Applied Data Science Capstone Project

Aldo Ortiz

Space Y Data Scientist

July 5, 2022

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Space Y is interested in competing against Space X on a bid for future space race contracts.
- Understanding the potential outcome of the re-use of Falcon 9 first stage is critical for Space Y project decision making.
- A detailed analysis of the Space X launch performance data campaign has been performed which included:
  - Data Collection and Data wrangling.
  - Exploratory Data Analysis with SQL and Visual Interaction.
  - Interactive and real time Dashboard.
  - Machine Learning Pipeline.
- The re-use of the Falcon 9 first stage is a feasible alternative, and its cost is estimated in $ 62 MM.

INTRODUCTION

- Commercial Space age is here. There are several companies making a lot of progress in suborbital flights, satellites and reusable rockets.
- SpaceX stands out for key accomplishments:
  - Sending spacecraft to the International Space Station.
  - Starlink, a satellite internet constellation providing satellite Internet access.
  - Sending manned missions to Space.
- One reason SpaceX can do this is the rocket launches are relatively inexpensive because SpaceX can reuse the first stage.
- This project objective is to determine if SpaceX will successfully reuse the first stage and its cost.
- This will be achieved by creating a machine learning model and use public information to predict if SpaceX will reuse the first stage.

METHODOLOGY

- Data Collection and Data Wrangling
- Data Wrangling
- Perform Exploratory Data Analysis (EDA)
  - SQL and Visualization (Pandas and Matplotlib)
  - Determine what attributes are correlated with successful landings
- Interactive Visual Analytics and Build Interactive Dashboard
  - Interactive Map - Explore the map with key markers and discover any patterns.
  - Interactive Dashboard - Plotly Dash
- Machine Learning Prediction
  - Predict if the first stage of the Falcon 9 lands successfully

METHODOLOGY

Data Collection

- SpaceX launch data that was gathered from an API and from Web Scrapping.
  - SpaceX REST API. This API provided data about Falcon 9 launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Falcon 9 launch data Wiki pages web scrapping.
- Modified Payload Null Values for Mean Values

Data Wrangling

- Convert the landing outcome of the launch to Class data either 0 or 1 depending on the outcome. (0 bad outcome, 1 good outcome)

METHODOLOGY

EDA and Interactive Visual Analytics

- Determine the attributes associated with successful landings using SQL.
- Categorical values are converted using one hot encoding.
- Use Folium interactive map with launch site markers, identify patterns.
- Use Plotly Dash app and build a dashboard to explore Space X data on real time.

METHODOLOGY

Predictive Analysis

- Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.
- In Summary the steps include:
  - standardize data
  - train_test split
  - train model
  - perform Gridsearch
  - determine model with best accuracy using Score Method.
- Models tested: Logistic regression, Decision Tree Classifier, K-Nearest Neighbors, and Support Vector machines.

RESULTS

EDA with SQL

- Unique Launch Sites CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- Average payload mass carried by booster version F9 v1.1 (6,138 kg)
- Date when the first successful landing outcome in ground pad was achieved. (6-4-2010)
- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 (1022, 1026, 1021.2, 1031.2)
- Total number of successful and failure mission outcomes (101).
- Booster_versions which have carried the maximum payload mass:

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

RESULTS

EDA with SQL

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

RESULTS

EDA with SQL

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

RESULTS

EDA with SQL

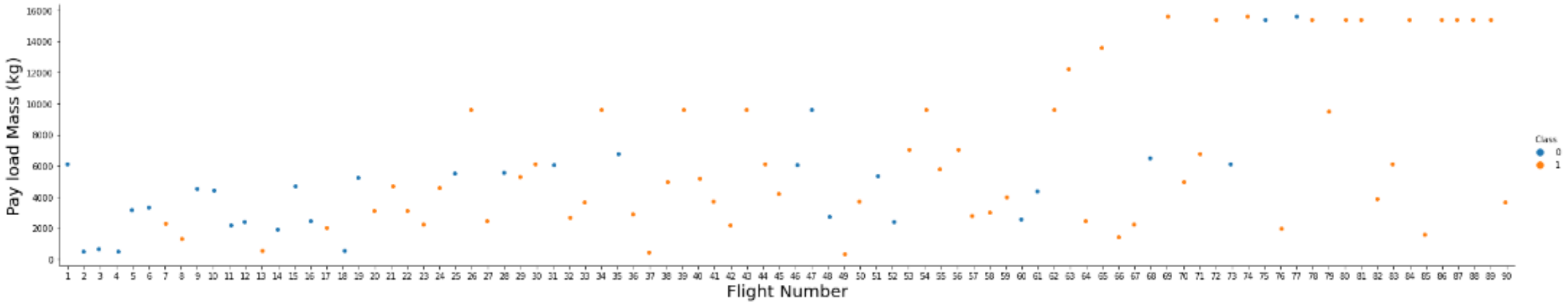
- Example of Friday Date Launches (Limit to 5)

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
4/18/2014	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
3/4/2016	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
4/8/2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)

RESULTS

EDA with Visualization

- Flight Number vs Payload Mass

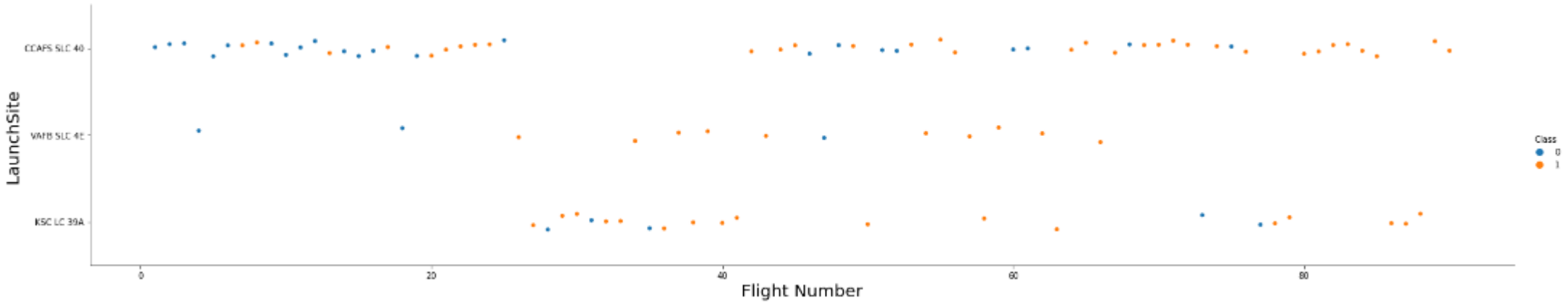


- As The flight number increases, the first stage is more likely to land successfully
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

RESULTS

EDA with Visualization

- Flight Number vs Launch Site

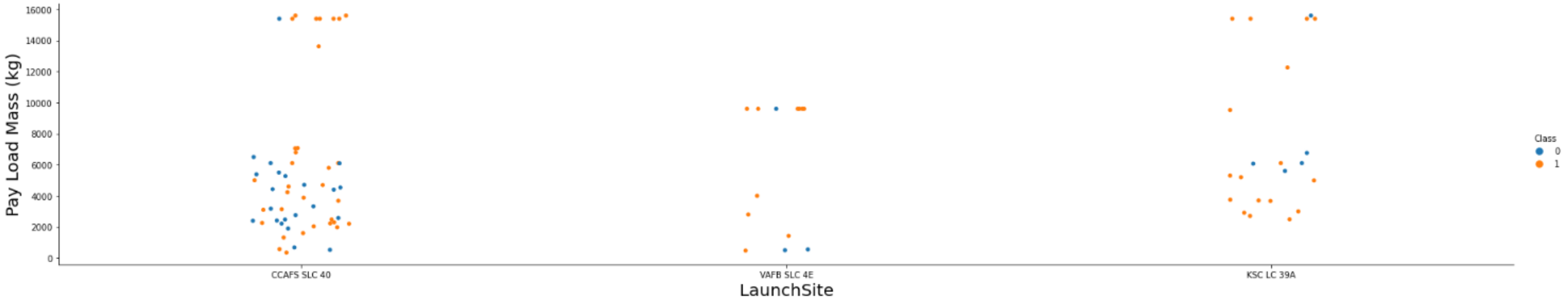


- SLC-40 is the launch site with the most launching attempts, while SLC-4E is the one with the fewest attempts (most of them failed)
- The initial flights from SLC-40 were successful while the last ones are failed attempts.

RESULTS

EDA with Visualization

- Launch Site vs Payload Mass

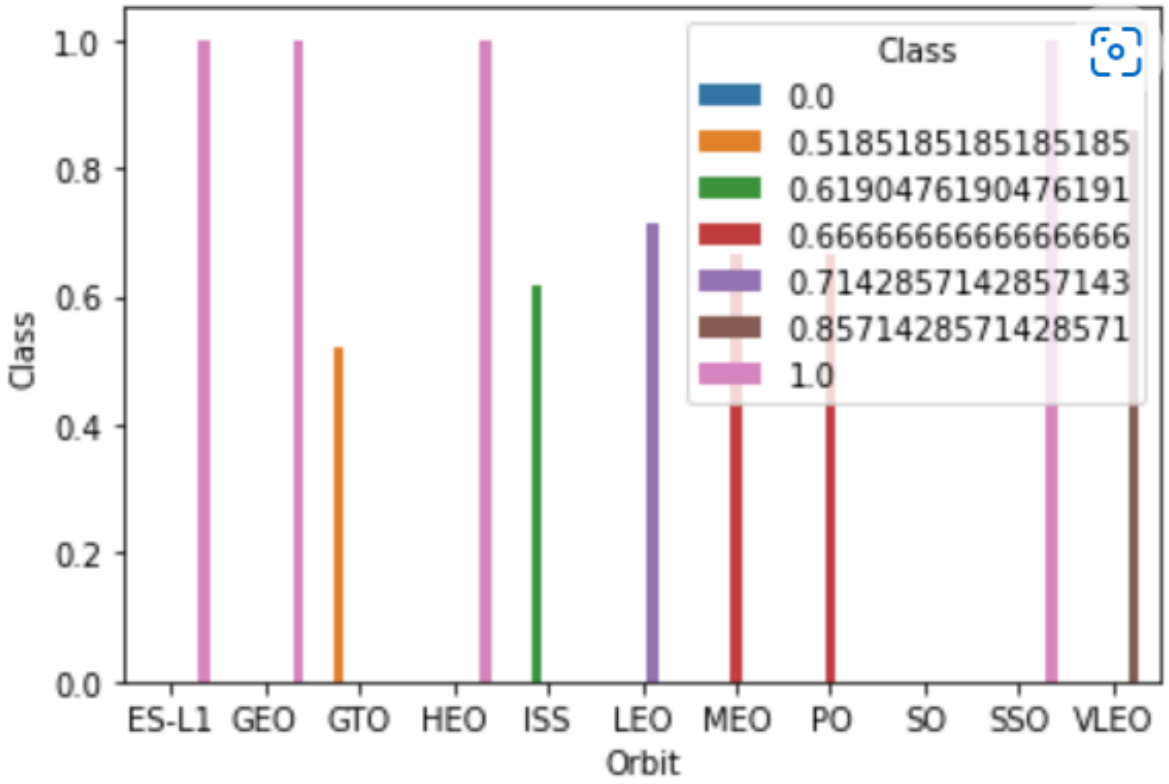


- LC-39A site has a wider range of payload tested when compared to SLC-40 and SLC-4E
- SLC-40 site has the most attempted flights with a payload less than 6000 kg.
- SLC-4E site has the least numbers of flights attempted and most of them failed.
- SLC-40 and LC-39A are the only sites testing payloads around 16000 kg.

RESULTS

EDA with Visualization

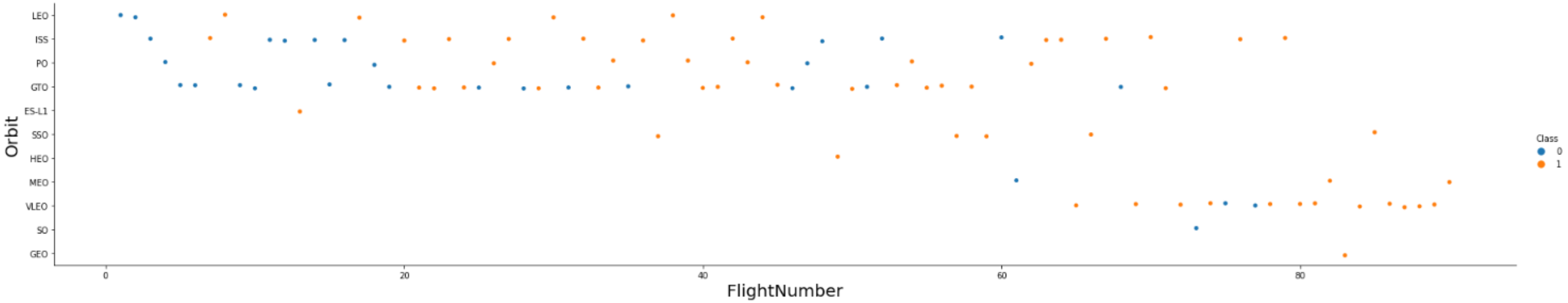
- Success Rate vs Orbit Type
  - 4 orbits display the highest success rate:
  - ES-L1, GEO, HEO, SSO



RESULTS

EDA with Visualization

- Flight Number vs Orbit

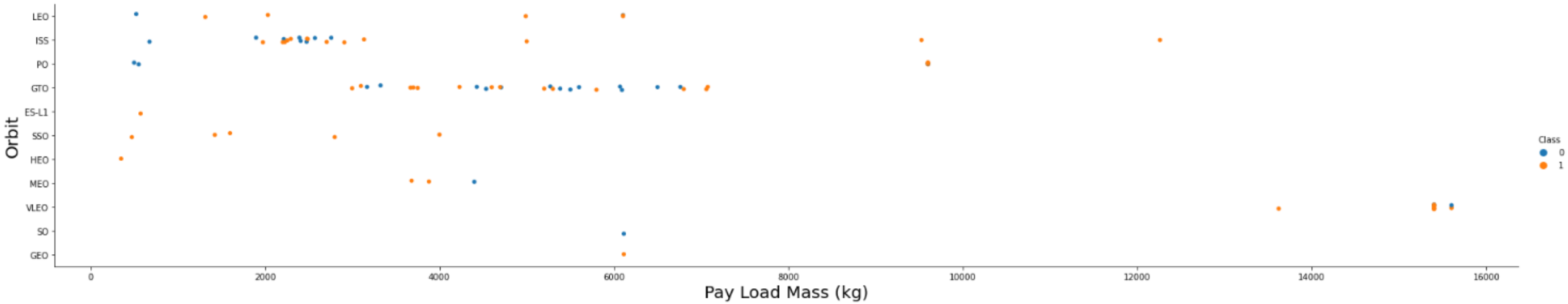


- LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit.

RESULTS

EDA with Visualization

- Payload vs Orbit

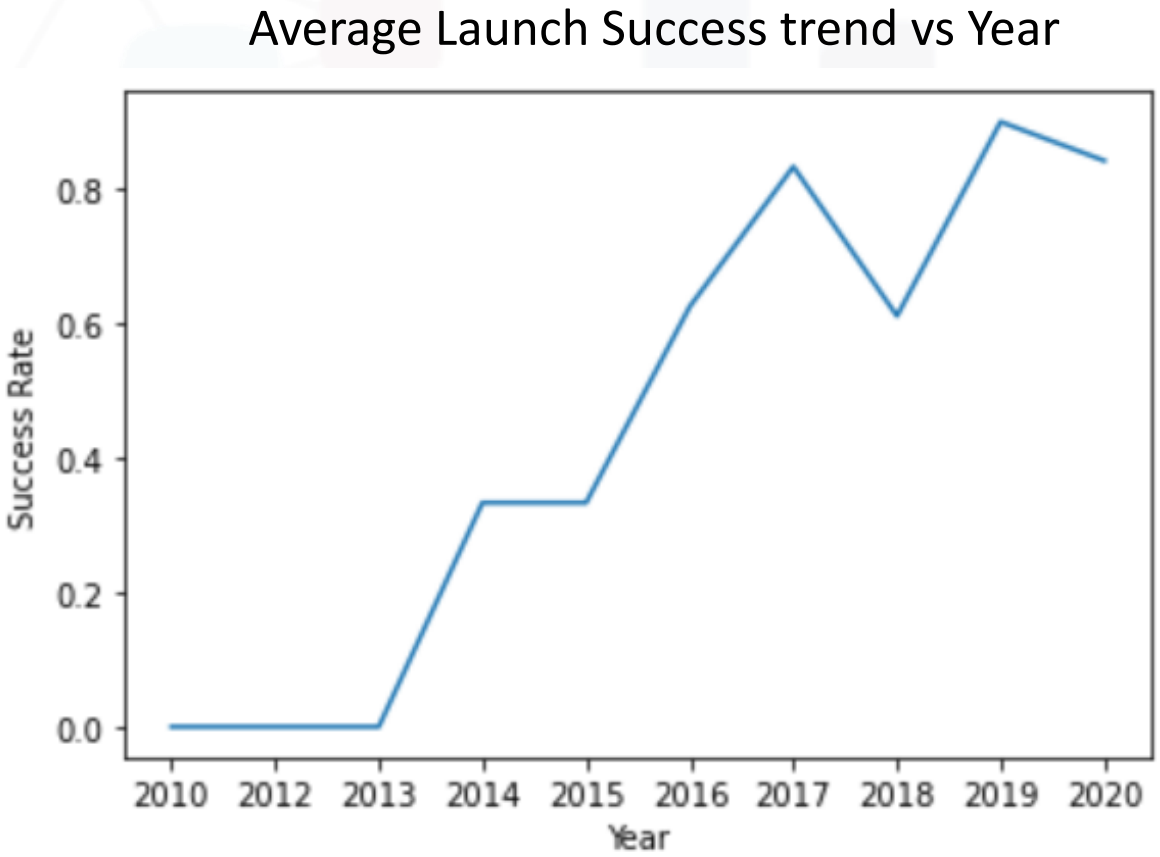


- With heavy payloads (over 8000 Kgs) the successful landing or positive landing rate are more for Polar, LEO and ISS.
- GTO orbit has a mixed bag of results with payloads between 3000 and 7000 Kgs.

RESULTS

- **EDA with Visualization**

- Increasing trend since 2013
- Sharp decrease in success rate between 2017 and 2018
- Highest success rates observed after 2018



RESULTS

- **EDA with Visualization**
  - Identified insights about how each important variable would affect the success rate.

FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
1	6104.959412	LEO	CCAFS SLC 40	1	FALSE	FALSE	FALSE	NaN	1	0	B0003
2	525	LEO	CCAFS SLC 40	1	FALSE	FALSE	FALSE	NaN	1	0	B0005
3	677	ISS	CCAFS SLC 40	1	FALSE	FALSE	FALSE	NaN	1	0	B0007
4	500	PO	VAFB SLC 4E	1	FALSE	FALSE	FALSE	NaN	1	0	B1003
5	3170	GTO	CCAFS SLC 40	1	FALSE	FALSE	FALSE	NaN	1	0	B1004

RESULTS

- **EDA with Visualization**
  - Used the function get_dummies and features dataframe to apply OneHotEncoder to the Orbits, LaunchSite, LandingPad, and Serial.
  - Assigned the value to the variable features_one_hot.

FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Serial_B1050	Serial_B1051	Serial_B1054	Serial_B1056	Serial_B1058	Serial_B1059	Serial_B1060	Serial_B1062
1	6104.959412	1	FALSE	FALSE	FALSE	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	525	1	FALSE	FALSE	FALSE	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	677	1	FALSE	FALSE	FALSE	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	500	1	FALSE	FALSE	FALSE	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	3170	1	FALSE	FALSE	FALSE	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
86	15400	2	TRUE	TRUE	TRUE	5	2	0	0	...	0	0	0	0	0	0	0	0	1	0
87	15400	3	TRUE	TRUE	TRUE	5	2	0	0	...	0	0	0	0	0	0	1	0	0	0
88	15400	6	TRUE	TRUE	TRUE	5	5	0	0	...	0	0	0	1	0	0	0	0	0	0
89	15400	3	TRUE	TRUE	TRUE	5	2	0	0	...	0	0	0	0	0	0	0	0	1	0
90	3681	1	TRUE	FALSE	TRUE	5	0	0	0	...	0	0	0	0	0	0	0	0	0	1

  - Since the DF contains numbers, the entire dataframe is set to float64 data type.

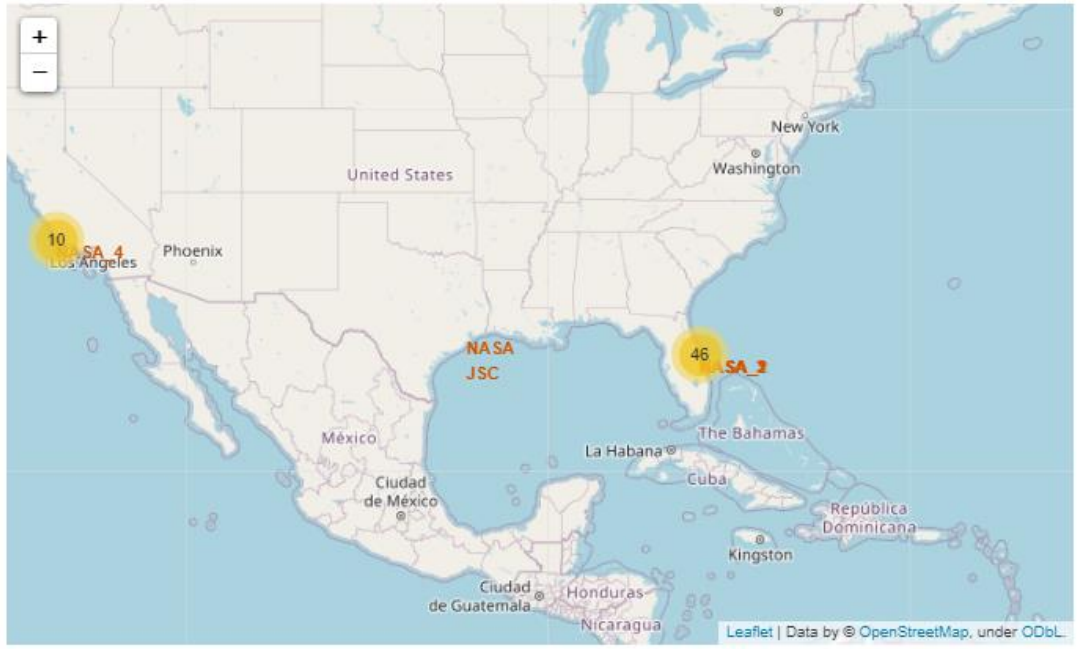
RESULTS

- **Interactive Map with Folium**
- **Launch Sites**
  - Launch Sites are not in proximity to the Equator Line but closer to the Tropic if Cancer
  - Launch Sites are in very close proximity to the coast.
  - Launch Sites located in Coast states California and Florida.

RESULTS

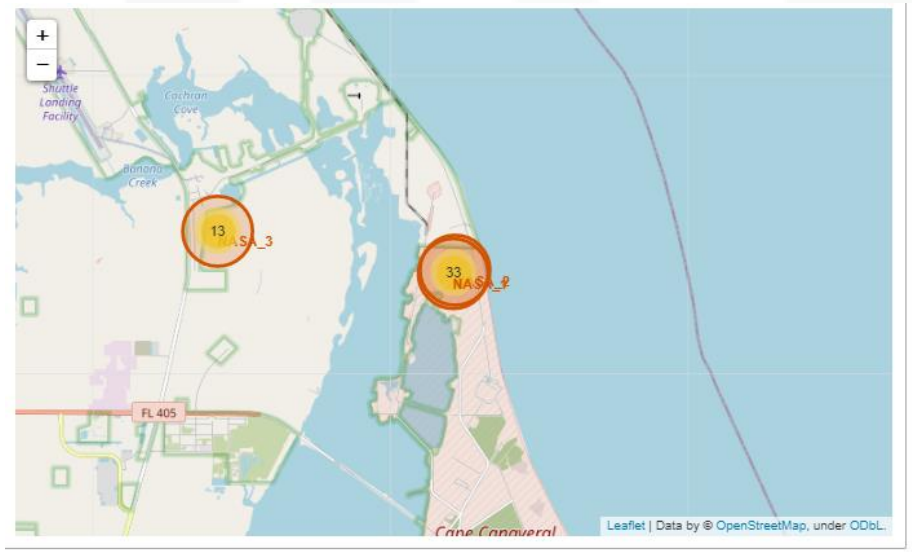
- **Interactive Map with Folium**

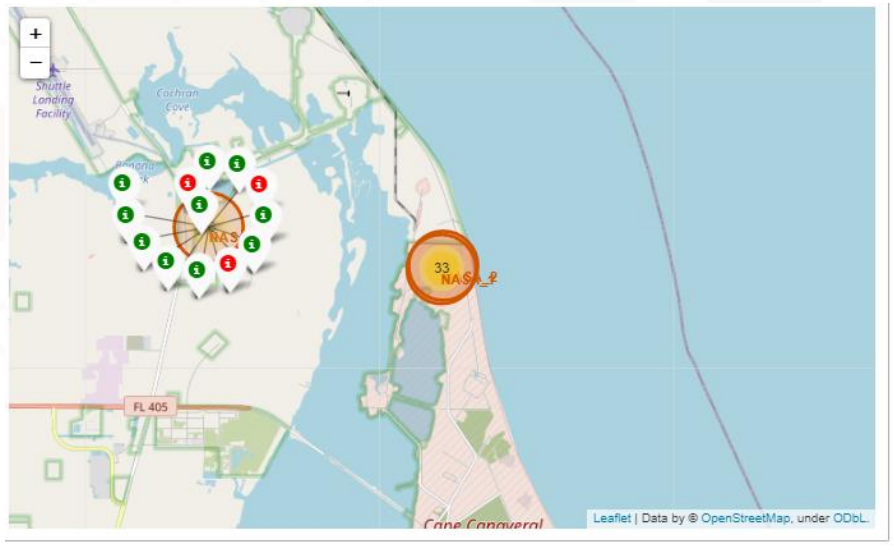
As shown in the next slides, from the color-labeled markers in marker clusters, it is easy to identify KSC LC 39A launch site has relatively high success rates.



RESULTS

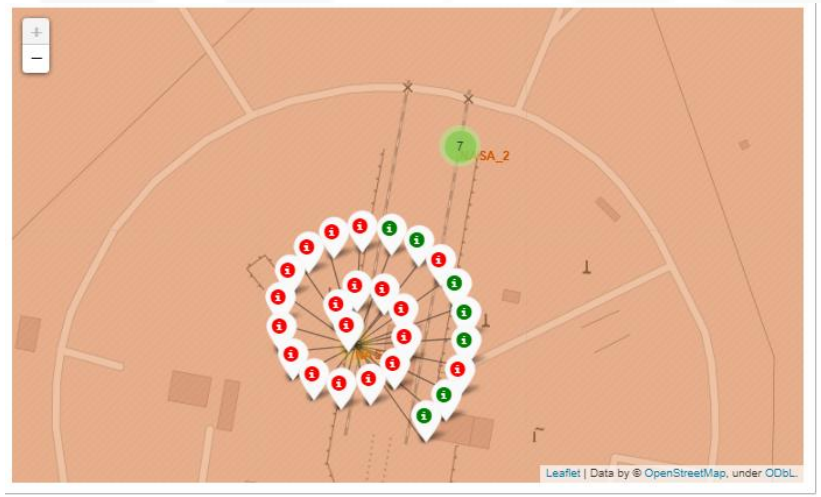
- Interactive Map with Folium, KSC LC-39A





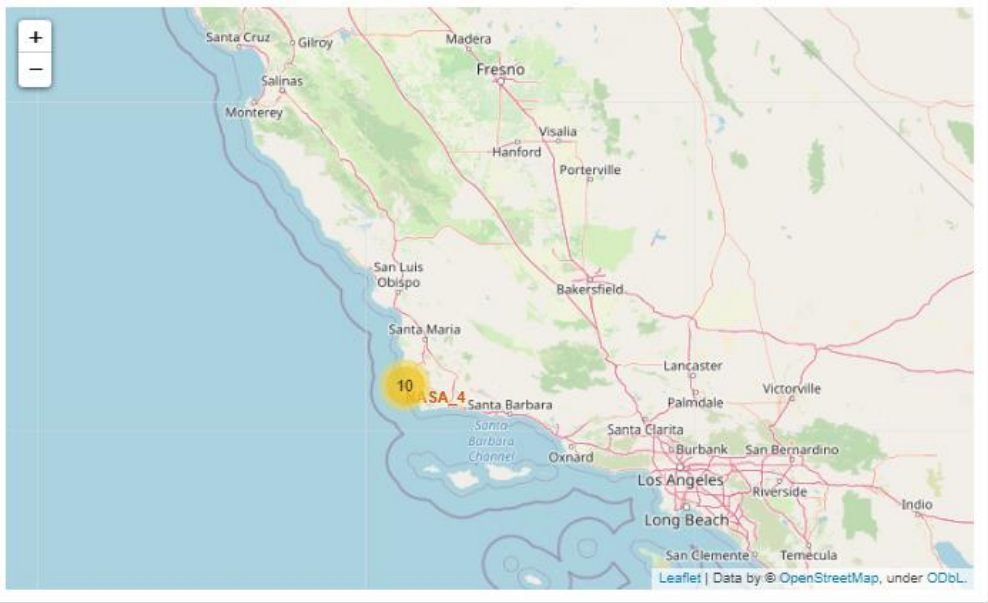
RESULTS

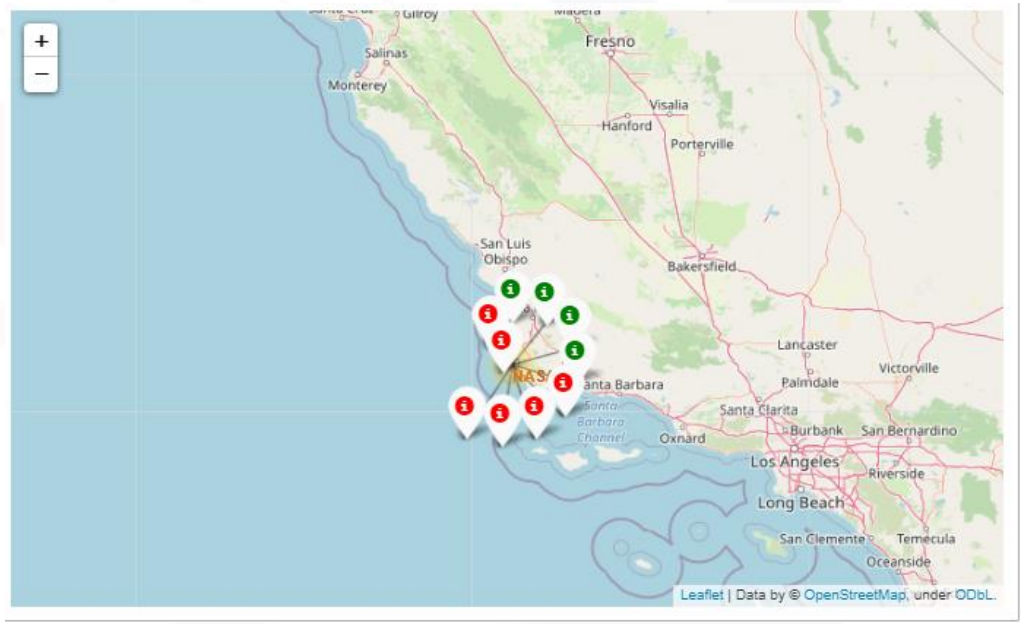
- Interactive Map with Folium CCAFS LC-40, CCAFS SLC-40



RESULTS

- Interactive Map with Folium VAFB SLC-4E



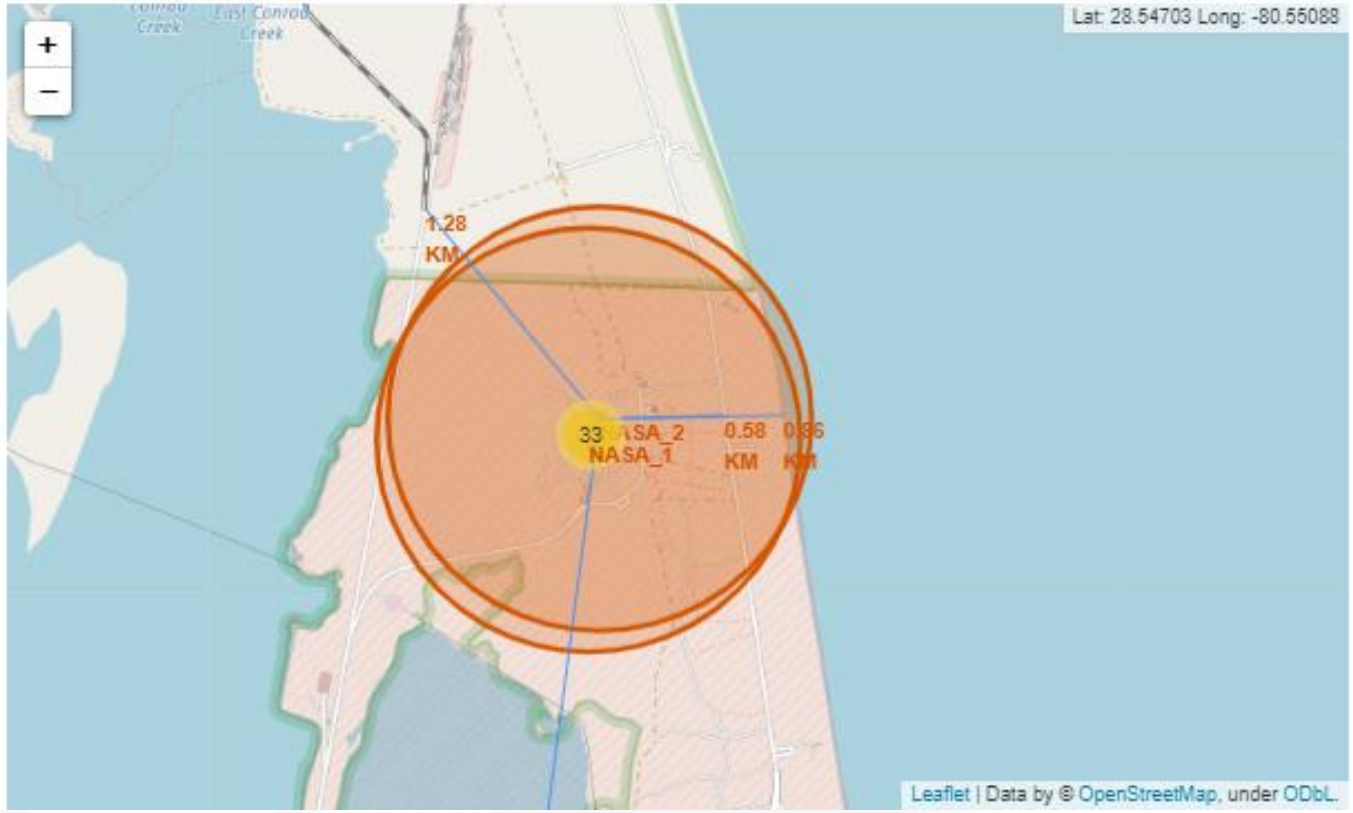


RESULTS

- **Interactive Map with Folium**

The following results are based on Launch Site CCAFS SLC-40.

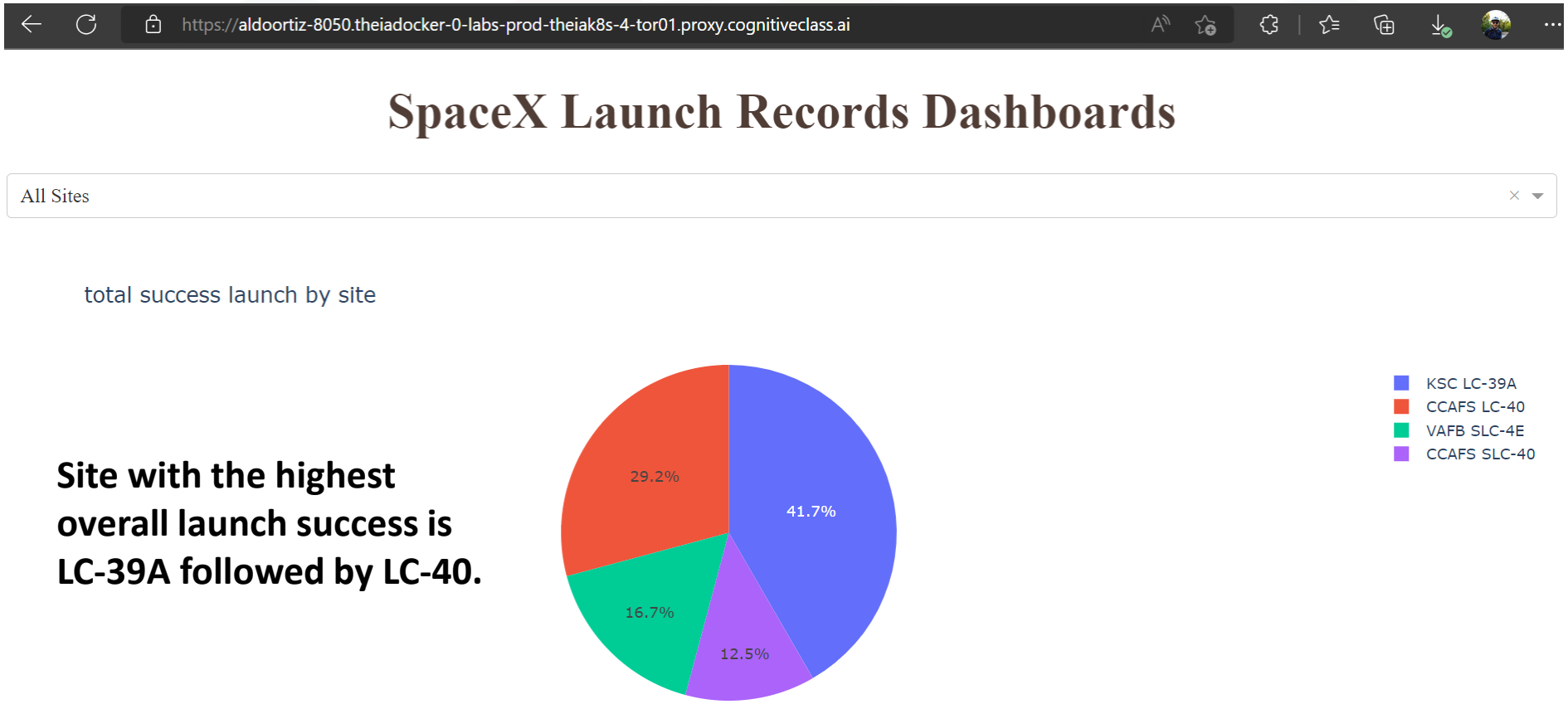
- Launch site is in close proximity to railways: 1.28 kms away
- Launch sites in close proximity to highways: 0.58 Kms away
- Launch sites in close proximity to coastline: 0.86 Km away
- Launch sites keep certain distance away from cities: 50+ Kms away is Melbourne.



RESULTS

- **Plotly Dash Interactive Dashboard**
  - Generated a dashboard that allowed to dissect the data in an interactive and real time way.
  - Dashboard features included drop down list, pie and scatter point charts, slider.
  - Allowed the analysis of the Space X Data by launch location, % of launch success, payload vs outcome for different booster versions.
  - Interaction with the drop-down list, allow visualization of % of success for each launch site and payload vs class in pie chart and scatter plot.

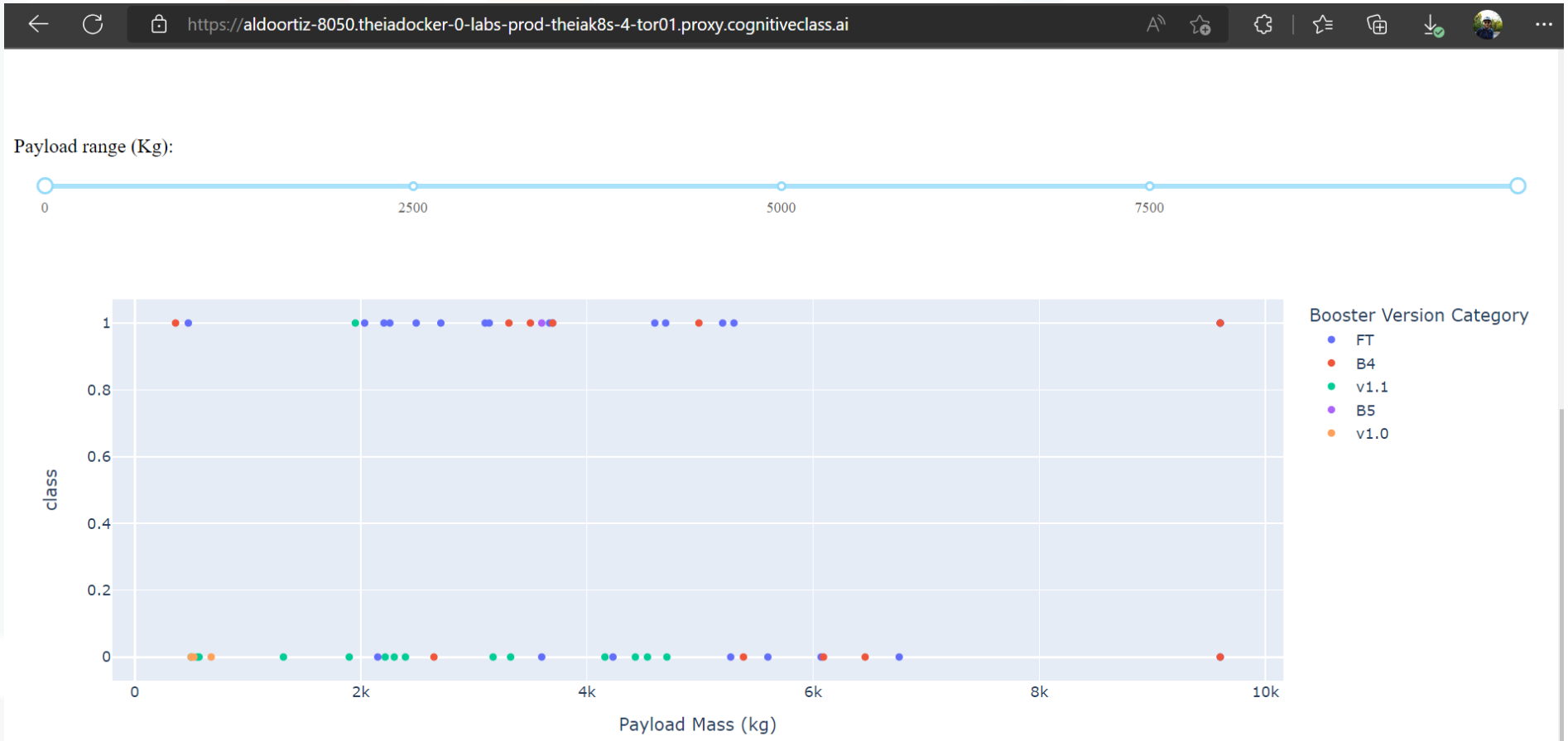
RESULTS



Site with the highest overall launch success is LC-39A followed by LC-40.

RESULTS

- **FT Booster version has the highest payload range with a positive outcome.**
- **Tested payload ranges over 5.5k Kgs were not successful except with B4 booster.**



RESULTS

- **Predictive Analysis**

  The results from the models that were tested were similar and practically no difference observed between them in terms of accuracy.

Model	Tuned HyperP accuracy	Score (accuracy on test data)
Logistic Regression	0.8472	0.83
Support Vector Machines	0.8472	0.83
Decision Tree Classifier	0.8888	0.88
K Neighbors Classifier	0.8472	0.83

CONCLUSION

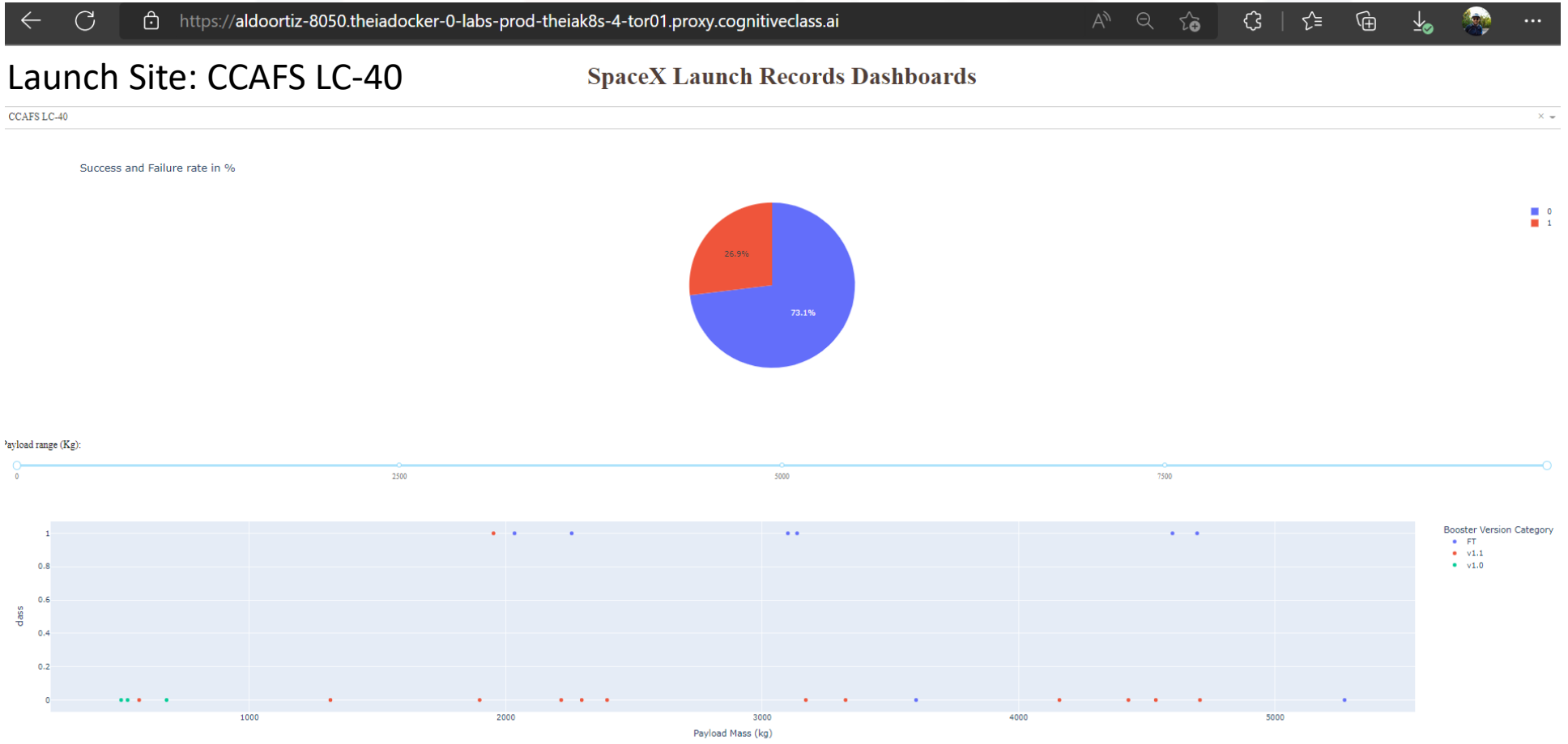
- Space X Data collected from API and web scrapping has been prepared to analyze the actual outcome of reusing the first stage and predict potential outcome in the future.
- The overall data analysis performed on Space X data included exploratory data analysis with SQL, visualization, Plotly dash dashboard and test of classifier models through a ML pipeline.
- Launch site LC-39A has the highest success rates with payloads up to 5.5k Kgs. The orbits with the highest success rates are ES-L1, GEO, HEO, SSO.
- Based on the data analysis performed, there is a high percentage chance that the first stage of the Falcon 9 lands successfully. This translates into a significant reduction of capital expenses ($62 MM) making it an attractive opportunity for Space Y.

APPENDIX

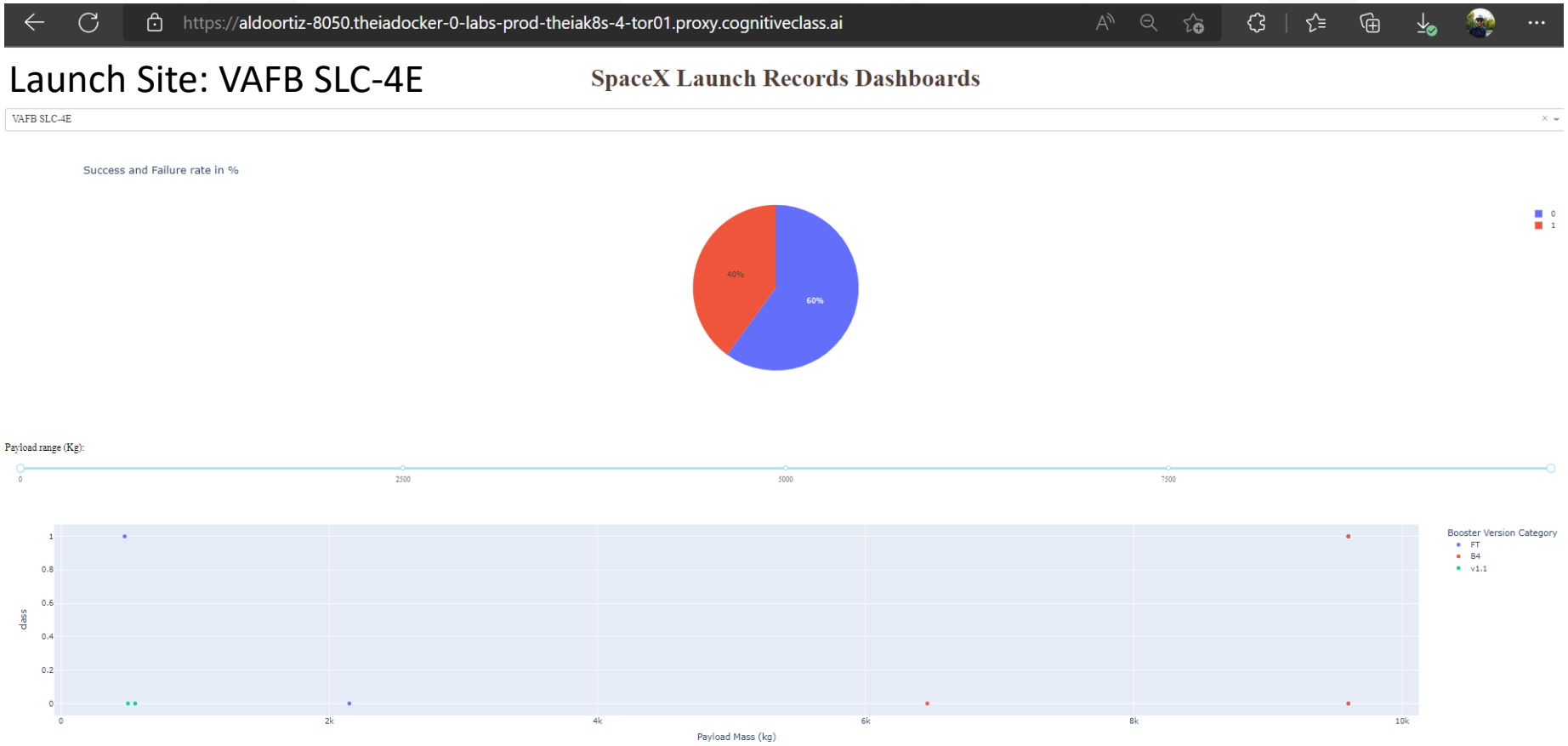
Dashboard selections for Each Launch Site

- Pie Chart depicting % of success rate
- Scatter plot of outcome vs payload.
- Slider displaying the Payload Range.

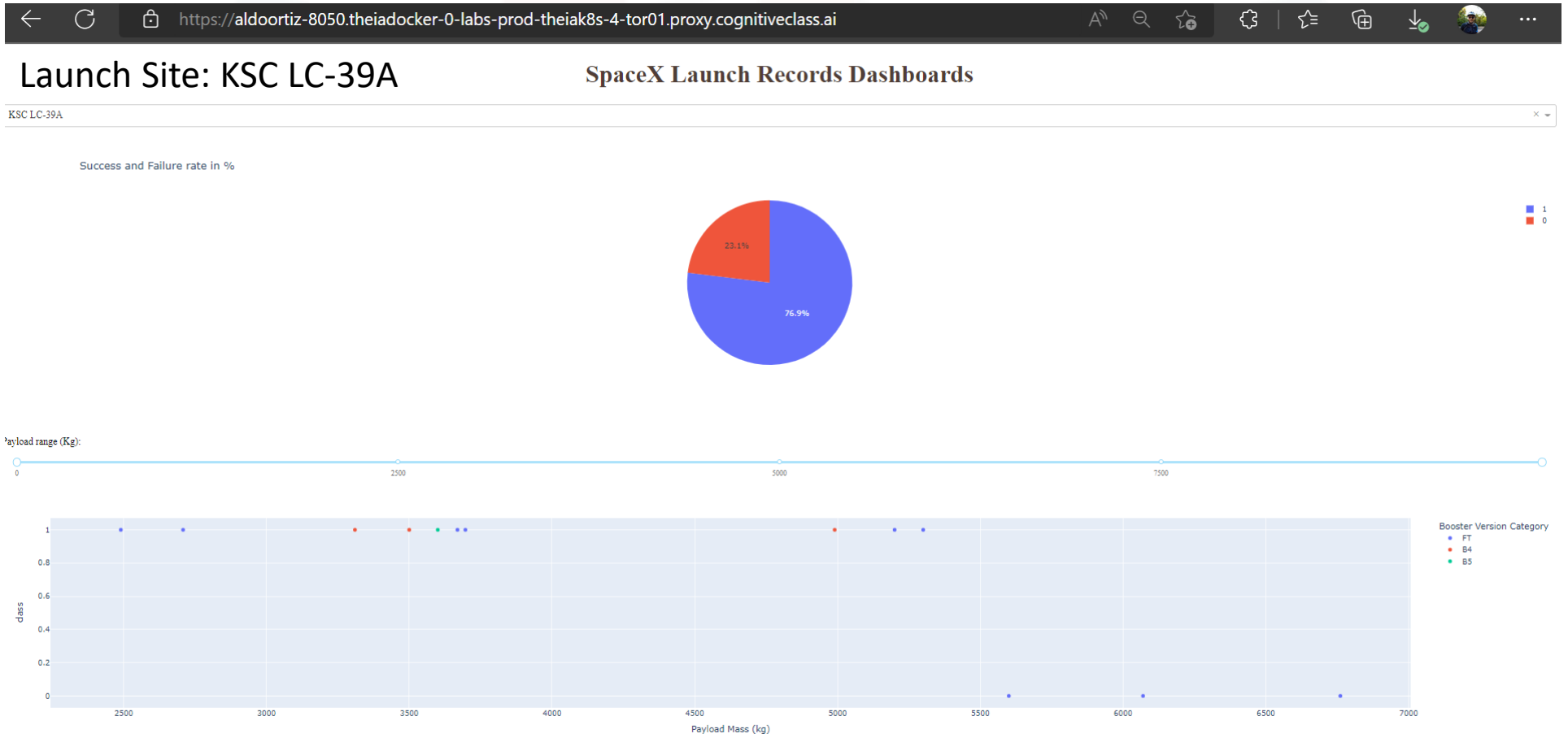
RESULTS


https://aldoortiz-8050.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai
Launch Site: CCAFS LC-40
SpaceX Launch Records Dashboards
CCAFS LC-40
Success and Failure rate in %
26.9%
73.1%
0
1
Payload range (Kg):
0
2500
5000
7500
class
Payload Mass (kg)
Booster Version Category
FT
v1.1
v1.0

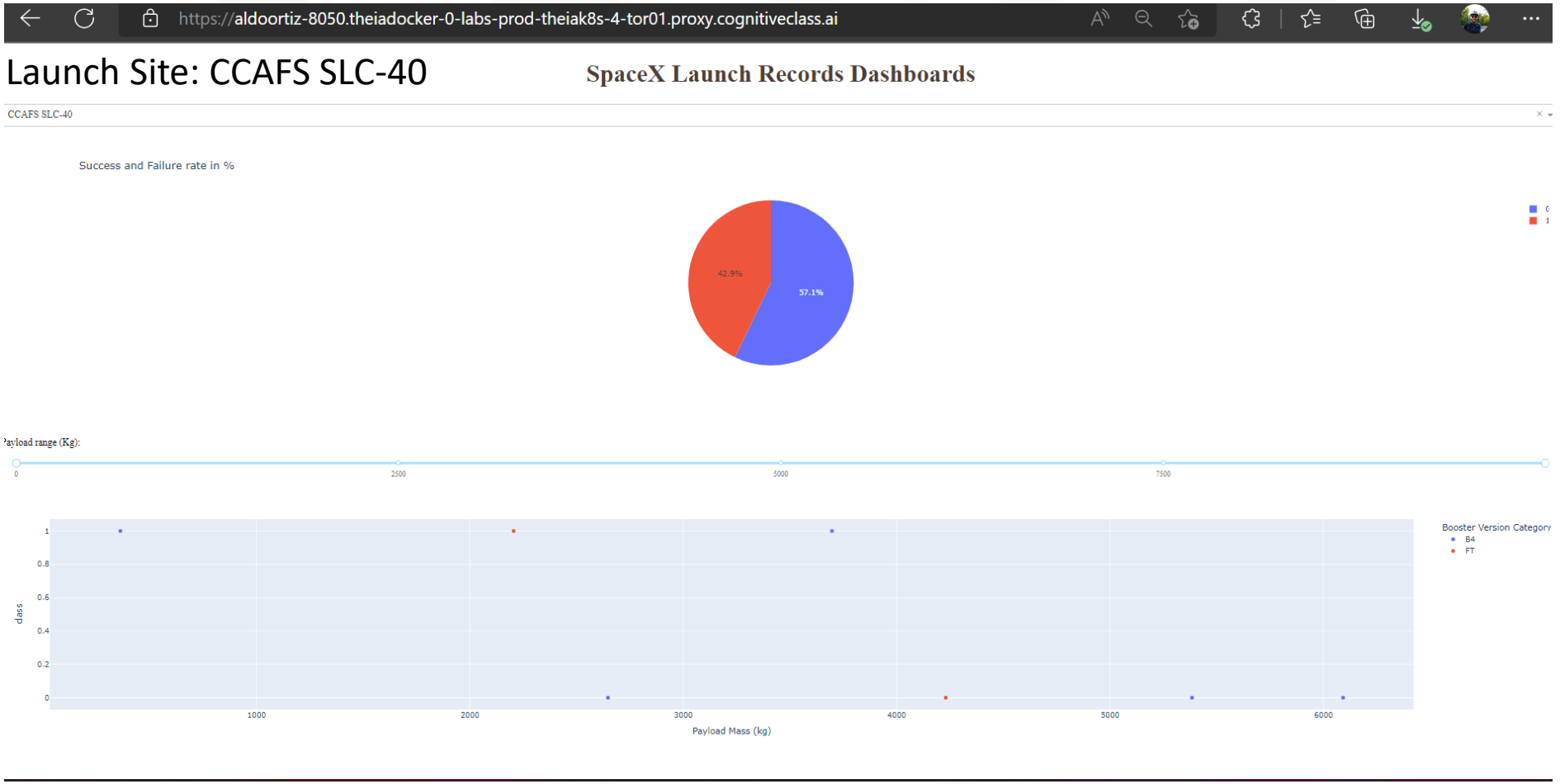
RESULTS


https://aldoortiz-8050.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai
Launch Site: VAFB SLC-4E
SpaceX Launch Records Dashboards
VAFB SLC-4E
Success and Failure rate in %
40%
60%
0
1
Payload range (Kg):
0
2500
5000
7500
Booster Version Category
FT
B4
v1.1
class
Payload Mass (kg)
0
2k
4k
6k
8k
10k

RESULTS


https://aldoortiz-8050.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai
Launch Site: KSC LC-39A
SpaceX Launch Records Dashboards
KSC LC-39A
Success and Failure rate in %
23.1%
76.9%
1
0
Payload range (Kg):
0
2500
5000
7500
class
Payload Mass (kg)
Booster Version Category
FT
B4
B5

RESULTS


https://aldoortiz-8050.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai
Launch Site: CCAFS SLC-40
SpaceX Launch Records Dashboards
CCAFS SLC-40
Success and Failure rate in %
42.9%
57.1%
Payload range (Kg):
0
2500
5000
7500
class
Payload Mass (kg)
Booster Version Category
B4
FT

RESULTS

[ortizaldo2569/capstone (github.com)]