IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Pablo Vicente González
08/04/2024

Outline

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

Executive Summary

Summary of methodologies

- We extracted the data by web scrapping both its wiki tables("List of Falcon 9 and Falcon Heavy launches") and its API.
- Once the data was parsed and completed, we converted success labels into binary labels so as the data was accommodated for exploratory data analysis.
- EDA query and visualization proved, through statistically reliable data, some trends that will be exposed later on the presentation.
- We trained and discussed machine learning models to face the predictability of launch success.

Summary of all results

- Success rate increase can be attributed to SpaceX's know-how development: orbit election, payload mass used, launch site selected, ….
- Models are suitable for prediction, although false positive bias is present.

Introduction

As SpaceX Falcon 9 rockets have proved trusty and relatively cheap in comparison to its market counterparts, we want to study every launch data we have access to so as we understand better the way they have been developed and consequently gather insights about their success.

In this project we are putting weight into questions about which variables contribute more to launch success, how those variables were developed over time, which decisions were made that contributed to the growing reliability of the Flacon 9 rocket. Overall, our job here is to work around the data in order to predict launch outcomes based on the data available.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected using Pandas, Requests and BeautifulSoup packages's usual methods to collect, parse and tabulate the data.

- Perform data wrangling

  Missing data was assessed as the mean of the variable in quantitative cases.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  General: standarized our independent variables and splited our data into test and training data.

  Per model: fitted the model with training data and tuned it with grid search so as we got the best predictive parameters.

Data Collection

SpaceX REST API

https://api.spacexdata.com/v4/launches/past
https://api.spacexdata.com/v4/rockets/
https://api.spacexdata.com/v4/launchpads/
https://api.spacexdata.com/v4/payloads/
https://api.spacexdata.com/v4/cores/

Requests + Pandas

"List of Falcon 9 and Falcon Heavy launches" -Wikipedia

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

BeautifulSoup + Pandas

FINAL DATA

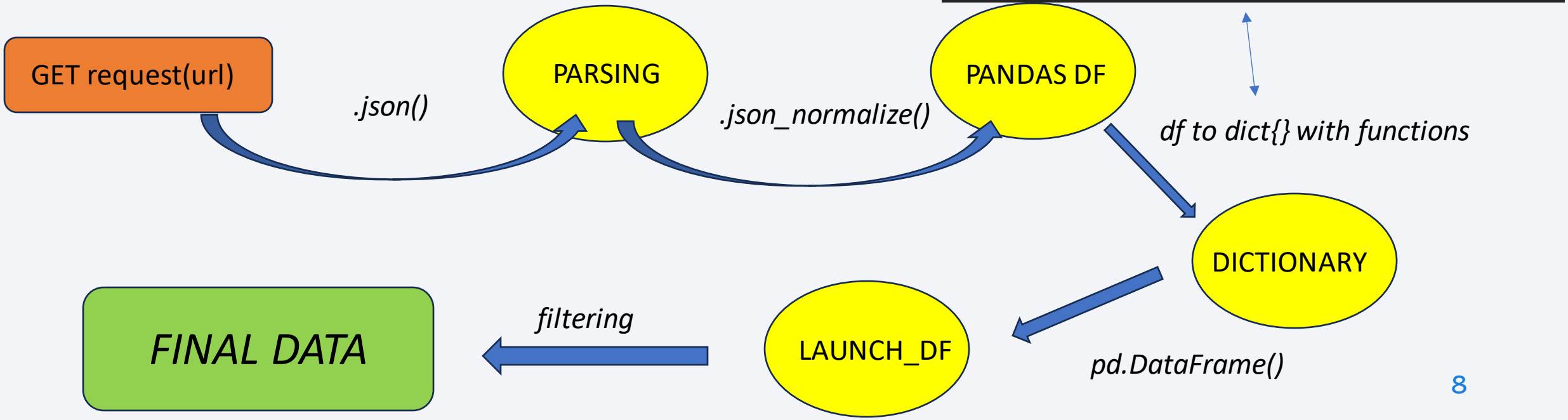
Data Collection – SpaceX API

GitHub Link with full development:

https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_SpaceX_API_calls_notebook.ipynb

```
def getLaunchSite(data):
    for x in data['launchpad']:
        if x:
            response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
            Longitude.append(response['longitude'])
            Latitude.append(response['latitude'])
            LaunchSite.append(response['name'])
```

Data collected using Requests and Pandas packages:



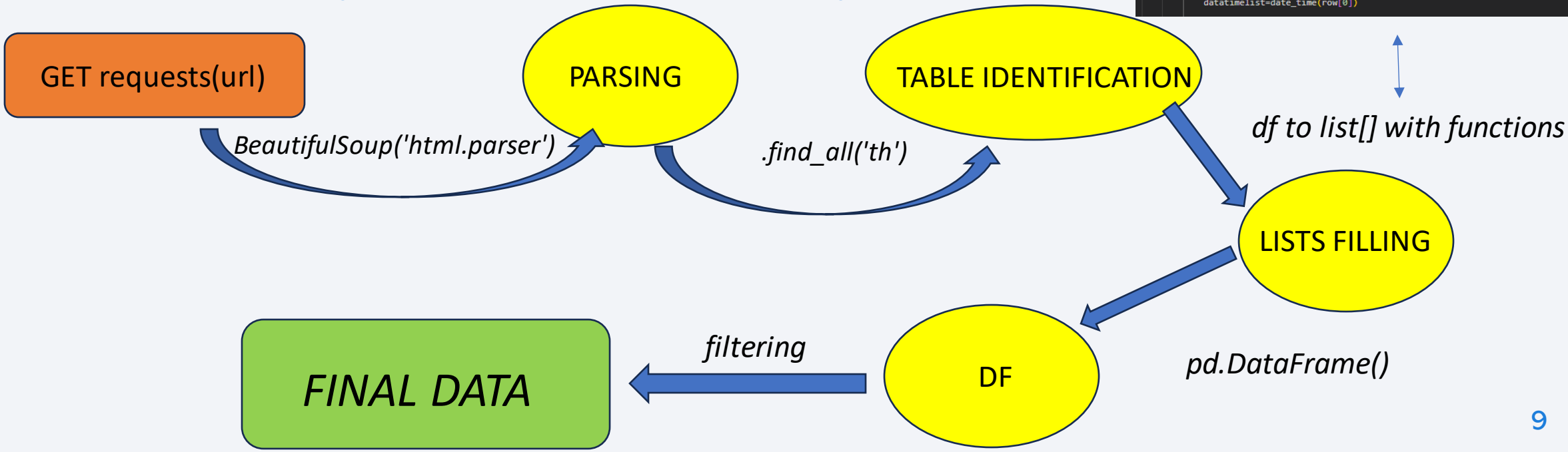
Data Collection - Scraping

GitHub Link with full development:

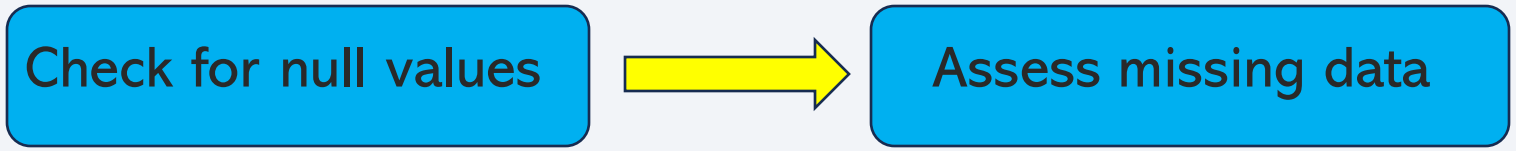
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_TableScrap(Wiki).ipynb

```
extracted_row = 0
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
    for rows in table.find_all("tr"):
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
        print(flag)
        row=rows.find_all('td')
        if flag:
            extracted_row += 1
            launch_dict={}
            launch_dict['Flight No.'] = flight_number
            datatimelist=date_time(row[0])
```

Data collected using Requests and Pandas packages:



Data Wrangling



-PayloadMass (5)	Replacing with mean value
-Landing pad (28)	Non parseable variable (no data available)

GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_SpaceX_API_calls_notebook.ipynb
&
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_DataWrangling.ipynb

EDA with Data Visualization

Charts plotted

- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Success Rate vs. Orbit Type
- Flight Number vs. Orbit Type
- Payload vs. Orbit Type
- Launch Success Yearly Trend

These charts were chosen based on the creation of a narrative that allow us to explain the changes occurred to Falcon 9 launches

GitHub Link with full development:

https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_Visual_EDA.ipynb

EDA with SQL

SQL queries performed

- Names of the unique launch sites.
- 5 records where launch sites begin with `CCA`.
- Total payload carried by boosters from NASA (CRS).
- Average payload mass carried by booster version F9 v1.1 .
- Dates of the first successful landing outcome on ground pad.
- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- Total number of successful and failure mission outcomes.
- Names of the booster which have carried the maximum payload mass .
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

These queries where perfomed in the same line as the visualizations.

GitHub Link with full development:

https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_SQL(EDA).ipynb

Build an Interactive Map with Folium

- Created launch site marks containing the name and a circle range for easy visualization.
- Added marker clusters that count the number of launches for each site, describing its success or failure (with color code) and cause of these.
- Developed the code to display the distance between launch sites and any other point on Earth, discussing launch site accessibility.

GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_Visual_EDA_Folium.ipynb

Build a Dashboard with Plotly Dash

Dash tool contains

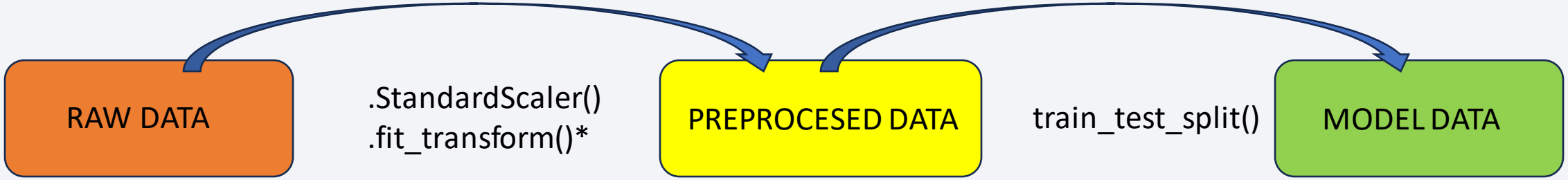
- Pie chart that visualizes the count of successes for each launch site ('All' selected) and the individual success ratio (by selecting an individual launch site).
- A scatter plot showing the relation between payload mass and launch success.
- Range slider that conditions the rockets selected, in the pie chart and the scatter plot, by the payload mass carried by the rocket.

This dash tool lets you freely switch parameters from the data in order to visualize them organically.

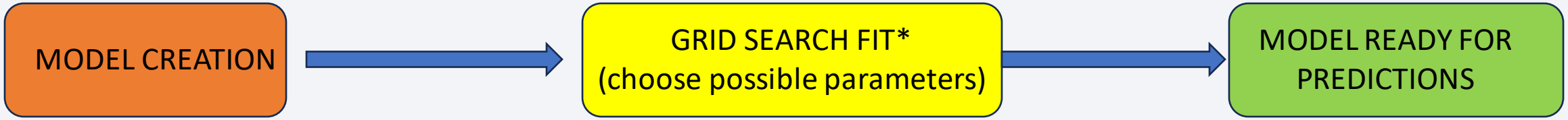
GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_Dash.ipynb

Predictive Analysis (Classification)

Data arranging:



Model creation:



*We fit with training data and predict over testing data.

GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_MachineLearningForPredictiveAnalysis.ipynb

Results

Exploratory data analysis results

* Success rate increase can be attributed to SpaceX's know-how development: orbit election, payload mass used, launch site selected, ….

Interactive analytics demo in screenshots

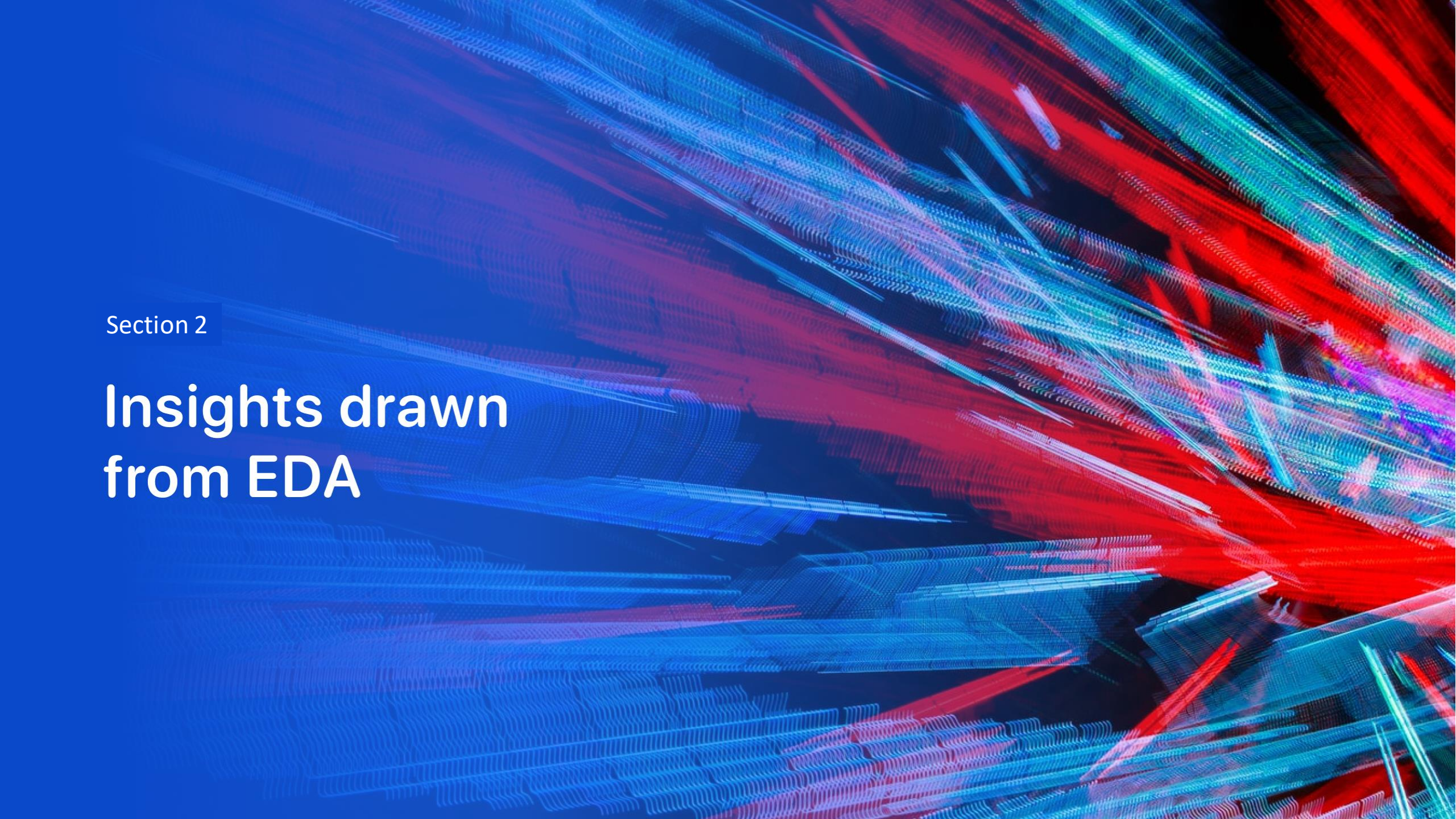
* Dash description and analysis will be later discussed.

Predictive analysis results

* Models are suitable for prediction, although false positive bias is present.

Section 2

Insights drawn from EDA

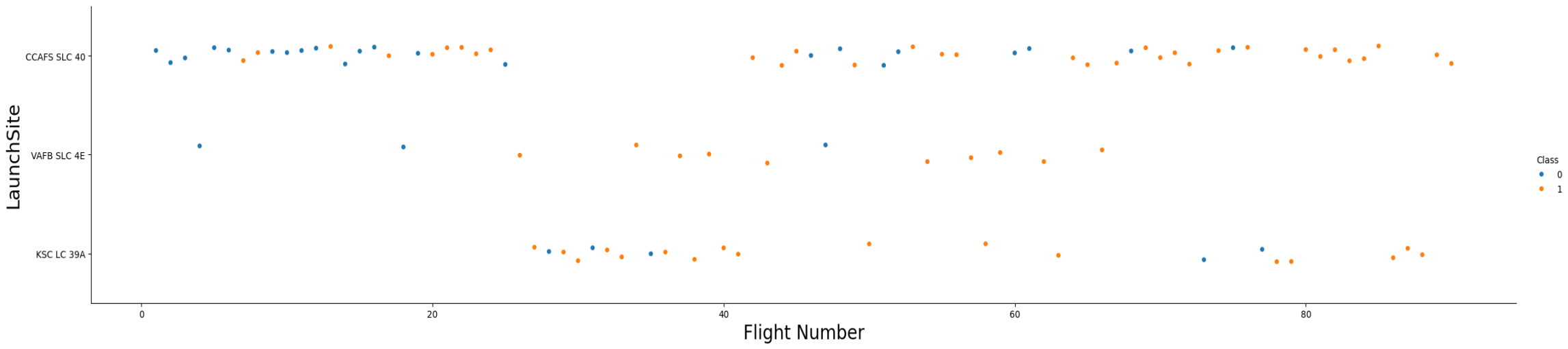


Section 2.1: VISUAL Extractive Data Analysis

Insights drawn from EDA

GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_Visual_EDA.ipynb

Flight Number vs. Launch Site

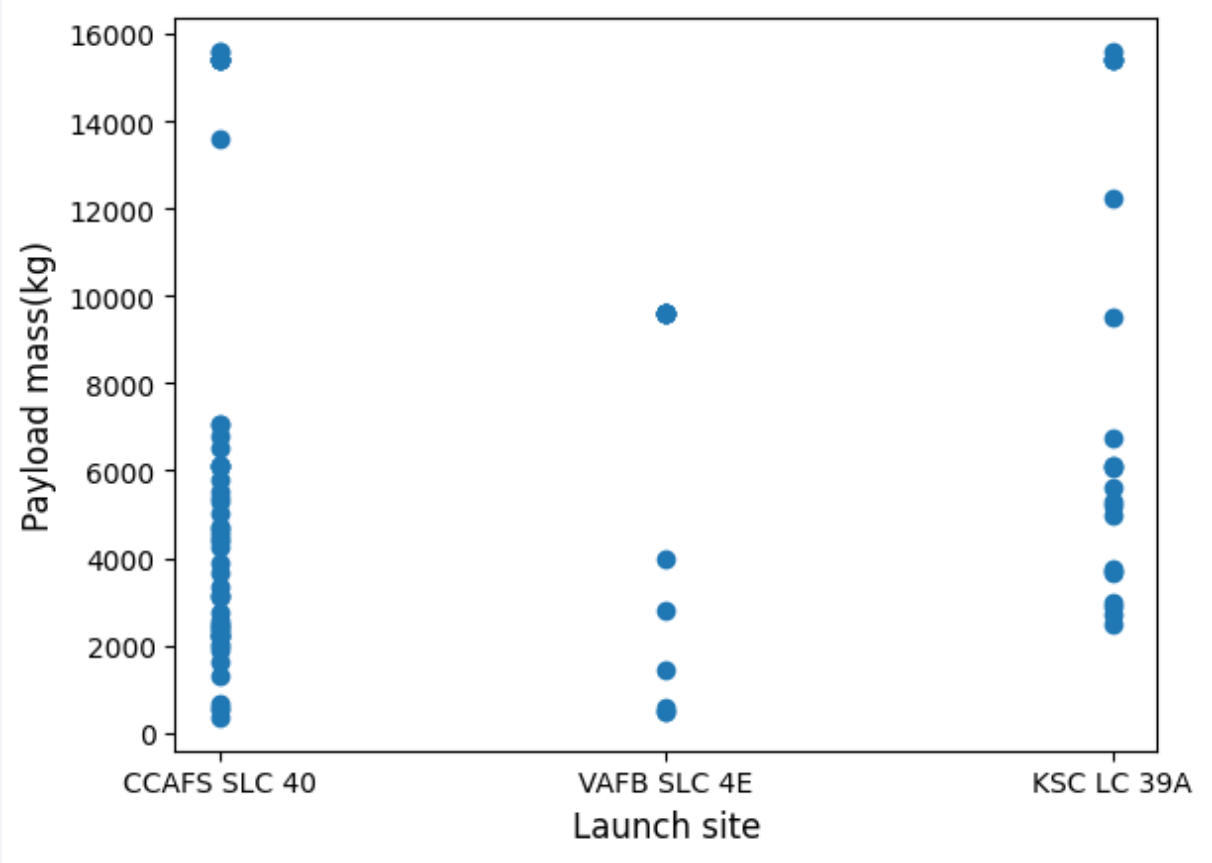


- Success rate is increased over time.
- CCAFS SLC 40 is the launch site with the highest count of rockets deployed, being the one with less success rate (60%) mainly because of it housing the first launches.
- Both KSC LC-39A and VAFB SLC 4E show a 77% success rate, note that VAFB SLC 4E has stopped launching Falcon 9 rockets.

Payload vs. Launch Site

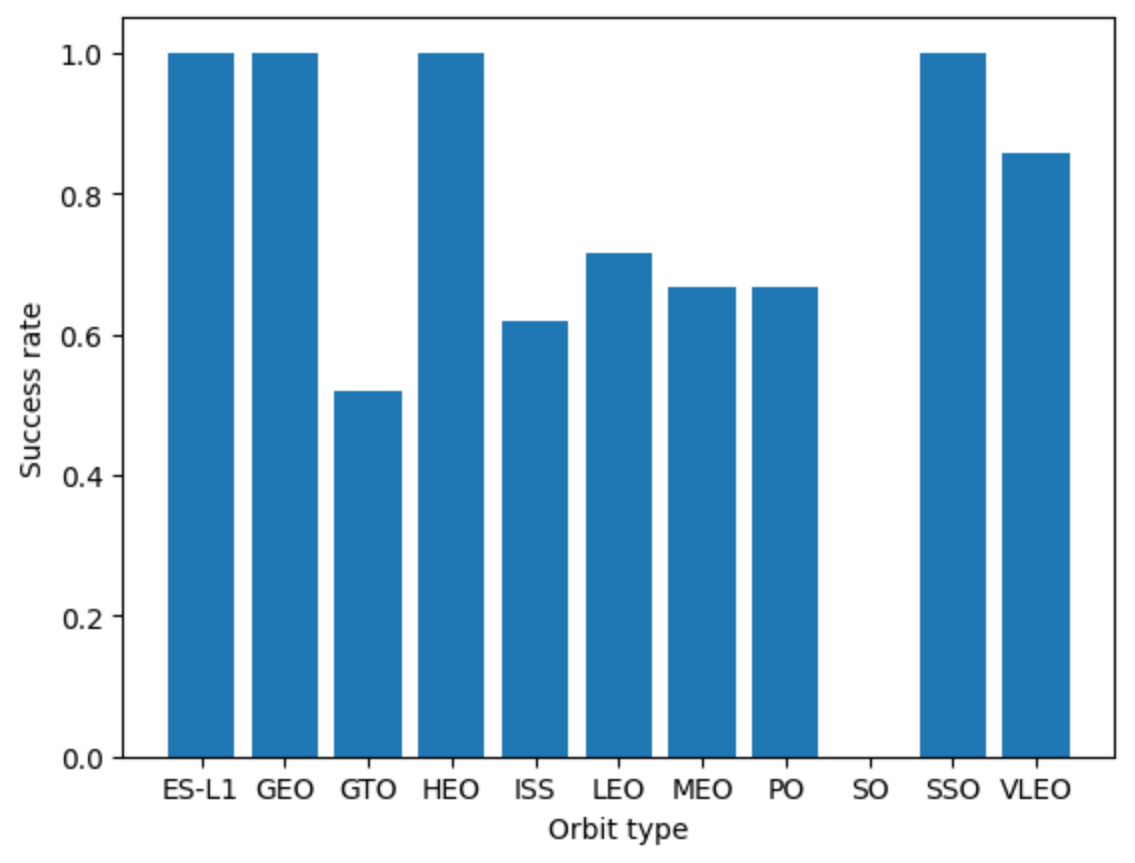
- Every launch site has a bias toward payloads with less than 10000kg of payload mass, specially VAFB SLC 4E.

- CCAFS SLC 40 is the place with a highest payload range tested whilst KSC LC 39 A is arguably the place where higher mean mass was tested.



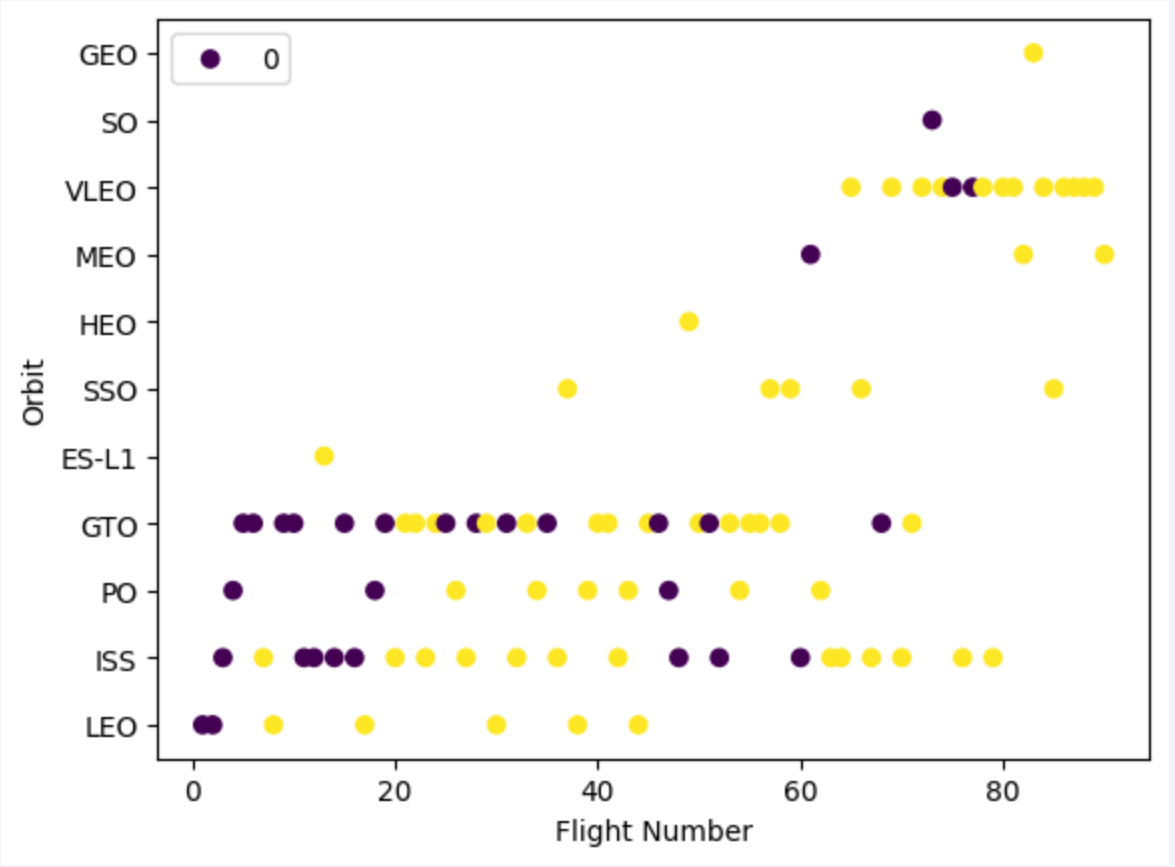
Success Rate vs. Orbit Type

- ES-L1, GEO, GTO and SSO account for the higher success rates (of 100%)
- SO orbit has never been successfully orbited.
- The remaining orbits show a success rate of 50% or higher.



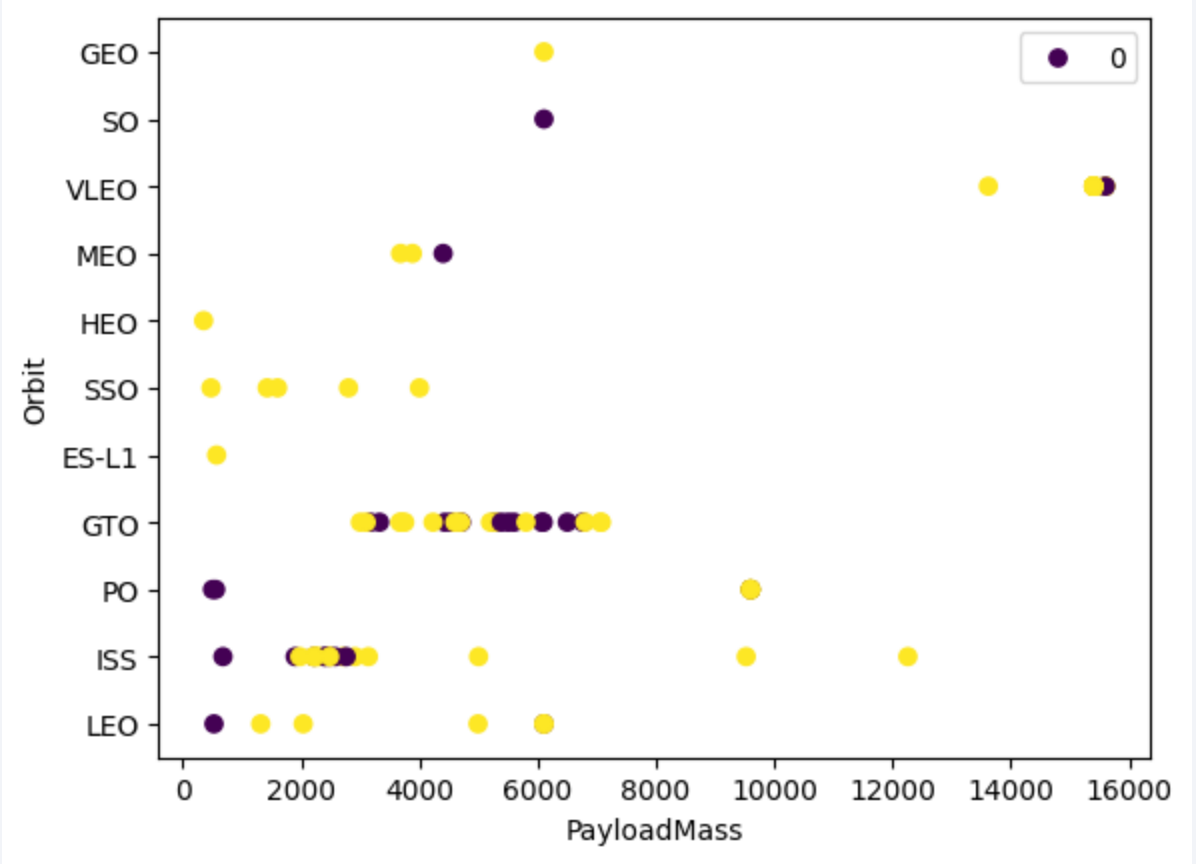
Flight Number vs. Orbit Type

- Flight number variable shows a progressive move between orbits that goes from old reliable proven orbits, such as LEO and ISS, towards new high success rate orbits, like SSO and VLEO.



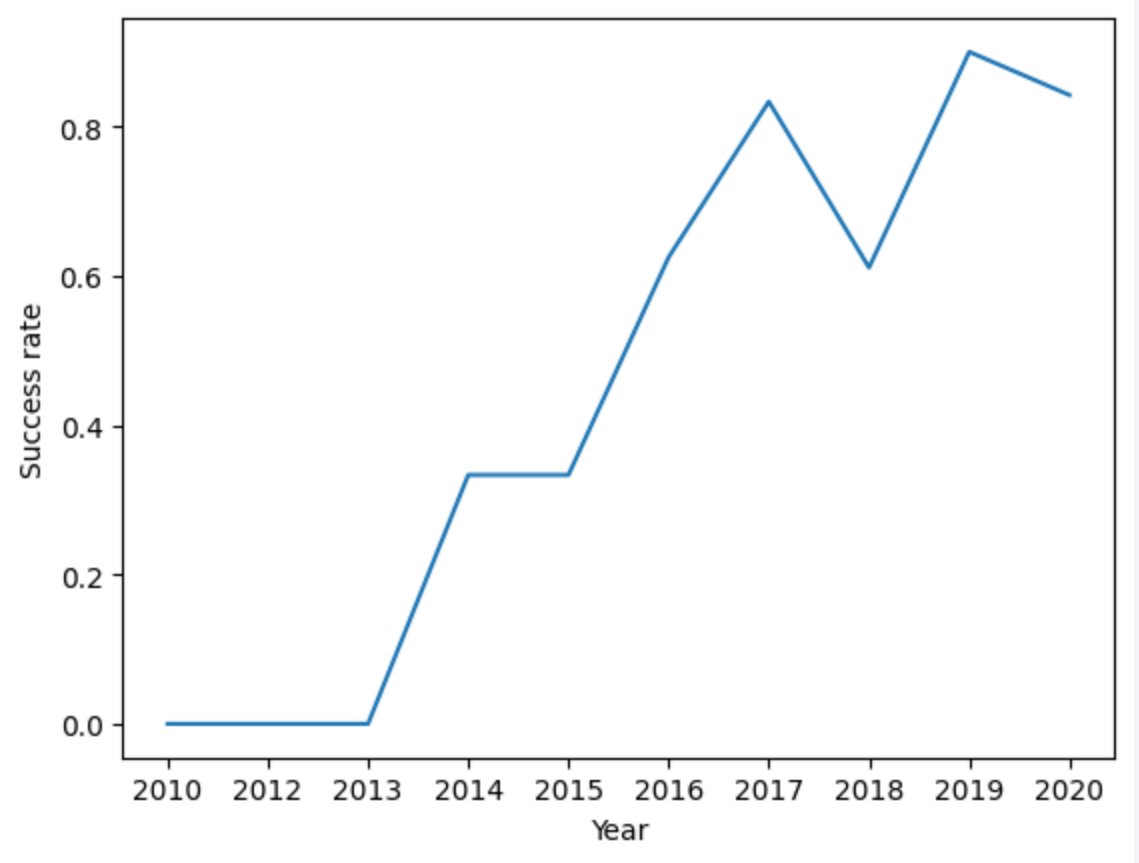
Payload vs. Orbit Type

- Most launches present a payload mass of 2000<X<7000 kg.
- Continuing previous slide, VLEO and SSO present high(>13000) and low (<5000) kg of payload mass, which shows that launches have diverged into two different types of launch.



Launch Success Yearly Trend

- Yearly launch success rates have increased over time up to 84%.
- Success rate starts growing in 2014.
- Highest year-on-year growth happens between 2013-2014 with 36pp.
- The only recession rates happen from years 2017-18 and 2019-20.



Section 2.2: SQL Extractive Data Analysis

Insights drawn from EDA

*Code logic may be repeated throughout the different queries, therefore any unexplained code has already been explained in previous queries.

GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_SQL(EDA).ipynb

All Launch Site Names

- *Find the names of the unique launch sites*

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

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

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

- We select the disctinct values for the variable "Launch_Site" from our table.
- This command performs similarly to how Pandas's .unique() method would work.

Launch Site Names Begin with 'CCA'

- *Find 5 records where launch sites begin with `CCA`*

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

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

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

- "like '%CCA%' " filters variable results containing 'CCA' between any characters.
- "limit 5" selects the first five results.

Total Payload Mass

- *Calculate the total payload carried by boosters from NASA (CRS)*

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

✓ 0.0s

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

sum(PAYLOAD_MASS__KG_)
45596

- Sum(PAYLOAD_MASS_KG_) gives the total payload mass.
- Filtered 'Customer' variable to only account for the payload mass of 'NASA (CRS)' (the customer asked for).

Average Payload Mass by F9 v1.1

- *Calculate the average payload mass carried by booster version F9 v1.1*

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like '%F9 v1.1%'
 ✓ 0.0s

 * sqlite:///my_data1.db
Done.

avg(PAYLOAD_MASS__KG_)
        2534.6666666666665
```

- Avg(PAYLOAD_MASS_KG_) calculates the mean value for the variable
- Where Booster_Version like '%F9 v1.1' filters the name of the booster version selected

First Successful Ground Landing Date

- *Find the dates of the first successful landing outcome on ground pad*

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome like 'Success (ground pad)'
 ✓ 0.0s

 * sqlite:///my_data1.db
Done.

 min(Date)
 2015-12-22
```

- Min(Date) obtains the oldest date.
- In this case, like performs the same result as '='.

Successful Drone Ship Landing with Payload between 4000 and 6000

- *List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000*

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000
```

✓ 0.0s

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Where acts as a conditional and 'and' adds sequent conditions.

Total Number of Successful and Failure Mission Outcomes

- *Calculate the total number of successful and failure mission outcomes*

```
%%sql SELECT
SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS Successful_Outcomes,
SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS Failed_Outcomes
FROM SPACEXTBL;
```

✓ 0.0s

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

Successful_Outcomes	Failed_Outcomes
100	1

- 'CASE' statement will add 1 to the sum of 'Successful_Outcomes' if the 'Mission_Outcome' variable presents the 'Success' characters.
- Same logic is applied to the 'Failed_Outcomes' case.

Boosters Carried Maximum Payload

- *List the names of the booster which have carried the maximum payload mass*

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

✓ 0.0s

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

```
%sql select substr(Date, 6,2) as Month,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTBL where (Landing_Outcome like '%Failure%' and substr(date,0,5)='2015')
```

✓ 0.0s

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

Month	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

- 'substr' slices the Date column to obtain the month (in the first case) and the year (in the second case).

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(*) from SPACEXTBL where Date>'2010-06-04' and Date<'2017-03-20' group by Landing_outcome order by count(*) DESC
```

✓ 0.0s

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

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

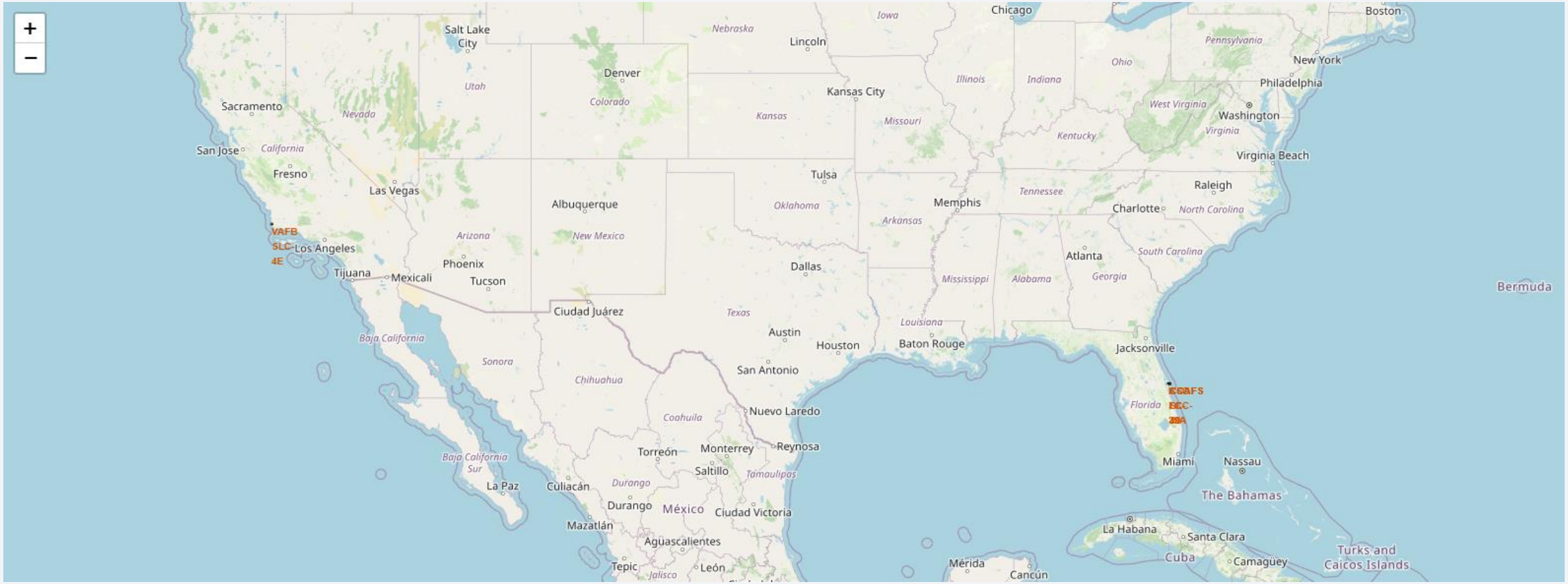
- "Order by count(*) DESC" will order landing outcomes based on their count in descending order.

Section 3

Launch Sites Proximities Analysis

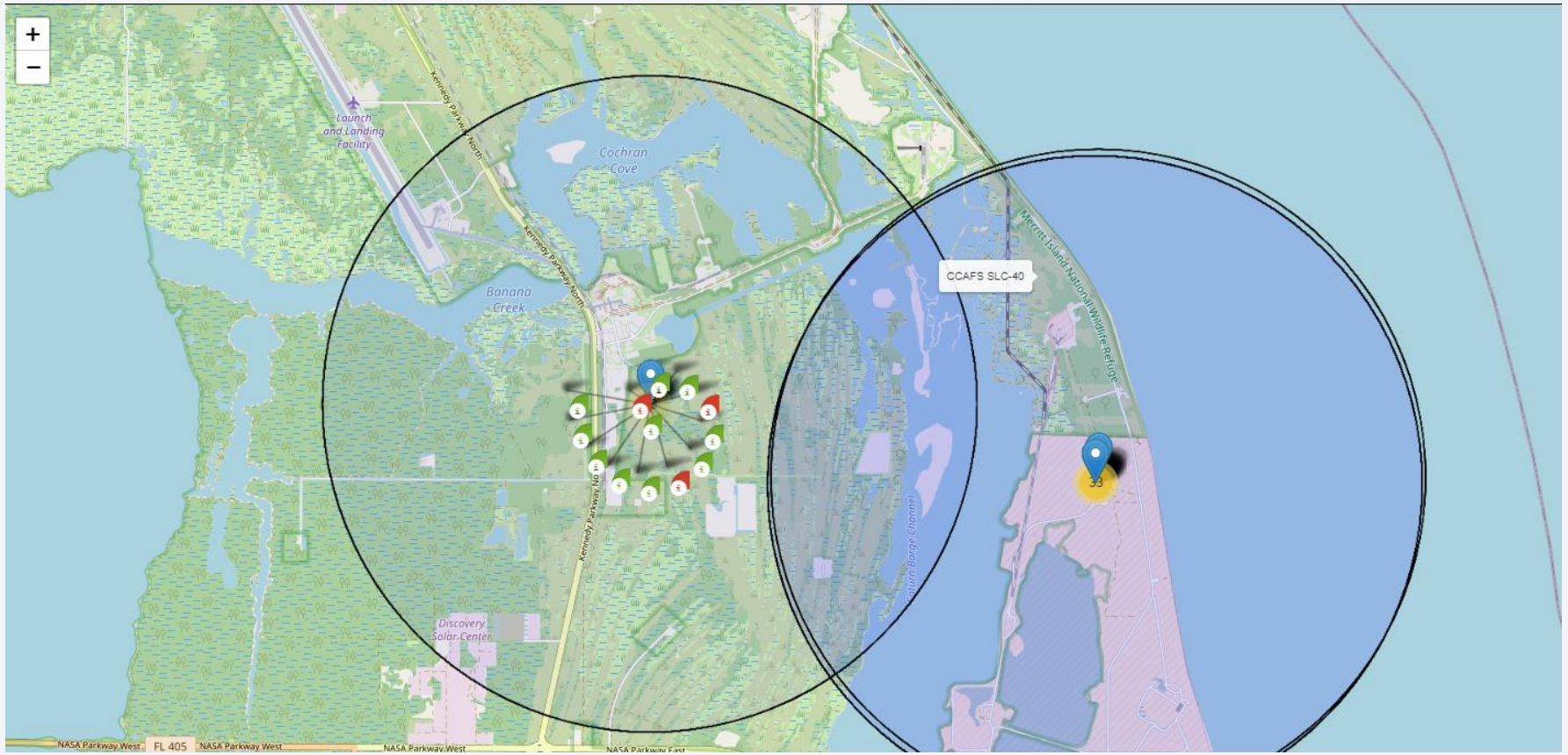
GitHub Link with full development:
https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_Visual_EDA_Folium.ipynb

Folium: Launch sites locations



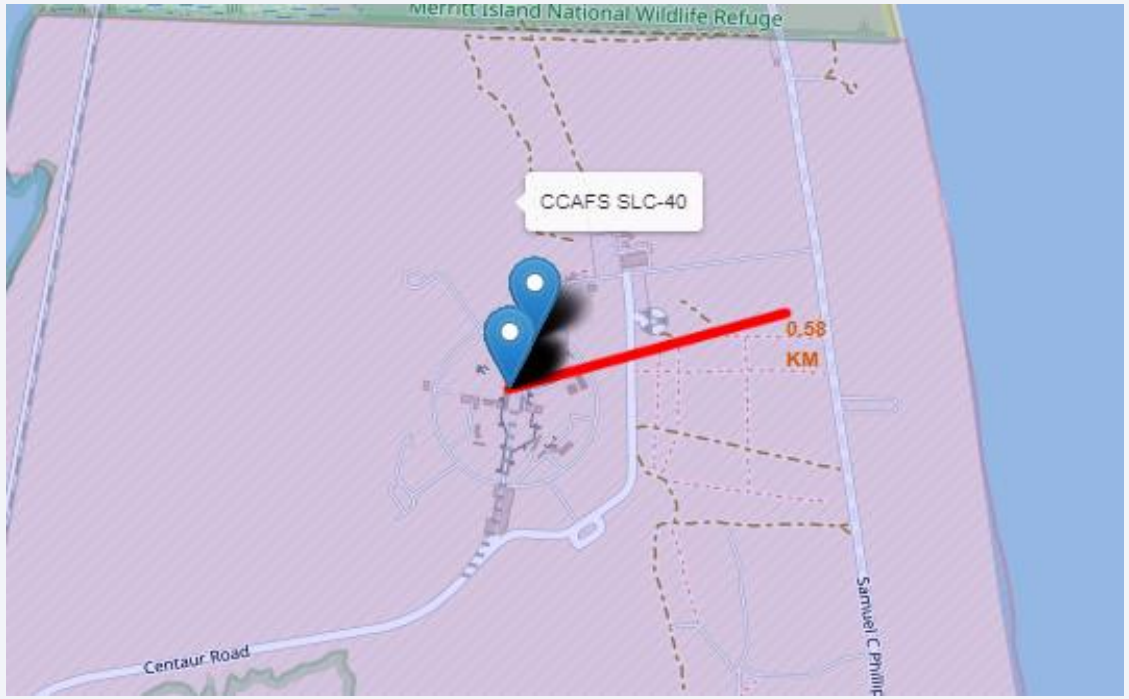
Three of the launch sites are in the state of Florida (Easts Coast), in Cape Canaveral and Merritt Island.
The remaining one, VAFB SLC 4E, is seated in USA's West Coast, in the state of California.

Folium: Launch site outcomes



KSC LC 39-A (Merritt Island) is the most succesful launch site.

Folium: Launch site accessibility



This is CCAFS SLC-40, it is one of Cape Canaveral's launch sites and it's located 0.58km away from its nearest coastline.

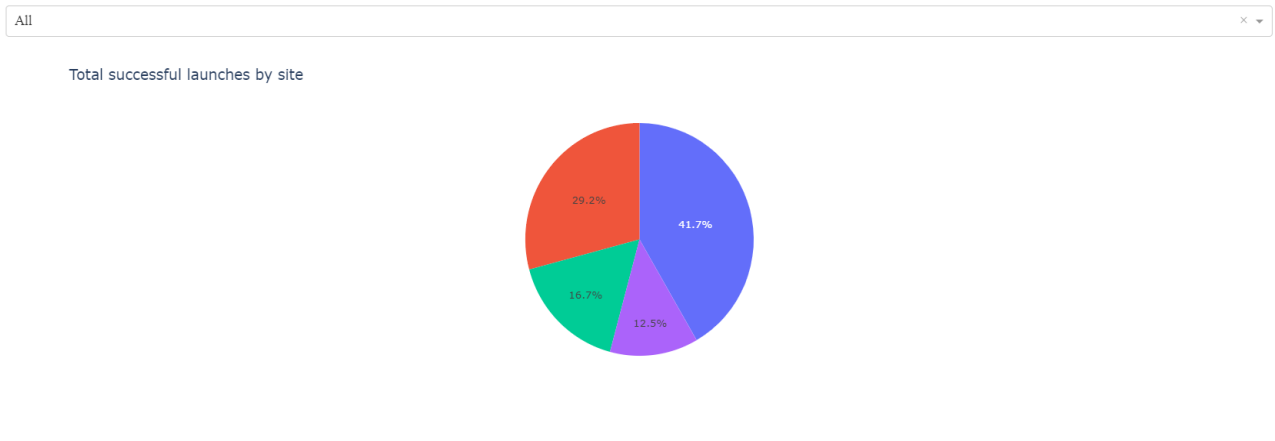
The importance of the launch site being close to a coastline, railway or a highway lies on the fact that these places need to be heavyly supplied for its proyects to go on.
It's not just a matter of success but also a matter of failure, as comunication and physical accesibility is also necessary in case the mission goes wrong (e.g., an engine fails and sets the rocket on fire).

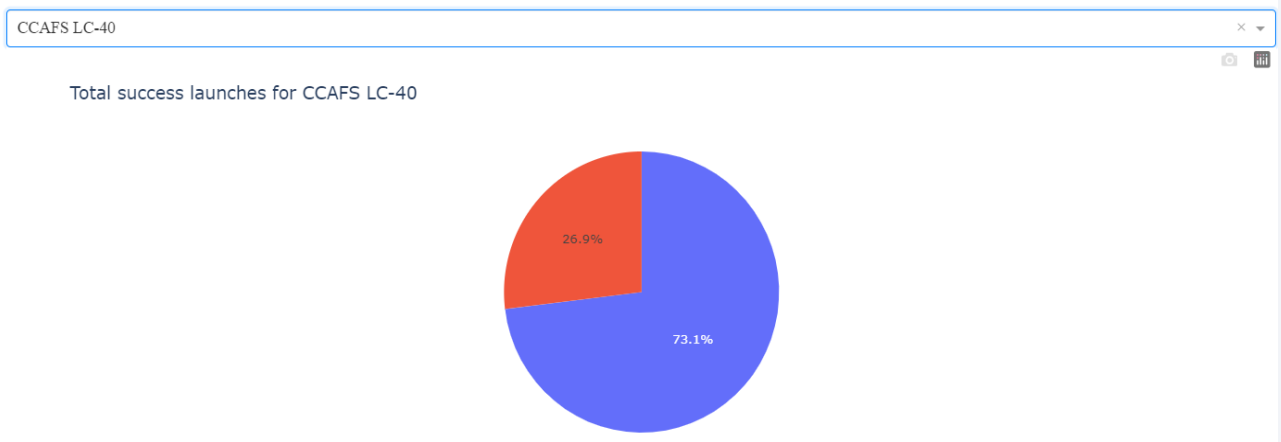
Section 4

Build a Dashboard with Plotly Dash

GitHub Link with full development:

Launch success pie chart



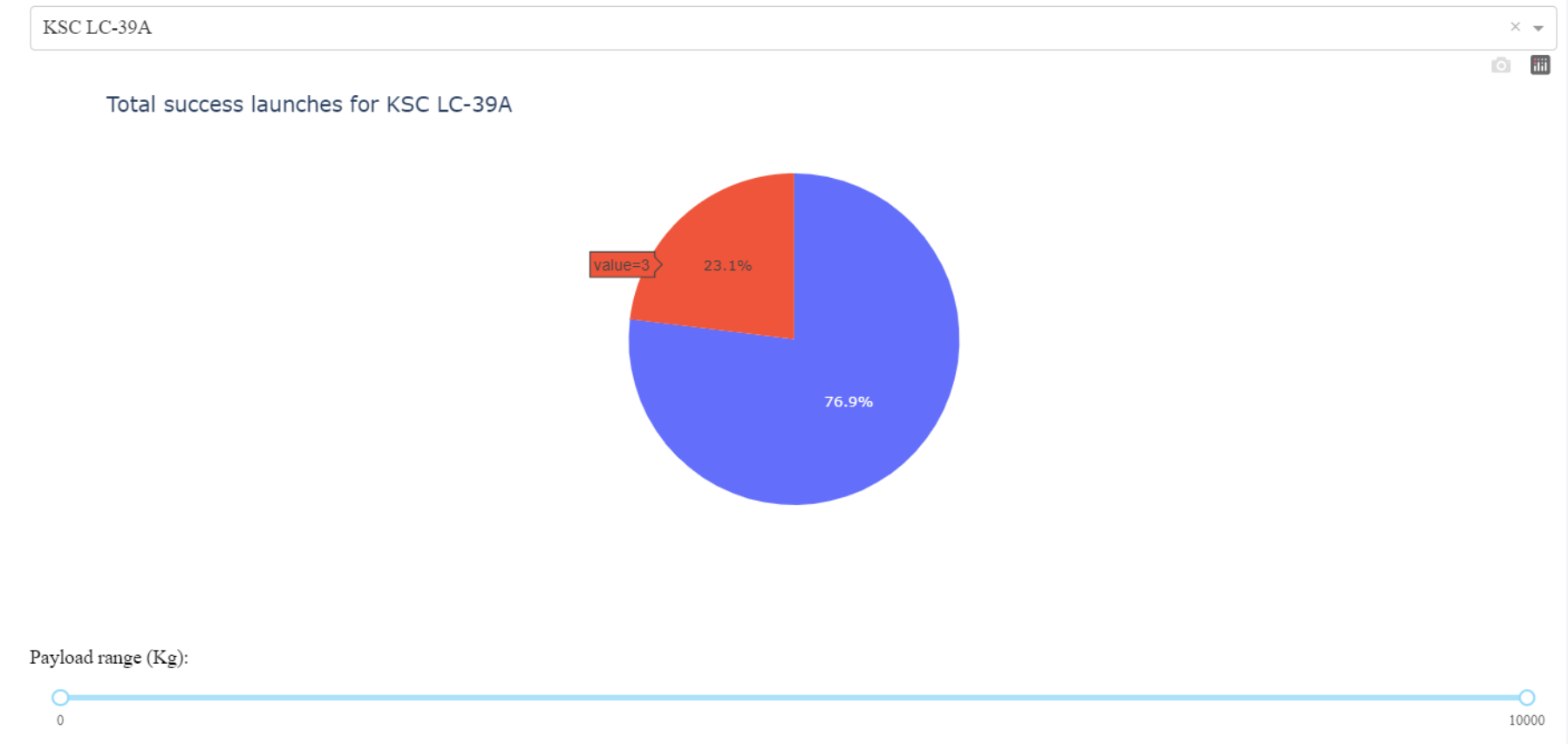


When 'All' is selected the chart shows the count of successes for each launch site.
When a single launch site is selected it show the relation of successes (blue) to failures (red).

Most important findigs

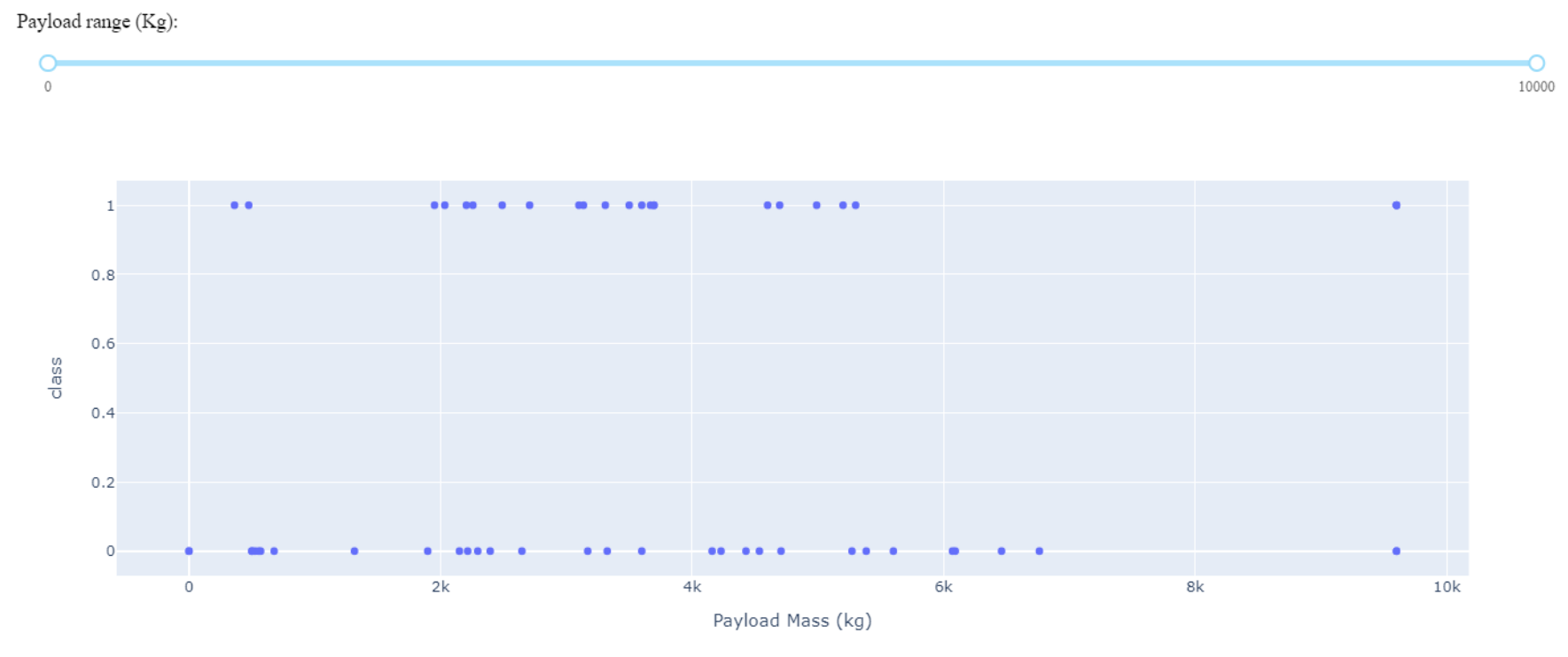
- KSC LC 39A is the most succesful launch site from any of them with a 76.9% individual success rate (as we saw previously with Folium).
- The rest show a individual success rate over 57% each.

KSC LC-39A launch success pie chart



KSC LC-39A is important because it shows the detrimental consequences of carrying too much payload mass, as the 3 failed launches are above the 5500 kg mark.

Launch success and payload range ('All' selected)



Most successes are concentrated around the 2k to 5.5k kg of payload mass.
Failures show a wider range "payload mass wise" even though they tend to happen less.

Launch success and payload range ('All' selected)

This slide goes on to show that higher payload mass carries a greater danger *range-wise*, those rockets with a payload mass different from 9.6 tons are more prone to fail.
This is a reality caused by model standarization.

This pattern has already been detected when studying KSC LC-39A launch site.

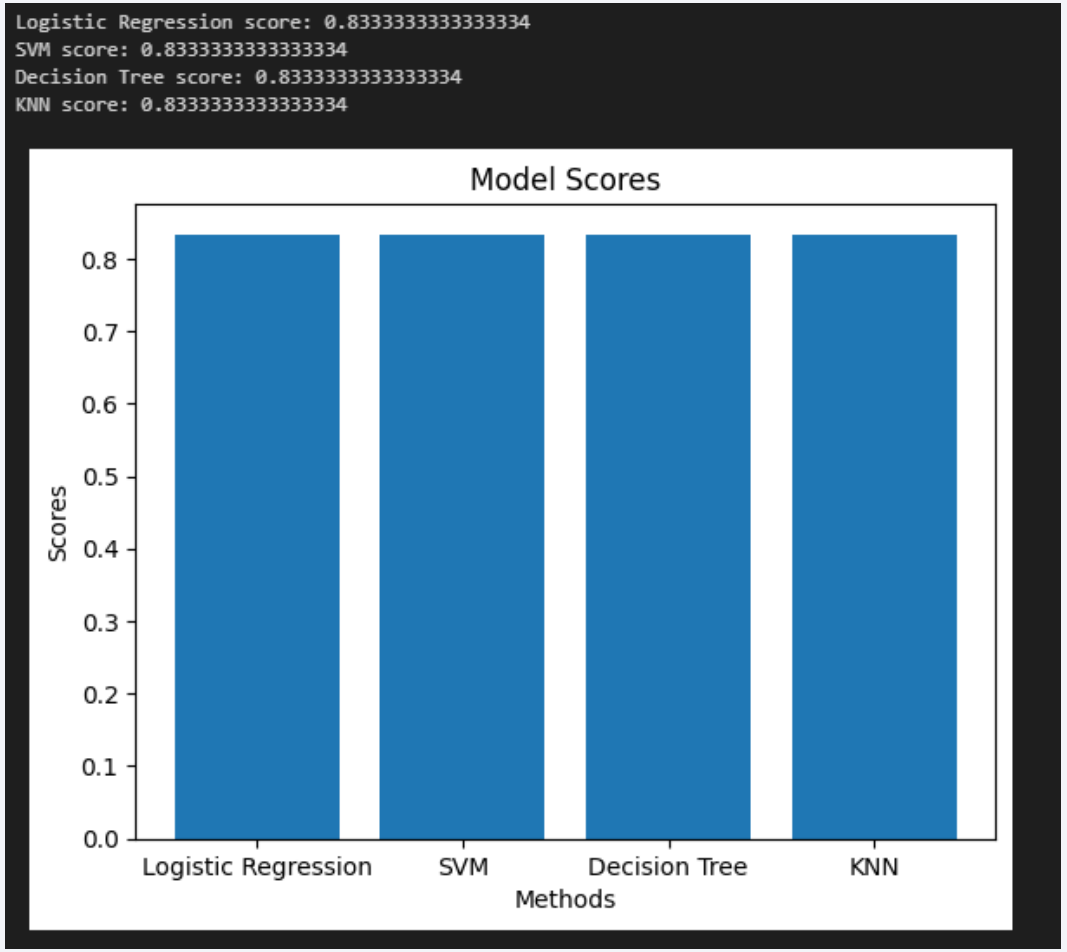
Section 5

Predictive Analysis (Classification)

GitHub Link with full development:

https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study/blob/main/Falcon9_MachineLearningForPredictiveAnalysis.ipynb

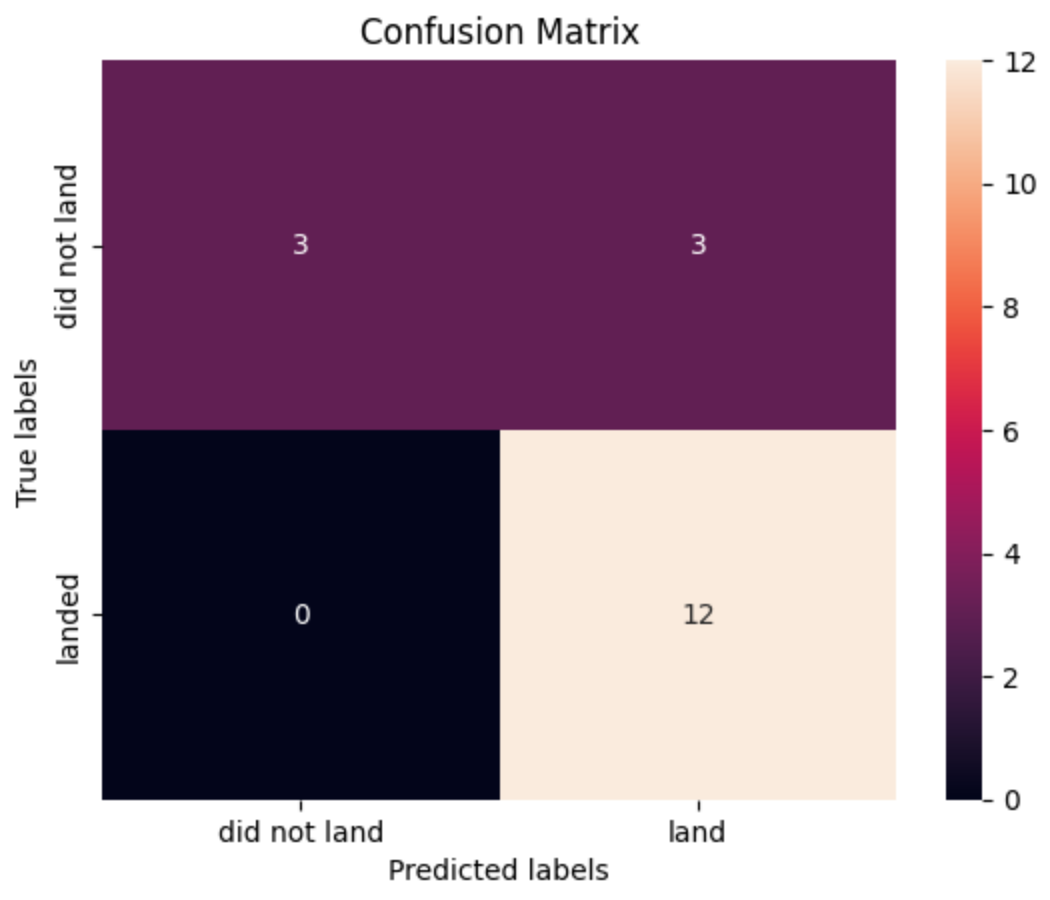
Classification Accuracy



After applying GridSearch, in order to obtain the best parameters for our prediction, every model shows the same score at 0.833334.

Since every method here is capable of performing cathegorical predictions any of them is as valuable as any other.
However, this results may change on the future(e.g. when the sample gets updated)

Confusion Matrix



This is what the confusion matrix looks like for every method.

To note that even the same kind of error is committed each time (3 false positives).

This result may vary with data updates (e.g., appearance of false negatives)

Conclusions

- Success rate is increased over time.
- Launches have progressed to newer and higher success rate orbits, although causality direction is yet to be proven.
- KSC LC-39A is the most successful launch site.
- The payload mass range for successful launches is between 2000 and 5500kg, as F9 V1.1 (2013) mean mass is 2534.66kg, mean mass has increased over time.
- Launch types have diverged into lower and higher than the mean payload mass.

- There is a current prediction bias towards false positives.
- We can use any of the models described to make a prediction as they all perform the same R2 (83.333).

Appendix

Every code is written and described in the attached GitHub:

https://github.com/DrepoBlapo/Applied-Data-Science-Capstone-a-Falcon-9-study

**Some descriptions may be written in spanish as it's my mother language and code serves personal purposes too.*

Thank you!