IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abolfazl Jafari
2022/09/12

Outline

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

Executive Summary

- Collected data from public SpaceX API and SpaceX Wiki page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, dashboards, and folium maps. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one-hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models.

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results
- Four machine learning models were produced:
  - Logistic Regression,
  - Support Vector Machine,
  - Decision Tree Classifier
  - K Nearest Neighbors.

Introduction

- **Project background and context**
  - Commercial Space Age is Here
  - 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
- **Problems you want to find answers**
  - Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery
  - Correlations between each rocket variables and successful landing rate
  - Conditions to get the best results and ensure the best successful landing rate

Section 1

Methodology

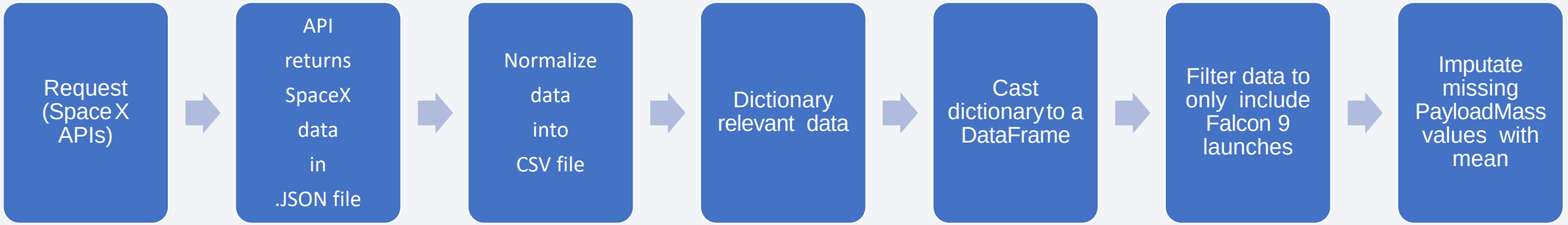
Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - • Convert outcomes into Training Labels with the booster successfully/unsuccessful landed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

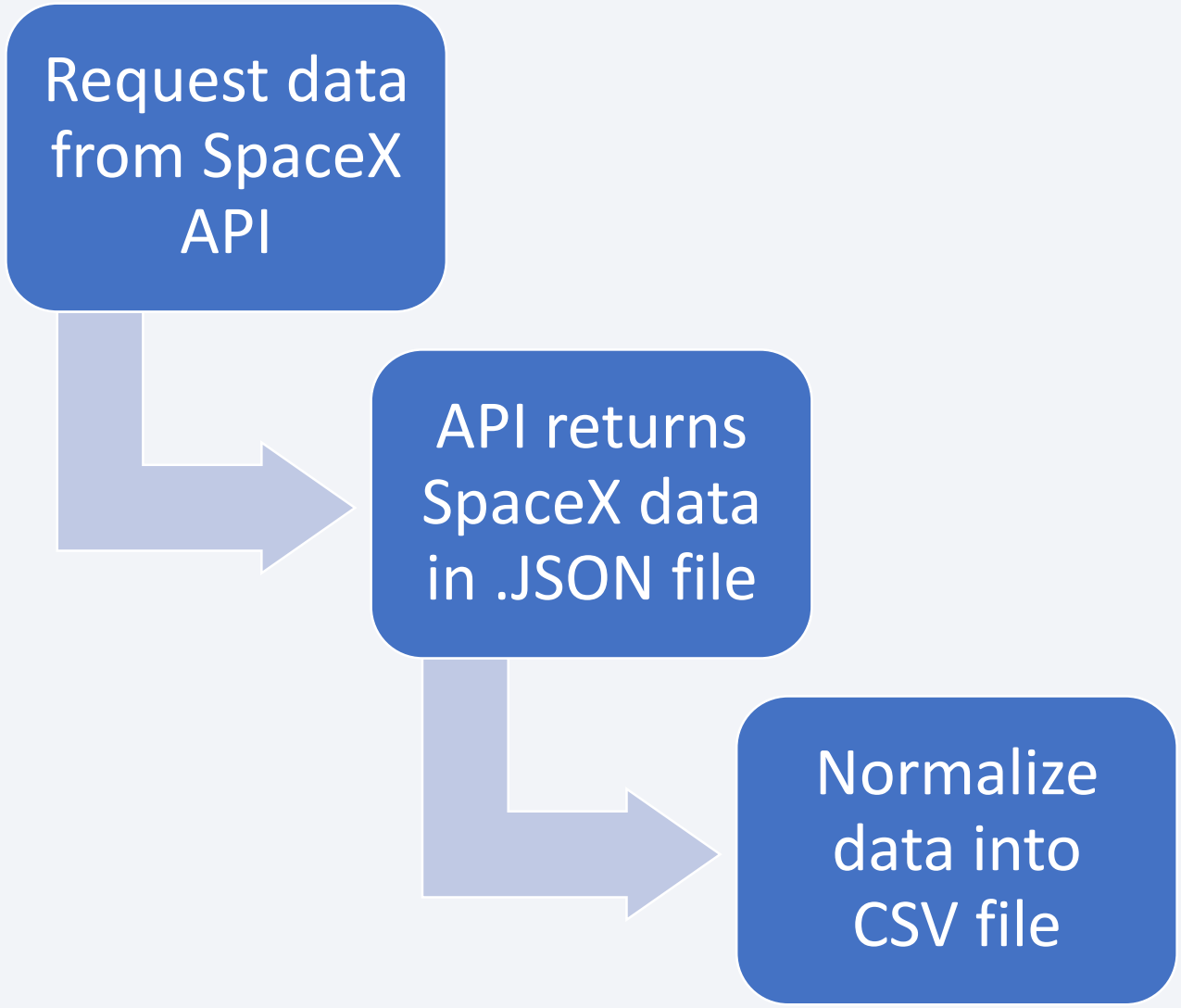
- The data collection process includes a combination of API requests from the SpaceX API and web scraping data from a table in the Wikipedia page of SpaceX, Falcon 9 and Falcon Heavy Launches Records.
  - SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Wikipedia Web Scrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time



Data Collection – SpaceX API

SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

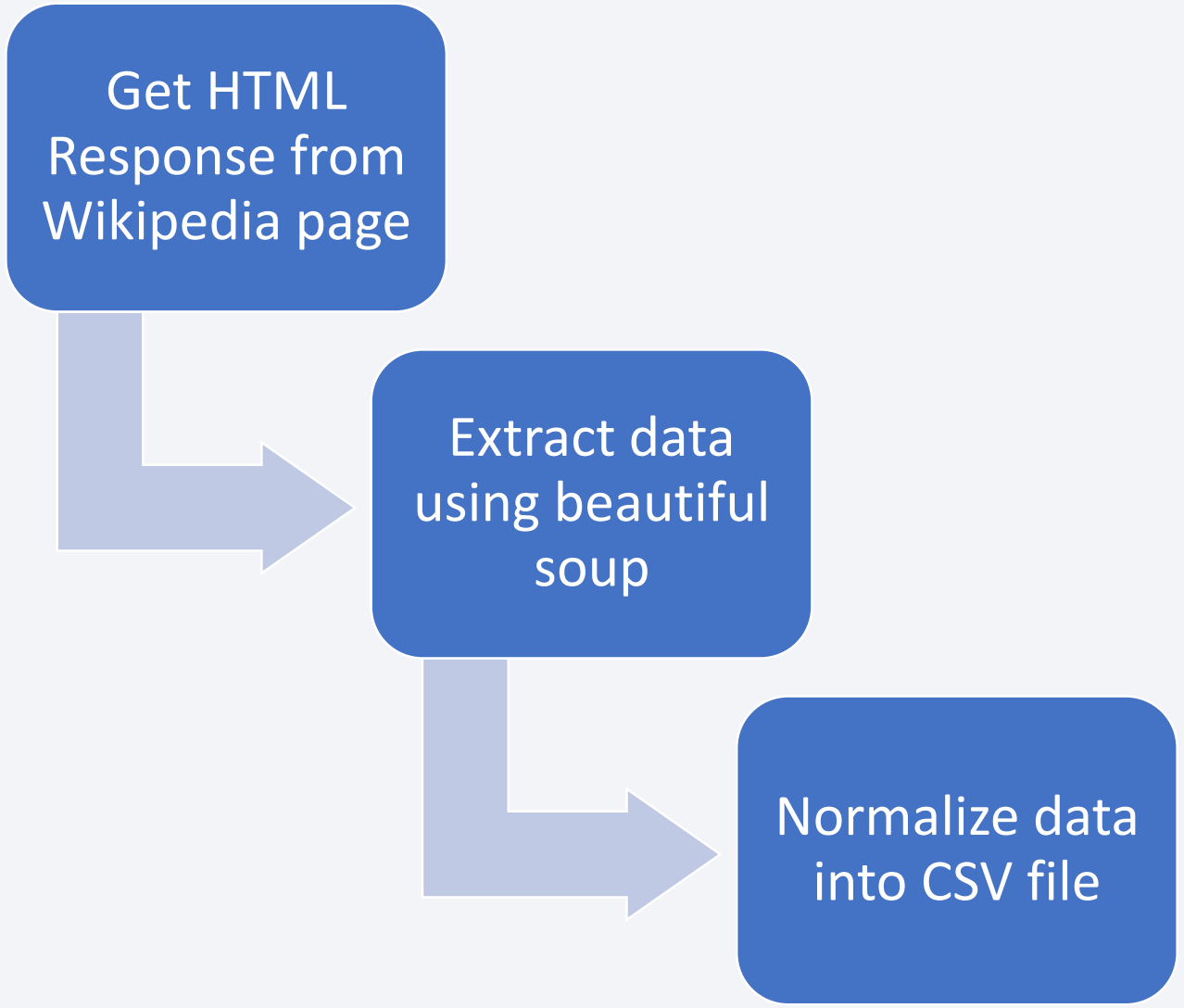
Github



Data Collection - Scraping

Wikipedia Web Scrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Github



Data Wrangling

- There are several cases in which the booster failed to successfully land on the dataset, and sometimes it attempted to land but failed because of accident.
  - True Ocean: the mission result has successfully landed in a specific area of the ocean
  - False Ocean: the mission result has not successfully landed in a specific area of the ocean
  - True RTLS: the mission result successfully landed on the ground pad
  - False RTLS: the mission result has not successfully landed on the ground pad
  - True ASDS: the mission result has successfully landed on the drone ship
  - False ASDS: the mission result has not landed on the drone ship
  - Converting these results into training labels:
  - 1 = successful / 0 = failure

Github

EDA with 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

Github

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

Github

Build an Interactive Map with Folium

- Objects created and added to a folium map:
  - Markers that show all launch sites on a map
  - Markers that show the success/failed launches for each site on the map
  - Lines that show the distances between a launch site to its proximities
- By adding these objects, following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes

Github

Build a Dashboard with Plotly Dash

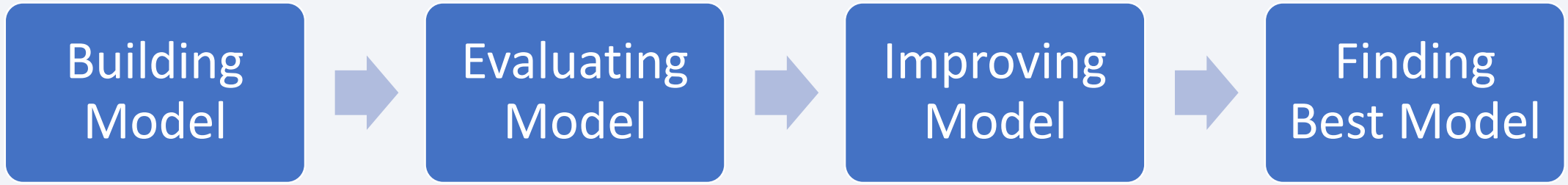
- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and
- booster version category.

- Github

Predictive Analysis (Classification)

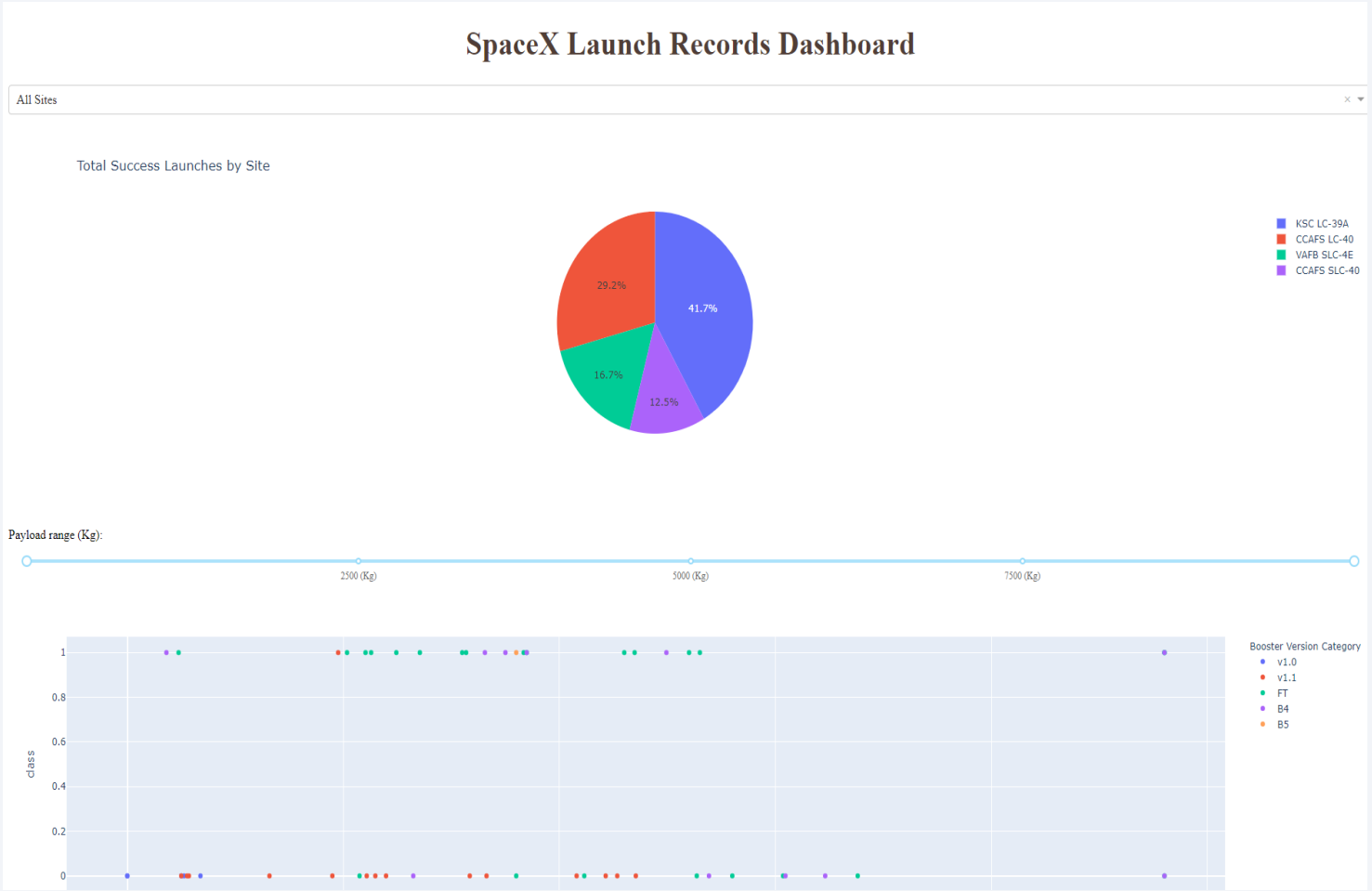
- Perform exploratory Data Analysis and determine Training Labels
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data

Github



Results

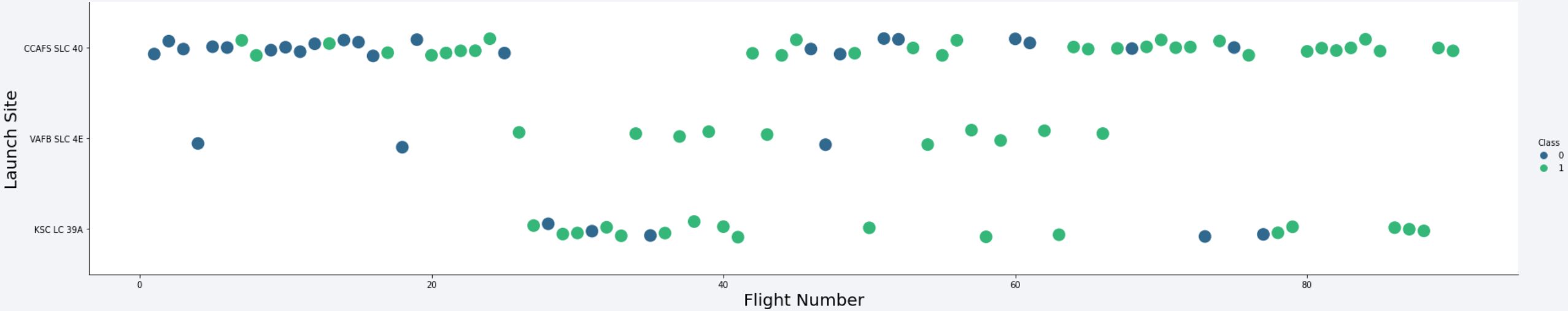
- The left screenshot is a preview of the Dashboard with Plotly Dash.
- The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and Interactive Dashboard will be shown in the next slides.
- Comparing the accuracy of the four methods, all return the same accuracy of about 83% for test data.



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



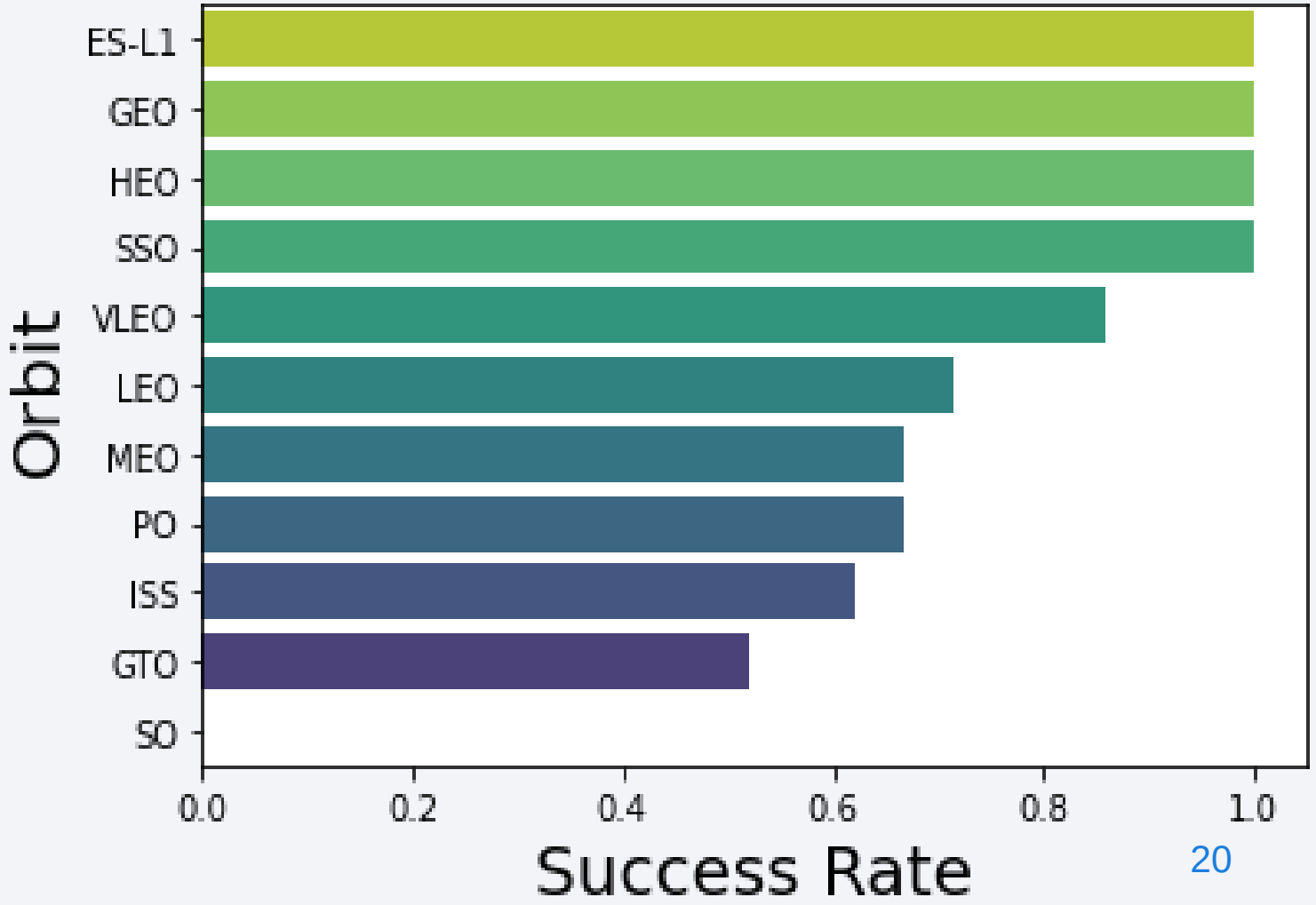
- Class 0 (blue) represents unsuccessful launch, and Class 1 (green) represents successful launch.
- This figure shows that the success rate increased as the number of flights increased.
- As the success rate has increased considerably since the 20th flight, this point seems to be a big breakthrough.

Payload vs. Launch Site

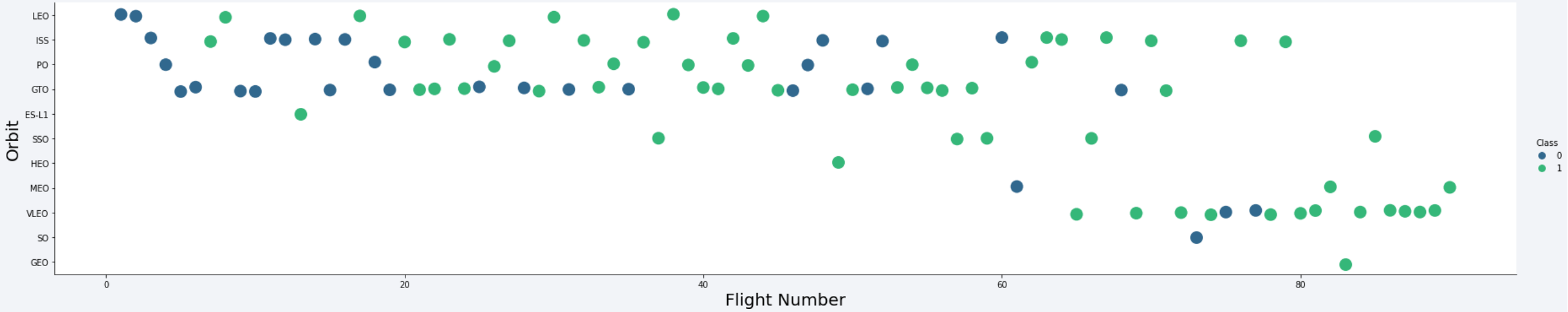
- • Class 0 (blue) represents unsuccessful launch, and Class 1 (green) represents successful launch.
- At first glance, the larger pay load mass, the higher the rocket's success rate, but it seems difficult to make decisions based on this figure because no clear pattern can be found between successful launch and Pay Load Mass.

Success Rate vs. Orbit Type

- • Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%).
- On the other hand, the success rate of orbit type GTO is only 50%, and it is the lowest except for type SO, which recorded failure in a single attempt.

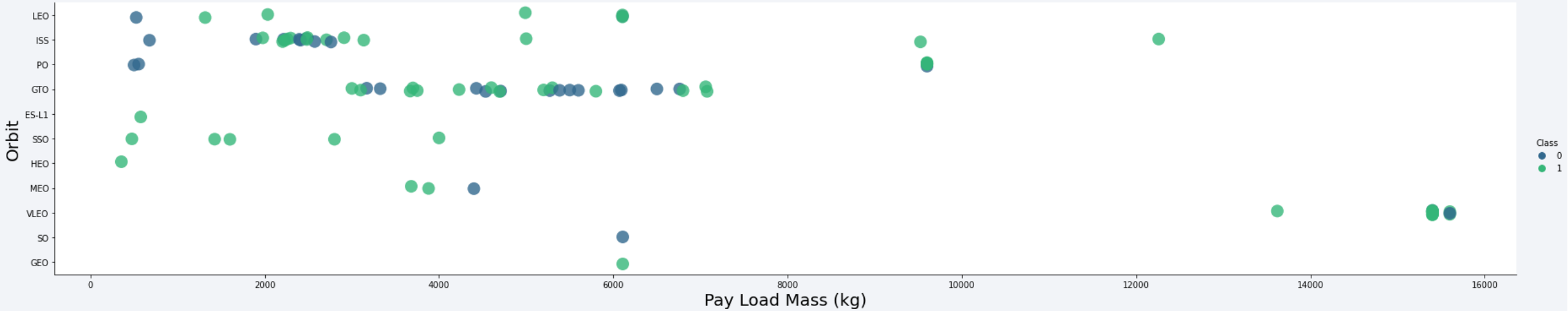


Flight Number vs. Orbit Type



- Class 0 (blue) represents unsuccessful launch, and Class 1 (green) represents successful launch.
- In most cases, the launch outcome seems to be correlated with the flight number.
- On the other hand, in GTO orbit, there seems to be no relationship between flight numbers and success rate.
- SpaceX starts with LEO with a moderate success rate, and it seems that VLEO, which has a high success rate, is used the most in recent launches.

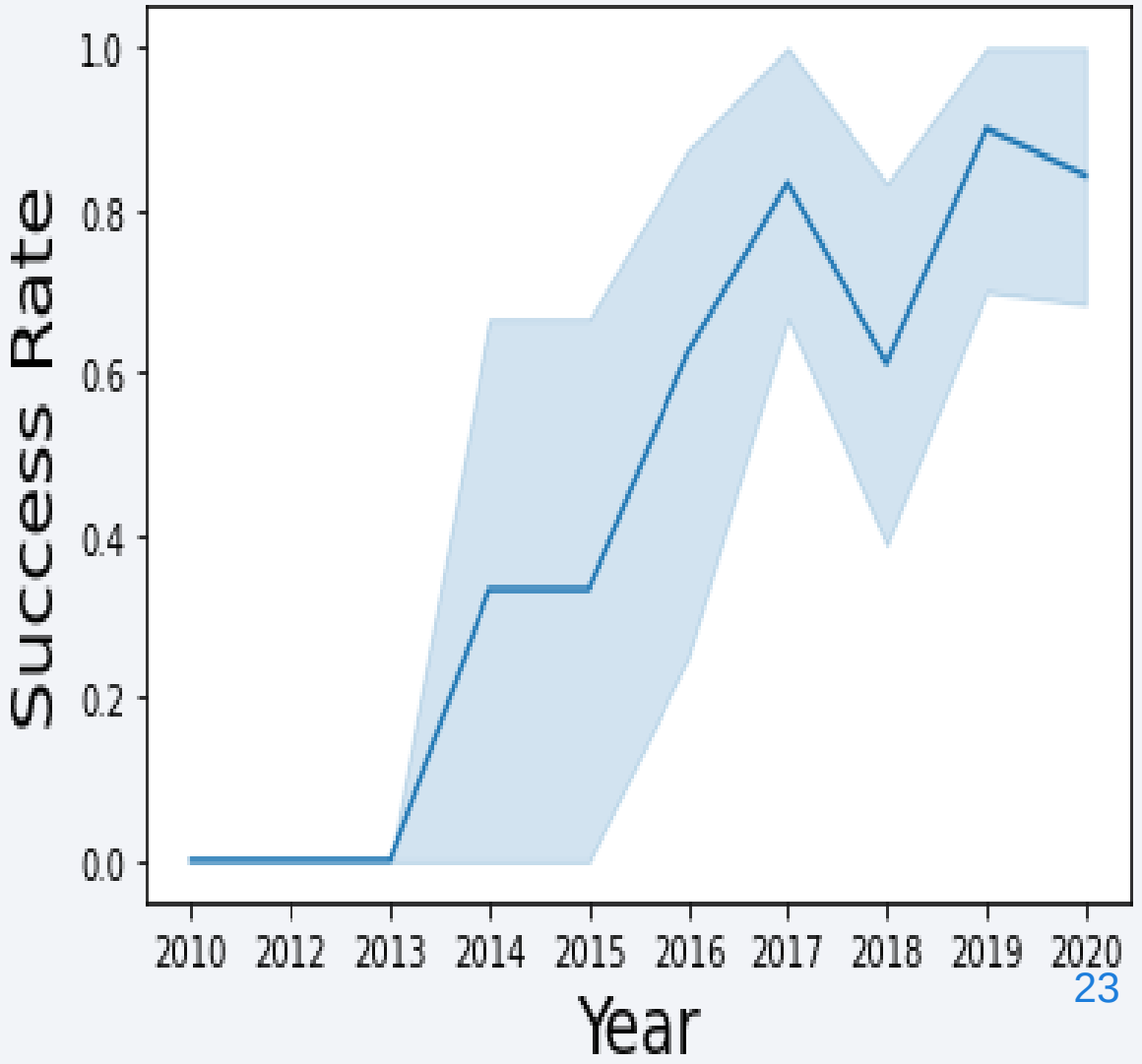
Payload vs. Orbit Type



- • Class 0 (blue) represents unsuccessful launch, and Class 1 (green) represents successful launch.
- With heavy payloads the successful landing or positive landing rate are more for LEO and ISS.
- However, in the case of GTO, it is hard to distinguish between the positive landing rate and the negative landing because they are all gathered together.

Launch Success Yearly Trend

- Since 2013, the success rate has continued to increase until 2017.
- The rate decreased slightly in 2018.
- Recently, it has shown a success rate of about 80%.



All Launch Site Names

- Query
  - SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL

- When the SQL DISTINCT clause is used in the query, only unique values are displayed in the Launch_Site column from the SpaceX table.
- There are four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

- Query
  - SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%‘ LIMIT 5
- Only five records of the SpaceX table were displayed using LIMIT 5 clause in the query.
- Using the LIKE operator and the percent sign (%) together, the Launch_Site name starting with CAA could be called

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

- Query
  - SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'

- Using the SUM() function to calculate the sum of column PAYLOAD_MASS__KG_.

sum_payload_mass_kg
45596

- In the WHERE clause, filter the dataset to perform calculations only if Customer is NASA (CRS)

Average Payload Mass by F9 v1.1

- Query
  - SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1‘

avg_payload_mass_kg
2928

- Using the AVG() function to calculate the average value of column PAYLOAD_MASS__KG_.

- In the WHERE clause, filter the dataset to perform calculations only if Booster_version is F9 v1.1.

First Successful Ground Landing Date

- Query
  - SELECT MIN(DATE) AS first_successful_landing_date FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)‘

- Using the MIN() function to find out the earliest date in the column DATE.

first_success
2015-12-22

- In the WHERE clause, filter the dataset to perform a search only if Landing__outcome is Success (ground pad).

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query
  - SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
- In the WHERE clause, filter the dataset to perform a search if Landing__outcome is Success (drone ship).
- Using the AND operator to display a record if additional condition PAYLOAD_MASS__KG_ is 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

- Query
  - SELECT MISSION_OUTCOME, COUNT(*) AS total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME
- Using the COUNT() function to calculate the total number of columns
- Using the GROUP BY statement, groups rows that have the same values into summary rows to find the total number in each Mission_outcome.
- According to the result, SpaceX se[illegible] completed nearly 99% of its missic[illegible]

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

Boosters Carried Maximum Payload

- Query
  - SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

- • Using a subquery, first, find the maximum value of the payload by using MAX() function, and second, filter the dataset to perform a search if PAYLOAD_MASS__KG_ is the maximum value of the payload..

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

- Query
  - SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015‘

- In the WHERE clause, filter the dataset to perform a search if Landing__outcome is Failure (drone ship).
- In 2015, there were two landing failures on drone ships.

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

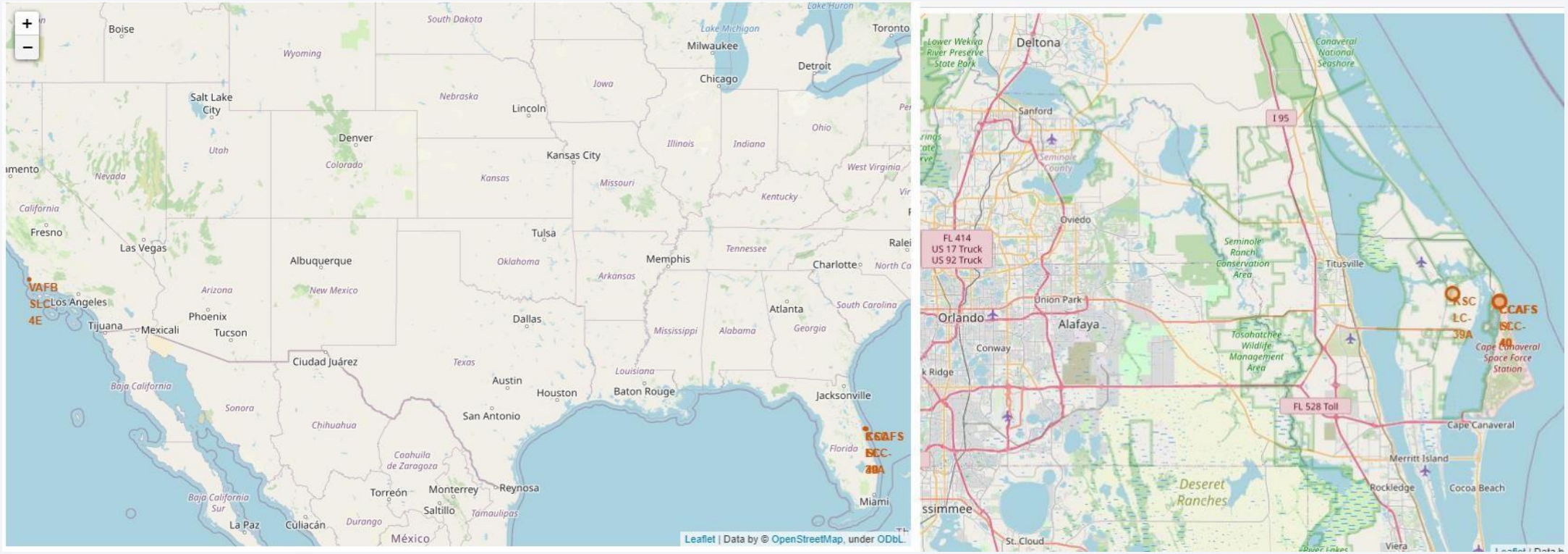
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Query
  - SSELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY total_number DESC
- In the WHERE clause, filter the dataset to perform a search if the date is between 2010-06-04 and 2017-03-20.
- Using the ORDER BY keyword to sort the records by total number of landing, and using DESC keyword to sort the records in descending order.
- According to the results, the number of successes and failures between 2010-06-04 and 2017-03-20 was

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


Section 3
Launch Sites
Proximities Analysis

All Launch Sites’ Locations

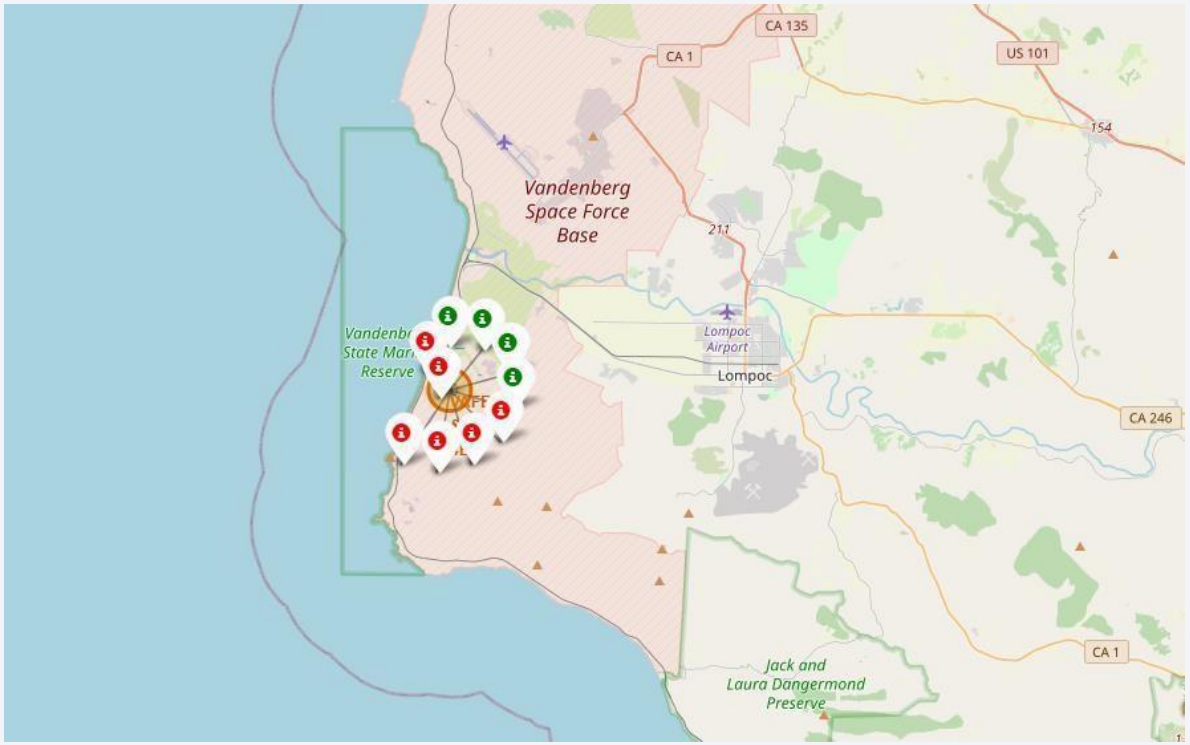


- The left map shows all SpaceX launch sites, and the right map also shows that all launch sites are in the United States.
- As can be seen on the map, all launch sites are near the coast.

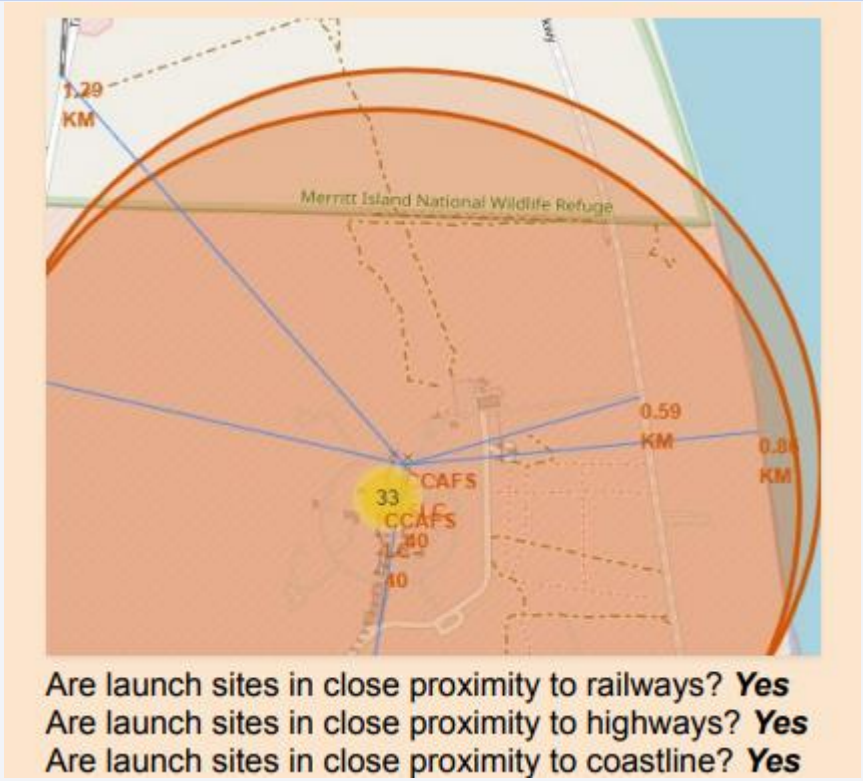
Color-Coded Launch Markers

- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed

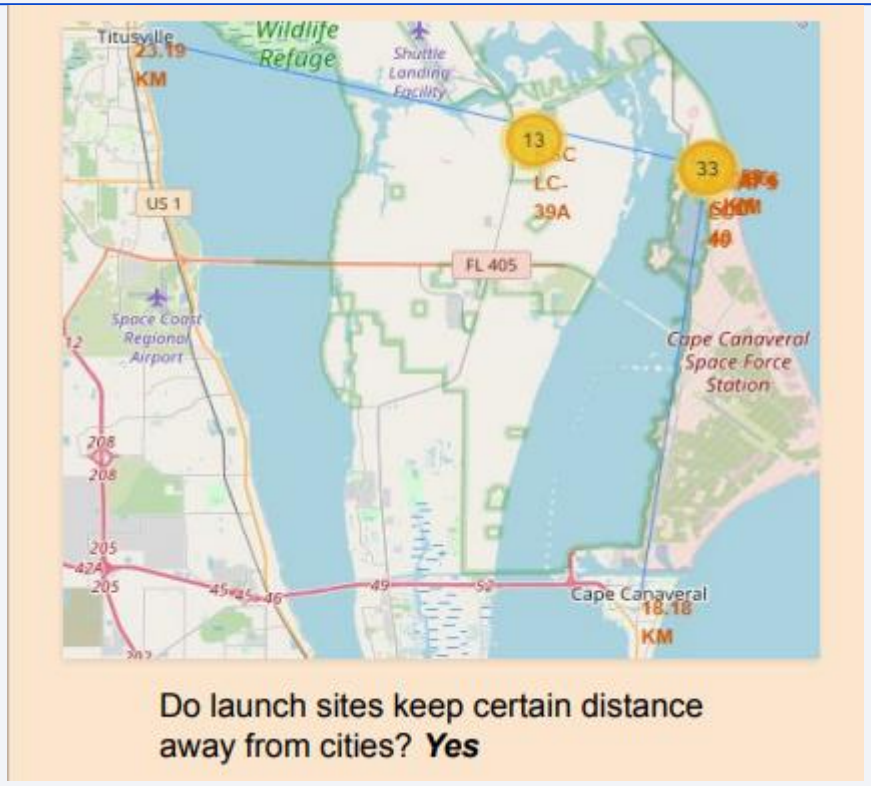
- landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



Proximities of Launch Sites

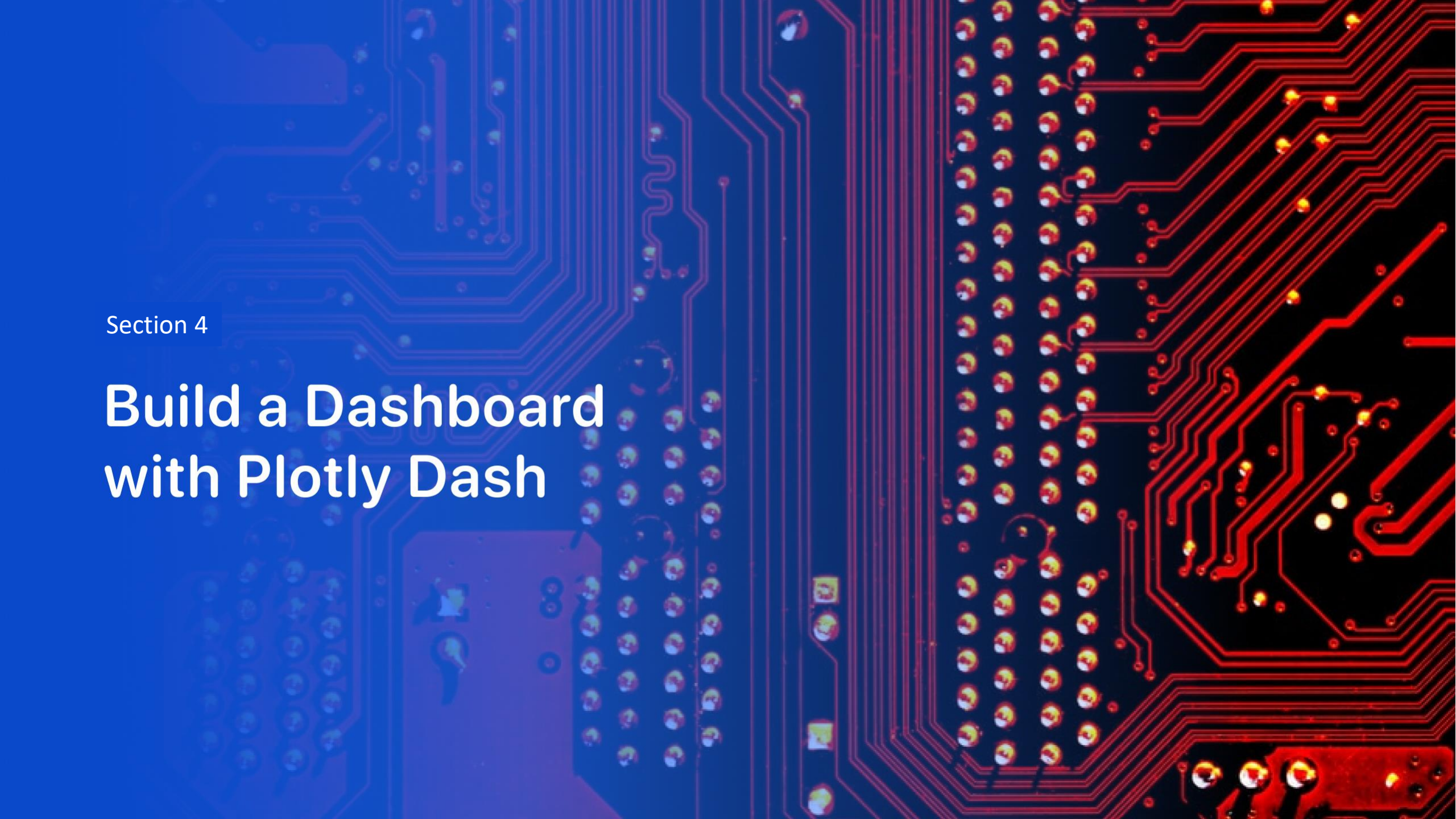


Are launch sites in close proximity to railways? ***Yes***
Are launch sites in close proximity to highways? ***Yes***
Are launch sites in close proximity to coastline? ***Yes***



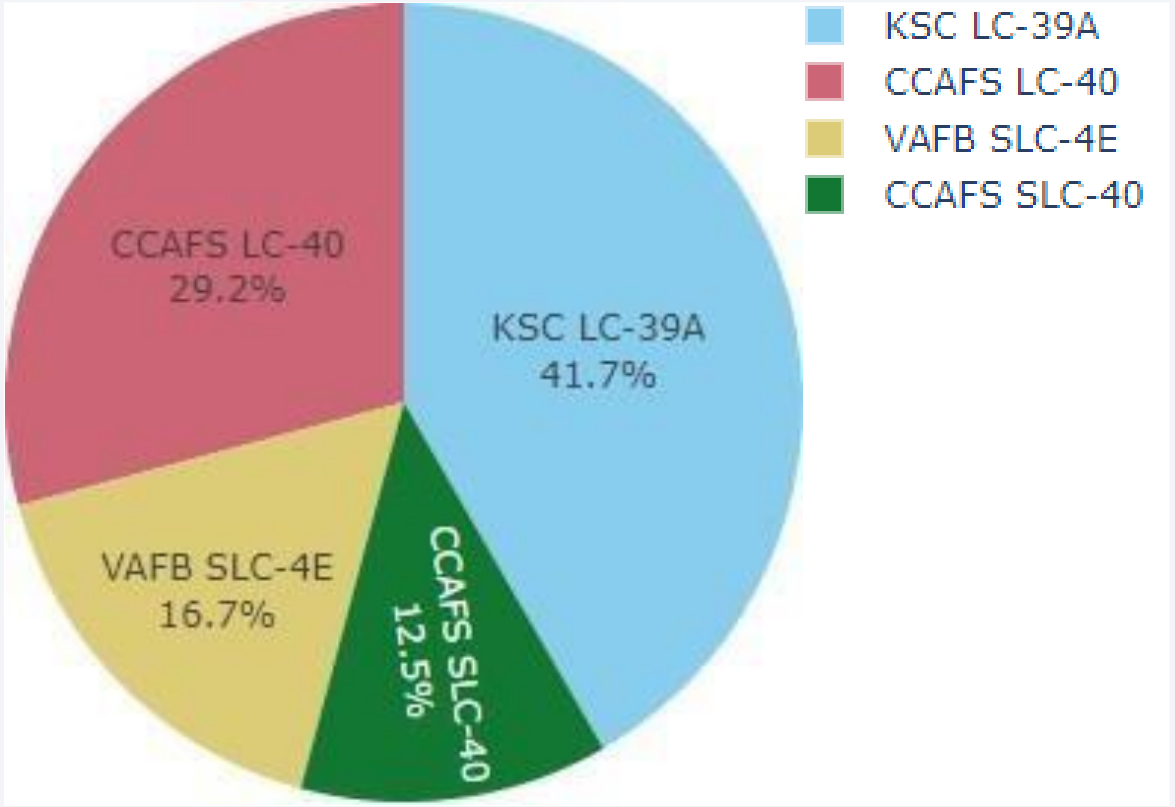
Do launch sites keep certain distance away from cities? ***Yes***

- • It can be found that the launch site is close to railways and highways for transportation of equipment or personnel, and is also close to coastline and relatively far from the cities so that launch failure does not pose a threat.


Section 4
Build a Dashboard with Plotly Dash

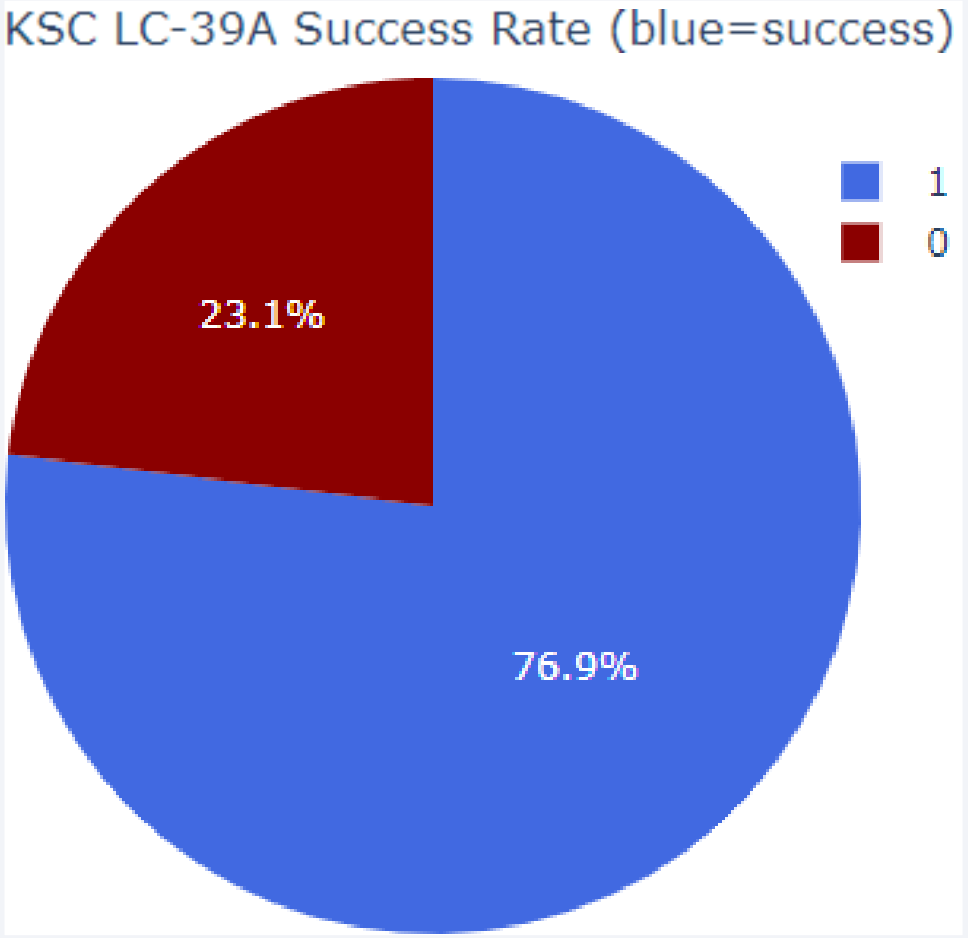
Total Success Launches By all sites

- 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.



Highest Success Rate Launch Site

- KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%).



Payload Mass vs. Success vs. Booster Version Category

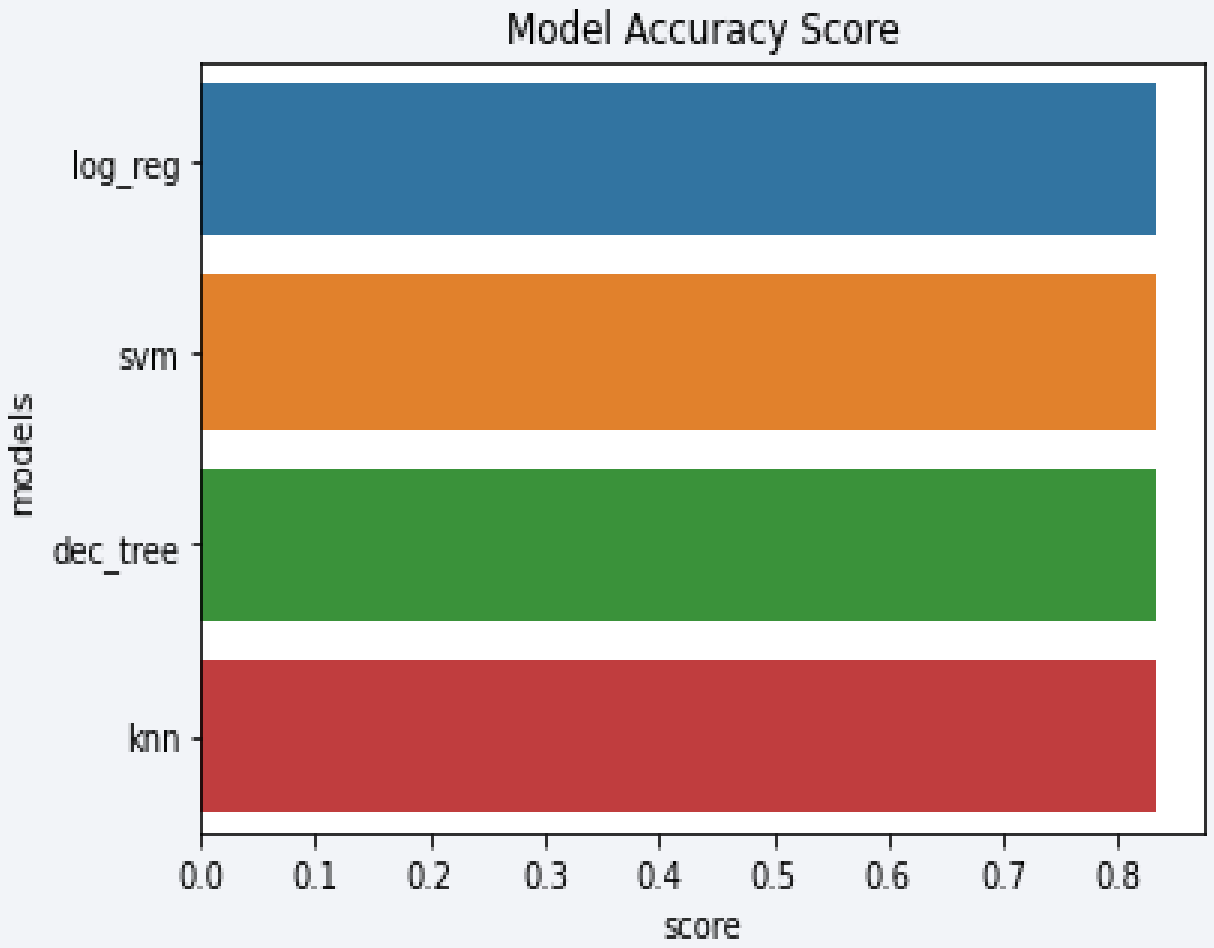
- These figures show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is higher than that of heavy weighted payloads(5000-10000 kg).

Section 5

Predictive Analysis (Classification)

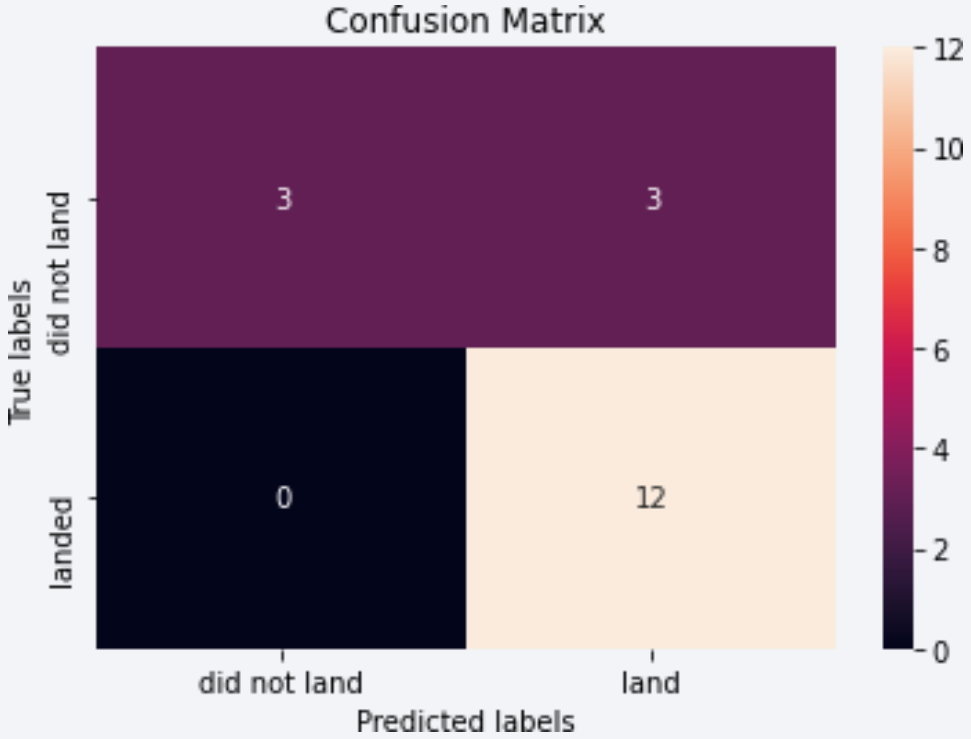
Classification Accuracy

- 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.



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.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- As the number of flights increased, the success rate increased, and recently it has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%). • The launch site is close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.

Appendix

- GitHub repository url:
  - https://github.com/abolfazll94/Coursera_Capstone

Thank you!