DATA SCIENCE CAPSTONE PROJECT

Mitchell Xanders – August 21, 2021

OUTLINE

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

EXECUTIVE SUMMARY

- For this report, I trained machine learning models to predict what makes a successful landing of the first phase using data from our competitors at SpaceX.
  - Using their open-source API and publicly available data via Wikipedia, I determined how impactful certain launch factors were, including booster version, payload mass, and launch site.
- I found that a Decision Tree Classifier works best in correctly identifying conditions that make for successful first phase landings.

INTRODUCTION

- SpaceY would like to become a formidable competitor in the rocket science field. One way we can get a strong foot in the door is to learn from our competitors' mistakes.

- How costly is each launch?
- What can we learn from our competitors?
- Can we predict when our competitors were able to reuse the first stage of their rockets?

METHODOLOGY

- Data collection methodology:
  - REST API
  - Web Scraping
- Perform data wrangling
  - Describe how data were processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating classification models

Methodology

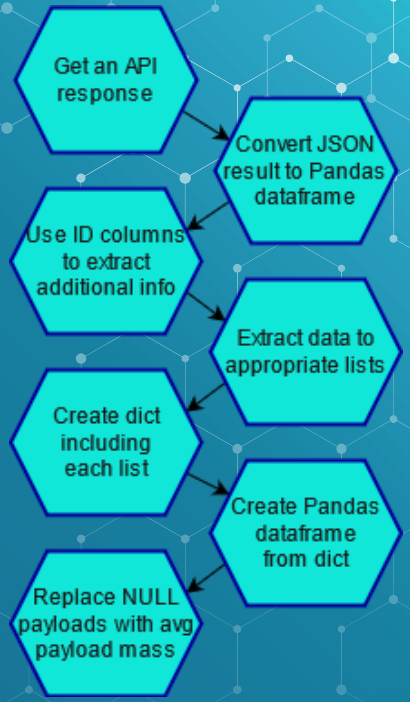
Data Collection, Visualization, and Analysis

DATA COLLECTION

- I used two methods of collecting SpaceX launch data:
  - REST API
  - Web Scraping

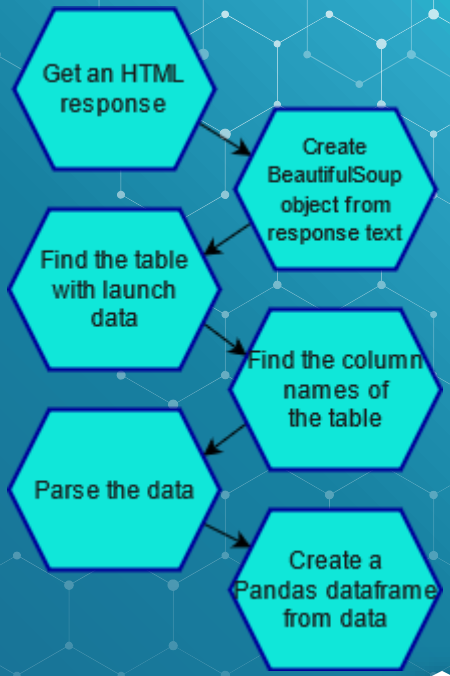
DATA COLLECTION – SPACEX API

- SpaceX has an open-source API from which I can make calls to via Python. It will return a JSON object of data depending on the specific calls I make.
- Some data returns ID numbers for certain items, like ‘rocket’, in which case I can use the ID number to make another API call to get specific information I need from that particular rocket, like ‘booster version’.
- Once I’ve narrowed down the data that is relevant, I can put it all into an easier-to-read dataframe format.
- Some values may be NULL. Substituting these values for values such as the average of the column are good practice for refinery. Another option is removing the whole row, but with a relatively small dataset as is, I thought it important to keep these rows.
- Data Collection via API Notebook on GitHub



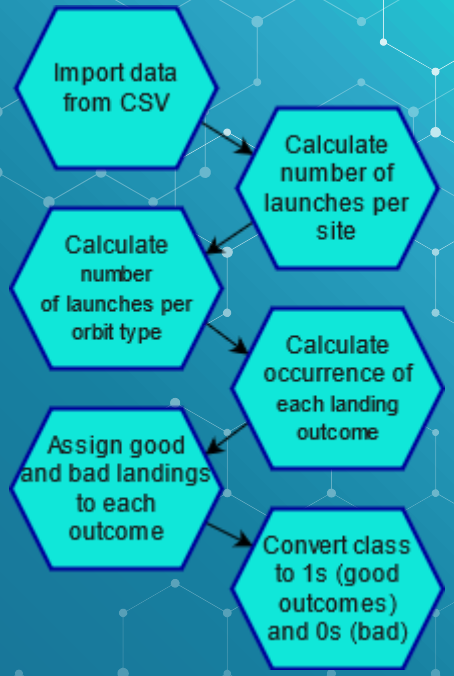
DATA COLLECTION – WEB SCRAPING

- I was able to find publicly available launch data in a nice table format via Wikipedia.
- Using the BeautifulSoup python library, it was easy to gather the data in HTML format on the Web page and put it into a clean dataframe.
- [Data Collection via Web Scraping Notebook on Github]



DATA WRANGLING

- After collecting our data, I exported it into a Comma-Separated Values (CSV) file, which allows for easy saving and accessing of organized data.
- Since the crux of the project is to predict successful and failed launches, data wrangling was necessary to convert 8 different outcomes into good (1) and bad (0) outcomes.
- This new classification system allows for easier computing of success rates, and simpler overall outcomes to predict.
- [Data Wrangling Notebook on Github]



EDA WITH DATA VISUALIZATION

- Charts and graphs are easier to comprehend quickly than SQL queries and the columns and rows of text they display.
- I used scatter plots to find the following relationships:
  - Payload Mass and Flight Number
  - Launch Site and Flight Number
  - Launch Site and Payload Mass
  - Orbit Type and Flight Number
  - Orbit Type and Payload Mass
- I also used a bar chart to compare the success rates of launches into each orbit.
- The last thing I charted was how successful SpaceX was in landing first phases over time via line chart.
- EDA with Data Visualization Notebook on Github

EDA WITH SQL

- I performed SQL queries to find the following:
  - Each launch site location
  - Every launch from a Cape Canaveral site
  - How much payload NASA Commercial Resupply Services sent in launches
  - The average payload mass carried by a F9 v1.1 booster
  - When the first successful landing via ground pad occurred
  - Which boosters that carried between 4,000 and 6,000kg of payload successfully landed via drone ship
  - How many landing succeeded and how many failed
  - Which boosters carried the maximum payload amount
  - Which boosters failed drone pad landings in 2015 and in which months
  - How many successful landings on drone ships vs. how many successful landings on ground pads between June 4, 2010 and March 20, 2017
- EDA with SQL Notebook on Github

BUILD AN INTERACTIVE MAP WITH FOLIUM

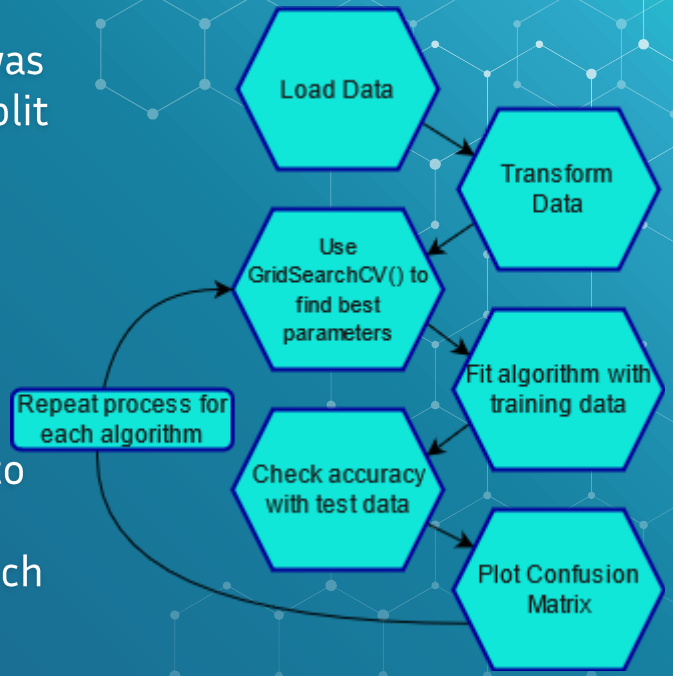
- Using Folium, I plotted locations where SpaceX launch sites were and added visual markers of each launch showing whether the landing succeeded or failed. I also chose to show the distance between the launch sites and nearby public features (ex: highways, coasts)
- My goal was to give a nice visual aid of where launch sites work. By color coding successful and failed landings at each site, we can make note of what kind of locations have shown most promising results, taking the environment around them into consideration.
- Interactive Map with Folium Notebook on Github

BUILD A DASHBOARD WITH PLOTLY DASH

- On the Dashboard I wanted to include the following:
  - A pie chart that showed rates of success among each and all sites.
  - A scatter plot of success rates among booster versions and how much payload they were carrying.
  - A slider to allow for payload ranges to be emphasized/zoomed in on.
- Having interactive visuals like these makes it a lot easier to digest important information such as which boosters, what payloads, and which sites resulted in successful landings.
- Plotly Dash Lab on Github

PREDICTIVE ANALYSIS (CLASSIFICATION)

- When building the models, the first order of business was to load and transform the data I had recorded. Then I split the data into training and testing sets for each of the algorithms to use. Using GridSearchCV(), I found parameters of each algorithm that maximized their respective scores before fitting the dataset to the algorithm for training.
- In evaluating each model, I checked the accuracy score with the training data, and plotted a Confusion Matrix to show True and False Positives and Negatives.
- Once accuracy scores were compared, I determined which algorithm had the highest success rate.
- [Machine Learning Prediction Notebook on Github]



RESULTS

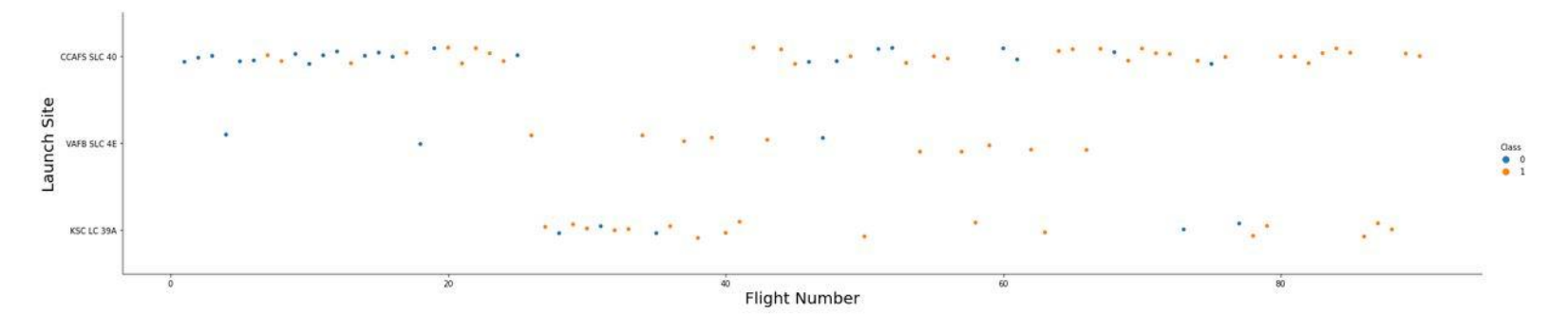
- Exploratory Data Analysis Results
- Interactive Analytics Demo (Screenshots)
- Predictive Analysis Results

EDA WITH VISUALIZATION

Data Analysis with Charts and Graphs

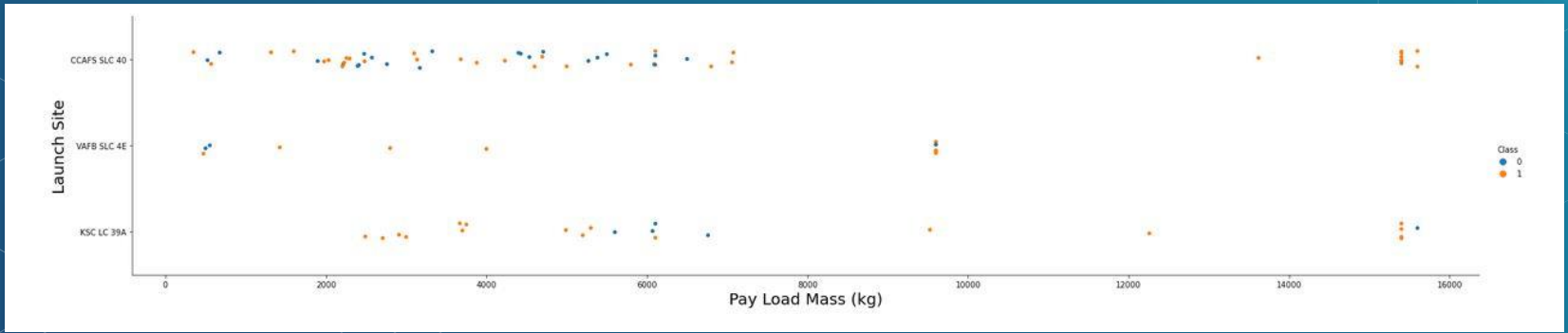
FLIGHT NUMBER vs. LAUNCH SITE

- Through early analysis, I found that rockets launched from the KSC LC 39A had an exceptional rate of success in landing the Falcon 9 first stage considering the number of flights.
- As expected, through trial and error, each site was able to more consistently get successful landings as they launched later rockets. As of today, each site has successfully landed at least 5 flights in a row.



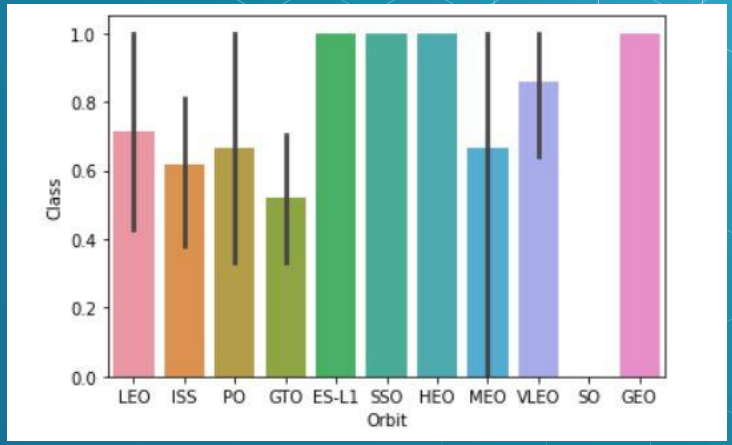
PAYLOAD vs. LAUNCH SITE

- Through early analysis, I found that Falcon 9 launches with heavy payloads fared exceedingly well in landing successfully, with a high success rate among payloads greater than 10,000kg.
- The range between 4,000 and 6,000 kg appears to be the most concentrated with flights and, accordingly, with failed landings.



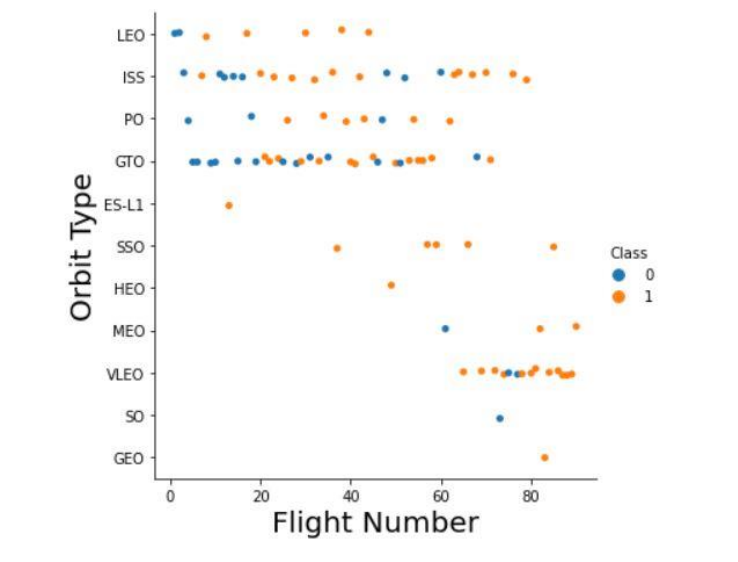
SUCCESS RATE vs. ORBIT TYPE

- The Falcon 9 found success in phase one landings from ES-L1, SSO, HEO, and GEO orbits, but none with SO orbits.
- One thing of note is that the orbits above plus MEO orbit each have a sample size smaller than n=4.
- Of orbits with a sizeable sample size, VLEO (14 launches) found the highest success rate (.857).



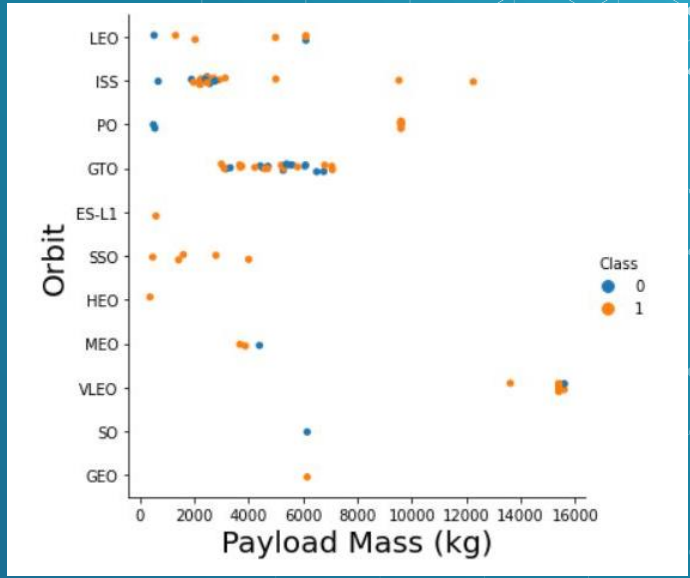
FLIGHT NUMBER vs. ORBIT TYPE

- As Falcon 9 launches have gone on, there has been an increased focus on VLEO orbits, which has resulted in a higher first phase landing success rate.
- GTO and ISS orbits have the most flights, with mixed landing results. ISS orbits lately have been much more consistently successful, though.
- This chart more clearly demonstrates the sample size discrepancies noted in the previous slide.



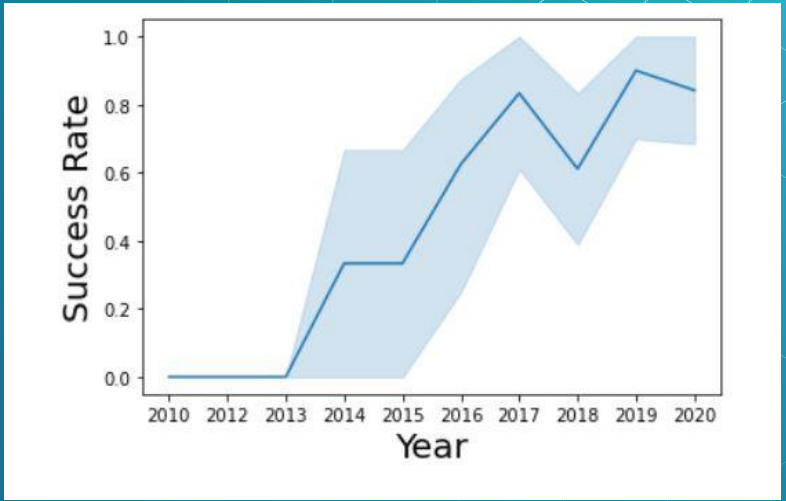
PAYLOAD vs. ORBIT TYPE

- Most of the successful landings from launches with heavy payloads came from VLEO orbit launches.
- Most of the ISS orbit launches had payloads between 2000 and 4000kg and were mostly successfully landings.
- The GTO orbit launches all fall between 2000 and 8000kg – again, with varying levels of success.



LAUNCH SUCCESS YEARLY TREND

- As a fledgling company, there was little to no success.
- Success rates started to increase form 2013 until 2017.
- After a dip in 2018, success rate was at its peak in 2019 before dropping again in 2020.



EDA WITH SQL

Data Analysis with Database Queries

ALL LAUNCH SITE NAMES

- SpaceX launched Falcon 9 rockets from 4 distinct locations.

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

LAUNCH SITE NAMES BEGINNING WITH ‘CCA’

- There have been 60 launches at sites beginning with ‘CCA’ (Cape Canaveral). These are the first 5 records of them.

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

TOTAL PAYLOAD MASS

- By adding up every entry launched by NASA Commercial Resupply Services, I found that 45,596kg of payload have been launched in those Falcon 9 missions.

NASA_Payload
45596

AVERAGE PAYLOAD MASS BY F9 v1.1

- By averaging out the payload masses of every entry with a F9 v1.1 booster, I found that launches conducted with F9 v1.1 boosters carried an average of 2,928kg in payload.

FIRST SUCCESSFUL GROUND LANDING DATE

- In querying the database, I found that the first successful landing outcome didn't happen until 2015 – 5 years after their first launch attempt.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 600

- These were the boosters that found success in drone ship landing within the 4,000 – 6,000kg payload threshold.
- If you'll recall, early data analysis showed this payload range was the most concentrated with failed landings.

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

TOTAL NUMBER OF SUCCESSFUL AND FAILED MISSION OUTCOMES

- Despite many landing failures, I found that the mission for each launch almost always turned out to be a success were it not for one in-flight failure.

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

BOOSTERS CARRIED MAXIMUM PAYLOAD

- These are the boosters that were tasked with carrying the maximum payload of 15,600kg.

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

- 2015 saw two failed landings at the CCAFS LC-40 site: one in January, and one in April.

MONTH	booster_version	launch_site	landing__outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

- From June 4, 2010 to March 20, 2017, the Falcon 9 saw 8 successful landings – 5 on drone ship and 3 on ground pad.

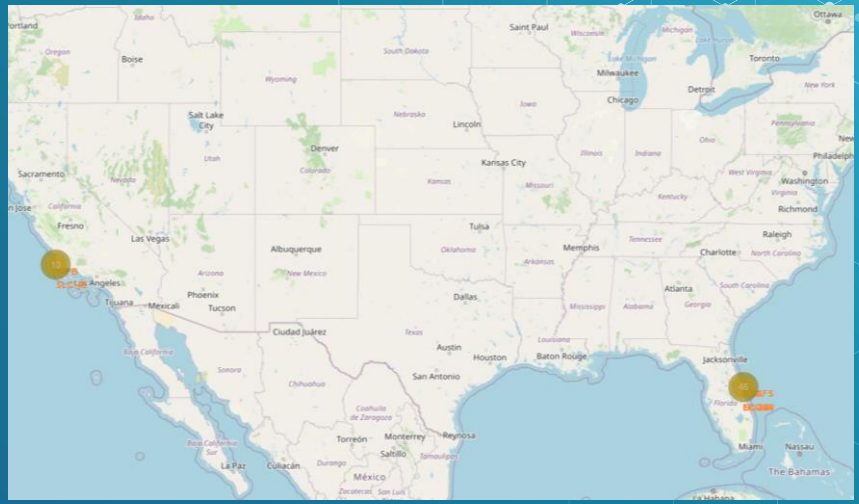
landing__outcome	COUNT
Success (drone ship)	5
Success (ground pad)	3

INTERACTIVE MAP WITH FOLIUM

Interactive Analytics with Maps

SPACEX LAUNCH LOCATIONS

- SpaceX's launch locations can be found along the coasts of CA and FL.
- Sensible reasons for these spots are warmer climate, closeness to water (for landing pad), and keeping distance from main highways.

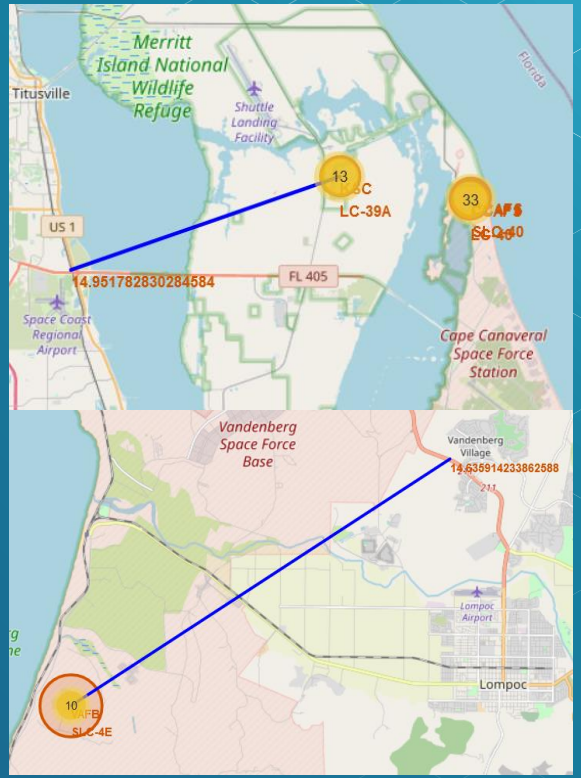


SUCCESS RATE AT LAUNCH SITES

- Here you can see the success rate at each site mapped out.
- Notably, the KSC LC-39A has the highest success rate, the only site eclipsing 50% success.

PROXIMITY TO LOCAL FEATURES

- As you can see from the map snippets, SpaceX prefers to keep their launch sites about 15km from prominent features.
- The KSC LC-39A site is approximately 15km from the mainland FL coast.
- The VAFB SLC-4E launch site is approximately 15km from the nearest stretch of public roadway.



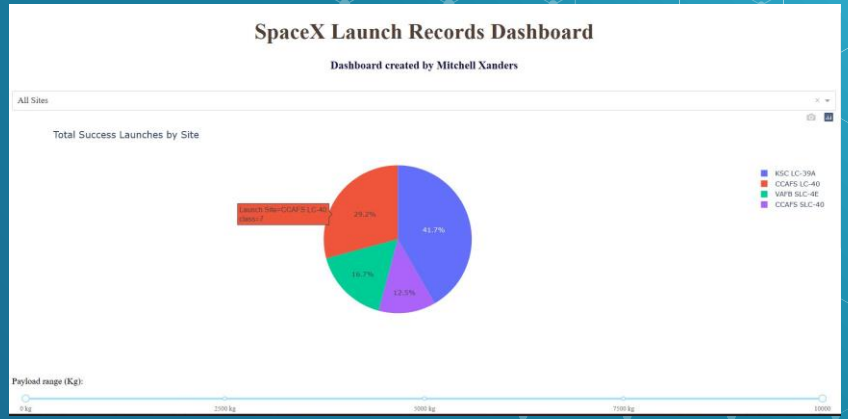
DASHBOARD WITH PLOTLY DASH

Interactive Analytics with Dashboard

SUCCESSES AT LAUNCH SITES

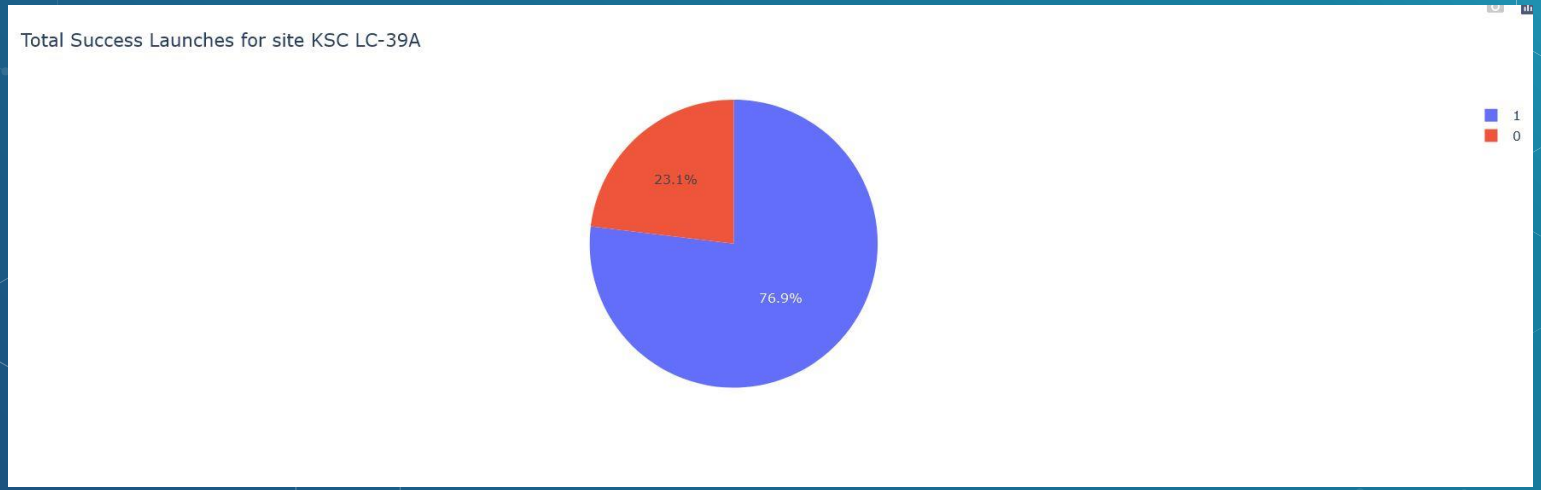
- The dashboard shows that the KSC LC-39A launch site (41.7%) and the CCAFS LC-40 site (29.2%) are responsible for 70.9% of successful Falcon 9 landings.

- The third FL site (CCAFS SLC-40) was responsible for the least number of successful landings (12.5%).



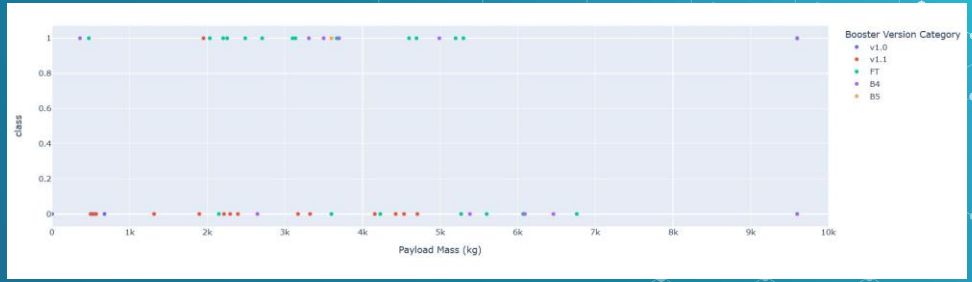
MOST SUCCESSFUL SITE

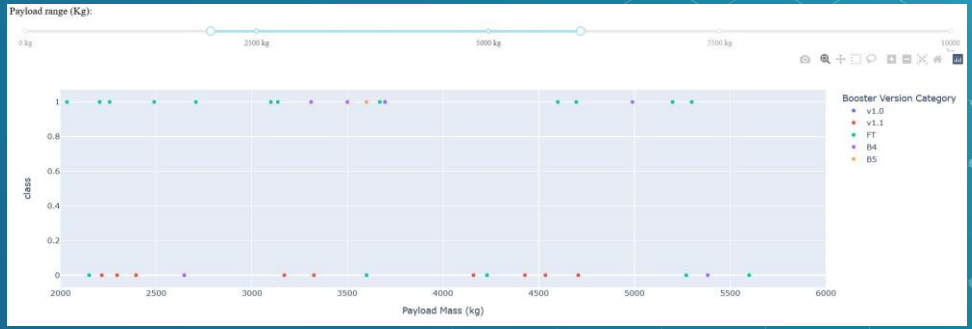
- Not only did the KSC LC-39A launch site see the most successful launches, it also had the highest success rate at 76.9%.



PAYLOAD vs. LAUNCH OUTCOME

- In the top image, you can see the how successful launches were at different payload amounts.
- In the bottom image, we zoom in on the range of 2000-6000kg, where most of the data points are.
- From this zoomed in plot, we can see the FT booster saw a positive success rate, while the v1.1 booster saw no success.



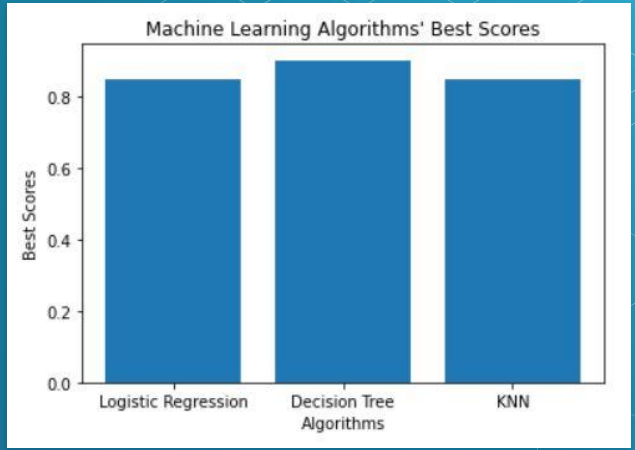


PREDICTIVE ANALYSIS (CLASSIFICATION)

Machine Learning Predictions & Analysis

CLASSIFICATION ACCURACY

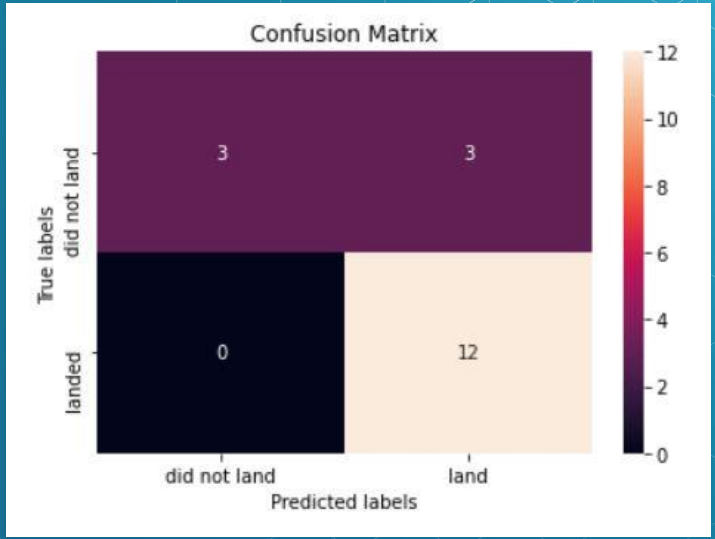
- I tried using three machine learning algorithms to best predict the success or failure of phase one landings: Logistic Regression, Decision Tree, and K-Nearest Neighbor (KNN).
- While they were all close, the Decision Tree had the highest accuracy score when using the training data at just over 90%.



Algorithm	Best Score
Logistic Regression	0.846429
Decision Tree	0.901786
KNN	0.848214

CONFUSION MATRIX

- This confusion matrix shows the accuracy of the Decision Tree model with the testing data.
- We can see it correctly predicted 12 successful landings and 3 failed landings.
- The top-right region shows it unsuccessfully predicted 3 first phases would land successfully when they actually did not.



CONCLUSION

- ISS orbit launches with lighter payloads (2,000 – 4,000kg) found success landing.
- Launches with heavier payloads (>10,000kg) found great success landing from VLEO orbits.
- The KSC LC-39A launch site in Florida found the most success in landing first phases.
- A Decision Tree Classifier will be the most accurate predictor of successes/failures in landing first phases.
- As we get more experience launching rockets, our successful landings will increase, as will our predicting algorithms.

APPENDIX

- Complete Github repository for this project can be found here.