IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Johannes Pollmanns

16.08.2023

Outline

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

Executive Summary I

- Methods:
    - Data collection through API and web scraping. Sources: SpaceX API (spacexdata.com) and Wikipedia (Falcon9 Launch page)
    - Data wrangling
    - Exploratory Data Analysis (EDA) with SQL + data visualization
    - Interactive map (with Folium)
    - Dashboard (with Plotly Dash)
    - Predictive analysis

Executive Summary II

- Results:
  - Some features of the launches were correlated with a positive outcome (i.e. successful landing).
  - Most recent flights, mainly to orbit VLEO, had a good success rate.
  - Booster version with highest success rate is version FT.
  - Payload mass might play a role for outcome (with worse outcome for midrange payload mass) but is also dependent from target orbit.
  - Decision tree is the best machine learning algorithm for prediction of launch outcome.

Introduction

Project background and context

SpaceY wants to bid against SpaceX for rocket launches. SpaceX advertises Falcon9 rocket launches 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 SpaceY can determine if the first stage will land, they can determine the cost of a launch. The goal of this project is to predict if the first stage of a Falcon9 will land successfully.

Research questions

- What are predictors of a successful landing of the first stage?
- What are the relationships between those predictors and the outcome?

Section 1

Methodology

Methodology

Executive Summary

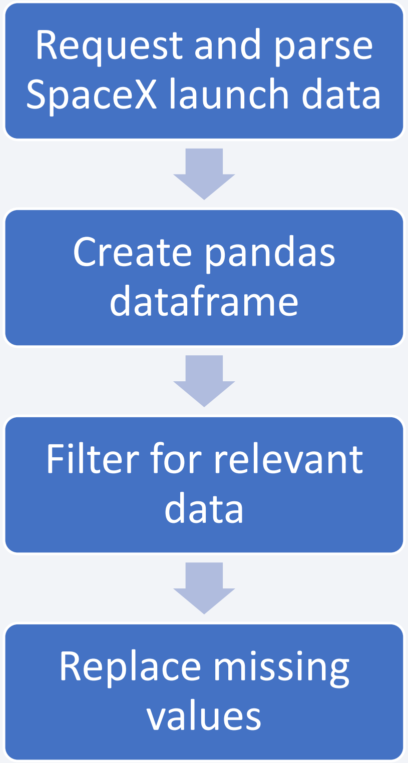
- Data collection methodology:
  - SpaceX API (api.spacexdata.com), Web Scraping (Falcon9 Launch Wikipedia page)
- Perform data wrangling
  - Handling missing values, One-Hot-Encoding for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, Decision tree, KNN

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

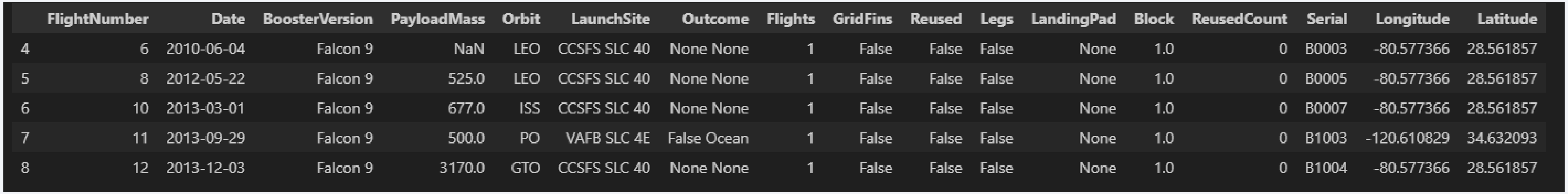
- API used: https://api.spacexdata.com/v4/launches/past
- Data was request and parsed using a get request
- JSON result was converted into a pandas dataframe
- Data was filtered to include only Falcon9 launches
- Missing values in PayloadMass were imputed with Mean
- GitHub-URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/01_jupyter-labs-spacex-data-collection-api.ipynb


Request and parse SpaceX launch data

Create pandas dataframe

Filter for relevant data

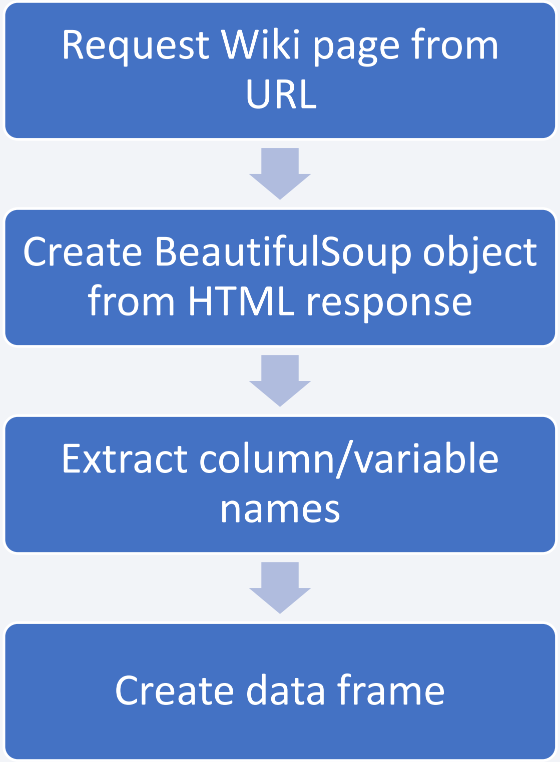
Replace missing values



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- Source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- For Scraping, the BeautifulSoup library has been used
- Data was converted into a Pandas dataframe
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/02_jupyter-labs-webscraping.ipynb

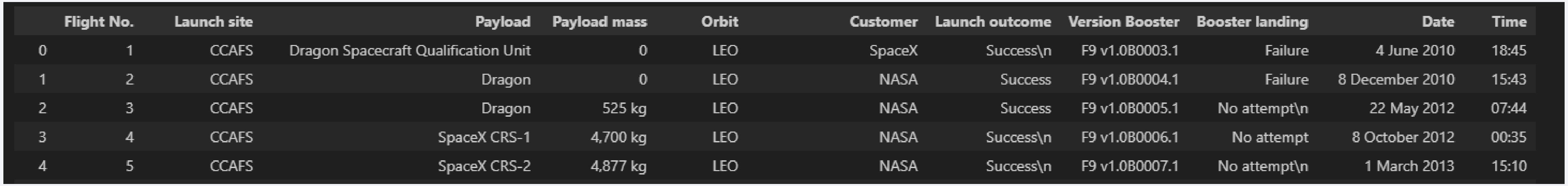


Request Wiki page from URL

Create BeautifulSoup object from HTML response

Extract column/variable names

Create data frame



	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

1. Calculated the number of launches at each site
2. Calculated the number and occurrence of each orbit
3. Calculated the number and occurrence of mission outcome/orbit type

➢ Pandas value_counts()-method has been used

➢ Finally, a landing outcome label has been created from the outcome column (bad outcome = 0, otherwise = 1)

- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/03_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Scatter charts for:
  - Flight Number / Launch Site
  - Payload / Launch Site
  - Flight Number / Orbit type
  - Payload / Orbit type
  - ➢ Visualizes relationship of those variables and outcome class
- Bar chart: Success rate of each orbit type
  - ➢ Analyzes relationship between success rate and orbit type
- Line chart: Trend of launch success over time
  - ➢ Yearly launch success rate
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/04_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Data has been queried using SQL for descriptive analysis:
  - Names of unique launch sites
  - Records where launch sites begin with ‘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
  - Booster names with have success in drone ship and payload between 4000 and 6000
  - Total number of successful and failure mission outcomes
  - Names of booster versions which have carried the maximum payload mass
  - Selected records for 2015
  - Count of landing outcome for selected dates
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/05_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Coordinates were used for each launch site to create circle markers around each launch site with name labels to identify those launch sites.
- The launch outcomes for each launch site are displayed markers on the map (using MarkerCluster)
- Distance between launch sites and landmarks (like nearest coastline) has been plotted with PolyLine.
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/06_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A Dashboard has been build using Plotly Dash to regularly show the data in an easy-to-understand manner. This includes:
  - Pi Charts for successful launches by site and launch outcomes for each site to visualize relationships
  - Scatter graph for outcome and payload mass for different booster versions showing relationships and range of data
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)

- For predictive analysis, the scikit-learn library has been used
- Preprocessing of data included standardization of variables and splitting the data into training and test data
- The following machine learning methods have been used for modelling: Logistic regression, Support vector machines (SVM), Decision tree and K-nearest neighbor (KNN)
- The models have been fitted on the training data
- Hyperparameter grid search has been used to find optimal parameters
- Models have been evaluated based on accuracy and confusion matrix
- GitHub URL for notebook: https://github.com/j-poll/IBM_DS_Capstone/blob/main/08_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

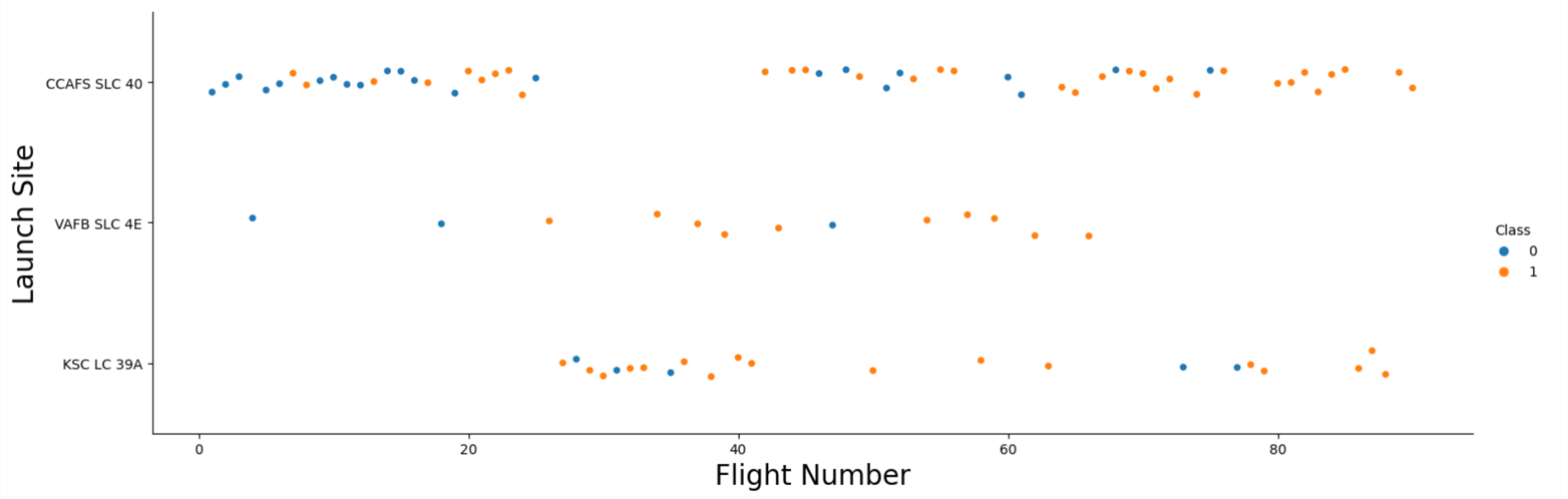
Results

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

Section 2

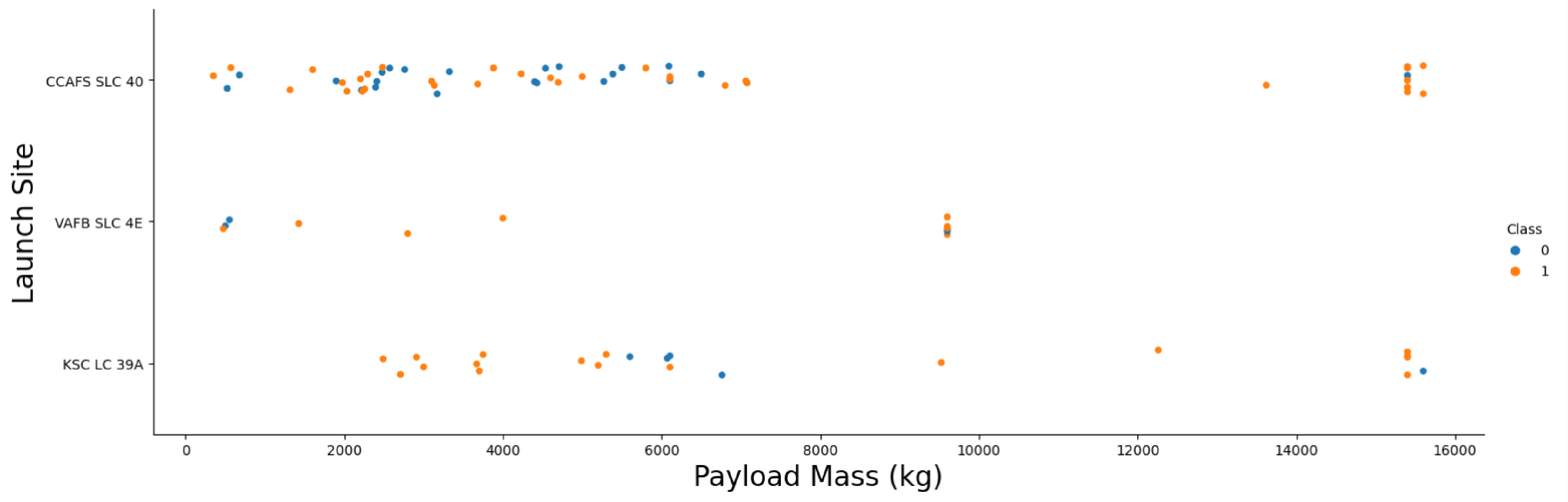
Insights drawn from EDA

Flight Number vs. Launch Site



Flight Number and classes seems to be correlated. Later flight numbers show more successes than early flight numbers (class 1 = success). Most of the early flights were started at CCAFS SLC 40.

Payload vs. Launch Site

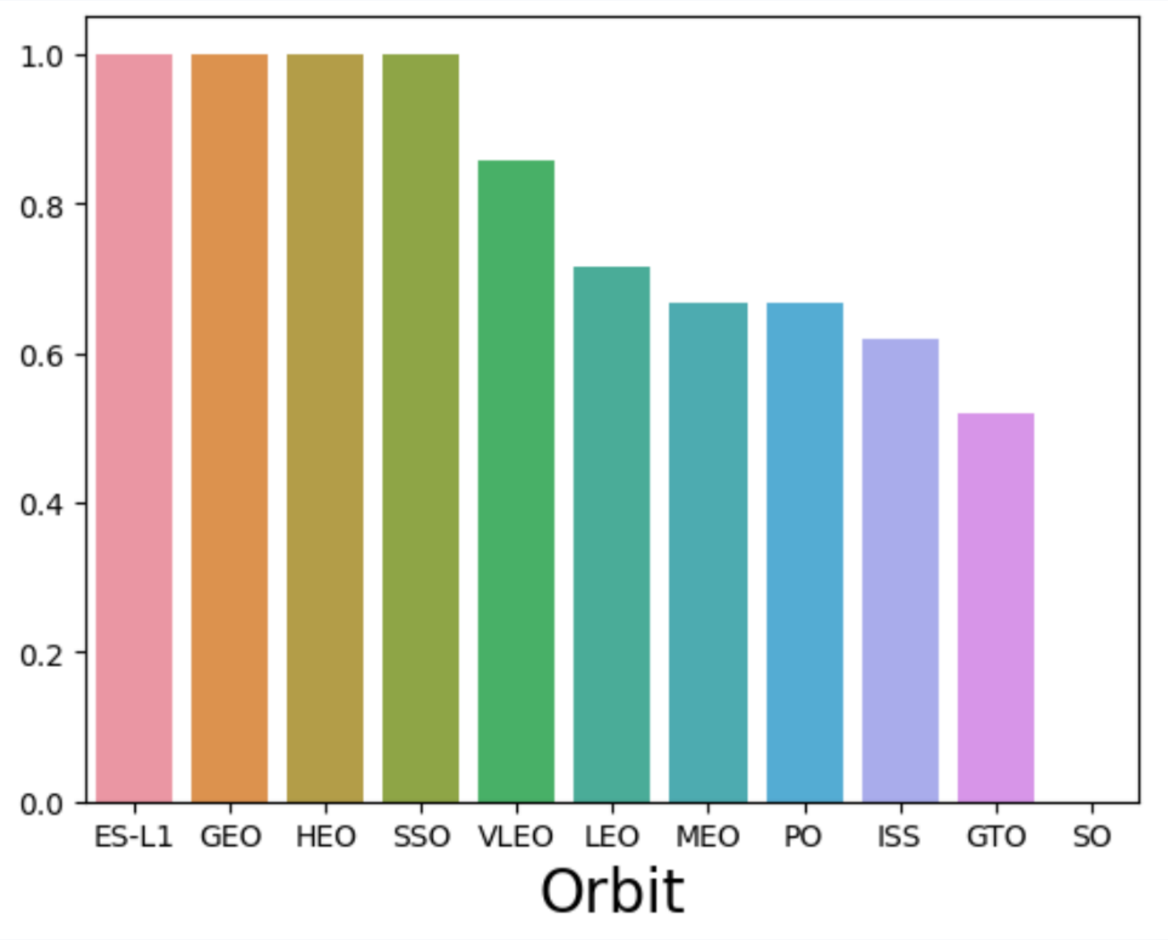


Flights with a higher Payload Mass at Launch Site CCAFS had a higher success rate than flights with lower Payload Mass. However, at KSC LC was a much better success rate for flights with lower Payload Mass.

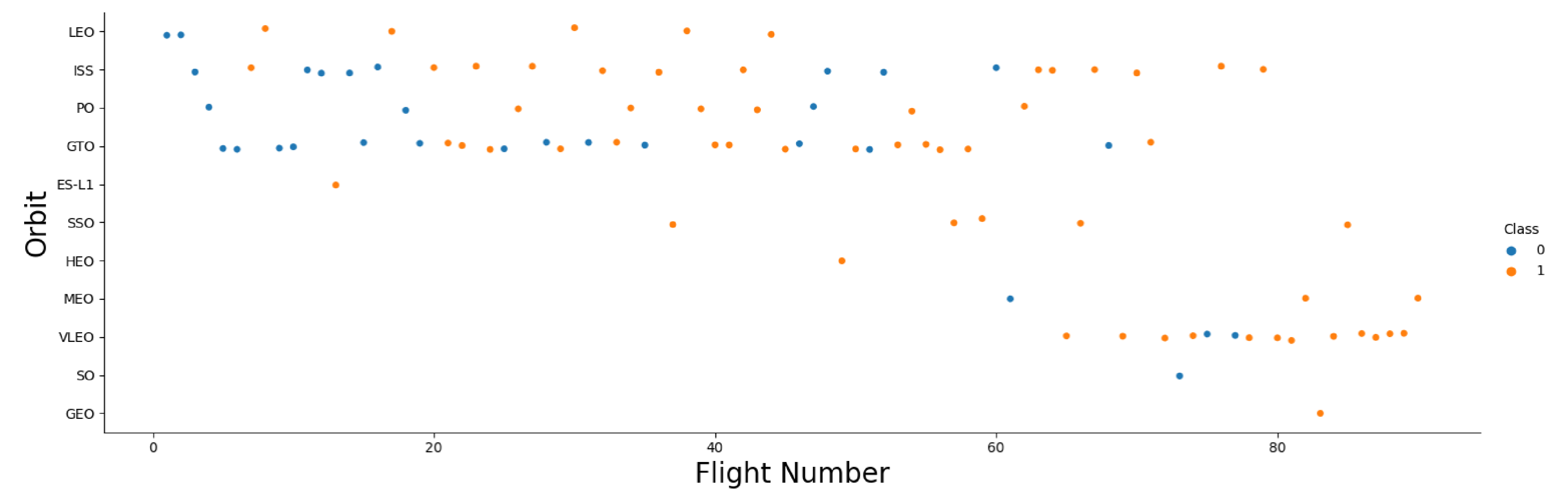
Success Rate vs. Orbit Type

There was a 100% success rate for flights into orbits ES-L1, GEO, HEO and SSO.

Other orbits had a lower success rate. There was no successful launch for orbit SO.

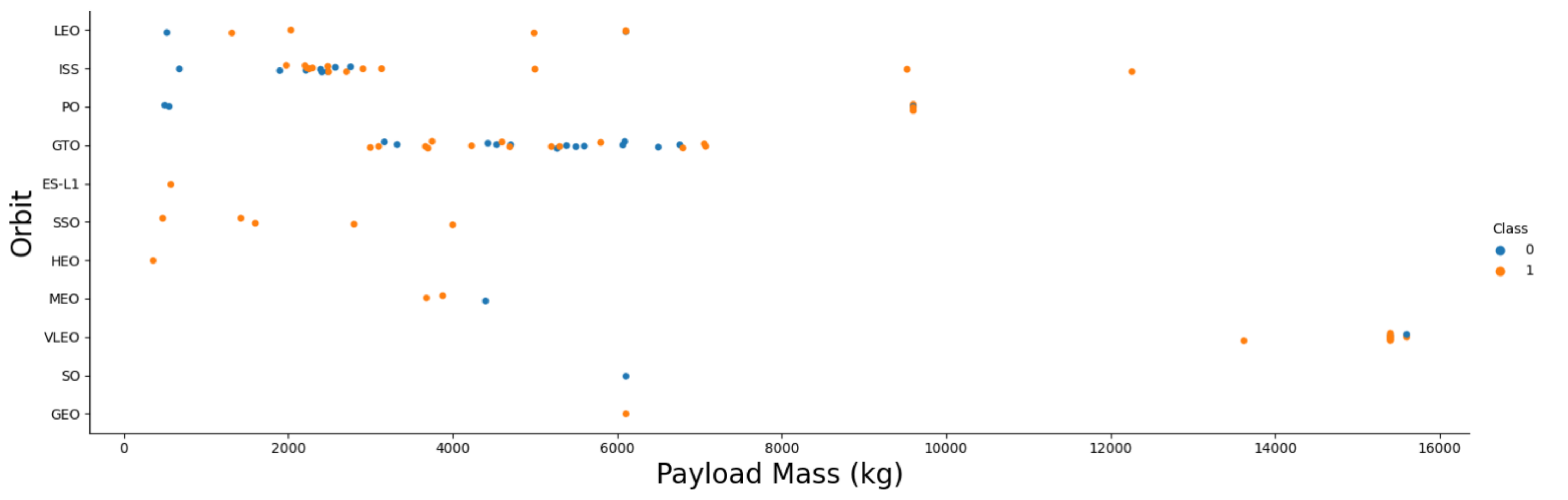


Flight Number vs. Orbit Type



Early flights were focused primarily on orbits ISS and GTO (with a low success rate) while later flights focused on orbit VLEO (with a good success rate).

Payload vs. Orbit Type

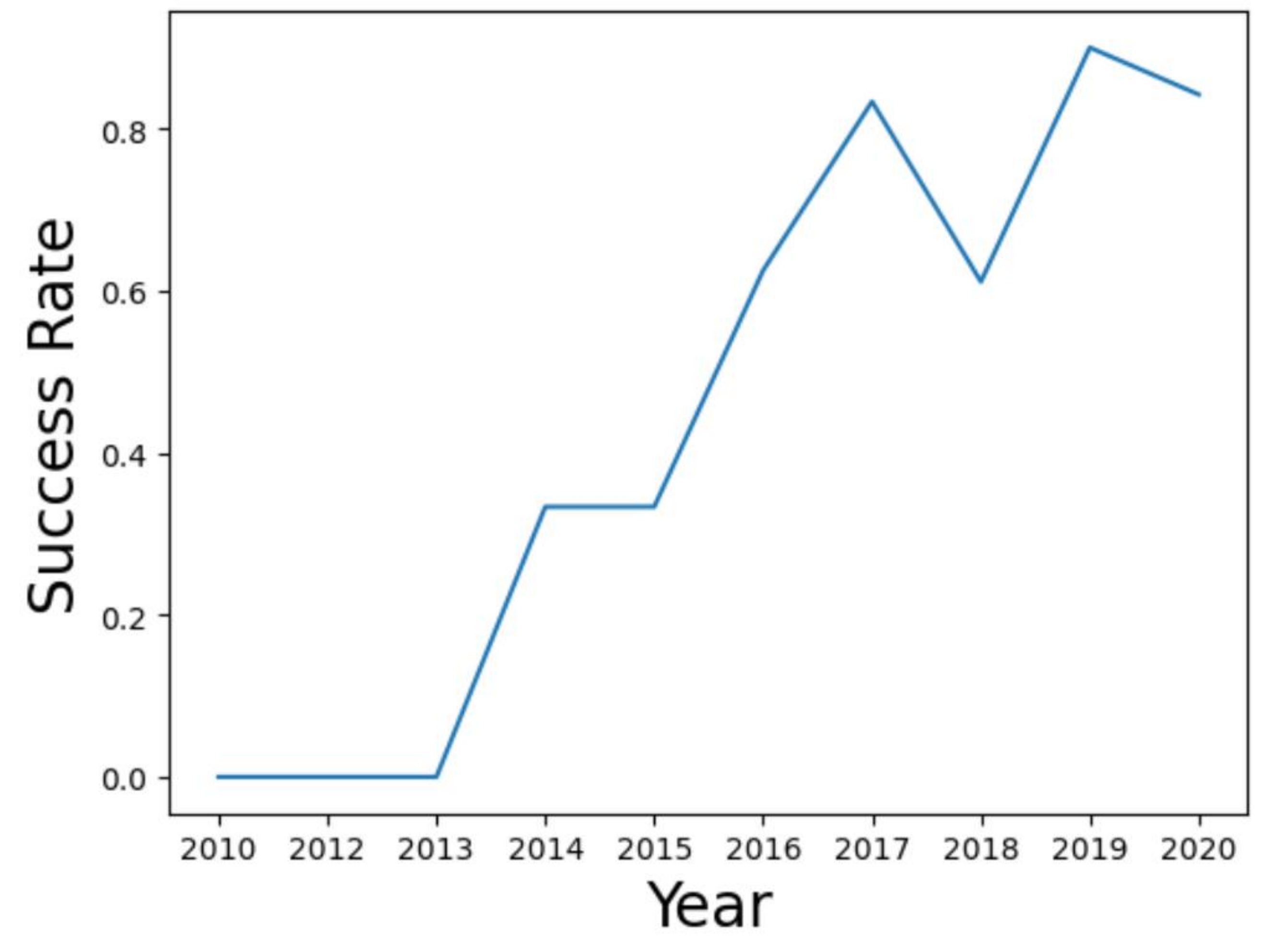


Flights to orbit ISS had rather low payloads. Flights to GTO had significantly higher payloads while flights to orbit VLEO had the highest payloads.

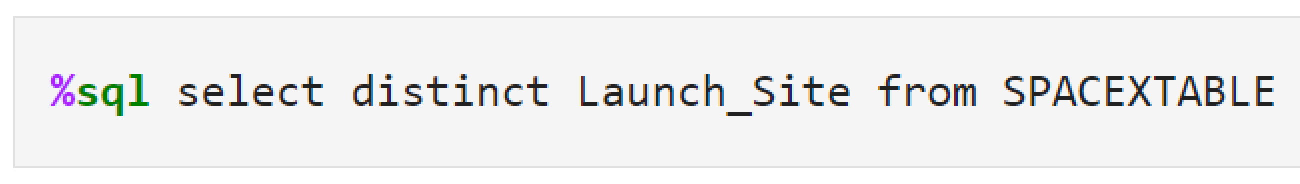
Launch Success Yearly Trend

Overall, we can see a positive trend for success rate over time. Later flights had on average a higher success rate than early flights.

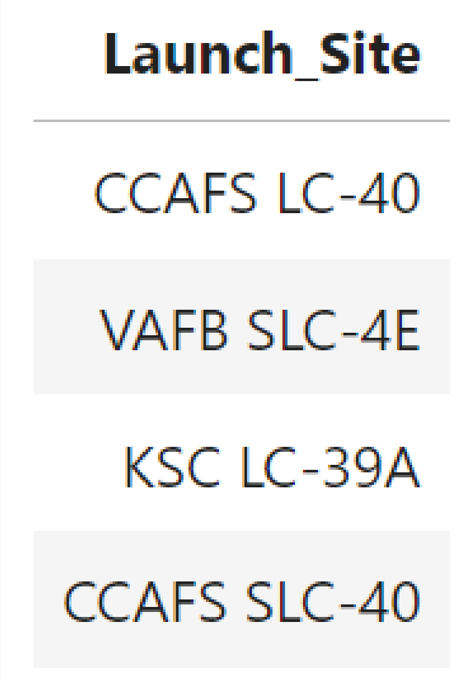
We can see a drop in success rate in 2018 but it remains unclear if this difference is statistically significant.



All Launch Site Names



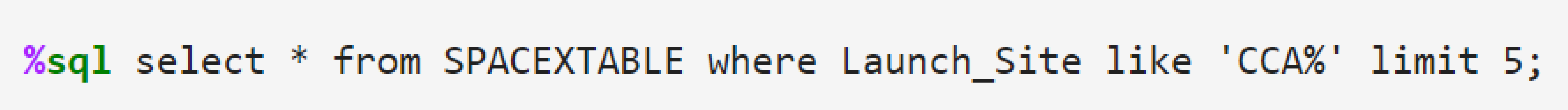
```
%sql select distinct Launch_Site from SPACEXTABLE
```



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

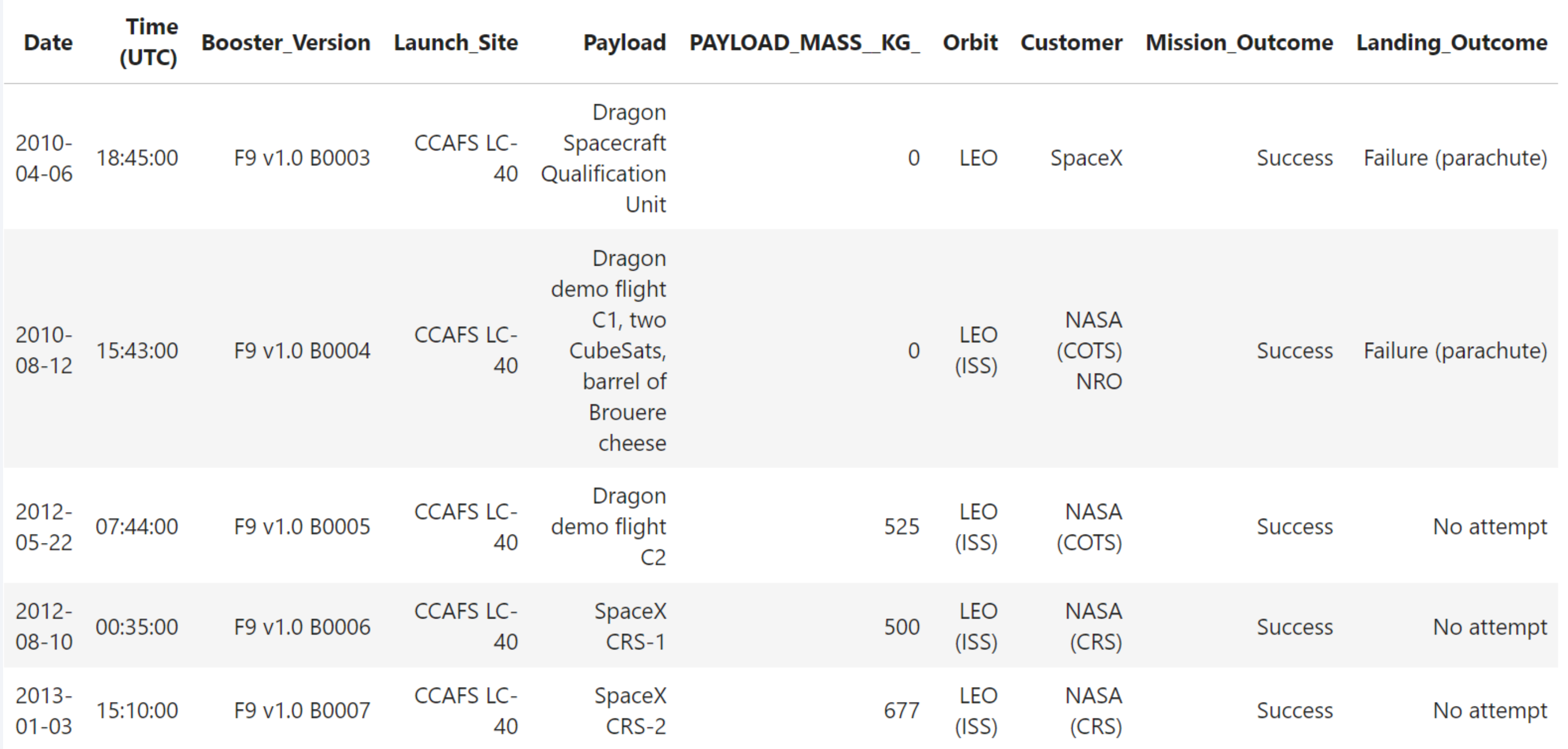
There are four different launch sites in the dataset.

Launch Site Names Begin with 'CCA'



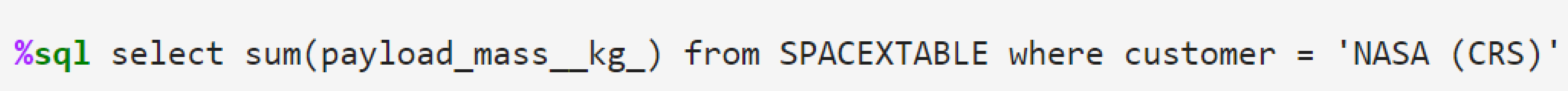
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;
```

5 records where launch sites begin with `CCA`



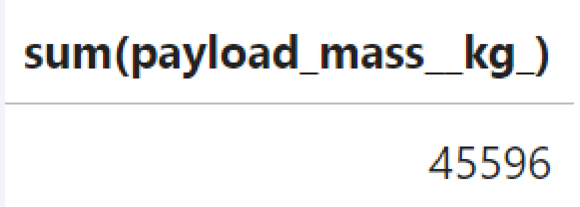
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass



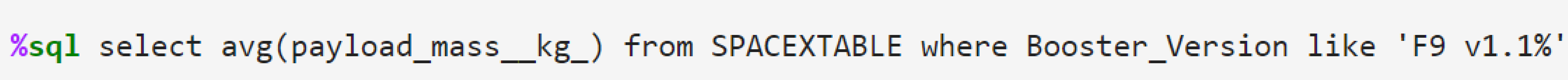
```
%sql select sum(payload_mass__kg_) from SPACEXTABLE where customer = 'NASA (CRS)'
```

The total payload carried by boosters from NASA is 45,596 kg.



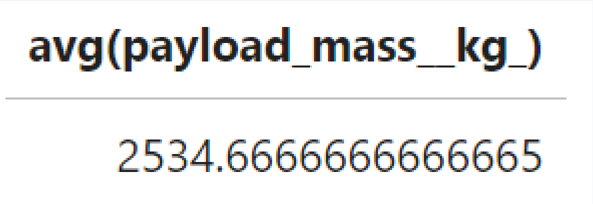
sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1



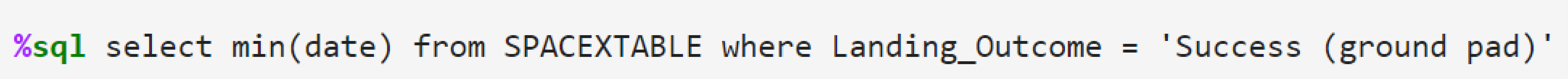
```
%sql select avg(payload_mass__kg_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

The average payload mass carried by booster version F9 v1.1 is 2534.67 kg.



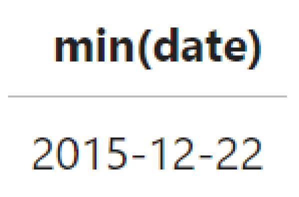
avg(payload_mass__kg_)
2534.6666666666665

First Successful Ground Landing Date



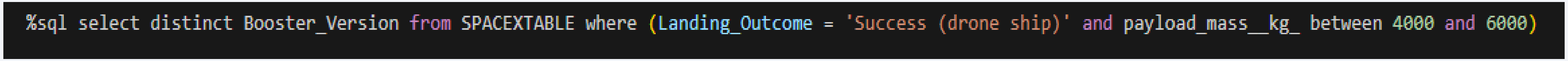
```
%sql select min(date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

The first successful landing outcome on ground pad was on 22.12.2015.



min(date)
2015-12-22

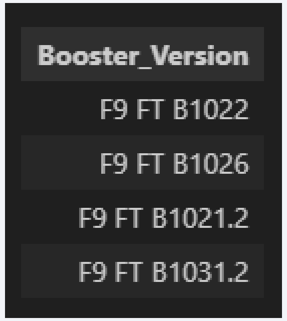
Successful Drone Ship Landing with Payload between 4000 and 6000



```
%sql select distinct Booster_Version from SPACEXTABLE where (Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000)
```

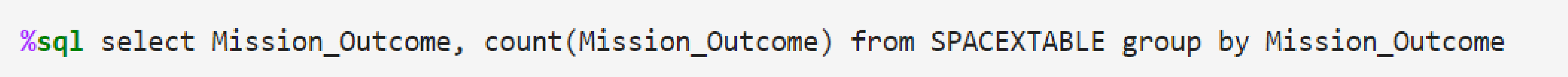
Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

There are four different booster versions which met this condition.



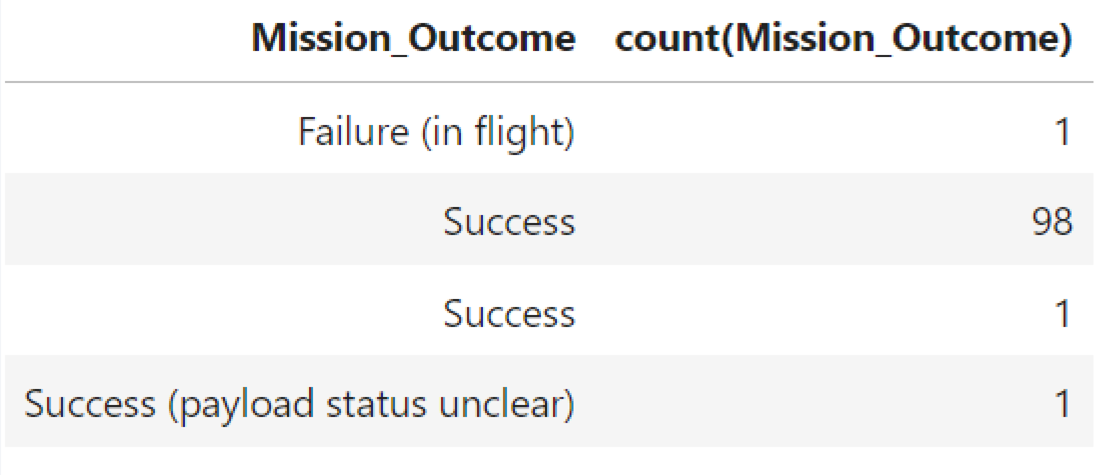
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



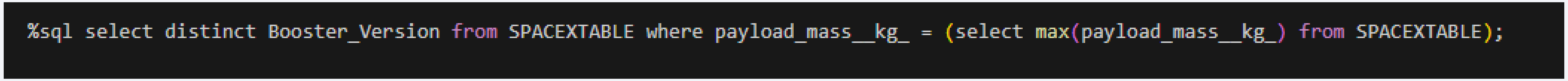
```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

There are 100 successful and 1 failure mission outcomes in the dataset.



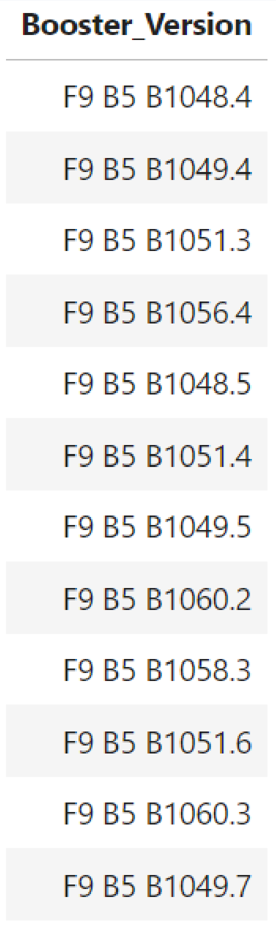
Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload



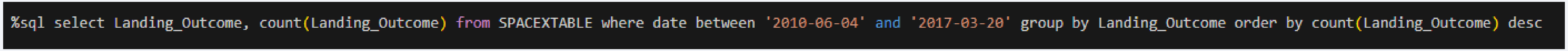
```
%sql select distinct Booster_Version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

12 different boosters have carried the maximum payload mass (names see table).



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

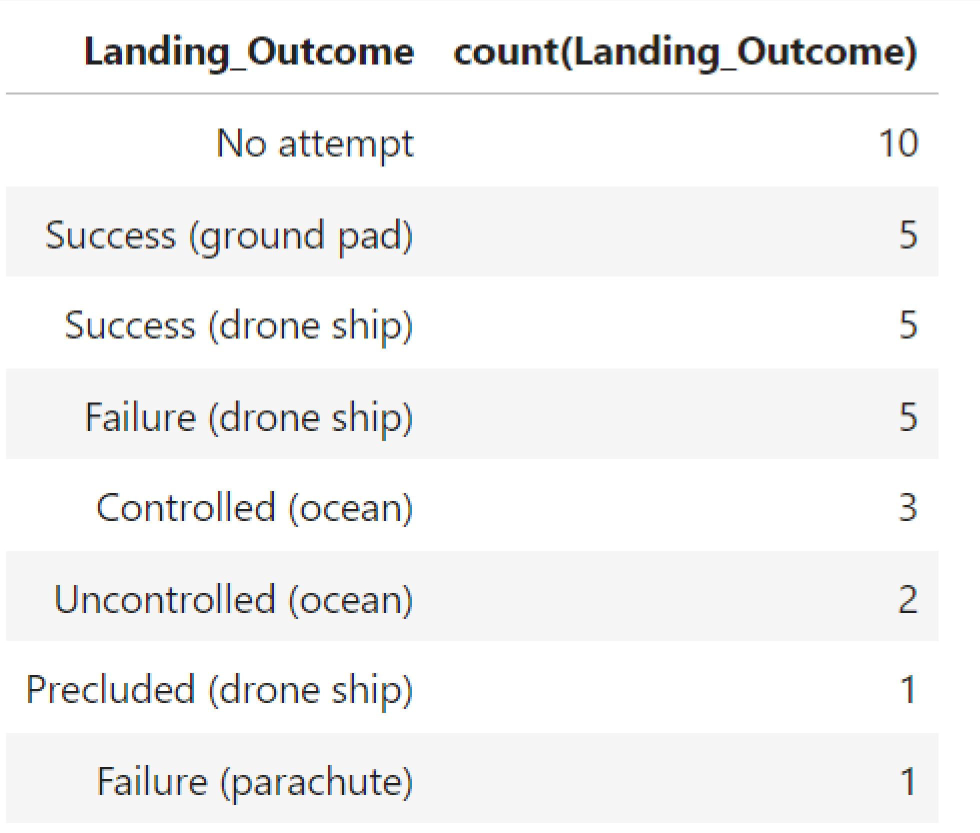
2015 Launch Records



```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc
```

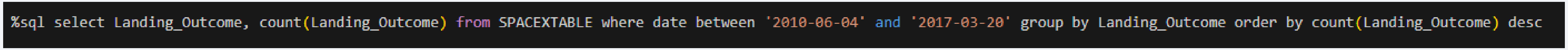
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here



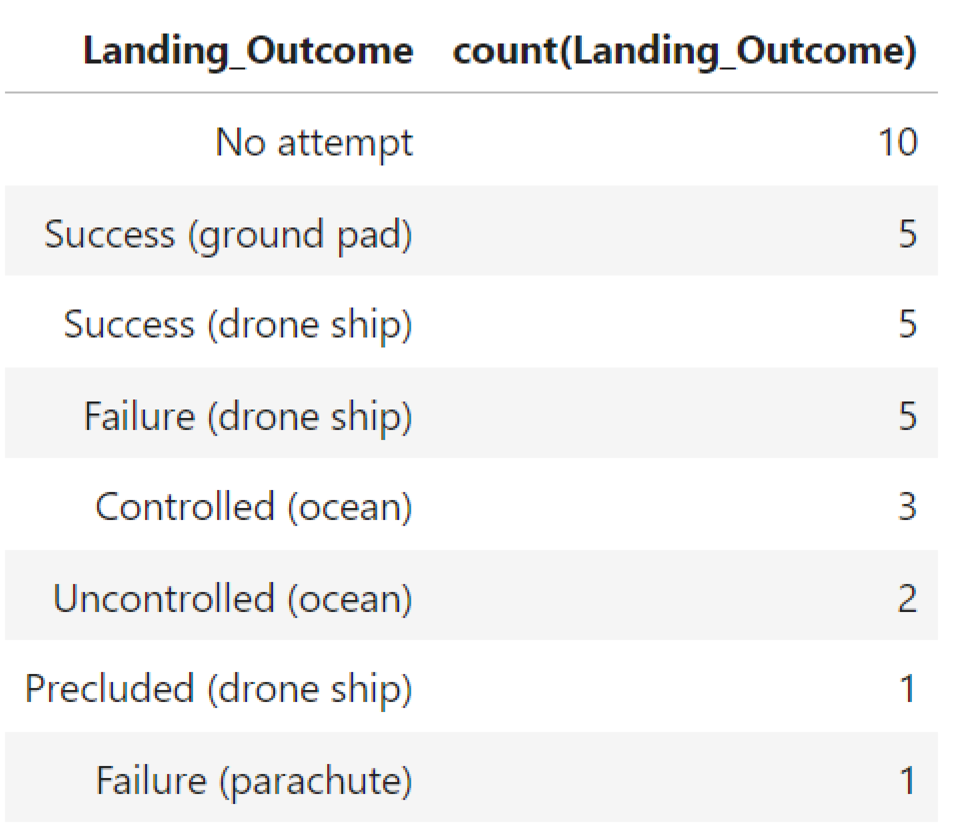
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

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



```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc
```

In landing outcomes between the date 2010-06-04 and 2017-03-20 we found 5 successes on ground pad, 5 successes on drone ship and 5 failures on drone ships. There was only 1 failure with parachute. 10 flights had no attempt to landing.



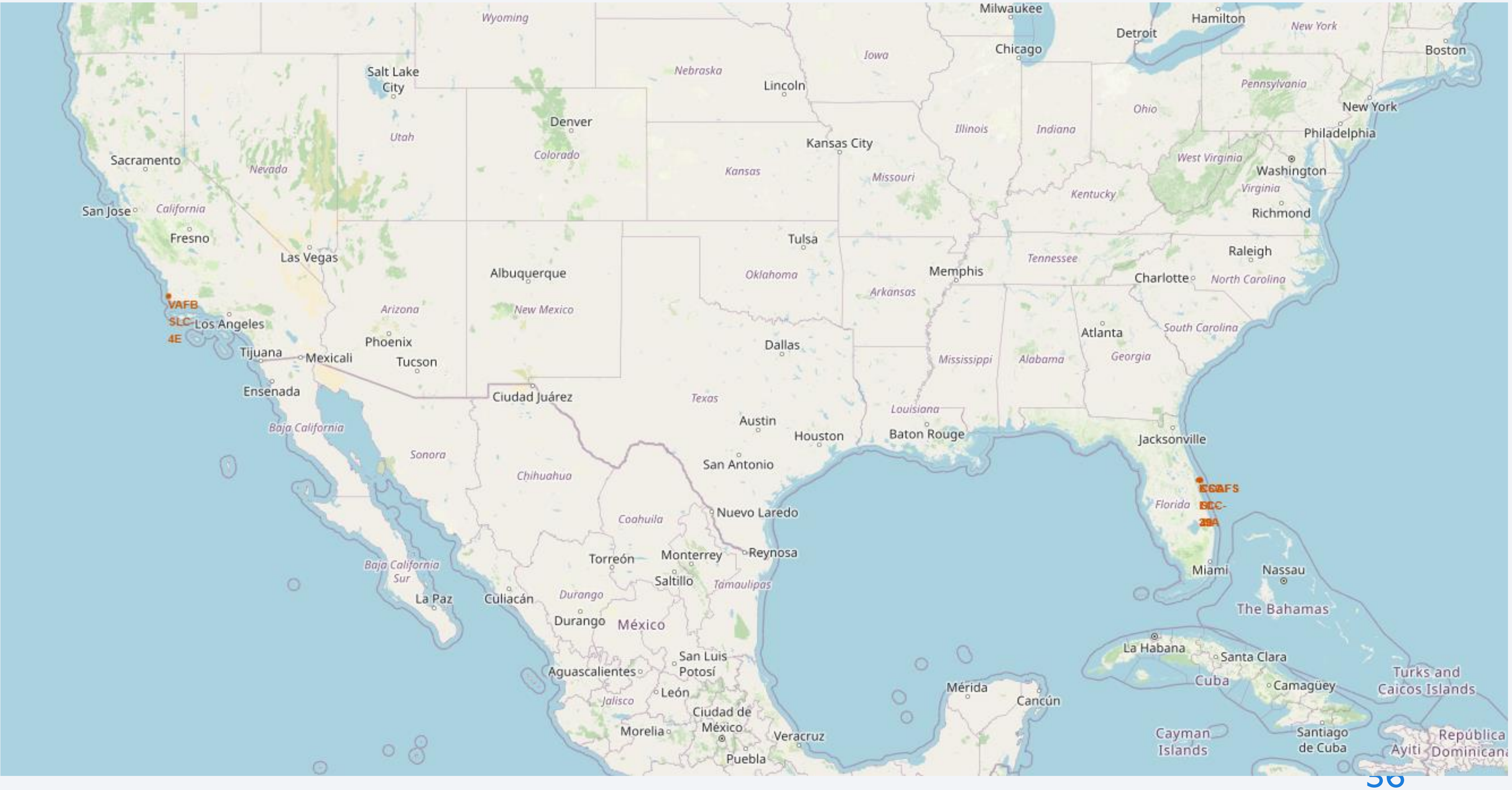
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Location of all launch sites

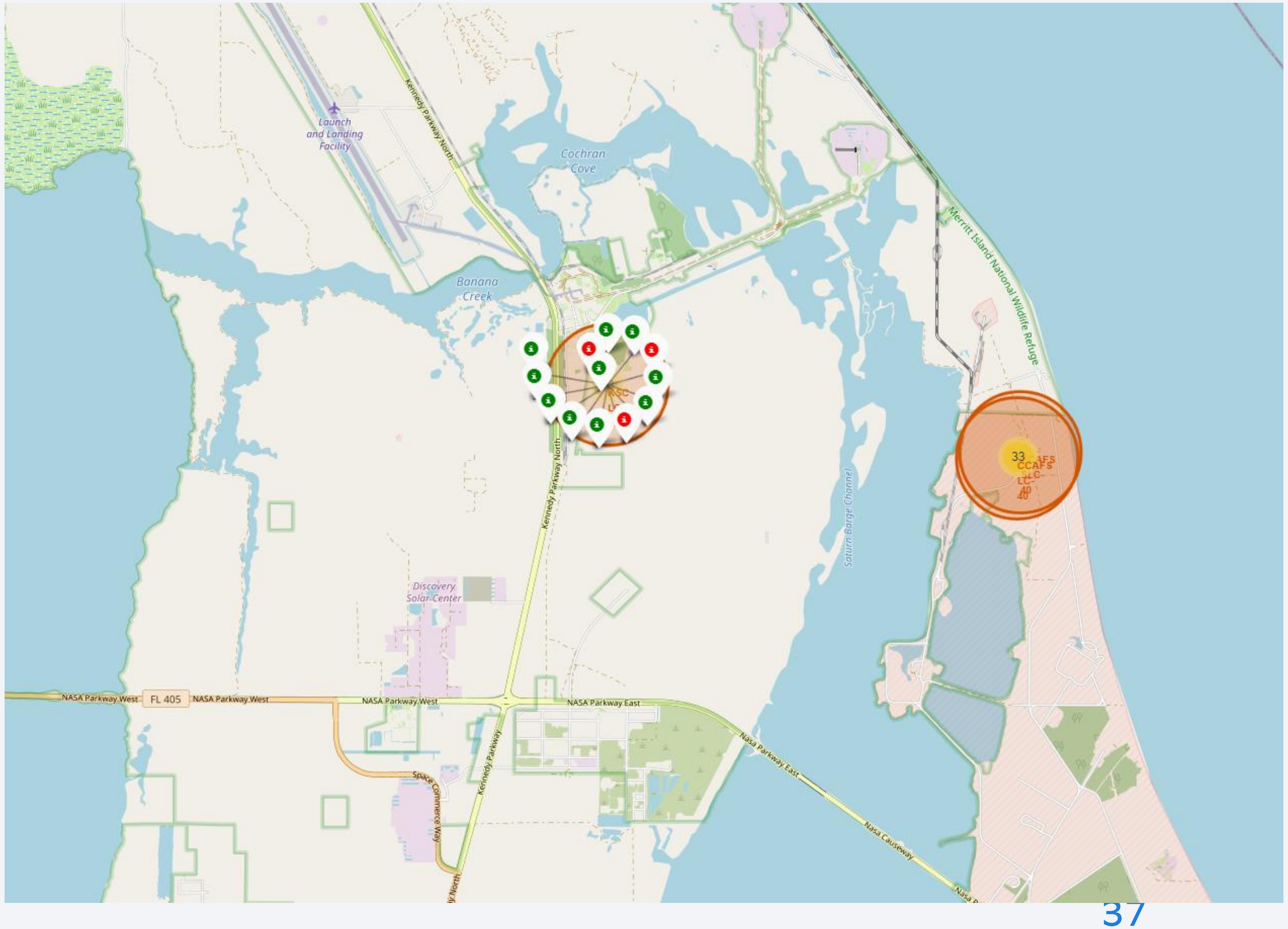
The map show all launch sites in the U.S. (red markers, 2 in Florida, 1 in California).



Launch outcomes for launch site

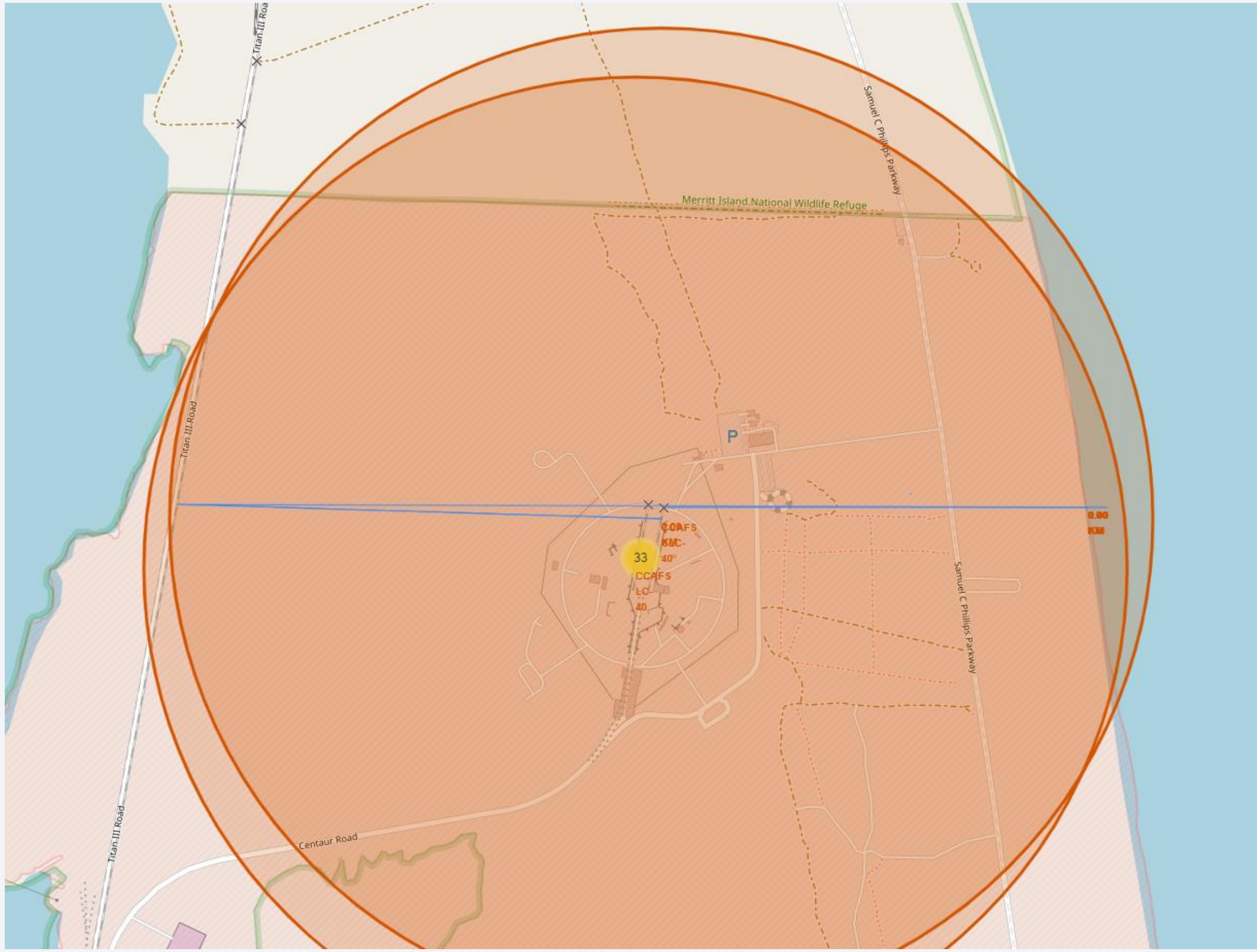
The map shows the outcome for each launch on each launch site (green = success).

In the screenshot we can see all the outcomes for launch site KSC LC 39A. Success rate for this launch site is 10/13 = 77 %.



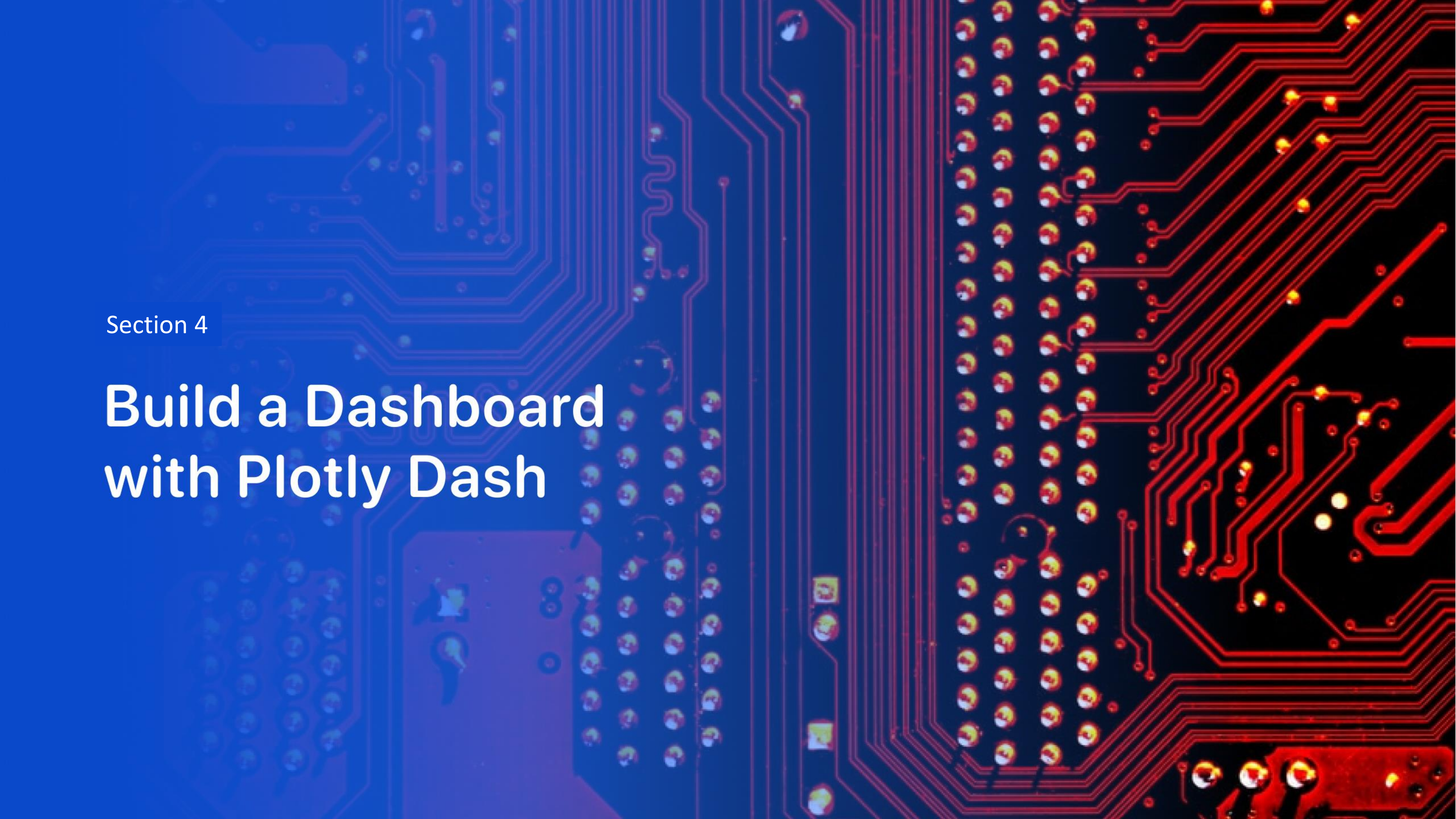
Distance from launch site

The generated folium map shows distance from a launch site to different points railway, highway or coastline. In the screenshot we can see distances from CCAFS to the nearest coastline and railway.

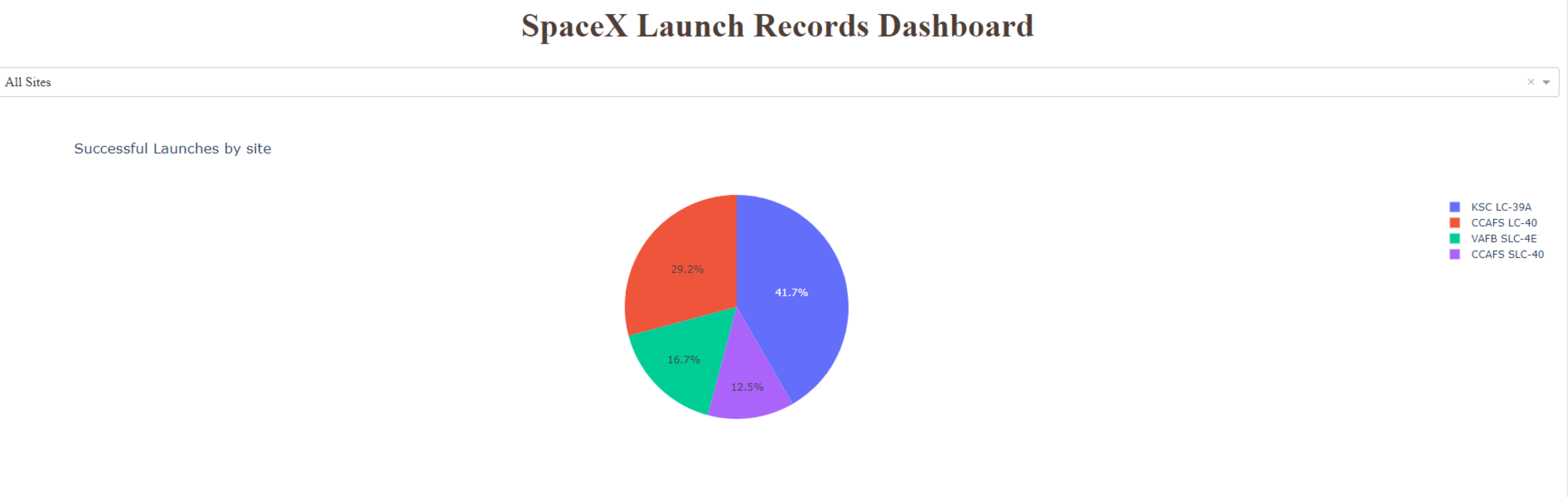

Titan III Road
Merritt Island National Wildlife Refuge
Samuel C Phillips Parkway
P
33
CCAFS
LC-
40
CCAFS
LC-
40
0.00
KM
Samuel C Phillips Parkway
Titan III Road
Centaur Road

Section 4

Build a Dashboard with Plotly Dash



Successful launches by site



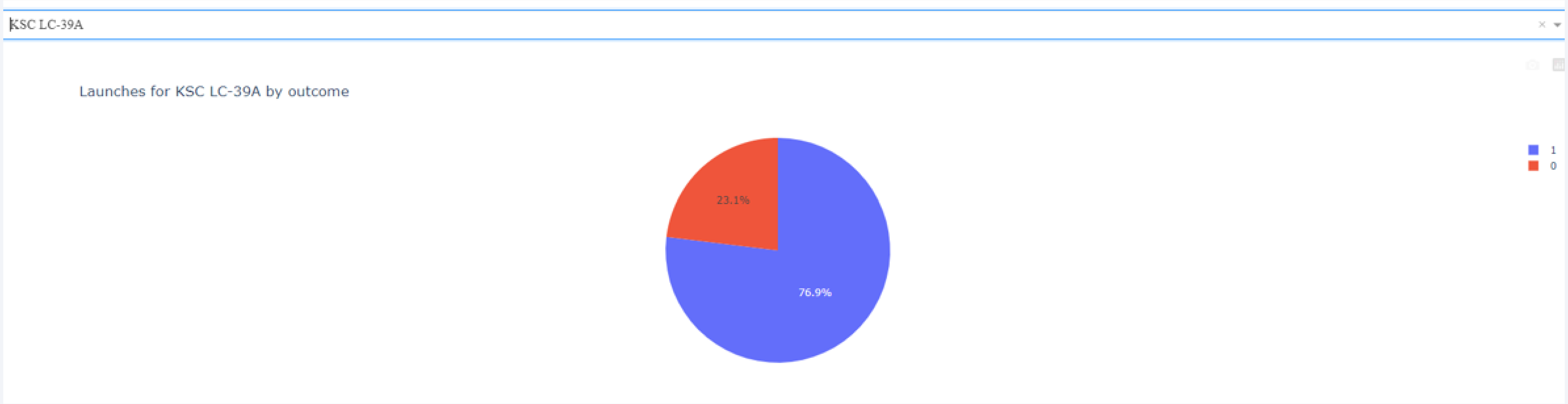
SpaceX Launch Records Dashboard

All Sites

Successful Launches by site

In the screenshot we can see a pie chart for successful launches by site. Most successful launches were recorded for KSC (41.7%) while CCAFS had the fewest successful launches (12.5%).

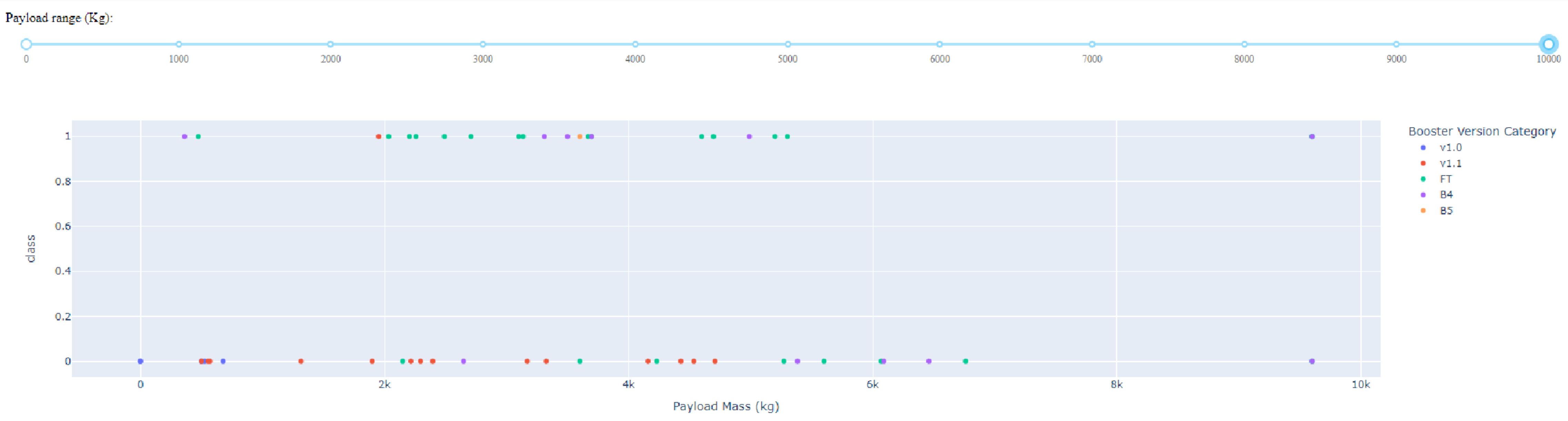
Highest launch success ratio



KSC LC-39A

Launch site KSC had the highest launch success ratio. 76.9% of all launches from this site had a successful outcome.

Payload vs. Launch Outcome (by Booster Version)



We can see a better success rate for lower Payload Mass (< 4000 kg) than for midrange Payload Mass (between 4000 and 8000 kg).

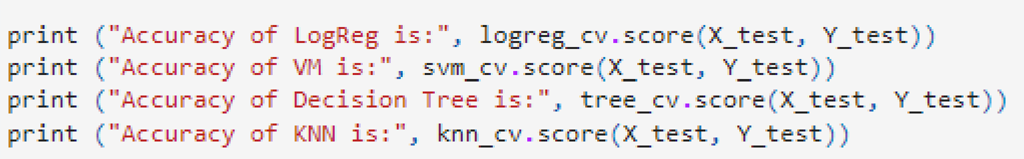
Booster Version FT has a very good success rate. Booster Version v1.1 has a very low success rate.

Section 5

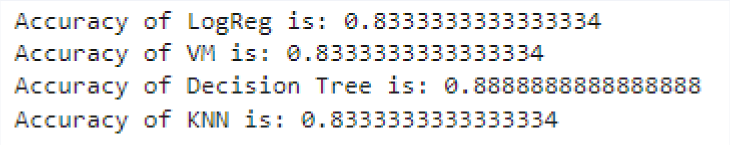
Predictive Analysis (Classification)

Classification Accuracy

From our models, decision tree had the highest classification accuracy on the test data (0.89 compared to 0.83 in all other models).



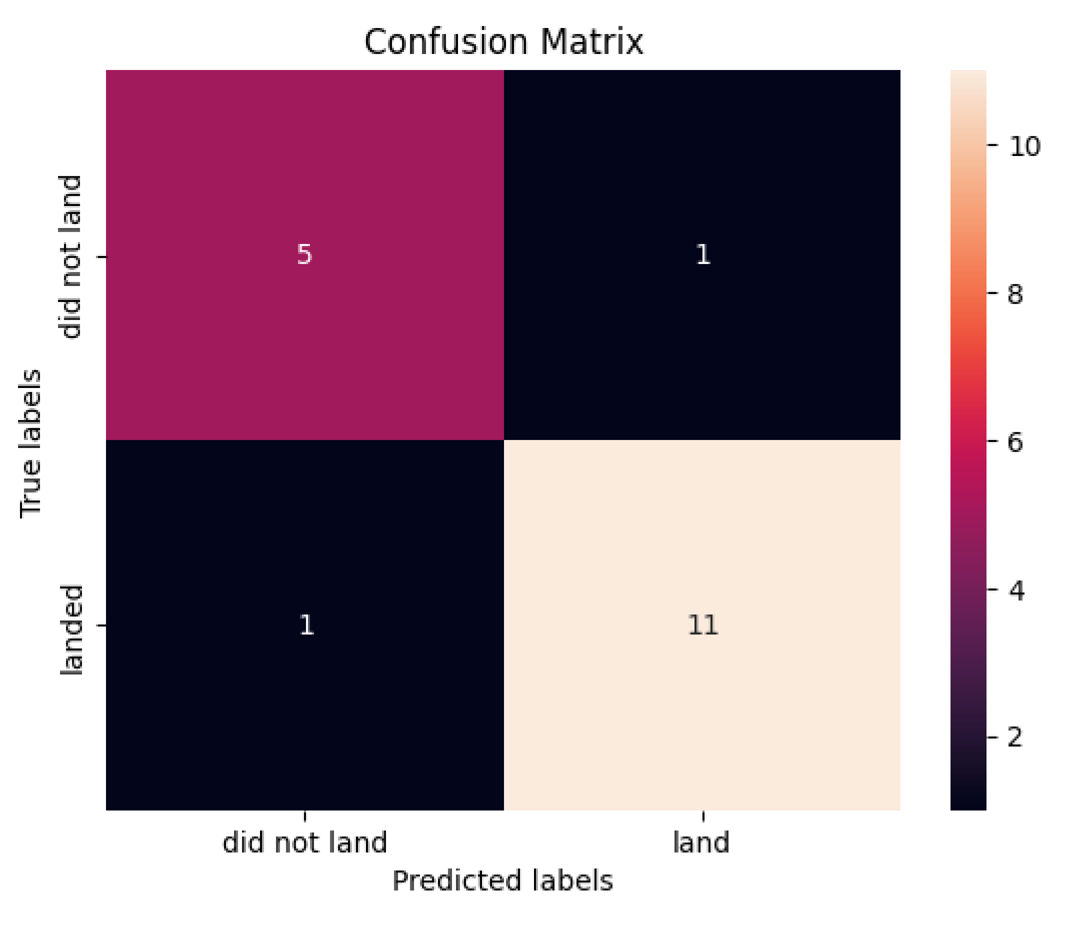
```
print ("Accuracy of LogReg is:", logreg_cv.score(X_test, Y_test))
print ("Accuracy of VM is:", svm_cv.score(X_test, Y_test))
print ("Accuracy of Decision Tree is:", tree_cv.score(X_test, Y_test))
print ("Accuracy of KNN is:", knn_cv.score(X_test, Y_test))
```



```
Accuracy of LogReg is: 0.8333333333333334
Accuracy of VM is: 0.8333333333333334
Accuracy of Decision Tree is: 0.8888888888888888
Accuracy of KNN is: 0.8333333333333334
```

Confusion Matrix for decision tree

- We had 18 launches in our test data.
- From all 18 launches, 16 were classified correctly with our decision tree model.
- One launch was classified false positive and one launch was classified false negative.



Conclusions

- We found that a few features might be correlated with a positive outcome of rocket launches.
- More recent launches had a higher success rate, possibly connected with new booster versions like FT.
- Payload mass might play a role for mission outcome but is dependent from target orbit. For orbit VLEO we found a very good success rate for missions with high payload mass but had no data on VLEO launches with low payload mass. Newer booster versions might also be able to carry a higher payload mass.
- Machine learning algorithms might help to predict a future outcome. We found high accuracy (80%<) for all tested models.
- Decision tree had the best outcome (accuracy: 0.89%) in predicting future outcomes.

Thank you!