IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ildze Freiberga
16 July 2024

Outline

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

Executive Summary

- Performing data collection, data wrangling, exploratory data analysis, interactive visual analytics and predictive analysis, we are able to predict SpaceX's ability to reuse the first stage of Falcon 9 rocket
- Results:
  - Falcon 9 first stage landing success depends on:
    - Launch site
    - Orbit
    - Payload mass
    - Booster version
  - The best machine learning classifier can predict the first stage landing success rate with an accuracy ~94%

Introduction

- SpaceX is an American aerospace company founded in 2002 by Elon Musk that helped usher in the era of commercial spaceflight. Its name in full is Space Exploration Technologies Corporation.
- SpaceX advantages:
    - Less expensive rocket launches comparing to other companies
    - SpaceX's ability to reuse the first stage of the rocket
- How can we predict whether the first stage will land to estimate the cost of a launch?
- This data can be utilized if another firm wishes to compete with SpaceX for a rocket launch.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Collecting data, requesting SpaceX API
  - Collecting data - web scraping from Wikipedia
- Perform data wrangling
  - Dealing with missing values, transforming and mapping data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare Logistic Regression, Support Vector Machine classifier, Decision tree classifier and k-Nearest Neighbor classifier

Data Collection

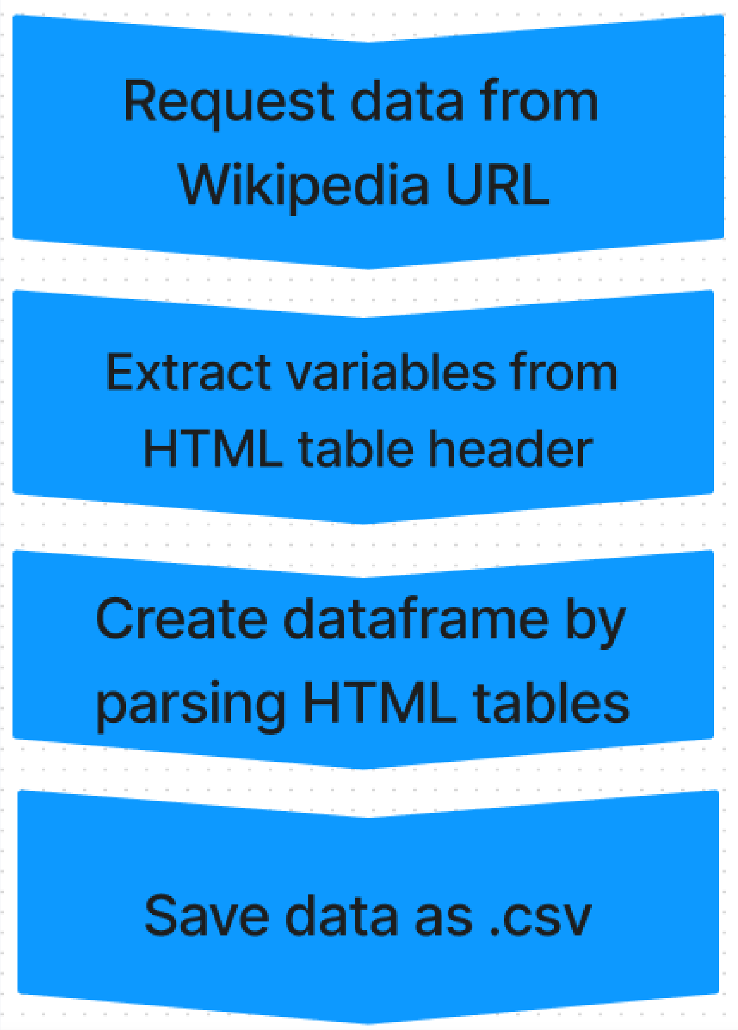
- Data was collected:
  - Requesting SpaceX API
    - https://api.spacexdata.com/v4/launches/past
  - By web scraping from Wikipedia
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- Request to the SpaceX API
- Clean the requested data
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p1_jupyter-labs-spacex-data-collection-api.ipynb

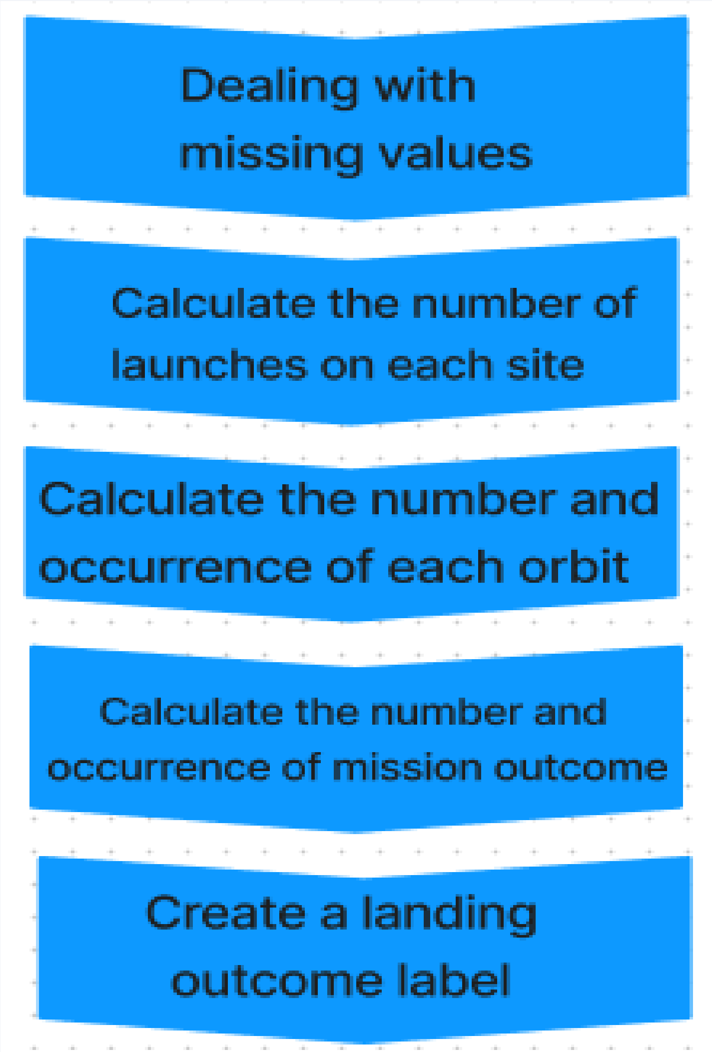
Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p2_jupyter-labs-webscraping.ipynb



Data Wrangling

- Dealing with Missing Values
- Perform Exploratory Data Analysis
- Determine Training Labels
- GitHub URL:
    - https://github.com/IldzeFre/Data-science/blob/main/p3_labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The relationship between Flight Number and Launch Site - scatter plot
- The relationship between Payload and Launch Site - scatter plot
- The relationship between success rate of each orbit type - bar plot
- The relationship between Flight Number and Orbit type - scatter plot
- The relationship between Payload and Orbit type - scatter plot
- The launch success yearly trend - line chart
- Charts advantages:
  - Scatter plot – for describing the relationship between two categorical data
  - Bar plot – for comparing categorical data
  - Line plot – for showing the time series data
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p5_edadataviz%20(1).ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass.
- 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.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p4_jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
  - to the coast
  - to the railway
  - to the highway
  - to the city
- Launch sites close to the cost and away from populated areas and infrastructure - to minimize the risk of having any debris dropping or exploding near people
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p6_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A launch site drop-down input component
- A callback function for success-pie-chart based on the selected site dropdown
- A range slicer to select payload
- A callback function for success-payload-scatter-chart scatter plot based on the selected site dropdown
- To perform interactive visual analytics on launch data in real-time in order to understand relationship between success rate and launch site/payload mass
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p7_spacex_dash_app.py

Predictive Analysis (Classification)

- Data wrangling and standardization
- Train_test_split function
- Predictive model evaluation:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- GitHub URL:
  - https://github.com/IldzeFre/Data-science/blob/main/p8_SpaceX_Machine%20Learning%20Prediction_Part_5%20.ipynb

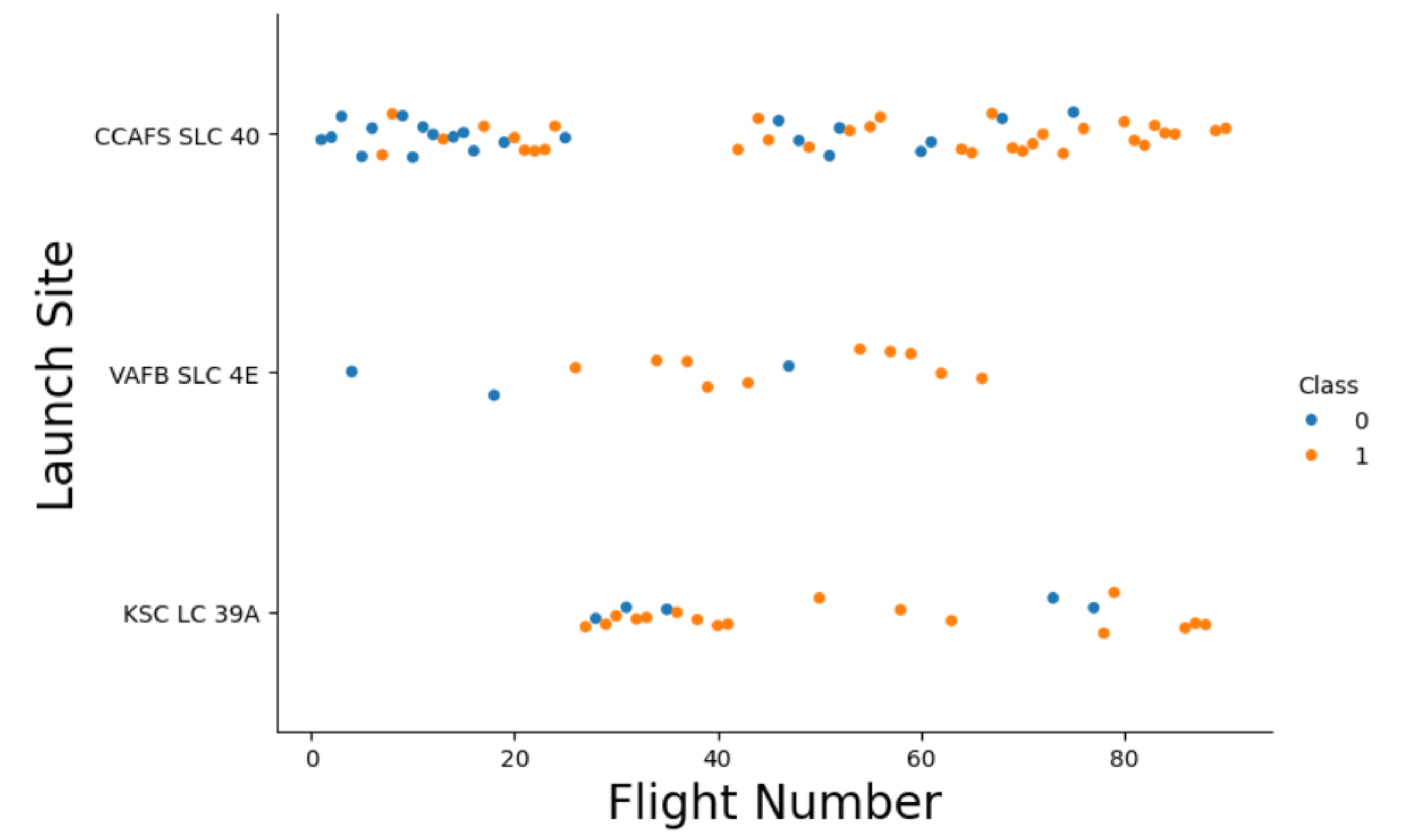
Results

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

Section 2

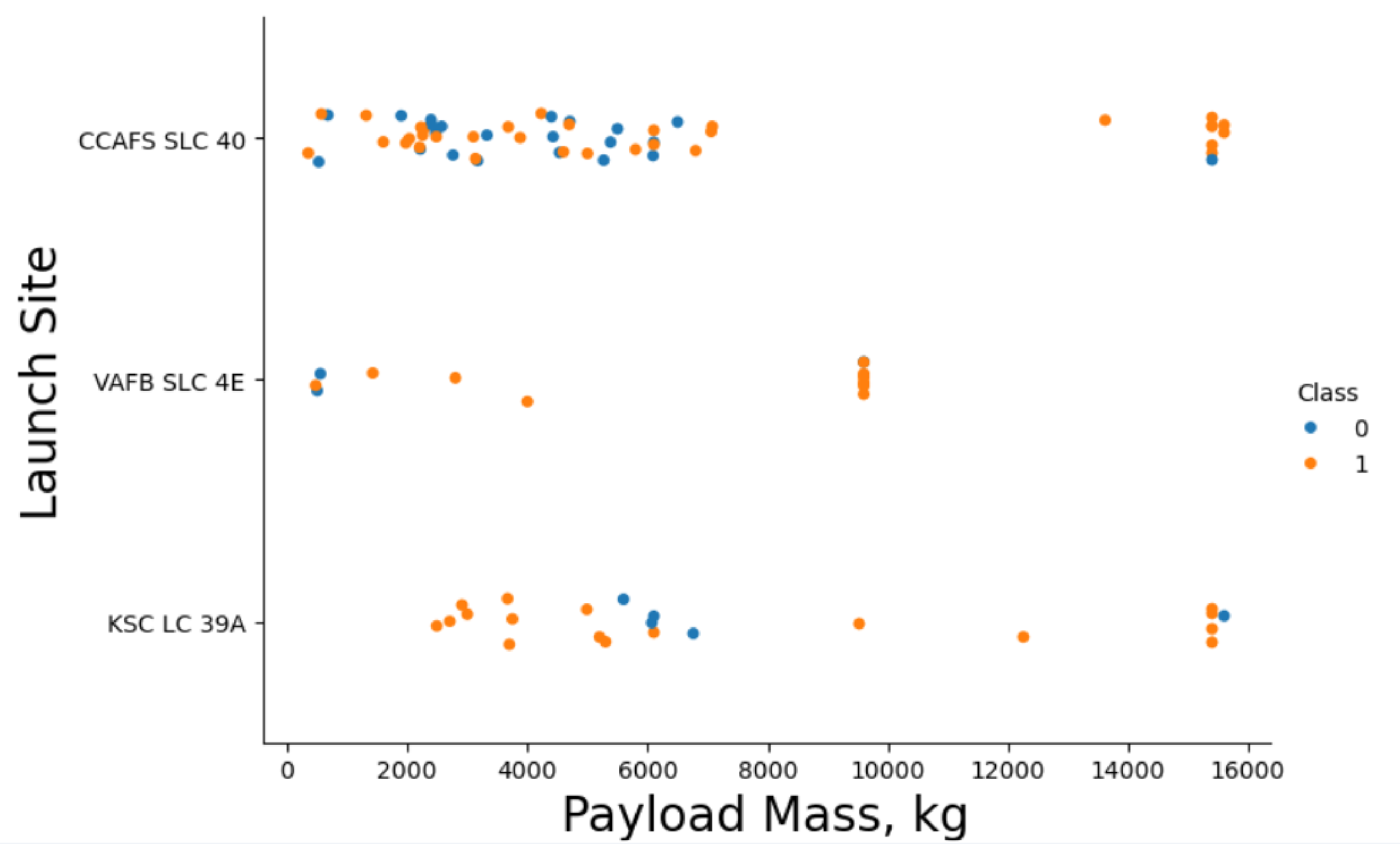
Insights drawn from EDA

Flight Number vs. Launch Site



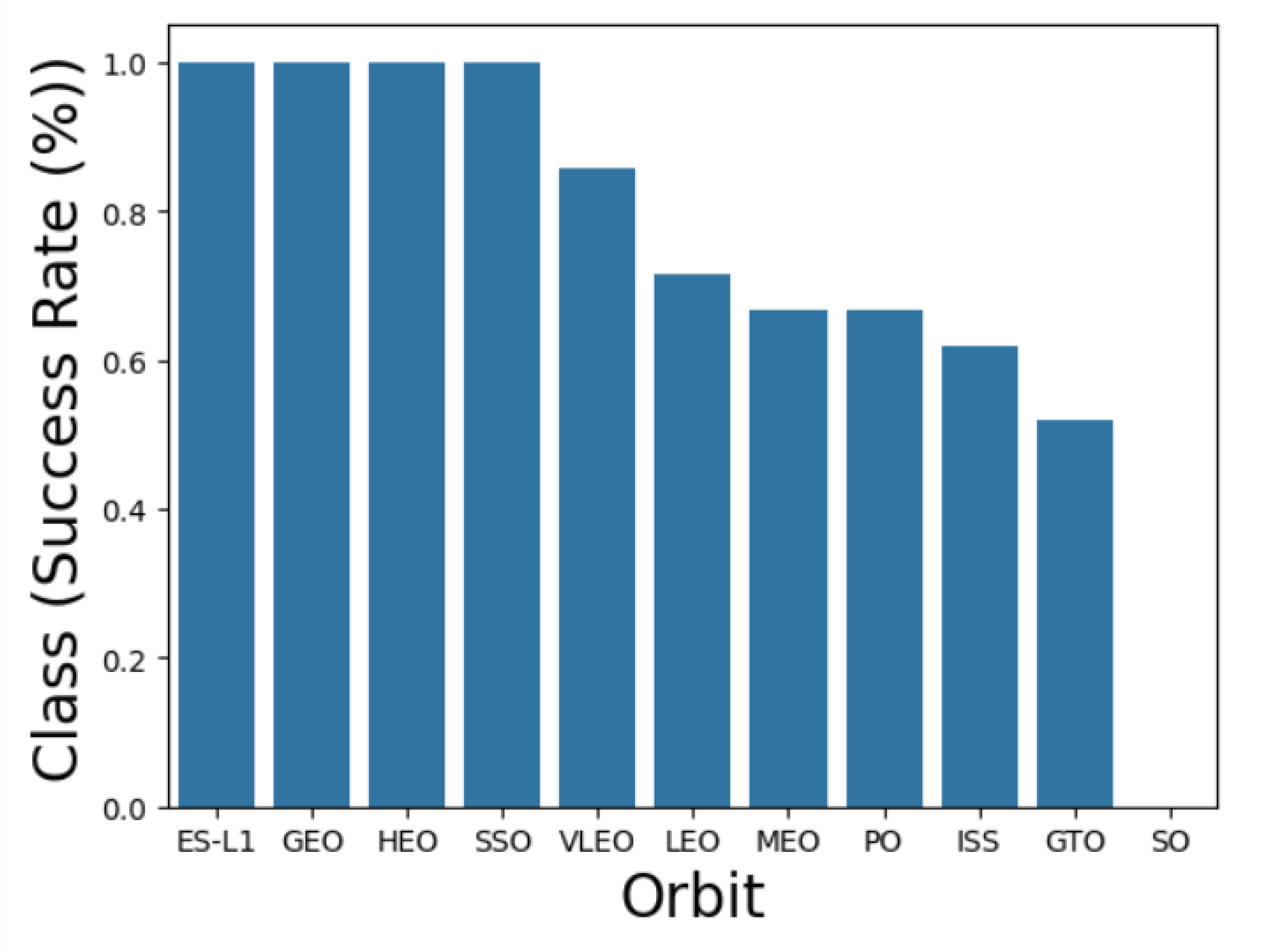
- CCAFS SLC 40 is the most used launch site.
- Different launch sites have different success rates.
- CCAFS SLC-40 has the lowest success rate
- KSC LC-39A has the highest success rate

Payload vs. Launch Site



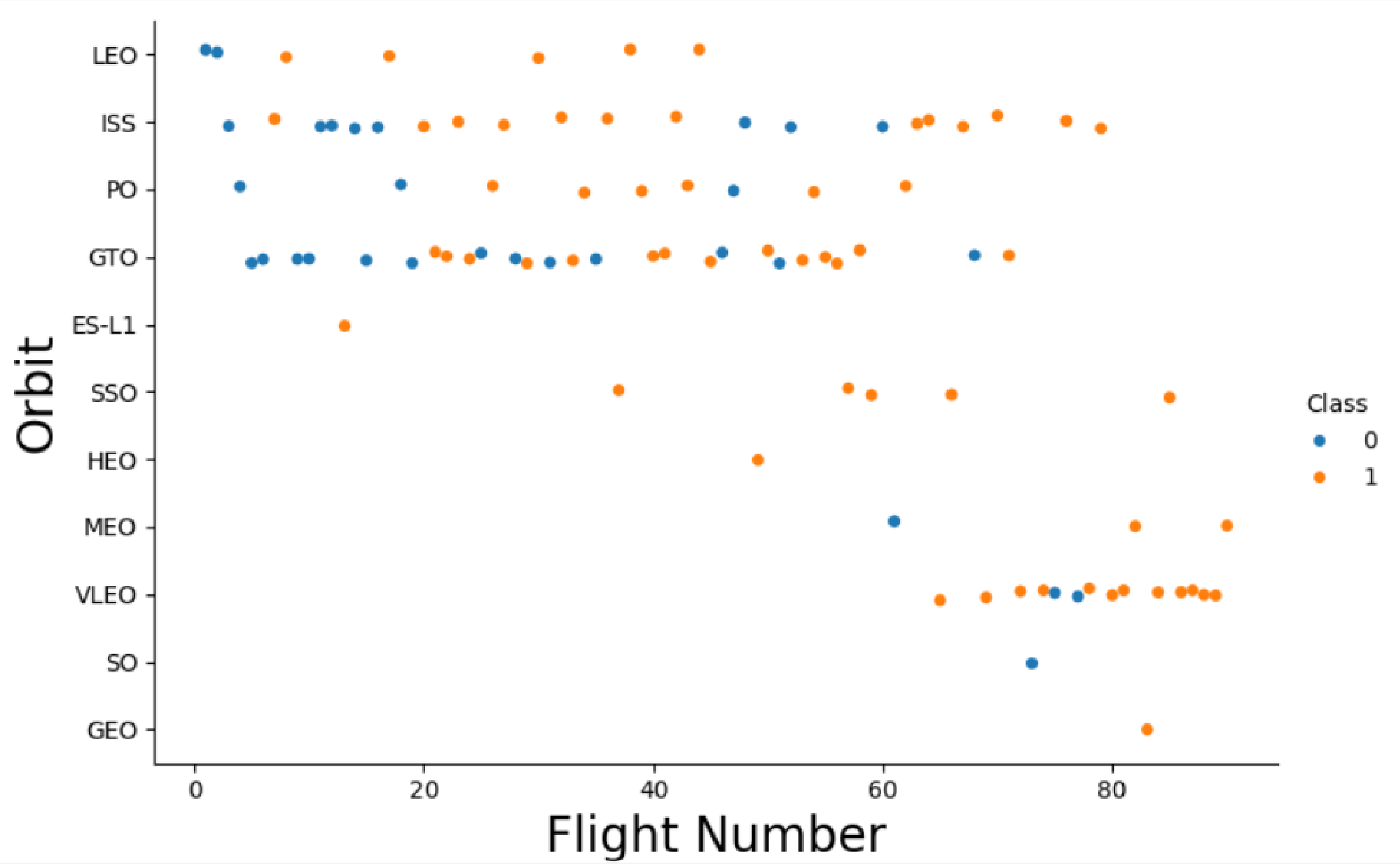
- VAFB-SLC – no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



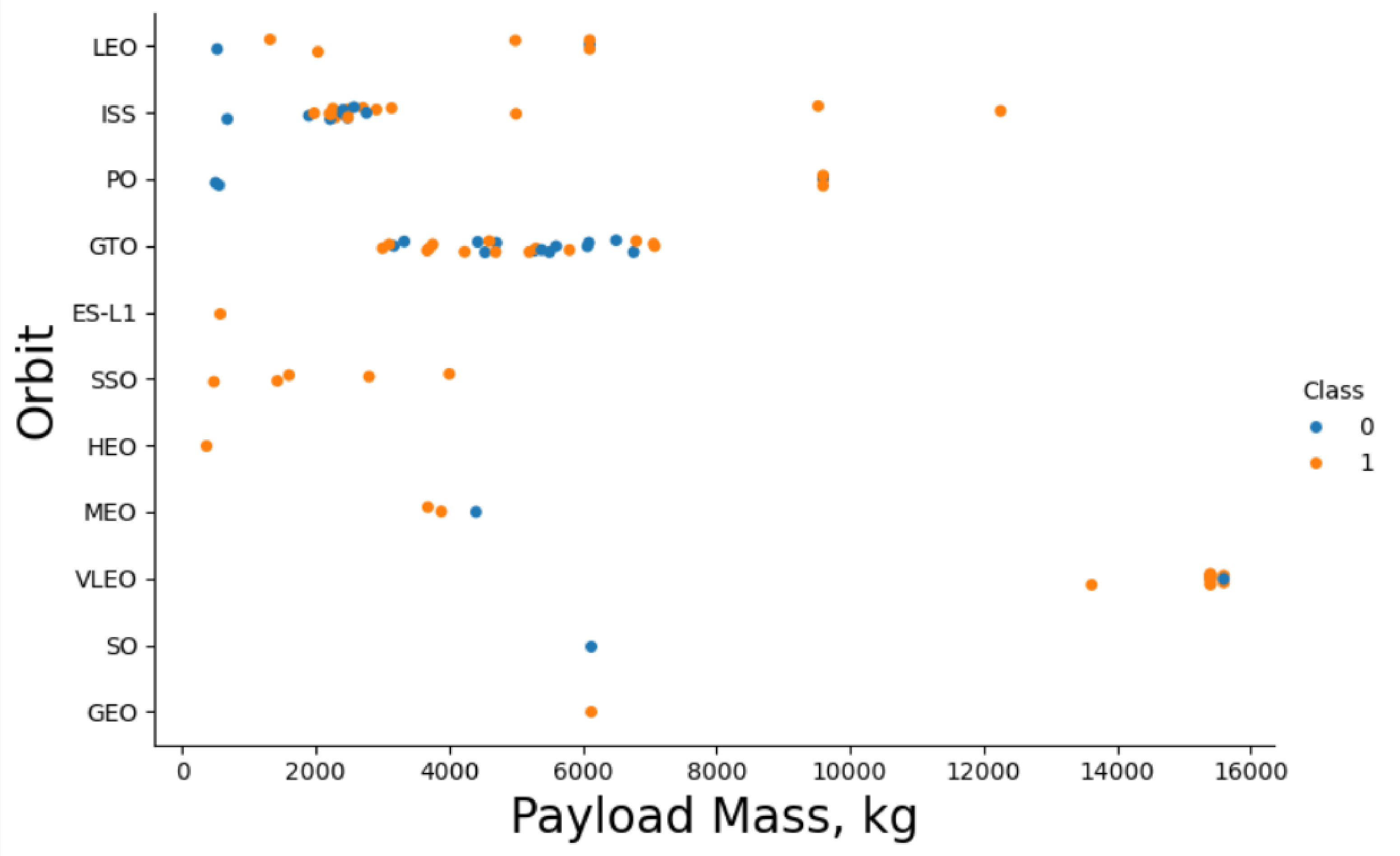
- Orbit types ES–L1, GEO, HEO and SSO have the highest success rate

Flight Number vs. Orbit Type



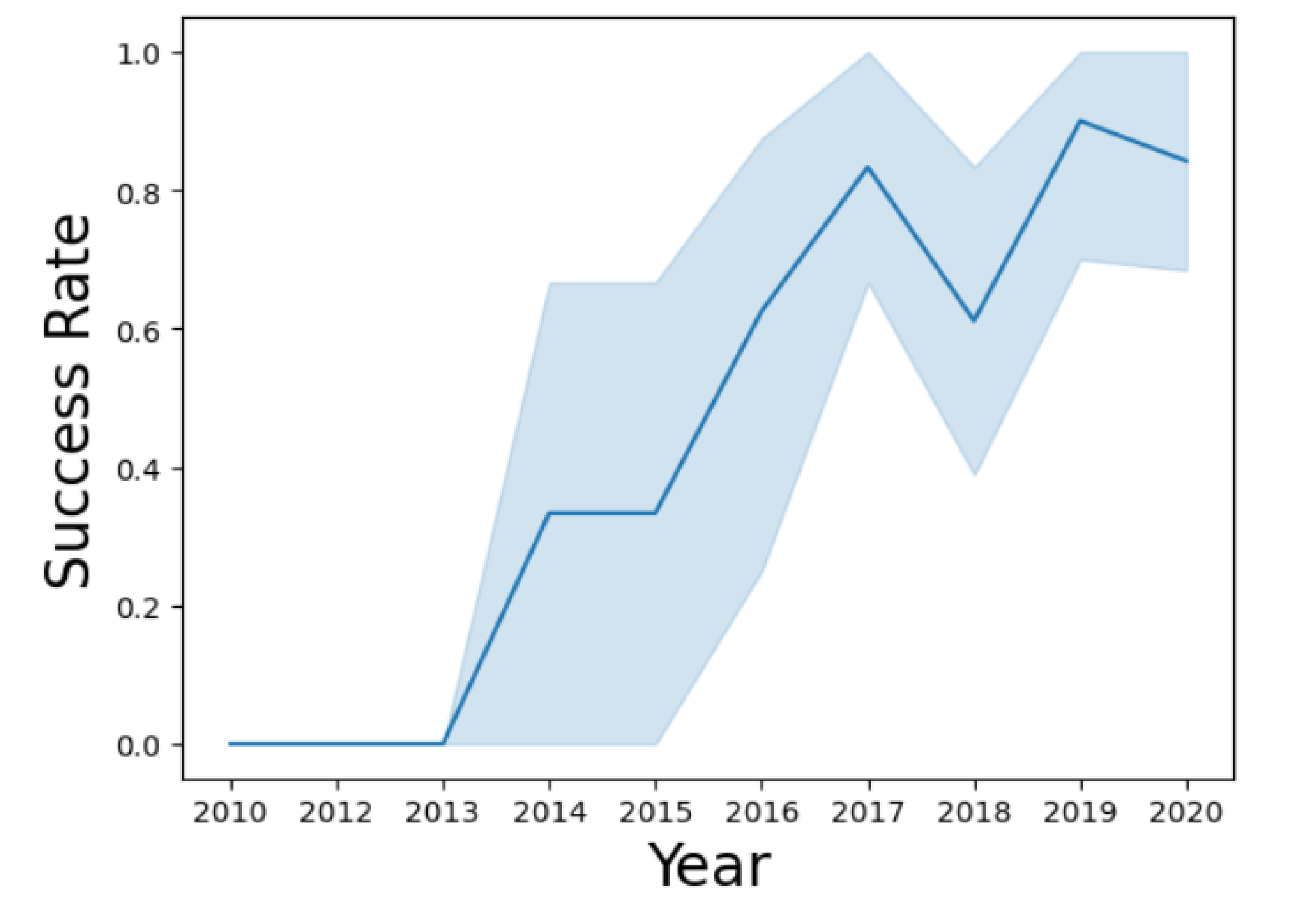
- In the LEO orbit the Success appears related to the number of flights
- No relationship between flight number in the GTO orbit

Payload vs. Orbit Type



- Positive landing rate for Polar, LEO and ISS orbits with heavy payloads
- For GTO orbit – no clear relationship

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select distinct Launch_Site as "Launch_Sites" from SPACEXTABLE;
```

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

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

- There are 4 different launch sites

Launch Site Names Begin with 'CCA'

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

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as "Total payload mass, kg", Customer from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

Total payload mass, kg	Customer
48213	NASA (CRS)

- The total payload mass for NASA (CRS) launches is 48 213 kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as "Average payload mass, kg", Booster_Version from SPACEXTABLE \
    where Booster_Version like '%f9 v1.1%';
```

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

Average payload mass, kg	Booster_Version
2534.6666666666665	F9 v1.1 B1003

- The average payload mass carried by booster version F9 v1.1 is 2534.67 kg

First Successful Ground Landing Date

```
%sql select min(Date) as "Date", Landing_Outcome from SPACEXTABLE where Landing_Outcome="Success (ground pad)";
```

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

Date	Landing_Outcome
2015-12-22	Success (ground pad)

- The first successful landing outcome on ground pad was on 22.12.2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE \
    where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- There are 4 different 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 Mission_Outcome, count(Mission_Outcome) as "Total Mission_Outcomes" from SPACEXTABLE group by Mission_Outcome;
```

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

Mission_Outcome	Total Mission_Outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The total number of successful (99), successful with unclear payload status (1), and failure (1) mission outcomes

Boosters Carried Maximum Payload

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE \
    where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

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

- There are 12 different boosters which have carried the maximum payload mass

2015 Launch Records

```
%sql select Date, substr(Date,0,5) as "Year", substr(Date,6,2) as "Month", Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE \
    where Landing_Outcome="Failure (drone ship)" and substr(Date,0,5)='2015';
```

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

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015

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

```
%sql select Date, Landing_Outcome, count(Landing_Outcome) as "Total Landing_Outcomes" from SPACEXTABLE \
    where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc;
```

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

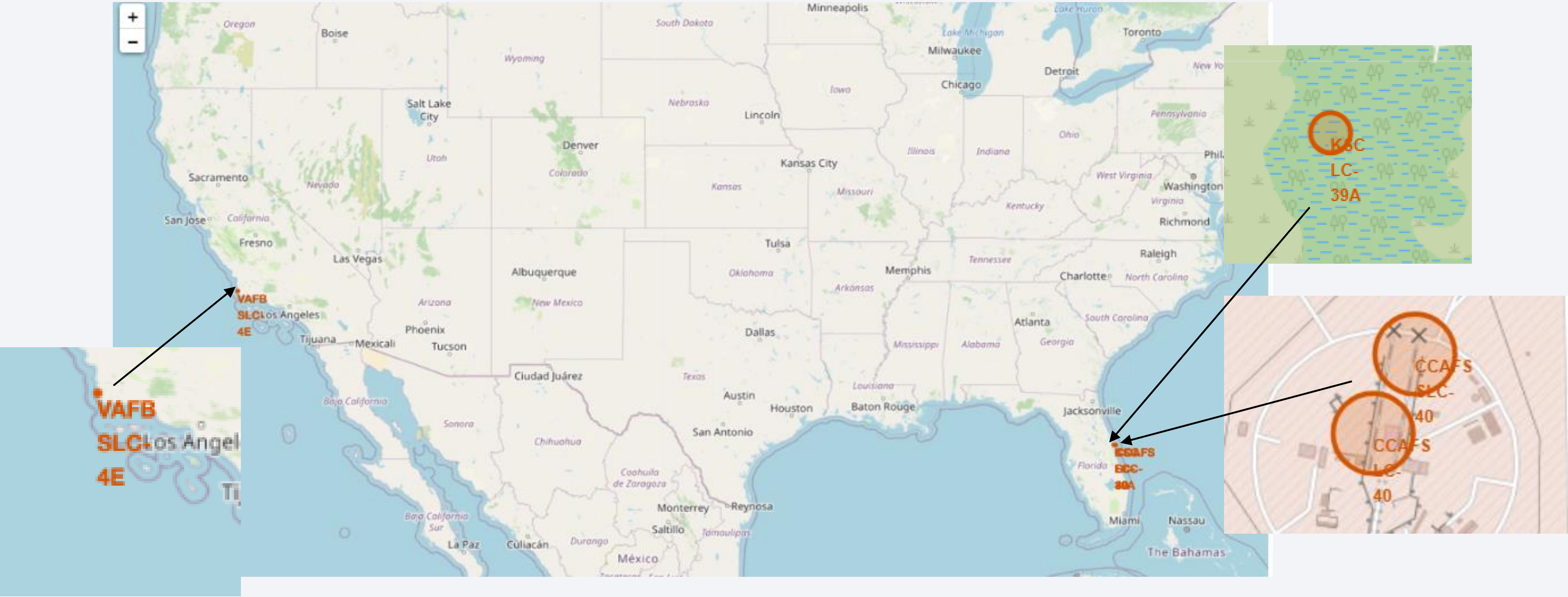
Date	Landing_Outcome	Total Landing_Outcomes
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

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

Section 3

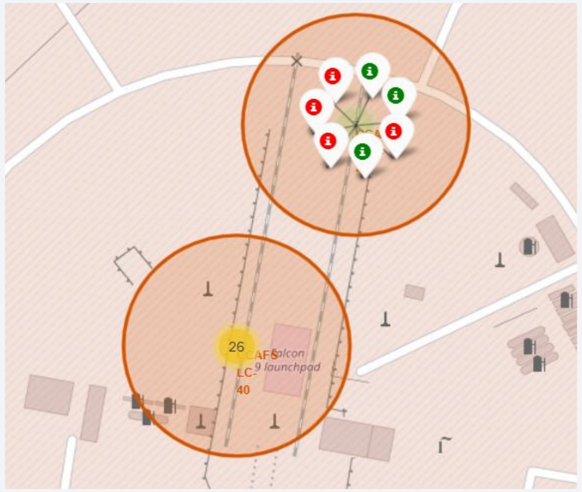
Launch Sites Proximities Analysis

Launch Sites



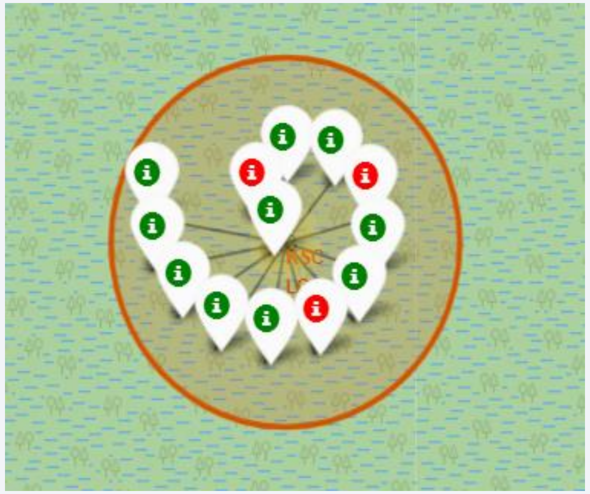
- All launch sites are in the close proximity to the Equator line as that provides some additional orbital speed to the launch vehicle
- All launch sites are in very close proximity to the coast to minimize the risk of having any debris dropping or exploding near people

Launch Outcomes of different Launch Sites

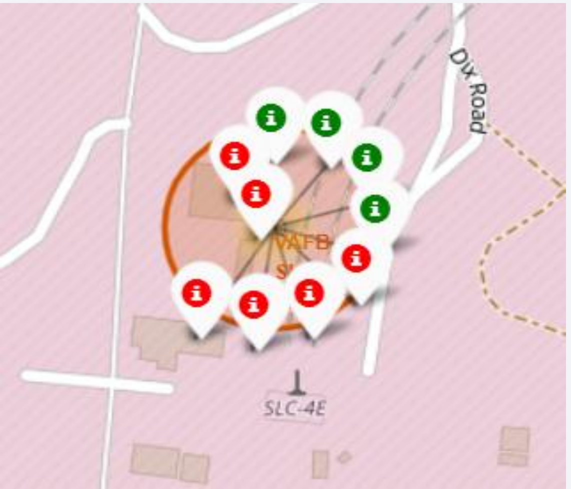


CCAFS SLC-40

CCAFS LC-40



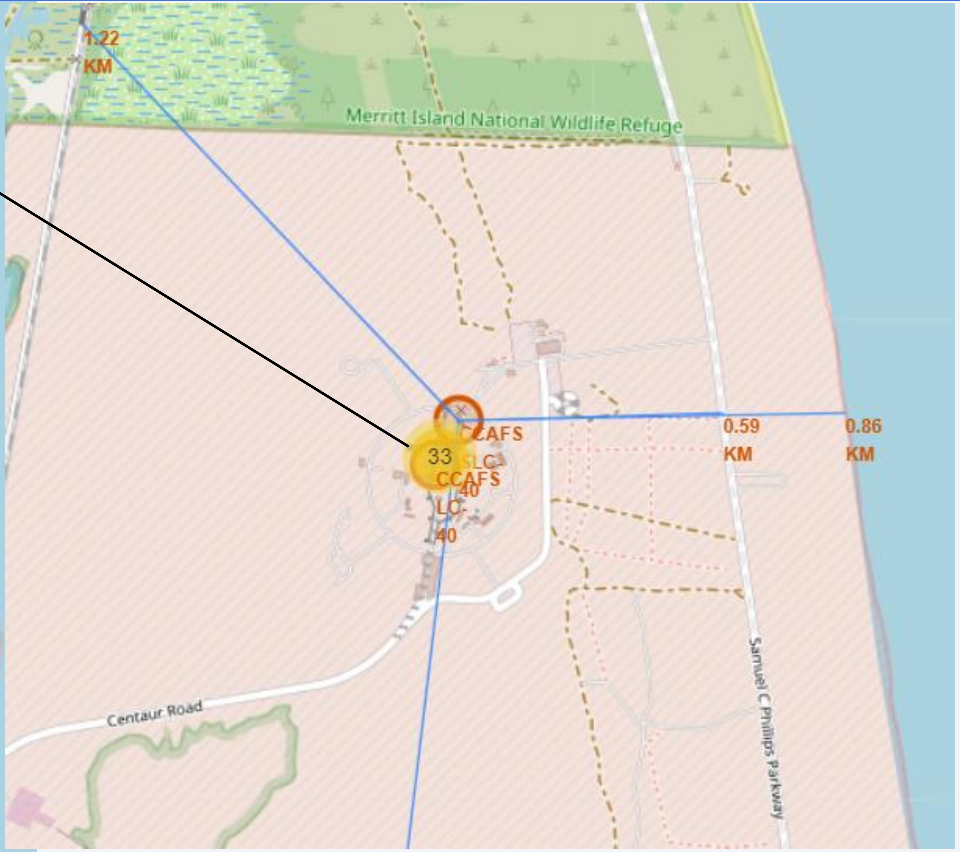
KSC LC-39A



VAFB SLC-4E

- CCAFS SLC 40 is the most used launch site.
- Different launch sites have different success rates.
- CCAFS SLC-40 has the lowest success rate
- KSC LC-39A has the highest success rates.

Distance from CCAFS SLC-40 to its proximities



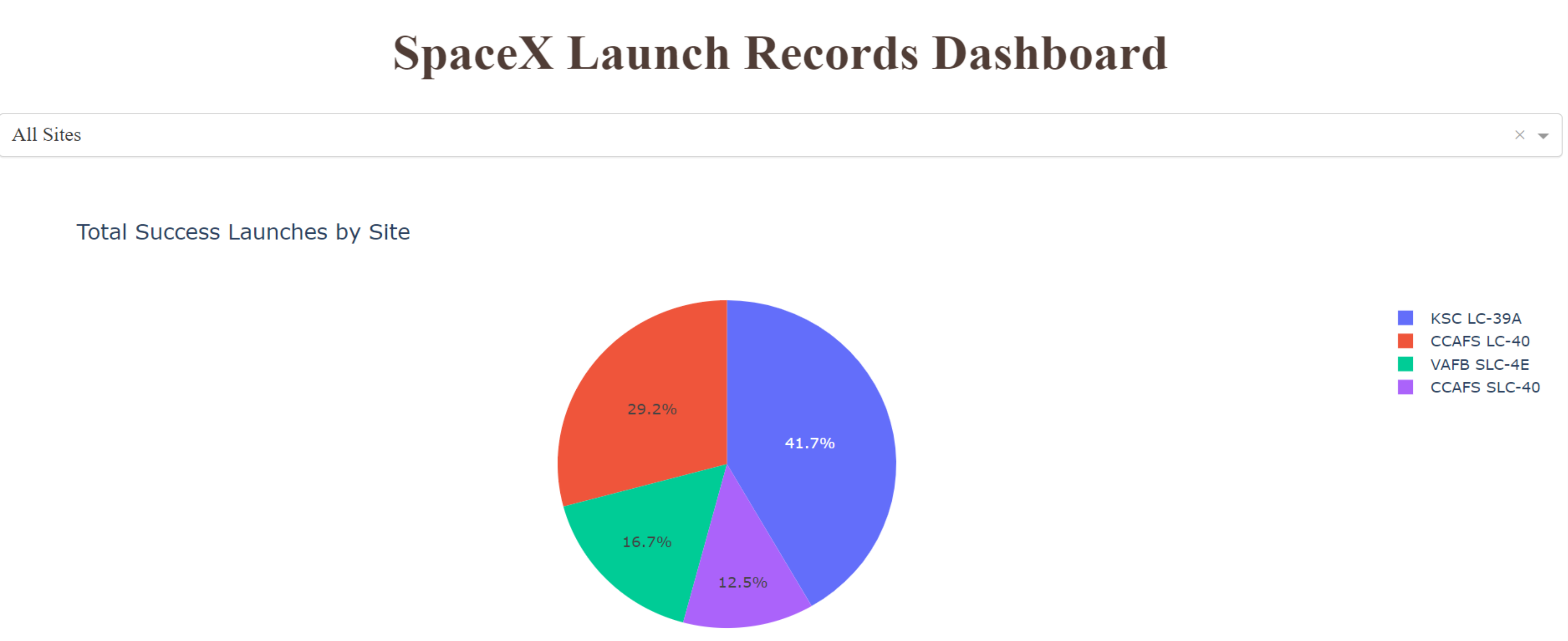
- Distance from CCAFS SLC-40 launch site to its proximities:
  - Coastline - 0.86km
  - Highway - 0.59km (Samuel C Phillips Parkway)
  - Railway - 1.22km (NASA Railroad)
  - City - 18.24km (Cape Canaveral)

- Launch sites are close to the cost and away from populated areas and infrastructure – to minimize the risk of having any debris dropping or exploding near people

Section 4

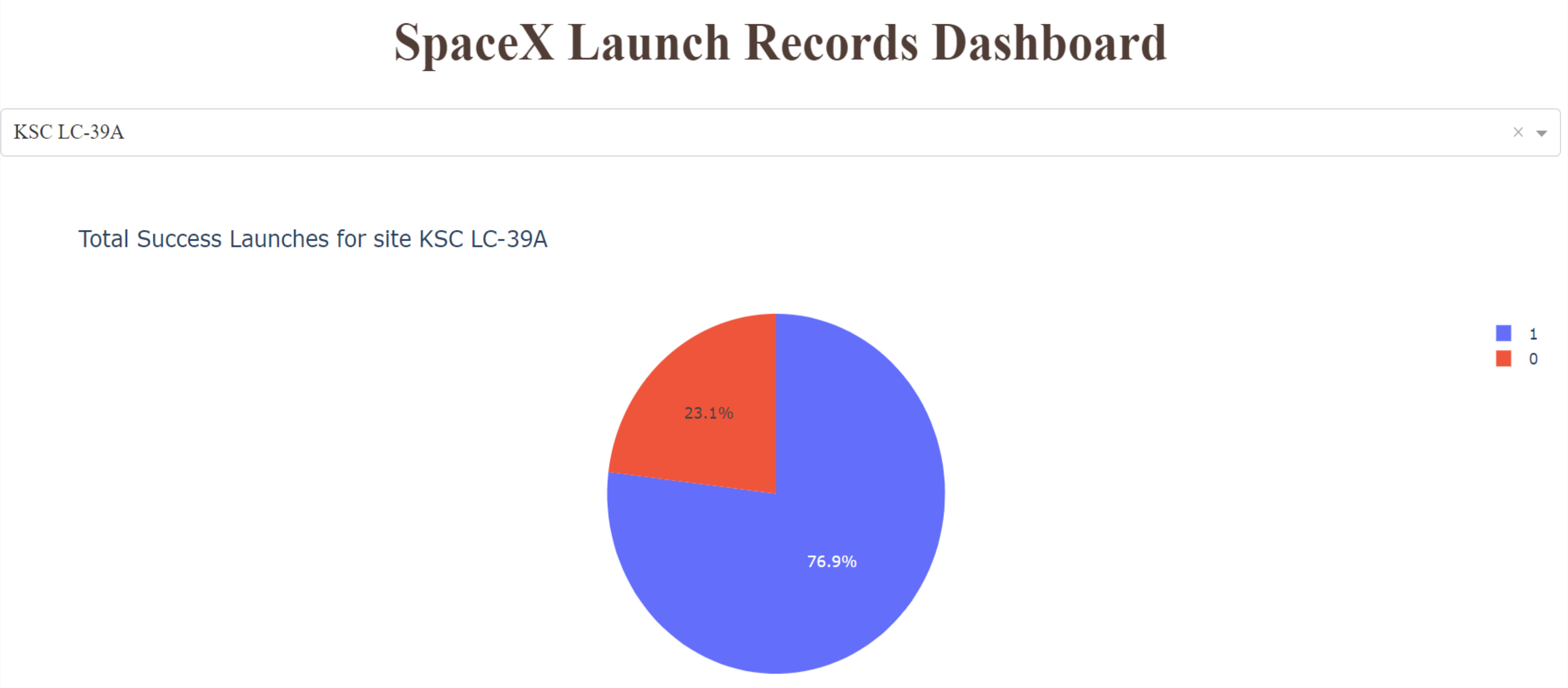
Build a Dashboard with Plotly Dash

Pie chart of Launch Success for all Sites



- KSC LC-39A: 41.7%
- CCAFS LC-40: 29.2%
- VAFB SLC -4E: 16.7%
- CCAFS SLC-40: 12.5%

Pie chart for the Launch Site with highest Success ratio



- Launch site with highest launch success ratio - KSC LC-39A

- Success ratio - 76.9%

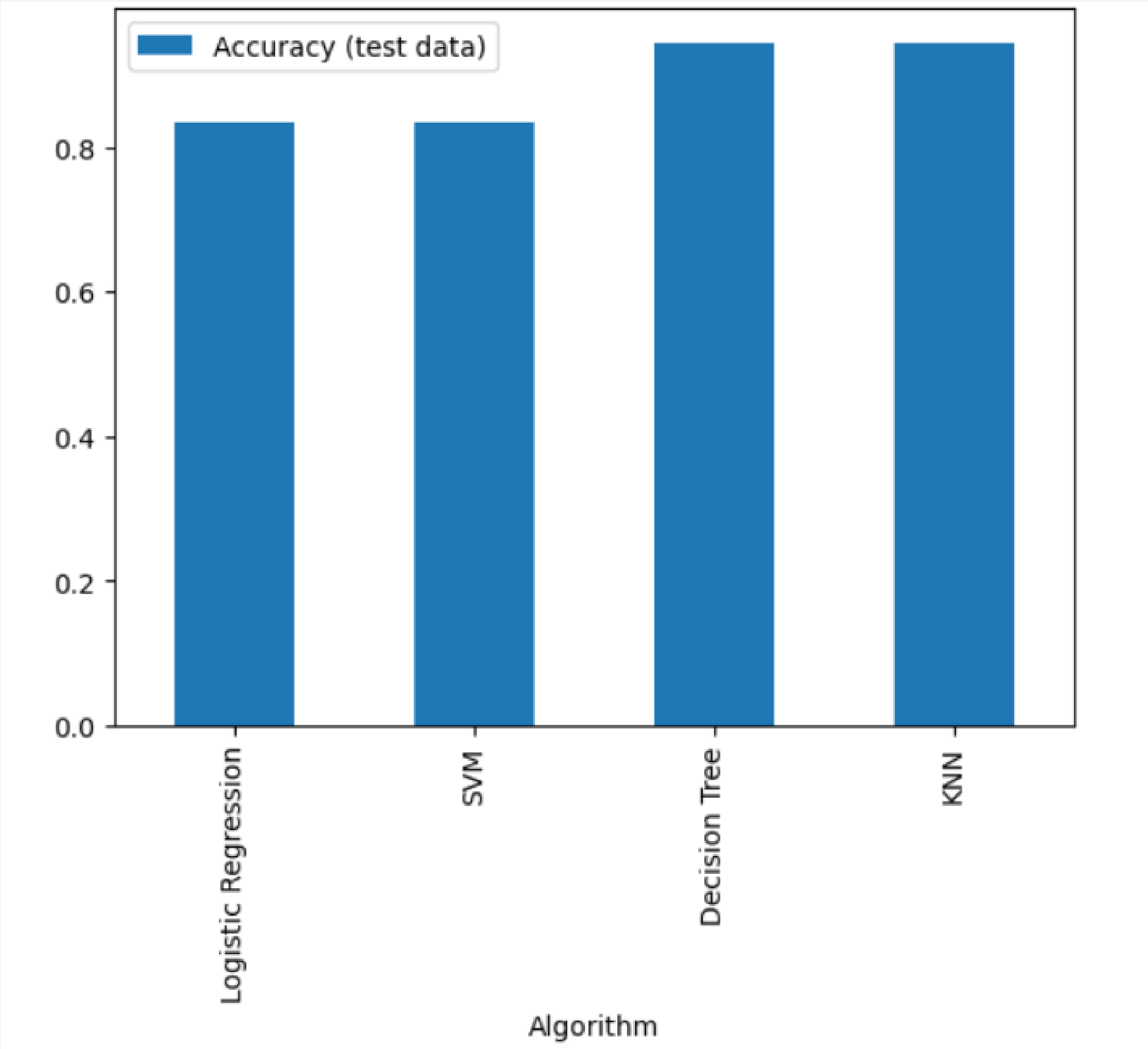
Payload vs. Launch Outcome scatter plot

- Booster versions v1.0 and v1.1 have the lowest success rate
- FT booster has the highest success rate for payload under 6000kg.
- Heavy payload decreases success ratio.

Section 5

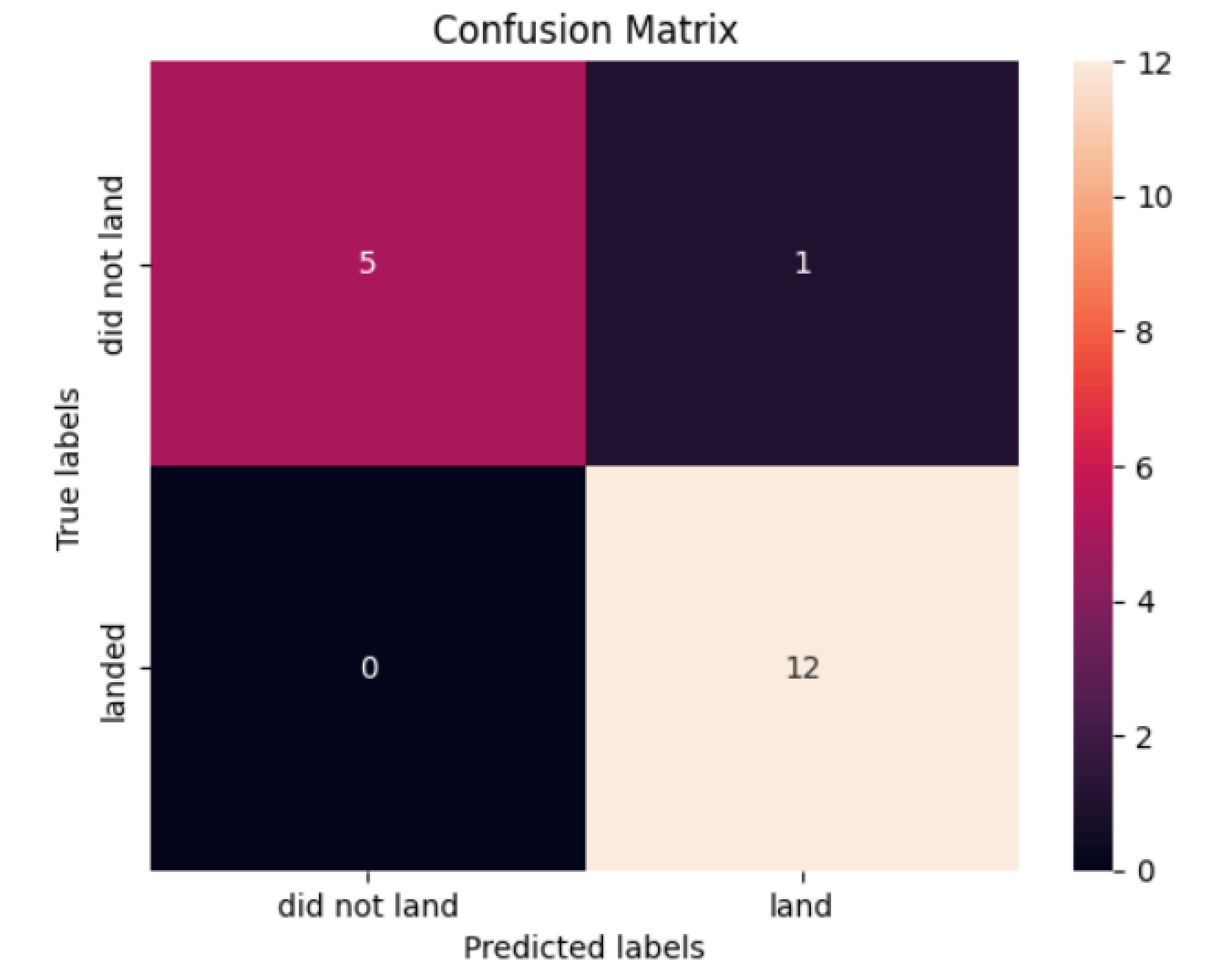
Predictive Analysis (Classification)

Classification Accuracy



- Decision tree and k-Nearest Neighbor algorithms have the highest classification accuracy (~94%)
- However, the test set is too small

Confusion Matrix



- Confusion matrix of the best performing model – Decision tree
- There are 5 true positive, 12 true negative, 1 false positive and 0 false negative predictions
- Decision tree can distinguish bet ween the different classes. Insignificant problem is 1 false positive prediction

Conclusions 1/2

- Prediction of SpaceX's ability to reuse the first stage of Falcon 9 rocket success rate depends on:
  - Launch site:
    - CCAFS SLC 40 is the most used launch site
    - Different launch sites have different success rates.
    - CCAFS SLC-40 has the lowest success rate
    - KSC LC-39A has the highest success rate (77%)
  - Orbit:
    - In the LEO orbit the Success appears related to the number of flights
    - Orbit types ES–L1, GEO, HEO and SSO have the highest success rate
  - Payload mass:
    - Positive landing rate for Polar, LEO and ISS orbits with heavy payloads
    - VAFB-SLC – no rockets launched for heavy payload mass (greater than 10000)
    - Heavy payload decreases success ratio.
  - Booster version:
    - Booster versions v1.0 and v1.1 have the lowest success rate
    - FT booster has the highest success rate for payload under 6000kg

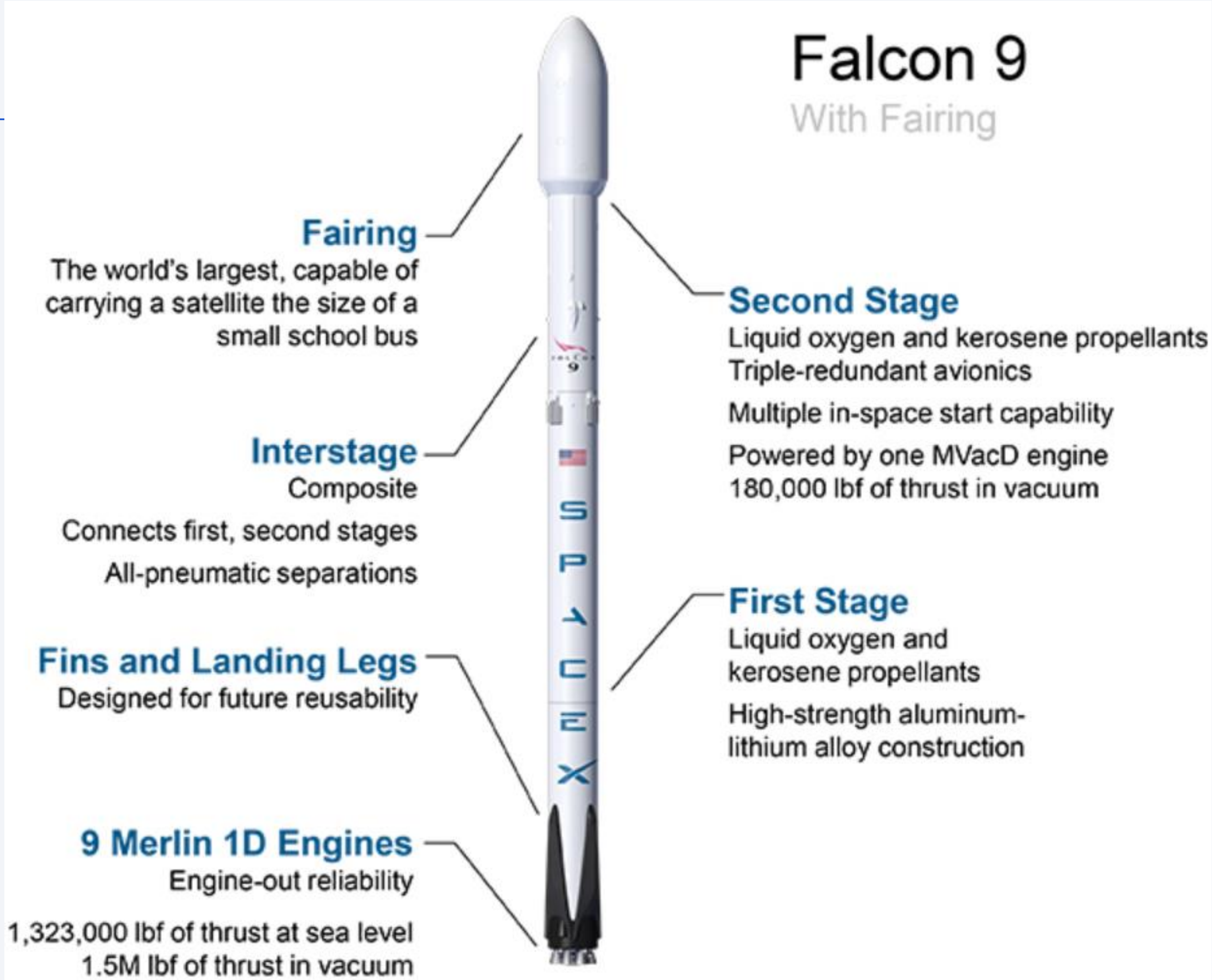
Conclusions 2/2

- The success rate since 2013 kept increasing
- The best machine learning classifiers Decision tree and k-Nearest Neighbor can predict the first stage landing success rate with an accuracy ~94%

Appendix 1/2

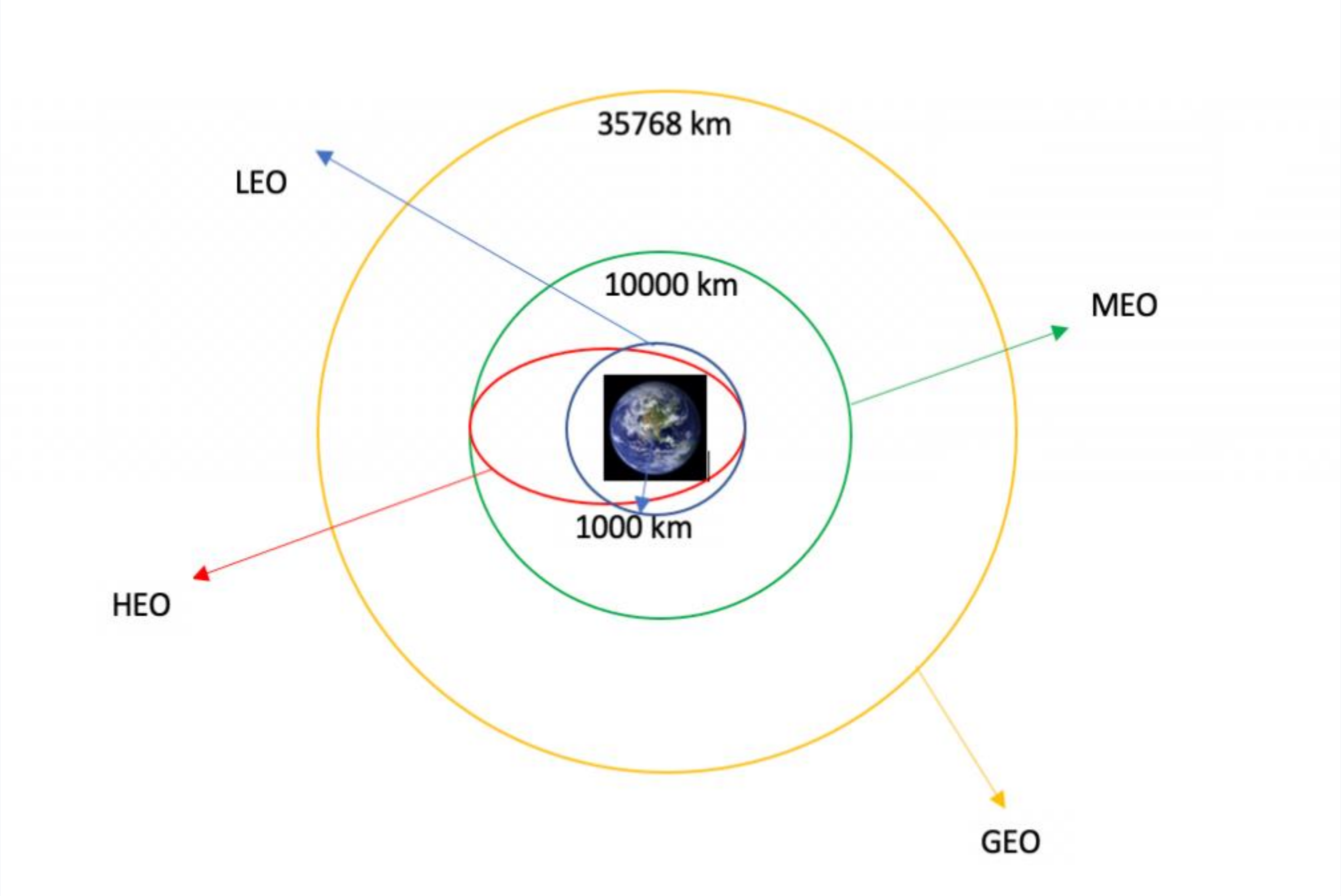
- Falcon 9 rocket

https://asd.gsfc.nasa.gov/archive/tess/images/spacex/spacex_falcon9_stage_image_lg.jpg



Appendix 2/2

- Orbit types



- GitHub project link https://github.com/IldzeFre/Data-science/tree/main

Thank you!