IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ritayan
2022/11/06

Outline

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

Executive Summary

The objective of this project was to predict the success rate of SpaceX Falcon 9 first stage landing.

Methodologies

- Data collection using API.
- Data wrangling.
- Exploratory Data Analysis using SQL.
- Interactive Maps applying Folium.
- Interactive Dashboard with Plotly.
- Predictive Analysis through classification.

Findings

- Exploratory data analysis results.
- Data visualizations.
- Best predictive model.

Introduction

Background

SpaceX is one of the most prestigious private companies in the world, that is involved in designing, manufacturing and launching rockets and satellites to space. SpaceX has contracts with NASA and the US DoD. One of the biggest achievements of SpaceX was to decrease the launch prices of rockets by more than 60%, pushing prices down in the entire industry. They were able to do that by reusing the first stage boosters, instead of expending them.

For this project, we have tried to predict if the Falcon 9's first stage will land successfully for a given launch, helping us to get a more accurate idea of the cost of launch.

Objectives

We want to determine how successful is the first stage and the probabilities of landing. In order to conduct the analysis the following questions need to be addressed.

1. What factors influence in making a landing successful?
2. Which variables are the most influential for the success rate?
3. In order to have the best results and ensure a successful landing, what conditions are required?
4. What is the most optimal rocket to use for a successful landing?

Section 1

Methodology

Methodology

The methodology used for this project is divided into five different stages ranging from compiling data, cleaning, to visualization and finally model prediction.

1. **Data collection**
    - SpaceX Rest API
    - Web Scrapping from Wikipedia
2. **Data wrangling**
    - One hot encoding data fields for Machine Learning and Dropping irrelevant columns.
3. **Exploratory data analysis (EDA) using visualization and SQL**
    - The use of diverse plots such as bar, scatter plots, to reveal the variable's relationship and provide patterns inside the data.
4. **Interactive visual analytics using Folium and Plotly Dash**
    - Use Folium and Plotly to create maps and Dashboards.
5. **Predictive analysis using classification models**
    - Build, train and evaluate models including Logistic Regression, SVM, Decision Tree and KNN.

Data Collection

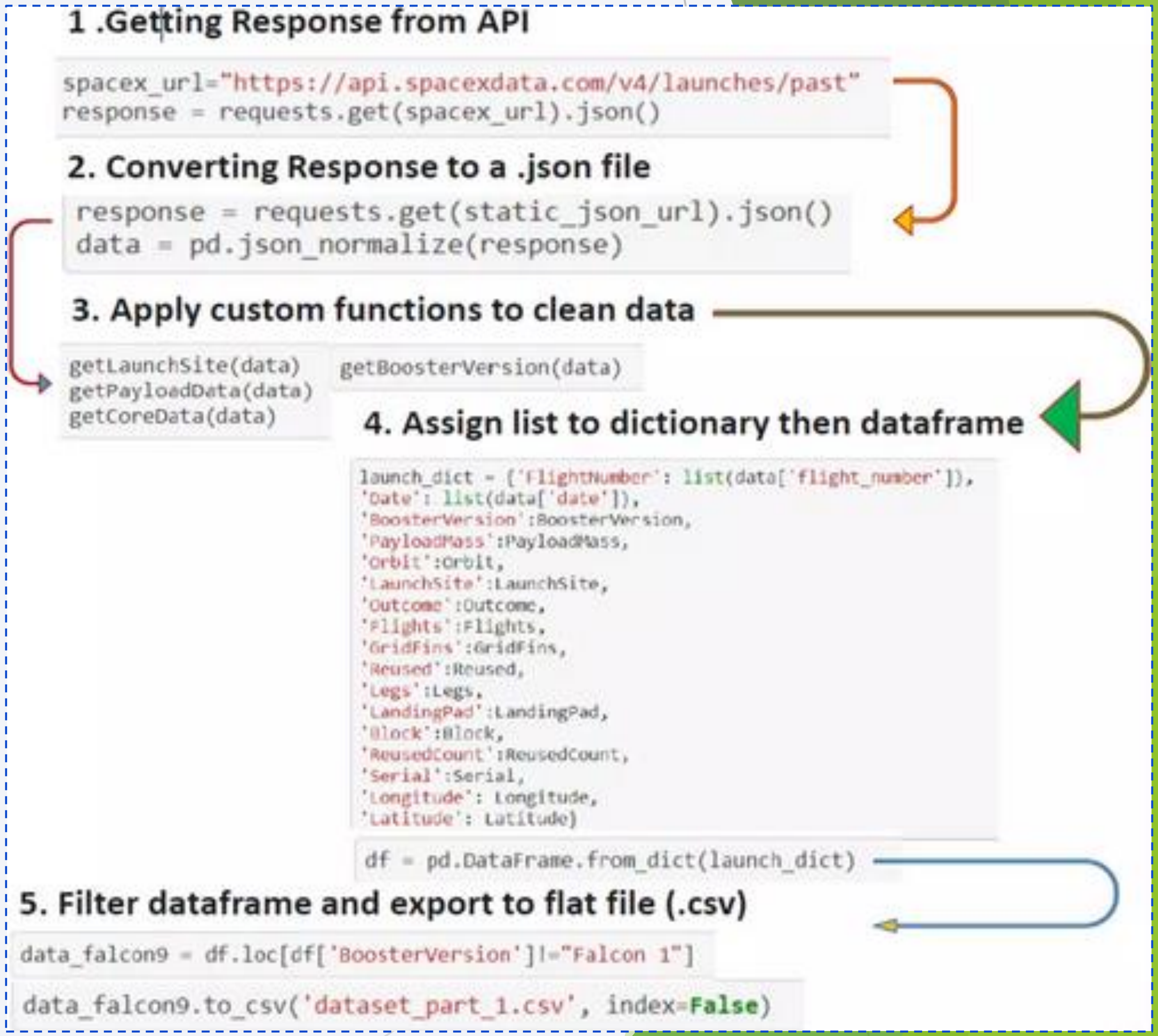
Dataset was collected using the Space X REST API which covers different information such as flight number, date, payload mass, and landing pad, among others. The main objective of using this dataset was to predict whether Space X will be able to successfully land the stage 1 boosters.

In order to start working with the dataset. We need to :

- Get Data from the API or Website by scrapping.
- Transform the data into a data frame.
- Filter the data frame as required before starting to work with the valuable information .
- Export the final dataset.

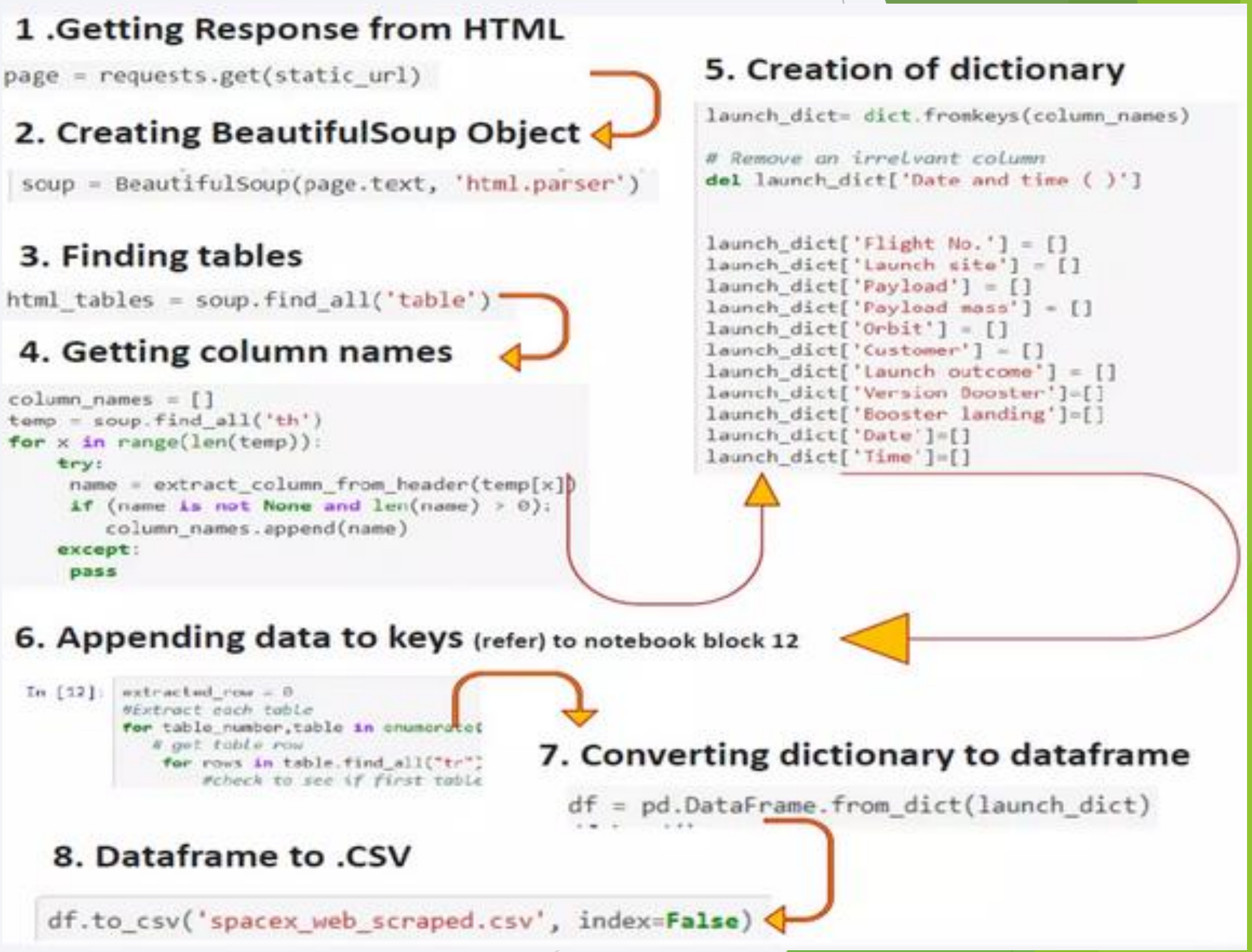
Data Collection – SpaceX API

- The API used is https://api.spacexdata.com/v4/rockets/
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- Every missing value in the data is replaced the mean the column that the missing value belongs to.
- We end up with 90 rows or instances and 17 columns or features.



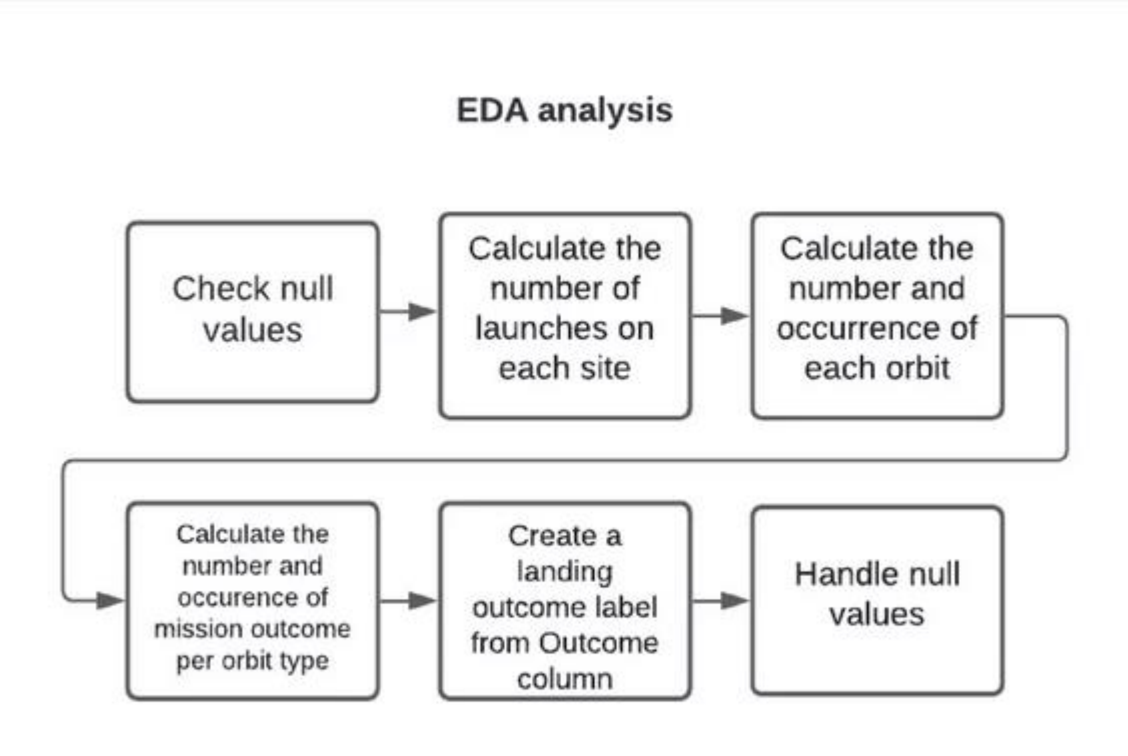
Data Collection - Scraping

- The data is scraped from : https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features.



Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.



EDA with Data Visualization

We use EDA to summarize the main characteristics of our data set and to find different patterns amongst the data. In order to summarize these characteristics, it is necessary to use statistical tools such as graphs. A list of the graphs plotted :

Relationships between groups using Scatter plots:

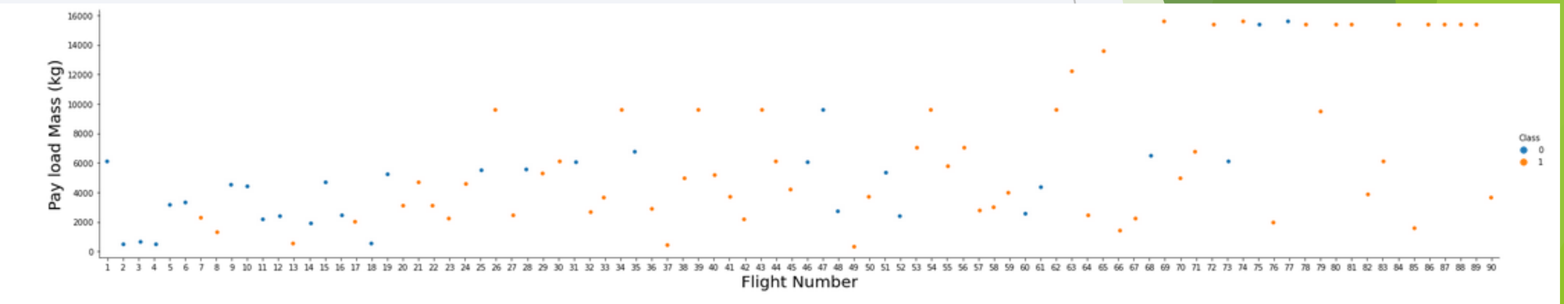
- Flight number and Payload Mass
- Flight number and Launch Site
- Payload and Launch Site
- Orbit and Flight number
- Payload and Orbit
- Orbit and Payload Mass

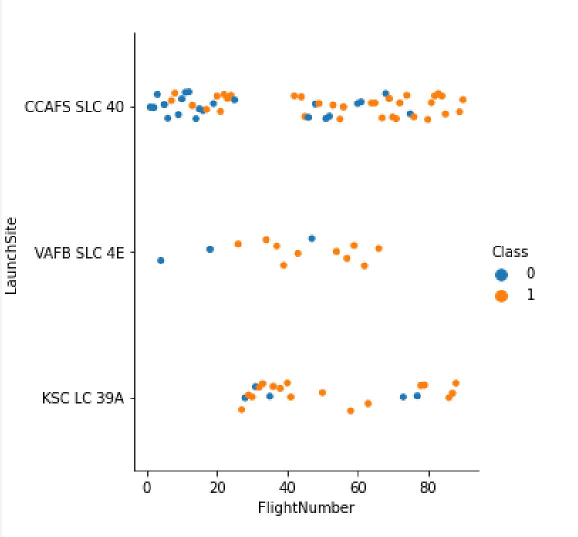
Patterns between groups using Bar plots

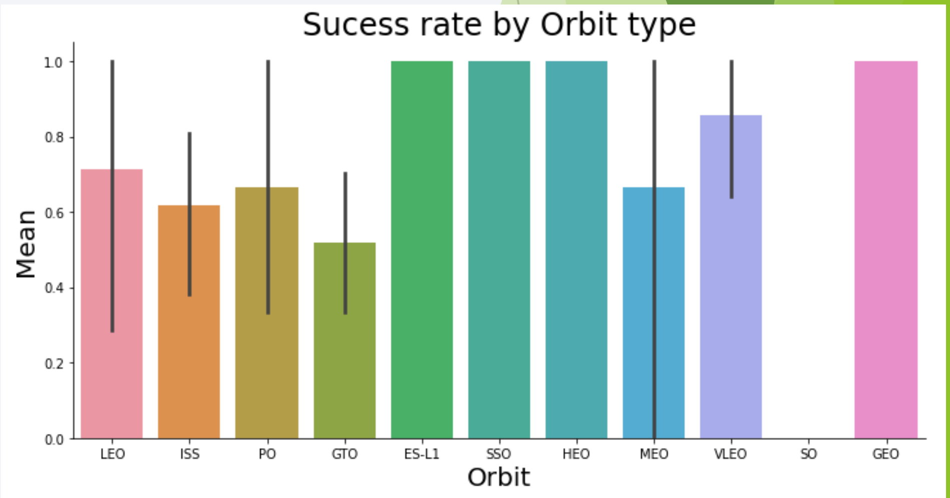
- Orbits and the mean

Success rates using Line plots

- Success rate by year





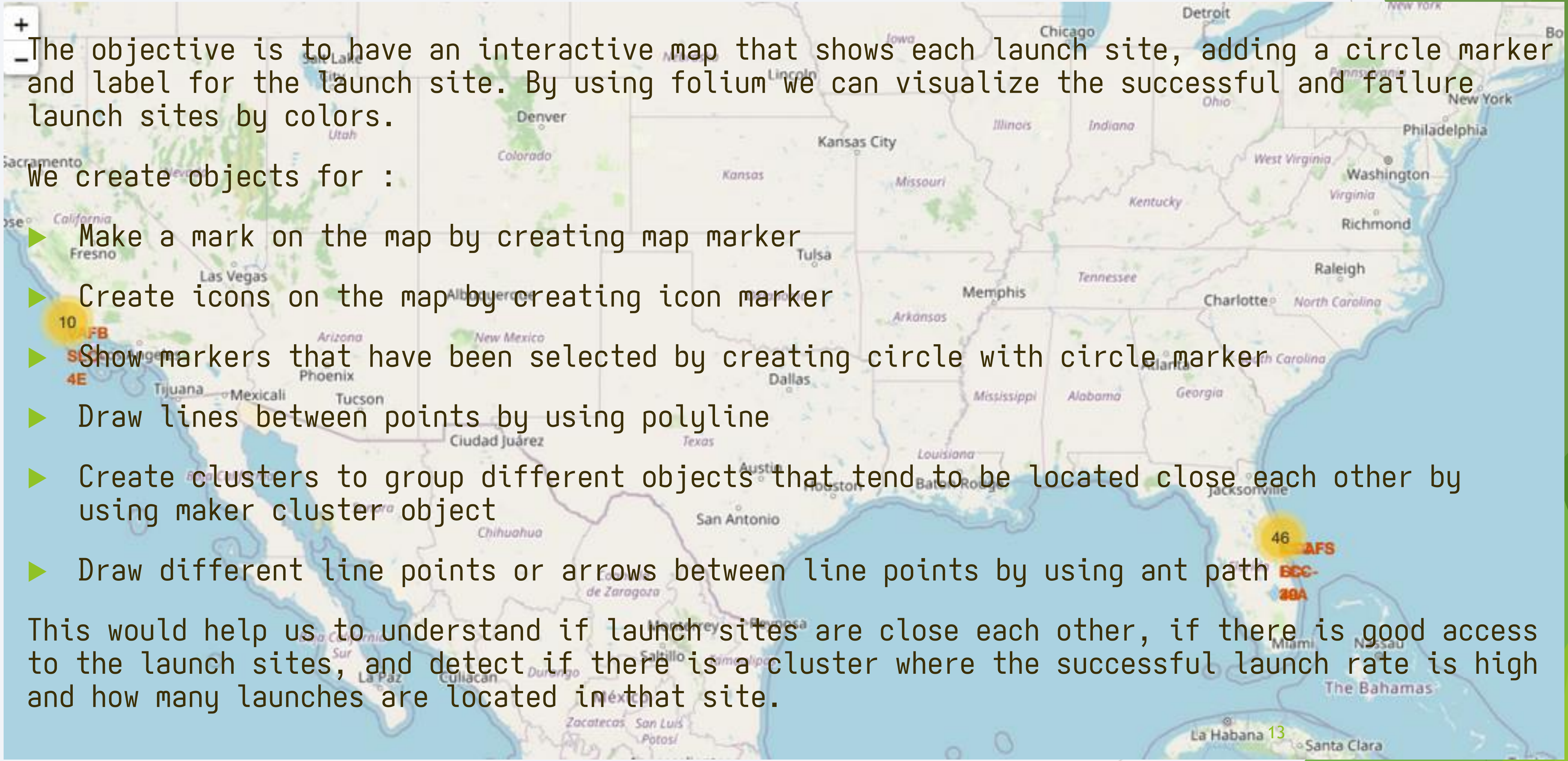


EDA with SQL

Summary of the SQL queries executed on IBM's DB2 database.

- The names of the unique launch sites in the space mission.
- 5 records where launch sites begin with 'CCA'.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by booster version F9 v1.1.
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium



The objective is to have an interactive map that shows each launch site, adding a circle marker and label for the launch site. By using folium we can visualize the successful and failure launch sites by colors.

We create objects for :

- Make a mark on the map by creating map marker
- Create icons on the map by creating icon marker
- Show markers that have been selected by creating circle with circle marker
- Draw lines between points by using polyline
- Create clusters to group different objects that tend to be located close each other by using maker cluster object
- Draw different line points or arrows between line points by using ant path

This would help us to understand if launch sites are close each other, if there is good access to the launch sites, and detect if there is a cluster where the successful launch rate is high and how many launches are located in that site.

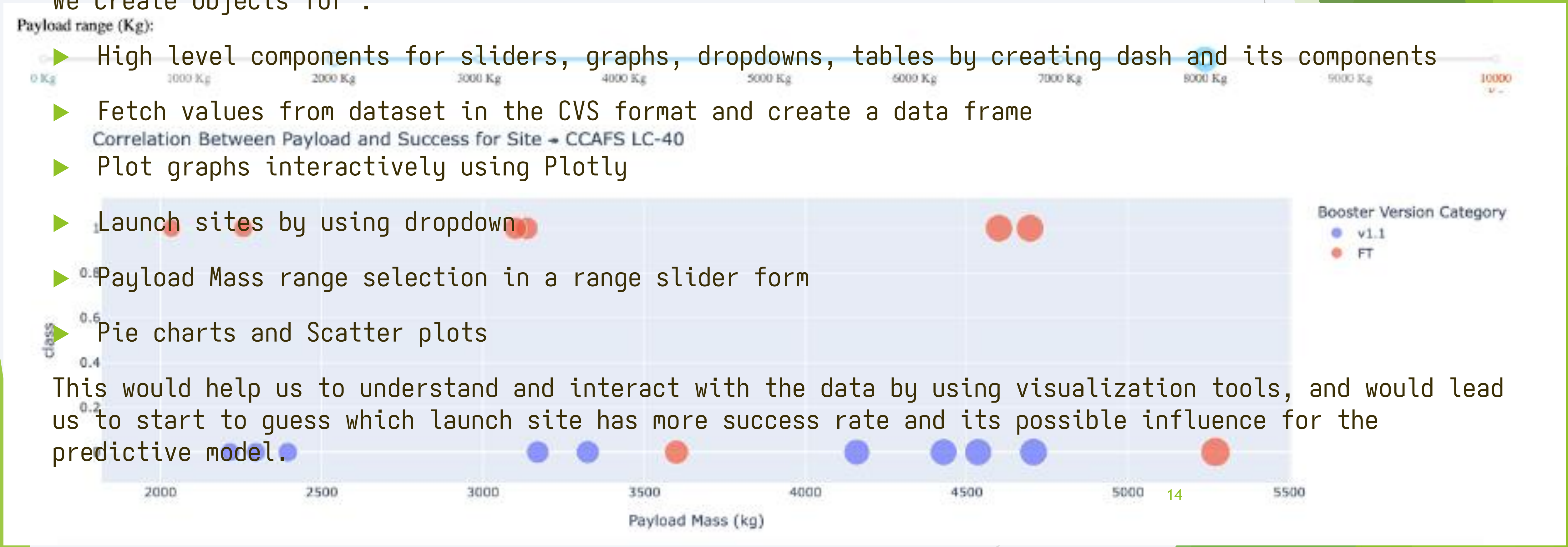
Build a Dashboard with Plotly Dash

The objective is to have an interactive dashboard with two different plots, a pie chart and a scatter plot. The former will show the percentage of success in relation with the launch sites while the later the relationship between the success rate and the booster version category.

We Create objects for :

- High level components for sliders, graphs, dropdowns, tables by creating dash and its components
- Fetch values from dataset in the CVS format and create a data frame
- Plot graphs interactively using Plotly
- Launch sites by using dropdown
- Payload Mass range selection in a range slider form
- Pie charts and Scatter plots

This would help us to understand and interact with the data by using visualization tools, and would lead us to start to guess which launch site has more success rate and its possible influence for the predictive model.



Predictive Analysis (Classification)

Functions from the Scikit-learn library are used to create our machine learning models.
The steps to build and train the models include :

- Building the model
  - Load dataset into a Pandas data frame
  - Transform and standardize the data
  - Split data into training and test
  - Select the type of machine learning algorithm which include (Logistic regression, Support vector machine (SVM), Decision tree, K nearest neighbors (KNN))
  - Fit the models on the training data set
  - Find the best combination of hyperparameters for each model
- Evaluating Model
  - Check accuracy for each model
  - Plot Confusion Matrix

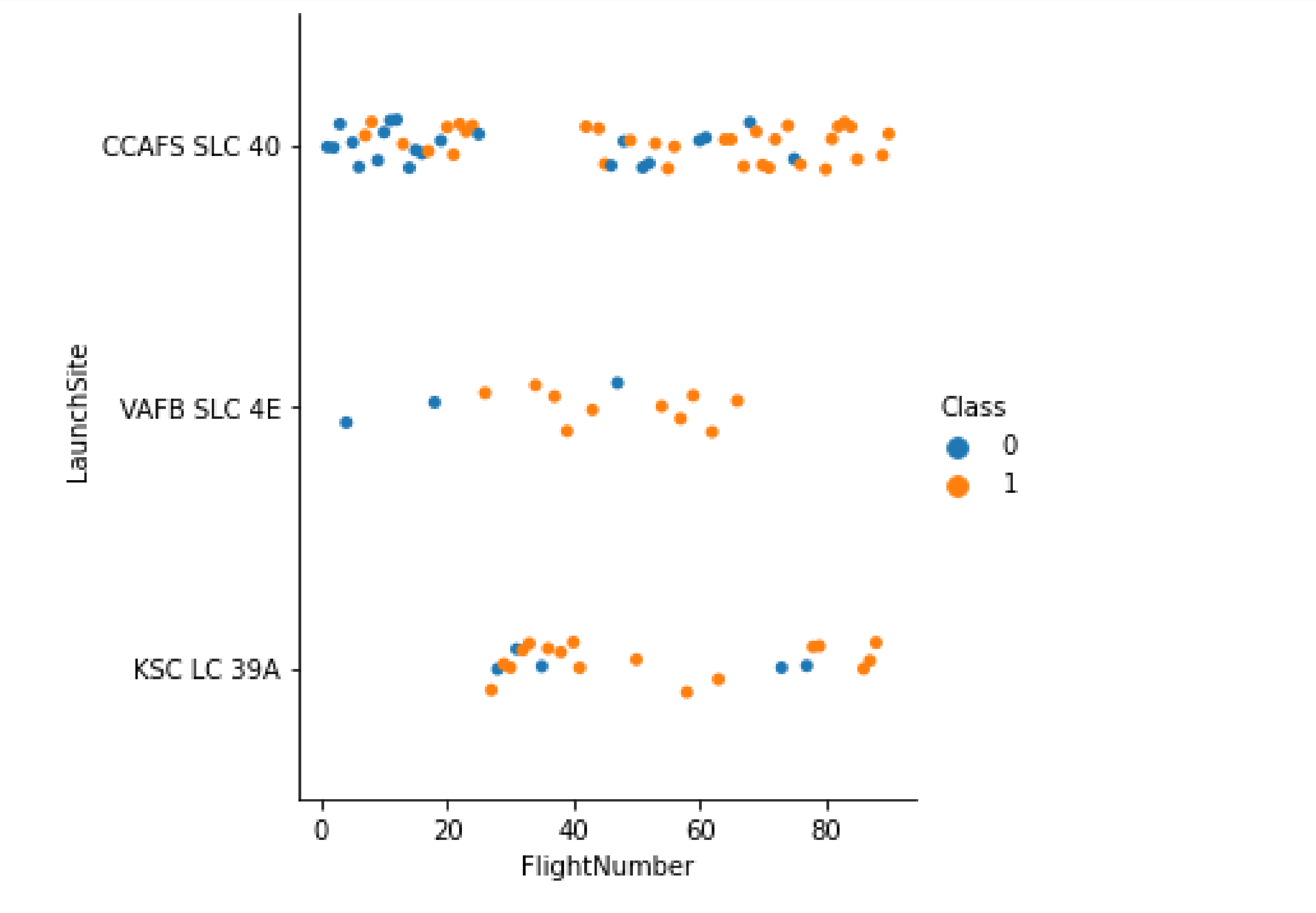
Results

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

Section 2

Insights drawn from EDA

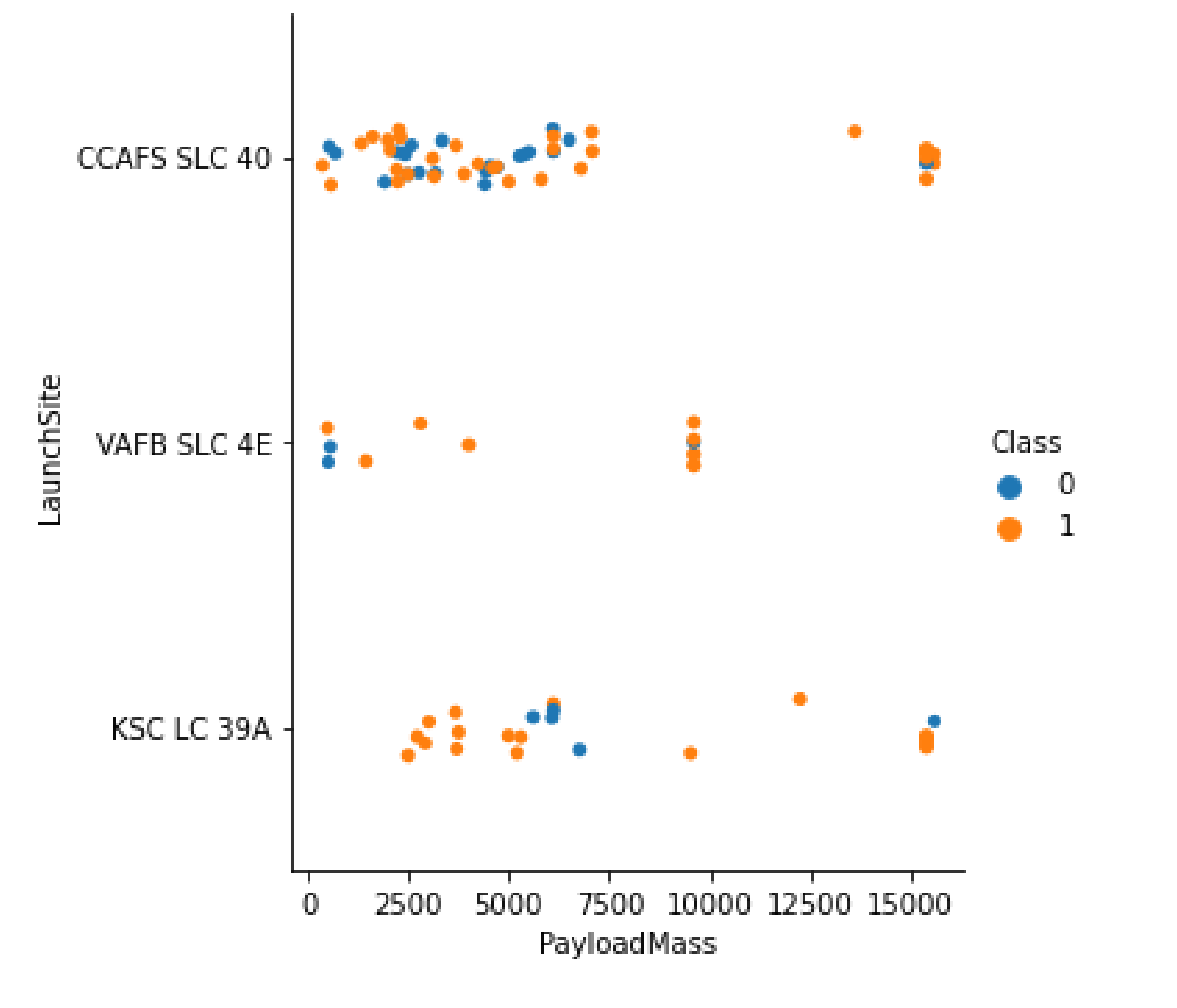
Flight Number vs. Launch Site



Launches from the site of CCAFS SLC 40 are significantly higher than the other launch sites.

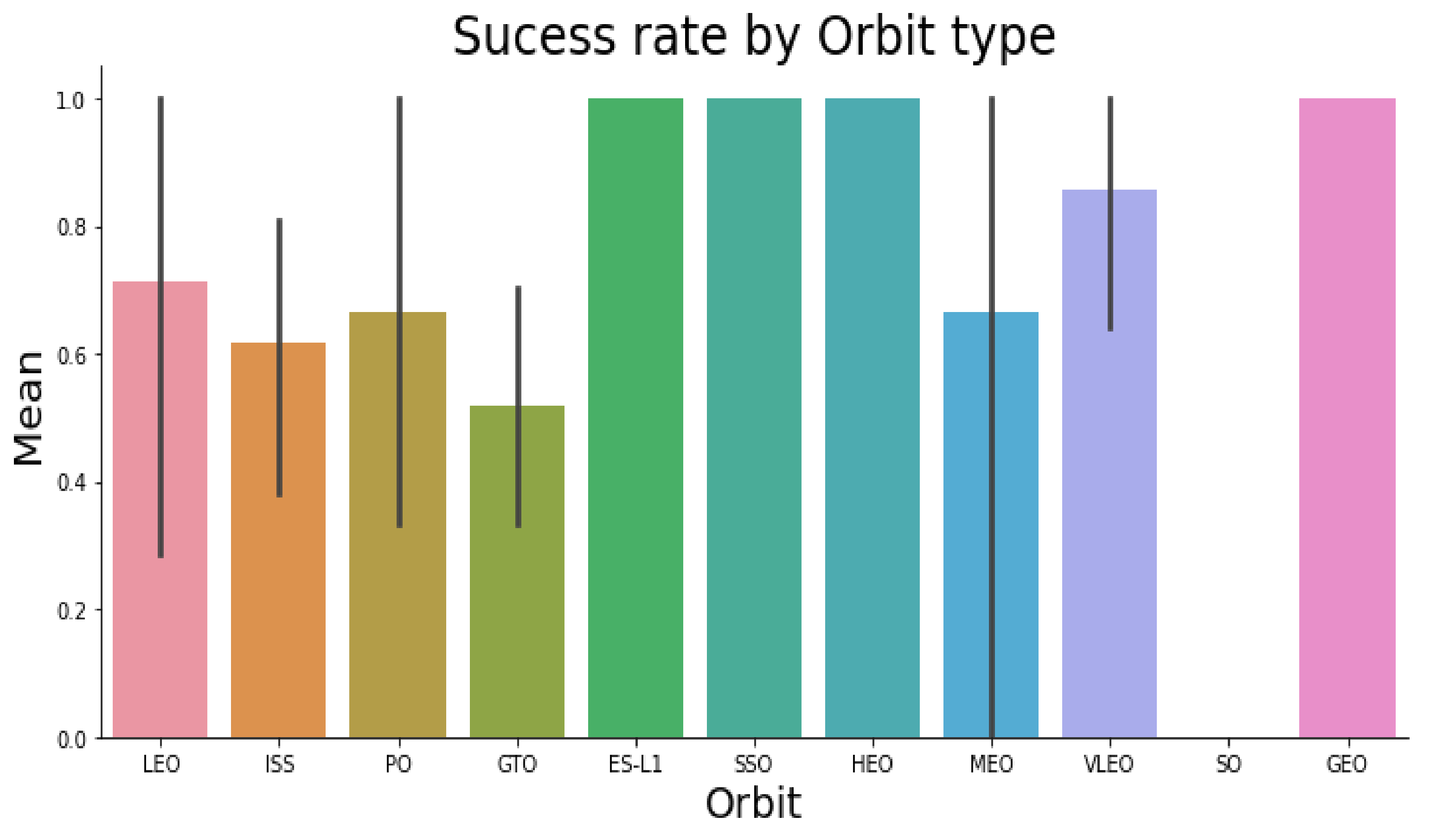
Payload vs. Launch Site

- VAFB-SLC launch site has no rockets launched with a heavy payload mass(greater than 10000).
- The majority of rockets launched with a payload less than 10000 have been launched from the CCAFS SLC 40 launch site.



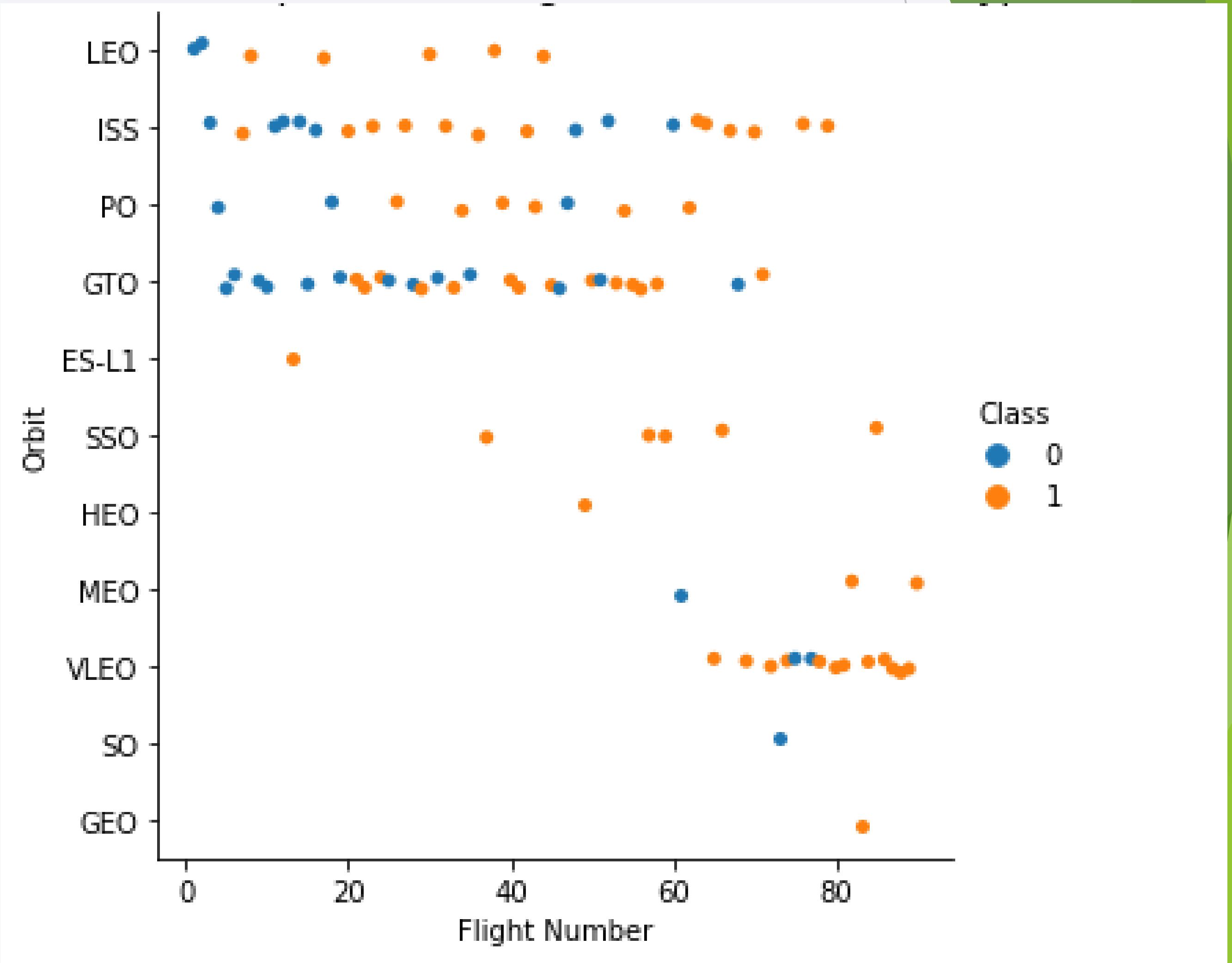
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have a success rate of 100%.
- SO has a success rate of 0%.



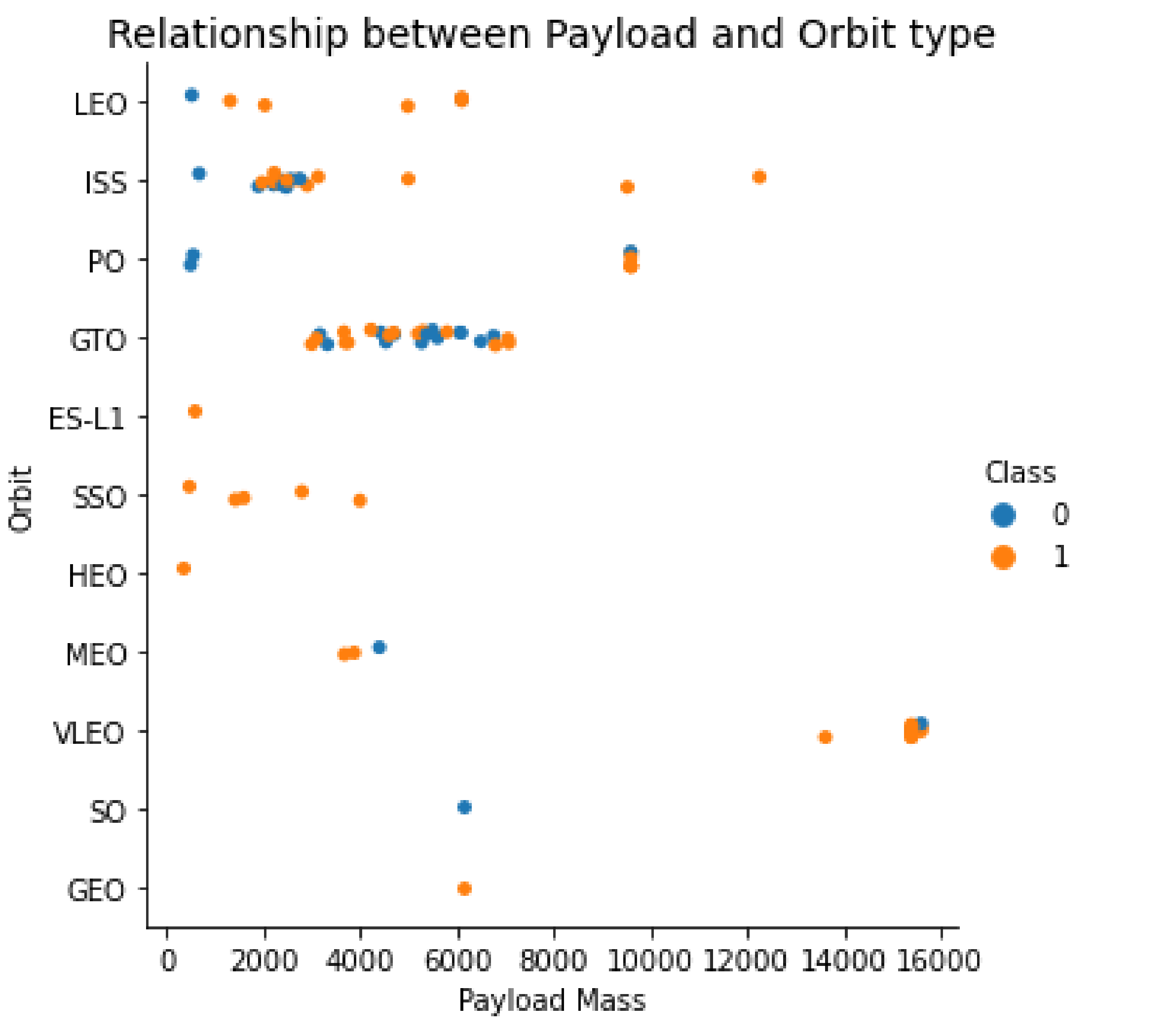
Flight Number vs. Orbit Type

- A trend can be observed of shifting to VLEO launches in recent years.



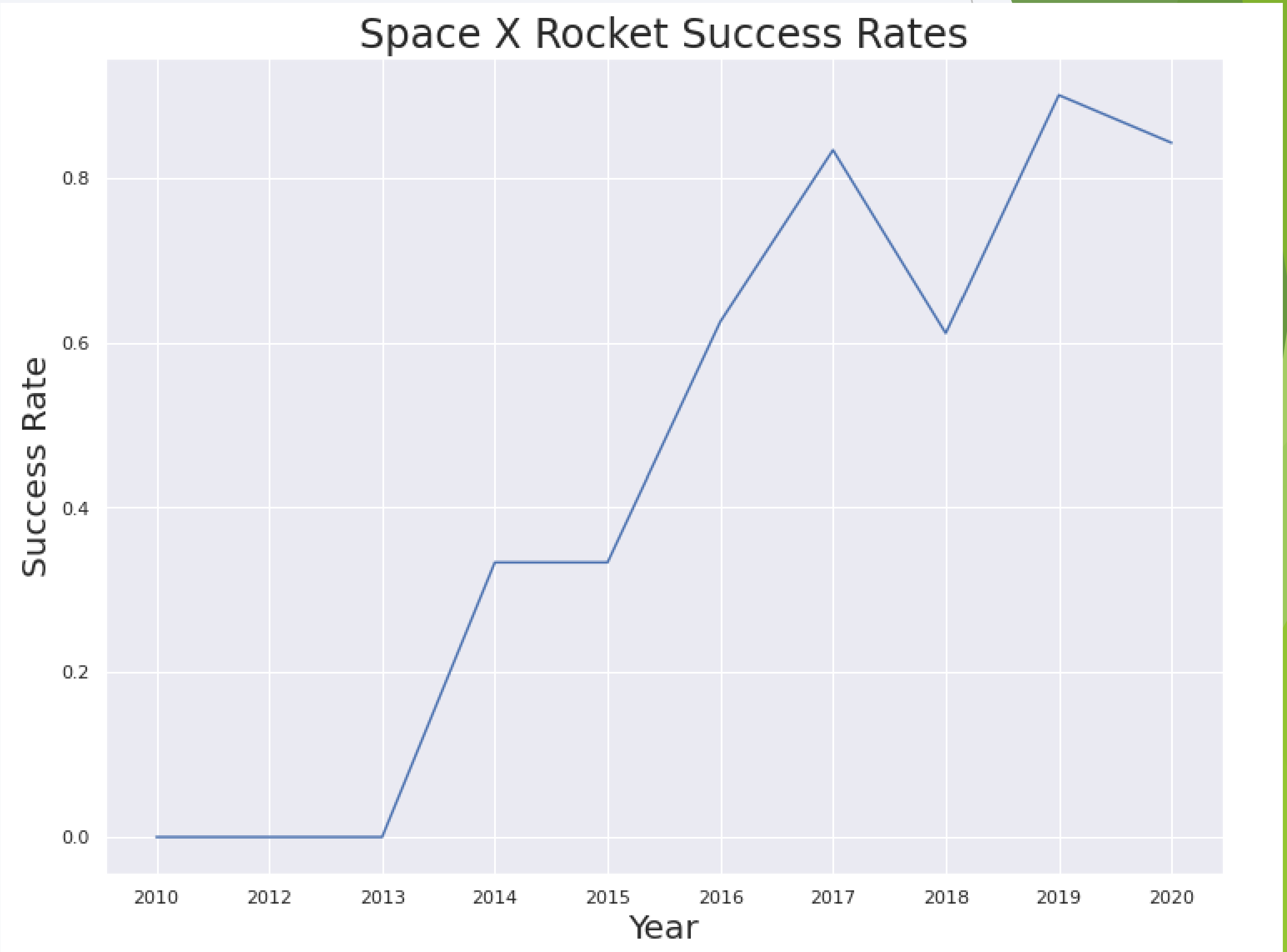
Payload vs. Orbit Type

- We can see the concentration of ~2500 payload mass at the ISS orbit, and payloads between 4000-7000 at the GTO orbit.



Launch Success Yearly Trend

- Since 2013 there has been a meteoric rise in the success rate.



All Launch Site Names

- The names of the unique launch sites in the space mission

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA’

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

- The total payload mass carried by boosters launched by NASA (CRS)

TOTAL PAYLOAD MASS BY NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 v1.1
2928

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad was achieved

First Succesful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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

- The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

- The names of the booster versions which have carried the maximum payload mass

BOOSTER VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS
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

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

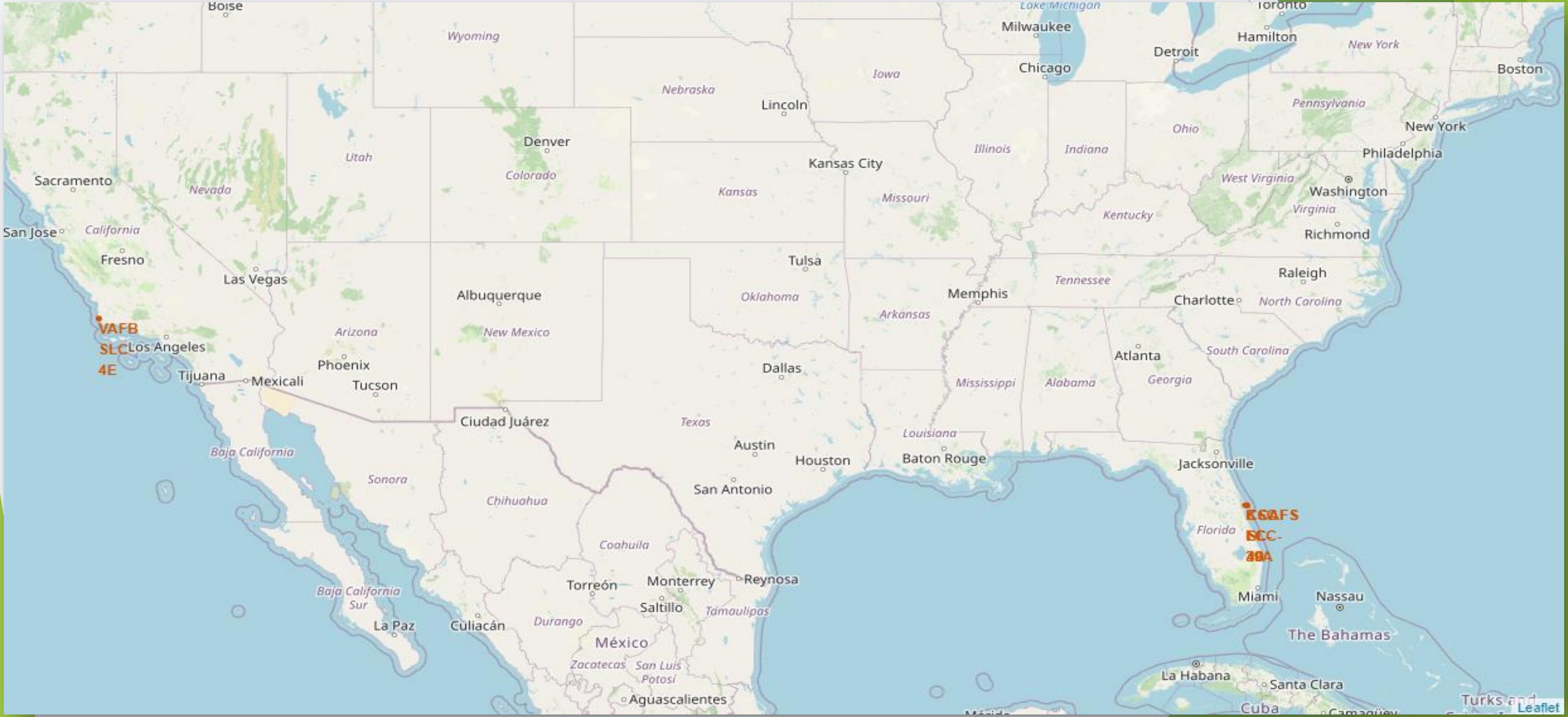
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

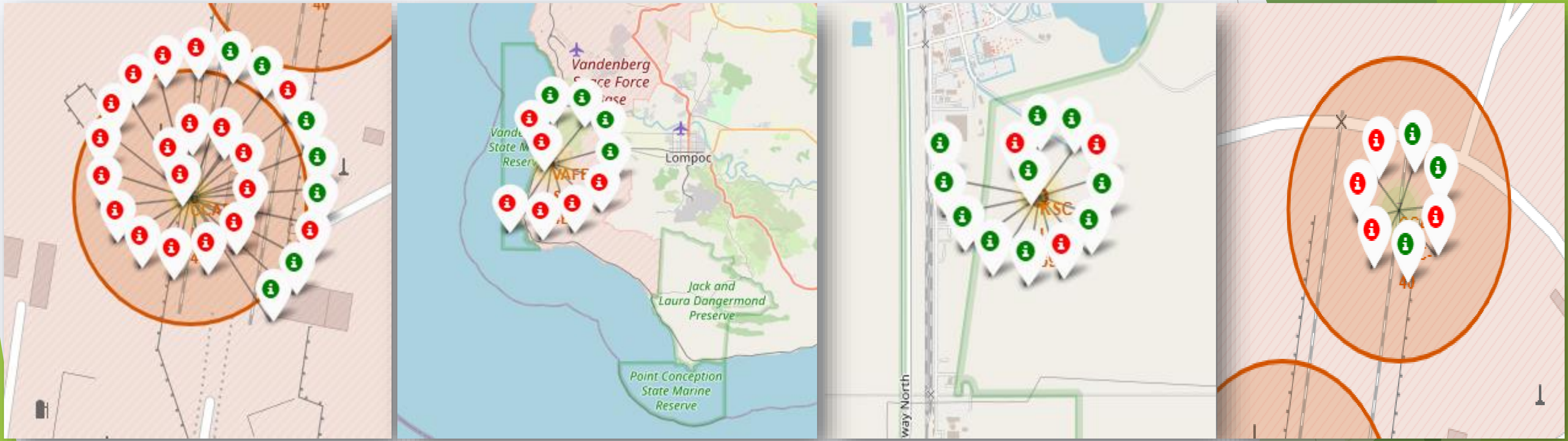
Section 3

Launch Sites Proximities Analysis

All launch sites on the map

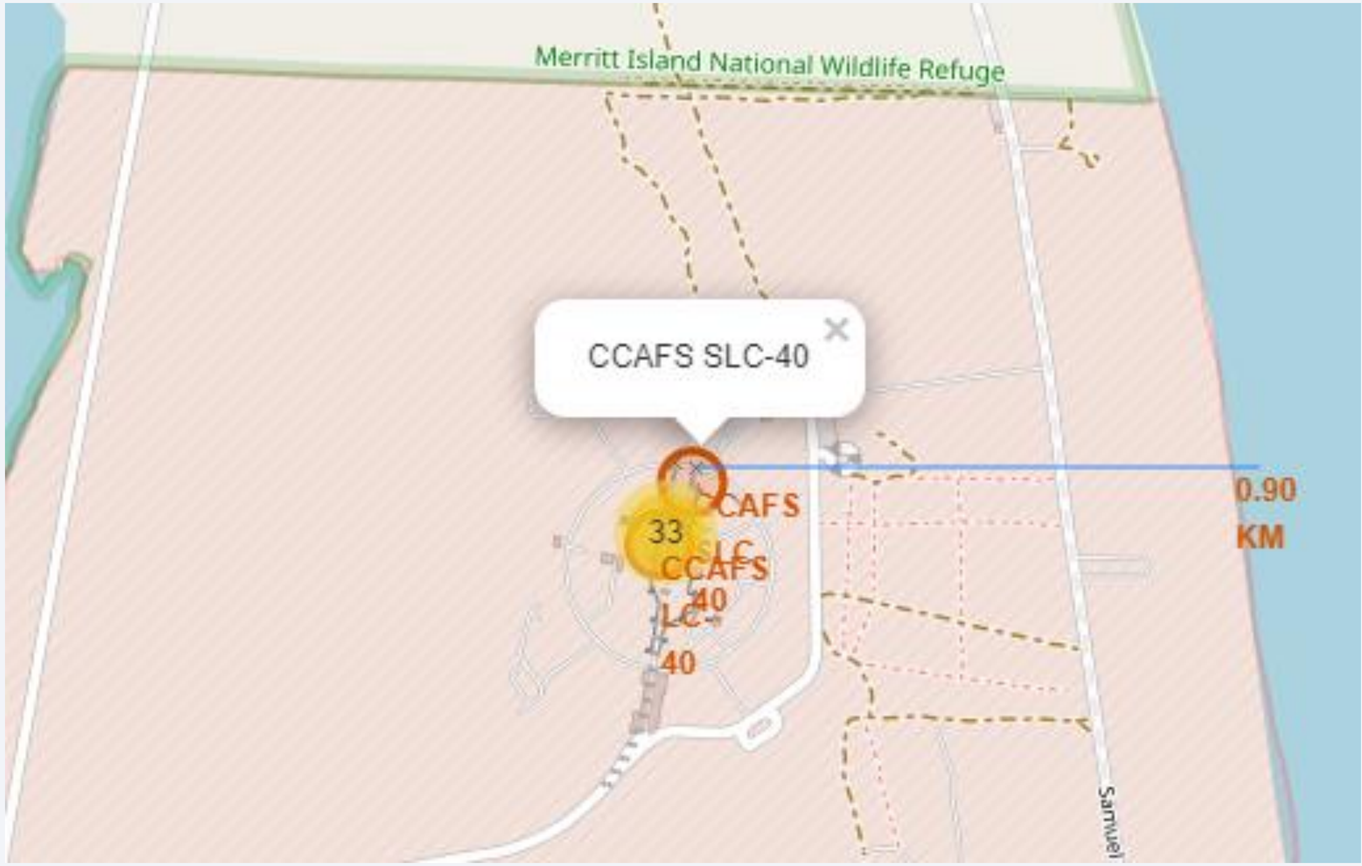


Success/Failed launches by site



- Launches, decoded by their outcome, across launch sites, in the following order:
  - CCAFS LC - 40
  - VAFB SLC – 4E
  - KSC LC – 39A
  - CCAFS SLC - 40

Distance between launch sites to its proximities

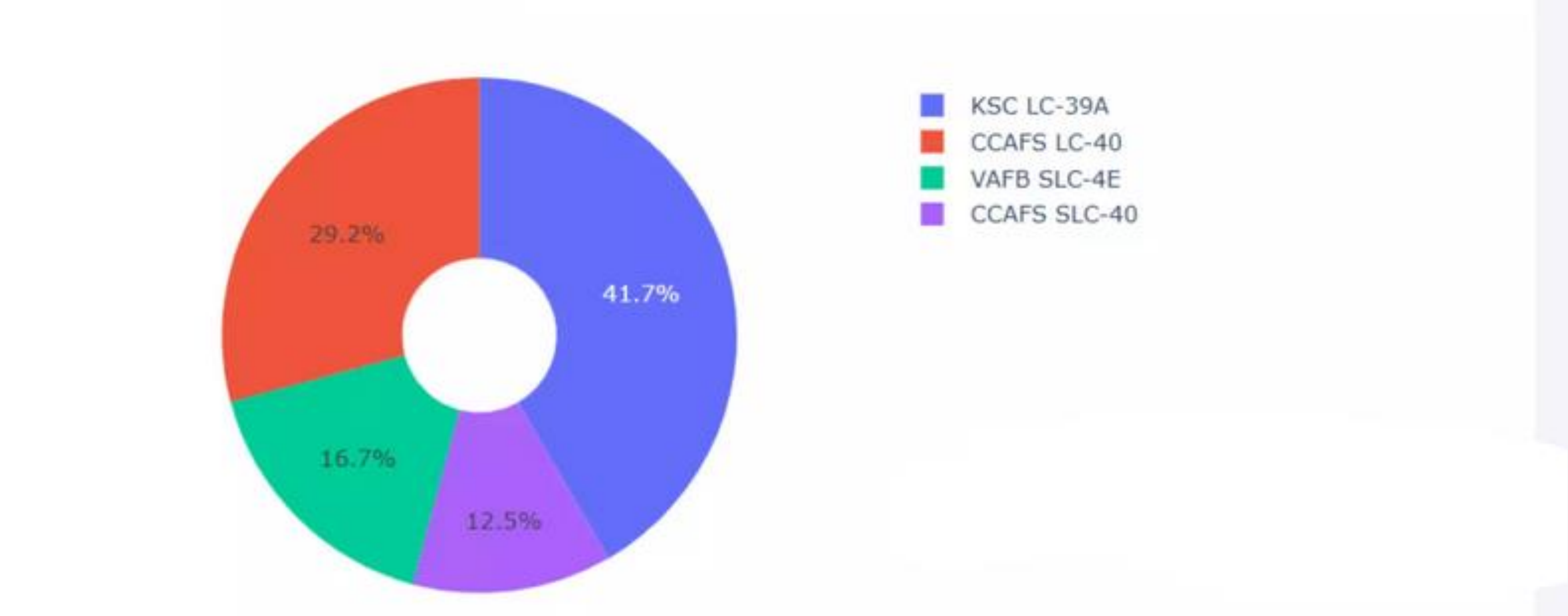


- The picture shows the distance between the CCAFS SLC-40 launch site and the nearest coastline.

Section 4

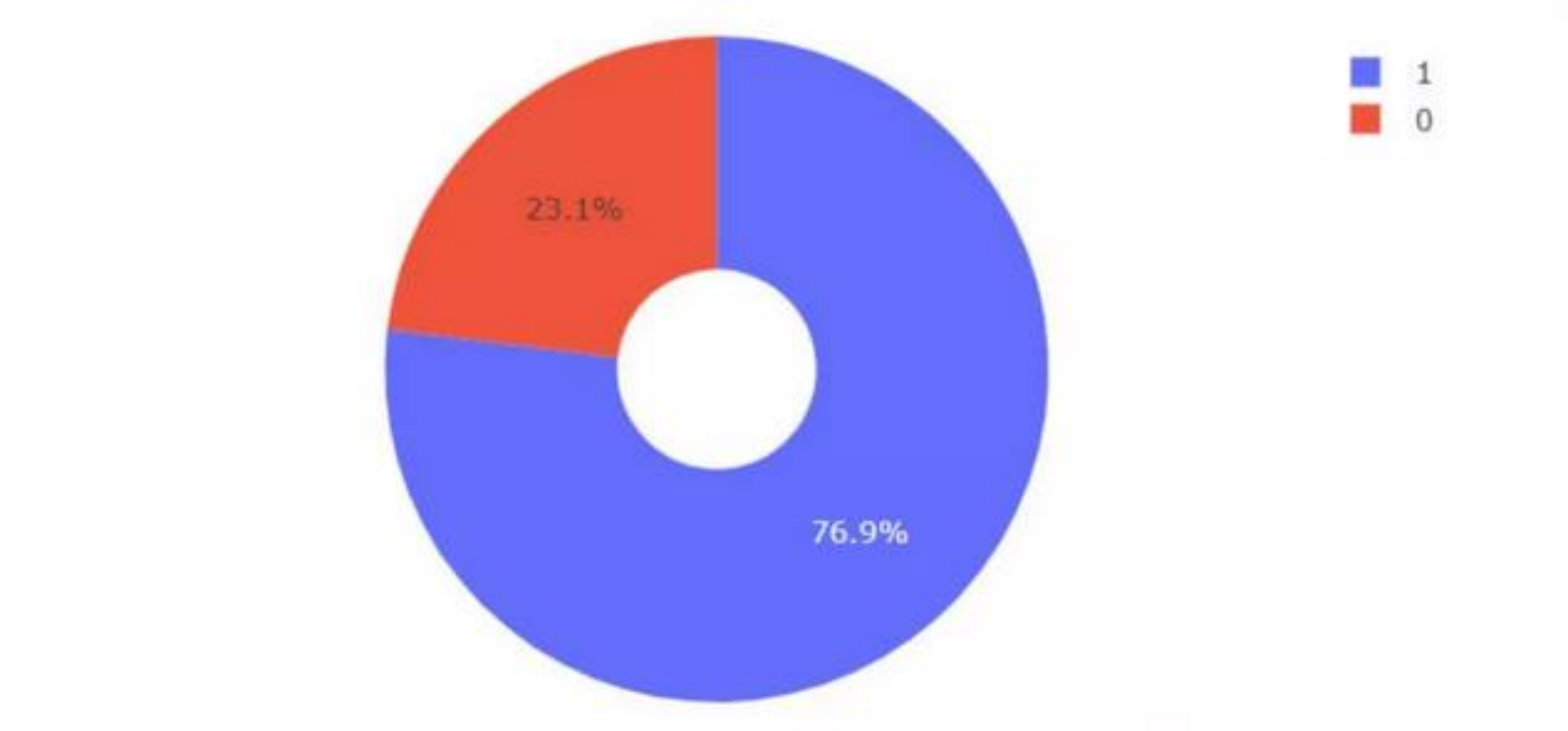
Build a Dashboard with Plotly Dash

Launch Success Count



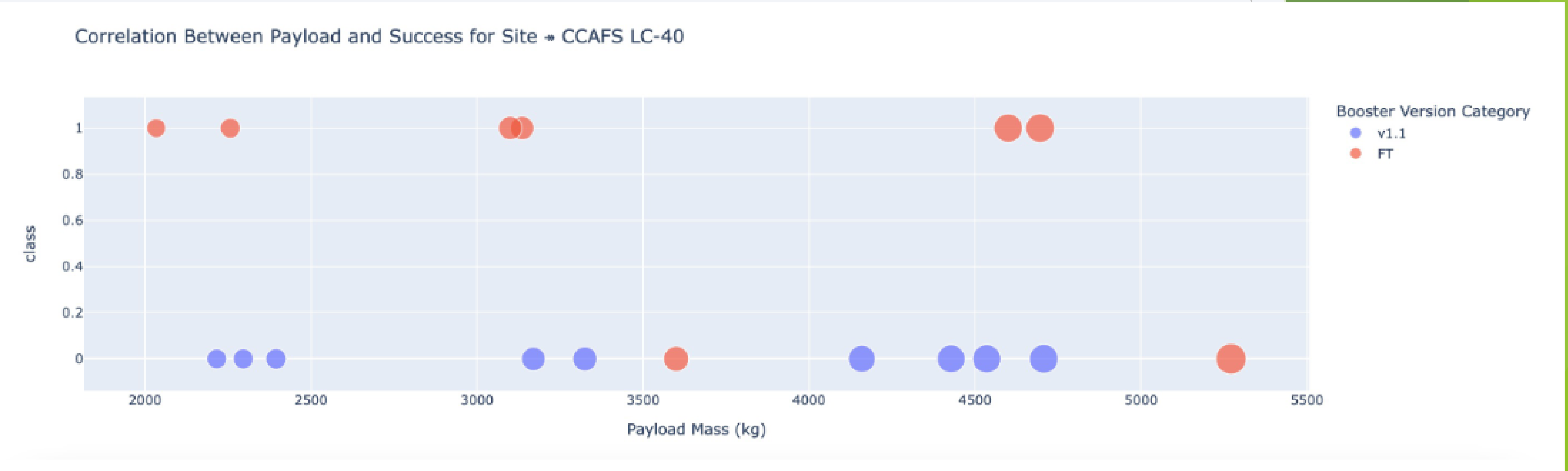
- KSC LC-39A has the highest success score with 41.70%.
- CCAFS LC-40 comes next with 29.20%.
- Finally, VAFB SLC-4E and CCAFS SLC-40 with 16.70% and 12.50% respectively.

Success rate by launch site



- LSC LC-39A has the highest score with 76.9% with payload range of 2000kg – 10000kg, and FT booster version has the highest score.

Payload vs. Launch Outcome

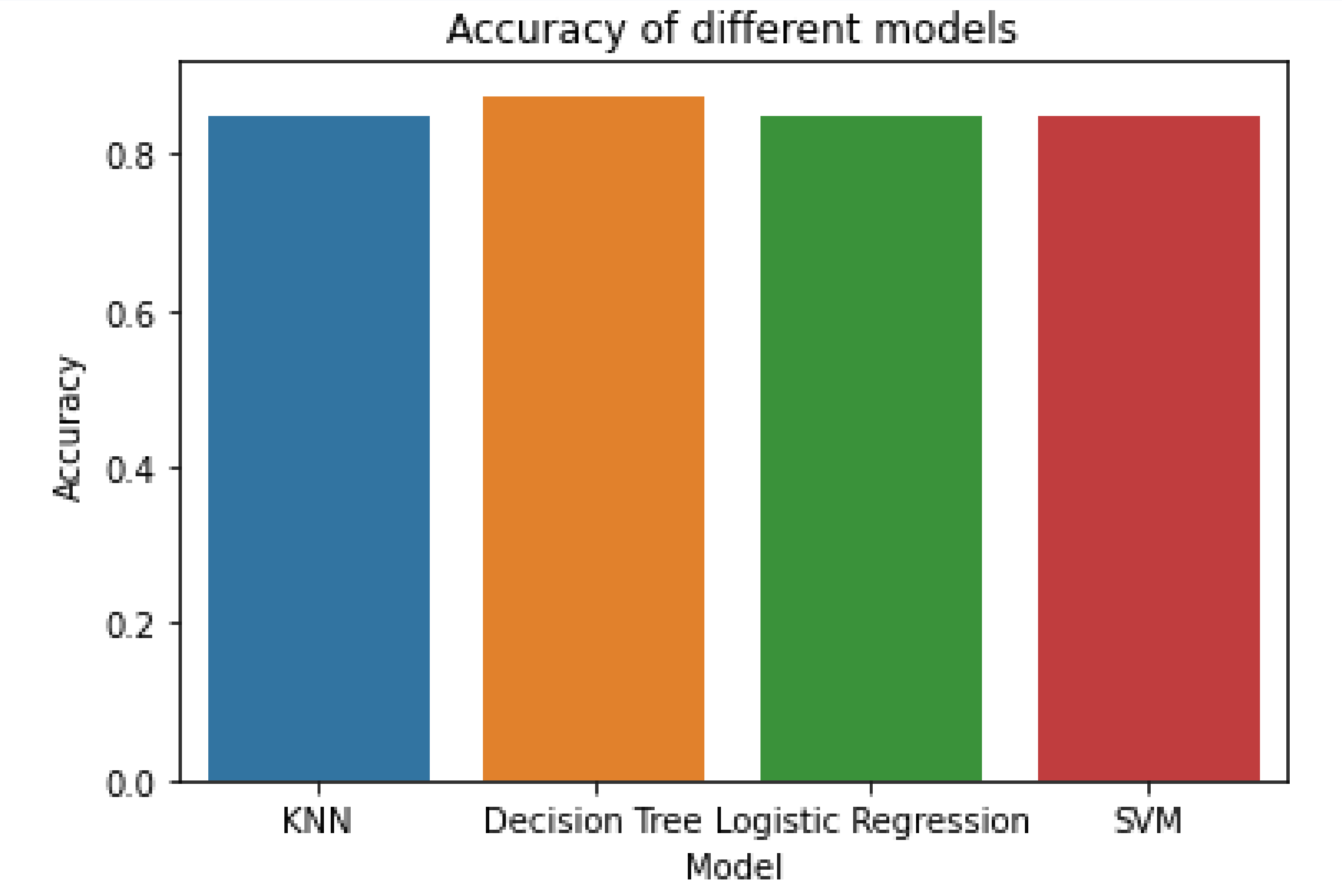


- We can see the success rates for low weighted payloads is higher than the heavy weighted payloads.

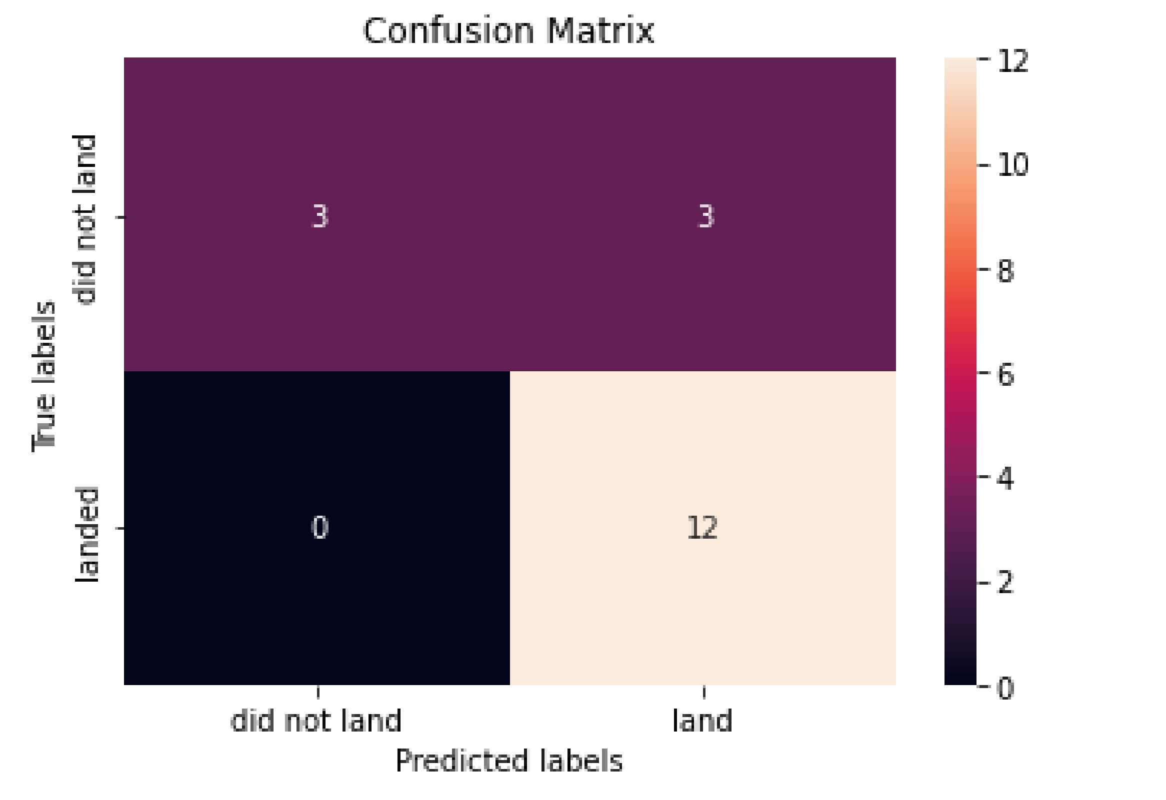
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



- Put side by side all the models have the same accuracy and confusion matrix on the test data.
- Therefore, we use their GirdSearchCV scores to rank them. Based on that Decision Tree Classifier has the best accuracy off the 4 models.
  The accuracies are :
  Logistic Regression - 0.846429
  SVM - 0.848214
  KNN - 0.848214
  Decision Tree - 0.875000

Conclusions

- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed, with an accuracy of 87.5%.
- Launch site KSC LC 39A, and orbits GEO, HEO, SSO, ES L1 have the highest success rates.

Thank you!