IBM Developer
SKILLS NETWORK

SPACE Y

Winning Space Race with Data Science

Ivan Lam
4 December 2022

Outline

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

Executive Summary

- How can we predict if the Falcon 9 first stage will land?
- Methodologies:
  - Data collection using public APIs and web scraping public web pages
  - Data wrangling
  - Exploratory data analysis
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using machine learning
- Results:
  - Use visualization to identify features common to successful launches
  - Find most accurate machine learning algorithm

Introduction

- Space Y is a newly formed rocket company
- By analyzing publicly available data from our competitor, Space X, using data science and machine learning, we can identify features that are characteristic of a successful rocket launch
- By focusing on these features, we can optimize the probability of success and minimize the cost of a launch, allowing us to make competitive bids

Section 1

Methodology

Methodology

Executive Summary

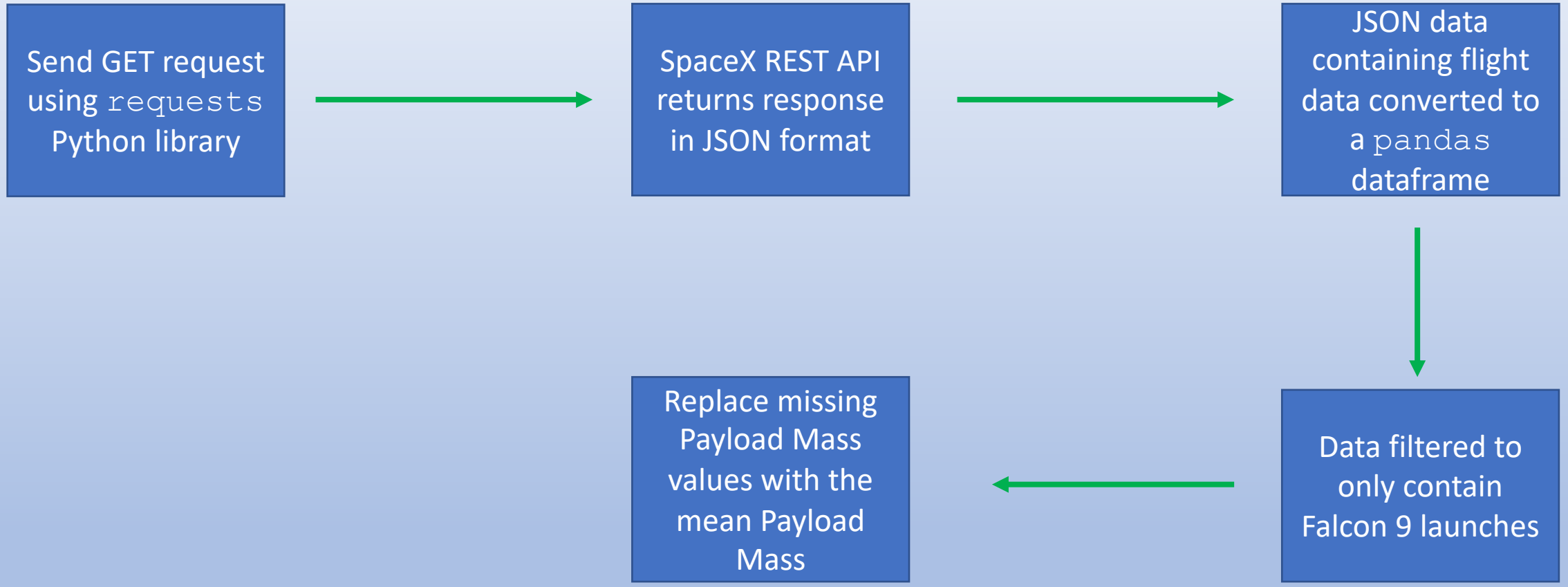
- Data collection methodology:
  - Data collected from SpaceX public API and web scraping relevant Wikipedia pages
- Perform data wrangling
  - Filter to only include relevant data
  - Replace missing data with mean values
  - One-hot encoding to convert categorial columns to numerical columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree and KNN models tested and accuracy determined

Data Collection

- Rocket launch data obtained from the SpaceX API: https://api.spacexdata.com/v4/launches/past

- Historical Falcon 9 and Falcon Heavy launch data obtained from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922”

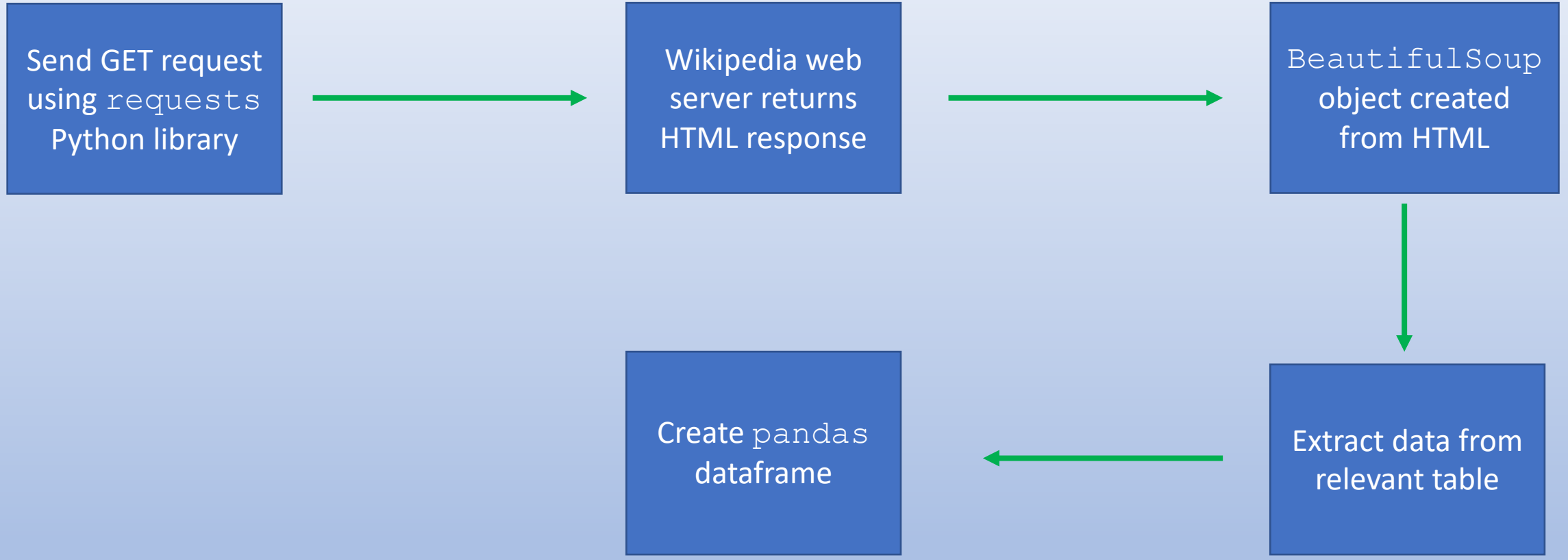
Data Collection – SpaceX API



GitHub URL of the completed SpaceX API calls notebook:

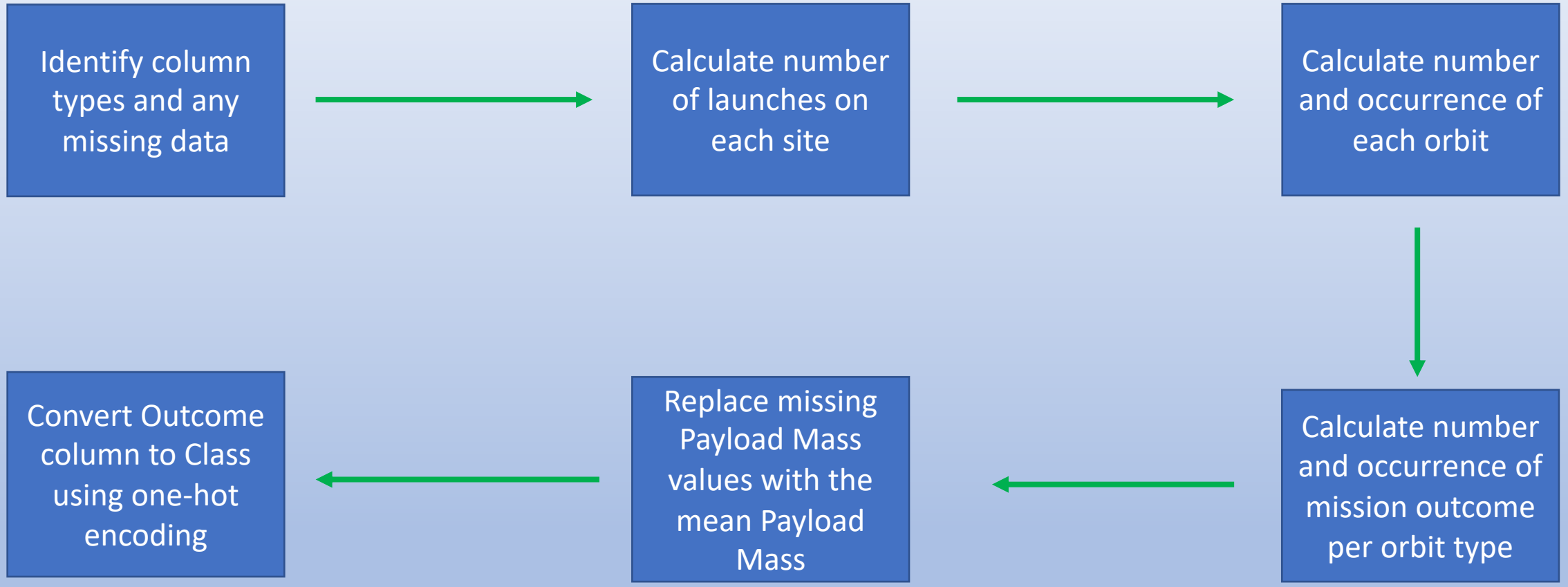
https://github.com/ictl/IBM-DS0720EN-Project/blob/master/Data%20Collection%20API%20Lab.ipynb

Data Collection – Web Scraping



GitHub URL of the completed web scraping notebook: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb

Data Wrangling and EDI



GitHub URL of completed data wrangling notebook: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/EDA%20Lab.ipynb

EDA with Data Visualization

- Flight Number vs Payload Mass – scatter plot using `seaborn catplot`
- Flight Number vs Launch Site – scatter plot using `seaborn catplot`
- Payload Mass vs Launch Site – scatter plot using `seaborn catplot`
- Orbit Type vs Success Rate – bar chart using `pandas`
- Flight Number vs Orbit Type – scatter plot using `seaborn catplot`
- Payload Mass vs Orbit Type – scatter plot using `seaborn catplot`
- Year vs Success Rate – line graph using `pandas`

Completed EDA with data visualization notebook: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL (1)

1. Find unique launch sites:
   - `SELECT UNIQUE LAUNCH_SITE FROM SPACEX;`
2. Find 5 records where launch site begins with ‘KSC’:
   - `SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5;`
3. Find total payload mass carried by boosters launched by NASA (CRS):
   - `SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)’;`
4. Find average payload mass carried by booster version F9 v1.1:
   - `SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';`

EDA with SQL (2)

5. Find the date where the first successful landing outcome in drone ship was achieved:

   - `SELECT DATE FROM SPACEX WHERE DATE = (SELECT MIN(DATE) FROM SPACEX);`

6. Find names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

   - `SELECT UNIQUE BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;`

7. Find total number of successful and failure mission outcomes:

   - `SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEX GROUP BY MISSION_OUTCOME;`

8. Find names of the booster versions which have carried the maximum payload mass:

   - `SELECT UNIQUE BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);`

EDA with SQL (3)

9. Find records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017:

   - `SELECT MONTHNAME(DATE) AS MONTH, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)' AND DATE LIKE '2017%';`

10. Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.:

    - `SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS SUCCESSFUL_LANDING_OUTCOMES FROM SPACEX WHERE LANDING__OUTCOME LIKE 'Success%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY SUCCESSFUL_LANDING_OUTCOMES DESC;`

GitHub URL of completed EDA with SQL notebook: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

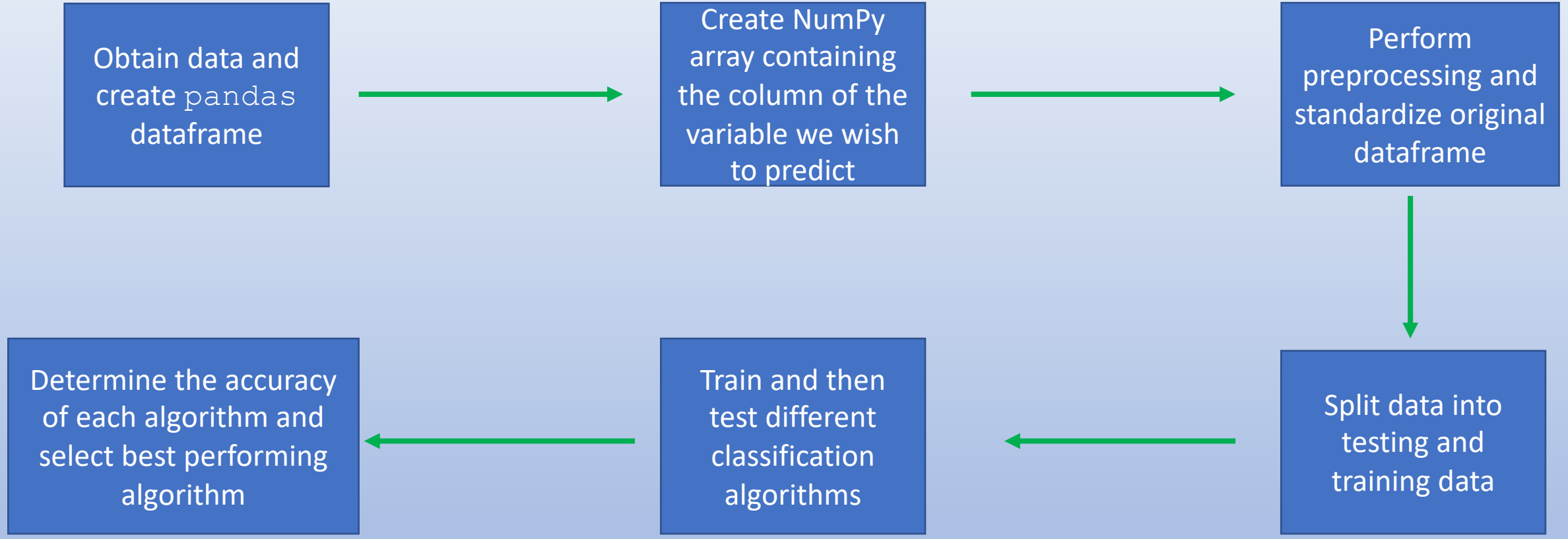
- Folium markers and circles added to identify launch sites
- Marker cluster added to identify successful launches, many of which will have the same coordinate
- Markers and lines added to show the distance between a launch site and the nearest highway and railroad routes which materials can be delivered to the launch site
- Markers and lines added to show the distance between a launch site and the nearest coastline and city – areas where a rocket could potentially land if launch not successful
- GitHub URL of completed interactive map with Folium map: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20Lab.ipynb
- Folium maps do not display in Jupyter notebooks uploaded to GitHub; the original file in IBM Watson can be found here: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/40c96d89-cc44-4e9a-88e7-cbabbbea7e70/view?access_token=5f146469b875b526f4370d145d310dbf4df7cc321ef3bda12ababde630a2ea83

Build a Dashboard with Plotly Dash

- Plotly dashboard contained a pie chart showing the proportion of successful launches by site
- Dropdown menu allows focusing the pie chart to show the success rate of a particular site
- Scatter plot showing correlation of payload and success rate for different boosters and launch sites
- Payload range slider allows focusing the scatter plot to a particular range of payload mass

- GitHub URL of completed Plotly Dash lab: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)



- Algorithms tested: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN)
- GitHub URL of completed predictive analysis lab: https://github.com/ictl/IBM-DS0720EN-Project/blob/master/Machine%20Learning%20Prediction%20Lab.ipynb

Results

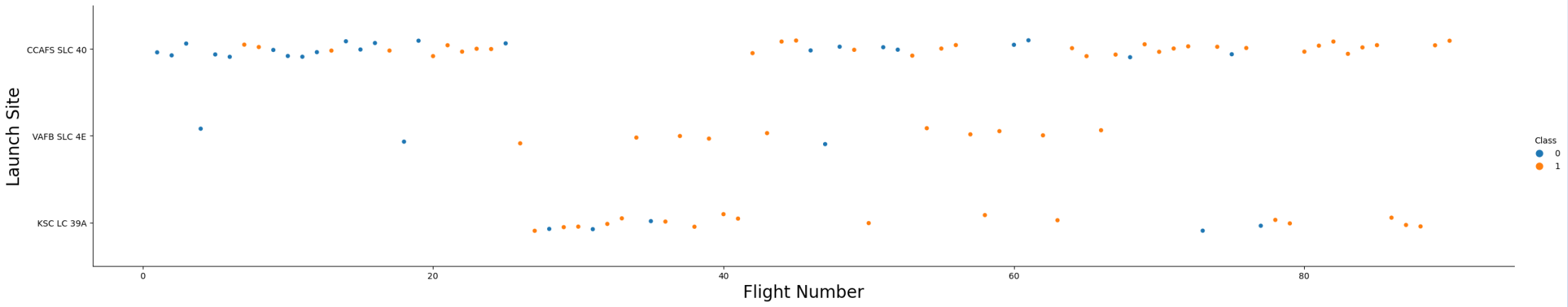
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

SPACE Y

Section 2

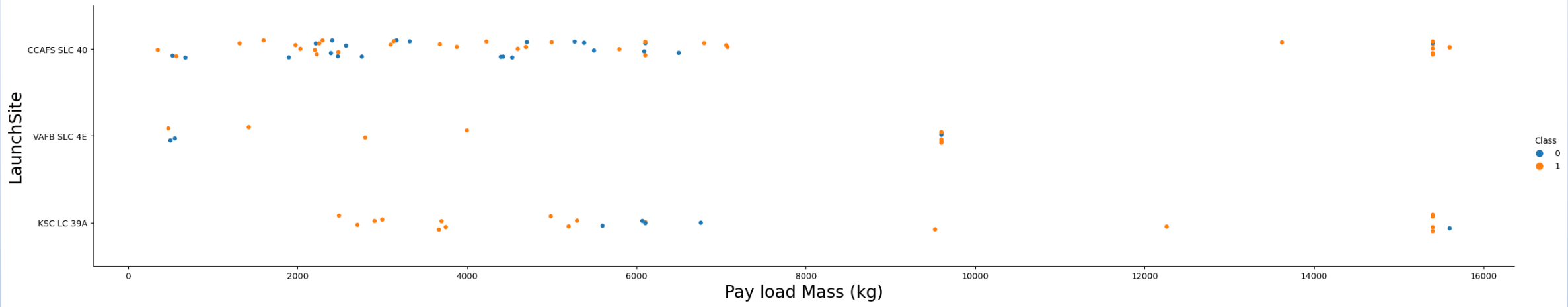
Insights drawn from EDA

Flight Number vs. Launch Site



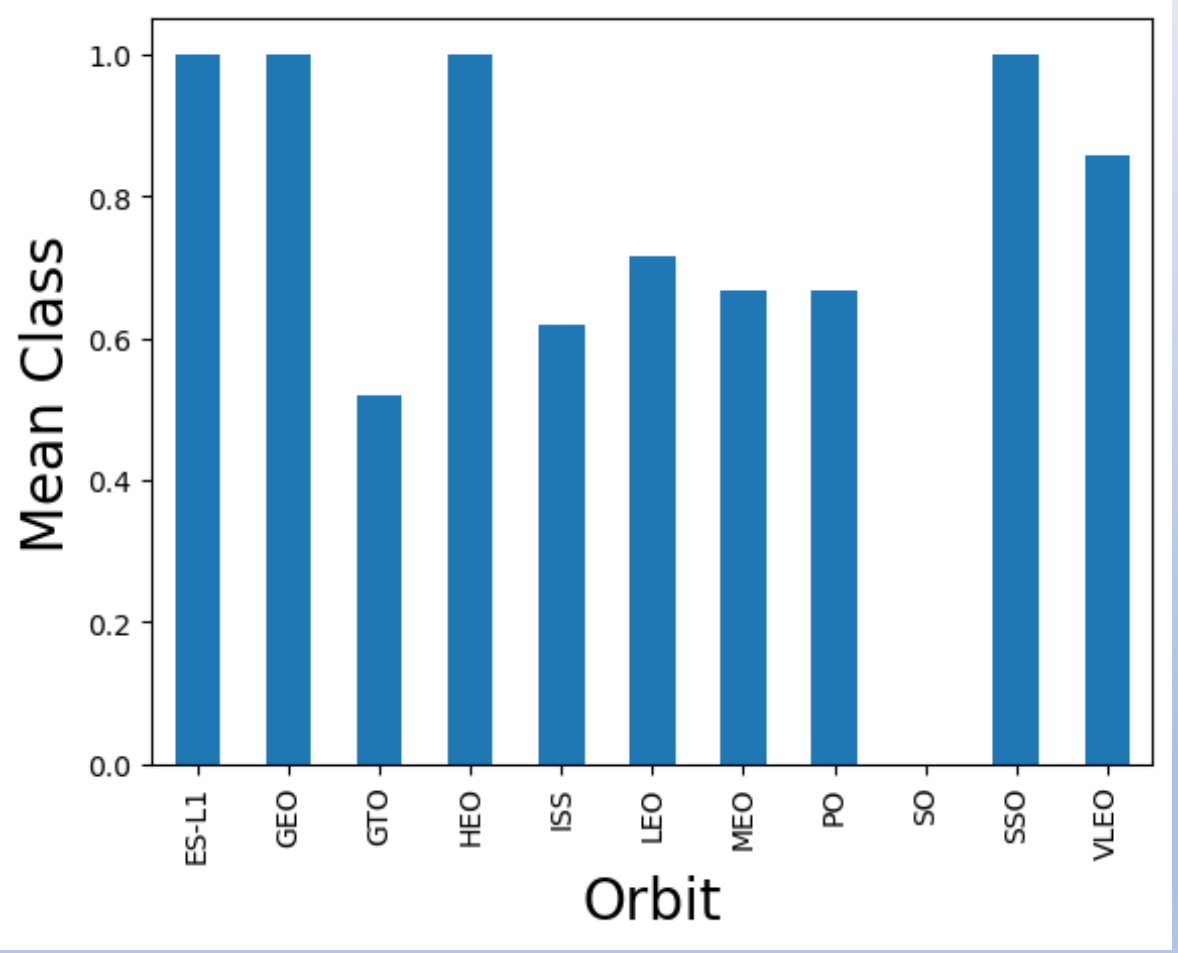
- Higher flight numbers represent more recent launches
- The most recent flight launches are more likely to be successful and launched from CCAPS SLC 40 or KSC LC 39A

Payload vs. Launch Site



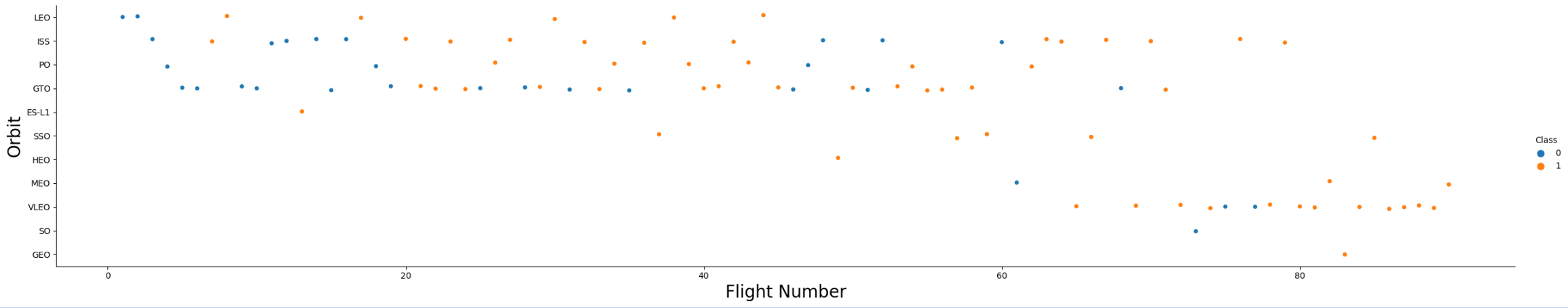
- Higher payload mass is more likely to result in a successful launch
- There are no launches at VAFB SLC 4E with payload mass > 10,000 kg. It may be the case that this launch site is not capable of supporting such a high payload mass

Success Rate vs. Orbit Type



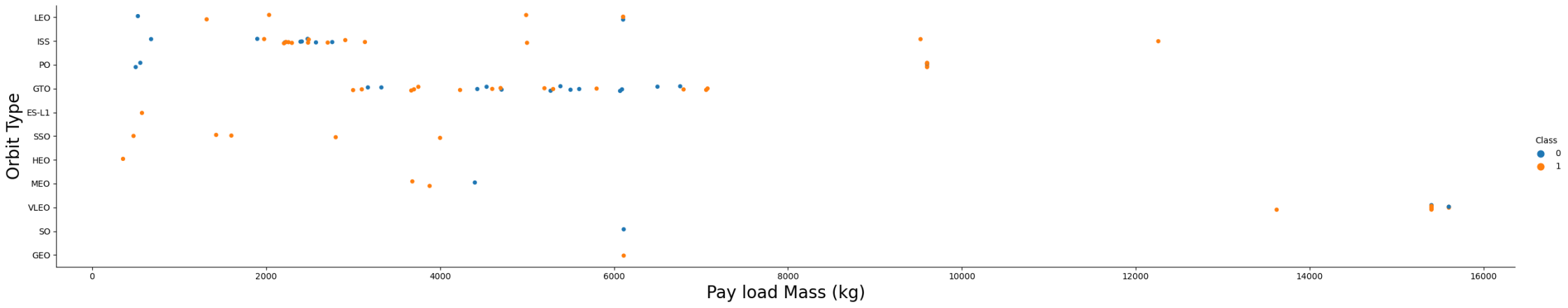
- ES-L1, GEO, HEO and SSO have the highest success rate
- Orbits which have a low altitude tend to have a lower success rate, with the exception of VLEO and GTO
- Limited conclusions can be drawn from this graph alone, as number of launches and date of launches are not represented

Flight Number vs. Orbit Type



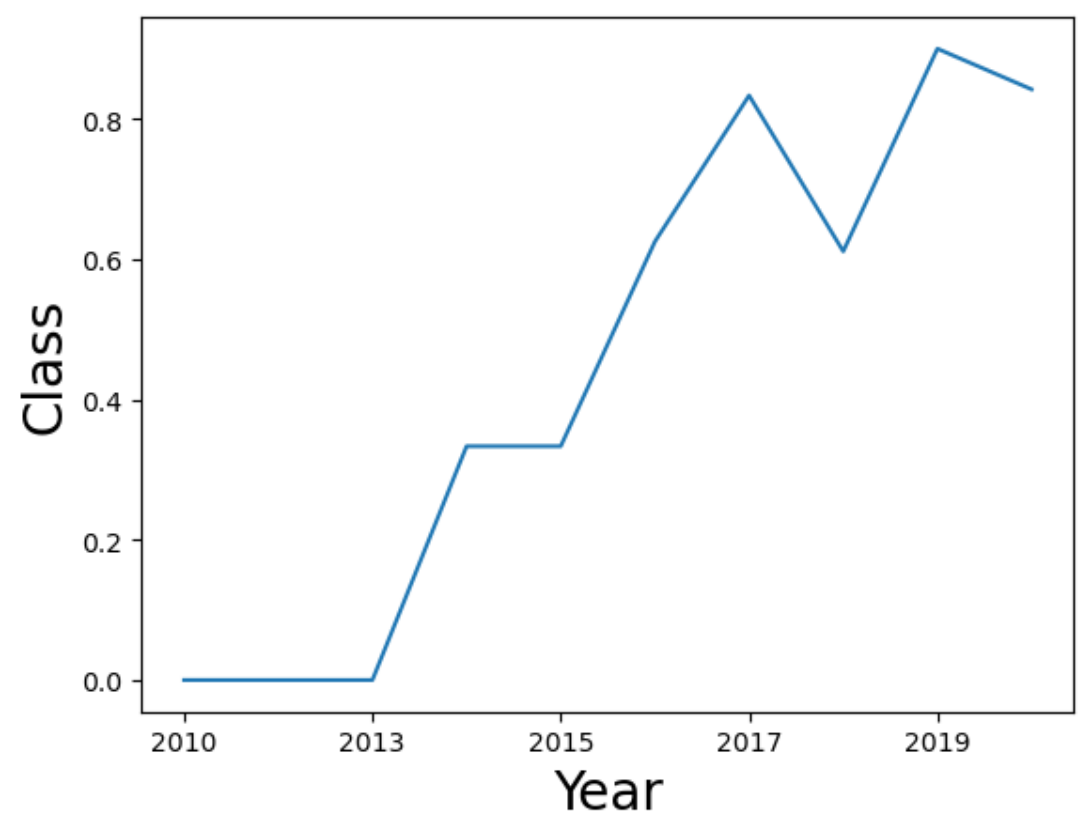
- More recent flights tend to have an orbit type of VLEO, which has a very low altitude
- Recent technological advances improve the success rate of lower altitude orbits
- Low altitude orbits allow more effective observation from Earth and are therefore more useful

Payload vs. Orbit Type



- The highest payload masses are used in the VLEO orbit type
- VLEO has a very low altitude – orbit types with a higher altitude may be more difficult to achieve with a very high payload mass

Launch Success Yearly Trend



- The success rate generally increased year-on-year from 2013 to 2020

All Launch Site Names

- `SELECT UNIQUE LAUNCH_SITE FROM SPACEX;`

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

- Use of ‘UNIQUE’ returns only unique results, removing any duplicates

Launch Site Names Begin with 'KSC'

- `SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5;`

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	06:00:00	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

- ‘LIKE ‘KSC%’’ will search for string beginning with ‘KSC’, as % is a wildcard character
- ‘LIMIT 5’ restricts number of records returned to 5.

Total Payload Mass

- `SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';`
- Result: 45596 kg

total_payload_mass
45596

- ‘SUM(PAYLOAD_MASS__KG)’ calculates total payload mass
- ‘WHERE CUSTOMER = ‘NASA (CRS)’ filters to only include NASA (CRS)

Average Payload Mass by F9 v1.1

- `SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';`
- Result: 2928 kg

average_payload_mass
2928

- 'AVG(PAYLOAD_MASS__KG)' calculates average payload mass
- 'WHERE BOOSTER_VERSION = 'F9 v1.1' filters to only include booster version F9 v1.1

First Successful Ground Landing Date

- `SELECT DATE FROM SPACEX WHERE DATE = (SELECT MIN(DATE) FROM SPACEX);`
- Result: 2010-06-04

DATE
2010-06-04

- 'SELECT MIN(DATE) FROM SPACEX' finds the lowest (i.e. earliest) date

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT UNIQUE BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

booster_version
F9 B4 B1040.1
F9 B4 B1043.1
F9 FT B1032.1

- 'AND' used to only return results matching both conditions
- 'PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000' used to find payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEX GROUP BY MISSION_OUTCOME;

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

- 'COUNT(MISSION_OUTCOME)' when combined with 'GROUP BY MISSION_OUTCOME' returns the number of occurrences of each unique value in MISSION_OUTCOME

Boosters Carried Maximum Payload

- `SELECT UNIQUE BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);`

- ‘SELECT UNIQUE BOOSTER_VERSION FROM SPACEX’ returns only unique booster versions, removing any duplicates
- ‘SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX’ finds the maximum payload mass
- Subquery is used to link the two statements together

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

2017 Launch Records

- SELECT MONTHNAME(DATE) AS MONTH, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)' AND DATE LIKE '2017%';

MONTH	landing__outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- ‘MONTHNAME(DATE)’ returns the month name for the given date
- ‘LANDING__OUTCOME = 'Success (ground pad)’ ’ returns results only where the landing outcome is successful on the ground pad
- ‘DATE LIKE '2017%'' returns results only where the date begins with ‘2017’ (i.e. year is 2017)

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

- `SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS SUCCESSFUL_LANDING_OUTCOMES FROM SPACEX WHERE LANDING__OUTCOME LIKE 'Success%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY SUCCESSFUL_LANDING_OUTCOMES DESC;`

landing__outcome	successful_landing_outcomes
Success (drone ship)	5
Success (ground pad)	3

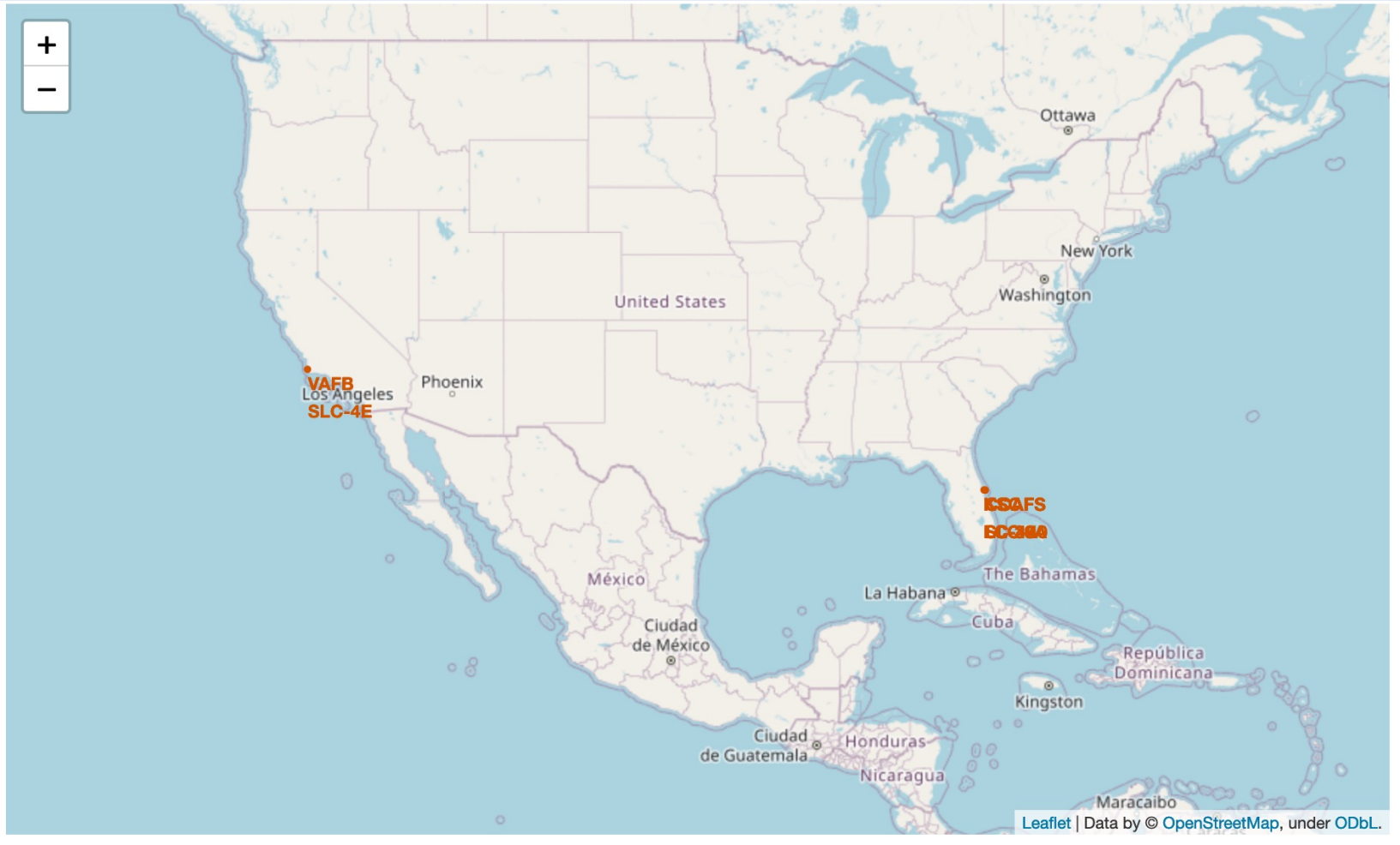
- ‘COUNT(LANDING__OUTCOME)’ when combined with ‘GROUP BY LANDING__OUTCOME’ returns the number of occurrences of each unique value in LANDING__OUTCOME
- ‘LANDING__OUTCOME LIKE 'Success%'’ returns results where LANDING__OUTCOME begins with ‘Success’.
- ‘DATE BETWEEN '2010-06-04' AND '2017-03-20' ’ returns results between the two dates
- ‘ORDER BY SUCCESSFUL_LANDING_OUTCOMES DESC’ orders in descending order

SPACE Y

Section 3

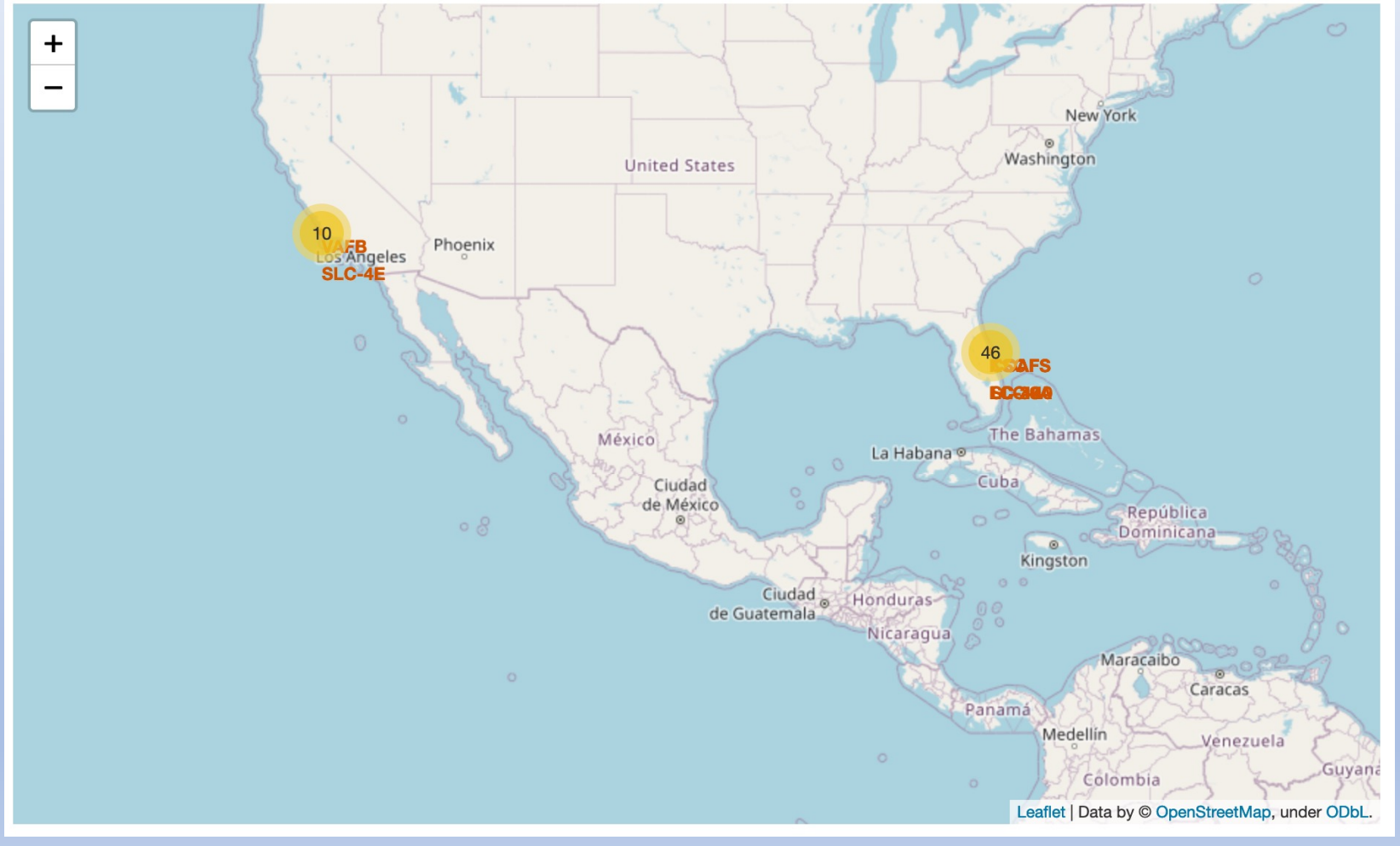
Launch Sites Proximities Analysis

Location of Launch Sites



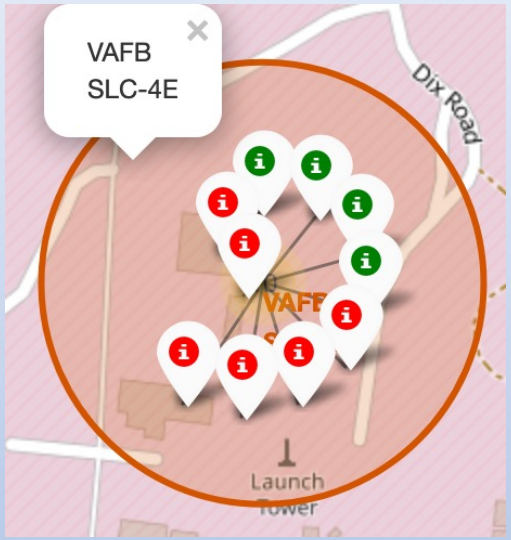
- Launch sites are close to the sea
- If a rocket launch fails, it will be more likely to land in the sea than on land, where it may cause damage to buildings etc.

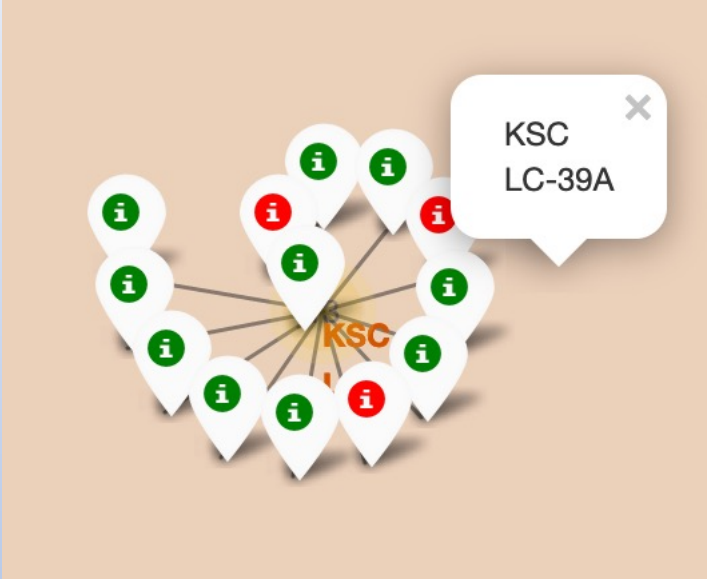
Number of launches per location

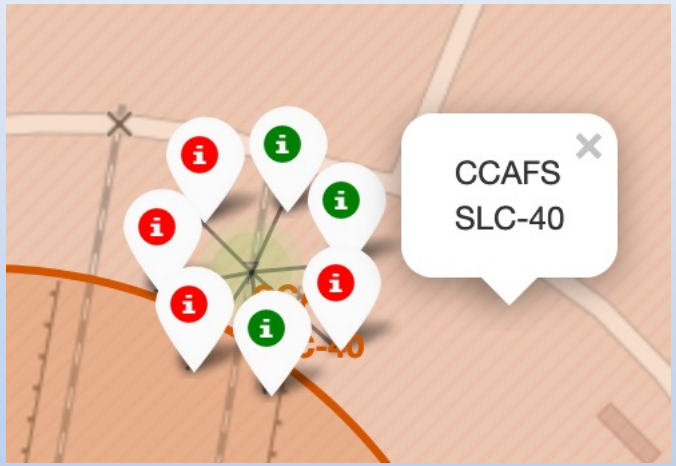


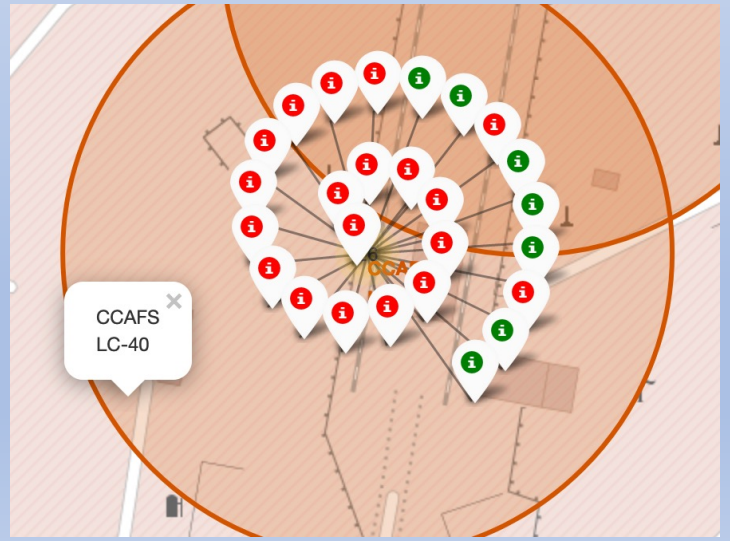
- Most launches took place on the East coast, rather than the West coast

Launch success rate per location



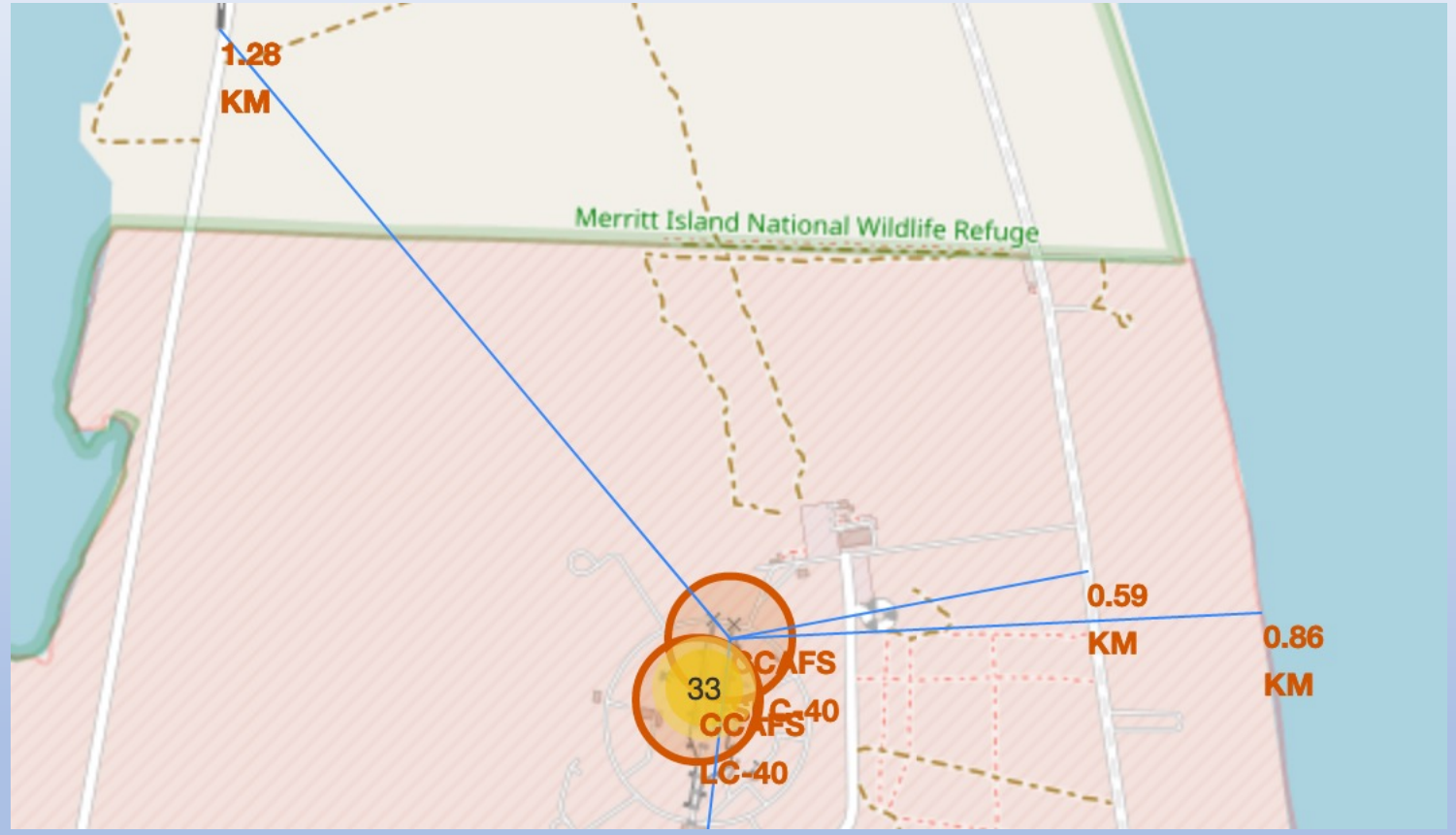






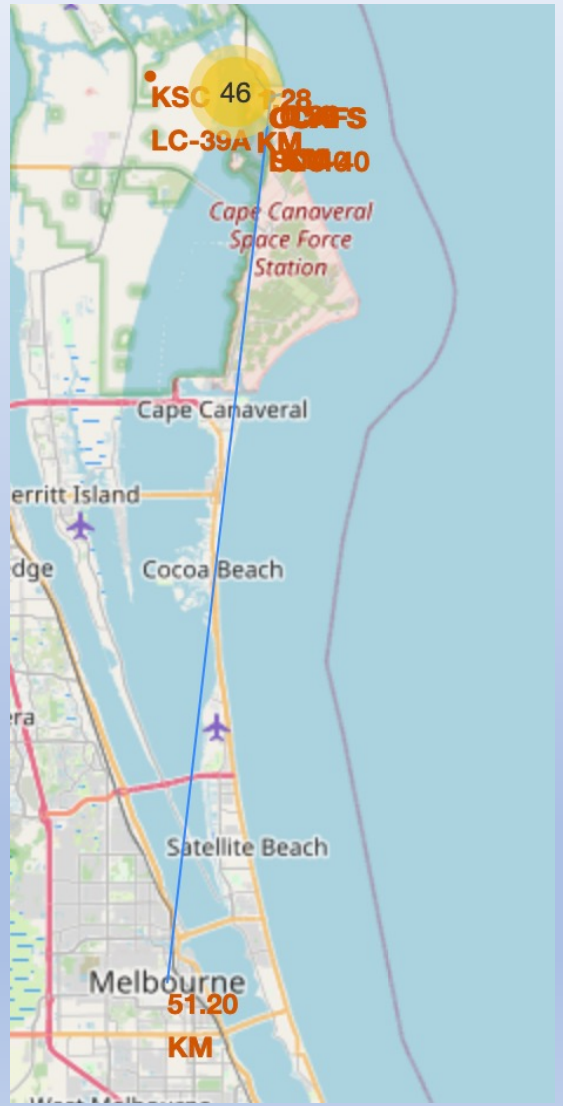
- KSC LC-39A has the highest success rate of the four launch sites

Distance between launch sites and points of interest



- Launch sites are close to highways and railroads
  - Allows easy delivery of heavy materials
- Launch sites are close to the coastline
  - Rockets more likely to fall into the ocean, causing less damage

Distance between launch sites and points of interest

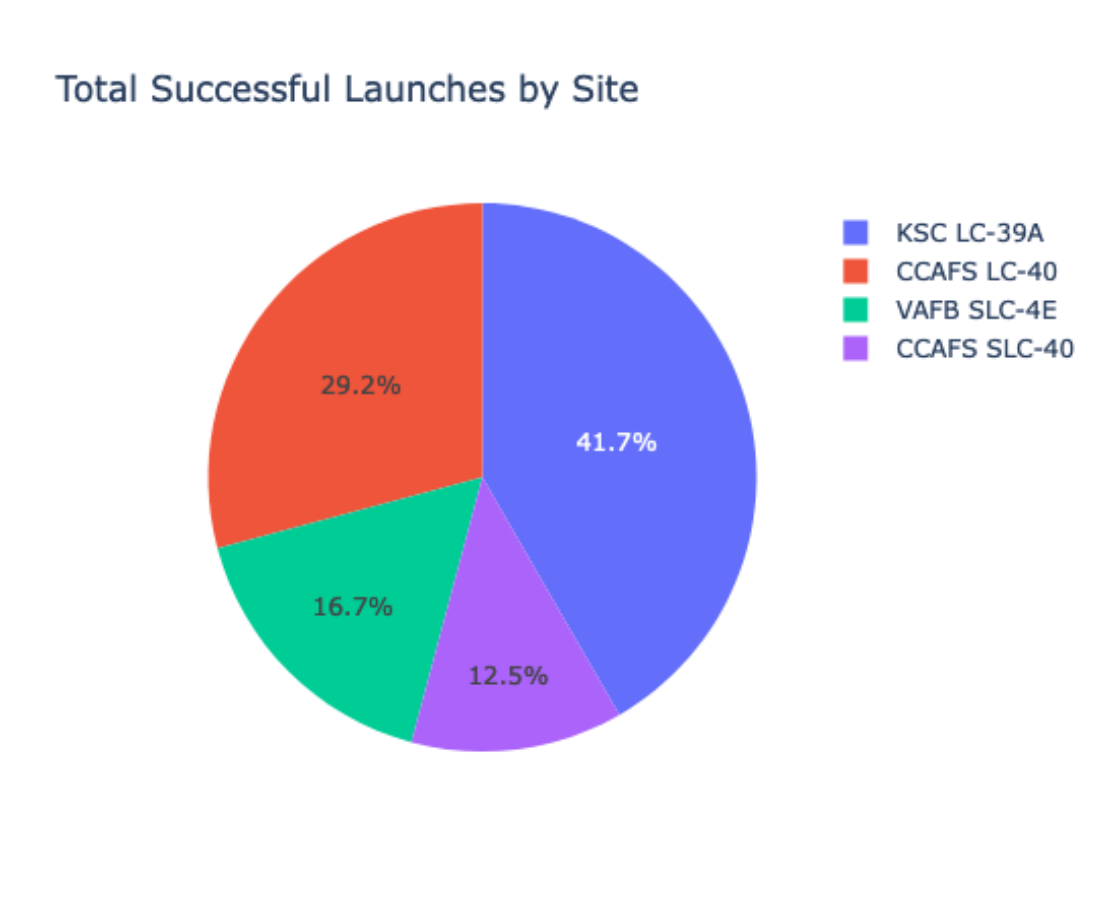


- Launch sites are far away from other cities
  - Rockets less likely to land in cities if they are further away, minimizing potential damage

Section 4

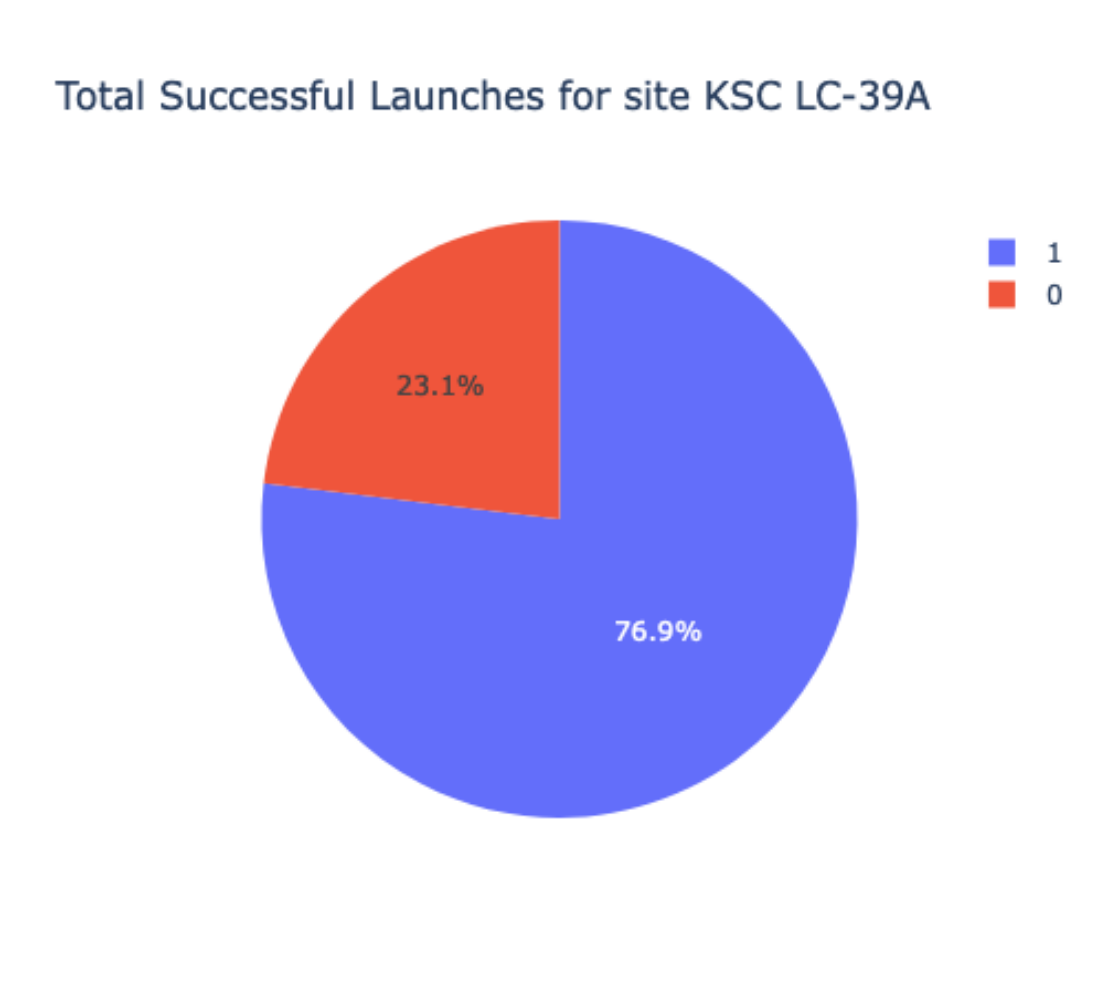
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



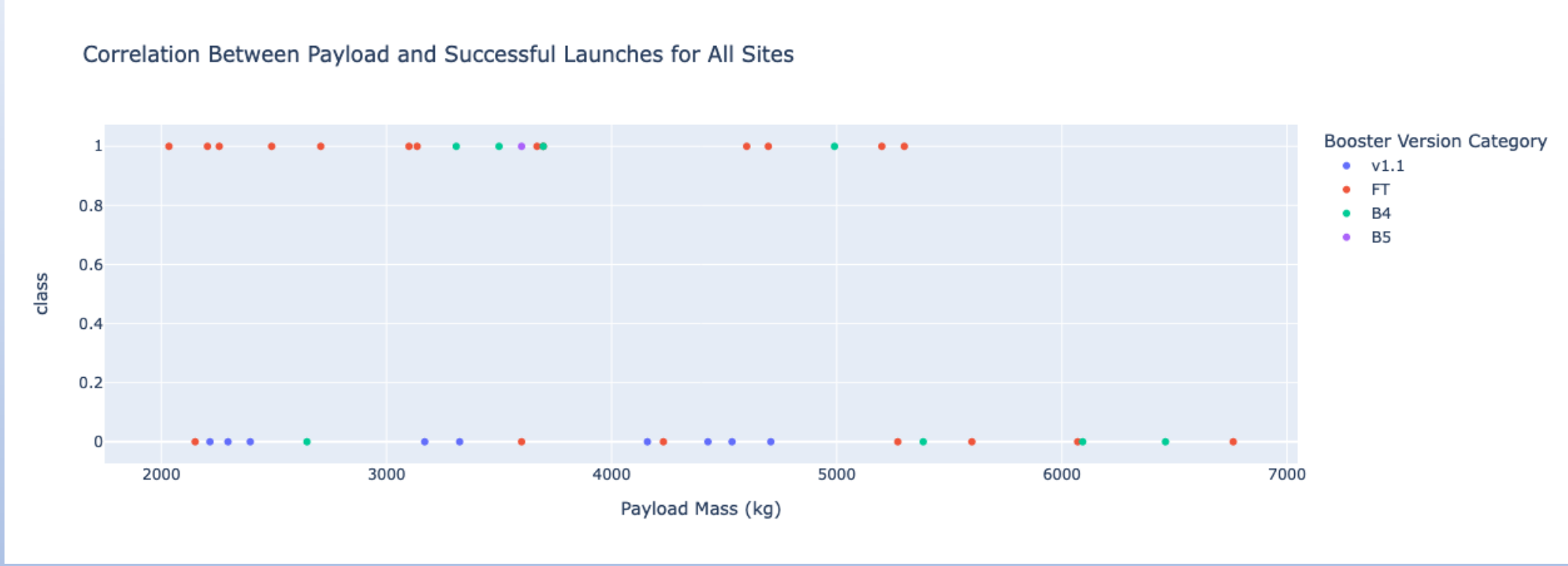
- The majority of successful launches were from the KSC LC-39A site

Total Successful Launches for Site KSC LC-39



- The majority of launches from site KSC LC-39A were successful (76.9%)

Payload vs Success Rate for Payload between 2000 and 7000 kg



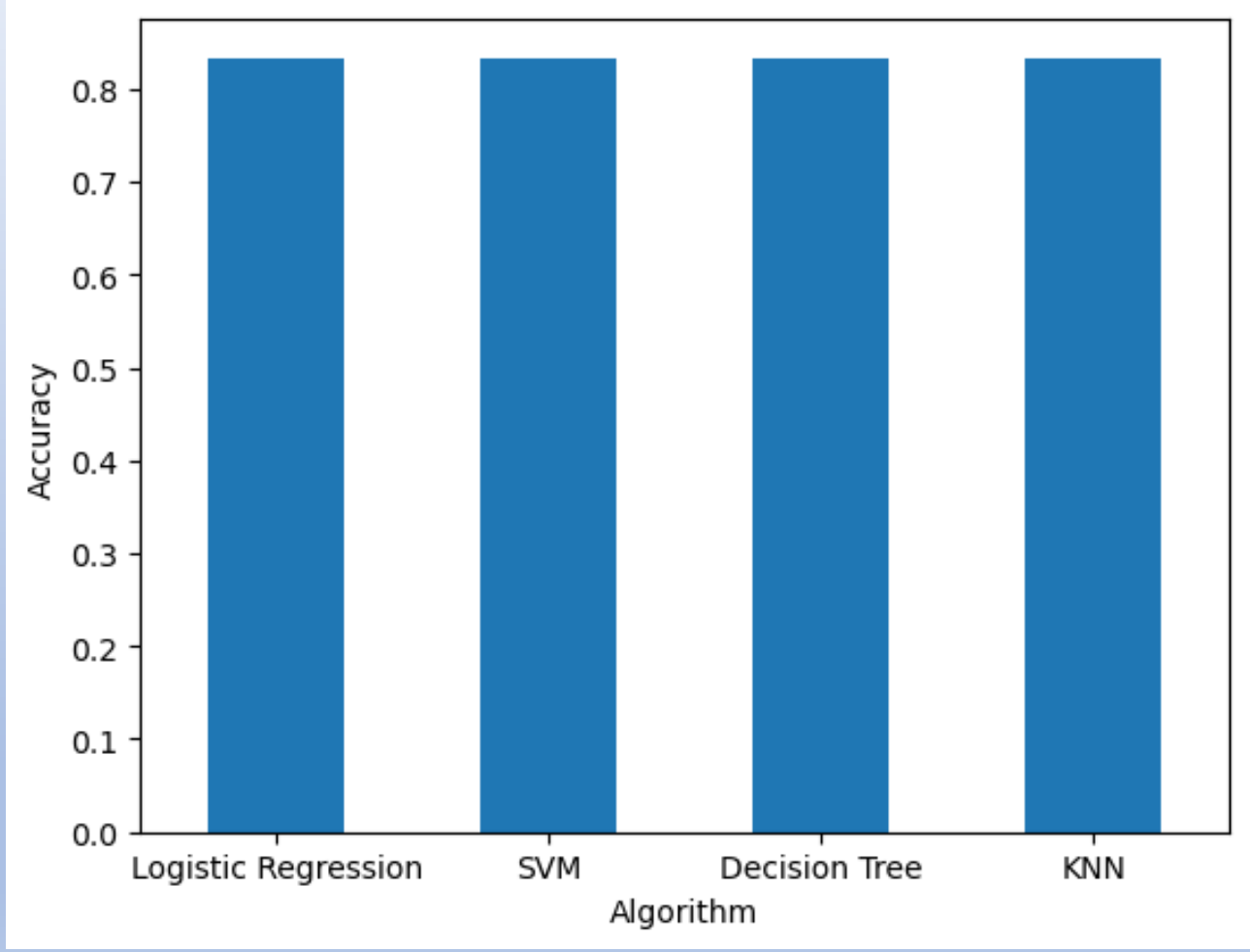
- Booster v1.1 had the lowest success rate
- Booster FT had the highest success rate when payload mass < 5300 kg but not between 3800 and 4300 kg
- Success rate for all boosters with payload mass between 5300 and 7000 kg was 0.

SPACE Y

Section 5

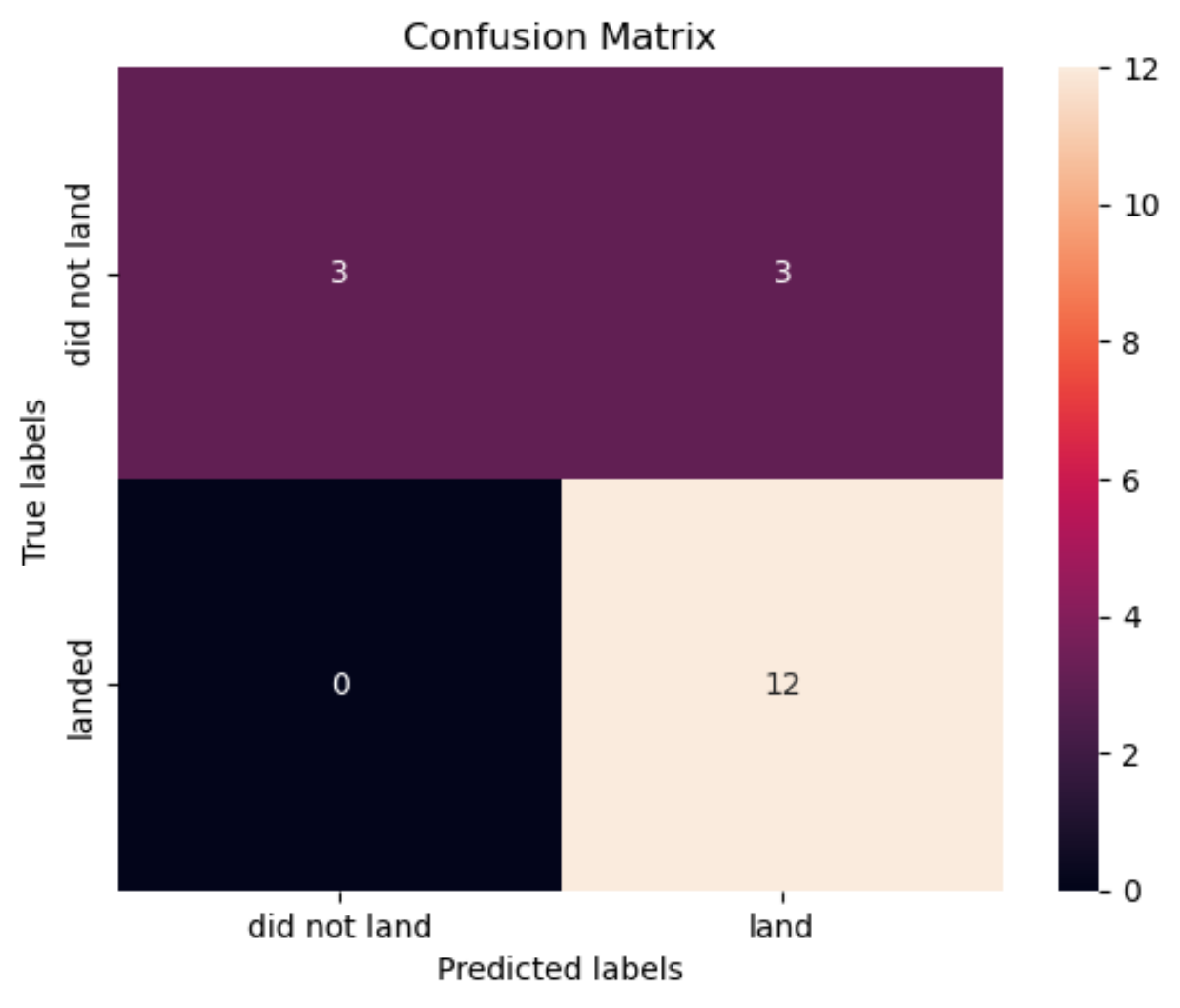
Predictive Analysis (Classification)

Classification Accuracy



- 4 algorithms were trained, tested and evaluated
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbour (KNN)
- All 4 algorithms had the same accuracy (83.33%)

Confusion Matrix



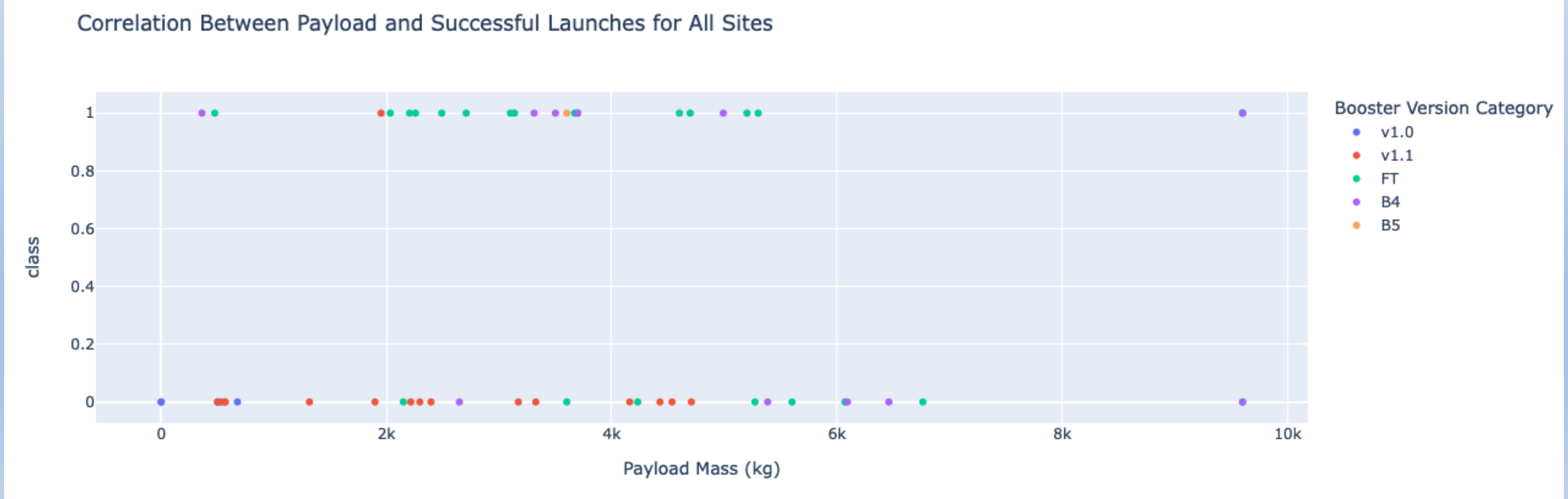
- Confusion matrix for all 4 algorithms was the same
- All algorithms predicted successful landings well, but were less accurate at predicting unsuccessful landings

Conclusions

- Purely from visual EDA, the following characteristics appear to be more likely to result in a successful launch:
  - From the KSC LC-39A launch site
  - Using VLEO orbit
  - Using payload mass of ~15,000 kg
- Machine learning should be used to predict success rate as it can take into account multiple variables
- All 4 machine learning algorithms tested (LR, SVM, Decision Tree, KNN) had the same accuracy
- All 4 algorithms can be used to predict if the launch will be successful, however, they may predict some false positives
- If we can maximize the success rate, we can reduce the cost of launches and make more competitive bids

Appendix

- GitHub repository: https://github.com/ictl/IBM-DS0720EN-Project
- Payload Mass vs Success Rate for all sites and payload mass between 0 and 10,000 kg:



Thank you!