IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michal Fornalski
2023-02-16

Outline

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

Executive Summary

Summary of methodologies

- Data collection using SpaceX API and Wikipedia Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL and Data Visualization
- Interactive visual analytics with Folium maps
- Plotly Dash Dashboard
- Machine learning predictive analysis

Summary of all results

- EDA Results
- Interactive analysis
- Predictive analysis

Introduction

Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems you want to find answers

- What makes the successful launch a success.
- Can we predict if the Falcon 9 first stage will land successfully based on available data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The data was collected from 2 sources

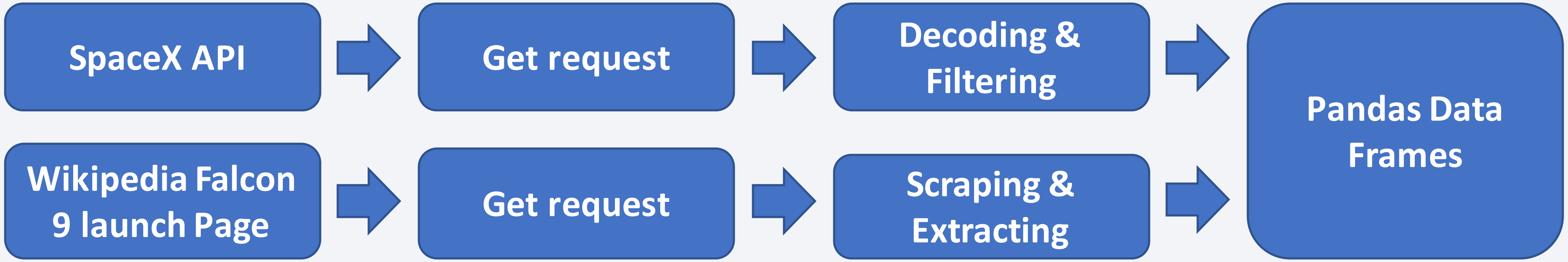
  - SpaceX API (https://api.spacexdata.com/v4)
  - Web scraping (Wikipedia page)

- Perform data wrangling
  - Basic data cleaning, One Hot encoding for landing outcome as preparation for machine learning.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, SVM, DT & KNN methods were used to determine if the first stage of Falcon 9 will land successfully.
  - Best parameters, accuracy & confusion matrixes were established for each of the models.

Data Collection

The data was collected from 2 sources:

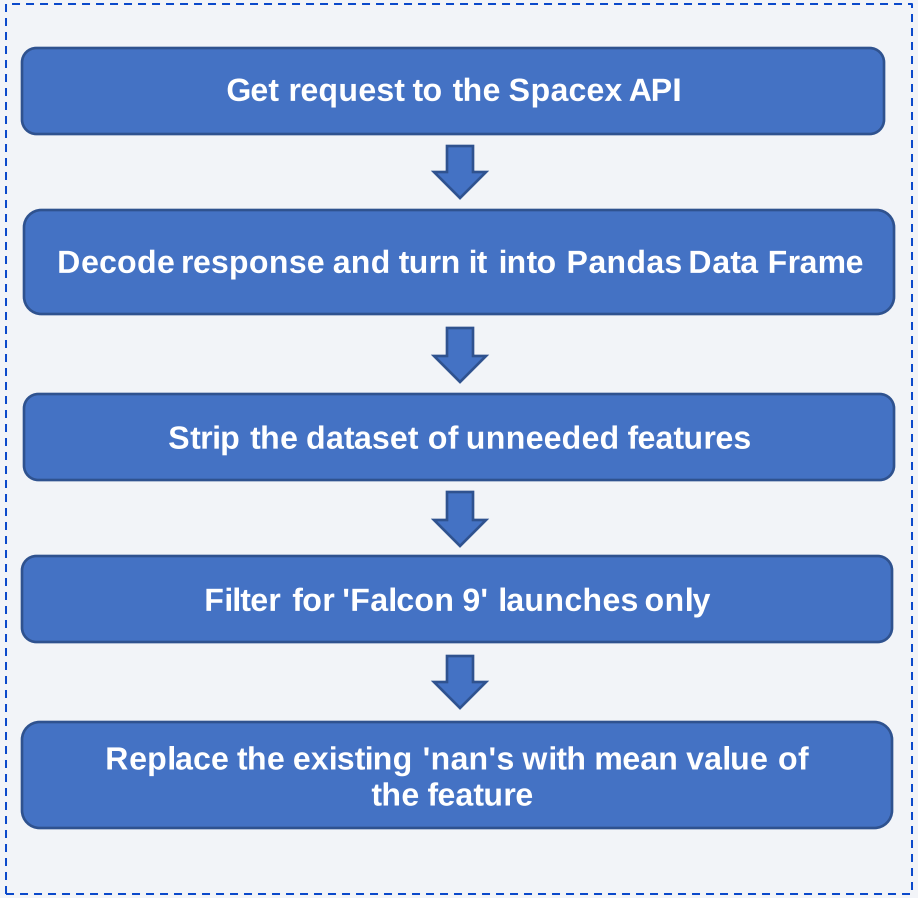
- SpaceX API (https://api.spacexdata.com/v4)
- Web scraping (Wikipedia page)



Data Collection – SpaceX API

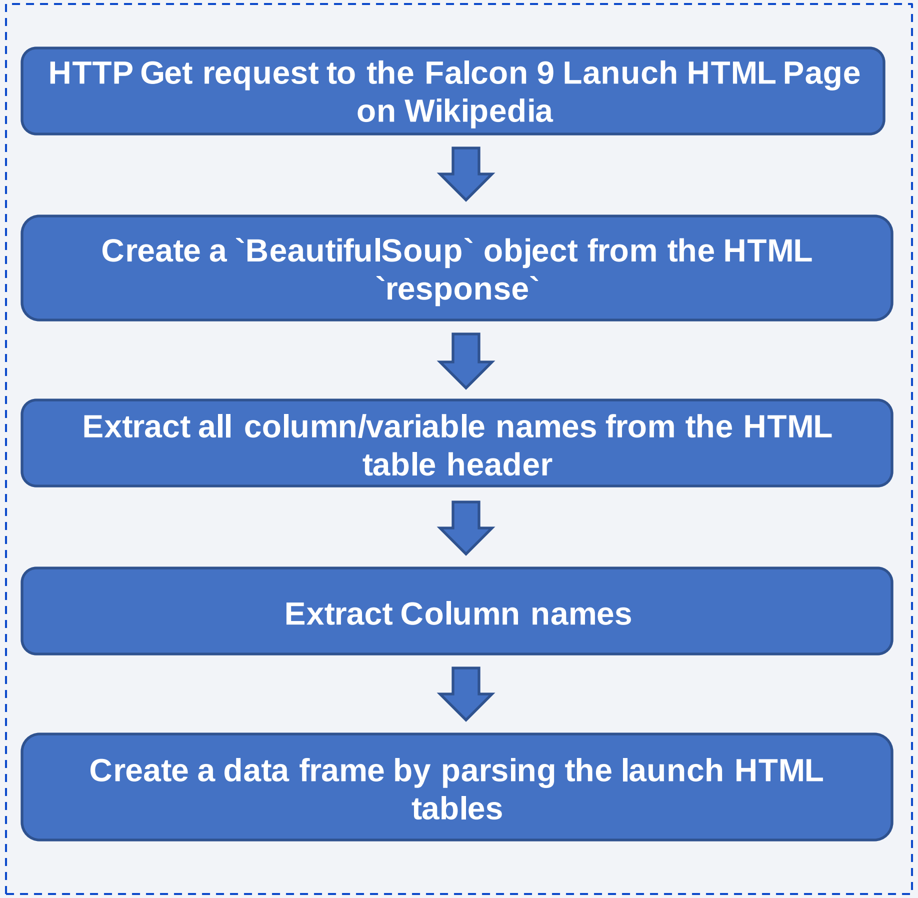
1. Request rocket launch data from SpaceX API
2. Decoding the response content as a Json using .json() and turning it into a Pandas dataframe using .json_normalize()
3. Filtering the data dataframe using the BoosterVersion column to only keep the Falcon 9 launches.
4. Dealing with Missing Values

For more detailed information see the **Jupyter notebook**


Get request to the Spacex API
Decode response and turn it into Pandas Data Frame
Strip the dataset of unneeded features
Filter for 'Falcon 9' launches only
Replace the existing 'nan's with mean value of the feature

Data Collection - Scraping

1. HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
2. Create a `BeautifulSoup` object from the HTML `response`
3. Extract all column/variable names from the HTML table header
4. Extract Column names
5. Create a data frame by parsing the launch HTML tables



HTTP Get request to the Falcon 9 Lanuch HTML Page on Wikipedia

Create a `BeautifulSoup` object from the HTML `response`

Extract all column/variable names from the HTML table header

Extract Column names

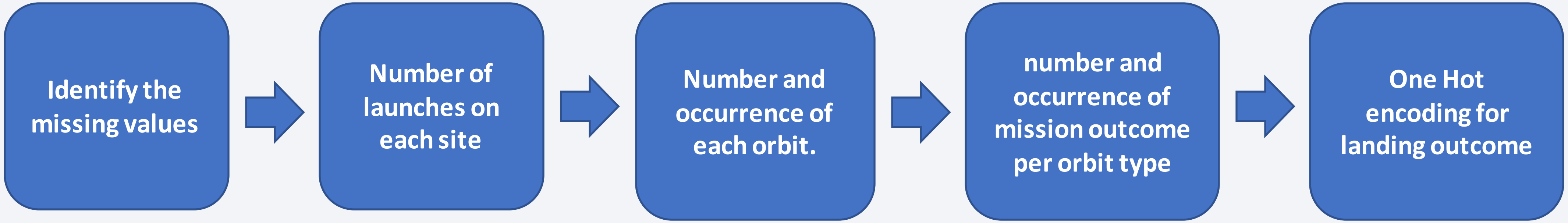
Create a data frame by parsing the launch HTML tables

For more detailed information see the **Jupyter notebook**

Data Wrangling

Data processing steps

- Identification and calculation of the percentage of the missing values in each attribute to determine the quality of the data.
- Calculation of the number of launches on each site.
- Calculation of the number and occurrence of each orbit.
- Calculation of the number and occurrence of mission outcome per orbit type.
- One Hot encoding for landing outcome as preparation for machine learning.



Identify the missing values → Number of launches on each site → Number and occurrence of each orbit. → number and occurrence of mission outcome per orbit type → One Hot encoding for landing outcome

For more detailed information see the **Jupyter notebook**

EDA with Data Visualization

Charts used

- Scatter point chart visualizing the relationship between Flight Number and Payload
- Scatter point chart visualizing the relationship between Flight Number and Launch Site
- Scatter point chart visualizing the relationship between Payload and Launch Site
- Bar chart visualizing the Success rate of each orbit type.
- Scatter point chart visualizing the relationship between Orbit and Flight Number
- Scatter point chart visualizing the relationship between Payload and Orbit Type
- Line chart visualizing launch success yearly trend

For more detailed information see the **Jupyter notebook**

EDA with SQL

SQL queries performed

- %sql select Unique(LAUNCH_SITE) From SPACEX
- %sql select * From SPACEX Where LAUNCH_SITE Like 'CCA%' limit 5
- %sql select Sum(payload_mass__kg_) From SPACEX Where Customer = 'NASA (CRS)'
- %sql select AVG(payload_mass__kg_) From SPACEX Where booster_version like 'F9 v1.1%'
- %sql select min(Date) From SPACEX Where landing__outcome = 'Success (ground pad)'
- %sql select booster_version From SPACEX Where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
- %sql select mission_outcome, count(mission_outcome) AS Count From SPACEX Group By mission_outcome
- %sql Select booster_version From SPACEX Where payload_mass__kg_ = (Select max(payload_mass__kg_) From SPACEX)
- %sql Select booster_version, launch_site From SPACEX Where landing__outcome = 'Failure (drone ship)' and year(DATE) = '2015'
- %sql select landing__outcome, count(landing__outcome) as COUNT from SPACEX Where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome) DESC

For more detailed information see the **Jupyter notebook**

Build an Interactive Map with Folium

Different objects such as: markers, circles and lines were added to the maps with explicit purpose:

- Markers type 1 - added to indicate different Launch Sites
- Markers type 2 – added to indicate all launches locations
- Circles - added to indicate different Launch Sites
- Lines – added to indicate distance from Launch site to: coast line, closest highway, closest railway & closest city.

For more detailed information see the **Jupyter notebook**

Build a Dashboard with Plotly Dash

Using the skeleton app provided by the course as a base there were few modification that were made:

- Adding a Launch Site Drop-down Input Component
- Adding a callback function to render `success-pie-chart` based on selected site dropdown
- Adding a Range Slider to Select Payload
- Adding a callback function to render the `success-payload-scatter-chart` scatter plot

Charts:

- Pie chart of the total successful launches by site
- Scatter plot showing correlation between Payload and success for all sites

For more detailed information see the **Jupyter notebook**

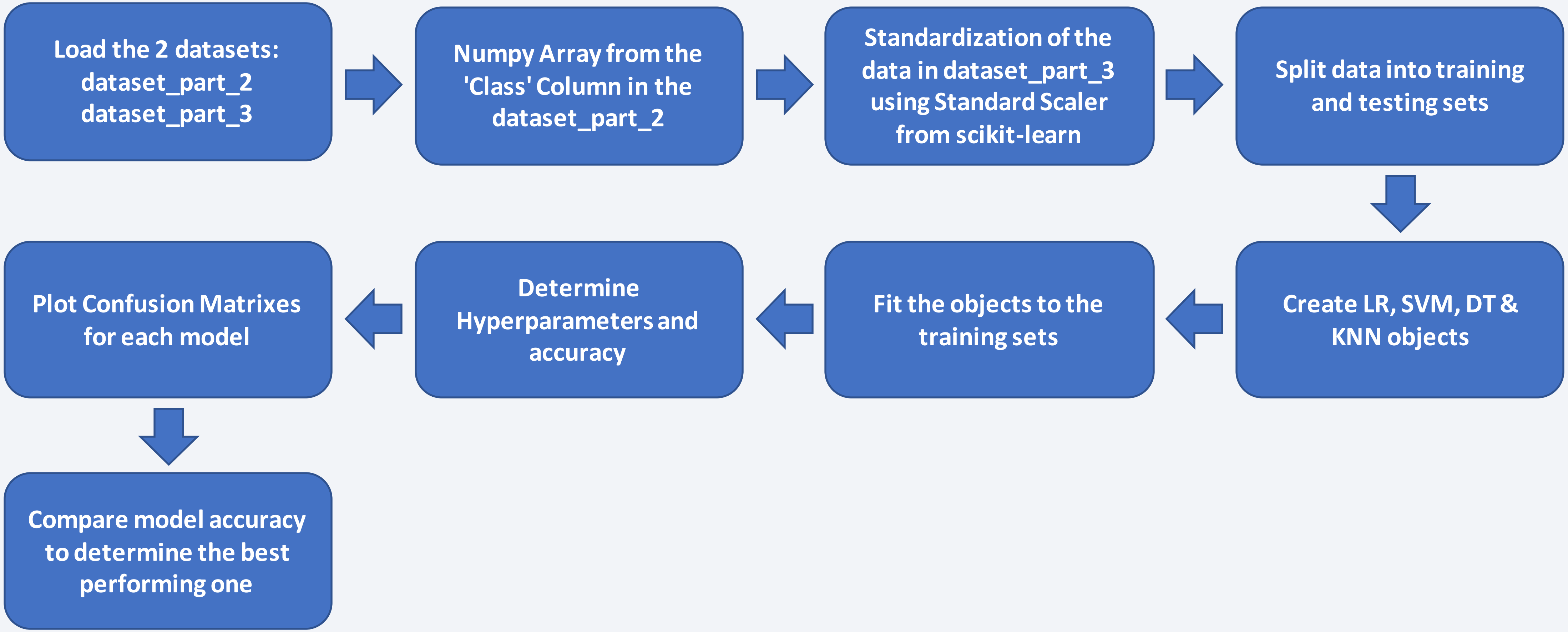
Predictive Analysis (Classification)

Summary of the predictive analysis steps

- Creation of a Numpy Array from the 'Class' Column in the dataset
- Data standardization using Standard Scaler from scikit-learn
- Split data into training and testing sets (test_size=0.2, random_state=2)
- Create, Fit & Find the hyperparameters for each of the models (LR, SVM, DT & KNN)
- Plot Confusion Matrixes for each model
- Compare model accuracy

For more detailed information see the **Jupyter notebook**

Predictive Analysis (Classification)



Load the 2 datasets:
dataset_part_2
dataset_part_3

Numpy Array from the 'Class' Column in the dataset_part_2

Standardization of the data in dataset_part_3 using Standard Scaler from scikit-learn

Split data into training and testing sets

Create LR, SVM, DT & KNN objects

Fit the objects to the training sets

Determine Hyperparameters and accuracy

Plot Confusion Matrixes for each model

Compare model accuracy to determine the best performing one

For more detailed information see the **Jupyter notebook**

Results

Exploratory data analysis results:

- **4** different launch sites were used for all the launches in the dataset
- Total payload mass carried by boosters launched by NASA (CRS) = **45 596 kg**
- Average payload mass carried by booster version F9 v1.1 = **2 928 kg**
- First successful landing outcome on the ground pad was achieved on **2015-12-22**
- Boosters which have successfully landed on a 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
- Total number of successful and failure mission outcomes:
  Failure (in flight) - 1, Success – 99, Success (payload status unclear) - 1
- Names of the booster_versions which have carried the maximum payload mass:
  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

Results

Exploratory data analysis results cont.:

- Failed landing outcomes on the drone ship, their booster versions, and launch site names for year 2015:
  F9 v1.1 B1012 - CCAFS LC-40, F9 v1.1 B1015-CCAFS LC-40

- Landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

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

Results

Interactive analytics demo in screenshots

- Displayed and explained in the Section 2 of the presentation

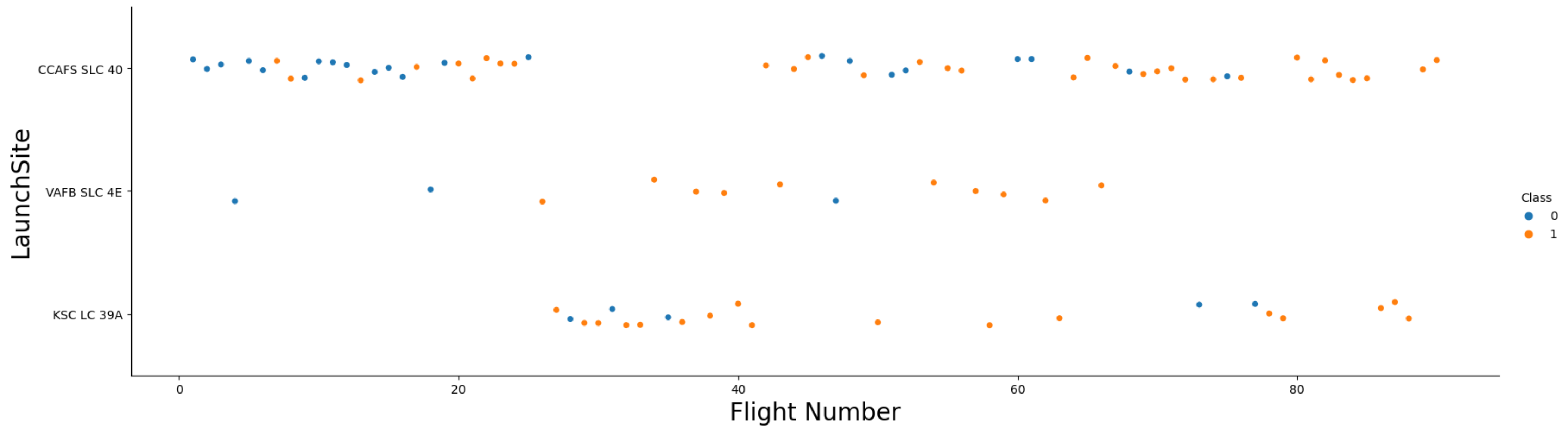
Predictive analysis results

- Displayed and explained in the Section 5 of the presentation

Section 2

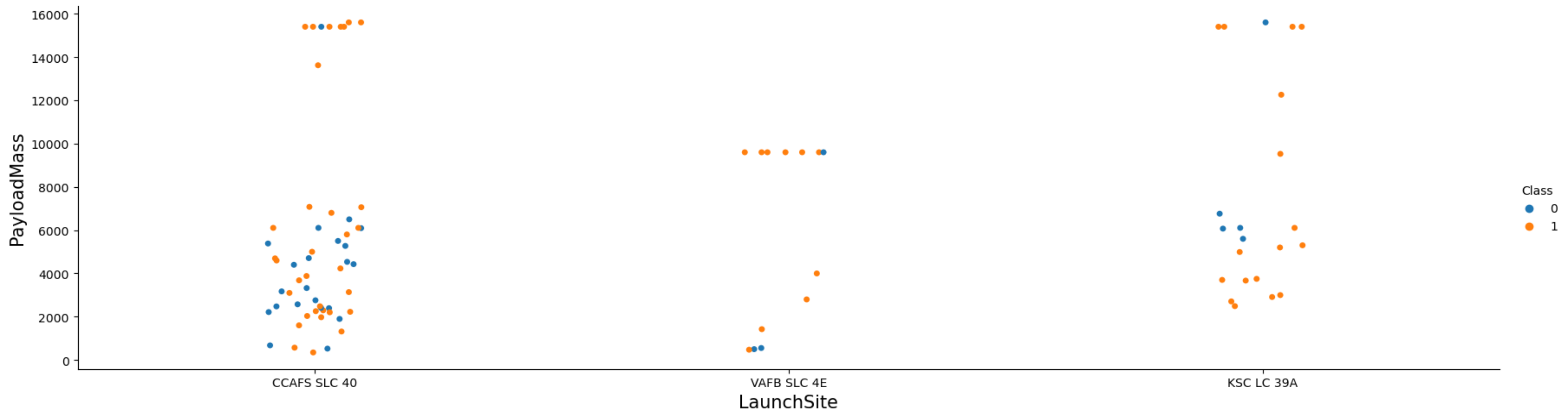
Insights drawn from EDA

Flight Number vs. Launch Site



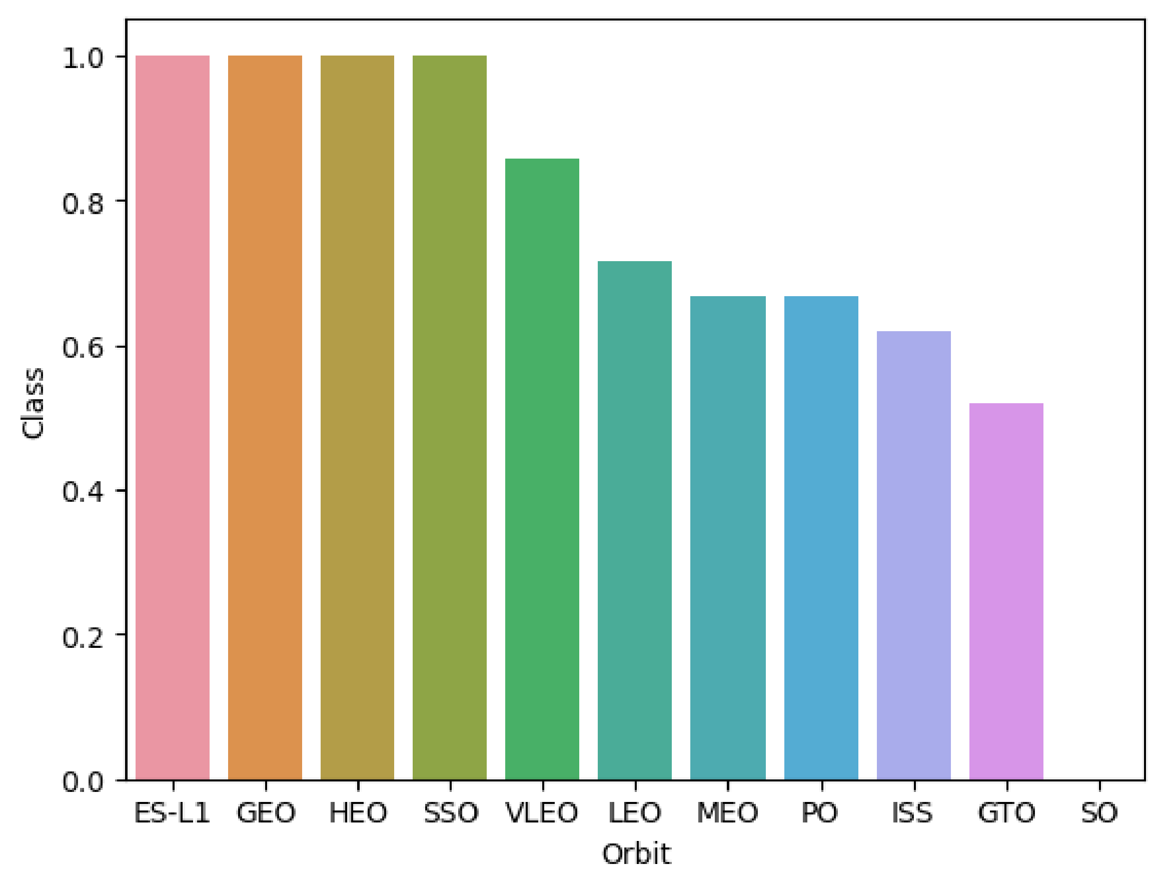
- General success rate improved over time
- Most Launches took place at CCAFS SLC 40

Payload vs. Launch Site



- For the VAFB-SLC launch site there are no rockets launched for payload mass greater than 10000

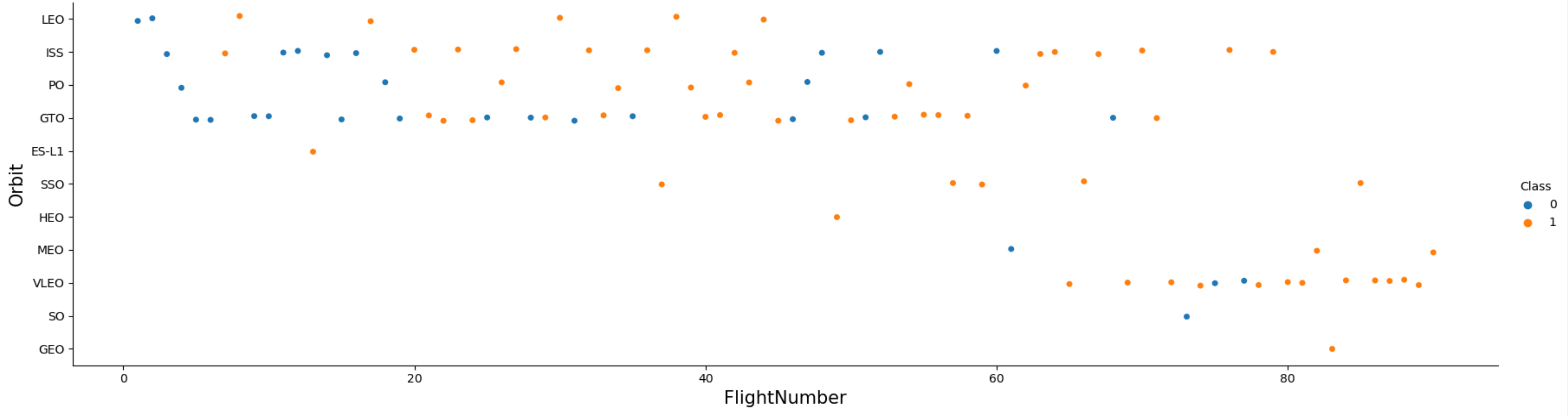
Success Rate vs. Orbit Type



- ES-L1, GEO, HEO & SSO are the most successful orbit types launches.

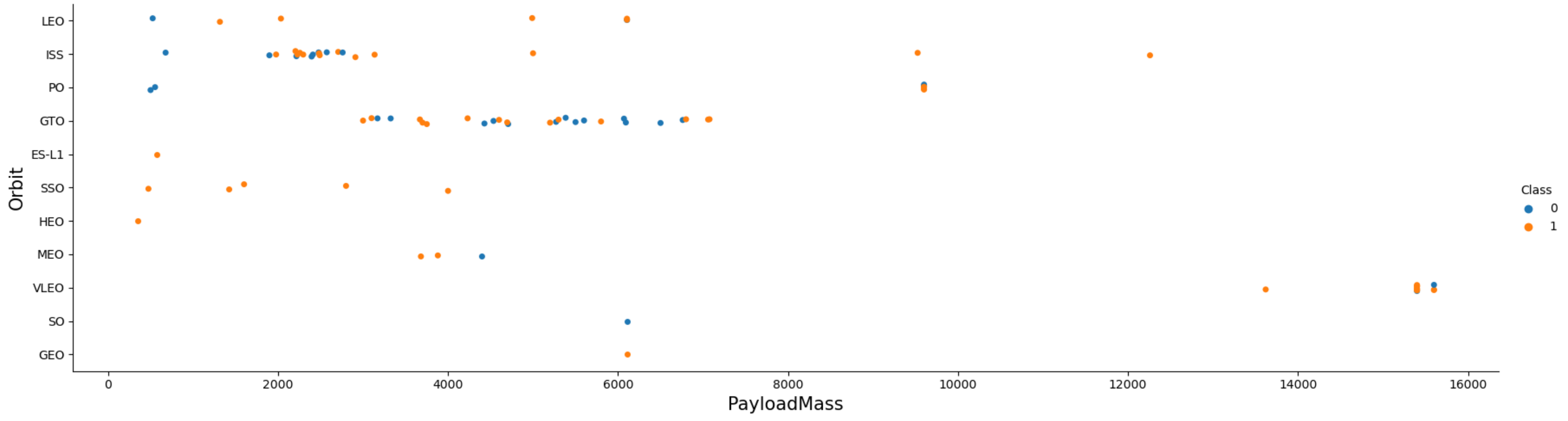
- SO, GTO & ISS have the lowest success rate of all orbit types launches.

Flight Number vs. Orbit Type



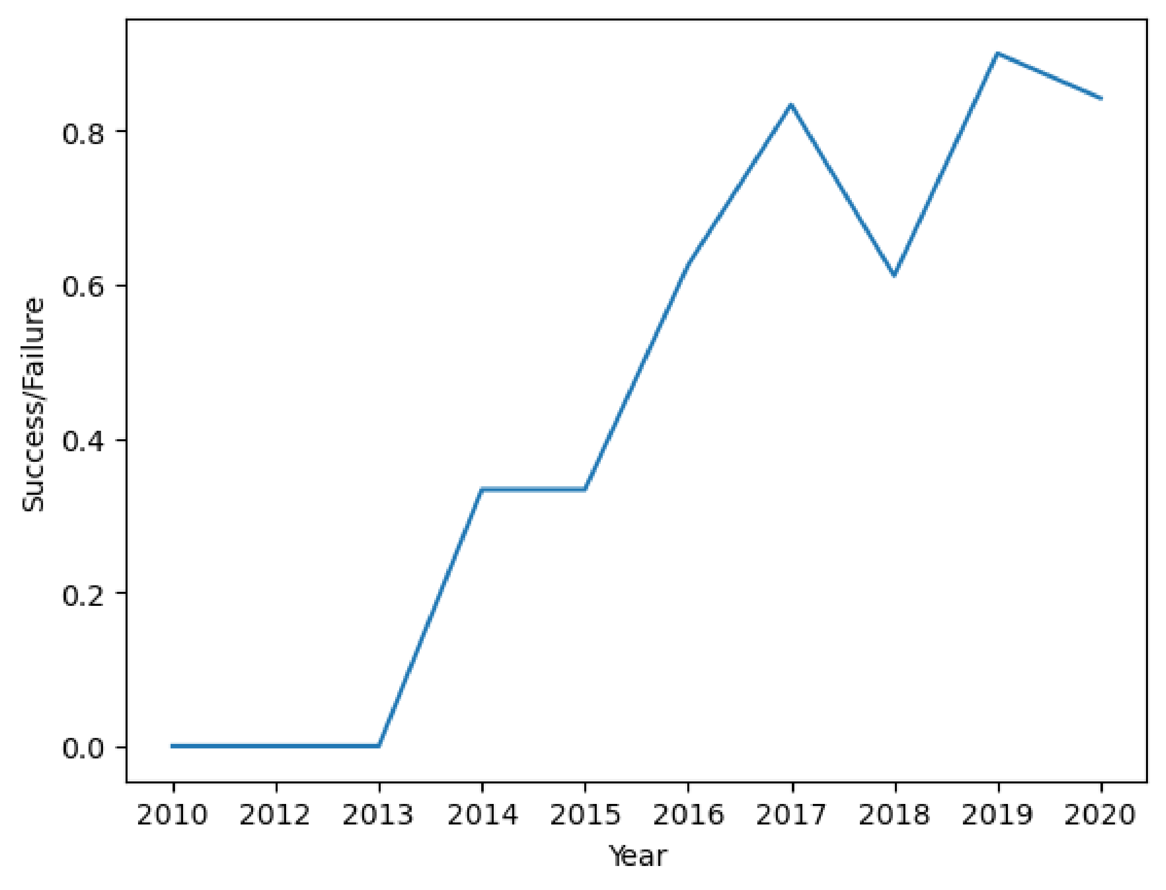
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Only a small number of launches exceed Payload Mass of 8000.
- There are only few launches for orbits SO and GEO.
- Orbit ISS has the widest range of payload mass.

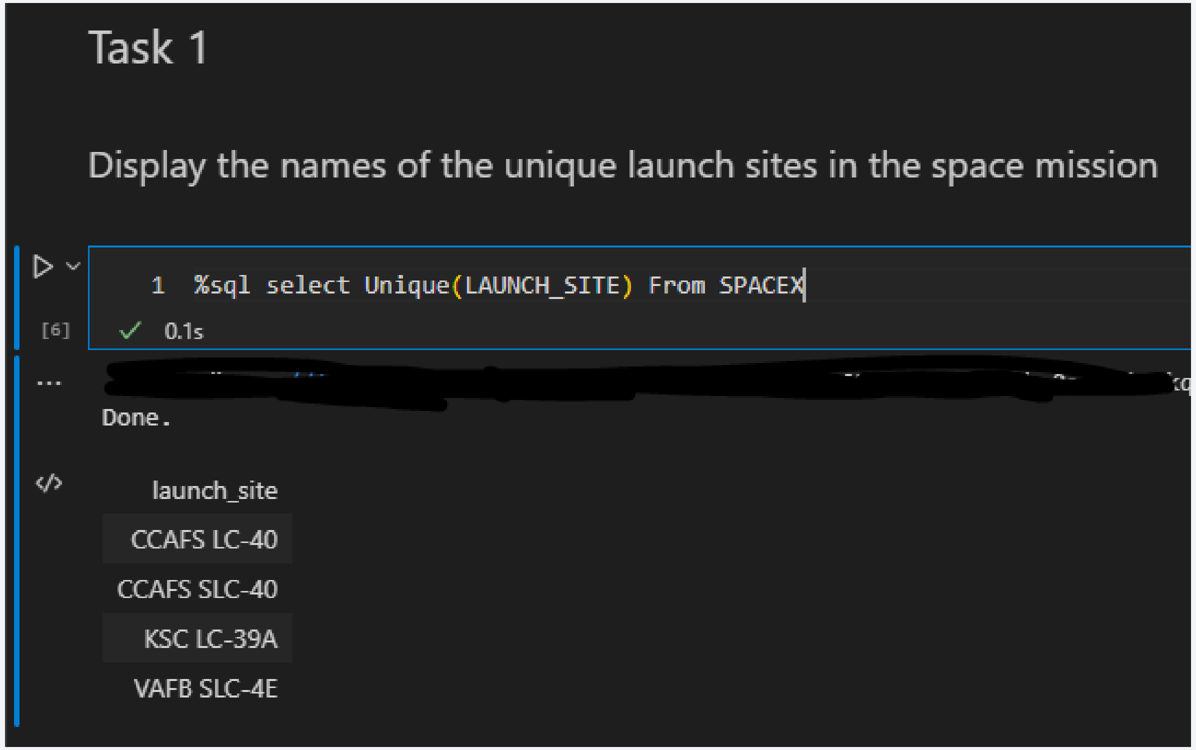
Launch Success Yearly Trend



- Sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- There are 4 unique launch sites.



Task 1

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

```
%sql select Unique(LAUNCH_SITE) From SPACEX
```

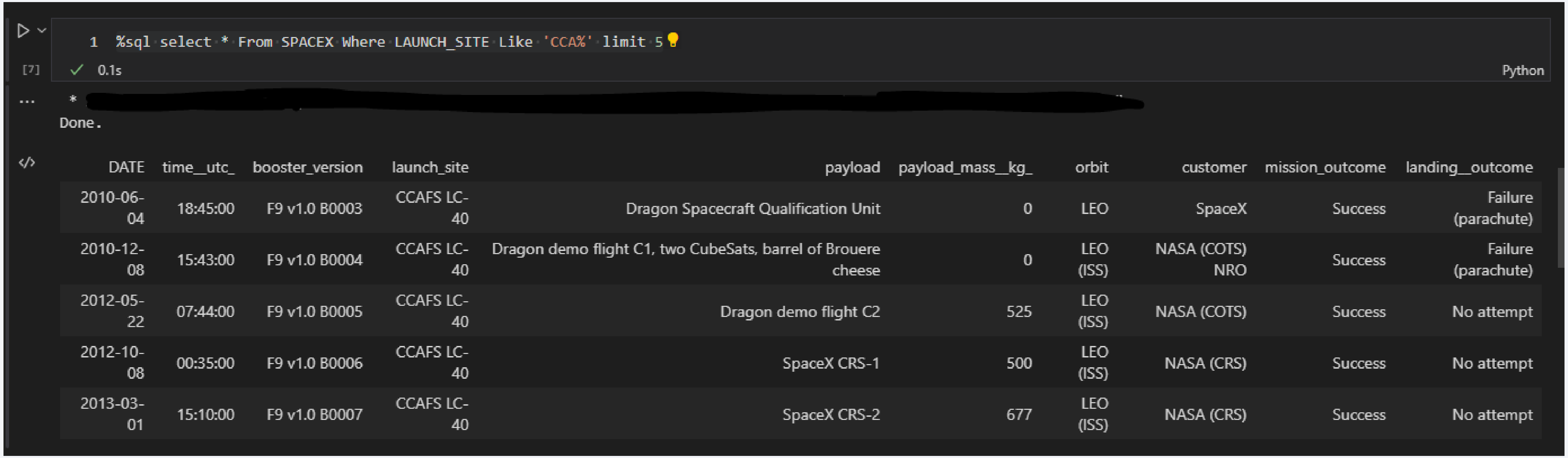
[redacted]

```
Done.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`



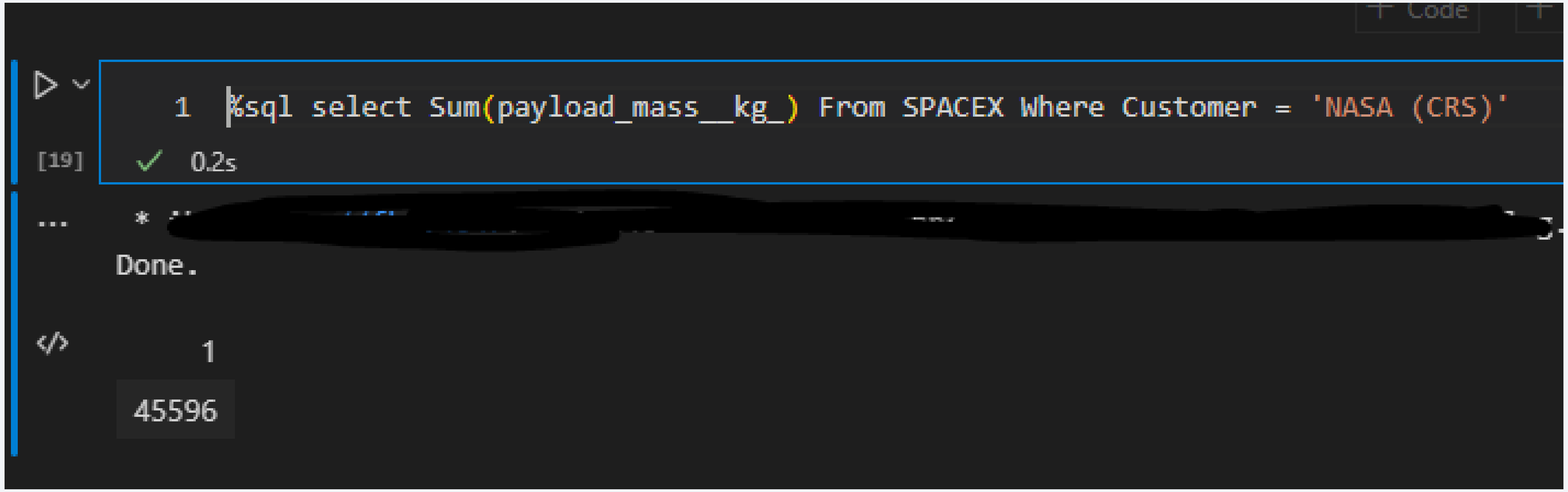
```
%sql select * From SPACEX Where LAUNCH_SITE Like 'CCA%' limit 5
```

```
* [redacted]
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA = 45 596 kg



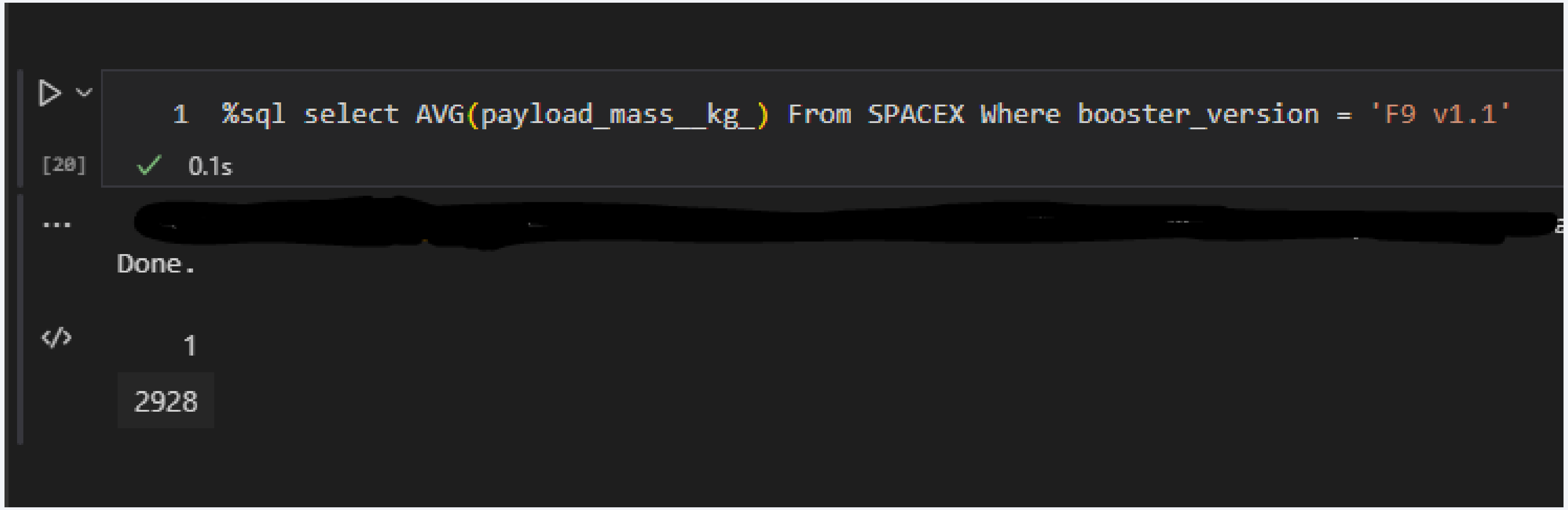
```
%sql select Sum(payload_mass__kg_) From SPACEX Where Customer = 'NASA (CRS)'
```

```
 * [redacted]
Done.

     1
 45596
```

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2 928 kg



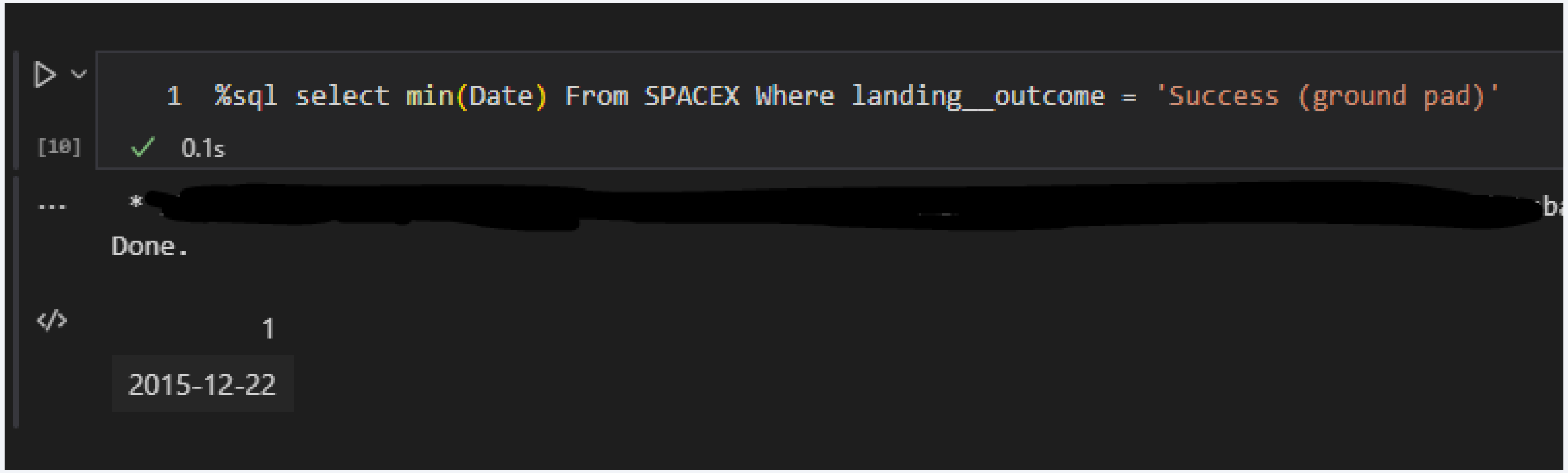
```
%sql select AVG(payload_mass__kg_) From SPACEX Where booster_version = 'F9 v1.1'
```

```
[redacted]
Done.

   1
2928
```

First Successful Ground Landing Date

- First successful landing outcome on the ground pad took place on 2015-12-22



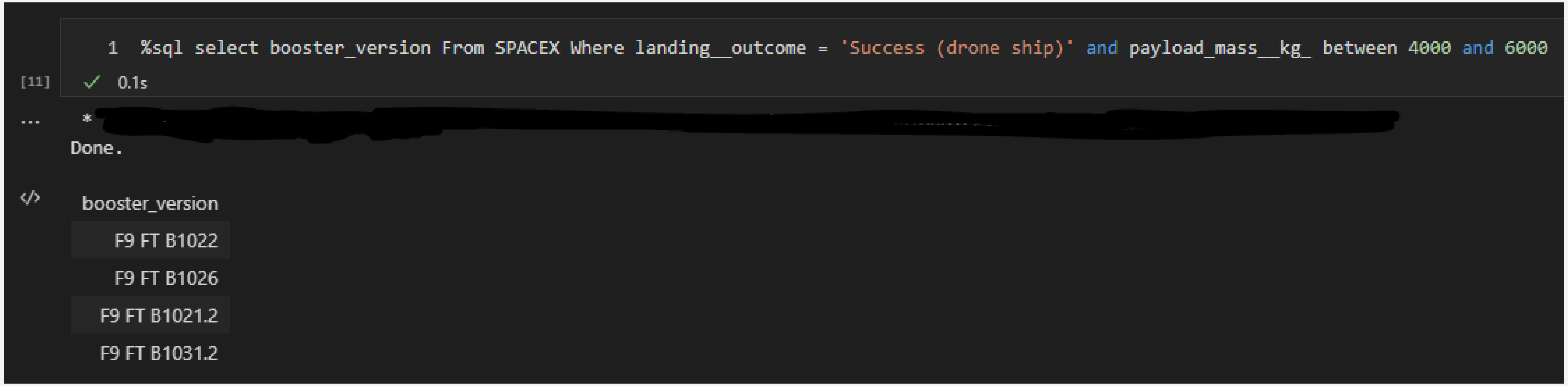
```
%sql select min(Date) From SPACEX Where landing__outcome = 'Success (ground pad)'
```

```
* [redacted]
Done.

         1
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



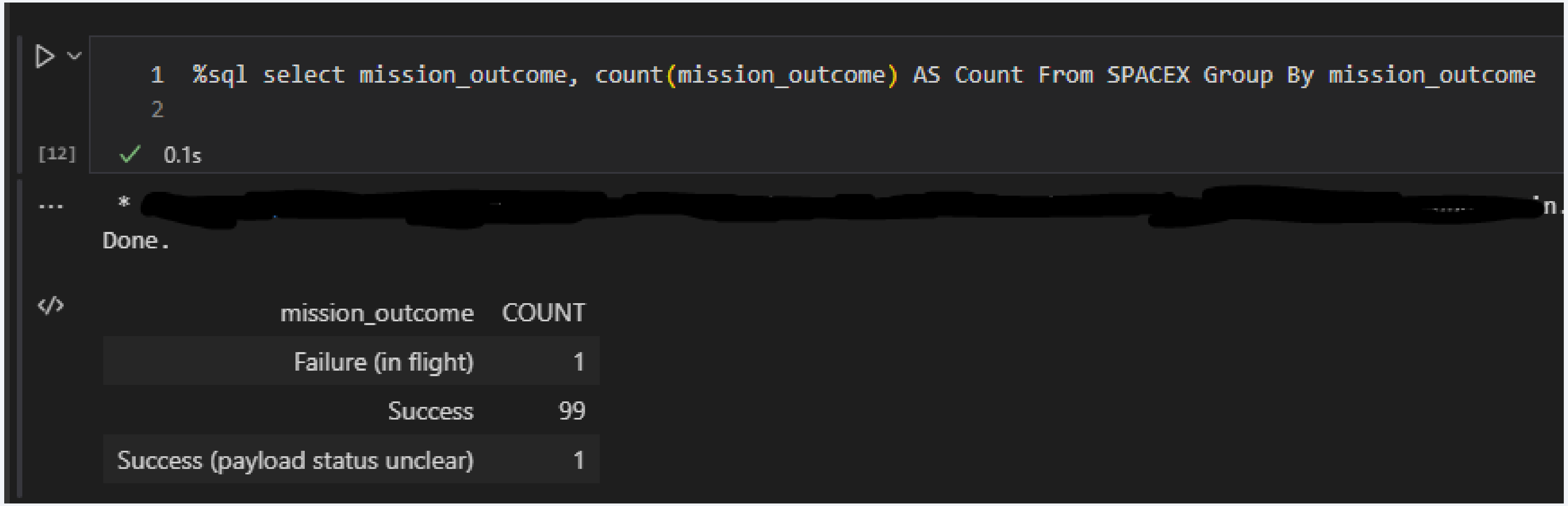
```
%sql select booster_version From SPACEX Where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
* [redacted]
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Success: 100
- Failure: 1

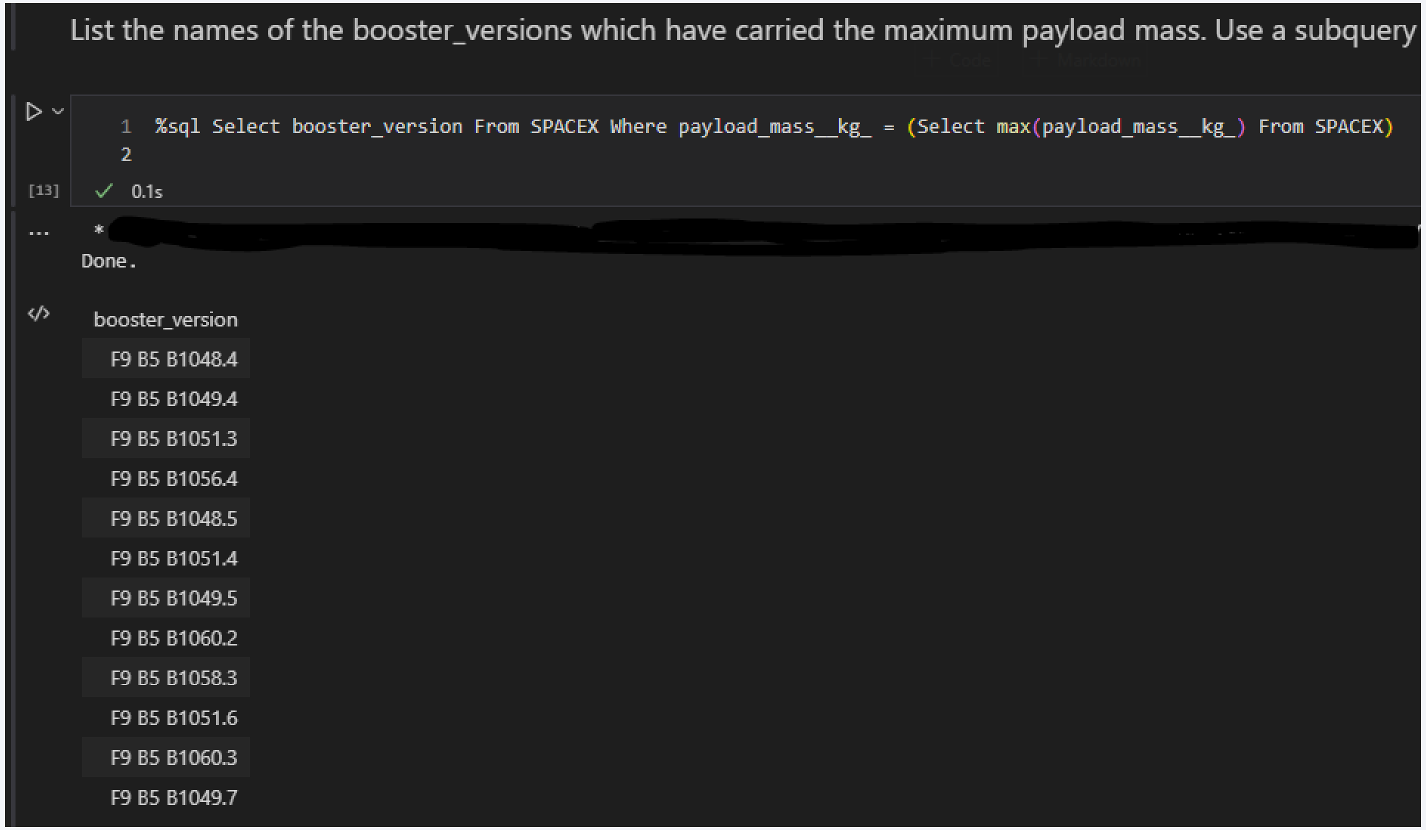


```
%sql select mission_outcome, count(mission_outcome) AS Count From SPACEX Group By mission_outcome
```

```
* [redacted]
Done.
```

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

Boosters Carried Maximum Payload



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

```
%sql Select booster_version From SPACEX Where payload_mass__kg_ = (Select max(payload_mass__kg_) From SPACEX)

```

[13] ✓ 0.1s

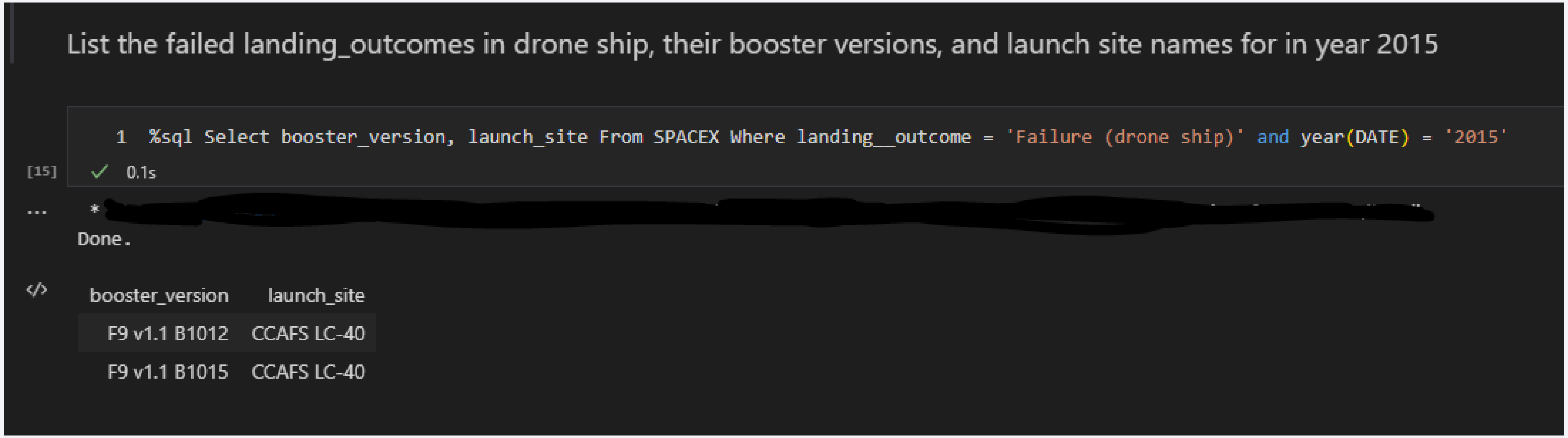
```
* [redacted]
Done.
```

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

- 12 versions of boosters have carried the maximum payload mass.

2015 Launch Records

- Failed landing_outcomes on a drone ship, their booster versions, and launch site names for a year 2015



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

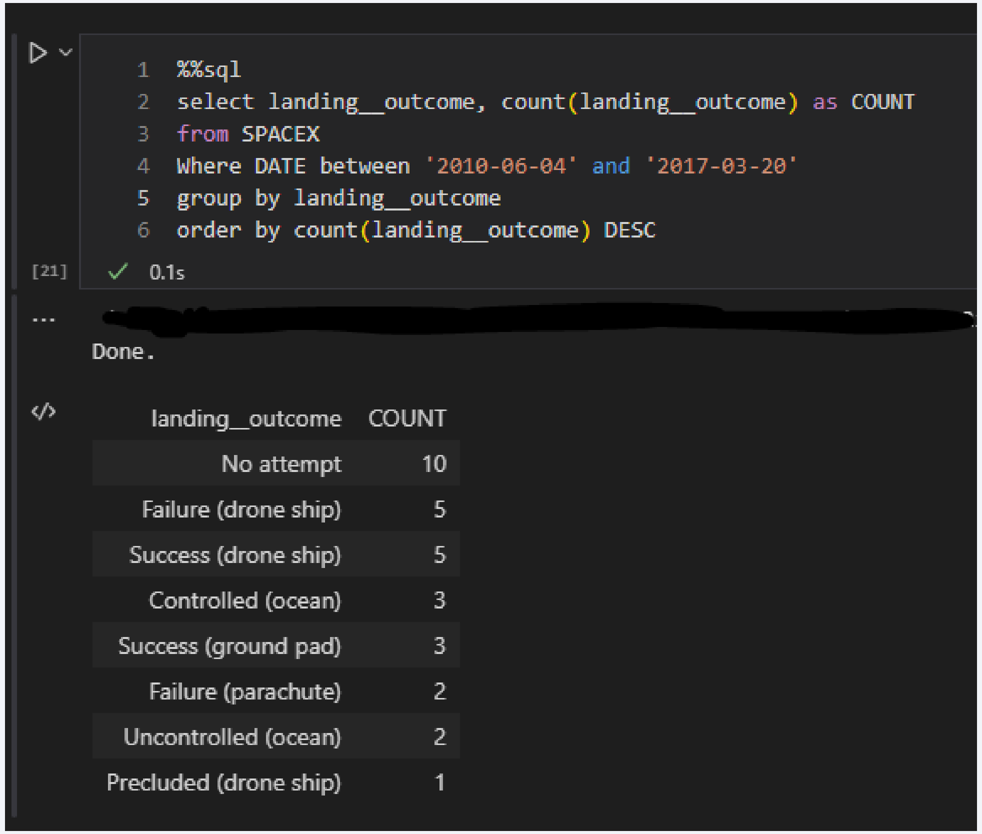
```
%sql Select booster_version, launch_site From SPACEX Where landing__outcome = 'Failure (drone ship)' and year(DATE) = '2015'
```

```
* [redacted]
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Number of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.



```
%%sql
select landing__outcome, count(landing__outcome) as COUNT
from SPACEX
Where DATE between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by count(landing__outcome) DESC
```

[21] ✓ 0.1s

[redacted]

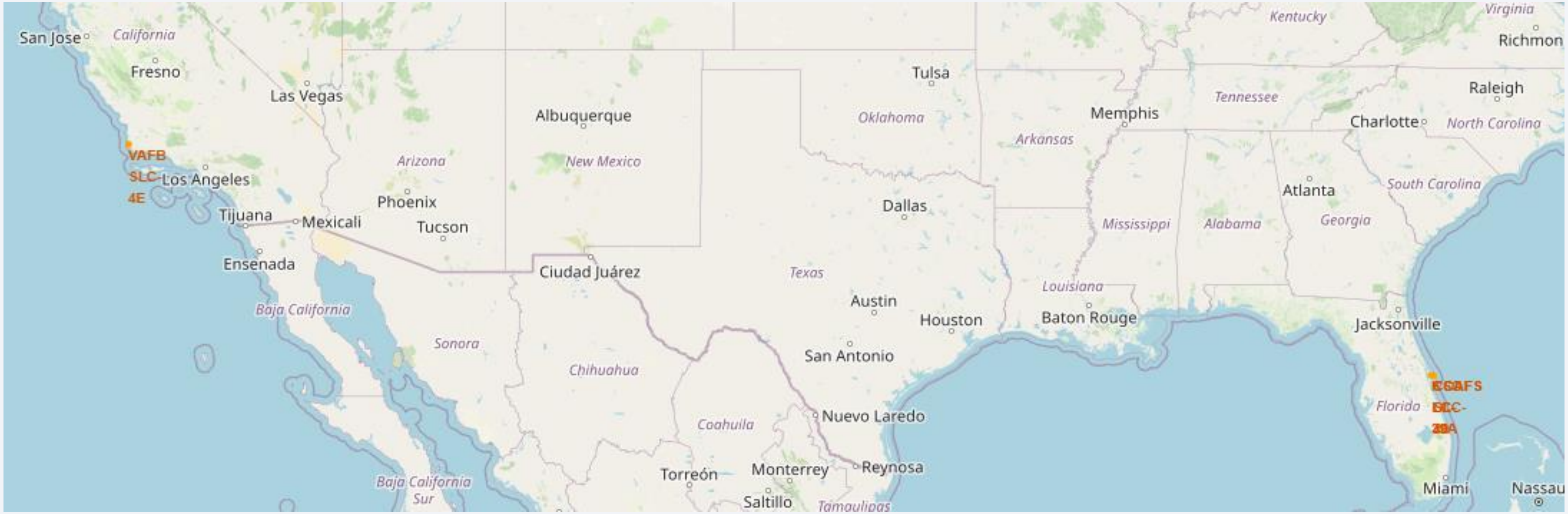
Done.

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

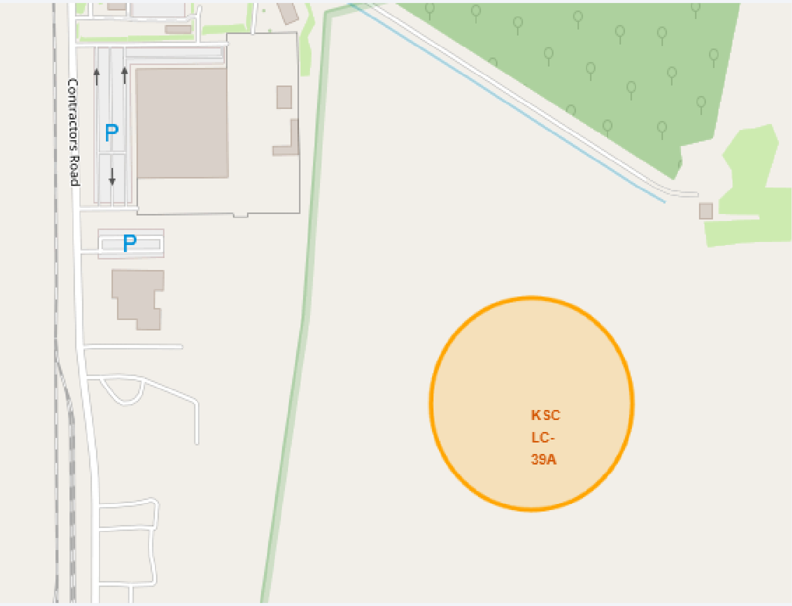
Section 3

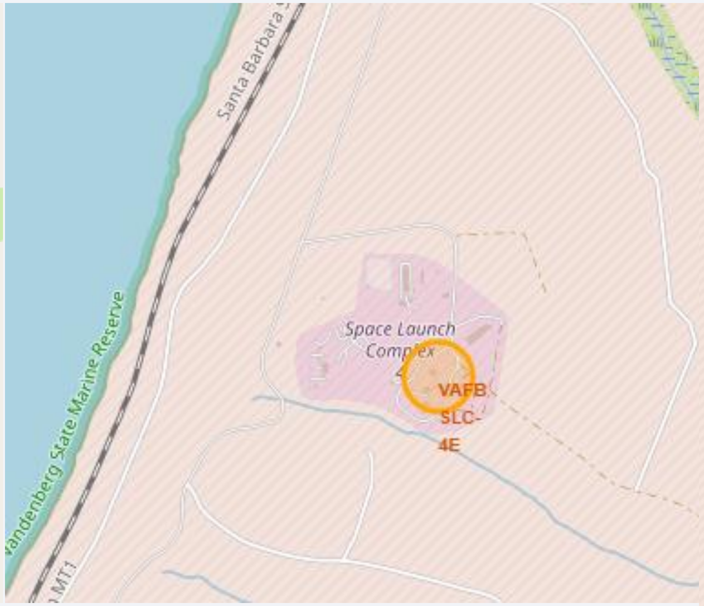
Launch Sites Proximities Analysis

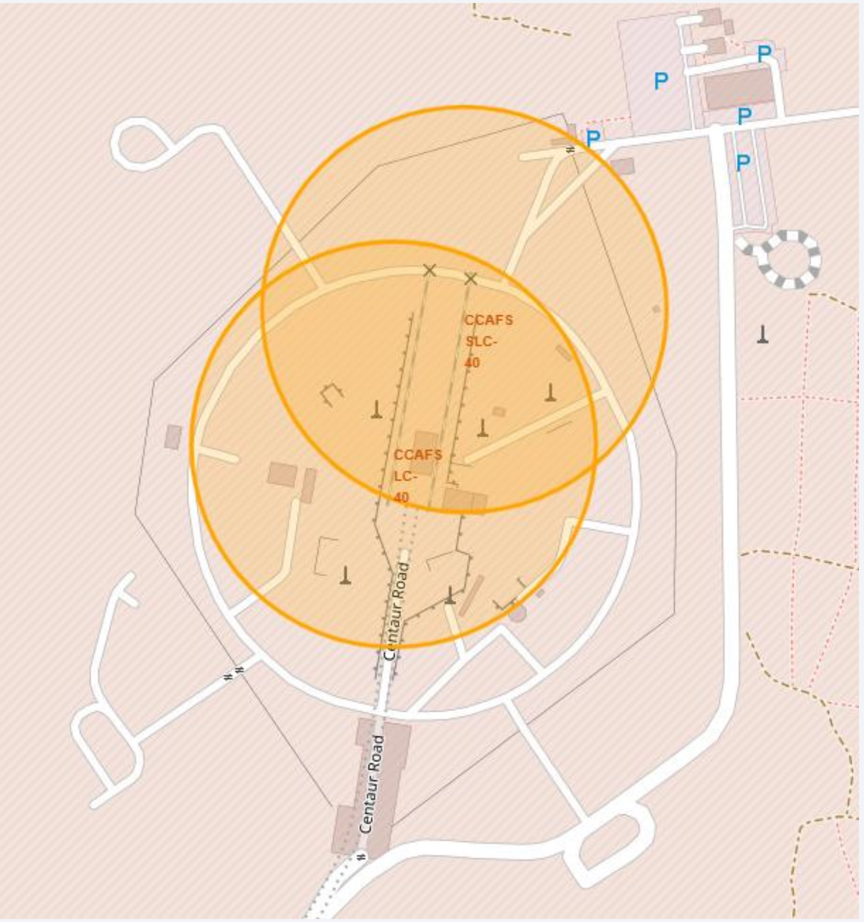
Launch sites Map



Launch sites Map

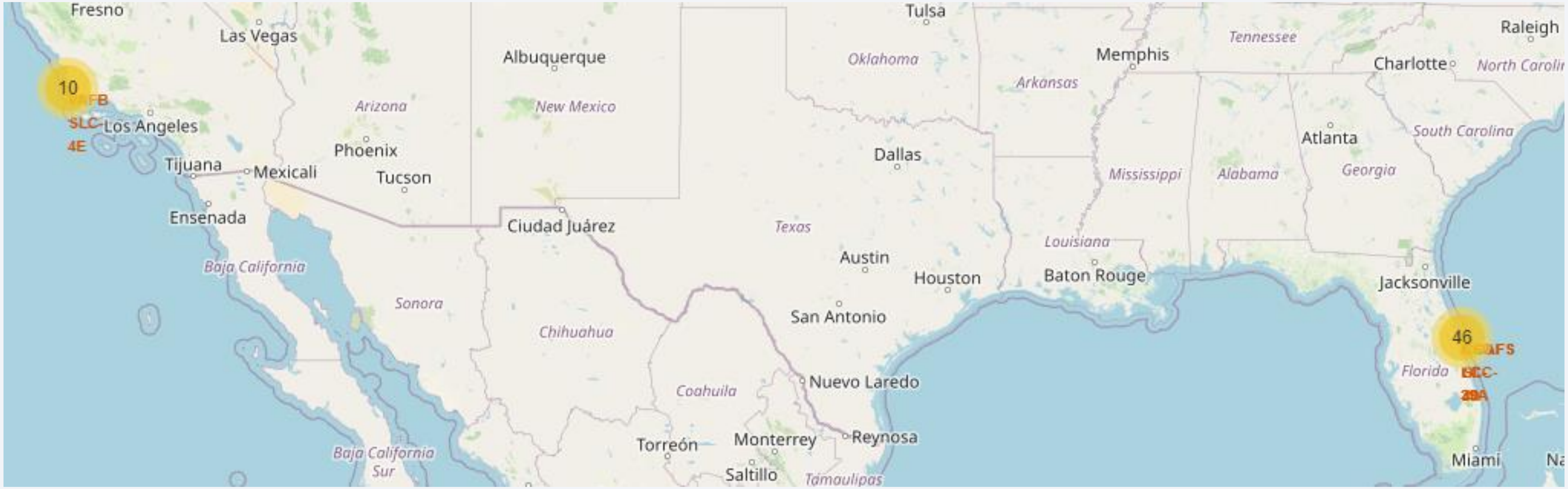




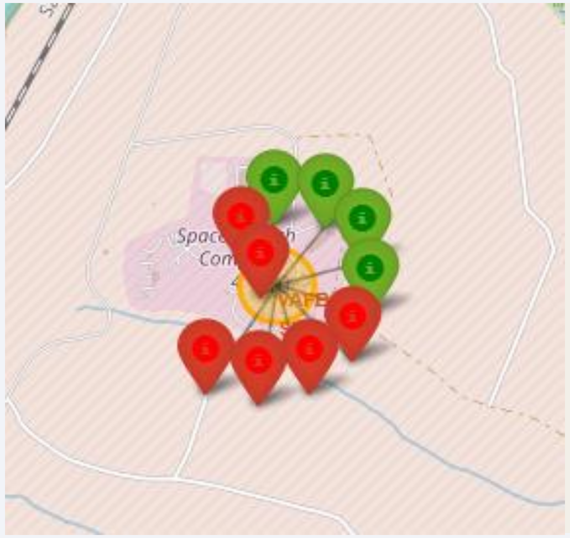


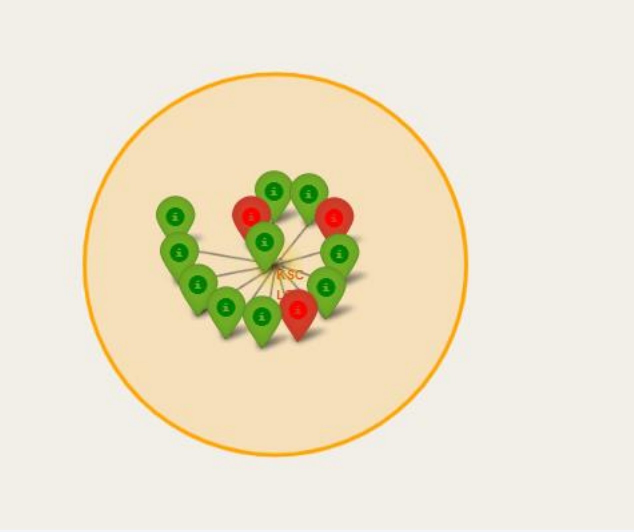
- There are 4 Launch sites 3 on the East coast and 1 on the West coast
- The sites are located near the sea, but also close to the key infrastructure like railroads and highways.

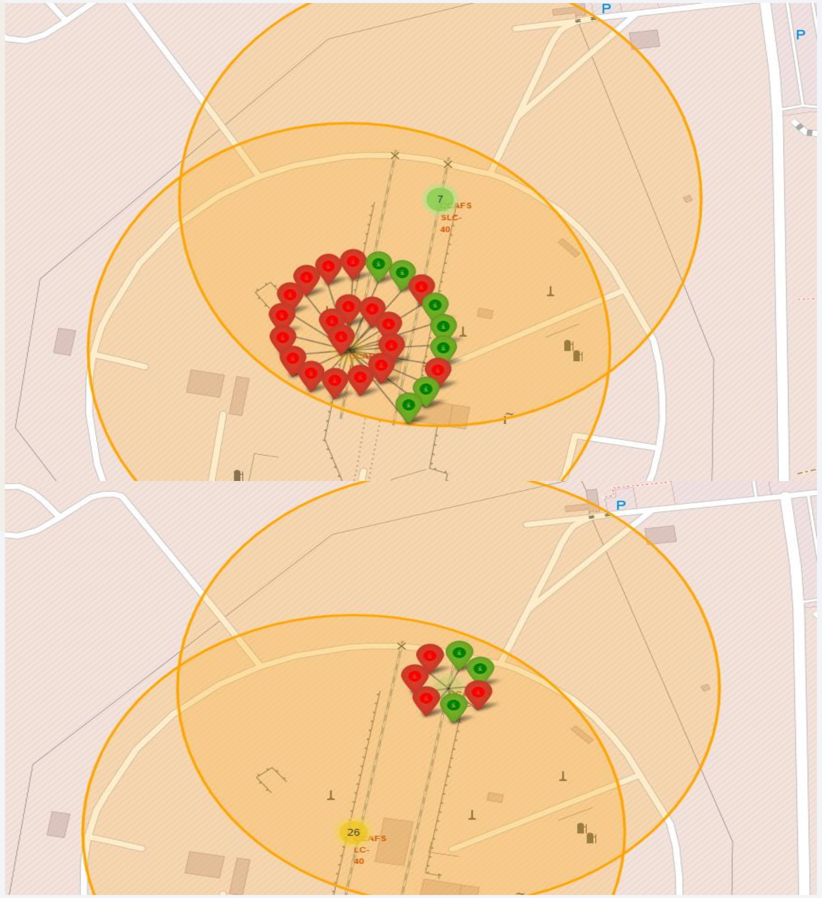
Launch outcomes map



Launch outcomes map







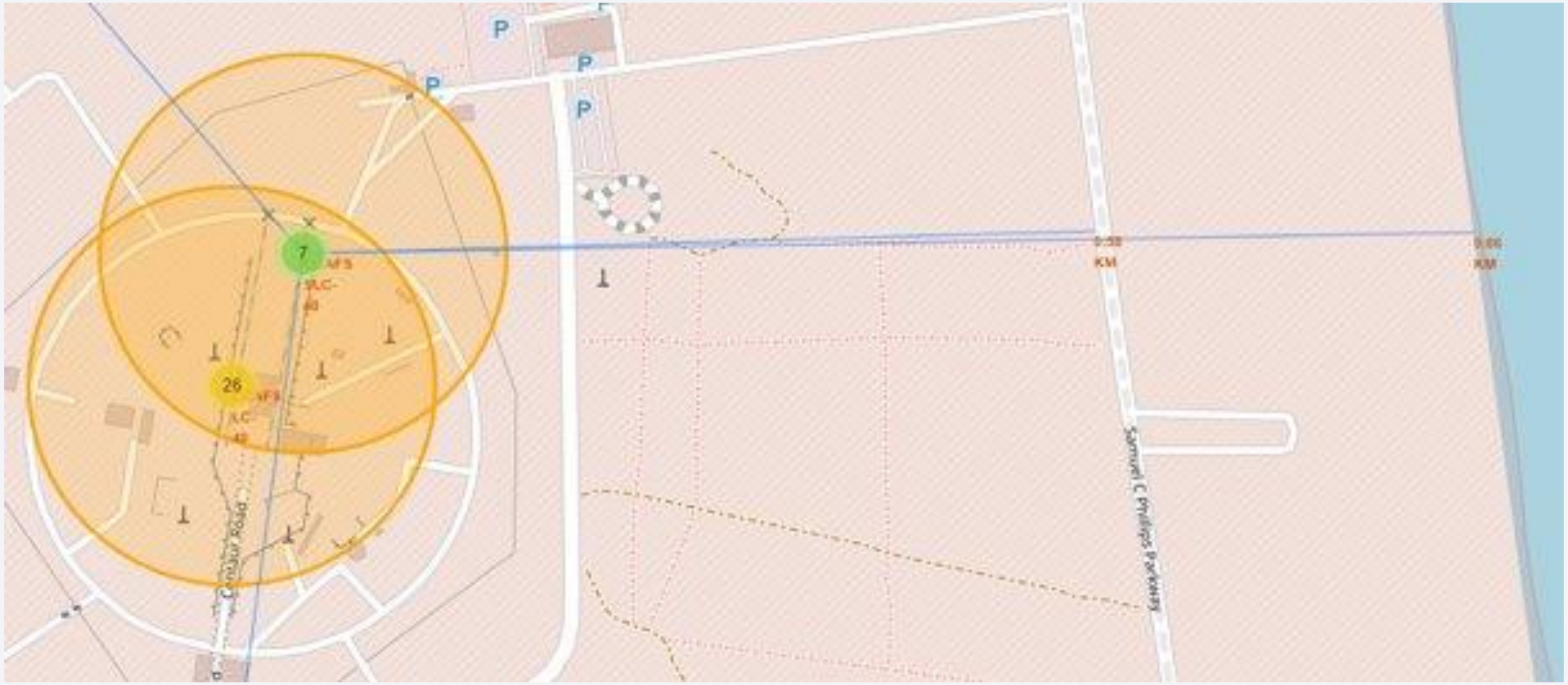
- Green markers indicate success and red ones failure.

Distance to the closest City

General Aviation Drive
General Aviation Drive
General Aviation Drive
North Apollo Boulevard
Charles D
Lorrain
9L/27R
51.43
KM
9R/27L
5/23
oulevard

Distance to the Closest City – 51.43 km

Distance to the closest highway and sea



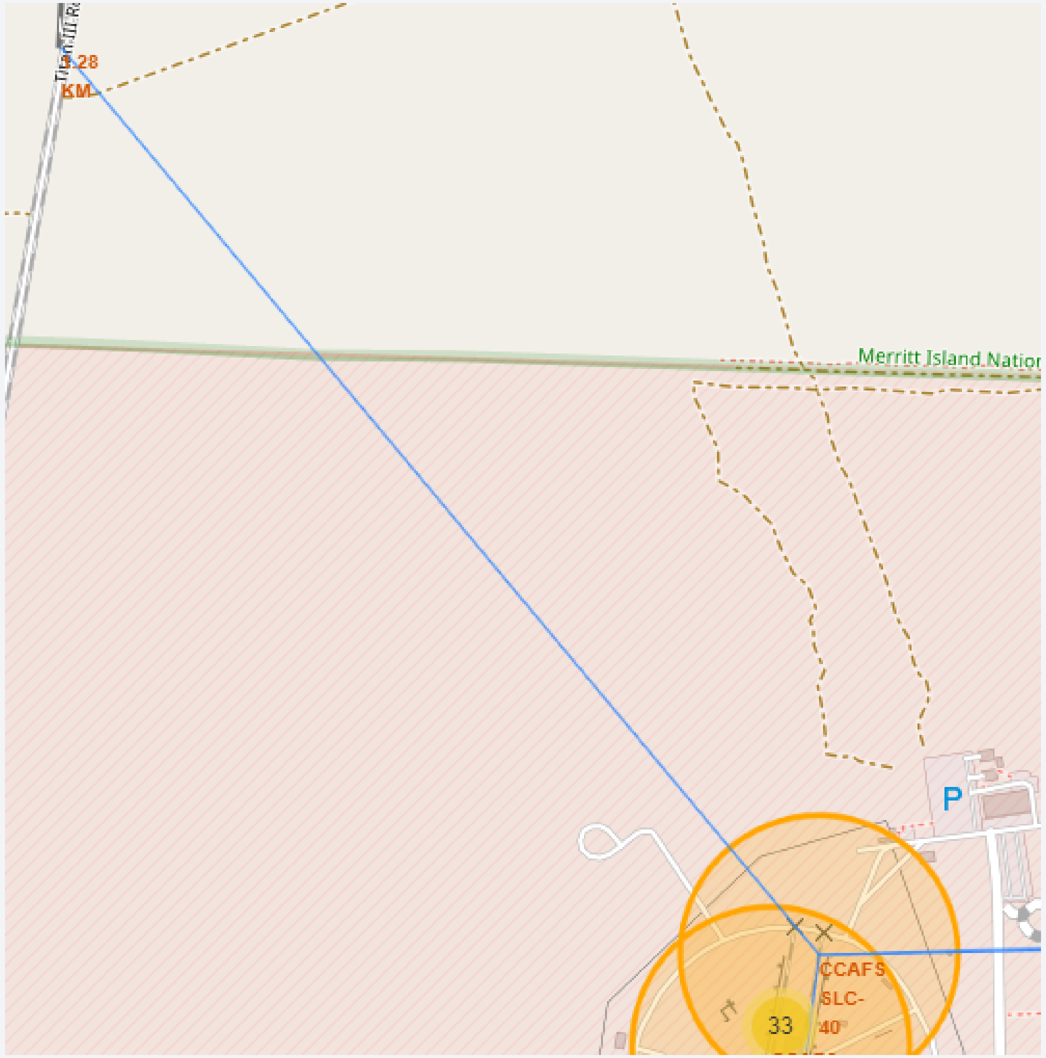
Distance to the closest highway and sea:

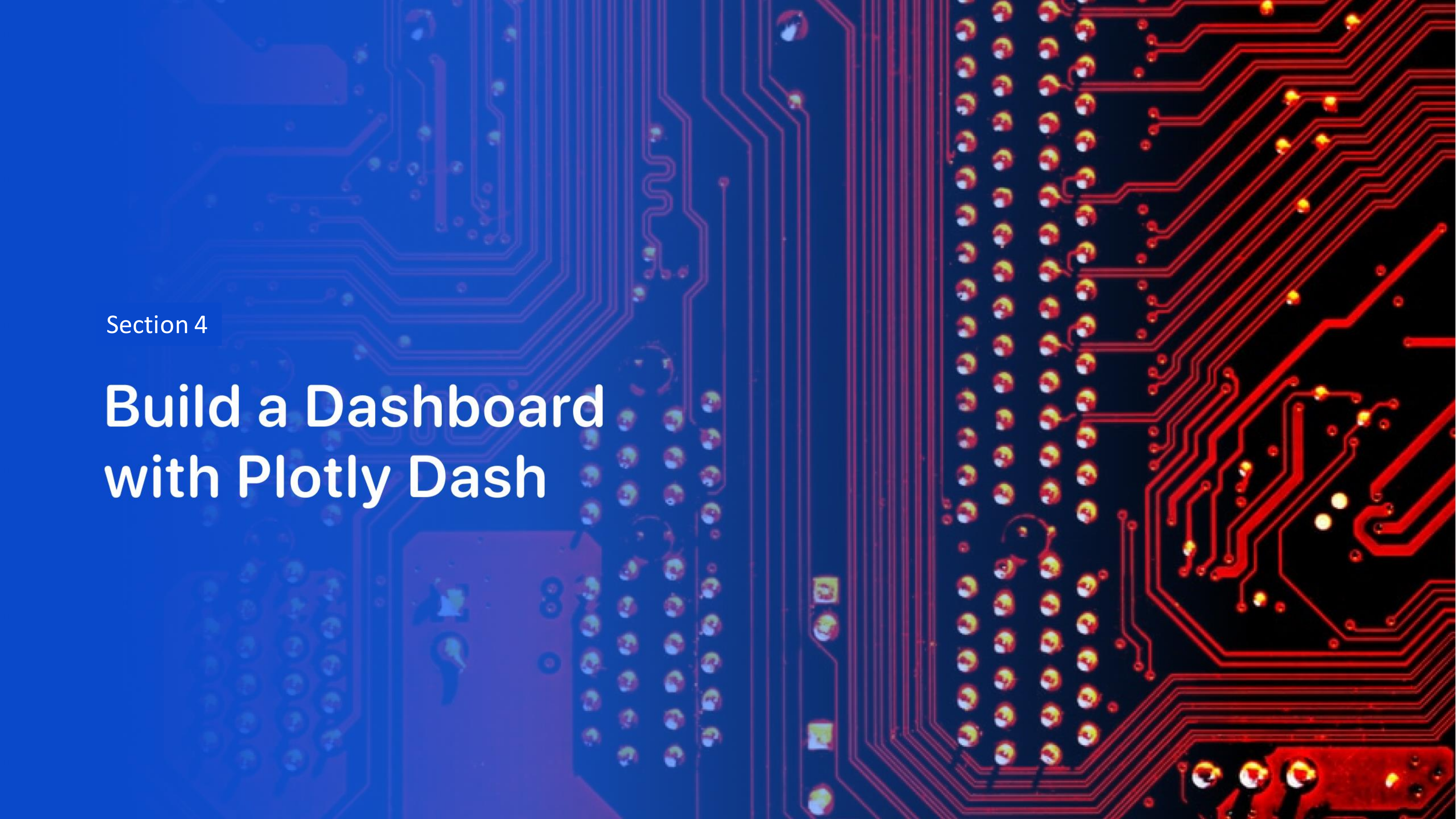
- Highway – 0.58 km
- Sea – 0.86 km

Distance to the closest rail road

Distance to the closest railroad

- Rail road– 1.28 km

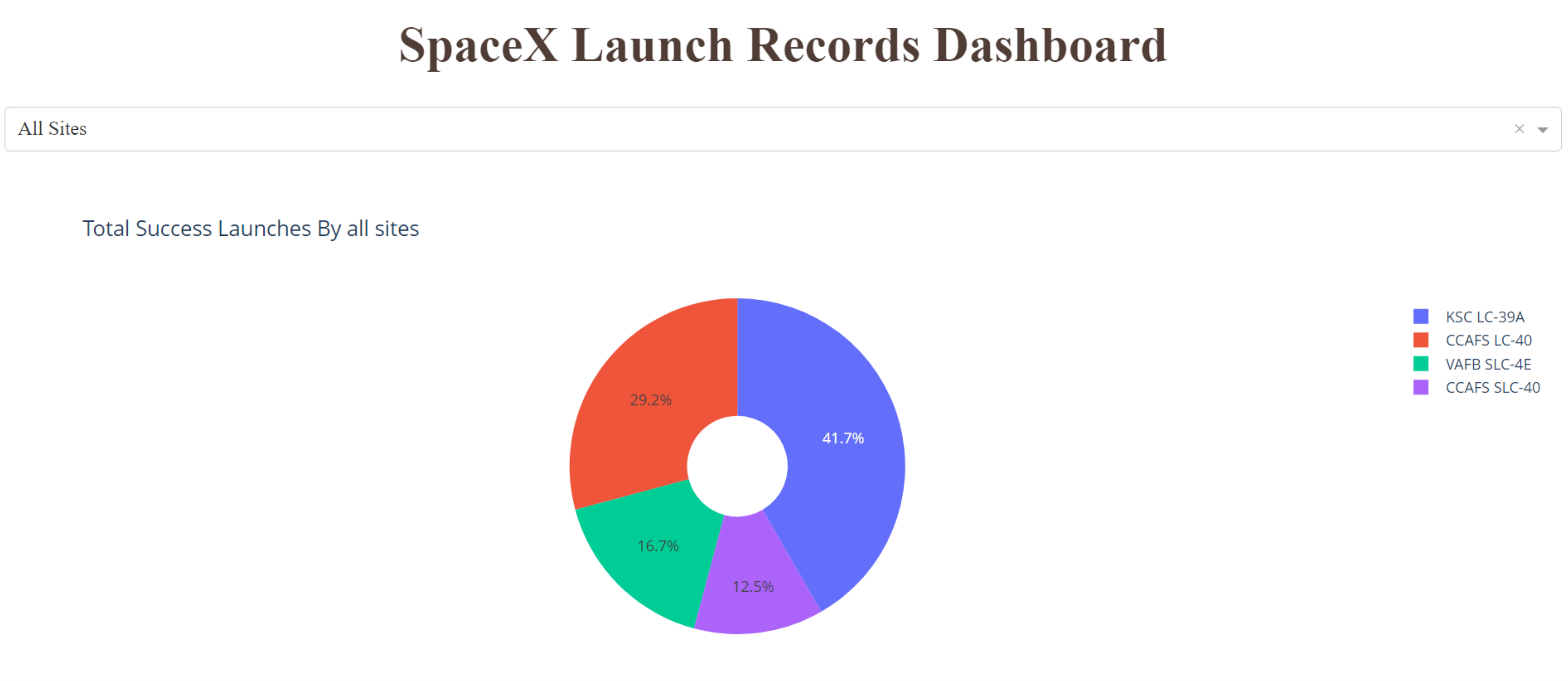




Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

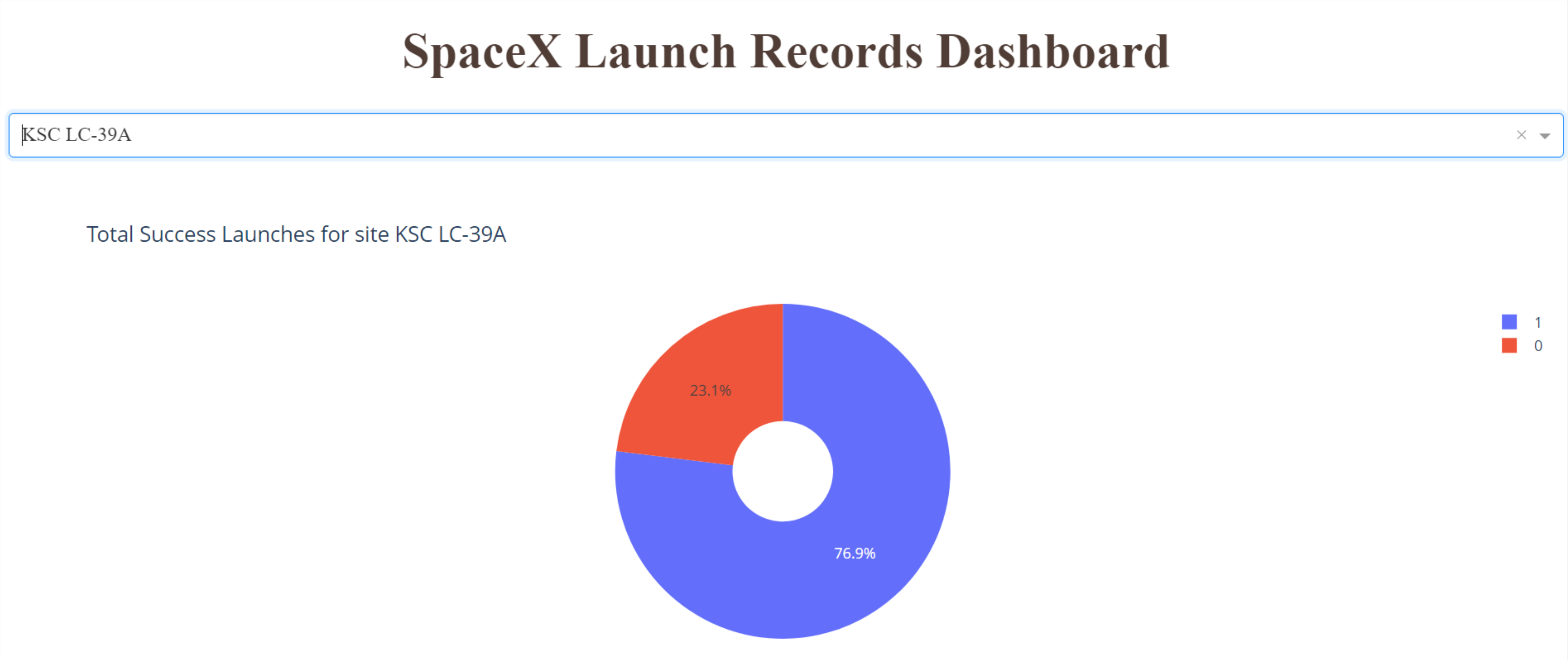


SpaceX Launch Records Dashboard

All Sites × ▾

- KSC LC-39A has the highest success rate of all the launch sites

SpaceX Launch Records Dashboard



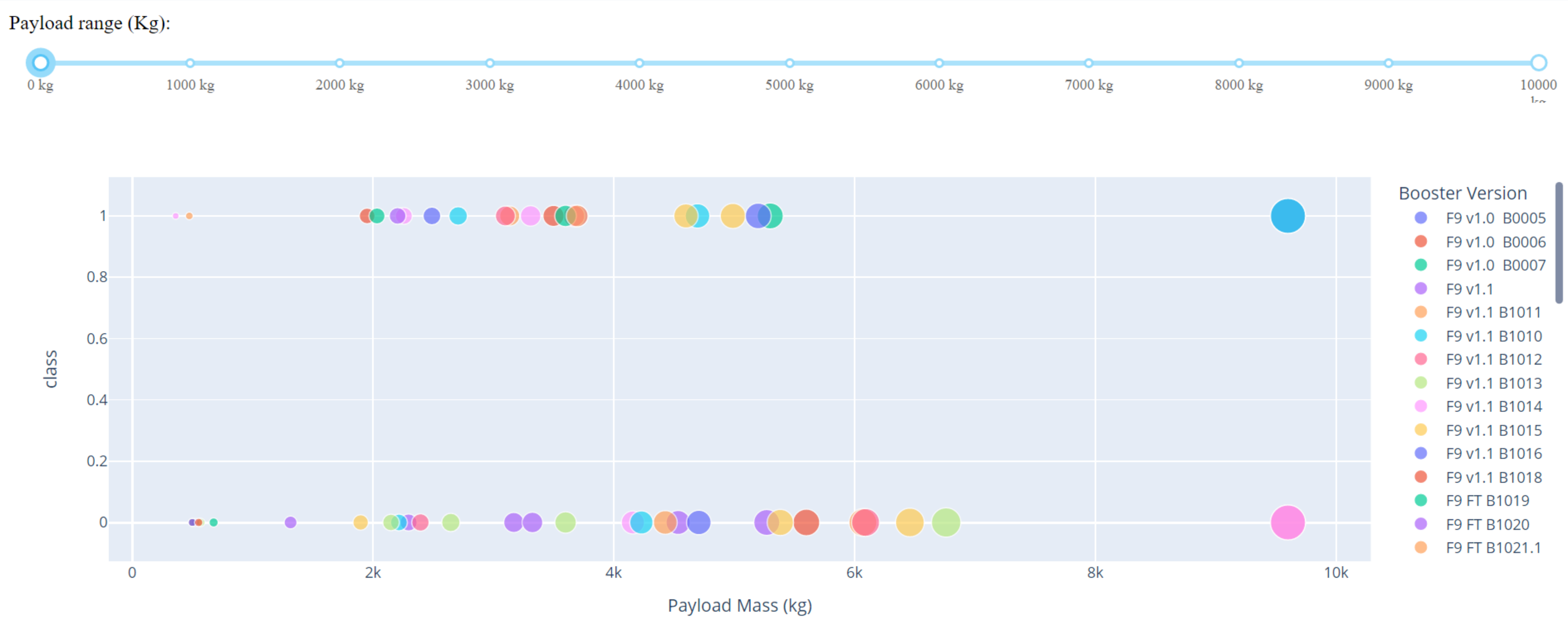
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

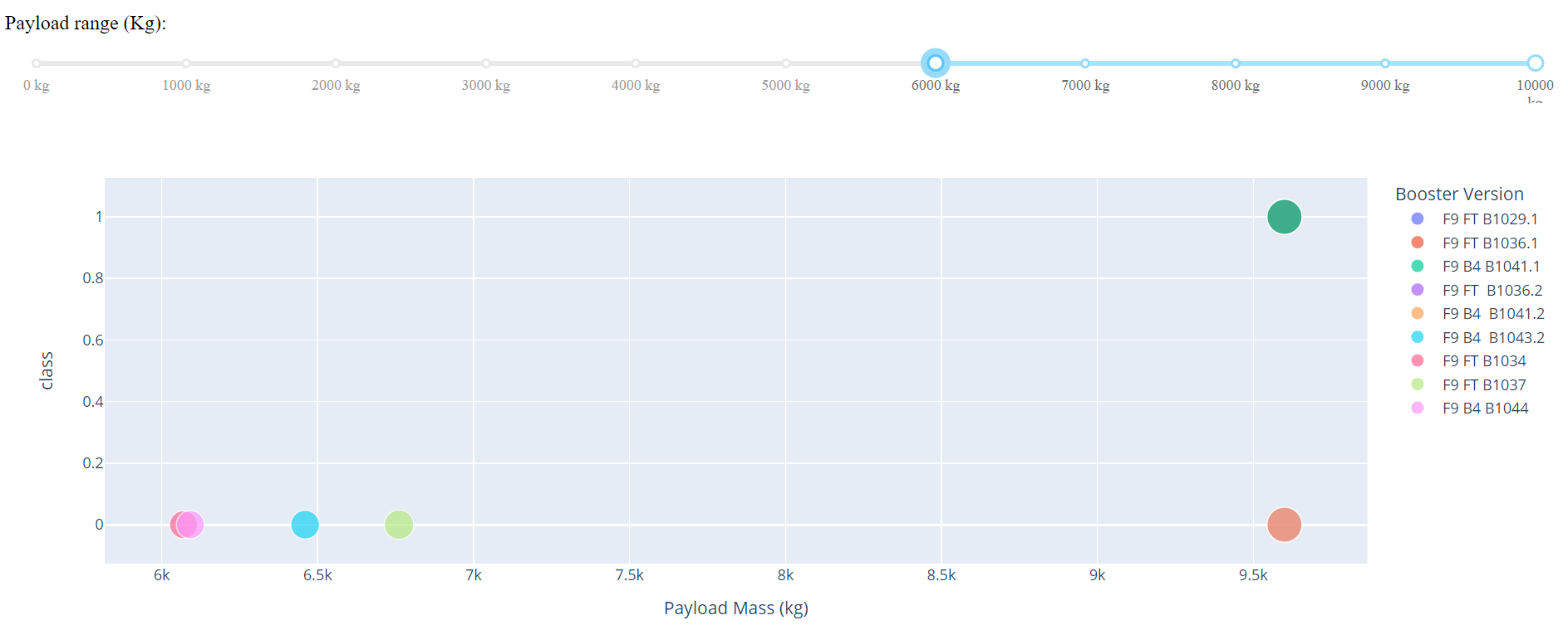
- KSC LC-39A has a 76,9% launch success rate.

SpaceX Launch Records Dashboard



- Most of the successful launches seem to be in the payload range of 2000kg to 6000kg

SpaceX Launch Records Dashboard



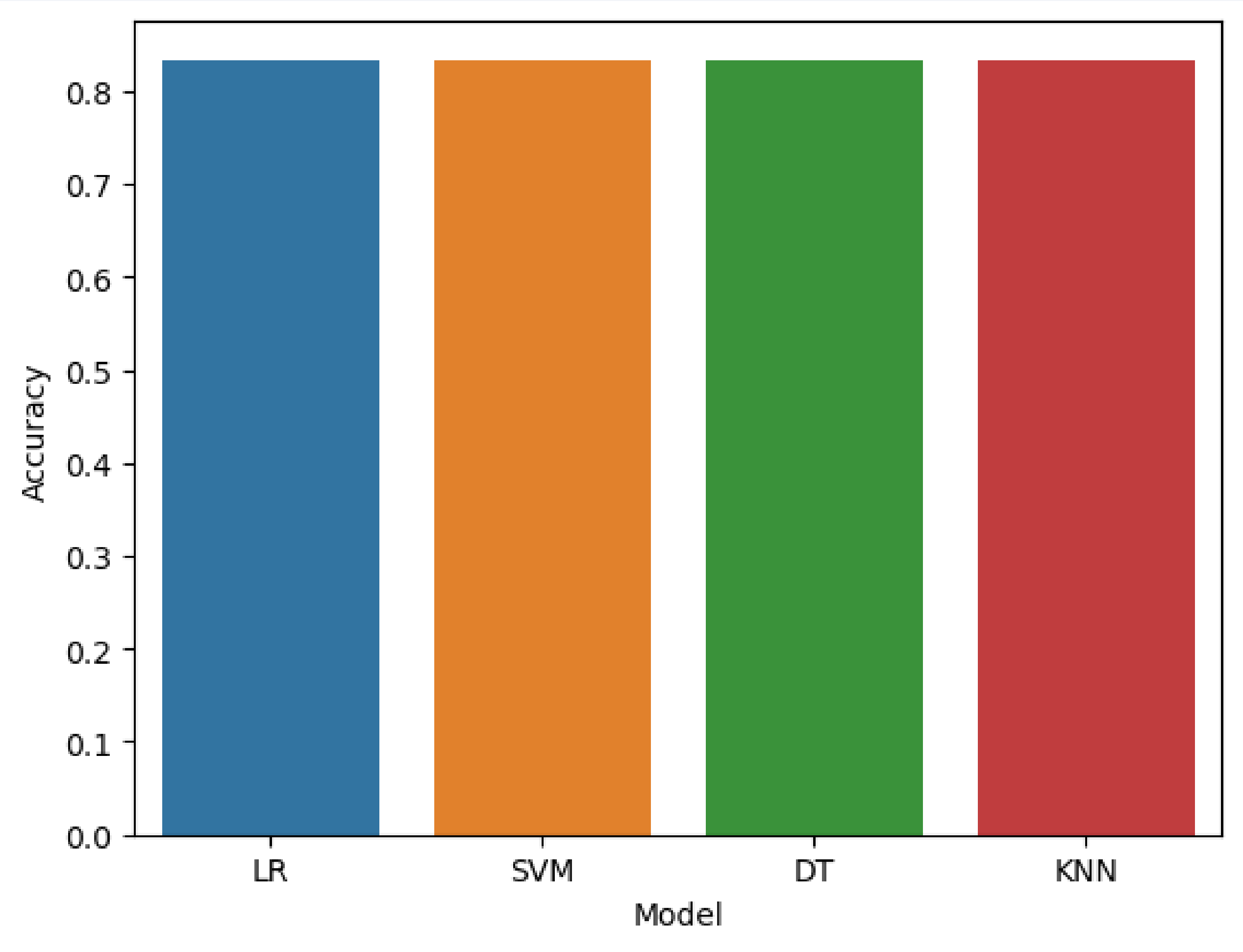
Payload range (Kg):

- There has been only 1 successful launch in the payload range of 6000kg to 10000kg

Section 5

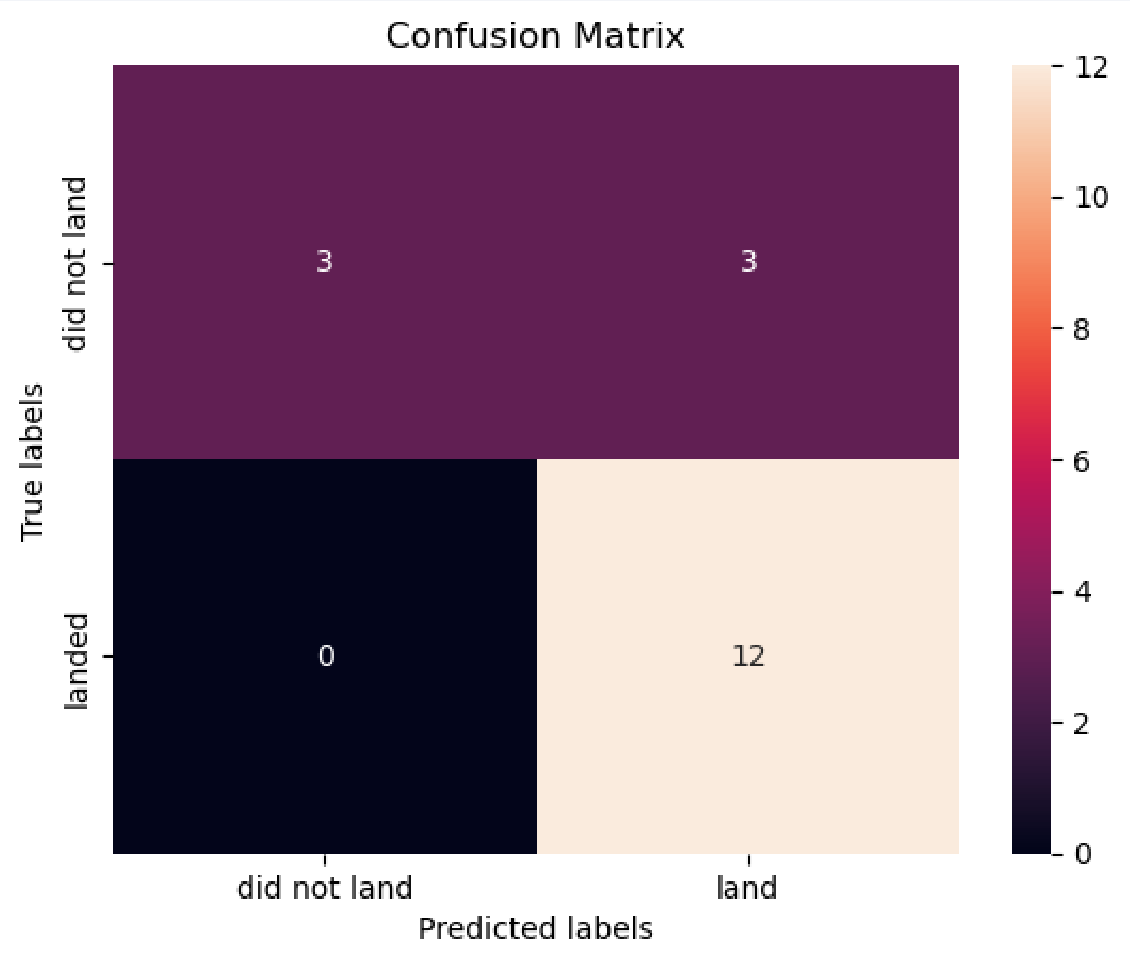
Predictive Analysis (Classification)

Classification Accuracy



- All models have a similar accuracy of 83%

Confusion Matrix



- Model's inaccuracy comes from the significant number of False Positives.

Conclusions

- Although all models displayed similar accuracy Decision tree should hardly be trusted. It has displayed significant variations for concurrent runs which leads me to the conclusion that the model is overfitting. This is most likely caused by having too many features and to little instances.
- Launch Sucess rate since 2013 kept increasing till 2020.
- ES-L1, GEO, HEO & SSO are the most successful orbit types launches.
- KSC LC-39A site has a 76,9% launch success rate which makes it the most successful site.
- General success rate improved over time
- Most Launches took place at CCAFS SLC 40

Thank you!