IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stefan Schmidt
05.01.2022

Outline

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

Executive Summary

Rocket launches are expensive, but costs can be reduced if a rocket's first stage can be reused. The projects' goal is to predict the landing outcome SpaceX' Falcon 9 rockets and therefore improve the estimation of a mission's costs.

- Methodologies
  - Data Collection was performed using the publicly available SpaceX API and web scraping Wikipedia's Falcon 9 launch page.
  - Data Wrangling and Exploratory Data Analysis: Data was cleaned, analyzed and visualized using various Python libraries, Folium, Plotly Dash and SQL.
  - During Predictive Analysis four different classification models were created and compared regarding their performance.
- Results
  - The selected classification models showed that the landing outcome and possible reuse of a rocket's first stage can be predicted with 83% accuracy.

Introduction

- Background
  - 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 due to the fact that SpaceX can reuse the first stage.
- Problems to find answers:
  - Can the landing outcome of a rocket's first stage be modelled and predicted? And with which accuracy?
  - Which variables (payload, orbit, launch site, booster version) have the highest impact on a missions landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the publicly available SpaceX API and web scraping Wikipedia's Falcon 9 launch site using Python library BeautifulSoup.
- Perform data wrangling
  - Data was processed and missing values were handled.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different classification models were built, tuned and evaluated.

Data Collection

- SpaceX API
  - Basic rocket launch data was collected from SpaceX launches API
  - IDs replaced with more detailed data from specific SpaceX APIs (booster version, launch site, payload, core)
  - Falcon 9 data was filtered and missing values replaced
- Web Scraping with Python library BeautifulSoup
  - Launch tables were identified
  - Launch data extracted and stored as csv file

Data Collection – SpaceX API

- Python libraries were used to request the public SpaceX API to retrieve lauch, rocket, launchpad payload and landing data.
- The dataset was filtered, missing values replaced and transformed into a data frame for further processing.
- Details of the data collection process are available at the hands-on lab published on GitHub

retrieve basic launch data

↓

replace IDs with detailed data from specific SpaceX APIs

↓

filter Falcon 9 data only

↓

replace missing payload values from classified missions with mean value

↓

create Pandas data frame

Data Collection - Scraping

- Python libraries requests and BeautifulSoup library were used to retrieve and parse the Falcon 9 launches Wikipedia page.
- Relevant tables were filtered, columns identified, data extracted and stored to a Pandas data frame.
- Details of the web scraping process are available at the hands-on lab published on GitHub

retrieve raw HTML data

↓

identify HTML tables to parse

↓

extract column names from HTML table headers

↓

retrieve HTML table data

↓

Store data in Pandas data frame

Data Wrangling

- Falcon 9 landing outcomes were identified and classified as either success or failure. An additional binary column (based on this categorisation) was added to the data frame for further processing.
- Details of the data wrangling process are available at the hands-on lab published on GitHub

identify all possible landing outcomes

↓

categorize landing outcomes in either ‘Success’ or ‘Failure’

↓

create an additional column ‘Class’ based on landing outcome

(‘Success’=1 / ‘Failure’=0)

↓

add new binary column ‘Class’ was to the data frame

EDA with Data Visualization

- Scatter plots (including 3rd variable class for success/failure) were used to visualize dependencies between:
  - Payload / launch site / orbit versus flight number: to visualize changes over time.
  - Launch site / orbit vs. Payload: to show additional dependencies.
- A bar chart revealed that landing success rates differ depending on the targeted orbit.
- A line chart visualized the increasing success rate over time.
- Details of the exploratory data analysis (EDA) are available at the hands-on lab published on GitHub

EDA with SQL

- Launch data was loaded into a sqlite database and SQL queries performed to:
    1. Identify four distinct launch sites.
    2. Retrieve five records with launch site starting with CCA.
    3. Calculate the total payload carried for customer NASA (CRS).
    4. Compute the average payload carried by booster F9 v1.1.
    5. Find the date of the first successful landing on ground.
    6. Identify boosters that successfully landed on drone ships.
    7. Sum up the total number of successful and failed missions.
    8. Find the boosters that carried maximum payload.
    9. Identify the 2015 missions that failed landing on drone ship.
    10. Rank the count of successful landing outcomes for a time span in descending order.
- Details are available at the hands-on lab published on GitHub

Build an Interactive Map with Folium

- Three interactive maps were created using Python library Folium to:
  - Mark the four launch sites (circle and marker) using their geospatial data and Folium Markers and Circles.
  - Visualize successful and failed launches for every launch site using MarkerClusters holding Markers in different colors (depending on landing outcome).
  - Calculate the distance between launch sites and surrounding points of interest (closest city, railway, highway) using Markers and PolyLines.
- Further details are available at the hands-on lab published on GitHub

Build a Dashboard with Plotly Dash

- A dashboard app has been created using Plotly Dash to provide further insights to end users with no programming skills.
- The dashboard app shows two figures:
  - A pie chart showing either the distribution of all launch sites, or the proportion of successful launches for a selected site.
  - A scatter chart showing payload versus landing outcome with color coded booster version for either the selected or all launch sites. The payload of interest (upper/lower limit) can be selected using a slider.
- Further details are available at the here on GitHub

Predictive Analysis (Classification)

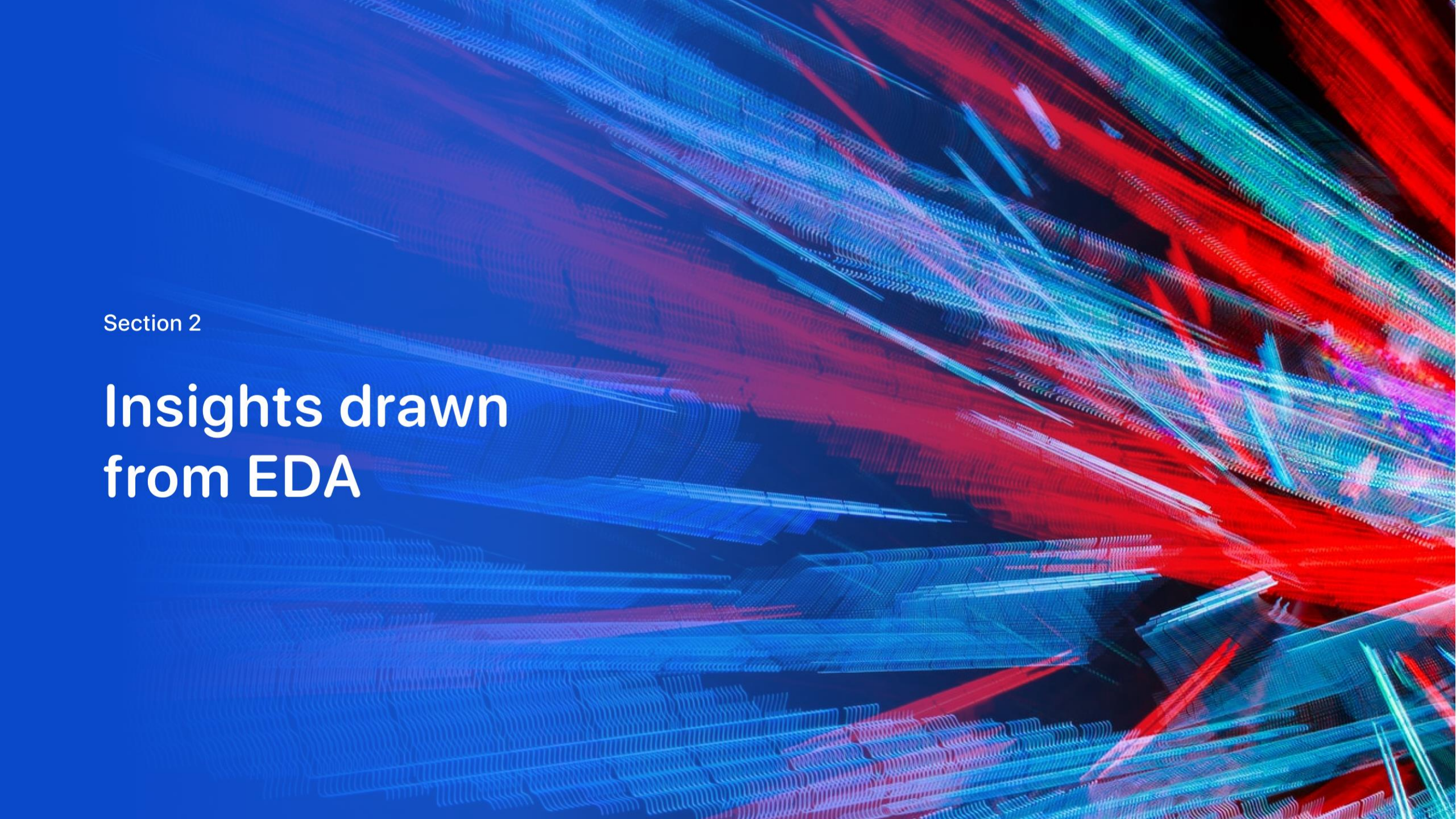
- The dataset was standardized and split in training and test sets.
- Four different classification models were created, trained and best parameters identified. Then they were tested for their accuracy.
- All four models (logistic regression, support vector machine, KNN, decision tree) showed similar performance on test data.
- Details of the predictive analysis are available at the hands-on lab published on GitHub

Results

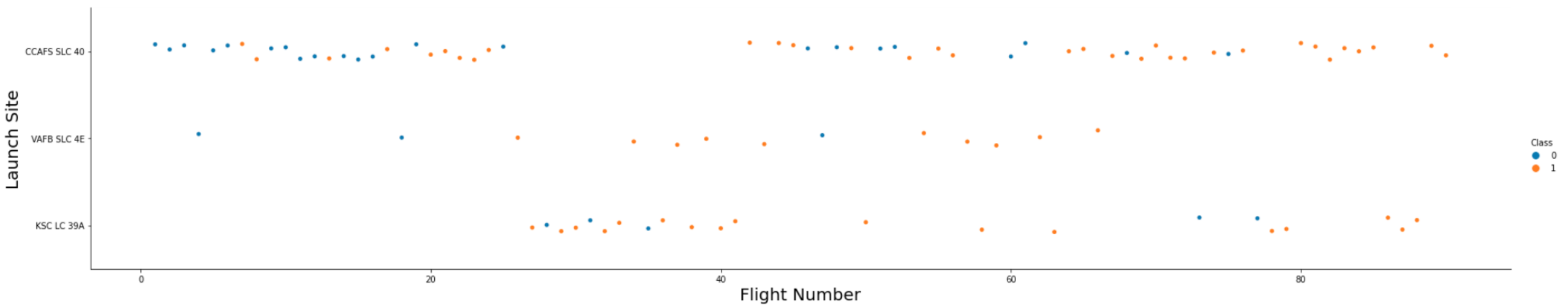
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

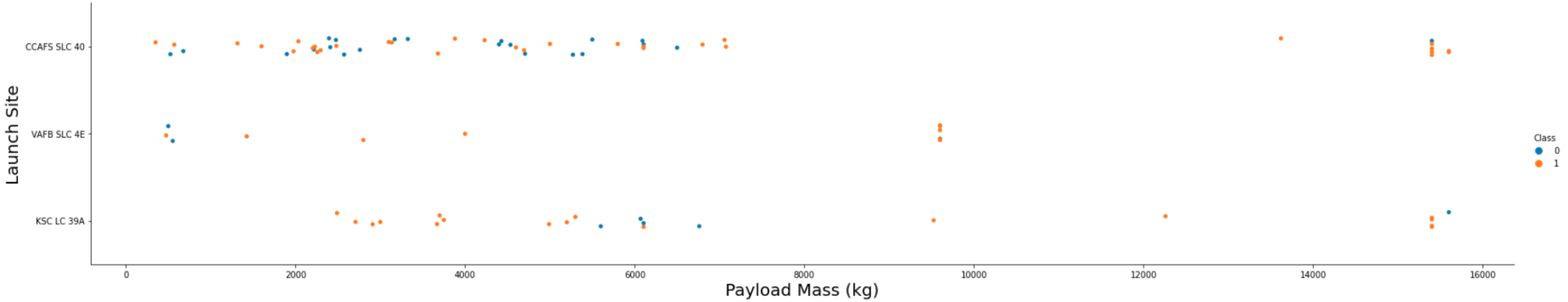


Flight Number vs. Launch Site



- The sequential flight number reveals that CCAFS SLC 40 was the first launch site put into service, VAFB SLC 4E followed, last one was KSC LC 39A.
- It is also visible that the success rate increased over time (successful landing=orange).

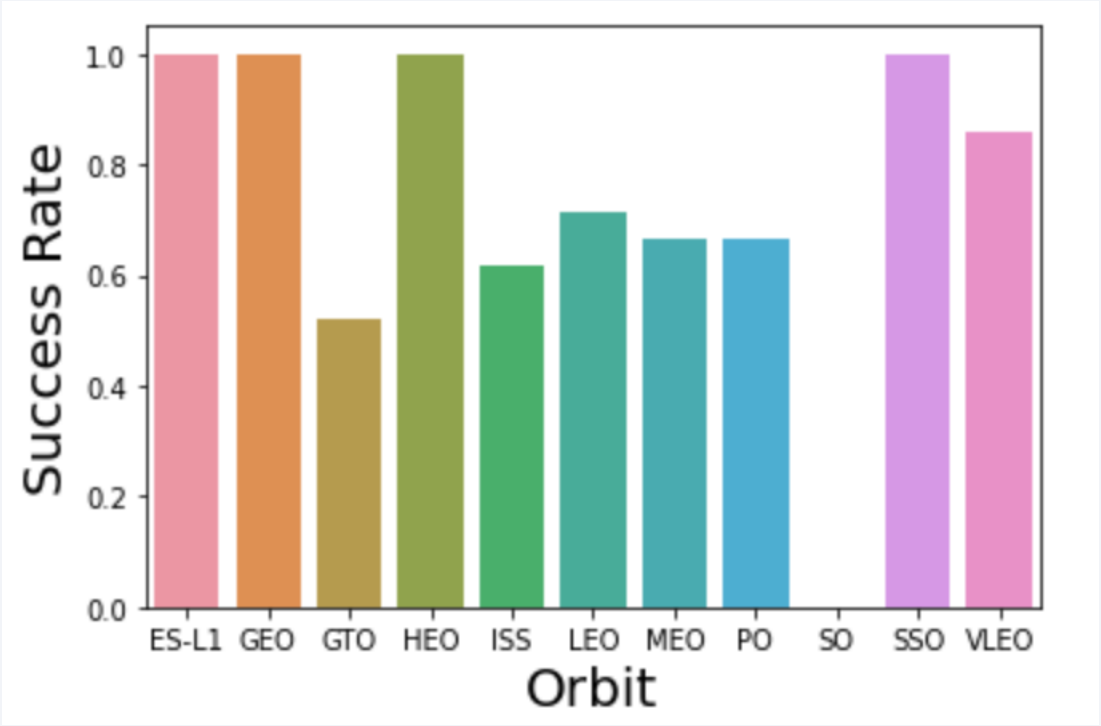
Payload vs. Launch Site



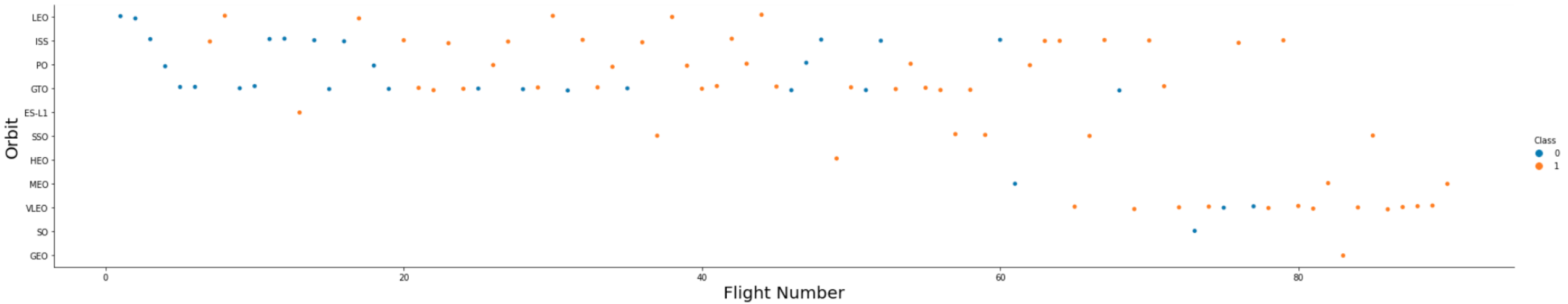
- Rockets with very high payloads are only launched from sites CCAFS SLC 40 and KSC LC 39A.
- The higher the payload, landing success (orange) increases.

Success Rate vs. Orbit Type

- Landing success rate depends on the targeted orbit.
- ES-L1, GEO, HEO and SSO have a 100% success rate.

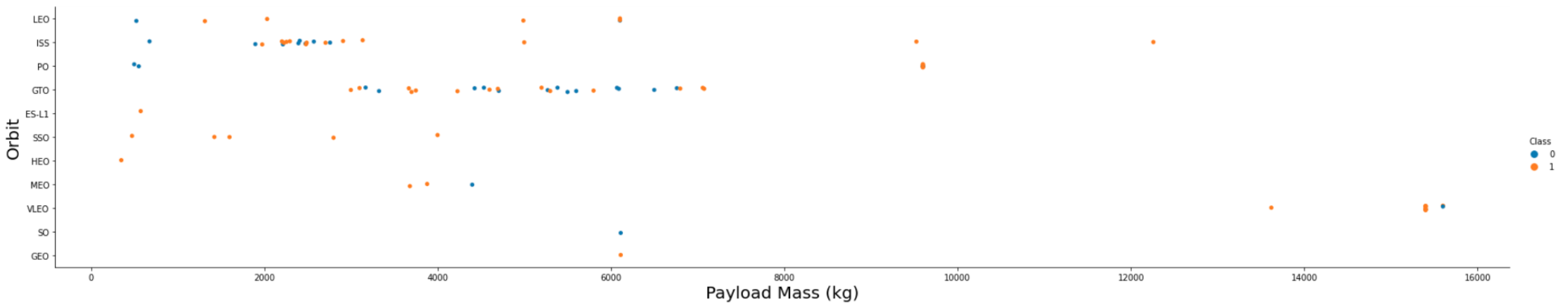


Flight Number vs. Orbit Type



- The targeted orbit changed over time. VLEO has only recently been targeted.

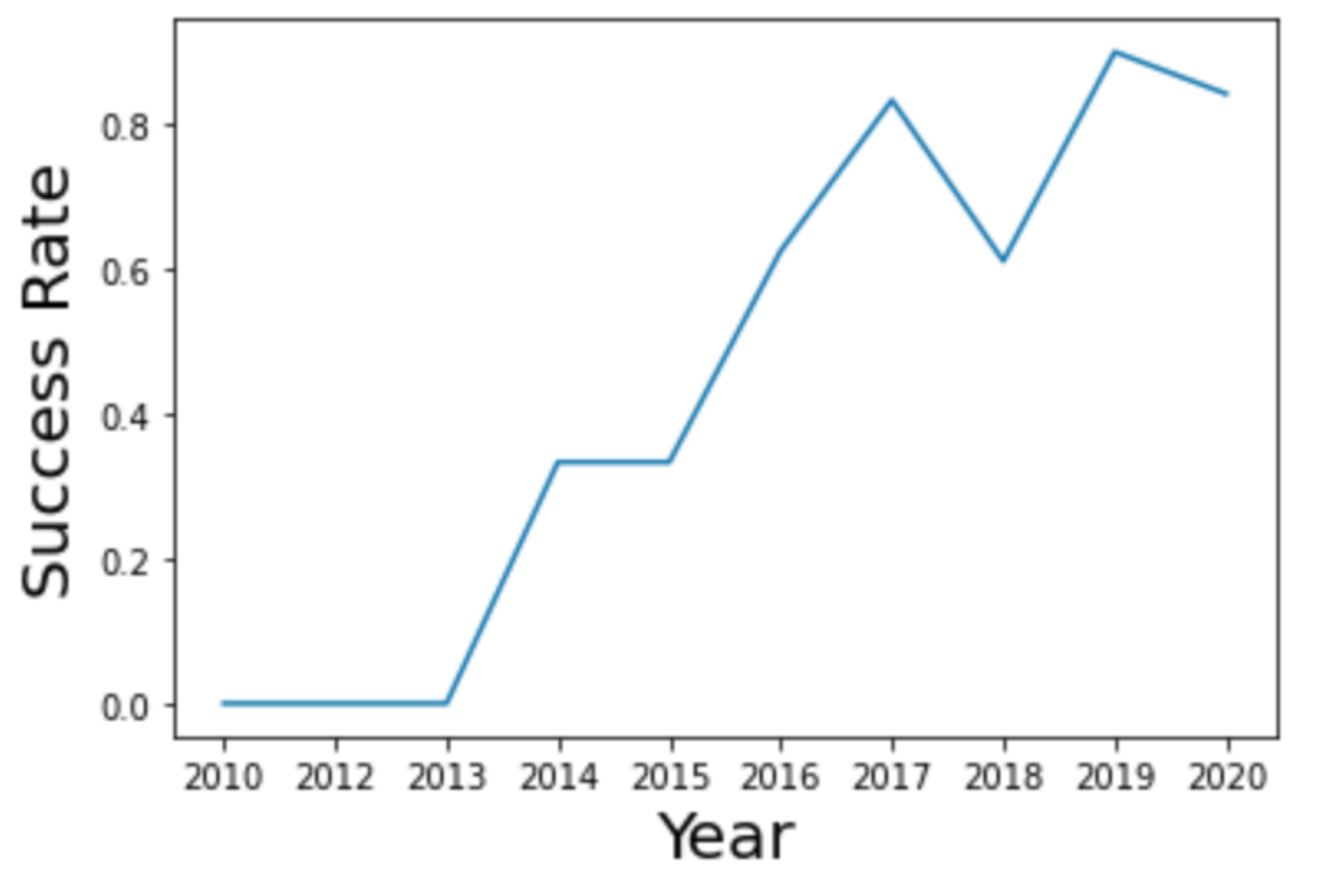
Payload vs. Orbit Type



- Launches with very high payloads (> 8’000kg) target PO, ISS or VLEO.

Launch Success Yearly Trend

- The rate of successful landings increased significantly over time: from 0% until 2013 to over 80% since 2019.



All Launch Site Names

- SpaceX rockets are launched from four different sites.

```
%%sql
select distinct Launch_Site
from spacextbl;
```

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

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

Launch Site Names Begin with 'CCA'

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

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

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

- The above statement presents the first 5 records where launch site begins with `CCA`

Total Payload Mass

- The total payload mass carried for customer NASA (CRS) is 45’596 kg

```
%%sql
select sum(Payload_Mass_kg)
from spacextbl
where Customer = 'NASA (CRS)';
```

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

sum(Payload_Mass_kg)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by F9 v1.1 boosters is 2’928.4 kg

```
%%sql
select avg(Payload_Mass_kg)
from spacextbl
where Booster_Version = 'F9 v1.1';
```

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

avg(Payload_Mass_kg)
2928.4

First Successful Ground Landing Date

- The first successful ground landing happened on 22.12.2015

```
%%sql
select min(Date)
from spacextbl
where Landing_Outcome = 'Success (ground pad)';
```

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

min(Date)
2015-12-22 00:00:00

Successful Drone Ship Landing with Payload between 4000 and 6000

- On the right is the list of booster versions that carried payload masses between 4’000 and 6’000 kg and successfully landed on a drone ship.

```
%%sql
select Booster_Version
from spacextbl
where Payload_Mass_kg between 4000 and 6000
and Landing_Outcome = 'Success (drone ship)';
```

```
 * 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

- Only 1 out of 101 missions failed, 100 are considered successful.

```
%%sql
select Mission_Outcome, count(*)
from spacextbl
group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- The list of all booster versions that carried the maximum payload mass.

```
%%sql
select Booster_Version
from spacextbl
where Payload_Mass_kg = (select max(Payload_Mass_kg) from spacextbl);
```

```
 * 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

- Month, landing outcome, booster version and launch site of failed 2015 drone ship landings.

```
%%sql
select substr(Date, 6, 2) as Month, Date, Landing_Outcome, Booster_Version, Launch_Site
from spacextbl
where substr(Date, 1, 4) = '2015'
and Landing_Outcome = 'Failure (drone ship)';
```

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

Month	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10 00:00:00	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14 00:00:00	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranked count of successful landing outcomes that took place between 04.06.2010 and 20.03.2017.

```
%%sql
    select Landing_Outcome, count(*) as count
    from spacextbl
    where Date between '2010-06-04' and '2017-03-20'
    and Landing_Outcome like 'Success%'
    group by Landing_Outcome
    order by count desc;
```

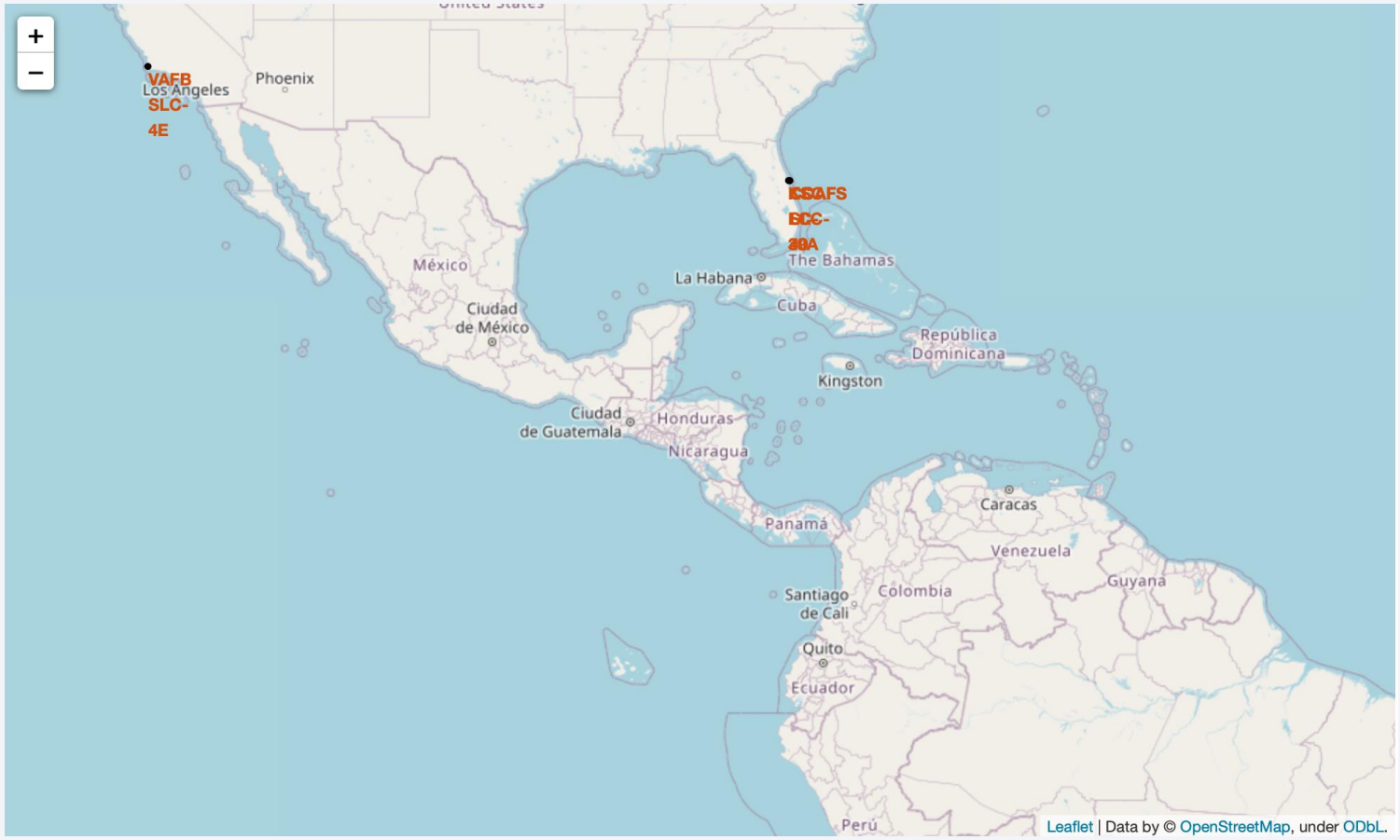
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count
Success (drone ship)	5
Success (ground pad)	3

Section 4

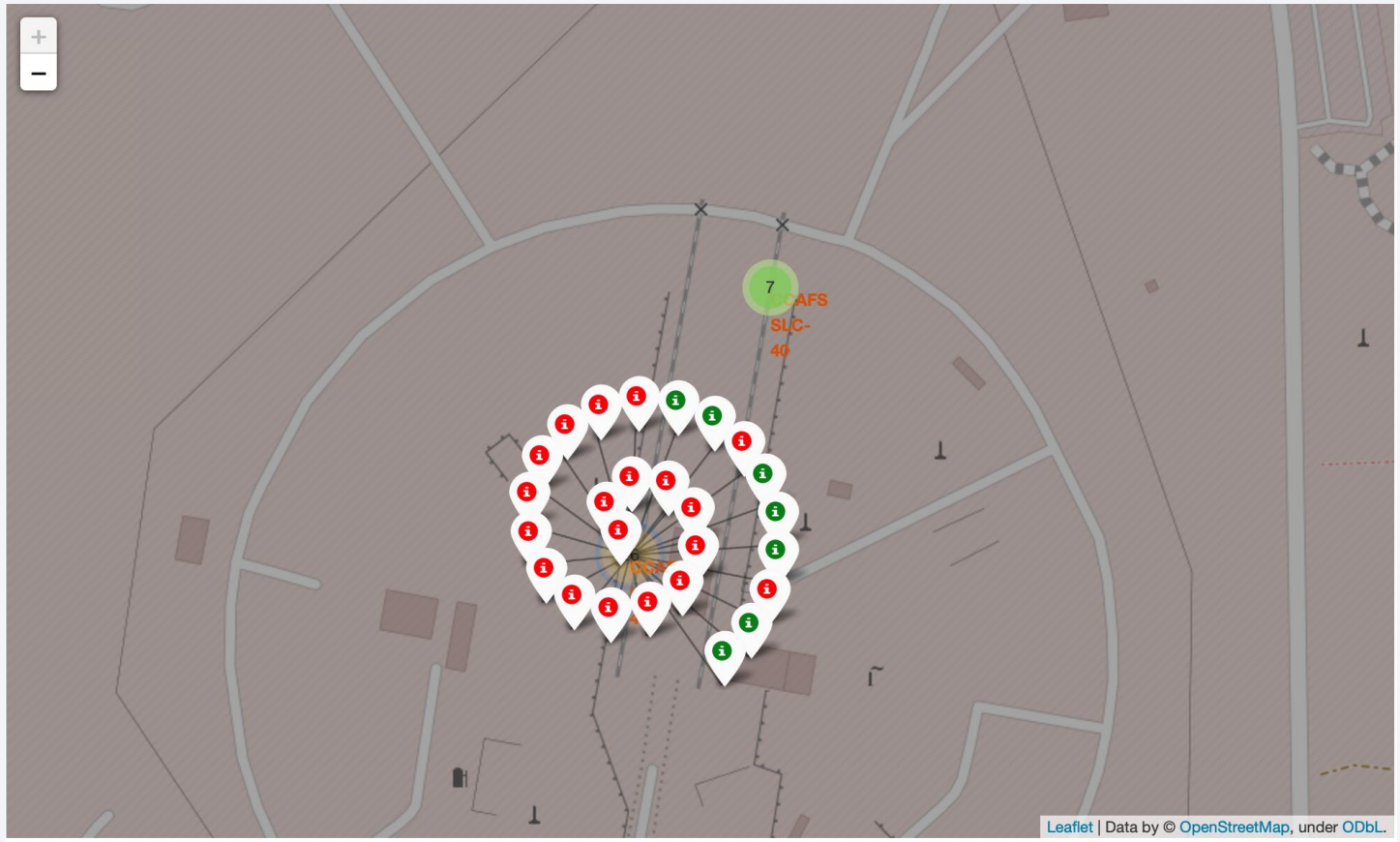
Launch Sites Proximities Analysis

Launch Site Map



- All launch sites are located near the coast and on US territory but as close as possible to the equatorial line (fuel savings based on earth rotation).

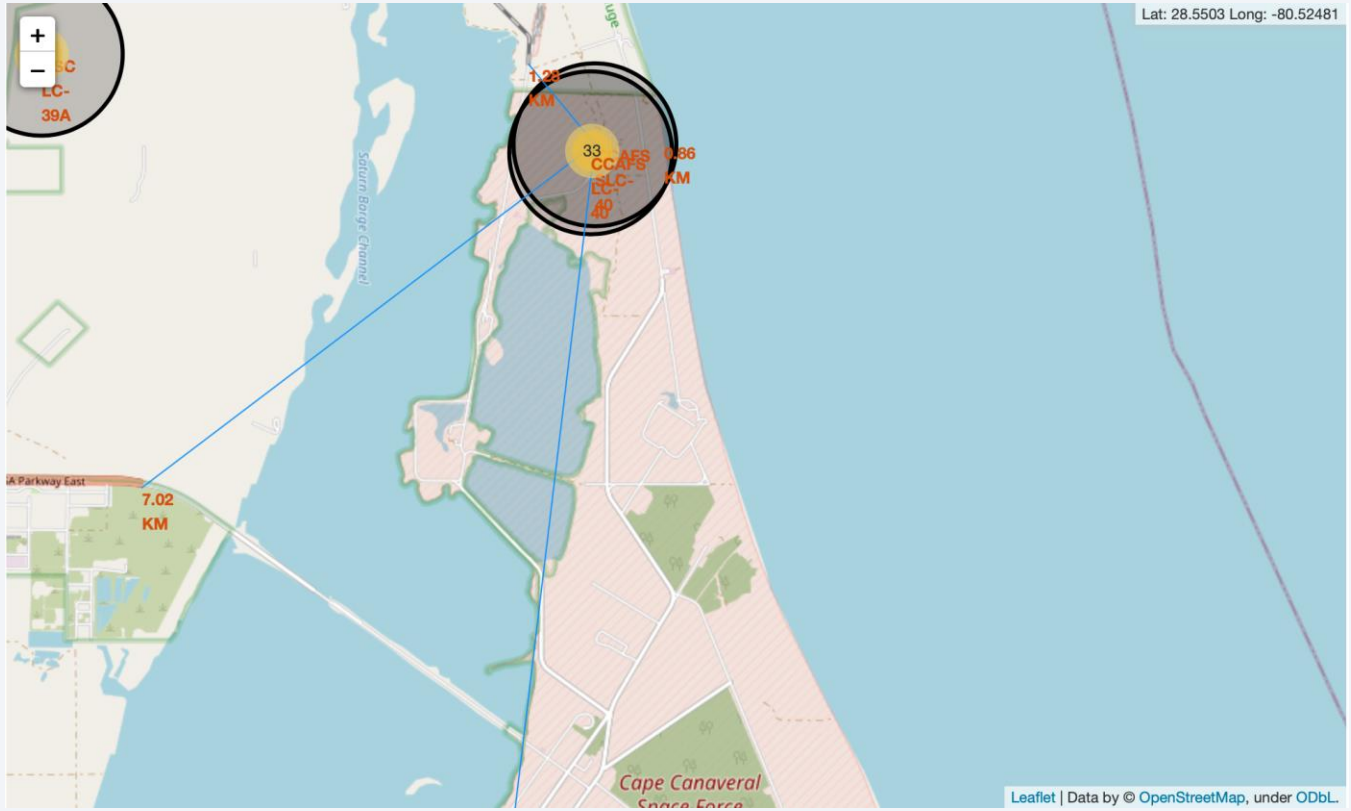
Launches with successful and failed Landings



CCAFS LC-40 launches with successful (green) versus failed landings (red).

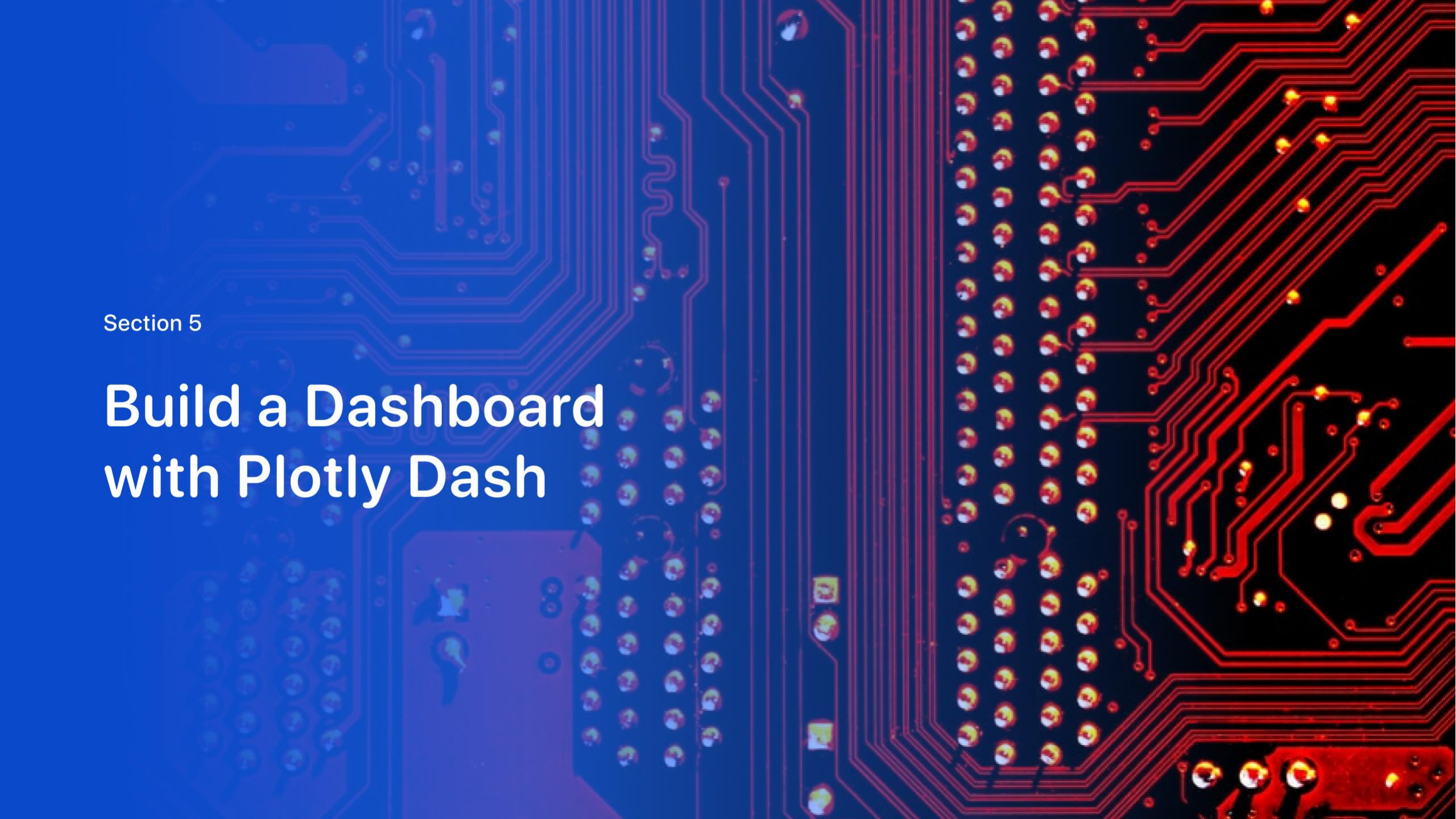
Launch Site and Distances to Proximities

- Launch site CCAFS SLC-40 and distances to its proximities:
  - Coast line: 0.86 km
  - Rail way: 1.28 km
  - Highway: 7.02 km

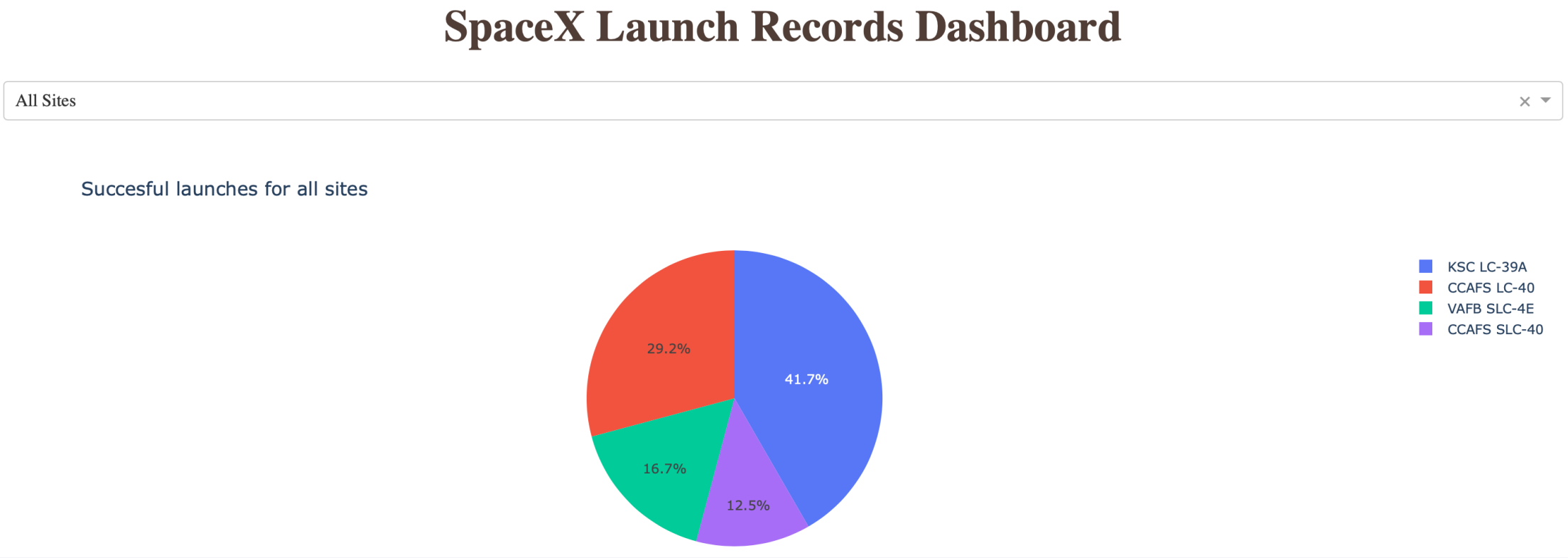


Section 5

Build a Dashboard with Plotly Dash

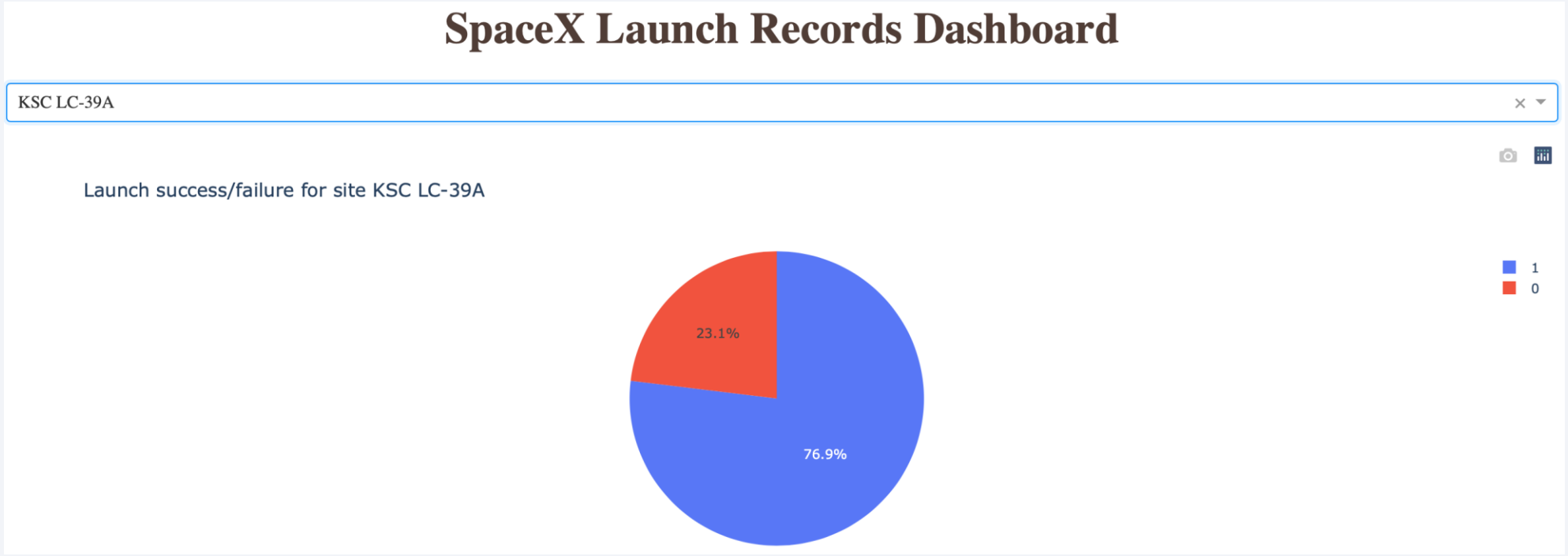


Successful launches for all Sites



- The pie chart shows the percentage each launch site contributed to the total of all successful launches.

Launch Site with Highest Success Rate



- Site KSC LC-39A has the highest success rate.

Payload, Booster Version and Success Rate

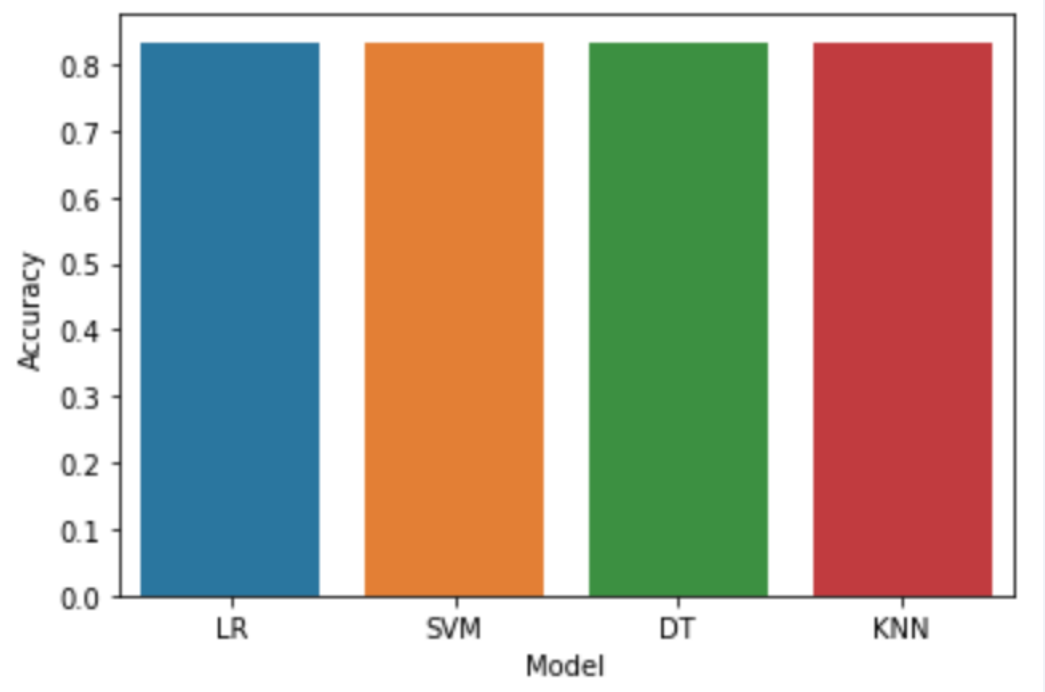
- Different payloads are carried by different booster versions, but also have different success rates:
  - Top: all kinds of payloads
  - Middle: high payloads
  - Bottom: low payloads

Section 6

Predictive Analysis (Classification)

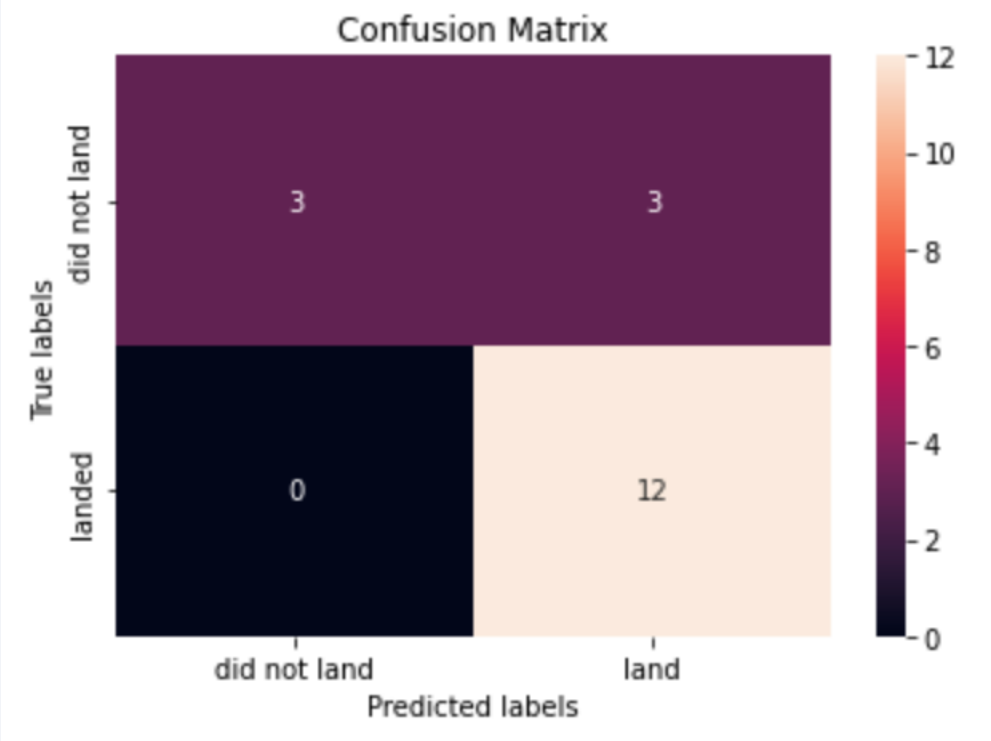
Classification Accuracy

- All classification models have the same test set accuracy of 83.33%



Confusion Matrix

- Also the confusion matrix for all of the evaluated models look exactly the same.
- 15 out of 18 test samples have been predicted correctly. There were no false negatives, but 3 out of 18 false positives (predicted as successful landing, but failed in reality).



Conclusions

- Publicly available data (Wikipedia and SpaceX API) is sufficient to make predictions for the 1st stage landing success of future SpaceX Falcon 9 missions.
- The compared predictive models (Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbors) based on historic data have comparable, high accuracy of 83.33%.
- There are great open source tools available, not only to create predictive models, but also to collect, wrangle, store and visualize data; to create dashboards and process geospacial information.

Appendix

- GitHub repository: https://github.com/qualityland/IBM_Applied_Data_Science_Capstone

Thank you!