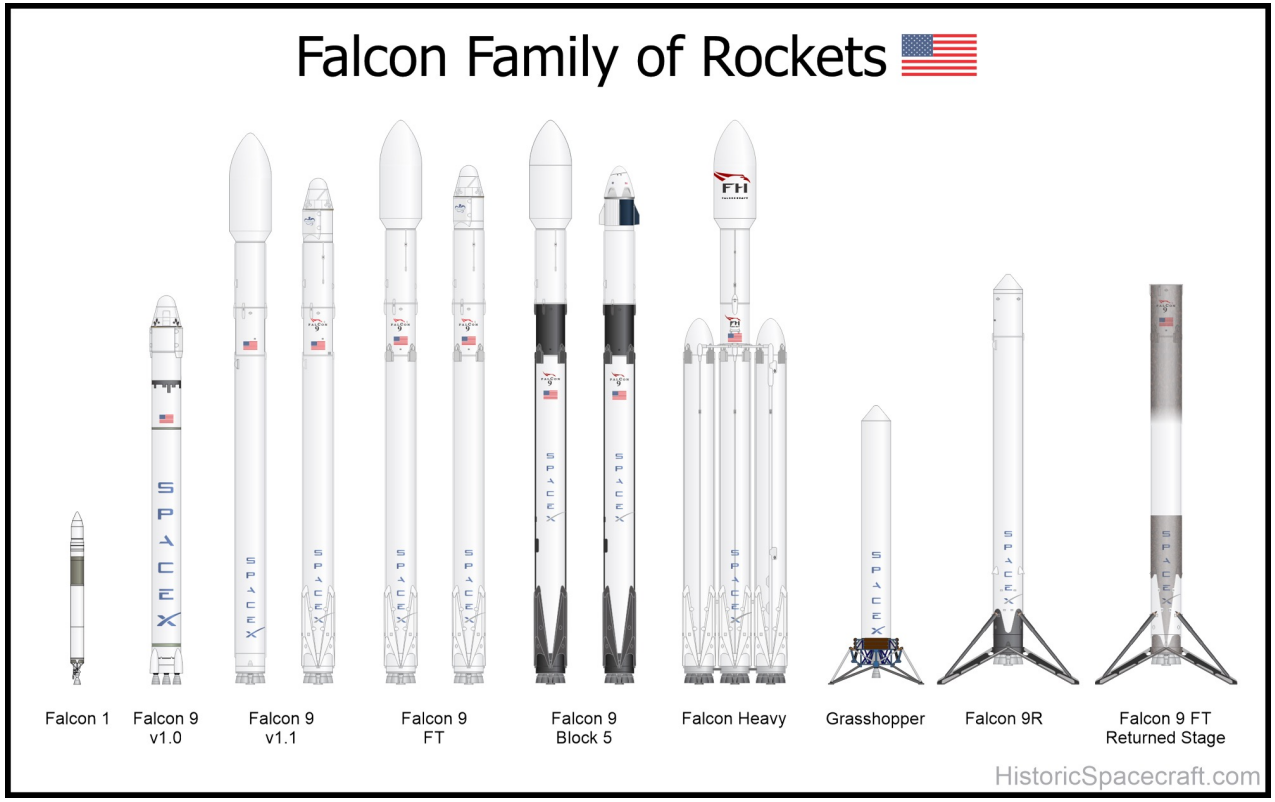

Space X Falcon 9 First Stage Landing Prediction
EdX/IBM DS0720EN
Data Science and Machine Learning Capstone Project
Anne Fengyan Shi 01/04/2026

OUTLINE

- Executive summary
- Introduction
- Methodology
  - Data collection
  - Data wrangling
  - Data exploration
  - Data analyses
  - Data visualization
  - ML Prediction
- Conclusion
- Appendix



EXECUTIVE SUMMARY

- Overall objective: using data to predict Falcon 9 first stage landing
- Data collection and wrangling
  - Extract data from SpaceX API
  - Web scraping launch record from Wikipedia
- Exploratory data analysis (EDA)
  - EDA with SQL
  - EDA with visualization
- Interactive visual analytics
  - With Folium
  - Dashboard with Ploty Dash
- Prediction by machine learning methods

Introduction

- Background
  - This project predicts if the Falcon 9 first stage will land successfully.
  - SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars. In comparison, other providers cost upward of 165 million dollars each. Much of the savings is owing to the fact that SpaceX can reuse the first stage.
  - If we can determine if the first stage will land, we can determine the cost of a launch. An alternate company can use this information to bid against SpaceX for a rocket launch.
- Strategy
  - Determine factors facilitating/obstructing landing success
  - Predict landing success rate with varied machine learning algorisms

Methodology

- Data collection
- Data wrangling
- EDA with SQL
- EDA with visualization
- Visualization with Folium
- Visualization via dashboard
- Predictive analysis with machine learning models

Data Collection From SpaceX API

- Extract rocket launch data from SpaceX API on:
  - Rocket
  - Payloads
  - Launchpad
  - Cores
- Turn data to Pandas dataframe
- Filter data to include only Falcon 9 launches
- Replace missing values of payloadMass with its mean
- Save data in csv

Data Collection With Web scraping

- Request Falcon 9 launch Wikipedia pages from its URL
- Extract column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML table
- Save data in csv

Data Wrangling

- Perform preliminary exploratory data analysis (EDA)
  - Calculate number of launches by site
  - Calculate number and occurrence by orbit
  - Calculate number and occurrence of outcome by orbit type
- Create a landing outcome label
  - Landing_class == 1: successful landing
  - Landing_class == 0: unsuccessful landing

Exploratory Data Analysis (EDA) with SQL

Query on:

- Names of launch sites
- Records where launch sites begin with 'KSC'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster F9 v1.1
- Date of the first successful landing in drone ship
- Booster versions that have carried the maximum payload mass
- Boosters which have success in ground pad with payload mass > 4000 but < 6000
- Total number of successful and failed outcomes
- Landing outcomes by drone ship, booster version and launch site name
- Months , successful landing outcomes in ground pad, booster versions, and launch sites in 2017

EDA with SQL Results

1. Names of the launch sites
   - CCAFS LC-40
   - VAFB SLC-4E
   - KSC LC-39A
   - CCAFS SLC-40

2. First 5 records where launch sites begin with the string 'KSC'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

EDA with SQL Results (cont.)

3. Total payload mass carried by boosters launched by NASA (CRS)
    - 45596 KG

4. Average payload mass carried by booster version F9 v1.1
    - 2928.4 KG

5. Date of the first successful landing outcome in drone ship
    - 2016-04-18

6. Boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
    - F9 FT B1032.1
    - F9 B4 B1040.1
    - F9 B4 B1043.1

EDA with SQL Results (cont.)

7. Total number of successful and failure mission outcomes

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

8. Booster versions that have carried the maximum payload mass

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

EDA with SQL Results (cont.)

9. Months , successful landing outcomes in ground pad, booster versions, and launch sites in 2017

month	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

EDA with SQL Results (cont.)

10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

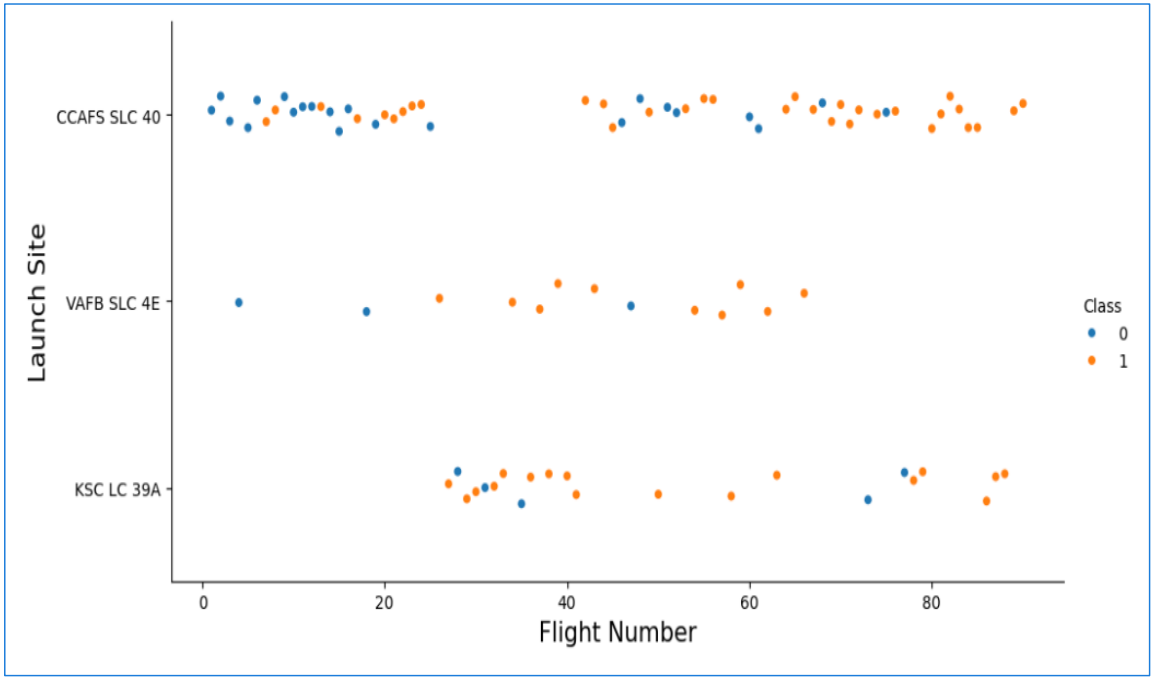
Landing_Outcome	outcome_count	rank
Precluded (drone ship)	1	8
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Success (ground pad)	3	4
Controlled (ocean)	3	4
Success (drone ship)	5	2
Failure (drone ship)	5	2
No attempt	10	1

Exploratory Data Analysis (EDA) with Visualization

- Visualizing relationship between
  - Flight number and launch site
  - Payload mass and launch site
  - Success rate and orbit type
  - Flight number and orbit type
  - Payload mass and orbit type
- Visualizing the launch success yearly trend

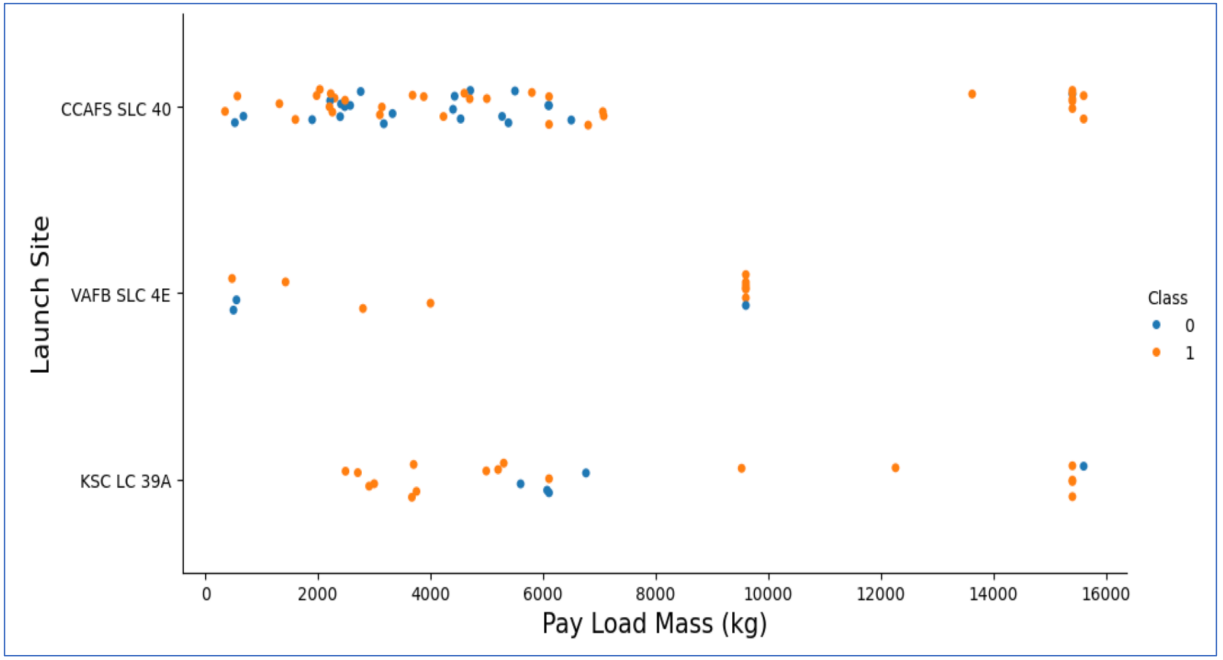
EDA with Visualization

Relationship between flight number and launch site



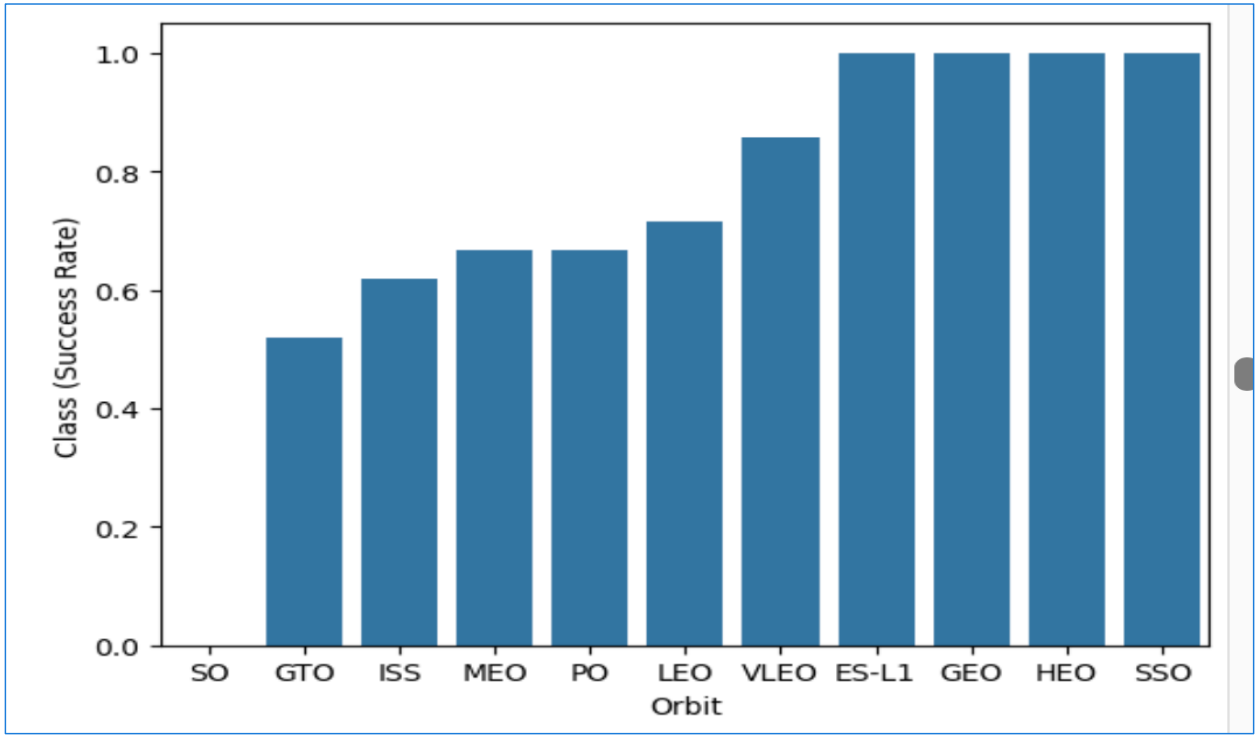
EDA with Visualization

Relationship between payload mass and launch site



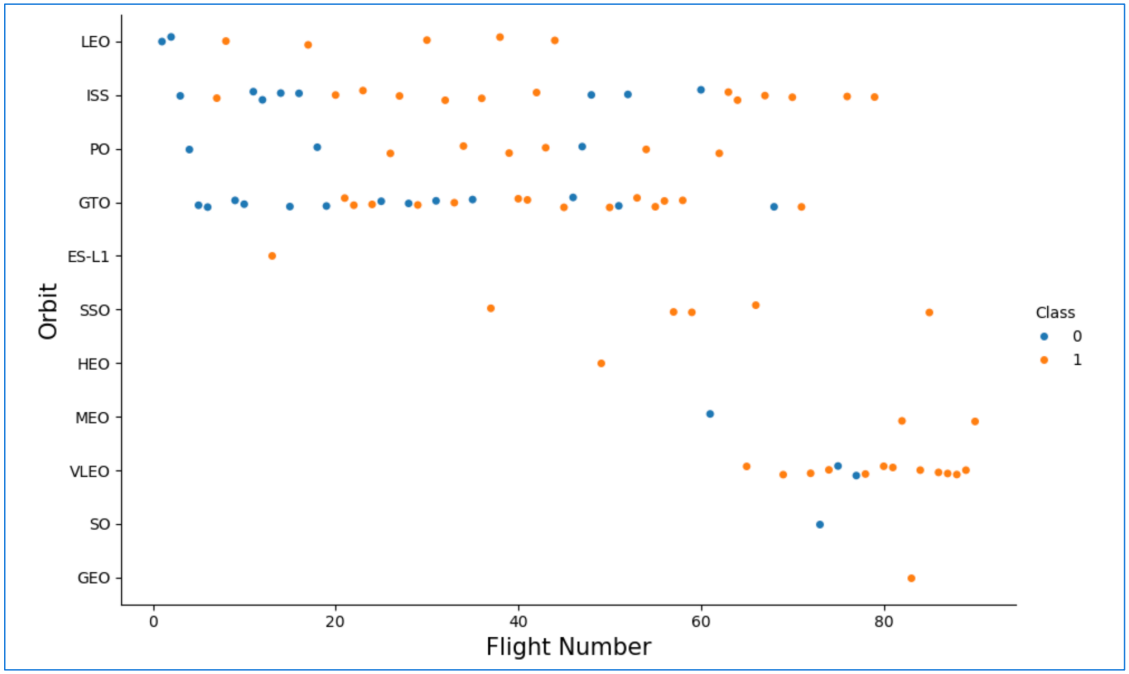
EDA with Visualization

Success rate by orbit type



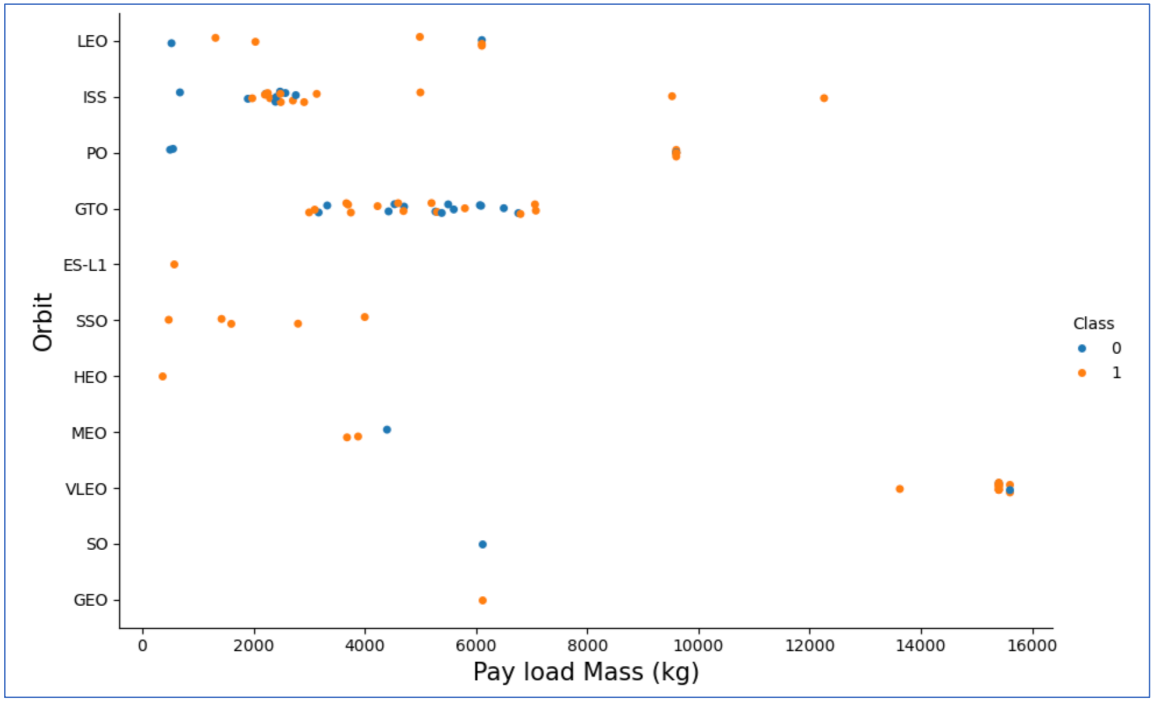
EDA with Visualization

Relationship between flight number and orbit type



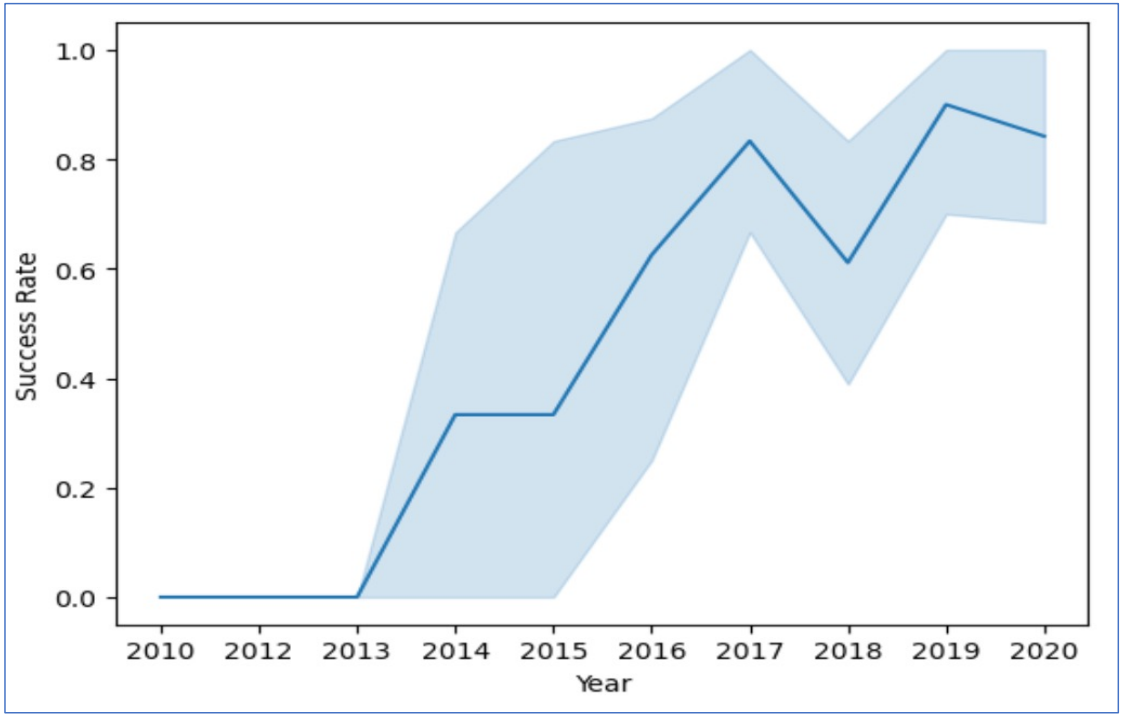
EDA with Visualization

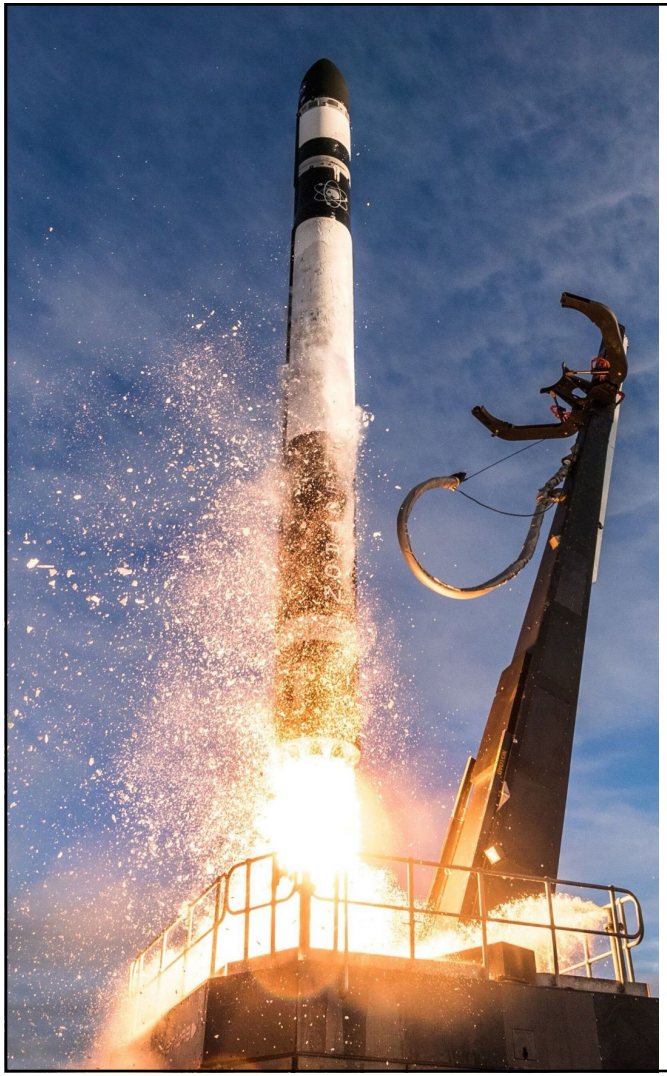
Relationship between payload mass and orbit type



EDA with Visualization

Launch success trend, 2010-2020



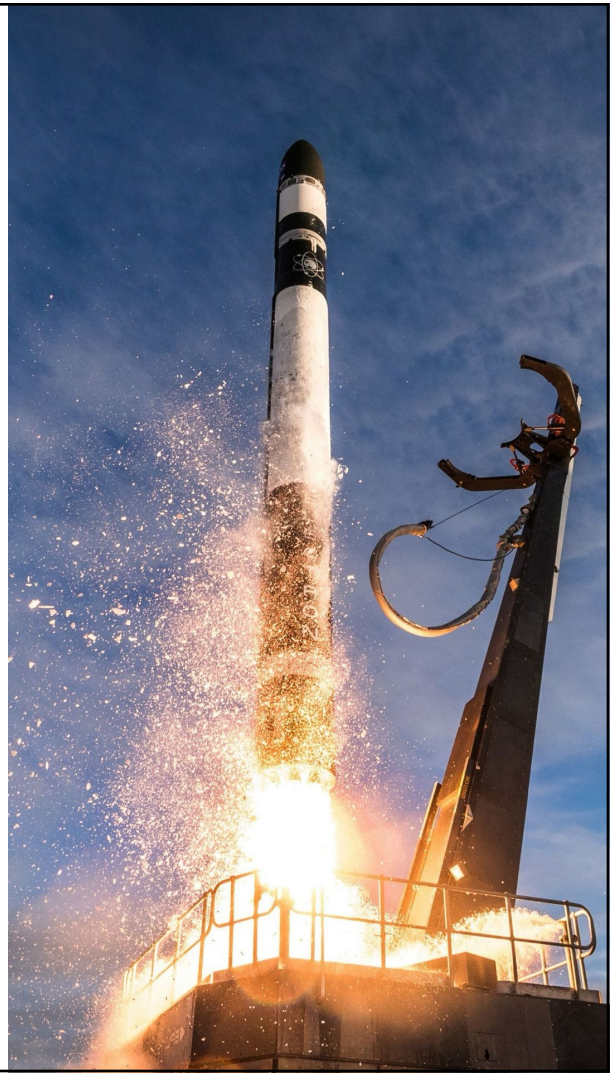


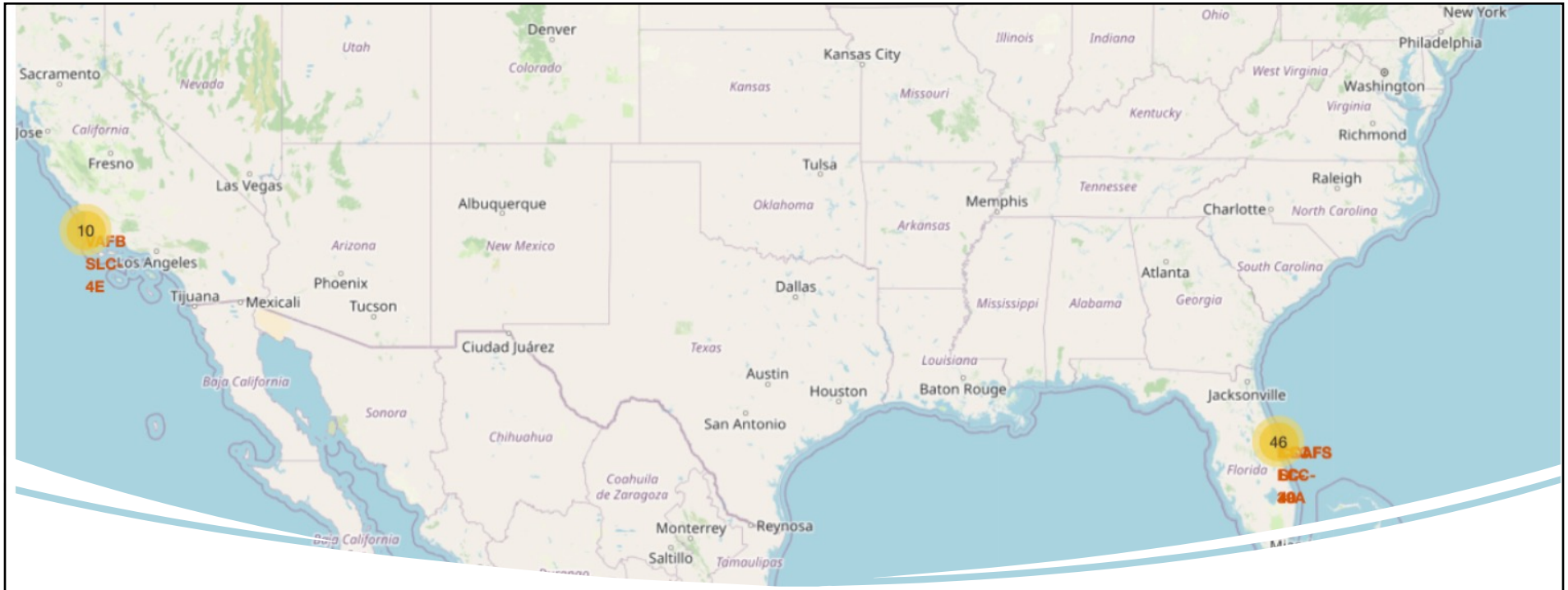
EDA with Visualization - Primary Findings

- Chances of success has improved with the increasing flight numbers
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass
- Success rate varied by orbit type
  - Orbits ES-L1, GEO, HEO & SSO have 100% success rate
  - Orbit SO has 0% success rate
  - success rates for other orbits vary between 50% to nearly 90%

EDA with Visualization
- Primary Findings (cont.)

- In the Low Earth Orbit (LEO) missions, success is correlated with the number of flights. Conversely, in the Geostationary Transfer Orbit (GTO) missions, there is no apparent relationship between number of flights and success
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present
- Between 2013 and 2020, the success rate kept increasing except a little dip in 2017

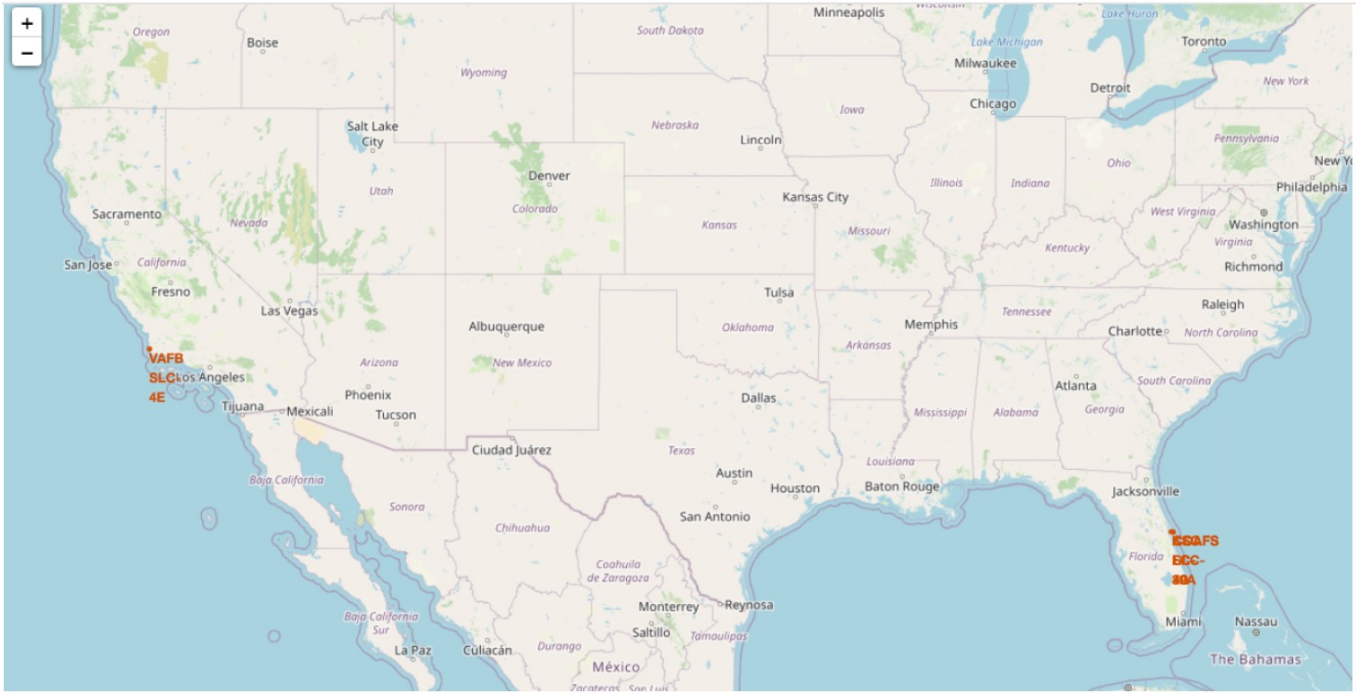


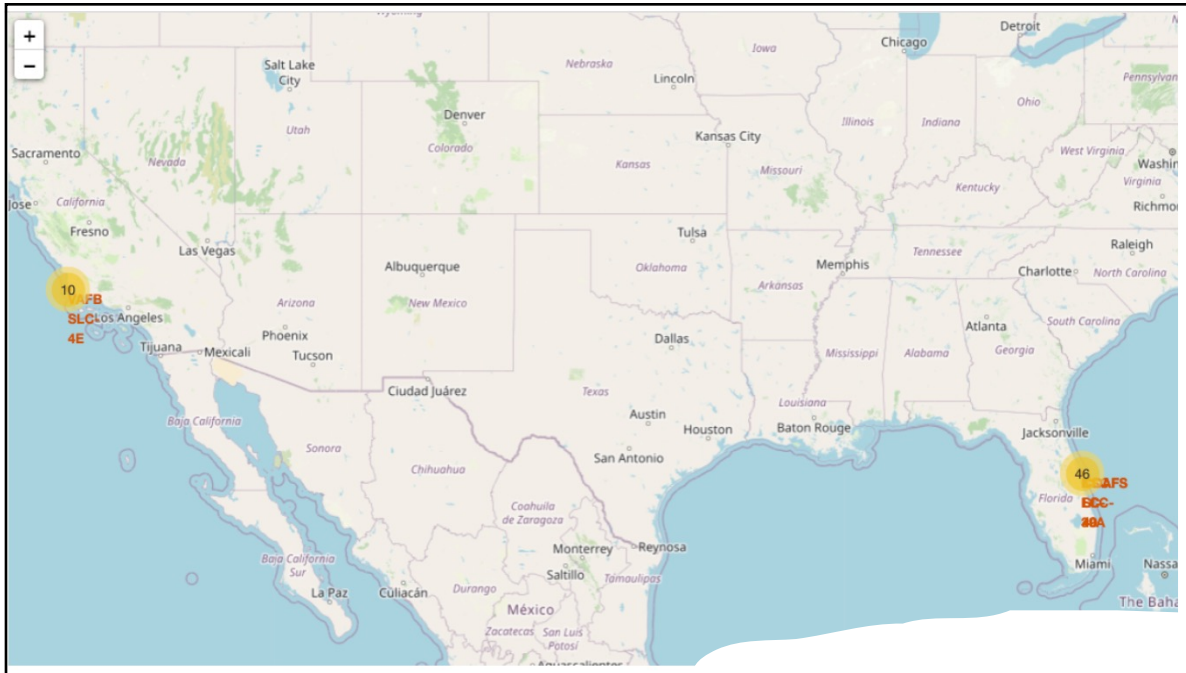


Interactive Visual Analytics with Folium (IVAF)

- Mark all launch sites on the map
- Mark successful/failed launches on the map
- Calculate distances between a launch site to its proximities

IVAF Results: Launch Sites





IVAF Results: Successful and Failed Launches

Interactive Visual Analytics with Folium (IVAF)

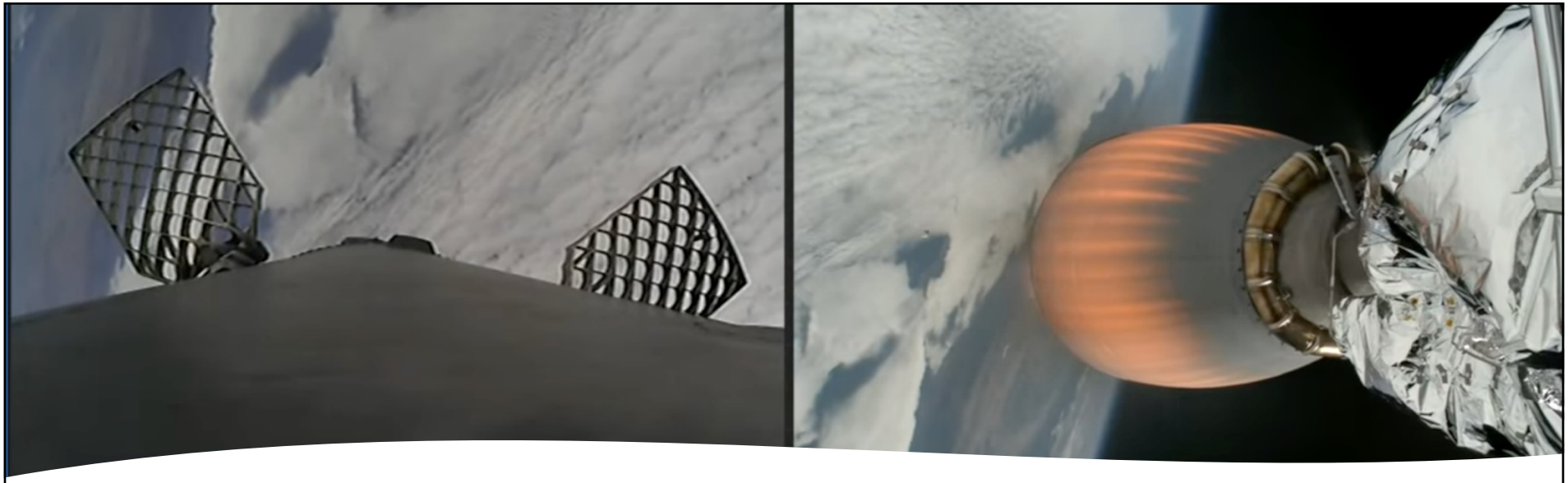
Distances between a launch site to its proximities

- To highway: 0.58 KM
- To coastline: 0.88 KM
- To Railroad: 1.28 KM
- To nearest city: 51.43 KM

Interactive Visual Analytics with Folium (IVAF)

Findings

- Launch sites are close to highway and railroad to facilitate transportation
- Launch sites are close to coastline and equator to utilize the Earth rotation
- Launch sites are distant from cities to keep population and property safe

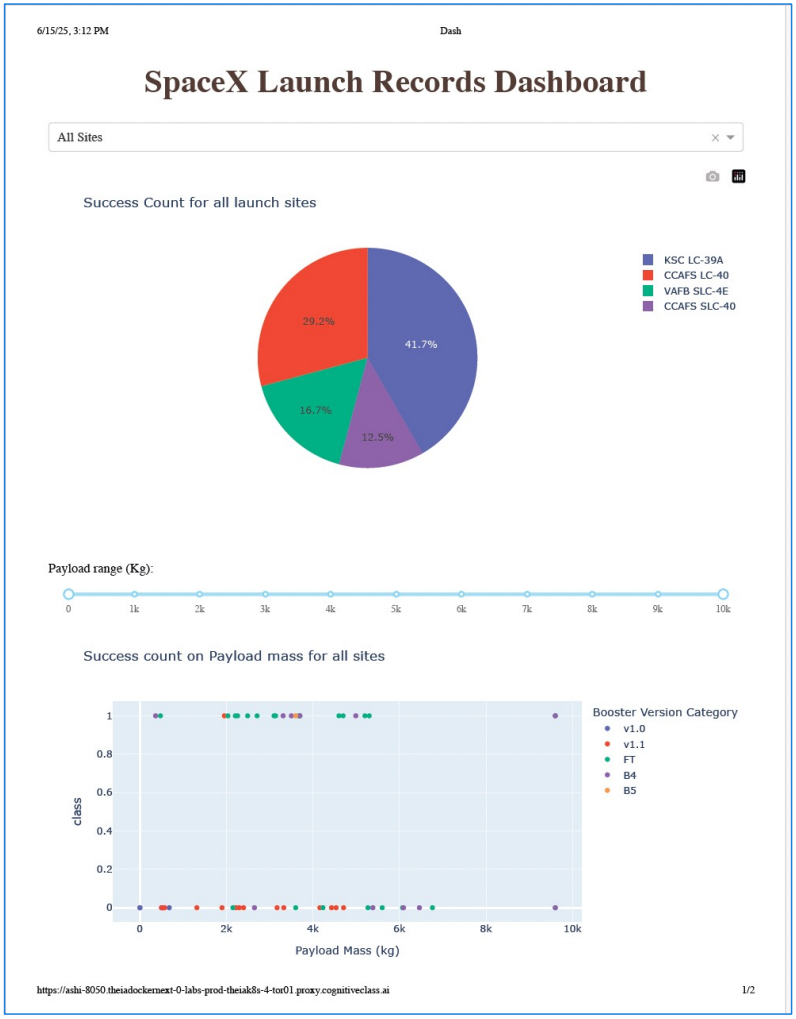


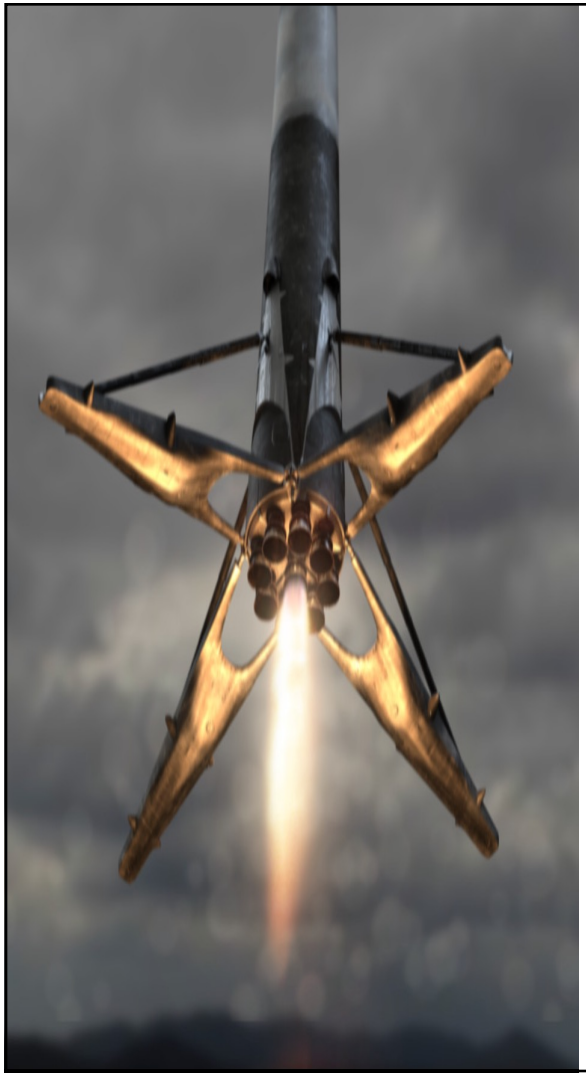
Interactive Visual Analytics on Dashboard

- Building a Plotly Dash application
- Build a dropdown list of launch sites
- Add callback function for success pie chart by selecting site dropdown
- Add a range slider for selecting payload
- Add callback function for scatter plot of success payload

Plotly Dash Dashboard

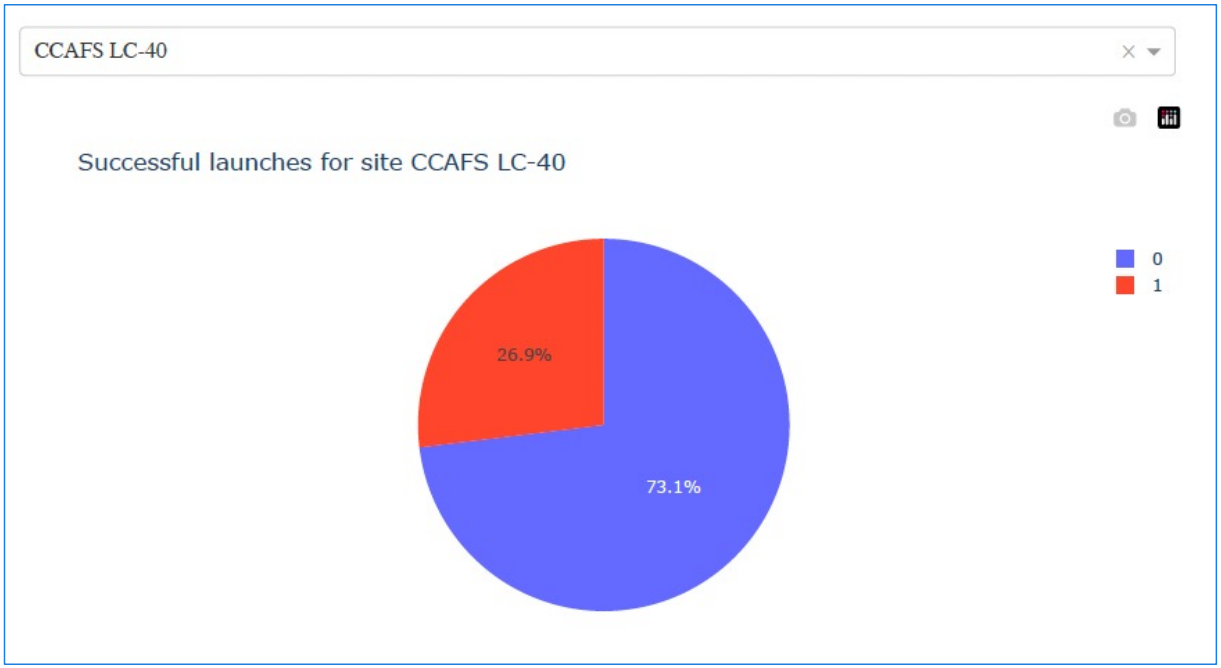
Snapshots Slide 1

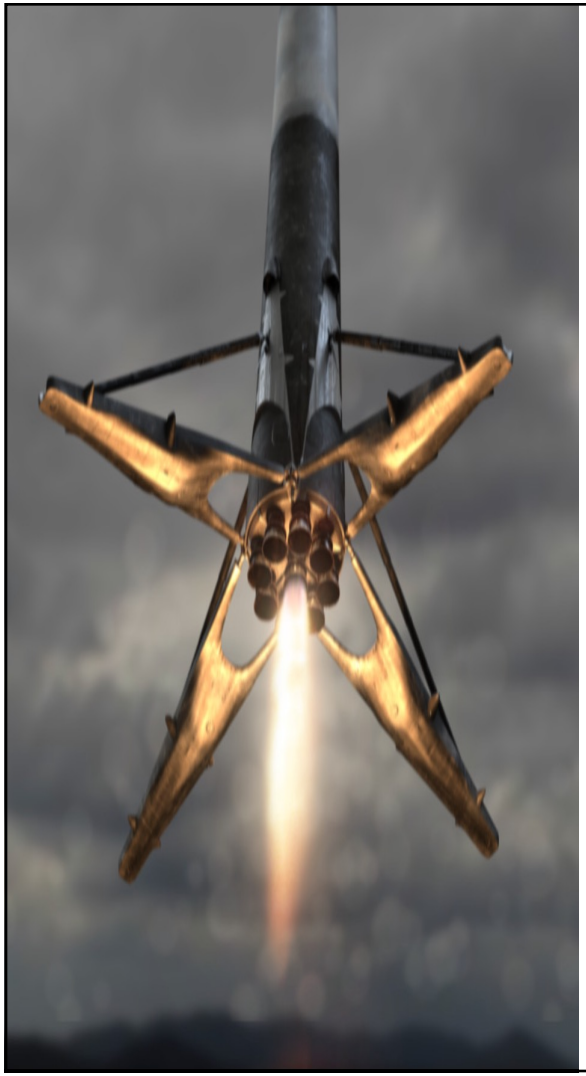




Plotly Dash Dashboard

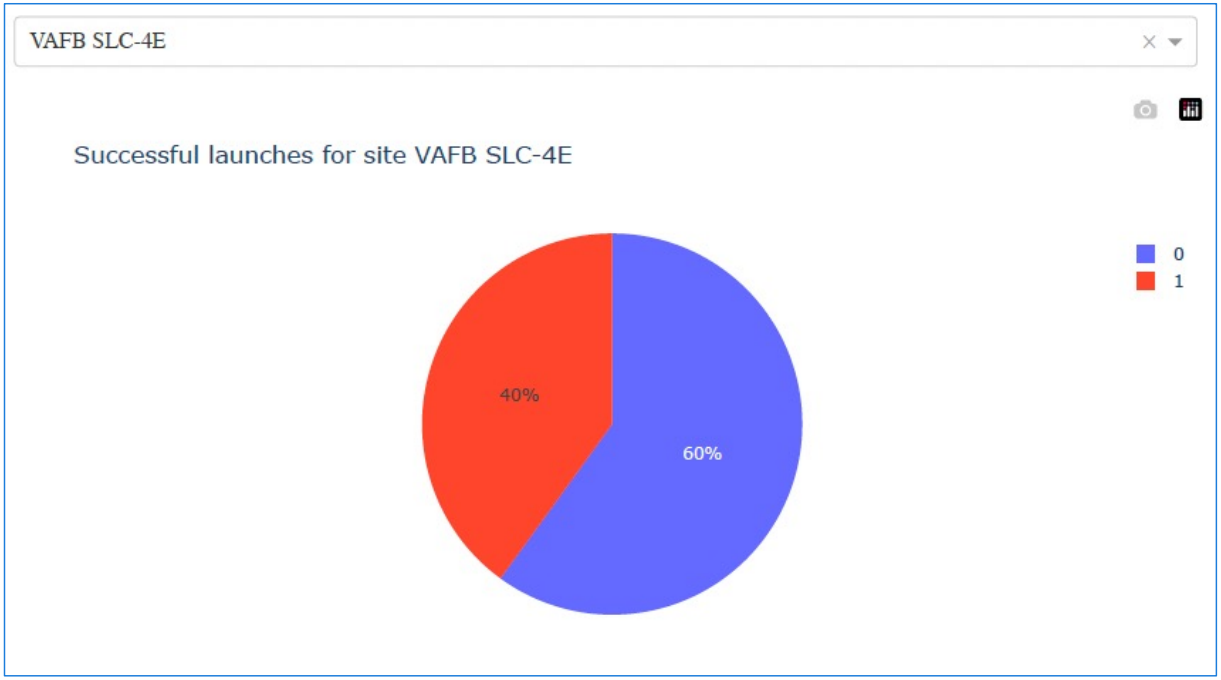
Snapshots Slide 2

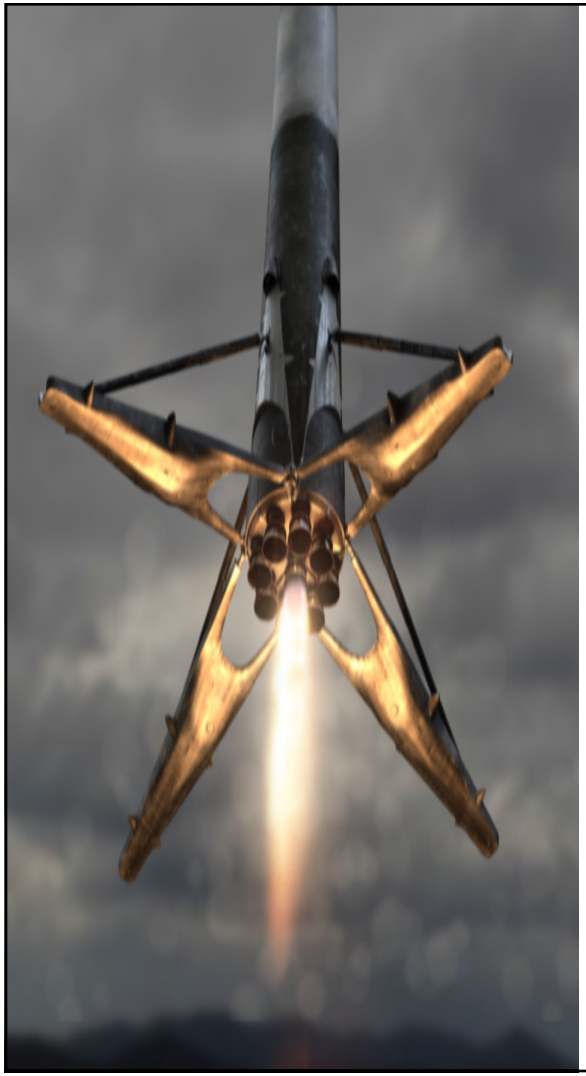




Plotly Dash Dashboard

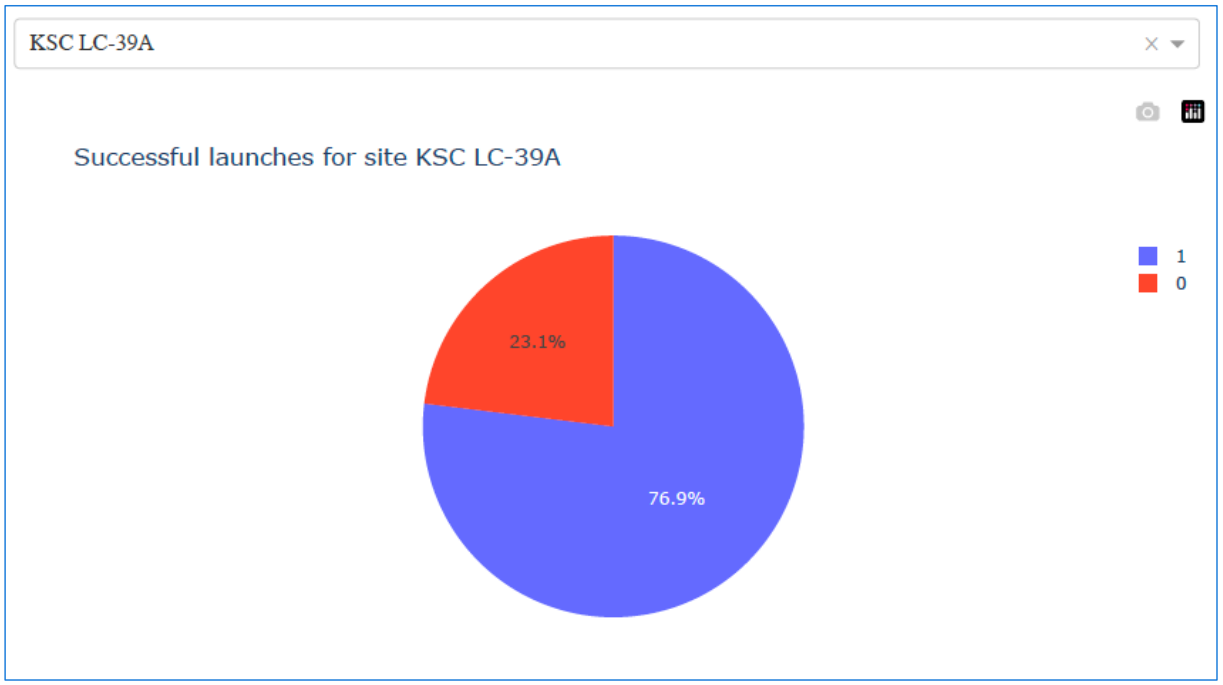
Snapshots Slide 3

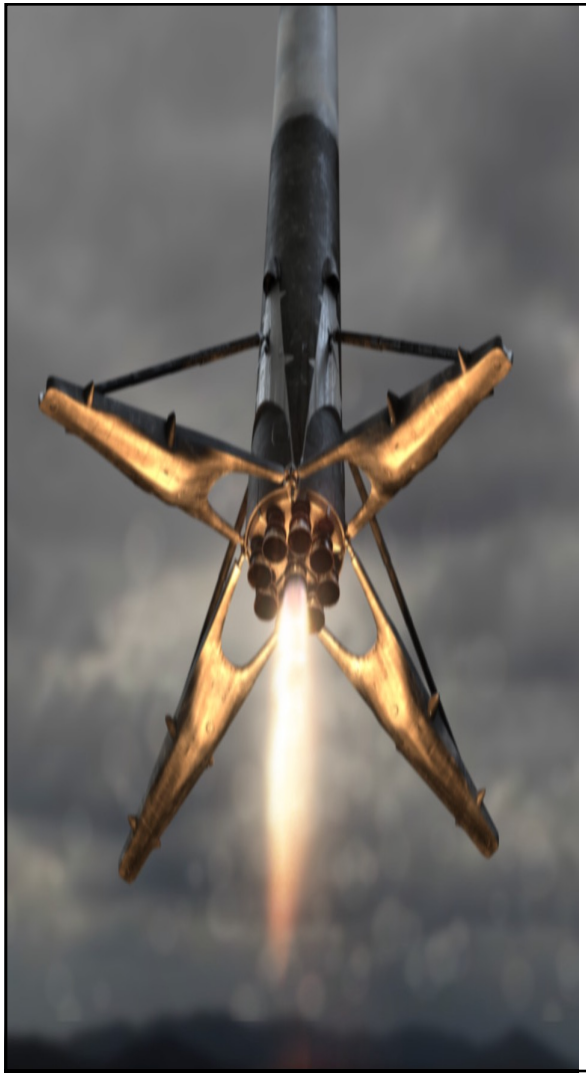




Plotly Dash Dashboard

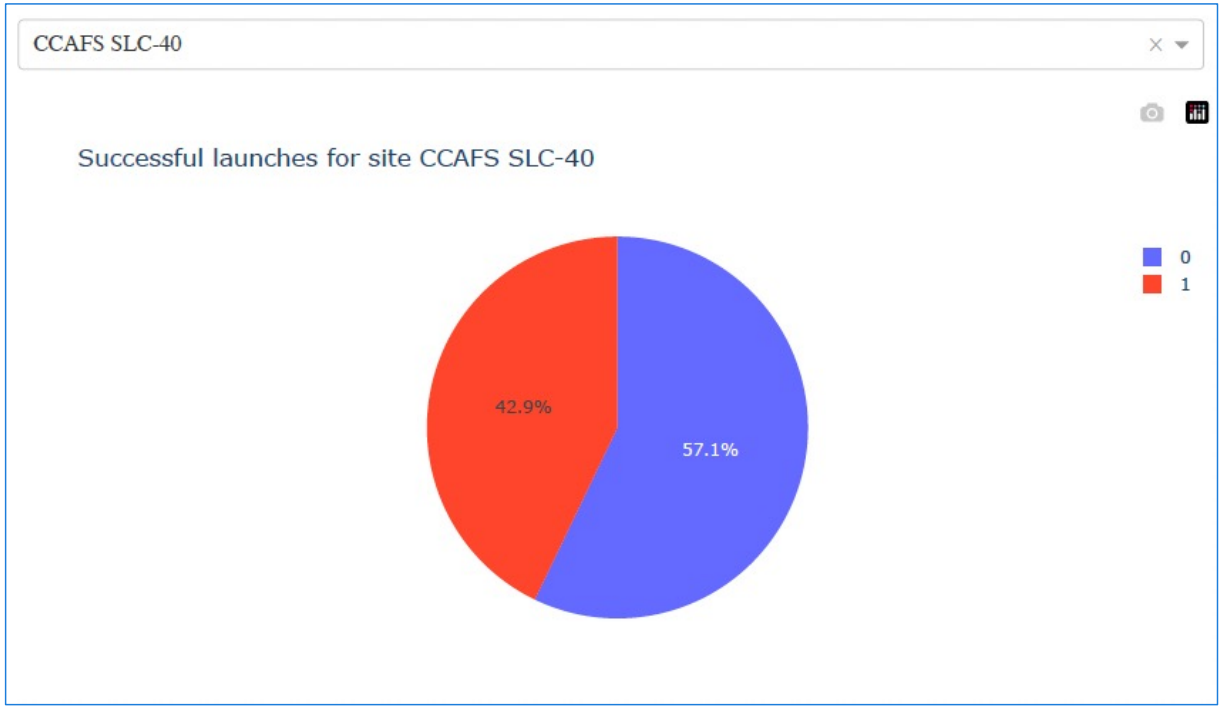
Snapshots Slide 4





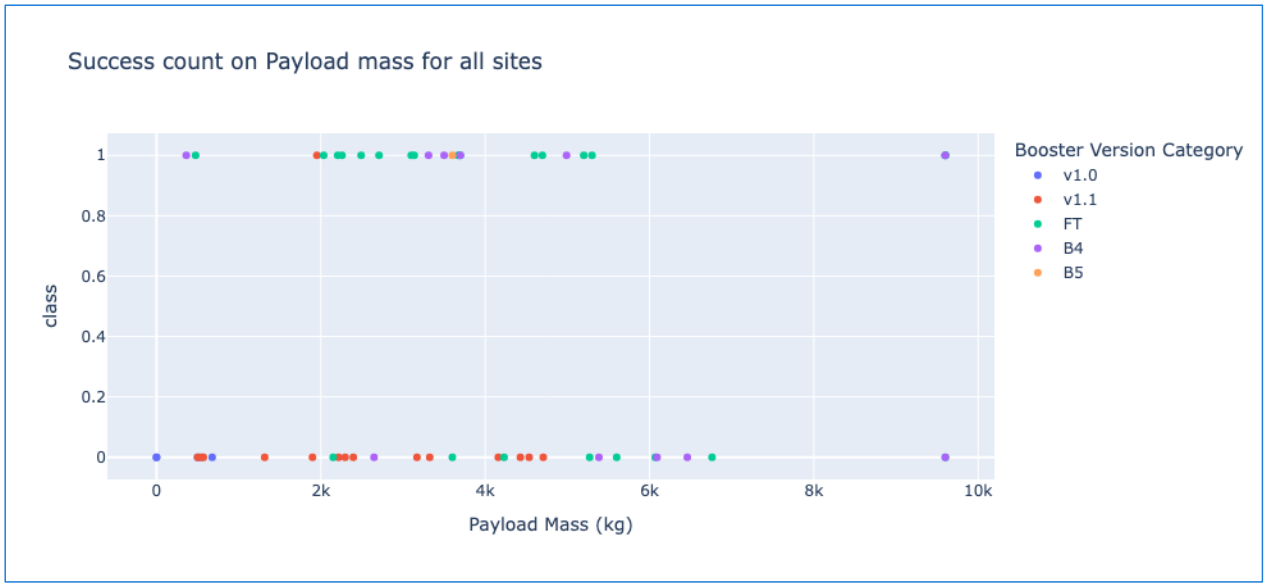
Plotly Dash Dashboard

Snapshots Slide 5



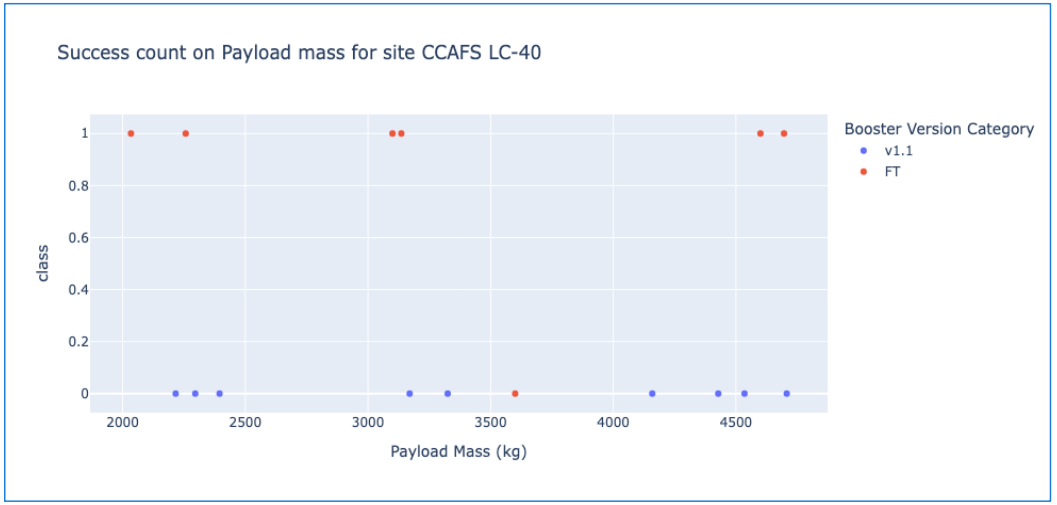
Plotly Dash Dashboard

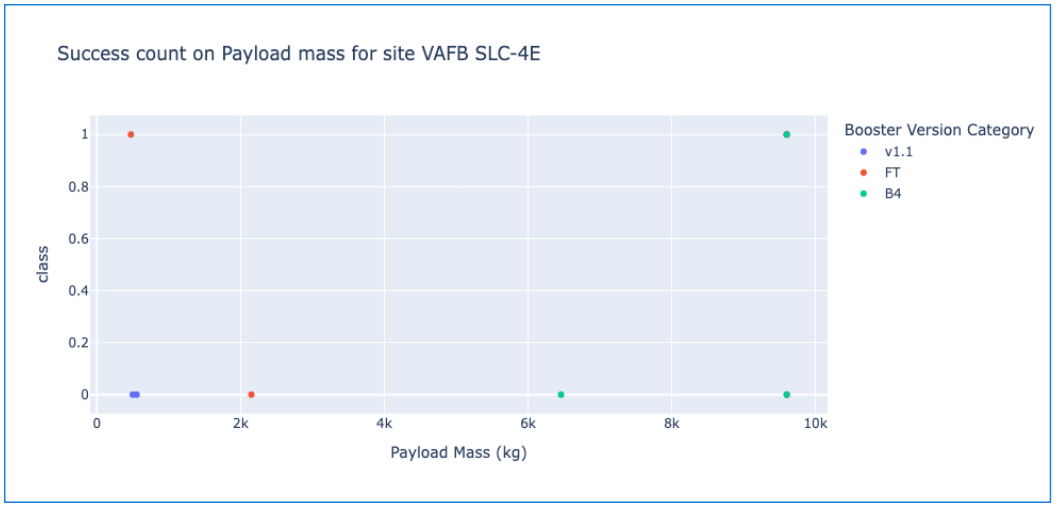
Snapshots Slide 6



Plotly Dash Dashboard

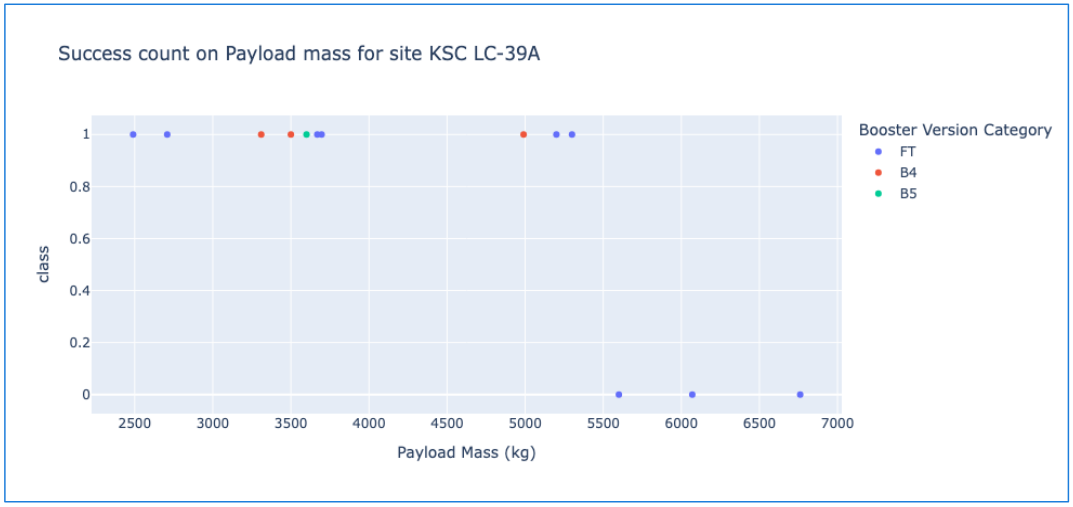
Snapshot Slide 7

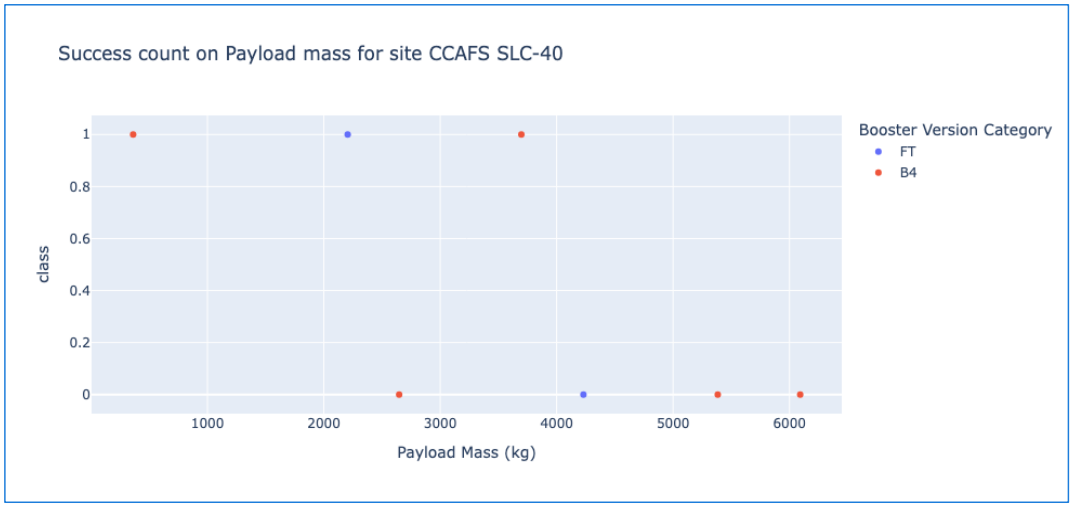




Plotly Dash Dashboard

Snapshots Slide 8





Dashboard Analytics Findings

- KSC LC-39A has largest success count (10 out of 24 successes, 41.7%)
- KSC LC-39A also has the highest success rate (76.9%)
- Payload range 2K – 6K and 9K-10K have the highest launch success rate
- Payload range 0-2K and 6-9K have the lowest success rate

Dashboard Analytics Findings (cont.)

- F9 Booster version B5 has the highest success rate, followed by FT and B4
- V1.1 has very low success rate (6.7%) and none of V1.0 succeeded (0% success)

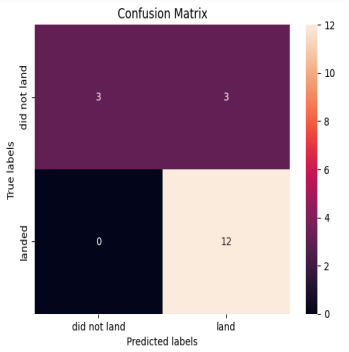
Predictive Analytics with Machine Learning

- Load data
- Standardize data
- Split data into training and testing sets
- Built classification models and test accuracy on:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbors (KNN)
- Identify the model with best performance

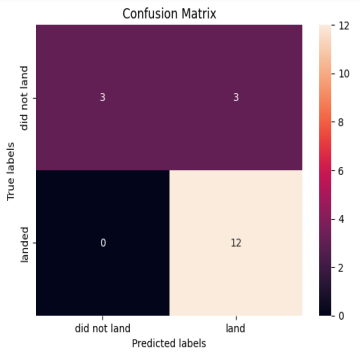
Predictive Analytics with Machine Learning

Results in confusion matrixes

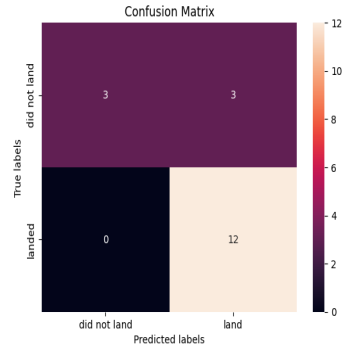
Logistic Regression



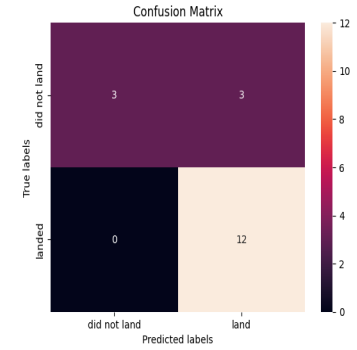
SVM

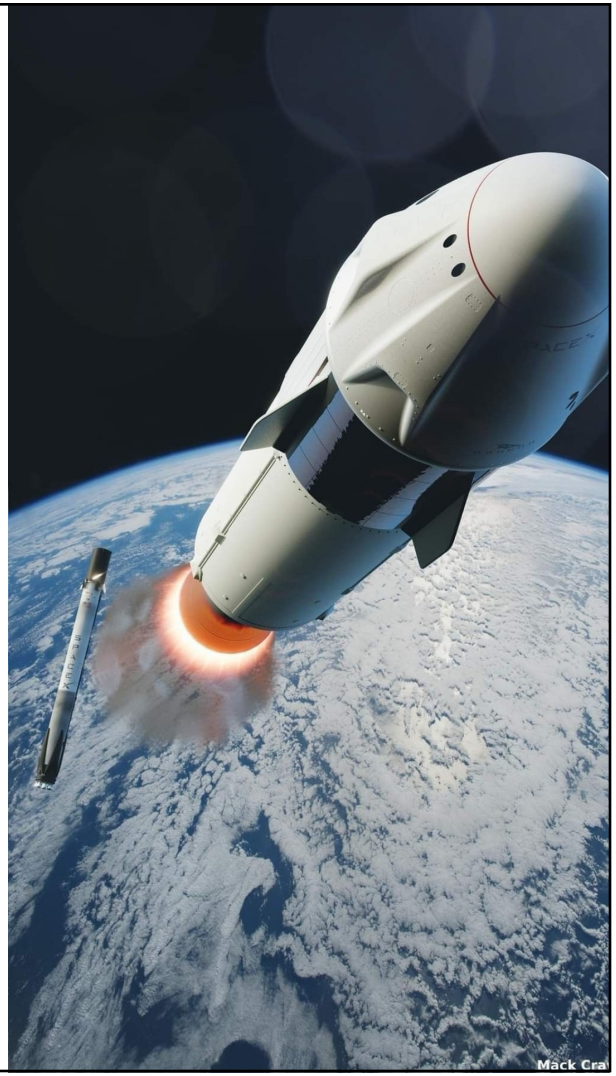


Decision Tree



KNN





Predictive Analytics with Machine Learning

- Results in confusion matrixes explained

The confusion matrixes for Logistic Regression, SVM, Decision Tree and KNN shows identical numbers of correct and incorrect predictions on the test set of 18 cases:

- 12 true positive (TP): correctly predicted successful landing
- 3 true negative (TN): correctly predicted failed landing
- 3 false positive (FP): incorrectly predicted failures as successes
- 0 false negative (FN): incorrectly predict success as failure

All 4 models have the same test accuracy rate of:

(TP+TN) / (TP+TN+FP+FN) = 83.83%

ML Model with Best Performance

	ML Method	Accuracy_on_test (%)	Accuracy_on_validation (%)	Difference (pp)
0	Logistic Regression	83.33	84.64	1.31
1	Support Vector Machine	83.33	84.82	1.49
2	Decision Tree	83.33	88.93	5.60
3	K Nearest Neighbour	83.33	84.82	1.49

When comparing machine learning models with identical test metrics (accuracy, precision, F1 score), the difference between validation and test accuracy, expressed in percentage points (pp), is a key indicator of model generalization and potential overfitting. The model with the smallest performance gap between the validation and test sets is the winner:

- Best performance: Logistic Regression (1.31 pp)
- Running up tied: SVM & KNN (1.49 pp)
- Worst performance: Decision Tree (5.60 pp)

ML Model with Best Performance (discussion)

- **Best Performance (Logistic Regression):** The smaller the gap between validation and test accuracy, the better the model's ability to generalize to new, unseen data, as it suggests the model's performance on the validation set is a better representation of its real-world performance. Logistic Regression has a gap of 1.3 percentage point, smaller than those for all other three models.
- **Greatest Overfitting Risk (Decision Tree):** A large gap (as seen with the Decision Tree's 5.6 percentage points difference) indicates that the model performed significantly better on the validation data than on the test data. This is a strong sign of **overfitting**, meaning the model has likely learned specific nuances or "noise" in the validation data that do not exist in the general population of data.

Conclusion

The objective of this project is to predict if the Falcon 9 first stage will land successfully. The procedures are conducted in Python, including:

1. Collect info from SpaceX API and Wikipedia and wrangle data
2. Explore and analyze data
   - Use SQL to derive info on launch sites, payload, successful and failed outcomes
   - Visualize data, which show the success rate has increased with the number of flights, particularly in LEO missions, and over time
3. Build interactive visualization maps and dashboard
   - Maps from Folium show that launch sites are close to highway and railroad for easy transportation, close to coastline and equator to utilize Earth rotation, but distant from cities to keep population and property safe
   - Results from Plotly Dashboard show KSC LC-39A were most successful; Version B5 were most successful; and Payload range 2K – 6K and 9K-10K have the highest success rate
4. Predict success rate by machine methods including SVM, Logistic Regression, Decision Tree and KNN and determine the model with the best performance
   - All four models has the identical test metrics (accuracy, precision, F1 score). it is impossible to determine the best performance based on the test scores alone.
   - Given the identical test accuracy, the winner goes to Logistic Regression because it has the smallest difference between validation and test accuracy.

Appendix

List of Python programs and their web address at Github

Module 1
- M1_1_jupyter-labs-spacex-data-collection-api.ipynb
- M1_2_jupyter-labs-webscraping.ipynb
- M1_3_labs-jupyter-spacex-Data wrangling-v2.ipynb

Module 2
- M2_1_jupyter-labs-eda-sql-edx-sqllite.ipynb
- M2_2_jupyter-labs_eda-data-visualization.ipynb

Module 3
- M3_1_lab_jupyter_folium_launch_site_location.ipynb
- M3_2_spacex-dash-app.py

Module 4
- M4_SpaceX_Machine_Learning_Prediction.ipynb

https://github.com/AnneShi/IBM-ML-Capstone


Thank you!