IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wilson Voelker
10/03/2024

Outline

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

Executive Summary

Data Science Capstone Project which will utilize data Available through the SpaceX API and Wikipedia pages to determine if it is possible to predict SpaceX Rocket Landing Success.

Project Scope

- Data Collection with Python and Beautiful Soup
- Data Wrangling with Python
- Exploratory Data Analysis with SQL and Visualization
- Dashboard with Plotly Dash
- Predictive Model Development with Machine Learning and Scikit Learn

Executive Result Summary

- Data Collected from Available Sources
- Relevant Features determine through exploratory Data Analysis
    - E.g. Payload Mass, Launch Site, Orbit
- Predictive Model Developed with high accuracy score

Introduction

Space Y is a new company breaking into the space Race. Our goal is to compete with established space companies, like SpaceX. This Project is designed to help us compete against larger and better funded companies.

Primary Business Question

We would like to design a cost efficient space program which will rely heavily on cost control. For this we are performing market intelligence to find key features of the existing program at SpaceX

- What characteristics of Launch Site contribute to safety and efficiency
- What factors contribute to state 1 booster recovery success
- How can we reasonable predict costs by predicting the likelihood of stage 1 boosters landing successfully

Section 1

Methodology

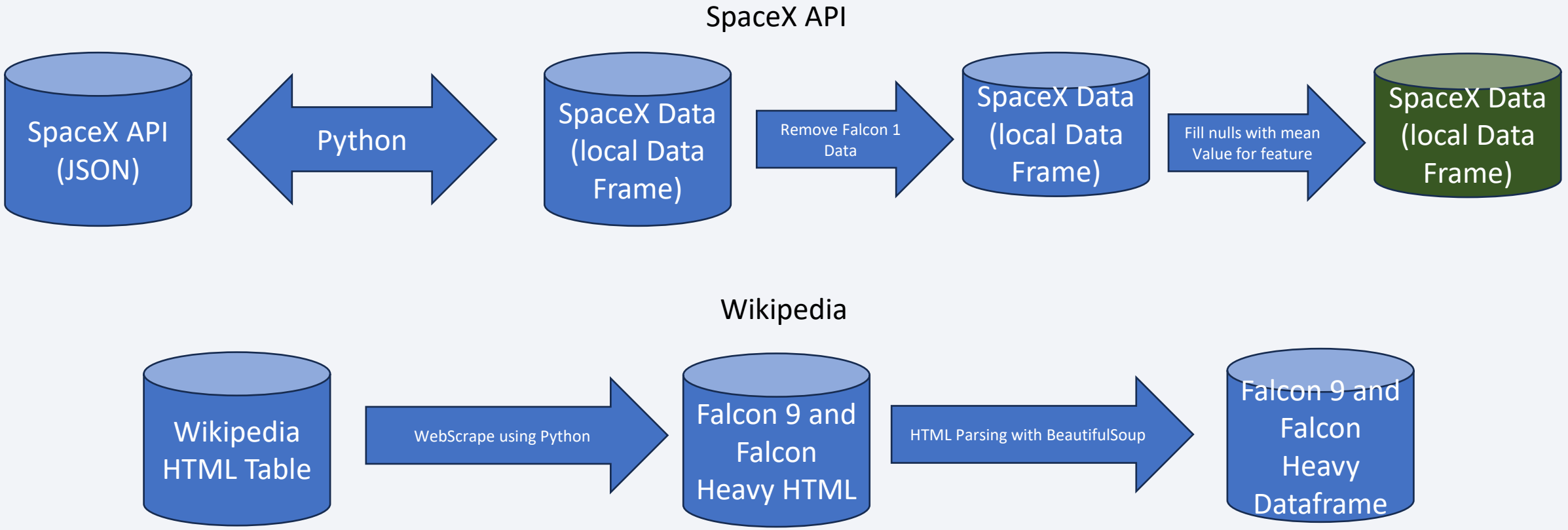
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and Wikipedia using Open Source Tools (Python)
- Perform data wrangling
  - Data was prepared, cleaned, and consolidated from multiple JSON downloads into a single data frame.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple predictive classification models were evaluated and tuned for optimum accuracy

Data Collection

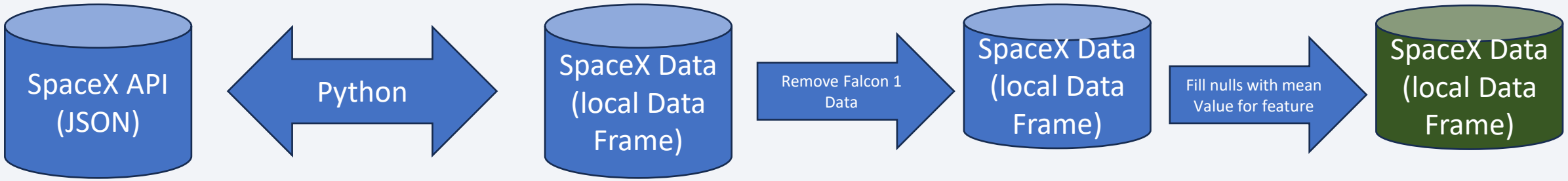
- Market research data collected from Publicly available data provided by SpaceX and Wikipedia.



Data Collection – SpaceX API

- Data Collected from SpaceX API
- Data Parsed into a Data Frame using Pandas JSON_normalize
- Specific Features selected and assembled into a single data frame from Launch and Core data

See the completed SPACEX API calls process here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/39eac2557a2541d28f2eab9bc67859b3fbbd5932/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Launch Data Available in Wikipedia
- Used BeautifulSoup to parse HTML Scraped from Wikipedia
- Selected Features were loaded to a data frame

See the complete Web-Scraping Process here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/39eac2557a2541d28f2eab9bc67859b3fbbd5932/jupyter-labs-webscraping%20(1).ipynb

Data Wrangling

Note: Data was processed using python and pandas

Landing Outcome Examination

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
False Ocean	2
None ASDS	2
False RTLS	1

Landing Outcome categorized by landing site and success or failure

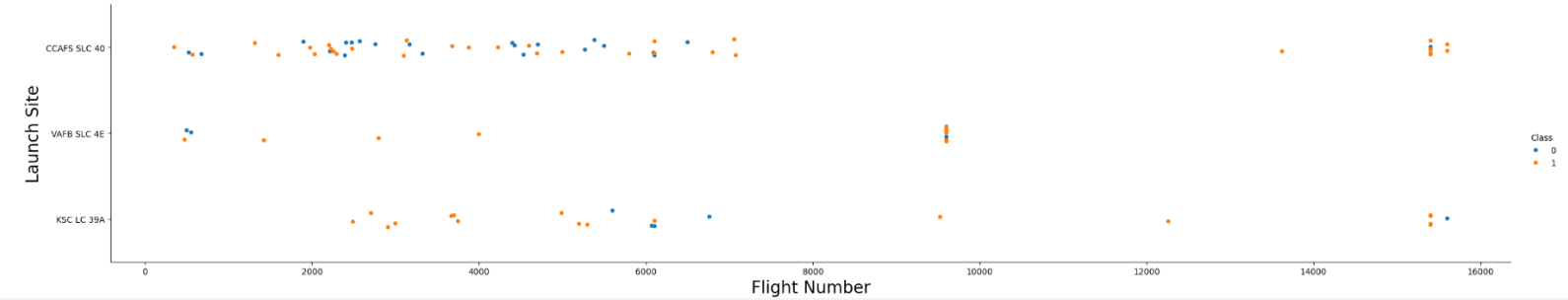
Binary True/False (1,0) Classification was needed for Machine learning models

landing_class	Outcome	
1	True ASDS	41
0	None None	19
1	True RTLS	14
0	False ASDS	6
1	True Ocean	5
0	False Ocean	2

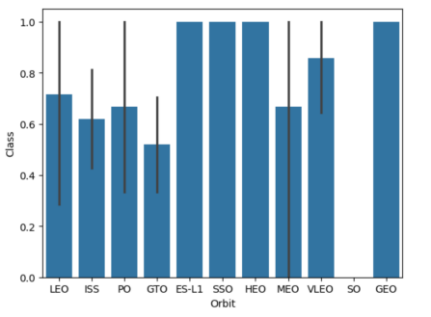
- See the code here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/956ef900640bbbfba9b6c6ca2546eac6569557e3/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

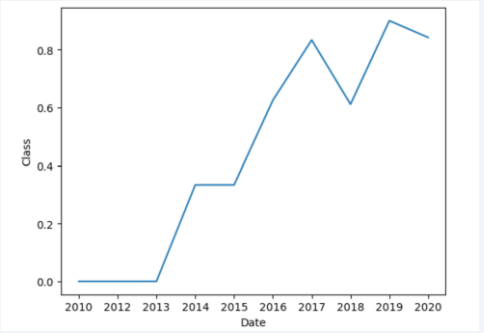
- Various charts were plotted to determine the relation ship of the available features to the landing outcome
- See the complete analysis here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/1d87fe5e51ad7bd348c35106d690f2bbfe768d3f/edadataviz.ipynb



With increasing payload masses and landing successes, it appears launch site VAFB SLC-4E is not used for payloads over 10000 KG. Launch site and payload mass may also hove a correlation with landing success



Analysis suggests that the intended orbit of the SpaceX rocket has correlation to the success rate



The analysis shows that SpaceX has more successful launches in recent years

EDA with SQL

Queries Performed to gain understanding of the Launch Data

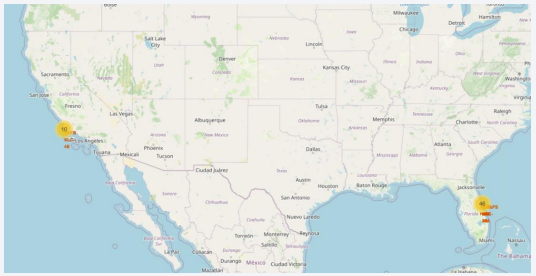
- Display the names of the unique launch sites
- Display 5 records where launch sites begin with the string 'CCA'
- display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 201
- Rank 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

Landing_Outcome	count(*)	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

Outcome	count(*)
Failure	40
Success	61

- View the analysis workbook here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/73cd838ab3403815a8ef549a0a29851e30e67458/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

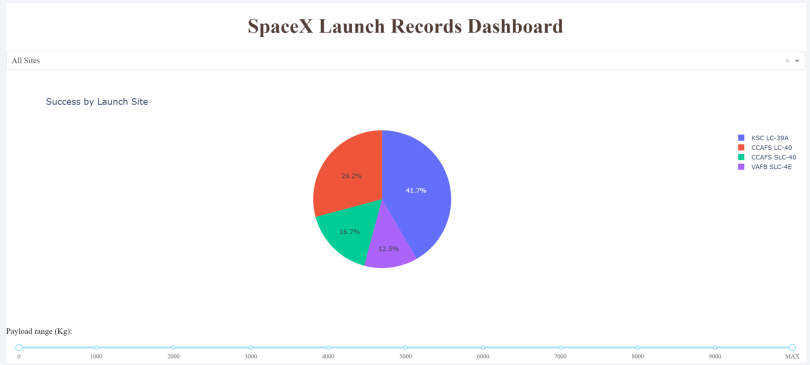


Folium Map Launch Site Analysis

- Circles were created for each launch site
- A Marker was added for each launch and color coded Red for Failed Landings and Green For Successful Landings (note these markers are grouped and will unfold when clicking on the yellow highlighted area)
- A distance measurement marker was added indicating the distance from one launch site to the nearest coast, highway, railroad, and city (calculated via latitude and longitude for each point)
- Lines were drawn from the launch site to the distance markers to better understand launch site placement and safety
- See the complete script here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/0fb25b13dbaa8294f210895c6883d8da215bd3da/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plotly Dash SpaceX Dashboard – Examine Launch Success by launch site and Payload mass with the ability to focus the analysis on specific sites and mass ranges to better understand the impact of these critical factors on landing success.



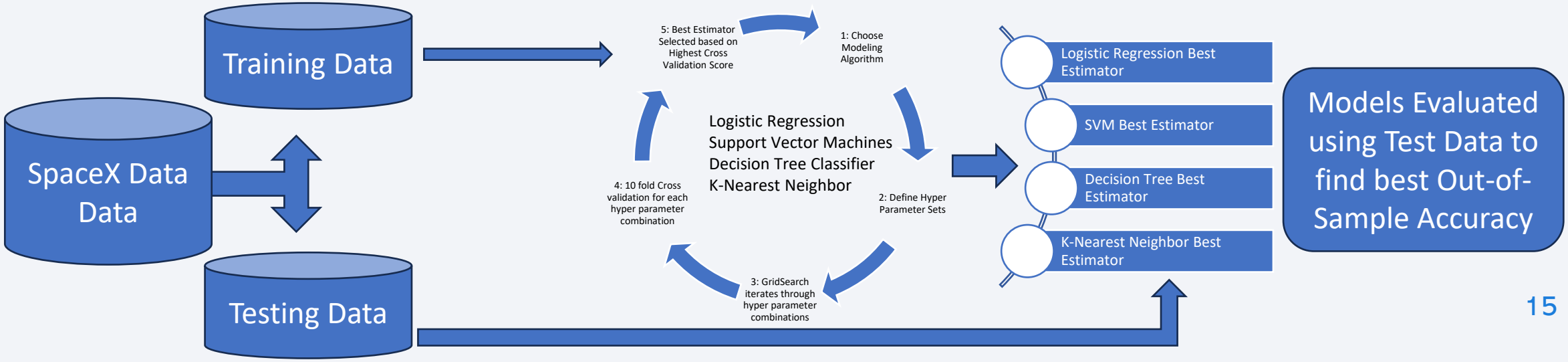
Available visualizations
- Launch Success Pie-Chart
- Payload/Success Scatter Chart

Dashboard Filters/Controls
- Launch Site Selector – Focus the Dashboard content on a selected launch site
- Payload mass Range Selector – Focus the Payload mass analysis to specified ranges of payload mass

View the code here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/0fb25b13dbaa8294f210895c6883d8da215bd3da/spacex_dash_app.py

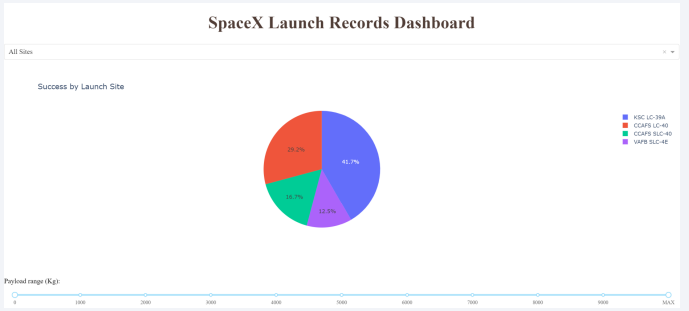
Predictive Analysis (Classification)

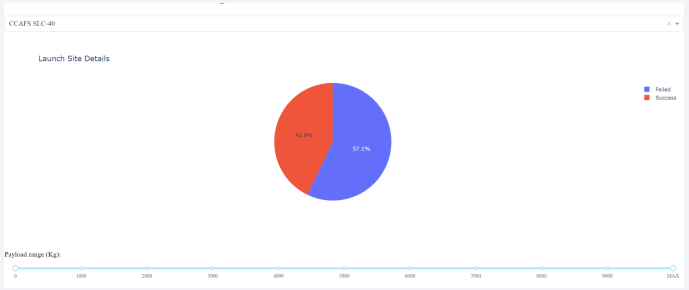
- SpaceX Data was split into training and Testing data Sets. Multiple Modeling Methods were used along with Hyperparameter GridSearch to find the optimal predictive model
- See the code here: https://github.com/GDaddySoul/Data-Science-Capstone/blob/d9da471ffa48d6112aee2d1f40bb860b28af1a04/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

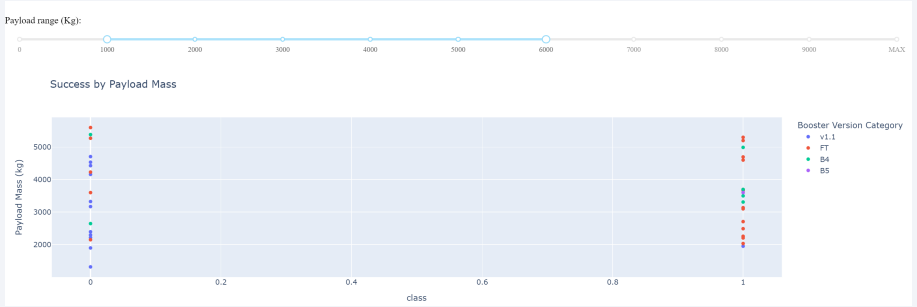


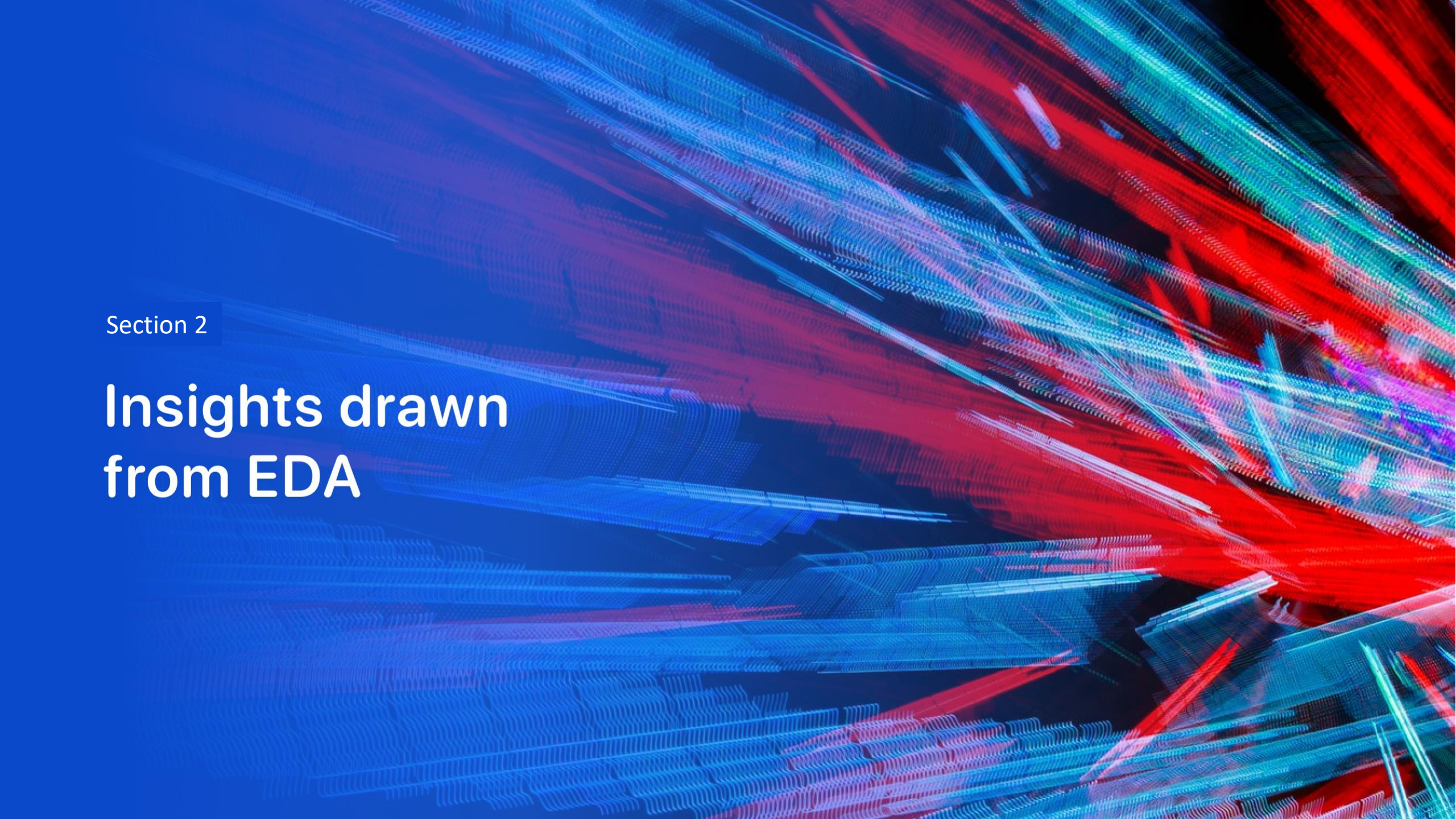
Results

- EDA Research offers several conclusions
  - Launch site economics likely require proximity to rail and highway
  - Launch site safety factors include larger distances from highly populated areas and proximity to coastlines
  - Payload mass, launch site, and orbit have strong correlation to landing success
  - SpaceX Landing success has increased over time indicating technology and methodology has advanced over the last 10 years
- Predictive analysis results
  - Model Accuracy Score .84 for Out of Sample Accuracy
  - Tendency towards false positives





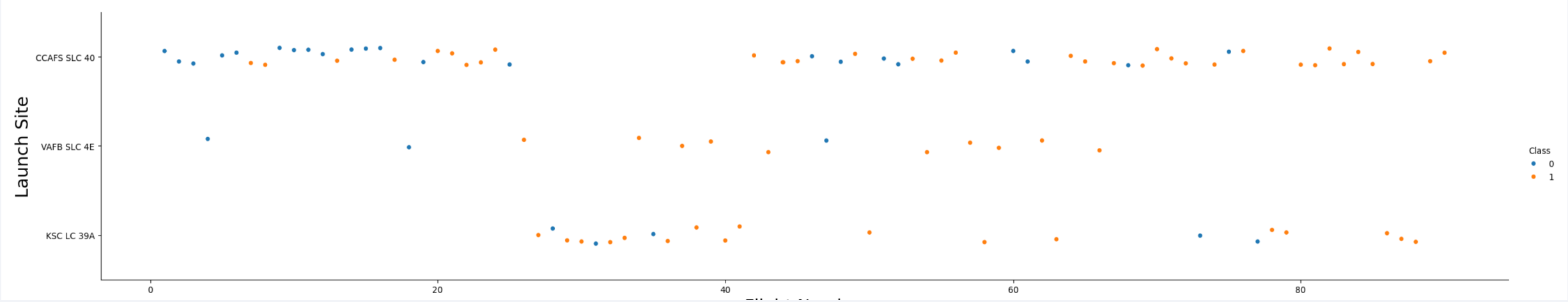



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

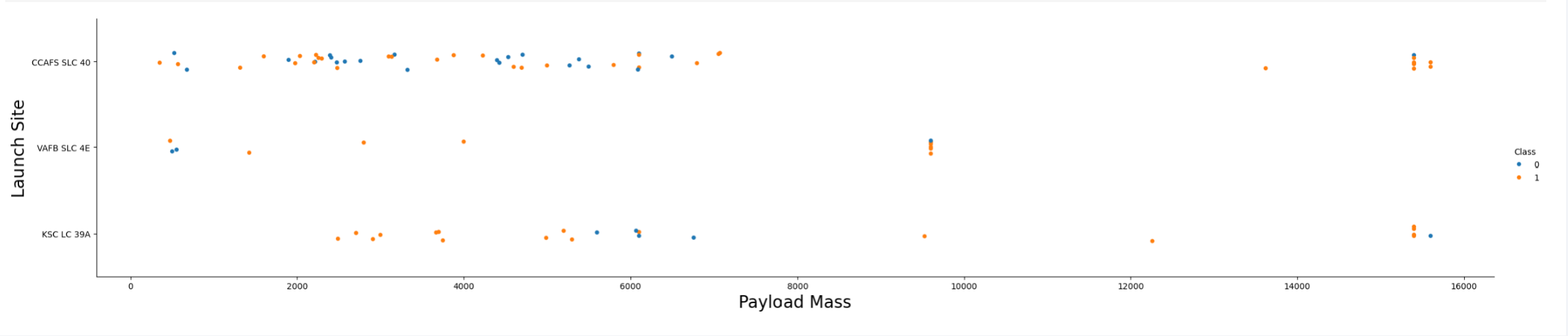
Progression from blue (failed) landings to orange (successful) landings as flight number increases indicates greater likelihood of success in later launches

Success Rate for the CCAFS SLC-40 launch site is considerably higher, indicating that there are factors at the launch site contributing to landing success



Payload vs. Launch Site

- VAFB SLC 4E is unable to handle heavier payloads
- Higher Payload masses have a higher landing success rate as indicated by the higher likelihood of orange(successful) launches

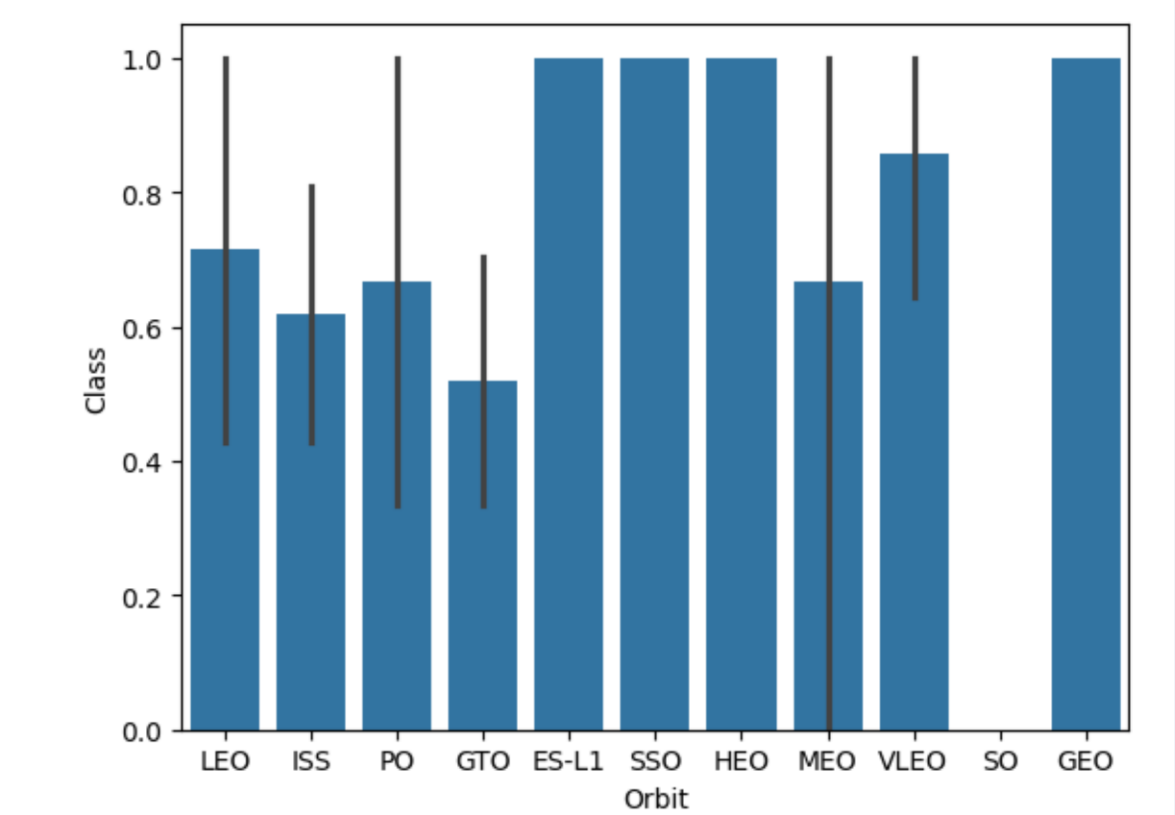


Success Rate vs. Orbit Type

The intended orbit seems to have a high bearing on landing success

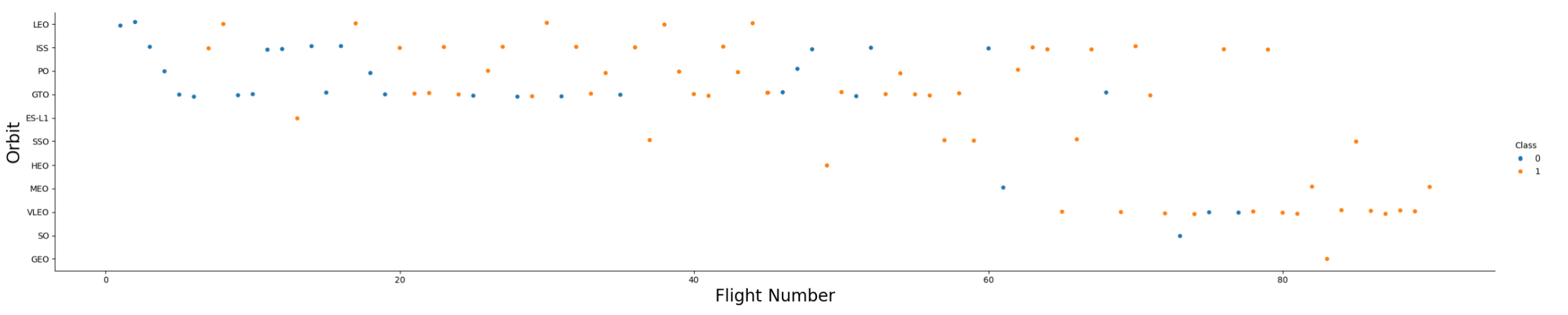
Most Successful Orbits:

- ES-L1 (low occurrence)
- SSO (low occurrence)
- HEO (low occurrence)
- VLEO (More Common recently)



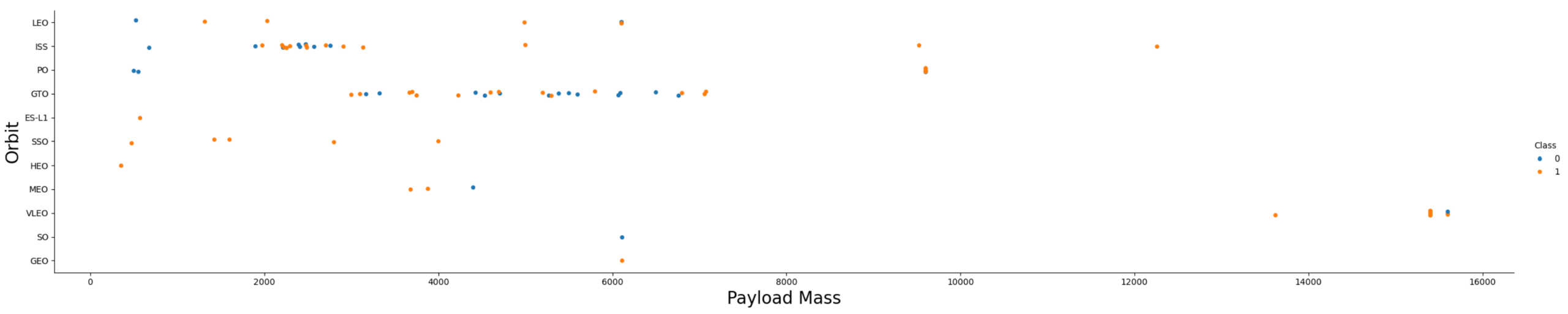
Flight Number vs. Orbit Type

- SpaceX has launch rockets with multiple orbit Types
- Later flights tend toward the VLEO orbit type. Previous chart indicates this is one of the most successful orbits for Booster landing
- GTO and ISS were more common in the past but have limited landing success



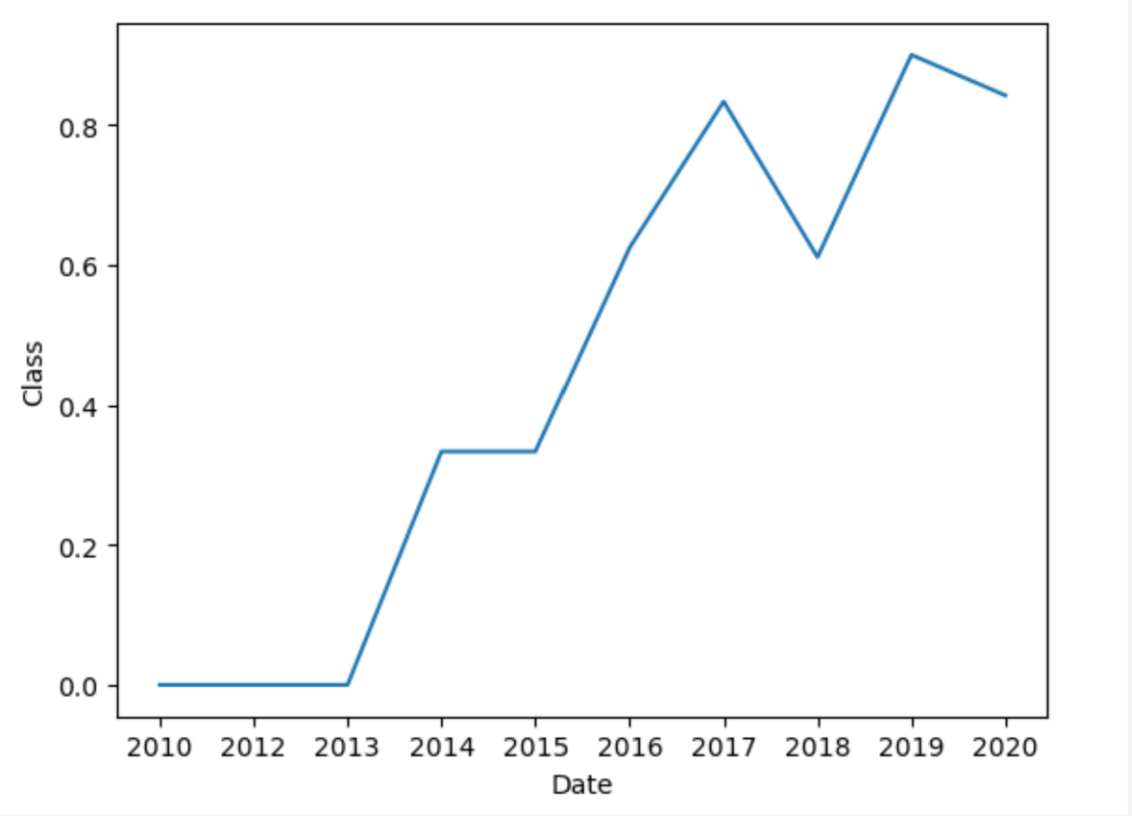
Payload vs. Orbit Type

- Chart below shows the successful and failed landings as they are distributed across payload masses for each attempted orbit



Launch Success Yearly Trend

- Over time SpaceX landing success has increased



All Launch Site Names

The below table lists the launch sites included in the analyzed data set.

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

Launch Site Names Begin with 'CCA'

The below table shows records where launch sites 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

Total Payload Mass

The Below total is the total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Below is the average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

SpaceX Data included launches beginning at the below date.

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

WE have listed below the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

Below we show that ratio of success to failed launches in our dataset.

Outcome	count(*)
Failure	40
Success	61

Boosters Carried Maximum Payload

Listed here are the names of the booster which have carried the maximum payload mass

Booster_Version
F9 B5 B1056.4
F9 B5 B1048.5

2015 Launch Records

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Launch_Site	Landing_Outcome	Booster_Version	month_num
CCAFS LC-40	Failure (drone ship)	F9 v1.1 B1012	01
CCAFS LC-40	Failure (drone ship)	F9 v1.1 B1015	04
CCAFS LC-40	Precluded (drone ship)	F9 v1.1 B1018	06

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

Rank 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

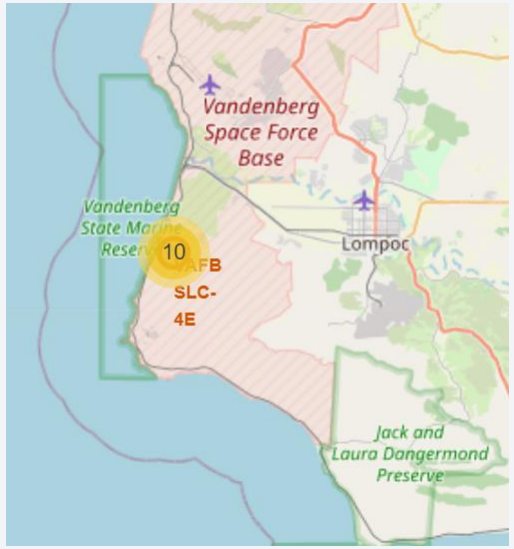
Landing_Outcome	count(*)	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

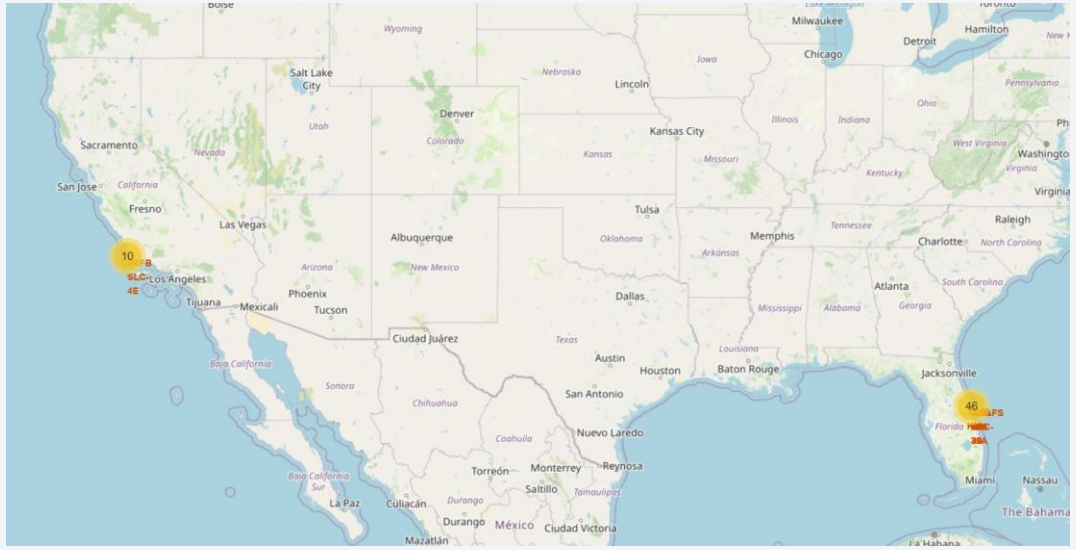
Section 3

Launch Sites Proximities Analysis

Launch Site Geo-Location Analysis

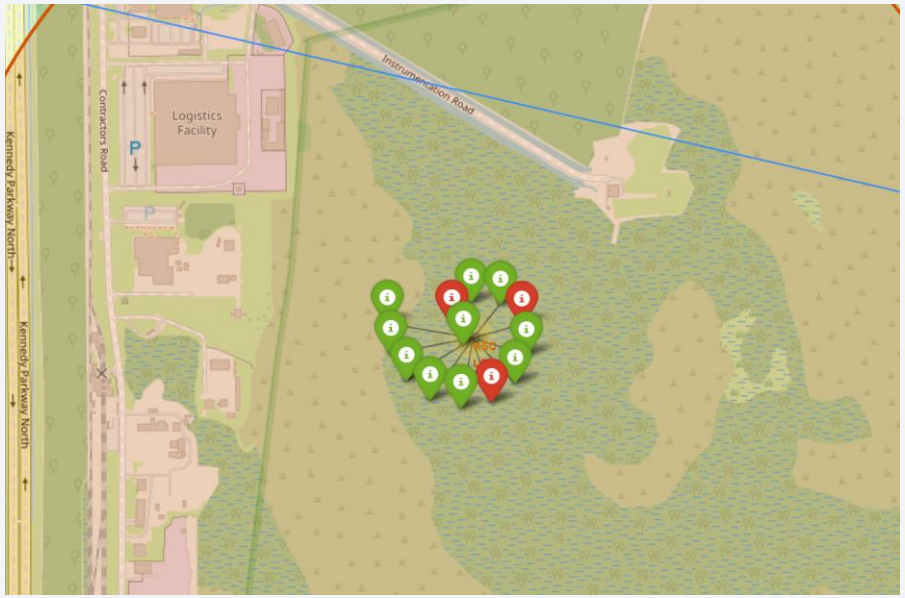
Launch Sites were mapped using Folium to perform an analysis of launch site attributes and the landing success and failure at each site.

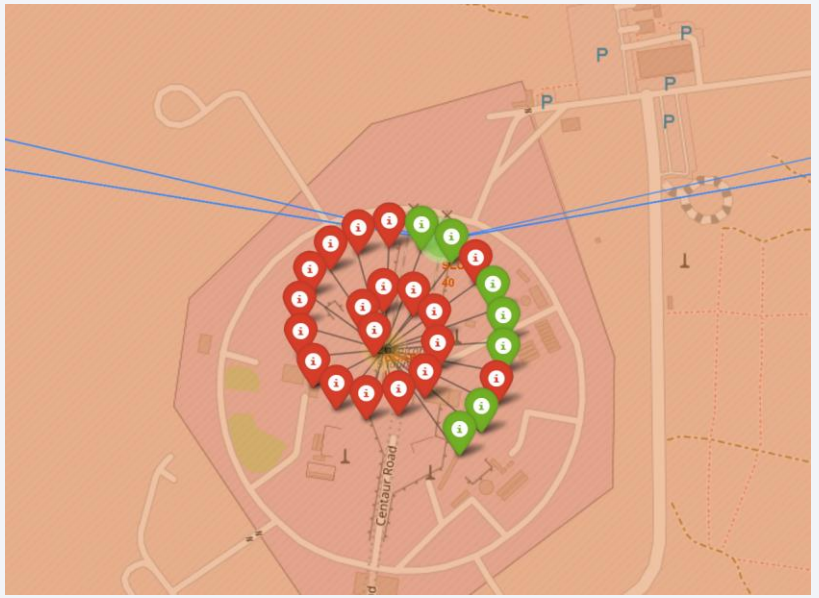




Landing Success Geo-Location Analysis

Density of successful landings depicted for each site shows significance of the launch location for landing success(Folium Marker Groups)

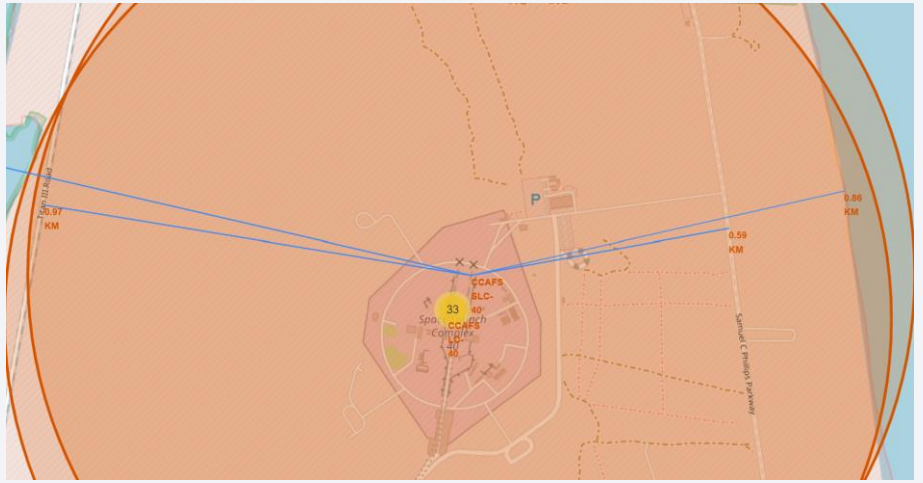


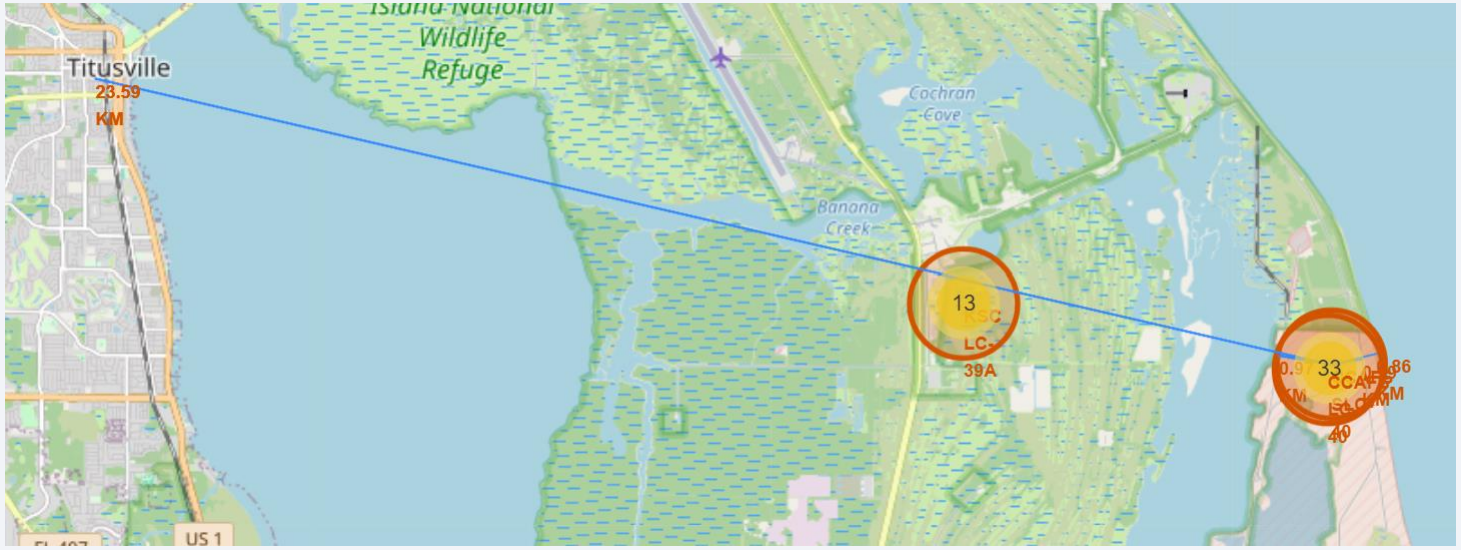


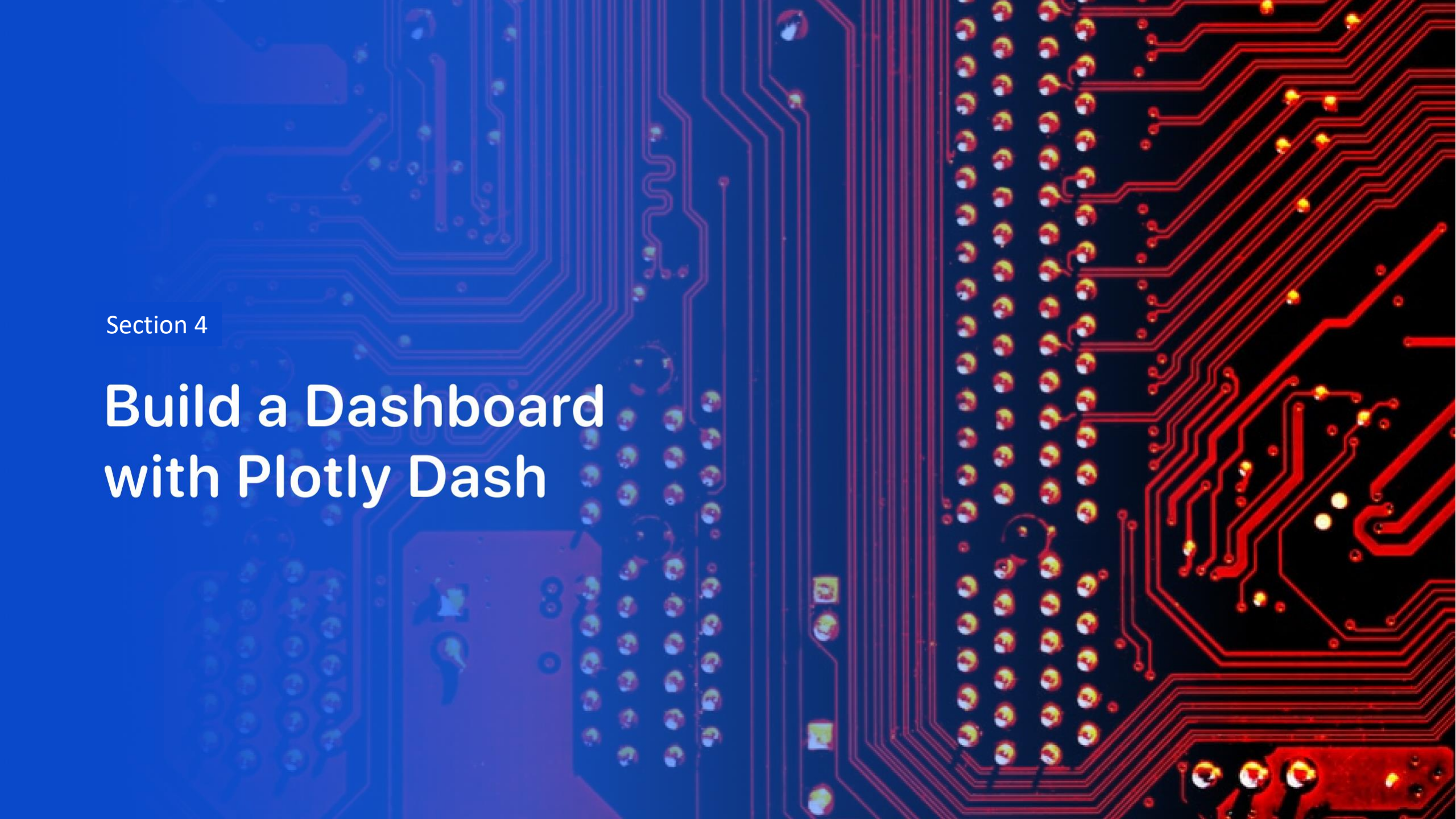
Launch Site Positioning Analysis

Key Takeaways regarding Launch Site Location

- Near Coast, Highways, and Railways
- Significant distance from population centers like cities.



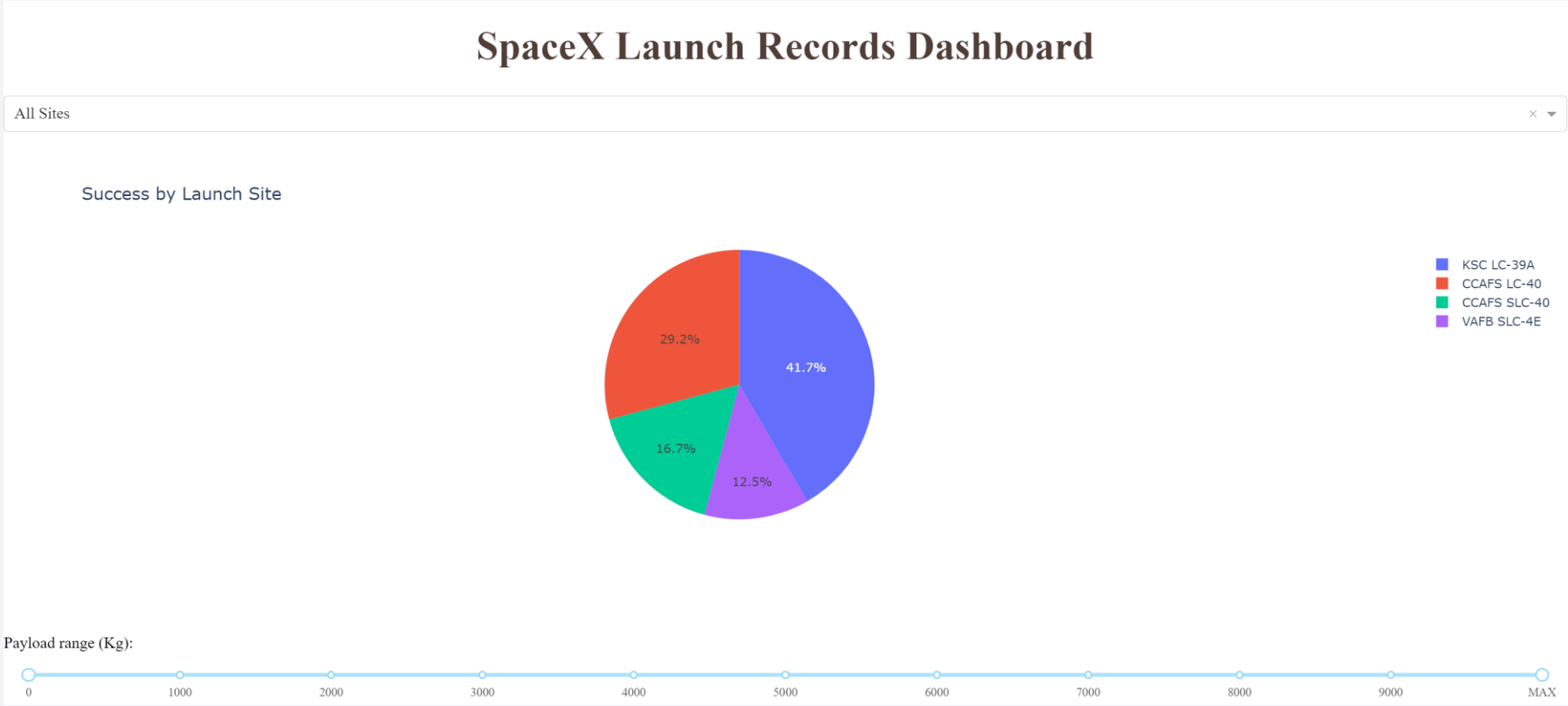



Section 4
Build a Dashboard with Plotly Dash

SpaceX Dashboard with Plotly Dash – Launch Sites

Pie Chart of Landing Success by Launch Site

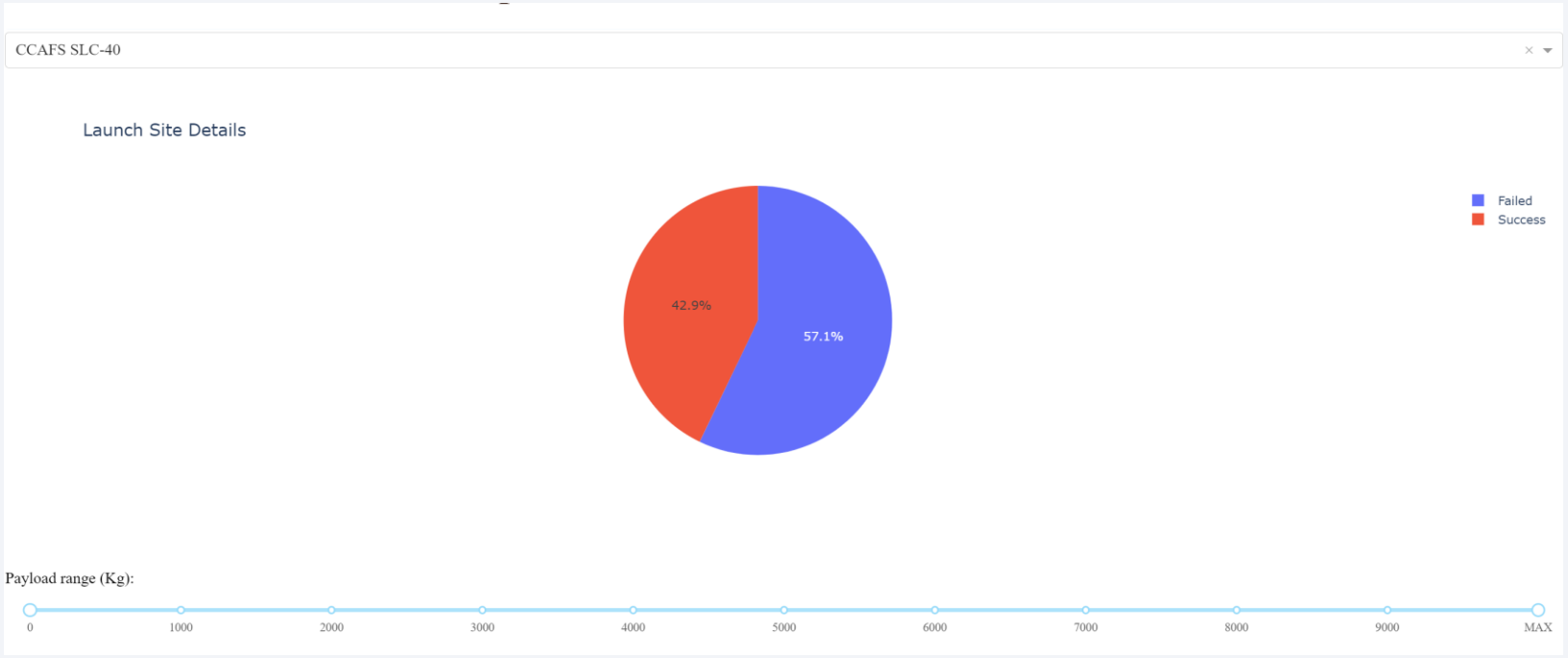
- KSC LC-39A has the largest number of successful landings
- Note: KSC LC-39A also has the largest number of launches



SpaceX Dashboard with Plotly Dash – Launch Sites

Site wit the highest Landing success

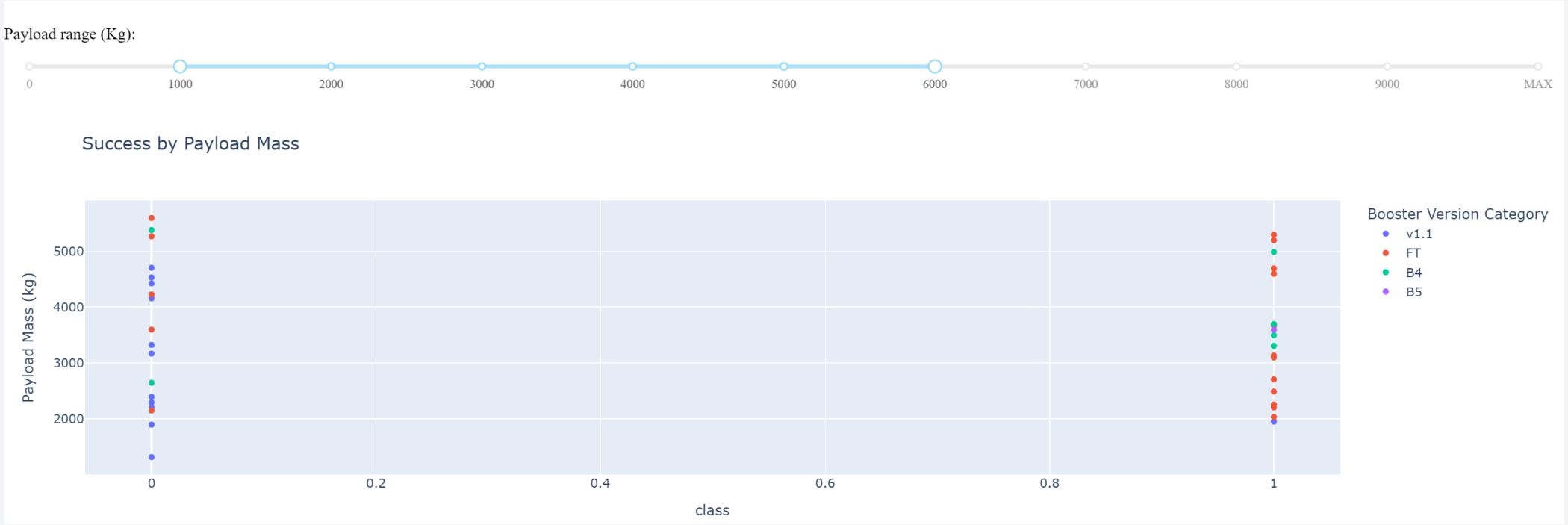
- CCAFS SLC-40 has the highest launch success
- Best Practice Benchmarking should focus on practices with this launch site



SpaceX Dashboard with Plotly Dash – Payload Mass

We examined the landing success rates for various payload mass ranges

The Bulk of launches had a payload mass ranging from 1000 to 6000 Kg.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Out-Of-Sample Accuracy

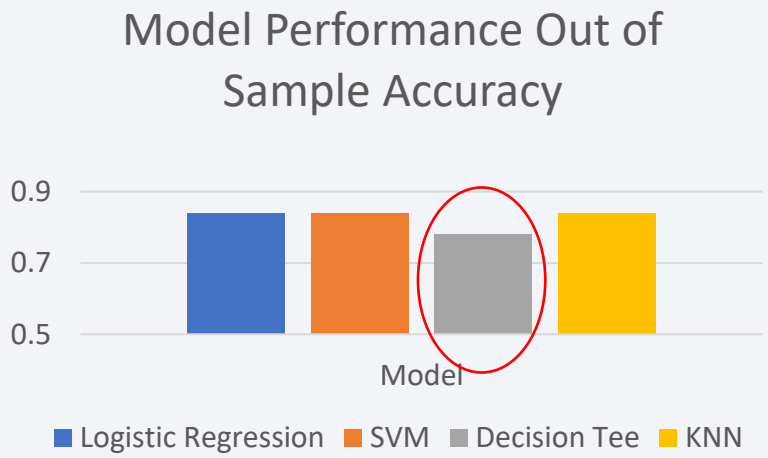
- Decision Tree Classifier underperforms as compared to other models .77 vs. .84
- Logistic Regression, SVM, and KNN models score equally well with identical confusion matrices tending towards false positive

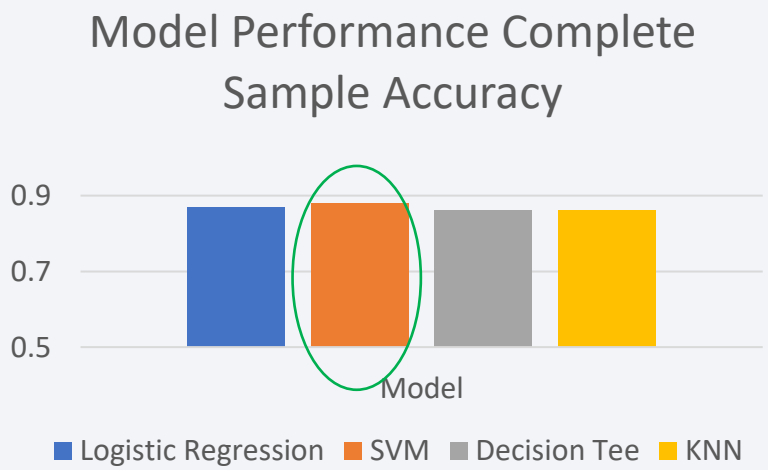
Complete Sample Accuracy

- Support Vector Machine out-performs other models across the complete sample - .88 accuracy score
- Confusion continue to show a tendency towards false positive for the complete sample

Recommendation

- Move forward using Support Vector Machine Learning for predictive needs



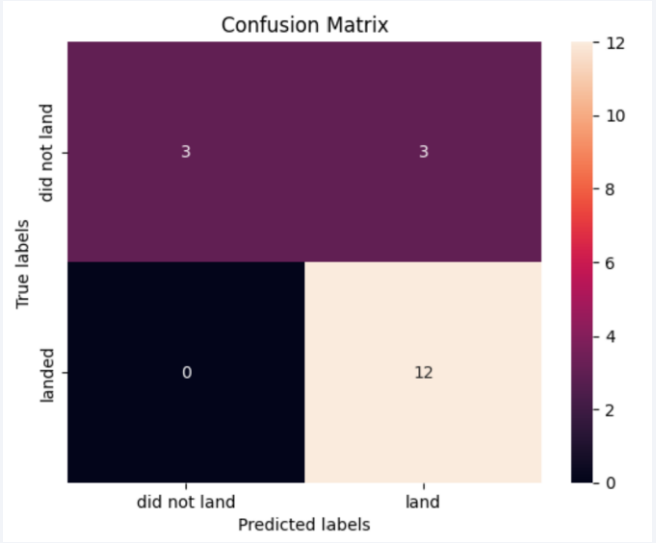


Confusion Matrix

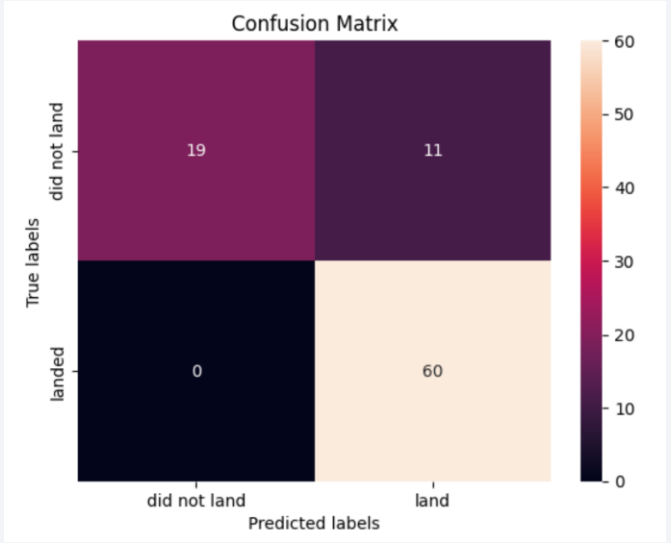
Recommended Modeling Method – Support Vector Machine Learning

- Confusion Matrix and Accuracy Score match the best performing models for Out of Sample Evaluation
- Highest accuracy Score and best confusion Matrix across the complete sample

Out-Of-Sample



Complete Sample



Conclusions

- Support Vector Machine Learning model was the most accurate predictive methodology
- The accuracy of the model allows for base budget planning based on launch parameters
- All models have a tendency towards false positives, indicating we will budgetary considerations for potential launch failure regardless of the predicted outcome

Appendix

All related code can be found here - https://github.com/GDaddySoul/Data-Science-Capstone

Thank you!