Data science capstone project

- Created by:

El Wafi Akram

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Acquired information from the public SpaceX API and SpaceX Wikipedia page. Established a 'class' column in the dataset to categorize successful landings. Examined the data through SQL queries, visualizations, folium maps, and dashboards. Selected pertinent columns as features for analysis. Employed one-hot encoding to convert categorical variables into binary format. Standardized the data and employed GridSearchCV to determine optimal parameters for machine learning models. Illustrated the accuracy scores of all models through visualization.

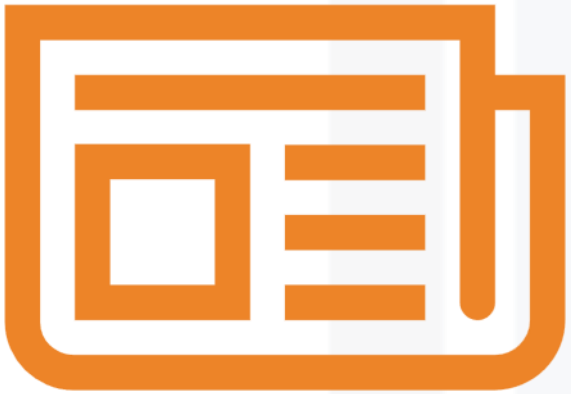
- Four machine learning models, namely Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All exhibited comparable outcomes, achieving an accuracy rate of approximately 83.33%. Notably, each model demonstrated a tendency to overpredict successful landings. It is suggested that acquiring additional data would contribute to refining model determination and enhancing overall accuracy.

INTRODUCTION

- Background:
  - Space X has best pricing ($62 million vs. $165 million USD)
  - Largely due to ability to recover part of rocket (Stage 1)
  - Space Y wants to compete with Space X

- Problem:
  - Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

METHODOLOGY



- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

METHODOLOGY: Data collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.
- **Space X API Data Columns:**
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
- Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- **Wikipedia Webscrape Data Columns:**
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

METHODOLOGY: Data collection

- There are numerous ways to collect data:

- **Web scraping**: Start off by requesting the wikipedia html and parse the response using bs4, then create a dictionary to extract the data, and finaly create a dataframe using the extracted data

- **SPACEXAPI:** It starts with a request to API of SPACEX which generates a json file which in itself is used to create a dataframe. From here we proceed the same way as the previous method and cast a dictionary to the dataframe while only selecting the column where Falcon9 is associated. This is data is treated in a way that it deals with missing values by putting a the mean value of the rows in those missing values.

METHODOLOGY: Data wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. Value Mapping:
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

METHODOLOGY: EDA Data visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

- **Plots Used:**
- 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
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

METHODOLOGY: EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

METHODOLOGY: Predictive analysis

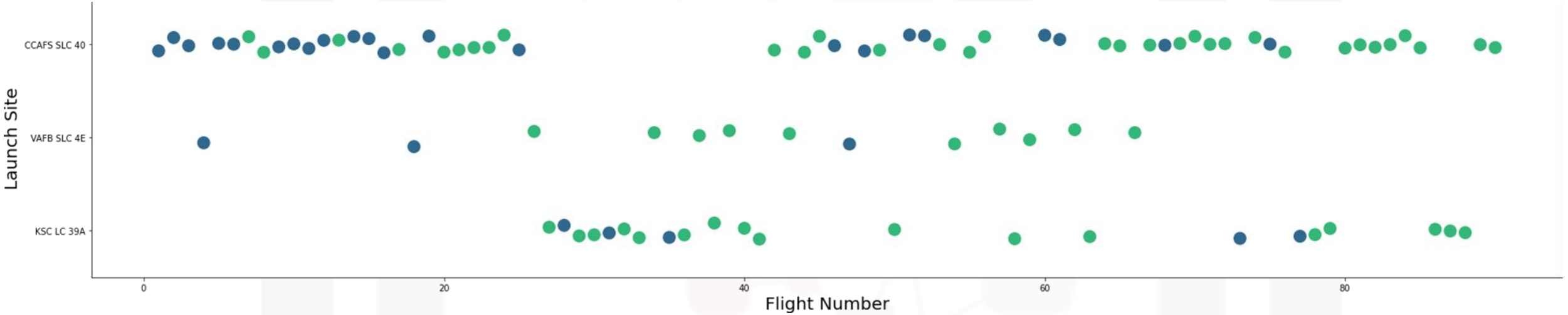
- Label a column "class" in the dataset
- Fit features using a standard scaler
- Train and test split the data
- Use a GRIDSEARCHCV to find optimal values
- SVM, DECISION TREE and KNN models
- Score models on split test set
- Create a confusion matrix
- Barplot to compare the models

METHODOLOGY: Predictive analysis

- Label a column "class" in the dataset
- Fit features using a standard scaler
- Train and test split the data
- Use a GRIDSEARCHCV to find optimal values
- SVM, DECISION TREE and KNN models
- Score models on split test set
- Create a confusion matrix
- Barplot to compare the models

RESULTS: EDA with visulization

Flight number & LaunchSite

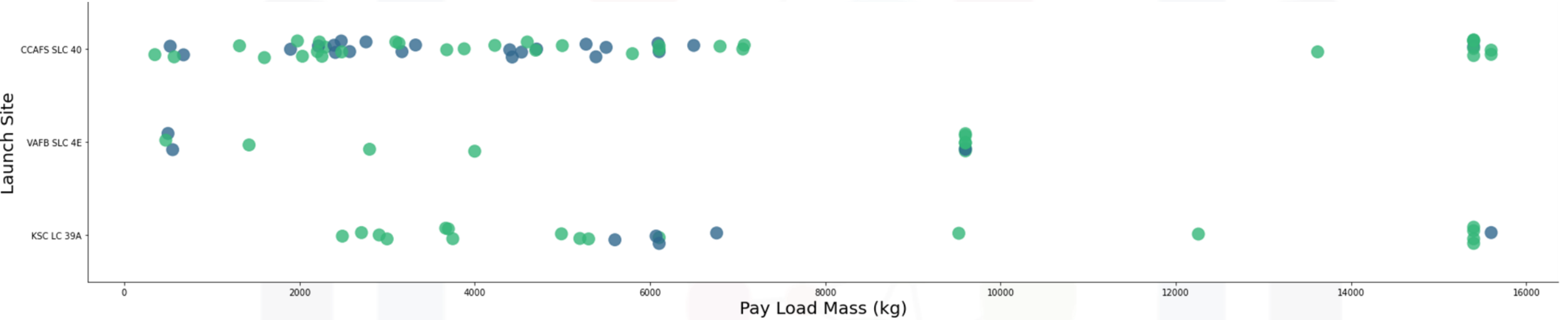


Interpretation:

- Graphic suggests an increase in success rate over time (indicated in Flight Number).
- Likely a big breakthrough around flight 20 which significantly increased success rate.
- CCAFS appears to be the main launch site as it has the most volume.

RESULTS: EDA with visulization

Payload & LaunchSite

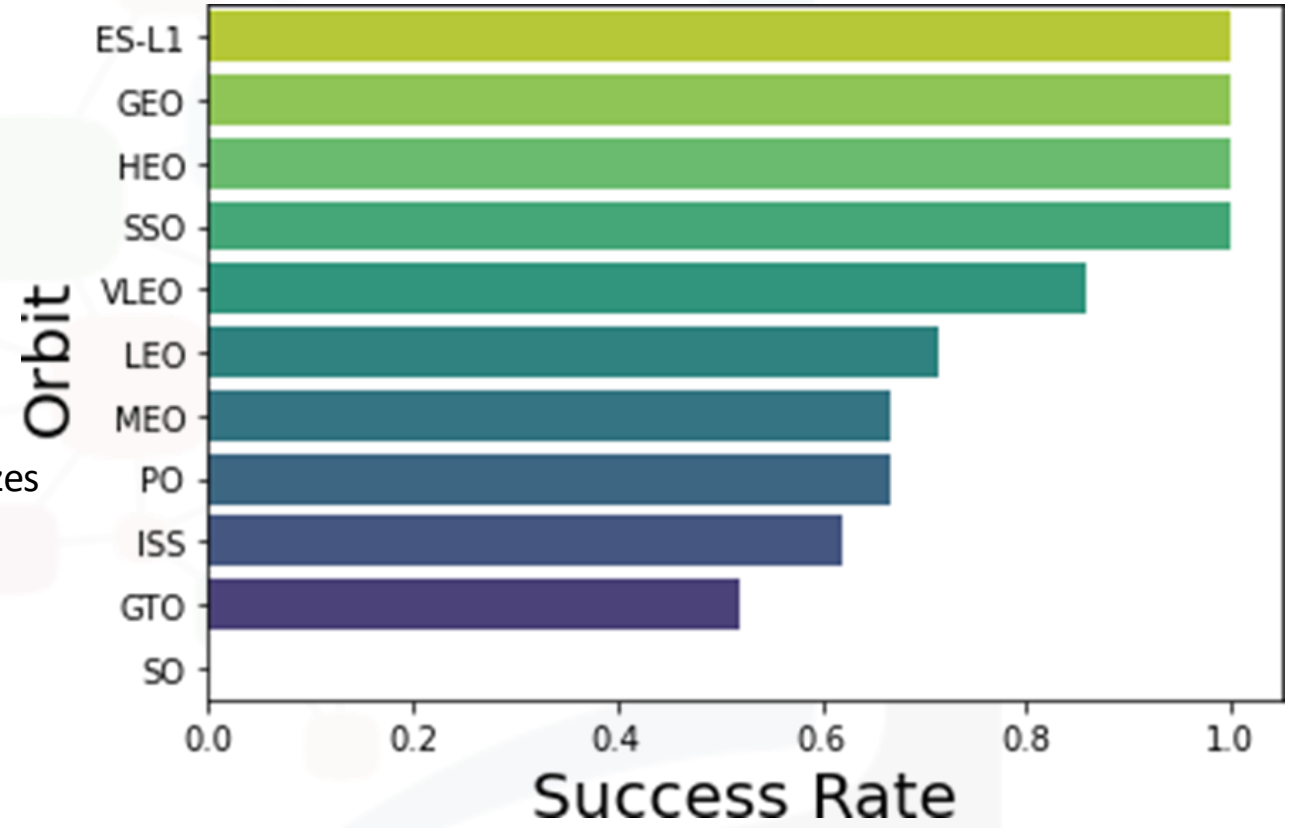


Interpretation:

- Payload mass appears to fall mostly between 0-6000 kg.
- Different launch sites also seem to use different payload mass

RESULTS: EDA with visulization

Sucess rate & Orbit

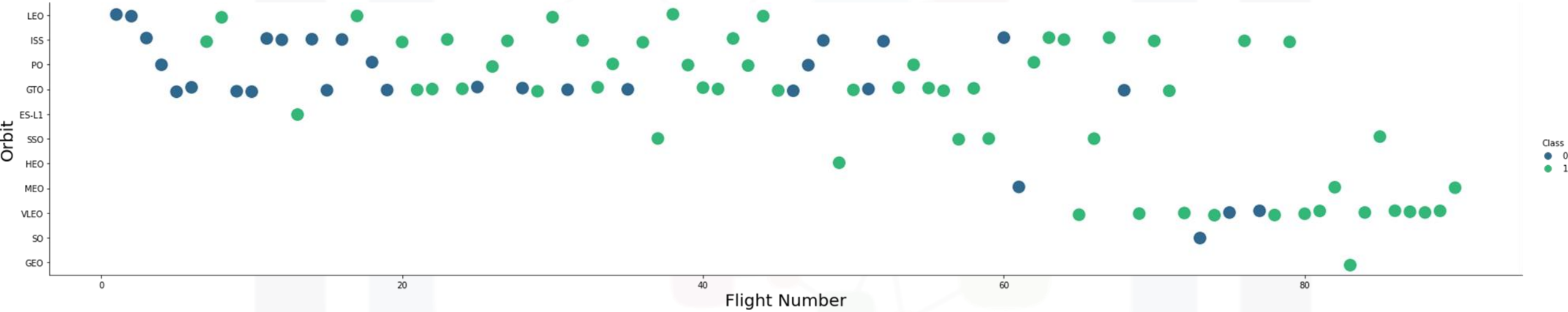


Interpretation:

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

RESULTS: EDA with visulization

Flight number & Orbit

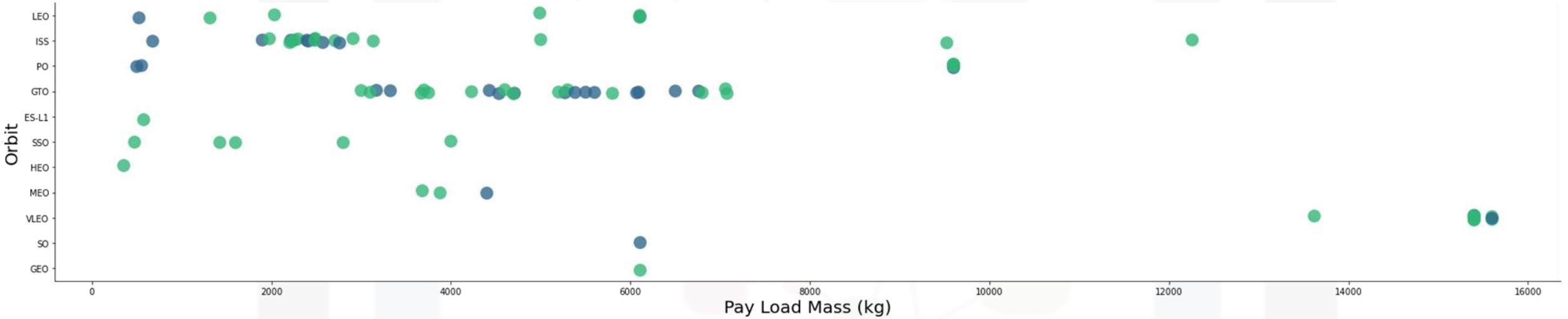


Interpretation:

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

RESULTS: EDA with visulization

Payload & Orbit

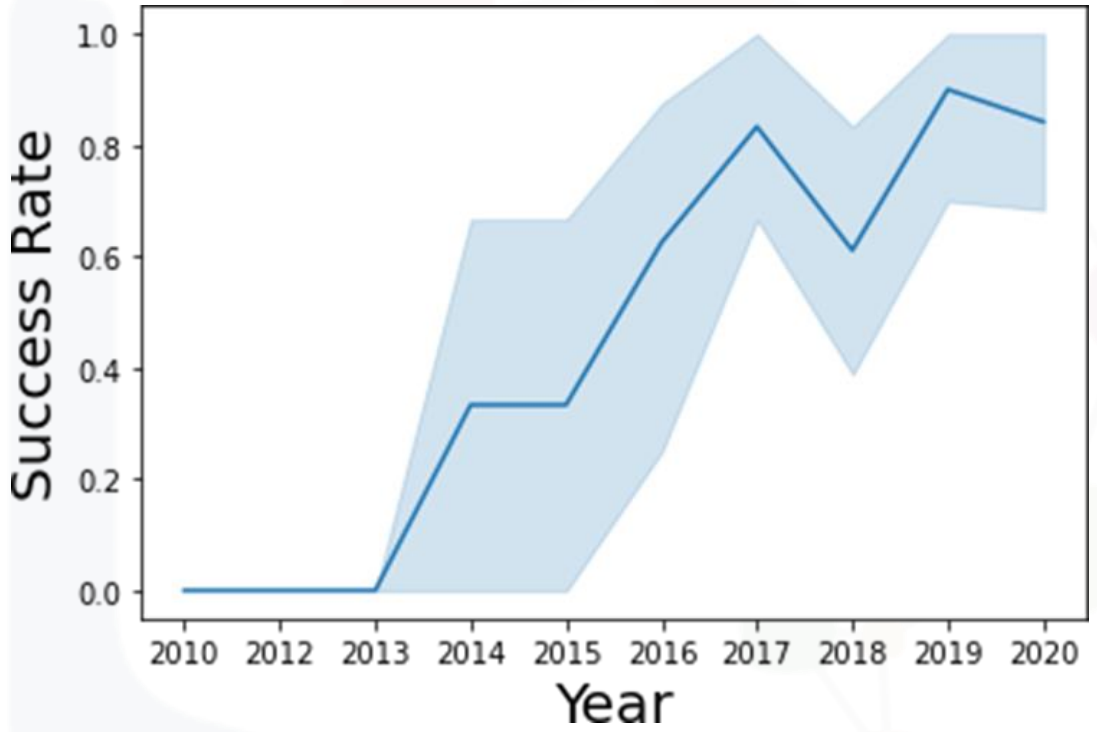


Interpretation:

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

RESULTS: EDA with visulization

Year & Sucess rate



Interpretation:

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%

RESULTS: EDA with SQL

All launch sites names

```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

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

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

- Query unique launch site names from 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.
- Only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

RESULTS: EDA with SQL

Launch sites names beginning 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

RESULTS: EDA with SQL

Average Payload Mass by F9 v1.1

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

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

avg_payload_mass_kg
2928

This query calculates the average payload mass or launches which used booster version F9 v1.1
Average payload mass of F9 1.1 is on the low end of our payload mass range

RESULTS: EDA with SQL

First Successful Ground Pad Landing Date

```
%%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

This query returns the first successful ground pad landing date.
First ground pad landing wasn't until the end of 2015.
Successful landings in general appear starting 2014.

RESULTS: EDA with SQL

Successful Drone Ship Landing with Payload Between 4000 and 6000

```
%%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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

RESULTS: EDA with SQL

Total Number of Each Mission Outcome

```
%%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

This query returns a count of each mission outcome. SpaceX appears to achieve its mission outcome nearly 99% of the time.

RESULTS: EDA with SQL

Boosters that Carried Maximum Payload

```
%%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

This query returns the booster versions that carried the highest payload mass of 15600 kg. These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

RESULTS: EDA with SQL

Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20

```
%%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;
```

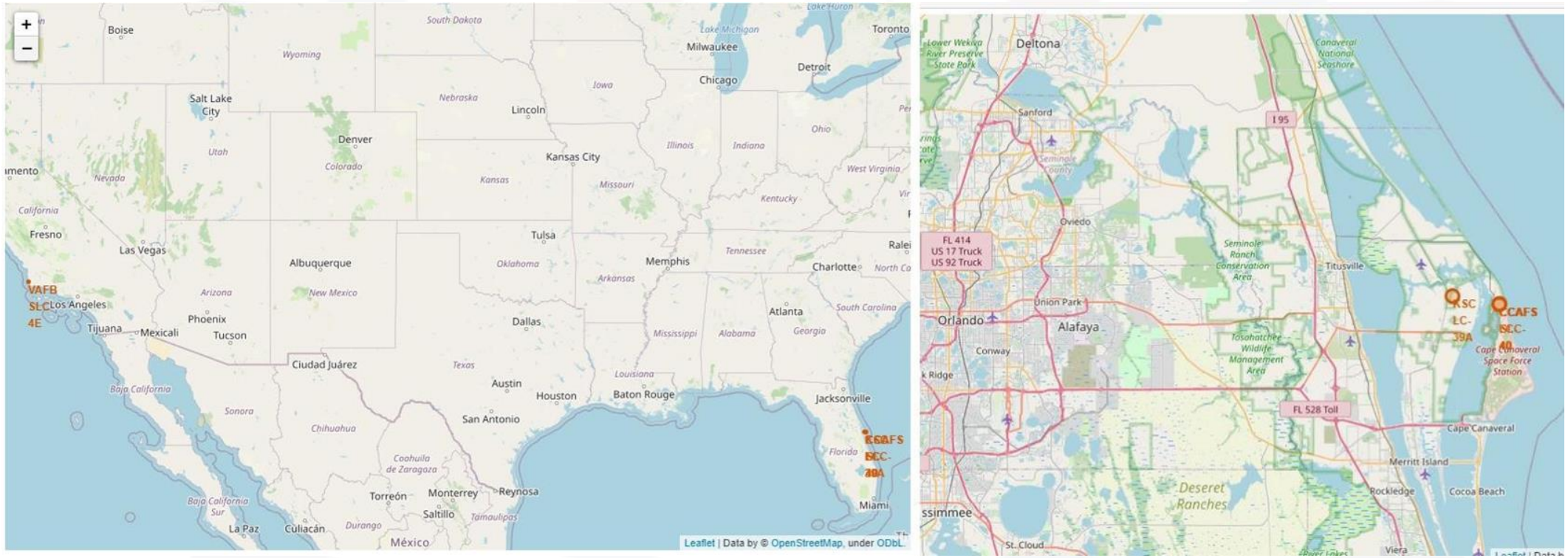
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

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

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively. There are two types of successful landing outcomes: drone ship and ground pad landings.

RESULTS: Interactive map FOLIUM

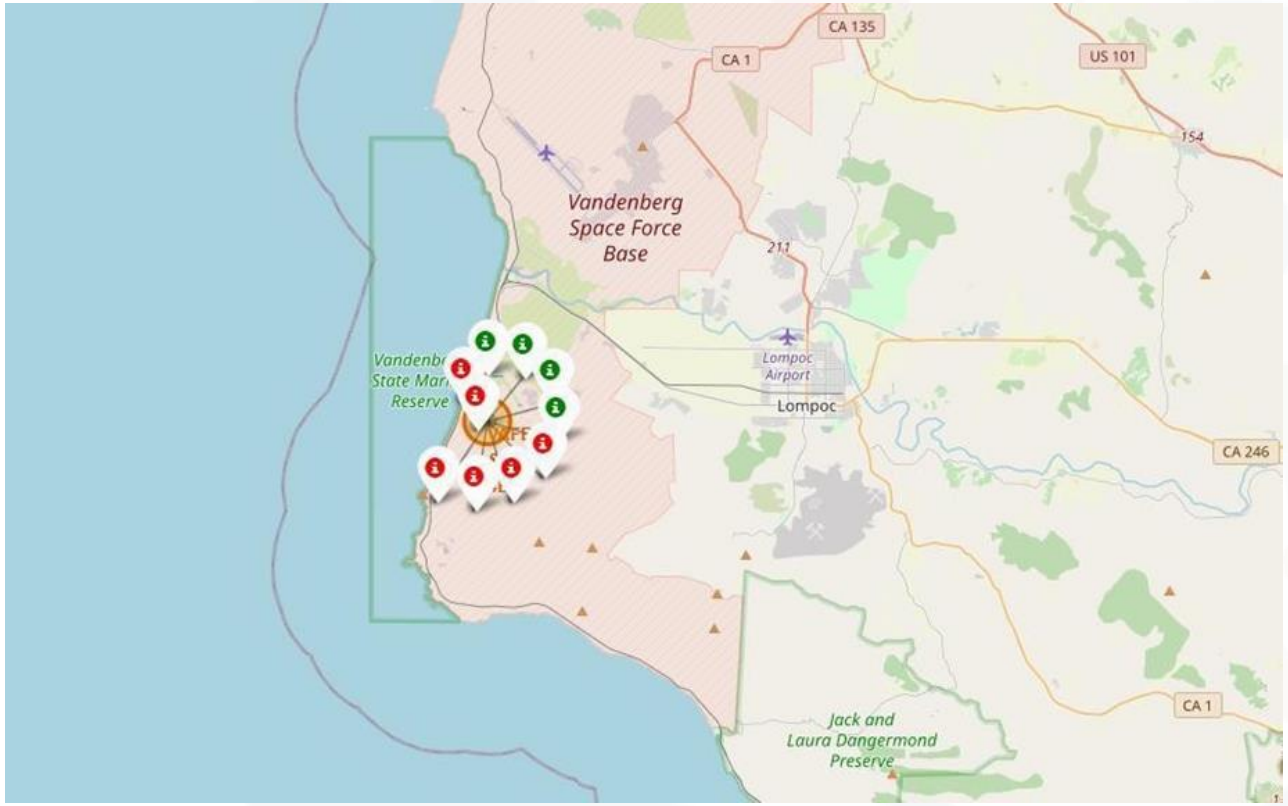
Launch site locations



This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively. There are two types of successful landing outcomes: drone ship and ground pad landings.

RESULTS: Interactive map FOLIUM

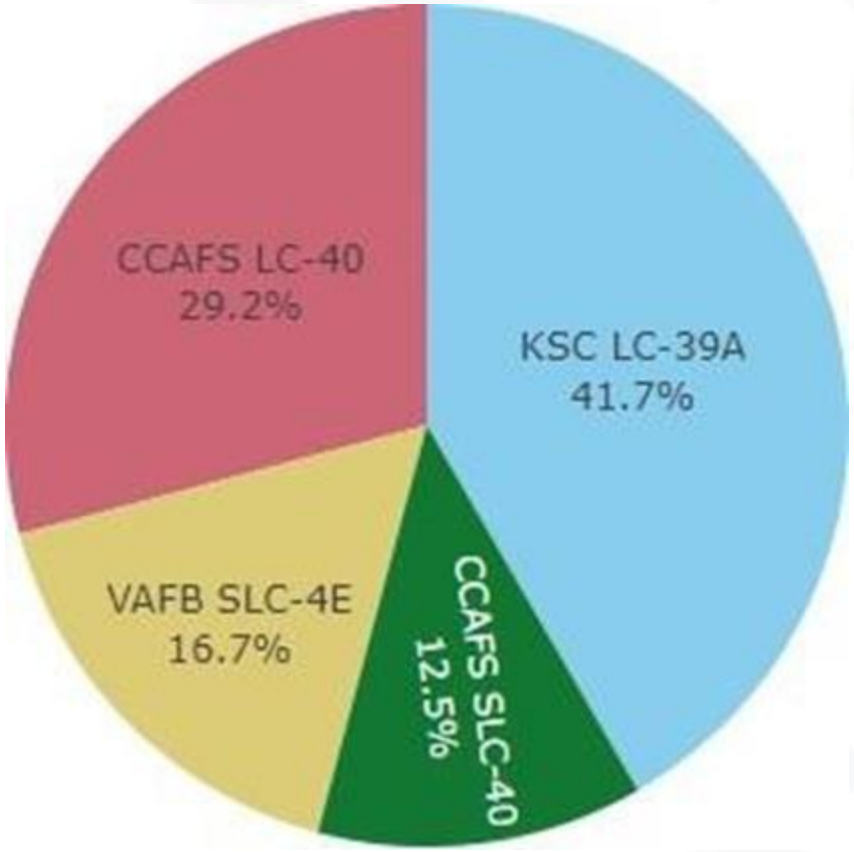
Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).
The map shows: VAFB SLC-4E shows 4 successful landings and 6 failed landings.

RESULTS: Dashboard with DASH

Successful launches across launch sites

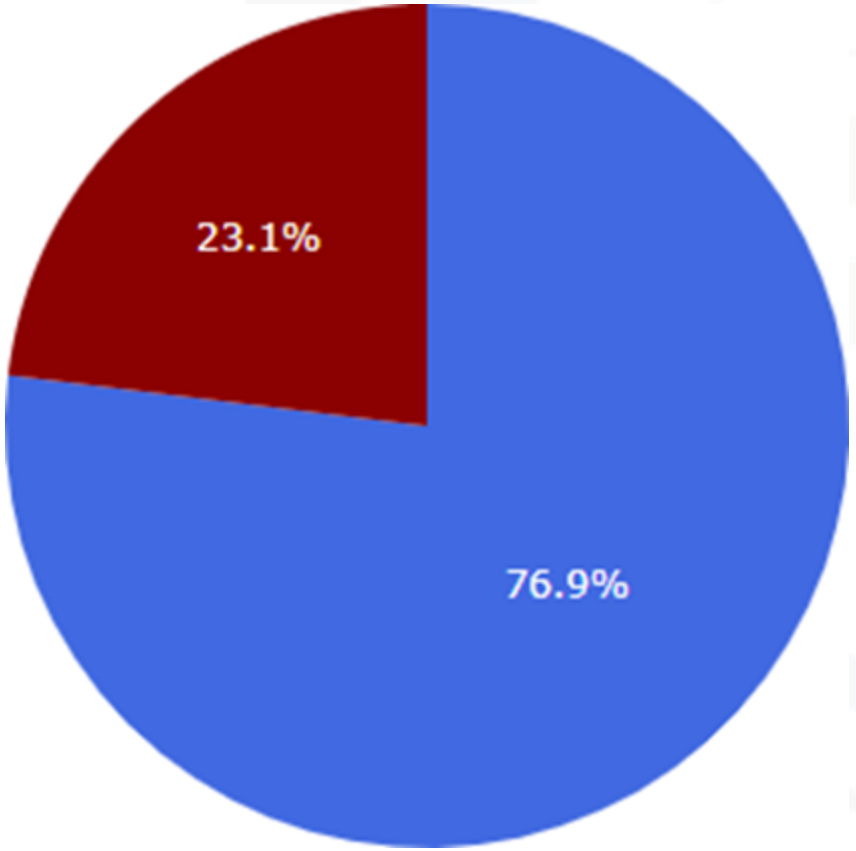


This is the distribution of successful landings across all launch sites:

- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change.
- VAFB has the smallest share of successful landings.
- This may be due to smaller sample and increase in difficulty of launching in the west coast.

RESULTS: Dashboard with DASH

Highest success rate launche site



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

RESULTS: Dashboard with DASH

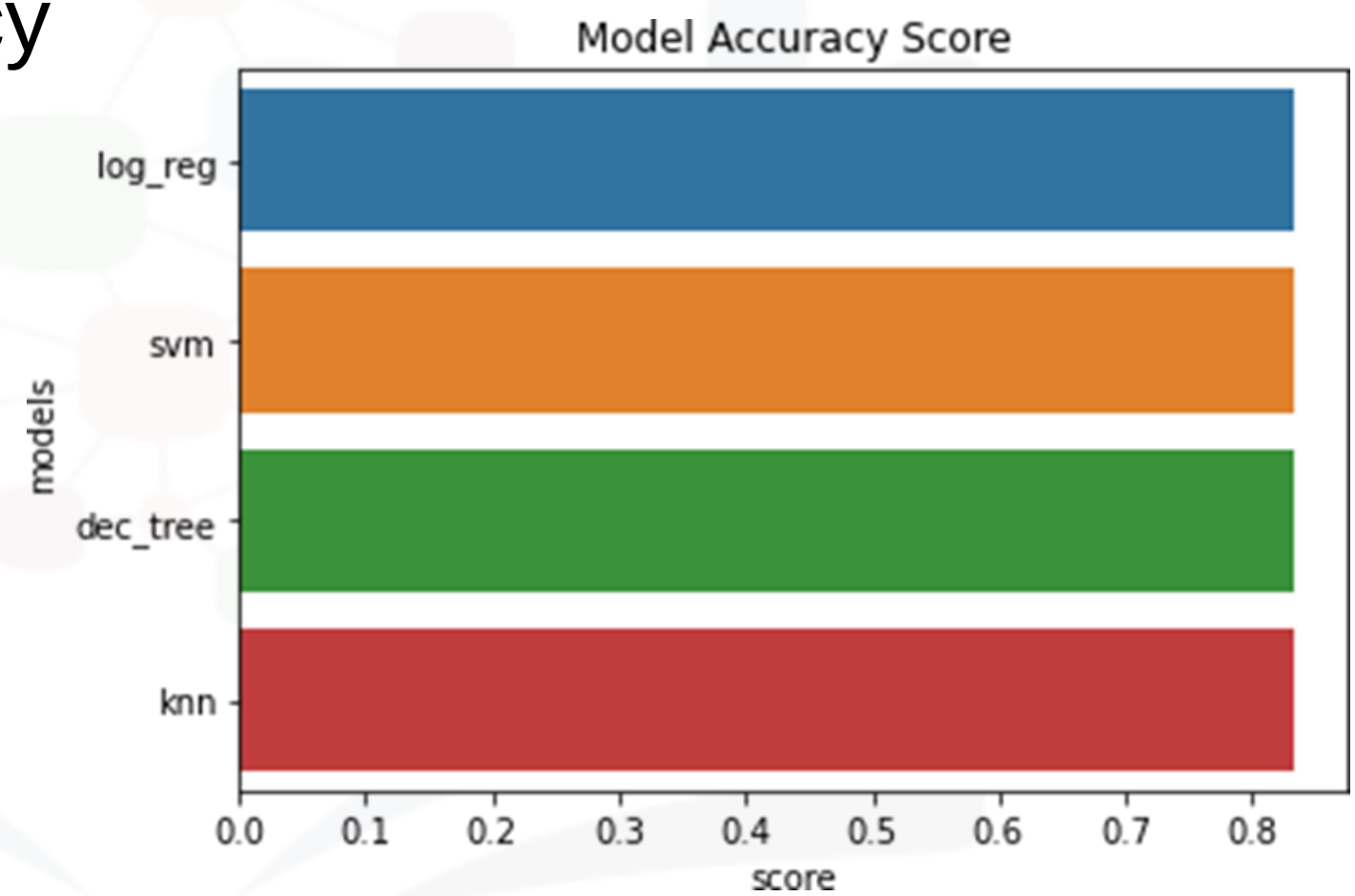
Payload Mass vs. Success vs. Booster Version Category

Plotly dashboard has a Payload range selector.
However, this is set from 0-10000 instead of the max Payload of 15600.
Class indicates 1 for successful landing and 0 for failure.

RESULTS: Predictive analysis

Classification Accuracy

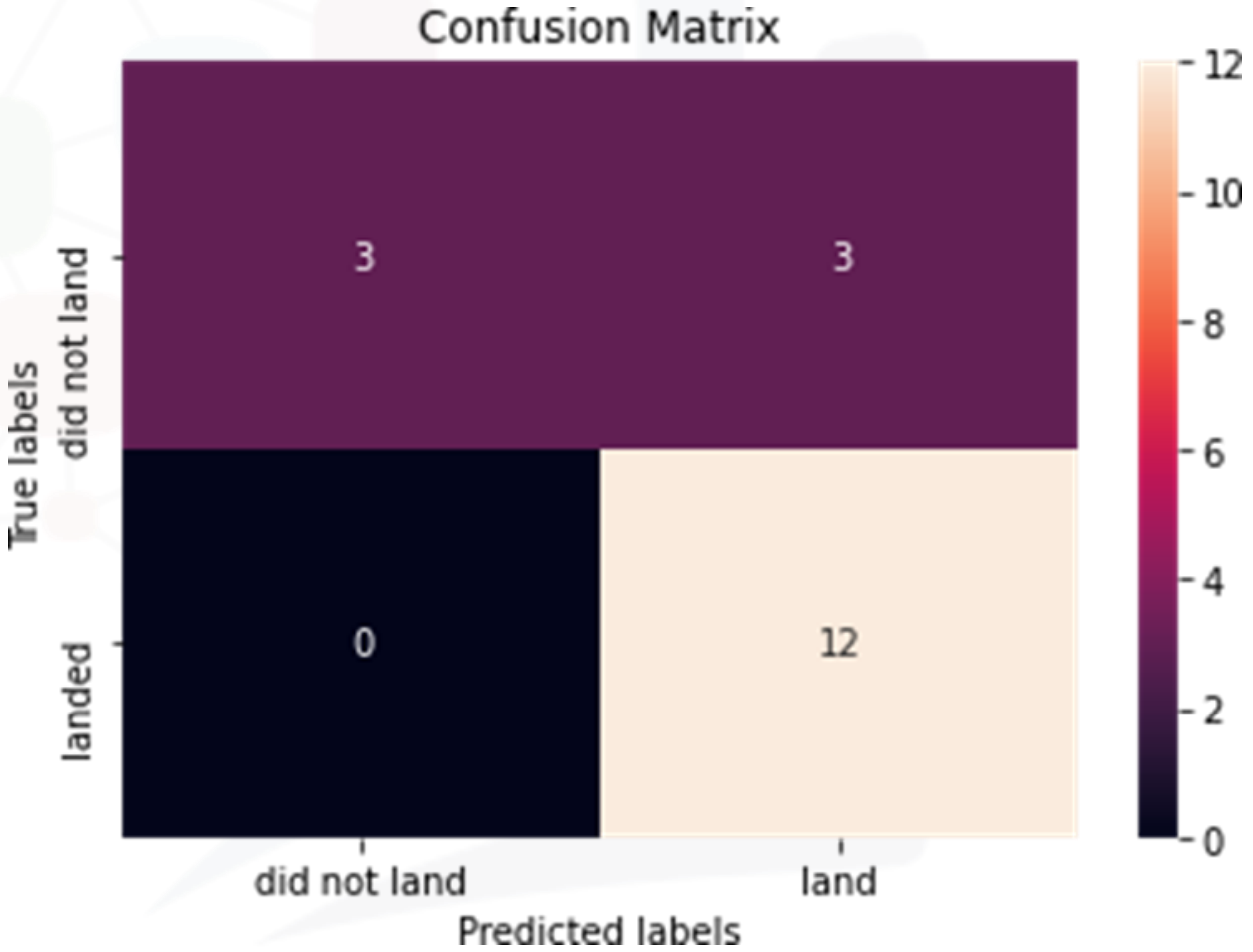
- All models had virtually the same accuracy on the test set at 83.33% accuracy.
- It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.



RESULTS: Predictive analysis

Confusion Matrix

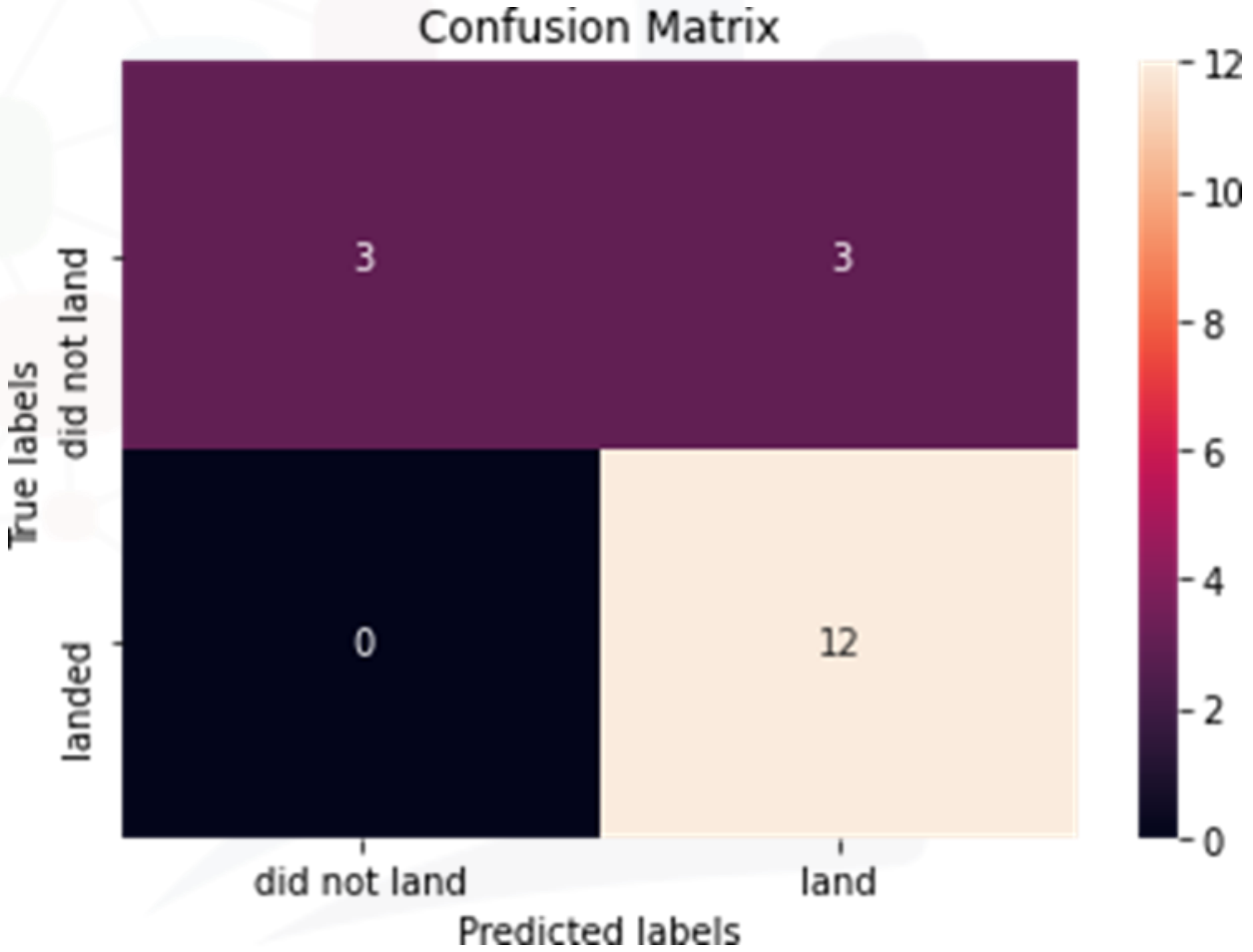
•Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
•The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.



RESULTS: Predictive analysis

Confusion Matrix

•Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
•The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.



CONCLUSION

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- If possible more data should be collected to better determine the best machine learning model and improve accuracy