IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ihsan Gunay
06/27/22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection - Api
  - Data Collection - Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Visual Analytics with Plotly
  - Machine Learning
- Summary of all results
  - SQL findings
  - Visual insights
  - Machine learning predictions

Introduction

- SpaceX has gained worldwide attention for a series of historic milestones.
    - It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
    - 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.
- Target Knowledge
    - Conditions required for a successful landing
    - Features that increase the probability of a successful landing
    - Cross coupling between relevant features

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from both the SpaceX api and the relevant Wikipedia page.
- Perform data wrangling
  - One-hot encoded categorical data and threw out useless features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used cross validation for selecting the best model.

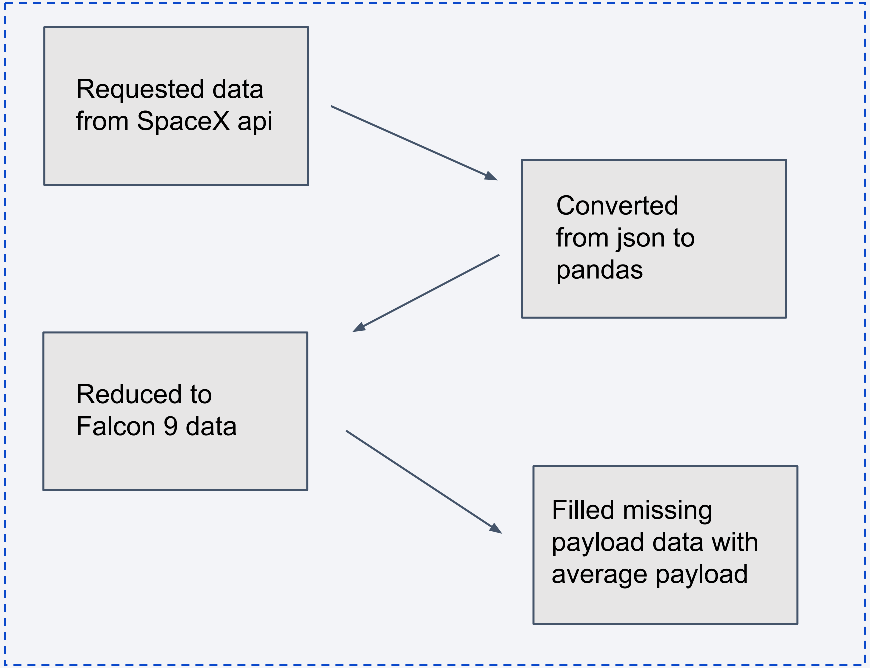
Data Collection

- Requested all of the launch data from the SpaceX api
- Loaded the data to a Pandas dataframe
- Cleaned up the data types and the inconsistencies
- Extracted the relevant Wikipedia page for supplemental public data
- Parsed and loaded the data to another dataframe
- Stored the data in csv files

Data Collection – SpaceX API

- Collected Falcon 9 data shared by SpaceX API.

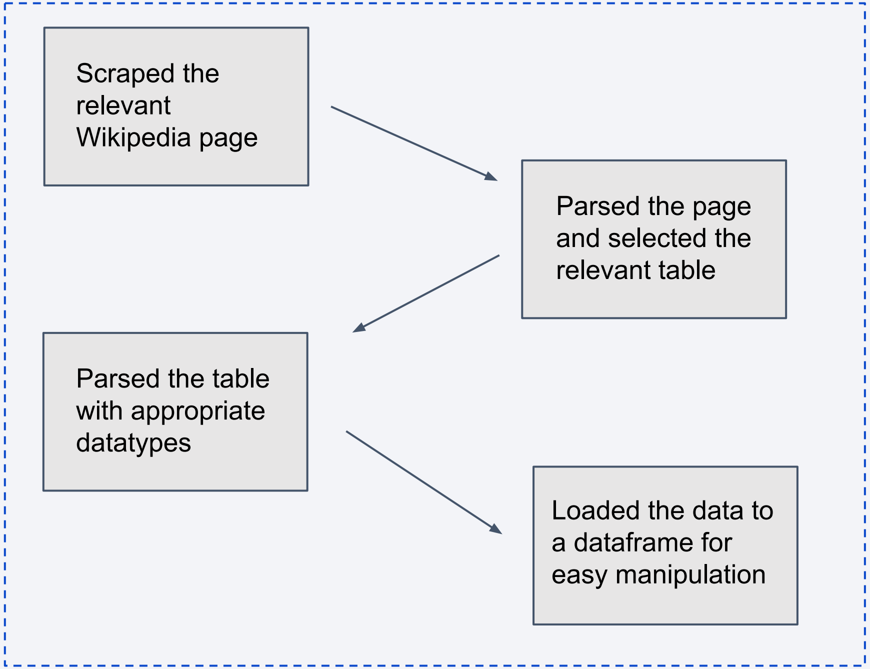
https://github.com/IhsanGunay/capstone/blob/main/1-collect-data-api.ipynb



Data Collection - Scraping

- Scraped the relevant Wikipedia page for supplementary information.

https://github.com/IhsanGunay/capstone/blob/main/2-collect-data-wiki.ipynb



Data Wrangling

- Put everything into a DataFrame
- Fixed the datatypes
- Filled missing payload values with average payload
- Deleted non Falcon9 rocket launches
- Deleted irrelevant columns

https://github.com/IhsanGunay/capstone/blob/main/3-wrangle-data.ipynb

EDA with Data Visualization

- Success vs Launch Site and Flight Number
- Success vs Launch Site and Payload
- Success rate vs Orbit
- Success vs Orbit and Flight Number
- Success vs Orbit and Payload
- Success rate vs Year

https://github.com/IhsanGunay/capstone/blob/main/4-eda-pandas.ipynb

EDA with SQL

- Number of distinct launch sites
- Launch sites that begin with ‘CCA’
- Total payload mass
- Average payload mass
- First successful ground landing date
- Booster names with mid payload mass
- Launch count by landing outcome
- Boosters with maximum payload
- Failed landings on drone ships
- Outcome rankings between 2010 and 2017

https://github.com/IhsanGunay/capstone/blob/main/5-eda-sql.ipynb

Build an Interactive Map with Folium

- Marked all launch sites with circles and their names
- Created clusters of all launches
- Color coded launches with successful landings
- Marked the distance to nearest highway, railroad, city, and coast

https://github.com/IhsanGunay/capstone/blob/main/6-visualize-data.ipynb

Build a Dashboard with Plotly Dash

- Built a pie chart option for launch site division of successful landings
- Built a pie chart option for success rate at each launch site
- Built a scatter plot with custom payload range and mission success

https://github.com/IhsanGunay/capstone/blob/main/dashboard.py

Predictive Analysis (Classification)

- Standardized the data
- Split the dataset into training and testing set
- Grid searched best parameters for logistic regression, svc, decision tree and k nearest neighbor algorithms
- Selected best performing model

https://github.com/IhsanGunay/capstone/blob/main/7-model-data.ipynb

Results

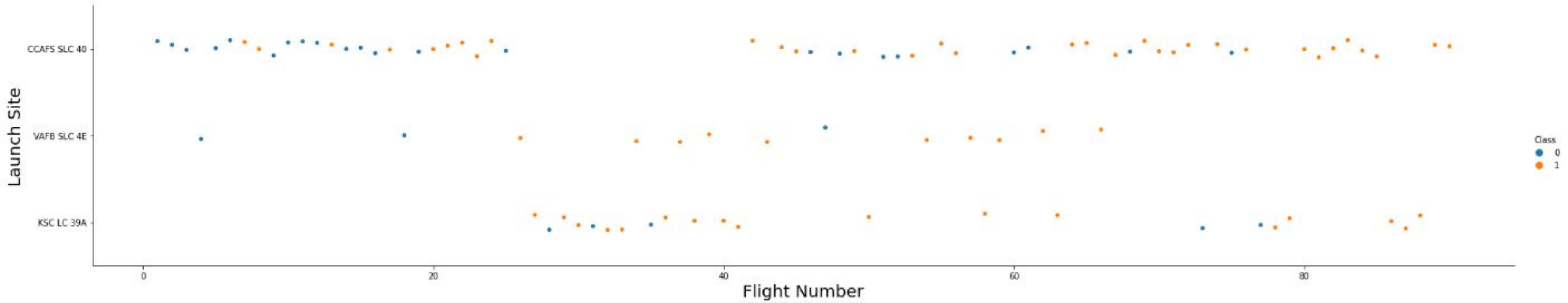
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

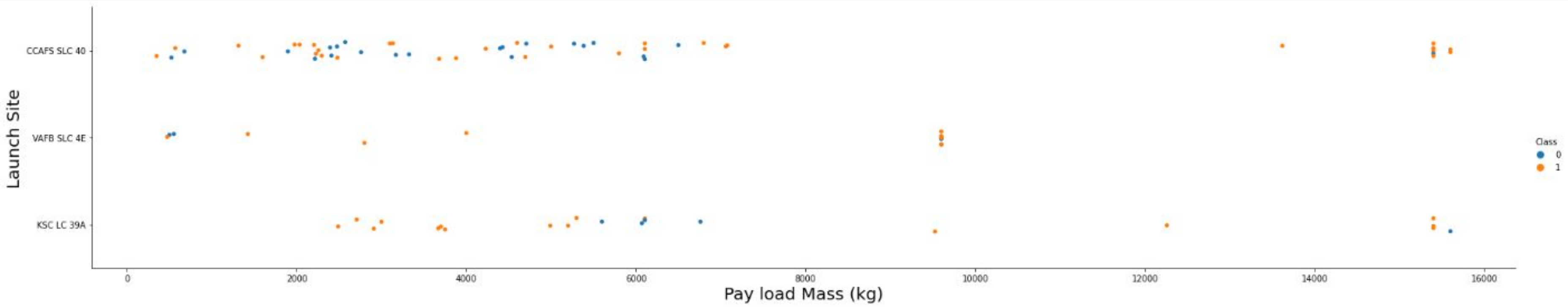
Flight Number vs. Launch Site

- As SpaceX launches more rockets, the probability of a successful launch increases at every launch site



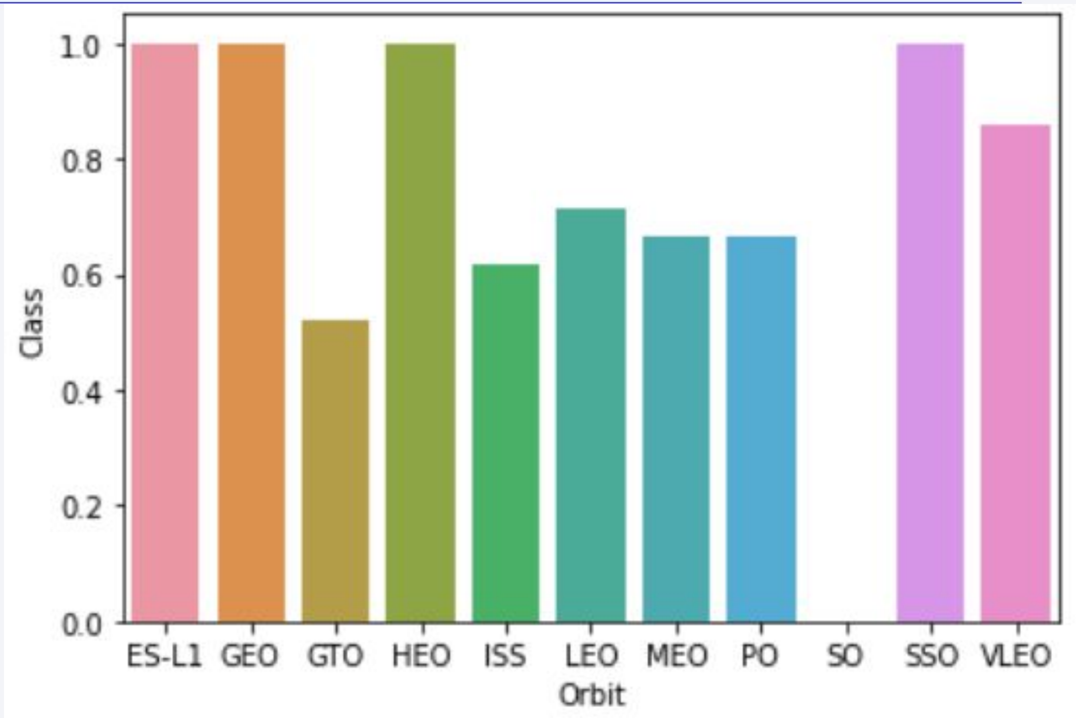
Payload vs. Launch Site

- Medium payloads have low probability of success at KSC LC 39A



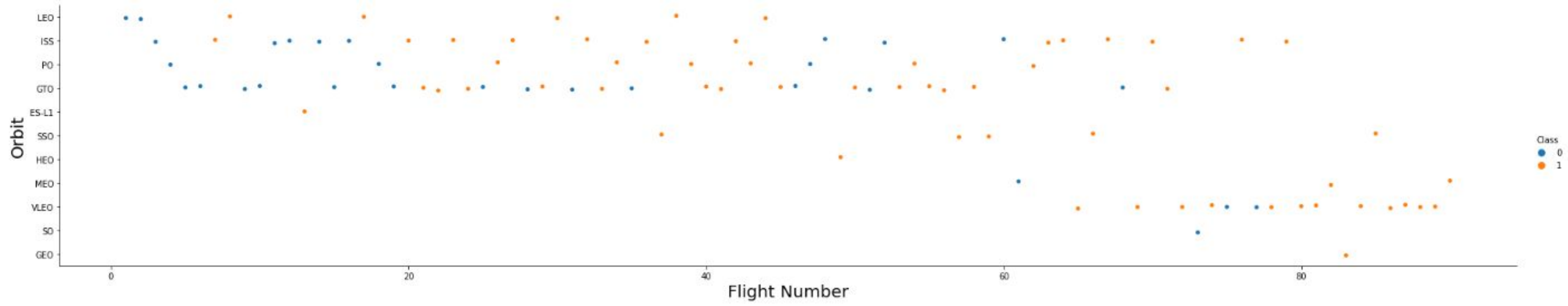
Success Rate vs. Orbit Type

- Target orbit affects the success rate.



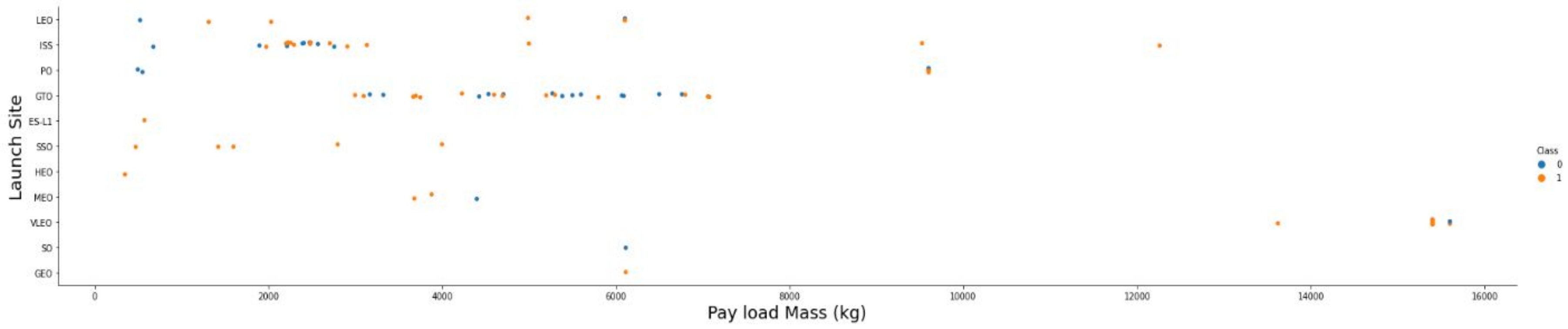
Flight Number vs. Orbit Type

- As SpaceX gained experience, they started to send payloads to more orbits.



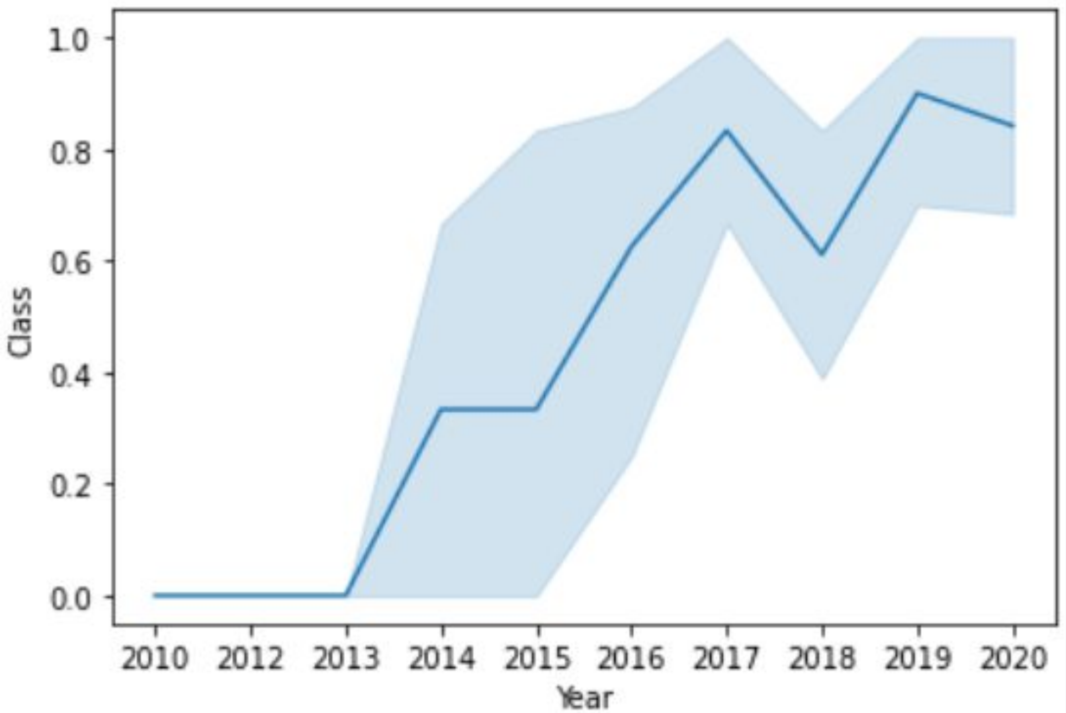
Payload vs. Orbit Type

- For some orbits, sending a lower payload lowers the probability of a successful landing.



Launch Success Yearly Trend

- Success rate has been increasing since 2013.



All Launch Site Names

- There are 4 distinct launch sites:
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40

Display the names of the unique launch sites in the space mission

In [5]:
```
%%sql
SELECT DISTINCT(LAUNCH_SITE)
FROM DATA
```

```
 * duckdb:///:memory:
```

Out[5]:

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [6]:

```
%%sql
SELECT *
FROM DATA
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
 * duckdb:///:memory:
```

Out[6]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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

- NASA sent 107,010 kg of payload using SpaceX rockets.

Display the total payload mass carried by boosters launched by NASA (CRS)

In [7]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM DATA
WHERE CUSTOMER LIKE '%NASA%'
```

```
 * duckdb:///:memory:
```

Out[7]:

	sum("PAYLOAD_MASS__KG_")
0	107010

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [8]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM DATA
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

```
 * duckdb:///:memory:
```

Out[8]:

	avg("PAYLOAD_MASS__KG_")
0	2534.666667

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [9]:
```
%%sql
SELECT MIN(DATE)
FROM DATA
WHERE LANDING_OUTCOME LIKE 'Success (ground pad)'
```

```
 * duckdb:///:memory:
```

Out[9]:

	min("DATE")
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [10]:
```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_, LANDING_OUTCOME
FROM DATA
WHERE LANDING_OUTCOME LIKE 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ > 4000
    AND PAYLOAD_MASS__KG_ < 6000
```

```
 * duckdb:///:memory:
```

Out[10]:

	Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
0	F9 FT B1022	4696	Success (drone ship)
1	F9 FT B1026	4600	Success (drone ship)
2	F9 FT B1021.2	5300	Success (drone ship)
3	F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [11]:

```
%%sql
SELECT LANDING_OUTCOME, COUNT(*)
FROM DATA
GROUP BY LANDING_OUTCOME
```

```
 * duckdb:///:memory:
```

Out[11]:

	Landing_Outcome	count_star()
0	Failure (parachute)	2
1	No attempt	21
2	Uncontrolled (ocean)	2
3	Controlled (ocean)	5
4	Failure (drone ship)	5
5	Precluded (drone ship)	1
6	Success (ground pad)	9
7	Success (drone ship)	14
8	Success	38
9	Failure	3
10	No attempt	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [12]:

```
%%sql
SELECT BOOSTER_VERSION
FROM DATA
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM DATA)
```

```
 * duckdb:///:memory:
```

Out[12]:

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	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

In [13]:
```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME, DATE
FROM DATA
WHERE LANDING_OUTCOME LIKE 'Failure (drone ship)'
    AND YEAR(DATE) = 2015
```

```
 * duckdb:///:memory:
```

Out[13]:

	Booster_Version	Launch_Site	Landing_Outcome	Date
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

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

In [14]:
```
%%sql
SELECT LANDING_OUTCOME, COUNT(*) AS COUNT
FROM DATA
WHERE DATE > '2010-06-04'
AND DATE < '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY COUNT DESC
```

```
 * duckdb:///:memory:
```

Out[14]:

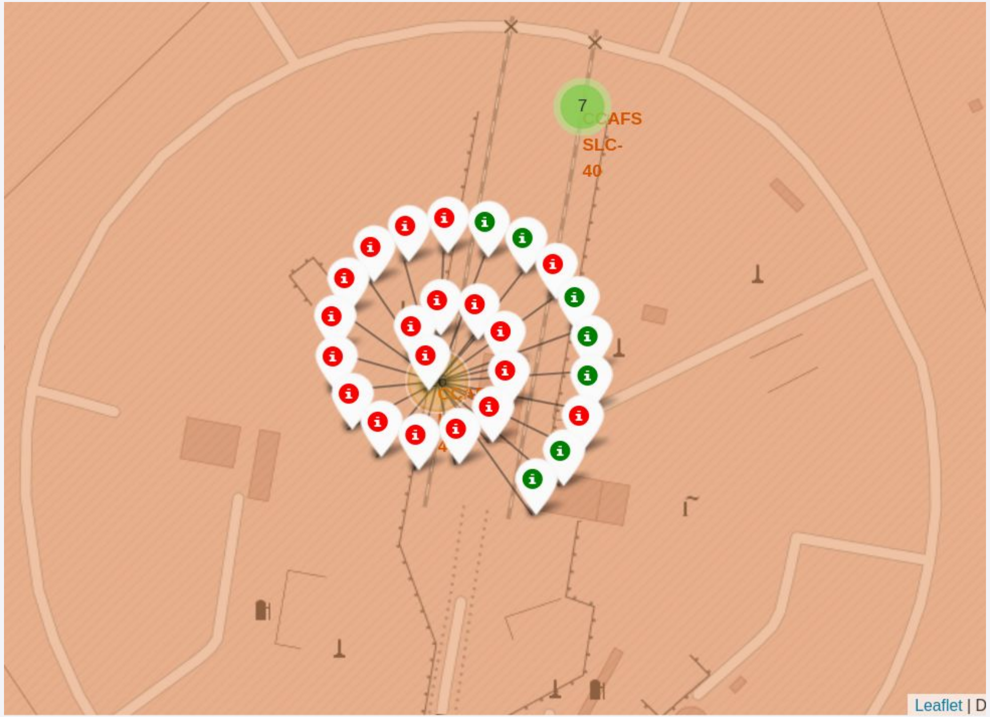
	Landing_Outcome	COUNT
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	1
7	Precluded (drone ship)	1

Section
3

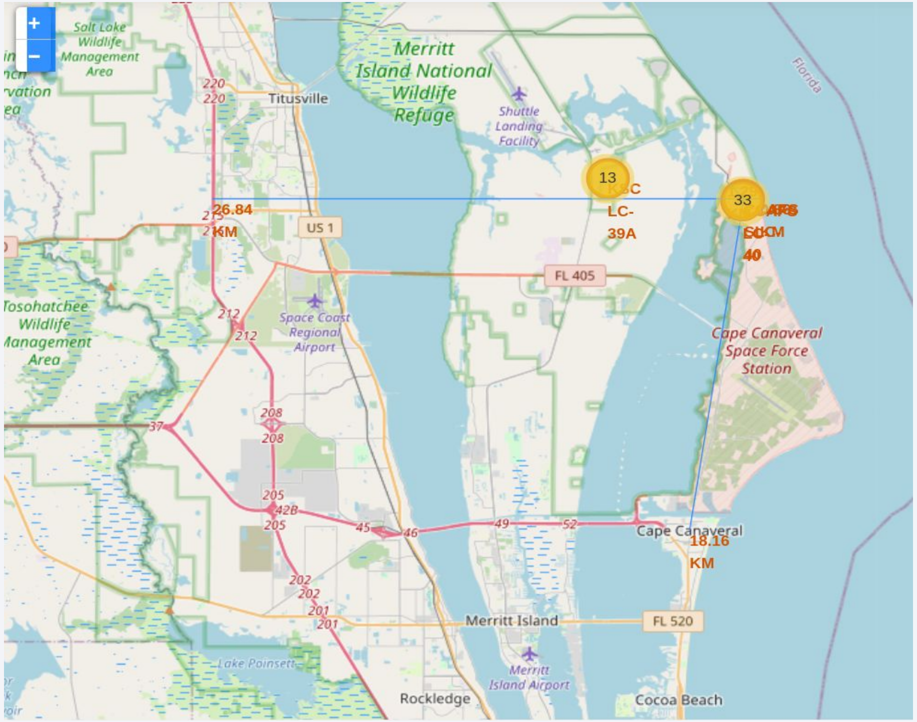
Launch Sites Proximities Analysis

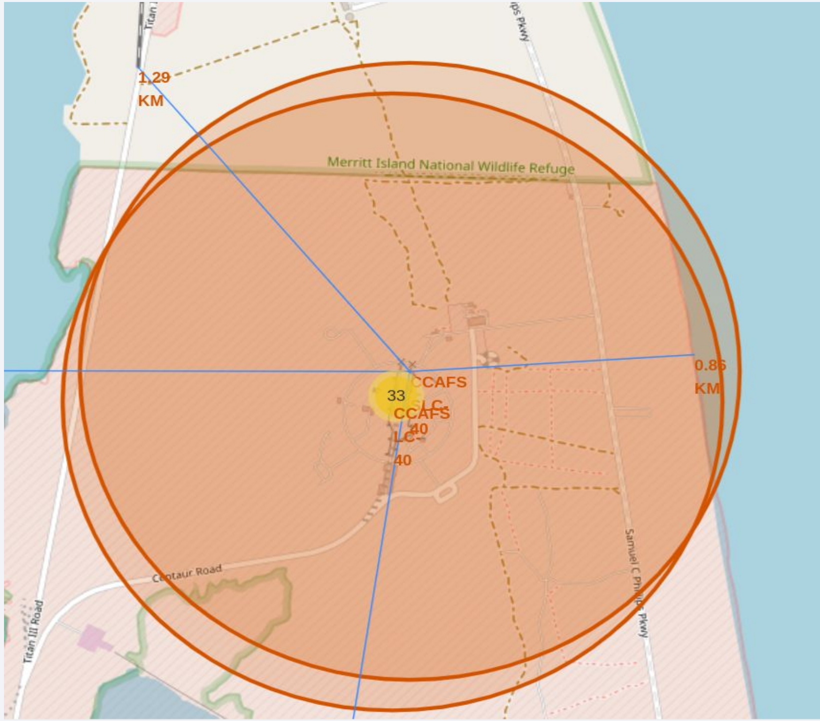
Locations of the launch sites

Cluster of launch outcomes at each site



Proximity to railroads, highways, cities and coast



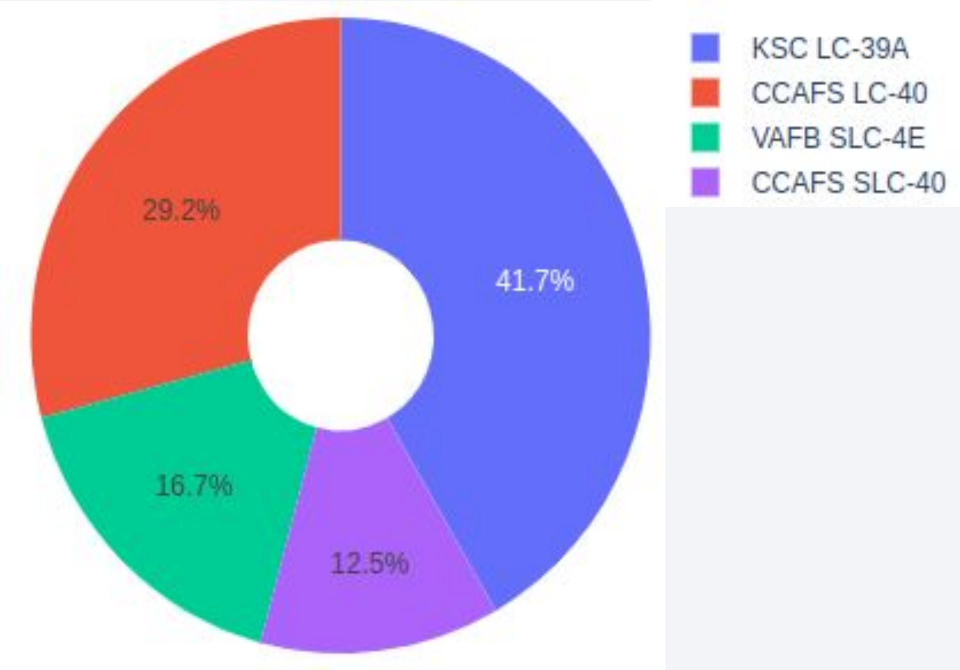


Section
4

Build a Dashboard with Plotly Dash

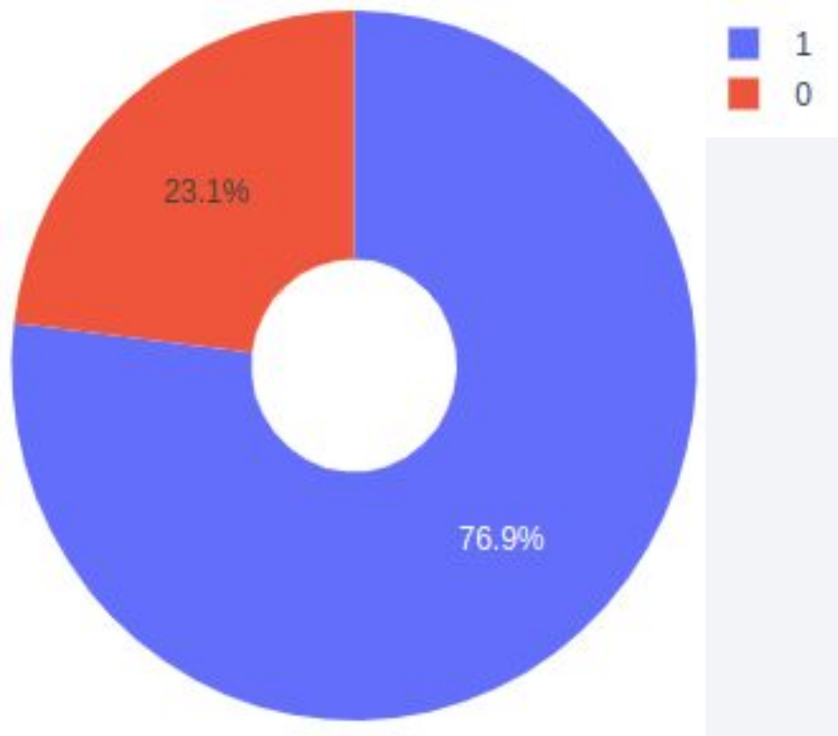
Successful launches by launch sites

- 70% of the successful launches were from 2 launch sites.



Success rate at the most successful site

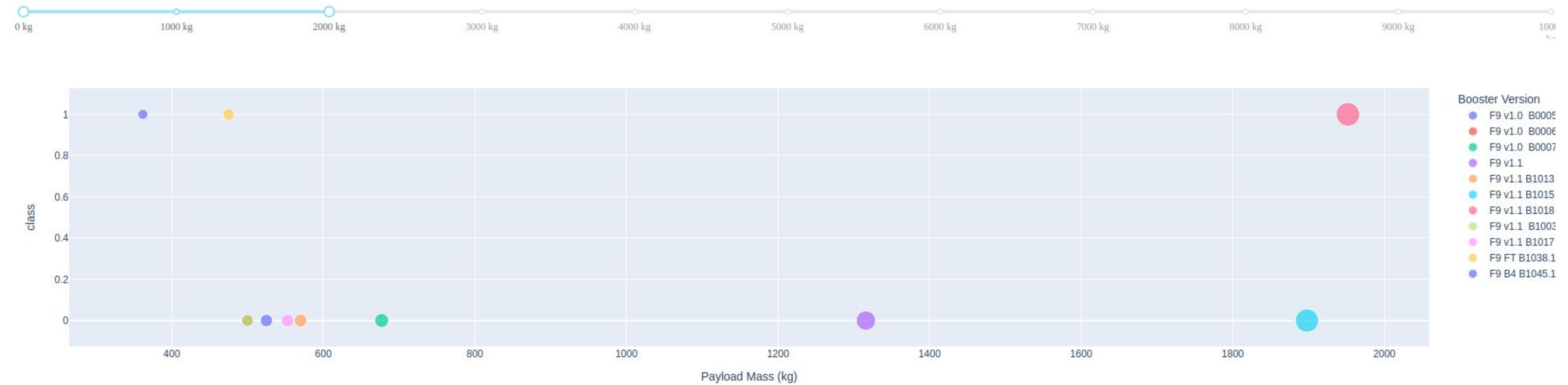
- Over three quarters of the launches from KSC LC-39A landed successfully.



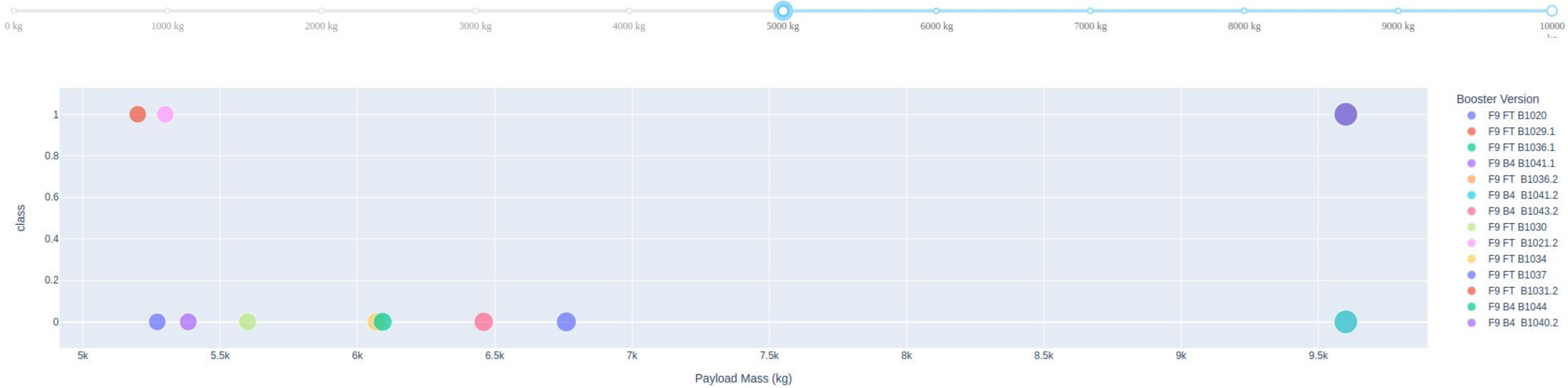
Success Rate by Payload

- Success rate decreases with too much and too little payload.

Payload range (Kg):



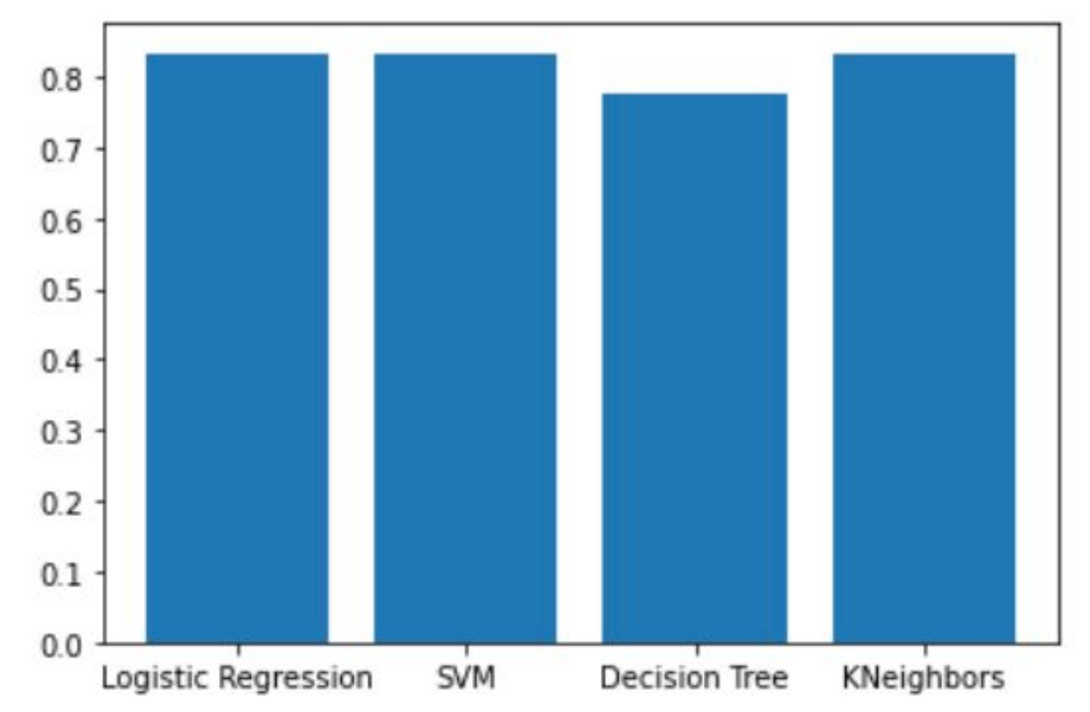
Payload range (Kg):



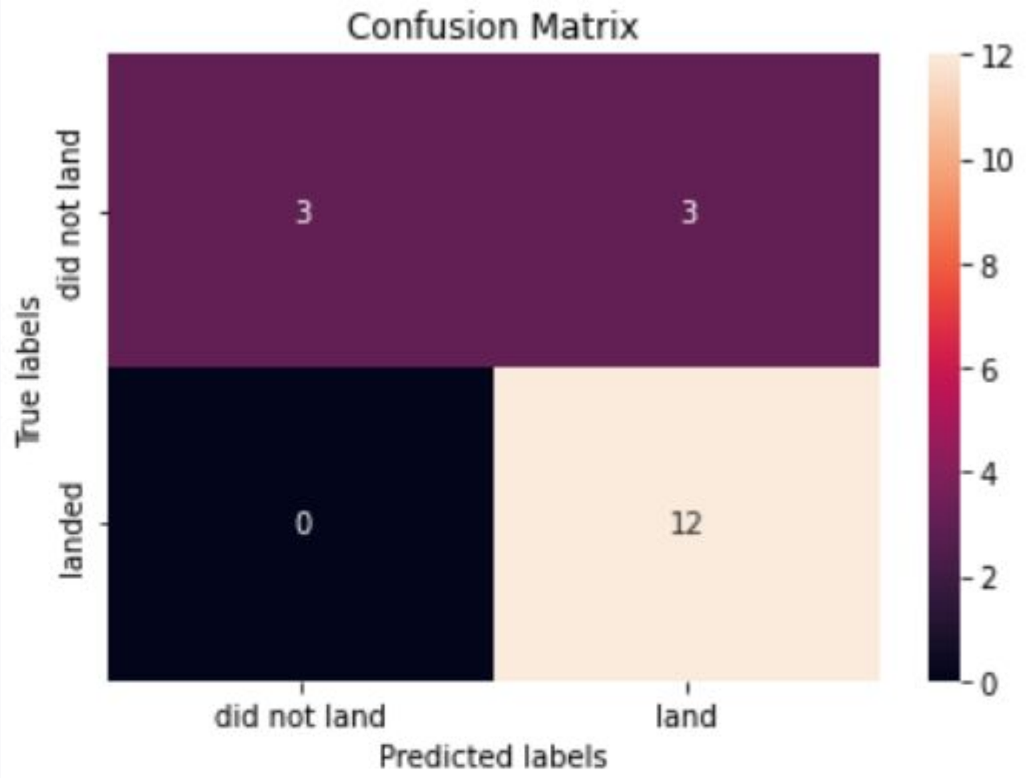
Section
5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- KSC LC-39A has the highest success rate
- Success rate increases with the number of launches
- Success rate increases continuously from 2013
- Some orbits have higher success rate than others
- Classification algorithms perform similarly

Thank you!