IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Allen N
April 2025

Outline

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

Executive Summary

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most success and successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K -nearest neighbor (KNN

Exploratory Data Analysis:
• Launch success has improved over time • KSC LC-39A has the highest success rate among landing sites • Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics: • Most launch sites are near the equator, and all are close to the coast Predictive Analytics: • All models performed similarly on the test set. The decision tree

Introduction

- SpaceX aims to reduce the cost of space travel by recovering the first stage of its rockets. When successful, this recovery can result in savings of up to $100 million per launch. However, these savings are only realized if the first stage is successfully recovered.
- Using historical launch data, we seek to predict the likelihood of first stage recovery by analyzing several key factors:
  - What impact does payload size have on recovery success?
  - How does the launch site influence the outcome?
  - How has the probability of recovery changed over time?
  - Does the type of booster affect the chances of a successful recovery?

Section 1

Methodology

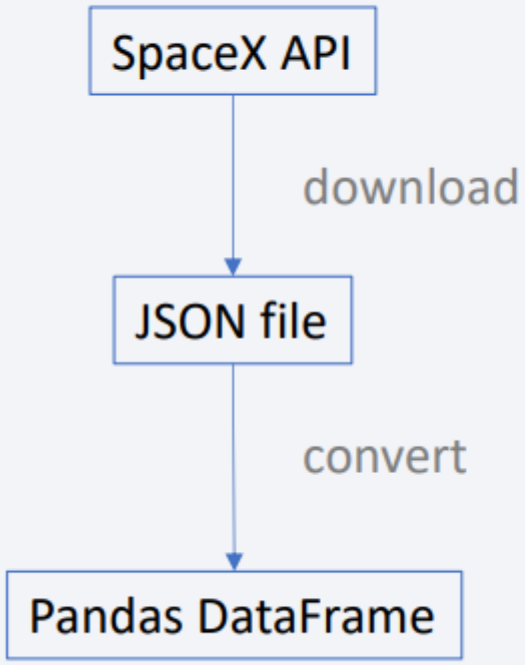
Methodology

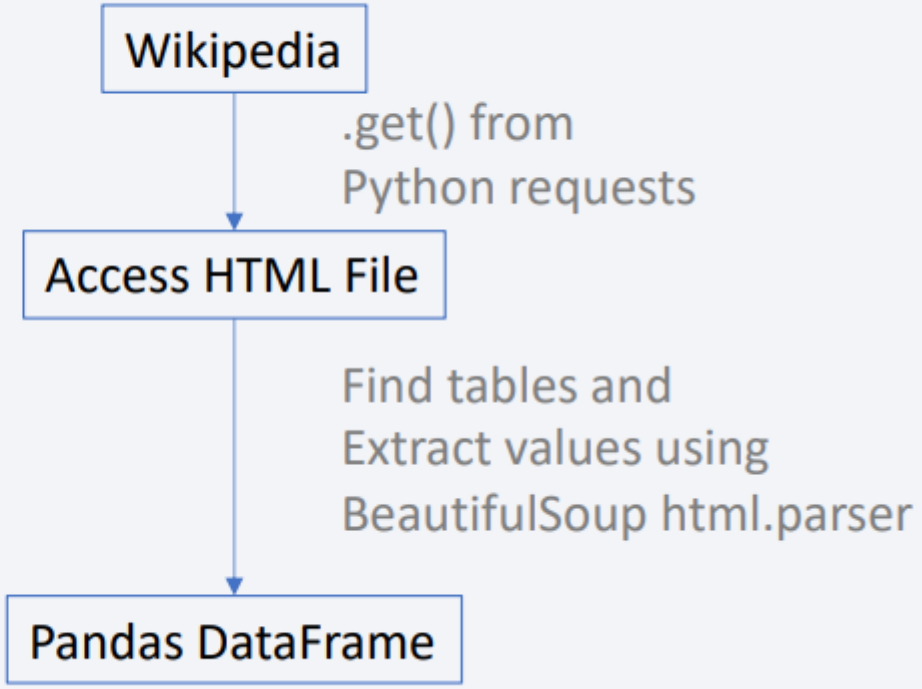
Executive Summary

- Data collection methodology:
  - Using SpaceX's publicly API (https://api.spacexdata.com/)
  - Utilizing launch history by scraping data from Falcon 9 Launch Wikipedia pages
- Perform data wrangling
  - Transformed data is easier to analyze and fill in missing data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuning model data for prediction and validation training

Data Collection

- Utilizing publicly available data sources (Space X API and webscraping Falcon 9 Launch Wikipedia

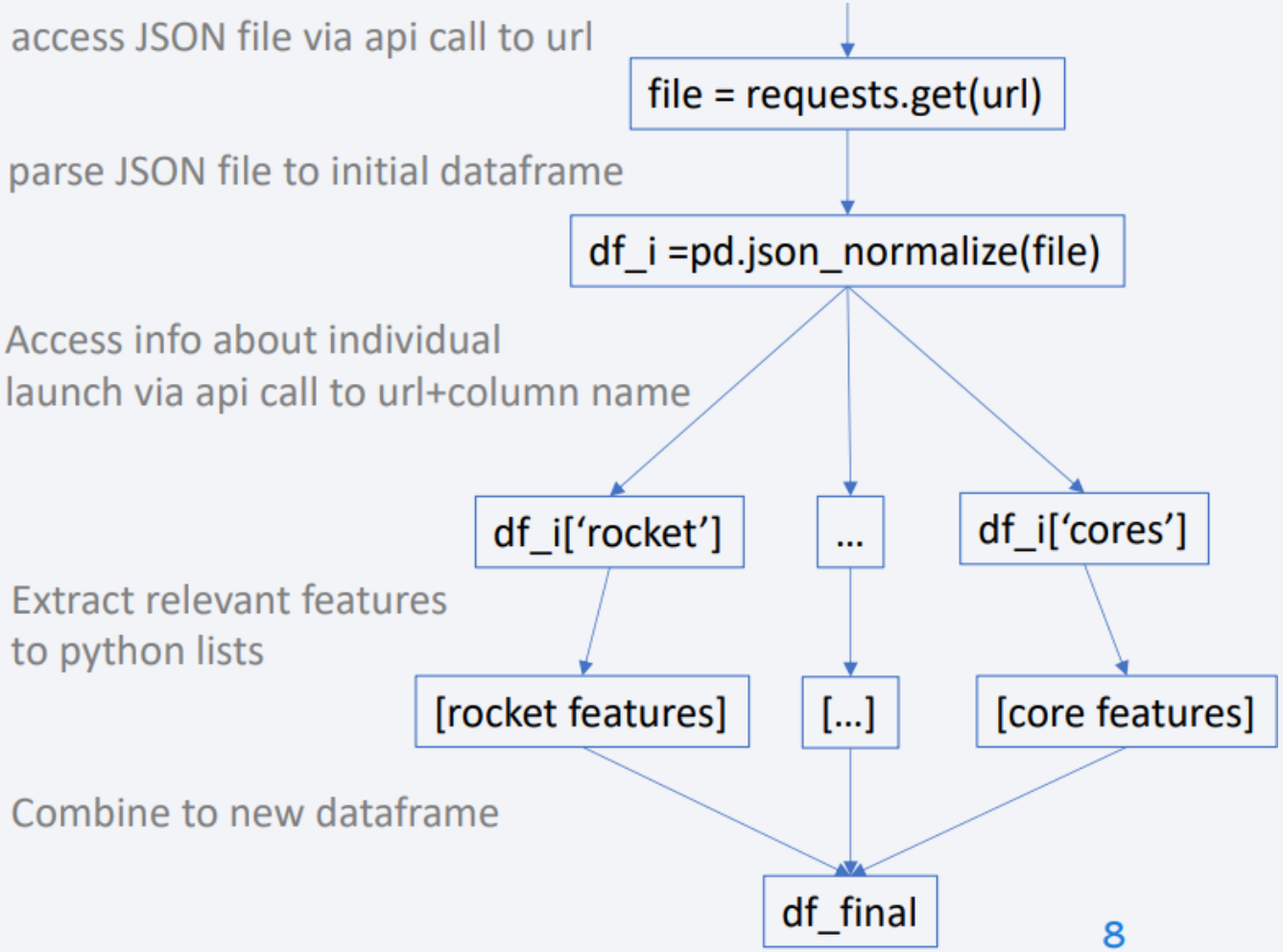




Data Collection – SpaceX API

- Data collection via SpaceX API
  - Using Pandas dataframe to pull data, accessing main level API where JSON files contain entries to be deciphered using specific API Calls

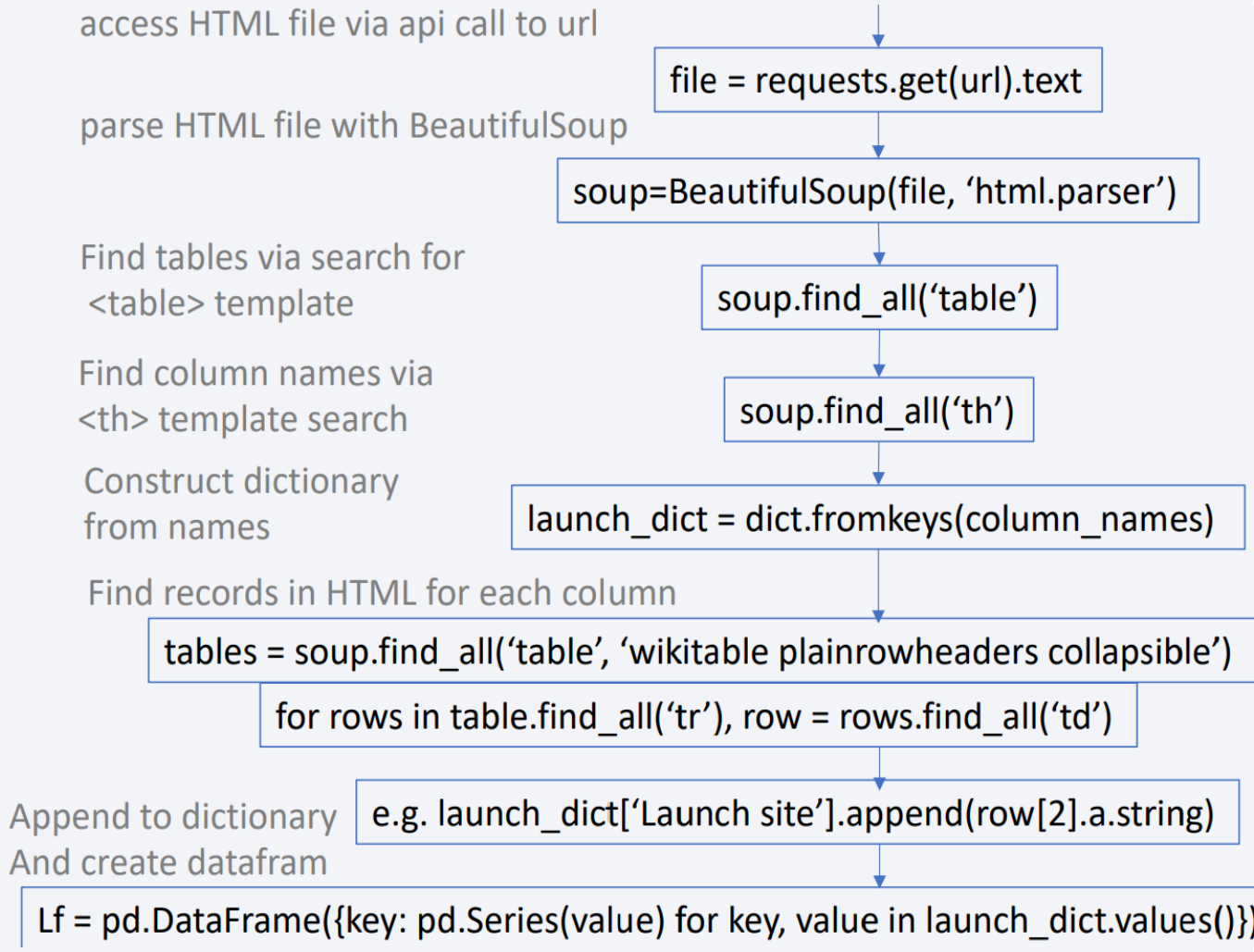
- https://github.com/shinobida/Coursera-Final/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

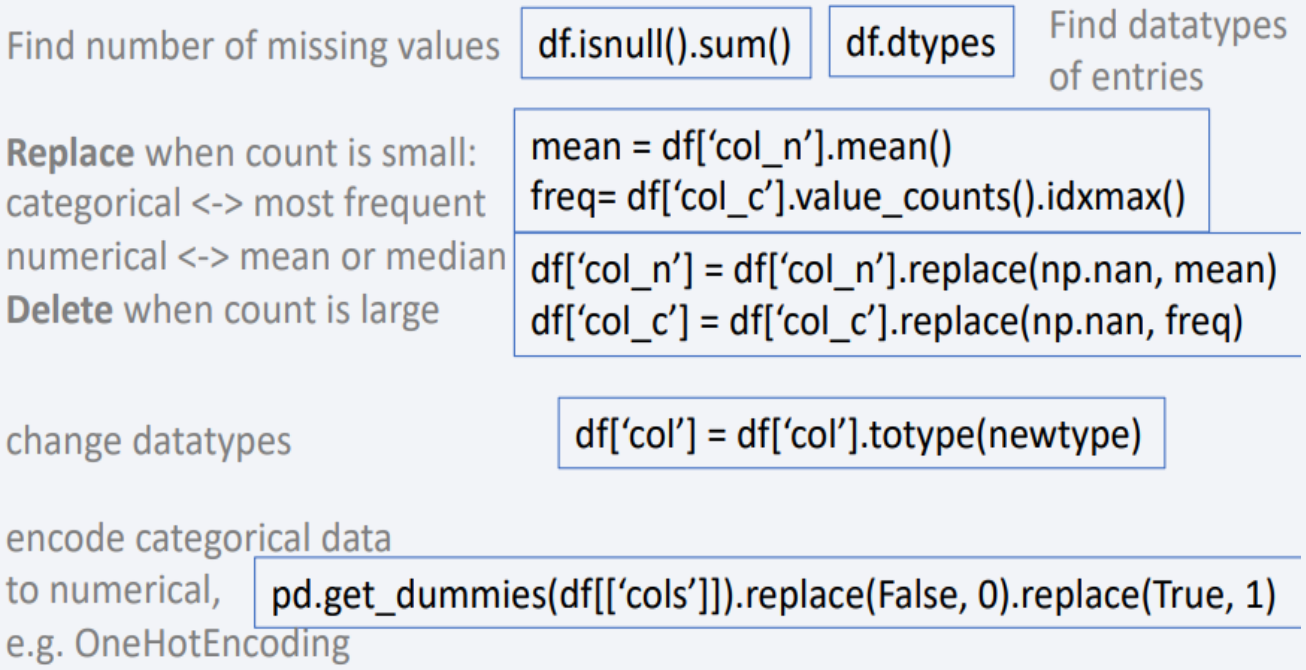
- Parsing HTML data from Falcon 9 Launch Wikipedia

- https://github.com/shinobida/Coursera-Final/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Cleaning datasets to replace/rescale categorical data with numerical values and change types in Data Wrangling
  - Missing values: 28% of LandingPad(categorical) missing. Choose to delete
  - Landing outcome is categorical is targeted, setting False and True to 0 and 1, respectively
- https://github.com/shinobida/Coursera-Final/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Utilizing various plats to get a better visualization of data structure

• Scatterplots (all colored by launch outcome)

  • Flight Number vs. Payload Mass This helps us see that more recent flights carried up to 8x more weight than early flights and had more successful outcomes
  • Flight Number vs. Launch Site. We see that Vandenberg is rarely used compared to Kennedy Space Center. Early flights from CCAFS SLC 40 were less successful than later flights.
  • Payload Mass vs. Launch Site: most low weight and high weight flights leave from KSC, while intermediate weight flights leave from VAFB
  • Flight Number vs. Orbit: we see that early flights went to LEO, ISS, PO, GFO while later flights went to HEO, VLEO, SO, GEO. Furthermore, LEO success seems to increase with flight number.
  • Payload vs. Orbit: heavy payloads tend to have more success with PO, LEO, and ISS though for GTO, the data is less clear

• Barplots

  • Orbit vs. outcome: we plot the percent of successful flights based on orbit type. ES-L1, GEO, HEO, SSO all have success rates of 1, while GTO has the lowest success rate of 0.5

• Lineplots

  • Success vs. year: we see the success rate tend to increase over the years, with a dip in 2018 and no change from 2010-2013 and from 2014-2015

- https://github.com/shinobida/Coursera-Final/blob/main/edadataviz.ipynb

EDA with SQL

- List unique launch sites:: %sql select DISTINCT "Launch_Site" from SPACEXTABLE
- Display 5 records where launch sites begin with string ‘CCA’: %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
- Display total payload mass carried by boosters launched by NASA (CRS) : %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like 'NASA (CRS)%’
- Display average payload mass carried by booster version F9 v1.1: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%’
- List the date when the first successful landing outcome in ground pad was achieved: %sql select min(Date) from SPACEXTABLE where "Landing_Outcome" like 'Success (ground pad)’
- List the names of the booster w/ success in drone ship and mass between 4000 and 6000: %sql select distinct "Booster_Version" from SPACEXTABLE where (("Landing_Outcome" like 'Success (drone ship)') and (4000<=PAYLOAD_MASS__KG_<=6000))
- Display total success and failure: %sql select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome"
- List all the booster versions that have carried the maximum payload: %sql select distinct "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
- List records from the year 2015: %sql select substr(Date, 6, 2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where (substr(Date, 0, 5)='2015' and "Landing_Outcome" like 'Failure (Drone Ship)’)
- Rank the landing outcomes between 2010-06-04 and 2017-03-20 in descending order: %sql select "Landing_Outcome", count(*) as Count from SPACEXTABLE where (Date > '2010-06-04' and Date < '2017-03-20') Group By "Landing_Outcome" Order By Count DESC

https://github.com/shinobida/Coursera-Final/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

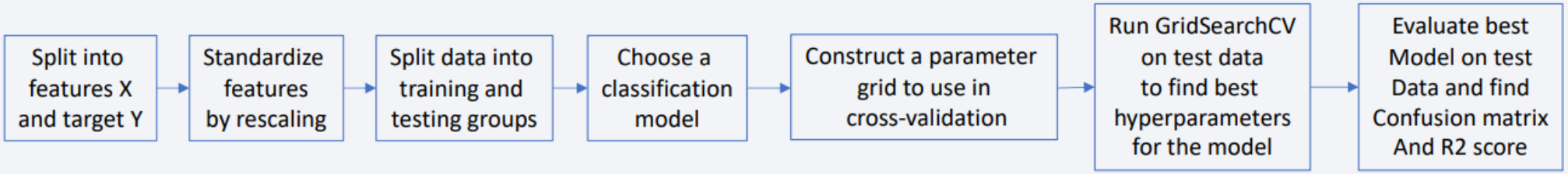
- Creating an interactive map using Folium. MarkerClusters grouping for markers at Vandenburg and Kennedy Space Center
  - Color coding launches by success or failure
- Showing additional landmarks to indicate why launch sites are chosen (isolation from population, railroad access, optimal flight locations
- https://github.com/shinobida/Coursera-Final/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Using Plotly Dash, created a dashboard that
  - Pie plot showing success rate at specific sites
  - Scatterplot showing success rate vs payload range
- Including graphics that would compare individual results to overall results
- https://github.com/shinobida/Coursera-Final/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

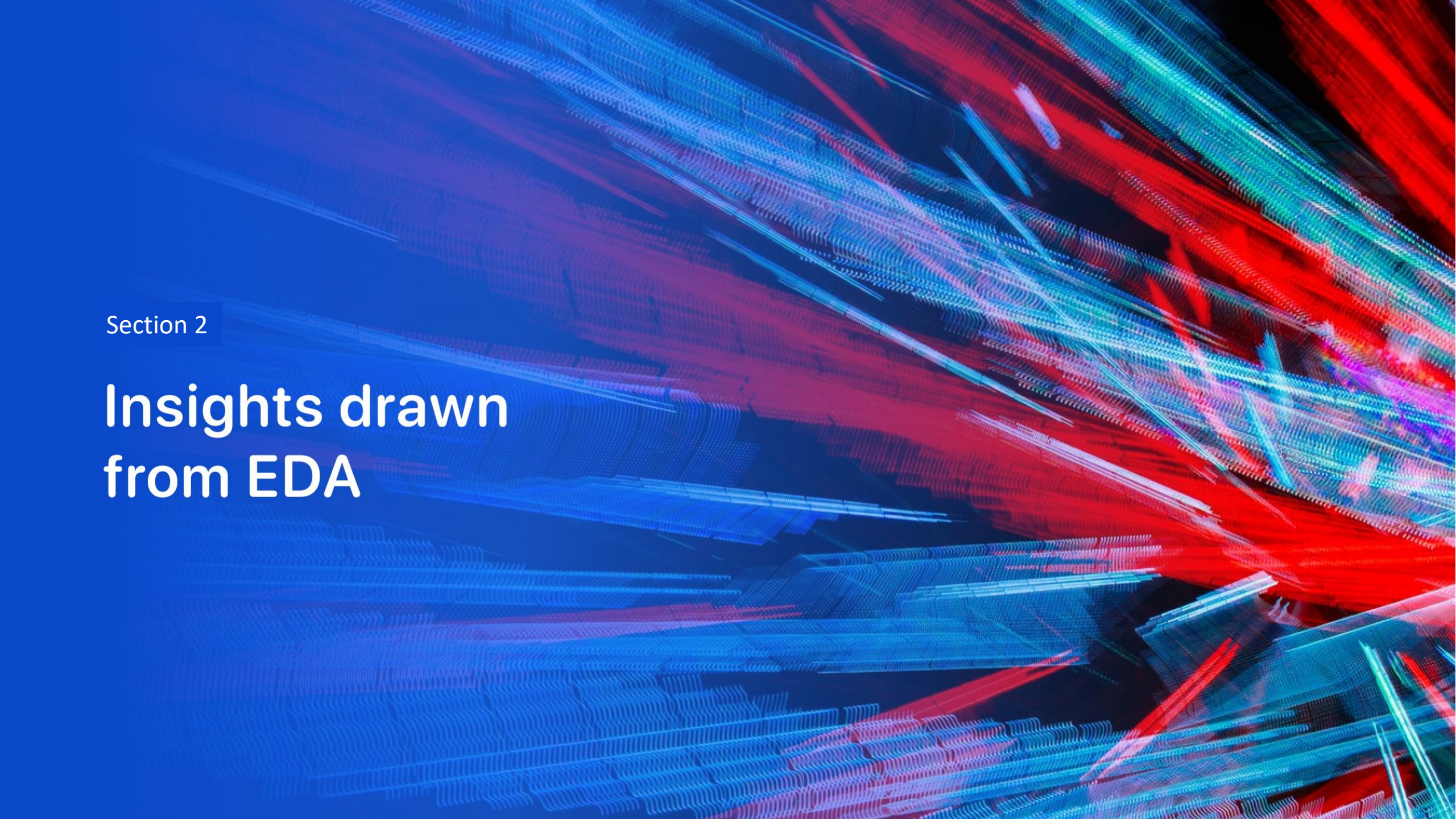
- Our objective was to predict the success of rocket launches based on a range of influencing factors. To achieve this, we evaluated several classification algorithms, including logistic regression, support vector machines, decision trees, and K-nearest neighbors. We optimized each model's hyperparameters using GridSearchCV with cross-validation, integrated within streamlined model pipelines. Our workflow followed these key steps:



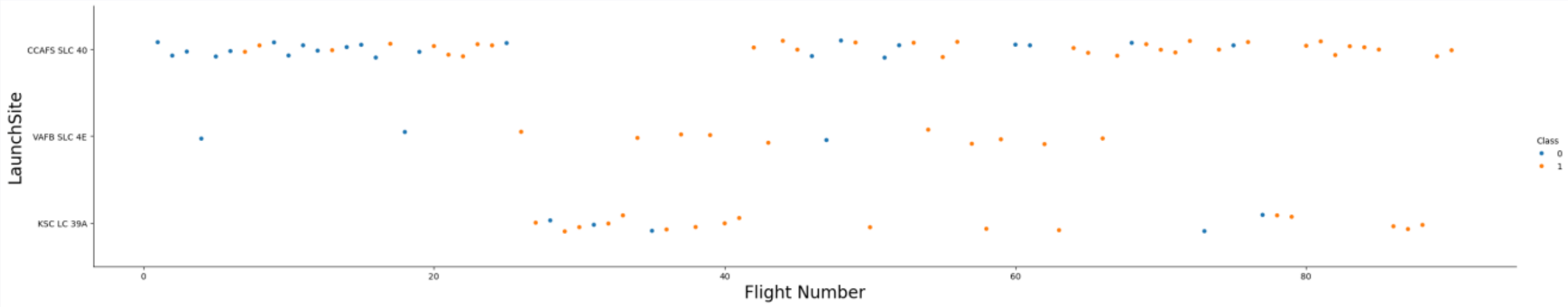
- https://github.com/shinobida/Coursera-Final/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

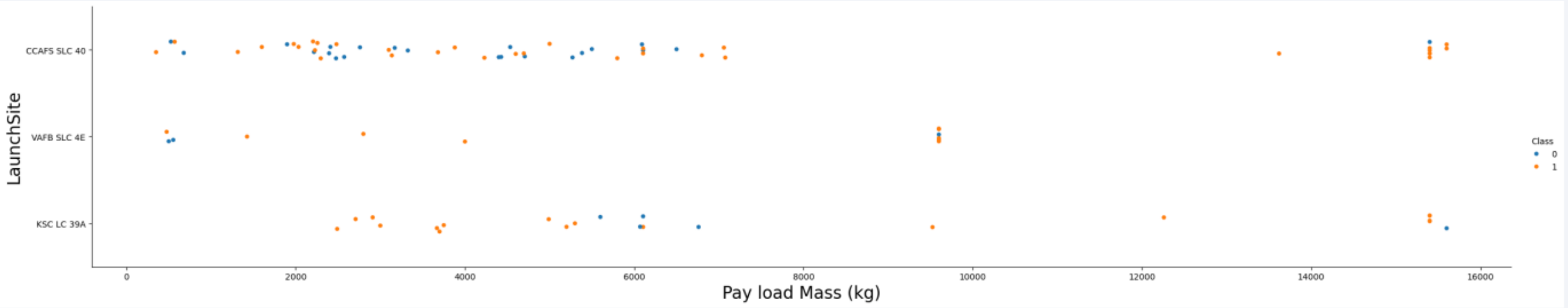

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



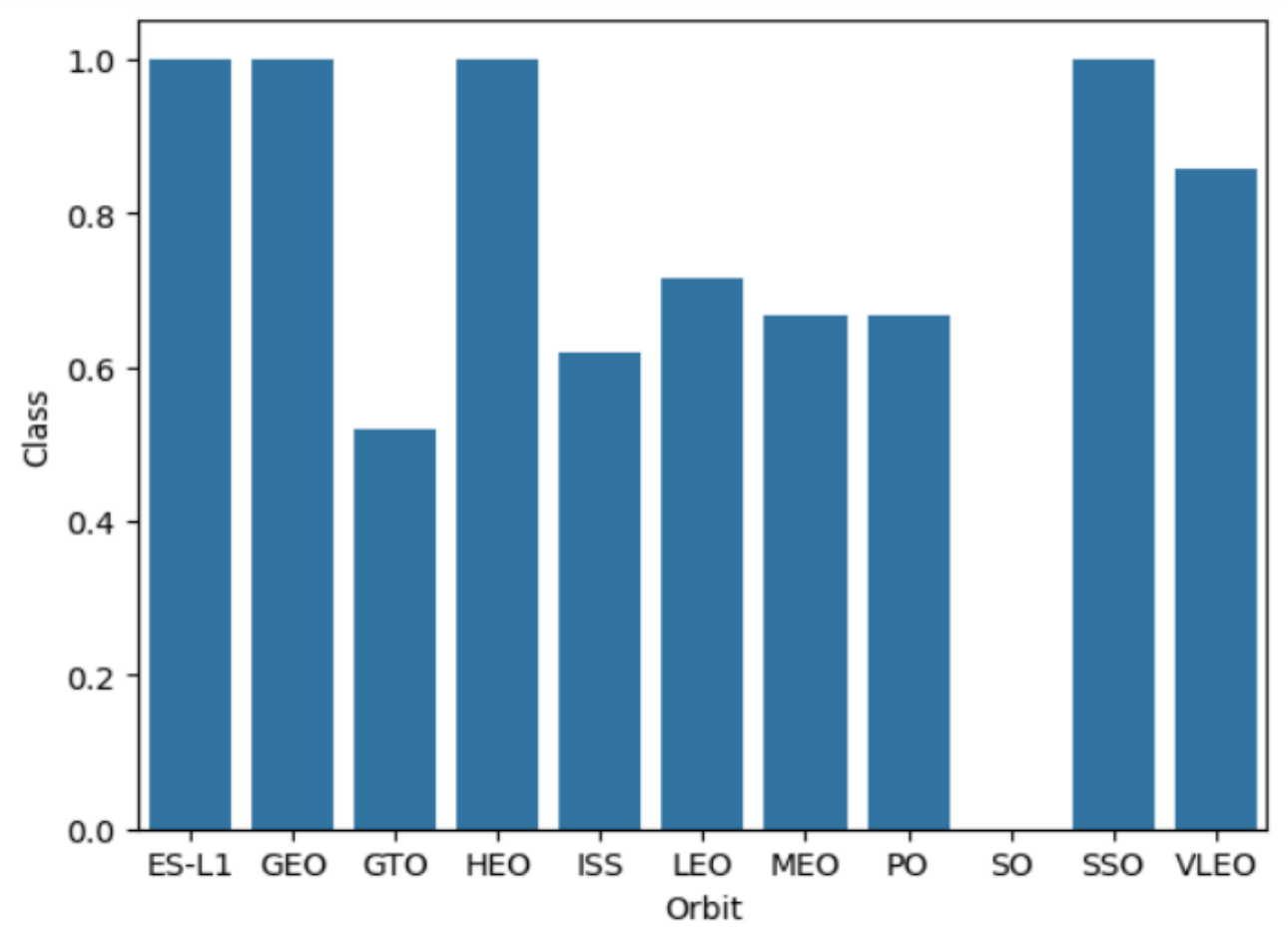
Here we plot flight number vs. Launch Site. CCAFS SLC 40 was the most common launch site early on and remains heavily used, while KSC LC 39A is now also frequently used. VAFB SLC 4E was common in Intermediate times. Success is fairly equal among sites recently, though early flights were unsuccessful

Payload vs. Launch Site



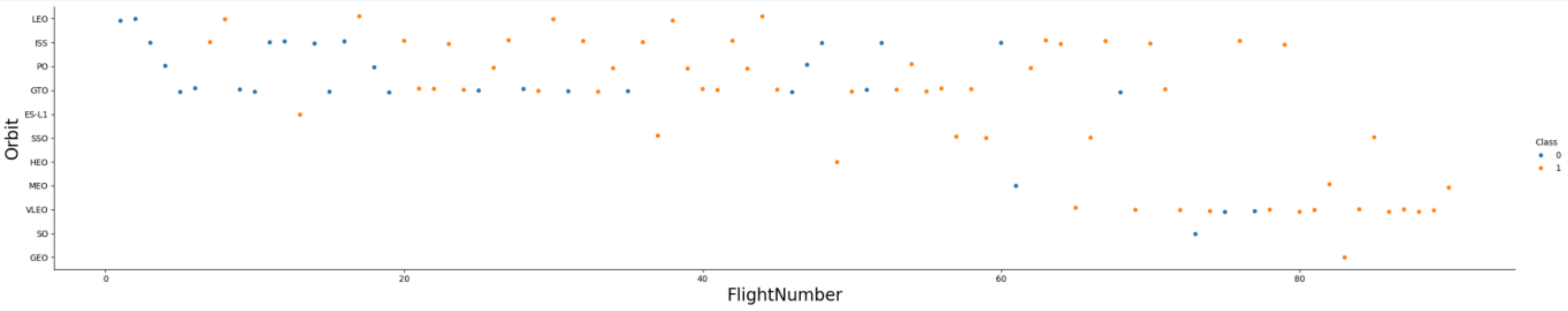
Here we plot Payload vs. Launch Site. Heavy payloads are exclusively launched from Kennedy Space Center and tend to be successful. Light payloads are launched from VAFB and CCAFS with varying success. Few intermediate mass payloads are launched from KSC, though they tend to be successful.

Success Rate vs. Orbit Type



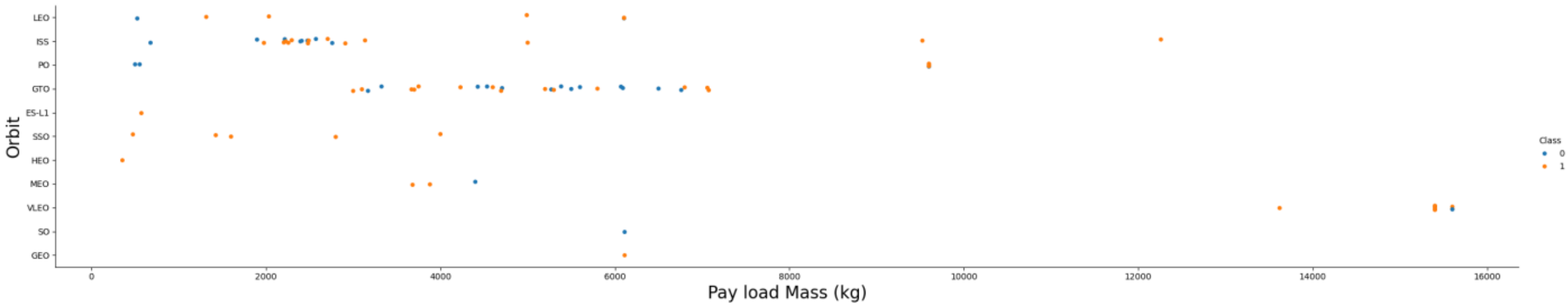
ES-L1, GEO, HEO, SSO have total success, whereas GTO is only 50% successful.

Flight Number vs. Orbit Type



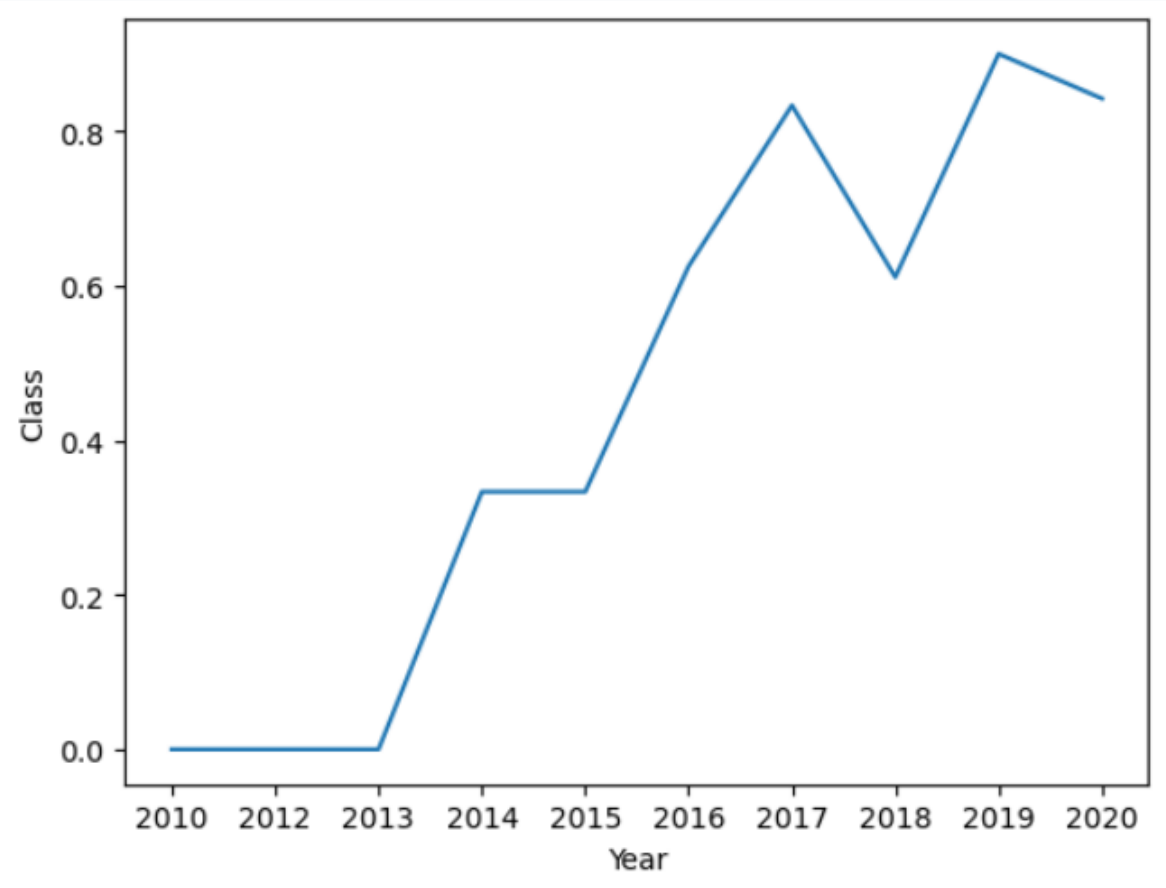
Early flights were restricted to LEO, ISS, PO, GTO, and ES-L1. Recently, Flights to VLEO, SO, MEO, and GEO have occurred. Flight success is more correlated with time than orbit, though GTO has middling success whereas VLEO is more successful.

Payload vs. Orbit Type



Light payloads are regularly sent to most orbits, though no VLEO, SO, GEO orbits. GTO has the widest range of payload masses, while VLEO has the heaviest payloads. There is large clustering of payload mass to the ISS.

Launch Success Yearly Trend



Launches have become more successful over time, though there was a decrease in 2018 and stagnation from 2010-2013 and 2014-2015.

All Launch Site Names

- Use SQL query to find unique launch site names

```
select DISTINCT "Launch_Site" from SPACEXTABLE
```

- Using “DISTINCT” on the Launch_Site column ensures results are unique to the field

Launch Site Names Begin with 'CCA'

- select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
- This will pull 5 records with launch site beginning with CCA
- The * symbol means to choose all columns from the table

Total Payload Mass

- select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like 'NASA (CRS)%’
- SQL query will calculate total payload carried by boosters from NASA (CRS)
- The ”sum()” function sums all values in the column PAYLOAD_MASS_KG_ that have a customer matching NASA (CRS).

Average Payload Mass by F9 v1.1

- select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%’
- Calculates average payload mass
- “avg()” function calculates the average of the column PAYLOAD_MASS__KG_ for records that have a Booster_Version starting with F9 v1.1 using % as the identifier

First Successful Ground Landing Date

- select min(Date) from SPACEXTABLE where "Landing_Outcome" like 'Success (ground pad)’
- Finds dates of successful first landing on Ground Pad
- The “min()” function can parse dates to find the smallest (earliest) date among the entries where the “Landing_Outcome” matches “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

- select distinct "Booster_Version" from SPACEXTABLE where (("Landing_Outcome" like 'Success (drone ship)') and (4000<=PAYLOAD_MASS__KG_<=6000))
- This lists boosters which have successfully landed on drone ship with payload mass over 4000 but under 6000
- “distinct” searches for specific/unique names utilizing “and” to combine queries

Total Number of Successful and Failure Mission Outcomes

- select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome“
- Calculates total number of mission outcomes (successful and failures)

- select count("Mission_Outcome") as "Successful Missions" from SPACEXTABLE where Mission_Outcome like 'Success%’
- Splits mission outcomes to exclude failures and group all Successes

Boosters Carried Maximum Payload

- %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

- Lists names of booster which carried max payload

2015 Launch Records

- select substr(Date, 6, 2) as month, "Landing_Outcome" , "Booster_Version" , "Launch_Site" from SPACEXTABLE where (substr(Date, 0, 5)='2015' and "Landing_Outcome" like 'Failure (Drone Ship)’)
- Lists failed landing_outcomes in droneship, booster version, launch site names in 2015
- Sqllite syntax requires that we use substr(Date, 6, 2) to look for months and substr(Date, 0, 5) to look for years.

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

- To 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, use the sql query:
- select "Landing_Outcome", count(*) as Count from SPACEXTABLE where (Date > '2010-06-04' and Date < '2017-03-20') Group By "Landing_Outcome" Order By Count DESC
- Ranks landing outcome counts between dates 2010-06-04 and 2017-03-20, in descending order
- We can group records by Landing_Outcome and then calculate the total counts in each group using count(*), renaming the column ”Count.”, order the results by the new column “Count” in descending order using “order by Count desc"

Section 3

Launch Sites Proximities Analysis

All Launch Sites

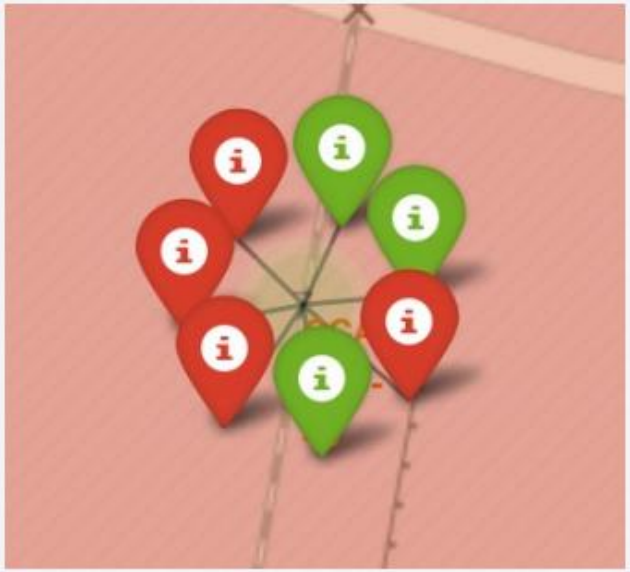
- Launch sites for SpaceX launches. There are three locations, one on the west coast of the United States at Vandenburg SFB and two at the Kennedy Space Center (overlapping on the map).

Launch Outcomes

VAFB SLC-4E

KSC LC-39A

CCAFS SC-40



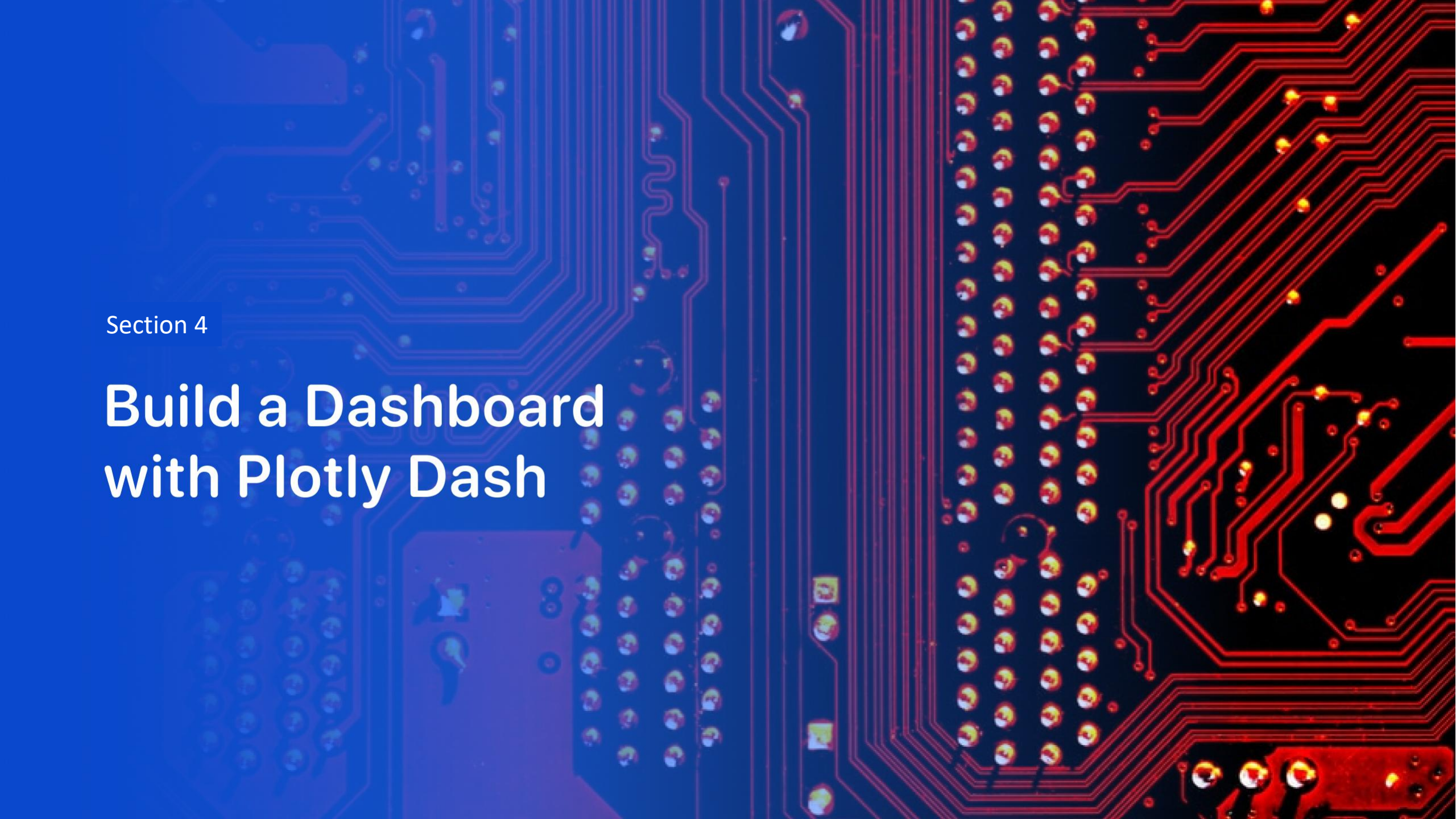
CCAFS LSC-40

Kennedy Space Center

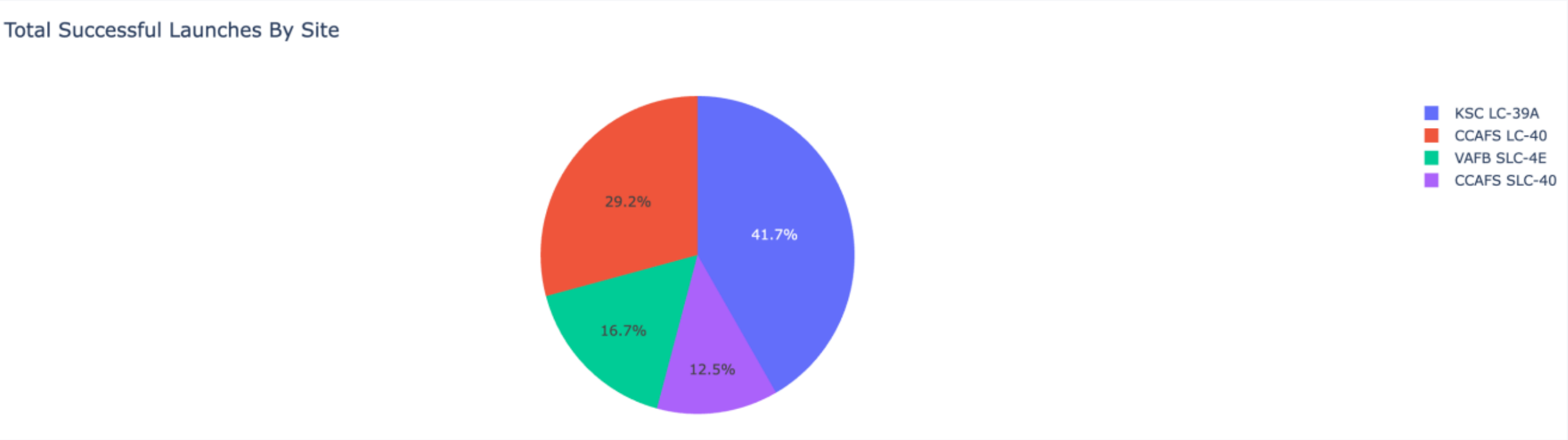
- The most successful launch site was KSC LC-39A, while the least successful was CCAFS SC-40

Distance to nearby landmarks from CCAFS LC-40

- Distance to nearest railroad and Titusville > 20 km
- Distance to Coastline and Highway < 1 km
- Launch sites are chosen to be close to the ocean for optimal flight path and to parkways for access to launch site, but far from towns and railroads to not disrupt infrastructure or pose a risk to civilians

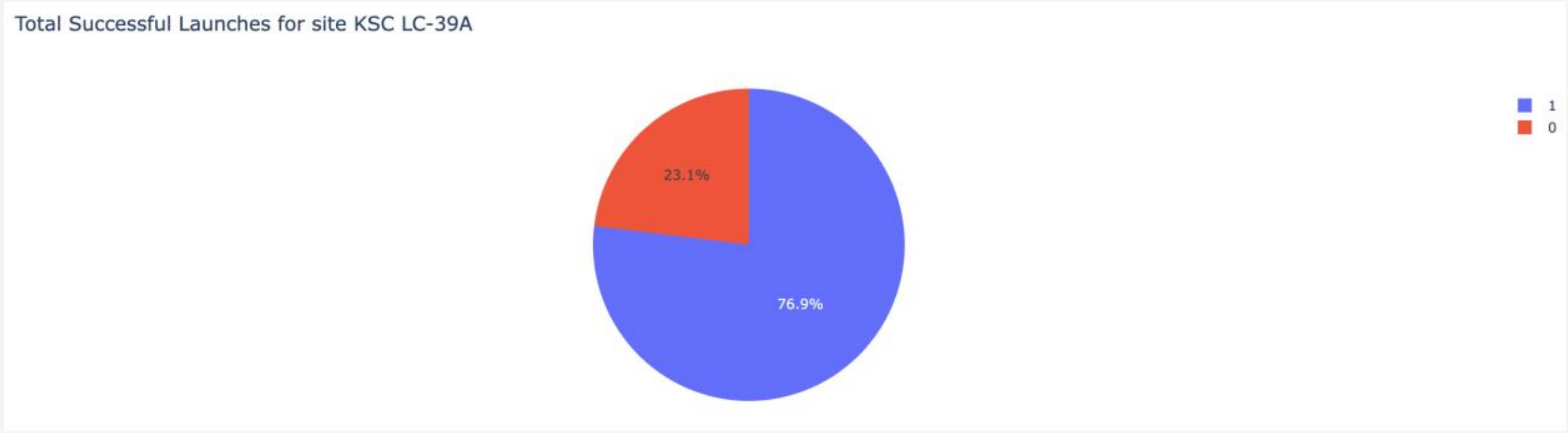

Section 4
Build a Dashboard with Plotly Dash

Overall launch success



- The most successful launch site was KSC LC-39A, and the least successful was CCAFS SLC-40

Most successful site



- KSC LC-39A saw the most success with 76.9% of launches successful (1) compared to a 23.1% failure rate (0)

Payload effect of launch outcomes

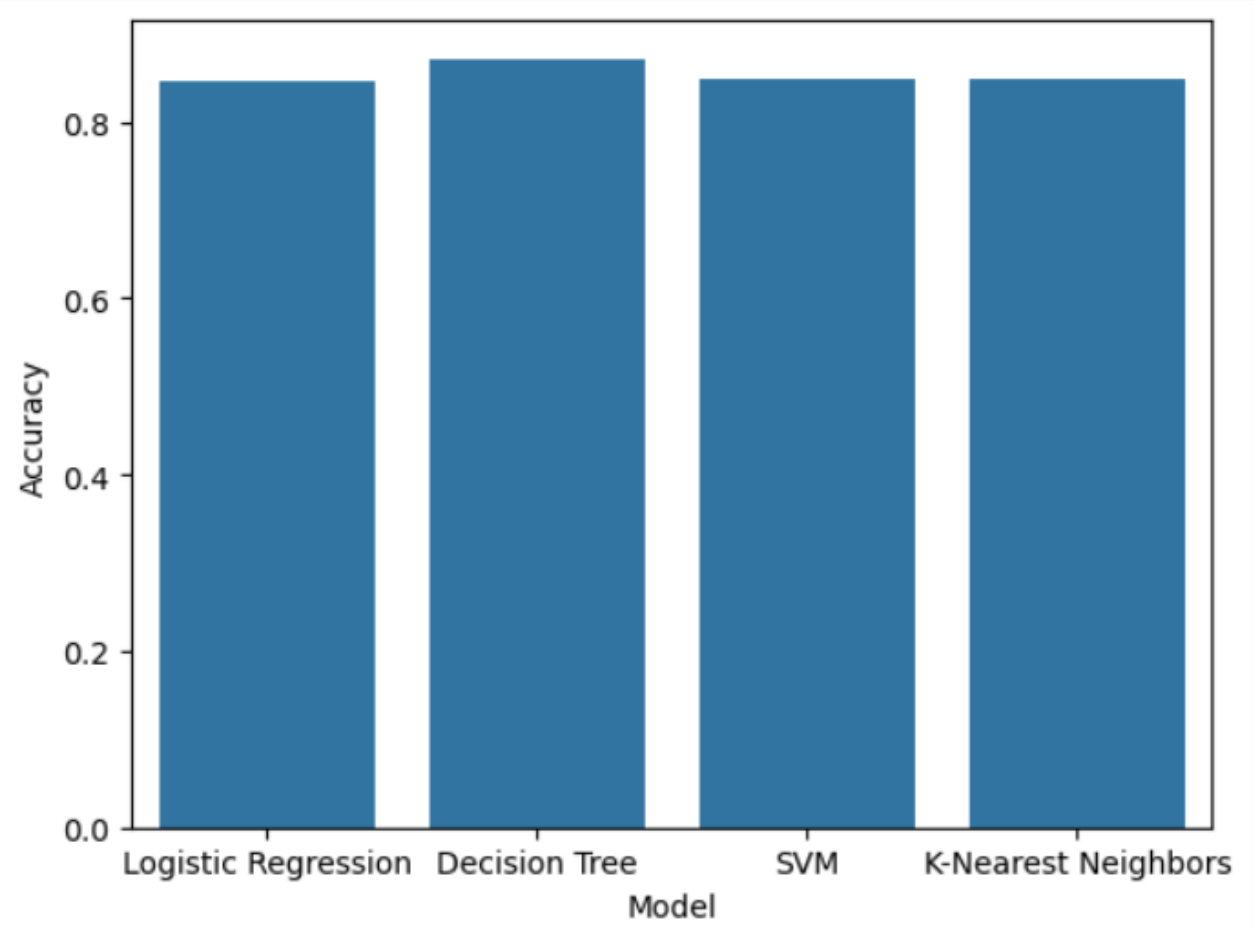
Zoomed in to highlight successful launches

- Most successful launches are in the intermediate payload range on FT boosters. Unsuccessful launches tend to be light payloads and v1.1 boosters

Section 5

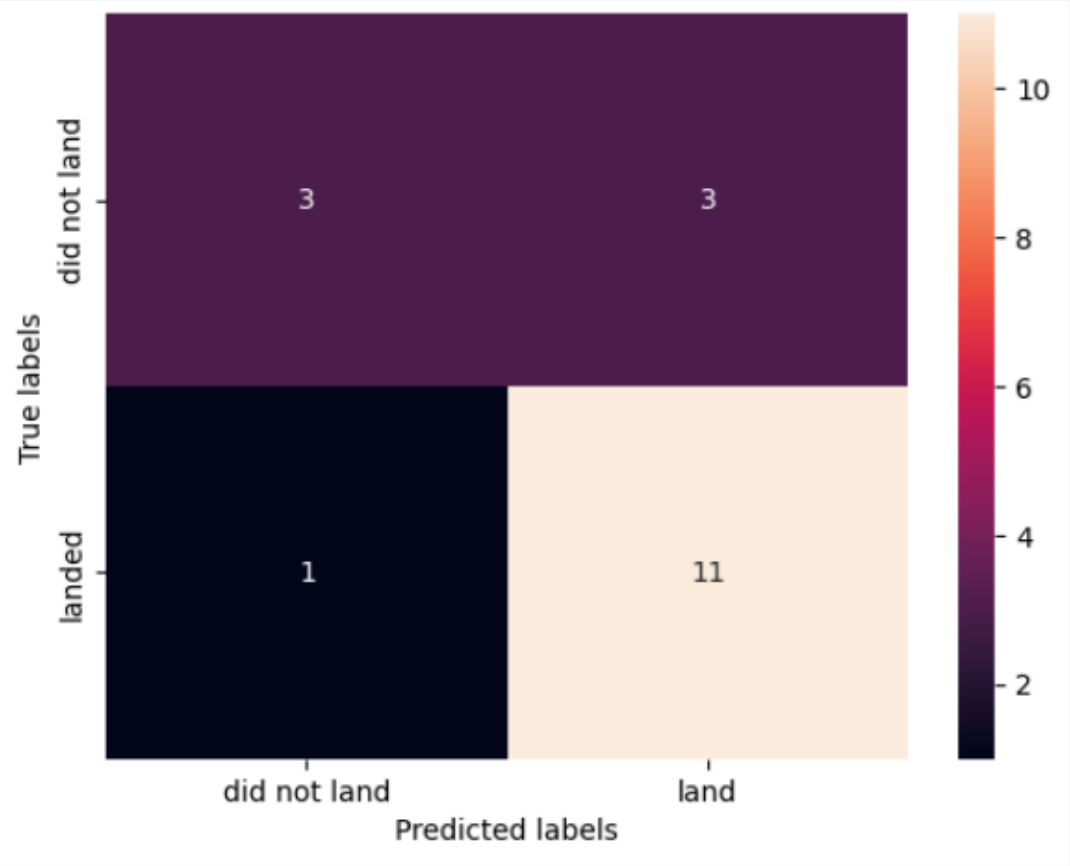
Predictive Analysis (Classification)

Classification Accuracy



The most accurate model was a decision tree with an accuracy of 87% compared to the other models with accuracies of ~85%

Confusion Matrix



The best performing model, a decision tree, predicted 3 false positives (predicted land when no landing occurred) and 1 false negative (predicted no land when a land occurred). It predicted 3 true negative and 3 true positives.

Conclusions

•SpaceX launches have shown increasing success rates over time and are now capable of carrying heavier payloads.

•Launch activity at Vandenberg Air Force Base (VAFB) has declined, with Kennedy Space Center emerging as the preferred site for both heavy and light payloads.

•While there is no strong direct correlation between payload weight and launch success, heavier payloads have seen more successful outcomes—likely due to improvements in booster technology and a strong correlation with time.

•Launch success can be effectively predicted using various machine learning models, with the best models achieving an accuracy of approximately 87%.

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!