IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Max Hamel
9/2/2025

Outline

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

Executive Summary

This research and analysis tries to determine if the first stage rocket will have a successful landing, and the factors involved with doing so. This will help determine the cost of a successful launch. Below is a summary of the methods used to reach the conclusions shown in the rest of this presentation.

Summary of methodologies

- **Data Collection & Wrangling**: Data was collected with the SpaceX REST API and web scraping, then cleaned and formatted to use in the rest of the analysis.
- **Exploratory Analysis**: Initial exploratory analysis was done with SQL and visuals considering factors such as launch sites, payload mass, booster versions and mission outcome
- **Visualizations**: The data was visualized using Folium Lab and Plotly Dash, illustrating launch site locations and outcomes on geographic maps. An Interactive dashboard was created visualizing the percentage of launches for all sites, with payload masses and booster versions.
- **Predictive Analysis**: Machine learning models were trained and tested to predict if the first stage landings are successful. The models used were logistic regression, support vector machines, decision tree classifier and K-nearest neighbor.

Results:

- Over time, the number of successful launches has increased. Launch site KSC LC-39A had the highest success rate. The following orbits also had a 100% success rate: ES-L1, GEO, HEO and SSO.
- With the machine learning models, all 4 performed quite similarly, with the decision tree performing slightly better than the others.

Introduction

Background

SpaceX is a world leader in the space industry, with the goal to make space travel more affordable for everyone. They regularly launch spacecraft into orbit and the International Space Station, along with successfully launching a satellite internet constellation called Starlink, which can be seen in the night sky. One reason for SpaceX's success is the relatively inexpensive cost of their rocket launches compared to other providers, with a Falcon 9 launch costing around $62 million dollars vs. $165+ million dollars from competitors. SpaceX can offer such a steep discount because they can reuse the first stage rockets. This analysis seeks to predict the success of the first stage landing and therefore aid in determining the cost of launches.

Goals

- Determine how factors such as payload mass, launch site, number of flights, landing type and orbits impact the landing success
- Calculate the rate of successful landings over time
- Train and test various machine learning models to find the best model to use to predict successful landings

Section 1

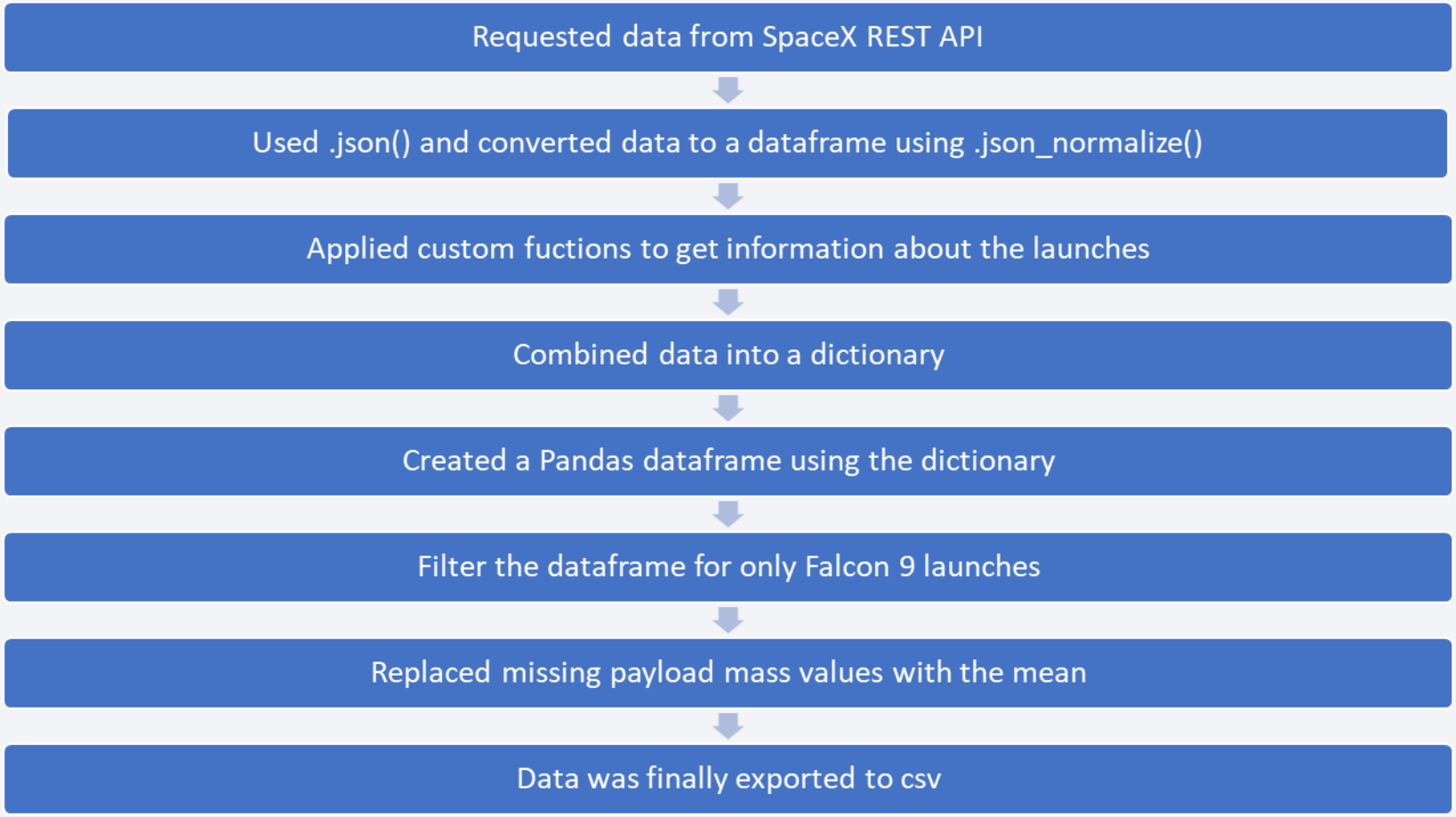
Methodology

Methodology

Executive Summary

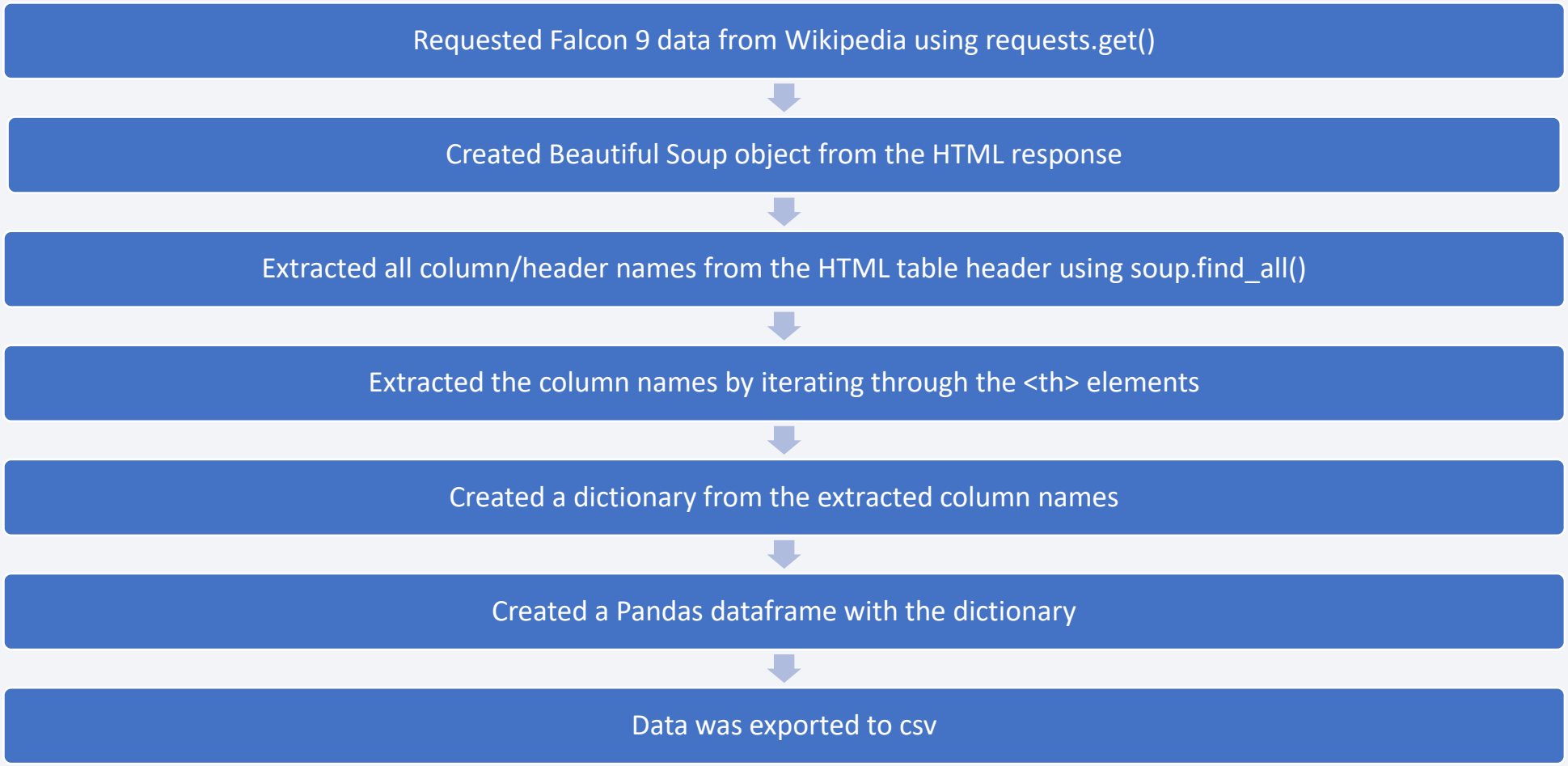
- Data collection methodology:
  - Completed using the SpaceX REST API and web scraping
- Perform data wrangling
  - After gathering the data, it was cleaned, filtered for Falcon 9 launches and placed into a Pandas dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four models were built, trained and tested to predict future landing outcomes

Data Collection – SpaceX API Overview



https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping Overview



https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping-bak-2025-09-02-19-42-04Z%20(1).ipynb

Data Wrangling

Loaded dataset from previous steps using pd.read_csv(), Identified missing values and which columns are numerical or categorical

Calculated # of launches at each site, # of orbit types and occurences, # of mission outcomes and occurences outcomes for each orbit type using the method value._counts()

Created a bad_outcomes set using the following types of landing outcomes: False ASDS, False Ocean, False RTLS, None ASDS and None None

Created a landing outcome label from the outcome column and bad_outcomes set using a loop with an if/else statement, the results were then assigned to the variable landing_class, column name Class. 1 = successful outcome 0 = bad outcome

df["Class"].mean() was then used to calculate a success rate of roughly 66.67%

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts

Scatter Point

- Help visualize the relationships between the X and Y variables. Identifying relationships between variables could useful for machine learning models
  - Flight Number and Payload Mass
  - Flight Number and Launch Site
  - Payload Mass and Launch Site

Bar Chart

- Shows comparison between categories and a value
  - Orbit Type and Success Rate

Line Chart

- Displays the change in a variable over a period of time
- Year and Success Rate

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

Queries

- Displayed list of unique launch site names
- Displayed 5 records where the launch site began with “CCA”
- Summed the total payload mass carried by NASA launched boosters
- Calculated average payload mass carried by F9 v1.1 boosters
- Listed the date of the first successful ground pad landing
- Listed the boosters that have had successful drone ship landings with a payload mass between 4000 and 6000
- Listed each mission outcome and the number of occurrences for each
- Displayed all booster versions that have carried the maximum payload mass
- Listed records along with month names for each month in the year 2015
- Ranked count of landing outcomes between June 4th 2010 and March 20th 2017 in descending order

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Started by marking NASA Johnson Space Center's coordinates on the map with a pop label showing its name and a blue circle
- Added markers for all launch sites with red circles and a popup label
- Launch outcome markers were added for the launch sites color coded green for successful outcomes and red for unsuccessful outcomes
- Distance from CCAFS SLC-40 and its closest coastline, railway, highway and city were calculated, and lines were added to the map.

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/Interactive%20visual%20analytics%20with%20folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown input component for the launch sites to allow users to select all or a specific site to filter the dashboard by
- Created a pie chart with a callback function based on the site selected in the dropdown, that displays successful and unsuccessful launches as a percent of total
- Added a range slider of payload mass between 1000 and 10,000 kg. Filters the other visuals as well
- Created a scatter chart with a callback function displaying payload mass vs. success rate, based on the launch site selected

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/plotly%20dash%20code.txt

Predictive Analysis (Classification)

- Created a Numpy array from the Class column by applying to_numpy()
- Standardized the data with a transform StandardScaler
- Split the data into training and test data using train_test_split
- Created a GridSearchCV object to find the best parameters
- GridSearchCV was then applied to each of the models, LogReg, Support Vector Machine, Decision Tree and K Nearest Neighbor
- Accuracy was calculated for each model type and plotted on a confusion matrix for each
- The best model was identified comparing the results of these matrices

https://github.com/Max45848/IBM-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

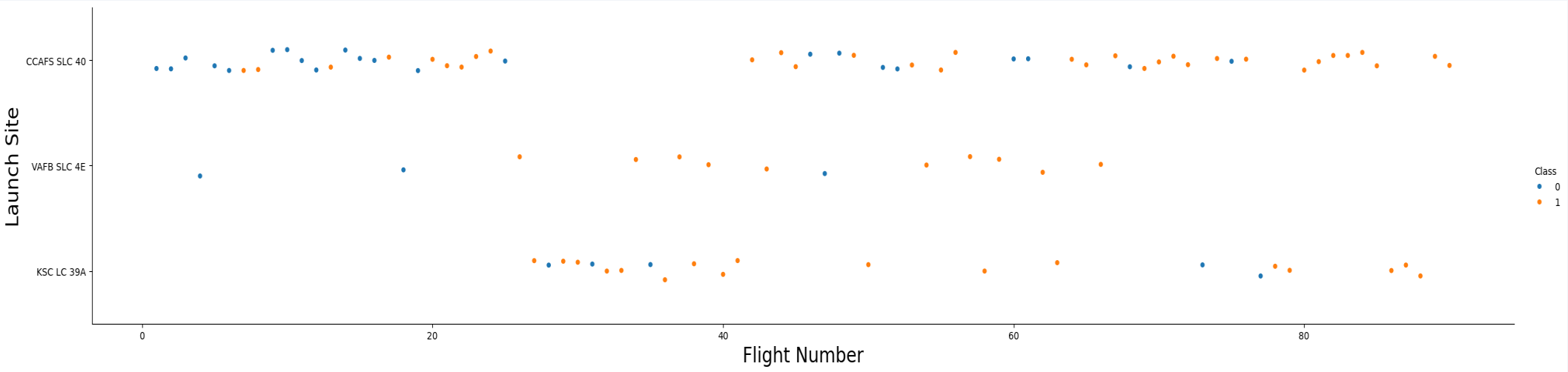
- Exploratory data analysis results
  - Launch success has improved over time
  - The site with the highest success rate was found to be KSC LC-39A
  - Orbits ES-L1, GEO, HEO and SSO had 100% success rates
- Interactive analytics demo in screenshots
  - Launch sites are located near the equator to take advantage of the Earth's rotational speed. They are near the coastline and far enough from population centers to reduce change of damage in the event of failed launches
- Predictive analysis results
  - All four models had similar results, however the decision tree was the best model overall

Section 2

Insights drawn from EDA

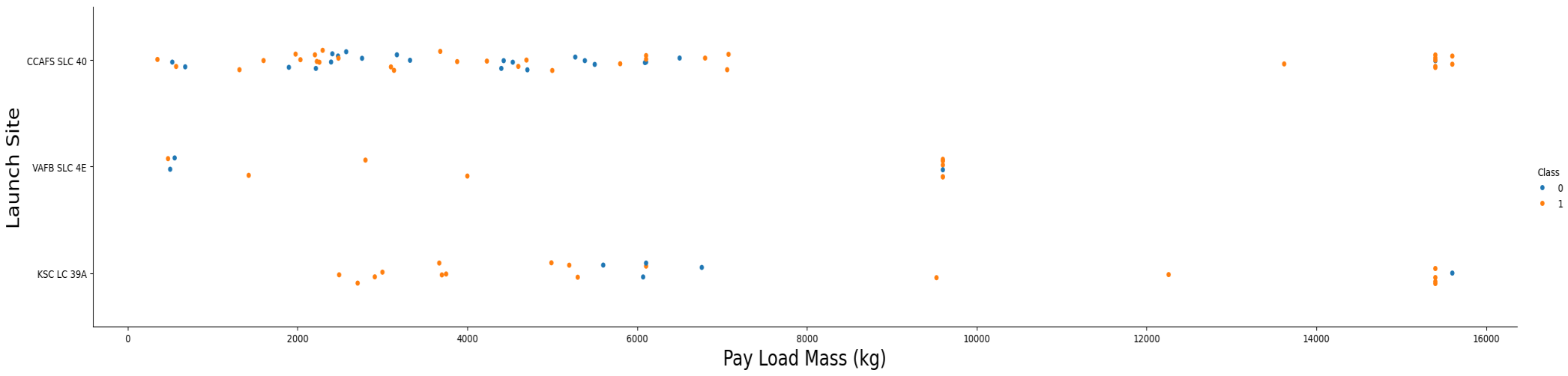
Flight Number vs. Launch Site

- Successful launches increased over time
- Most launches were from CCAFS SLC 40



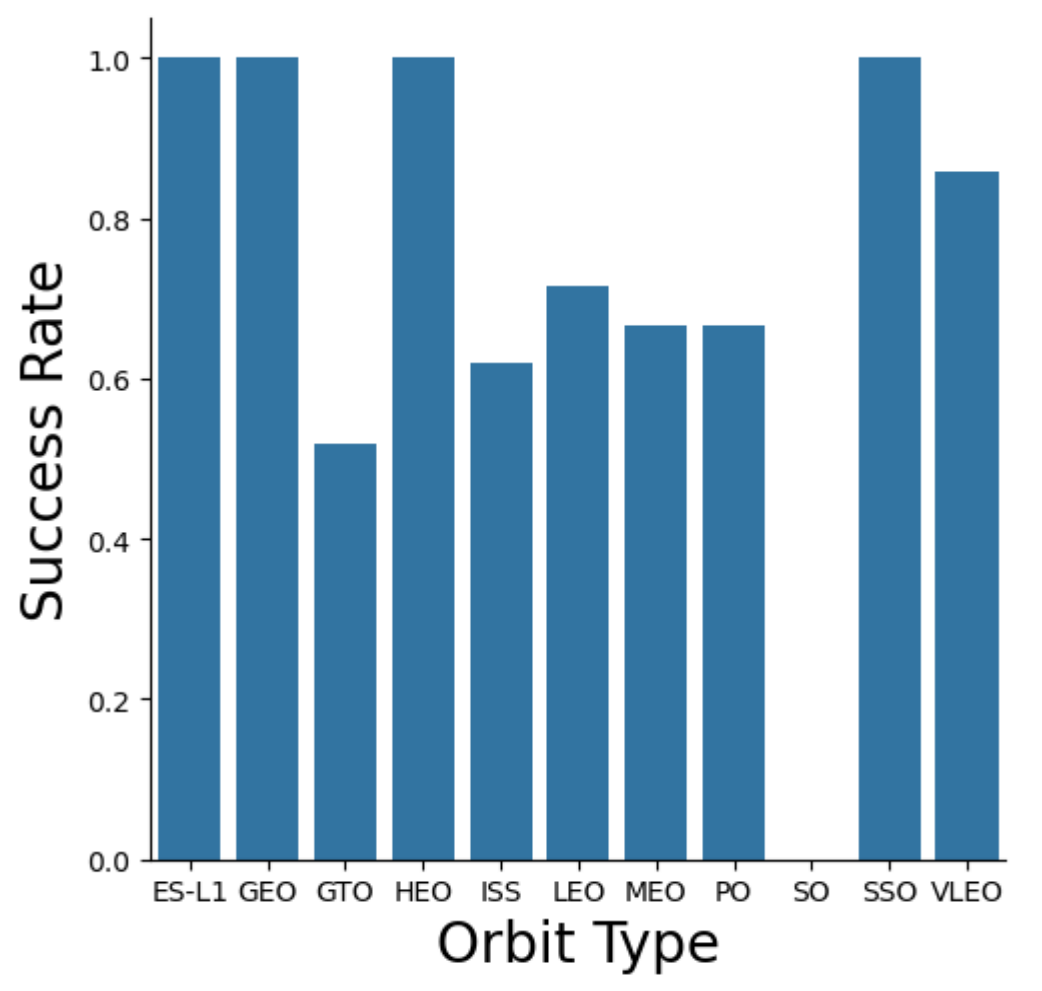
Payload vs. Launch Site

- Success rate appears to increase with payload mass
- The majority of launches were below 8000 kg
- VAFB SLC 4E had no launches over 10,000 kg



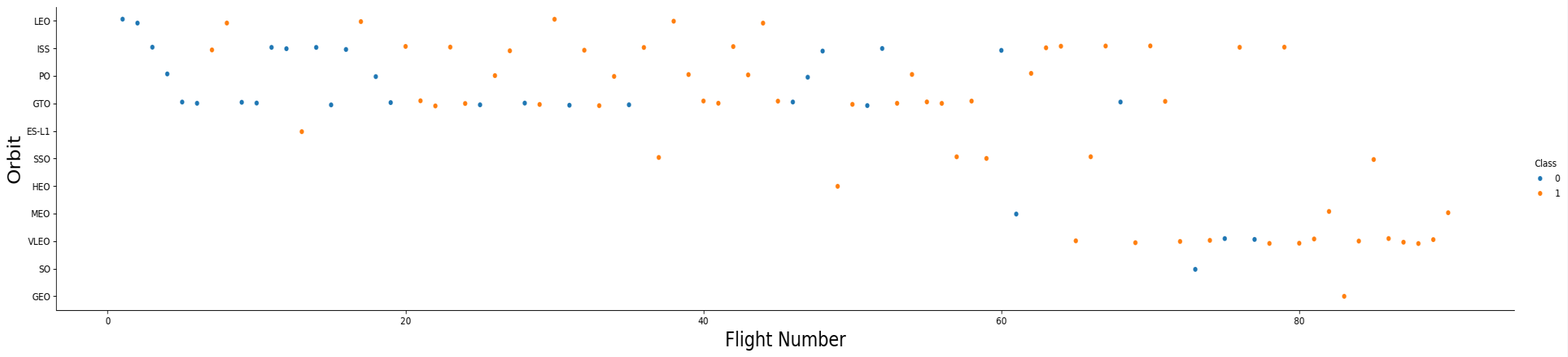
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO had 100% success rates
- Most other orbits had around 60-80% success rates
- GTO was below 60%, where as SO had a 0% success rate



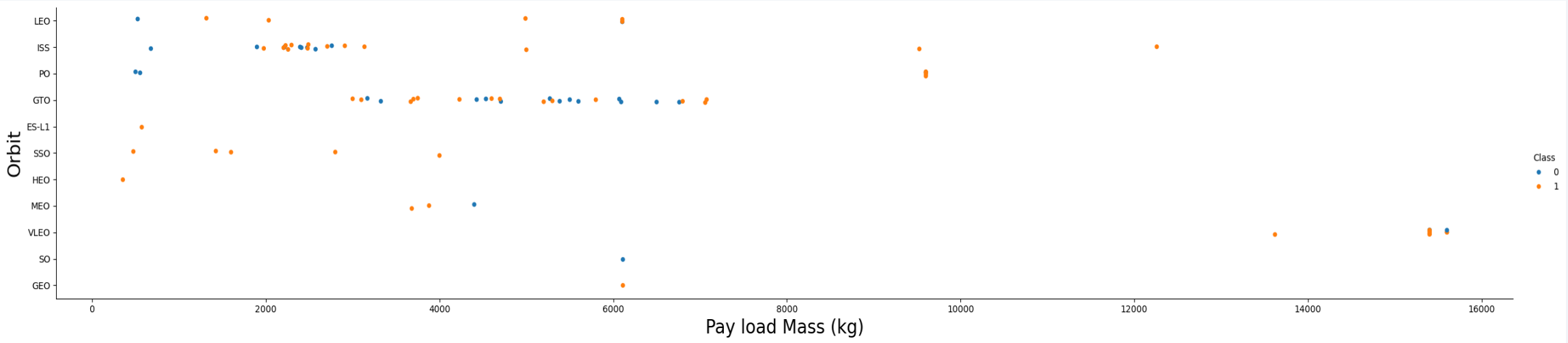
Flight Number vs. Orbit Type

- Successful orbits appear to increase over time
- Additional orbit types appear over time as well, especially around flight number 60



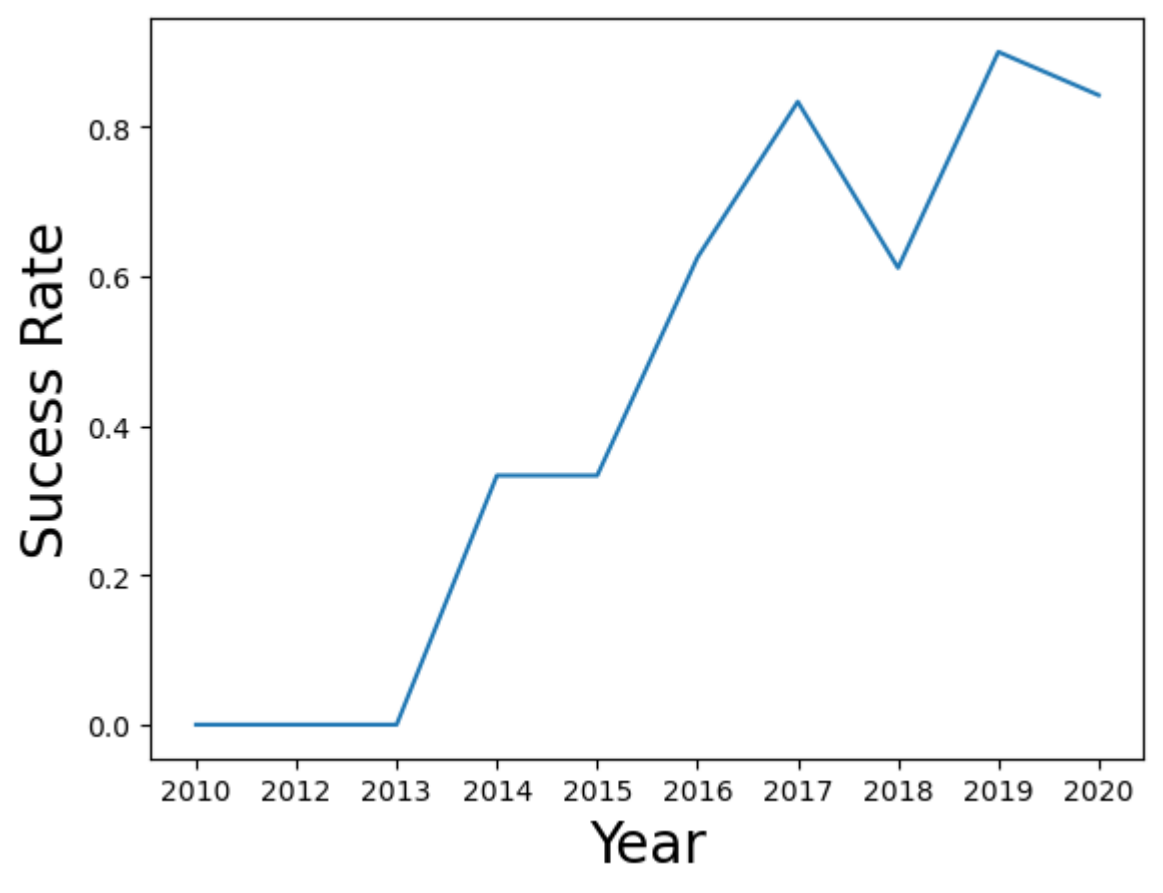
Payload vs. Orbit Type

- Certain payload masses favor different orbit types
- Such as GTO from roughly 3000 to 7000 kg and the ISS from 2000 to under 4000 kg



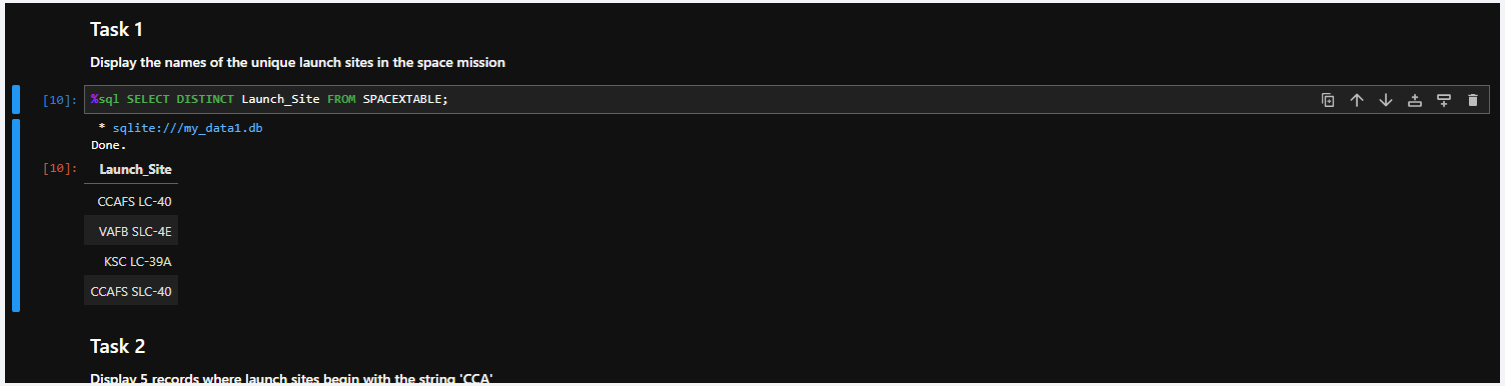
Launch Success Yearly Trend

- Launch success rate saw a sharp increase from 2013 to about 2017
- From 2017-18 a temporary decrease in success rate occurred before leveling back out



All Launch Site Names

- Launch Sites
  - All distinct names from database table



Task 1

Display the names of the unique launch sites in the space mission

```
[10]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

[10]:

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

Task 2

Display 5 records where launch sites begin with the string 'CCA'

Launch Site Names Begin with 'CCA'

- Displayed first 5 rows where launch site starts with CCA

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
[11]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

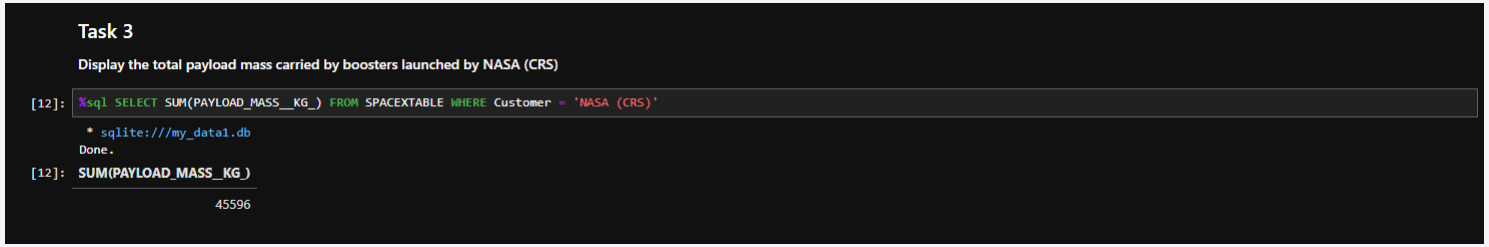
```
 * sqlite:///my_data1.db
Done.
```

[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

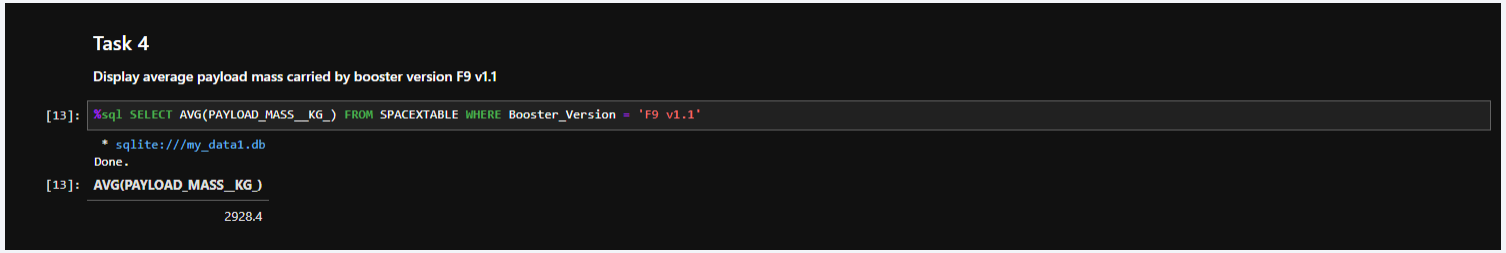
Total Payload Mass

- Summed payload mass where customer = NASA CRS
  - 45,596 kg



Average Payload Mass by F9 v1.1

- Average payload mass = 2928.40 kg



Task 4

Display average payload mass carried by booster version F9 v1.1

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

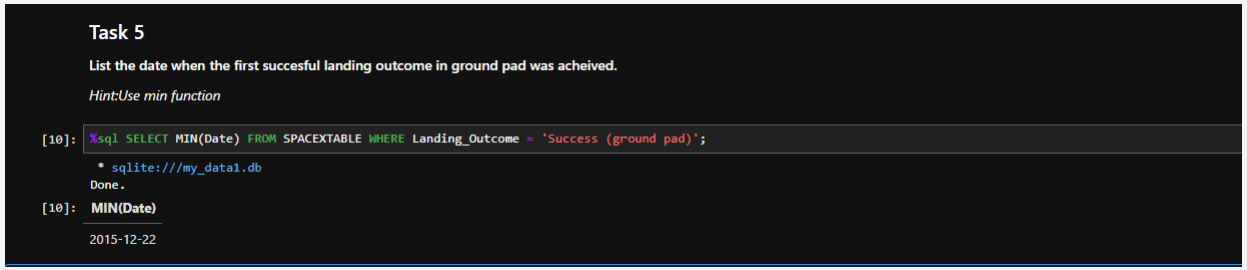
```
 * sqlite:///my_data1.db
Done.
```

[13]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First ground pad success was December 22nd, 2015



Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters with successful drone ship landings with payloads between 4000 and 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- By far most missions result in success based on mission outcome

Task 7

List the total number of successful and failure mission outcomes

```
[16]: %sql SELECT COUNT(Mission_Outcome), Mission_Outcome FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

[16]:

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

Boosters Carried Maximum Payload

- Filtered the SPACEXTABLE for only payload masses that equal the maximum amount in the table

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
[17]: %sql SELECT 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

- There were failed drone ship landings in January and April of 2015

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
[18]: %sql SELECT substr(Date,6,2) AS Month, Date, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015';
```

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

[18]:

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

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

- No attempt had the highest occurrence rate, with second place tied with drone ship successes and failures

Task 10

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.

```
[19]: %sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
```

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

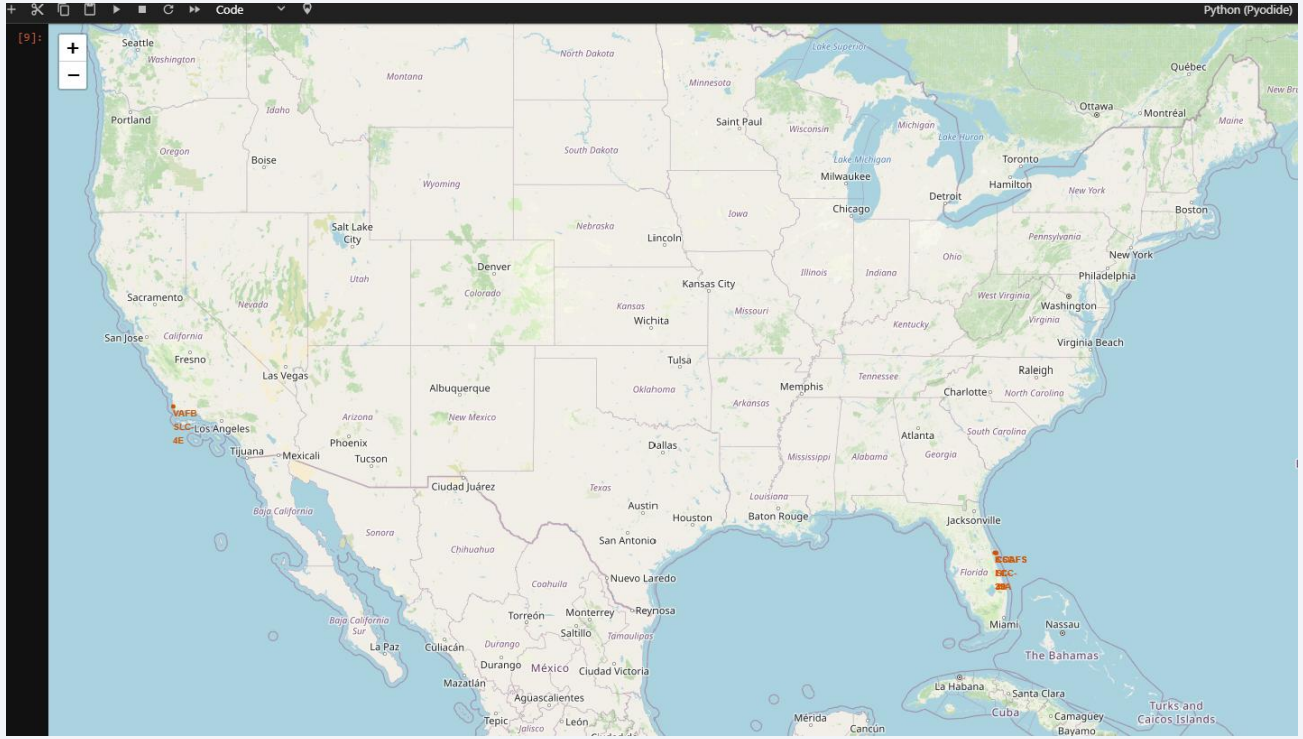
[19]:

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


Section 3
Launch Sites
Proximities Analysis

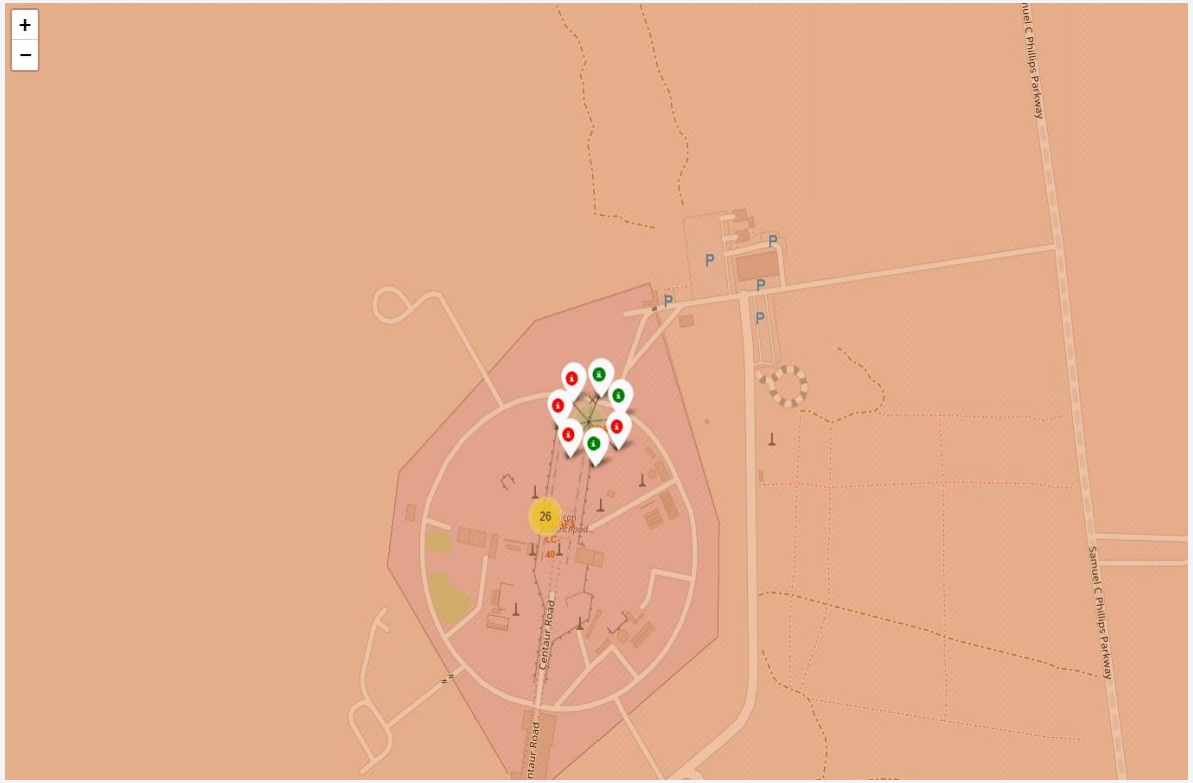
Launch Sites with Markers

- Markers with labels were placed on all launch sites
- Launch sites tend to be near the equator to use the Earth's rotational speed for additional speed while exiting the atmosphere, which therefore can save costs



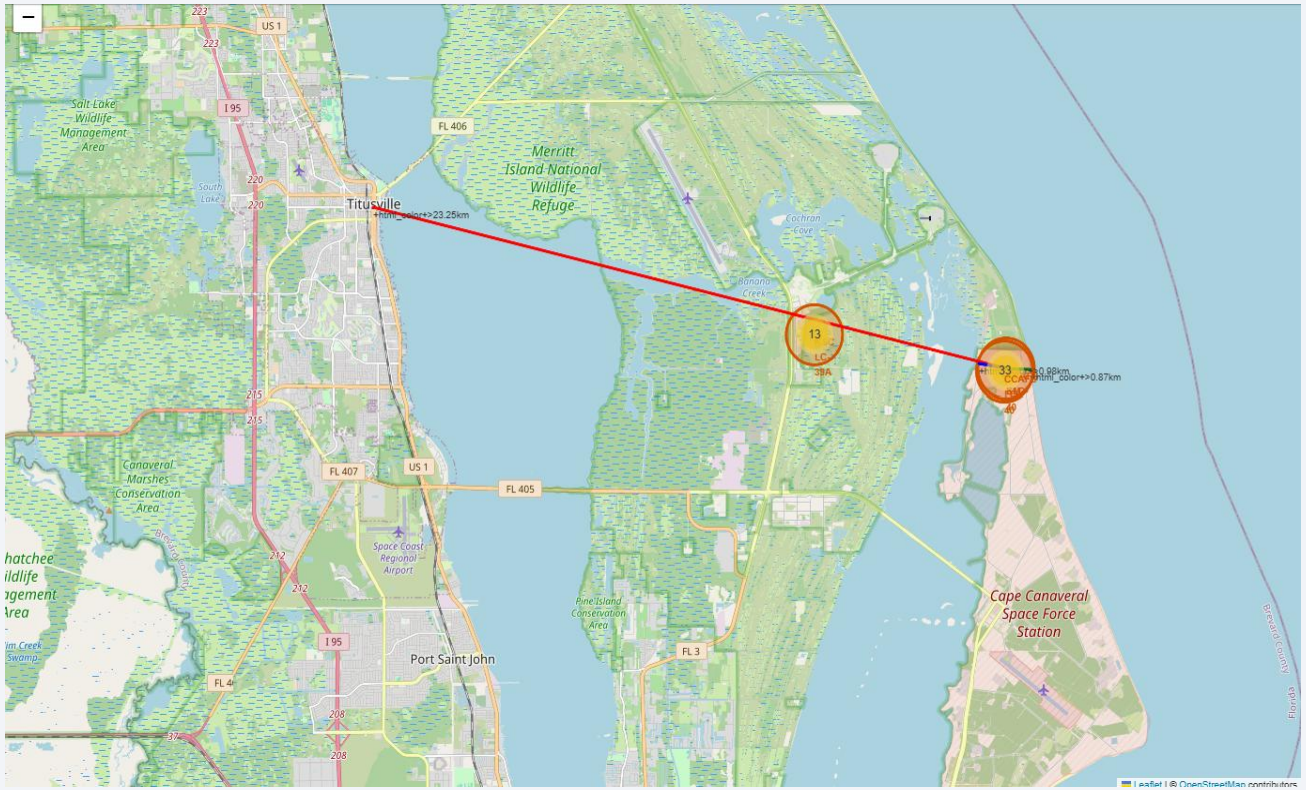
Launch Outcomes

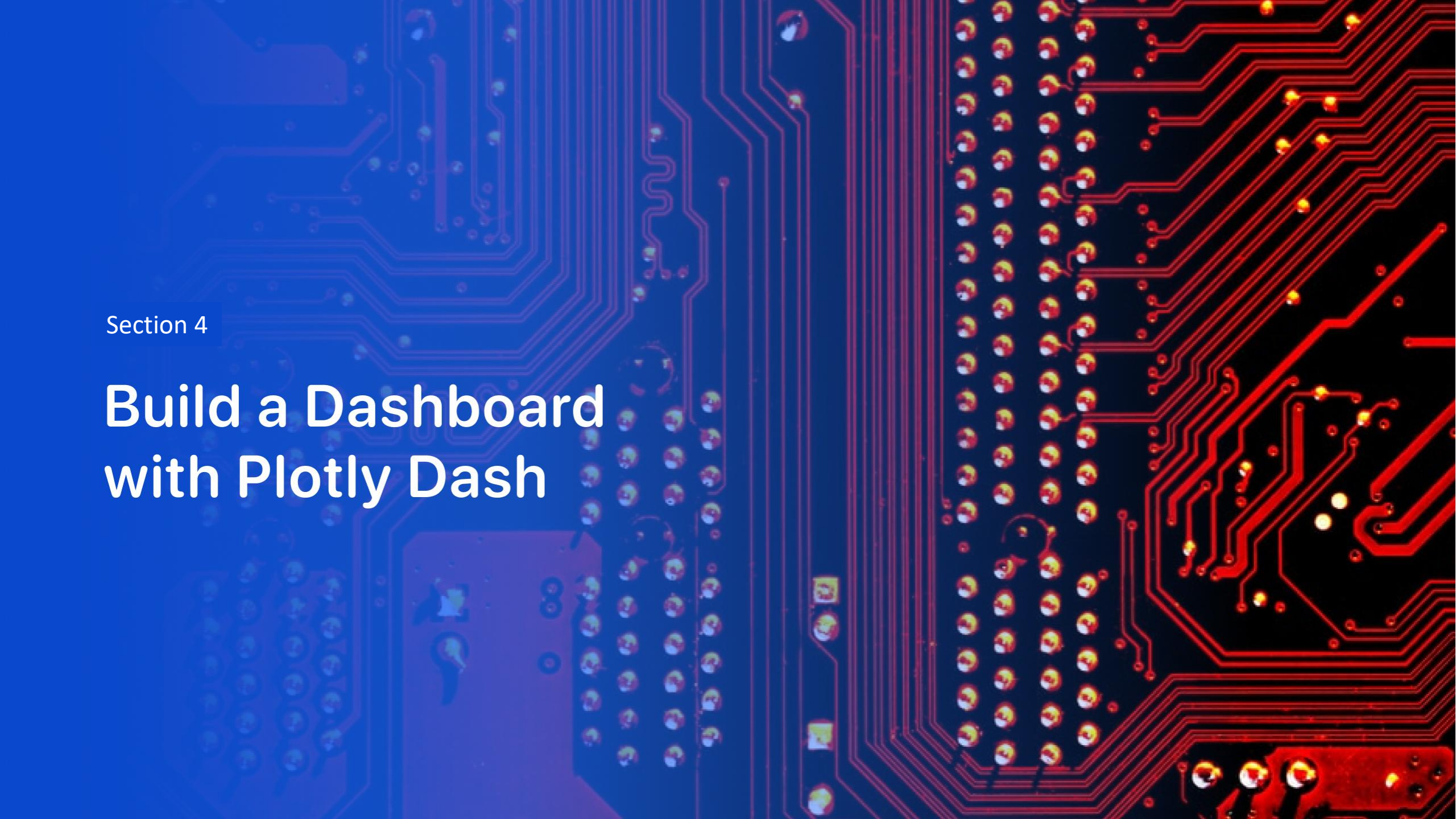
- Each launch site has pop up markers for launch outcomes
- Green = successful
- Red = unsuccessful



Distance to Proximities

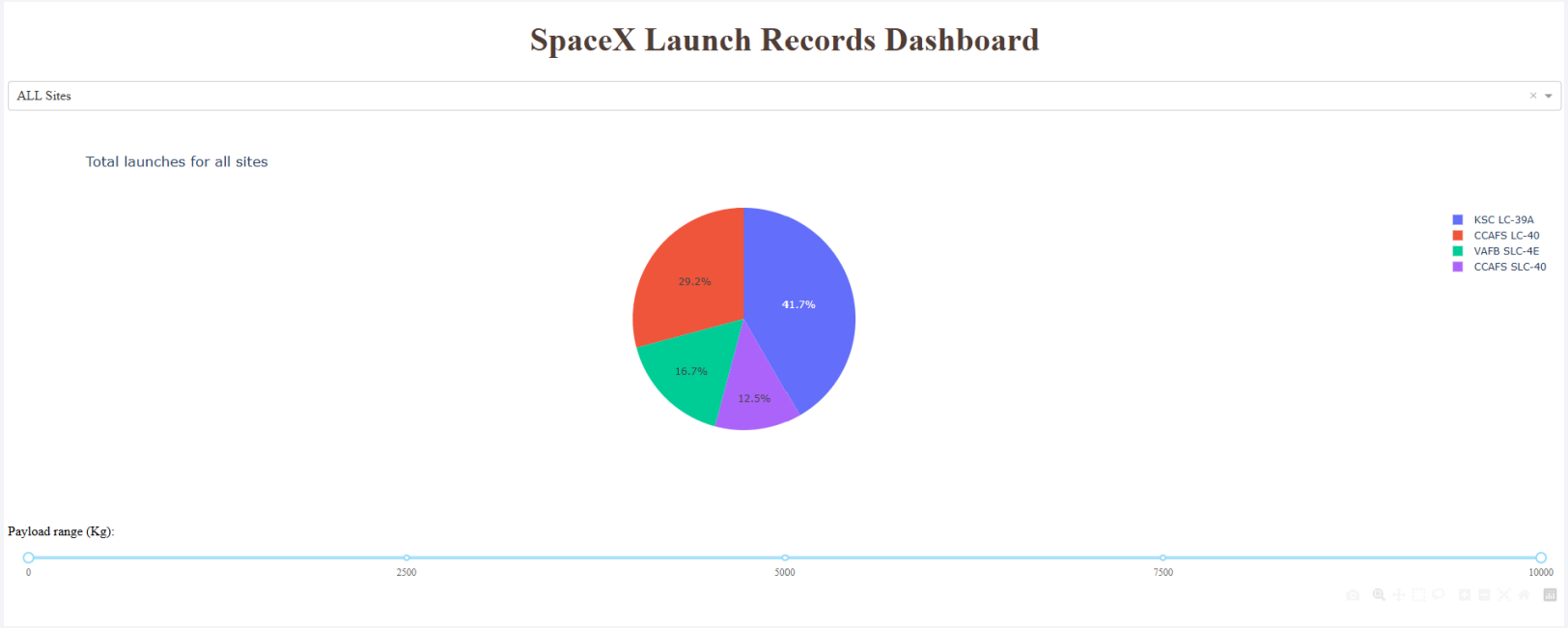
- Colored lines are added to denote the distance to the near city, railway, highway and coastline
- Red = City
- Blue = Railway/Highway
- Green = Coastline
- Launch sites tend to be far enough from population centers, roads and railways to avoid safety issues during unsuccessful launches/landings but close enough to be able to ship in material and people who work there




Section 4
Build a Dashboard with Plotly Dash

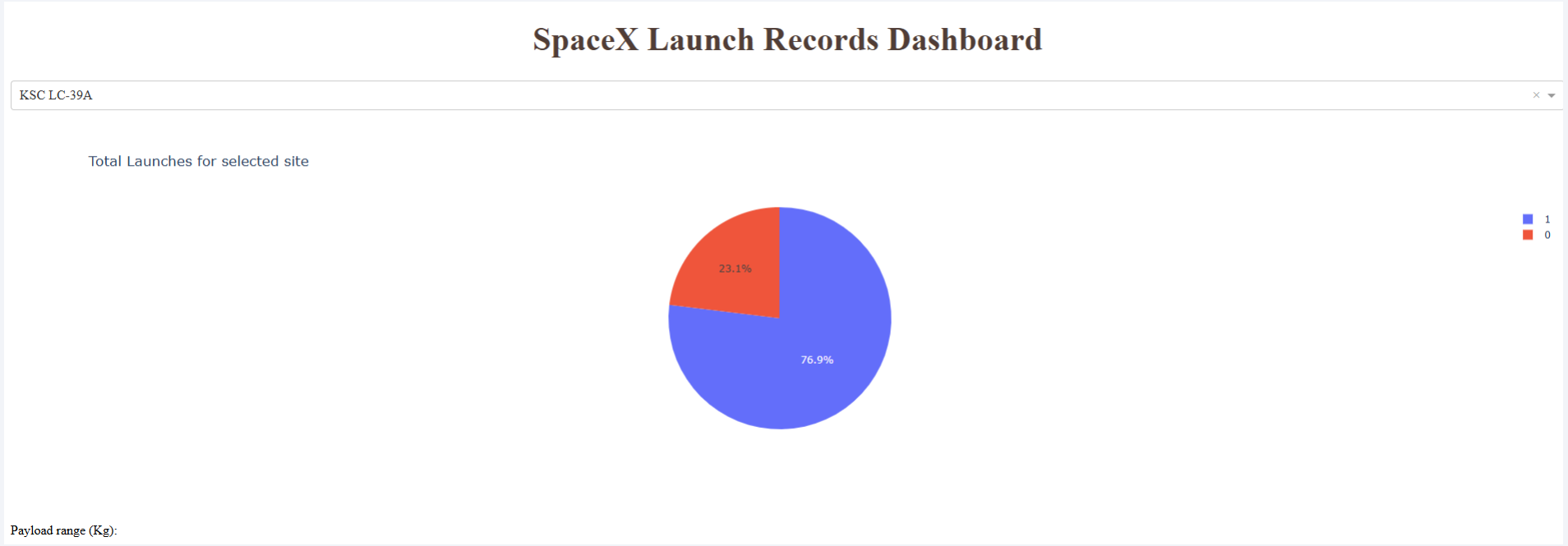
Success Rates by Launch Site

- KSC LC-39A was the site with the highest success rate at 41.7%



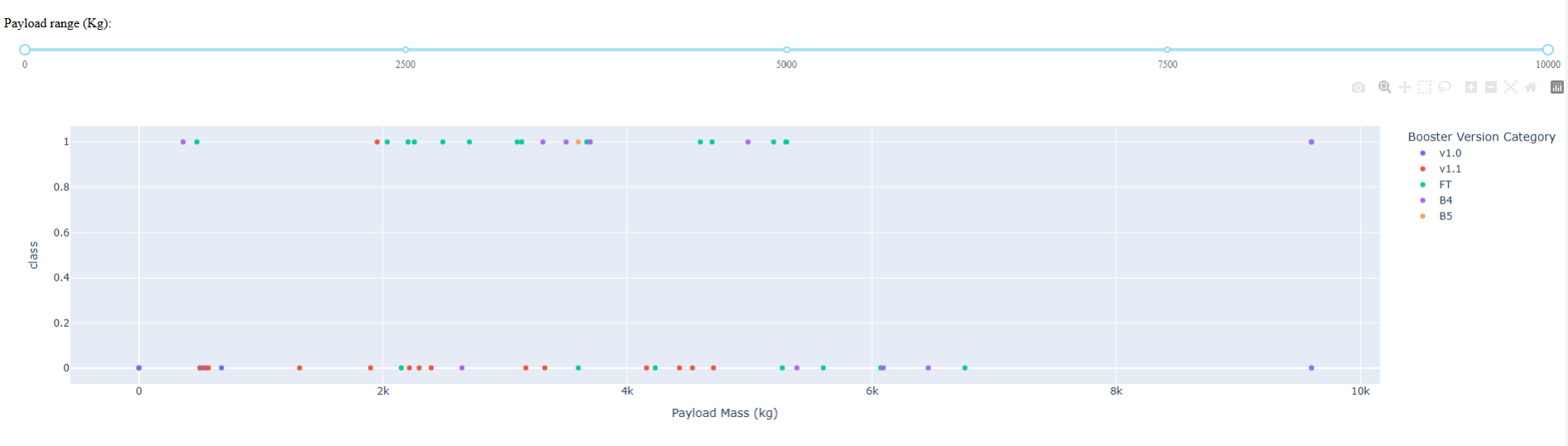
Most Successful Site

- KSC LC-39A was the most successful site with 76.9% of its launches resulting in success



Payload Mass and Mission Success

- Most successful launches had payloads between 2000 and under 6000 kg
- Booster Version FT appears to have the most successes out of the categories
  - 1 = successful 0 = unsuccessful



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All four models performed quite closely, but the decision tree had the highest accuracy of the four models at around 86%

TASK 12

Find the method performs best:

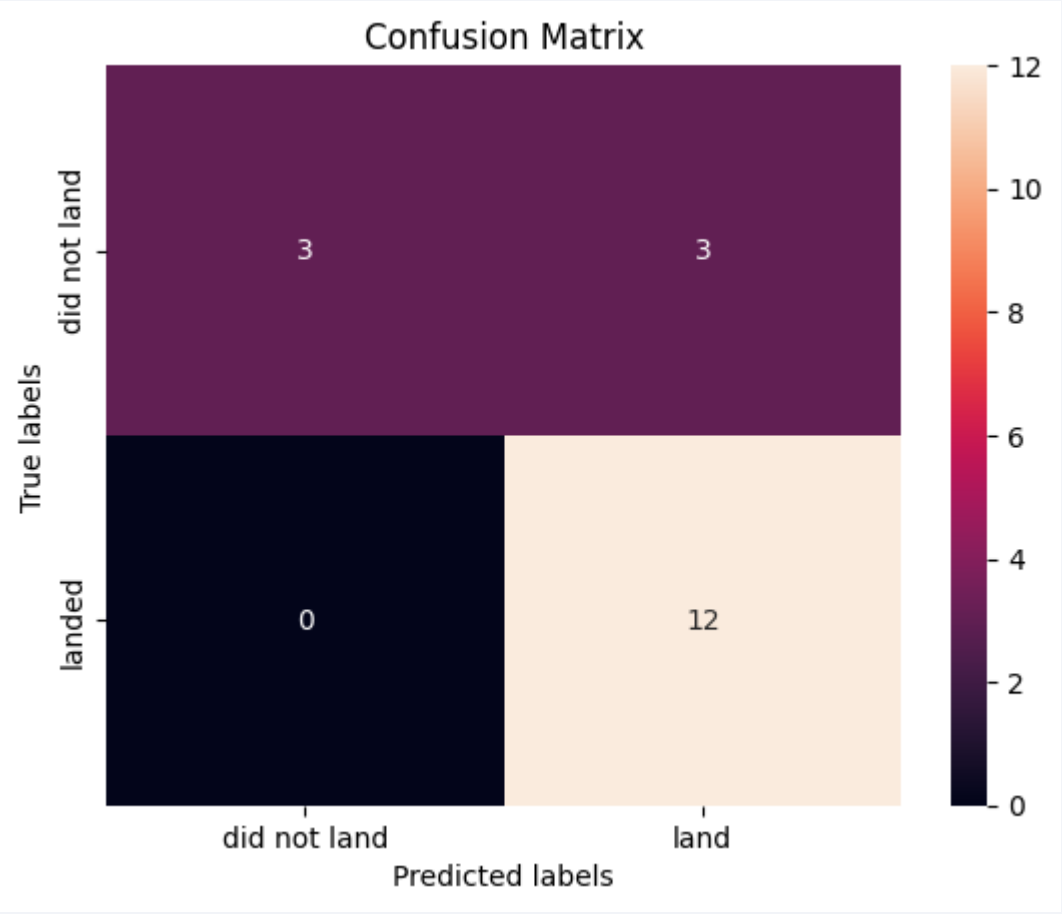
[49]:
```
best_model = {'LogReg':[logreg_cv.best_score_], 'SVM':[svm_cv.best_score_], 'Decision Tree': [tree_cv.best_score_], 'KNN':
df_columns = pd.DataFrame.from_dict(best_model, orient='index', columns=['Accuracy'])
sorted_columns = df_columns.sort_values(by='Accuracy', ascending=False).reset_index()
sorted_columns
```

t[49]:

	index	Accuracy
0	Decision Tree	0.860714
1	KNN	0.848214
2	SVM	0.848214
3	LogReg	0.846429

Confusion Matrix

- The Decision tree confusion matrix
  - True Positives = 12
  - True Negatives = 3
  - False Positives = 3
  - False Negatives = 0
- Precision = 12 / (12 + 3) = .8
- Recall = 12 / (12 + 0) = 1
- Accuracy = (12 + 3) / (12 + 3 + 3 + 0) = .83



Conclusions

- Launch success has improved over time
- The site with the highest success rate was found to be KSC LC-39A
- Orbits ES-L1, GEO, HEO and SSO had 100% success rates
- Launch sites are located near the equator to take advantage of the Earth's rotational speed. They are near the coastline and far enough from population centers to reduce change of damage in the event of failed launches, yet close enough to ship in people and material for work
- All four machine learning models had similar results, however the decision tree was the best model overall

- Considerations:
  - The data set is somewhat small and could have an impact on the results of this analysis
  - Other features may be considered to be included in future analysis to help improve model accuracy if possible

Thank you!