IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Carson Xie>

<2023-04-10>

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection using Api and web scraping
  - Data wrangling -
  - EDA with Data Visualization and SQL
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction and summary results
- **Summary of all results**
  - To Predict the success rate of Spacex's rocket first stage landing, therefore reduce total cost of launching.
  - To Determine which features have most impact on predicting the cost.(location/etc.)
  -

Introduction

- Project background and context: The goal is to assess if Space Y can compete with SpaceX, who offers affordable space travel by reusing the first stage of its Falcon 9 rockets. We will predict if SpaceX will reuse the first stage using public data and machine learning.

- Problems you want to find answers:
  - The success rate of landings of the first stage of rockets in different type, site.
  - what are the factors affect the success rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using API: https://api.spacexdata.com
  - Web scraping from wiki page Space X Falcon 9
- Perform data wrangling
  - Calculate the number of launches on each site, different orbit type statistic
  - Create a landing outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

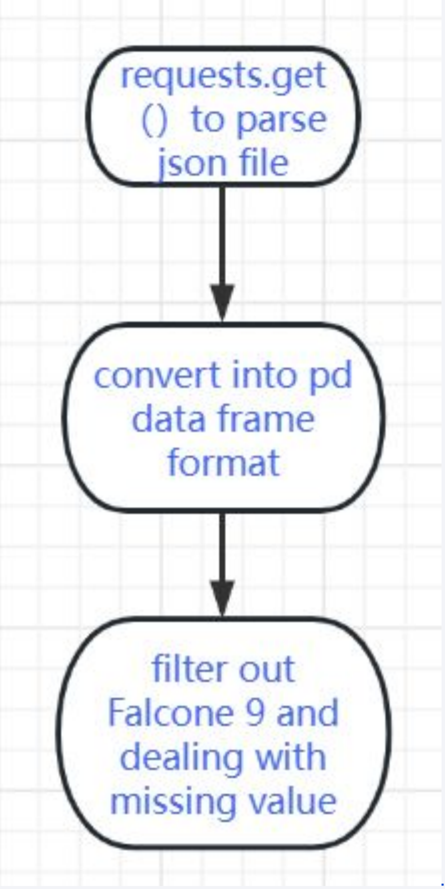
Data Collection

- Data sets were collected using API or web scraping
- API: using request package, Space X API will return json type file (https://api.spacexdata.com/v4/rockets/)
- Web scraping: from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches).

Data Collection – SpaceX API

- Flowcharts of Data collection with SpaceX REST calls:

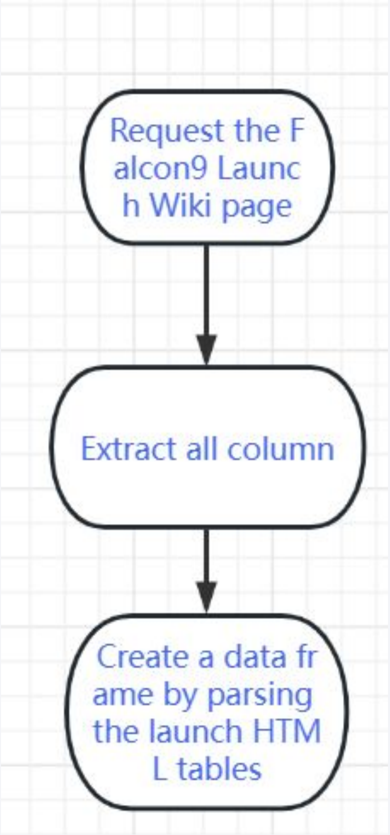
- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb


requests.get () to parse json file
convert into pd data frame format
filter out Falcone 9 and dealing with missing value

Data Collection - Scraping

- Another way to download data from Wikipedia page in html format then parse by beautifulsoup

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/2_labs-jupyter-spacex-scraping.ipynb

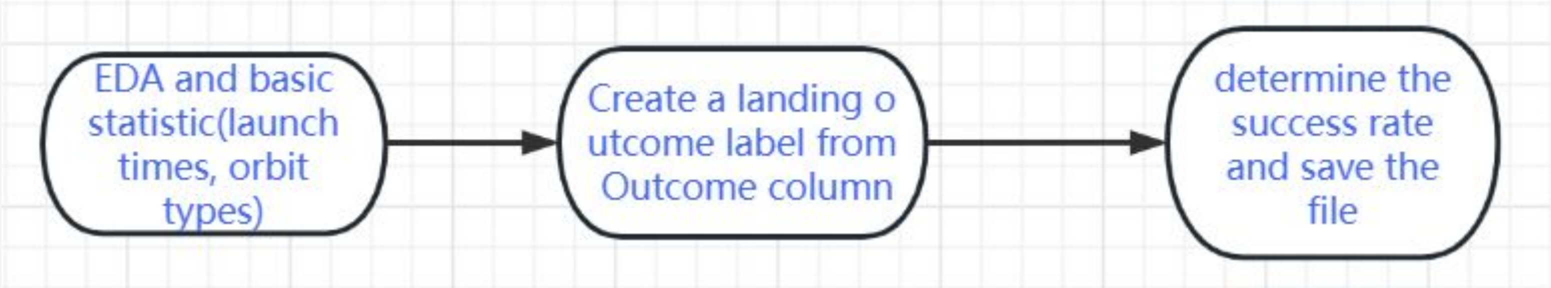


Request the Falcon9 Launch Wiki page

Extract all column

Create a data frame by parsing the launch HTML tables

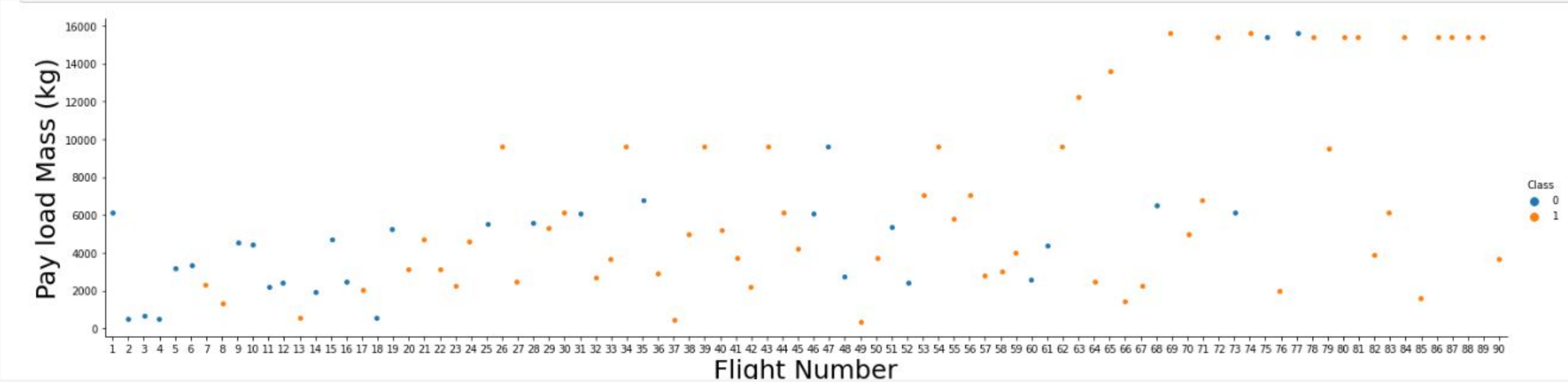
Data Wrangling

- First Perform exploratory Data Analysis and determine Training Labels.
- 
EDA and basic statistic(launch times, orbit types)
Create a landing o utcome label from Outcome column
determine the success rate and save the file

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots and bar plots were used to visualize the data to discover the correlation and trend between features
- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/4_jupyter-labs-eda-dataviz.ipynb



EDA with SQL

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

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/5_jupyter-labs-eda-sql-.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were added in Folium Maps
- Markers used for different launch sites
- Circles indicate NASA Johnson Space Center's coordinate with a popup label
- Marker clusters show success/failed launches for each site on the map
- Lines are used to indicate distances between two coordinates.

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/6_lab_jupyter_launch_site_location_Folium.ipynb

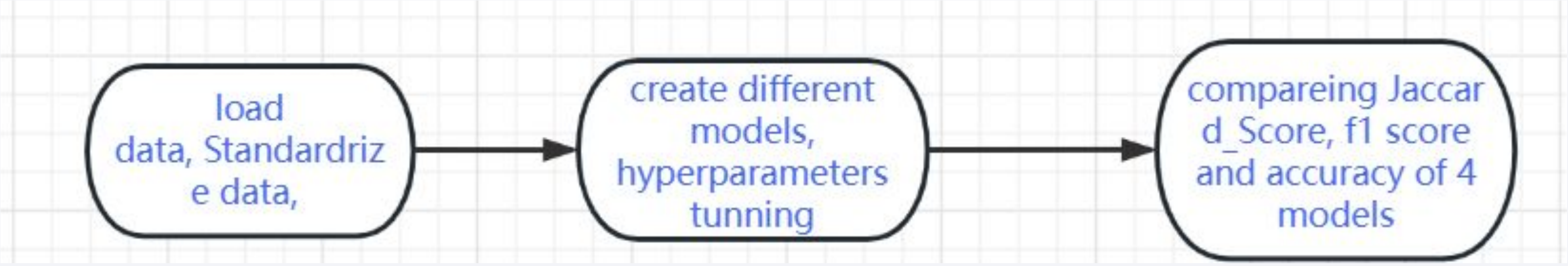
Build a Dashboard with Plotly Dash

- The following graphs and plots were used in the dash
  - Pie chart about Percentage of launches by site
  - Payload range and correlation between payload and success

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models we train 4 models: logistic regression, support vector machine, decision tree and knn.

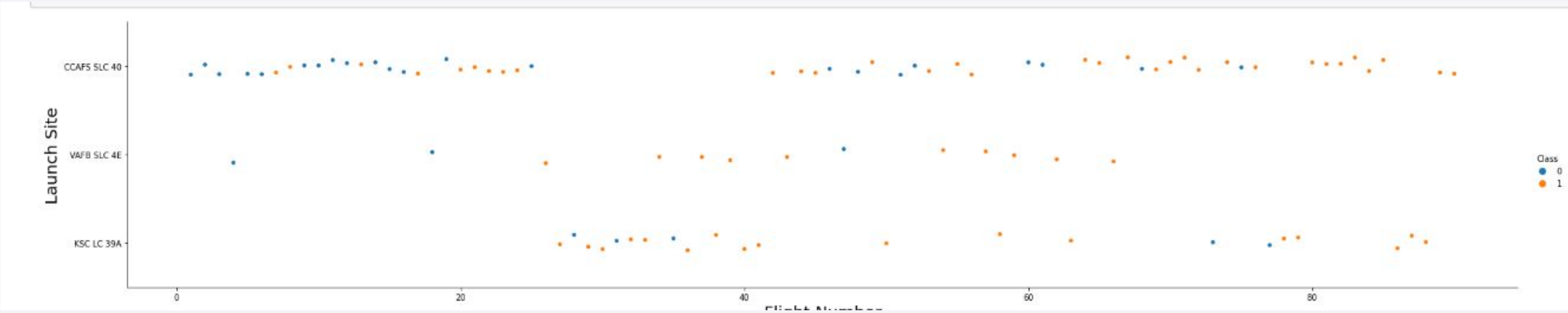

load data, Standardrize data,
create different models, hyperparameters tunning
compareing Jaccard_Score, f1 score and accuracy of 4 models

- https://github.com/carsonxie/IBM-Applied-Data-Science-Capstone-Project/blob/main/7_SpaceX_Machine%20Learning%20Prediction.ipynb

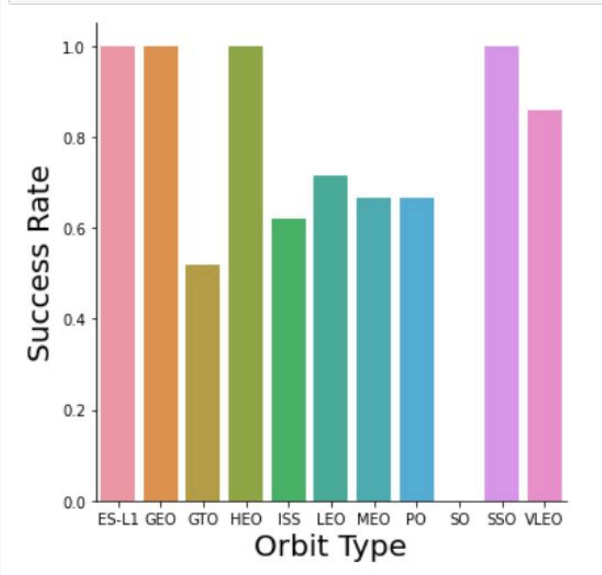
Results

- Exploratory data analysis results

correlation plot:



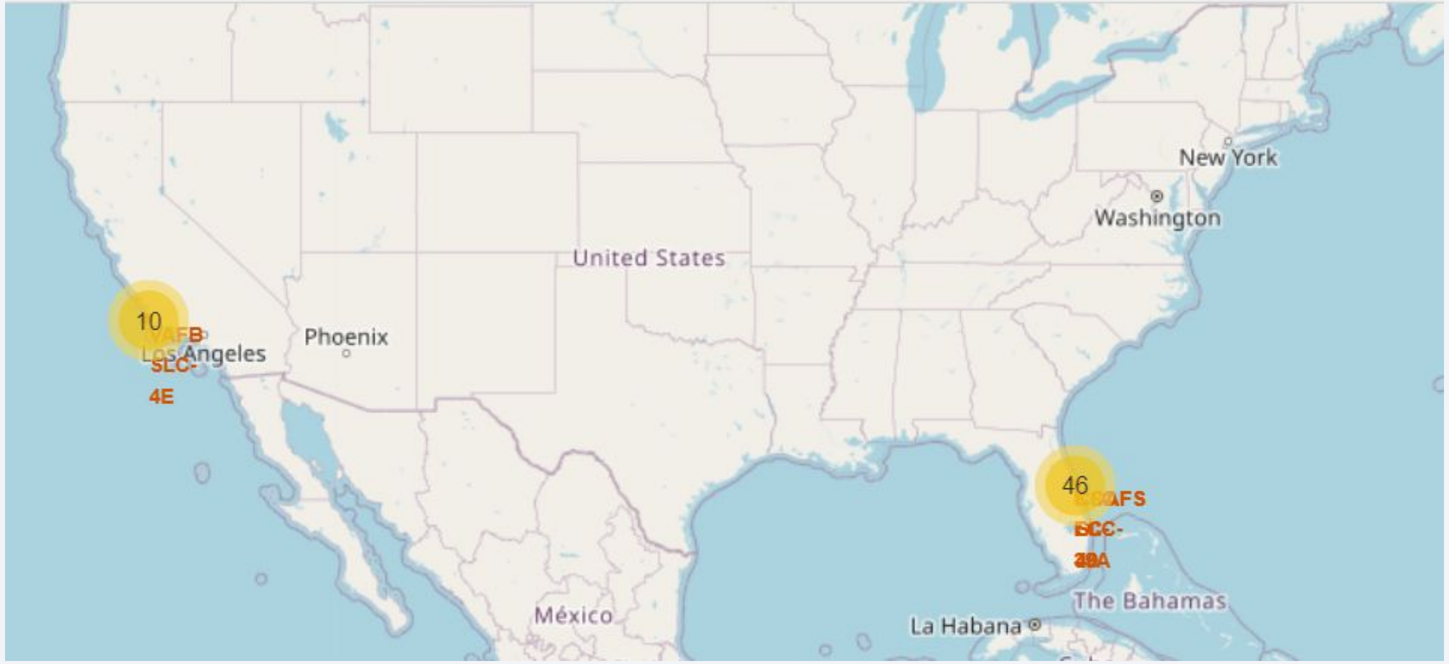
orbit type vs success rate:



Results

- Interactive analytics demo in screenshots

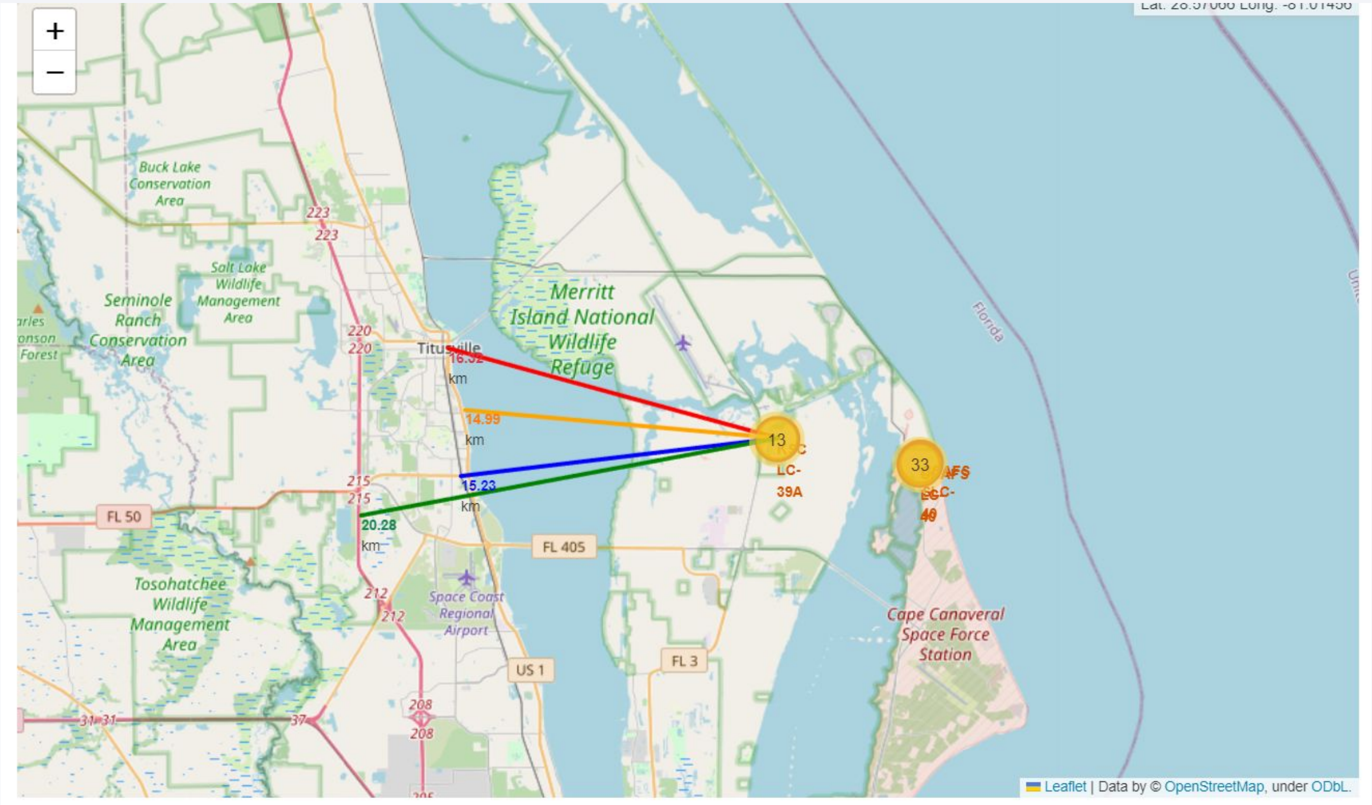
`different launch sites`



Results

- Interactive analytics demo in screenshots

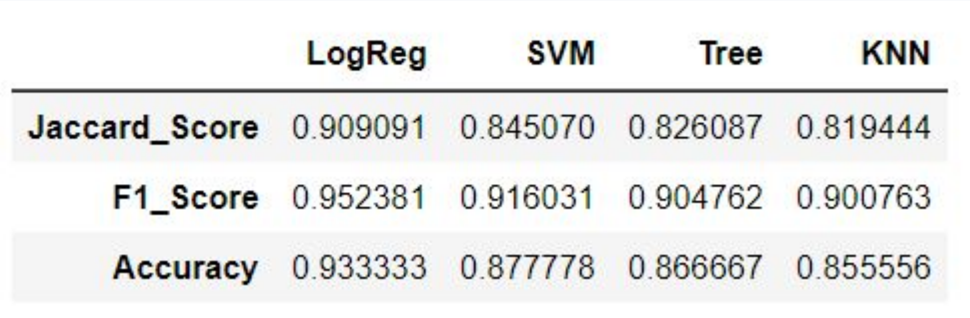
launch site to the closest coastline


Lat: 28.57066 Long: -81.01456
+
−
Buck Lake Conservation Area
Salt Lake Wildlife Management Area
Seminole Ranch Conservation Area
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
LC-39A
33
AFS
LC-40
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Cape Canaveral Space Force Station
Florida
FL 50
FL 405
FL 3
US 1
223
220
215
212
208
37
31
Leaflet | Data by © OpenStreetMap, under ODbL.

Results

- Predictive analysis results

We can see that overall logistic regression outperforme other 3 models in terms of jaccard score, f1 score and accuracy



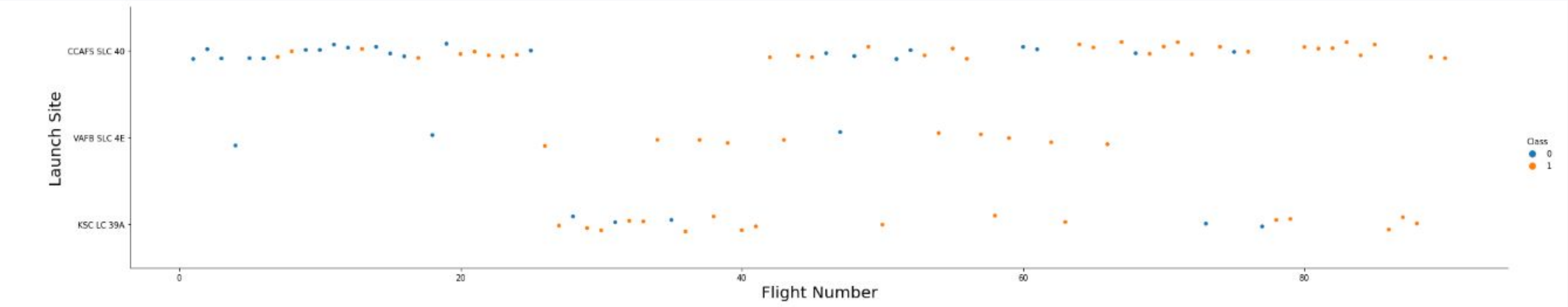
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.909091	0.845070	0.826087	0.819444
F1_Score	0.952381	0.916031	0.904762	0.900763
Accuracy	0.933333	0.877778	0.866667	0.855556

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

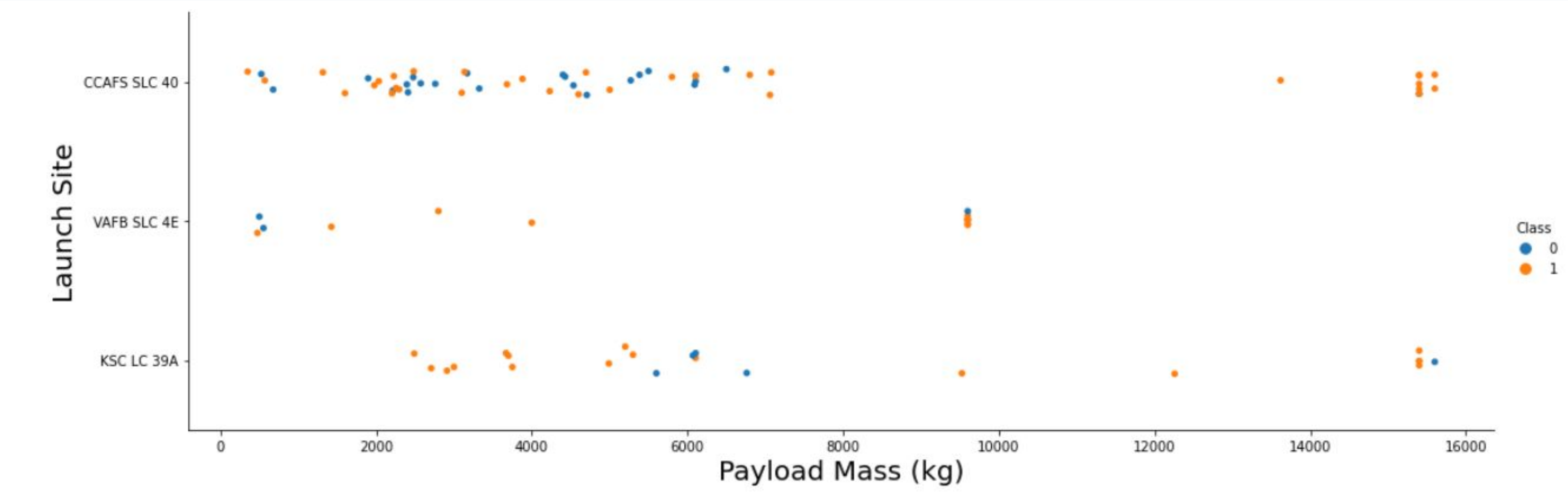
- Show a scatter plot of Flight Number vs. Launch Site

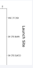


- We can see that in CCAF5 SLC 40 location, most of launches were successful

Payload vs. Launch Site

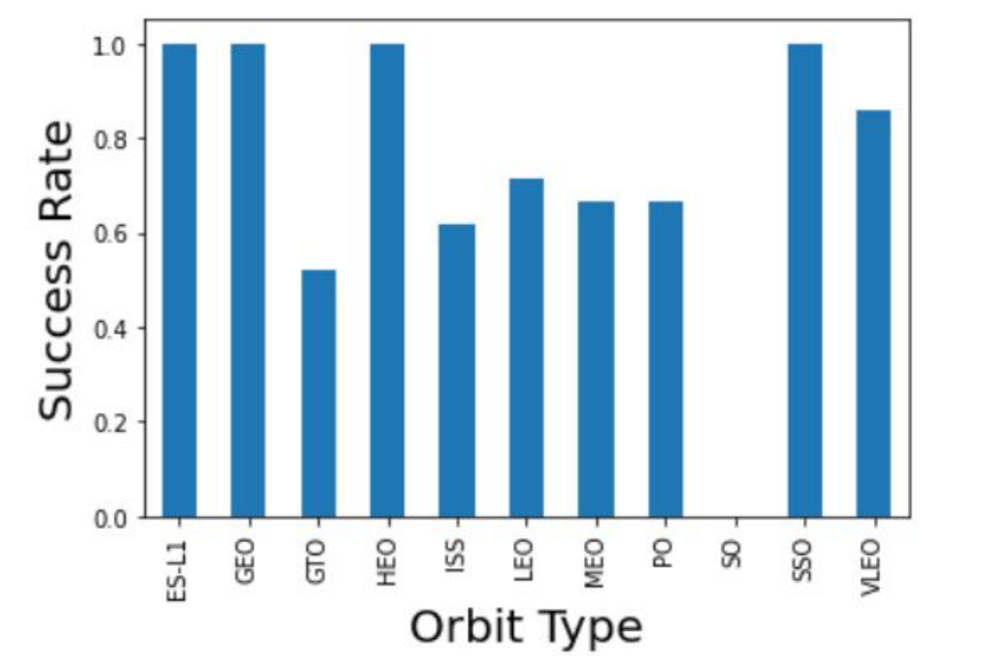
- Show a scatter plot of Payload vs. Launch Site





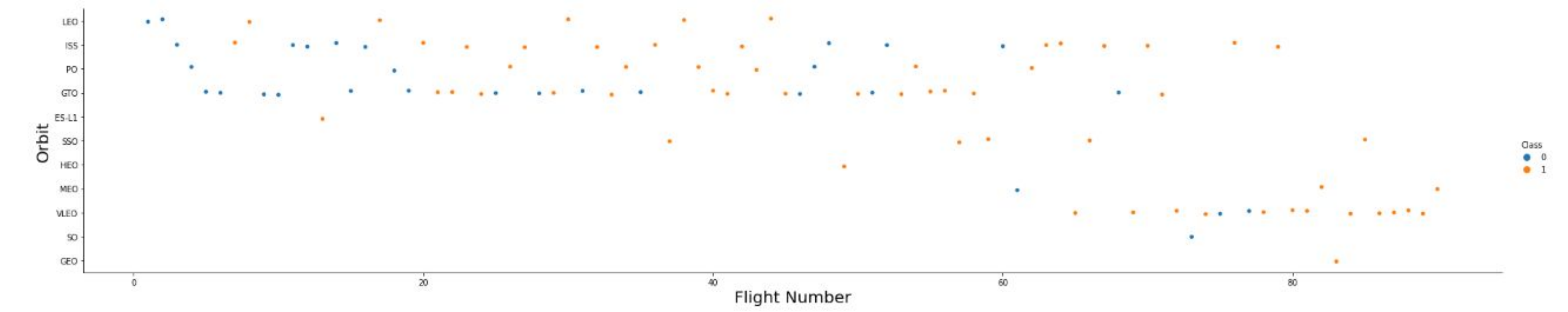
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type:



Flight Number vs. Orbit Type

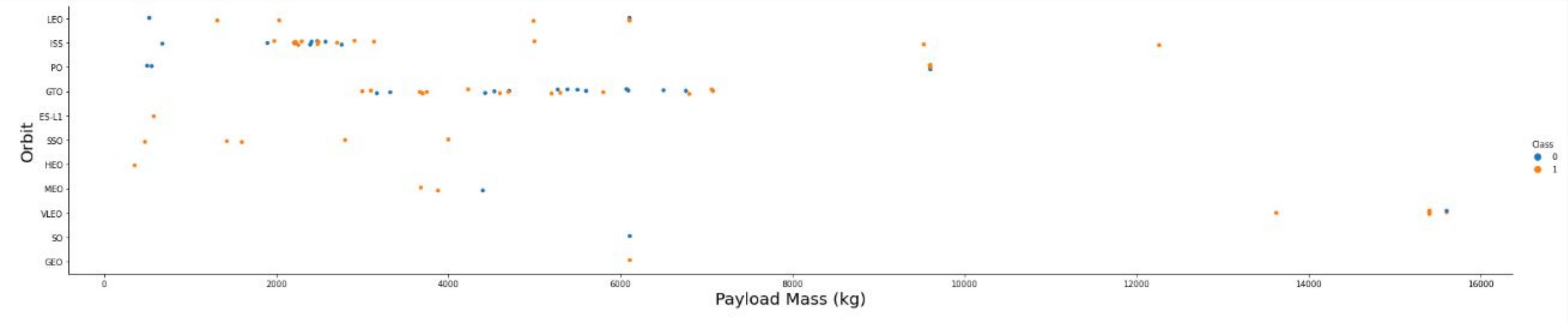
- Show a scatter point of Flight number vs. Orbit type



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

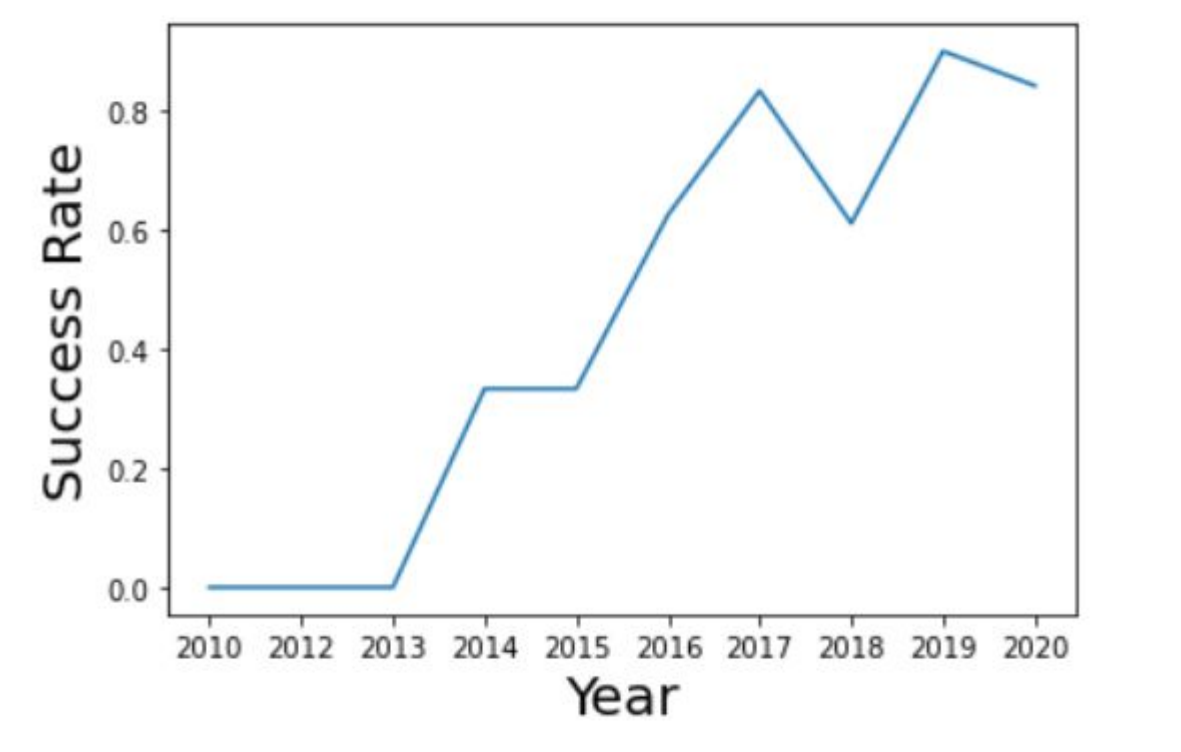
We can see that most payload are under 8000kg



Launch Success Yearly Trend

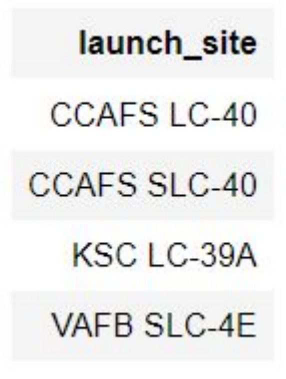
- Show a line chart of yearly average success rate

Over all success rate increase from 2013 to 2020, but drop at 2018.



All Launch Site Names

- The names of the unique launch sites are:



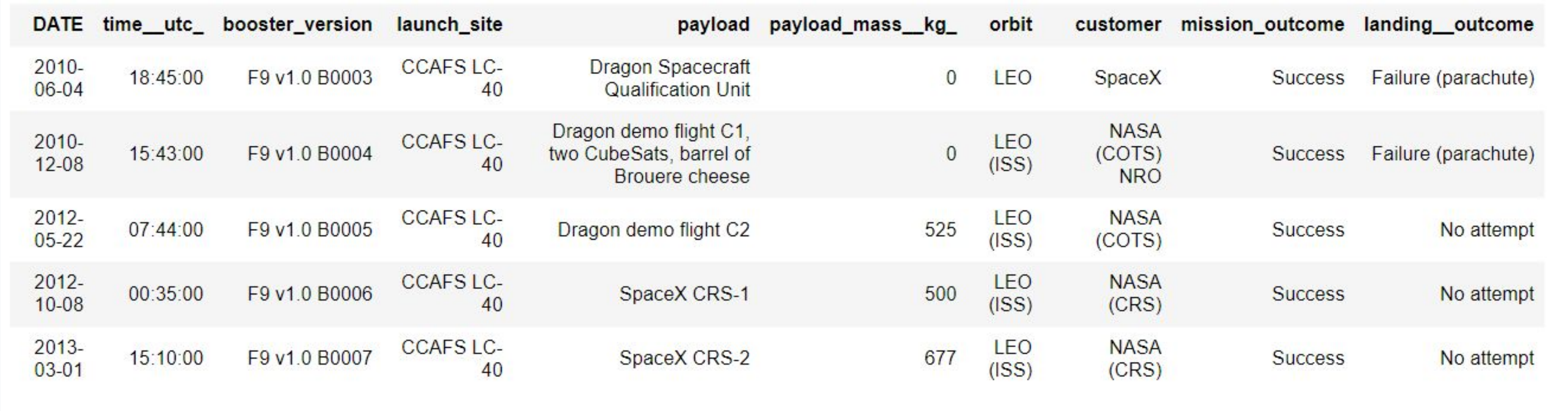
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Use ‘distinct’ keyword to filter out duplicate

Launch Site Names Begin with 'CCA'

- The table shows 5 records where launch sites begin with `CCA`

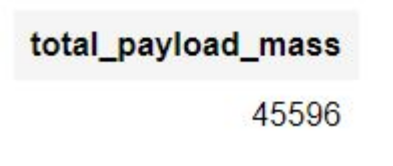
This SQL query retrieves data from the table `SPACEXDATASET` where the `launch_site` column value starts with the characters 'CCA'



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

- Calculate the total payload carried by boosters from NASA



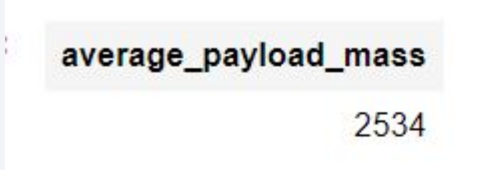
total_payload_mass
45596

This SQL query retrieves data from the table `SPACEXDATASET` and sum the play_mass where the `customer` column value is 'NASA (CRS)'

Average Payload Mass by F9 v1.1

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

This SQL query retrieves data from the table `SPACEXDATASET` where the `booster_version` column value contains the string 'F9 v1.1' and then calculate the average of the payload mass

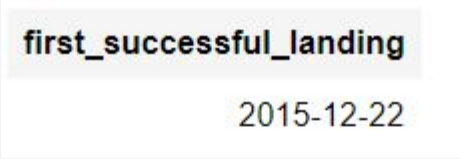


average_payload_mass
2534

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

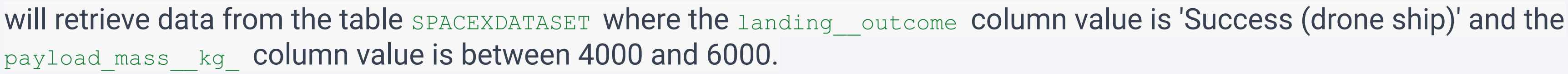
  we need to use min() function to get the first date.



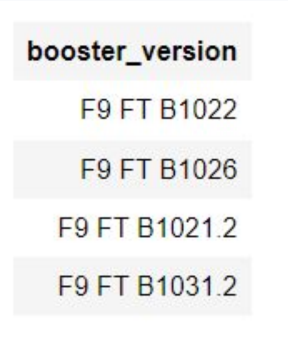
first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



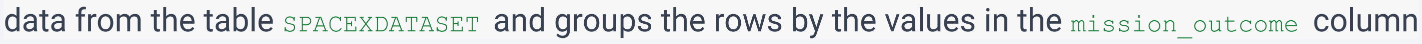
will retrieve data from the table `SPACEXDATASET` where the `landing__outcome` column value is 'Success (drone ship)' and the `payload_mass__kg_` column value is between 4000 and 6000.



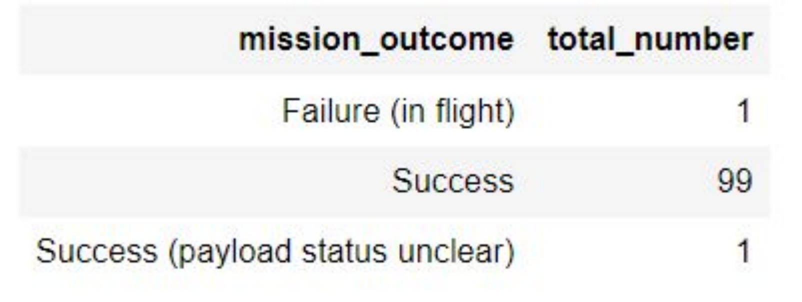
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



data from the table `SPACEXDATASET` and groups the rows by the values in the `mission_outcome` column

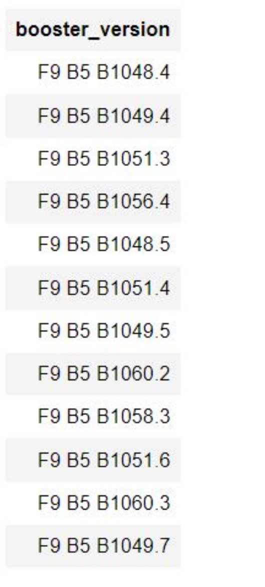


mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

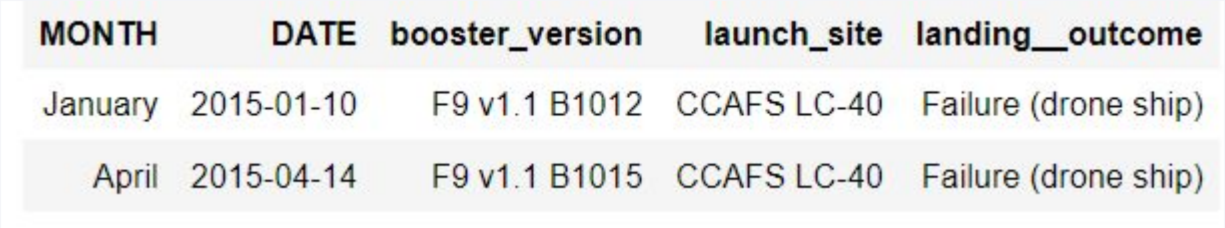
the `payload_mass__kg_` column value is equal to the maximum value of `payload_mass__kg_` in the same table, the max value is from subquery by max() function.



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

2015 Launch Records

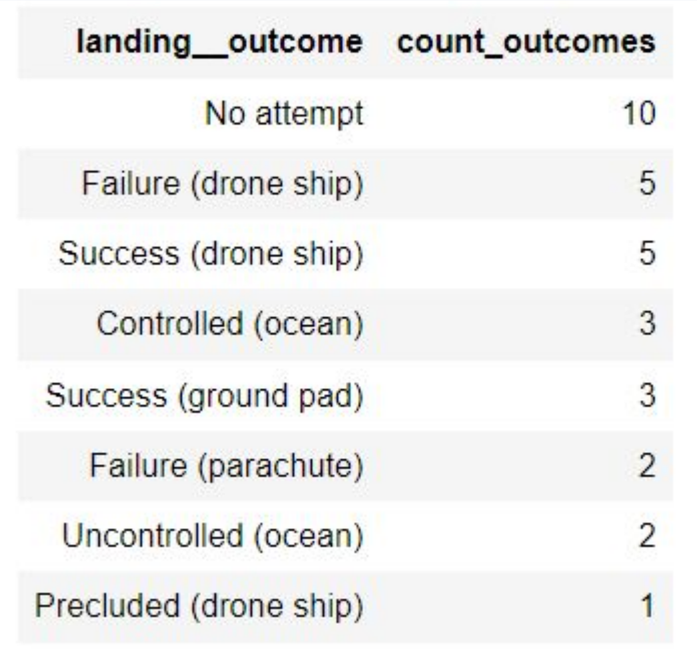
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

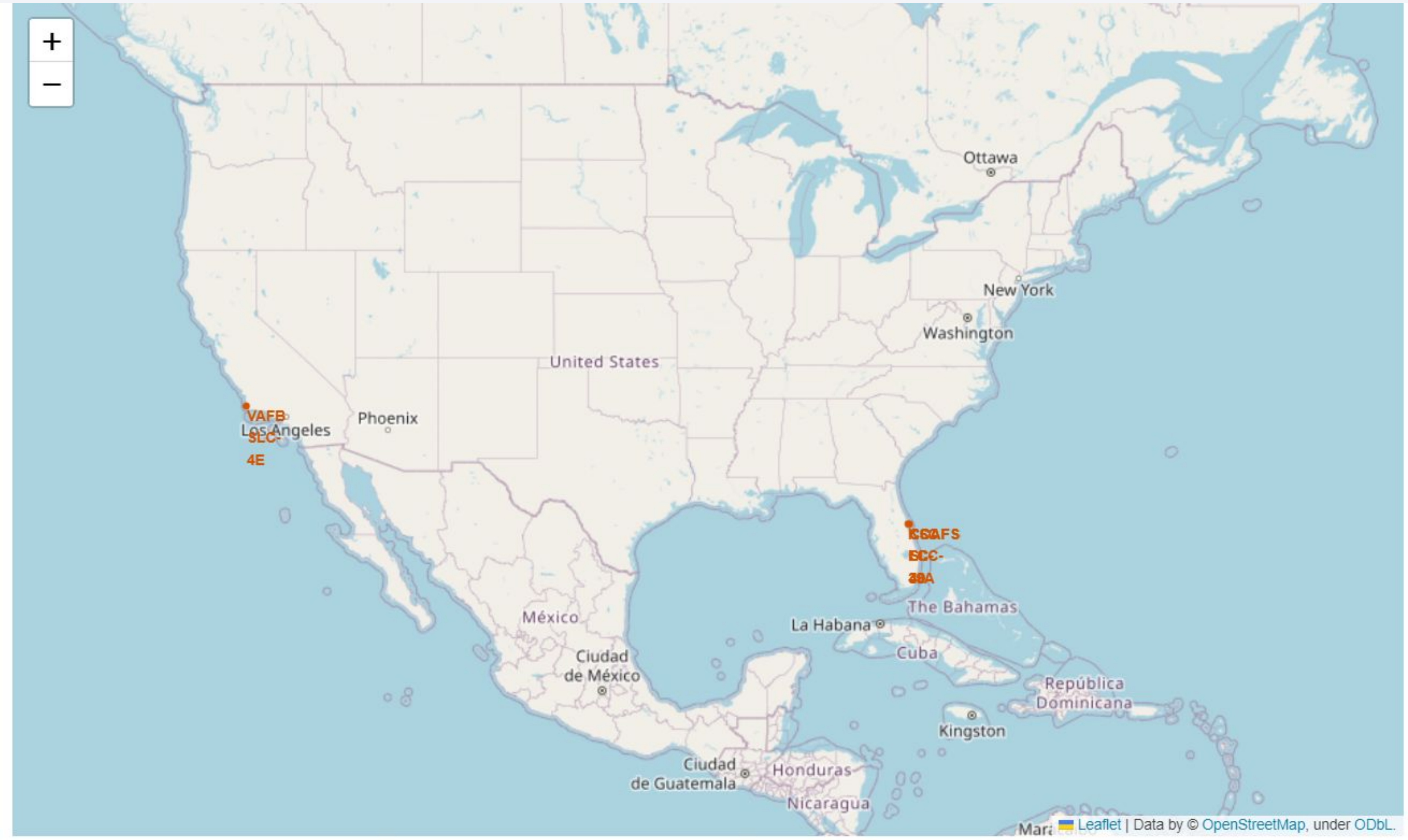


landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

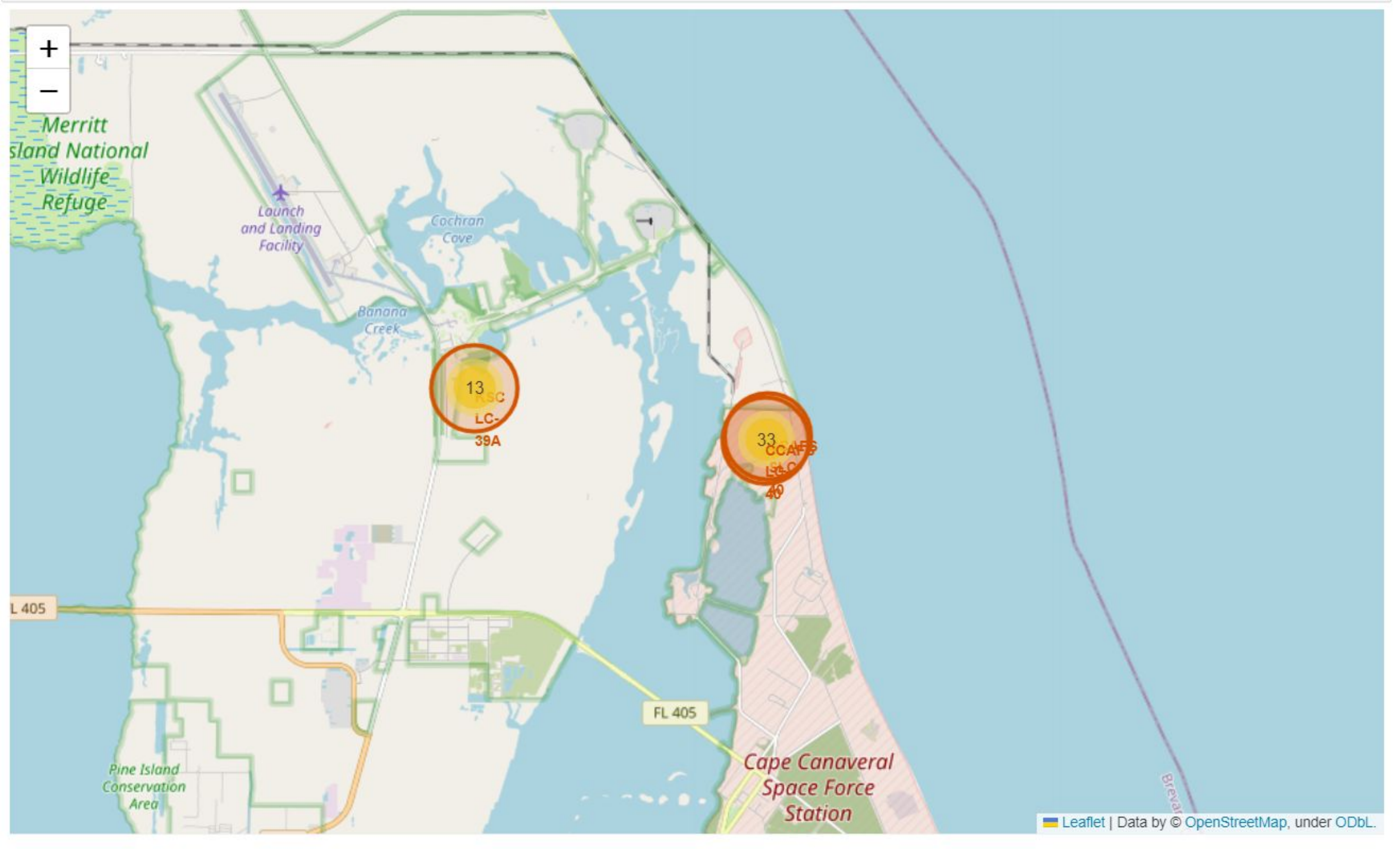
Section 3

Launch Sites Proximities Analysis

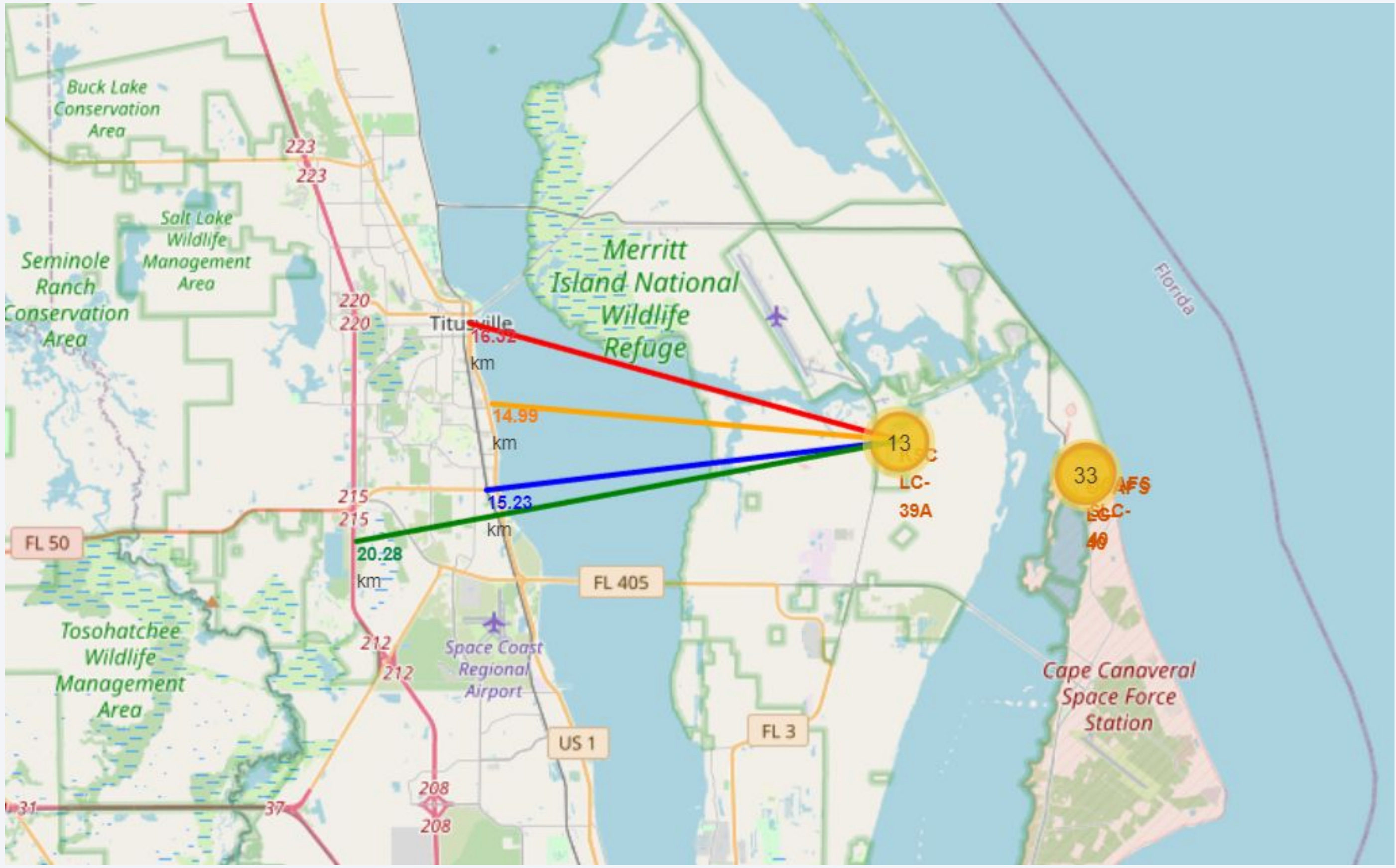
<All launch sit in the US>


+
−
Ottawa
New York
Washington
United States
VAFB
Los Angeles
SLC-
4E
Phoenix
CCAFS
SLC-
39A
The Bahamas
La Habana
Cuba
México
Ciudad de México
República Dominicana
Kingston
Ciudad de Guatemala
Honduras
Nicaragua
Leaflet | Data by © OpenStreetMap, under ODbL.

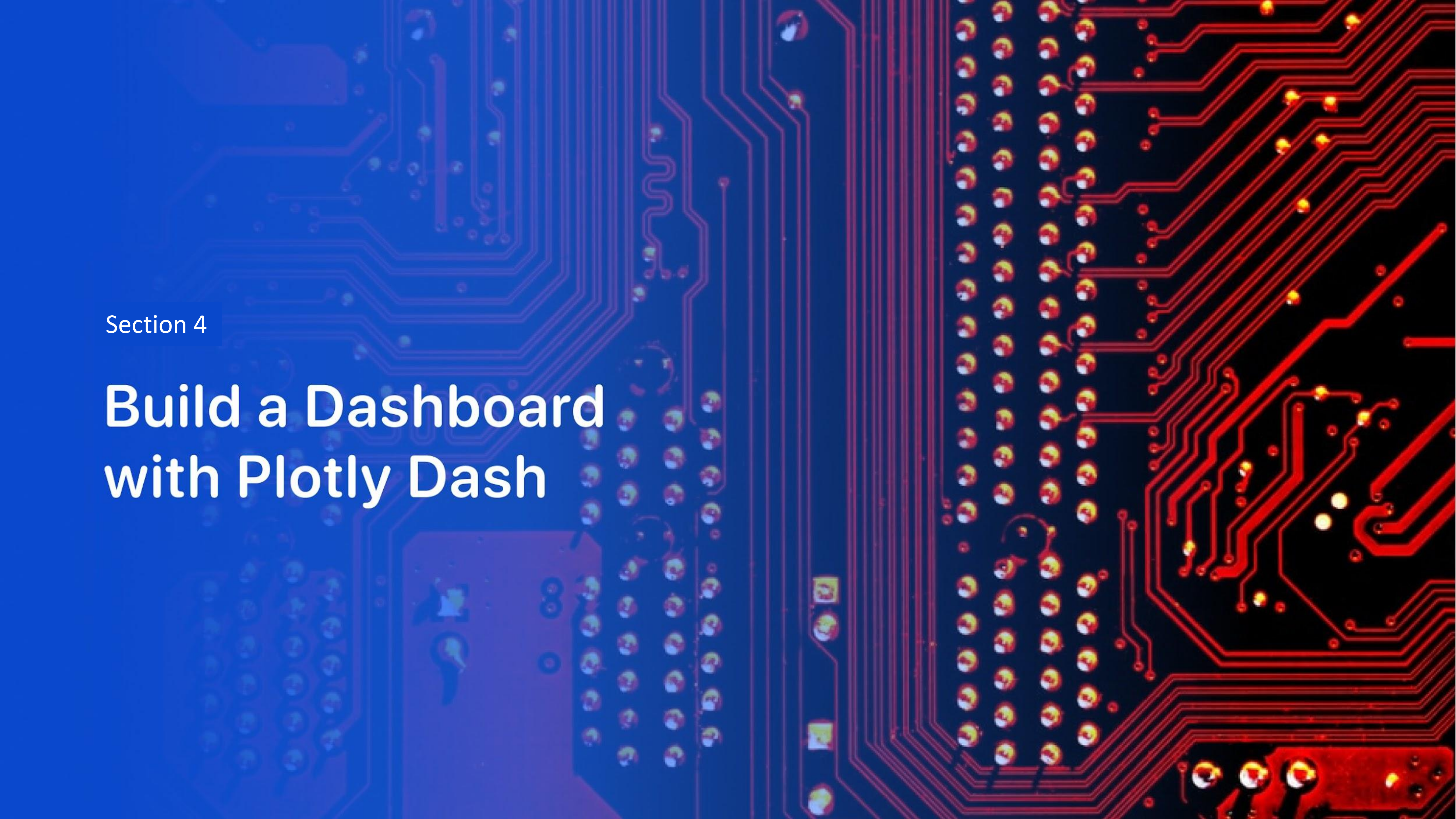
<Launch site clusters>


+
−
Merritt Island National Wildlife Refuge
Launch and Landing Facility
Cochran Cove
Banana Creek
13
KSC LC-39A
33
CCAFS LC-40
L 405
FL 405
Pine Island Conservation Area
Cape Canaveral Space Force Station
Leaflet | Data by © OpenStreetMap, under ODbL.

<Distances between a launch site to its proximities >



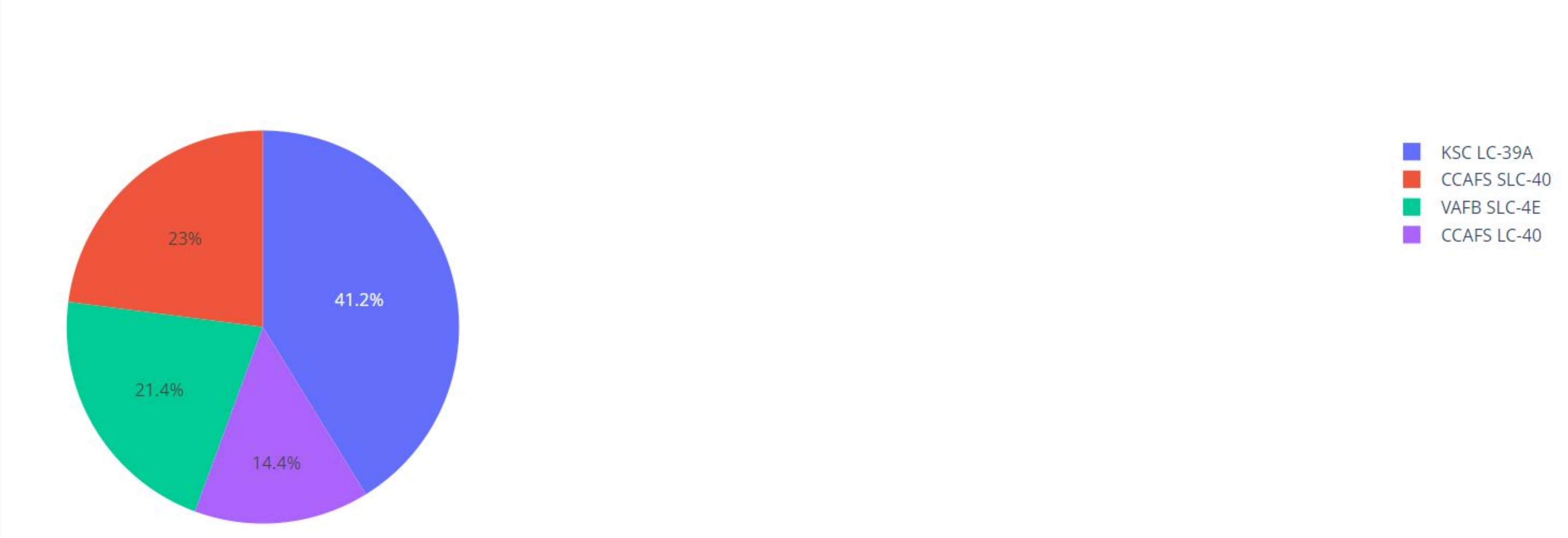
Buck Lake Conservation Area
Salt Lake Wildlife Management Area
Seminole Ranch Conservation Area
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
LC-39A
33
LC-40
Florida
223
220
215
212
208
37
31
FL 50
FL 405
FL 3
US 1
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Cape Canaveral Space Force Station



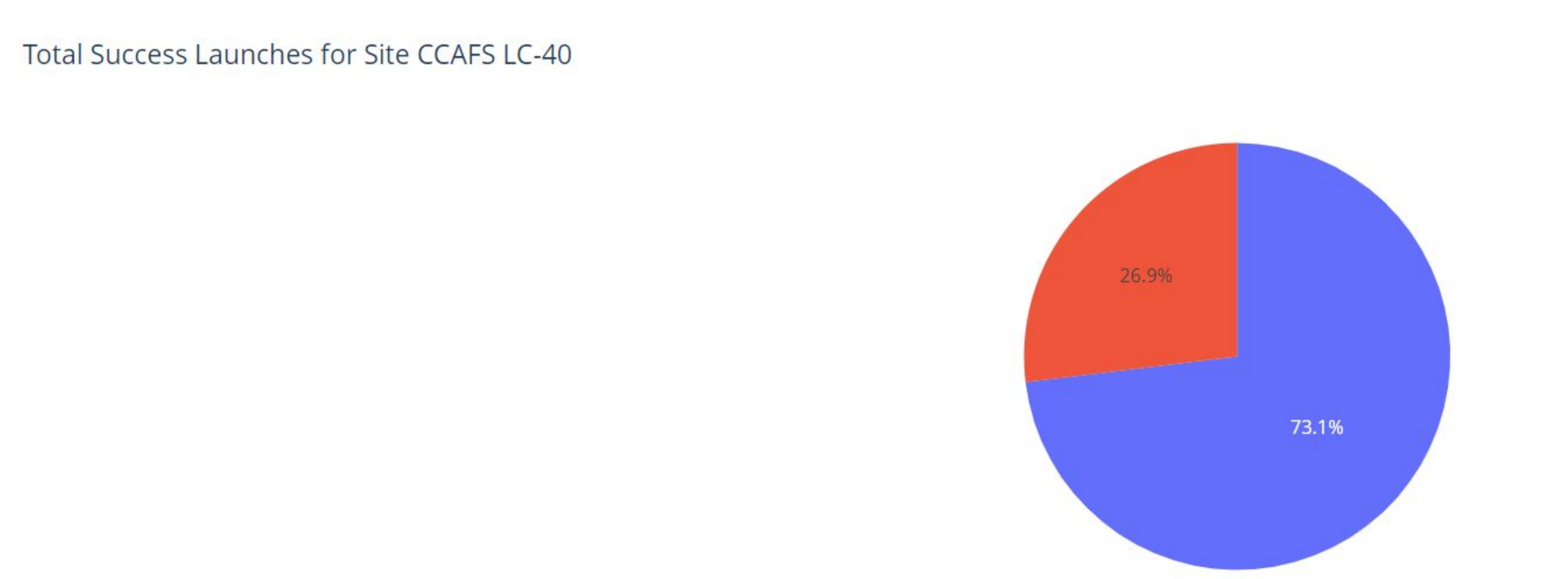
Section 4

Build a Dashboard with Plotly Dash

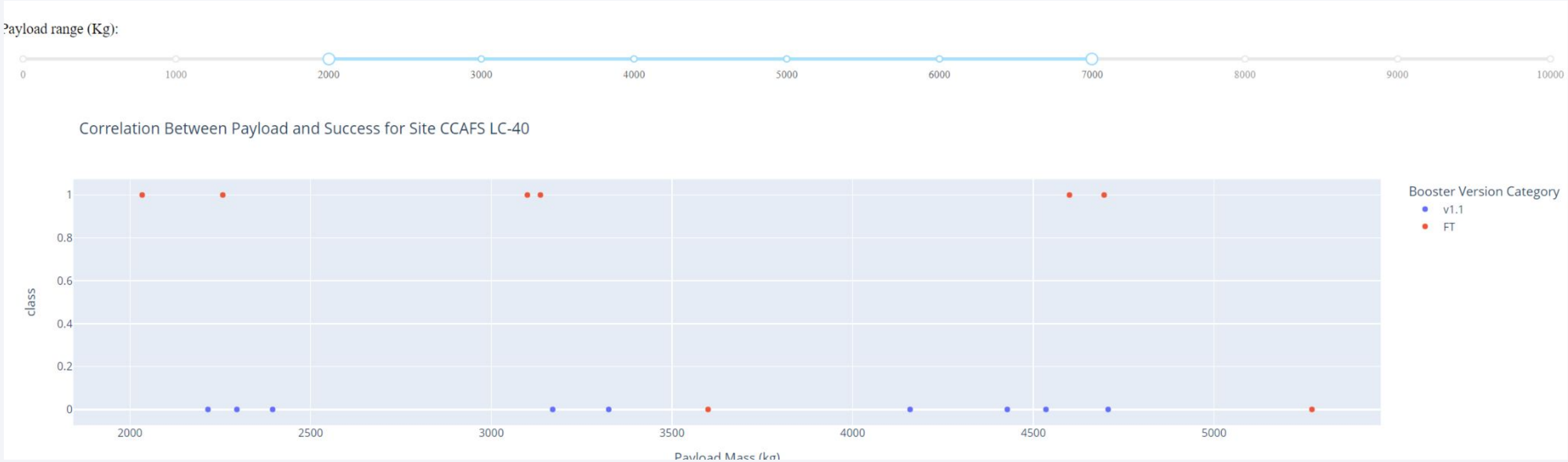
<Success rate of by sites>


41.2%
23%
21.4%
14.4%
KSC LC-39A
CCAFS SLC-40
VAFB SLC-4E
CCAFS LC-40

<Total success rate of the highest site>


Total Success Launches for Site CCAFS LC-40
26.9%
73.1%

<Payload vs. Launch Outcome>



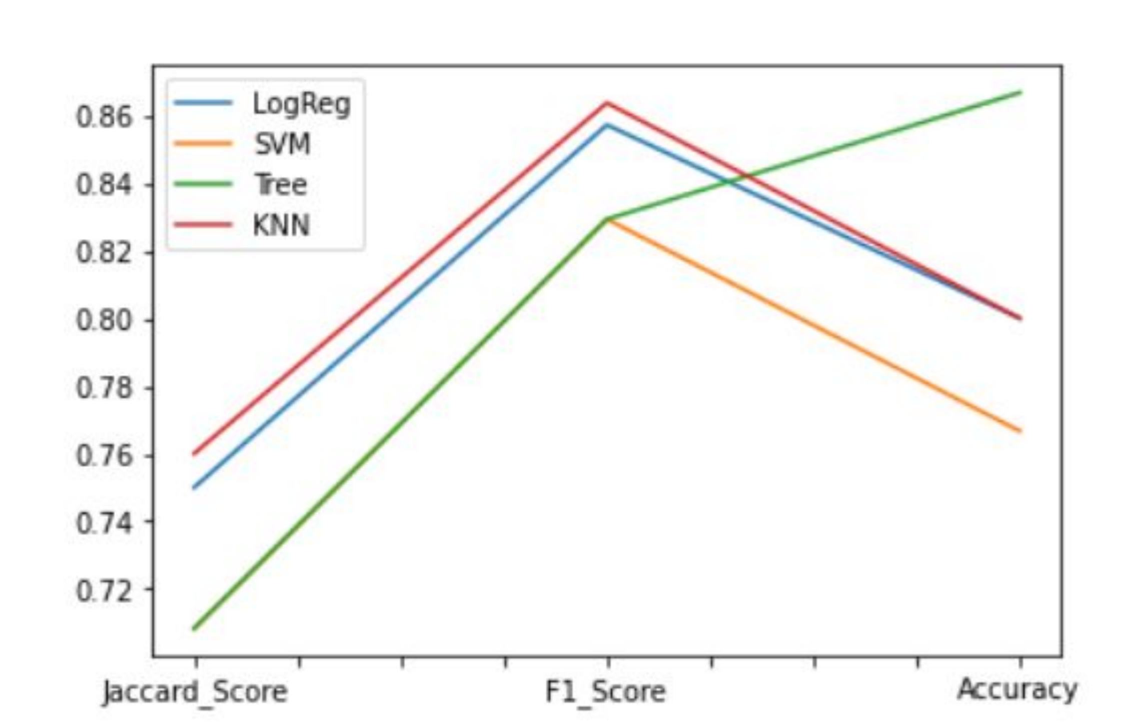
Payload range (Kg):

Section 5

Predictive Analysis (Classification)

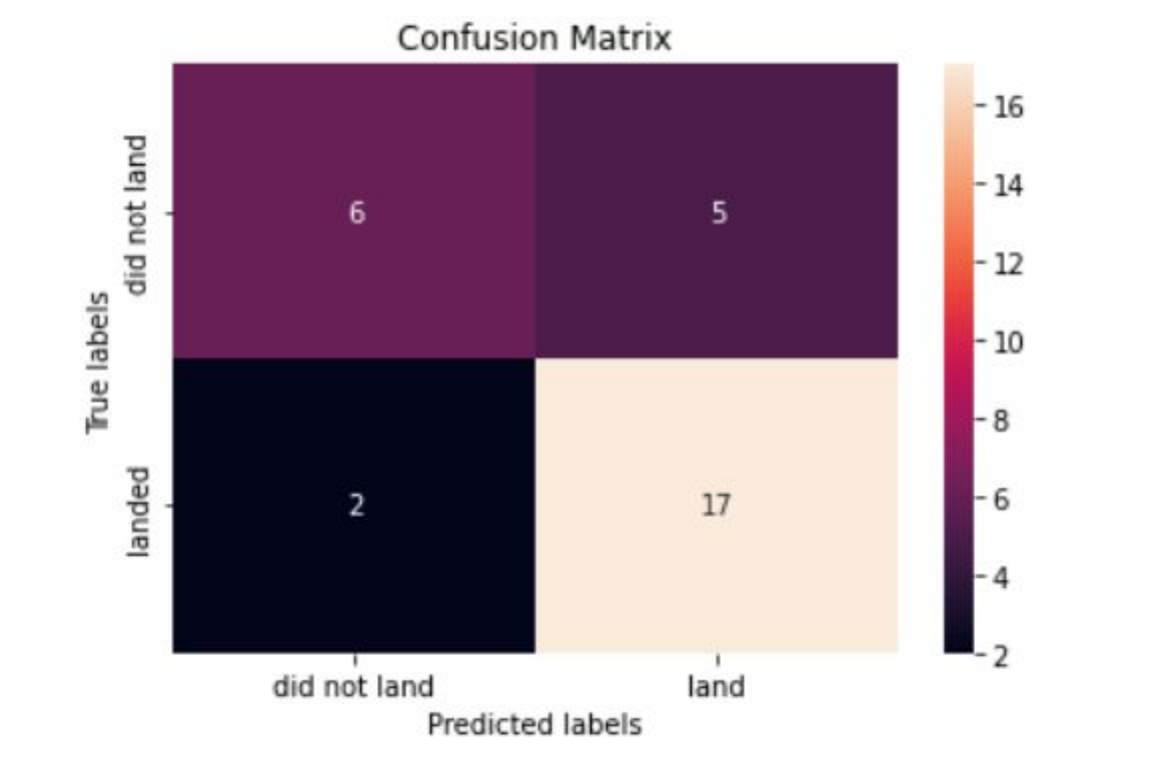
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a line chart
- The green line, tree model has the highest acc.



Confusion Matrix

- Confusion matrix of Tree model, it shows true positive and true negative compared to the false ones.



Conclusions

- Decision Tree Model performs the best for this dataset, for further exploration we can train other ensemble model base on trees, like XGboost
- GTO orbit type has the lowest success rate
- Larger payload mass tend to have higher success rate.
- The success rate of launches increases over the years.

Appendix

- Resource:

  scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation
  python-visualization.github.io

Thank you!