IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tony Bantupalli
05-May-2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data was collected from the SpaceX public API and publically available data on Wikipedia. Data wrangling included extracting launch outcome information to serve as the dependent variable in the Machine Learning models.
  - SQL queries and data visualizations (static plots, interactive maps, and an interactive dashboard) were created to discover insights about the data set and answer questions.
  - Predictive analysis was pursued using Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) Machine Learning models.

- **Summary of all results**
  - Launch data include info about flight number, date of launch, payload mass, orbit type, launch site, mission outcome and other variables.
  - Logistic Regression, SVM (Support Vector Machine), and KNN (k-Nearest Neighbors) all perform equally well for Machine Learning models on this dataset.

Introduction

- In competition with SpaceX, a rival rocket launch company wants to make predictions about the success/failure of SpaceX Falcon 9 rocket first stage landings.
- What is the nature and extent of the data that we have on SpaceX Falcon 9 first stage landings?
- Which machine learning model would work best (have the highest accuracy) to predict the outcome of a Falcon 9 first stage landing from a future launch?
- Will a future Falcon 9 first stage landing be successful?

Section 1

Methodology

Methodology

Executive Summary

- SpaceX API and Wikipedia launch table data was collected.
- Data was cleaned in preparation for visualizations, queries and machine learning model creation.
- Exploratory data analysis (EDA) was done using visualization and SQL.
- Interactive visual analytics were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done.

Data Collection

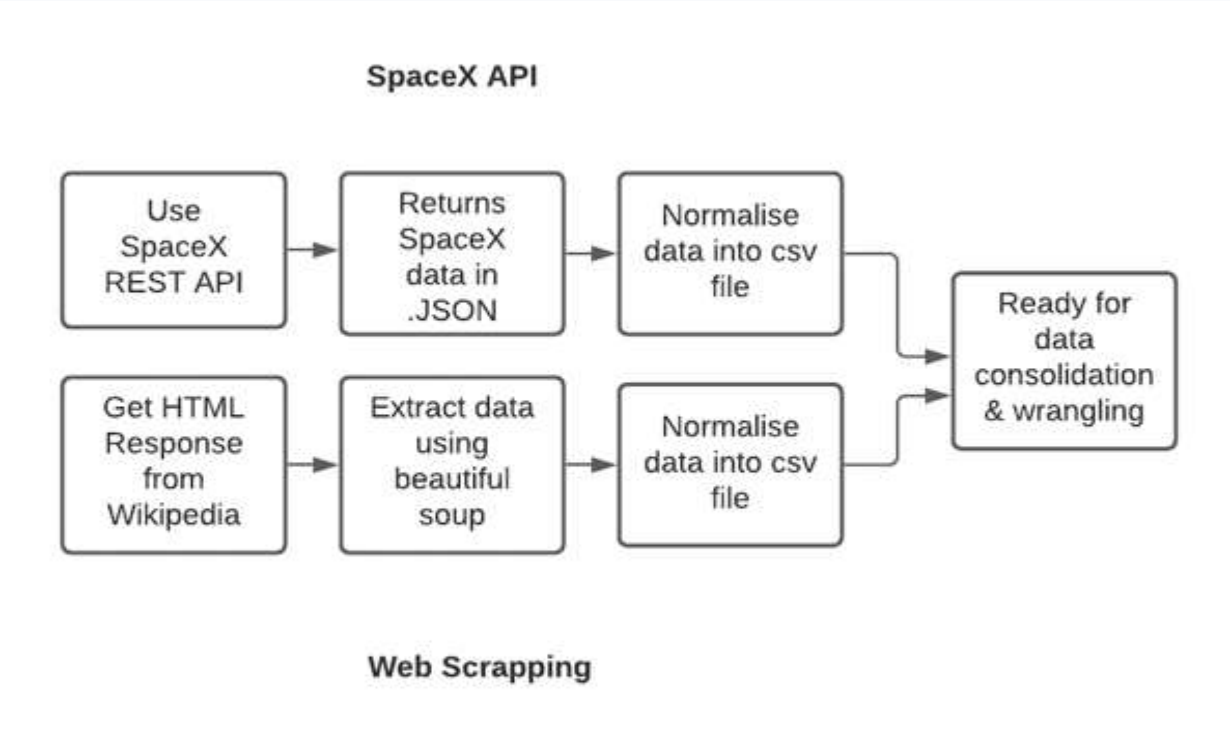
The following datasets was collected:

• SpaceX launch data that is gathered from the SpaceX REST API.

• This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

• The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

• Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup



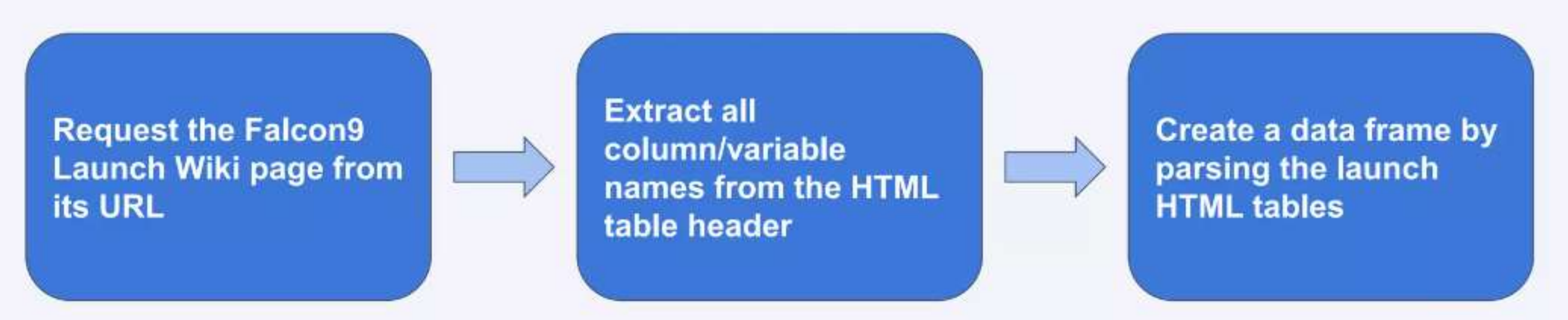
Data Collection – SpaceX API

Data collection with SpaceX REST calls

[Data Collection- Spacex API GITHUB LINK]

Data Collection - Scraping

- Web Scrapping from Wikipedia

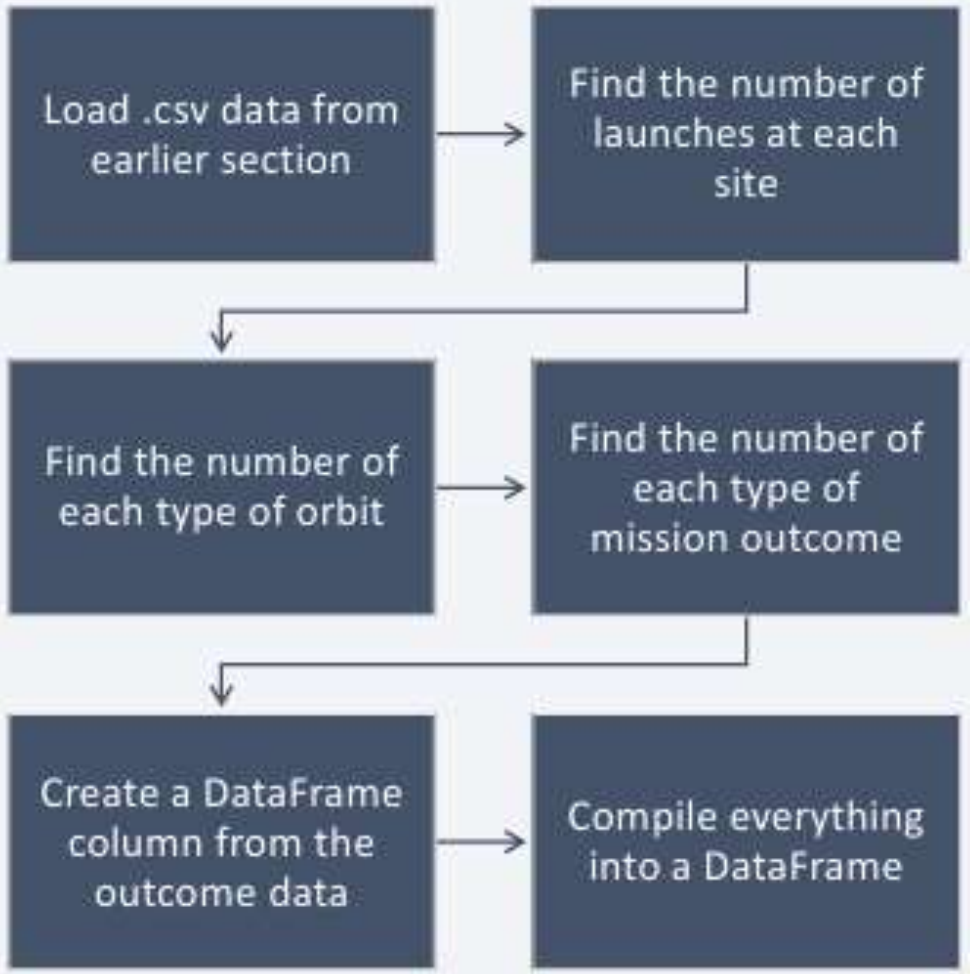


[Data Collection – Scraping GITHUB LINK]

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

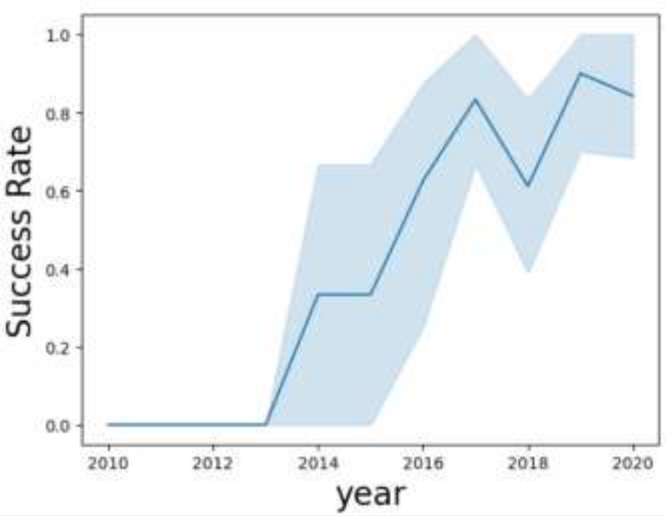
[Data Wrangling   GITHUB LINK]

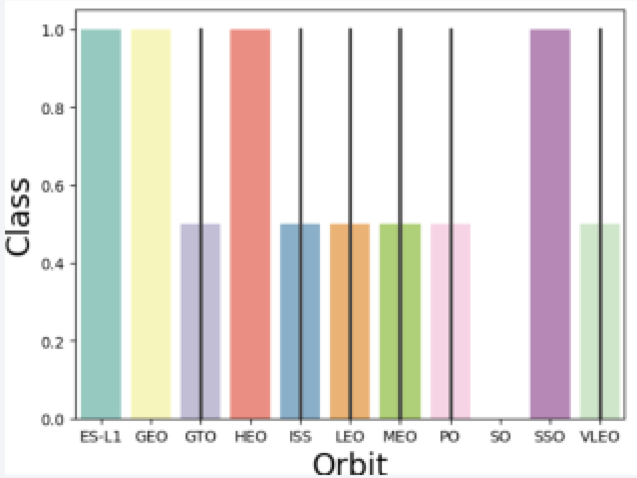


EDA with Data Visualization

- The following charts were created to look at Launch Site trends
    - **Scatterplot** to see mission outcome relationship split by Launch Site and Flight Number.
    - **Scatterplot** to see mission outcome relationship split by Launch Site and Payload.
- The following charts were created to look at Orbit Type trends
    - **Bar chart** to see mission outcome relationship with Orbit Type.
    - **Scatterplot** to see mission outcome relationship split by Orbit Type and Flight Number.
    - **Scatterplot** to see mission outcome relationship split by Orbit Type and Payload.
- The following chart was created to look at trends based on time
    - **Line plot** to see mission outcome trend by year.

[EDA with Data Visualization GITHUB LINK]





EDA with SQL

SQL queries performed include:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster
- versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010 06 04 and 2017 03 20 in descendingorder

EDA with SQL GITHUB LINK

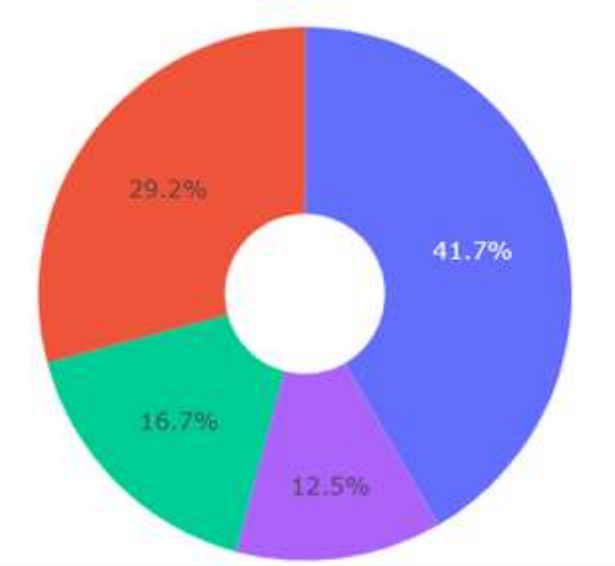
Build an Interactive Map with Folium

- 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

[Interactive Map with Folium   GITHUB LINK]

Build a Dashboard with Plotly Dash

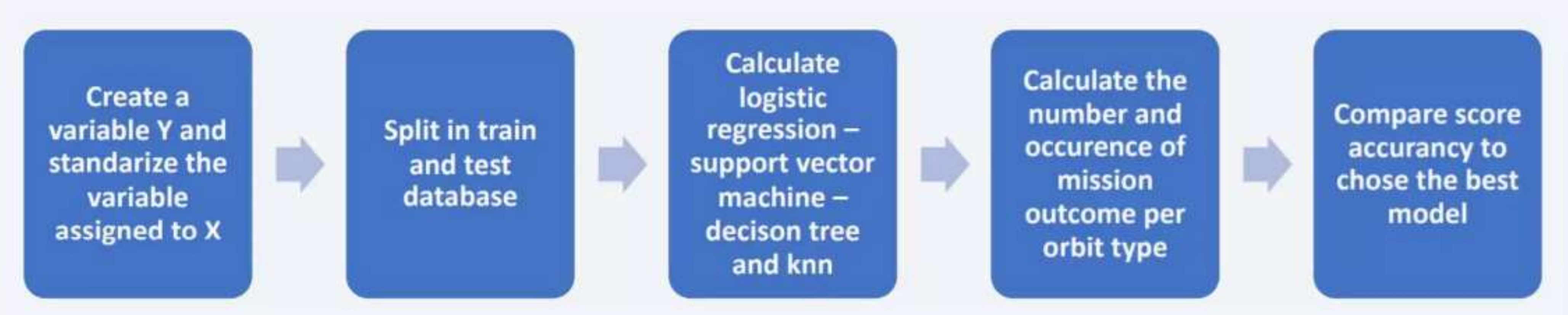
- The input dropdown is used to select one or all launch sites for the pie chart and scatterplot.
- The pie chart displays one of two things:
  - For All Sites –the distribution of successful Falcon 9 first stage landings between the sites
  - For One Site –the distribution of successful and failed Falcon 9 first stage landings for that site
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displays the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.



[Dashboard with Plotly Dash GITHUB LINK]()

Predictive Analysis (Classification)

- The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958.



[Predictive Analysis   GITHUB LINK]

Results

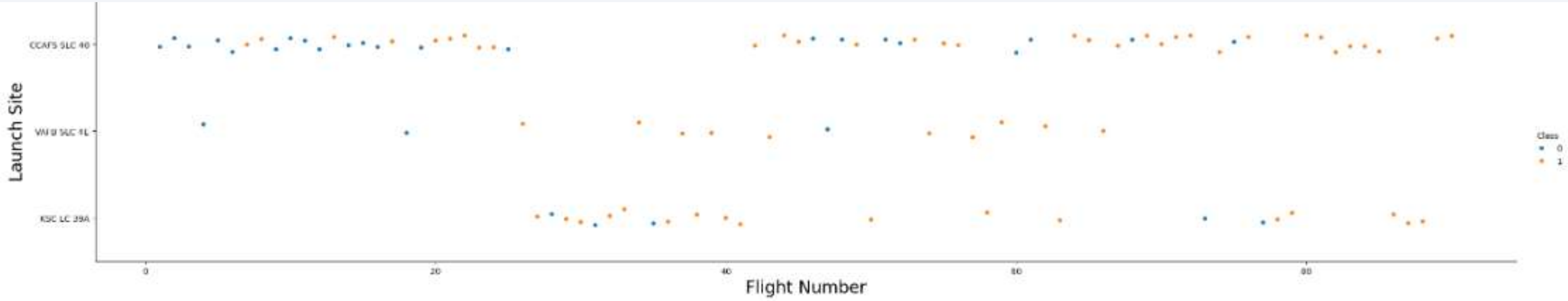
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites. • Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Success rate varies noticeably with launch site.
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases



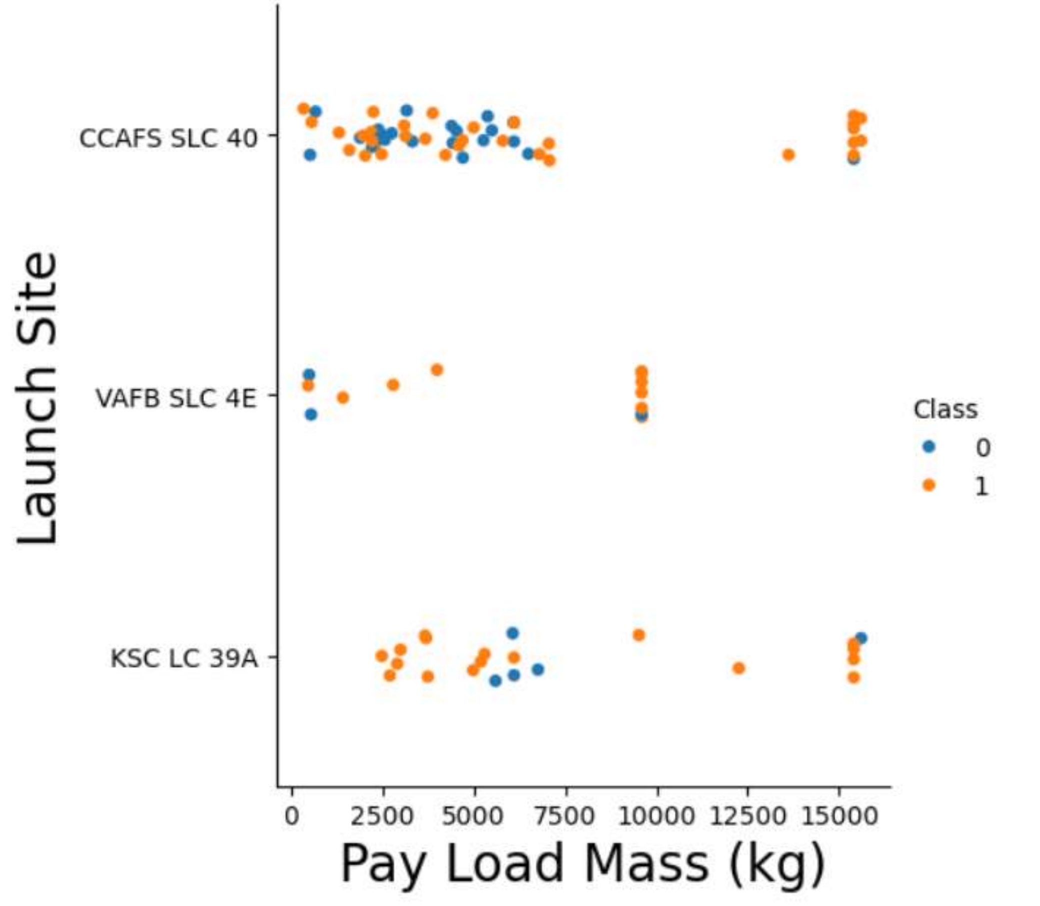
- Falcon 9 first stage failed landingsare indicated by the ‘0’ Class (Blue) and successful landings by the ‘1’ Class (Orange)

Payload vs. Launch Site

- The majority of lPay Loads with lower Mass have been launched from CCAFS SLC 40

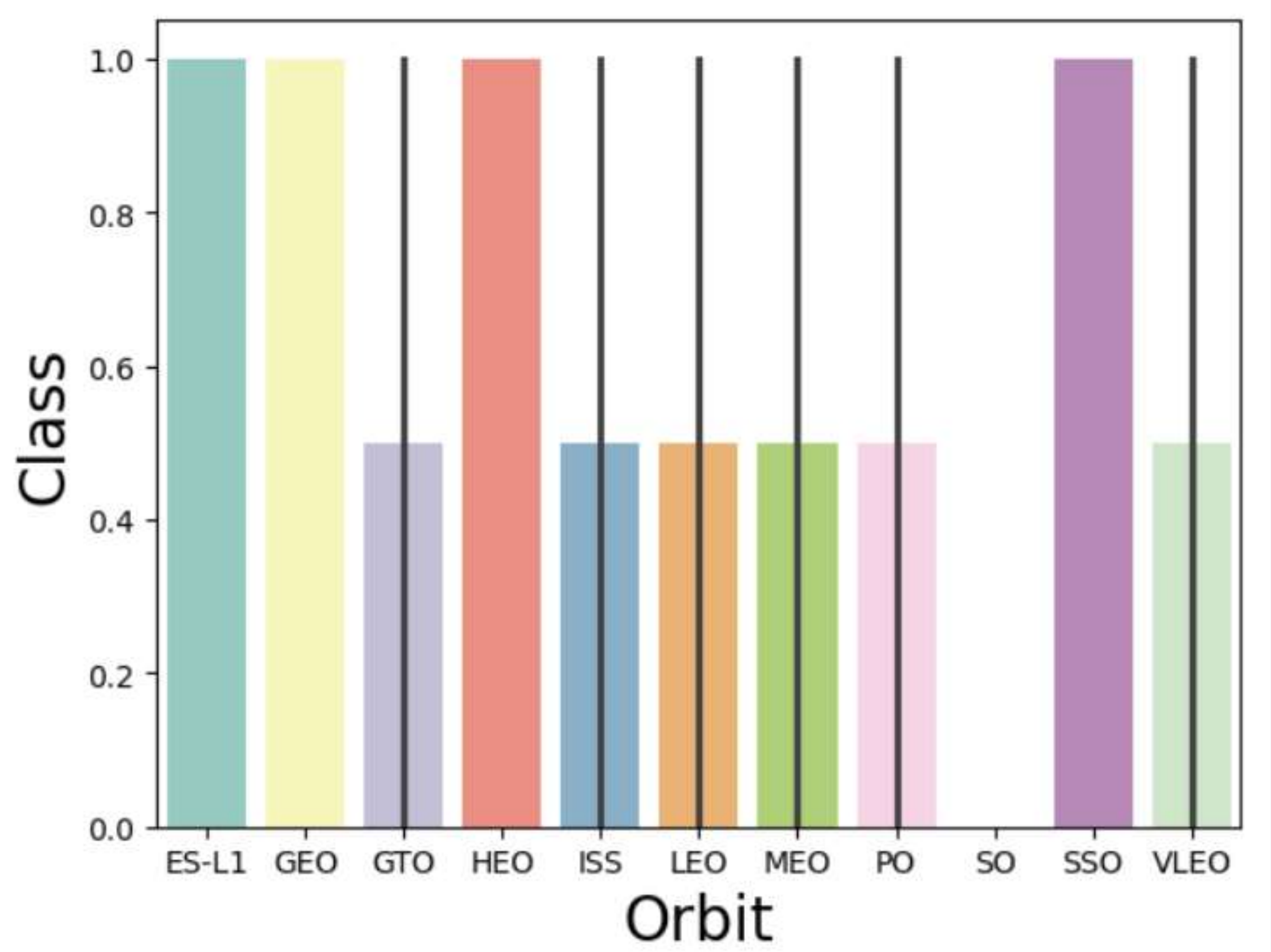
- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.

- The failed landings at the KSC LC 39A launch site are all grouped around a narrow band of payload masses.



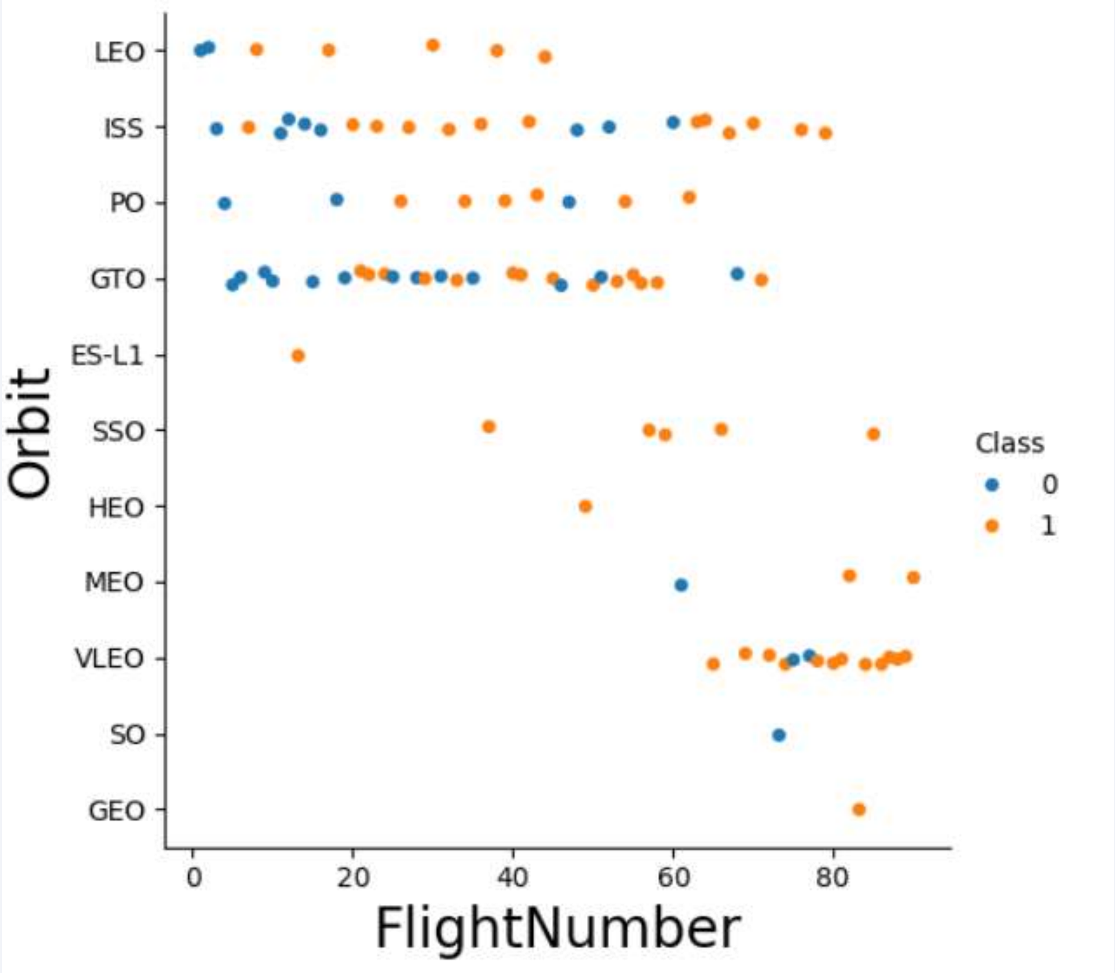
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.



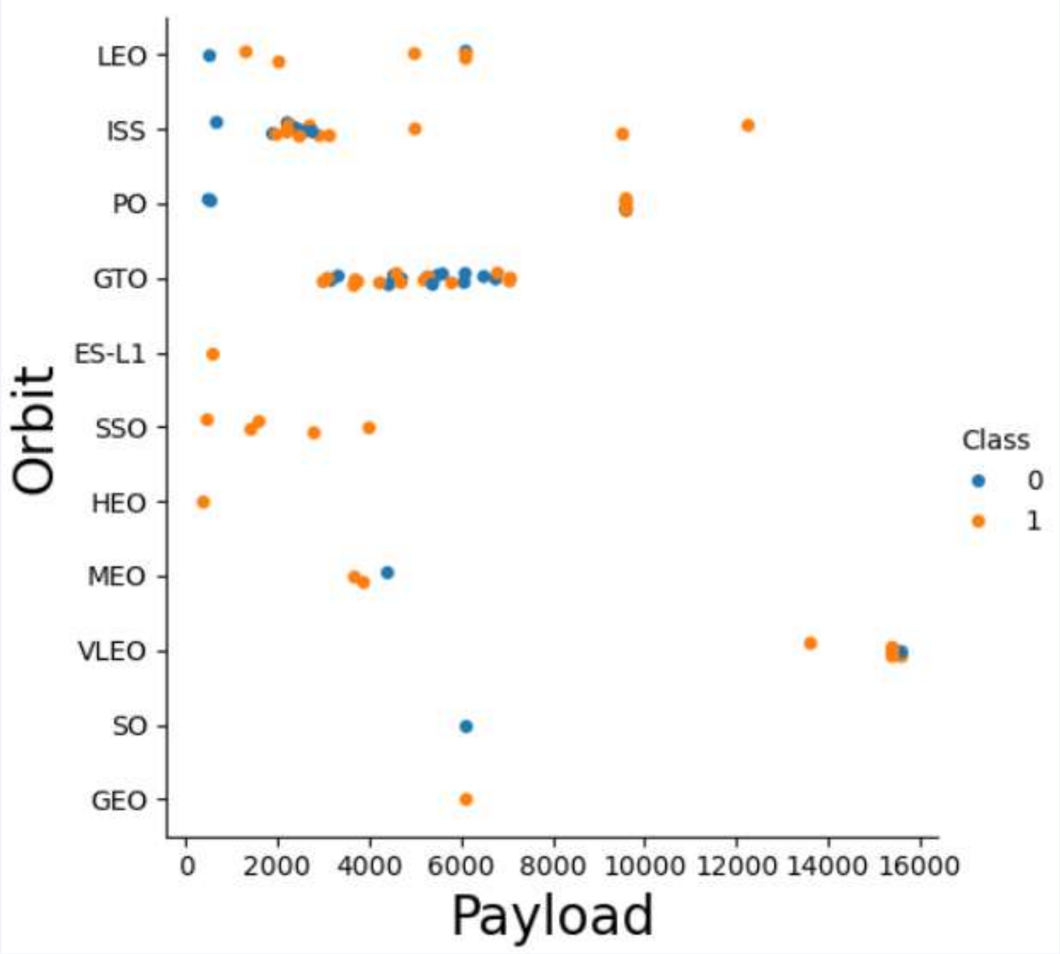
Flight Number vs. Orbit Type

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



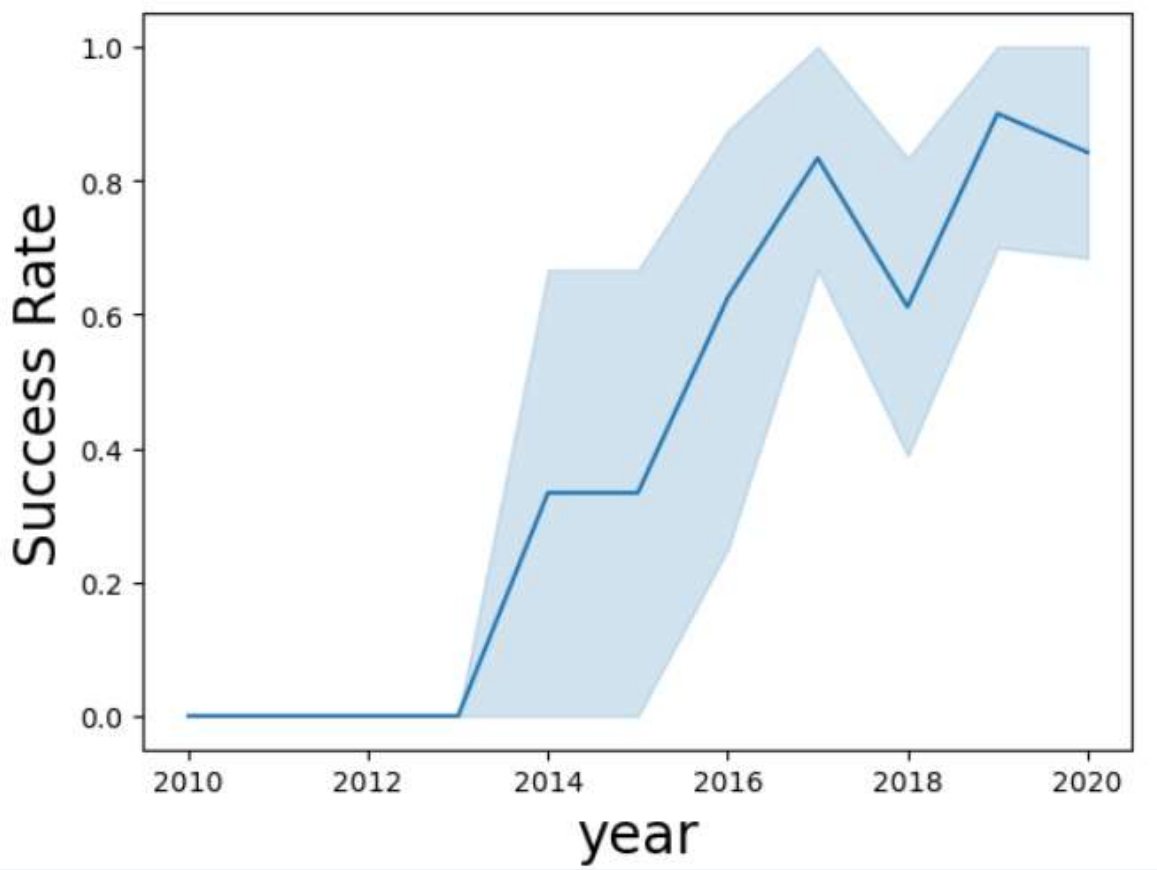
Payload vs. Orbit Type

- Some orbit types have better success rates than others.
- Success rate appears to have no obvious correlation with payload mass.



Launch Success Yearly Trend

- Launch success rate has increased significantly since 2013 and has stablised since 2019, potentially due to advance in technology and lessons learned.



All Launch Site Names

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

- There are four unique launch sites.

Launch Site Names Begin with 'KSC'

%sql SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'KSC%' LIMIT 5;

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

- 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 "Total Payload Mass (kg)" FROM SPACEXTABLE WHERE customer LIKE '%NASA (CRS)%';

Total Payload Mass (kg)
48213

- The total payload carried by boosters from NASA is 48,213 kg.

Average Payload Mass by F9 v1.1

%sql SELECT sum(payload_mass__kg_) / count(payload_mass__kg_) AS "Average Payload Mass (kg)" FROM SPACEXTABLE WHERE booster_version LIKE 'F9 v1.1';

Average Payload Mass (kg)
2928

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

First Successful Ground Landing Date

%sql SELECT min(DATE) AS "First Successful Landing Outcome Date" FROM SPACEXTABLE WHERE landing_outcome LIKE 'Success (ground pad)';

First Successful Landing Outcome Date
2015-12-22

- 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 DISTINCT booster_version FROM SPACEXTABLE WHERE landing_outcome = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 and 6000;
```

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

- 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 (SELECT count(*) FROM SPACEXTABLE WHERE LOWER(landing_outcome) LIKE '%success%') AS "Success", count(*) AS "Failure" FROM SPACEXTABLE WHERE LOWER(landing_outcome) NOT LIKE '%success%';
```

Success	Failure
61	40

- There were 61 successful and 40 failed mission outcomes.

Boosters Carried Maximum Payload

%sql SELECT booster_version, payload_mass__kg_ FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTABLE);

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

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

2017 Launch Records

%sql SELECT strftime('%Y-%m', DATE) AS "Month", landing_outcome, booster_version, launch_site FROM SPACEXTABLE WHERE strftime('%Y', DATE) = '2017';

- There were two failed landing outcomes with a drone ship in 2015. Both launched from CCAFS LC-40.

Month	Landing_Outcome	Booster_Version	Launch_Site
2017-01	Success (drone ship)	F9 FT B1029.1	VAFB SLC-4E
2017-02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
2017-03	No attempt	F9 FT B1030	KSC LC-39A
2017-03	Success (drone ship)	F9 FT B1021.2	KSC LC-39A
2017-05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
2017-05	No attempt	F9 FT B1034	KSC LC-39A
2017-06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
2017-06	Success (drone ship)	F9 FT B1029.2	KSC LC-39A
2017-06	Success (drone ship)	F9 FT B1036.1	VAFB SLC-4E
2017-07	No attempt	F9 FT B1037	KSC LC-39A
2017-08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
2017-08	Success (drone ship)	F9 FT B1038.1	VAFB SLC-4E
2017-09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
2017-10	Success (drone ship)	F9 B4 B1041.1	VAFB SLC-4E
2017-10	Success (drone ship)	F9 FT B1031.2	KSC LC-39A
2017-10	Success (drone ship)	F9 B4 B1042.1	KSC LC-39A
2017-12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40
2017-12	Controlled (ocean)	F9 FT B1036.2	VAFB SLC-4E

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

```
%sql SELECT landing_outcome, count(landing_outcome) AS "Count" FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY count(landing_outcome) DESC;
```

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

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

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

•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.

Distance from Launch Site to 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.

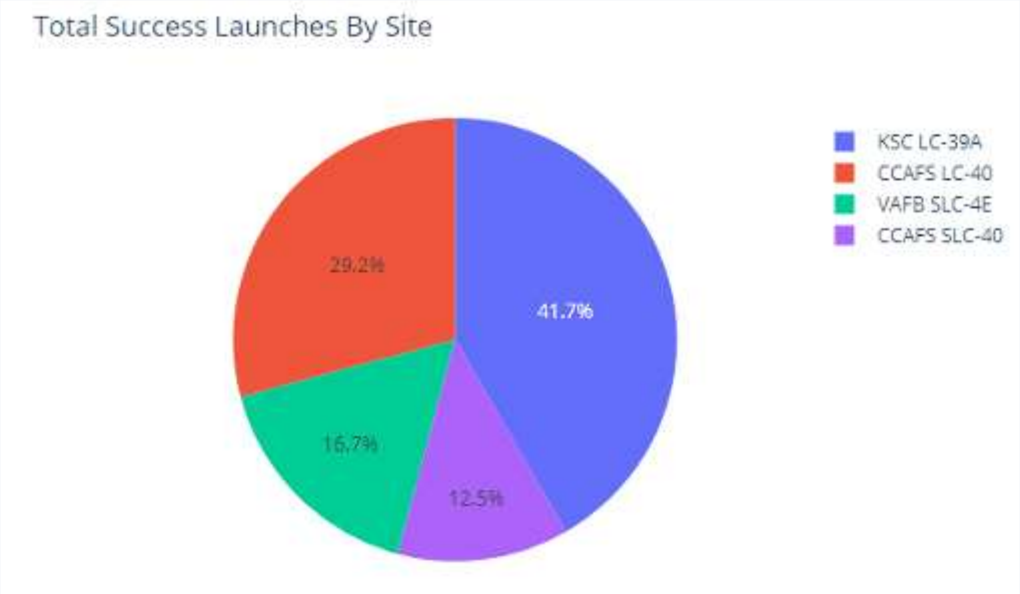
•The rail line is 1.33 km away from CCAFS LC-40.

Section 4

Build a Dashboard with Plotly Dash

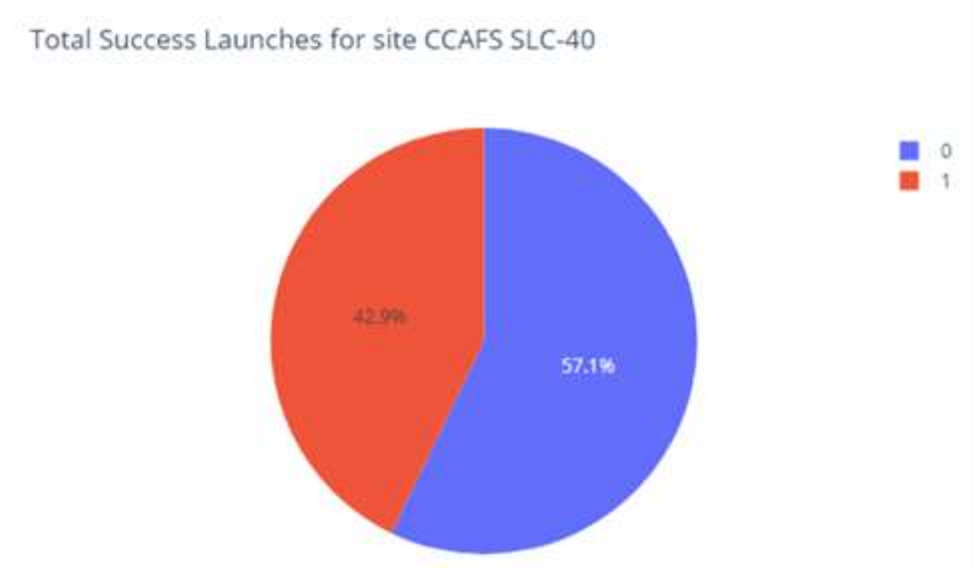
Total success launches by all sites

- The pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.



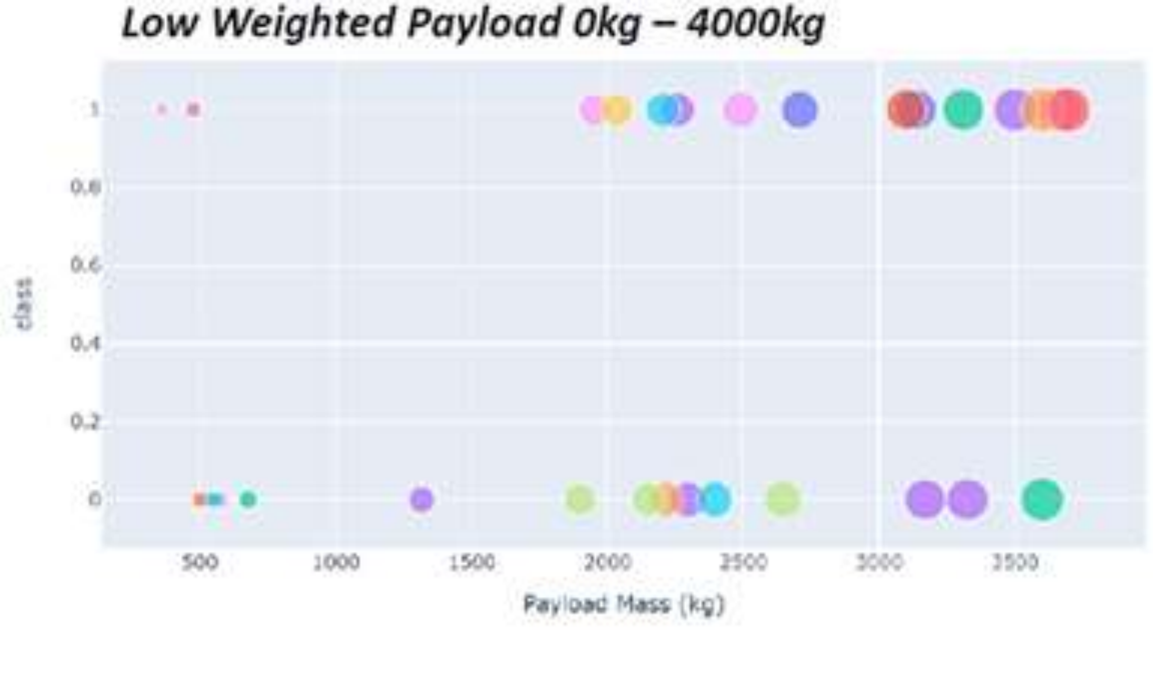
Success rate by site

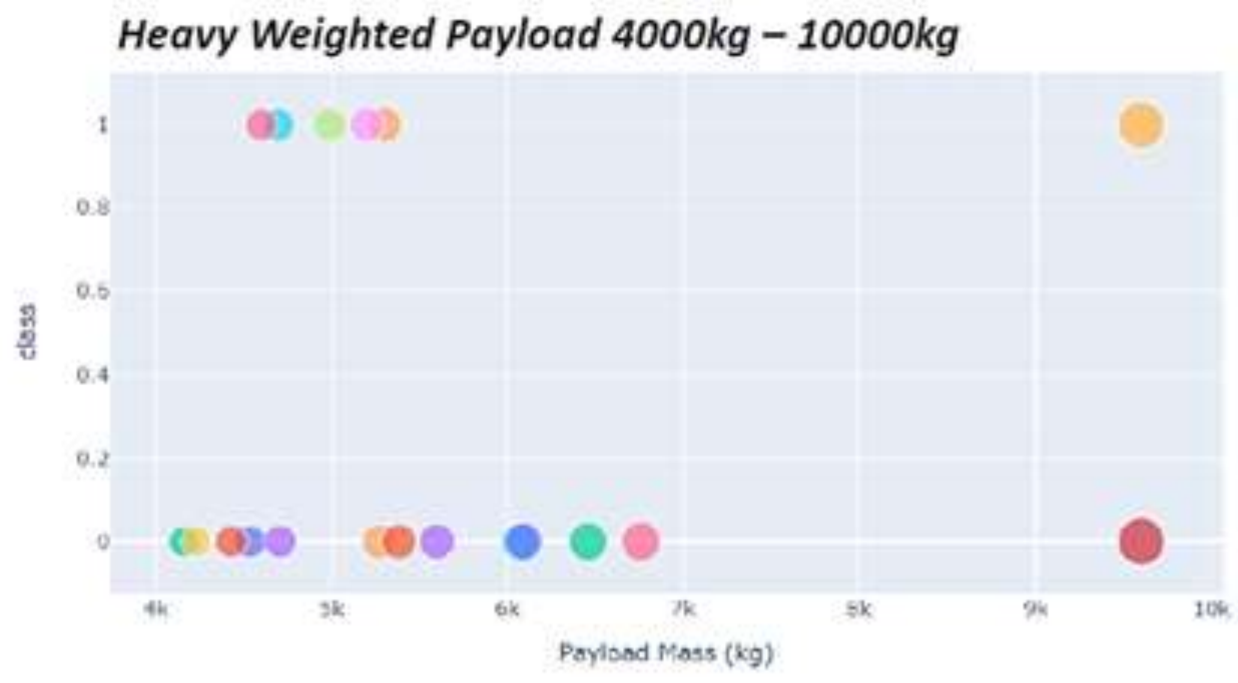
- Falcon 9 first stage failed landings are indicated by the ‘0’ Class (blue wedge in the pie chart) and successful landings by the ‘1’ Class (red wedge in the pie chart).
- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%)



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.



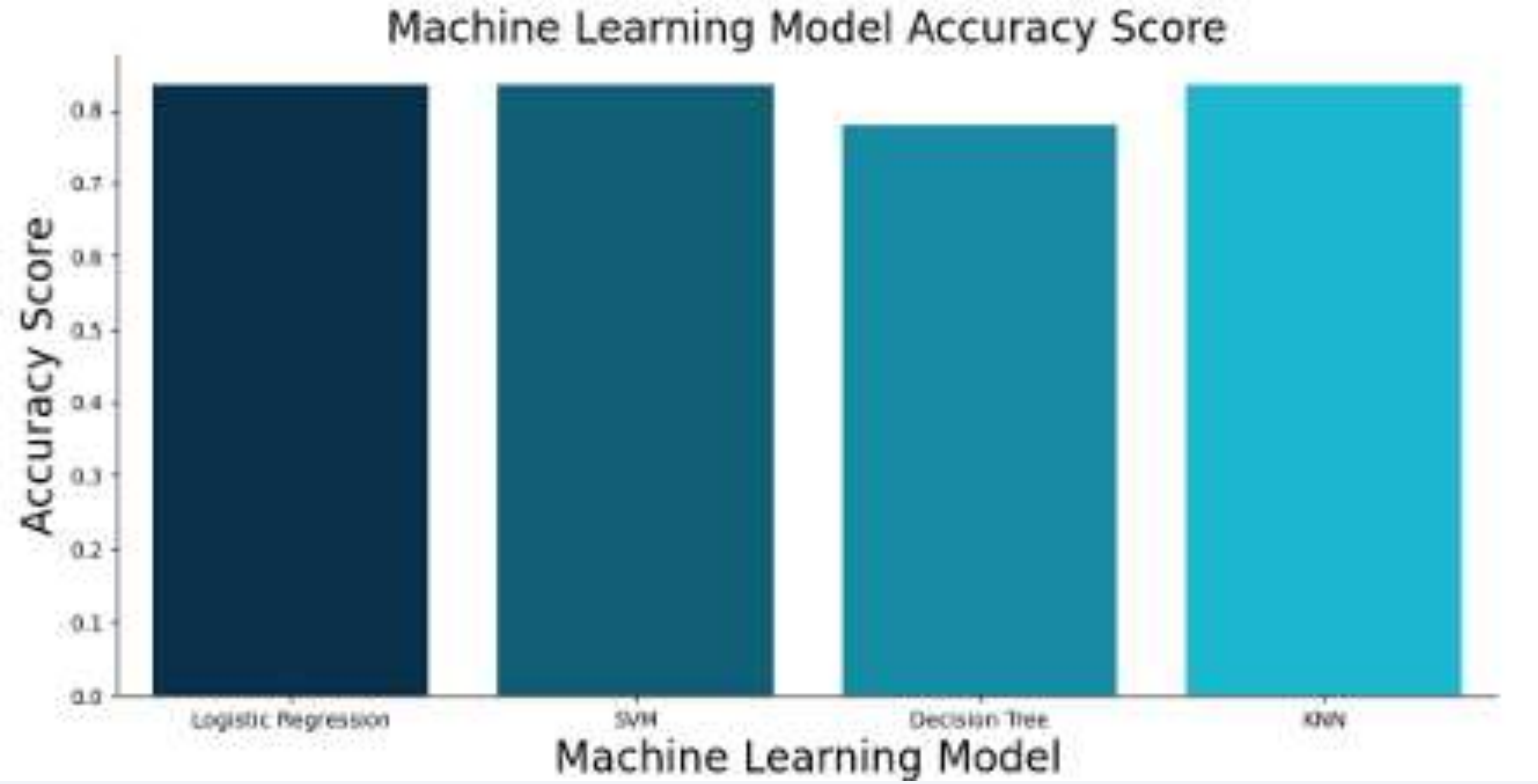


Section 5

Predictive Analysis (Classification)

Classification Accuracy

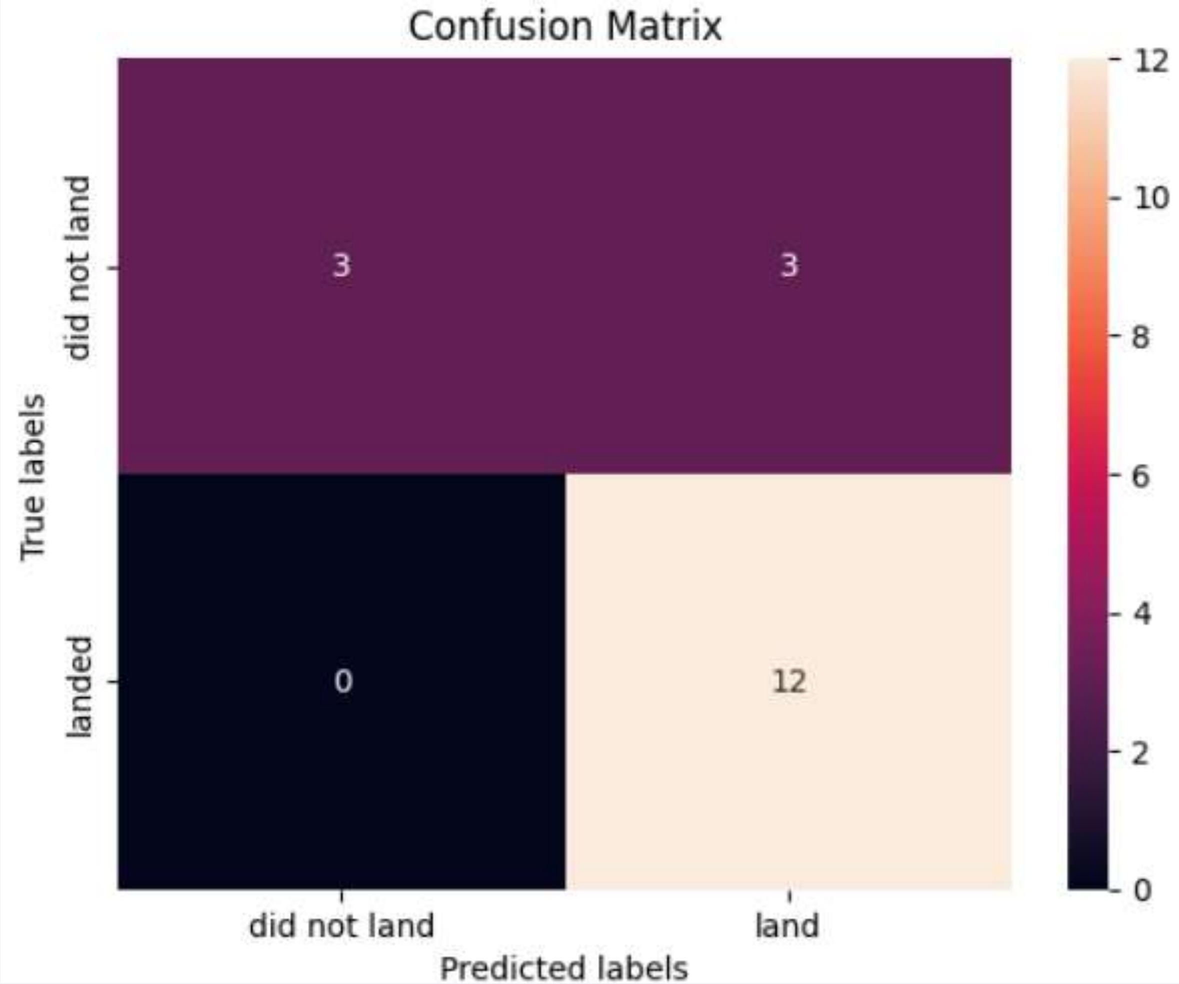
- All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.



Confusion Matrix

- The 4 models has the same confusion matrix as they has the same accuracy test percentage the main problem of this models is false positivity

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



Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Appendix

DATASETS

- GitHub URL (CSV 1)
- GitHub URL (CSV 2)
- GitHub URL (CSV 3)
- GitHub URL (scrapped)

JUPYTER NOTEBOOKS AND DASHBOARD PYTHON FILE

- GitHub URL (Data Collection)
- GitHub URL (Web Scraping)
- GitHub URL (Data Wrangling)
- GitHub URL (EDA with SQL)
- GitHub URL (EDA with Data Visualization)
- GitHub URL (Folium Maps)
- GitHub URL (Dashboad File)
- GitHub URL (Machine Learning)

Thank you!