IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Claudia Nestmeyer
November 22nd 2022

Outline

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

Executive Summary

- Summary of methodologies

  - Data collection with API
  - Data collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction

- Summary of all results

  - Results of Exploratory Data Analysis
  - Interactive Analytics
  - Results of Predictive Analysis

Introduction

- Project background and context

  Rocket launches of the Falcon 9 are advertised with a cost of 62 million dollars by Space X while for other providers it costs up to 165 million dollars each. The difference in cost is mainly because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land successfully, we can determine the cost of a launch. This information can be used if an alternate company wants to compete against Space X in terms of costs for a rocket launch. The goal of the project is therefore to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - Which interaction amongst various features determines the success rate of a successful landing?
  - What operating conditions need to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- 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

The data was collected using API and web scraping methods

- Data collection was performed by using get request to the SpaceX API.
- In the next step, the response content was decoded as a JSON file using .json() function call and turn it into a pandas dataframe using .json_normalize().
- Afterwards, the data was cleaned and checked for missing values, missing values were filled with mean values where applicable.
- More data was collected with the Web Scraping method from Wikipedia for Falcon 9 launch records with BeautifulSoup.
- Launch records were extracted as HTML table, parsed into a table and converted into a pandas dataframe for further analysis.

Data Collection – SpaceX API

- The get request was used to collect data from the SpaceX API, further the requested data was cleaned, basic data wrangling was applied, and the data was formatted

- The URL to the notebook is: https://github.com/yippiyayai/project/blob/master/Data%20Collection%20API%20lab.ipynb

1. Get request for rocket launch data using API

In [6]:
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

In [7]:
```
response = requests.get(spacex_url)
```

2. Use json_normalize method to convert json result to data frame

In [11]:
```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

3. Perform data cleaning and filling in the missing values

In [43]:
```
data_falcon9.isnull().sum()
```

In [41]:
```
# Calculate the mean value of PayloadMass column
mean = data_falcon9["PayloadMass"].mean()

# Replace the np.nan values with its mean value
data_falcon9["PayloadMass"].replace(np.nan, mean, inplace=True)
```

Data Collection - Scraping

- Web scraping was applied to the Website of Falcon 9 launch records with BeautifulSoup
- The table was parsed and converted into a pandas dataframe
- The URL to the notebook is https://github.com/yippiyayai/project/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

1. Apply HTTP get method to request the Falcon 9 rocket launch page

In [4]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

In [5]:
```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
response.status_code
```

2. Create a BeautifulSoup object from the HTML response

In [27]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.text, 'html.parser')
```

In [13]:
```
# Use soup.title attribute
soup.title
```

3. Extract all column names from the HTML table header

In [16]:
```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header() to get a column name
# Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called column_names

element = soup.find_all('th')
for row in range(len(element)):
    try:
        name = extract_column_from_header(element[row])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

4. Create a data frame by parsing the HTML launch tables

Data Wrangling

- Exploratory data analysis was performed, and the training labels determined
- The number of launches at each site, the number & occurrence of each orbit type, and number & occurrence of mission outcome per orbit type were calculated
- A landing outcome label was created from the outcome column and the results exported to a csv file.
- The link to the notebook is: https://github.com/yippiyayai/project/blob/master/Data%20Wrangling:%20EDA%20lab.ipynb

In [5]:
```
# Apply value_counts() on column LaunchSite
df.value_counts('LaunchSite')
```

In [6]:
```
# Apply value_counts on Orbit column
df.value_counts('Orbit')
```

In [7]:
```
# landing_outcomes = values on Outcome column
landing_outcomes = df.value_counts('Outcome')
```

In [9]:
```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

In [15]:
```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []

for key, value in df['Outcome'].items():
    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

EDA with Data Visualization

- We visualized the data to see the relationship between
  - flight number and launch site
  - payload and launch site
  - success rate and orbit type
  - flight number and orbit type
  - payload and orbit type
  - launch success throughout the years
- This was done to see if we can see which factors have an influence on a successful landing outcome.
- This is the link to the notebook: https://github.com/yippiyayai/project/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- EDA with SQL was applied to get various insights into the data. With the help of queries, we got information on:
  - The names of unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date of the first successful outcome in a drone ship was achieved
  - The booster names of success in ground pad with a mass between 4000 & 6000kg
  - The total number of successful & failure mission outcomes
  - The names of booster versions with maximum payload mass
  - All successful landing outcomes in ground pad for 2017 with detailed information
- The link to the notebook is: https://github.com/yippiyayai/project/blob/master/jupyter-labs-eda-sql-edx_rev2.ipynb

Build an Interactive Map with Folium

- All launch sites were marked on the map, additionally all successful/failed launches for each site were marked on the map, lines were added to measure distance to its proximities
- Using the color-labeled marker cluster, we could identify which launch sites have a relatively high success rate
- Through calculating the distance between a launch site and its proximities, the following questions could be answered:
  - Are launch sites highways, railways and coastlines?
  - Are launch sites situated far away from cities?
- The link to the notebook is: https://github.com/yippiyayai/project/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was built with Plotly dash
- Pie charts were plotted showing the total launches for each site, and showing the total number of successful/failing landing outcomes by a certain site, to see which launch site has the highest success rate and to see how many launches were done at each site
- A scatter graph was plotted showing the relationship between Landing Outcome and Payload Mass (kg) for different booster versions, to see if the Landing Outcome is influenced by the Payload mass
- The link to the notebook is: https://github.com/yippiyayai/project/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was loaded using numpy and pandas, then transformed and splitted into training and testing data
- Different machine learning models were built and different hyperparameters tuned using GridSearchCV
- Accuracy was used as the metric for our model, and the model was improved using feature engineering and algorithm tuning
- One of the models was found to be the best performing classification model in our project
- The link to the notebook is: https://github.com/yippiyayai/project/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

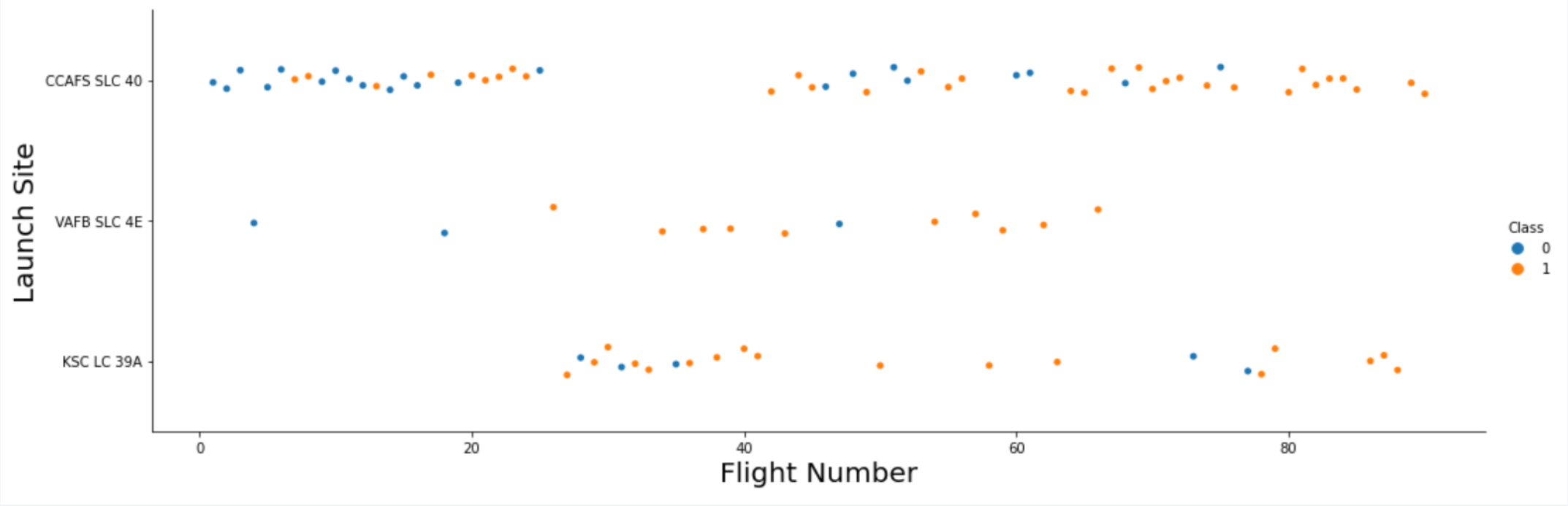
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

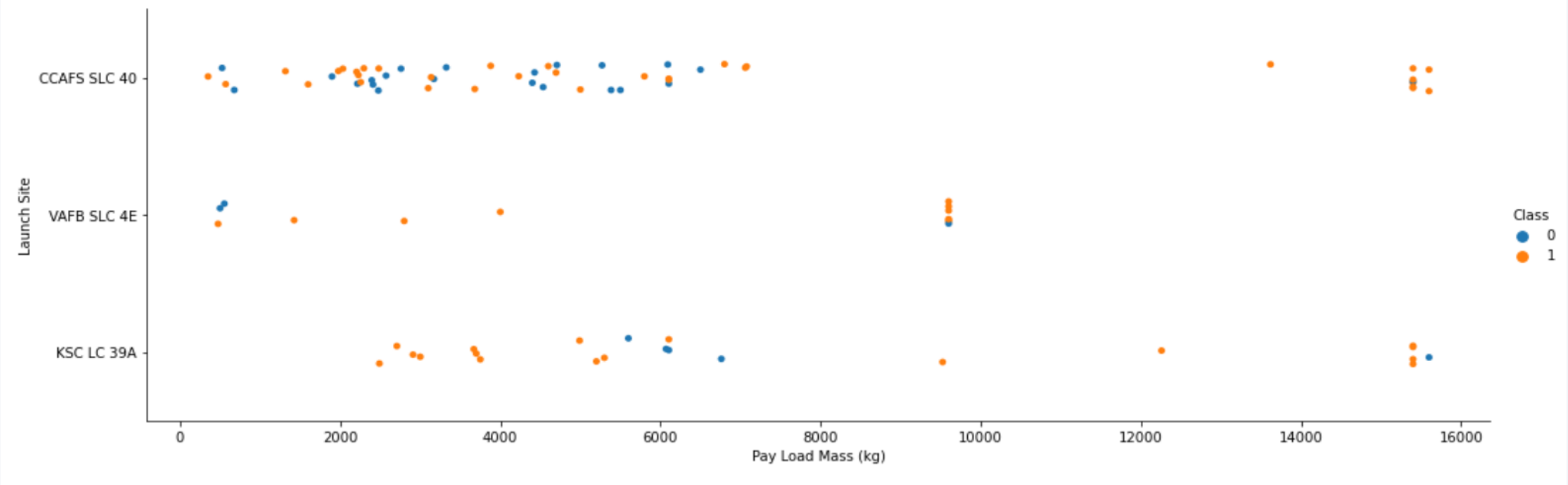
It can be observed that success rate increases with the increasing number of flights at each launch site.



Payload vs. Launch Site

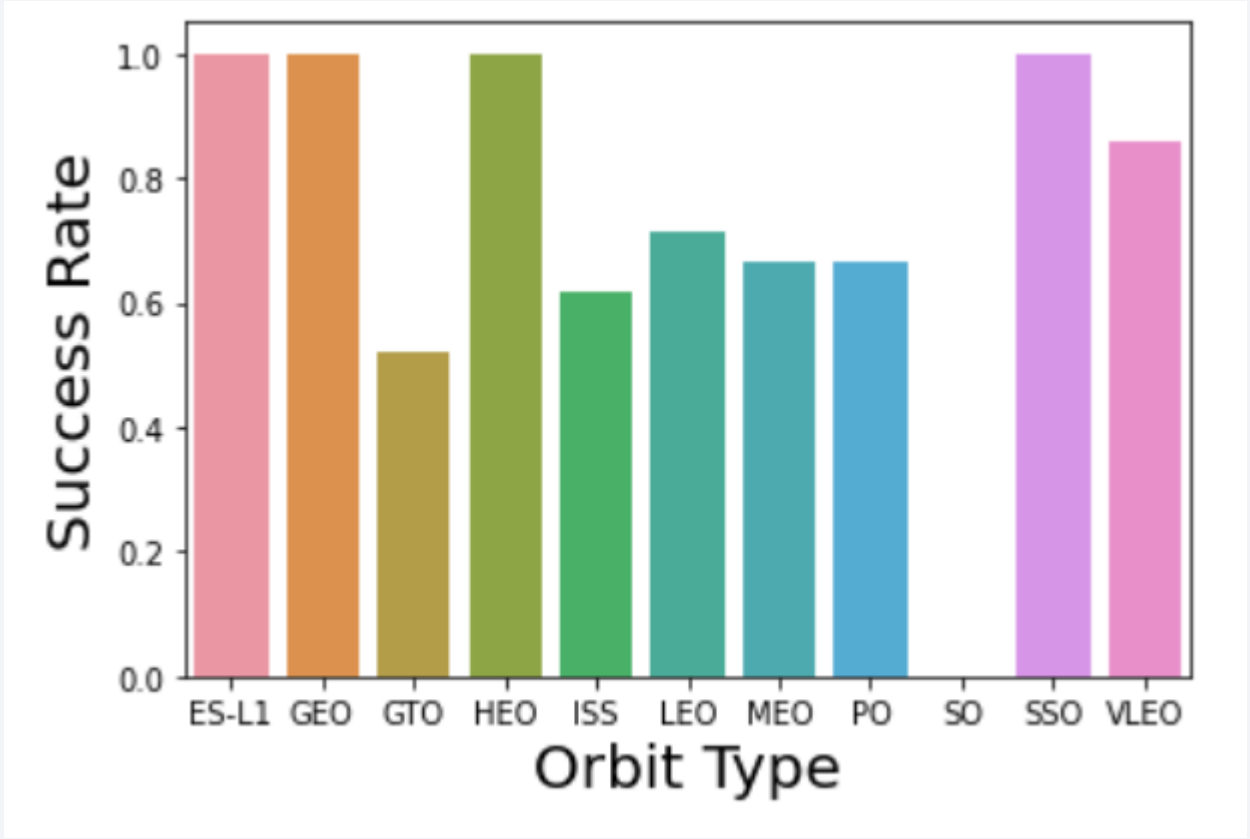
It can be observed that at the VAFB SLC 4E there are no rockets launched with a heavy payload mass over 10000kg.

At the CCAFS SLC 40 the success rate increases with greater payload mass of the rocket.



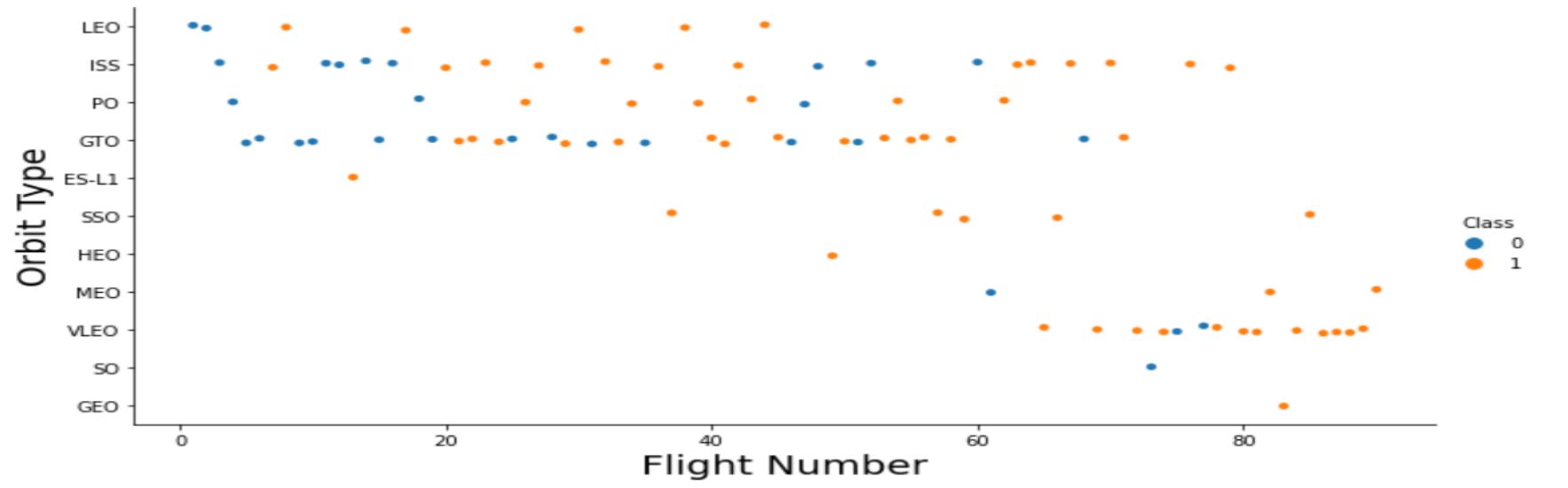
Success Rate vs. Orbit Type

- Highest success rate (100%) for the orbit types ES-L1, GEO, HEO and SSO
- High success rate (>80%) also for orbit type VLEO
- Lowest success rate (0%) for orbit type SO



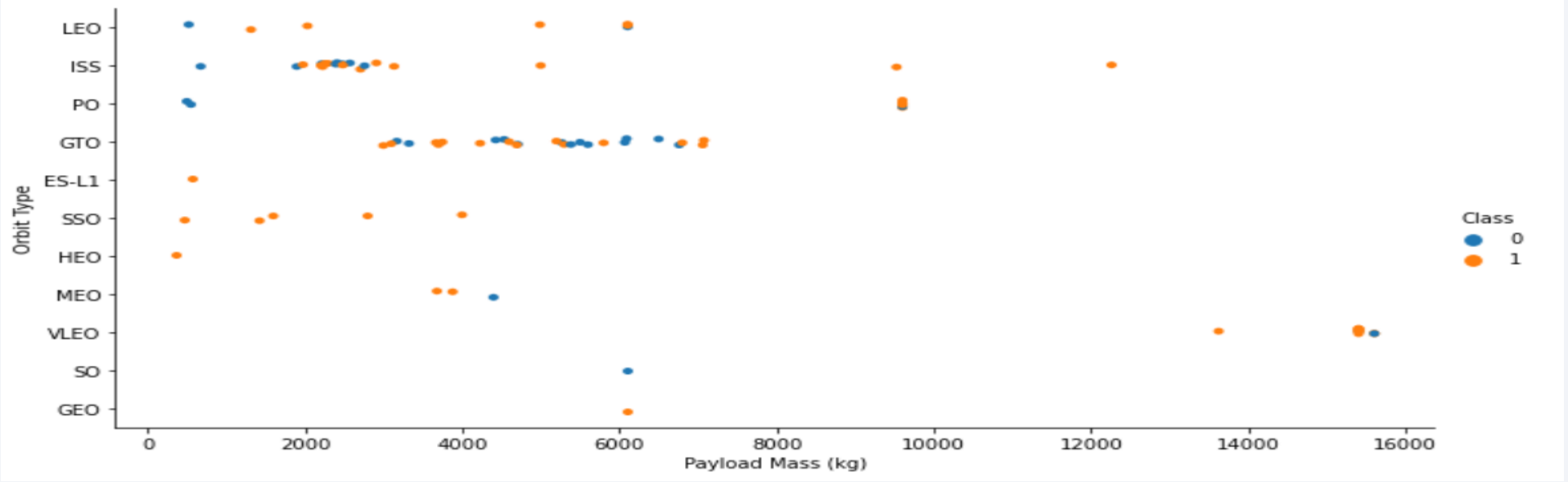
Flight Number vs. Orbit Type

- In orbit type LEO success is related to the number of flights, whereas in the GTO orbit, no relationship can be found between flight number and orbit type.
- No correlation between flight number and orbit type can be examined for orbit types ES-L1, HEO, GEO, and SO, due to lack of flights to those orbits.



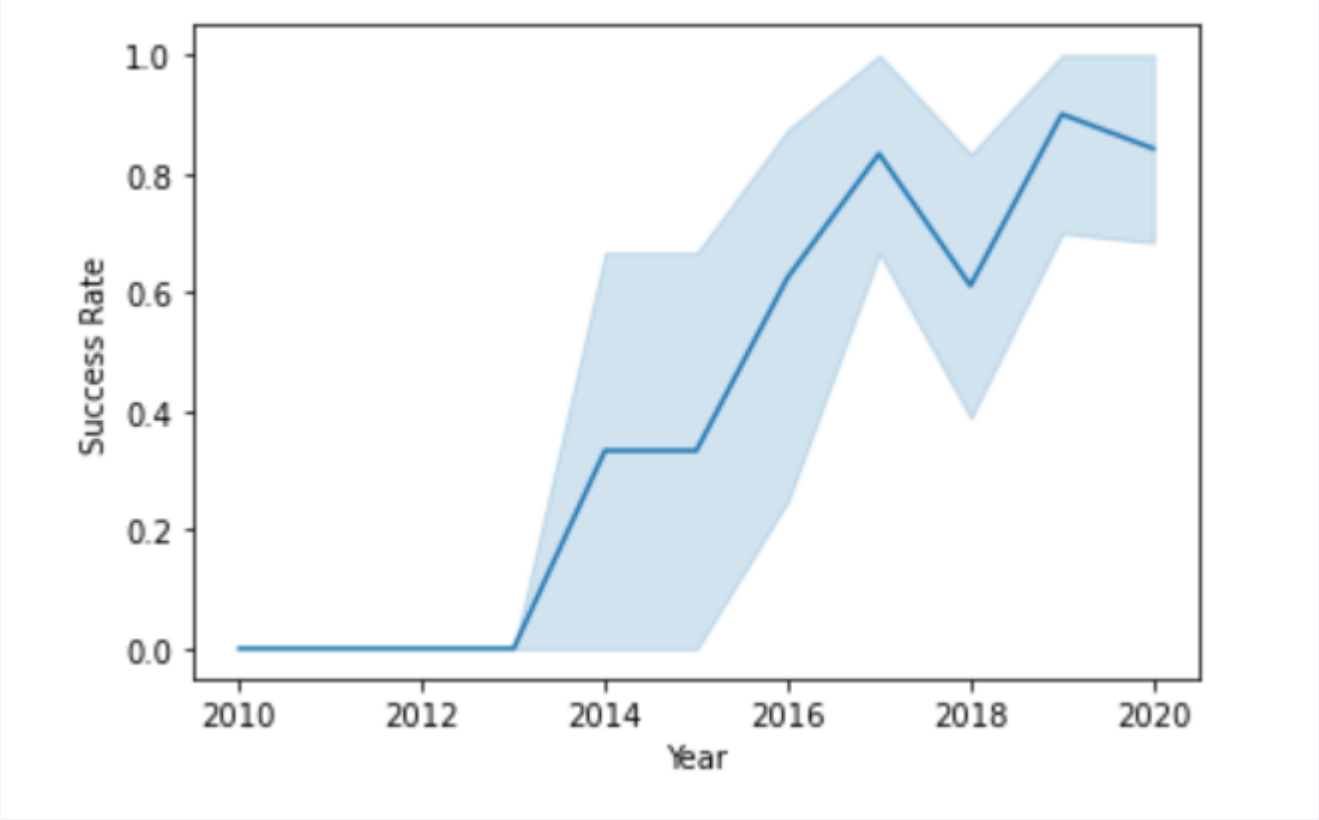
Payload vs. Orbit Type

- With heavy payloads the successful landing rate are more for PO, LEO, and ISS.
- For GTO on the other hand, no correlation between payload mass and orbit type can be found as both successful and unsuccessful landings can be found over whole span of different payload masses.



Launch Success Yearly Trend

It can be observed that the success rate for launches increased since 2013 until 2020.



All Launch Site Names

The key word DISTINCT together with the SELECT command was used to show only the unique launch sites from the SpaceX data set.

Task 1

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

```
%sql SELECT DISTINCT Launch_site FROM SPACEXDATASET
```

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

Launch Site Names Begin with 'KSC'

Task 2

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

```
%sql SELECT * FROM SPACEXDATASET WHERE Launch_site LIKE 'KSC%' LIMIT 5
```

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

The Query above was used to display 5 records of launches at the launch site that starts with ‘KSC’.

Total Payload Mass

Task 3

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

```
%sql SELECT SUM(Payload_Mass__Kg_) FROM SPACEXDATASET WHERE CUSTOMER = 'NASA (CRS)'
```

1
45596

The total payload mass carried by boosters launched by NASA (CRS) was calculated as 45596 kg using the query above.

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG(Payload_Mass__Kg_) FROM SPACEXDATASET WHERE Booster_Version = 'F9 v1.1'
```

1
2928

The average payload mass carried by booster version F9 v1.1 was calculated as 2928 kg with the query above.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql SELECT min(Date) FROM SPACEXDATASET WHERE Landing__outcome = 'Success (drone ship)'
```

1
2016-04-08

The first successful landing outcome for a drone ship was achieved on the 8th of April 2016.

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

```
%sql SELECT Booster_Version FROM SPACEXDATASET WHERE Payload_Mass__Kg_ between 4000 and 6000 AND landing__outcome = 'Success (ground pad)'
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- With the combination of the WHERE clause and the AND condition the boosters could be detected which have success in ground pad and have a payload mass between 4000 kg and 6000 kg.
- The list above shows the three booster in fulfilling both conditions.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(Mission_outcome) AS Mission_outcome FROM SPACEXDATASET WHERE Mission_outcome LIKE '%Success%' UNION SELECT COUNT(Mission_ou
```

mission_outcome
1
100

The total number of successful mission outcomes is 100, while the total number of failure mission outcomes is only 1.

Boosters Carried Maximum Payload

Task 8

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

```
%sql SELECT Booster_Version FROM SPACEXDATASET \
WHERE Payload_Mass__Kg_ = (SELECT MAX(Payload_Mass__Kg_) FROM SPACEXDATASET)
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- The table on the left shows all booster versions which varied the maximum payload mass.
- ´The maximum payload mass for booster versions was determined by using a subquery with the MAX() function in the WHERE clause.

Successful Landing Outcomes (ground pad) 2017

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

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'Month') AS MONTH_NAME,\
Landing__outcome AS Landing__outcome, \
Booster_Version AS Booster_Version, \
Launch_site AS Launch_site \
FROM SPACEXDATASET WHERE Landing__outcome = 'Success (ground pad)' AND "DATE" LIKE '%2017%'
```

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

- With WHERE, AND and LIKE the successful landings for ground pad in 2017 were filtered and with TO_CHAR brought into a nice table showing the month name, outcome, booster version and launch_site

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.

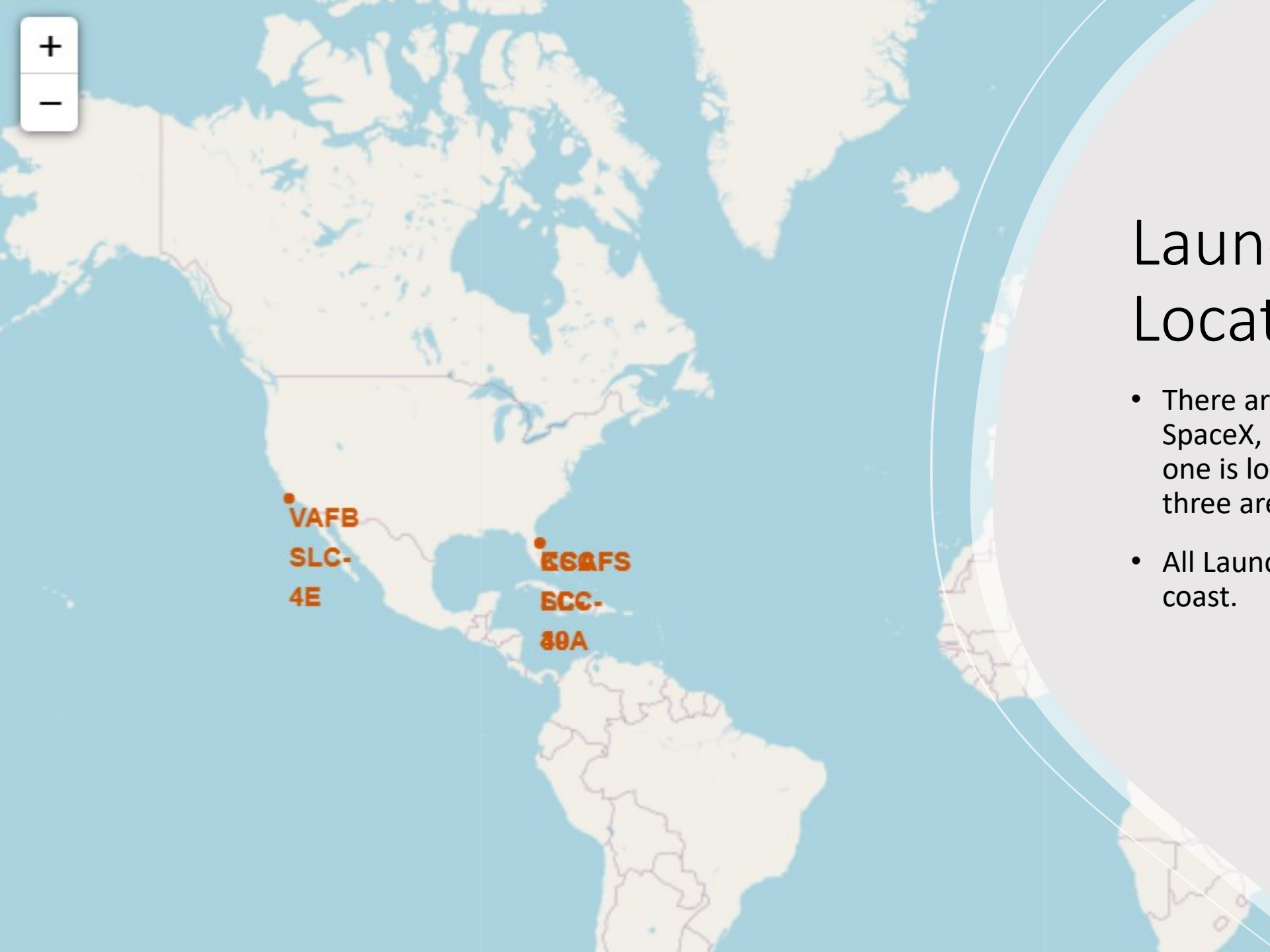
```
%sql SELECT "DATE", COUNT(Landing__outcome) AS COUNT FROM SPACEXDATASET \
WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND Landing__outcome LIKE '%Success%' \
GROUP BY "DATE" ORDER BY COUNT(Landing__outcome) DESC
```

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

The first successful landing outcome was first achieved in 2015.

Section 3

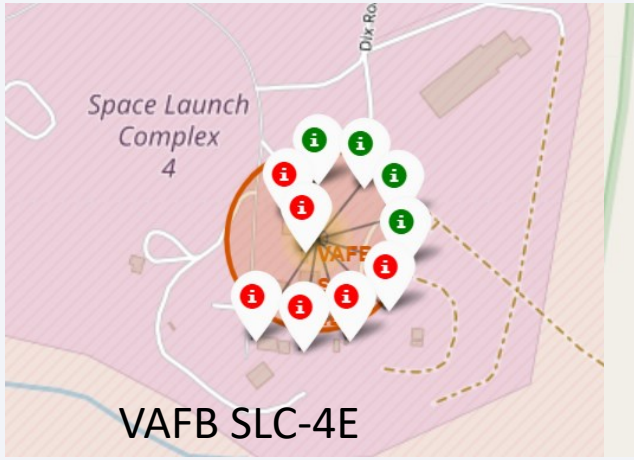
Launch Sites Proximities Analysis

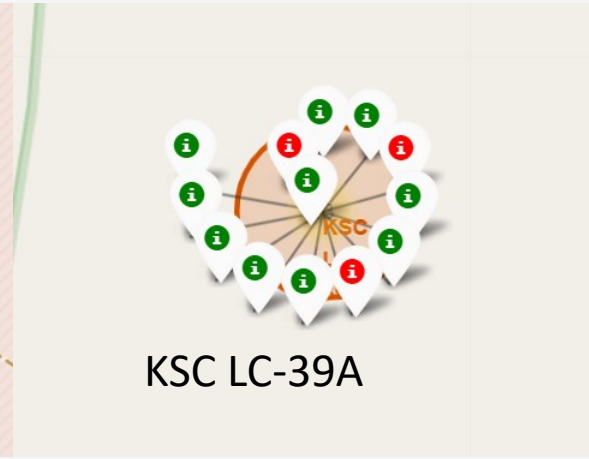


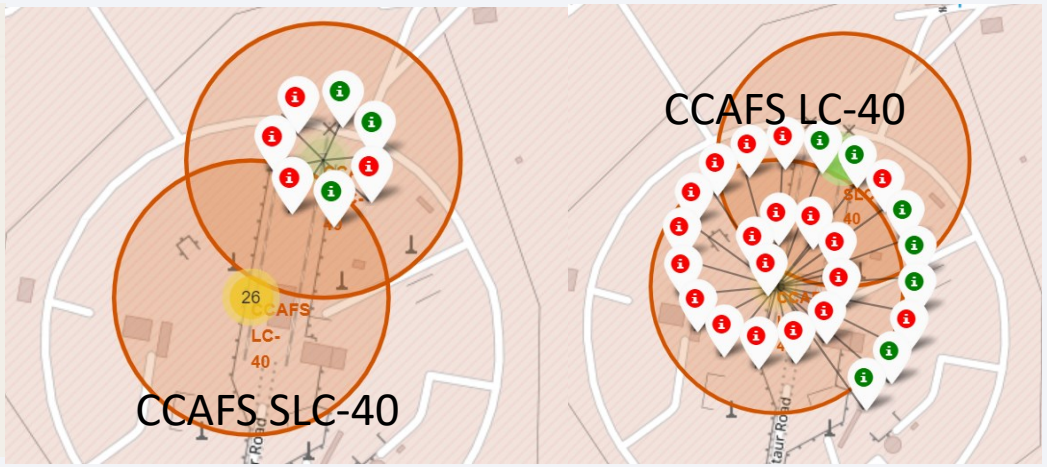
Launch Site Locations

- There are four Launch Sites of SpaceX, all are located in the US, one is located in California and three are located in Florida.
- All Launch Sites are near the coast.

Launch Outcomes per Launch Site

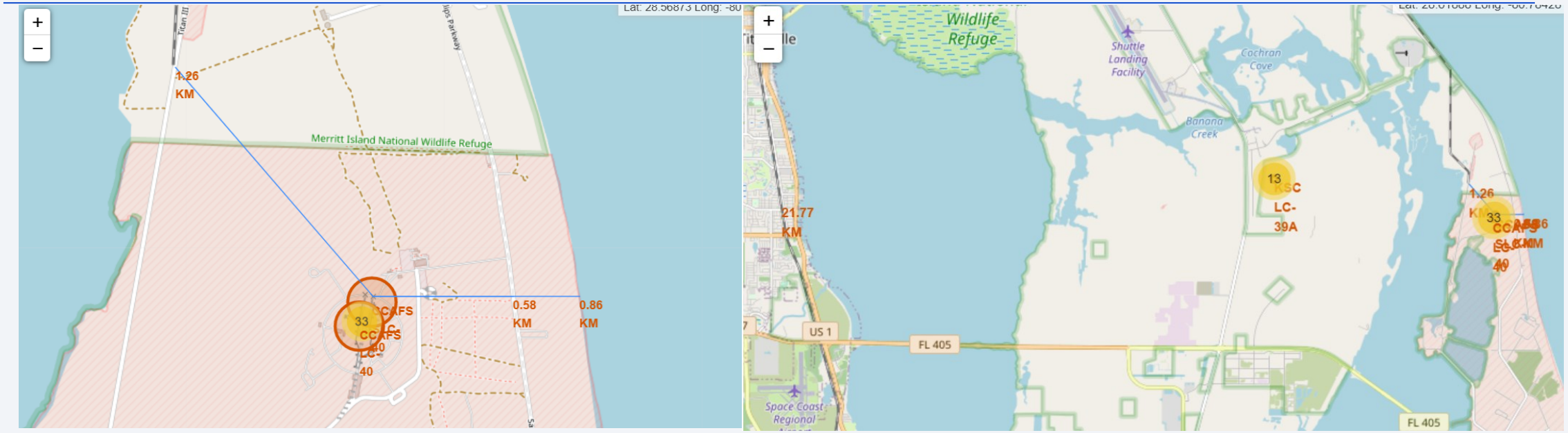




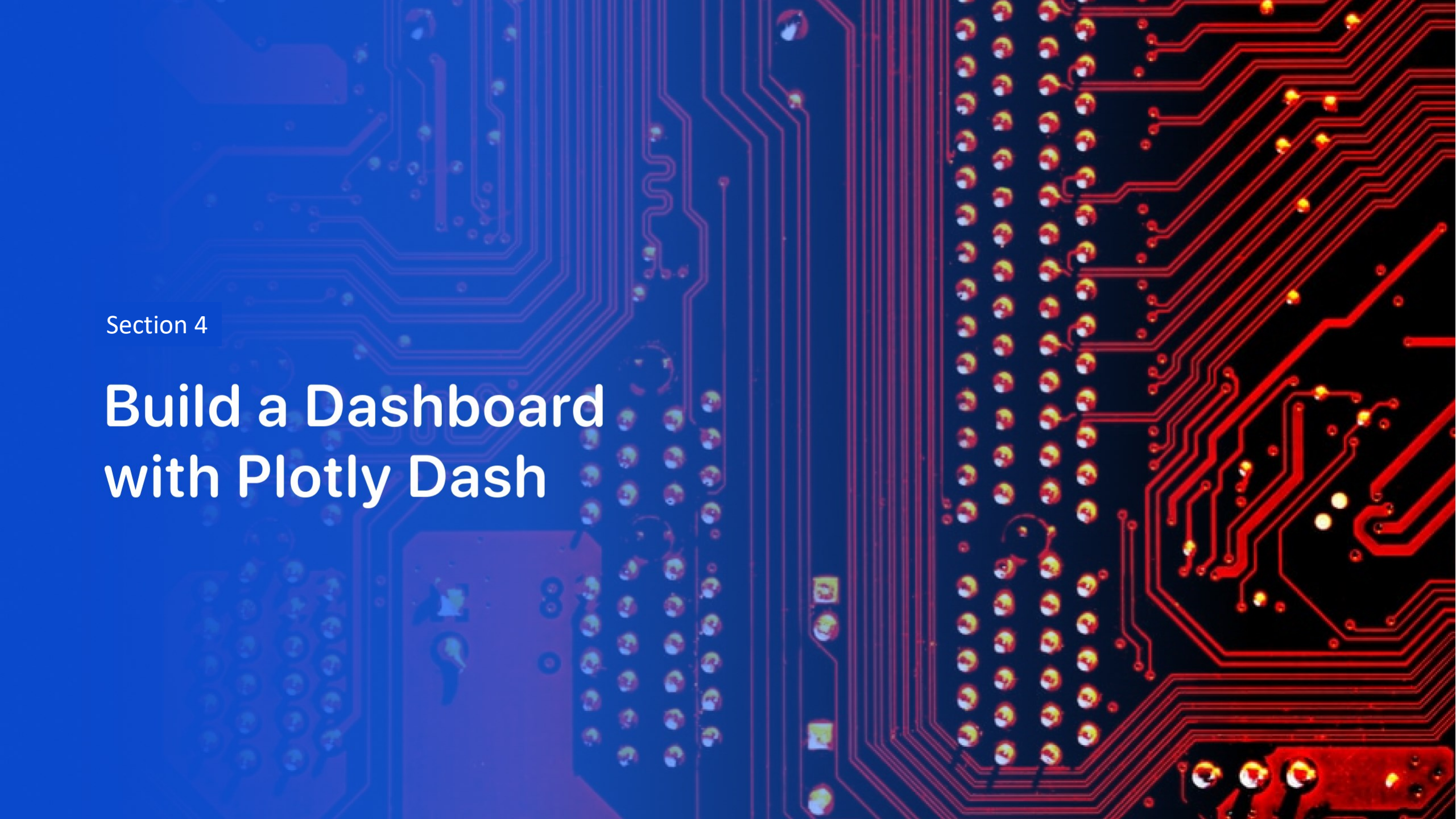


The red markers show the failed launches, and the green markers show the successful launches.

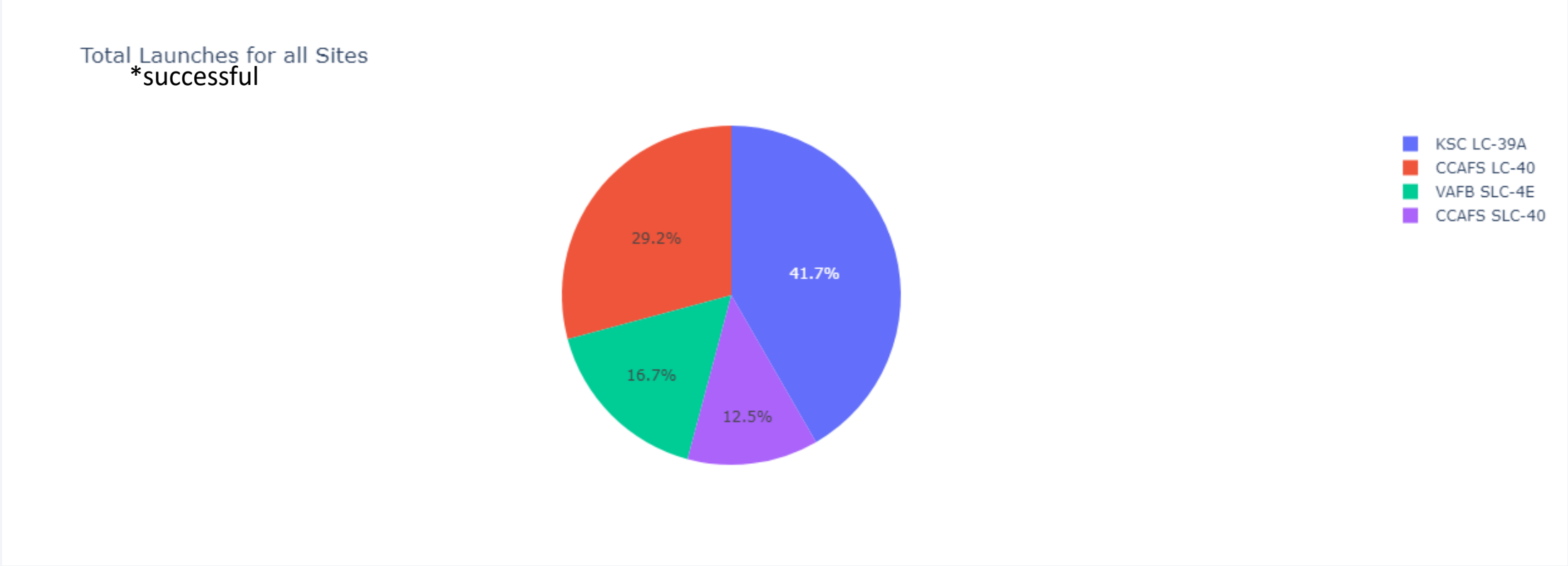
The Launch Site to its proximities



The distance from the launch site to the coast, railway, and highway are quite short, while they are located far away from cities.

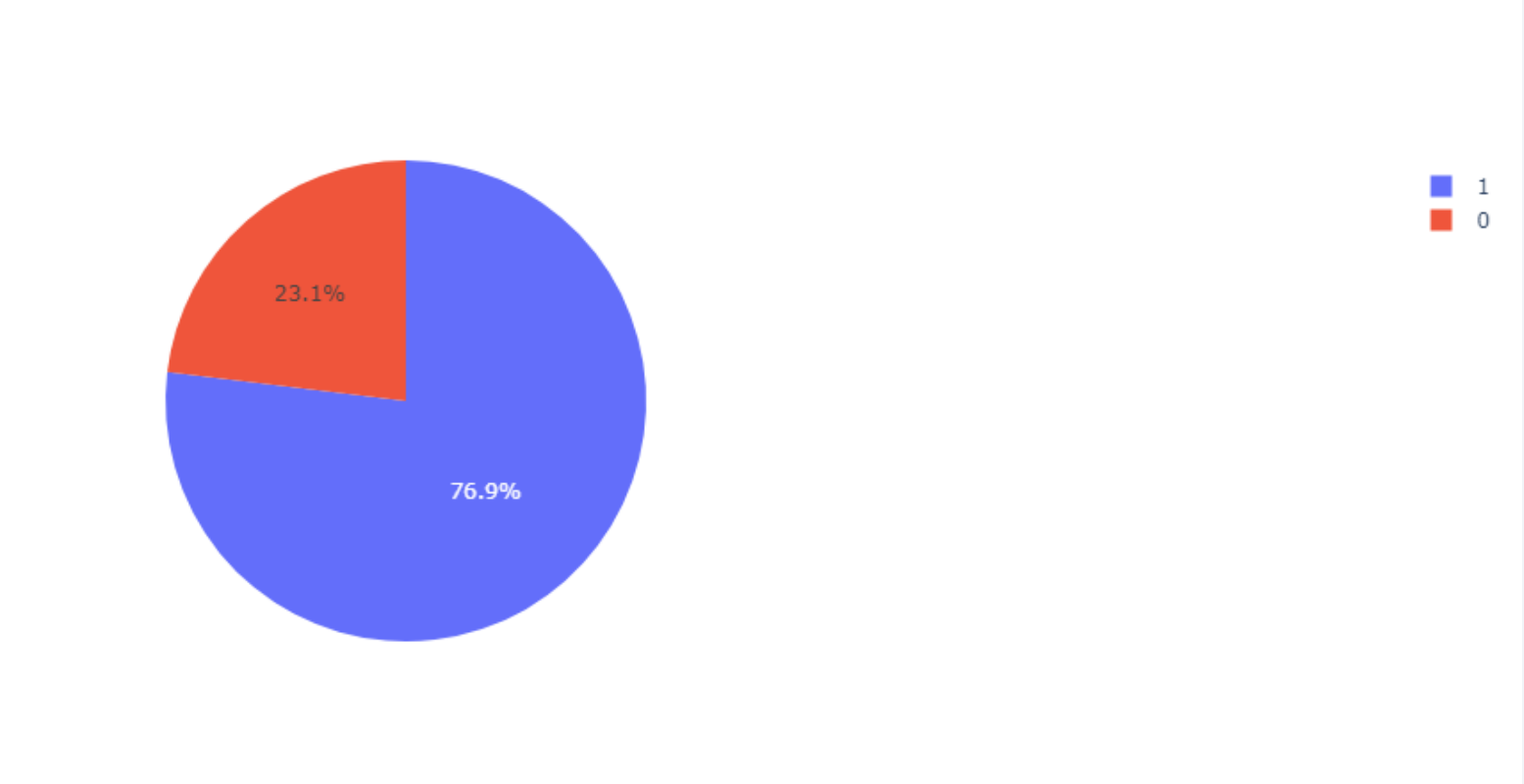

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



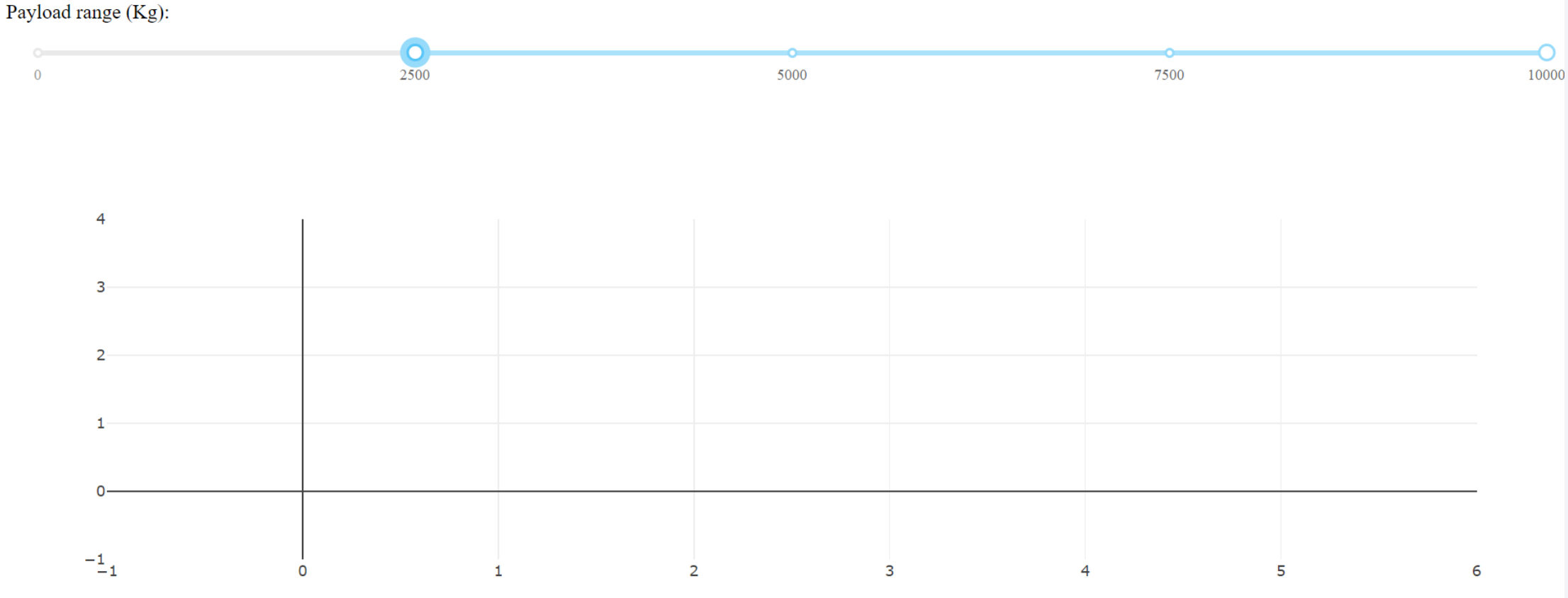
The pie chart is showing the total successful launches for all sites. We can see that KSC LC-39A has the highest success rate of all launch sites.

Pie chart showing the launch site with the highest success rate



The KSC LC-39A achieved a success rate of 76.9% while getting a failure rate of 23.1% for its launches.

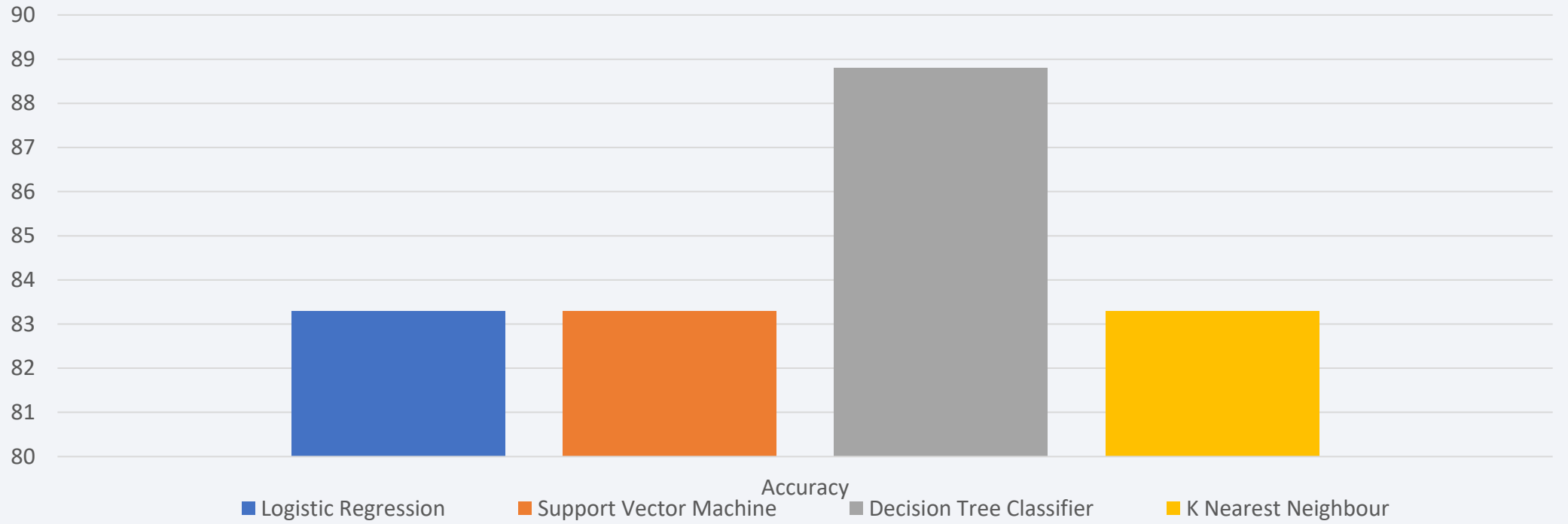
Scatter plot of Payload vs Launch Outcome for all sites for different payloads



Section 5

Predictive Analysis (Classification)

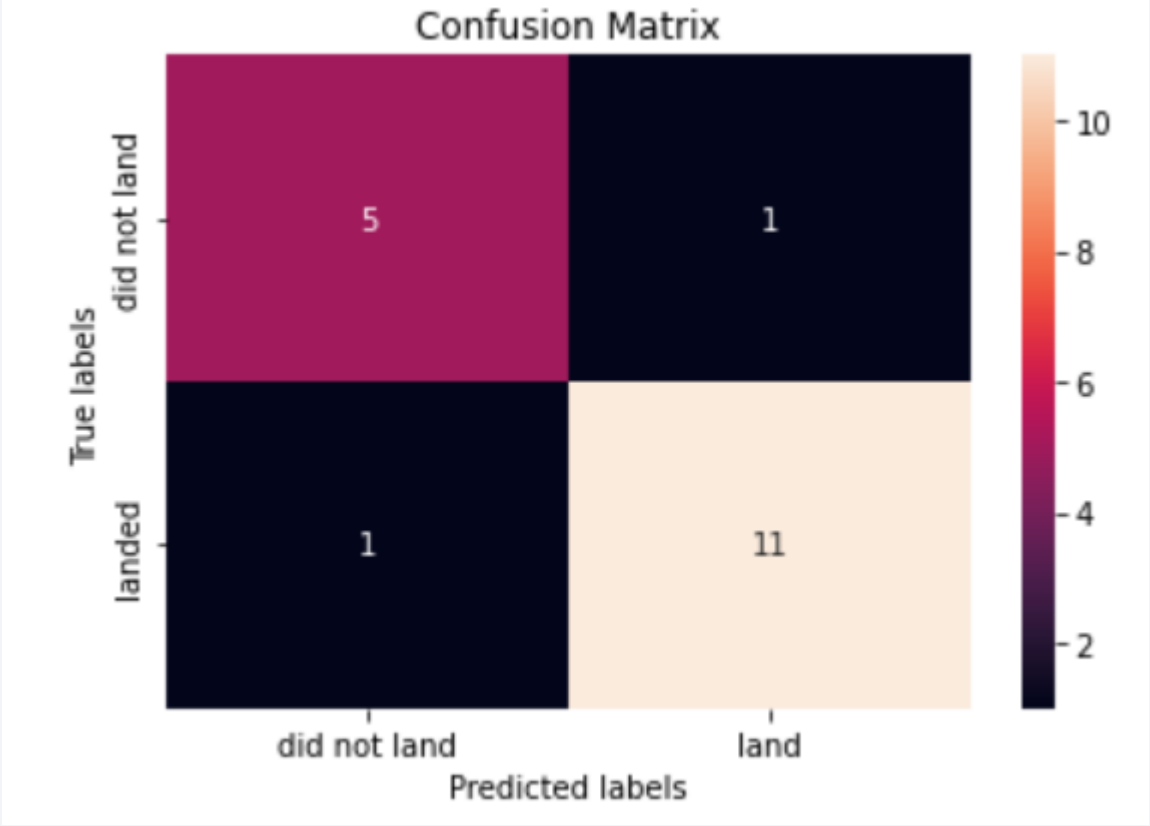
Classification Accuracy



We can see that the Decision Tree Classifier has the highest accuracy with 88.8%.

Confusion Matrix

The decision tree model had the highest accuracy. The confusion matrix shows that the model is able to distinguish between the different classes. Only 2 out of 18 values were predicted wrong (1 false positive, 1 false negative) which is a very good result, because a small error rate increases the reliability of the model's results.



Conclusions

- Since 2013 the success rate was steadily increasing
- KSC LC-39A has the highest success rate of launches compared to the other sites.
- In regard to the number of flights the orbits with the highest success rate are LEO, SSO and VLEO.
- For rockets with a heavy payload mass, orbit types PO, LEO and ISS have a higher success rate
- At launch site CCAFS SCC-40 success rate is increasing with greater payload mass.
- The decision tree classifier model is the best machine learning algorithm for this task.

Thank you!