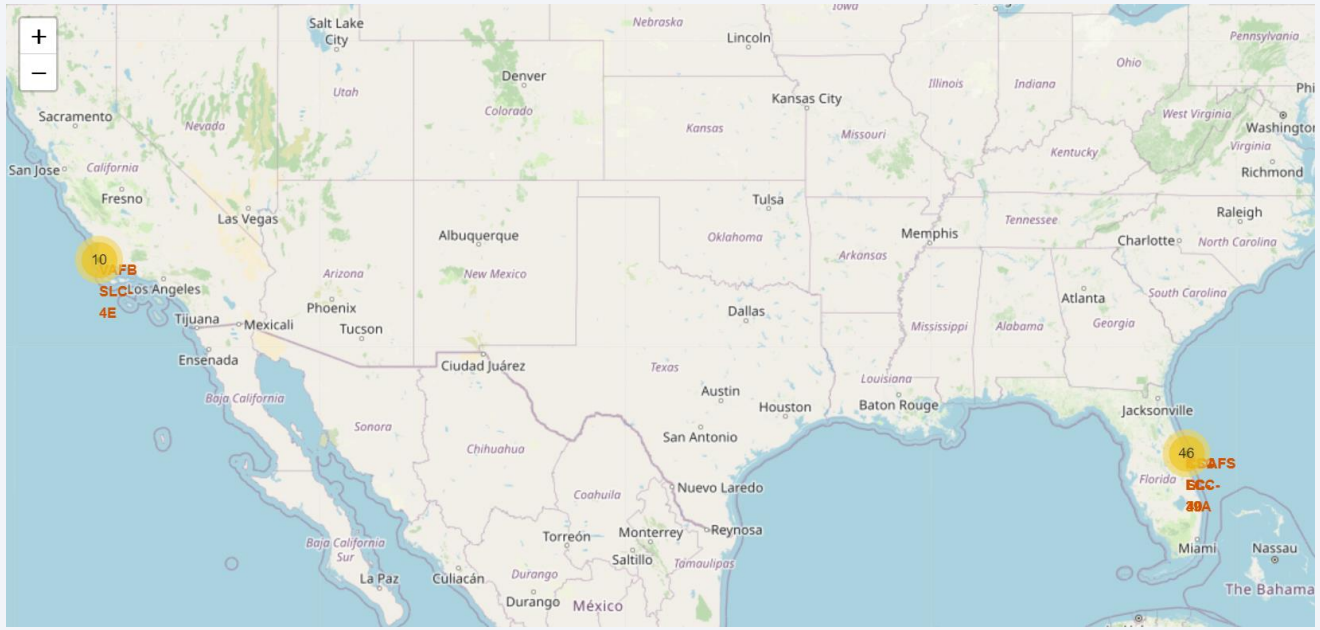
Launch Sites Proximities Analysis

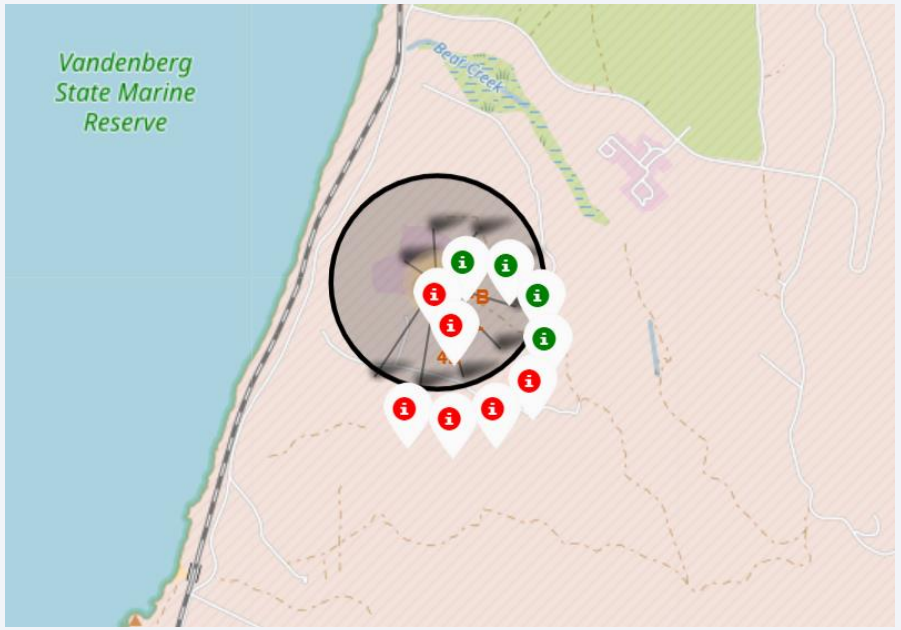
Mark all launch sites on a map



- All launch sites are near the coastline of USA.

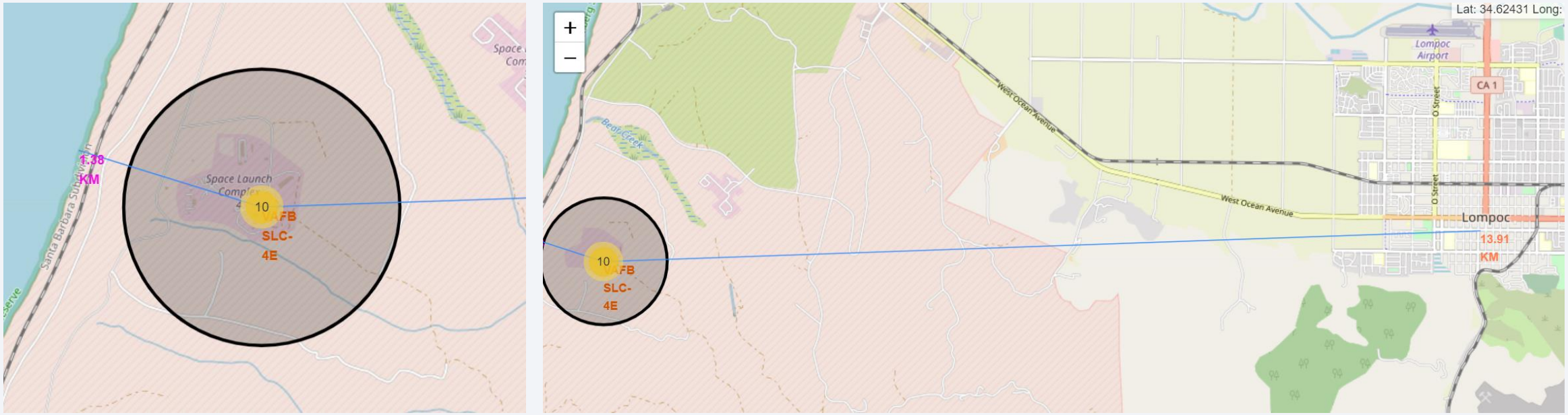
Mark the success/failed launches for each site on the map



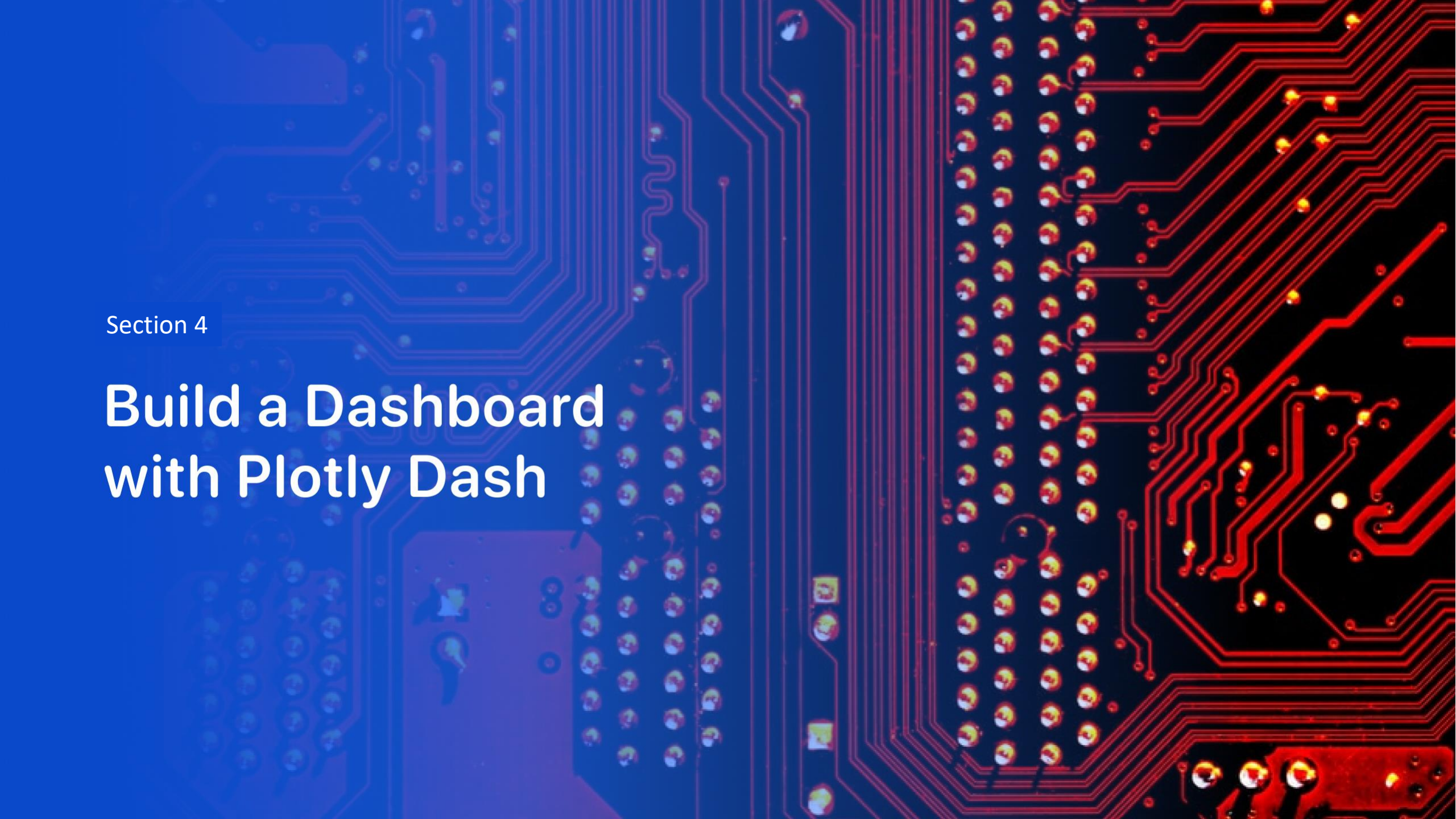


- Enhance the map by adding the launch outcomes for each site
- Click on the site icon to show detailed outcomes

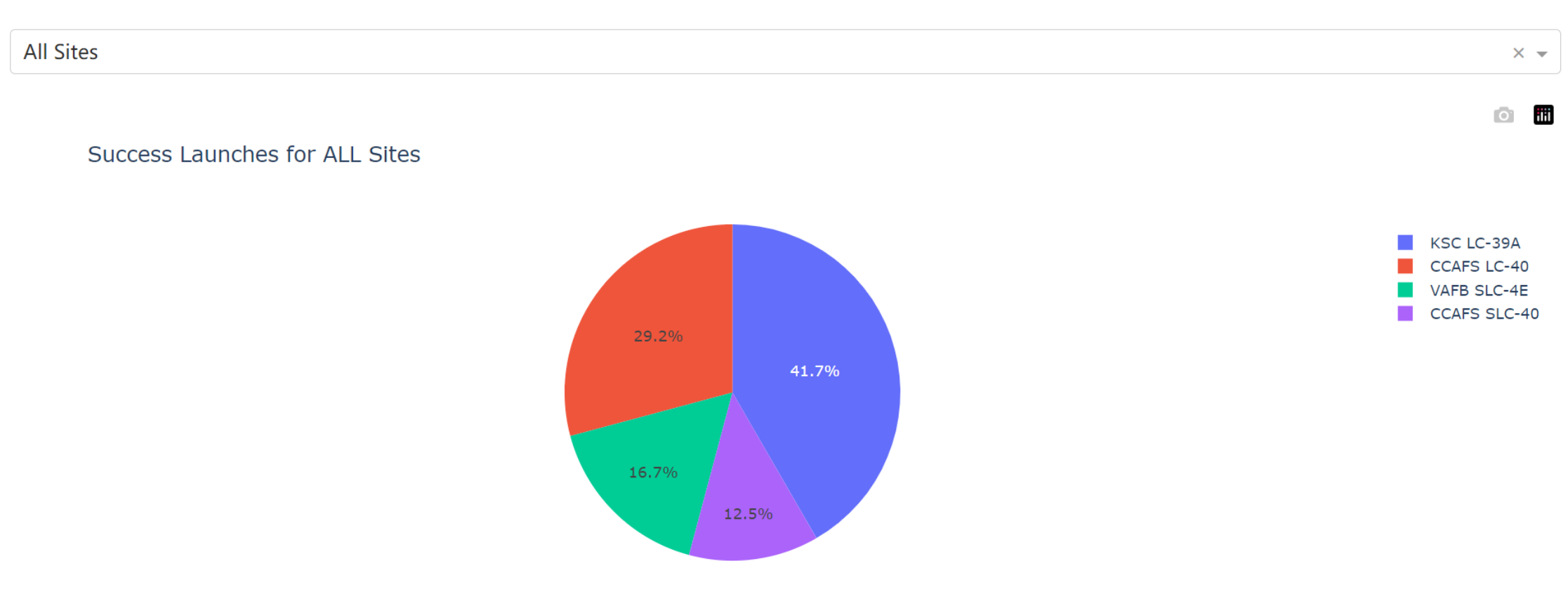
Calculate the distances between a launch site to its proximities



- Choose the site VAFB SLC-4E.
- The site is located around 1.38km away from the coastline, and 13.91km away from the nearest town Lompoc.

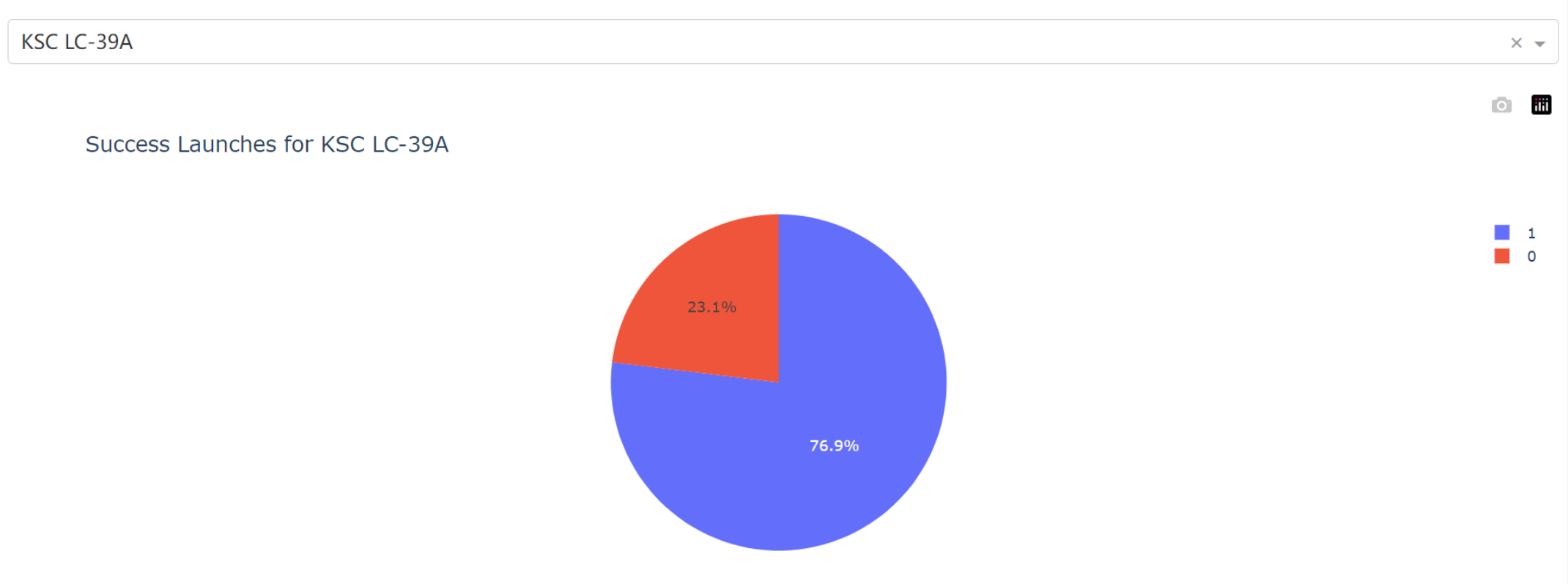

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



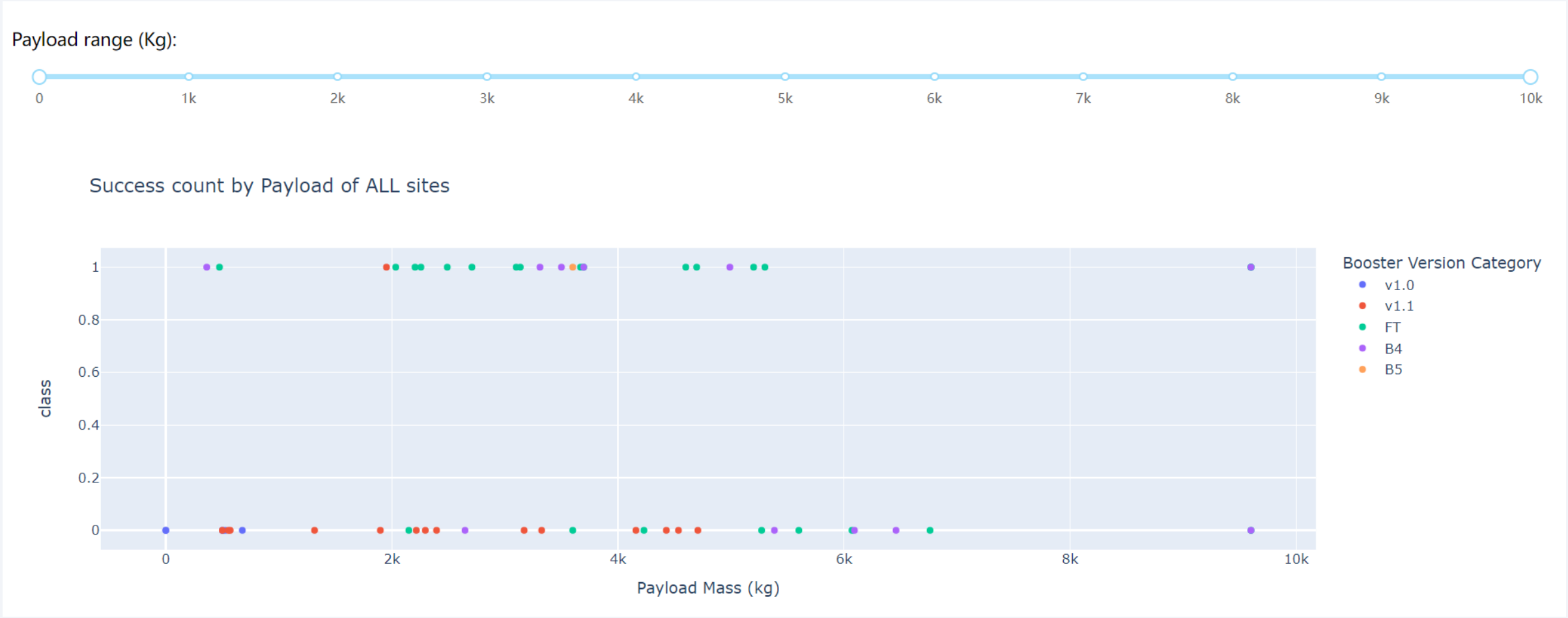
- KSC LC-39A takes up 41.7% of all success launches among all sites.
- CCAFS SLC-40 ranks bottom, at 12.5%.

Launch site with highest launch success ratio



- KSC LC-39A has the highest success rate, at 76.9%.

Payload vs. Launch Outcome scatter plot for all sites

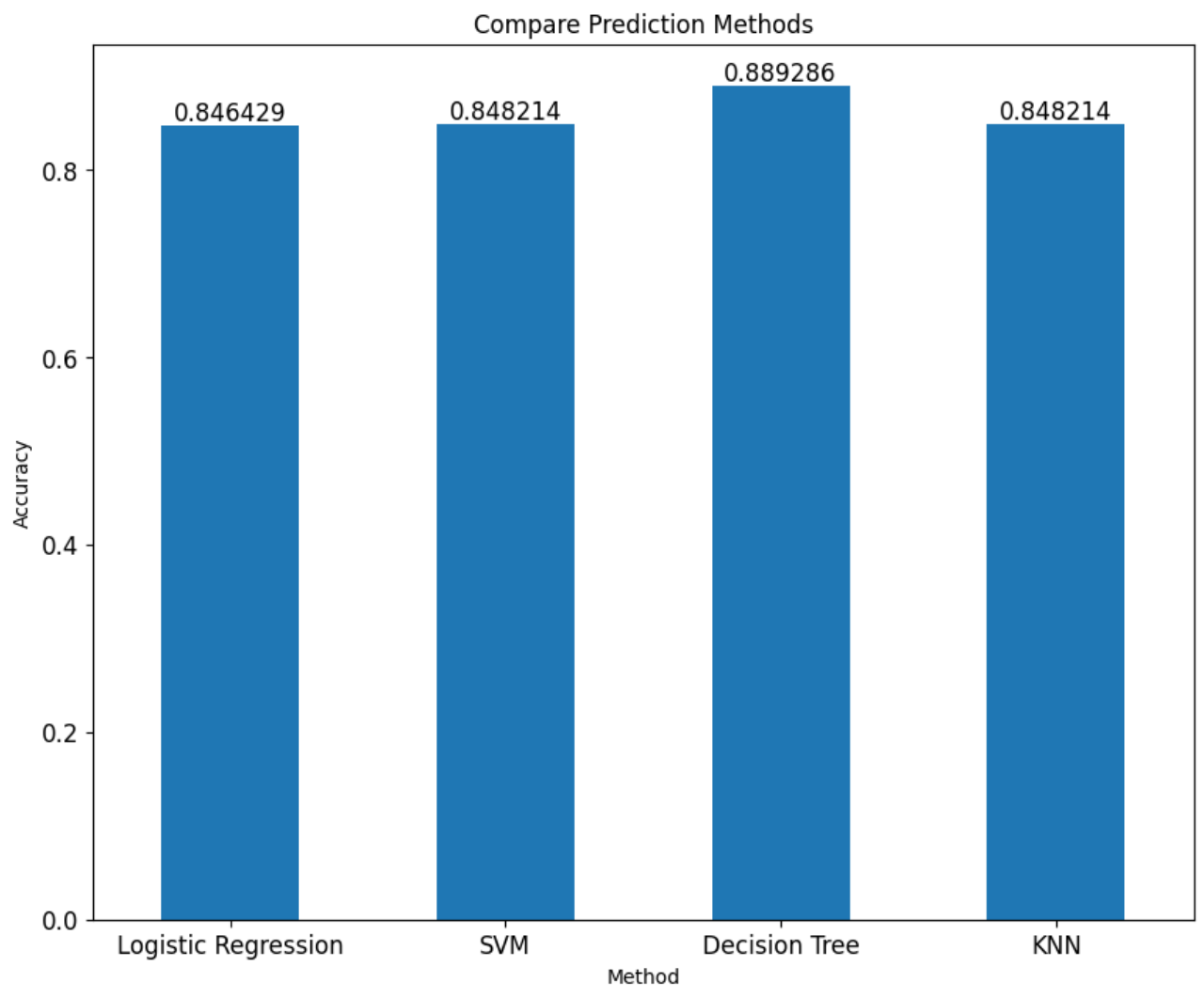


- When payloads are low (below 6000kg), the success rate increases.

Section 5

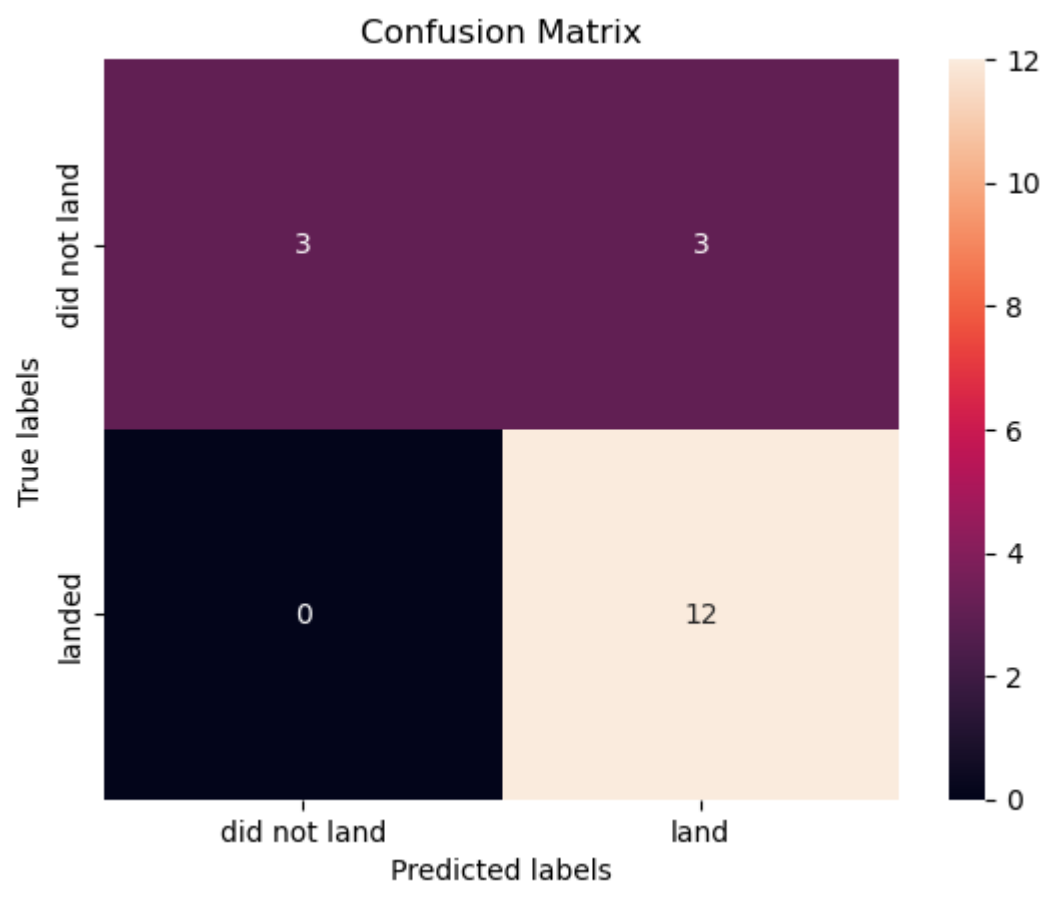
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree Classifier is the most accurate model, with the accuracy at 0.889286.

Confusion Matrix of Decision Tree



- The confusion matrix shows that Decision Tree Classifier can classify different classes effectively.

Conclusions

- The overall success rate has improved since 2013.
- KSC LC-39A is the best launch site.
- Low payloads mass (below 6000kg) have better landing outcomes than heavy payloads.
- The Decision Tree is the best Machine Learning method for this dataset.

Thank you!