Winning Space Race with Data Science

Mari Arzan
17 October 2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

On the road to affordable space travel, an essential issue is the reliable reuse of the rockets, so their successful landings matter.

- Summary of methodologies: *Exploratory* data analysis was used to get a grasp of the ongoing process, *Interactive* analysis was performed to illustrate the key points, *Predictive* analysis was applied to select the best classification model per the available data on the SpaceX successful landings.
- Summary of all results: Successes increase with the launch counts and the payload mass, reaching about 80% rate by 2017. For Falcon 9, booster version F9 B5 B1048.4 carries the maximum payload mass, with success to failure ratio of 61:10. KSC LC-39A site has over 2/5 of the total number of successful launches, while CCAFS LC-40 site showed the worst results despite hosting the highest number of launches. Decision Tree model is the most accurate Machine Learning technique to assess if the first stage will land based on the available data – with the key issue being false positive predictions.

Introduction

- Project background and context: The new developments in the area of space launches include operationalization of innovative technologies that demand less resources to take off the Earth and deliver the artisan spaceship to the existing destinations in space (planets, space stations, maybe comets).
- Problems needing answers: The tangible questions worth answering can be estimation of these launches' cost and their chances of successful landing so that the first stage rocket could be reliably reused – as shown below.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Request and parse SpaceX launch data, filter the resulting dataframe, deal with the missing values (with application of API and/or BeautifulSoup methods).
- Perform data wrangling: e.g. calculate the number of launches on each site, the number and occurrence of each orbit and of mission outcome per orbit type, create a landing outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: after the data are cleaned and standardized, apply the established classification models (e.g. logistic regression, support vector machine, decision tree classifier, K-nearest neighbours), calculate their accuracy, and chose the model with the best indicator values

Data Collection

- How data sets were collected: The SpaceX REST Application Programming Interface (API) was used for this exercise. The output contained the data on rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Data collection process: The API's specific "endpoint" through its URL is accessed by GET request, to receive past launch data. The resulting JSON files are then converted to dataframes and the next step is "cleaning" the data. Alternatively, the data was collected using BeautifulSoup method – see the links below.

Data Collection (API, Scrapping), Data Wrangling

- Data Collection – API (GitHub link)

Request and parse SpaceX launch data via its API and convert it from JSON to a Pandas dataframe.

- Data Collection – Scrapping (GitHub link)

Alternatively: apply BeautifulSoup method to scrap the data from the relevant webpages and convert it from HTML to a Pandas dataframe.

- Data Wrangling (GitHub link)

Clean and format the data, deal with the missing values, standardize and normalize the data. Also, get the key features of the acquired dataset, e.g. elicit the prevalent occurrences and possible outliers.

Exploratory Data Analysis with Data Visualization

- Number of flights and Payload mass versus Launch site and Orbit type: these four graphs aim to characterize the relation between the throughput features and the setting features of the trials.
- Success rate per Orbit type and per year: these two graphs aim to characterize the success placement and trend.

Details are available online (GitHub link)

Exploratory Data Analysis with SQL

- Launch site names
- Total and Average payload masses
- Launch outcomes per date and per site
- Success and failure totals
- Maximum payload by booster version
- First success date
- Success by above average payload mass
- Rank landing outcomes

Details are available online (GitHub link)

Build an Interactive Map with Folium

- Certain map objects (markers, circles, lines) were created and added to the Folium maps.
- These were used to show the placement of the Launch sites within the USA and basic success and failure statistics per site.
- The line from a selected Launch site to the nearby railway point demonstrates the proximity of the key infrastructure.

Details are available online (GitHub link)

Build a Dashboard with Plotly Dash

- Pie charts and Scatter plots were added to the dash-board.
- These aim to illustrate the success rate per site and the relation between the Payload mass and the Launch outcome per site.

Details are available online (GitHub URL)

Predictive Analysis (Classification)

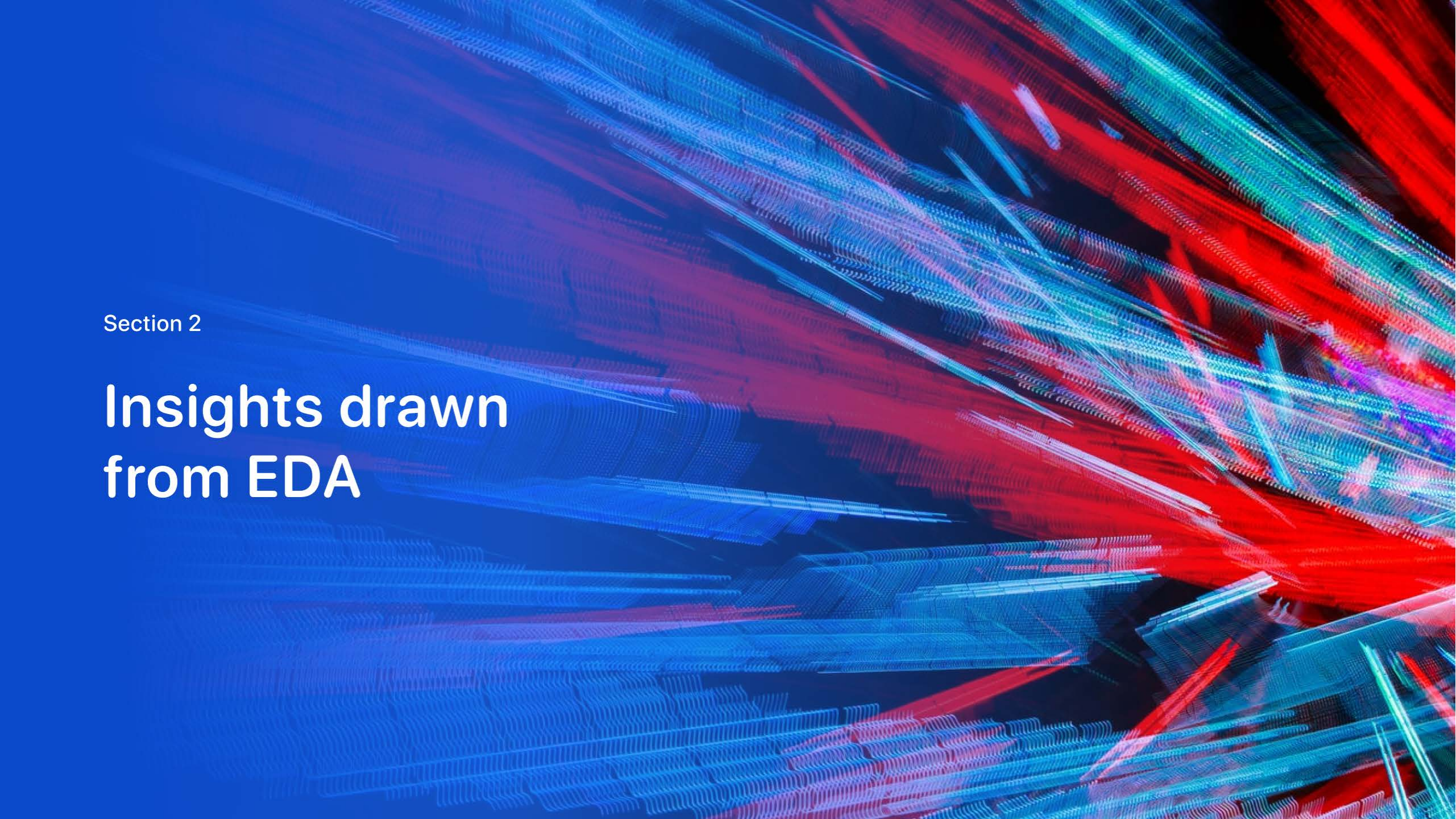
- Classification methods were deployed to perform the necessary predictive analysis: Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbors
- The flowchart starts at getting the data, cleaning and normalizing it, then loops through four Machine Learning techniques to yield their accuracy assessment values, then compares the values to point out the least erroneous landing outcome prediction method.

Details are available online (GitHub link)

Results

- Exploratory data analysis provided a perception of the nature of the data and its peculiarities.
- Interactive analytics demonstrated the revealed features in a reader-friendly manner (see demo in screenshots below).
- Predictive analysis delivered the best classification method to choose the prediction model.

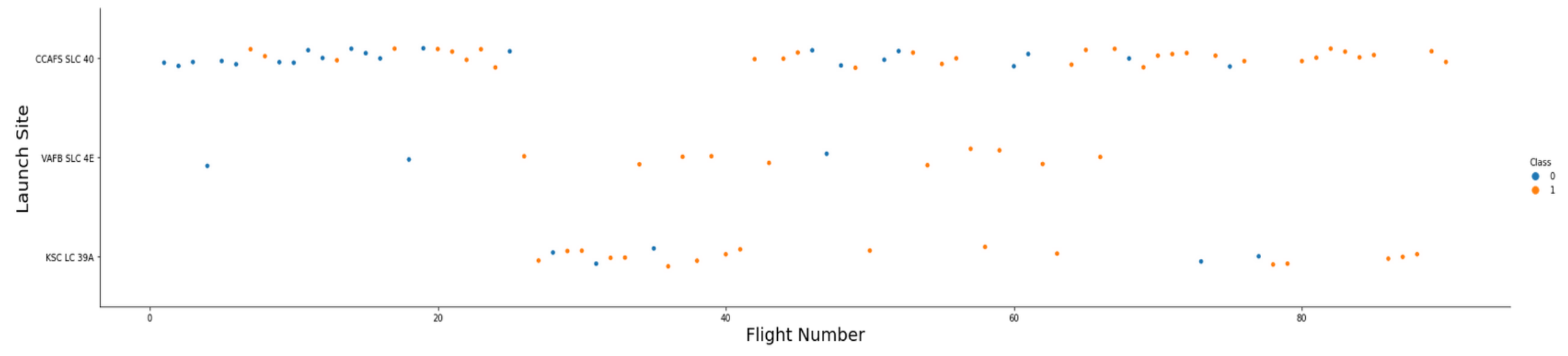
The complete project exercises are available online (GitHub link)



Section 2

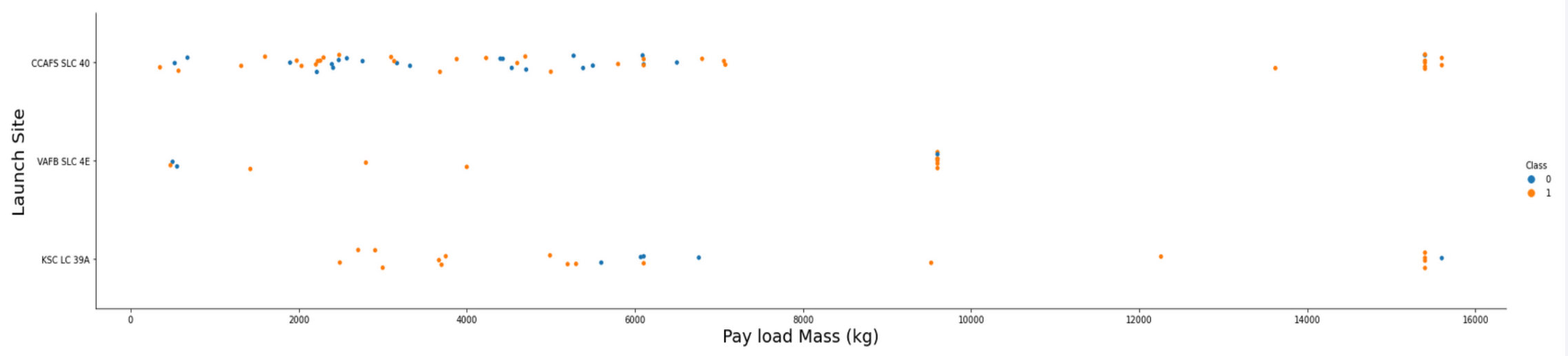
Insights drawn from EDA

Flight Number vs. Launch Site



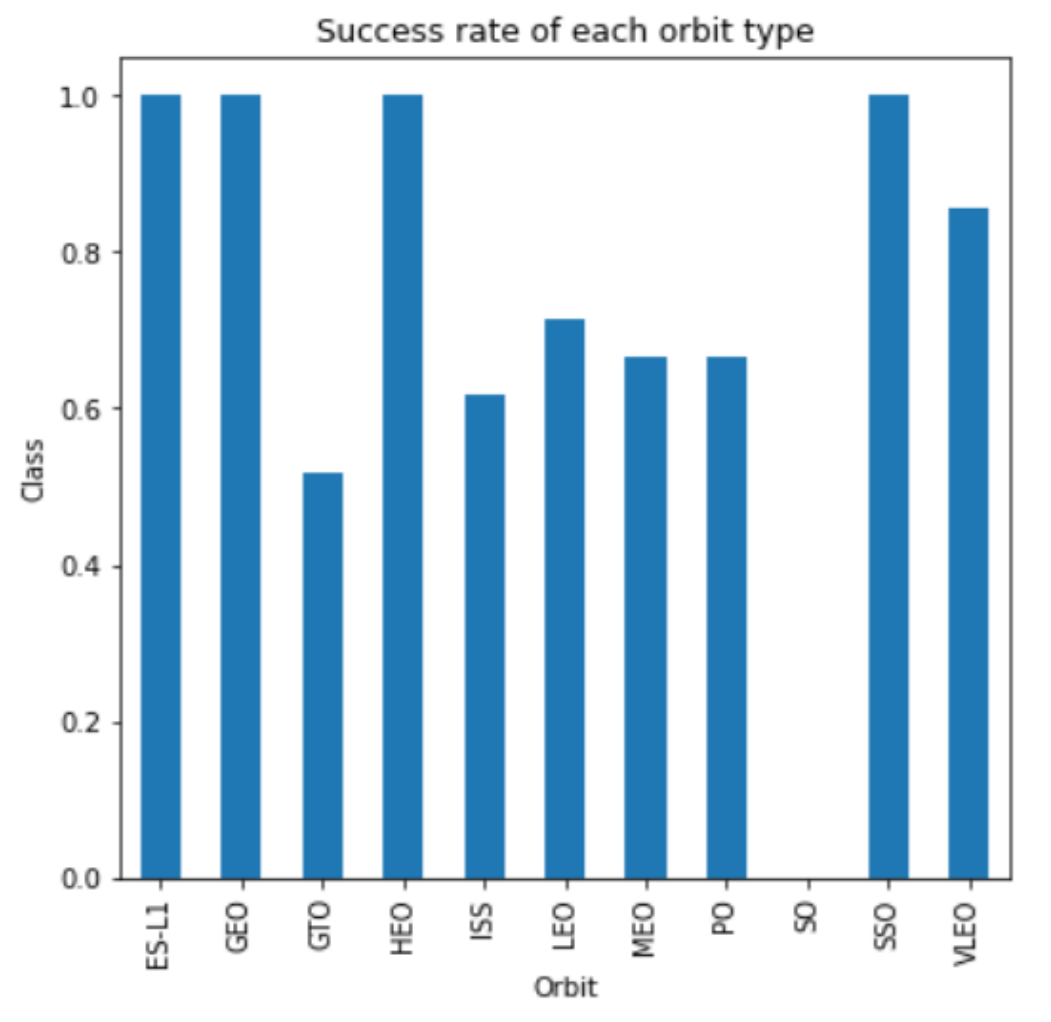
- It appears that success rate generally improves with the increase in launch count.
- Most of the launches were performed from CCAFS SLC 40 site with mixed results.

Payload vs. Launch Site



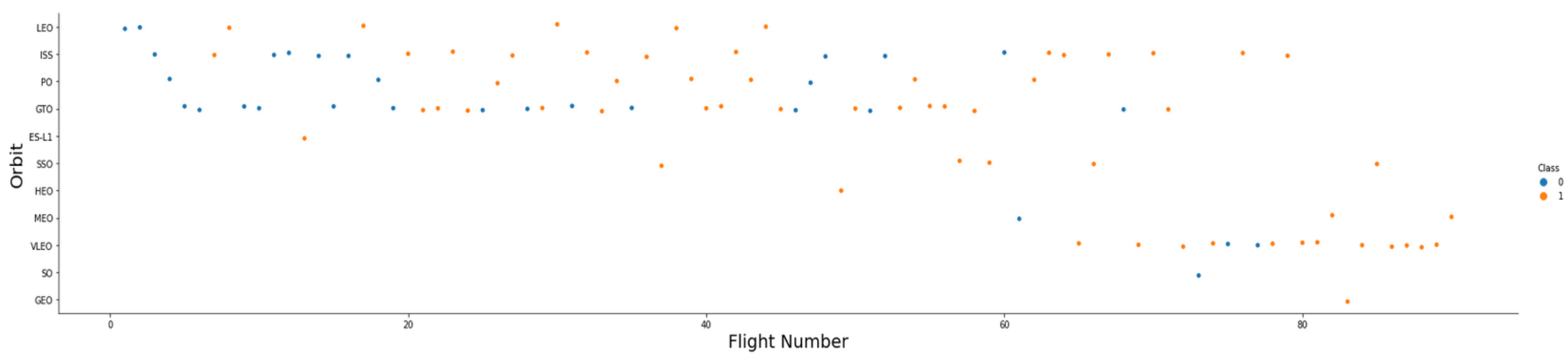
- Rockets with payload mass below about 7 thousand kg were launched mostly from CCAFS SLC 40 site with mixed results.
- Rockets with pay load mass closer to the maximal appear to have launched mostly successfully.

Success Rate vs. Orbit Type



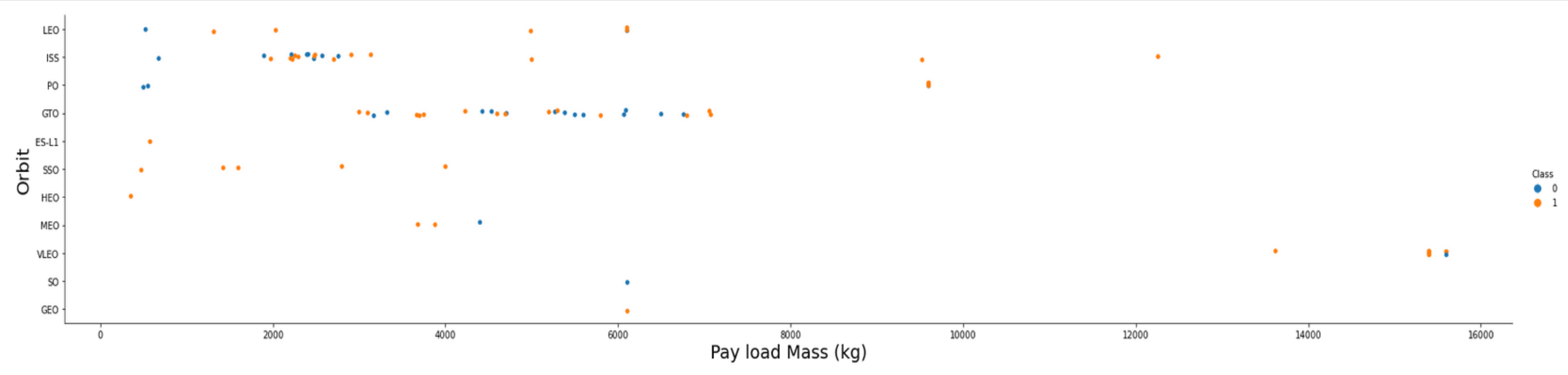
- The highest success rates are registered at orbits ES-L1, Geo, HEO, SSO
- With the exception of VLEO orbit, the less successful ones show significantly lower success rates
- SO orbit seems to be an outlier

Flight Number vs. Orbit Type



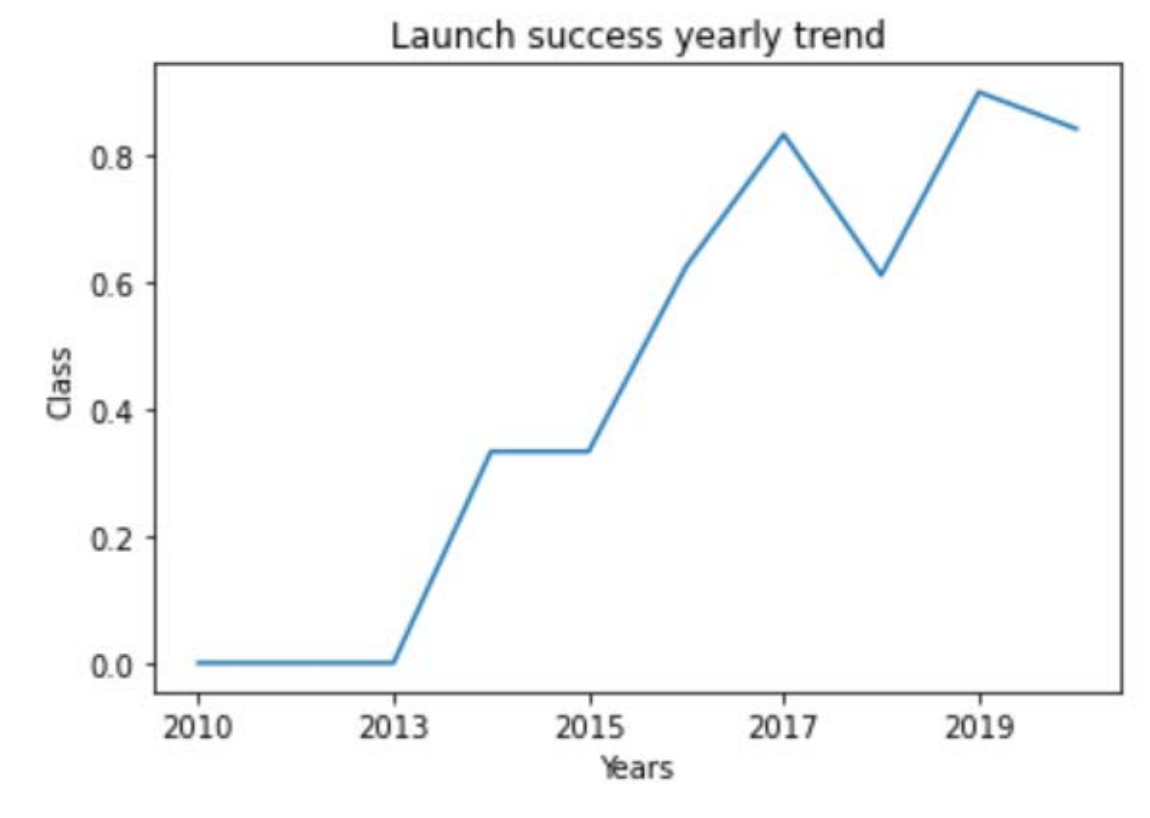
- Success rate seems to increase with growing number of flights
- Mixed results in GTO orbit.

Payload vs. Orbit Type



- High payloads show lower success rates in GTO orbit
- Only one failure is visible for payload masses above about 7 thousand kg
- LEO and ISS orbits show better results for higher payloads

Launch Success Yearly Trend



- The trials are successful since 2015
- By 2017, about 80% success rate was achieved
- Since then the growth fluctuates (anyway, you cannot exceed 100%)

All Launch Site Names

```
%%sql
select distinct launch_site from spacex;
```

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

Launch Site Names Beginning with 'CCA'

```
%%sql
select launch_site from spacex where launch_site like 'CCA%' limit 5
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%%sql
select sum(PAYLOAD_MASS__KG_) as total from spacex where customer = 'NASA (CRS)'
```

total
45596

Average Payload Mass by F9 v1.1

```
%%sql
select avg(PAYLOAD_MASS__KG_) as average
        from spacex where booster_version = 'F9 v1.1'
```

average
2928

First Successful Ground Landing Date

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

beginning
2015-12-22

Successful Drone Ship Landing (Payload 4000 to 6000)

```
%%sql
select booster_version from spacex where Landing_Outcome = 'Success (drone ship)'
        and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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 count(Landing_Outcome) as done from spacex
        where Landing_Outcome like '%uccess%'
```

done
61

```
%%sql
select count(Landing_Outcome) as failed from spacex
        where Landing_Outcome like '%ailure%';
```

failed
10

Boosters Carried Maximum Payload

```
%%sql
select distinct booster_version,
PAYLOAD_MASS__KG_ as heavy
        from spacex where
        PAYLOAD_MASS__KG_
        in (select max(PAYLOAD_MASS__KG_)
        from spacex);
```

booster_version	heavy
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

```
%%sql
select Landing_Outcome, booster_version, launch_site, date
        from spacex where Landing_Outcome = 'Failure (drone ship)' and year(date) = 2015;
```

landing_outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

Rank Landing Outcomes (2010-06-04 to 2017-03-20)

```
%%sql
select Landing_Outcome, count(Landing_Outcome) as counts
        from spacex where date >= '2010-06-04' and date <= '2017-03-20
        group by Landing_Outcome order by counts desc;
```

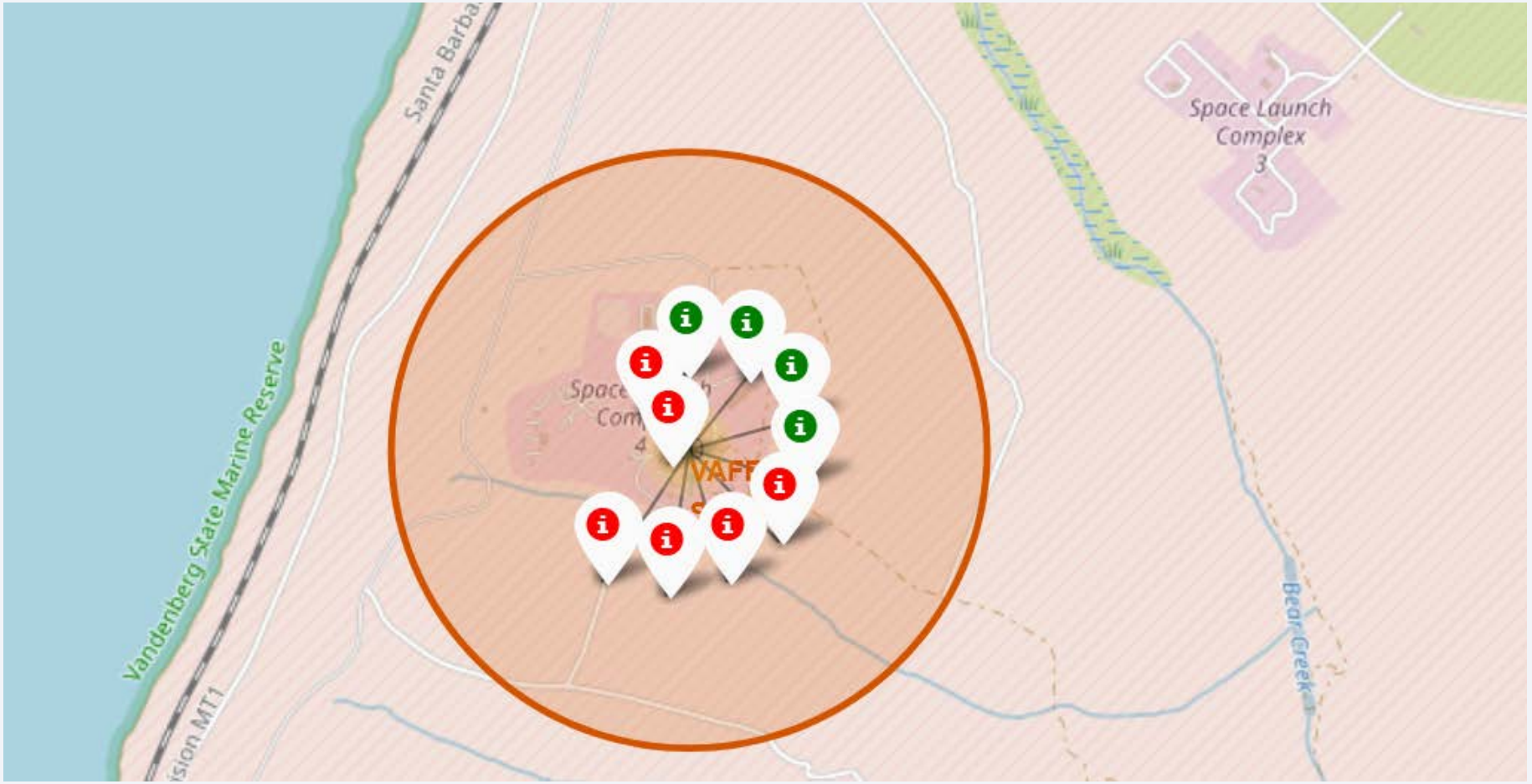
landing_outcome	counts
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

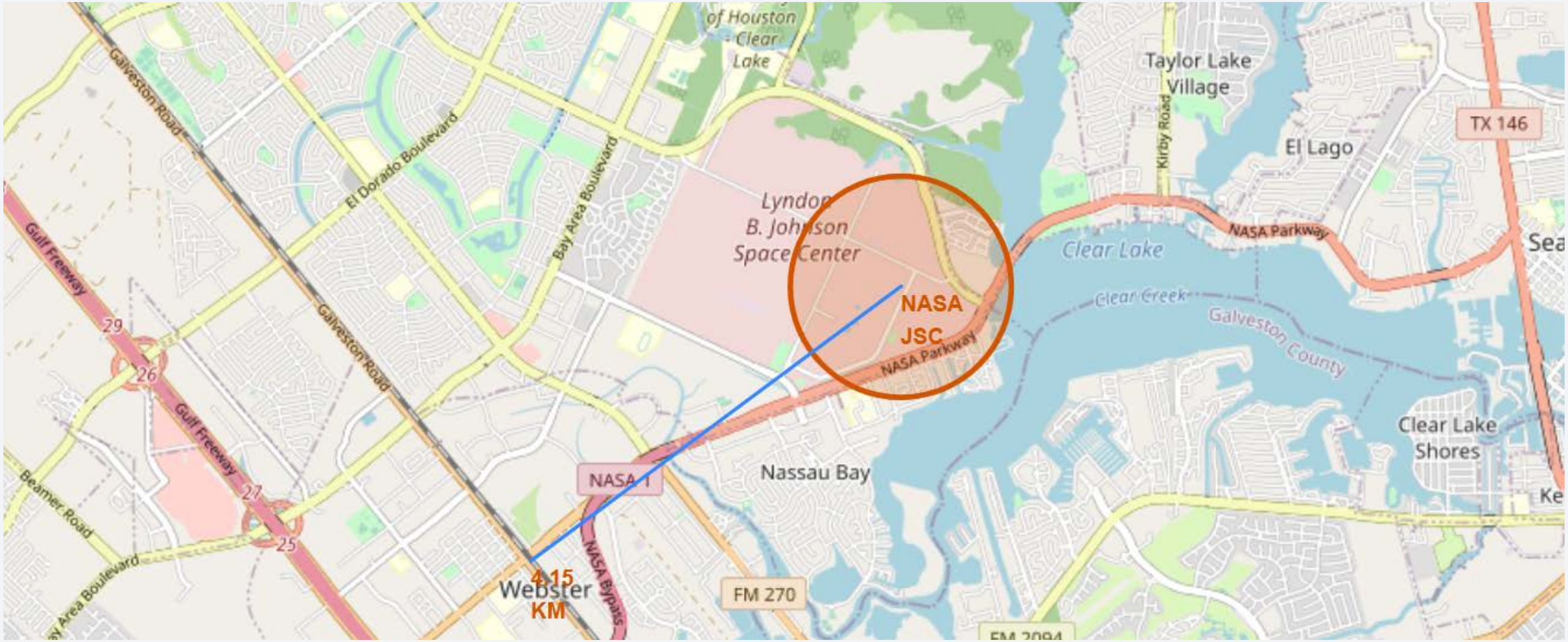
Launch Sites Proximities Analysis

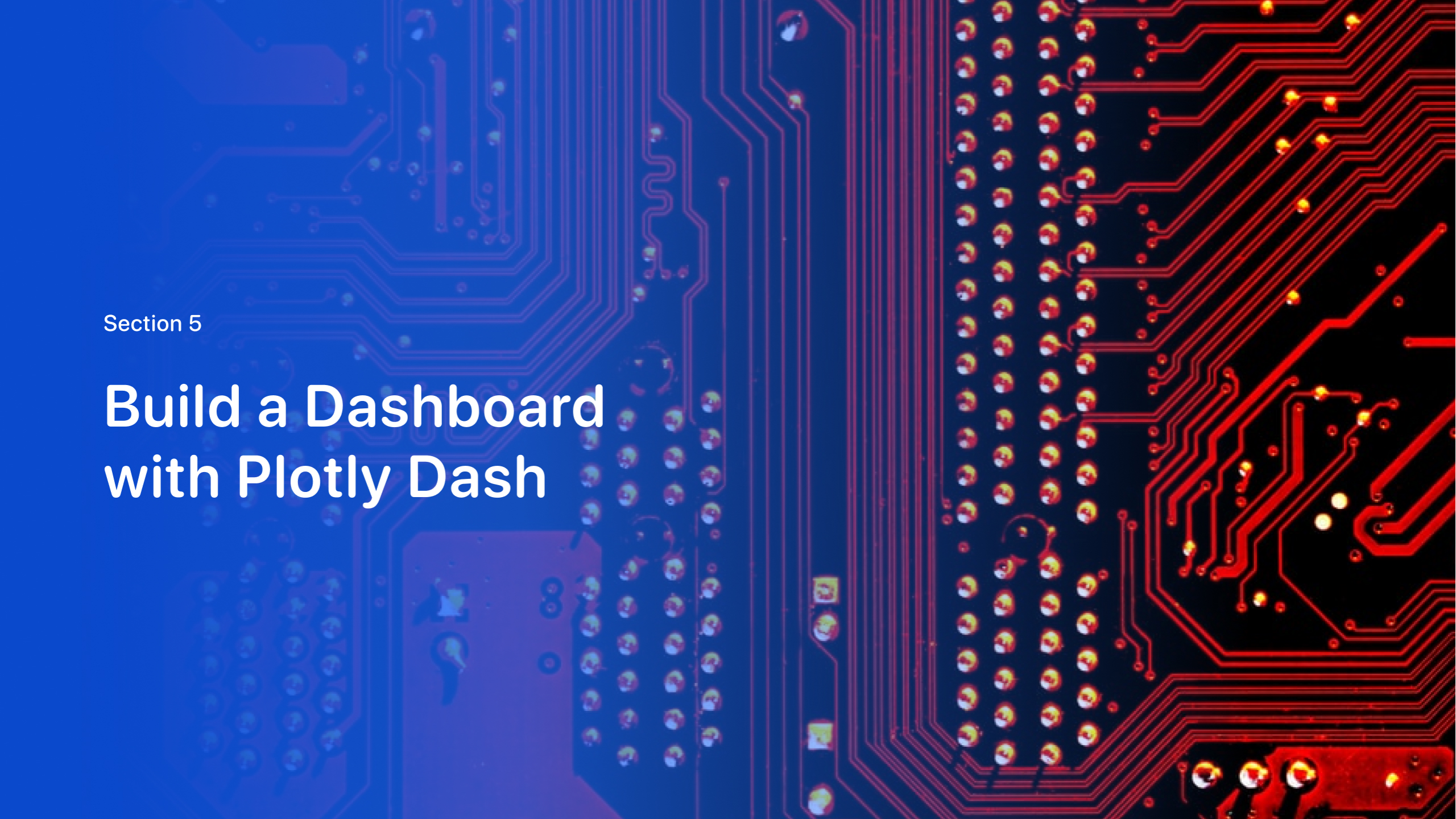
All launch sites' location markers

Color-labeled launch outcomes at VFB SLC 4E site



Launch site near Houston linked to a selected railway point

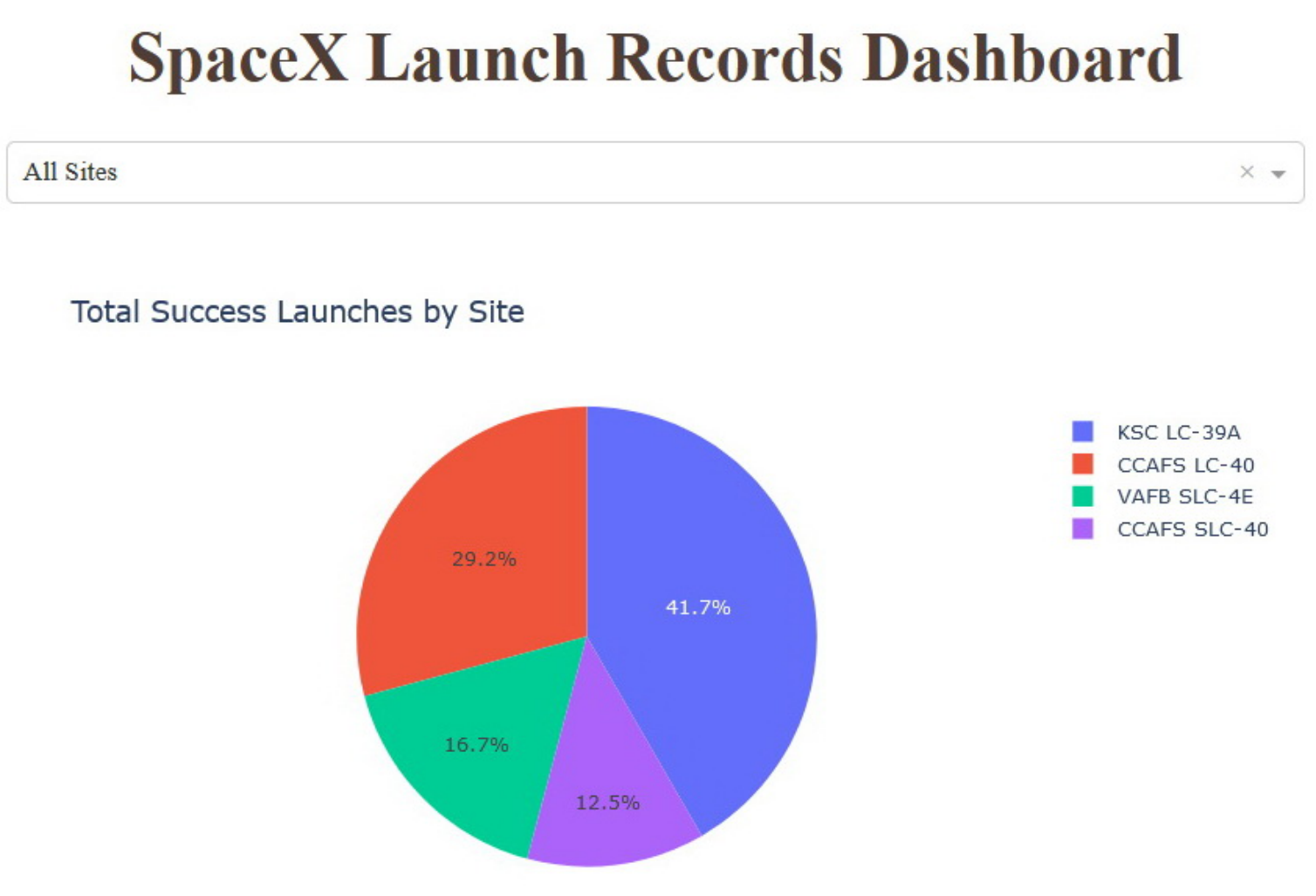




Section 5

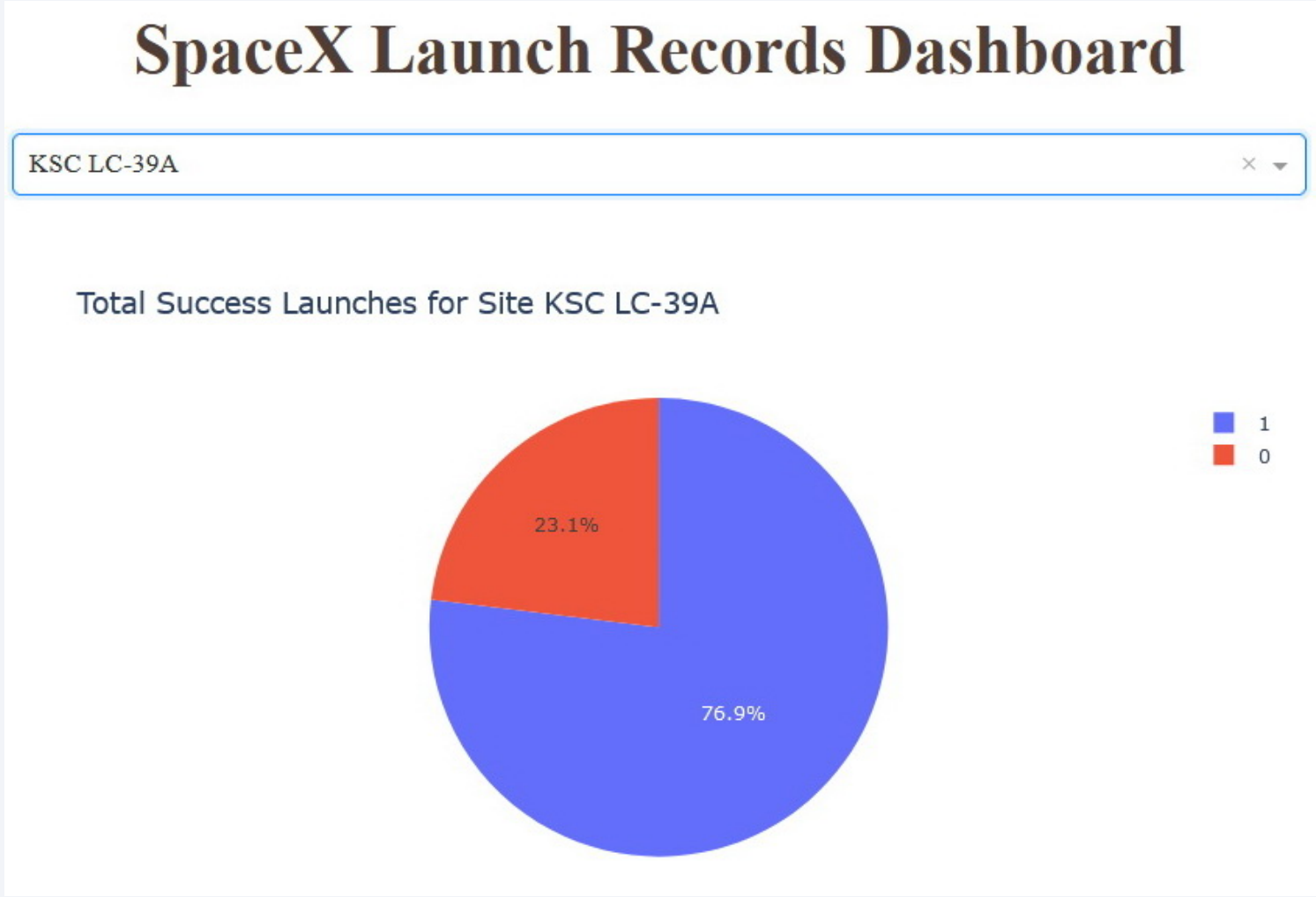
Build a Dashboard with Plotly Dash

Launch success count for all sites



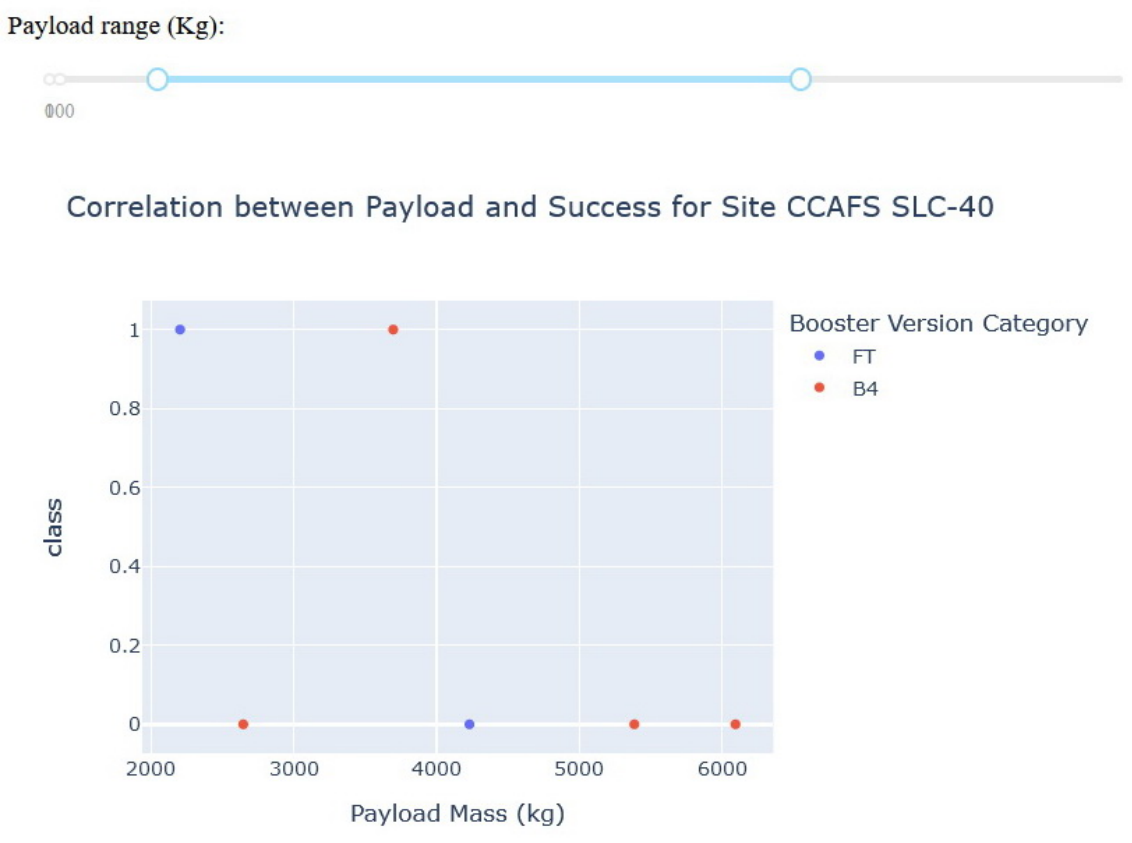
- The pie chart represents the share of each site in the total number of successful launches.
- The pie chart does not compare the success rates of individual launch sites.

Launch site with highest launch success ratio

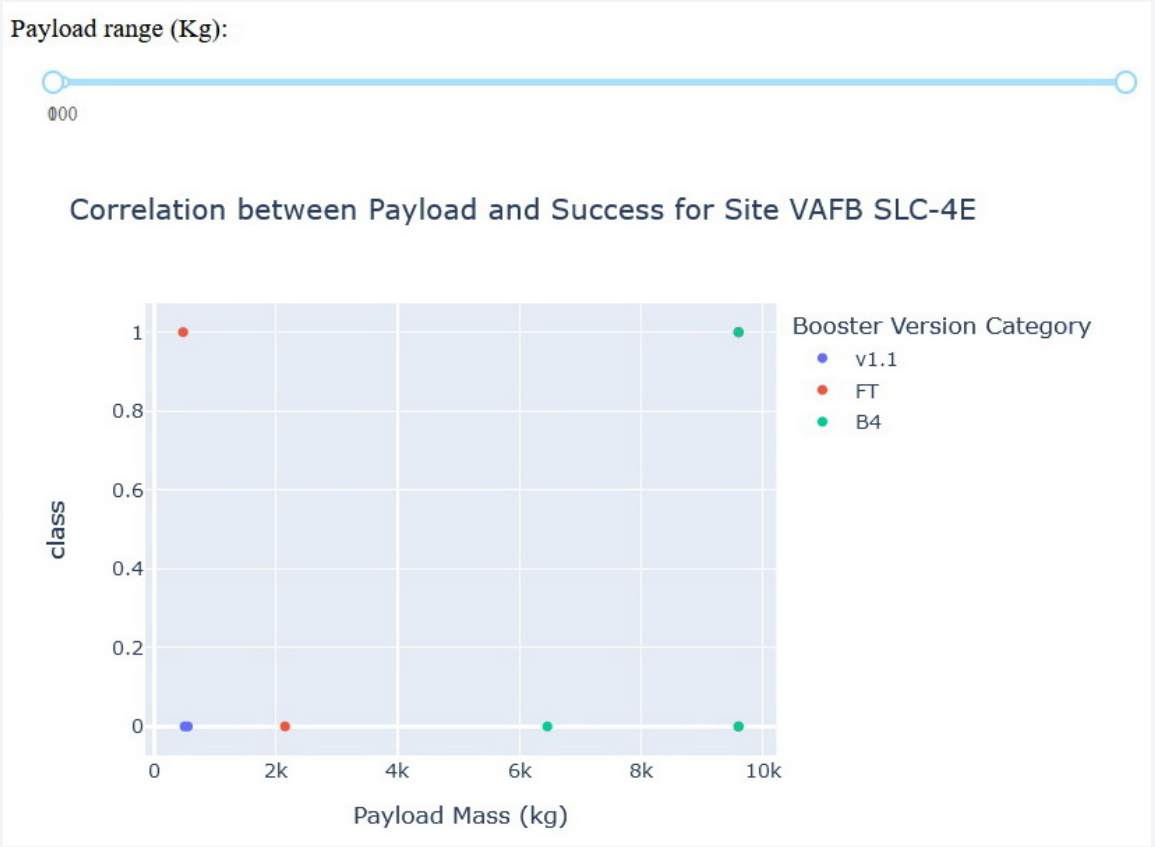


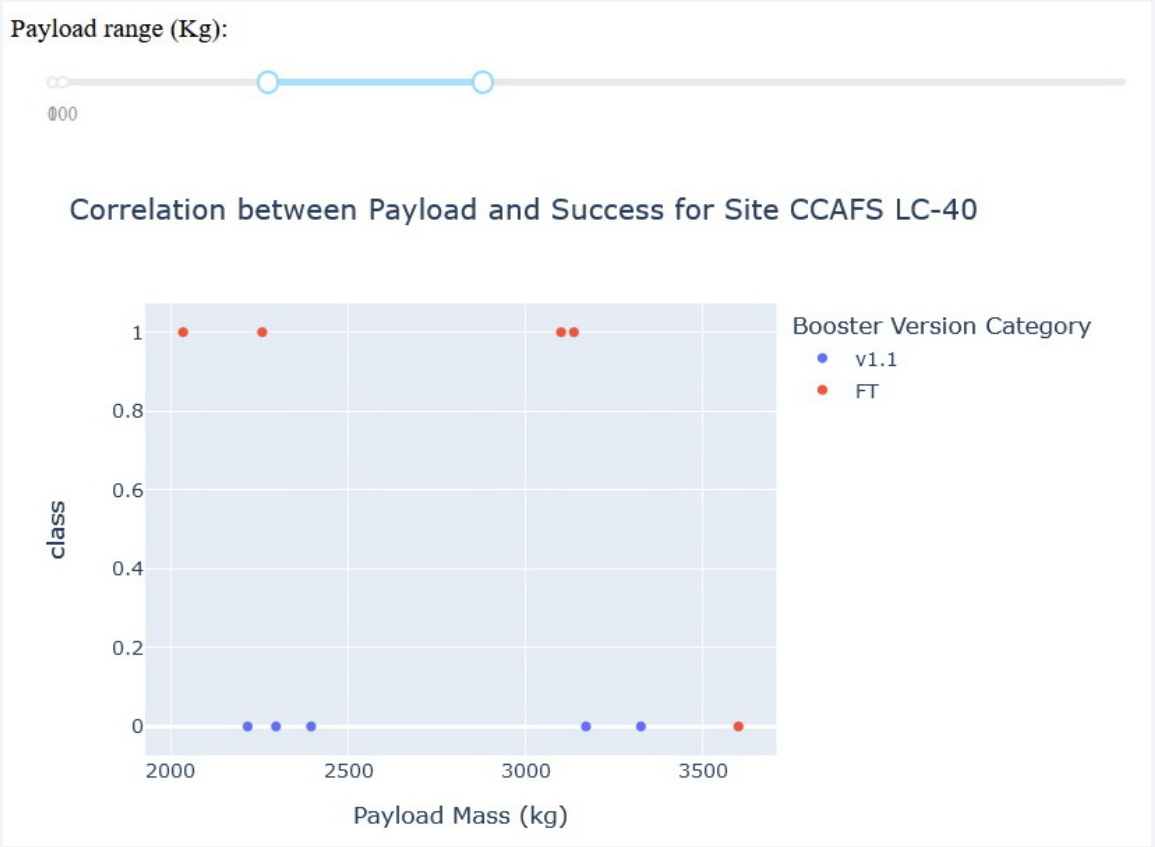
- The presented ratio refers to 10 successful launches out of 13.
- Next comes CCAFS SLC-40 with 3 of 7
- The third is VAFB SLC-4E with 4 of 10
- The fourth is CCAFS LC-40 with 7 of 26

Payload vs. Launch Outcome scatter plots 1 and 2



Payload vs. Launch Outcome scatter plots 3 and 4

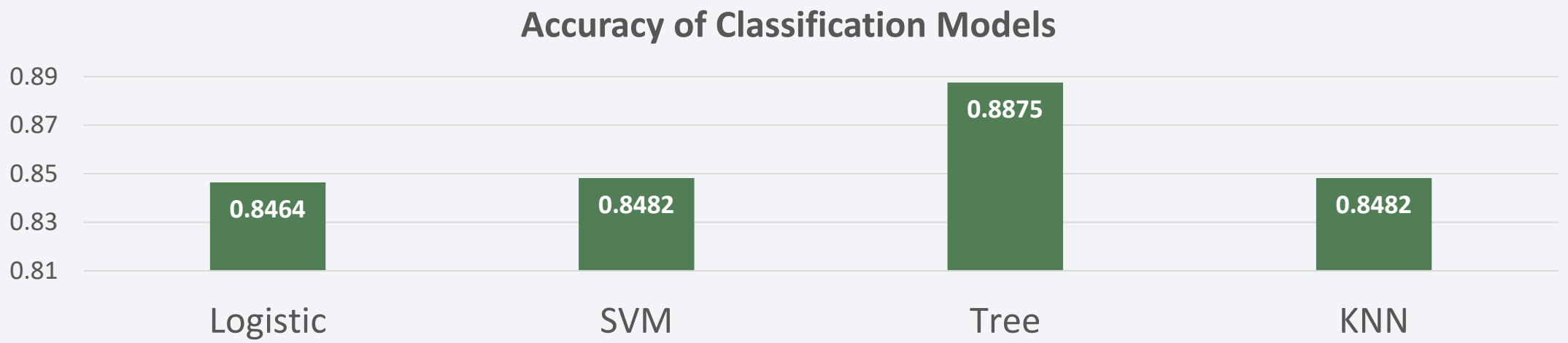




Section 6

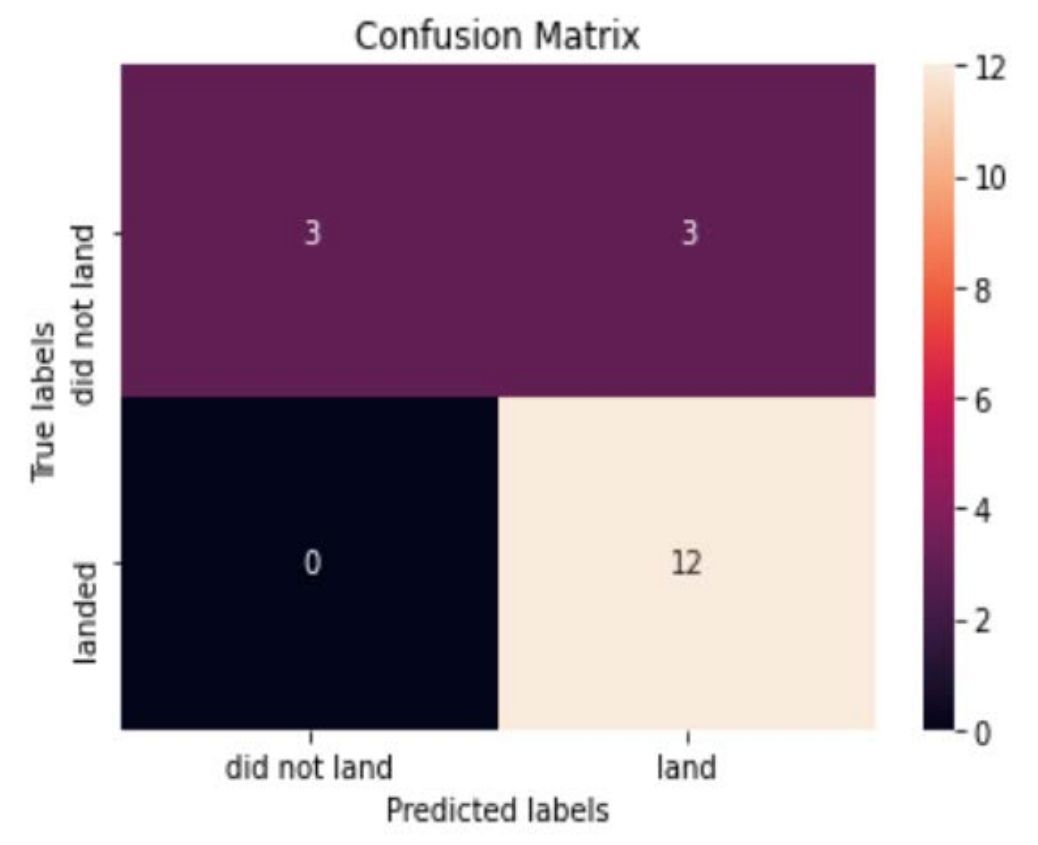
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree model is the most accurate one to assess if the first stage will land based on the available data.
- Other Machine Learning techniques used: Logistic Regression, Support Vector Machine, K Nearest Neighbors

Confusion Matrix



- The key issue are the false positive predictions (according to the upper right corner of the Confusion matrix)
- No false negative predictions observed.

Conclusions

- Successes increase with the launch counts and the payload mass
- Since the first success in 2015, about 80% success rate was achieved by 2017
- For Falcon 9, booster version F9 B5 B1048.4 carries the maximum payload mass, while success to failure ratio for landings is 61:10
- KSC LC-39A site has over 2/5 of the total number of successful launches
- While hosting the highest number of launches, CCAFS LC-40 site showed the worst results
- Decision Tree model is the most accurate Machine Learning technique to assess if the first stage will land based on the available data – with the key issue being false positive predictions

Thank you!