Winning Space Race with Data Science

Kenneth V Miu

March 4th, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
  - Data Wrangling -- Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
  - Execute SQL queries to understand the Spacex DataSet
  - Exploratory Data Analysis and Feature Engineering
  - **Launch Sites Locations Analysis with Folium**
  - **Build a Plotly Dash Application**
  - **Machine Learning Prediction**

Executive Summary

- Summary of all results
  - We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
    - Number of launched success have improved over the years
    - Booster Version FT is the most successful
  - Decision Tree model had the highest accuracy at 89%

Introduction

- Project background and context
  - Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond. Falcon 9 is the world's first orbital class reusable rocket. Reusability allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access.
- Problems you want to find answers
  - Project goal is to predict if the Falcon 9 first stage will land successfully. SpaceX 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 SpaceX can reuse the first stage. 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 SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe was collected from SpaceX API and web scraping from Wiklpedia and SpaceX
- Perform data wrangling
  - Data was summarized and analyzed through Python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different supervised machine learning techniques were used to predict the outcomes.

Data Collection

- SpaceX API Data Columns:
  - https://github.com/durham430/Capstone-Project/blob/main/spacex-data-collection-api.ipynb

- Web Scraping on Wikipedia
  - https://github.com/durham430/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Perform Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurence of mission outcome per orbit type
  - Create a landing outcome label from Outcome column
- https://github.com/durham430/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
  - Exploratory Data Analysis
  - Preparing Data Feature Engineering
- To vsualize the relationship between Flight Number and Launch Site =>
  - sns.catplot(y="LaunchSite", x="FlightNumber", data= df, hue="Class", aspect = 5)
- To visualize the relationship between Payload and Launch Site =>
  - sns.catplot(x='PayloadMass', y='LaunchSite', data=df, hue = "Class", aspect=4)
- To visualize the relationship between success rate of each orbit type =>
  - sns.barplot(x='Orbit', y='Class', data=relation, hue='Orbit')
- To visualize the relationship between FlightNumber and Orbit type =>
  - sns.catplot(x='FlightNumber', y='Orbit', data=df, hue='Class')

EDA with Data Visualization

- To visualize the relationship between Payload and Orbit type =>
  - sns.catplot(x='PayloadMass', y='Orbit', data=df, hue='Class', aspect=4)
- To visualize the launch success yearly trend =>
  - sns.lineplot(data=df_copy, x='Extracted_year', y='Class')

- https://github.com/durham430/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(1).ipynb

EDA with SQL

- Objectives are:
  - Understand the Spacex DataSet
  - Load the dataset into the corresponding table in a Db2 database
  - Execute SQL queries to answer assignment questions
- Display the names of the unique launch sites in the space mission
  - %sql select distinct Launch_Site from SPACEXTBL
- Display 5 records where launch sites begin with the string 'CCA'
  - %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
- Display the total payload mass carried by boosters launched by NASA (CRS)
  - %sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
- Display average payload mass carried by booster version F9 v1.1
  - %sql select avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'

EDA with SQL

- **List the date when the first successful landing outcome in ground pad was acheived.**
  - %sql select min(DATE) from SPACEXTBL WHERE landing__outcome = 'Success (ground pad)'
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - %sql select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)'\and payload_mass__kg_ between 4000 and 6000
- List the total number of successful and failure mission outcome
  - %sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - %sql select booster_version, payload_mass__kg_ from SPACEXTBL\ where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - %sql select booster_version, launch_site from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015

EDA with SQL

- 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
  - %sql select count(landing__outcome), landing__outcome from SPACEXTBL \ where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome\ order by count(landing__outcome) desc
- https://github.com/durham430/Capstone-Project/blob/main/EDA%20with%20sql-coursera.ipynb

Build an Interactive Map with Folium

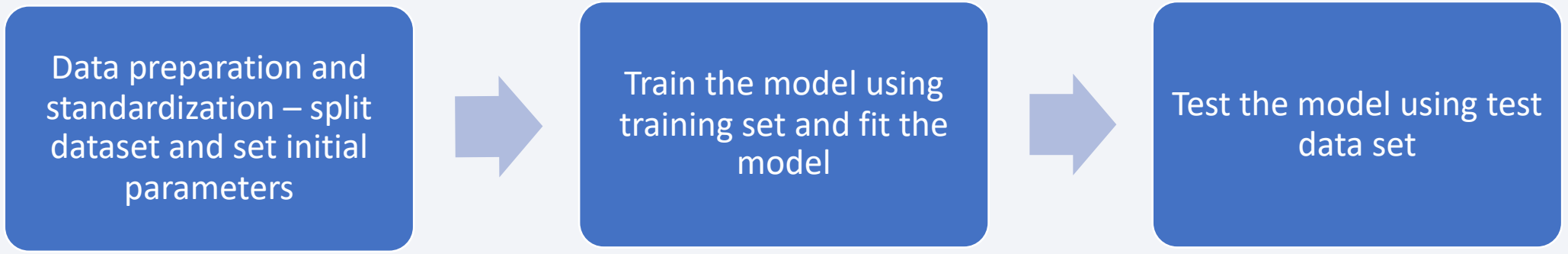
- We will performing more interactive visual analytics using Folium
- To determine the success rate dependence on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories
- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
- https://github.com/durham430/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- To show the success rate relationship between launch sites, Booster Version and Payload, we will build an interactive Dashboard that includes a pie chart and scatter plot
- Pie Chart to show the breakdown of success rate of All sites, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40
- Scatter Plot is to show interactively the success or failure by sites, Booster Version Category and payload between 0 and 10000kg using slideruler.
- https://github.com/durham430/Capstone-Project/blob/main/spacex_dash_app.py

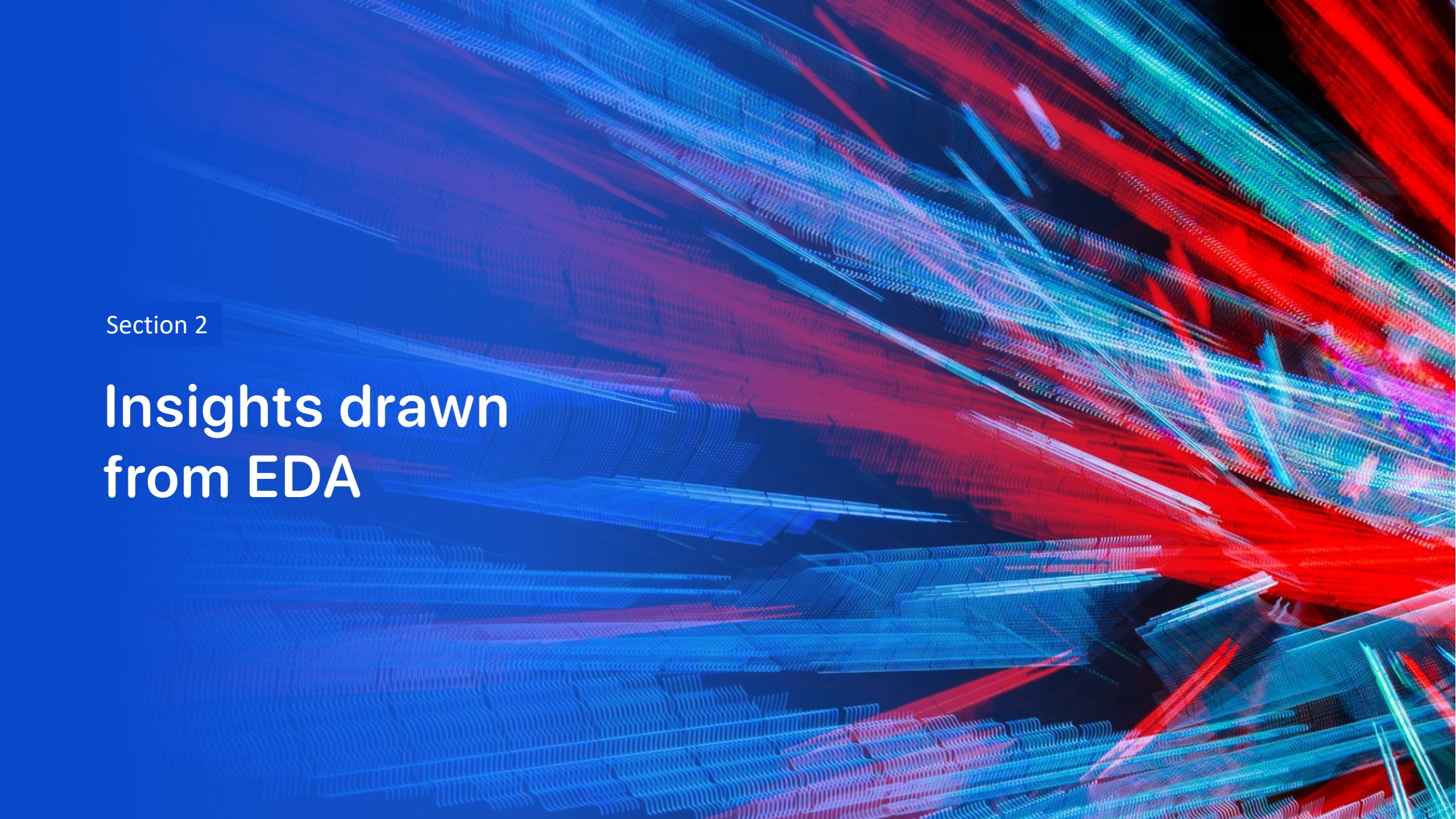
Predictive Analysis (Classification)

- We evaluated four different supervised classification model – logistic regression, support vector machine, decision tree and k nearest neighbors – to determine their performances.
- https://github.com/durham430/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb



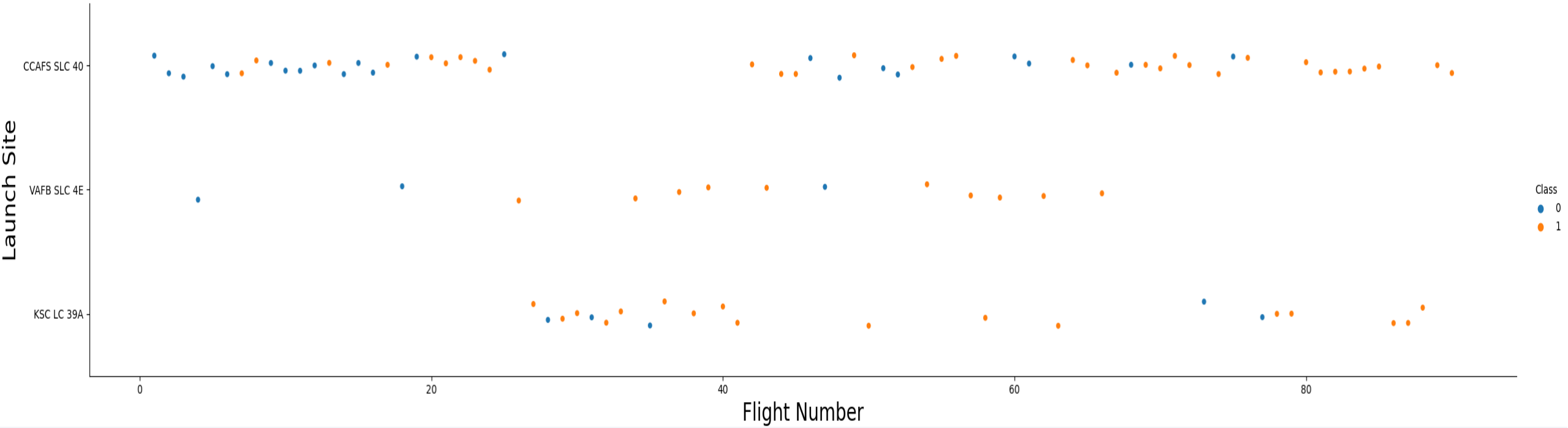
Results

- We see that 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.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- you can observe that the sucess rate since 2013 kept increasing till 2020
- Model accuracy:Logistic Regression Accuracy : 0.8464, SVM Accuracy : 0.8482, Decision Tree Accuracy : 0.889, KNN Accuracy : 0.8482


Section 2
Insights drawn from EDA

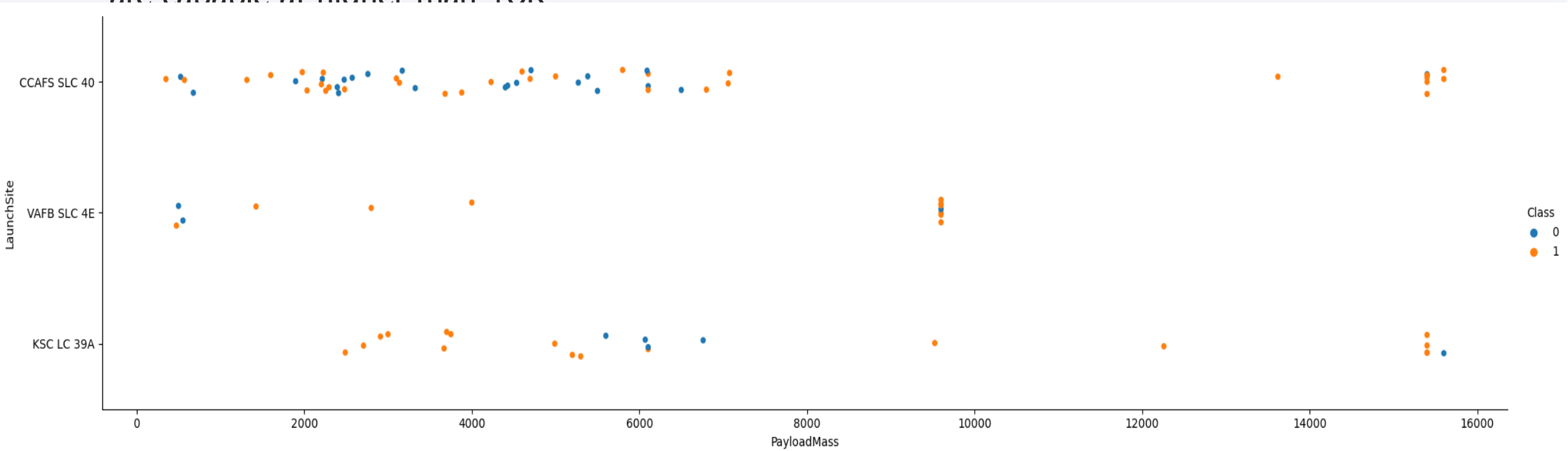
Flight Number vs. Launch Site

- scatter plot of Flight Number vs. Launch Site
- It can be concluded that recent flights success rate has improved for all launch sites



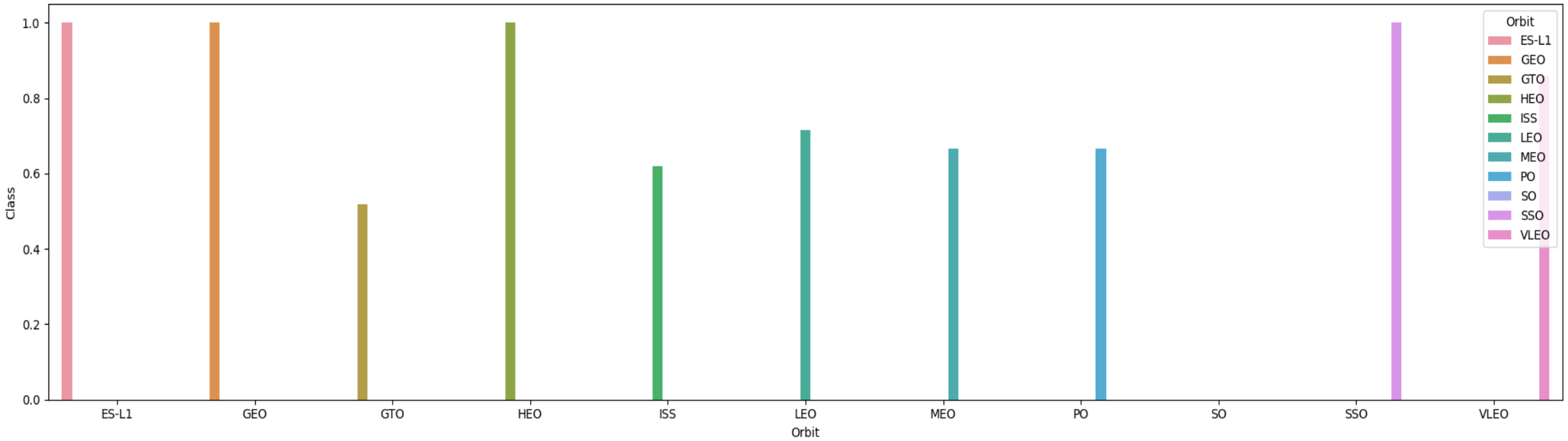
Payload vs. Launch Site

- scatter plot of Payload vs. Launch Site
- Higher than 9K has very high success rate. It seems only CCAFS SLC-40 and KSC LC 39A are capable at higher than 10K



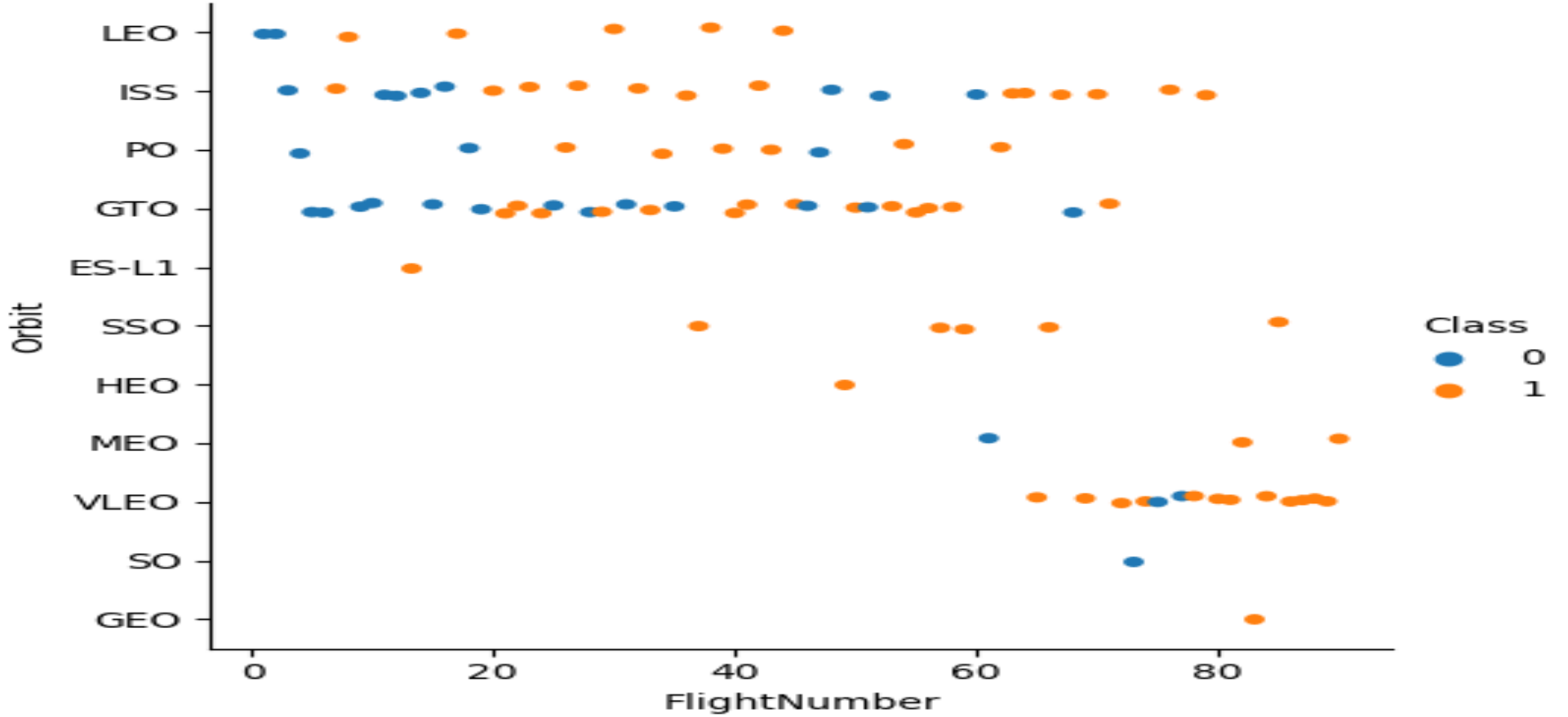
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO, and SSO has 100% success rate, others are better than 50%



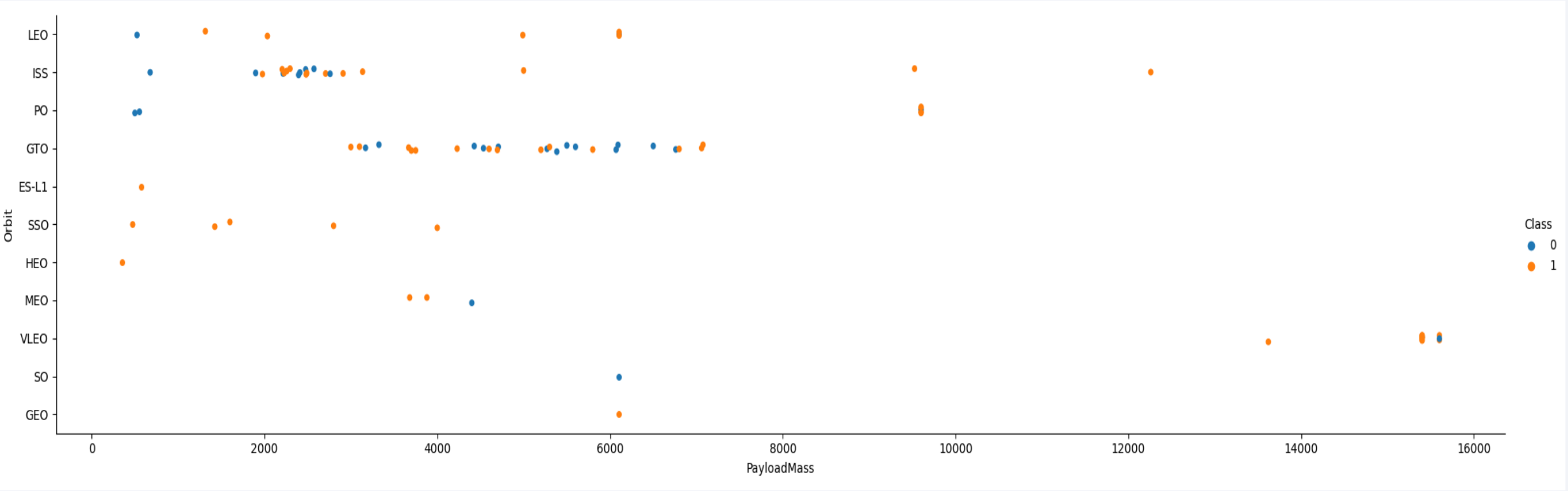
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type
- Success rate for all orbit has improved overtime, with VLEO being the most popular recently



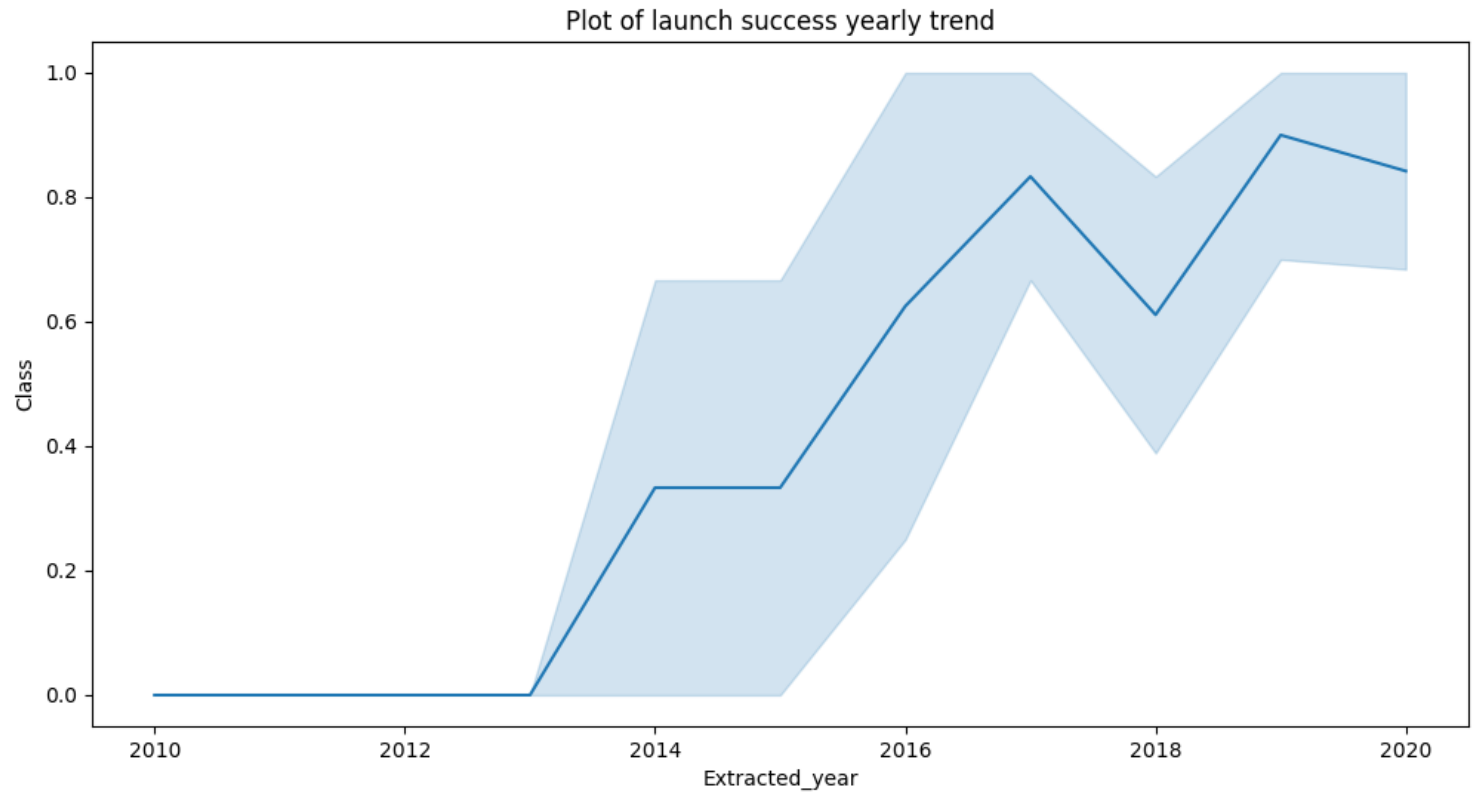
Payload vs. Orbit Type

- a scatter point of payload vs. orbit type
- GTO is often used between 3K and 7K, ISS is used with the widest range and VLEO is used for higher payload (>13k)



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Success rate has steady increased over the years, almost 90% at 2020



All Launch Site Names

- names of the unique launch sites
- %sql select distinct Launch_Site from SPACEXTBL

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39
AVAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`, or Cap Canaveral
- %sql select * from SPACEXTBL 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

- Calculate the total payload carried by boosters from NASA
- %sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- %sql select avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- %sql select min(DATE) from SPACEXTBL WHERE landing__outcome = 'Success (ground pad)'

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- %sql select booster_version from SPACEXTBL 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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- %sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- %sql select booster_version, payload_mass__kg_ from SPACEXTBL\where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- %sql select booster_version, launch_site from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

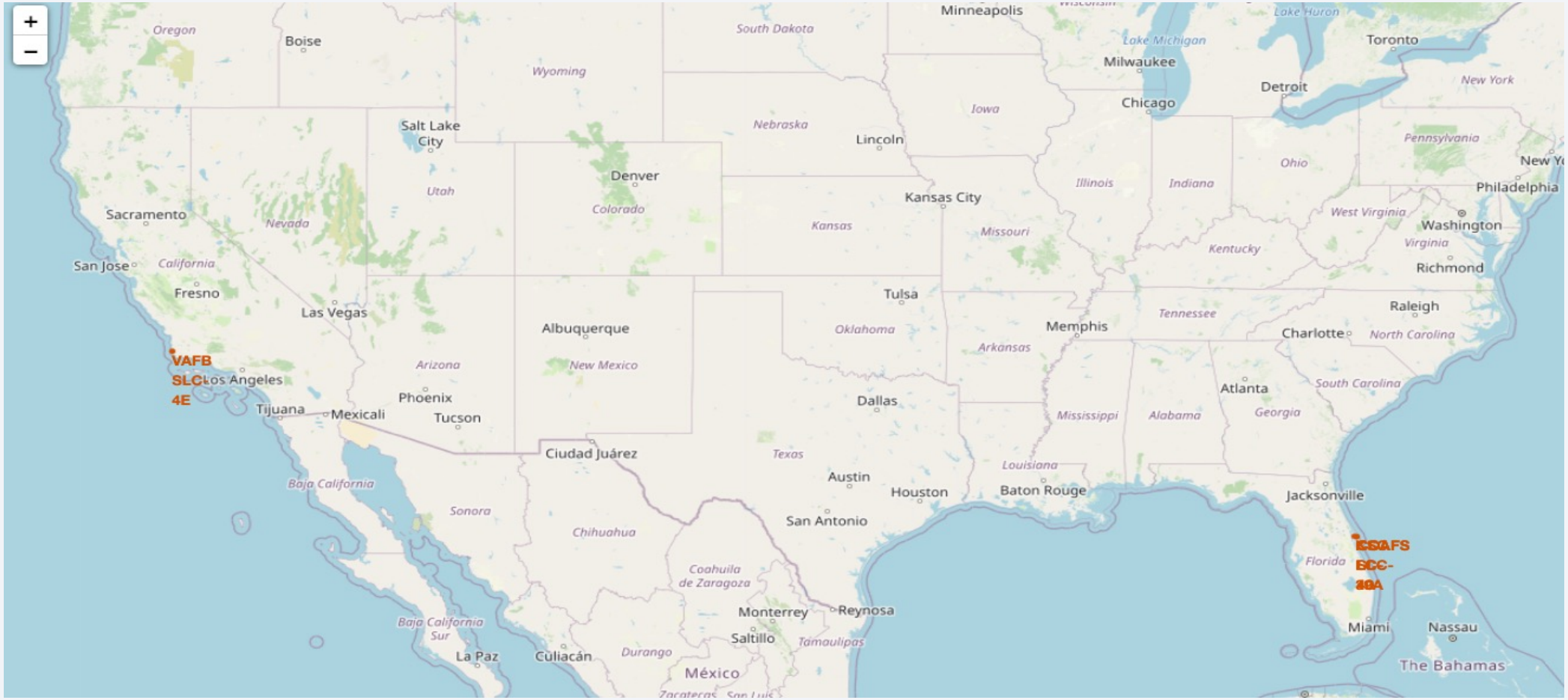
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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
- %sql select count(landing__outcome), landing__outcome from SPACEXTBL \ where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome\ order by count(landing__outcome) desc

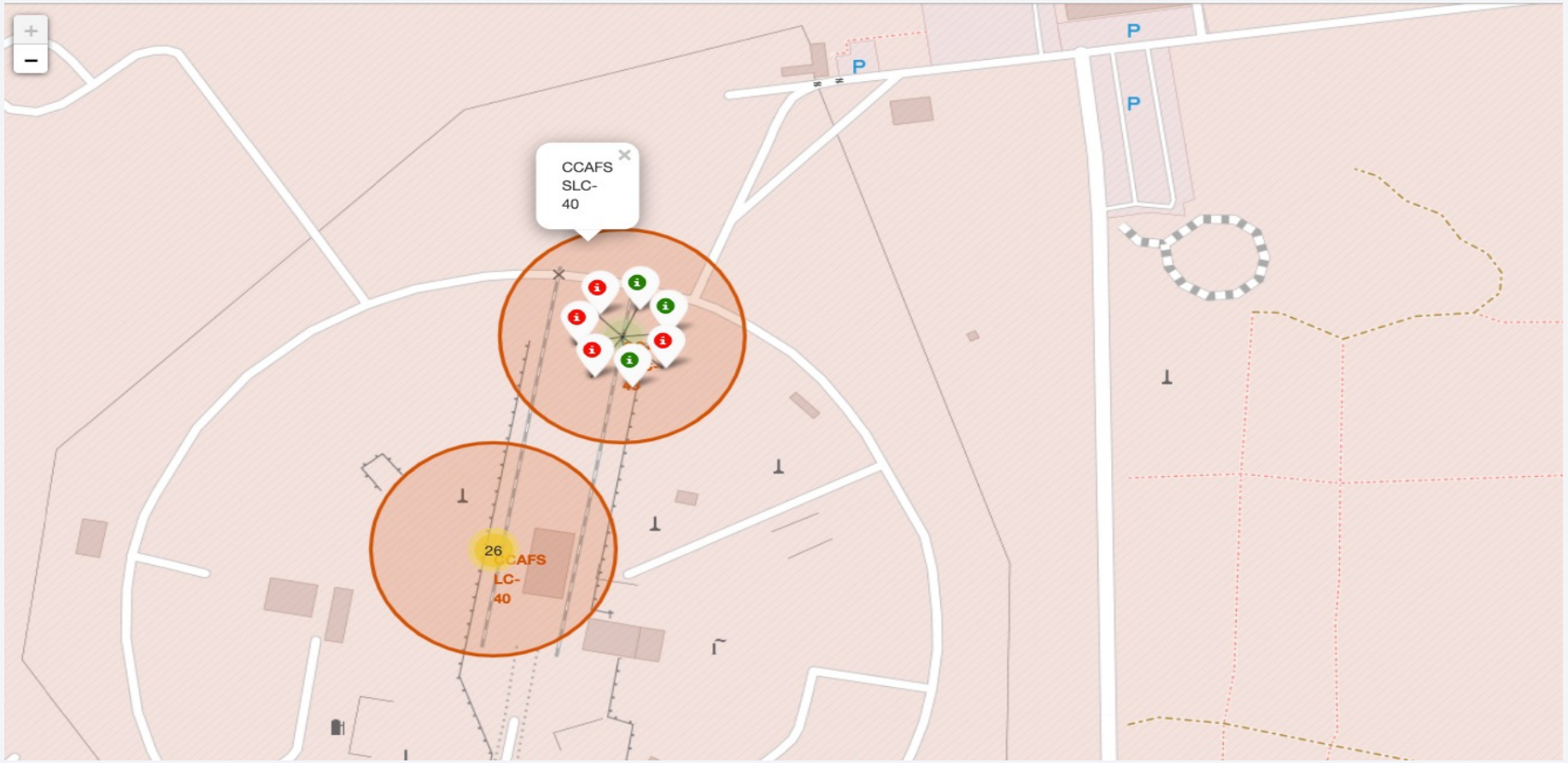
1	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)


Section 3
Launch Sites
Proximities Analysis

All Launch sites with global map markers



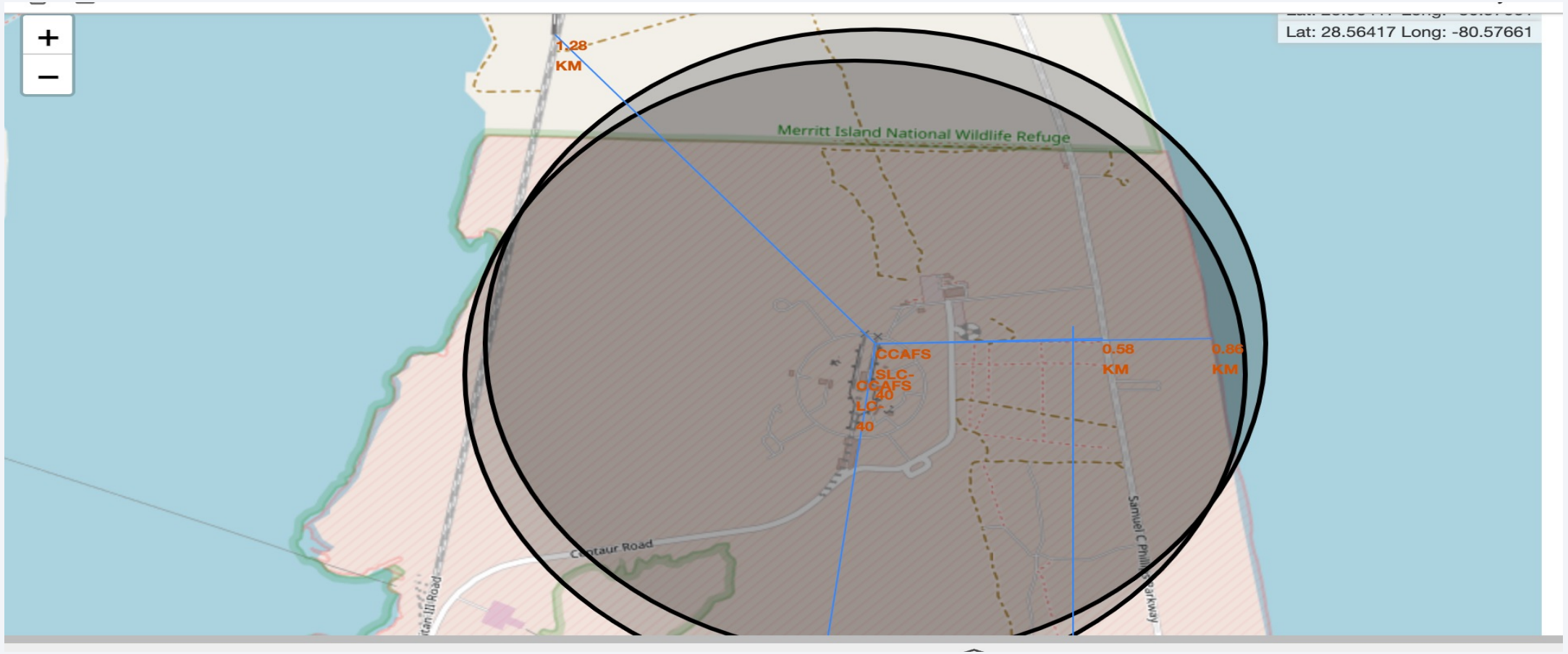
Launch sites with markers and color labels



Green is success and red and failure

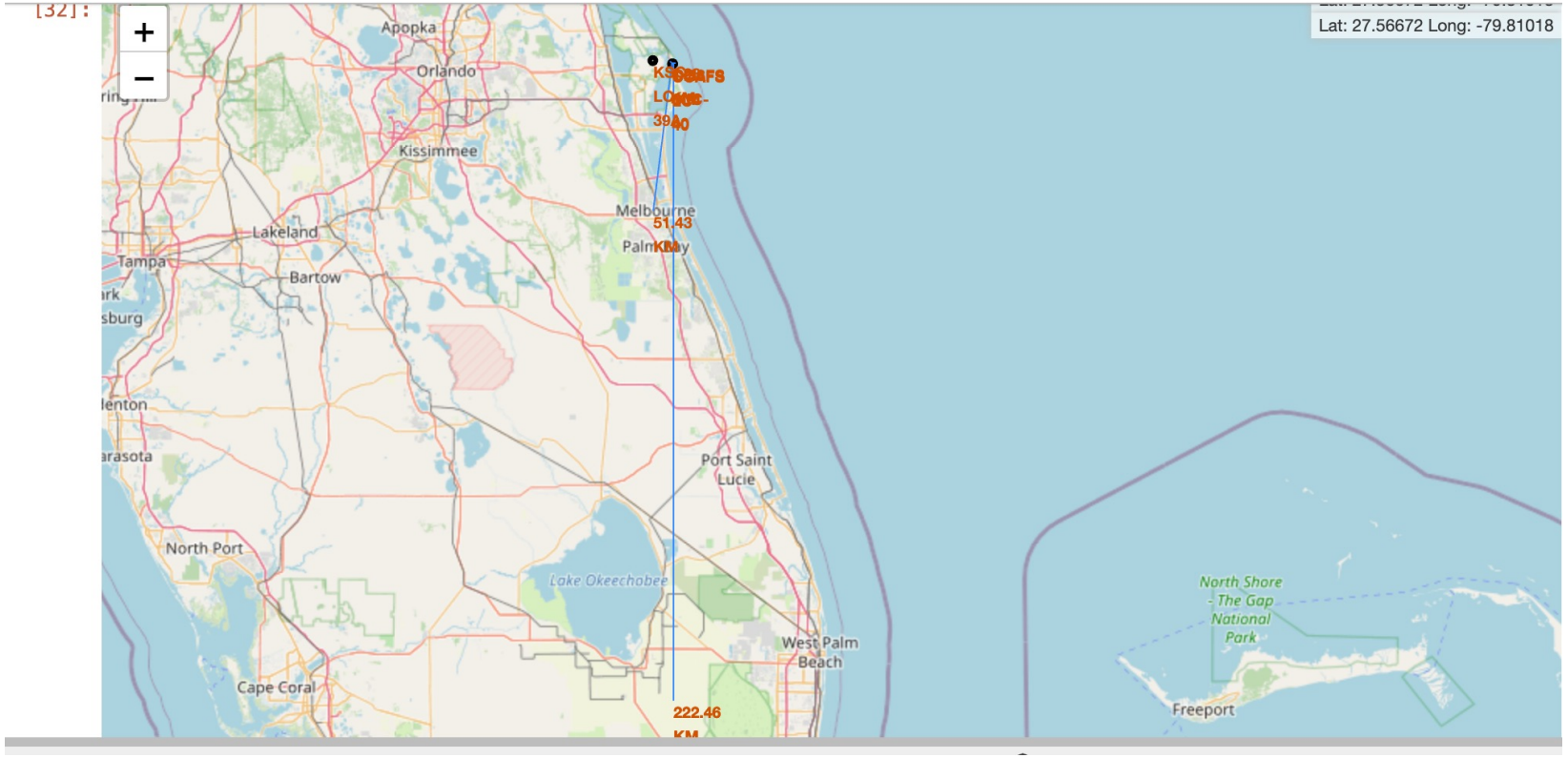
Launch sites to landmarks distance

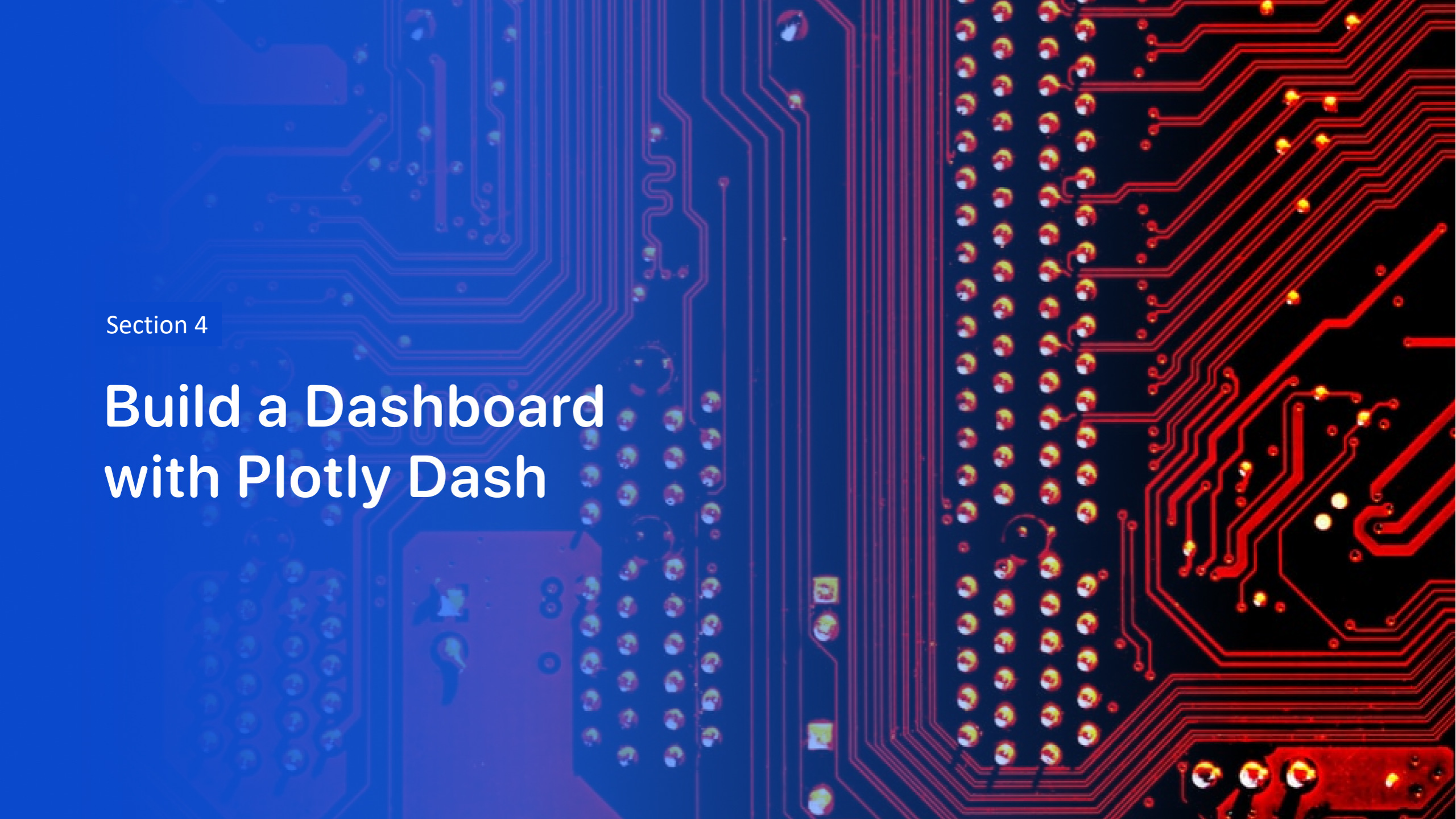
Launch sites to railroad and cost line – not close



Launch sites to landmarks distance

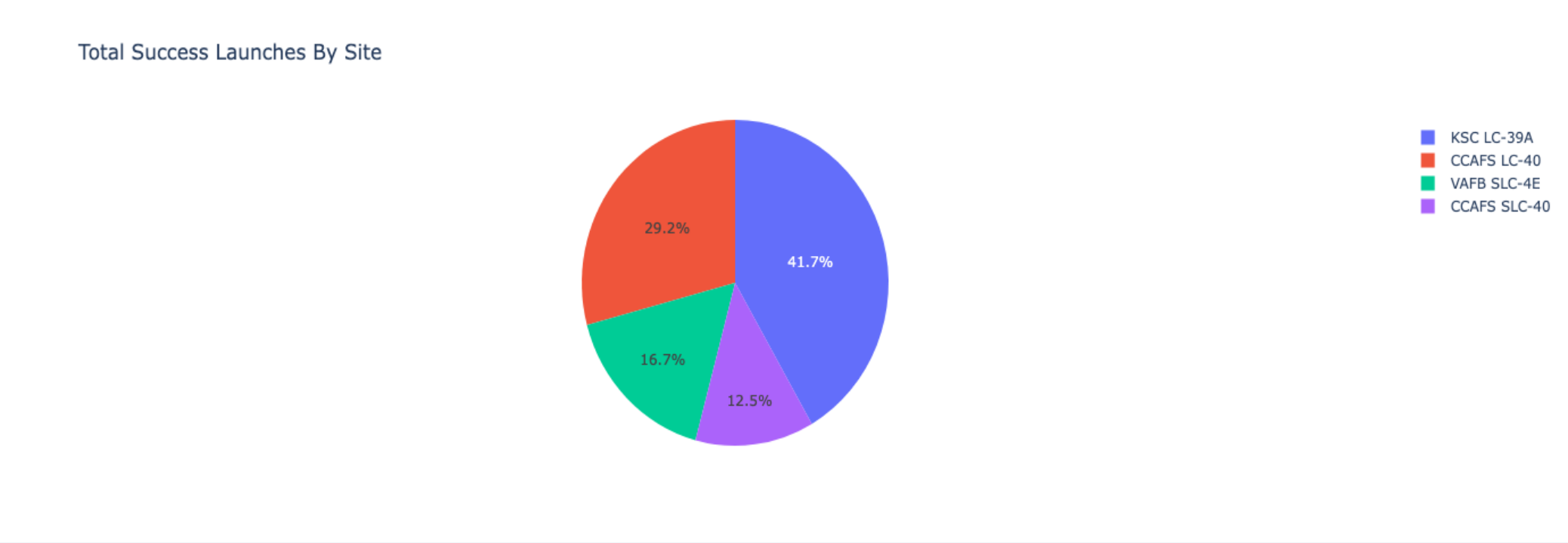
Launch sites to town and equator – not close




Section 4
Build a Dashboard with Plotly Dash

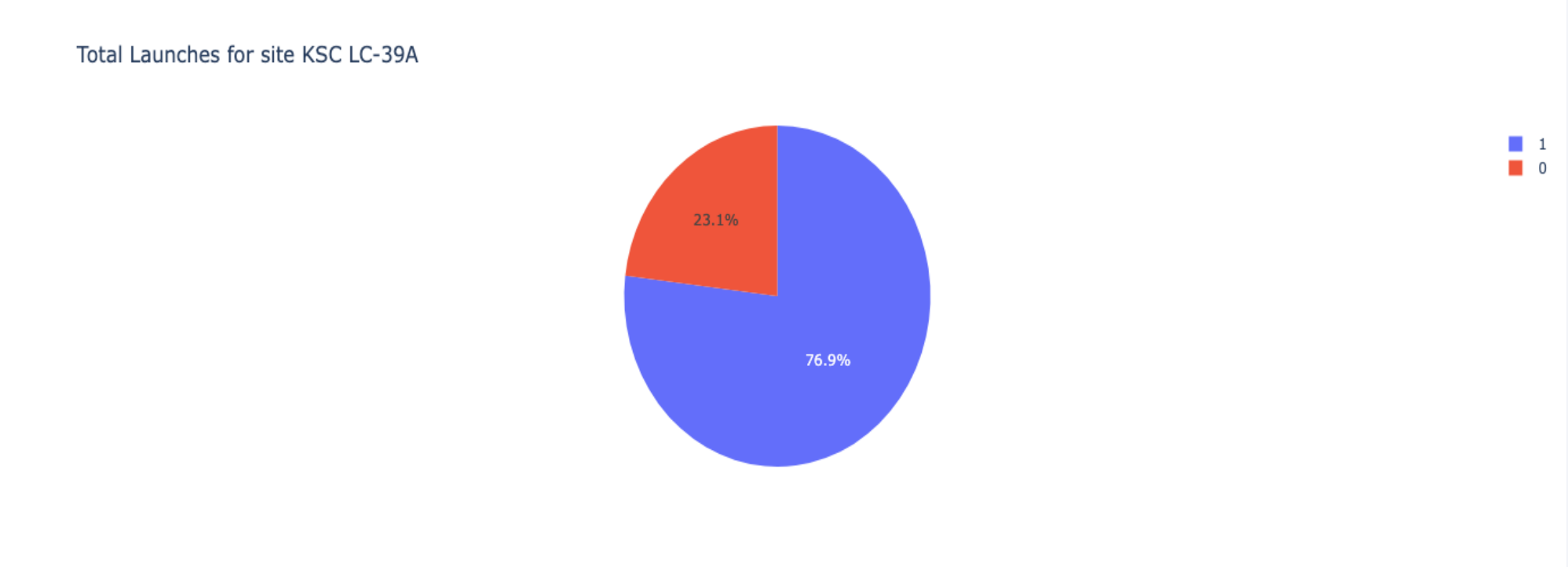
Success Rate for all launch sites

KSC LC-39A has the highest success rate

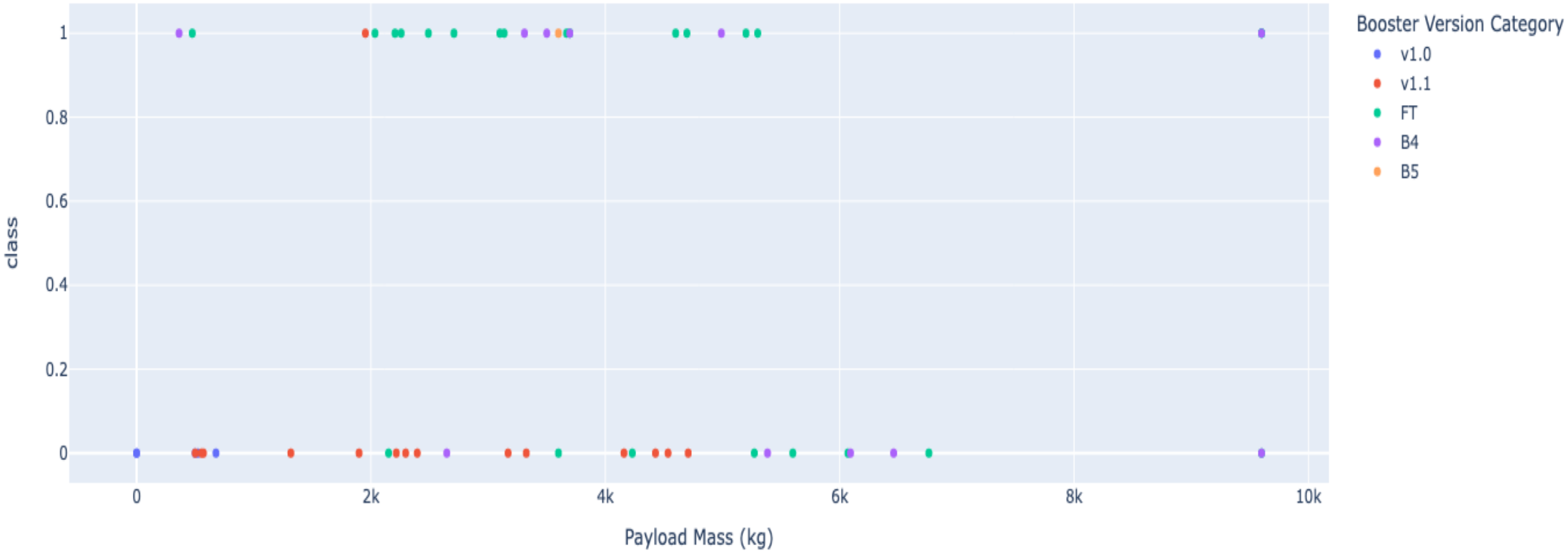


Highest launch site success rate

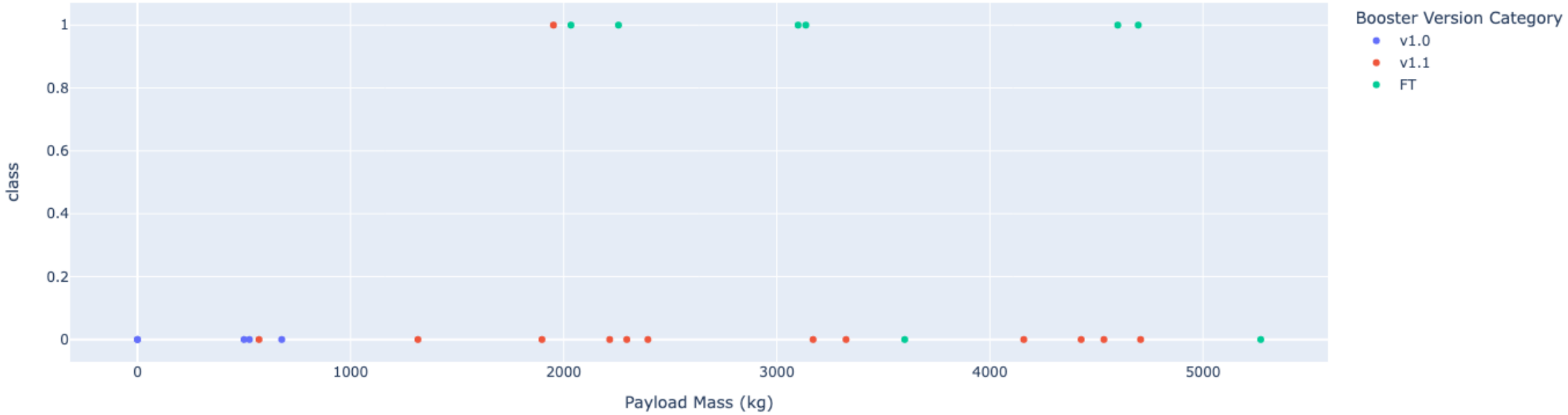
- KSC LC-39A has close to 77% success rate



Success rate of all launch sites

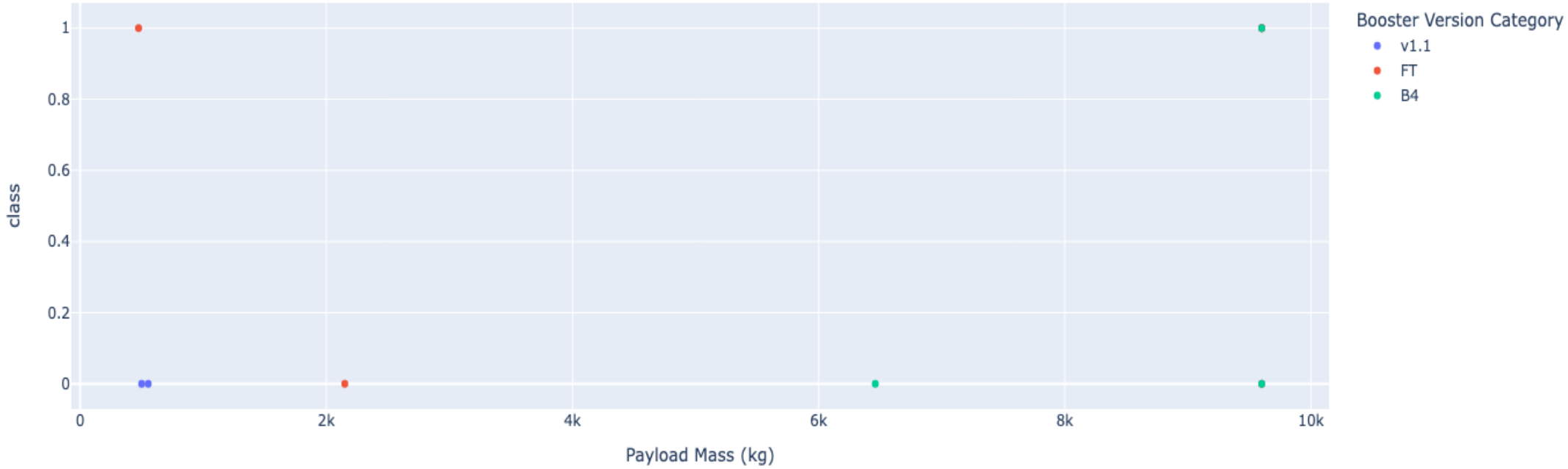


Success rate of CCAFS LC-40



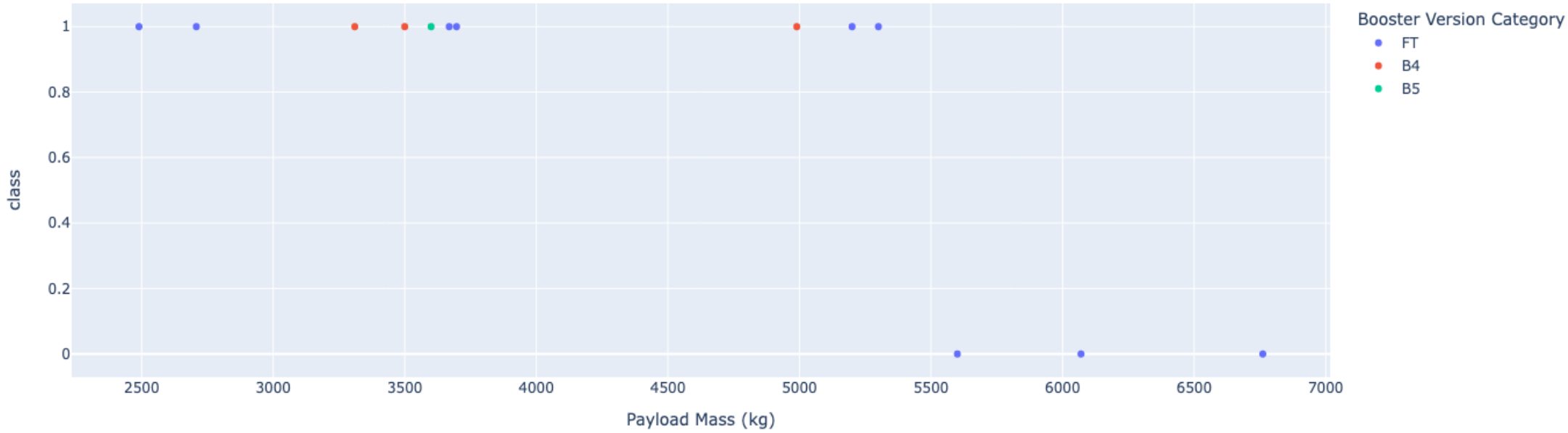
Payload not to exceed 60K, Booster FT is very successful

Success rate of VAFB SLC-4E



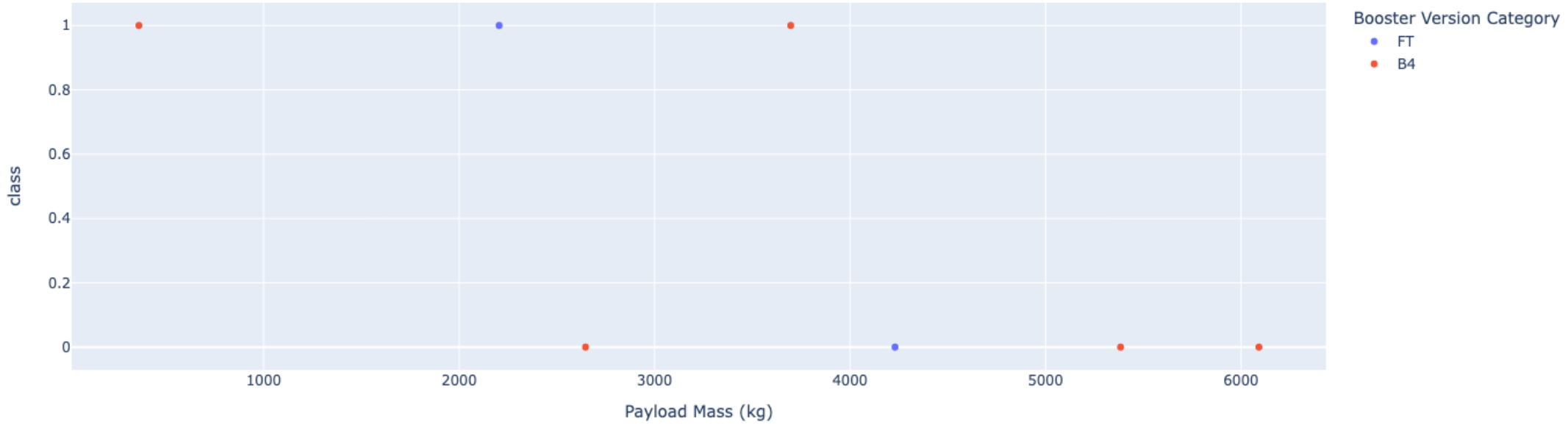
Payload almost all the way to 10k, very little data

Success rate of KSC LC-39A



Zero success rate above 5.5K, all are FT booster version

Success rate of CCAFS SLC-40



No significant difference in Booster, zero success rate above 4K

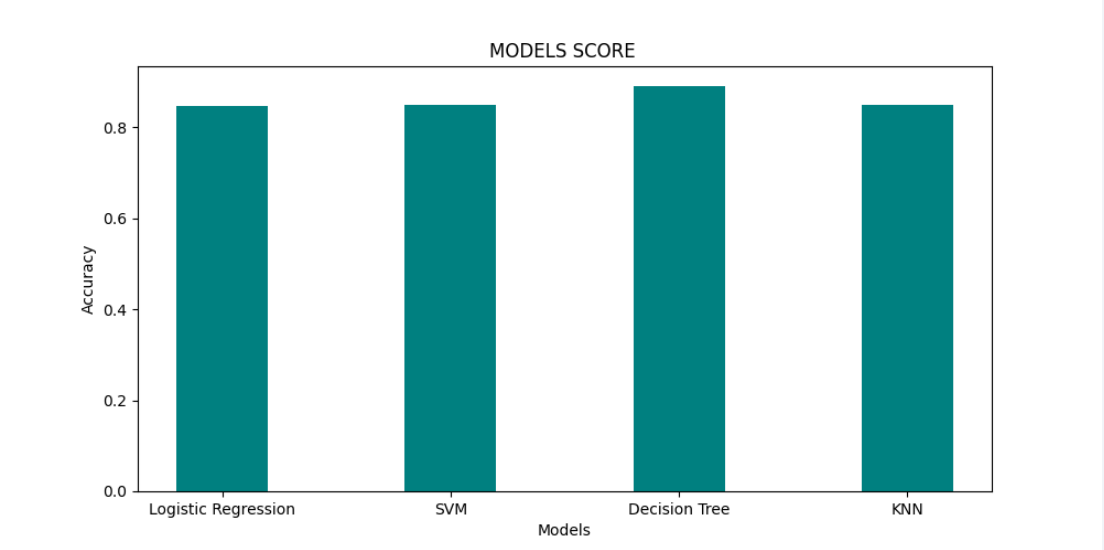
Section 5

Predictive Analysis (Classification)

Classification Accuracy

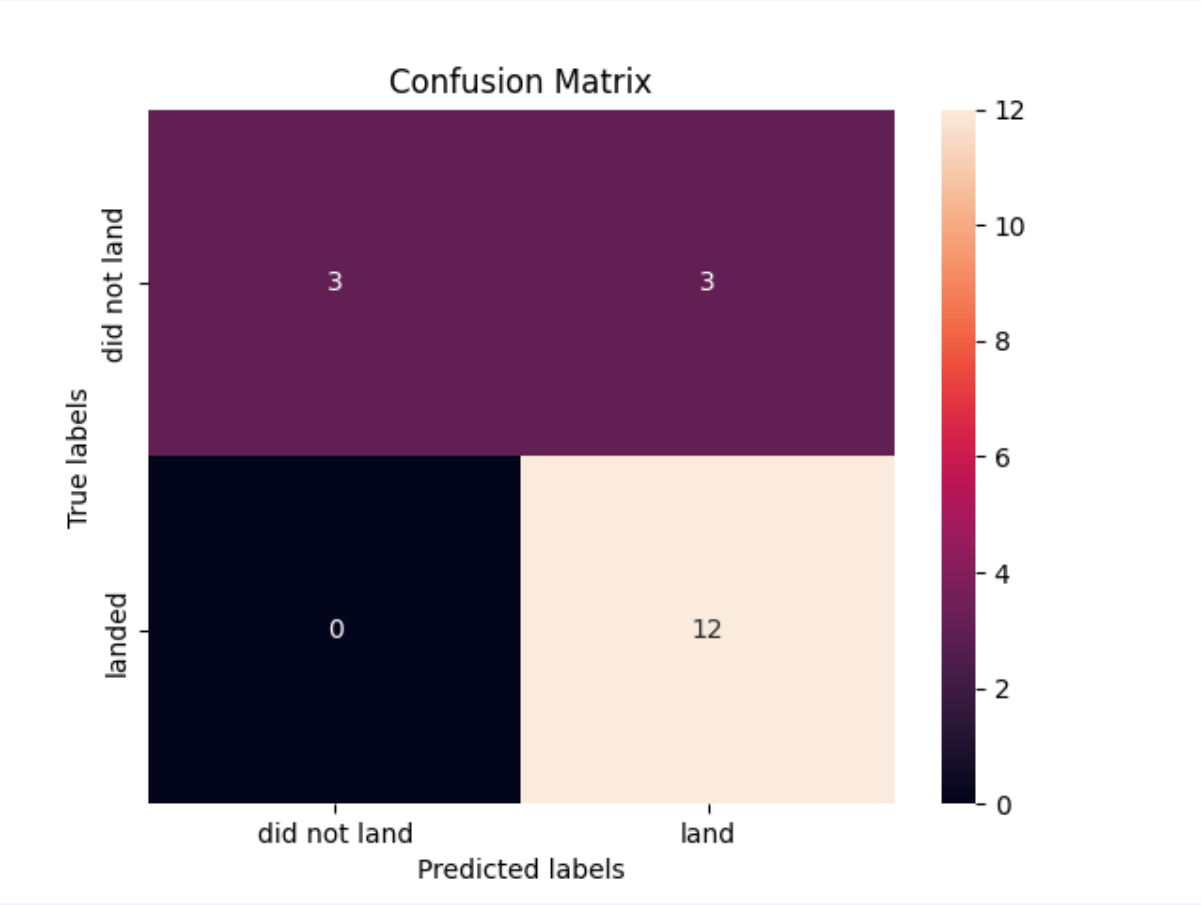
- Decision Tree has the highest accuracy

Logistic Regression Accuracy :	0.8464
SVM Accuracy :	0.8482
Decision Tree Accuracy :	0.8892
KNN Accuracy :	0.8482



Confusion Matrix

- Decision Tree confusion matrix – 15/18 was predicted correctly



Conclusions

- Success rate has improved over the years with 90% at 2020
- Success rate improved with higher payloads
- ES-L1, GEO, HEO, and SSO has 100% success rate, others are better than 50%
- KSC LC-39A is the top launch sites with the highest success rate
- Decision tree classifier has highest accuracy

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!