IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cristian Castañeda

July 14, 2024

Outline

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

Executive Summary

- SpaceY is a new commercial rocket launch provider who wants to bid against SpaceX.

- Given mission parameter such as payload mass and desired orbit, the models produced in this report were able to predict the first stage rocket booster landing successfully with an accuracy level of 83.3%.

- As a result, SpaceY will be able to make more informed bids against SpaceX by using first stage landing predictions as a proxy for the cost of a launch.

Introduction

Background

- This report has been prepared as part of the Applied Data Science Capstone course in the IBM Data Science Professional Certificate in Coursera.

Business Problem

- SpaceX advertises that the first stage of their Falcon 9 rocket launches can be reused.
- This report aims to accurately predict the likelihood of the first stage rocket landing successfully as a proxy for the cost of a launch.

Section 1

Methodology

Methodology

The methodology for this report was as follows:

- Data collection
- Data wrangling
- Exploratory data analysis
- Data visualization
- Model development

Data Collection

- API
  - Historical launch data form Open Source REST API for SpaceX

- Web Scraping
  - Historical launch data from Wikipedia: ‘List of Falcon 9 and Falcon Heavy Launches’

Data Wrangling

- Explored data to determine the label for training the model
  - Number of launches per site
  - Number and occurrence of orbits
  - Number and occurrence of mission outcome

- Created landing outcome data for training the model
  - Class 0: first stage booster did not land successfully
  - Class 1: first stage booster landed successfully

EDA

- EDA with SQL
  - Information on launch sites
  - Information of payload masses
  - Information on booster version and landings
  - Information on mission outcomes

- EDA with Visualization
  - Visualizations using Matplotlib and Seaborn

Interactive Map with Folium

- Launch Sites Location Analysis
  - Marked all sites on a map
  - Marked the successful/failed launches for each site on map
  - Calculated the distances between a launch site to its proximities

Dashboard with Plotly Dash

- Pie chart showing success rate
- Scatter plot showing payload mass vs. landing outcome
- Drop-down menu to choose between all sites

Predictive Analysis (Classification)

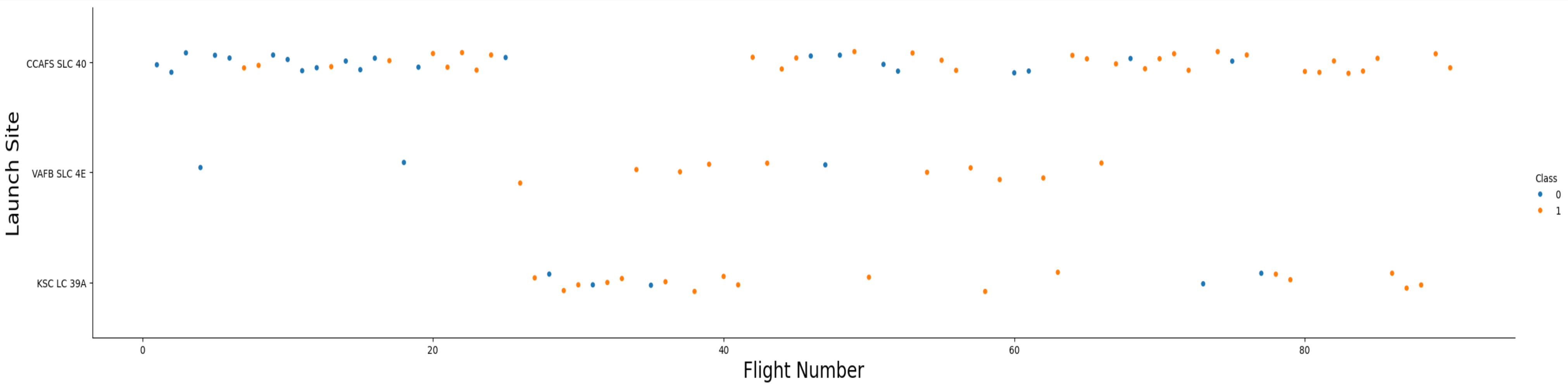
- Standardized the data
- Split the data into training and test data sets
- Fit the training data to various model types
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier
- Cross-validation for hyperparameters selection
- Evaluated accuracy of each model

Section 2

Insights drawn from EDA

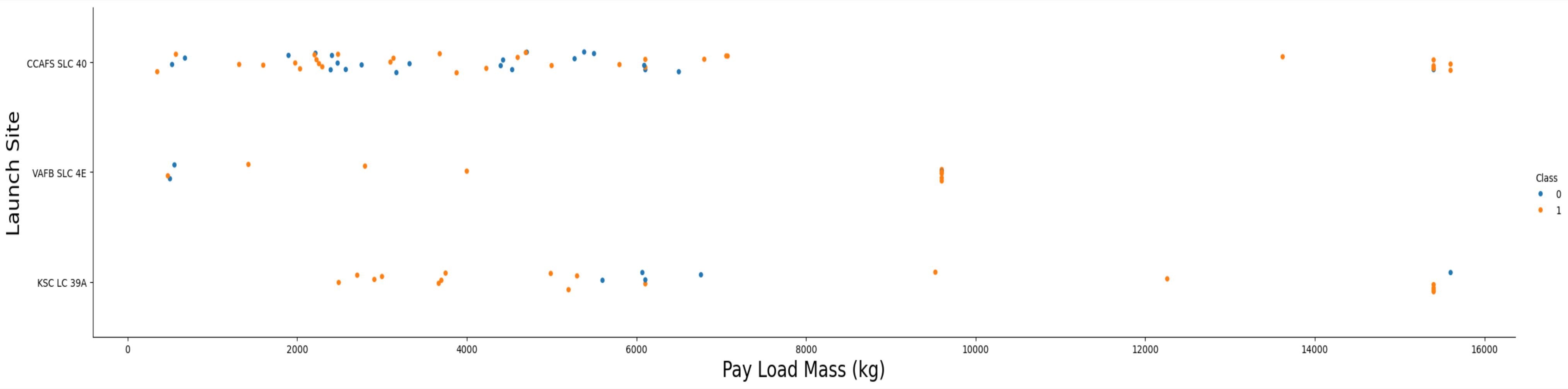
Flight Number vs. Launch Site

- We can observe that CCAFS SLC 40 had more flights than the other launch sites.



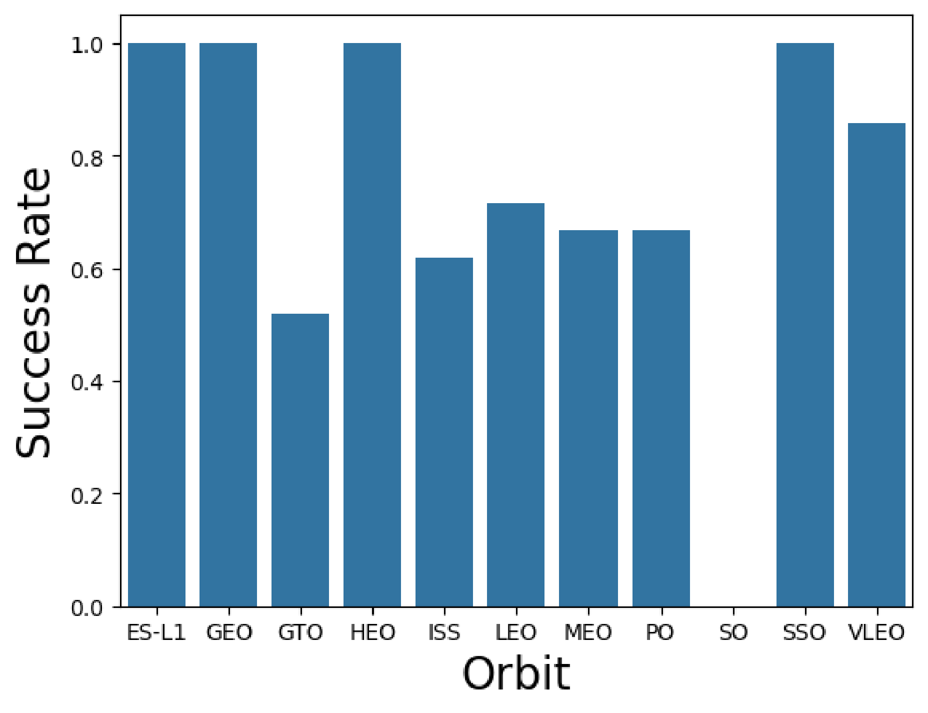
Payload vs. Launch Site

- The VAFB-SLC launch site has no rocket launches for heavy payload mass.



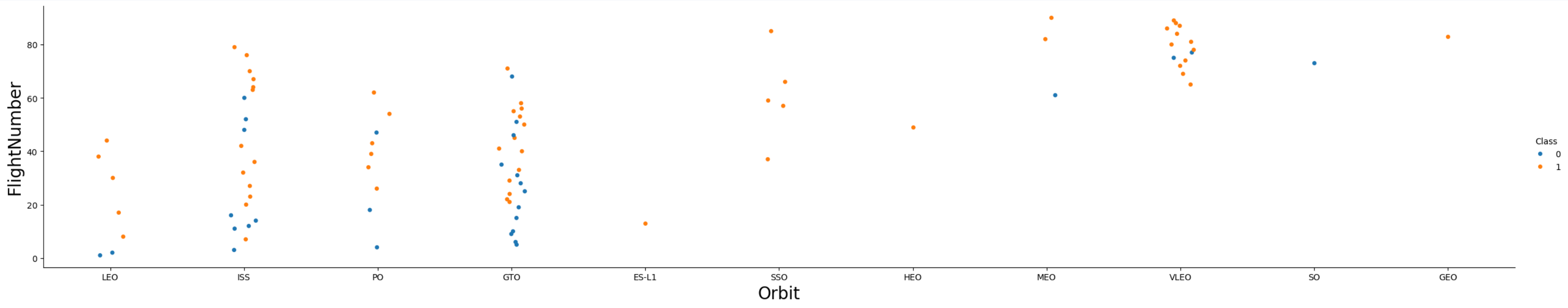
Success Rate vs. Orbit Type

- Missions ES-L1, GEO, SSO and HEO had the highest success rates



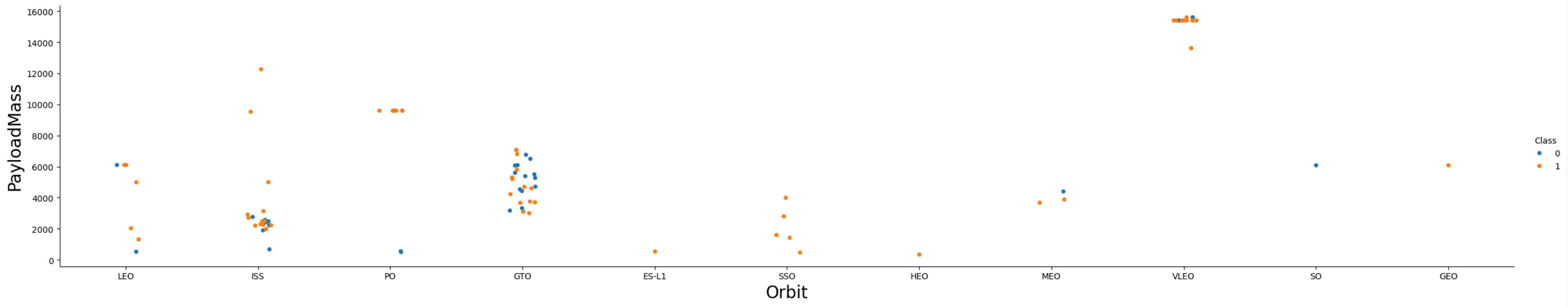
Flight Number vs. Orbit Type

- With the LEO orbit, the success seems to be related to the number of flights.



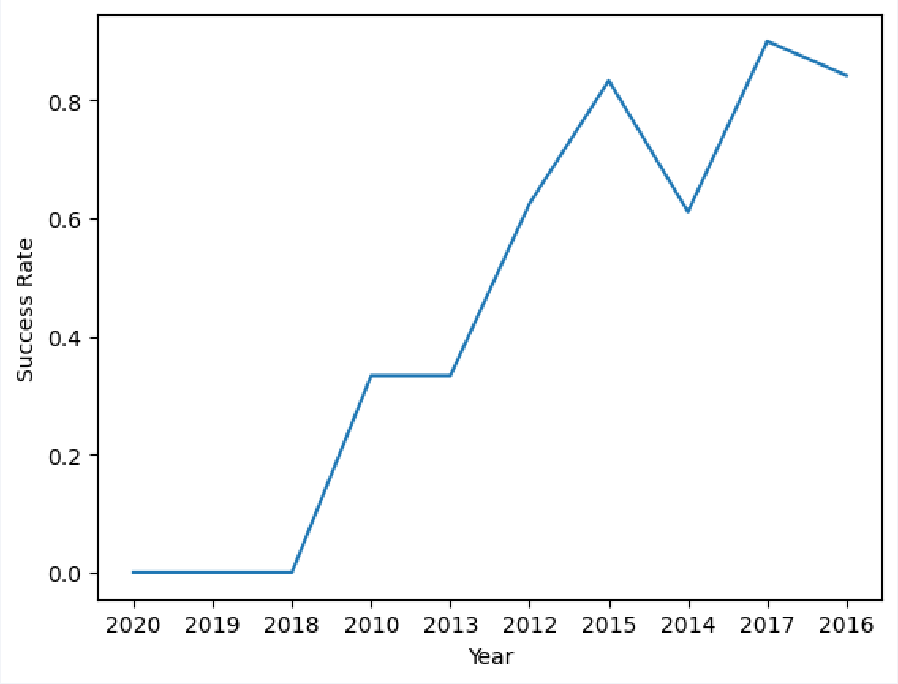
Payload vs. Orbit Type

- With heavy loads Polar, LEO and ISS have more successful landings.



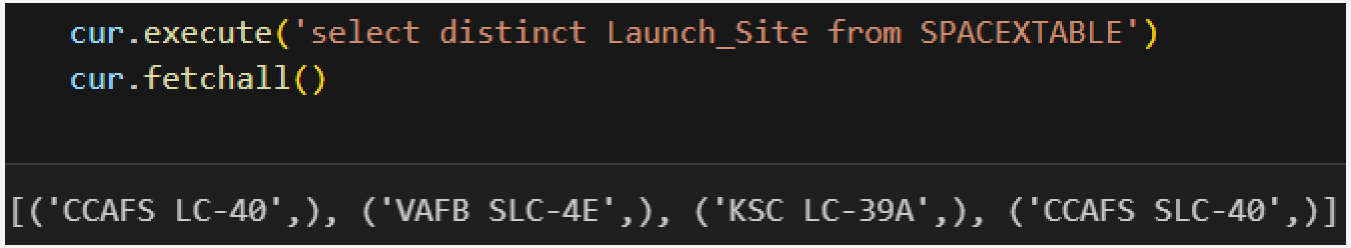
Launch Success Yearly Trend

- The yearly success rate seems to be increasing every year, with a few exceptions.



All Launch Site Names

- We can use the SELECT DISTINCT statements to retrieve the unique values of the launch sites.

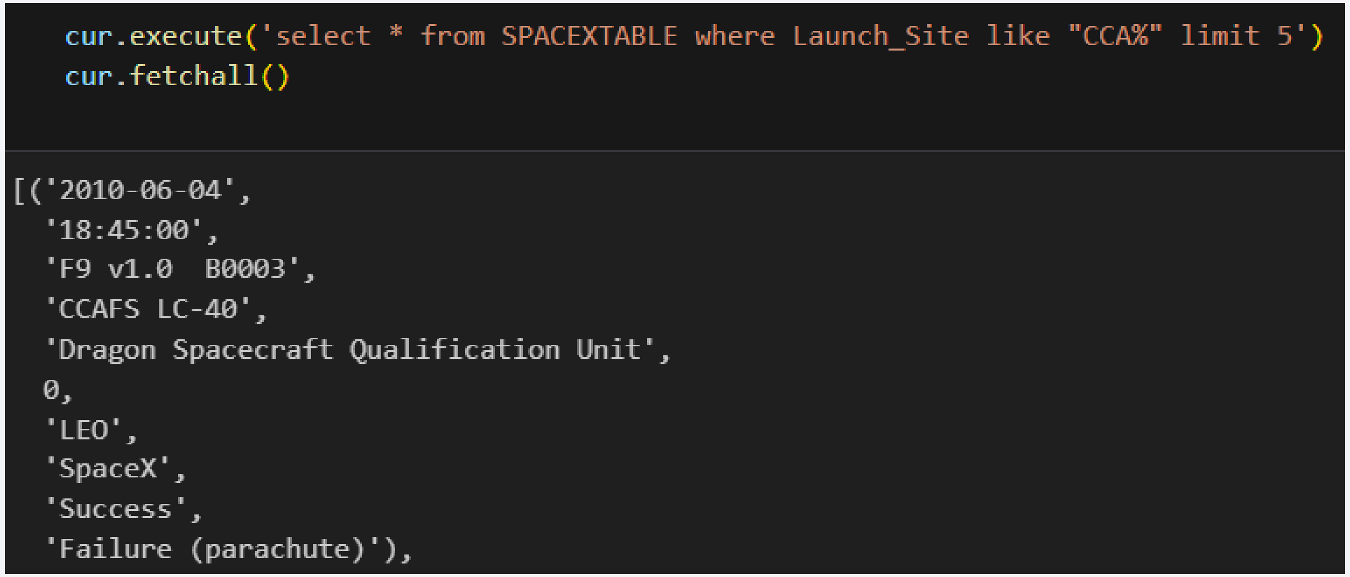


```
cur.execute('select distinct Launch_Site from SPACEXTABLE')
cur.fetchall()
```

```
[('CCAFS LC-40',), ('VAFB SLC-4E',), ('KSC LC-39A',), ('CCAFS SLC-40',)]
```

Launch Site Names Begin with 'CCA'

- We select the information of the first five launch sites that begin with CCA

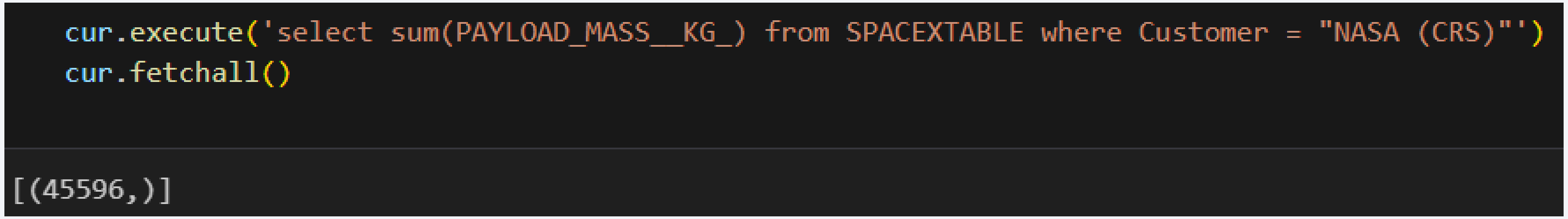


```
cur.execute('select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5')
cur.fetchall()
```

```
[('2010-06-04',
  '18:45:00',
  'F9 v1.0  B0003',
  'CCAFS LC-40',
  'Dragon Spacecraft Qualification Unit',
  0,
  'LEO',
  'SpaceX',
  'Success',
  'Failure (parachute)'),
```

Total Payload Mass

- We use the SUM statement to get the total payload mass

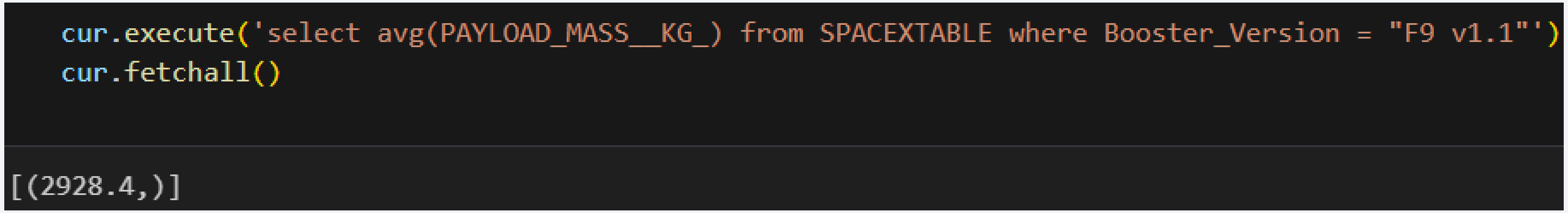


```
cur.execute('select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"')
cur.fetchall()
```

```
[(45596,)]
```

Average Payload Mass by F9 v1.1

- We can use the AVG statement to get the average payload mass for the F9 v1.1

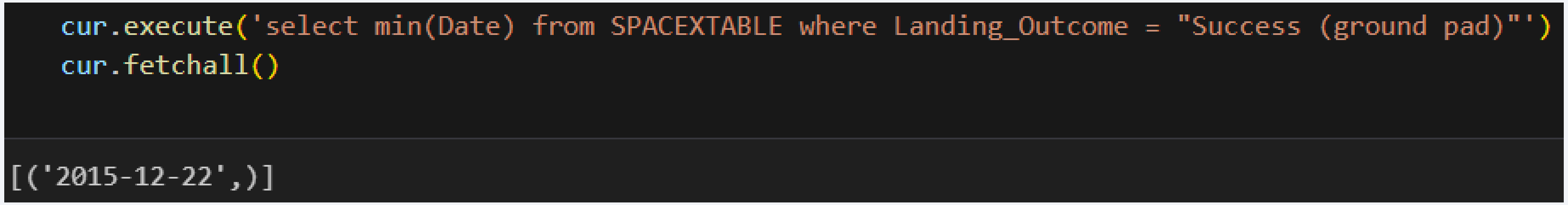


```
cur.execute('select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"')
cur.fetchall()
```

```
[(2928.4,)]
```

First Successful Ground Landing Date

- We use the MIN statement to get the earlier date with a successful landing

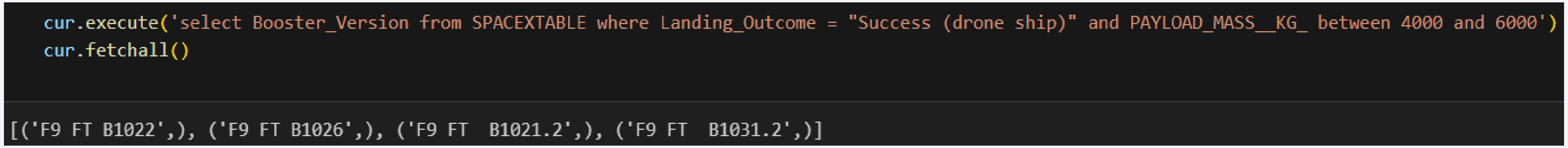


```
cur.execute('select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"')
cur.fetchall()
```

```
[('2015-12-22',)]
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- We use the BETWEEN statement to get the information in the interval declared and see the success landing sites

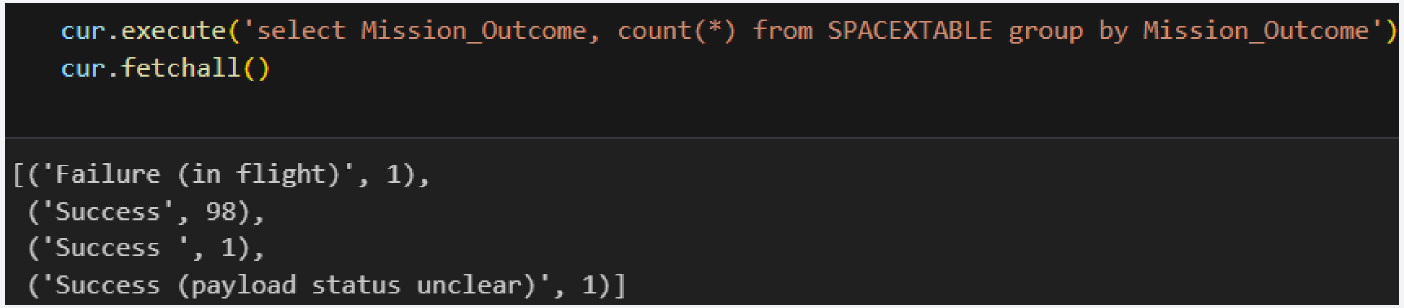


```
cur.execute('select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000')
cur.fetchall()
```

```
[('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT  B1021.2',), ('F9 FT  B1031.2',)]
```

Total Number of Successful and Failure Mission Outcomes

- We see the number of mission outcome using the GROUP BY statement

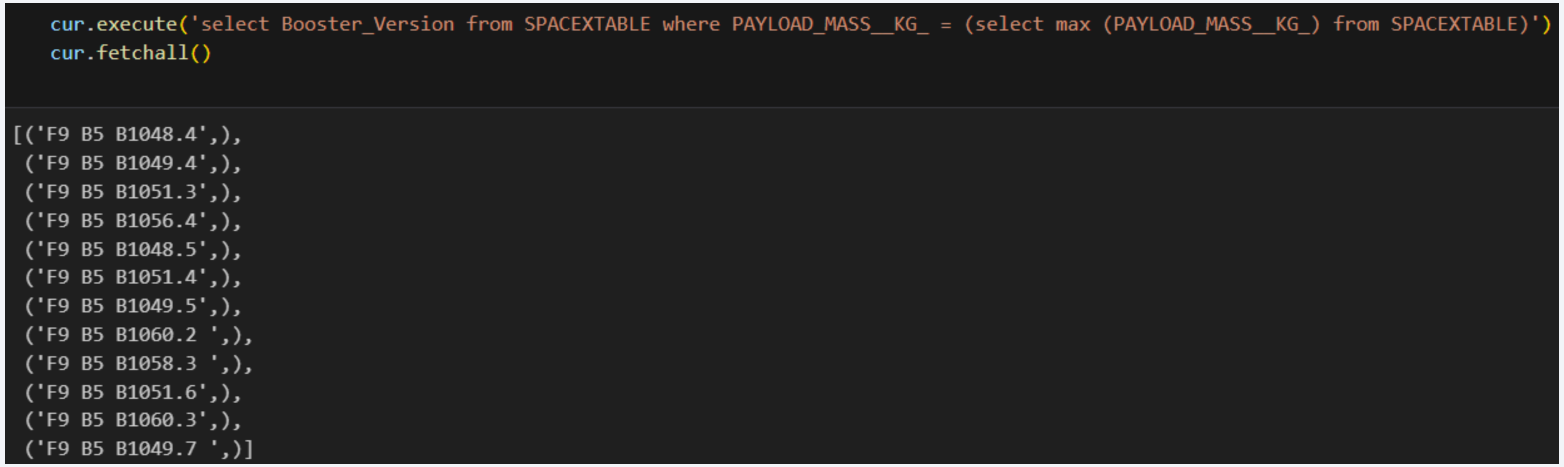


```
cur.execute('select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome')
cur.fetchall()
```

```
[('Failure (in flight)', 1),
 ('Success', 98),
 ('Success ', 1),
 ('Success (payload status unclear)', 1)]
```

Boosters Carried Maximum Payload

- We see the boosters maximum payload carried

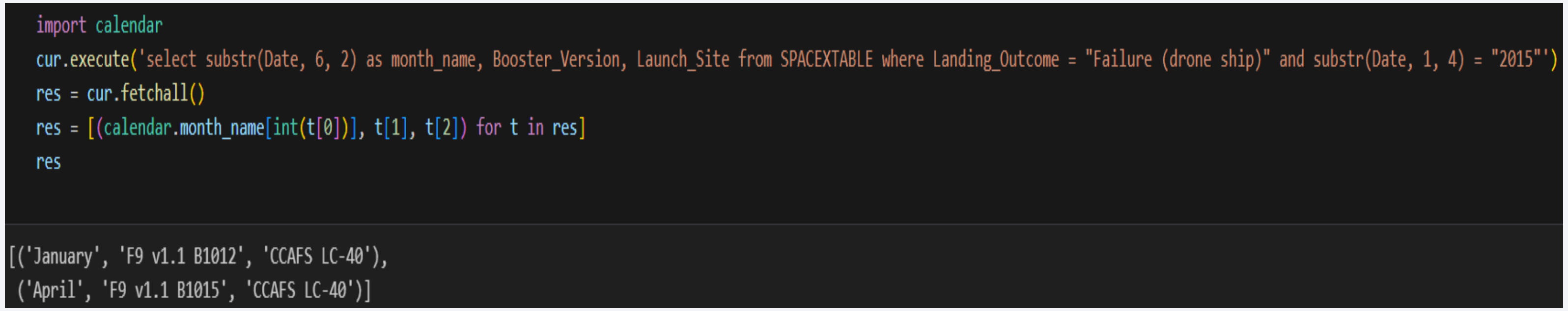


```
cur.execute('select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max (PAYLOAD_MASS__KG_) from SPACEXTABLE)')
cur.fetchall()
```

```
[('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

- We see the 2025 launch records

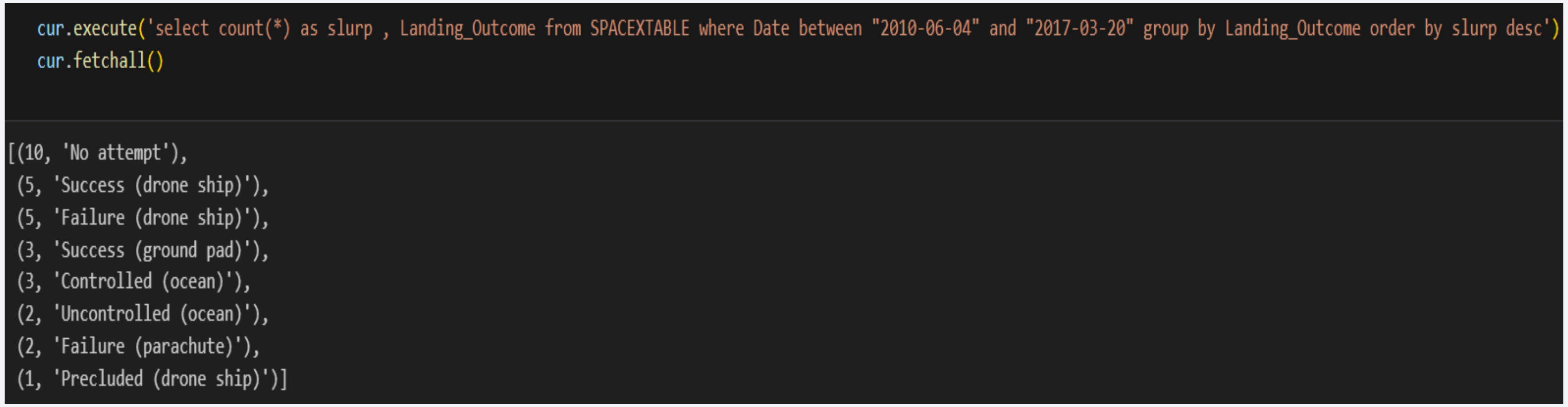


```
import calendar
cur.execute('select substr(Date, 6, 2) as month_name, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome = "Failure (drone ship)" and substr(Date, 1, 4) = "2015"')
res = cur.fetchall()
res = [(calendar.month_name[int(t[0])], t[1], t[2]) for t in res]
res
```

```
[('January', 'F9 v1.1 B1012', 'CCAFS LC-40'),
 ('April', 'F9 v1.1 B1015', 'CCAFS LC-40')]
```

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

- We see the different landing outcomes ranked between the given dates



```
cur.execute('select count(*) as slurp , Landing_Outcome from SPACEXTABLE where Date between "2010-06-04" and "2017-03-20" group by Landing_Outcome order by slurp desc')
cur.fetchall()
```

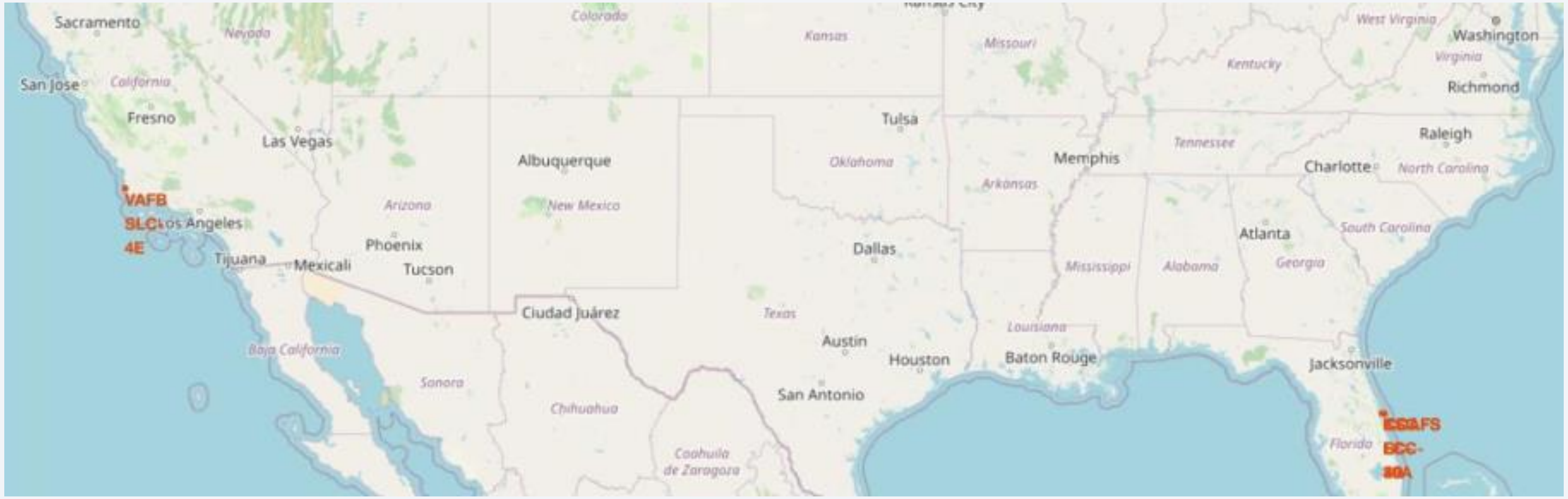
```
[(10, 'No attempt'),
 (5, 'Success (drone ship)'),
 (5, 'Failure (drone ship)'),
 (3, 'Success (ground pad)'),
 (3, 'Controlled (ocean)'),
 (2, 'Uncontrolled (ocean)'),
 (2, 'Failure (parachute)'),
 (1, 'Precluded (drone ship)')]
```

Section 3

Launch Sites Proximities Analysis

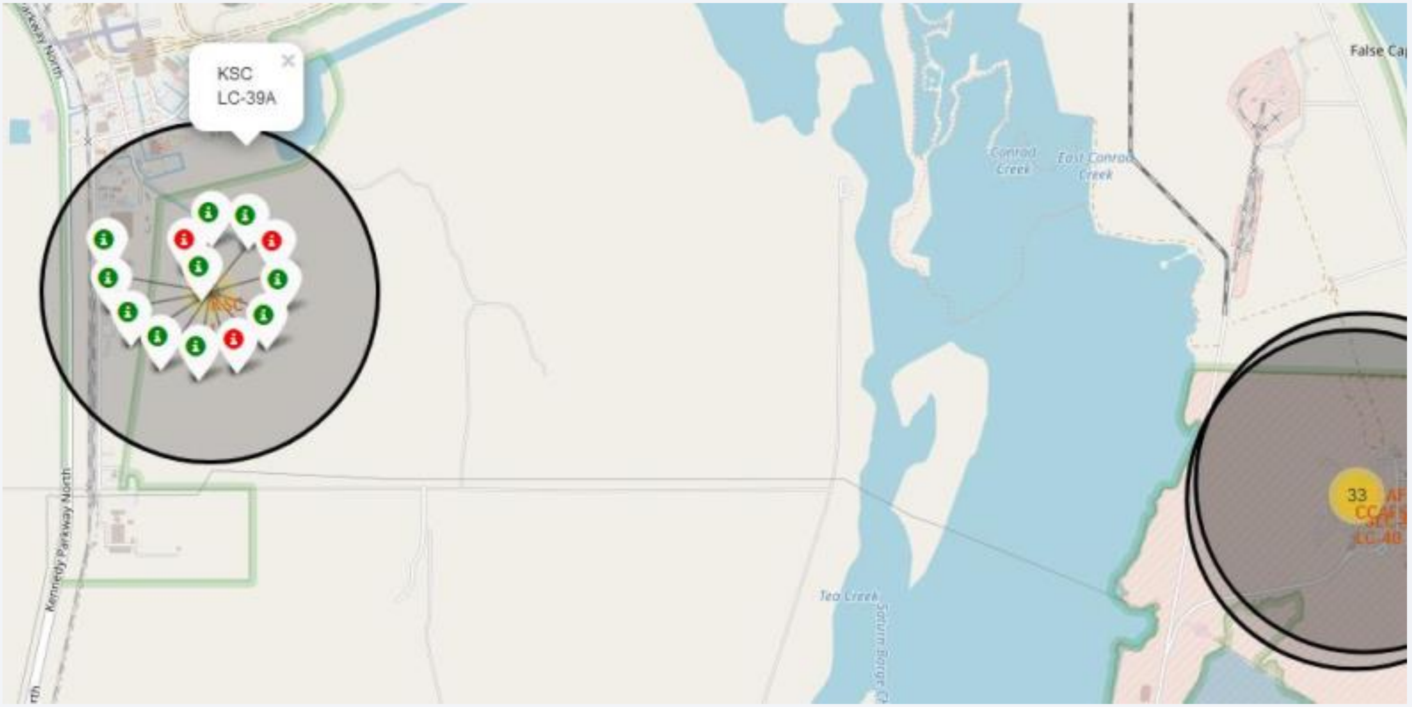
Launch Sites Locations

- Map with all the SpaceX launch sites



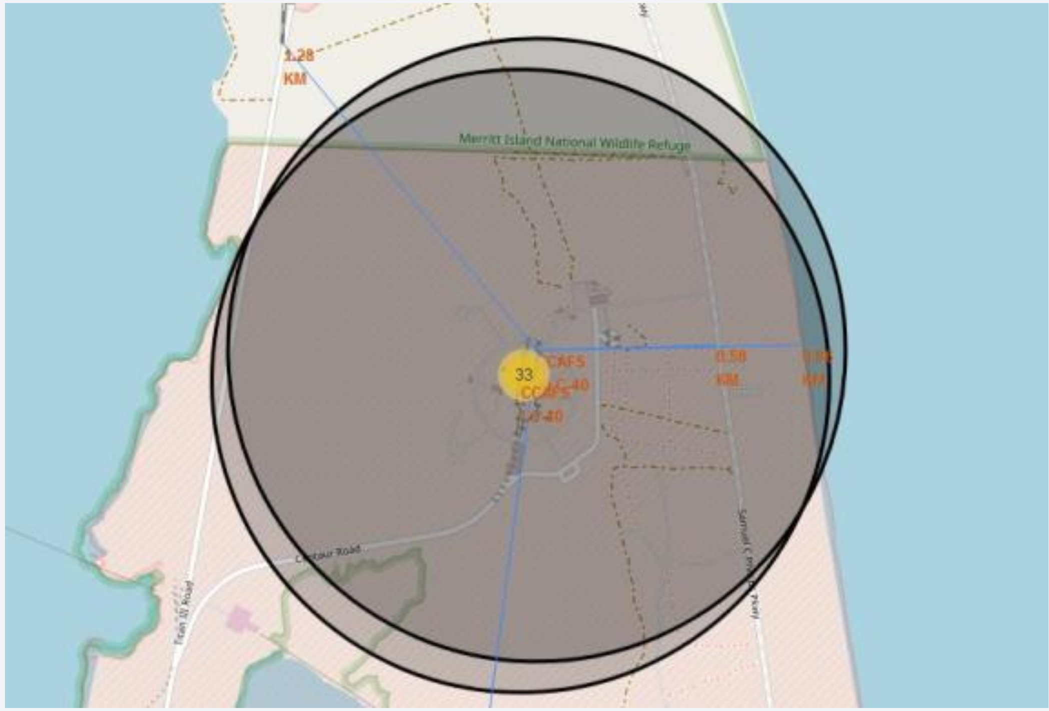
Successful/Failed Landings

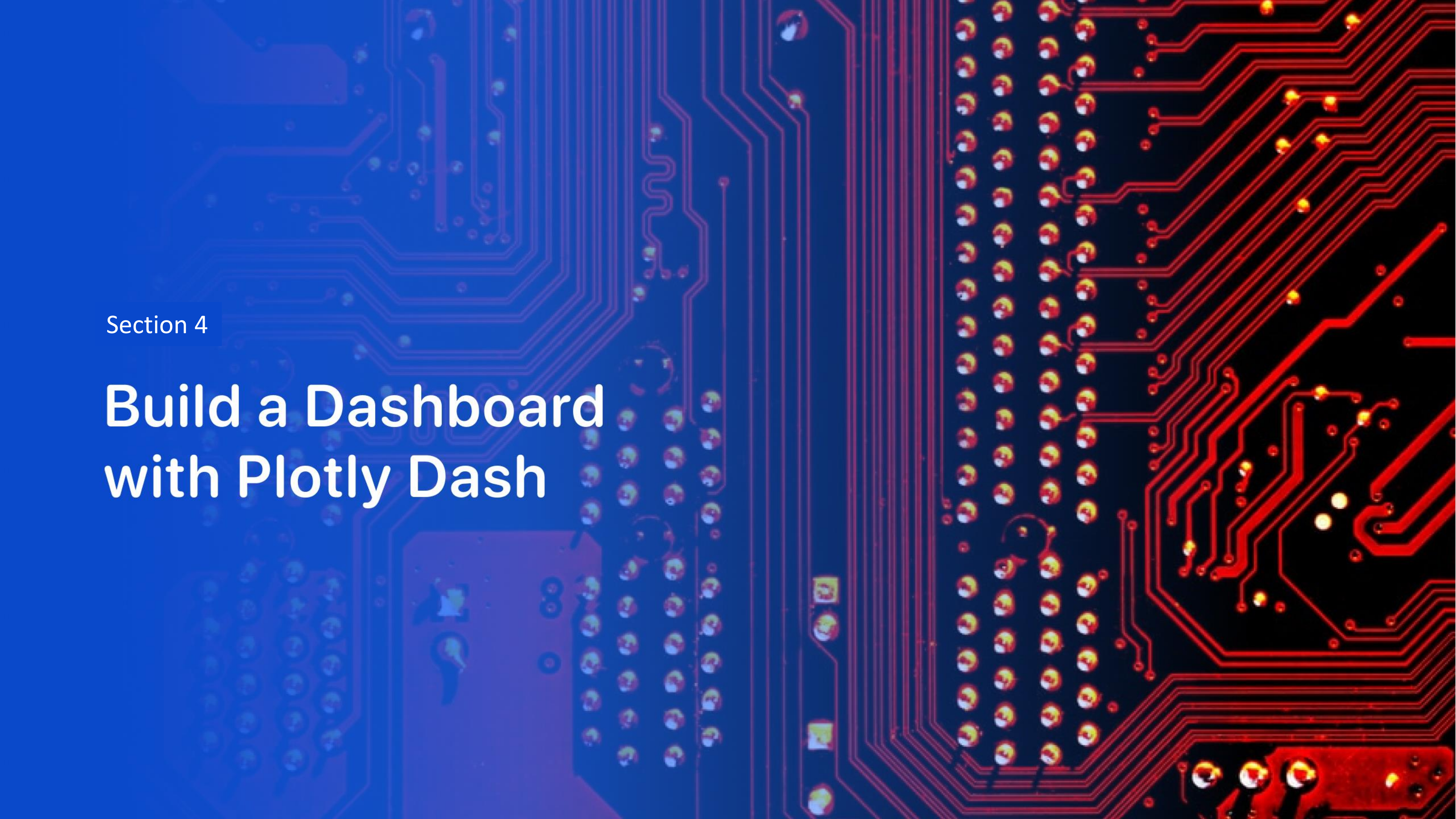
- Launch outcomes for each landing site



Railway, Highway, Coastline Proximities

- We see the proximity of the sites to different railways, highways and coastlines.



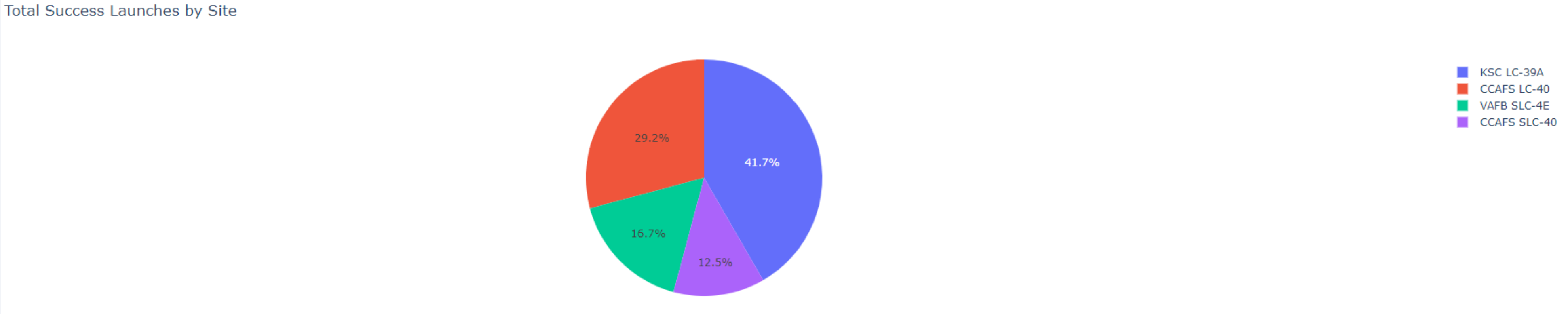


Section 4

Build a Dashboard with Plotly Dash

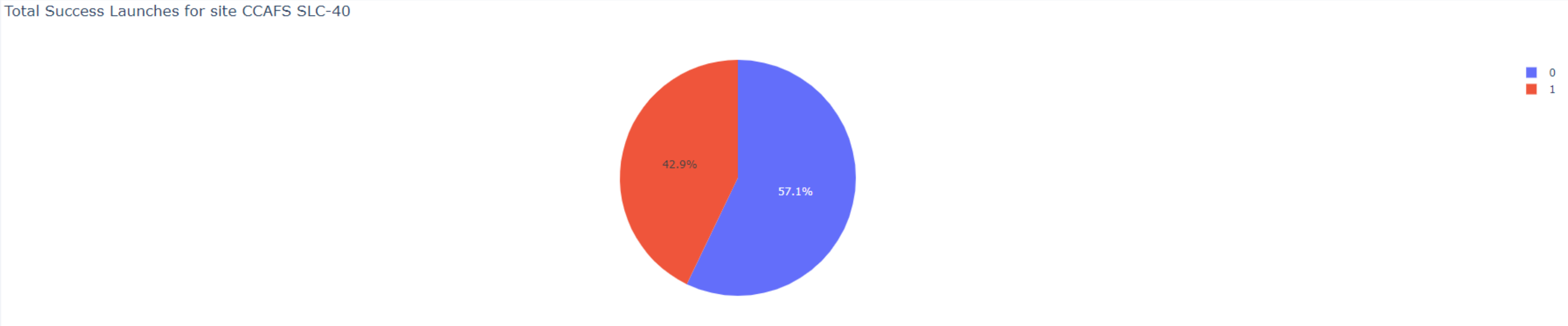
Total Success Launches by Site

- We see that the most successful launch site is KSC LC-39C with almost 42% of successful launches.



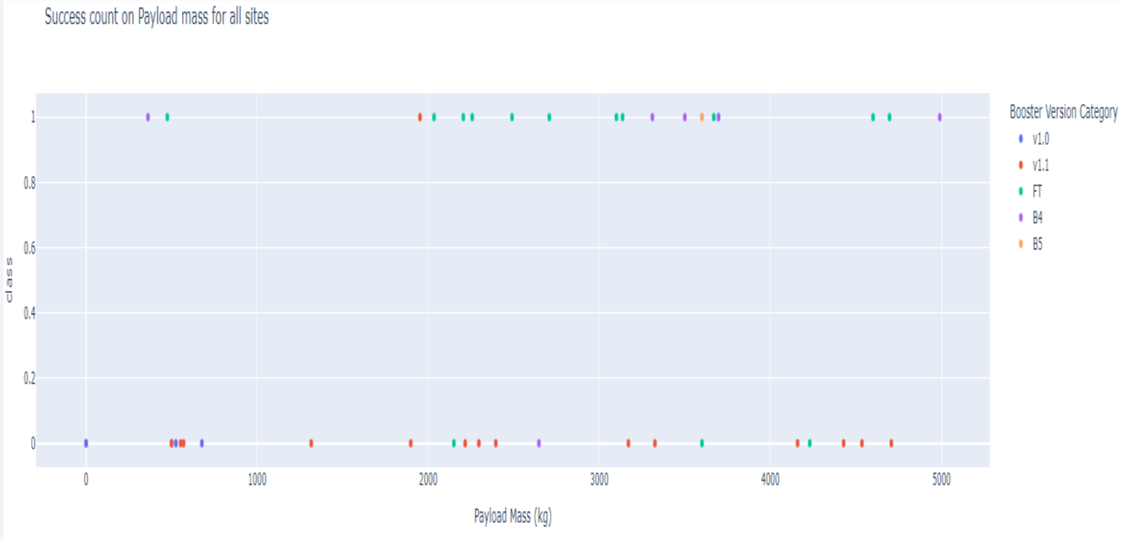
Highest Success Ratio

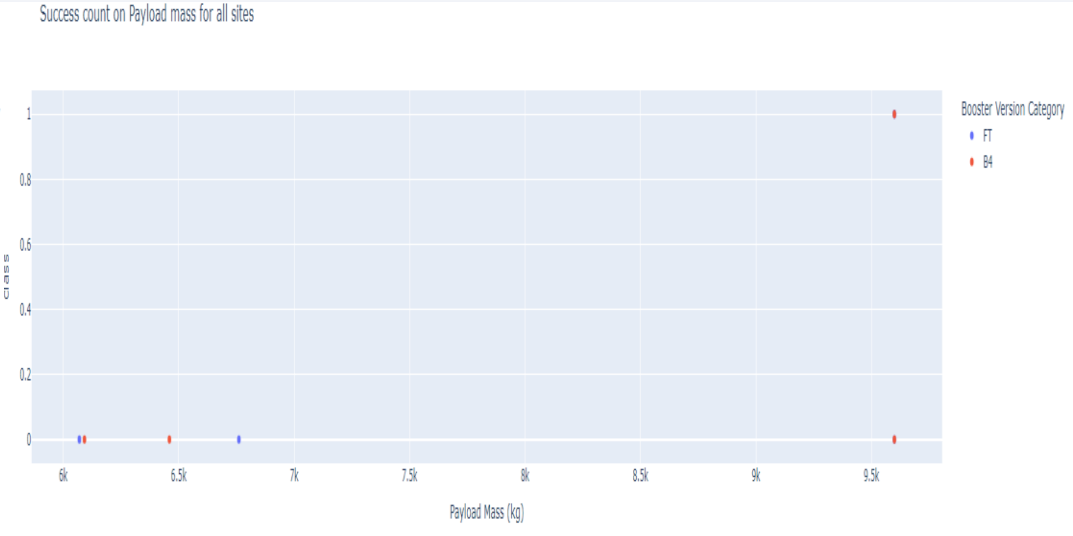
- We see that the site with the highest success ratio, with a 42.9% success, is the site CCAFS SLC-40



Success Count on Payload

- We see the success count on payload for all sites. On the left we see payloads between 0k-5k and on the right payloads between 6k-10k.



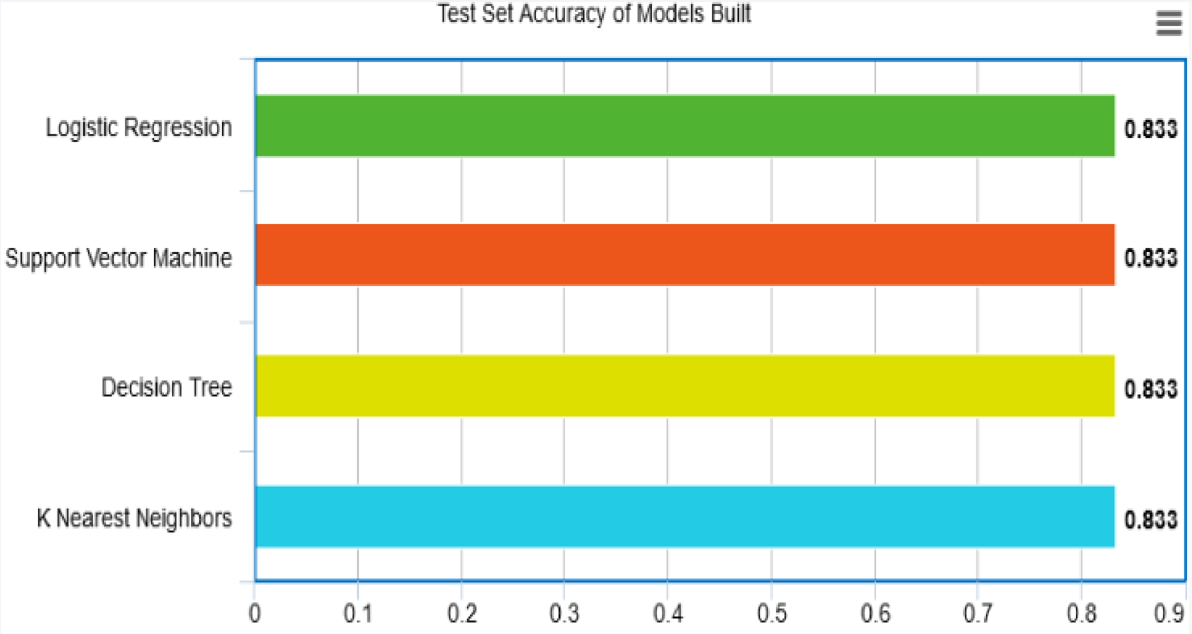


Section 5

Predictive Analysis (Classification)

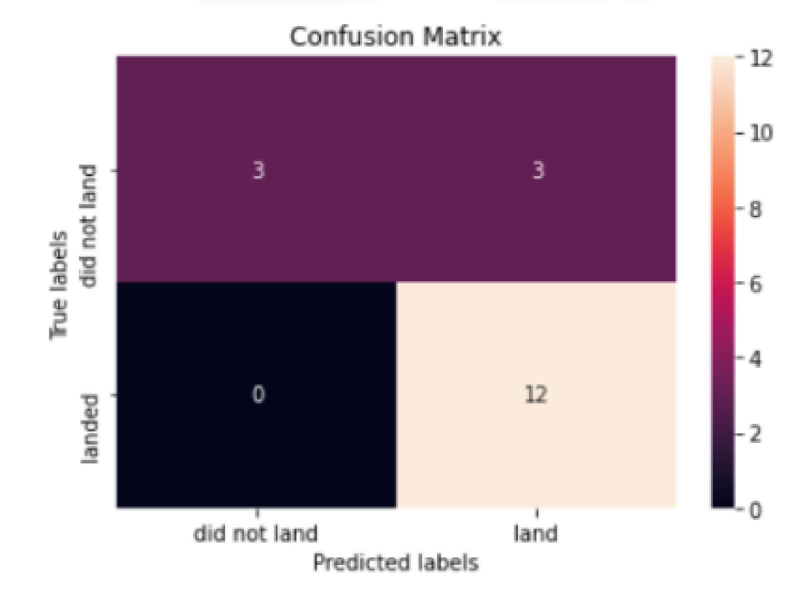
Classification Accuracy

- Every model came with an accuracy of 83.33%.



Confusion Matrix

- All the models have the same confusion matrix.
- We see that the major problem are false positives.



Conclusions

- Using the models from this report, SpaceY can predict when SpaceX will successfully land the first stage booster with 83.3% accuracy.
- We identified that CCAFS SLC-40 had the highest number of launches among the sites considered.
- An upward trend in payload mass over the years was observed.

Appendix

- All the codes and this presentation can be found in the next link

https://github.com/cristoforo93/AppliedDataScienceCapstone

Thank you!