IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Georgi Naumov
22/07/2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- Project background and context

SpaceX is a leading company that makes space travel more affordable. They offer Falcon 9 rocket launches at a competitive price of $62 million, which is significantly cheaper than other providers who charge upwards of $165 million per launch. One reason for this cost savings is that SpaceX is able to reuse the first stage of their rockets. By analyzing whether the first stage will land, we can estimate the overall cost of a launch. To make this prediction, we'll use publicly available data and machine learning models.

- Answers to problems we need to find

How the variables in the collected data affect the rocket launch result?

What is the trend of rocket launch results?

What algorithm to use for classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By using Space X REST API
  - By scraping Wikipedia
- Perform data wrangling
  - The data was filtered and cleaned up
  - Missing values were replaced
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were build, tuned, evaluated

Data Collection

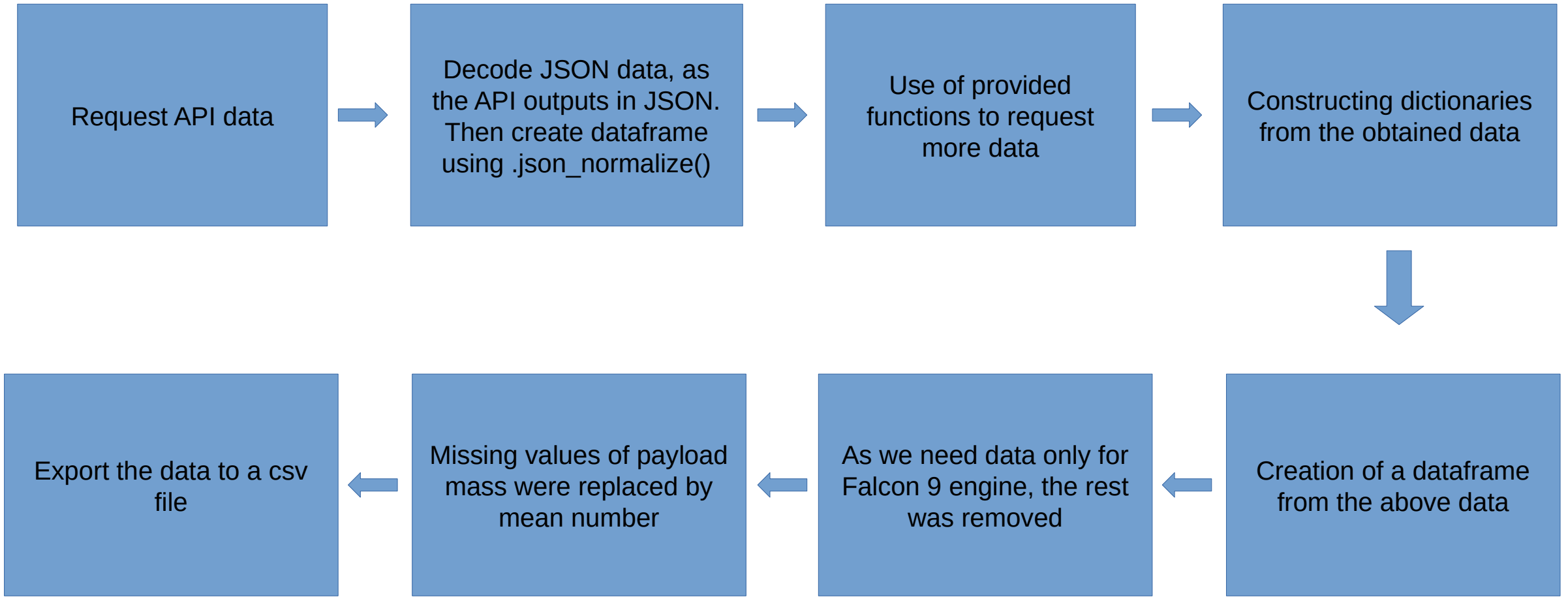
Data was collected from 2 separate source, in order to achieve completeness:

From the Space X REST API and from Wikipedia by web scraping.

Columns obtained from REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

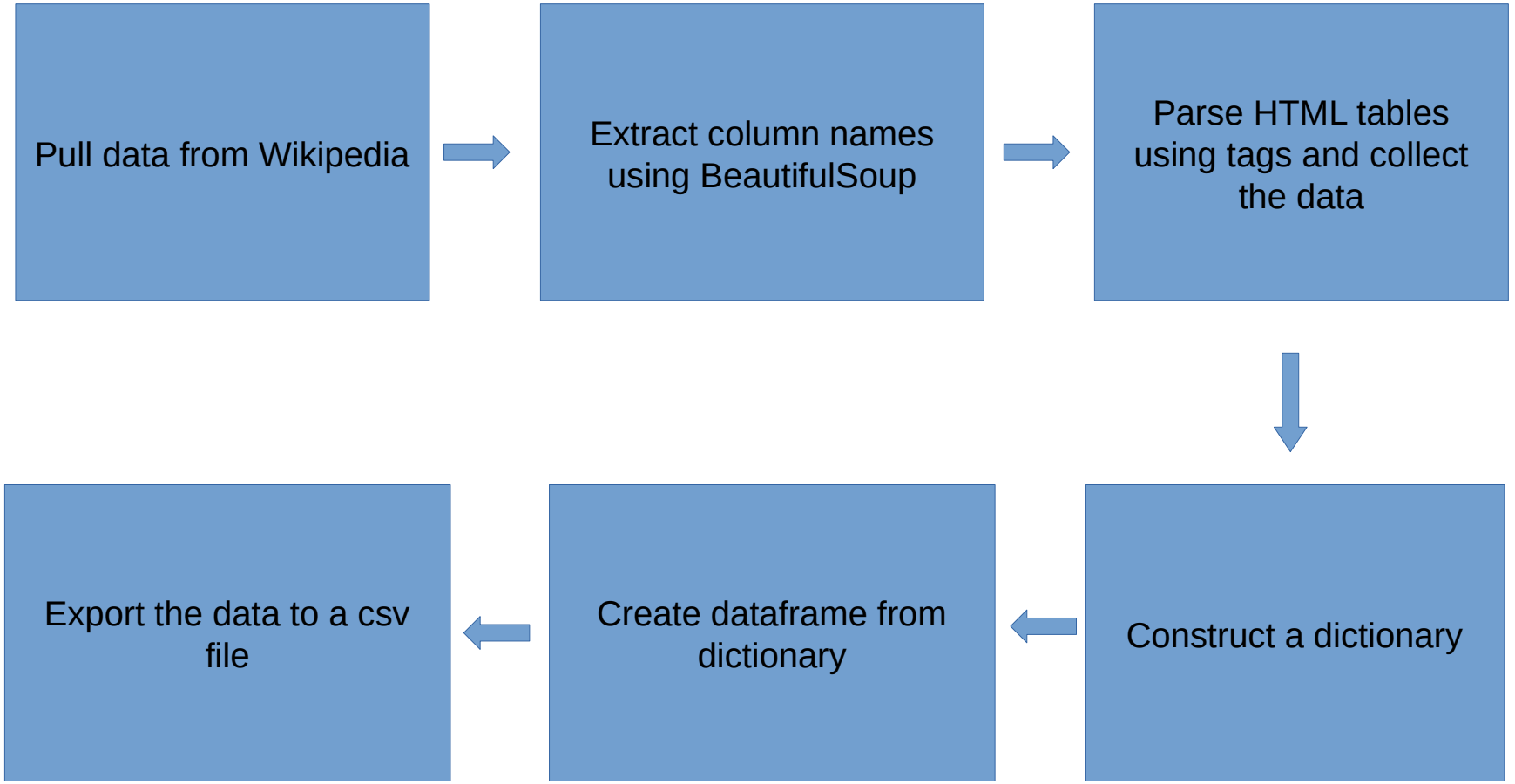
Columns obtained from Wikipedia: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection flowchart – Space X REST API



https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/01%20Data%20collection%20API.ipynb

Data collection flowchart – Wikipedia



https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/02%20Webscraping.ipynb

Data Wrangling

The following have to be calculated:

- Number of launches per site
- Orbit statistics
- Mission outcome per orbit type
- From the above, labels must be created
- Export the data to csv file

In order to proceed with the data analysis process, the data must be normalized. For all cases where the booster did not land successfully Below is an explanation of the relevant cases:

Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We have to convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/03%20Data%20wrangling.ipynb

EDA with Data Visualization

Chart were created in order to visualize and understand the data:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

Scatter plots show the relationship between variables.

Bar charts show comparisons among discrete categories.

Line charts show trends in data over time (time series).

https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/05%20Visualization.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count 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

https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/04%20SQL.ipynb

Build an Interactive Map with Folium

Geographic Visualization of Launch Sites

We've created markers to represent all launch sites, featuring:

- The NASA Johnson Space Center, marked with a circle, popup label, and text label, using its latitude and longitude coordinates as a starting point.
- Each launch site, marked with a circle, popup label, and text label, showcasing its geographical location and proximity to the Equator and coastlines.

Colored Markers for Launch Outcomes:

- Successes are represented by green markers, while failures are shown in red.
- Marker clustering helps identify launch sites with high success rates.

Distances between Launch Sites and Their Proximities:

We have added colored lines to illustrate the distances between launch site KSC LC-39A (used as an example) and its proximities, such as railways, highways, coastlines, and closest cities.

https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/06%20Folium.ipynb

Build a Dashboard with Plotly Dash

We have implemented several interactive features to enhance our visualization:

A dropdown list allows users to select specific launch sites for further analysis.

A pie chart provides a breakdown of successful launches across all sites, as well as for individual sites when selected.

A slider enables users to filter payload mass ranges.

A scatter chart illustrates the correlation between payload mass and success rate for different booster versions.

https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/dashboard.py

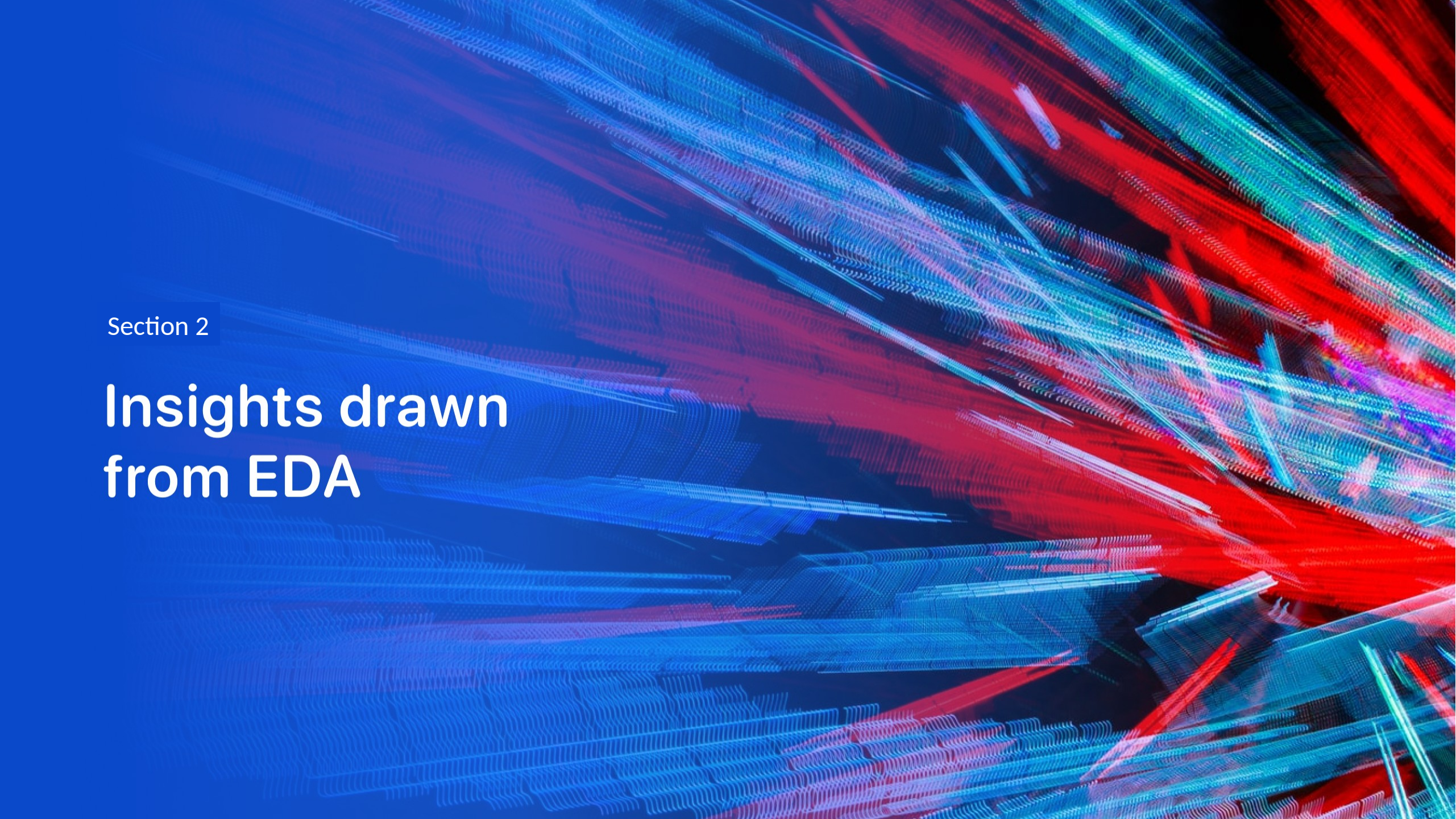
Predictive Analysis (Classification)

1. Create a numpy array from “Class” column
2. Standarize data with StandatrdScaler
3. Fit and transform the data
4. Split the data into 2 sets: training and testing
5. Find the best parameters with GridSearchCV
6. Apply the results of GridSearchCV to LogReg, SVM, Decision tree and KNN models
7. Calculate the accuracy of the above
8. Examine the confusion matrix
9. Use Jaccard _score and F1_Score to find how the models performed

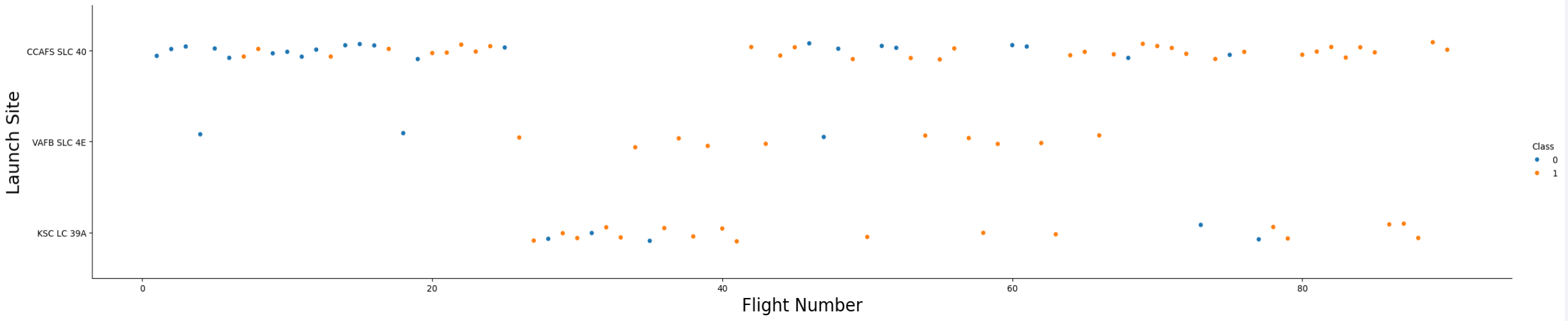
https://github.com/geonaumov/IBM_Applied_data_science_capstone/blob/main/08%20Machine%20learning.ipynb

Results

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

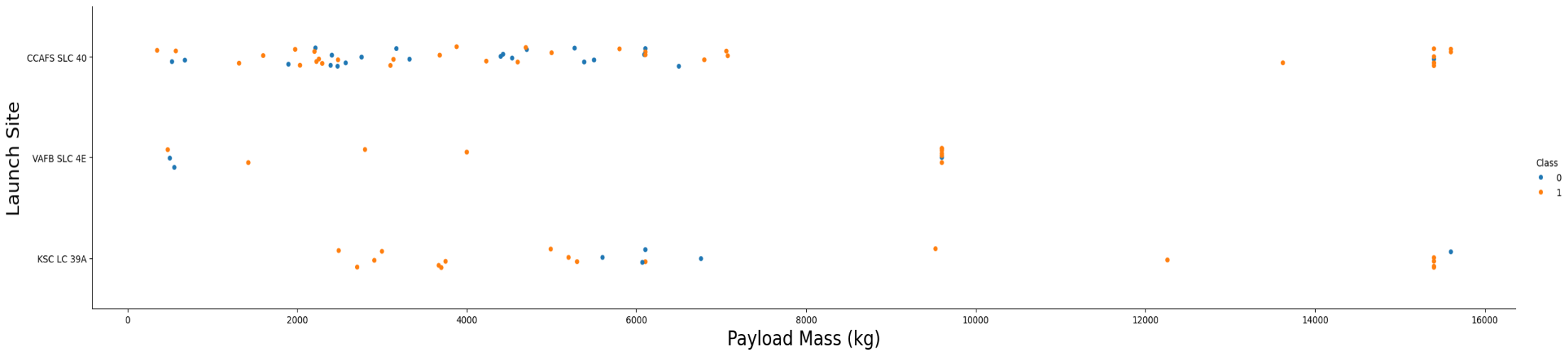

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



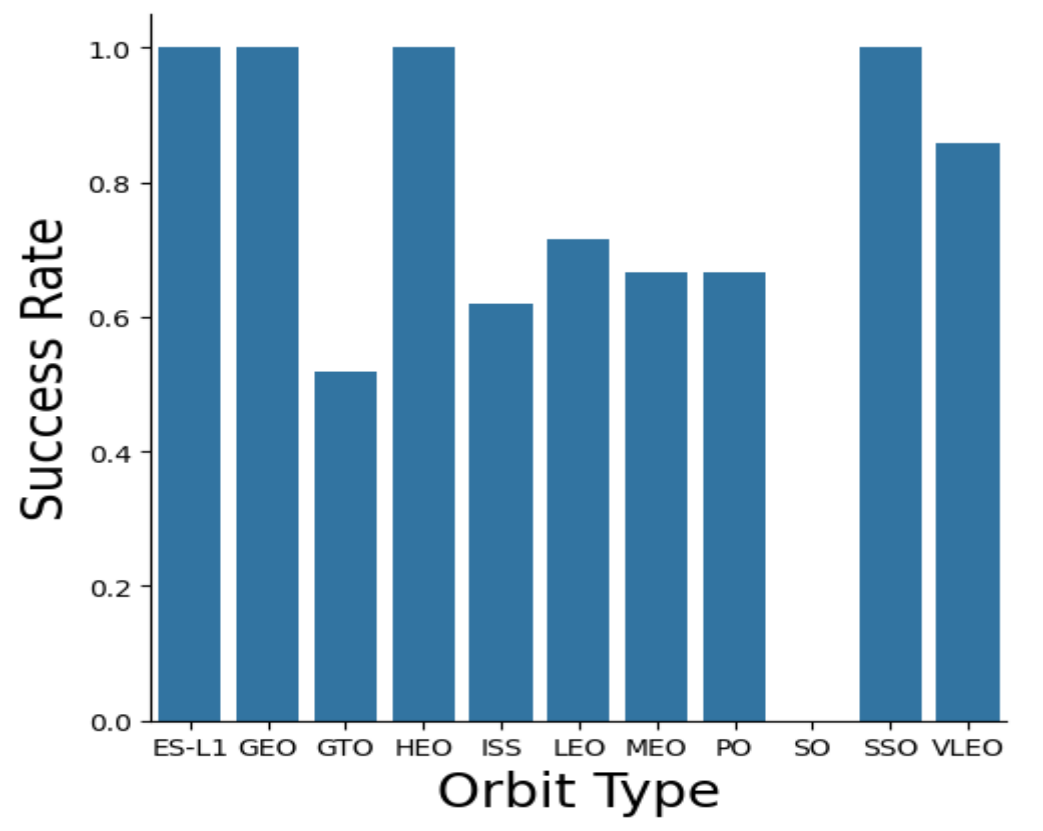
- There is a positive trend of success
- Most flights are from CCAFS SLC 40
- The rest of the launch sites have higher success rate compared to CCAFS.

Payload vs. Launch Site



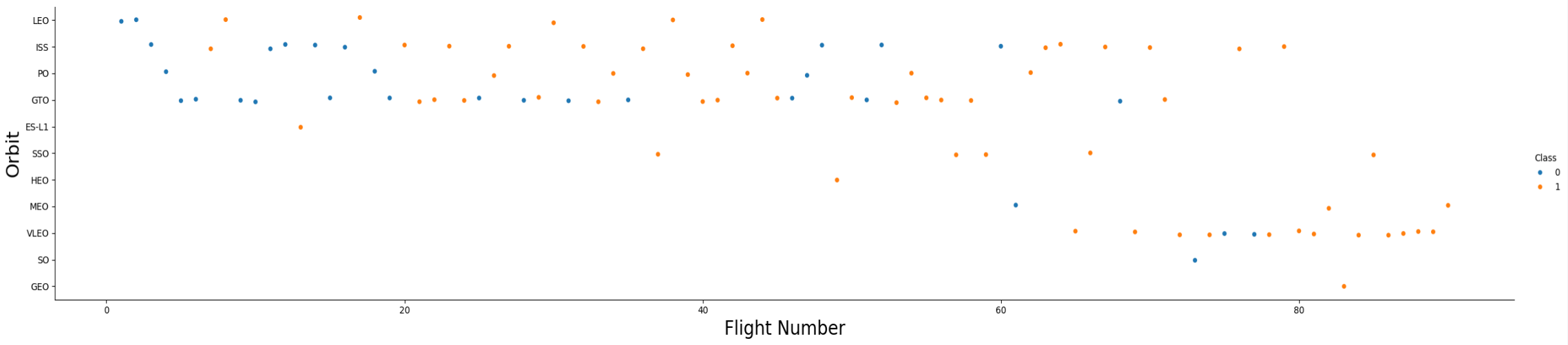
- Large payloads have higher success rate, however this is probably due to correct risk management by Space X
- Lower payload launches are probably experimental and have varying results

Success Rate vs. Orbit Type



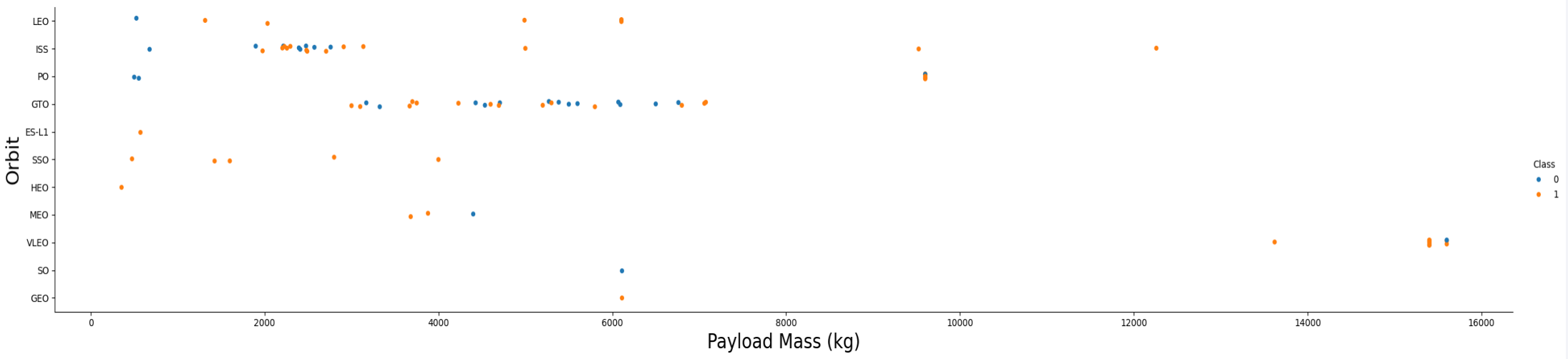
Orbit types have varying results, however orbit SO has 0% success rate!

Flight Number vs. Orbit Type



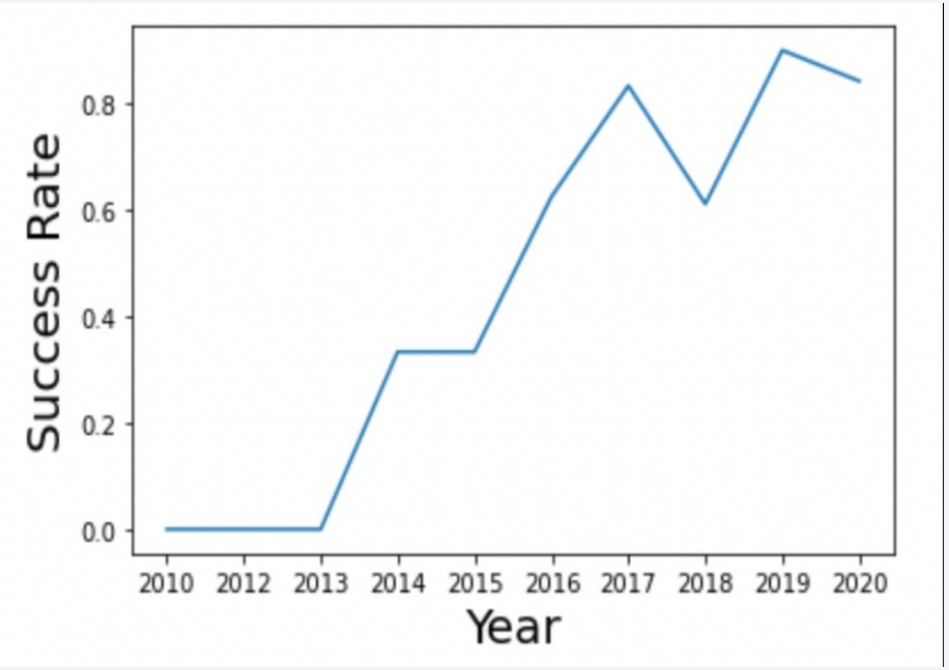
- LEO orbit has become successful as flight numbers increase
- SO has one flight, so it’s zero success rate is not that bad
- Newer sites are generally more successful as time passes.

Payload vs. Orbit Type



- ISS orbit is successful with high payloads
- SSO is successful with low payloads

Launch Success Yearly Trend



Success rate is increasing up to 2020, with small fluctuation in 2018

All Launch Site Names

```
[18] 1 %sql select distinct "Launch_Site" from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

SpaceX has 4 launch sites

Launch Site Names Begin with 'CCA'

```
[11] 1 %sql select * from SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

Executed at 2024.07.24 18:40:17 in 7ms

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_(
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

```
[14] 1 %sql select sum("PAYLOAD_MASS__KG_") as "payload_mass_kg" from SPACEXTABLE where customer = 'NASA (CRS)';
Executed at 2024.07.24 18:42:23 in 6ms
 * sqlite:///my_data1.db
Done.
```

payload_mass_kg
45596

Total payload mass in kilograms for client NASA (CRS)

Average Payload Mass by F9 v1.1

```
[15] 1 %sql select avg("PAYLOAD_MASS__KG_") as avg_payload_F9 from SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%';
Executed at 2024.07.24 18:44:20 in 7ms
 * sqlite:///my_data1.db
Done.
```

avg_payload_F9
2534.6666666666665

Average payload mass in kilograms for booster F9 V1.1

First Successful Ground Landing Date

```
[21] 1 %sql select min("Date") as "first_succesful_landing" from SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)';
     2
```

Executed at 2024.07.24 18:51:44 in 5ms

```
 * sqlite:///my_data1.db
Done.
```

first_succesful_landing
2015-12-22

OR

```
[20] 1 %sql select "Date" as "first_succesful_landing" from SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)' ORDER
       BY "Date" LIMIT 1;
```

Executed at 2024.07.24 18:51:43 in 7ms

```
 * sqlite:///my_data1.db
Done.
```

first_succesful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[22] 1 %sql select booster_version from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and payload_mass__kg_
   between 4000 and 6000;
Executed at 2024.07.24 18:55:35 in 9ms
```

```
 * sqlite:///my_data1.db
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

```
[23] 1 %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

Executed at 2024.07.24 18:55:41 in 8ms

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

```
[24] 1 %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

Executed at 2024.07.24 18:59:32 in 5ms

```
 * sqlite:///my_data1.db
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

2015 Launch Records

```
[28] 1 %%sql
     2 select  substr(Date, 6,2) as month, date, booster_version, launch_site, Landing_Outcome from SPACEXTABLE
     3 where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

Executed at 2024.07.24 19:20:16 in 4ms

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
[29] 1  %%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
     2  where date between '2010-06-04' and '2017-03-20'
     3  group by Landing_Outcome
     4  order by count_outcomes desc;
```

Executed at 2024.07.24 19:20:17 in 5ms

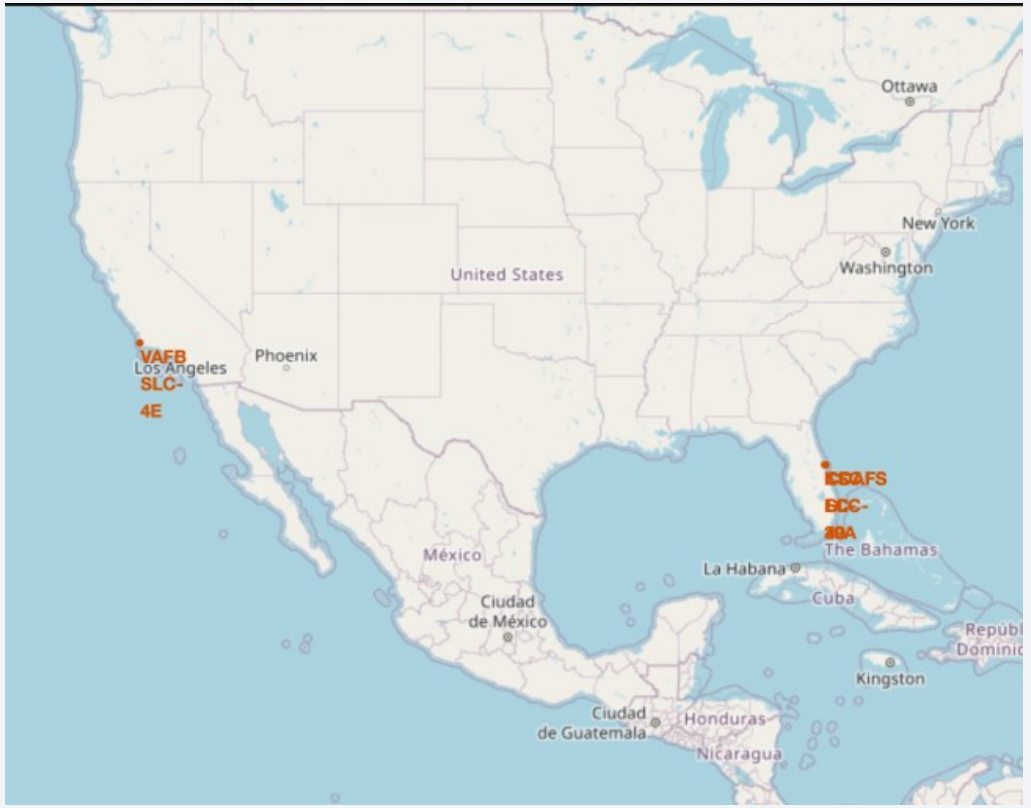
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
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

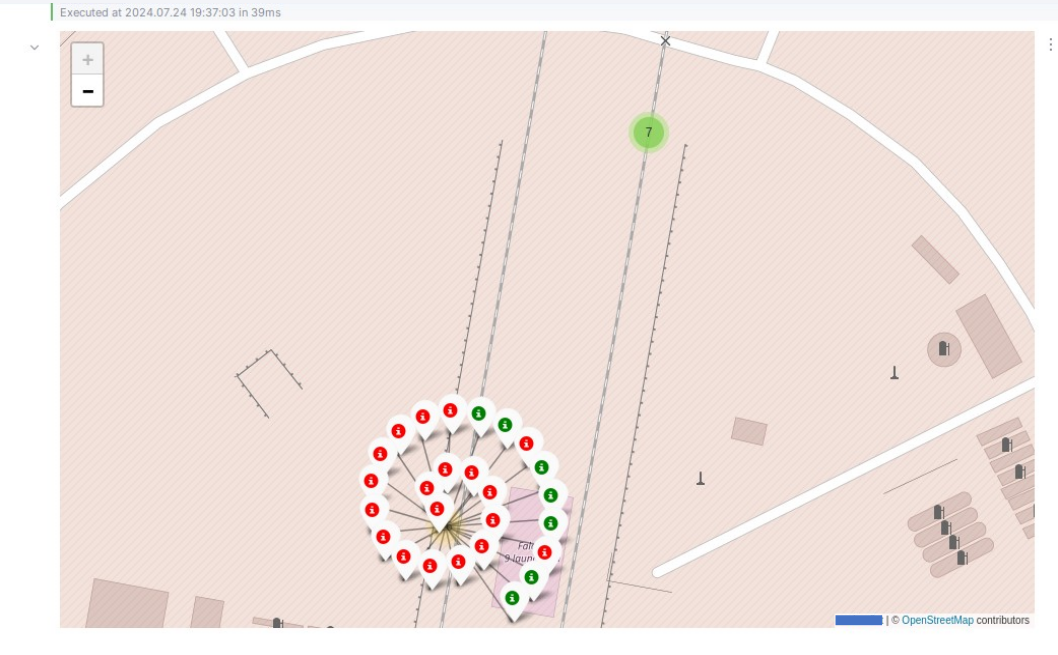
Section 3

Launch Sites Proximities Analysis

All launch sites on a map

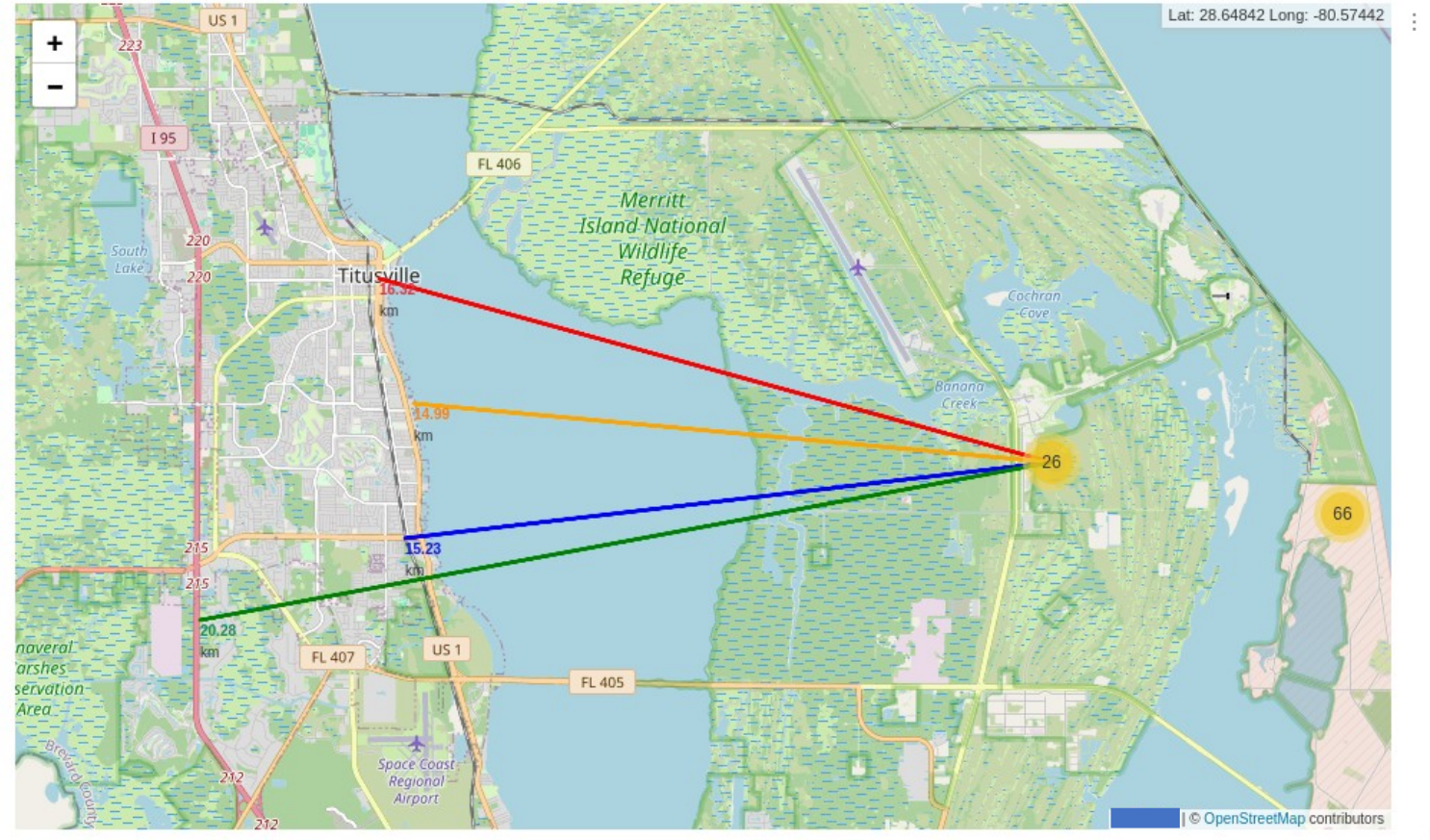


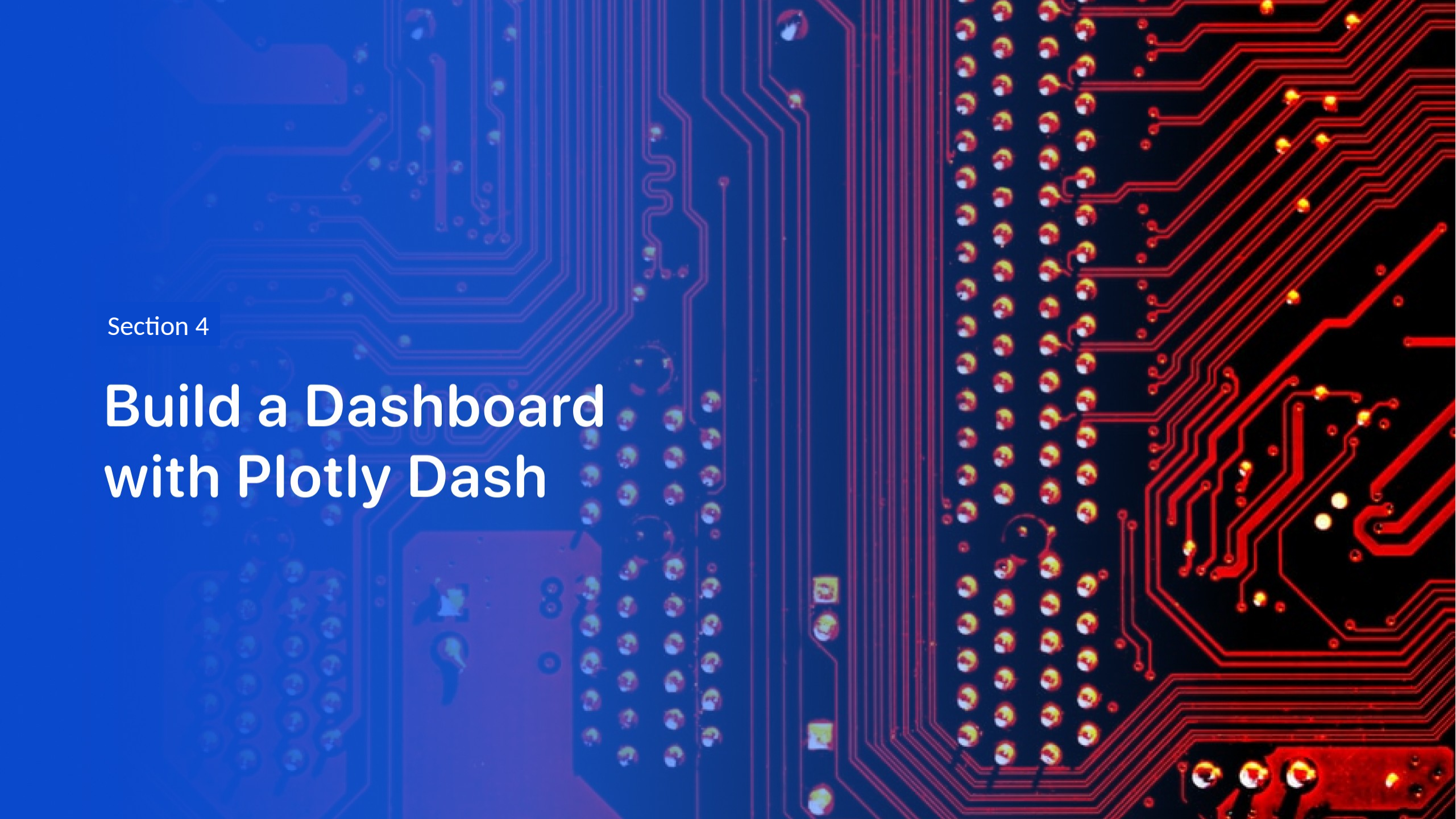
Colour-labeled launch markers on the map



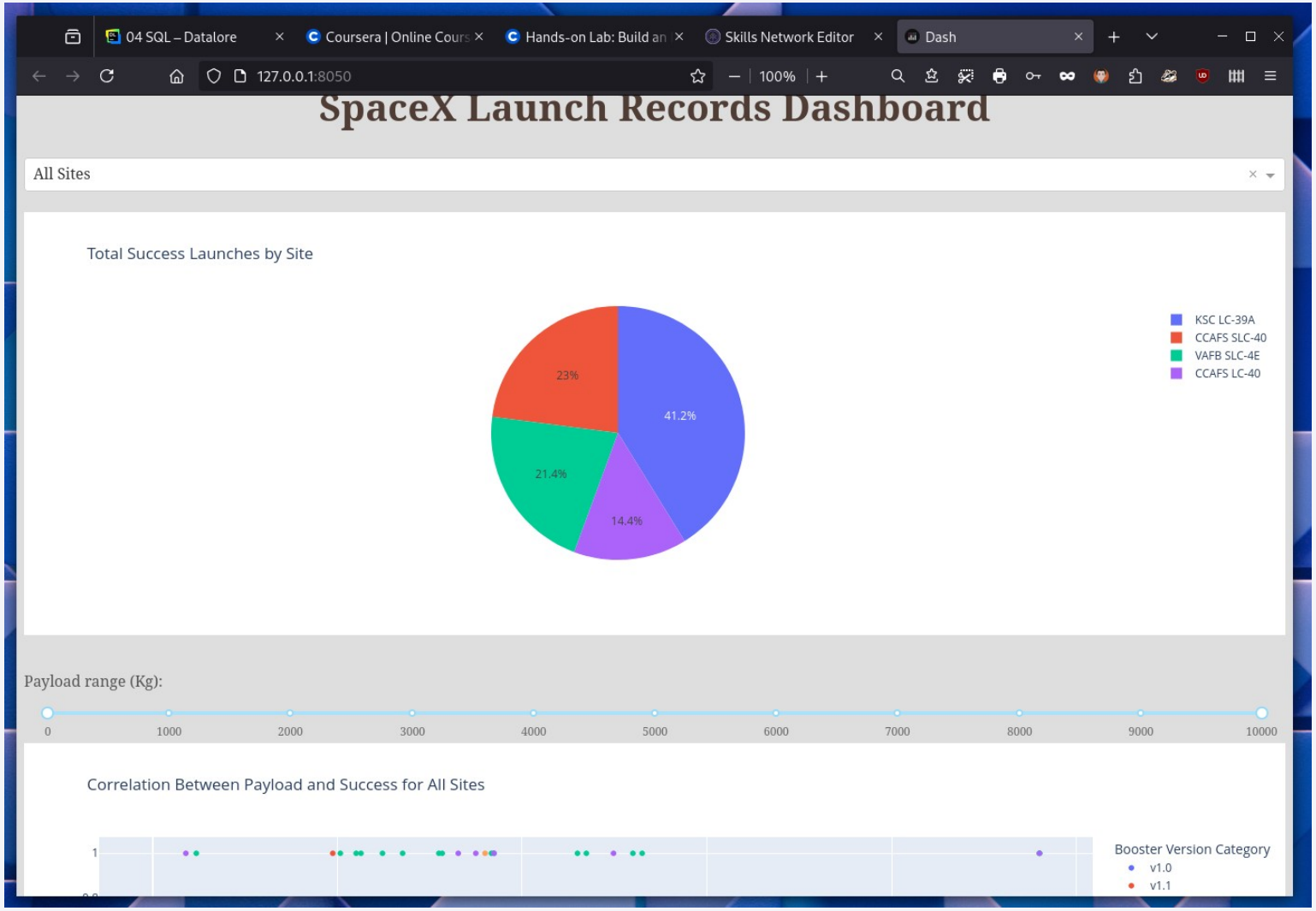
Green markers represent successful launches
Red ones represent failed ones

Proximities of launch site KSC LC-39A

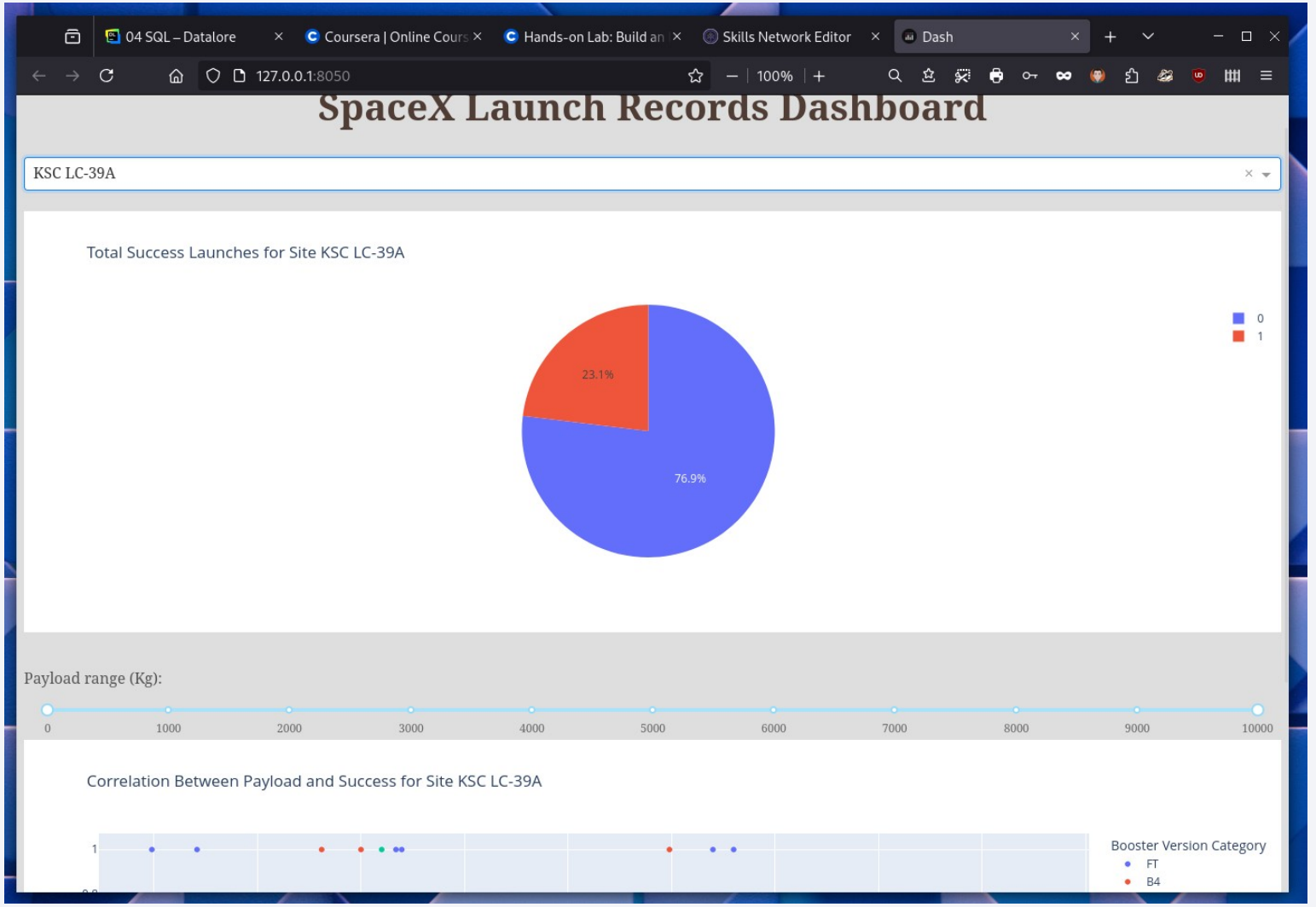



Section 4
Build a Dashboard with Plotly Dash

Launch success per launch site



KSC LC-39A


04 SQL – Datalore
Coursera | Online Cours
Hands-on Lab: Build an
Skills Network Editor
Dash
127.0.0.1:8050
100%
SpaceX Launch Records Dashboard
KSC LC-39A
Total Success Launches for Site KSC LC-39A
23.1%
76.9%
0
1
Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Correlation Between Payload and Success for Site KSC LC-39A
1
Booster Version Category
FT
B4

Payload mass success rate

Payloads between 5000 and 7000 kg are the most successful

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The test sample is very small
The most appropriate model is the Decision Tree

[32]

```
from sklearn.metrics import jaccard_score, f1_score

accuracy = [logreg_accuracy, svm_accuracy, tree_accuracy, knn_accuracy]

scores = pd.DataFrame(np.array([accuracy]), index=['Accuracy'] , columns=['LogReg', 'SVM', 'Tree', 'KNN'])
scores
```

Table Raw Visualize Statistics

	LogReg	SVM	Tree	KNN
Acc…	0.8333333333333334	0.8333333333333334	0.8333333333333334	0.8333333333333334

1 row x 4 columns Jump to top Jump to bottom

[33]

```
accuracy = [logreg_cv.score(X, Y), svm_cv.score(X, Y), tree_cv.score(X, Y), knn_cv.score(X, Y)]

scores = pd.DataFrame(np.array([accuracy]),
                      index=['Accuracy'],
                      columns=['LogReg', 'SVM', 'Tree', 'KNN'])
scores
```

Table Raw Visualize Statistics

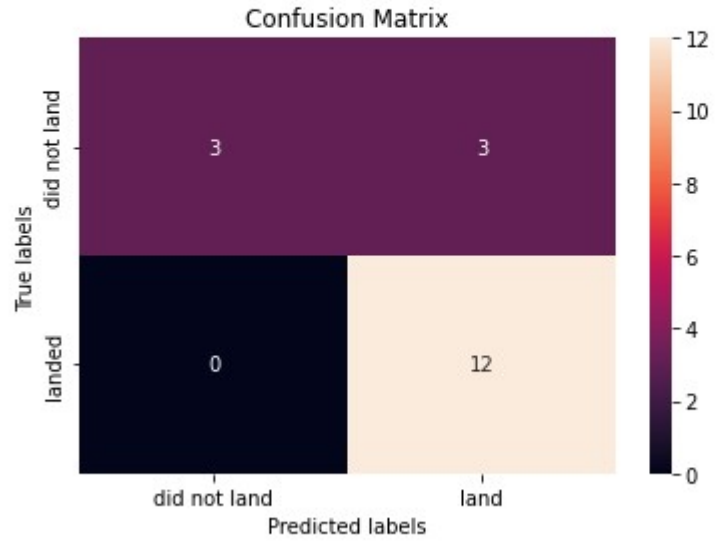
	LogReg	SVM	Tree	KNN
Acc…	0.8666666666666667	0.8777777777777778	0.8777777777777778	0.8555555555555555

1 row x 4 columns Jump to top Jump to bottom

Confusion Matrix

[26]

```
1 yhat = tree_cv.predict(X_test)
2 plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Decision Tree Model is the top choice for predicting launch outcomes.
- Low-payload launches tend to perform better than those with heavier payloads.
  Most launch sites are located near the Equator and along the coast.
  Launch success rates have been increasing over the years.
  Kennedy Space Center LC-39A has the highest success rate among all launch sites.
  Orbits ES-L1, GEO, HEO, and SSO have achieved very good success rates.

Thank you!