IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Brett Shelley
August 2022

Outline

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

Executive Summary

- The goal of analysis is to build a model to predict whether or not a SpaceX Falcon 9 stage 1 rocket will be landed successfully so that it is able to be reused
- Data was compiled on Falcon 9 launches, including multiple factors and variables regarding the rocket used and the circumstances of the launch, including launch site, payload mass, orbit type, and flight number
- The data was processed by using one hot encoding to turn categorical data into numerical data, normalizing the data, and splitting the data into train and test sets
- Logistic regression, K nearest neighbor, classification tree, and support vector machine algorithms were fit to the training data and evaluated for accuracy in predicting the test data set outcomes.
- All of the algorithms, excluding K nearest neighbor, predicted landing outcomes with greater than 83% accuracy

Introduction

- SpaceX is able to keep rocket launch costs for the Falcon 9 rocket low by reusing the first stage of the rocket when it can be landed successfully
- By being able to predict if the first stage can be reused or not, we can more accurately determine the true cost of individual rocket launches
- Using publicly available data and machine learning algorithms, we will attempt to build a classification model to predict whether or not the Falcon 9 rocket first stage will be successfully landed so it can be reused

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was collected directly from SpaceX using their API to retrieve launch data
  - Data was also collected from publicly available records online using web scraping
- Perform data wrangling
  - The data was cleaned to remove unnecessary and null values
  - Used one hot encoding to transform categorical data into numerical data for analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built multiple classification models using different machine learning algorithms to evaluate and choose the best predictive model

Data Collection

- Data was collected from two sources, directly from SpaceX using their public facing API, as well as from public data available on Wikipedia.org

- Data was compiled to include Date, Flight Number, Payload Mass, Orbit Type, Launch Site, Outcome, and Landing Pad for Falcon 9 rocket launches

- The resulting data was prepared for processing by replacing 5 missing values of Payload Mass with the mean of all payload masses

Data Collection – SpaceX API

- Use REST get request to https://api.spacexdata.com/v4 to retrieve past rocket launch information as JSON data
- Read JSON data in pandas dataframe
- Filter the resulting dataframe to only include data for Falcon 9 launches
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/Data_Collection_SpaceX_API.ipynb

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Use REST get request to https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches to retrieve supplementary data
- Use beautifulsoup to parse web page data into a pandas dataframe
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/Data_Collection_Web_Scraping.ipynb

Place your flowchart of web scraping here

Data Wrangling

- Used one hot encoding to create a numerical label for successful Falcon 9 rocket landings from the categorical data of multiple different landing outcomes
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/Data_Wrangling.ipynb

EDA with Data Visualization

- Used scatter plots, bar charts, and line plots to explore the relationships between different launch variables and the landing outcome to help identify which features may be most significant
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/EDA_Visualizations.ipynb

EDA with SQL

- Used SQL queries to pull identify specific records from the dataset to help understand the overall dataset and trends within the data

- `select unique(LAUNCH_SITE) from SPACEXDATASET`
- `select avg(payload_mass__kg_) from spacexdataset where booster_version = 'F9 v1.1'`
- `select date from spacexdataset where landing__outcome = 'Success (ground pad)' order by date asc limit 1`
- `select mission_outcome, count(*) as count from spacexdataset group by mission_outcome`

- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Used Folium to add markers to and display a map visualizing the different launch sites used for the Falcon 9 rockets and their corresponding landing success rates
- Clusters were used to display multiple launches for a single site since there are many individual records for each site.
- Lines and distance measurements were also added to the map to examine the possible influencing factor of nearby geographical features
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/Visualizations_Folium.ipynb

Build a Dashboard with Plotly Dash

- Created a dashboard to be able to see landing success rates for all sites, and for each individual site.
- The dashboard also contained a selection criteria for payload mass which allows a user to select different payloads and see the resulting effect on landing outcomes
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/5b18e84dabf519072dfa6e62bdebd3cef1c73910/Visualizations_Plotly.py

Predictive Analysis (Classification)

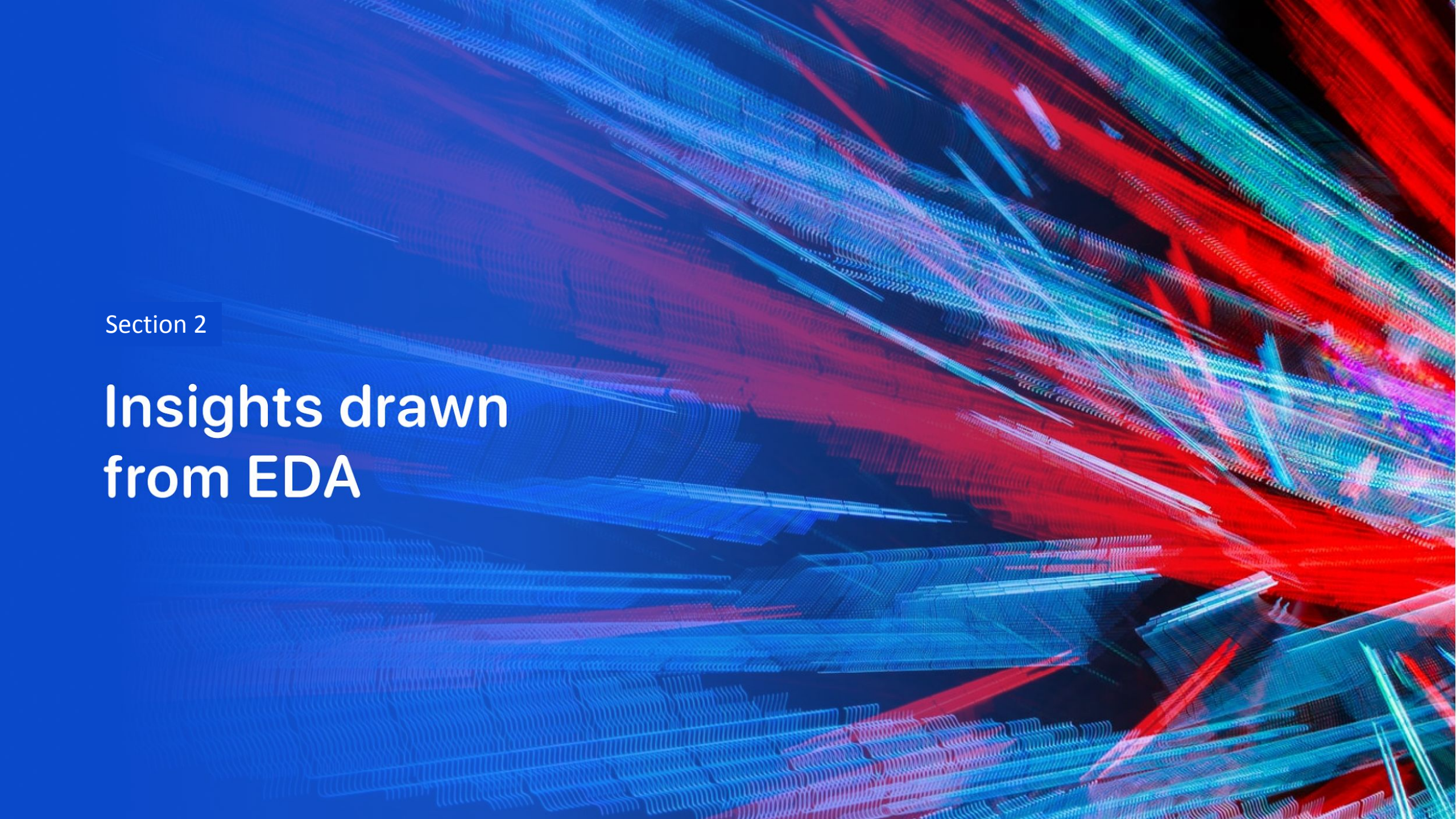
- Using cross fold validation, multiple models were trained to predict the landing outcome of Falcon 9 rockets. The models were tested on subsets of our historical data to evaluate which has the best accuracy.
- The models created were K nearest neighbors, classification tree, support vector machine, and logistic regression
- https://github.com/brettshelley/IBM_Coursera_Capstone/blob/f9faa45eddc844811384cee2d7ef0e16533ad78a/Predictive_Analysis.ipynb

Results

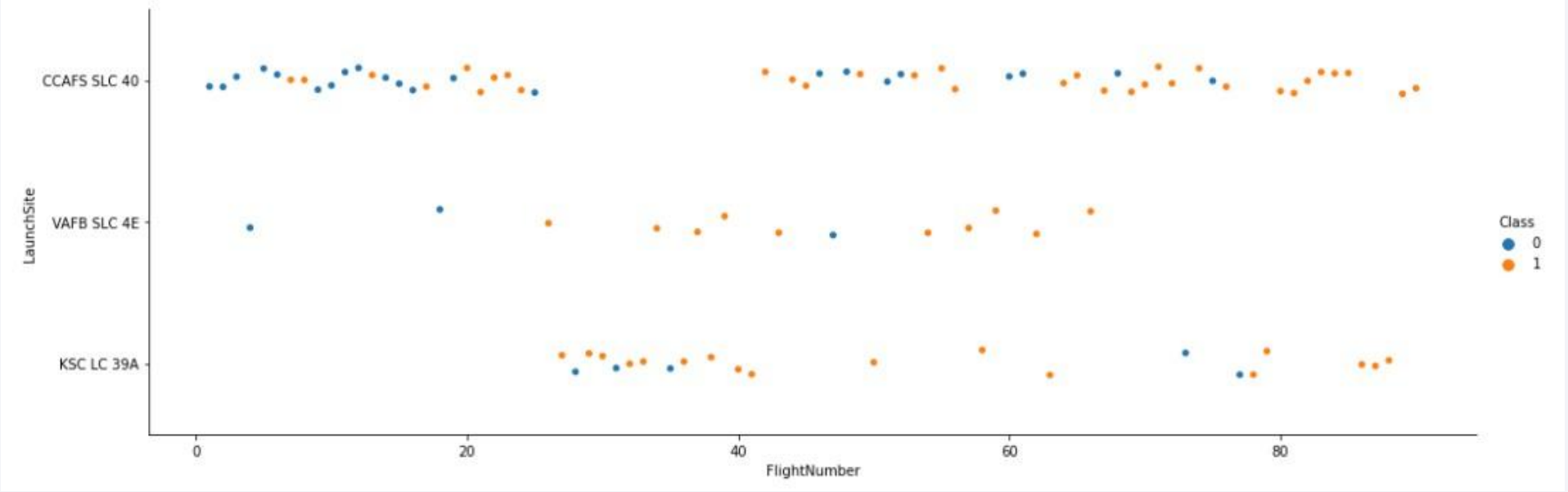
- With the public data available, using multiple feature analysis and machine learning algorithms, it was determined that whether or not a Falcon 9 rocket launch would result in a successful landing of the first stage could be predicted with greater than 83% accuracy
- Classification tree, logistic regression, and support vector machine models all performed similarly on test data with approximately 83% accuracy, while the k nearest neighbors model performed worst with a 78% accuracy score.

Section 2

Insights drawn from EDA

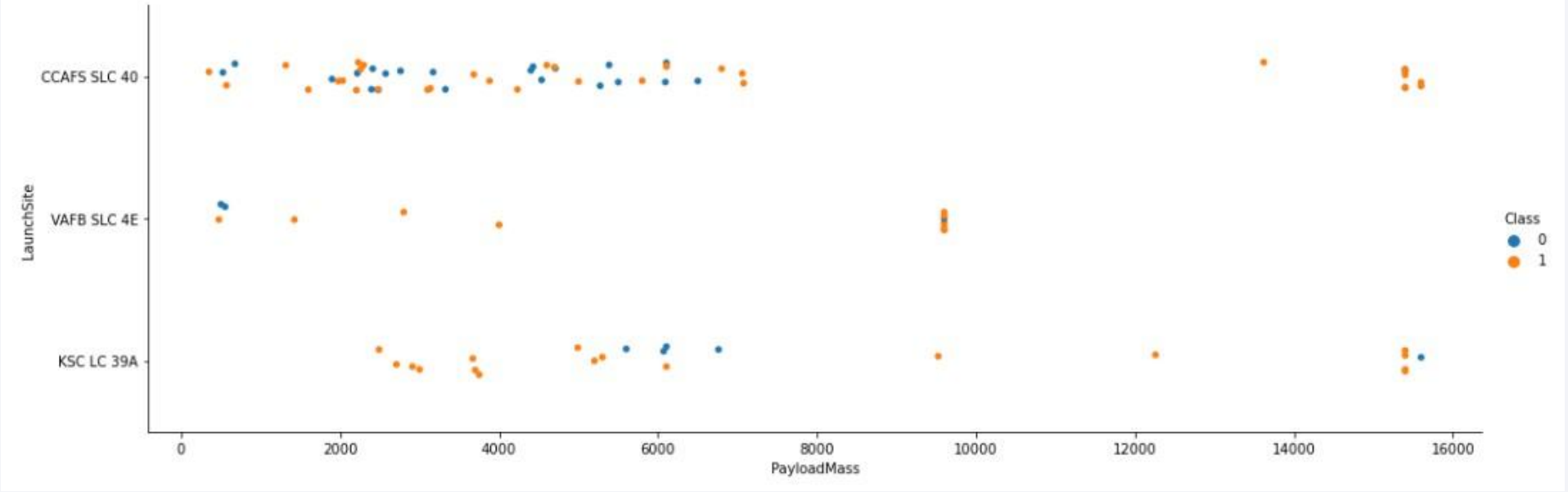


Flight Number vs. Launch Site



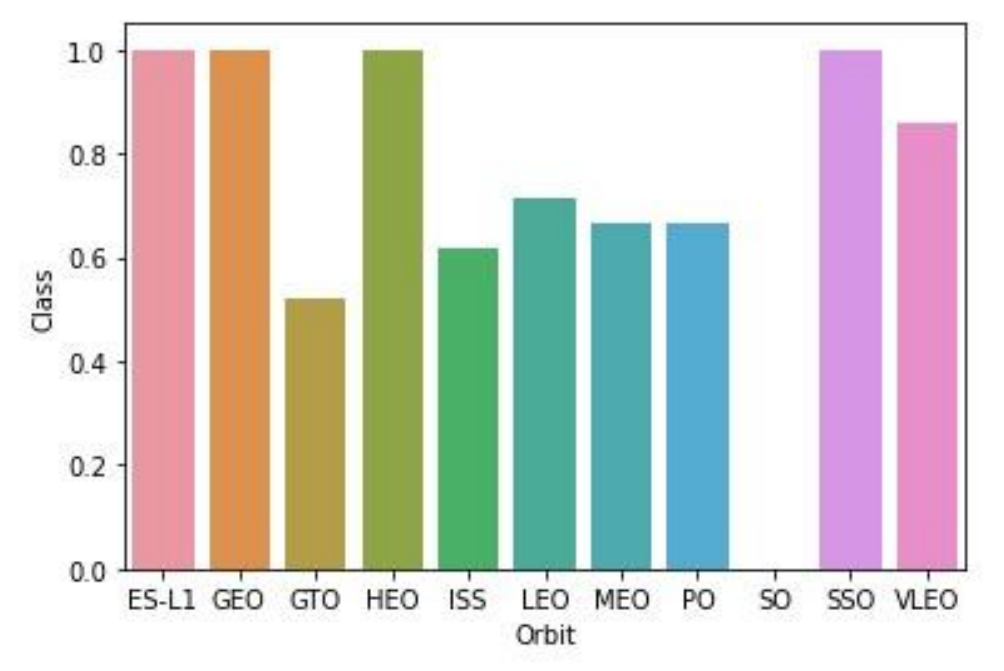
- The plot shows the flight number of each launch plotted against its launch site, as well as the landing outcome indicated by color.
- This can give us insight into the fact that earlier flights had much fewer successful landings

Payload vs. Launch Site



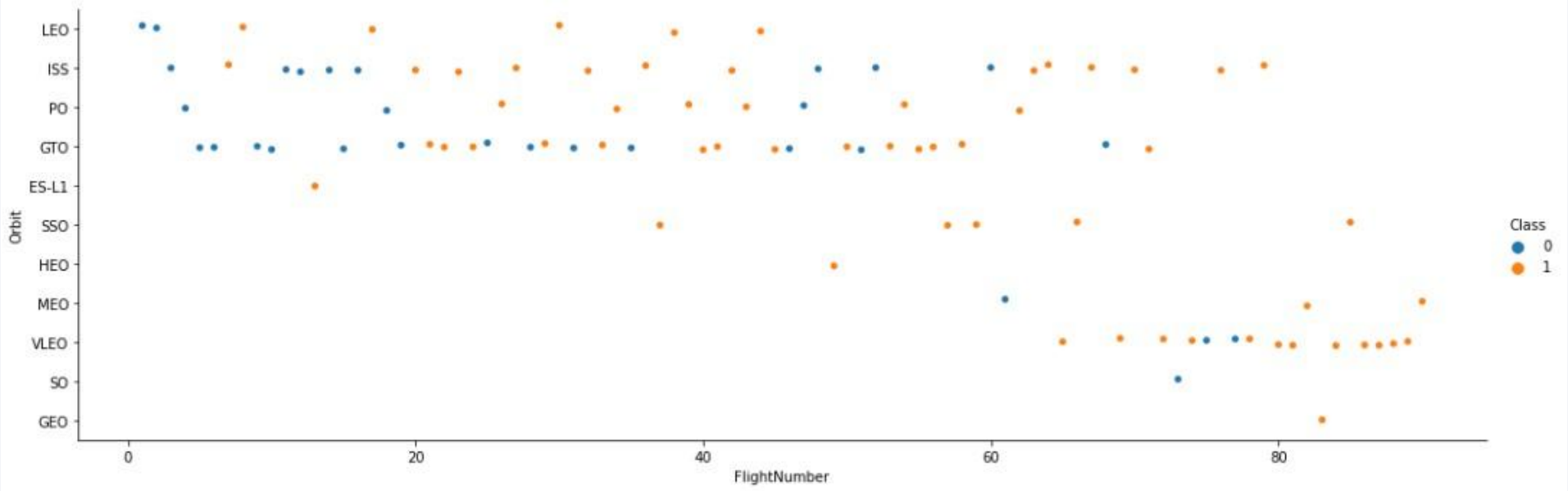
- The plot shows the payload mass of each launch plotted against its launch site, as well as the landing outcome indicated by color.
- This plots shows a trend that indicate that heavier payloads are more likely to result in a successful landing

Success Rate vs. Orbit Type



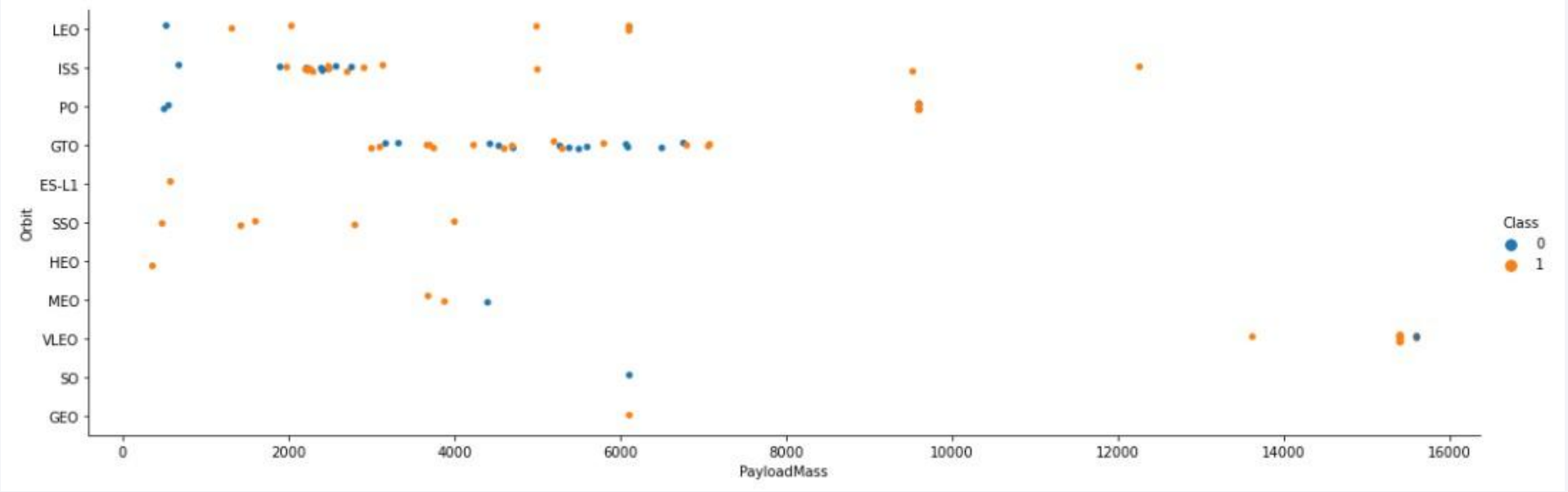
- The plot shows the orbit type of each launch plotted with its average rate of successful landings
- This can give us insight into the fact that certain orbit types have significantly higher success rates

Flight Number vs. Orbit Type



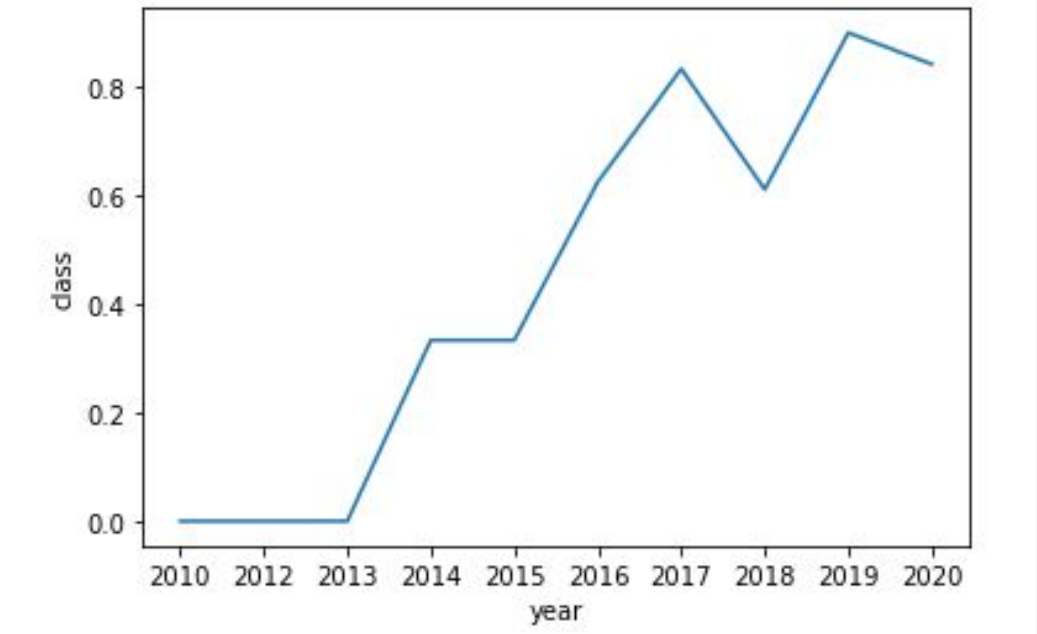
- The plot shows the flight number of each launch plotted against its orbit type as well as the landing outcome indicated by color.
- This plots shows how the types of orbit have varied over time, with earlier launches primarily targeting different orbits than later launches

Payload vs. Orbit Type



- The plot shows the payload mass of each launch plotted against its orbit type as well as the landing outcome indicated by color.

Launch Success Yearly Trend



- The plot shows the rate of successful landings by year, showing a general increase in success rate over time

All Launch Site Names

- SELECT unique(launch_site) FROM spacexdataset

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM spacexdataset WHERE launch_site LIKE 'CCA%' limit 5

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

- SELECT sum(payload_mass__kg_) FROM spacexdataset
  WHERE customer = ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

- SELECT avg(payload_mass__kg_) FROM spacexdataset
  WHERE booster_version = 'F9 v1.1'

1
2928

First Successful Ground Landing Date

- SELECT date FROM spacexdataset
  WHERE landing__outcome = ‘Success (ground pad)’
  ORDER BY date ASC LIMIT 1

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT booster_version, payload_mass__kg_ FROM spacexdataset WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- SELECT mission_outcome, count(*) as count FROM spacexdataset GROUP BY mission_outcome

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

Boosters Carried Maximum Payload

- SELECT distinct(booster_version) FROM spacexdataset
  WHERE payload_mass__kg_ =
    (SELECT max(payload_mass__kg_) FROM spacexdataset)

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- SELECT landing__outcome, booster_version, launch_site
  FROM spacexdataset
  WHERE landing__outcome = 'Failure (drone ship)' and year(date) = '2015'

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

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

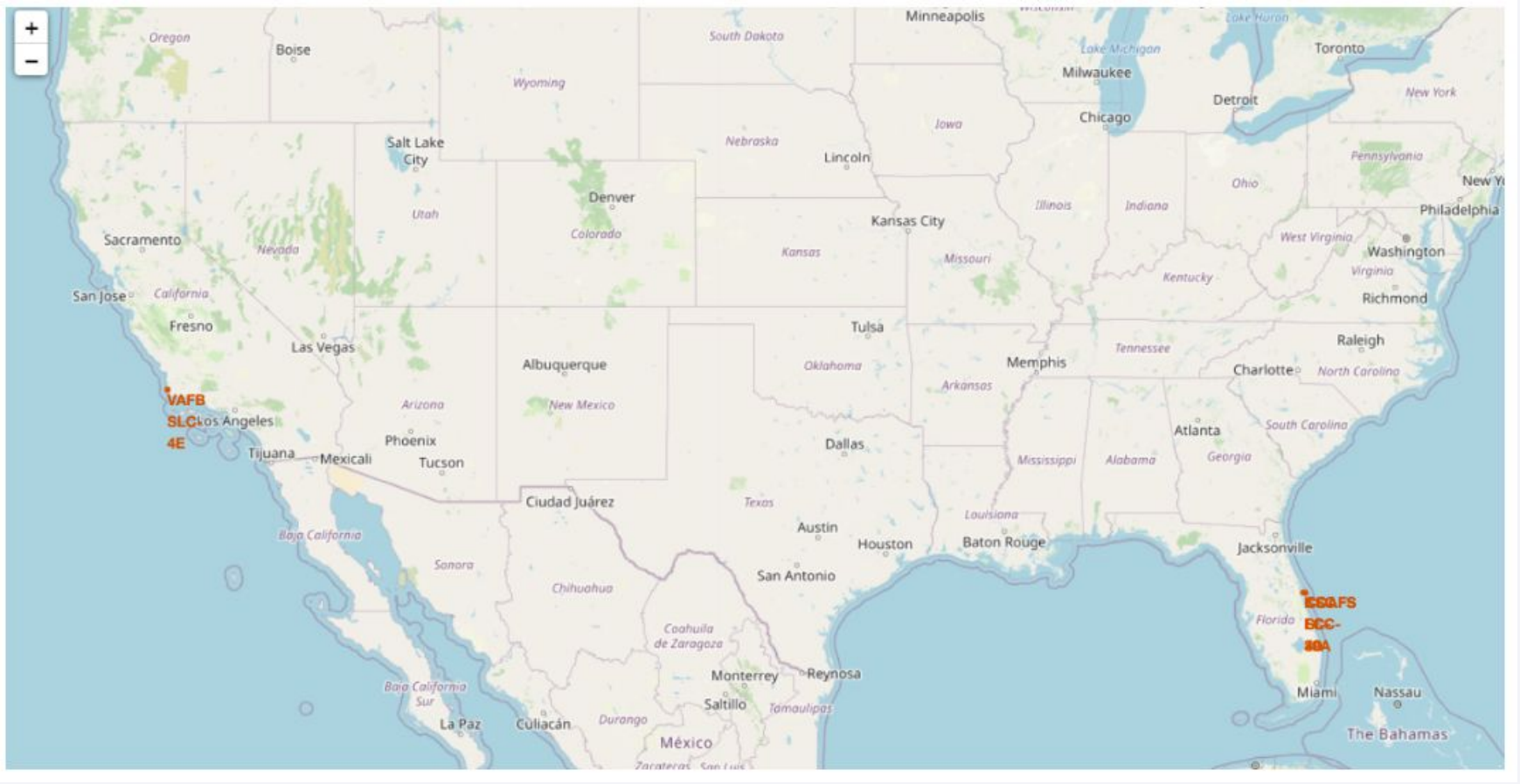
- SELECT landing__outcome, count(*) as count FROM spacexdataset
  WHERE date > '2010-06-04' and date < '2017-03-20'
  GROUP BY landing__outcome
  ORDER BY count desc

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


Section 3
Launch Sites
Proximities Analysis

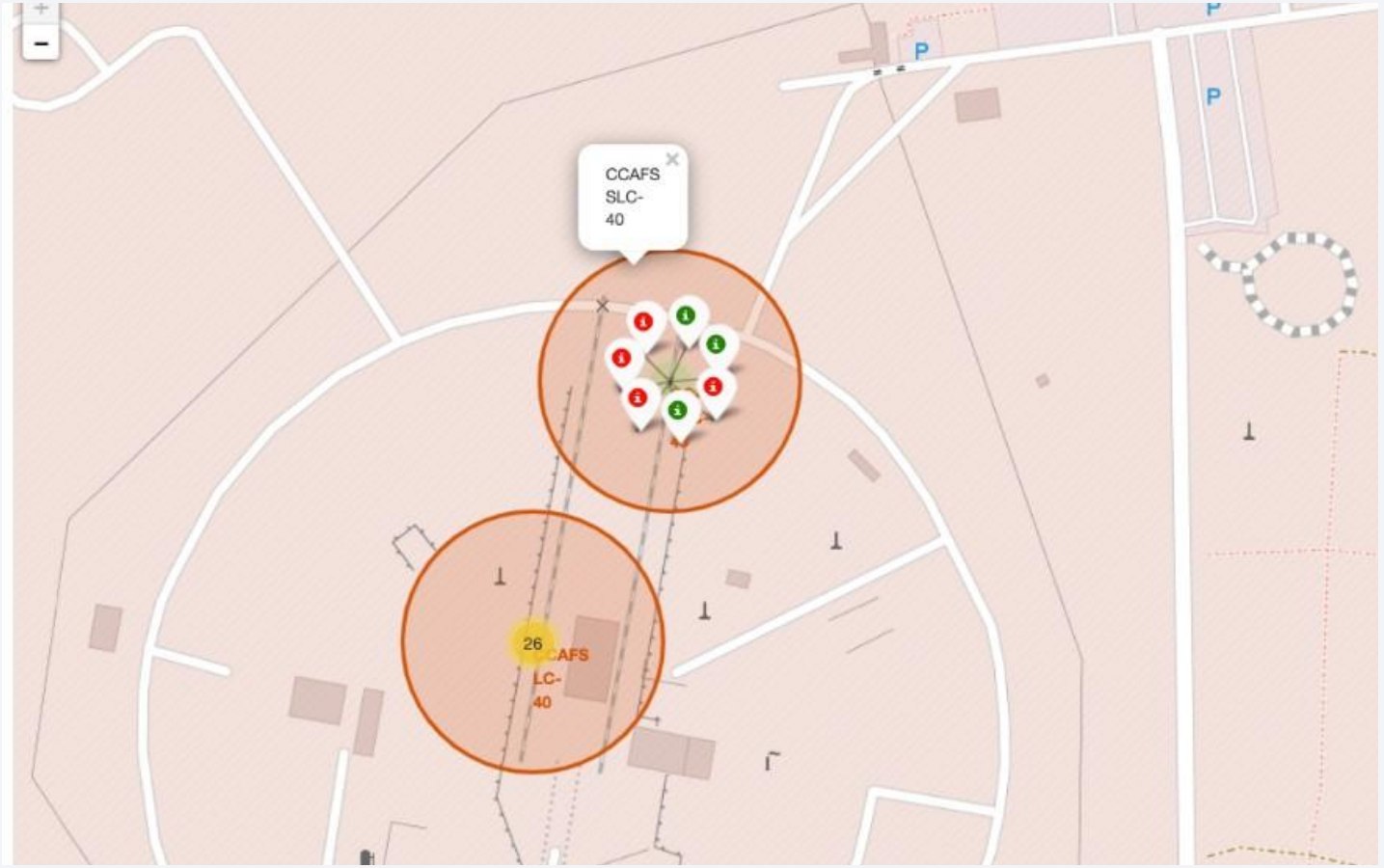
Falcon 9 Launch Sites

- This map shows the locations of the different sites for Falcon 9 launches



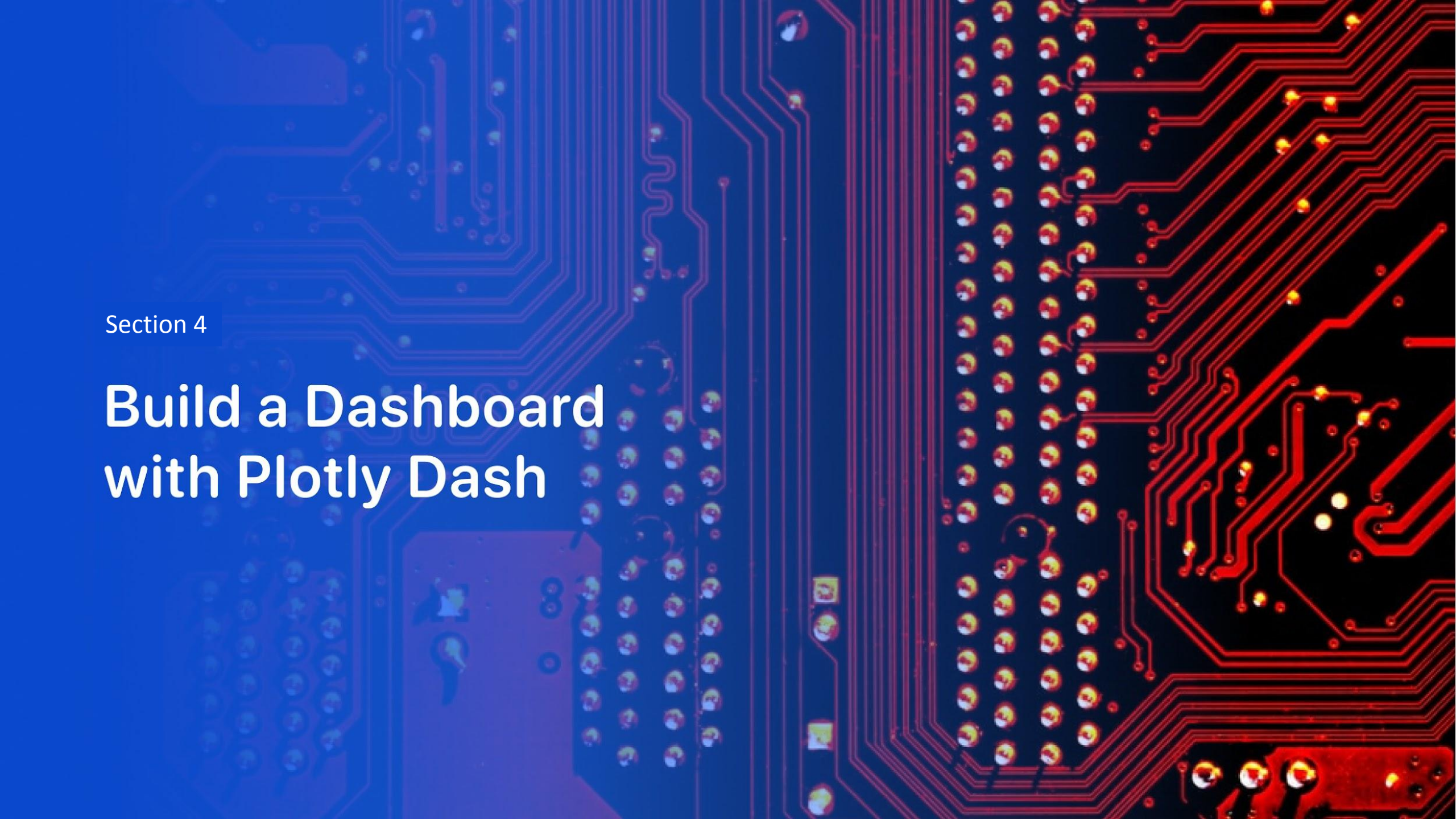
Labeled Clusters

- This map shows a close up of a launch site as well as color coded labels indicating the landing success of failure of individual launches

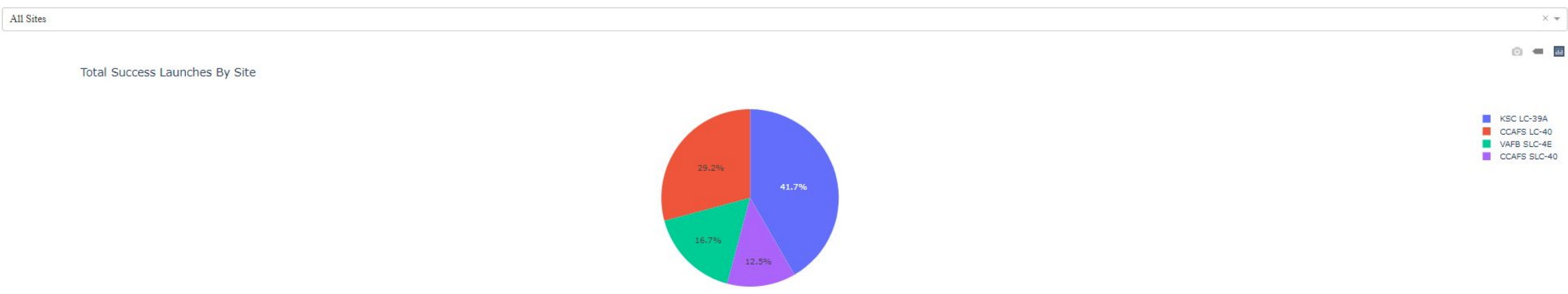


Proximate Distances

- This map shows the locations of features nearby to launch sites and allows for the calculation of distances

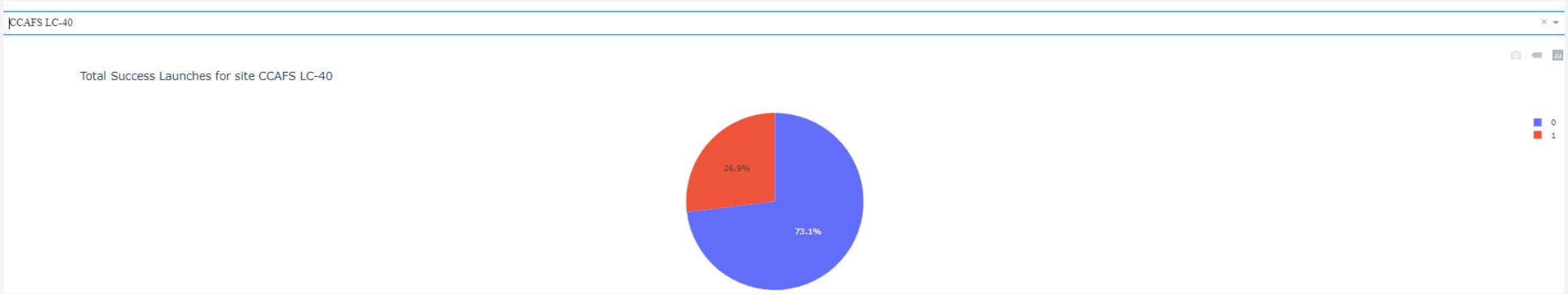

Section 4
Build a Dashboard with Plotly Dash

Overall Success Rates



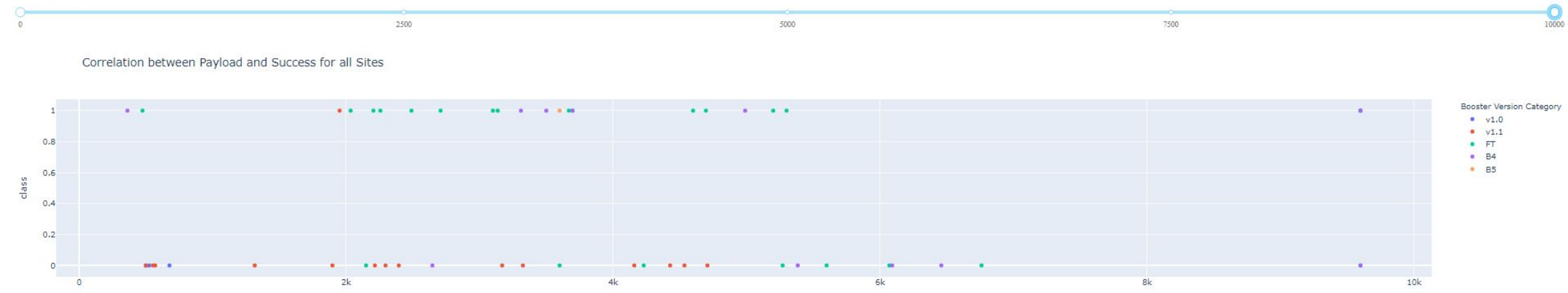
- This pie chart shows the number of total successful launches broken down by launch site, as well as the drop down menu to choose a launch site for further analysis

Individual Site Success Rate



- This pie chart shows the success rate of landings for an individual launch site, chosen from the drop down menu of all different sites

Impact of Payload on Success Rates



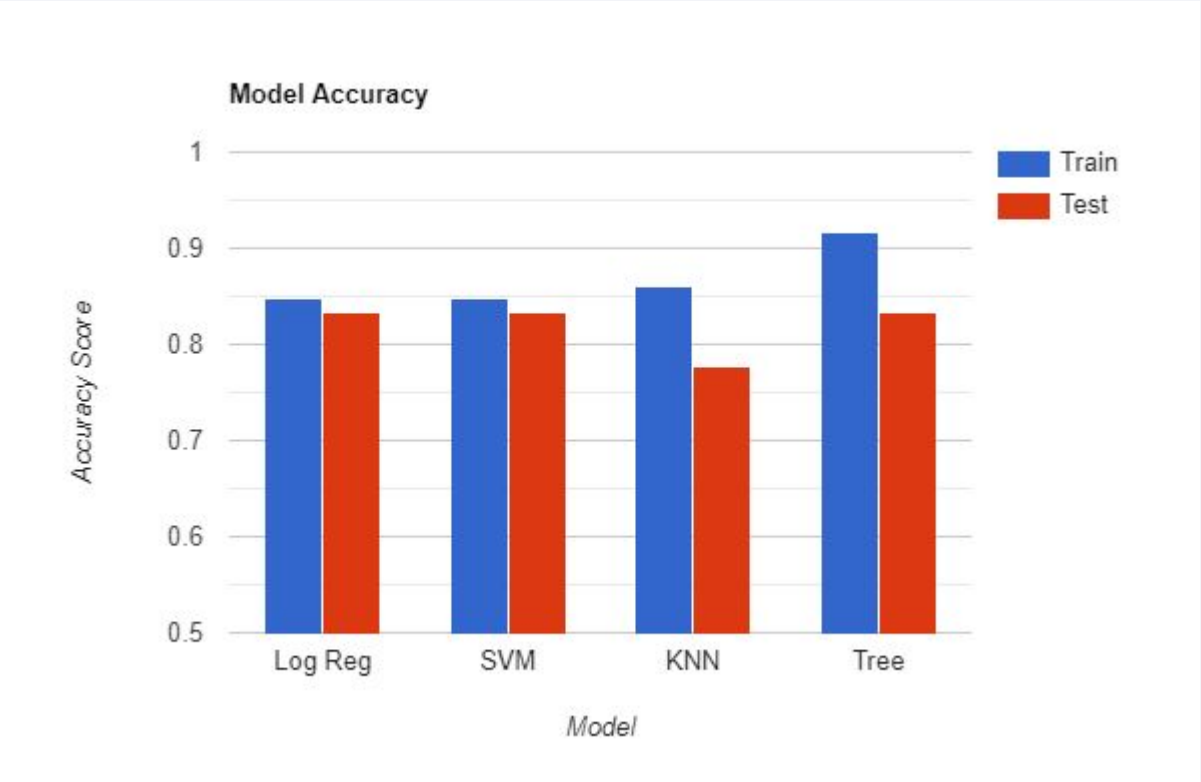
- This scatter plot shows the success rate of landings plotted against the payload mass of each launch

Section 5

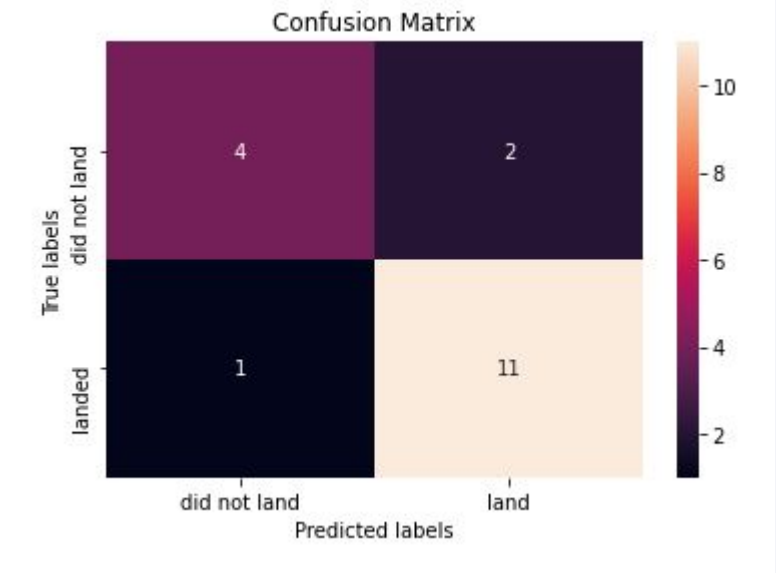
Predictive Analysis (Classification)

Classification Accuracy

- 3 out of 4 models had the same testing accuracy of 83.33%
- K nearest neighbors had the lowest test accuracy of 77.77%
- Classification tree had the highest overall accuracy between both training and testing sets



Confusion Matrix



- The decision tree classifier confusion matrix show a high rate of accuracy with only a very small amount of both false positives and false negatives

Conclusions

- It is possible to predict whether or not a Falcon 9 rocket will be landed successfully with greater than 83% accuracy
- This information allows for better estimation of the true costs of each rocket launch based on if the first stage will be able to be reused
- Launch variables can be used by competing companies to provide evidence of higher than estimated costs if it is predicted that there will not be a successful landing, which is how SpaceX keeps its costs lower

Thank you!