IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

VYTAS K.
2024-10-13

Outline

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

Executive Summary

METHODOLOGIES

- Data extraction: Gather relevant data via SpaceX publicly available API. Gather additional data about rocket launches using web scraping method via Wiki.
- Data transformation: Deal with missing values. Transform categorical values to numerical.
- Visualization: Use graphing and dashboards to conduct exploratory analysis.
- Modeling: Try various modeling techniques to get best accuracy for predicting launch success.
- Evaluation: Evaluate built models and present results.

RESULTS

- Mean launch success of Falcon9 rockets have increased drastically since 2013 till 2020 (mean success rate in 2020 reached ~84%): Results(1).
- Despite differences in payload mass and launch site, launches success rate increases as SpaceX conducts more flights: Results(2).
- Decision Tree and Support Vector Machine (SVM) models performed best in predicting launch success: Results(3).

Introduction

- Space X 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 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 space X for a rocket launch.
- Therefore, this study achieves the task of predicting whether SpaceX Falcon9 launches first stage will land successfully or not.

Section 1

Methodology

Methodology

Executive Summary

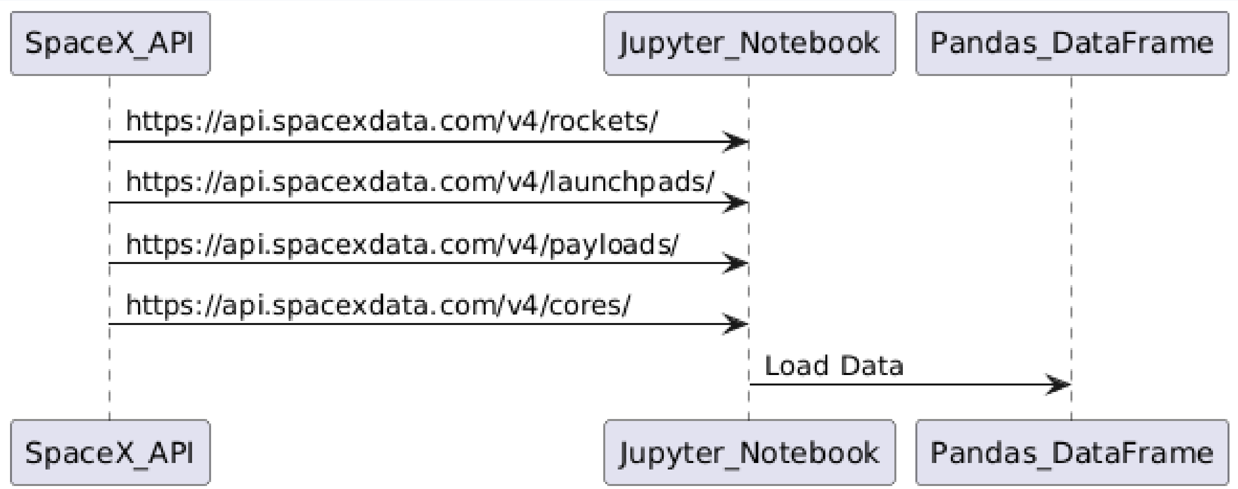
- Data collection methodology:
    - Gather relevant data via SpaceX publicly available API. Gather additional data about rocket launches using web scraping method via Wiki.
- Perform data wrangling
    - Deal with missing values. Transform categorical values to numerical.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Select model algorithms, find best tuning parameters using GridSearchCV, train and test model on best tuned parameters, summarize best performers by accuracy score.

Data Collection

- Part of data were collected using publicly available SpaceX API (https://github.com/r-spacex/SpaceX-API/tree/master/docs)
- Four API endpoints were used: /payloads; /cores; /rockets; /launchpads.
- Additional data were gathered via web scraping webpage (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose - https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb




```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = df_launch[df_launch['BoosterVersion'] == 'Falcon 9']
```

[37] Python

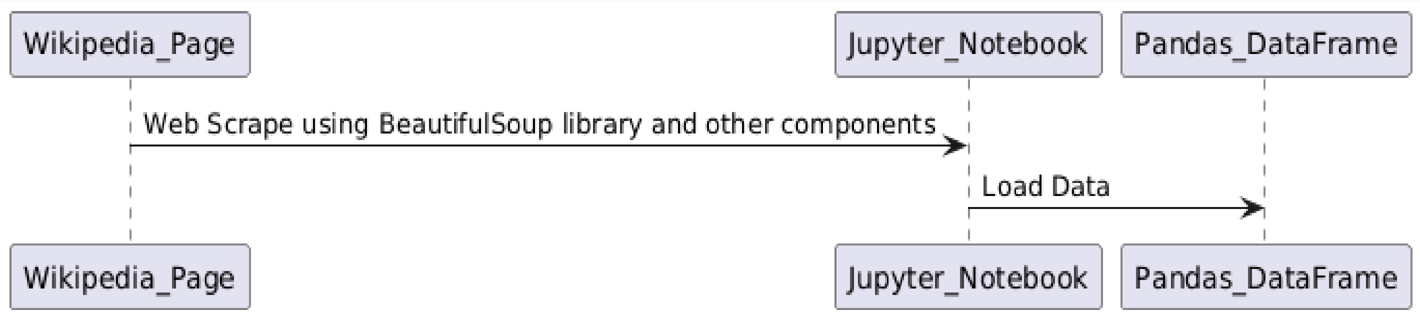
```
data_falcon9.head()
```

[38] Python

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Seri
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B00
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B00
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B00
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B10
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B10

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose - https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/jupyter-labs-webscraping.ipynb



After you have fill in the parsed launch record values into `launch_dict`, you can create a dataframe from it.

```python
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

Python

```python
df.head()
```

Python

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
2	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0003.18	No attempt\n	4 June 2010	18:45
3	2	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0004.18	No attempt	8 December 2010	15:43
4	3	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44

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

Data Wrangling

- Dealing with missing values:
  - # Calculate the mean value of PayloadMass column
  - mean_payloadmass = data_falcon9['PayloadMass'].mean()
  - # Replace the np.nan values with its mean value
  - data_falcon9['PayloadMass'].replace(np.nan, mean_payloadmass, inplace=True)
- Filter Falcon9 launches:
  - data_falcon9 = df_launch[df_launch['BoosterVersion'] == 'Falcon 9']
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb

EDA with Data Visualization

- Plotted:
  - Rocket payload mass versus Flight number – as SpaceX conducted more flights payload mass and success increased.
  - Launch site versus Flight number – as SpaceX conducted more flights success rate increased by launch site.
  - Launch site versus Payload mass – VAFB SLC 4E did not launched rockets having larger payload mass (<10000 kg)
  - Success Rate versus Orbit type – ES-L1, GEO, HEO, SSO Orbit types had the best success rate.
  - Flight number versus Orbit type – as SpaceX conducted more flights they chose to launch to further orbits.
  - Others – presented in Insights section
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/edadataviz.ipynb

EDA with SQL

- I used these SQL queries:
  - SELECT command with filtering (WHERE, GROUP BY, LIMIT, ORDER BY)
  - Various methods like SUM(), COUNT(), MIN(), MAX(), DISTINCT.
  - Nested SELECT queries.
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added circles on the map where launch sites are located.
- Added markers in each launch site for successful and unsuccessful launches.
- Added line to measure distance from launch site to closest railway and city.
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart by launch site and success rate.
- Success rate by payload mass filtered by range of payload mass in kg and launch site.
- These plots were added to dashboard to see if there are some interesting relations between launch site, payload mass of rocket and success rate of landing.
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- Created object for various models.
- Searched for best tuned hyperparameters using GridSearchCV.
- Calculated accuracy score of each model.
- Summarized the data.
- https://github.com/VytasKer/IBM-Data-Science-Certification/blob/main/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

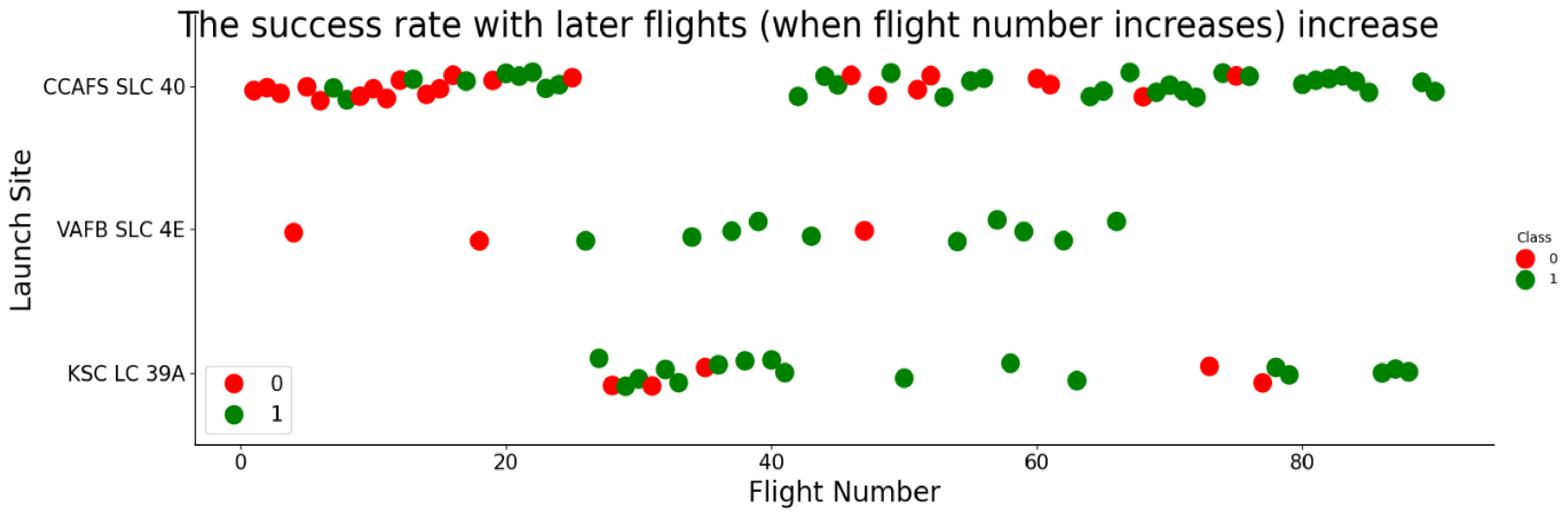
Results

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

Section 2

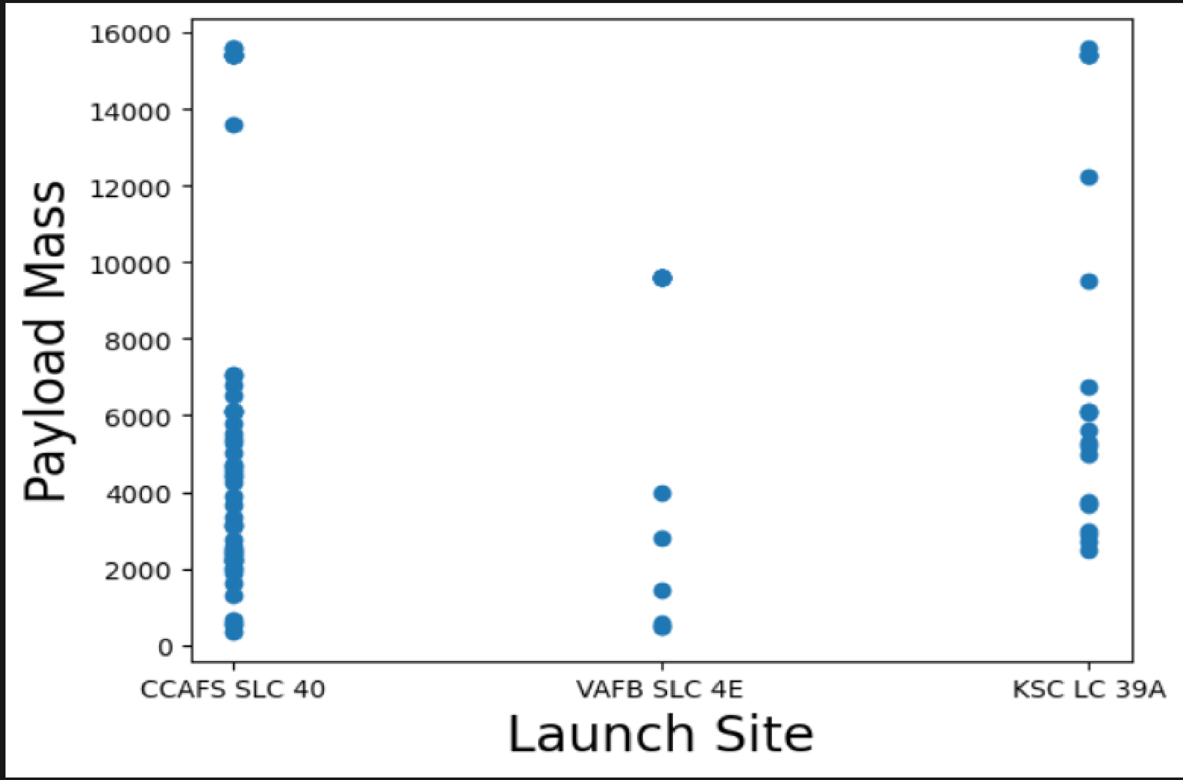
Insights drawn from EDA

Flight Number vs. Launch Site



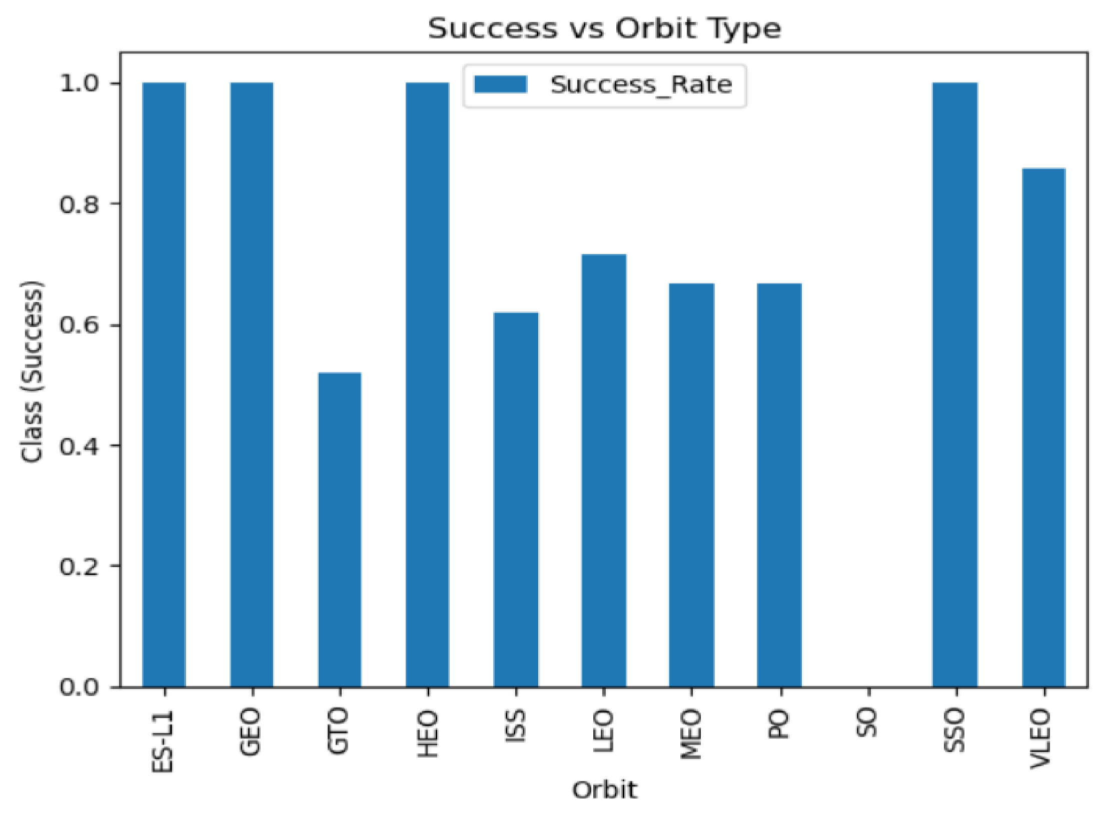
Success rate increases with flight number: The graph shows a clear trend that the success rate of SpaceX launches increases as the flight number increases. This suggests that SpaceX has made significant improvements in its launch technology over time. SpaceX has made significant improvements in its launch technology over time, leading to a higher success rate with later flights.

Payload vs. Launch Site



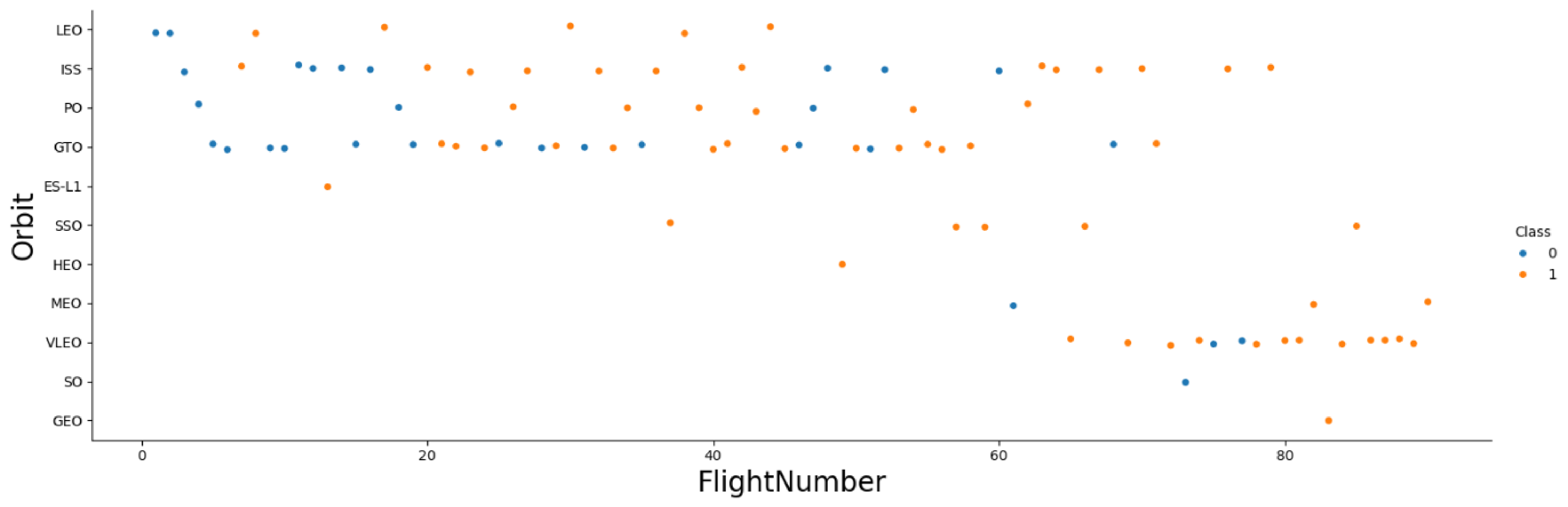
NowCCAFS SLC 40 is the most versatile launch site, capable of handling a wide range of payload masses. VAFB SLC 4E is suitable for moderate payloads, while KSC LC 39A is primarily used for lighter payloads.

Success Rate vs. Orbit Type



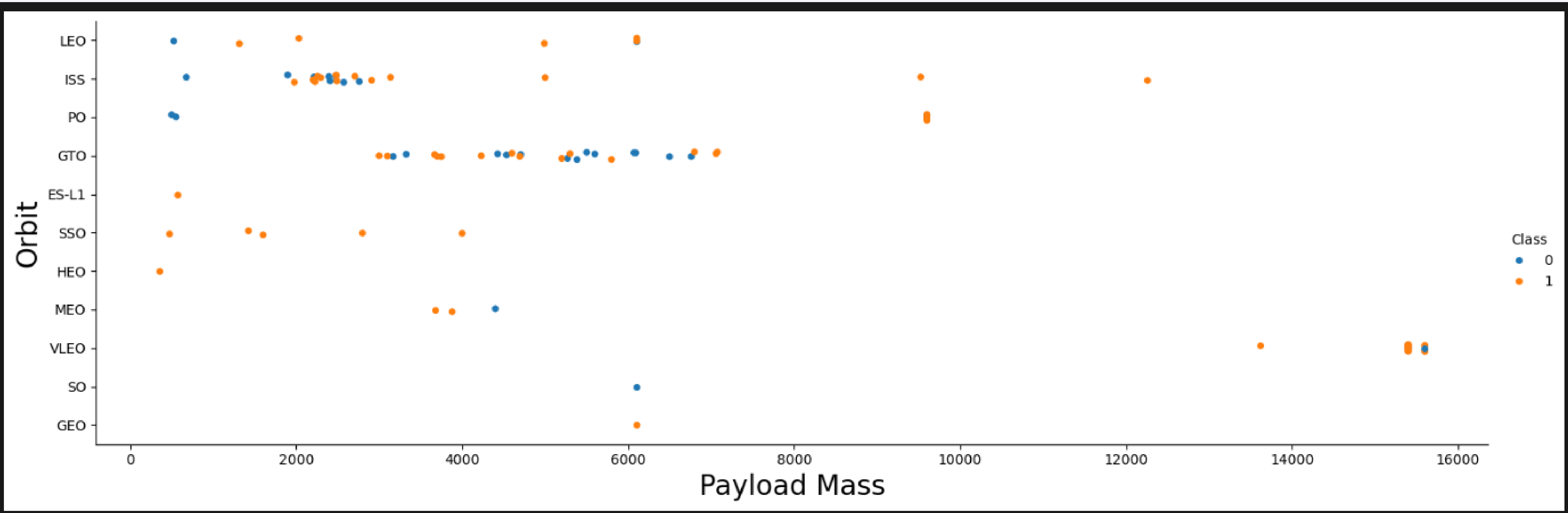
SpaceX has achieved high success rates for launches to ES-L1, GEO, HEO, SSO, and VLEO orbits. However, launches to GTO, ISS, LEO, MEO, and PO orbits have lower success rates. Further analysis would be needed to understand the reasons for these differences.

Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

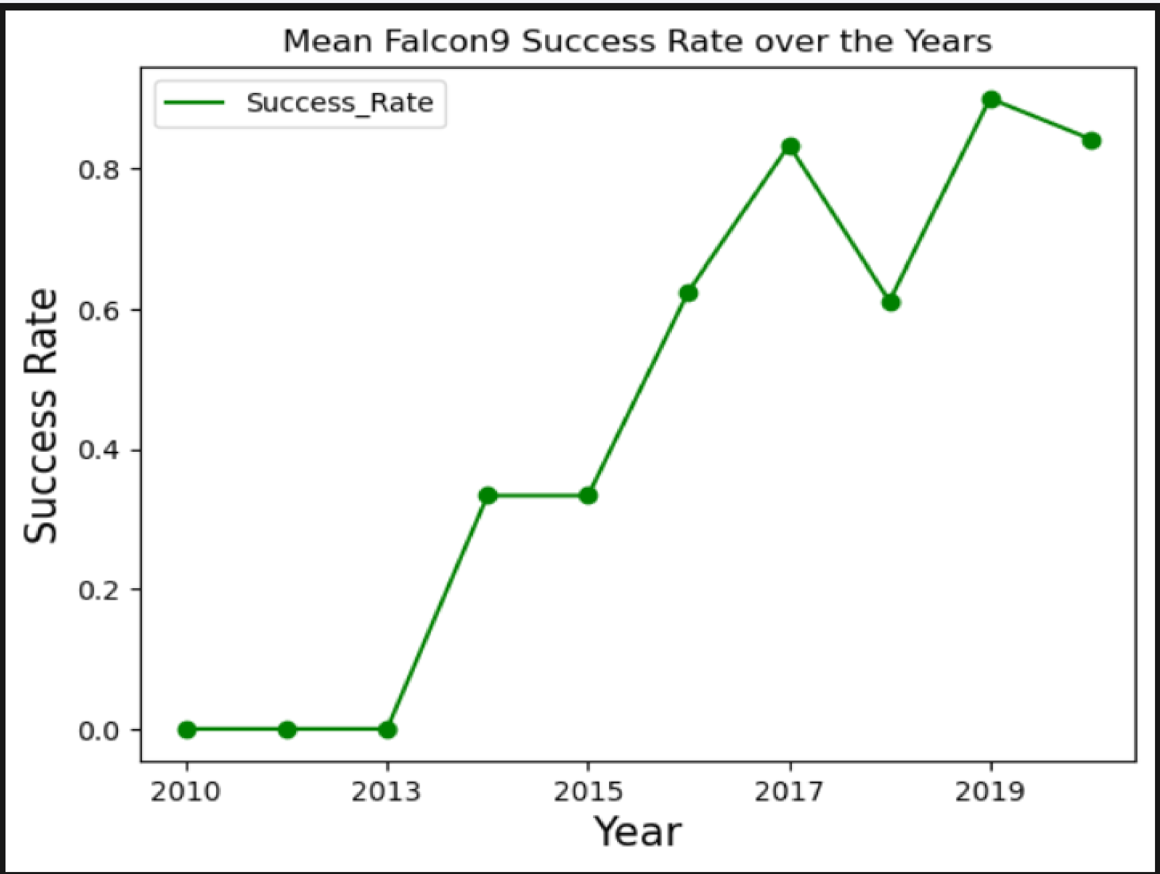
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



You can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
[13]
%%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
```

[14] 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

Total Payload Mass

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

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

22]

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

Total_Mass
45596

Average Payload Mass by F9 v1.1

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

[23]

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

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

Average_Mass
2928.4

First Successful Ground Landing Date

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

Hint:Use min function

[26]

```
%%sql
SELECT MIN("Date") AS First_Successful_Launch FROM SPACEXTABLE
WHERE "Mission_Outcome" = 'Success'
```

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

First_Successful_Launch
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[14]

```
 * 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

```
%%sql
SELECT DISTINCT("Mission_Outcome") AS Mission_Outcome, COUNT("Mission_Outcome") AS Number_of_Launches
FROM SPACEXTABLE
GROUP BY "Mission_Outcome"
```

[18]

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

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

Boosters Carried Maximum Payload

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

[19]
```
%%sql
SELECT "Booster_Version" FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" IN (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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
SELECT substr("Date", 6,2) AS Month_Number, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE
WHERE substr("Date",0,5)='2015'
AND "Landing_Outcome" = 'Failure (drone ship)'
```

2] Python

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

Month_Number	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") AS Number_of_Launches FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
--AND "Landing_Outcome" = 'Success (ground pad)'
GROUP BY "Landing_Outcome"
ORDER BY Number_of_Launches DESC
```

[31] Python

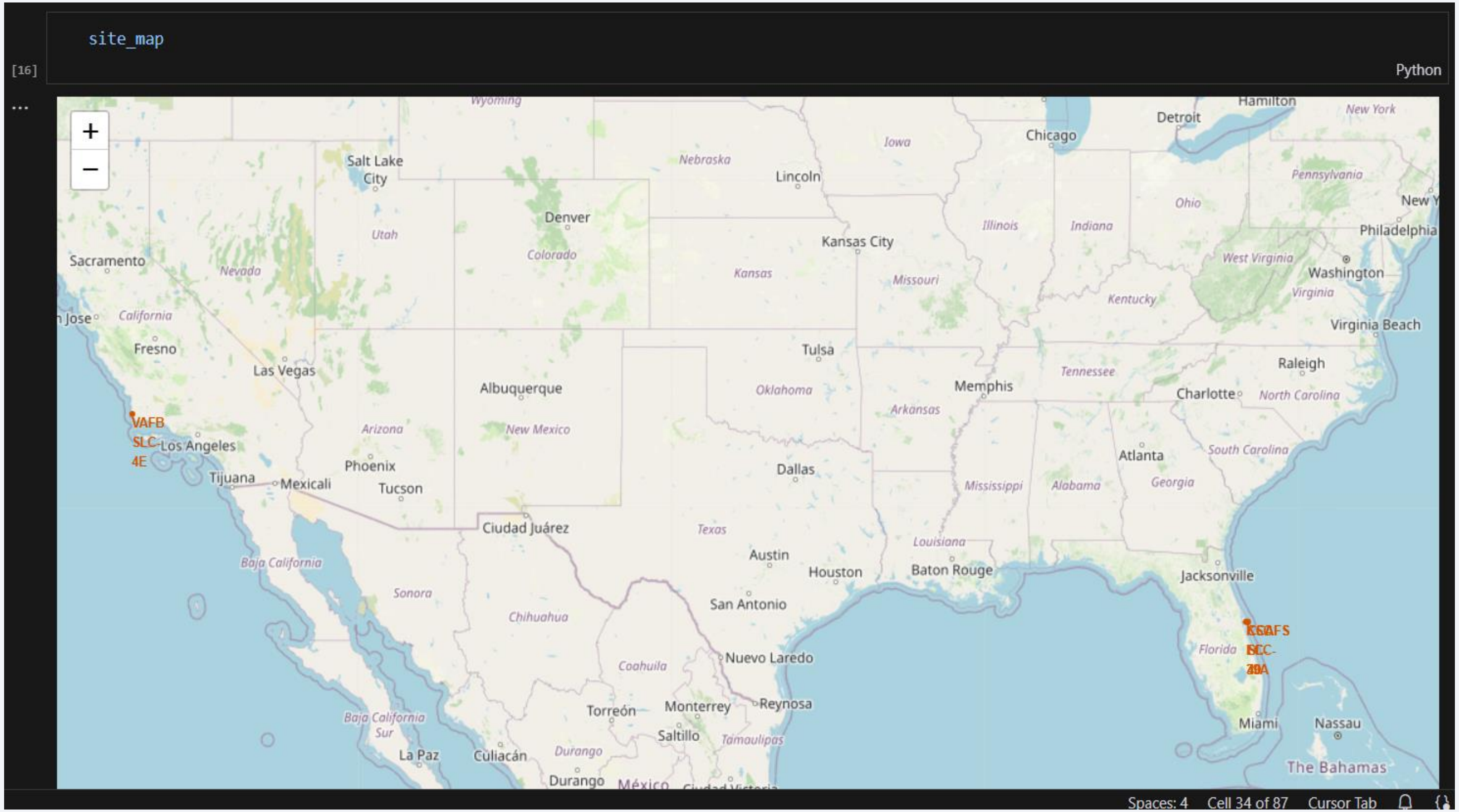
* sqlite:///my_data1.db
Done.

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

Section 3

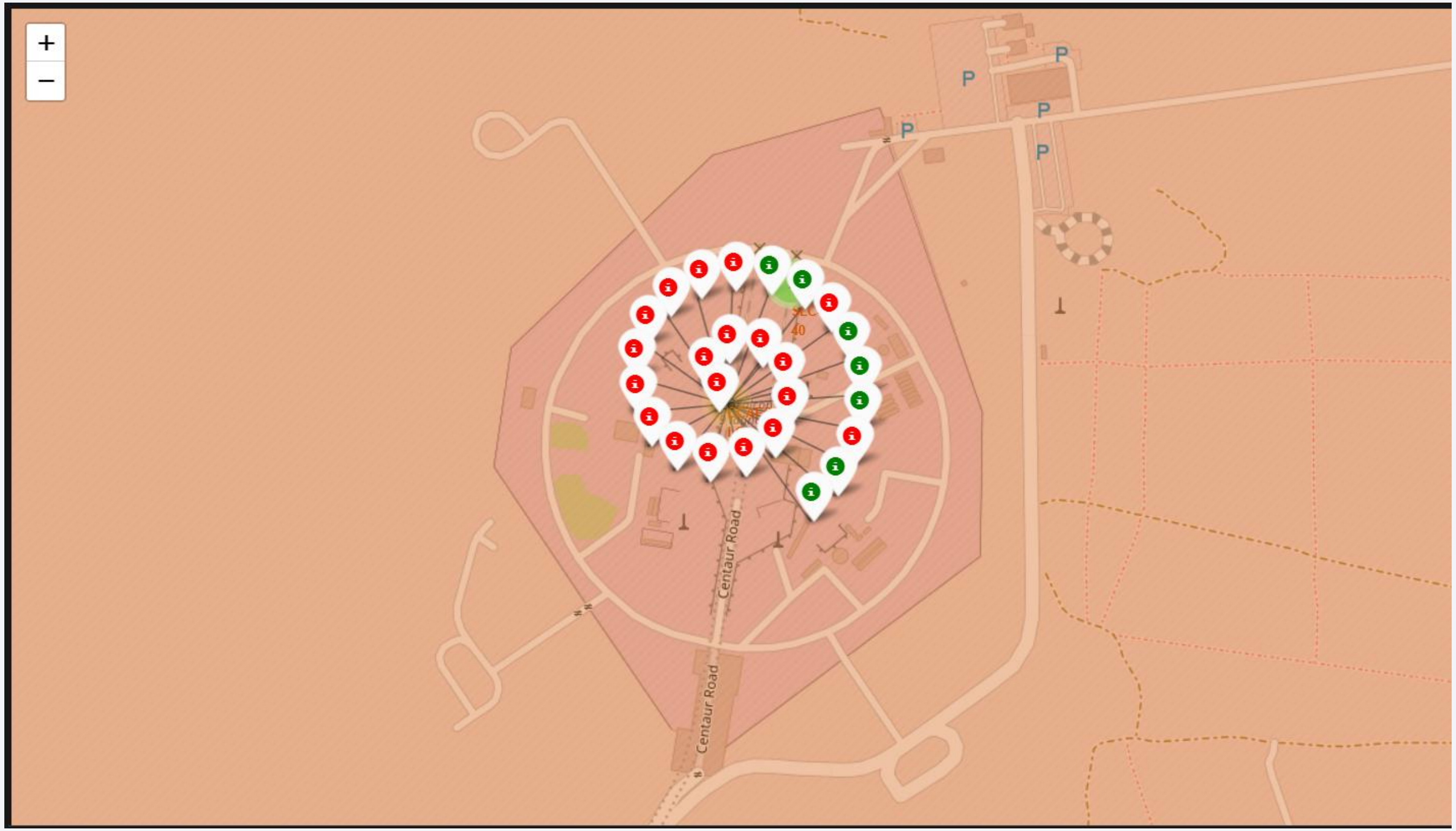
Launch Sites Proximities Analysis

SpaceX rocket launch sites map



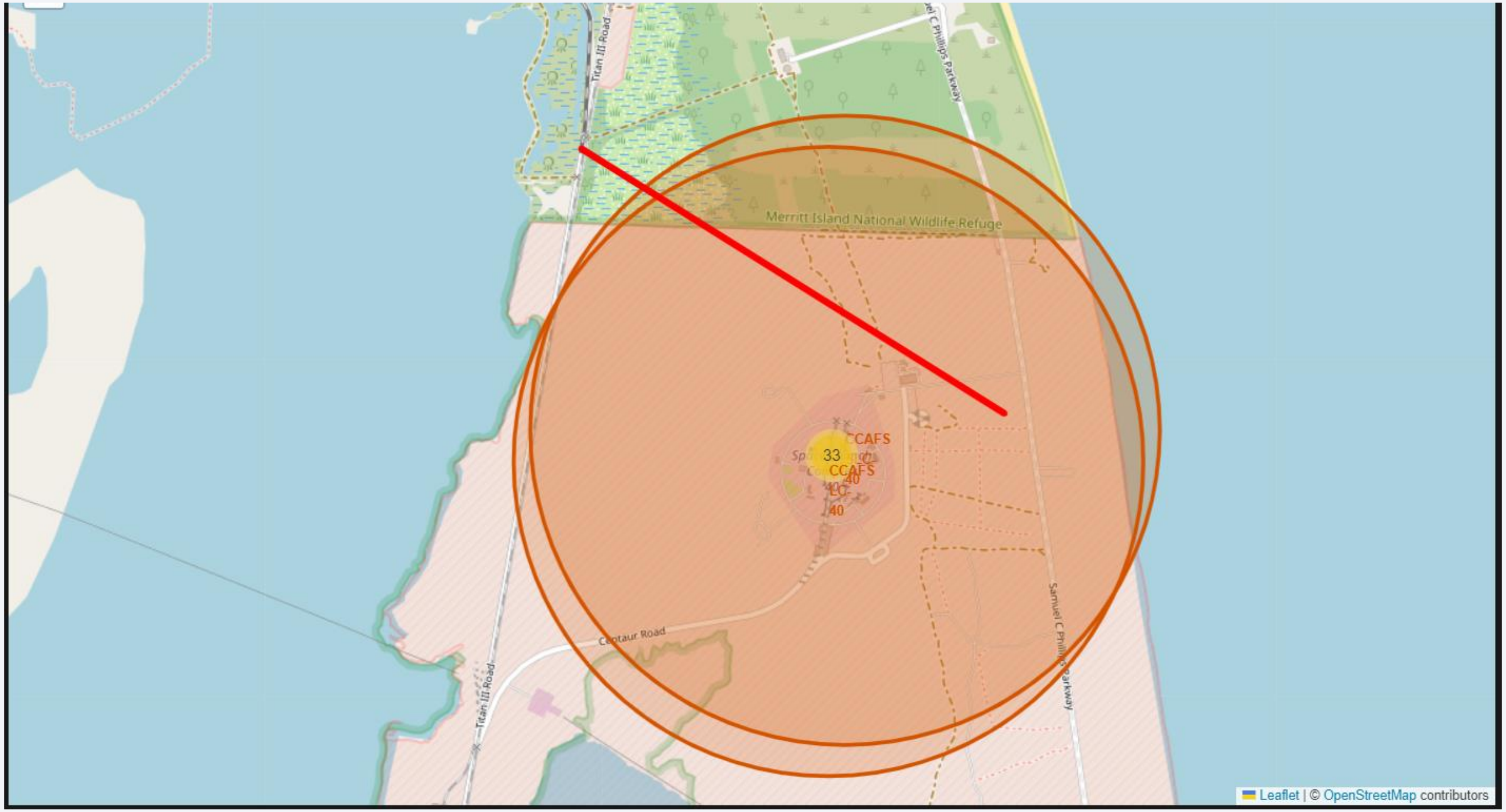
Launch sites are in the eastern and western parts of USA (Florida and California).

Marked successful and unsuccessful launches in a site



One can see spiraled markers of successful and unsuccessful launhes in a particular site (Florida – Kennedy Space Center).

Line to the closest railway



Marked line extends roughly 1.5km from the site coordinate to the closest railway.

Section 4

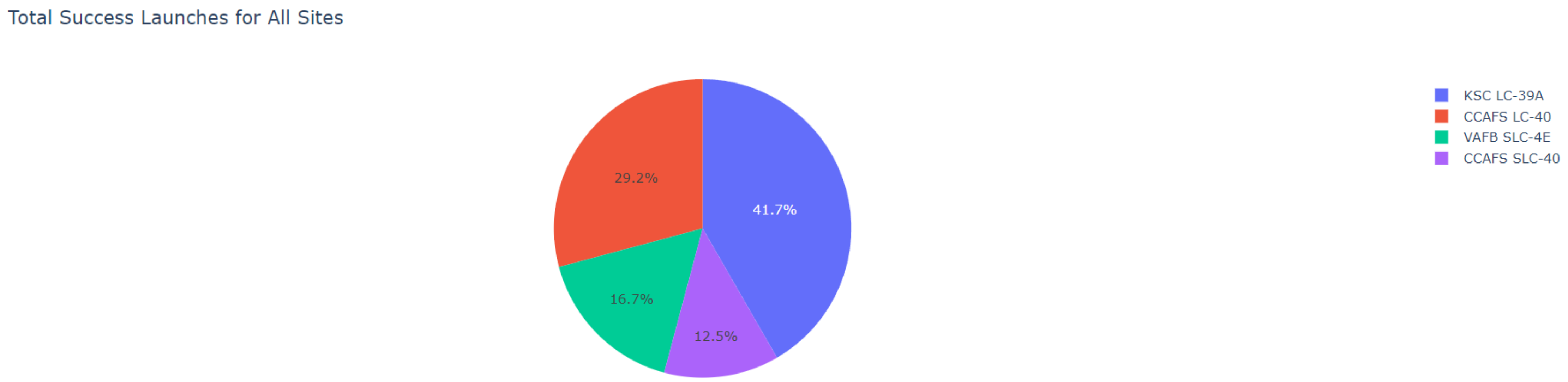
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

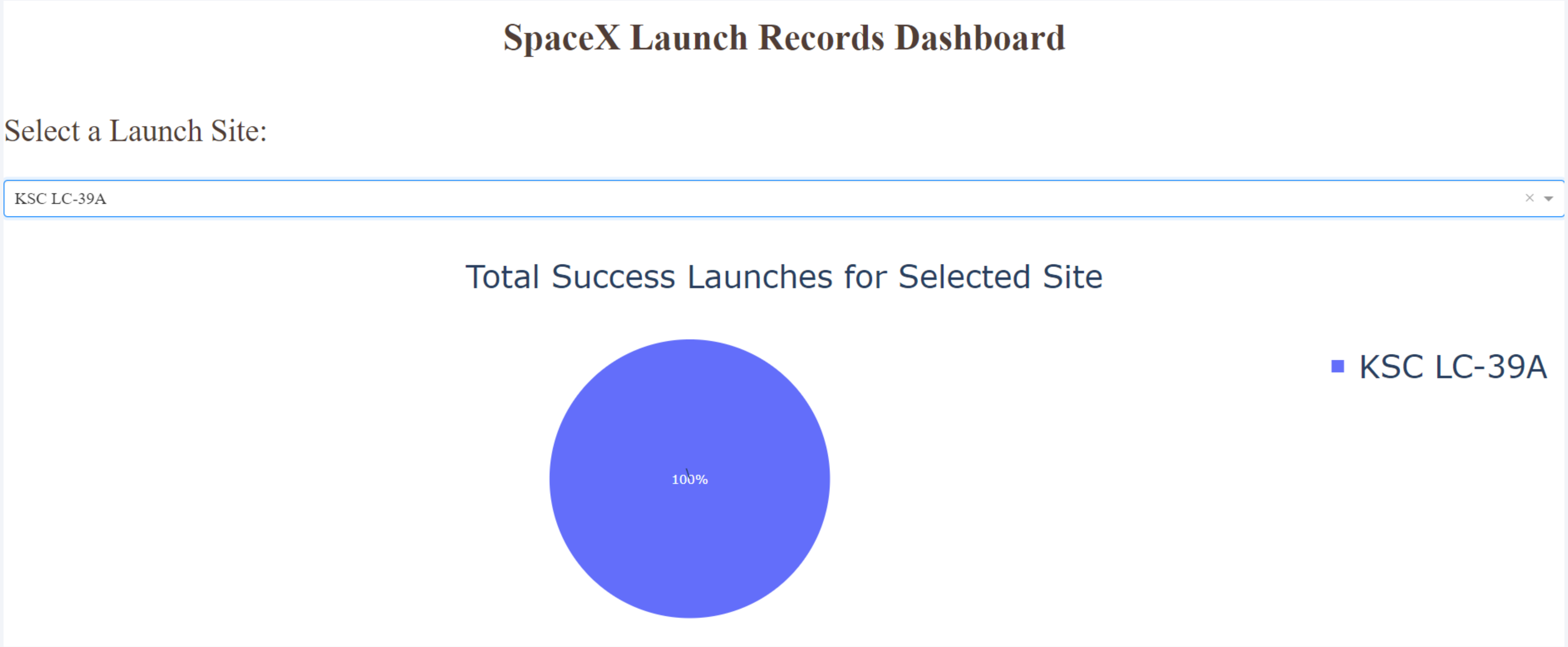
SpaceX Launch Records Dashboard

Select a Launch Site:

All Sites



Success rate for particular site



KSC LC-39A seems to have the highest success rate

Success rate vs Payload mass

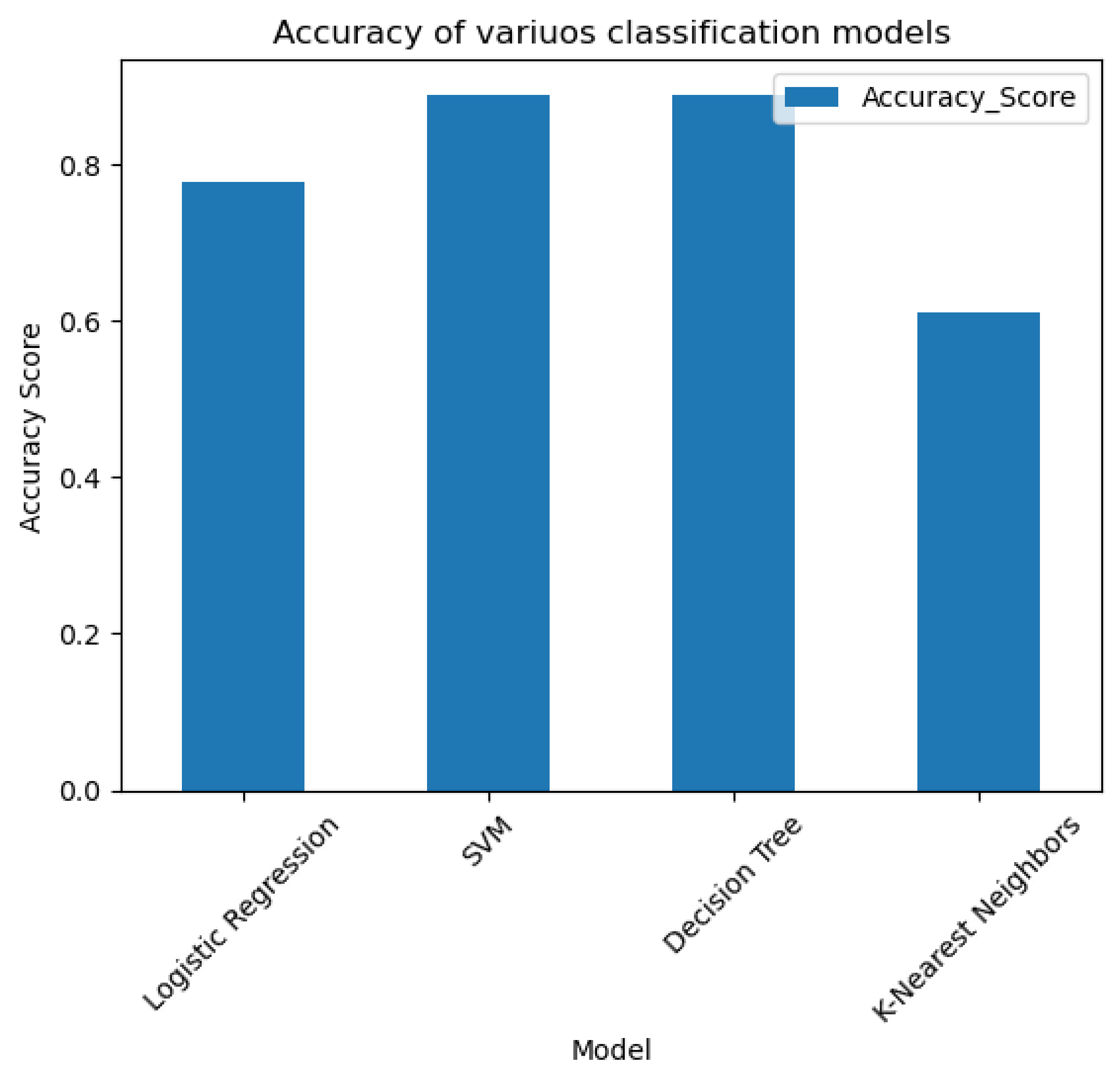
It seems that for range 2000 to 6000 kg success rate of landing is the best.

Section 5

Predictive Analysis (Classification)

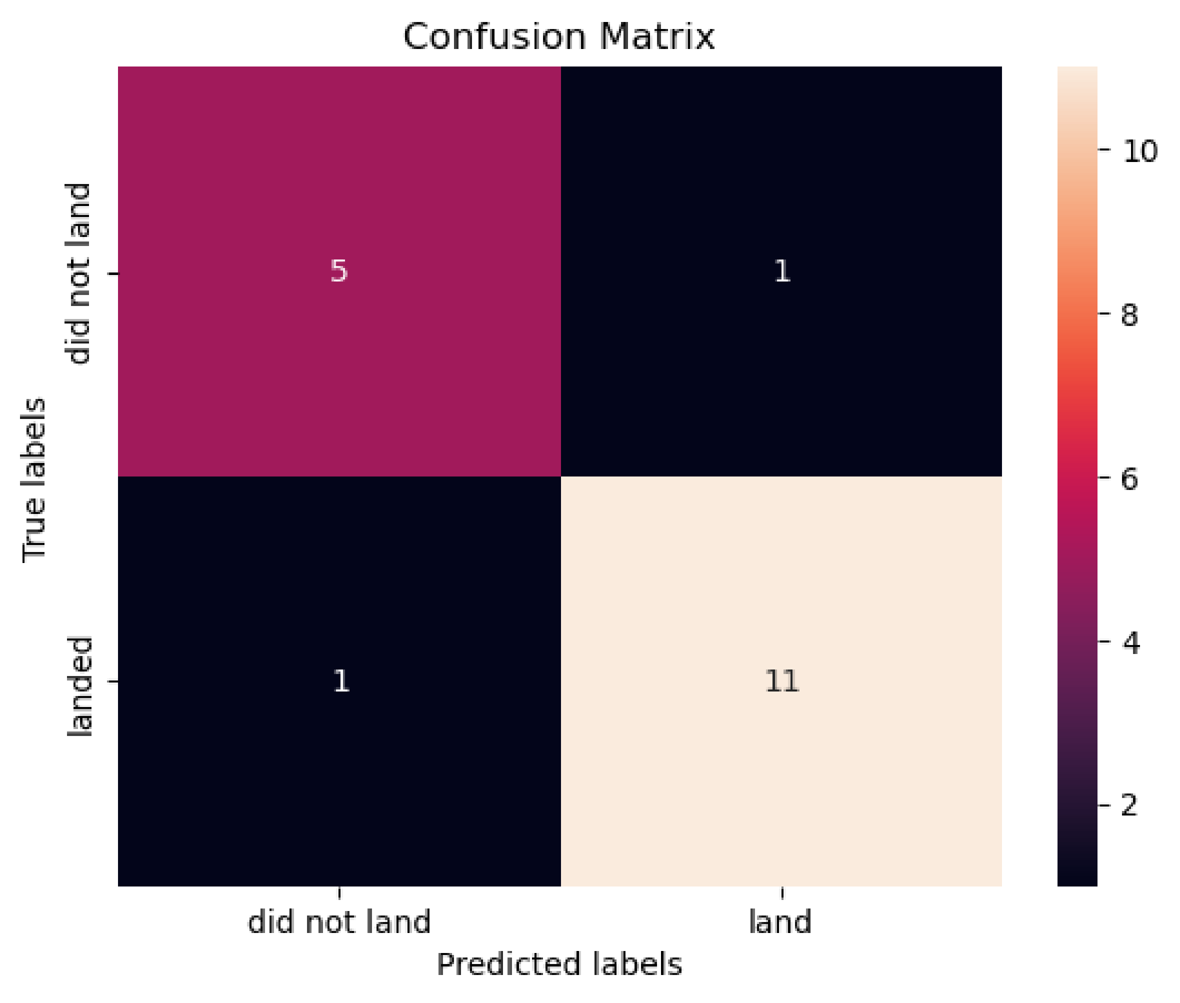
Classification Accuracy

- SVM and Decision Tree have the highest accuracy score.



Confusion Matrix

- SVM confusion matrix – only one false positive and one false negative case in each group.



Conclusions

- Mean success rate of Falcon9 rockets launches have increased drastically since 2013 till 2020 (mean success rate in 2020 reached ~84%).
- Despite differences in payload mass and launch site, launches success rate increases as SpaceX conducts more flights.
- Flights with payload mass between roughly 2000 and 6000kg have the highest chance of successful landing.
- Decision Tree and Support Vector Machine (SVM) models performed best in predicting launch success with accuracy of about ~0,89.

Appendix

- Github URL to all project material - https://github.com/VytasKer/IBM-Data-Science-Certification/tree/main/Capstone

Thank you!