Winning Space Race with Data Science

Chakrit Phongwithayalert
1 April 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collect Space X data by using SpaceX API and Web Scraping
  - Perform data wrangling using Pandas and EDA using visualization and SQL
  - Create interactive visual analytics using Folium and Plotly Dash
  - Predict analysis using classification models in Scikit-learn library

- Summary of all results
  - The best classification model is Decision Tree Classifier model with accuracy of 94.44%
  - From the confusion matrix, Decision Tree Classifier can distinguish between the different classes, but have false positives as a major problem

Introduction

- Project background

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

In this capstone, I am a data scientist working for a new rocket company named “Space Y” that would like to compete with SpaceX. My responsibilities is to determine the price of each launch by gathering information about Space X and creating dashboards. also determine if SpaceX will reuse the first stage by training a machine learning model and use public information to predict if SpaceX will reuse the first stage

- Problems
  - Which factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.
- Goal
  - Determine if Space Y should reuse the first stage rocket based on machine learning model, trained using Space X data to predict if the first stage will land successfully

Section 1

Methodology

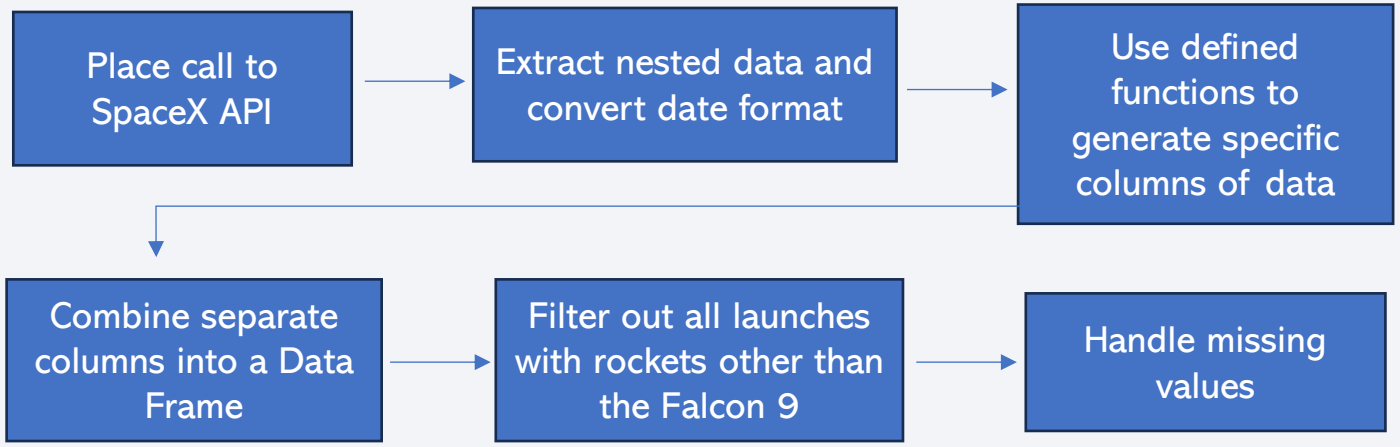
Methodology

Executive Summary

- Data collection methodology:
  - Collecting Space X data by using SpaceX API and Web Scraping.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Analyzing and Cleaning data using Pandas library.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using Scikit-learn library.

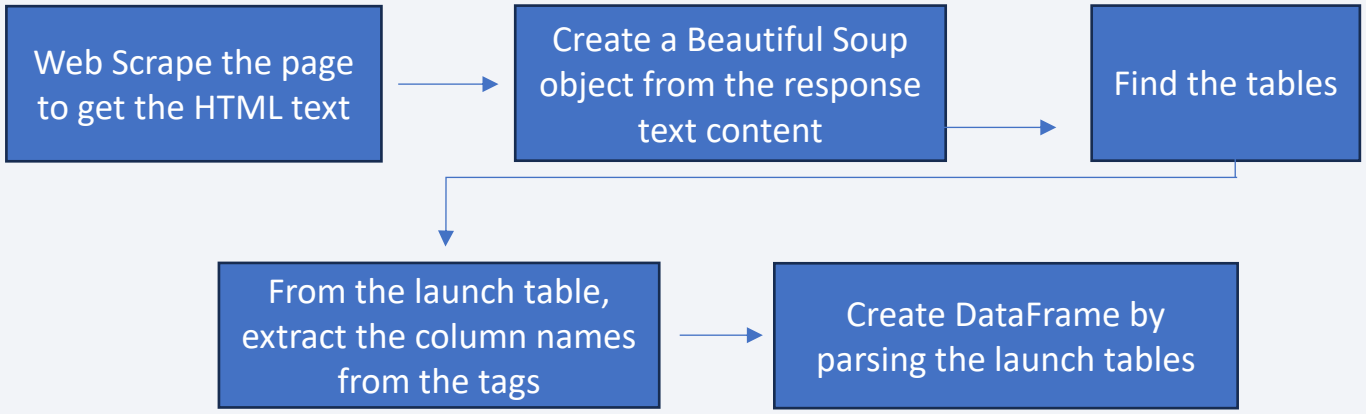
Data Collection – SpaceX API

- The SpaceX API has data available publically.
- Once a GET request has been made to the SpaceX API and the response received, the data can be placed into a Pandas Data Frame for further analysis.
- GitHub URL (Data Collection): https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/1.collection-spacex-data-collection-api.ipynb



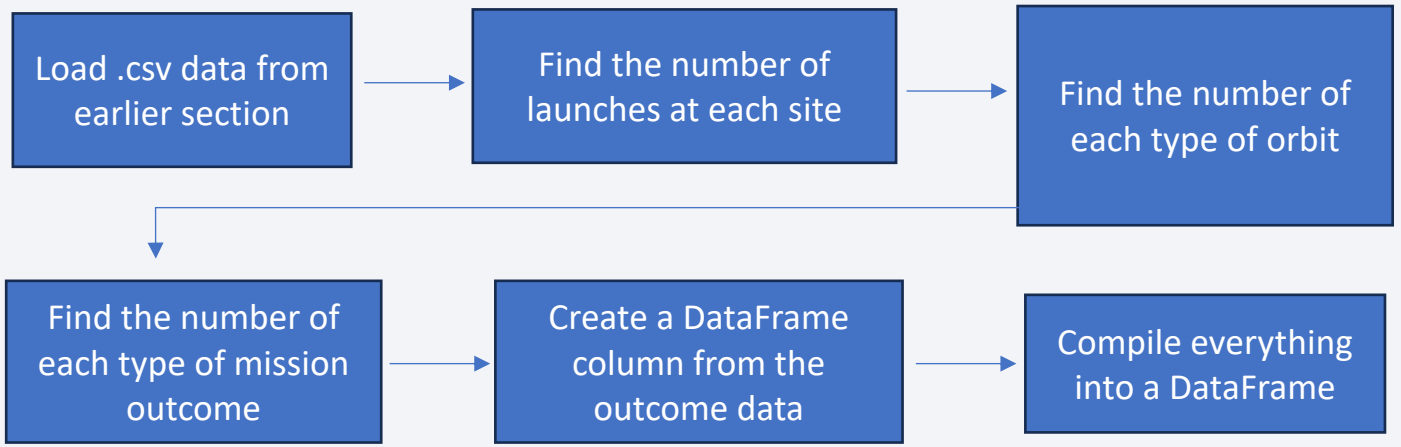
Data Collection - Scraping

- Wikipedia has a page that has tables of data about SpaceX launches.
- These tables can be scraped to extract launch data that can be put into a Pandas DataFrame for further analysis.
- GitHub URL (Web Scraping) https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The .csv file from the first section contains the data that needed to be cleaned.
- The launch sites, orbit types and mission outcomes were cleaned up.
- The handful of mission outcome types were converted to a binary classification where 1 means that the Falcon 9 first stage landing was a success and 0 means that it was a failure.
- The new classification was added to the DataFrame for further analysis
- GitHubURL (Data Wrangling): https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summary
  - Use scatter plot to visualize the relationship between Flight Number, Payload, Launch Site and Orbit type
  - Use bar plot to visualize the relationship between success rate of each orbit type
  - Use line plot to visualize the launch success yearly trend

- GitHub URL (EDA with Data Visualization) :

https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/vizualize-jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- GitHub URL (EDA with SQL):

https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/sql-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Markers were added for launch sites and for the NASA Johnson Space Center
  - Circles were added for the launch sites.
  - Lines were added to show the distance to the nearby features:
    - Distance from CCAFS LC-40 to the coastline
    - Distance from CCAFS LC-40 to the rail line
    - Distance from CCAFS LC-40 to the perimeter road
- GitHub URL (Folium Maps) :

https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-launchsize-folium.ipynb

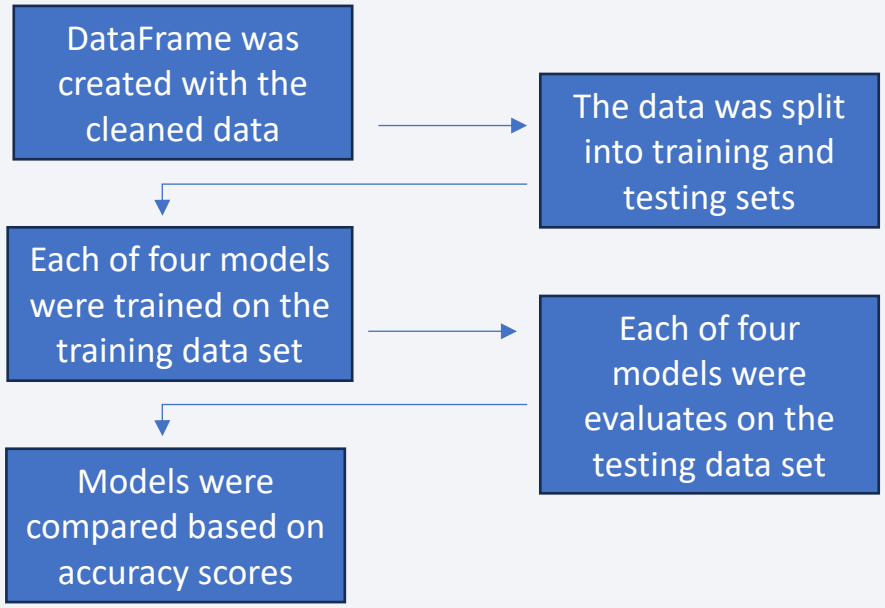
Build a Dashboard with Plotly Dash

- Summarize dashboard
  - Create dashboard with 4 components including dropdown menu, pie chart, slider, and scatter plot.
- Explain plots and interactions
  - Dropdown menu for selecting launch sites
  - Pie chart to visualize success rate in each launch site
  - Slider to select payload range
  - Scatter plot to visualize relationship launch site, payload, and booster version
- GitHub URL (Dashboard File) :

https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/spacex_dash_app.py

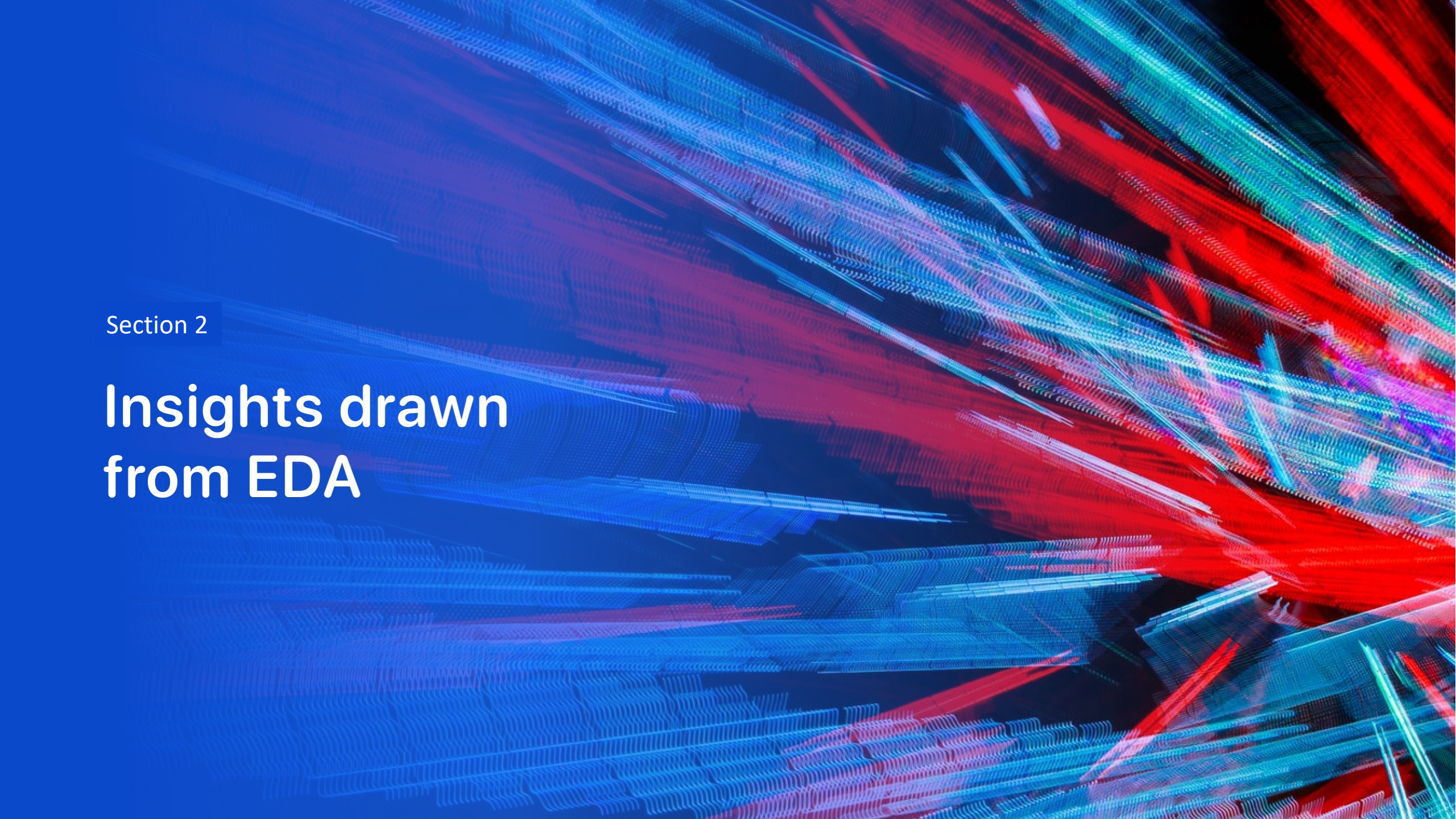
Predictive Analysis (Classification)

- The dataset was split into training and testing sets.
- Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) machine learning models were trained on the training data set.
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using ".best_params_".
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.
- GitHub URL (Machine Learning): https://github.com/chabiw1/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/SpaceX-ML-Prediction.ipynb



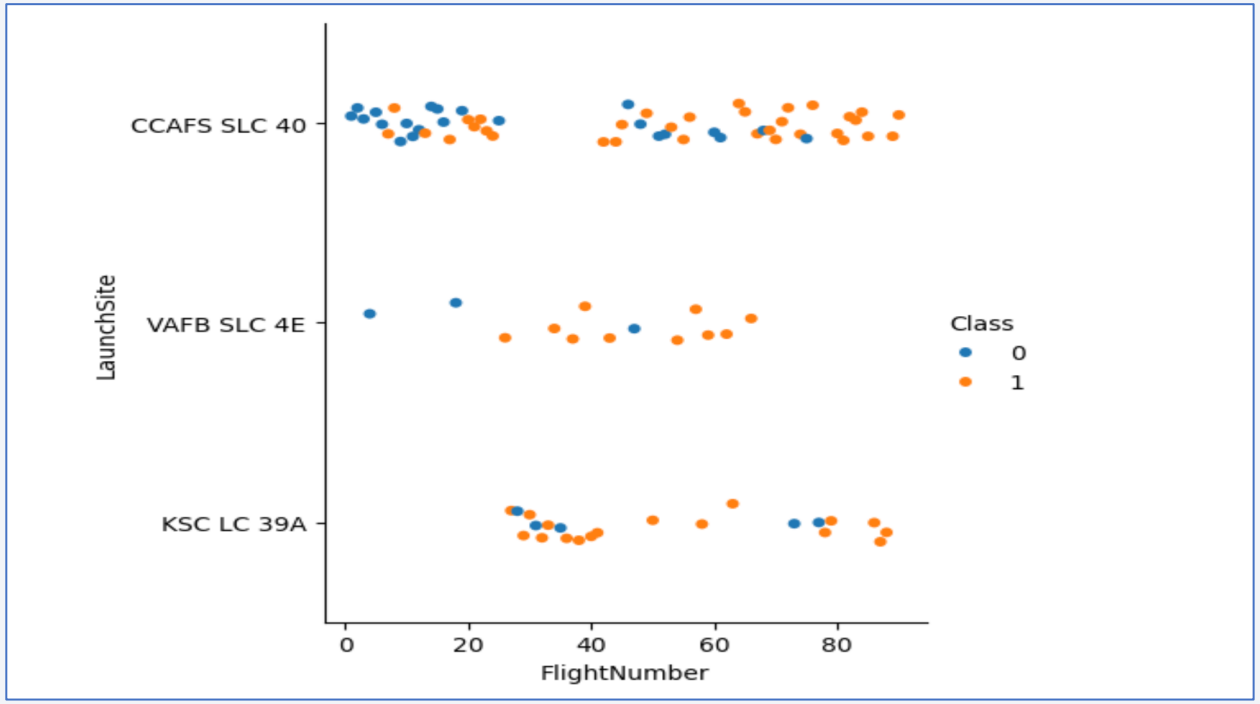
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site

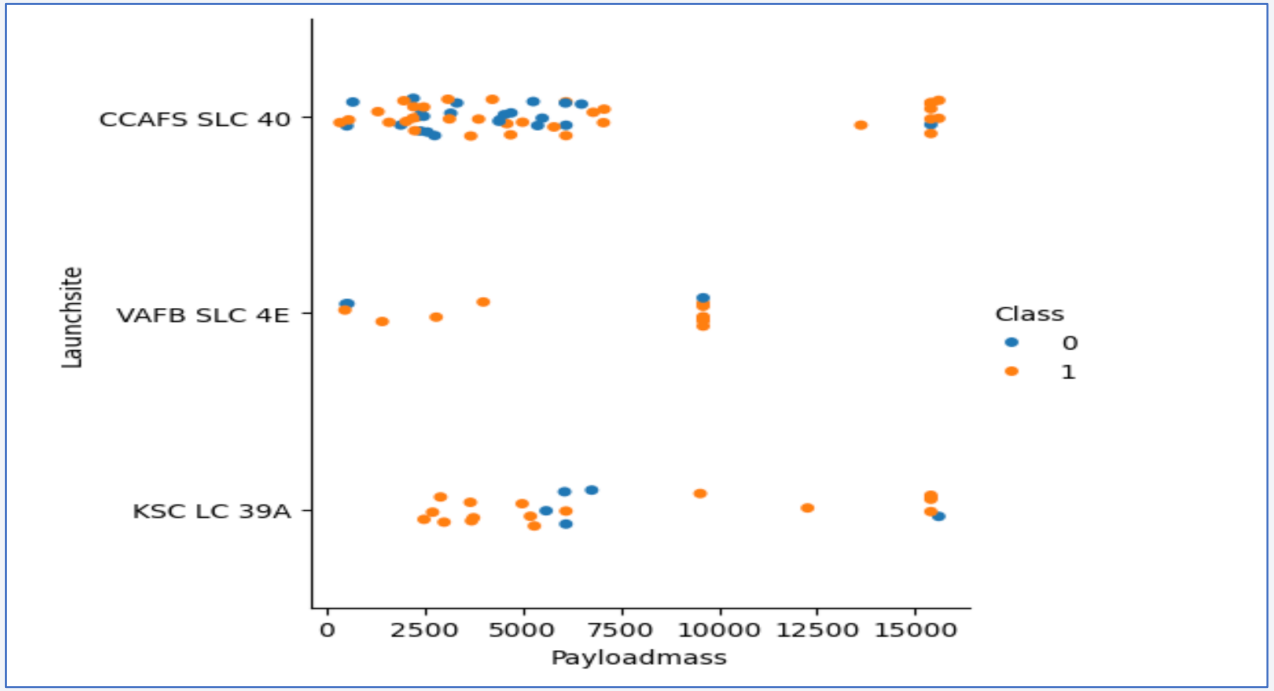


Falcon 9 first stage failed landings are indicated by the '0' Class (● blue markers) and successful landings by the '1' Class (● orange markers).

- Launch Site CCAFS SLC 40 has the most number of launch.
- Class 0 (Failed) is often founded at early flight Number.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

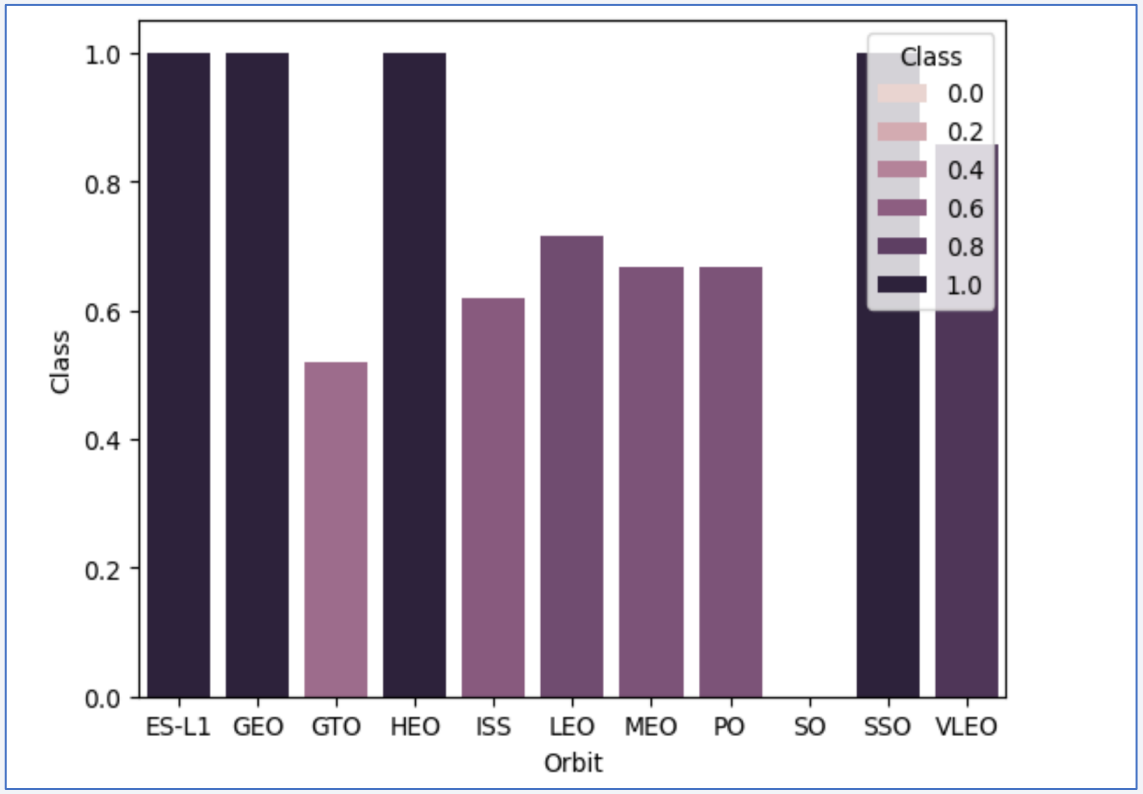


- Launch Site VAFB SLC 4E do not have launch that have pay load mass more than 10,000 kg
- The failed landings at the KSC LC 39A launch site are all grouped around a narrow band of payload masses.

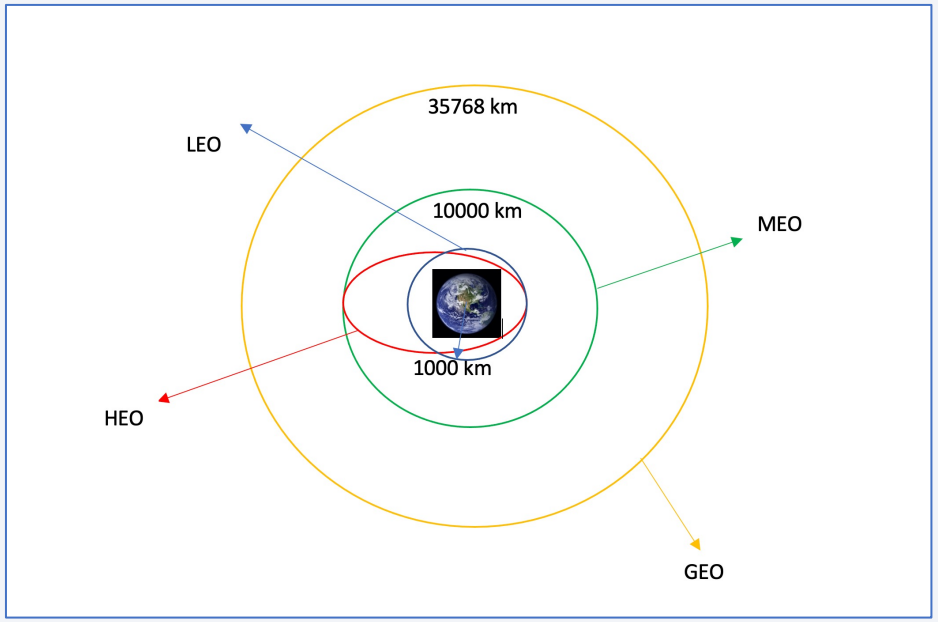
Falcon 9 first stage failed landings are indicated by the '0' Class (● blue markers) and successful landings by the '1' Class (● orange markers).

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type



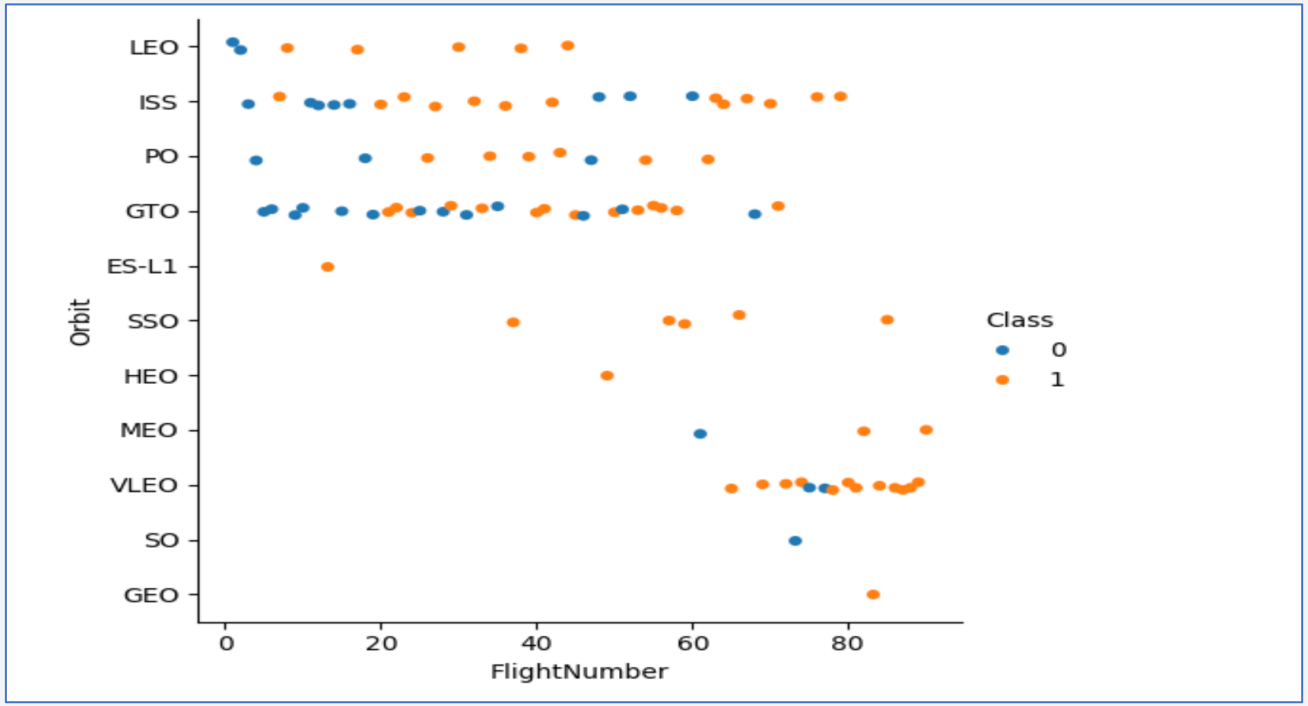
Falcon 9 first stage Landing Success Rate by Orbit Type



- Orbit ES-L1, GEO, HEO, SSO have the best success rate
- SO orbits have no successful first stage landings.

Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type



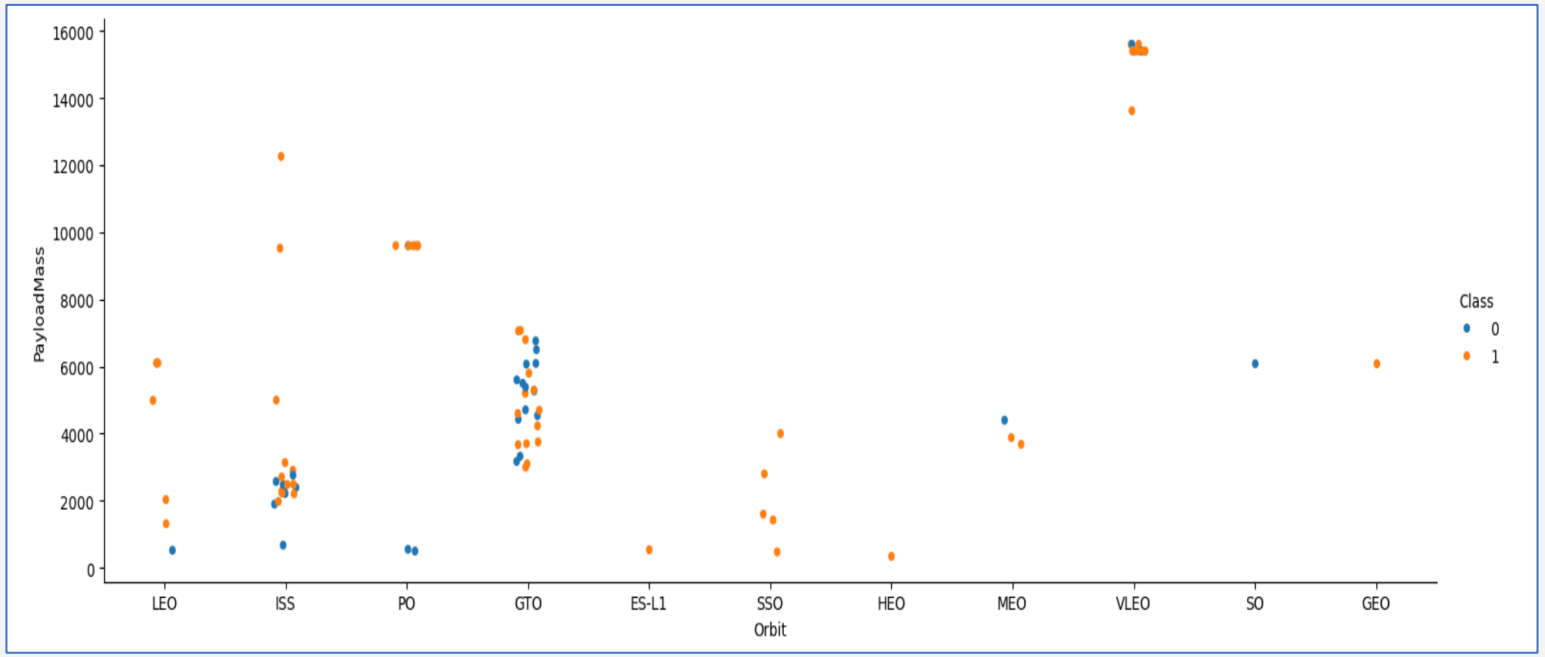
Falcon 9 first stage failed landings are indicated by the '0' Class (● blue markers) and successful landings by the '1' Class (● orange markers).

- There is a correlation between flight number and success rate with larger flight numbers being associated with higher success rates.

- Orbits MEO,VLEO,SO and GEO are founded at flight Number 60

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

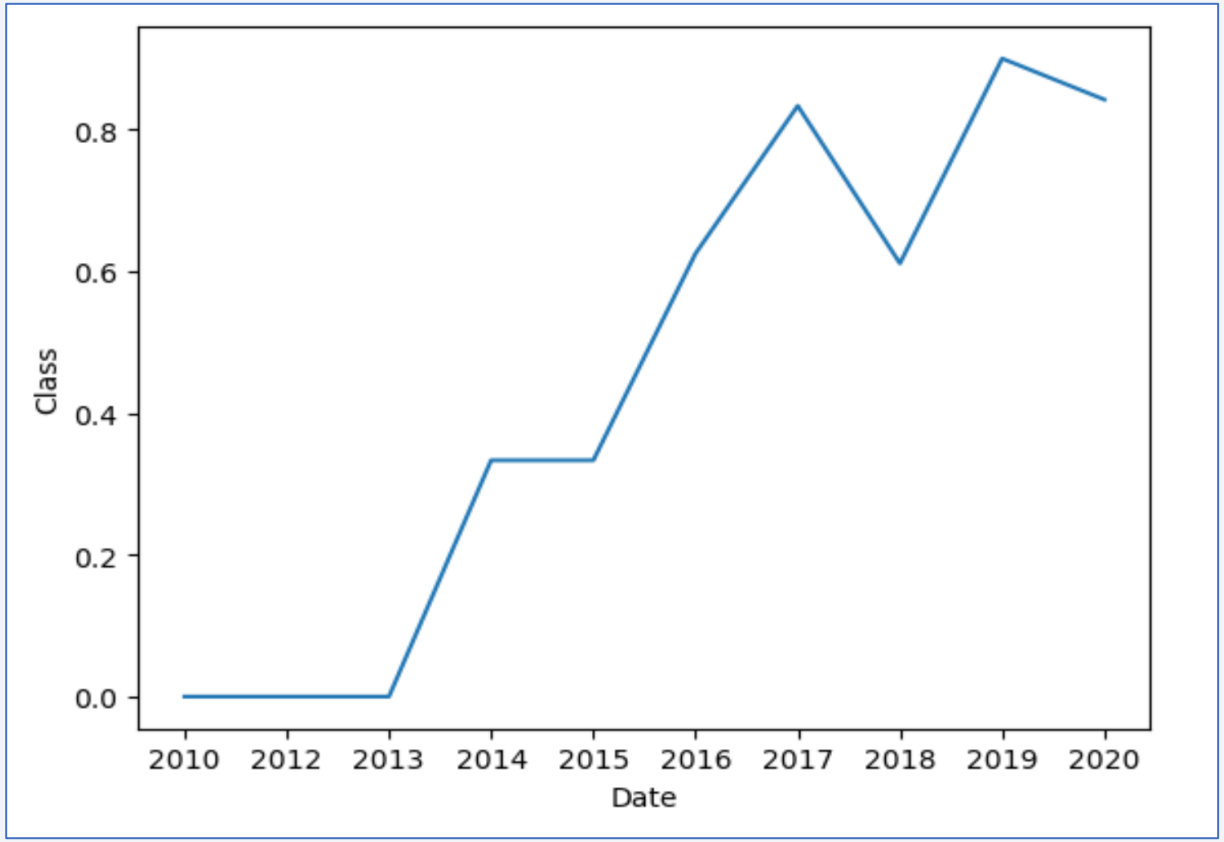


Falcon 9 first stage failed landings are indicated by the '0' Class (● blue markers) and successful landings by the '1' Class (● orange markers).

- Some orbit types have better success rates than others.
- Success rate appears to have no obvious correlation with payload mass.
- Orbit VLEO use the most Payload Mass

Launch Success Yearly Trend

- Line chart of yearly average success rate



Falcon 9 First Landing Success Rate by Year ,Y axis represent success rate

- The success rate has increased significantly over the years.

- Success rate in 2018 has decreased to 60% from 80% success rate in previous year(2017).

All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTBL
```

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

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

- Explanation: There are four unique launch sites.

Launch Site Names Begin with 'CCA'

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

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	M
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

- Explanation: This is a fairly straightforward sampling mechanism used to gain a sense of the data contained in the database table.

Total Payload Mass

```
%sql select *,SUM(PAYLOAD_MASS__KG_) as sumpayload from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

ter_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	sumpayload
v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt	45596

- Explanation: The total payload carried by boosters from NASA (CRS) is 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- Explanation: The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

```
%sql select min(Date),Landing_Outcome from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

- The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version,PAYLOAD_MASS__KG_,Landing_Outcome  from SPACEXTBL where Landing_Outcome is 'Success (
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Explanation: The four booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg are listed above.

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(*),Mission_Outcome from SPACEXTBL where Mission_Outcome like 'Failure%'
```

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

count(*)	Mission_Outcome
1	Failure (in flight)

- Explanation: There were 61 successful and 40 failed mission outcomes.

Boosters Carried Maximum Payload

```
%sql select max(PAYLOAD_MASS__KG_),Booster_Version as 'Booster_Version with max payload' from SPACEXTBL group by
```

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

max(PAYLOAD_MASS__KG_)	Booster_Version with max payload
2647	F9 B4 B1039.2
5384	F9 B4 B1040.2
9600	F9 B4 B1041.2
6460	F9 B4 B1043.2
3310	F9 B4 B1039.1
4990	F9 B4 B1040.1
9600	F9 B4 B1041.1
3500	F9 B4 B1042.1
5000	F9 B4 B1043.1
6092	F9 B4 B1044

- Explanation: The maximum payload mass carried in this dataset is 15,600 kg. Twelve (12) separate Falcon 9 boosters carried this amount of payload mass.

2015 Launch Records

```
%sql select substr(Date,6,2) as month,Landing_Outcome,Booster_Version,Launch_Site as month from SPACEXTBL where L
```

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

month	Landing_Outcome	Booster_Version	month_1
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 landing outcomes with a drone ship in 2015. Both launched from CCAFS LC-40. One occurred in January and the other in April.

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

```
%sql select Landing_Outcome,count(Landing_Outcome) as 'landing number' from SPACEXTBL where Date between '2010-06
```

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

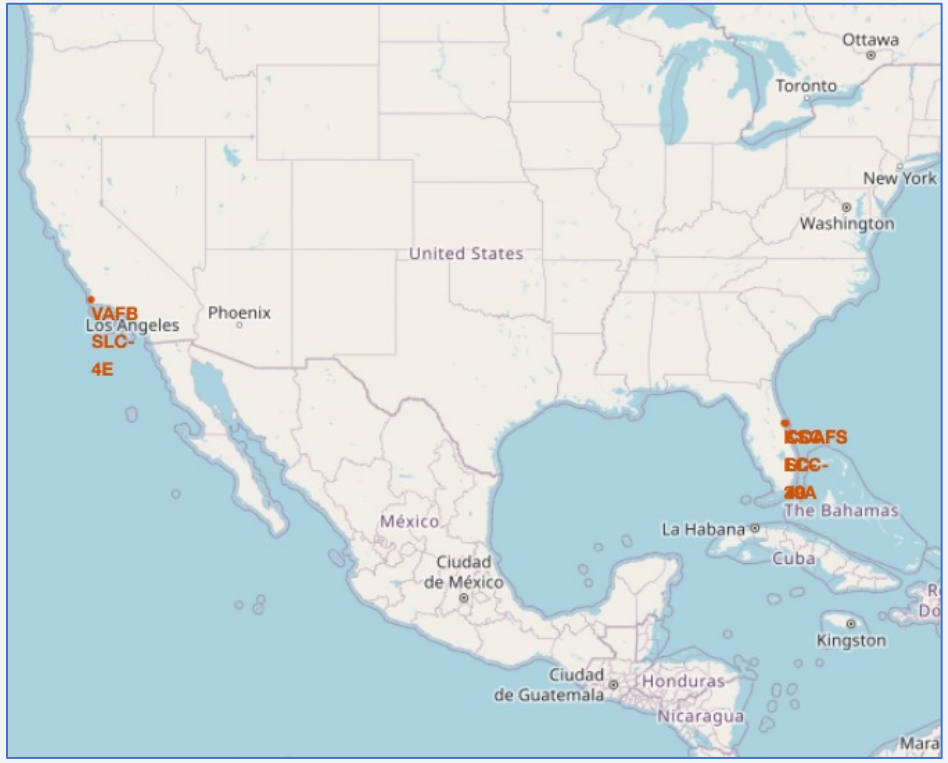
Landing_Outcome	landing number
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Explanation: The most common landing outcome was ‘not attempted’.


Section 3
Launch Sites
Proximities Analysis

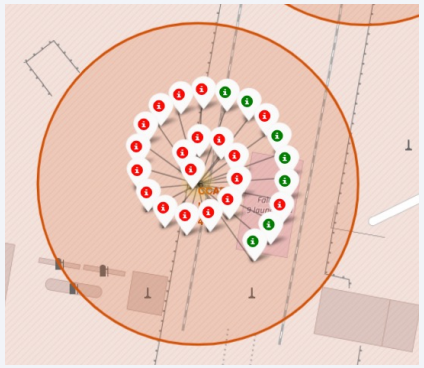
Falcon 9 Launch Site Locations

- VAFB SLC-4E (California, USA) :
  Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA) :
  Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA) :
  Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA) :
  Cape Canaveral Air Force Station Space Launch Complex 40

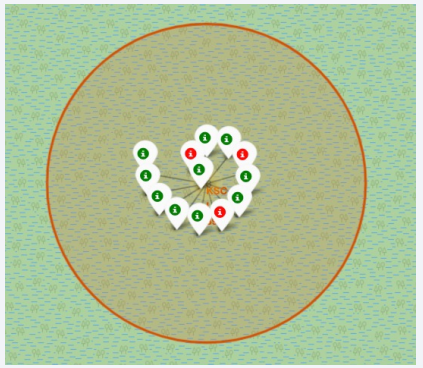


Map Markers of Success/Failed Landings

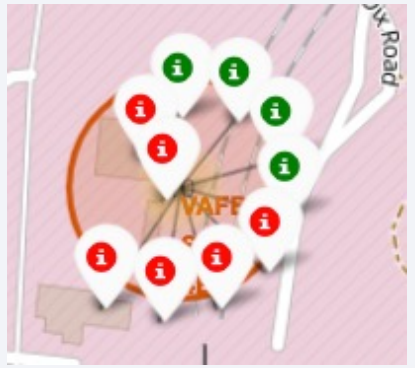
CCAFS SLC-40



CCAFS LC-40



KCS LC-39A

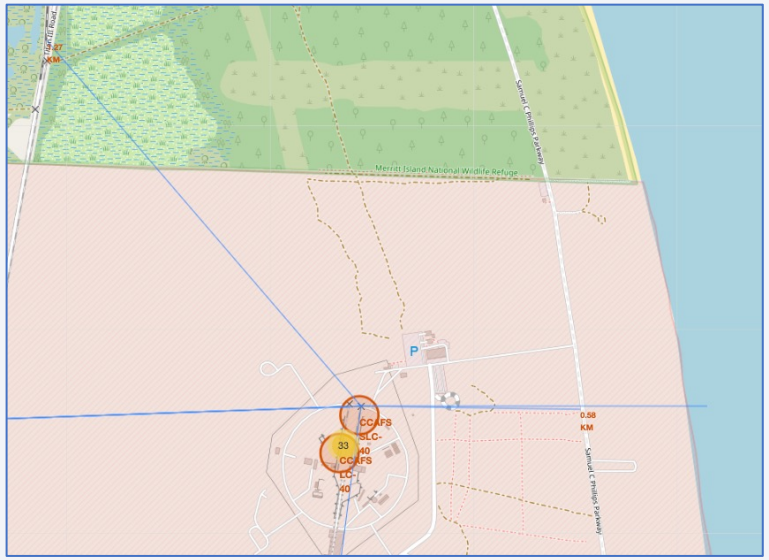


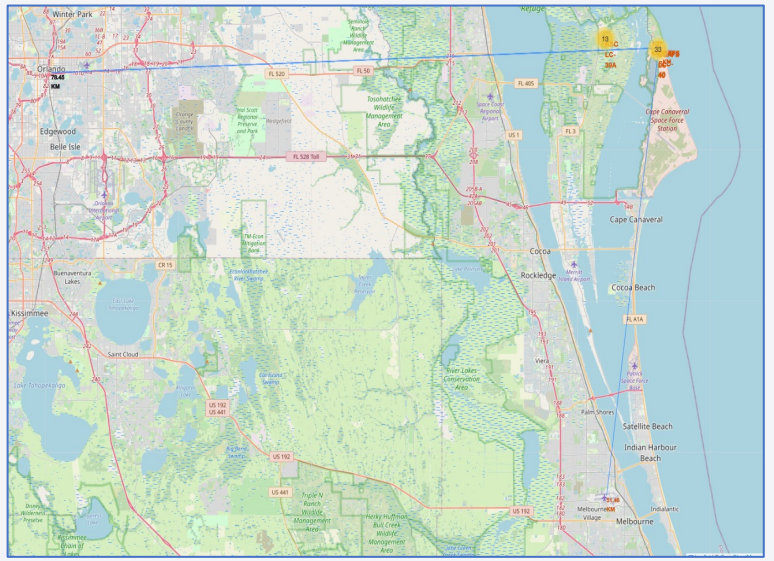
VAFB SLC-4E

- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.

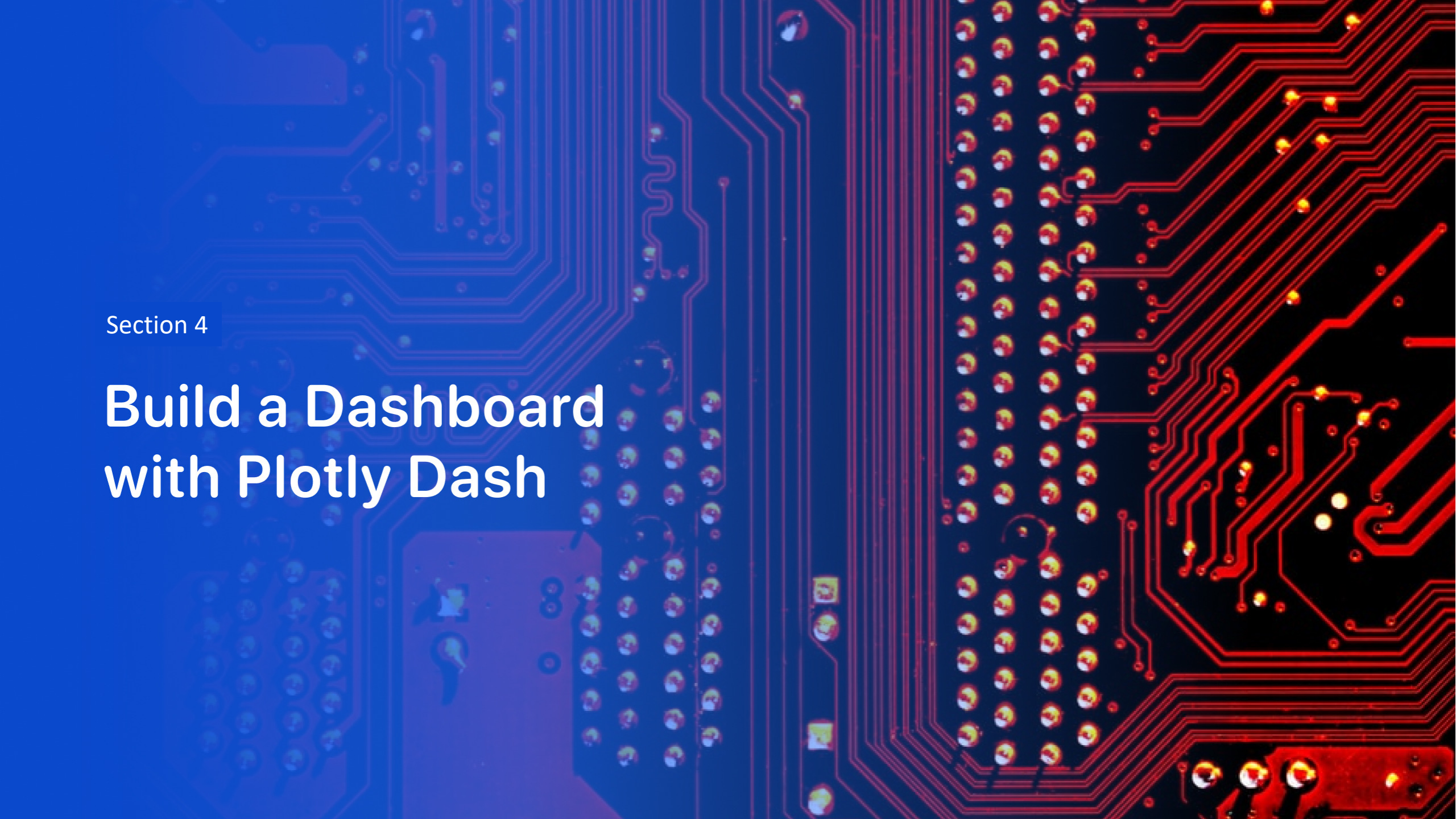
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.

Distances between a Launch Site to its Proximities

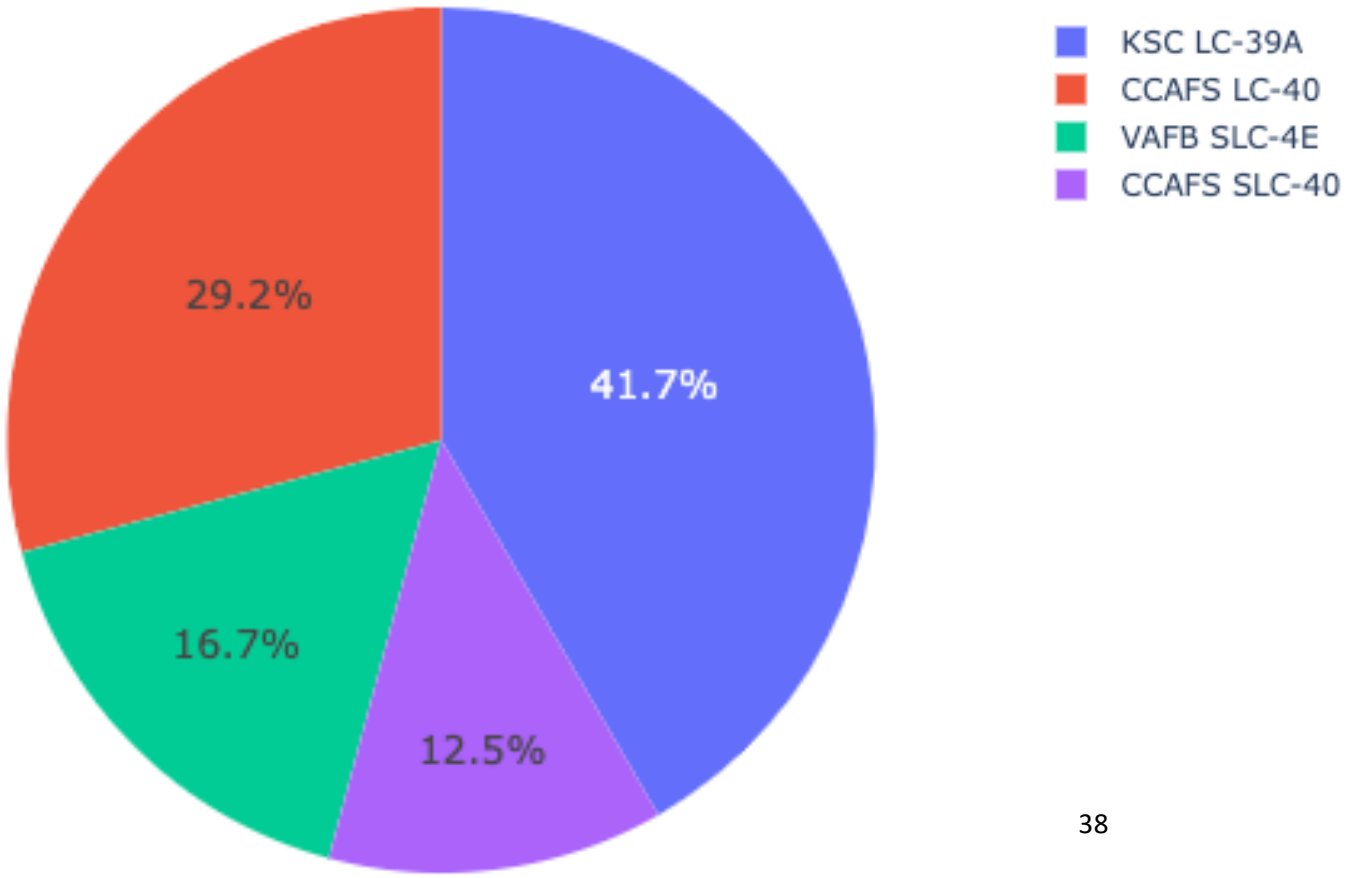




- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being, but are not exactly, right on top of each other.
- The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- The coastline is 0.92 km away from CCAFS LC-40.
- The rail line is 1.33 km away from CCAFS LC-40.


Section 4
Build a Dashboard with Plotly Dash

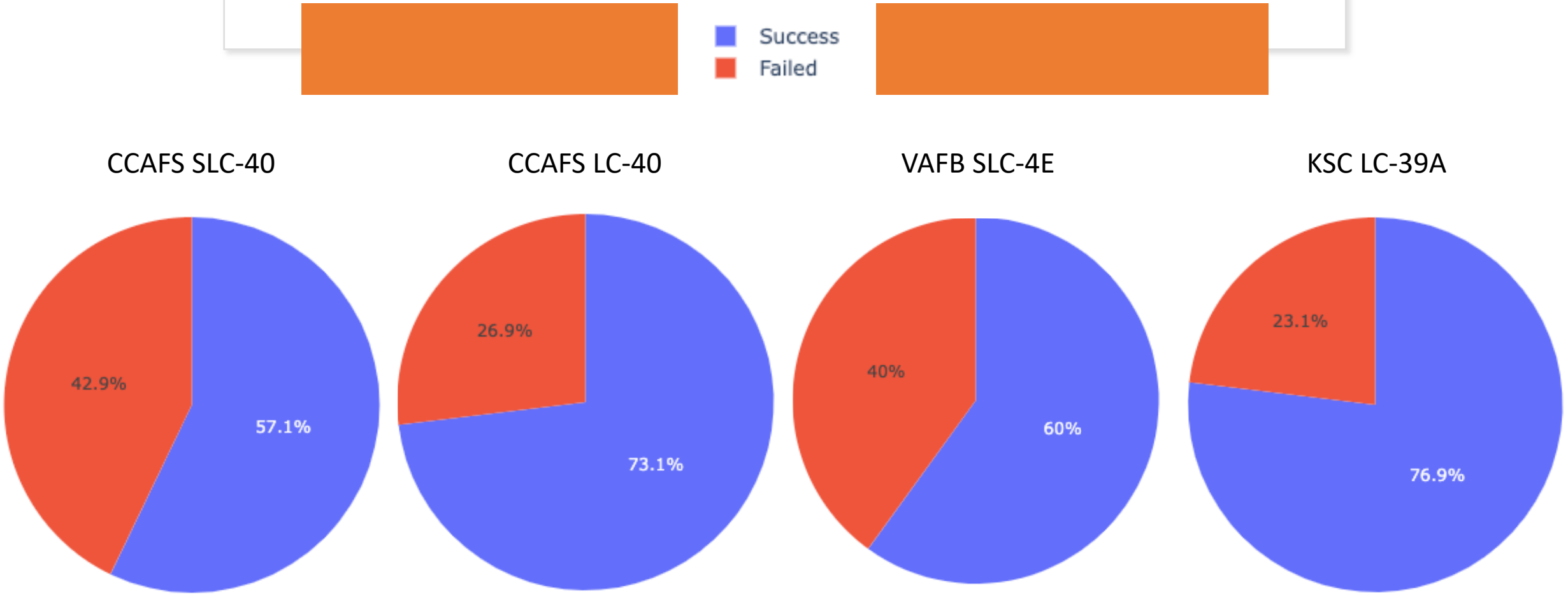
Launch Success Count for All Sites



- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.

- With all launch sites selected, the pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.

Launch Success Count for All Sites



- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).

Payload vs. Launch Outcome

- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.

- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.

- The ‘FT’ booster version category has the largest success rate.

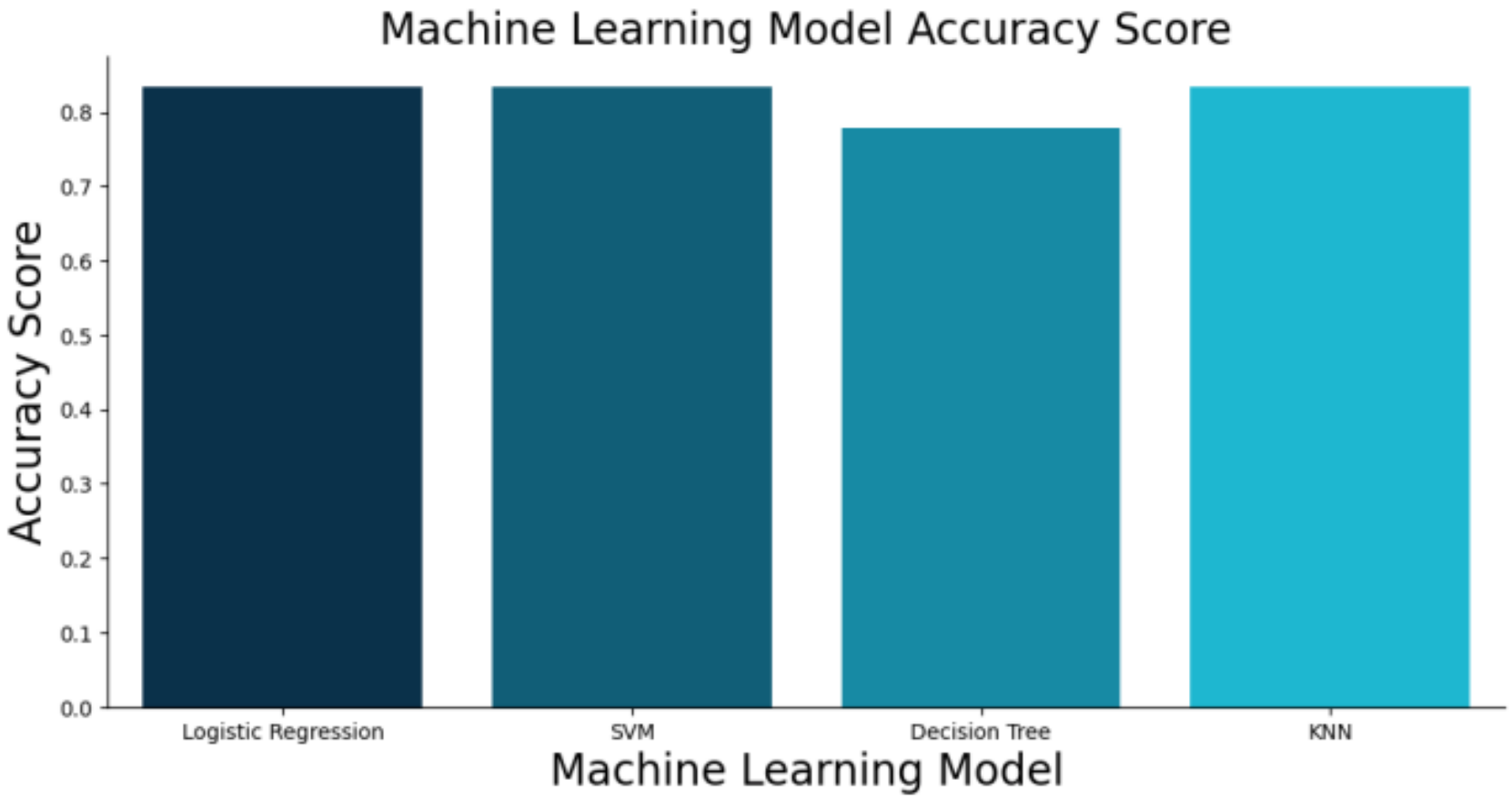
Section 5

Predictive Analysis (Classification)

Classification Accuracy

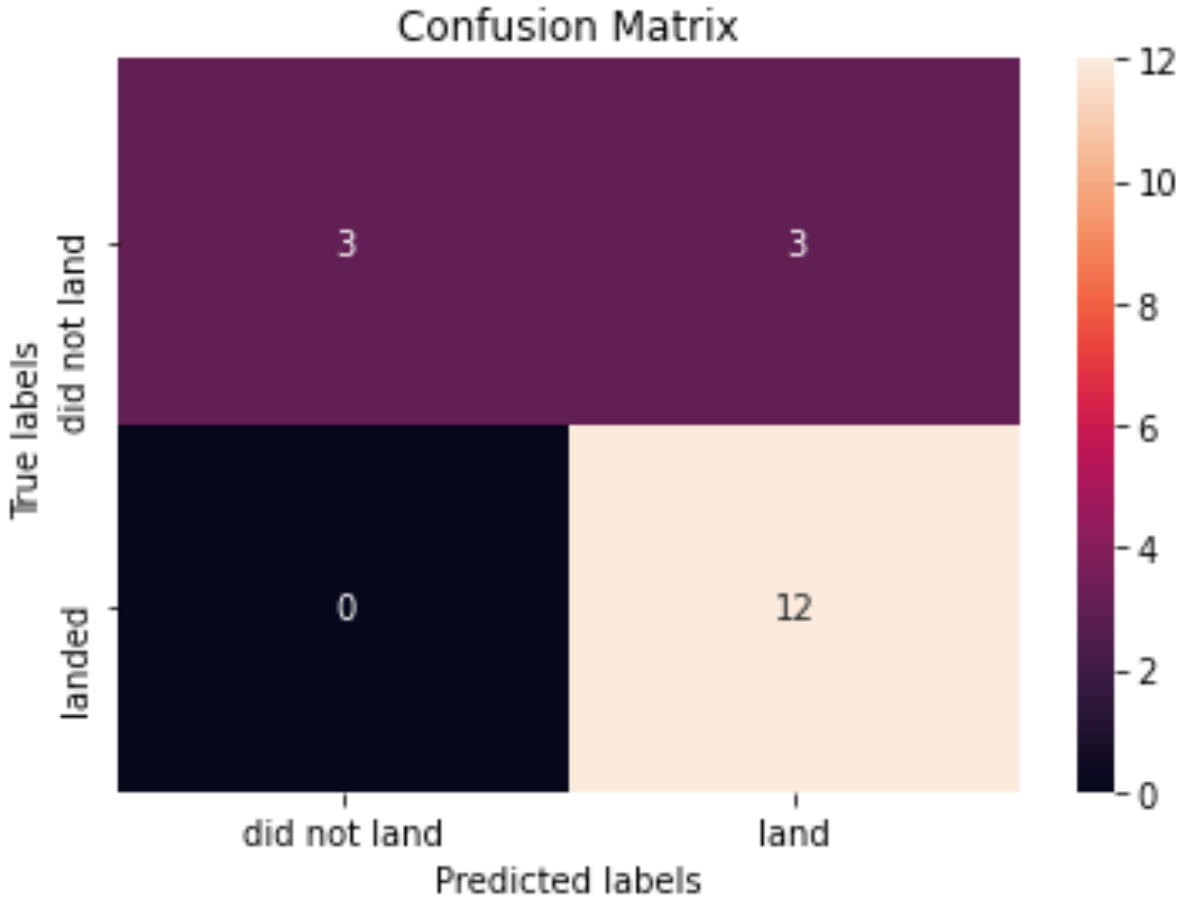
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.6111111111111112
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```



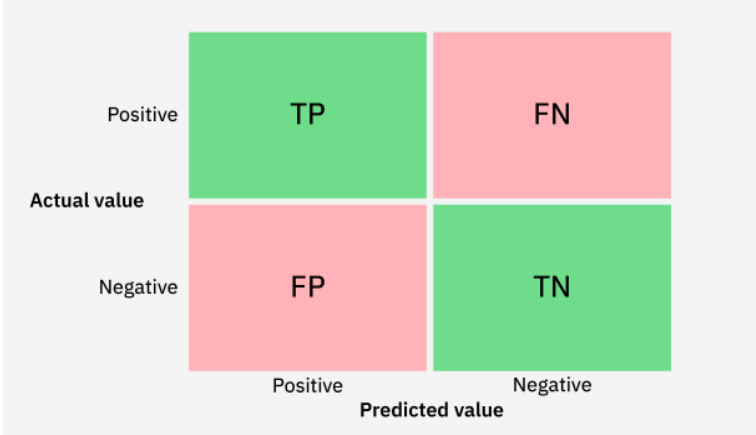
- All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.

Confusion Matrix



- Shown here is the confusion matrix for the Logistic Regression model.

- Confusion matrices can be read as:



- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives

Conclusions

- SpaceX does not have a perfect track record of Falcon 9 first stage landing outcomes
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made.
- The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

Appendix

- Initial Data Sets
  - SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - SpaceX (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
  - Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  - Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv
- Data Sets (.csv files)
  - GitHub URL (CSV 1): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_1.csv
  - GitHub URL (Web Scraped): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_web_scraped.csv
  - GitHub URL (CSV 2): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_2.csv
  - GitHub URL (SpaceX): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/Spacex.csv
  - GitHub URL (CSV 3): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_3.csv
  - GitHub URL (Launch Geo): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_launch_geo.csv
  - GitHub URL (Launch Dash): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_launch_dash.csv

Thank you!