IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivaylo Gardev
30.11.2025

Outline

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

Executive Summary

- The research was done using standard data science methodology
  - data collection (from SpaceX API and websites)
  - data preparation,
  - model building and testing
- Summary of all results
  - We could predict the success of the first stage landing with 83% acuracy

Introduction

- Project background and context
  - SpaceX first stage landing and reuse saves a lot of money
  - First stage landing is not always successful
  - Predicting the first stage landing success is important for saving money and future improvements
- Problems
  - What are the reasons for the first stage landing success/failure
  - Can we predict if the first stage landing will be a success

Section 1

Methodology

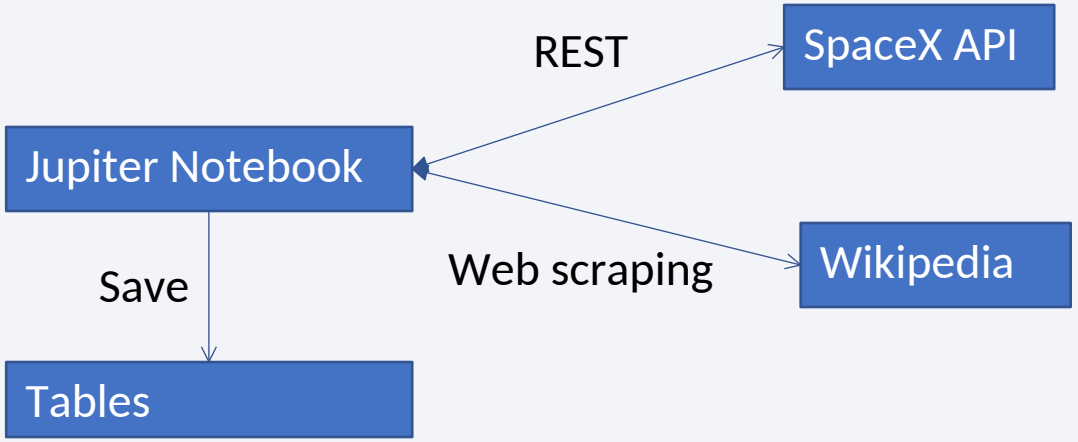
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API and web scraping
- Perform data wrangling
  - The data was transformed into tables, which have all the details for the launch and its success and are good 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
  - Build, tune and evaluate classification models

Data Collection

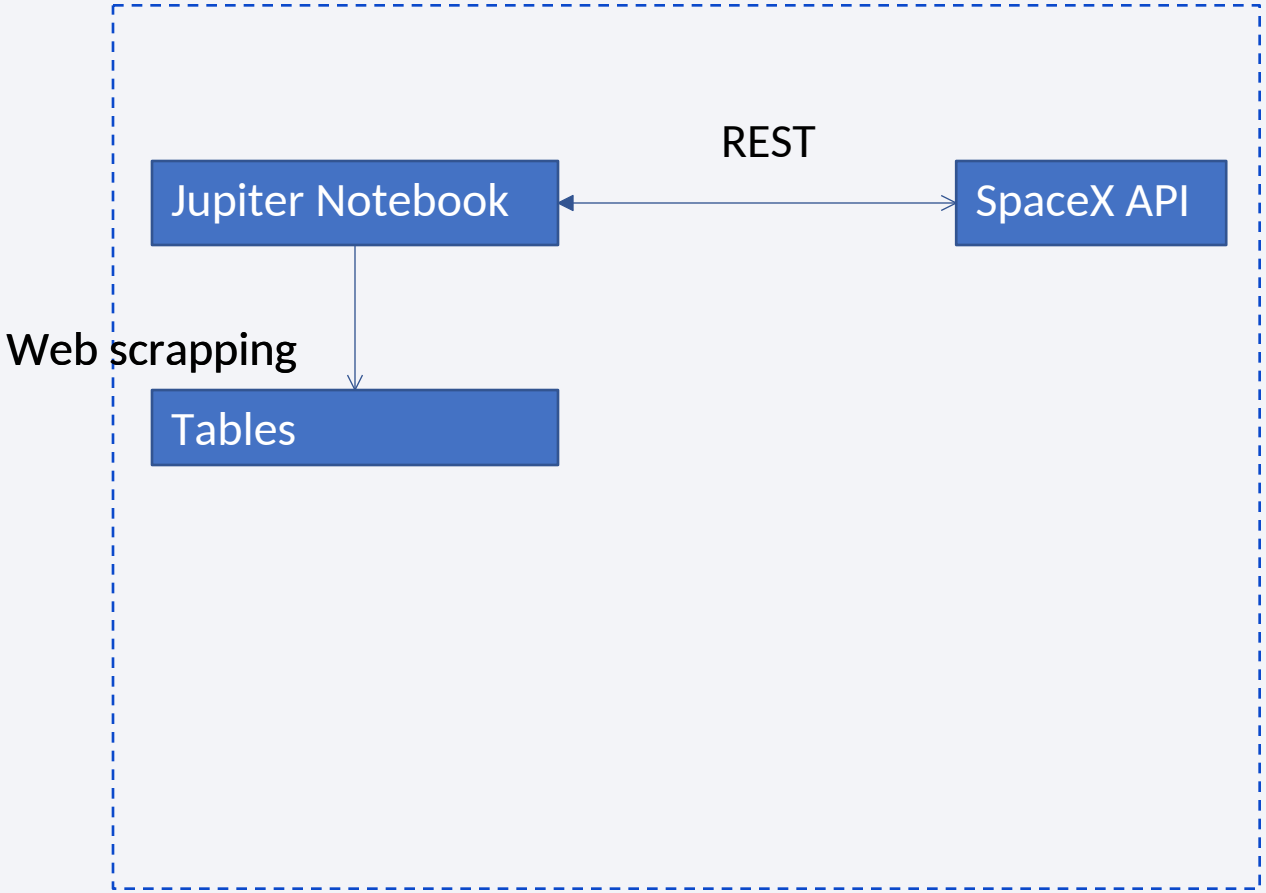
- SpaceX API
  - Most of the data was collected from SpaceX API: https://api.spacexdata.com/v4
- Web scraping
  - Additional data was collected from web (Wikipedia)
- The process:



Data Collection – SpaceX API

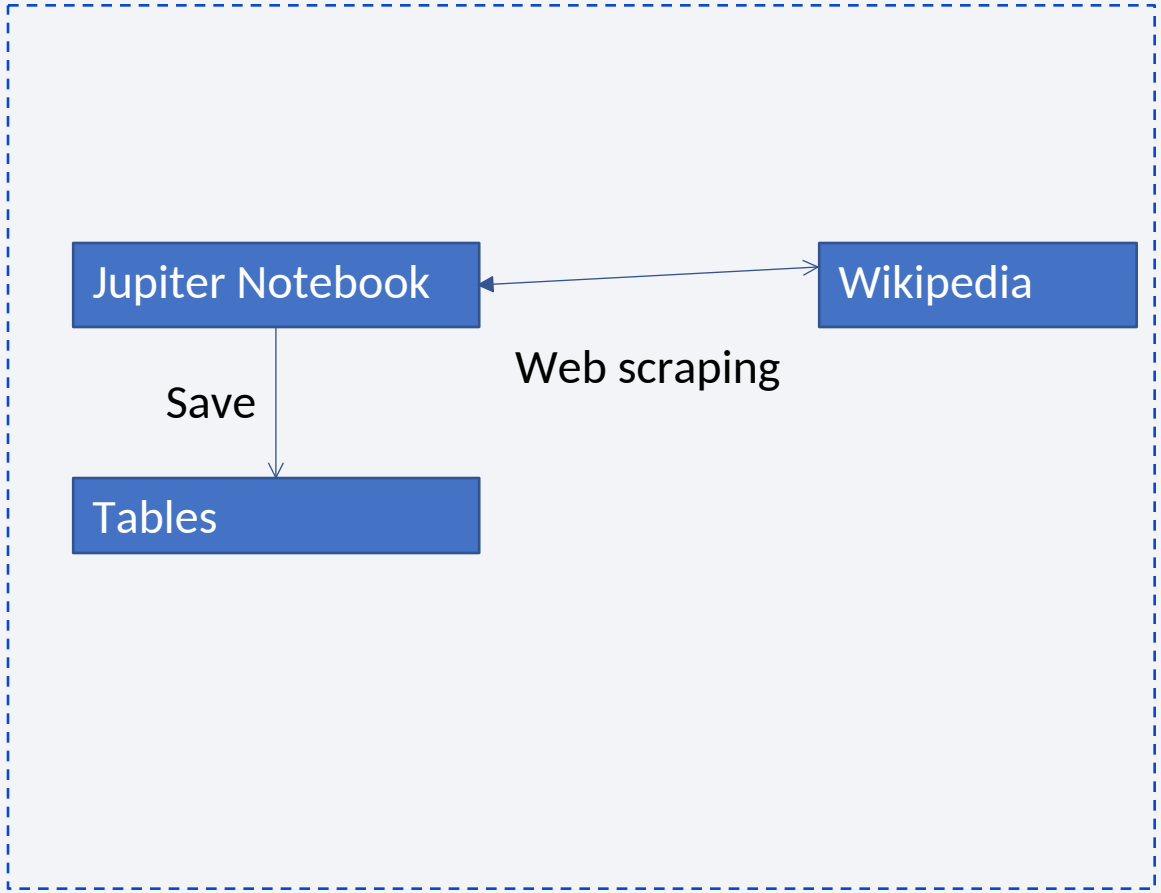
- REST calls are sent from Jupiter Notebook to SpaceX API
- SpaceX API returns launch details as json
- Jupiter Notebook transforms the json data to tables and saves it as .csv files (later tables)

- More details - https://github.com/igardev/datascience/blob/main/capstone/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

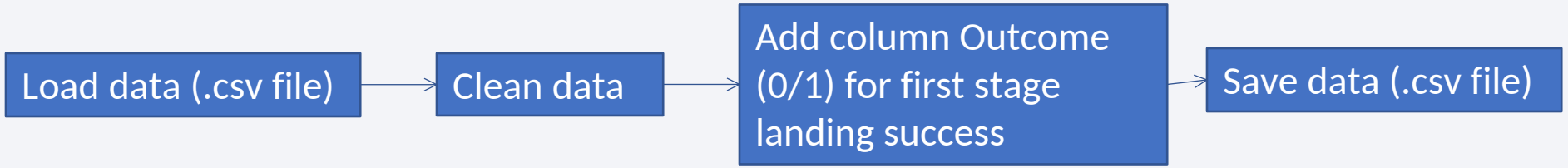
- Jupiter Notebook reads web page from Wikipedia
- The returned HTML is processed and saved to .csv files (later tables)
- More details - https://github.com/igardev/datascience/blob/main/capstone/jupyter-labs-webscraping.ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) is used to find some patterns in the data and determine what would be the label for training supervised models.
- All data fields with information about the success of the first stage landing converted into Training Labels
  - 1 means the booster successfully landed
  - 0 means it was unsuccessful.

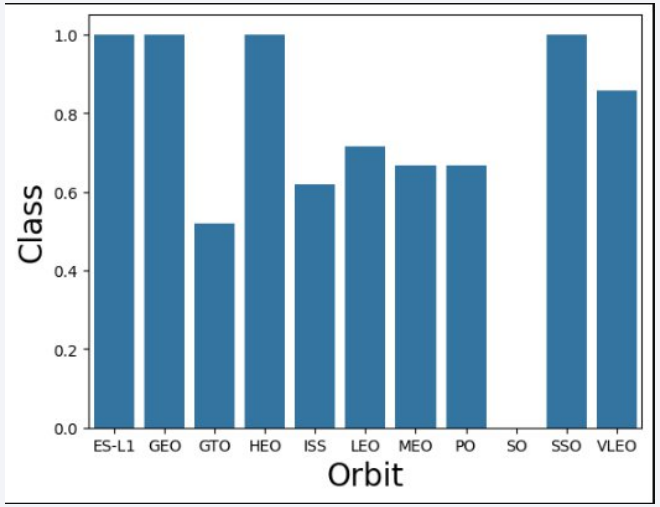
- Process:



- More details - https://github.com/igardev/datascience/blob/main/capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- The cleaned and labeled data was visualized with charts



Success rate of each orbit launch

- More details - https://github.com/igardev/datascience/blob/main/capstone/edadataviz.ipynb

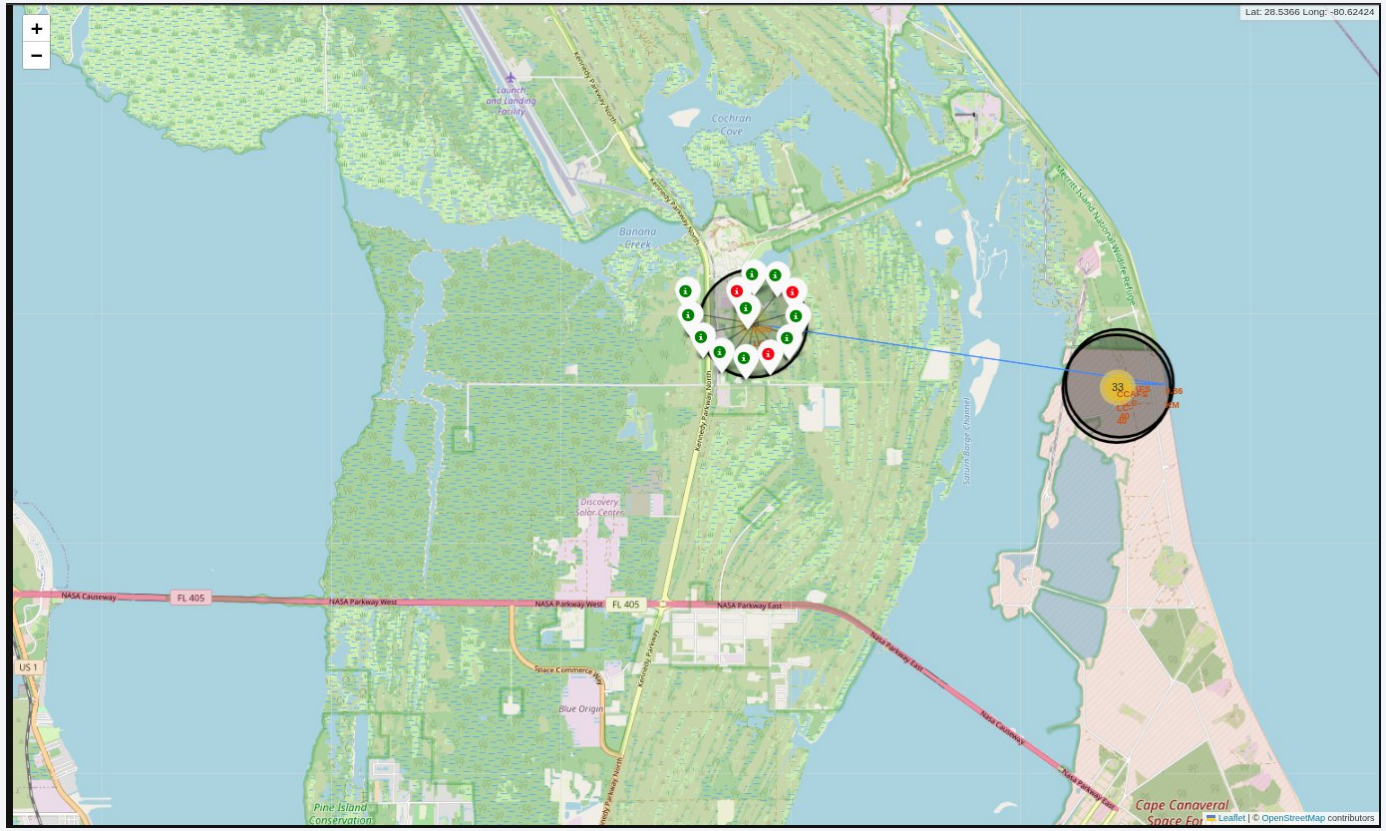
EDA with SQL

- SPACEXTABLE was created and SQL queries were performed, e.g.:
  - Names of the unique launch sites: *select distinct launch_site from spacextable*
  - Unique mission outcomes: *select distinct mission_outcome from spacextable*
  - Average payload mass carried by booster version F9 v1.1: *select avg(PAYLOAD_MASS__KG_) from spacextable where Booster_Version = 'F9 v1.1'*
  - Total number of successful and failure mission outcomes: *select mission_outcome, count(*) from spacextable group by mission_outcome*
  - Month names, failure landing outcomes in drone ship for the months in year 2015: *select substr(Date, 6,2), landing_outcome, booster_version, launch_site from spacextable where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'*
- More details

Interactive Map

- Visualizations on map were also prepared
  - Markers, cluster markers, circles, lines and distances were added (on folium map)
- Adding objects map visualizations helps understand the context better

- More details

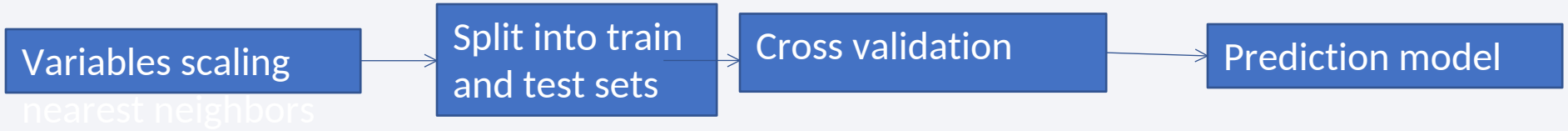


Dashboard with Plotly Dash

- On a Dashboard were added following interactive charts
  - Dropdown for selecting a site (or all sites), which affects the charts
  - Pie chart of the successful landings
  - Payload slider
  - Scatter plot for success based on payloadhttps://github.com/igardev/datascience/blob/main/capstone/spacex--dash-app.py
- Interactive visual elements are useful for exploring the relationships among the success and other variables like site and payload
- More details

Predictive Analysis (Classification)

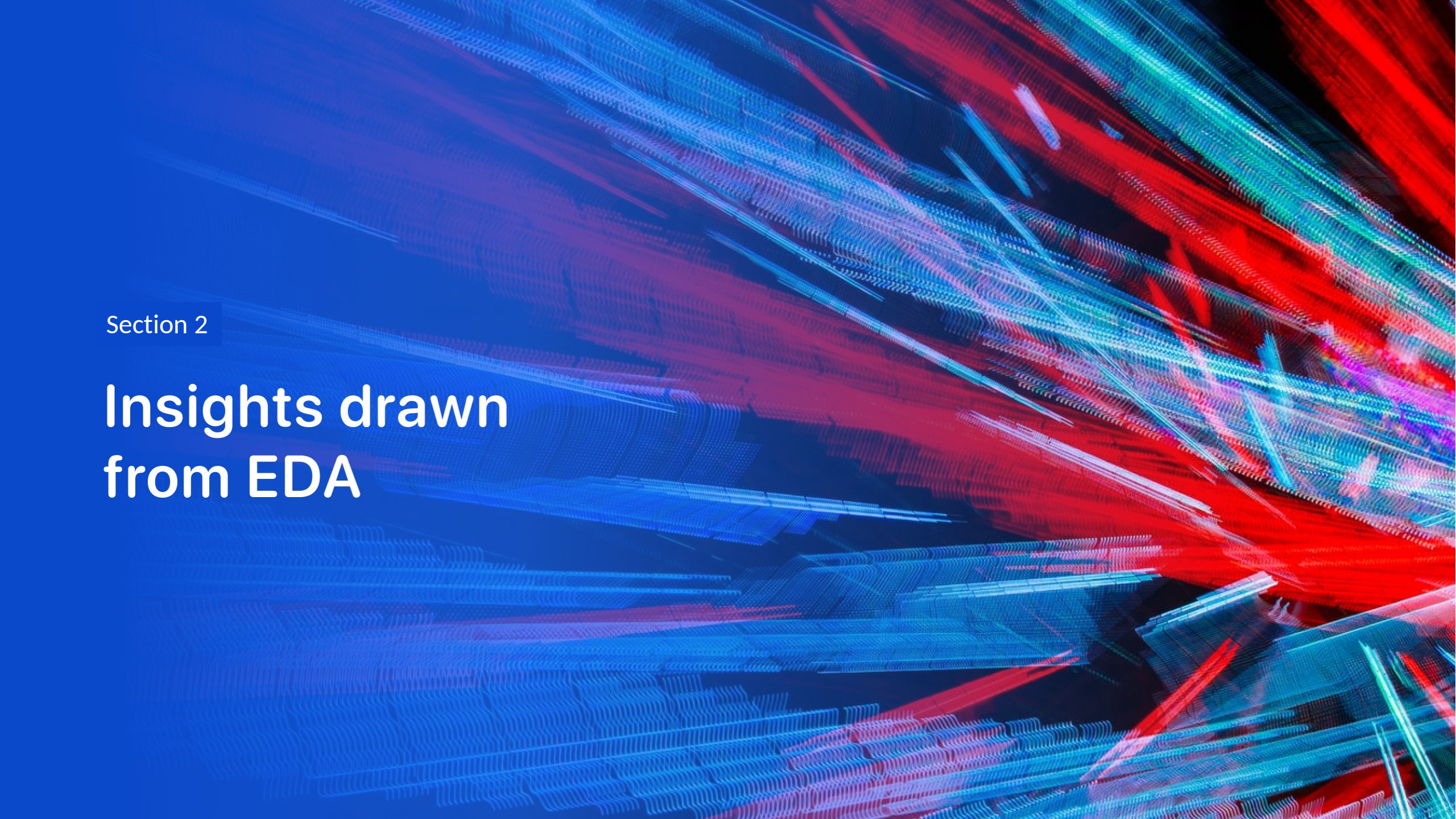
- The variables were scaled, the available labeled data was split into train and test sets (80%/20%) and cross validation was used to train success prediction models
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K nearest neighbors
- Process



- More details

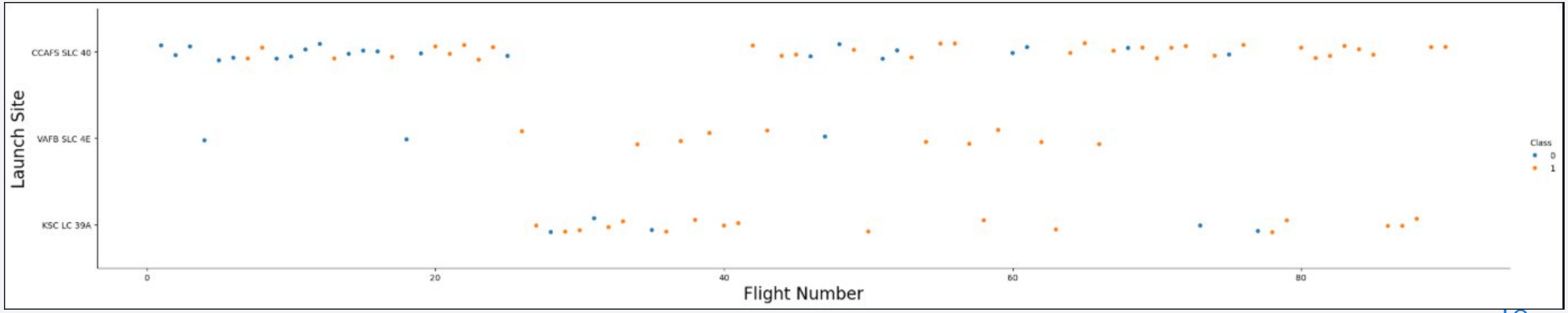
Results

- Exploratory data analysis results
- Interactive analytics
- Predictive analysis results


Section 2
Insights drawn from EDA

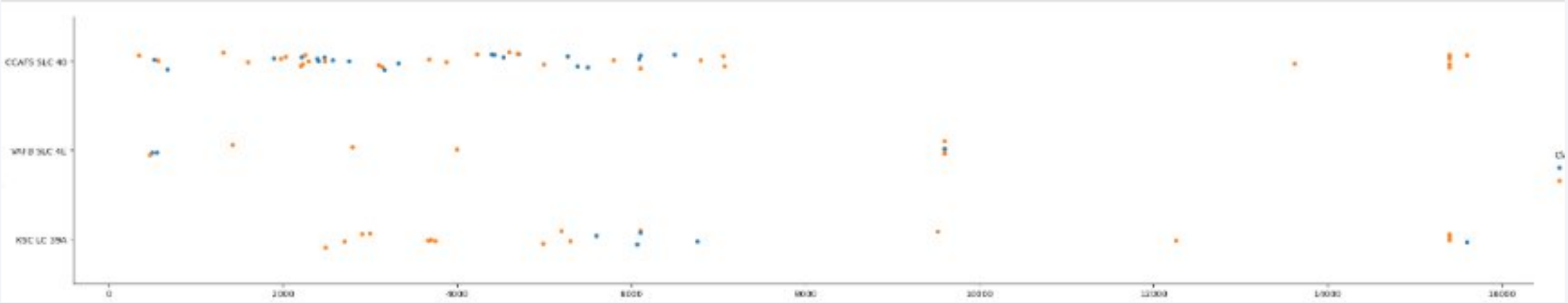
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- The scatter plot shows that
  - Fist 25 flights were mainly from one site
  - The remaining flights were distributed more eveny across the sites



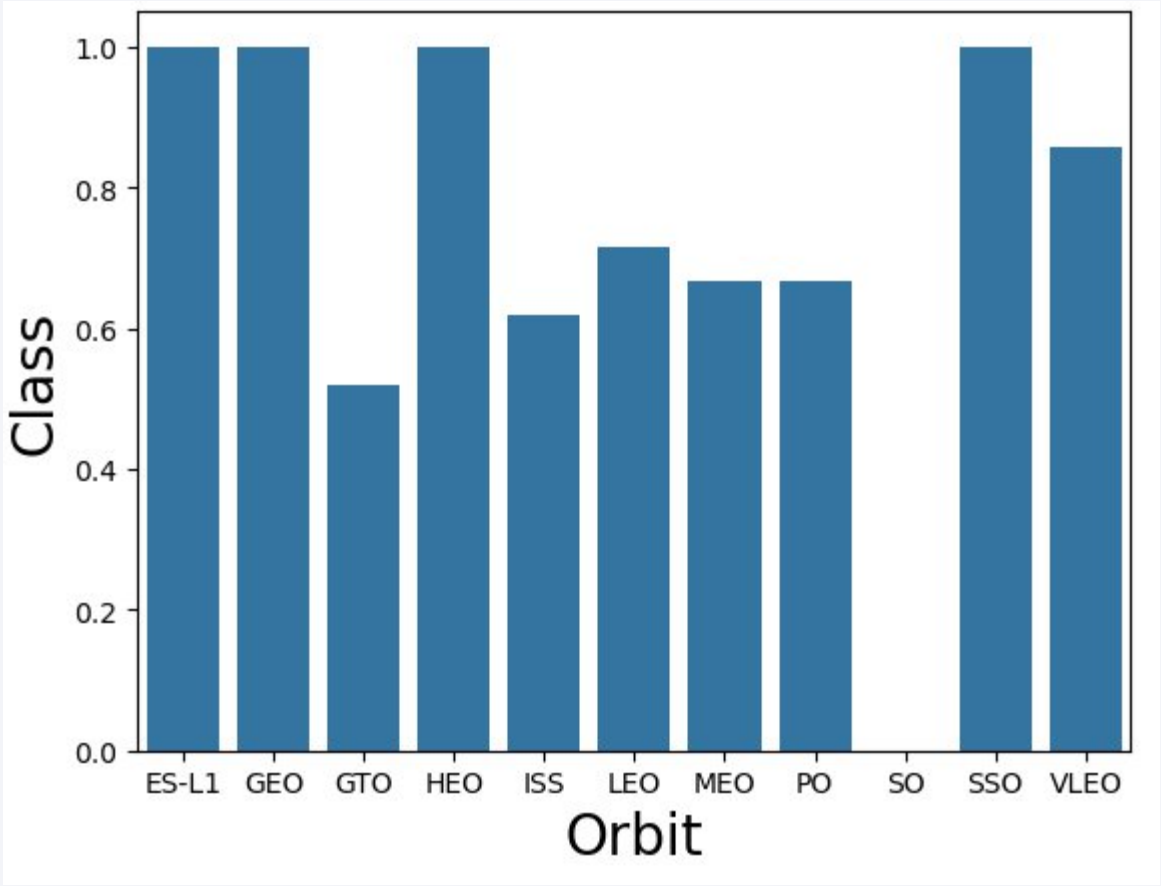
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site
- For site VAFB-SLC there are no rockets launched for heavy payload mass(greater than 10000).



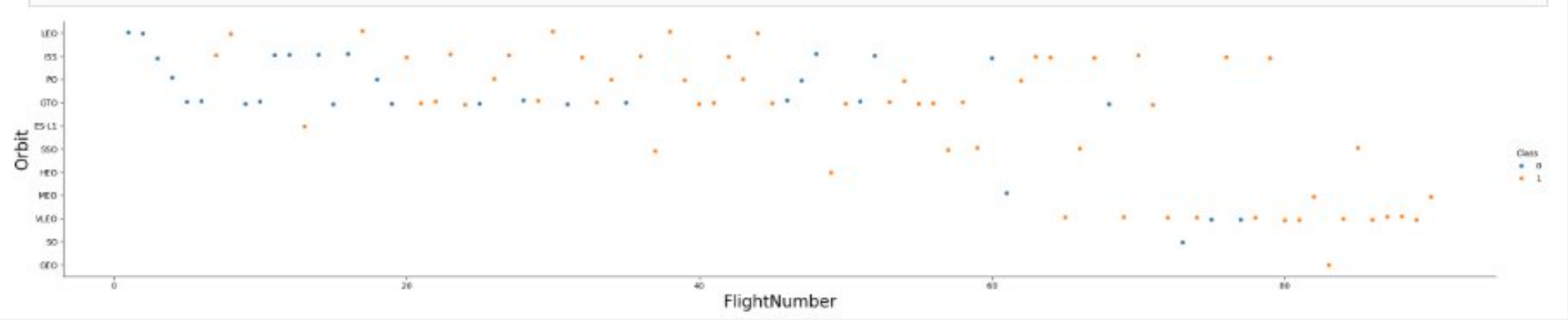
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- The highest success rate have orbits ES-L1, GEO and SSO



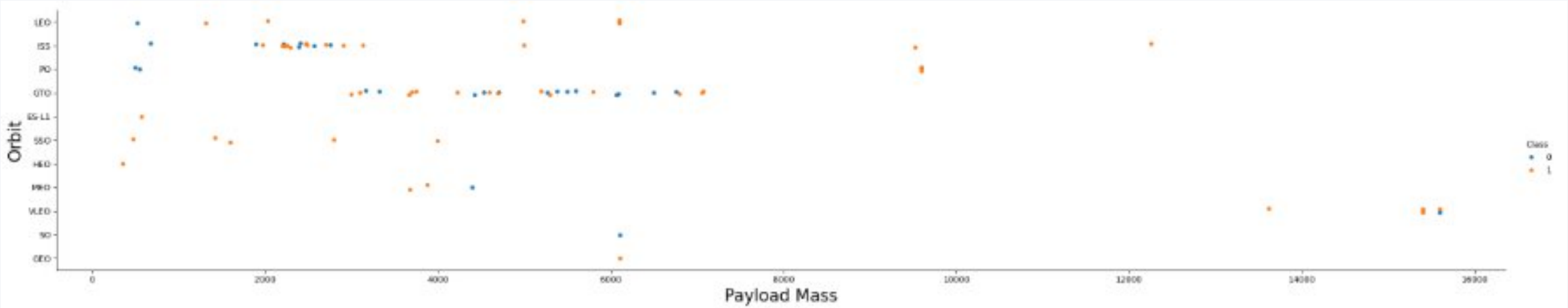
Flight Number vs. Orbit Type

- Scatter plot of Flight number vs. Orbit type
- LEO orbit - success seems to be related to the number of flights.
- GTO orbit - there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

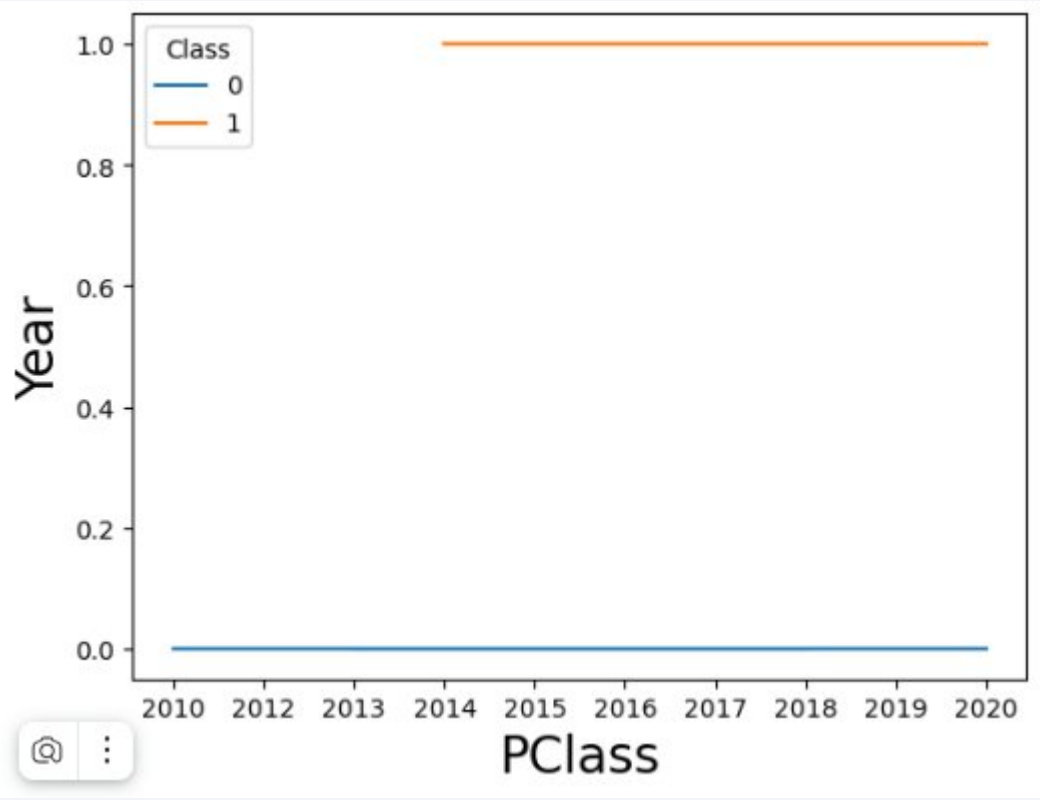
- Scatter point of payload vs. orbit type
- With heavy payloads the success rate is better for Polar, LEO and ISS
- For GTO both outcomes are present



Launch Success Yearly Trend

- Yearly average success rate

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

- Unique launch sites

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

- Unique launch sites were extracted with
  - *select distinct launch_site from spacextable*

Launch Site Names With Prefix 'CCA'

- 5 records where launch sites begin with `CCA`
- The records were extracted with: *select * from spacextable where launch_site like 'CCA%' limit 5*

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcon
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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 (parachut
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

- Total payload carried by boosters from NASA
  - 45596 kg

- This was calculated with the following sql:
  select sum(PAYLOAD_MASS__KG_) from spacextable where customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
  - 2928.4 kg

- The average payload was extracted with the following sql query:
  - *select avg(PAYLOAD_MASS__KG_) from spacextable where Booster_Version = 'F9 v1.1'*

First Successful Ground Landing Date

- First successful landing outcome on ground pad was on
  - 2015-12-22

- First successful landing outcome was extracted with the sql query:
  - *select min(Date) from spacextable where Landing_Outcome = 'Success (ground pad)'*

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- Extracted with sql query:
  - *select Booster_Version from spacextable where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 and Landing_Outcome = 'Success (drone ship)'*

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

- Extracted with sql query:
  - *select mission_outcome, count(*) from spacextable group by mission_outcome*

Boosters Carried Maximum Payload

- Names of the booster 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

- Extracted with sql query
  - *select distinct booster_version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable)*

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Extracted with sql query:
  - *select substr(Date, 6,2), landing_outcome, booster_version, launch_site from spacextable where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'*

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

- Rank of 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	rank
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

- Extracted with sql query:
  - *select landing_outcome, count(*) rank from spacextable where date >= '2010-06-04' and date <= '2017-03-20' group by landing_outcome order by rank desc*

Section 3

Launch Sites Proximities Analysis

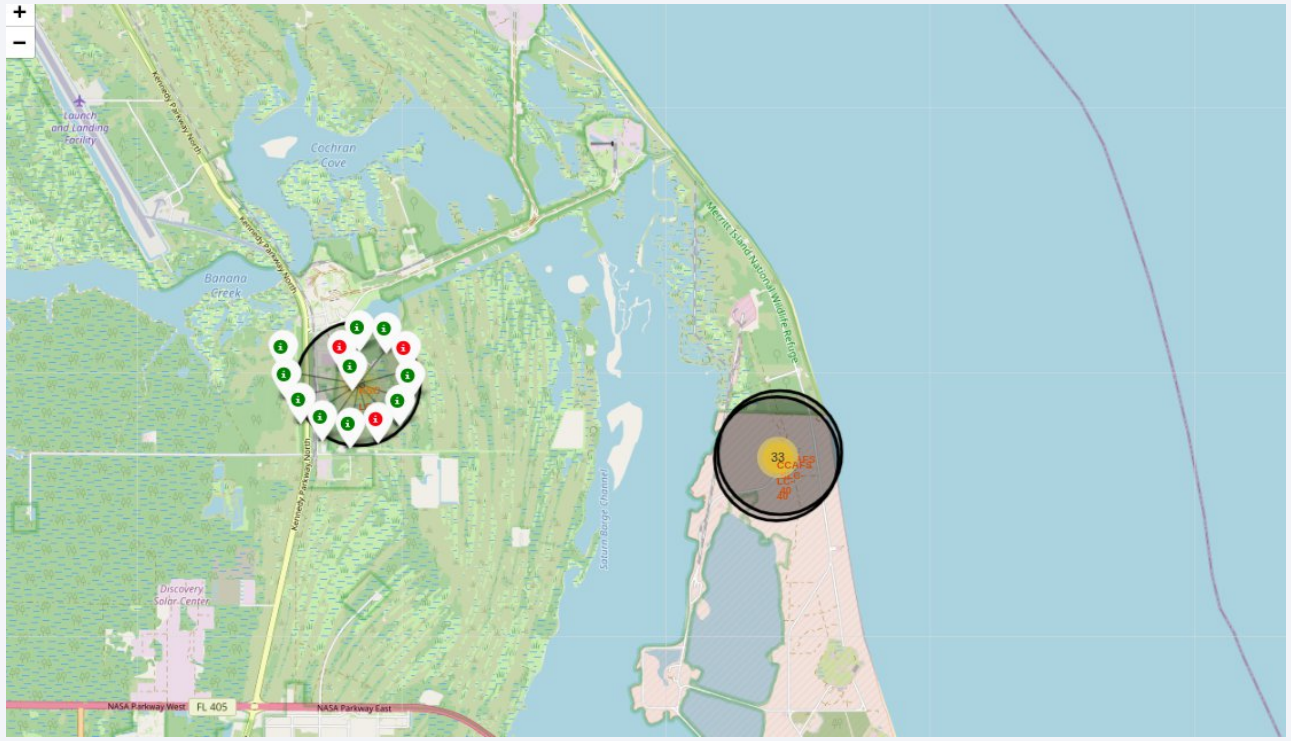
All launch sites’ locations

- All launch sites’ location markers on a global map
- Launch sites are on the east coast and west coast. No locations inside the country

Launch outcomes

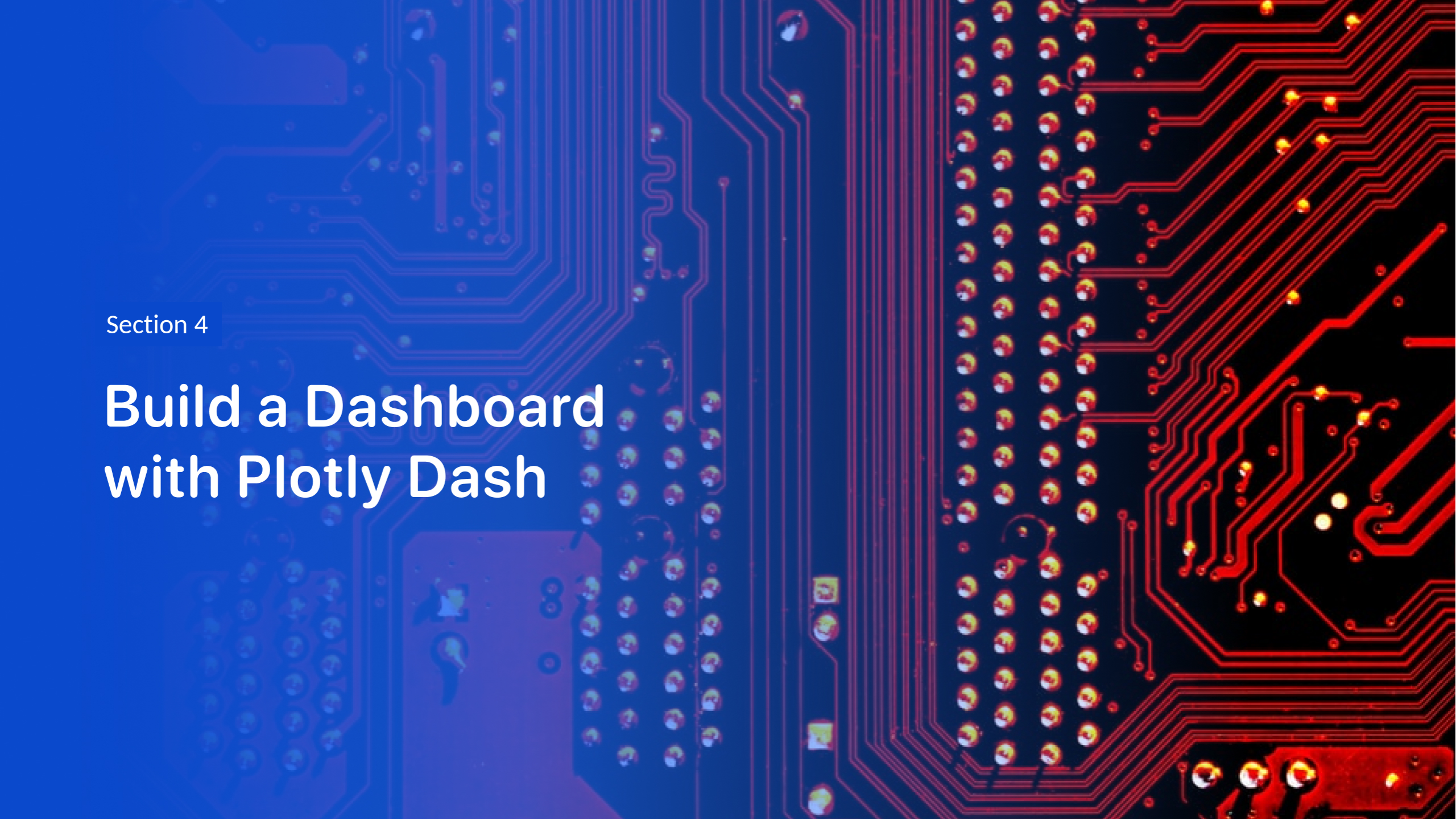
- Color-labeled launch outcomes on the map

- KSC LC 39A site has highest success rate



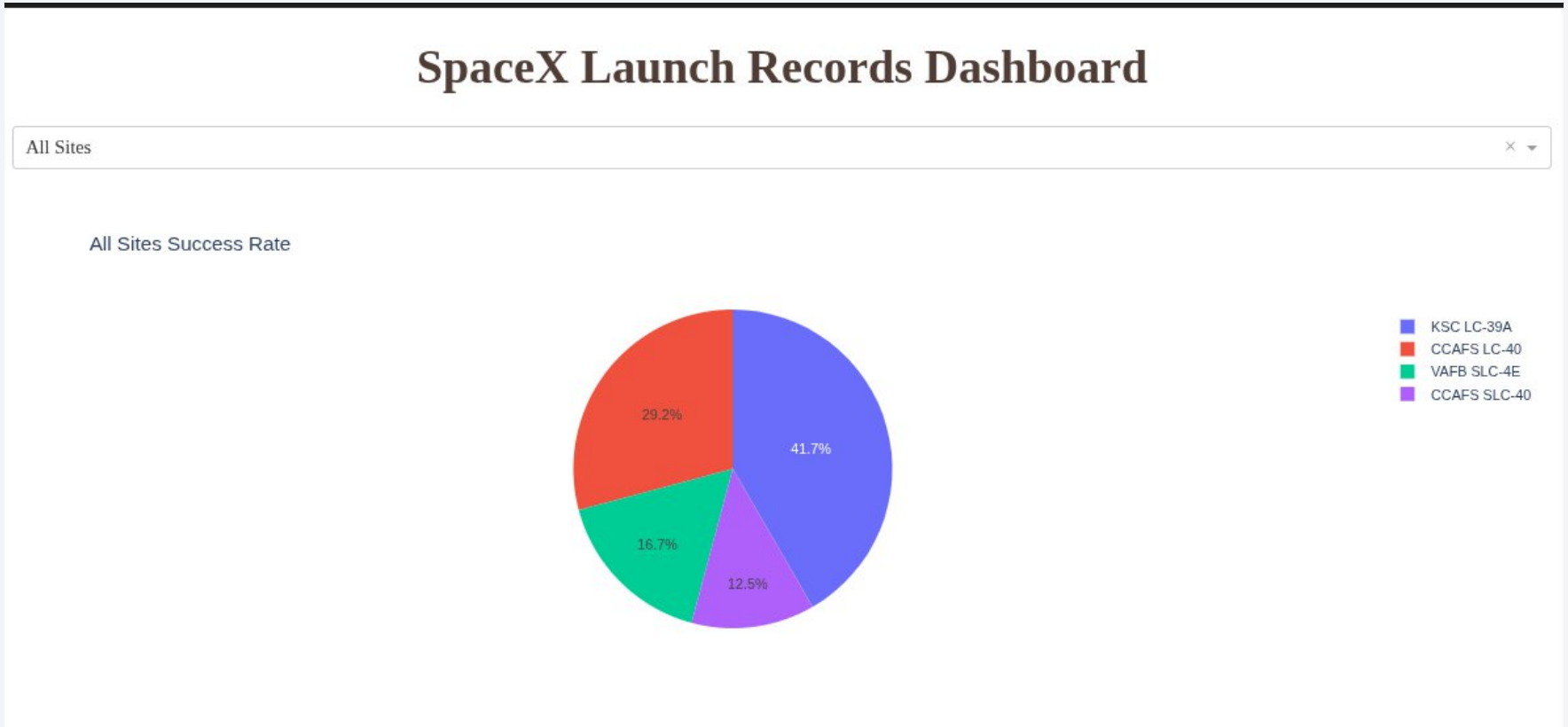
Sites proximities to railway, highway, coastline

- Map with launch site proximities to railway, highway, coastline, with distance calculated and displayed
- Launch sites are close to coastline, railway and highway


Section 4
Build a Dashboard with Plotly Dash

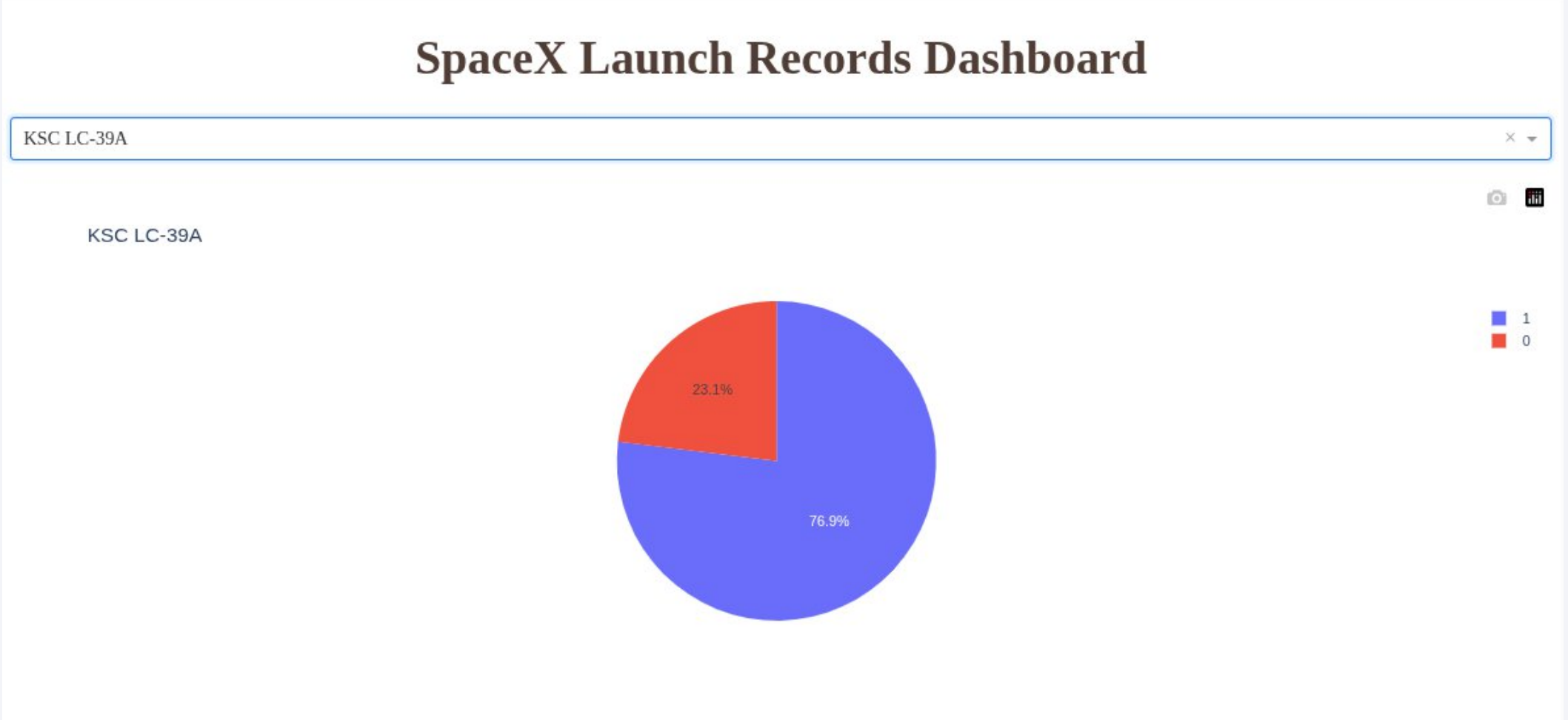
Success rate for all sites

- Launch success count for all sites, in a pie chart
- Site KSC LC 39A has highest number of successful first stage landings (41.7%)



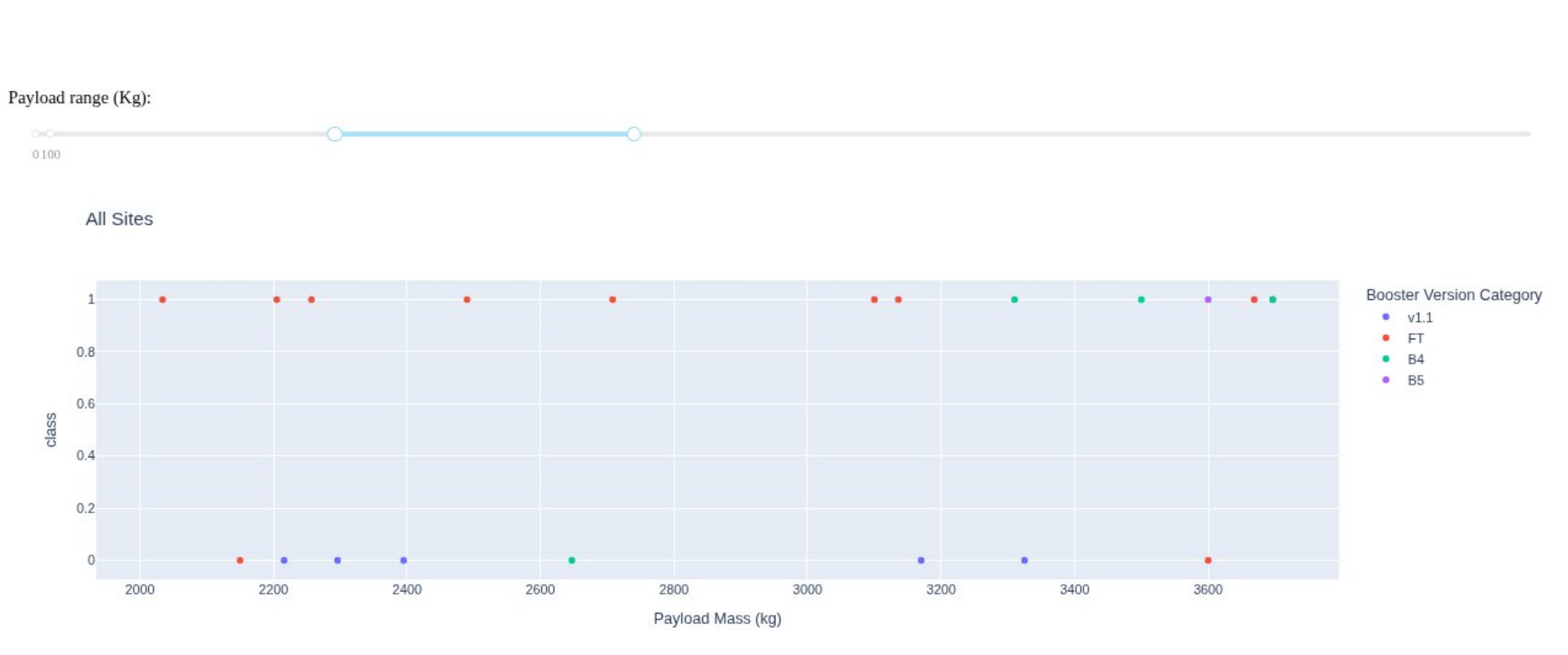
Site with highest launch success ratio

- Pie chart for the launch site with highest launch success ratio - KSC LC 39A
- Site KSC LC 39A success rate is 76.9%



Payload vs. Launch Outcome

- Payload vs. Launch Outcome scatter plot for all sites, with different payloads
- Payload Range 2000 - 4000 is the range with highest success rate



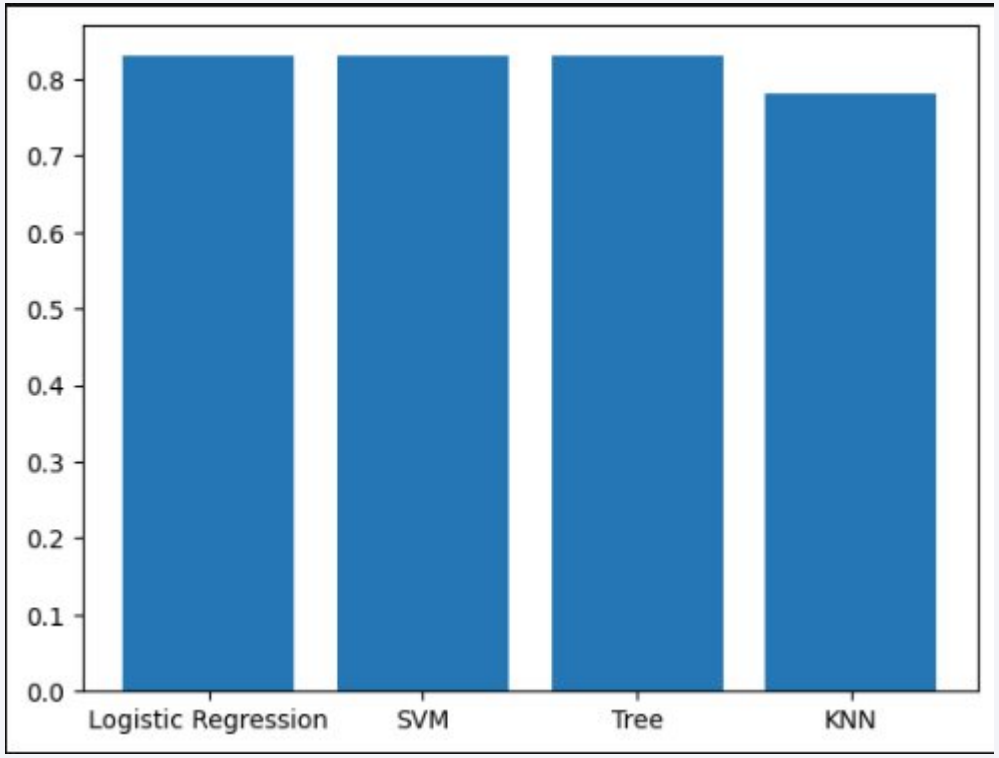
Section 5

Predictive Analysis (Classification)

Classification Accuracy

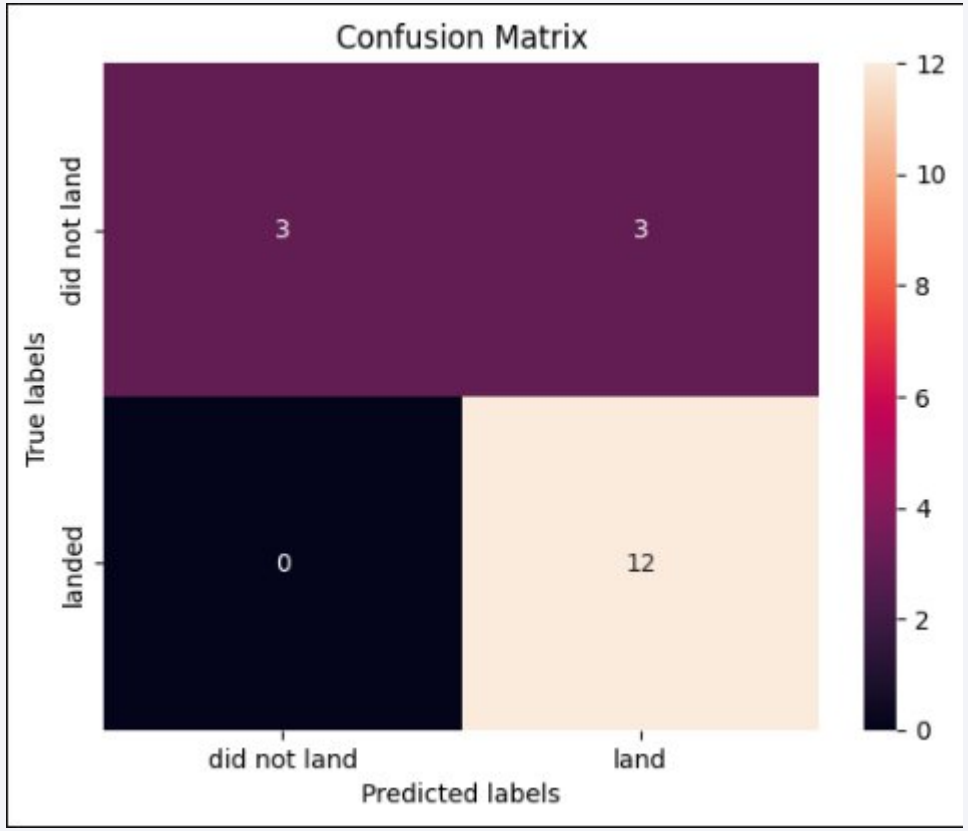
- Classification modes accuracy
  - Logistic Regression
  - SVM
  - Tree
  - KNN

- Logistic Regression and SVM have the highest prediction accuracy (0.83). Other models are very close.



Confusion Matrix

- Confusion matrix of Logistic Regression model



Conclusions

- First successful landing outcome on ground pad was on - 2015-12-22
- Payload Range 2000 - 4000 is the range with highest success rate
- Site KSC LC 39A is the site with the highest success rate - 76.9%
- Logistic regression and SVM models both perform equally well (83%) on predicting the success rate of the first stage landing

Appendix

Below are Jupiter Notebooks (with sql, queries, charts, maps, etc.), python flies and .csv files, which were used during the project:

- SpaceX_Machine Learning Prediction_Part_5.ipynb
- edadataviz.ipynb
- jupyter-labs-eda-sql-coursera_sqllite.ipynb
- jupyter-labs-spacex-data-collection-api.ipynb
- jupyter-labs-webscraping.ipynb
- lab_jupyter_launch_site_location.ipynb
- labs-jupyter-spacex-Data wrangling.ipynb
- spacex--dash-app.py
- spacex_launch_dash.csv

Thank you!