IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rodrigo Quezada
March 27th, 2022

Outline

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

Executive Summary

- Summary of methodologies

  Data Science methodology is in place as we start with problem to solve which is to Predict if the Falcon 9 first stage will land successfully. Then we move into data collection using both REST APIs and Web Scraping. Then we proceed with data wrangling. Right after we do Exploratory Data Analysis (EDA) to both visualize and understand data. After that four different machine learning models were used and compared to produce conclusions and provide the findings (an answer to the problem).

- Summary of all results

  Falcon 9 first stage will land successfully is expected 66.6% of the time currently. However, with current report I am **making suggestions for the Falcon 9 first stage to increase landing success**.

Introduction

- Project background and context
  - Company market position (secure more customers) and financials (profits) are impacted by the quotes made to potential customers based on boosters reusage capability which in turn needs to be estimated.

- Problems you want to find answers
  - Predict if the Falcon 9 first stage will land successfully.
  - What conditions can change in order for Falcon 9 first stage to increase landing successfully.

Section 1

Methodology

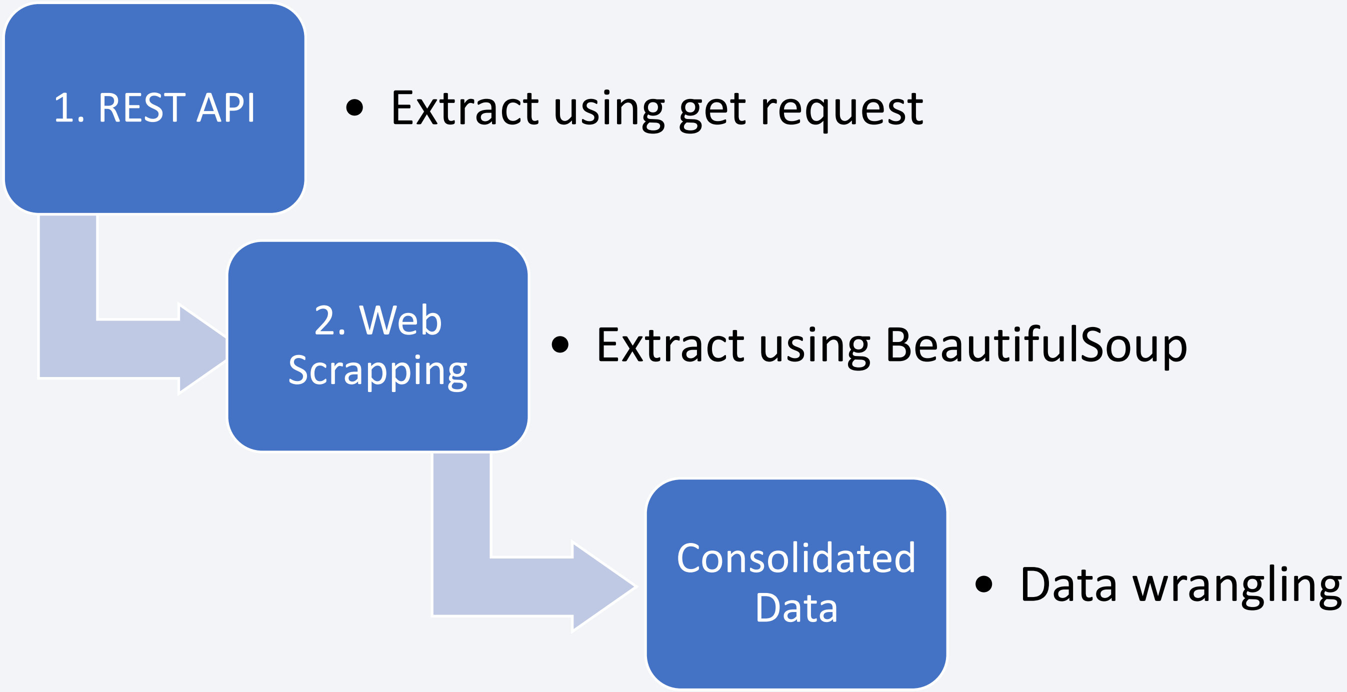
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using both SpaceX REST API through a get request and Web Scrapping with BeautifulSoup at Wiki pages for Falcon 9 launch records.
- Perform data wrangling
  - Data was processed as consolidating into a single Data Frame, addressing missing values, formatting text to numbers, and removing unneeded data *(ie Falcon 1 observations)*.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models were built, tuned, and evaluated.

Data Collection

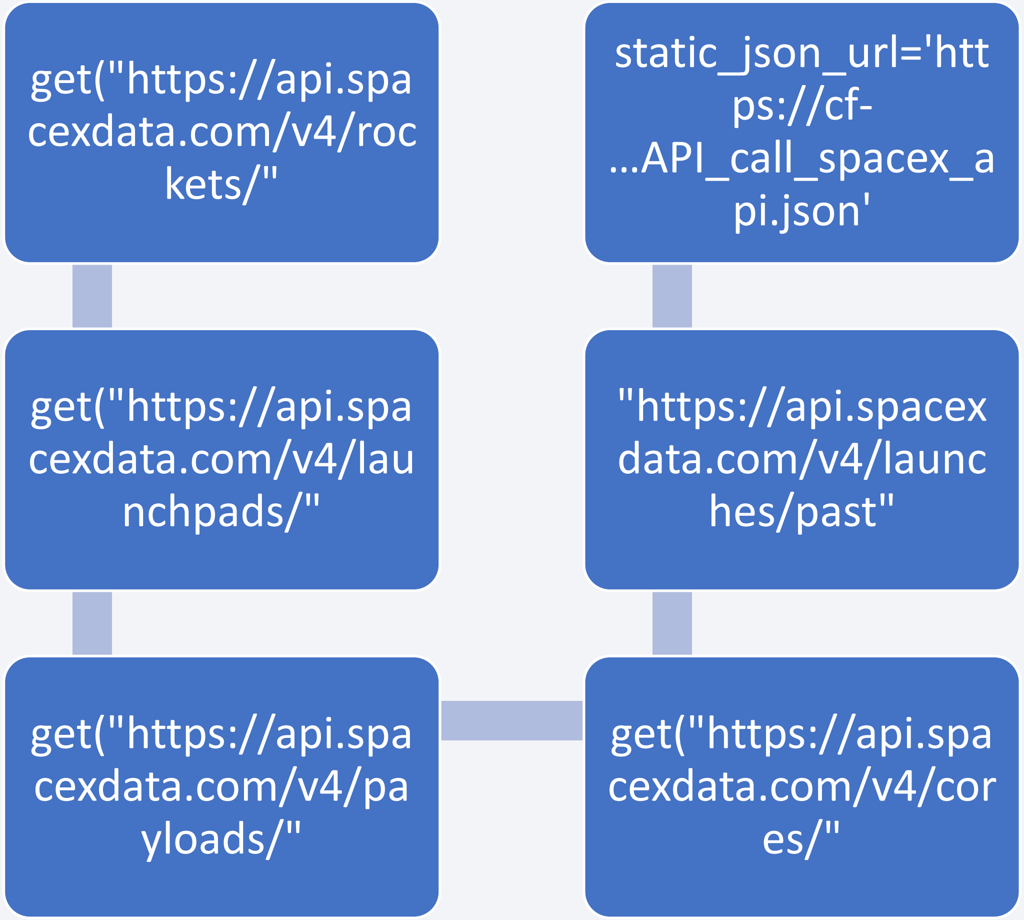
- Data sets collection can be described in 2 phases as we can appreciate in the flowchart below that lead to the next natural step of Data Wrangling:



Data Collection – SpaceX API

- To the right we see more detail on the flowchart of SpaceX API calls data collection with SpaceX REST calls.

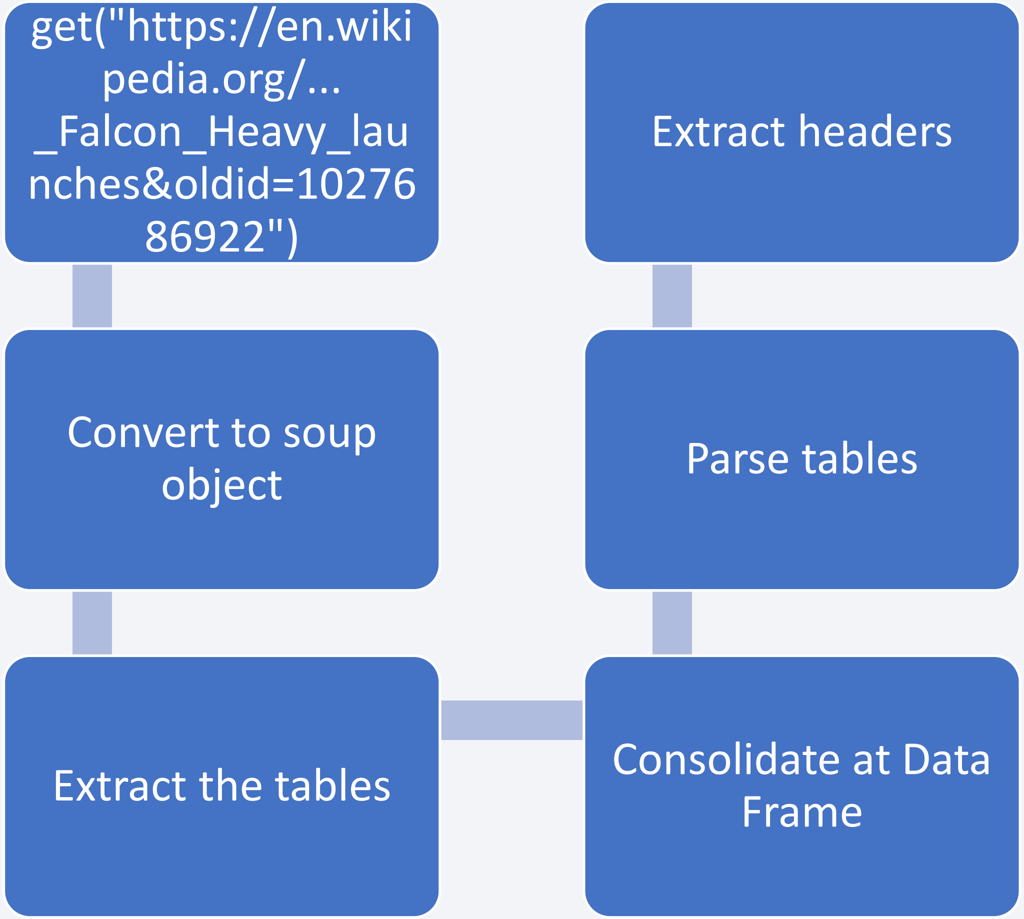
- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed SpaceX API calls notebook which are completed including code cells and outcome cells, as an external reference and peer-review purpose



Data Collection - Scraping

- To the right we see more detail on the flowchart of web scraping data collection with SpaceX REST calls.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed web scraping notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose



Data Wrangling

- To the right we see more detail on the flowchart of web scraping data collection with SpaceX REST calls.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed data wrangling notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose

1
- Address missing values

2
- Format text into numbers

3
- Remove unnecessary data

EDA with Data Visualization

- To the right we see bullet point format on the charts that were plotted to visualize and understand data as an exploratory approach.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed EDA with data visualization notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose

- 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

EDA with SQL

- To the right we see bullet point format on the SQL queries you performed to understand data as an exploratory approach.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed EDA with SQL notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose

- 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 acheived.
- 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

Build an Interactive Map with Folium

- To the right we see bullet point format on the map objects such as markers, circles, lines, etc. you created and added to a folium map to understand spatial relationship.

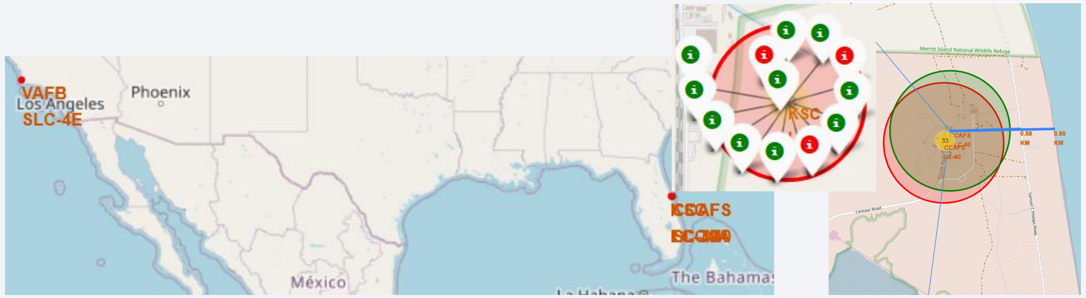
- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed EDA with SQL notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose

- All Site Locations Markers on Global Map

- East Locations Markers Observation

- Color-labeled Launch Outcomes maps

- Site Proximities Map

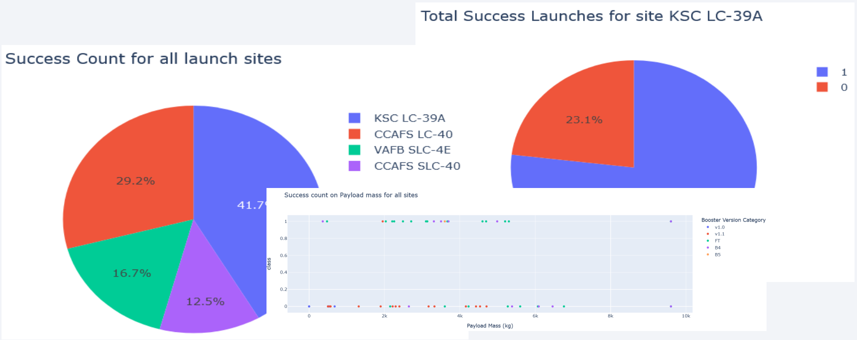


Build a Dashboard with Plotly Dash

- To the right we see bullet point format on the map objects such as markers, circles, lines, etc. you created and added to a folium map to understand spatial relationship.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed EDA with Plotly Dash lab Python file, as an external reference and peer-review purpose

- Launch Success Count per Site
- Launch Site with Highest Launch Success
- Overall Payload Vs Launch Outcome Review
- Specific Payload Vs Launch Outcome Review



Predictive Analysis (Classification)

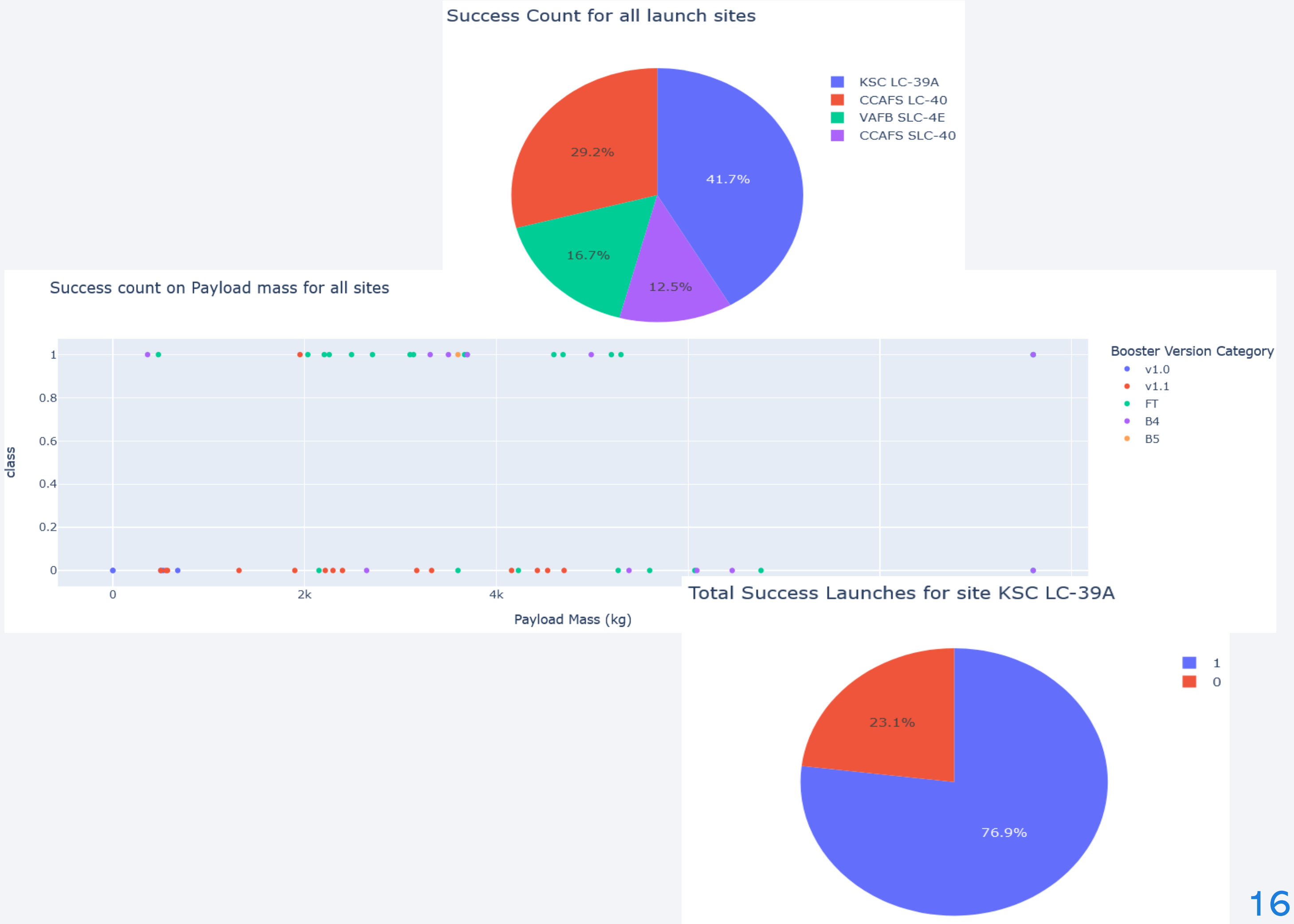
- To the right we see more detail on the flowchart of the modeling process for predictive analytics.

- And here is the GitHub URL https://github.com/Rod256/Rod_SpaceX_DS_Assignment of the completed predictive analysis notebook, which are completed including code cells and outcome cells, as an external reference and peer-review purpose

1. Preprocessing
2. Split Train and Test data sets
3. Grid Search
4. Determine Model with best accuracy
5. Visualize Confusion Matrix

Results

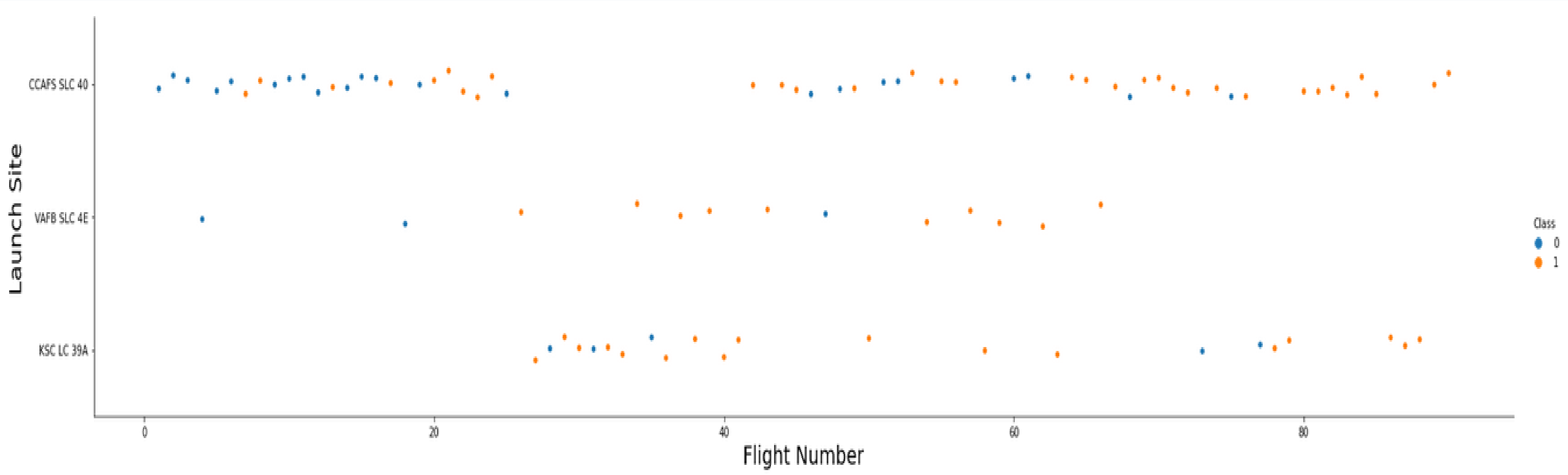
- To the right interactive analytics demo in screenshots.
- Exploratory data analysis results details in upcoming section.
- Predictive analysis results details in upcoming section.



Section 2

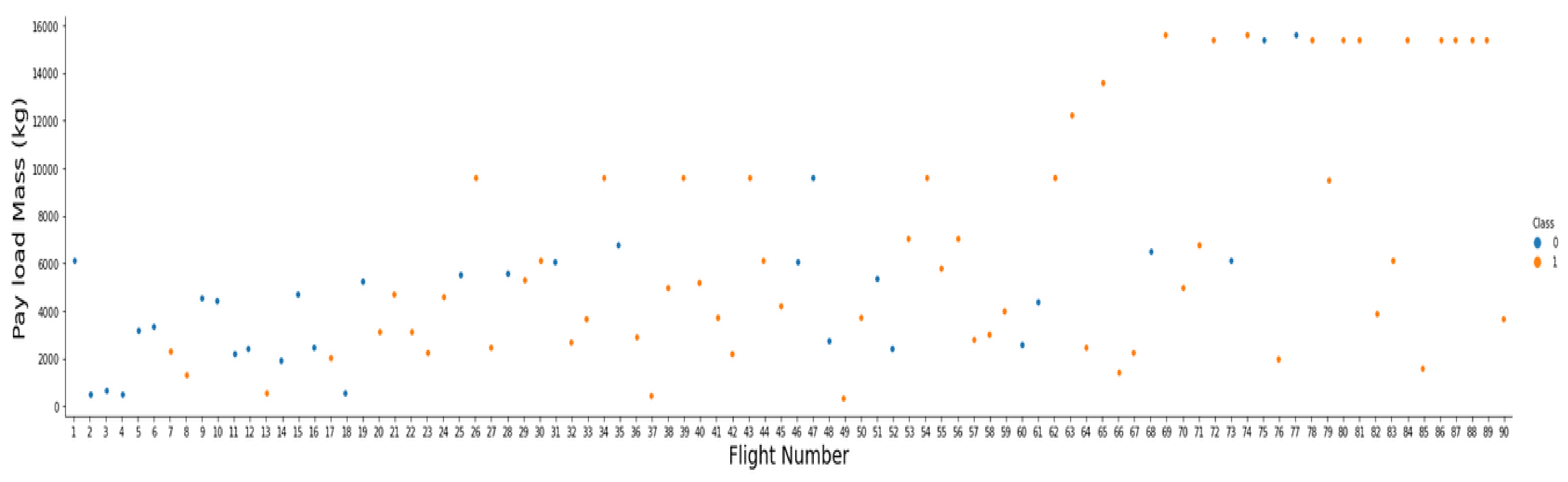
Insights drawn from EDA

Flight Number vs. Launch Site



Scatter plot above shows Payload vs. Launch Site. Here we can appreciate that there is a significant correlation on launch site related to launching success rates.

Payload vs. Launch Site

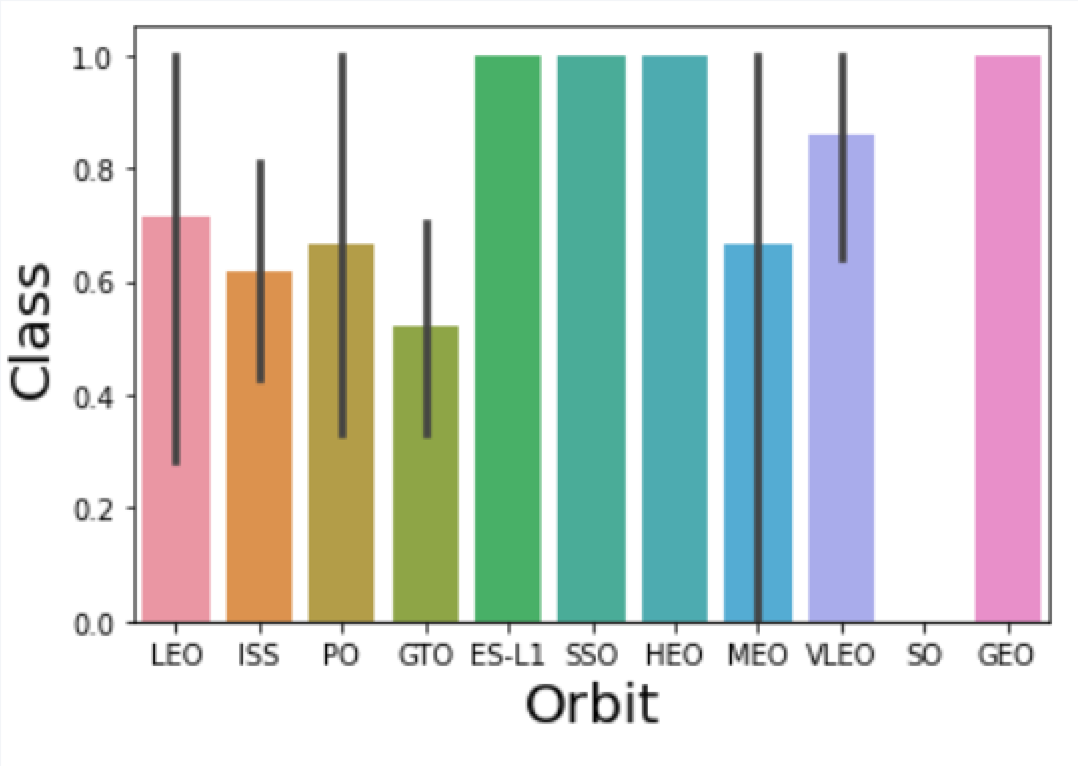


Scatter plot above shows Flight Number vs. Launch. Thus, combines data of success/failure per flight number related to the Payload Mass (kg). On this exploration, we can see an overall increased number of successful flights with Payload Mass above 8,000 kg. Yet less of a correlation than launch site itself.

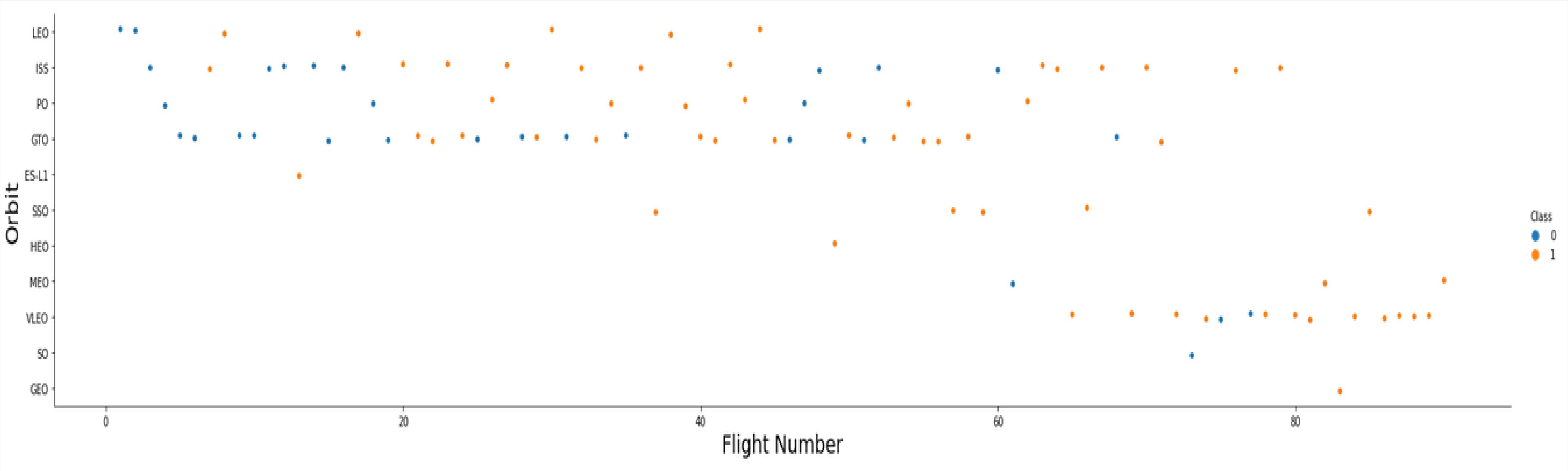
Success Rate vs. Orbit Type

- To the right we can see a bar chart for the success rate of each orbit type.

- We can appreciate 4 of the orbits have been consistently successful *(ES-L1, SSO, HEO, GEO)*.

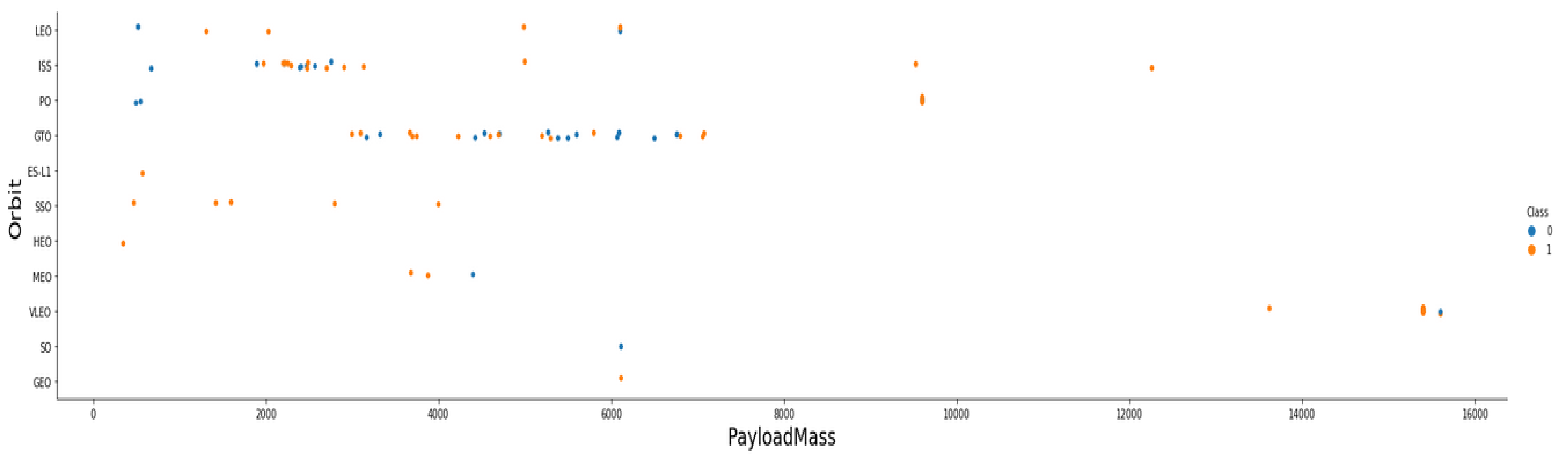


Flight Number vs. Orbit Type



Scatter plot above shows Flight Number vs. Orbit type. Thus, combines data of success/failure per flight number related to the Orbit. We already knew there are 4 which are always successful yet overall, there is a low informational gain on this relationship.

Payload vs. Orbit Type

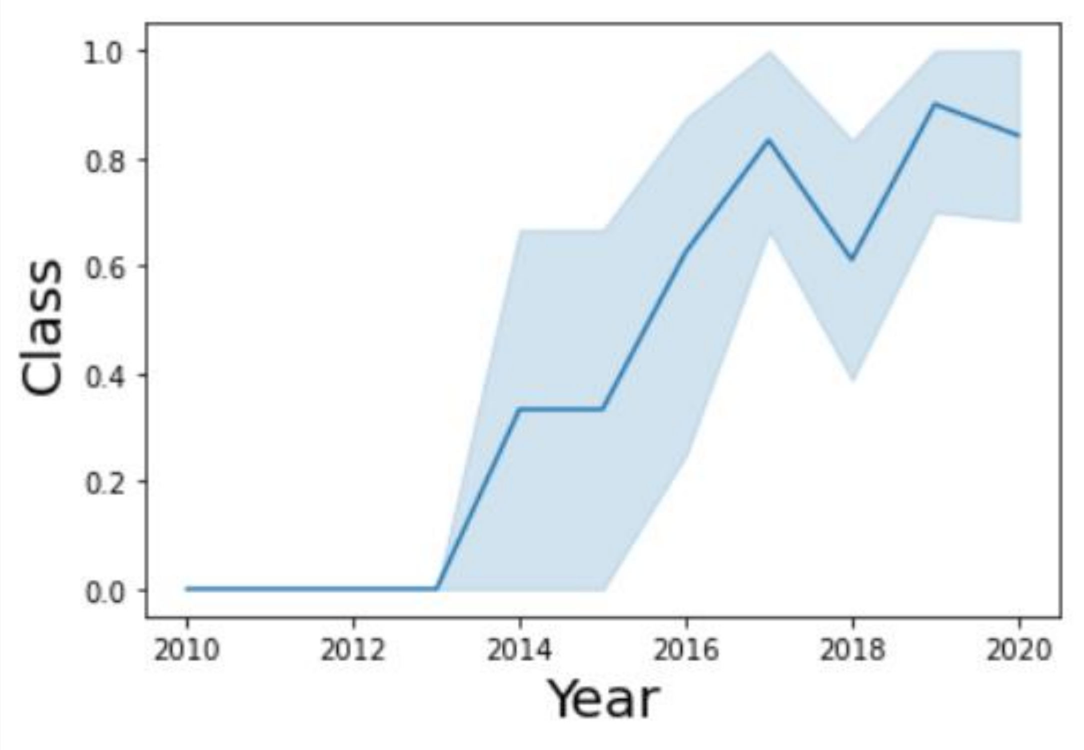


Scatter plot above shows Payload Mass (Kg) vs. Orbit type. Thus, combines data of success/failure per Payload Mass related to the Orbit. Here we also see a low informational gain on this relationship.

Launch Success Yearly Trend

- To the right we see a line chart of yearly average success rate.

- It helps us understand there is a positive trend starting at year 2014 with a slight fall by 2018 but still an overall positive trend for successful launches.



All Launch Site Names

- At the image below we can find the names of the unique launch sites:

Display the names of the unique launch sites in the space mission

```
unique_launch_sites = df.Launch_Site.unique()
unique_launch_sites
```

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

- Thus, the query result helps us understand there are only 4 launch sites in total.

Launch Site Names Begin with 'CCA'

- At the image below we can find 5 records where launch sites begin with `CCA`:

```
0    CCAFS LC-40
1    CCAFS LC-40
2    CCAFS LC-40
3    CCAFS LC-40
4    CCAFS LC-40
Name: Launch_Site, dtype: object
```

- Thus, the query result helps us understand the first CCA launch sites displayed.

Total Payload Mass

- At the image below we can see the calculated total payload carried by boosters from NASA:

```
Total payload mass carried by boosters launched by NASA (CRS) is:  619967 kg
```

- Thus, the query result helps us understand total amount of kilograms that have been attempted to go into space.

Average Payload Mass by F9 v1.1

- At the image below we can see the calculate the average payload mass carried by booster version F9 v1.1:

```
The average payload mass carried by booster version F9 v1.1 is:  PAYLOAD_MASS__KG_    2928.4
dtype: float64 kg
```

- Thus, the query result helps us understand average amount of kilograms that have been attempted to go into space per F9 v1.1 boosters.

First Successful Ground Landing Date

- At the image below we can see the dates of the first successful landing outcome on ground pad:

	Date	Class
8	2014-04-18	1
9	2014-07-14	1

- Thus, the query result helps us understand timeline of the successful landing outcomes started on April 4, 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

- At the image below we can see the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

	Booster_Version	Landing _Outcome	Class	PAYLOAD_MASS__KG_
27	F9 FT B1026	Success (drone ship)	1	4600
23	F9 FT B1022	Success (drone ship)	1	4696
42	F9 FT B1031.2	Success (drone ship)	1	5200
31	F9 FT B1021.2	Success (drone ship)	1	5300

- Thus, the query result helps us understand there are four boosters that 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

- At the image below we can see the calculated total number of successful and failure mission outcomes:

```
Total number of successful mission outcomes is:  100
Total number of failure mission outcomes is:  1
```

- Thus, the query result helps us understand that only 1 in 101 mission was unsuccessful which means less than 1% is not.

Boosters Carried Maximum Payload

- At the image below we can see the list the names of the booster which have carried the maximum payload mass:

- Thus, the query result helps us understand that a significant amount boosters have the maximum payload mass of 15,600 kg.

	Booster_Version	PAYLOAD_MASS__KG_
90	F9 B5 B1049.6	15440
82	F9 B5 B1048.5	15600
83	F9 B5 B1051.4	15600
93	F9 B5 B1058.3	15600
74	F9 B5 B1048.4	15600
92	F9 B5 B1060.2	15600
80	F9 B5 B1056.4	15600
77	F9 B5 B1049.4	15600
94	F9 B5 B1051.6	15600
95	F9 B5 B1060.3	15600
85	F9 B5 B1049.5	15600
79	F9 B5 B1051.3	15600
99	F9 B5 B1049.7	15600

2015 Launch Records

- At the image below we can see the list the failed *landing_outcomes* in drone ship, their booster versions, and launch site names for in year 2015:

	Booster_Version	Landing _Outcome	Class	PAYLOAD_MASS__KG_
16	F9 v1.1 B1015	Failure (drone ship)	0	1898
13	F9 v1.1 B1012	Failure (drone ship)	0	2395
20	F9 v1.1 B1017	Failure (drone ship)	0	553
21	F9 FT B1020	Failure (drone ship)	0	5271
25	F9 FT B1024	Failure (drone ship)	0	3600

- Thus, the query result helps us understand that 5 landing outcomes unsuccessful were unsuccessful using drone ship.

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

- At the image to the right, we can see the ranked 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:

- Thus, the query result helps us understand success and failure at a given snapshot in time.

	Date	Landing _Outcome	Class
0	2010-04-06	Failure (parachute)	0
1	2010-08-12	Failure (parachute)	0
2	2012-05-22	No attempt	0
3	2012-08-10	No attempt	0
4	2013-01-03	No attempt	0
6	2013-03-12	No attempt	0
5	2013-09-29	Uncontrolled (ocean)	0
8	2014-04-18	Controlled (ocean)	1
10	2014-05-08	No attempt	0
7	2014-06-01	No attempt	0
11	2014-07-09	No attempt	0
9	2014-07-14	Controlled (ocean)	1
12	2014-09-21	Uncontrolled (ocean)	0
15	2015-02-03	No attempt	0
16	2015-04-14	Failure (drone ship)	0
17	2015-04-27	No attempt	0
18	2015-06-28	Precluded (drone ship)	0
13	2015-10-01	Failure (drone ship)	0
14	2015-11-02	Controlled (ocean)	1
19	2015-12-22	Success (ground pad)	1
20	2016-01-17	Failure (drone ship)	0
21	2016-04-03	Failure (drone ship)	0
24	2016-05-27	Success (drone ship)	1
23	2016-06-05	Success (drone ship)	1
25	2016-06-15	Failure (drone ship)	0
26	2016-07-18	Success (ground pad)	1
22	2016-08-04	Success (drone ship)	1
27	2016-08-14	Success (drone ship)	1
32	2017-01-05	Success (ground pad)	1
28	2017-01-14	Success (drone ship)	1
29	2017-02-19	Success (ground pad)	1
34	2017-03-06	Success (ground pad)	1
30	2017-03-16	No attempt	0

Section 3

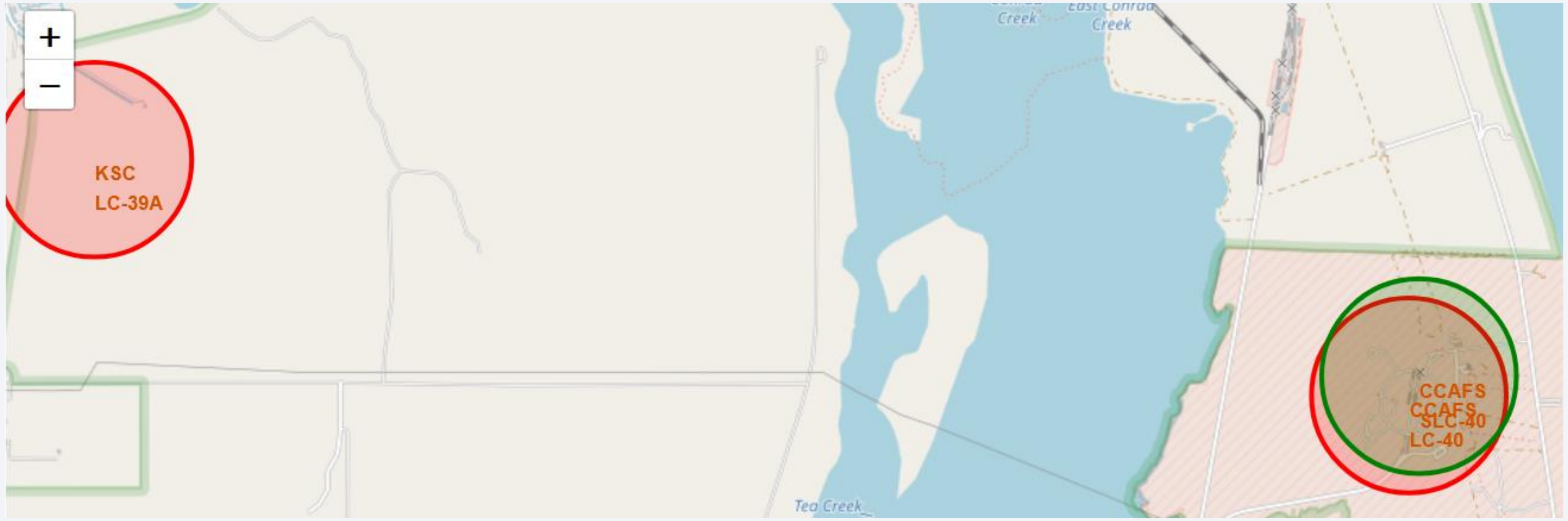
Launch Sites Proximities Analysis

All Site Locations Markers on Global Map

Per the map displayed above we can pinpoint all launch site locations are in the USA both at east *(3 locations)* and west *(just 1 location -VAFB SLC-4E-)*.

They also share the commonality of both being located at the south which makes them closer to the equator for physics reasons and that both are near the coast for safety reasons.

East Locations Markers Observation

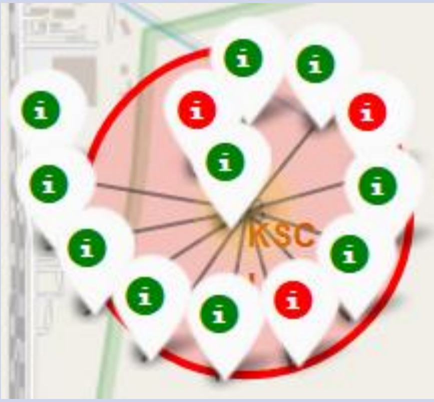
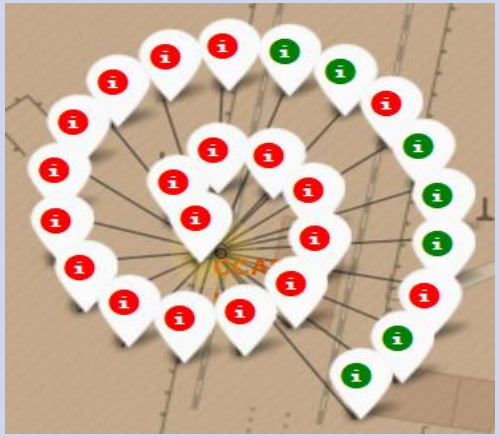
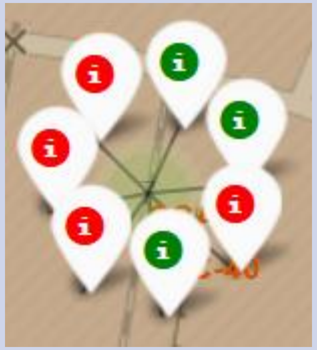


The east locations are in turn very close to each other.

CCAFS LC-40 and CCAFS SLC-40 are so close to each other that the marker on the map overlaps so only with close zoom in can we tell them apart.

Color-labeled Launch Outcomes maps

We can appreciate on the charts below a visual display of successful and unsuccessful launchings per site. Therefore, spotting that East and West do not mark a difference, yet specific sites do for best (KSC LC-39A) and worst (CCAFS LC-40) outcomes.

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

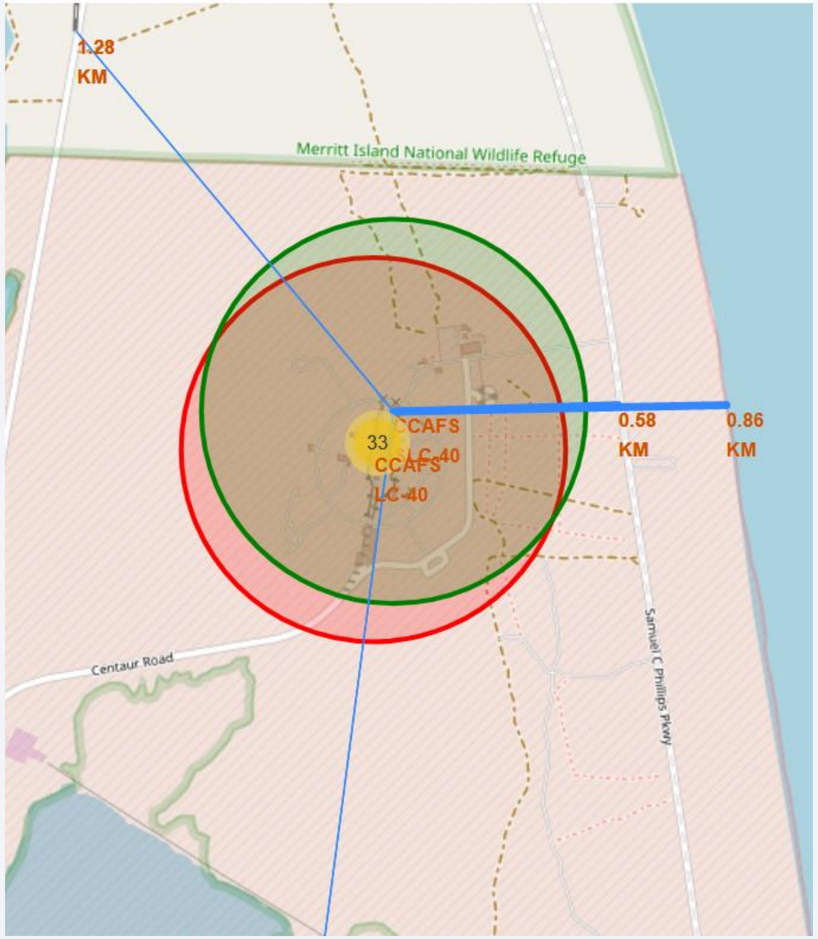
Site Proximities Map

As we can see at the chart to the right, what is close to the launch sites which as expected, cities are far away for safety reasons for rocket not to land near people.

And coastlines for safety reasons as well are far away for rocket not to land near people.

And highways, and railroads are close for materials handling.

Interesting yet not as deterministic for our core subject of successful launchings.



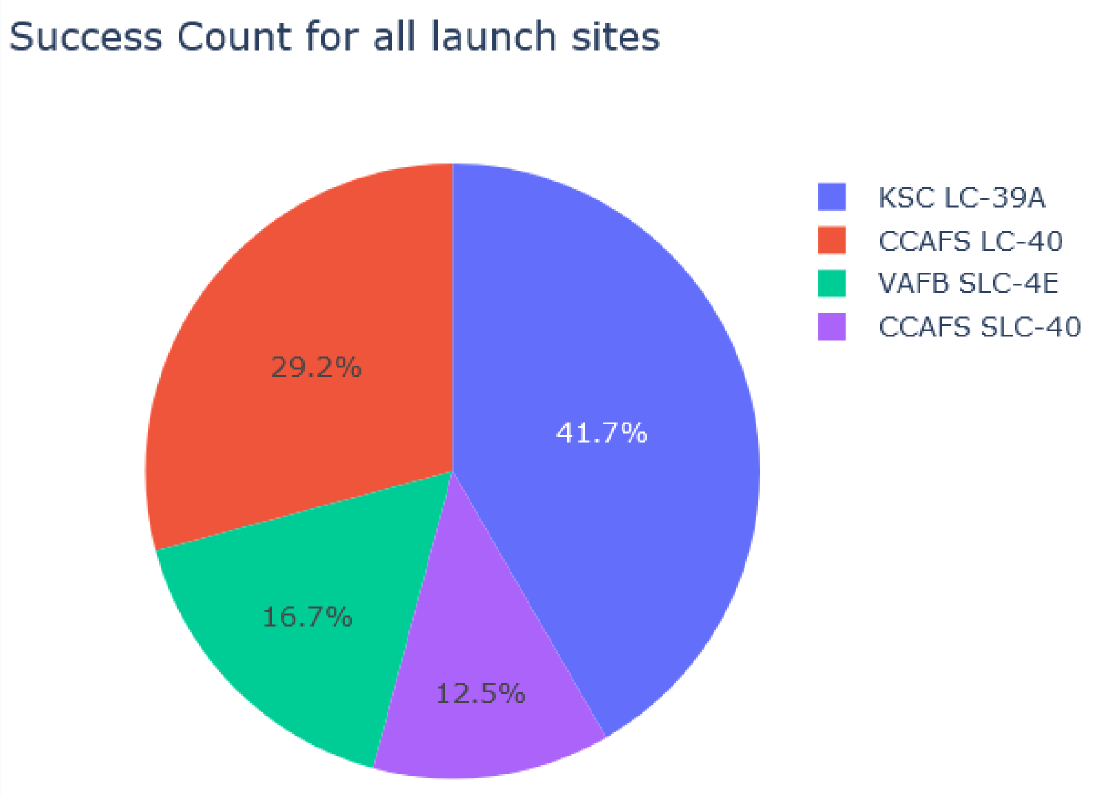
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count per Site

The chart on the right displays the launch success count for all sites in a pie chart format.

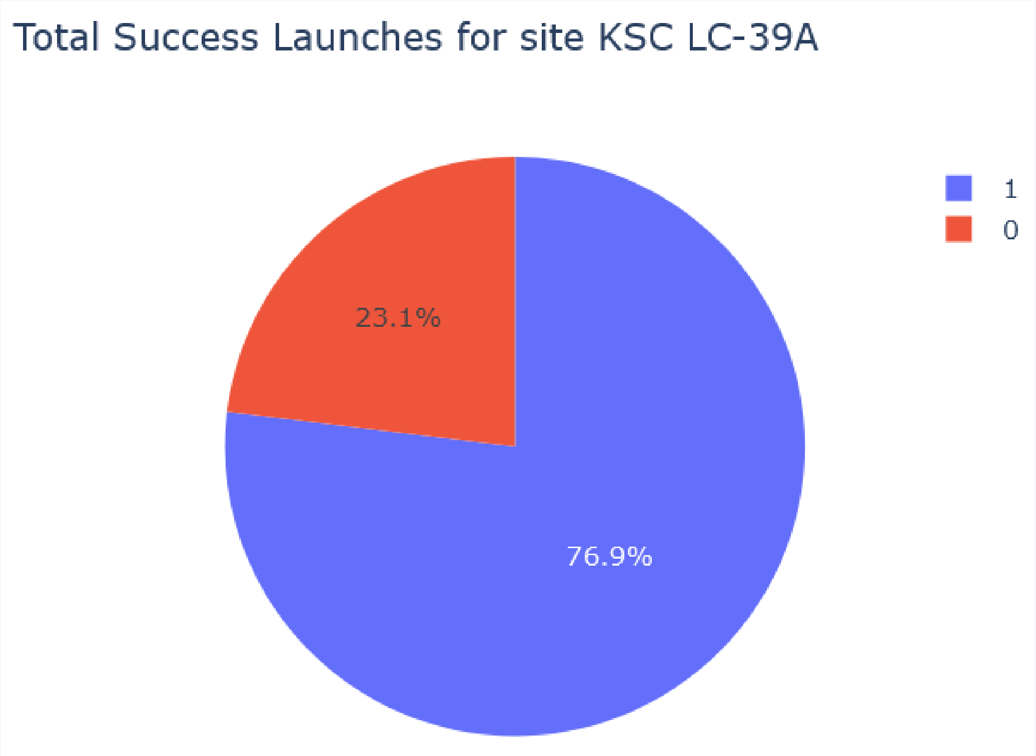
We can visually explore that it is not a successfully distributed scenario. KSC LC-39A represents 41.7% of all successful launches. Quite the opposite, CCAFS SLC-40 only 12.5%. Therefore, launch site matters a lot.



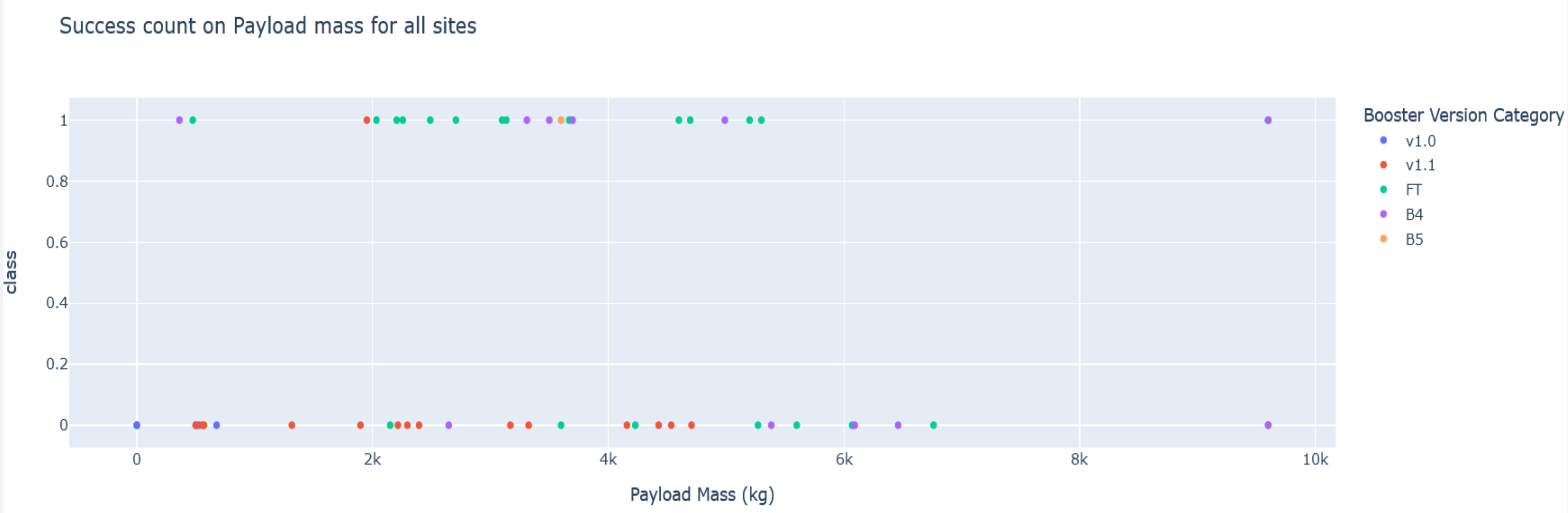
Launch Site with Highest Launch Success

So now we see on the right that drilling deeper into the more successful launching site there is a 76.9% success rate within that site.

Therefore, site selection as KSC LC-39A increases probability for a successful outcome.



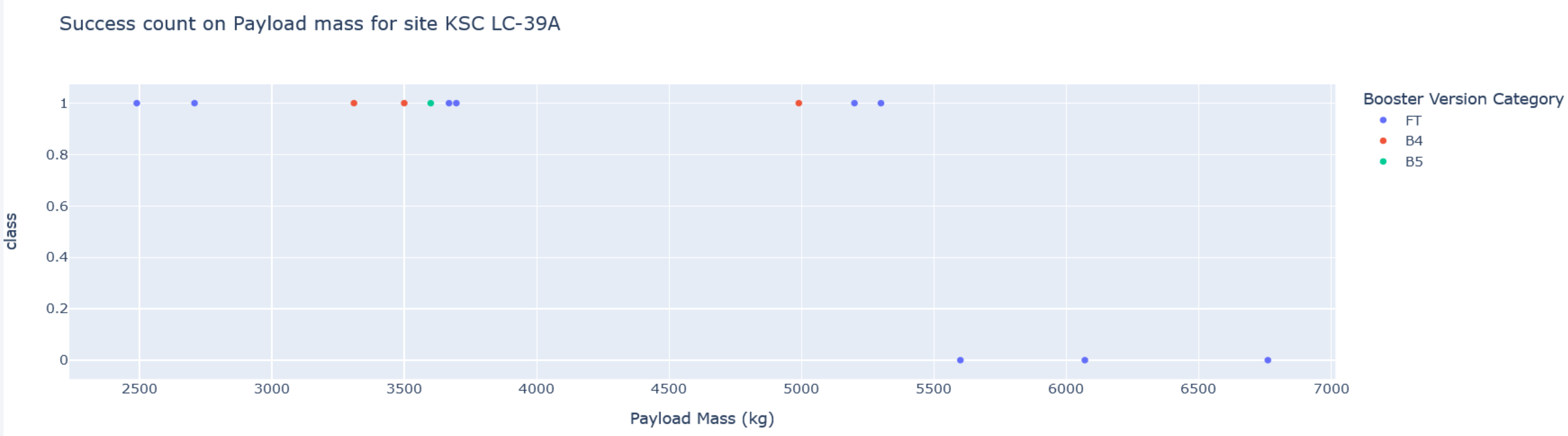
Overall Payload Vs Launch Outcome Review



Per chart above with Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider we can appreciate that Booster Version **V1.1 has a low success outcome** hindering success rate regardless of payload.

Plus, once payload is beyond the 5,500 kg the success rate declines sharply.

Specific Payload Vs Launch Outcome Review



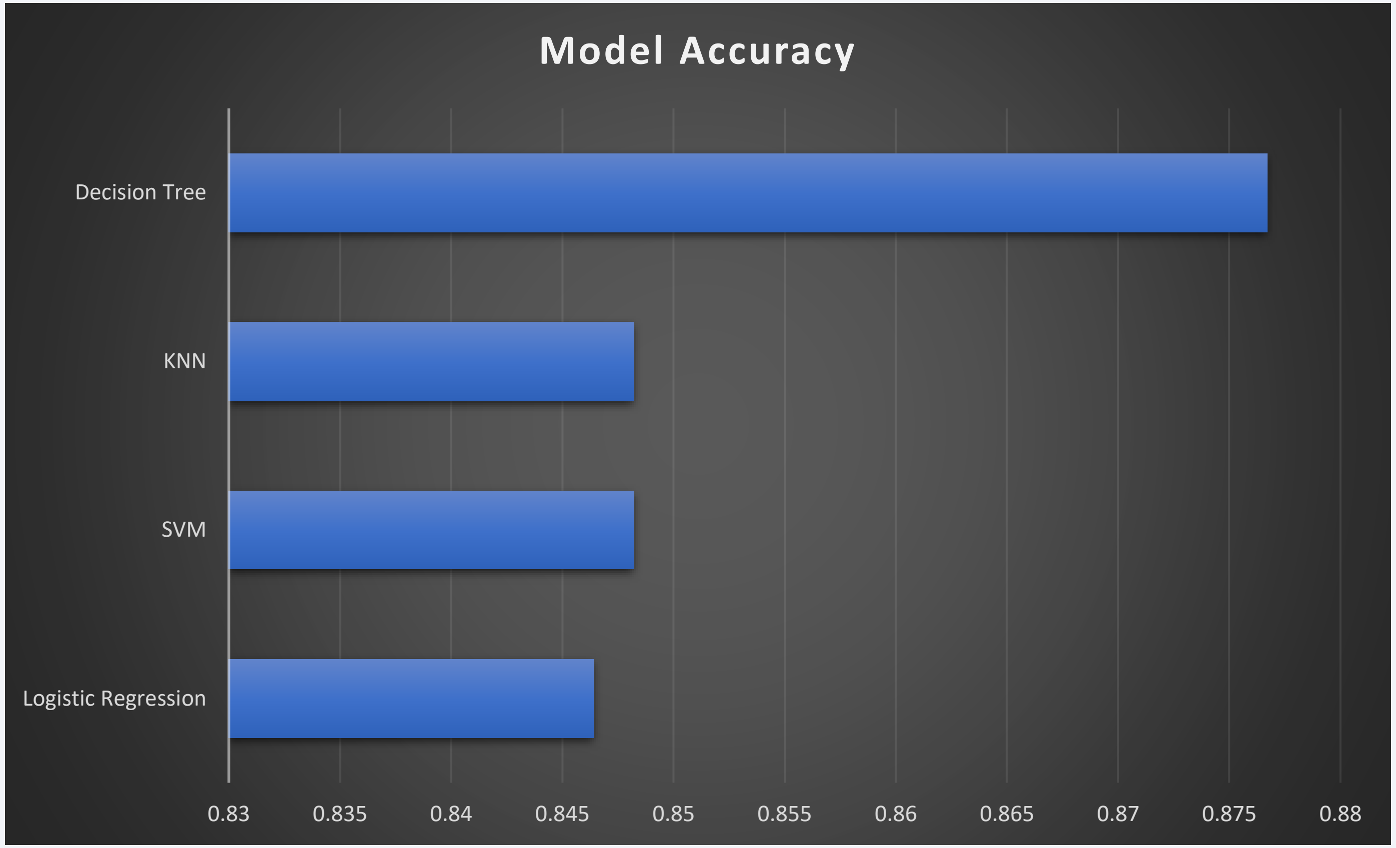
Specific for KSC LC-39A Payload vs. Launch Outcome scatter plot above, we first realize 2 boosters do not launch here which might relate to the success rate.

More important still, **if payload is below 5,500 kg, we can appreciate a high proximity to 100% success rate with any of the FT, B4 or B5 Booster Versions at this site.**

Section 5

Predictive Analysis (Classification)

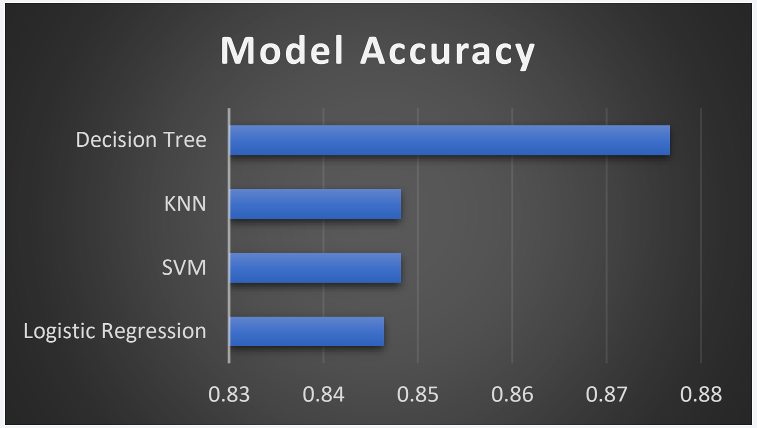
Classification Accuracy



On the chart above we can visualize the built model accuracy for all built classification models, in a bar chart and therefore, point out that the model that has the highest classification accuracy once hyperparameters are tuned is Decision Tree.

Classification Accuracy Observation

While indeed the Decision Tree expects a better outcome per the graph to the right, the 4 models *(Logistic Regression, SVM, Decision Tree, and KNN)* ended up providing the same confusion matrix thus 83.3% accuracy regardless of the model per the outcome below.




```
Logistic regression accuracy      :  0.8333333333333334
SVM accuracy                      :  0.8333333333333334
Decision Tree regression accuracy:  0.8333333333333334
KNN accuracy                      :  0.8333333333333334
```

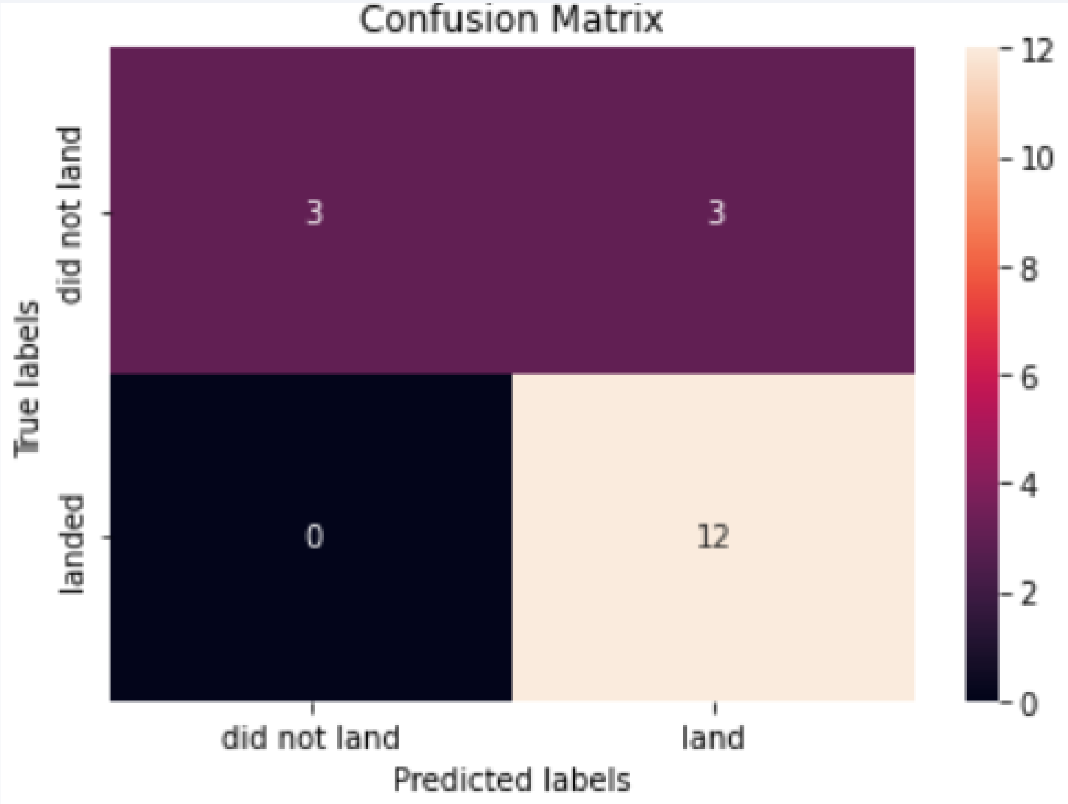
Believe this is due to the small data set of only 89 observations. Would expect that given a larger data set the accuracy over y hat will grow to the 88% level.

Confusion Matrix

On the chart to the right, we see the confusion matrix of the best performing model which helps us understand we do not have Type II errors.

We appreciate 15 properly predicted values and 3 false negatives *(predicted as not landed yet they did)*.

Per the former slide, this was the Confusion Matrix outcome using all 4 models.

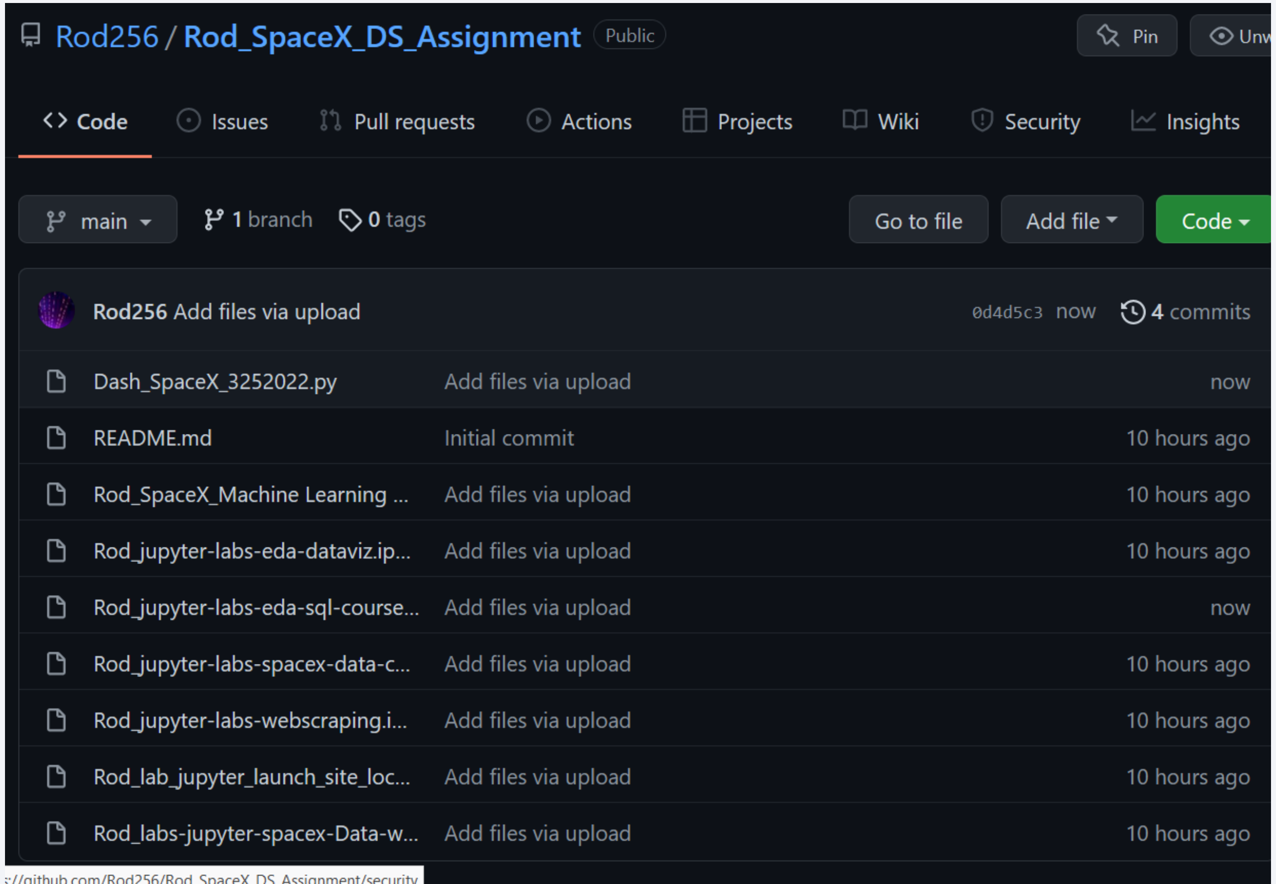


Conclusions

- Falcon 9 first stage will land successfully is expected 66.6% currently. However, if the following suggestions below are followed that number will raise:
    - Launch site to be KSC LC-39A.
    - Keep Payload Mass below 5,500 kg.
    - Use orbits as either ES-L1, SSO, HEO, or GEO.
    - Use booster as either FT, B4 or B5.
- All models are providing 83% accuracy. And none is displaying type II errors.
- To make predictions use the Decision Tree model with tuned hyperparameters as it has the highest accuracy.

Appendix

- All files, including the MS PowerPoint presentation saved as PDF are located at the GitHub repository for additional information:



Thank you!