IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Data Science Capstone Project

Bahareh Ahmadzadeh 07/14/2023

Outline

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

Executive Summary

- Data gathered from the SpaceX Wikipedia page and open SpaceX API. Labels column 'class' was created to categorise successful landings. used SQL, visualisation, folium maps, and dashboards to explore the data. compiled pertinent columns for use as features. used a single hot encoding to convert all categorical variables to binary. GridSearchCV was used to discover the ideal parameters for machine learning models using standardised data. Display the accuracy rating for each model.
- Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbours are the four machine learning models that were created. All gave identical results, with an average accuracy percentage of 83.33%. Successful landings were anticipated by all models. For improved model determination and accuracy, more data is required.

Introduction

Project background and context:

- It's the Commercial Space Age now.
- The most affordable option is Space X ($62 million vs. $165 million USD).
- mostly because of the rocket's ability to be recovered (Stage 1)
- Space Y aspires to rival Space X.

Problems you want to find answers:

- We have been given the task by Space Y to develop a machine learning model that can forecast successful Stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was compiled using the SpaceX Wikipedia page and the public SpaceX API.
- Perform data wrangling
  - Perform data manipulation by identifying successful and unsuccessful true landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using classification models, perform predictive analysis: tuned models with GridSearchCV.

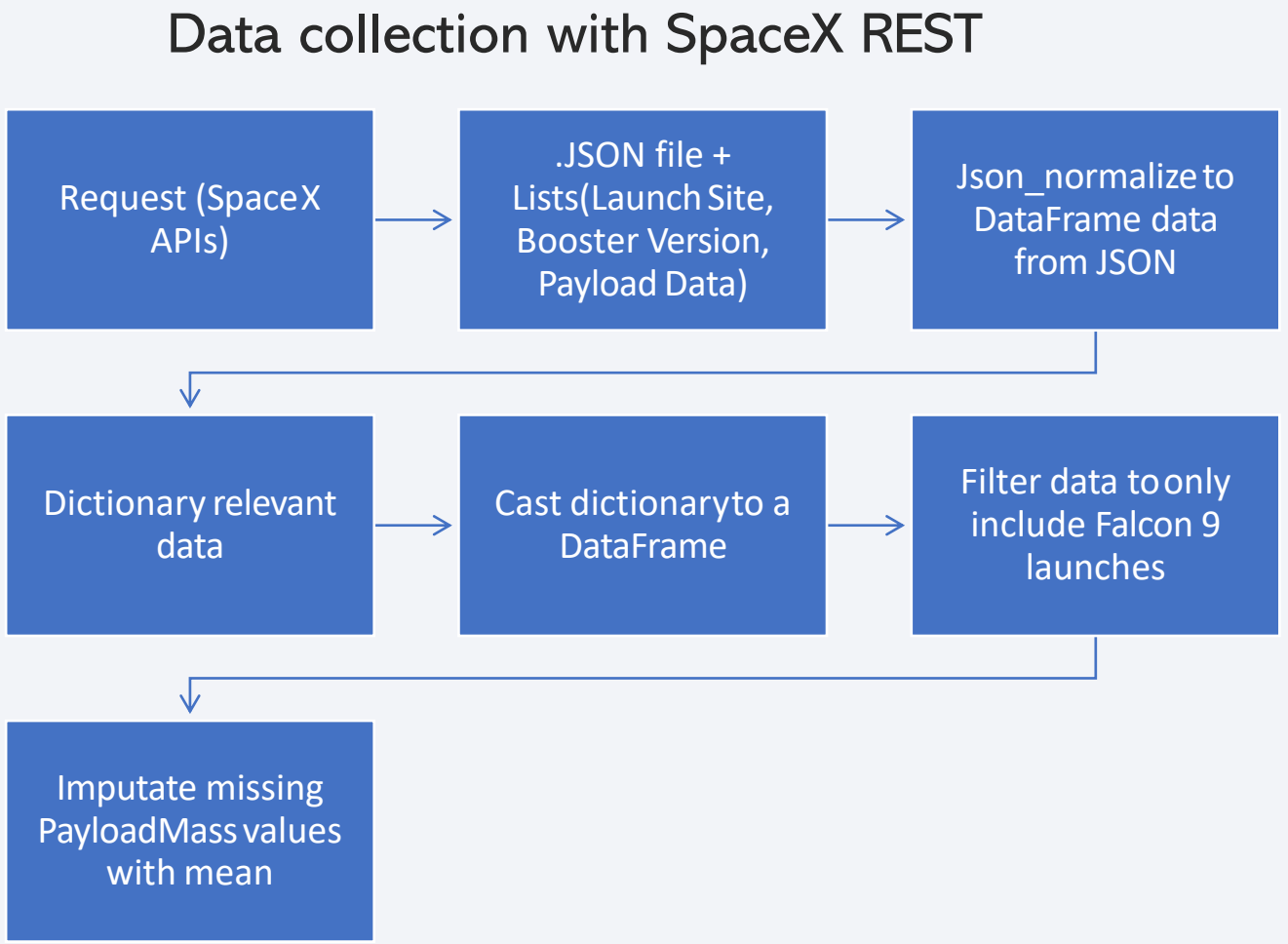
Data Collection

- A combination of API queries from Space X's public API and web scraping data from a table in Space X's Wikipedia entry were used in the data collection procedure.
- The flowchart for data collection from an API is shown on the following slide, and the flowchart for data collection via web scraping is shown on the slide after that.
- Wikipedia Webscrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- Space X API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data Collection – SpaceX API

GitHub URL of the completed SpaceX API:

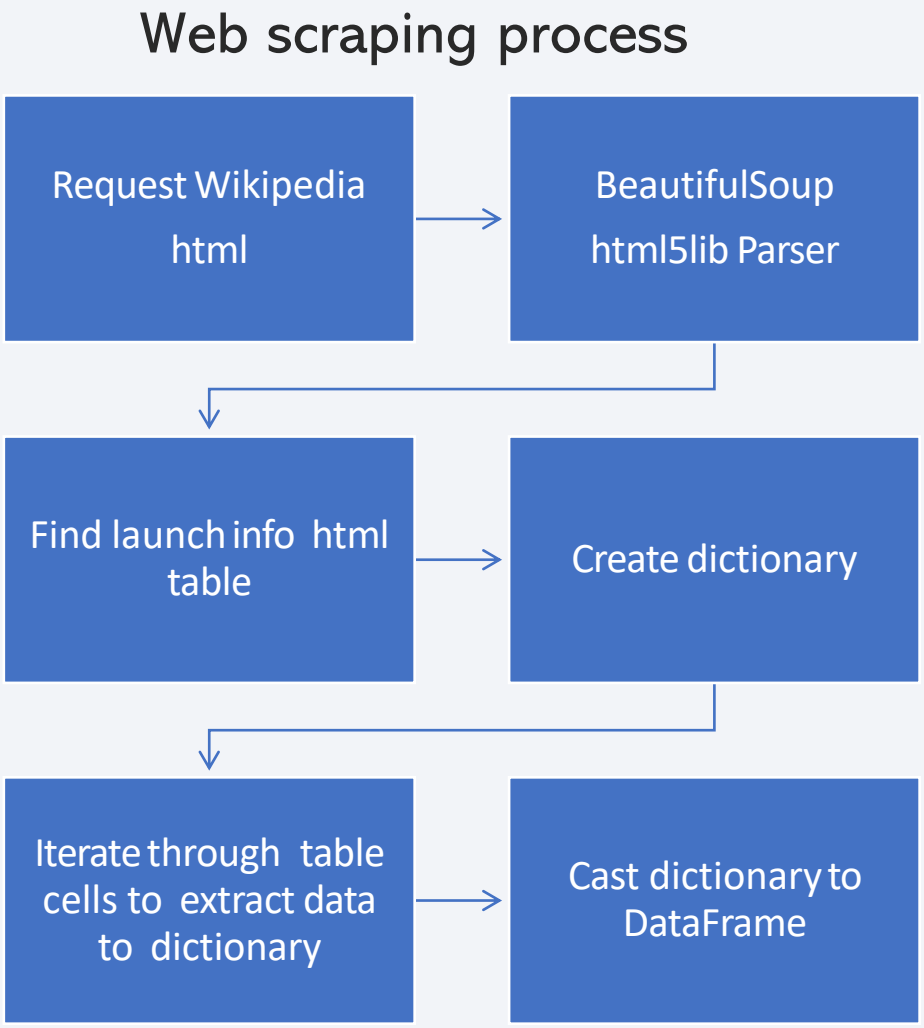
https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week1/Data%20Collection%20API.ipynb



Data Collection - Scraping

GitHub URL of the completed web scraping notebook:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week1/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Training label with landing outcomes: successful = 1 , failure = 0.
- Outcome column: ‘Mission Outcome’, ‘Landing Location’
- New training label ‘class’: 1 if ‘Mission Outcome’ is True and 0 otherwise.
- 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/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week1/Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.
- In order to determine whether a relationship between two variables exists so that it may be used in training the machine learning model, relationships between variables were compared using scatter plots, line charts, and bar plots.
- Plot for 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

GitHub URL:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week2/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- loaded IBM DB2 Database with a data set.
- Using SQL Python integration to query.
- To understand the dataset better, queries were run.
- Asked for details on launch site names, mission results, different customer and booster payload amounts, and landing results.
- GitHub URL:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch sites, successful and unsuccessful landings, and examples of important destinations that are close by are marked on folium maps, including railways, highways, coasts, and cities.
- This enables us to comprehend potential reasons for the positioning of launch sites. visualizes successful landings in relation to their location as well.
- GitHub URL:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

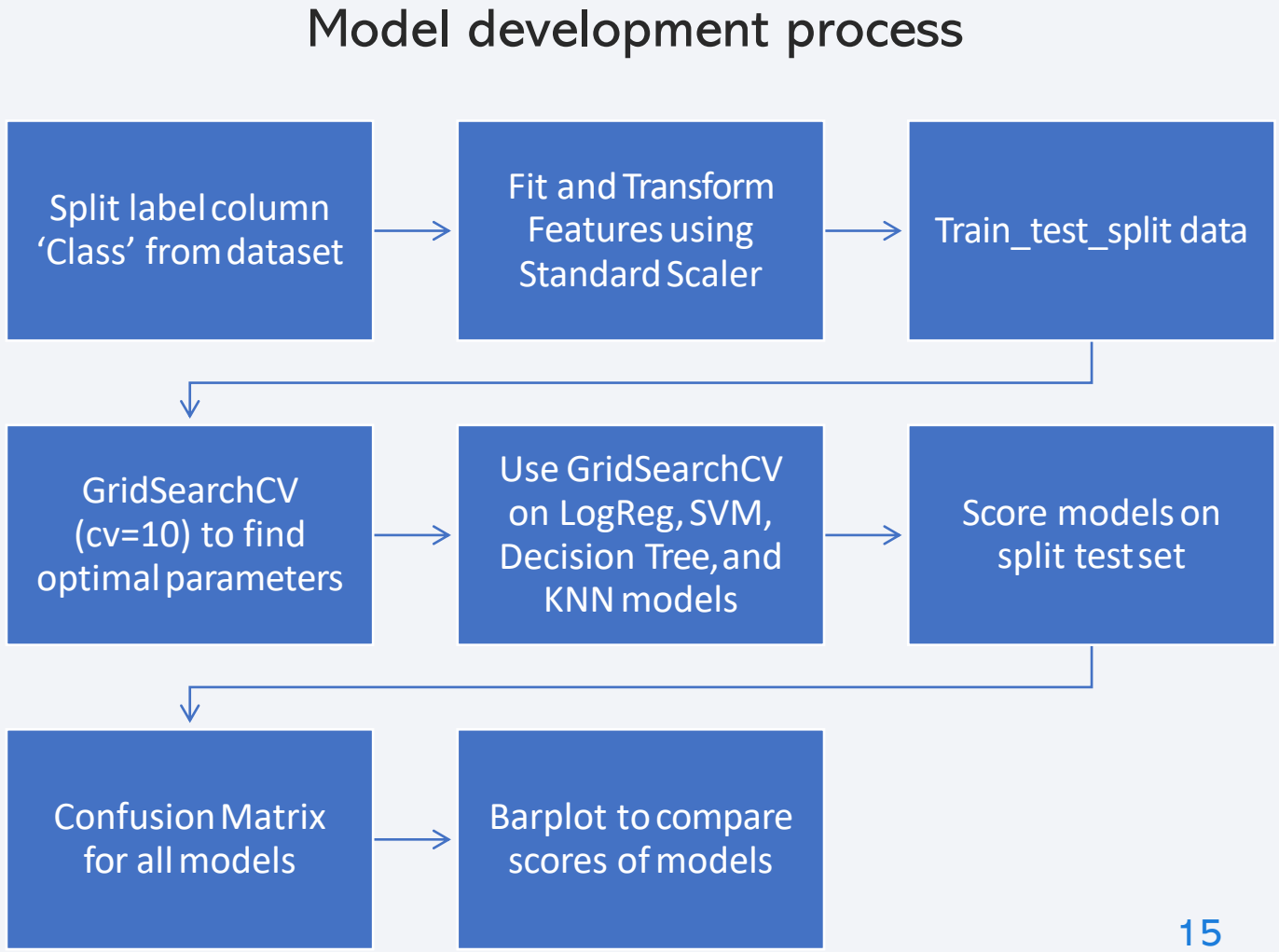
- Dashboard has a scatter plot and a pie chart.
- Pie charts can be chosen to display the distribution of successful landings among all launch locations as well as the success rates of individual launch sites.
- The payload mass on a slider between 0 and 10,000 kg, and either all sites or a specific site, are the two inputs for the scatter plot.
- The success rate of the launch site is displayed via a pie chart.
- We can examine how success varies among launch sites, payload tonnage, and booster version categories using the scatter plot.
- GitHub URL:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week3/Dashboard.ipynb

Predictive Analysis (Classification)

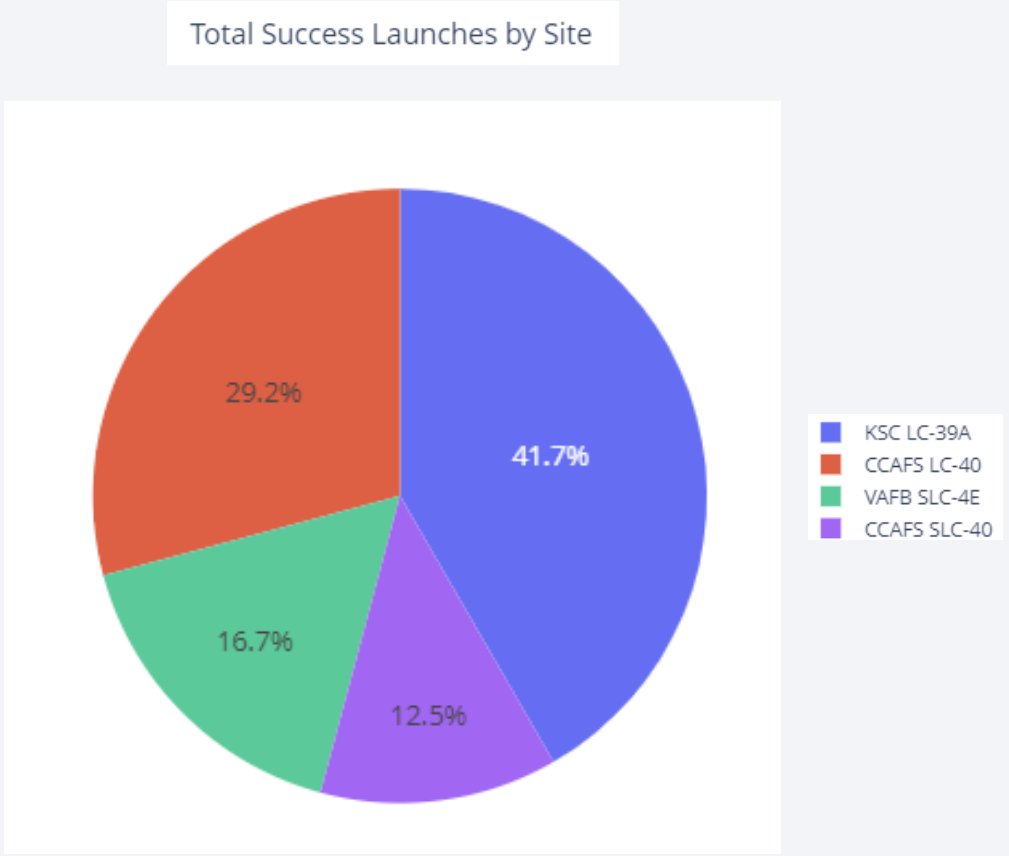
- GitHub URL:

https://github.com/BaharehAhz/IBM-Capston-Gitfiles/blob/master/week4/Machine%20Learning%20Prediction.ipynb



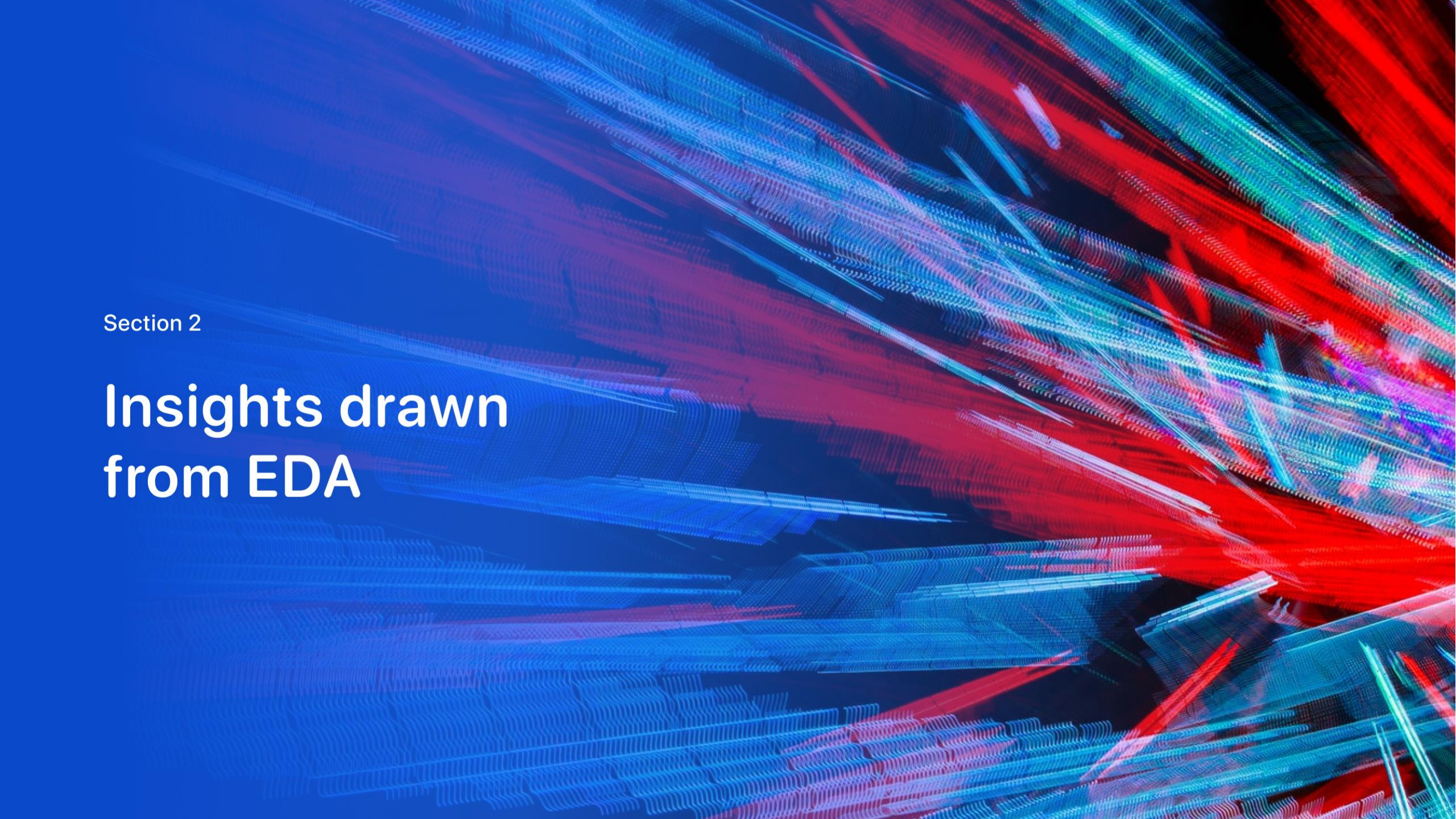
Results

- The Plotly dashboard preview is shown here. The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the outcomes of our model with an approximate 83% accuracy are displayed on the following sides.
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low-weighted payloads perform better than heavier payloads.
- The success rates for SpaceX launches are directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, and ES L1 has the best Success Rate.



Section 2

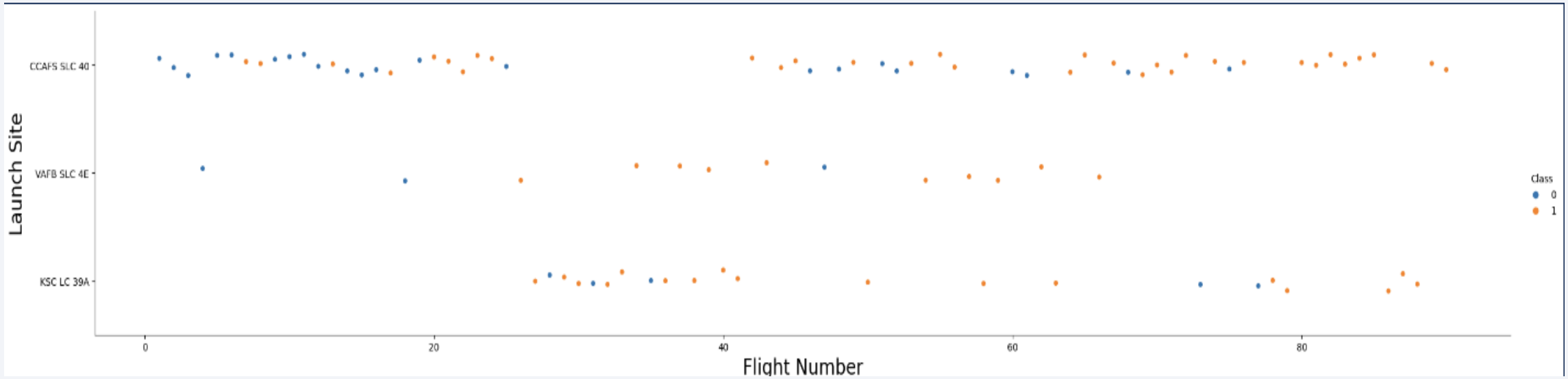
Insights drawn from EDA



Flight Number vs. Launch Site

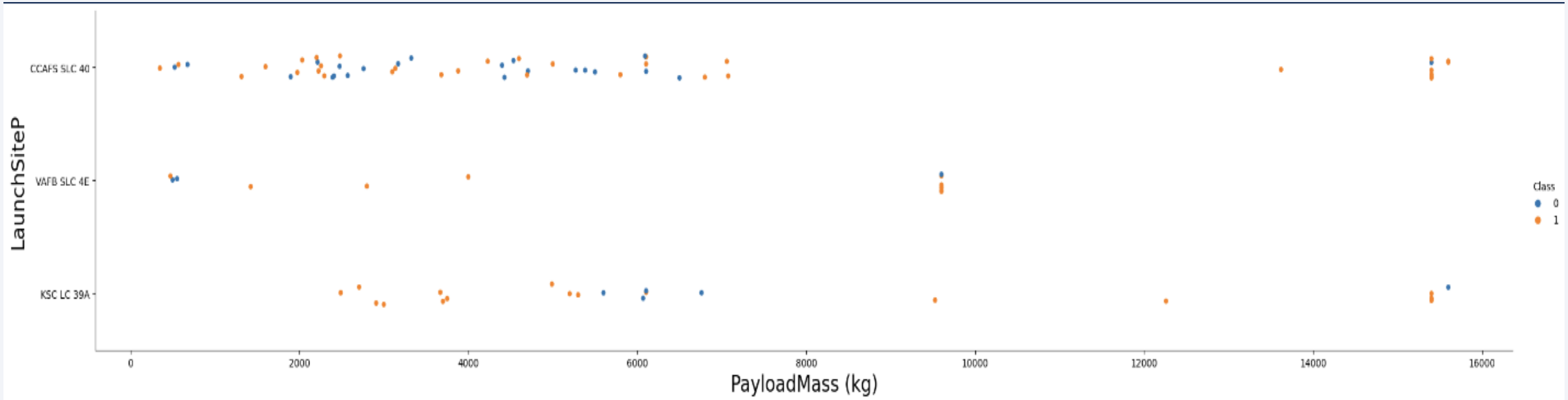
- Orange points denote successful launch and blues indicates failed launch.
- Graphic demonstrates an increase in success rate over time (shown in Flight Number).

There was probably a huge advancement around Flight 20, which greatly raised the success rate. CCAFS looks to be the main launch point as it has the largest volume.



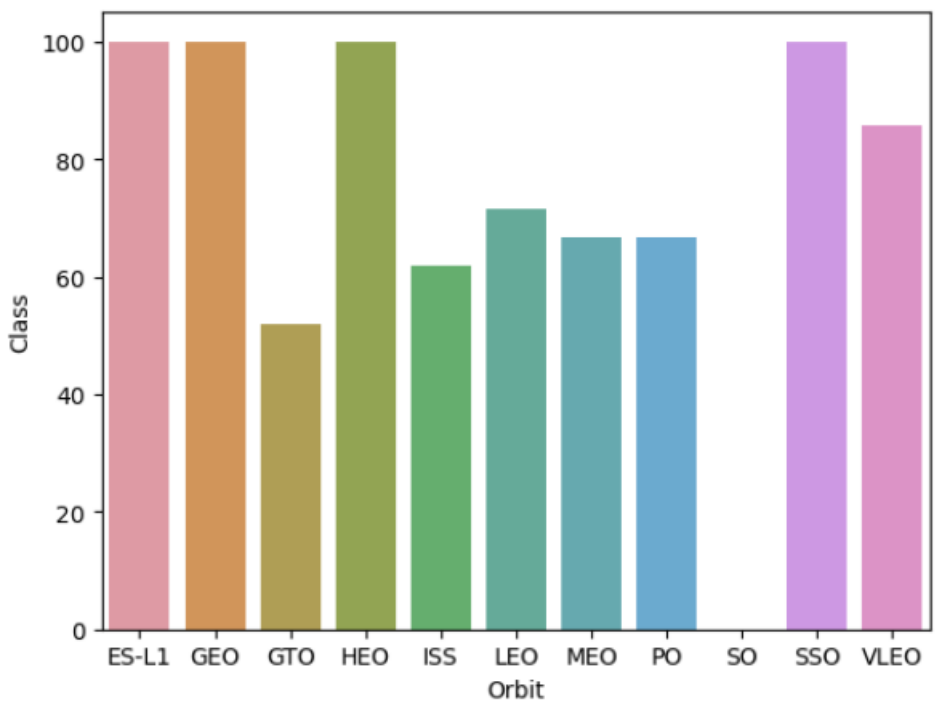
Payload vs. Launch Site

- Orange points denote successful launch and blues indicates failed launch.
- Payload mass appears to fall largely between 0-6000 kg. varying launch sites also seem to use varying payload mass.



Success Rate vs. Orbit Type

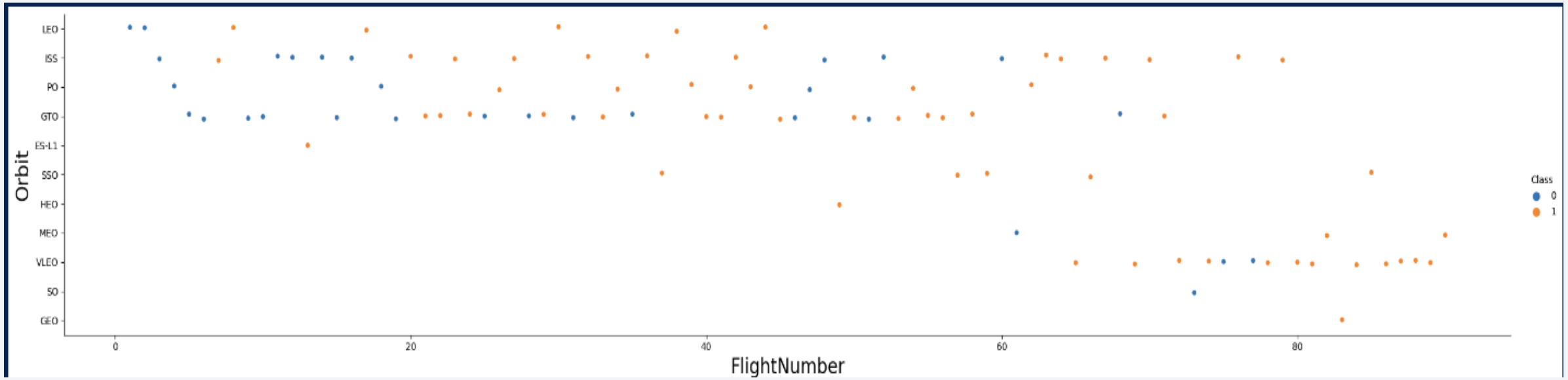
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- SSO (5) has a 100% success rate
- VLEO (14) has a decent success rate and attempts
- SO (1) has a 0% success rate
- GTO (27) has around 50% success rate but the largest sample



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

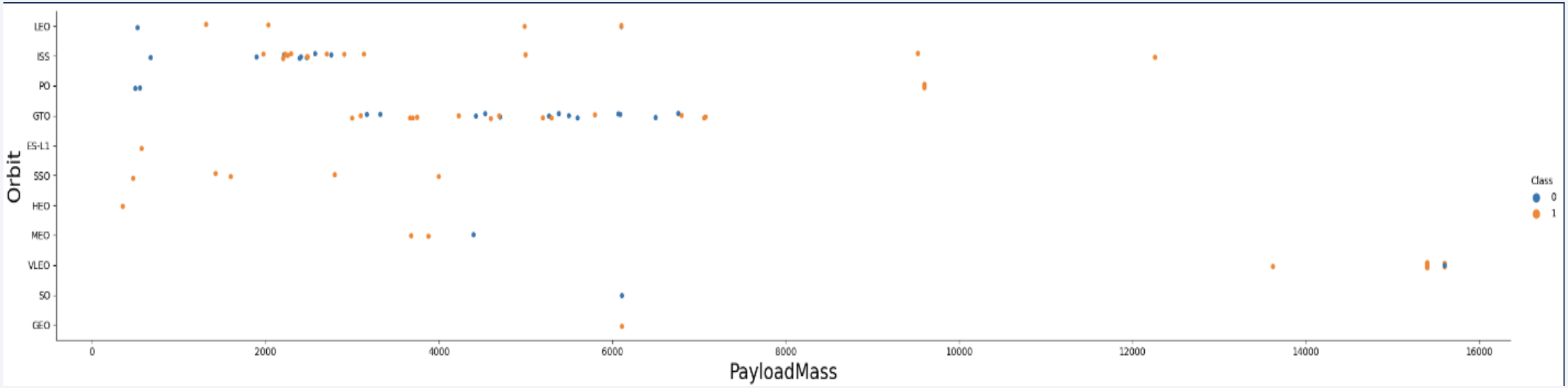
Flight Number vs. Orbit Type

- Launches that are successful are shown in orange, whereas failures are shown in blue.
- Flight Number was preferred over Launch Orbit. This preference appears to be correlated with Launch Outcome.
- SpaceX started with LEO orbits which experienced considerable success In recent launches, LEO and VLEO were switched. It seems that SpaceX performs better in lower or Sun-synchronous orbits.



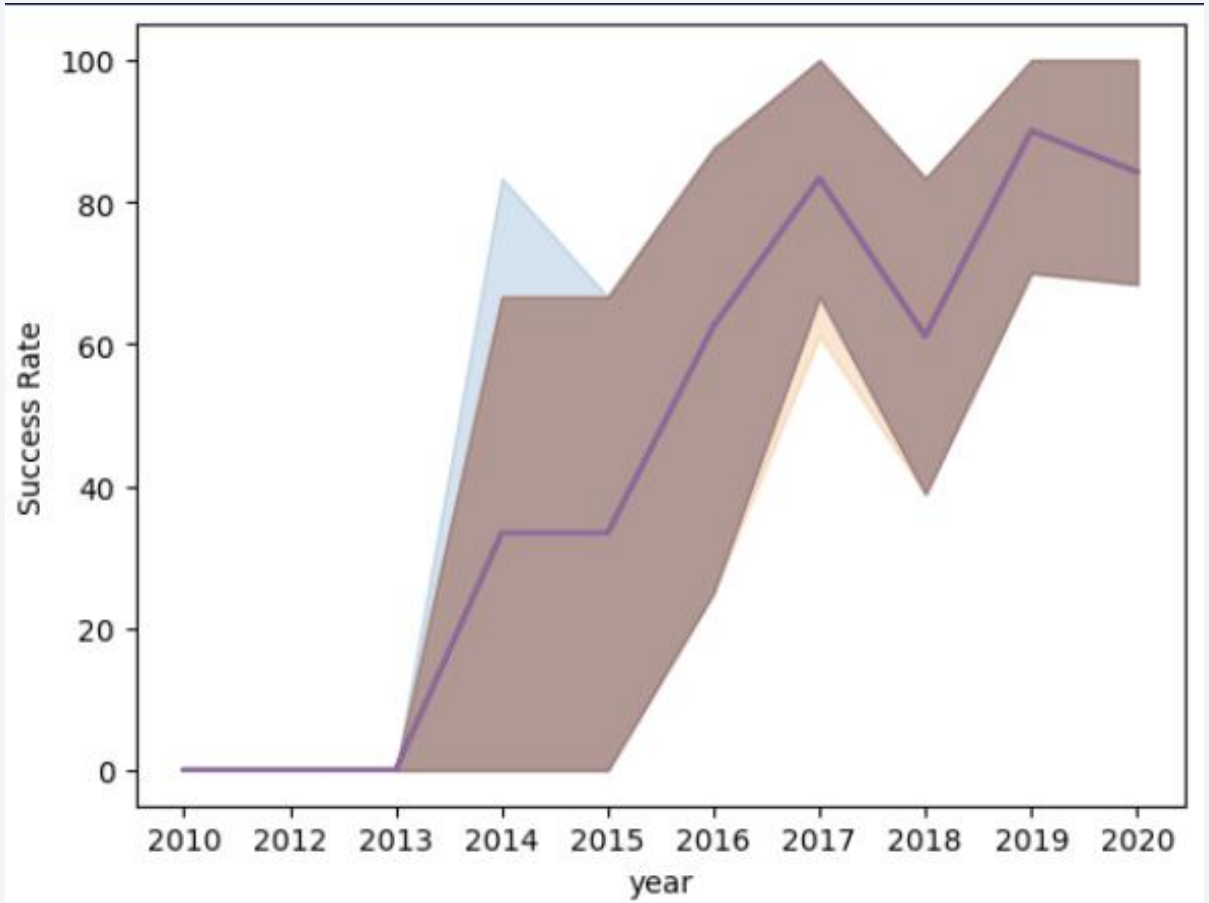
Payload vs. Orbit Type

- Launches that are successful are shown in orange, whereas failures are shown in blue.
- Payload mass and orbit appear to be correlated, with LEO and SSO having relatively modest payload masses. Only payload mass values at the upper end of the range are available for the other most successful orbit VLEO.



Launch Success Yearly Trend

- Since 2013, success has typically increased with a little decline in 2018. Success has been about 80% in recent years.



95% confidence interval (light blue shading)

All Launch Site Names

EDA with SQL: EXPLORATORY DATA ANALYSIS WITH SQL DB2 INTEGRATED IN PYTHON WITH SQLALCHEMY

Query unique launch site names from the database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
launch site with data entry errors.

CCAFS LC-40 was the previous

name. Likely only 3 unique

launch_site values: CCAFS SLC-

40, KSC LC-39A, VAFB SLC-4E

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

- The first five records in the database have Launch Site names that start with CCA.

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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

- When NASA was the customer, this query adds up the total payload mass in kilograms.
- These payloads were delivered to the International Space Station (ISS), which is known by the abbreviation CRS.

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- The average payload mass for launches using the booster version F9 v1.1 is determined by this query. The average payload mass of F9 v1.1 is at the low end of our payload mass range.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

First Successful Ground Landing Date

- The first successful ground pad landing date is returned by this query.
- It took till the end of 2015 for the initial ground pad landing.
- In general, successful landings start to occur in 2014.

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The four booster types with successful drone ship landings and a payload mass between 4,000 and 6,000 are returned by this search.

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- This search yields a count of each mission result. It seems like SpaceX completes its missions almost 99% of the time.
- This indicates that the majority of landing mishaps are deliberate. Interestingly, one launch's payload status is unknown, and regrettably, one launch failed in flight.

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

Boosters Carried Maximum Payload

- The results of this search are the booster iterations that could lift a payload of up to 15600 kg. These booster variants are all of the F9 B5 B10xx.x variety and are remarkably similar.
- This suggests that the payload mass and the booster design are related.

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

- The results of this search are the 2015 launches where stage 1 failed to land on a drone ship, along with the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site. There were two instances like this.

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

- A list of successful landings between 2010-06-04 and 2017, inclusive, are returned by this query.
- Drone ship landings and ground pad landings are the two different forms of successful landing results.
- There were a total of 8 successful landings during this time.

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

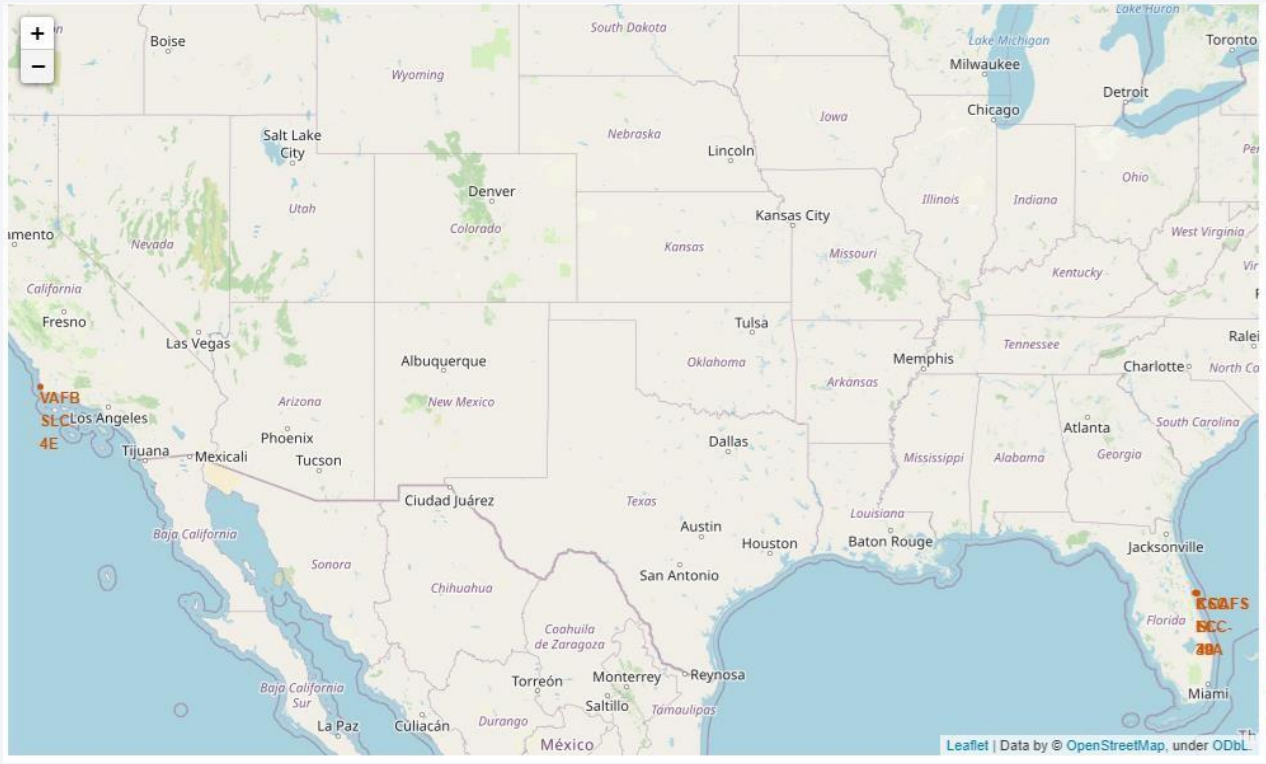
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

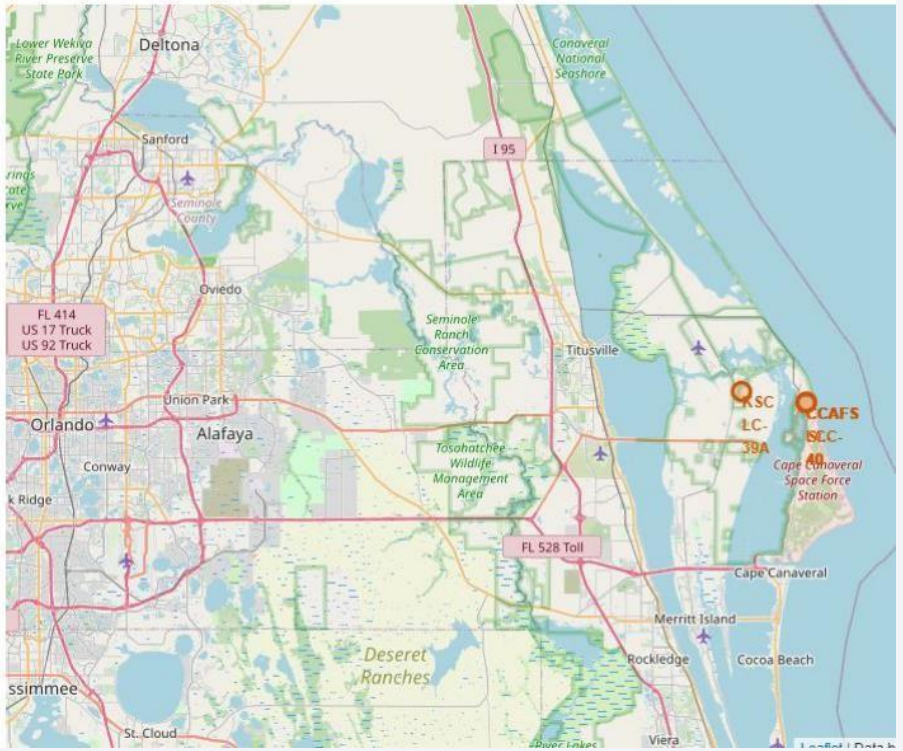
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

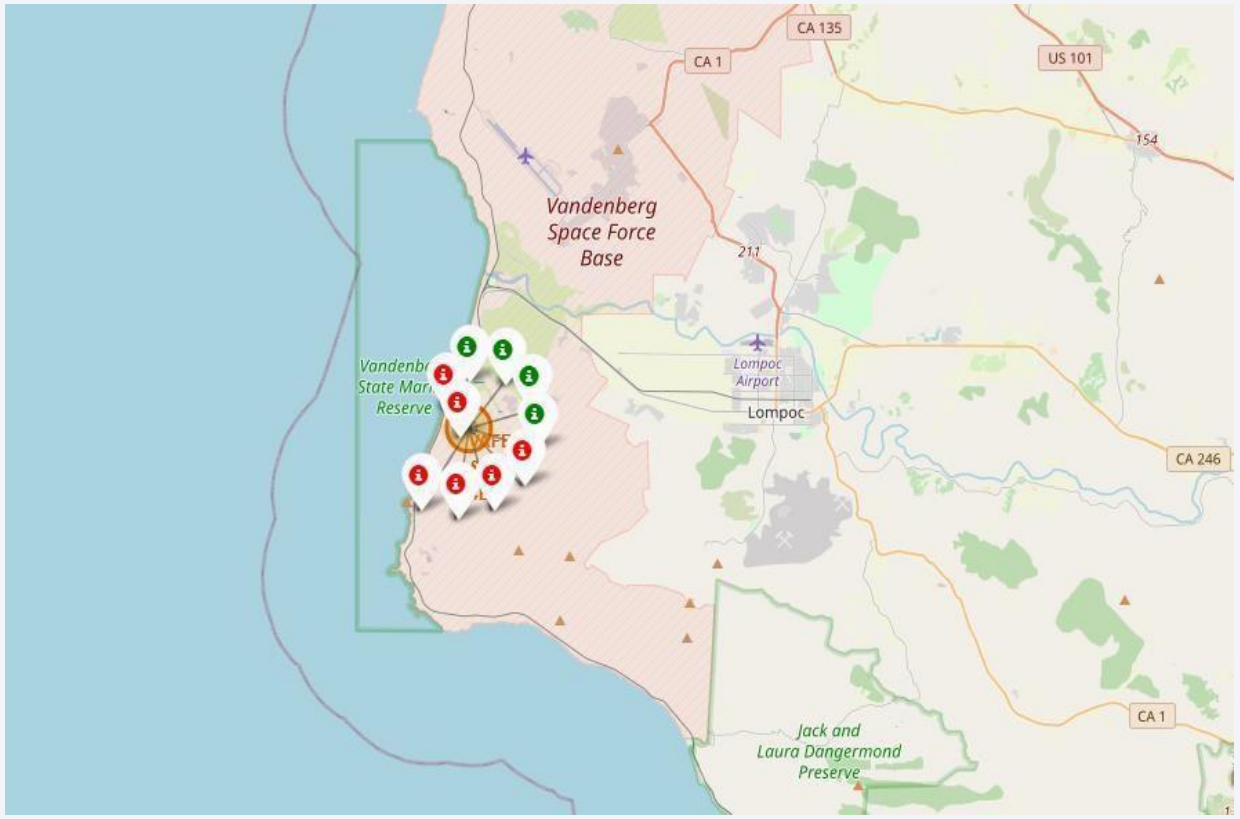
- The left map displays a relative US map with all launch sites. Due to their proximity, the two Florida launch sites are shown on the right map. Every launch place is close to the water.





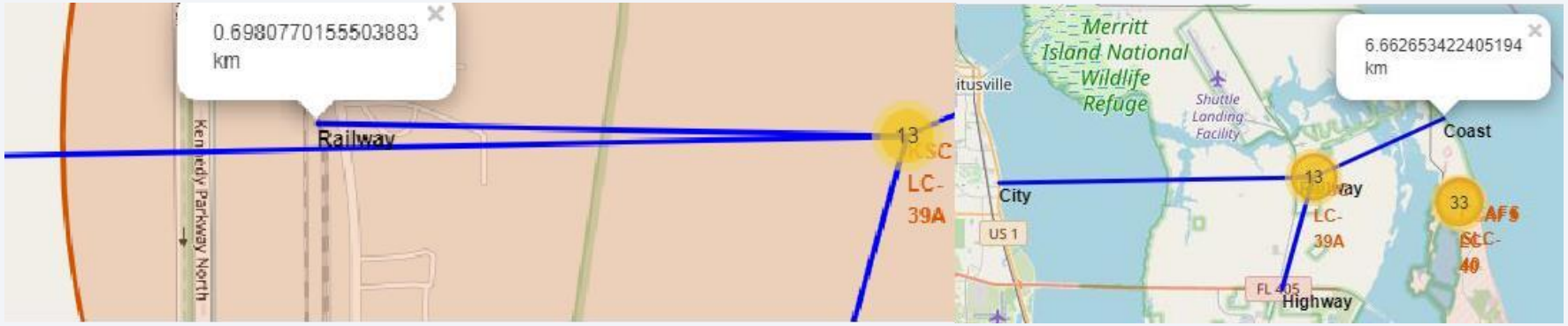
Color-Coded Launch Markers

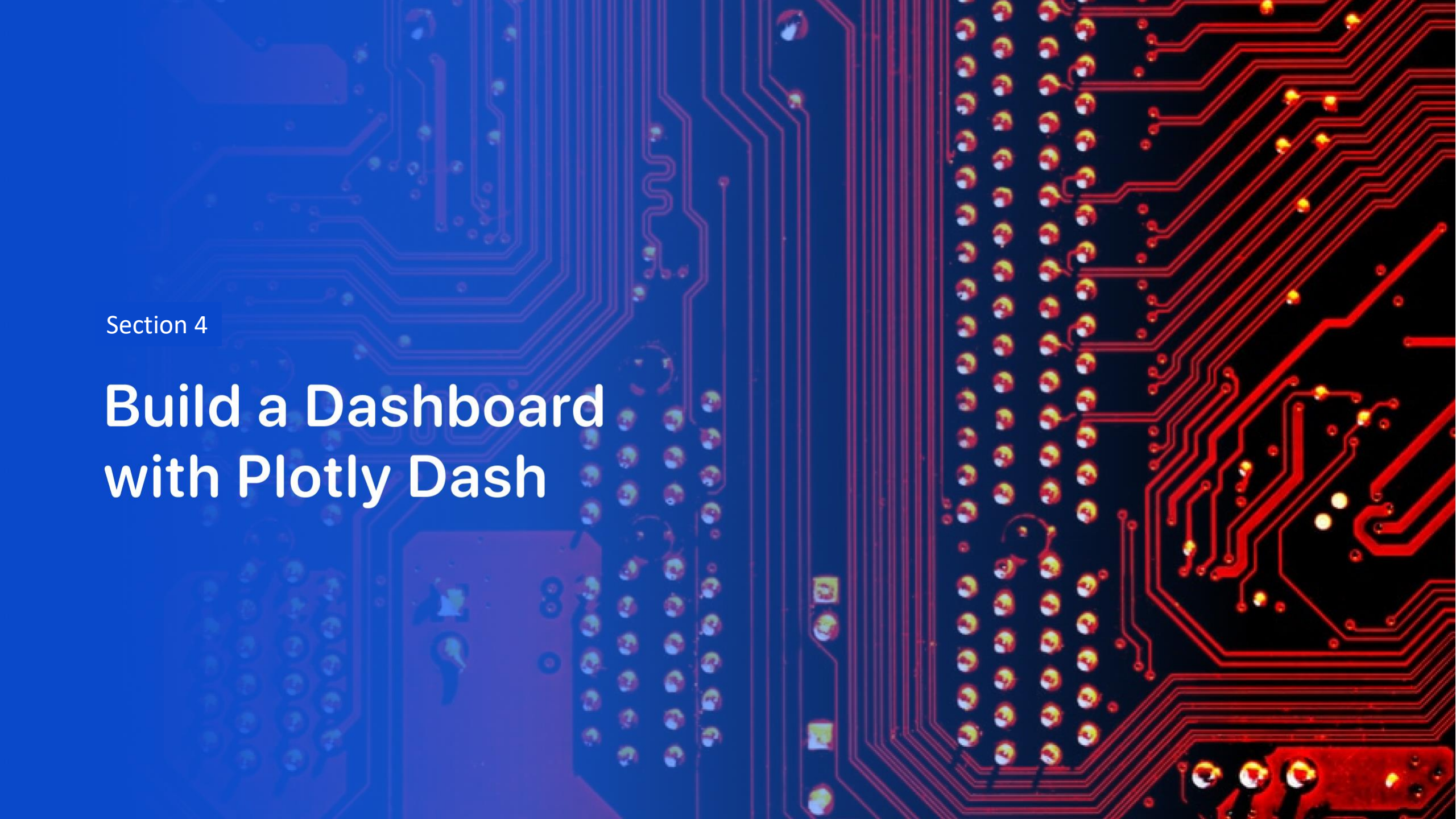
- Clicking on clusters on the Folium map will show each successful (green icon) and unsuccessful (red symbol) landing. VAFB SLC-4E has 4 successful landings and 6 unsuccessful landings in this instance.



Key Location Proximities

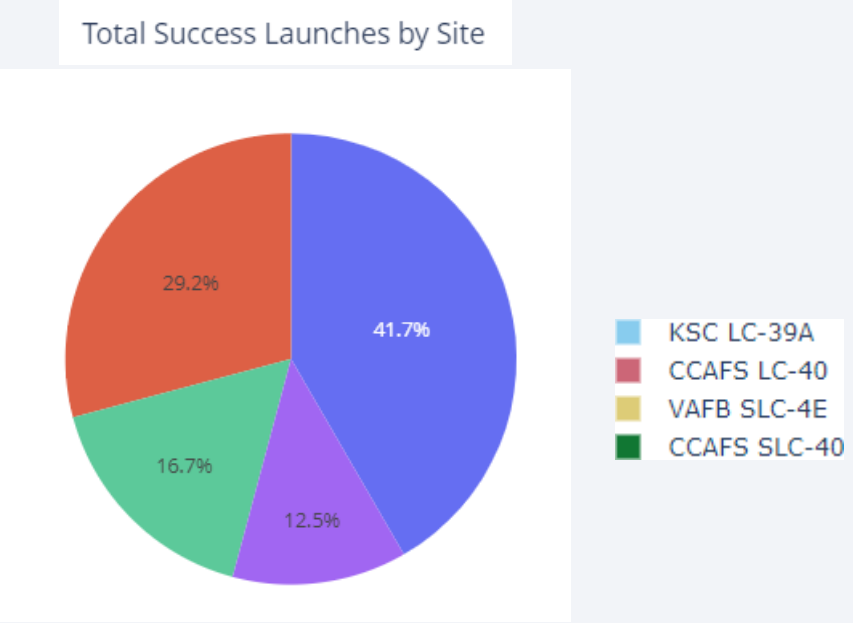
- For the most part and supply transportation, launch sites are located relatively close to railroads, using KSC LC-39A as an example. Highways for the transportation of people and supplies are around launch sites. In order to prevent rockets from falling on densely populated regions, launch locations are also located close to coasts and away from major towns.




Section 4
Build a Dashboard with Plotly Dash

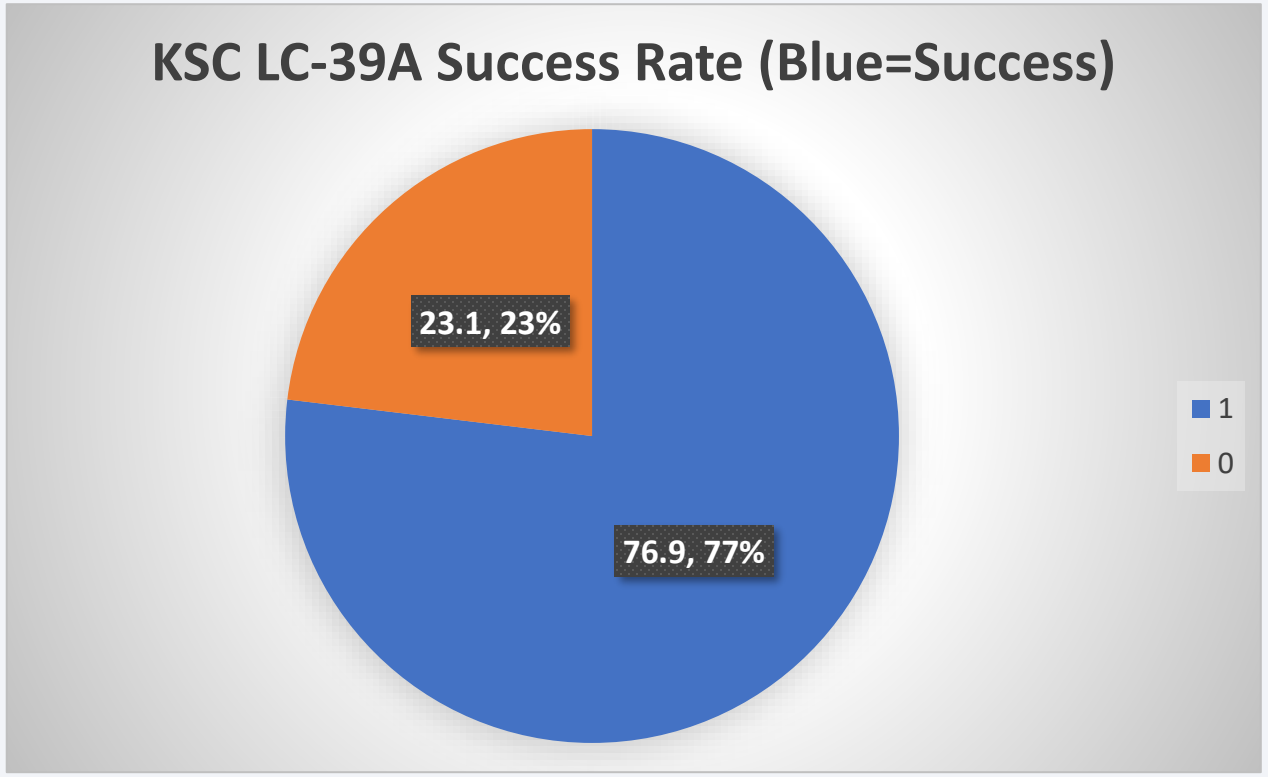
Successful Launches Across Launch Sites

- This graph shows how successful landings have been distributed throughout all launch sites. The number of successful landings for CCAFS and KSC is equal, however the majority of them took place before to the name change because CCAFS LC-40 was the previous name for CCAFS SLC-40. The least number of successful landings occur at VAFB. This might be because the sample size was lower and launching was more challenging on the west coast.



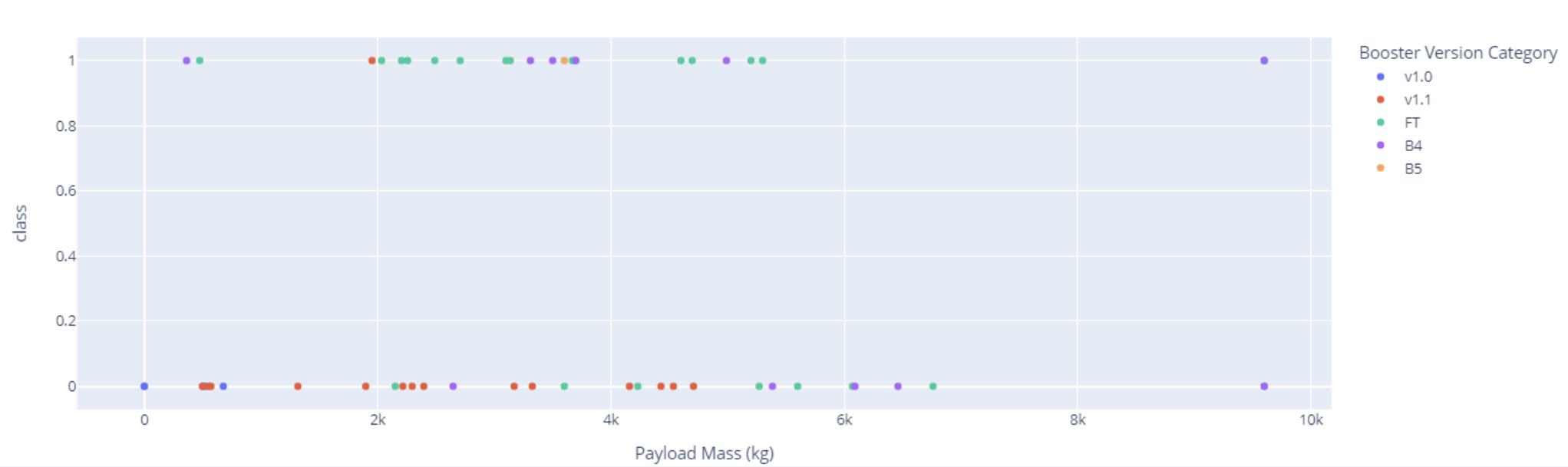
Highest Success Rate Launch Site

- The most successful landings were made by KSC LC-39A, with 10 successful attempts and 3 unsuccessful ones.



Payload Mass vs. Success vs. Booster Version Category

- Using the Payload range selector on the Plotly dashboard. However, instead of the maximum Payload of 15600, this is set from 0 to 10,000. Class displays 1 for a successful landing and 0 for an unsuccessful one. The booster version category in colour and the number of launches in point size are also taken into account by the scatter plot. It's noteworthy that there have been two failed landings in this specific range of 0-6000 with payloads of zero kilograms.

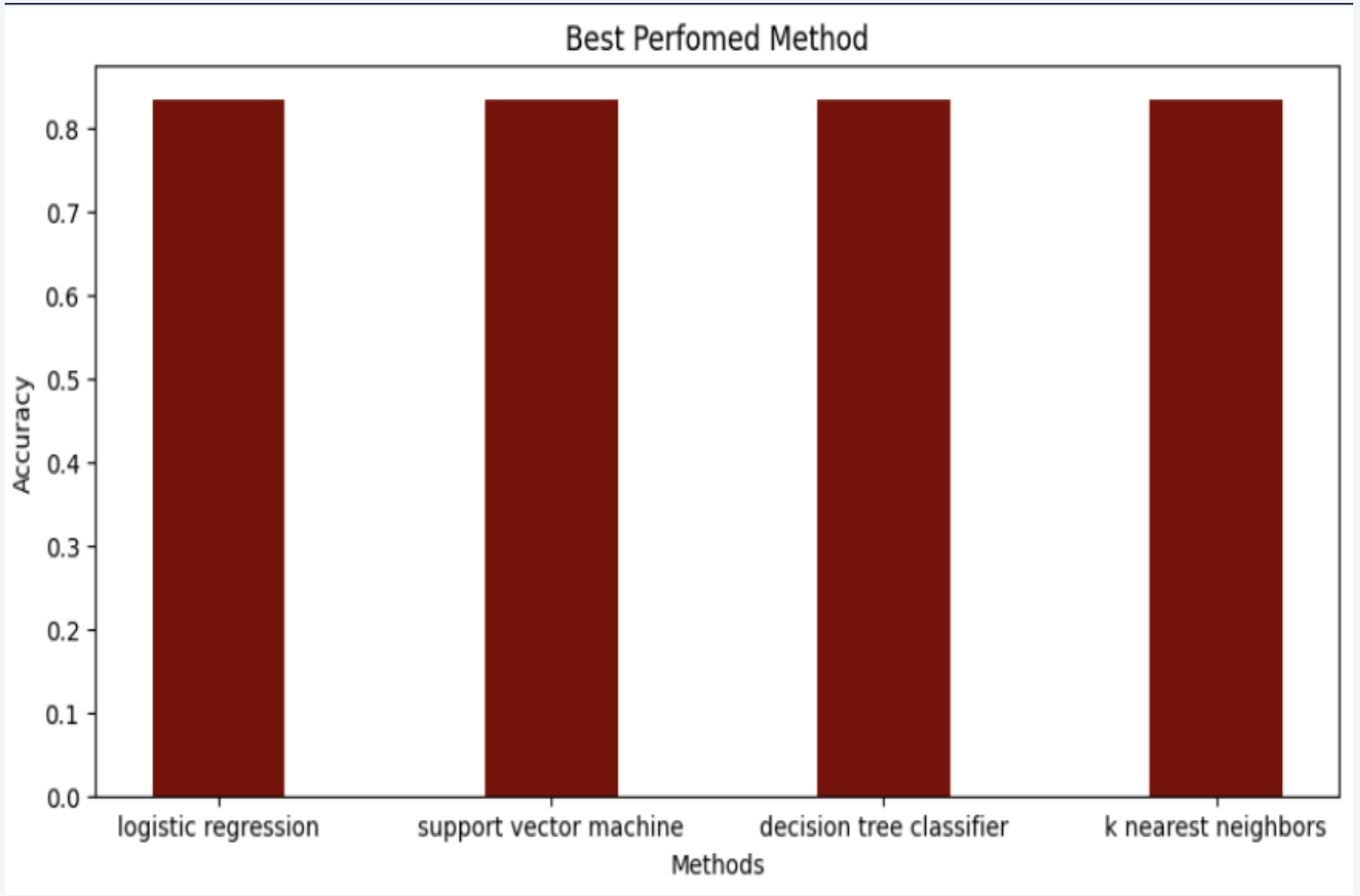


Section 5

Predictive Analysis (Classification)

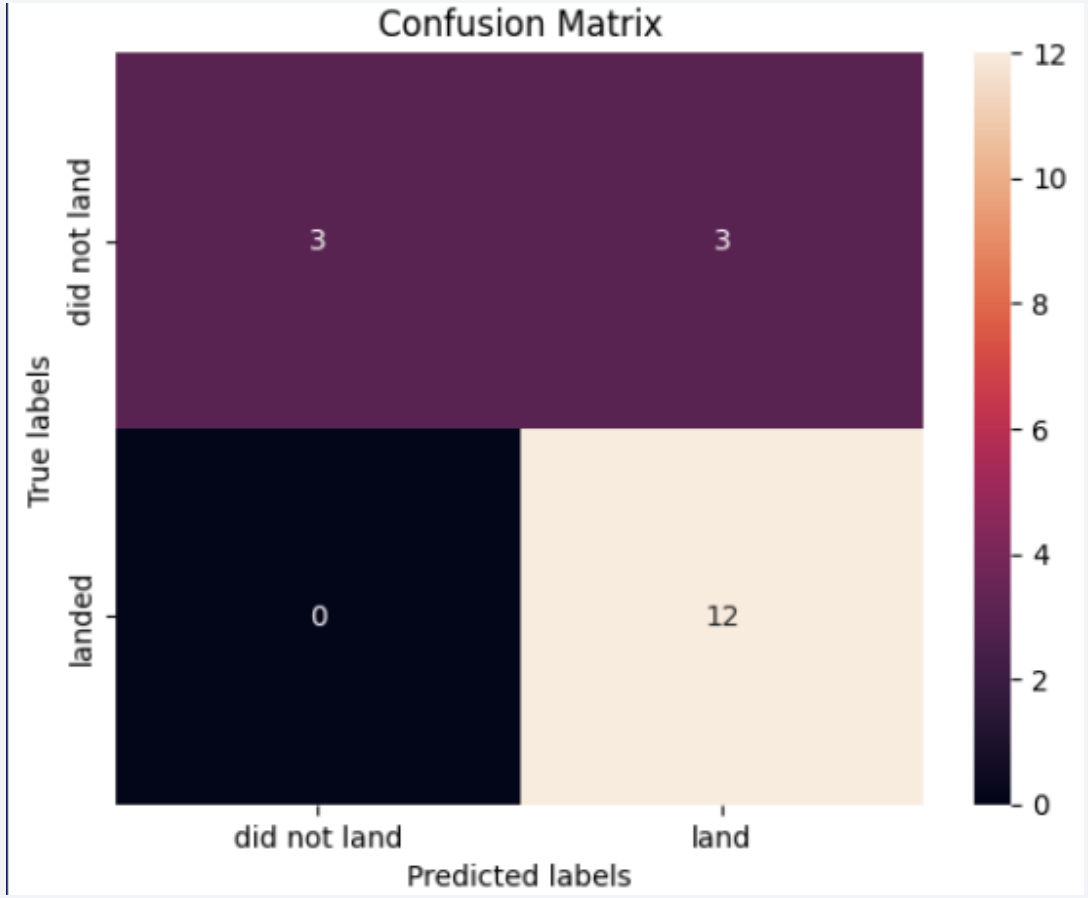
Classification Accuracy

- On the test set, all models' accuracy was 83.33% or essentially the same. It should be emphasized that the sample size of 18 is a modest test size.
- When a decision tree classifier model is used repeatedly, this can result in significant variance in accuracy outcomes.
- To choose the optimal model, we probably require more information.



Confusion Matrix

- The confusion matrix is the same for all Logistic regression, SVM, and KNN models because their performance on the test set was identical. When the actual label was successfully landing, the models projected 12 successful landings.
- When the actual label was failure landing, the models projected three unsuccessful landings.
- When the actual label was unsuccessful landings, the models incorrectly predicted three successful landings (false positives). Our forecasts overestimate the success of landings.



The diagonal of correct predictions run from top left to bottom right.

Conclusions

- Our job is to create a machine learning model that can forecast when Stage 1 will successfully land in order to save Space Y, which wants to compete against SpaceX, about $100 million USD.
- Used information from the SpaceX Wikipedia page and a public SpaceX API.
- Built a dashboard for visualization, created data labels, and added data to a DB2 SQL database.
- Our machine-learning model had an 83% accuracy rate.
- In order to decide whether or not to proceed with a launch, Allon Mask of SpaceY can use this model to forecast, with a fair amount of accuracy, if a launch will have a successful Stage 1 landing.
- To improve accuracy and choose the optimum machine learning model, more data should ideally be gathered.

Appendix

Special Thanks to All Instructors

My GitHub repository URL:

https://github.com/BaharehAhz

Thank you!