IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sorawit Lertwilaikulnatee
31/07/2024

Outline

Executive Summary

Entering into the space business is a big decision like the spaceship itself. You need to know what the current players in the business are doing and the best one is SpaceX. It can reuse some parts of the rocket from flights to flights which can cut down the cost of operation tremendously. This is very interesting business point and needed to be studied.

First, we can study the statistics of landing outcome which can infer the major cost of operation. To study this topic, we can obtain various data from public sources and apply data science methodology to get insights from them. Then, we calculate the statistics with python programming that apply machine learning algorithm to help us extract the valuable information better than we do by ourself.

The result shows us that various topics are related to the landing outcome through many visualizations and accuracy computations from different types of machine learning model.

Introduction

- Project background and context
  - SpaceX can cut cost of launching spaceship more than other providers that might cost upward of 165 million dollars. If other companies want to enter into this business, the statistics of SpaceX launching could be vital information for them to use for the business decision.
- Problems you want to find answers
  - Prediction of landing outcome according to various factors

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data from public resources with API.
- Perform data wrangling
  - Clean and normalize data to prepare it for further computation.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use GridSearchCV to help us computing different types of classification model.
  - Compare result from each model and choose the best one.

Data Collection

- There are 2 datasets used in this analysis

1. Data from “SpaceX API” : acquire data from calling API to get the data in "https://api.spacexdata.com/v4/launches/past"

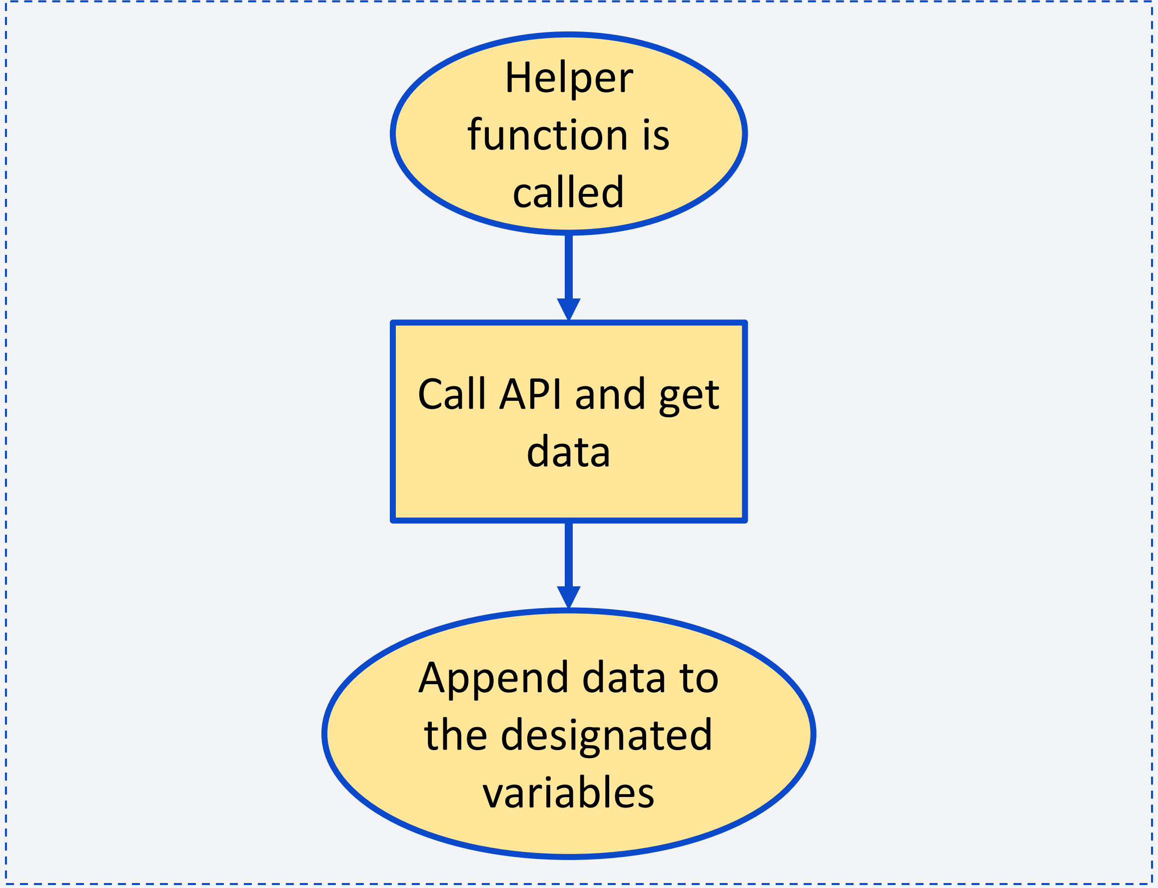
2. Data from tables in Wikipedia website* : acquire by using web scraping method. The tables contain collective series of launching information.

*https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

Acquire data from calling API to get the data in "https://api.spacexdata.com/v4/launches/past and use them in helper function getBoosterVersion(), getLaunchSite(), getPayloadData() and getCoreData().

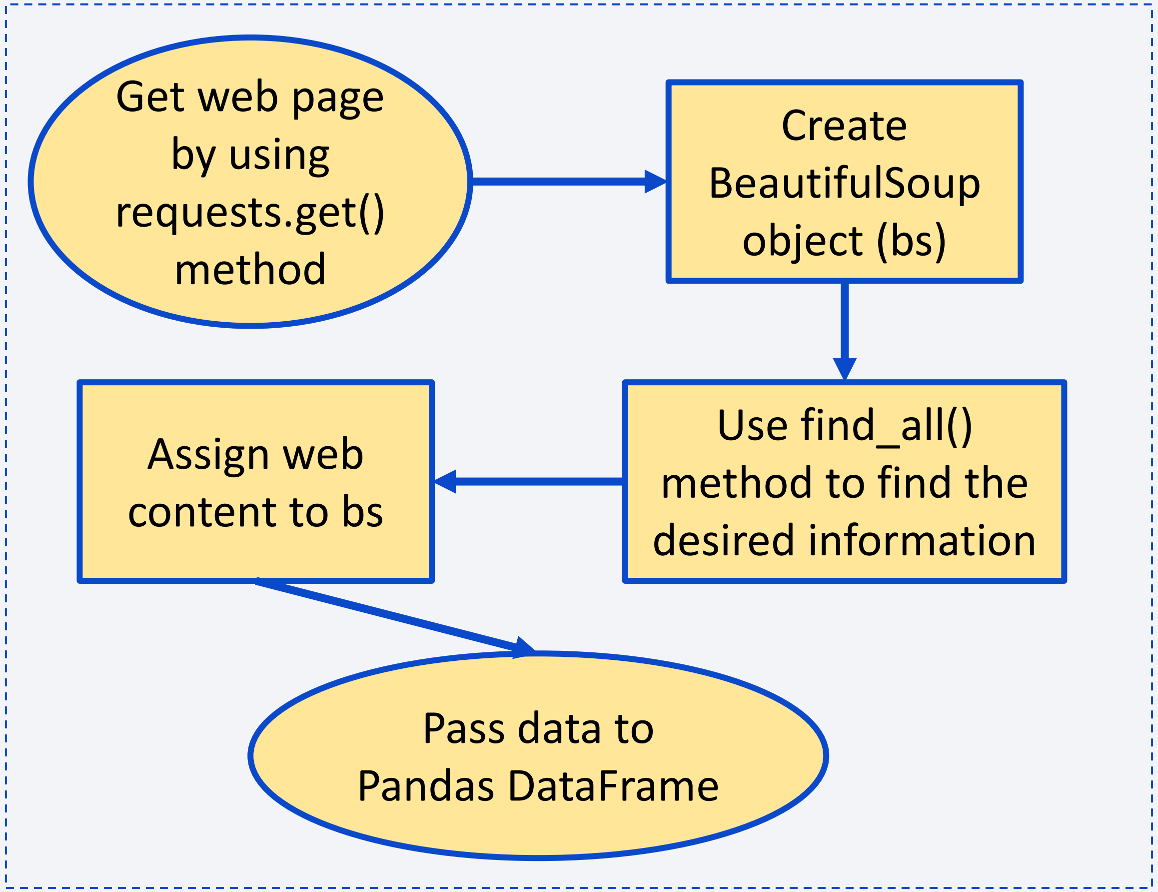
GitHub URL of the completed SpaceX API calls notebook : https://github.com/promptsrw/IBM-data-capstone/blob/693c3a87a7aa8848ea70fa50e1313ae072c086d3/C10_Lab%201_M1_%20Collecting%20the%20data.ipynb



Data Collection - Scraping

- Use BeautifulSoup library for web-scraping process to extract a Falcon 9 launch records table from Wikipedia, parse the table and convert it into a Pandas data frame for further analysis

GitHub URL of the completed web scraping notebook : https://github.com/promptsrw/IBM-data-capstone/blob/dfd9669eb5c8219e5d568582eb16f656f3595ff9/C10_Lab%202_M1_%20Space%20X%20Falcon%209%20First%20Stage%20Landing%20Prediction.ipynb



Data Wrangling

- Data was first loaded into Panda DataFrame. Then check data types of each column and availability. Calculations were performed with various method
  1. Calculate the number of launches on each site using .value_counts() method
  2. Calculate the number and occurrence of each orbit using .value_counts() method
  3. Calculate the number and occurrence of mission outcome of the orbits using .value_counts() method
  4. Create a landing outcome label from Outcome column using .mean() method

GitHub URL of your completed data wrangling related notebooks : https://github.com/promptsrw/IBM-data-capstone/blob/f0a36383ed6247995c2649ed9a0169418f427c3a/C10_Lab%203_M2_%20Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

1. Scatter plot between Flight number and Launch Site : there are improvement of success rate in every launch sites over flight number
2. Scatter plot between Pay load mass and Launch site : some launch site have no heavy payload
3. Bar plot between success rate and orbit type : success rate vary by each orbit type
4. Scatter plot between Flight number and orbit type : there are improvement of success in almost every orbit type over flight number
5. Scatter plot between Payload mass and orbit type: some orbit types have 100% success rate
6. Line plot between success rate over time (year) : improvement start from 2013

Add the GitHub URL of your completed EDA with data visualization notebook : https://github.com/promptsrw/IBM-data-capstone/blob/792d351a72755040f1ef0e7f0e033be2c24a13ff/C10_Lab%205_M2_%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

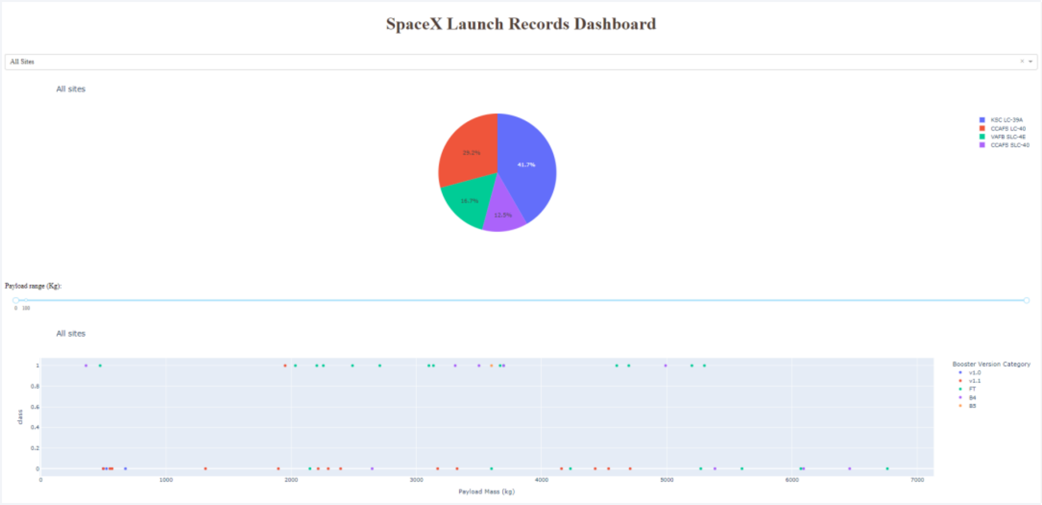
- Using bullet point format, summarize the SQL queries you performed
  - Task 1 : List unique launch site names with SELECT DISTINCT
  - Task 2 : Display 5 records where launch sites begin with the string 'CCA' with LIKE clause
  - Task 3 : Calculate total payload mass of NASA (CRS) using function SUM()
  - Task 4 : Calculate average payload mass by booster version F9 v1.1 using function AVG()
  - Task 5 : List the date when the first succesful landing outcome in ground pad was achieved using MIN() function
  - Task 6 : List the names of the boosters with conditions
  - Task 7 : List the total number of successful and failure mission outcomes
  - Task 8 : List the names of the booster_versions using subquery
  - Task 9 : List month name and failure landing
  - Task 10 : Ranking the landing outcomes
- GitHub URL of your completed EDA with SQL notebook : https://github.com/promptsrw/IBM-data-capstone/blob/c8a343a4d67b9a23aca7ba600e509d56daf0b6a4/C10_Lab%204_M2_%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map was added with
  1. NASA marker (circle) : for locating the starting point of the map
  2. Launch site markers and circles : for locating the launch sites in the dataset
  3. Marker cluster : for indicating the landing outcomes of each site
  4. Mouse position : for showing the coordination at which the mouse is currently pointing to
- GitHub URL of your completed interactive map with Folium map : https://github.com/promptsrw/IBM-data-capstone/blob/785488746d4ec3d9735bd143a28ad9fd381ec0a4/C10_Lab%206_M3_Folium_Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- “SpaceX Launch Records Dashboard” was built with Plotly Dash. It composes of 2 inputs and 2 graph. First we choose the launch site to show all of them or specific site. Then, select pay load mass. Pie chart will show the portion of success rate and the scatter plot will show the landing outcome against pay load mass categorized by Booster version
- GitHub URL of your completed Plotly Dash lab : https://github.com/promptsrw/IBM-data-capstone/blob/491a9030af869b85039c7d4b2313d5df111e67f1/C10_Lab%207_M3_Hands-on%20Lab%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

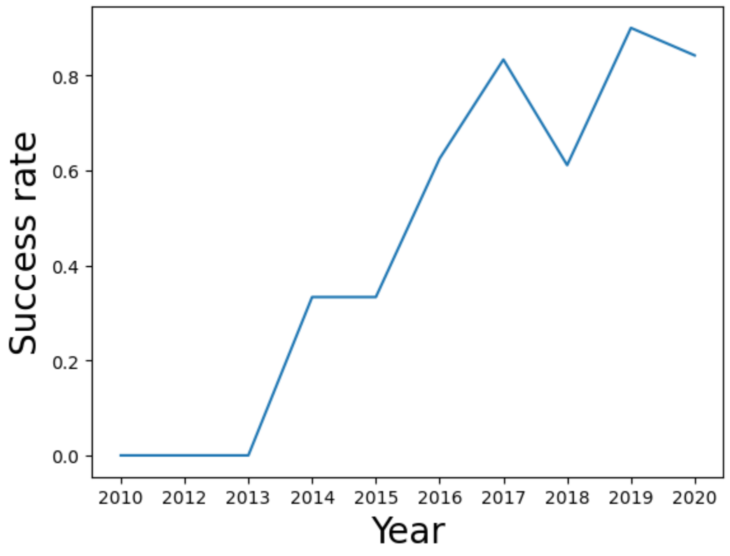


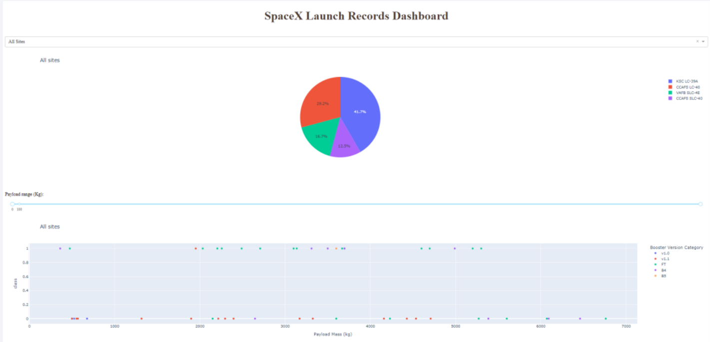
Predictive Analysis (Classification)

- Model was made by
    1. Load data into DataFrame
    2. Standardize the data values
    3. Split data into train-test datasets
    4. Using GradSearchCV to make model and evaluate the accuracy with different parameters
    5. Calculate the accuracy for each model with the best parameter from GridSearchCV
    6. Compare the accuracy of the different types of model and plot confusion matrix
- GitHub URL of your completed predictive analysis lab : https://github.com/promptsrw/IBM-data-capstone/blob/e1ecee4b16d7b7de40959e4805a17951748fac0e/C10_Lab%207_M4_Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results **fig.1**
- Interactive analytics demo in screenshots **fig.2**
- Predictive analysis results **fig.3**





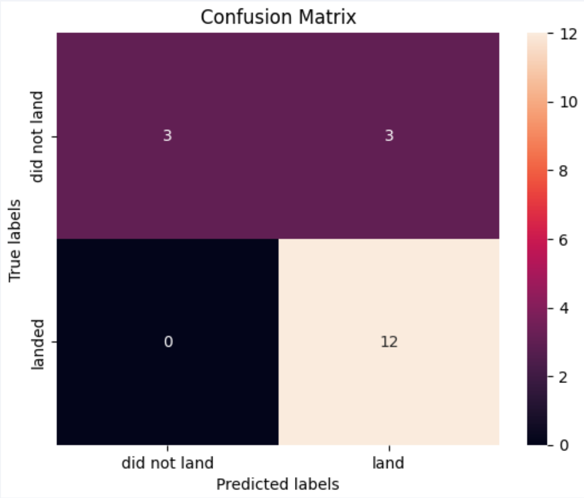


fig.1

fig.2

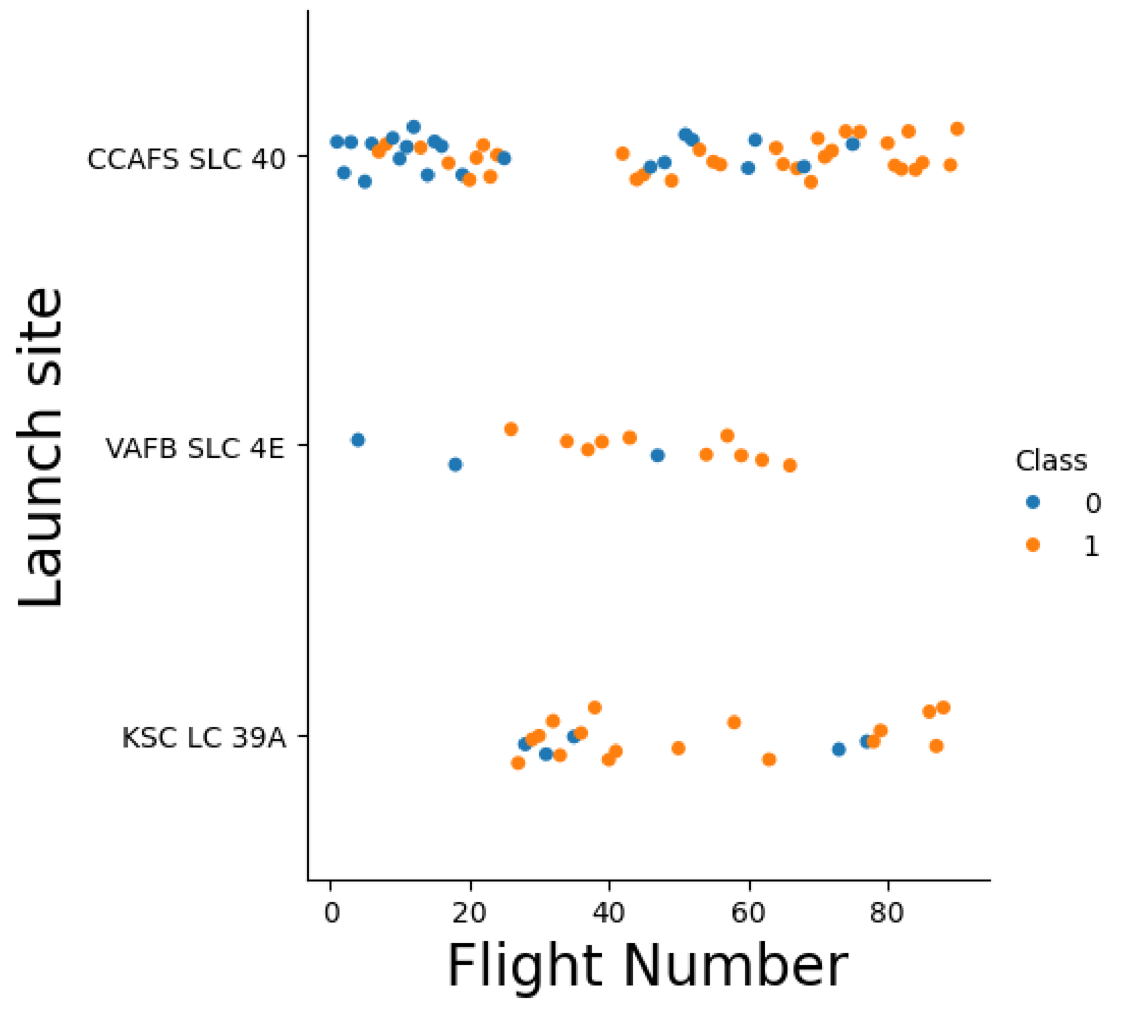
fig.3

Section 2

Insights drawn from EDA

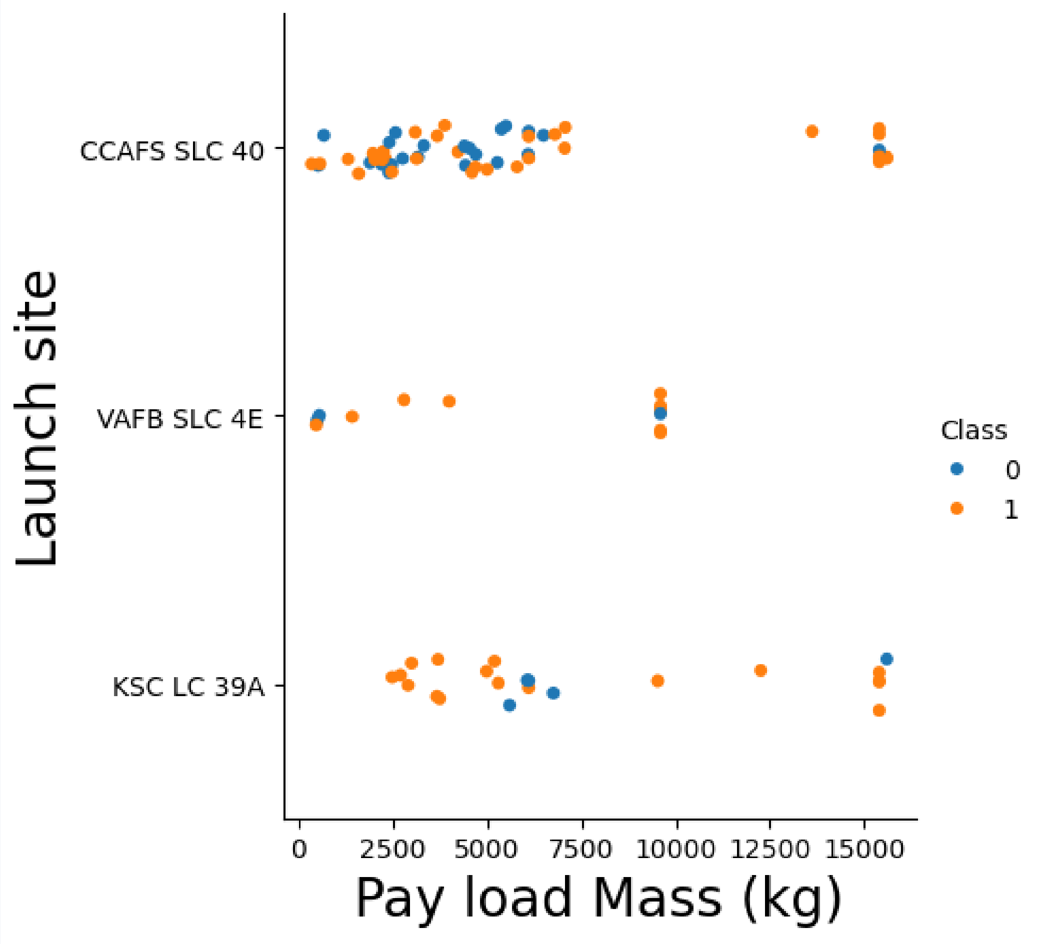
Flight Number vs. Launch Site

Scatter plot between Flight number and Launch Site : there are improvement of success rate in every launch sites over flight number



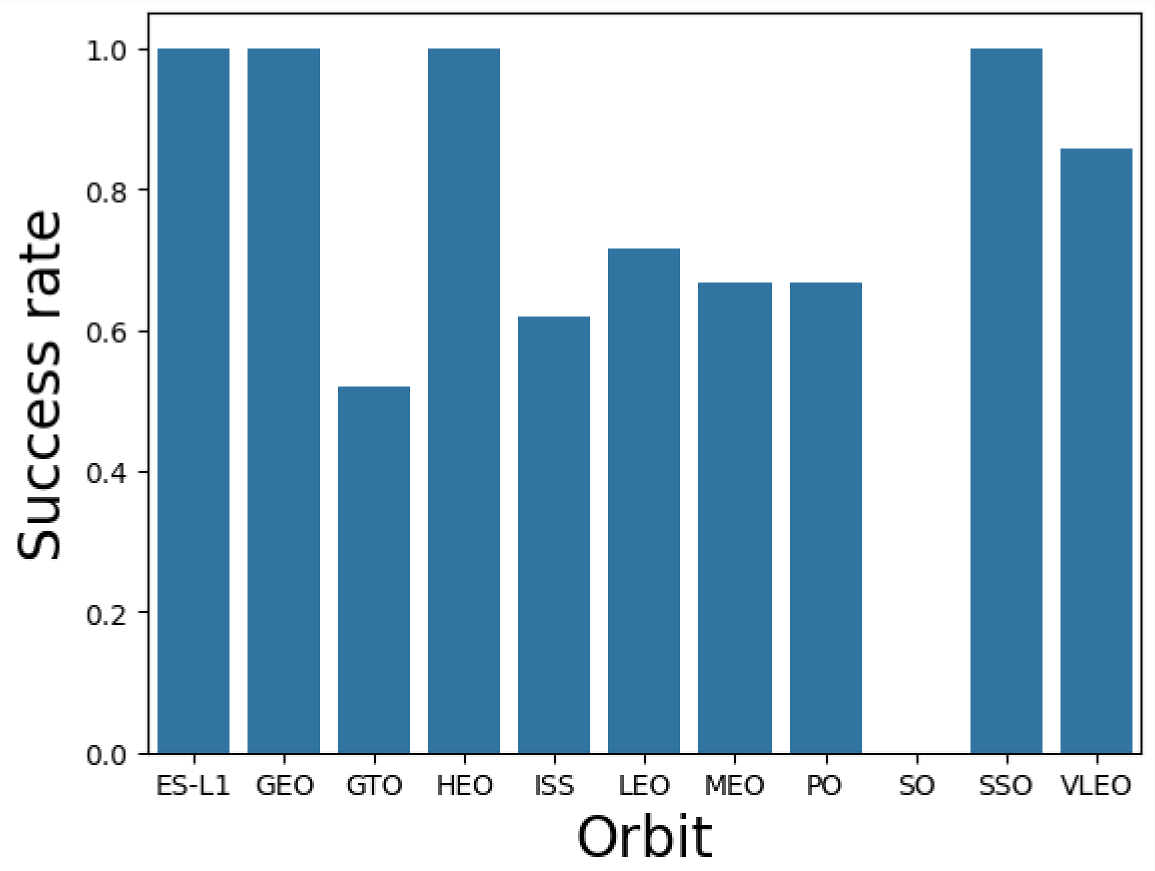
Payload vs. Launch Site

Scatter plot between Pay load mass and Launch site : some launch site have no heavy payload



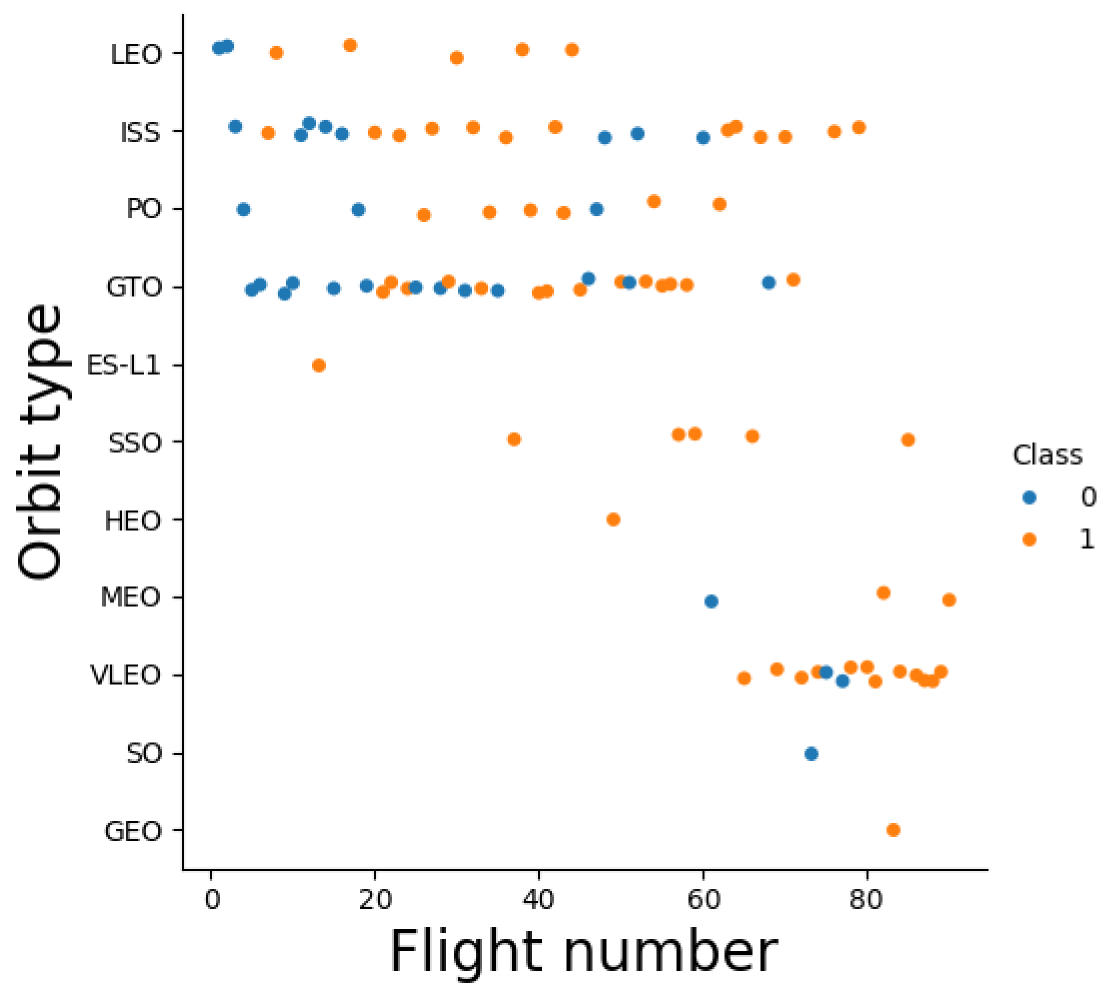
Success Rate vs. Orbit Type

Bar plot between success rate and orbit type : success rate vary by each orbit type



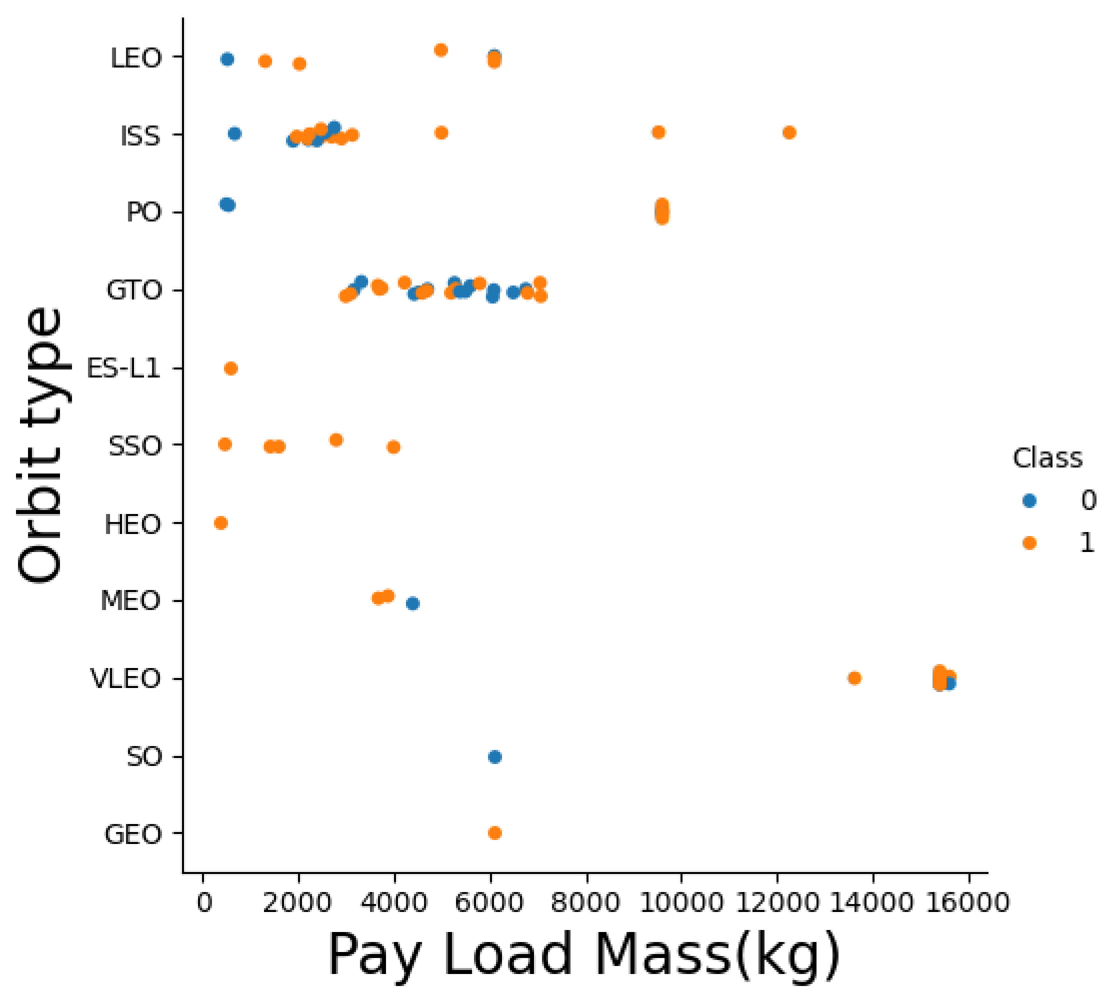
Flight Number vs. Orbit Type

Scatter plot between Flight number and orbit type : there are improvement of success in almost every orbit type over flight number



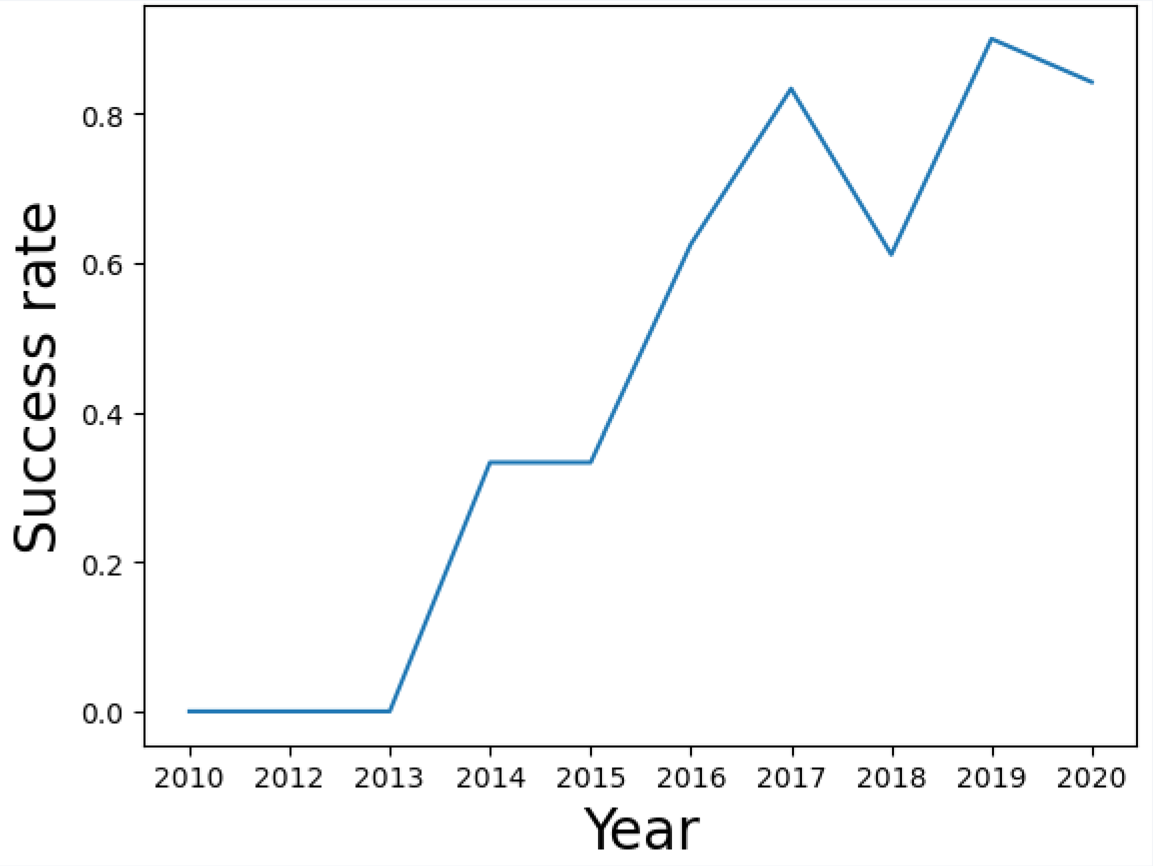
Payload vs. Orbit Type

Scatter plot between Payload mass and orbit type: some orbit types have 100% success rate



Launch Success Yearly Trend

Line plot between success rate over time (year) : improvement start from 2013



All Launch Site Names

- Use SELECT DISTINCT to list the launch site names showed in the screenshot.

```
[14]  %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE

...  * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

Python

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Use LIKE clause to list 5 records where launch sites begin with `CCA`

Total Payload Mass

- Use SUM() function to calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Use AVG() function to calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Use MIN() function to find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Use condition to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Use GROUP BY clause to calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Use sub-query to list the names of the booster which have carried the maximum payload mass

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

```
%sql SELECT SUBSTR(Date,6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,1,4) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

SUBSTR(Date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM (SELECT * FROM SPACEXTABLE WHERE Date > "2010-06-04" and Date < "2017-03-20")
GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC ;
```

```
 * sqlite:///my_data1.db
Done.
```

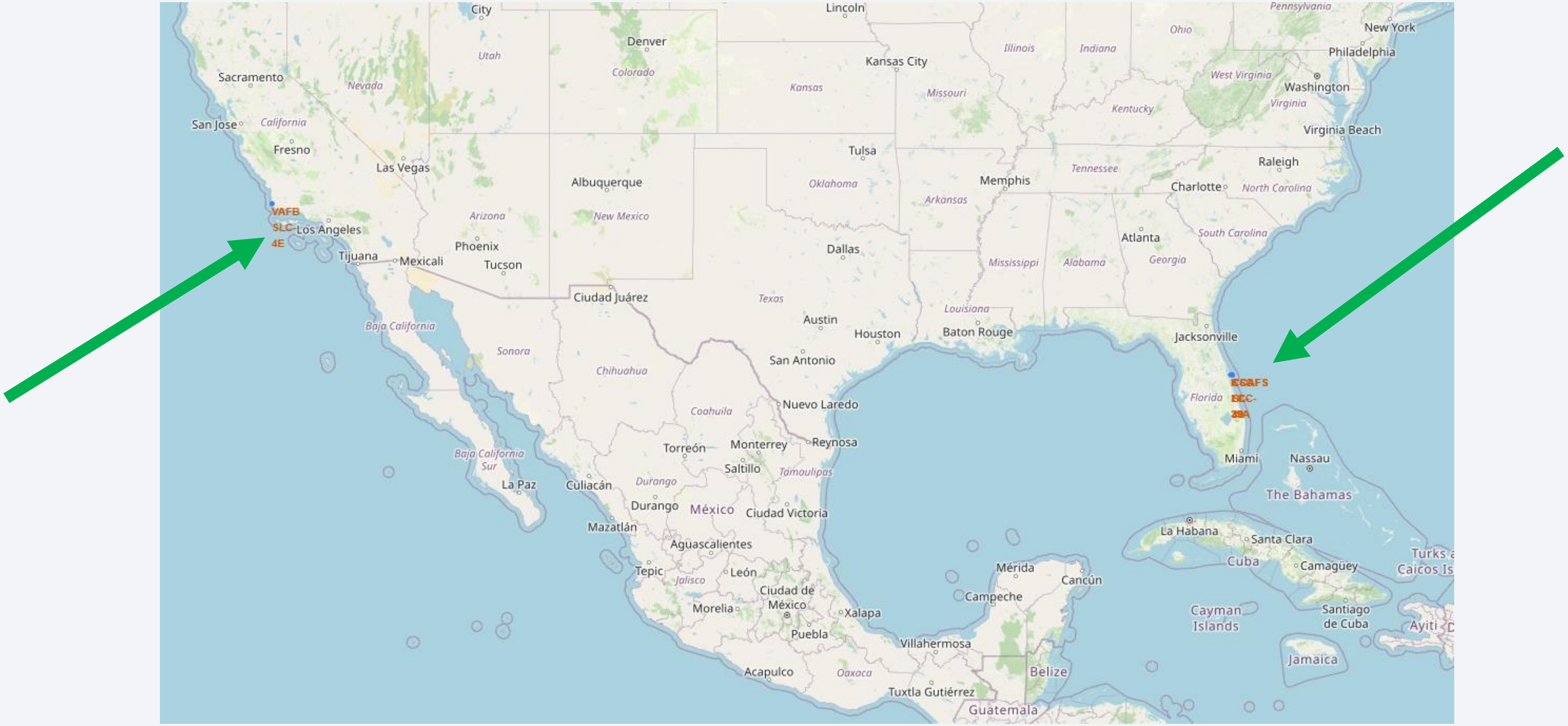
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

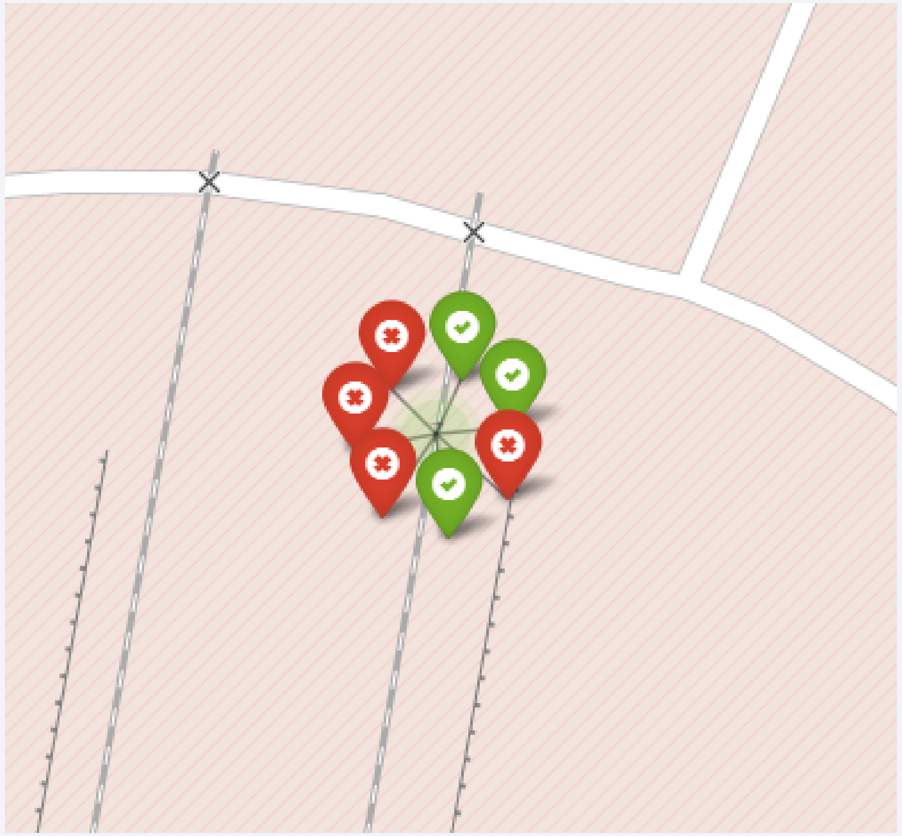
All launch site locations

- Mark on the world map of all launch site location in datasets with circles and texts



Landing outcome markers

- Mark the landing outcome of each location with guided signs that show the outcomes of each location



<Folium Map Screenshot 3>

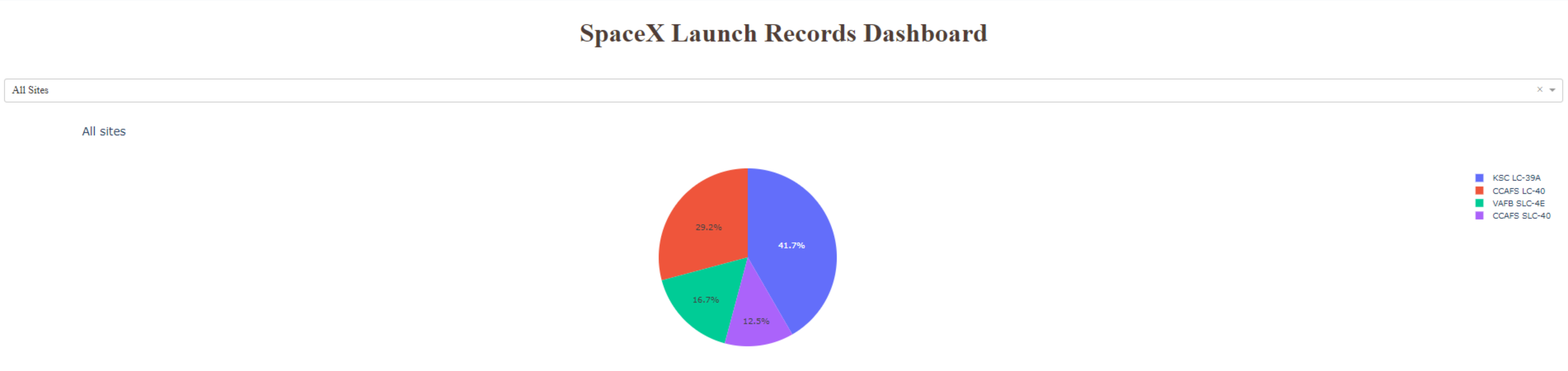
- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash

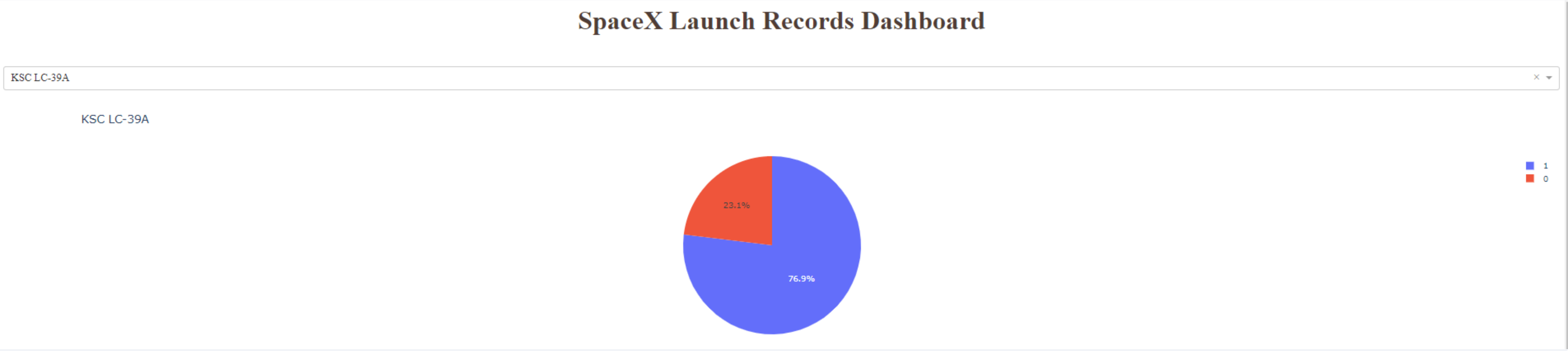
Success rate visualization

- Use pie chart to visualize the portion of success of each site has in a single pie chart



KSC LC 39A : the launch site with highest launch success ratio

- Launch site KSC LC 39A has highest launch success ratio among the group



Scatter plot of Payload vs. Launch Outcome for all sites

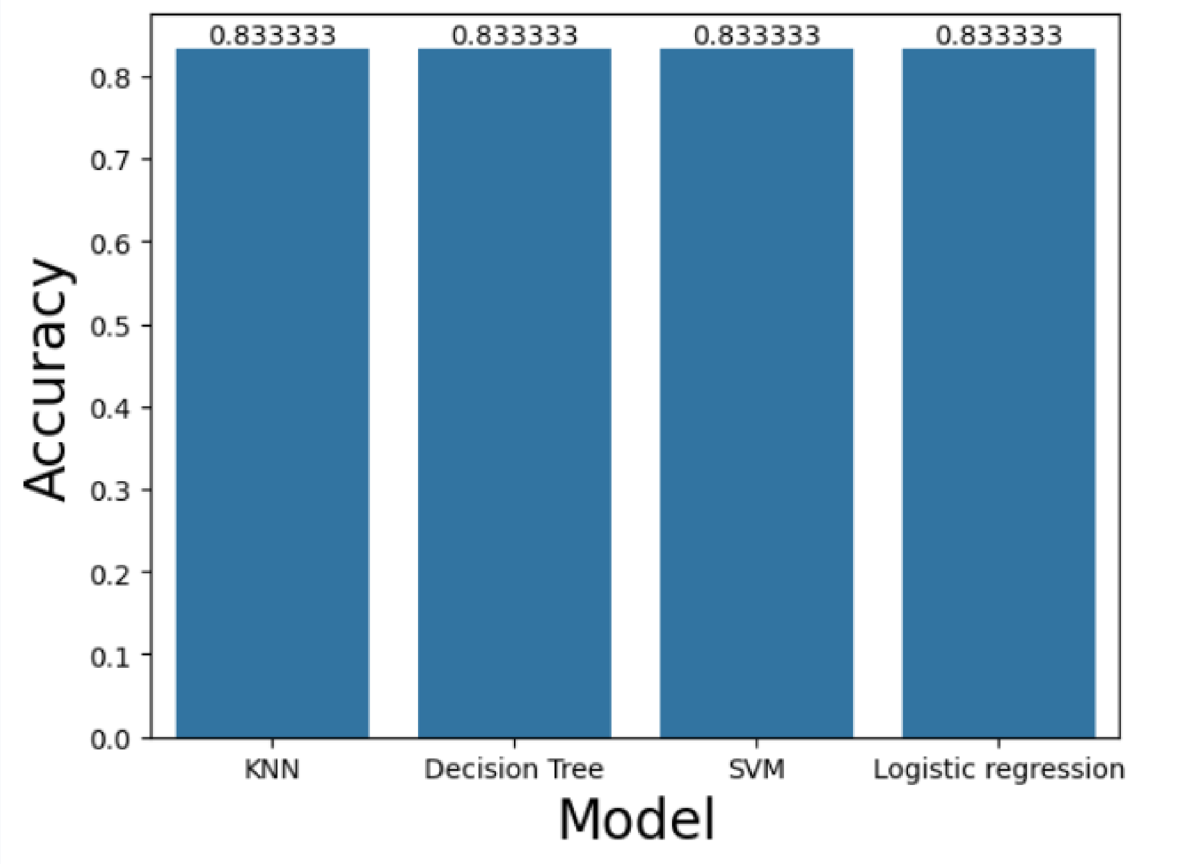
- Scatter plot shows the successful/failed landing outcome varied with payload and booster versions

Section 5

Predictive Analysis (Classification)

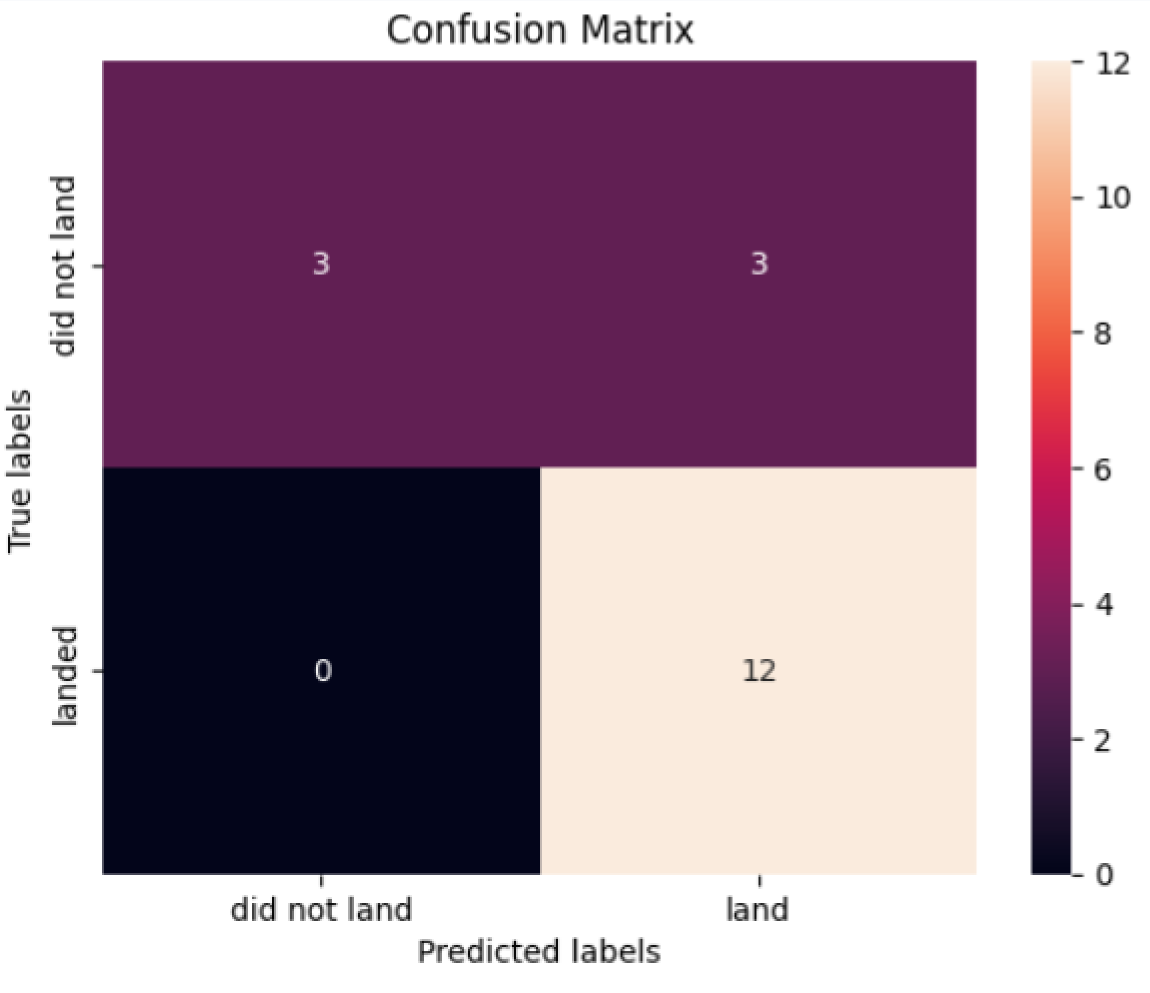
Classification Accuracy

- All models have the same classification accuracy when performed prediction on the test dataset



Confusion Matrix

- All models have the same performance



Conclusions

- To predict the outcome of landing of the spaceship various steps are performed as following list :
  - Data is collected with over public sources with API
  - Data preparation is performed for consistency of the analysis
  - Exploratory data analysis is performed by computing statistical measures and through various visualizations to get the brief introduction to the datasets.
  - Different types of prediction model give different accuracy in training dataset but have the same accuracy on testing set

Appendix

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False

A part of dataset

Thank you!