IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<William Chaudoin>
<04/12/23>

Outline

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

Executive Summary

- Methods used for this project
  - Data Collection through webscrapping and API
  - Explanatory Data Analysis with data visualization, interactive mapping, and machine learning
  - Various machine learning models such as KNN, SVM, LogisticReg, and Decision Trees
- Results Summary
  - Data Collected through methods revealed useful and valuable information for analysis
  - Data visualization techniques provided unique insights into SpaceX mission outcomes
  - Machine Learning provided accurate prediction models with some performing better than others.

Introduction

- The purpose of this project is to analyze historical launch data associated with SpaceX to determine success rates through various machine learning models while also using data visualization to learn more about the collected dataset.
- This is performed to evaluate whether establishing a new company "Space Y" is a viable option.
- Some information I seek to find is:
  - The success rate dealing with launches.
  - The number of successful missions
  - The number of successful landings along with the number of failures.
- With more information I can determine whether attempting to land the first stage will succeed.

Introduction

- Models used include Logistic Regression, Support Vector Machines, Decision Trees, and K Nearest Neighbor.
- I used mapping techniques to learn more about launch locations and which have the highest success rates.
- Launch data was graphed with various tools.
- An interactive dashboard was created to learn more about launch sites.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was collected by using webscrapping via Wikipedia and through the SpaceX API
- Performed data wrangling
  - Data was cleaned and analyzed to find a way to enrich it.
  - This was performed by creating new variables through grouping, sorting, and based on landing outcomes.
- Performed exploratory data analysis (EDA) using visualization and SQL

Methodology

- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Utilized multiple supervised and unsupervised machine learning models
  - Split obtained data into training and testing sets to refine models
  - Tested model accuracy to find best predictor.

Data Collection

- Data was collected from two locations:
  - The SpaceX API: https://api.spacexdata.com/v4/rockets/
  - Via Webscrapping Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Contact and request information from API to establish connection
- Import dataset and prepare for analysis.
- Handle any data type issues, missing values, and create new labels for further analysis.
- Github: [Data Collection notebook]

Connect to API

- Request data from API
- Allocate data to variable

Clean Data

- Filter and Sort data
- Create new variables and labels

Analyze and Identify remaining issues

- Fix missing values
- Ensure accuracy and correct data types

Data Collection - Scraping

- Request website information through HTML, identify tables and column headings, use information to create dataframe.
- Github: Webscrapping notebook

Establish Connection
- Request information from website

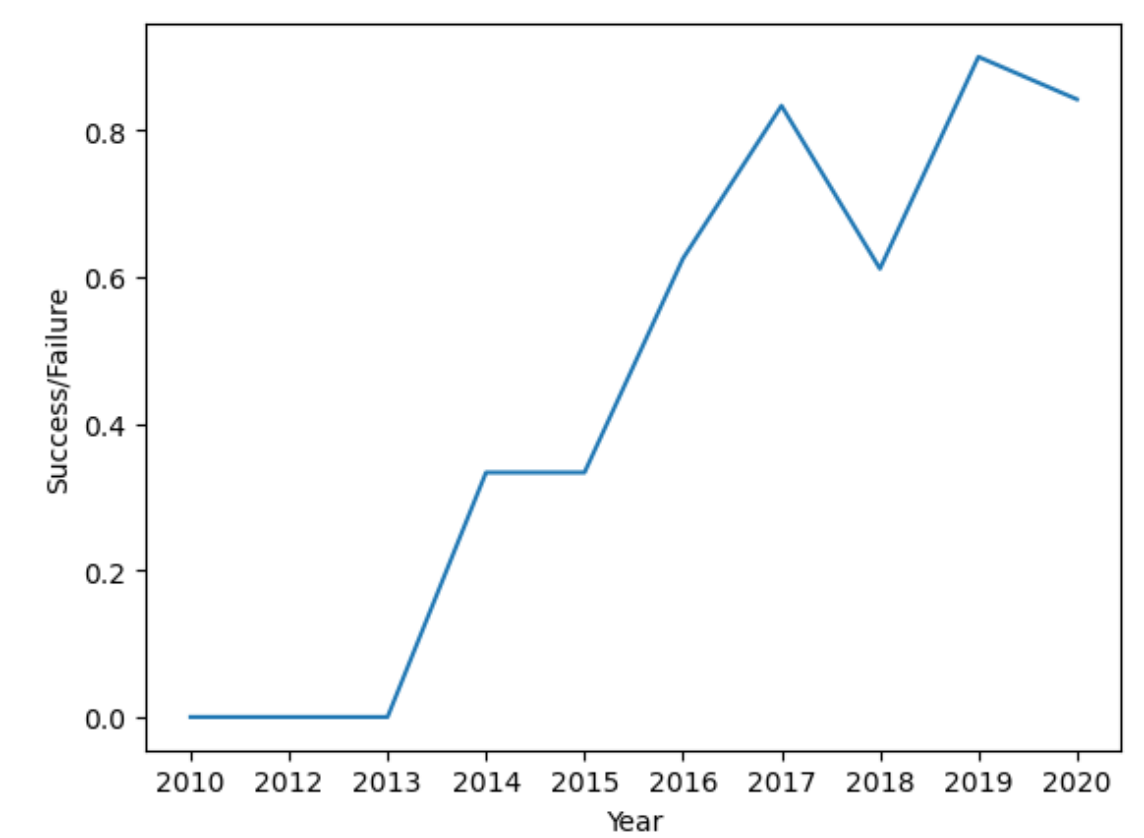
Gather HTML data
- Get HTML data and identify table columns

Create Dataframe
- Create dataframe by parsing HTML tables

Data Wrangling

- Explanatory Data Analysis was performed to create a number of statistical outcomes.
- The values for Launch Site, Orbit, and Outcome was determined, with outcome being attached to a new variable Landing Outcome.
- Landing Outcome was used to form Landing Class which was used to determine the success rate of a mission.
- Github: Data Wrangling

EDA with Data Visualization



- Data was visualized using various scatterplots to graph Orbit, Flight Number, and Payload Mass to see the relationships and determine further context.
- These plots were color coded by whether a landing was successful or not to grasp more knowledge regarding success rates.
- Success rate was graphed over time via line chart (pictured)
- Github: Data Visualization Notebook

EDA with SQL

- I utilized SQL magic to run various queries on the dataset.
  - A query to id distinct launch sites
  - Total payload on boosters contracted through NASA
  - Determine the average payload for Falcon 9 first stage boosters
  - The earliest date where a successful ground pad landing occurred.
  - Where a booster carrying a payload between 4000 and 6000 kg successfully landed on a drone ship
  - Total number of mission outcomes and by result
  - The names of the boosters that carried the maximum payload
  - For 2015, all failed landings on a drone ship, with their booster version and launch site.
  - All landing outcomes between June 4, 2010 and March 20, 2017 by descending order.
- Github: EDA SQL

Build an Interactive Map with Folium

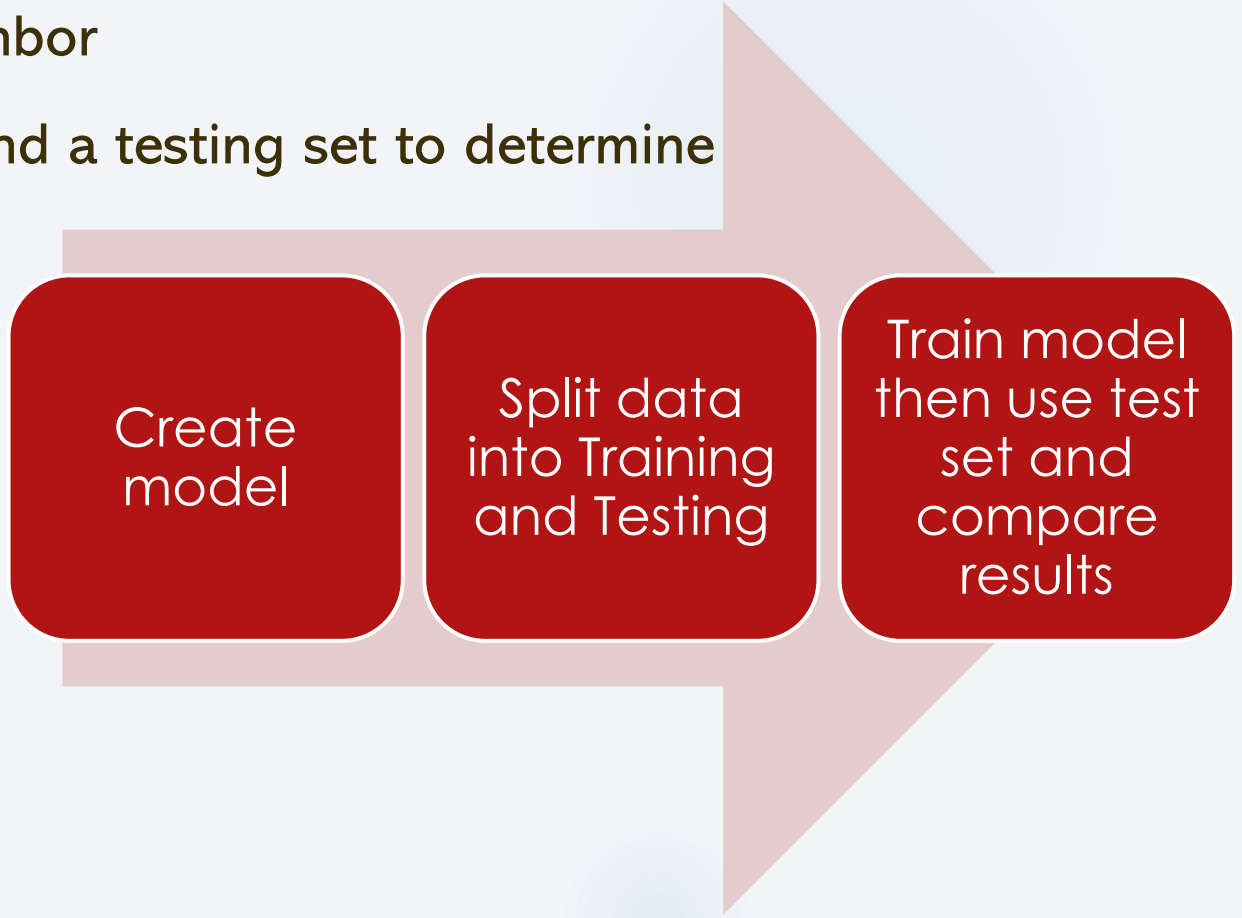
- Utilized folium to create map objects such as markers, circles, lines, etc.
- Objects added using longitude and latitude data in order to determine distances from landmarks and to determine where launch sites were located.
- Aided in understanding where the most successful launches occurred.
- Folium Notebook

Build a Dashboard with Plotly Dash

- Utilized Plotly Dash to create a dashboard to display launch site data with a pie chart and a scatter plot that could be interacted with via drop down menus.
- Pie chart depicts launch site and success rates for each while the scatter plot depicts launch site and payload mass.
- Plotly Dash Code

Predictive Analysis (Classification)

- Four models were used in this analysis: Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbor
- The dataset was split between a training set and a testing set to determine the most accurate model.
- Machine Learning Notebook



Results

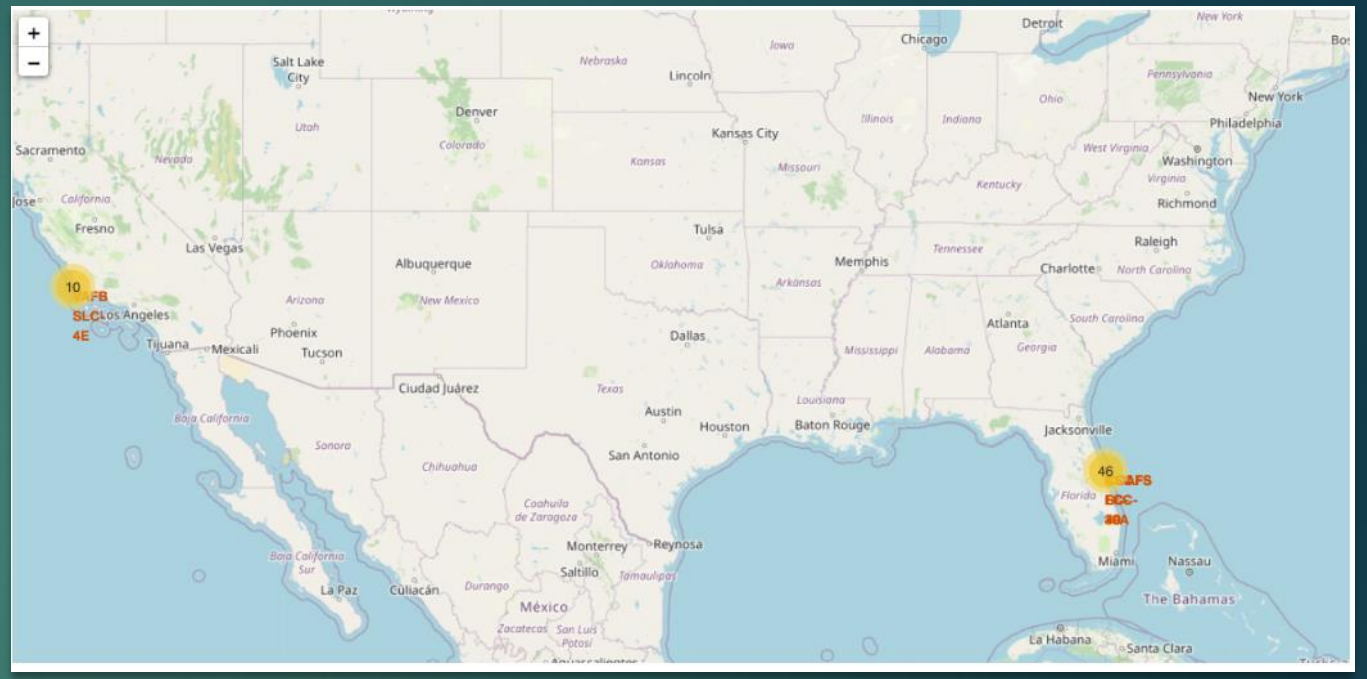
- SpaceX uses 4 different launch sites on the East and West Coast with launches done for NASA and SpaceX purposes
- Total Payload for all NASA (CRS) contracting flights is 45,596 kg
- The Average Payload of a Falcon 9 version 1.1 booster is 2,928 kg
- The earliest launch with a successful ground pad landing was on December 22, 2012.
- There are 4 boosters that have successfully landed on a drone ship and carried between 4000 and 6000 kgs of payload
- There have been a total of 101 flights, with 1 failure and 100 varying levels of success flights.
- Interactive analytics demo in screenshots
- Predictive analysis results

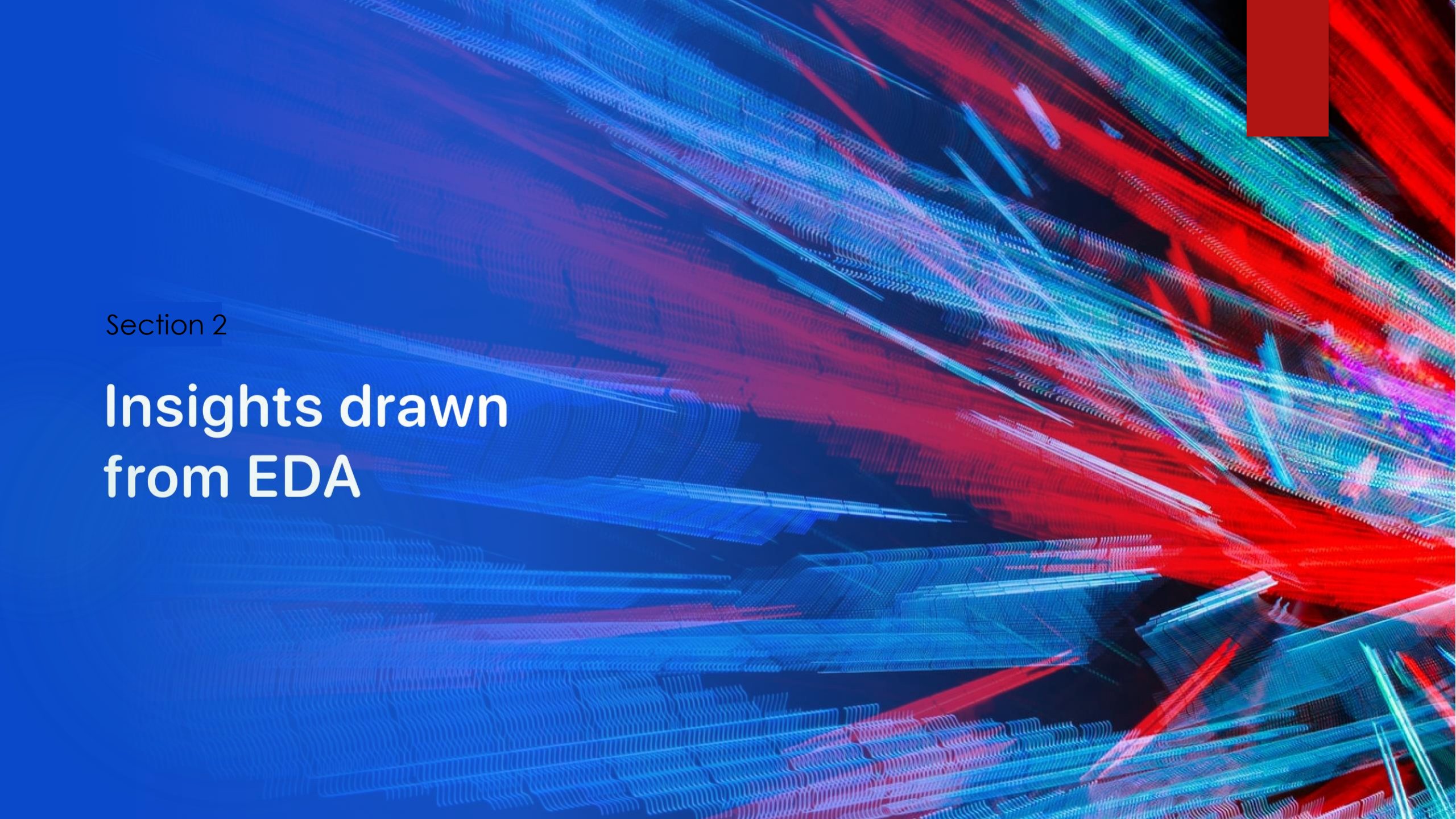
Results

- The majority of launches occur on the East coast
- Using the testing set data and the built-in score function found that the Logistic Regression, SVM, and KNN models all performed roughly the same with Decision Trees performing the worst.
- Using the built-in best_score function found conflicting results though, with the Decision Tree model performing better than the other three.

Results

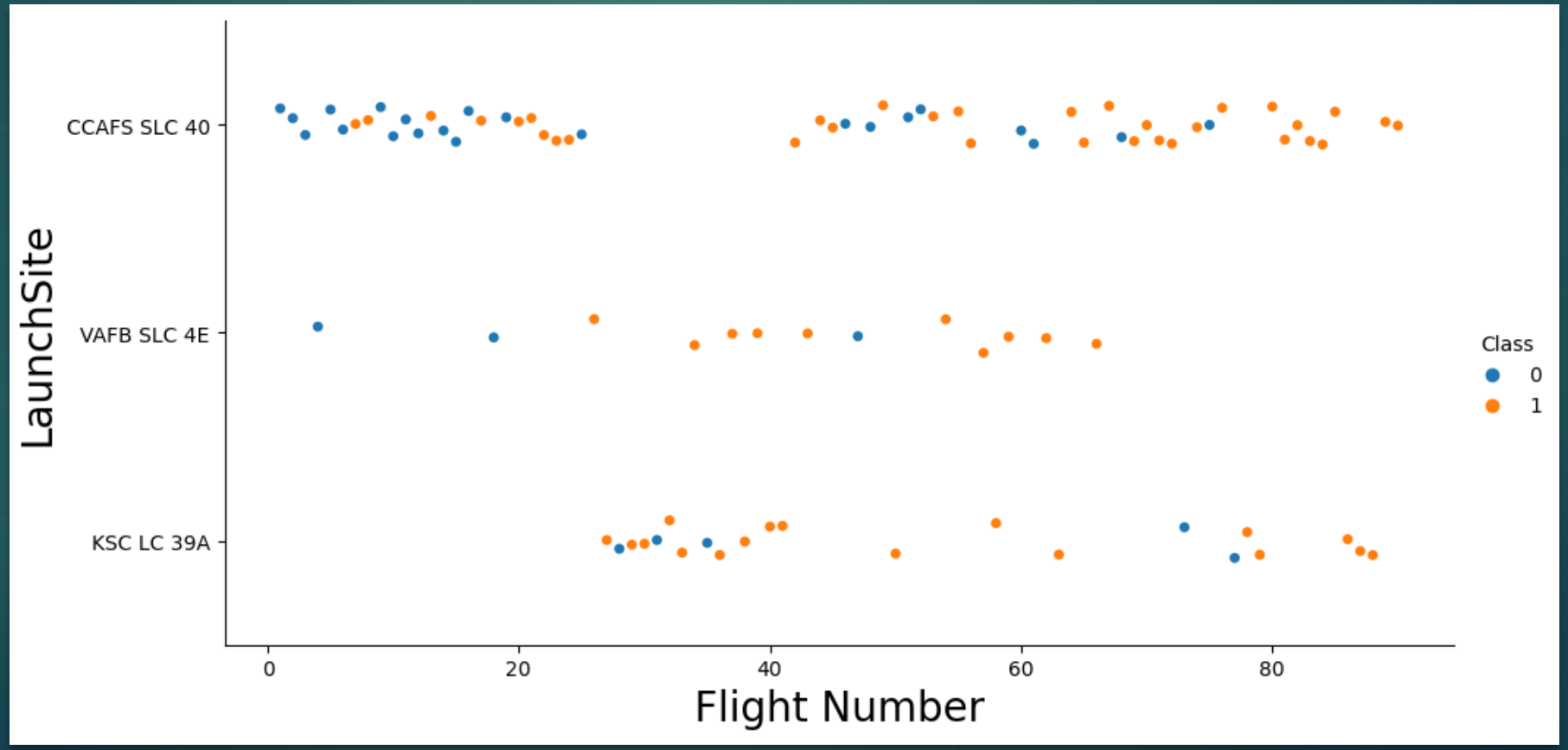
- Using interactive visual analytics aided in understanding the launch site data. Specifically, the location of the majority of sites and their distance to surrounding landmarks.




Section 2
Insights drawn
from EDA

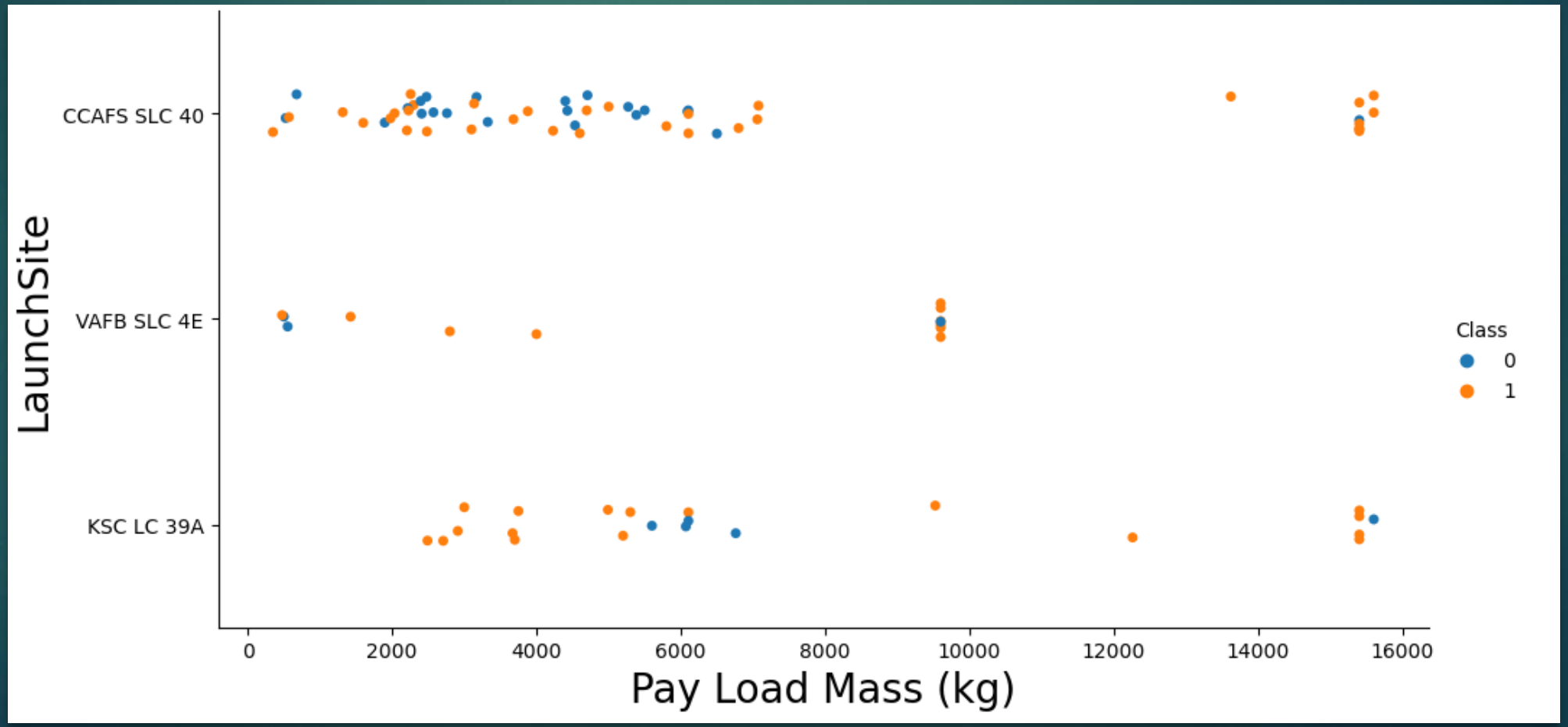
Flight Number vs. Launch Site

- Per the scatterplot below, the success rate of flights increased over time as more flights occurred.
- The site with the most flights is CCAFS SLC 40 with a higher success rate than others included.



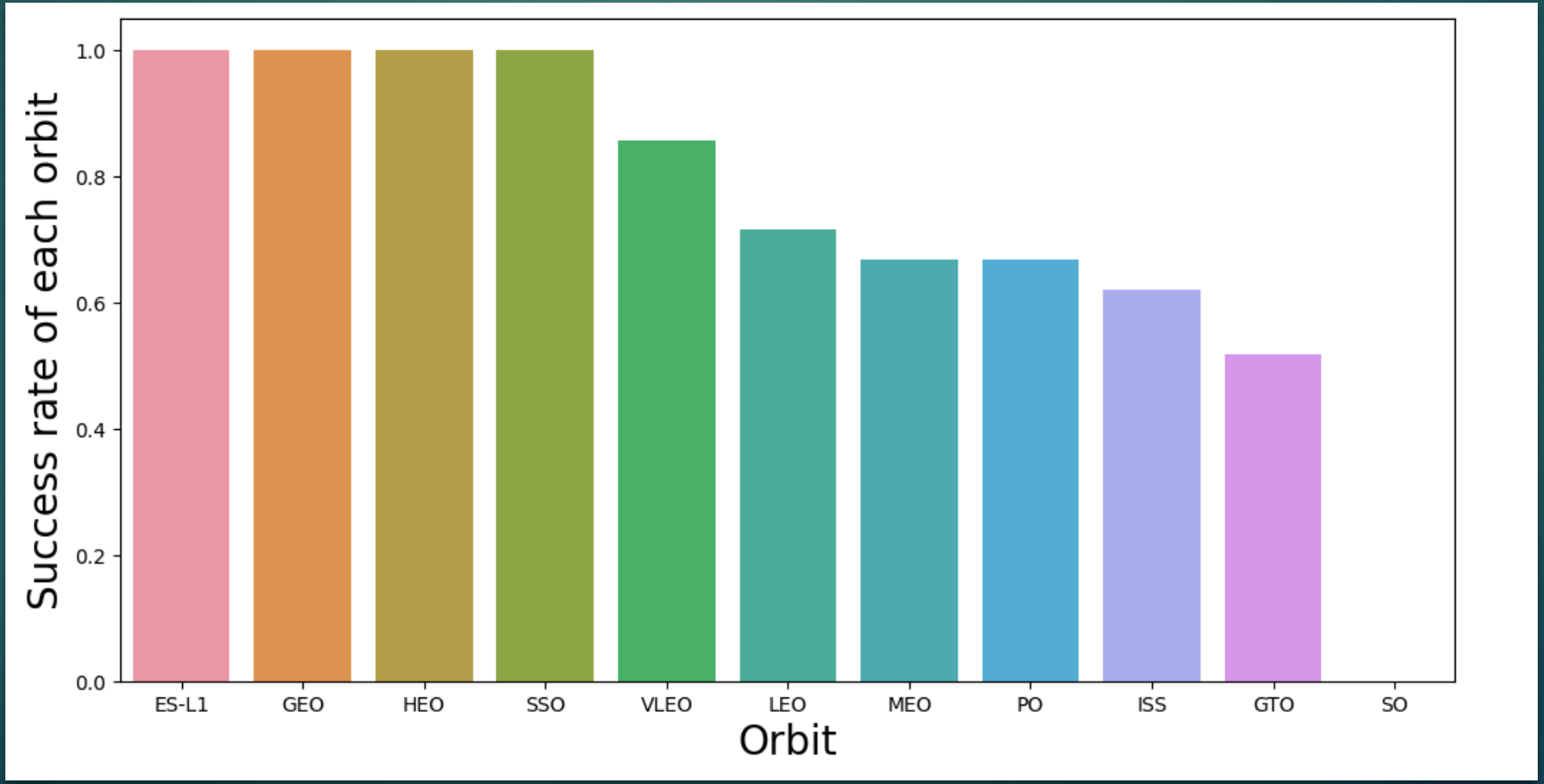
Payload vs. Launch Site

- Payload mass tends to be less than 8000 kg for most flights with CCAFS SLC 40 having the majority of flights with it and KSC LC 39A having the highest payload launches.



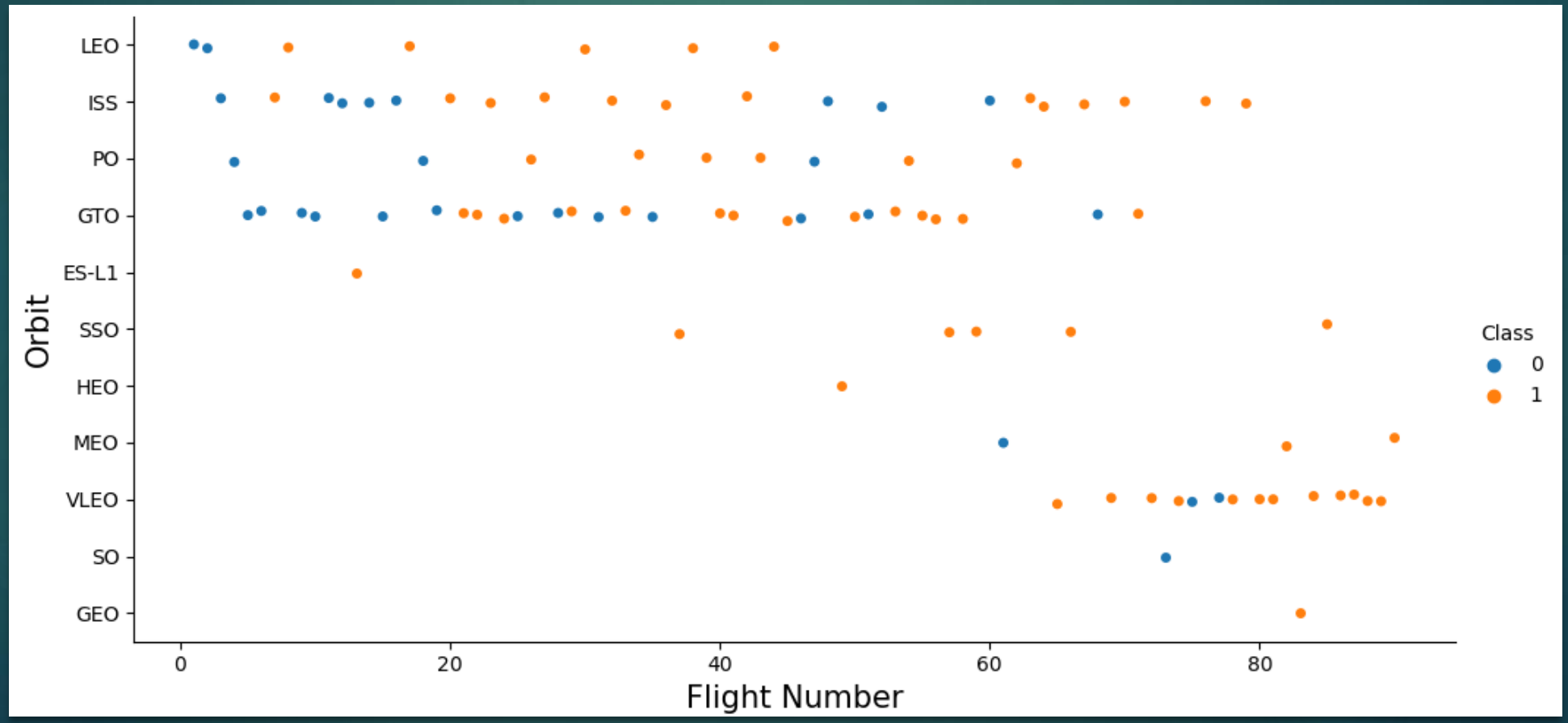
Success Rate vs. Orbit Type

- The orbit types with the highest success rate for each orbit are:
  - ES-L1, GEO, HEO, and SSO with VELO following the pack.



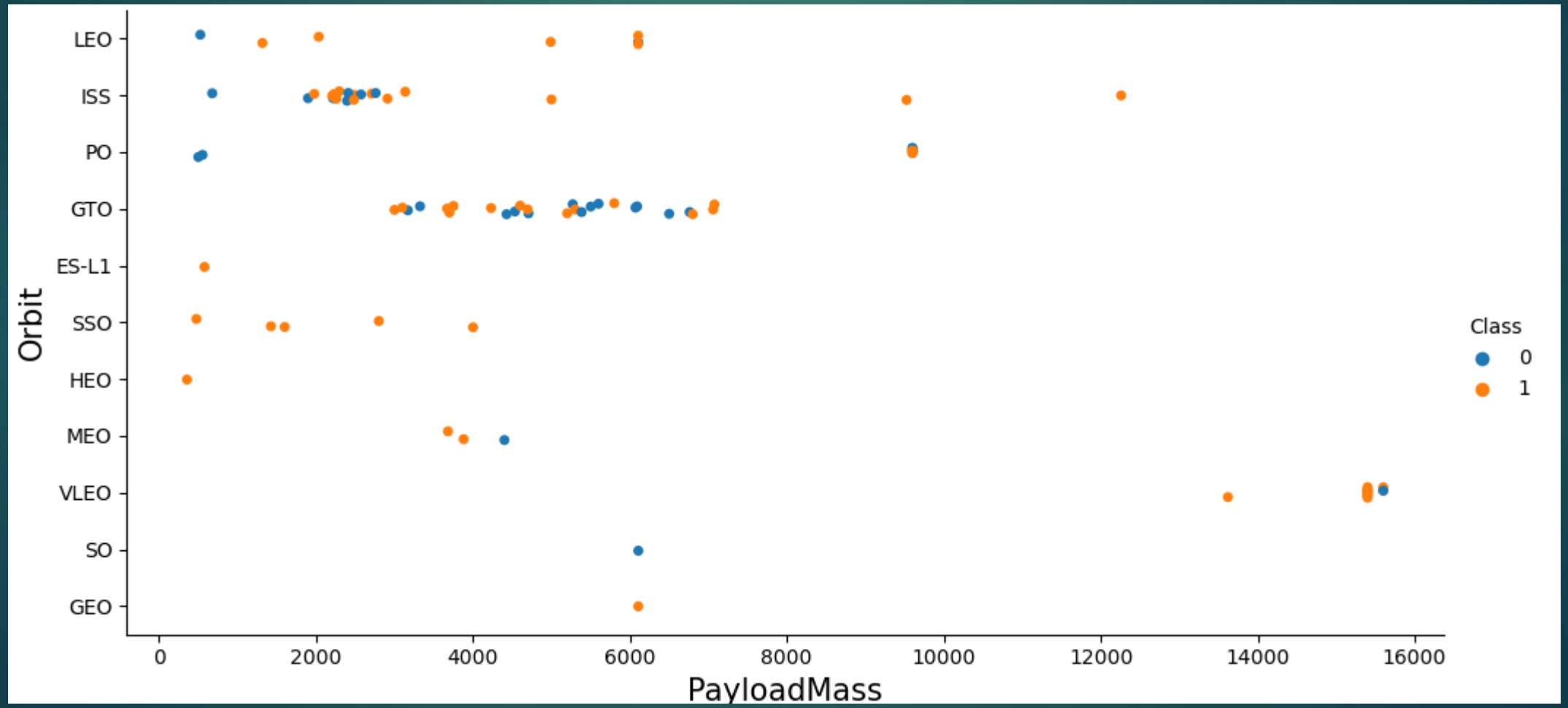
Flight Number vs. Orbit Type

- Different orbits have been attempted over time. As the flight number increased, new orbits were attempted while others were continued or discontinued.



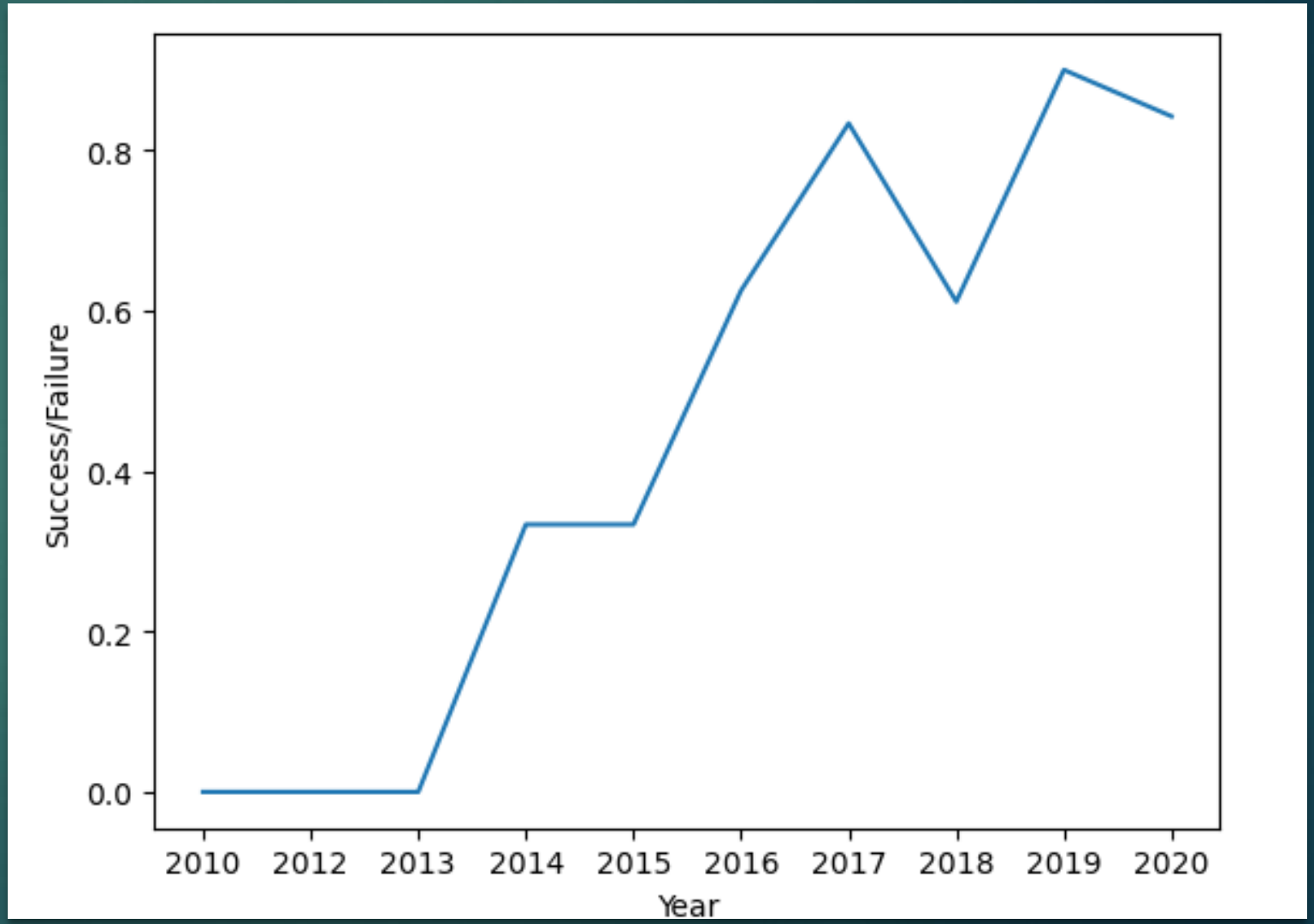
Payload vs. Orbit Type

- GTO was used for mid-range payload mass launches while VELO was used for flights with the highest payload masses.
- ISS was used for several lower payload mass flights.



Launch Success Yearly Trend

- Success rates over time tend to increase with a slight stabilization or drop in success during specific periods.
- Rates increased from 2013 to 2017 before seeing a decline in 2018 and then increasing in 2019.



All Launch Site Names

- There were four launch sites used by SpaceX for flights contracted through NASA (CRS).
- These sites were located on the West Coast and East Coast of the United States.

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

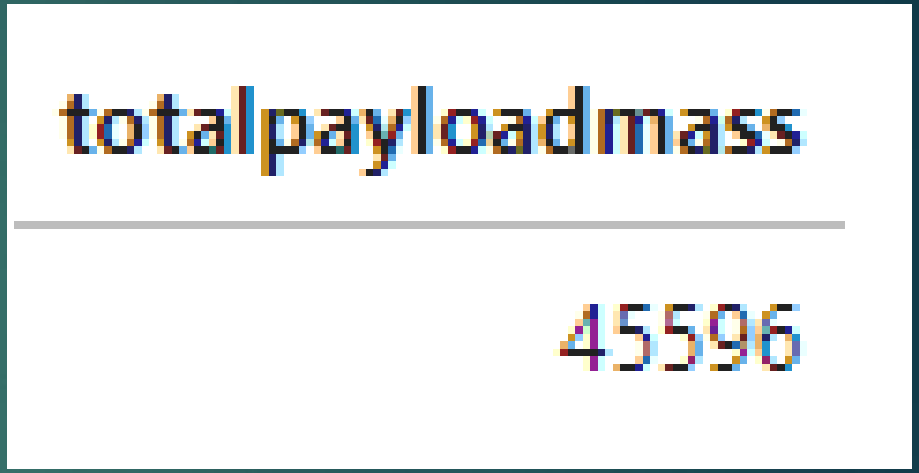
Launch Site Names Begin with 'CCA'

- There were 5 flights using the CCAFS LC-40 launch site contracted with NASA (CRS)

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The total payload mass for flights contracted with NASA performed by SpaceX was 45,596 kg.



Average Payload Mass by F9 v1.1

- The average payload for a Falcon 9 V1.1 booster is 2,928 kg.

meanpayloadmass
2928

First Successful Ground Landing Date

- The first successful ground pad landing for a flight occurred on December 22, 2015.

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 4 boosters have successfully landed on a drone ship after carrying a payload between 4000 and 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

- There have been 101 total flights with:
  - 1 failure
  - 99 successes
  - 1 success with an unknown payload status

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

Boosters Carried Maximum Payload

- a maximum payload
each was not checked
could discover this.

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

- There were 2 failed booster landings on a drone ship in 2015.
  - One occurred on January 10 and the other occurred on April 14, with both taking off from the CCAFS LC-40 launch site and both were Falcon 9 v1.1 boosters.

DATE	booster_version	launch_site	landing_outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Grand Marshal
- Of the 31 flights, there w
  - 3 successful ground pad l
  - 5 successful drone ship land
  - 5 failed drone ship landings
  - 2 failed parachute deployme
  - 3 controlled landings into the
  - 2 uncontrolled landings into
  - 10 flights with no attempted
  - 1 classified as precluded or

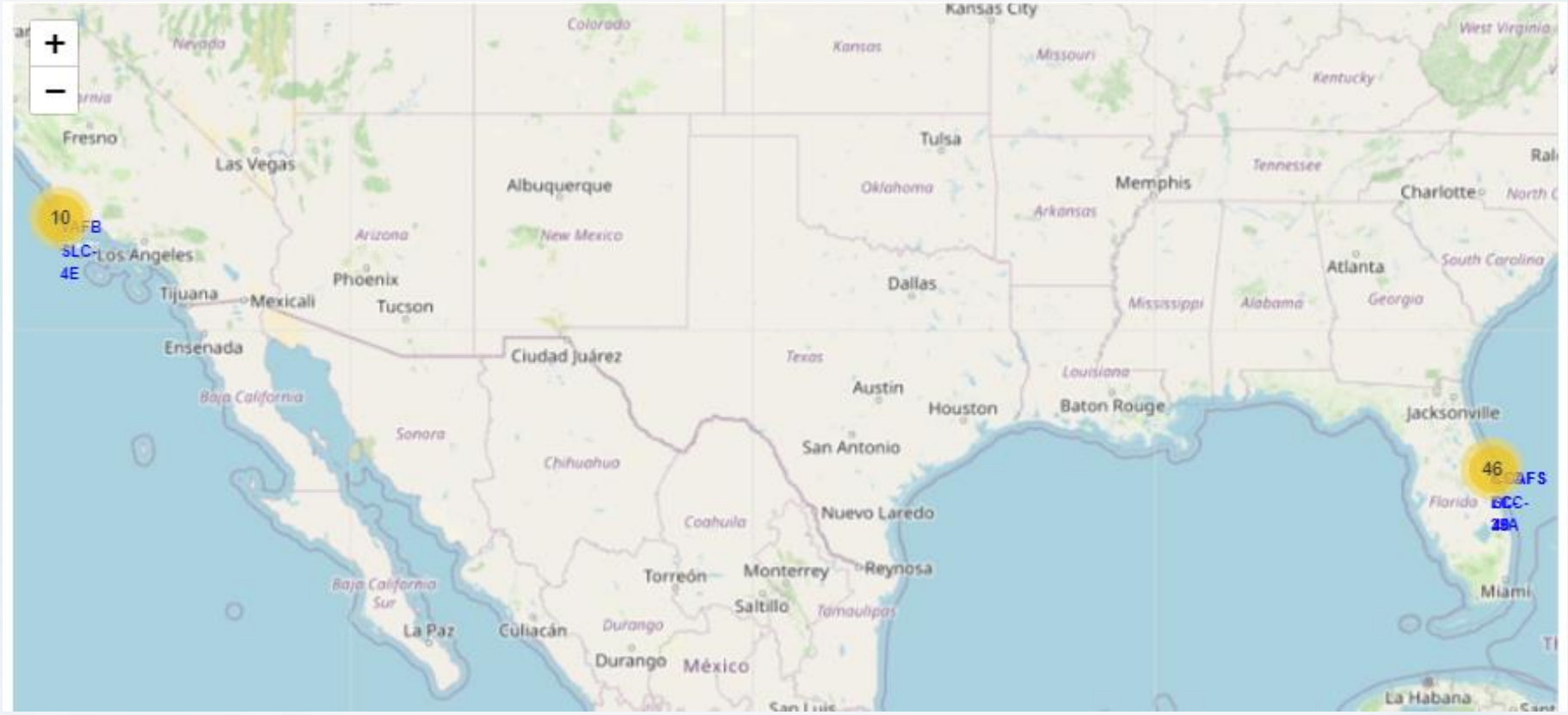
DATE	landing_outcome
2017-03-16	No attempt
2017-02-19	Success (ground pad)
2017-01-14	Success (drone ship)
2016-08-14	Success (drone ship)
2016-07-18	Success (ground pad)
2016-06-15	Failure (drone ship)
2016-05-27	Success (drone ship)
2016-05-06	Success (drone ship)
2016-04-08	Success (drone ship)
2016-03-04	Failure (drone ship)
2016-01-17	Failure (drone ship)
2015-12-22	Success (ground pad)
2015-06-28	Precluded (drone ship)
2015-04-27	No attempt
2015-04-14	Failure (drone ship)
2015-03-02	No attempt
2015-02-11	Controlled (ocean)
2015-01-10	Failure (drone ship)
2014-09-21	Uncontrolled (ocean)
2014-09-07	No attempt
2014-08-05	No attempt
2014-07-14	Controlled (ocean)
2014-04-18	Controlled (ocean)
2014-01-06	No attempt
2013-12-03	No attempt
2013-09-29	Uncontrolled (ocean)
2013-03-01	No attempt
2012-10-08	No attempt
2012-05-22	No attempt
2010-12-08	Failure (parachute)
2010-06-04	Failure (parachute)

Section 3

Launch Sites Proximities Analysis

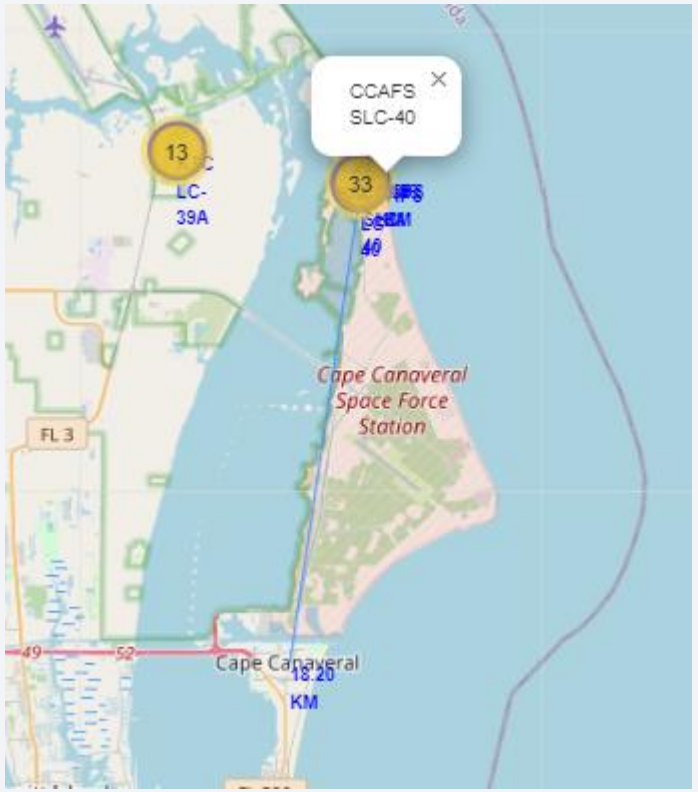
Geographic Visualizations

- There were 46 launches on the east coast and 10 launches on the west coast.



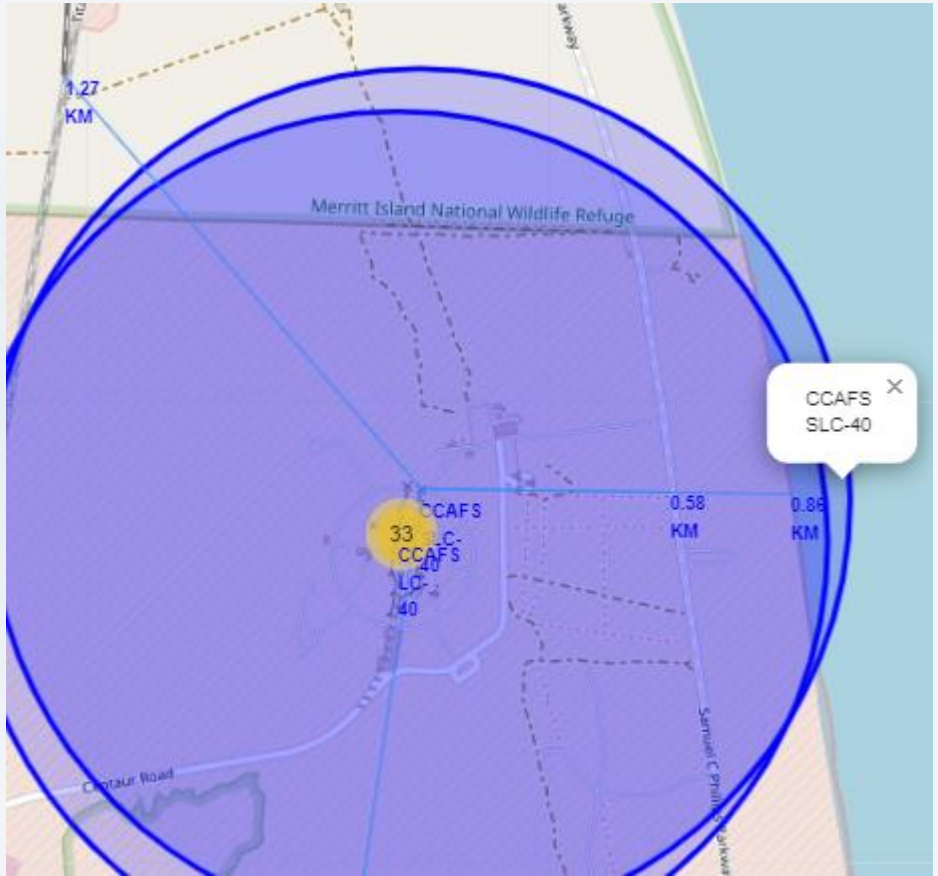
Geographic Visualizations

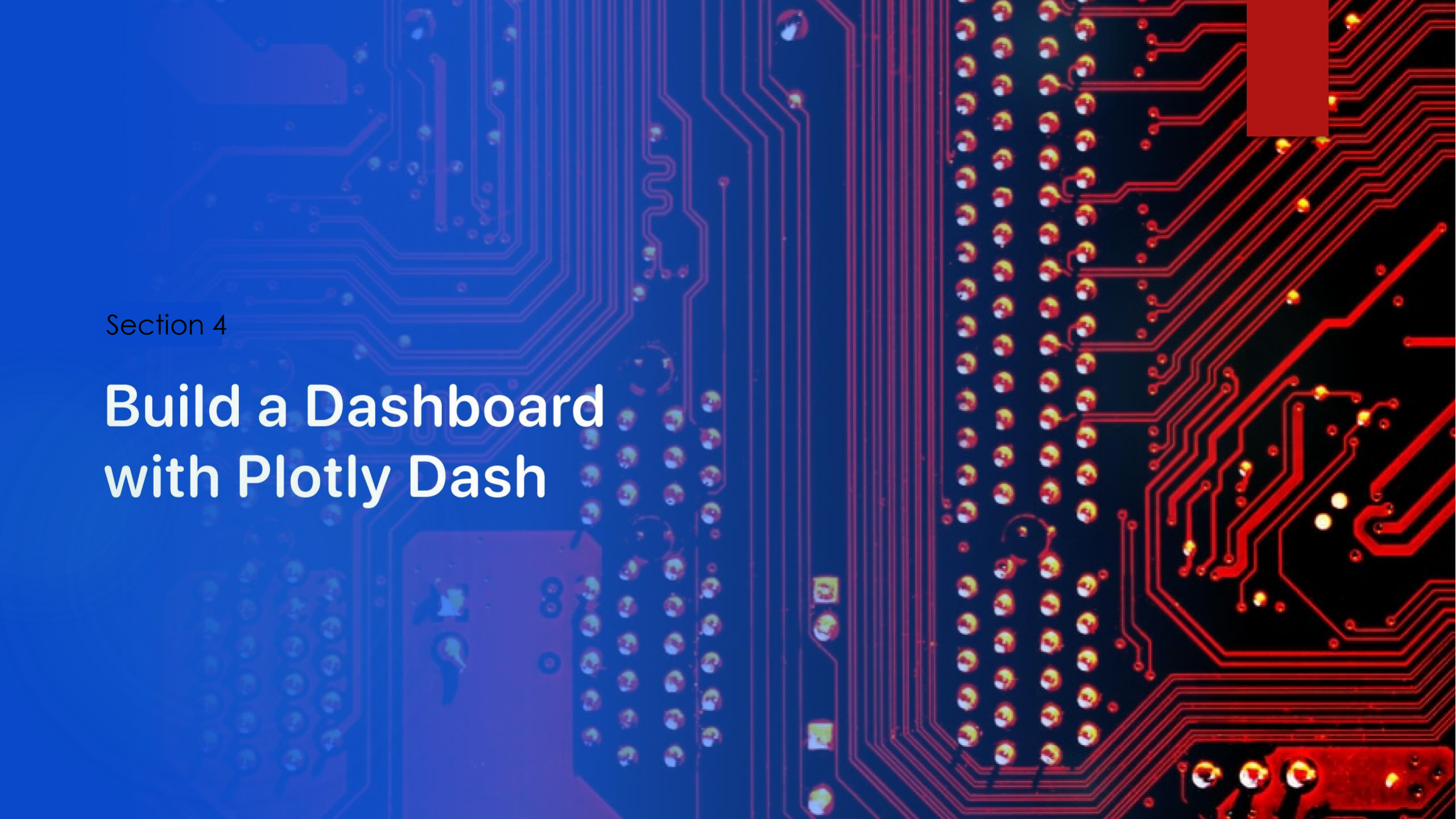
- Folium was used to create land markers regarding the launch sites and to local fixtures such as Cape Canaveral to see to distance between the city and the launch site.
- This determine the distance between the launch site and these fixtures using longitudinal and latitudinal data.



Geographic Visualizations

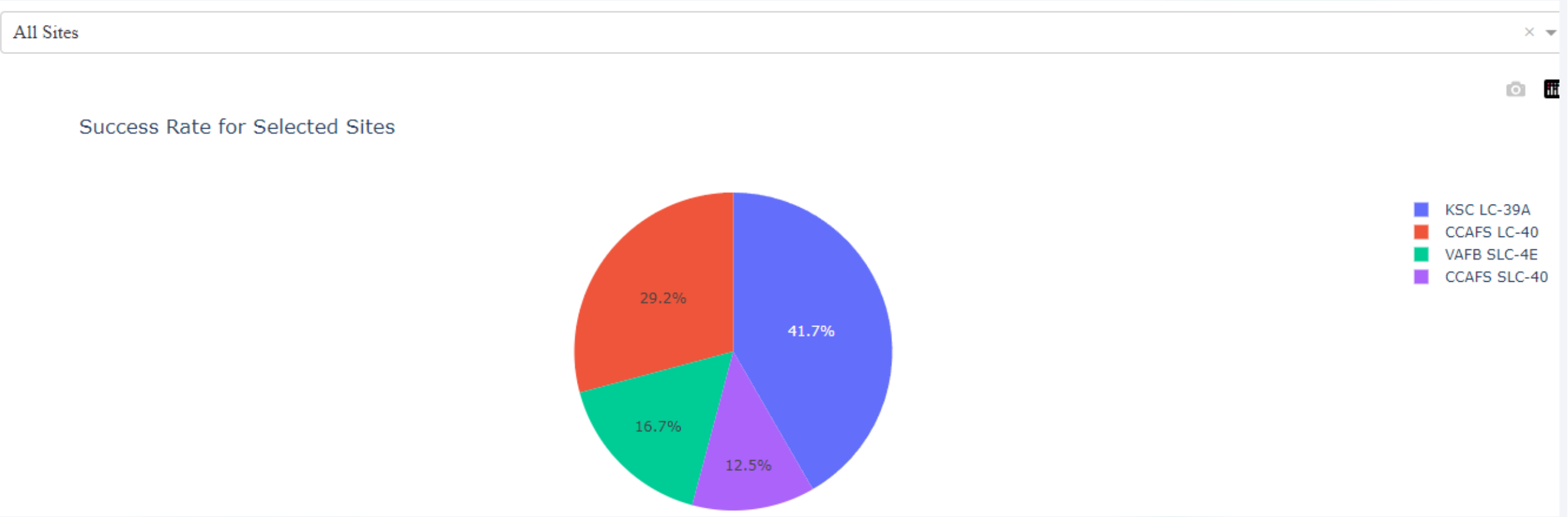
- Folium was used to create land markers regarding the launch sites and to local fixtures such as the coast, the nearest highway, and railroad.
- This determine the distance between the launch site and these fixtures using longitudinal and latitudinal data.




Section 4
Build a Dashboard with Plotly Dash

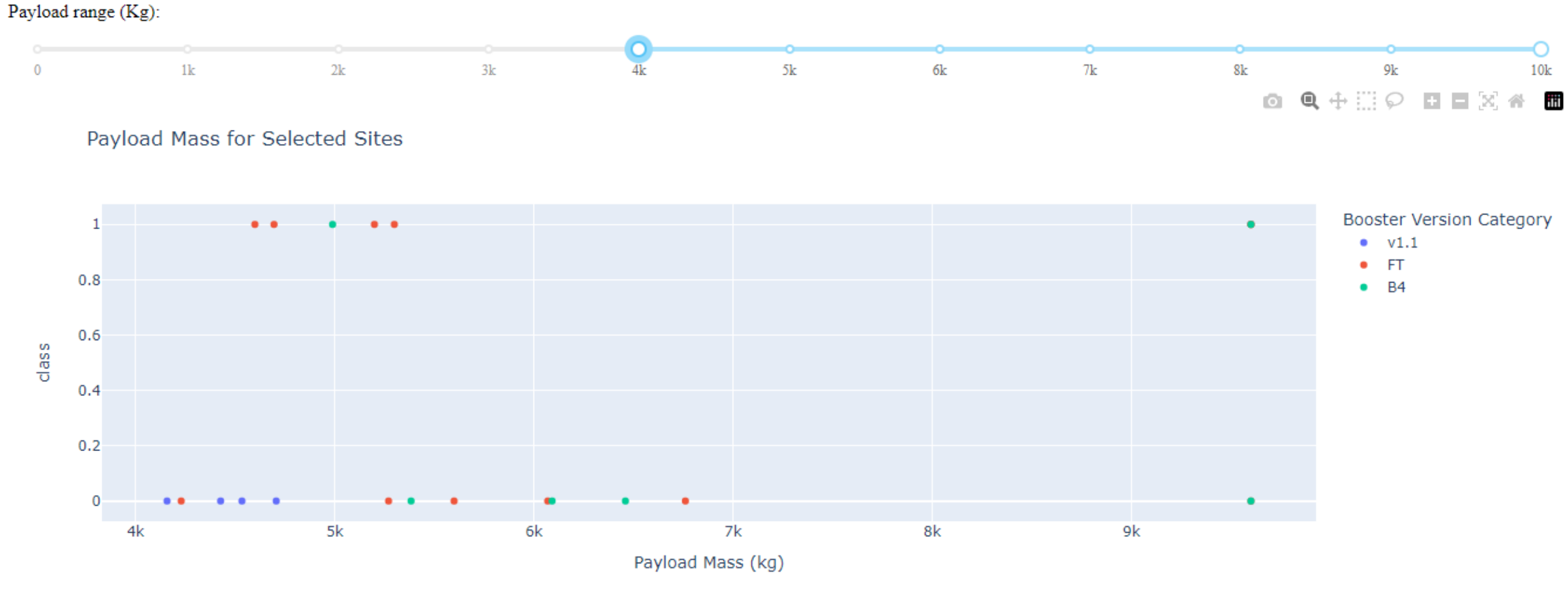
Launch Success for All Sites

- The success rate for all sites varies by site with KSC LC-39A having the highest at 41.7%. Followed by CCAFS LC-30, VAFB SLC-4E, and CCAFS SLC-40



Payload Mass for All Boosters

- The payload mass of 4000 kg and above for all booster versions used.
  - Only the V1.1, FT, and B4 boosters were used for payload masses above 4000 kg.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Of the 4 models used, the Decision Tree model had the weakest accuracy. While the remaining models all tied for accuracy.

```
print('Logistic Regression Accuracy', logreg_cv.score(X_test, Y_test))
print('Support Vector Machine Accuracy', svm_cv.score(X_test, Y_test))
print('Decision Tree Accuracy', tree_cv.score(X_test, Y_test))
print('K Nearest Neighbor Accuracy', knn_cv.score(X_test, Y_test))
```

```
Logistic Regression Accuracy 0.8333333333333334
Support Vector Machine Accuracy 0.8333333333333334
Decision Tree Accuracy 0.7222222222222222
K Nearest Neighbor Accuracy 0.8333333333333334
```

Classification Accuracy

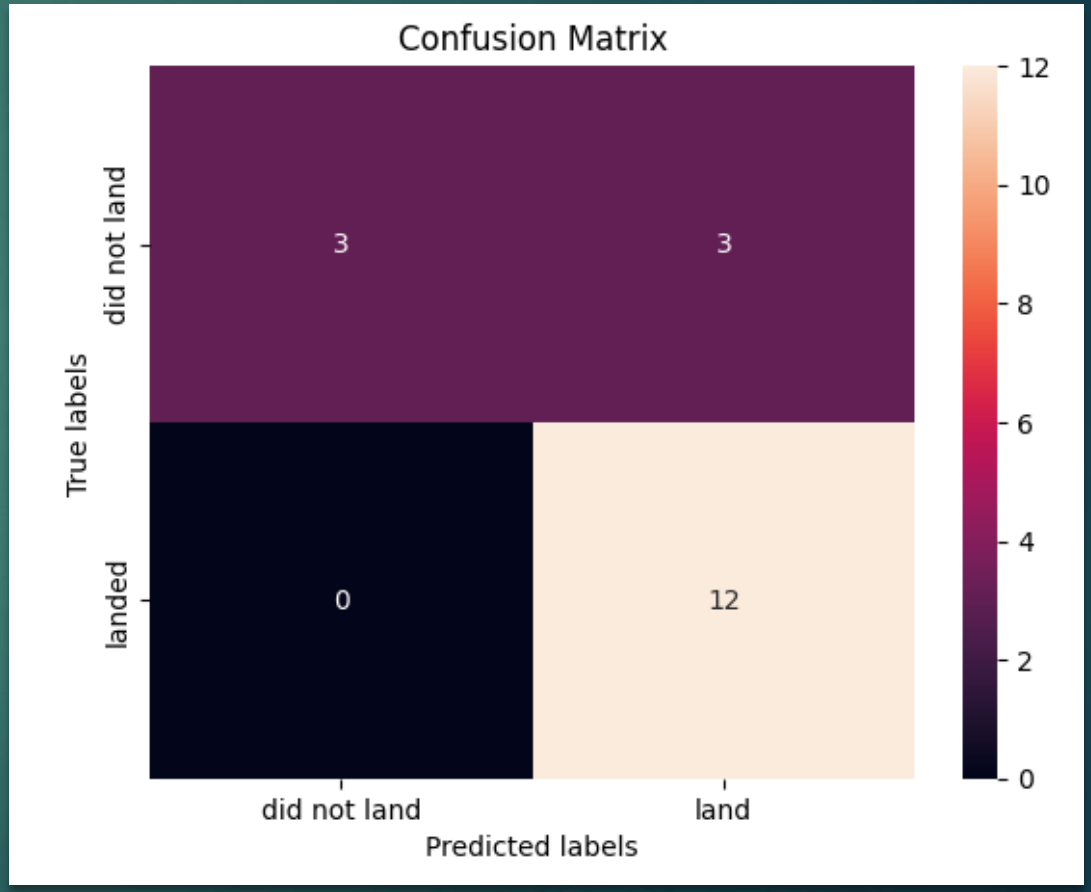
- When using a different built-in score function, the Decision Tree model produced the best accuracy of the four models with KNN and SVM tying and the Logistic Regression placing last.

```
print('Logistic Regression Best Accuracy', logreg_cv.best_score_)
print('Support Vector Machine Best Accuracy', svm_cv.best_score_)
print('Decision Tree Best Accuracy', tree_cv.best_score_)
print('K Nearest Neighbor Best Accuracy', knn_cv.best_score_)
```

```
Logistic Regression Best Accuracy 0.8464285714285713
Support Vector Machine Best Accuracy 0.8482142857142856
Decision Tree Best Accuracy 0.8732142857142857
K Nearest Neighbor Best Accuracy 0.8482142857142858
```

Confusion Matrix

- The Decision Tree model predicted that 12 would land and 0 would fail to land, when 3 actually failed to land.



Conclusions

- There are several launch sites used by SpaceX for NASA contracted missions
- Payload mass varies based on booster, orbit, launch site, and year.
- Landing outcomes vary based on mission parameter, specifically if it is landing on a ground pad or drone ship
- Data Visualizations aid in understanding the relationship between orbit, landing outcome, payload mass, and booster type.

Conclusions

- Machine Learning results vary by model with accuracy fluctuating depending on how it is measured.
- Even the most accurate model will have a measurement error and not accurately predict 100% of the time.
- However, given the results of these models an estimate can be made regarding the landing outcome of a given flight based on booster, year, and payload mass.

Appendix

- Data Collection
- Webscrapping
- Data Wrangling
- EDA SQL
- Data Visualization
- Folium
- Plotly Dash
- Machine Learning

Thank you!