Winning Space Race with Data Science

Kohei Suzuki
23.02.2022

Outline

- **Executive Summary**
- **Introduction**
- Section1: **Methodology**
- Section2: **Results** from exploratory data analysis
- Section3: **Results** from launch sites proximities analysis
- Section4: **Results** from interactive dashboard analysis
- Section5: **Results** from machine learning predictive analysis
- **Conclusion**
- **Appendix**

Executive Summary

Methodologies

Our data is collected by using SpaceX REST API and web scraping related Wikipedia pages. The following data analyses are conducted after data wrangling process: exploratory data analysis (EDA) using visualization and SQL, interactive visual analysis using Folium and Plotly Dash, and predictive analysis using machine learning (ML) classification models.

Results

Our data analyses clarify some patterns in successful launch conditions. Our Decision Tree model shows good classification performance. This model can be used to predict mission outcome, and therefore, it allows to estimate the cost of a launch beforehand.

Introduction

Background and context

Our new rocket company **SpaceY** would like to compete with **SpaceX**, the most successful company in aerospace today. Key success factor of SpaceX is that they could drastically reduce the rocket launch cost. SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Objectives

Determine the cost of a launch by gathering information about SpaceX and creating dashboards. ***Determine if SpaceX will reuse the first stage*** by training a ML model and using public information.

Section 1

Methodology

Methodology

- Data collection methodology:

  Using SpaceX REST API and web scraping related Wikipedia pages

- Perform data wrangling:

  Using an API, sampling data and dealing with nulls in raw data

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

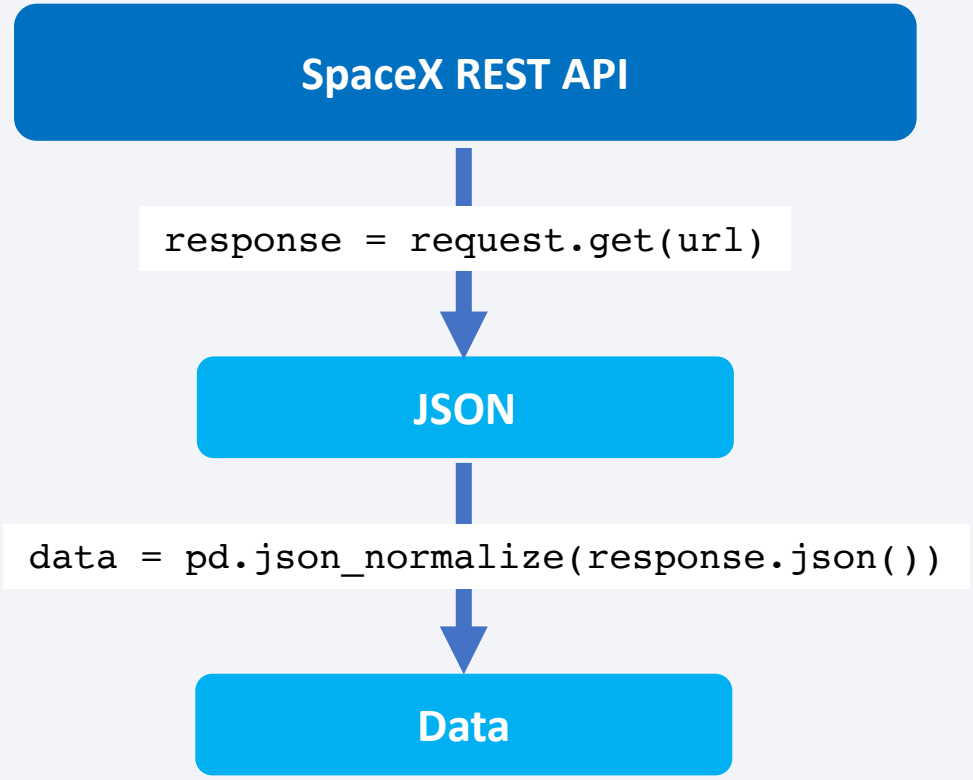
- Perform predictive analysis using machine learning (ML) classification models:

  Building a ML pipeline (preprocessing, train-test-split and grid search) and confusion matrix to determine the model with the best accuracy

Data Collection – SpaceX API

- Using request library to obtain Falcon 9 launch data in JSON from SpaceX API.

- Using json_normalize function to convert the JSON data into data frame.

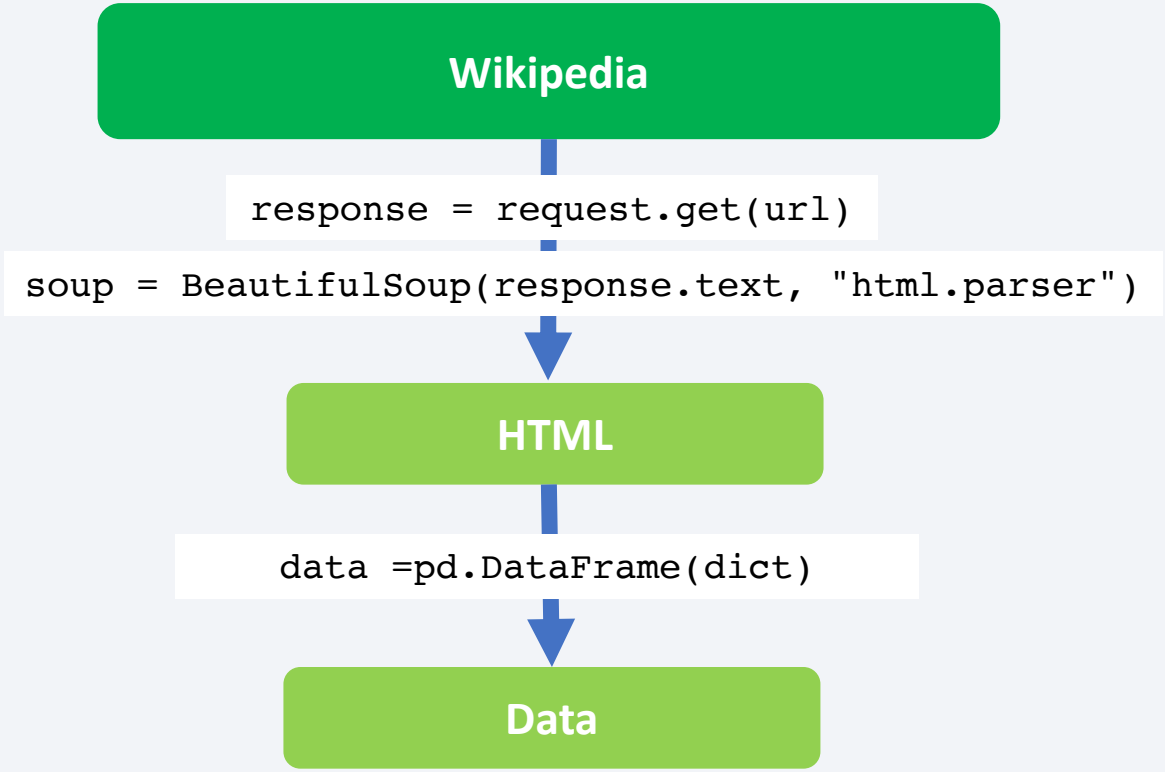


* Link to the completed SpaceX API calls notebook

Data Collection - Scraping

- Using BeautifulSoup package to web scrape HTML tables with Falcon 9 launch records from Wikipedia pages.

- Parsing the records extracted from HTML tables into dictionary, and then converting the dictionary into data frame.

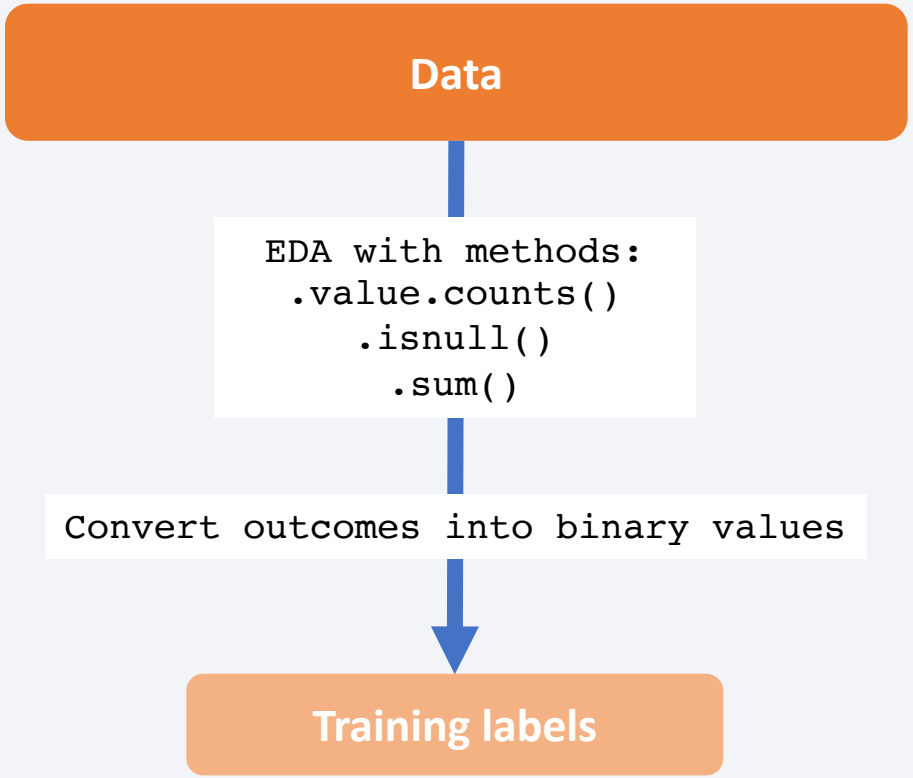


* Link to the completed web scraping notebook

Data Wrangling

- Performing Exploratory Data Analysis (EDA) to find some patterns in the data.

- Converting mission outcomes into training labels with binary numbers:
1 (successful landing) or 0 (unsuccessful landing).



* Link to the completed data wrangling notebook

EDA with SQL

- Understanding the SpaceX dataset (.csv file).
- Loading the dataset into the corresponding table in Db2 database.
- Executing SQL queries to explore mission outcomes according to launch sites, landing places, payload mass and booster versions.

* Link to the completed EDA with SQL notebook

EDA with Data Visualization

In order to see which variable would affect the launch outcome, following six plots are created by using Pandas and Matplotlib:

1. Scatter point chart with x axis to be Flight Number and y axis to be the launch site
2. Scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site
3. Bar chart for the success rate of each orbit
4. Scatter point chart with x axis to be Flight Number and y axis to be the Orbit
5. Scatter point chart with x axis to be Payload and y axis to be the Orbit
6. Line chart with x axis to be the extracted year and y axis to be the success rate

* Link to the completed EDA with data visualization notebook

Build an Interactive Map with Folium

- Marking the launch site locations on an interactive map with Folium by adding highlighted circle area with text label on coordinates for each site.

- Adding the launch outcomes for each site to visualize its success rate.

- Exploring and analyzing the proximities of launch sites (its closest city, railway, highway, etc.) in order to explain how to choose an optimal launch site.

* Link to the completed interactive map with Folium notebook

Build a Dashboard with Plotly Dash

Building a Plotly Dash application with the following components: launch site dropdown input component, callback function to render success-pie-chart based on selected site dropdown, range slider to select payload, and callback function to render the success-payload-scatter-chart scatter plot.

The following five questions will be answered with this dashboard : 1. Which site has the largest successful launches? 2. Which site has the highest launch success rate? 3. Which payload range(s) has the highest launch success rate? 4. Which payload range(s) has the lowest launch success rate? 5. Which F9 Booster version has the highest launch success rate?

* Link to the completed Plotly Dash lab notebook

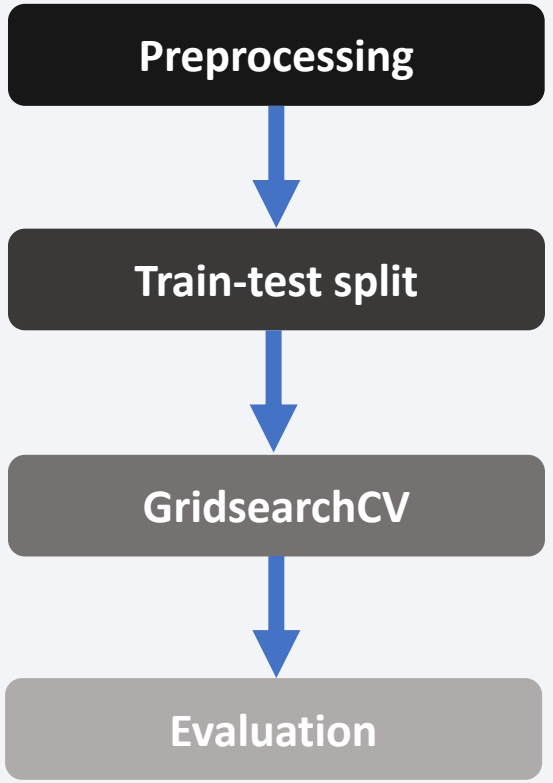
Predictive Analysis (Classification)

1. Creating target (Y) and standardizing features (X).

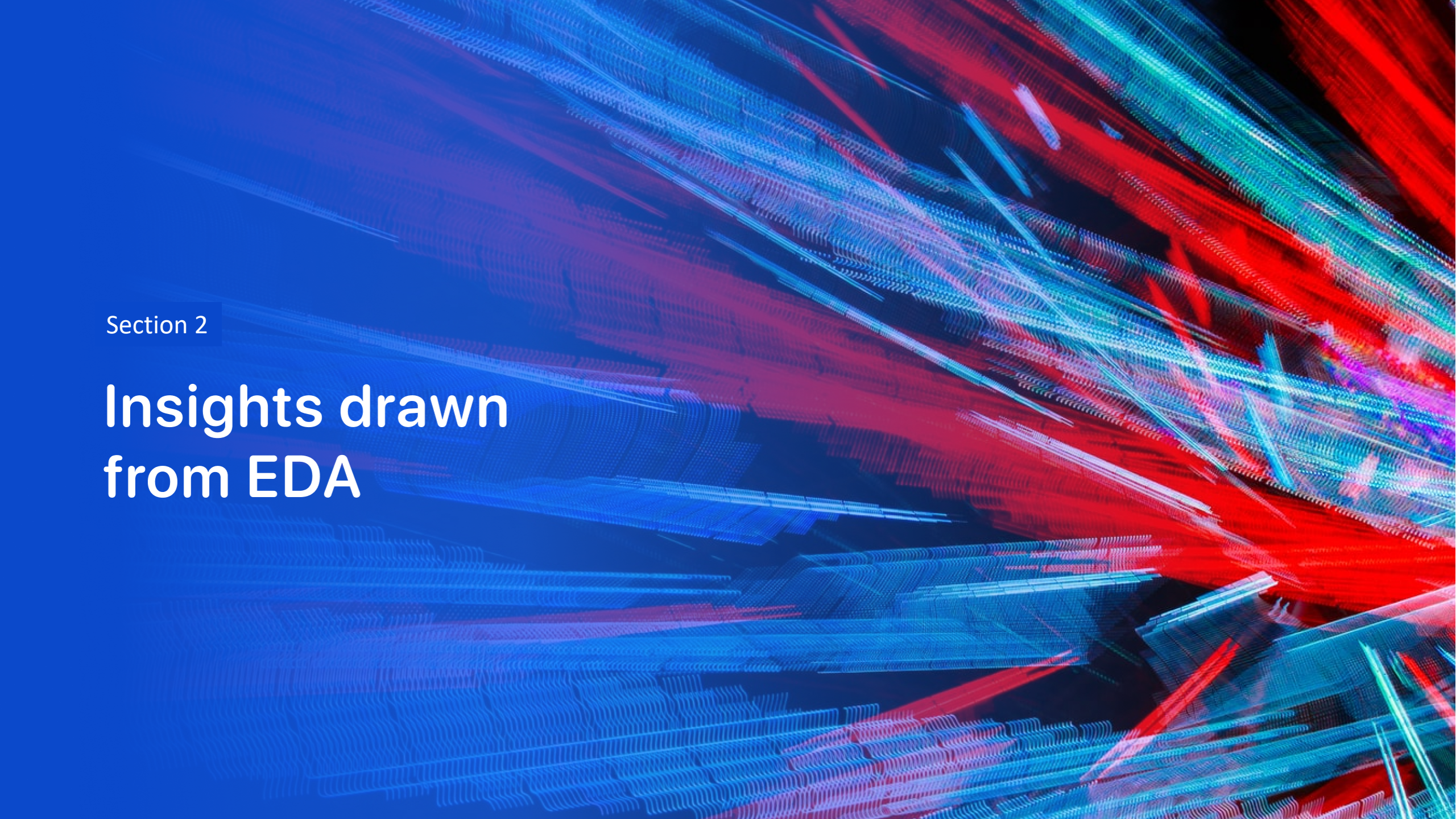
2. Splitting the data into training and testing data by using the function train_test_split.

3. Using grid search cross-validation to find the best parameters for each model (SVM, Classification Trees, KNN Classifier and Logistic Regression).

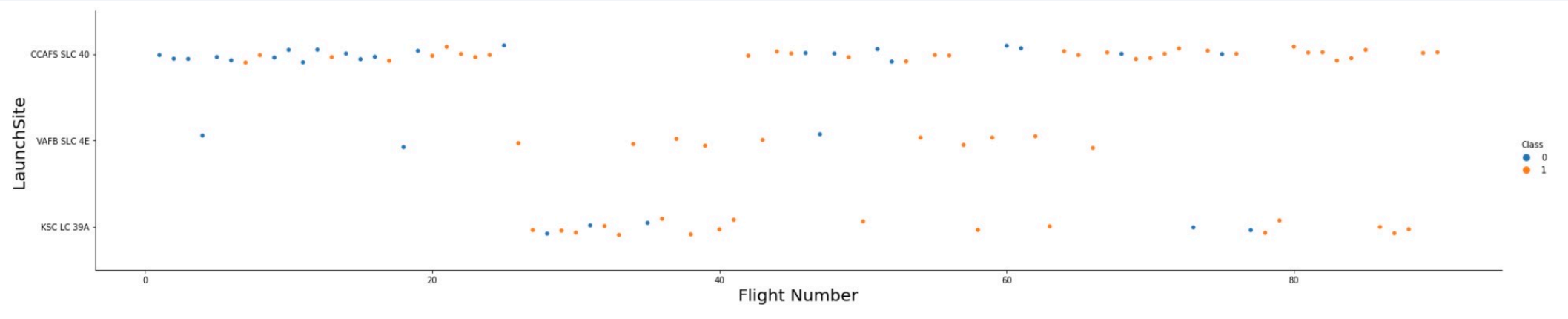
4. Evaluating the model performance with the best parameters on the testing data and plotting the confusion matrix.



* Link to the completed ML prediction notebook


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

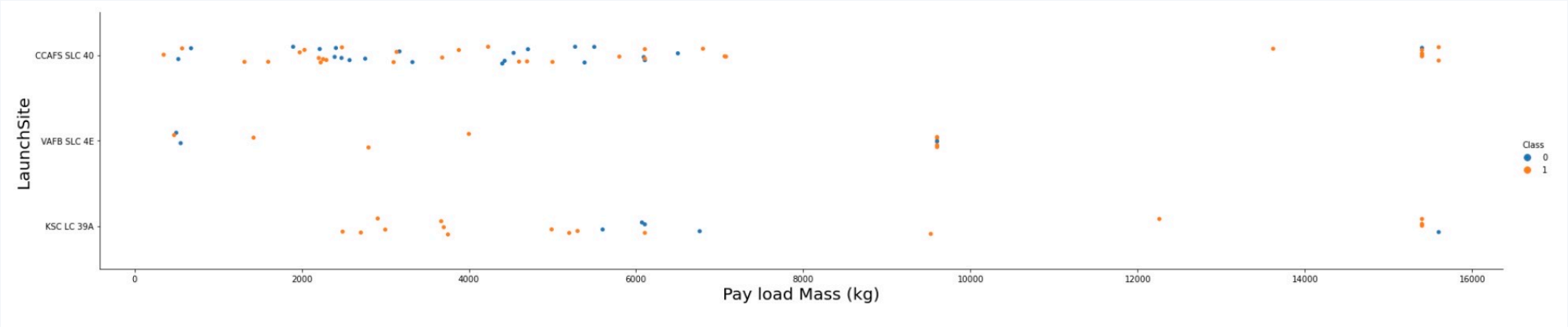


Scatter plot of Flight Number vs. Launch Site

(Class 1 means the booster successfully landed, class 0 means it was unsuccessful)

The last 13 flights landed all successfully. The VAFB-SLC launch site is not used for these flights.

Payload vs. Launch Site

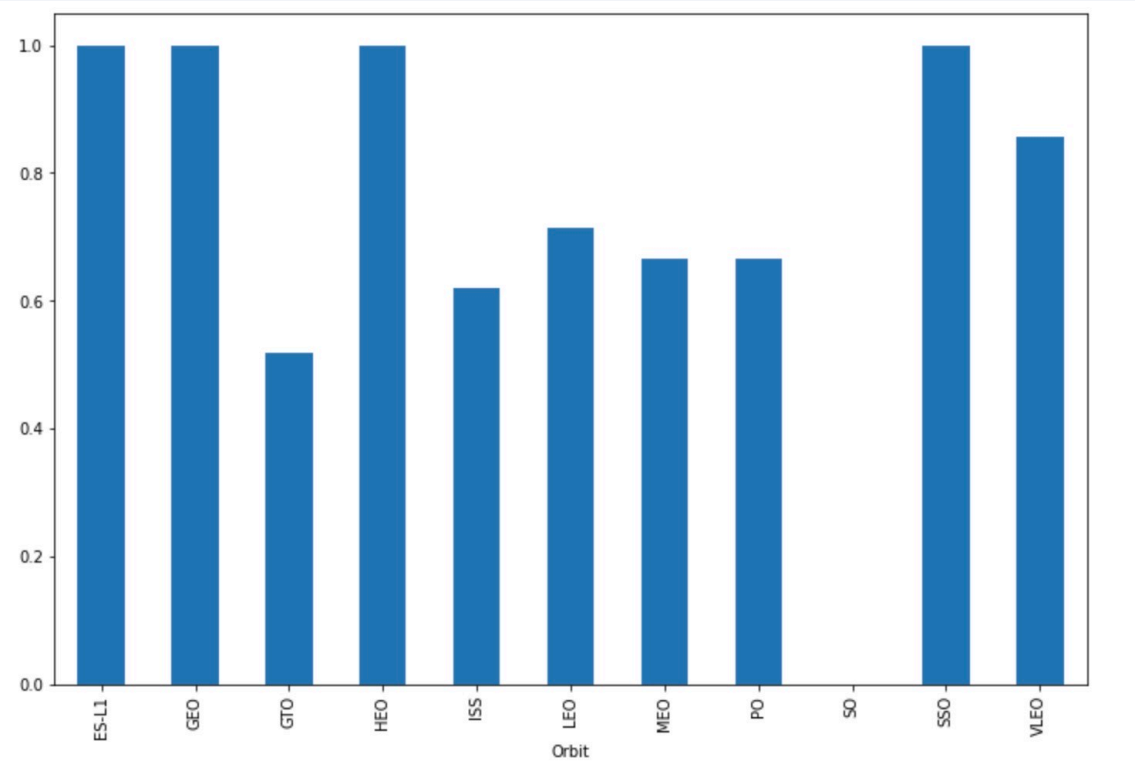


Scatter plot of Payload vs. Launch Site

(Class 1 means the booster successfully landed, class 0 means it was unsuccessful)

There are no rockets launched for heavy payload mass from the VAFB-SLC launch site.

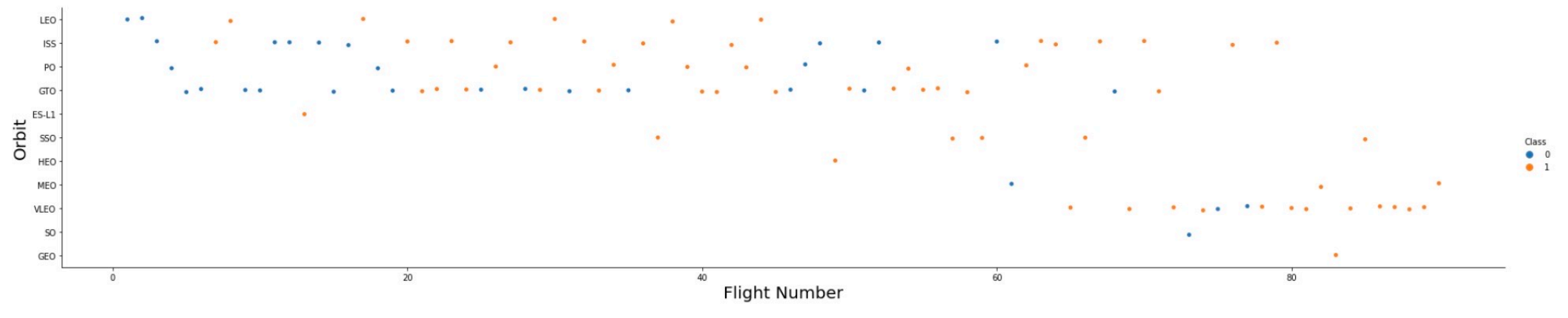
Success Rate vs. Orbit Type



Bar chart for the success rate of each orbit type

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

Flight Number vs. Orbit Type

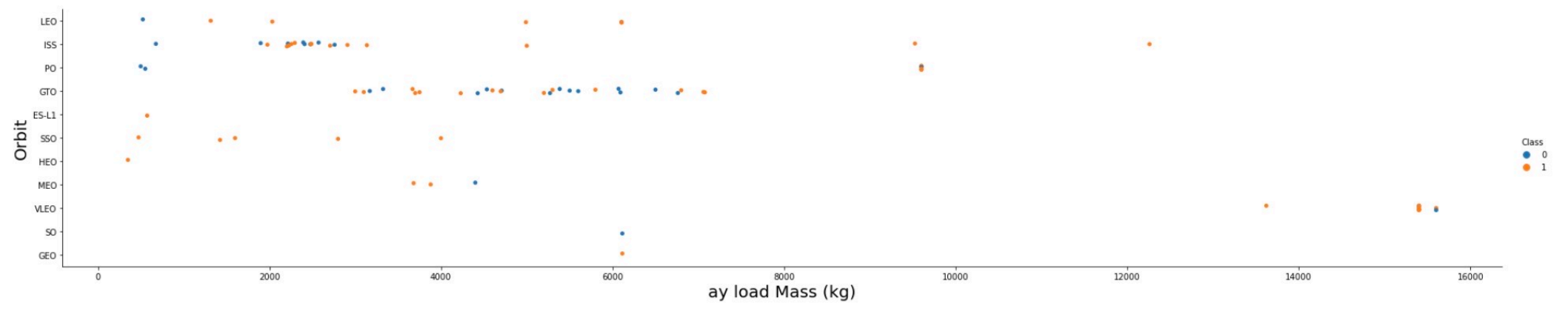


Scatter plot of Flight number vs. Orbit type

(Class 1 means the booster successfully landed, class 0 means it was unsuccessful)

There seems to be correlation between the LEO orbit outcome the number of flights. However, there seems to be no such clear correlation in other orbits.

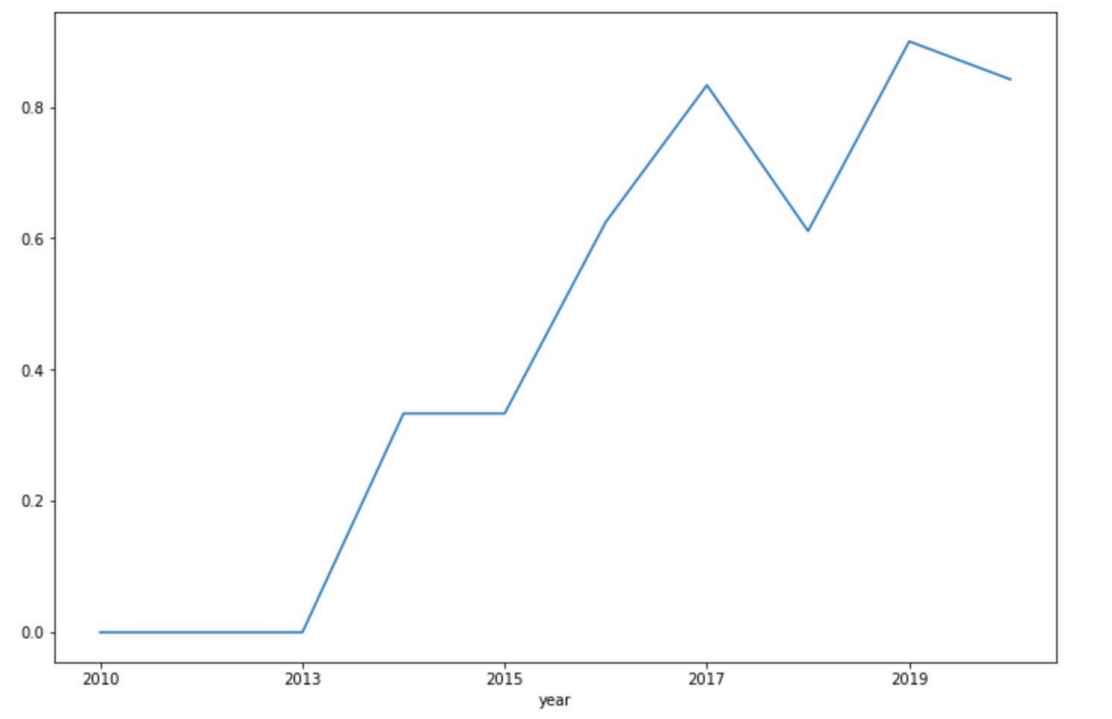
Payload vs. Orbit Type



Scatter plot of payload vs. orbit type

With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.

Launch Success Yearly Trend



Line chart of yearly average success rate

The success rate kept increasing since 2013.

All Launch Site Names

Our data contains several SpaceX launch facilities:

1. Cape Canaveral Air Force Station (CCAFS) Space Launch Complex 40 (SLC-40)
2. Cape Canaveral Air Force Station (CCAFS) Launch Complex 40 (LC-40)
3. Vandenberg Air Force Base (VAFB) Space Launch Complex 4E (SLC-4E)
4. Kennedy Space Center (KSC) Launch Complex 39A (LC 39A)

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXDATASET WHERE launch_site like '%CCA%' Limit 5
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

First 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer = 'NASA (CRS)'
```

1
45596

Calculating the total payload carried by boosters from NASA

> 45,696 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version like 'F9 v1.1%'
```

1
2534

Calculating the average payload mass carried by booster version F9 v1.1

> 2,534 kg

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE landing__outcome like 'Success (ground pad)'
```

1
2015-12-22

Finding the dates of the first successful landing outcome on ground pad

> 2015.12.22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE (landing__outcome like 'Success (drone ship)') & (payload_mass__kg_ > 4000) & (payload_mass__kg_ < 6000)
```

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

Finding the names of boosters which have successfully landed on drone ship, and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Distinct mission_outcome, COUNT(mission_outcome) as total
FROM SPACEXDATASET
GROUP BY mission_outcome
```

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Calculating the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, payload_mass__kg_
FROM SPACEXDATASET
WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXDATASET)
```

booster_version	payload_mass__kg_
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

Finding the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql
SELECT booster_version, launch_site
FROM SPACEXDATASET
WHERE (YEAR(DATE) = 2015) & (landing__outcome = 'Failure (drone ship)')
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015

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

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) as total
FROM SPACEXDATASET
WHERE (DATE between '2010-06-04' and '2017-03-20')
GROUP BY landing__outcome
ORDER BY total DESC
```

landing__outcome	total
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

Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order


Section 3
Launch Sites
Proximities Analysis

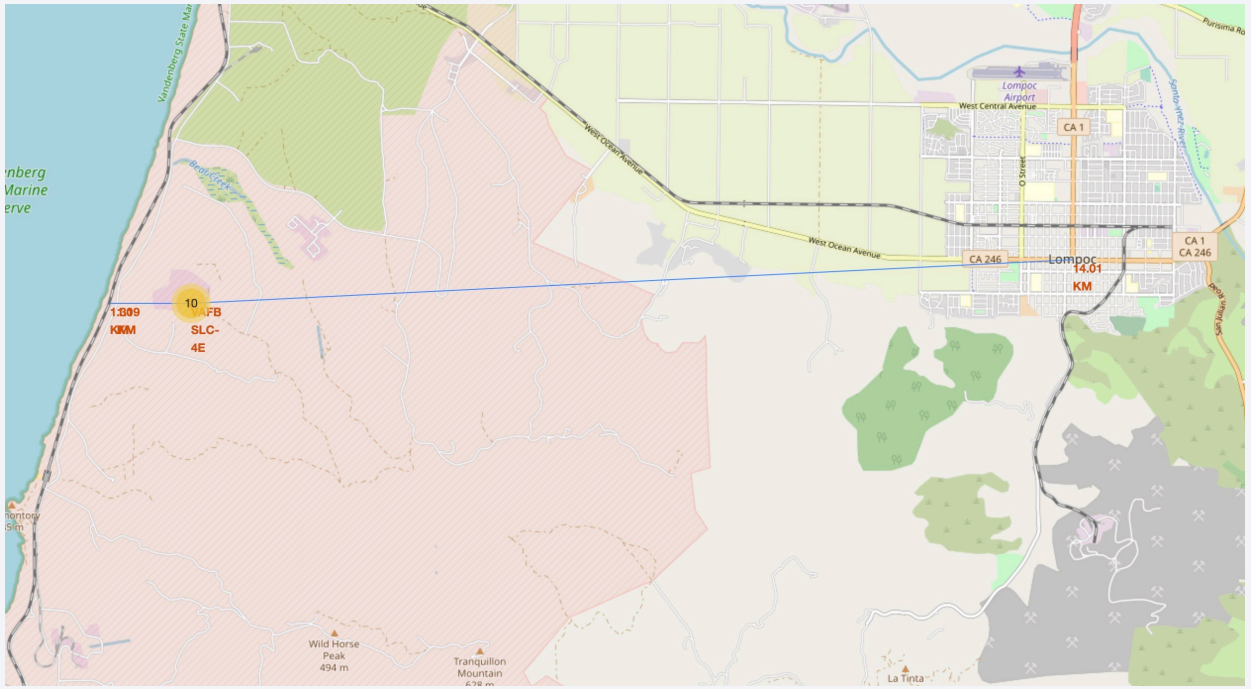
All launch sites on a global map

All launch sites are in proximity to the equator and the coast.

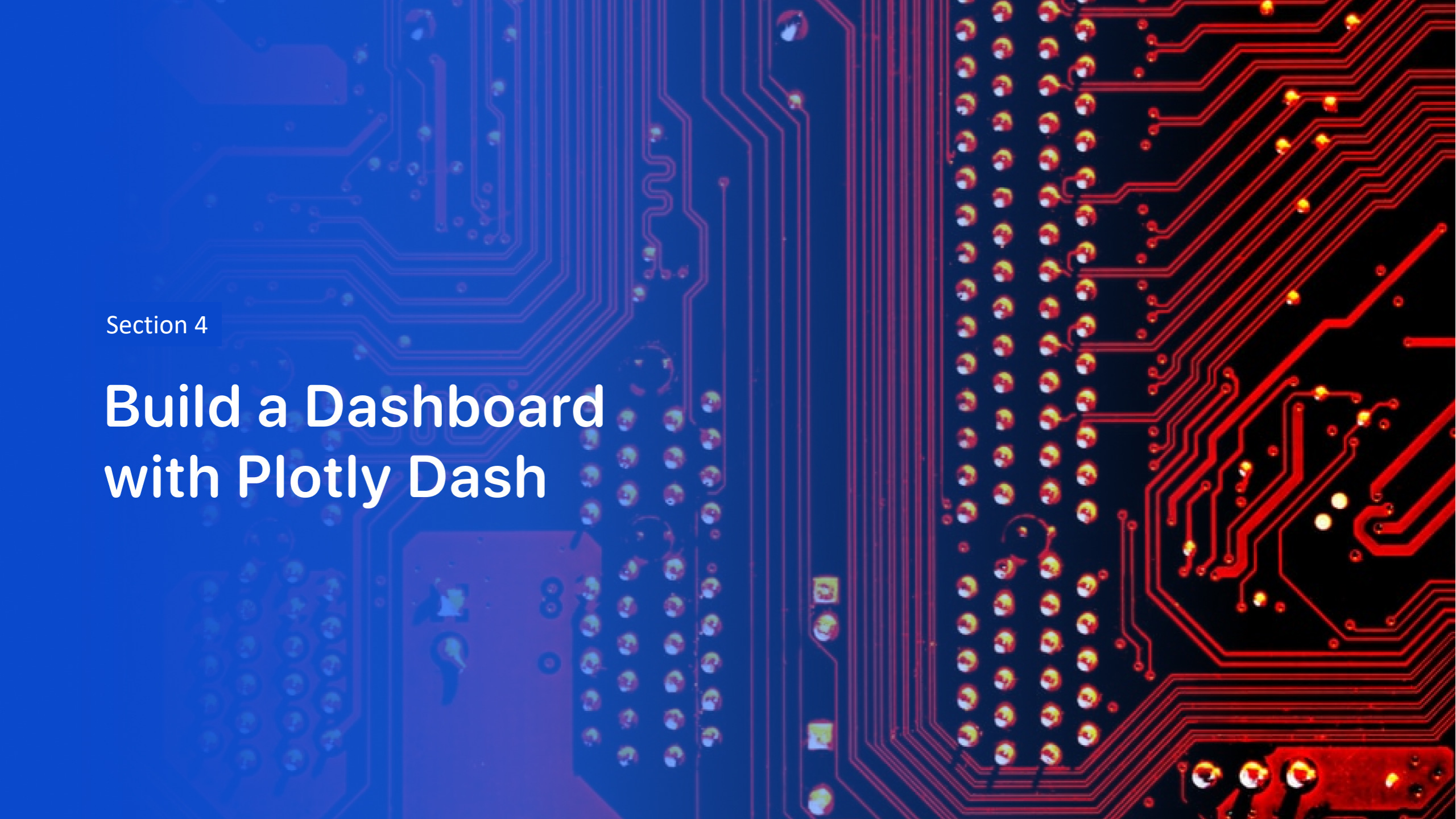
Color-labeled launch outcomes on the map

It can be easily identified which launch sites have relatively high success rates by using color-labeled markers in marker clusters.

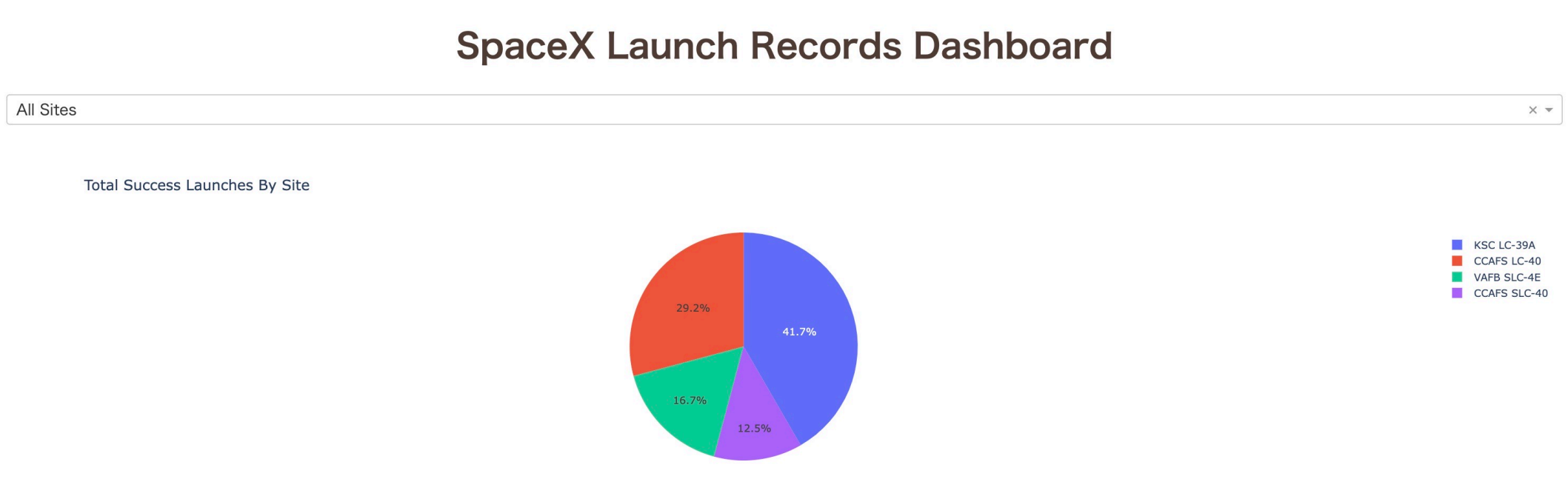
A selected launch site to its proximities



Calculating and displaying the distance between the Vandenberg Air Force Base Space Launch Complex 4E (**VAFB SLC-4E**) and its proximities such as railway, highway, coastline and city.

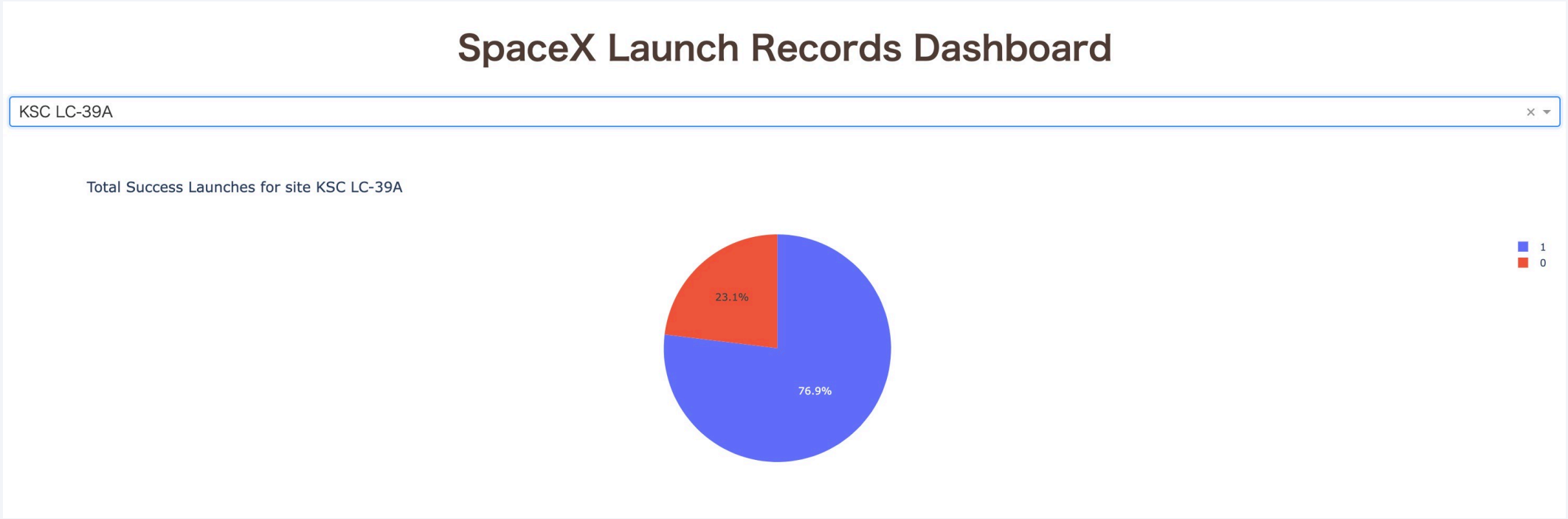

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites in pie-chart



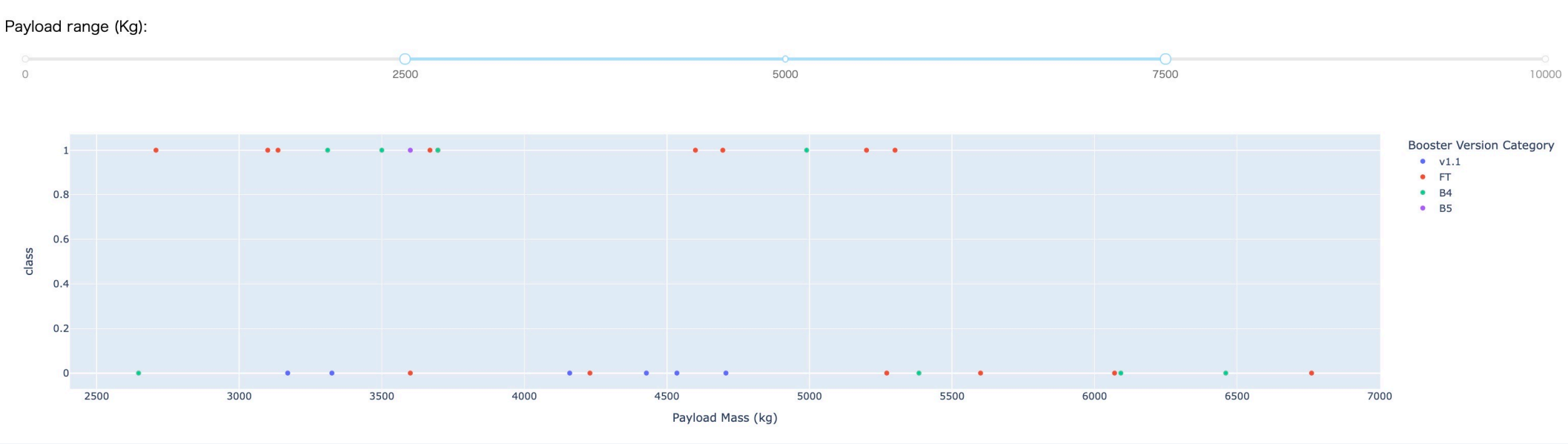
Pie-chart graph showing the total success launches by site.

Pie-chart for the most successful launch site



Kennedy Space Center Launch Complex 39A (**KSC LC-39A**) has the highest launch success ratio.

Scatter plot with payload range slider



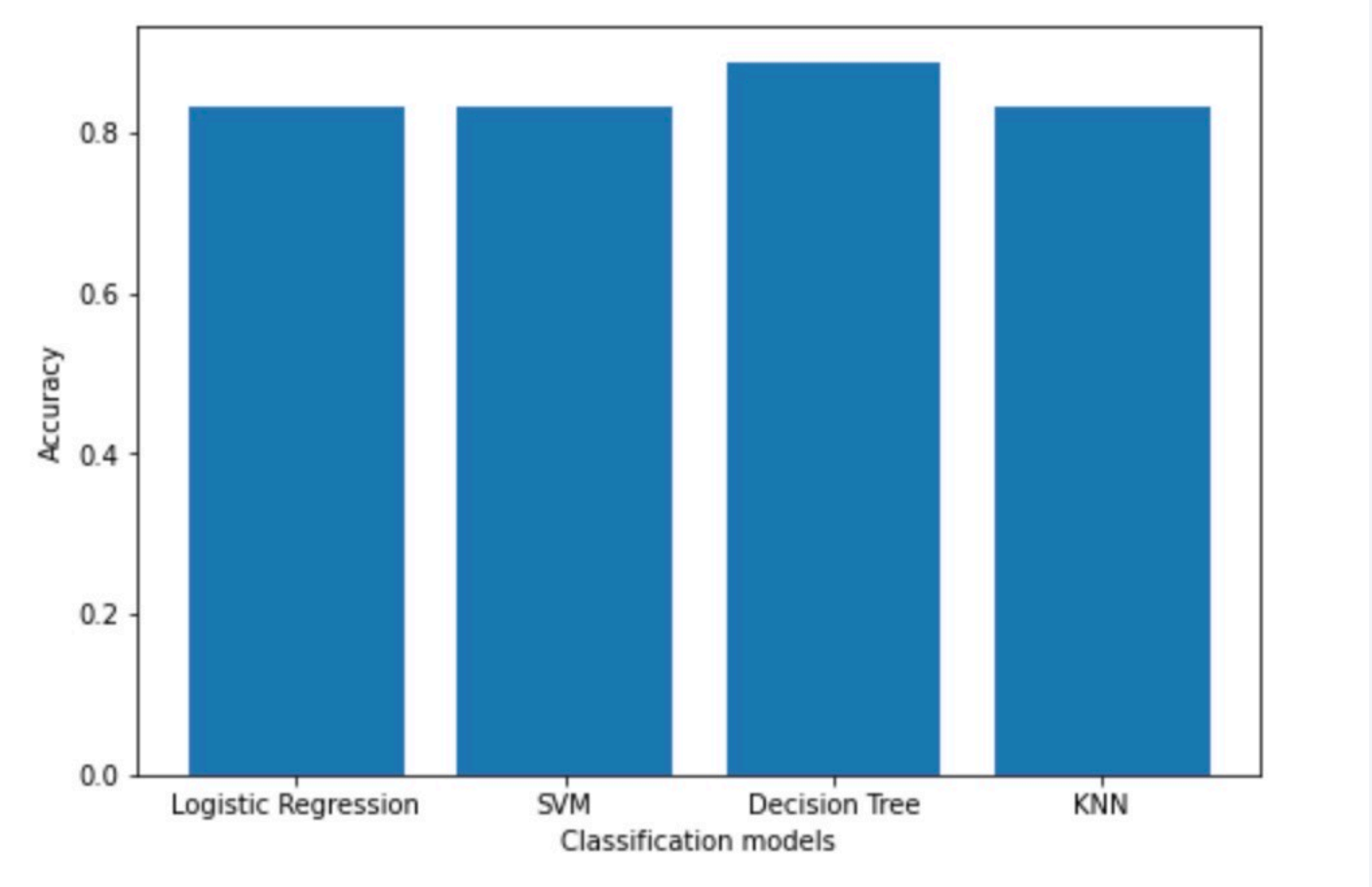
Scatter plot of Payload vs. Launch Outcome in selected payload range (here 2.5k-7k).

There seems to be no clear visual pattern detected in payload range up to 5,500.

Section 5

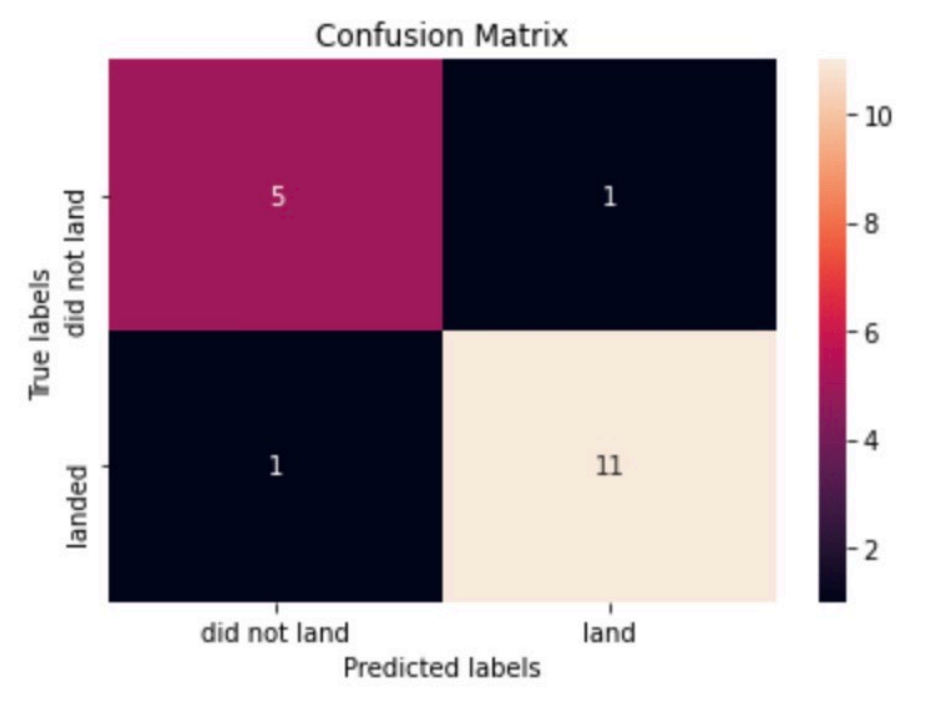
Predictive Analysis (Classification)

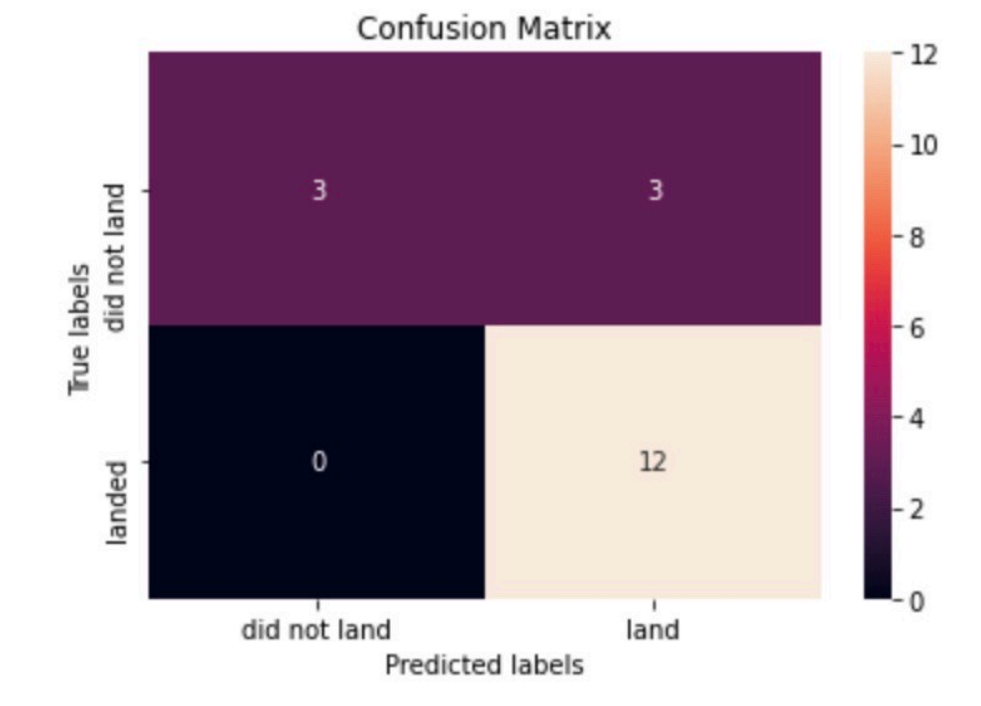
Classification Accuracy



The Decision Tree classification model has the best **accuracy** score.

Confusion Matrix



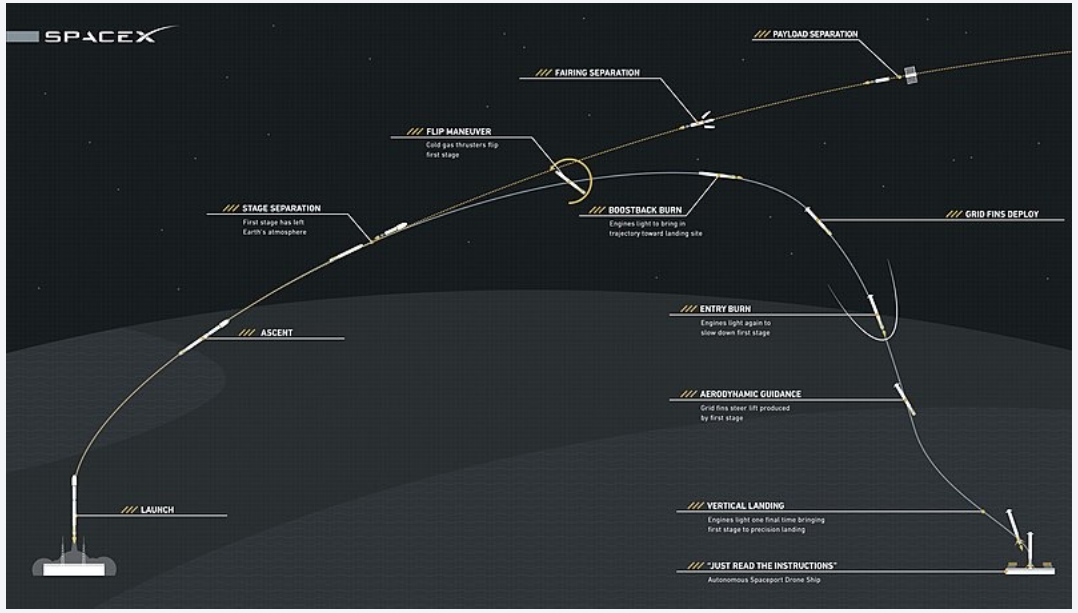


The Decision Tree classification model (**left**) has higher **pression** than other models (**right**).

Conclusion

- The Cape Canaveral Air Force Station (CCAFS) is suitable launch site for flights with light payload and the Kennedy Space Center (KSC) is for flights with heavy payload.
- Flights aiming to the geosynchronous orbit (GTO) could be a good business opportunity for our company, because SpaceX has the least success rate in this area so far.
- Successful launch sites should be located on the coast near the Equator. However, they should be distant from big cities.
- The Decision Tree classification model could be useful to predict mission outcomes, and to estimate the cost of a launch.

Appendix 1



SpaceX conducts the **SpaceX reusable launch system development program** since 2011 "*to develop a set of new technologies for an orbital launch system that may be reused many times in a manner similar to the reusability of aircraft*."

"SpaceX reusable launch system development program," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=SpaceX_reusable_launch_system_development_program&oldid=1060557886

See also: Lars Blackmore, "Autonomous Precision Landing of Space Rockets," 2016.

Appendix 2-1

	Class
Block	0.416015
ReusedCount	0.466584
GridFins_False	-0.642540
GridFins_True	0.642540
Legs_False	-0.673825
Legs_True	0.673825
Class	1.000000

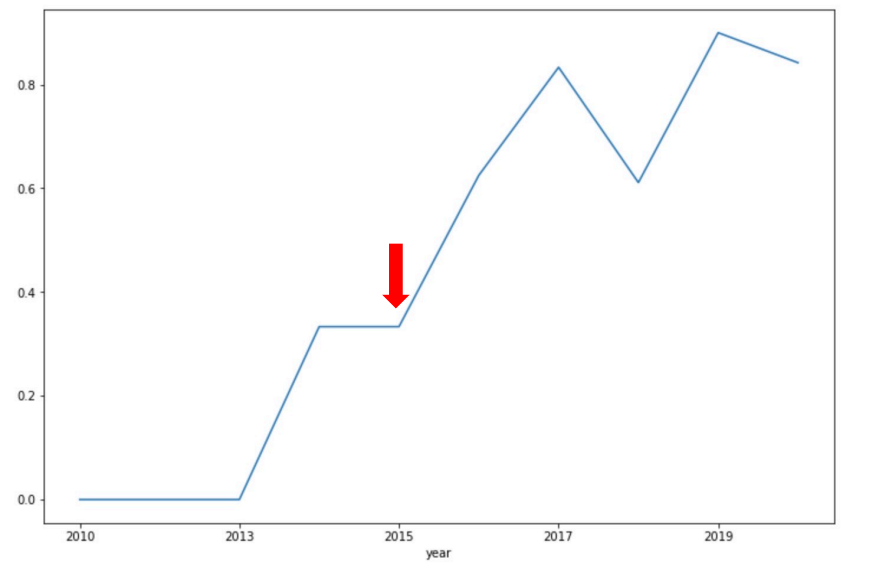
Correlation (**Pearson correlation coefficient**) in the data which was used for the Machine Learning Prediction lab is explored.

• Among 83 features, grid fins (**GridFins**) and landing legs (**Legs**) show especially strong correlation with target (**Class**).

• These landing technologies are developed within the the **SpaceX reusable launch system development program** (s. Appendix 1).

* Link to the notebook for this extra EDA

Appendix 2-2



Line chart of yearly average success rate

Our data shows that grid fins are first used for the Flight 12 on 10.01.2015, and landing legs for the Flight 7 on 18.04.2014.

These technological developments correspond exactly with a major turning point in the Falcon 9 launch history.

Thank you!