IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Conor Shanahan
12 NOV 2024

Outline

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

Executive Summary

- Methods
  - Data was collected using the SpaceX Web API and by scraping the SpaceX Wikipedia page containing Falcon 9 Launches. Data wrangling was performed primarily to encode the target variable specifically whether the booster landed successfully for different launches.
  - SQL queries and data visualization were used to explore the data, allowing for the generation of insights.
  - Lastly, predictive analysis was completed using four different modeling methods: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree.
- Summary of all results
  - Maybe add something
  - Each of the predictive models was found to perform equally well on the validation data with each model receiving an accuracy score of 83.33%

Introduction

- In recent years there has been a growing number of companies in the private sector which have shown the ability to make space exploration possible. SpaceX is a company headed by Elon Musk that has been arguably most successful in this area. This can be attributed to SpaceX's ability to reuse the first stage of their rockets, specifically the Falcon 9 rocket. As someone working for a rival space exploration company, we are interest in determining if the first stage of a rocket will land, generating insight into the cost of the launches.
- Key Questions:
  - What data is available on Falcon 9 launches that can be used for such modeling?
  - What insights can be generated from EDA and data visualization?
  - Does the data need to be transformed and/or engineered for modeling?
  - Can a model predict whether the first stage of a Falcon 9 rocket will successfully land and if so, which model performs the best?

Section 1

Methodology

Methodology

Executive Summary

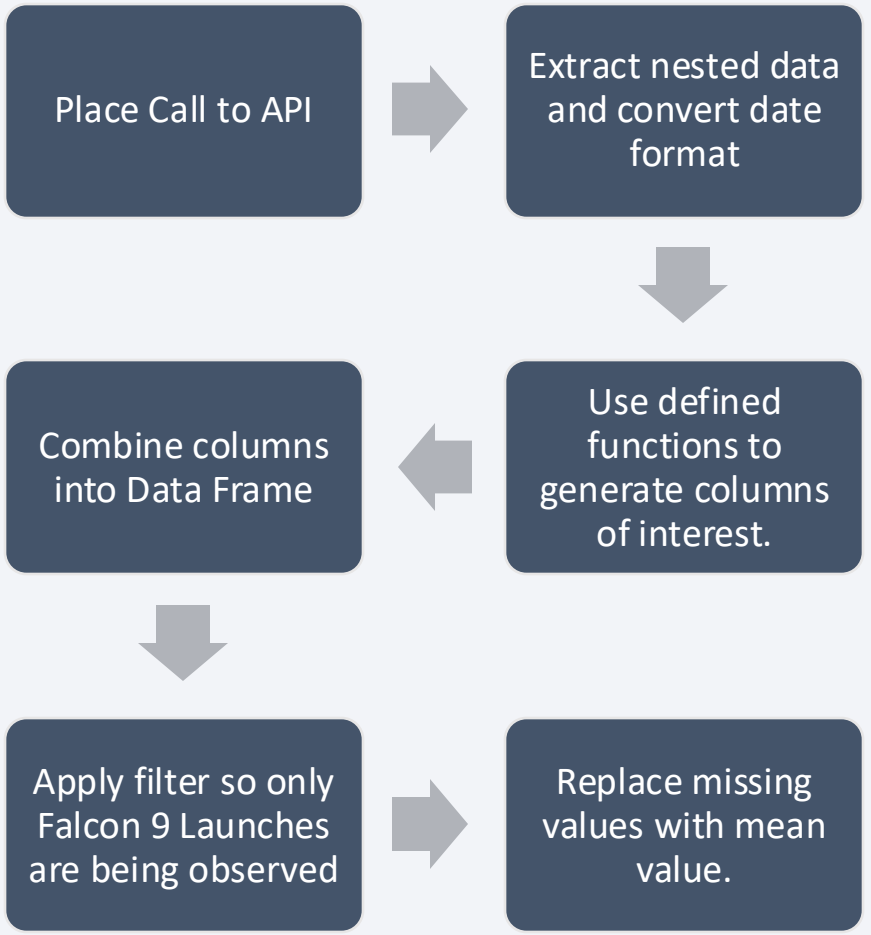
- Data collection methodology:
  - Data was collected from the SpaceX API and the Falcon 9 Wikipedia Launch Table.
- Perform data wrangling
  - Data was cleaned and target variable was encoded in preparation for modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were tuned using GridSearchCV with 10 validation folds and a parameter dictionary.
  - Models were then scored on the validation data to assess performance.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

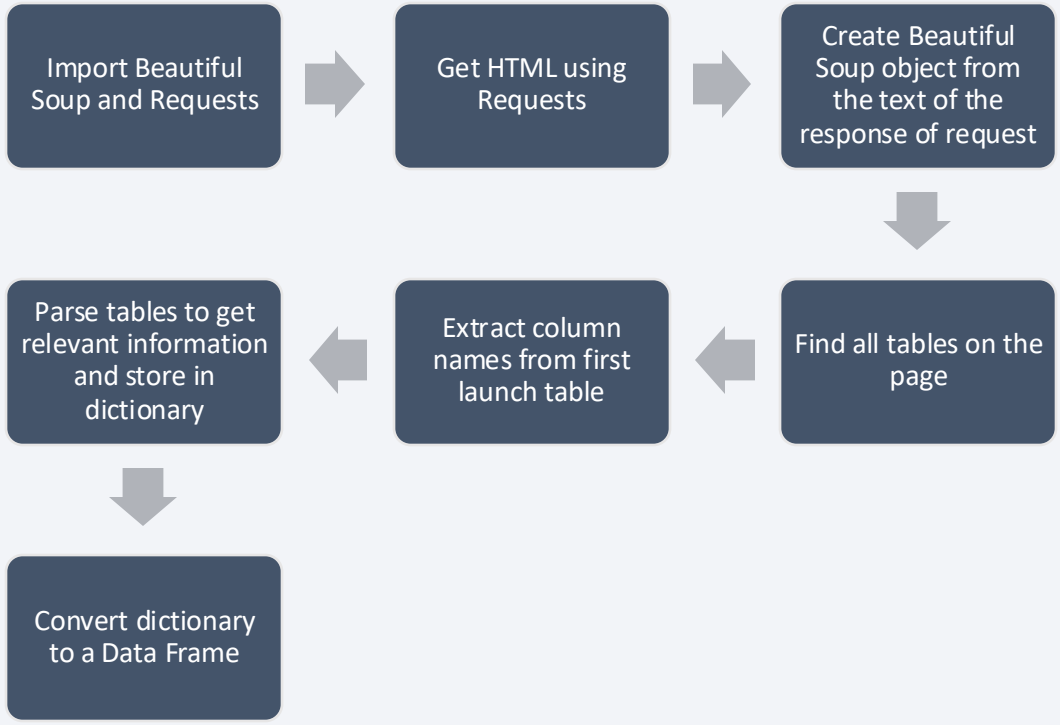
Data Collection – SpaceX API

- SpaceX has a publicly available API which can be used to retrieve information about their launches.
- Extract data from API response and convert date format
- Generate columns of interest and populate columns with data from API response
- Combine data into data frame
- Filter only Falcon 9 launches and replace missing column values with mean
- https://github.com/CWS01/Predicting-Successful-Stage-1-Return/blob/262cbafb1cb23771eb0b8ca75f089a7edde1581f/Data_Collection_Using_SpaceX_API.ipynb



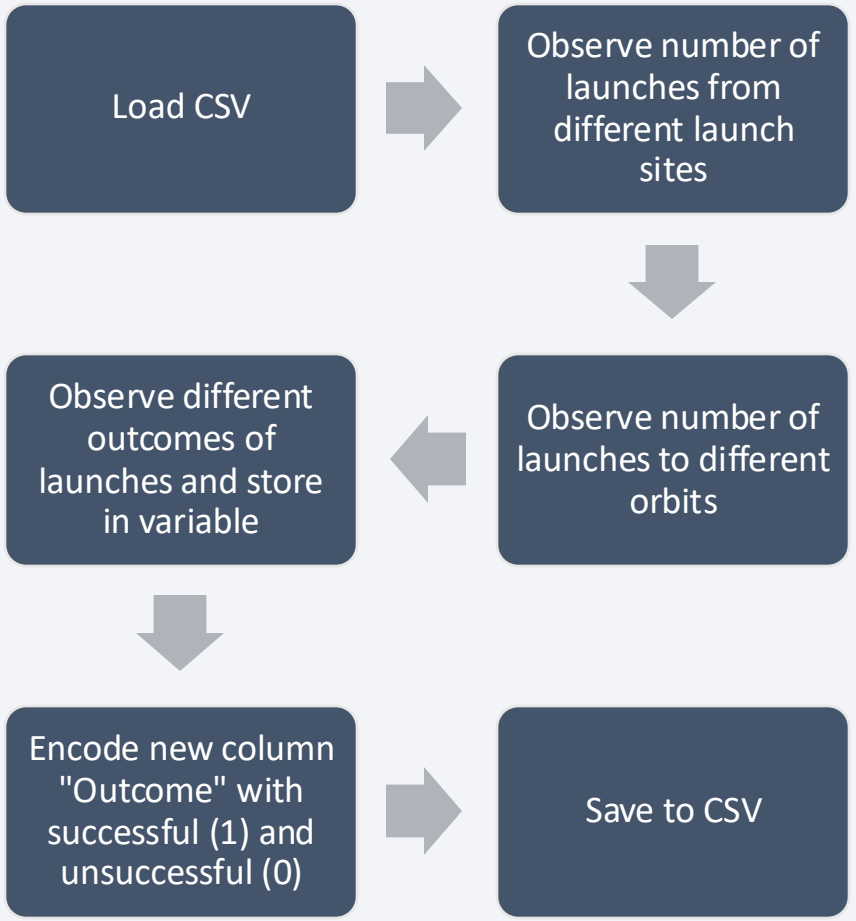
Data Collection - Scraping

- Tables with launch data for various SpaceX launches are available on Wikipedia.
- These tables can be scraped using Beautiful Soup to extract relevant information that can then be stored in a pandas data frame for further analysis.
- Link: https://github.com/CWS01/Predicting-Successful-Stage-1-Return/blob/019f042b317eb4d90b81f7f11de9cedf946fe68c/Data_Collection_Web_Scraping.ipynb



Data Wrangling

- The CSV file generated in the previous section was first loaded as a data frame.
- The number of launches from the different launch sites was then observed. Followed by the number of rockets launched into the different orbits.
- The different types of landings were looked at the count of each was assigned to a variable. This variable was then used to create a new column titled “Outcome” in the data frame.
- Unsuccessful landings were encoded as a 0 and successful landings were encoded as a 1.
- Link:



EDA with Data Visualization

- Launch Site Trends
  - Landing outcome was observed as a function of flight number and launch site using a scatter plot
  - Landing outcome was observed as a function of payload mass and launch site using a scatter plot
- Orbit Type Trends
  - Landing success rate by orbit type was visualized using a bar chart
  - Landing outcome was observed as a function of flight number and orbit type using a scatter plot
  - Landing outcome was observed as a function of payload mass and orbit type using a scatter plot
- Yearly Trends
  - Average landing outcome each year since 2010 was plotted as line chart
- Link: https://github.com/CWS01/Predicting-Successful-Stage-1-Return/blob/226fb7831caa047d0fe1278f2b681b410303e6b4/EDA_Data_Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Unique launch sites
  - 5 launch records from a site beginning with “CCA”
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass carried by booster version “F9 v1.1”
  - Date of first successful ground pad landing
  - Successful drone ship landings with payloads between 4000 and 6000 kg
  - Total number of successful vs unsuccessful landings
  - Boosters which have carried the maximum payload mass
  - Drone ship landing failures in 2015
  - Count of different landing outcomes between the date 06/04/2010 and 03/20/2017
- Link: https://github.com/CWS01/Predicting-Successful-Stage-1-Return/blob/66d457c1a92c7459c77afc82b6afa2f07df759e9/EDA_SQL.ipynb

Build an Interactive Map with Folium

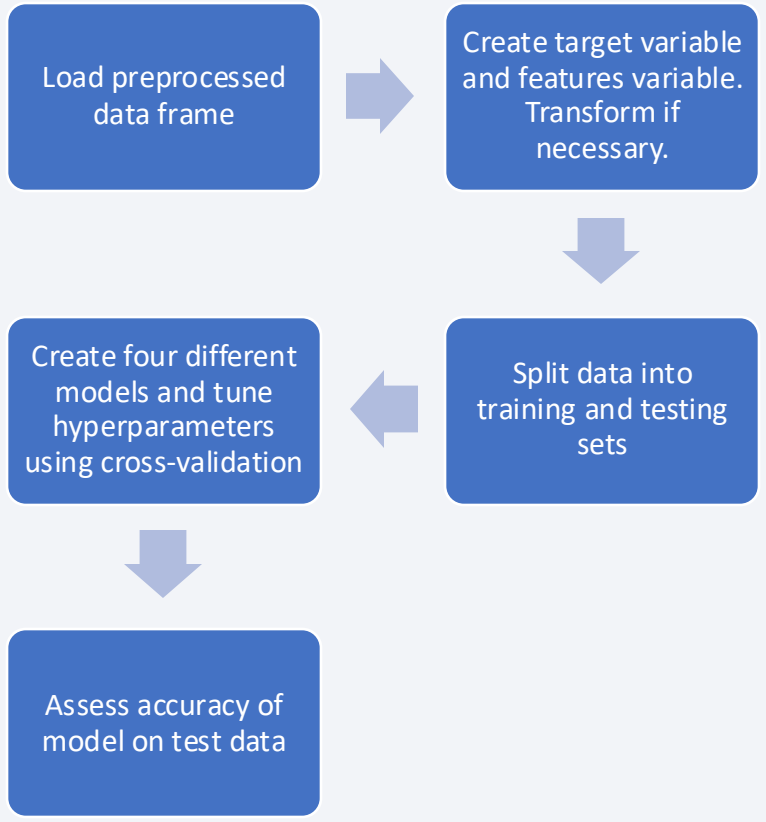
- Folium Map Objects
  - Circles were added to mark the different launch sites for the rockets
  - Markers were added to label each of the launch sites with their respective names
  - Landing outcomes were encoded with either “red” or “green” to mark successful vs. unsuccessful launches
  - Marker clusters were used to show the different landing outcomes on the map
  - Lines were added to the map to show the distance from a specific launch site to different landmarks. The landmarks of interest were:
    - Coastline
    - Highway
    - Railroad
    - City
- Link: https://github.com/CWS01/Predicting-Successful-Stage-1-Return/blob/baafb8b46f60d53deabb0fb68a3fb93f04943376/Data_Visualization_Folium_Map.ipynb

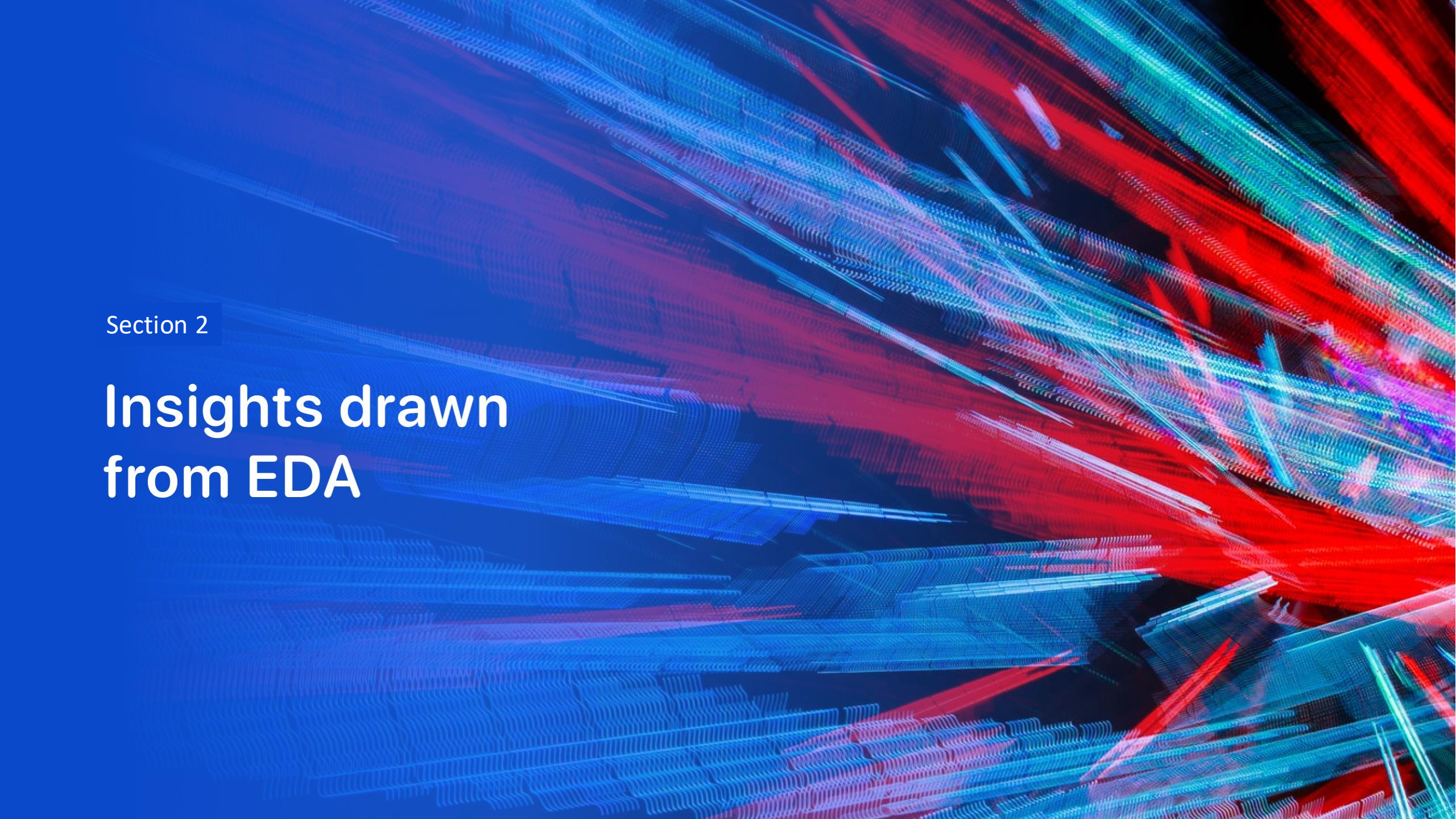
Build a Dashboard with Plotly Dash

- A dashboard was created for interactive visualization of some rocket launch information.
- A dropdown menu was added to select a single launch site or all launch sites
- The first graph is a pie chart
  - When all sites are selected, the pie chart shows the percentage of successful landings that can be attributed to individual sites.
  - When one site is selected the pie chart shows the ratio of successful vs. unsuccessful landings at that specific site.
- A slider was added to filter on different ranges of payload mass carried by a mission
- The second graph is a scatter plot
  - The scatter plot shows the landing outcome as a function of payload mass and booster version. Note: the payload mass can be adjusted using the slider.
- Link:

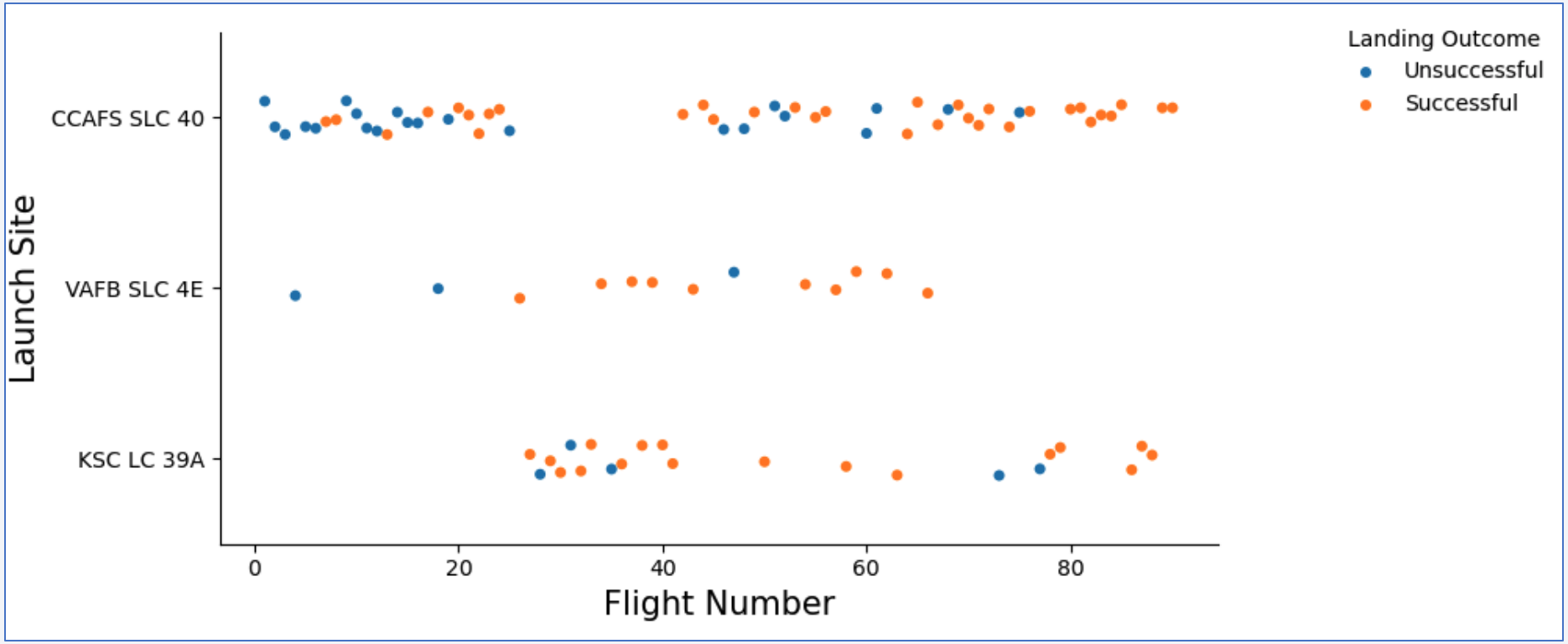
Predictive Analysis (Classification)

- Data was loaded, split into target and feature arrays, and transformed if necessary.
- Data was split into training and testing sets
- Four models were created and trained on training data:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- Hyperparamters to test were defined and cross validation using GridSearchCV was completed.
- Accuracy was then assessed on test data using the hyperparameters selected during cross-validation
- Link:




Section 2
Insights drawn from EDA

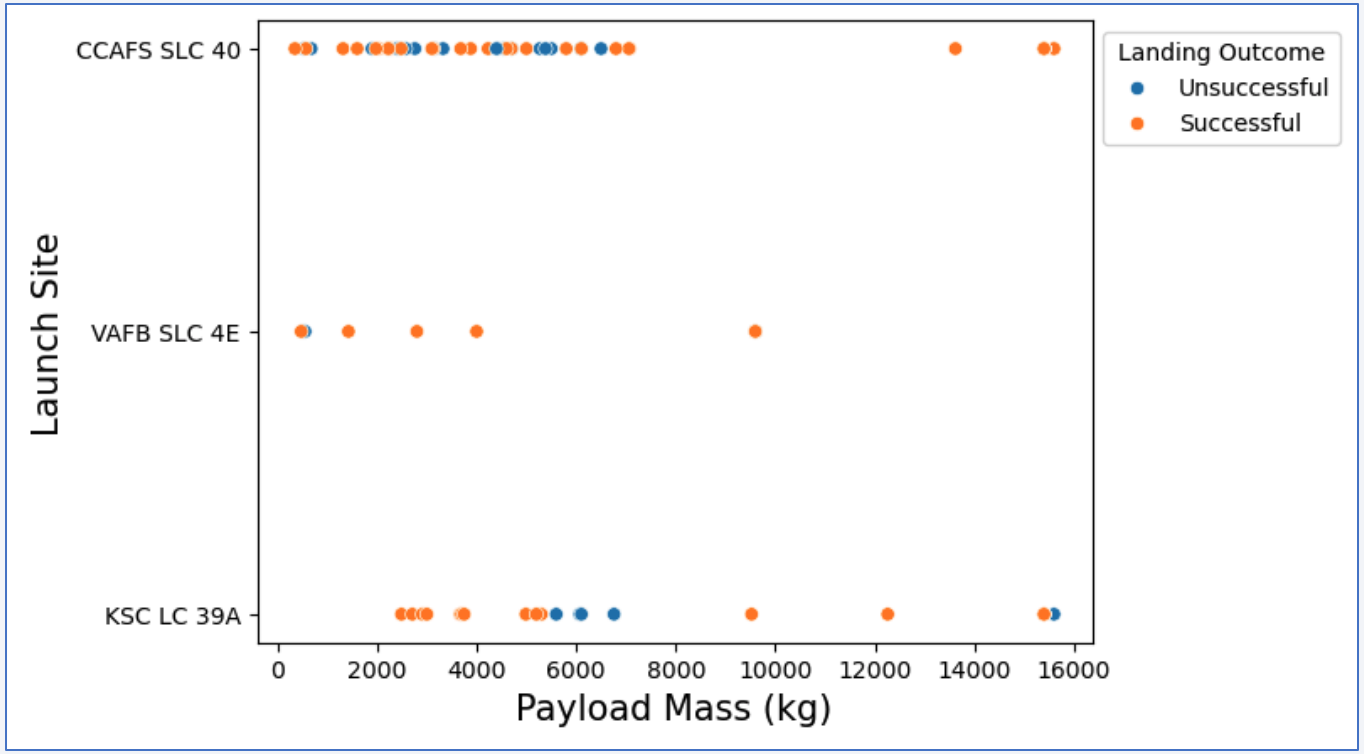
Flight Number vs. Launch Site



- The above scatter plot shows the landing outcomes based on flight number and launch site.
- Successful launches appear to increase with increasing flight number.

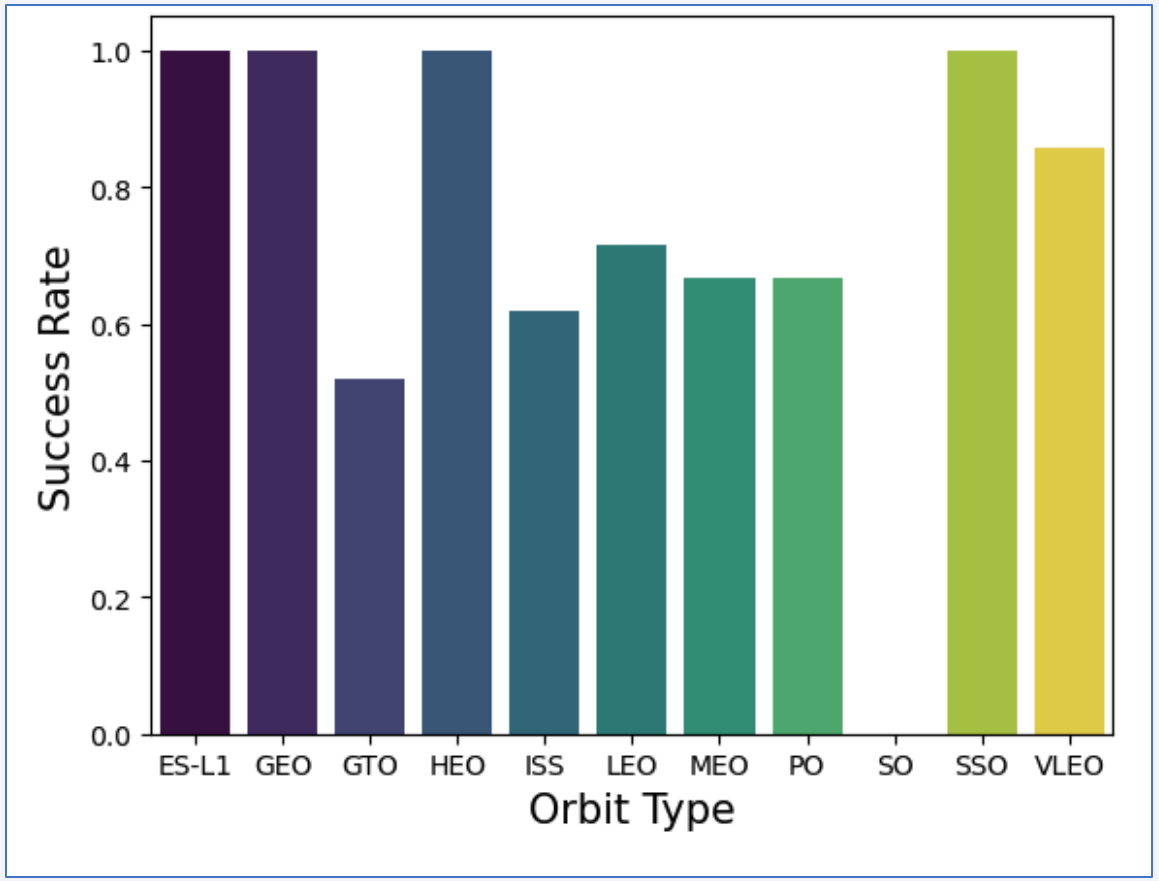
Payload vs. Launch Site

- The scatter plot to the right shows landing outcome as a function of launch site and payload mass.
- There does not appear to be any specific trends. However, there does appear to be more failures around a 6000kg payload mass when launching from KSC LC 39A



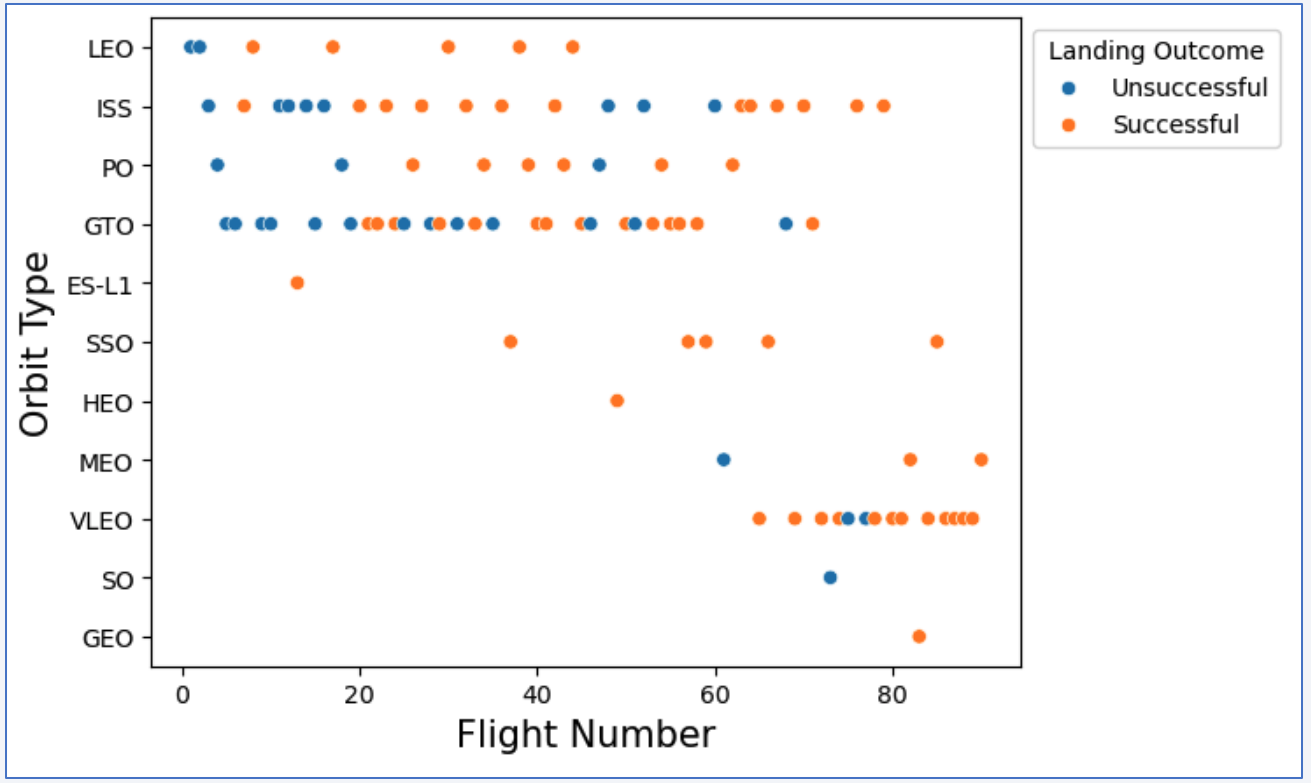
Success Rate vs. Orbit Type

- The bar chart to the right shows the success rate of landings based on the destination orbit.
- ES-L1, GEO, HEO, and SSO target orbits have resulted in 100% successful landings.
- SO target orbit has resulted in 0% successful landings.



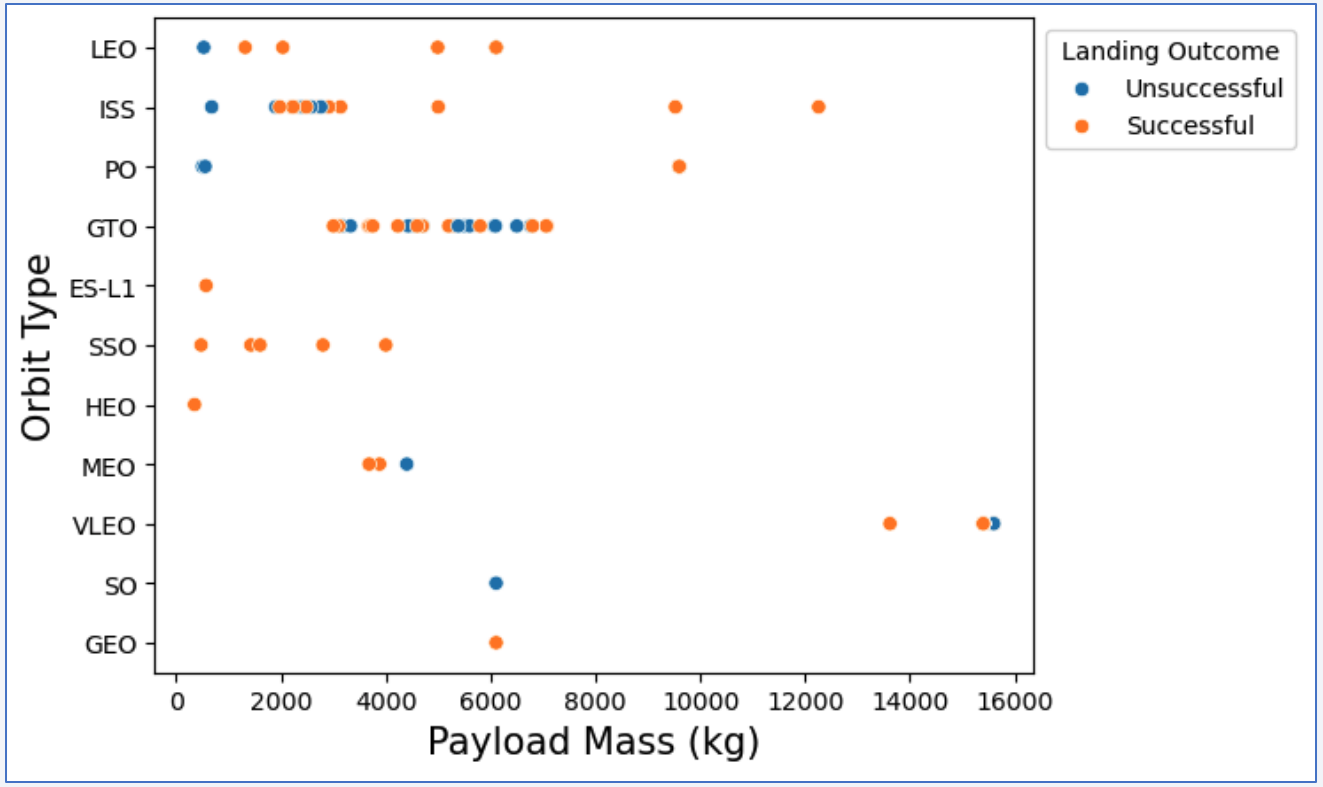
Flight Number vs. Orbit Type

- The scatter plot to the right shows the effects of flight number and orbit type on landing outcome.
- This plot again reiterates the fact that as flight number increases there appears to be more successful landings.



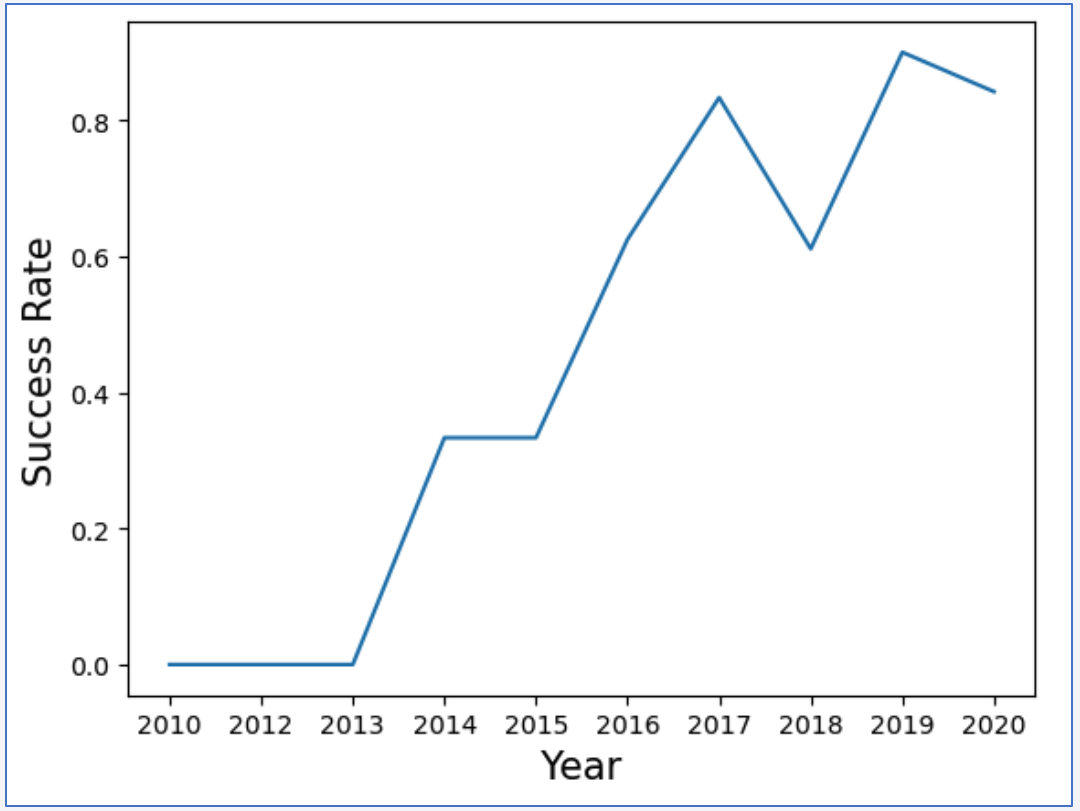
Payload vs. Orbit Type

- The effects of orbit type and payload mass on landing outcome are shown to the right.
- Most failures in the graph appear to be associated with smaller payloads and the GTO orbit, but no clear correlation is observed.



Launch Success Yearly Trend

- The landing success rate as a function of year is shown to the right.
- The success rate has grown significantly since 2010, reaching a high of ~85% in 2019.



All Launch Site Names

- Query: SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
- Result:

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

- Explanation: The query selects all the distinct launch sites in the data frame, of which there are four.

Launch Site Names Begin with 'CCA'

- **Query**: SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- **Result**:

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

- **Explanation**: Used to further explore the data.

Total Payload Mass

- **Query**: SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)%'
- Result:

SUM(PAYLOAD_MASS__KG_)
48213

- Explanation: The sum of payload mass carried by boosters from “NASA (CRS)” is 48,213kg.

Average Payload Mass by F9 v1.1

- **Query**: SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
- Result:

AVG(PAYLOAD_MASS__KG_)
2928.4

- Explanation: The average payload mass carried by booster version "F9 v1.1" is 2928.4kg.

First Successful Ground Landing Date

- Query: SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
- Result:

min(Date)
2015-12-22

- Explanation: The first successful landing outcome on ground pad was found to be on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
SELECT Booster_Version FROM SPACEXTABLE

    WHERE (Landing_Outcome = 'Success (drone ship)') AND
          (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

- Result:

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

- Explanation: There were four booster versions carrying a payload between 4000 and 6000kg which had a successful landing outcome on a drone ship.

Total Number of Successful and Failure Mission Outcomes

- Query:

```
SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE
    GROUP BY Mission_Outcome
```

- Result:

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

- Explanation: There were 100 successful missions and only one mission failure.

Boosters Carried Maximum Payload

- Query:

```
SELECT Booster_Version FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM
        SPACEXTABLE)
```

- Results:

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

- Explanation: There were 12 booster versions that have carried the maximum payload

2015 Launch Records

- Query:

SELECT substr(DATE, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 0, 5) ='2015'

- Results:

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

- Explanation: There were two failed drone ship landings in the year 2015. The first occurred in January and the second occurred in April

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

- Query:

```
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Launch_Count FROM SPACEXTABLE
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY Landing_Outcome
    ORDER BY Launch_Count DESC
```

- Result:

Landing_Outcome	Launch_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

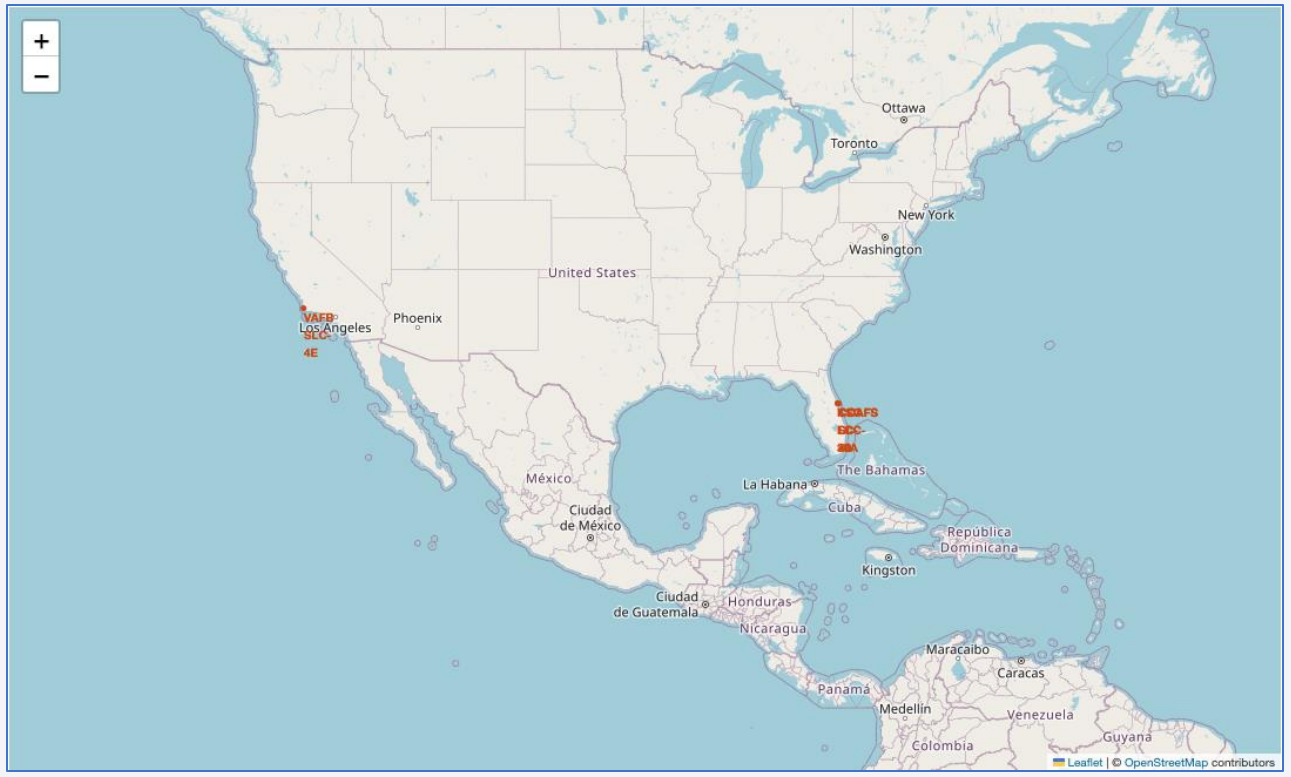
- Explanation: There were 31 total landing attempts between June 6, 2010 and March 20, 2017. The most common was “No attempt.”

Section 3

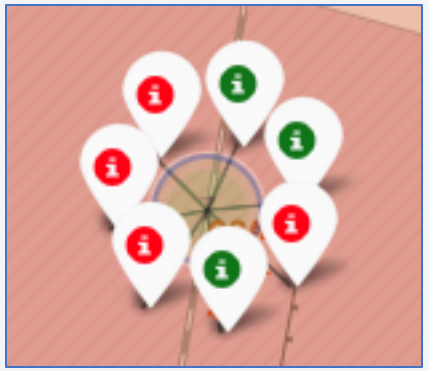
Launch Sites Proximities Analysis

Launch Site Locations

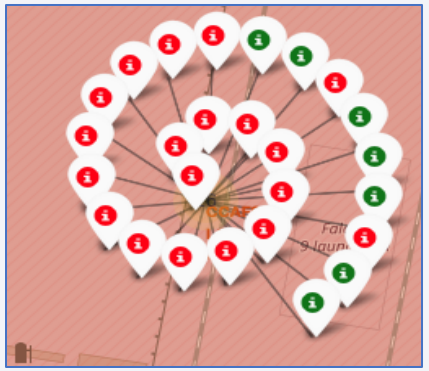
- VAFB SLC-4E (California, USA)
- KSC LC-39A (Florida, USA)
- CCAFS LC-40 (Florida, USA)
- CCAFS SLC-40 (Florida, USA)



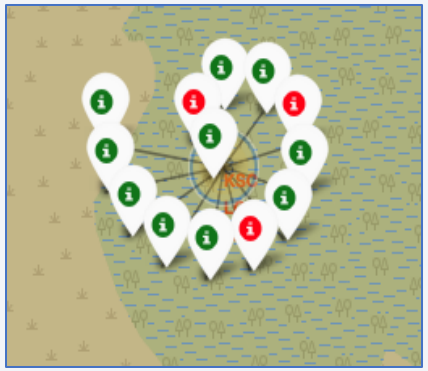
Successful/Unsuccessful Landings per Launch Site



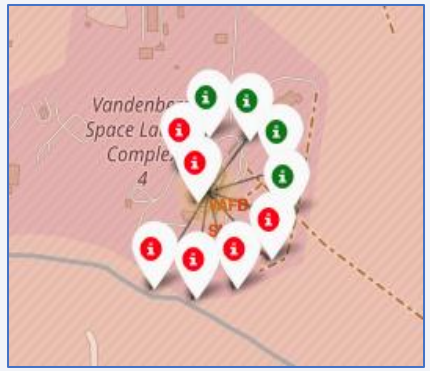
CCAFS SLC-40



CCAFS LC-40



KSC LC-39A

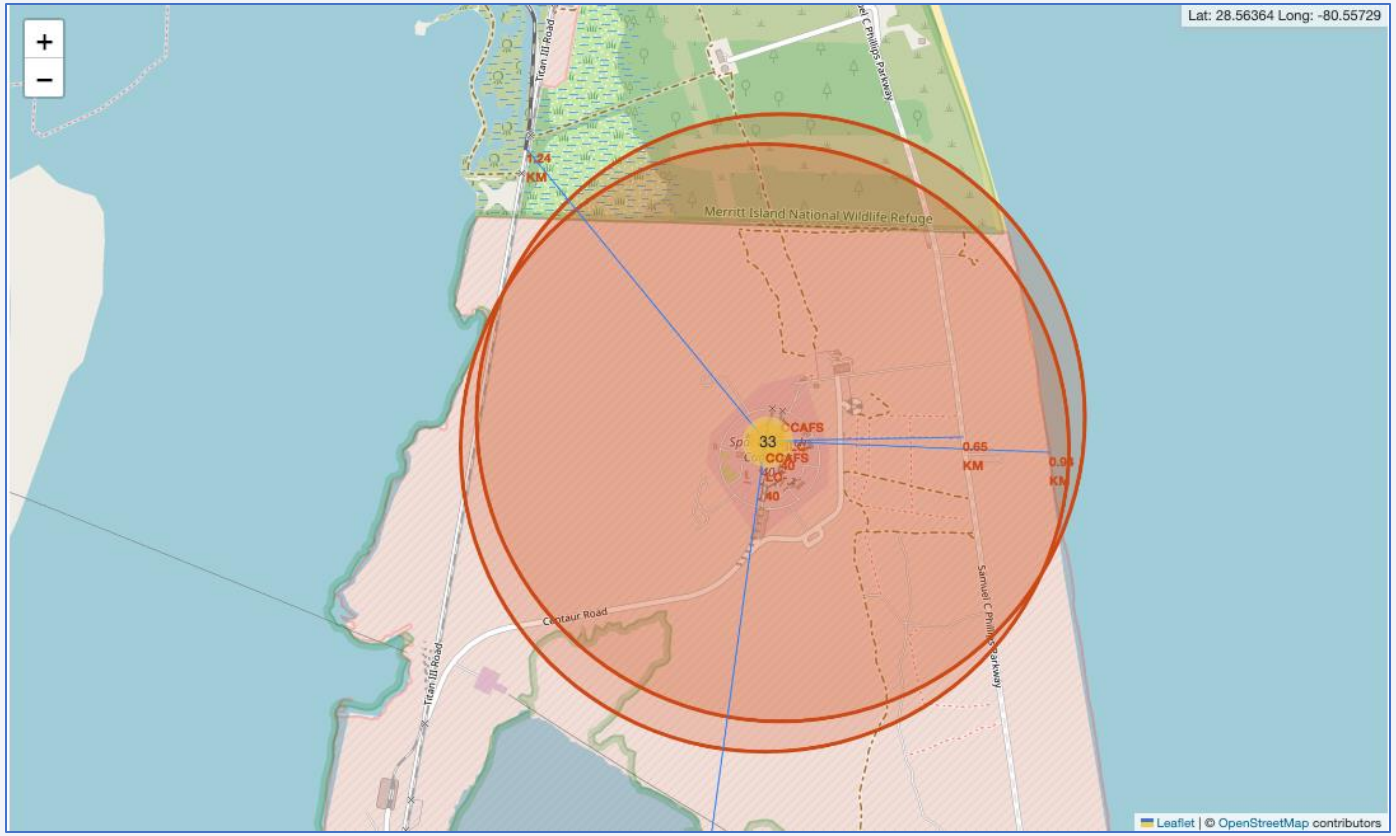


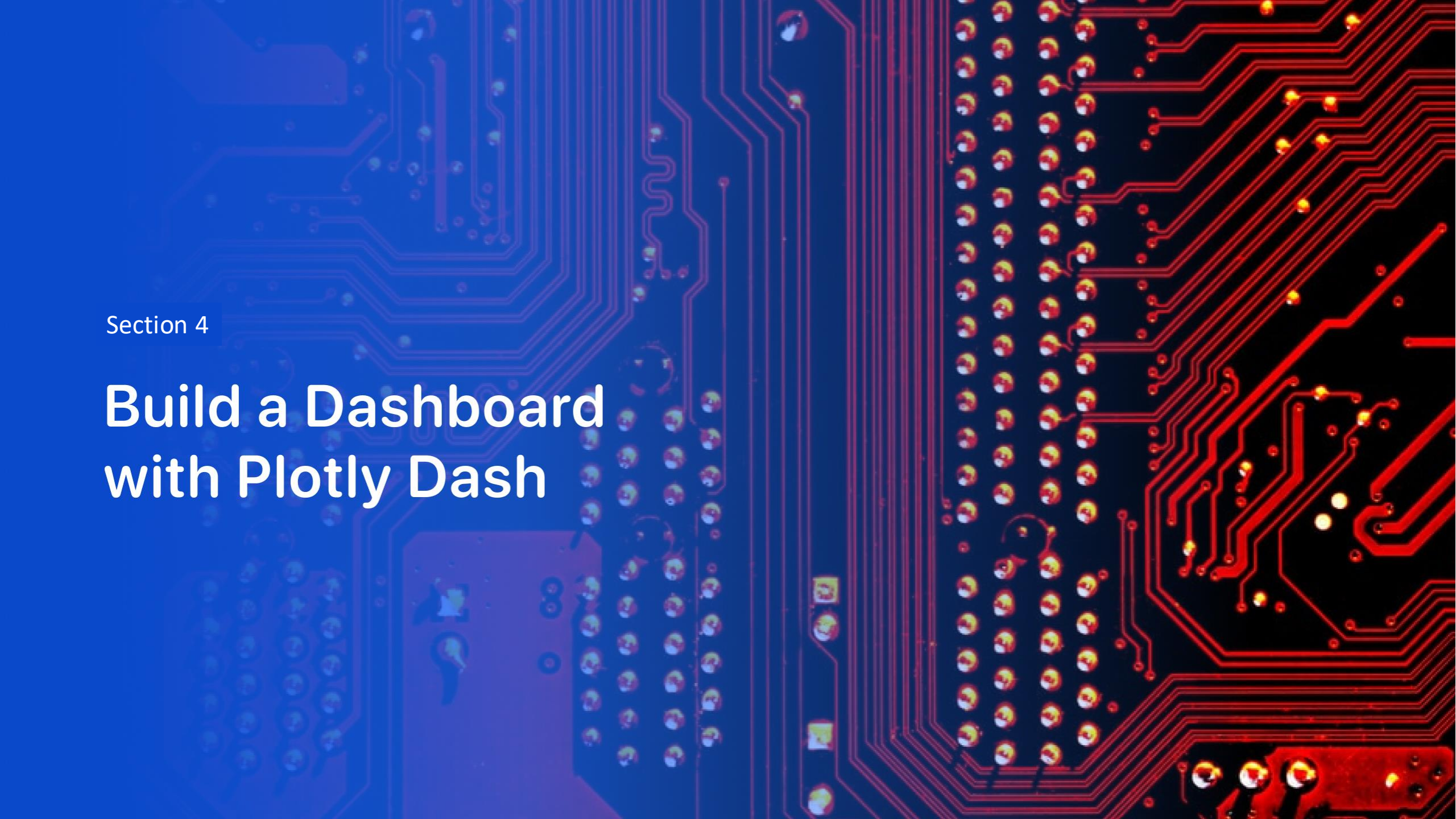
VAFB SLC-4E

- The above screen shots show the different landing outcomes, successful (green) and unsuccessful (red), for the different launch sites.
- The relative landing outcome success rate per launch site can be inferred from this information.

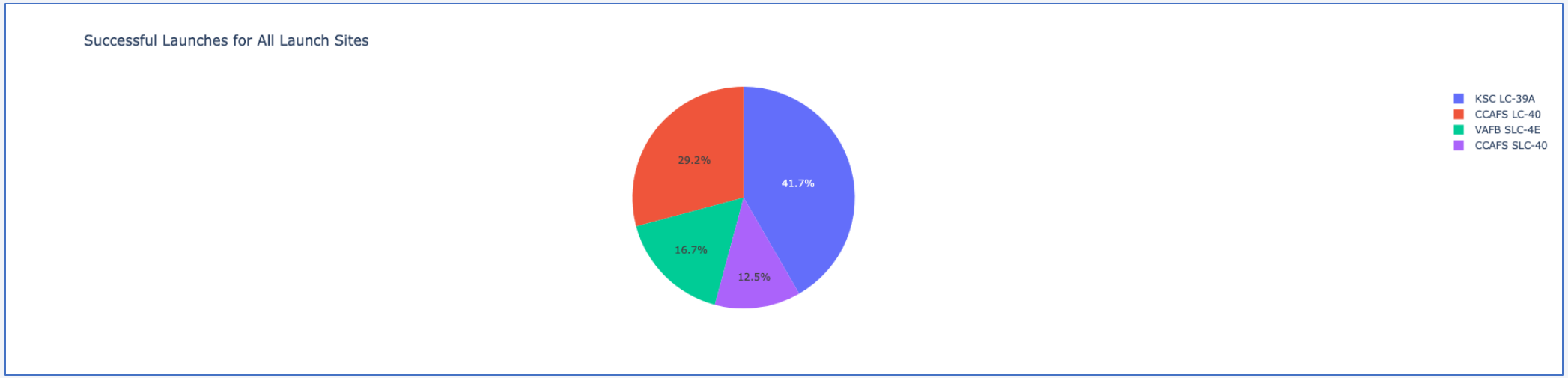
Distance from Launch Site to Different Landmarks

- The CCAFS LC-40 launch site is ~0.94 km away from the nearest coastline.
- The launch site is ~0.65km away from the nearest highway.
- The launch site is ~1.24km away from the nearest railroad.
- The launch site is ~19.82km away from the nearest city.



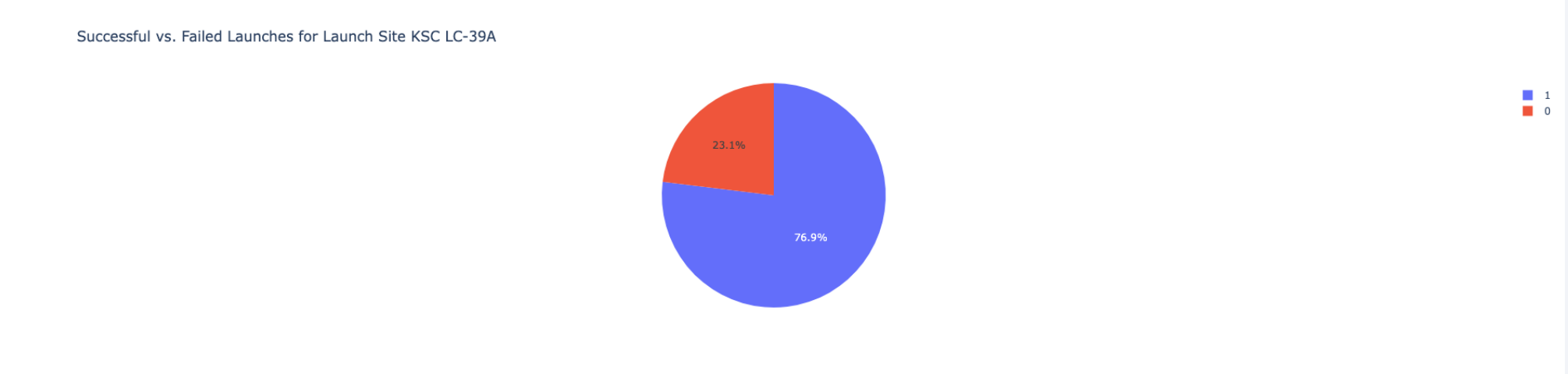

Section 4
Build a Dashboard with Plotly Dash

Landing Outcome Success Distribution by Launch Site



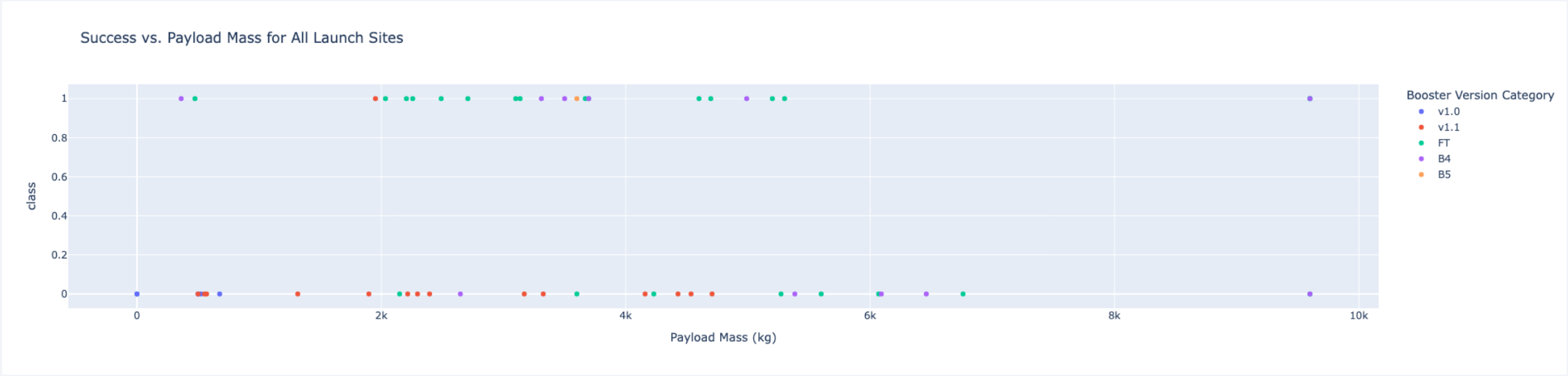
- The above pie chart shows the distribution of successful landings according to launch site.
- The launch site with the most successful landings is KSC LC-39A with 41.7% of the total successful landings.

Landing Outcome Distribution for Launch Site KSC LC-39A



- The above plot shows the distribution of unsuccessful and successful launches at the launch site: KSC LC-39A.
- Out of all attempted landings, 76.9% were successful and 23.1% were unsuccessful.

Landing Outcome vs. Payload Mass



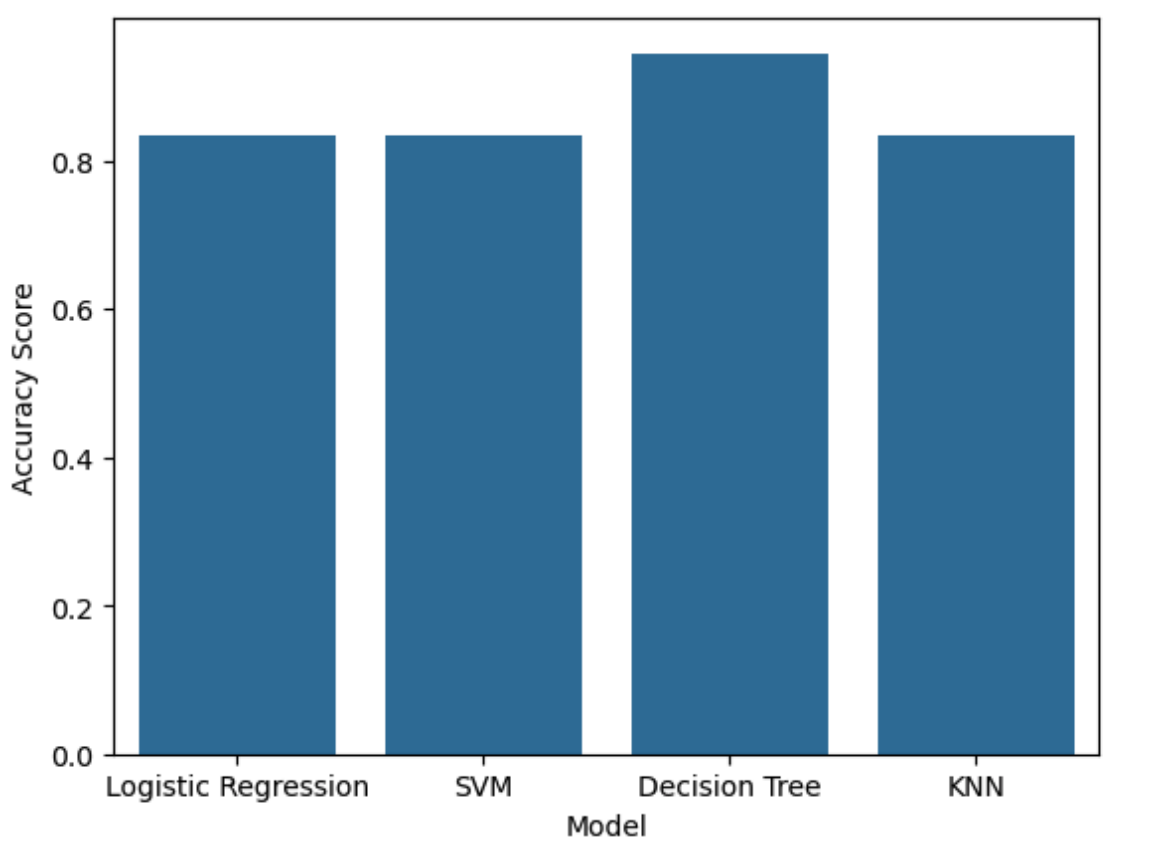
- The above plot shows the landing success as function of payload mass and booster version.
- This plot can be filtered based on payload mass and launch site (not shown but functionality is present in dashboard)
- Most successful launches occurred with a payload mass between 2000 and 6000kg and booster version FT

Section 5

Predictive Analysis (Classification)

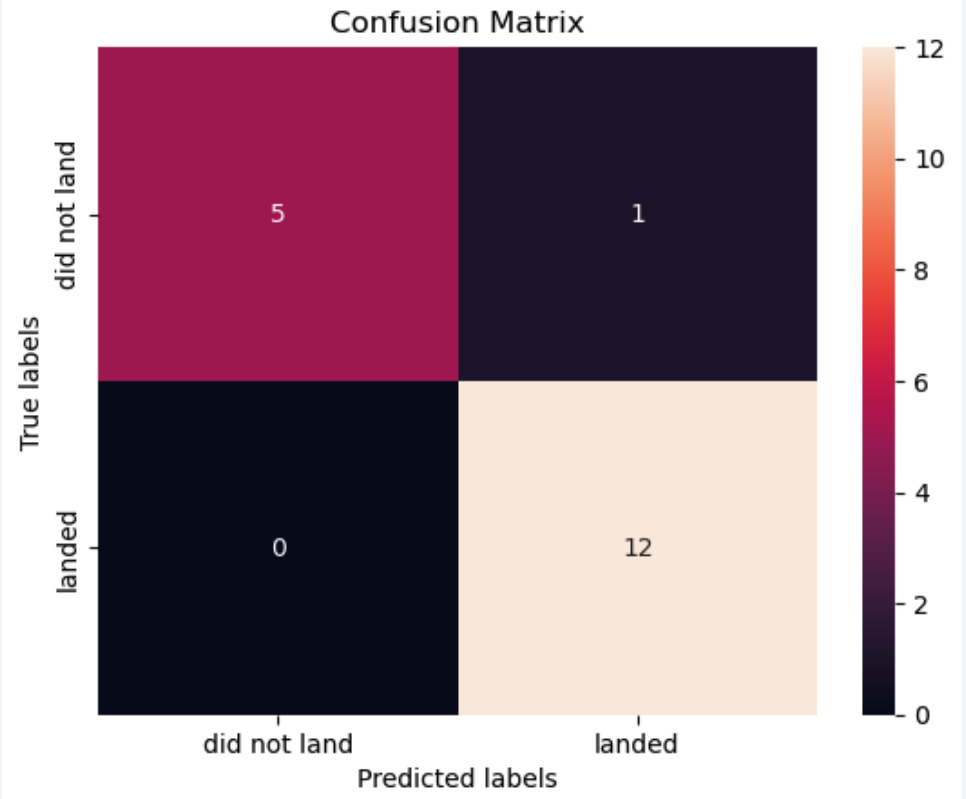
Classification Accuracy

- The accuracy of each model on the test data is shown in the bar graph on the right.
- The logistic regression, SVM, and KNN models each had an accuracy score of 83.33%.
- The Decision Tree model had an accuracy score of 94.44%, the highest of any of the models.



Confusion Matrix

- The confusion matrix for the decision tree model is shown on the right.
- The model had 17/18 outcomes correctly predicted.
- The one incorrect classification was a false positive where the model predicted a successful landing but in truth the landing was unsuccessful.



Conclusions

- Sufficient data can be extracted from the SpaceX API and publicly available data such as launch tables on Wikipedia.
- As both more launches are made (and time goes on) SpaceX has shown that successful Falcon 9 landings are increasing.
- Payload doesn't seem to have a high correlation with landing outcome, but many successful landings were achieved with payload between 2000 and 6000kg.
- Launch sites are in close proximity to coastlines, highways, and railroads, but are not in close proximity with cities.
- The decision tree machine learning model can be used to predict successful landing with 94.4% accuracy. The other models can predict successful landing with 83.33% accuracy.

Thank you!