IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Erin Jones

1/25/2023

Outline

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

Executive Summary

- Summary of methodologies
  - ➢ Data collection with API
  - ➢ Data collection with Web Scraping
  - ➢ Data Wrangling
  - ➢ Exploratory Data Analysis(EDA) with SQL
  - ➢ Exploratory Data Analysis(EDA) with Data Visualization
  - ➢ Interactive Map with Folium
  - ➢ Dashboard creation with Plotly Dash
- Summary of all results
  - ➢ EDA Results
  - ➢ Interactive analytics
  - ➢ Predictive Analytics

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website which cost 62 million dollars while other providers can be upward of 165 million. The cost differential is largely due to SpaceX innovation which allows them to reuse the first stage. If the first stage success rate can be predicted, then cost per launch could be calculated.

- Problems you want to find answers

1. What is the first stage success rate (successful re-landings)? What factors might contribute to this?
2. What is the cost per launch?
3. Is there a way to make future landings more successful based on findings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From SpaceX API, web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding on categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

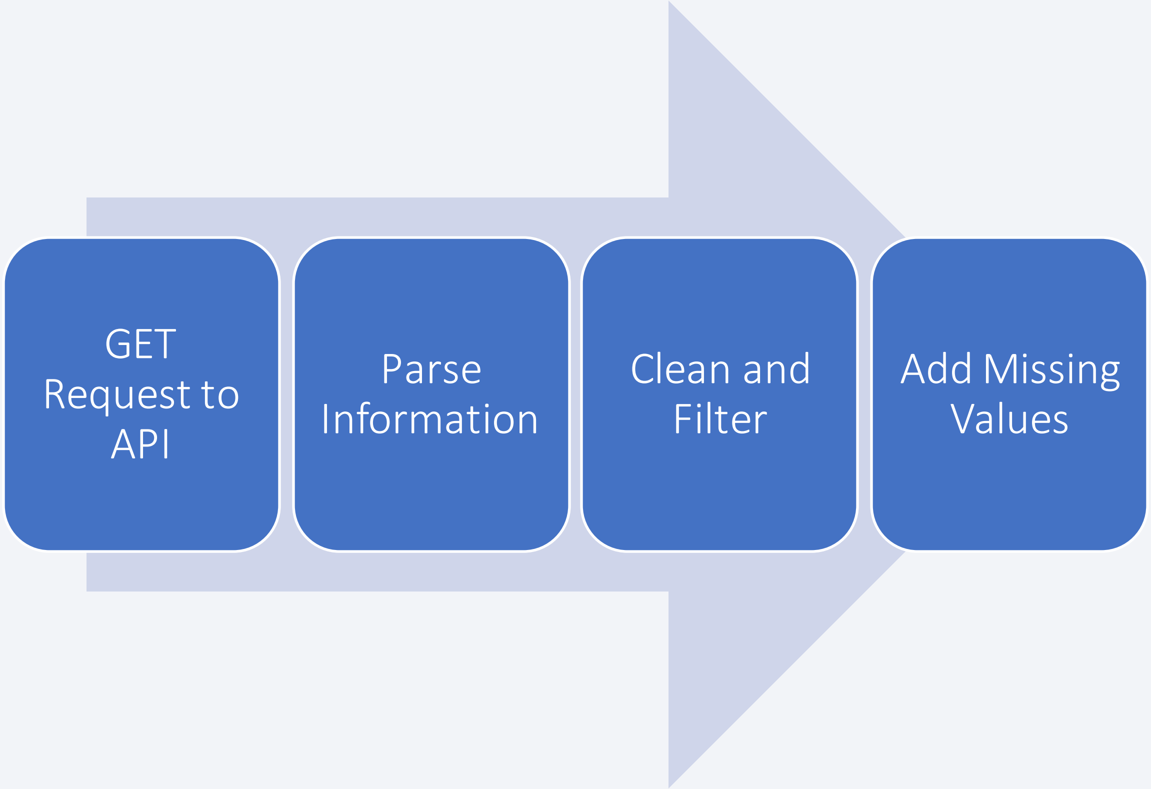
Data Collection

Using helper functions, the API was used to extract information with the identification numbers in the launcher data. The extracted information was decoded and turned into a pandas dataframe using the normalize() function for json files. It was then cleaned and any missing values were added. Some of the information collected from the database were the rocket names, the launchpad site, and the mass of the payload and where it was headed. Cores were also collected to determine the outcomes of landings, if the rocket was reused, and how is coincided with the previous variables listed. Accuracy of test data was also collected.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/data-collection-.ipynb



GET Request to API

Parse Information

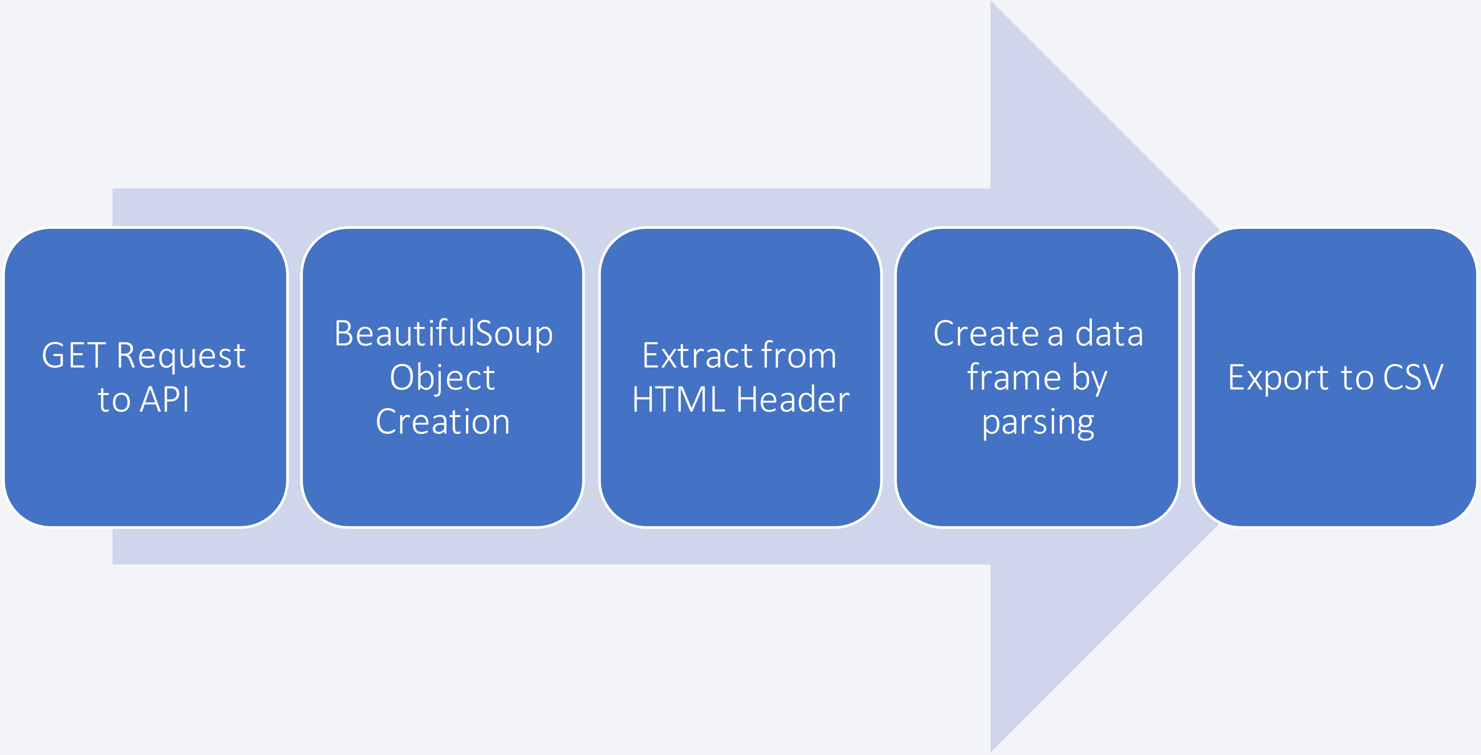
Clean and Filter

Add Missing Values

Data Collection - Scraping

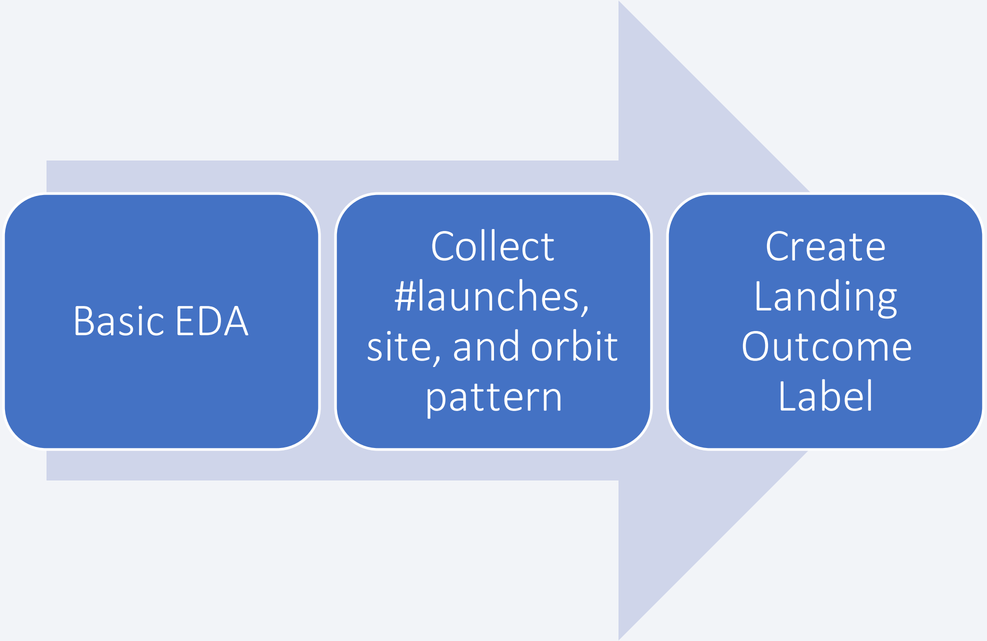
- Using BeautifulSoup we web scrapped the Falcon 9 Launch Records

- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/webscraping.ipynb



Data Wrangling

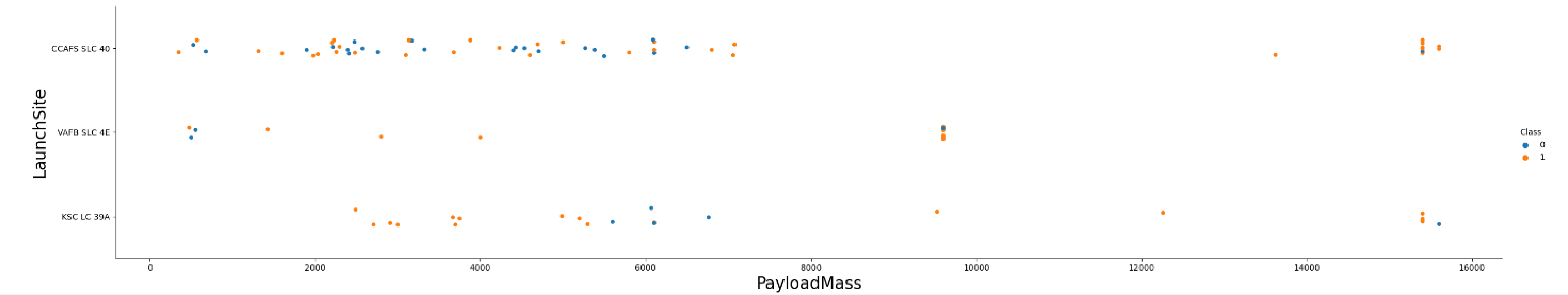
- Basic EDA was preformed and training sets were determined. Then the amount of launches at each site as well as their orbit pattern was determined. This allowed for a landing outcome label to be created
- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/data_wrangling.ipynb


Basic EDA
Collect #launches, site, and orbit pattern
Create Landing Outcome Label

EDA with Data Visualization

- The most prominent chart used to view the data were scatterplots followed by barplots. Some of the features compared were: Launch Site x Payload Mass, Launch Site x Flight Number, Payload Mass x Flight Number, Orbit and Flight Number, and Payload x Orbit

- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/EDA.ipynb



EDA with SQL

- The chronological order of SQL queries performed:
  - ➢ Names of the unique launch site
  - ➢ Top 5 launch sites beginning with 'CCA'
  - ➢ Total payload mass of boosters launched by NASA(CRS)
  - ➢ Date of first achieved landing
  - ➢ Booster names with a payload mass between 4000-6000 kg
  - ➢ Number of success/failure
  - ➢ Amount of Boosters with max payload
  - ➢ Failed landings for the drone ship, booster version, and launch site for 2015
  - ➢ Rank of landing outcomes between 2010 and 2017
- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/eda-sql.ipynb

Build an Interactive Map with Folium

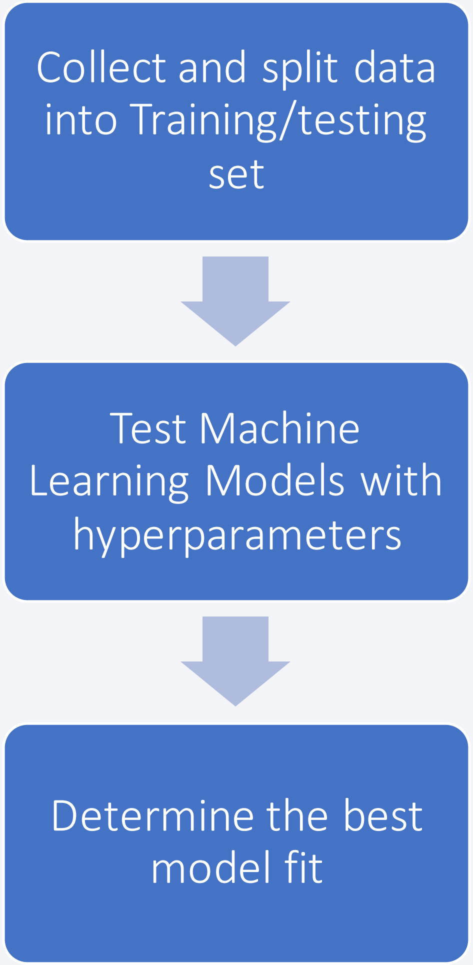
- The following map objects were used: Markers, Circles, Lines, and Marker Clusters
  - ➢ Markers: launch sites
  - ➢ Circles: specific locations (coordinates) of important places (Space center, etc.)
  - ➢ Lines: distance between coordinates
  - ➢ Marker Clusters: groups event at coordinates (ex. Launches)
- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/4ef870d28154ded8f4d09721e5cf76891c78f342/Interactive%20Visual.ipynb

Build a Dashboard with Plotly Dash

- Scatter Plots and Pie charts were used to visualize the data. The two graphs were the percentage of launches per site, and the payload range. By analyzing this the best launch site per payload can be determined.
- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/2629a02c57a309a923afb4c1bea15ec5132725a3/PlotlyDashApp.py

Predictive Analysis (Classification)

- After the data was loaded, numpy and pandas was used to transform and split the data into a training and testing set.
- This allowed for different Machine-Learning Models to be built and tuned under different hyperparameters. Accuracy was used in order to improve the model and tune it.
- The Machine Learning Models we viewed were logistic regression, decision tree, support vector, and k nearest neighbor.
- GitHub Notebook URL: https://github.com/ErinJones2234/Data-Science-Projects/blob/2629a02c57a309a923afb4c1bea15ec5132725a3/Mahine%20Learning.ipynb



Collect and split data into Training/testing set

Test Machine Learning Models with hyperparameters

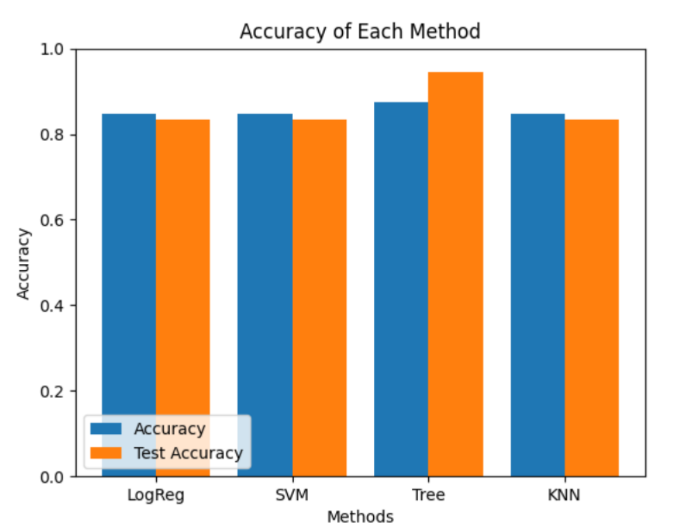
Determine the best model fit

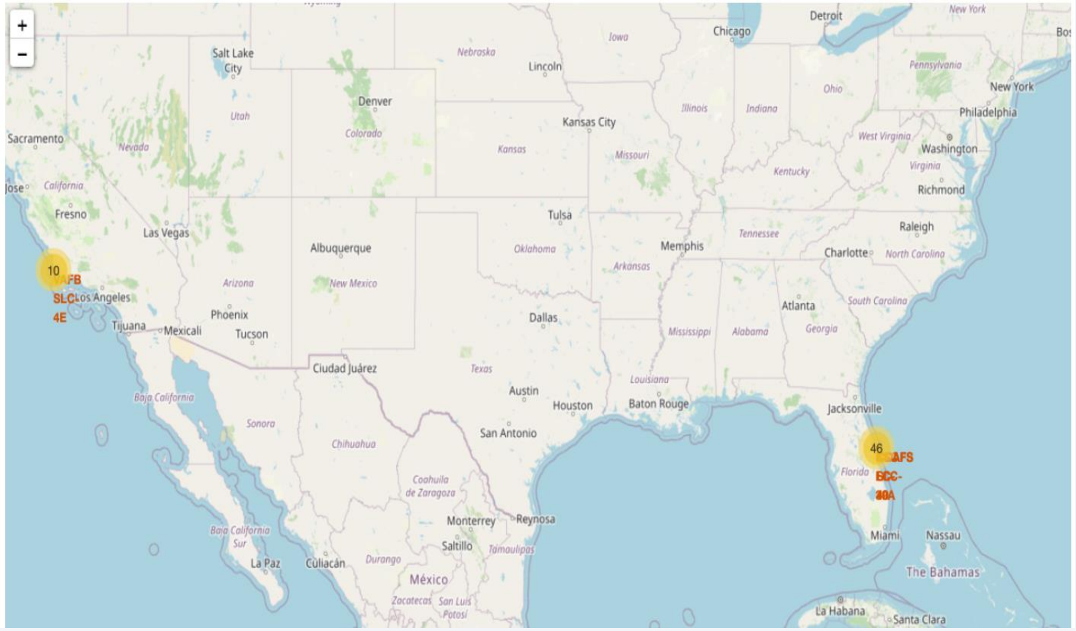
Results

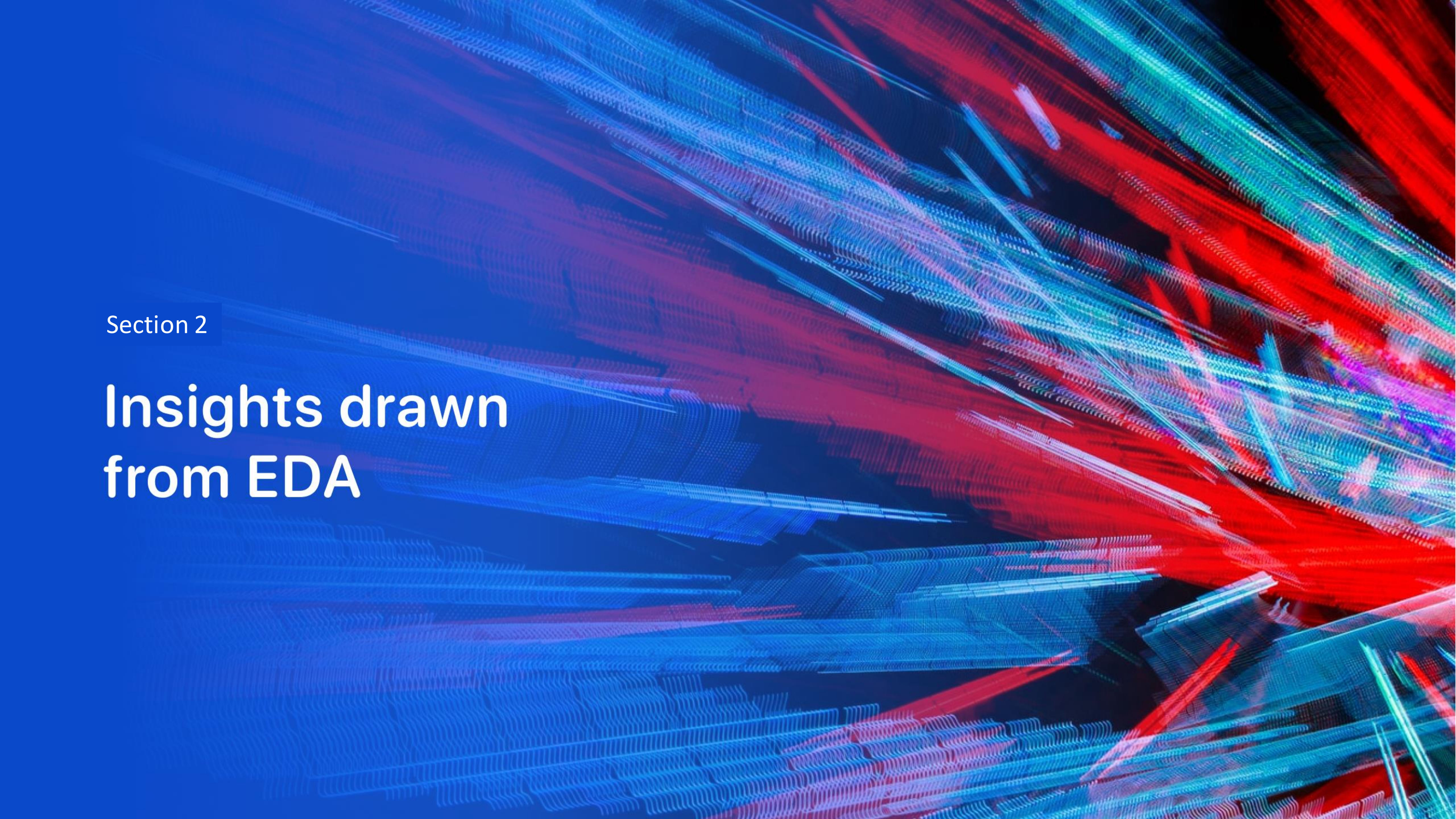
- Exploratory data analysis results:
  - 4 different launch sites were used
  - First successful landing was in 2015, but two boosters F9's failed landing this year
  - Average payload was 2,928 kg
  - F9's could handles heavy payloads and successfully land at drone ships
  - As years increased so did success rate, with almost 100%

Results

- The interactive analytics showed the best launch sites were located on the coast, and the most launches were on the east coast as seen on the map below
- Using predictive analysis the best model predict successful landings was the decision tree model as show in the figure below.



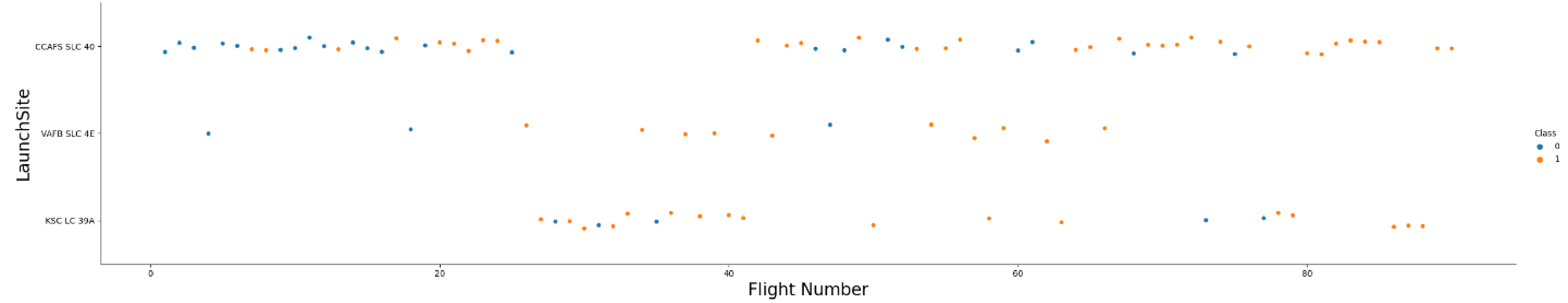




Section 2

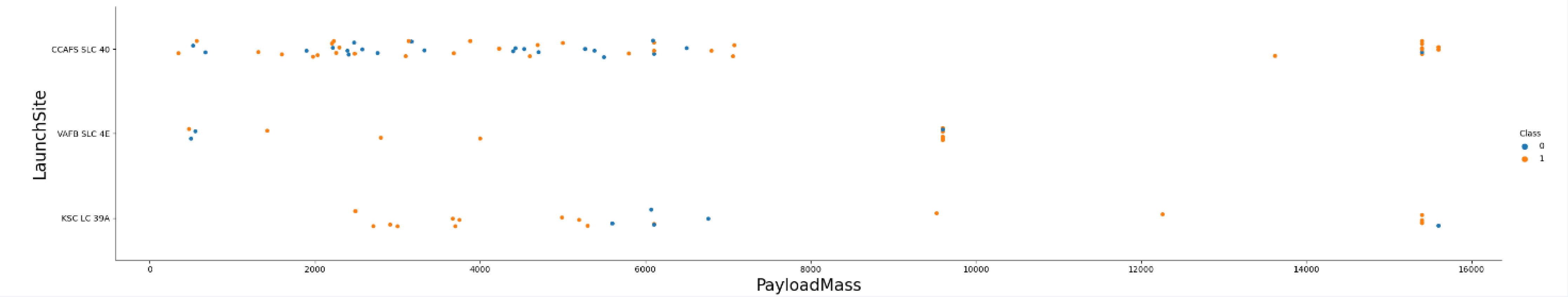
Insights drawn from EDA

Flight Number vs. Launch Site



- From the scatterplot of Launch site x Flight Number we can determine that CCAF5 SLC 40 is the most used launch site, followed by VAFB SLC4E.

Payload vs. Launch Site



- In this graph Launch Site and Payload Mass were compared. The higher the payload, especially at 9,000kg had a high success rate and typically launches from VAFB SLC4E.

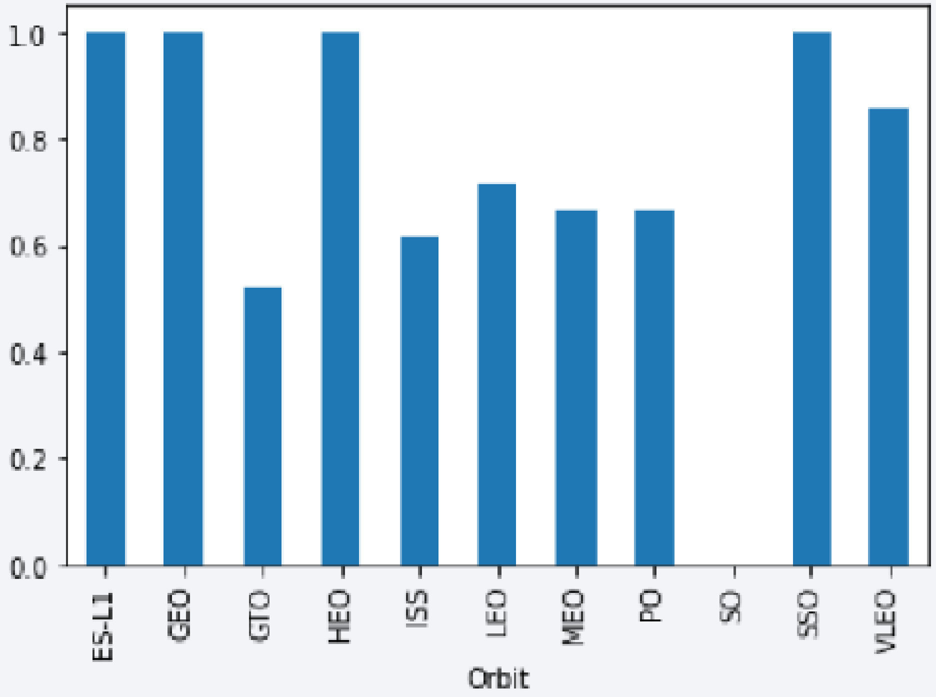
- CCAFS SLC40 launched the most payloads below 7,000kg.

- Heavier payloads (above 10,000) were launched from CCAFS SLC40 and KSC LC39A

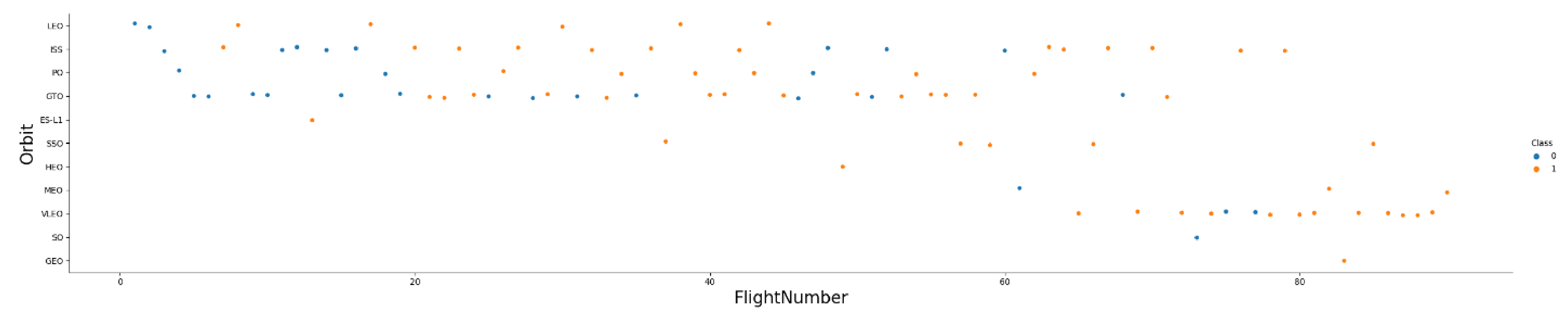
Success Rate vs. Orbit Type

- Four orbit patterns had the highest success rate:
  - ➢ ES-L1
  - ➢ GEO
  - ➢ HEO
  - ➢ SSO

- SO orbit pattern had a success rate of 0%.
- The second lowest orbit pattern was GTO.



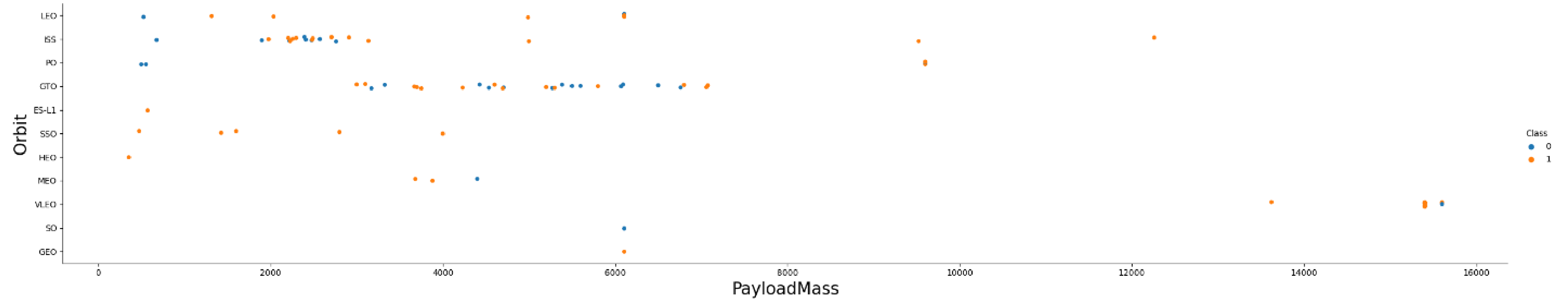
Flight Number vs. Orbit Type



- In this scatter plot Orbit type and Flight Number were compared. GEO flight orbit that was previously mentioned had a 100% success rate notably because it only has one launch from Flight number 83.

- GTO which had the second lowest success rate also has the most flight numbers. VLEO had a high success rate above 80% and has a decent amount of flights.

Payload vs. Orbit Type

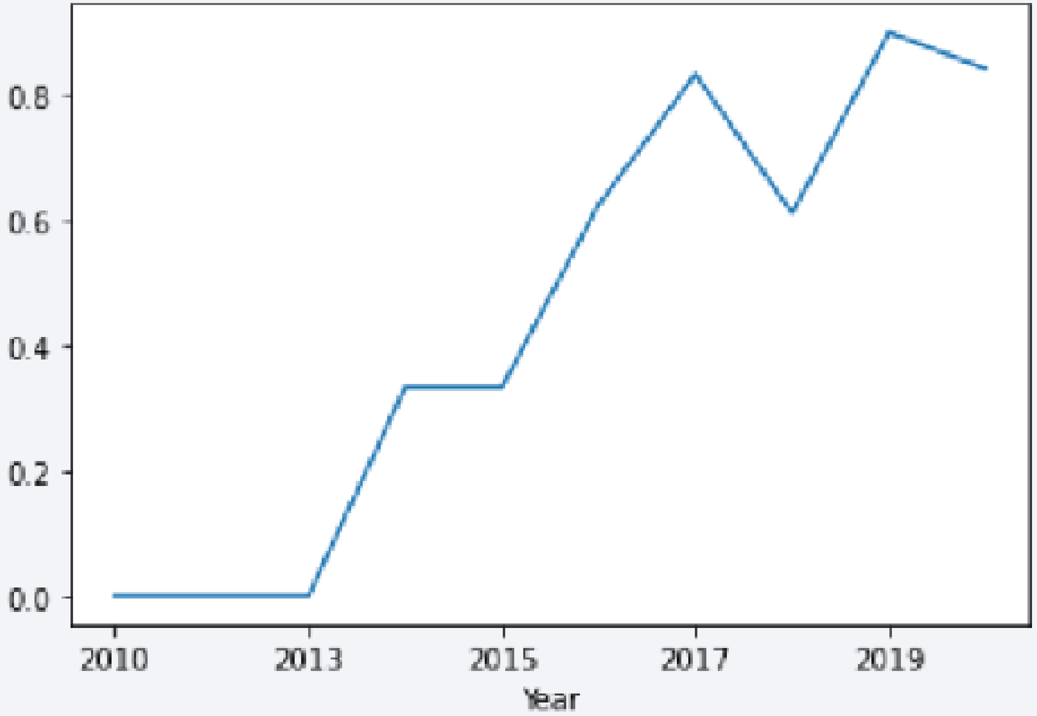


- Heavy payloads over 8,000kg went into ISS, PO, and LEO orbits.
- Payload mass between 4,000-6,000 kg preferred the GTO orbit.
- ISS was also a notable orbit pattern for payloads between 2,000-4,000kg

Launch Success Yearly Trend

- A significant increase in success rate was seen in 2013 after the first launch. It plateaued in 2014 but in 2015 it jumped again.

- Note:2015 had the first successful landing.

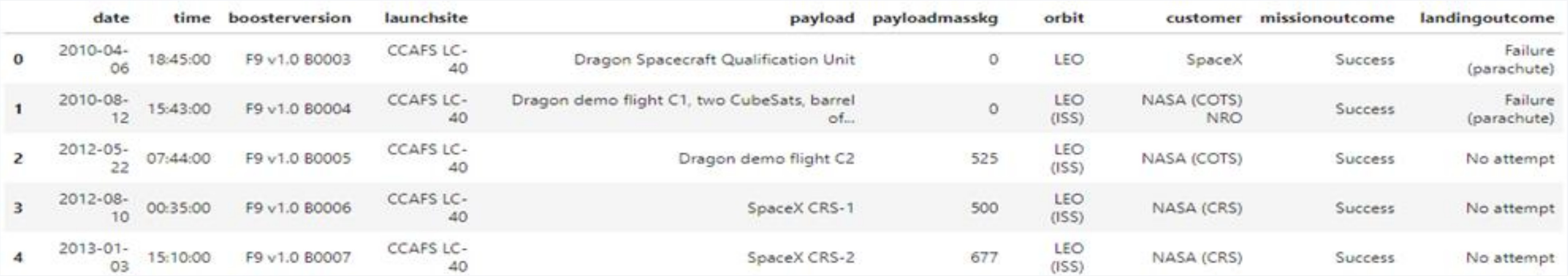


All Launch Site Names

- There are four unique launch sites as seen by the figure.
- The unique launch sites were obtained by selecting unique 'launch_site' values

Launch Site(s)
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'



	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- 5 records where launch sites begin with `CCA`
- In order to find CCA unique launch sites we used a SELECT, WHERE (CCA), and a LIMIT function to select the dataset, put a clause for CCA, and limit it to five records.

Total Payload Mass

- The total payload mass carried by boosters from NASA was 45,596 kg
- This was collected by using a SUM function of the payload mass from the data set.

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928.4 kg
- Here we used an AVG function on the payload mass and a WHERE function to make it only F9 v1.1 from the data set

Total Payload Mass
2928.4

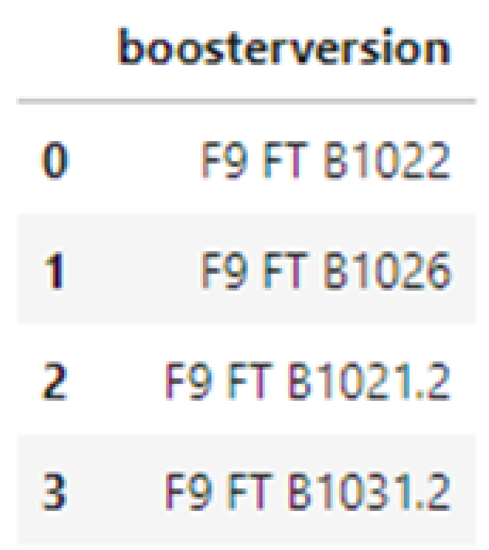
First Successful Ground Landing Date

- The first successful landing on a grounding pad was December 22, 2015.
- In order to find this we used a MIN function on the date and a WHERE function on the LandingOutcome where the landing was successful

First Successful Landing
12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed in the figure.
- In order to find this we first used a SELECT statement on the version, FROM the SpaceX dataset. Then we used a WHERE clause in order to specify a succesful landing from the landing outcome. We also added AND conditions to put the range between 4,000-6,000 kg.



	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is listed in the figure below
- The WHERE clause was used on the MIssionOutcome variable.
- Notably: One of the successful missions had an unclear payload status

Success	Failure
100	1

Boosters Carried Maximum Payload

- The names of the booster versions which have carried maximum payload mass are listed in the figure
- This query was obtained by using a WHERE clause and MAX function on payload mass.

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

- Looking at 2015 Launch records, wo boosters failed their landing in 2015 at the CCAFS LC-40 launch site which is one of the most used landing sites.
- In order to find the failed launches a WHERE clause was used to find the landing outcomes that resulted in failure. The BETWEEN condition was used to specify 2015 launches.

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

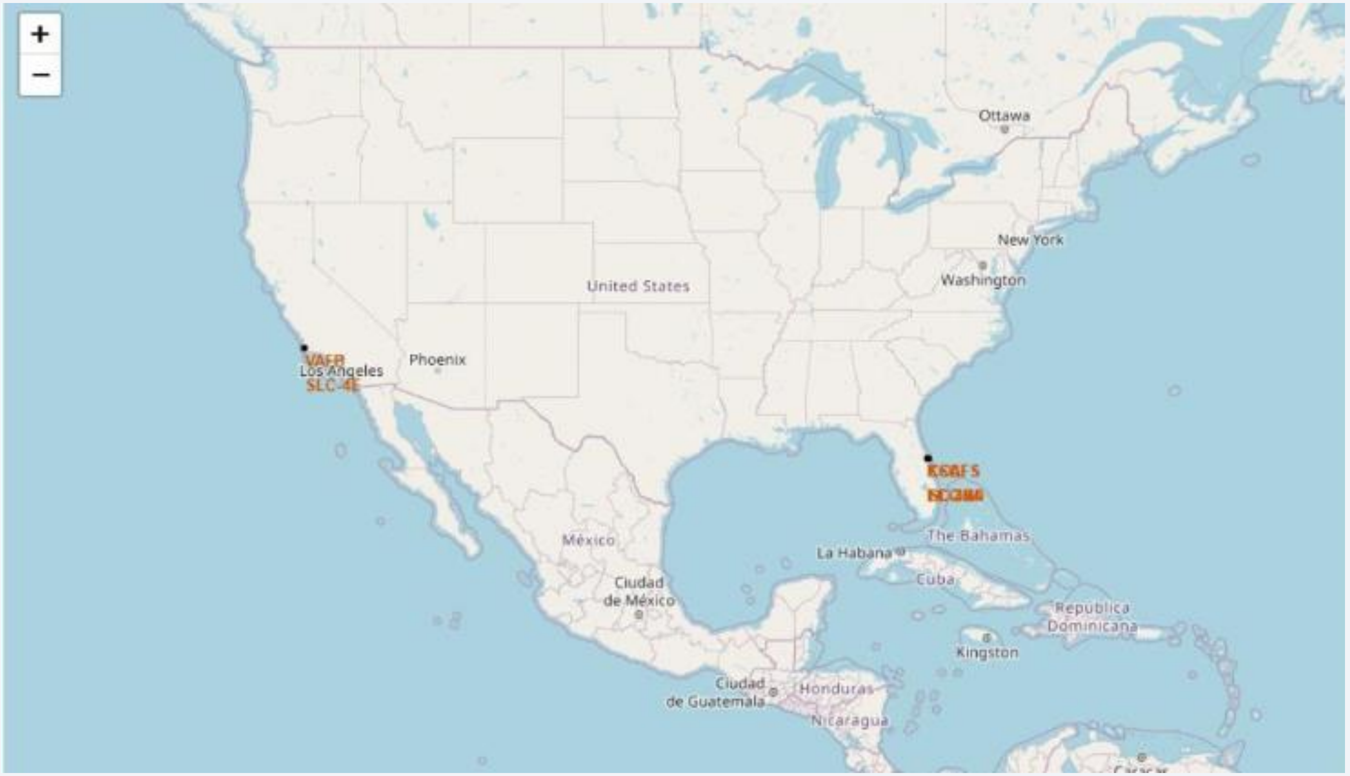
- The landing outcomes between 04/06/2010 - 03/20/2017 were ranked according to the frequency of the outcome as seen in the table.
- The COUNT function was used on the landing outcomes variable, and WHERE was used for the specified data range. In order to assemble the list in descending order, the ORDER BY function was used.

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

Section 3

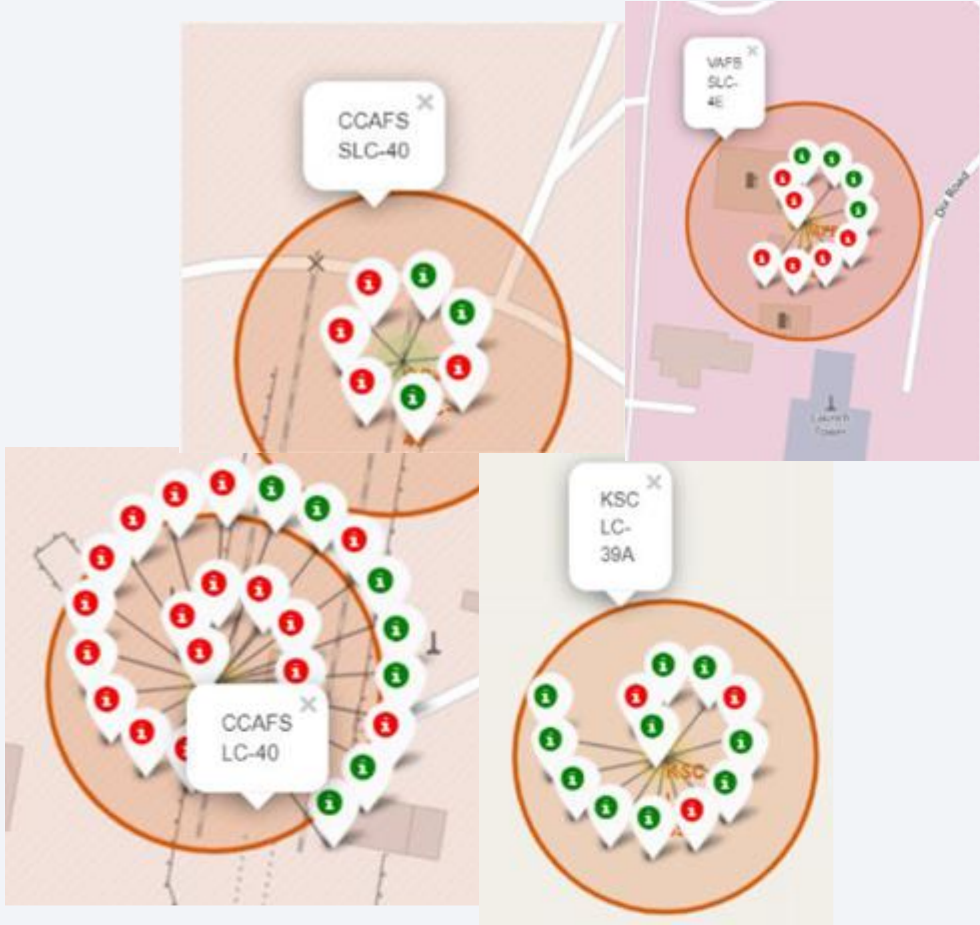
Launch Sites Proximities Analysis

Map of West and East Coast Launch Sites



The launch sites on the West and East coast of the USA are shown. Although both sites are used, the launch site in Florida contains the most used launch site at Cape Canaveral and also where the first rocket was launched in 2015

Launch Sites with Markers



Note: Green markers = successful launch, red = failure

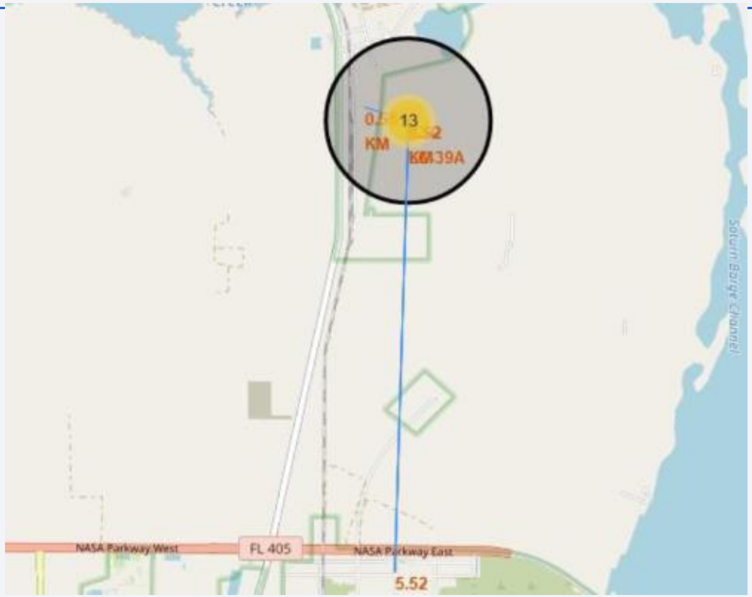
CCAFS SLC-40 and CCAFS LC-40 are extremely close to one another. KSC LC-39A is also located on the East coast. Surprisingly KSC has the most successful launches when compared to the other east coast sites.

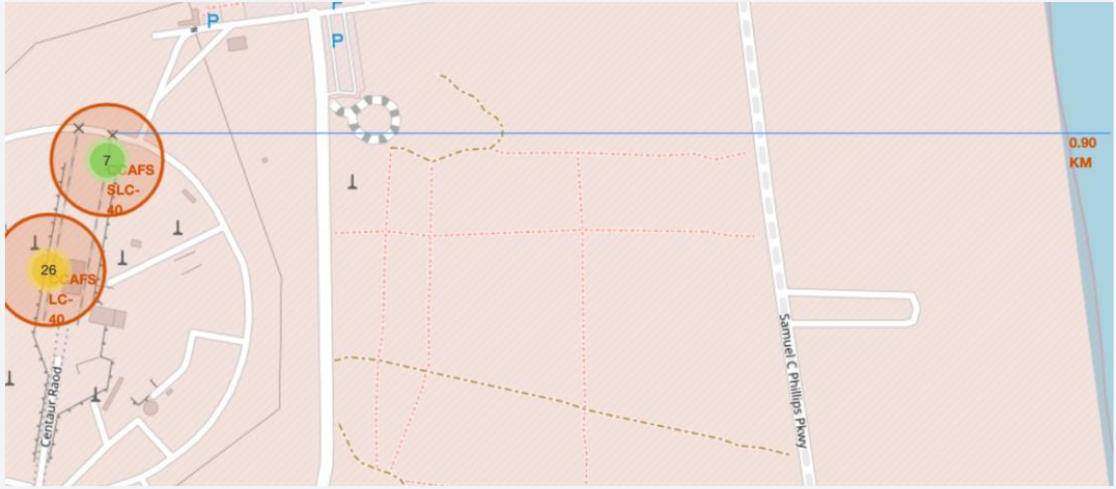
VAFB SLC-4E is located on the west coast in California and has a rocky success rate.

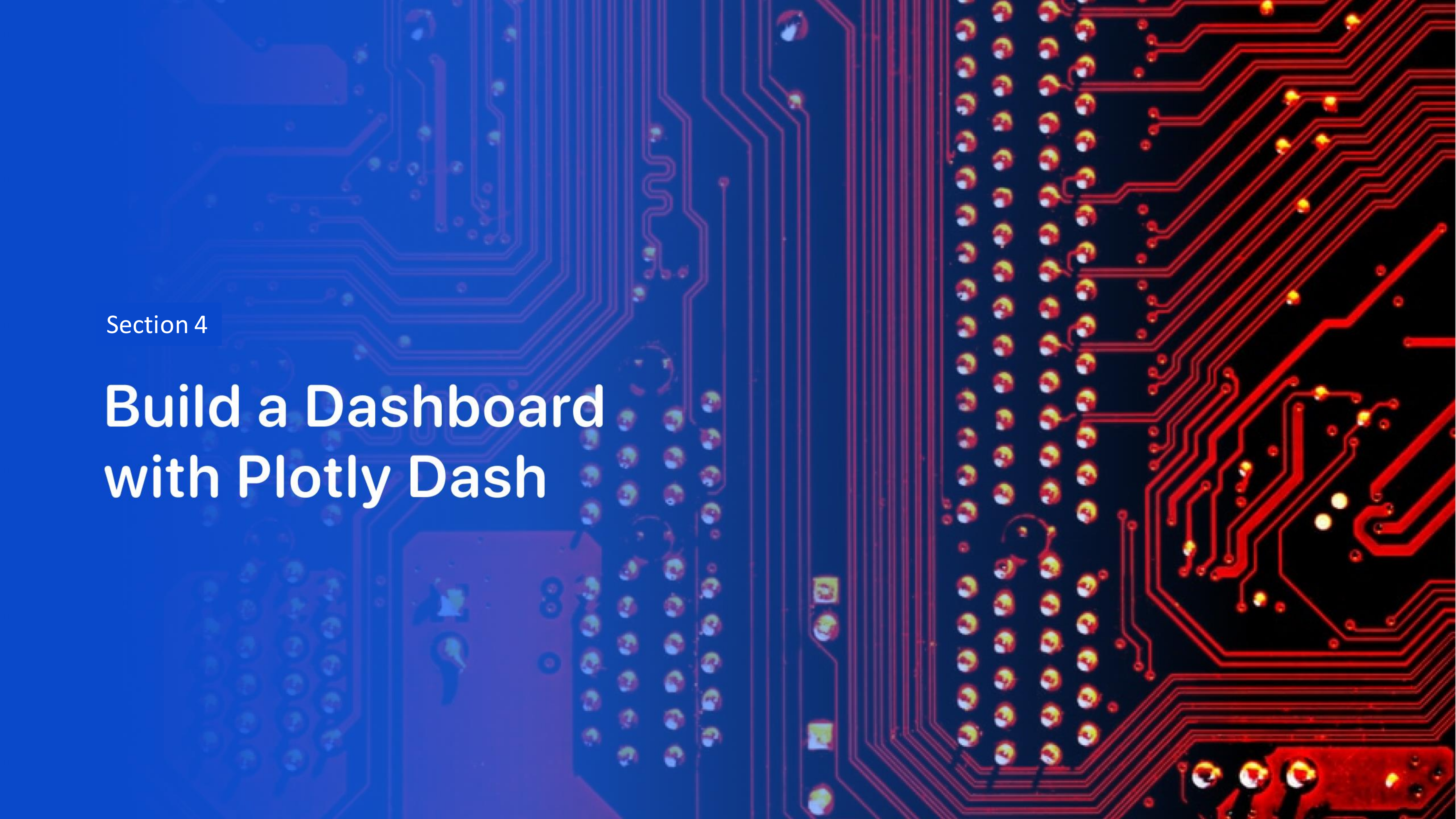
Launch Site Geospatial Safety

- By looking at the maps we can see that the landing sites in Florida are about 0.90 KM from the coast. Aside from the SpaceX compound it is far enough from civilization as to not cause issues, and far enough from the coast for some hurricane precautions.

- The KSCLC-39A seen in the upper right figure is 5.52 KM from the nearest parkway. Though it is near a road and a railroad it doesn't seem to be near any inhabited neighborhoods or marketplaces.



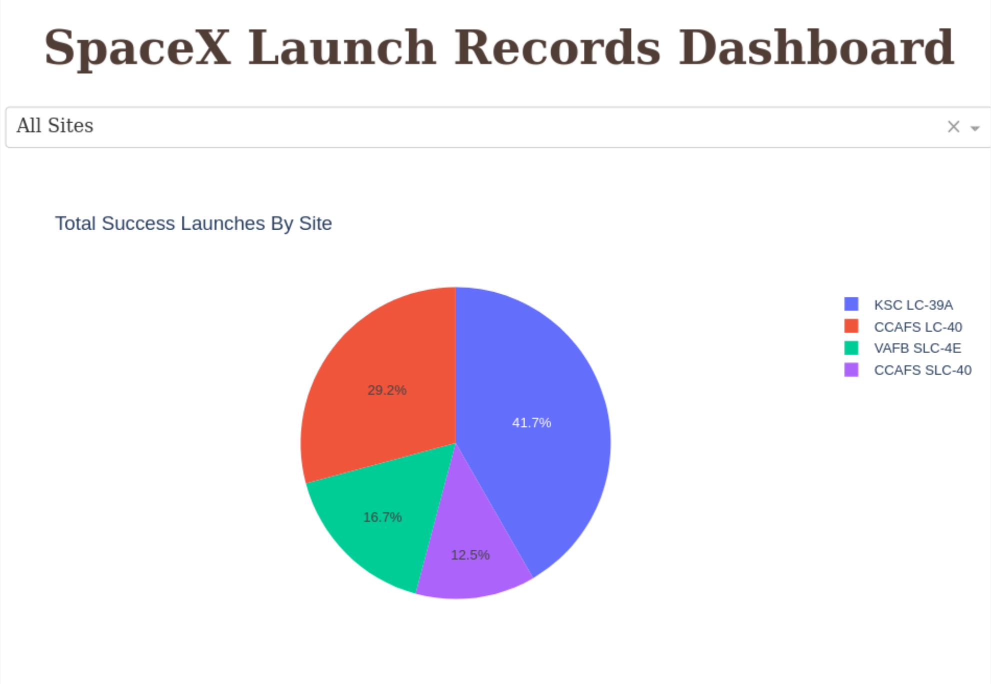




Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Launch Site

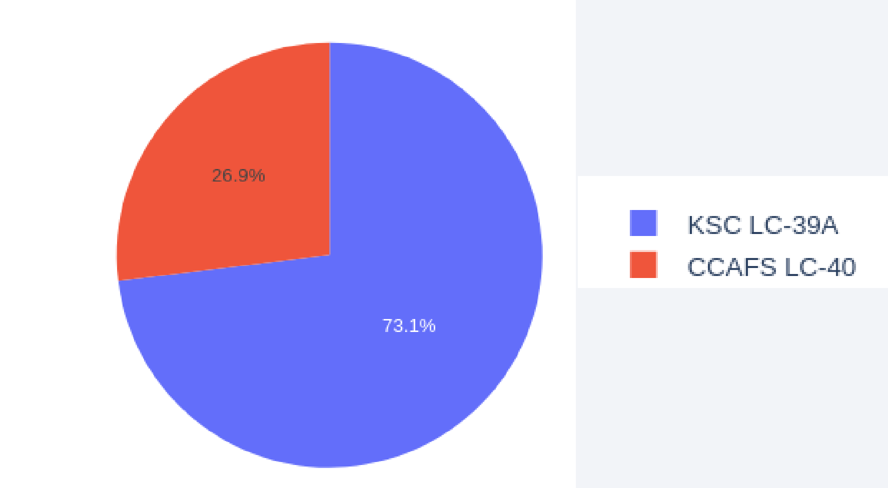


SpaceX Launch Records Dashboard

All Sites

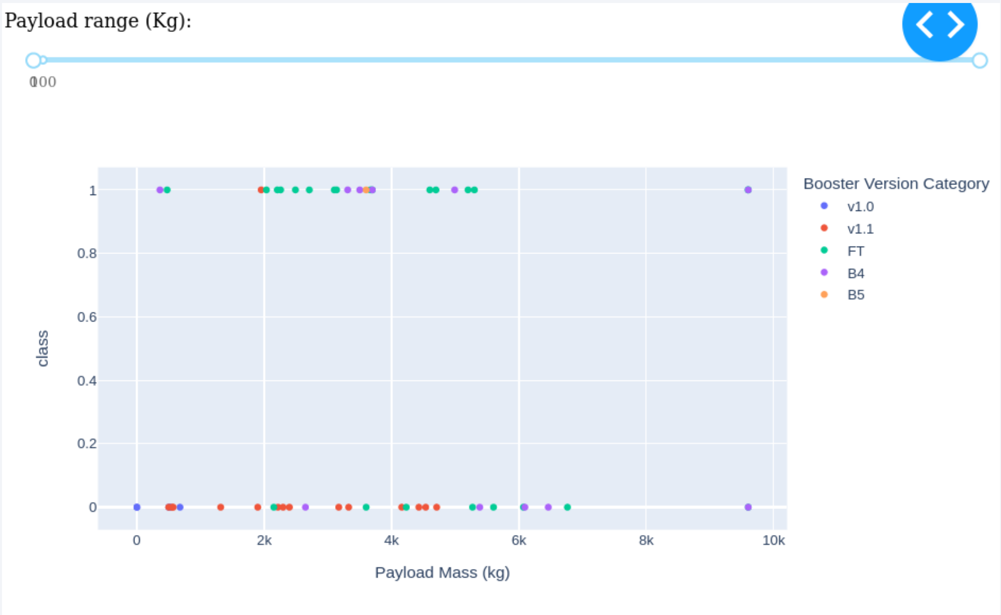
- KSC LC-39A was the most successful launch site followed by CCAFS LC-40. Both of these sites had high launch rates as well, seen by previous analysis.
- Note: CCAFS LC-40 was the first site and thus will notably have more failures contributing to its low percentile.

KSC LC-39A Launch Site succes rate



- As mentioned in the previous slide, KSC had the highest success rate followed by CCAFS LC-40. However more analysis may need to be done to see the difference in launch times, as CCAFS LC-40 was the first launch site. However, despite the site the Florida launch site locations show the most success.

Payload vs. Launch Outcome



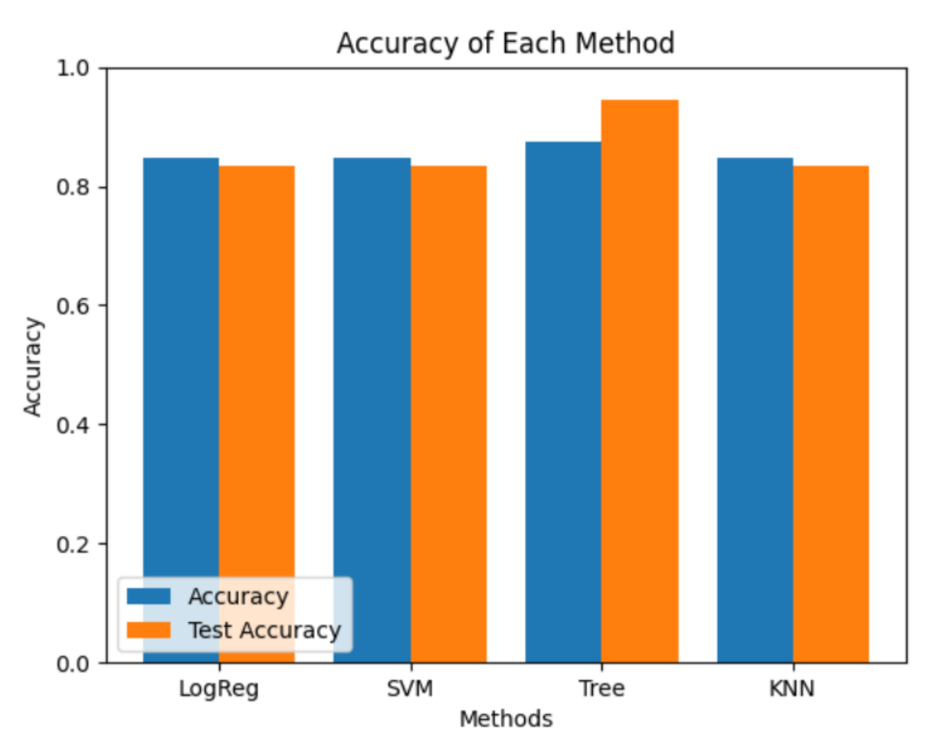
- High payload mass typically was assigned to B5 boosters. FT boosters were widely used for payloads between 2,000-6,000kg. Notably, v1.1 booster versions closely followed typically carrying a payload no more than 5,000kg.

Section 5

Predictive Analysis (Classification)

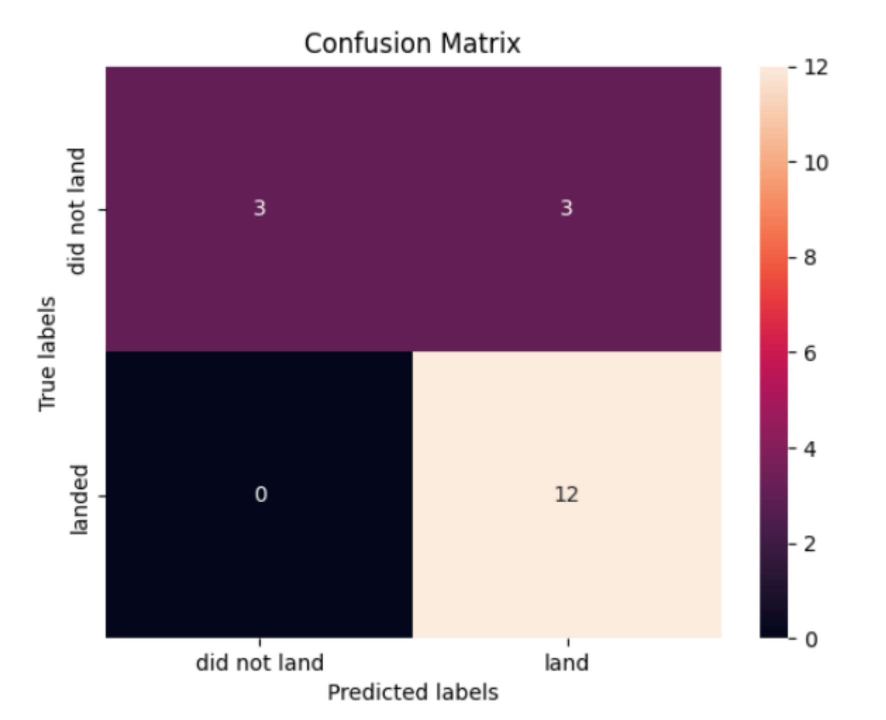
Classification Accuracy

- Using predictive analysis the best model to predict successful landings was the decision tree model as show in the figure.



Confusion Matrix

- The confusion matrix of the Decision Tree Model showed there were no false negatives but at least three false postives.



Conclusions

- Going back to the initial objective of this project was to find success rate. Some notable factors are orbit pattern, booster version, payload mass, and launch site. Though some variables such as orbit pattern had clear winners per percentage calculated (ES-L1) the frequency of use needed to be considered. Though it had the highest success rate, it had only been used once. For this reasons my conclusions are based on context of the rate of occurrence, the time of occurrence, and the success rate.
- Payloads over 8,000 kg have the highest success in ISS, PO, and LEO orbits and should be launched from CCAFS SLC40 and KSC LC39A in a F9 B5 boosters. Payloads with a mass between 4,000-6,000 kg should be in a GTO orbit in a F9 v1.1 or F9 FT and launched from CCAFS SLC40 and KSC LC39A. Payloads below 5,000 kg should in an ISS orbit in a F9 v1.1 and can be launched from any site, though, CCAFS SLC40 is the most used.
- KSC LC39A has the most successful launches but CCAFS SLC40 was the first launch site which may contribute to its lower percentage due to earlier failed launches. However, since it's first launch in 2015, the process has been refined enough now to where all launches are almost 100% successful.
- Note: The best model chosen to represent this project is a decision tree classifier

Appendix

- Some maps in the folium lab did not properly load and unfortunately screenshots of charts and graphs lost quality upon being enlarged.

Thank you!