IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

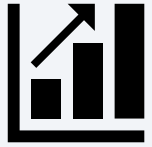
Fabian Zemke
25.08.2023

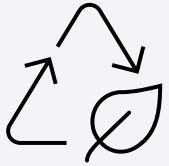
Outline

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

Introduction

- Space travel is an upcoming topic of importance.
- The Space Exploration Technologies Corp. (SpaceX) accelerated this topic.
- Cost-efficiency due to reusibility leads to competition.





- Here, hypothetical company SpaceY cost calculation.
- Machine learning (ML) if a rocket of SpaceX can be reused.
- How should SpaceY set its price to be competitive?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API / https://api.spacexdata.com/v4/
  - Webscraping of Wikipedia pages.
- Perform data wrangling
  - Data might be in JSON format and needs normalization.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

Option 1:

- SpaceX launch data from SpaceX REST API.
  - Payload, launch and landing specifications, landing outcome.
  - https://api.spacexdata.com/v4/launches/past.
- Usage of the *requests* python library.

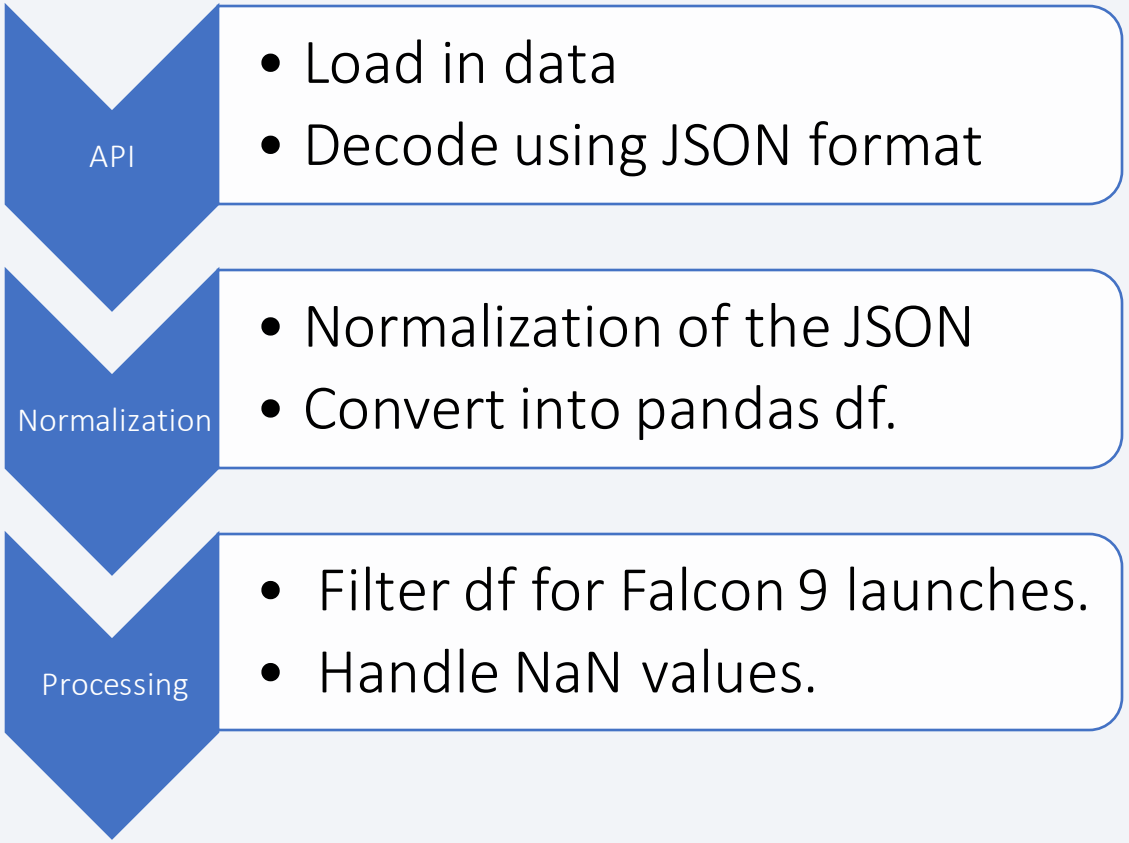
Option 2:

- Webscraping the Wiki pages via pythons *BeautifulSoup*.
- Further handling using python *pandas*.

Data Collection – SpaceX API

- Using the SpaceX API (https://api.spacexdata.com/v4/launches/past) the data is loaded in JSON format.
- The data is normalized: df_json = pd.json_normalize(json).
- The df consists of the "BoosterVersion", "PayloadMass", "Orbit", "LaunchSite", "Outcome", "Flights", "GridFins", "Reused", "Legs", "LandingPad", "Block", "ReusedCount", "Serial", "Longitude", and "Latitude".
- The df is filtered to only include Falcon 9 launches.
- NaN values are replaced by the mean of its column.

Github link:
https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%201%20-%20Data%20Collection%20API.ipynb



Data Collection - Scraping

- Using *BeautifulSoup* a text is extracted from the HTML of the "List of Falcon 9 and Falcon Heavy launches" Wikipedia page.
- For better visibility, *prettify()* is used.
- All tables of the webpage are extracted and the third one chosen, which contains the launch records.
- Column names are extracted.
- Creating a dataframe by first using a dictionary with the data.
- Github link (might require some loading time): https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%201%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The data is explored to find parameters of interest for further investigation.
- The percentage of missing values is determined.
- The type of the individual columns is shown.
- Value counts of the launch sites are evaluated.
- The successful and failed landing outcomes are set to 1 and 0, respectively.

Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%201%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- A scatter plot (*sns.catplot*) of the Flight number and Payload mass show the relevance of a higher FlightNumber and lower PayloadMass on successful landing.
- The scatter plot (*sns.catplot*) of FlightNumber and LaunchSite, indicate that site CCAFS SLC 40 has a high dependency of the Flight Number and the most Launches, whereas KSC LC-39A has a high success rate.
- Another scatter plot (*sns.scatterplot*) of PayloadMass and LaunchSite shows that fewer launches occur with a higher payload.
- A bar chart demonstrates the highest success rate for Orbits: ES-L1, GEO, HEO, and SSO.
- The relationship between the flight number and Orbit becomes apparent for the scatterplot of the two, where LEO is seemingly dependent on the Flight number, whereas GTO and ISS are not. Although, the data points of LEO are fewer.
- The scatterplot of PayloadMass and Orbit shows that some Orbits were tested with higher payload. Although, a direct correlation is difficult to conclude since one mass cannot be compared with another.

Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%202%20-%20EDA%20with%20Visualization.ipynb

EDA with SQL

- The SQL database is loaded into the Jupyter notebook.
- The names of all individual launch sites are listed.
- Five records are shown where the launch site starts with KSC are shown.
- The total payload launched by NASA is calculated.
- The average payload mass of Booster version F9v1.1 is determined.
- The first successful landing outcome is shown.
- All names of boosters which successfully landed and had a certain payload mass are listed.
- Furthermore, the total successful and failed mission outcomes are shown.
- The booster version which handled the maximum payload mass is evaluated.
- The months, successful landing outcomes in ground pad, with the booster versions and launch sites are listed for 2017.
- A list of landing outcomes between a certain date are given in descending order.
- Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%202%20-%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- A *folium.Circle* and *folium.Marker* object was used to show the coordinates of the individiual launch sites on a map.
- A MarkerCluster() object was used to summarize/simplify the individual successful and failed landings.
- The distance to coastlines, railways, highways, and cities was visualized with a PolyLine.
  - While coastlines are great to securely detonate the payload, and railways are great to transport heavy equipment.
  - Highways and the distance to cities must be considered to avoid casualties.

Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%203%20-%20InteractiveVisualAnalyticswithFolium.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created with plotly.
- A dropdown menu allows for selecting specific launch sites.
- The successful landings are shown in a pie chart.
- Furthermore, a scatter plot is used to show the Payload mass and Booster version category for the selected launch sites.
- Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%203%20-%20Interactive%20Dashboard%20Ploty%20Dash.py

Predictive Analysis (Classification)

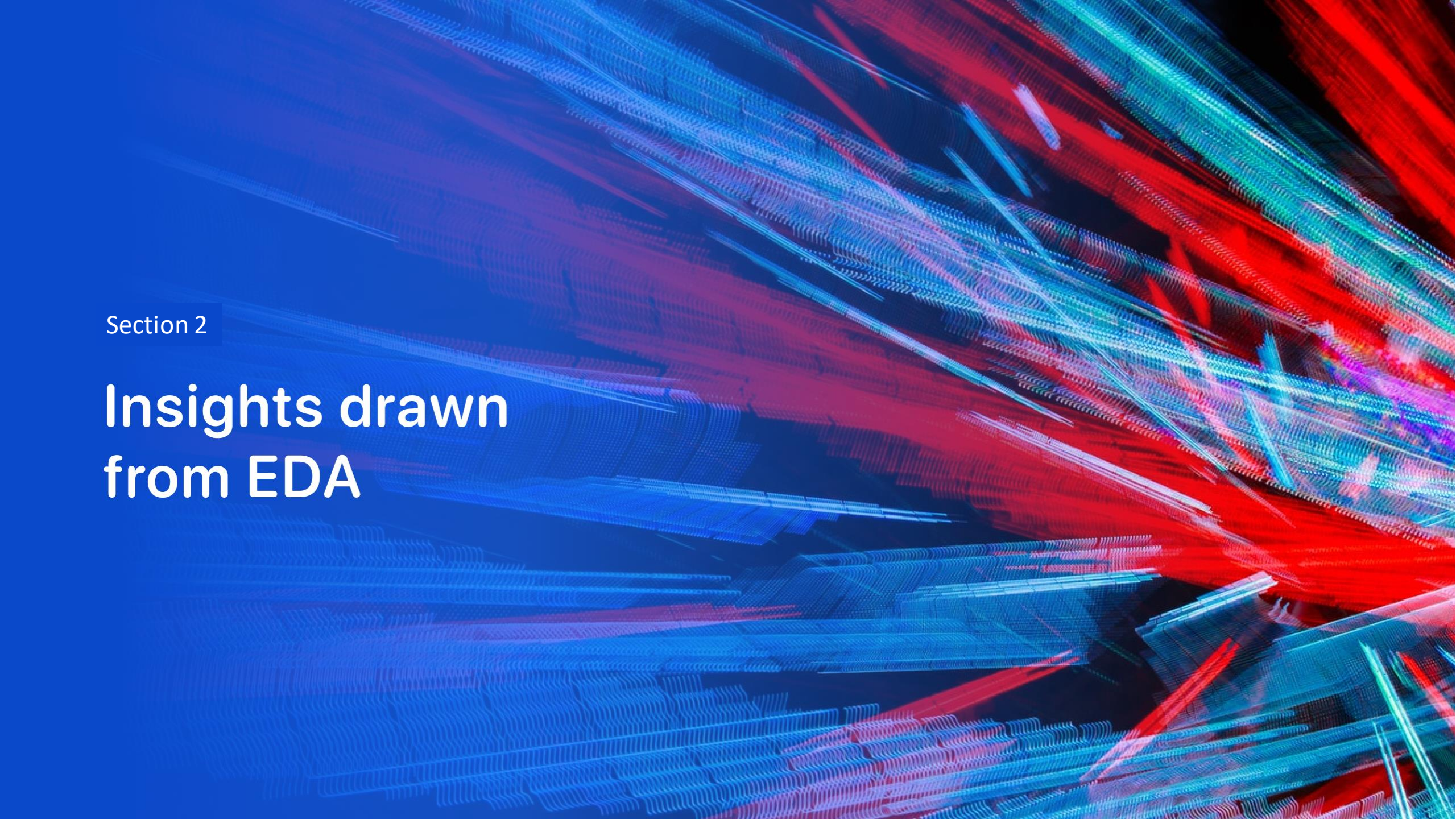
- The pandas dataframe of the SpaceX data was scaled to standardize the dataset: *preprocessing.StandardScaler()*
- The dataset was split into a training and test set, using 80 % of the data for training: train_test_split().
- For all classification methods, a GridSearchCV object was used to determine the best parameters for the individual model.
- The data was analyzed using logistic regression, support vector machine, decision tree and K nearest neighbors.
- Each ML prediction was evaluated using the score function, as well as a confusion matrix.
- Github link: https://github.com/OganessonOg118/MLCapstoneProjectRepo/blob/master/Module%204%20-%20Complete%20the%20ML%20Prediction%20Lab.ipynb

Results

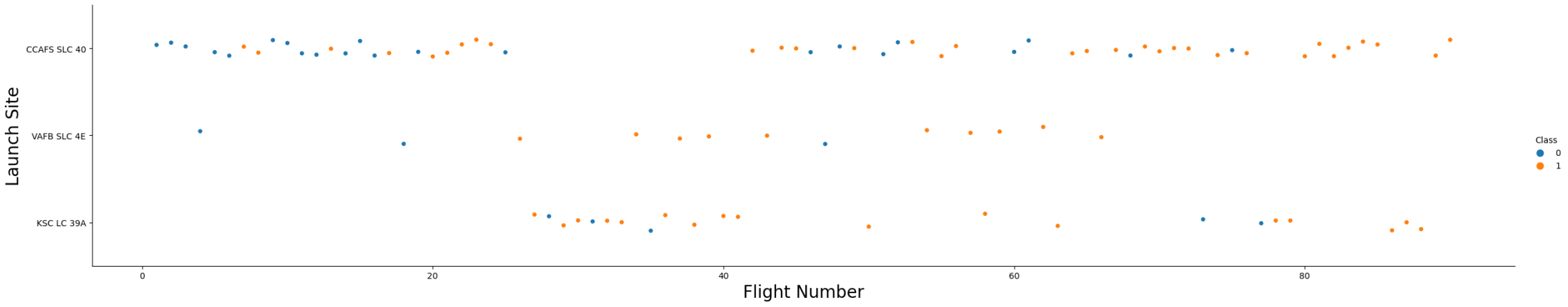
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



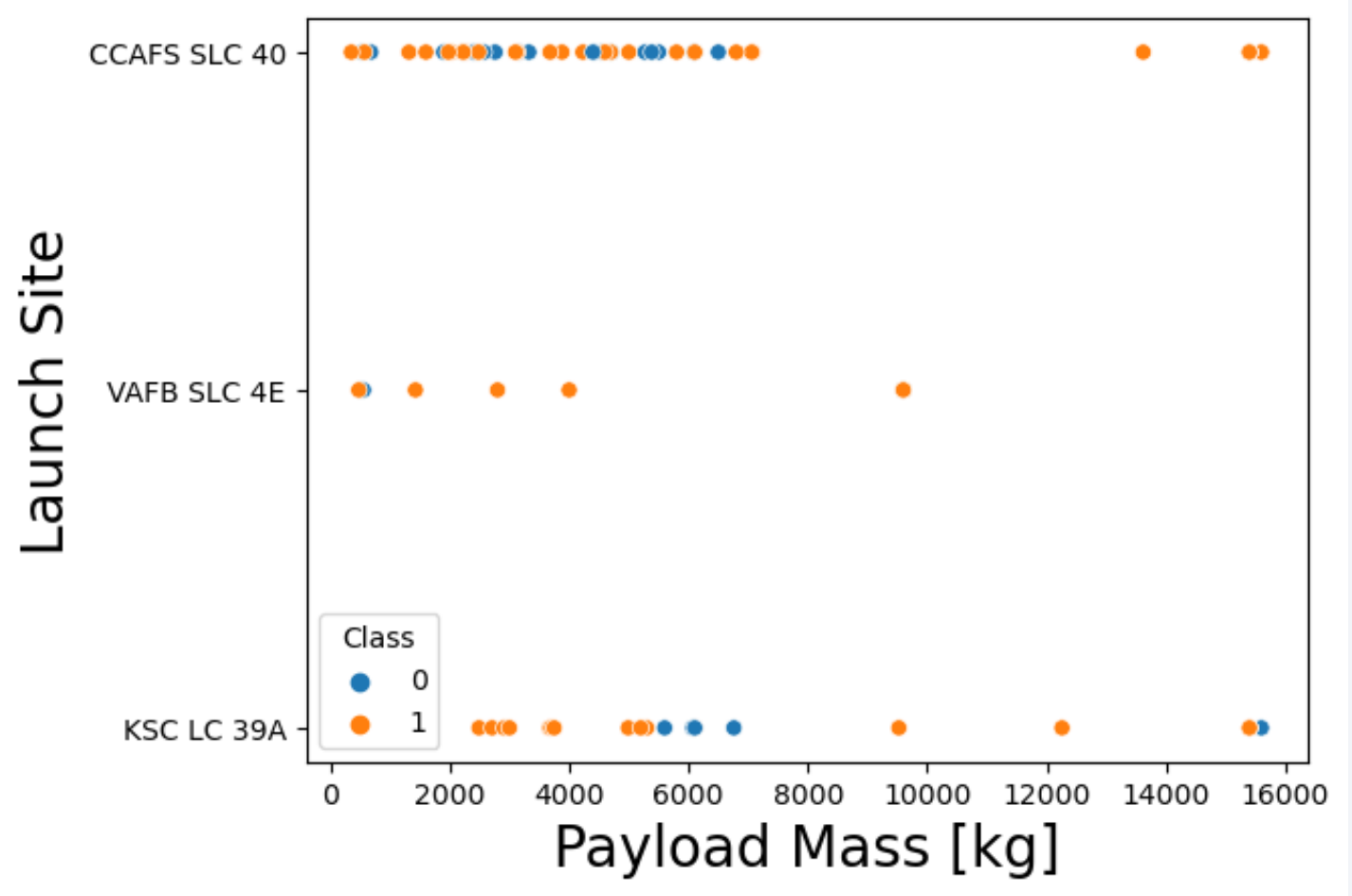
Flight Number vs. Launch Site



- For CCAFS SLC 40 the success rate seems to increase with the flight number.
- VAFB SLC 4E would need more data points to give a good estimate about a correlation.
- KSC LC 39A seemingly is less influenced by the flight number.

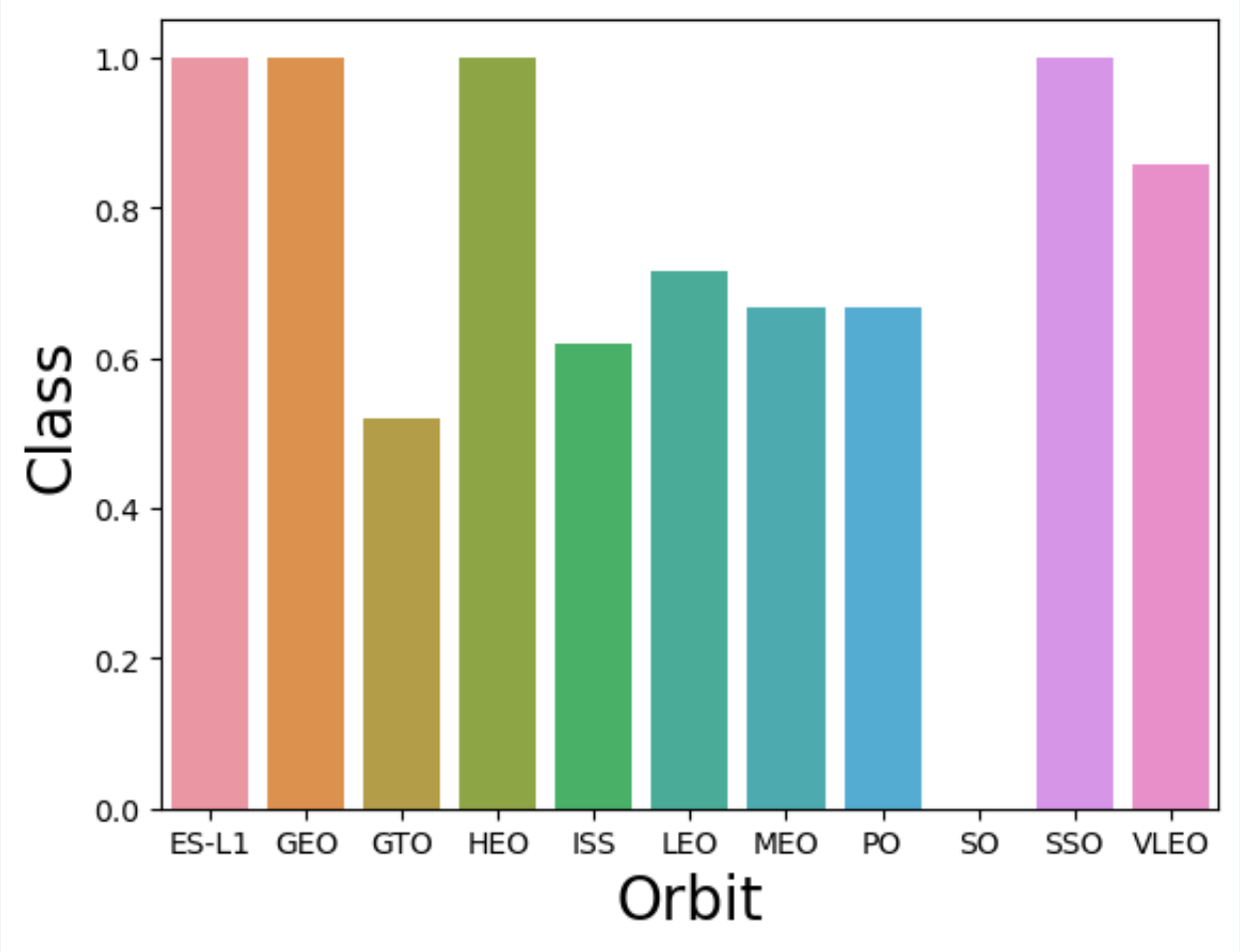
Payload vs. Launch Site

- CCAFS SLC 40 has a high number of tests and is seemingly successful in the landing of high payloads.
- VAFB SLC 4E on the other hand has only few data points and none above a payload of 10000 kg.
- There is an indication that KSC LC 39A is inferior in landing high payloads in comparison to CCAFS.



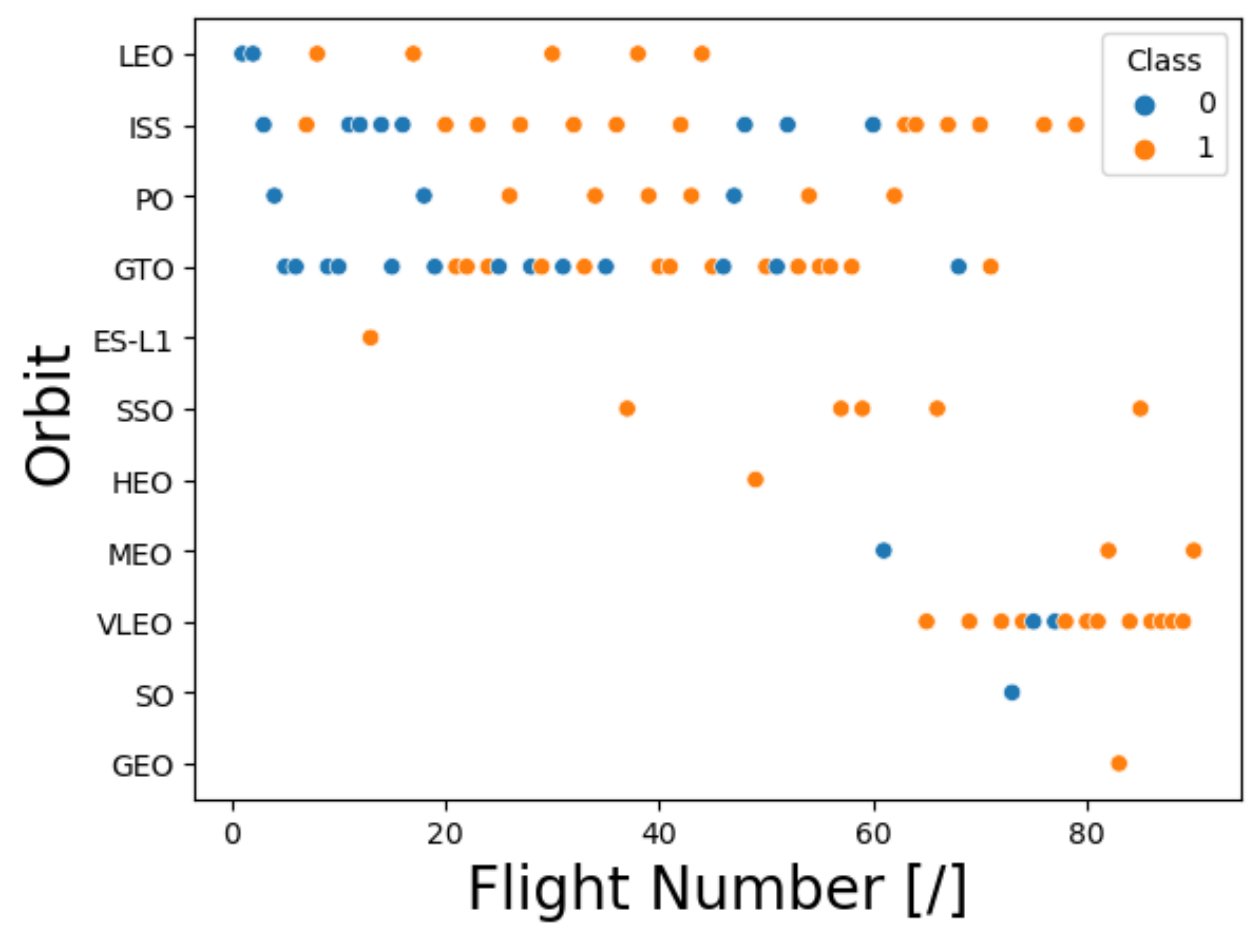
Success Rate vs. Orbit Type

- The highest success rate was reported for Orbits ES-L1, GEO, HEO, and SSO.
- There are no successful landings with Orbit SO.



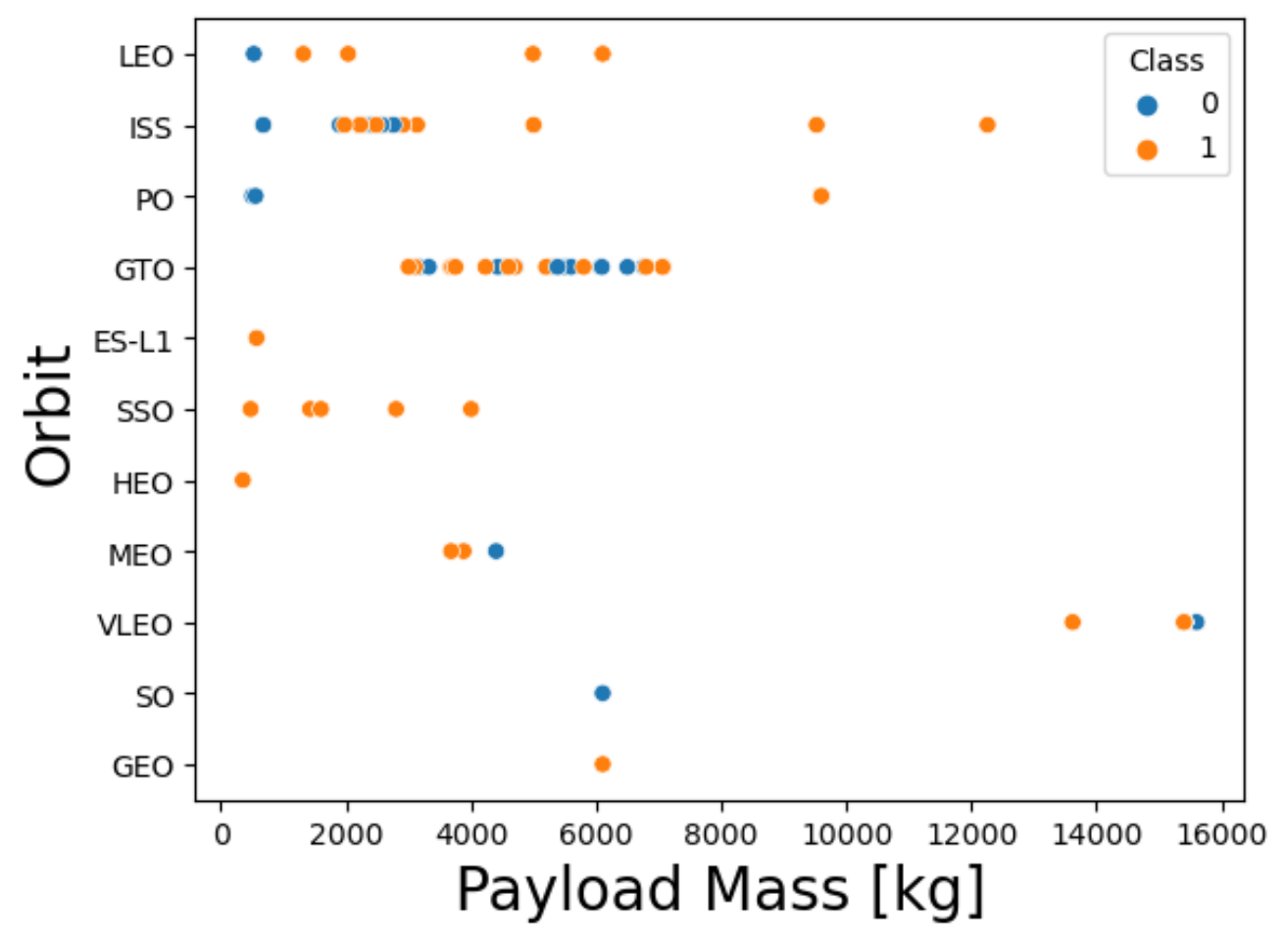
Flight Number vs. Orbit Type

- While LEO seemingly is influenced by the flight number.
- However, this orbit demonstrates a severely lower number of tests, and does not provide good statistics.
- Despite that, the success rate with the flight number is not related to the orbit.



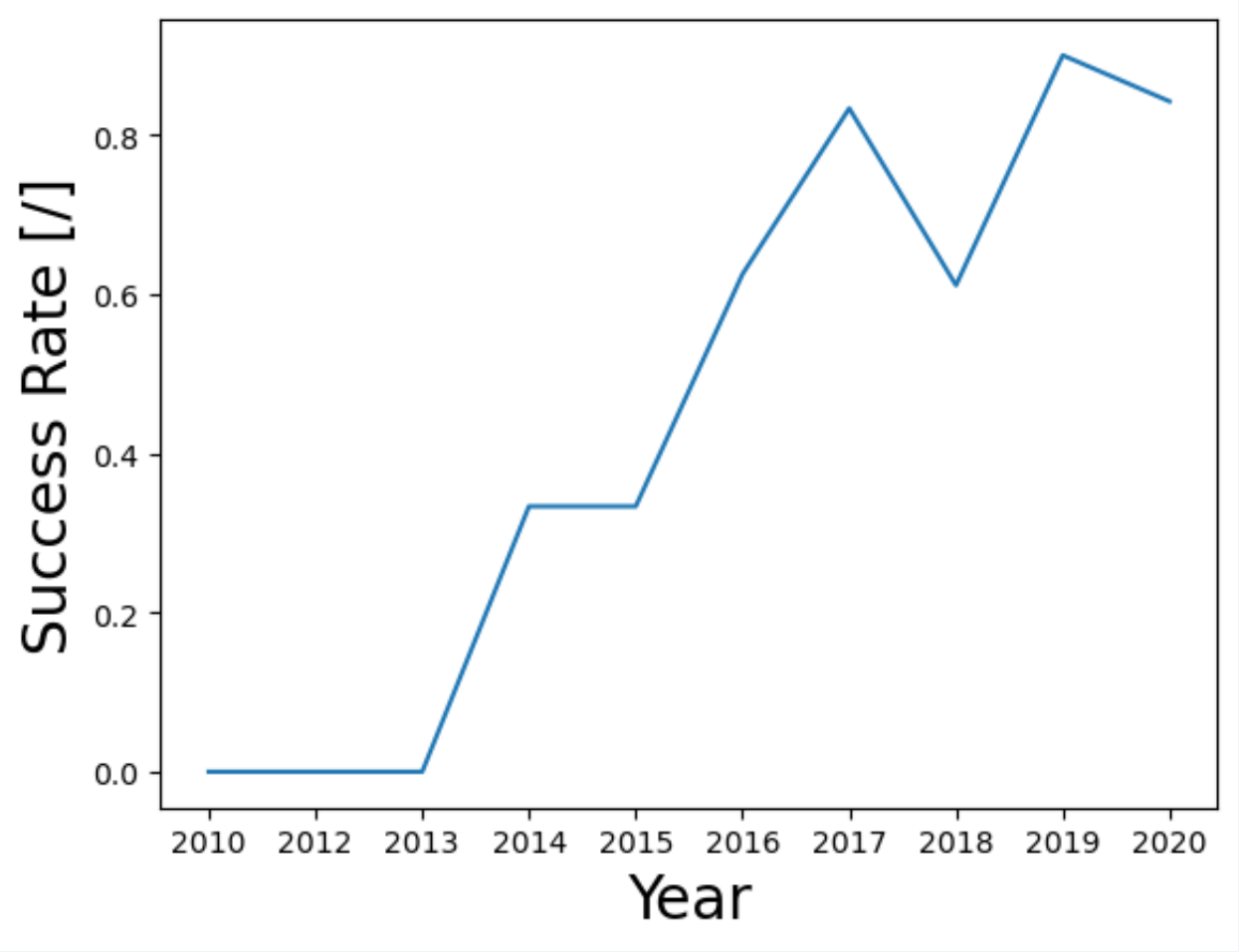
Payload vs. Orbit Type

- While it could be argued that ISS and PO are especially suited for higher payloads, they were not tested for masses as high as for VLEO.
- VLEO transported the highest payloads.
- There is no correlation observed for GTO and its success rate with the payload mass.



Launch Success Yearly Trend

- The yearly success rate increased over the years.
- Before 2013 there were no successful tests.



All Launch Site Names

```
In [8]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

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

- The launch sites are denoted as: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB Sl C-4E.

Launch Site Names Begin with 'KSC'

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5;
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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	6:00:00	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

- There are 5 entries where the Launch Site starts with "KSC" from the year 2017.
- Three of them were successful, whereas two were not attempted to land.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS_KG_) AS SUM_PAYLOADMASS FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

sum_payloadmass
45596

- The total amount of payload which was loaded in at NASA is 45596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS_KG_) AS AVG_PAYLOADMASS FROM SPACEX WHERE BOOSTER_VERSION= 'F9 v1.1';
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

avg_payloadmass
2928

- The average payload mass, where Booster version F9 v1.1 was used, is 2928 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS FROM SPACEX WHERE LANDING_OUTCOME LIKE 'Success%';
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_success
2015-12-22

- The first successful landing was reported on the 22.12.2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING_OUTCOME='Success (ground pad)' AND PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- Booster versions F9 FT B1032.1, F9 B4 B1040.1, and F9 B4 B1043.1 lead to successful missions with a payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) AS COUNT_OUTCOME FROM SPACEX GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- Overall, the mission was almost always successful.
- Only one time each there was a failure in flight or the payload status unclear.

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS_KG_=(SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEX)
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- Above is a list of the booster versions which carried the maximum payload mass.

2015 Launch Records

```
%sql SELECT MONTHNAME(DATE) AS MONTH_NAME, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING_OUTCOME='Success (ground pad)' AND YEAR(DATE)=2017;
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- The successful landing outcomes at the ground pad in the year 2017 were mostly reported for the KSC LC-39A launch site.

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

```
%sql SELECT LANDING_OUTCOME FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

```
 * ibm_db_sa://rjr97973:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

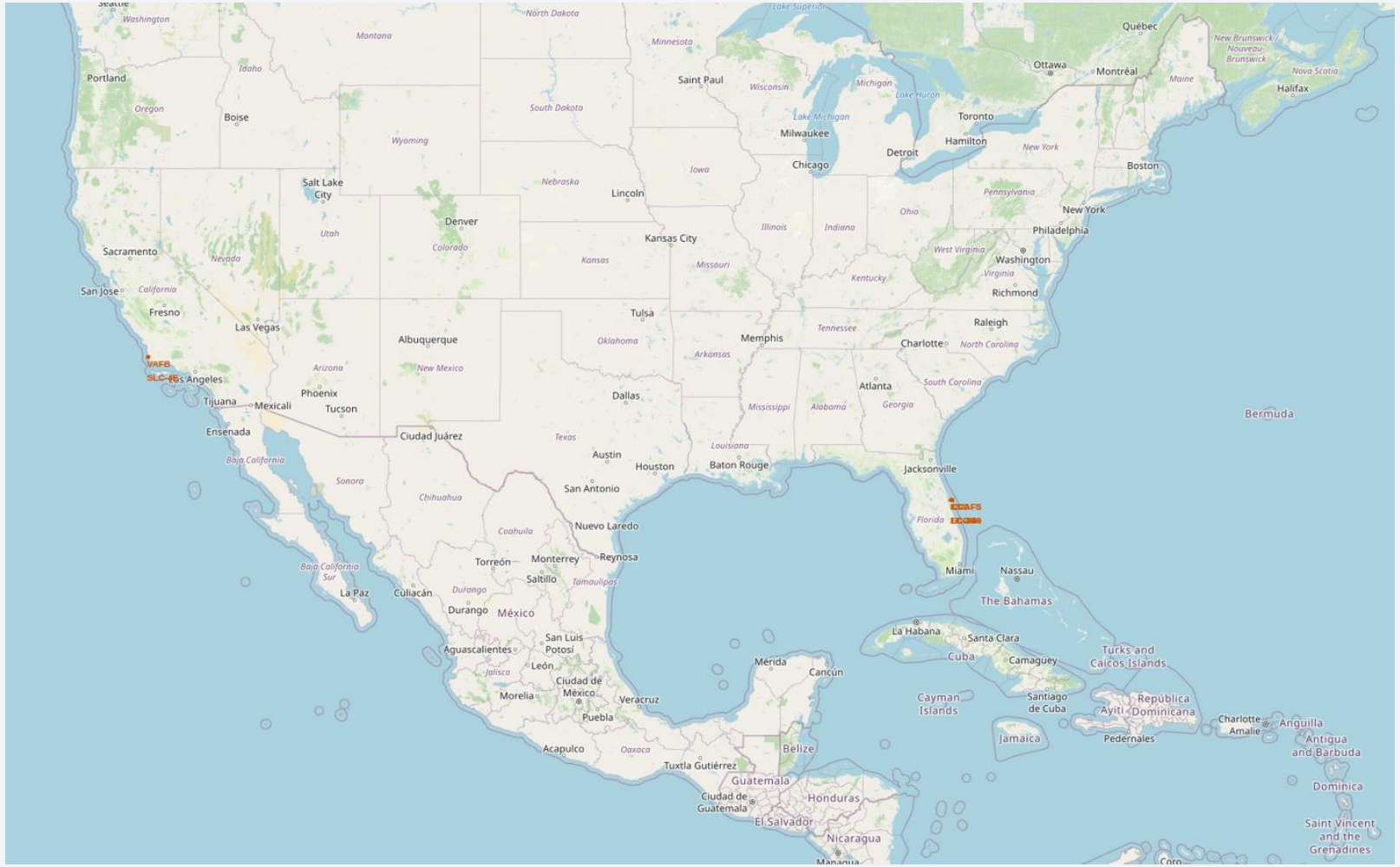
landing_outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

- The landing outcomes between 04.06.2010 and 20.03.2017 are listed in descinding order.

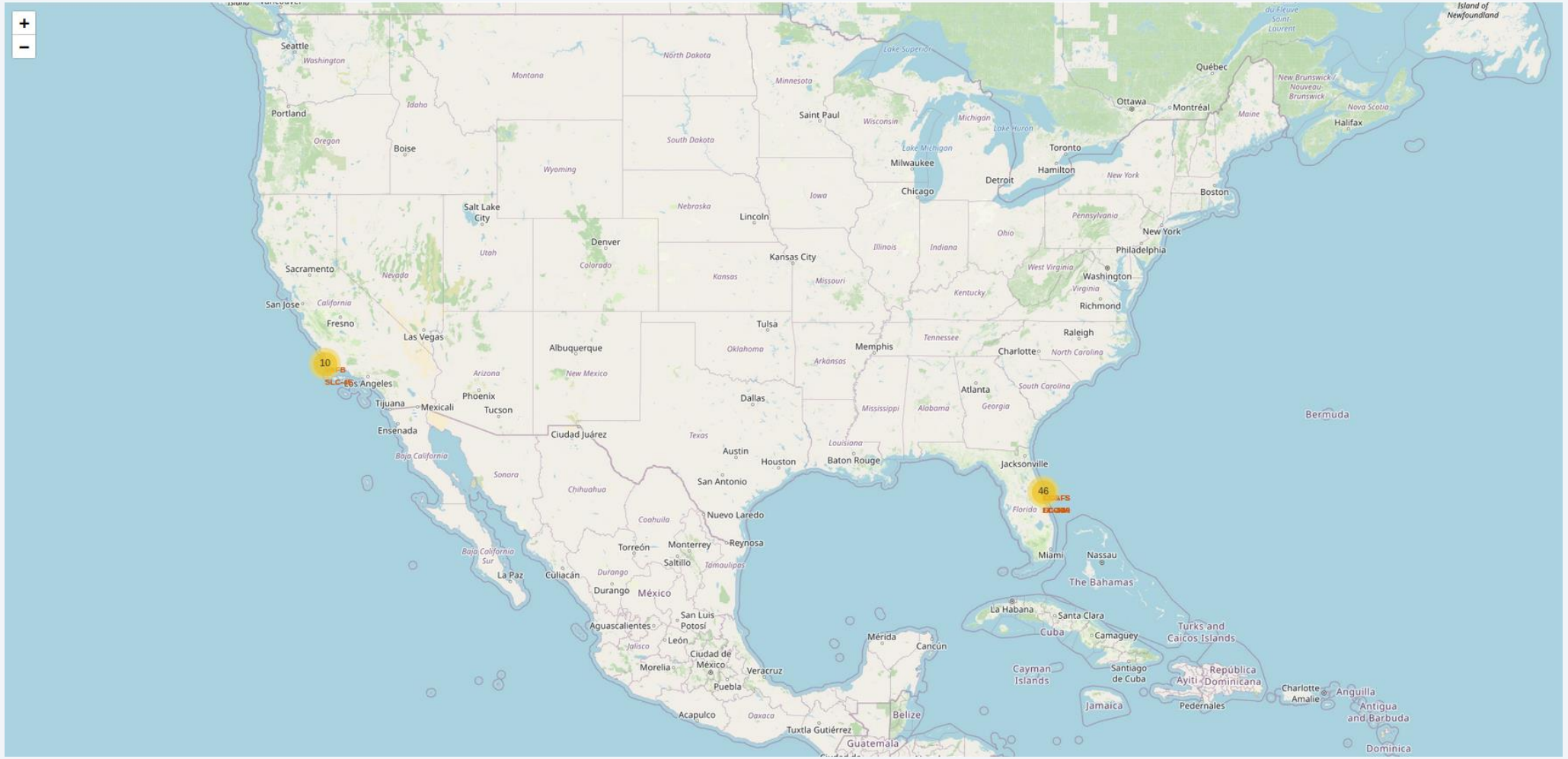

Section 3
Launch Sites
Proximities Analysis

Launch site locations

- There are a few launch sites clustered at the east and west coast of the USA.

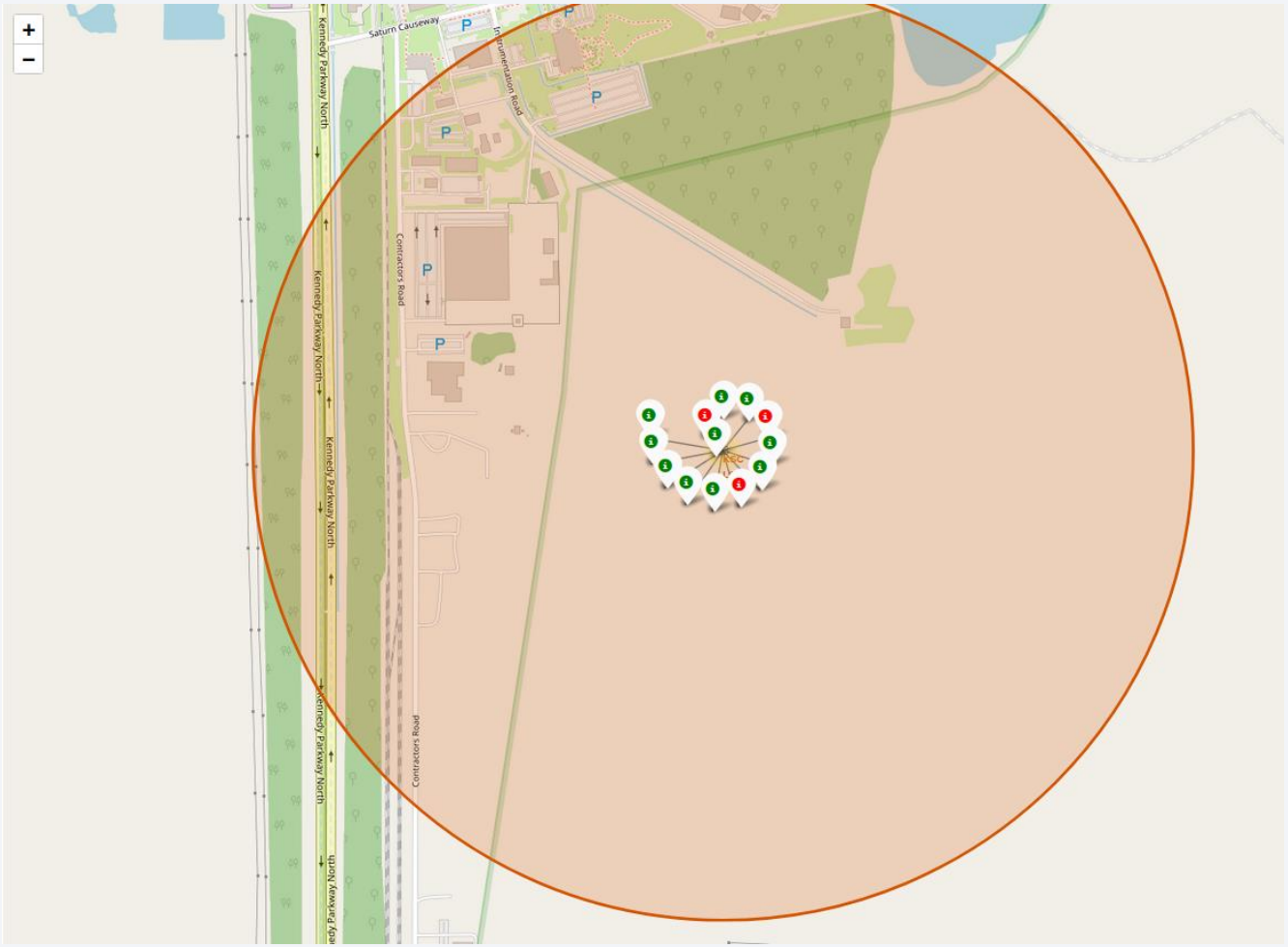


Marker cluster of launch outcomes



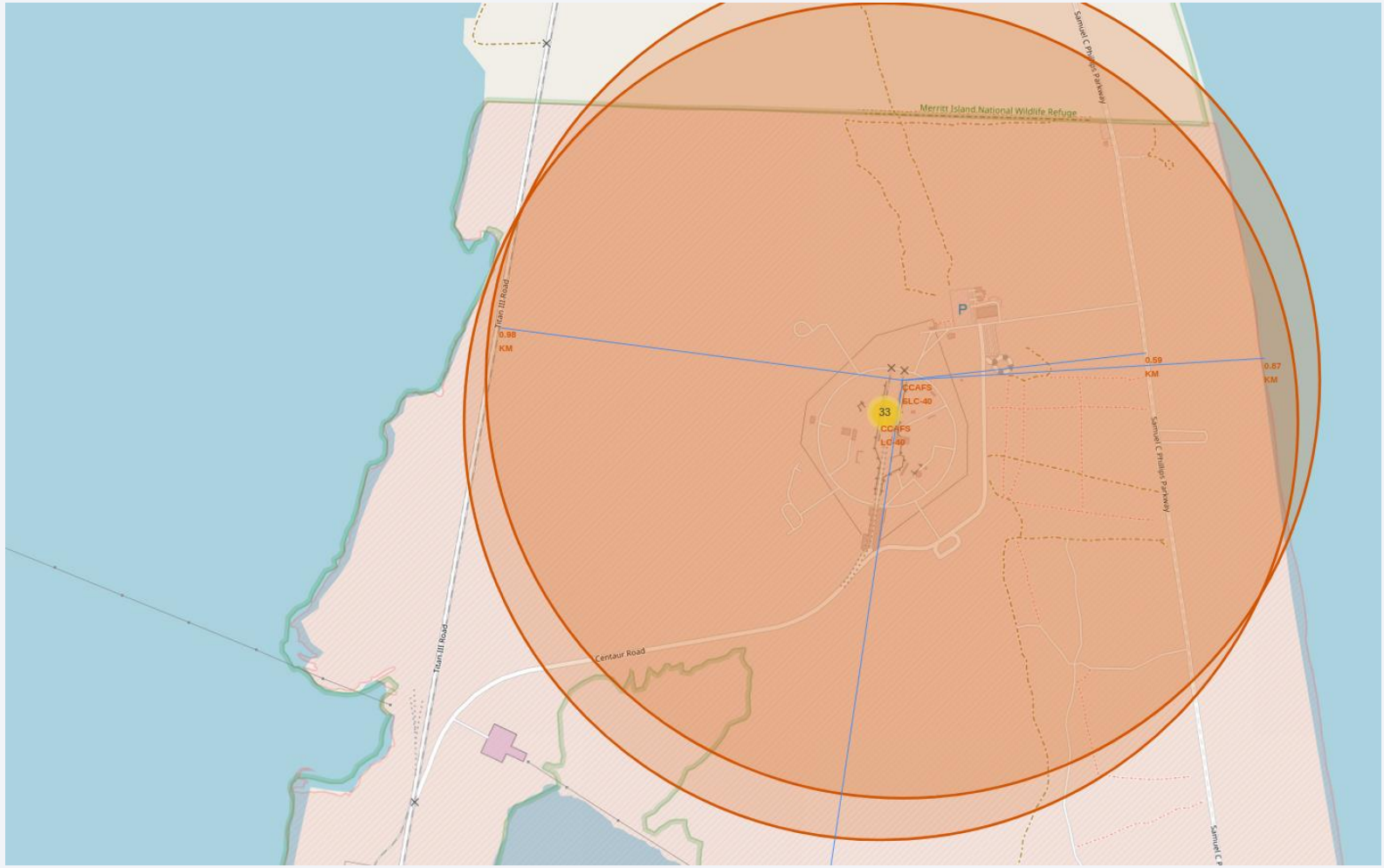
Marker cluster of launch outcomes

- For better visualization the launch outcomes are clustered around one location.
- The launch outcome is divided into failure (red) and success (green).



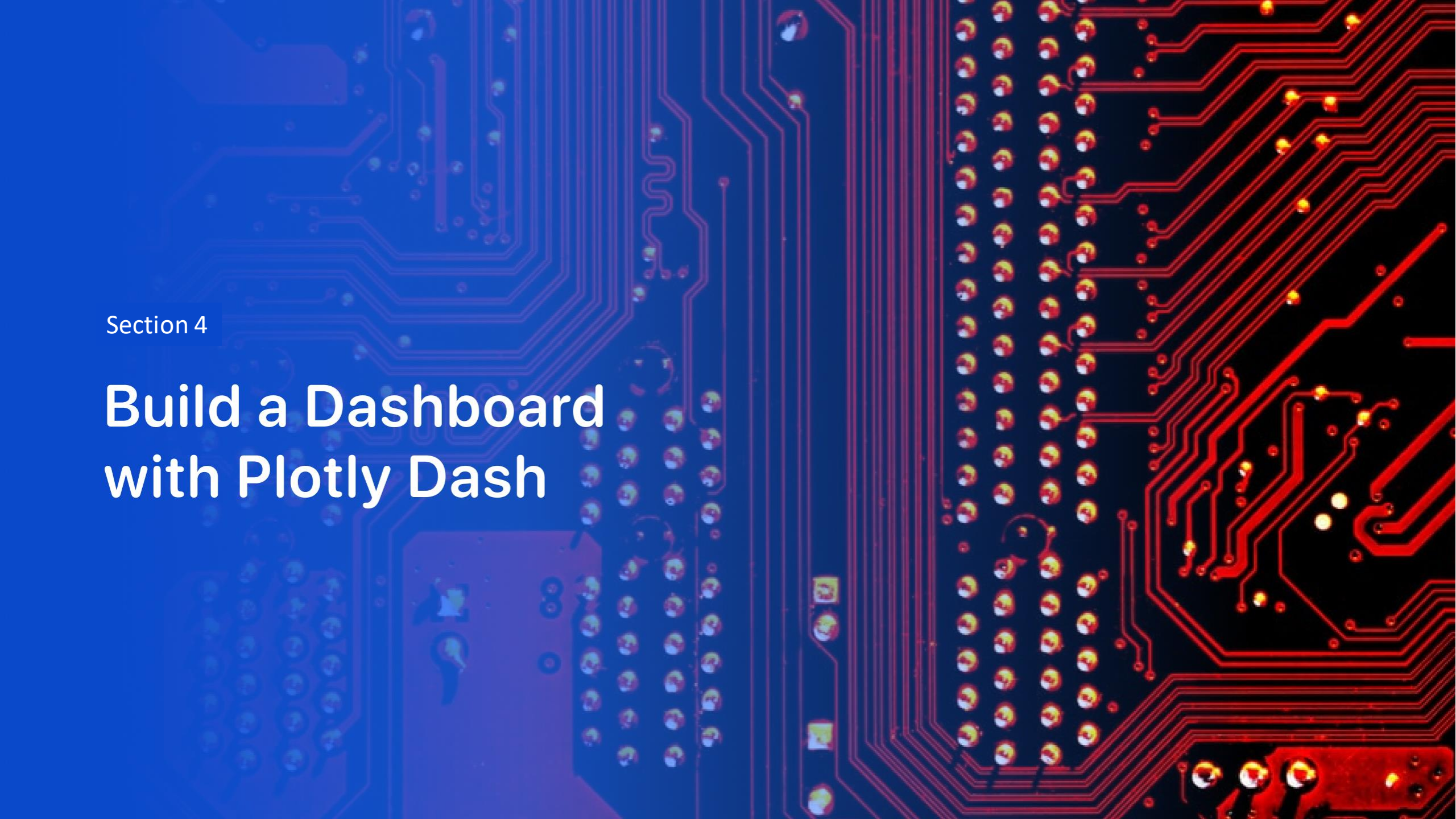
Launch site distance to coast line, highway, railway.

- The distance to the coast is preferably low to send the rocket into uninhabited areas in case of failure.
- Railways should be kept close to transport heavy equipment.
- The distance to highways would have to be evaluated in comparison to the overall density of highways in the USA.

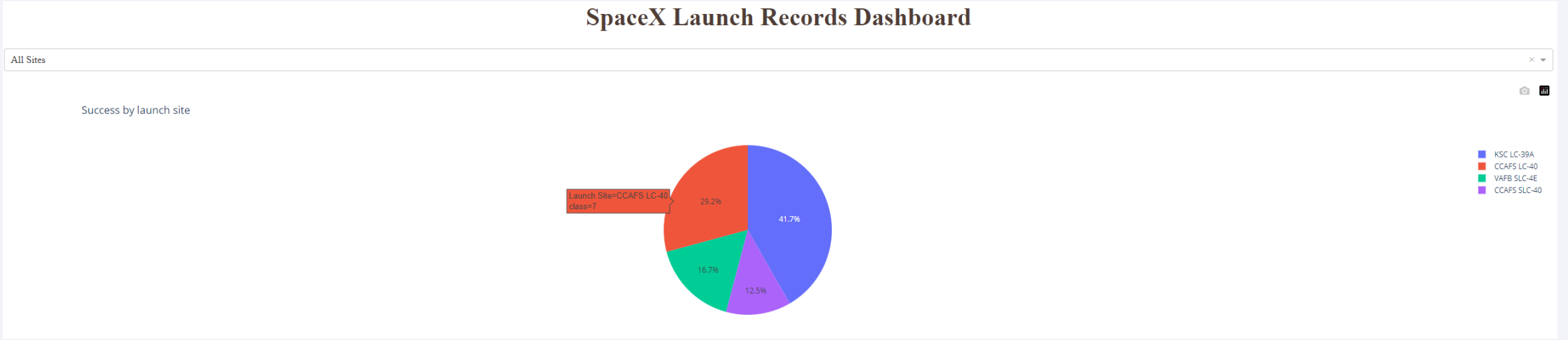


Launch site distance to city

- The distance between a launch site and a city should be kept relatively far.
- In case of failures or unforeseen circumstances casualties must be avoided at all costs.

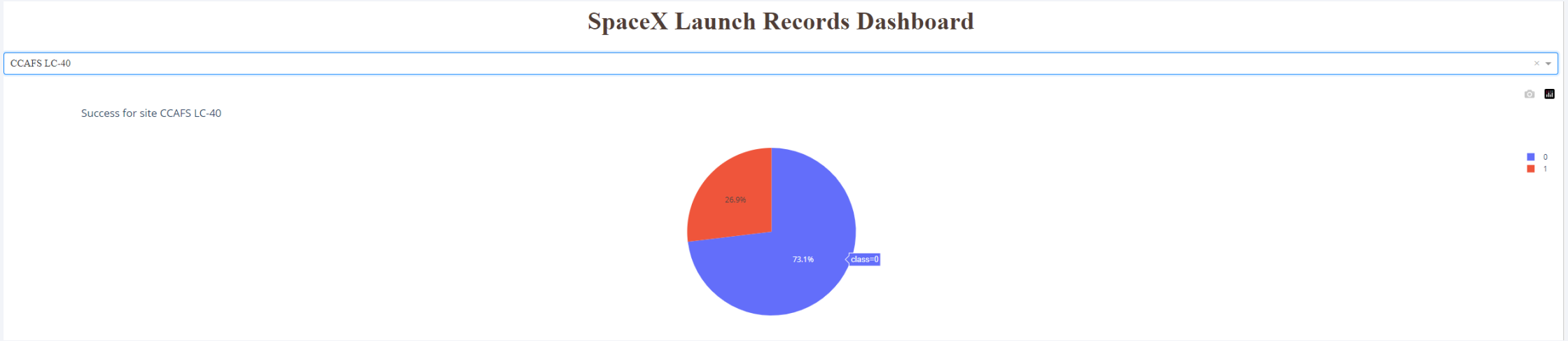

Section 4
Build a Dashboard with Plotly Dash

SpaceX Successful landing by launch site



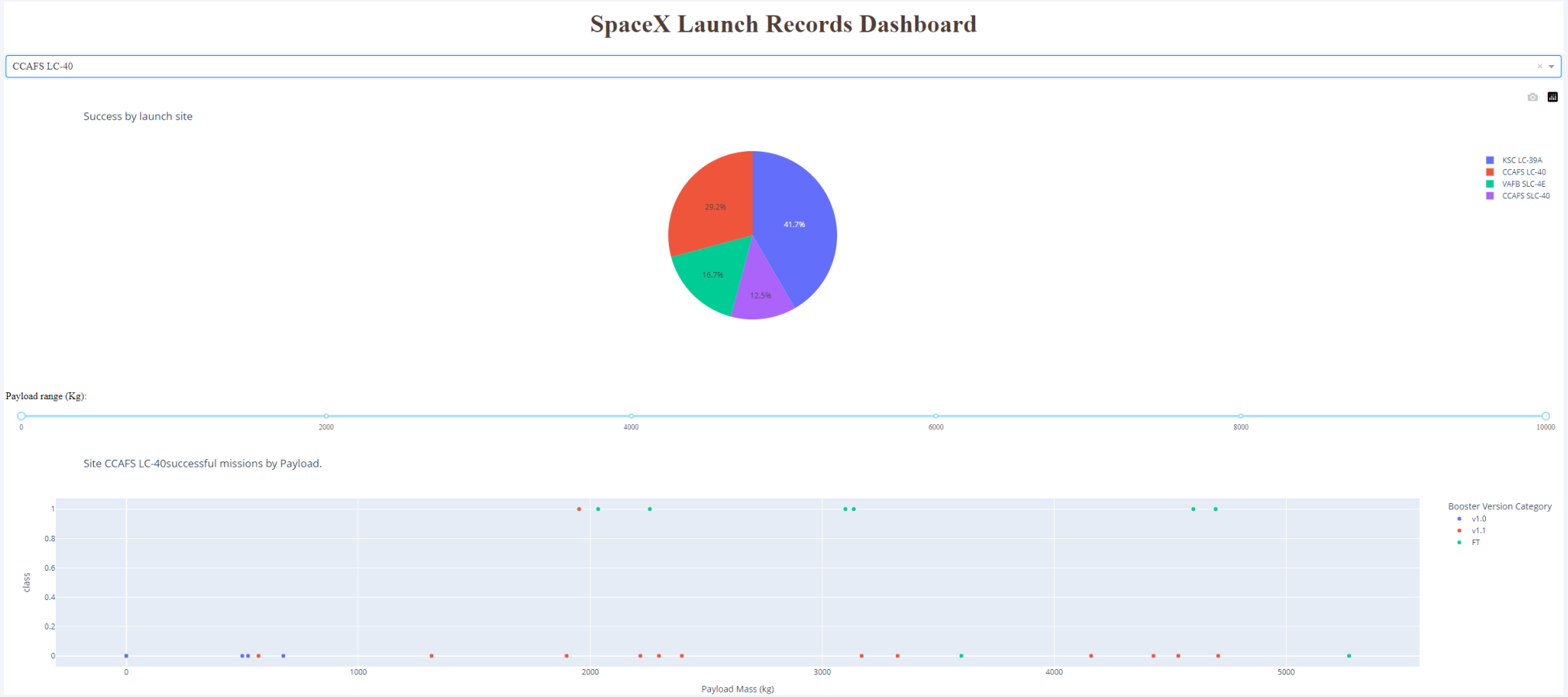
- The launch site KSC LC-39A showed the highest success rate with 41.7 % of all successful launches coming from this site.

Success ratio for CCAFS LC-40



- The ratio of successful landings for CCAFS LC-40 was 73.1 % of failures to 26.9 % of success.

SpaceX Successful missions by Payload for CCAFS



- CCAFS LC-40 showed a high amount of failures with succesful landings mainly reported for booster category FT.

Section 5

Predictive Analysis (Classification)

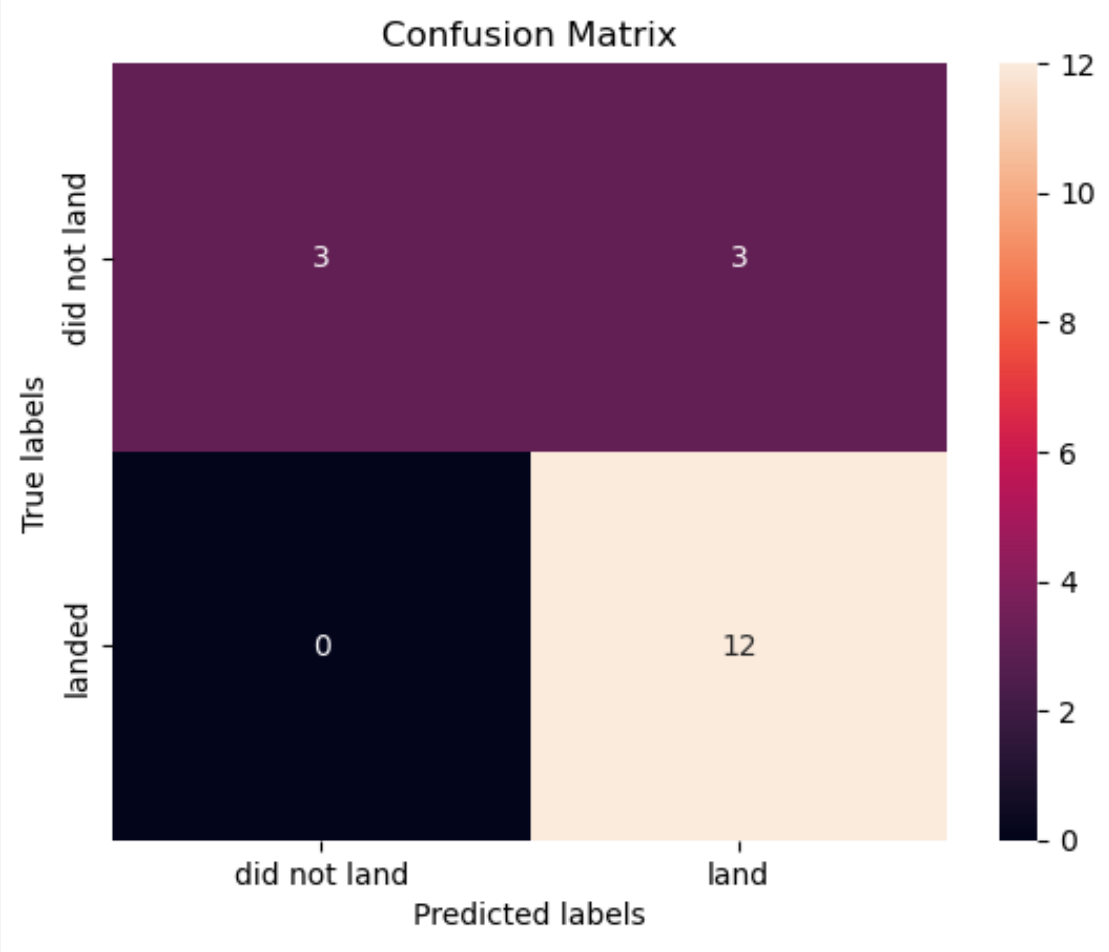
Classification Accuracy

```
The score of logreg was: 0.8333333333333334
The score of svm was: 0.8333333333333334
The score of decision tree was: 0.8333333333333334
The score of knn was: 0.8333333333333334
```

- A classification accuracy of 83.33 % was determined for the logistic regression, support vector machine, decision tree and K nearest neighbors.
- This identical result made a visualization in form of a bar graph irrelevant.
- This outcome is likely due to a comparably small dataset, which would have to be extended in the future.

Confusion Matrix

- The confusion matrix for the different classification models was identical.
- There was no false negative of a prediction of "did not land". This means that with high certainty it can be estimated if a rocket would not land.
- On the other hand, in a few cases a successful landing was predicted, which was not the case.
- Overall, the predictions were very reliable.



Conclusions

- Within this project data of SpaceX launches were collected.
- The data was processed and divided into successful and failed launches.
- Visualization of the data unraveled dependencies of the success with e.g. Payload masses, flight numbers.
- ML algorithms were trained to predict the landing outcomes.
- Within the used classification methods (logistic regression, support vector machine, decision tree and K nearest neighbors) there was no difference in their ability for prediction.
- Further datasets are needed to improve the models.

Appendix/References

- The information provided in this presentation was mainly taken from:
  - IBM DS0720EN Data Science and Machine Learning Capstone Project, edx.org, 2023.
  - https://api.spacexdata.com/v4/
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!