Winning Space Race with Data Science

Xiaobing Shang
May 11 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis Using SQL
  - Exploratory Data Analysis for Data Visualization
  - Interactive Visual Analytics and Dashboard with Folium
  - Machine Learning Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Visual Analytics Visualization and Dashboard Results
  - Predictive Analytics (Classification) results

Introduction

- Project background and context

  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. In this lab, you will collect and make sure the data is in the correct format from an API.

- To predict if the first stage will land given the data from the preceding labs
- To determine what factors determine on the land success

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - REST API – SPACEX
  - Web Scraping - Wikipedia
- Perform data wrangling
  - One-hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Numpy - SciKitLearn
  - logistic regression, decision tree, support vector machine and K-nearest neighbor models
  - identify the best performing model for prediction

Data Collection

Data was collected through

- SPACEX API and
- Wikipedia Webpage.

Data Collection – SpaceX API

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

```
200
```

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = launch_df[launch_df['BoosterVersion'] != 'Falcon 9']
data_falcon9
```


Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · umgcjuan/Capstone (github.com)

Data Collection - Scraping

Data Request → Extract Data → Pandas DF

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
# assign the response to a object
html_data = requests.get(static_url)
html_data.status_code
```

TASK 2: Extract all column/variable names from the HTML table header

Next, we want to collect all relevant column names from the HTML table header

Let's try to find all tables on the wiki page first. If you need to refresh your memory about `BeautifulSoup`, please check the external refe
this lab

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

Starting from the third table is our target table contains the actual launch records.

```
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

- Capstone/jupyter-labs-webscraping.ipynb at main · umgcjuan/Capstone (github.com)

Data Wrangling

- Performed Exploratory Data Analysis (EDA) on the data set to determine training labels,
- Calculated the launches by location, and the numbers and occurrence of orbits
- Calculated the number and occurrence of mission outcome per orbit type
- Created landing outcome label from Outcome column

Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb at main · umgcjuan/Capstone (github.com)

EDA with Data Visualization

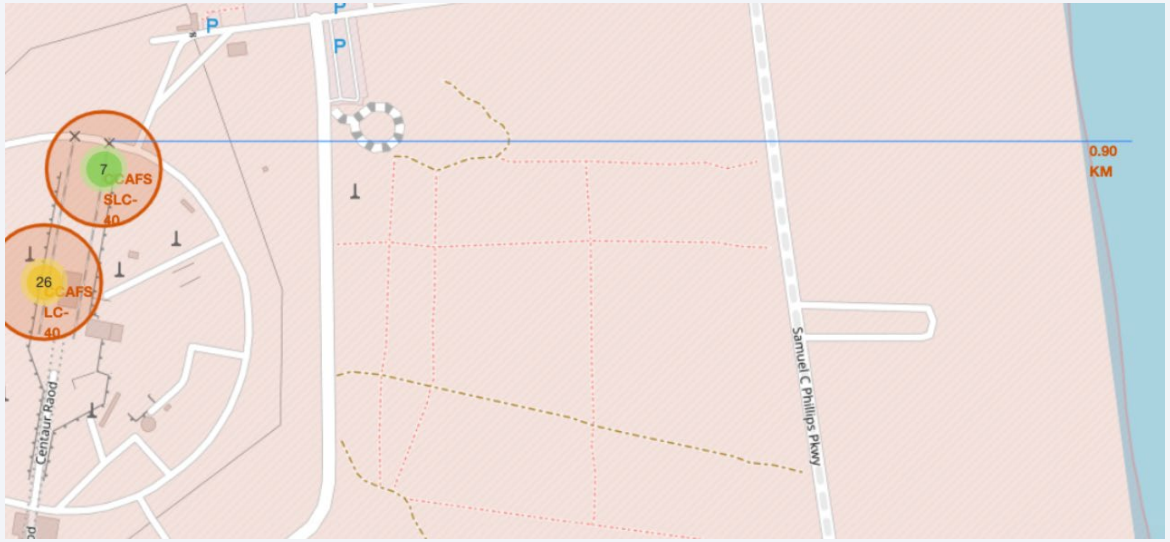
- Successful and failed landings were plotted to detect the relationships between variables.
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload and Launch Site
  - Success rate of each orbit type
  - FlightNumber and Orbit type
  - Payload and Orbit type
  - Launch success yearly trend
- Capstone/jupyter-labs-eda-dataviz.ipynb at main · umgcjuan/Capstone (github.com)

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Names of the unique launch sites in the space mission
  - Launch sites begin with the string 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship
  - Total number of successful and failure mission outcomes
  - Use subquery to query the names of the booster versions with the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- https://github.com/umgcjuan/Capstone/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Map objects created: markers, circles, lines, etc.
- 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
- https://github.com/umgcjuan/Capstone/blob/main/lab_jupyter_launch_site_location.ipynb



Build a Dashboard with Plotly Dash

- Add Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-scatter-chart scatter plot
- https://github.com/umgcjuan/Capstone/blob/main/app.py

Predictive Analysis (Classification)

- create a column for the class
- Standardize the data
- Split into training data and test data
- Find the method performs best using test data - Best model is DecisionTree with a score of 0.875
- https://github.com/umgcjuan/Capstone/blob/main/SpaceX_MachineLearningPrediction_Part_5.ipynb

Results

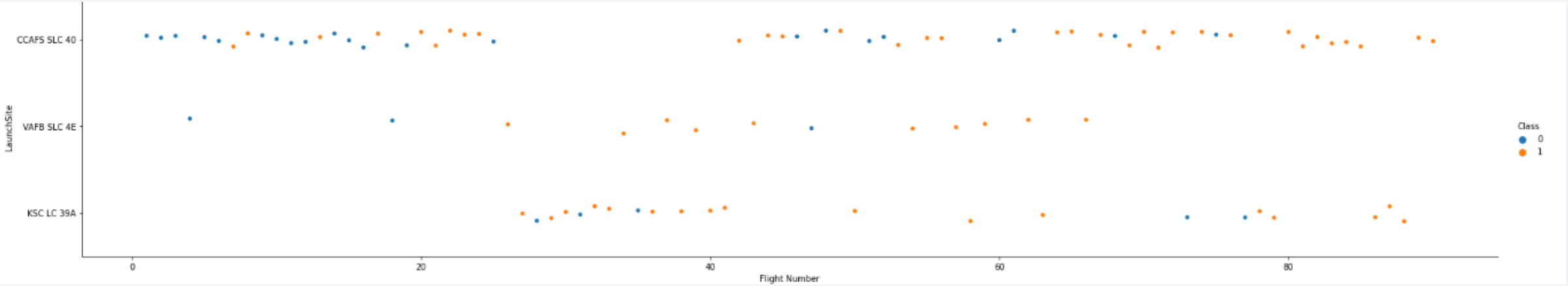
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

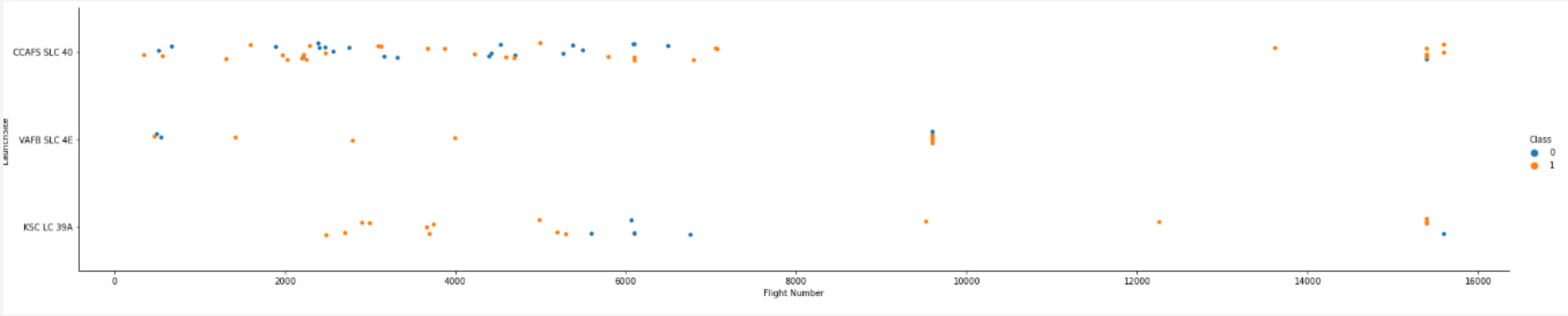
Flight Number vs. Launch Site

- The plot shows the flight amount at a launch site has a positive impact on the success rate at a launch site.



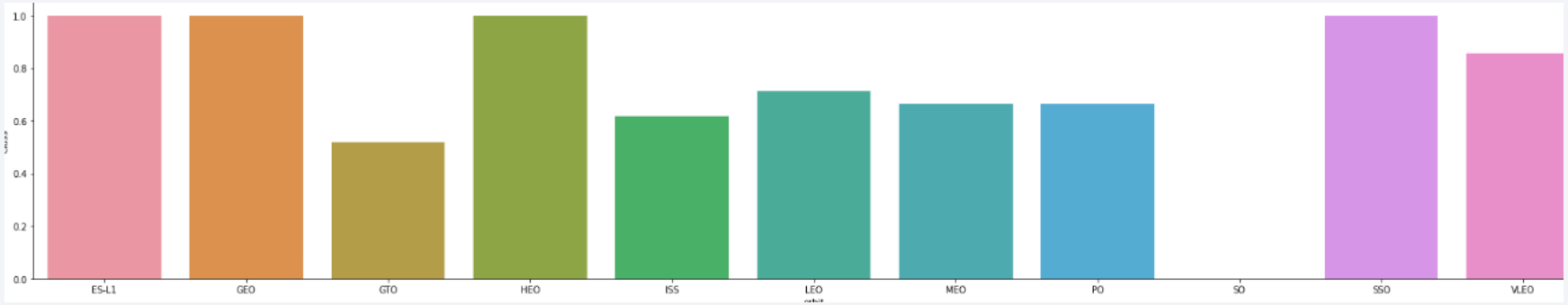
Payload vs. Launch Site

- The plot shows those have higher payload mass result higher success landing rates
- Site CCAFS SLC 40 has more lower payload launches



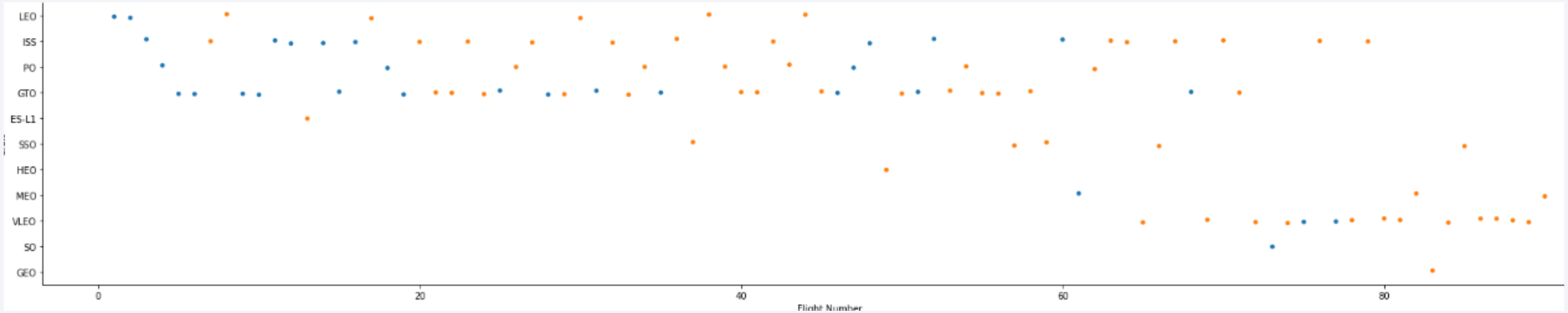
Success Rate vs. Orbit Type

- Landing success rates are various based on orbit types.



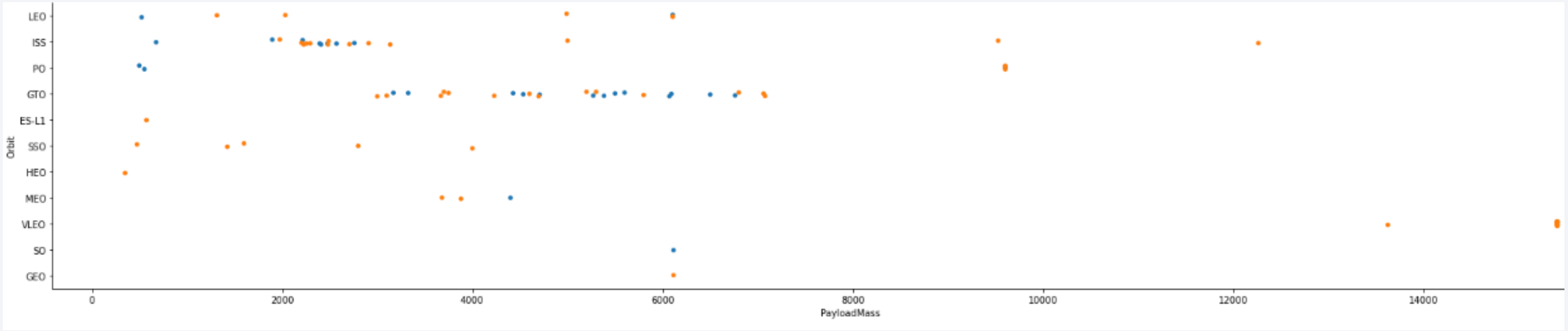
Flight Number vs. Orbit Type

- Landing attempts were failed among earlier flights
- LEO launches had more successful landing during early stage.



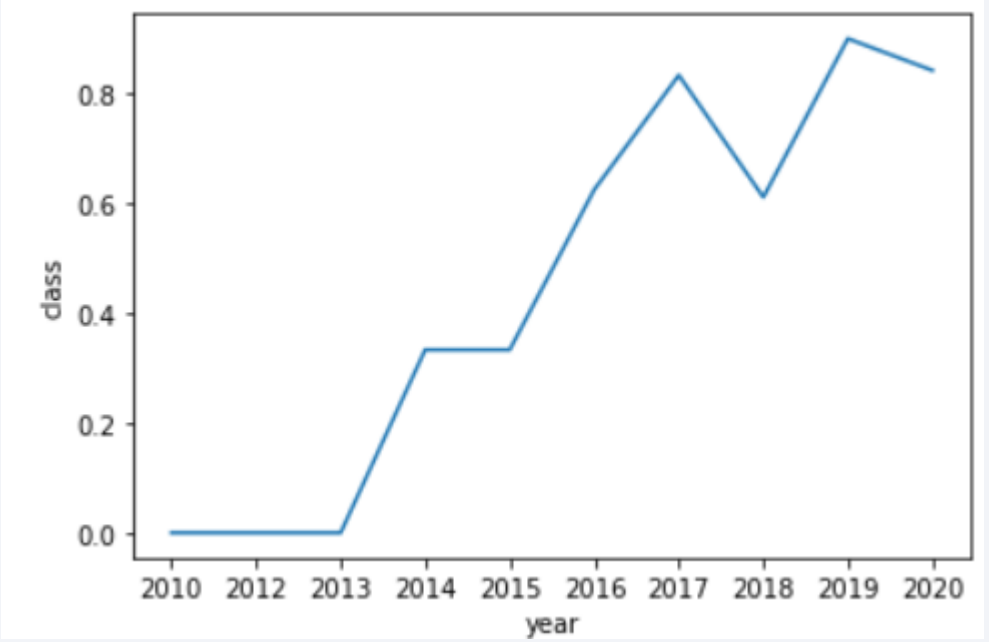
Payload vs. Orbit Type

- Below plot shows there are more successful landsing for those PO, ISS, and LEO orbits, which have heavier payloads.



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

```
%sql select distinct launch_site from spacex
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

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

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

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcor
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
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 (parachu
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

Task 3

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

```
%sql select sum(payload_mass__kg_) from spacex where customer like 'NASA%
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0
Done.
```

1
99980

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_) from spacex where booster_versio
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n
Done.
```

Out[20]:

1
2534

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) from spacex where landing__outcome like '%ground pad%'
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnr
Done.
```

Out[21]:

1
2015-12-22

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

In [22]:
```
%sql select distinct booster_version from spacex where (payload_mass__kg_ < 6000 AND payload_mass__kg
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdom
Done.
```

Out[22]:

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1020
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [23]:
```
%sql SELECT mission_outcome, count(*) AS total FROM spacex GROUP BY mission_outcome
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98
Done.
```

Out[23]:

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

Task 8

Boosters Carried Maximum Payload

Task 8

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

In [24]:
```
%sql SELECT DISTINCT (booster_version) from spacex where payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUD
Done.
```

Out[24]:

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

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 [25]:
```
%sql select booster_version, launch_site from spacex where landing__outcome = 'Failure (drone ship)' AND DATE like '2015%'
```

```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[25]:

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

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, i

[26]:
```
%sql SELECT * FROM spacex where DAYNAME(DATE)='Friday' LIMIT 5
```

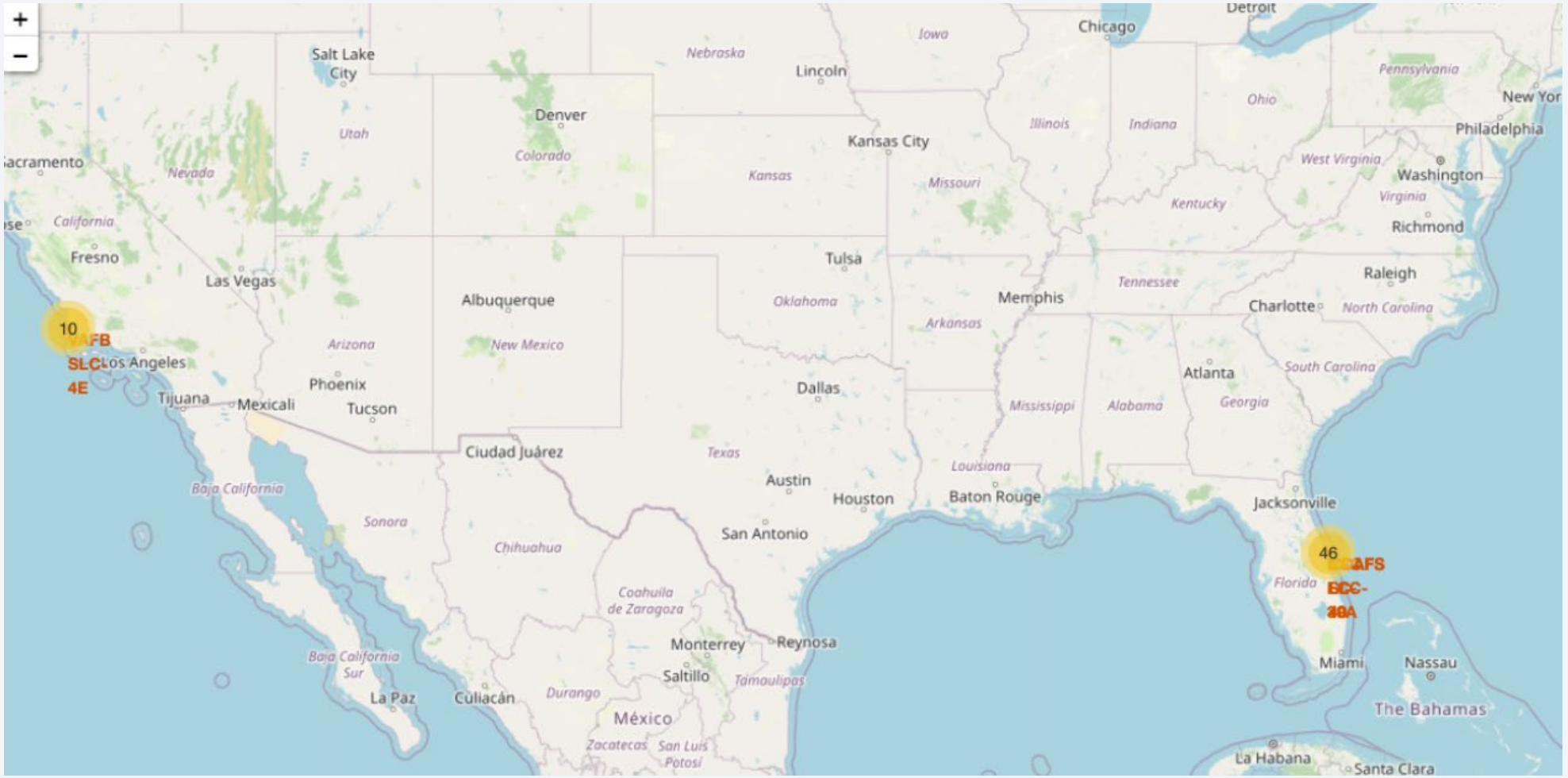
```
 * ibm_db_sa://glq09342:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

t[26]:

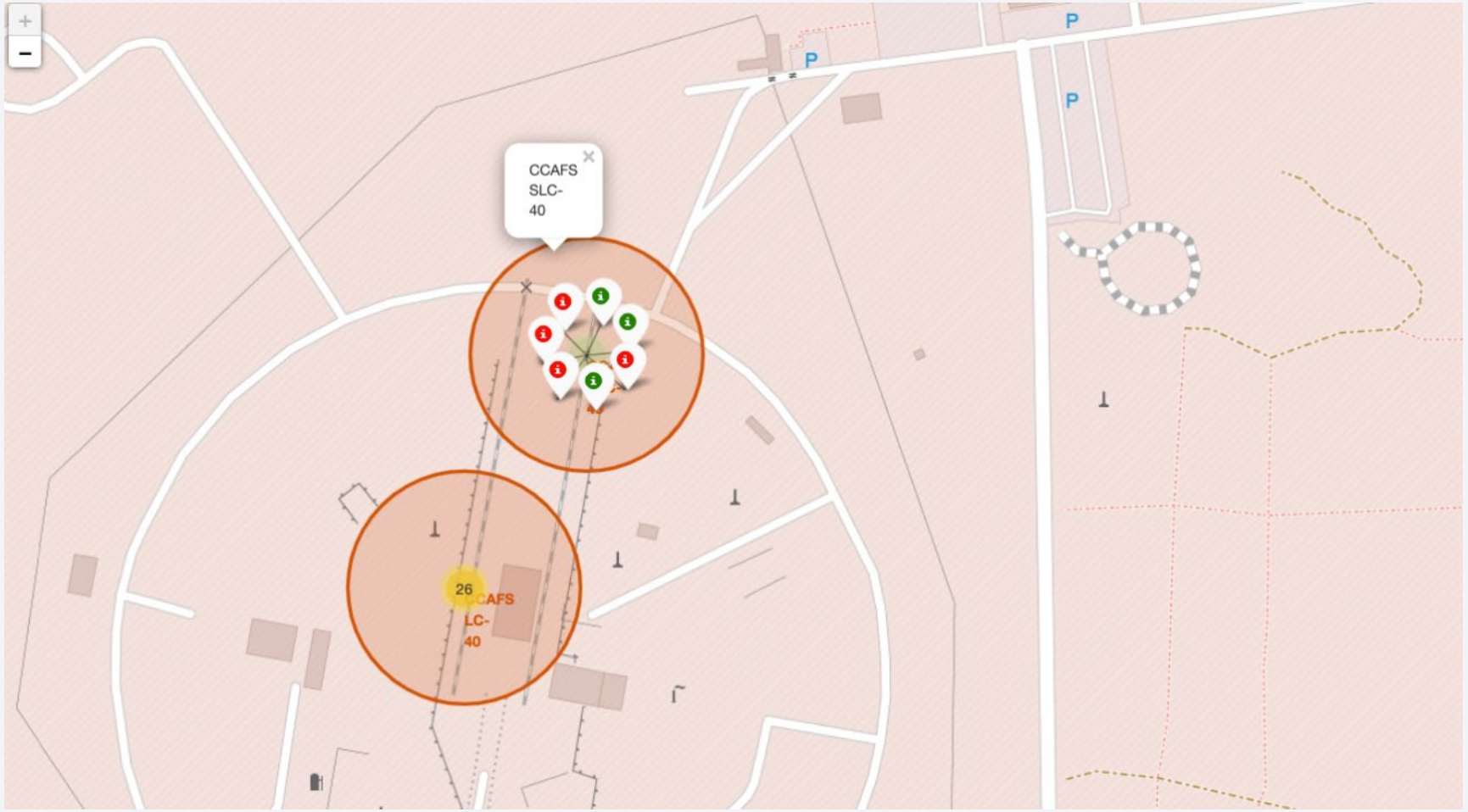
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success
2014-04-18	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success


Section 3
Launch Sites
Proximities Analysis

Launch Sites Locations Analysis with Folium 1



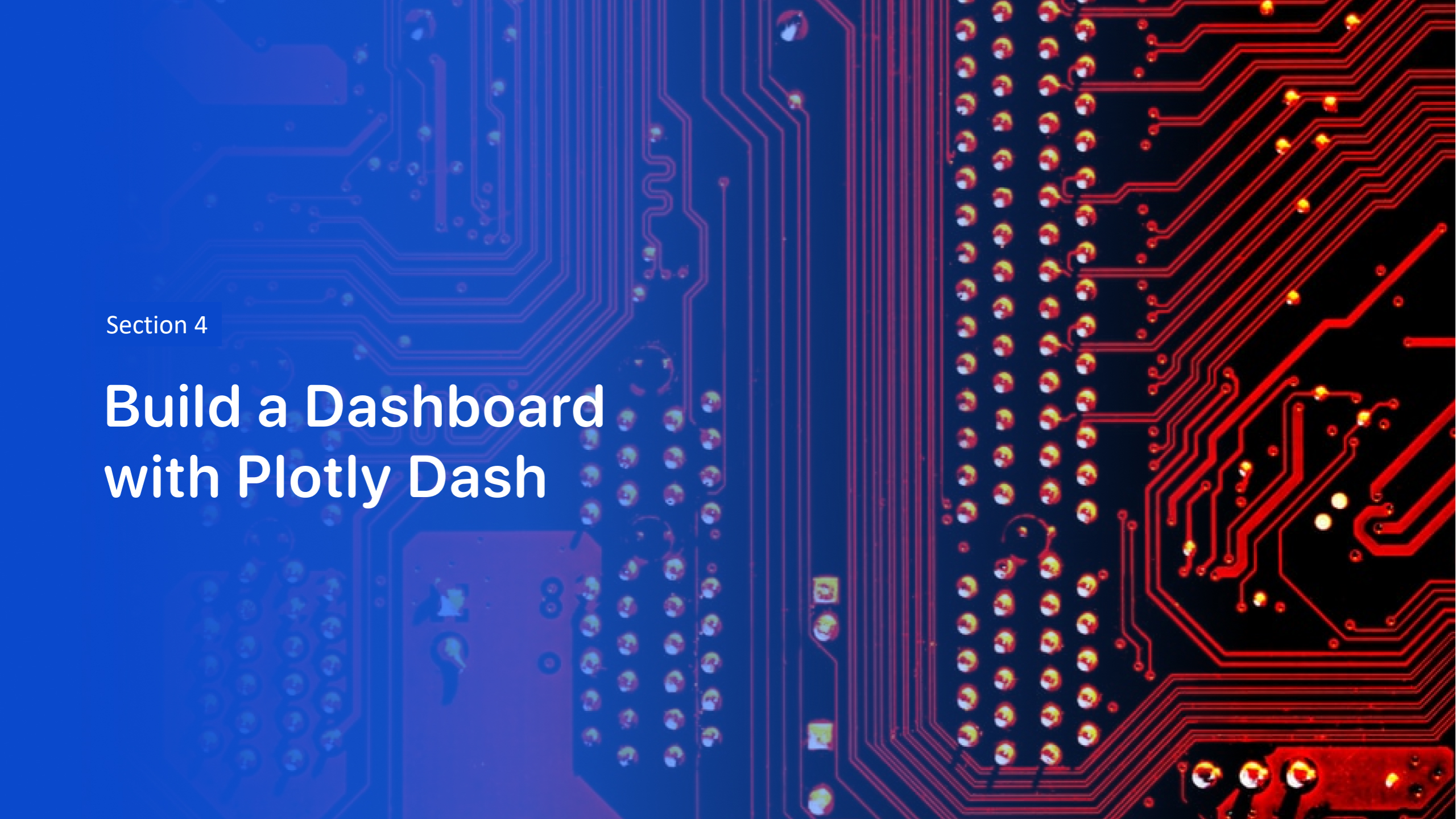
Launch Sites Locations Analysis with Folium 2



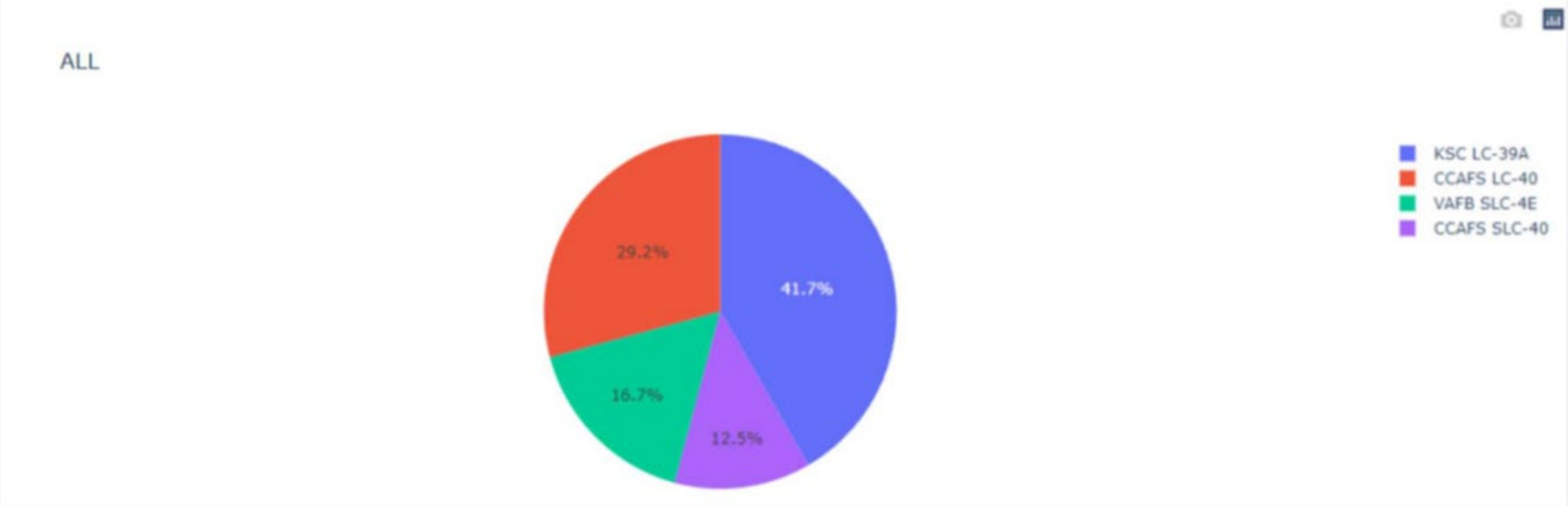
- Green Marker shows success; and red marker shows failed.

Launch Sites Locations Analysis with Folium 3

- Launch site close to the ocean and railway

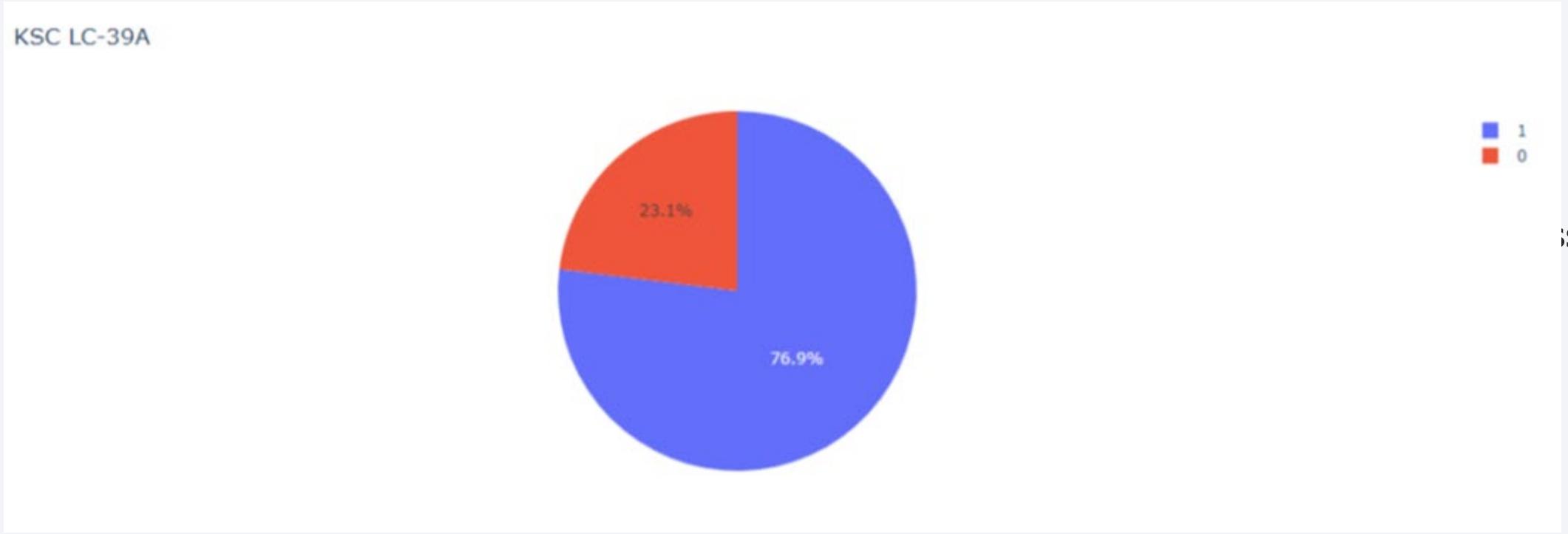

Section 4
Build a Dashboard with Plotly Dash

Interactive visual analytics on SpaceX launch 1



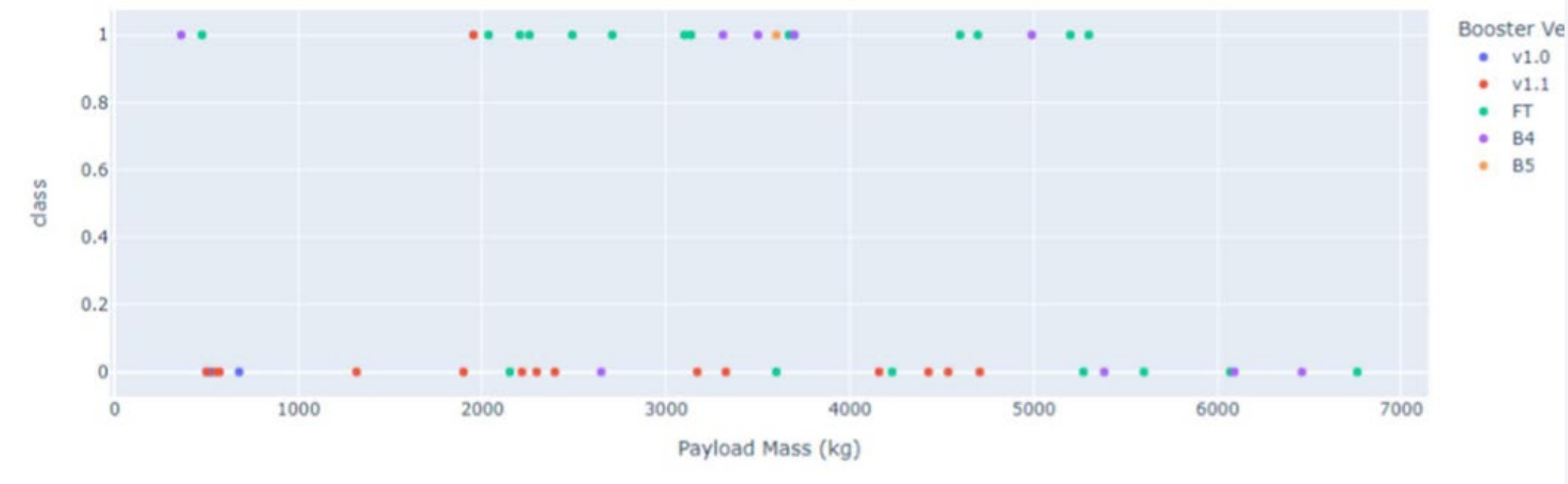
- KSCLC39A has highest number of successful launch with booster landing

Interactive visual analytics on SpaceX launch 2



- 76.9% of launches has successful landings at KSCLC-39A

Interactive visual analytics on SpaceX launch 3



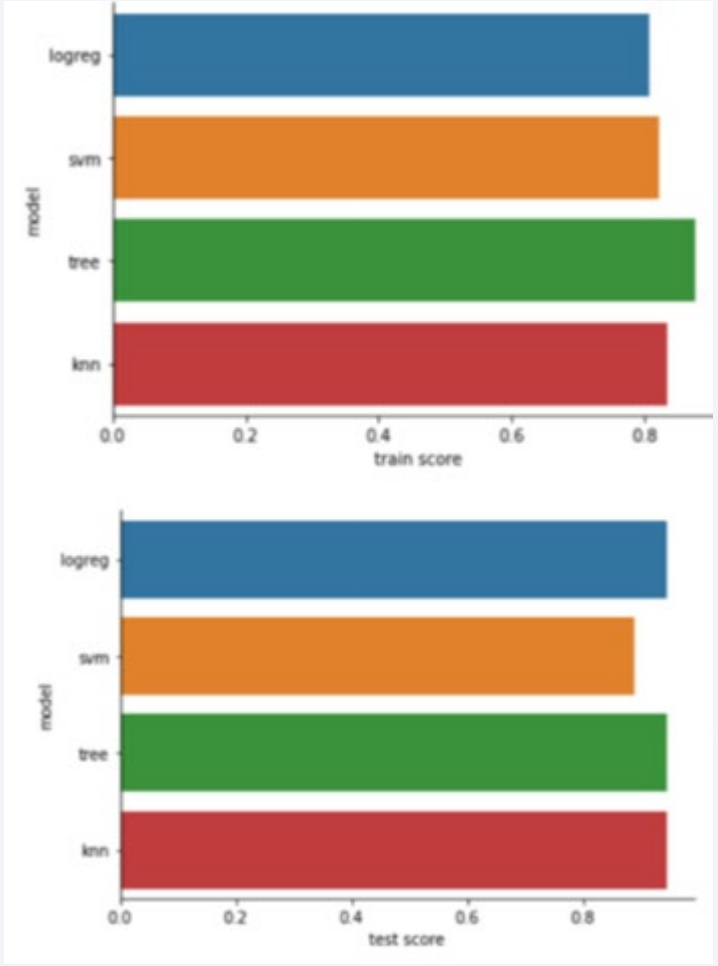
- Between 2000 to 4000 kg payload has higher success rate than launches with size above 4000 kg.

Section 5

Predictive Analysis (Classification)

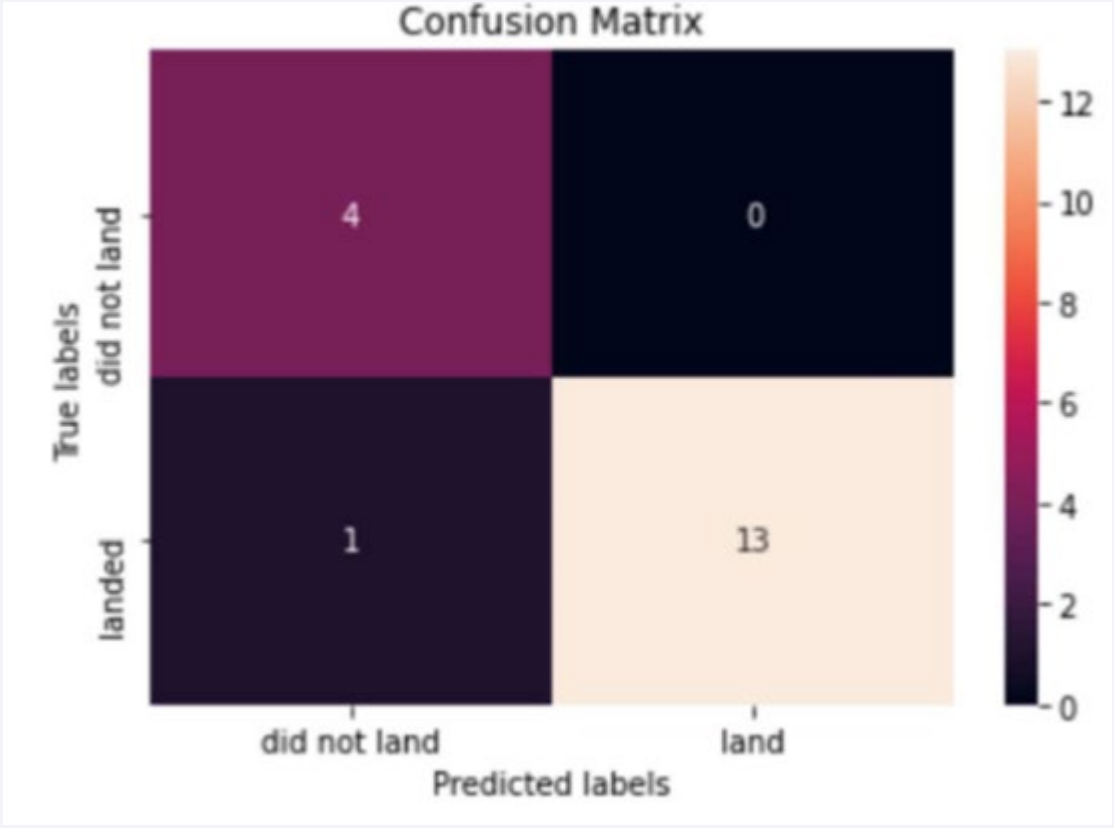
Classification Accuracy

Decision Tree model has the highest classification accuracy



Confusion Matrix

- The model correctly predicted 13 landings and 4 failed landing outcomes. The model incorrectly labeled one landing as a failed landing outcome.



Conclusions

- Decision tree was determined to be the best predictive
- EDA showed the flight numbers, launch site, orbit type, and payload weight had an impact on landing.
- Launch success rate started to increase in 2013 till 2020.
- Physical parameters may have impact on landing success was not included in the decision tree model.

Thank you!