IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sean Murphy
April 10, 2025

Outline

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

Executive Summary

- Data was loaded from the SpaceX API and Wikipedia
- Data was cleaned and processed for model building
- Four different classification models were built
- All models performed similar
- The most important parameter for predicting landing success is the sequential launch number
- Since SpaceX learned from their mistakes to improve their landing success, any model build from the SpaceX data is not likely to be useful for another company to predict landing success.

Introduction

Project Background and Context

- SpaceX advertises Falcon 9 launches at $62 million, significantly cheaper than competitors who charge up to $165 million, primarily because SpaceX can reuse the first stage if it lands successfully.
- By predicting landing success, we can estimate the launch cost and provide valuable insights for an alternate company bidding against SpaceX.
- This project focuses on predicting whether the first stage of SpaceX's Falcon 9 rocket will land successfully after a launch.
- The task involves collecting data from an API, formatting it correctly, and building a predictive model.

Problems to Solve

- Can we accurately predict if the Falcon 9 first stage will land successfully?
- How well does the model perform on unseen data (test set accuracy)?

Section 1

Methodology

Methodology

Executive Summary

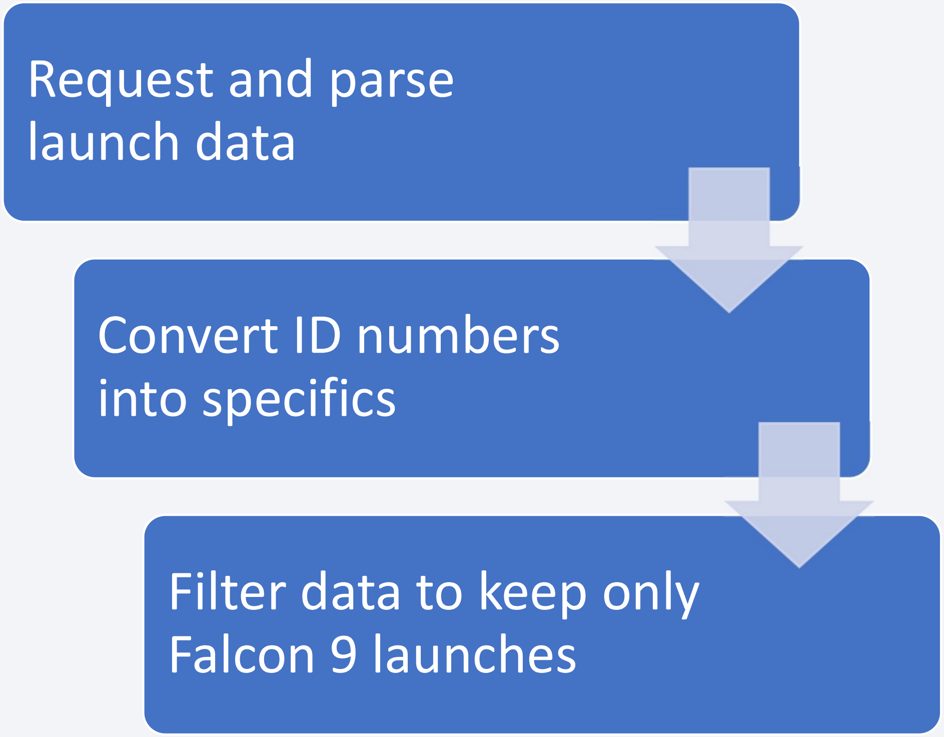
- Data collection methodology:
  - Data was collected from the SpaceX API and from Wikipedia
- Performed data wrangling
  - The data was cleaned. Failure and success were assigned to each launch.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Four models were built and tuned and the best one selected

Data Collection

1. SpaceX API
   - Past launches downloaded from api.spacexdata.com/v4/launches/past
   - IDs for rocket, payloads, launchpad and cores were used to download specifics from the API
   - Data was filtered to only examine Falcon 9 launches (90 instances)
2. Wikipedia Web Scraping
   - Wikipedia page for List_of_Falcon_9_and_Falcon_Heavy_launches downloaded
   - Data from main table was extracted (121 instances)

Data Collection – SpaceX API

- Data collected from https://api.spacexdata.com/
- Removed launches with multiple cores and multiple payloads
- Limited data to launches before Nov. 13, 2020
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-01-spacex-data-collection.ipynb



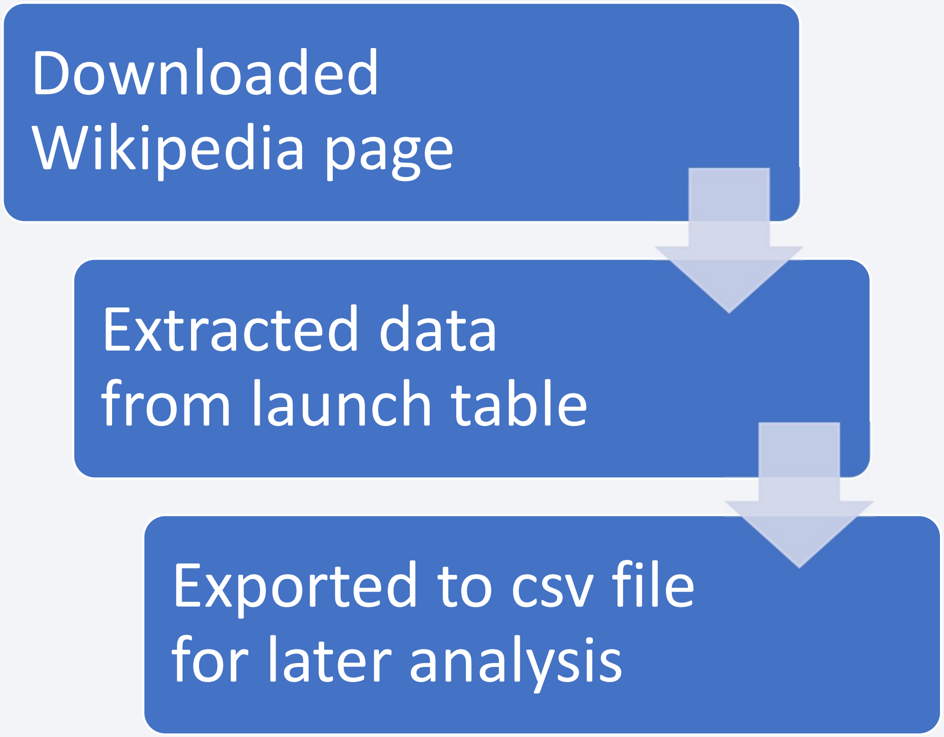
Request and parse launch data

Convert ID numbers into specifics

Filter data to keep only Falcon 9 launches

Data Collection - Scraping

- Data collected from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Included launch site, payload, orbit, and customer
- Limited data to launches before June 9, 2021
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-02-webscraping.ipynb



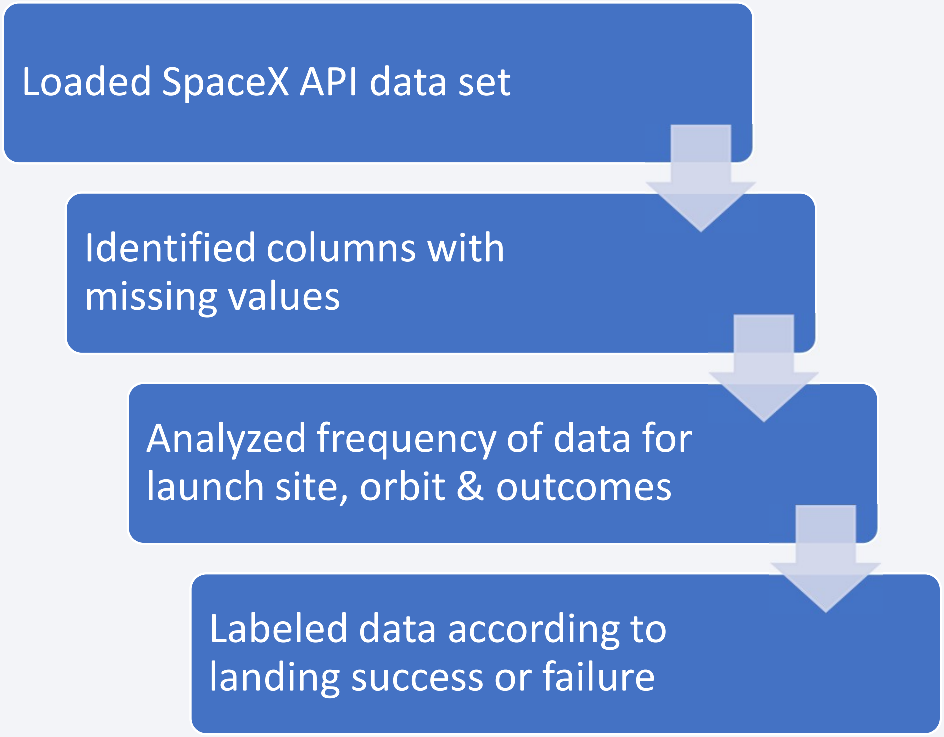
Downloaded Wikipedia page

Extracted data from launch table

Exported to csv file for later analysis

Data Wrangling

- Data collected from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Landings that weren't attempted were labeled as failures
- The data set contains 67% successful landings
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-02-webscraping.ipynb



Loaded SpaceX API data set

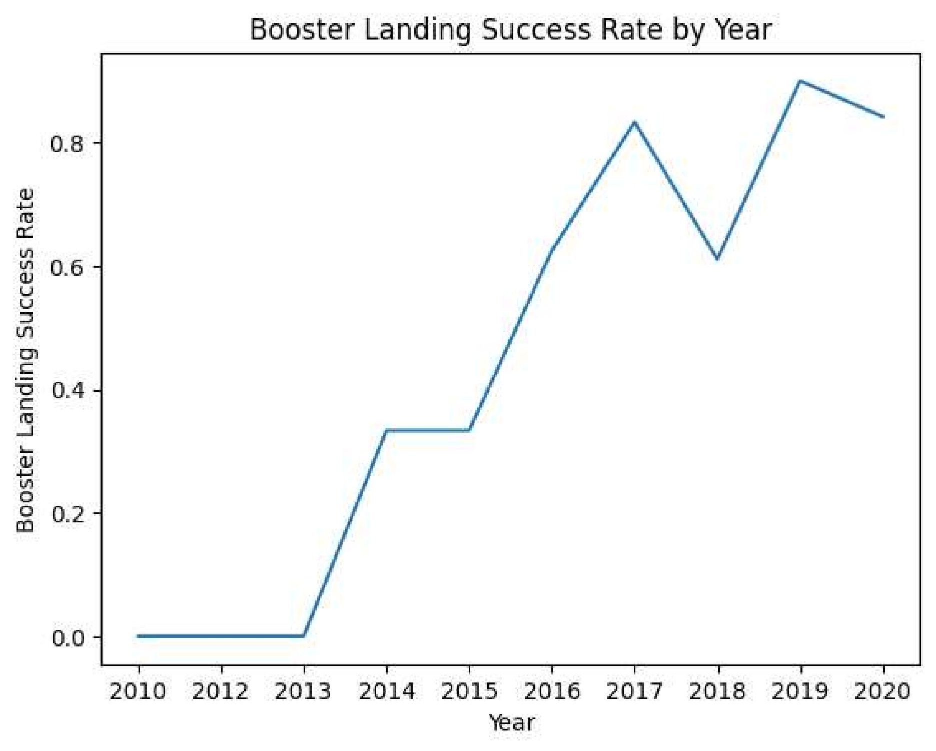
Identified columns with missing values

Analyzed frequency of data for launch site, orbit & outcomes

Labeled data according to landing success or failure

EDA with Data Visualization

- Visualizations were created to examine the dependence of landing success on:
  - flight number
  - payload mass
  - launch site
  - orbit type
  - year of launch
- The dependence on two variable was also examined
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-05-eda-dataviz.ipynb

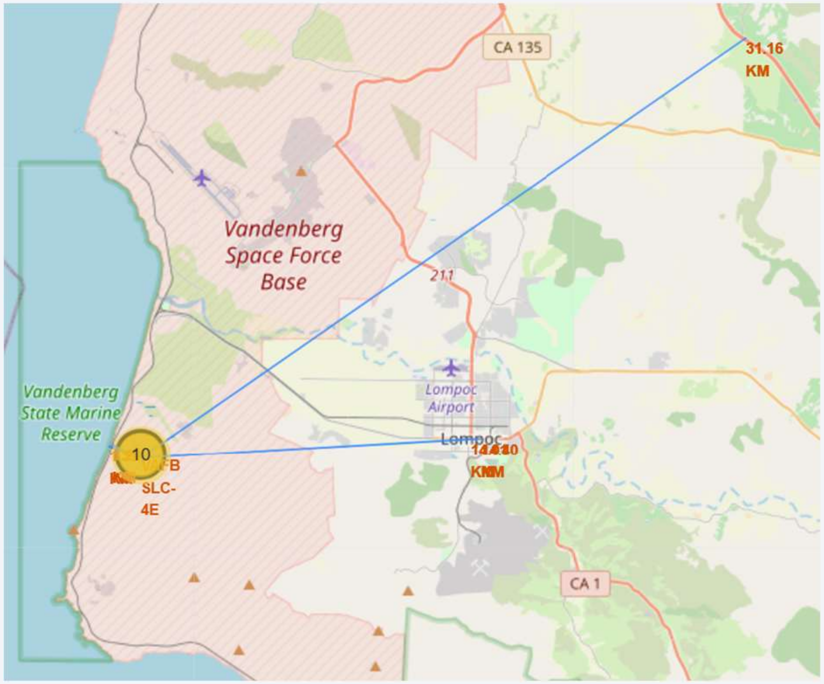


EDA with SQL

- Data was interrogated to discover:
  - the unique launch sites
  - the launches from Cape Canaveral launch sites
  - the total payload mass launched by NASA (CRS)
  - the average payload mass carried by the v1.1 booster version
  - the date of the first successful landing
  - the names of the boosters with successful landings on drone ships and with payload masses between 4000 and 6000 kgs
  - the total number of successful missions and the total number of failed missions
  - the booster versions that have carried the maximum payload mass
  - the launches which failed on drone ships in 2015
  - a ranked list of landing outcomes between two specific dates
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-04-EDA-SQL.ipynb

Build an Interactive Map with Folium

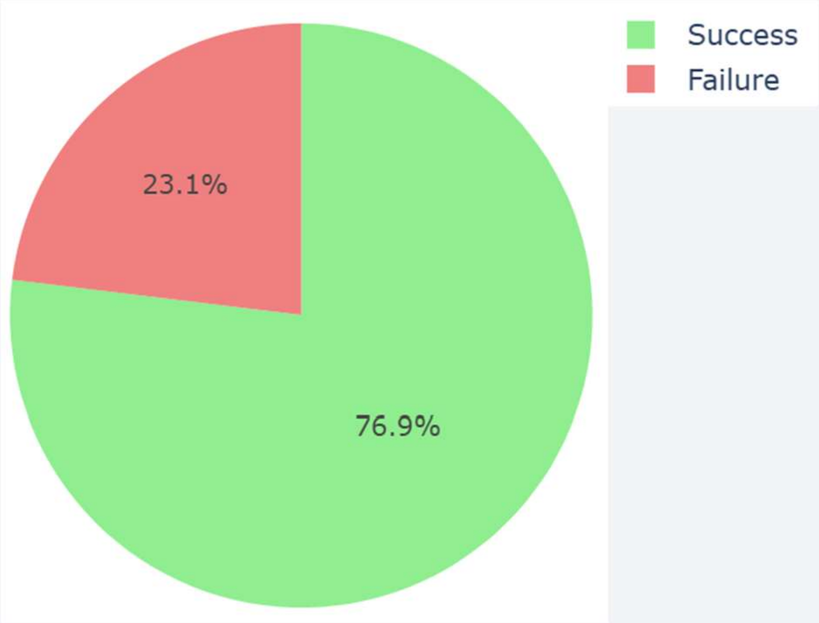
- Maps were created with:
  - launch sites as markers and circles to show where the launch sites are
  - a marker and a line to the coast to show how far away the coast is
  - a marker and a line to the nearest city to show how far away the nearest city is
  - a marker and a line to the nearest railroad to show how far away the nearest railroad it
  - a marker and a line to the nearest highway to show how far away the nearest highway is
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-06-launch-site-location.ipynb



Build a Dashboard with Plotly Dash

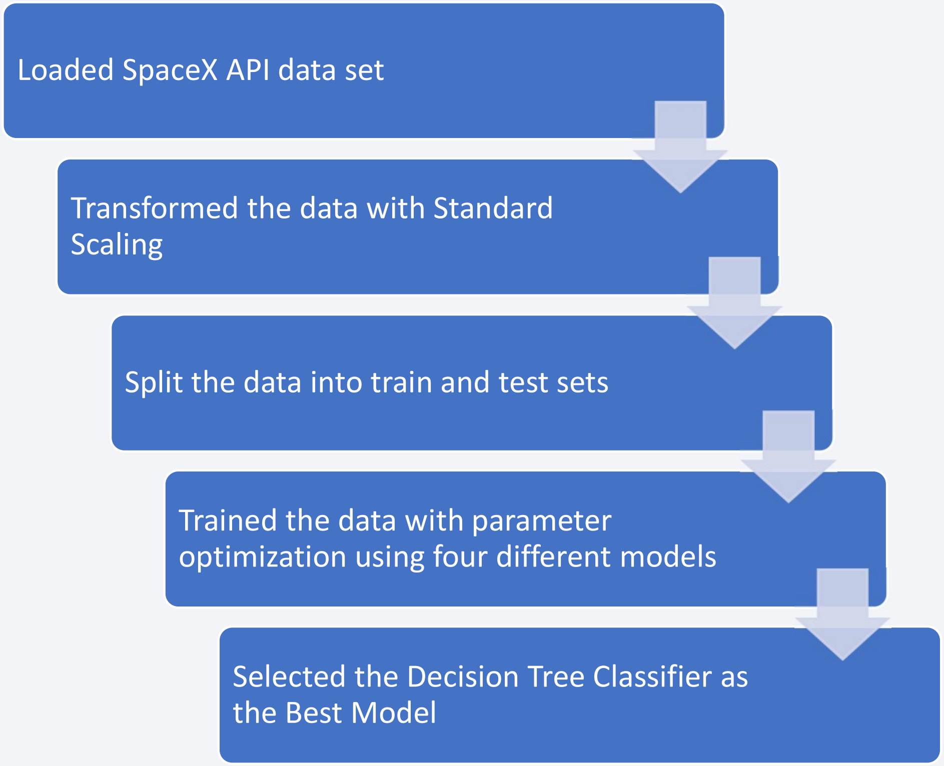
- A dashboard was created with
  - A dropdown selector for launch sites
  - Pie charts showing either the percentage of launches per site or the success/failure percentages for each launch site
  - A slider for the payload mass in kg
  - A scatter plot showing the payload vs. landing success
- These plots allow the user to view the data interactively
- Link to Python file: https://github.com/seandata88/space-y/blob/main/Course10-lab-07-spacex-dash-app.py

Success Rate for Landings after Launching from the Kennedy Space Center Site 39A



Predictive Analysis (Classification)

- A well perfoming model was selected using the flow chart at right
- Four models were tested:
  - Logistic Regression
  - Support Vector Machine
  - K-Nearest Neighbors
  - Decision Tree
- Link to Jupyter Notebook: https://github.com/seandata88/space-y/blob/main/Course10-lab-08-ML-Prediction.ipynb



Loaded SpaceX API data set

Transformed the data with Standard Scaling

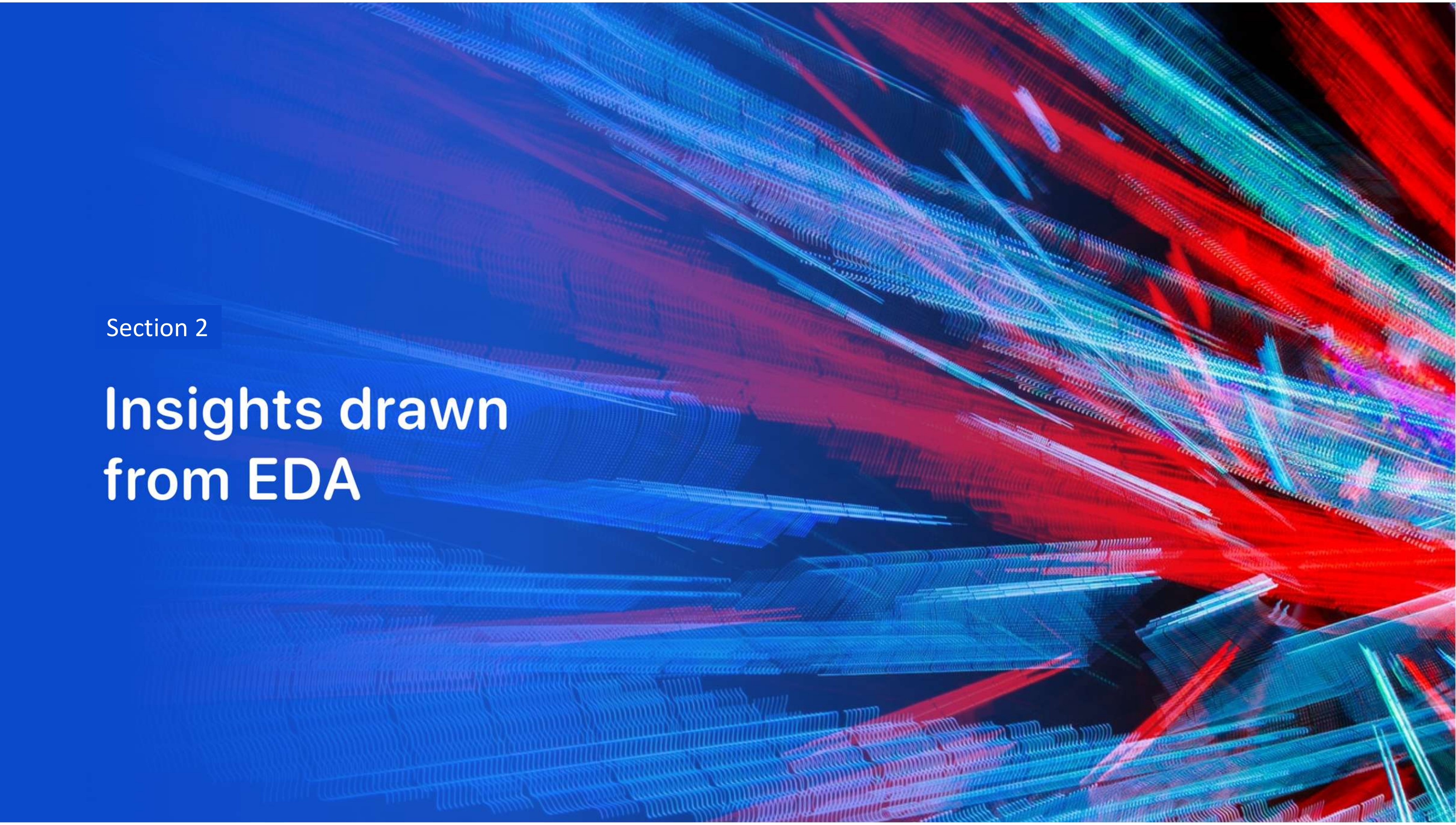
Split the data into train and test sets

Trained the data with parameter optimization using four different models

Selected the Decision Tree Classifier as the Best Model

Results

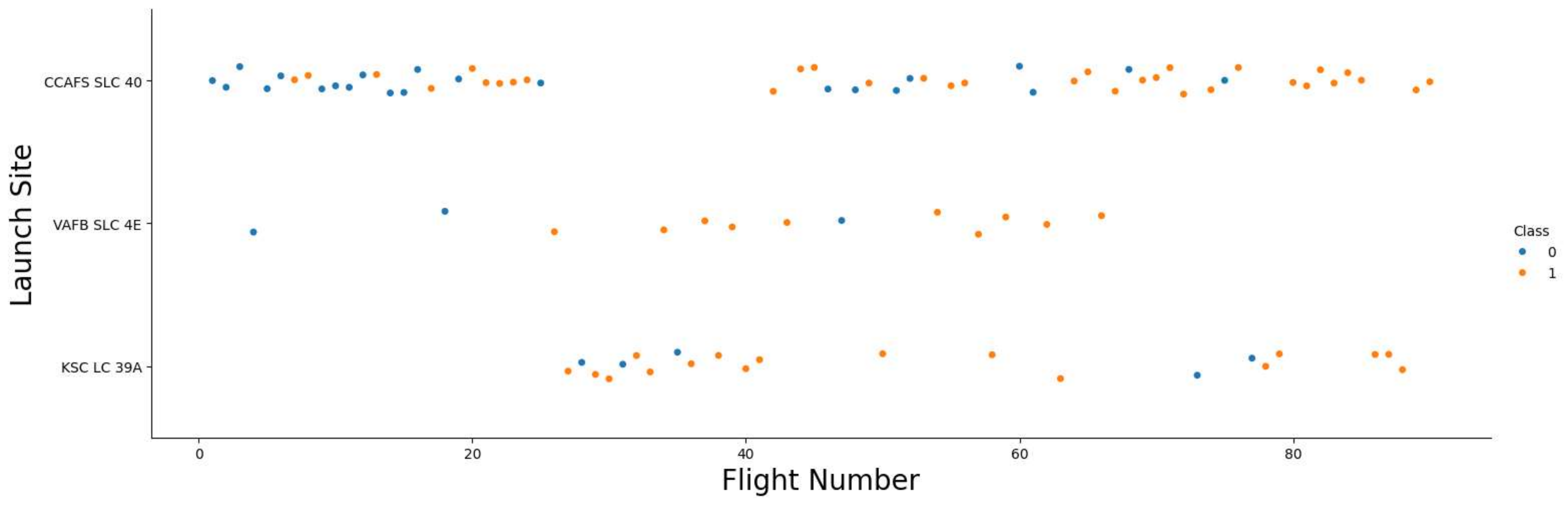
- Data collection methodology:
  - Data was collected from the SpaceX API and from Wikipedia
- Performed data wrangling
  - The data was cleaned. Failure and success were assigned to each launch.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Four models were built and tuned and the best one selected



Section 2

Insights drawn from EDA

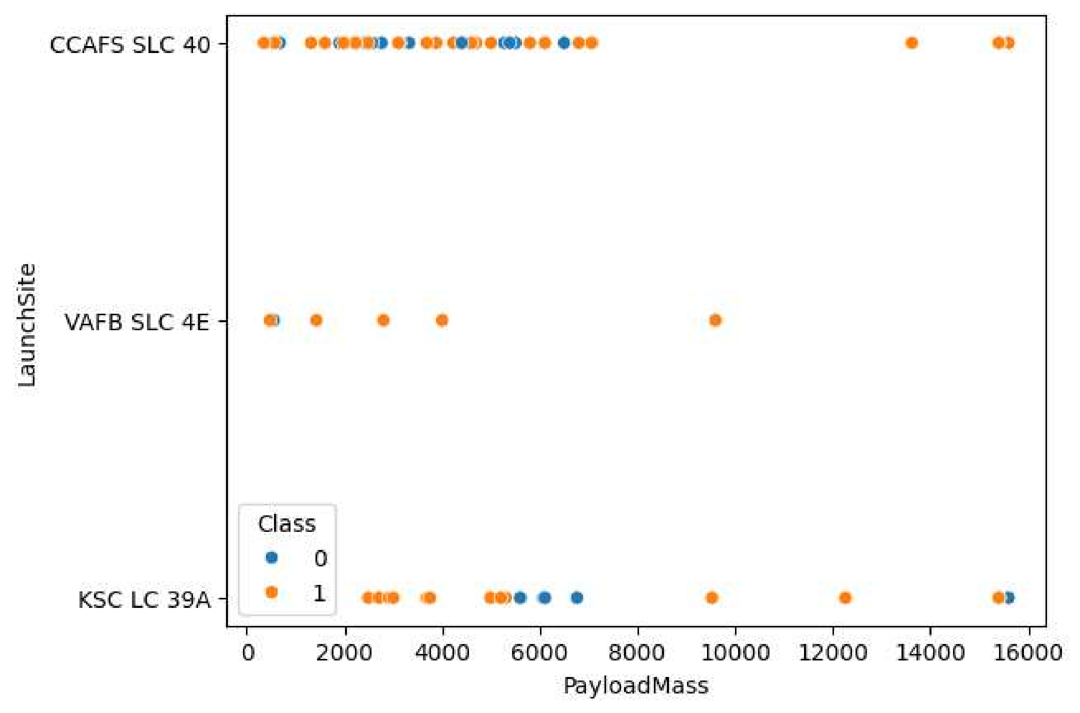
Flight Number vs. Launch Site



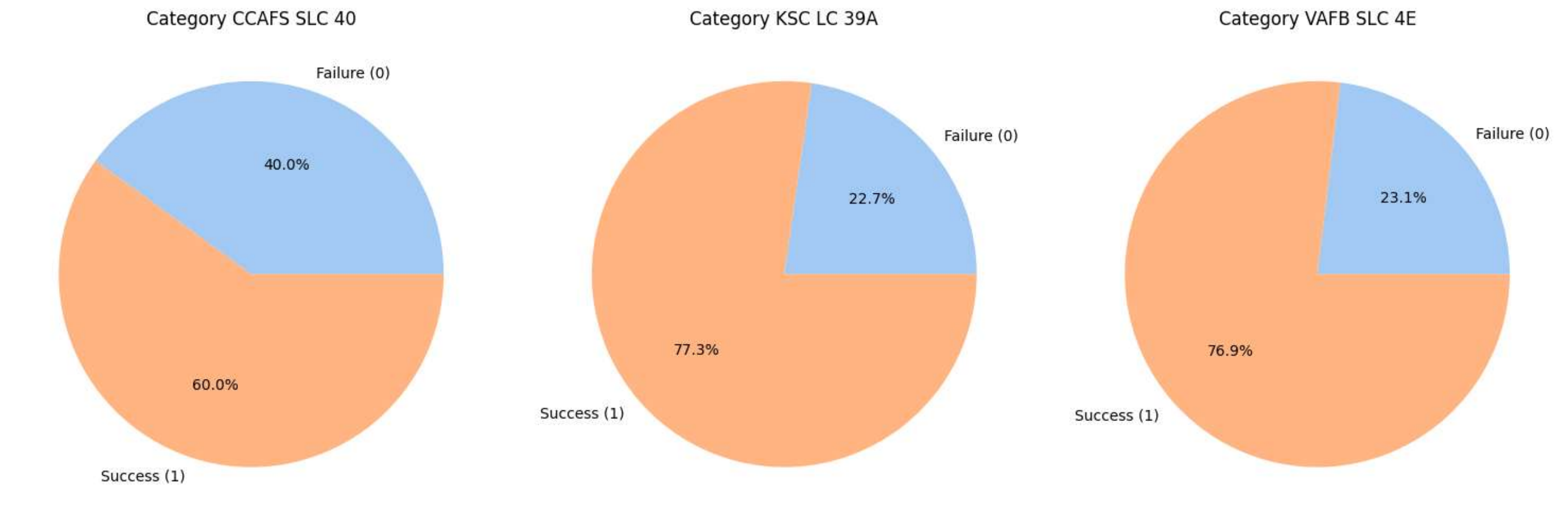
- The successful landings are shows in Orange
- All three launch sites had more success with later flight numbers

Payload vs. Launch Site

- Successful landings shown in orange
- Most larger payloads were successful
- Vandenburg has the fewest number of failed landings
- Some dots are overlapping so it is difficult to see the success rate at each site



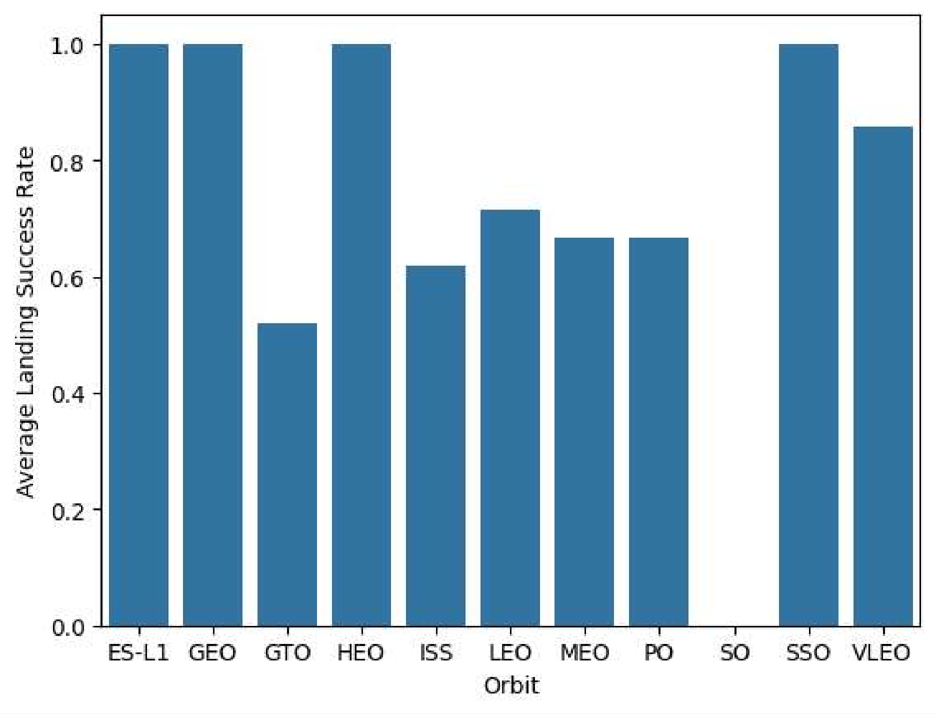
Success Rate for Each Launch Site



- Success rate for each site is better visuallized with Pie Charts
- Both Kennedy Space Center and Vandenburg had high success rates

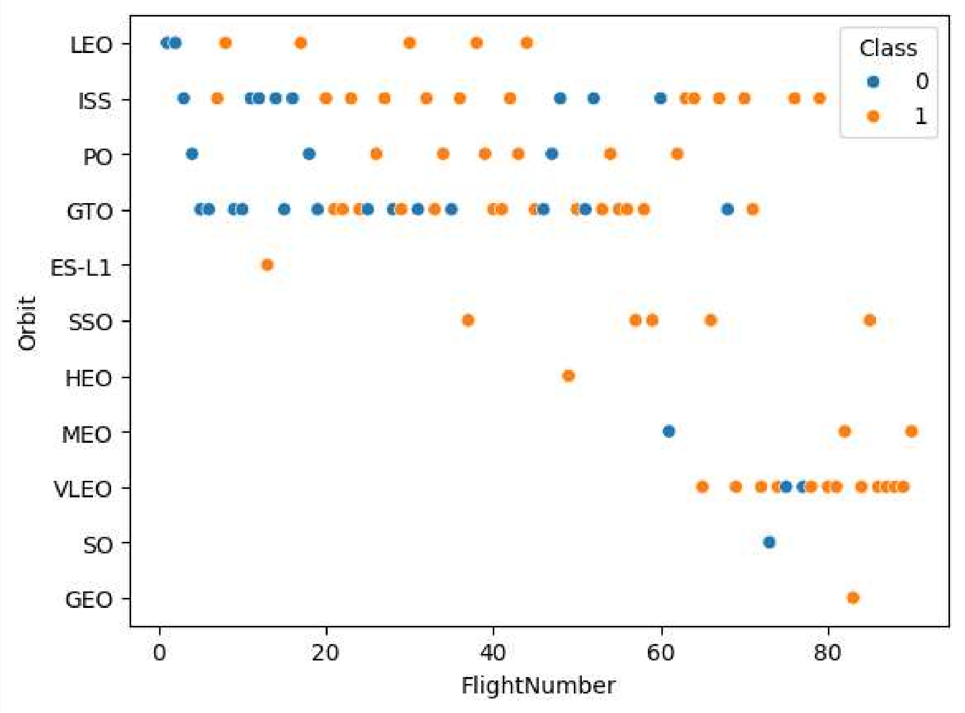
Success Rate vs. Orbit Type

- Orbital types with perfect success rates were ES-L1, GEO, HEO and SSO
- SO orbit type had no successes (but only 1 launch)
- GTO and ISS had the next lowest success rates



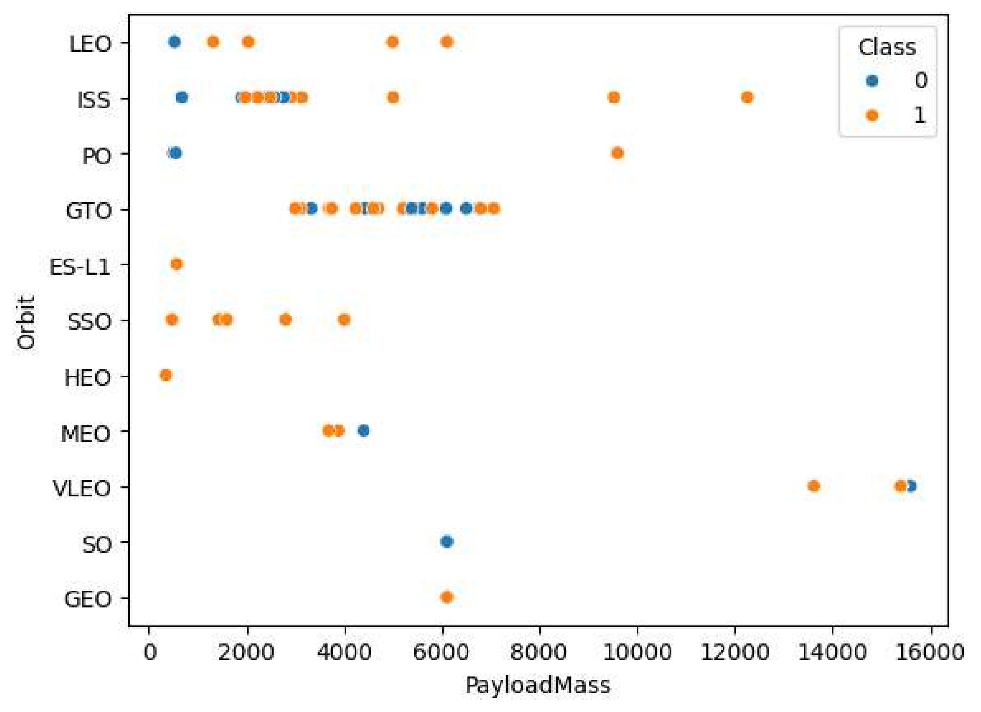
Flight Number vs. Orbit Type

- Successful landings shown in orange
- Very Low Earth Orbit (VLEO) missions increased with later flight numbers
- ISS flights had a perfect landing success rate after flight number 60



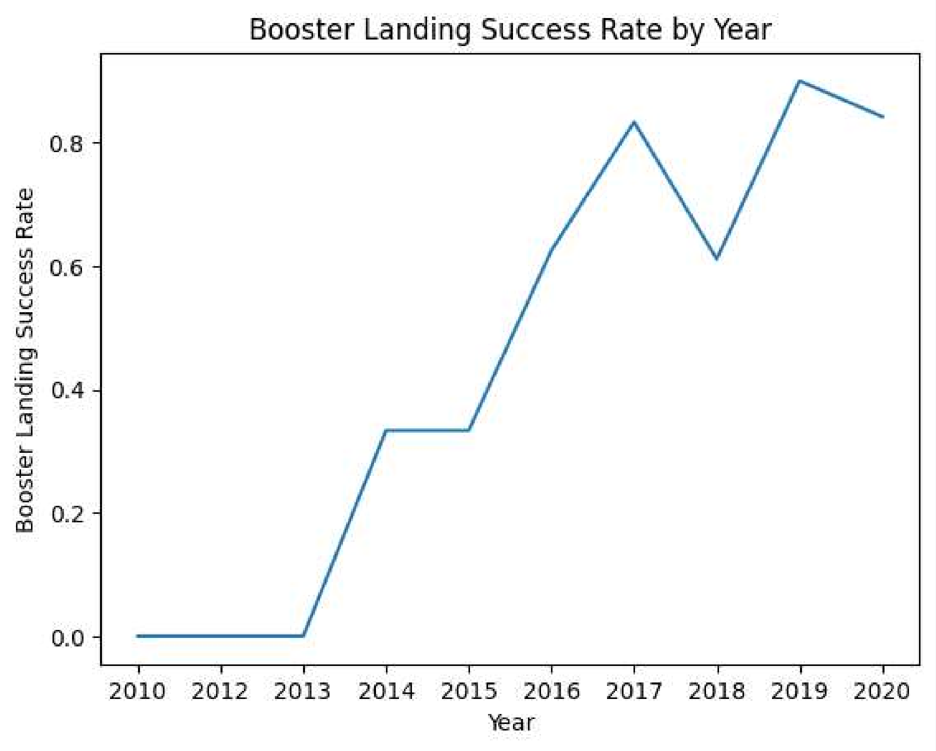
Payload vs. Orbit Type

- Successful landings shown in orange
- The largest payloads were on ISS, PO and VLEO orbits
- All SSO orbits had small payloads and successful landings
- ISS payloads have the greatest range



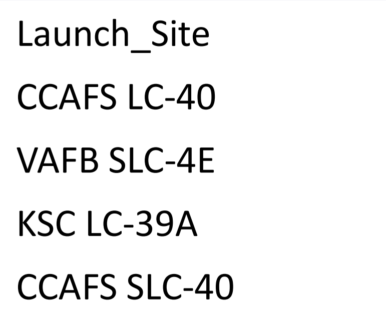
Launch Success Yearly Trend

- Landing success generally increases with year
- The year 2018 showed a noticeable drop in success rate
- The rate was constant from 2014 to 2015



All Launch Site Names

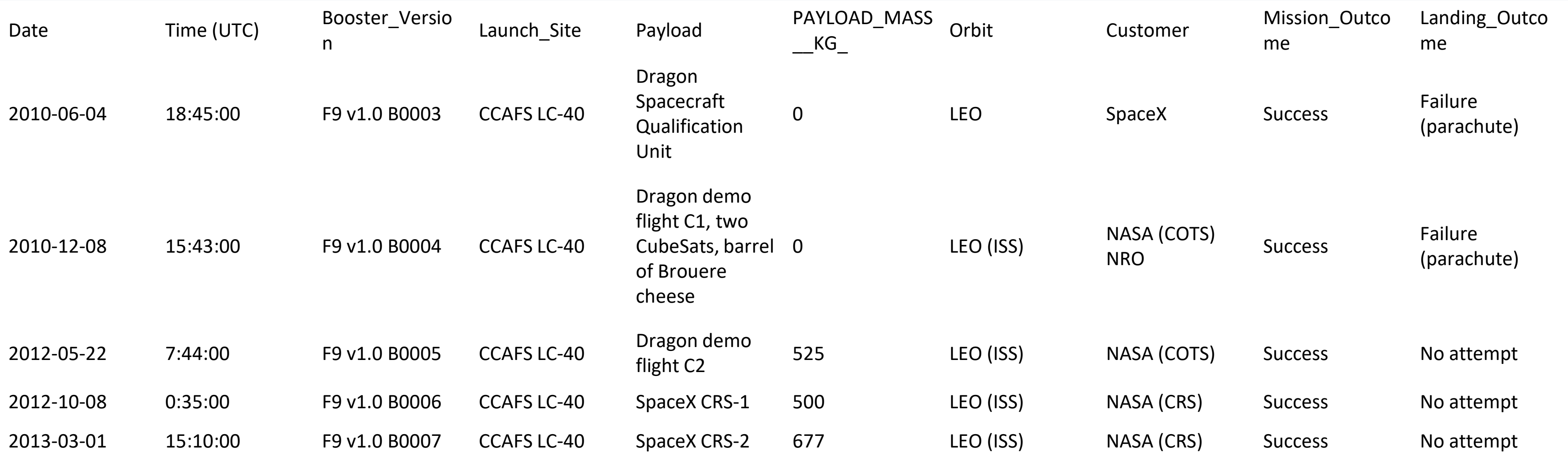
- All Falcon 9 launches in the data set launched from one of these four sites:



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

Launch Site Names Begin with 'CCA'

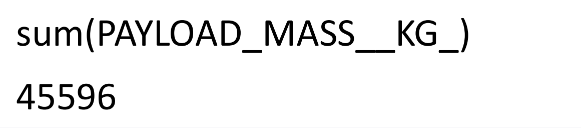
- 5 records where launch sites begin with `CCA` as shown here:



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

Total Payload Mass

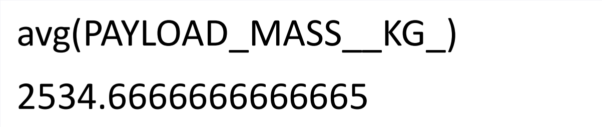
- The total payload carried by boosters from NASA (CRS) is shown here:



```
sum(PAYLOAD_MASS__KG_)
45596
```

Average Payload Mass by F9 v1.1

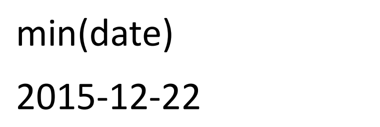
- The average payload mass carried by booster version F9 v1.1 is shown below:



avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

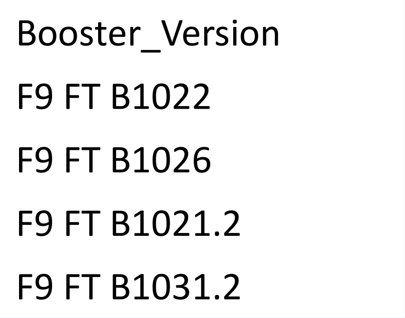
- The date of the first successful landing outcome on ground pad is:



min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

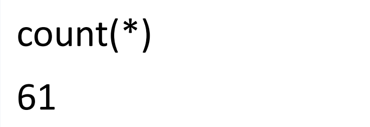
- The list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:



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

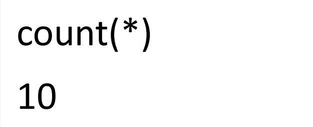
Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is:



```
count(*)
61
```

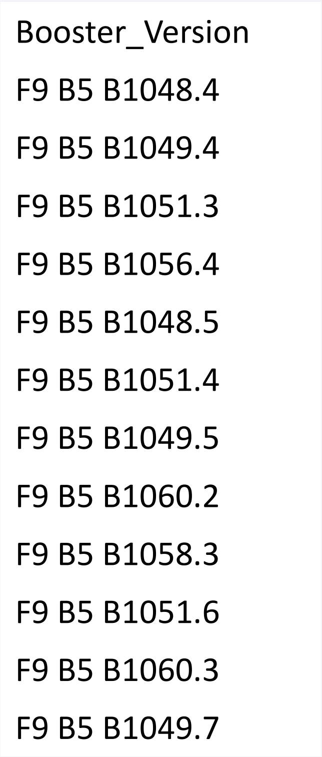
- The total number of failure mission outcomes is:



```
count(*)
10
```

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are:



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

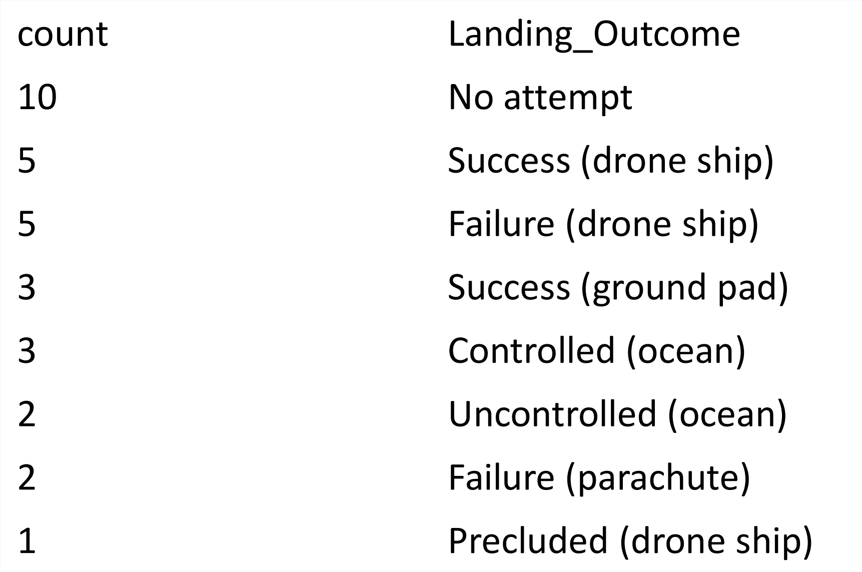
2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are:

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 ranked 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, is:

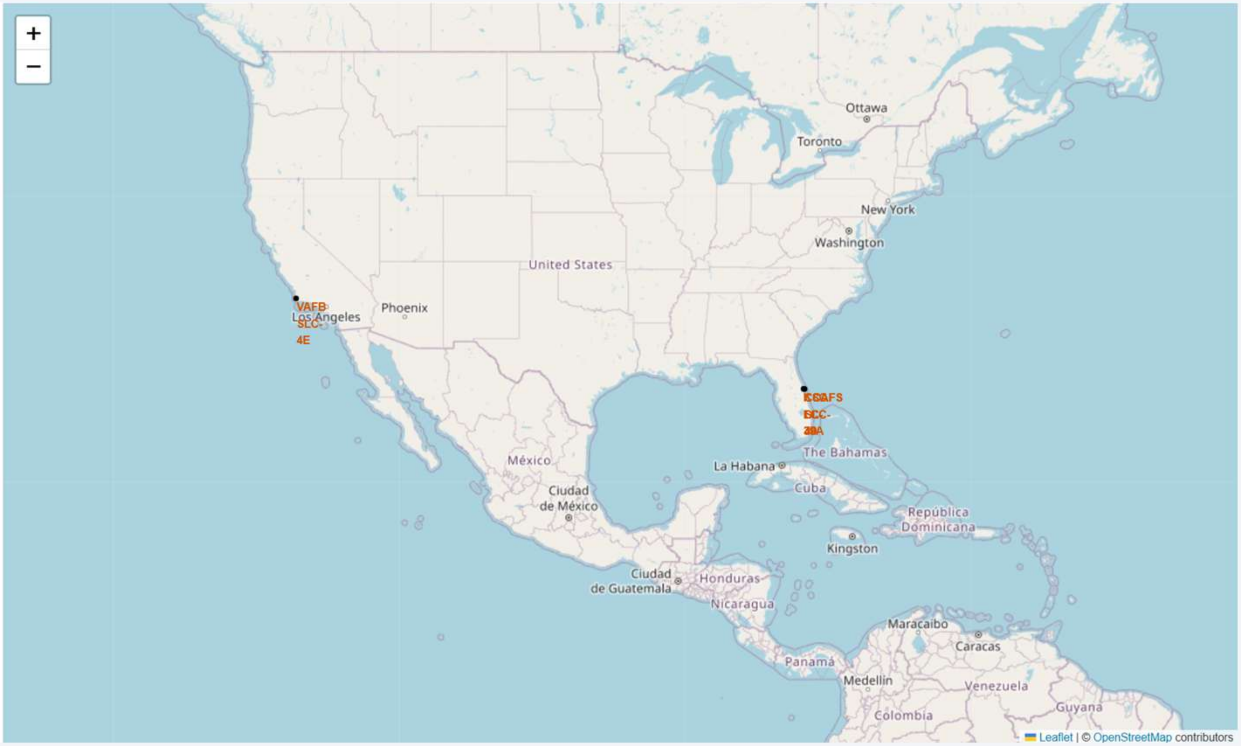


count	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

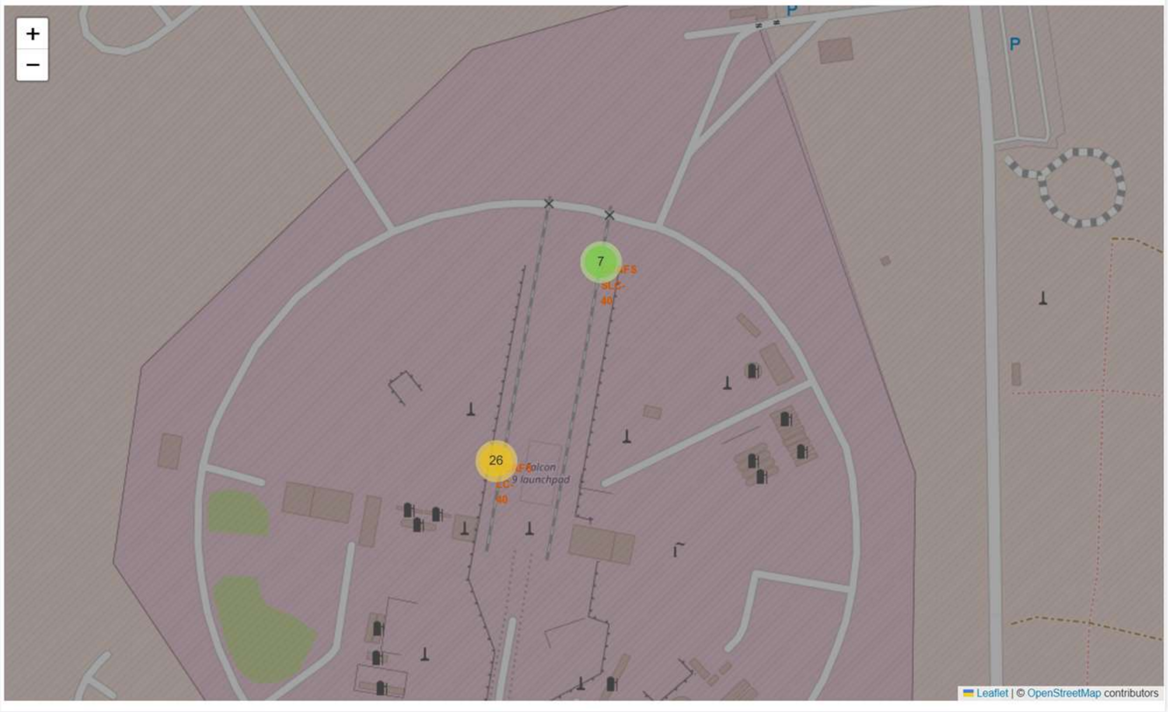
Launch Sites Proximities Analysis

Launch Sites for Falcon 9 Rockets



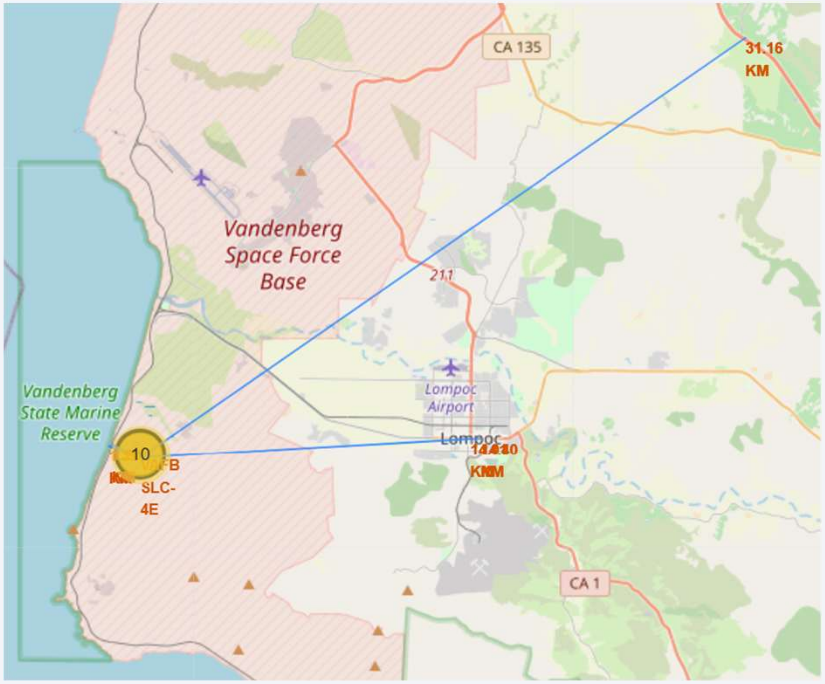
- The launch sites are on either coast of the United States
- They are in southern locations closer to the equator

Color-coded Launch Sites

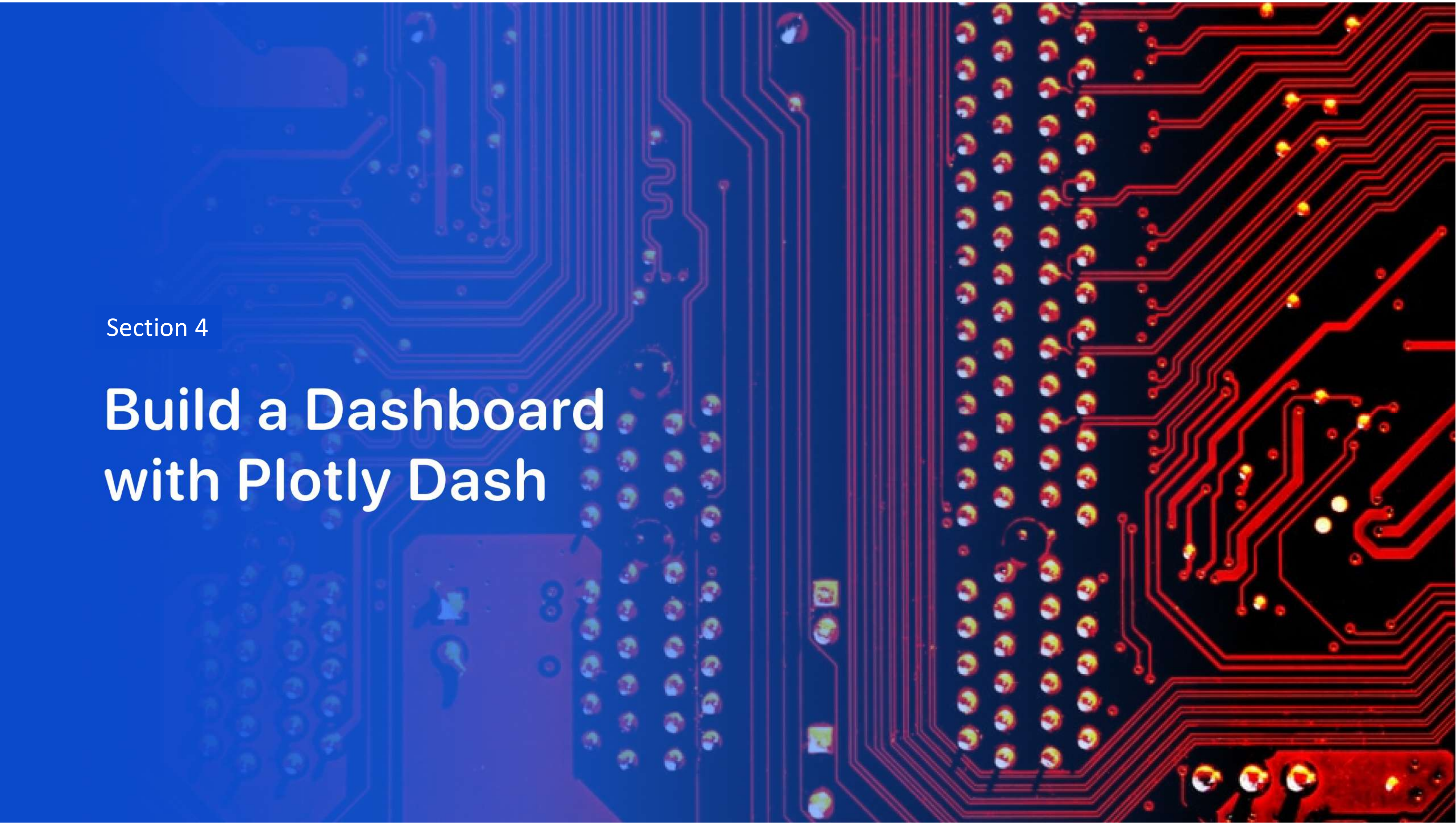


- Launch sites colored by success (green) or failure (red) of the booster landing
- Multiple launches from the same site with different success colors are shown in yellow
- Numbers at each launch site represent the numbers of launches from that site

Vandenburg Site Distance to Nearby Places



- Locations from Vandenburg launch site to:
  - nearest highway: 31.16 km
  - nearest city: 14.01 km
  - nearest Starbucks: 14.40 km
  - nearest railroad: 1.29 km
  - nearest coastline: 1.39 km

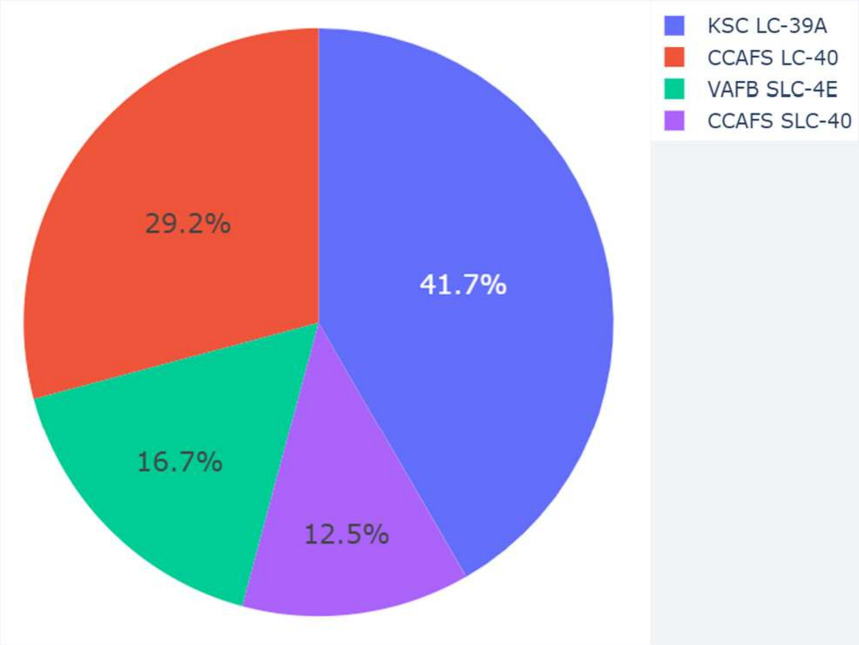


Section 4

Build a Dashboard with Plotly Dash

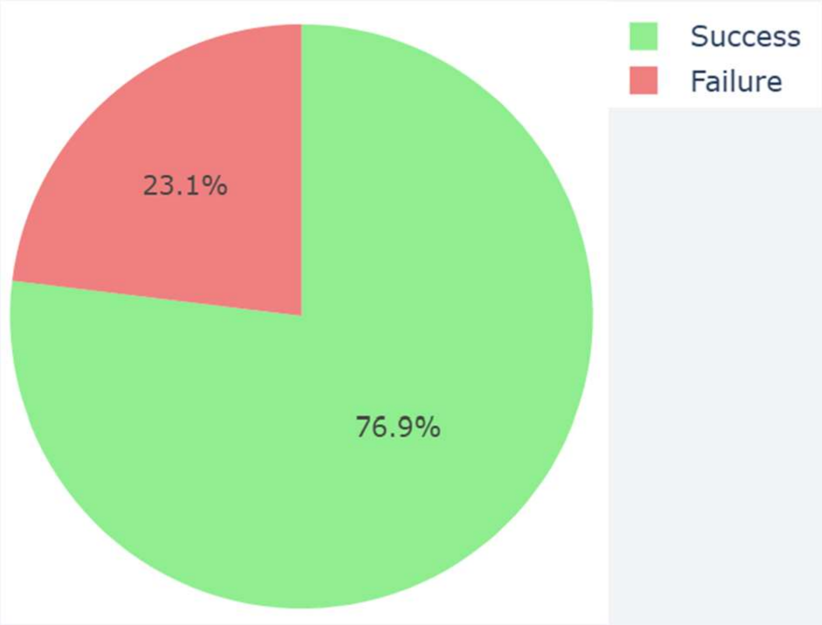
Total Success Launches By Site

- The Kennedy Space Center site 39A had the most successful landings
- The Cape Canaveral Space Launch Complex 40 had the fewest
- The dashboard allows the user to get the success rate for each site by selecting it in the drop-down menu

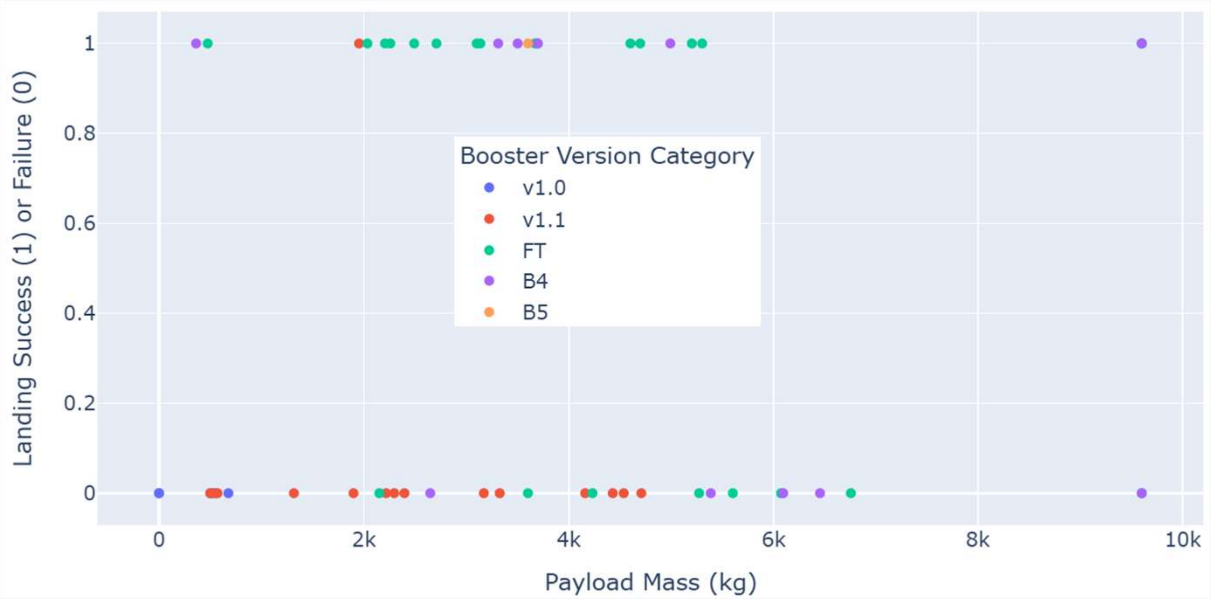


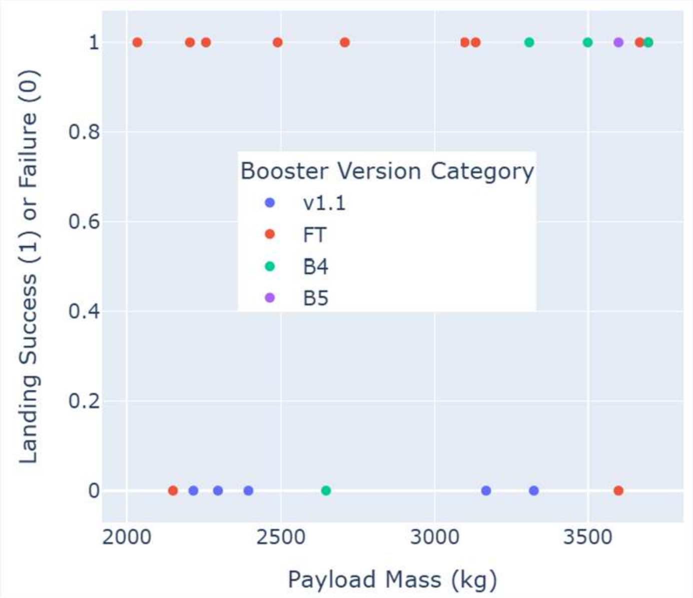
Success Rate for Kennedy Space Center Site 39A

- The Kennedy Space Center Site 39A has the highest success rate of 77%
- The other sites had the following success rates:
  - CCAFS LC-40: 27%
  - VAFB SLC-4E: 40%
  - CCAFS SLC-40: 43%



Success by Payload Range and Booster Type





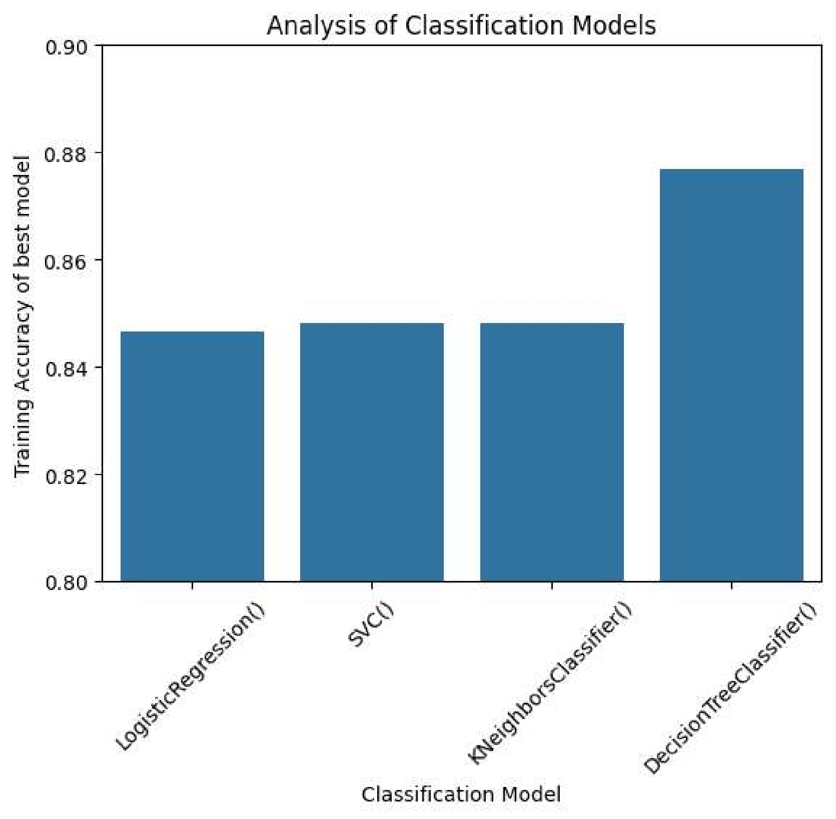
- For all payloads, FT had the most successful landings and v1.1 had the most failures
- Payloads above 6000 kg have much higher failure rate

- For payloads from 2000 to 4000 kgs, FT had the most successful landings and v1.1 had the most failures

Section 5

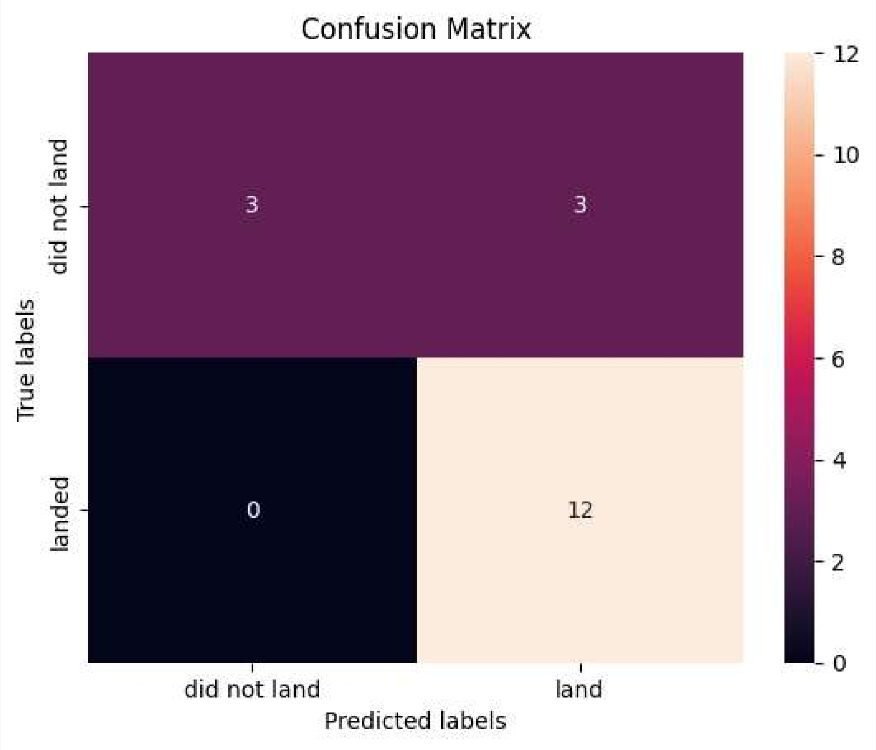
Predictive Analysis (Classification)

Classification Accuracy



- All models had similar accuracy for the best model
- The Decision Tree Classifier had the best accuracy score

Confusion Matrix



- The best model has an accuracy of 83.33 for the test set
- There are 12 true positives and 3 true negatives
- There are no false negatives
- There are 3 false positives

Conclusions

- All of the classification models performed about the same
- The most important factor in landing success is the sequential launch number which represents the ability of SpaceX to learn from their mistakes
- The increased success of SpaceX is not expected to translate to another company because they don't have the same knowledge
- If a new company truly wants to predict whether or not their rocket boosters will land, they should rely on their own experience and not the historical SpaceX data

Thank you!