IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Erik Blake
March 10, 2025

Outline

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

Executive Summary

This analysis aims to determine factors which predict whether a Falcon 9 First Stage will land or not. By leveraging exploratory analysis and machine learning, we identify features most critical to the overall success of the launch

Key findings:

- **Experiential Learning** - Landing success improved significantly over time suggesting that SpaceX learned from past failures
- **Orbit Complexity** - There is a significant difference in landing success based on orbit type
    - 82% of failed landings occurred for either GTO or ISS obit types, suggesting unique challenges for these trajectories
- **Launch Site Differences** – Landing success appeared to be influenced by where the rocket launched, suggesting regional factors such as weather, launch pad, launch team, etc.
- **Model Accuracy** – A reliable machine learning model was developed to predict first stage landing success with 87% accuracy to unseen data

Impact

These findings suggest that we can not only understand and predict whether the first stage lands successfully, but we can also predict the cost associated with future launches allowing Space Y to make competitive bids moving forward. Approximate cost could approach $35M (see appendix for calculation)

Introduction

- Commercial space travel is becoming more affordable and is seeing increased demand. Space X has been able to reduce the cost per launch significantly from ~$165M to ~$62M mainly due to figuring out how to reuse the First Stage. Space Y is looking to become more competitive in the market and has hired us to create a prediction model. If a model can determine if a First Stage can land, the cost to launch and land successfully can be calculated
- The objective of this project is to
  - Explore Space X historical data to better understand patterns and trends
  - Develop a Machine Learning model that can predict if the First Stage will land successfully
  - Approximate the cost of a launch

Section 1

Methodology

Methodology

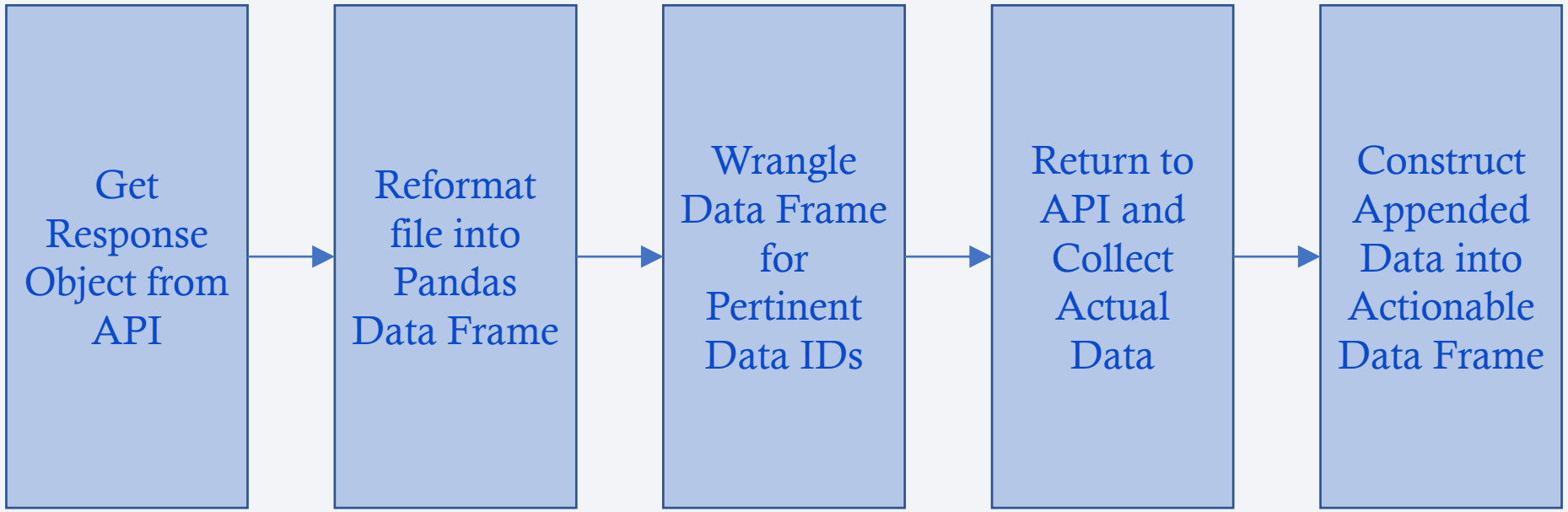
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, and evaluate classification models

Data Collection

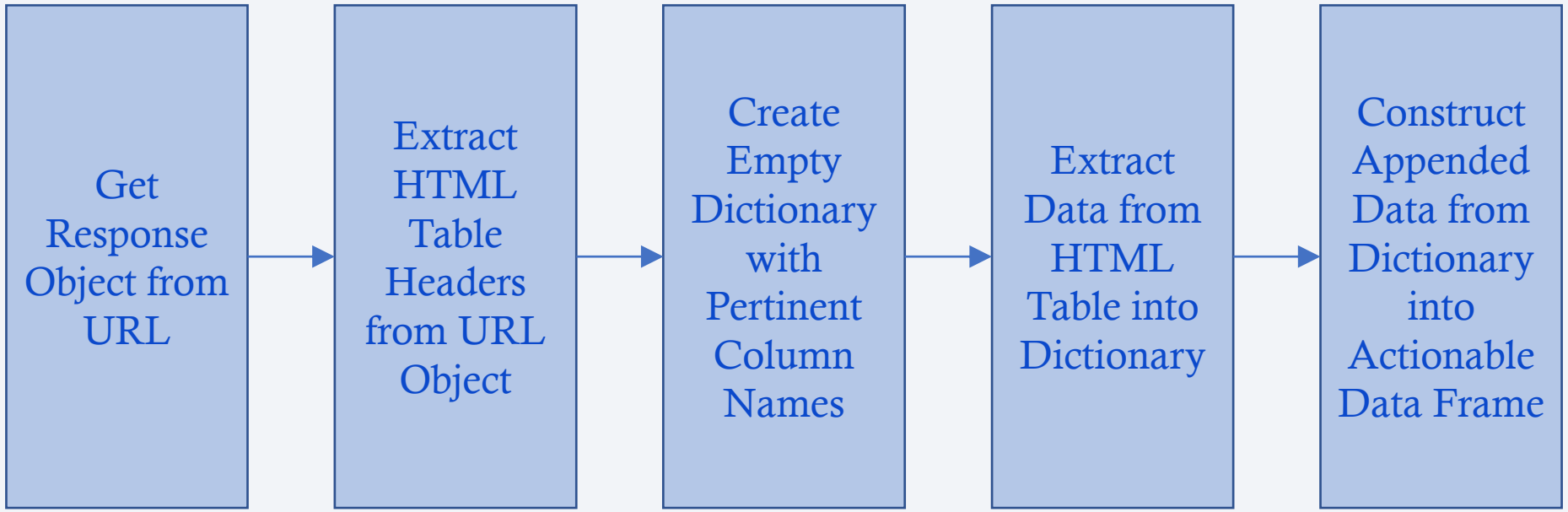
- Data Collection of the SpaceX data was executed in two separate ways
  - Data was retrieved from an Application Programming Interface
  - Data was scraped from Wikipedia page using Beautiful Soup
- API retrieval – High level summary
  - Retrieve a preliminary set of data IDs in the form of a Json file using a get function
  - Turn the IDs into a data frame to perform some preliminary data wrangling
  - Go back into the API with user designed functions to retrieve the desired information
- Web scraping
  - Create a response object from a url page
  - Use Beautiful Soup to extract the table of interest from the object
  - Extract the desired information from each table header and append into a pandas df construct

Data Collection – SpaceX API



View data collection details from SpaceX API in notebook on GitHub

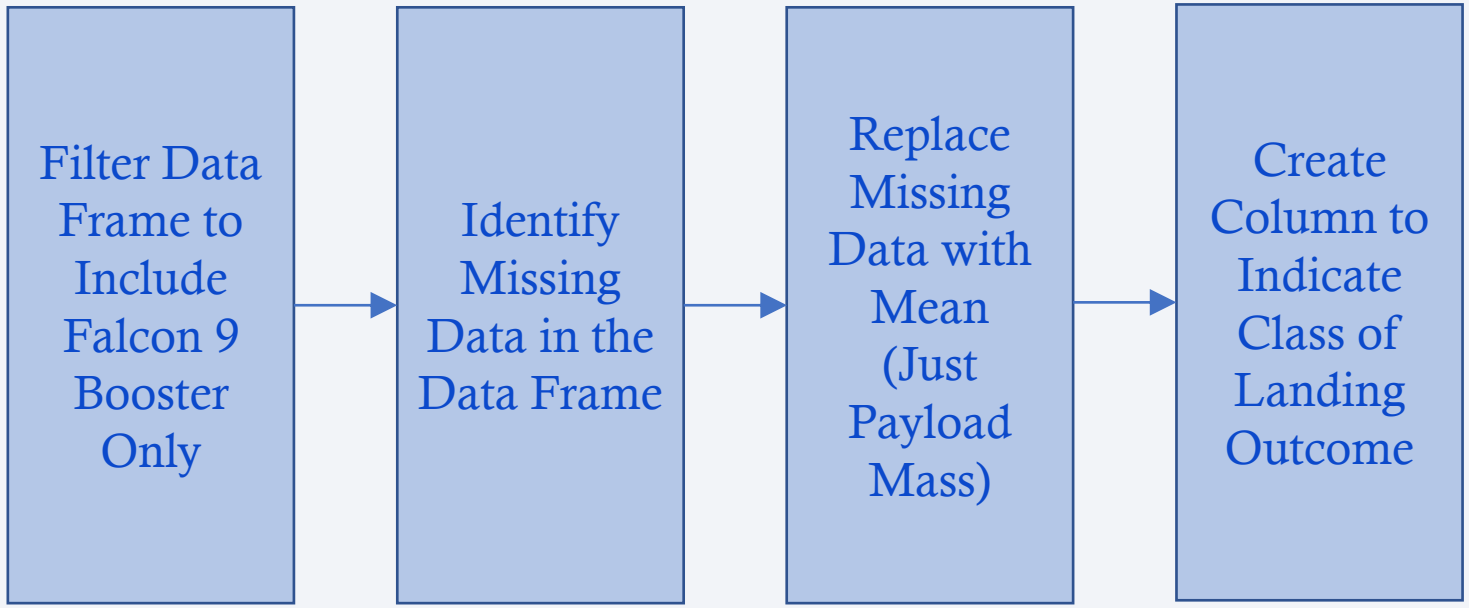
Data Collection - Scraping



View data collection details from SpaceX web scraping in notebook on GitHub

Data Wrangling

- After collection from the API, data was processed as seen in the flowchart below



View SpaceX data wrangling details in notebook on GitHub

EDA with Data Visualization

- Data was visualized multiple ways to gain a preliminary understanding of the data
- Scatter plots were used to visualize the following relationships
  - Payload Mass vs. Flight Number
  - Launch Site vs. Payload Mass
  - Orbit vs. Flight Number
  - Orbit vs. Payload Mass
- Bar chart was used to show Success Rate vs. Orbit
- Line chart was used to show Success Rate per year between 2010 - 2020

View SpaceX data visualization details in notebook on GitHub

EDA with SQL

Queries performed to explore the data were:

- Unique names of the different launch sites
- First five records that begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA CRS
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on a ground pad
- Names of booster carrying between 4000 – 6000 kgs with successful drone ship landings
- Number of successful vs failed mission outcomes

View SpaceX data EDA details in notebook on GitHub

EDA with SQL (continued)

Additional queries performed to explore the data were:

- Names of the boosters carrying maximum payload mass
- Records of month name, landing outcome, booster version, and launch site, where the outcome was a failed drone ship landing in year 2015
- Count of every landing outcome category within a specific date range, ranked in descending order

View SpaceX data EDA details in notebook on GitHub

Build an Interactive Map with Folium

Folium map was created to explore location details. The following objects were added to the map

- Folium circles and markers were added to label and highlight the exact location of each site
- Marker Clusters were added to show the successful and failed outcomes for each site
- Polylines with distance markers were added to show the distance between various landmarks
  - Coastline
  - Train tracks
  - Hwy
  - City

View SpaceX Folium map details in notebook on GitHub

Build a Dashboard with Plotly Dash

Dashboard was created to explore various factors in the data interactively. The following components and graphs were added to the dashboard

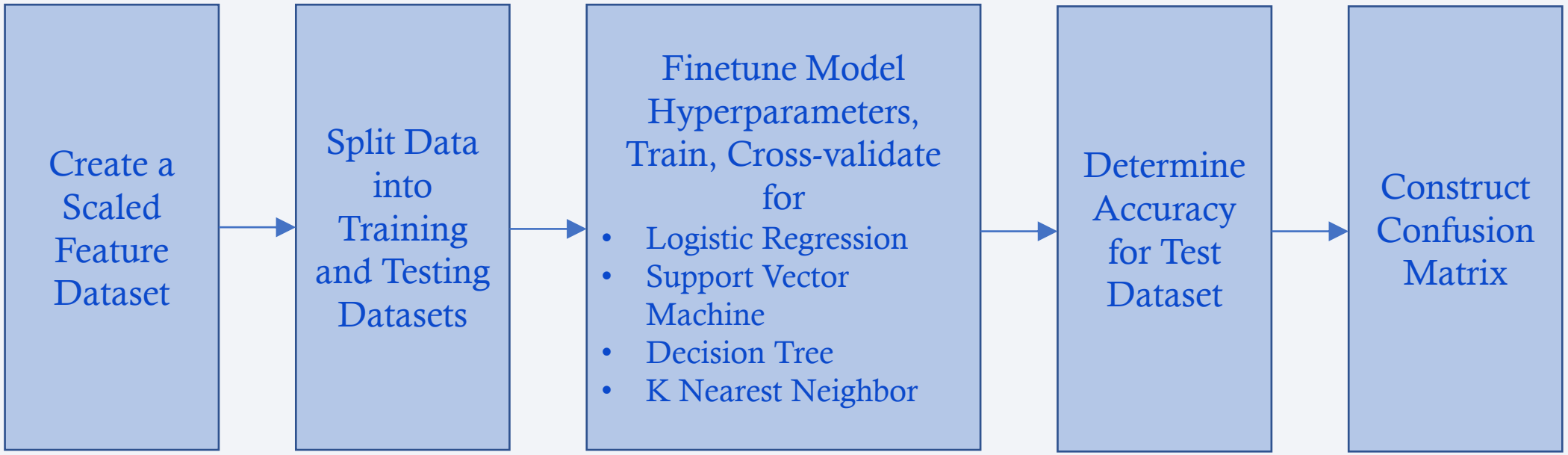
- Dropdown menu allowing user to select all sites or specific launch site
- Pie chart that changes based on the dropdown menu selection
  - If 'All' selected, pie chart shows every launch site with their percentage of successful landings
  - If a specific site is selected, pie chart shows that sites proportion of successful to failed landings
- Range Slider allowing user to adjust the payload range
- Scatter Plot of Class vs. Payload Mass by Booster Version
  - Plot adjust based on both dropdown for site, and the range slider for payload

View SpaceX Dashboard on GitHub

View SpaceX Dash code on GitHub

Predictive Analysis (Classification)

- Prediction Analysis consisted of data preprocessing, model selection, model training, and model evaluation. The flowchart below shows more detailed steps



View SpaceX Machine Learning details in notebook on GitHub

Results

- Landing success improved over time (implies learning from each launch)
- 16% of launches did not attempt to land *(Only used launches after first successful landing)*
  - 73% of all failed landings never attempted to land
- 60% of failed launches – Legs did not deploy
  - ~20% of launches for CCAFS/KSC
  - 10% of launches for VAFB (Though smaller sample - 1 out of 11)

Site

- 50 of 84 (60%) launches took place at CCAFS SLC-40 (66% Success Rate)
  - 25% launches at KSC (77% Success Rate)
  - 15% launches at VAFB (83% Success Rate)
- 77% of all failed landings occurred launching from CCAFS

Results (continued)

Orbit

- 50% of failed launches were of GTO orbit type
- 82% of failed launches were of either GTO or ISS orbit type

Orbit/Launch Site

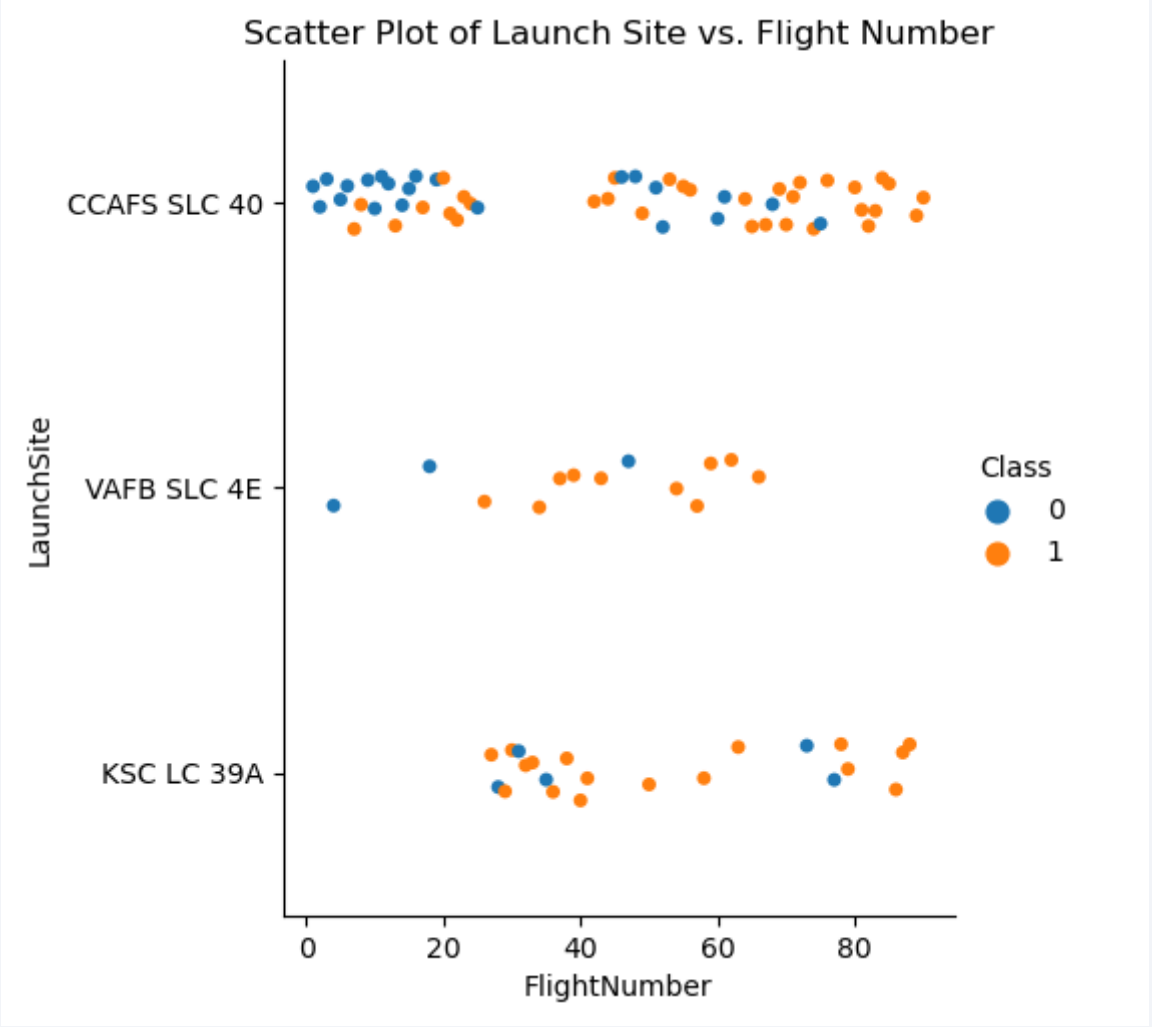
- 90% of failures were either Launched at CCAFS site or destined to GTO orbit

Section 2

Insights drawn from EDA

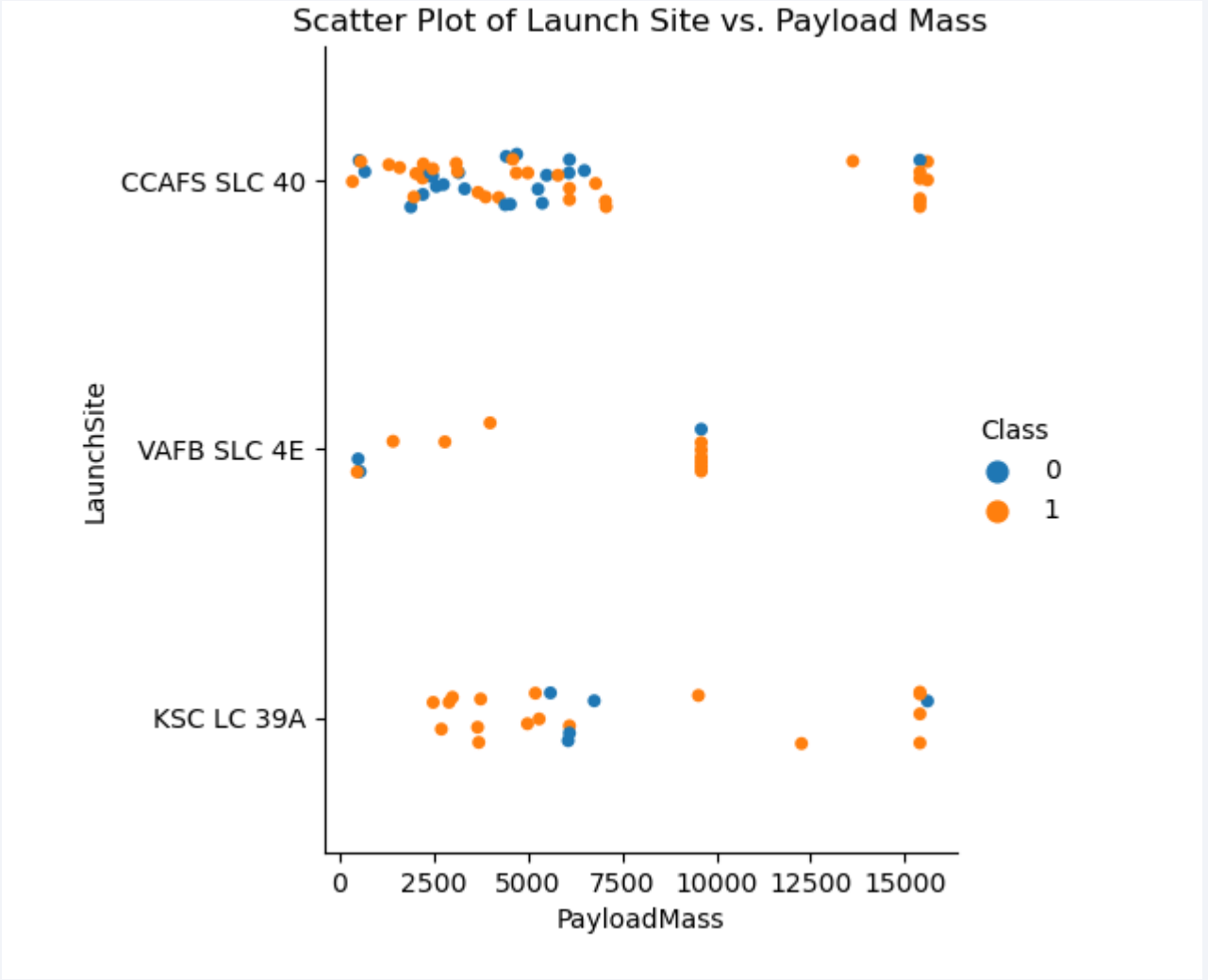
Flight Number vs. Launch Site

- Graph of relationship between launch site and flight number
  - Most of the launches came from CCAFS SLC40
  - Percentage of failures appears to be higher for CCAFS SLC40
  - Most failed landings came from the earliest launches



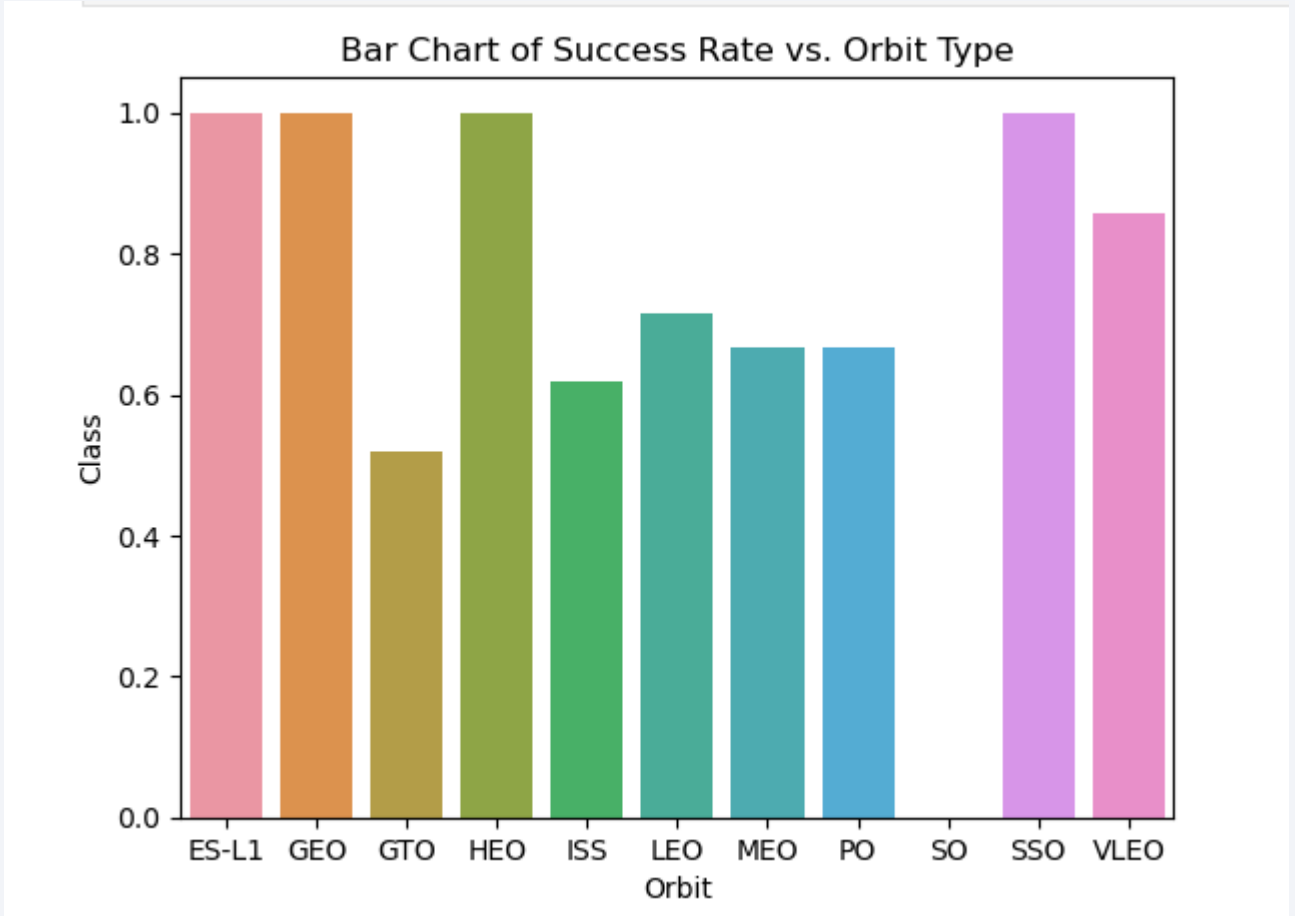
Payload vs. Launch Site

- Graph of relationship between launch site and payload mass
  - Appears to show less failed launches above 8000 kgs



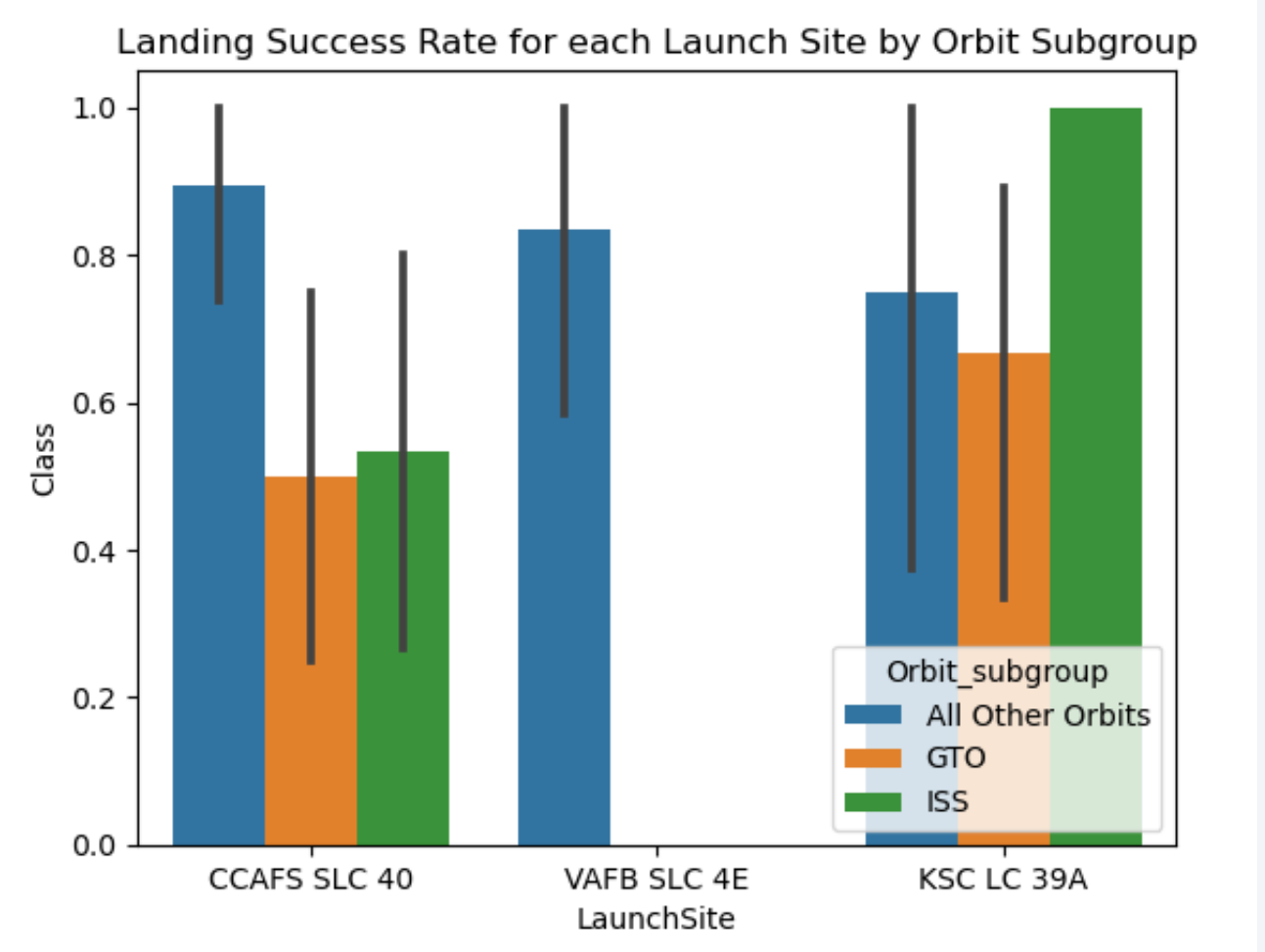
Success Rate vs. Orbit Type

- Graph of success rate for each orbit type
  - Orbit ‘SO’ had no successful launches
  - Orbit ‘ES-L1’, ‘GEO’, ‘HEO’, and ‘SSO’ have 100% success rate



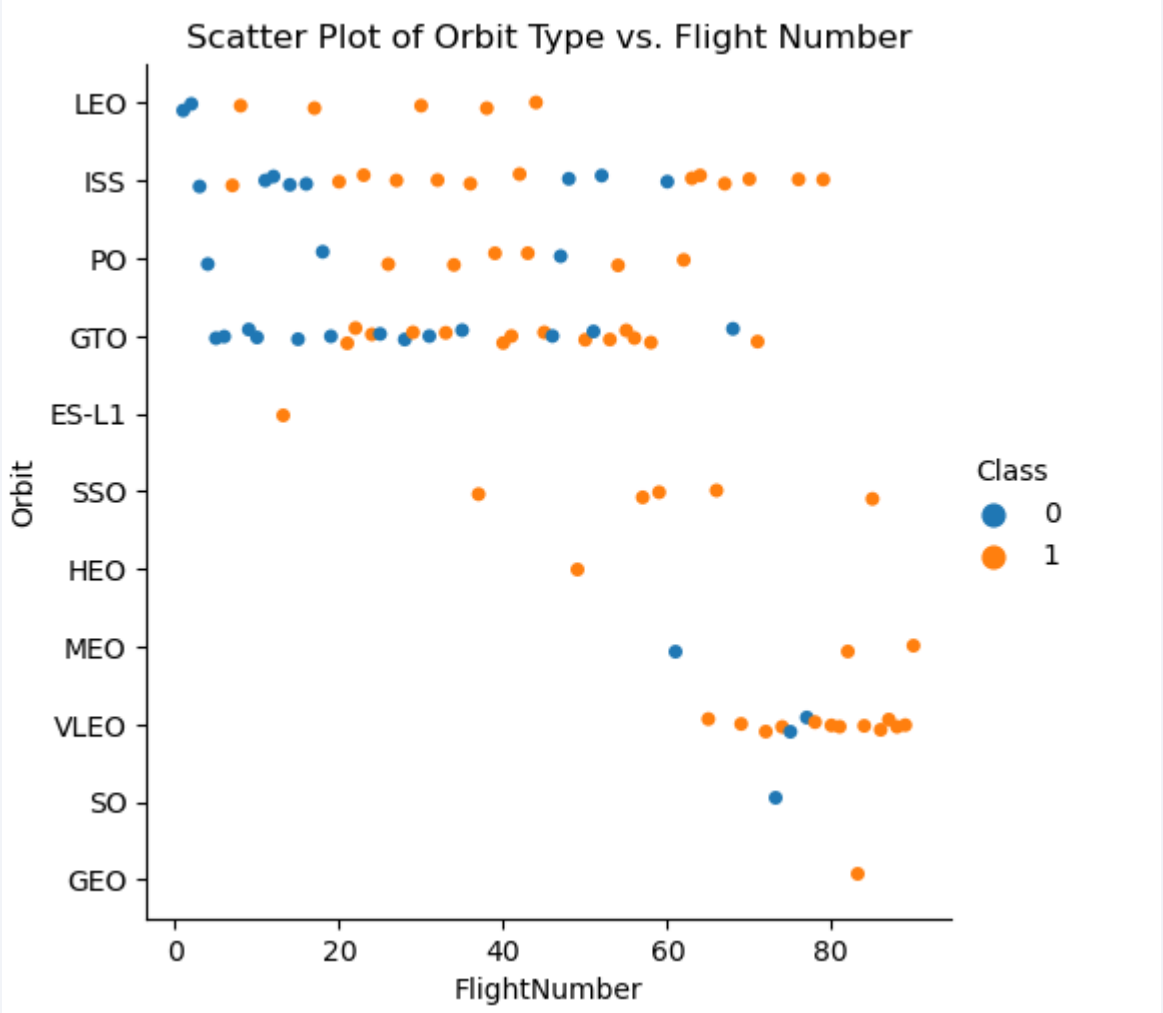
Success Rate vs. Launch Site (by Orbit)

- Graph of success rate for each launch site by orbit
  - CCAFS has the lowest overall landing success rate
  - GTO orbit type has a low success rate regardless of site
  - ISS orbit type has a low success rate for one site but 100% for another



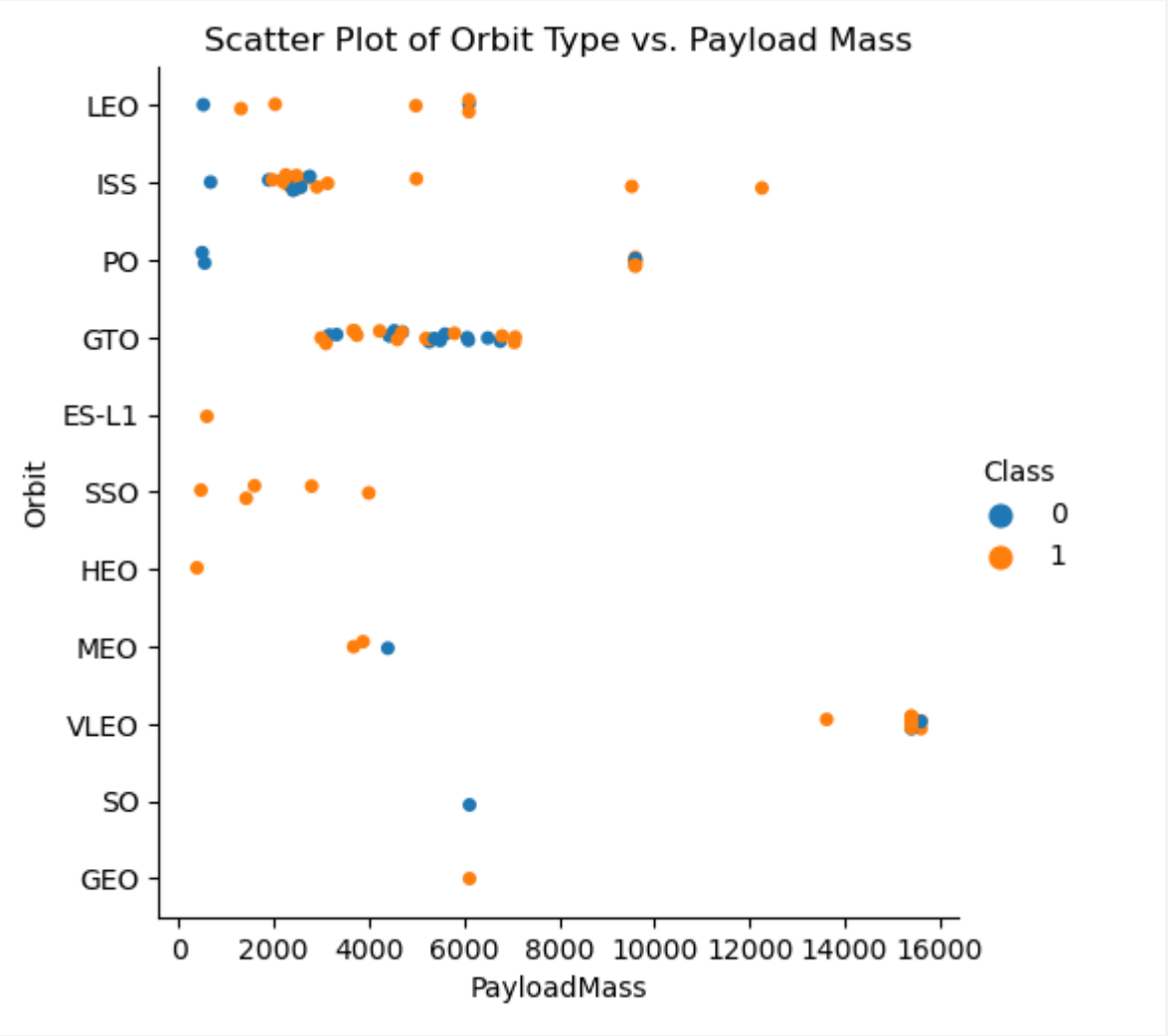
Flight Number vs. Orbit Type

- Graph of relationship between orbit type and flight number
  - ‘ES-L1’, ‘HEO’, ‘SO’, and ‘GEO’ only had 1 launch
  - ‘SSO’ only multi-launch orbit with 100% success rate



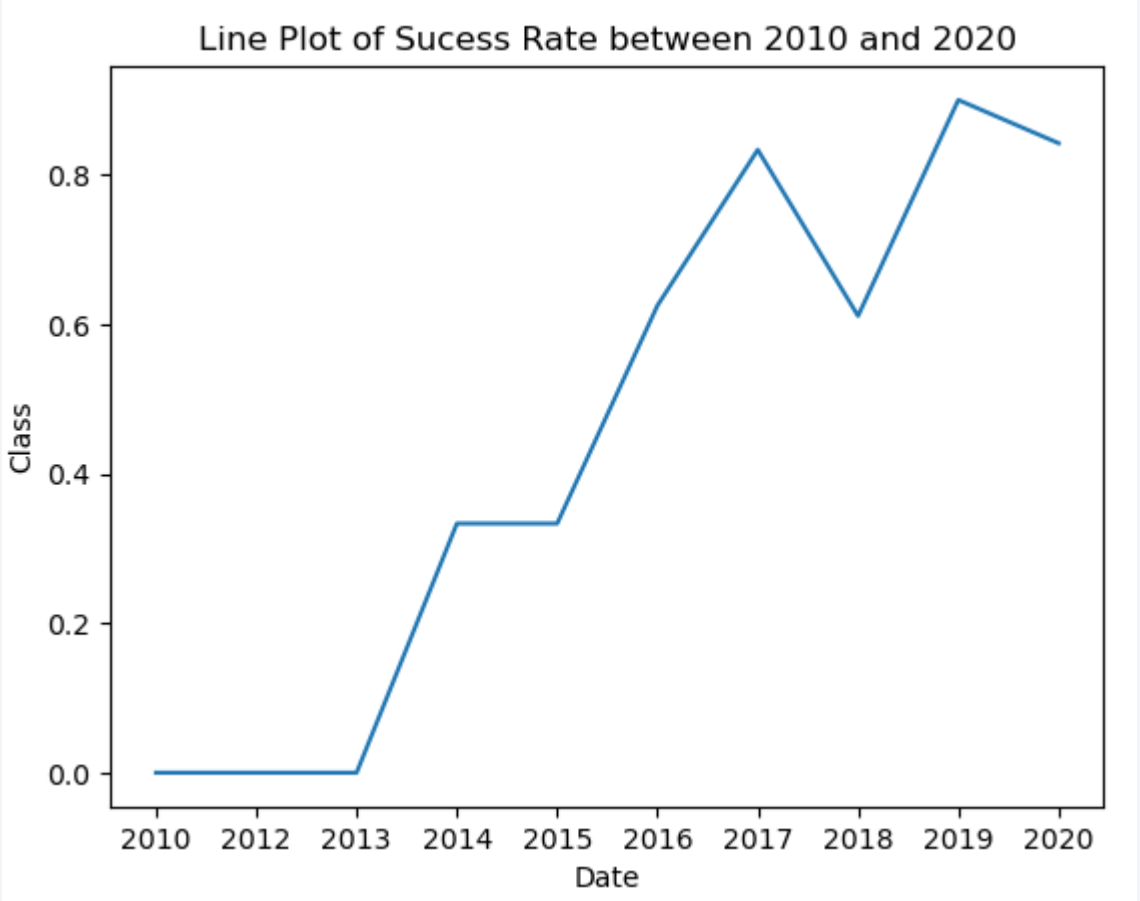
Payload vs. Orbit Type

- Graph of relationship between orbit type and payload mass
  - Majority of launches carried less than 8000 kgs



Launch Success Yearly Trend

- Graph of the launch success rate over time (2010-2020)
  - No successful landings from the first three years
  - Upward trend in successful landings over time
  - 2019 most successful year (~90%)



All Launch Site Names

- Query of the unique SpaceX launch sites

[11]:
```
%%sql
/*
PRAGMA table_info(SPACEXTABLE);
*/

SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE;
```

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

[11]:

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

Launch Site Names Begin with 'CCA'

[11]:
```
%%sql

SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

[11]:

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

- Query of the first five records that begin with ‘CCA’
  - Every record is from the same site (‘CCAFS LC-40’)

Total Payload Mass

- Query of the total payload carried in kgs for all launches between the years 2010 and 2020

```
[12]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER LIKE 'NASA (CRS)%'
```

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

[12]:

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- Query of the average payload carried in kgs for all launches between 2010 and 2020

```
[13]: %%sql
      SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'

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

[13]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Query of the first successful landing date

```
[14]: %%sql
      SELECT MIN(DATE) FROM SPACEXTABLE WHERE MISSION_OUTCOME LIKE 'SUCCESS%'

       * sqlite:///my_data1.db
      Done.
[14]: MIN(DATE)
      2010-06-04
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

[15]:
```
%%sql

SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE
WHERE LANDING_OUTCOME LIKE '%SUCCESS (DRONE SHIP)%'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

[15]:

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

Total Number of Successful and Failure Mission Outcomes

- Query of total number of successful and failed mission outcomes

[16]:
```
%%sql
SELECT
    COUNT(CASE WHEN MISSION_OUTCOME LIKE 'SUCCESS%' THEN 1 END) AS Mission_Success_Count,
    COUNT(CASE WHEN MISSION_OUTCOME LIKE 'FAILURE%' THEN 1 END) AS Mission_Failure_Count
FROM SPACEXTABLE;
```

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

[16]:

Mission_Success_Count	Mission_Failure_Count
100	1

Boosters Carried Maximum Payload

- Query of names of boosters which have carried the maximum payload mass

```
[17]: %%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[17]:

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

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

[18]:
```
%%sql
SELECT CASE
    WHEN strftime('%m', DATE) ='01' THEN 'JANUARY'
    WHEN strftime('%m', DATE) ='02' THEN 'FEBRUARY'
    WHEN strftime('%m', DATE) ='03' THEN 'MARCH'
    WHEN strftime('%m', DATE) ='04' THEN 'APRIL'
    WHEN strftime('%m', DATE) ='05' THEN 'MAY'
    WHEN strftime('%m', DATE) ='06' THEN 'JUNE'
    WHEN strftime('%m', DATE) ='07' THEN 'JULY'
    WHEN strftime('%m', DATE) ='08' THEN 'AUGUST'
    WHEN strftime('%m', DATE) ='09' THEN 'SEPTEMBER'
    WHEN strftime('%m', DATE) ='10' THEN 'OCTOBER'
    WHEN strftime('%m', DATE) ='11' THEN 'NOVEMBER'
    WHEN strftime('%m', DATE) ='12' THEN 'DECEMBER'
    END AS Month_name, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE
WHERE LANDING_OUTCOME LIKE '%FAILURE (DRONE SHIP)%' AND strftime('%Y', DATE) = '2015';
```

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

[18]:

Month_name	Landing_Outcome	Booster_Version	Launch_Site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Query of ranked counts of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
  - No landing attempt had highest count during this period
  - No drone ship failures during this period

```
%%sql
SELECT 'Failure_Parachute' AS OUTCOME,
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%FAILURE (PARACHUTE)%' THEN 1 END) AS Count
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'No_Attempt',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%NO ATTEMPT%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Uncontrolled_Ocean',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE 'UNCONTROLLED (OCEAN)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Controlled_Ocean',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%CONTROLLED (OCEAN)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Failure_Drone_Ship',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%FAILUE (DRONE SHIP)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Precluded_Drone_Ship',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%PRECLUDED (DRONE SHIP)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Success_Ground_Pad',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%SUCCESS (GROUND PAD)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'Success_Drone_Ship',
    COUNT(CASE WHEN LANDING_OUTCOME LIKE '%SUCCESS (DRONE SHIP)%' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'SUCCESS',
    COUNT(CASE WHEN LANDING_OUTCOME = 'SUCCESS' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
UNION ALL

SELECT 'FAILURE',
    COUNT(CASE WHEN LANDING_OUTCOME = 'FAILURE' THEN 1 END)
FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'

ORDER BY Count DESC;
```

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

OUTCOME	Count
No_Attempt	10
Controlled_Ocean	5
Success_Drone_Ship	5
Success_Ground_Pad	3
Failure_Parachute	2
Uncontrolled_Ocean	2
Precluded_Drone_Ship	1
Failure_Drone_Ship	0
SUCCESS	0
FAILURE	0

OUTCOME	Count
No_Attempt	10
Controlled_Ocean	5
Success_Drone_Ship	5
Success_Ground_Pad	3
Failure_Parachute	2
Uncontrolled_Ocean	2
Precluded_Drone_Ship	1
Failure_Drone_Ship	0
SUCCESS	0
FAILURE	0

Section 3

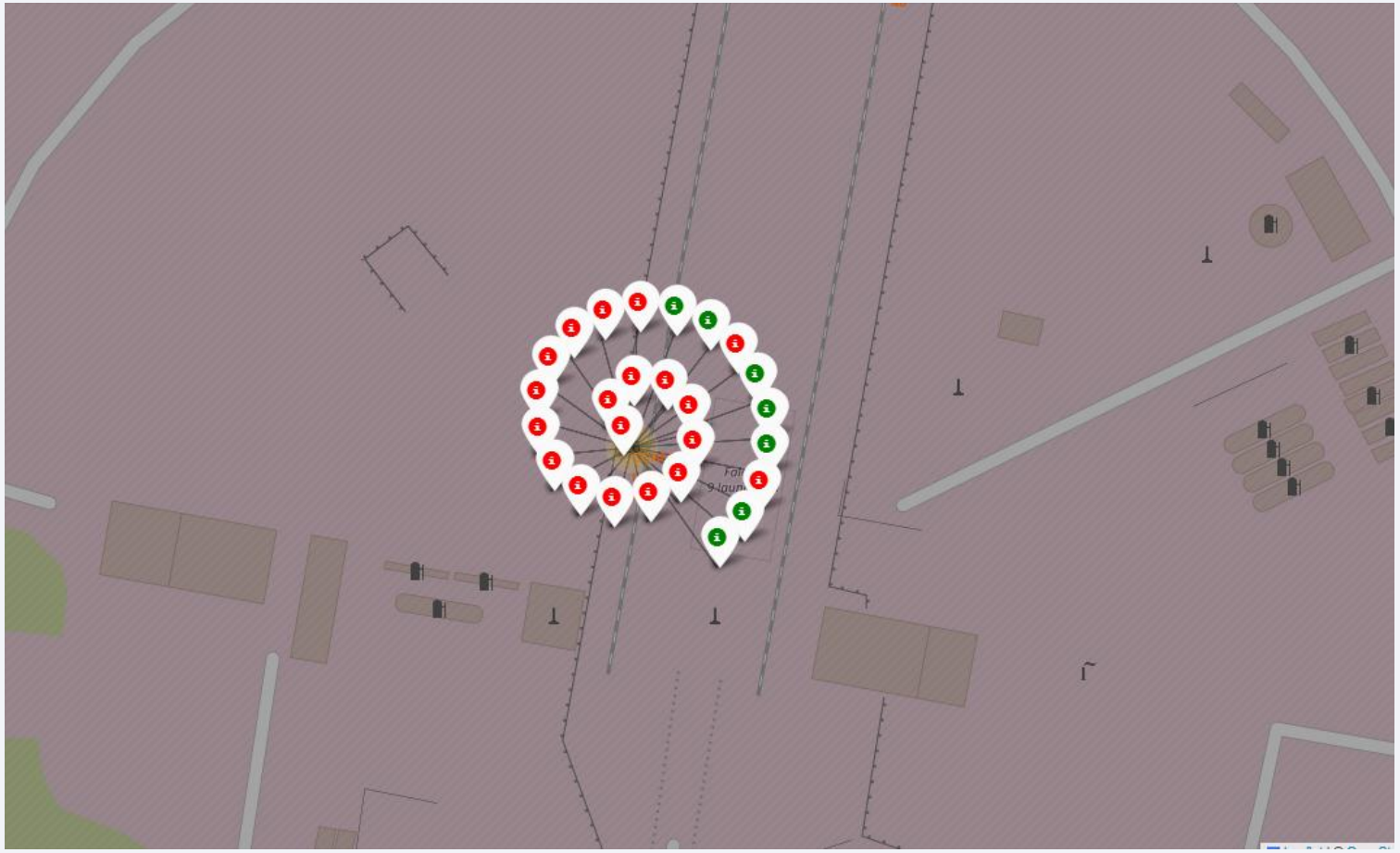
Launch Sites Proximities Analysis

Location of Launch Sites

- Image of markers on a Folium map showing the locations of the Launch Sites
  - Circles are present also and are visible when zoomed in to each location

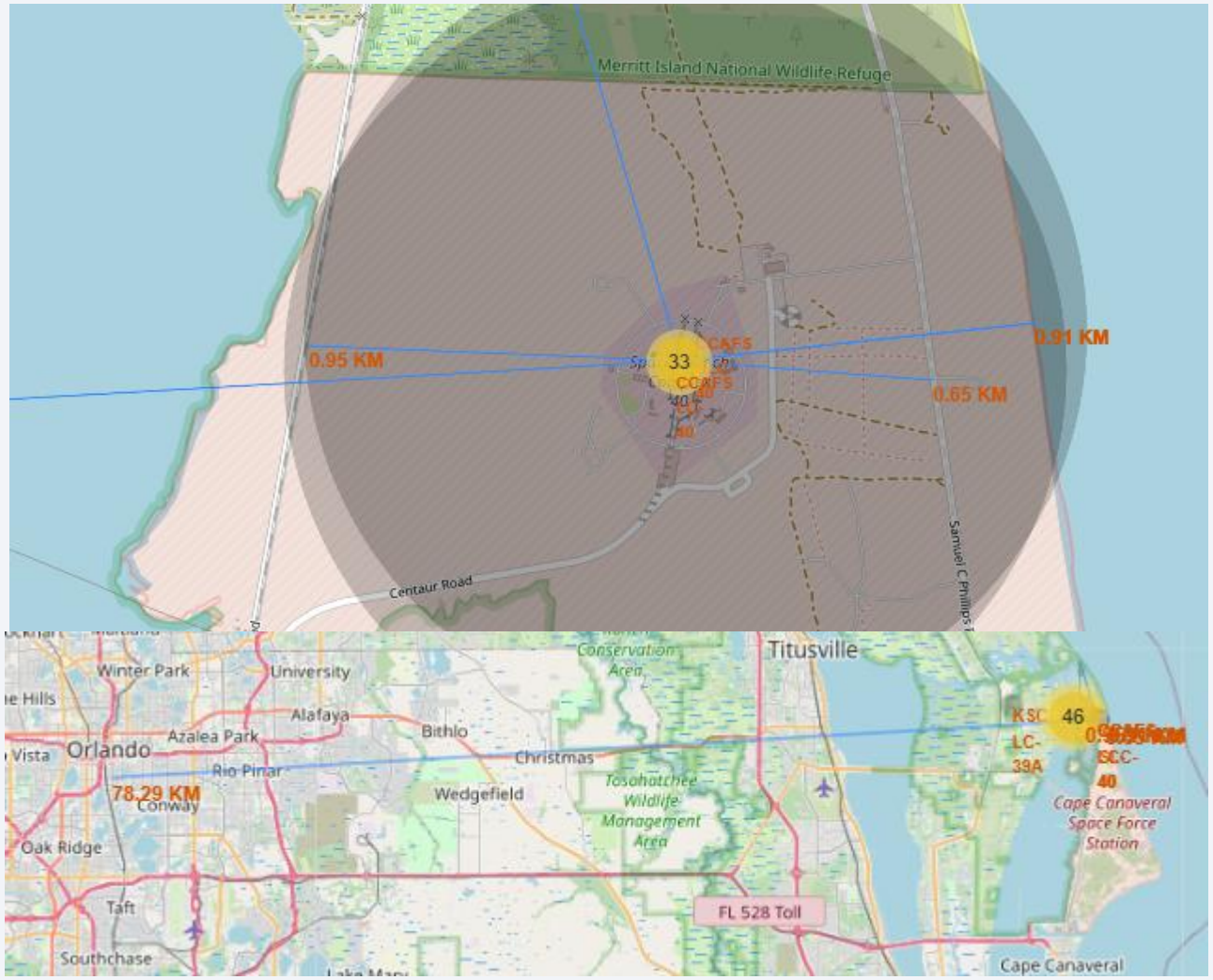
Marker Cluster of Landing Outcomes

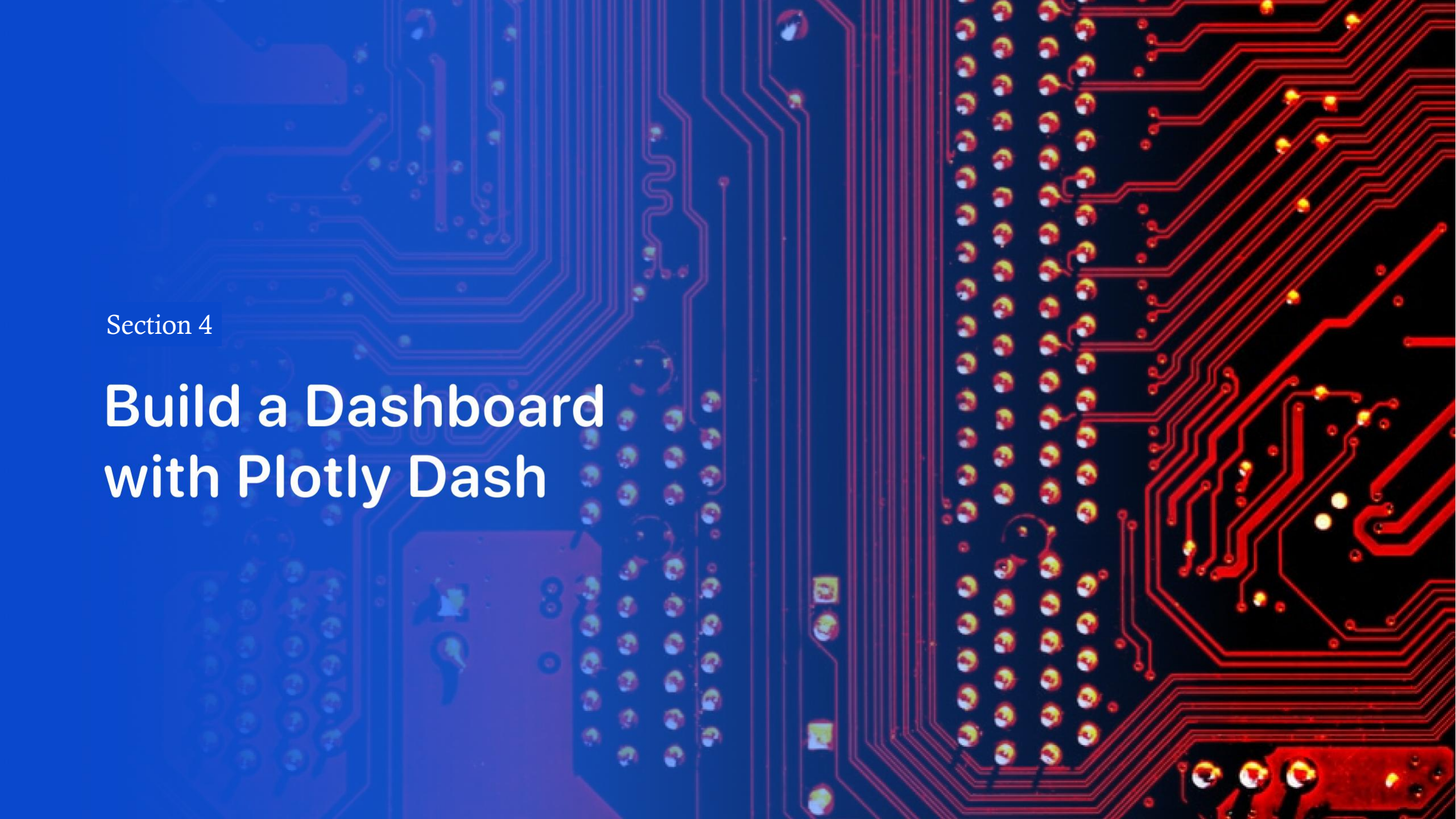
- Image showing the Landing outcomes for one of the launch sites
  - Red – Failed landing
  - Green – Successful landing



Launch Site Distance to Landmarks

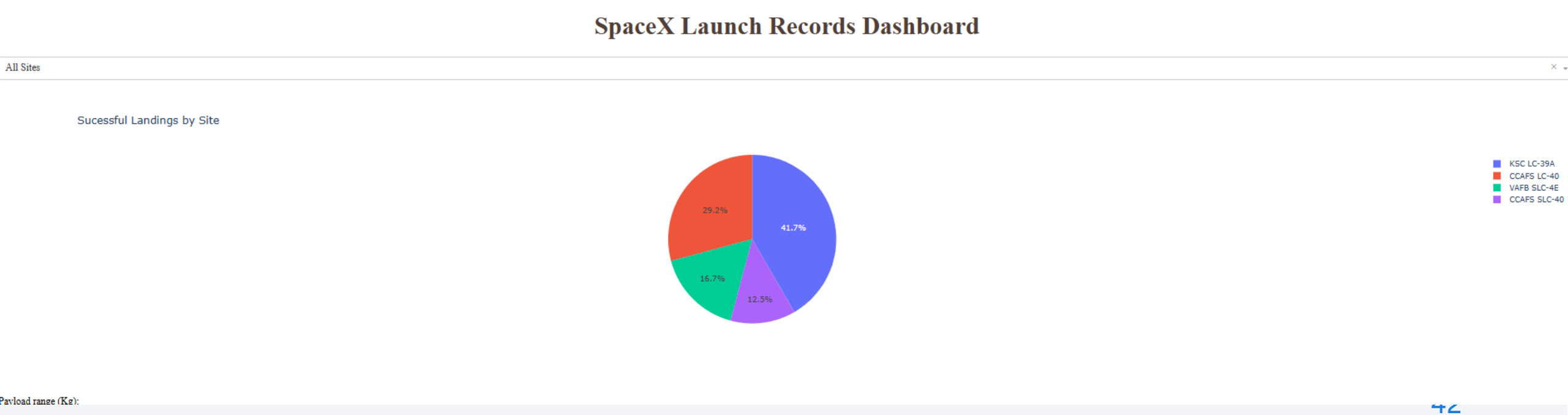
- Image showing a launch site with attached polyline marking distance between coastlines, train tracks, highways, and cities
  - All sites are relatively close to coastlines and train tracks
  - All sites are relatively distance from major cities




Section 4
Build a Dashboard with Plotly Dash

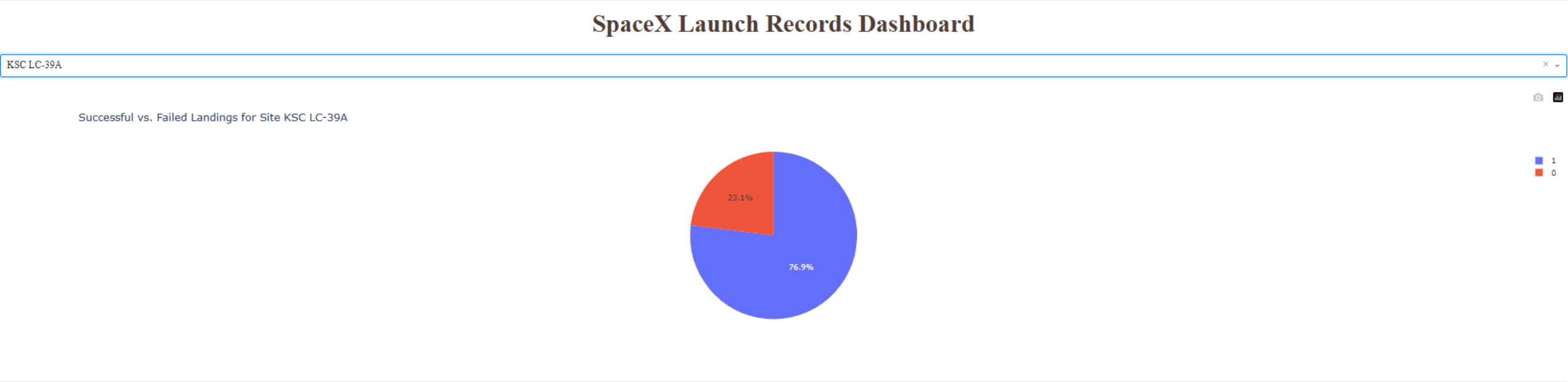
Dashboard – Successful Landing for All Sites

- Image shows the percentage of successful landings across the four launch sites
  - KSC LC-39A has percentage of successful landings with 47%
  - VAFB has the lowest percentage of successful landings with 16%



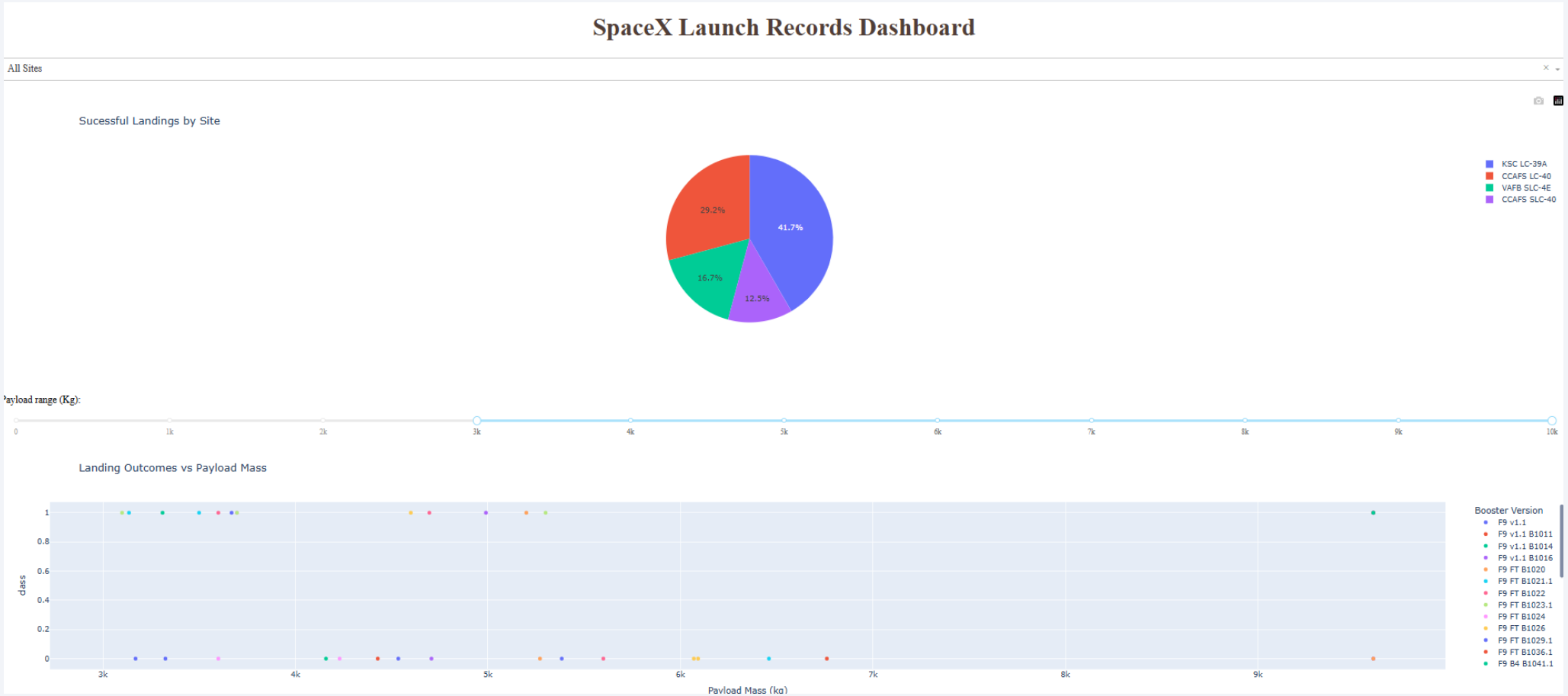
Dashboard – Successful Landings for KSC LC-39A

- Image showing the launch site with the highest percentage of successful landings



Dashboard

- Image showing previously shown pie chart along with scatter plot that shows class vs. payload mass by booster version.

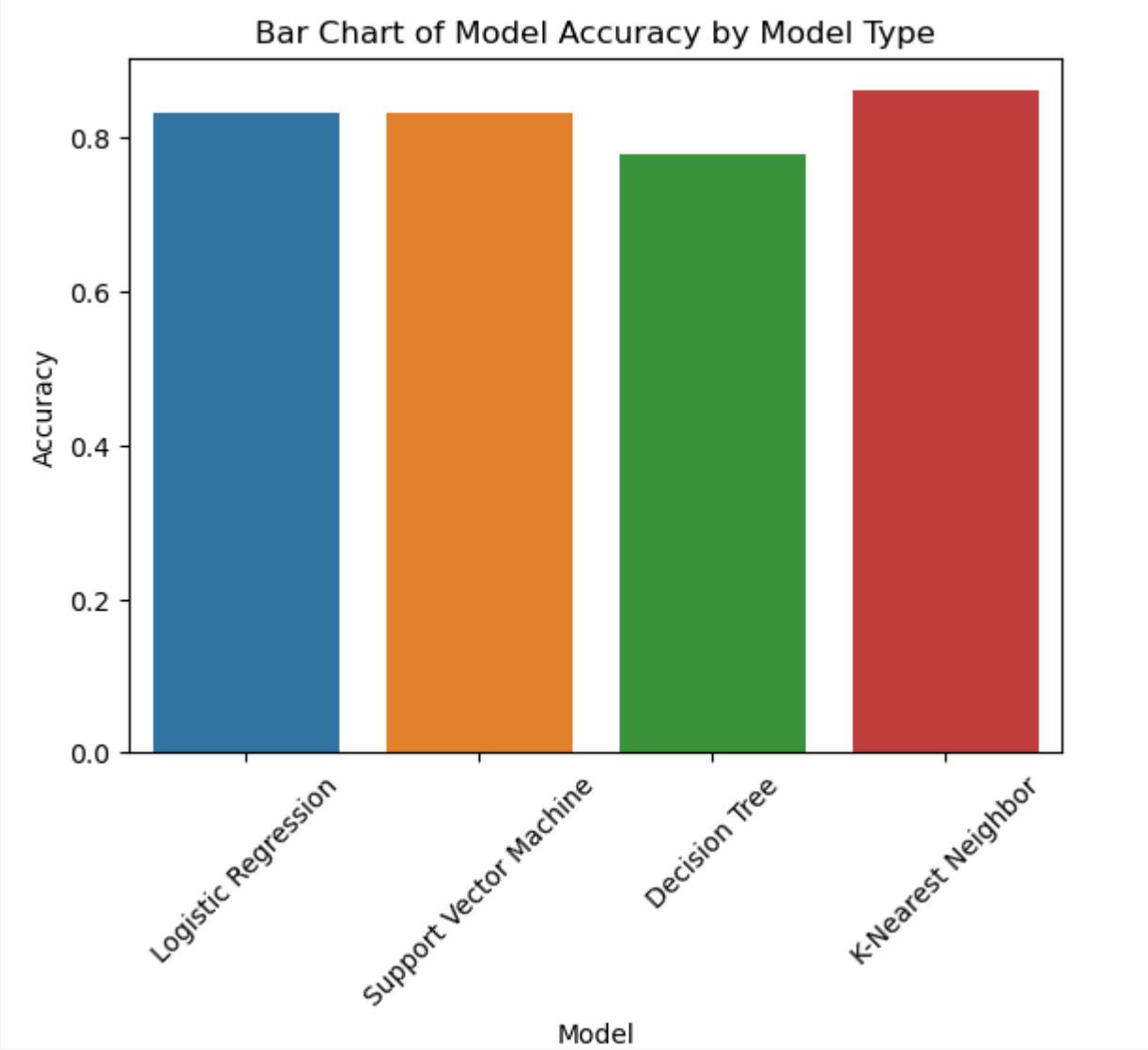


Section 5

Predictive Analysis (Classification)

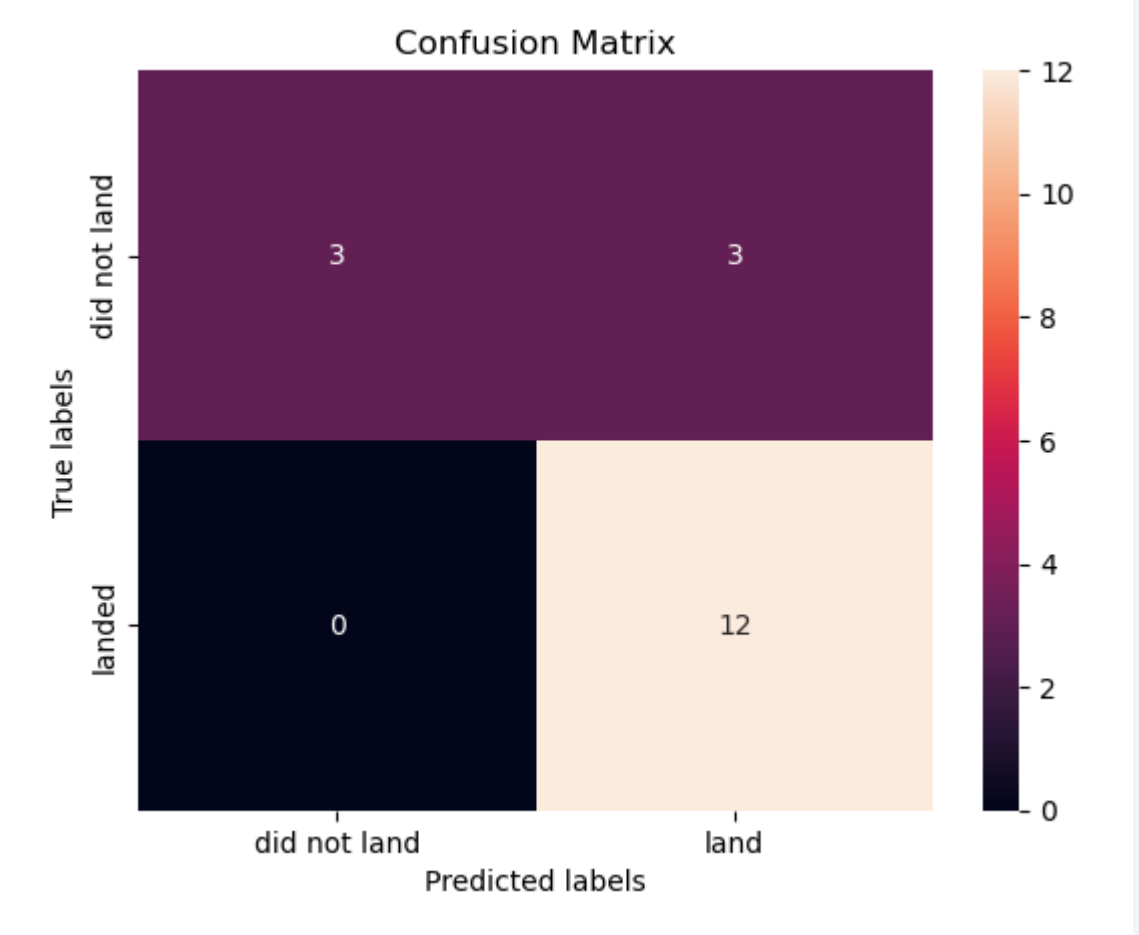
Classification Accuracy

- Graph showing model accuracy for each of the model types
  - K Nearest Neighbor model was most accurate

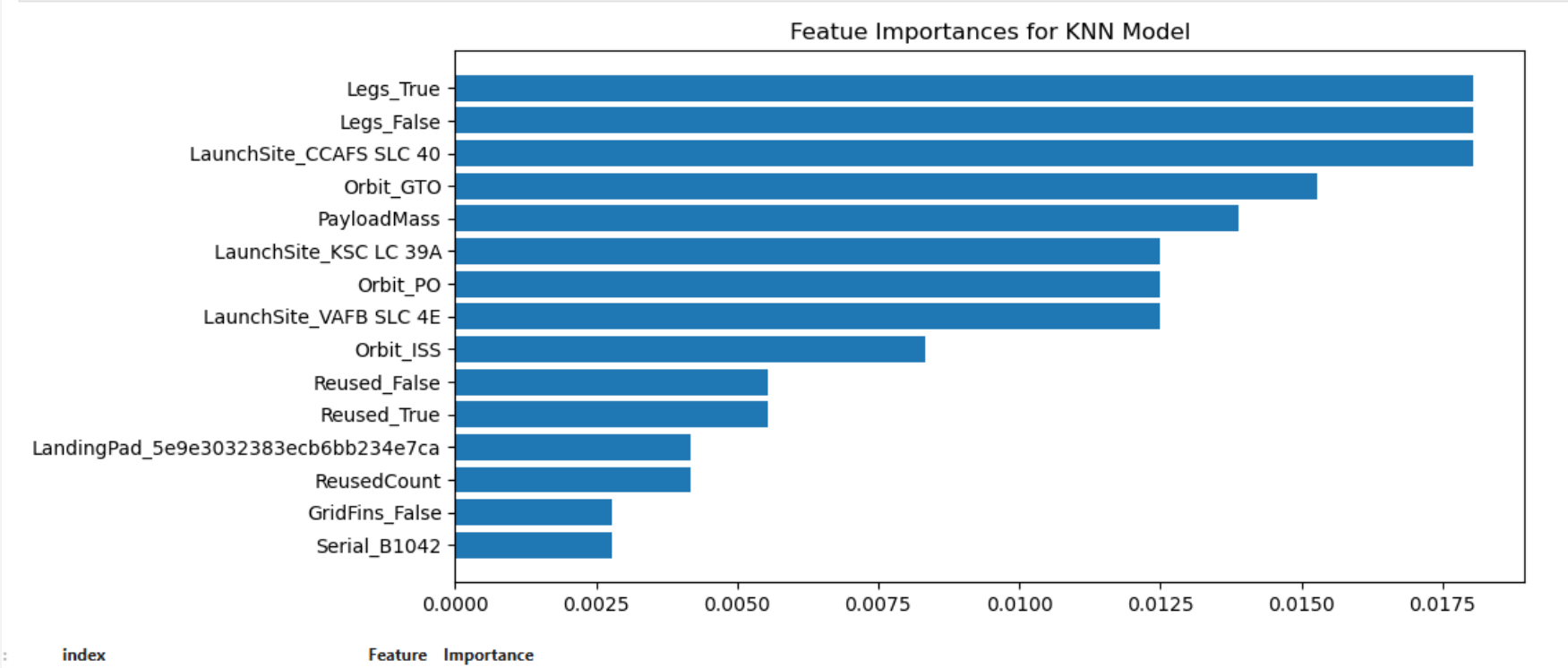


Confusion Matrix

- Confusion matrix of the best model



Feature Importance Graph



- Graph of feature importance based on amount of change in class given the changes in the various features (Top 15 of 82 features)
  - Key features are Leg Deployment, Launch Site, Orbit, and Payload

Conclusions

- The success rate of launches increased over time suggesting iterative learning per launch
- Success rate increases when payload gets higher than 8000 kgs
- Certain orbit types are more reliable than others
  - Low, Very Low, and Sun Synchronous orbits are the most reliable orbits
  - Geostationary transfer orbit is the least reliable orbit
- Launch site plays an important role in whether a first stage lands successfully
  - CCAFS had lowest success rate
  - VAFB had highest success rate
- A Machine Learning model can be used to reliably predict whether Falcon 9 will land or not

Appendix

Approximating Launch Calculation

- Known information:
  - Approximate cost per launch is $62M ($5,580 M total for 90 launches)
  - Number of Boosters used out of 90 launches – 53
    - 53 times, the cost was $A(cost of launch with ability to reuse First Stage)
    - 37 times, the cost was $A + x(cost of launch plus the additional cost replace rocket)

- Approximation Calculation:

53A + 37(A+x) = 90A +37x = $5,580 (where x is the cost to replace a falcon 9 approximated at $67M)

90A + 37($67) = 5580

90A = 3101

A = ~$35M per Launch

Thank you!