IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ilir Snopche
February 14, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of methodologies**
- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

- **Summary of all results**
- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result from Machine Learning Lab

Introduction

- **Project background and context**
  - Space X 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 Space X can reuse the first stage. Therefore, in order to determine the cost of a launch, it is crucial to determine if the first stage will land successfully. The goal of this project is to create a machine learning pipeline to predict the landing outcome of the first stage in the future. This would be useful for any company that would like to bid against space X for a rocket launch.
- **Problems we want to find answers**
  - Determine which features influence the landing outcome.
  - Determine the relationship between various features (variables) and how these features affect the landing outcome.
  - Identify which conditions will increase the probability for successful landings of the first stage of rockets.
  - Determine where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - Data was collected using:
    - Space X API: https://api.spacexdata.com/v4/rockets/
    - Web Scraping:

      https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- **Perform data wrangling**
  - For categorical features one-hot encoding was applied.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - How to build, tune, evaluate classification models

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia, using web scraping

(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- In case of the SpaceX API, we collected the data using get request. Then we used .json() function call to decode the response content as a Json and used json_normalize() to turn it into a pandas dataframe. Next we cleaned the data and dealt with the missing values.

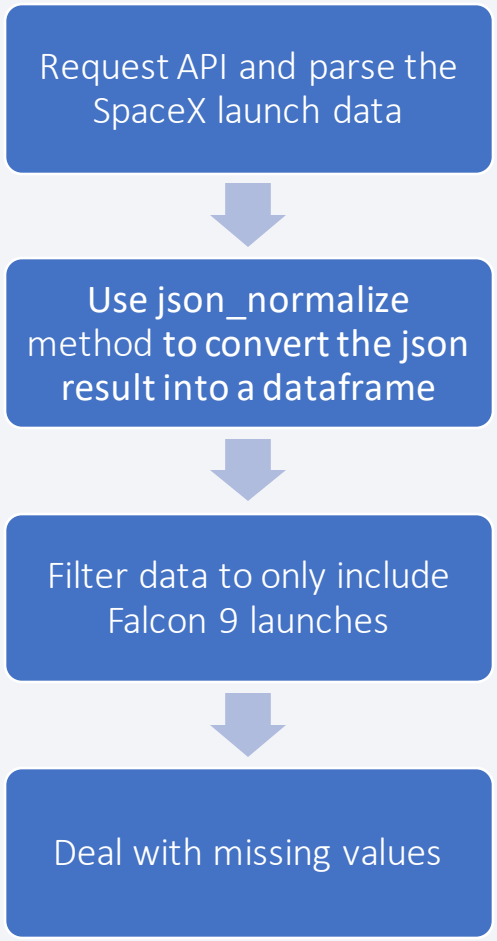
We used BeautifulSoup for web scraping; the objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for further analysis.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and used. We used this API according to the flowchart on the right.

- **Source code:**

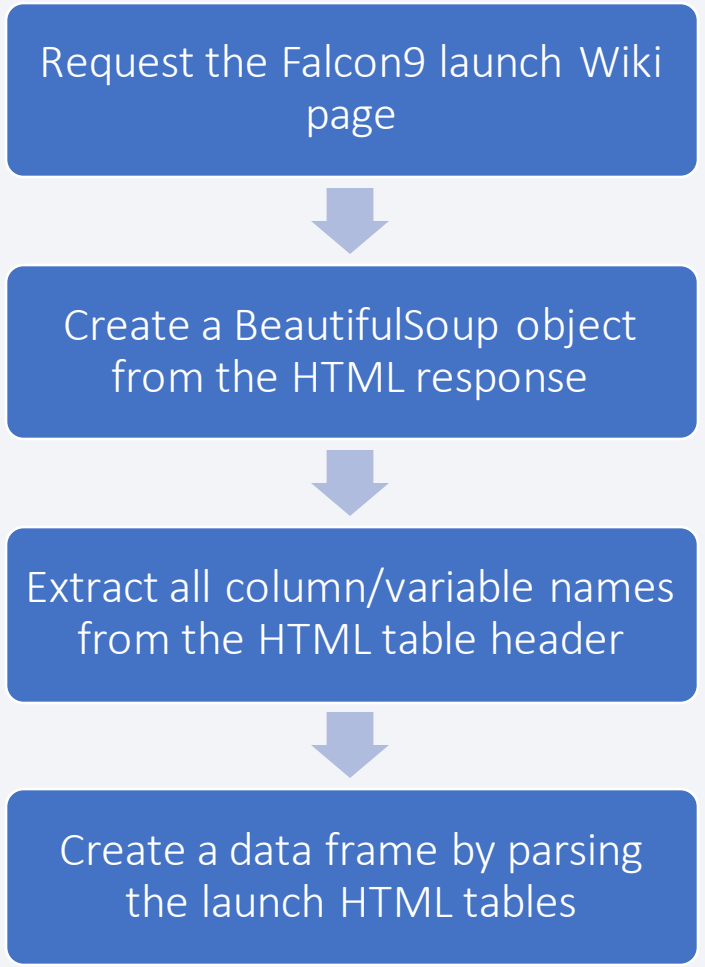
https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/spacex_api-data-collection_snopche.ipynb



Data Collection - Scraping

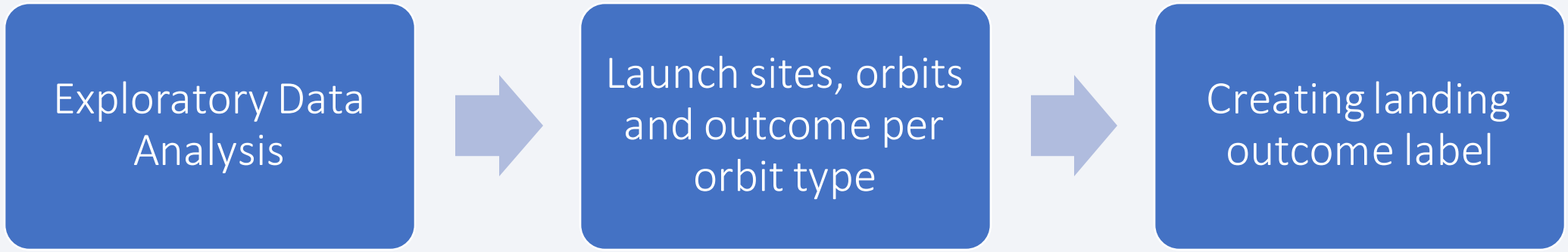
- Data about SpaceX launches one can obtain from Wikipedia as well. We downloaded data from Wikipedia according to the flowchart on the right.
- **Source code:**

https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/web scraping_data-collecction_snopche.ipynb



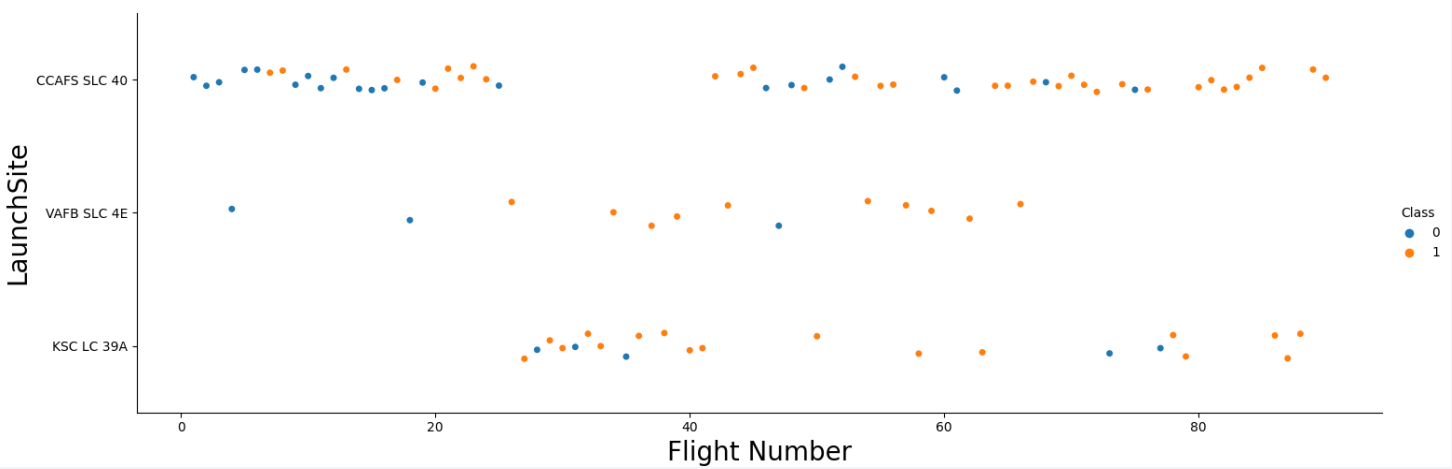
Data Wrangling

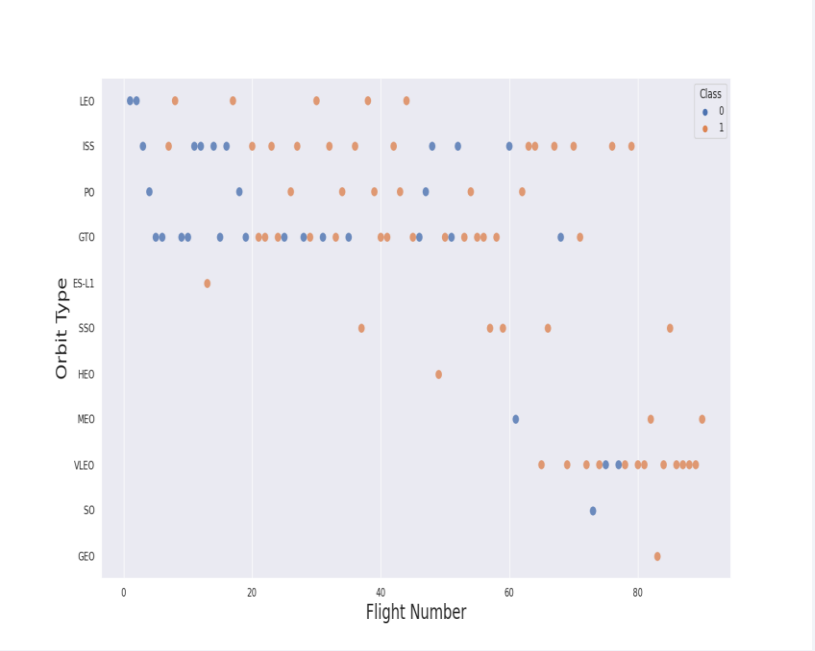
- Initially we performed on the dataset some Exploratory Data Analysis.
- Next we calculated the number of launches at each site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Finally, a landing outcome label was created from Outcome column.



Sorce code: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/Data_Wrangling_Snopche.ipynb

EDA with Data Visualization





We used scatter plots to observe and visualize the relationship between different attributes, such as between:

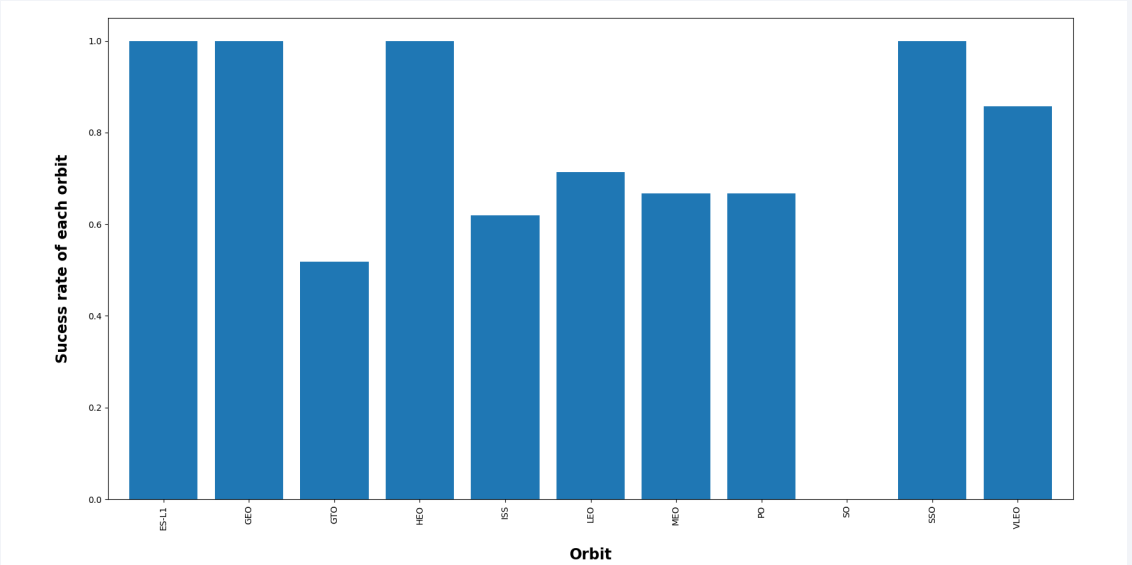
- Payload and Flight Number.
- Flight Number and Launch Site.
- Payload and Launch Site.
- Flight Number and Orbit Type.
- Payload and Orbit Type.

Source: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/EDA%20with%20Data%20Visualization_snopche.ipynb

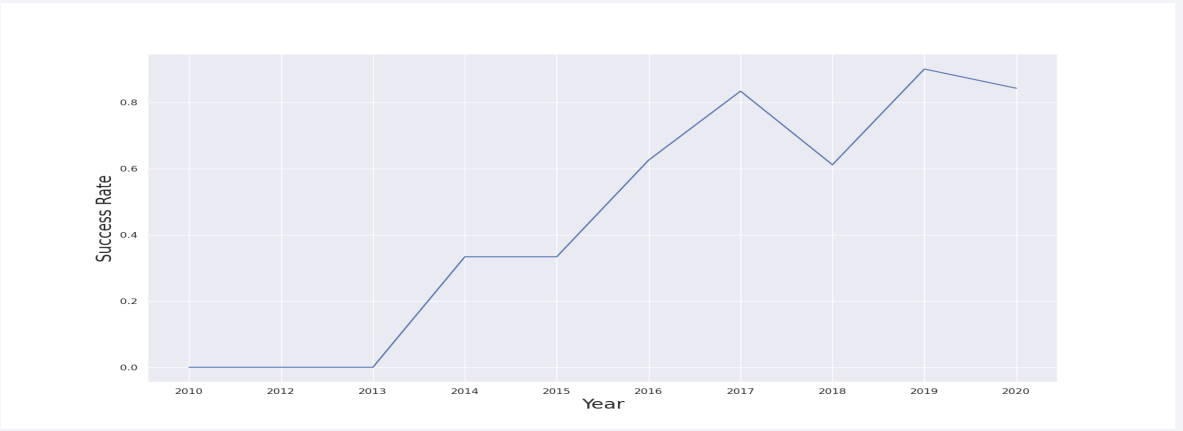
EDA with Data Visualization

In order to check if there are any relationship between success rate and orbit type we used a bar chart.

Source: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/EDA%20with%20Data%20Visualization_snopche.ipynb



In order to visualize the launch success yearly trend we used a line chart.



EDA with SQL

The following SQL queries were performed:

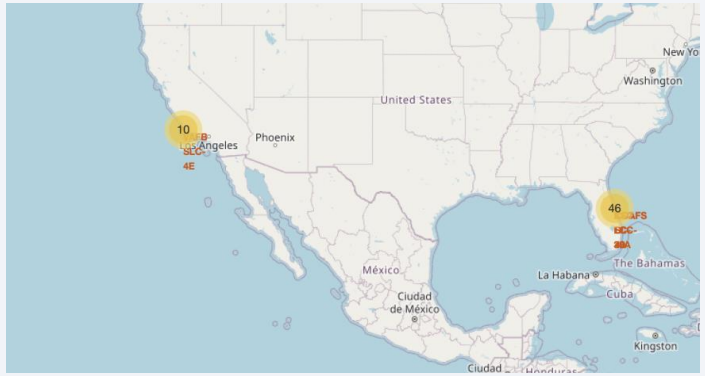
- Displaying the names of the unique launch sites in the space mission;
- Displaying 5 records where launch sites begin with the string 'CCA';
- Displaying the total payload mass carried by boosters launched by NASA (CRS);
- Displaying average payload mass carried by booster version F9 v1.1;
- Listing 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 between 4000 and 6000 kg;
- 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, launch sites names for the months in 2015; and
- 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.

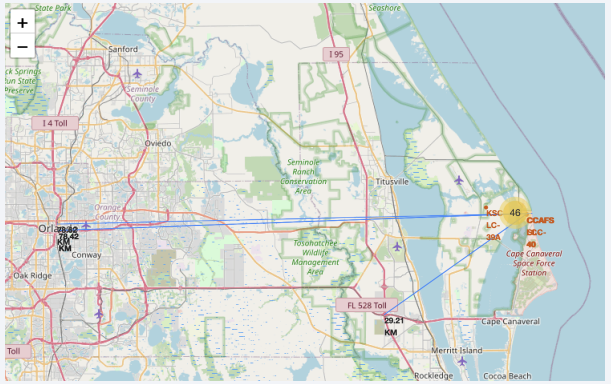
Source code:

https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/EDA%20with%20SQL_snopche.ipynb

Build an Interactive Map with Folium

- We marked all launch sites on a map by adding a circle and a marker around each launch site with a label of the name of the launch site.
- Next we marked the success/failed launches for each site on the map. To mark a successful launch we used a *green marker,* and we used a *red marker* if a launch was failed.
- Finally, we drew a line from a launch site to a selected closest coastline point as well as to its closest city, railway, highway, and we calculated the respective distances.





Source code: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/Interactive%20visual%20analytics%20with%20folium_snopche.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly Dash. It contains a Pie-chart showing the total launches for all sites. Moreover, there is a site-dropdown which allows us for each launch site to choose a pie chart visualizing launch success counts.
- We also plotted a scatter graph showing the correlation between Payload Mass (kg) for the different booster version and Outcome. It contains a Range Slider which allows us easily to select different payload range and see if we can identify some visual patterns.

Source code: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/spacex_dash_app_snopche.py

Predictive Analysis (Classification)

- We built and compared four classification models: Logistic Regression, Support Vector Machine, Decision Tree and k-Nearest Neighbors.

Building the model

- Load the dataset into Pandas and create Numpy arrays
- Standardize the data
- Split the data into training and test datasets
- Build each model

Evaluating the model

- Tune different hyperparameters using GridSearchCV
- Determine the best parameters
- Calculate the accuracy using the method *score*
- Plot the confusion matrix

Comparison of results and determining the best model

Source code: https://github.com/ilirsnopche/Applied-Data-Science-Capstone-SpaceX_Rocket_Launch/blob/main/Machine_Learning_Prediction_snopche.ipynb

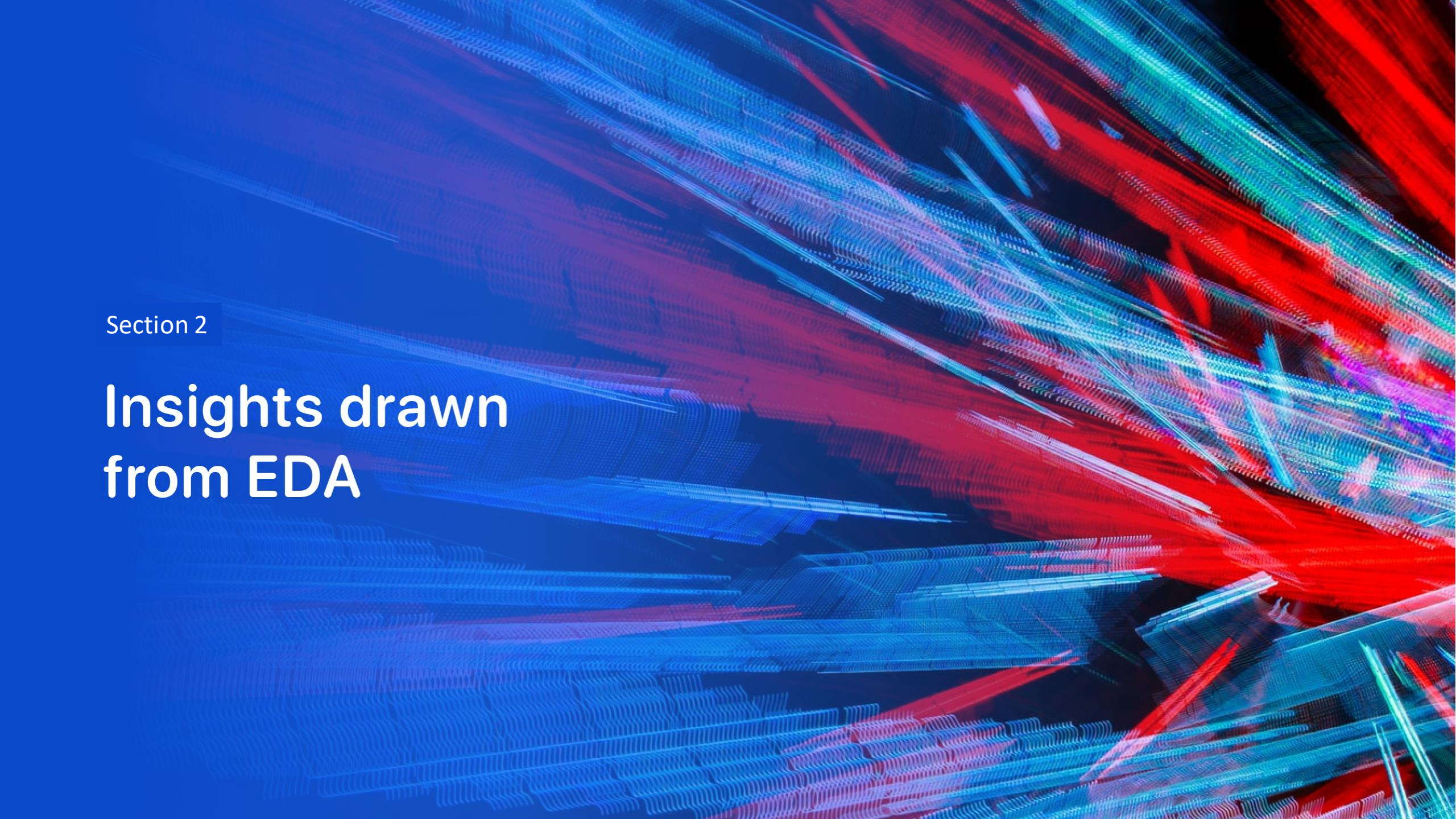
Results

Our results will be categorized in three groups:

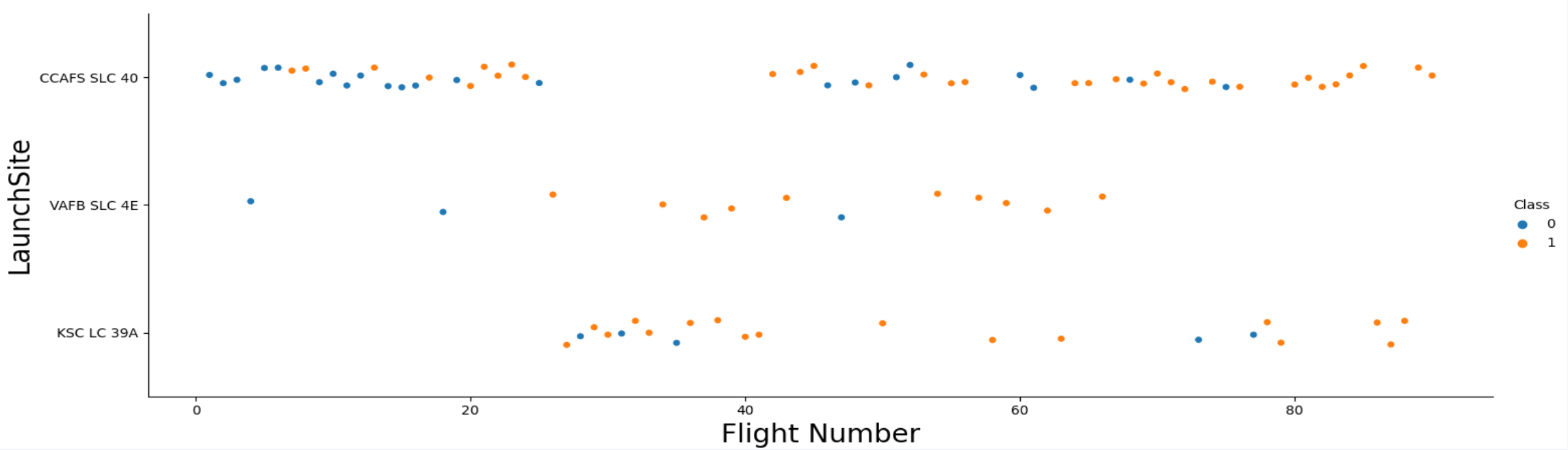
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

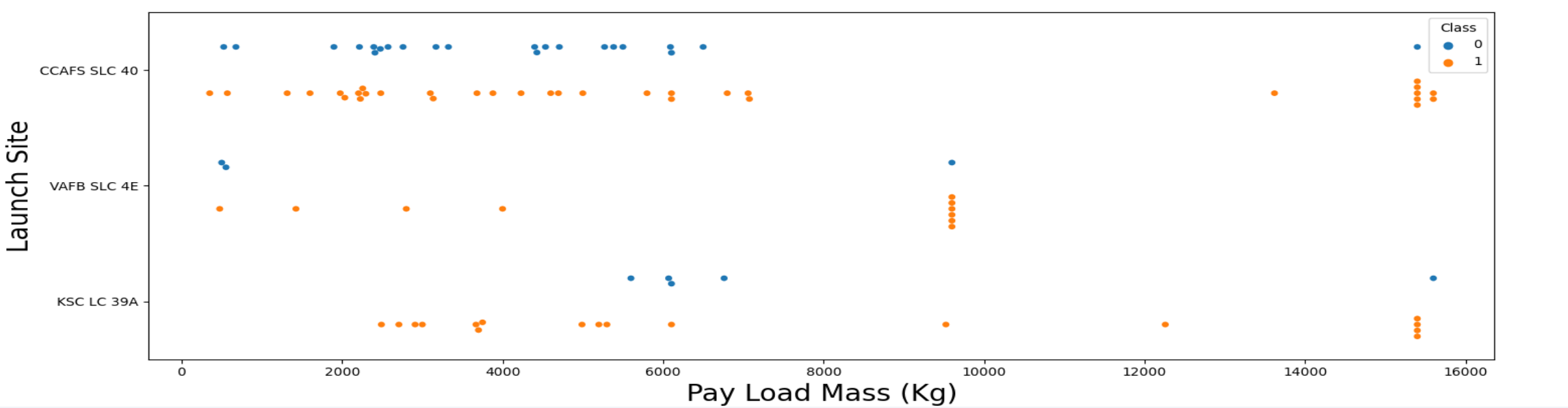


Flight Number vs. Launch Site



- From this scatter plot we can conclude that as the flight number increases, the first stage is more likely to land successfully. Hence, the general success rate improved over time.
- The plot also indicates that the most popular launch site nowadays is CCAF5 SLC40, where most of recent launches were successful.

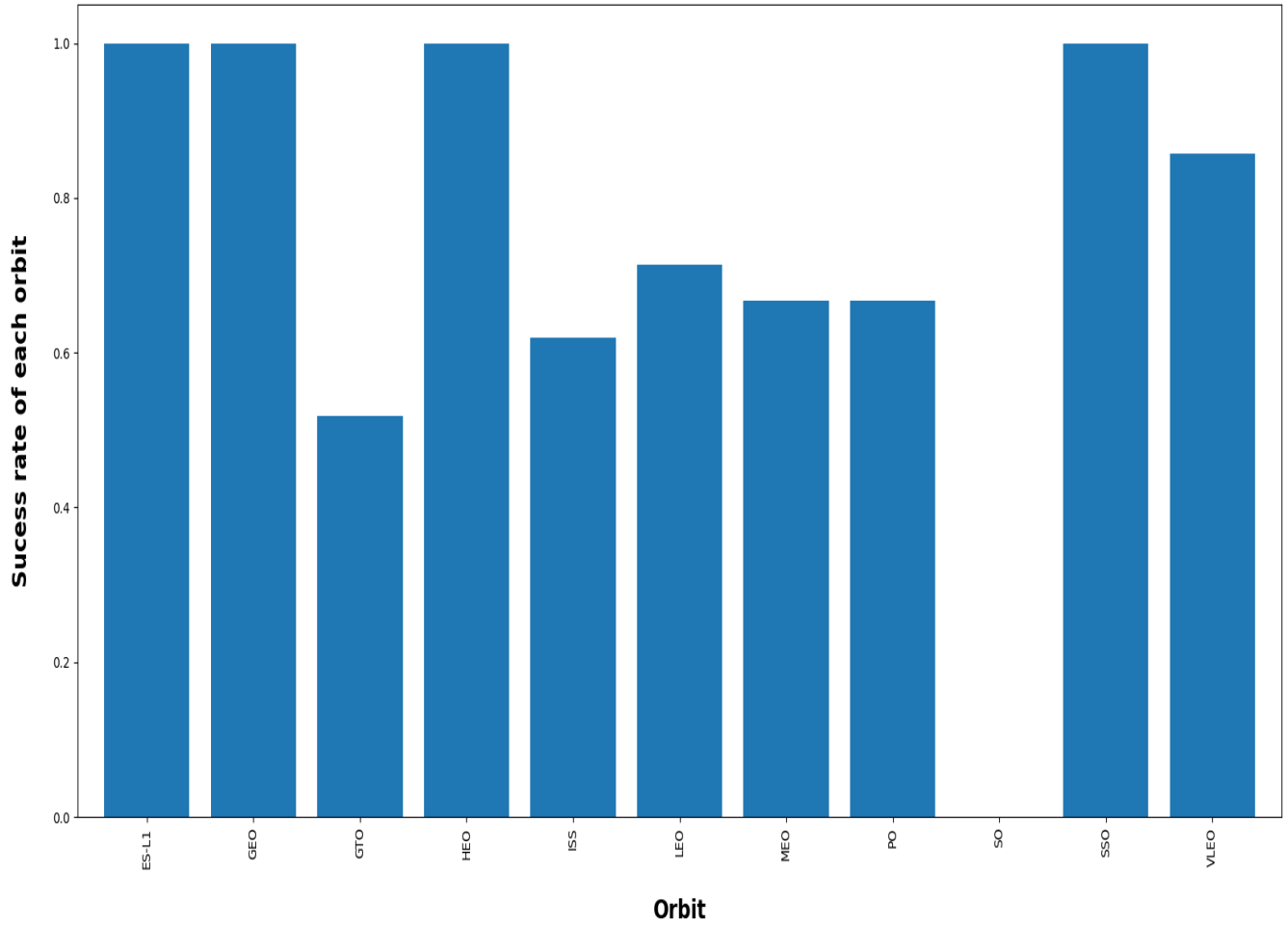
Payload vs. Launch Site



- From the scatter plot we conclude that the success rate of landing grows significantly once the payload mass is greater than 7000kg. In particular payloads over 9,000kg have excellent success rate.
- When the payload mass is less than 7000kg the success rate on the launch site CCAFS SLC 40 is lower than the success rate on the other two sites.
- We can also observe that when the payload mass is over 10,000kg, there are no rockets launched on the launch site VAFB SLC 4E.

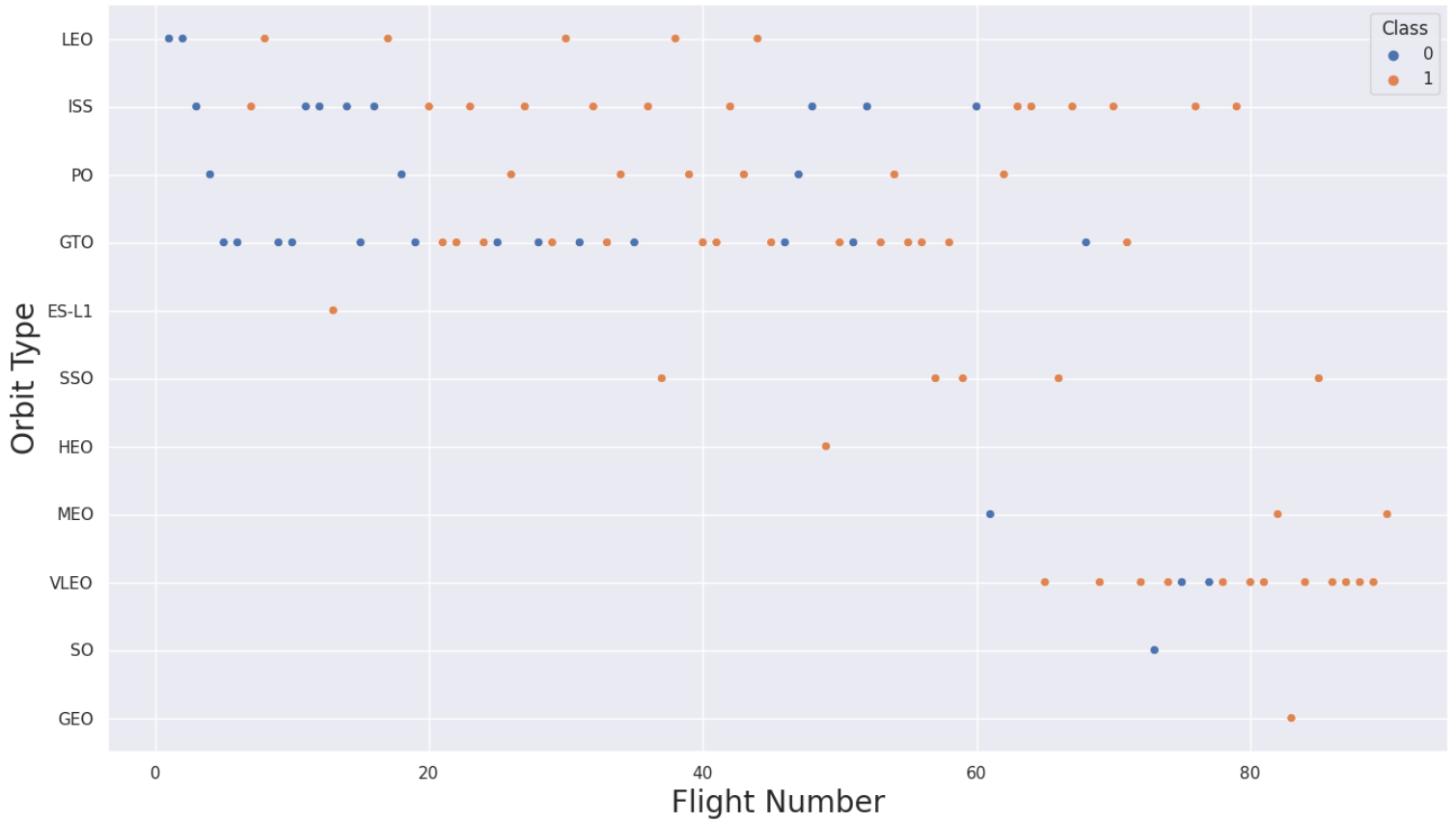
Success Rate vs. Orbit Type

- The bar chart shows that the orbit type influences the landing outcome. For instance some orbits, such as SSO, HEO, GEO and ES-L1 have 100% success rate, while SO orbit produced 0% rate of success.
- However, a further analysis shows that each of the orbits HEO, GEO and ES-L1 has only 1 occurrence, which is not a good evidence to make any significant conclusion.
- We can observe that the orbit VLEO occurs more often and it has a success rate above 80%.



Flight Number vs. Orbit Type

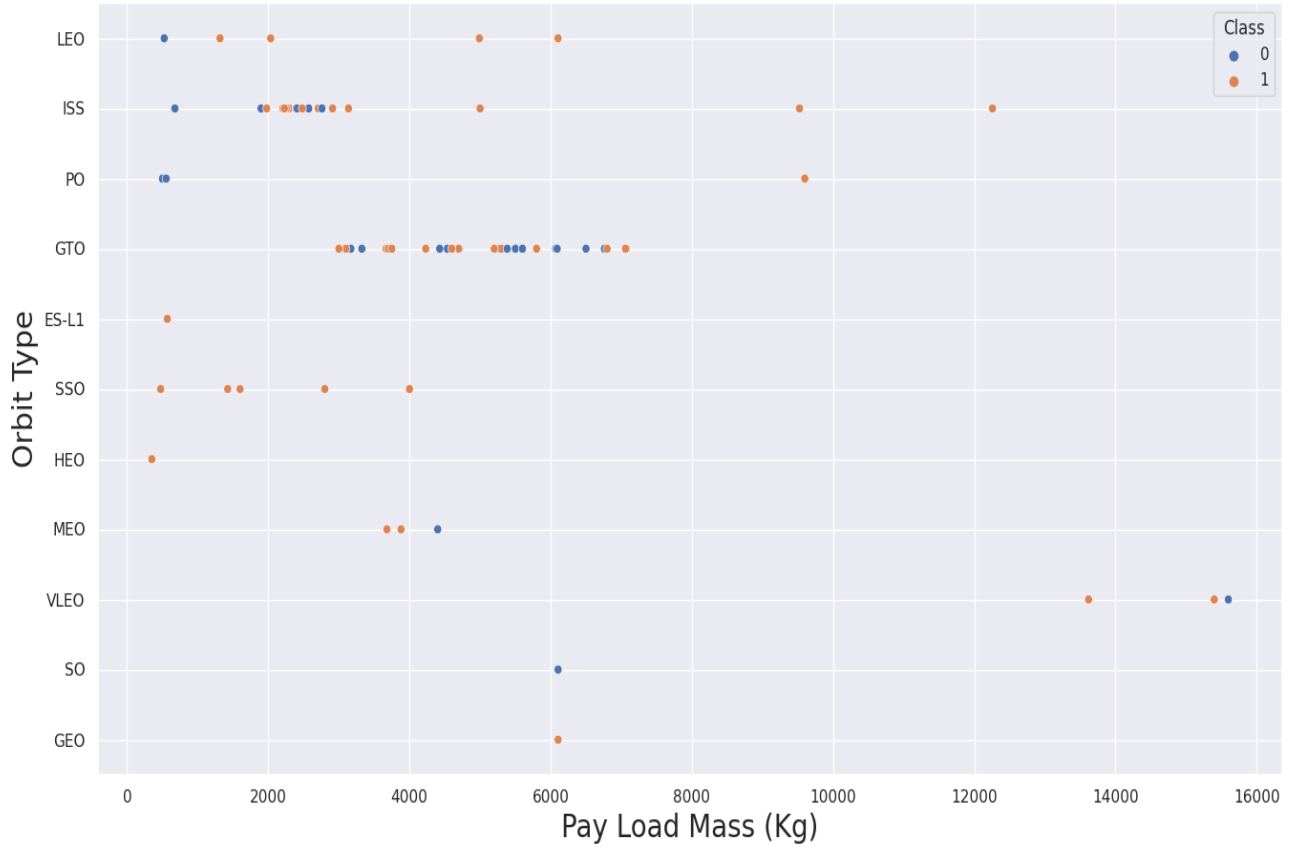
- The plot shows that the success rate improved over time in most of the orbits, in particular in the LEO orbit.
- However, in the GTO orbit there is no relationship between the flight number and success.
- We can also observe that the frequency of the VLEO orbit increased significantly lately, so the VLEO orbit is becoming more popular than the other orbits.



Payload vs. Orbit Type

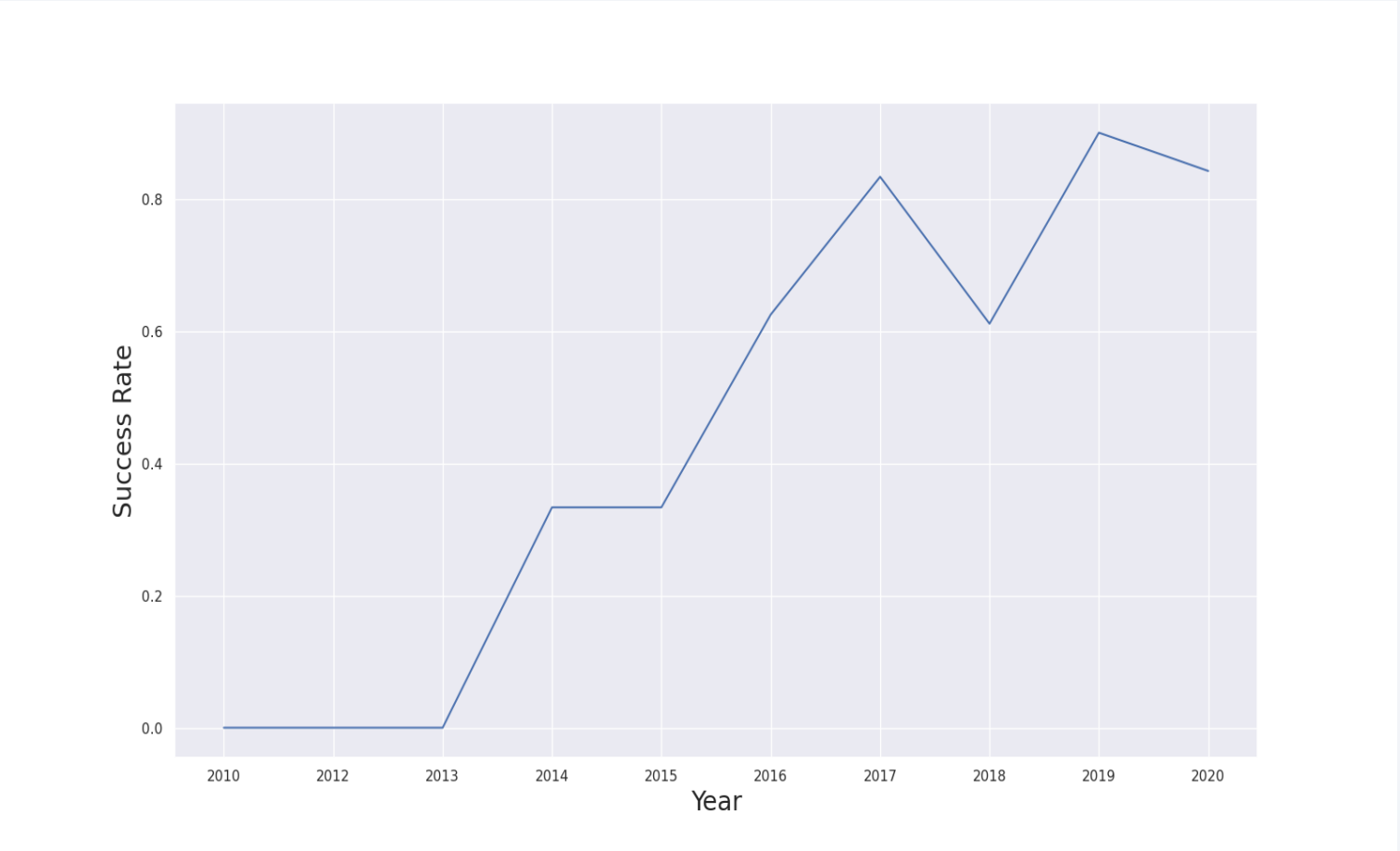
- From the scatter plot we can conclude that with heavy payloads the successful landing rate is high for Polar, LEO and ISS.

- On the other hand, for GTO we cannot distinguish this well as both positive landing rate and negative landing rate occur here and there.



Launch Success Yearly Trend

- The line chart shows that the success rate started increasing in 2013 and kept increasing until 2020 (with exception in 2018, where there was a slight decline).
- We may hope that that this increasing trend will continue during the next few years until eventually reaching a success rate close to 100%.



All Launch Site Names

According to data, there are four launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with 'CCA'

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

Total Payload Mass

- The total payload mass carried by boosters from NASA is 45596 kg.
- We calculated the total payload mass using the query below.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%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 booster version F9 v1.1 is 2928.4 kg.
- We calculated this amount using the query below.

```
%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

- First successful landing outcome on ground pad was on December 22, 2015.
- To determine this date we used the query below.

```
%sql SELECT DATE FROM SPACEXTBL WHERE Landing_Outcome='Success (ground pad)' ORDER BY date(DATE) LIMIT 1;
```

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

Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg are the following: FT B1022, F9 FT B1026, F9 FT B1021.2 and FT B1031.2, F9.
- We obtained the result using the following query.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * 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

- We used the following query to calculate the total number of successful and failure mission outcomes.

```
%sql SELECT sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful Mission", \
    sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \
FROM SPACEXTBL;
```

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

Successful Mission	Failure Mission
100	1

Boosters Carried Maximum Payload

- The following query yields the names of the boosters which have 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

- The following query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for year 2015.

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT substr(Date, 4, 2) AS MONTH_NAME, \
    Landing_Outcome AS LANDING_OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

MONTH_NAME	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

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

- We use the following query to rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

Landing Outcome	Total Count
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

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

- The following query ranks the successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT Date, COUNT(Landing_Outcome) as "Total Count" FROM SPACEXTBL \
  WHERE Date BETWEEN '2010-06-04' and '2017-03-20' AND Landing_Outcome LIKE '%Success%' \
  GROUP BY Date\
   ORDER BY COUNT(Landing_Outcome) DESC
```

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

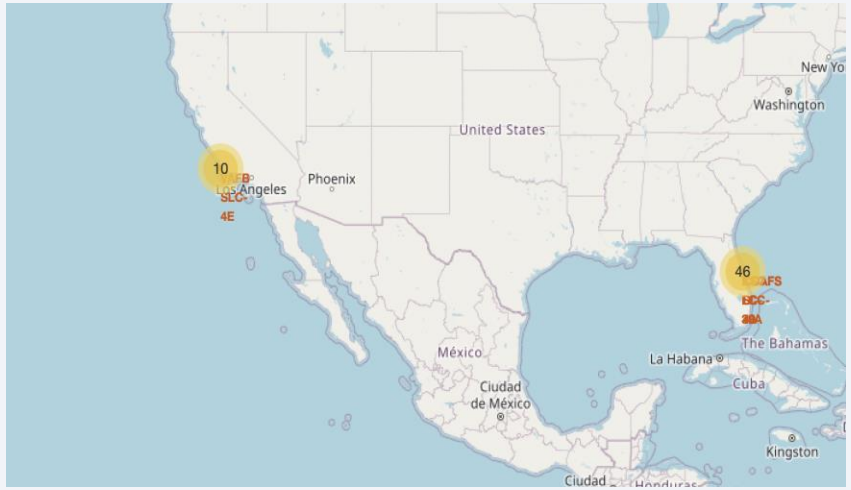
Date	Total Count
2017-02-19	1
2017-01-14	1
2016-08-14	1
2016-07-18	1
2016-05-27	1
2016-05-06	1
2016-04-08	1
2015-12-22	1

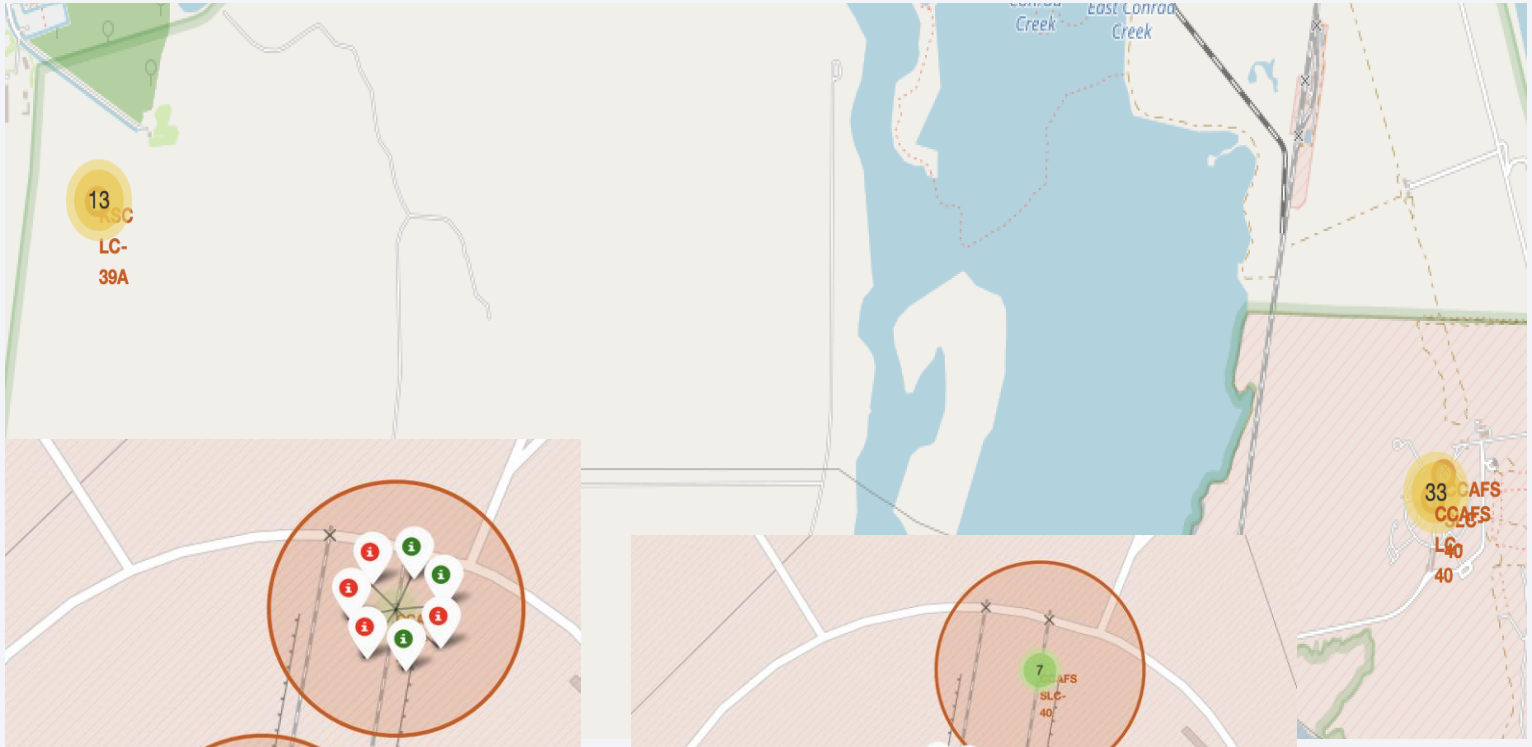

Section 3
Launch Sites
Proximities Analysis

Locations of launch sites

- From the map we can observe that SpaceX launch sites are in the USA, more precisely, in the states of California and Florida.
- All the launch sites are near the coast.

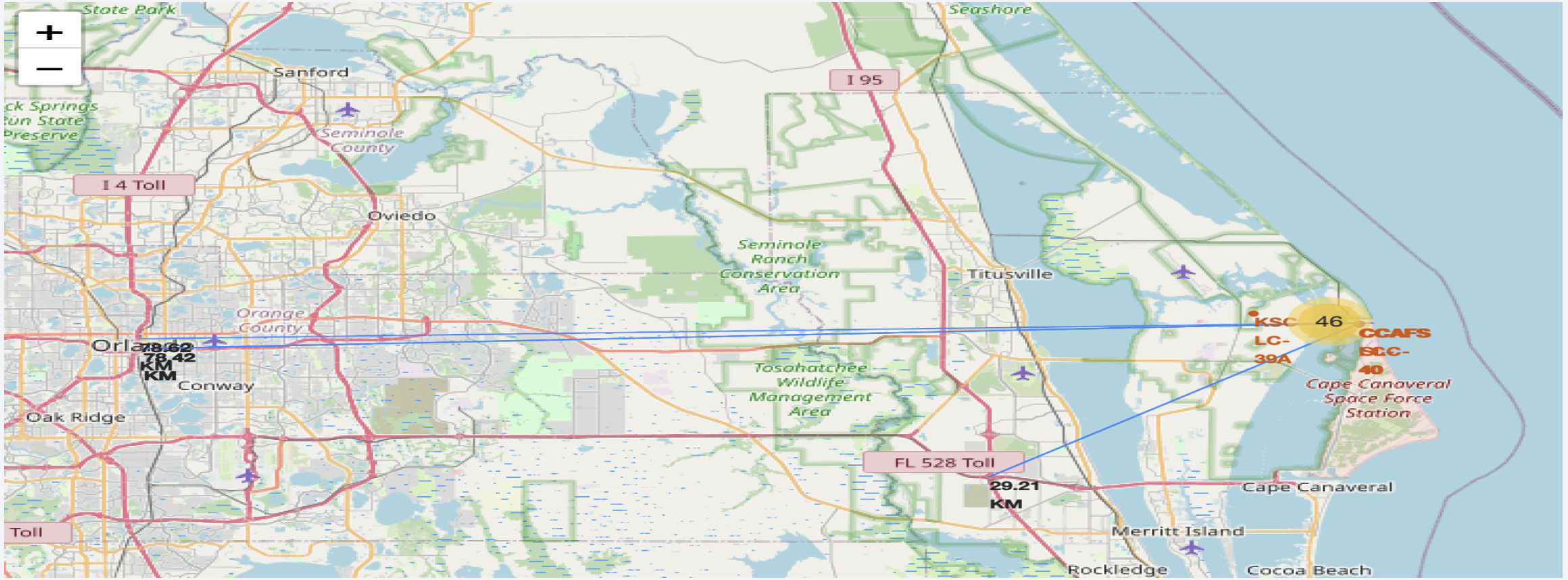
Markers showing Launch Sites Outcomes



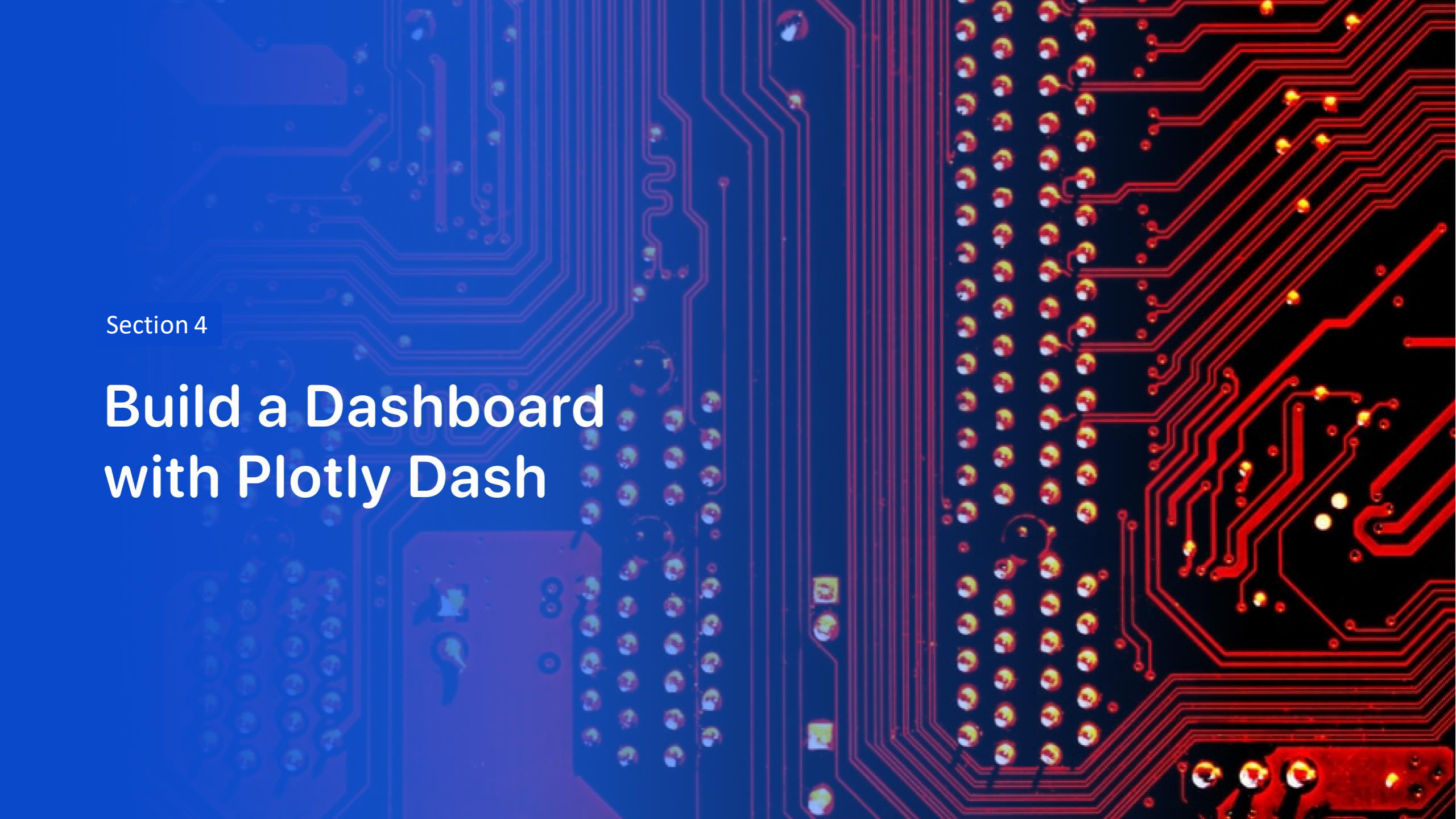


Green Markers show successful launches and Red Markers show failures.

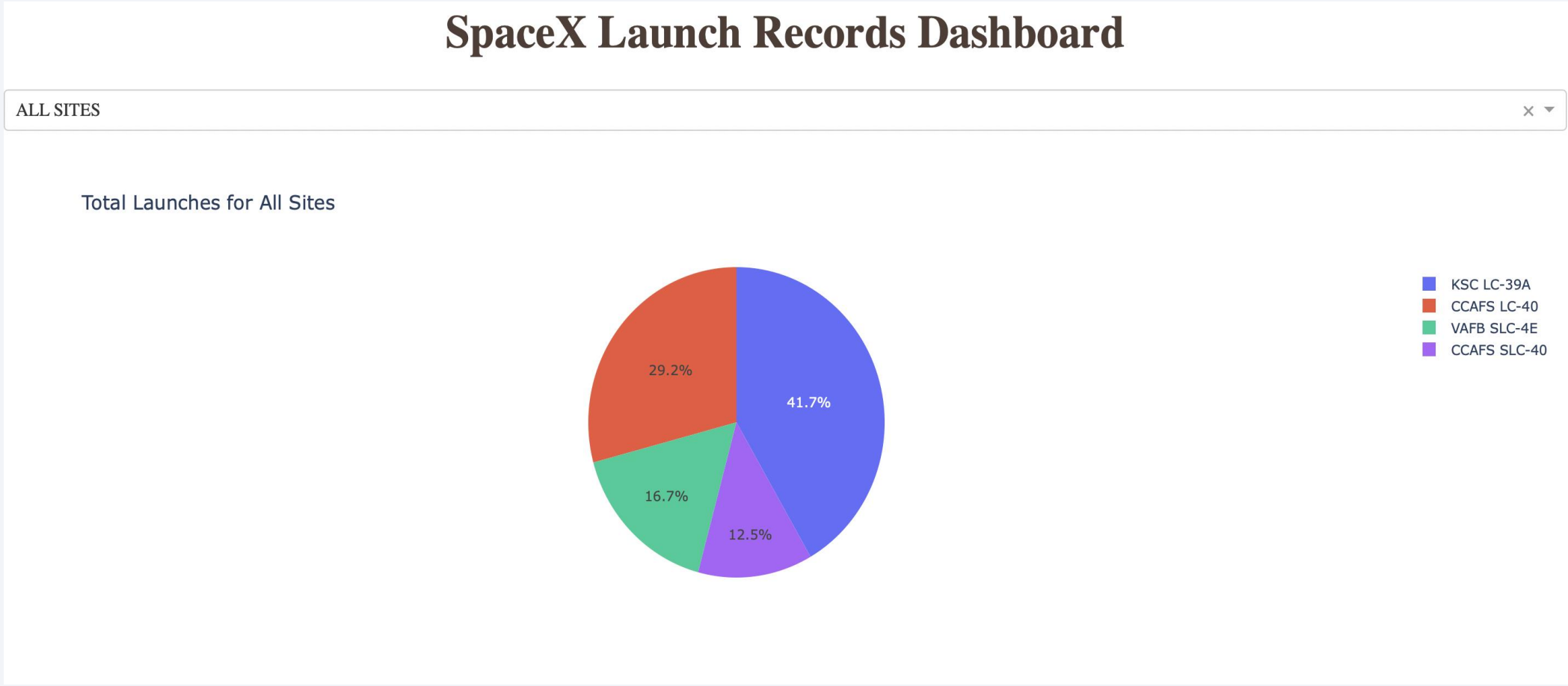
Launch site distance to landmarks



- Exploring the generated folium map we observe the distance of a selected launch site (here CCAFS SLC-40) to its proximities such as closest city (here Orlando), railway, highway and coastline, with distance calculated and displayed.
- Launch sites are near the coast and keep certain distance away from the cities.

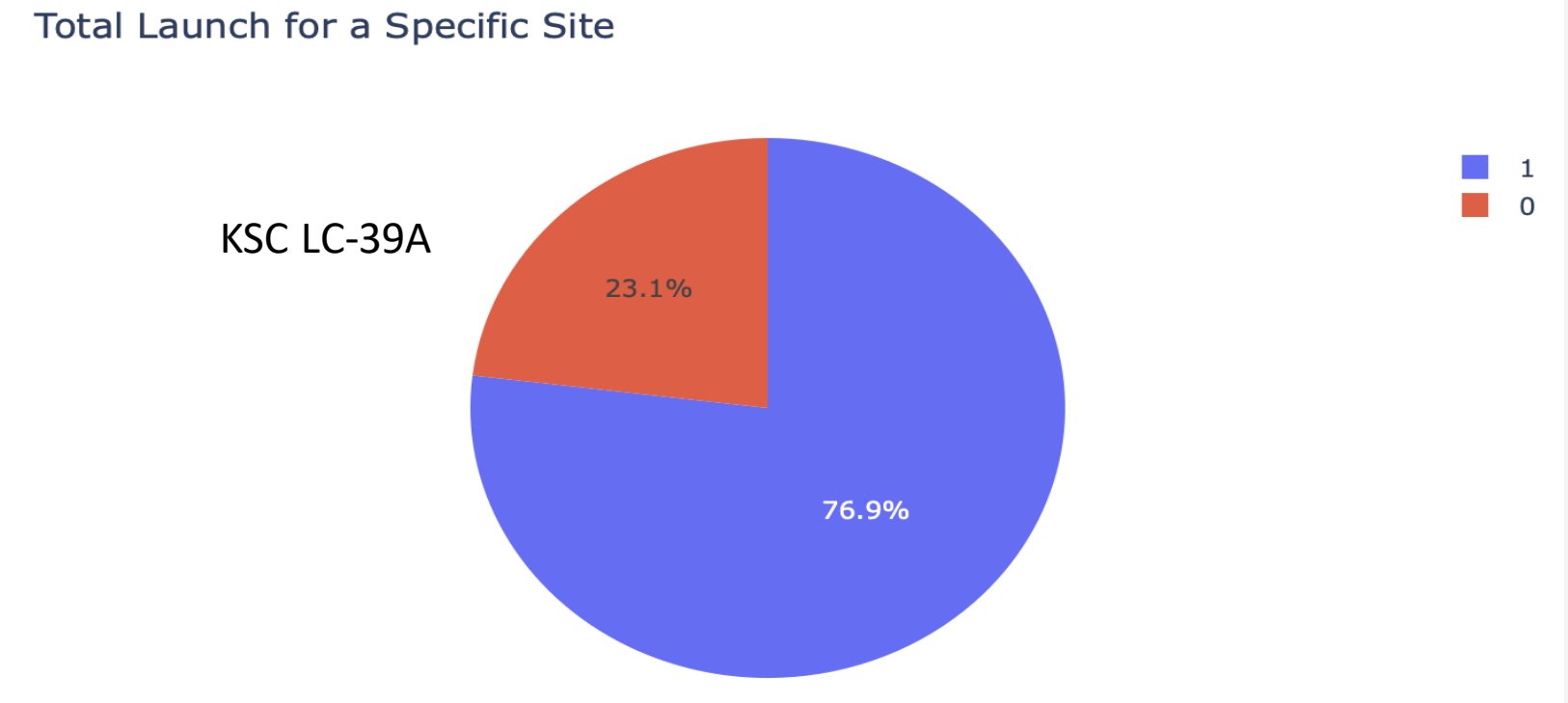

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



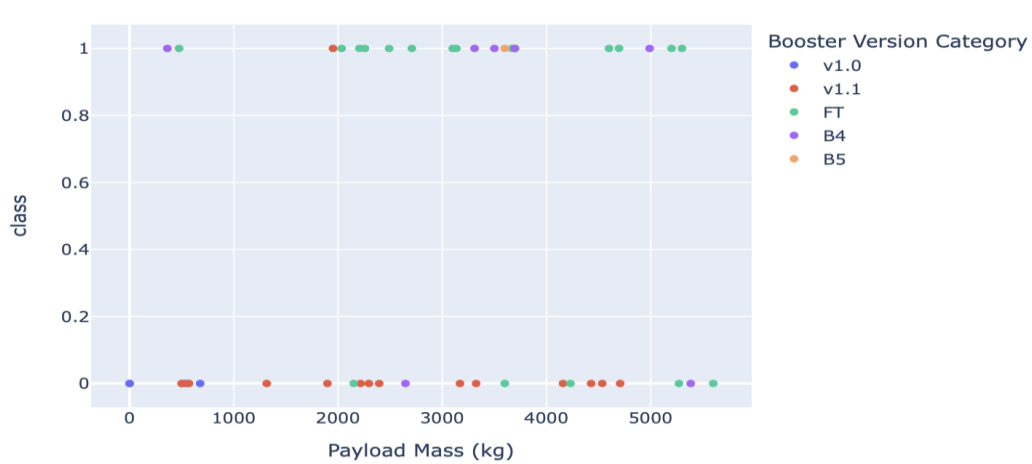
The highest number of successful launches have done from the launch site KSC LC-39A.

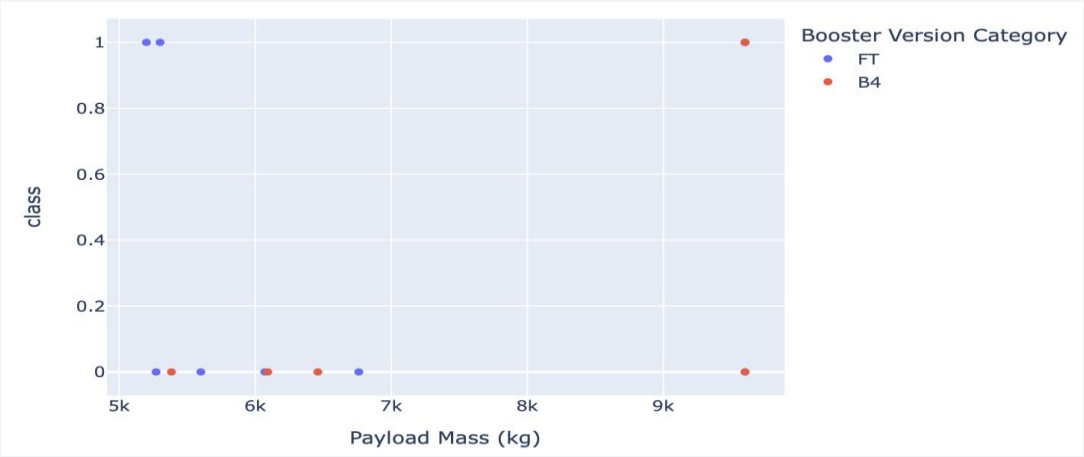
Launch Success Ratio for KSCLC-39A



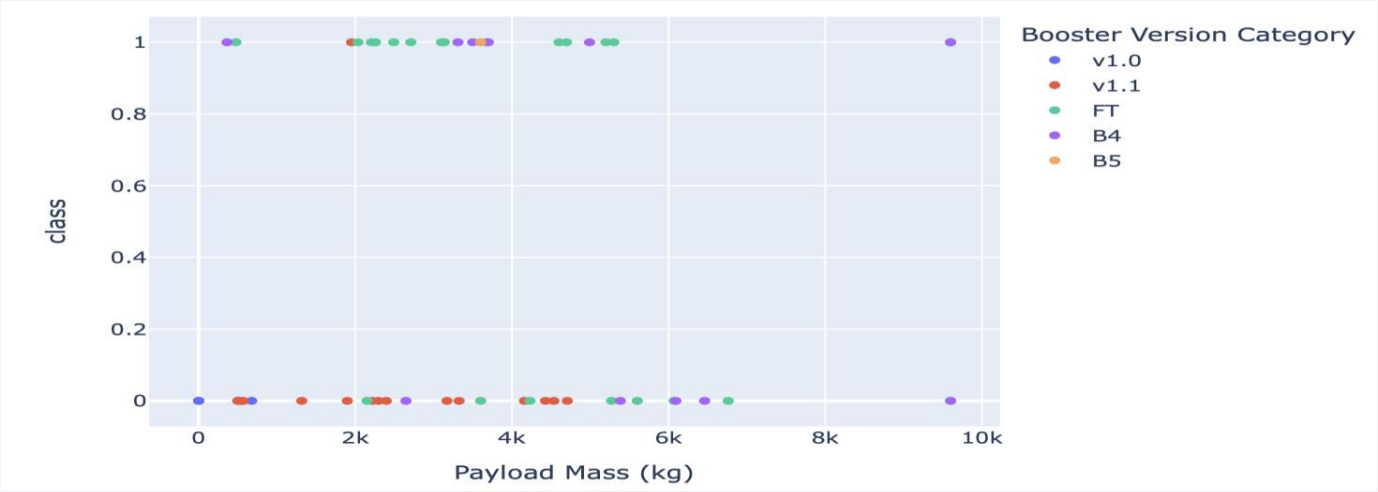
- KSC LC-39A is the launch site with highest launch success.
- 76.9% of launches are successful in this site.

Payload Mass vs. Launch Outcome





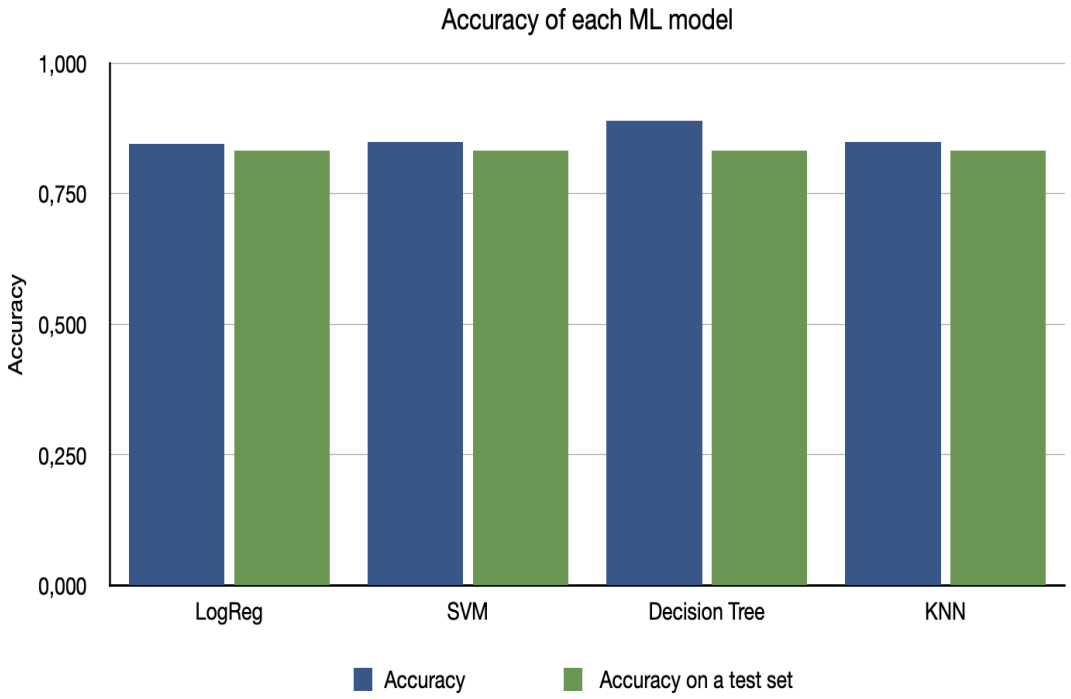
- For low weighted (under 6000 kg) payloads booster version FT has the largest success rate.
- Only booster versions FT and B4 occur with payloads above 5000 kg, and the success rate for such payloads is not good.



Section 5

Predictive Analysis (Classification)

Classification Accuracy



- We built and tested four classification models. The bar chart on the left shows the corresponding accuracies.
- The model with the highest classification accuracy is the Decision Tree Classifier, which has accuracy of 88.9 %.

Classification Accuracy

- The query on the right found the best performing model.

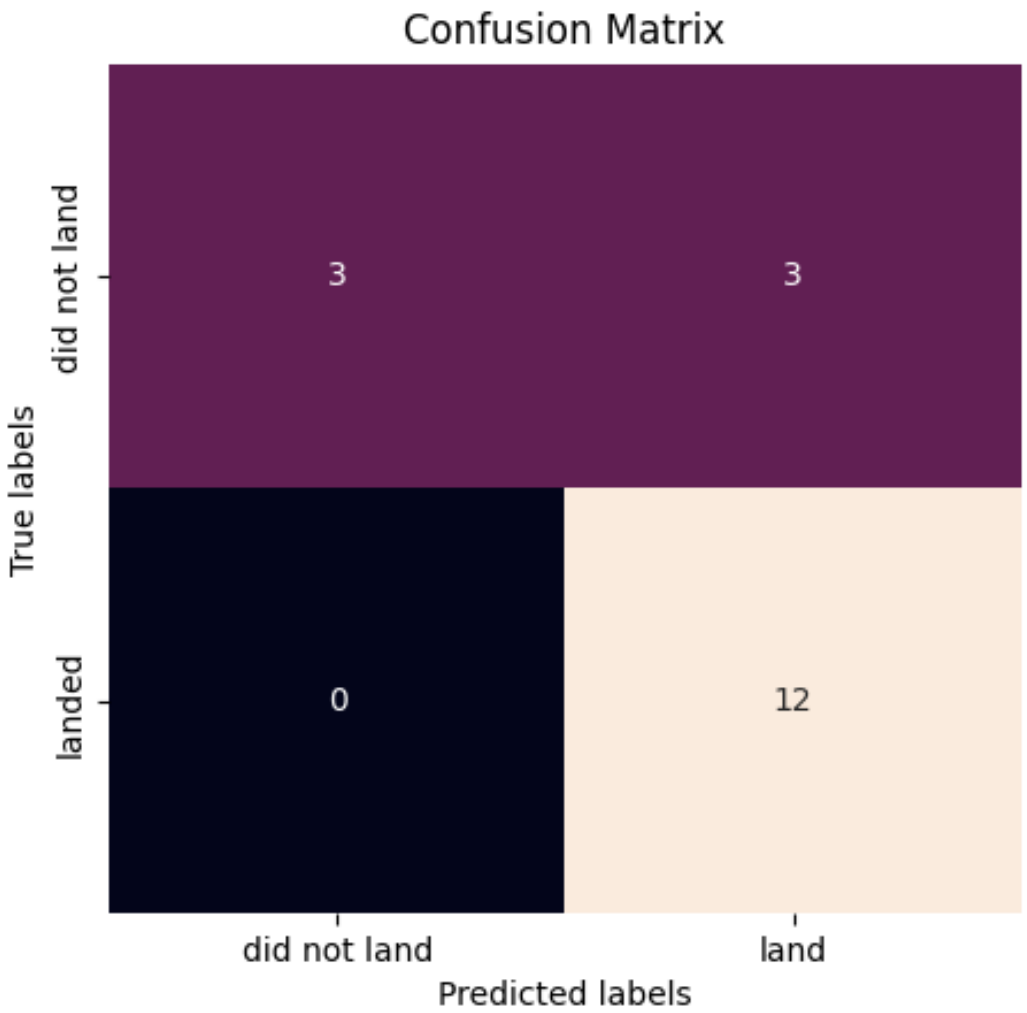
```
models = {'KNN':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8892857142857145
Best params is : {'criterion': 'entropy', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 2,
'min_samples_split': 10, 'splitter': 'best'}
```

Confusion Matrix

- On the right we have the confusion matrix of the Decision Tree Classifier, which is the best performing model.
- Analyzing the confusion matrix, we see that the Decision Tree can distinguish between the different classes; the major problem is false positives.



Conclusions

From our analysis we derive the following conclusions:

- The success rate of launches started increasing in 2013 and kept increasing until 2020 (with exception in 2018, where there was a slight decline). We may hope that that this increasing trend will continue during the next few years until eventually reaching a success rate close to 100%. That would allow to reuse the first stage and reduce the cost significantly.
- KSC LC-39A is the launch site with highest launch success (76.9%).
- The orbit SSO has a success rate of 100% (and more than 1 occurrence).
- For some orbits (such as Polar, LEO and ISS) with heavy payloads the successful landing rate is high, while for some other orbits (such as GTO) we cannot distinguish this.
- The Decision Tree Classifier is the best Machine Learning model (with the best accuracy score) for our purpose. One should use this ML model to predict successful first stage landings.
- Hence, using the right launch site, orbit and payload mass one may achieve a very high success rate on the first stage, allowing us to reduce the cost to a great extent and offer better prices than the competitors.

Thank you!