IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Le Minh Tuan
19/02/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building map with Folium
  - Building dashboard with Dash
  - Predictive analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this project, we will be provided with an overview of the problem and the tools you need to complete this project.

- Problems you want to find answers

- Use data science methodologies to define and formulate a real-world business

- Correlation between each rocket variables and successful landing rate

Section 1

Methodology

Methodology

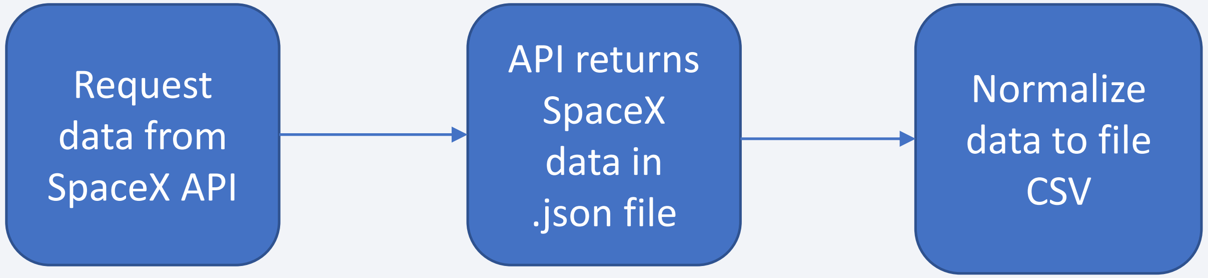
Executive Summary

- Data collection methodology:
  - Collect from SpaceX API and Web Scraping (Wikipedia)
- Perform data wrangling
  - Convert outcome into training labels with the booster successful/fail 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, Logistic Regression

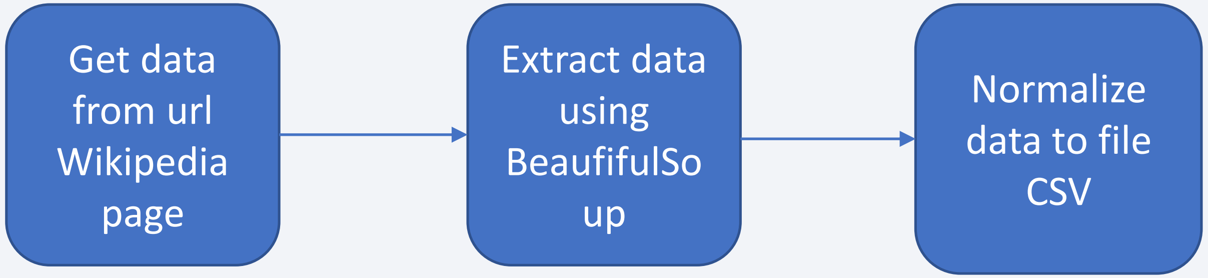
Data Collection

- Data collect from SpaceX API and web scraping data from Wikipedia page with title List of Falcon 9 and Falcon Heavy launches

- Folwchart get data from SpaceX API


Request data from SpaceX API
API returns SpaceX data in .json file
Normalize data to file CSV

- Folwchart get data from SpaceX API


Get data from url Wikipedia page
Extract data using BeaufifulSo up
Normalize data to file CSV

Data Collection - SpaceX API

1. Request data from API : *https://api.spacexdata.com/v4/launches/past*
2. Convert response to json file

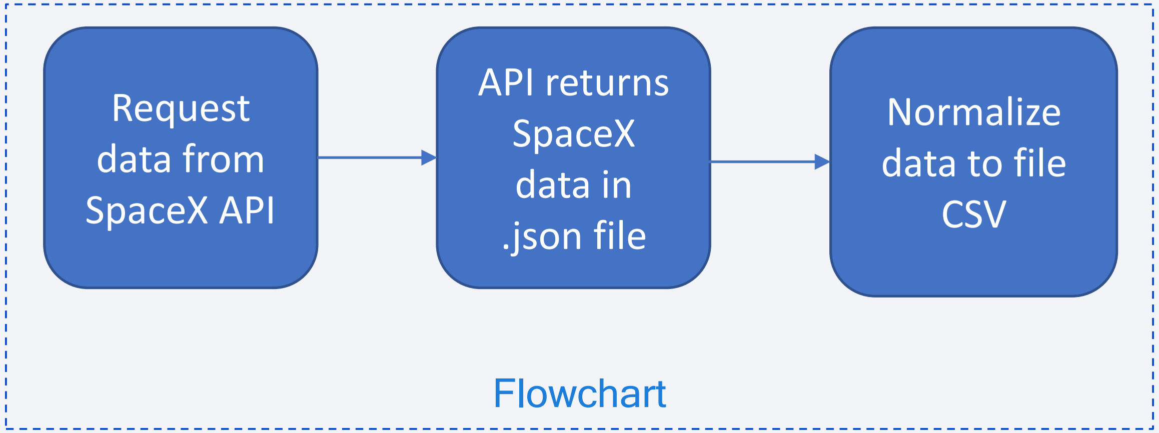
   data = pd.json_normalize(response.json())
3. Clean Data

   getBoosterVersion(data)

   getLaunchSite(data)

   getPayloadData(data)

   getCoreData(data)
4. Normalize to CSV *data_falcon9.to_csv('dataset_part_1.csv', index=False)*


Request data from SpaceX API
API returns SpaceX data in .json file
Normalize data to file CSV

Flowchart

- GitHub

Data Collection - Scraping

1. Request data from url : *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*

2. Extract data using BeautifulSoup

   soup = BeautifulSoup(response.text, 'html.parser')

   html_tables = soup.find_all('table')

   first_launch_table = html_tables[2]

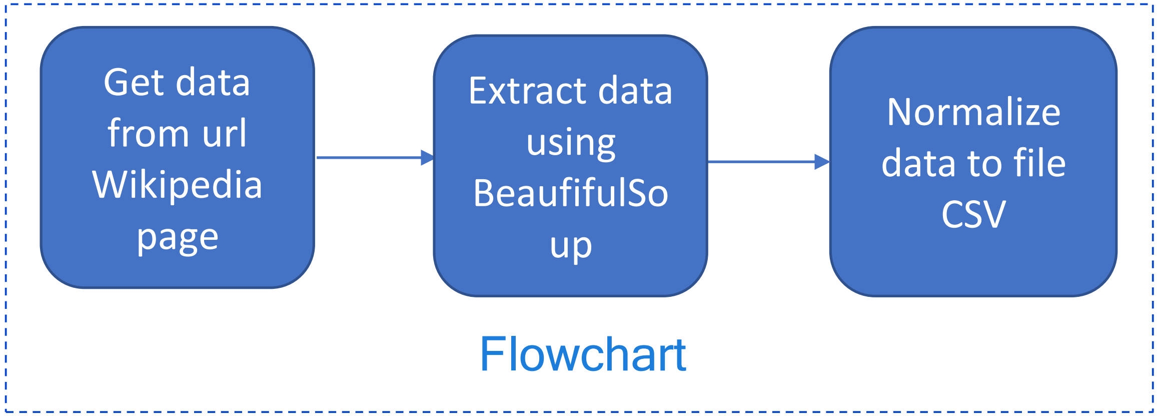
3. Create a data frame by parsing the launch HTML table

   launch_dict= dict.fromkeys(column_names)

4. Normalize to CSV

   df.to_csv('spacex_web_scraped.csv', index=False)

GitHub


Get data from url Wikipedia page
Extract data using BeaufifulSoup
Normalize data to file CSV
Flowchart

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident
  - True Ocean: means the mission outcome was successfully landed to a specific region of the ocean while
  - False Ocean: means the mission outcome was unsuccessfully landed to a specific region of the ocean.
  - True RTLS: means the mission outcome was successfully landed to a ground pad
  - False RTLS: means the mission outcome was unsuccessfully landed to a ground pad.
  - True ASDS: means the mission outcome was successfully landed on a drone ship
  - False ASDS: means the mission outcome was unsuccessfully landed on a drone ship.
- Convert result into training labels
  - 1=successful / 0=fail

- GitHub

EDA with Data Visualization

- Scatter chart : A scatter chart show relationship between two variables
  - Flight Number and Launch Site
  - Payload and Launch Site
  - FlightNumber and Orbit type
  - Payload and Orbit type
  - A scatter chart show relationship between two variables.
- Bar chart: use for compare data between multi variables
  - Success rate of each Orbit type
- Line chart : use show trend or predict
  - Year and Success Rate
- GitHub

EDA with SQL

- Summarize the my SQL queries
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - 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 names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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
- Github

Build an Interactive Map with Folium

- Objects created and added to folium map
  - Markers that show all launch sites on a map
  - Markers that show the success/fail launches for earch site on the map
  - Lines that show the distances between a launch site to its proximites
- Can find launch sites in map
- GitHub

Build a Dashboard with Plotly Dash

- Pie chart
  - Show total success launches by sites
- Scatter chart
  - Show the relationship between Outcomes and Payload mass(kg) by different boosters
- GitHub

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels

  * Create a column for the class

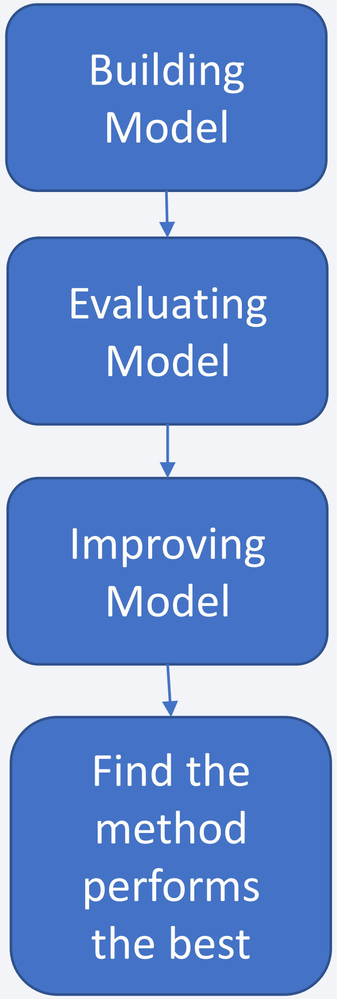
  * Standardize the data

  * 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]



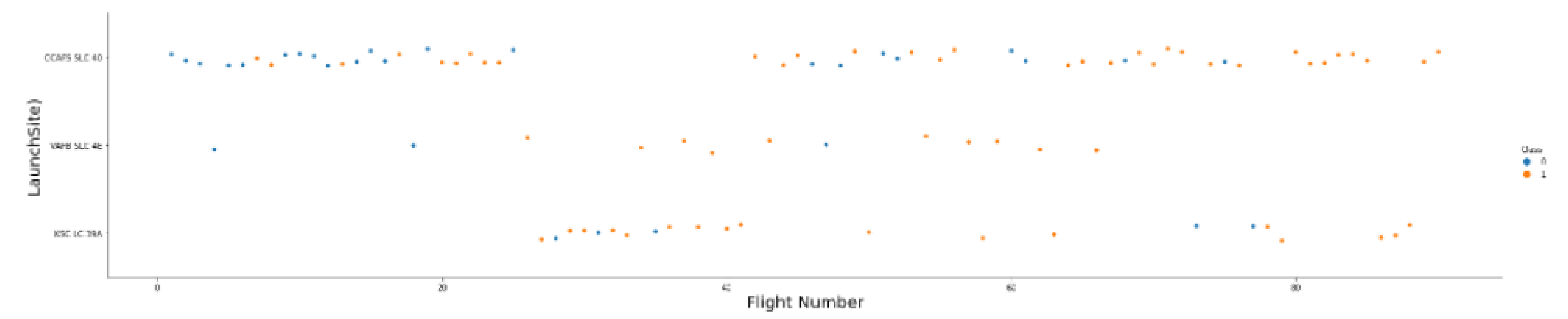
Results

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

Section 2

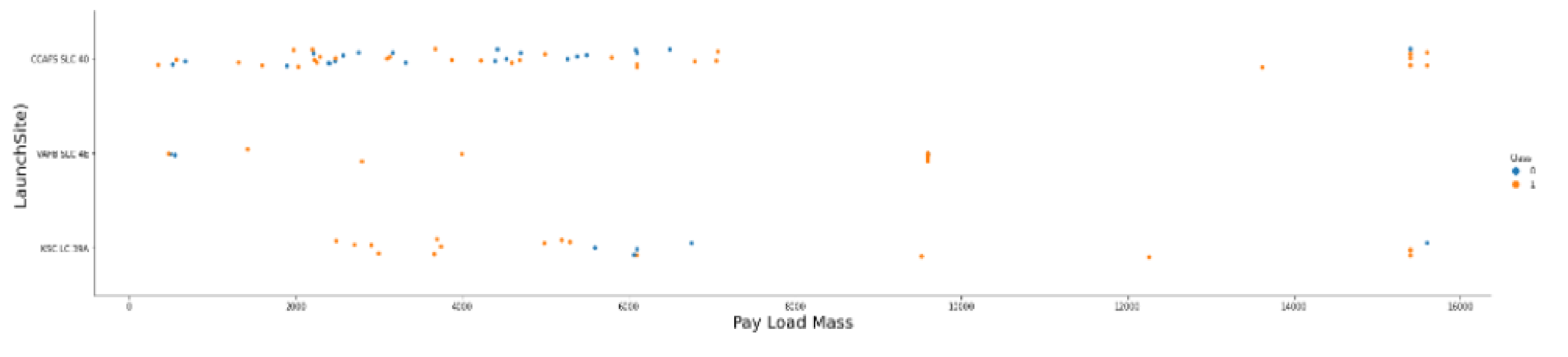
Insights drawn from EDA

Flight Number vs. Launch Site



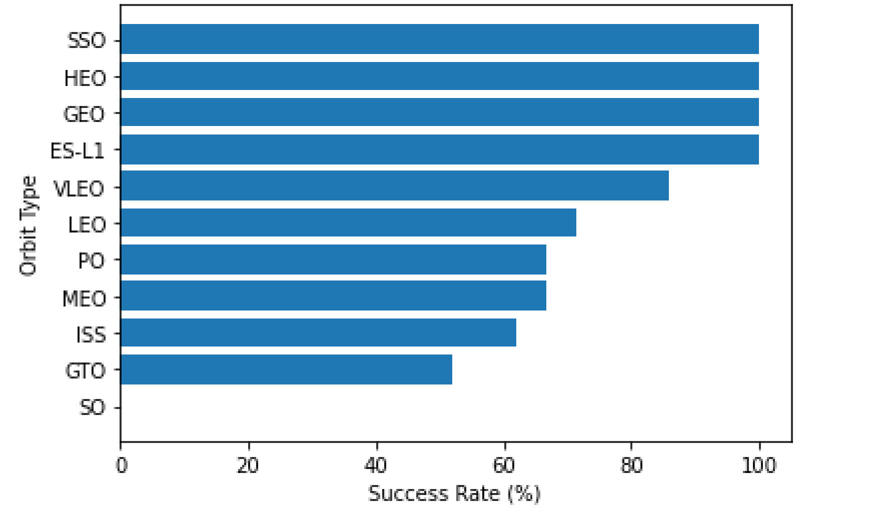
- Class 0 (blue color): fail launch , Class 1 (orange color) : successful launch
- This figure show that the success rate increased as the number of flights increased

Payload vs. Launch Site



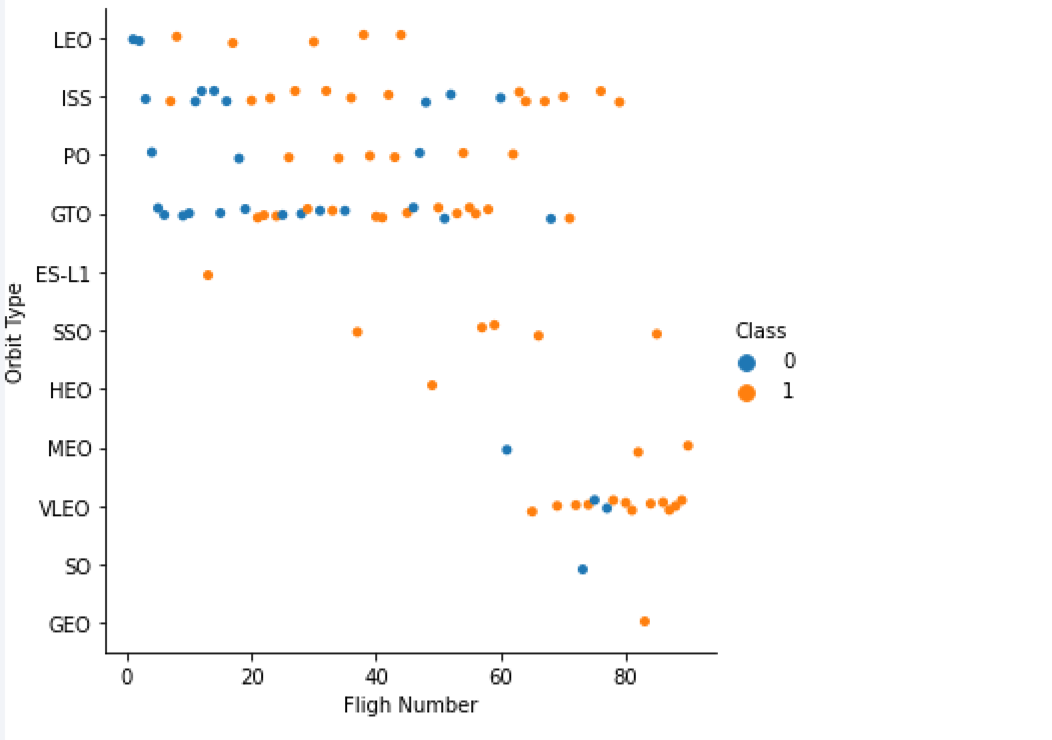
- Class 0 (blue color): fail launch , Class 1 (orange color) : successful launch

Success Rate vs. Orbit Type



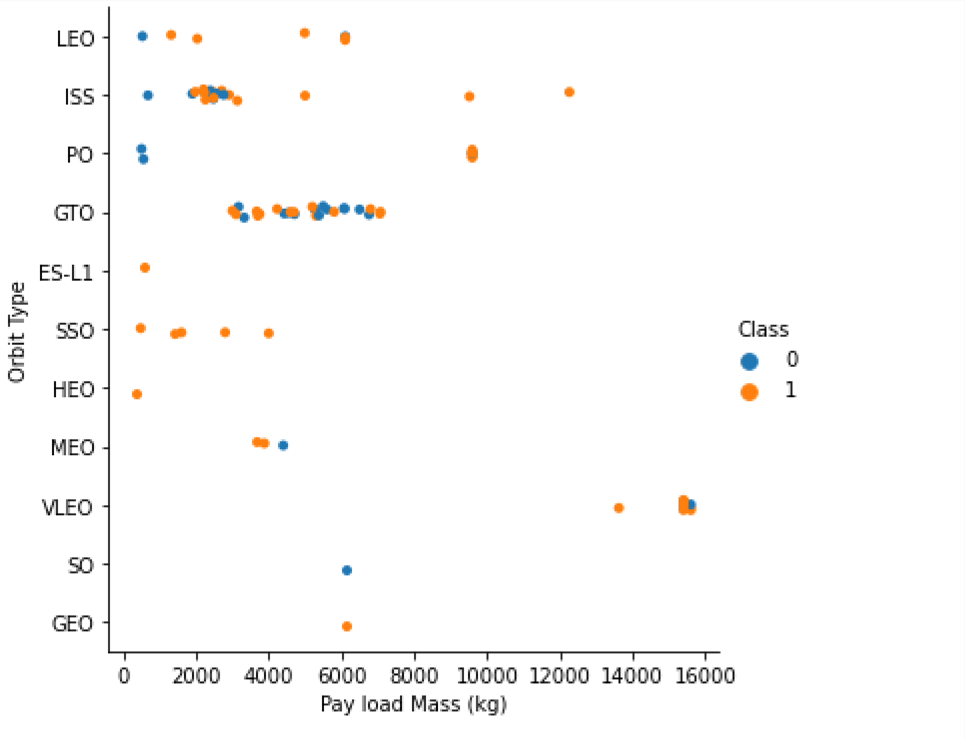
- SSO, HEO, GEO, ES-L1 is the highest
- SO is zero

Flight Number vs. Orbit Type



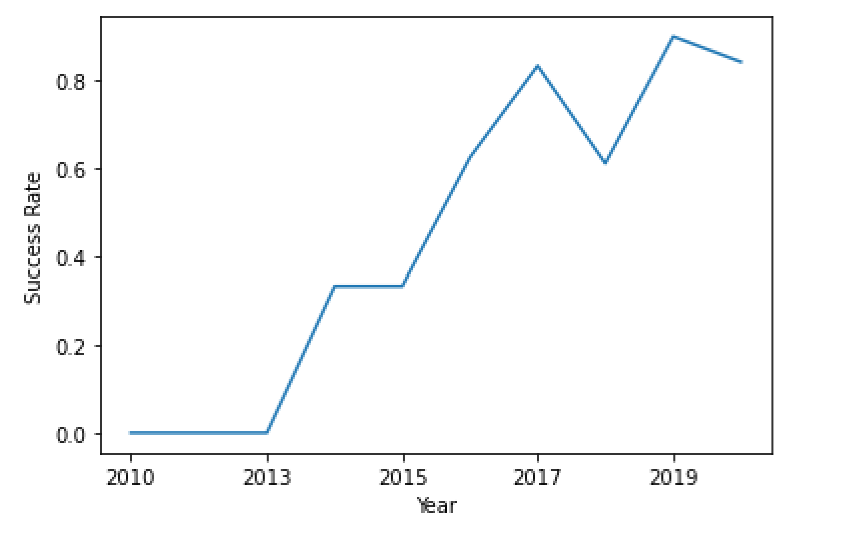
- Class 0 (blue color): fail launch , Class 1 (orange color) : successful launch
- LEO have rate success the highest

Payload vs. Orbit Type



- Class 0 (blue color): fail launch , Class 1 (orange color) : successful launch

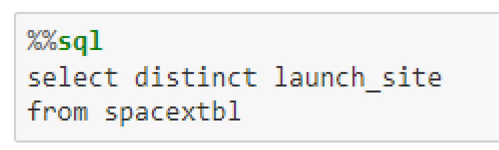
Launch Success Yearly Trend



- From 2013, rate increase very good
- 2018 decreased and now, Rate is good, more 80%

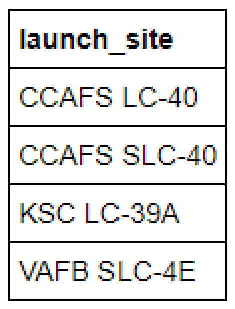
All Launch Site Names

- Query



```
%%sql
select distinct launch_site
from spacextbl
```

- Result

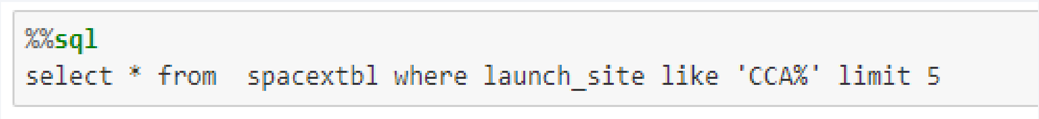


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Have 4 results
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Use Distinct get unique values in column launch_site

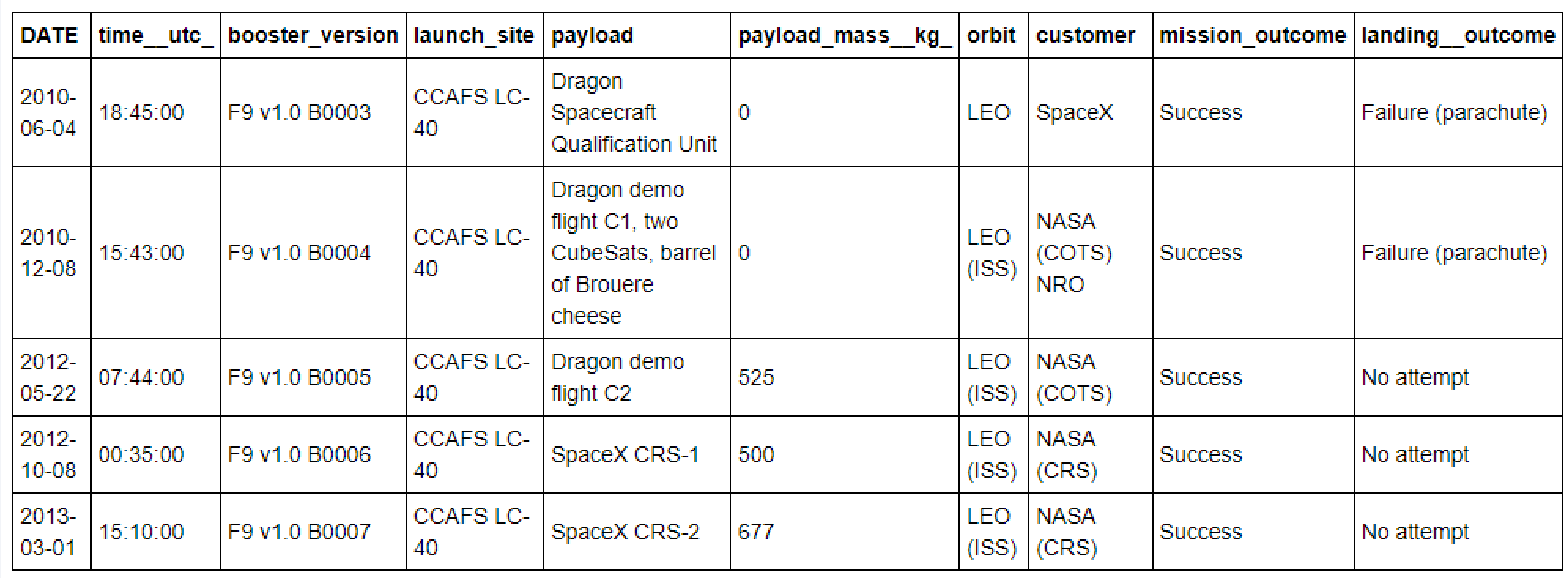
Launch Site Names Begin with 'CCA'

- Query



```
%%sql
select * from  spacextbl where launch_site like 'CCA%' limit 5
```

- Result



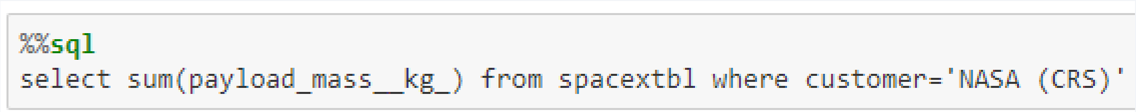
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Get 5 records with condition launch sites begin with ‘CCA’ using condition **like ‘CCA%’ limit 5**

- Show data

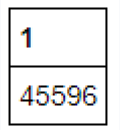
Total Payload Mass

- Query



```
%%sql
select sum(payload_mass__kg_) from spacextbl where customer='NASA (CRS)'
```

- Result

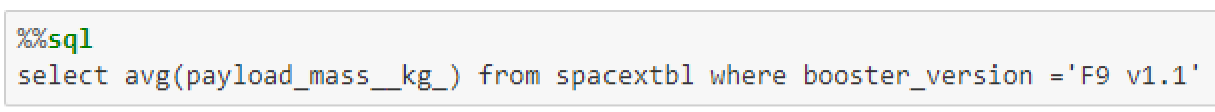


1
45596

- Calculate the total payload using function sum with condition customer=‘NASA (CRS)’
- Show result calculate total payload mass kg

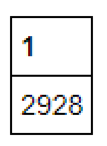
Average Payload Mass by F9 v1.1

- Query



```
%%sql
select avg(payload_mass__kg_) from spacextbl where booster_version ='F9 v1.1'
```

- Result

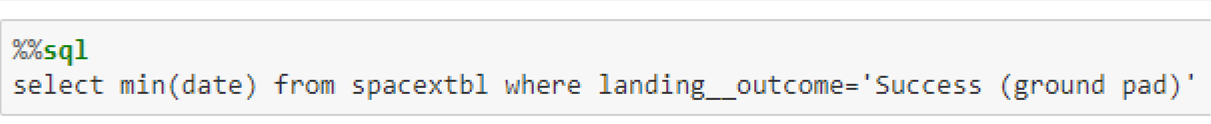


1
2928

- Calculate the average payload mass using function avg with condition booster_version =‘F9 v1.1’

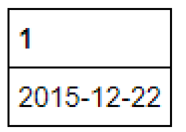
First Successful Ground Landing Date

- Query



```
%%sql
select min(date) from spacextbl where landing__outcome='Success (ground pad)'
```

- Result

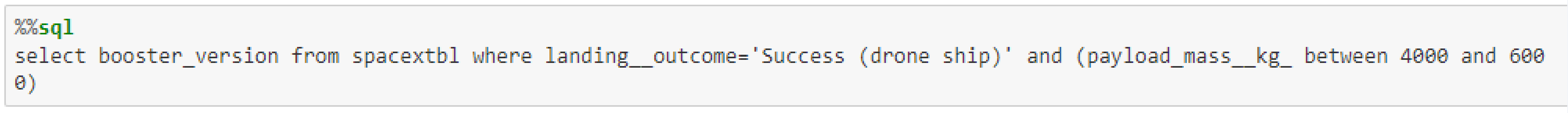


1
2015-12-22

- Use Min function find the dates of the first successful landing outcome on ground pad

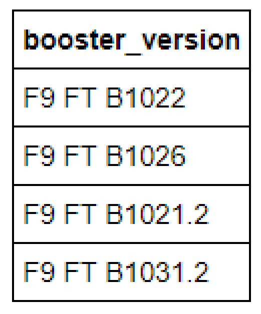
Successful Drone Ship Landing with Payload between 4000 and 6000

- Query



```
%%sql
select booster_version from spacextbl where landing__outcome='Success (drone ship)' and (payload_mass__kg_ between 4000 and 600
0)
```

- Result

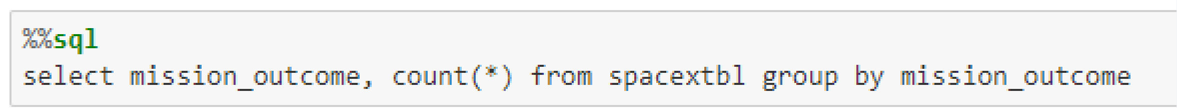


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Use condition ‘Success (drone ship)’ and payload_mass__kg_ between 4000 and 6000 for list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

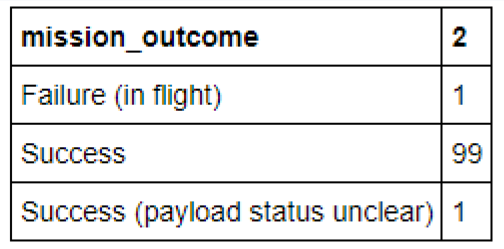
Total Number of Successful and Failure Mission Outcomes

- Query



```
%%sql
select mission_outcome, count(*) from spacextbl group by mission_outcome
```

- Result

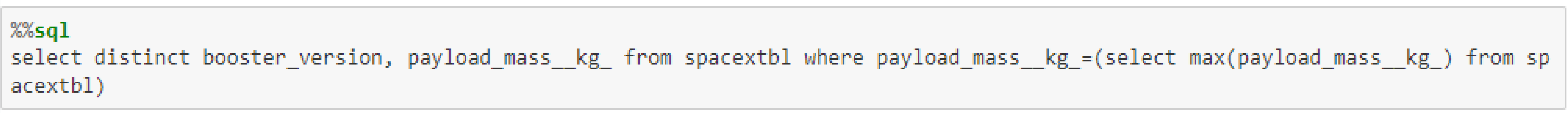


mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Use Count function for calculate
- Use GROUP BY for group rows with the same value

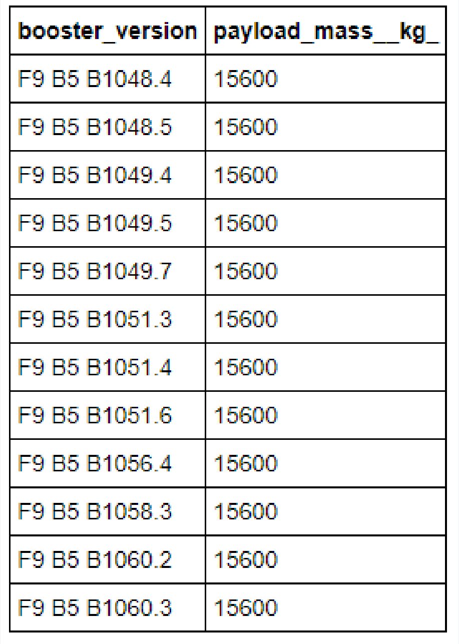
Boosters Carried Maximum Payload

- Query



```
%%sql
select distinct booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_=(select max(payload_mass__kg_) from sp
acextbl)
```

- Result

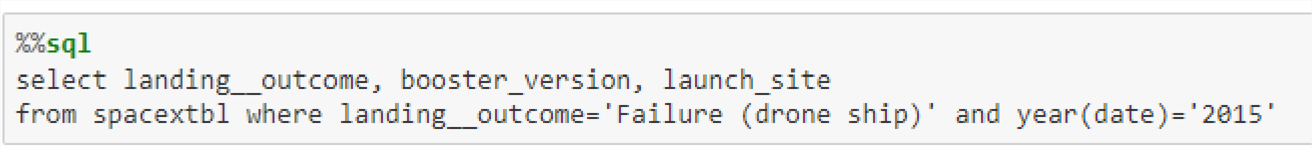


booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- Get Max value payload_mass_kg__
- Filter data in table spacextbl with condition equal max value from subsquery

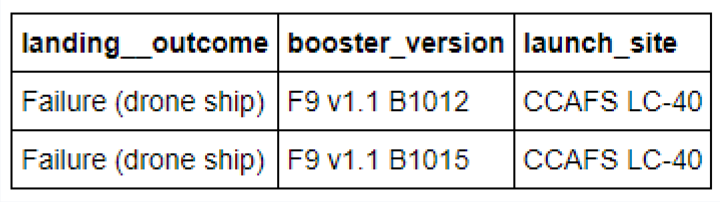
2015 Launch Records

- Query



```
%%sql
select landing__outcome, booster_version, launch_site
from spacextbl where landing__outcome='Failure (drone ship)' and year(date)='2015'
```

- Result

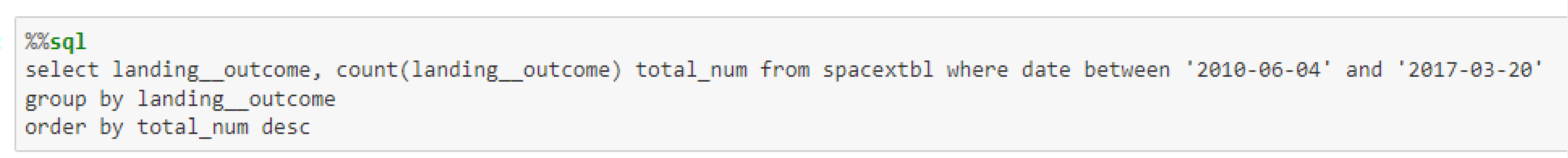


landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Use condition filter data column landing__outcome by value ‘Failure (drone ship)’ and year = 2015
- Function Year() return value year of data datetime

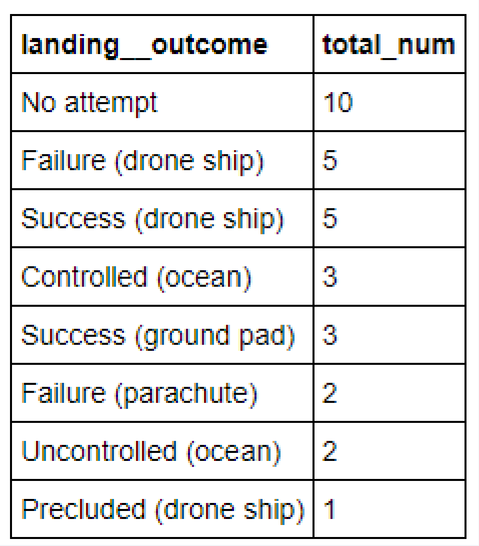
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Query



```
%%sql
select landing__outcome, count(landing__outcome) total_num from spacextbl where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by total_num desc
```

- Result



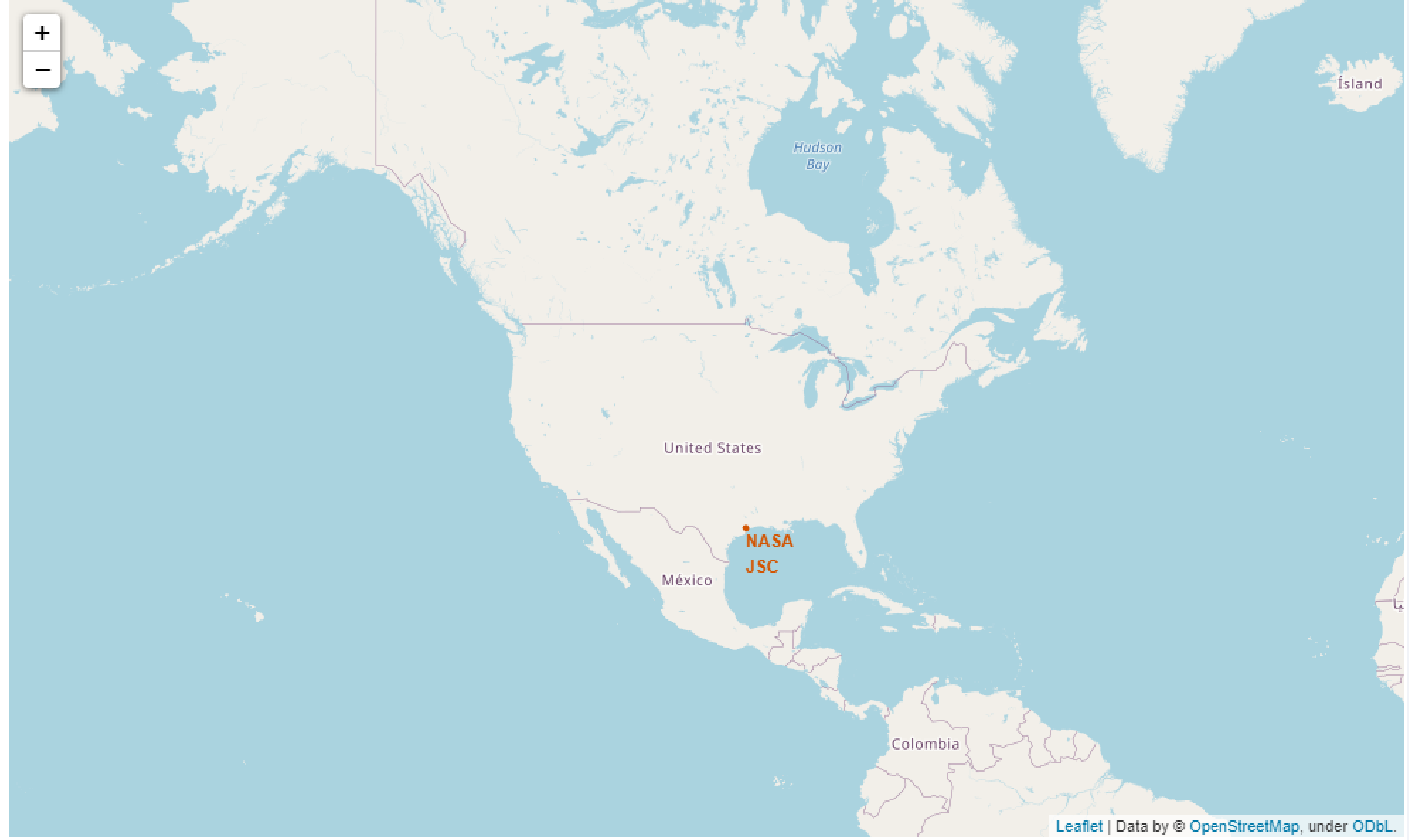
landing__outcome	total_num
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

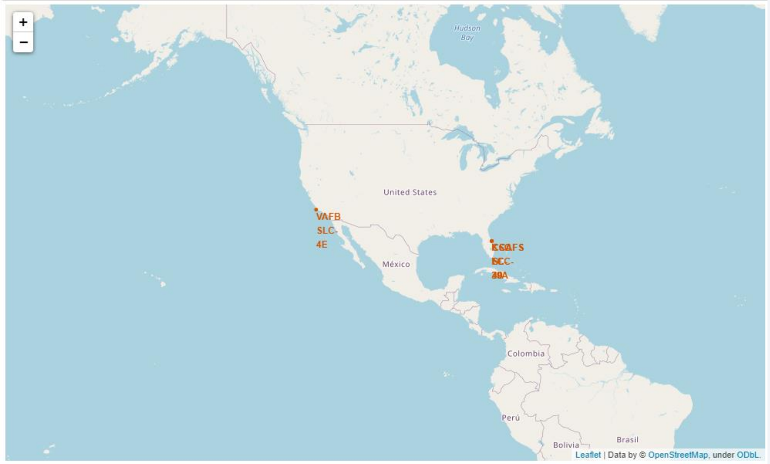
- Use condition filter date data between ‘2010-06-04’ and ‘2017-03-20’
- Use Count calculate data have the same in landing__outcome column.
- Use Order By to sort the record total_number landing

Section 3

Launch Sites Proximities Analysis

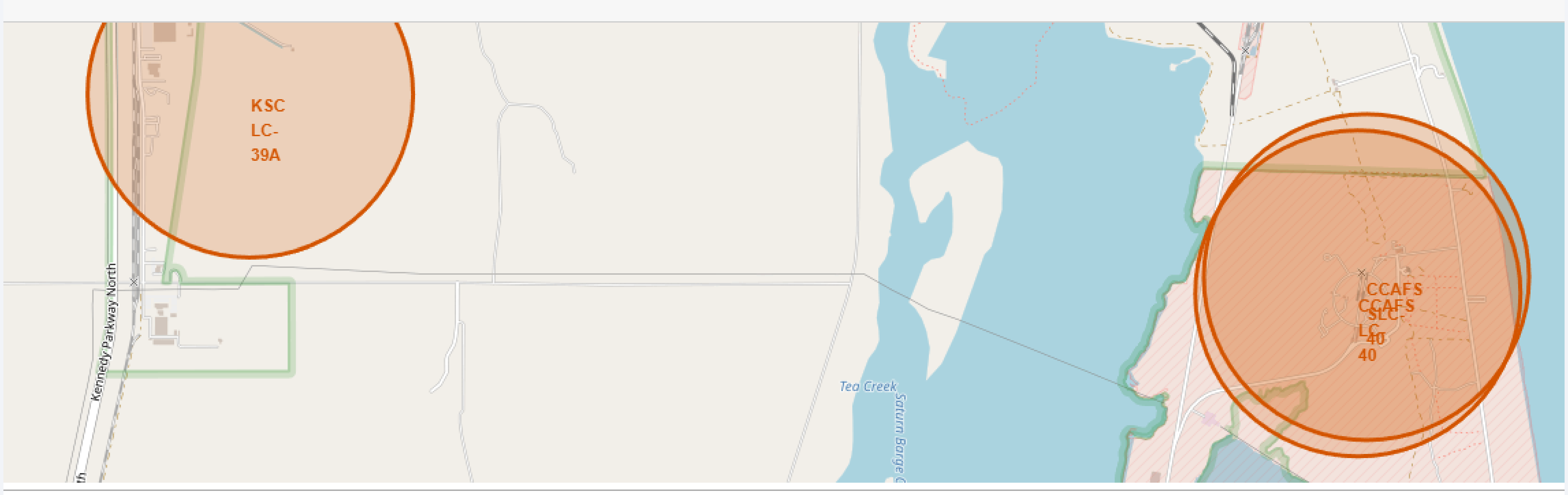
All Launch Site’s Locations



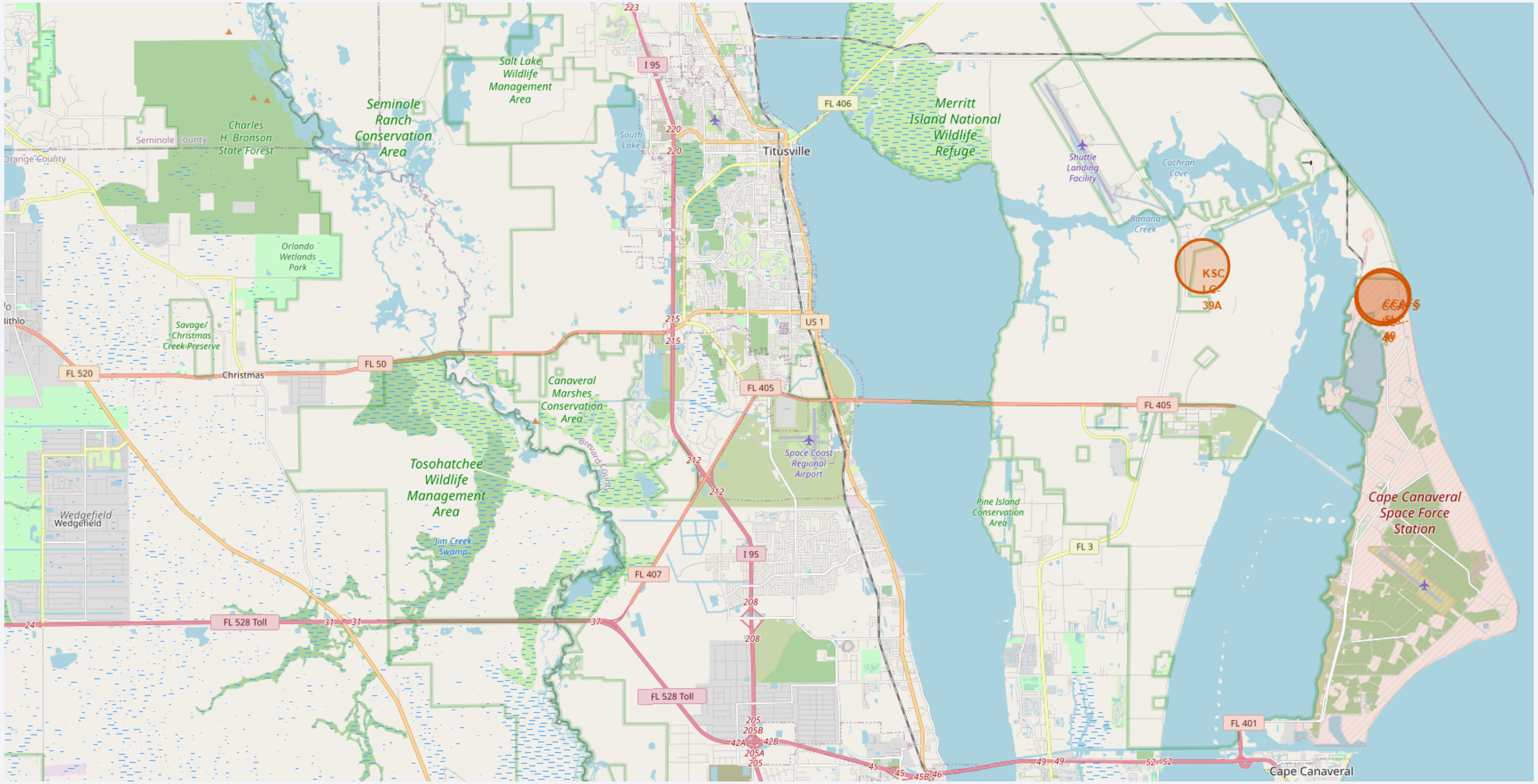


- The left map show all SpaceX launch site, the right map also show that all launch sites are in the US

Color-labeled Launch Outcomes

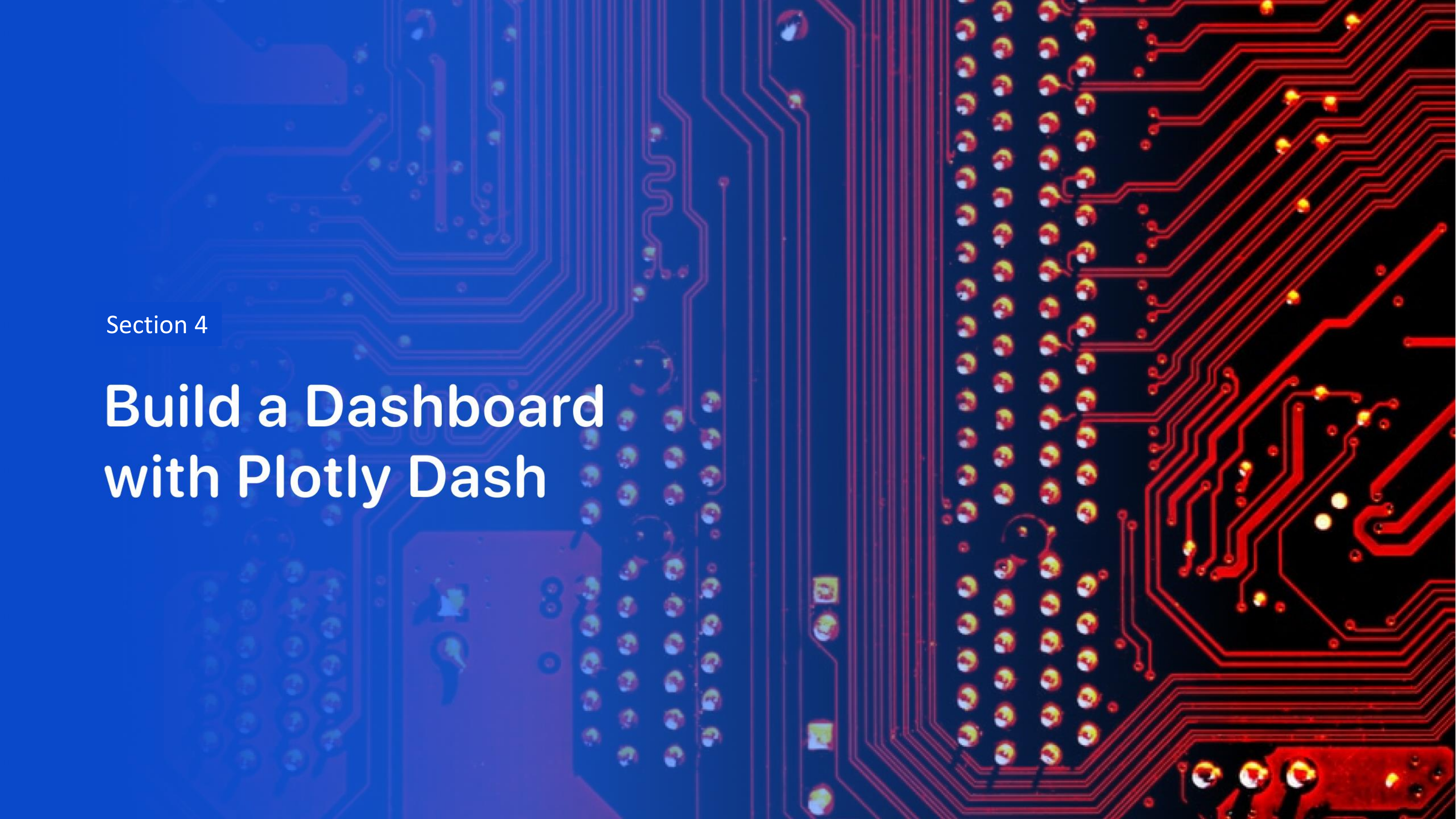


Proximites of Launch Sites

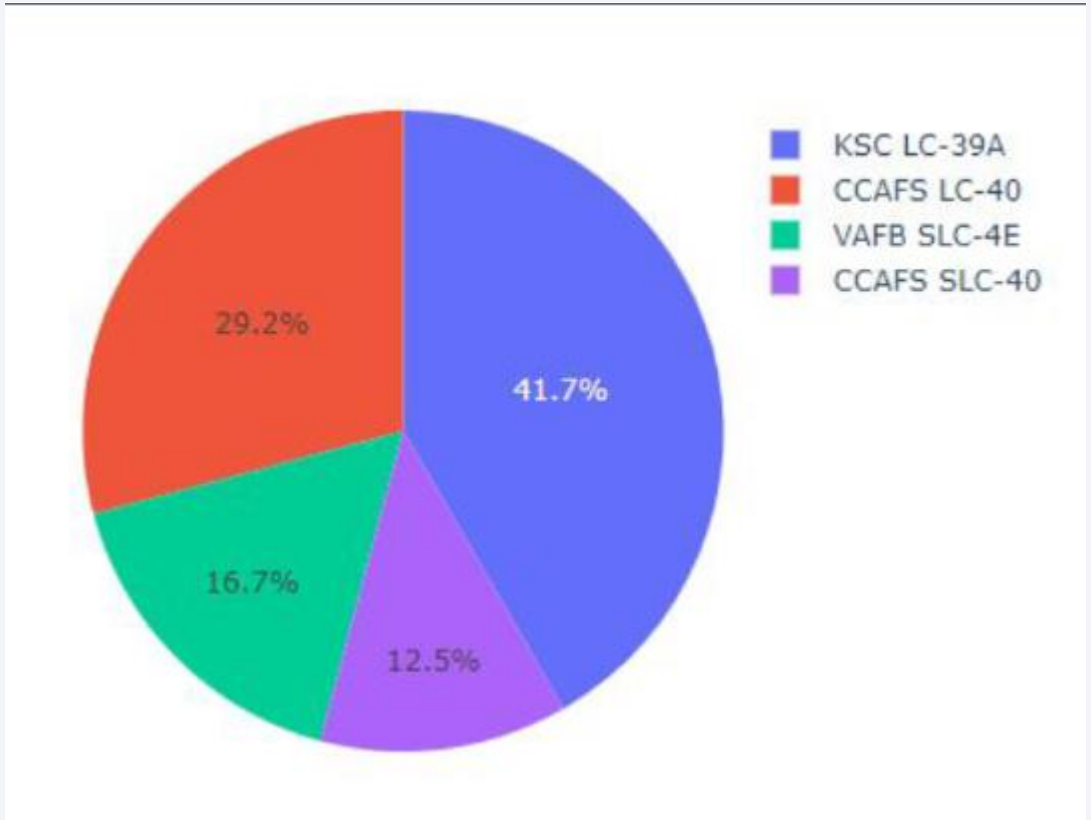

KSC LC-39A
CCAFS SLC-40
Titusville
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
Cochran Cove
Banana Creek
Cape Canaveral Space Force Station
Cape Canaveral
Seminole Ranch Conservation Area
Salt Lake Wildlife Management Area
Charles H. Bronson State Forest
Orlando Wetlands Park
Savage/ Christmas Creek Preserve
Christmas
Canaveral Marshes Conservation Area
Tosohatchee Wildlife Management Area
Jim Creek Swamp
Space Coast Regional Airport
Pine Island Conservation Area
Wedgefield
Seminole County
Orange County
Brevard County
South Lake
I 95
US 1
FL 406
FL 405
FL 50
FL 520
FL 407
FL 3
FL 401
FL 528 Toll

Section 4

Build a Dashboard with Plotly Dash

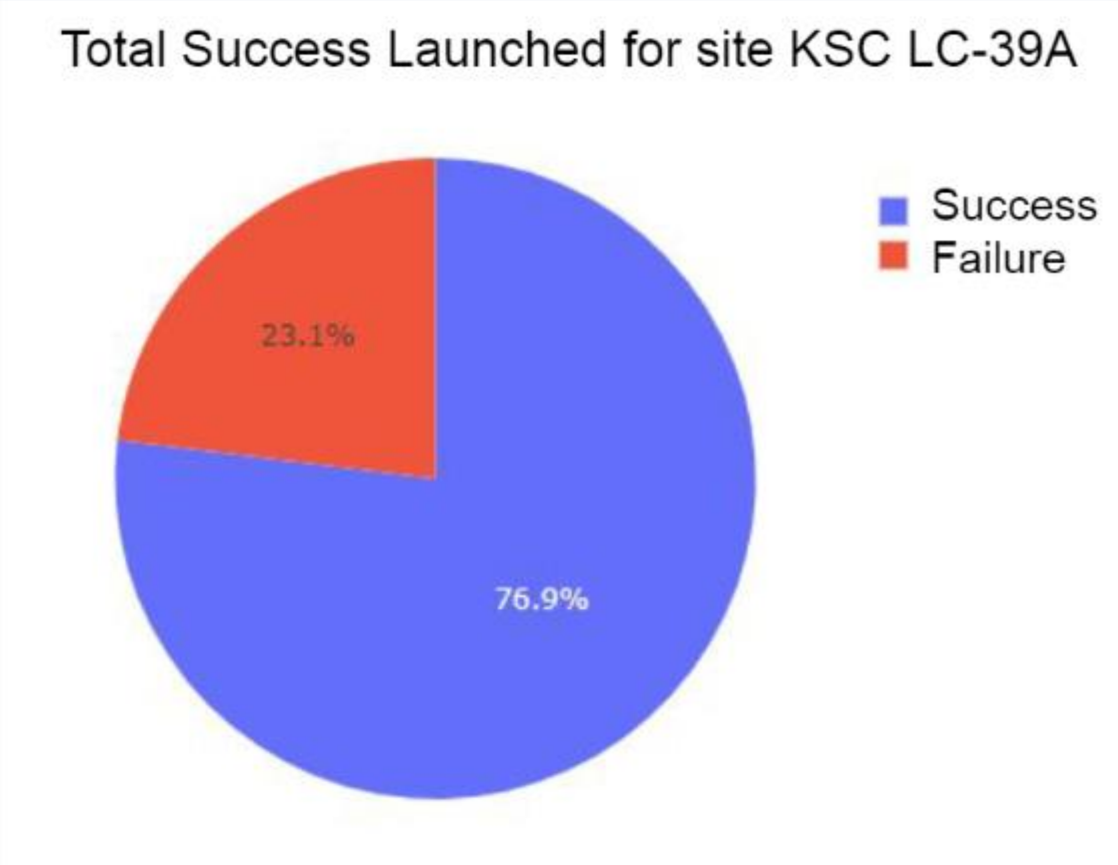


Total Success Launches By all sites



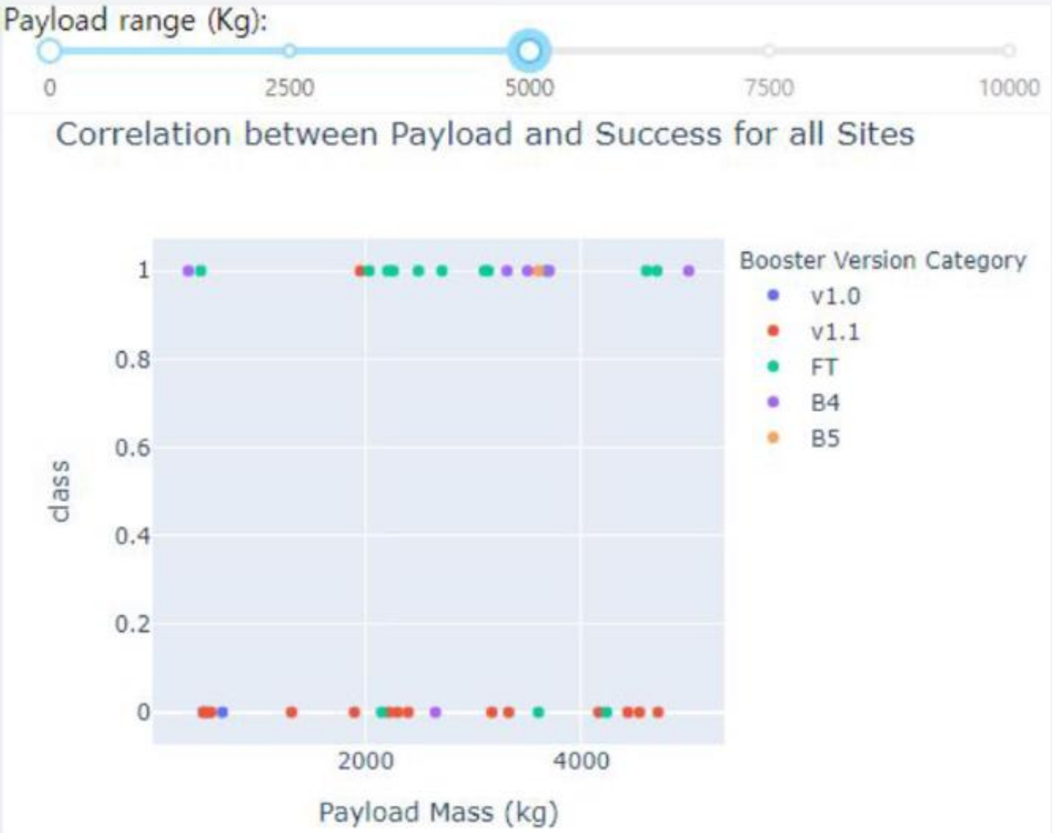
- KSC LC-39A : highest success
- CCAFS SLC-40 the fewest

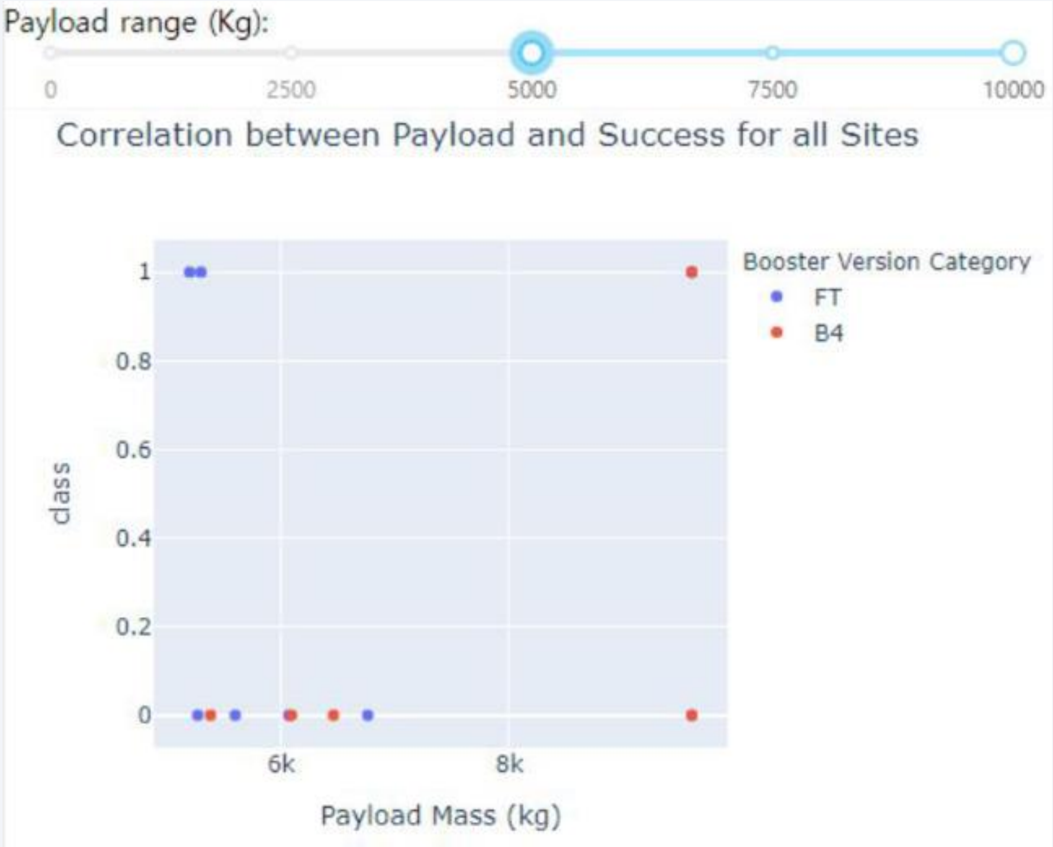
Launch Site with highest Launch Success Ratio



- KSLC-39A has the highest success ratio 76,9% and fail is 23,1%.

Payload vs Launch Outcome scatter plot for all sites





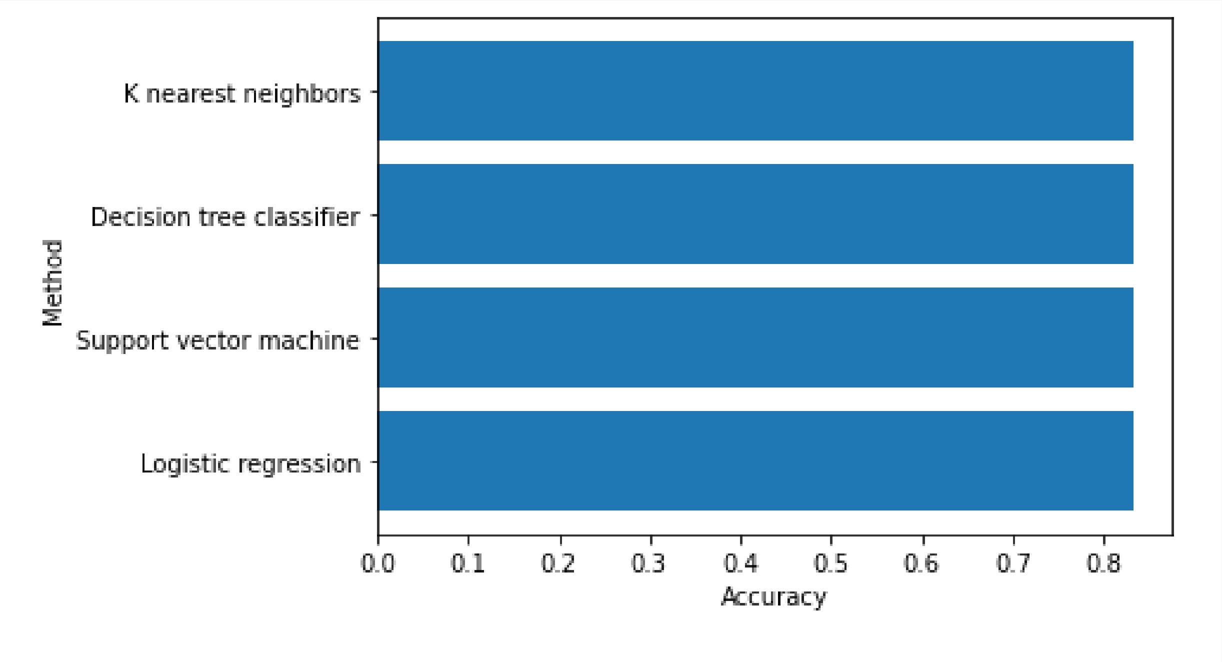
- These figure show that the launch success rate (class1) for low weighted payloads(0-5000kg) is higher than heavy weight payload

Section 5

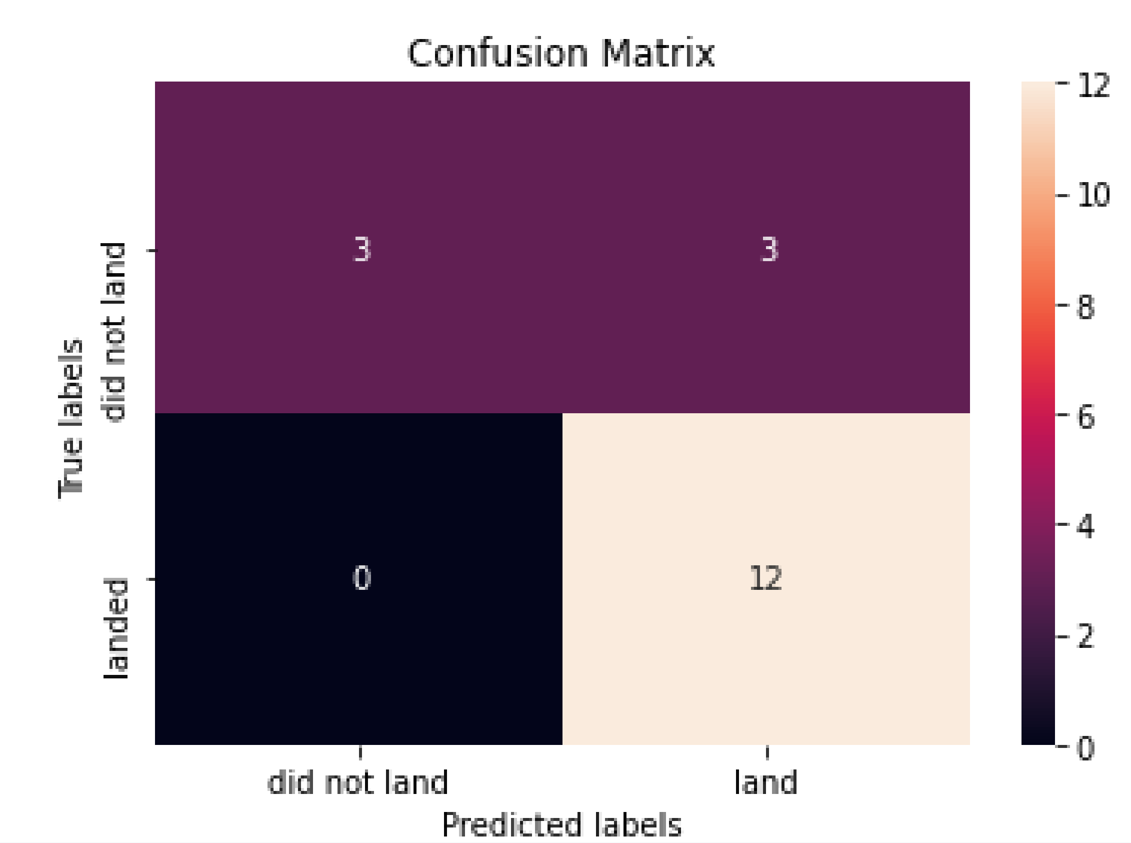
Predictive Analysis (Classification)

Classification Accuracy

- The accuracy of all models are the same 83,3% in test set, may test size is small



Confusion Matrix



- The confusion matrix is the same for all models because all models performed the same for the test set
- These models predict successful landings

Conclusions

- Orbital types SSO, HEO, GEO and ES-L1 have highest success rate
- The launch success rate of low weighted payload is higher than heavy weight payloads
- The success rate increased from 2013
- In this dataset , all models have the same accuracy 83.33%

Appendix

- GitHub
- Applied Data Science Capstone by IBM (Coursera)

Thank you!