IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aya Zmizem (https://github.com/snowieeeee)
September 23, 2023

Outline

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

Executive Summary

- We collected data from two primary sources: the public SpaceX API and by scraping SpaceX's Wikipedia page. The data acquisition process allowed us to compile a comprehensive dataset for analysis.
- A critical step involved the creation of a 'Class' column, which categorizes landings as either successful or not. This labeling was pivotal for our predictive analysis.
- We performed EDA using a combination of SQL queries, data visualization, interactive Folium maps, and informative dashboards. These techniques allowed us to gain insights into the data, identify trends, and visualize key metrics.
- Feature Engineering: To prepare our data for machine learning, we selected relevant columns to serve as features for our models. Categorical variables were transformed into binary form using One Hot Encoding.
- Data standardization was applied to ensure uniformity in model inputs. - We utilized GridSearchCV to fine-tune our machine learning models, searching for optimal parameters.
- Four machine learning models were employed: Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K Nearest Neighbors. - Each model produced similar results, achieving an accuracy rate of approximately 83.33%. However, it's important to note that all models exhibited a tendency to overpredict successful landings.

Introduction

In this capstone project, we delve into the fascinating world of space exploration and economics. At the heart of our endeavor lies the prediction of the successful landing of SpaceX's Falcon 9 first stage. SpaceX, a pioneering company in the commercial space industry, advertises Falcon 9 rocket launches on its website, boasting a cost of 62 million dollars per launch. In stark contrast, other providers demand upward of 165 million dollars for each launch. The key to this dramatic cost difference lies in SpaceX's groundbreaking ability to reuse the first stage of its rockets. Understanding the potential success of this critical stage of the launch process holds immense significance, not only in terms of cost determination but also for any aspiring company looking to compete with SpaceX. If Space Y, our client, aims to challenge SpaceX's dominance in the industry, the ability to predict the outcome of the first stage's landing is paramount.

Space Y has entrusted us with a mission: to harness the power of data science and machine learning to predict the success of Stage 1 recovery in SpaceX's Falcon 9 rocket launches. The implications of our success are profound, as accurate predictions will not only inform Space Y's strategic decisions but also contribute to the larger landscape of space travel economics.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Gathered data from SpaceX public API and by scrapping SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We tuned the models using GridSearchCV

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.

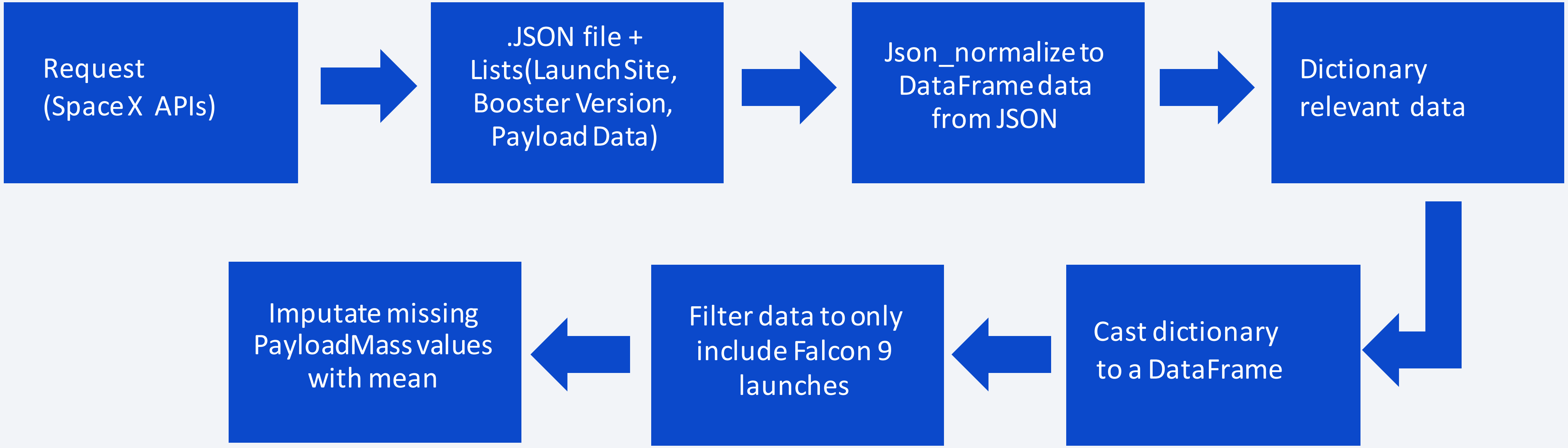
Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscraping Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



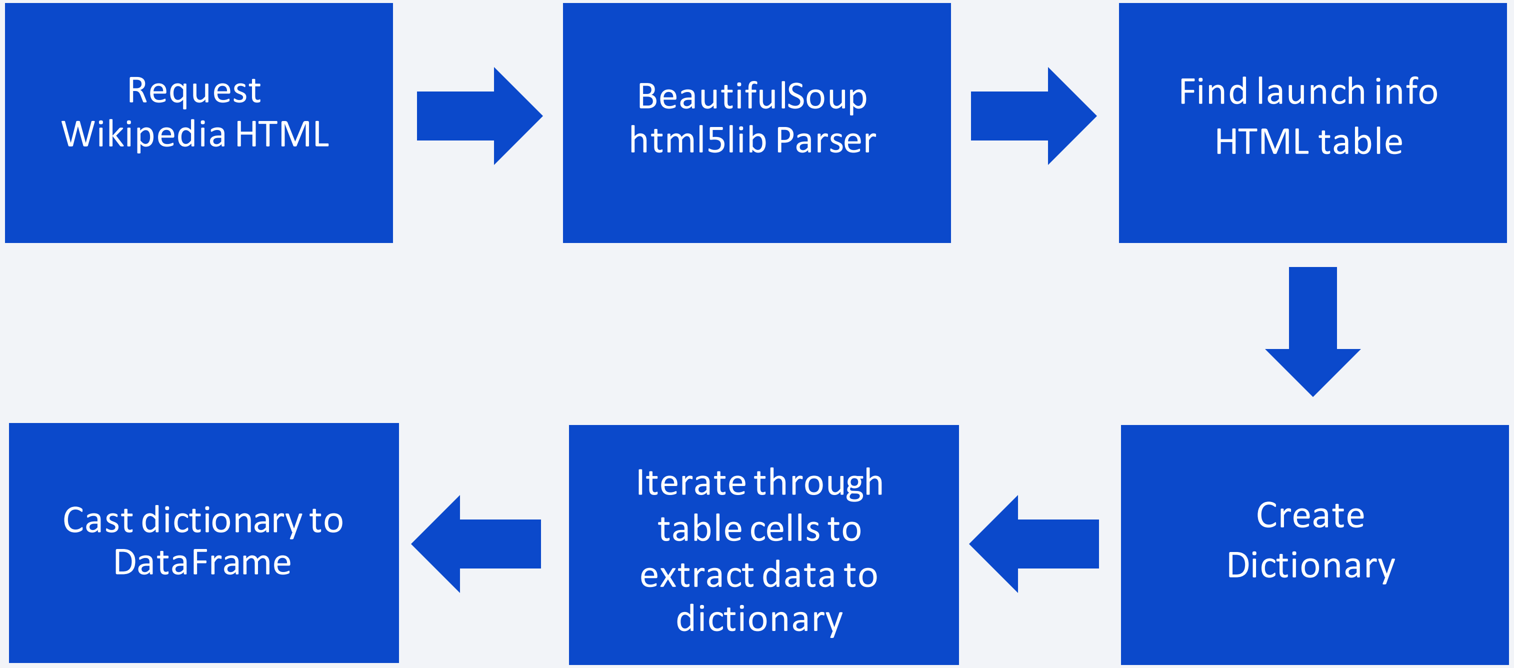
GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/Data_Collection_API.ipynb

Data Collection - Scraping

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/Data_Collection_Web_Scraping.ipynb



Data Wrangling

Create a training label with landing outcomes where successful = 1 & failure = 0.

Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping:

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/Data_Wrangling.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/EDA_with_SQL.ipynb

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to

decide if a relationship exists so that they could be used in training the machine learning model

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/EDA_with_Visualization.ipynb

Build an Interactive Map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

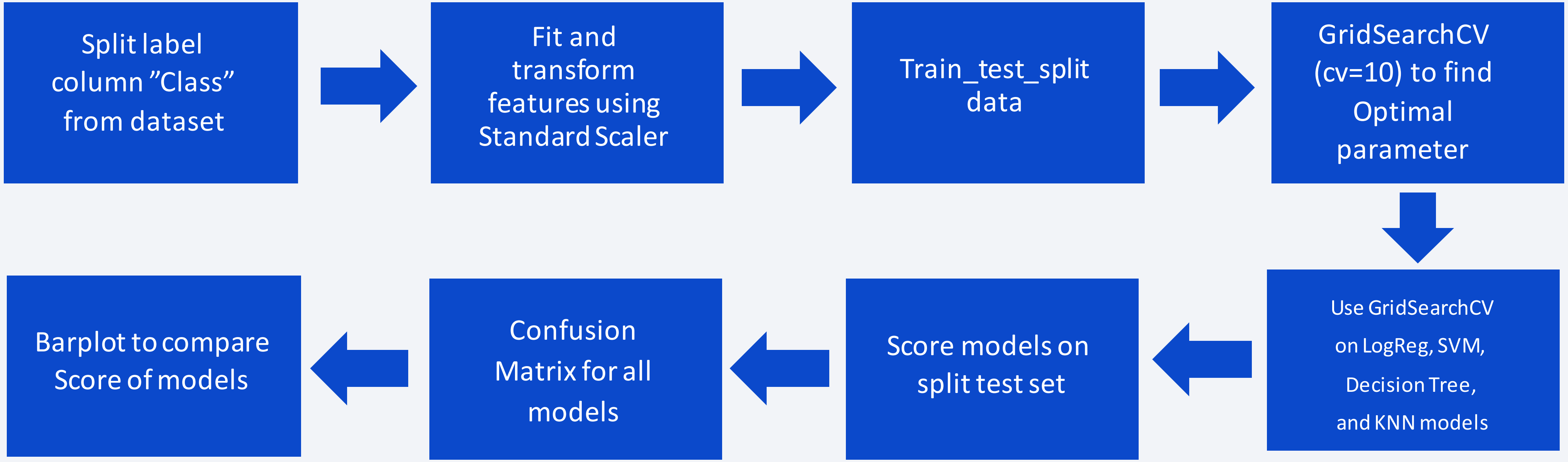
Dash and html components were used as they are the most important thing and almost everything depends on them, such as graphs, tables, dropdown and so on.

To simplify things, I used Pandas to create a dataframe. To plot the graphs I used Plotty. Scatter and pie chart were used as well. Dropdown was used to visualize launch sites.

GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/dashboard_application.py

Predictive Analysis



GitHub URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone/blob/main/Predictive_Analysis_.ipynb

Results

LogReg

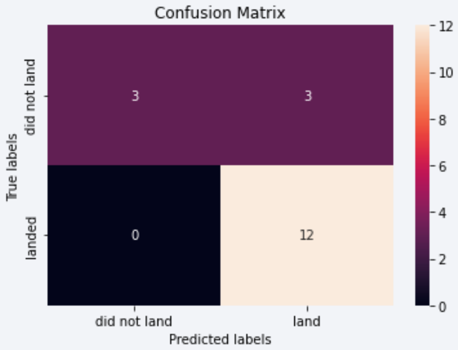
Accuracy: 83.33%

SVM

Accuracy: 83.33%

Decision Tree

Accuracy: 83.33%



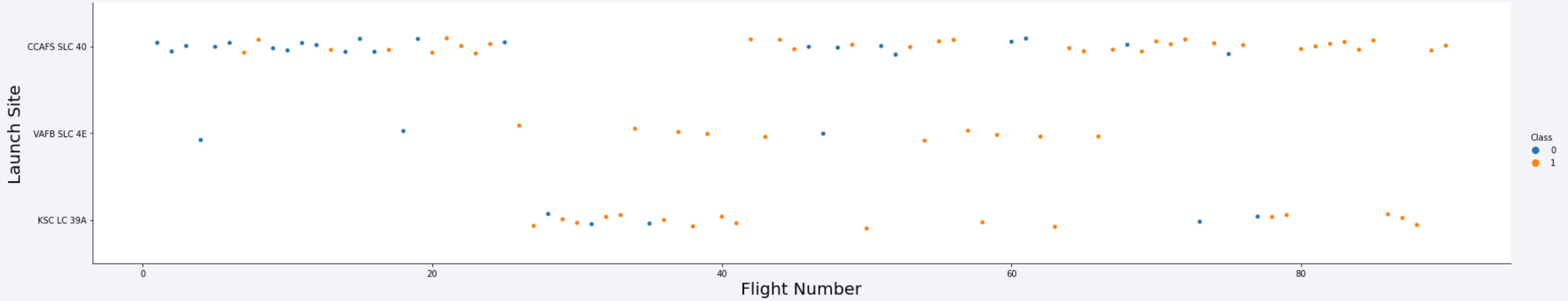
KNN

Accuracy: 83.33%

Section 2

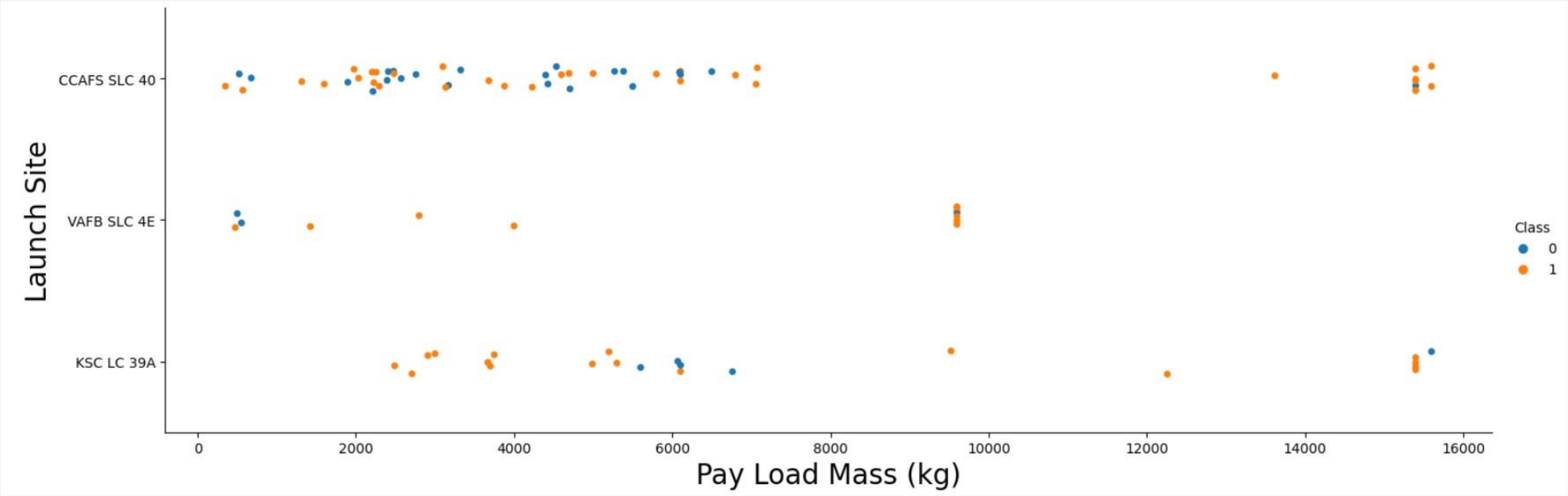
Insights drawn from EDA

Flight Number vs. Launch Site



Blue indicates successful launch and orange indicates unsuccessful launch. Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate.
CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



Blue indicates successful launch and orange indicates unsuccessful launch. Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

Success Rate vs. Orbit Type

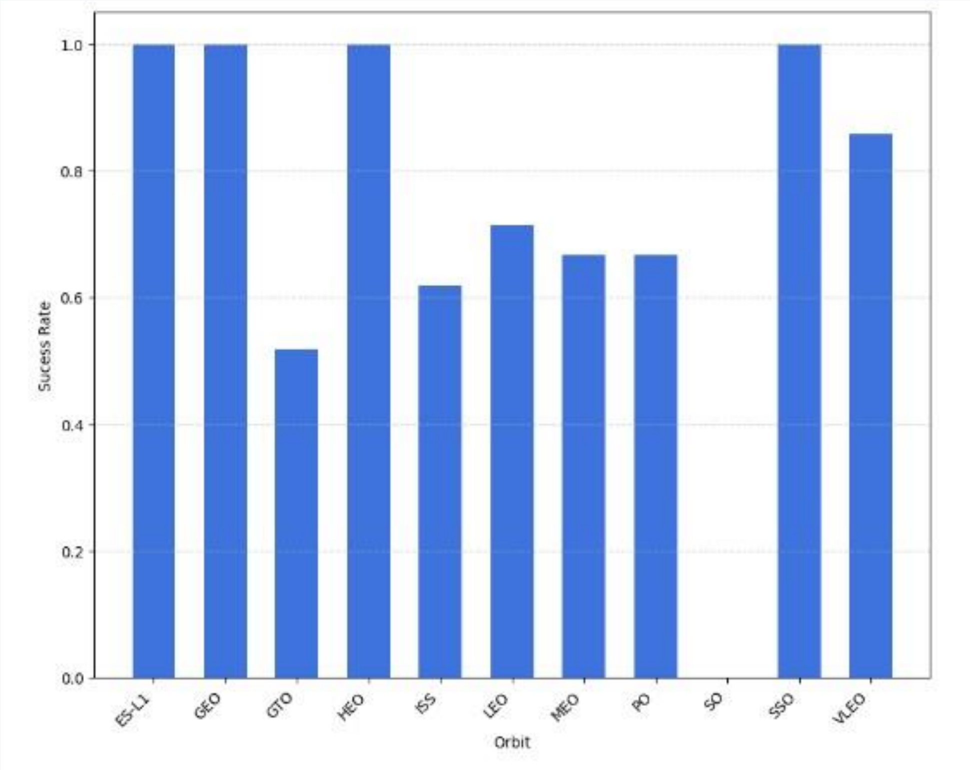
ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

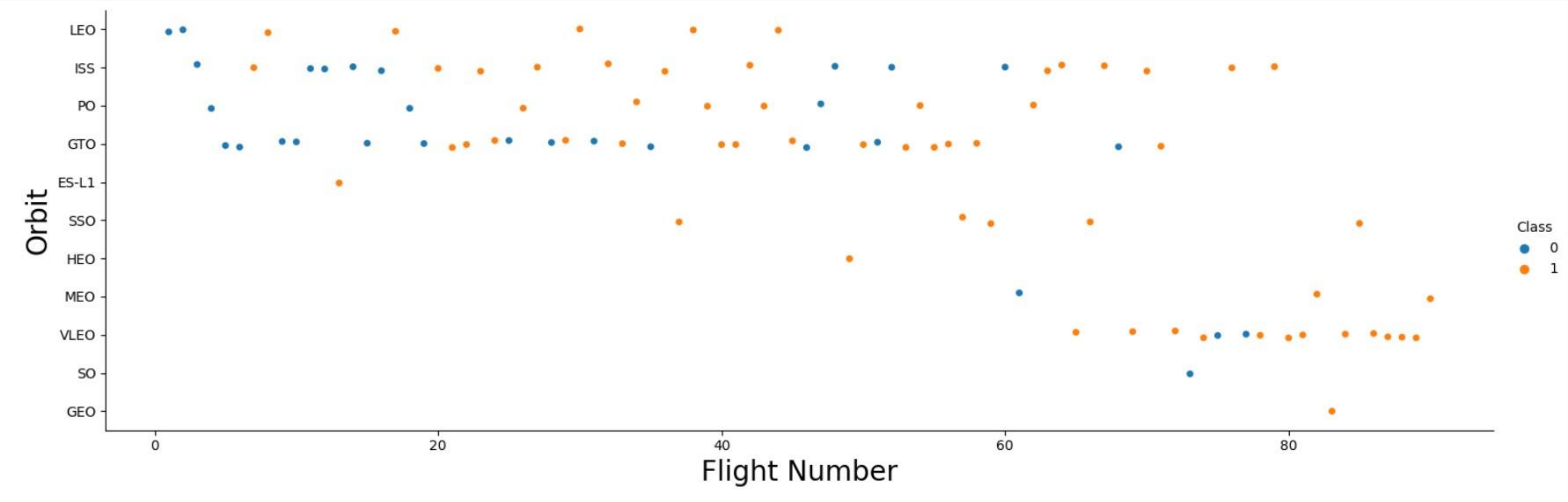
Success Rate



Orbit

Success Rate Scale with 0 as 0%
0.6 as 60% 1 as 100%

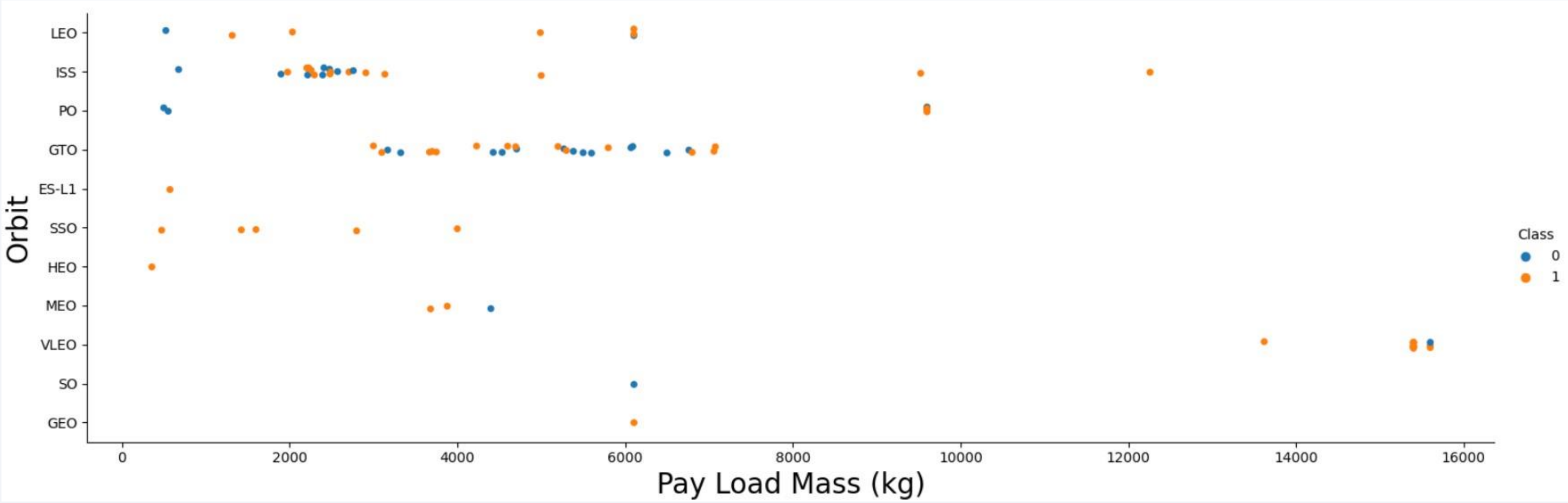
Flight Number vs. Orbit Type



Blue indicates successful launch and orange indicates unsuccessful launch.

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

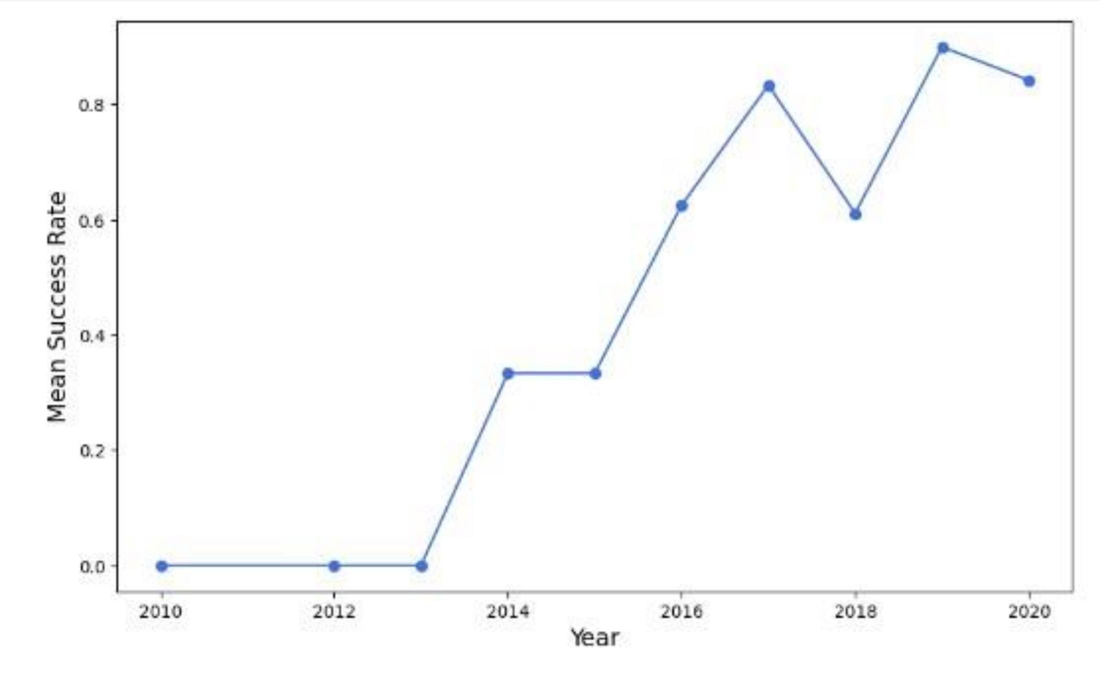


Blue indicates successful launch and orange indicates unsuccessful launch.

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



95% confidence interval (light blue shading)

All Launch Site Names

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

```
% %sql

SELECT DISTINCT
LAUNCH_SITE FROM
SPACEXDATASET;
```

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

Launch Site Names Begin with 'CCA'

```
% %sql

SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

sum_payload:

45596

```
% %sql

SELECT SUM (payload_mass__kg_) AS
SUM_PAYLOAD FROM SPACEXDATASET
WHERE customer = 'NASA (CRS)';
```

sum_payload
45596

Average Payload Mass by F9 v1.1

avg_payload:

2928

```
% %sql

SELECT AVG (payload_mass__kg_) AS
AVG_PAYLOAD FROM SPACEXDATASET
WHERE BOOSTER_VERSION =‘F9 v1.1';
```

avg_payload
2928

First Successful Ground Landing Date

min_date:

2015-12-22

```
% %sql

SELECT MIN(DATE) AS MIN_DATE FROM
SPACEXDATASET WHERE landing__outcome
= 'Success (ground pad)';
```

min_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

% %sql

SELECT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_ > '4000’ AND payload_mass__kg_ < '6000’ AND landing__outcome = 'Success (drone ship)';

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

Total Number of Successful and Failure Mission Outcomes

Success: 100

```
%%sql

SELECT COUNT(*) AS SUCCESS FROM
SPACEXDATASET WHERE
mission_outcome LIKE 'Success%';
```

success
100

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar.
- This likely indicates payload mass correlates with the booster version that is used.

```
%%sql

SELECT booster_version,(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET) AS MAX_Booster FROM SPACEXDATASET ;
```

booster_version	max_booster
F9 v1.0 B0003	15600
F9 v1.0 B0004	15600
F9 v1.0 B0005	15600
F9 v1.0 B0006	15600
F9 v1.0 B0007	15600
F9 v1.1 B1003	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1 B1011	15600
F9 v1.1 B1010	15600
F9 v1.1 B1012	15600
F9 v1.1 B1013	15600
F9 v1.1 B1014	15600

2015 Launch Records

% %sql

SELECT Date, booster_version, launch_site,landing__outcome FROM SPACEXDATASET WHERE landing__outcome = 'Failure (drone ship)’ AND YEAR(Date) = 2015;

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

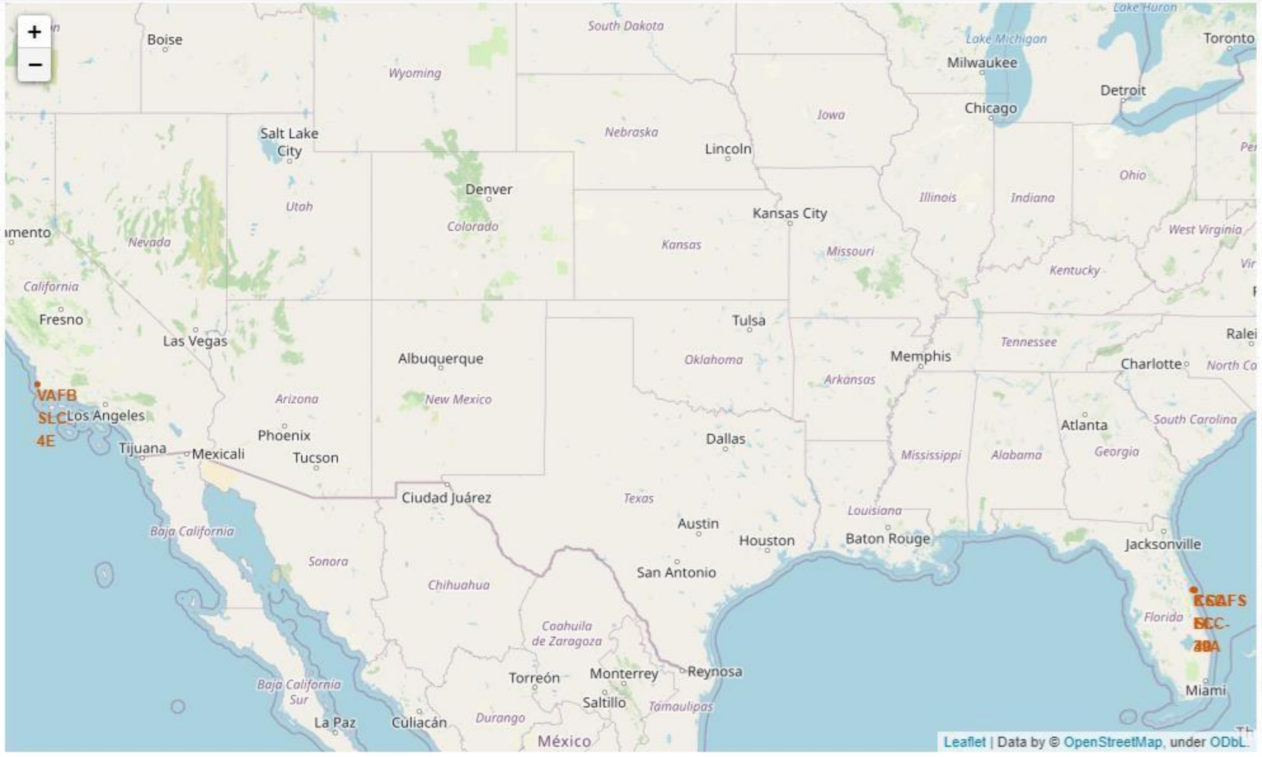
```
% %sql

SELECT landing__outcome FROM SPACEXDATASET WHERE
Date > '2010-06-04' AND Date < '2017-03-20' GROUP BY
landing__outcome ORDER BY COUNT(landing__outcome)
DESC;
```

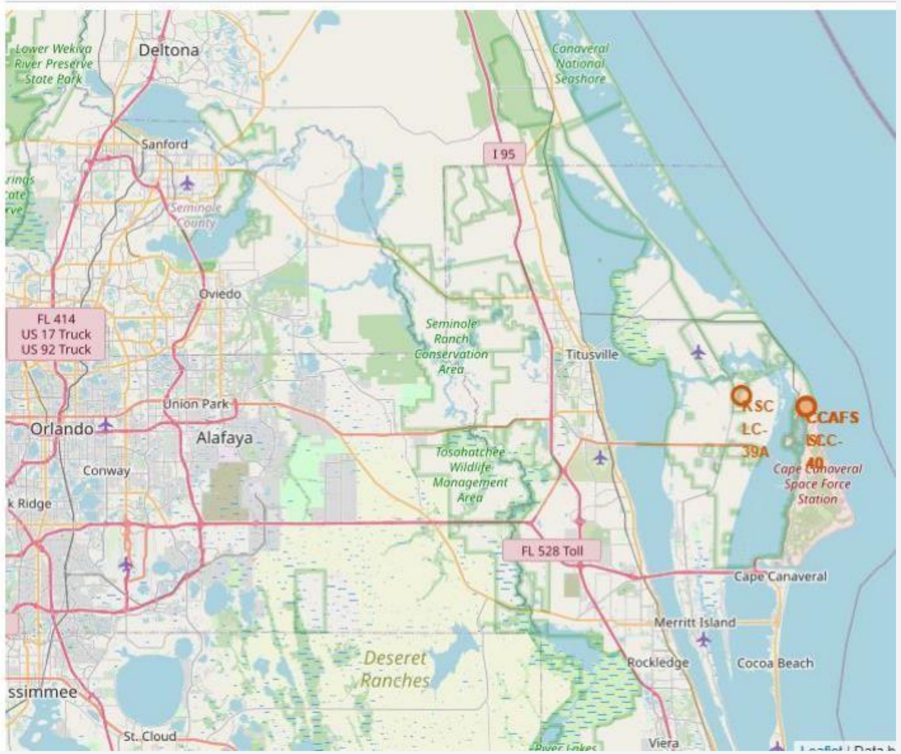
landing__outcome
No attempt
Failure (drone ship)
Success (drone ship)
Controlled (ocean)
Success (ground pad)
Uncontrolled (ocean)
Failure (parachute)
Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

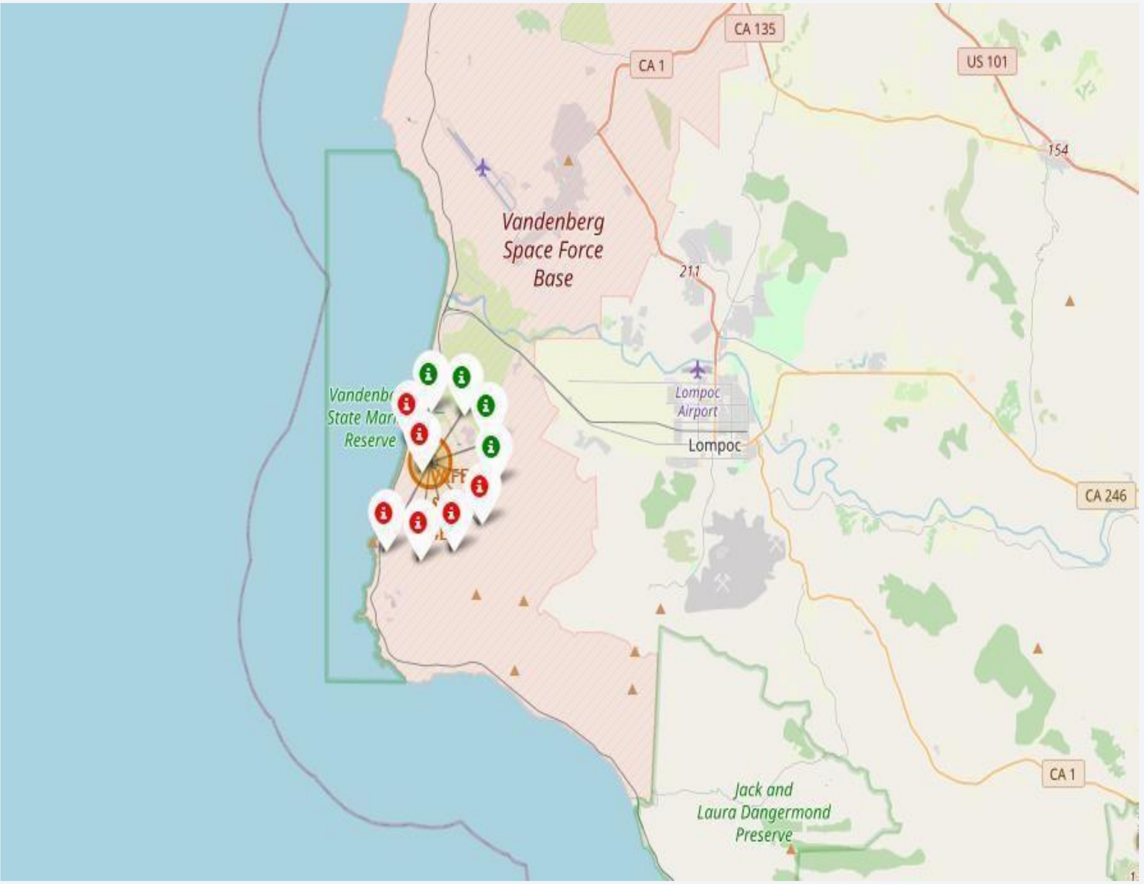
Launch Site Locations





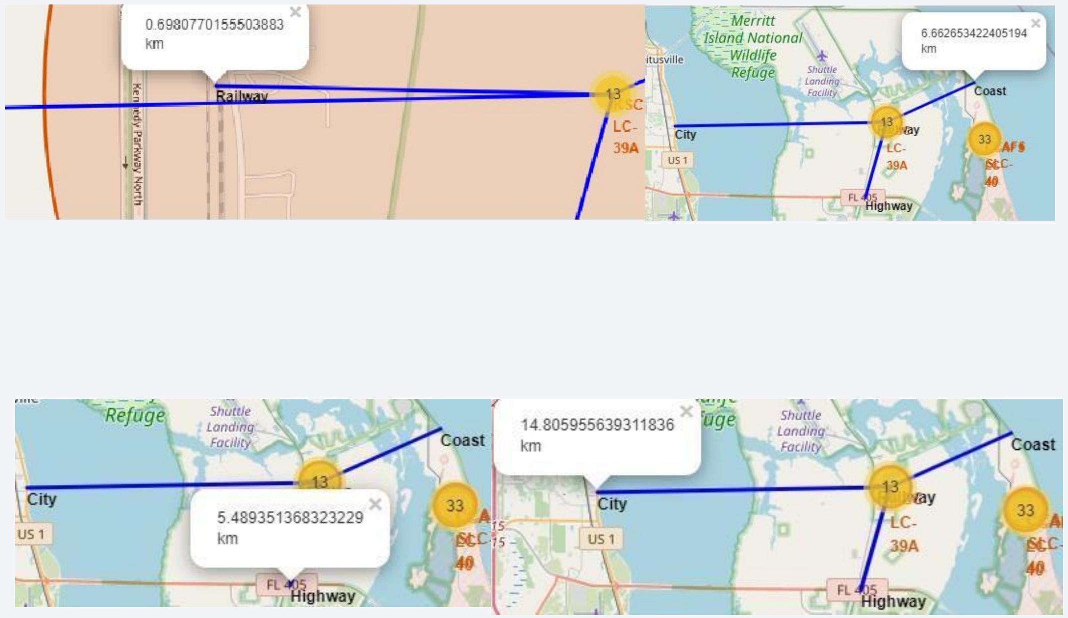
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

Key Location Proximities

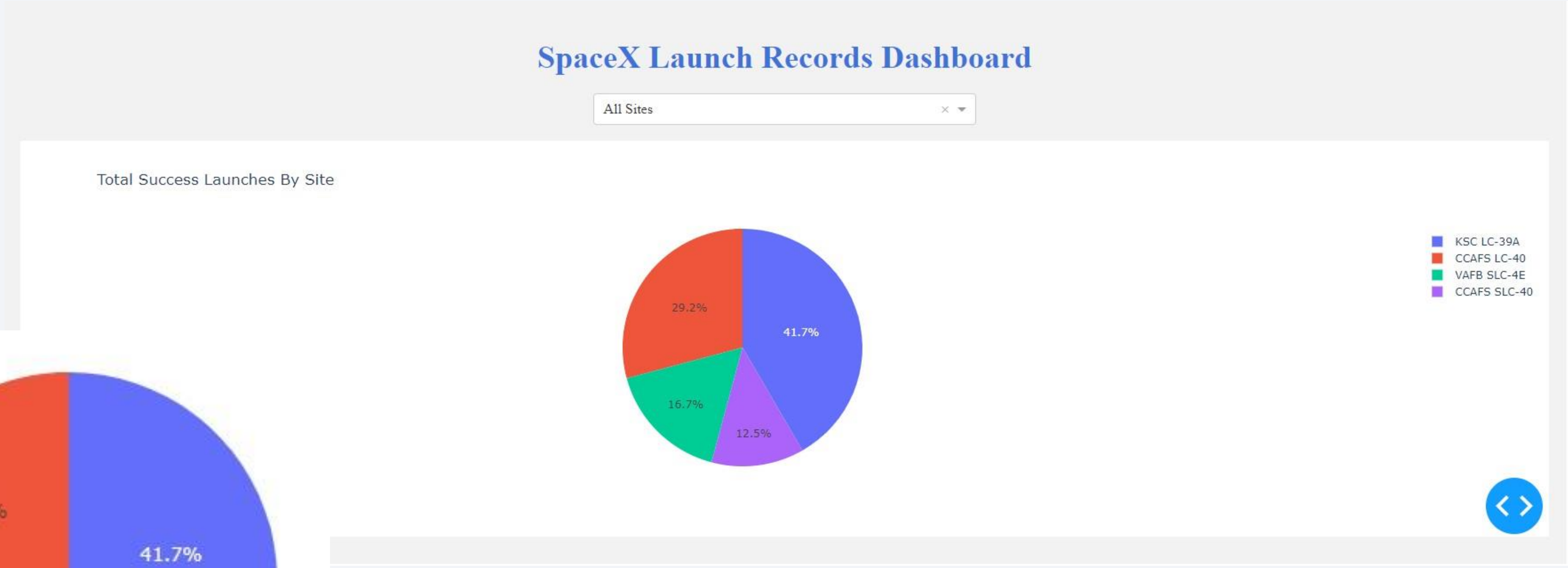


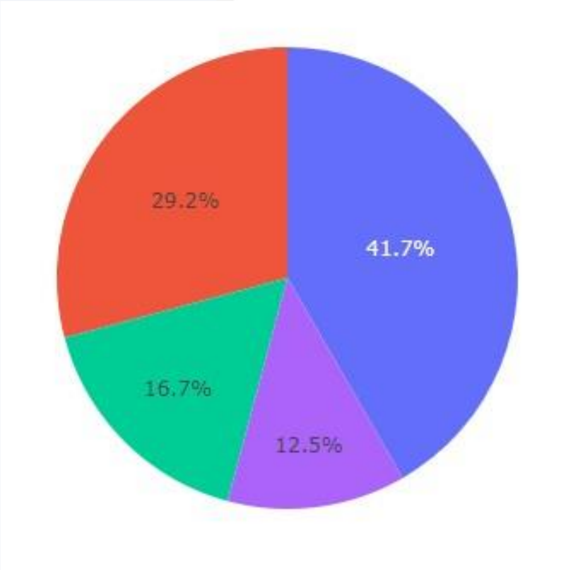
Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Section
4

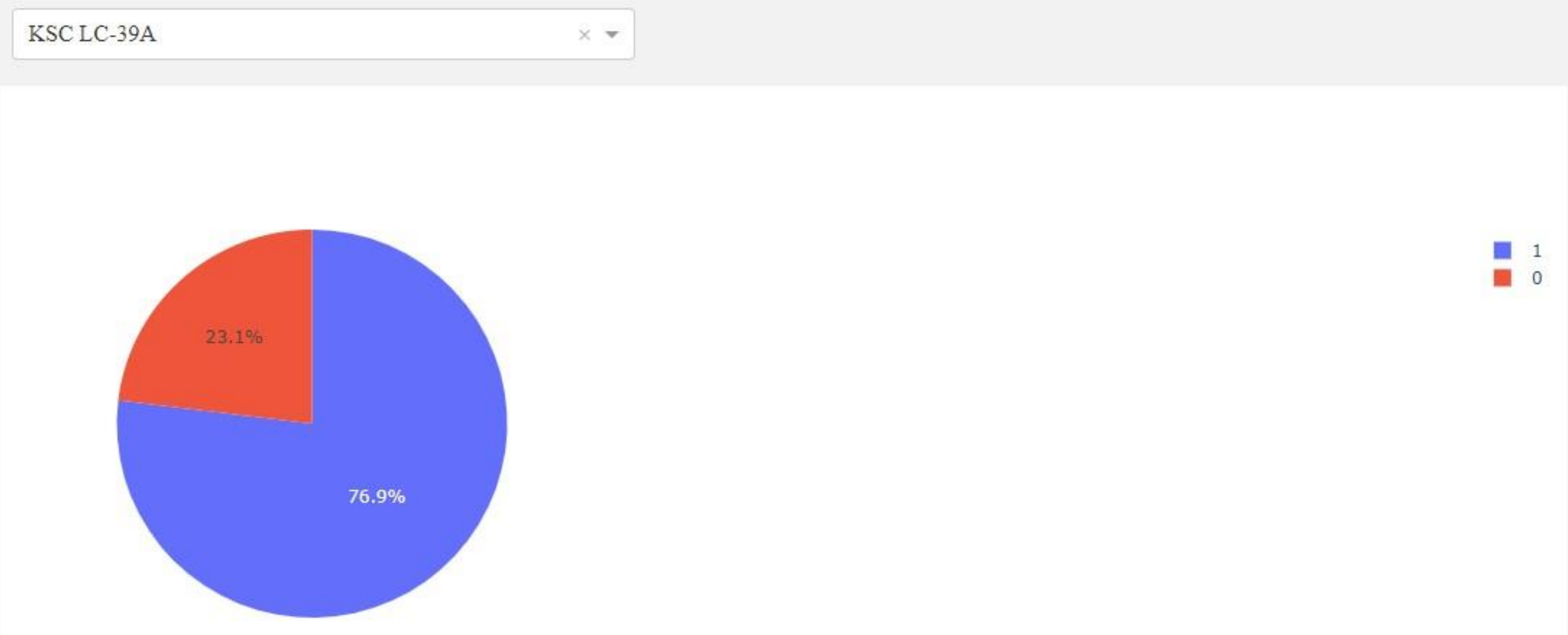
Build a Dashboard with Plotly Dash

Total success launches by all sites

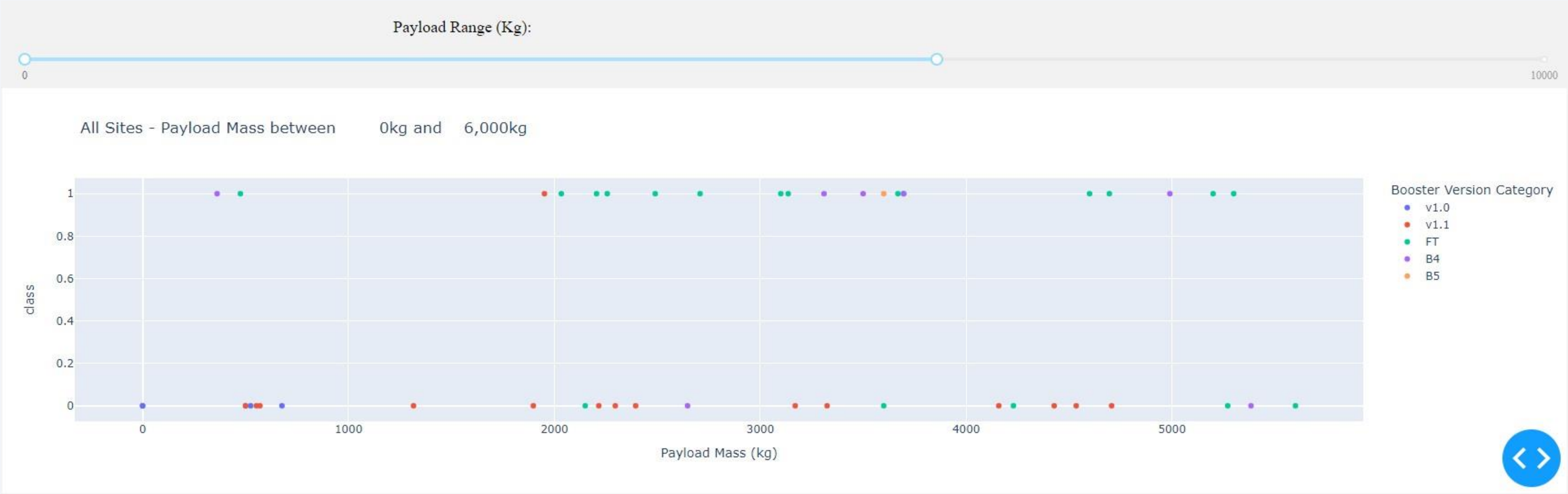




KSC LC-39A achieved a success rate of 76.9%



Payload vs Launch outcome



Section 5

Predictive Analysis (Classification)

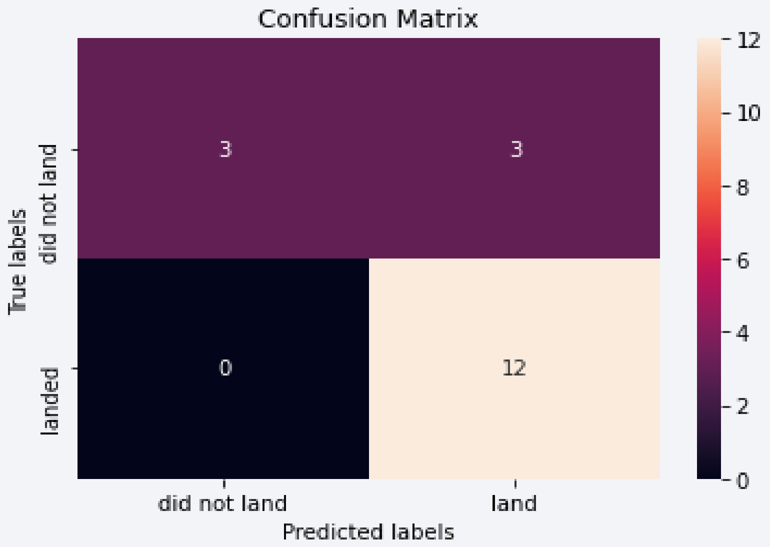
Classification Accuracy

All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

We likely need more data to determine the best model.

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
Our models over predict successful landings.

Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

GitHub Repository URL:

https://github.com/snowieeeee/IBM-Applied-Data-Science-Capstone

Instructors:

Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Special Thanks to All Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!