IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tom Muñoz
2025 September 18

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection (API, Web Scraping)
  - Data wrangling
  - Exploratory Data Analysis (SQL, Data Visualization)
  - Build an interactive map to present launch site distribution and successes. Folium library is used to create the map views.
  - Build a dashboard to effective, interactive analysis of results to decision making. Plotly library is used for the dashboard.
  - Crete a Machine Learning (ML) model for predictive analysis (Classification). Python and associated ML libraries are used.
- Summary of all results
  - Historical Success Rate
    - Success rate improvement was observed over time.
  - Launch Site Success Rate
    - KSC LC-39A shows the highest success rate of all launch sites.
  - Success Rate by Orbit
    - ES-L1, SSO, HEO, and GEO were observed to have the highest success rates.
  - Payload
    - Heavier payloads have a high failure rate early on with a significant improvement over time.
  - Predictive Analysis
    - The Decision Tree Classifier algorithm has proven highly accurate in predicting landing outcomes.

Introduction

- Project background and context
  - The purpose of this project is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully. Success/failure of a first stage landing impacts SpaceX's cost of a launch. With this information, we can decide if we want to submit a bid against SpaceX for a rocket launch.
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
  - Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - What is the historical success rate of the SpaceX Falcon 9 first stage landings?
  - How do launch/flight parameters (payload mass, launch site, number of flights, orbits, etc) affect the success of the first stage landing?
  - Does the landing success rate improve over time?
  - Which algorithm(s) are best suited to classify the success of the landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API endpoints were used to collect data directly from SpaceX
  - Web scaping from Wikipedia to extract tabular data was performed
- Perform data wrangling
  - Extracted relevant records, addressed missing values, flattened fields, converted categorical data to numerical values with one-hot-encoding
- Performed exploratory data analysis (EDA) using visualization and SQL
  - Explore relationships amongst parameters
- Performed interactive visual analytics using Folium and Plotly Dash
  - Mark all launch sites on a map identifying successful and failed launches
  - Calculate distances to landmarks near the launch site
  - Dashboard for interactive exploration of the data
- Perform predictive analysis using classification models
  - Build and evaluate classification models (Logistic Regression, SVM, KNN, Decision Tree)

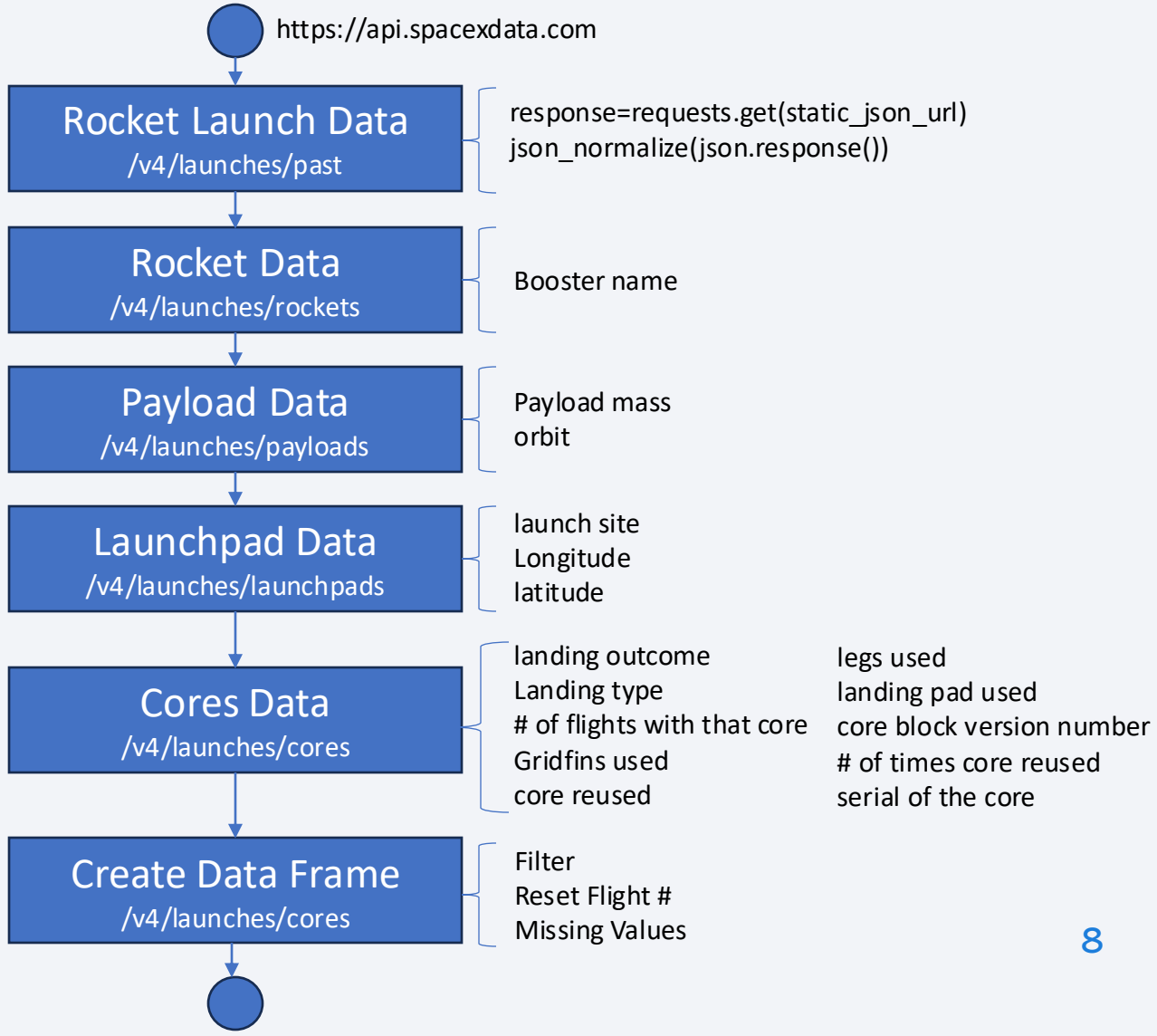
Data Collection

- Data was collected directly from SpaceX via their API endpoints and also by scraping wiki data regarding SpaceX launch records.
  - HTTP requests against various SpaceX API endpoints (https://api.spacexdata.com):
    - Specific data about the launch vehicles, launch sites, payloads, and cores was gathered from the following endpoints:
      - /v4/rockets
      - /v4/launchpads
      - /v4/payloads
      - /v4/cores
    - Launch data was obtained from /v4/launches/past
  - Historical launch records were scraped as tabular data from Wiki using BeautifulSou
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- Data from the API gathered included details related to Launch ata, Rocket data, Payload data, Launchpad data, and Cores data.
- This data was filtered to include Falcon 9 launches only and then some cleanup was performed.
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



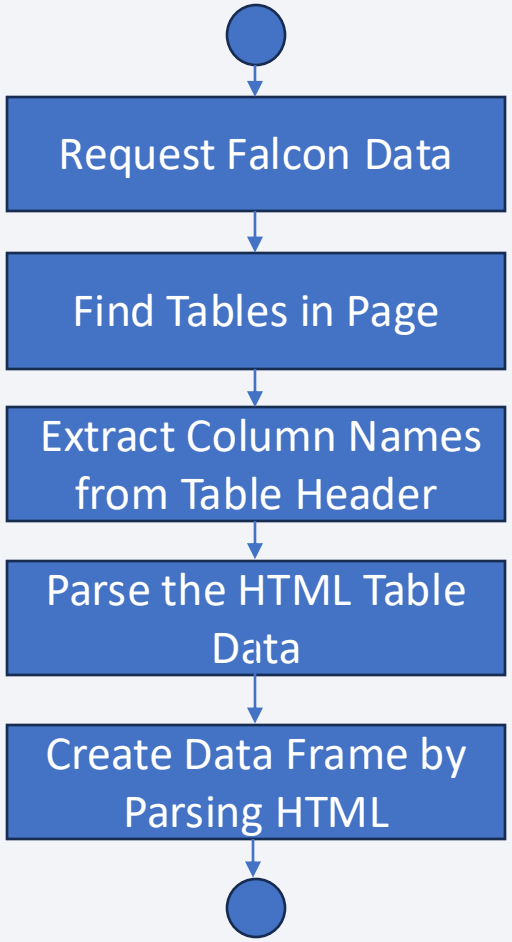
Data Collection - Scraping

- Data was scraped from the wiki site contained in tabular form.

  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- GitHub URL

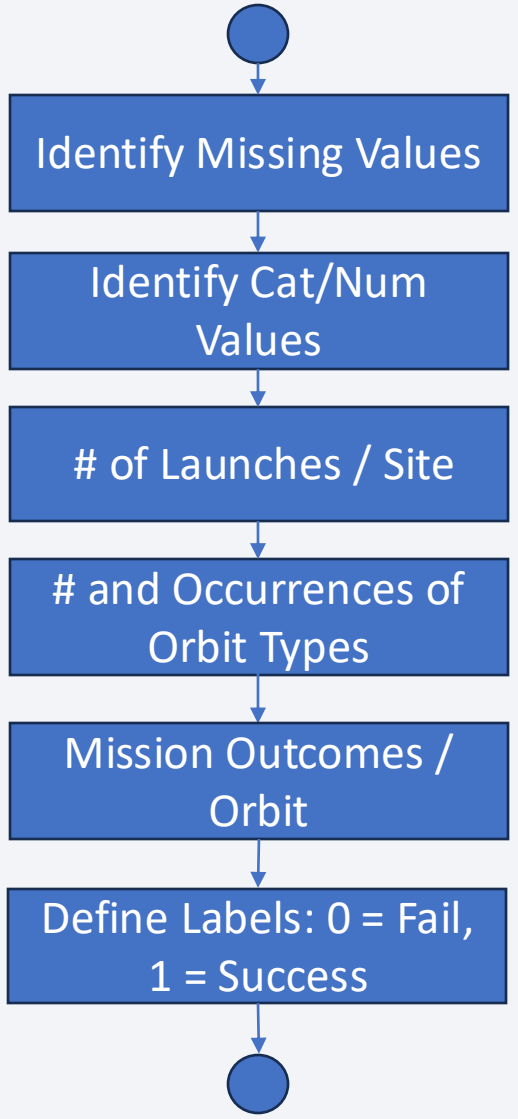
  https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Purpose of our data wrangling is to understand the data by performing Exploratory Data Analysis (EDA) to find patterns in the data and to determine the labels for training supervised models.
- EDA
  - Find missing values as a percentage of each attribute
  - Identify column types, numerical or categorical
  - Show launches per Launch Site
  - Show distribution of Orbit type
  - Explore outcomes, and group them by binary outcome (success or failure)
- Determine Labels
  - Represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Charts created to visualize data include:
  - Payload Mass vs Flight # : to see likelihood of success as flight and payload increases
  - Launch Site vs Flight #: to see if launch site has an impact on success and see launch site usage
  - Launch Site vs Payload Mass: to see which payloads are launched from which sites
  - Success Rate vs Orbit Type: to see if any orbit tend to be more successful than others
  - Orbit vs Flight #: to see if there is any trend for orbits and successful landings
  - Orbit vs Payload Mass: to see which orbits have more success for different payload masses
  - Success Rate vs Year (trend over time): Historical success rate trend
- Apply OneHotEncoder via get_dummies(features,..) function to convert categorical values and then cast to float64.
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Display the names of the unique launch sites in the space mission
  - 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 succesful 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 all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - 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

- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created a map with folium and included several objects
  - Markers – to add a label to a launch site
  - Circles – to identify launch sites locations
  - Lines – to show distances to other landmarks
  - Marker Clusters – to identify various launches at each launch site and outcome information
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

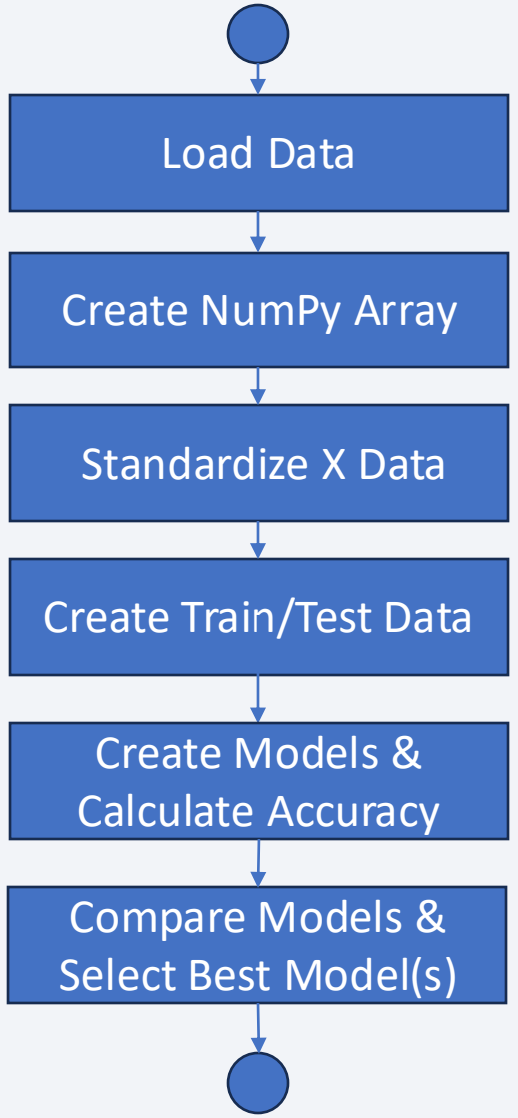
- A dashboard was created with Plotly Dash to present launch site, payload mass, and booster data for successful and failed landings.
- Specific objects/plots/graphs contained in the dashboard include:
  - Dropdown selector to choose a Launch Site or All Launch Sites
  - Pie chart
    - Breakdown of successful outcomes across all sites or selected site.
  - Scatter plot
    - Outcome by payload mass and booster version for all sites or a selected site.
  - Payload Mass range slider that to selected ranges of payload masses
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

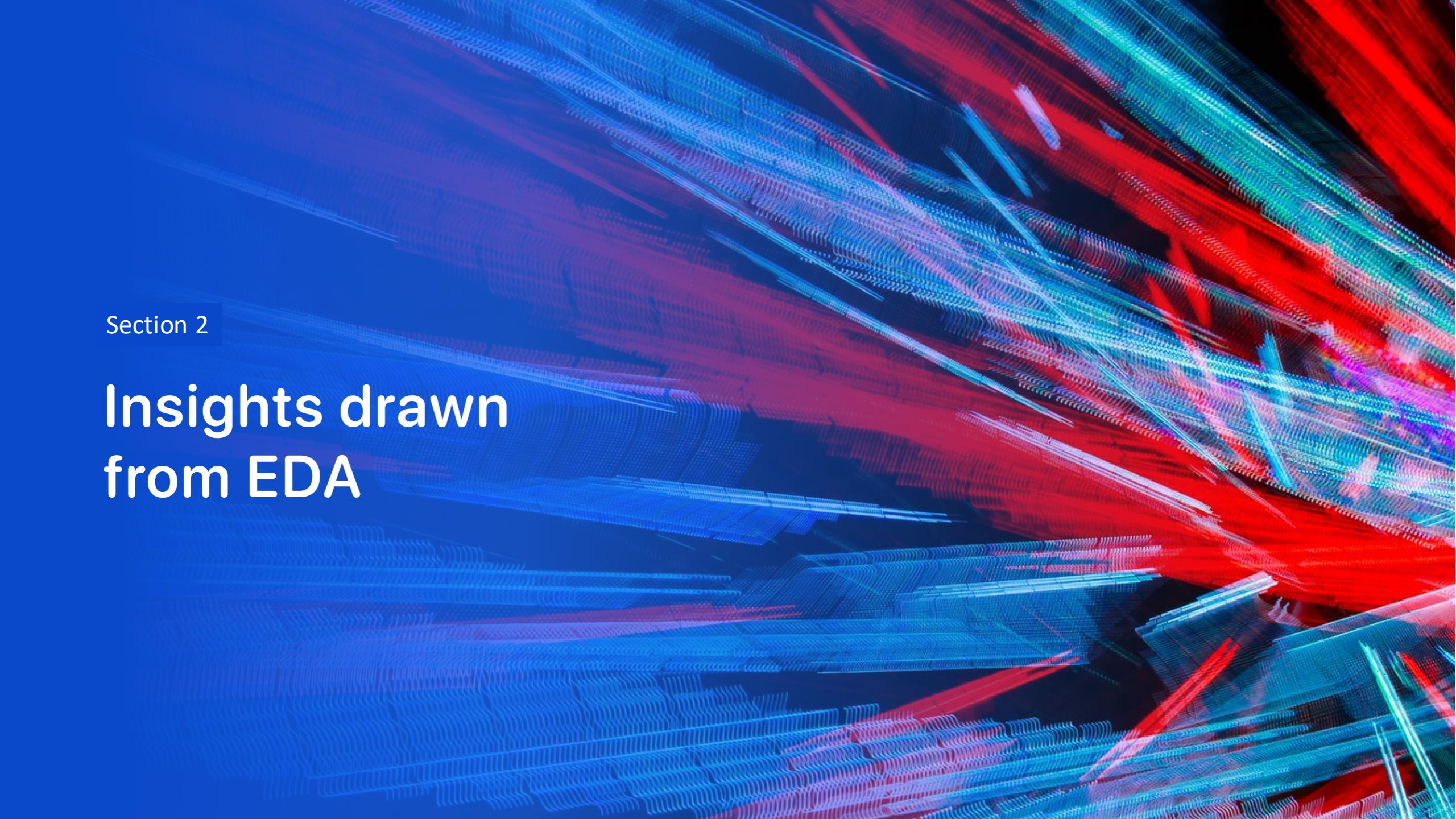
- Model development, evaluation, selection
  - Load data
  - Create a NumPy array from the column Class in data, by applying the method to_numpy() then assign it to the variable Y
  - Standardize the data in X then reassign it to the variable X
  - Create the train and test data
  - For each model, create the model objects, GridSearch objects, and fit objects to find the best parameters.
    - Logistic Regression
    - SVM
    - Decision Tree
    - K Nearest Neighbor
  - Calculate the model accuracy and created a confusion needed, as appropriate
  - Compare the model accuracies to select best model(s)
- GitHub URL

https://github.com/tomunoz/IBMDSProfessionalCertificate/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

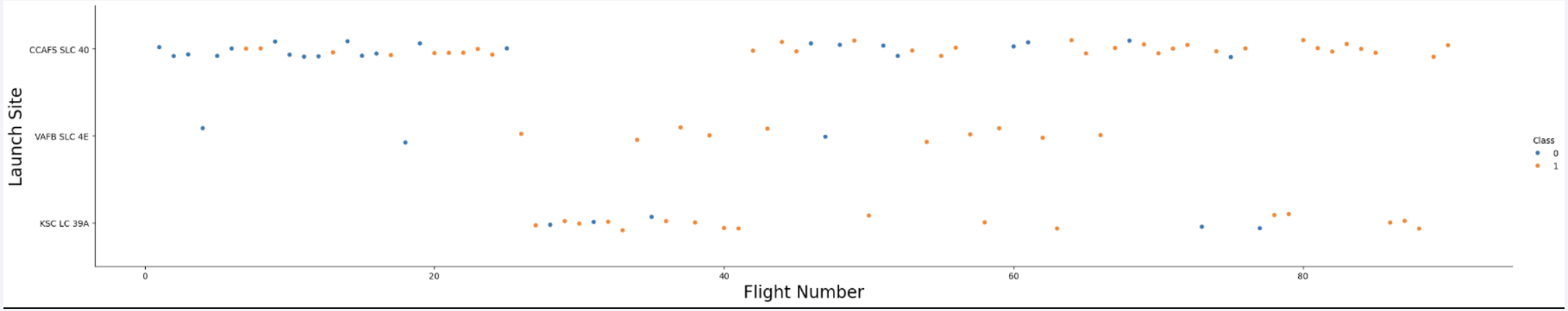


Results

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

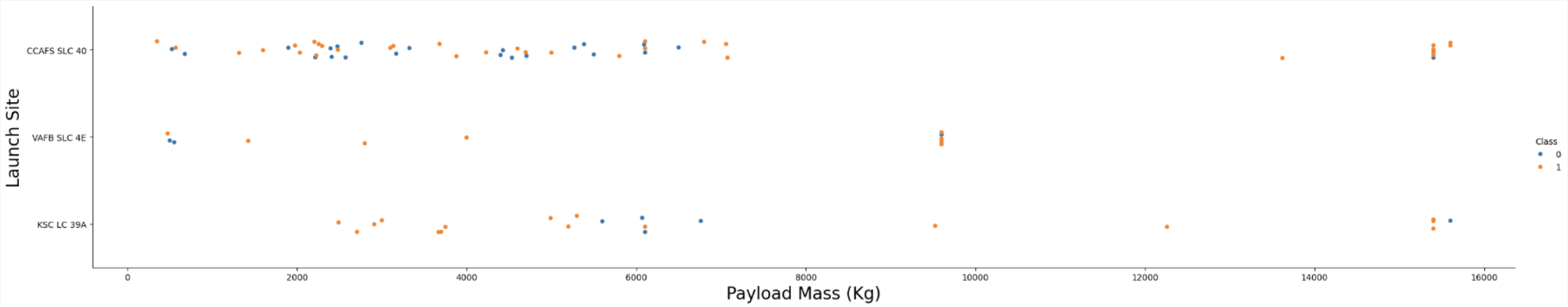

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- More flights failed earlier in the flight sequence and more successes in the later flight sequence.
- CCAFS SLC 40 launch site has the most flights.
- KSC LC 39A launch site started being used after about 2 dozen flights.
- Interesting that when flights started in KSC LC 39A, flights stopped in CCAFS SLC 40

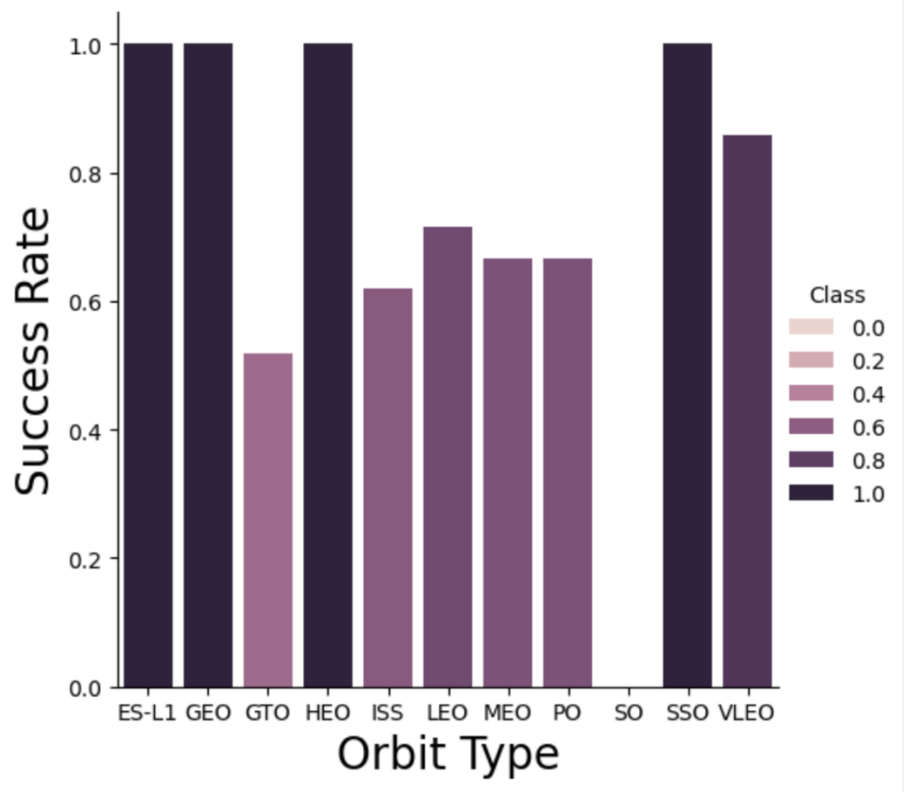
Payload vs. Launch Site



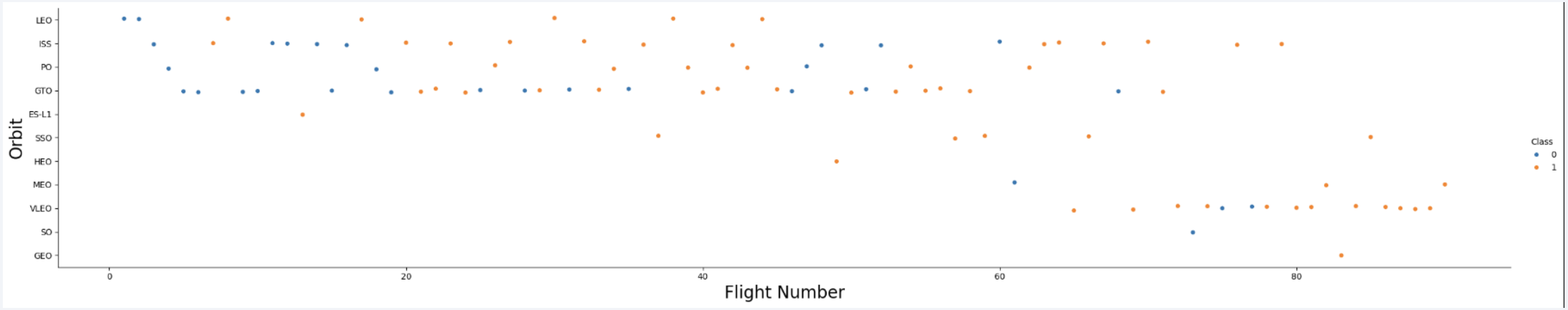
- Payload masses of more than 7,000 kg were mostly successful regardless of launch site
- CCAFS SLC 40 had the most launches while VAFB SLC 4E had the fewest
- KSC LC 39A was generally successful except for payload masses around 6,000 kg
- CCAFS SLC 40 had a spotty record of success/failures below 7,000 kg payload masses

Success Rate vs. Orbit Type

- Some orbits demonstrate high success
  - ES-L1, GEO, HEO, SSO, VLEO
- GTO orbit had the least success
- Remaining orbits success rates fared between 60% and 80%

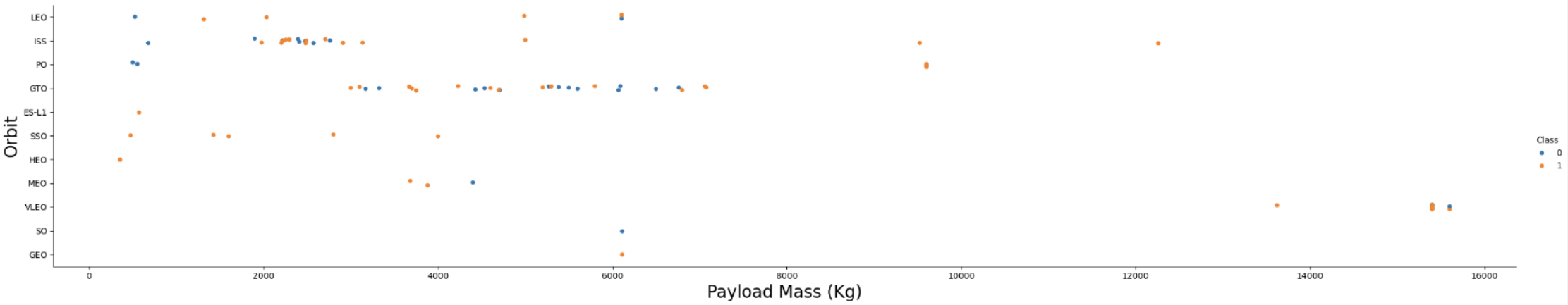


Flight Number vs. Orbit Type



- More successes as flight number increases
- Some orbits were not attempted until later flights (SSO, HEO, MEO, VLEO, SO, GEO)

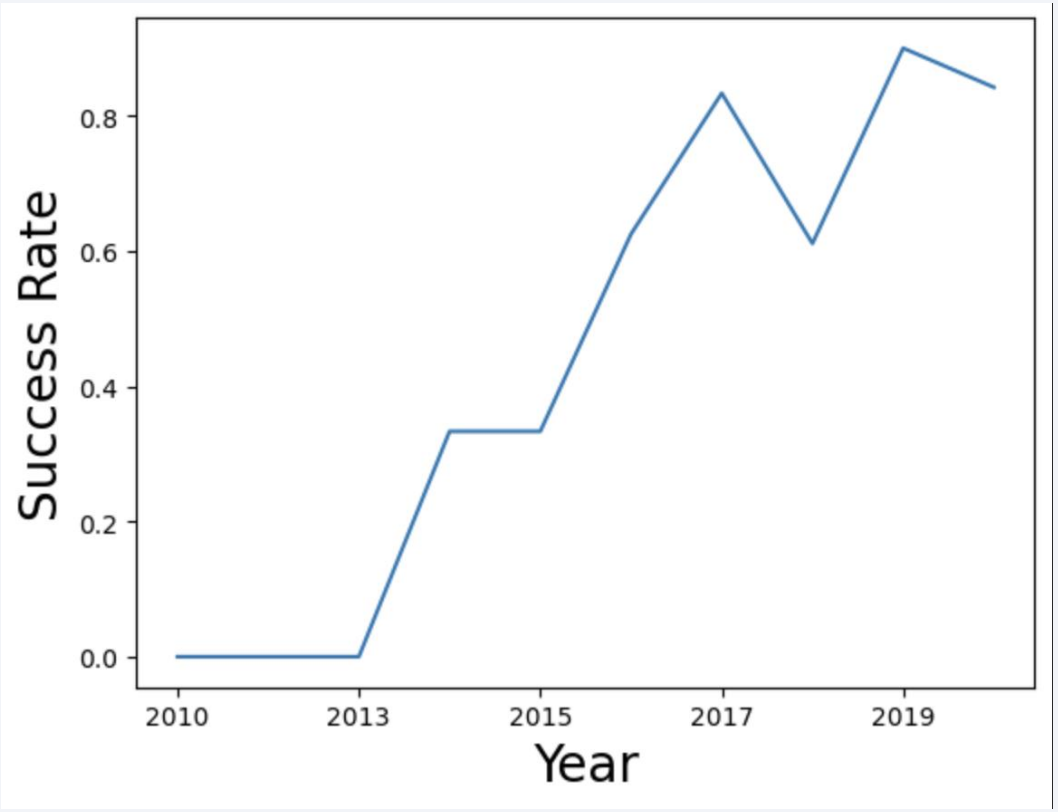
Payload vs. Orbit Type



- Fewer payload masses above 7,000 kg launched, and then only to limited orbits
- GTO and ISS orbits appear to have the most launches
- GTO and ISS orbits combined have the majority of failures

Launch Success Yearly Trend

- Success increases over time.
- 2018 had a noticeable decline in success rate (%) due to a lower overall number of launches in later flights as demonstrated in previous charts.



All Launch Site Names

- 4 unique launch sites were identified.

```
%sql select distinct launch_site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Total payload carried by boosters from NASA = 45596 kg

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2534.66 kg

```
%sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTABLE where "Booster_Version" like '%F9 v1.
 * sqlite:///my_data1.db
Done.
 average_payload_mass
 2534.6666666666665
```

First Successful Ground Landing Date

- First successful landing outcome on ground pad is December 22, 2015

```
%sql select min(date) as first_succesful_landing from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

first_succesful_landing
2015-12-22

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

```
%sql select "Booster_Version" from SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' and PAYLOAD_MASS_
```

```
 * sqlite:///my_data1.db
Done.
```

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
  - 100 successful missions with varying success labels and 1 failure

```
%sql select "Mission_Outcome", count(*) as total_number from SPACEXTABLE group by "Mission_Outcome";
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
  - **%sql** select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

```
%sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPAC
 * sqlite:///my_data1.db
Done.
```

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, their booster versions, and launch site names for in year 2015

%sql select case substr("Date",6,2) when '01' then 'January' when '02' then 'February' when '03' then 'March' when '04' then 'April' when '05' then'May' when '06' then 'June' when '07' then 'July' when '08' then 'August' when '09' then 'September' when '10' then 'October' when '11' then 'November' when '12' then 'December' else 'none' end as month, "Date", "Booster_Version", "Launch_Site", "Landing_Outcome" from SPACEXTABLE where "Landing_Outcome" = 'Failure (drone ship)' and substr(Date,0,5)='2015';

```
%sql select case substr("Date",6,2)  when '01' then 'January' when '02' then 'February' when '03' then 'March' wh
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

%sql select "Landing_Outcome", count(*) as count_outcomes from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by count_outcomes desc;

```
%sql select "Landing_Outcome", count(*) as count_outcomes from SPACEXTABLE where "Date" between '2010-06-04' and
 * sqlite:///my_data1.db
Done.
```

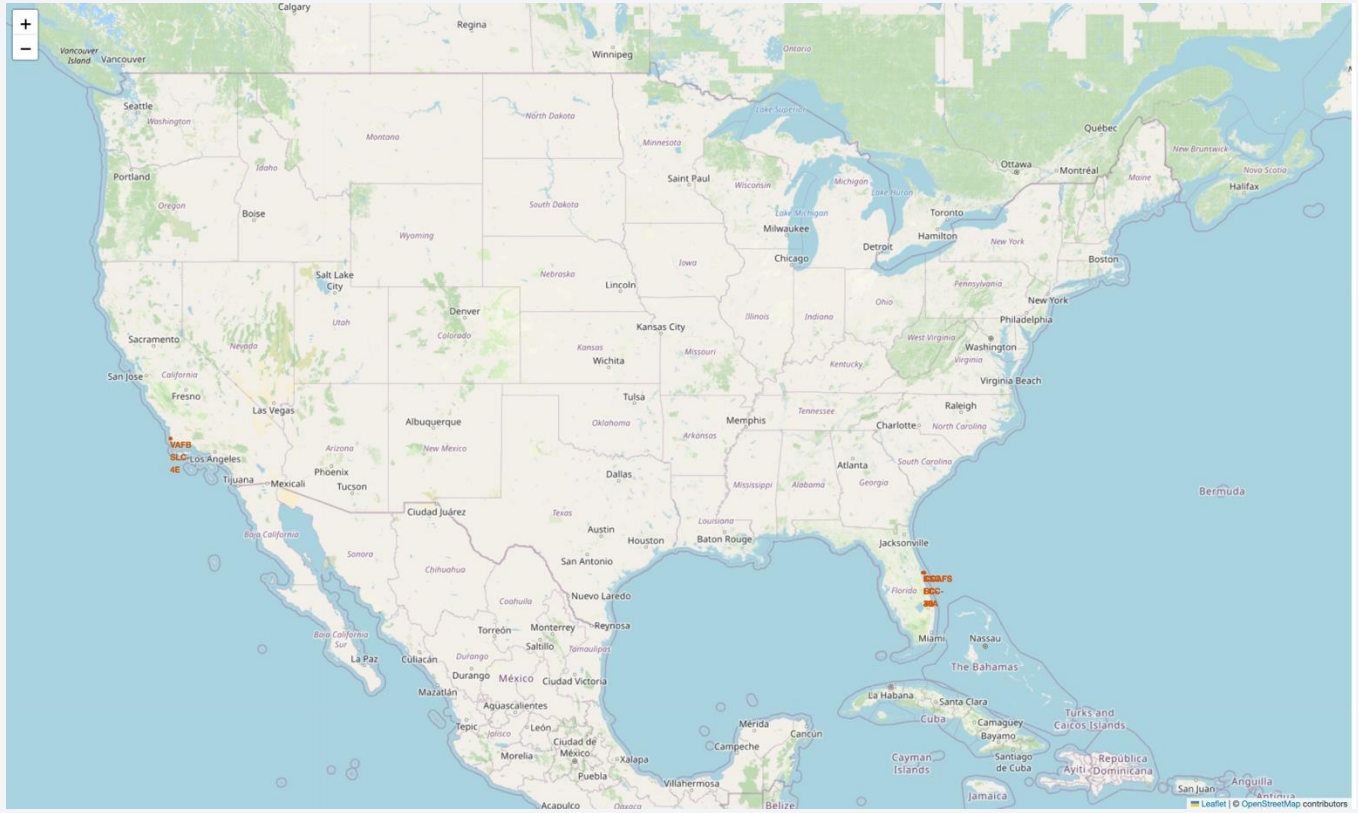
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

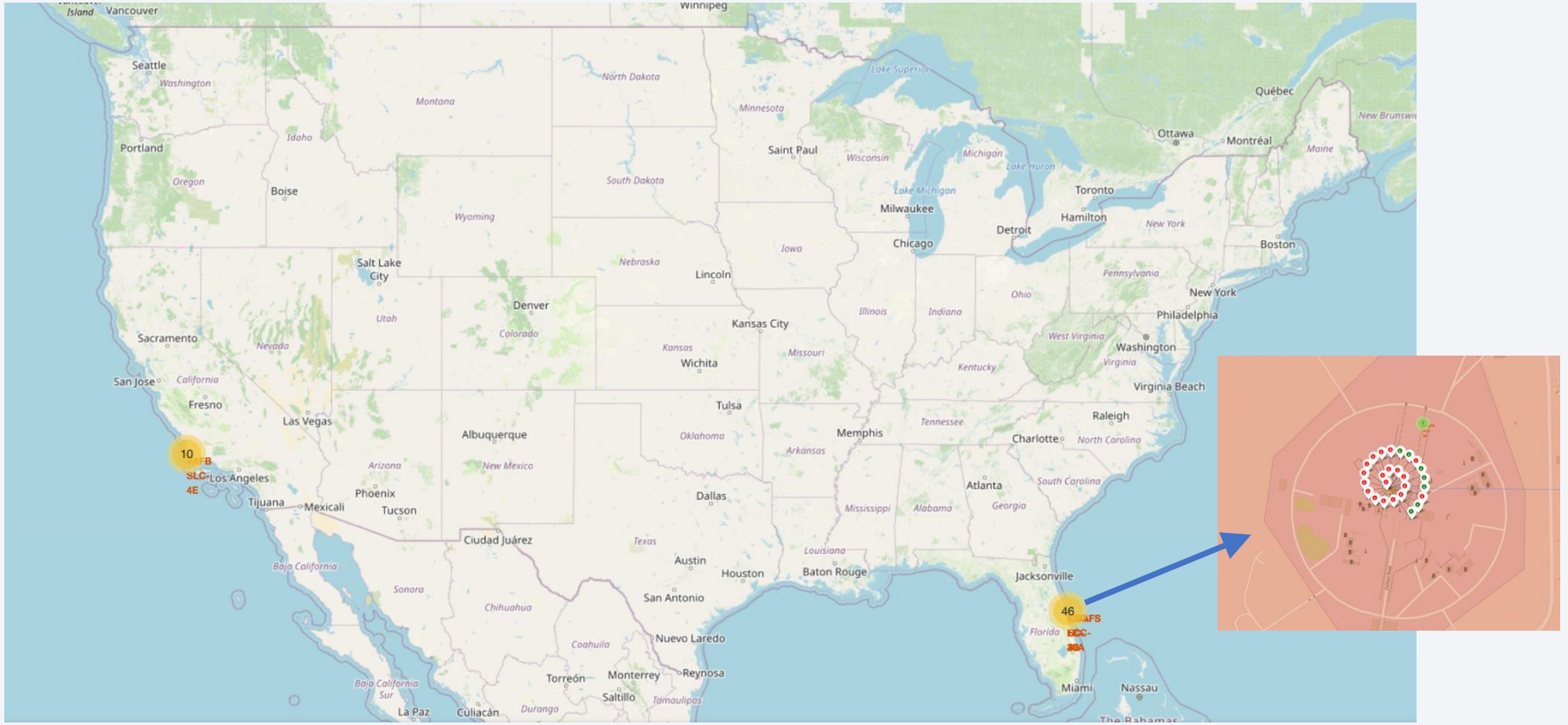
Launch Sites Proximities Analysis

Launch Site Locations

- Launch sites are near coastal regions, relatively close to the equator within the contiguous 48 states
- Proximity to oceans provides some safety such that debris can fall in sparsely populated areas



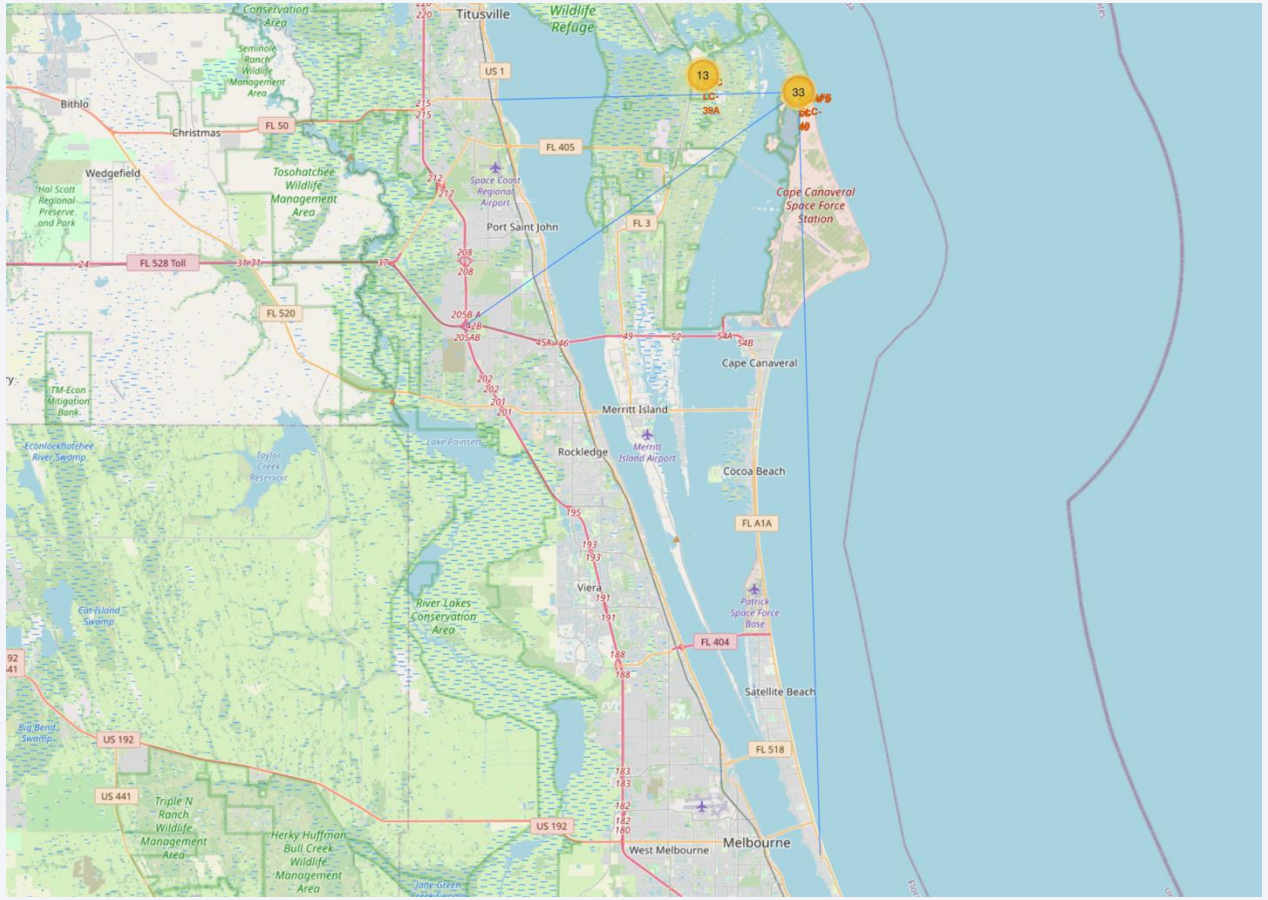
Launch Site Outcomes Map

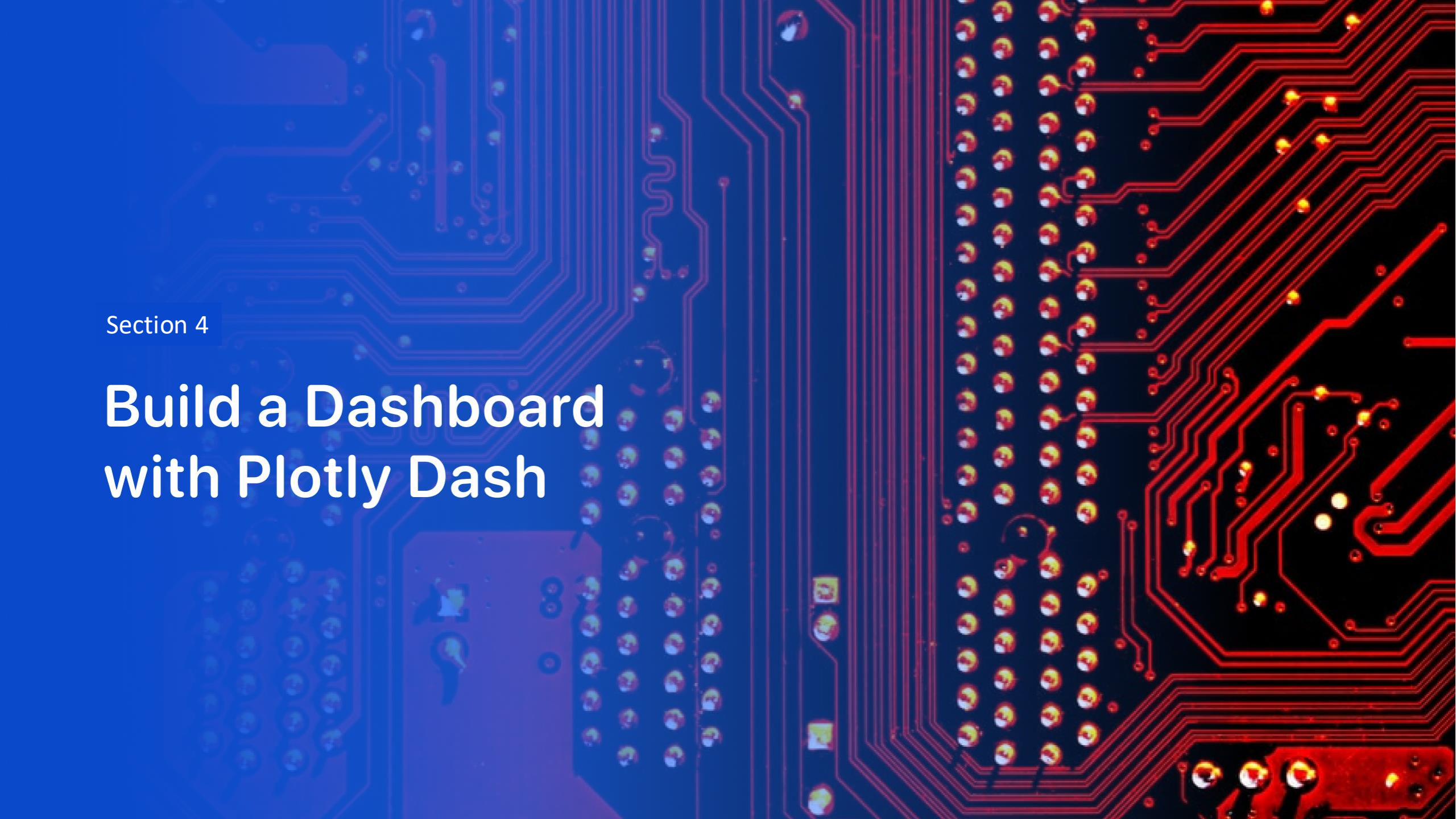


Proximity Landmarks

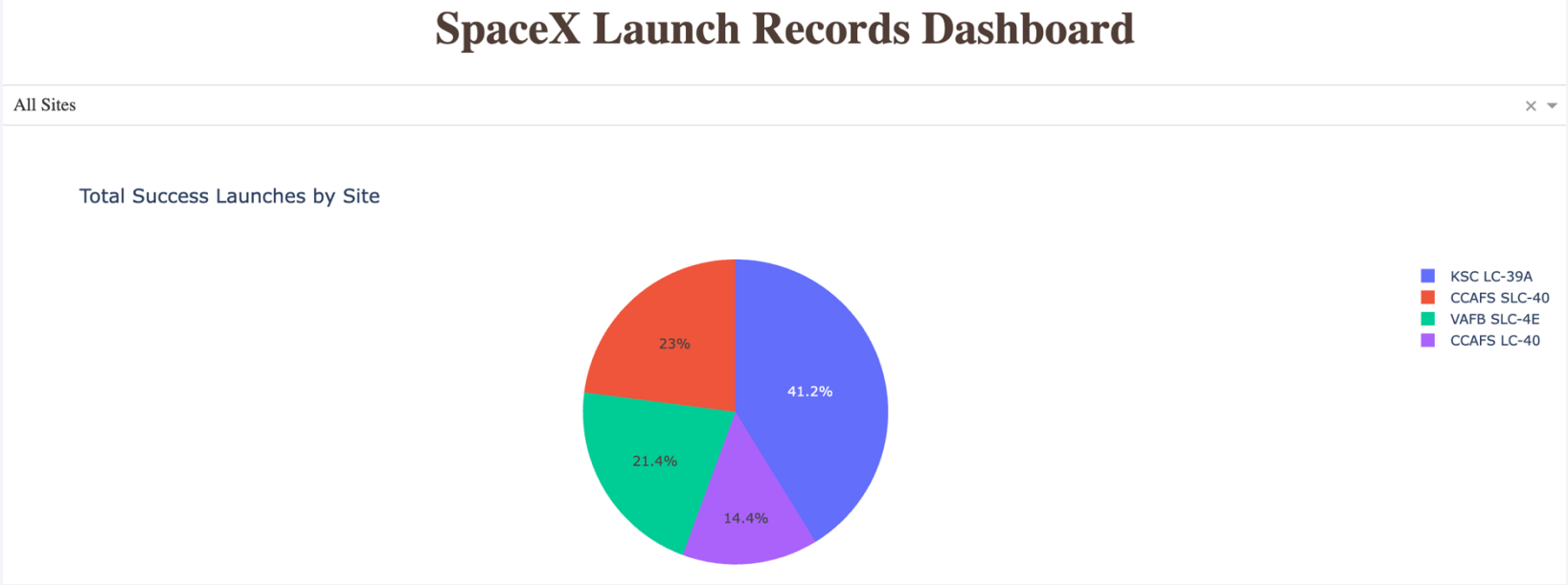
- Launch site to its proximities to major city, railway, and highway

Landmark	Distance to Landmark (KM)
Melbourne	54.76
Railway	21.91
Highway	29.09



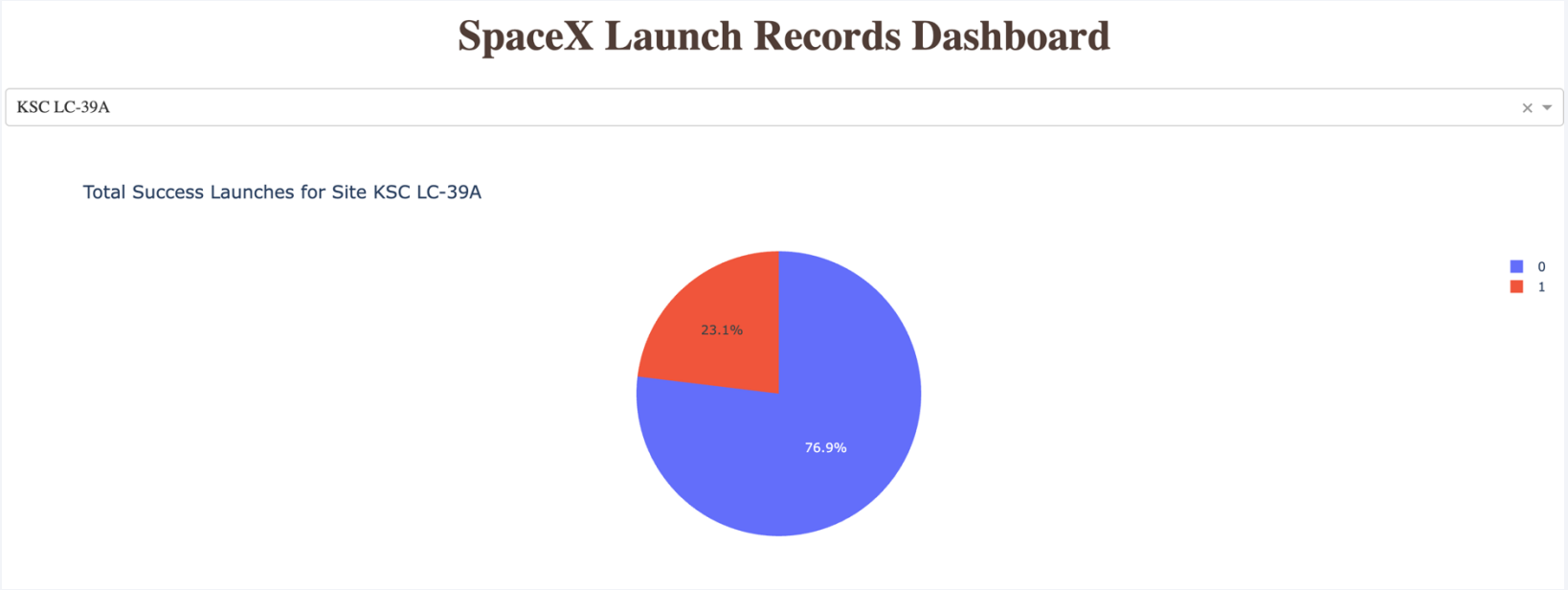

Section 4
Build a Dashboard with Plotly Dash

Launch Success Counts – All Sites



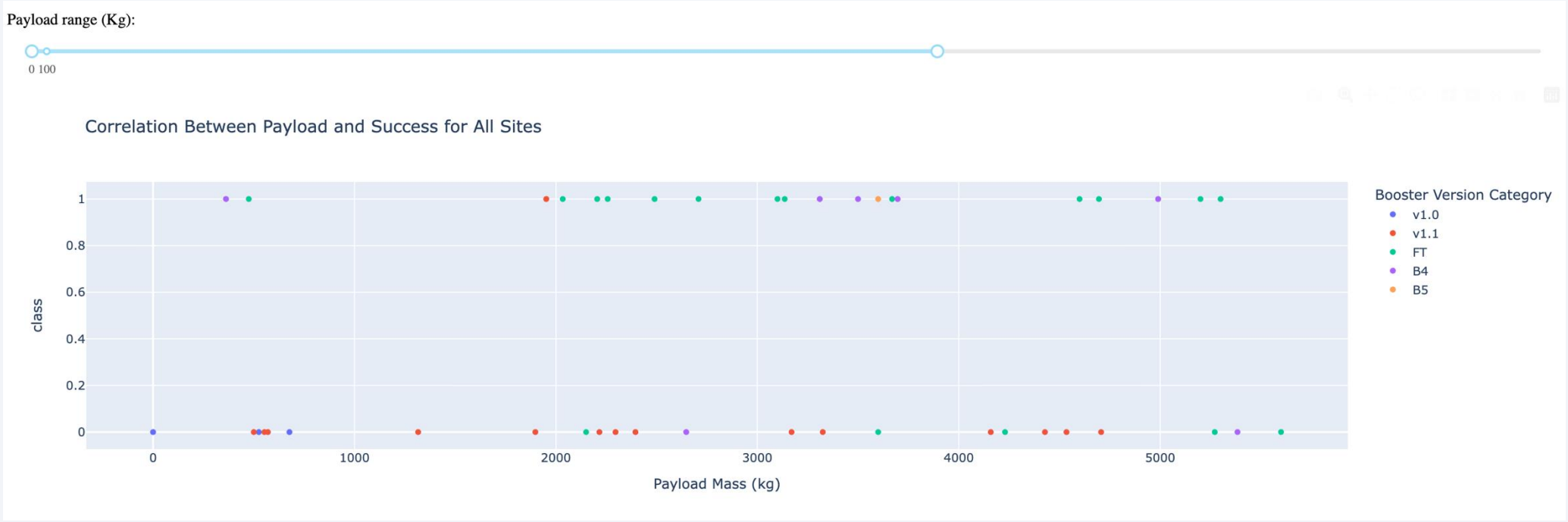
- KSC LC-39A is the most successful site with 41.2%
- CCAFS LC-40 is least successful with 14.4%

Launch Site – Highest Success Ratio



- KSC LC-39A had the highest success ration of 76.9%

Payload vs Launch Outcome – All Sites, <6000kg

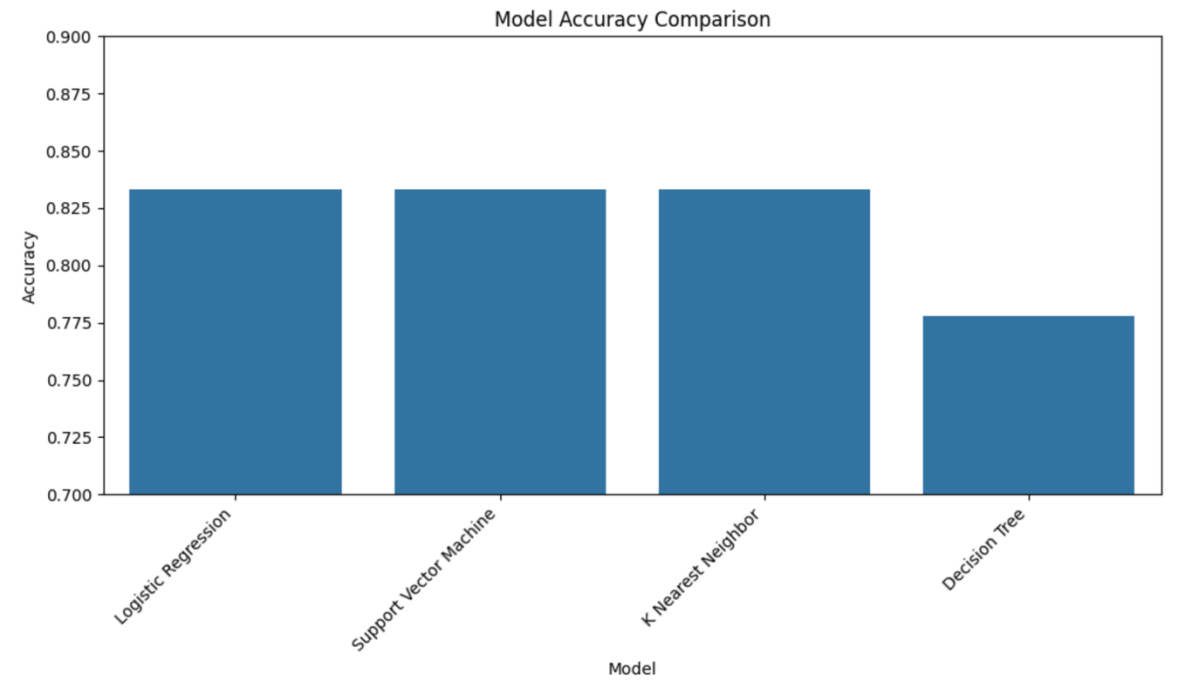


- In the range below 6000kg payload masses, the FT and B4 boosters had the most success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

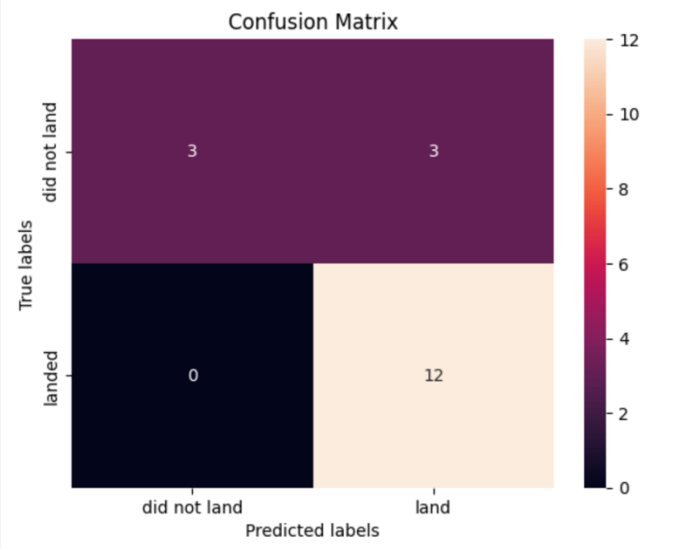


- All the models had the same accuracy except for the decision tree.
  - Note that an error appear when creating the data for the decision tree so that result may be a bit suspect

```
/lib/python3.12/site-packages/sklearn/model_selection/_validation.py:547: FitFailedWarning:
3240 fits failed out of a total of 6480.
The score on these train-test partitions for these parameters will be set to nan.
If these failures are not expected, you can try to debug them by setting error_score='raise'.
```

Confusion Matrix

- The confusion matrix for the log regression, SVN, and KNN are similar.



Conclusions

- Log regression, SVM, and KNN were all equally good models. Decision tree was the worst model though there may be an issue with the data imported/used.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Launches with a low payload mass show better results
- The success rate of launches improves over time.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!