IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stacy Dugan, SpaceY Sr. Decision Scientist
December 23, 2023

Outline

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

Executive Summary

- Business Case: To become a viable competitor in the Commercial Space Age we must be able to successfully land our first stage boosters, and thereby predict and control costs.
- Purpose of study: To predict if the Falcon 9 first stage will land successfully.
- Methodology: Freely available SpaceX data was collected from various sources on the web, exploratory data analysis and visualization tools were used to determine the most relevant features to including in our models, 4 machine learning algorithms were designed and compared. The best algorithm was selected based on accuracy, business application and ease of use.
- Results: A logistic regression model was selected as the best predictor algorithm. SpaceY is now in a strong position to bid on future launch missions.

Introduction - Project background and context

- The commercial space age has arrived as companies compete for the opportunity to launch satellites and rockets into space.

- SpaceX is the current space age industry leader, accomplishments include:
  - Sending spacecraft to the International Space Station.
  - Providing satellite Internet access via Starlink
  - Sending manned missions to Space.

- A driving factor in SpaceX's success is their ability to **reuse the first stage** of the rocket (see Appendix B- Anatomy of a Rocket)
  - This reuse decreases the cost of a launch by upwards of 40% saving over $100M in some cases.

Introduction – Problem to be Solved

- Space Y, founded by Billionaire Industrialize Allon Musk, would like to compete with SpaceX
- To be a viable competitor, we must be able to identify if a rocket’s first stage will land successfully (i.e. be available for reuse). This includes:
  - Identifying the dominant features of a launch to maximize a successful first stage landing
  - Estimating the cost of a launch in order to bid against SpaceX
- We've developed a machine learning model using publicly available data from SpaceX's Falcon 9 rockets
- This model aims to identify the optimal combination of features for maximizing the reusability potential of the first stage, thereby controlling costs

Section 1

Methodology

See Appendix B - Anatomy of a Rocket

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API/Get requests/JSON parser
  - Webscraping Falcon 9 Wiki pages using Beautiful Soup
- Perform data wrangling to better explore and understand the datasets
  - Determine the target variable, Y – landing success/failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using scikit-Learn Toolkit grid search cross validation

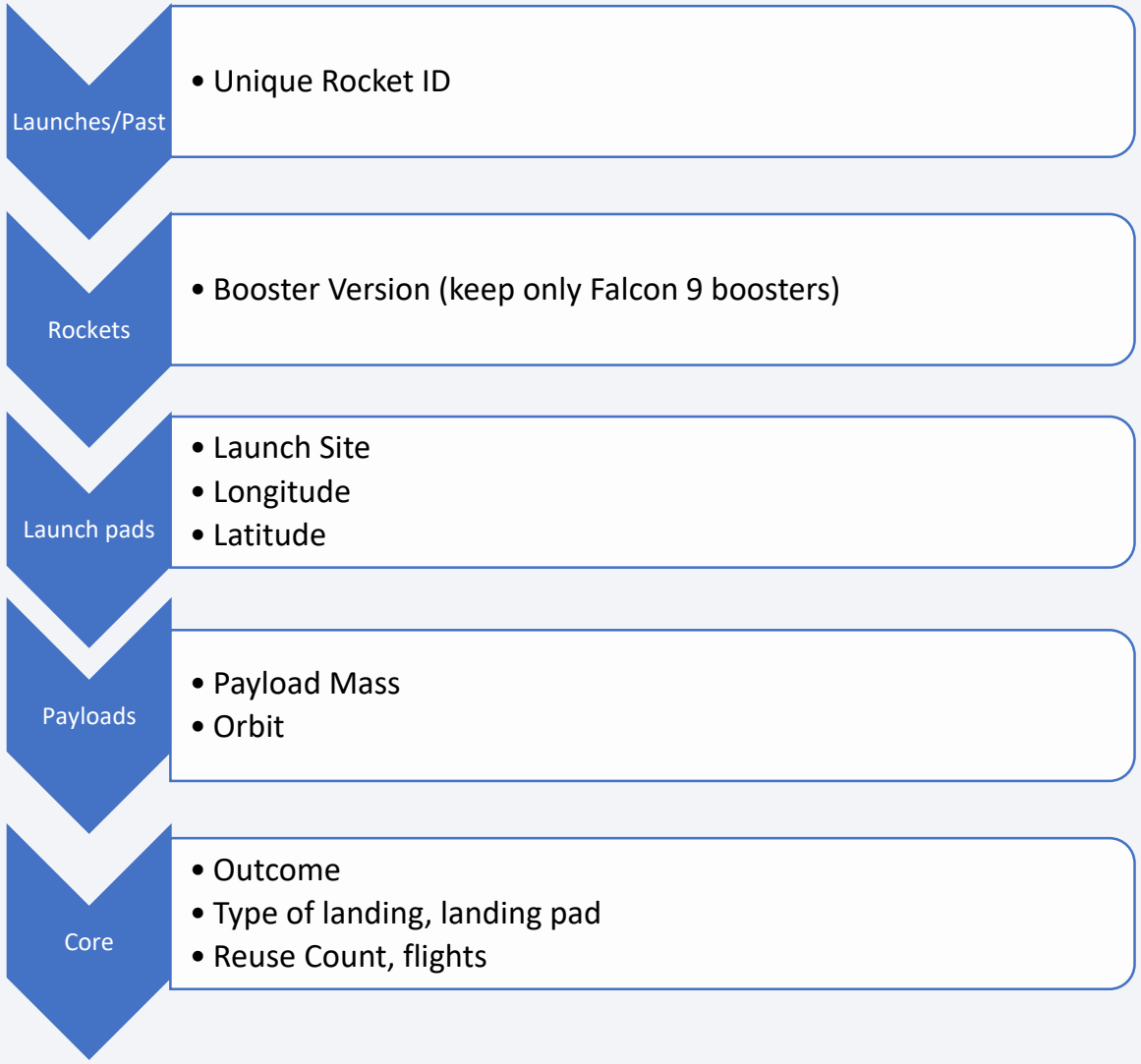
Data Collection – SpaceX launch data

- Data was collected from:
  - SpaceX REST API (api.spacexdata.com/v4/)
    - Includes: rocket used, payload delivered, launch specifications, landing specifications, and landing outcome
    - Get request performed which returns a list of JSON objects, each representing a launch
  - Wiki Page: “List of Falcon 9 and Falcon Heavy Launches”

    (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
    - Includes: Flight Number, Booster Version, Launch site, Payload, Payload Mass, Orbit, Launch outcome, Booster landing, Customer, and Date
    - Python BeautifulSoup package used to web scrape the Wiki HTML

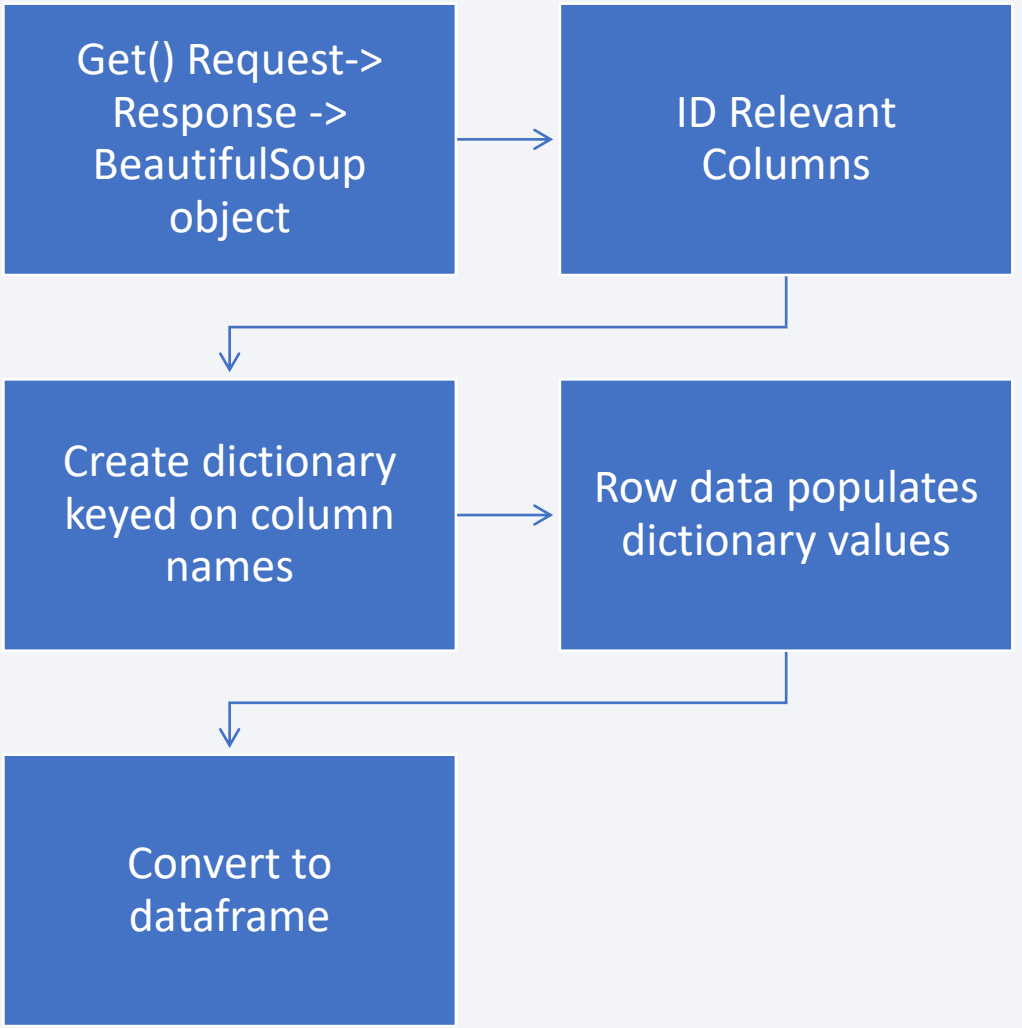
Data Collection – SpaceX API

- Request data from the following SpaceX API endpoints:
  - Launches/past, rockets, launchpads, payloads, core
- Store data as List -> Dictionary -> Pandas dataframe



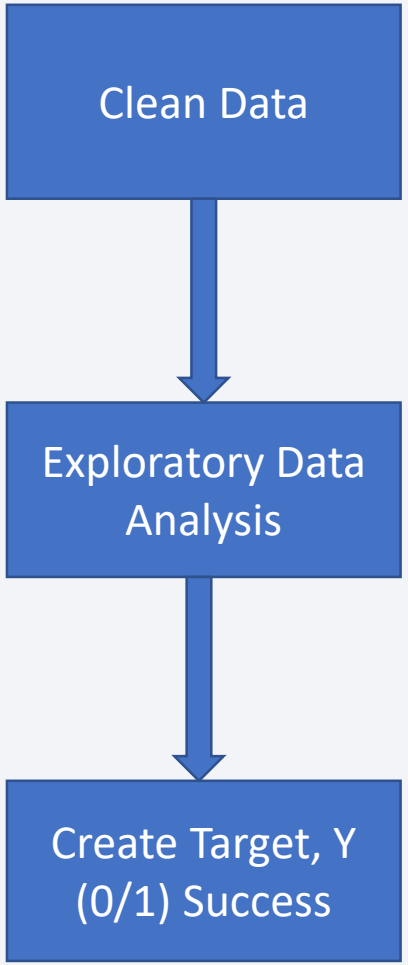
Data Collection – Web Scraping Wiki

- Request the Falcon 9 Launch Wiki page from its URL
- Create BeautifulSoup object from HTML Response
- Locate the launch records table and extract all relevant column names from the header
- Create an empty dictionary from the column names
- Read in table rows and populate the dictionary values
- Convert dictionary to a Pandas data frame



Data Wrangling

- Clean data
  - Remove null values
  - Remove irrelevant features
  - Remove rockets with specialized boosters and extra payloads
- Exploratory Data Analysis (EDA) to find some patterns in the data
  - Number of launches per launch site
  - Number and occurrence of each orbit type
  - Landing outcome total counts, by flight, and per orbit type
  - Success rate of all previous Falcon 9 landings
- Create classification variable, Y, representing the 0/1 outcome of each first stage landing
  - Derived from the 8 Landing Outcome categories
  - See Appendix C for more details on Landing Outcomes



EDA with SQL

- SQL queries performed (using SQLLite):
  - Display the names of the unique launch sites
  - Display 5 records where the flights begin with “CCA” (Cape Canaveral)
  - Compute the total payload mass carried by boosters launched by NASA (CRS)
    - NASA's Commercial Resupply Services (CRS) is a program managed by NASA that contracts commercial companies to transport cargo and supplies to the International Space Station (ISS). It's a key part of NASA's strategy to involve private companies in space exploration and support the needs of the ISS.
  - Compute the average payload mass carried by booster version F9 v1.1
    - Falcon 9 version 1.1
  - Find the date of the first successful ground pad landing was achieved

EDA with SQL

- SQL queries performed (using SQLLite), continued:
  - Display the names of the boosters which successfully landed on a drone ship and have payload mass between (4000, 6000)
  - Compute the total number of successful and failed mission outcomes
  - List the names of the boosters which have carried the maximum payload mass.
  - List months, booster versions,k and launch site for the year 2015 which had a failed drone ship landing.
  - Display 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.

Build an Interactive Map with Folium

- Goal: to identify a set of factors that might contribute to the optimal selection of a launch site.
- The landing success rate depends on many factors, including the location and proximities of a launch site.
- Highlighted success vs failed landings using green/red markers
- Highlighted proximity of launch sites to various geographical and urban features by
  - Computing the distance between the two locations
  - Drawing a poly line to connect the two locations
- Features of interest included coastlines, railroads, highways, major cities, etc.

Build a Dashboard with Plotly Dash

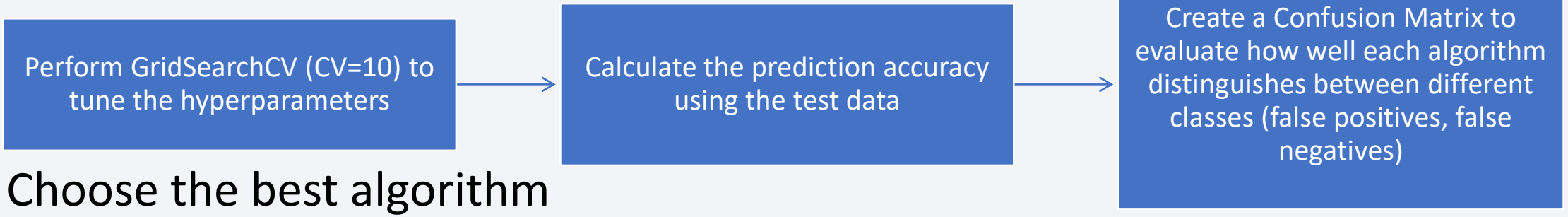
- Created a SpaceX Launch Records Dashboard which includes:
  - Pie Chart of 1st Stage Booster Landings
    - Default: displays the proportion of successful landings among all launch sites
    - Drill down into each launch site to show the ratio of landing success/failure
  - Scatterplot of Payload vs. Success/Failure Category by Booster Type
    - Default: displays the all payload mass values in the dataset and all launch sites
    - Drill down into each launch site to show the different booster types, payloads used, and their landing outcomes
- This dashboard provides insight into the success rates at each launch site
- It further provides insight into the affect of payload mass and booster type on successful landings

Predictive Analysis (Classification)

- Scikit-learn Toolkit used
- 4 Algorithms Created and Results Compared
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Classification Tree
  - K-Nearest Neighbor (KNN)
- Best Algorithm Selected

Predictive Analysis (Classification)

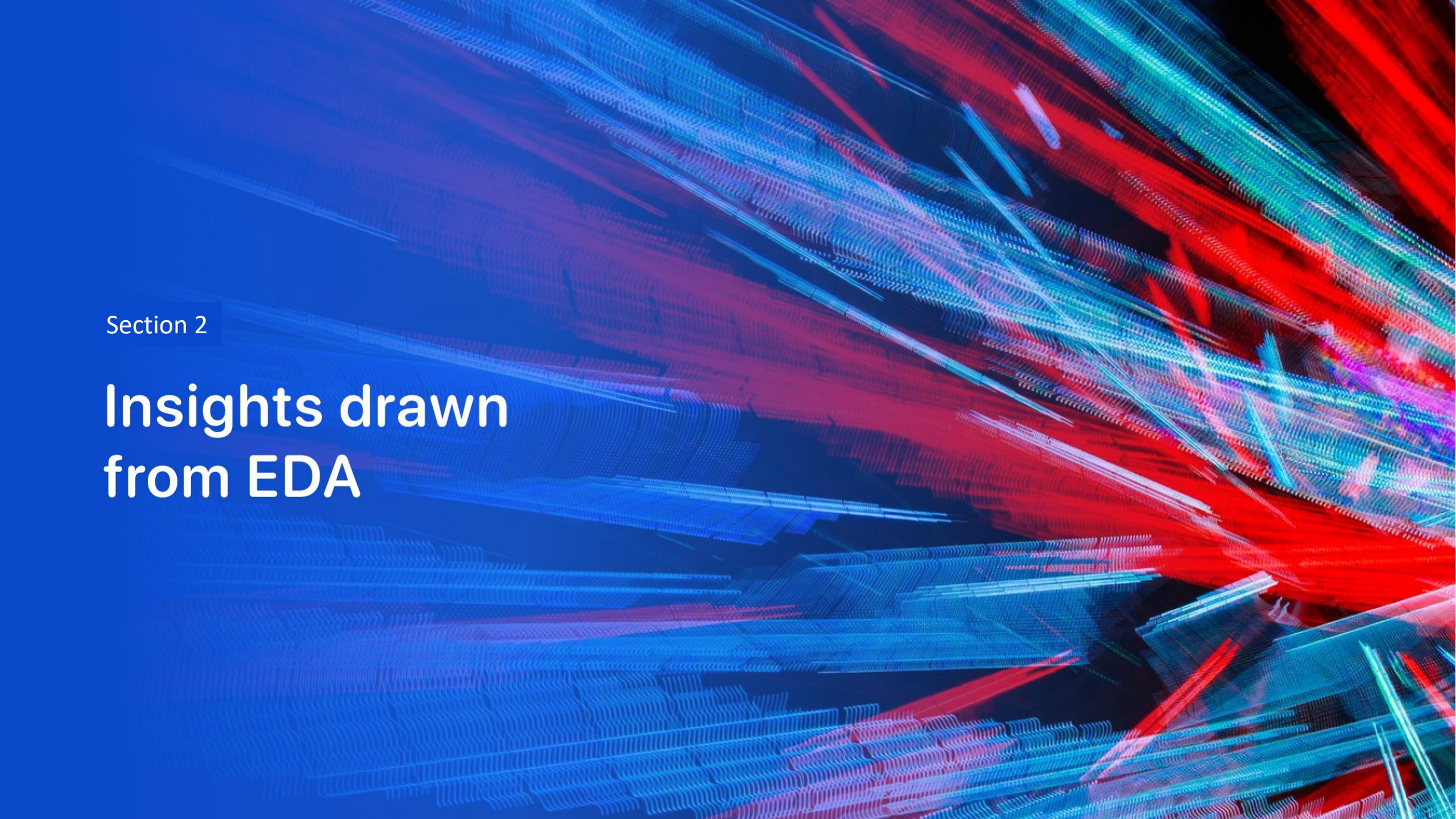
- Input SpaceX dataset
- Create Target Variable, Y, as 0/1 : Failure/Success in landing outcome
- Standardize the Feature Variables, X, using the default StandardScalar, fit_transform function
- Split the dataset into Training/Test sets in 80:20 ratio
- For each of the 4 algorithms:



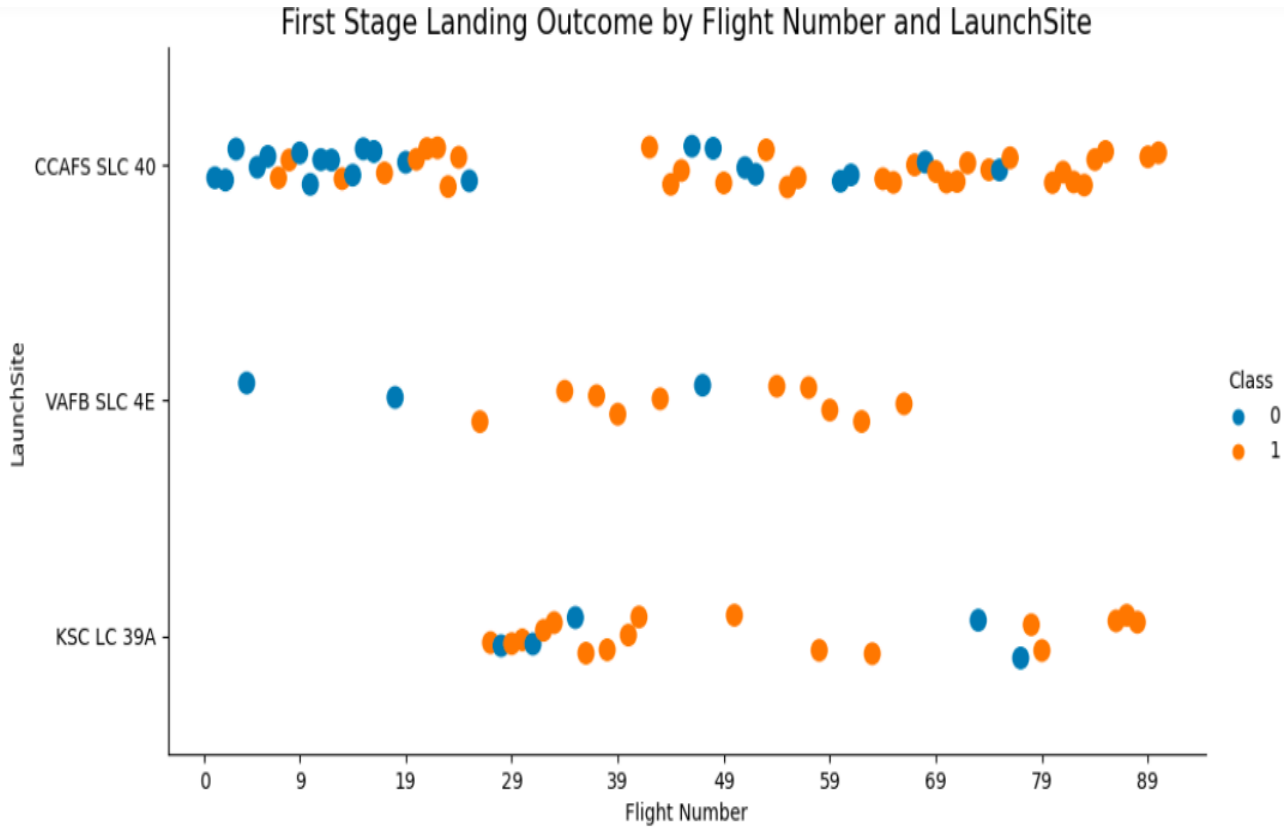
- Choose the best algorithm

Results

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

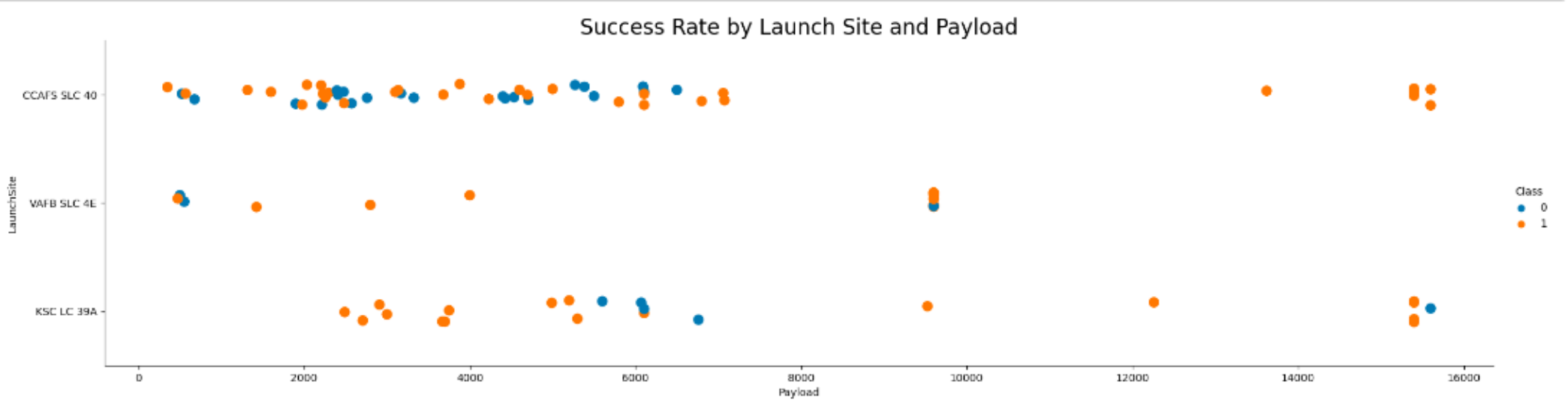

Section 2
Insights drawn from EDA

EDA with Data Visualization



- Different launch sites have different success rates.
- Overall:
  - CAFS SLC-40 = 60%
  - KSC LC-39A =77%
  - VAFB SLC 4E = 77%.
- Note, flight number is chronological. Thus, the larger the flight number, the more recent the launch.
- For flight numbers >=25:
  - CCAFS SLC 40 = 75%
  - KSC LC 39A = 77%
  - VAFB SLC 4E = 91%
- It appears CCAFS SLC 40 (Cape Canaveral) learned from its early mistakes and improved its success rate over time.

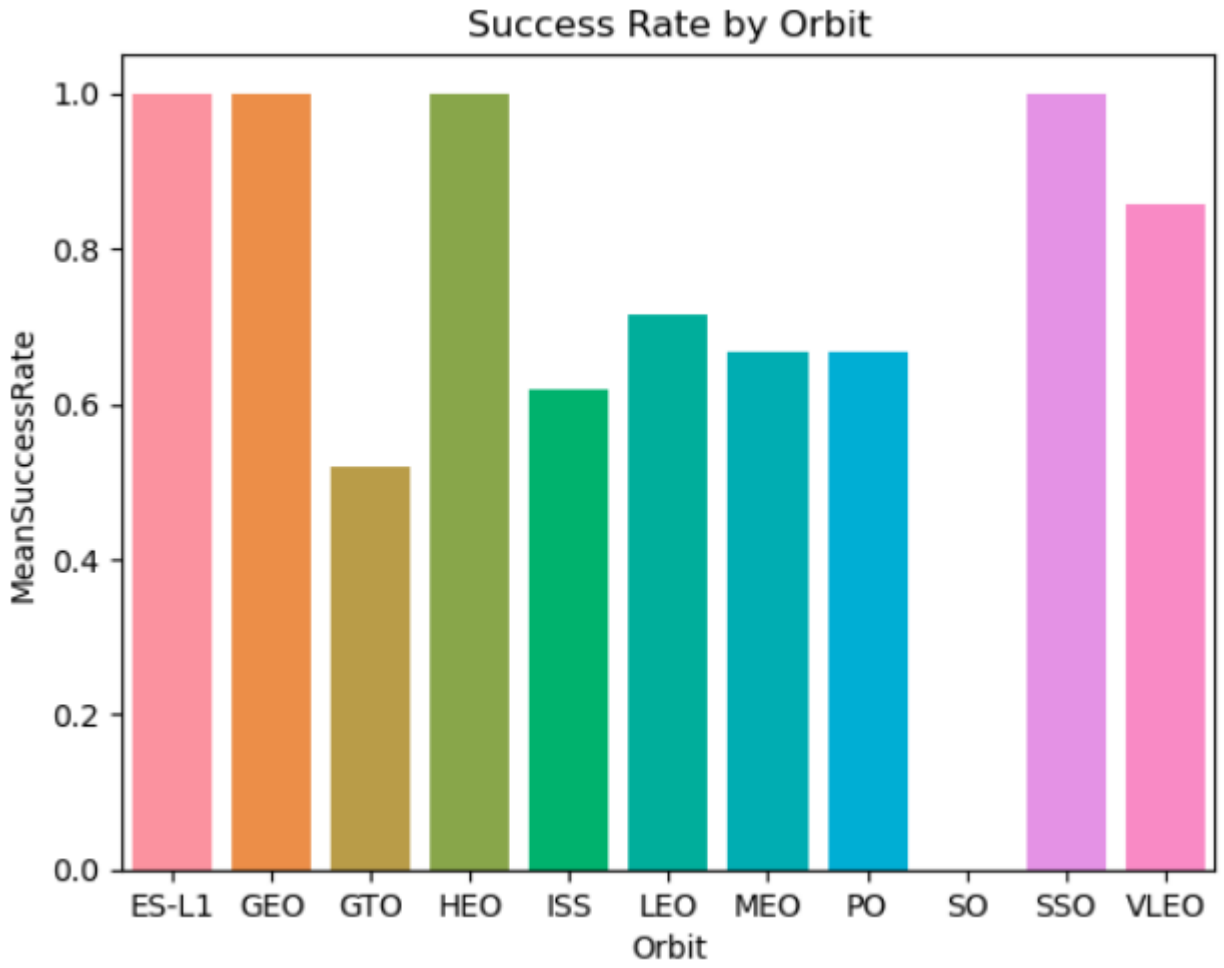
EDA with Data Visualization



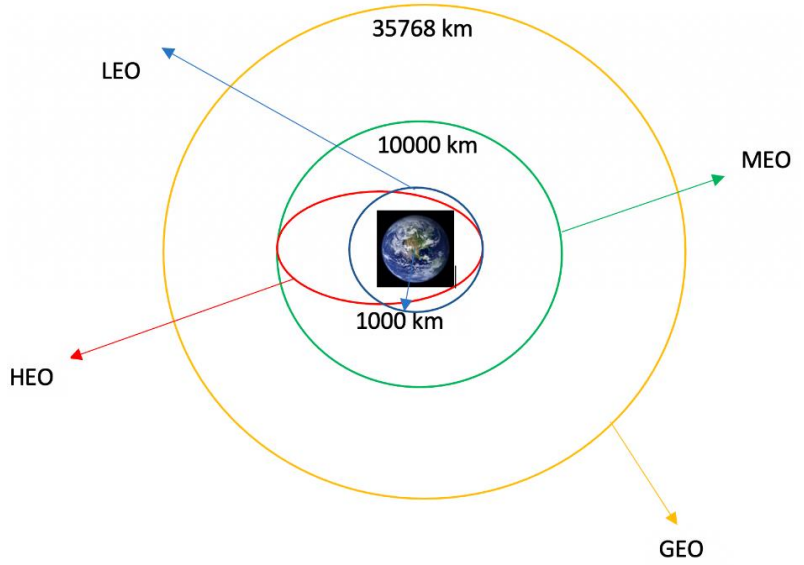
Observations:

- No rockets with payload mass > 10,000 are launched from VAFB-SLC (Vanderberg Air Force Base in CA.
- It is rare to have a payload mass > 10,000 but these seem to have the highest success rate

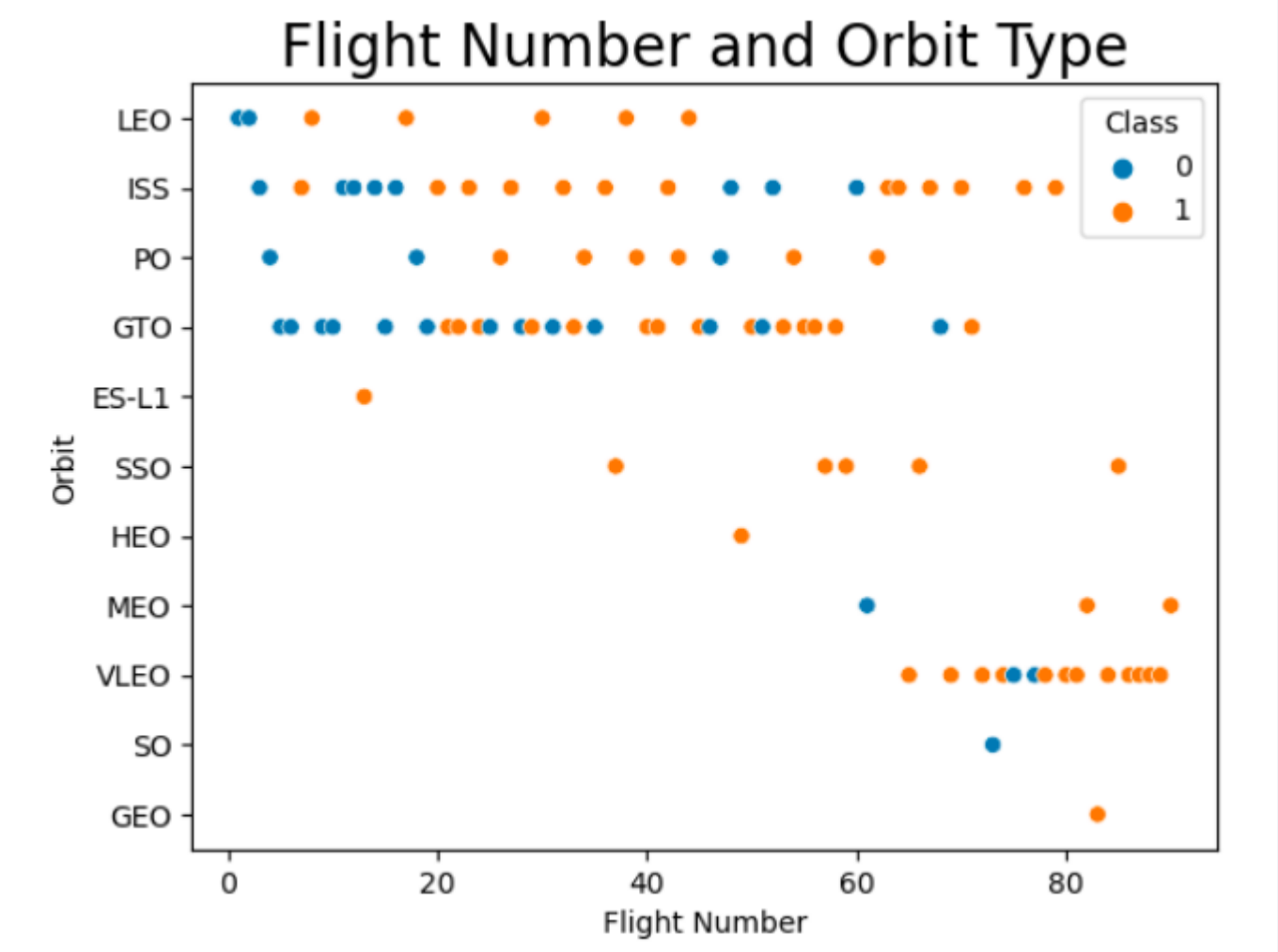
EDA with Data Visualization



Some of the orbits plotted:

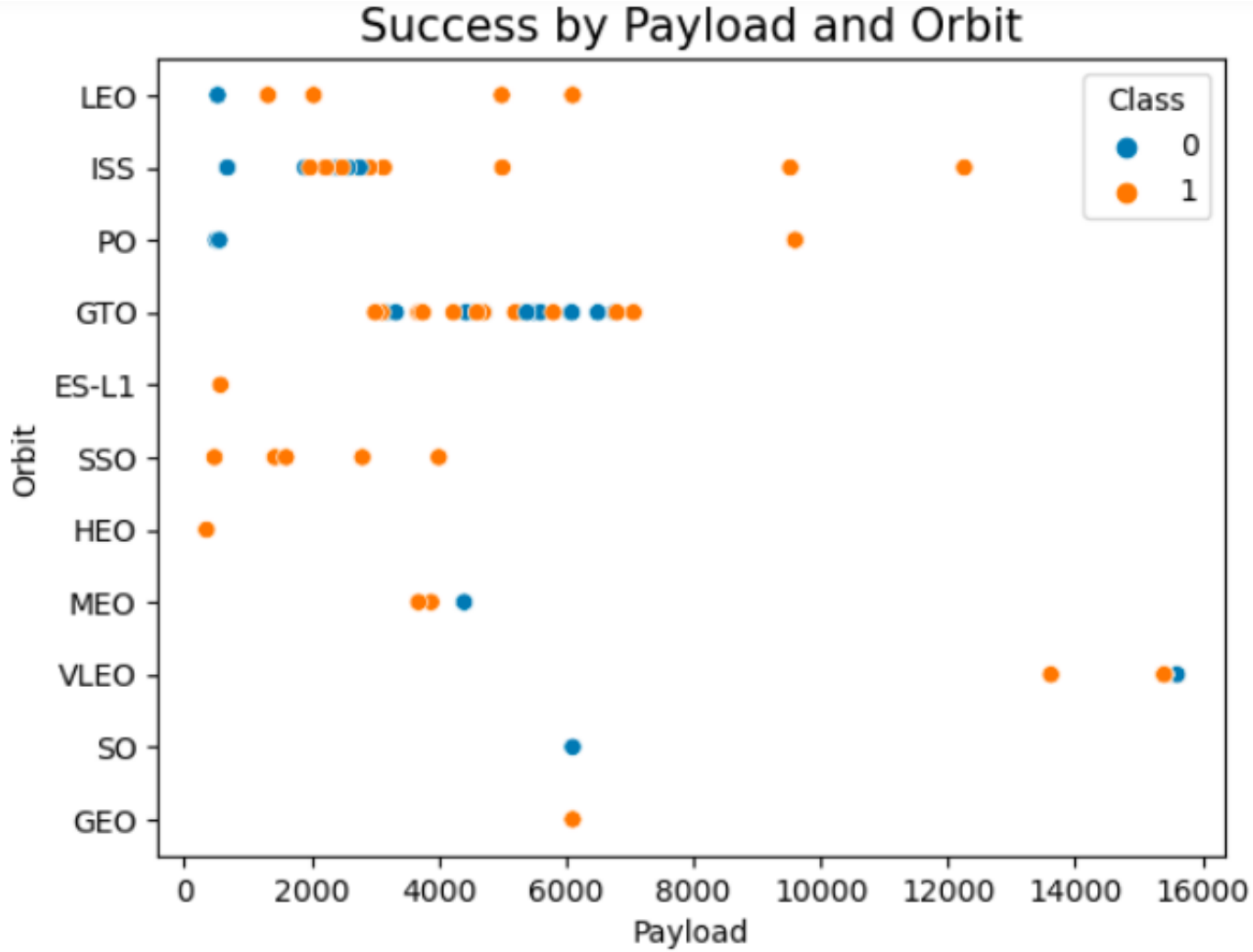


Flight Number vs. Orbit Type



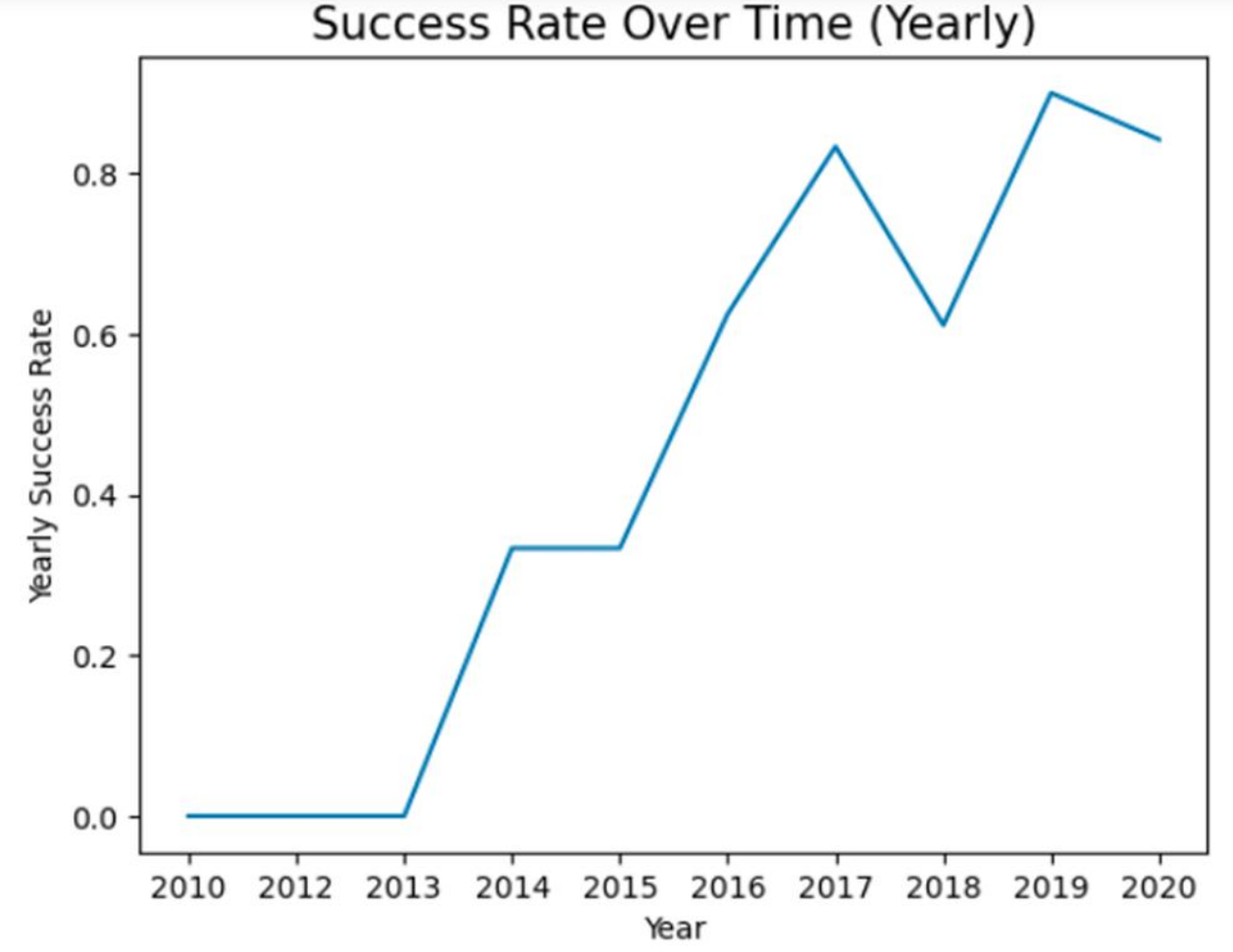
- Some observations:
  - LEO Orbit: success appears to be related to the flight number
  - The most common orbits are ISS and GTO for which there seems to be no relationship between flight number and success

EDA with Data Visualization



- Polar, ISS, and VLEO Orbits seem to be the only ones that are used for heavy (> 10,000) payloads.

Launch Success Yearly Trend



Observations:

- As one would expect, the overall trend in landing success increases over time
- Something happened in 2018 to cause a dip in the overall upward trend in success rate. Further investigation to explain this anomaly is needed.

All Launch Site Names

- The names of the unique launch sites in the space mission:
  - CCAFS SLC-40 and CCAFS LC-40
    - Note: In this exercise, I re-named all to CCAFS SLC-40 as LC-40 is the same site location, it is just a historical name
  - VAFB SLC-4E
  - KSC LC-39A

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

- This table shows 5 records from the SPACEX table where launch sites begin with `CCA` (i.e. Cape Canaveral)
- There are two outcome categories: Mission Outcome and Landing Outcome.
  - The Mission Outcome is a success in each case
  - The Landing Outcome, refers to just the successful landing of the 1st stage booster. In this example there were two failures to land (using parachute) and three in which no attempt was made to land the booster

Total Payload Mass

- The total payload mass carried by boosters launched by NASA is approximately 100,000 Kg

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,535 Kg

First Successful Ground Landing Date

- The date the first successful ground pad landing was achieved on Dec 12, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on a drone ship and have a payload mass between (4000, 6000)kg

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

Total Number of Successful and Failure Mission Outcomes

- Successful = 61
- Failed = 10

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass:

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

- Boosters, and their corresponding launch sites, which failed to land on a drone ship by month in 2015

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

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

- The occurrence of each of the 8 different landing outcomes between the date 2010-06-04 and 2017-03-20

Landing_Outcome	Cnts
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

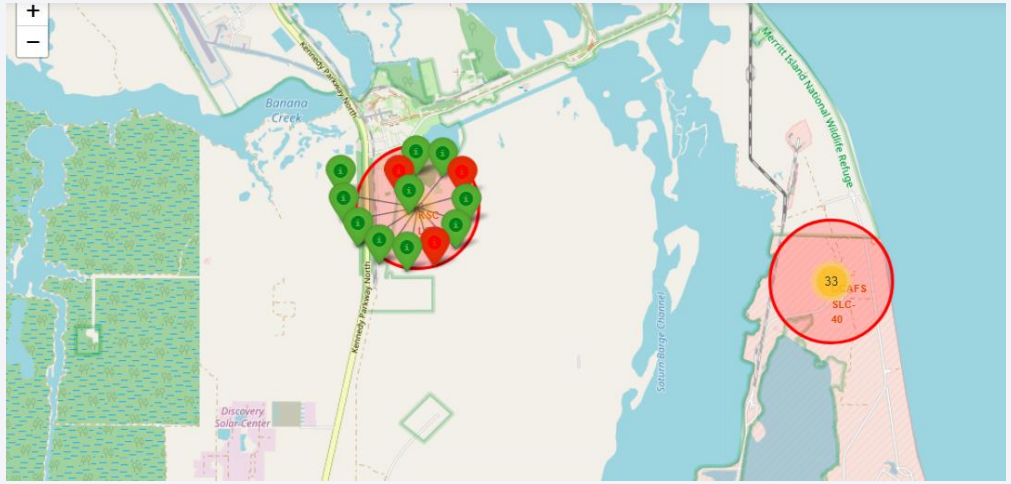
Section 3

Launch Sites Proximities Analysis

US Launch Site Locations

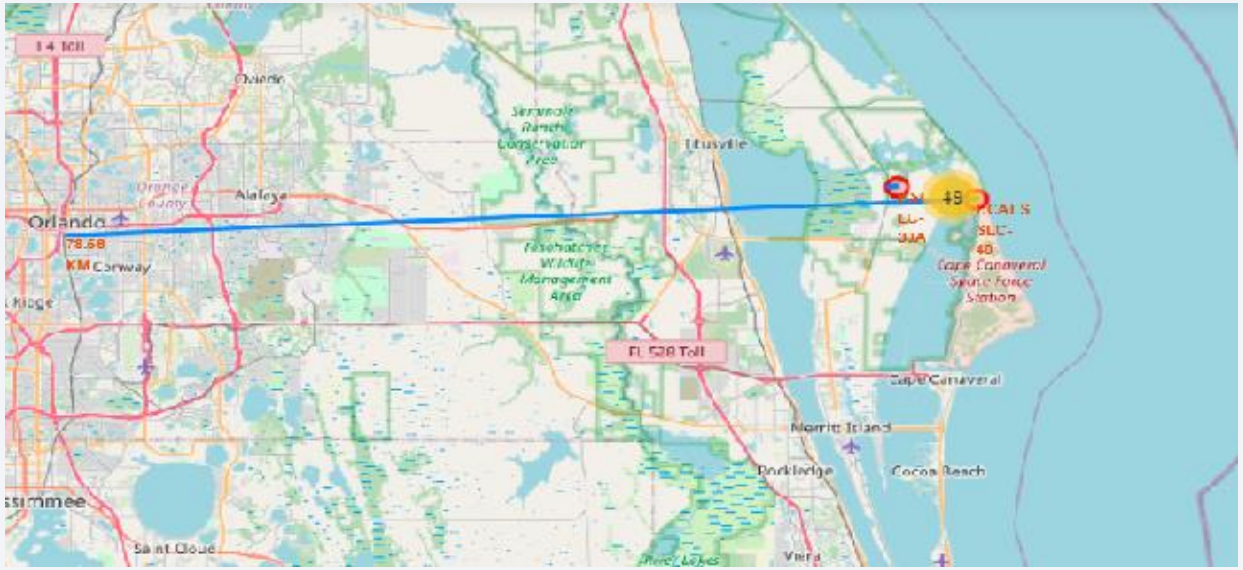
- The three main US launch sites were marked on a map:
  - Vanderburg Air Force Base
  - Kennedy Space Center
  - Cape Canaveral Air Force Station
    - Note: LC-40 entries were renamed to SLC-40 as they are identical site, LC-40 is the older, historical name

Launch Outcomes Success/Failure



- The two screenshots shown highlight successful and failed landing outcomes for Kennedy Space Center and Vanderburg Air Force Base
- Green markers indicate successful outcome
- Red markers indicate failed outcome

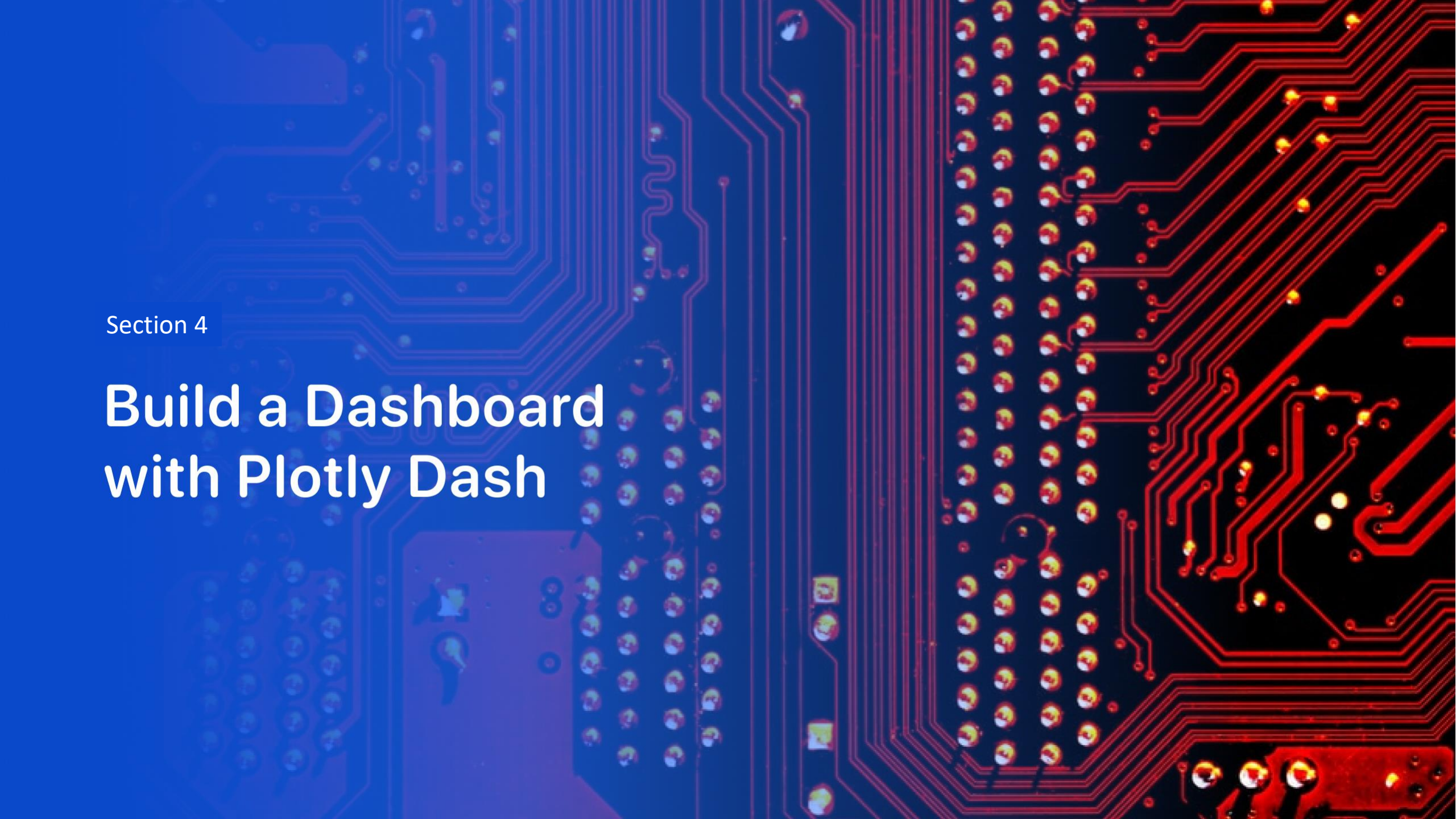
Launch Site Proximity to Geographical Features



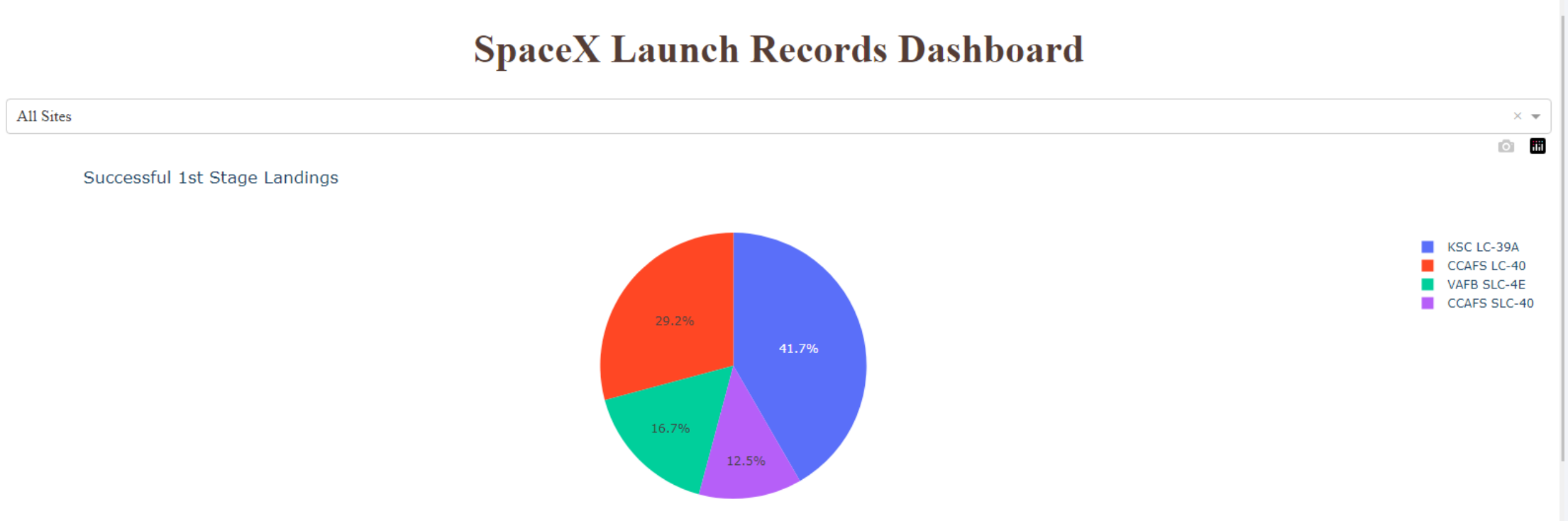
- These screenshots show the distance between Cape Canaveral from the Florida coastline and from the closest major city, Orlando.
- You can see that the coastline less than 1km away, whereas Orlando is over 75km away.

Launch Site Proximities Analysis Findings

- All three launch sites are close to the equator
  - More predictable weather patterns and temperatures
- Close Proximity to Railways and Highways
  - logistical support
  - transportation of equipment
  - transport of payloads and materials
- Close Proximity to Coastline:
  - ability to conduct launches over open water, enhancing safety in case of mission aborts
- Extended Distance from Cities:
  - distance from densely populated areas for safety reasons in case of launch failures or accidents


Section 4
Build a Dashboard with Plotly Dash

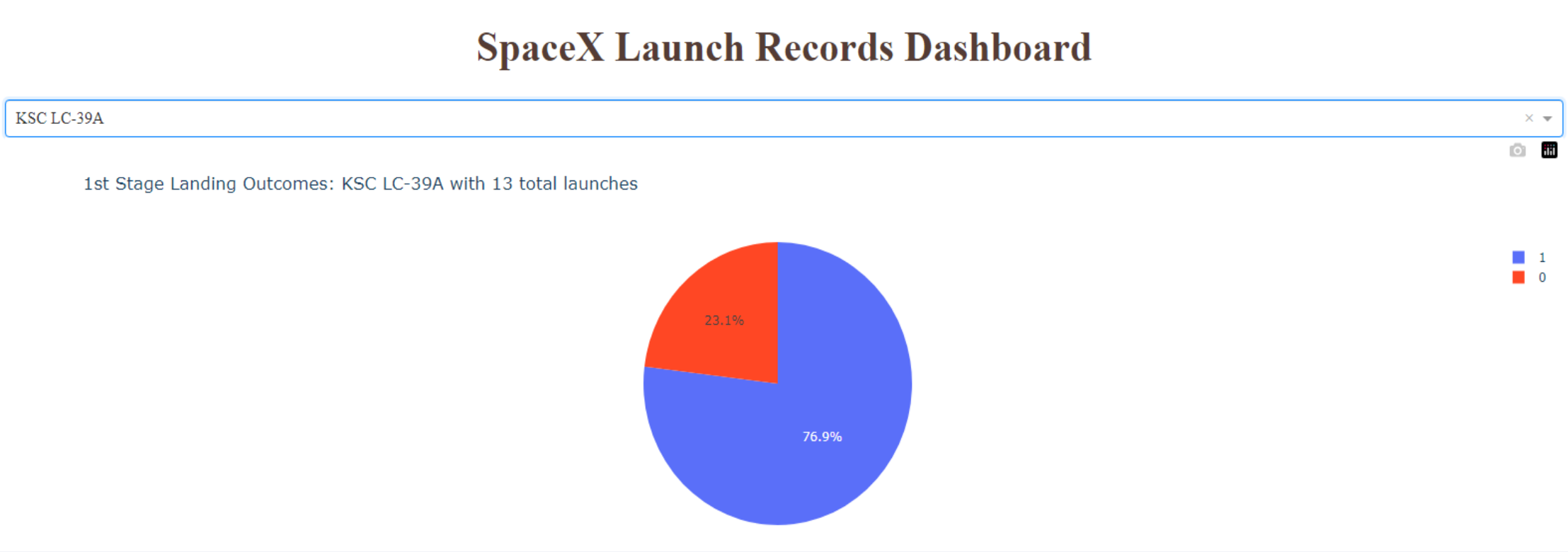
Successful 1st Stage Landings by Site



- Kennedy Space Center has the highest number of successful 1st stage booster landings.
  - Note CCAFS LC-40 and CCAFS SLC-40 are the same launch site. The LC-40 name was changed to SLC-40 in 2016 to better reflect the function of the launch site as a “space launch complex”. But, for the purposes of this study they are kept separate.

Highest Landing Success Ratio: KSC LC-39A

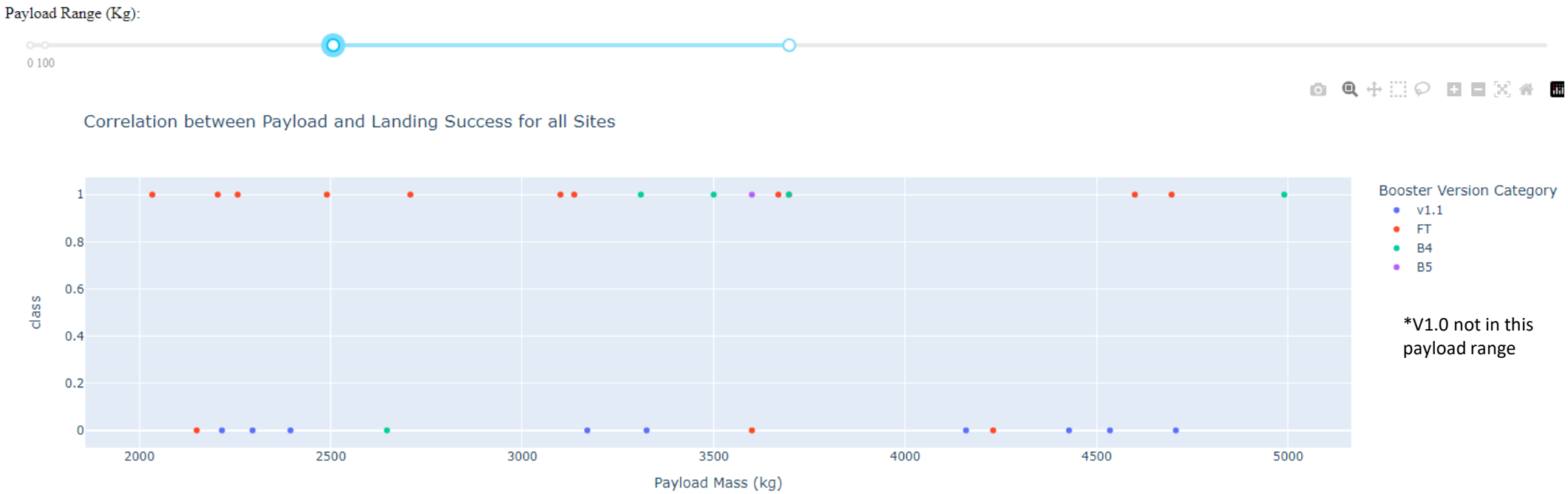
- The Kennedy Space Center, KSC LC-39A, has the highest success ratio of almost 80%.
- In the time period studied, this data represents 13 total launches.



Success Outcomes by Payload Mass, Booster Type

- For Payload Mass between (0, 10K) the FT Booster has the most number of successful landings
- The B4 Booster is the only booster that handles large payloads (>9K kg)
- The v1.0 booster was only used on a few low payload mass flights and the landings were unsuccessful

Payload Range with Both Highest & Lowest Success Rates

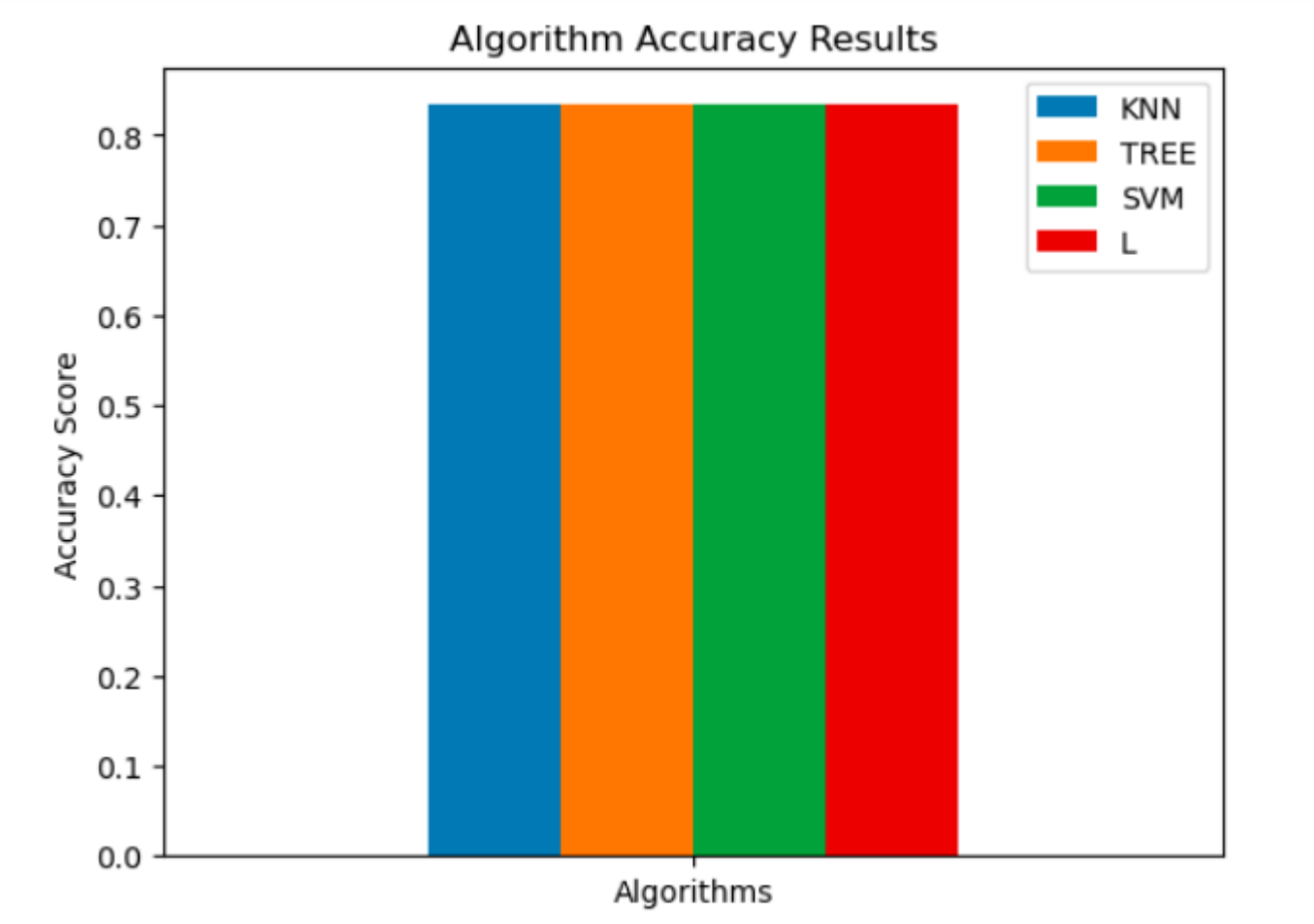


- It appears most rockets carry a payload between (2K, 5K kg)
- The number of successful and number of failed landings within this payload and for all launch sites* and boosters is about 50-50%.
  - Further analysis is needed to drill down into Flight # (earlier vs later time period flights), success by booster type, success by launch site combinations.

Section 5

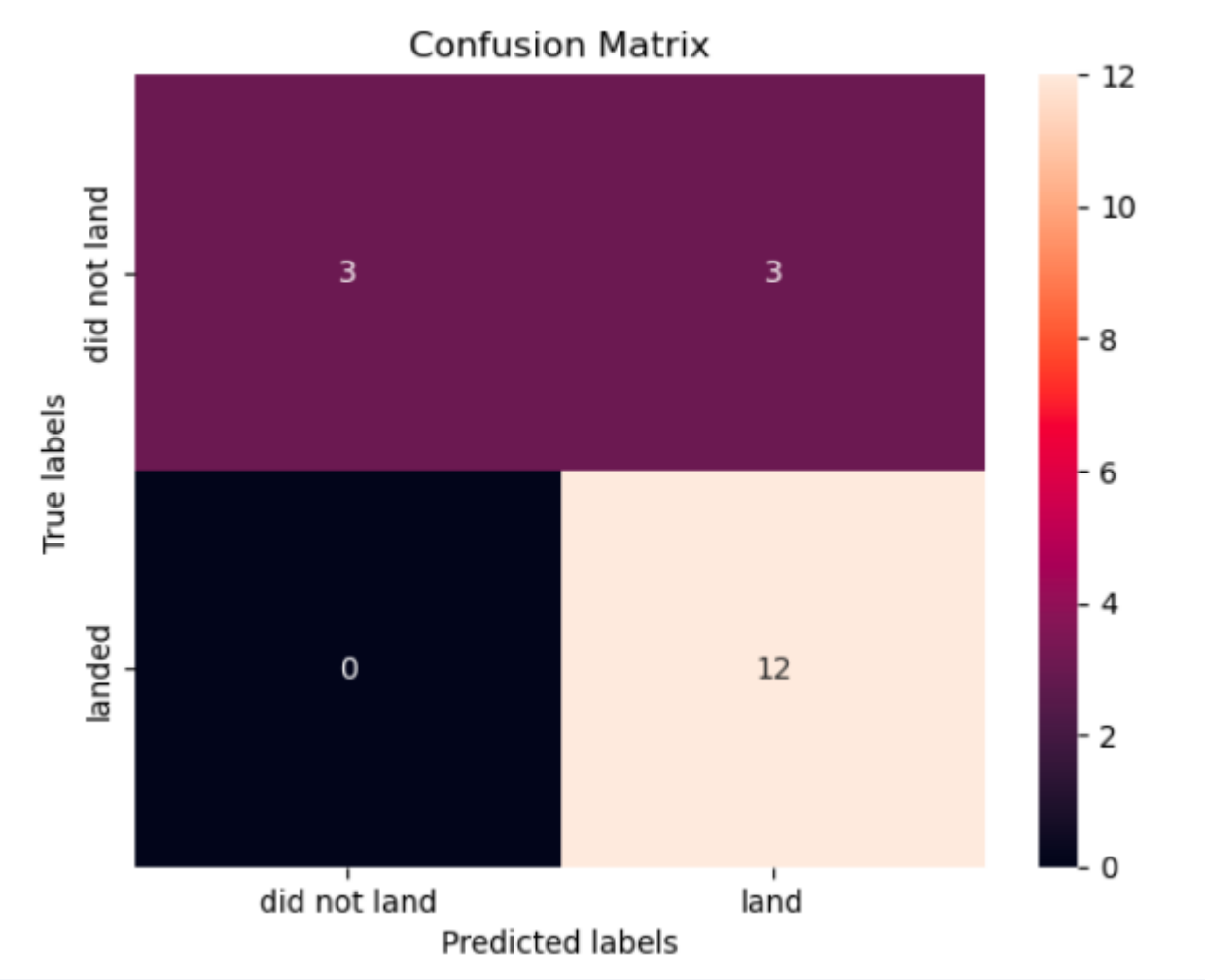
Predictive Analysis (Classification)

Classification Accuracy



- As you can see, all 4 algorithms had the same accuracy score.

Confusion Matrix



- In addition, all 4 algorithms had the same Confusion Matrix output
  - 3 false positives
- So, which algorithm is the best?
- Logarithmic Regression preferred as provides both prediction accuracy, and also measures the impact of features on the predicted outcome (Appendix E)

Appendix

Appendix A

- Our founder Billionaire Industrialist, Allon Mask
- Our company

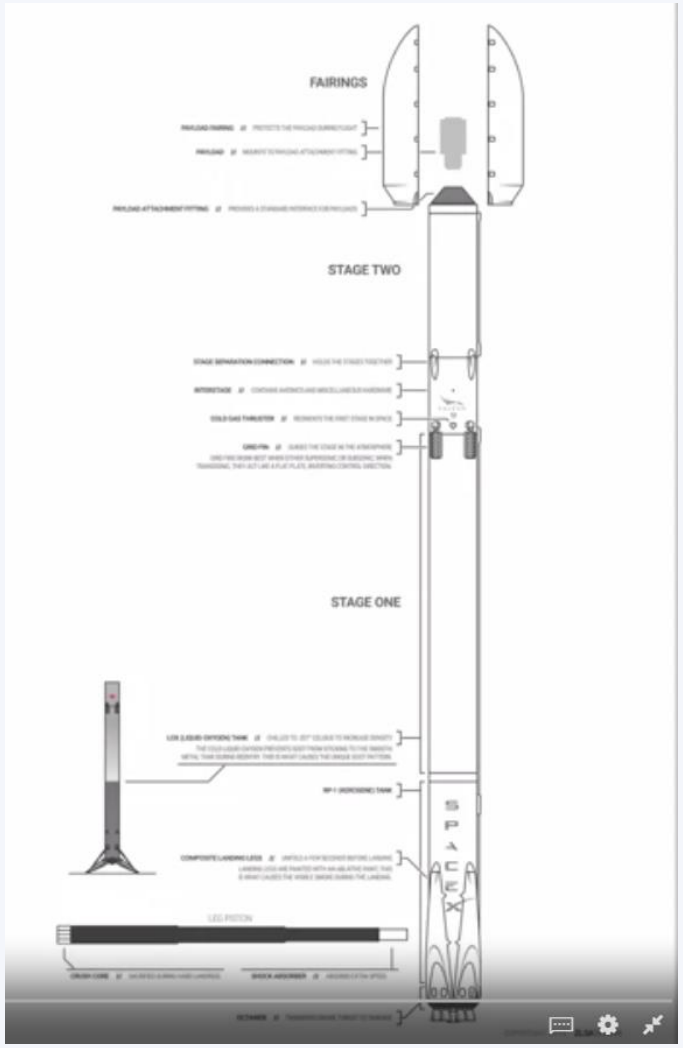
SPACE Y

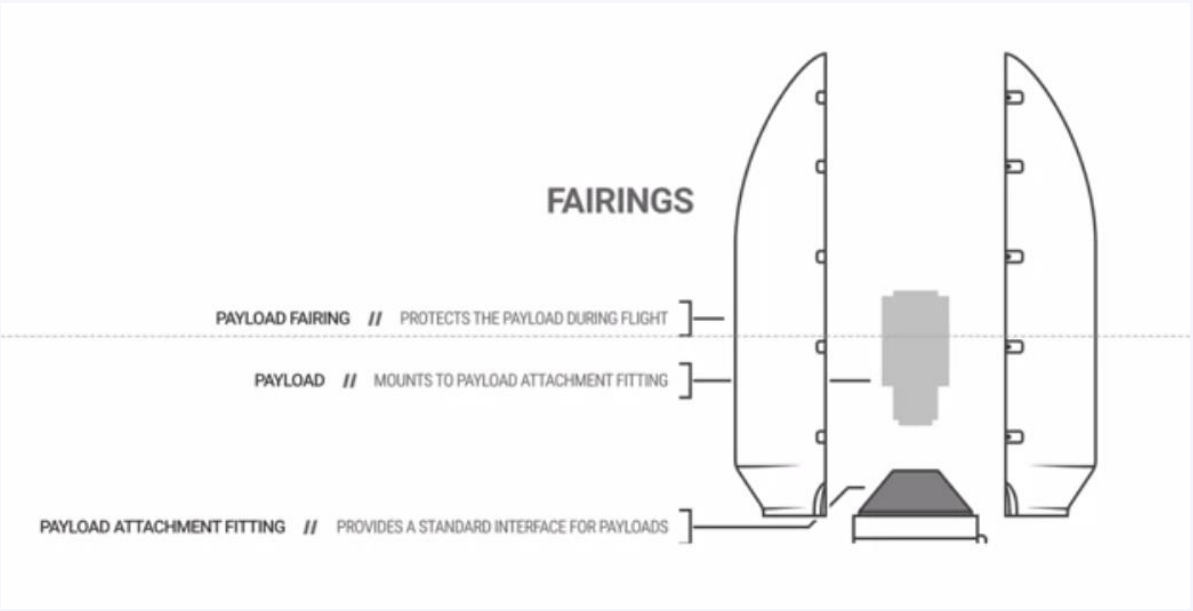
Allon Mask

Appendix B

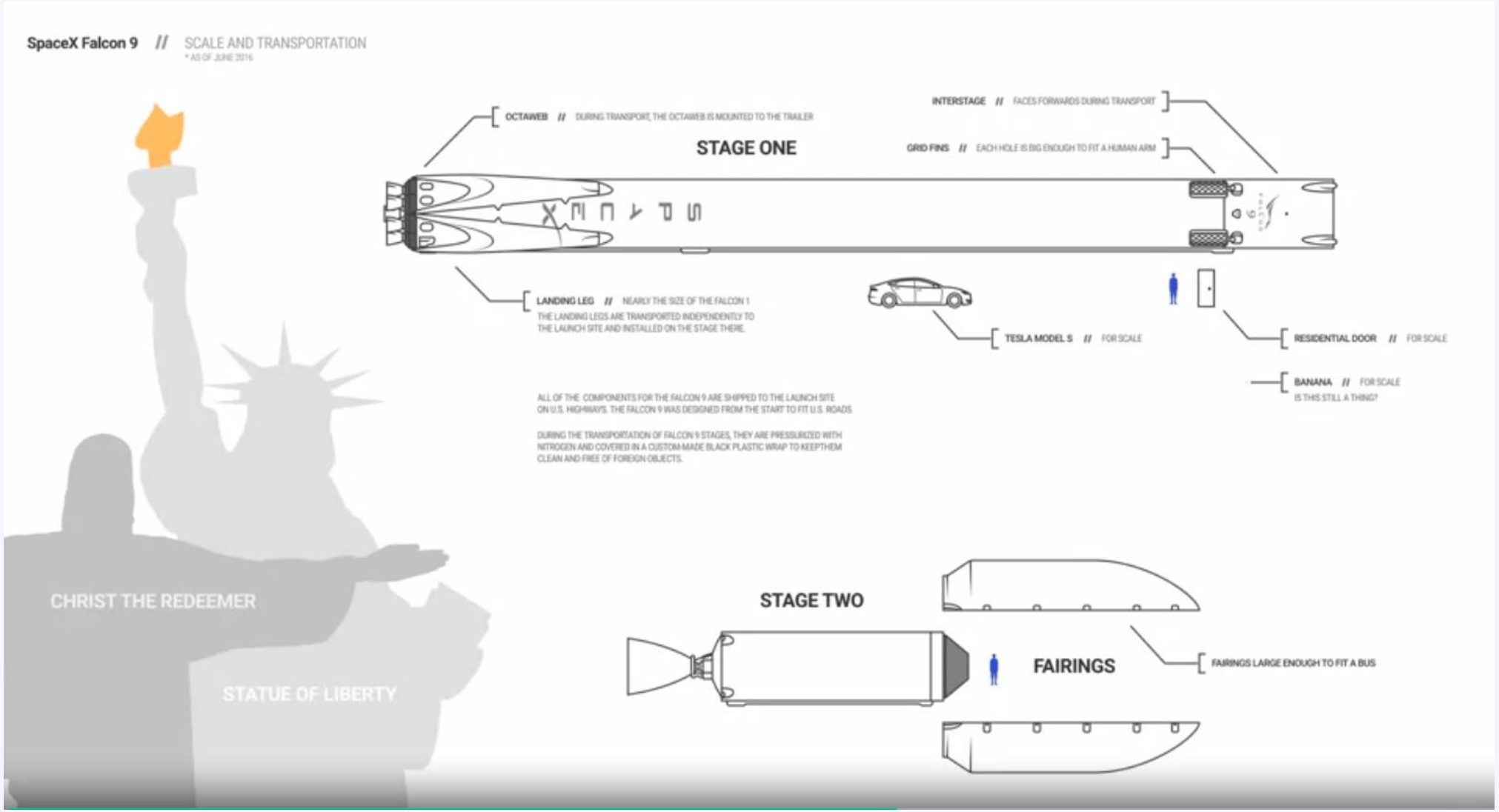
- Spaces X’s Falcon 9 launch like regular rockets. To help us understand the scale of the Falcon 9, we are going to use these diagrams from Forest Katsch, at zlsadesign.com. He is a 3D artist and software engineer. He makes infographics on spaceflight and spacecraft art. He also makes software. The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. The first stage is shown here. This stage does most of the work and is much larger than the second stage. Here we see the first stage next to a person and several other landmarks. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Appendix


FAIRINGS
STAGE TWO
STAGE ONE
SPACEX


FAIRINGS
PAYLOAD FAIRING // PROTECTS THE PAYLOAD DURING FLIGHT
PAYLOAD // MOUNTS TO PAYLOAD ATTACHMENT FITTING
PAYLOAD ATTACHMENT FITTING // PROVIDES A STANDARD INTERFACE FOR PAYLOADS

Appendix


SpaceX Falcon 9 // SCALE AND TRANSPORTATION
* AS OF JUNE 2016
OCTAWEB // DURING TRANSPORT, THE OCTAWEB IS MOUNTED TO THE TRAILER
INTERSTAGE // FACES FORWARDS DURING TRANSPORT
STAGE ONE
GRID FINS // EACH HOLE IS BIG ENOUGH TO FIT A HUMAN ARM
LANDING LEG // NEARLY THE SIZE OF THE FALCON 1
THE LANDING LEGS ARE TRANSPORTED INDEPENDENTLY TO THE LAUNCH SITE AND INSTALLED ON THE STAGE THERE.
TESLA MODEL S // FOR SCALE
RESIDENTIAL DOOR // FOR SCALE
BANANA // FOR SCALE
IS THIS STILL A THING?
ALL OF THE COMPONENTS FOR THE FALCON 9 ARE SHIPPED TO THE LAUNCH SITE ON U.S. HIGHWAYS. THE FALCON 9 WAS DESIGNED FROM THE START TO FIT U.S. ROADS.
DURING THE TRANSPORTATION OF FALCON 9 STAGES, THEY ARE PRESSURIZED WITH NITROGEN AND COVERED IN A CUSTOM-MADE BLACK PLASTIC WRAP TO KEEP THEM CLEAN AND FREE OF FOREIGN OBJECTS.
CHRIST THE REDEEMER
STATUE OF LIBERTY
STAGE TWO
FAIRINGS
FAIRINGS LARGE ENOUGH TO FIT A BUS

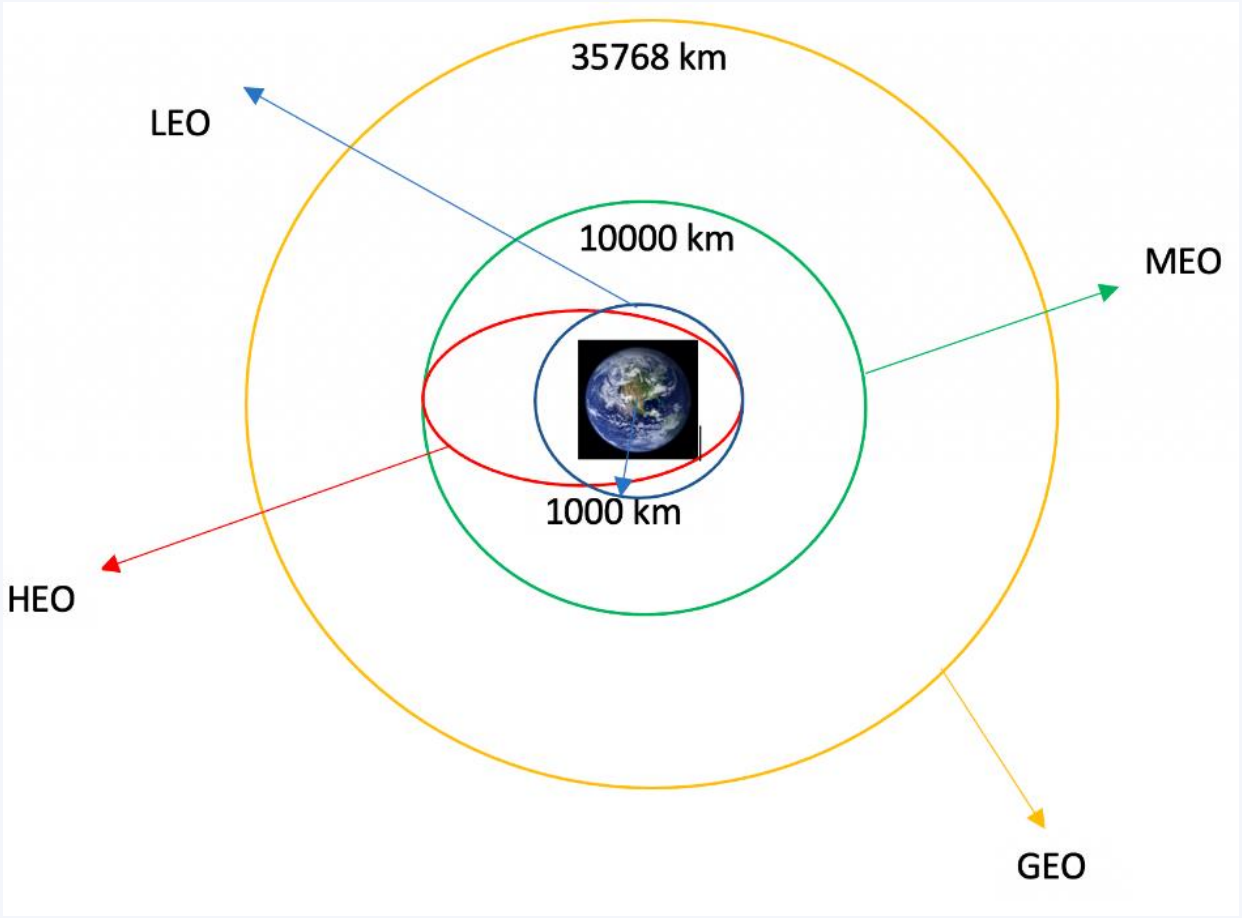
Appendix C

- Landing Outcome Definitions:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean
  - False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
  - True RTLS means the mission outcome was successfully landed to a ground pad
  - False RTLS means the mission outcome was unsuccessfully landed to a ground pad.
  - True ASDS means the mission outcome was successfully landed to a drone ship
  - False ASDS means the mission outcome was unsuccessfully landed to a drone ship.
  - None ASDS and None None represent a failure to land.

Appendix C - Orbits

- Each launch aims to an dedicated orbit, and here are some common orbit types:
- LEO: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].
- VLEO: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].
- GTO A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .
- SSO (or SO): It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .
- ES-L1 :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .
- HEO A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].
- ISS A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]
- MEO Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]
- HEO Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]
- GEO It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]
- PO It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11]

Appendix C - Orbits



Appendix D

- Python source code and datasets available on github at: https://github.com/csg47/DSCapstone

Appendix E

- Logistic Regression outputs additional information (i.e. coefficients) to measure the impact of features on the predicted outcome
- For example:
    - LaunchSite_CCAFS SLC 40, Coefficient: -0.024
    - LaunchSite_KSC LC 39A, Coefficient: 0.015
    - LaunchSite_VAFB SLC 4E, Coefficient: 0.015
- Because KSC and VAFB had the same success rates, their coefficients (i.e. impact on the model to predict success rates) are the same
- CCAFS had a lower success rate. Thus, its impact on the model prediction is negative (less likely for success) and larger than the KSC and VAFB coefficients.