



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marga Vancells
11th March 2022



Outline

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

Executive Summary

- Summary of methodologies
 - Data Collection API
 - Data Collection with Web Scraping
 - Data Wrangling
 - Exploratory data analysis (EDA) with SQL
 - Exploratory data analysis (EDA) with Visualization
 - Data Visualization with Folium
 - Interactive Dashboard with Plotly Dash
 - Predictive Analysis (Classification)
- Summary of all results
 - Data analysis results
 - Predictive Analysis results

Introduction

- Project background and context

SpaceX 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 SpaceX can reuse the first stage.

Therefore 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 SpaceX for a rocket launch.

The goal of the projects is to predict if the Falcon 9 first stage will land successfully.

- Problems we want to find answers

- Which features influence a successful landing?
- Which is the best model to predict a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - Requests to the SpaceX API
 - Web scraping from a Wikipedia page
- Perform data wrangling
 - Create a classification variable from a categorical column
- 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

Data Collection

- **Request** rocket launch **data** from
 - SpaceX API (via get request)
 - Wikipedia webpage (via webscrapping)
- Decode the response content and turn it into a **Pandas** dataframe
- **Filter** out not needed launches
- **Export** to a CSV file

Data Collection – SpaceX API

Request data from SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

Decode the response into a Pandas dataframe

```
data = pd.json_normalize(response.json())
```

Filter the dataframe

```
data_falcon9 = data_falcon9[data_falcon9['BoosterVersion']!='Falcon 1']
```

Deal with missing values

```
plm_mean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
# data_falcon9['PayloadMass'].fillna(value=plm_mean, inplace=True)
data_falcon9['PayloadMass'] = df['PayloadMass'].replace(np.nan, plm_mean)
```

Export to CSV

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```


Data Collection - Scraping

Request the Falcon9 Launch Wikipedia page

```
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)
```

Create a BeautifulSoup object from the HTML response

```
soup = BeautifulSoup(response.text)
```

Find all tables

```
html_tables = soup.find_all('table')
```

Get column names

```
ths = first_launch_table.find_all('th')
for th in ths:
    cn = extract_column_from_header(th)
    if type(cn) is str and cn != 'None' and len(cn) > 0 : column_names.append(cn)
```

Create a data frame by parsing the HTML tables

Export to CSV

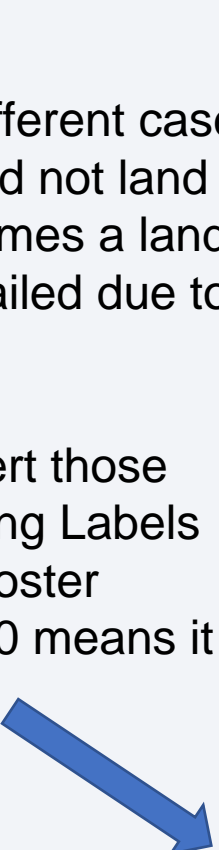
```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

True	ASDS	41
None	None	19
True	RTLS	14
False	ASDS	6
True	Ocean	5
None	ASDS	2
False	Ocean	2
False	RTLS	1

There are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident.

We will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

Calculate the number of launches for each site.

Calculate the number and occurrence of each orbit

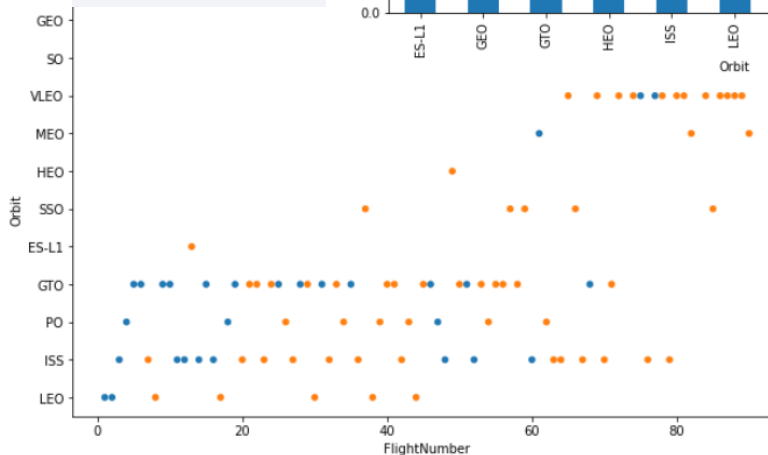
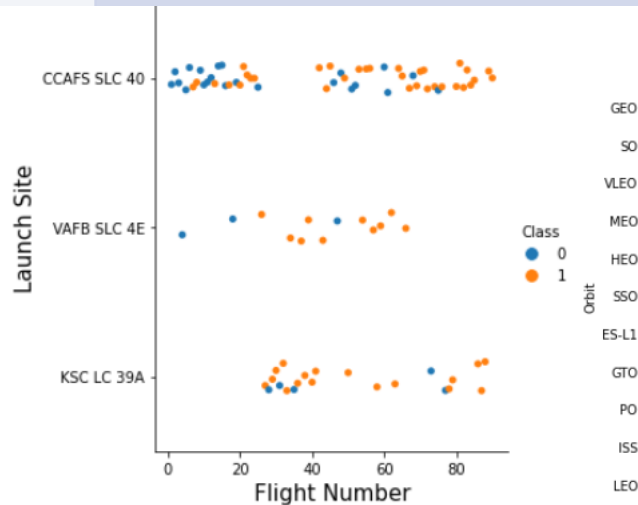
Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column.

EDA with Data Visualization

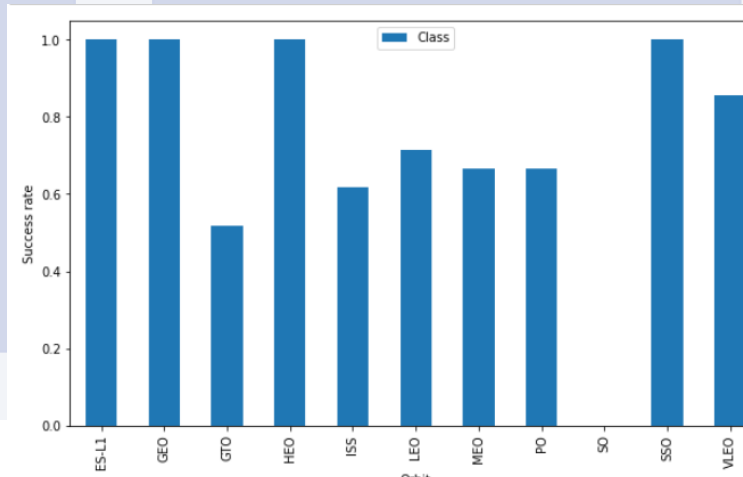
Scatter charts

- FlightNumber vs PayloadMass
- FlightNumber vs LaunchSite
- Payload vs Launch Site
- FlightNumber vs Orbit type
- Payload vs Orbit type



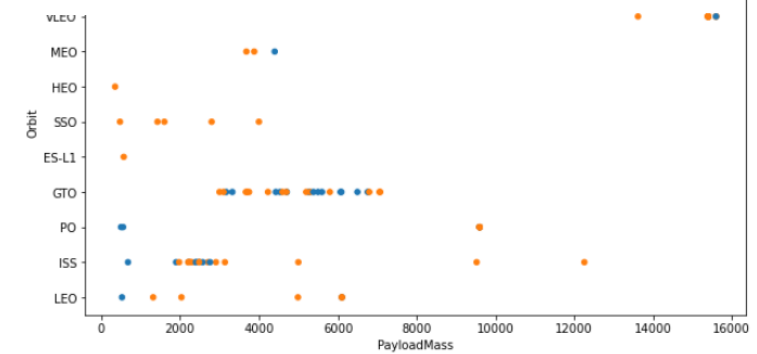
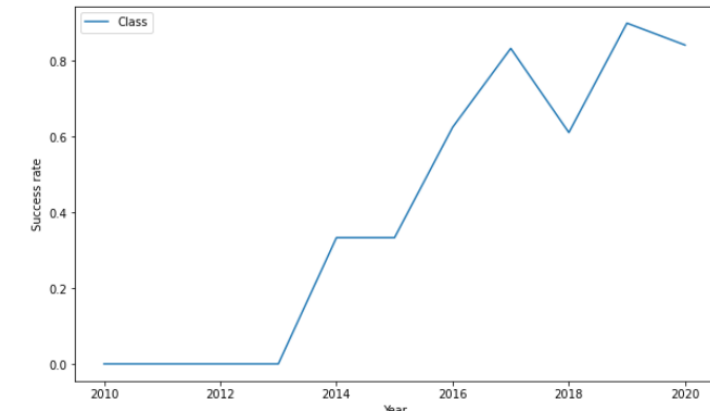
Bar charts

- success rate of each orbit



Line charts

- success yearly trend



EDA with SQL

```
%sql select landing__outcome, count(*) as num_lands from RGC73063.SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by num_lands desc
```

- Store the dataset in database table
- Write and execute SQL queries to do EDA:
 - Display the names of the unique launch sites in the space mission
 - Display 5 records where launch sites begin with the string 'CCA'
 - Display the total payload mass carried by boosters launched by NASA (CRS)
 - Display average payload mass carried by booster version F9 v1.1
 - List the date when the first successful landing outcome in ground pad was achieved
 - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
 - List the total number of successful and failure mission outcomes
 - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
 - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
 - 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

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

Build an Interactive Map with Folium

The launch success may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.

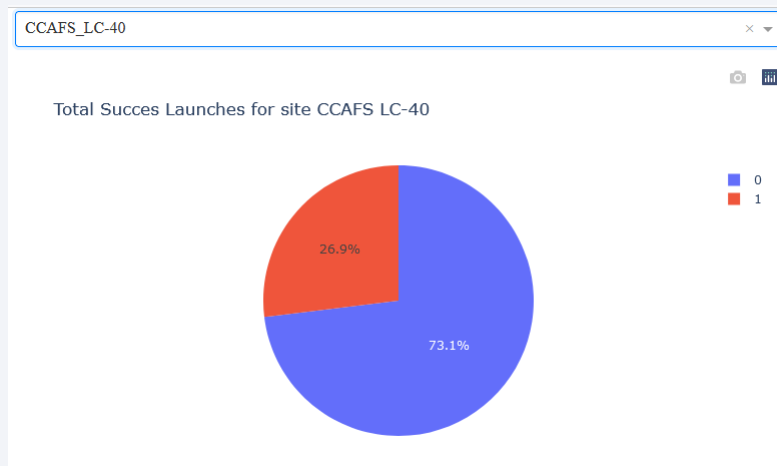
Finding an optimal location for building a launch site certainly involves many factors, and hopefully, we could discover some of the factors by analyzing the existing launch site locations.

Map Object	Use
Circle	To add a highlighted circle area with a text label on a specific coordinate
Marker	To make a mark on the map
Marker Cluster	To simplify the map containing many markers having the same coordinate
Mouse Position	To obtained a coordinate of a point
Poly Line	To draw a line between a launch site to the selected coastline point

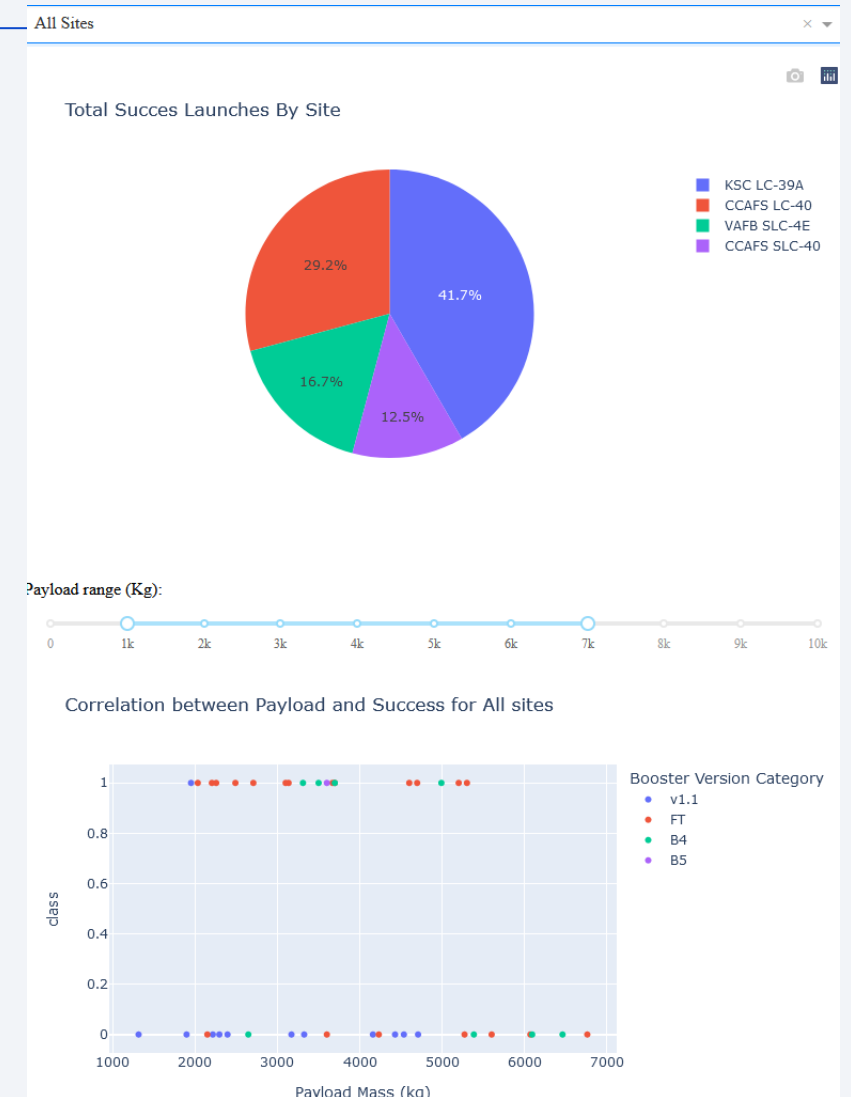
Build a Dashboard with Plotly Dash

We build a dashboard with these objects

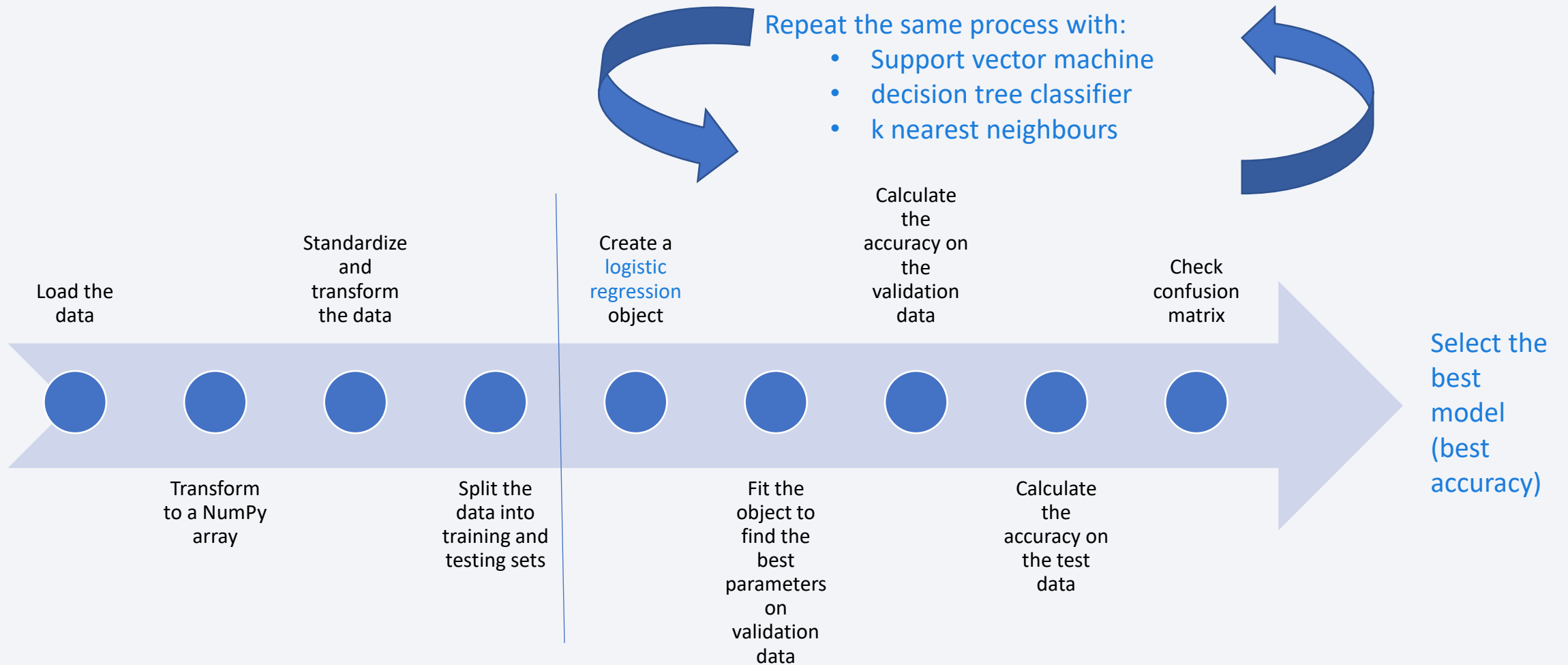
- A launch site drop-down list
- A success pie chart based on the selected site drop-down
- A range slider to select Payload
- A success payload scatter plot based on the selected site drop-down and the selected Payload range



With the dashboard we could analyze SpaceX launch data and answer some questions



Predictive Analysis (Classification)



Results

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

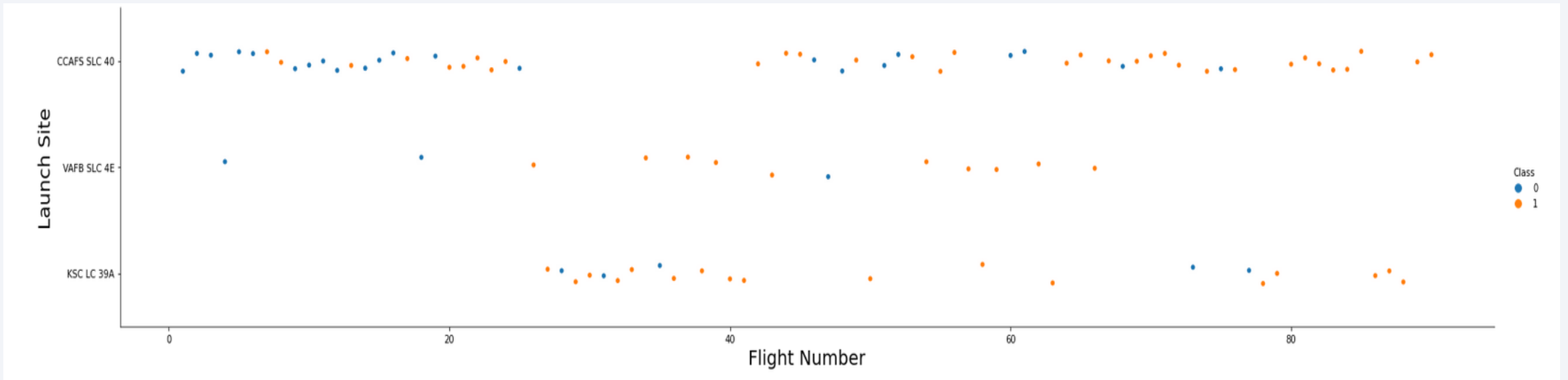
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

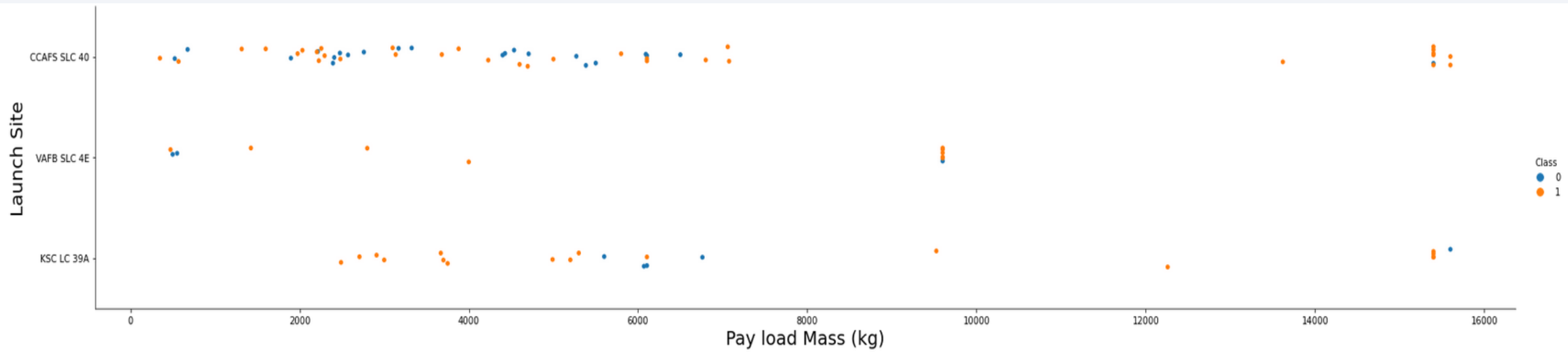
With more amount of flights, the success rate increases for each site



Payload vs. Launch Site

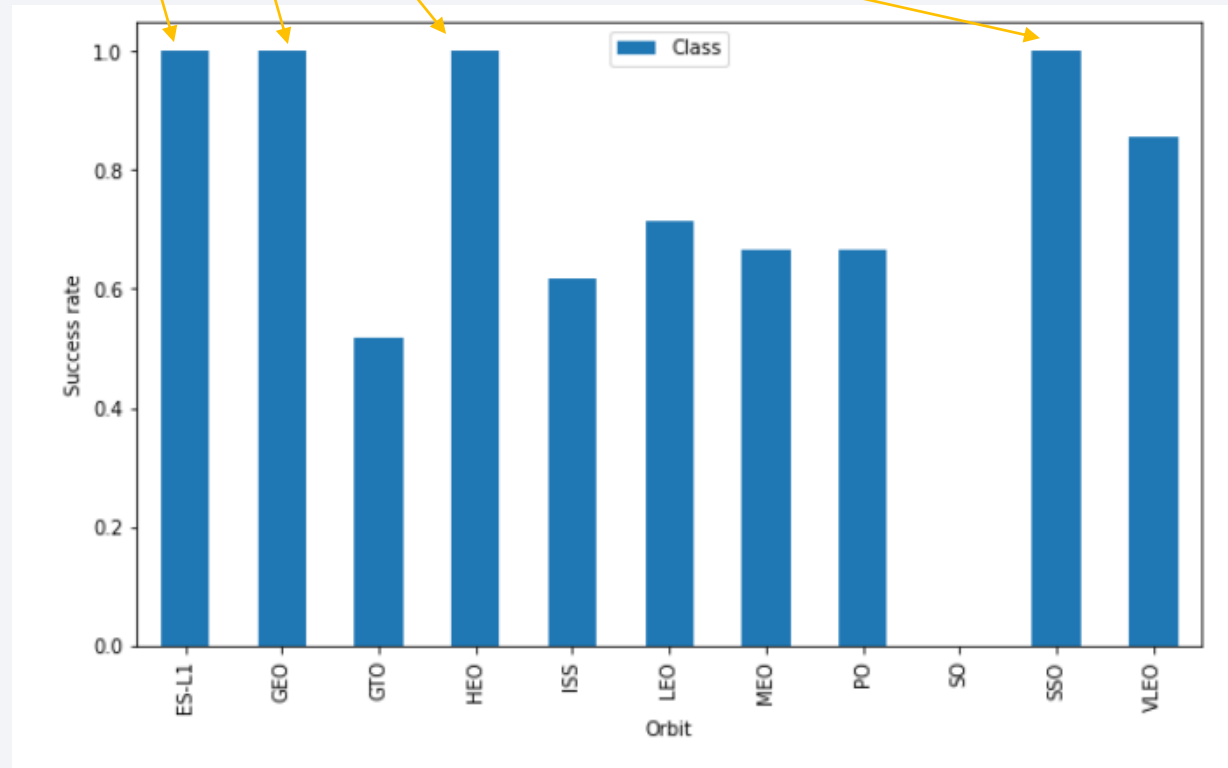
For the VAFB-SLC-4E launch site, there are no rockets launched for heavy payload mass (greater than 10000)

For CCAFS-SLC-40, the greater the payload mass, the higher the success rate



Success Rate vs. Orbit Type

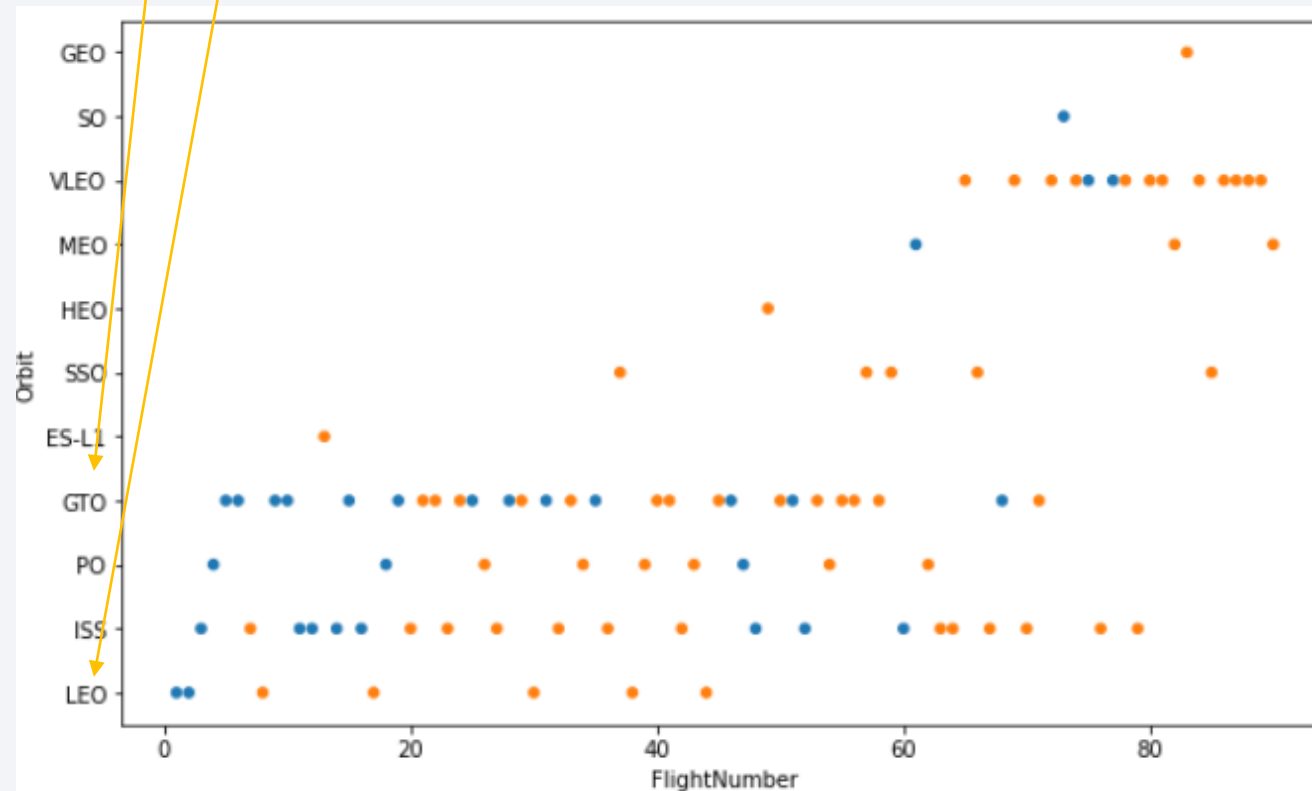
ES-L1, GEO, HEO and SSO orbits have the best success rates



Flight Number vs. Orbit Type

In the LEO orbit, the Success appears related to the number of flights

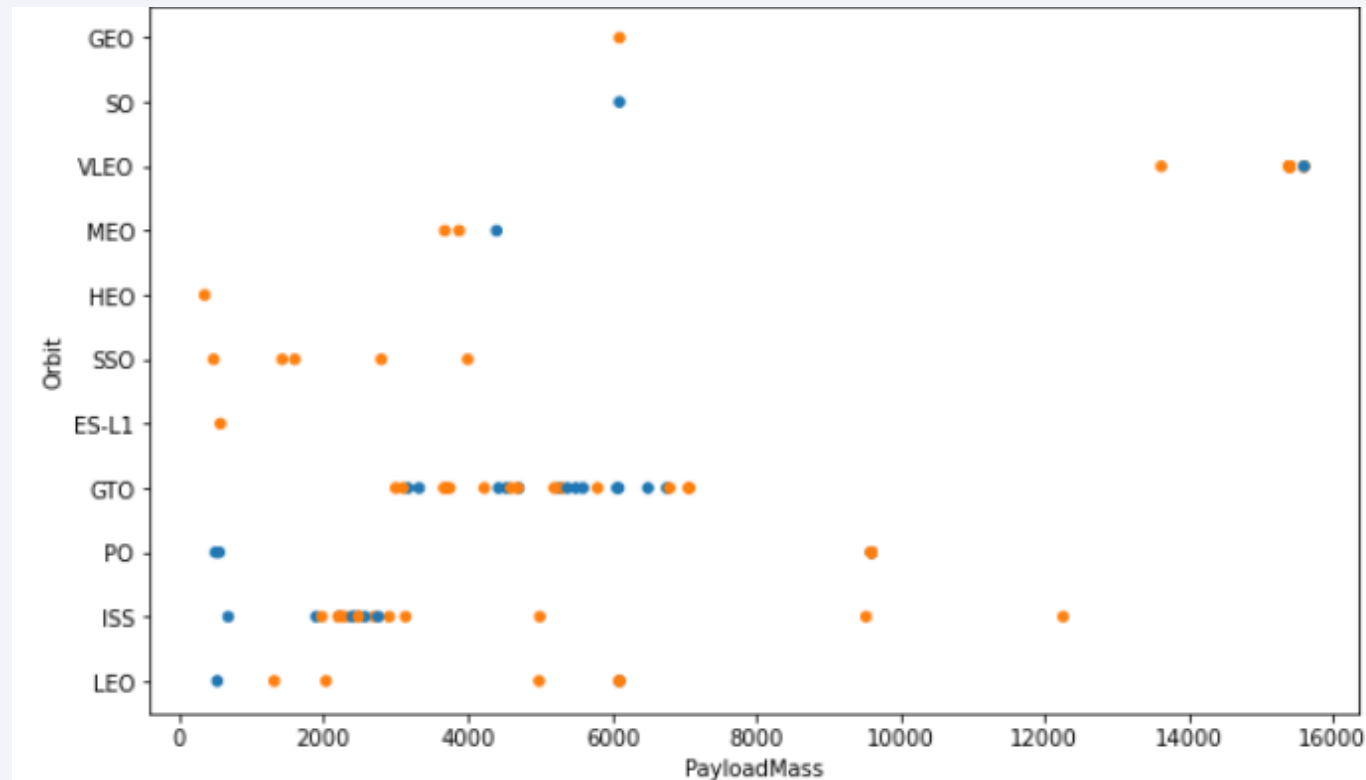
In GTO orbit, there seems to be no relationship between flight number



Payload vs. Orbit Type

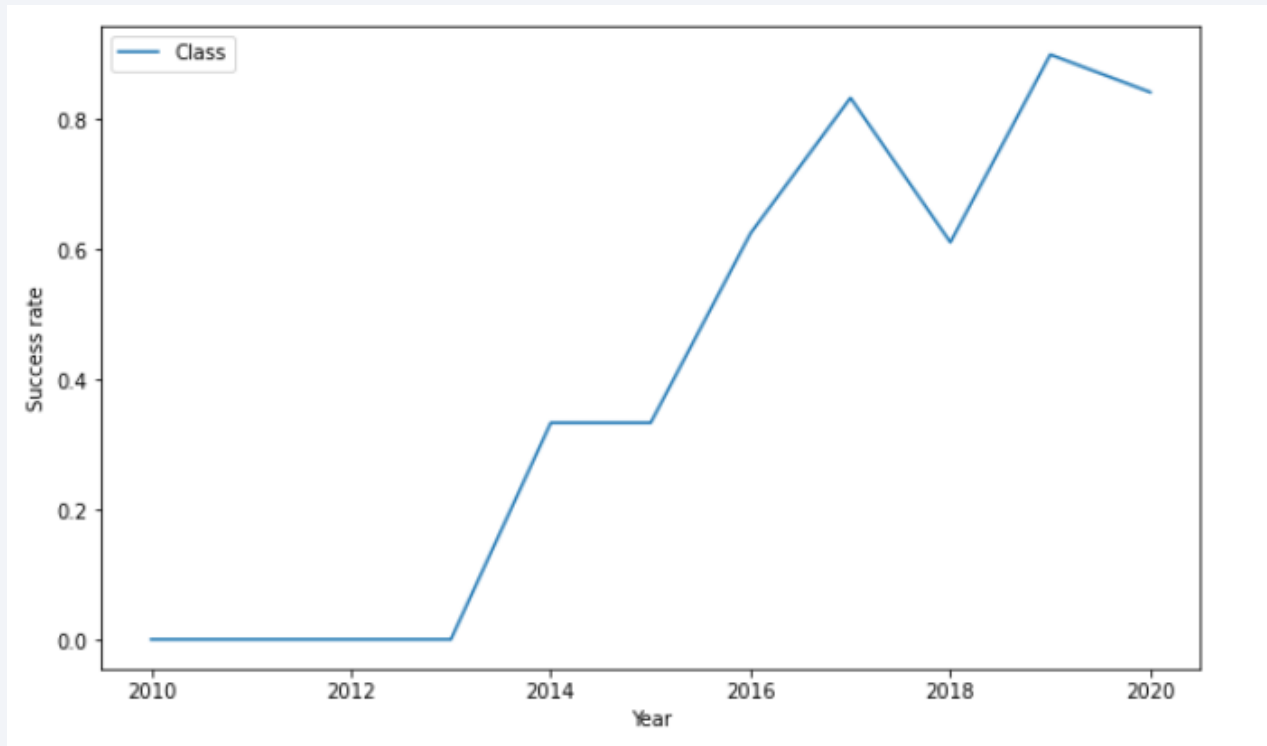
With heavy payloads, the successful landing or positive landing rate is more for Polar, LEO and ISS

For GTO, we cannot distinguish this well as both positive landing rate and negative landing



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

```
%sql select distinct Launch_Site from RGC73063.SPACEXTBL
```

Using the word **DISTINCT**
we get only unique values

launch_site

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%sql select * from RGC73063.SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- With **LIKE** and **CCA%**, we get the ones that begin with CCA
- With **LIMIT 5**, we get only 5 records

Total Payload Mass

```
%sql select sum(payload_mass__kg_) total_payload_mass from RGC73063.SPACEXTBL where customer='NASA (CRS) '
```

- With the **SUM**, function we calculate the total payload mass
- With the filter in the **WHERE** clause, we calculated only over 'NASA (CRS)' customers

total_payload_mass

45596

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) avg_payload_mass from RGC73063.SPACEXTBL where booster_version='F9 v1.1'
```

- With the **AVG** function, we calculate the average payload mass
- With the filter in the **WHERE** clause, we calculate only over the booster version F9 v1.1

avg_payload_mass
2928

First Successful Ground Landing Date

```
%sql select min(DATE) first_successful_landing from RGC73063.SPACEXTBL where landing__outcome ='Success (ground pad) '
```

- With the **MIN** function, we get the first date
- With the filter in the **WHERE** clause, we get it only over successful landing outcome on ground pad

first_successful_landing

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from RGC73063.SPACEXTBL where landing__outcome ='Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

- With the filter in the **WHERE** clause, we get only the boosters that have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000

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 sum(case when mission_outcome like 'Success%' then 1 else 0 end) Successful_Mission,\n            sum(case when mission_outcome like 'Failure%' then 1 else 0 end) Failure_Mission \nfrom RGC73063.SPACEXTBL;
```

- For successful_mission, with the **CASE** clause, we indicate 1 if outcome begins with 'Success' and 0 otherwise, and then we **SUM** it
- For failure_mission, we do the same but with the outcome beginning with 'Failure'

successful_mission	failure_mission
100	1

Boosters Carried Maximum Payload

```
%sql select distinct booster_version \
      from RGC73063.SPACEXTBL \
      where payload_mass__kg_=(select max(payload_mass__kg_) from RGC73063.SPACEXTBL )
```

- First, we get the maximum payload mass with the **MAX** function
- Then we filter the query with the max value returned with the **subquery**

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

```
%sql select DATE,booster_version, launch_site \  
      from RGC73063.SPACEXTBL \  
      where landing__outcome ='Failure (drone ship)' and EXTRACT(YEAR FROM DATE)=2015
```

We get only 2015 with the `EXTRACT(YEAR FROM [date])` function

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql select landing__outcome, count(*) as num_lands \
      from RGC73063.SPACEXTBL \
      where DATE between '2010-06-04' and '2017-03-20' \
      group by landing__outcome \
      order by num_lands desc
```

- With the **COUNT**, function we get the number of records
- With **GROUP BY**, we apply the count function to each landing outcome
- With **ORDER BY DESC**, we list them in descending order

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

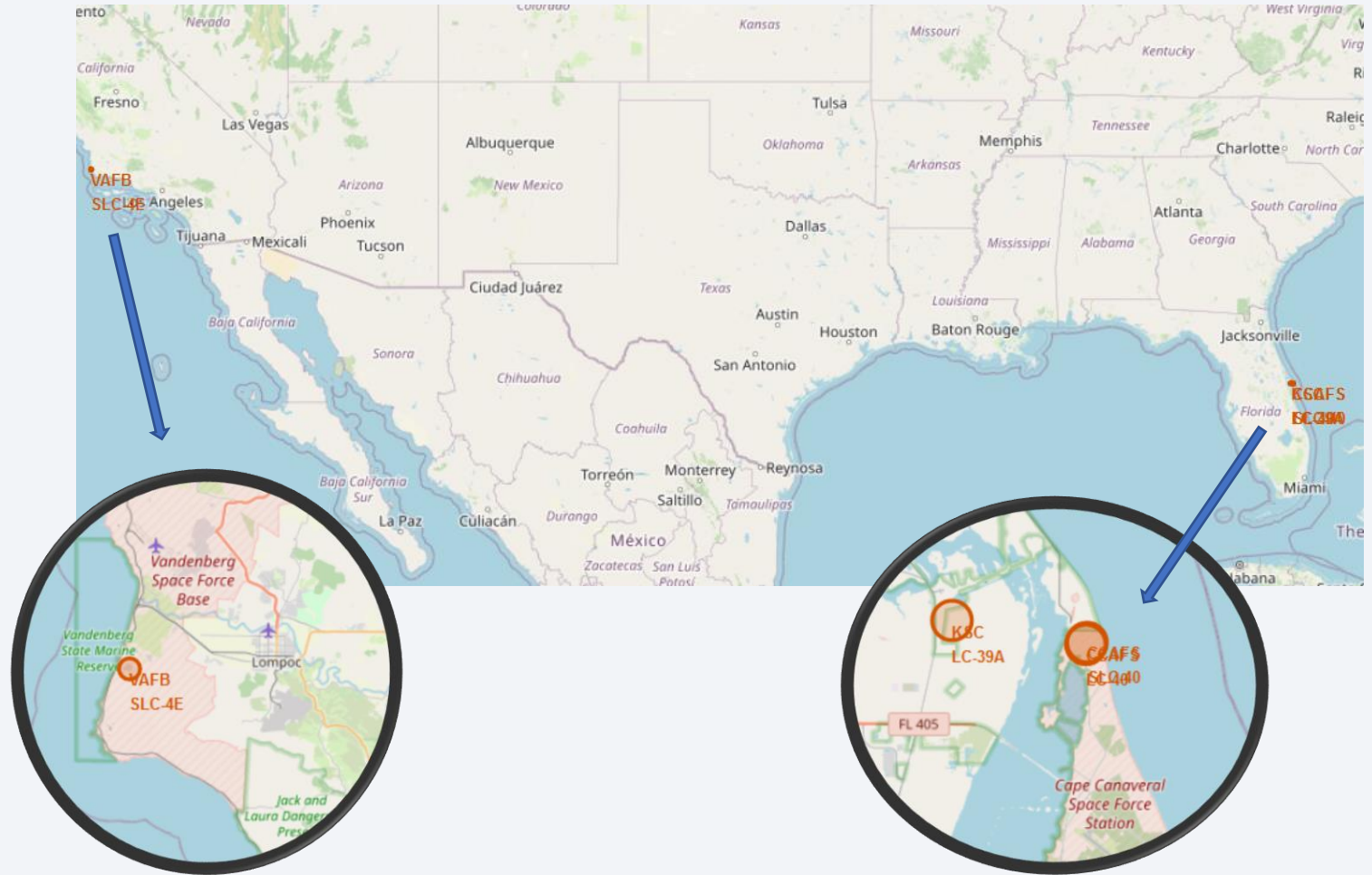
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

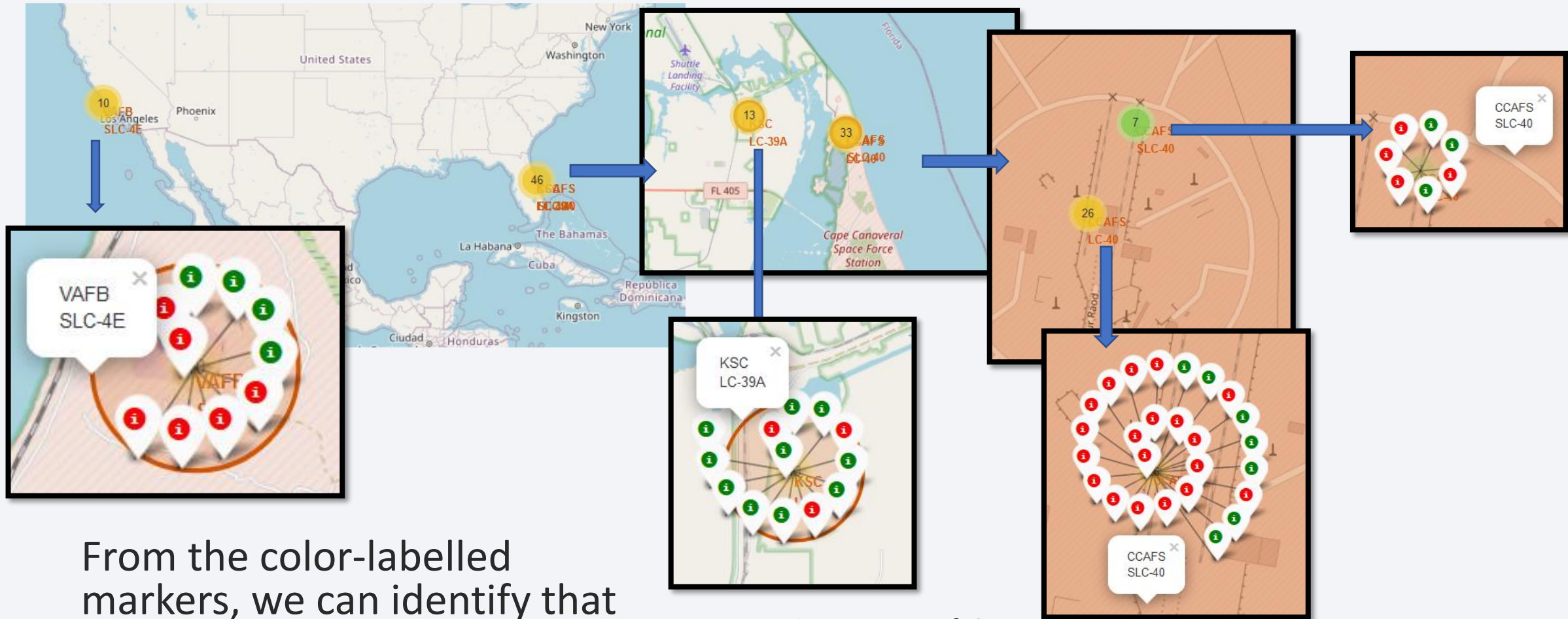
Launch Sites Proximities Analysis

All launch sites on a map

- All launch sites are in very close proximity to the coast
- All launch sites are in proximity to the Equator line



Color-labeled launch on the map

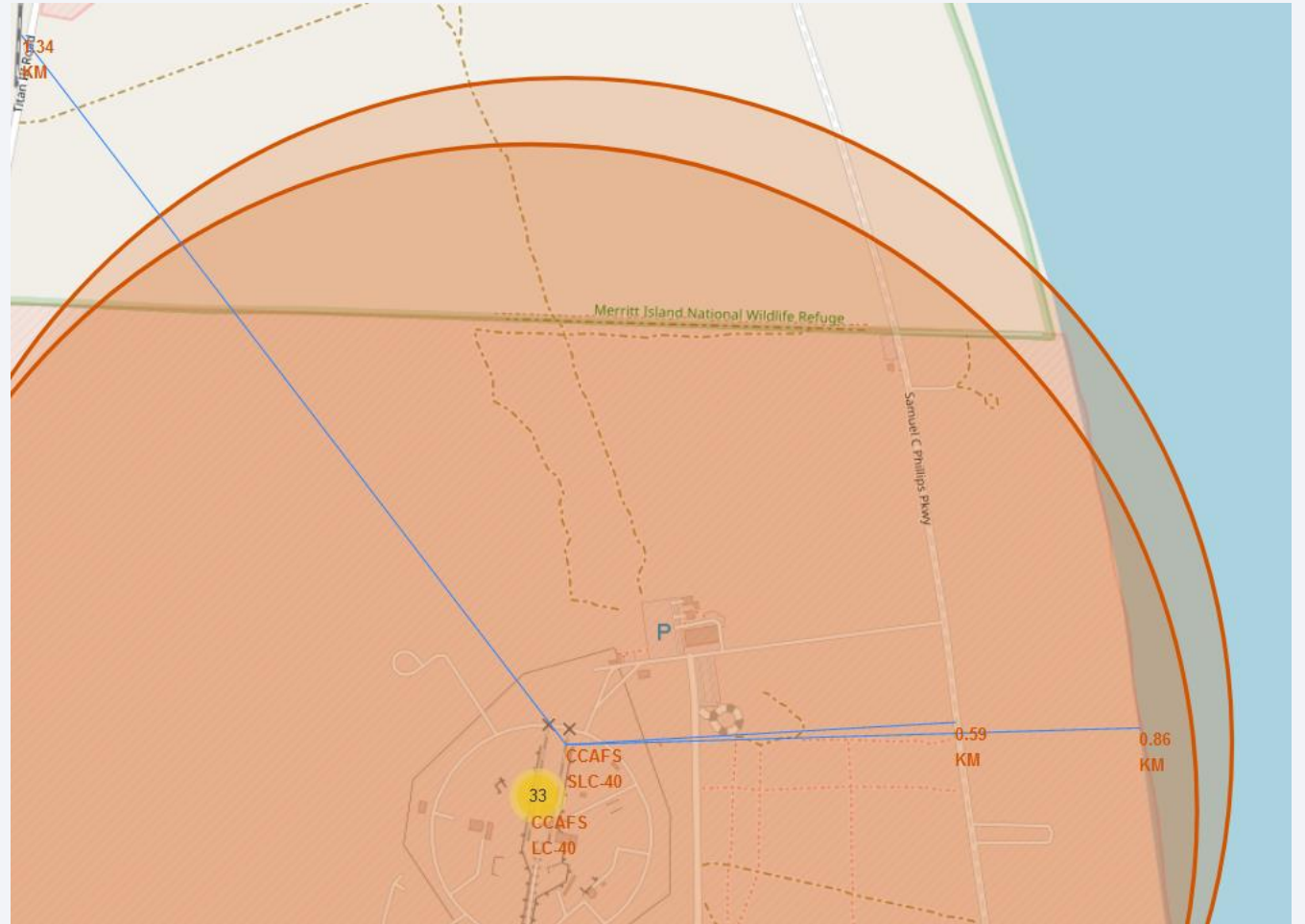


From the color-labelled markers, we can identify that launch site KSC LC-39A has relatively high success rates

Green marker: Successful
Red marker: Failure

Launch Site distances

Launch Site CCAFS SLC-40 is not far away from the railway, highway and coastline



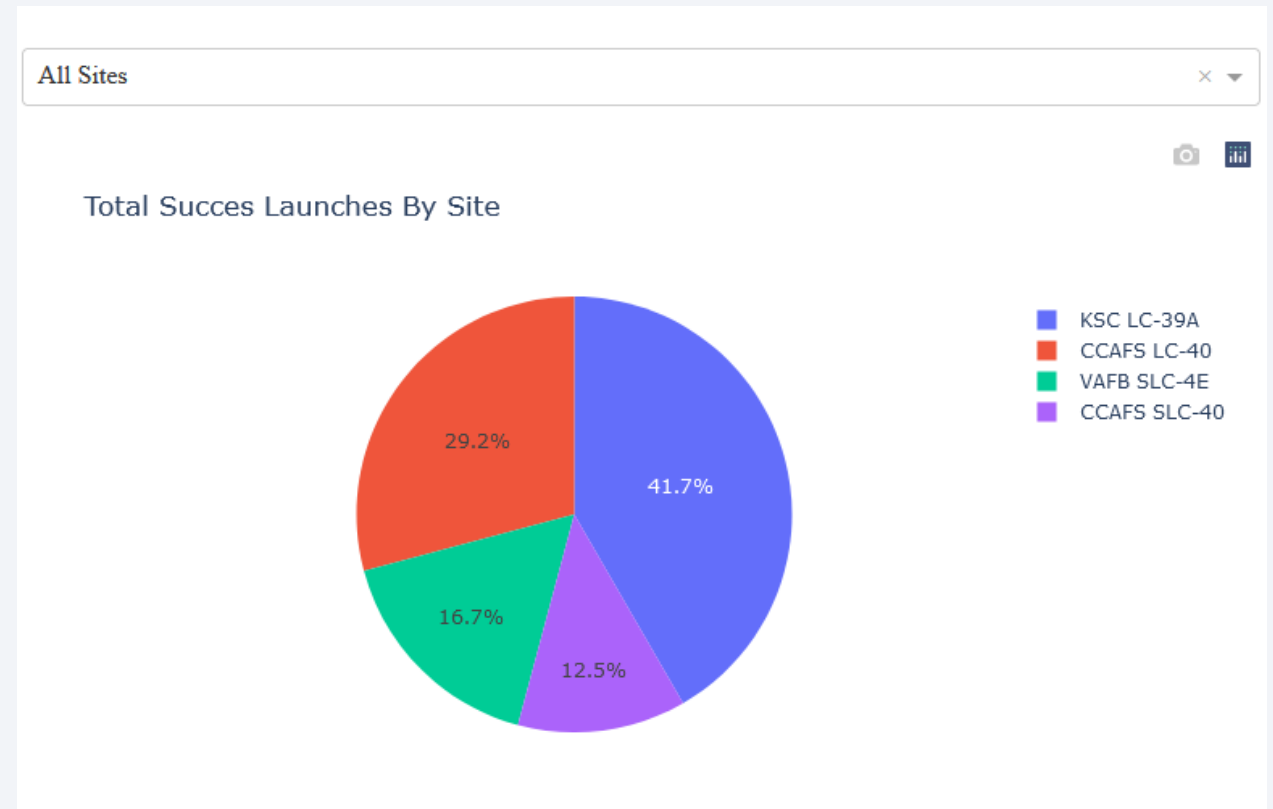


Section 4

Build a Dashboard with Plotly Dash

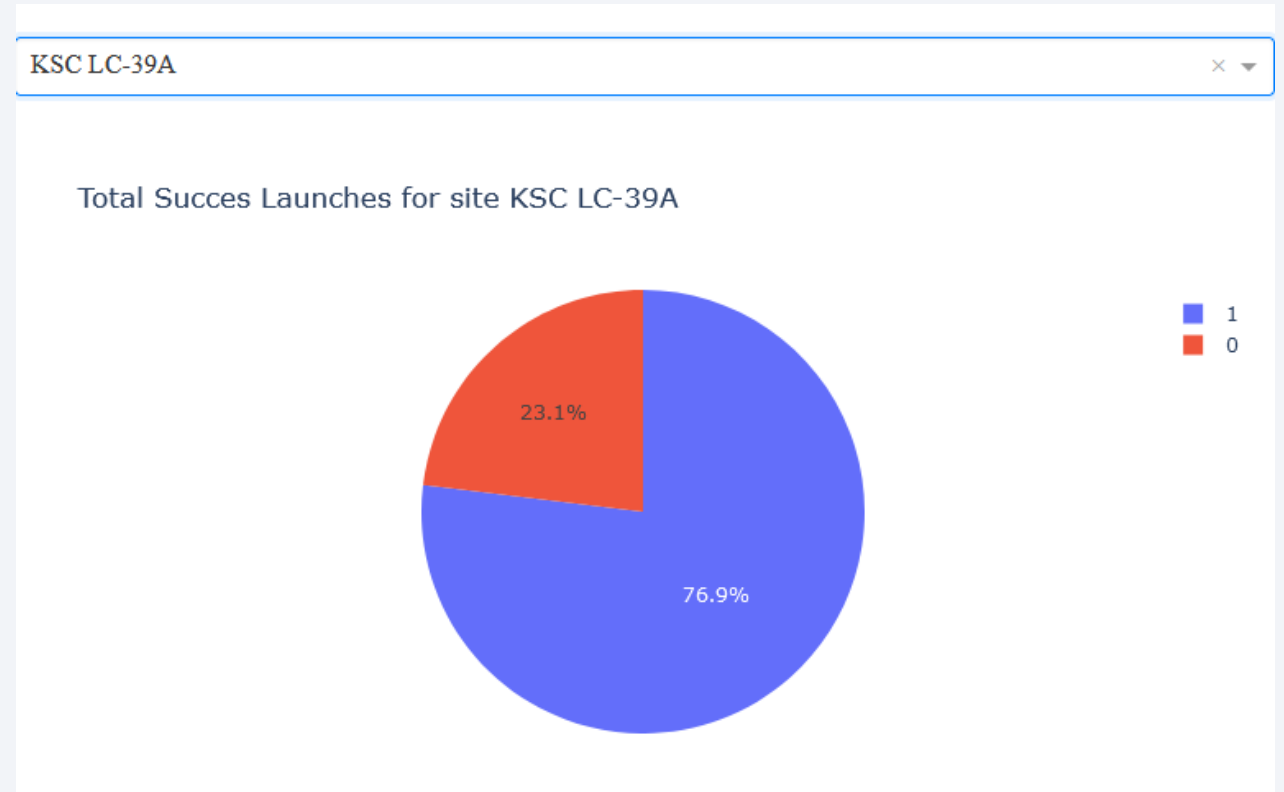
Total Success Launches by Site

KSC LC-39A launch site has the most success launch rate

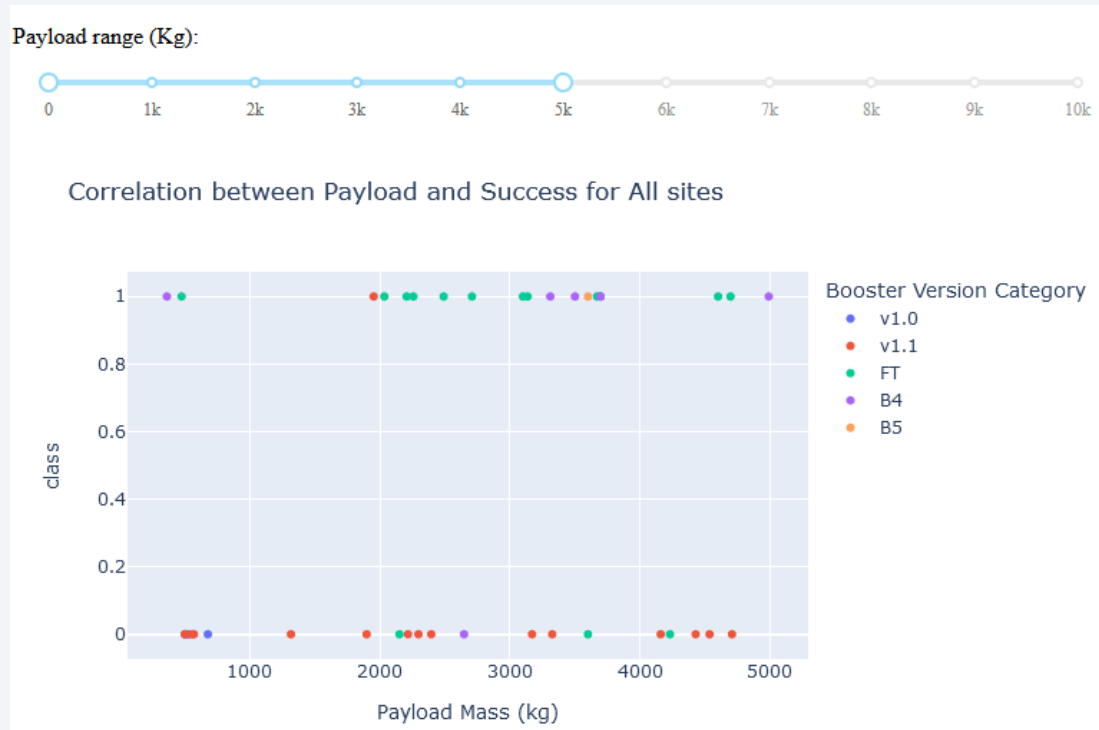


Launch site with highest launch success ratio

KSC LC-39A has a 76,9% success rate and a 23.1% failure rate



Payload vs. Launch Outcome plot for all sites



The success rate is higher for low weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

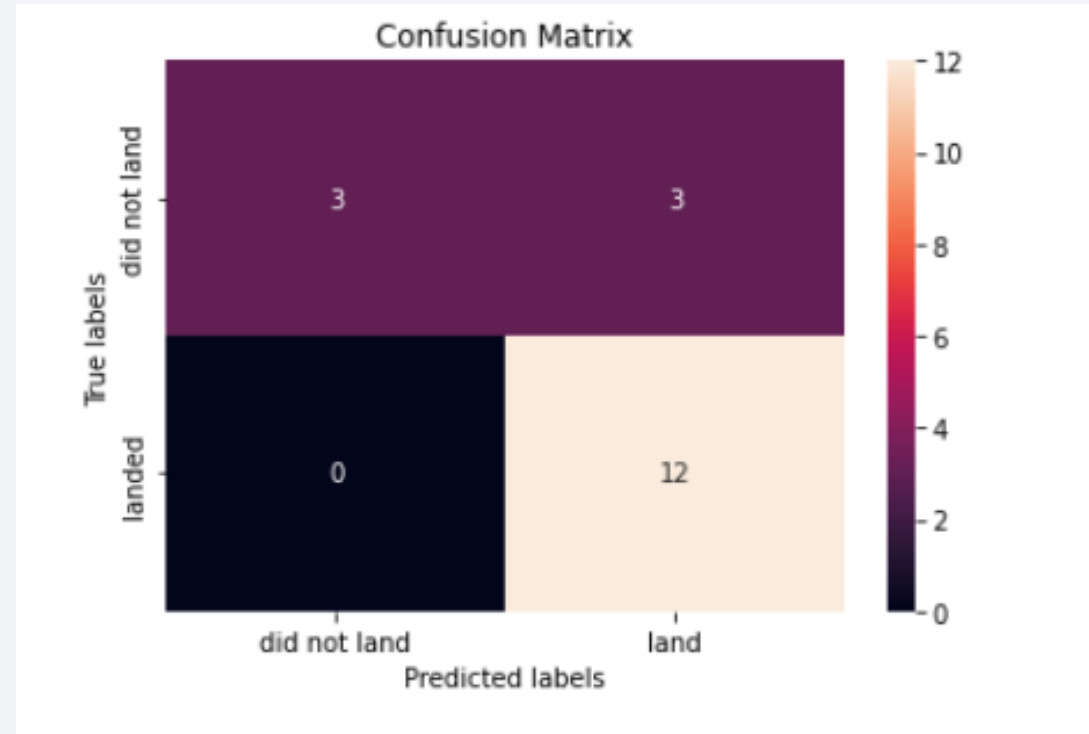
Algorithm	Accuracy	Accuracy on test data
Logistic Regression	0.8464285714285713	0.8333333333333334
Support Vector Machine	0.8482142857142856	0.8333333333333334
Decision Tree classifier	0.8875	0.8333333333333334
K nearest neighbors	0.8482142857142858	0.8333333333333334

The decision tree classifier is the model with the highest classification accuracy, with these parameters:

```
tuned hpyerparameters :(best parameters) {'criterion': 'gini',  
'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 2,  
'min_samples_split': 5, 'splitter': 'best'}  
accuracy : 0.8875
```

Confusion Matrix

The decision tree classifier can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- With more amount of flights, the success rate increases for each site
- ES-L1, GEO, HEO and SSO orbits have the best success rates
- The success rate has been increasing since 2013
- KSC LC-39A has the most success launch rate
- The success rate is higher for low weighted payloads
- The decision tree classifier is the model with the highest classification accuracy

Thank you!

