Data Science Capstone project

Oliver Freimuth

08/28/2021

Outline

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

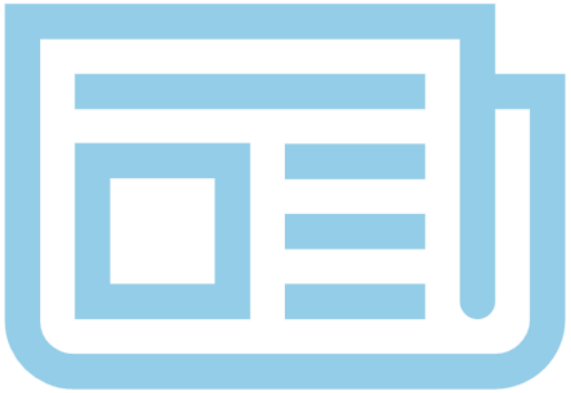
Executive Summary

- Objective: Figure out which useful insights Space X data can provide for Space Y
- Data: Obtained via API and Web Scraping
- Results:
  - Space X success increased over time: What did they change?
  - Orbit type and launch site are relevant for the success rate: What makes them different to each other?
  - Heavy Payloads seem to be problematic

Introduction

- Space X is unique since it has stage one parts for rockets
- Space Y wants to compete with Space X. Thus, we will try to figure out if first stages of a certain launch can be reused and which factors seem to determine this.

Methodology



- Data collection methodology:
  - Web Scraping
  - REST API
- Perform data wrangling
  - Filter Data Frame
  - Handle NAs
  - Create Outcome Lables
- 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

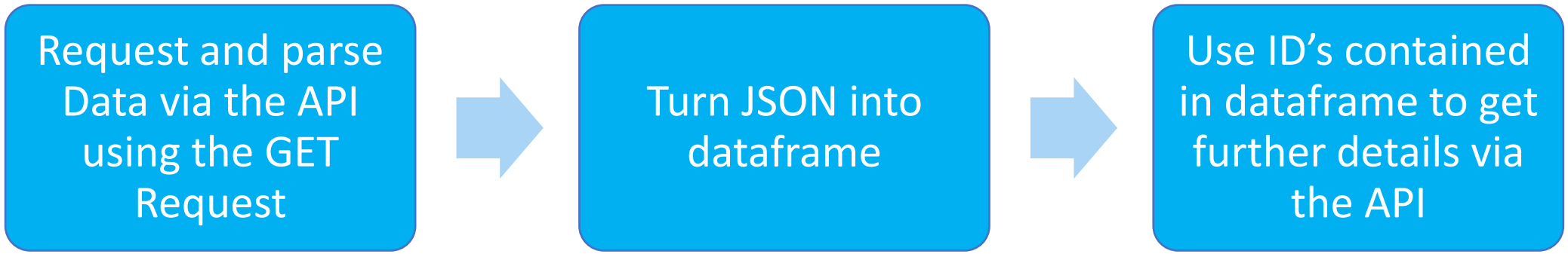
Methodology

Data collection

- Data was collected by API and Web Scraping

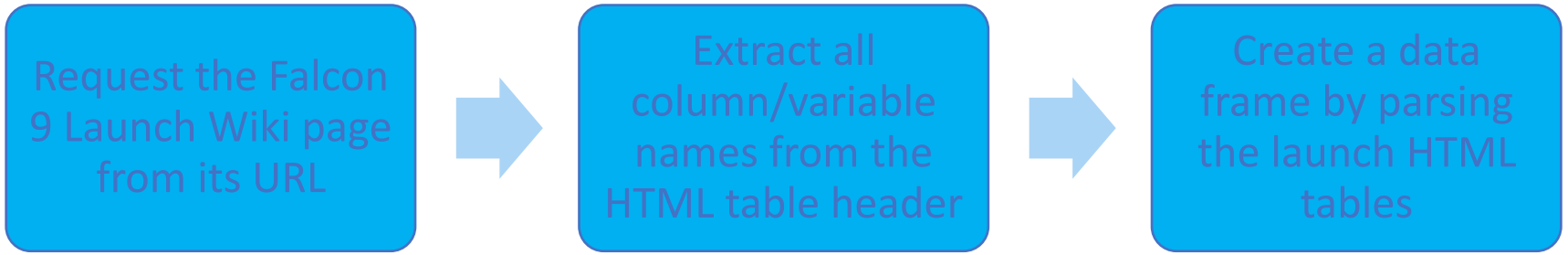
- See Flow Charts on the next slides

Data collection – SpaceX API



GitHub: https://github.com/OliFre94/PythonDSFinalProject/blob/master/01_DataCollectionAPI.ipynb

Data collection – Web scraping



GitHub URL:
https://github.com/OliFre94/PythonDSFinalProject/blob/master/02CompleteDataCollectionWithWebScraping.ipynb

Data wrangling

- Filter only Falcon 9 launches
- Dealing with missing values:

  Replace NAs in PayloadMass with it’s mean
- Create a landing Outcome label (0 = bad, 1 = good)

GitHub URL:

https://github.com/OliFre94/PythonDSFinalProject/blob/master/02CompleteDataCollectionWithWebScraping.ipynb

https://github.com/OliFre94/PythonDSFinalProject/blob/master/03DataWrangling.ipynb

EDA with data visualization

Chart Type	Chart Content	Key Insight
Scatterplot	Flight Number vs. Payload Mass	Success rate increases with flight number
Scatterplot	Launch Site vs. Flight Number	No recent launches from VAFB; Most launches from CCAFS;
Scatterplot	Launch Site vs. Payload Mass	No Pay Load over 10000 kg from VAFB; most unsuccessful launches with lower pay loads (possibly early launches had lighter weights)
Barplot	Success rate by Orbit	Rates vary between ~50% and 100% with the exception of SO (0%)
Satterplot	Flight Number vs. Orbit	Some orbits have success rates related to the flight number, others do not
Scatterplot	Orbit vs Payload Mass	Depending on the Orbit the influence of Payload Mass can be positive or negative
Linechart	Sucessrate by years	Upwards trend in success rate from 2013 to 2020

GitHub URL:

https://github.com/OliFre94/PythonDSFinalProject/blob/master/05EDAVisualization.ipynb

EDA with SQL

- Query for unique launch sites
- Query for 5 records where launch site begins with ‘KSC’
- Query for total payload mass (kg) launched by NASA (CRS)
- Query for average payload mass carried by F9 v1.1 boosters
- Query for dates where the landing on drone ship was successful
- Query for names of boosters landing successfully on ground pad with payload mass between 4000 kg and 6000 kg
- Query for total number of successful and failure mission outcomes
- Query for names of booster versions which carried the maximum payload mass
- Query for records with successful landing outcome in 2017
- Query for ranked count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20

GitHub URL: https://github.com/OliFre94/PythonDSFinalProject/blob/master/04EDAwithSQL.ipynb

Build an interactive map with Folium

- Circles were set to indicate:
  - Launch Sites (3x Florida, 1x California)
- Markers were set:
  - For each successful launch at the respective launch site (green)
  - For each unsuccessful launch at the respective launch site (red)
  - At the railway nearest to launch sites CCAFS LC-40 and SLC-40
- A PolyLine was used to show the distance between the railway marker and launch site CCAFS SLC-40

- By doing this, there is a visual overview of the locations of launch sites, their success rates and the distance to the next railway (exemplary for one launch site)

- GitHub: https://github.com/OliFre94/PythonDSFinalProject/blob/master/06InteractiveVisual%20Analytics(Folium).ipynb

Build a Dashboard with Plotly Dash

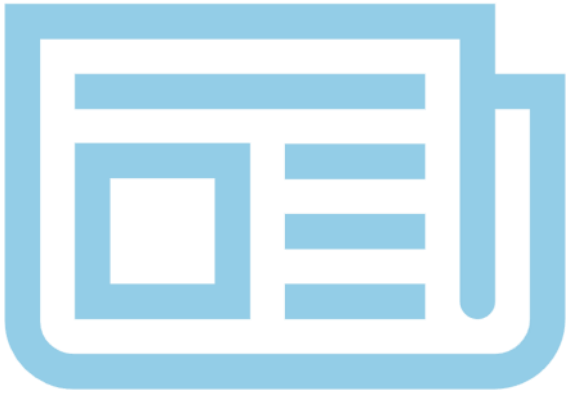
- User can select a launch site (or all) and a range of pay load
- A pie chart shows the share of successful launches per launch site or success rate per launch site
- A scatter plot shows the launches by success or fail and payload depending on the range of payload selected and launch siteselected
- Thus dynamic visual insights can be obtained

- GitHub: https://github.com/OliFre94/PythonDSFinalProject/blob/master/07dashboard_spacex.py

Predictive analysis (Classification)

GitHub: https://github.com/OliFre94/PythonDSFinalProject/blob/master/08PredictiveAnalysis.ipynb

Results

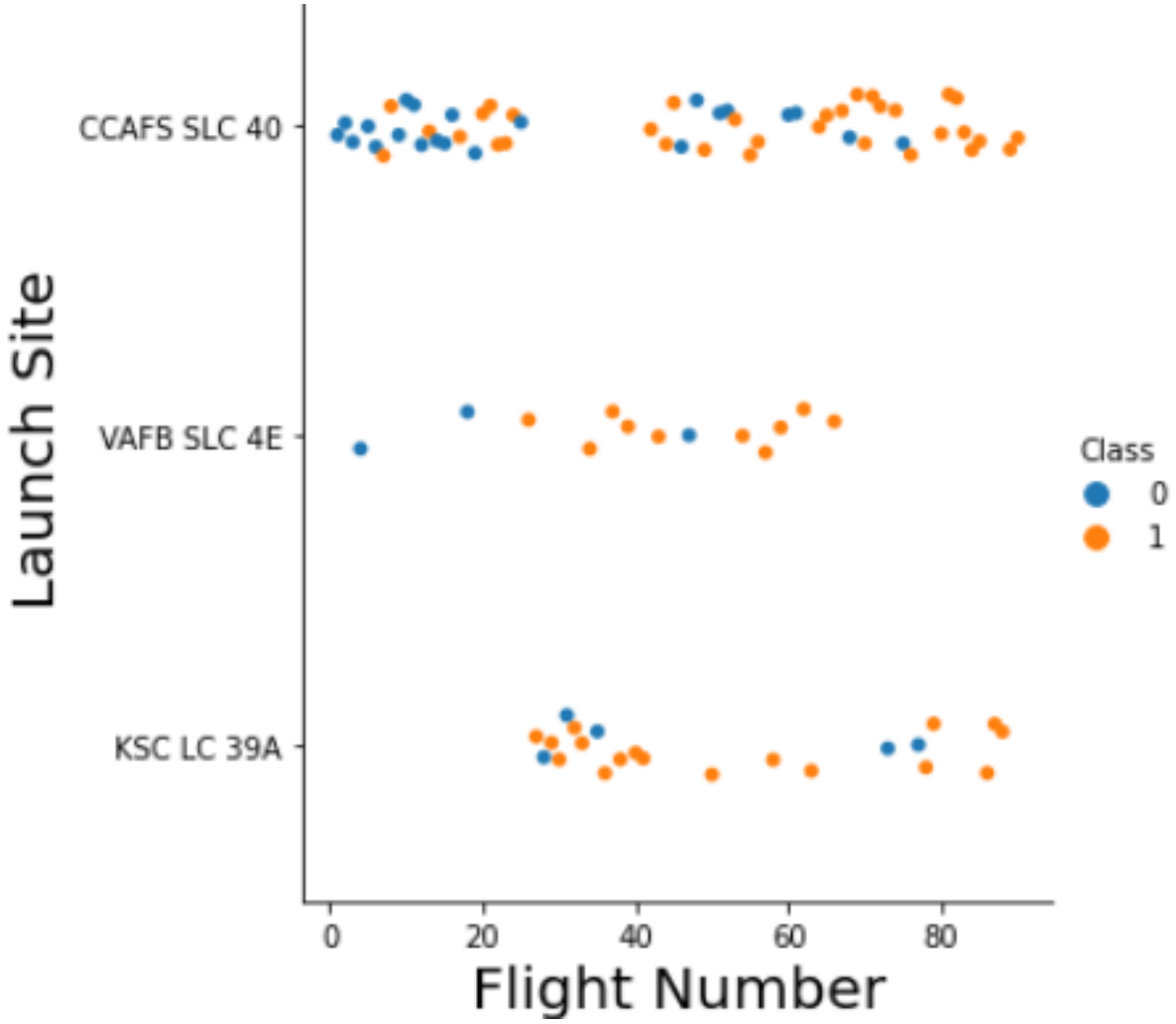


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

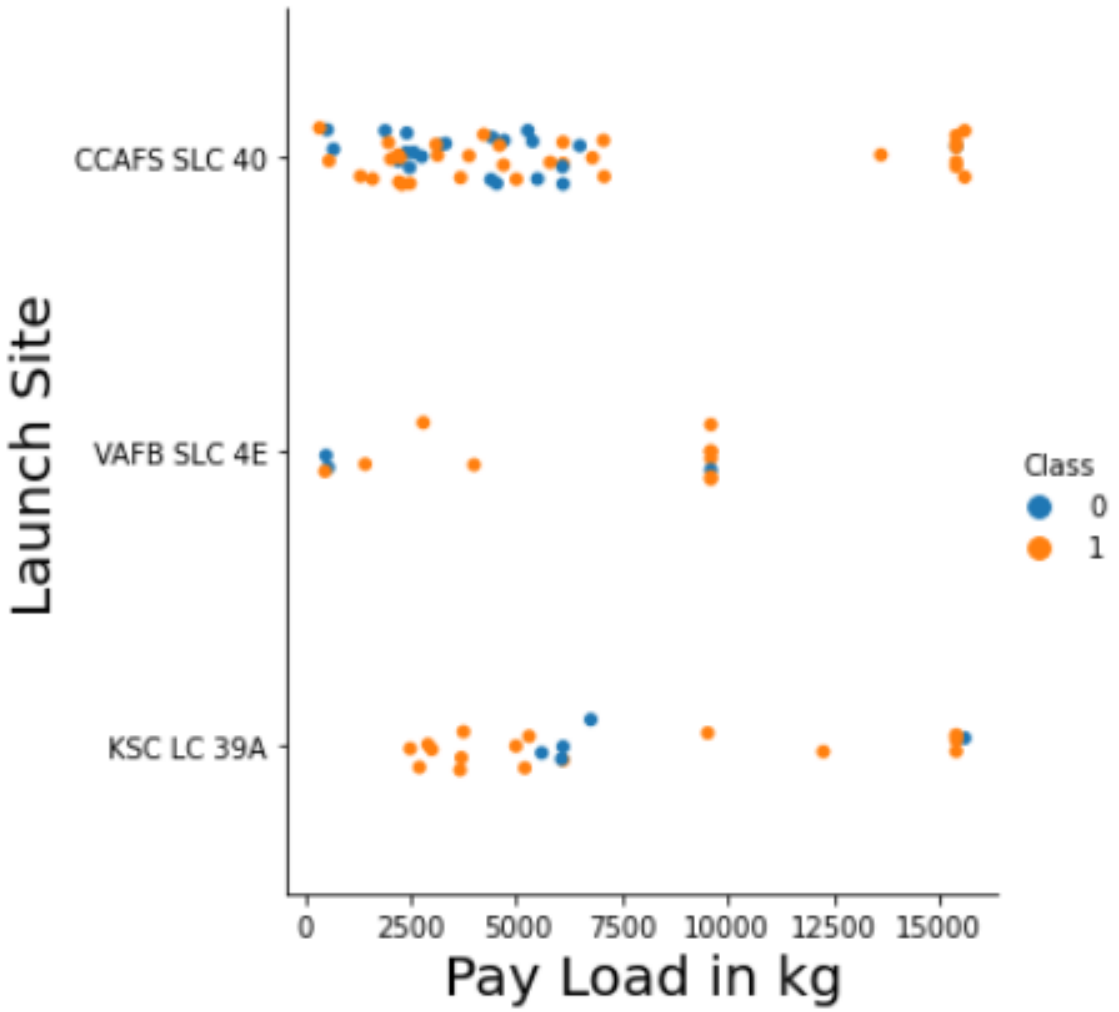
Flight Number vs. Launch Site

- More early flights had a lower success rate
- There are no recent launches at VAFB SLC 4E



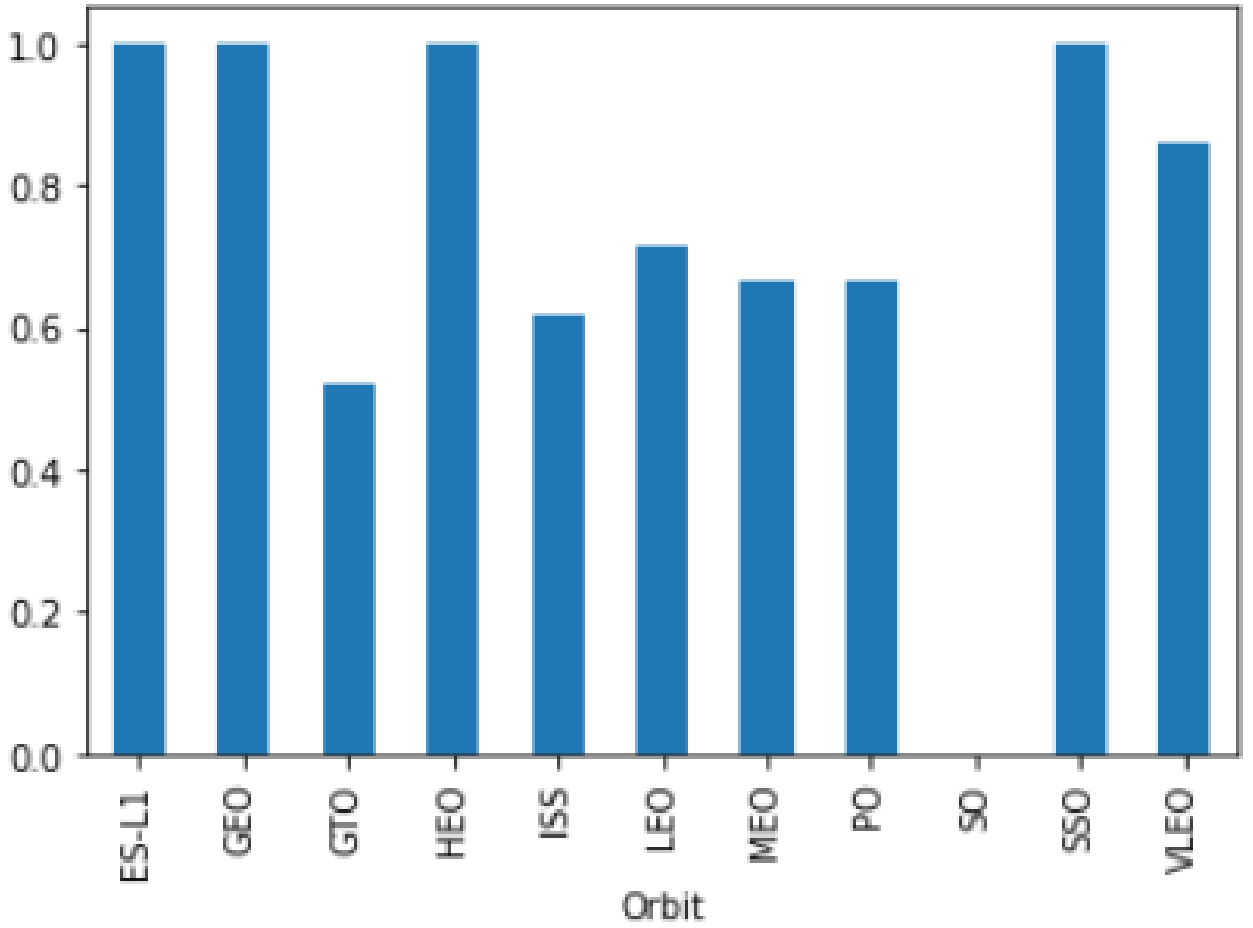
Payload vs. Launch Site

- Lower Pay Loads seem to correlate with lower success rates
- Possibly lower pay loads are also earlier flights
- No launches over 10000 kg from VAFB SLC 4E



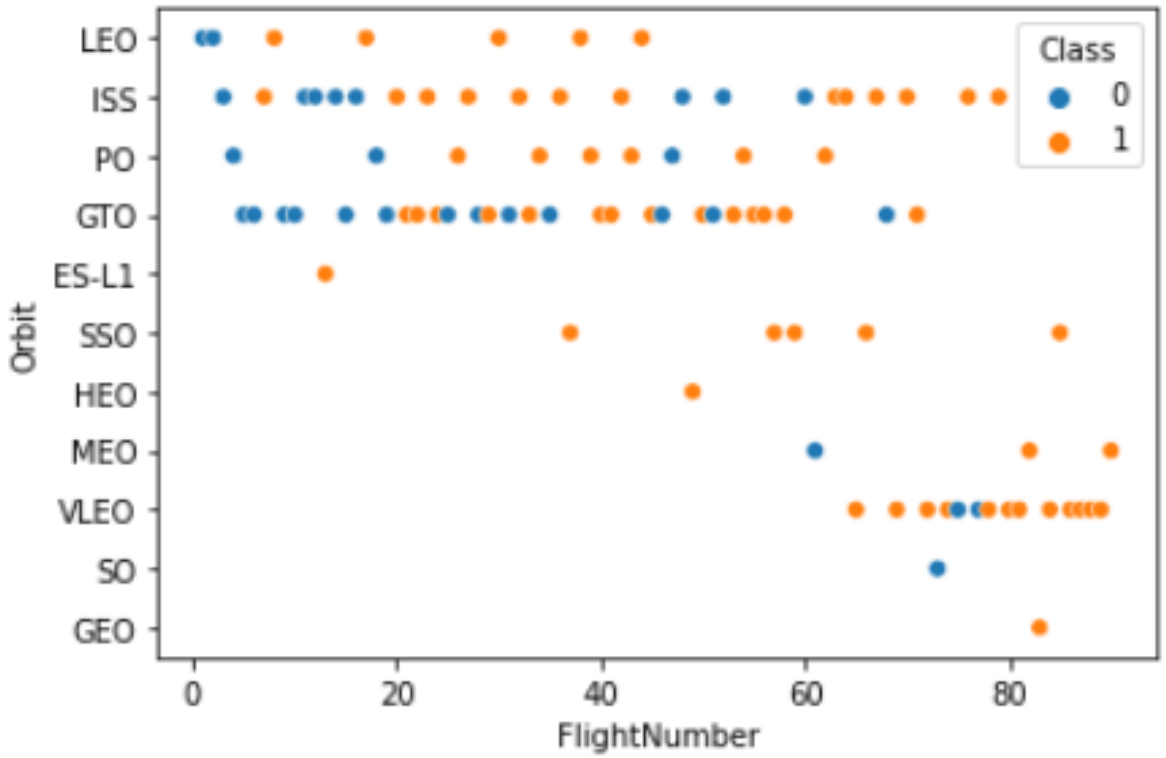
Success rate vs. Orbit type

- Most Orbit types have a success rate over 60%
- GTO is below 60%
- SO even at 0%



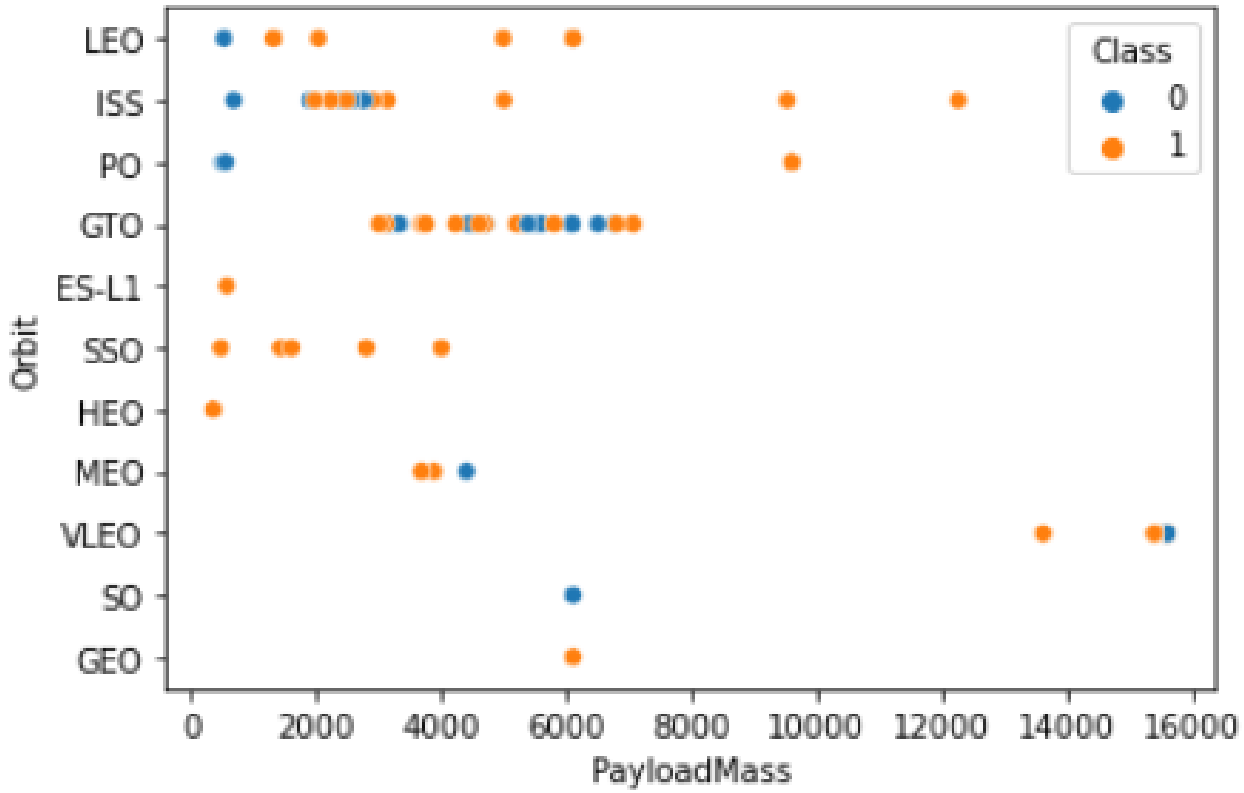
Flight Number vs. Orbit type

- Some orbits perform better at later launches (ISS
- Others still have problems (GTO)
- Some seemingly have no correlation (VLEO)



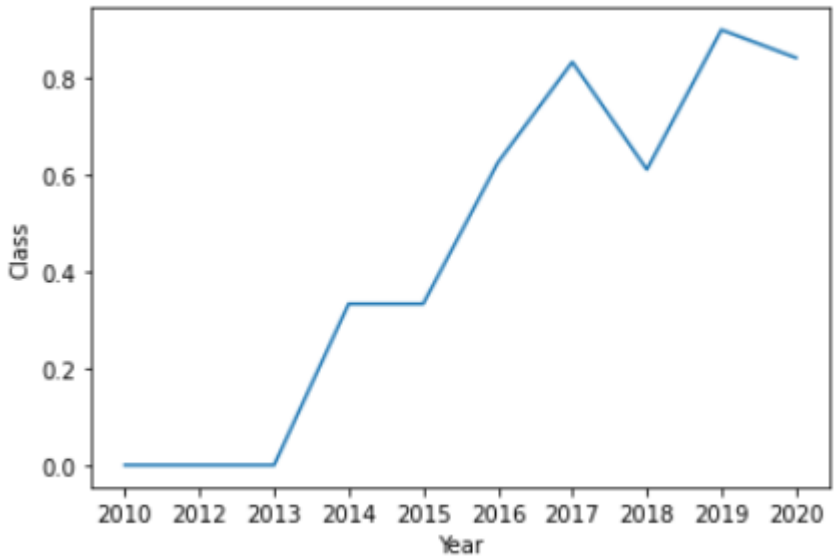
Payload vs. Orbit type

- GTO seems to perform worse with increasing PayloadMass
- ISS seems to perform better with increasing PayloadMass



Launch success yearly trend

- The success rate shows an increasing trend
- There is a considerable dip in 2018



EDA with SQL

All launch site names

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

Launch site names begin with `CCA`

Task 2

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

In [5]:
```
%%sql SELECT *
FROM TSN62941.SPACEXTBL
WHERE launch_site LIKE 'KSC%'
LIMIT 5
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total payload mass

Task 3

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

In [6]:
```
%%sql SELECT SUM(payload_mass__kg_)
FROM TSN62941.SPACEXTBL
WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

1
45596

Average payload mass by F9 v1.1

Task 4

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

In [7]:
```
%%sql SELECT AVG(payload_mass__kg_)
from TSN62941.SPACEXTBL
WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

1
2534

First successful ground landing date

Task 5

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

In [8]:
```
%%sql SELECT Min(Date)
FROM TSN62941.SPACEXTBL
WHERE Landing__Outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

1
2016-04-08

Successful drone ship landing with payload between 4000 and 6000

- F9 FT B1032.1
- F9 B4 B1040.1
- F9 B4 B1043.1

Total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

In [10]:
```
%%sql SELECT Mission_Outcome, COUNT(Customer)
FROM TSN62941.SPACEXTBL
GROUP BY Mission_Outcome
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	2
Failure (in flight)	1
Success	99
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

In [11]:
```
%%sql SELECT booster_version
FROM TSN62941.SPACEXTBL
WHERE payload_mass__kg_ = (
    SELECT MAX(payload_mass__kg_) FROM TSN62941.SPACEXTBL)
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdoma
Done.
```

Out[11]:

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 2017

In [12]:
```
%%sql SELECT MONTHNAME(date), landing__outcome, booster_version, launch_site
FROM TSN62941.SPACEXTBL
WHERE (YEAR(date) = 2017 and landing__outcome = 'Success (ground pad)')
LIMIT 5
```

```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od
Done.
```

Out[12]:

1	landing__outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A

Rank success count between 2010-06-04 and 2017-03-20

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

In [13]:
```
%%sql SELECT landing__outcome, count(mission_outcome)
FROM TSN62941.SPACEXTBL
WHERE (landing__outcome LIKE 'Suc%' and Date > '2010-06-04' and Date < '2017-03-20')
GROUP BY landing__outcome
ORDER BY count(mission_outcome) DESC
```

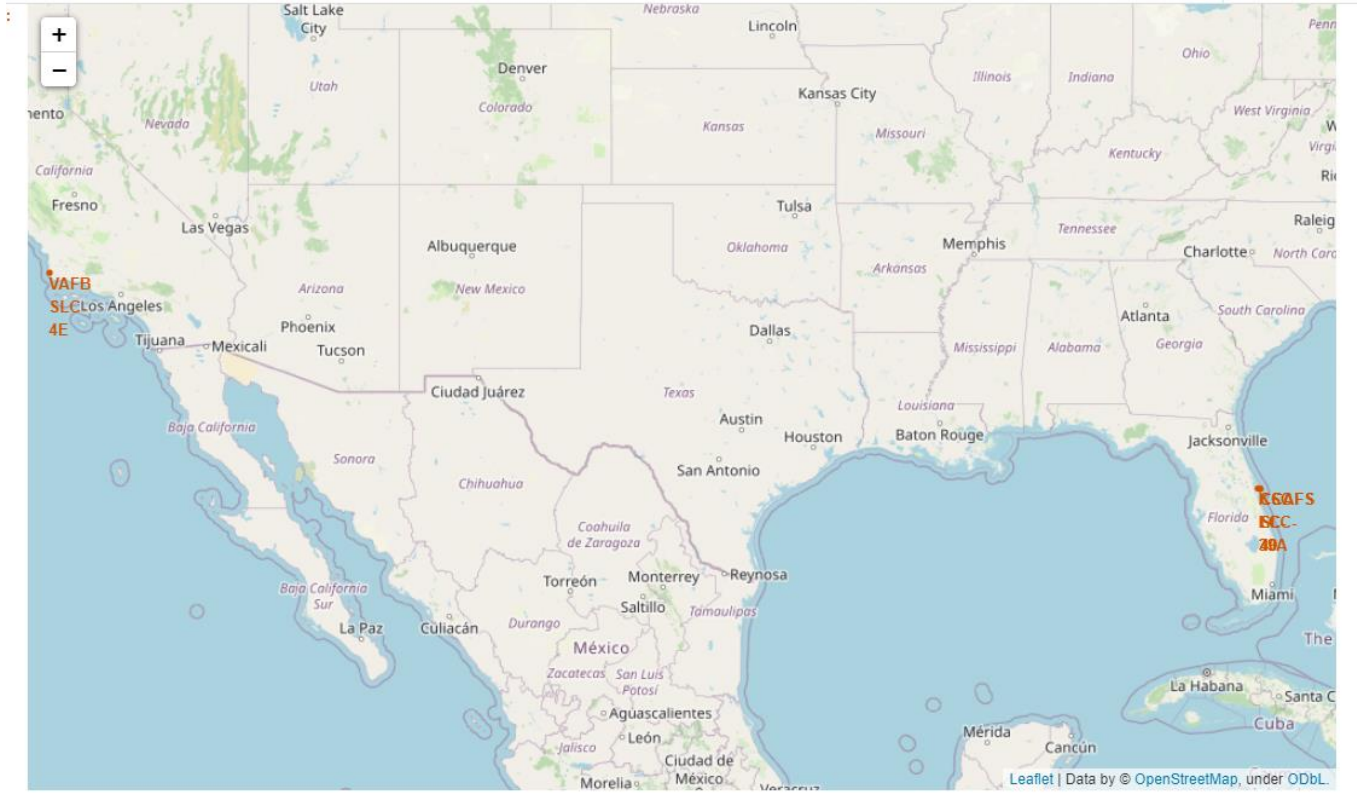
```
 * ibm_db_sa://tsn62941:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

landing__outcome	2
Success (drone ship)	5
Success (ground pad)	3

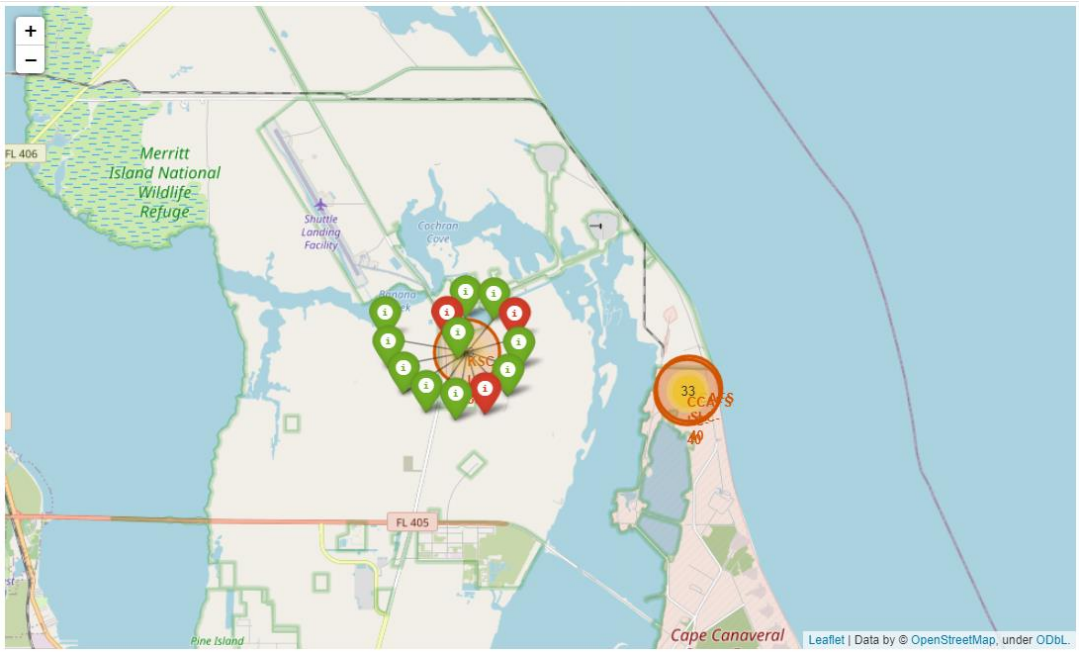
Interactive map with Folium

Folium: All Launch Sites



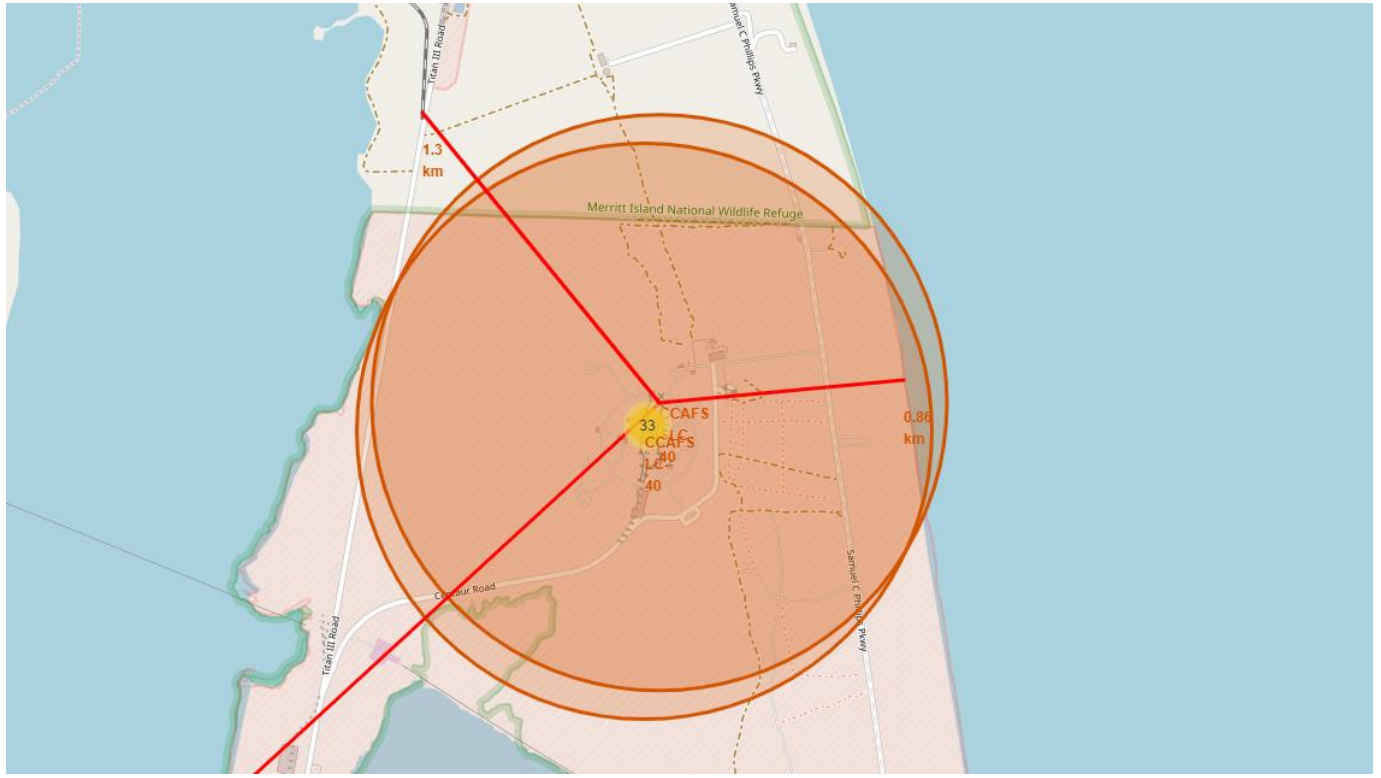
The launch sites in Florida and California can be seen on this map

Folium: Success Rate By Launch Site



Each green marker indicates a successful outcome, red markers indicate unsuccessful outcomes

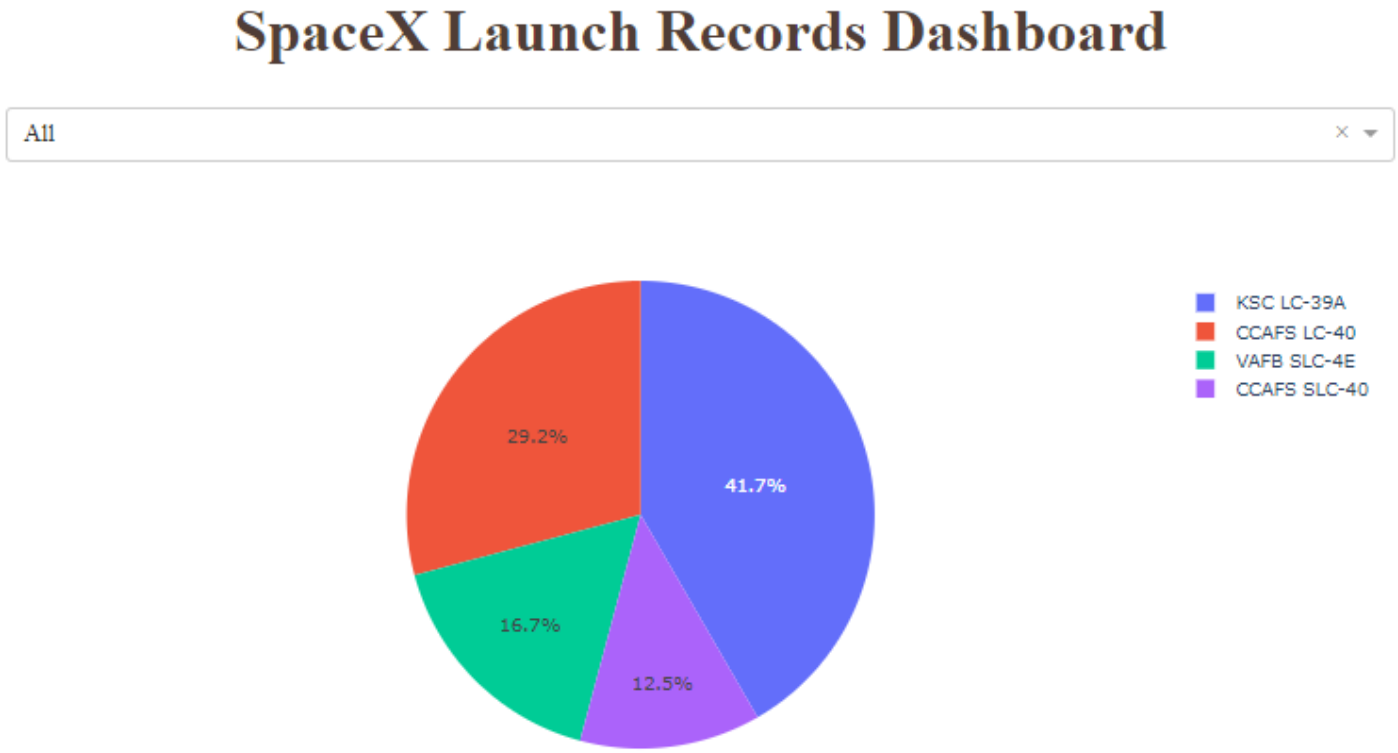
Folium: Distances from launch site



The lines indicate the distance between the nearest railway, coastline and city

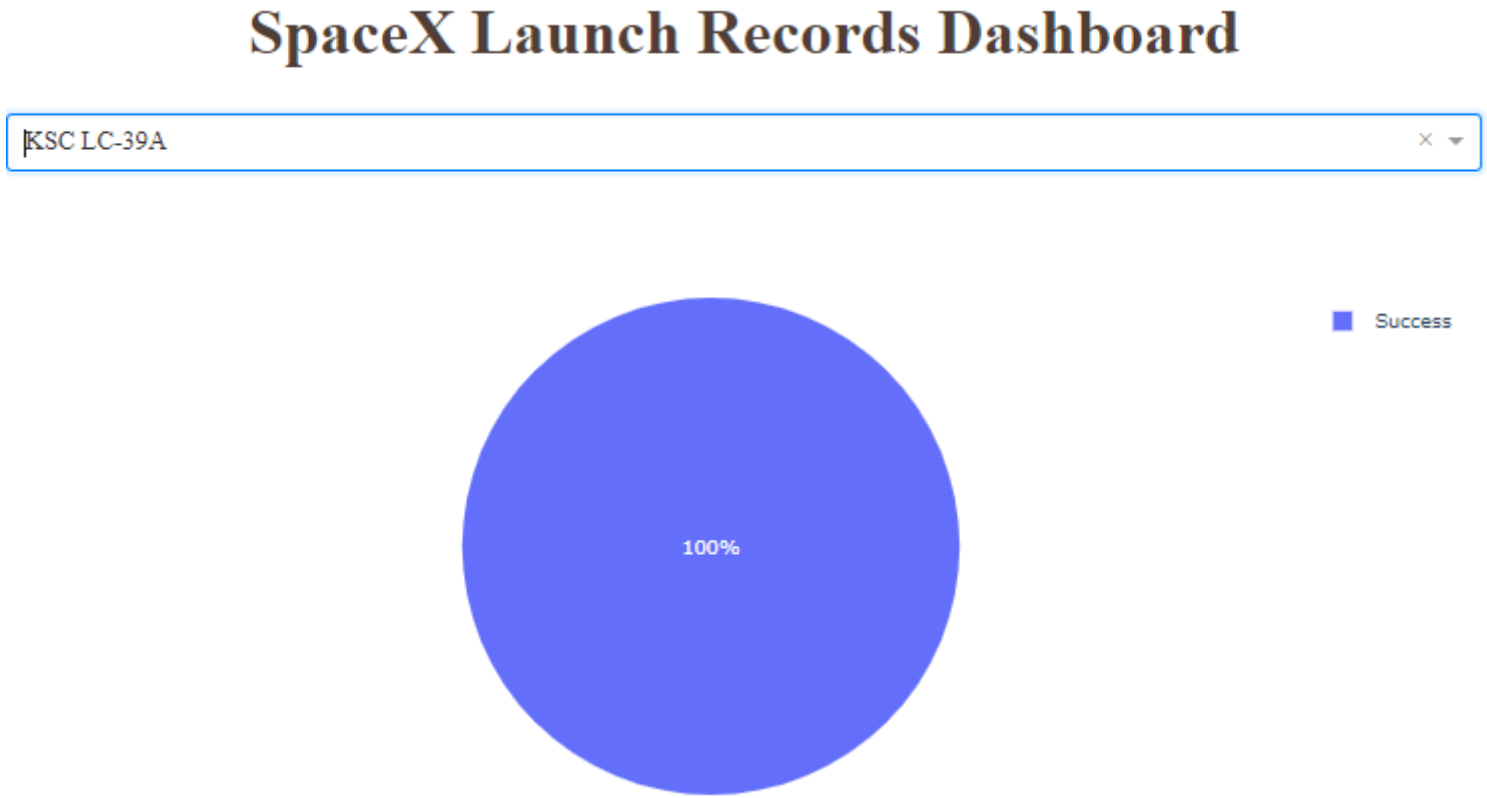
Build a Dashboard with Plotly Dash

Share of Successful Launches Per Launch Site



Pie chart shows the share of successful launches per launch site. Most successful launches are from KSC LC-39A

Highest Success Rates



The best success rates are found at KSC LC-39A and VAFB SLC-4E

Scatter Plot: Mission Outcome and Payload Mass

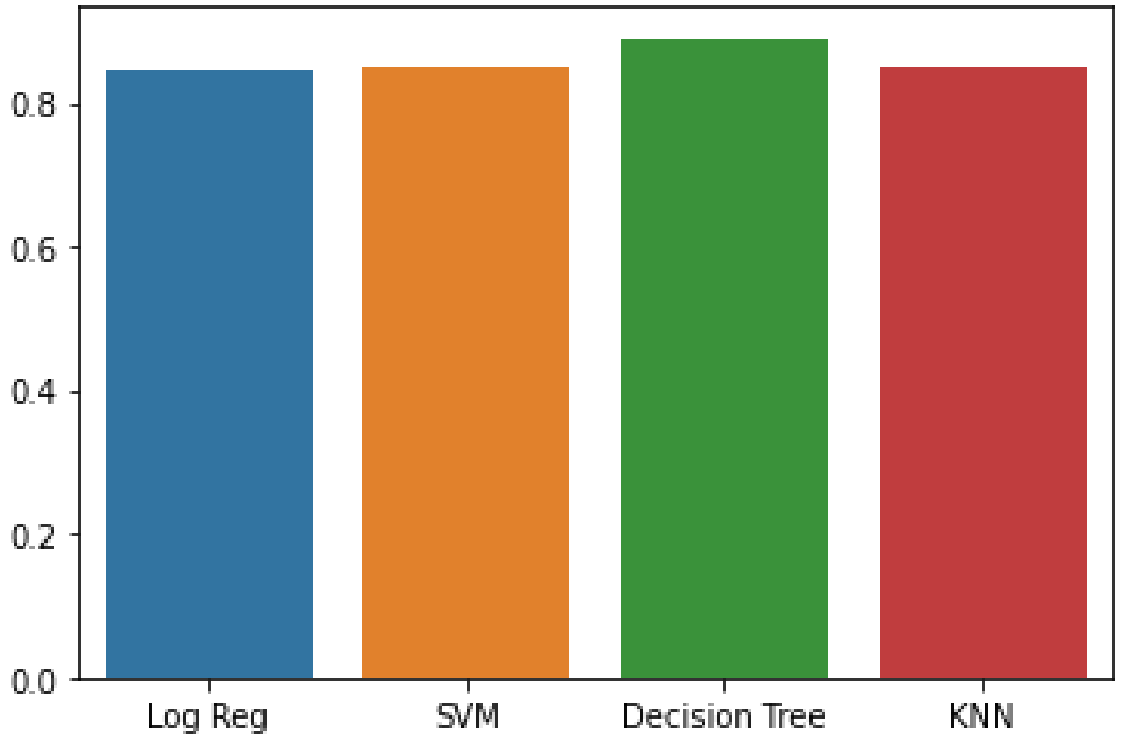
- 1 means success, 0 means failure
- The colors indicate the booster Type

Predictive analysis (Classification)

Classification Accuracy

This Chart shows the in sample accuracy for each mode since out of sample accuracy is equal for all models.

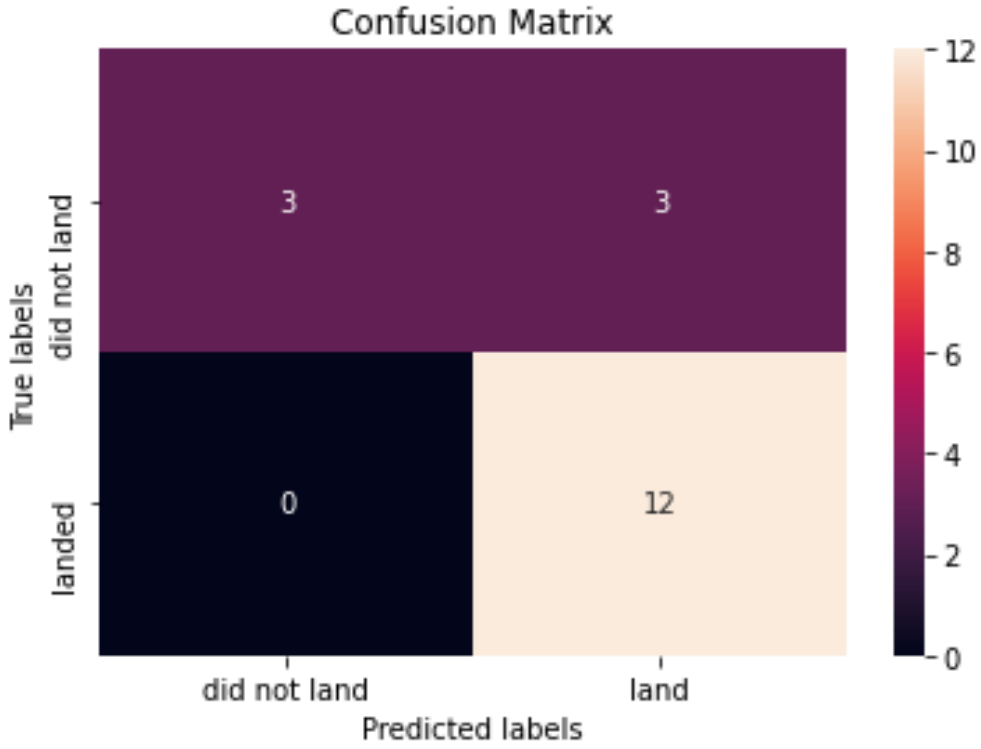
Decision Tree performs best.



Confusion Matrix

Decision Tree correctly classified 3 instances as “did not land” and 12 instances as “landed”.

It wrongly classified 3 instances as “did not land” despite they actually did land



CONCLUSION

- Very Heavy Pay Loads are Problematic
- Success rates differ by launch sites. We need to figure out what sets them apart.
- The type of orbit is relevant to the outcome
- Success increased over time. We need to figure out what changes were made by Space X.

APPENDIX

All GitHub-Files:

https://github.com/OliFre94/PythonDSFinalProject/tree/master