Winning Space Race with Data Science

Helio Murata
August 16, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with Web Scraping
  - Data Collection API with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with SQL
  - EDA with Visualization
  - Interactive Visual Analytics and Dashboards
  - Machine Learning Predictive Analysis
- Summary of all results
  - EDA results
  - Interactive Analysis results
  - Predictive Analysis results

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.
  - Machine learning will be used to predict if the first stage will land successfully.
- Problems to find answers
  - What is the best launch site with highest success rate?
  - What factors and conditions imply the success of launch?
  - What is the best way to predict a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from SpaceX obtained using:
    - SpaceX API: https://api.spacexdata.com/v4/
    - Wikipedia web scraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Exploratory Data Analysis was performed, and Training Labels was determined.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was standardized and splitted into training data and test data.
  - Classification models such as Logistic Regression, SVC, Decision Tree Classifier and K-Neighbors Classifier were used.
  - Tests are performed to determine the best classification model.

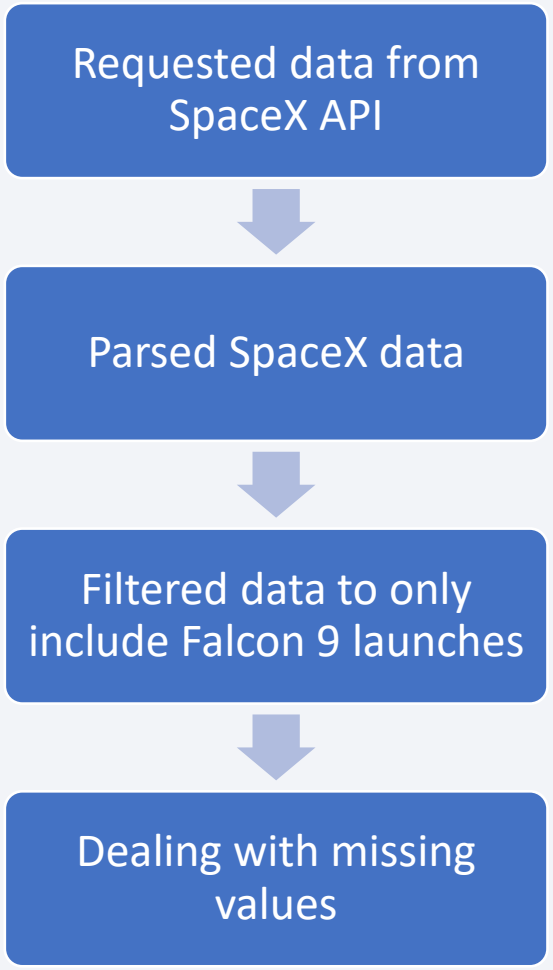
Data Collection

- Data from SpaceX was collected using:
  - SpaceX API: https://api.spacexdata.com/v4
  - Wikipedia web scraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX API data:
  - Boosted version: https://api.spacexdata.com/v4/rockets
  - Launch site: https://api.spacexdata.com/v4/launchpads/
  - Pay load data: https://api.spacexdata.com/v4/payloads/
  - Core data: https://api.spacexdata.com/v4/cores/
- Data get requested to the SpaceX API
- Cleaned the requested data
- Link for Notebook from Github:

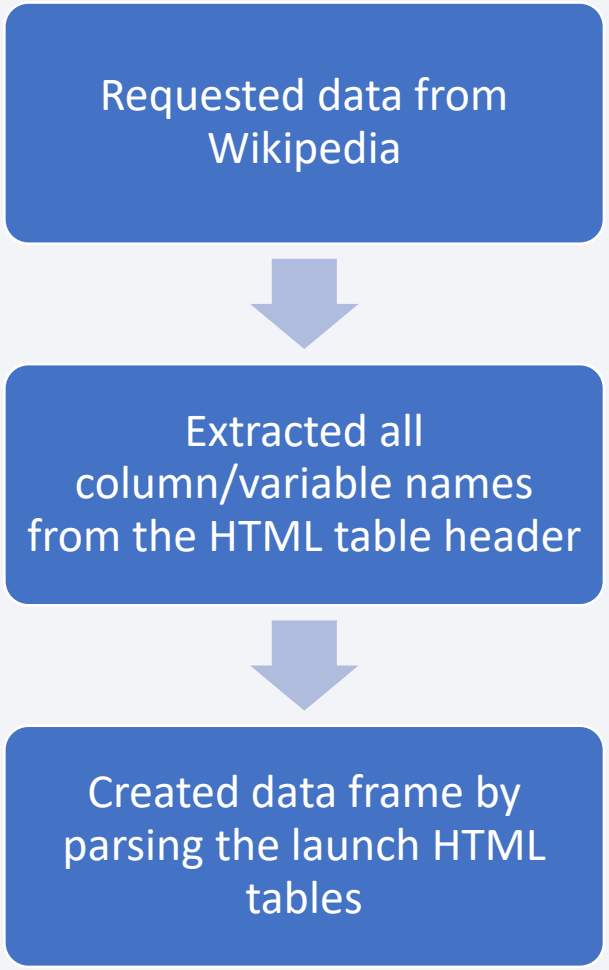
https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20SpaceX%20API.ipynb



Data Collection - Scraping

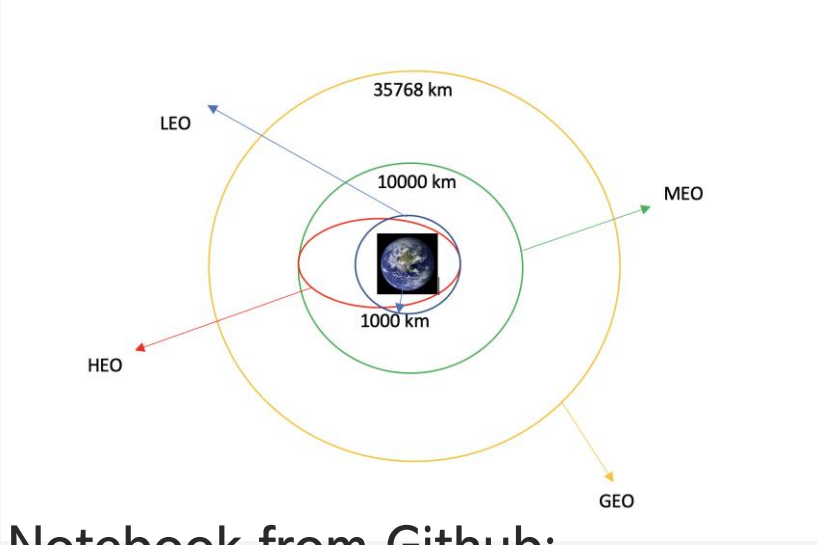
- SpaceX Data from Wikipedia:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Extract Falcon 9 launch records HTML table with BeautifulSoup
- Parse HTML table and convert it into Pandas data frame
- Link for Notebook from Github:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20Web%20scraping.ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) was performed, and Training Labels was determined.



- Link for Notebook from Github:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

Loaded SpaceX Dataset

Calculated the number of lauches on each site

Calculated the number and occurrence of each orbit

Created a landing outcome label from Outcome column

EDA with SQL

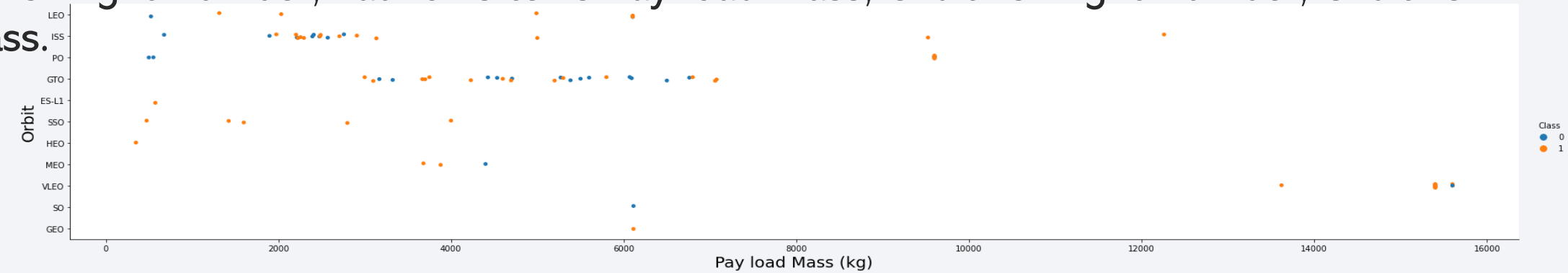
- The SpaceX dataset was loaded into SQLite database.
- The following SQL queries were performed:
  - The names of the unique launch sites in the space mission;
  - 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA(CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 and less than 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

- Link for Notebook from Github:

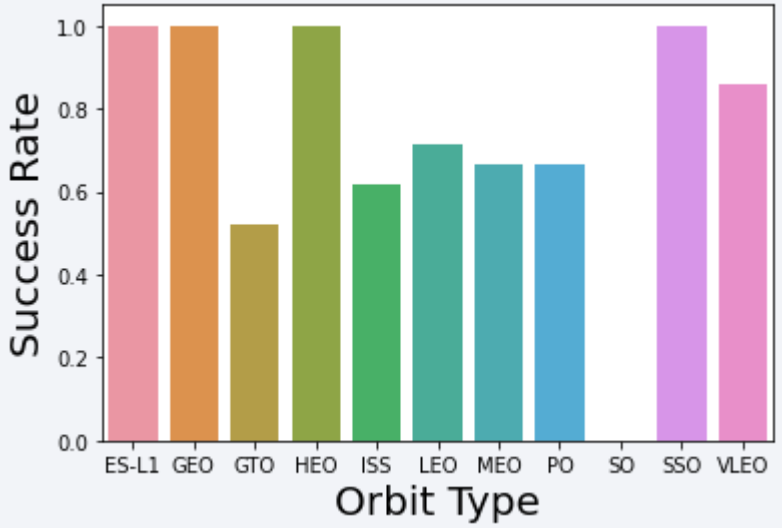
https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

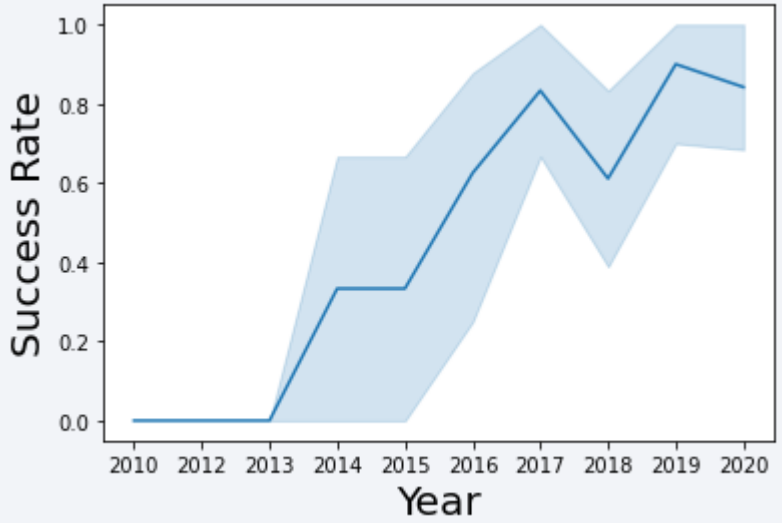
EDA with Data Visualization

- Scatter plot was used to visualize the relationship between: Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Pay load Mass, Orbit vs Flight Number, Orbit vs Pay load Mass.



- Bar plot was used to visualize the success rate of each orbit.
- Line plot was used to visualize the success rate of each year.
- Link for Notebook from Github:





https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

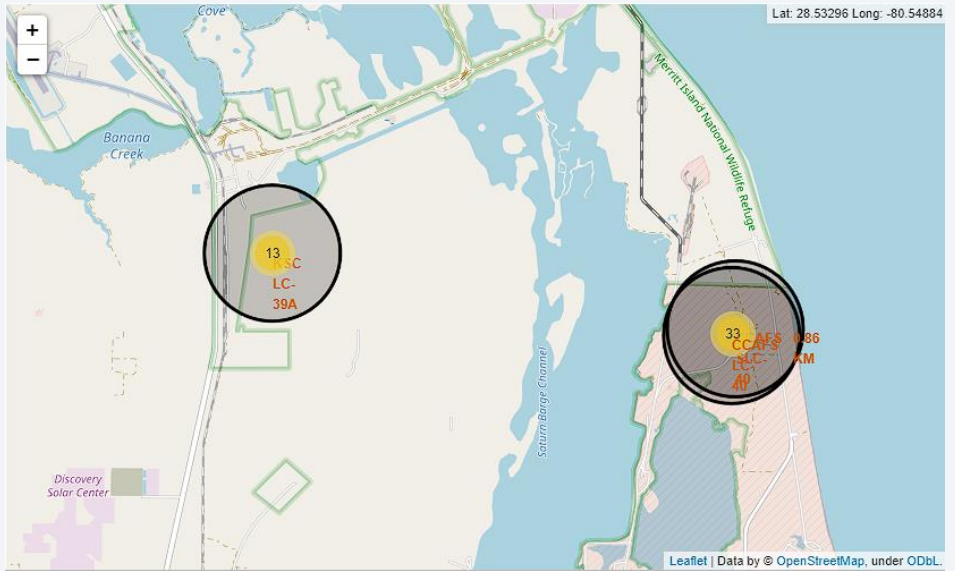
Build an Interactive Map with Folium

- All SpaceX launch sites were marked on map with different objects such as with markers, circles and lines.
  - Markers indicate launch site locations;
  - Circles indicate areas around specific coordinates;
  - Lines indicate distances between two coordinates.
- The success/failed launches for each site on the map were marked.
- The distances between a launch site to its proximities were calculated.
- Link for Notebook from Github:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

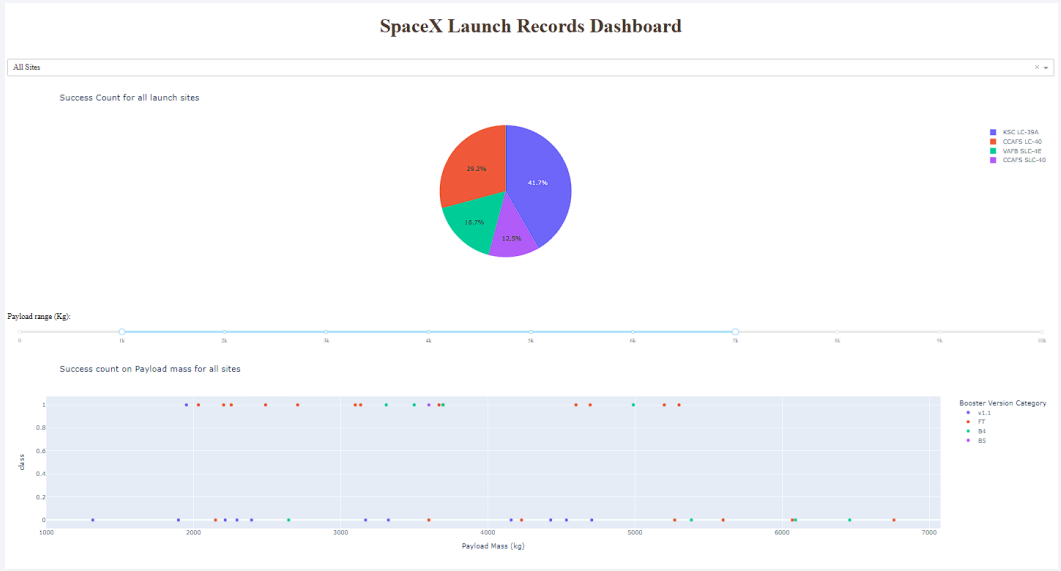
Some interactions at:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/tree/main/Interactive%20Visual%20Analytics%20with%20Folium



Build a Dashboard with Plotly Dash

- Interactive Dashboard was built with Plotly Dash
  - Added a Launch Site Drop-down input component
  - Added a callback function to render `success-pie-chart` pie chart based on selected Launch Site Drop-down
  - Added a Range Slider to select Payload range (kg)
  - Added a callback function to render the `success-payload-scatter-chart` scatter plot (Class vs Payload Mass (kg) for different booster version categories).
- Link for Notebook from Github:



https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

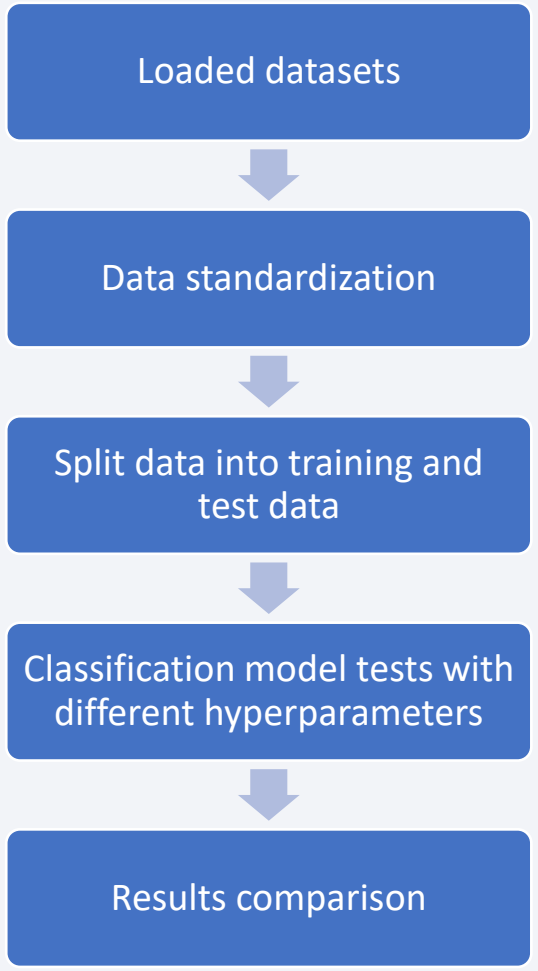
Some interactions at:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/tree/main/Interactive%20Dashboard%20with%20Plotly%20Dash

Predictive Analysis (Classification)

- Two datasets were loaded into Pandas Dataframe (`data` and `X` data frames).
- The column `Class` from `data` was assigned to a Numpy array (`Y` array).
- The data in `X` was standardized and transformed.
- The data `X` and `Y` were split into training and test data.
- Logistic Regression, Support Vector Machine (SVM), Decision Tree classifier and K-Nearest Neighbors classification models were applied using GridSearchCV and the respective accuracies on the test data were calculated.
- The best classification method performed was found.
- Link for Notebook from Github:

https://github.com/hmmurata/IBM-Applied-Data-Science-Capstone/blob/main/Predictive%20Analysis%20-%20Classification.ipynb



Results

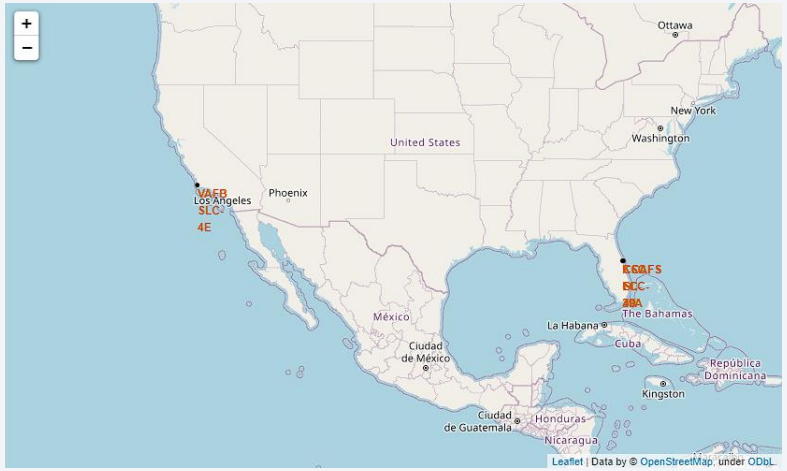
- Exploratory Data Analysis (EDA) with SQL results
  - The total payload mass carried by boosters launched by NASA (CRS) was 45,596 kg.
  - The average payload mass carried by booster version F9 v1.1 was 2,928.4 kg.
  - The first successful landing outcome in ground pad was achieved in 22-12-2015.
  - The boosters which have success in drone ship and have payload mass greater than 4,000 kg but less than 6,000 kg were: F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.
  - The total number of successful and failure mission outcomes was 101.
  - There were 12 different booster versions which have carried the maximum payload mass.

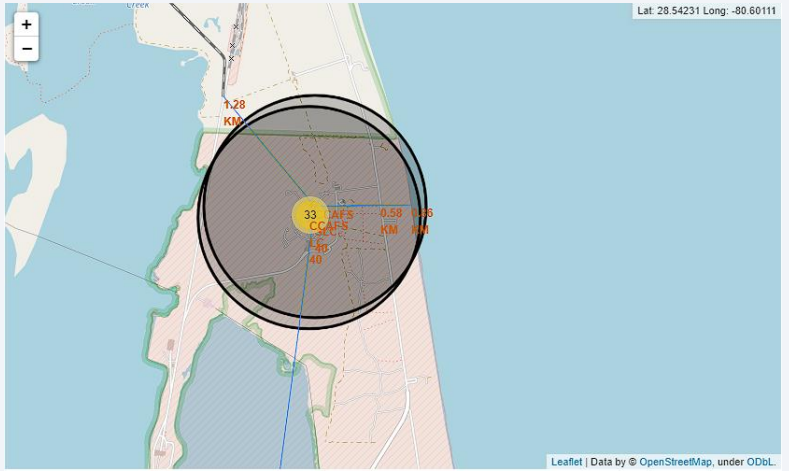
Results

- Exploratory Data Analysis (EDA) with Data Visualization results
  - We see that as the flight number increases, the first stage is more likely to land successfully.
  - We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %,while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - VAFB-SLC launch site there are no rockets launched for heavy pay load mass (greater than 10000).
  - ES-1, GEO, HEO, SSO and VLEO orbits have high success rate.
  - In the 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.
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - The success rate since 2013 kept increasing till 2020

Results

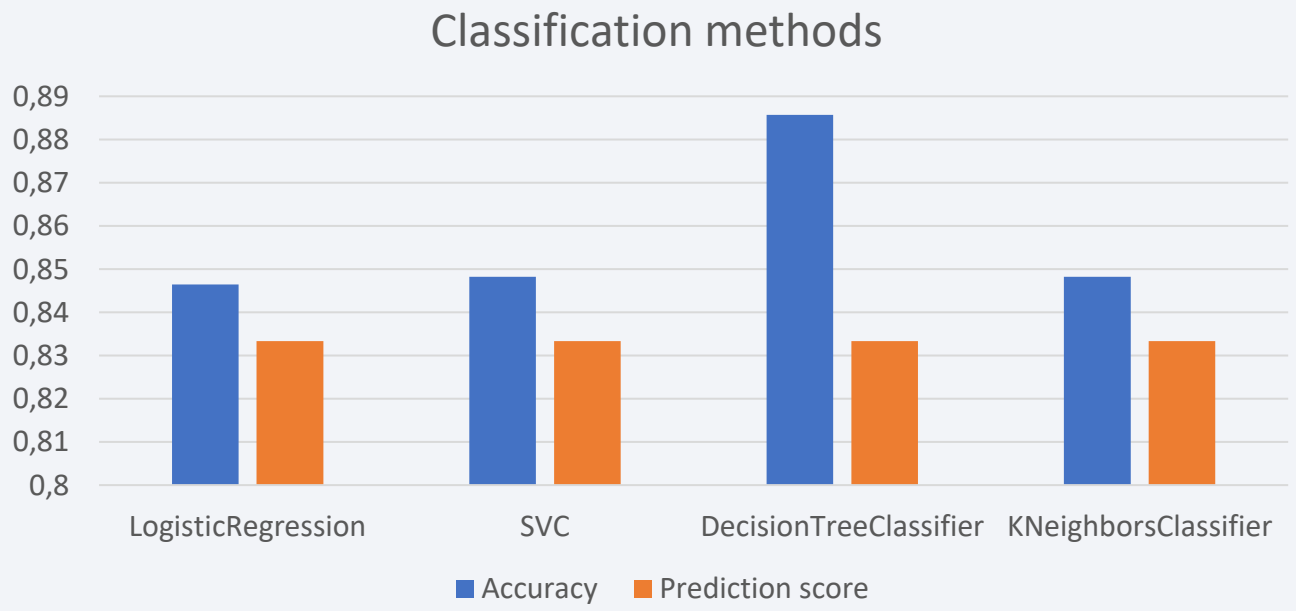
- Interactive analytics demo in screenshots
  - Was possible to identify launch sites near to coastline with certain distance away from big cities for safety
  - Good transport proximity as railway and highway for people and heavy cargo transport respectively.
  - Launch sites are near equator to minimize fuel consumption.

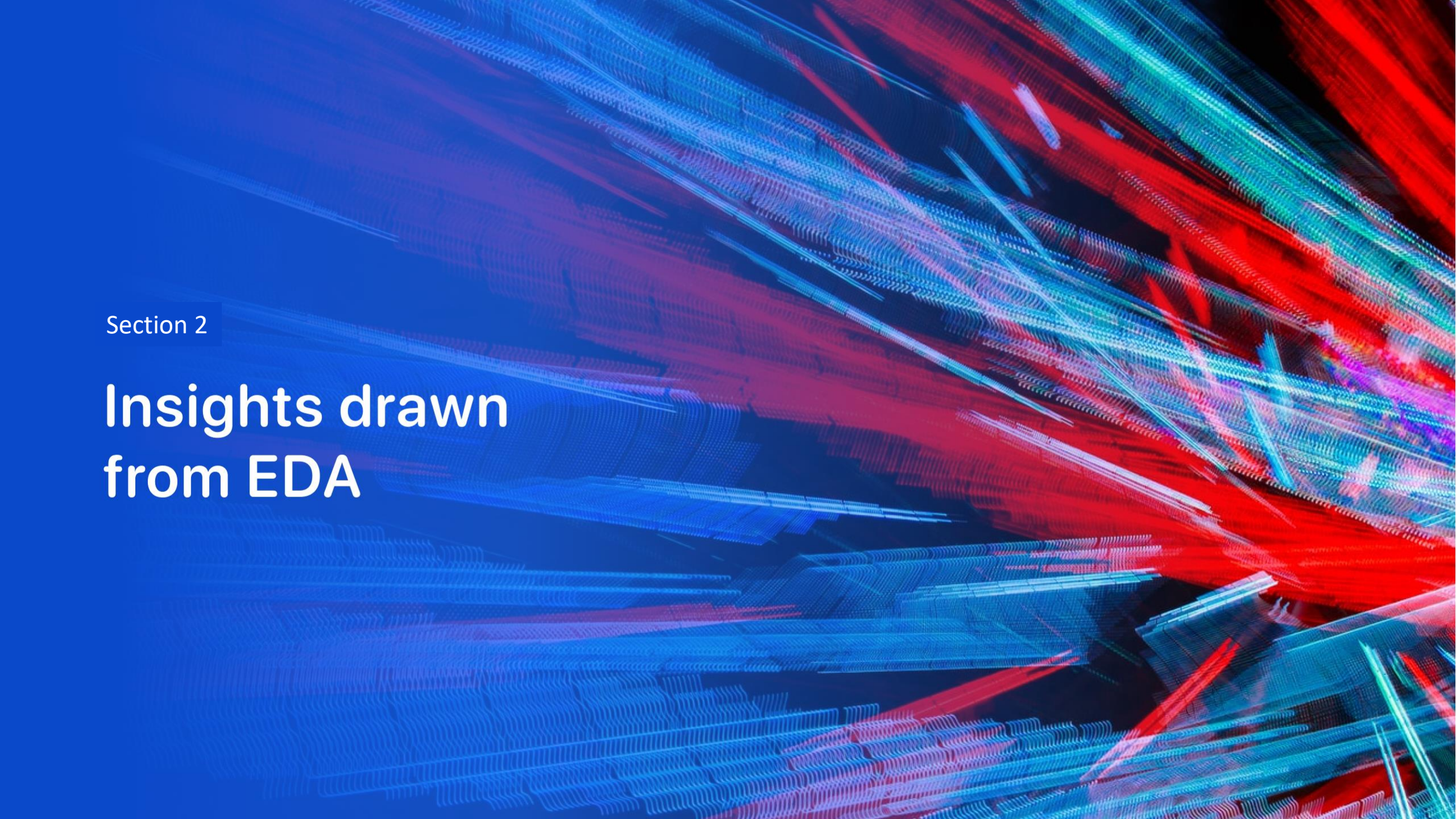




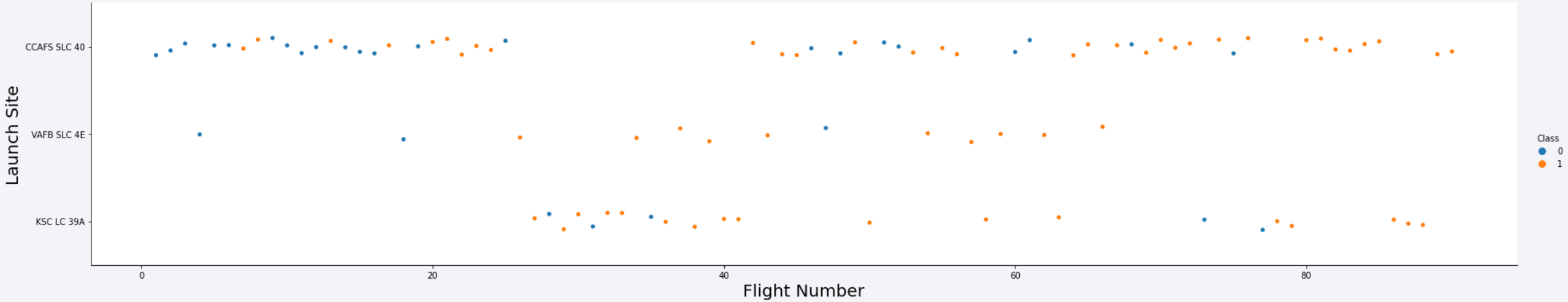
Results

- Predictive analysis results
  - The best classification method performed with SpaceX datasets was the Tree Decision Classifier with accuracy over approximately 88.6 % and prediction score over approximately 83.3 %.



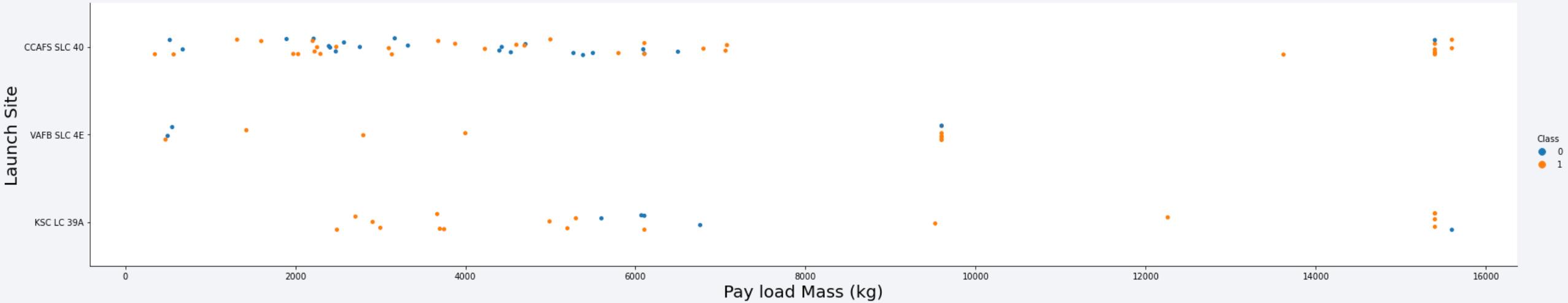

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The best launch site was CCAF5 SLC 40.
- The most recent launches were most successful.

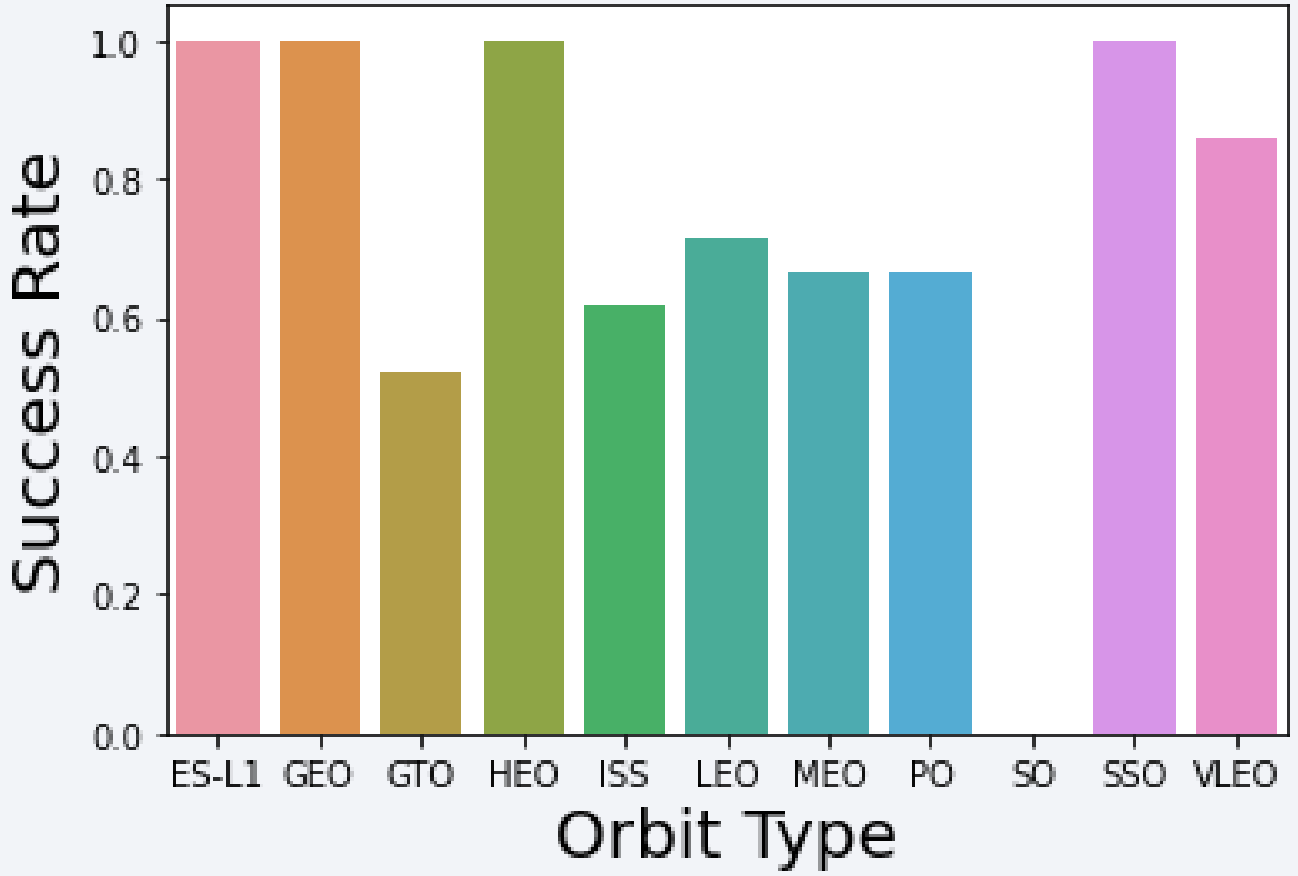
Payload vs. Launch Site



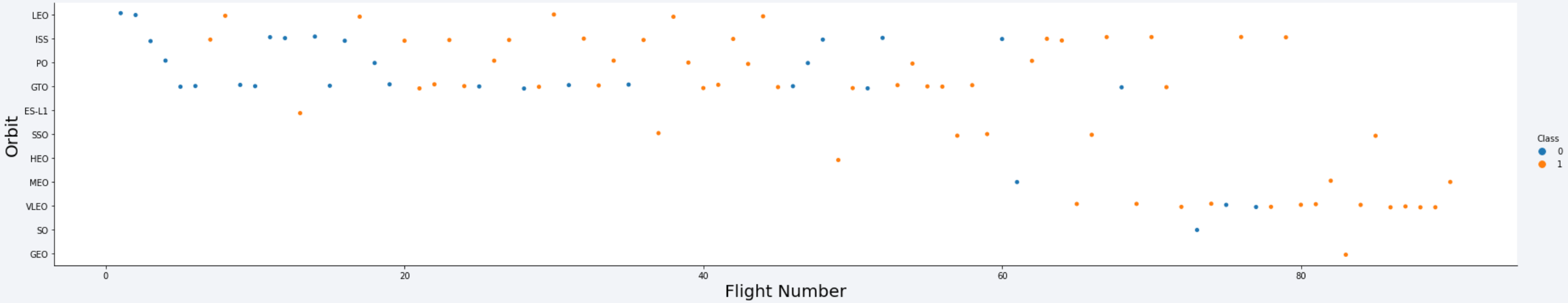
- Payload mass over 9,000 kg has most successful rate.
- The VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10,000 kg).

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits have the highest success rate.



Flight Number vs. Orbit Type



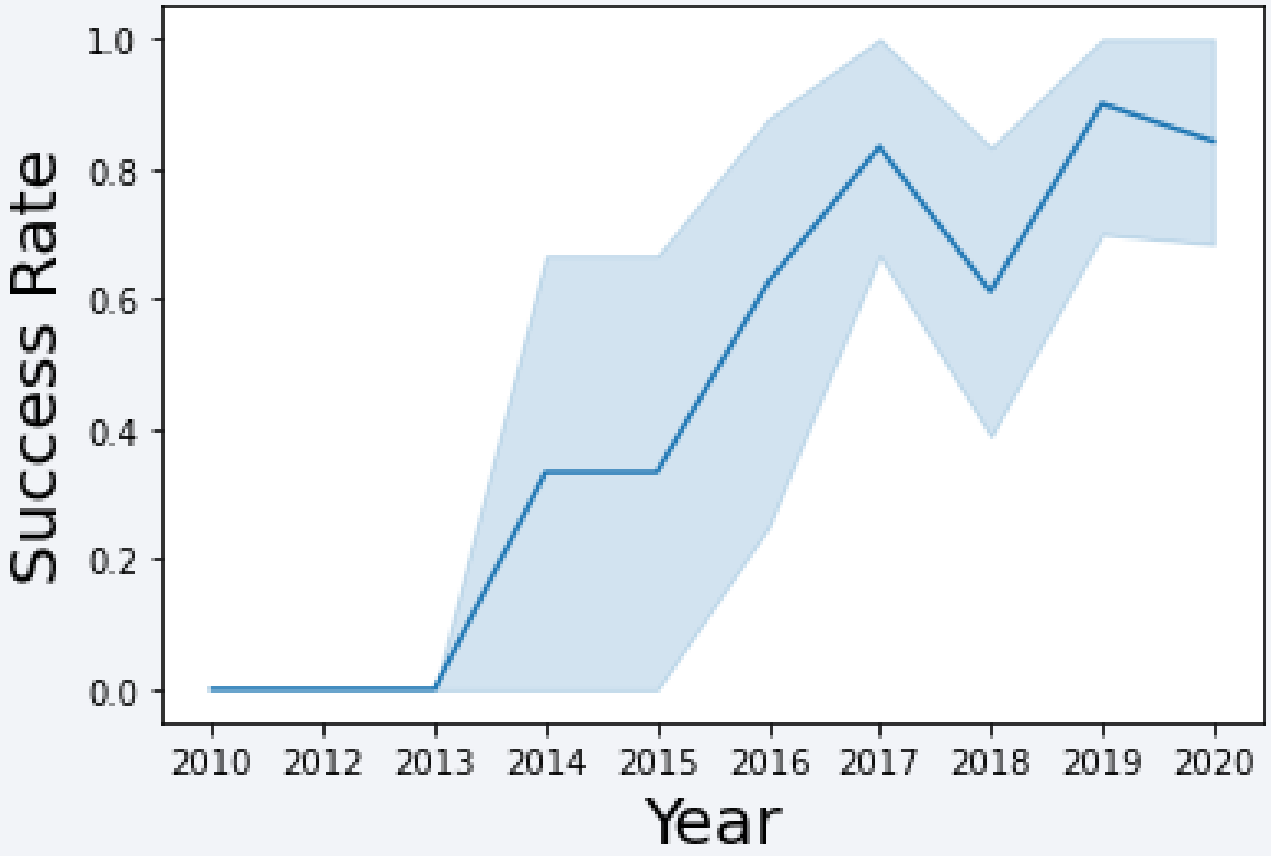
- In the LEO orbit the success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit. But in general, success rate improved over time for other orbits.

Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate were more for PO, LEO and ISS orbits.
- For GTO orbit, there were not a relationship with payload mass and success rate.
- There were few data for SO and GEO orbits related to the payload mass.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020
- There was no success rate data before 2013.



All Launch Site Names

- Query used:

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

- In SQLite, the unique name of launch site is given selecting DISTINCT occurrences from SPACEXTBL table.
- Results:

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

Launch Site Names Begin with 'CCA'

- Query used:

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

- Results:

Date	Time (UTC)	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 result is 5 records with information about launch site names begin with ‘CCA’
- There are 3 “no attempt” landing outcomes.

Total Payload Mass

- Query used:
```
%sql select sum(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL where CUSTOMER='NASA (CRS)';
```
- Result:

payload_mass
45596

- The total payload carried by boosters from NASA (CRS) was calculated by the sum function all payloads for NASA (CRS).

Average Payload Mass by F9 v1.1

- Query used:

```
%sql select avg(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL where BOOSTER_VERSION='F9 v1.1';
```

- Result:

payload_mass
2928.4

- The average payload mass carried by booster version F9 v1.1 was obtained by avg function

First Successful Ground Landing Date

- Query:

```
%sql select min(DATE) from SPACEXTBL where "LANDING _OUTCOME" = 'Success (ground pad)' and DATE > date('now','localtime');
```

- Result:

min(DATE)
22-12-2015

- In SQLite the column DATE is a text type. For use with min function was necessary to use the condition DATE > date('now', 'localtime')

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
%sql select BOOSTER_VERSION from SPACEXTBL where "LANDING _OUTCOME"='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

- Results:

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

- There were 4 booster versions as successful drone ship landing

Total Number of Successful and Failure Mission Outcomes

- Query:

```
%sql select count(MISSION_OUTCOME) as mission_outcomes from SPACEXTBL;
```

- Results:

mission_outcomes
101

- The total number of successful and failure mission outcomes was 101.

Boosters Carried Maximum Payload

- Query:

```
%sql select distinct(BOOSTER_VERSION) from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

- Results:

Booster_Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

- There were 12 results of booster versions which have carried the maximum payload mass.
- A subquery was used to select the maximum payload value. This value was used to select a unique booster version.

2015 Launch Records

- Query:

```
%sql select substr(Date, 4, 2) as month,"LANDING _OUTCOME",BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where substr(Date,7,4)='2015'
and "LANDING _OUTCOME"='Failure (drone ship)';
```

- Results:

month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- SQLite does not support monthnames, substr(Date, 4, 2) as month was used to get the months and substr(Date,7,4)='2015' for year.

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

- Query:

```
%sql select "LANDING _OUTCOME", count("LANDING _OUTCOME") as count from SPACEXTBL where
DATE(substr(DATE,7,4)||'-'||substr(DATE,4,2)||'-'||substr(DATE,1,2)) BETWEEN DATE('2010-06-04') AND DATE('2017-03-20')
GROUP BY "LANDING _OUTCOME" order by count desc;
```

- Results:

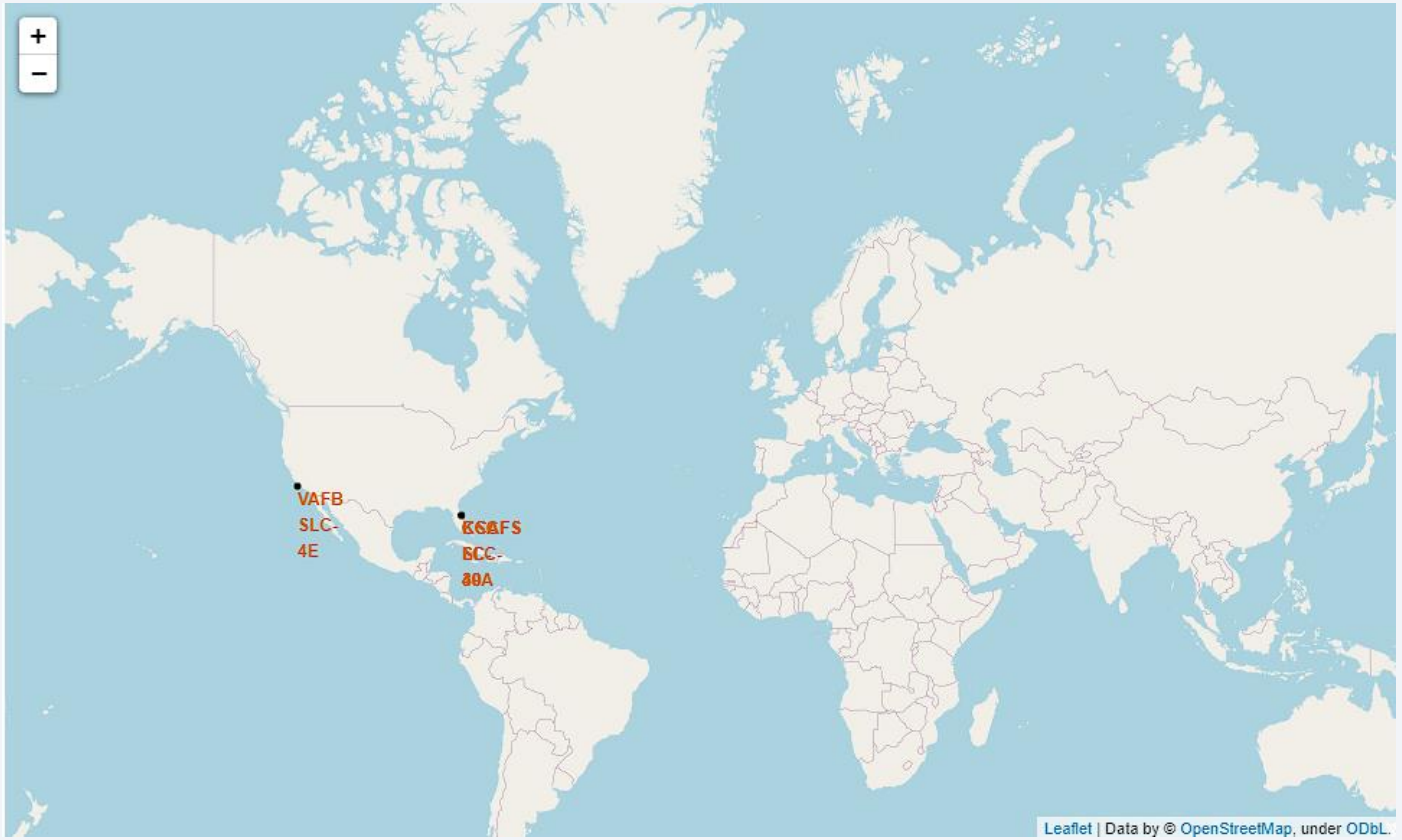
Landing _Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- SQLite does not support date. The function DATE was used to convert a date with format %Y-%m-%d with (substr(DATE,7,4)||'-'||substr(DATE,4,2)||'-'||substr(DATE,1,2)). The clause GROUP BY was used to group landing outcomes and ORDER BY count DESC was used to show the counts in descending order.

Section 3

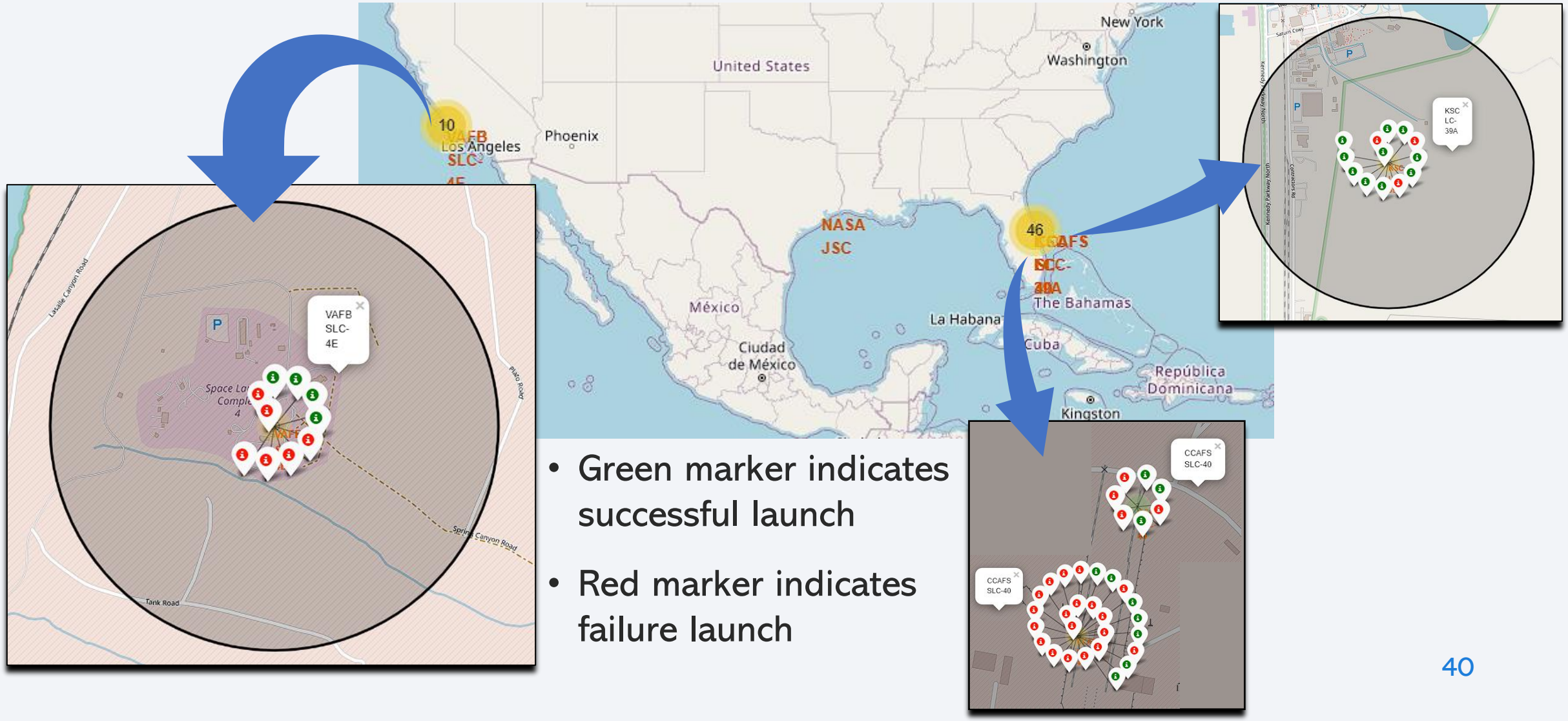
Launch Sites Proximities Analysis

All launch sites' location marker on global map



- SpaceX launch sites are in the USA near to coastline, highways and railroads.

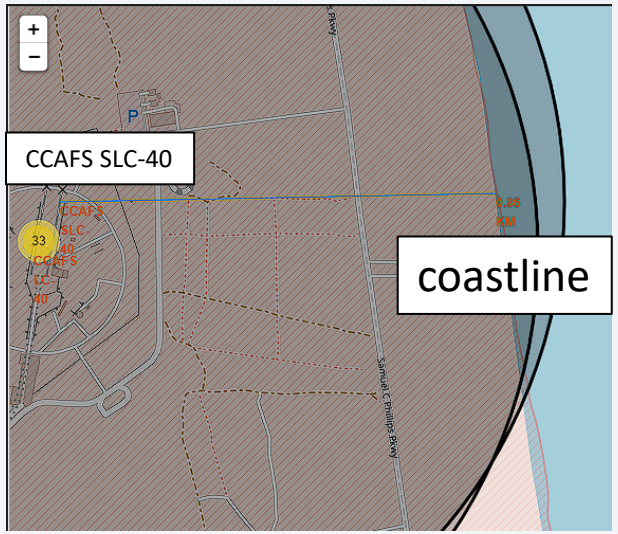
Launch Outcomes

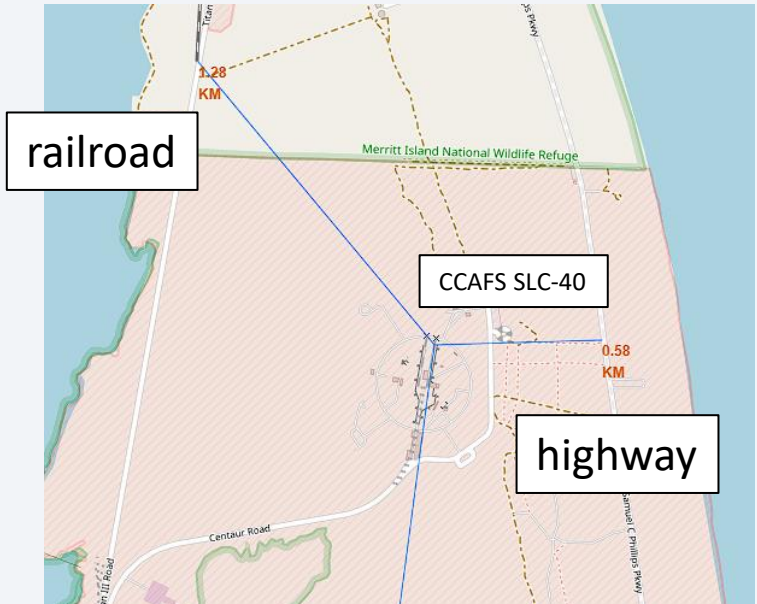


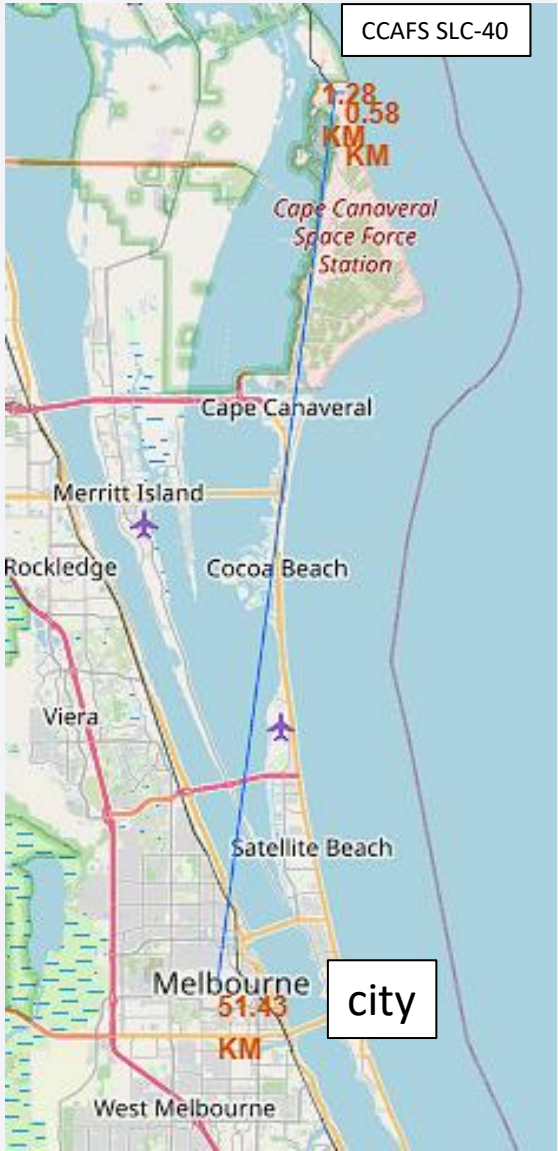
- Green marker indicates successful launch
- Red marker indicates failure launch

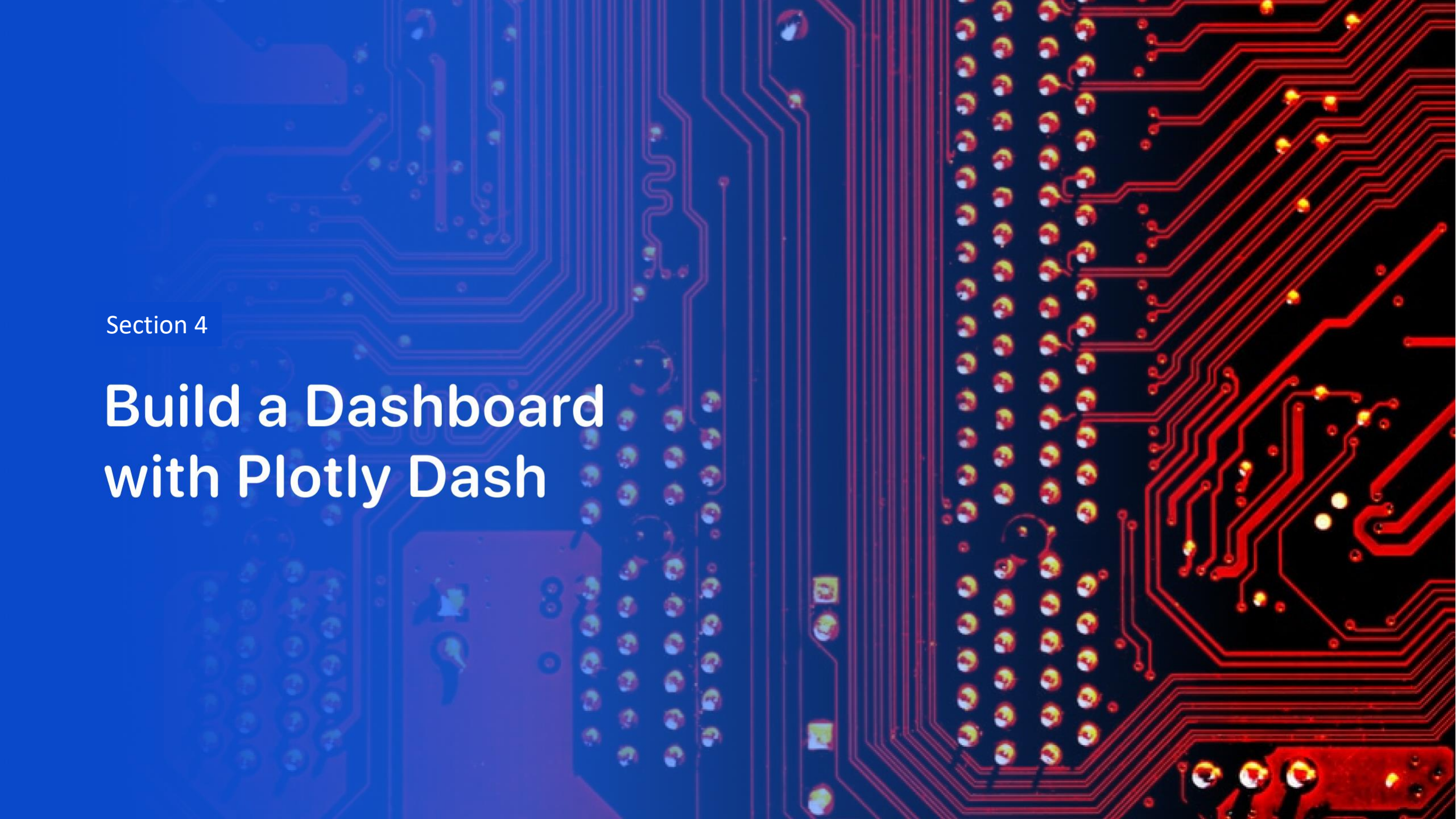
Launch site proximities

- Example of launch site proximities:
- Distance between CCAFS SLC-40 and:
  - Coastline: 0.86 km
  - Highway: 0.58 km
  - Railroad: 1.28 km
  - City: 51.43 km

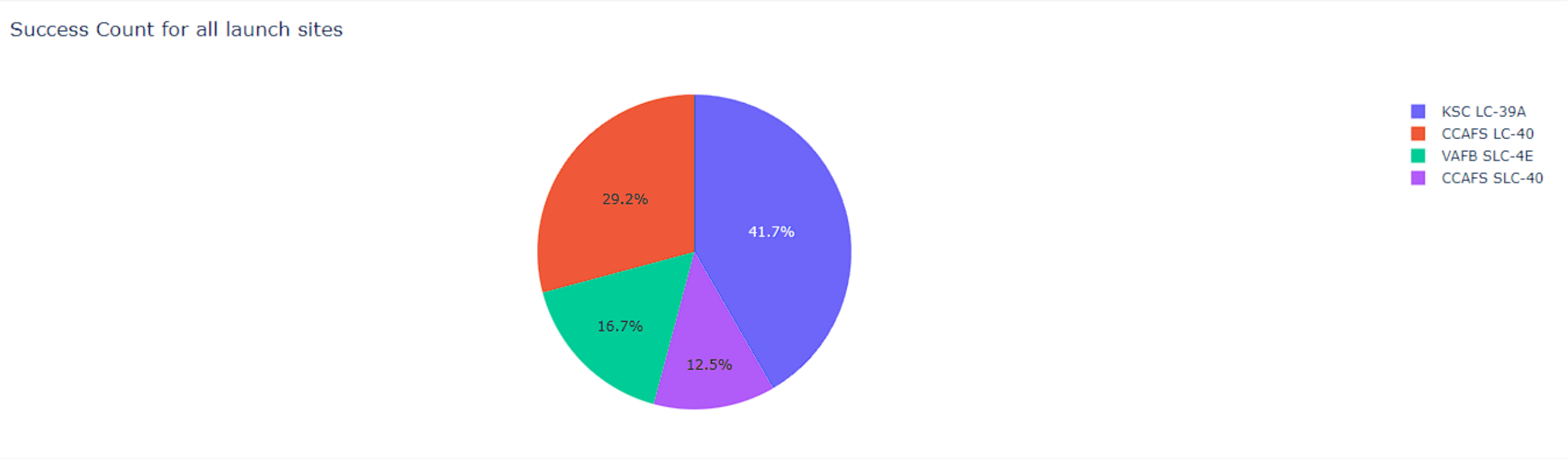






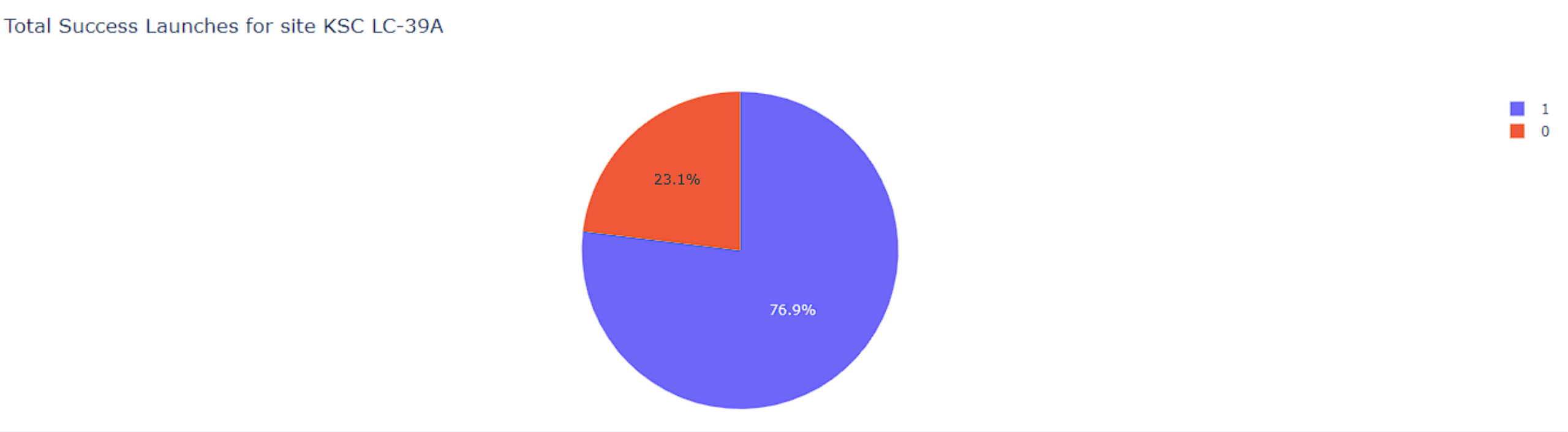

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



- KSC LC-39A launch site has the most successful launches

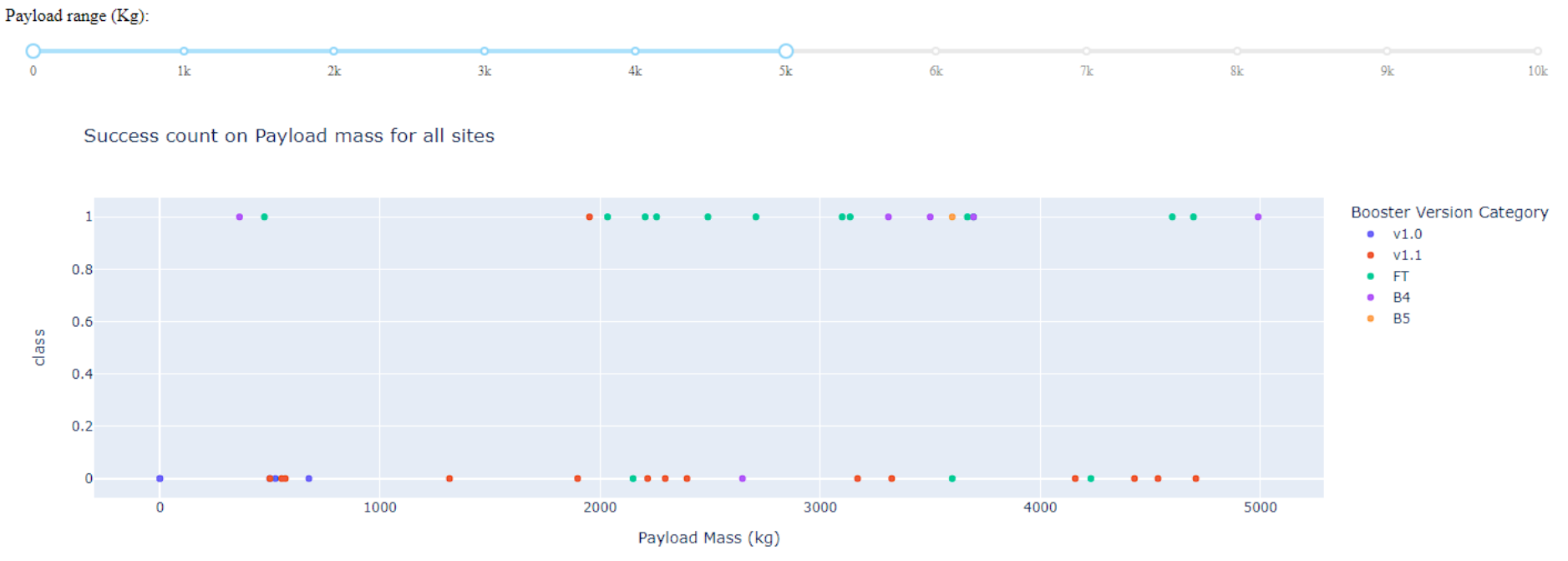
Launch site with highest launch success



- The KSC LC-39A launch site has the highest launch success with 76.9%

Payload vs. Launch Outcome for all sites, with different payload

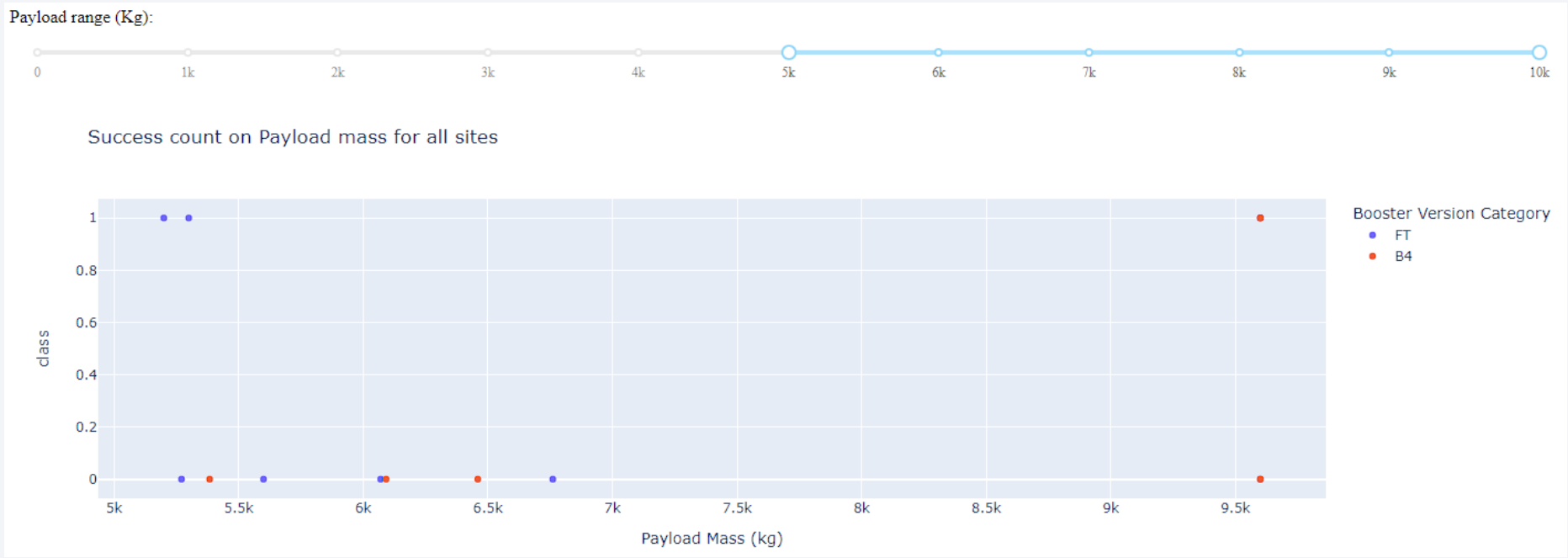
- Payload range: 0 to 5,000 kg



- Highest success rate for FT Booster version with payload mass between 2,000 and 5,000 kg

Payload vs. Launch Outcome for all sites, with different payload

- Payload range: 5,000 to 10,000 kg

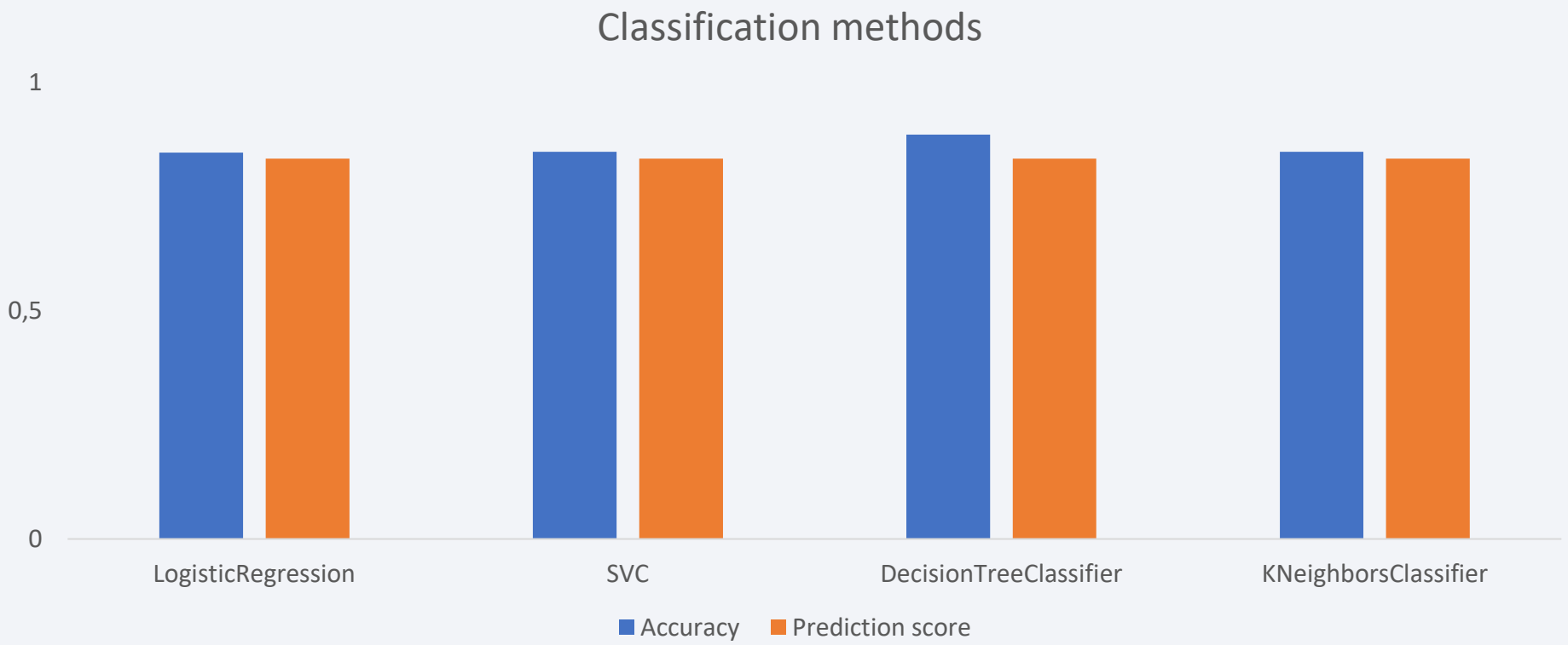


- Low success rate for FT Booster version with heavy weighted payload mass

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The model with most accuracy is the Decision Tree Classifier with approximately 88%

Confusion Matrix

- The confusion matrix for tree decision classifier shows that can distinguish between the different classes. The major problem is false positives.

Conclusions

- The launch success rate since 2013 kept increasing till 2020
- ES-L1, GEO, HEO and SSO orbits have the highest success rate
- KSC LC-39A launch site has the most successful launches
- An optimal location for building a launch site certainly involves many factors and some of the factors can be found by analyzing the existing launch site locations.
- Decision Tree Classifier is the best model to predict if the first stage will land given the data from SpaceX.

Appendix

- Some differences between Db2 and SQLite queries:
  - Db2:
    - %sql select EXTRACT(MONTH FROM DATE) as month,LANDING__OUTCOME,BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015' and LANDING__OUTCOME='Failure (drone ship)';
  - SQlite:
    - %sql select substr(Date, 4, 2) as month,"LANDING _OUTCOME",BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where substr(Date,7,4)='2015' and "LANDING _OUTCOME"='Failure (drone ship)';
  - Db2:
    - %sql select unique(BOOSTER_VERSION) from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
  - SQLite:
    - %sql select distinct(BOOSTER_VERSION) from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

Thank you!