Winning Space Race with Data Science

Keerthana T
1st October 2025

Outline

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

Executive Summary

- Collected data from SpaceXAPI to help SpaceY compete with SpaceX by predicting rocket reusability

- Built interactive dashboards to visualize launch costs, orbits, and outcomes
- Trained ML models predict first-stage reusability
- Accuracy of models: 83%

Introduction

- **SpaceX Falcon 9**: Advertised launch cost $62M vs. competitors at $165M+
- **Cost Advantage:** Achieved through reusing the first stage of the rocket
- **Falcon 9 structure:**
  - **Payload** enclosed in fairings
  - **Second stage** delivers payload to orbit
  - **First stage** does most of the work, is larger and more expensive

Section 1

Methodology

Methodology

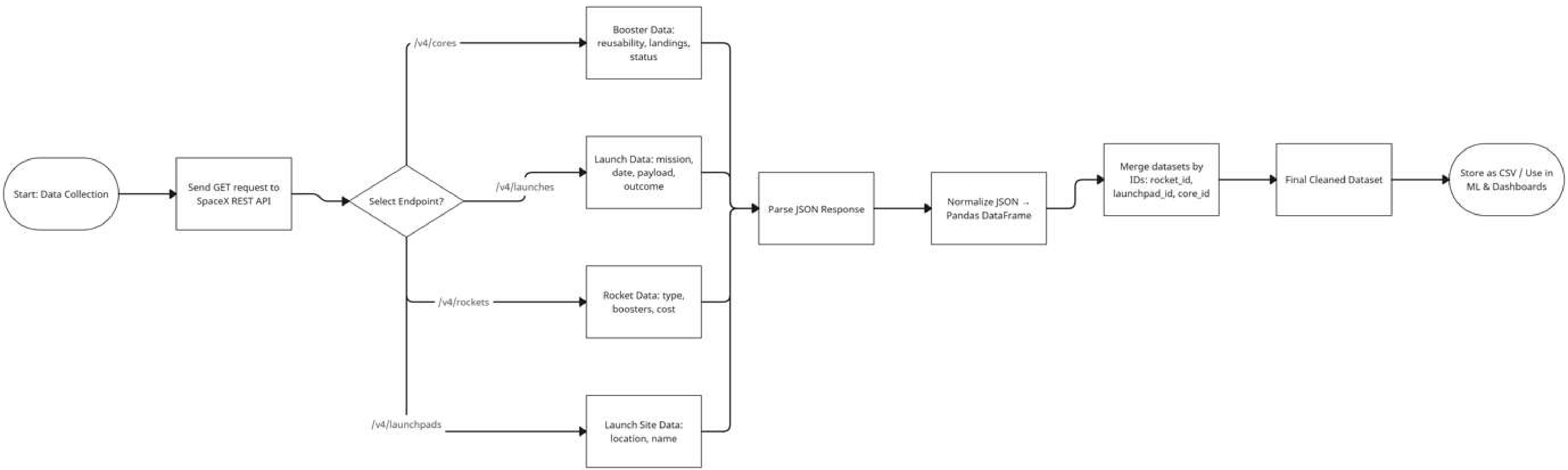
Executive Summary

- Data collection methodology:
  - Collected data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling

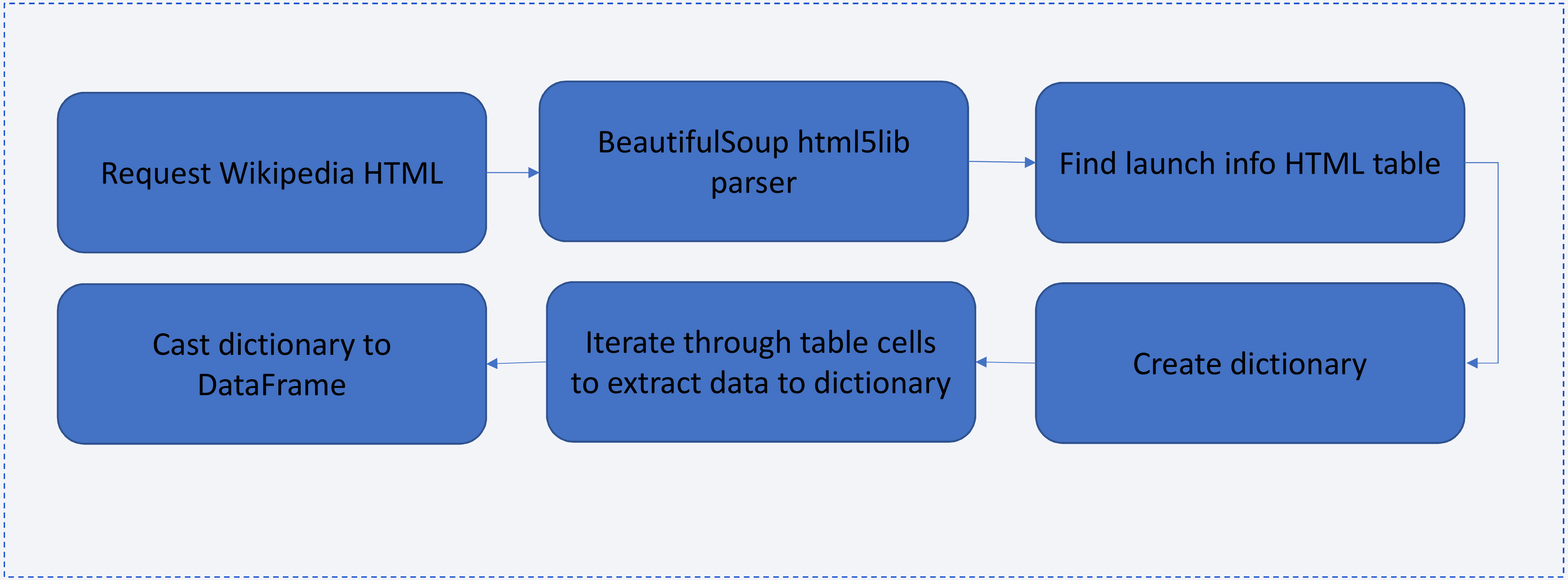
  * Classifying true landings as successful and unsuccessful otherwise
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

- Collected data using SpaceXAPI by doing some web scraping on the SpaceX website
- Space X API Data Columns:
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time


Start: Data Collection
Send GET request to SpaceX REST API
Select Endpoint?
/v4/cores
/v4/launches
/v4/rockets
/v4/launchpads
Booster Data: reusability, landings, status
Launch Data: mission, date, payload, outcome
Rocket Data: type, boosters, cost
Launch Site Data: location, name
Parse JSON Response
Normalize JSON → Pandas DataFrame
Merge datasets by IDs: rocket_id, launchpad_id, core_id
Final Cleaned Dataset
Store as CSV / Use in ML & Dashboards

Data Collection - Scraping



Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: 'Mission Outcome' 'Landing Location'
- New training label column 'class' with a value of 1 if 'Mission Outcome' is True and 0

otherwise. Value Mapping:

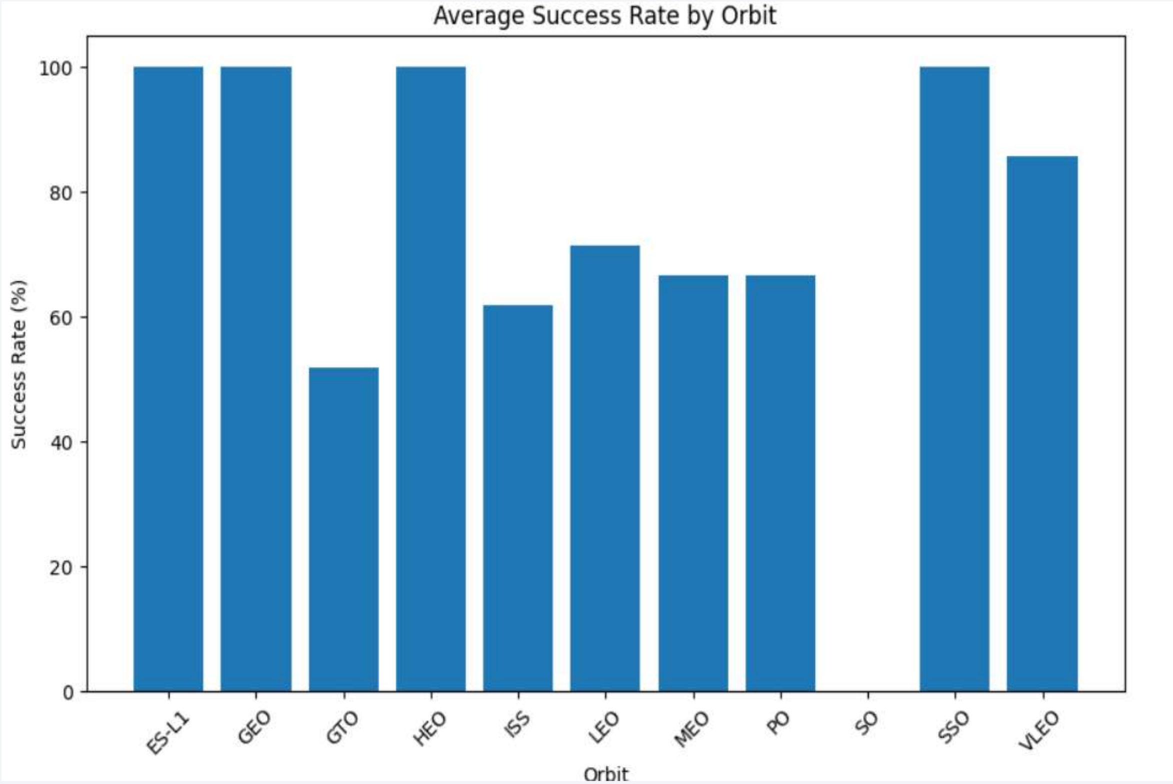
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

Github Link

https://github.com/kt-sudo/Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

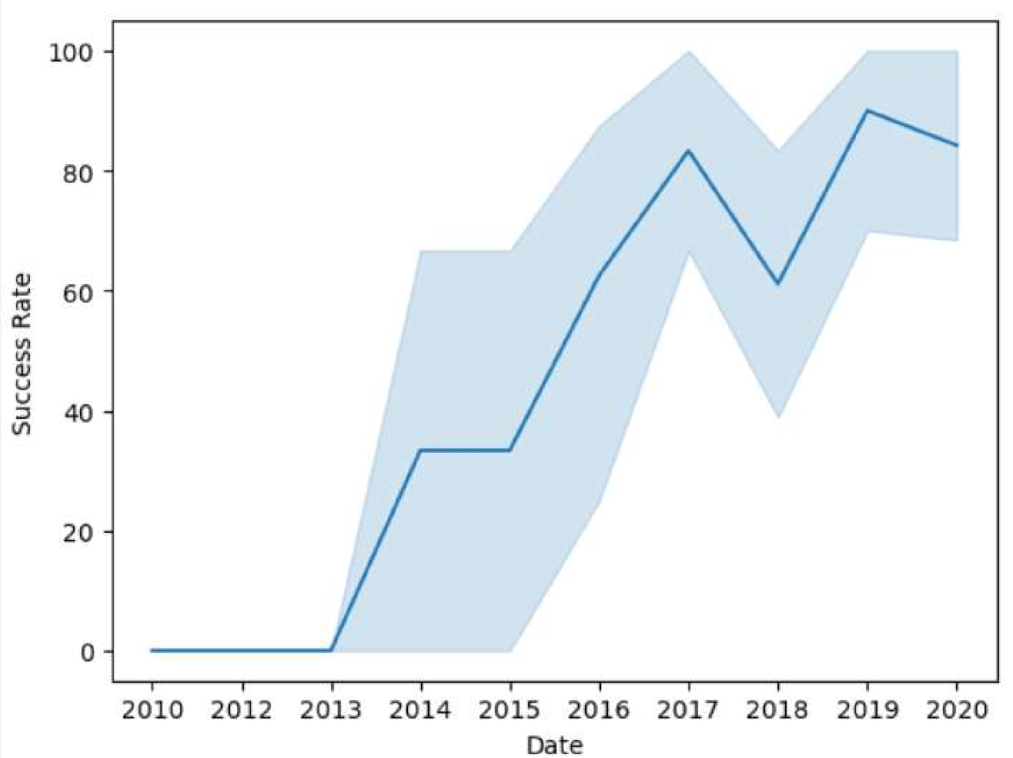
EDA with Data Visualization

- Visualized the relationship between
  - Payload and LaunchSite
  - Success rate of each orbit type (graph on the right)
  - FlightNumber and Orbit Type
  - Payload Mass and Orbit Type
  - Launch success and year



EDA with SQL

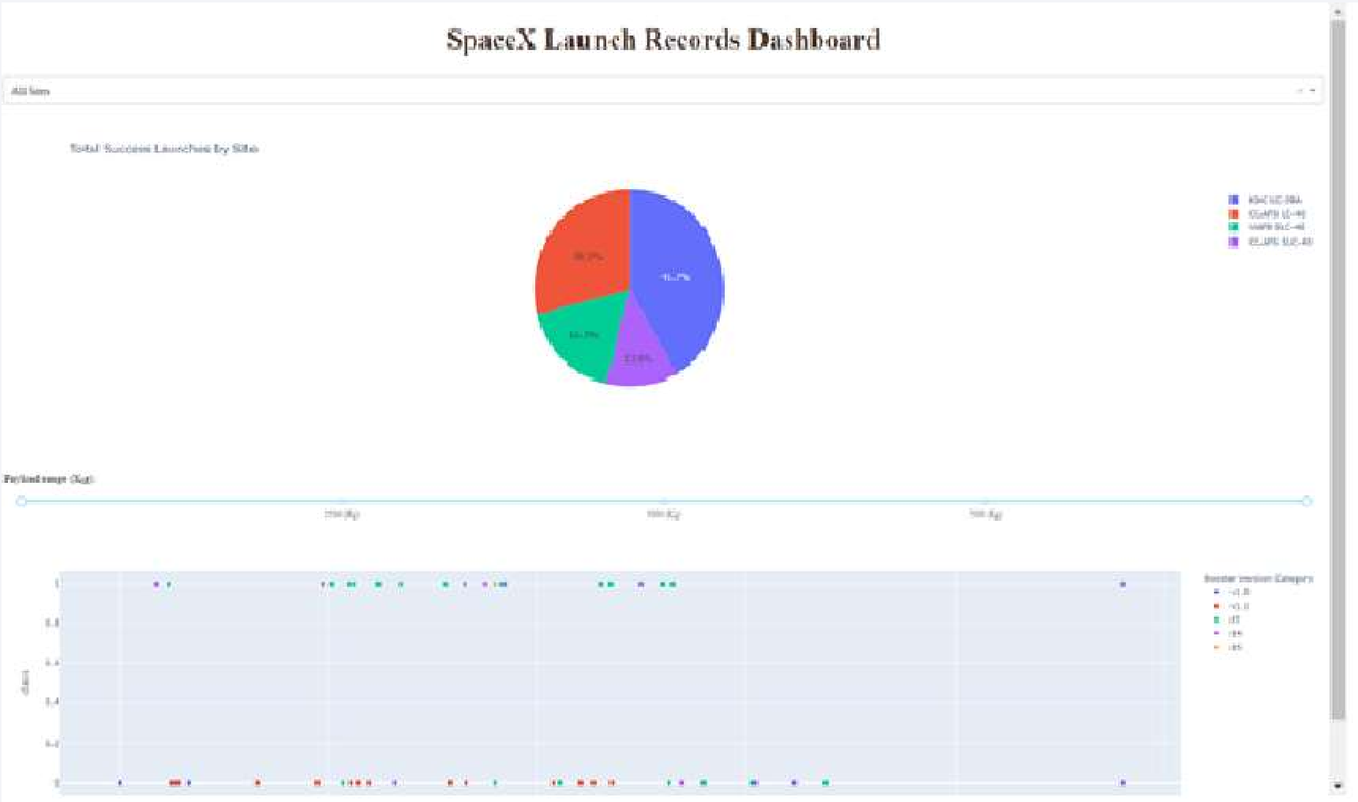
Launch Success Yearly Trend graph is shown on the right side



Build an Interactive Map with Folium

Optimal location for building launch site

Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

- Used Logistic Regression, Decision tree classifier, support vector machine and KNN models
- Used GridSearchCV to find accuracy on the test set at each case

Results

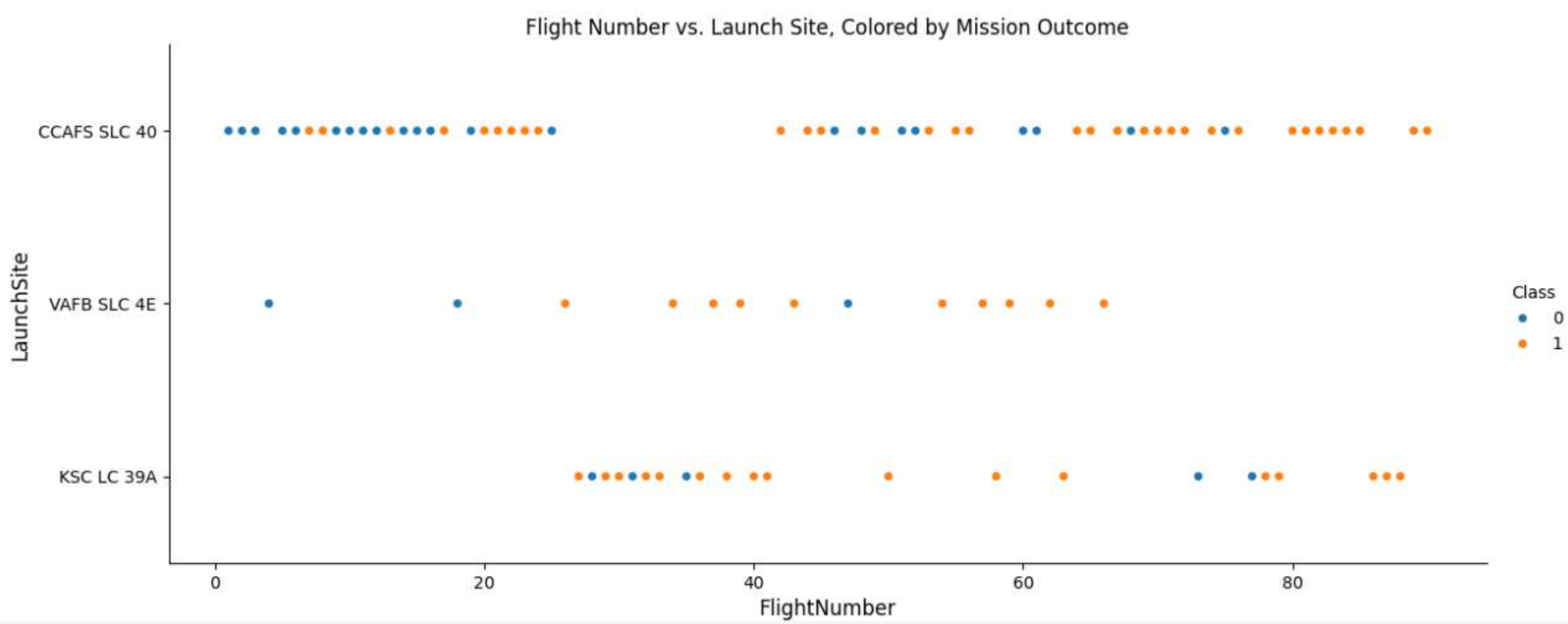
Accuracy for each of these models: KNN, SVM, logistic regression, decision tree classifier: 83%

Out of all, Decision Tree Classifier model is the best.

Section 2

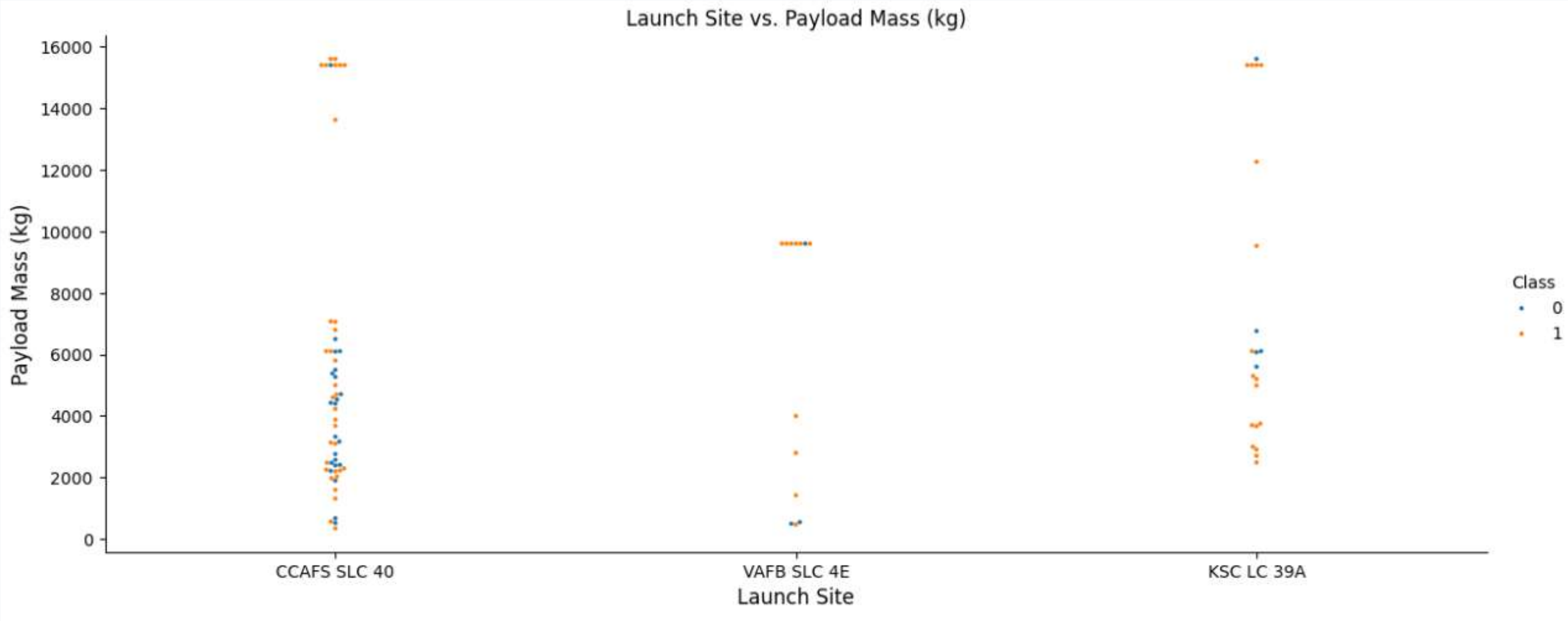
Insights drawn from EDA

Flight Number vs. Launch Site



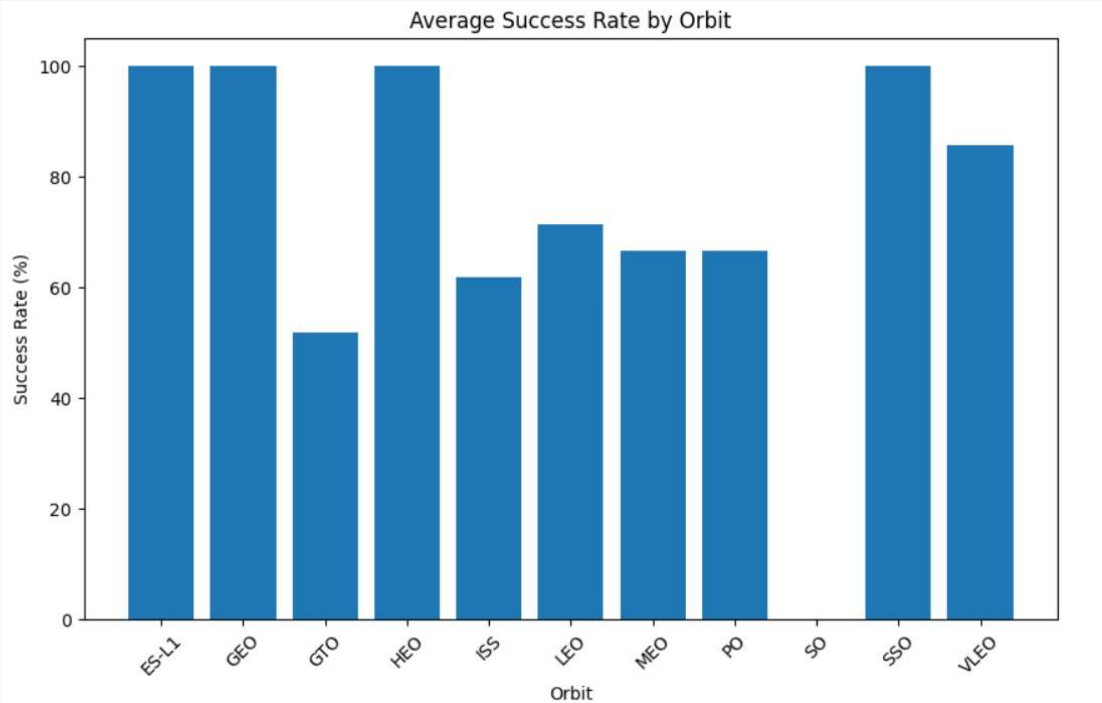
"CCAFS SLC 40" launch site has the highest number of flights with most successful launches.

Payload vs. Launch Site



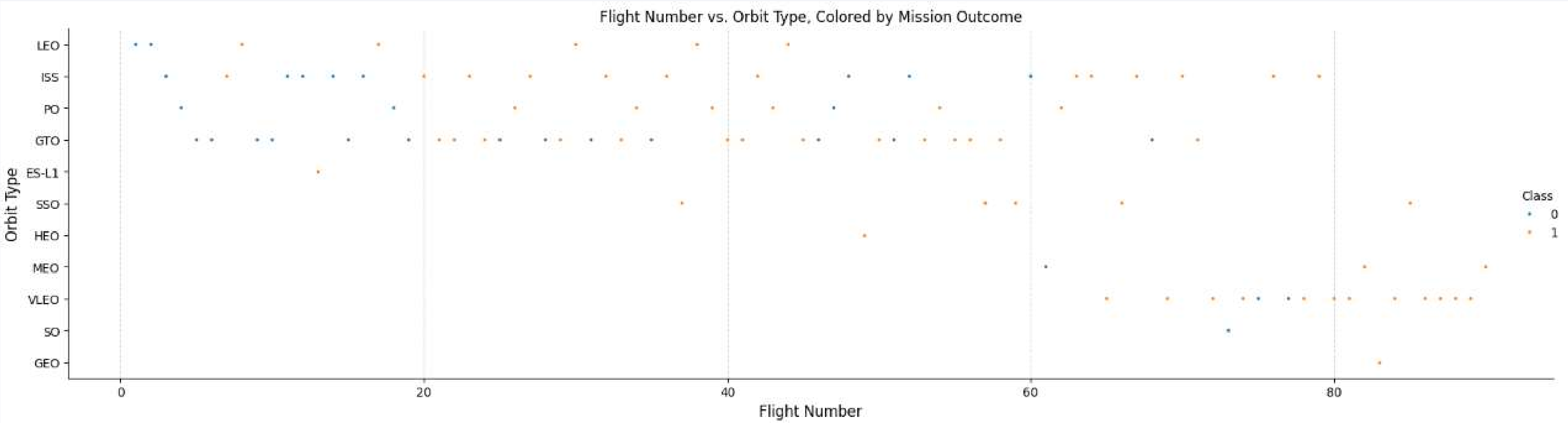
Highest payload mass for ‘CCAFS SLC 40’ launch site

Success Rate vs. Orbit Type



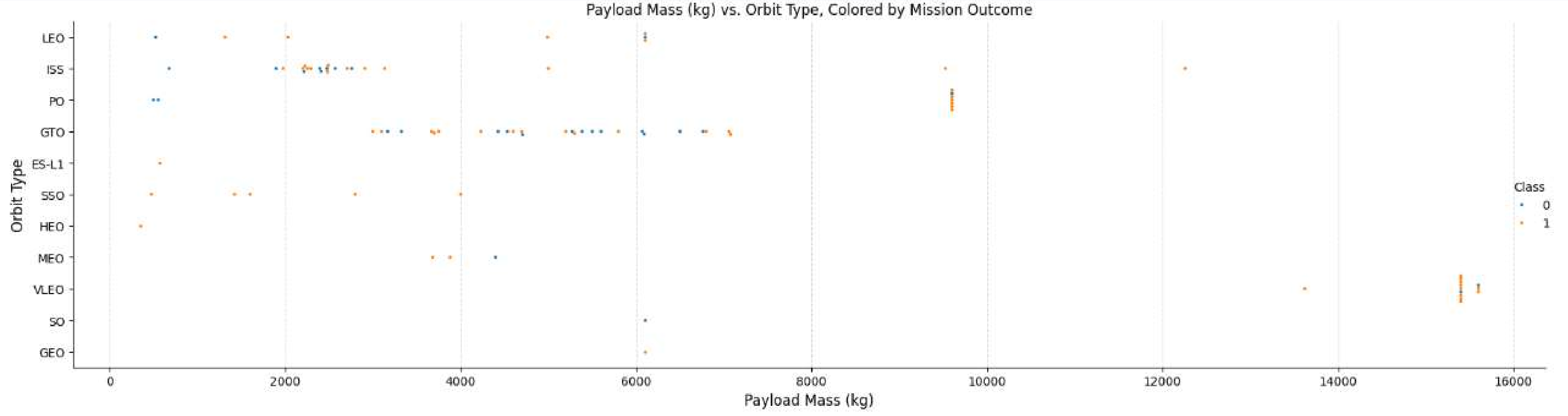
- Success rate for ES-L1, GEO, HEO, SEO is close to 100

Flight Number vs. Orbit Type



Maximum flights are of orbit type VLEO out of which 86% flights were successful

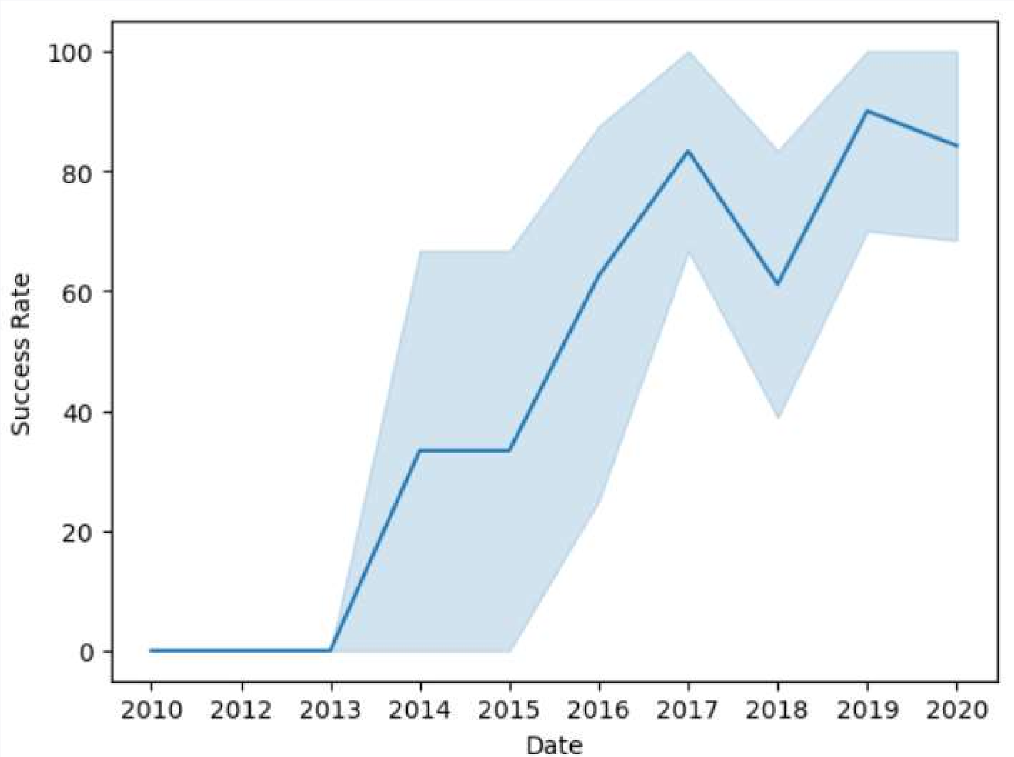
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Throughout 2013 and 2020 success rates have been steadily increasing



All Launch Site Names

The listed are the names of all launch site.

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

Launch Site Names Begin with 'CCA'

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

- Launch site names that begin with ‘CCA’

Total Payload Mass

TotalPayloadMass
45596

- Total payload mass is 45596

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass is 2928.4 in kilograms.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

- First Successful Ground Landing Date is 22nd December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Booster versions of successful drone ship landing with payload >4000 but <6000

Total Number of Successful and Failure Mission Outcomes

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

- Out of 103, 98+1 has success and 1 had success with payload status unclear

Boosters Carried Maximum Payload

- Booster versions that carried maximum payload

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

- Launches were made in the months of January, February, March, April, June and December with most launches made in April

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	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
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- 5 success (drone ship), 5 failure(drone ship)
- 3 success (ground pad)

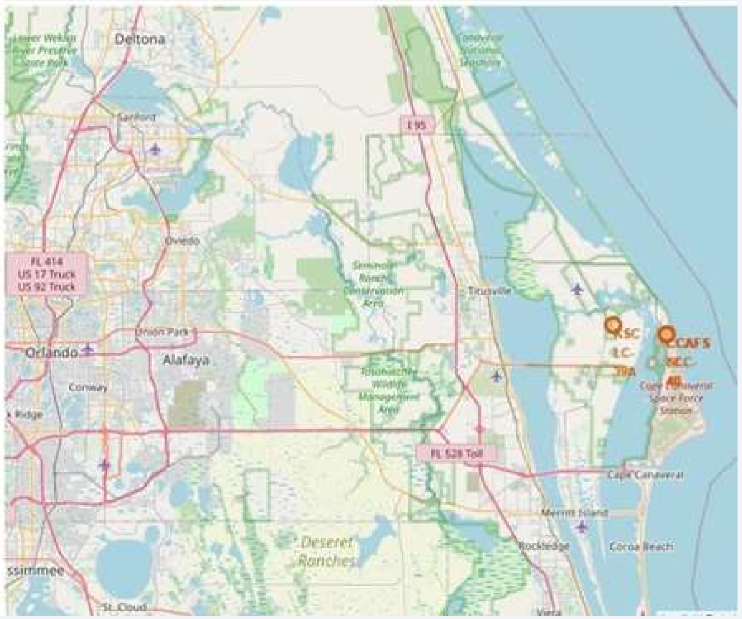
Attempts were made between 2010-06-04 and 2017-03-20

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

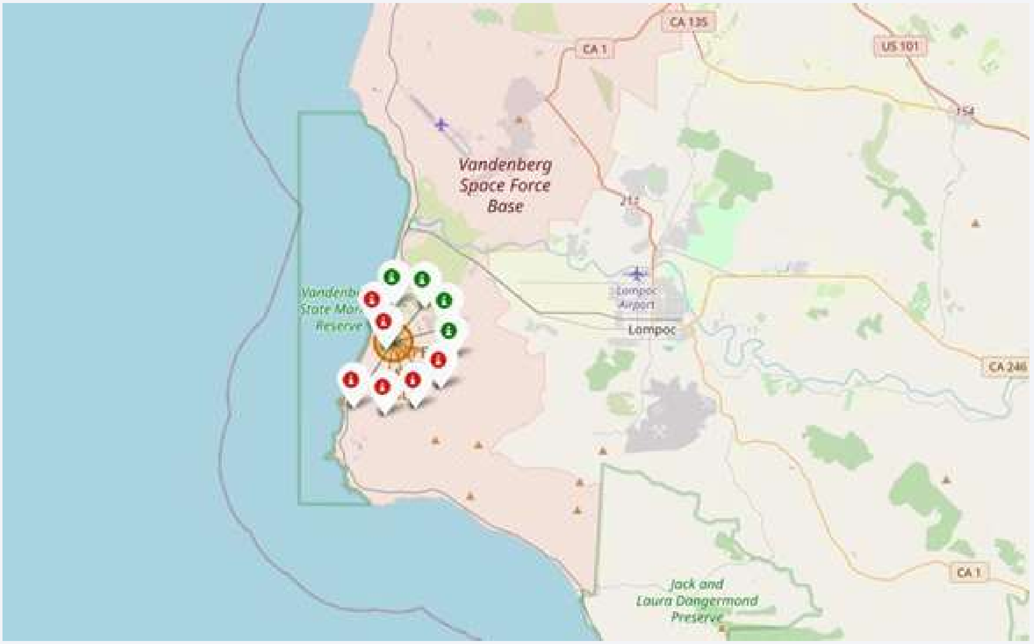
Launch Sites Proximities Analysis

Launch Site Locations



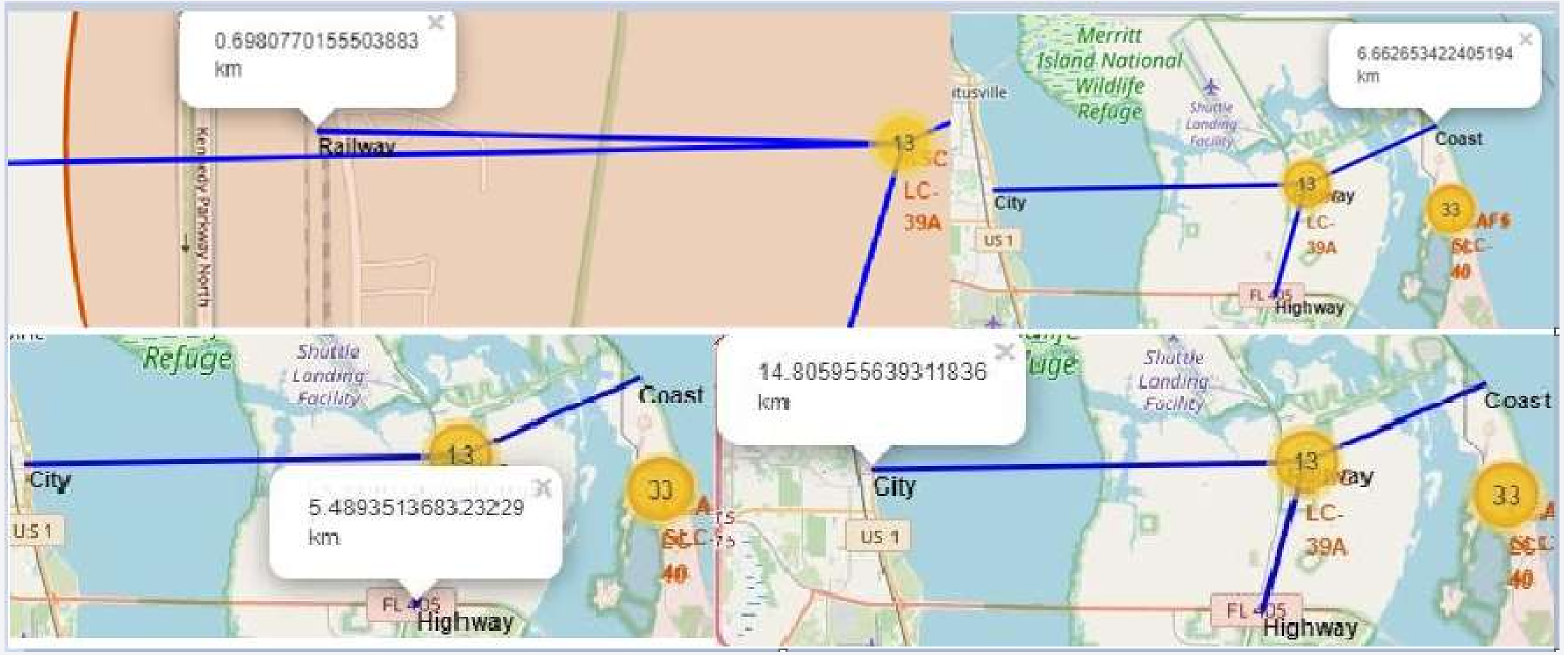
- The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-coded Launch markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

Key Location Proximities

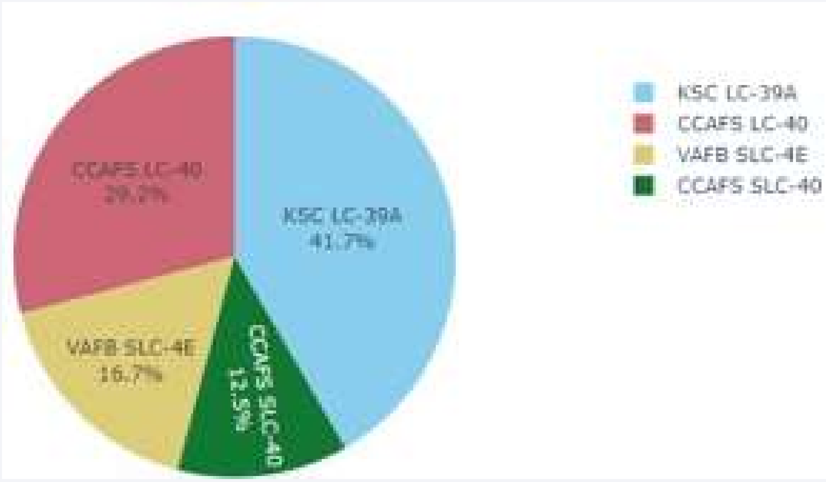


• Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Section 4

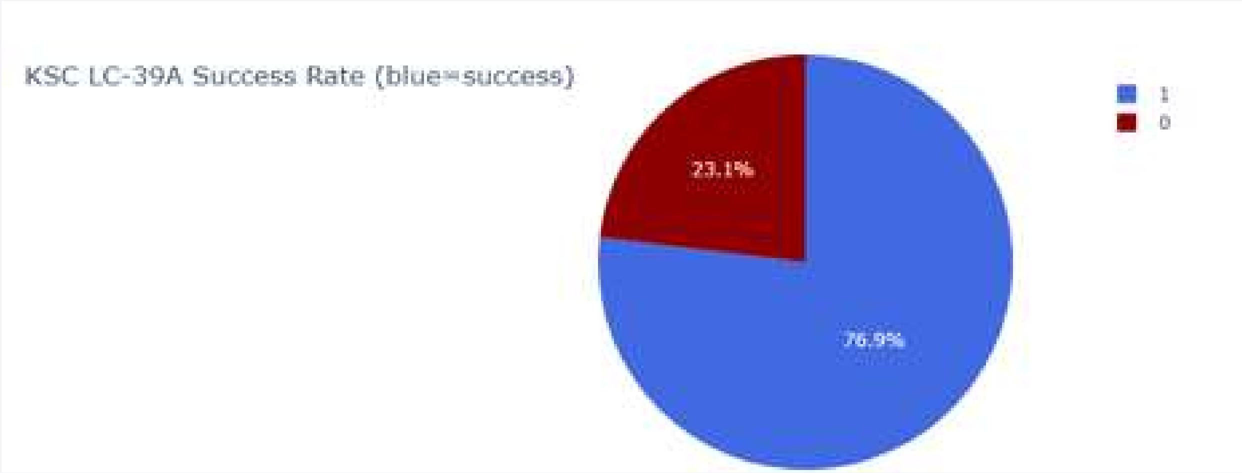
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



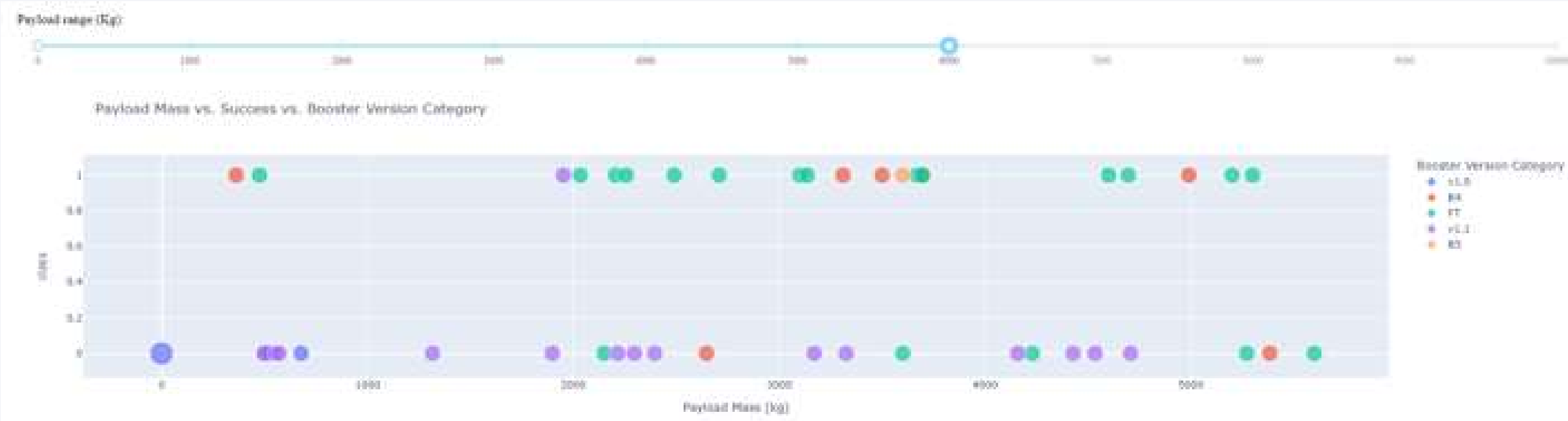
- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

Highest Success Rate Launch Site



- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Payload Mass vs Success vs Booster Version Category

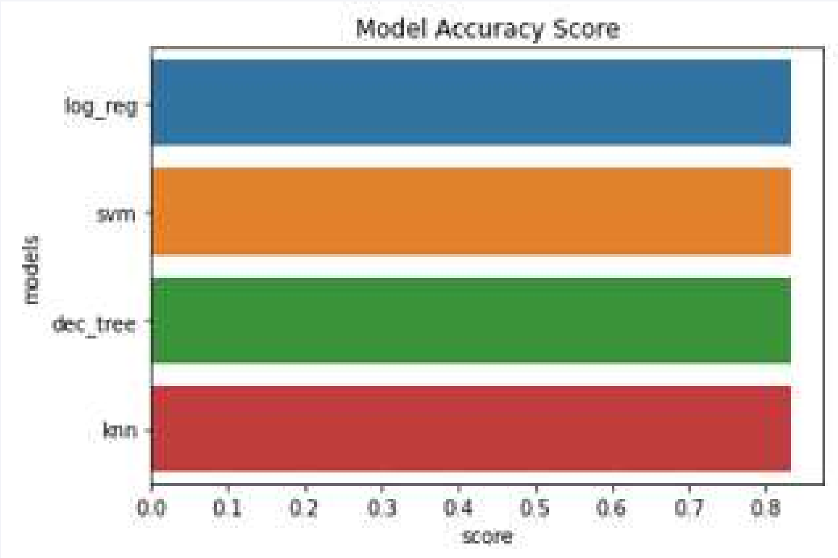


- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

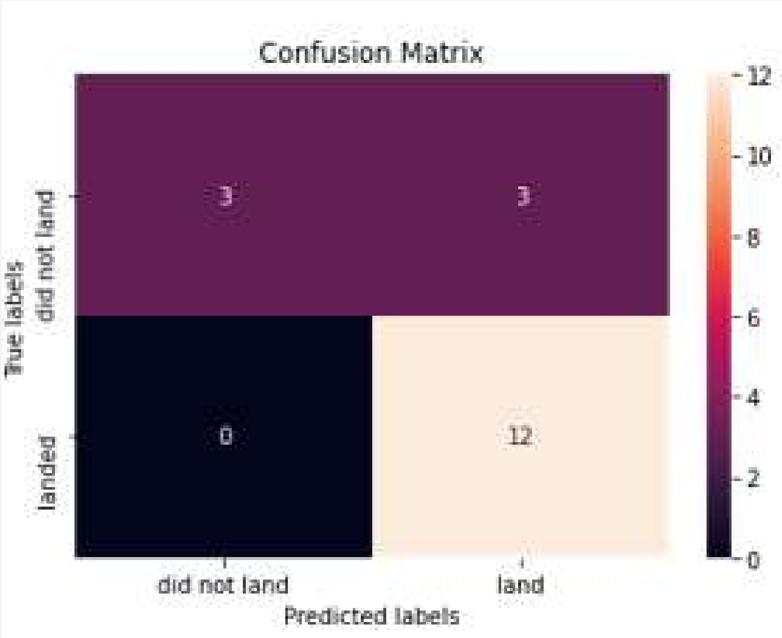


- All models had virtually the same accuracy on the test set at 83.33% accuracy.

It should be noted that test size is small at only sample size of 18.

- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix



- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

Special Thanks to All Instructors

Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo


Thank you!