IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

James Chenery
2022-11-11

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folum
  - Building a dashboard with Plotly Dash
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context

  Our startup aerospace company is competing with SpaceX, a manufacturer and operator of commercial space launch vehicles, and we are currently bidding against them for a space launch contract.

  SpaceX advertises Falcon 9 rocket at a cost of 62 million dollars while other providers cost upward of 165 million dollars each.

- Problems you want to find answers

  Much of the SpaceX's savings are due to reusing the first stage.

  We need to predict the probability of SpaceX Falcon 9 first stages landing successfully (and thus being reuseable) in order to make more informed bids against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

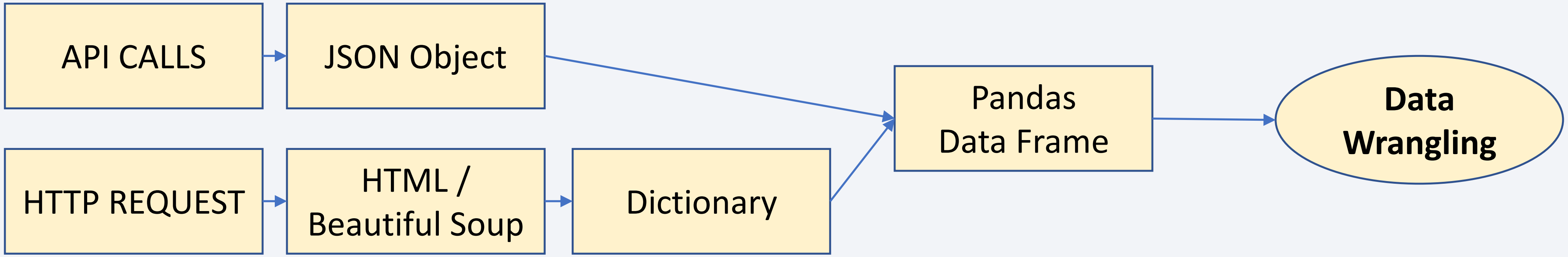
Executive Summary

- Data collection methodology:
  - JSON data was obtained via GET requests to the SpaceX API
  - HTML data was obtained from Wikipedia via Beautiful Soup
  - JSON/HTML data -> Normalized Pandas data frame
- Performed data wrangling
  - Removed non-Falcon 9 launch records, removed unneeded columns, handled null values by replacing 5 missing Pay Load Mass values with the mean average. Used One Hot Encoding to prepare data for machine learning
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - LR, KNN, SVM and DT models were compared for best classifier

Data Collection

SpaceX data was collected from two sources:

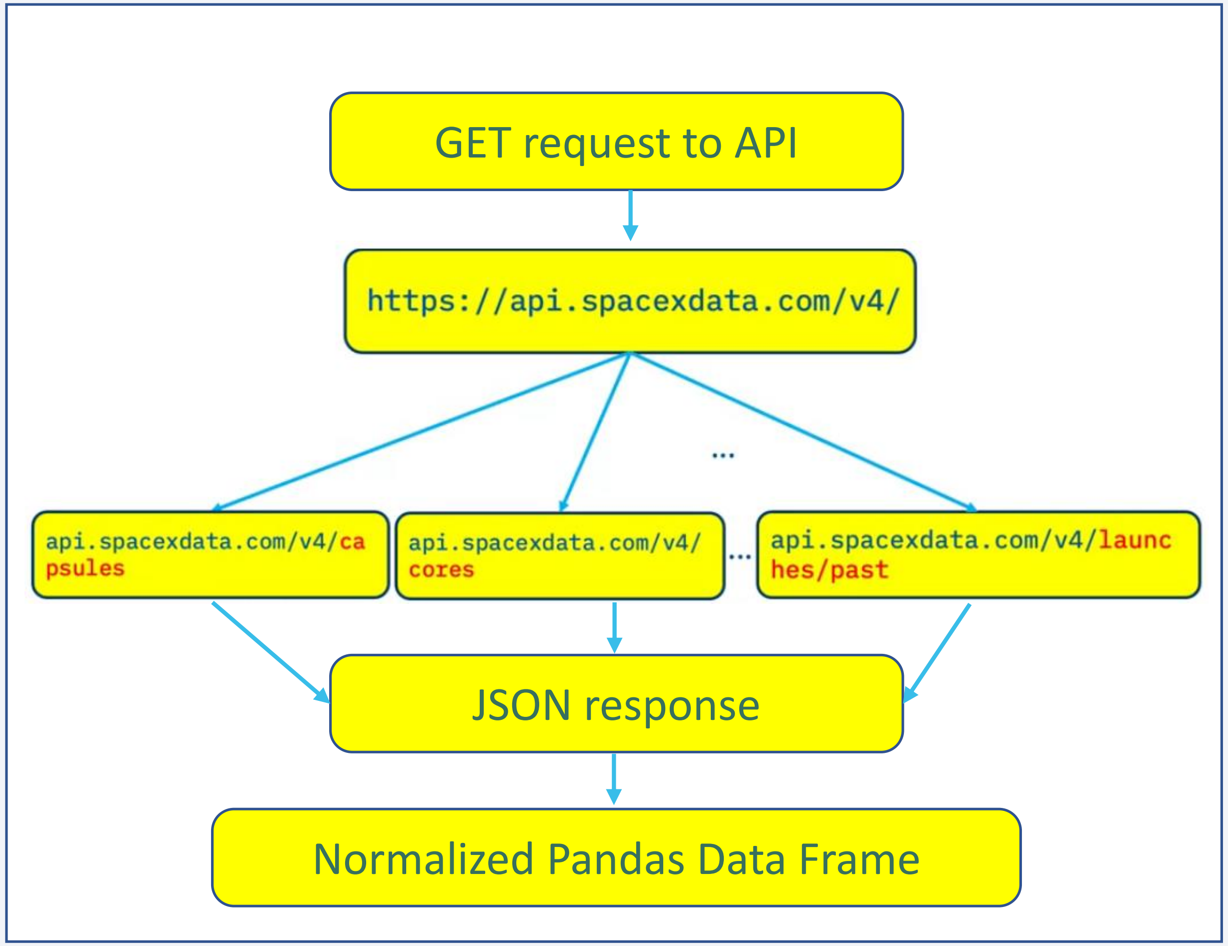
1. Spacexdata.com who provide SpaceX data via calls to the r/SpaceX API
   - https://docs.spacexdata.com/
2. Wikipedia.com's SpaceX Falcon 9 launches list page, which contains a table of launch data, which was scraped using Beautiful Soup
   - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Data Collection – SpaceX API

- API documentation: https://github.com/r-spacex/SpaceX-API/tree/master/docs#rspacex-api-docs
- REST call (GET) sent to four API endpoints:
    - Root, rocket, payload, launchpad, cores
    - Rocket, payload, launchpad and cores are supplementary calls used to append data to root response by using ID values (e.g. rocket) as key indexes via functions
- Each JSON response is normalized into a Pandas data frame, ready for wrangling in Python

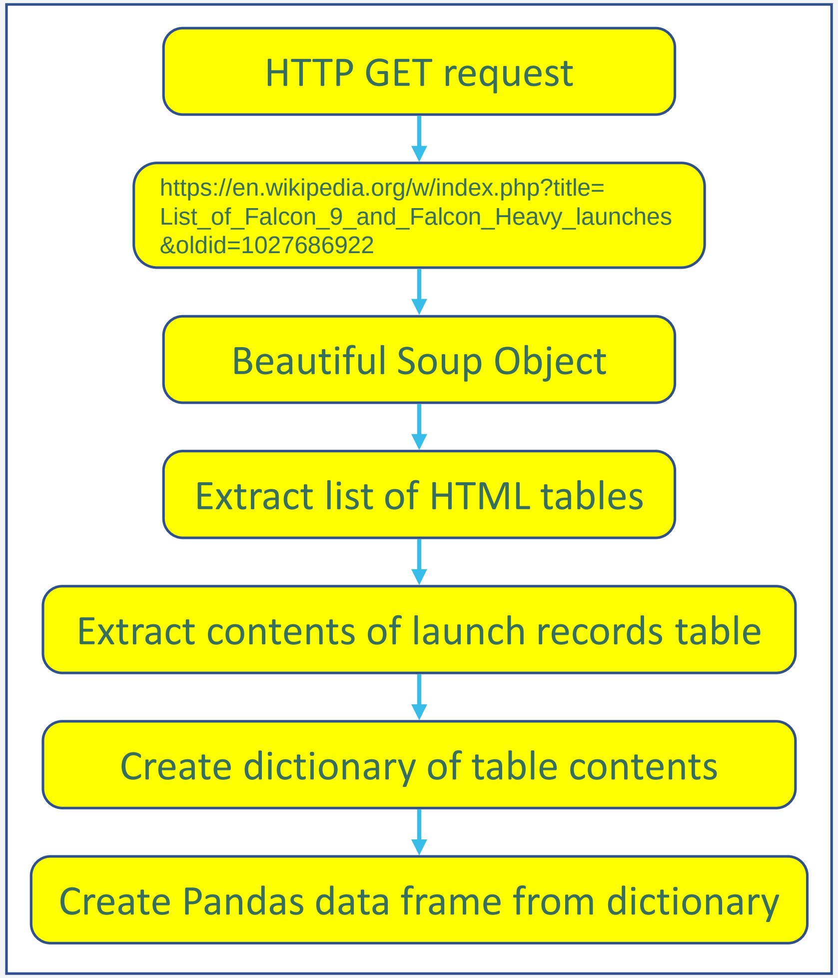
Python Notebook:
https://github.com/james-chenery/IBM-Data-Science/blob/22b68ad89cc1a5d5dec333927d1fecd790d35dd5/API-data-collection.ipynb



Data Collection - Scraping

Beautiful Soup documentation: https://beautiful-soup-4.readthedocs.io/en/latest/

- HTTP GET method requested the Falcon9 Launch Wikipedia HTML page, as an HTTP response
- Created a Beautiful Soup object from the HTML response
- Created a list of all HTML tables from the Beautiful Soup object by using the *find_all* function with element type *'table'*
- Identified and extracted the launch records table
- Used the *find_all* function with element type '*th*' to create a list of column values
- Initialized dictionary with column values, assigned an empty list to each dictionary value and then populated those lists from the table with *find_all*
- Created a Pandas data frame from the dictionary



Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/a95323eb315de1ffe238ea4dc9a7ec18a32b8354/Webscraping.ipynb

Data Wrangling

Steps of Data Wrangling

1. Check for null values
2. For each launch site, calculate the number of launches
3. For each orbit, calculate the number of launches
4. For each orbit, outcome and calculate the number of missions
5. Add a binary Class column for success/failure of landing outcomes
6. Remove null values

Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/106fd6ac33aa17fbbc8d3f337f6f0a1b3a71675f/wrangling.ipynb

EDA with Data Visualization

Summary of Data Visualizations

1. **Launch Site vs. Flight Number Scatter Plot:**
   To visualize the relationship between Flight Number and Launch Site
2. **Launch Site vs. Pay Load Scatter Plot:**
   To visualize the relationship between Payload and Launch Site
3. **Success vs. Orb Bar Chart:**
   To visualize the relationship between success rate of each orbit type
4. **Orbit vs. Flight Number Scatter Plot:**
   To visualize the relationship between FlightNumber and Orbit type
5. **Orbit vs. Pay Load Scatter Plot:**
   To visualize the relationship between Payload and Orbit type
6. **Success vs. Year Line Chart:**
   To visualize the launch success yearly trend

Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/366f6ed7d0dee8f09c18470ba37f61e2428823e6/EDA-DATAVIZ.ipynb

EDA with SQL

Queries performed:

1. Displayed the names of the unique launch sites in the space mission
2. Displayed 5 records where launch sites began with the string 'CCA'
3. Displayed the total payload mass carried by boosters launched by NASA (CRS)
4. Displayed the average payload mass carried by booster version F9 v1.1
5. Listed the date when the first successful landing outcome in ground pad was achieved
6. Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Listed the total number of successful and failure mission outcomes
8. Listed the names of the booster_versions, which have carried the maximum payload mass, with a subquery
9. Selected month names, failure landing_outcomes in drone ship ,booster versions and launch_site for months in year 2015
10. Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017, in descending order

Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/366f6ed7d0dee8f09c18470ba37f61e2428823e6/EDA-SQL.ipynb

Build an Interactive Map with Folium

A folium map was created with:

- Circle markers and labels to indicate launch sites
- Lines to indicate proximity to the nearest coast, railway line station/terminus, highway exit, city-designated residential area and civil airport from Cape Canaveral
- Distance values were appended to the proximity lines
- Markers were added to each launch site to indicate each launch from that site and whether or not it was successful.

The non-interactive versions in this report were supplemented with:

- Magnified views
- Circles and arrows were added to indicate launch region each magnified view related to

Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/31d1dce129b0959548c8ebb58ed251129d0749aa/EDA-Launch_Site.ipynb

Build a Dashboard with Plotly Dash

The dashboard is comprised of:

- Pie chart showing the total successful launches vs. launch site
    - Interactive drop down to select all sites or specific sites
- Scatter graph showing launch outcomes vs. booster version
    - Interactive slider to select a payload mass (kg) range

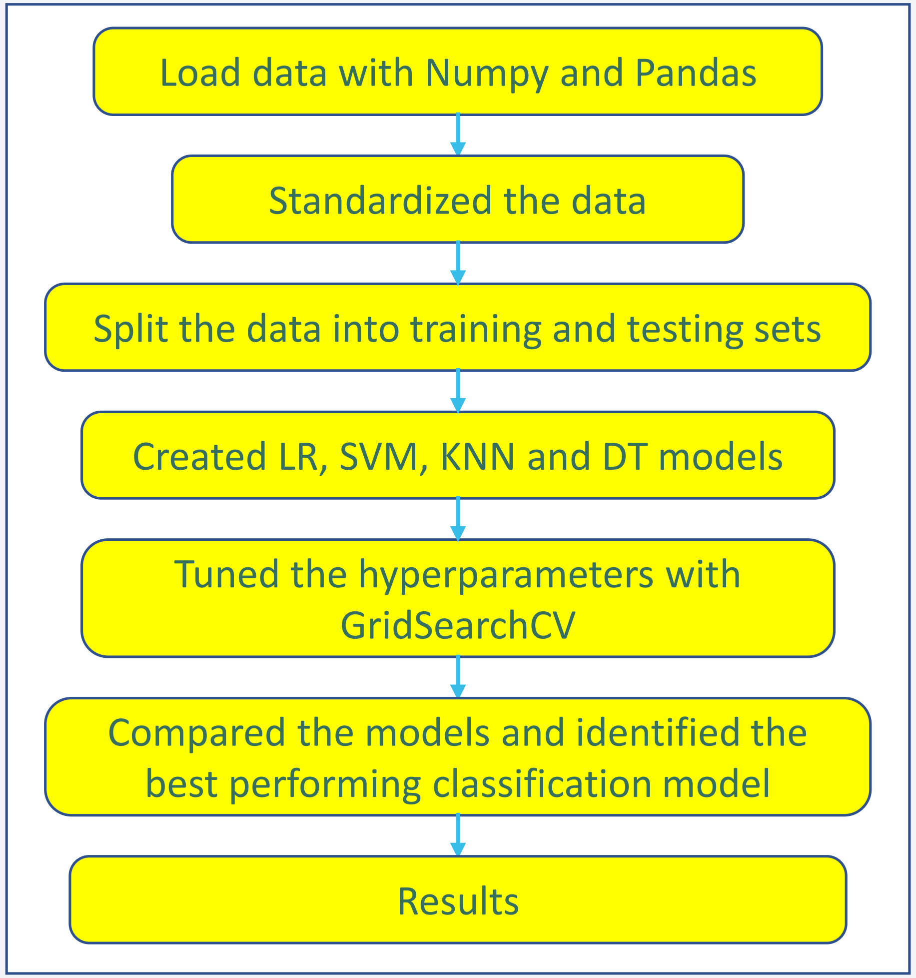
The dashboard was constructed in order to answer these questions:

1. Which site has the largest number of successful launches?
2. Which site has the highest launch success rate?
3. Which payload range(s) have the highest launch success rate?
4. Which payload range(s) have the lowest launch success rate?
5. Which F9 Booster version (`v1.0`, `v1.1`, `FT`, `B4`, `B5`, etc.) has the highest launch success rate?

Python File: https://github.com/james-chenery/IBM-Data-Science/blob/f8529265f6dfa534575b05d58849f4a480ba87c2/dashboard.py

Predictive Analysis (Classification)

- The data was loaded using numpy and pandas. It was transformed and split into sets for training and testing.
- Four different machine learning models were built and their hyperparameters tuned using GridSearchCV.
- The model with the highest accuracy was chosen as the best performing classification model.



Python Notebook: https://github.com/james-chenery/IBM-Data-Science/blob/96a07f7a5717480f9de24c19483b497727169ef2/ML-Prediction.ipynb

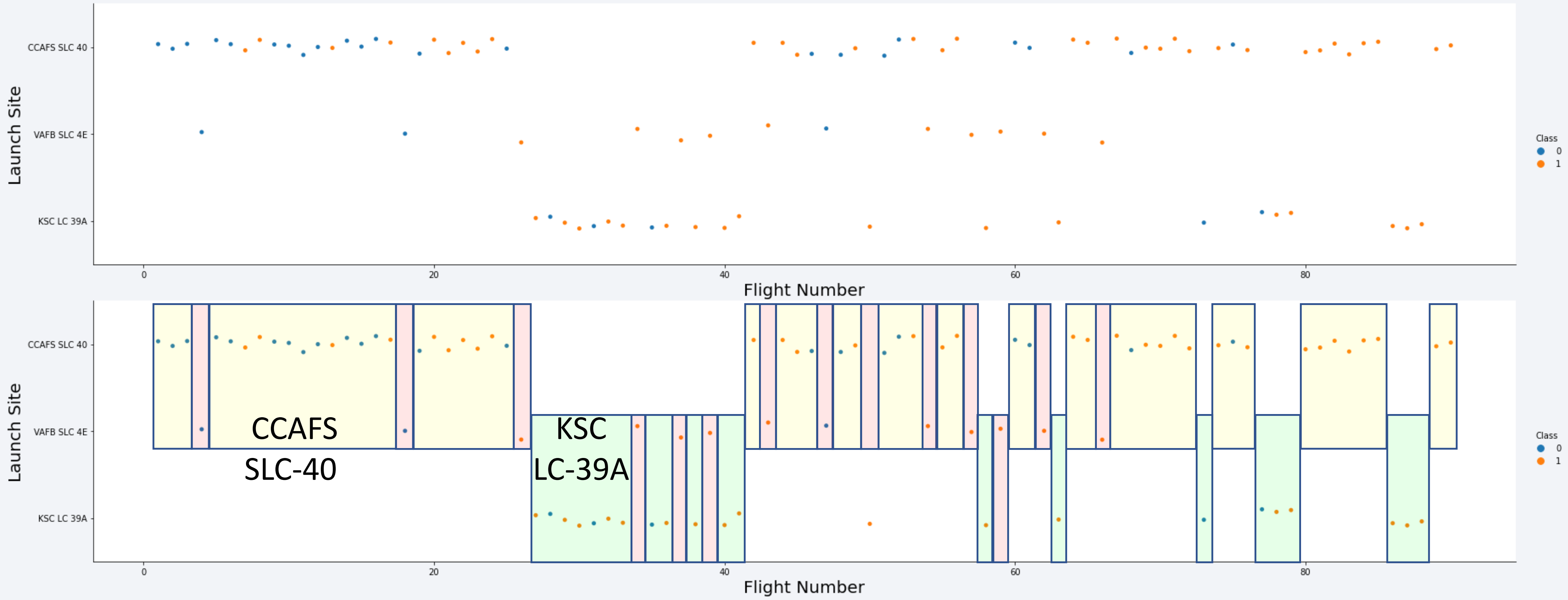
Results

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

Section 2

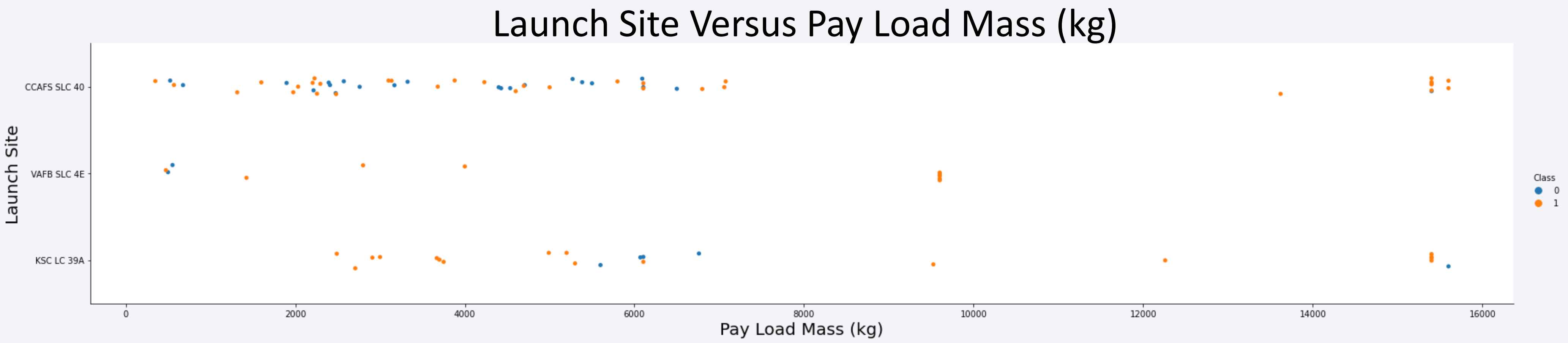
Insights drawn from EDA

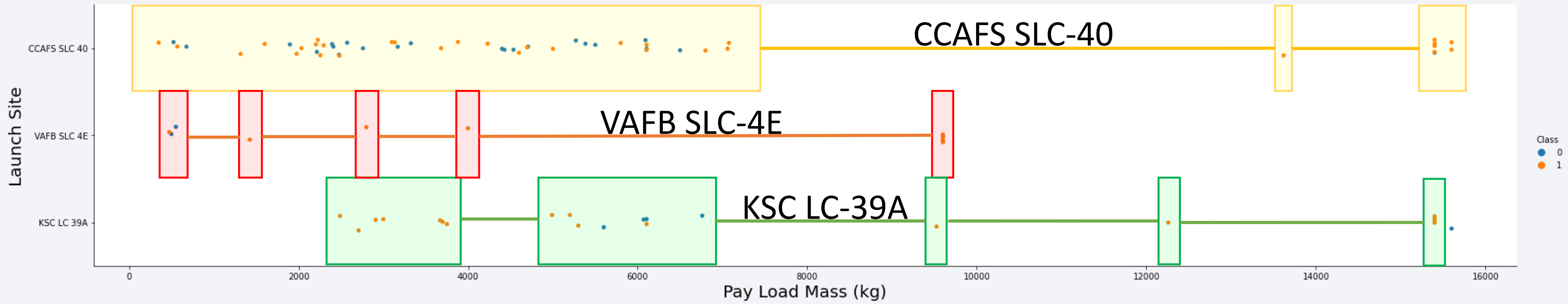
Flight Number vs. Launch Site



Flights alternate between CCAFS SLC-40 and KSC LC-39A, with occasional flights from VAFB SLC-4E.

Payload vs. Launch Site

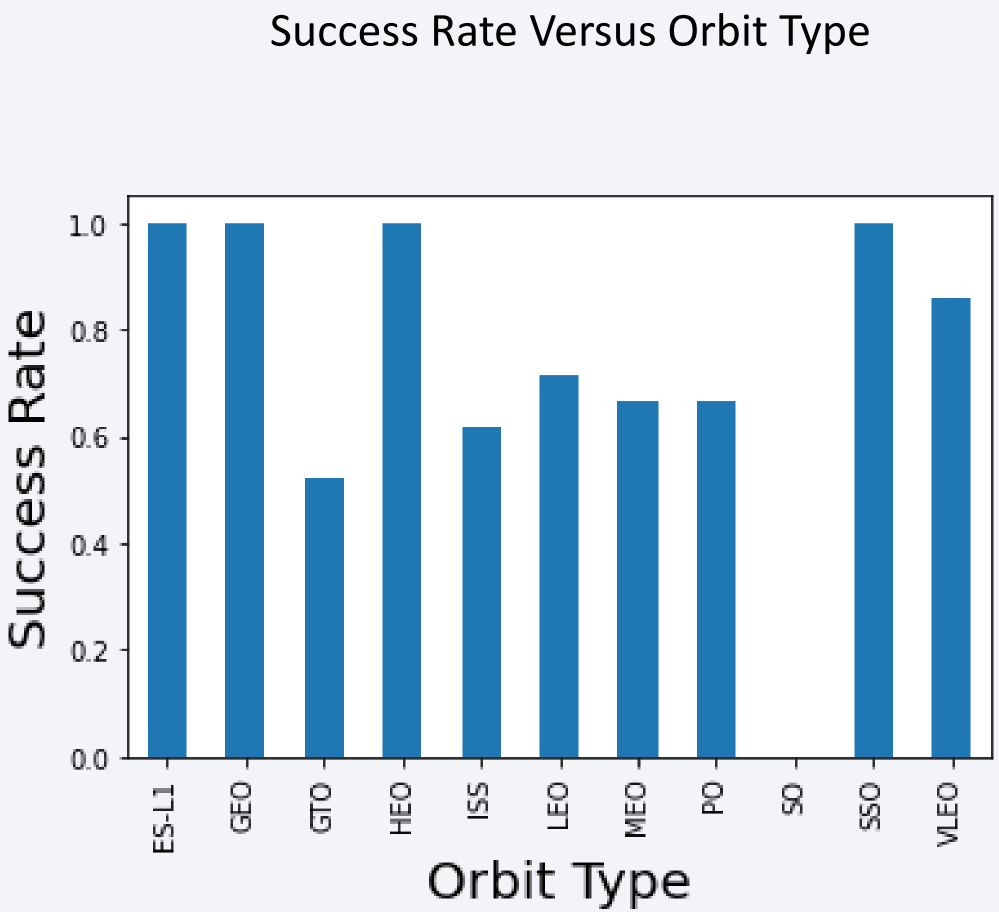




The main launch sites are CCAFS SLC-40 and KSC LC-39A, with what might be test flights from VAFB SLC-4E.

No heavy flights (10,000 kg or more) fly from VAFB SLC-4E.

Success Rate vs. Orbit Type

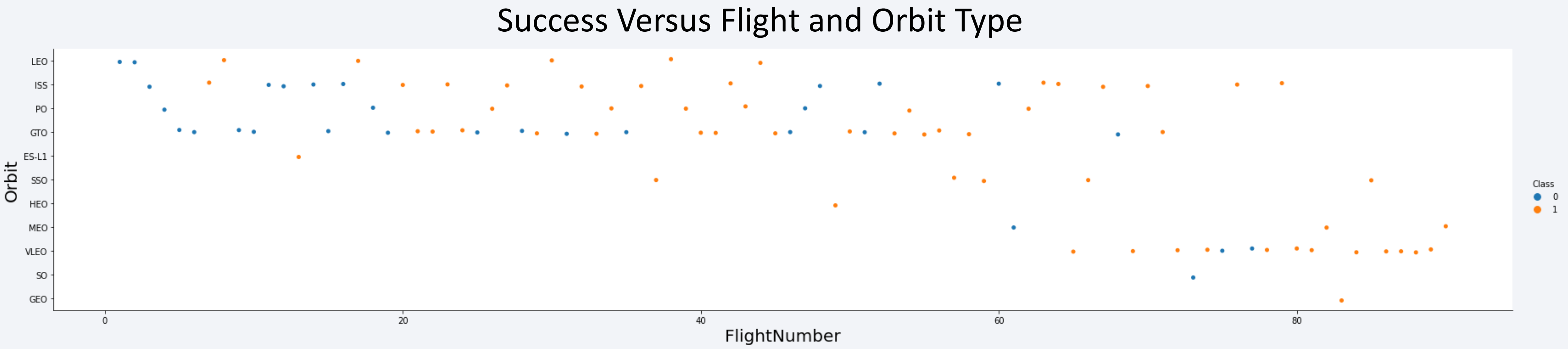


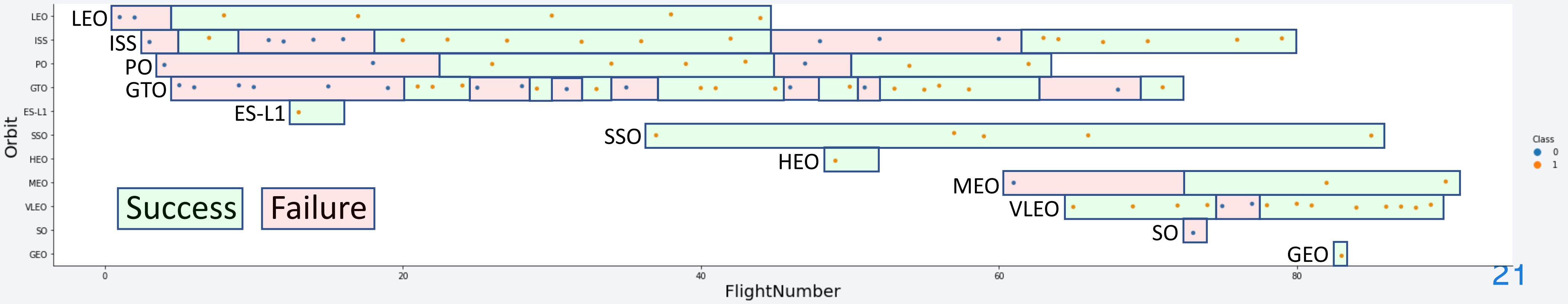
ES-L1, GEO, HEO and SSO have a 100% success rate. VLEO is over 80% success rate.

LEO, MEO, PO and ISS around a 65% success rate.

The GTO orbit has the worst success rate at 50%.

Flight Number vs. Orbit Type





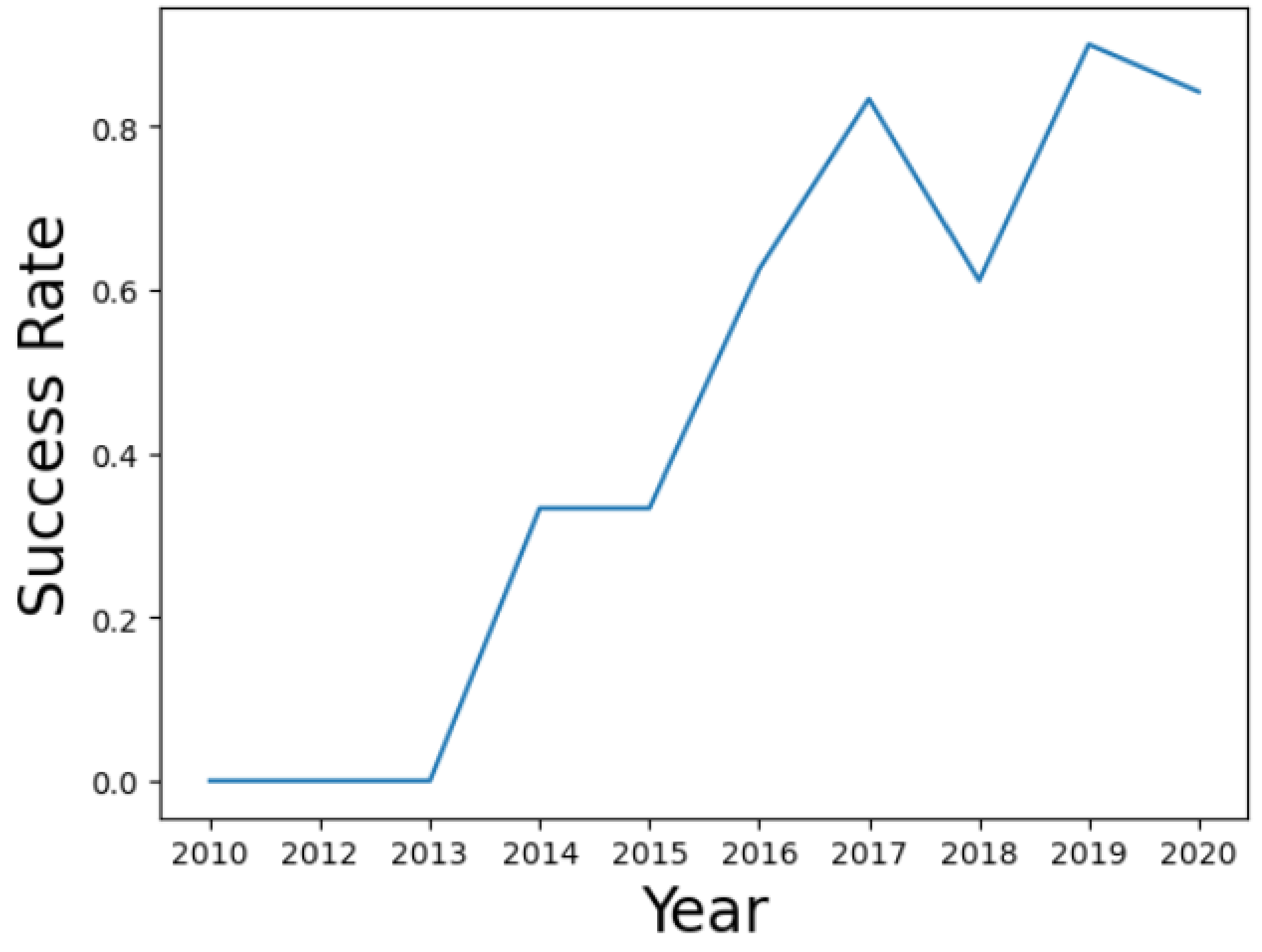
The earlier a flight is, among all SpaceX flights, the more likely it is to fail.
The earlier a flight is for a specific orbit, the more likely it is to fail.
There seems to be a strong correlation between experience and success.

Payload vs. Orbit Type

- LEO, ISS, PO and GTO began with a few light failures which may have been test/demo flights. PO and GTO have mixed performance for heavier payloads
- ES-L1, SSO, HEO, VLEO and GEO were successful regardless of payload.

Launch Success Yearly Trend

Line Chart of Yearly Average Success Rate



The success rate kept climbing until 2020.

All Launch Site Names

	Launch Site	Latitude	Longitude
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

The two launch sites beginning with "CCAFS" are located on the east coast, within the launch facilities of Cape Canaveral Space Force Station (formerly **C**ape **C**anaveral **A**ir **F**orce **S**tation.)

KSC LC-39A is located close to the **K**ennedy **S**pace **C**enter's Vehicle Assembly Building.

VAFB SLC-4E is located on the west coast, in Vandeburg Space Force Base (formerly **V**andenberg **A**ir **F**orce **B**ase.)

Launch Site Names Begin with 'CCA'

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

The first five launches from Cape Canaveral include three test flights.

Total Payload Mass

```
%sql SELECT "Customer", SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Customer" = "NASA (CRS)"
```

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

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

The total pay load carried by SpaceX for NASA (CRS) in the data is 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT "Booster_Version", AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1"
```

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

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

The total average pay load returned for the basic v1.1 version of the F9 booster was 2,928.4 kg.

First Successful Ground Landing Date

```
%sql SELECT "Landing _Outcome", MIN("Date") AS "Earliest Date" FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)"
```

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

Landing _Outcome	Earliest Date
Success (ground pad)	01-05-2017

The first successful F9 booster landing on a ground pad occurred on January 5^{th}, 2017.

Successful Drone Ship Landing with Payload between 4000 and 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.
```

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

- Four boosters have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") from SPACEXTBL GROUP BY "Mission_Outcome"
```

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

Mission_Outcome	COUNT("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- There were 101 successful missions, and just 1 failure.

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version", PAYLOAD_MASS__KG_ AS "Highest Payload" FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Booster_Version	Highest Payload
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

This list contains the boosters which have carried the maximum pay load mass of 15,600 kg.

2015 Launch Records

```
%sql SELECT SUBSTR(Date, 7, 4) AS "Year", SUBSTR(Date, 4, 2) AS "Month", "Booster_Version", "Launch_Site", "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" = "F
```

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

Year	Month	Booster_Version	Launch_Site	Landing _Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

This list contains the failed landing outcomes on drone ships in 2015, and the related booster versions, and launch site names

The query used was:

SELECT SUBSTR(Date, 7, 4) AS "Year", SUBSTR(Date, 4, 2) AS "Month", "Booster_Version", "Launch_Site", "Landing _Outcome"
FROM SPACEXTBL
WHERE "Landing _Outcome" = "Failure (drone ship)" AND SUBSTR(Date, 7, 4) = "2015"

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%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("La
```

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

Landing _Outcome	COUNT("Landing _Outcome")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

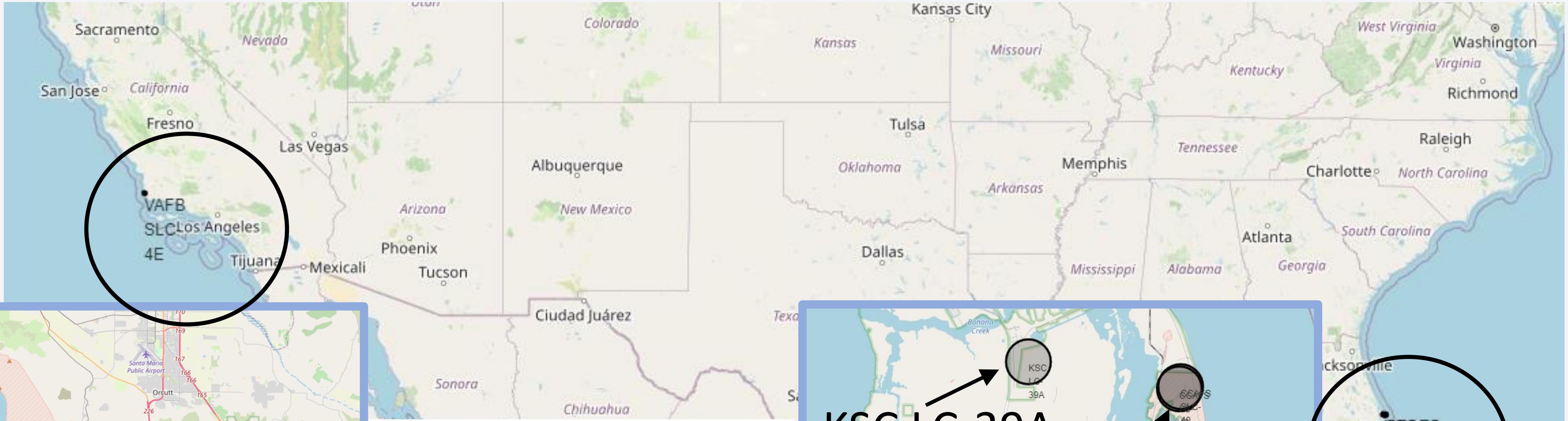
These landing results were obtained via the following query:

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

Section 3

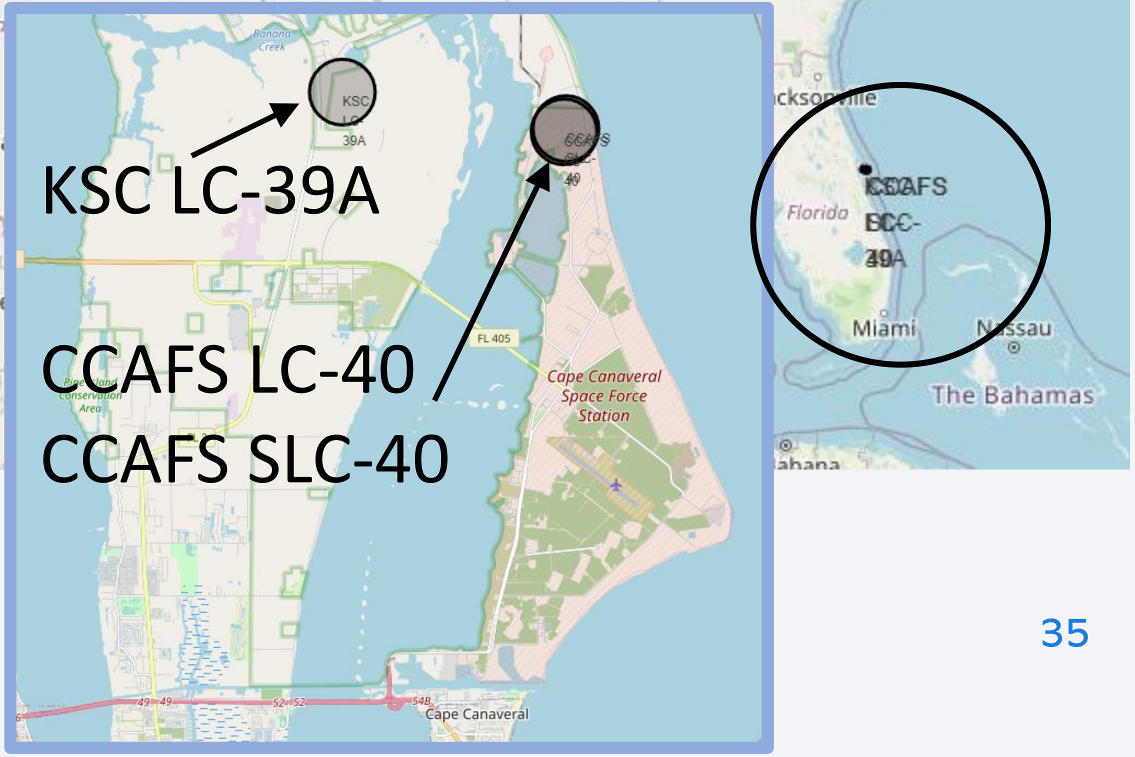
Launch Sites Proximities Analysis

Launch Site Locations



Launch sites are located at:

- Vandenburg Space Force Base
- Kennedy Space Center
- Cape Canaveral Space Force Station



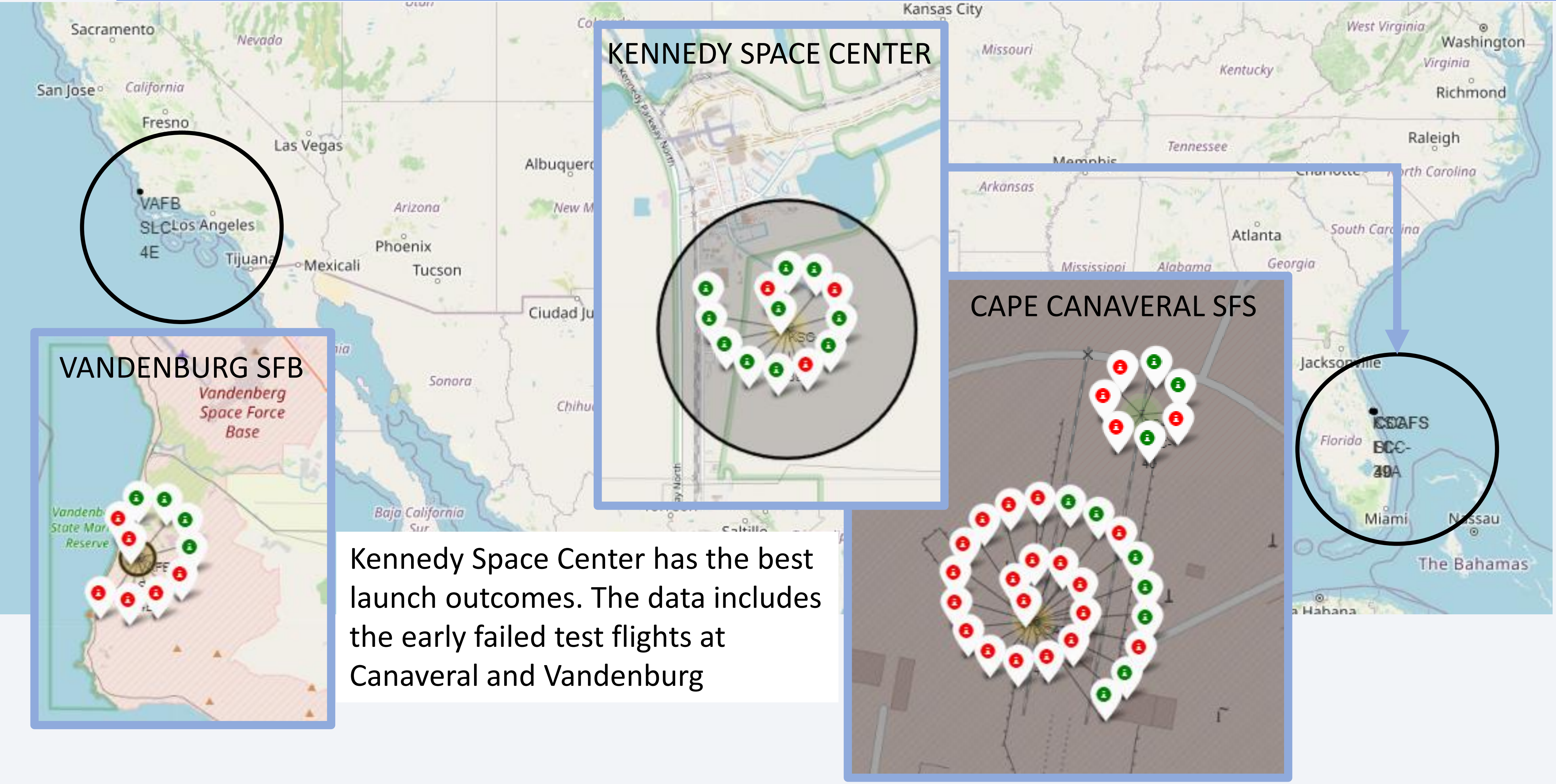
Launch Site Proximity to the Equator

	Launch Site	Latitude	Longitude	Distance from Equator
0	CCAFS LC-40	28.562302	-80.577356	3175.83 km
1	CCAFS SLC-40	28.563197	-80.576820	3175.93 km
2	KSC LC-39A	28.573255	-80.646895	3177.05 km
3	VAFB SLC-4E	34.632834	-120.610745	3850.81 km
4	**ELS-Soyuz**	**5.30217**	**-52.83337**	**589.55 km**

The launch sites are not very close to the equator, so the additional cost of attaining most orbits must be offset by other factors, e.g. locally skilled workforce, security, accessibility, etc.

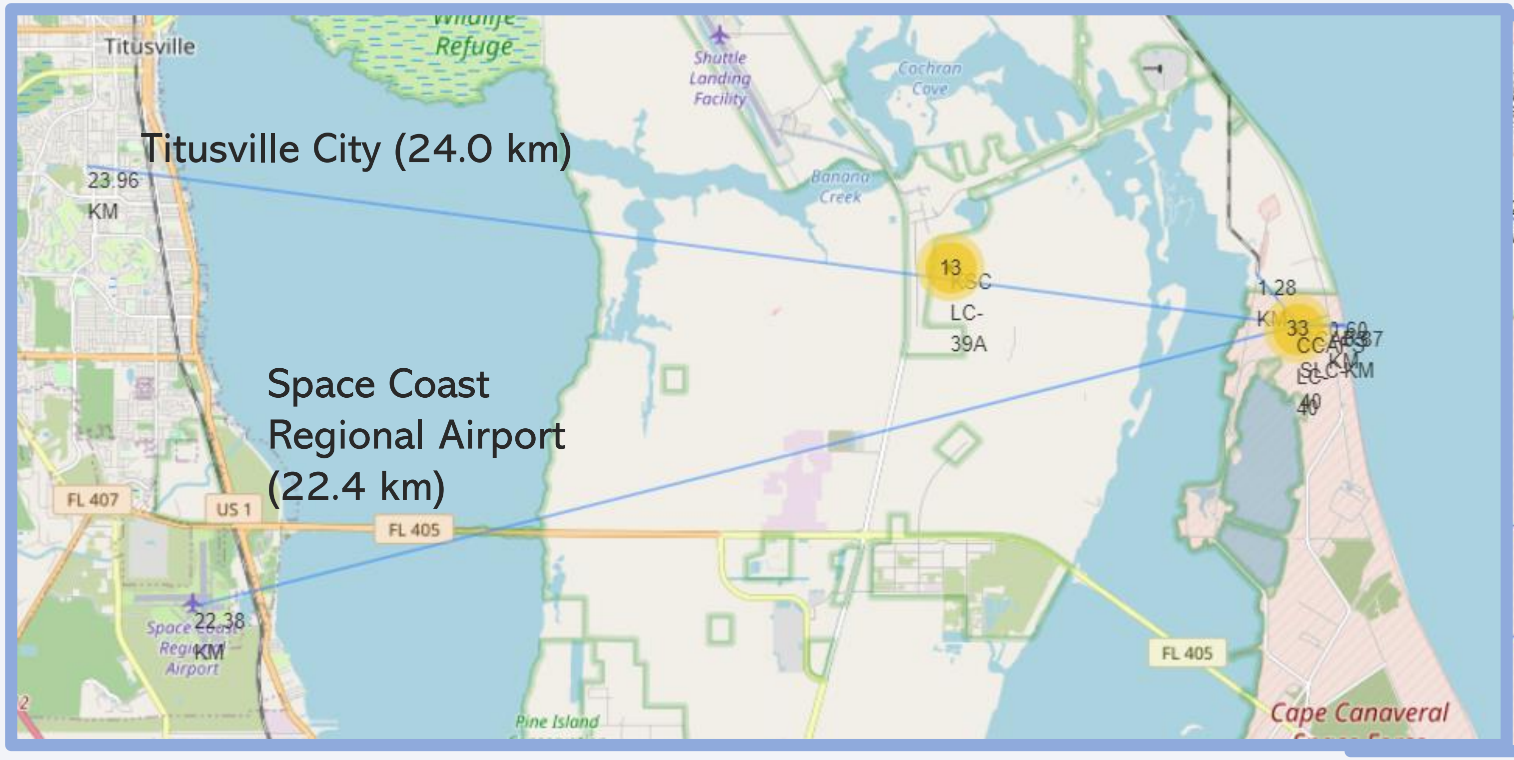
The **equatorial** ELS-Soyuz launch pad in French Guiana has been added for comparison.

Launch Site Outcomes



Kennedy Space Center has the best launch outcomes. The data includes the early failed test flights at Canaveral and Vandenburg

Cape Canaveral Launch Site Proximities Analysis



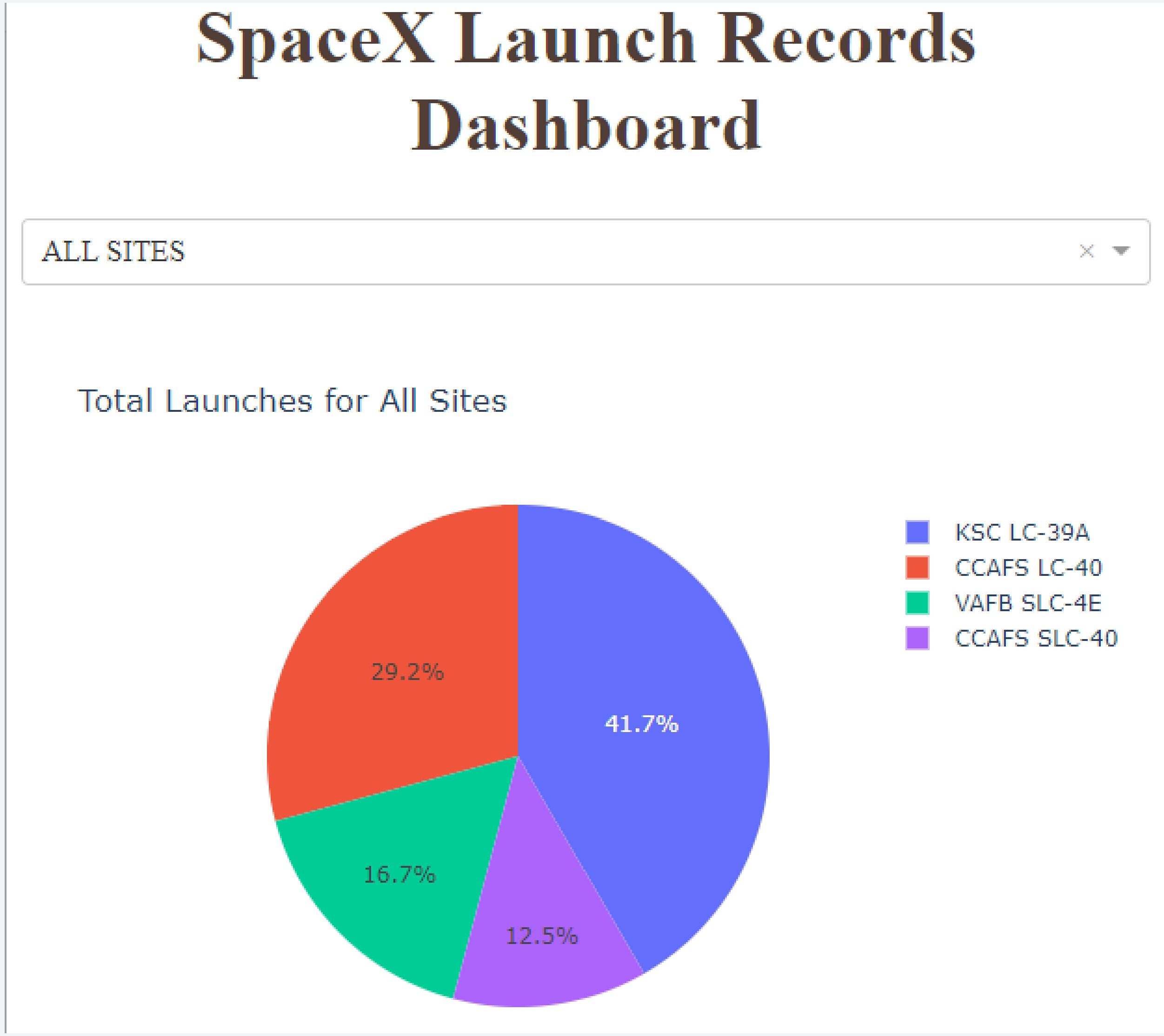
- **596.9 m** • Closest_highway = [28.56483, -80.571] # Closest highway exit
- **1.28 km** • Closest_railroad = [28.57205, -80.58524] # Railway Terminus
- **22.4 km** • Closest_airport = [28.513560, -80.798769] # Space Coast Regional Airport
- **24.0 km** • Closest_city = [28.591111, -80.82] # Titusville (City)

The launch site combines the benefits of excellent transportation links, a highly secure coastal location safely removed from heavily populated areas, and proximity to aerospace infrastructure and expertise.

Section 4

Build a Dashboard with Plotly Dash

Interactive Dashboard (All Launch Sites)

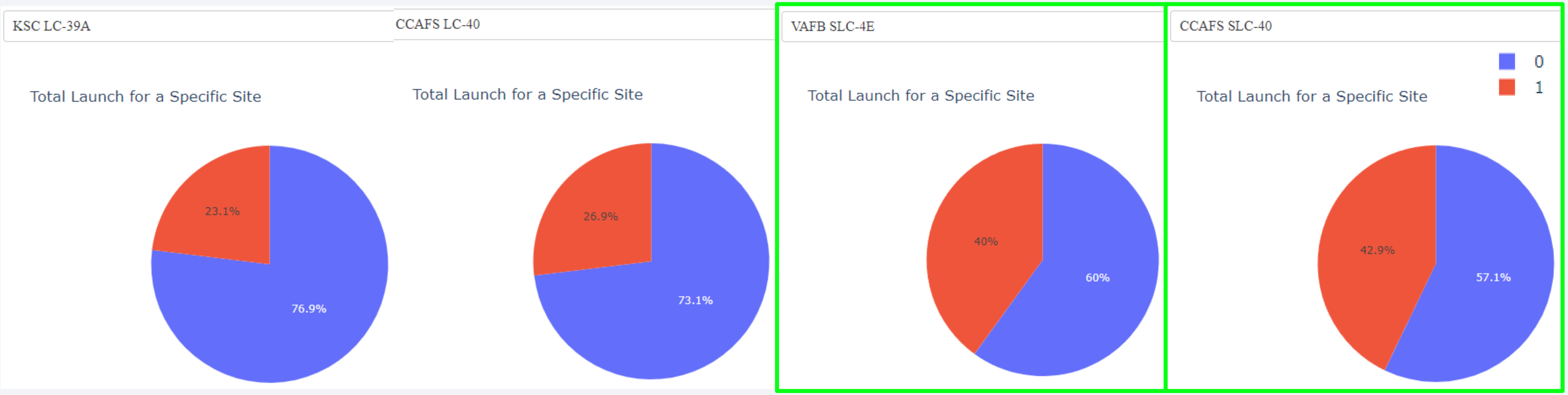


KSC has 41.7% of all launches, however, it is not the most.

Cape Canaveral renamed LC-40 to SLC-40, so the pad's total percentage is **also** 41.7%. KSC and CCAFS have therefore operated the same number of launches.

Vandenburg Space Force Base's SLC-4E launch site is the third most used, at 16.7%.

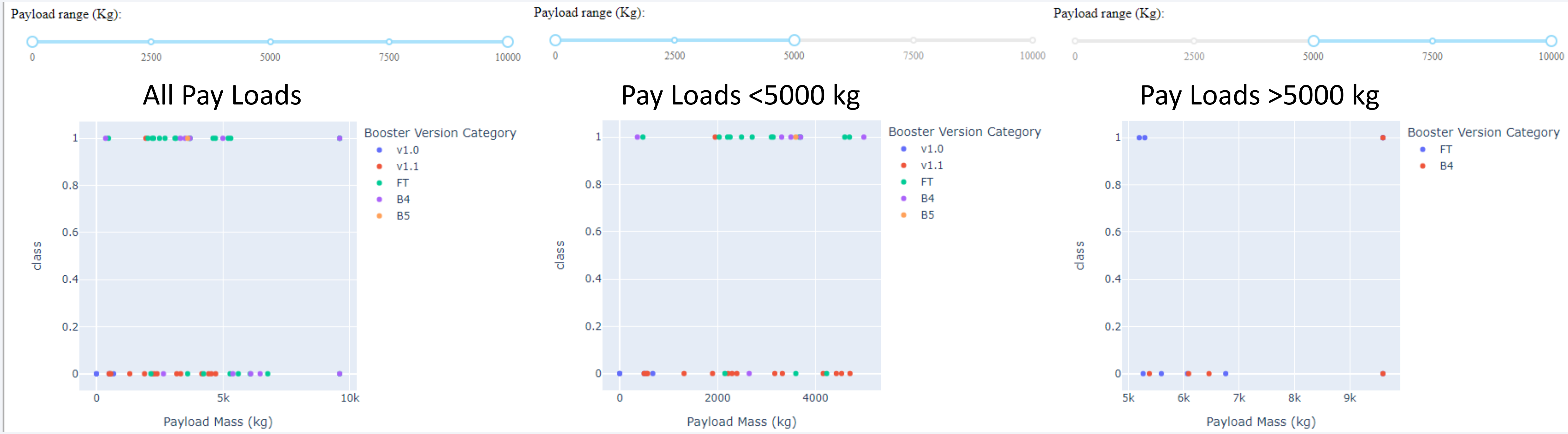
Interactive Dashboard (Landing Success Per Launch Site)



CCAFS SLC-40 has the best landing success rate at 42% (including test flights and deliberate no-landing launches.) CCAFS LC-40's statistics pre-date SLC-40's (due to the pad's name change) and therefore likely include a lot of early test/demo launches.

VAFB SLC-4E has the second best landing success rate at 40%. CCAFS LC-40 and KSC LC-39A have lost 73.1% and 76.9% of boosters, respectively.

Interactive Dashboard (Payload vs. Success Scatter)

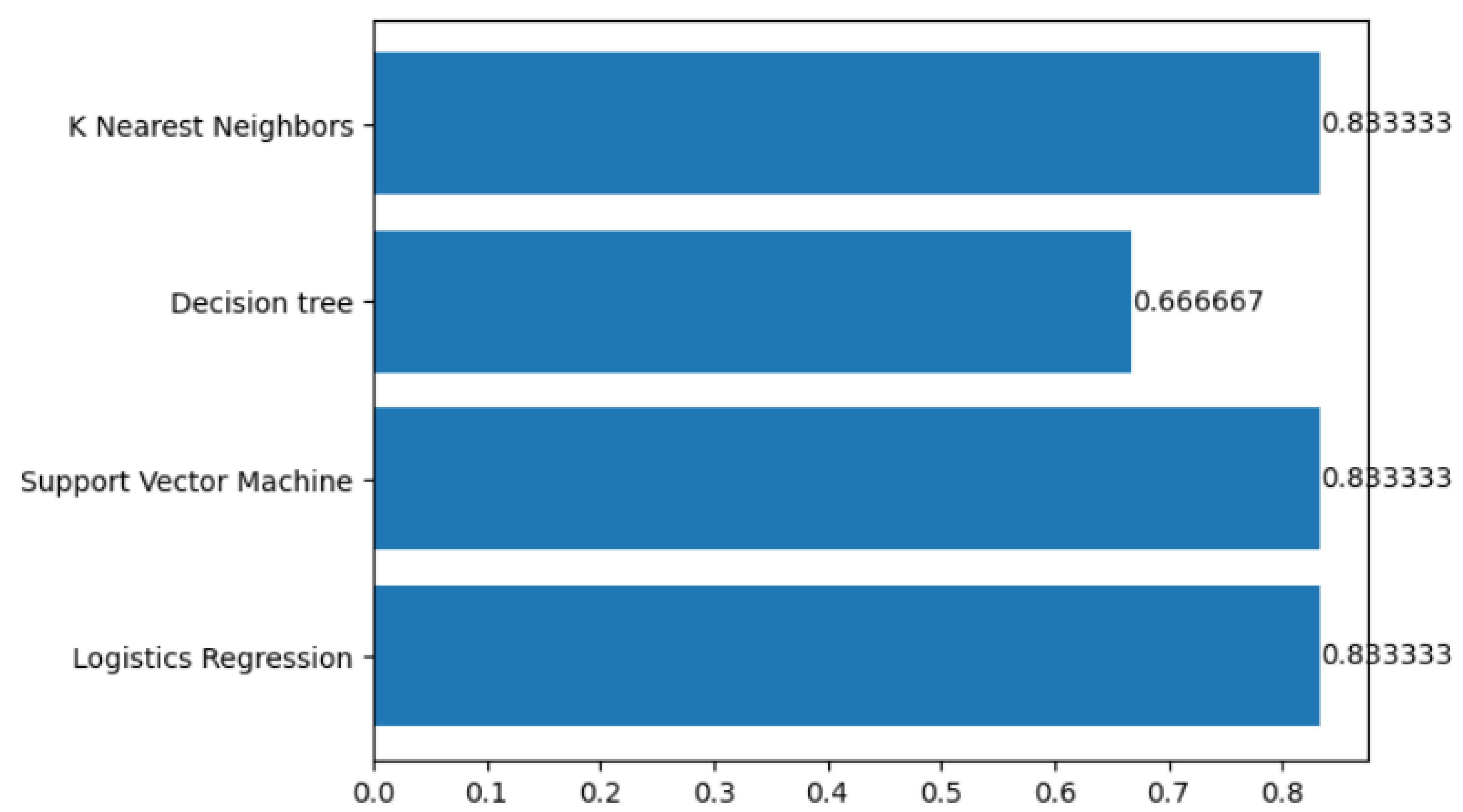


Successful landing of the first stage booster becomes increasingly unlikely as pay load increases, possibly indicating a lack of fuel for landing due to the additional mass.

Section 5

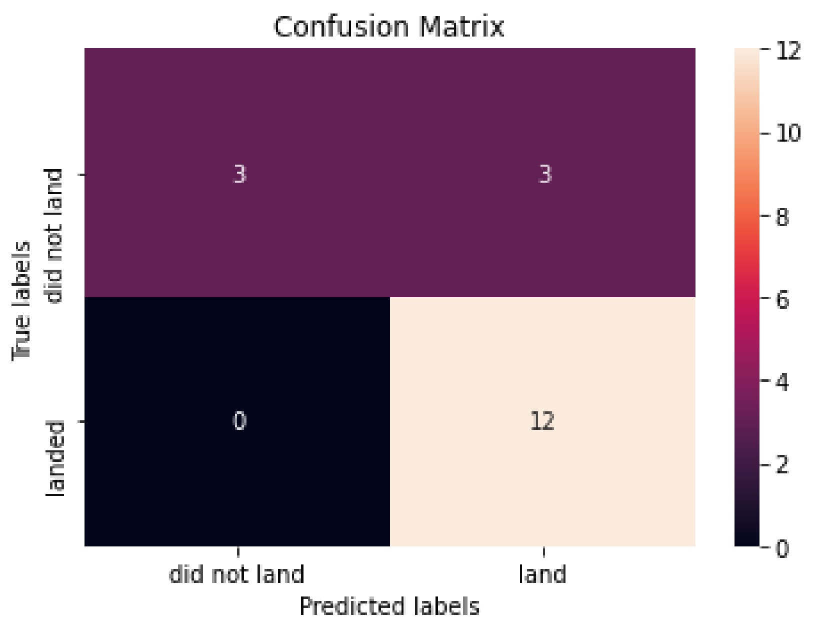
Predictive Analysis (Classification)

Classification Accuracy



The decision tree model has lower accuracy. The other three models were found to have equivalent accuracy.

Confusion Matrix



- LR, SVM and KNN models had the best accuracy.
- The confusion matrix shows that the classifier can distinguish between the different classes.
- The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions

- Initial launches to a given orbit had a high landing failure rate.
- The most successful orbits for landing are:
  ES-L1, GEO, HEO, SSO and VLEO
- The landing success rate increased from 2013 to 2020.
- The current landing rate for all launches SpaceX's Eastern launch facilities stands at 40-42%
- From 2020, more launches failed to land as payloads tended to be heavier.
- The exception are SSO orbit launches, which loft heavy loads, but have a flawless landing rate.
- Launches with heavier pay loads (>5000 kg) have a greater probability of failing to land their boosters.

Appendix

Additional content have been interspersed into this presentation in appropriate places, e.g.

3	VAFB SLC-4E	34.632834	-120.610745	3850.81 km
4	**ELS-Soyuz**	**5.30217**	**-52.83337**	**589.55 km**

The launch sites are not very close to the equator, so the additional cost of attaining most orbits must be offset by other factors, e.g. locally skilled workforce, security, accessibility, etc.

The **equatorial** ELS-Soyuz launch pad in French Guiana has been added for comparison.

or:

- Closest_highway = [28.56483, -80.571] # Closest highway exit
- Closest_railroad = [28.57205, -80.58524] # Railway Terminus
- Closest_airport = [28.513560, -80.798769] # Space Coast Regional Airport

The l
trans
locat
areas

Thank you!