IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Árpád Szenczi
19Feb2024

Outline

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

Executive Summary

Objective: Predicting SpaceX's Falcon-9 first stage reusability for cost-effective space launches.

Methodologies:

- Utilized SpaceX REST API and Python's BeautifulSoup for data collection.
- Conducted data wrangling, exploratory data analysis (EDA), and interactive visual analytics.
- Employed predictive analysis techniques.

Key Findings:

- Success rate improved, reaching approximately 80% during 2017-2020.
- Payloads within the range of 2000-6000 kg showed the highest likelihood of successful landings.
- Orbit type and launch site significantly influence landing outcomes.
- Logistic regression achieved high accuracy in predicting landing outcomes.

Introduction

- Project background and context:
  - Using past Falcon 9 launches data the task is to predict if the Falcon 9 first stage will land successfully in order to reuse it and decreasing the cost of launches.
- Problems you want to find answers:
  - Find the appropriate parameters which are the most suitable for performing predictive analysis from the obtained data
  - Visualizing and presenting the results

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data was requested from SpaceX API and converted into a Pandas DataFrame
  - Web scrape Wiki page containing launch data and converted into a Pandas DataFrame
- Perform data wrangling
  - Data transformation was done on the dataframe to create new variable for landing outcomes
- Exploratory data analysis (EDA) was applied using visualization and SQL on the cleaned dataframe and relevant correlation/trends between different parameters were revealed

Methodology

Executive Summary (cont.)

- Exploratory data analysis (EDA) was applied using visualization and SQL on the cleaned dataframe and relevant correlation/trends between different parameters were revealed
- Interactive visual analytics also applied using Folium and Plotly Dash and additional insights regarding launch sites locations identified (e.g. best performing site)
- Predictive analysis was done using classification models LR, SVM, DT and KNN. Classification accuracy was calculated for each and confusion matrix confirmed the best model to use: LR

Data Collection

API

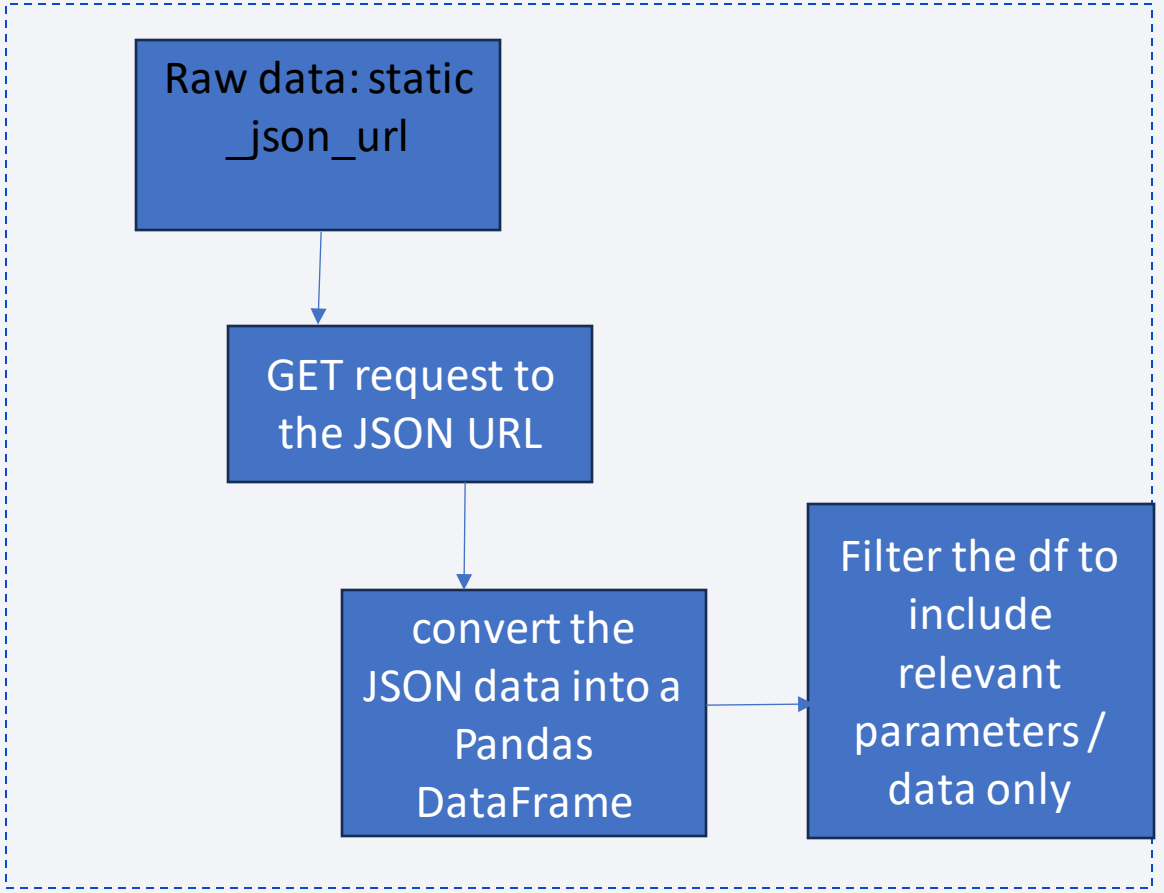
- Data was requested from SpaceX API with the following URL https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- get request applied to obtain relevant data
- json_normalize was used to convert the JSON data into a Pandas data frame

Webscraping

- Data was obtained from Wikipedia page https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Falcon 9 launch data extracted from HTML table
- After parsing the table it was converted into a Pandas data frame

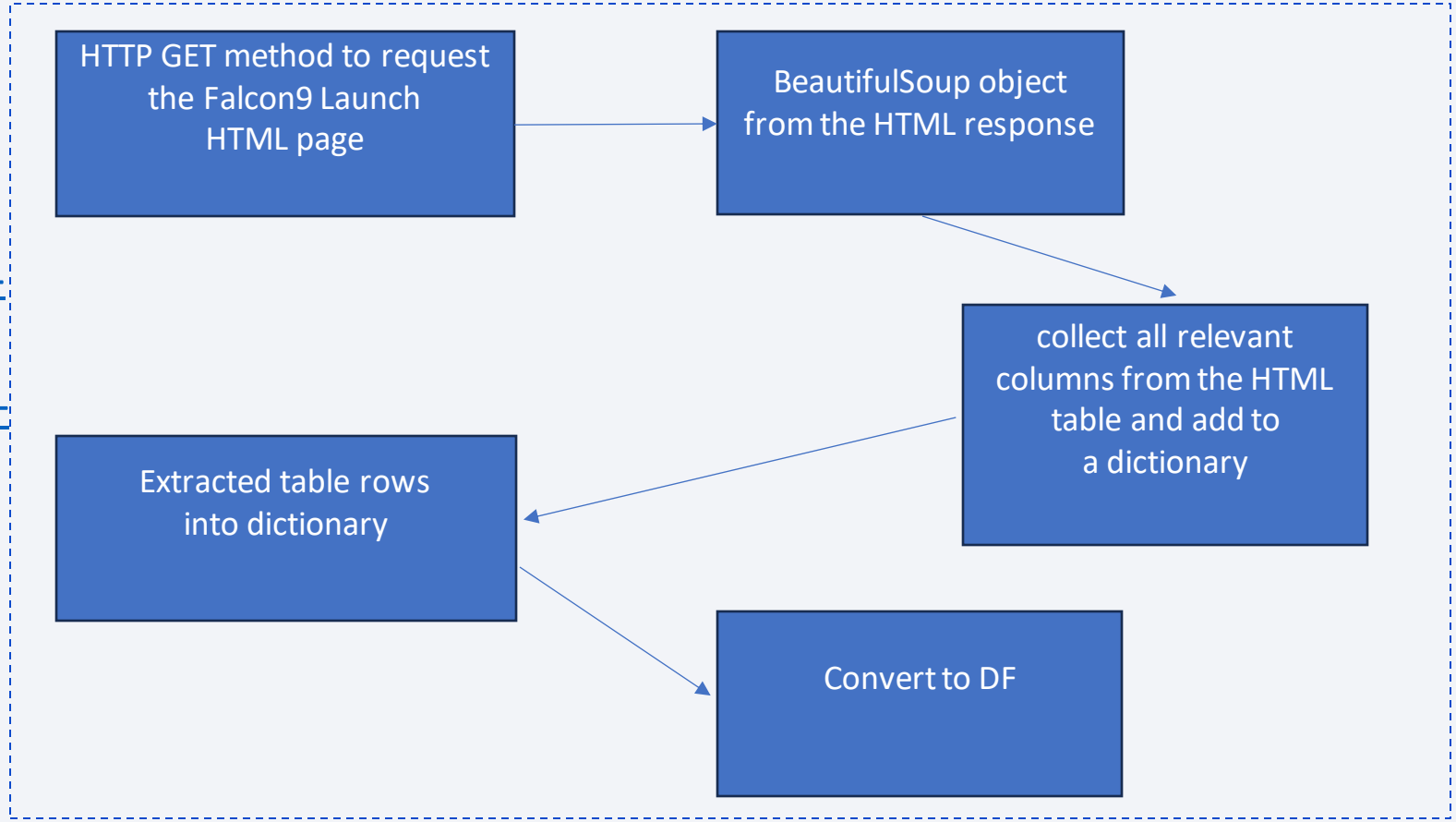
Data Collection – SpaceX API

- GitHub URL:
- https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/jupyter-labs-spacex-data-collection-api.ipynb



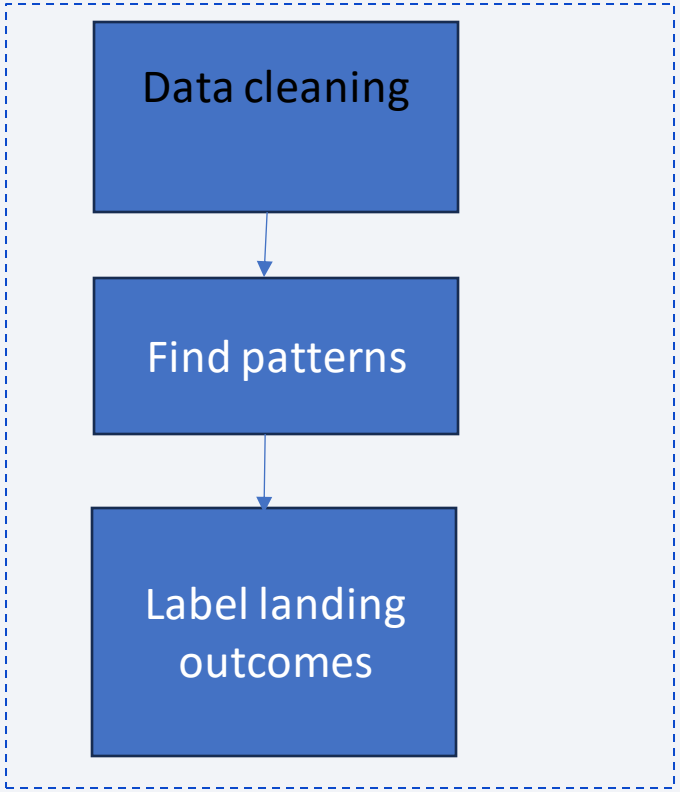
Data Collection - Scraping

- GitHub URL:
- https://github.com/mappy1945/Coursera-Capstone/blob/ff3c0fc4f63ef326c187c9d6348fc92900e44576/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- Aim is to label successful/unsuccessful landing in the dataframe to help building predictive models afterwards
- Find some patterns in the data (launch site, orbit type, landing outcomes statistics) to make it useful for analysis
- Landing outcomes scenarios were categorized 0 or 1 and added to DF
- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plot was used to see correlations between
  - flight number vs payload mass
  - flight number vs orbit type
  - flight number vs launch site
  - orbit vs payload mass
- Conclusion /statements were made as a result of adding landing outcomes to the visualization (see slides 20-25)
- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/jupyter-labs-eda-dataviz.ipynb.jupyterlite2.ipynb

EDA with SQL

- SQL queries performed
  - names of the unique launch sites
  - 5 records where launch sites begin with the string 'CCA'
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - List of date when the first successful landing outcome in ground pad was achieved.
  - List of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL (cont.)

- SQL queries performed (cont.)
  - List of total number of successful and failure mission outcomes
  - List of names of the booster_versions which have carried the maximum payload mass.
  - List of records displaying the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
  - Rank of 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.

- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map was generated and different map objects added to the folium map in order to find geographical patterns about launch sites
- Circles were added to the map to highlight the location and markers were applied to label the launch sites
- Circles were customized to visualize the successful launches in addition to the location
- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

- Interactive visual analytics on SpaceX launch data in real-time
- Dropdown list and a range slider were applied to interact with a pie chart and a scatter point chart
- Using these plots helped to answer the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version has the highest launch success rate?
- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/spacex_dash_app.py

Predictive Analysis (Classification)

- Column created for the class in the dataset
- Data standardization was made
- Dataset was split into training data and test data
- Best hyperparameter for SVM, Classification Trees and Logistic Regression, KNN found
- Method performs best using test data

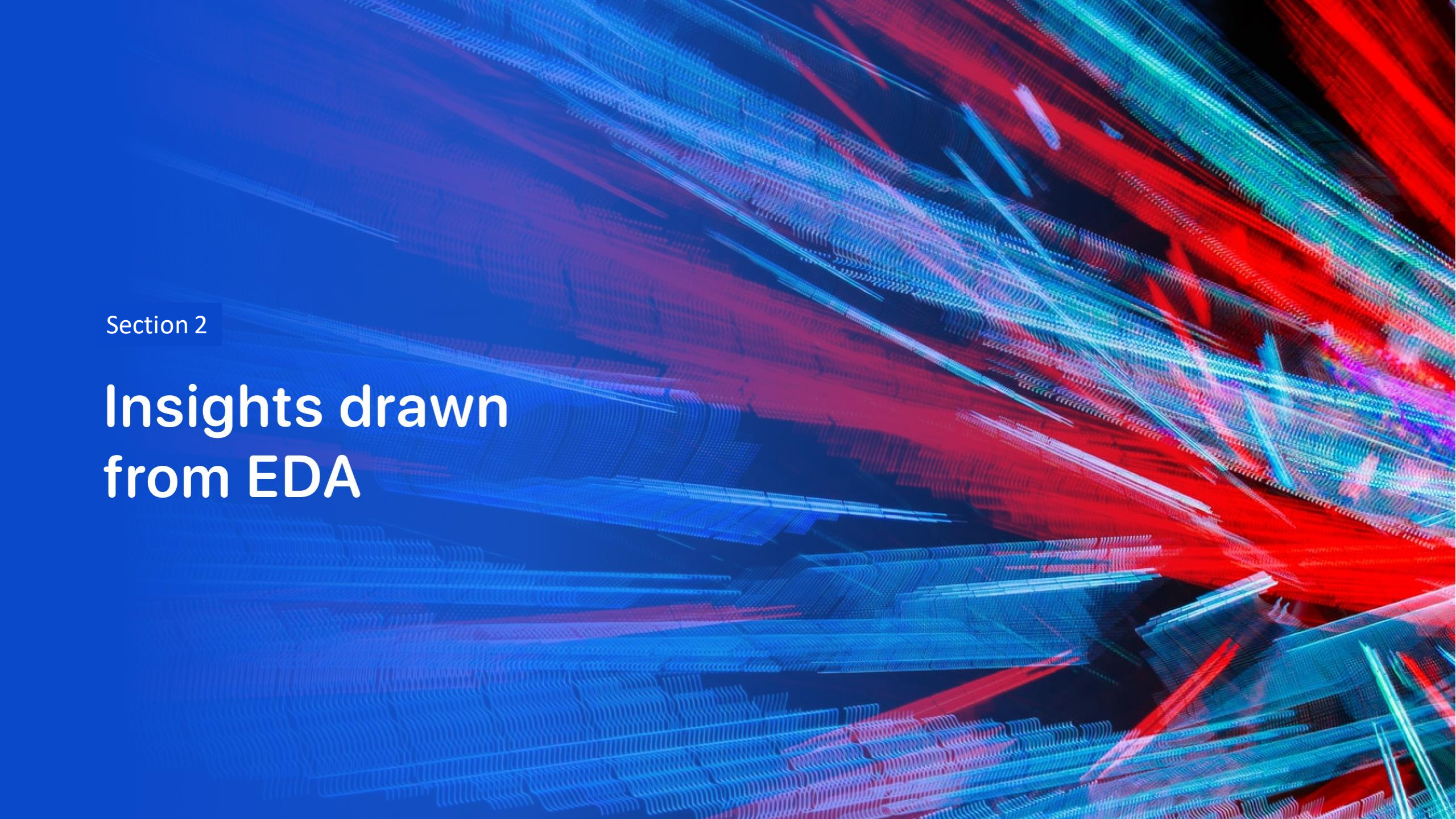
- GitHub URL: https://github.com/mappy1945/Coursera-Capstone/blob/982978c53b524bbebe3688aec3a054ae66ab323e/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

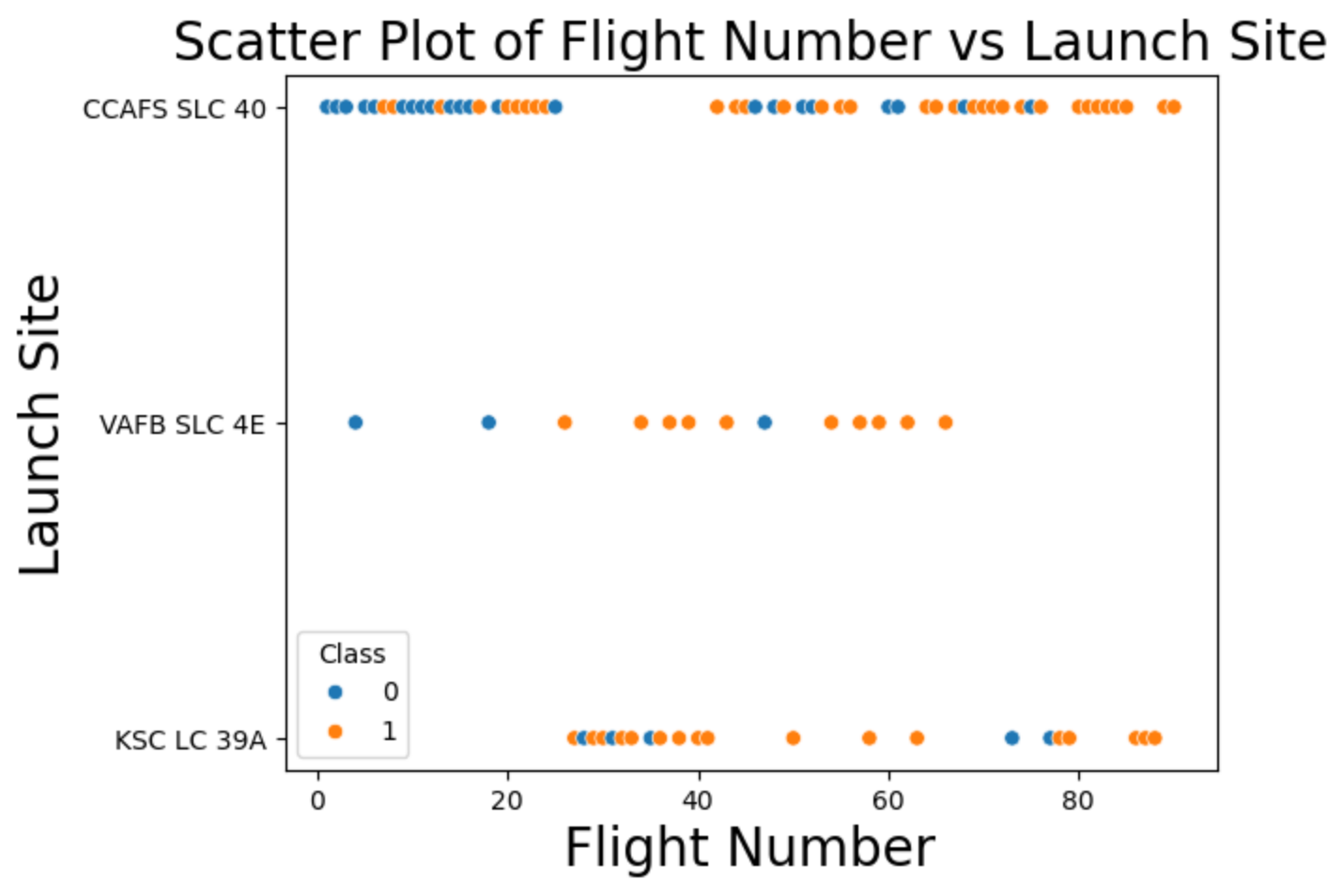
Section 2

Insights drawn from EDA



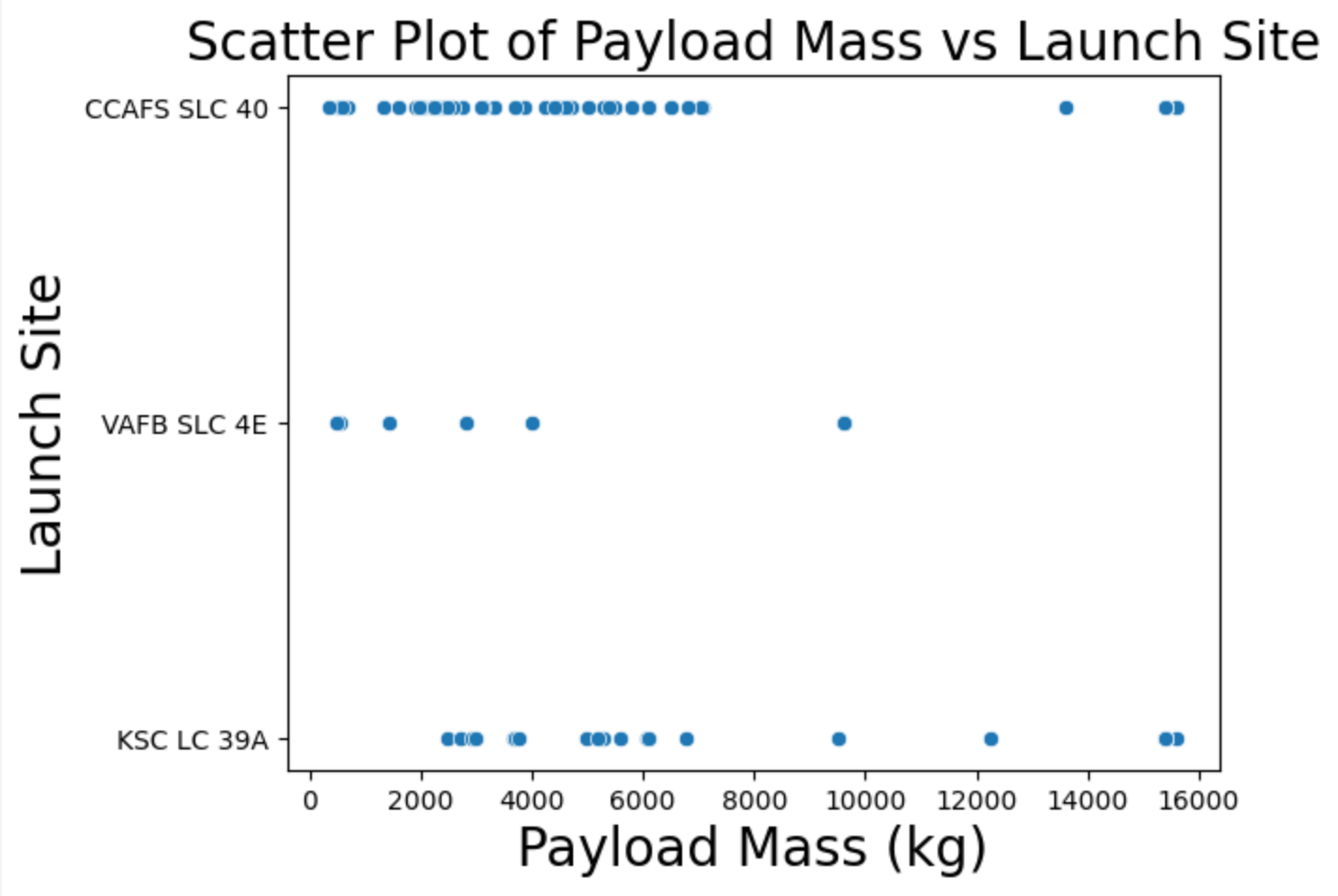
Flight Number vs. Launch Site

- We see that different launch sites have different number of launches and success rate
- Successful landing improved with time



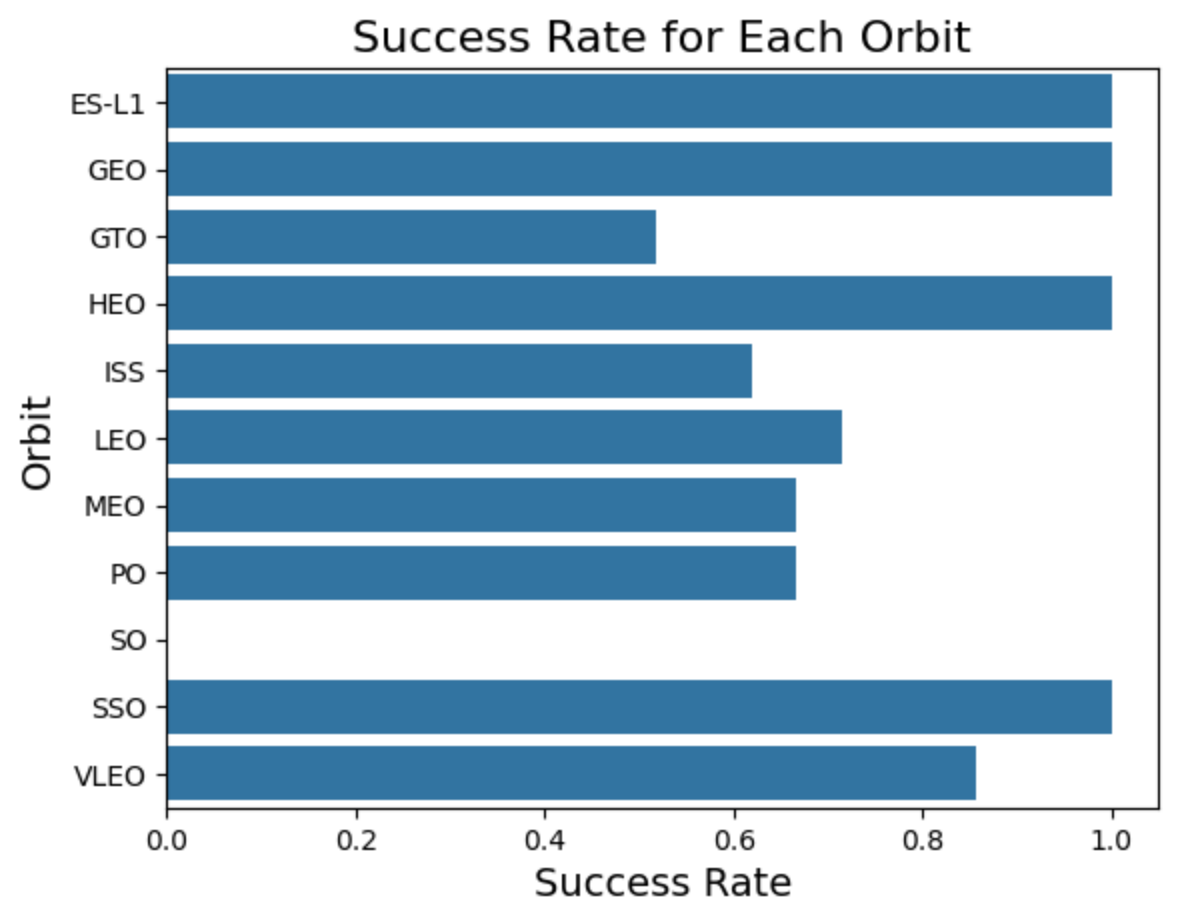
Payload vs. Launch Site

- there are no rockets launched for heavy payload mass (greater than 10000) at VAFB-SLC launchsite



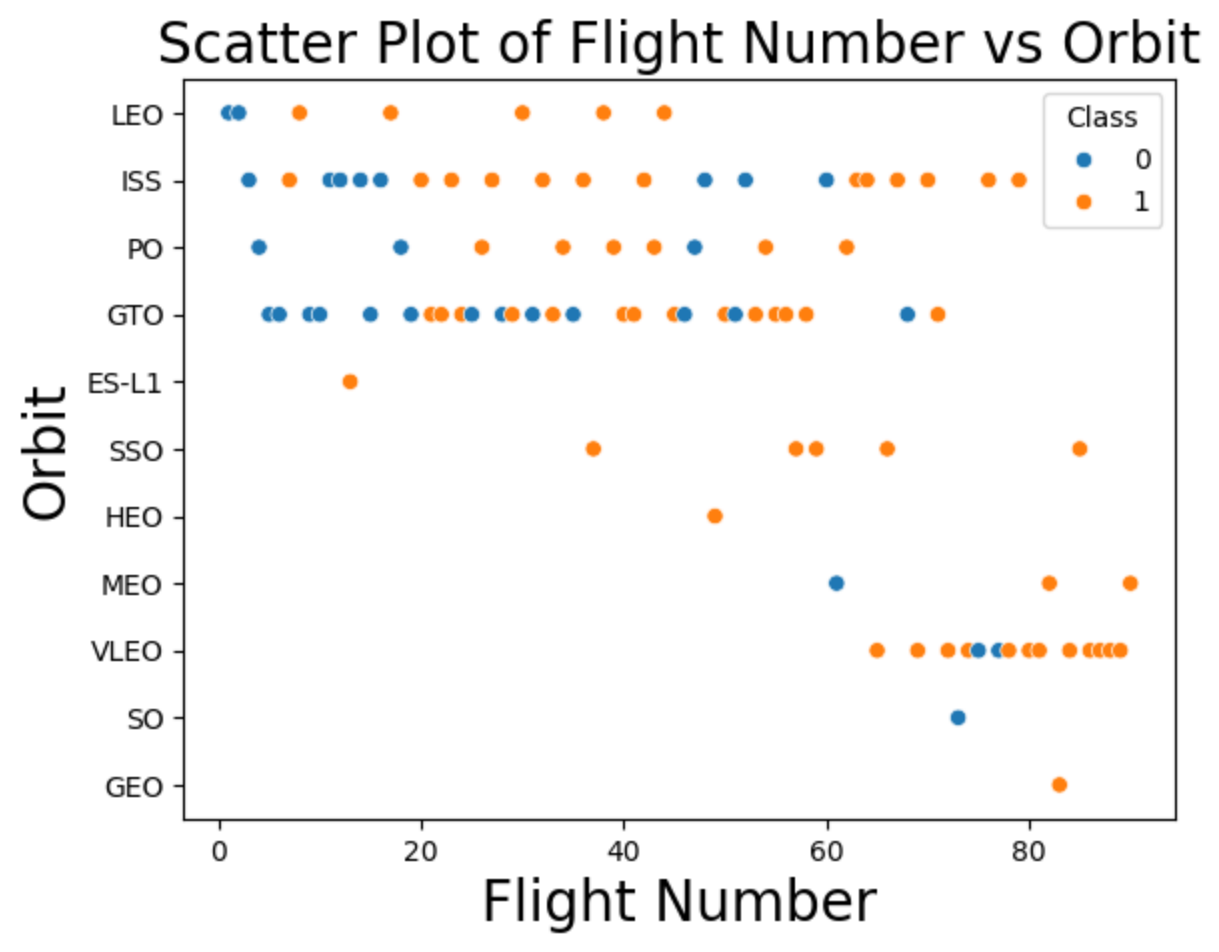
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbit have 100% success rate
- SO orbit has 0% success rate
- The remaining orbits have average 60-70% success rate



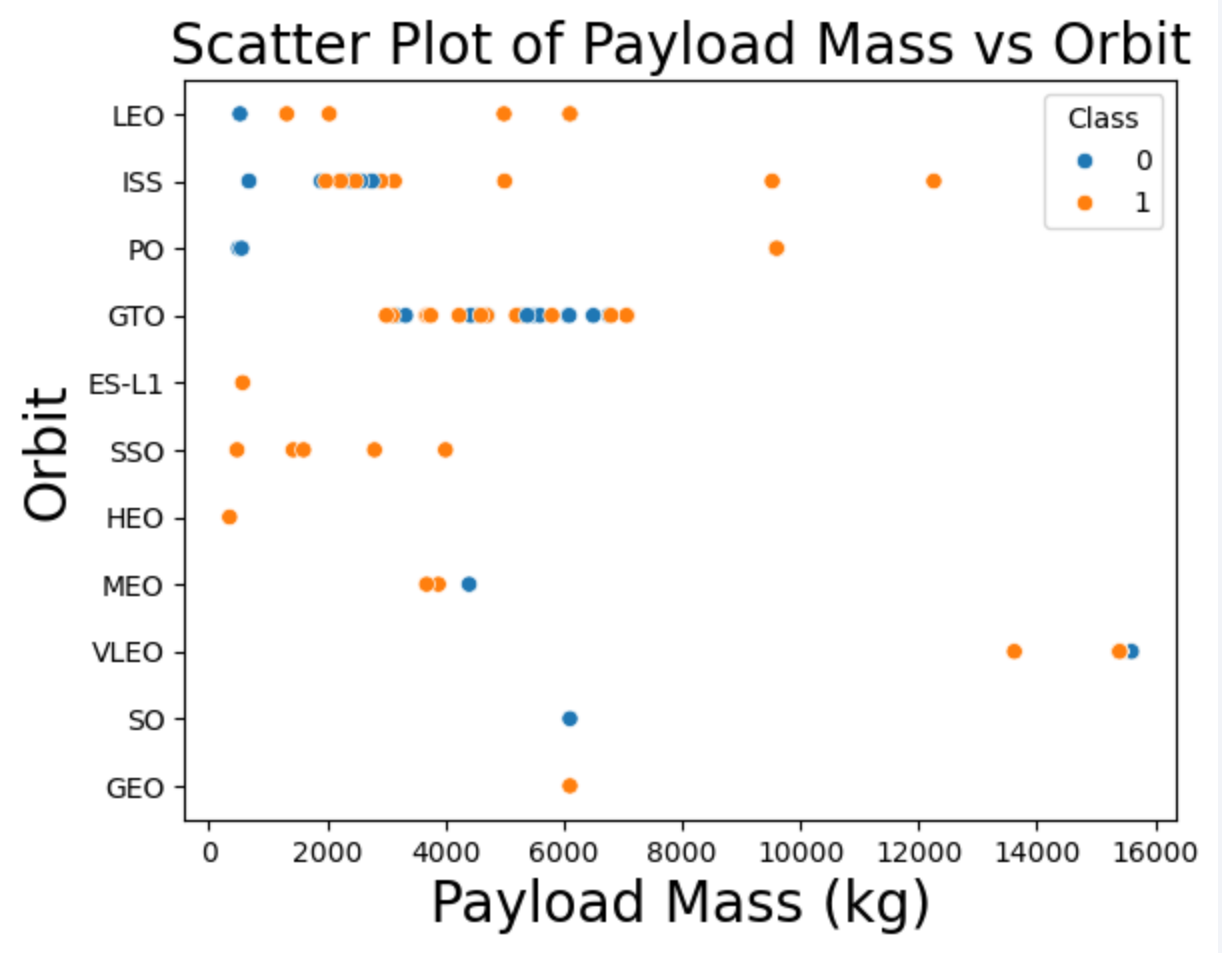
Flight Number vs. Orbit Type

- LEO orbit: the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- There are four different orbits with only one launch



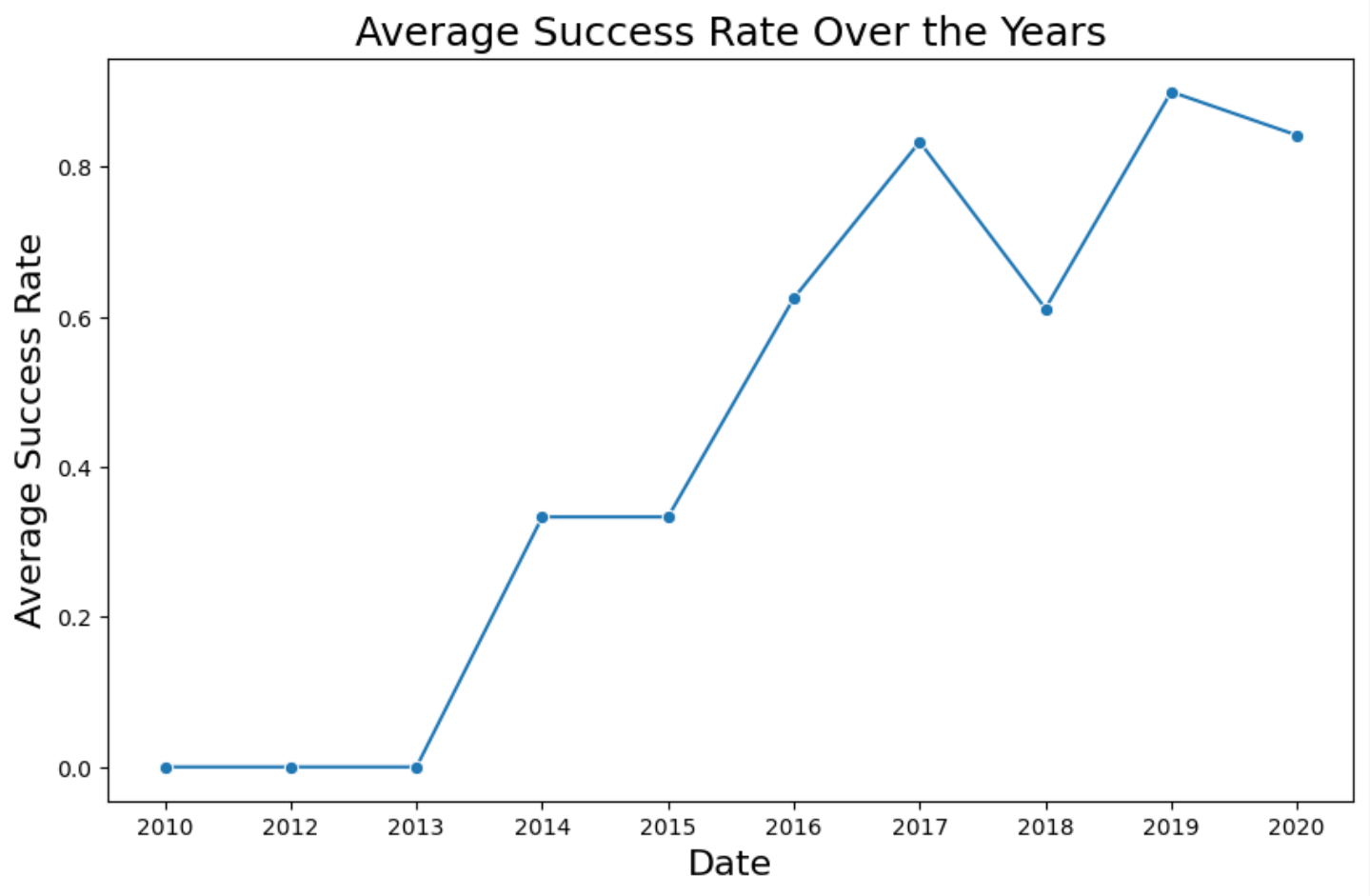
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, VLEO and ISS
- GTO we cannot distinguish this well as both positive landing rate and negative landing are both occurred



Launch Success Yearly Trend

- success rate since 2013 kept increasing till 2020



All Launch Site Names

- The following Falcon 9 launch sites were extracted from the dataset

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

Launch Site Names Begin with 'CCA'

- SQL query was limited to 5 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

- Total payload carried by boosters from NASA was calculated with SQL query - resulted 45 596 kg

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

AveragePayloadMass
2534.6666666666665

First Successful Ground Landing Date

- Date of first successful landing outcome on ground pad

FirstSuccessfulLandingDate
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 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

- Total number of successful and failure mission outcomes

Mission_Outcome	TotalCount
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

- Failed landing outcomes in drone ship with their booster versions, and launch site names in year 2015

Month	Mission_Outcome	Landing_Outcome	Booster_Version	Launch_Site
01	Success	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Success	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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
- In most of the cases, there was no landing attempt

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

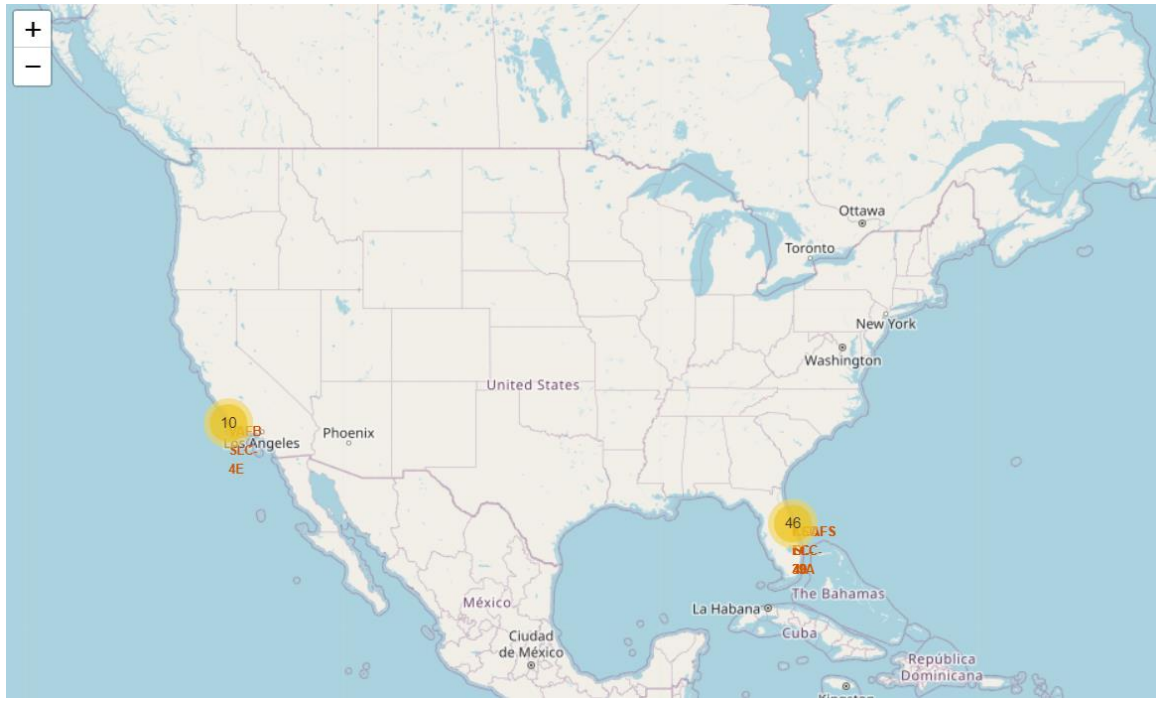

Section 3
Launch Sites
Proximities Analysis

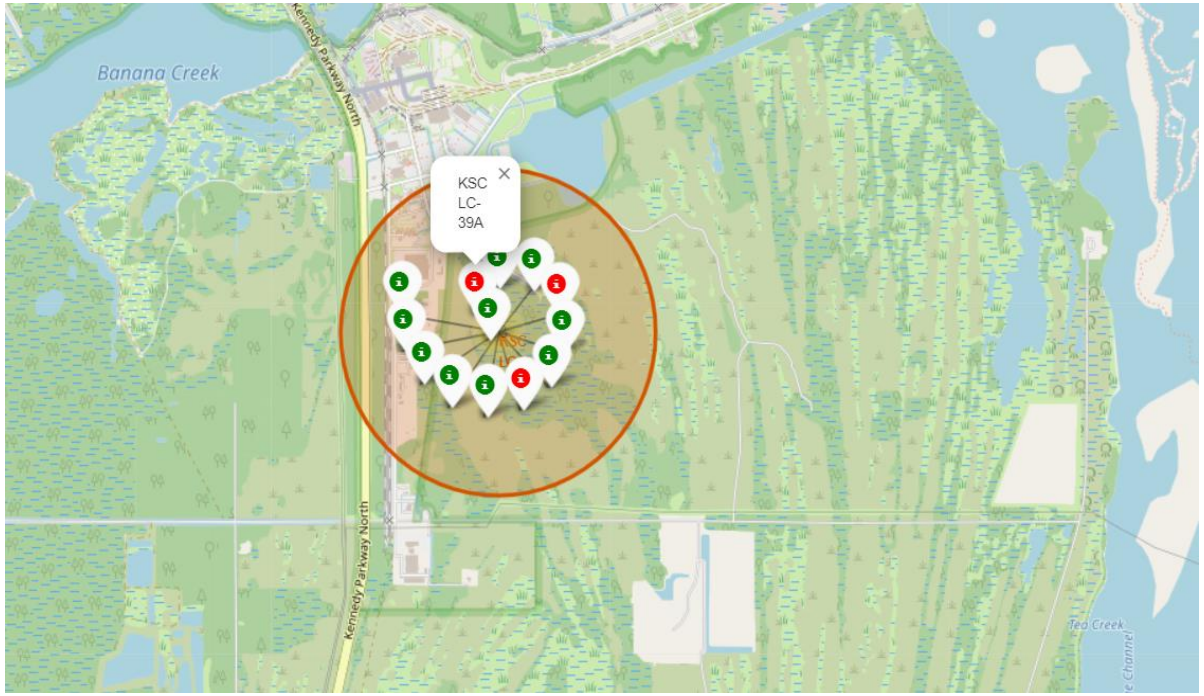
Launch Sites Location

- Launch sites are located
  - very close proximity to the coast
  - in proximity to the Equator line (but not very close)

Successful Launch Outcomes on map

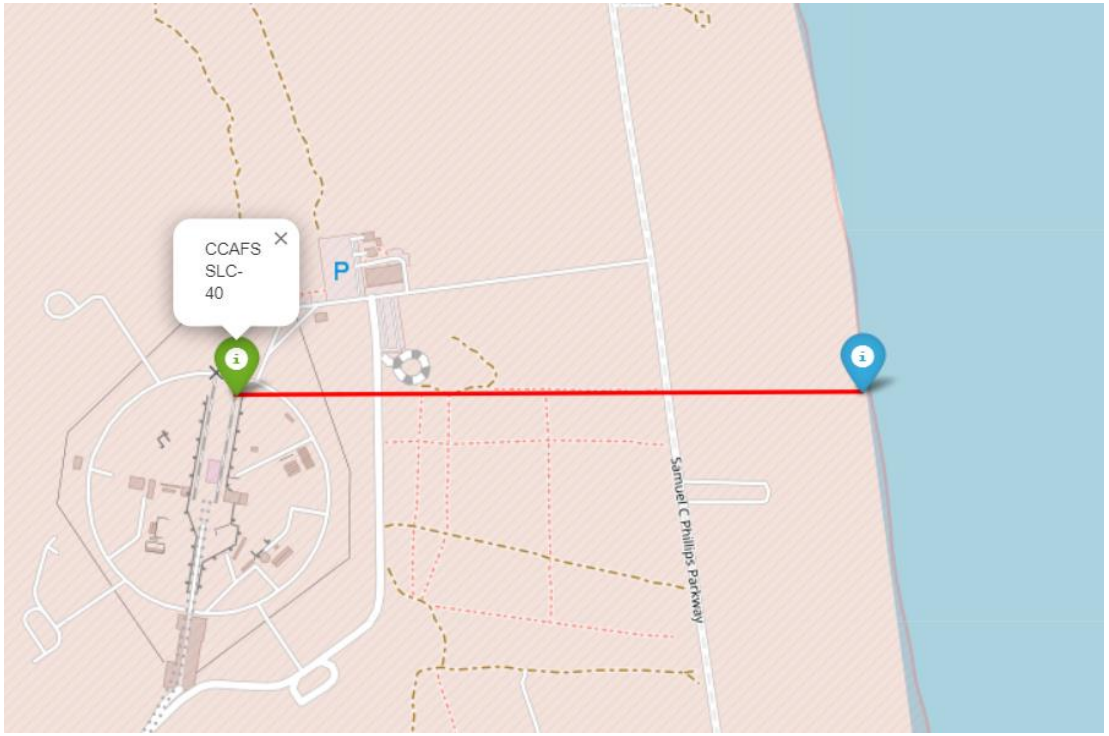
- KSCLC-39A site has the highest success rate

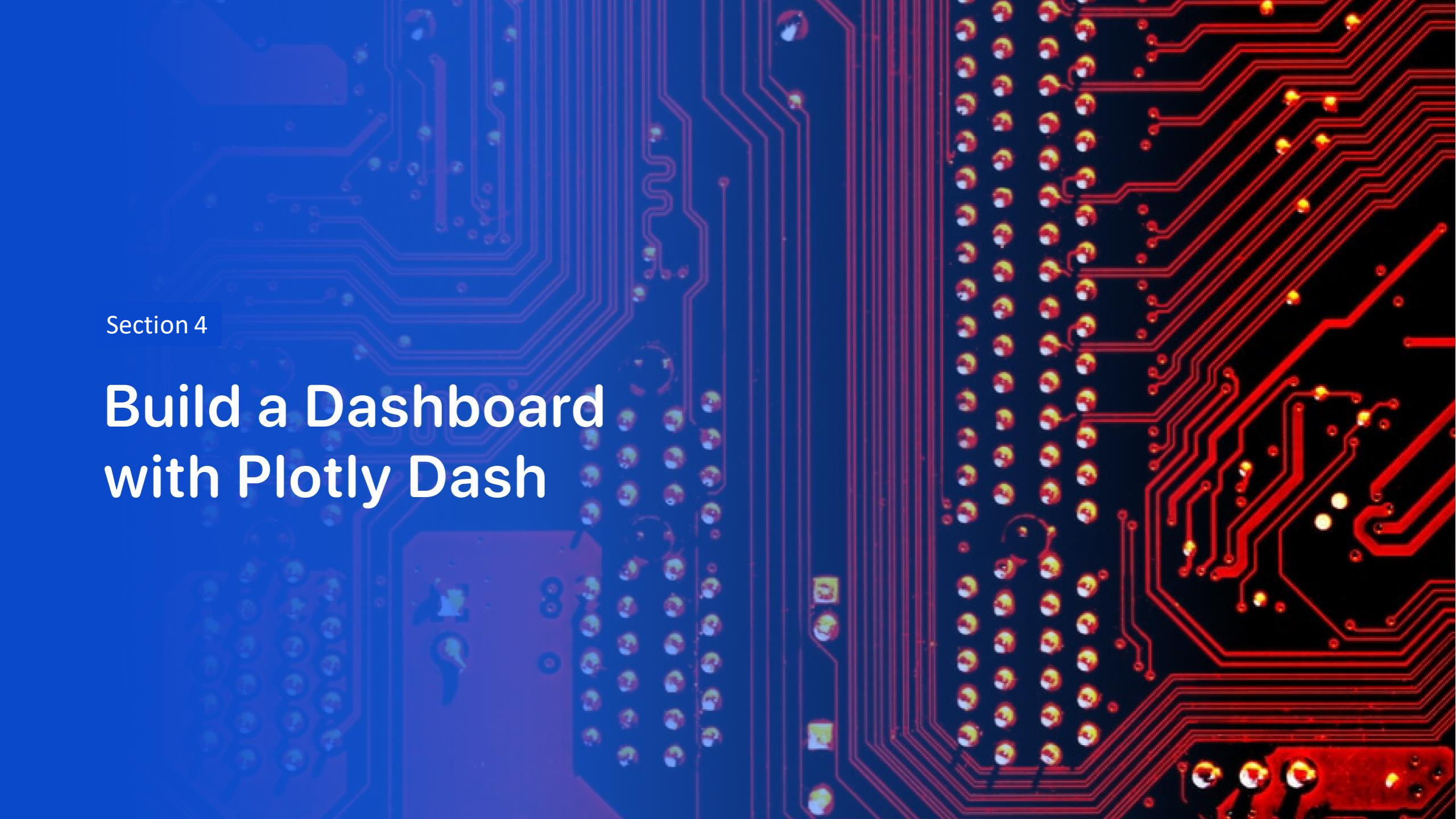




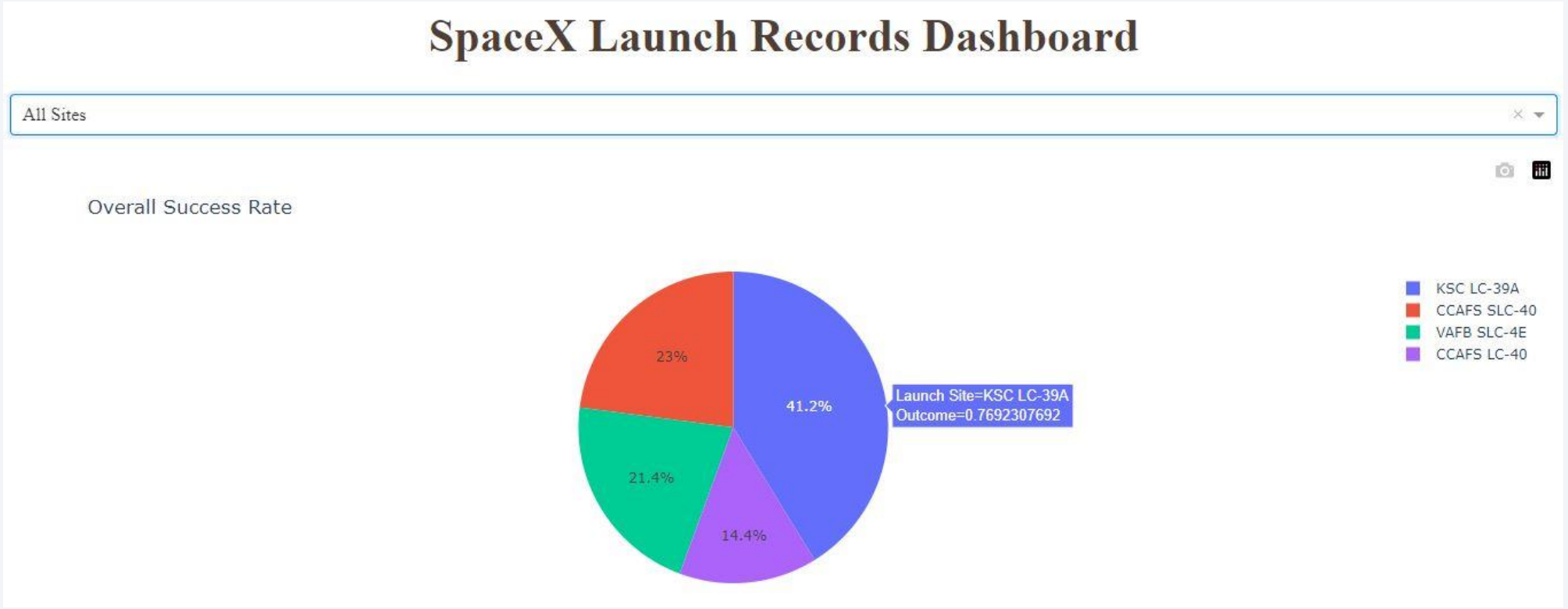
Distance to coastline

- Launch site CCAFS SLC-40 is <1 km from coastline
- No highway/railroads in close proximity




Section 4
Build a Dashboard with Plotly Dash

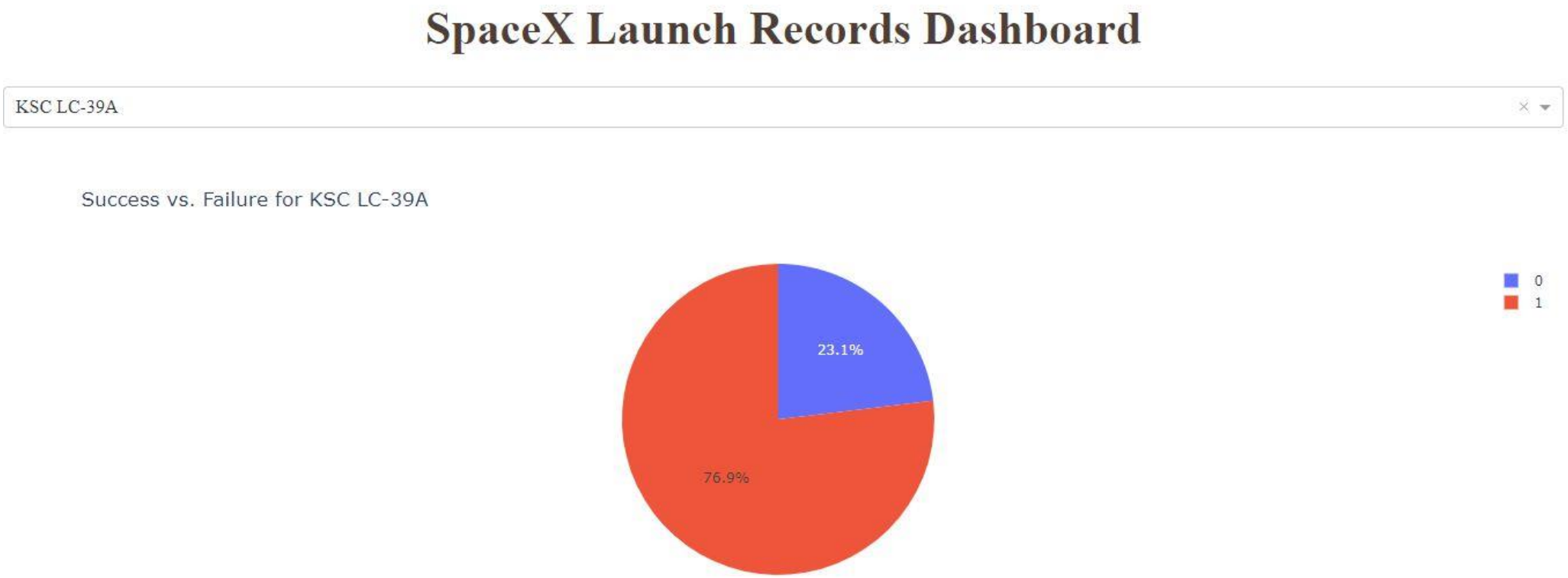
SpaceX Launch Records - Overall Success Rate



- 41% of all success landing occurred at site KSC LC-39A

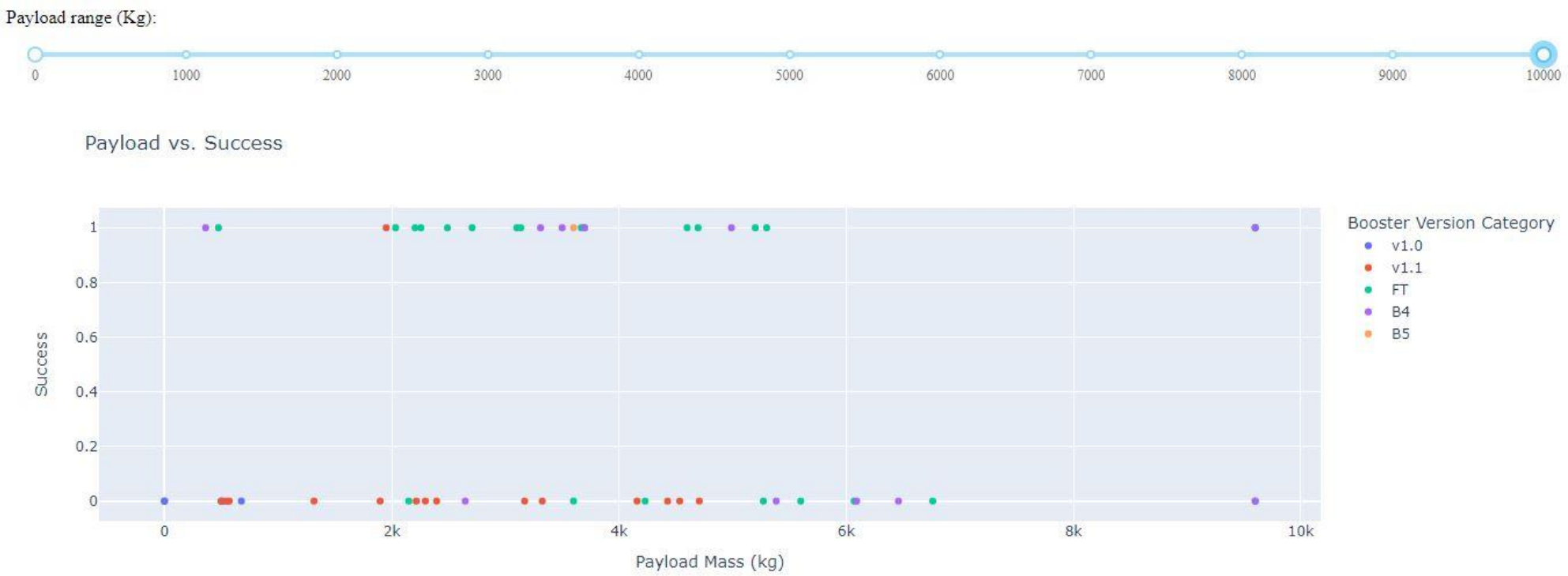
Launch site with highest launch success rate

- Launch site KSC LC-39A showed the highest successful landing rate with 76.9%



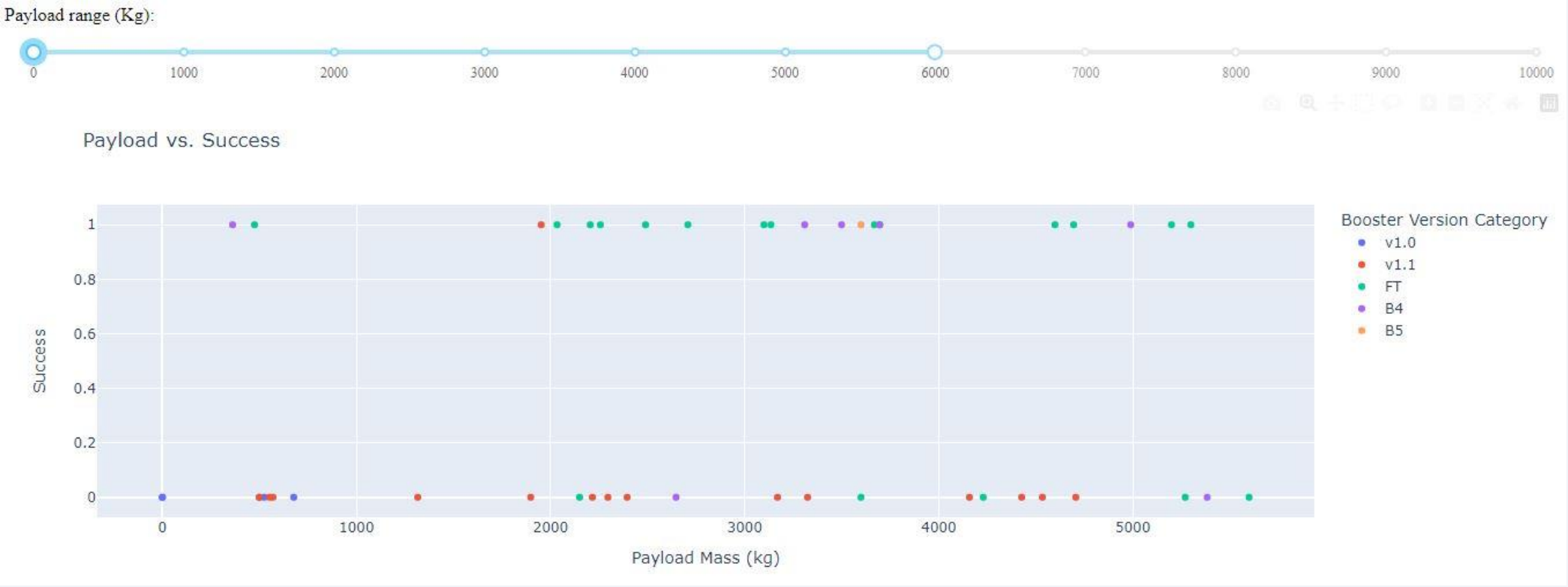
Payload vs Launch outcome

Whole payload range:



Payload vs Launch outcome

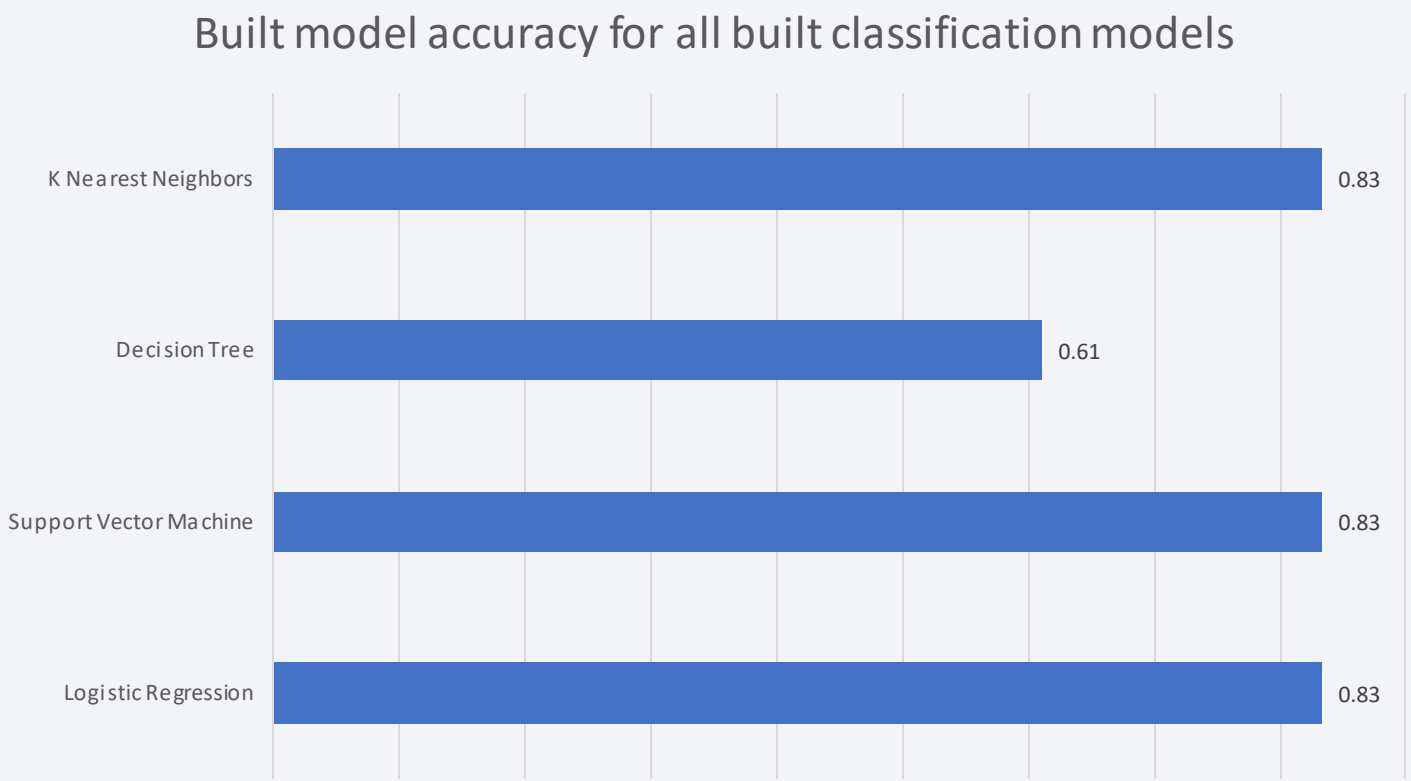
- Payload range 2000 – 6000 kg and Booster version FT show the highest success rate:



Section 5

Predictive Analysis (Classification)

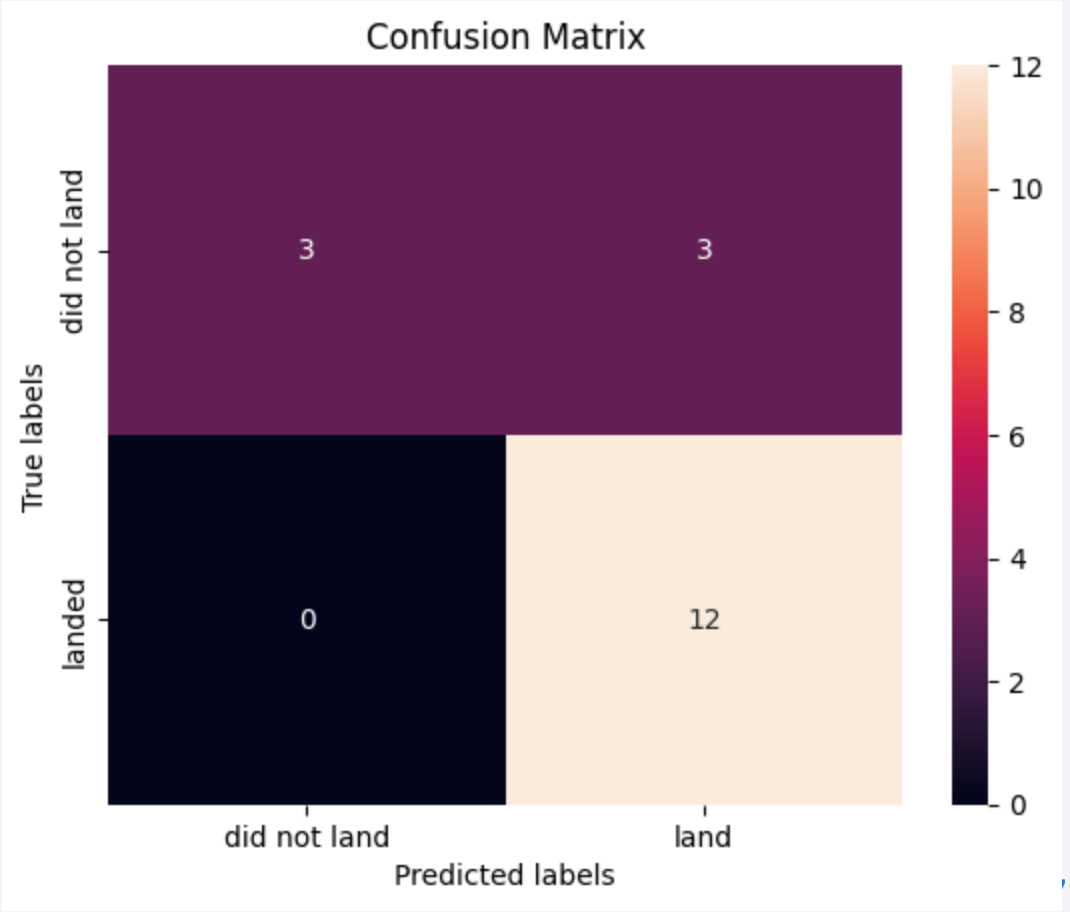
Classification Accuracy



- LR, SVM and KNN models have the highest classification accuracy equally with 0.83

Confusion Matrix of the best performing model

- True Positives (TP): 12 (Instances correctly predicted as landed)
- True Negatives (TN): 3 (Instances correctly predicted as not landed)
- False Positives (FP): 0 (No incorrect positive predictions)
- False Negatives (FN): 3 (Instances incorrectly predicted as landed)



Conclusions

- Confusion Matrix metrics all show high values confirming the model (Logistic Regression) well performing
  - Accuracy
  - Precision
  - Recall (Sensitivity)
  - F1-Score

Thank you!