IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neil de la Rey
11/06/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected using the SpaceX REST API
  - Data wrangling techniques used to create success / fail outcome variable
  - Data Visualisation techniques used to explore the payload, launch site, flight number and yearly trend factors
  - Data analysis performed using SQL
  - Launch site data visualised to display successful payload ranges
  - Model constructed to predict landing outcomes using SVM and KNN

- Summary of all results
  - Exploratory Analysis
    - Launch Success improved over time
    - Launch site KSC LC-39A has the highest success rate
  - Predictive Analysis
    - All models were similarly accurate, therefore more future data is required to make a decision in this regard
  - Launch sites
    - Most launch sites are near the coast and as close as possible to the equator

Introduction

- Perhaps the most successful commercial space agency is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station, Starlink, a satellite internet constellation providing satellite Internet access, & Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Spaces X's Falcon 9 launch like regular rockets.
- Therefore, if we can determine if the first stage will land, we can more accurately determine the cost of a launch.

Section 1

Methodology

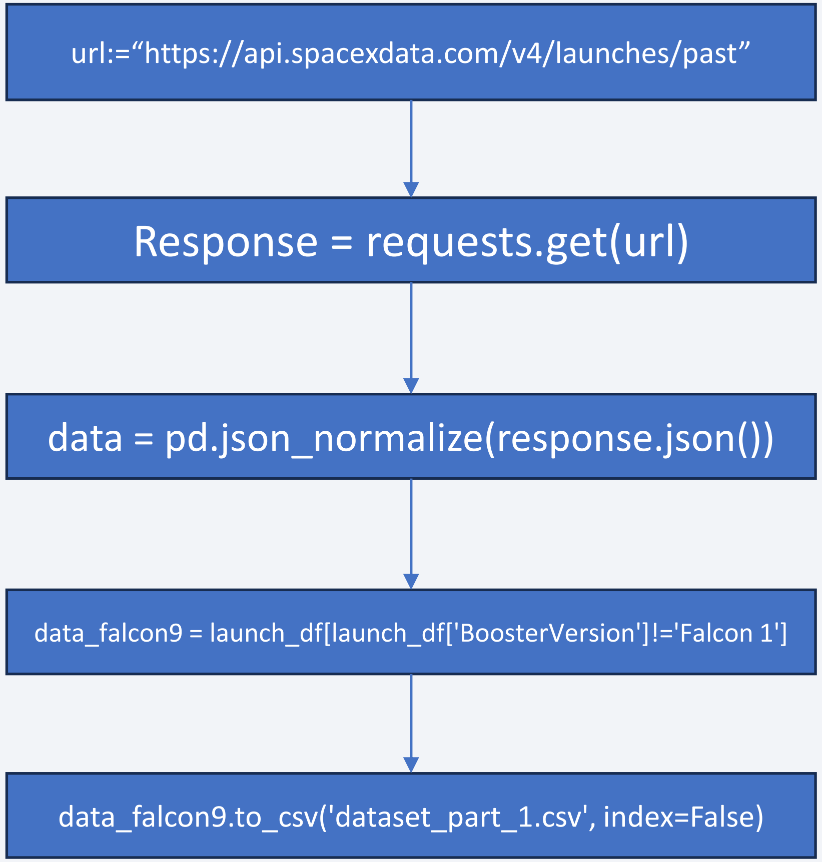
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX REST API
- Perform data wrangling
  - Missing values were handled during data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Landing models were produced using classification methods

Data Collection – SpaceX API

- Data collection with SpaceX REST API
  - Data response is in json format and loaded to a pandas dataframe
  - Dataframe is filtered to only display Falcon 9 launch data
  - Missing values replaced with calculated mean()
  - Data exported to dataset_part_1.csv



- GitHub URL of the completed SpaceX API calls notebook:

https://github.com/neildlr/coursera-dsmod10/blob/938191fee660c2df9689698ff7df376723b25cc2/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

- Data collection with Wikipedia web scraping
  - Using Beautifulsoup object from HTML response
  - Extract column names into table header
  - Missing values replaced with calculated mean()
  - Data exported to spacex_web_scraped.csv

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response = requests.get(static_url).text
soup = BeautifulSoup(response, 'html.parser')
df.to_csv('spacex_web_scraped.csv', index=False)
```

- GitHub URL of the completed web scraping notebook:

https://github.com/neildlr/coursera-dsmod10/blob/938191fee660c2df9689698ff7df376723b25cc2/jupyter-labs-webscraping.ipynb

Data Wrangling

- Perform EDA
- Calculate:
  - Num of launches for each site
  - Num and occurrence of orbit
  - Num and occurrence of mission outcome per orbit type
- Create binary landing outcome
- Export to csv file

- Landing outcomes:
  - Landings not always successful
  - True / False Ocean
  - True / False RTLS
  - True / False ASDS
- Outcomes converted to indicate 1 for successful and 0 for unsuccessful landing

- GitHub URL of the completed data wrangling notebook:

  https://github.com/neildlr/coursera-dsmod10/blob/938191fee660c2df9689698ff7df376723b25cc2/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Flight number vs Payload
- Flight number vs Launch Site
- Payload Mass vs Launch Site
- Payload Mass vs Orbit type

- GitHub URL of the completed EDA data visualisation notebook:

https://github.com/neildlr/coursera-dsmod10/blob/938191fee660c2df9689698ff7df376723b25cc2/edadataviz.ipynb

EDA with SQL

- Display
  - Names of unique launch sites
  - 5 records where launch site name begins with CCA
  - Total Payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1

- GitHub URL of the completed SQL data visualisation notebook:

  https://github.com/neildlr/coursera-dsmod10/blob/4e34a16c5893eb6945a173159b235b366d66fbec/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- List
  - Date when first successful ground pad landing was achieved
  - Names of boosters with drone ship landing success and payload between 4000kg and 6000kg
  - Total number of Success and Failure outcomes
  - Names of boosters that have carried maximum payload
  - Failed landing outcomes on drone ship, its booster version and launch site for each month in 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

- Markers Indicating Launch Sites
- Added circle at NASA Johnson Space Centre
- Added red circles at all launch sites
- Added coloured markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- Added coloured lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

- GitHub URL of the completed folium notebook:

https://github.com/neildlr/coursera-dsmod10/blob/4e34a16c5893eb6945a173159b235b366d66fbec/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown list with Launch Sites
  - Allow selection of one or multiple sites
- Pie Chart Showing Successful Launches
  - % of Successful vs Unsuccessful launches
- Slider of Payload mass range
  - Allow selection of payload mass range
- Scatter chart showing payload mass vs success rate by booster version
  - Visualise the relation between payload and launch success

- GitHub URL of the completed plotly dash code:

https://github.com/neildlr/coursera-dsmod10/blob/1f83ad38e5ec087be5cbd9af3d89c489ada48e73/Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- Create a NumPy array from the Class column in the data
- Standardise the data
- Split the data into training and test data
- Use logistic regression to find the best parameters
- Calculate the accuracy of the test data
- Assess the confusion matrix for all models
- Identify the best model using score and accuracy

- GitHub URL of the predictive analysis notebook:

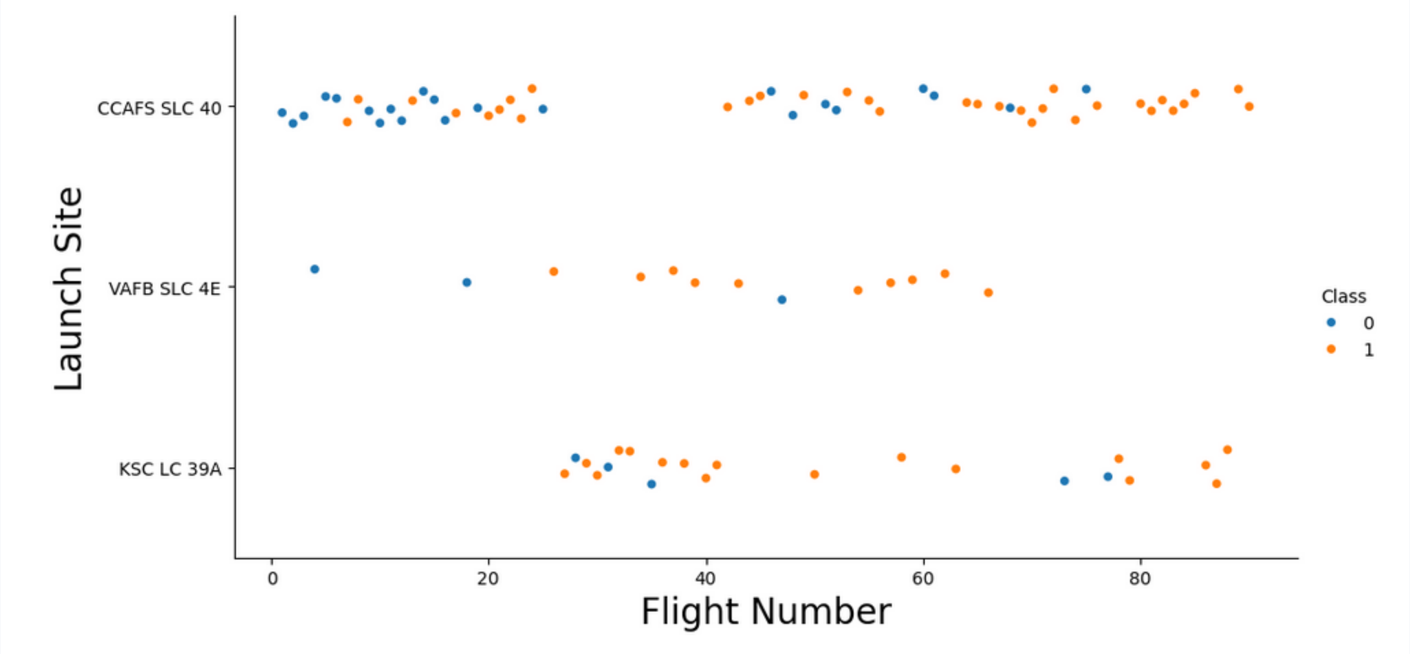
https://github.com/neildlr/coursera-dsmod10/blob/c9a95e3d4b9368384d258217aa860fc89b684f3b/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Section 2

Insights drawn from EDA

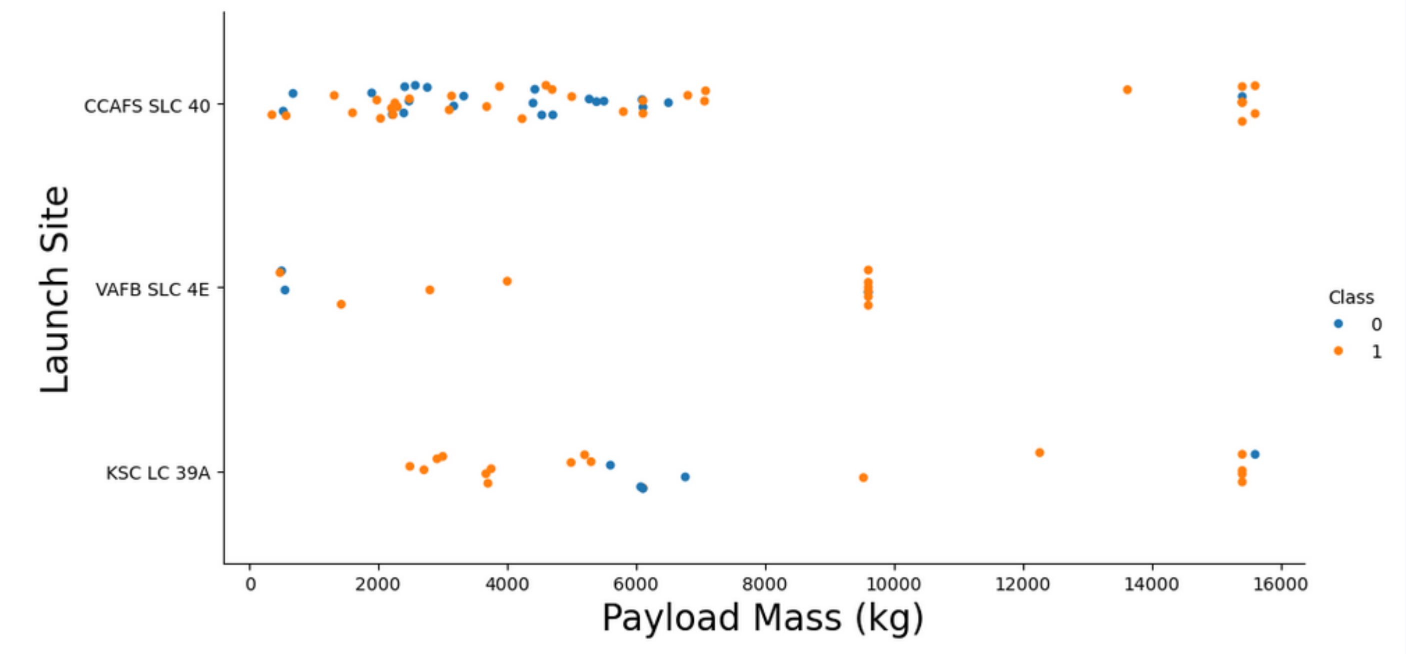
Flight Number vs. Launch Site

- Blue dots represent failures while orange dots represent success
- From this we can see that later flights tend to be more successful which displays launch system improvements over time



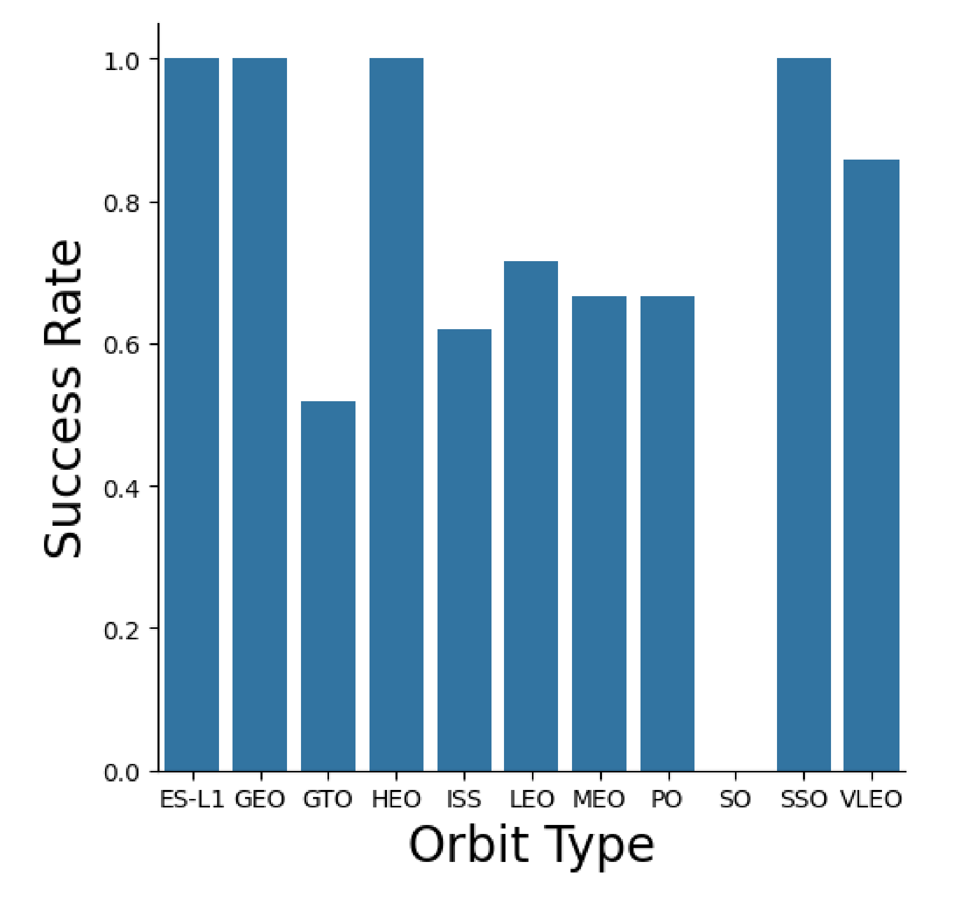
Payload vs. Launch Site

- Very few failures occurred with a payload mass larger than 8000kg
- Typically, we see a higher success rate with a higher payload mass



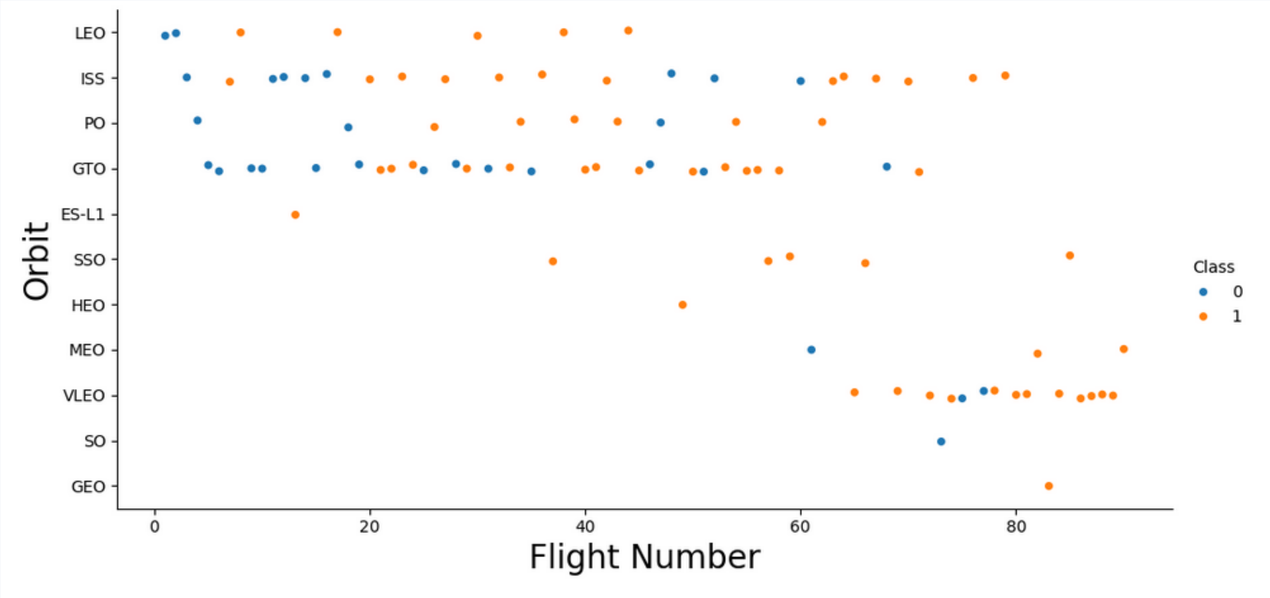
Success Rate vs. Orbit Type

- 100% Success Rate:
  - ES-L1, GEO, HEO, SSO
- 50-70% Success Rate:
  - GTO, ISS, LEO, MEO, PO
- 0% Success Rate:
  - SO



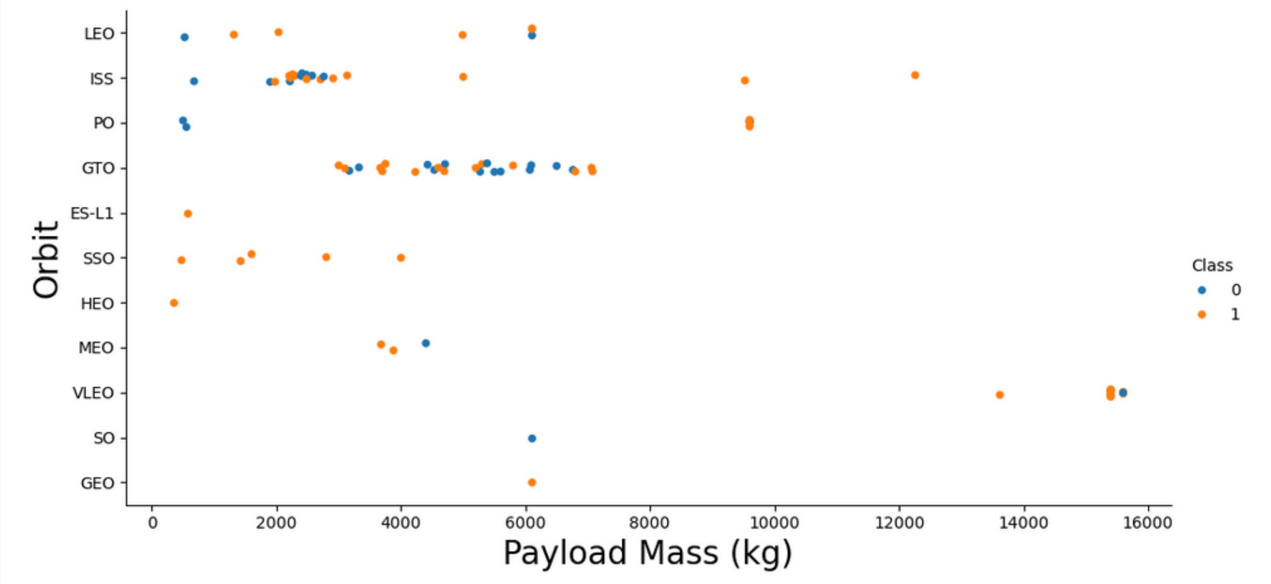
Flight Number vs. Orbit Type

- Success rate seems to increase with time
- With regards to LEO, this relationship is very accurate, but this is not so for GTO



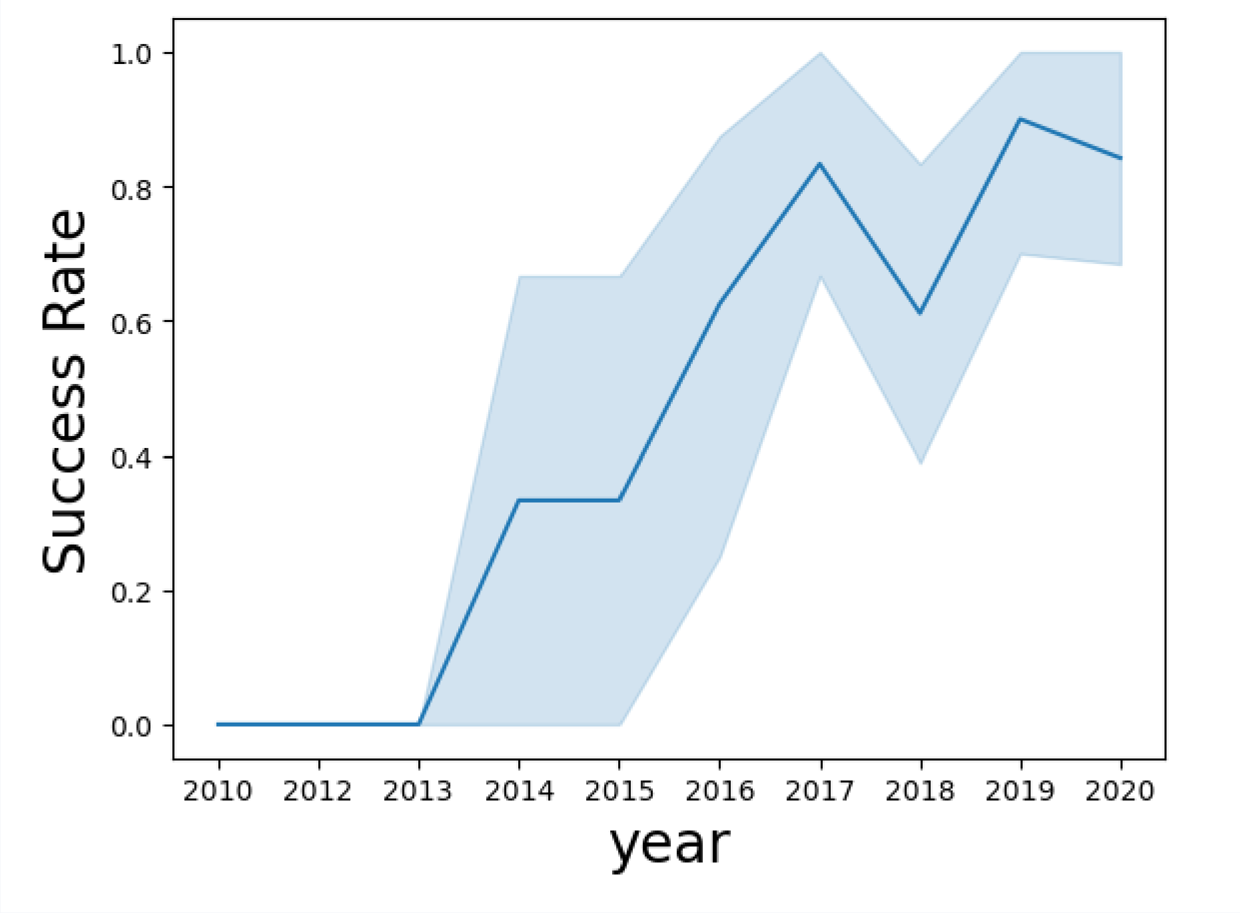
Payload vs. Orbit Type

- GTO orbit has mixed success
- SSO, HEO and E-L1 have 100% success rate
- ISS & PO have better success with higher payloads



Launch Success Yearly Trend

- Overall success rate increased over time
- We have an alarming dip in 2018 which needs to be investigated further



All Launch Site Names

- Unique Launch Sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

- Launch sites with CCA (Cape Canaveral Air Station) prefix

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

In [12]:
```
%sql select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
```

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

Out[12]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total Payload Mass for NASA has been **107 010** kg

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

```
In [13]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like '%NASA%';

 * sqlite:///my_data1.db
Done.

Out[13]:
```

sum(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is **2928.4** kg

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

In [14]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

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

Out[14]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First successful landing outcome on a ground pad

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

Hint:Use min function

In [15]:
```
%sql select min(Date), Landing_Outcome from SPACEXTBL where Landing_Outcome like '%ground pad%';
```

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

Out[15]:

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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 [16]:
```
%sql select Date, Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > '400
```

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

Out[16]:

Date	Booster_Version
2016-05-06	F9 FT B1022
2016-08-14	F9 FT B1026
2017-03-30	F9 FT B1021.2
2017-10-11	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [17]:
```
%sql select MISSION_OUTCOME, count(*) as total_number from SPACEXTBL group by MISSION_OUTCOME;
```

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

Out[17]:

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

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

In [18]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[18]:

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 of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
In [19]: %sql SELECT substr(Date,6,2) as month, date, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL where Landing_Outc

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

Out[19]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
01	2016-01-17	F9 v1.1 B1017	VAFB SLC-4E	Failure (drone ship)
03	2016-03-04	F9 FT B1020	CCAFS LC-40	Failure (drone ship)
06	2016-06-15	F9 FT B1024	CCAFS LC-40	Failure (drone ship)

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

- Ranked count of landing outcomes

In [23]:
```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

Out[23]:

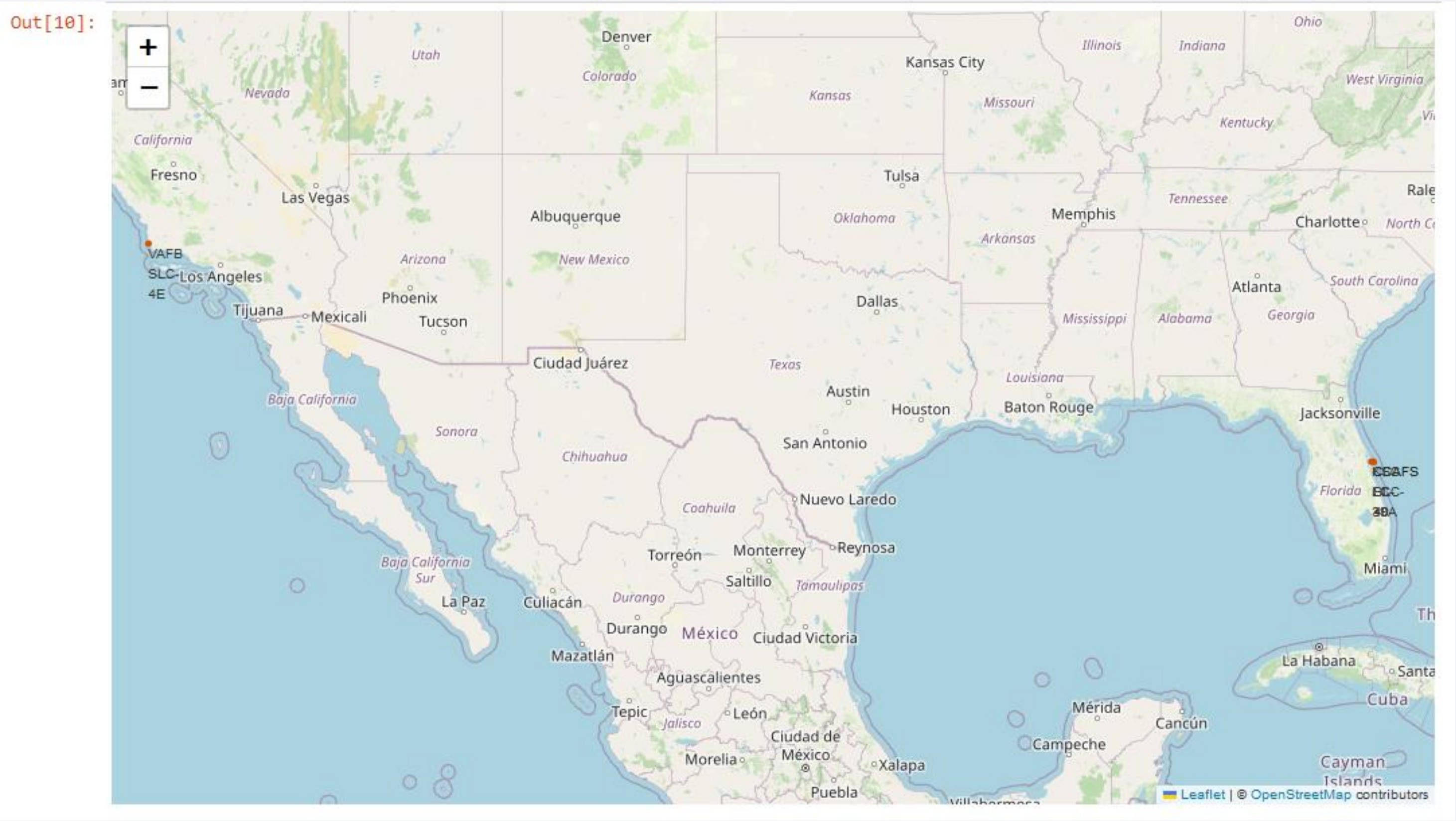
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

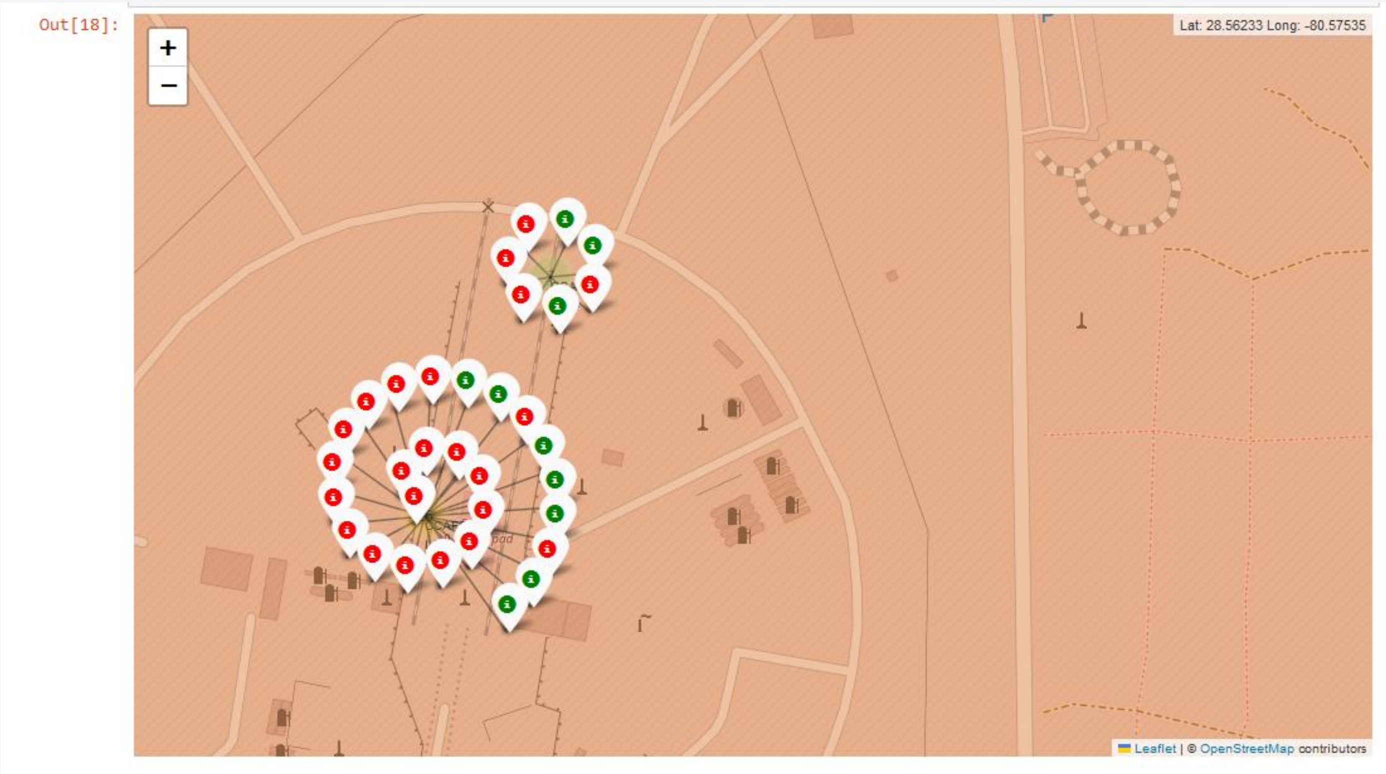
SpaceX Launch sites in USA

- Launch sites as close as possible to the equator to allow for the best orbit options
- Natural Earth rotation can assist in more fuel efficient launches



Launch outcomes at each site

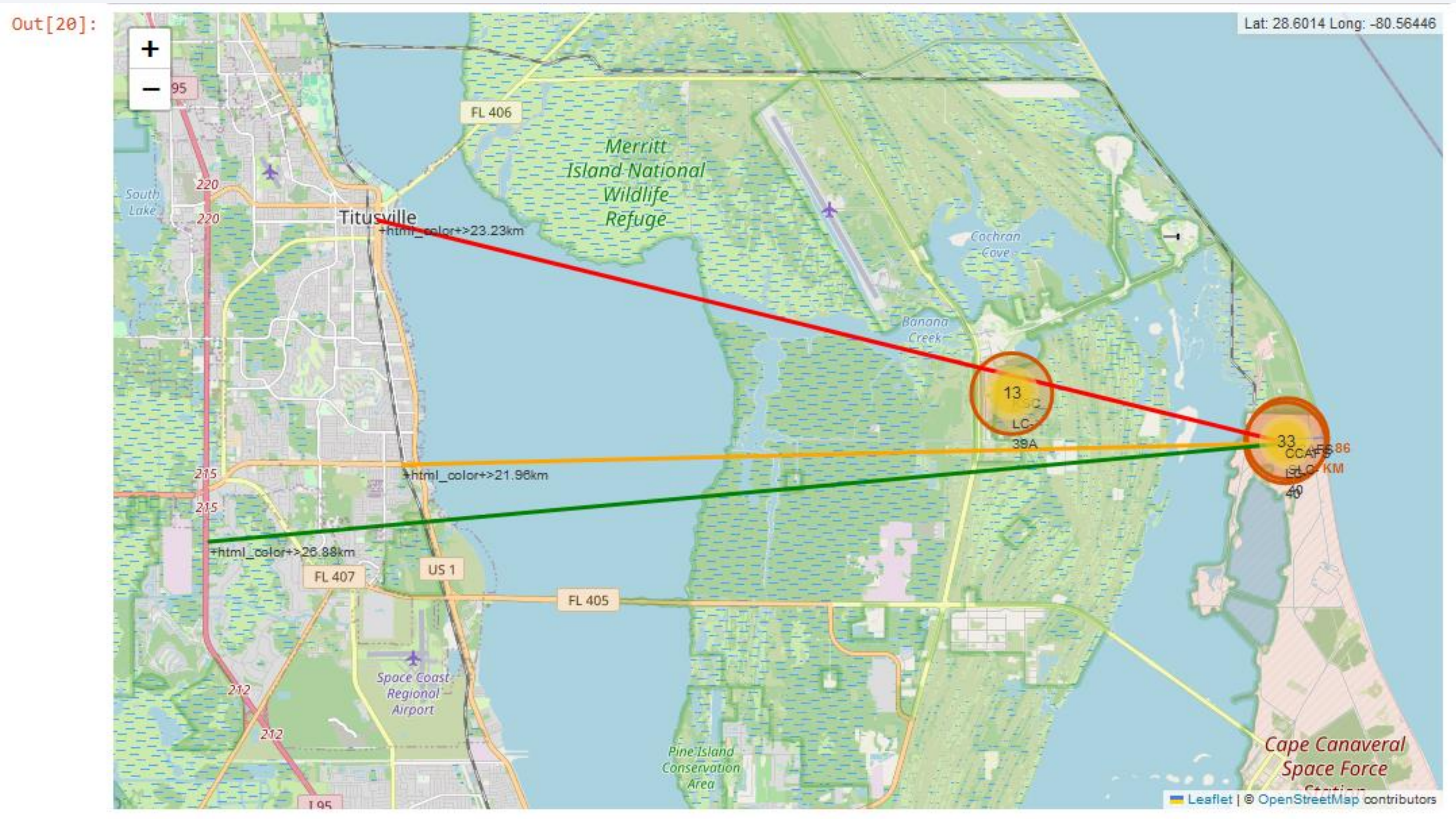
- Green markers show successful launches
- Red markers show launch failure



Nearby infrastructure

Proximities to:

- Railway: 21.96km
- Highway: 26.8km
- Coastline: 0.86km
- Town Center: 23.23km

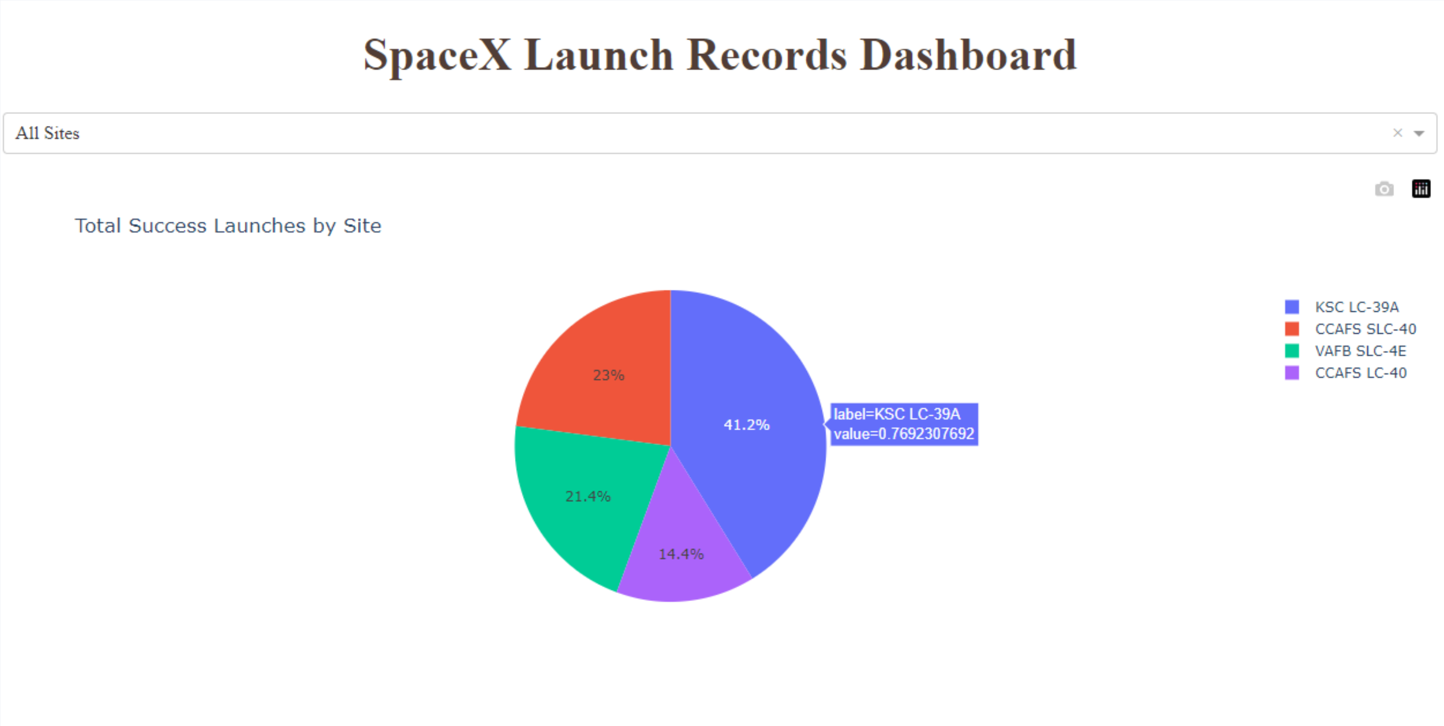


Section 4

Build a Dashboard with Plotly Dash

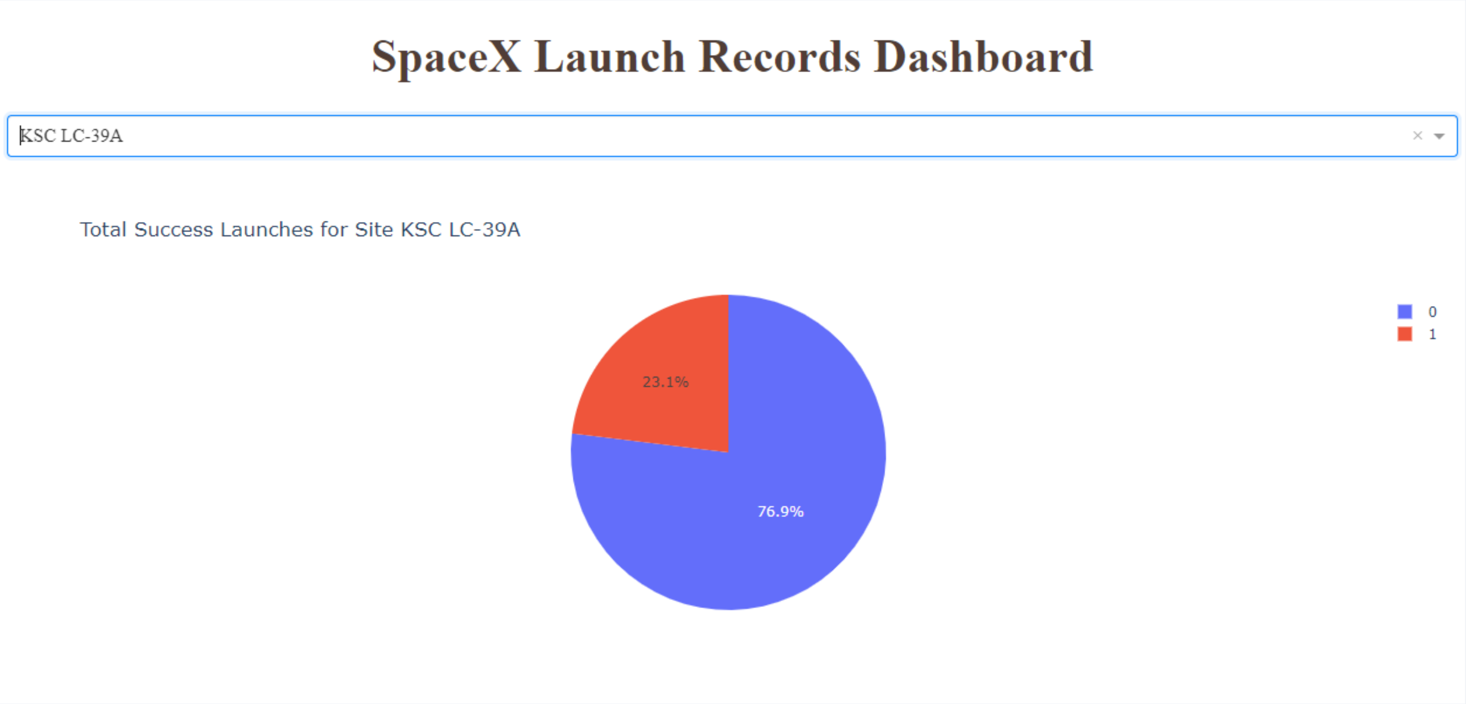
Launch Successes by Site

- KSC LC-39A is the most successful site amongst all launch sites with 41.2%



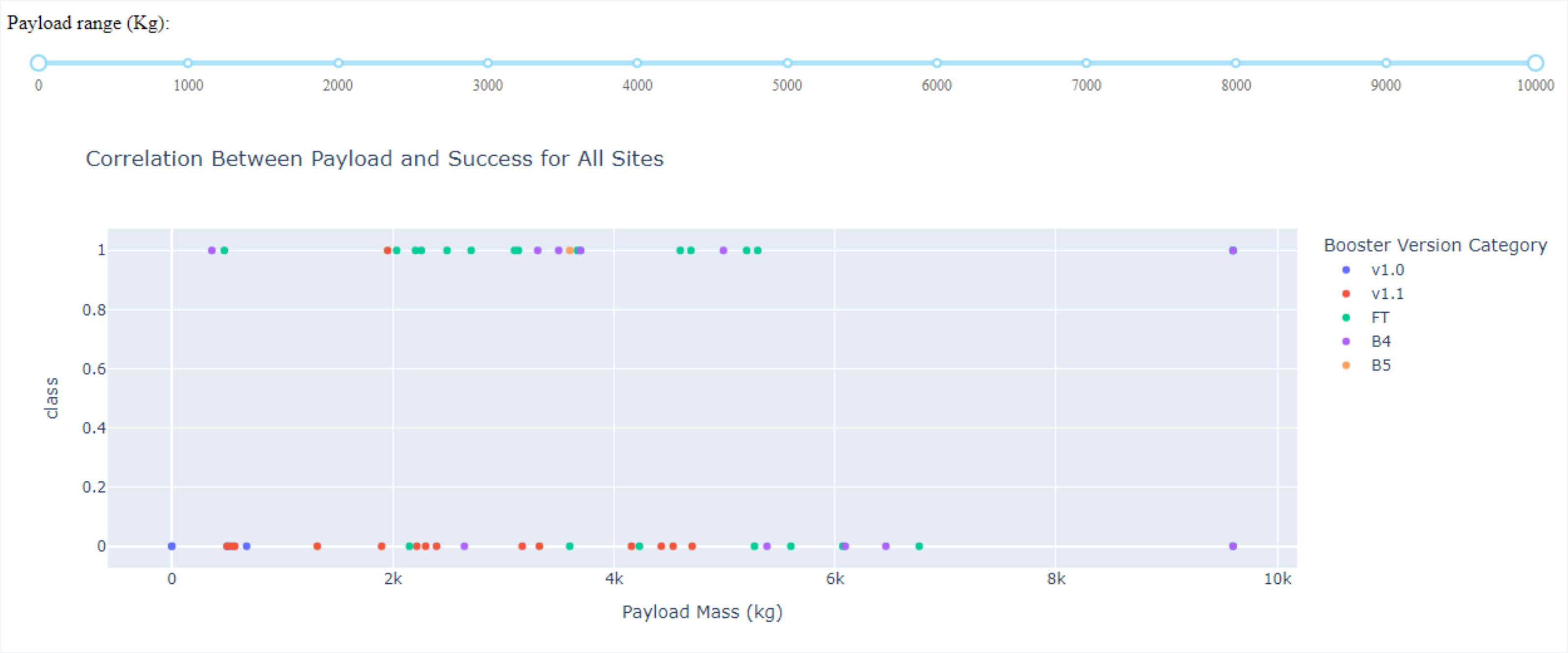
Launch Site Success Rate

- KSC LC-39A has the highest success rate amongst all launch sites at 76.9%



Success and Payload Mass

- Payloads between 2000kg and 5000kg:
  - 15 successful and 13 failed launches which results in a 53.57% success rate

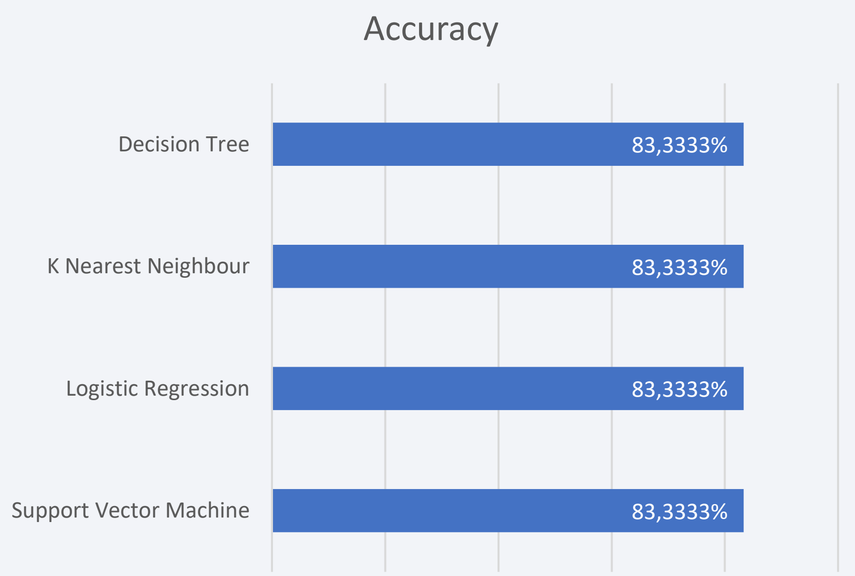


Section 5

Predictive Analysis (Classification)

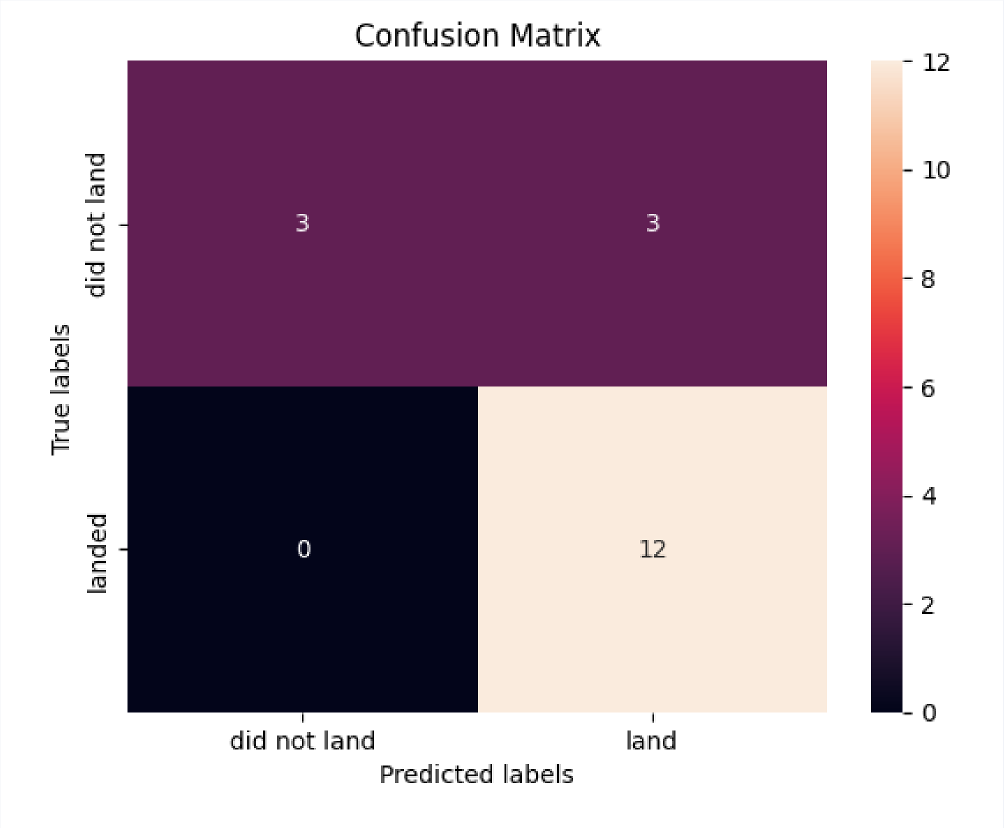
Classification Accuracy

- All models performed the same across all metrics. This may be due to a very small data set



Confusion Matrix

- 12 True Positives
- 3 True Negatives
- 3 False Positives
- 0 False Negatives



Conclusions

- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate
- Coast: All the launch sites are close to the coast for safe abort options
- Model Performance: All the models performed similarly

Considerations:

- A larger dataset will help build on the predictive analytics model

Thank you!