IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ruslan Almetov
March 23, 2023

Outline

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

Executive Summary

In this project, we aim to predict the successful landing of the Falcon 9 first stage using the methodologies learned in the course, including:

- Data collection via API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis
- Data visualization using Matplotlib, Seaborn and Dash
- Machine learning methods

As a result, we trained 4 machine learning models to predict the success of the landing, the accuracy in all cases was 83.3%

Introduction

- 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.
- So, 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 Project, we trained machine learning models for such predictions.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Request to the SpaceX API
  - Web scraping from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
- Perform data wrangling using Pandas
  - Calculated: the number of launches on each site, the number and occurrence of each orbit, mission outcome per orbit type;
  - Created a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

- Perform predictive analysis using classification models using sk-learn
  - Data was standardized (by StandardScaler), splited into training and testing (by train_test_split function);
  - After that, we found best Hyperparameters (GridSearchCV) and fited objects of SVM, Classification Trees, Logistic Regression, K-neighbors Models (by fit method);
  - We calculated the accuracy on the test data using the method score, and ploted the confusion matrix.

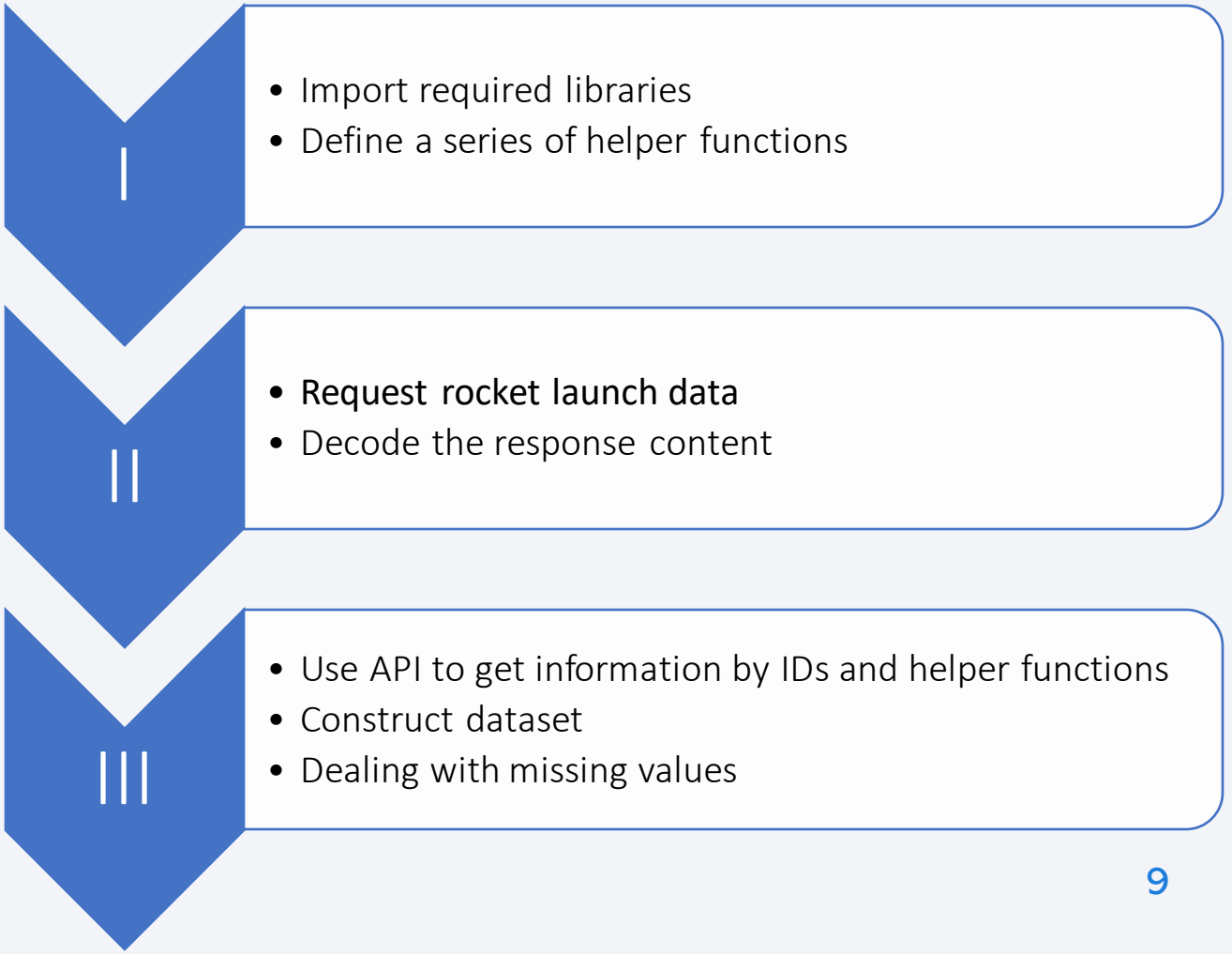
Data Collection

- Data sets were collected:
  - via requests to the SpaceX API
  - using Web scraping from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches'

Data Collection – SpaceX API

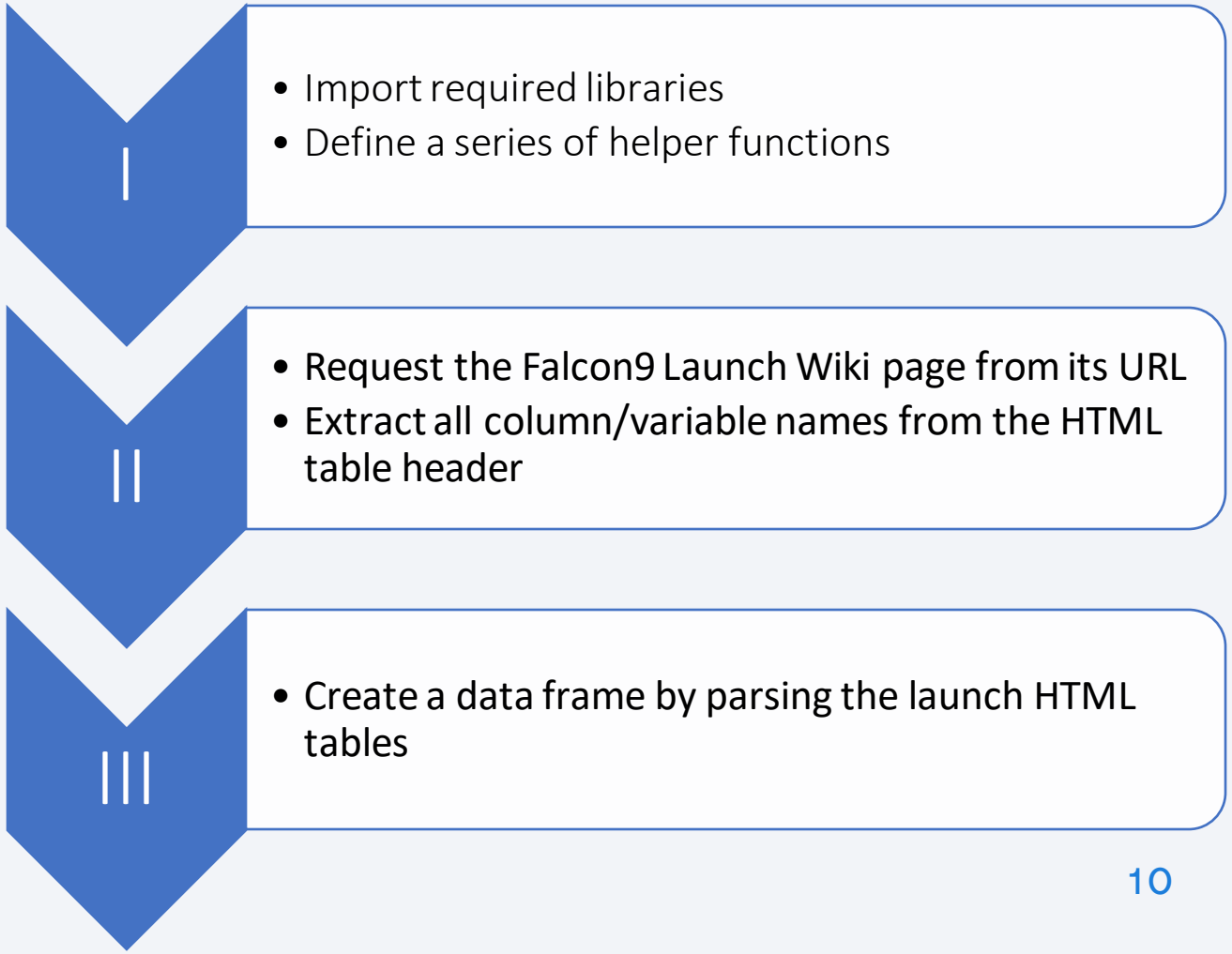
- SpaceX URL:
- https://api.spacexdata.com/v4

- GitHub URL of the completed SpaceX API calls notebook:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/7d715ed0ab6f938f636f879462a476ae674fb4ec/01_Data%20Collection%20API.ipynb



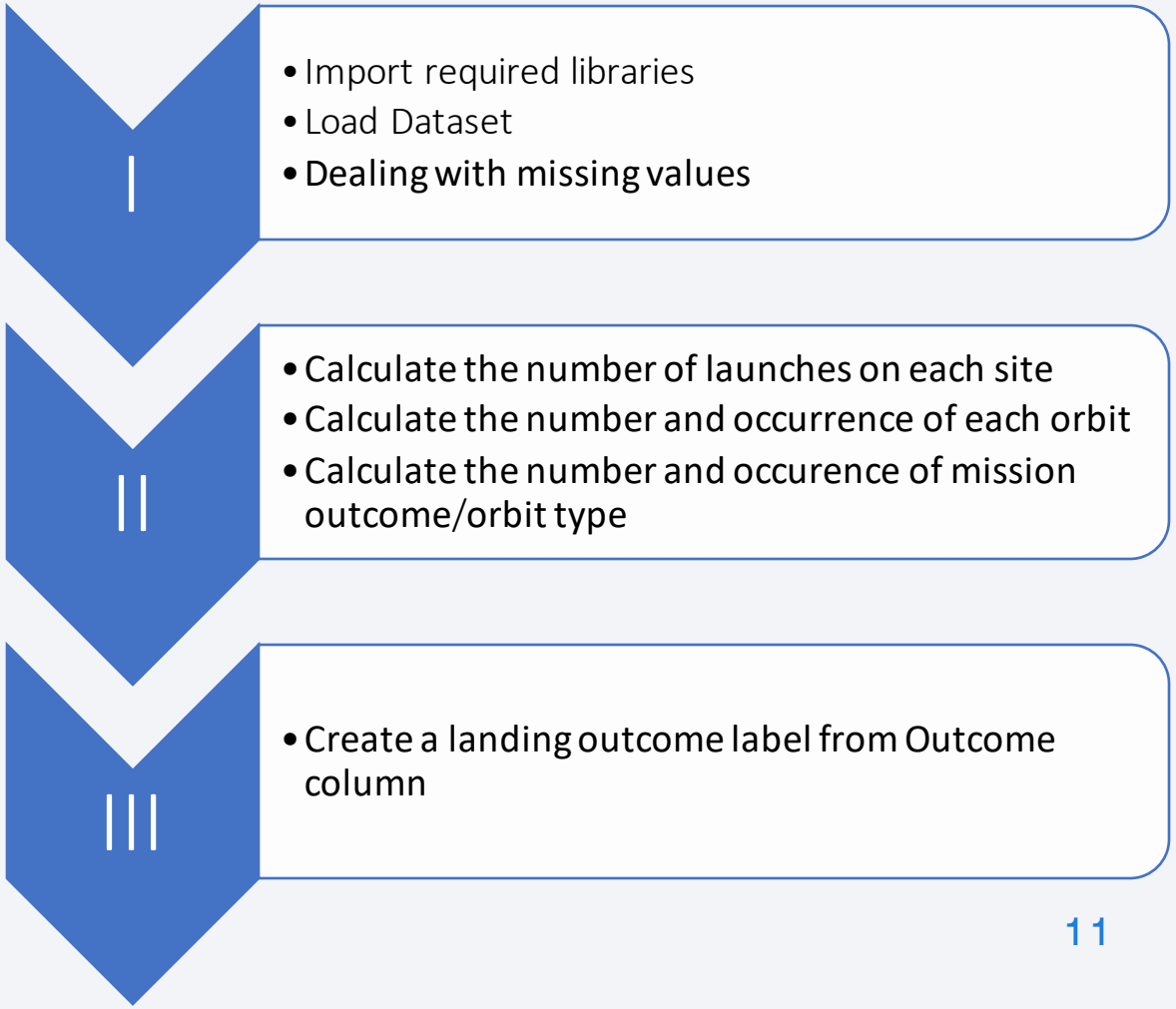
Data Collection - Scraping

- Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches` URL:
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub URL of the completed web scraping notebook:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/7d715ed0ab6f938f636f879462a476ae674fb4ec/02_Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine the label for training models, including:
  - Calculated: the number of launches on each site, the number and occurrence of each orbit, mission outcome per orbit type;
  - Created a landing outcome label from Outcome column
- GitHub URL of the completed data wrangling related notebooks:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/7d715ed0ab6f938f636f879462a476ae674fb4ec/03_Data%20wrangling.ipynb



EDA with Data Visualization

- As part of the implementation of the EDA, we have built the following charts:
- **Scatter point charts** showing relationships between pairs of the following features in various combinations: Flight Number, Payload, Launch Site, Orbit types;
- **Bar chart** showing the relationship between success rate of each orbit type;
- **Line chart** showing the launch success yearly trend

GitHub URL of the completed EDA with data visualization notebook:

https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/7d715ed0ab6f938f636f879462a476ae674fb4ec/05_EDA%20Pandas%20Matplotlib.ipynb

EDA with SQL

- The SQL queries performed as part of the implementation of the EDA:
  - 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 succesful 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 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL of the completed EDA with SQL notebook:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/7752c2d31474efe5a2fd9573e9f1f42d8bf1aad6/04_EDA%20SQL_sqllite_corr.ipynb

Build an Interactive Map with Folium

- We created and added to a folium map the following objects:
- **Markers and highlighted circle area** on a specific coordinates(Launch Sites)
- **MarkerCluster including two-color markers** on the success/failed launches for each site
- **Lines with markers** showing calculated distance to a closest railway, highway, coastline, port.

- GitHub URL of the completed interactive map with Folium map:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/45097040f6df8fdd4353337af89fcb261fb54c97/06_Data%20Visualization%20with%20Folium.ipynb

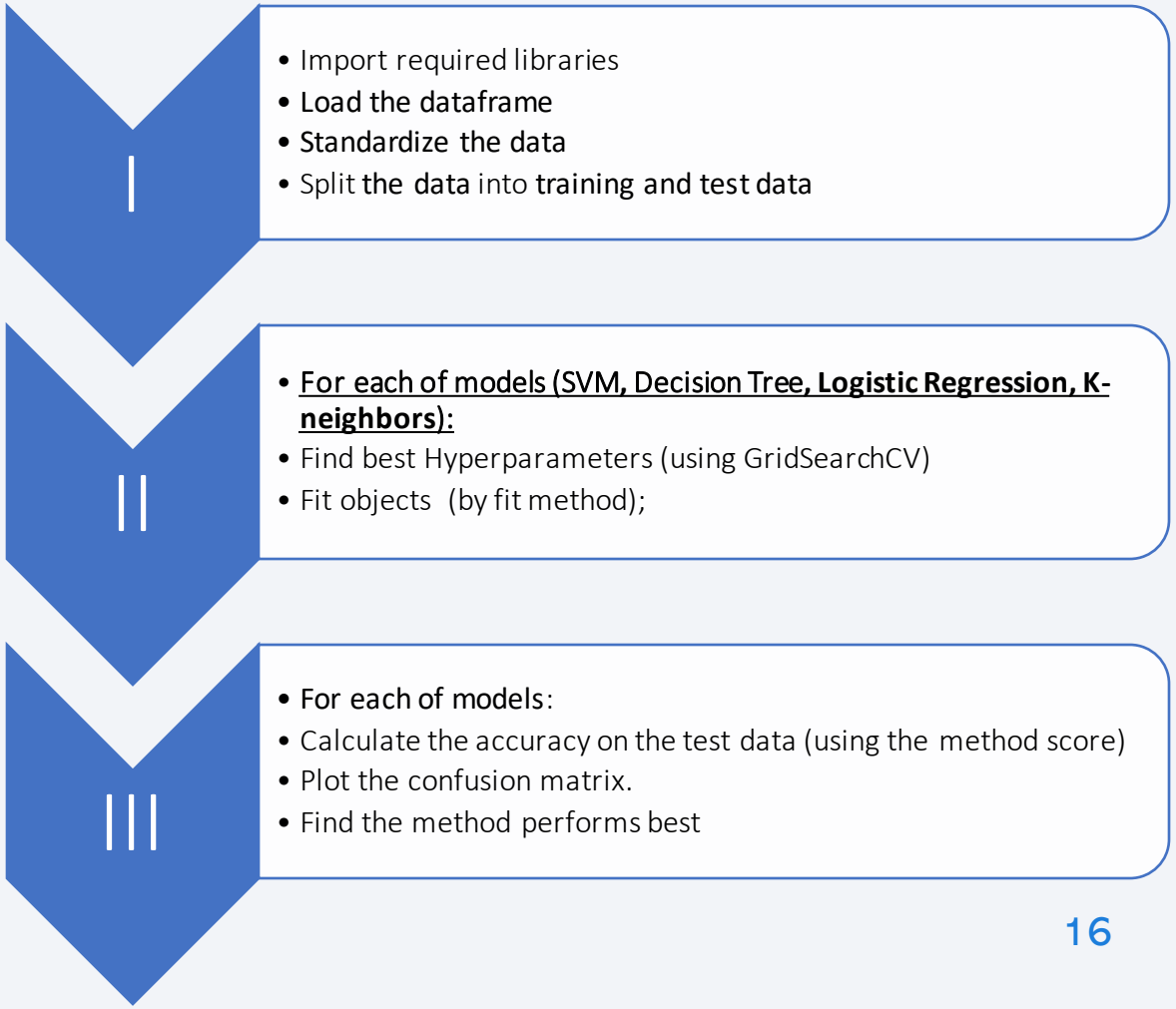
Build a Dashboard with Plotly Dash

- We created and added to a dashboard the following objects:
- A Launch Site Drop-down Input Component
- A Pie chart showing Total Success for selected by drop-down input site.
- A Range Slider to Select Payload
- A Scatter chart showing how payload (selected by range slider) correlate with mission outcomes for selected by drop-down input site.

- Add the GitHub URL of your completed Plotly Dash lab:
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/45097040f6df8fdd4353337af89fcb261fb54c97/07_Data%20Visualization%20Dash%20App.py

Predictive Analysis (Classification)

- We performed predictive analysis using following classification models in order to compare the results and choose the best model:
- Logistic regression
- Support vector machine
- Decision Tree
- K-Neares Neighbors
- **GitHub URL of the completed predictive analysis lab:**
- https://github.com/ralmetov/Applied-Data-Science-Capstone/blob/45097040f6df8fdd4353337af89fcb261fb54c97/08_Machine%20Learning%20Prediction.ipynb



Results

- **Exploratory data analysis results**
- There are 4 unique launch sites in the SpaceX mission (CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A)
- The total payload mass carried by boosters launched by NASA (CRS) - is 45 596 KG
- The average payload mass carried by booster version F9 v1.1 - is 2928.4 KG
- The first succesful landing outcome in ground pad was achieved at 22-12-2015
- There are 4 boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 (F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2)
- There are 99 Successful mission outcomes, 1 - Successful (payload status unclear), 1 - Failure (in flight)
- There are 12 booster_versions which have carried the maximum payload mass
- There are 2 failure landing outcomes in drone ship in year 2015 (both from CCAFS LC-40): F9 v1.1 B1012 in January, F9 v1.1 B1015 in April

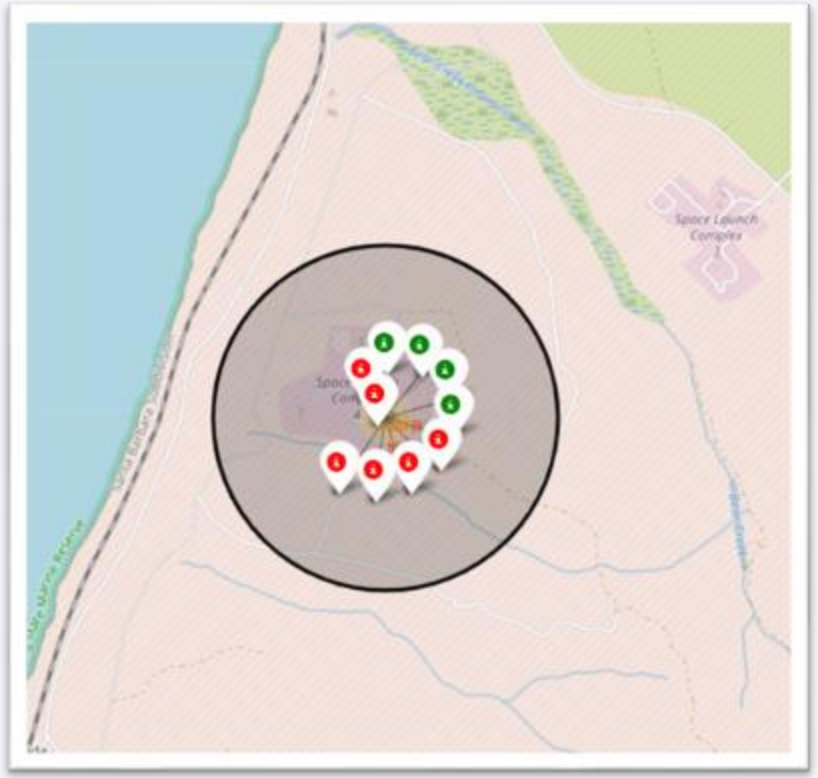
Results

- **Exploratory data analysis results**
- Launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Over time, the percentage of successful landings of each Launch Site improves
- The VAFB-SLC-4E: there are no rockets launched for heavy payload mass(greater than 10000)
- Orbits have different success rate. ES-L1, GEO, HEO and SSO has the best success rate of 100%, SO has the worst success rate of 0%
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- The success rate since 2013 kept increasing till 2020

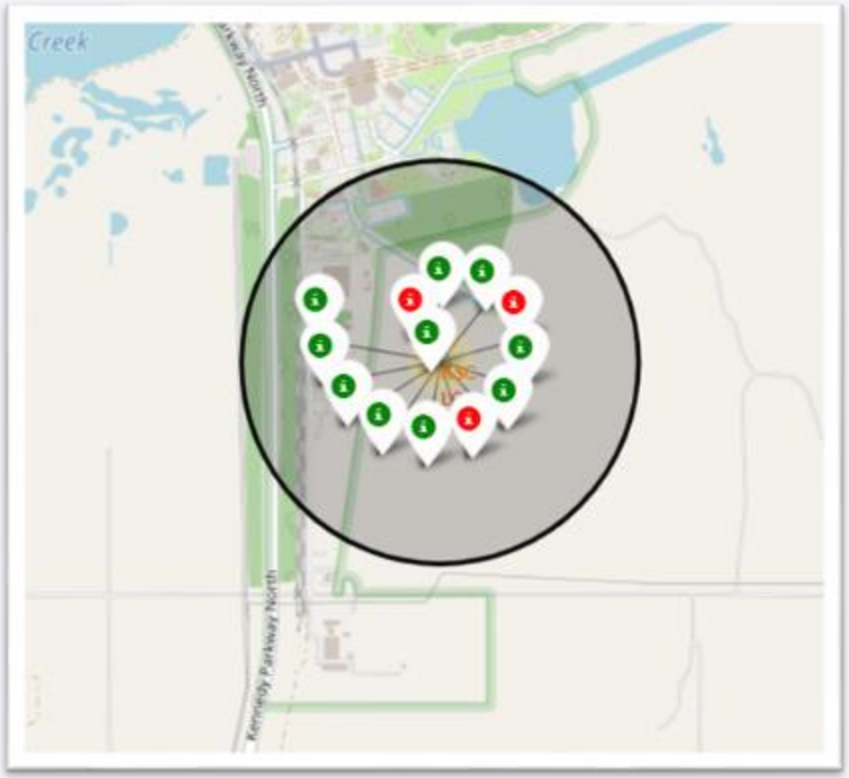
Results

- **Interactive analytics demo in screenshots**

Launch Site *VAFB-SLC-4E*



Launch Site *KSC LC-39A*



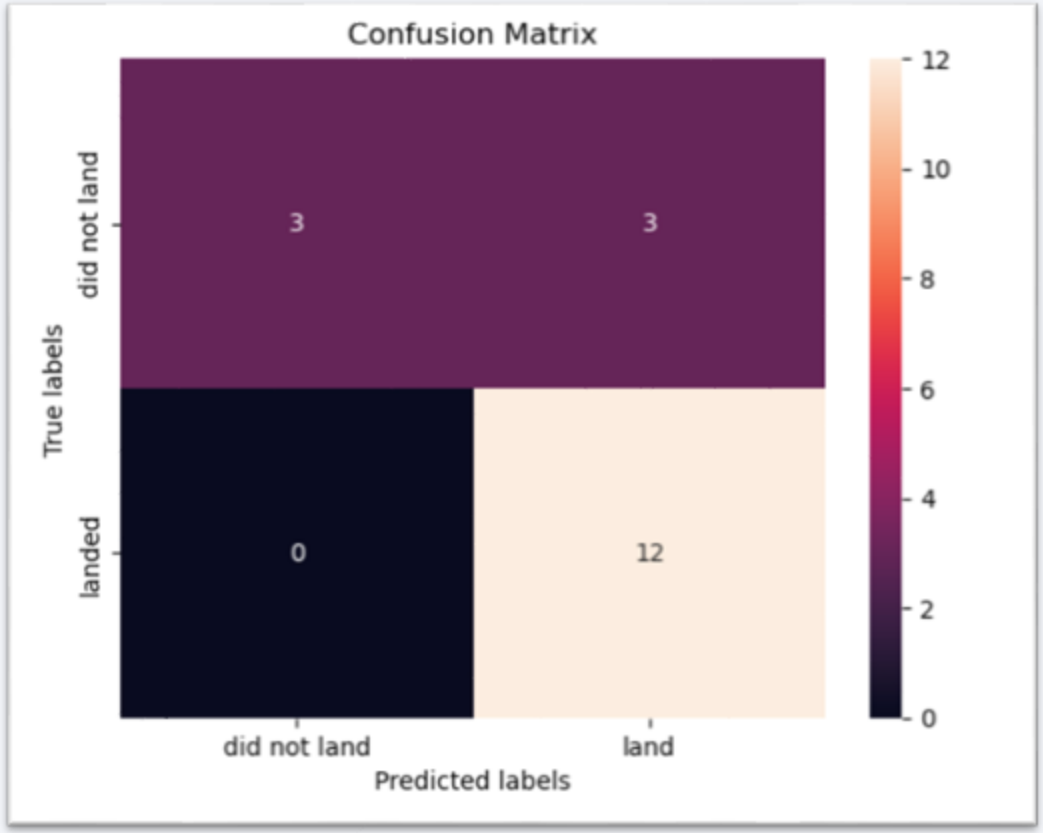
Results

- **Interactive analytics demo in screenshots**

Launch Sites: CCAFS LC-40 and CCAFS SLC-40

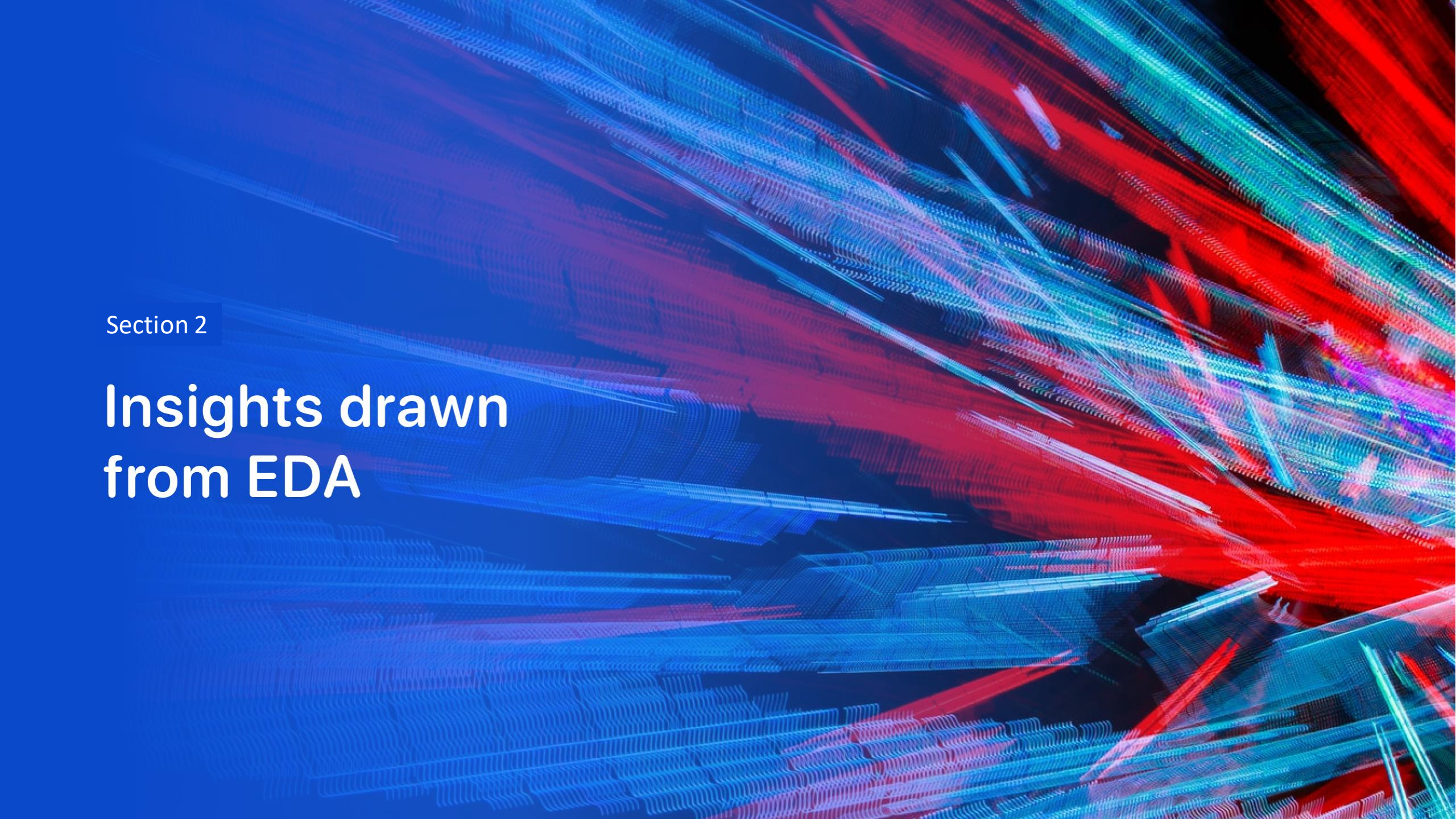
Results

- **Predictive analysis results**
  - Practically all trained models give the same result on the test data:
  - Accuracy = 0.833
  - Confusion Matrix:



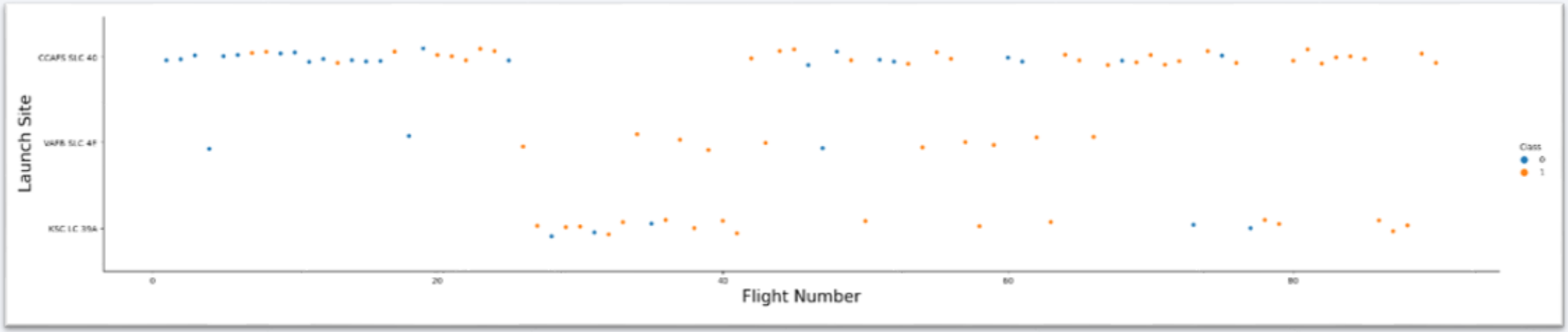
Section 2

Insights drawn from EDA



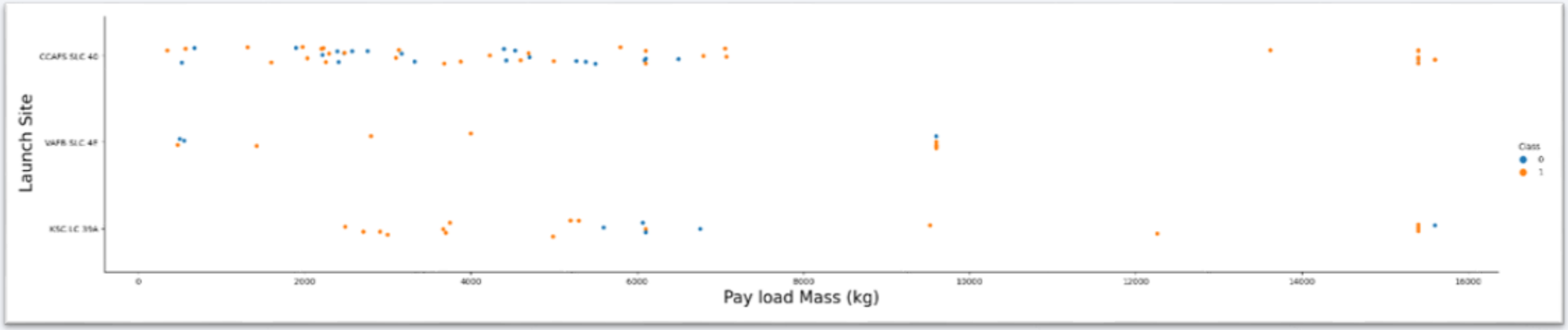
Flight Number vs. Launch Site

- **As you can see in the chart below:**
- Over time, the percentage of successful landings of each Launch Site improves
- The Launch Site VAFB-SLC-4E has not been used in at least the last 24 launches



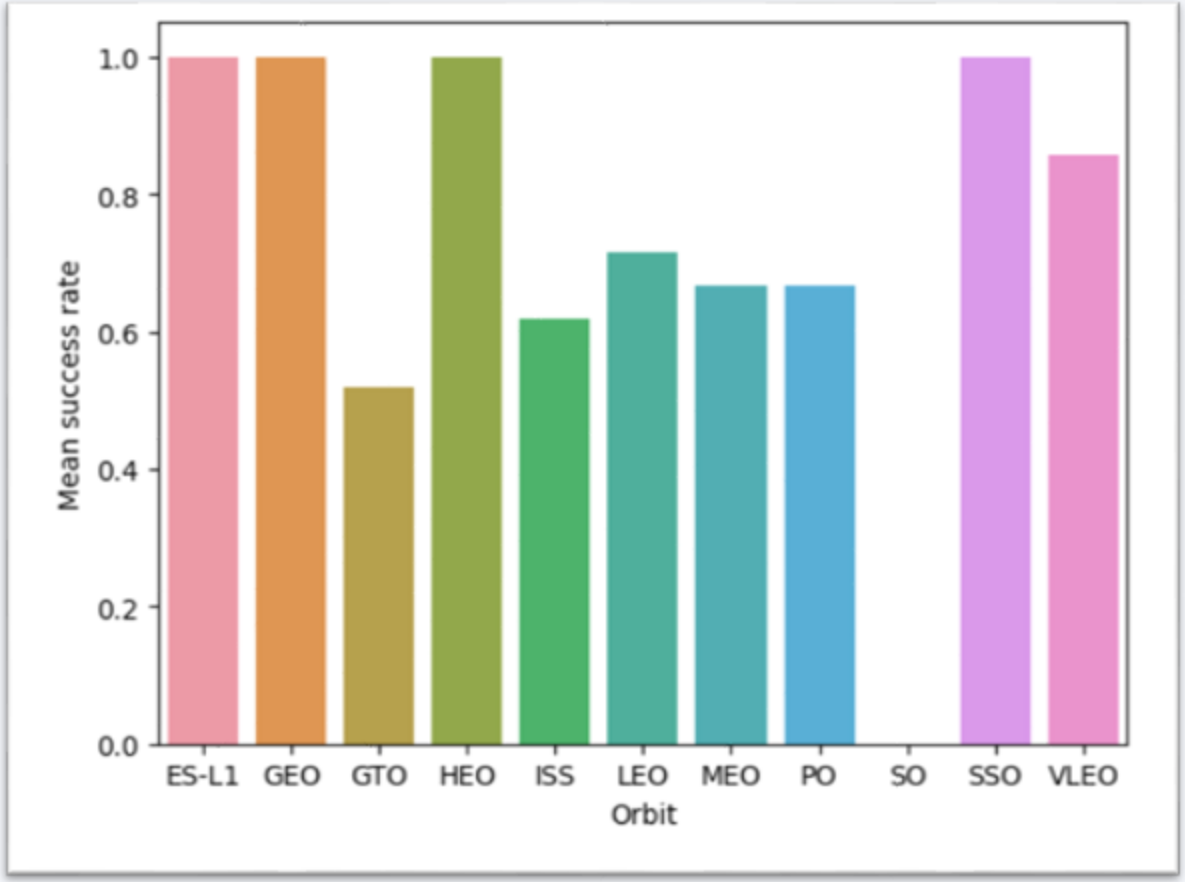
Payload vs. Launch Site

- **As you can see in the chart below:**
- The Launch Site VAFB-SLC-4E - there are no rockets launched for heavypayload mass (greater than 10000)
- Pay load Mass over 9 000 kg have better success rate than others



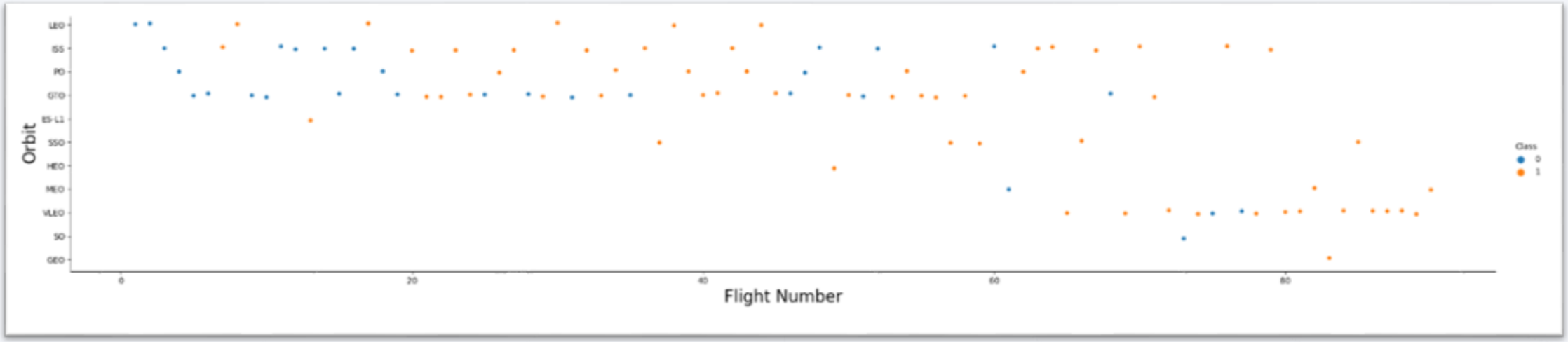
Success Rate vs. Orbit Type

- **As you can see in the chart on the right** Orbits have different success rate**:**
- ES-L1, GEO, HEO and SSO has the best success rate of 100%
- SO has the worst success rate of 0%



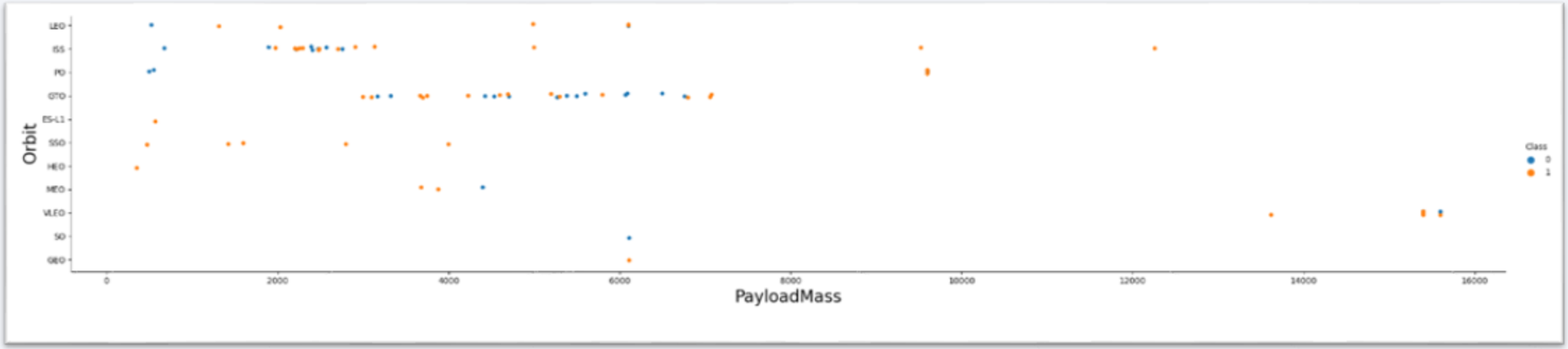
Flight Number vs. Orbit Type

- **As you can see in the chart below:**
- On the one hand, the LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency



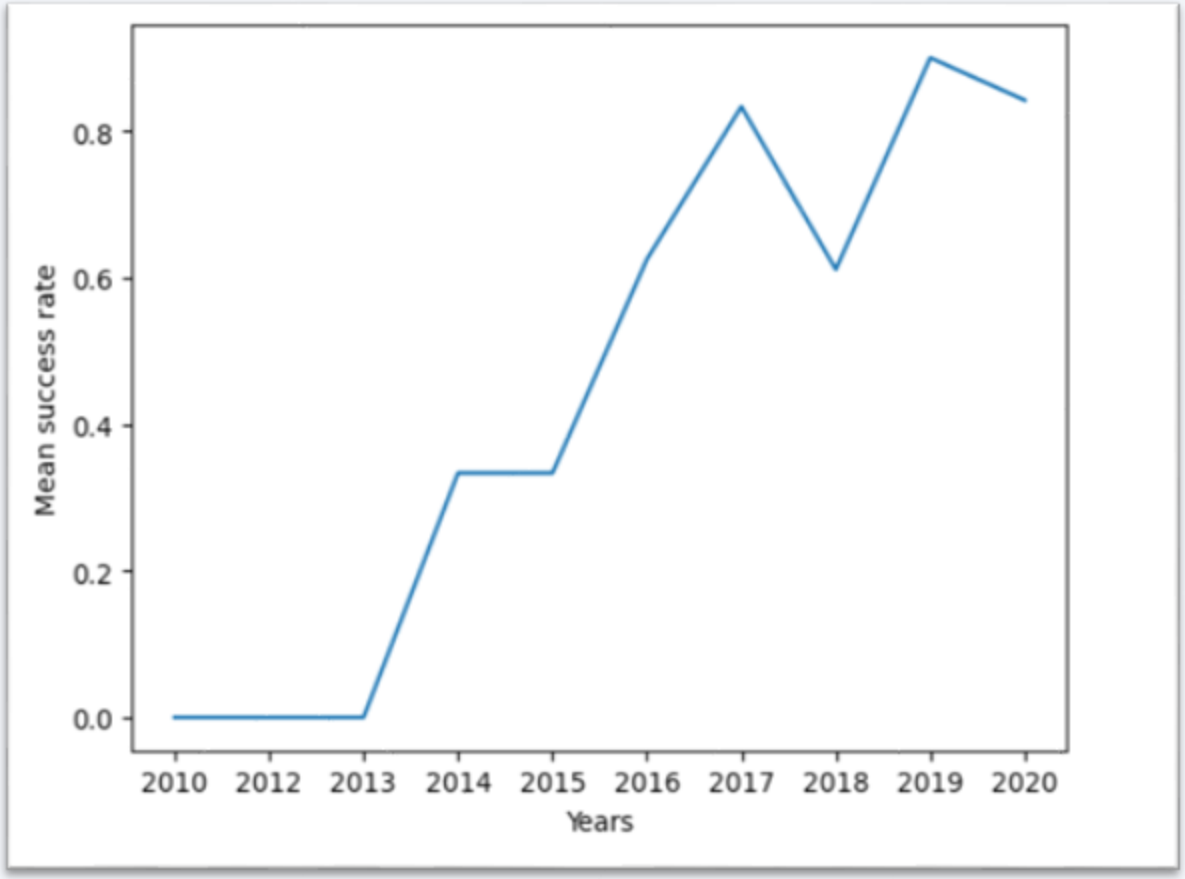
Payload vs. Orbit Type

- **As you can see in the chart below:**
- 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

- **As you can see in the chart on the right**
  The success rate since 2013 kept increasing till 2020



All Launch Site Names

- **As you can see in the table on the right,** there are 4 unique launch sites in the space mission (according to database).

- SQL query:
- SELECT DISTINCT Launch_Site FROM SPACEX

Out[11]:

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

Launch Site Names Begin with 'CCA'

- **You can see in the table below,** 5 database records where launch sites begin with the string 'CCA'. All of them is The Cape Canaveral Launch Complex 40

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- SQL query:
- SELECT * FROM SPACEX WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- **You can see in the table below** the total payload mass carried by boosters launched by NASA (CRS) (according to database)

Out[14]:

SUM(PAYLOAD_MASS__KG_)
45596

- SQL query:
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

- **You can see in the table below** the average payload mass carried by booster version F9 v1.1 (according to database)

Out[15]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- SQL query:
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE booster_version = 'F9 v1.1'

First Successful Ground Landing Date

- **You can see in the table below** the date when the first succesful landing outcome in ground pad was achieved (according to database)

Out[34]:

DATE
2015-12-22

- SQL query (SQLlite):

```
SELECT MIN(strftime('%Y-%m-%d', substr(DATE, 7, 4) || '-' || substr(DATE, 4, 2) || '-' || substr(DATE, 1, 2))) AS DATE
FROM SPACEX WHERE "Landing _Outcome" = "Success (ground pad)"
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- **You can see in the table below** the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 (according to database)

Out[36]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- SQL query:
- SELECT booster_version, PAYLOAD_MASS__KG_
- FROM SPACEX
- WHERE "Landing _Outcome" = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)

Total Number of Successful and Failure Mission Outcomes

- **You can see in the table below** the total number of successful and failure mission outcomes (according to database)

Out[37]:

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

- SQL query (SQLlite):

```
SELECT mission_outcome, COUNT(mission_outcome) as count FROM SPACEX GROUP BY mission_outcome
```

Boosters Carried Maximum Payload

- **You can see in the table on the right,** the names of the booster versions which have carried the maximum payload mass (according to database).

Out[38]:

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

- SQL query:
- SELECT booster_version FROM SPACEX
- WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)

2015 Launch Records

- **You can see in the table below** the records which display the failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015 (according to database)

Out[39]:

Month	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- SQL query (SQLlite):

```
SELECT substr(Date, 4, 2) as Month, booster_version, "Landing _Outcome", launch_site
FROM SPACEX WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

- **You can see in the table on the right,** the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order (according to database).

Out[42]:

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

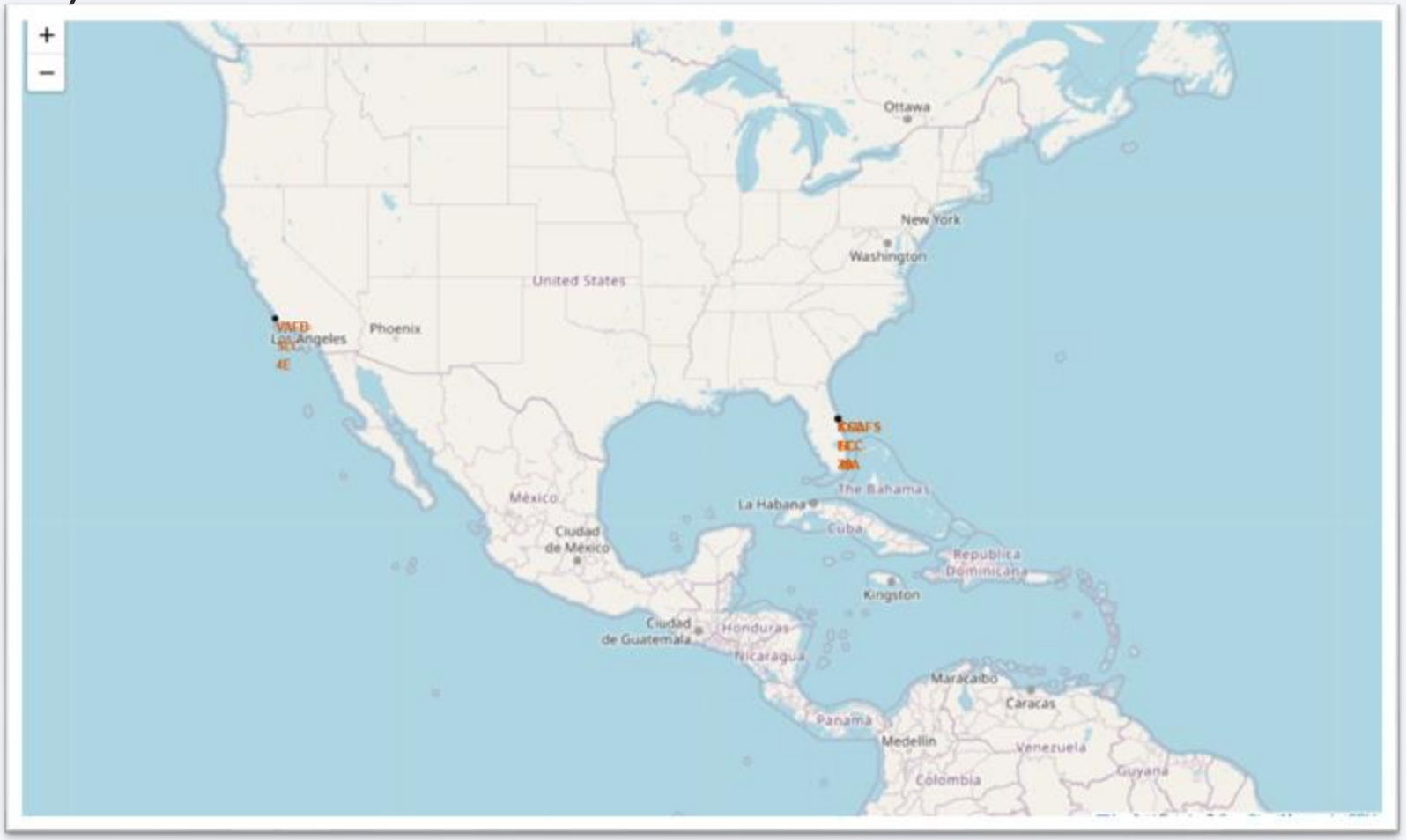
- SQL query (SQLlite):

```
SELECT "Landing _Outcome", COUNT("Landing _Outcome") as count
FROM SPACEX
WHERE strftime('%Y-%m-%d', substr(DATE, 7, 4) || '-' ||
substr(DATE, 4, 2) || '-' || substr(DATE, 1, 2)) BETWEEN '2010-06-04'
AND '2017-03-20'
GROUP BY "Landing _Outcome" ORDER BY count DESC
```


Section 3
Launch Sites
Proximities Analysis

All launch sites' location

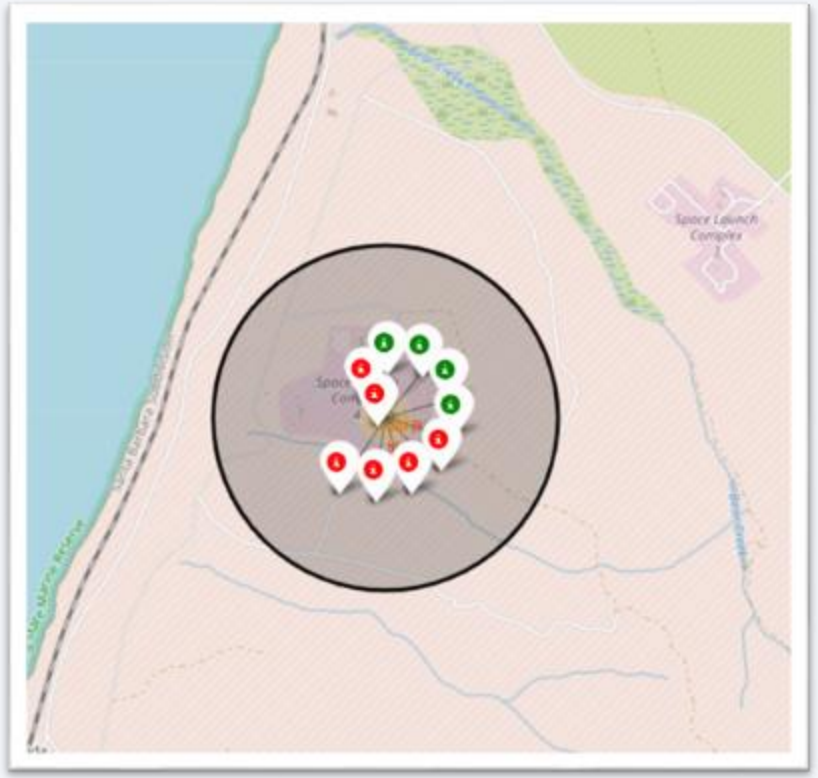
- **As you can see in the map below** All launch sites located close to coastline and approximately the same distance from the equator, but one of them on the west coast (VAFB SLC-4E):



Launch outcomes map

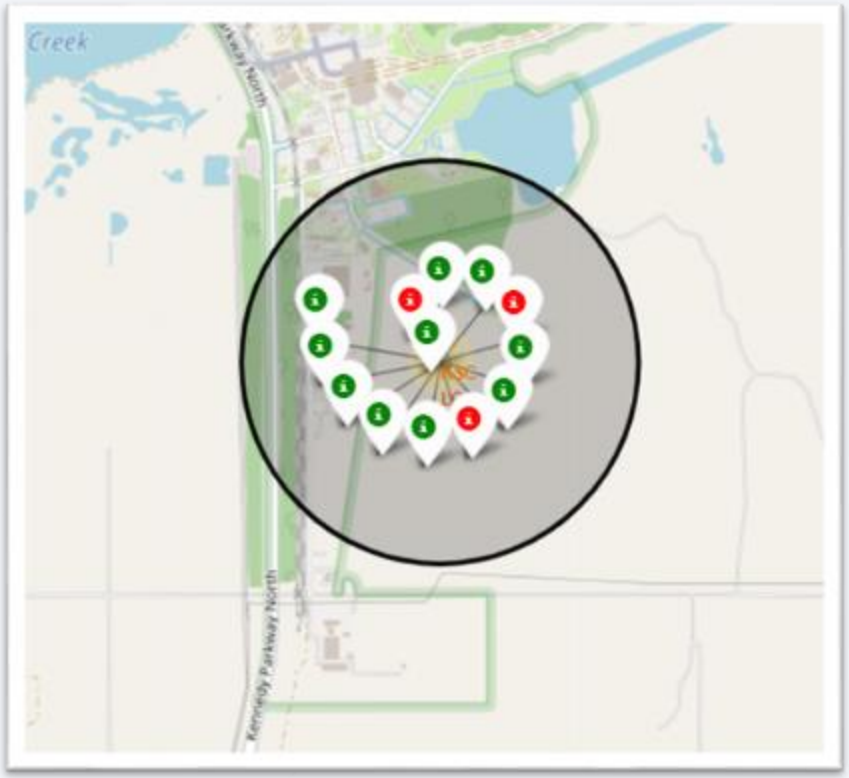
- **As you can see in the map below** All launch sites have a different ratio of successful (**green markers**) **and** failure launch outcomes (**red markers**)

Launch Site *VAFB-SLC-4E*



40% / 60%

Launch Site *KSC LC-39A*

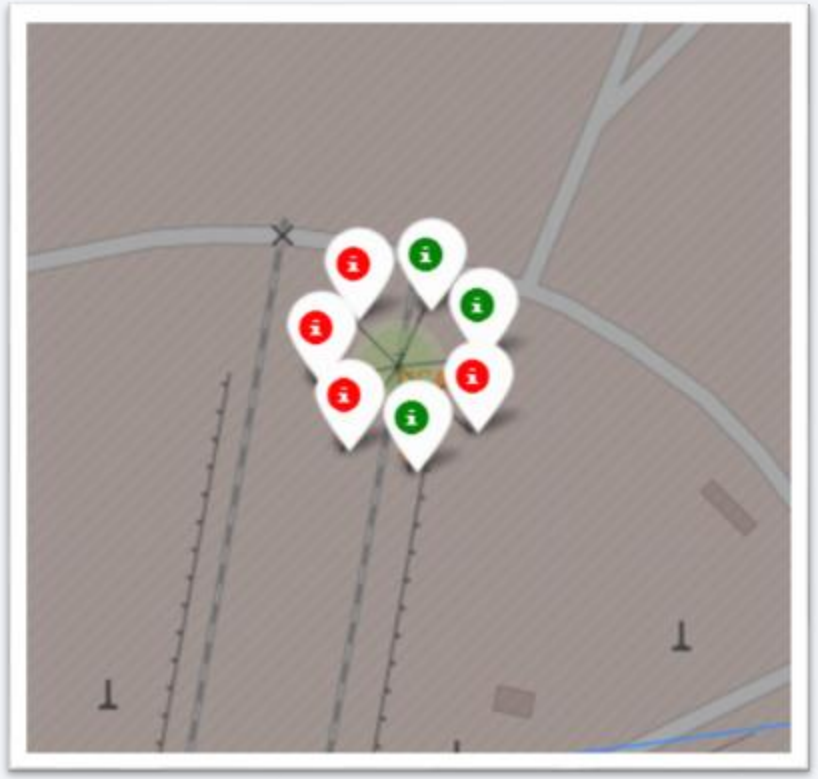


77% / 23%

Launch outcomes map

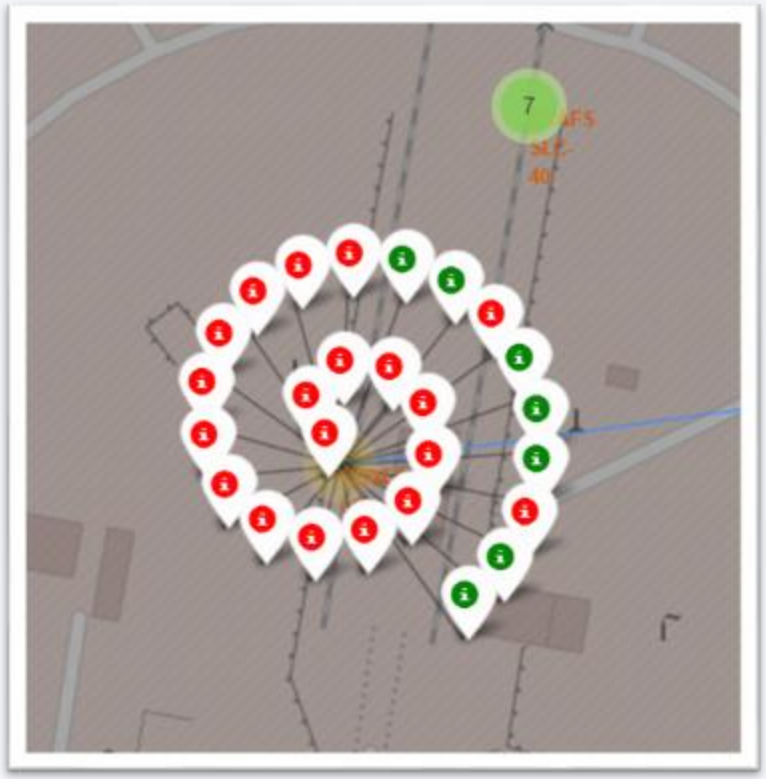
- **As you can see in the map below** All launch sites have a different ratio of successful (**green markers**) **and** failure launch outcomes (**red markers**)

Launch Site *CCAFS-SLC-40*



43% / 57%

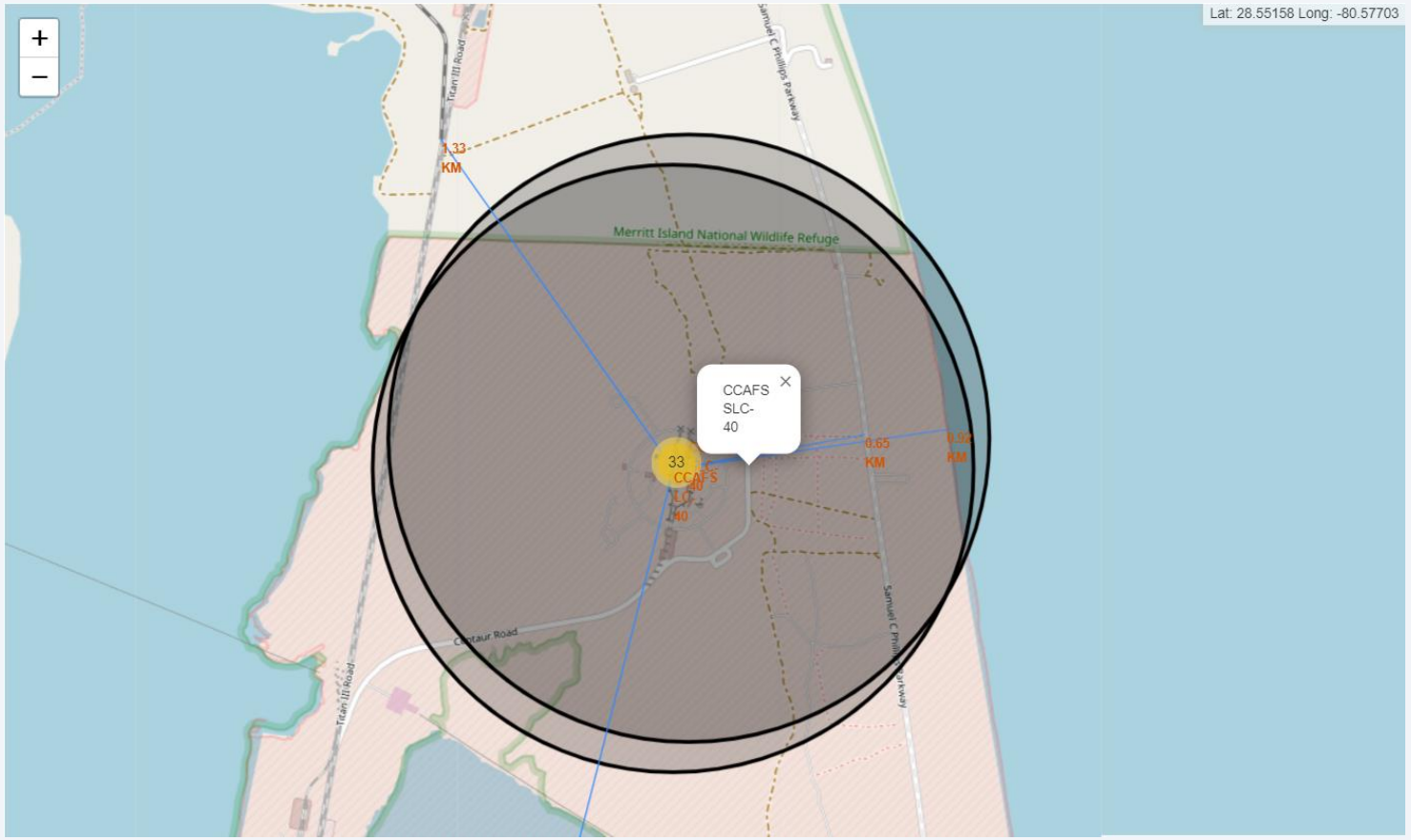
Launch Site *CCAFS-LC-40*

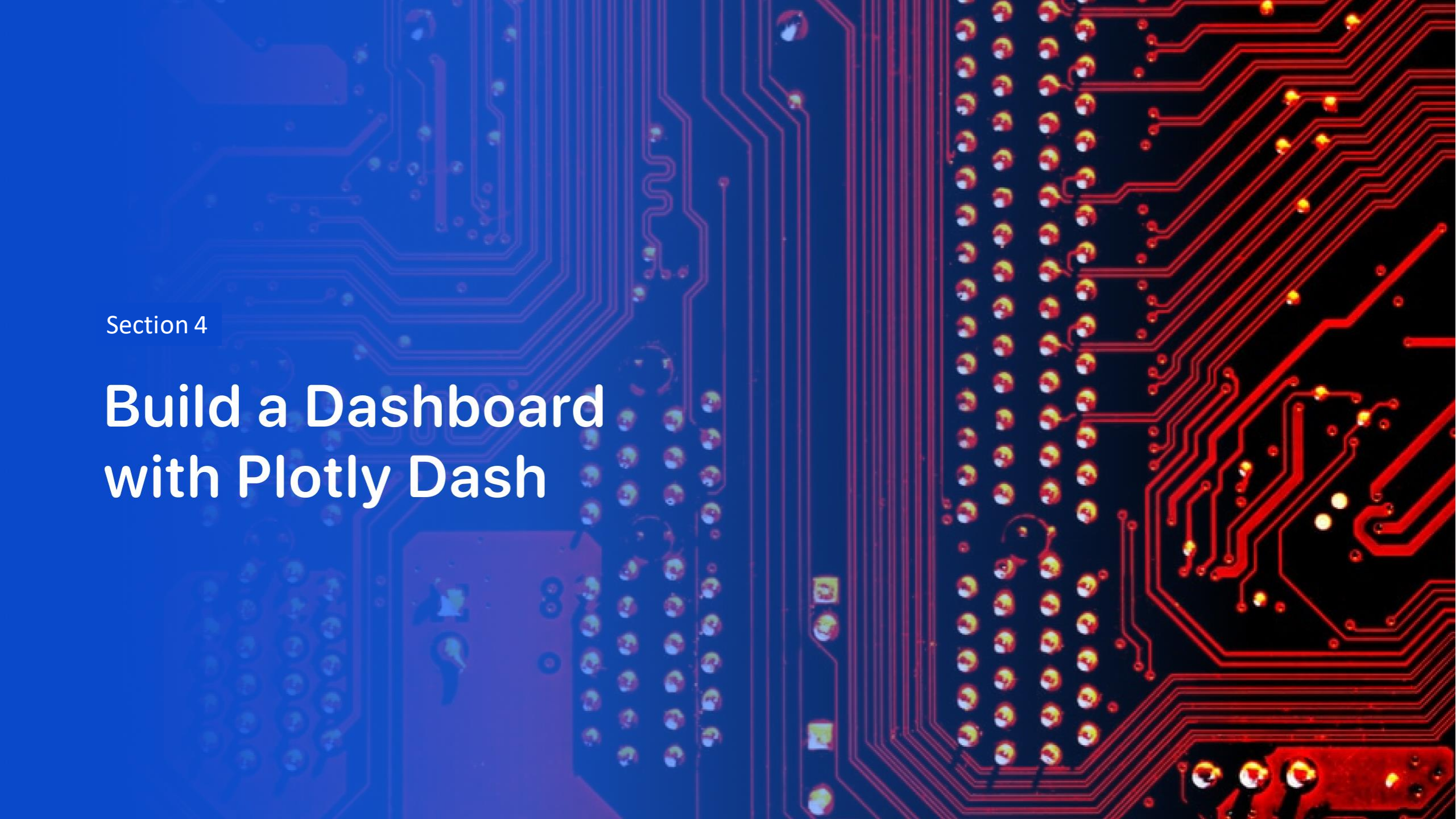


27% / 73%

Launch Site CCAFS-LC-40 (proximities)

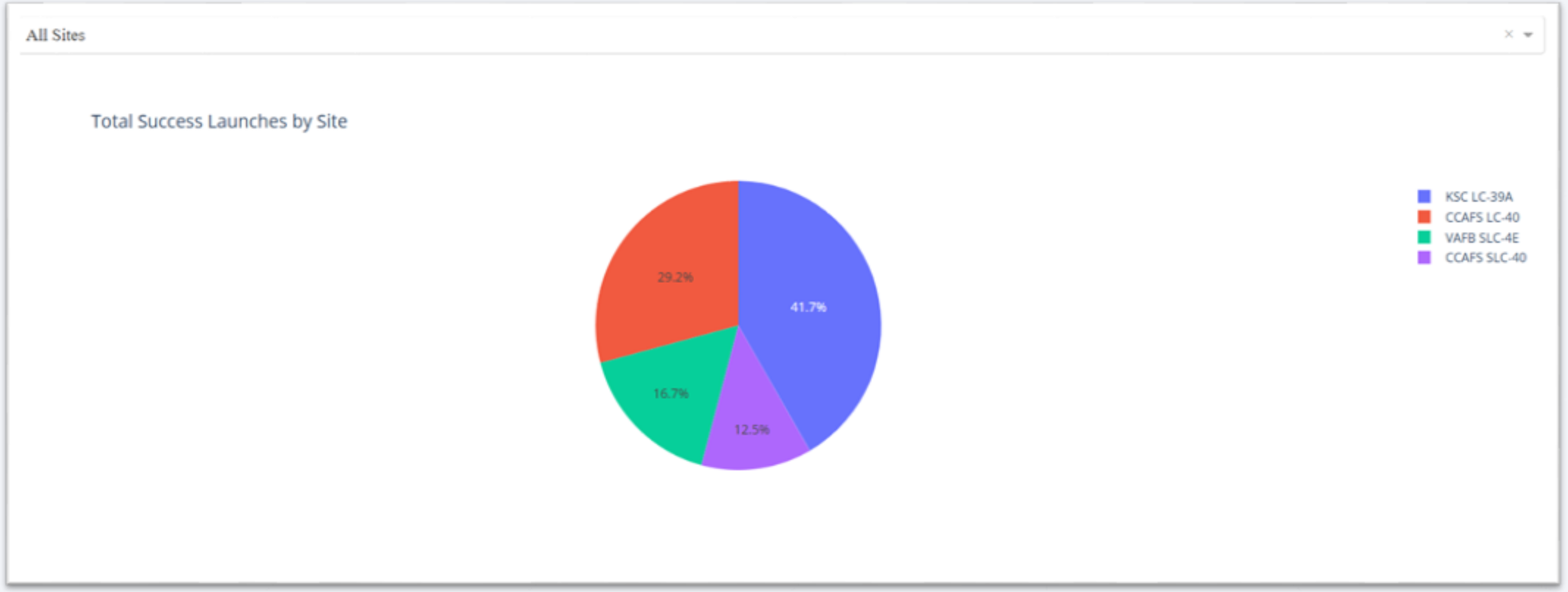
- **As you can see in the map below** CCAFS-LC-40 has next distance to proximities: Coastline – 0.92 KM, Highway – 0.65 KM, Railway – 1.33 KM, Port – 16.95 KM




Section 4
Build a Dashboard with Plotly Dash

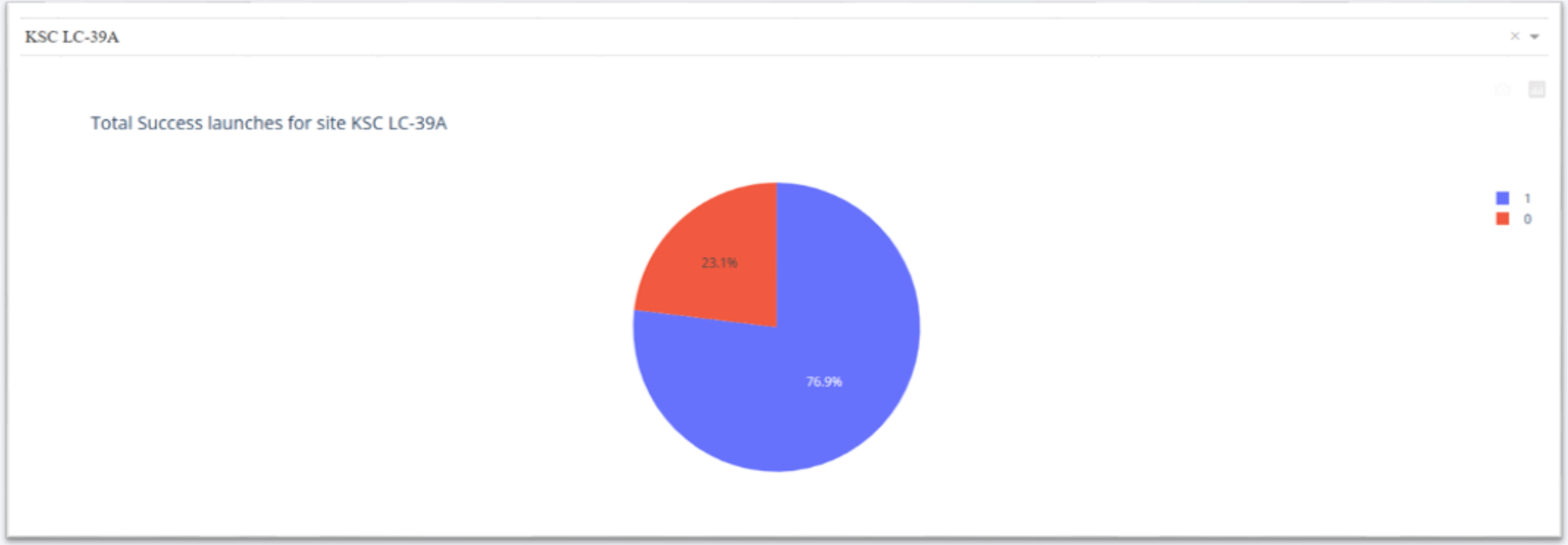
Total Success Launches by Site

- **As you can see in the pie chart below** KSC LC-39A has the largest number of success launches among all sites



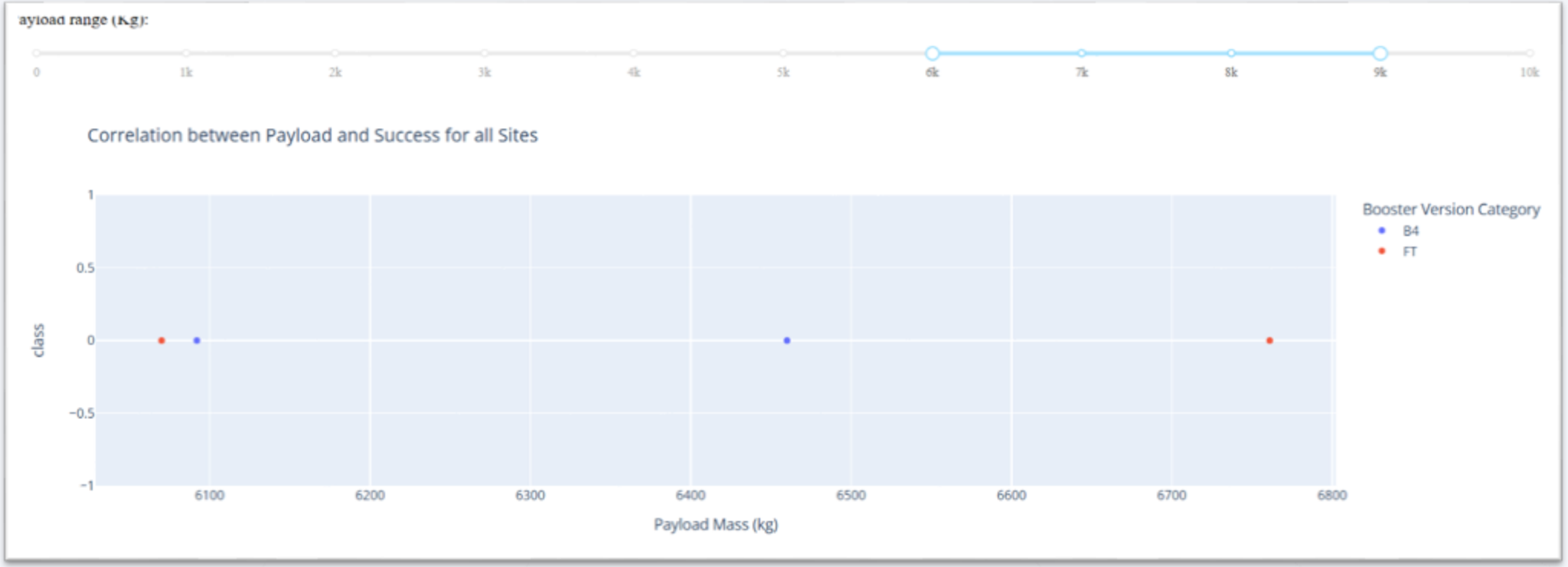
Launch Success Ratio for KSC LC-39A

- **As you can see in the pie chart below** there are 76.9% of all launches was successful



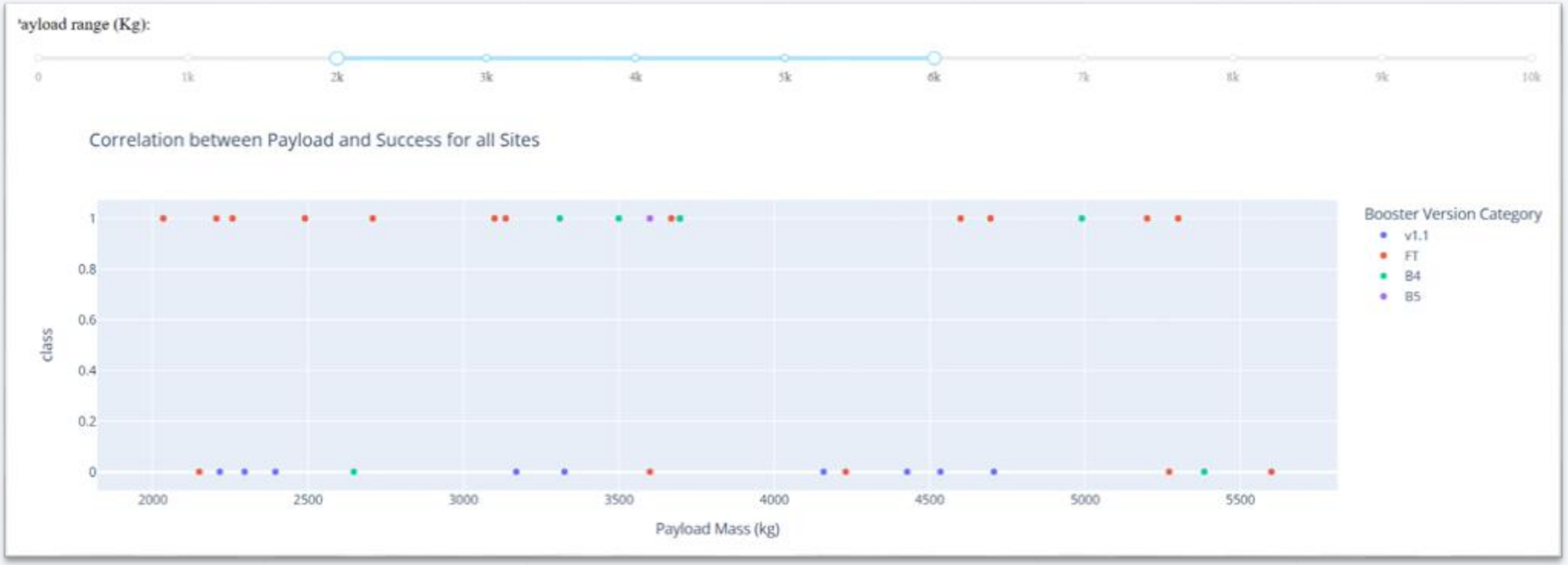
Correlation between Payload and Success for All Sites

- **As you can see in the scatter point chart below** there are not successful launch outcomes with Payload between 6000 KG and 9000 KG



Correlation between Payload and Success for All Sites

- **As you can see in the scatter point chart below, obviously** FT is the most successful booster version category among launches with Payload between 2000 KG and 6000 KG

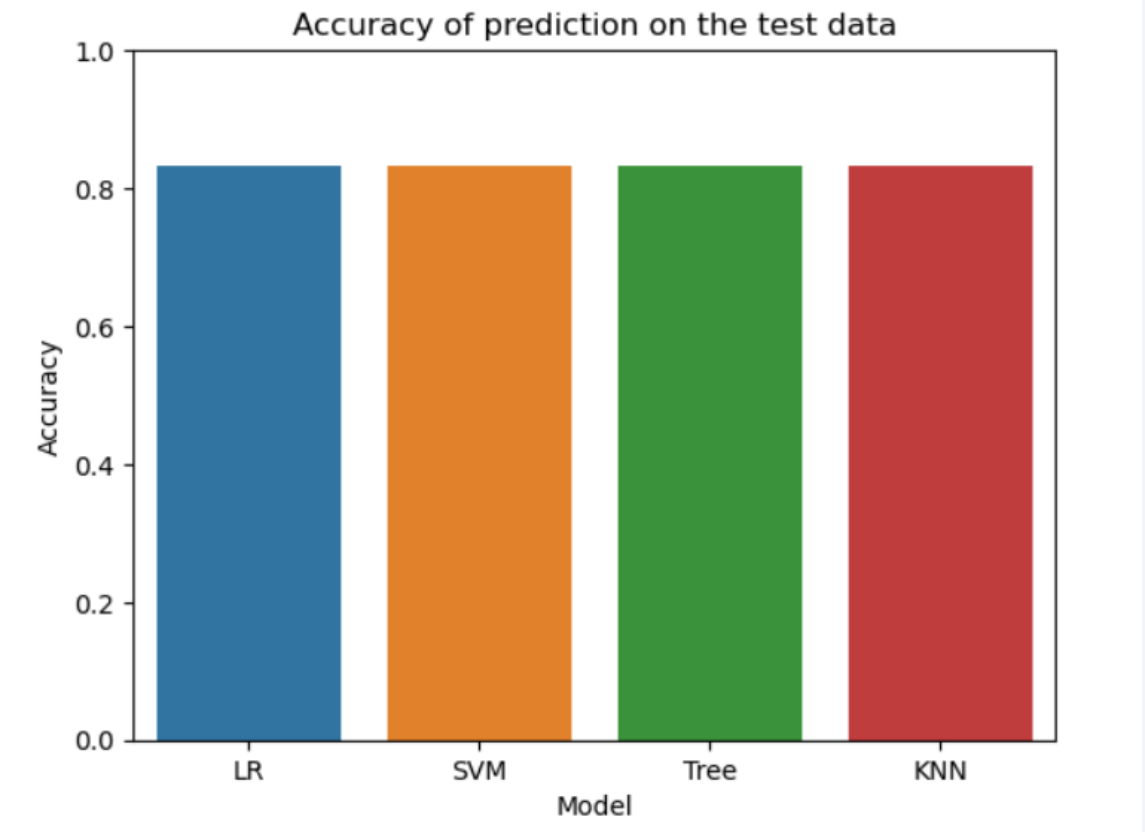


Section 5

Predictive Analysis (Classification)

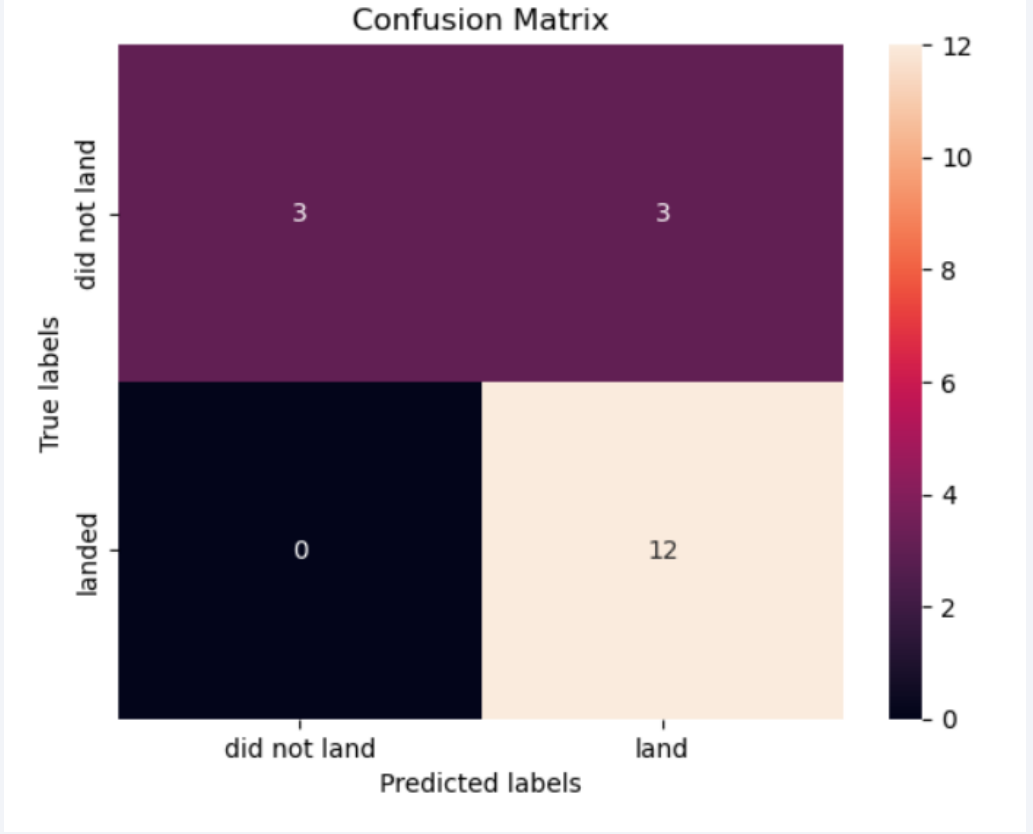
Classification Accuracy

- **As you can see in the bar chart on the right there are the same** built model accuracy on the test data for all built classification models
- All these algorithms give the same result 0.833:
- *Logistic Regression – LR*
- *Support Vector Machine – SVM*
- *Decision Tree – Tree*
- *K-nearest neighbors - KNN*



Confusion Matrix

- **Since there are the same** built model accuracy on the test data for all built classification models, the confusion matrix also matches.
- As you can see, the major problem is false positives.



Conclusions

- This project aimed to train machine learning models to predict whether the first stage of a Falcon 9 rocket launch by SpaceX would successfully land or not.
- The rationale for this was that knowing whether the first stage would land could help determine the cost of the launch, given that SpaceX is able to reuse the first stage and thus provide significant cost savings compared to other rocket providers.
- Our results show that machine learning models can be trained with reasonable accuracy (83%) to predict whether the first stage will land or not, based on various input features.
- Overall, this project demonstrates the potential for machine learning to contribute to the space industry by providing predictive insights that can inform decision-making and improve efficiency.

Appendix

- As I was unable to complete some steps of this project on Watson Studio and Db2, I used other tools instead: 'Skills Network Labs', VSCode, and SQLite
- It turns out that SQLite doesn't really have a date column type, so the DD-MM-YYYY formatted dates are sorted as strings. Additionally, SQLite does not support month names, which is why I needed to use the 'substr' function.

Thank you!