IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anjali Walisinghe
24 July 2024

Outline

Executive Summary

Project: Maximizing cost savings for SpaceY through Falcon 9 rocket launches

Methodology:

- Collection of SpaceX launch data from SpaceX and related Wiki sites
- Exploratory data analysis (EDA) and data visualization of SpaceX launch features, including rocket, launch site and landing features
- Predictive model exploration, testing and training to determine the best classification model to predict successful rocket launches

Results:

- Data collection and exploratory analysis allowed identification of rocket launch features of interest

Introduction

Background

- SpaceY recognizes the exciting new opportunities available with commercial innovations in space travel
- Competing with existing players in the market means understanding all aspects of space flight to provide the best and safest experience without an exorbitant price tag
- Existing commercial space flight providers have already tested and documented space flight with various different rockets and launchers
- **Problems**
- What equipment/work is needed to achieve space flight?
- Where can SpaceY minimize cost without impacting on space flight safety?

Section 1

Methodology

Methodology

Executive Summary

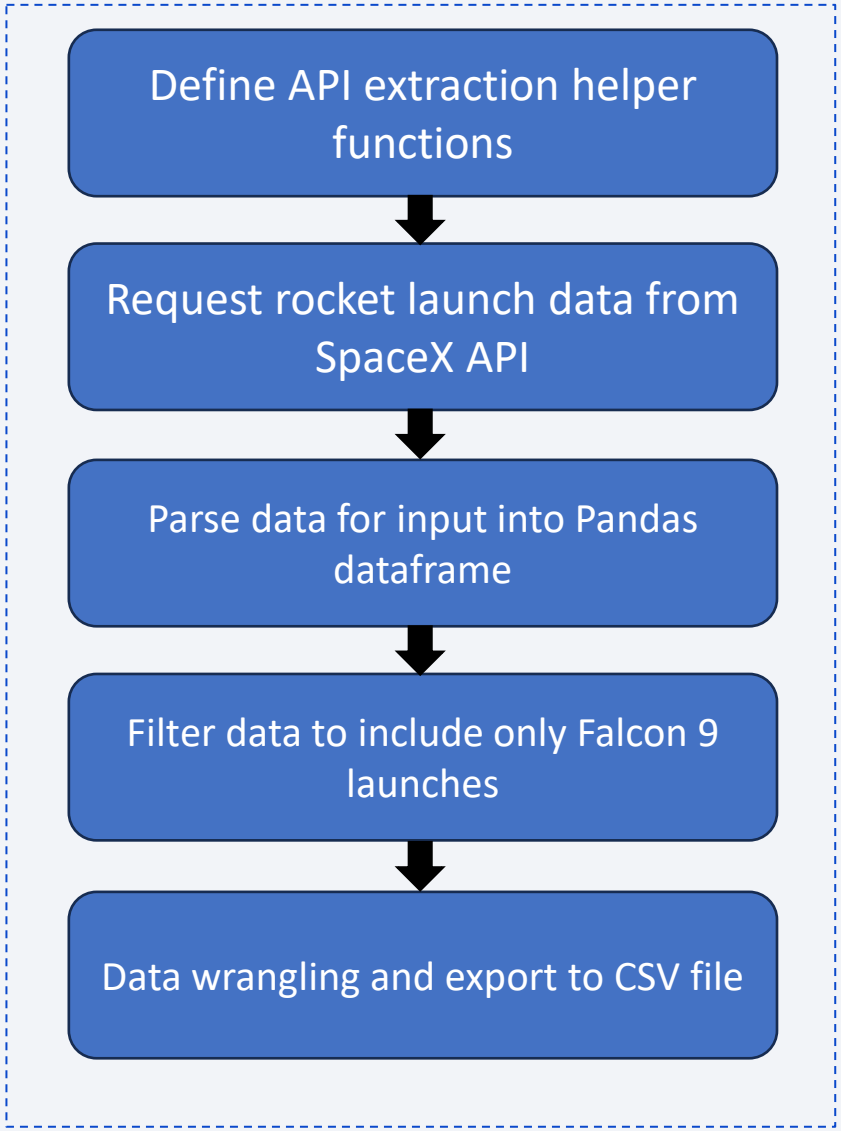
- Data sources
- Data collection methodology
- Data Wrangling
- Exploratory data analysis (EDA) and data visualization
- Interactive visual analytics
- Predictive analysis using classification models

Data Collection – Data Sources

- Data used was obtained from SpaceX's website via the SpaceX REST API (URL https://api.spacexdata.com/v4)
- SpaceX data obtained related to rocket booster name, launch site (including geographical coordinates), payload mass, orbit and launch outcomes
- Additional data was also obtained through scraping of a related Wikipedia page, "List of Falcon 9 and Falcon Heavy launches" (URL https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Launch data located in a HTML table on the page was used, which included data on flight date and time, rocket booster used, launch site, payload, orbit, customer and launch outcome
- SpaceX launches were isolated from the Wikipedia page for this analysis

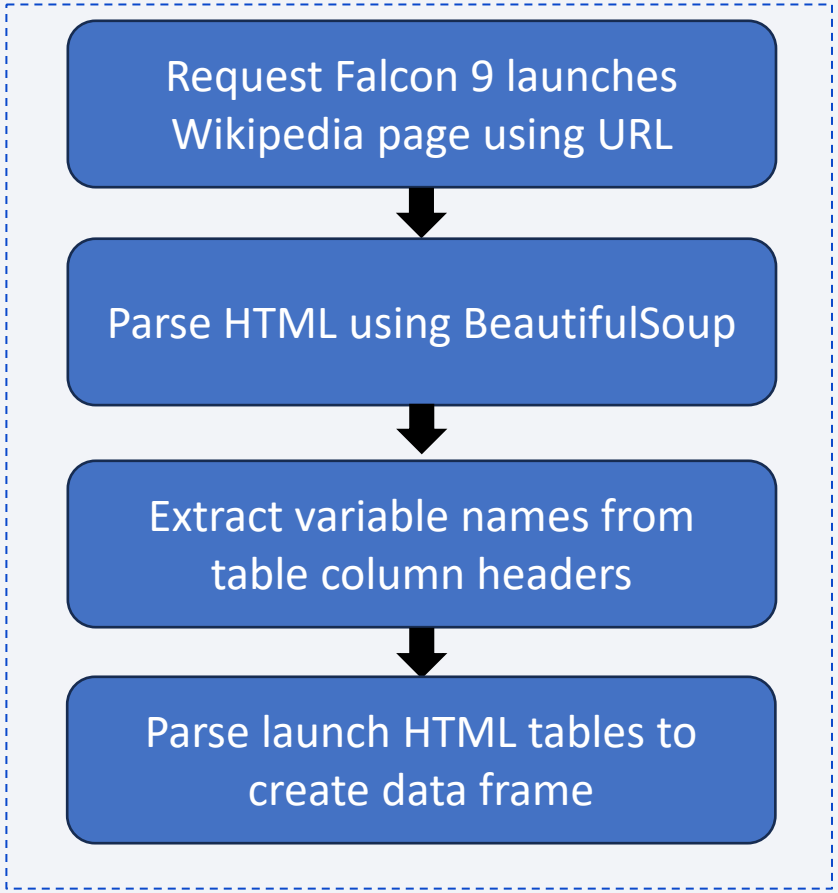
Data Collection – SpaceX API Flowchart

- Illustrated is the process for collecting and cleaning SpaceX Falcon 9 launch data.
- The full Python notebook detailing libraries used, API request, data parsing and wrangling methods can be found here



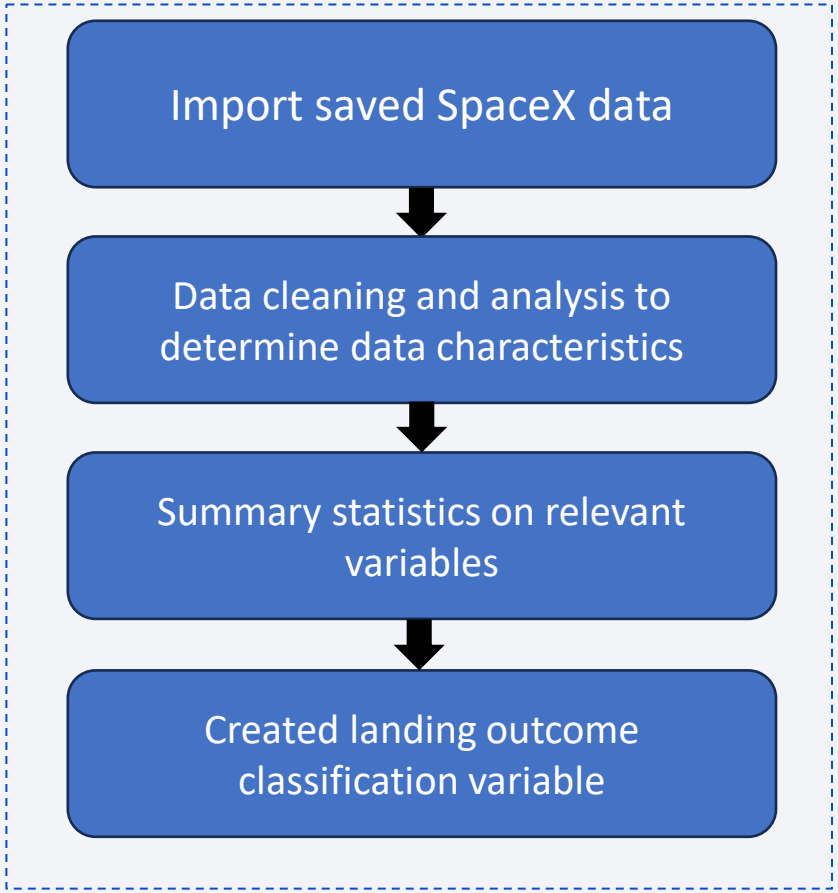
Data Collection – Scraping Flowchart

- Illustrated is the process for scraping and transforming data on historical Falcon 9 launches
- The full Python notebook containing libraries and functions used, data scraping and parsing methods can be found here



Data Wrangling Flowchart

- Once the SpaceX data CSV file was imported, data was examined to determine data types and missing values (including proportion of missing values) for all variables.
- Counts were calculated for each launch site location, orbit type and mission outcome.
- Landing outcomes were classified based on whether they were successful or unsuccessful
- The full Python notebook with all data wrangling steps can be found here



EDA and Data Visualization

Exploratory Data Analysis (EDA) conducted and data visualizations created to get insight into relationships between key variables and prepare for feature engineering (Python notebook here)

Visualizations:

- Launch outcomes by flight number and payload mass: how does payload mass affect launch success over time?
- Launch outcomes by flight number and launch site: how did launch success change over time at each site?
- Launch outcomes by payload mass and launch site: what payload masses were launched at each site, and how did likelihood of launch success change at different payloads for each site?
- Success rate by orbit type: Which orbits had the highest launch success rate?
- Launch outcomes by orbit: how does launch success change over time for each orbit type?
- Launch outcomes by payload mass and orbit type: how does payload mass affect launch successes for each orbit type?
- How did launch successes change over time?

Feature Engineering:

- Determine relevant features for predictive modelling based on visualizations
- Create dummy variables for categorical variables (One Hot Encoding)
- Change numeric columns to float (float64)

EDA with SQL

- SQL queries were written using SQL magic in Python
- Data exploration included searching for launch site names, calculating payload mass totals for different launch scenarios and summarizing landing outcomes
- Queries were also used to select specific records, to examine data features at a glance
- The full Python notebook can be found here

Interactive Visual Analytics: Folium Maps

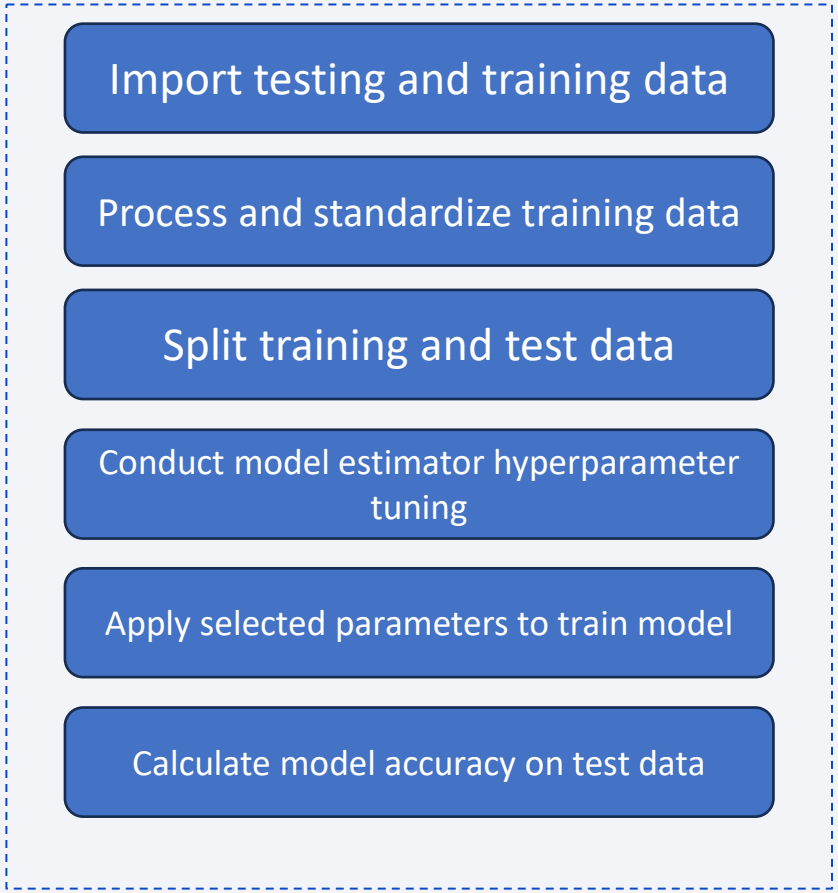
- Interactive maps of all launch sites were developed with Folium
- Circles and text labels were used to mark each site with its name
- Additional markers were added to display launch outcomes at each site, colour coordinated to indicate successful (green) or failed (red) landings
- To illustrate the relative proximity of one site (VAFB) to its surroundings, line markers were added with distance calculations to show the nearest natural and man-made land features
- The full Python notebook can be viewed here (Python notebook is stored here but maps cannot be rendered)

Interactive Visual Analytics: Plotly Dash

- Plotly Dash was used to build a dashboard illustrating landing outcomes by launch site
- A pie chart calculated the proportion of successful landings by launch site, with a further drill-down to illustrate the proportion of successful and failed landings at each site
- A scatter chart illustrated trends in launch outcomes by payload mass and booster version, which could be filtered via a payload mass slider to isolate trends for different ranges
- The Plotly Dash app code can be viewed here

Predictive Analysis (Classification)

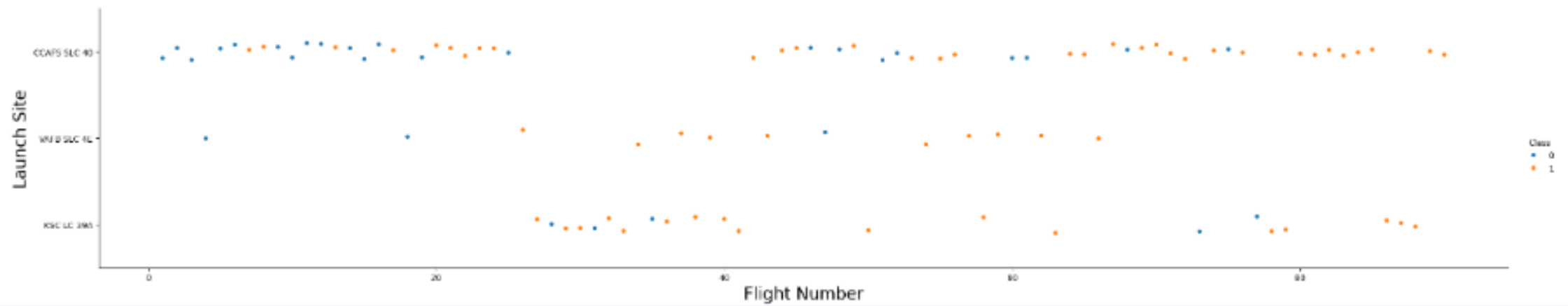
- SpaceX data was used to develop training and test datasets to train four types of classification models: Logistic Regression, Support Vector Machines (SVM), Decision Trees and K-Nearest Neighbors (KNN)
- Hyperparameter tuning was conducted to determine the most accurate parameters for each model
- Accuracy scores and confusion matrices were calculated for all trained models after testing with test data, to determine the best model type for predicting launch outcomes
- The full Python notebook can be found here



Section 2

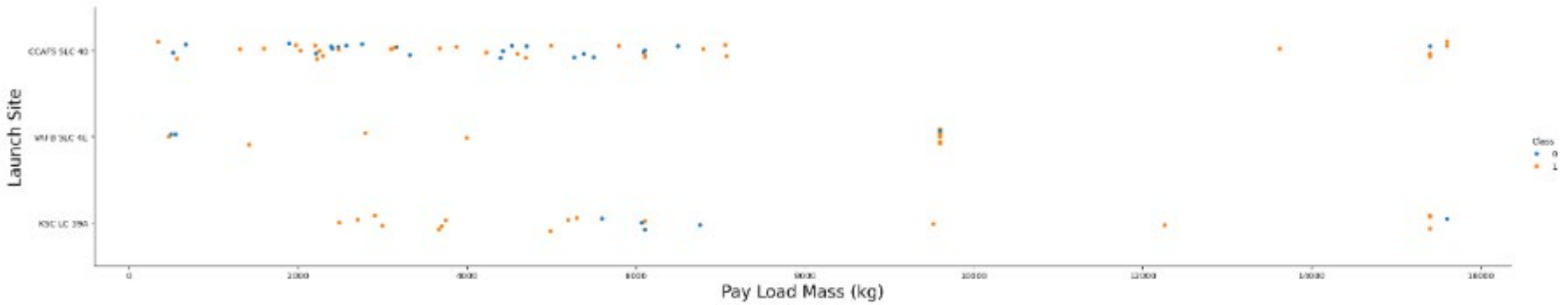
Insights drawn from EDA

Flight Number vs. Launch Site



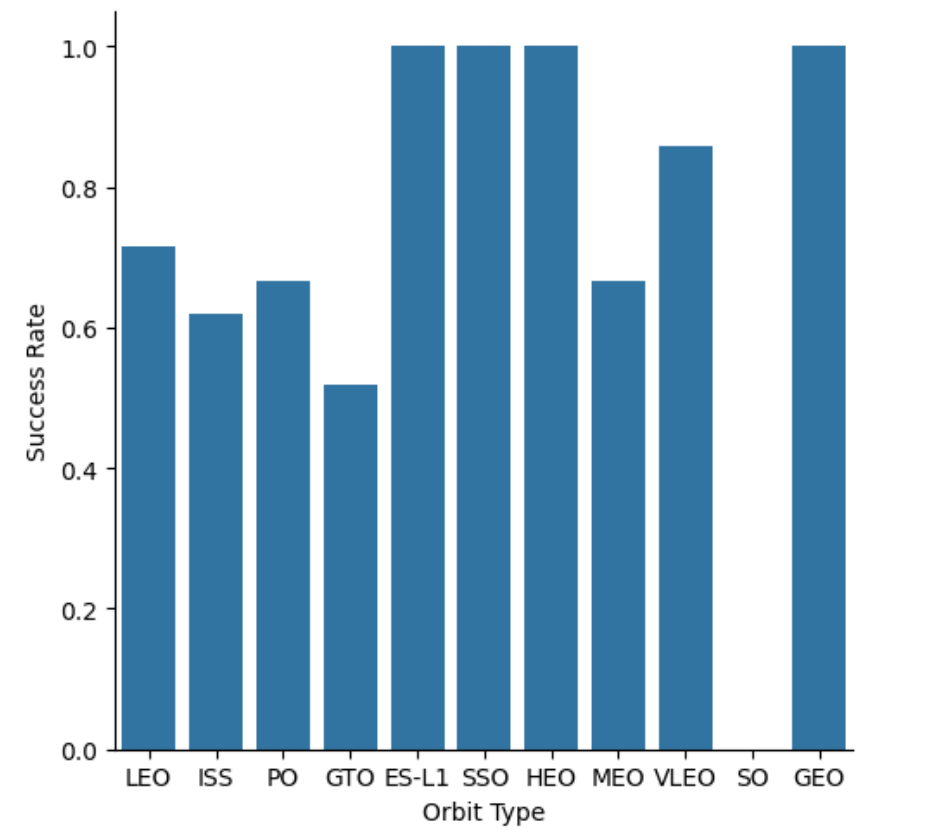
- Likelihood of launch success increased over time
- Most flight launches at VAFB and KSC were conducted after 25 or so launches were completed at CCAFS
- CCAFS had the most flight launches, and VAFB the least
- In terms of number of launches over time, CCAFS is the busiest launch site

Payload vs. Launch Site



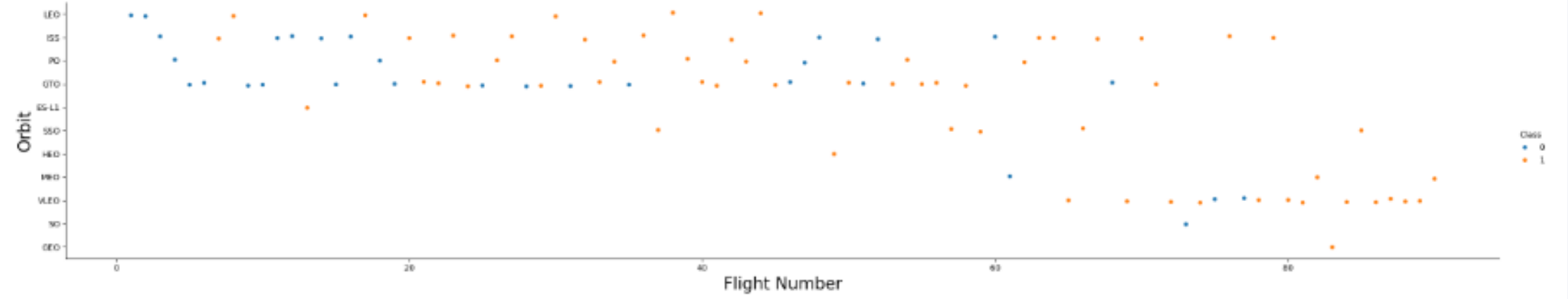
- Most launches at all sites had a payload mass less than 10000
- Higher payload mass is associated with higher likelihood of launch success at CCAFS and VAFB

Success Rate vs. Orbit Type



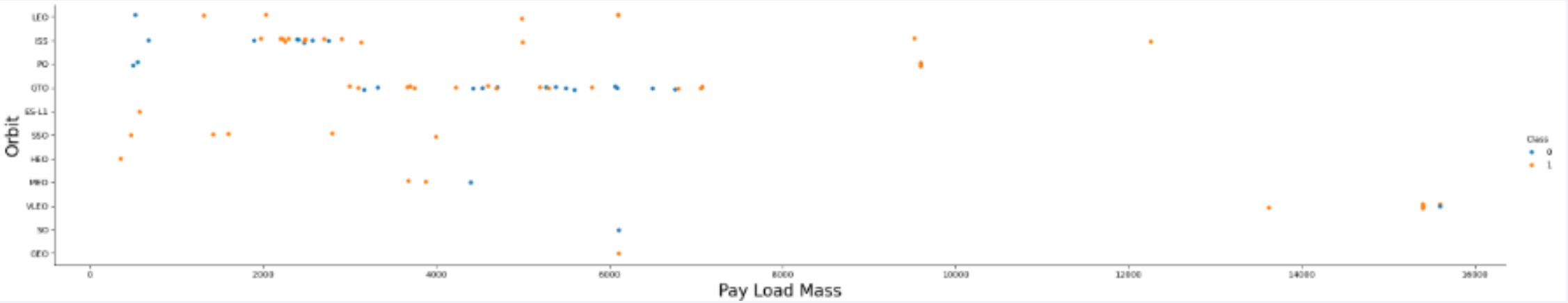
- GEO, HEO, SSO and ES-L1 have the highest success rates of all orbit types (NOTE: only one launch of orbit types GEO, SSO and ES-L1 have been recorded)
- Of the three orbits with the highest frequencies, VLEO has the highest success rate

Flight Number vs. Orbit Type



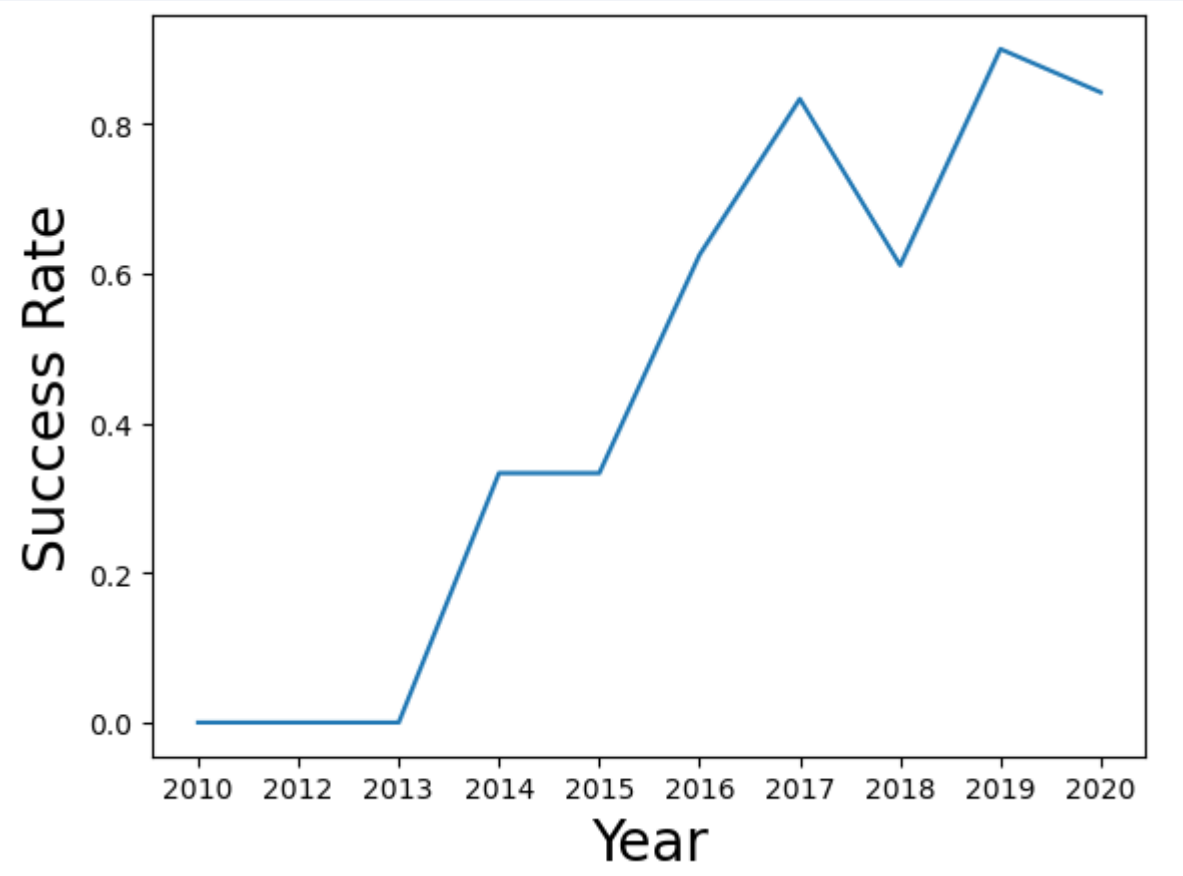
- For LEO, ISS AND PO orbits the more recent flights tend to have successful launch outcomes
- VLEO orbit flights have been launched more recently than the other orbit types, with most being successful

Payload vs. Orbit Type



- Most launches are for payloads under 10000, however increasing successful launches can be seen for LEO, ESS and PO orbits
- There is no clear trend for GTO
- SSO and VLEO have few launches, with the few unsuccessful launches being at higher payloads

Launch Success Yearly Trend



- By 2020 most launches are successful
- The number of launches also continues to increase over time, suggesting that changes to rocket features are likely to have improved likelihood of launch success

All Launch Site Names

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

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

- The above code requests the unique launch site names from SPACEXTABLE, the dataset created in the SQL database containing the SpaceX Falcon 9 launch data
- The code returned the four sites, including the two launch sites located at CCAFS

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

- The above code requests 5 records from sites beginning with ‘CCA’, to bring 5 records from SPACEXTABLE on rockets launched at CCAFS (useful for getting a quick glimpse of the data)
- The 5 records returned show no successful landing outcome for the launches

Total Payload Mass

Total_Payload_Mass
45596

```
%sql select sum(cast("PAYLOAD_MASS__KG_" as numeric)) as "Total_Payload_Mass" from SPACEXTABLE where Customer='NASA (CRS)'
```

- The above code calculates the total payload mass across all records in SPACEXTABLE for Falcon rockets launched by NASA (CRS)
- The payload mass variable was converted to a numeric variable for the calculation

Average Payload Mass by F9 v1.1

Avg_Payload_Mass
2928.4

```
%sql select avg(cast("PAYLOAD_MASS__KG_" as numeric)) as "Avg_Payload_Mass" from SPACEXTABLE where "Booster_Version"='F9 v1.1
```

- The above calculates the average payload mass carried across all Falcon F9 launches with booster version 1.1
- The payload mass variable was converted to a numeric format for calculation
- The average payload mass across all launches was just under 3000 kg

First Successful Ground Landing Date

First Sucessful Ground Pad Landing
2015-12-22

%sql select min(Date) as "First Sucessful Ground Pad Landing" from SPACEXTABLE where "Landing_Outcome"='Success (ground pad)'

- The above code selects the record with the earliest landing date out of all records of successful drone ship landings where payload mass carried was between 4000-6000 kg
- The payload mass variable was converted to numeric for records selection
- The first launch to result in a successful Ground Pad landing was conducted at the end of 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select distinct "Booster_Version" from SPACEXTABLE
where "Landing_Outcome"='Success (drone ship)' and cast("PAYLOAD_MASS__KG_" as numeric) between 4000 and 6000
```

- The above code brings the names of the boosters used for Falcon launches with a successful drone ship landing, carrying between 4000-6000kg payloads
- The payload mass variable was converted to numeric for records selection
- Only four booster versions were associated with successful drone ship landings for the above payload range

Total Number of Successful and Failure Mission Outcomes

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

```
%sql select "Mission_Outcome", count(distinct Date) as "Outcomes" from SPACEXTABLE group by "Mission_Outcome"
```

- The above code counts all Falcon 9 launches in the dataset and categorizes them by their outcome.
- Dates were counted as each date corresponds to a record of a launch
- Most mission outcomes were successful, however the payload status of one successful launch could not be determined
- The above findings illustrate a possible data cleaning requirement for Mission_Outcome

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

```
%sql select distinct "Booster_Version", "PAYLOAD_MASS__KG_" from
SPACEXTABLE where "PAYLOAD_MASS__KG_" =
        select max(cast("PAYLOAD_MASS__KG_" as numeric)) from
        SPACEXTABLE)
```

- The above code selects the Falcon booster versions from all launch records that have carried the largest recorded payload
- A subquery was used to determine the maximum payload mass launched
- 12 booster types have been used for launches with the maximum payload mass recorded

2015 Launch Records

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

```
%sql select case substr(Date, 6,2)
when '01' then 'Jan'
when '02' then 'Feb'
when '03' then 'Mar'
when '04' then 'Apr'
when '05' then 'May'
when '06' then 'Jun'
when '07' then 'Jul'
when '08' then 'Aug'
when '09' then 'Sep'
when '10' then 'Oct'
when '11' then 'Nov'
when '12' then 'Dec'
else substr(Date, 6,2) end as "Month", "Landing_Outcome", "Booster_Version",
"Launch_Site" from SPACEXTABLE
where "Landing_Outcome"='Failure (drone ship)' and substr(Date,0,5)='2015'
```

- This code selects records showing the month, booster version and launch site for failed drone ship launches in 2015
- Substr() was used to isolate the months from the launch date variable (returned month as a number from 01 to 12)
- All month numbers were changed to names for the printed table
- Two launches in 2015 were failed drone ship landings

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

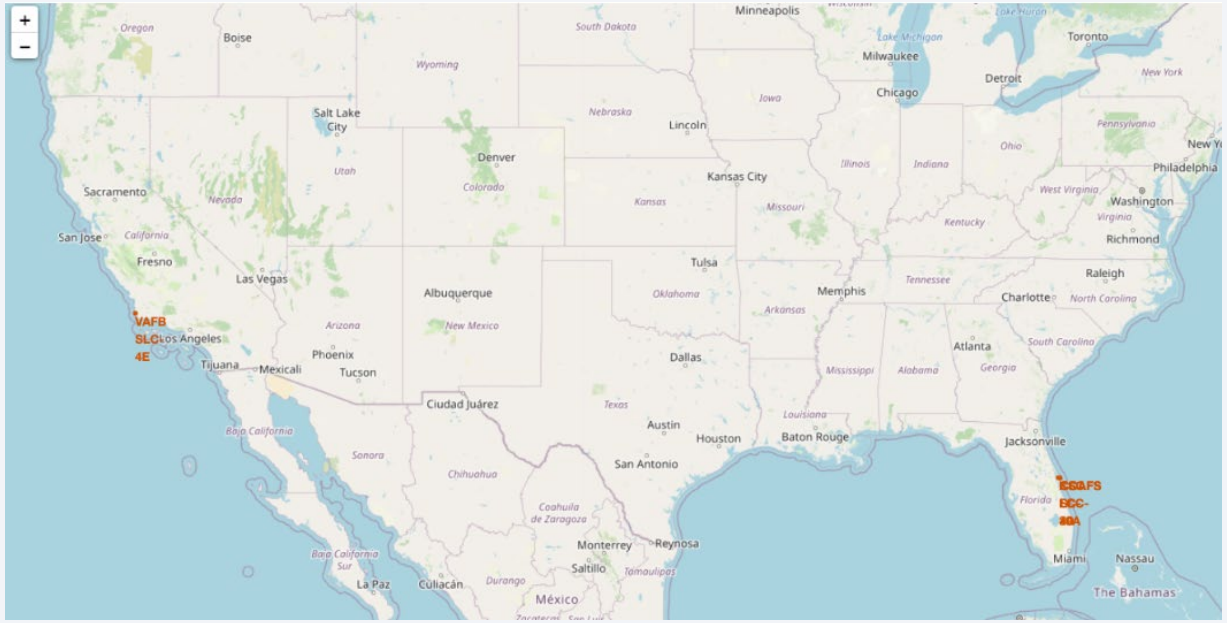
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

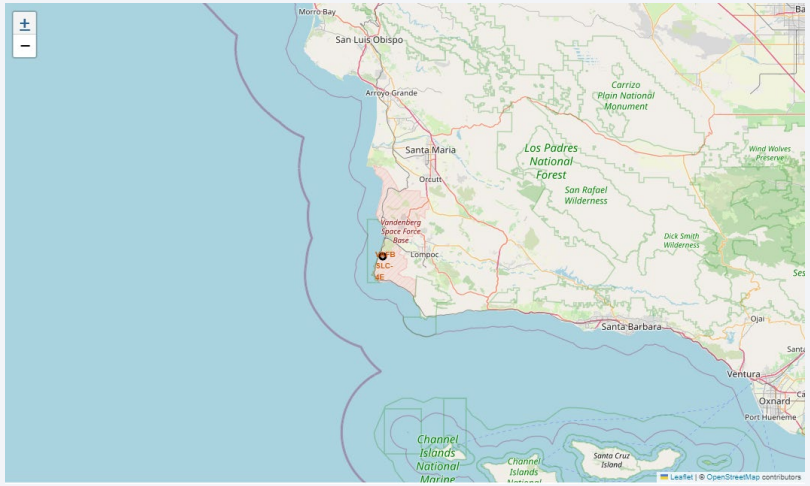
```
%sql select "Landing_Outcome", count(distinct Date) as "Count" from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20' group by "Landing_Outcome"
order by count(distinct Date) desc
```

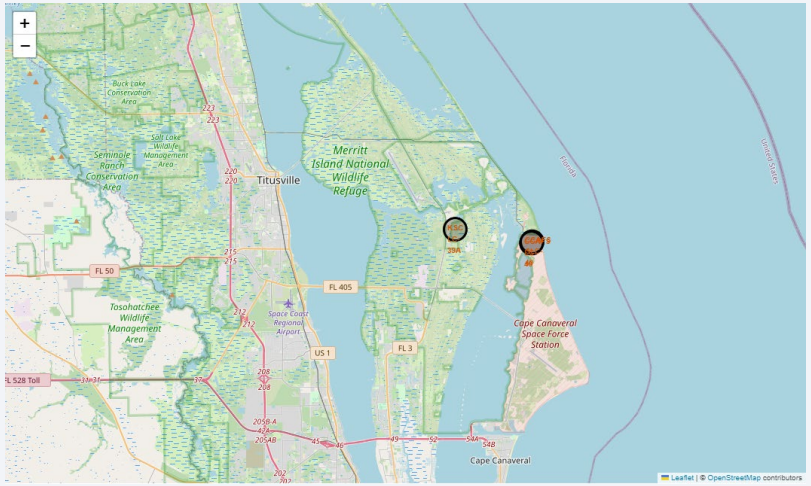
- The above code counts all landing outcomes for Falcon 9 rockets launched between 6 April 2010 and 20 March 2017, by launch outcome
- Results were ranked in descending order by the number of launches per outcome
- Most landings were not attempted, and drone ships had the most successful landings


Section 3
Launch Sites
Proximities Analysis

Falcon 9 Launch Sites Map

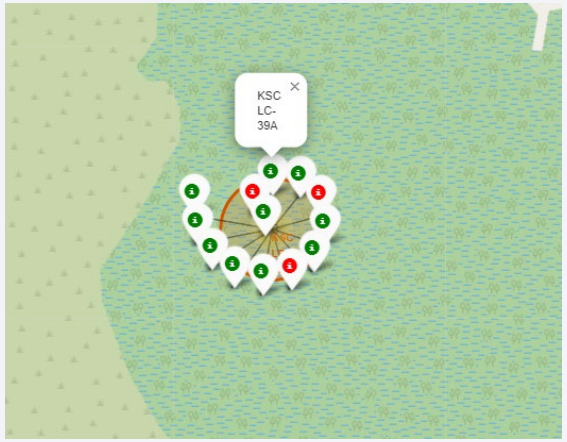


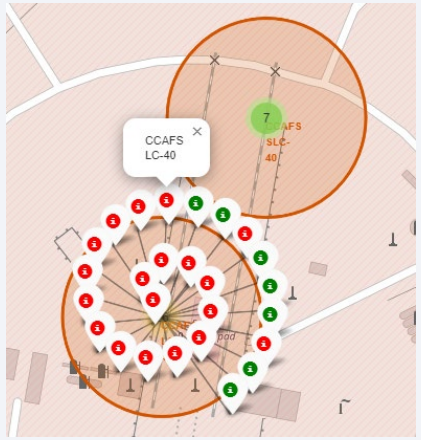




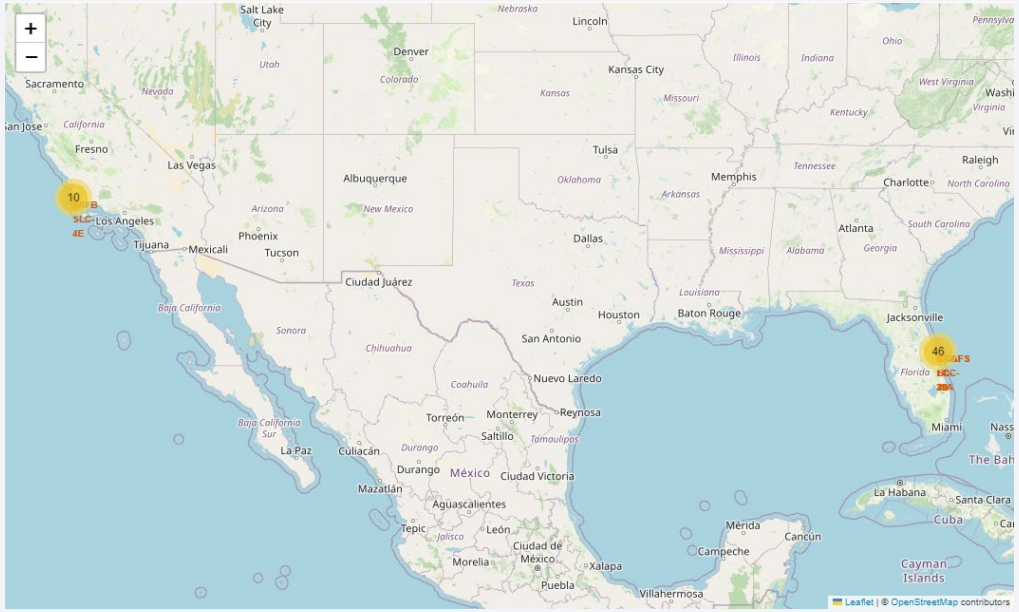
- The above map shows all four Falcon 9 launch site locations in the United States
- The maps to the left show the location of the VAFB (top), KSC and CCAFS sites (bottom)
- Three out of four launch sites are located on the east coast
- KSC is located inside Merritt Island National Wildlife Refuge, while the others are located on space force stations

Falcon 9 Launch Sites Landing Outcomes Map



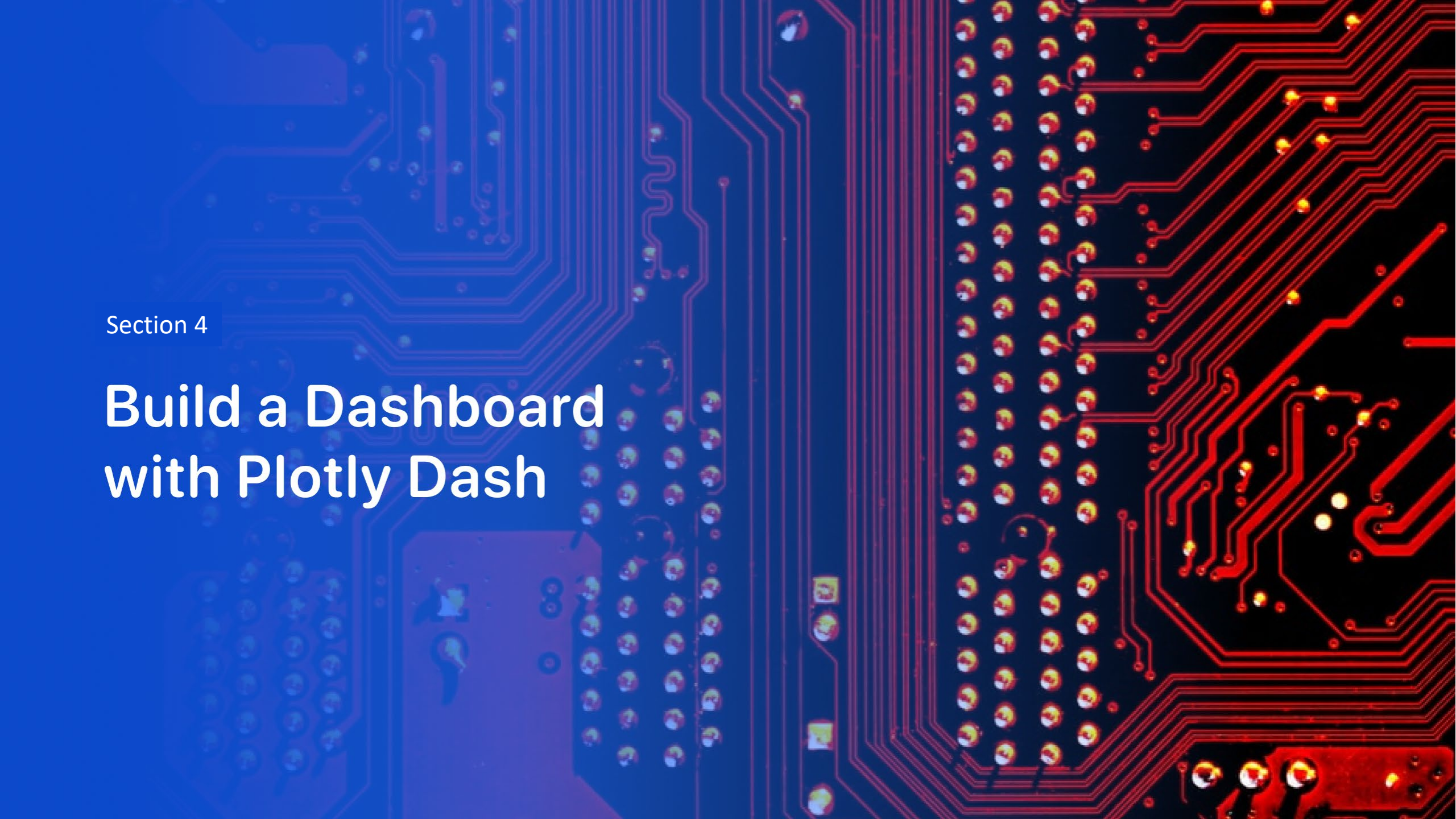


- The above maps show the successful and failed Falcon 9 landings for VAFB (left), KSC (second from left) and CCAFS (second from right and right)
- KSC had the most successful launches
- CCAFS LC-40 had a high number of unsuccessful launches compared to CCAFS-SLC-40
- All launch sites except KSC had more unsuccessful launches than successful ones

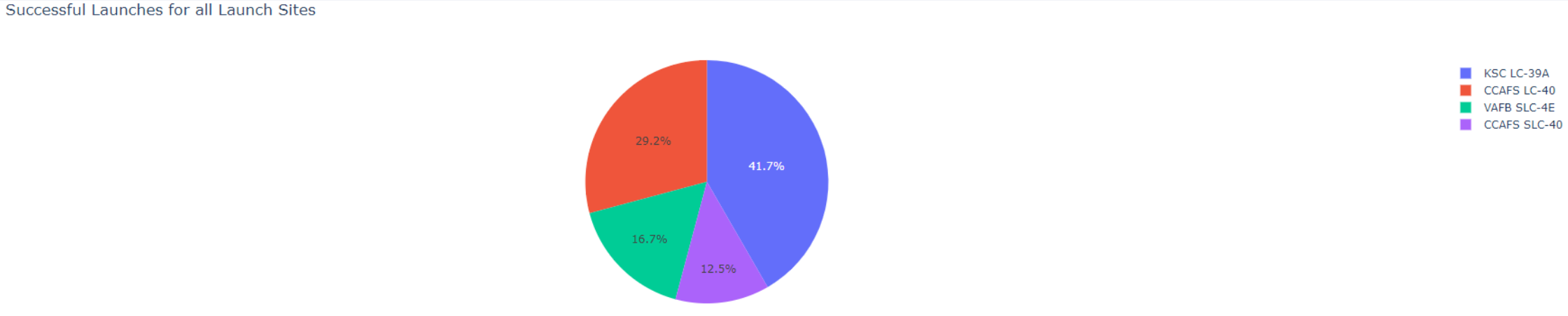


Launch Site Proximity Analysis: VAFB

- VAFB is located near the California coast, north of Los Angeles
- It is in a relatively isolated area surrounded by nature reserves and national parks
- The nearest natural feature to VAFB is the Vandenberg State Marine Reserve, approximately 1.4km west of the site
- The nearest railway line is the UP Santa Barbara Subdivision, approximately 1.25km west of the site

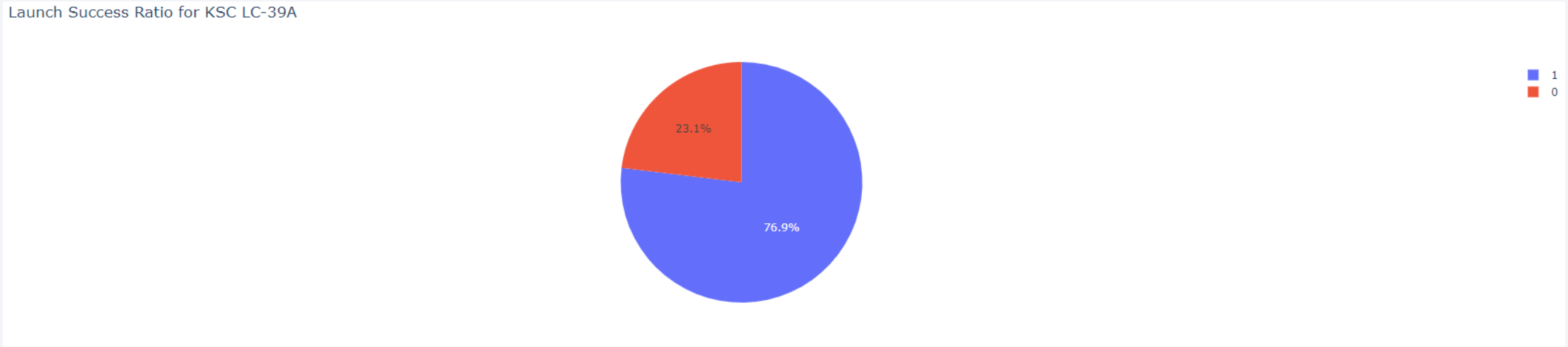

Section 4
Build a Dashboard with Plotly Dash

SpaceX Successful Launches for All Sites



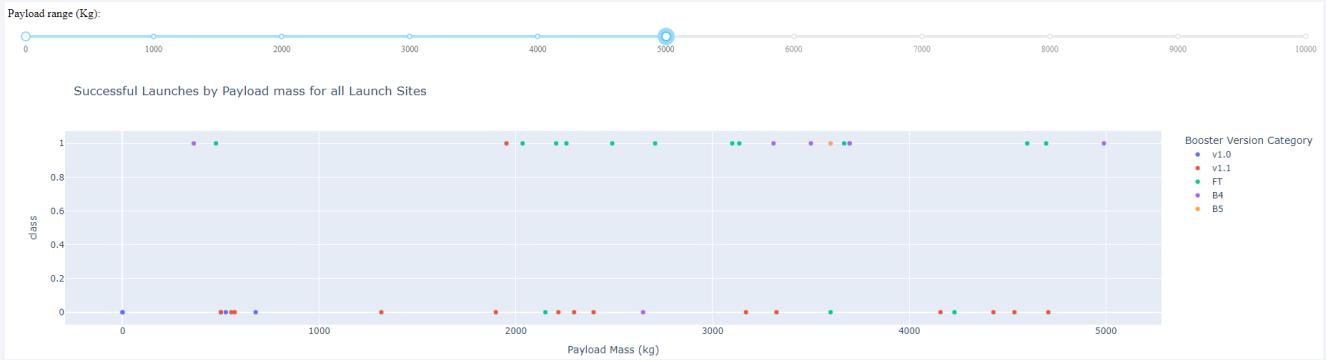
- The above chart shows the proportion of successful launches completed at each site
- KSC accounted for the most successful launches out of all sites at almost 42%
- CCAFS SLC-40 accounted for the least successful launches
- CCAFS LC-40 accounted for a much larger portion of successful launches compared to CCAFS SLC-40

SpaceX Launch Success Ratio for KSC LC-39A



- More than 75% of launches at KSC were successful
- KSC also recorded the least number of unsuccessful launches

Successful Launches by Payload: All Launch Sites

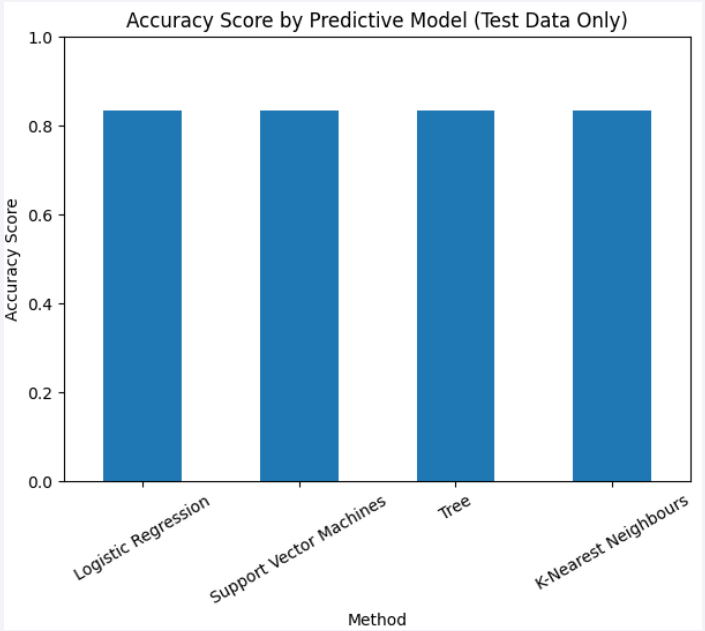


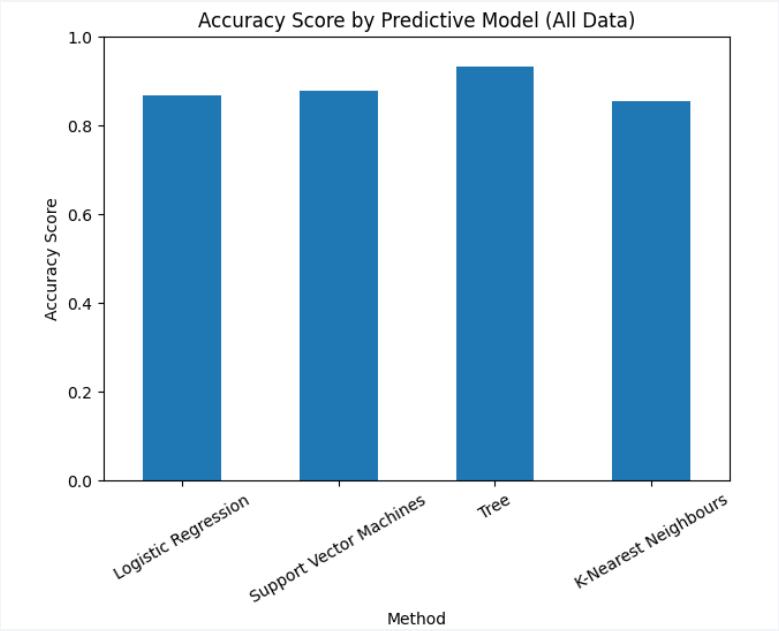
- Displayed is a comparison of successful launches at all launch sites for all payloads (top) for deployed payloads between 0-5000 (middle) and 5000-10000 (bottom), by booster version
- V1.1 has the highest number of unsuccessful launches at all payload weights, FT has the highest number of successful launches
- Most successful launches had payloads between 0 and 5000 kg.

Section 5

Predictive Analysis (Classification)

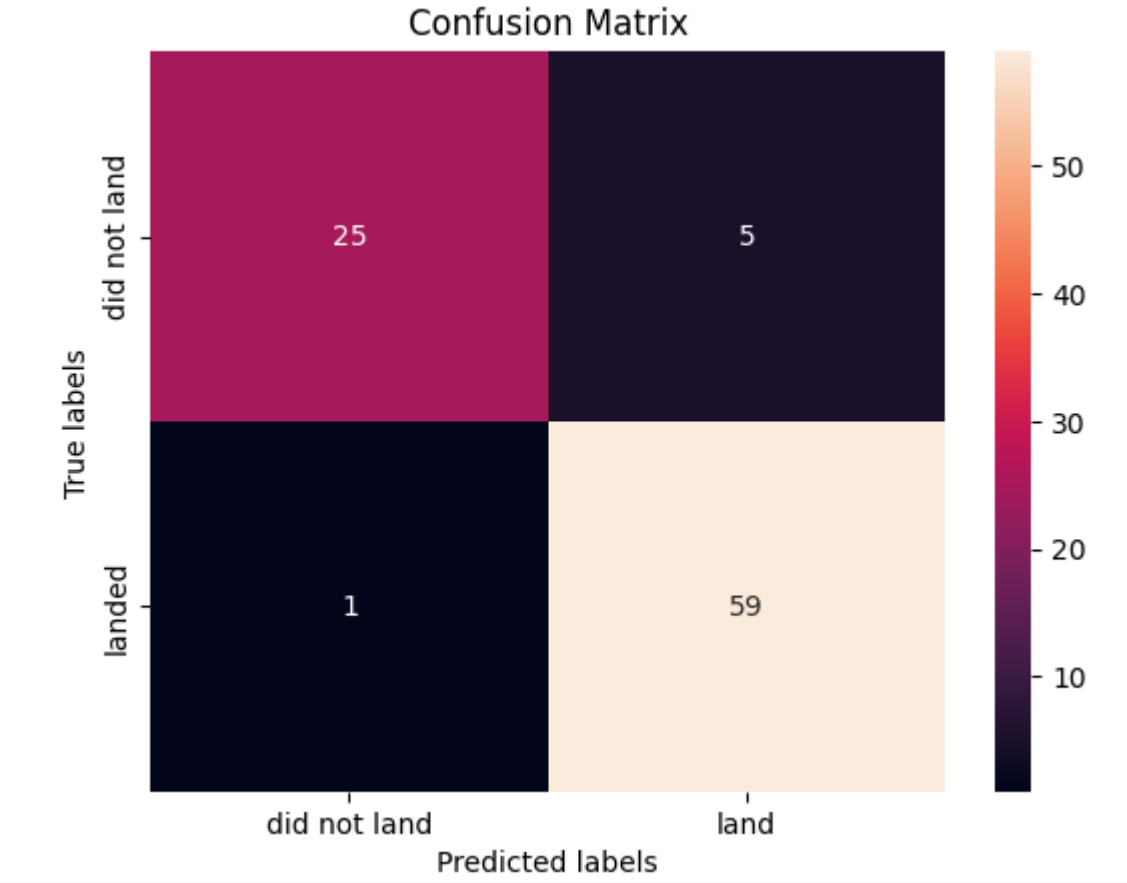
Classification Accuracy





- Four predictive classification models were trained using SpaceX data: Logistic Regression, Support Vector Machines (SVM), Decision Tree and K-Nearest Neighbors (KNN)
- Classification accuracy was high for all models, and the similarity of accuracy scores meant no one model could be identified as ideal for real-world prediction
- Classification accuracy was calculated after running trained models on all data, with Decision Tree scoring the highest classification accuracy score

Confusion Matrix



- The confusion matrix for the Decision Tree model (tested on all data) had the highest true positive classification rate (precision of 98%)
- This model also correctly labeled the most successful launch outcomes (recall of 92%)

Conclusions

- SpaceX data was collected and exploratory data analysis conducted to determine which launch features had the most impact on successful landings
- Successful Falcon 9 landings are influenced by launch site, payload mass carried, flight number (examining landing outcomes over time), rocket and orbit features
- Visualizations showed the impact of payload mass, flight number, rocket booster type, orbit type and launch site on successful landings
- Using the identified launch features, 4 classification models were trained on predicting successful landings
- Decision Trees appears to be the most precise model, as well as having the most accurate recall, for determining whether a Falcon 9 landing is successful
- Decision Trees therefore should be used to accurately forecast potential cost savings on Falcon 9 landings

Thank you!