IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Julia Lenc

21/03/2023

Outline

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

Executive summary

Methodology:

Data from public SpaceX API and SpaceX Wikipedia page =>SQL, Visualization, Folium Maps, Dashboards => Machine Learning models to predict success of future launches.

Results:

All models show similar, high accuracy in predicting launch results (~80%).

Despite certain characteristics (eg. light weight) Falcon 9 has chance for successful landing because Space X has mastered launches during last ten years, knowing best location, payload, orbits and has 80%+ success rate.

Introduction

Background:

SpaceX prices Falcon 9 app. 100MM$ lower than other companies. But does it have a chance of success?

Question:

Will Falcon 9 land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology - API requests and Web Scraping
- Data wrangling – recoding, unifying naming and cleaning (duplicates, Nas, etc.)
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models and recommending model(s) with best fit

Data Collection

Collection methods

1. API requests from SpaceX public API.

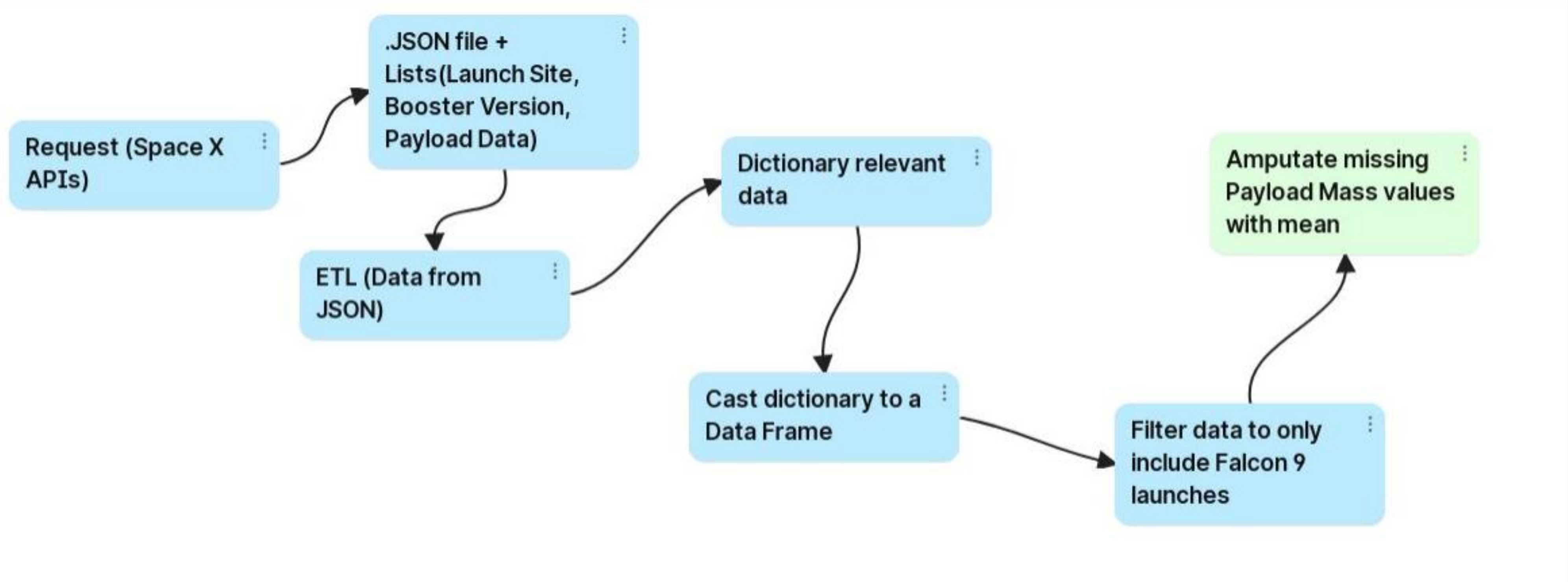
    Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights,

    Grid Fins, Reused, Legs, Landing, Block, Reused Count, Serial, Longitude, Latitude

2. Web scraping from SpaceX´s Wikipedia.

    Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Request (Space X APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

ETL (Data from JSON)

Dictionary relevant data

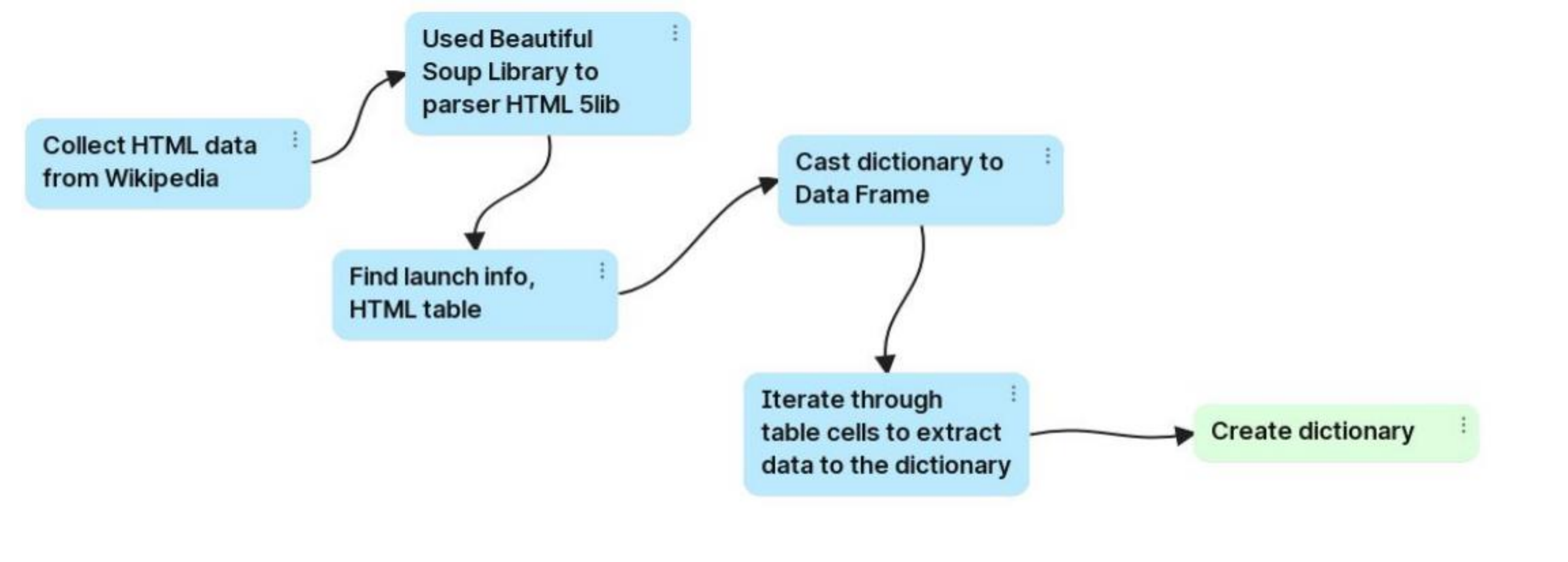
Cast dictionary to a Data Frame

Filter data to only include Falcon 9 launches

Amputate missing Payload Mass values with mean

Github: [IBM_Data-Science_Capstone-Project_SpaceX/W1_Data_Collection_API.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)]

Data Collection - Scraping



Collect HTML data from Wikipedia

Used Beautiful Soup Library to parser HTML 5lib

Find launch info, HTML table

Cast dictionary to Data Frame

Iterate through table cells to extract data to the dictionary

Create dictionary

Github: IBM_Data-Science_Capstone-Project_SpaceX/W1_Data_Collection_Web_Scraping.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)

Data Wrangling

- Training label with landing outcomes where success = 1 & failure = 0
- Outcome column
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is
- Mapping of categorical variables (True set to -> 1, False = 0).

Github: IBM_Data-Science_Capstone-Project_SpaceX/W1_Data_Wrangling.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Scatter plots: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit
- Line chart: Success Yearly Trend

Github: IBM_Data-Science_Capstone-Project_SpaceX/W2_EDA_Visualizations.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

Github: IBM_Data-Science_Capstone-Project_SpaceX/W2_EDA_SQL.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

Github: IBM_Data-Science_Capstone-Project_SpaceX/W3_Visual_Analytics_With_Folium.ipynb at main · julialenc/IBM_Data-Science_Capstone-Project_SpaceX (github.com)

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch

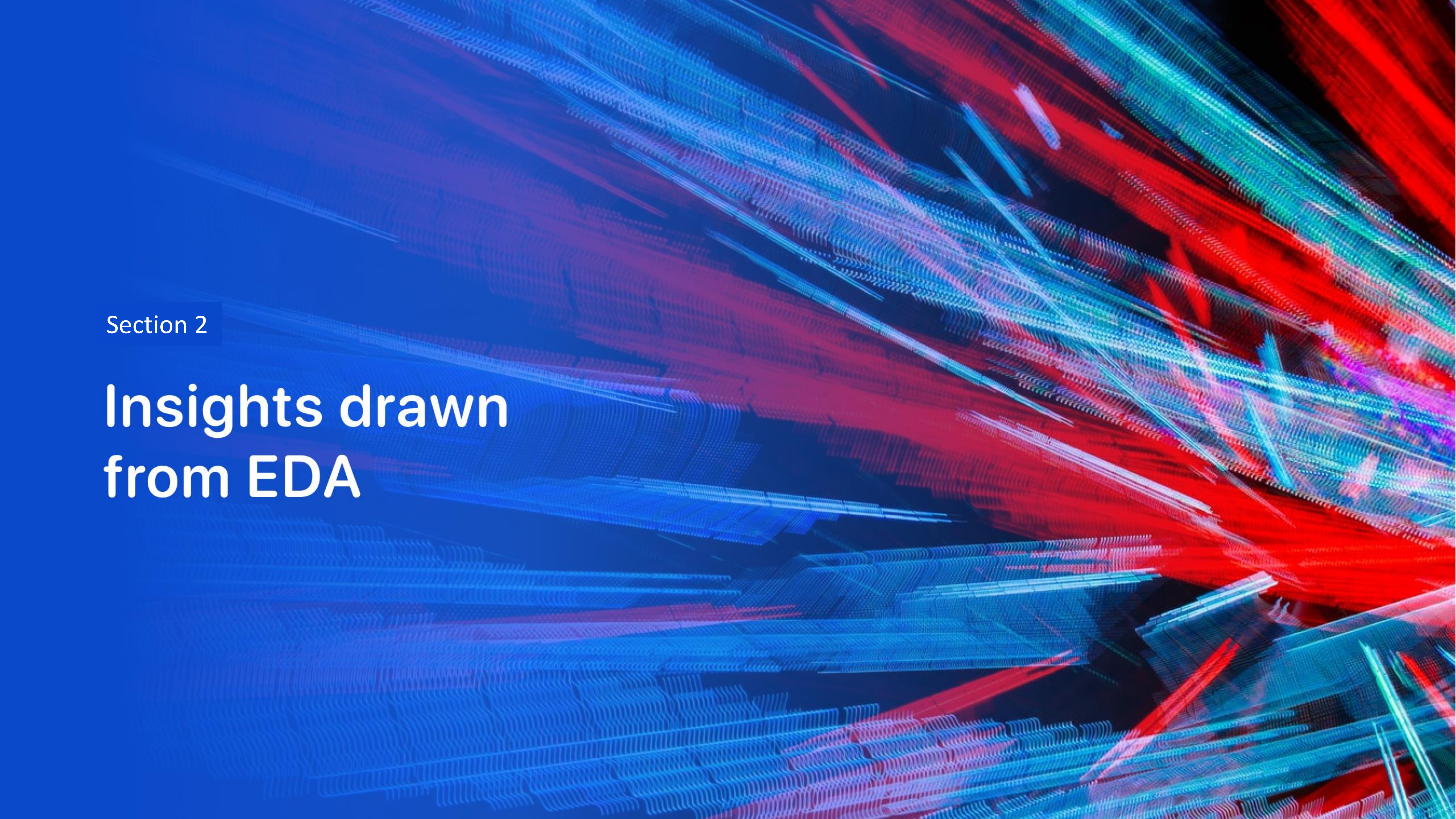
sites or separately.

- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

Github: https://github.com/julialenc/IBMDataScience/blob/main/Capstone%20Project/Week%203%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Section 2

Insights drawn from EDA

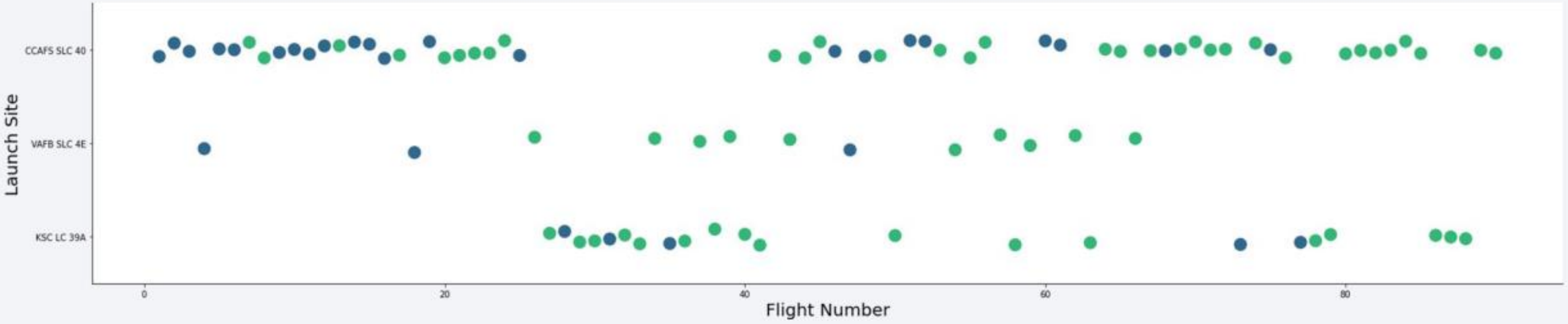


Flight Number vs. Launch Site

The chart below show distribution of launches by location (vertical axis) vs flight number (horizontal axis) vs outcome (green = success, blue = failure).

We can observe:

- Success rate increases over time (higher flight number = higher success rate)
- Predominant launch location is CCAFS. Due to many unsuccessful landings at the beginning of experimentation, location was moved to KSC; however, it is clear that not location but rather technical advancement is a determinant of success.

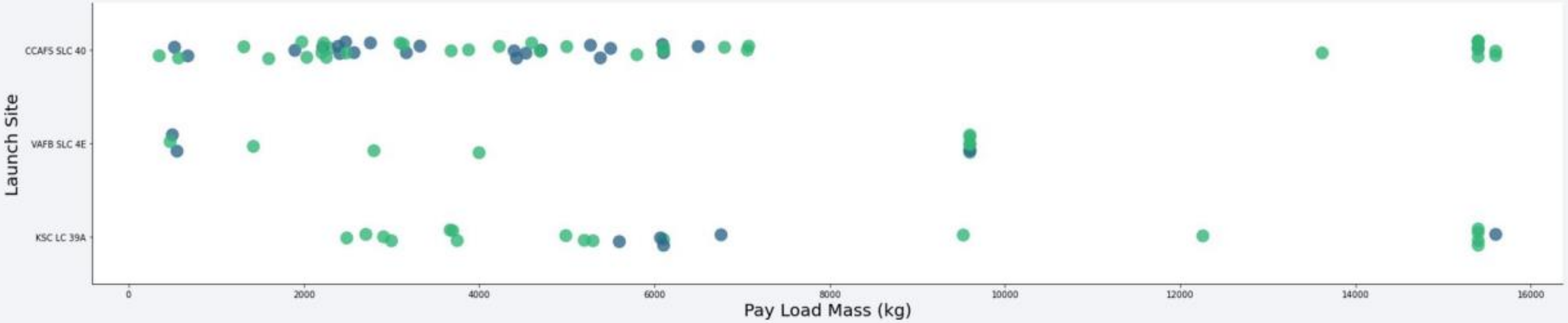


Payload vs. Launch Site

The chart below show distribution of launches by location (vertical axis) vs payload mass (horizontal axis) vs outcome (green = success, blue = failure).

We can observe:

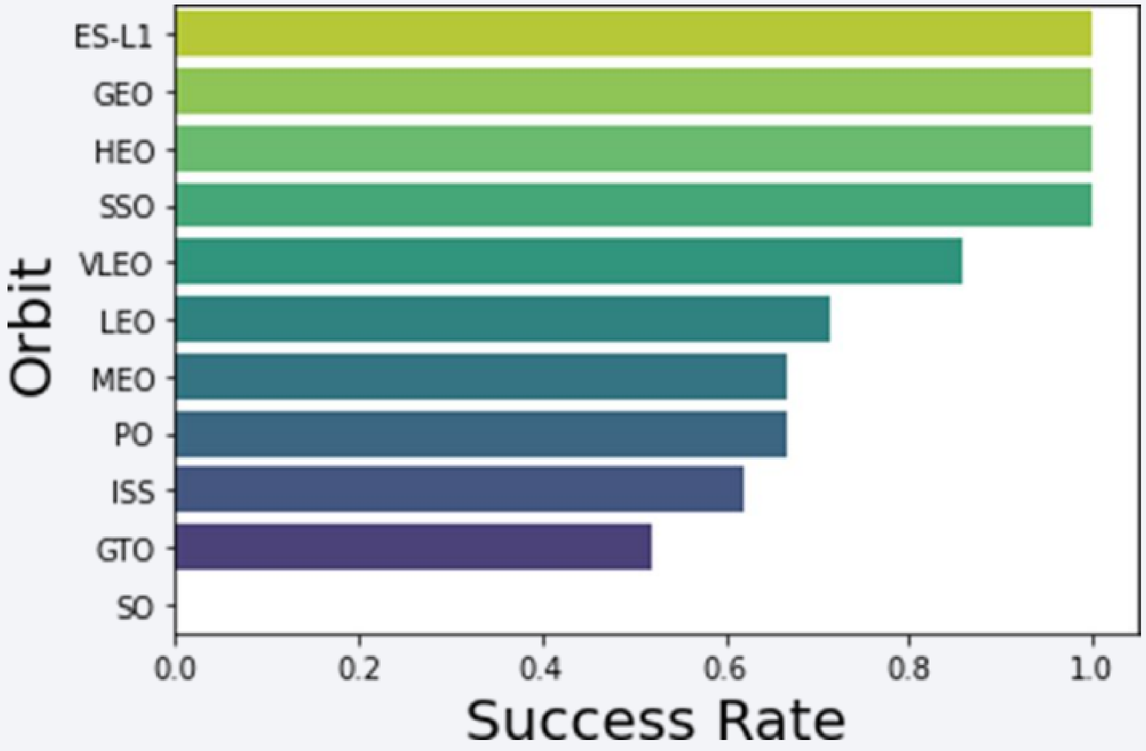
- Success rate seems to be higher for heavier payloads, independently on launch site.
- Predominant payloads for CCAFS and KSC are light, followed by heavy.
- VAFB “specialized” in medium payloads.



Success Rate vs. Orbit Type

Success rate differs by orbit:

- 100% success rate for ES-L1, GFO, HEO and SSO
- <60% success rate for GTO
- 60-80% success rate for remaining orbits.

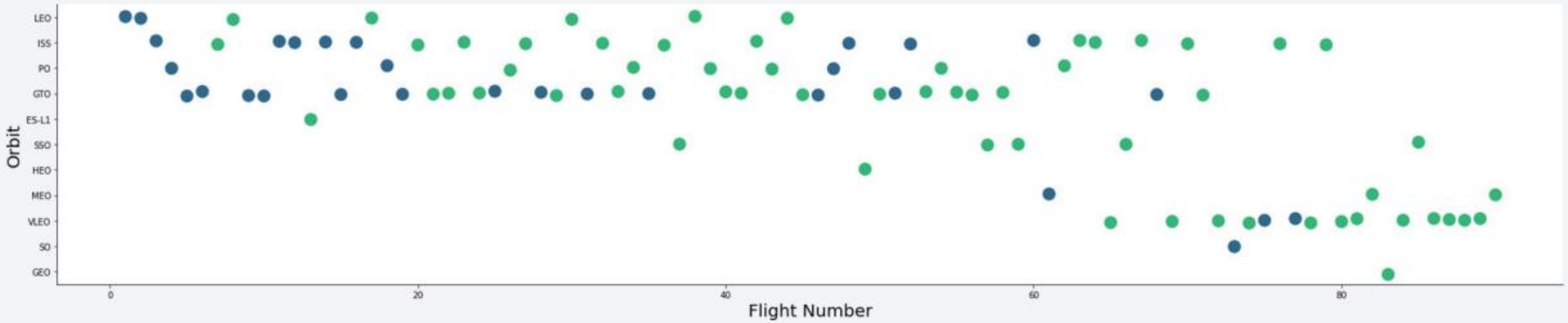


Flight Number vs. Orbit Type

The chart below show distribution of launches by orbit (vertical axis) vs flight number (horizontal axis) vs outcome (green = success, blue = failure).

We can observe:

- There is a clear strategy of testing different orbits and moving to those, where success rate is higher (higher flight numbers for orbits GEO, ES-L1)
- This is also clearly visible that first launches had zero success rate. This could be driven by orbit but also by technical advancement.

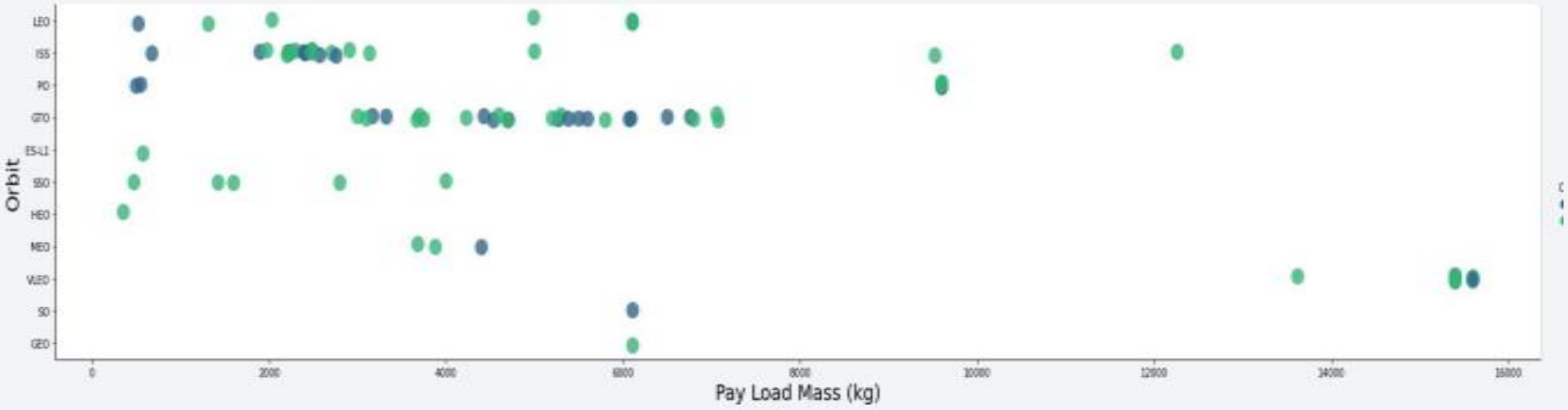


Payload vs. Orbit Type

The chart below show distribution of launches by orbit (vertical axis) vs payload (horizontal axis) vs outcome (green = success, blue = failure).

Taking into consideration previous slides, we can make a hypothesis:

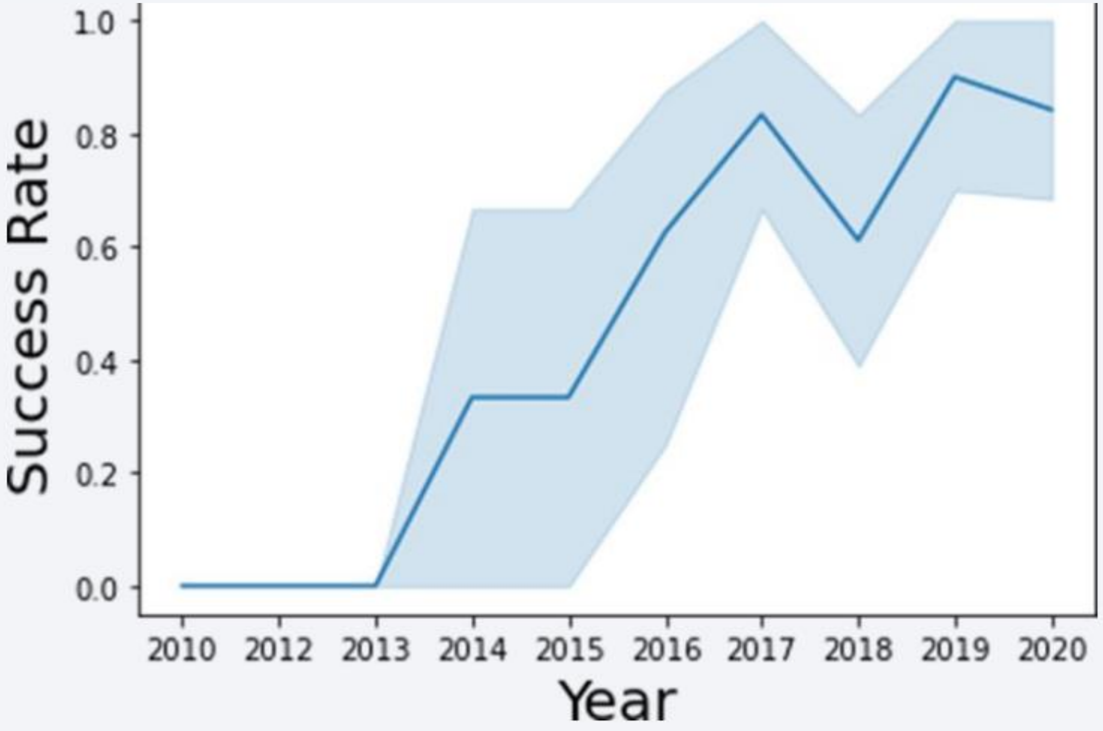
- Success rate depends less on payload and on orbit or technical advancement (flight number). Most successful orbits can be those with predominantly low payload (ES-L1, HEO) and those with high payload (GEO, VLEO).



Launch Success Yearly Trend

Success improves over time:

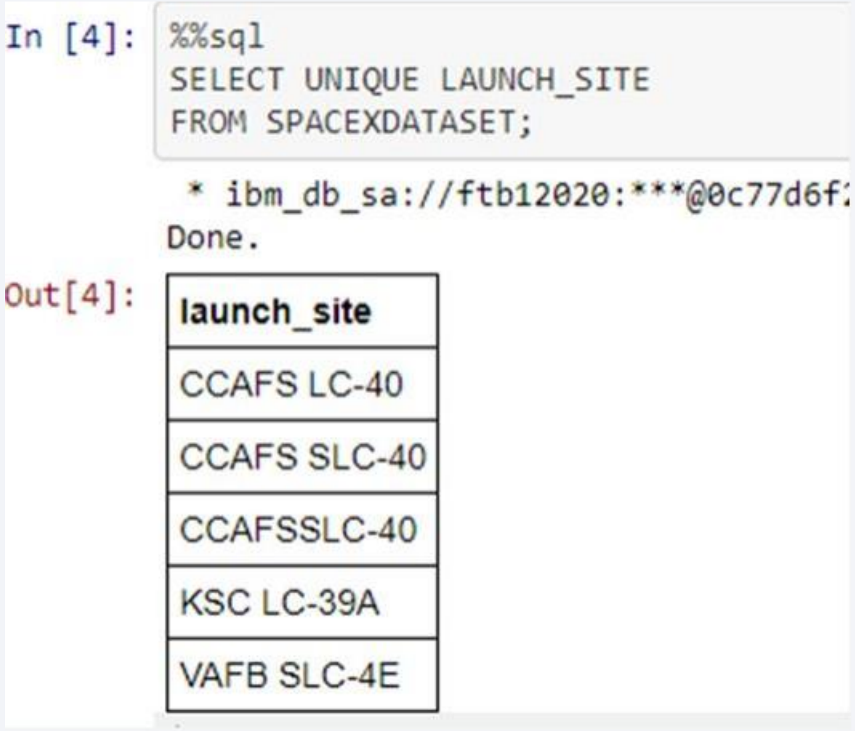
- 0% success rate during first three years of experiments
- Exponential growth from 0% to 80% in period 2014-2017
- Relative stabilization (plateau?) after 2017



All Launch Site Names

As confirmed by map, there are three unique launch sites, while due to errors in data entry the below outcome shows 5 sites.

The error is coming from CCAFS site; therefore the site name for all CCAFS should be unified and re-coded.



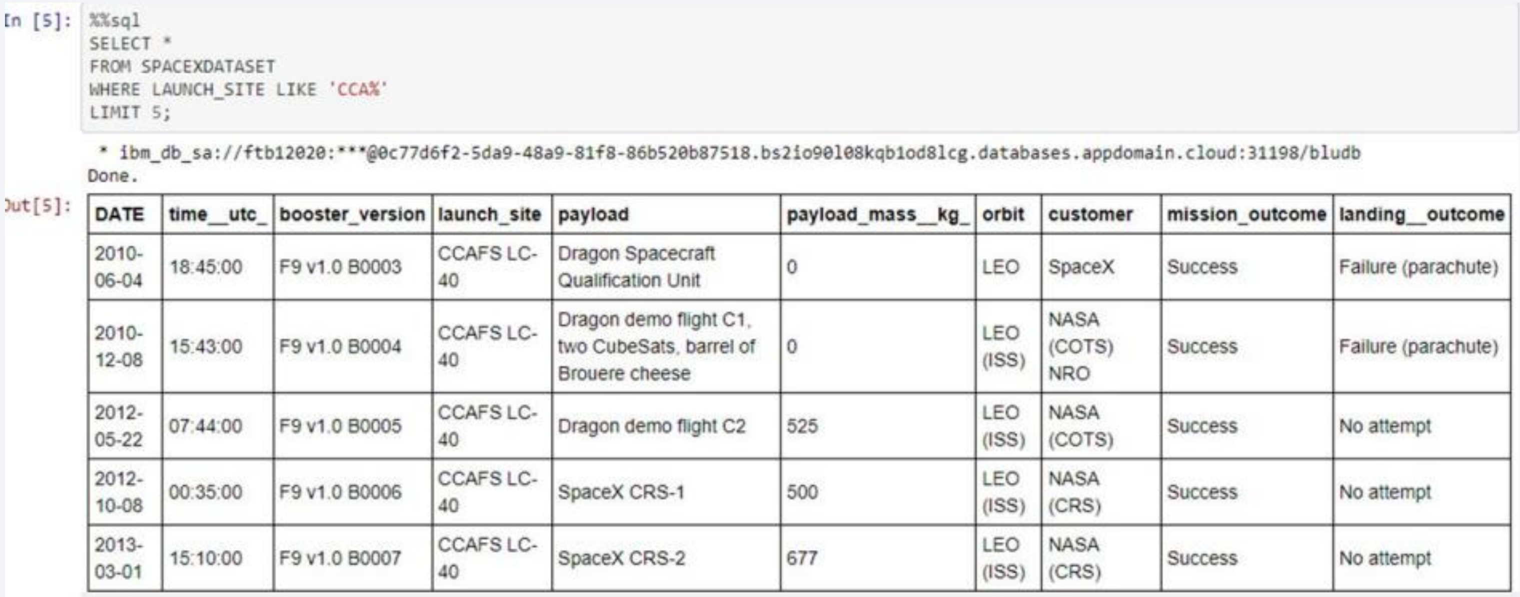
```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f
        Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

On team´s request, the table below presents first 5 entries with launch beginning with CCA

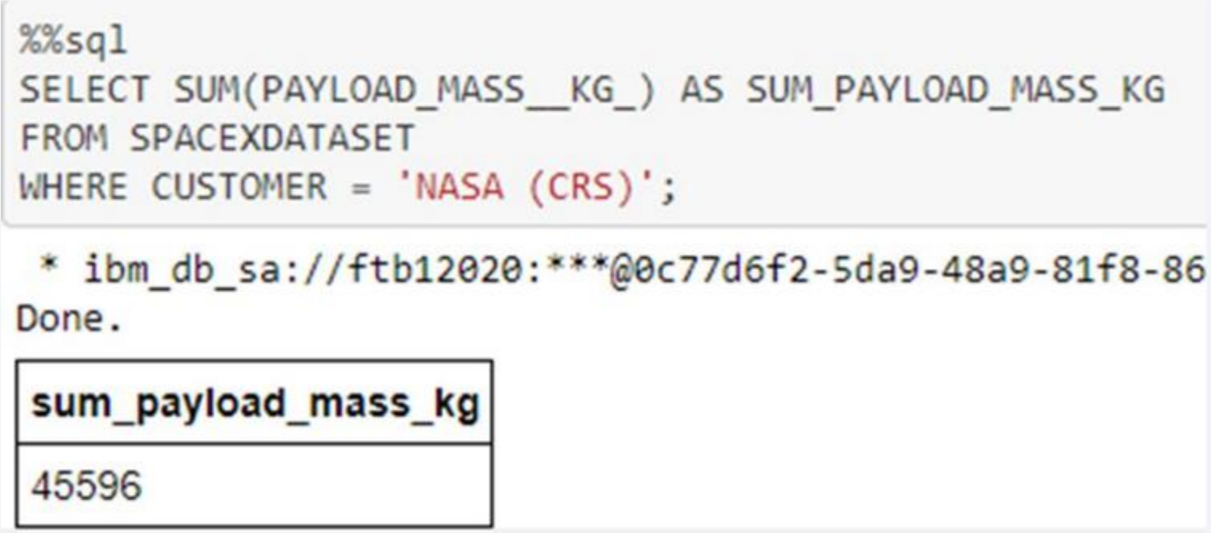


```
In [5]: %%sql
        SELECT *
        FROM SPACEXDATASET
        WHERE LAUNCH_SITE LIKE 'CCA%'
        LIMIT 5;

         * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[5]:
```

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

During experimentation time, total payload of launches done for NASA was 45,596kg.



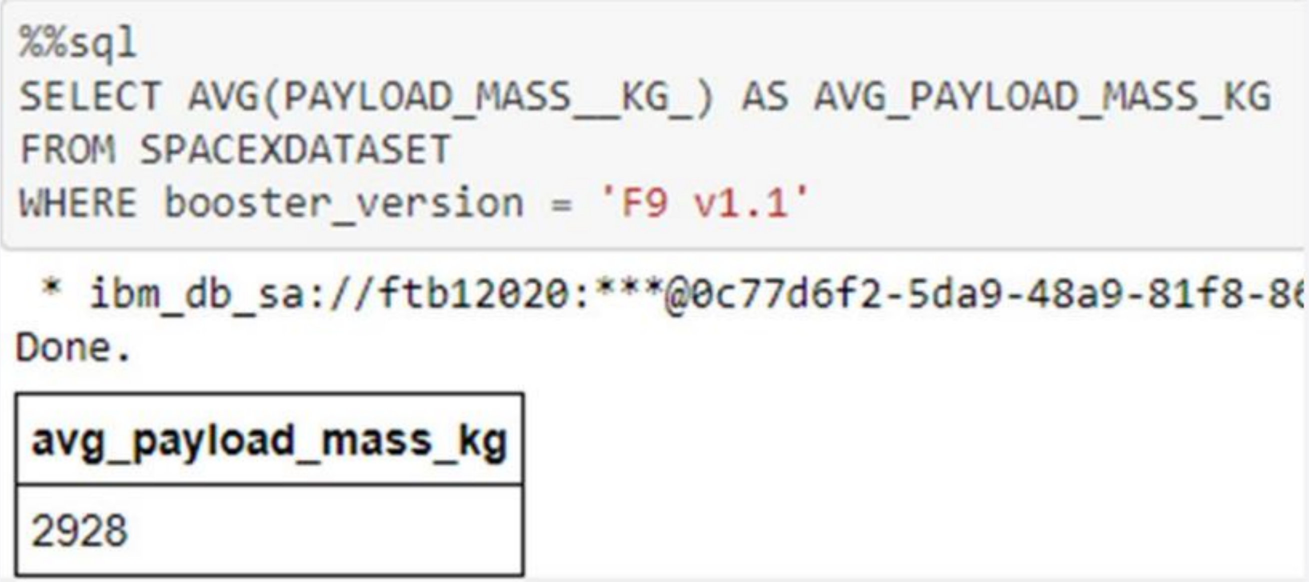
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

Average payload of F9 v1.1 is 2,928kg. This puts it at lower end of payloads during experimentation time.



```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

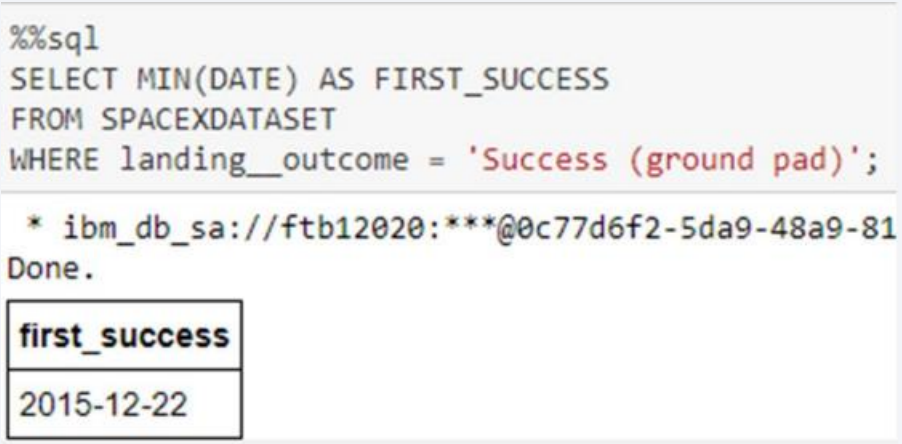
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

First Successful Ground Landing Date

First successful landing took place on Dec 22, 2015.

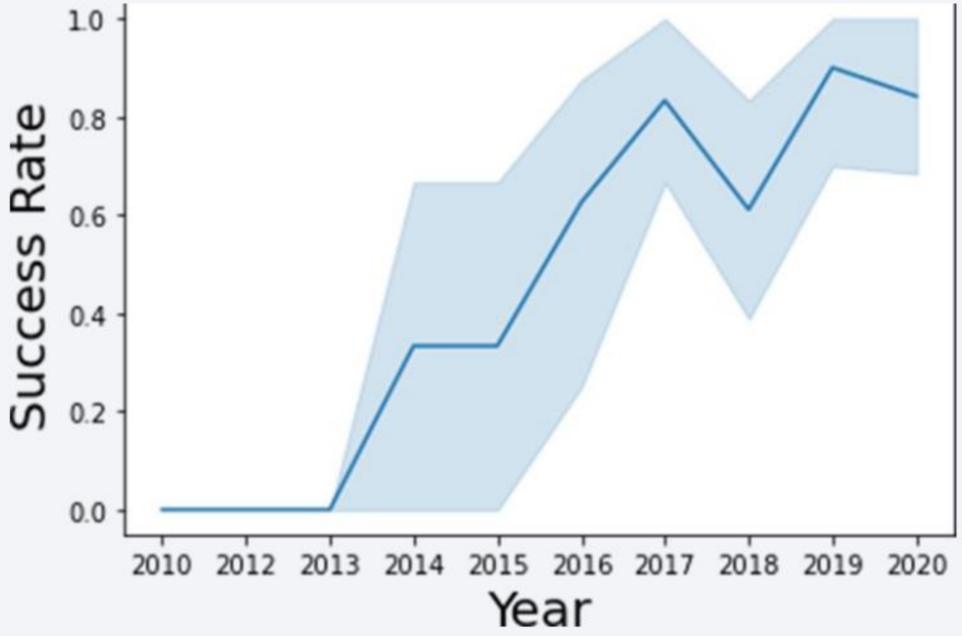
There were no successful landings during first 3 years of experimentation.



```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

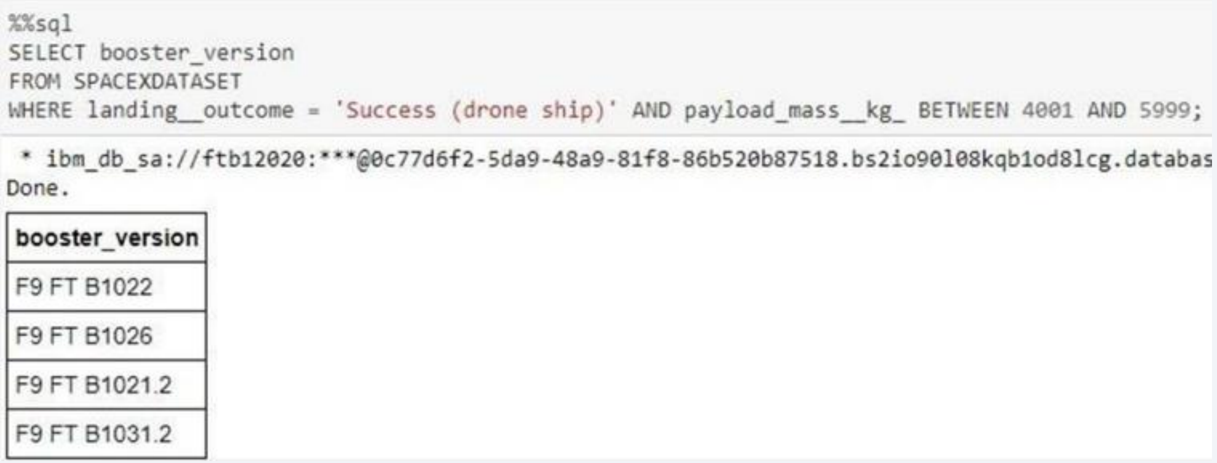
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22



Successful Drone Ship Landing with Payload between 4000 and 6000

Successful drone ship landing for payload 4000-6000kg were for boosters F9 FT B1 family.



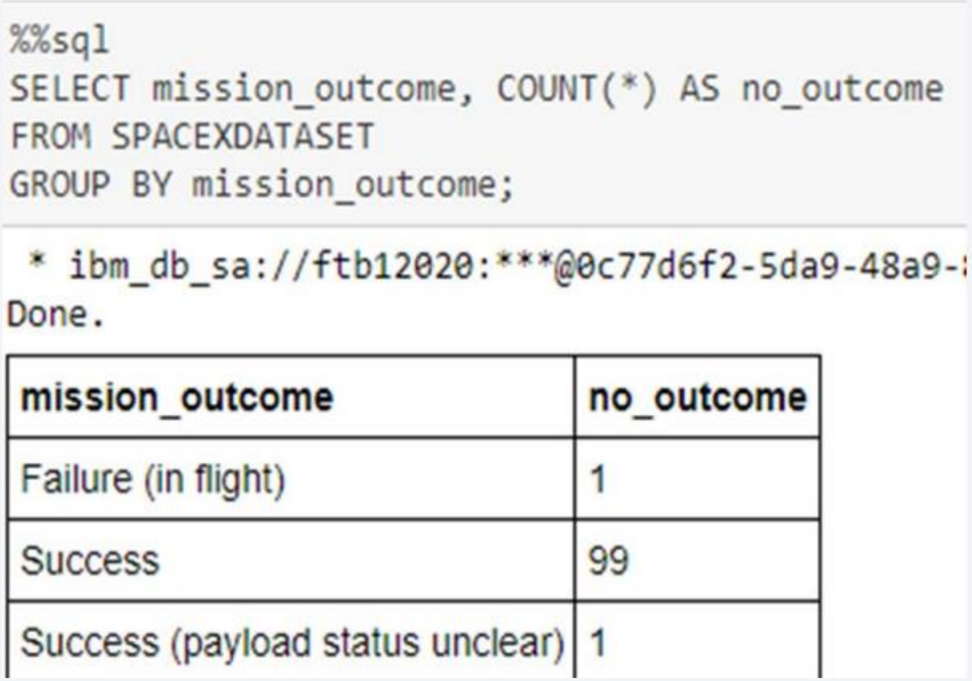
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databas
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

Only 1% of launches finished with failure (1 out of 101 launches).



```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

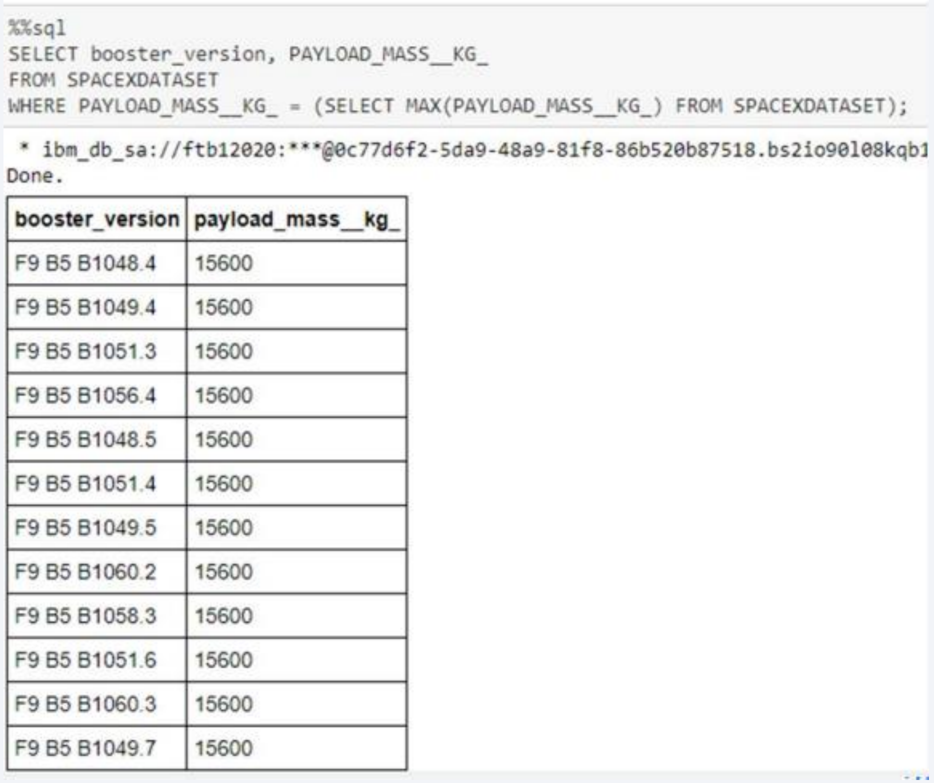
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

Boosters Carried Maximum Payload

The highest payload mass was 15600 kg and for F9 B5 B10--.- models.

We can make a hypothesis that booster type depends on payload.



```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

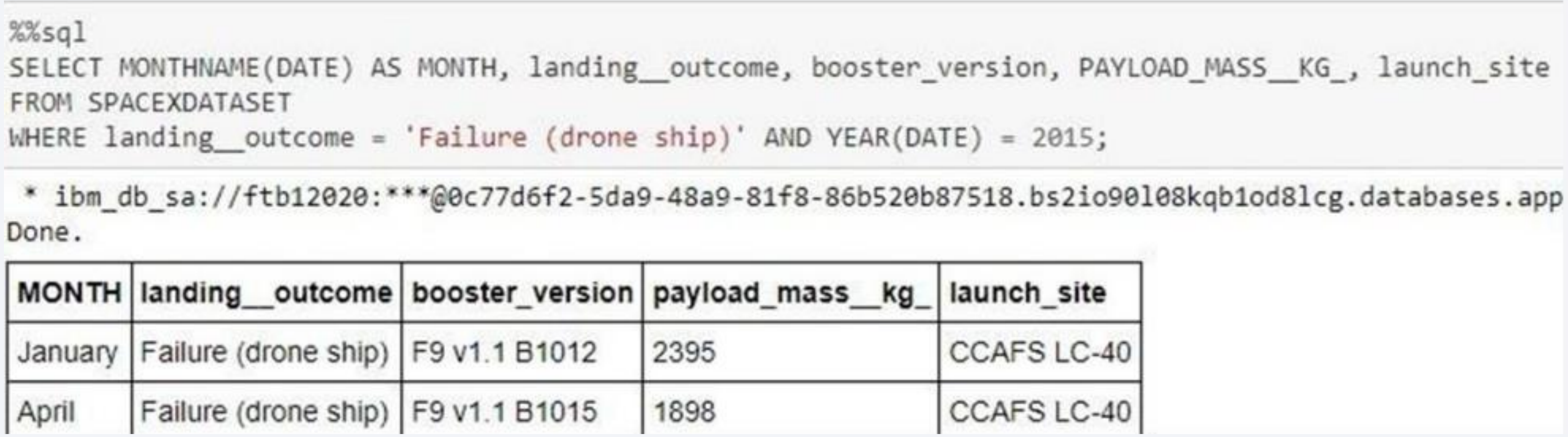
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

In 2015, two launches failed to land, both:

- Were launched from the same site (CCAFS LC-40)
- Both failed to land on a drone ship
- Both had low load (<3000kg)
- Both were version F9 v1.1. B10--



```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

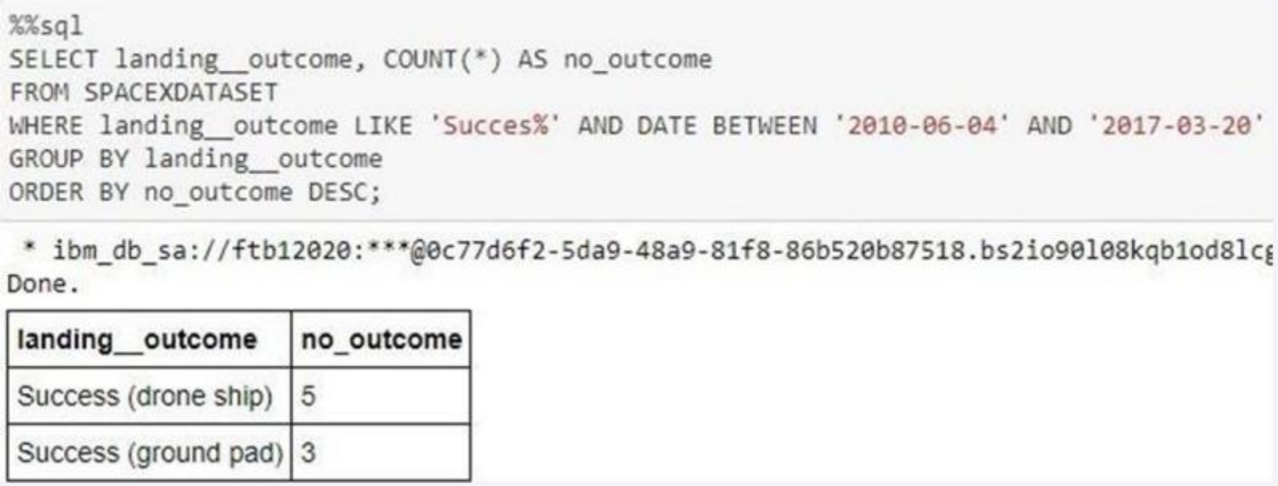
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively:

- 8 successful landings in total
- 5 landings on drone ship and 3 landings on ground pad.



```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

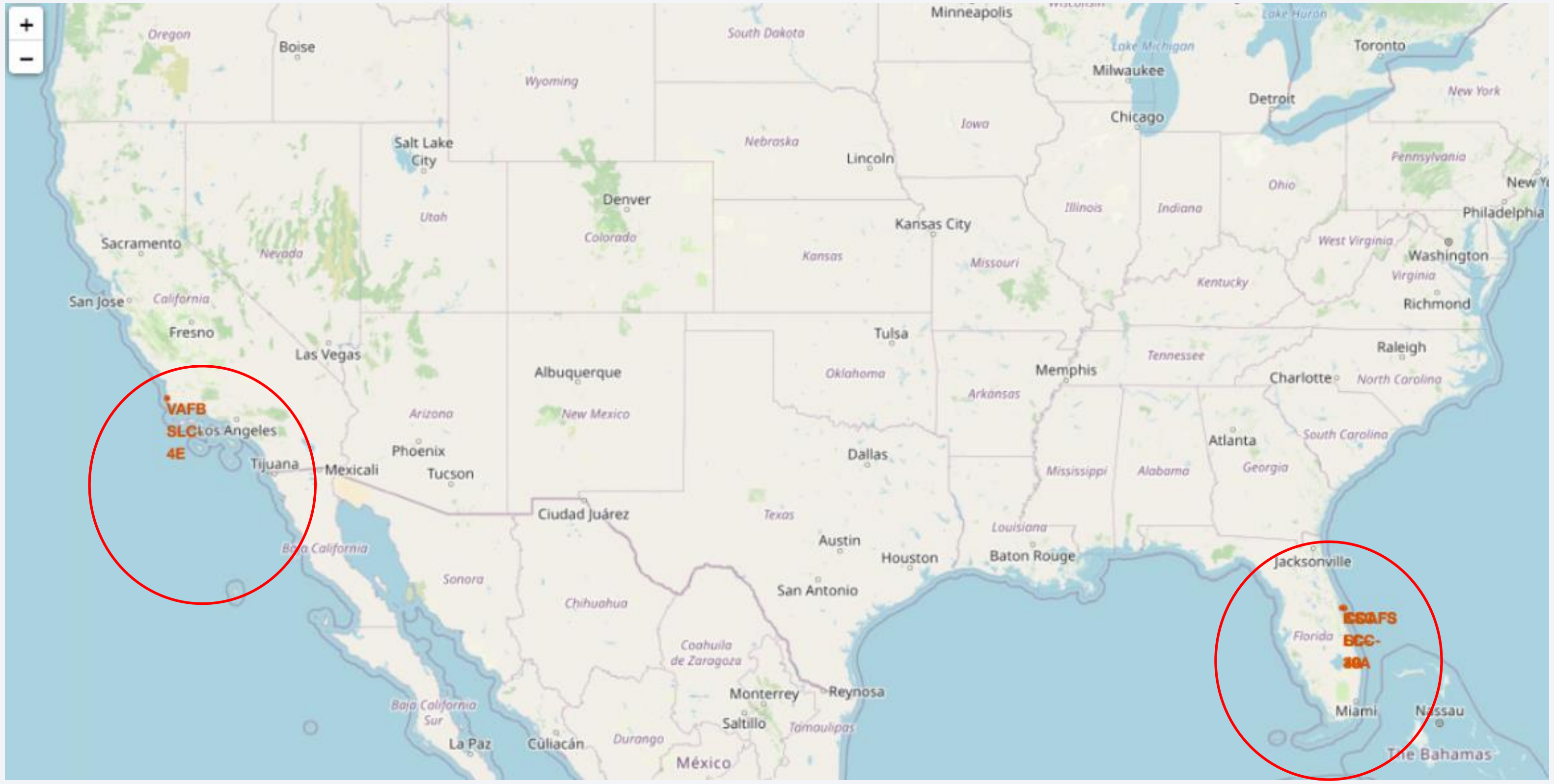
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

Launches geography

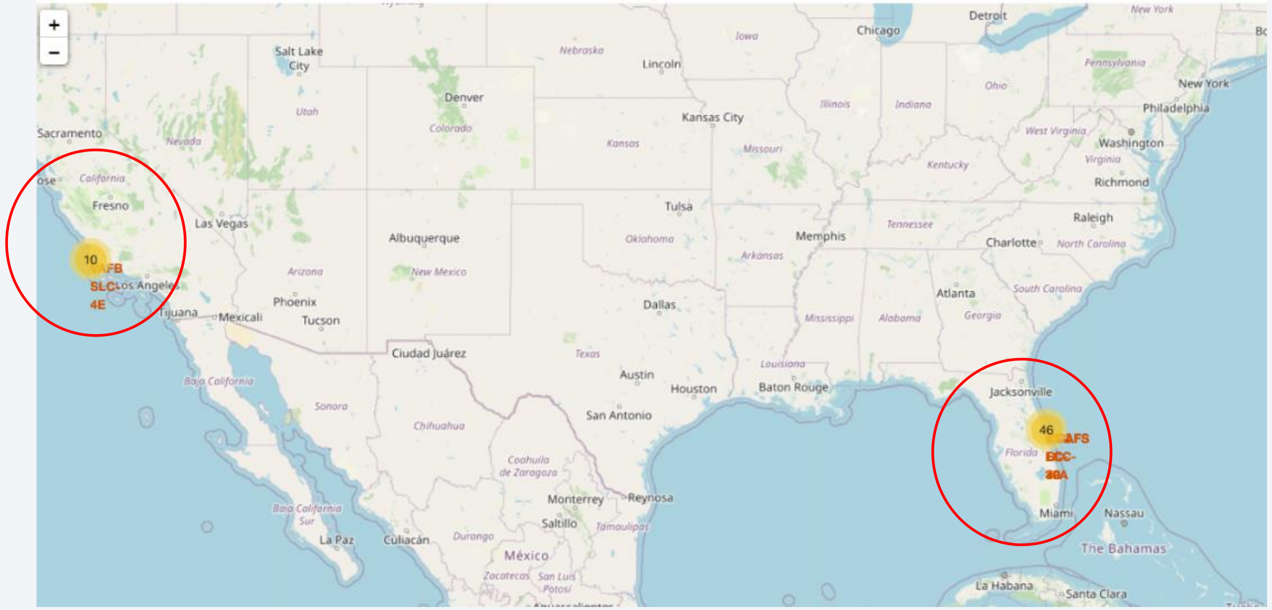
As shown on the map below, all launches took place near equator line and at coastal area, which makes the experiments comparable (similar impact of gravity).

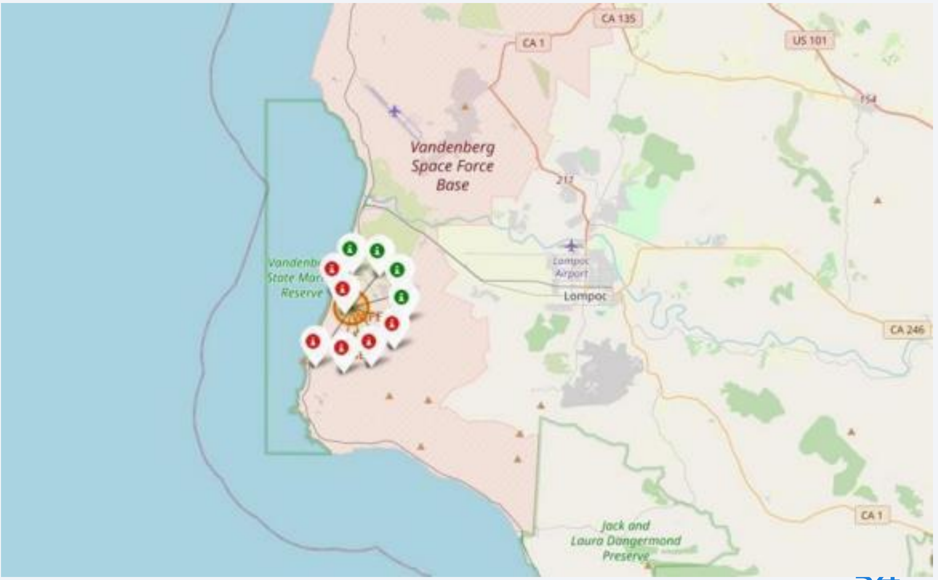


Launches geography and success

From the maps below we can see that majority of launches (46) took place from Florida, while 10 from West Coast. Importantly, all West Coast launches were VAFB.

Another map shows how successful (green) and unsuccessful (red) launches split for VAFB, we can see 4 successful and 6 unsuccessful launches.

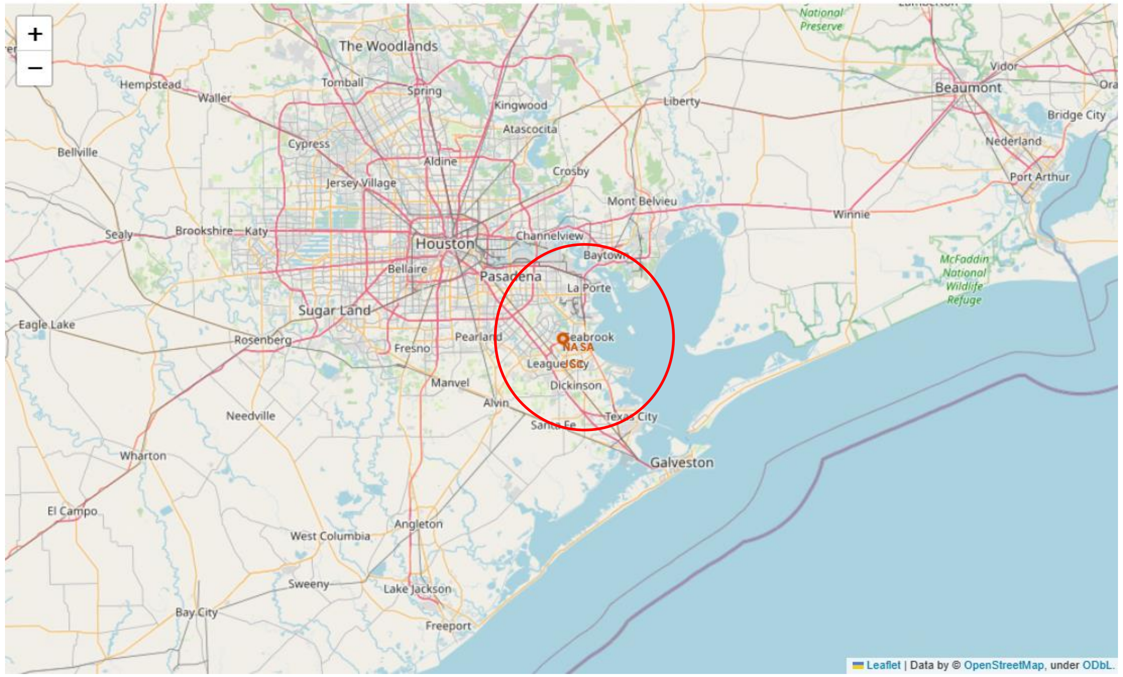


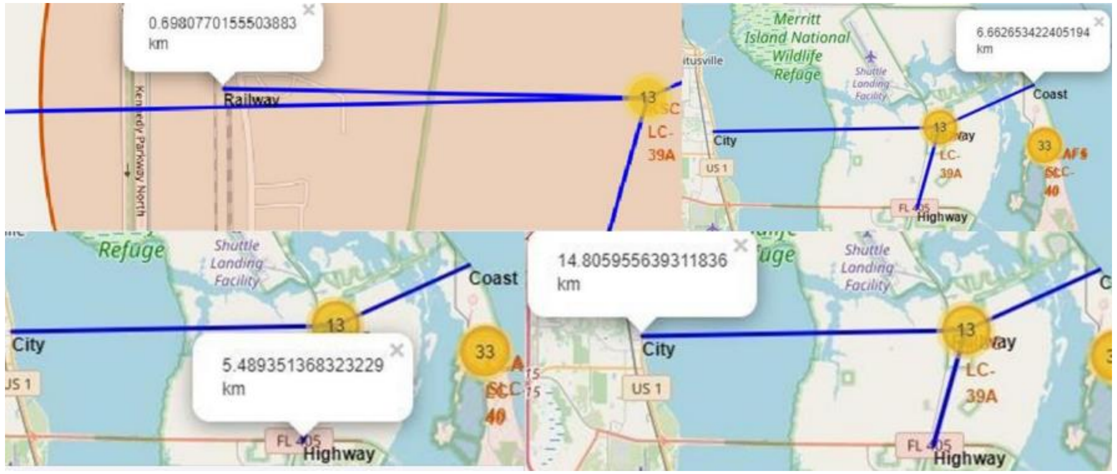


What is in proximity of launch place

Below maps demonstrate on example of Houston and Florida that launches take place near transport infrastructure (railway, highways), near coast but at reasonable distance from the city centers.

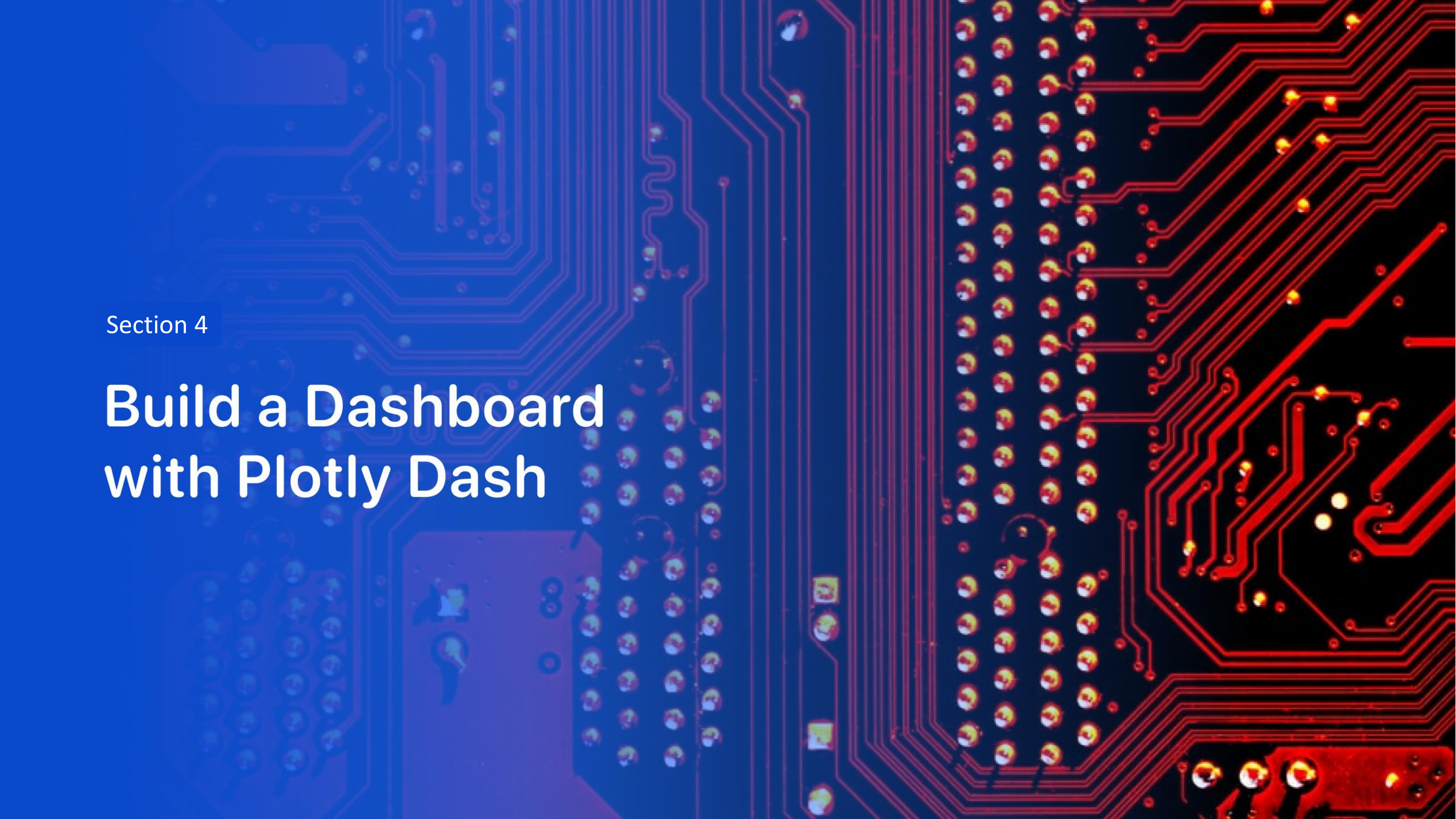
This is driven by the need to minimize the costs of fuel transport (infrastructure) and, in case of failure, enable as little harm to civilians and city infrastructure as possible.





Section 4

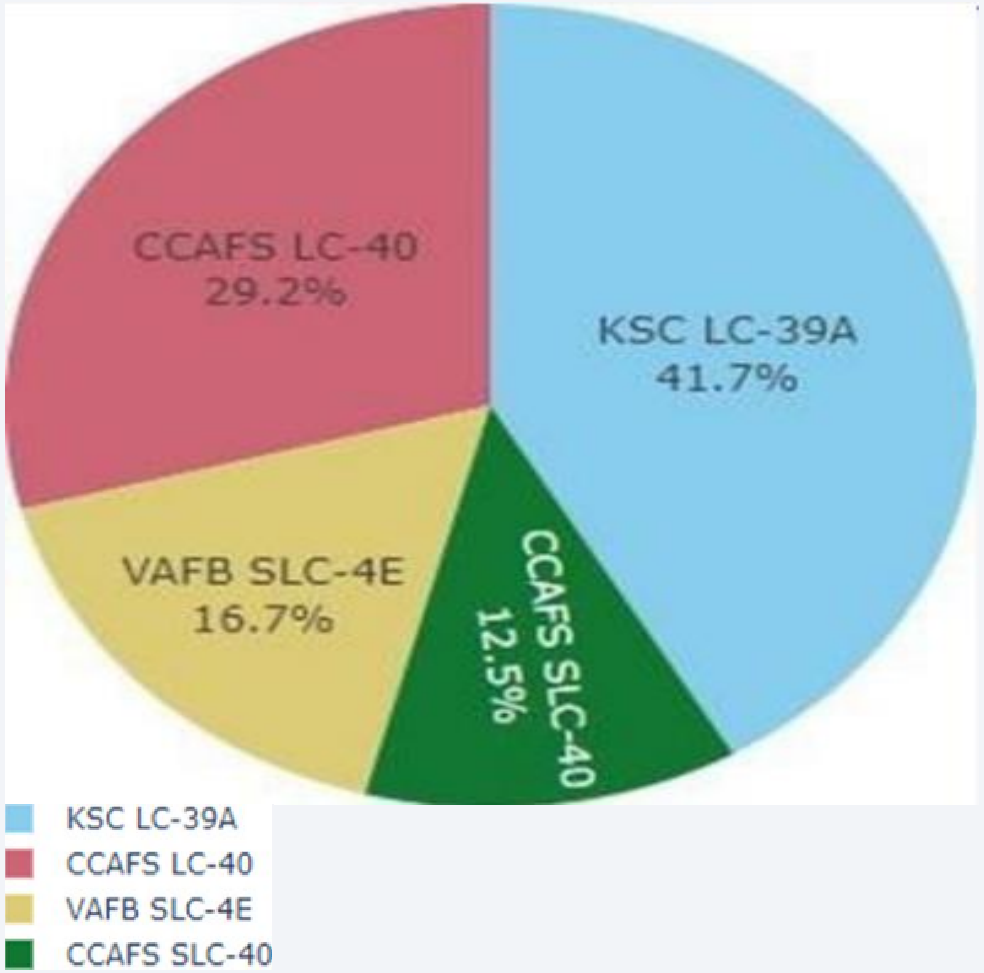
Build a Dashboard with Plotly Dash



Share of successful landings

We observed the following distribution of successful landings:

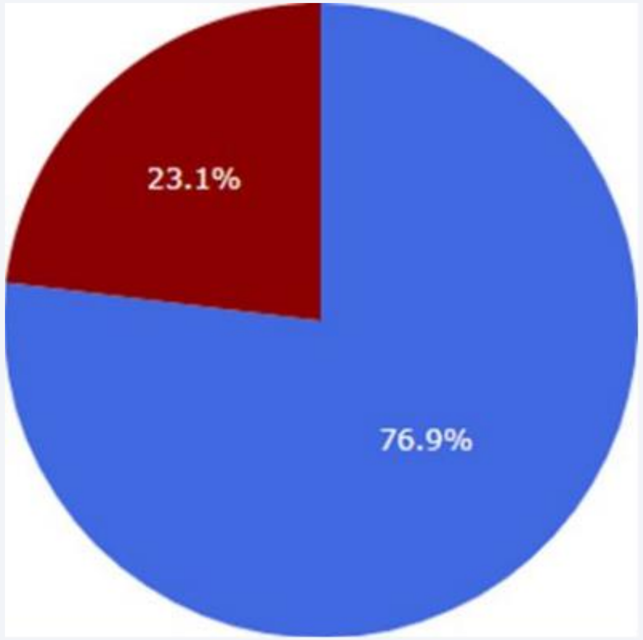
- KSC and CCAFS (LC-40 and SKC-40) have the same share of successful landings: 41,7%
- VAFB has the smallest share of successful landings: 16,7%. This should not be interpreted that VAFB has the *smallest chance* of successful landings but only that we should not make conclusions based on VAFB experiments.



KSC has the highest rate of successful landings

KSC LC-39A has the highest rate of successful landings with 10 successes (76,9%) and 3 fails (23,1%).

KSC LC-39A Success Rate (blue=success)

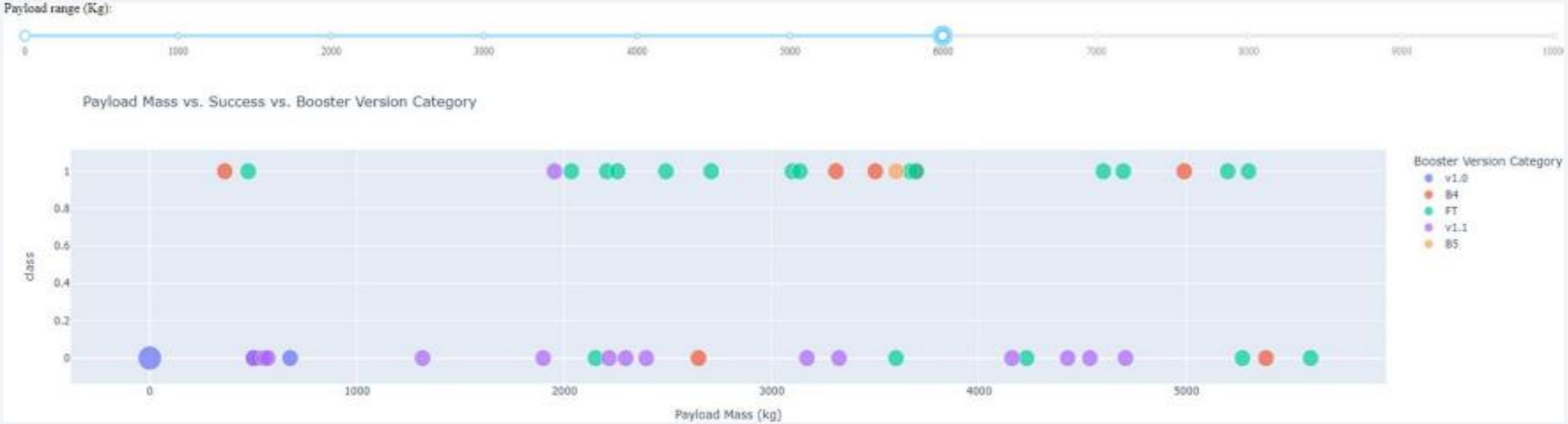


Booster version, not payload, is critical for success

The scatter chart below demonstrates launch success (vertical axis, 1 for success, 0 for failure) vs payload (horizontal axis) and takes into consideration booster version (color).

We can conclude that independently on payload, the boosters FT and B4 are most efficient, while v 1.1 is the lest efficient.

We can also conclude that booster version, not payload, plays critical role in launch success.



Section 5

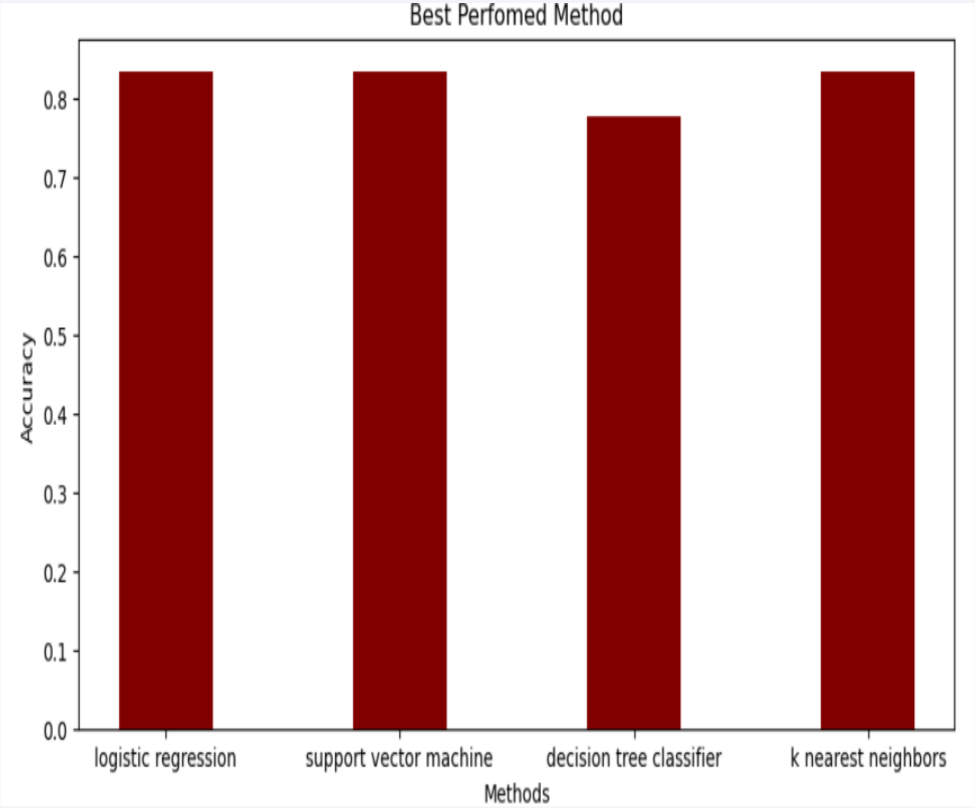
Predictive Analysis (Classification)

Classification Accuracy

Three out of four models have the same accuracy:

- 83% of Logistic regression, SVM and K nearest neighbors
- 78% for Decision tree classifier

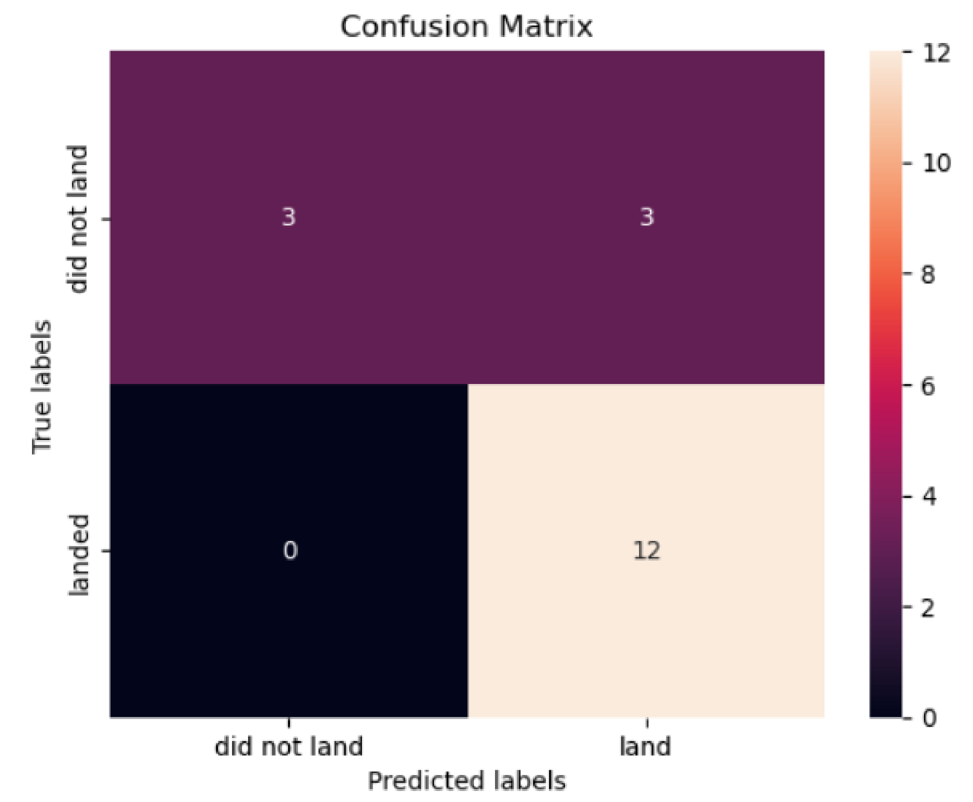
We should keep in mind that we are dealing with small sample size, which is driven by industry specifics (very high cost of trials).



Confusion Matrix

With 83% correct predictions, our models should be considered as accurate. However, we should keep in mind that they tend to over-predict successful landings:

- Predicted 12 successful landings when the true label was successful landing.
- Predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- Predicted 3 successful landings when the true label was unsuccessful landings (false positives).



Appendix

Special Thanks to All Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Github repository:

https://github.com/julialenc/IBM_Data-Science_Capstone-Project_SpaceX

Thank you!