IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jesse Huang, MPH PhD
July 21, 2013

Outline

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

Executive Summary

- **AIM:** Accurately predicting the whether SpaceX will reuse their rockets can provide important applications for estimating business operation costs.
- **Method**: SpaceX launch data was obtained by access of SpaceX REST API and from web scrapped of relevant Wikipedia tables of SpaceX launches. This includes numerous launch parameters, along with the outcome variable of whether the rockets were reused.
- In general, launch have become much more reliable (after first 50 launches), more success landing of rocket boosters.
- Large payloads (>5000kg) have to use FT or B4 booster versions, which have historically failed more often.
- Logistic Regression Models and SVM was able to predict Launch Success (booster landing) with an accuracy of 83%.

Introduction

- The commercialized space travel has become a possibility with recent developments from companies such as SpaceX and Virgin Galactic.
- Although still not commonly accessible due to its costs, recent technological developments, particularly with reusable booster rockets, have significantly reduced the cost of launching into space.
- For SpaceX, the recovery and reuse of the first stage rockets directly affects the cost of each launch.
- Therefore, accurately predicting the whether SpaceX will reuse their rockets can provide important applications for estimating business operation costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX rocket launch data are fetched from SpaceX REST API (requests) and Wikipedia Launch Data Tables (Webscraping).
- Perform data wrangling
  - Raw API request data and web-scrapped tables were cleaned and processed to remove NAs, metadata, unstructured/semi-structured data, and inconsistencies.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Comparing logistic regression, SVM, decision tree, KNN

Data Collection

- SpaceX rocket launch data are fetched from two different sources.
  - SpaceX REST API (requests)
  - Wikipedia Launch Data Tables (webscraping: BS4, BeautifulSoup)
- Variables – (Identifier/Predictors)
  - Rocket: Booster name (identifier)
  - Launchpad: launch site, longitude, and latitude
  - Payload: mass of the payload, the orbit
  - Cores: outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core, number of times this has been reused, and the serial of the core
  - Outcome (T/F): whether boosters successfully landed (regardless of condition)

Data Collection – SpaceX API

- First dataset is requested from the official SpaceX Rest API.
    - api.spacexdata.com/v4/launches/past
- The initial request fetches all historical launch data and related variables (up to 2021).
- Further requests fetch details of each variable, often linked by different identifiers (foreign keys).
- These include measures for the **rocket**, **launchpad**, **payloads**, and **cores**.
- Code for requesting SpaceX API is provided in (Part 1) of 1_SpaceXData_fetchData.ipynb.
    - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/1_SpaceXData_fetchData.ipynb

1. Initial API Request for past Launch Data

2. Extract and Process Launch Variables Ids (foreign keys in database)

3. Fetch additional measures using API

Data for ROCKET

Data for LAUNCHPAD

Data for PAYLOADS

Data for CORES

Data Collection - Scraping

- Additionally, launch data were also web-scrapped from HTML tables.
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The webpage is first downloaded and then processed using BeautifulSoup
- Each table (by year of launch) was extracted and converted to dataFrame.
- Code for requesting webscrapping is provided in (Part 2) of 1_SpaceXData_fetchData.ipynb.
    - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/1_SpaceXData_fetchData.ipynb

Download request of webpage (Wikipedia)

HTML file is imported and converted in BS4

HTML for each table is converted to pandas dataFrame using read_html (table are by year).

Tables are joined and cleaned to remove any unnecessary metadata.

Data Wrangling

- Raw API request data and web-scrapped tables were processed by:
  - NA/empty values (required measures)
  - metadata, unstructured/semi-structured data
  - inconsistencies in categories/values
  - removed unnecessary measures
- All JSON and HTML tables were structured and all relevant data were converted to pd.dataFrames prior to writing.
- Defining variables/outcomes:
  - Relevant predictor variables were defined and processed
  - Outcome variable was defined as binary (T/F), success/failure
- Code for data wrangling is provided in (Part 3) of 1_SpaceXData_fetchData.ipynb:
  - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa51420 6e9c9c82fff8ccba3c7fb1ae7c132/1_SpaceXData_fetchData.ipynb

EDA with Data Visualization

- Explore and visualize SpaceX launch data
  - Assess Launch Success(1/0) by launch site and flight number
  - Assess Launch Success (1/0) by payload mass and launch site
  - Compare success rate by orbit type
  - Assess Launch Success (1/0) by orbit and flight number
  - Assess Launch Success (1/0) by payload mass and orbit
  - Success rate over time (trend)
- Code for data exploration and initial visualization is provided in 2_SpaceXData_dataExplore_1.ipynb:
  - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/2_SpaceXData_dataExplore_1.ipynb

EDA with SQL

- Summary of SQL queries of data
  - Find the names of all launch sites (unique values)
  - Find first five records of 'CCA' launch sites
  - Calculate the total payload mass carried by boosters launched for NASA (CRS)
  - Calculate the average payload mass carried by booster version F9 v1.1
  - Fetch date of the first succesful landing outcome in ground pad
  - List boosters with success in drone ship and have payload mass between 4000 and 6000
  - Fetch number of all mission outcomes, incl. successful and failure events.
  - List booster_versions which have carried the maximum payload mass
  - Display the month names, landing_outcomes, booster versions, launch_site for 2015.
  - Sort the count of landing outcomes (2010-06-04 and 2017-03-20), in descending order.
- Queries and results are available in 2_SpaceXData_dataExplore_SQL.ipynb:
  - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/2_SpaceXData_dataExplore_SQL.ipynb

Build an Interactive Map with Folium

- Use interactive figures (Folium) to explore and better understand the available measures.
    - **USA Map of all SpaceX Launch Sites** - overview of locations of launches on the US map.
    - **USA Map: Proportion of Successful Launches by Location** - enhance the map by adding the launch outcomes for each site.
    - **Distance between Launch Site and Infrastructure** - to explore and analyze the proximities of launch sites.
- Code for interactive visualization using Folium is provided in 3_SpaceXData_CreateVisualDash.ipynb:
    - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/3_SpaceXData_CreateVisualDash.ipynb

Build a Dashboard with Plotly Dash

- Create interactive dashboards for exploring data relationships
  - Assessing payload mass (adjustable), launch site (adjustable), and booster versions
- Dashboards aim to assess:
  - Proportion of SpaceX launches at each Launch Location.
  - Identify launch site with the highest launch success rate.
  - Launch Success by payload mass and booster version.
- Code for interactive visualization using Plotly Dash is provided in 4_SpaceXData_InteractDash.py:
  - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/4_SpaceXData_InteractDash.py

Predictive Analysis (Classification)

- Data was further preprocessed
  - Predictor variables (w/ dummies) were prepared with outcome success (0/1)
  - Predictor variables were transformed (scaled).
  - Split into training and testing sets (20%, random state=2)
- Testing different predictive models, GridSearchCV (optimize parameters)
  - Logistic Regression
  - SVM
  - Decision tree
  - K-nearest neigh. (KNN)
- Code for testing each predictive model is provided in 5_SpaceXData_PredictSuccess.ipynb:
  - https://github.com/jhudev/SpaceX_BoosterLandingPredict/blob/3cc3aa514206e9c9c82fff8ccba3c7fb1ae7c132/5_SpaceXData_PredictSuccess.ipynb

Results

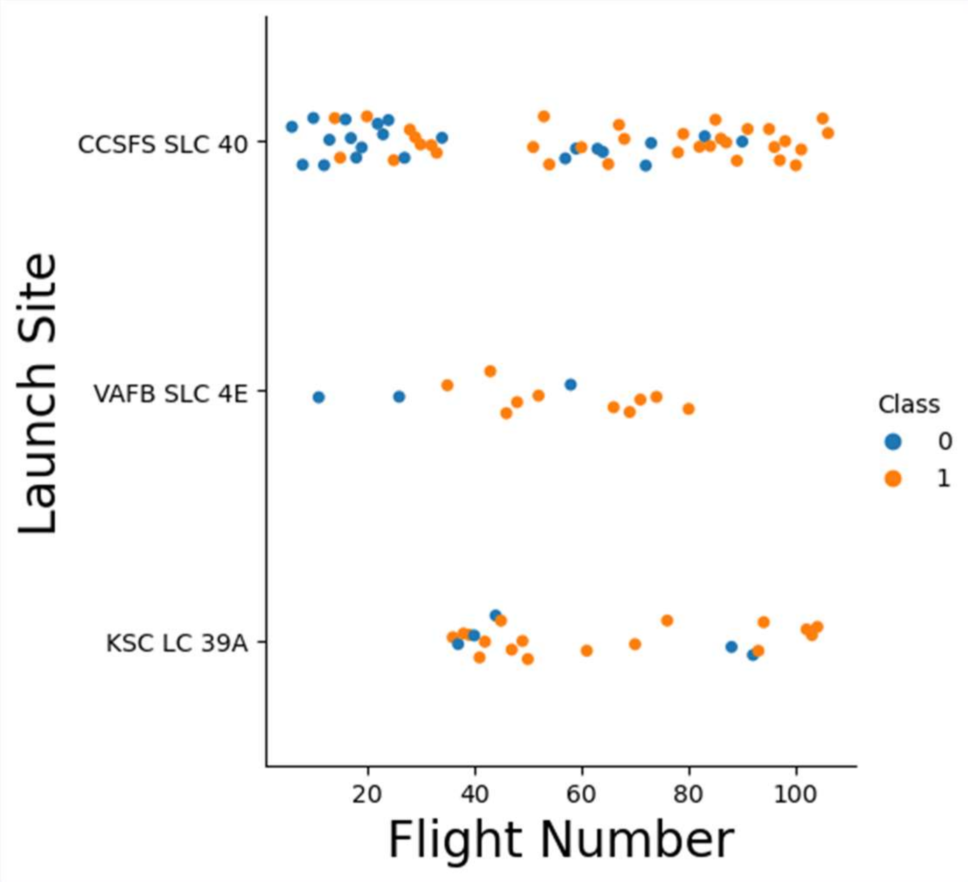
- Exploratory data analysis
  - Relationship between launch parameters (orbit, site, payload).
- Interactive analytics demo in screenshots
  - Folium – Geographic analysis of nearby structures
  - Plotly – Launch site, payload mass, and success
- Predictive analysis results
  - Confusion Matrix

Section 2

Insights drawn from EDA

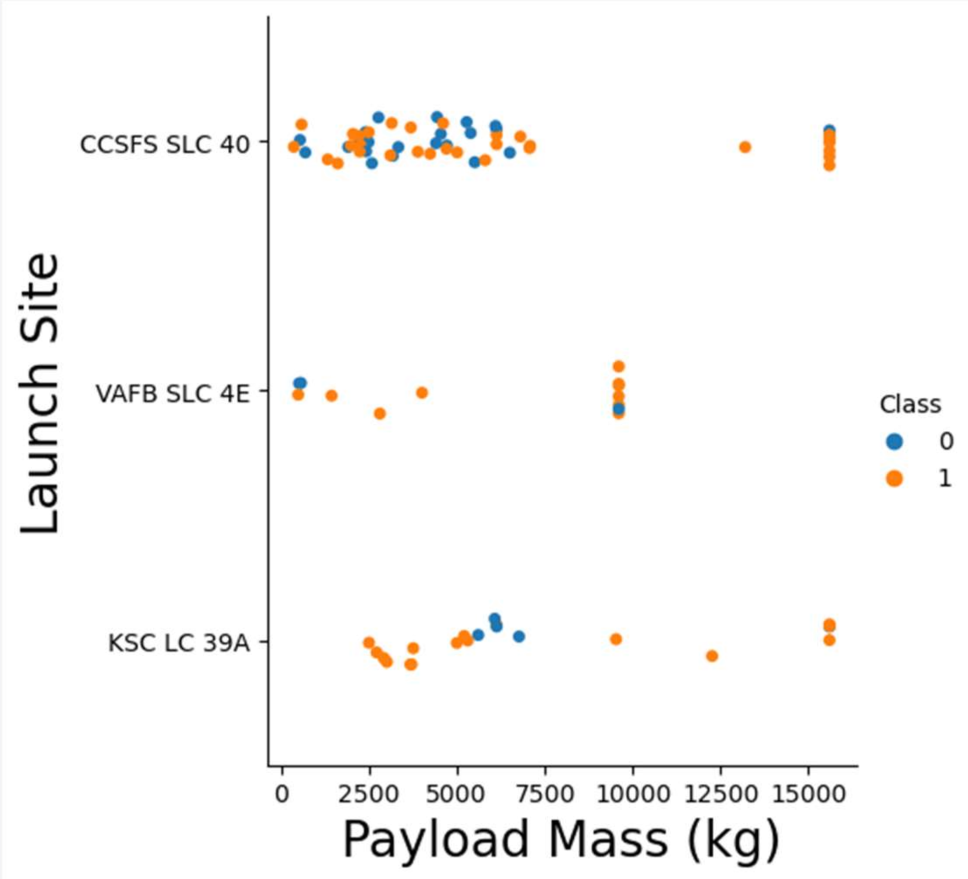
Flight Number vs. Launch Site

- Assess Launch Success (1/0) by launch site and flight number
  - In general, launch failures occurred most frequently in the first 50 flights (any sites), but have become much more reliable.
  - For each launch site, most launch failures occur in the first 10 launches per site.



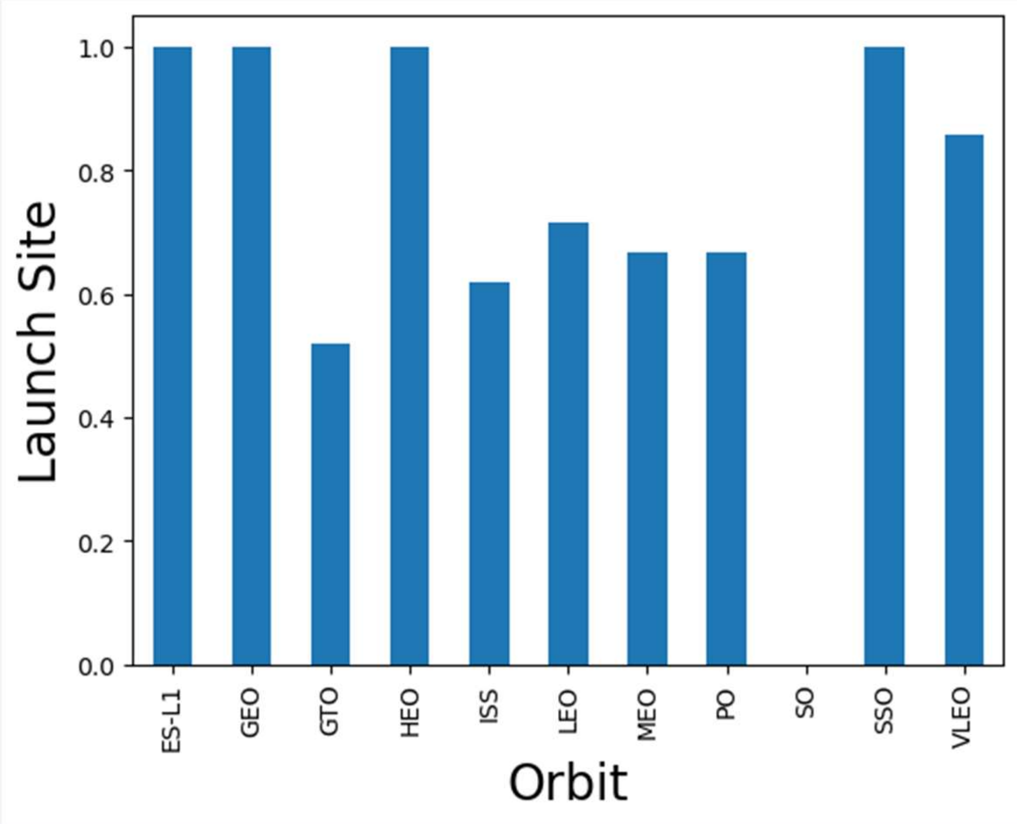
Payload vs. Launch Site

- Assess Launch Success (1/0) by payload mass and launch site
  - Different launch sites have different max payload masses. (ex. VAFB <1000kg)
  - There was no association between payload mass and launch success



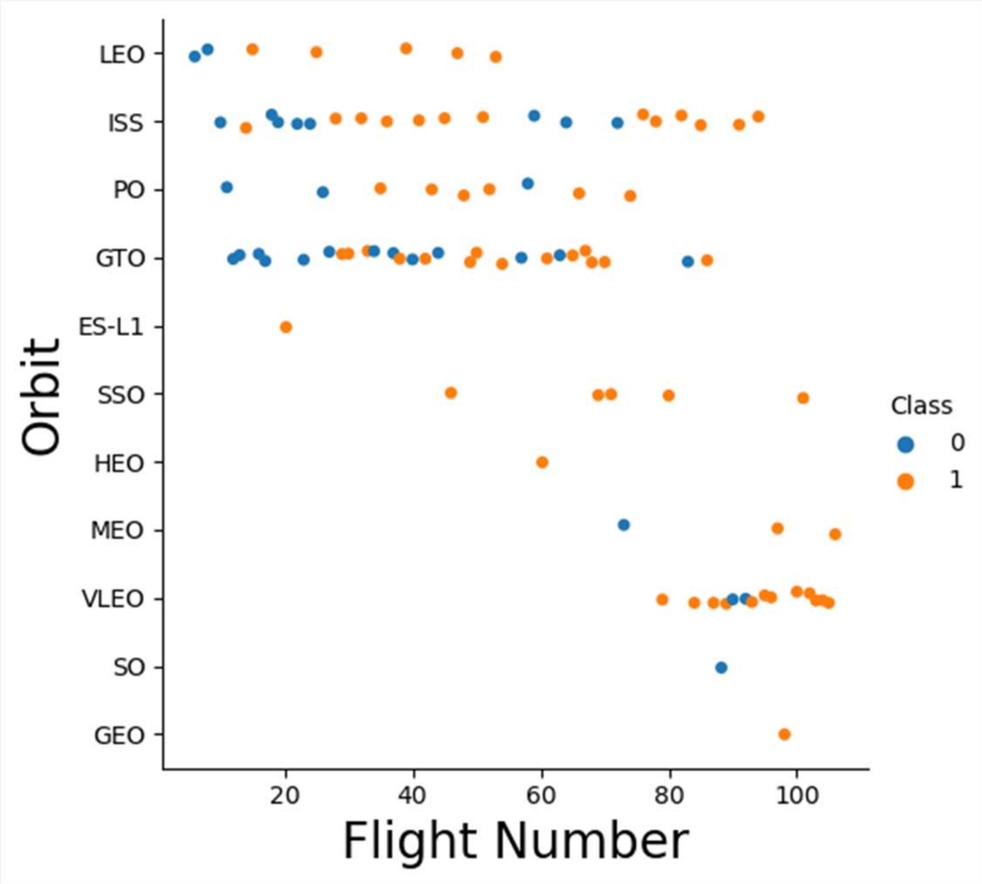
Success Rate vs. Orbit Type

- Compare success rate by orbit type
  - ES.L1, GEO, HEO, and SSO have the highest success rate.
  - GTO, ISS, and SO have the lowest success rate.
  - However, this is likely biased by the order of the flight number.



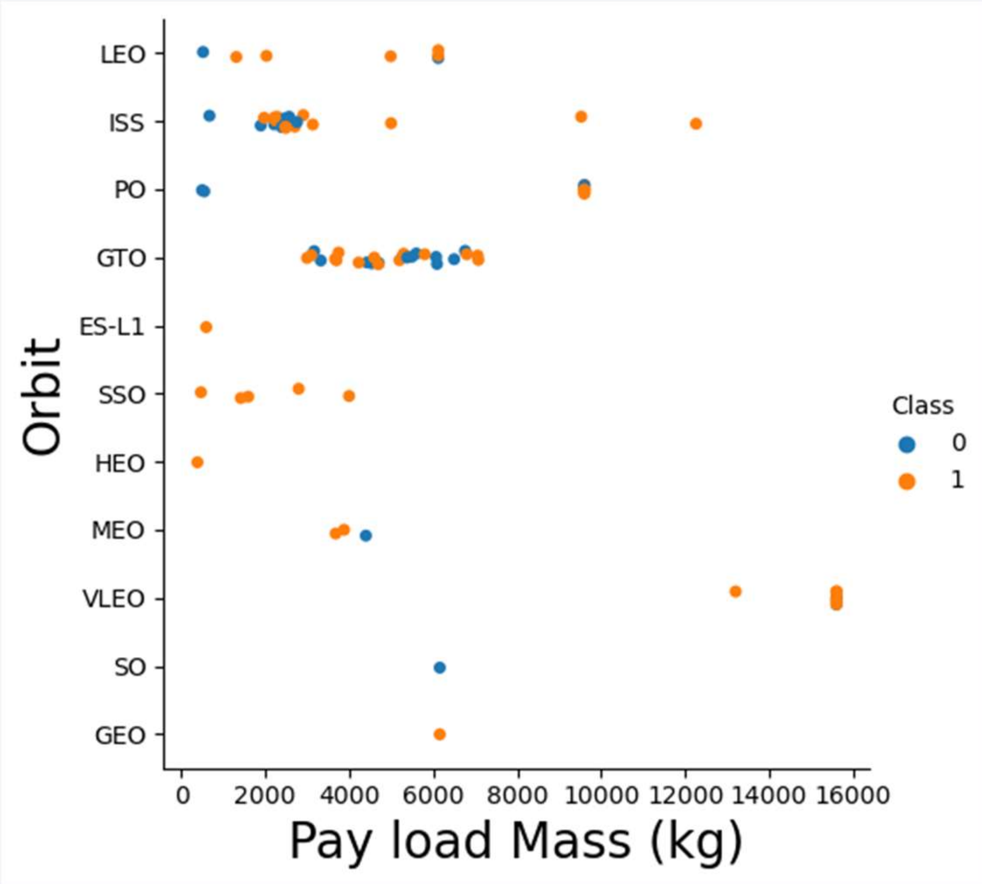
Flight Number vs. Orbit Type

- Assess Launch Success (1/0) by orbit and flight number¶
    - In the past, flights were more LEO, GTO, ISS, and PO.
    - However, more recent flights have been mostly VLEO.
    - For many, success improves with flights. However, GTO seems to show less obvious improvements and still have reoccurring failures.



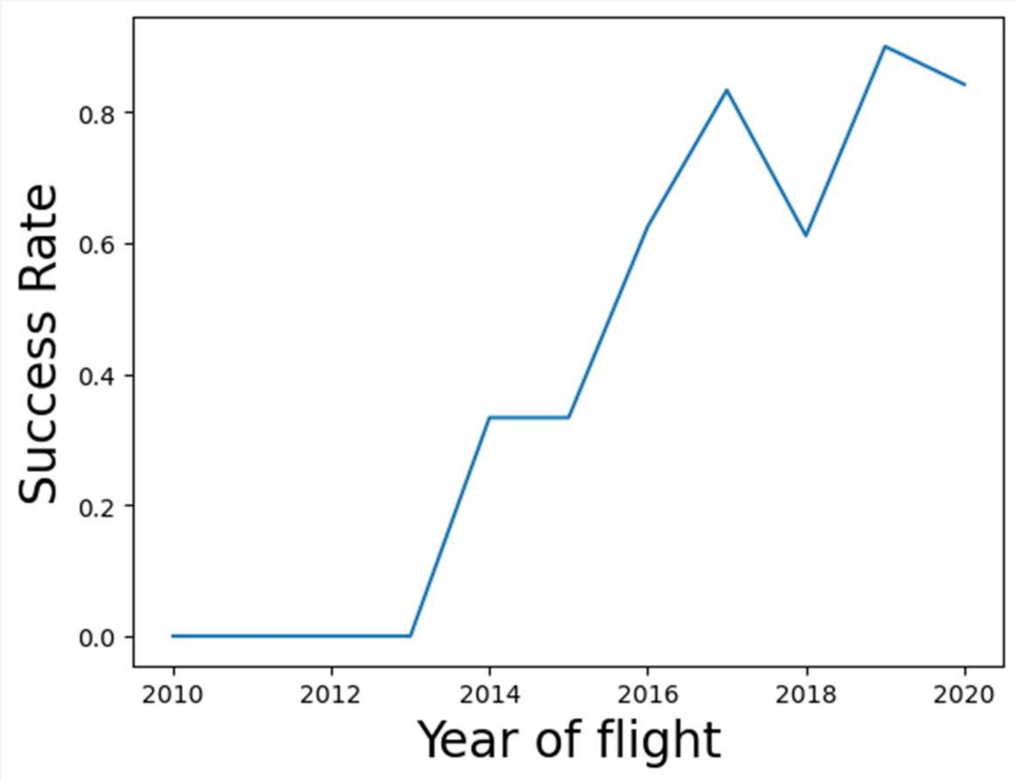
Payload vs. Orbit Type

- Assess Launch Success (1/0) by payload mass and orbit
  - For LEO, PO, and ISS, success occurs more with high payload mass, although this could be related to flight number.
  - Payload mass for GTO is less predictable for success.



Launch Success Yearly Trend

- Success rate over time
  - Success rate of flights have been growing significantly overtime.



All Launch Site Names

- Find the names of all launch sites (unique values)
  - Four different sites
- *Using DISTINCT*
- *%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL*

```
[20]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[20]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site 'CCA'

- Find first five records of 'CCA' launch sites (see below)
- Prefix “CCA%”, Limited to 5
- %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5

```
[27]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5
 * sqlite:///my_data1.db
Done.
[27]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass Launched for NASA (CRS)

- Calculate the total payload mass carried by boosters launched for NASA (CRS)
    - Total of 45596 Kg launched by NASA (CRS)
- Sum of payload mass for Customer="NASA (CRS)"
- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer="NASA (CRS)"

```
[33]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer="NASA (CRS)"
       * sqlite:///my_data1.db
      Done.
[33]: SUM(PAYLOAD_MASS__KG_)
                     45596.0
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - Average payload mass is 2534.6 Kg for F9 v1.1
- Find pattern “F9 v1.1%” in Booster_Version, Use AVG
- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1%"

[4] Calculate the average payload mass carried by booster version F9 v1.1

```
[58]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1%"

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

[58]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - 01/08/2018
- Outcome of “Success (ground pad)”, using MIN on Date
- %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome="Success (ground pad)"

[5] Fetch date of the first succesful landing outcome in ground pad

```
[34]: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome="Success (ground pad)"

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

[34]:

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the boosters that have success in drone ship and have payload mass between 4000 and 6000 (See below)
- Outcome of "Success (drone ship)“, between 4000-6000
- %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome="Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

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

[35]:

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

Total Number of Successful and Failure Mission Outcomes

- Fetch number of all mission outcomes, including successful and failure events.
- Count, Group by Mission Outcome
- %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome

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

[56]:

Mission_Outcome	COUNT(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List booster_versions which have carried the maximum payload mass
- Fetch max payload, then filter WHERE
- %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

[37]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- Display the month names, landing_outcomes, booster versions, launch_site for 2015.
- Fetch data parameters, Order by month
- %sql SELECT substr(Date,4,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)="2015" ORDER BY Month

Month	Landing_Outcome	Booster_Version	Launch_Site
02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- Sort the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- Use group by and order by
- %sql SELECT Landing_Outcome,COUNT(Landing_Outcome) FROM SPACEXTBL WHERE Date BETWEEN '04/06/2010' AND '20/03/2017' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

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

[55]:

Landing_Outcome	COUNT(Landing_Outcome)
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

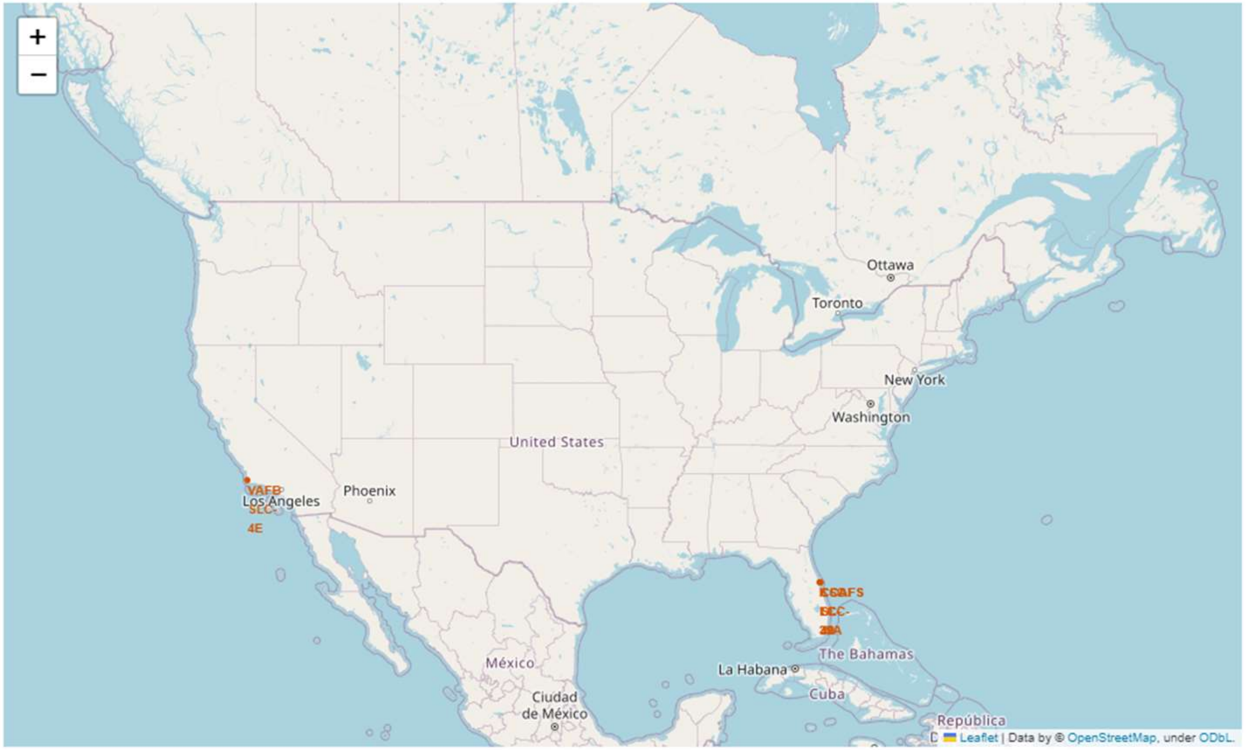
USA Map of all SpaceX Launch Sites

- Most launches are on the east-coast of Florida with only VAFB-SLC4E on the coast of California.

In [59]:

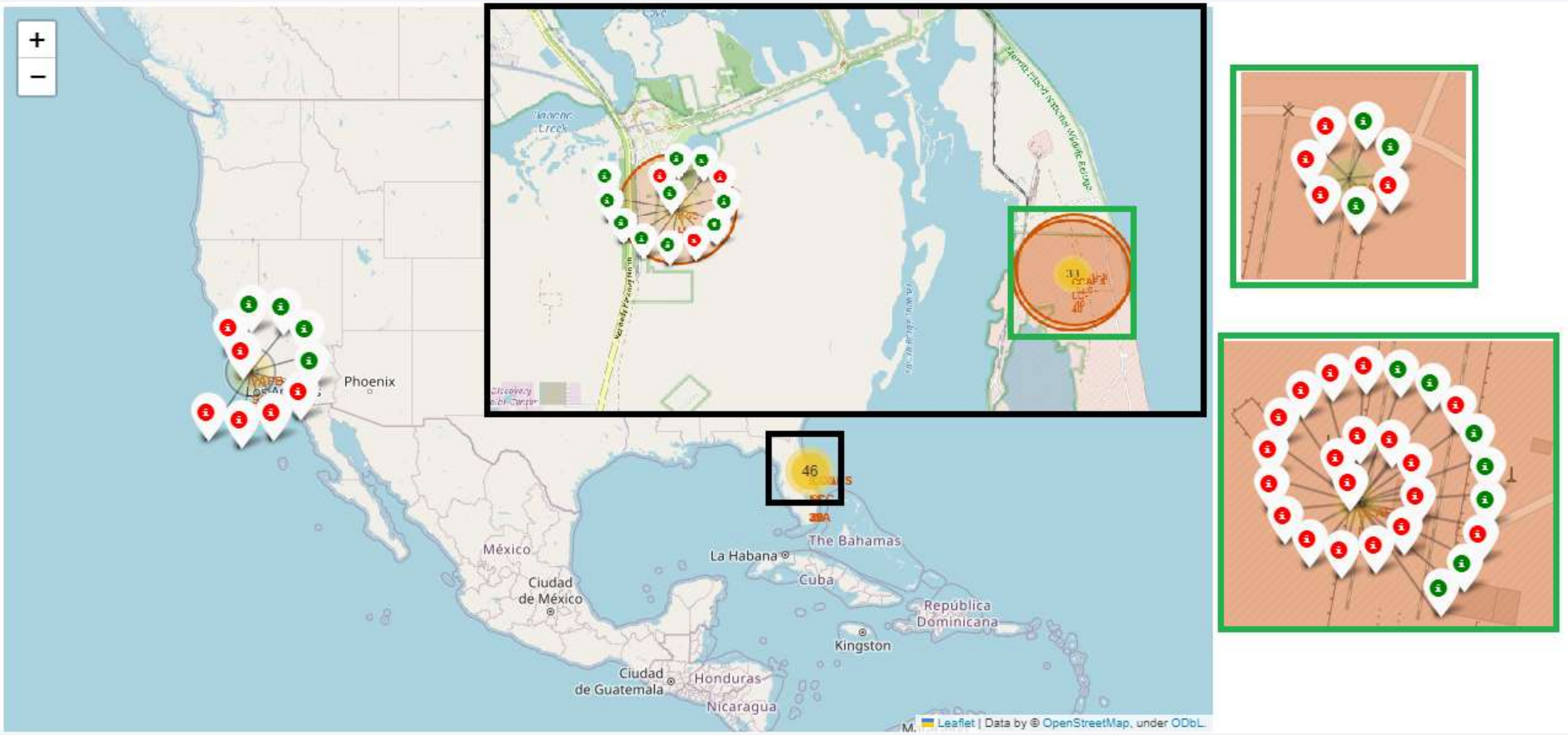
```
site_map = folium.Map(location=[29.5597, -95.0831], zoom_start=4.2) #<create map object
for name, coor in [[x[0],[x[1], x[2]]] for x in launch_coor.values.tolist()]:
    folium.Circle(coor, radius=1000, color='#d35400', fill=True).add_to(site_map)
    folium.Marker(coor, icon=DivIcon(icon_size=(20,20),icon_anchor=(0,0), html=f'<div style="font-size: 22; color:#d35400;">
site_map
```

Out[59]:

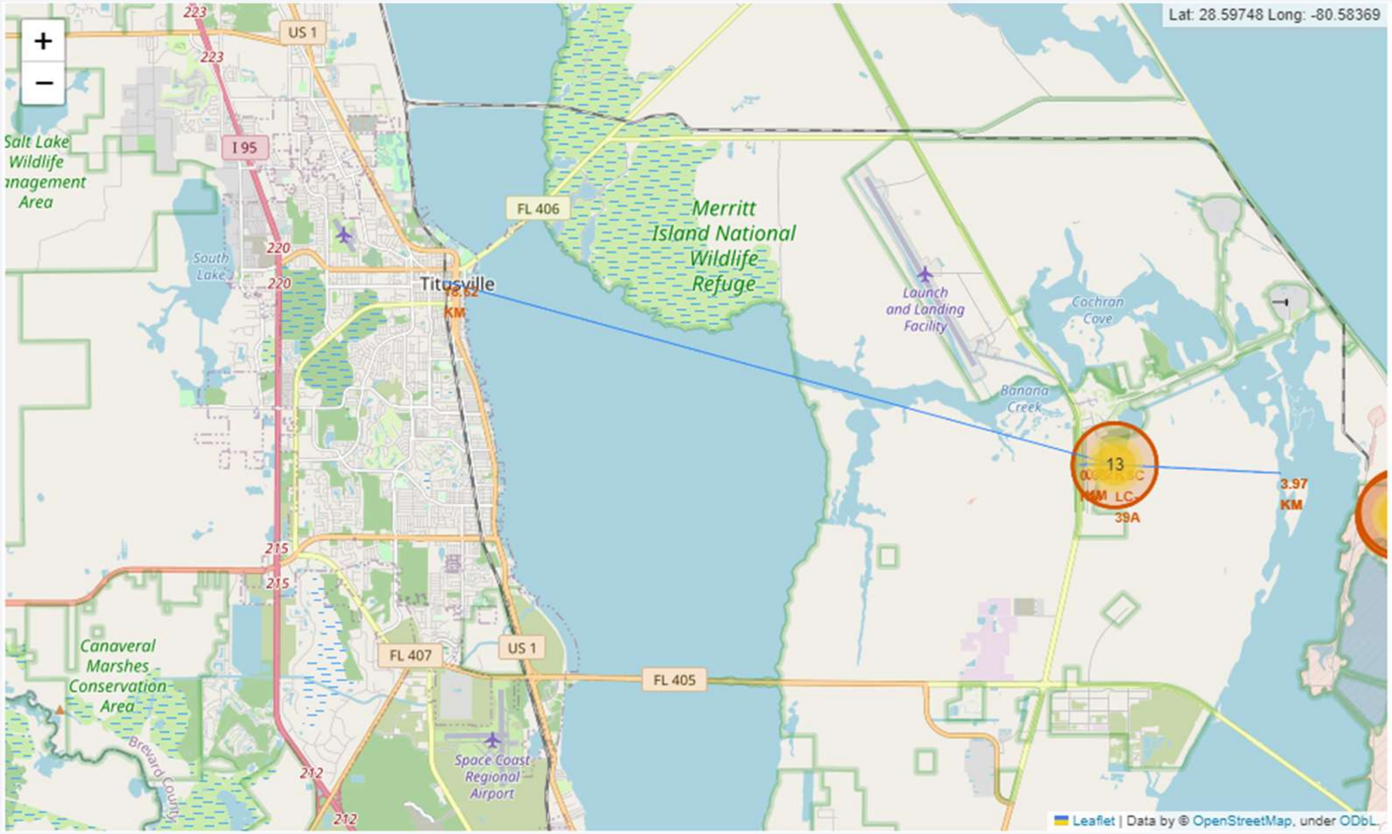


USA Map: Proportion of Successful Launches by Location

- Most successful launches came from Florida (KSC LC-39A)



Distance between Launch Site and Infrastructure



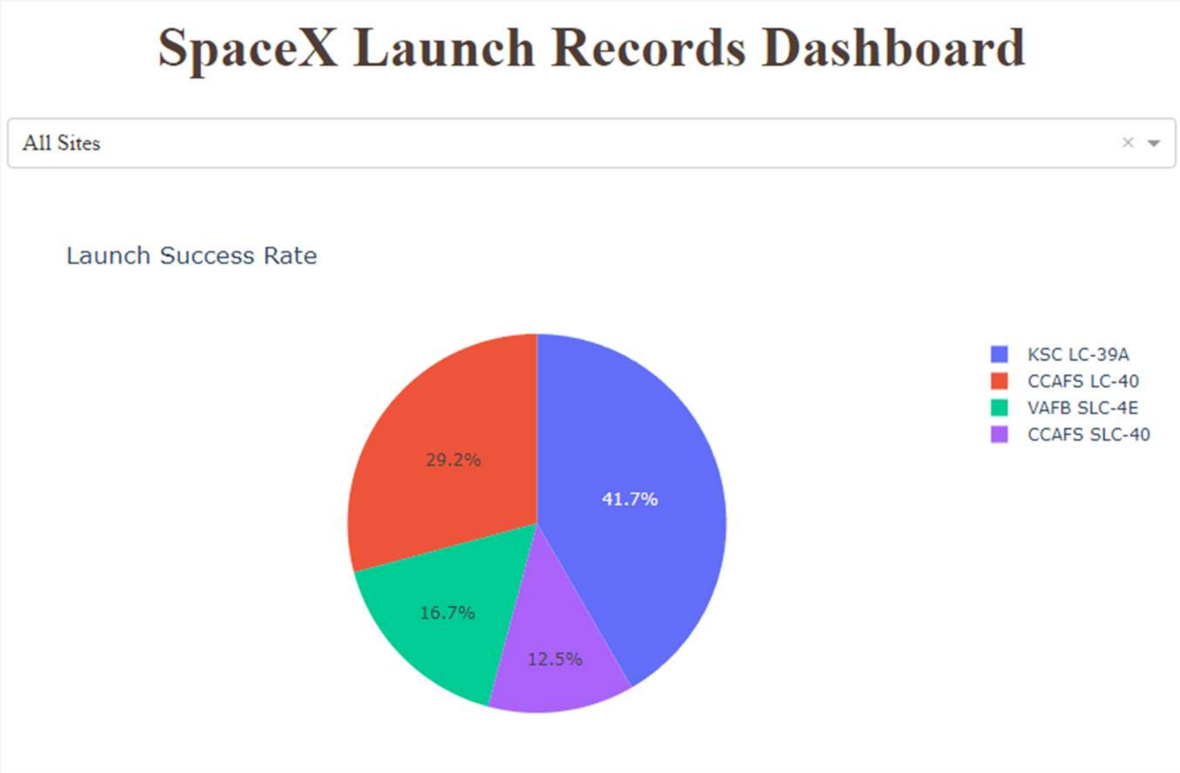
As expected, space rocket launch sites are chosen to be far from populated cities while being close to coast, railway, and highway for easier transportation of material.

Section 4

Build a Dashboard with Plotly Dash

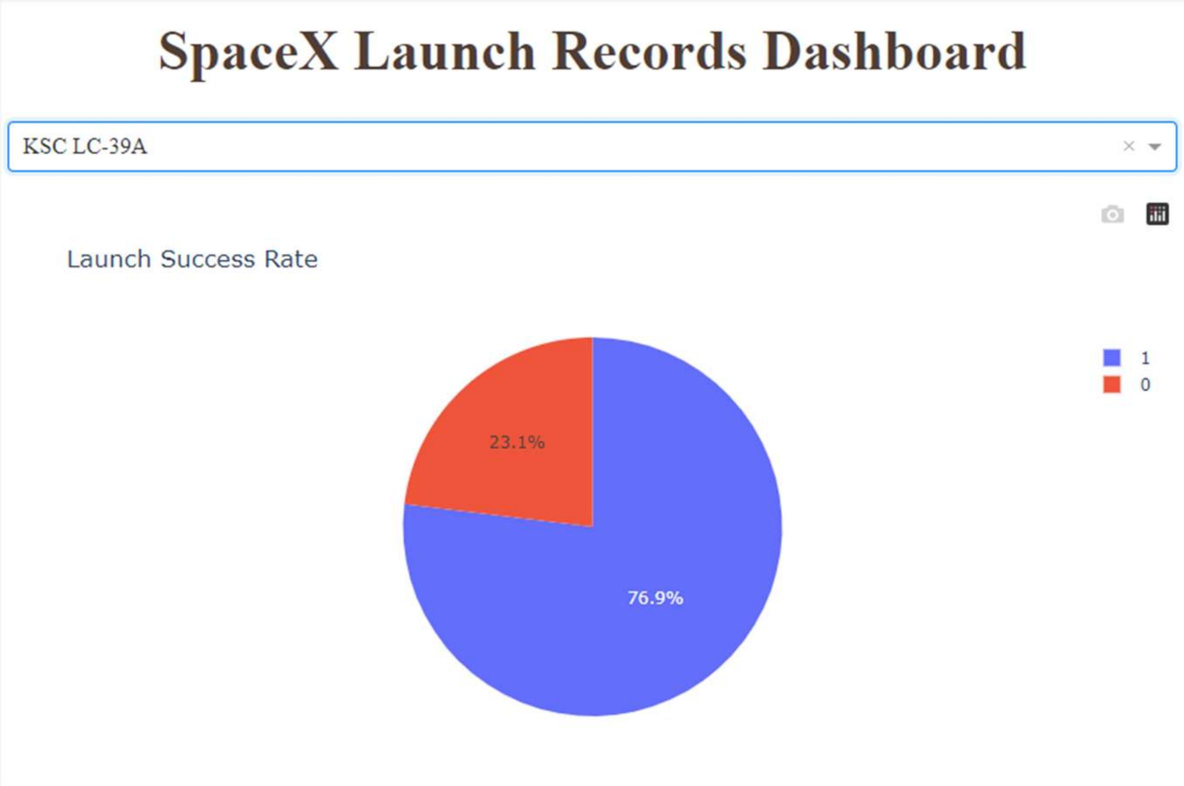
Proportion of Launches at each Site.

- Majority of launches occurred at KSC LC-39A and CCAFS LC-40.



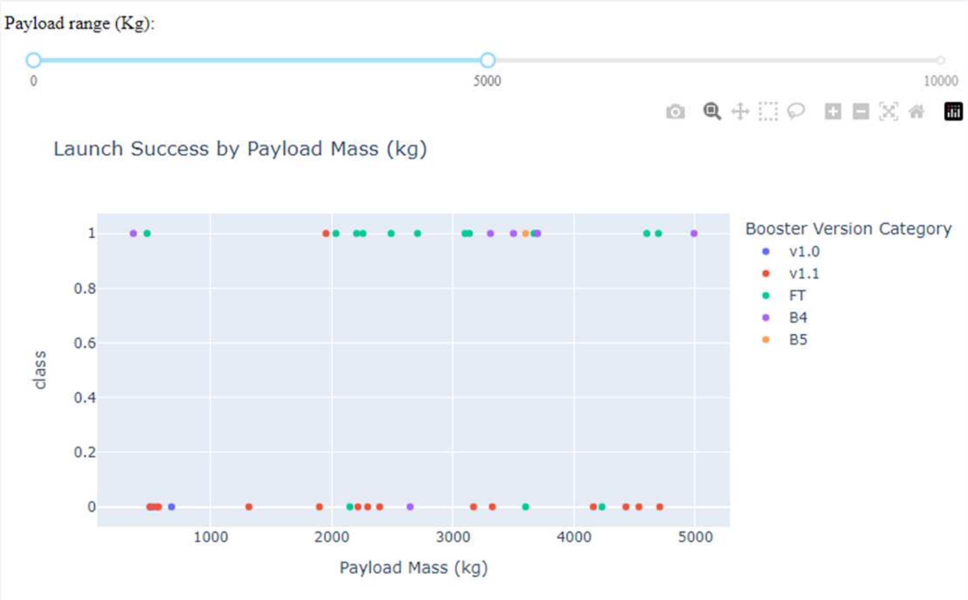
Launch Site with the highest launch success rate

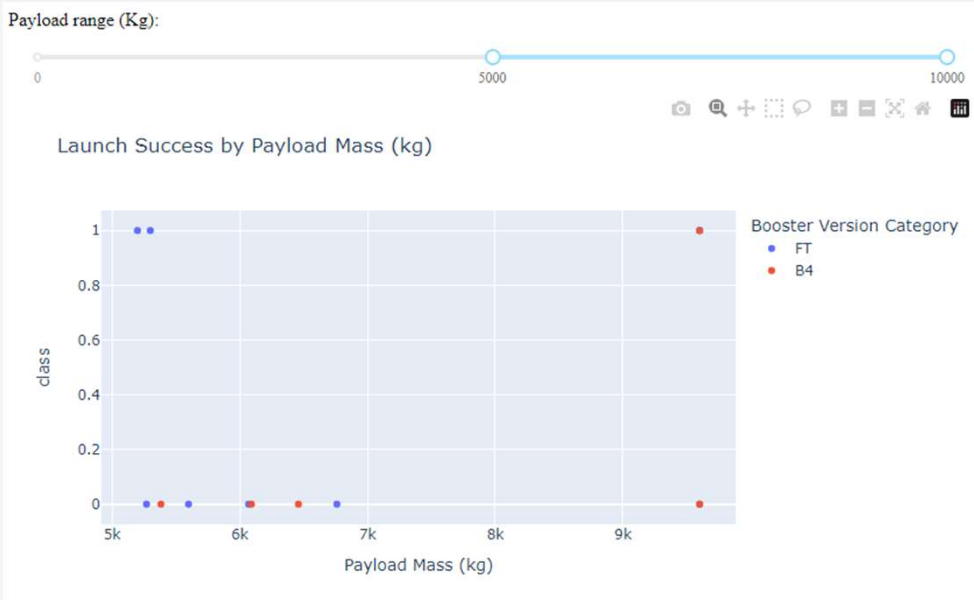
- The site KSC LC-39A had the highest launch success rate 76.9%.



Launch Success by payload mass and booster ver.

- Those with large payloads (>5000kg) have to use FT or B4 booster versions. These tend to have historically failed more often than lower payload rockets.



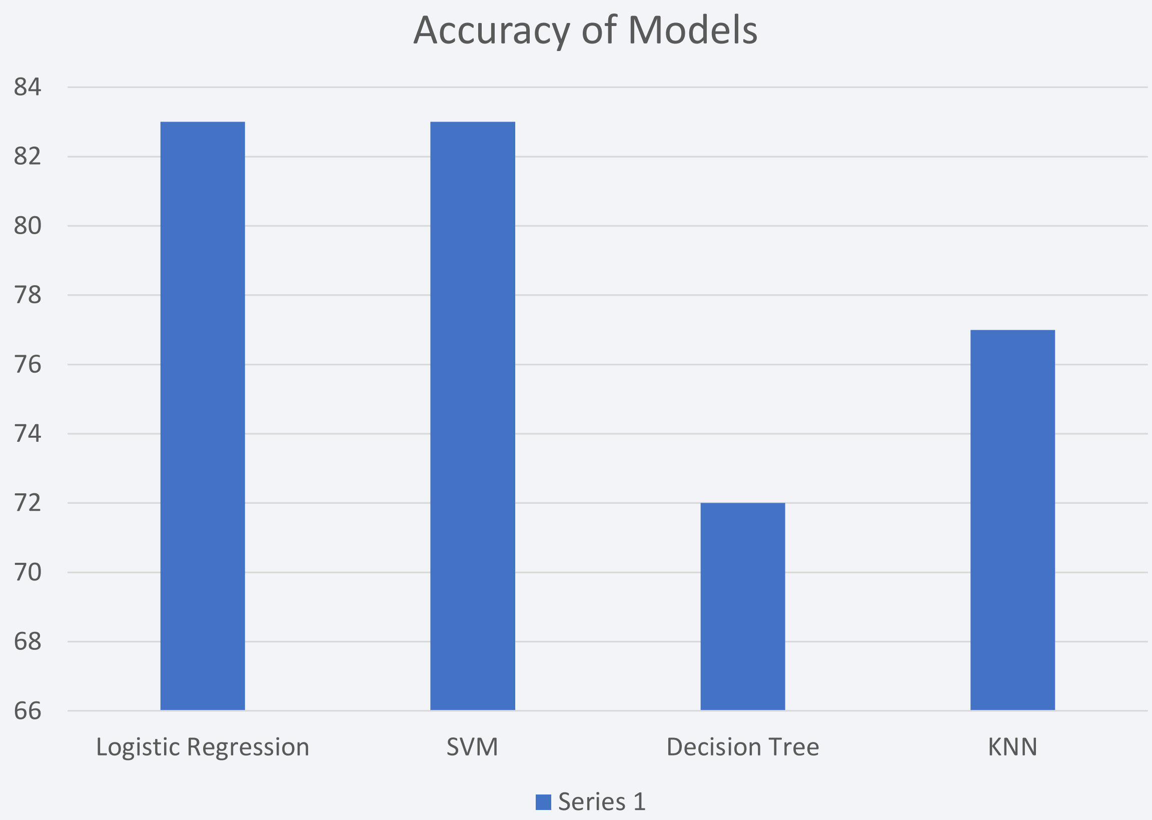


Section 5

Predictive Analysis (Classification)

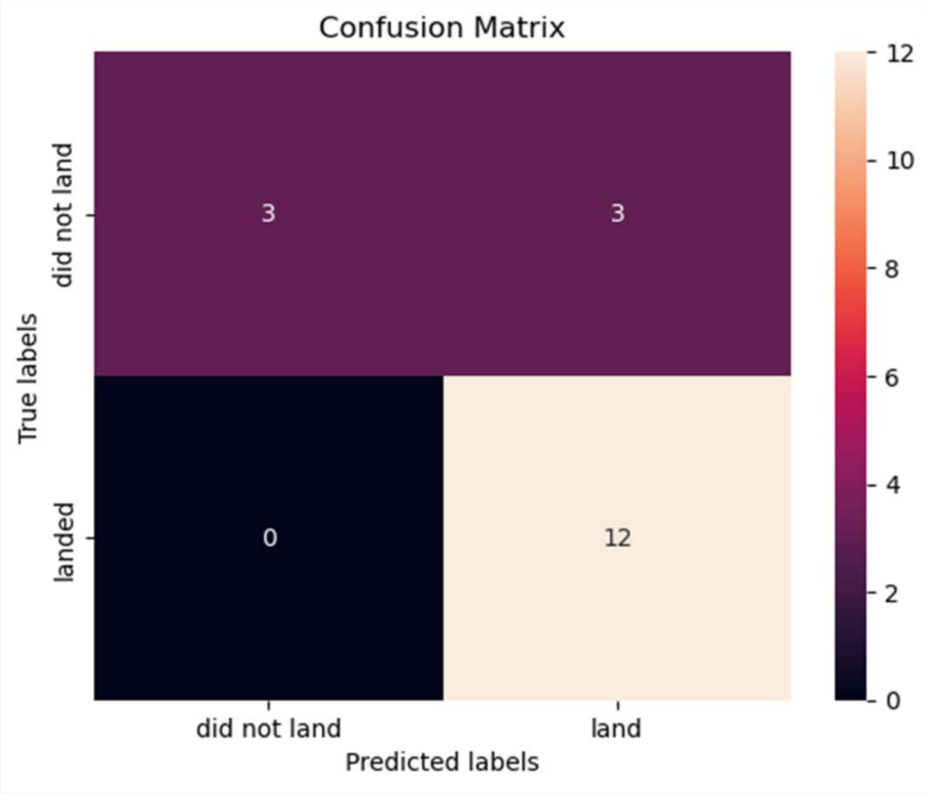
Classification Accuracy

- Logistic and SVM models showed the highest accuracy of (83%) for predicting the success of launch (booster landing).



Confusion Matrix

- Logistic and SVM models showed the highest accuracy of (83%) for predicting the success of launch (booster landing).
- Difficulty was dealing with false positives (those that were predicted to land, but did not land).



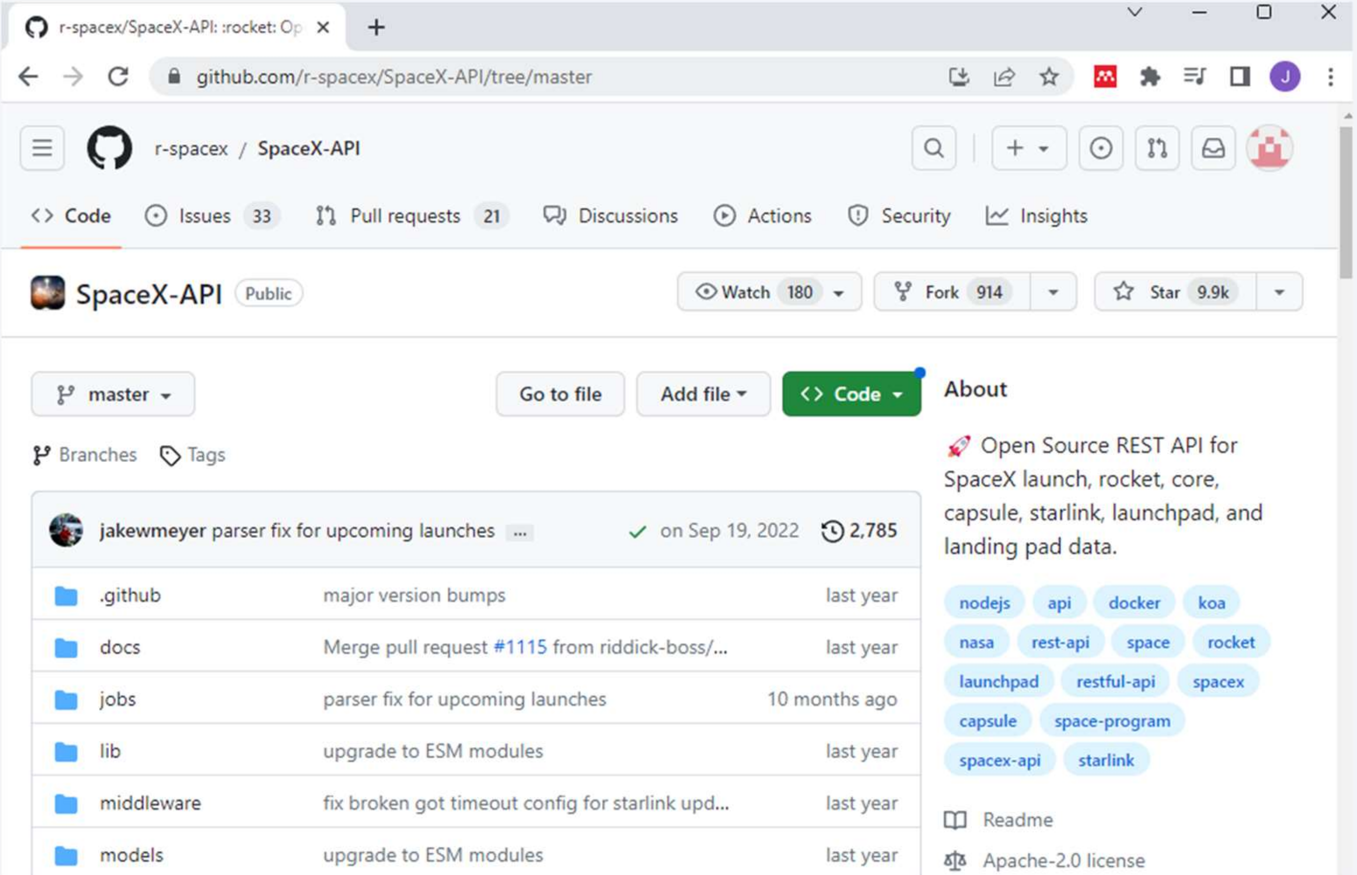
Conclusions

- In general, launch failures occurred most frequently in the first 50 flights (any sites), but have become much more reliable.
- The site KSC LC-39A had the highest launch success rate 76.9%.
- Large payloads (>5000kg) have to use FT or B4 booster versions, which have historically failed more often.
- Logistic Regression Models and SVM was able to predict Launch Success (booster landing) with an accuracy of 83%.
- However, one has to be cautious of false positives (not landing when predicted).

Appendix

- Data source: SpaceX-API
- Data source: Wikipedia – SpaceX Rockets
- Additional success rate measures by launch site.
- Additional confusion matrixes

Data source: SpaceX-API



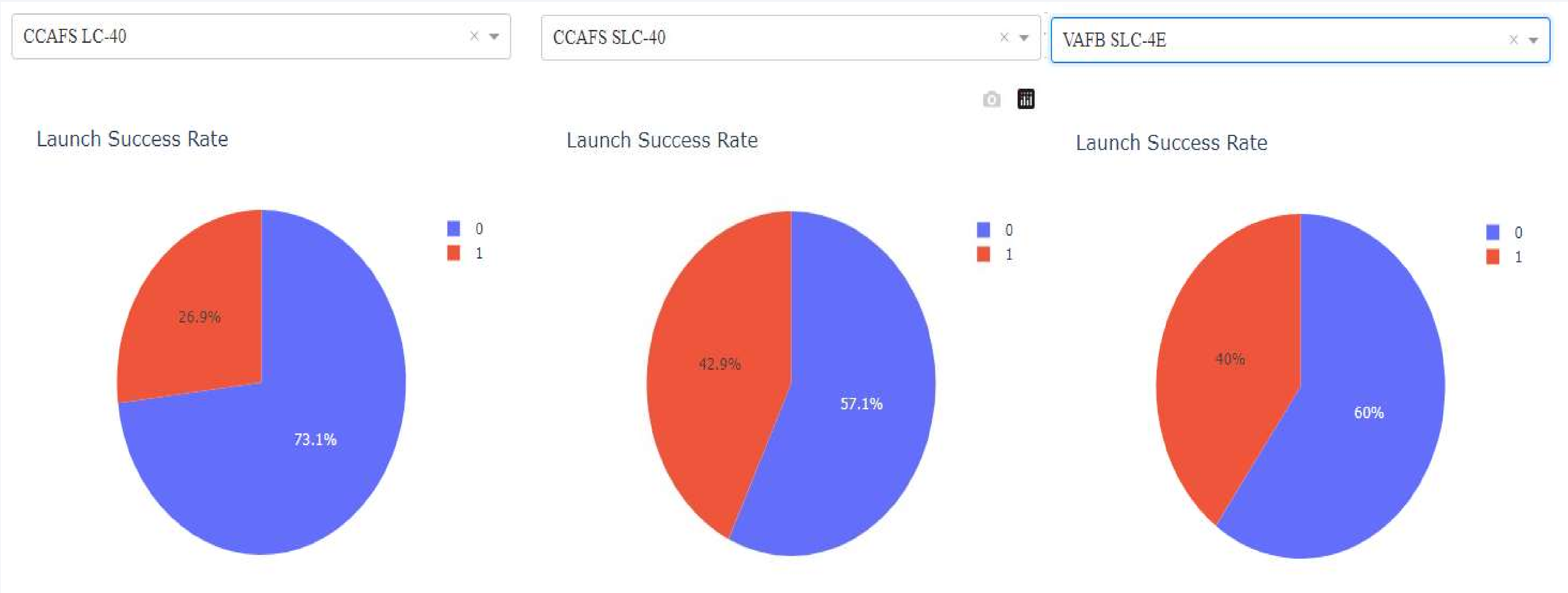
Data source: Wikipedia – SpaceX Rockets

Past launches

2010 to 2013

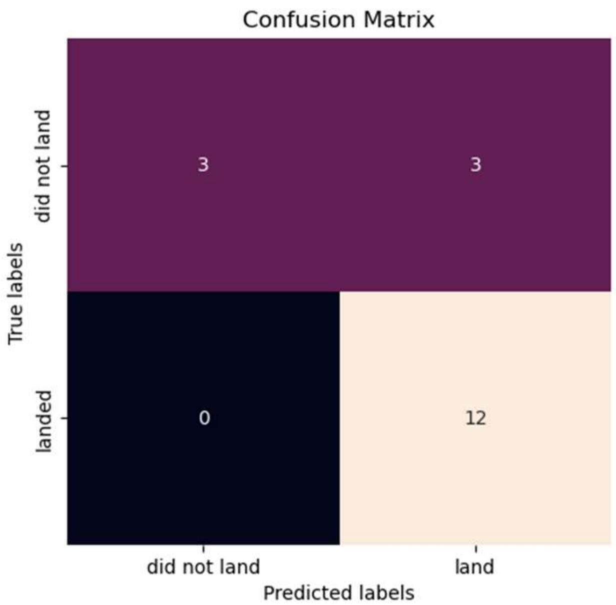
[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
				SpaceX CRS-1[22]	4,700 kg	LEO			

Additional success rate measures by launch site.

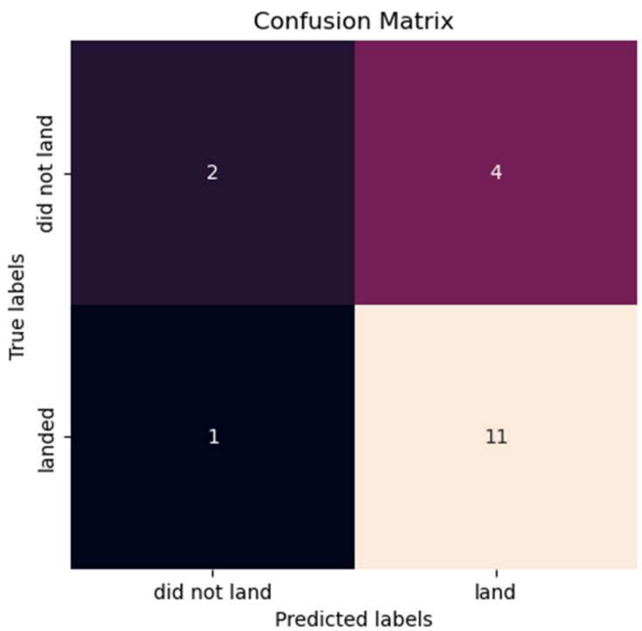


Additional confusion matrixes

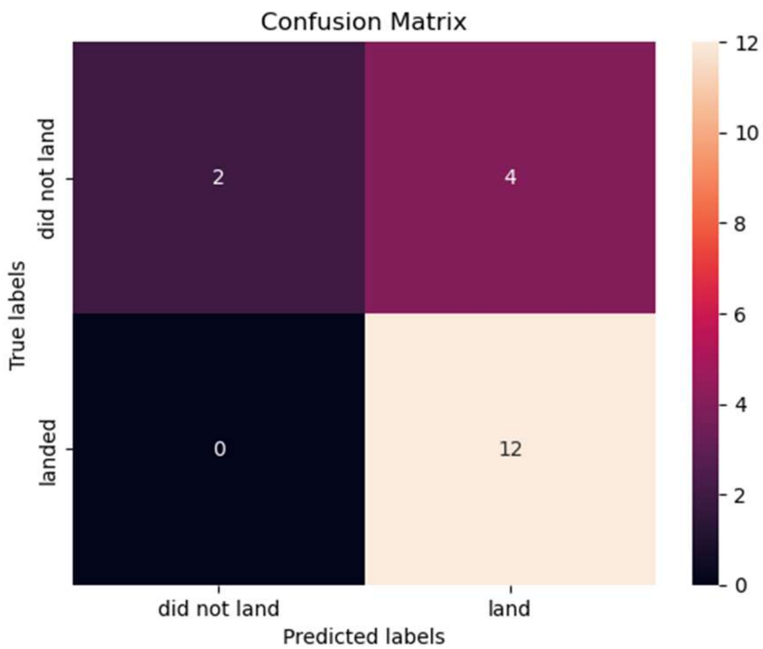
Logistic/SVM



Decision Tree



KNN



Thank you!