IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Panagiotis Karaoulanis>
<22/11/2024>

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

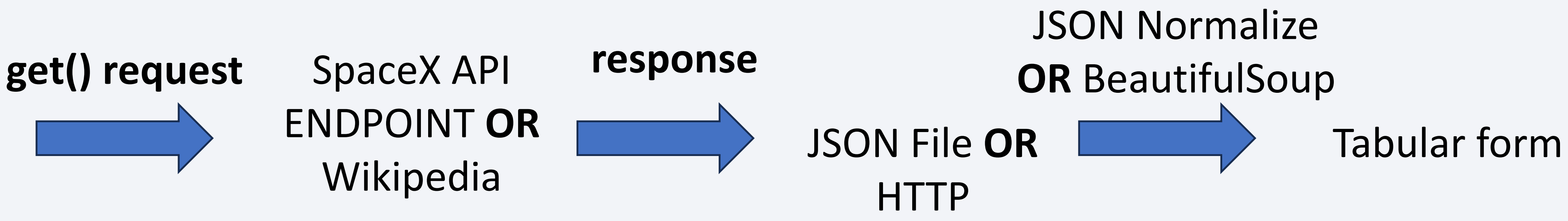
Methodology

Executive Summary

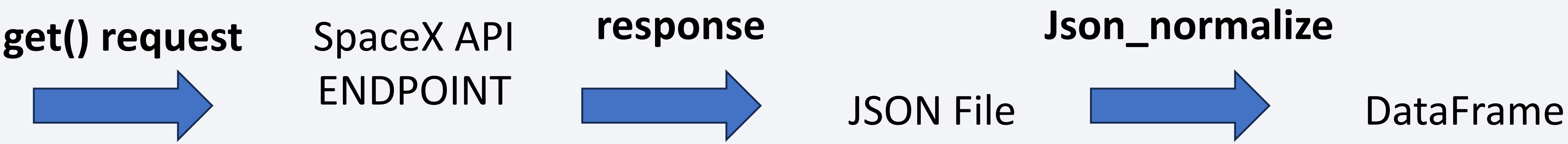
- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The primary method for collecting the data is through the SpaceX REST API
- Web scraping, from Wikipedia pages or similar sources that list Falcon 9 launches



Data Collection – SpaceX API



- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

get() request → Wikipedia **response** → HTTP BeautifulSoup → Tabular form

- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/jupyter-labs-webscraping.ipynb

Data Wrangling

- Creating a column in the dataframe to easily account for successful launches

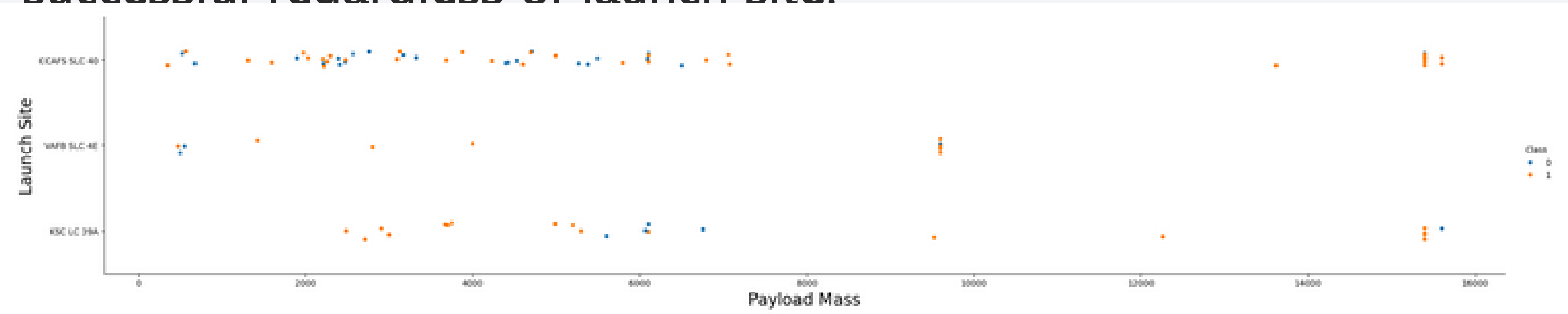
BINARY CLASSIFICATION OF LAUNCHES

DATAFRAME → **FINAL DATAFRAME**

- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots were used to visually determine the correlation (if any) between variables and to draw any conclusions regarding the relationships of the data features. For example when plotting launch site vs. Payload Mass we can easily conclude that after a certain weight the results tend to be more successful regardless of launch site.



- A bar plot was used to plot data that had a categorical value and a line chart was used to present data on a temporal axis

- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Query to create the table to work on
  - Query to display the names of the unique launch sites in the space mission
  - Query to display the average payload mass carried by booster version F9 v1.1
  - Query to list the total number of successful and failure mission outcomes
  - Query to List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - folium.Circle and folium.map.marker were used to add a highlighted circle area with a text label on a specific coordinate such as the coordinates of NASA or other launch sites. Using these objects can help the user identify the launch site and points of interest in a map. We could use these objects to show the number of successful launches in each site.
  - Lines were used to visually portray the distances between 2 points on the map. For example, the distance between a launch site and a highway or the coastline
- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/145a5ca567a8642a3ac6121b84bc50900a7f5239/lab_jupyter_launch_site_location(2).ipynb

Build a Dashboard with Plotly Dash

- Two elements controlled by the user were added:
  - Choosing launch site from a dropdown menu
  - Selecting payload mass range with a slider set from 0 to 10000 kg
- Two graphs were selected:
  - A **pie** chart that would display the success percentage attributed to each launch site (if in the dropdown menu the user selected "ALL") or the success rate in each launch site (if in the dropdown menu the user selected a specific launch site)
  - A scatter plot that would show the relationship between success and weight of the payload (either universally or for a given launch site)
- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/spacex_dash_app.py

Predictive Analysis (Classification)

1. Create a numpy array for the dependent variable which takes on values that indicate a success or failure **and** a dataframe holding the independent variables
2. Standardize the variables in the dataframe in order to perform predictive analysis
3. Split the data into test and train data
4. Train different models and use gridsearch algorithm to find best parameters for each model
5. Use a confusion matrix to quickly visualize the accuracy of the model
6. Calculate accuracy for each model given best parameters and choose the most accurate one

- https://github.com/PanosKaraoulanis/DATA-ANALYSIS-CAPSTONE/blob/0c064fc9b81b2ecae69775573a4b66641d09976f/SpaceX_Machine%20Learning%20Prediction_Part_5(1).ipynb

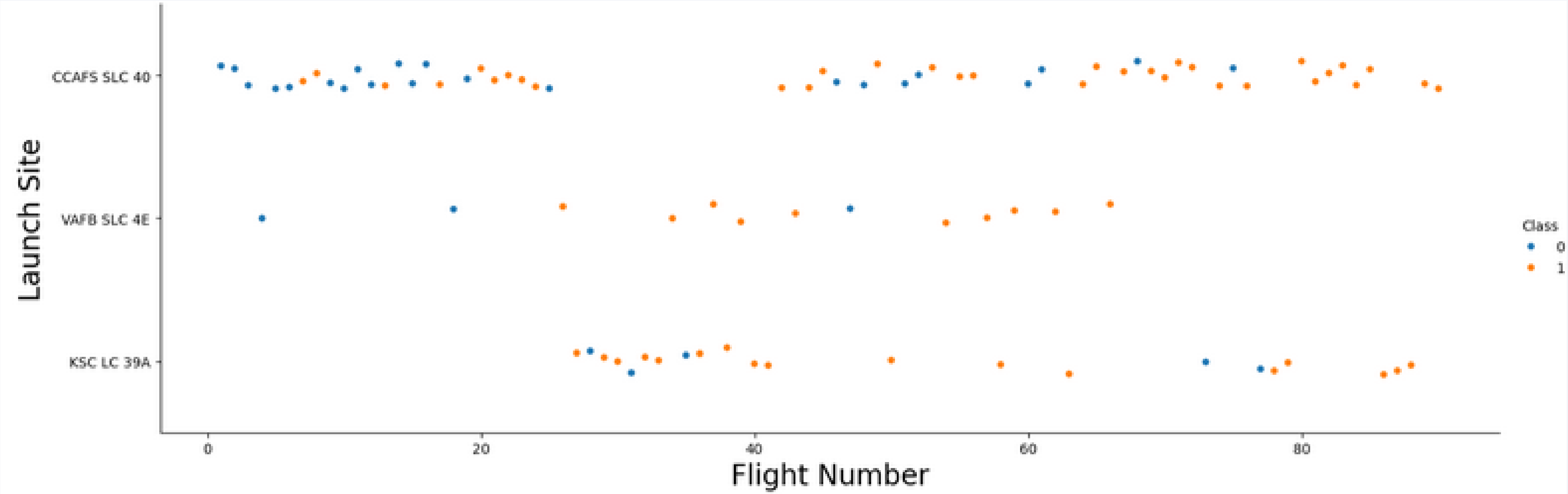
Results

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

Section 2

Insights drawn from EDA

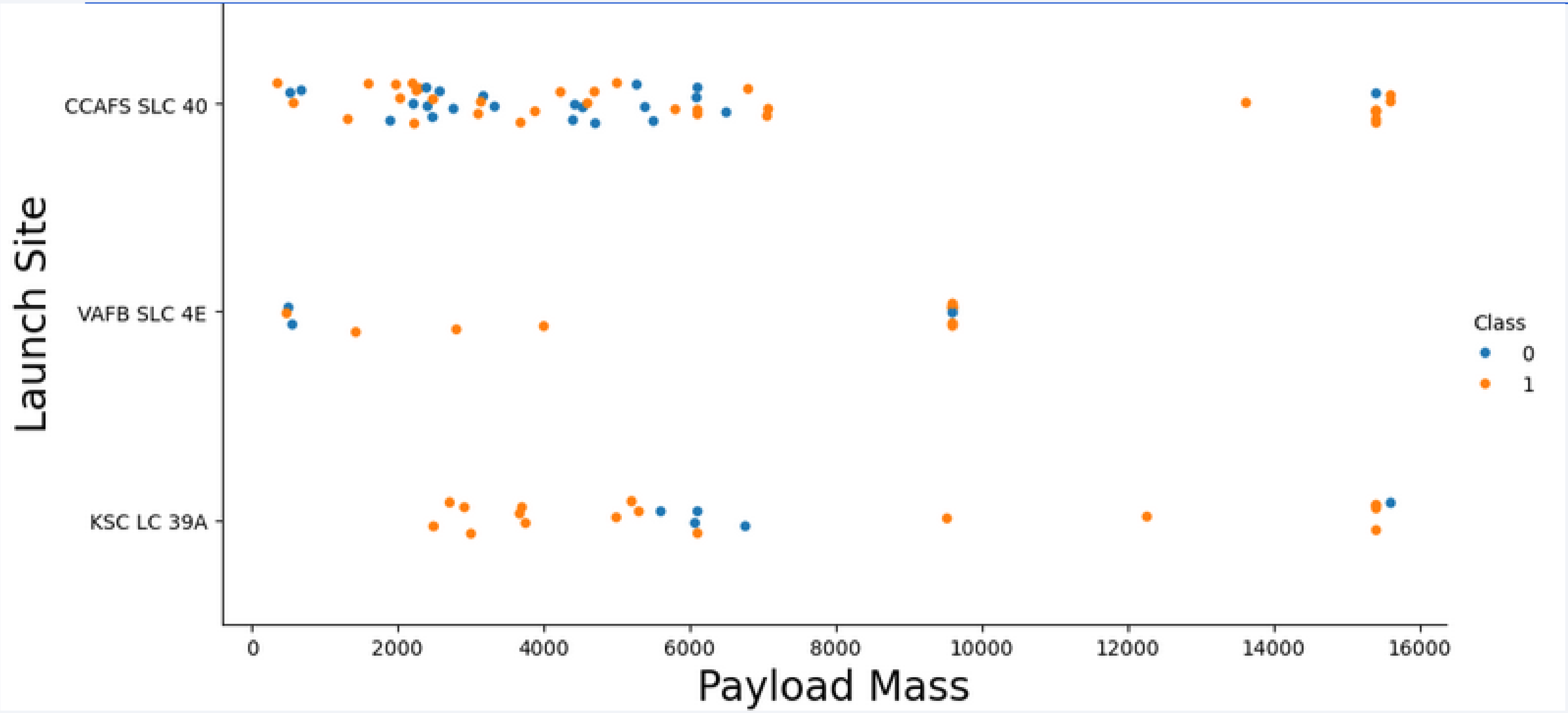
Flight Number vs. Launch Site



Explanation:

As flight number increases successes outnumber failures indiscriminately of location

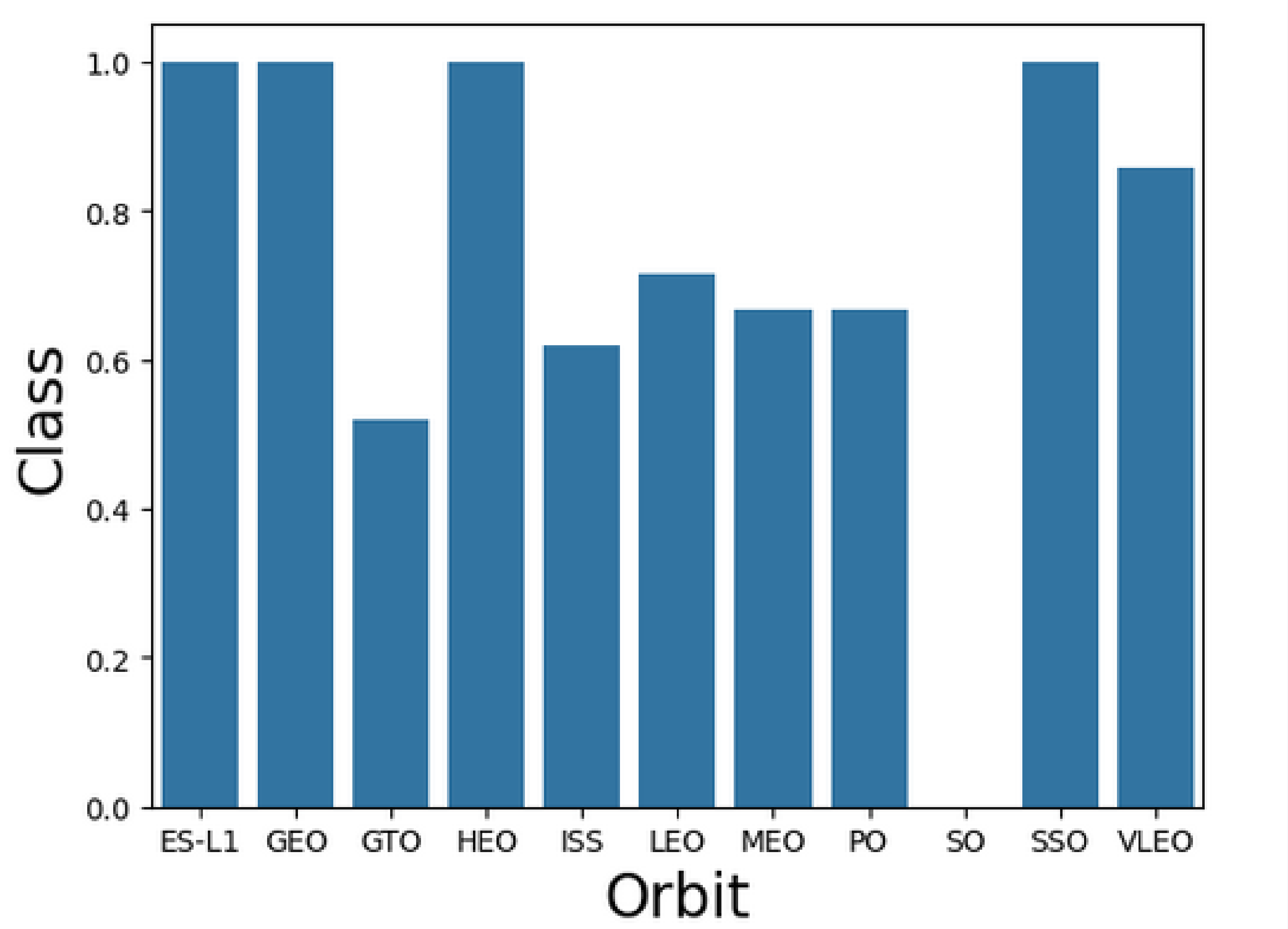
Payload vs. Launch Site



As payload mass increases successes outnumber failures indiscriminately of location

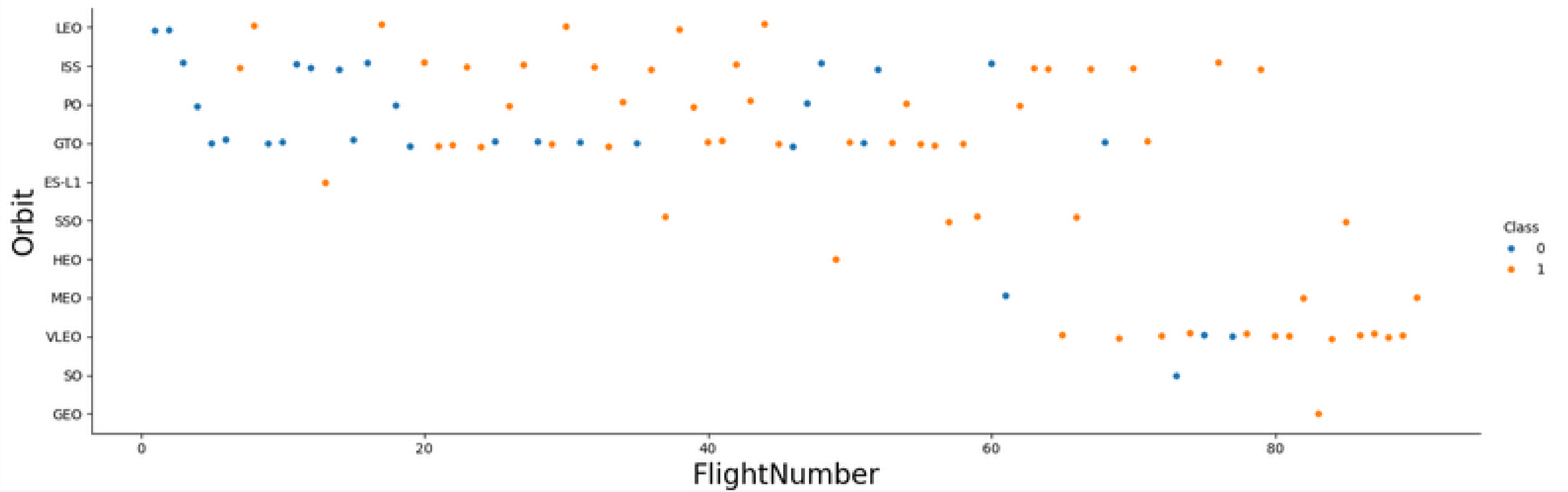
For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



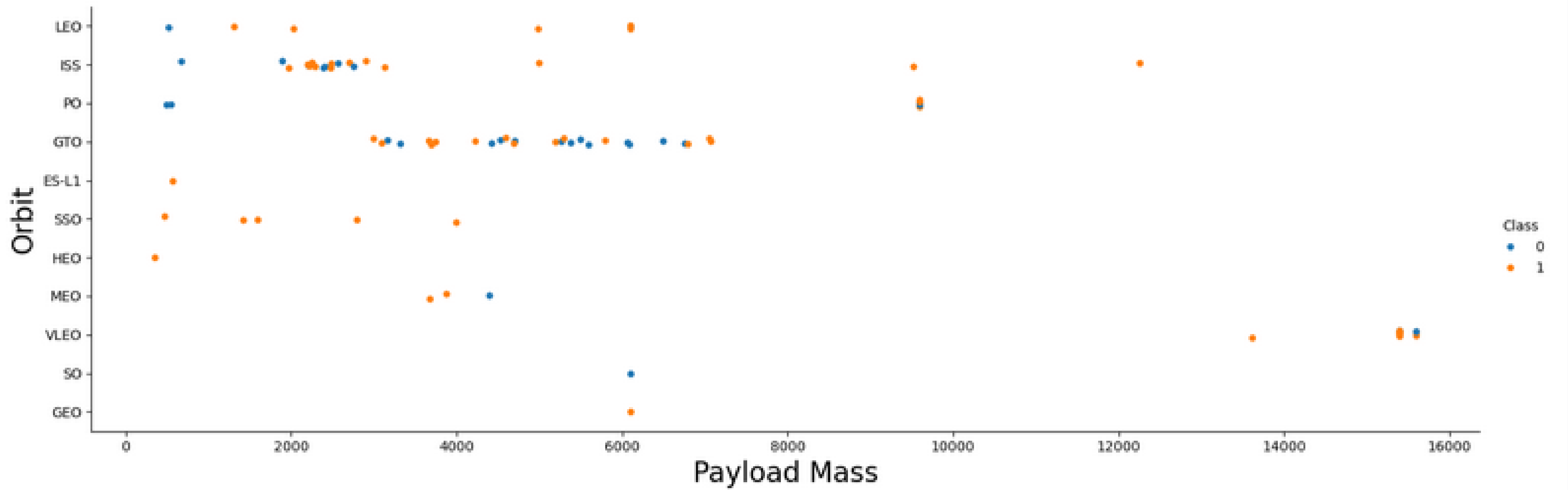
- ES-L1, GEO, HEO, SSO appear to be the orbits with the highest success rates

Flight Number vs. Orbit Type



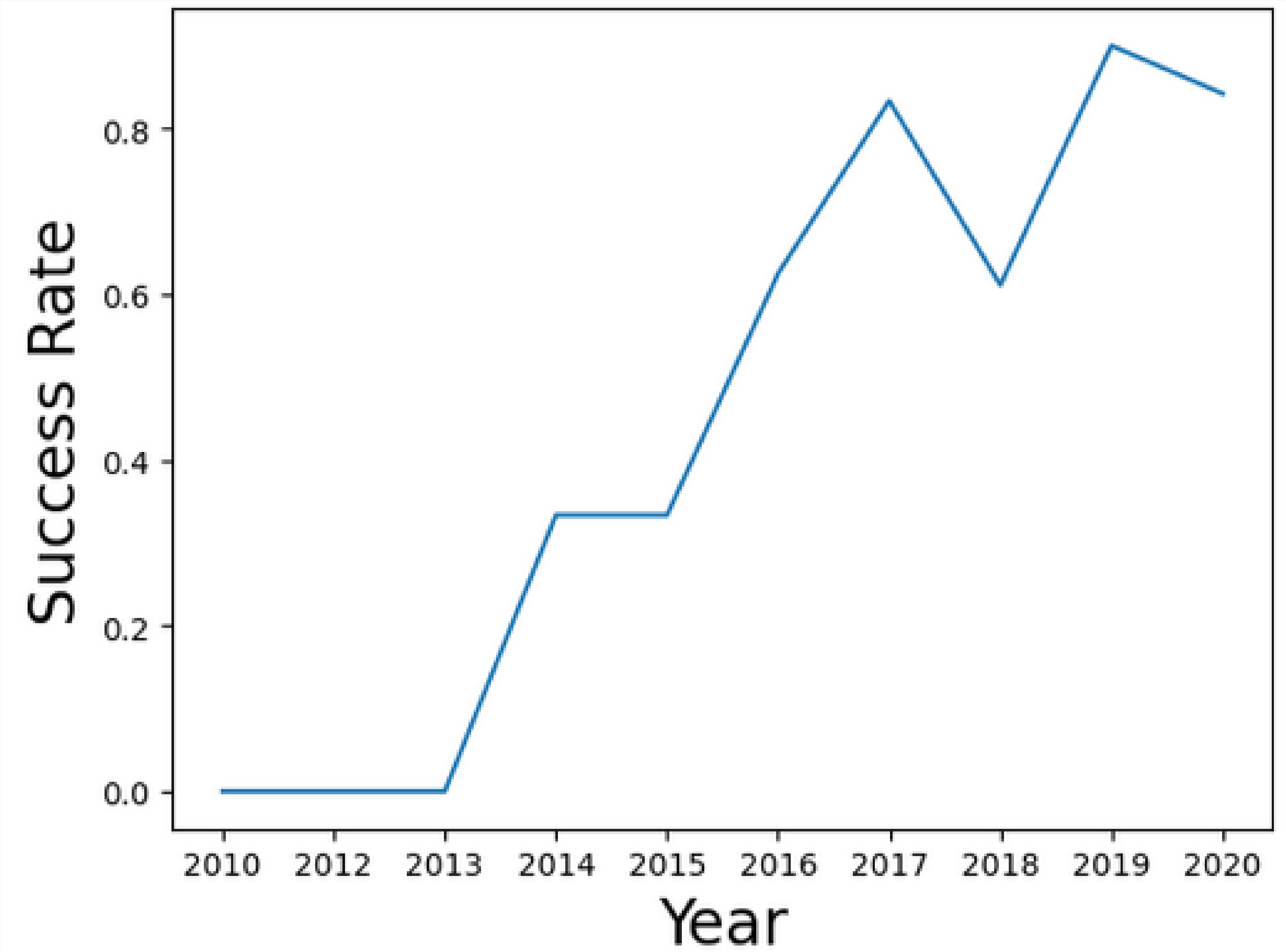
- GTO orbit appears to have lower success rate despite increasing Flight number which seems to be positively corelated to success of launch

Payload vs. Orbit Type



- GTO orbit presents mixed results concerning payload mass and success rate
- ISS orbit exhibits high success when handling lighter loads

Launch Success Yearly Trend



- There is noticeable growth in the success rate of launches as time passes

All Launch Site Names

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

QUERY USED IN SQL:

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

USED TO GET UNIQUE ROW ELEMENTS FROM THE COLUMN

COLUMN NAME

TABLE NAME

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Succe
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	Succe
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Succe
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Succe
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Succe

QUERY USED IN SQL:

```
%%sql

select * from SPACEXTABLE
where "Launch_Site" like "CCA%"
limit 5
```

RETURN EVERY COLUMN FROM TABLE

LIMITING SEARCH QUERY TO 5 RESULTS

FILTERING THE RESULTS USING THE SIMILARITY CLAUSE "LIKE"

Total Payload Mass

sum("PAYLOAD_MASS__KG_")
45596

QUERY USED IN SQL:

```
%%sql
select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE
where "Customer"=="NASA (CRS)"
```

RETURN THE SUM OF ELEMENTS FROM CHOSEN COLUMN

FILTERING THE RESULTS USING A BOOLEAN OPERATOR

Average Payload Mass by F9 v1.1

sum("PAYLOAD_MASS__KG_")/count(PAYLOAD_MASS__KG_)
2928

QUERY USED IN SQL:

RETURN THE AVERAGE OF ELEMENTS FROM CHOSEN COLUMN

```
%%sql
select sum("PAYLOAD_MASS__KG_")/count(PAYLOAD_MASS__KG_) from SPACEXTABLE
where "Booster_Version"=="F9 v1.1"
```

FILTERING THE RESULTS USING A BOOLEAN OPERATOR

First Successful Ground Landing Date

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

QUERY USED IN SQL:

RETURN 2 COLUMNS FROM TABLE

```
%%sql
select "Date","Landing_Outcome" from SPACEXTABLE
where "Landing_Outcome"=="Success (ground pad)"
order by "Date" asc
limit 1
```

FILTER RESULTS USING A BOOLEAN OPERATOR ON DESIRED COLUMN

RETURN RESULTS IN ASCENDING ORDER

RETURN 1ST RESULT

Successful Drone Ship Landing with Payload between 4000 and 6000

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

QUERY USED IN SQL:

RETURN 1 COLUMN FROM TABLE

```
%%sql
select "Booster_Version" from SPACEXTABLE
where "Landing_Outcome"=="Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

FILTER RESULTS USING A BOOLEAN OPERATOR AND THE BETWEEN-AND OPERATOR TO FILTER RESULTS BASED ON NUMERICAL VALUES

Total Number of Successful and Failure Mission Outcomes

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

QUERY USED IN SQL:

RETURN THE COLUMN FROM TABLE AND THE COUNT FOR EACH DISTINCT VALUE

```
%%sql
select "Mission_Outcome",count("Mission_Outcome") from SPACEXTABLE
group by "Mission_Outcome"
```

GROUP RESULTS BASED ON UNIQUE VALUES OF THE SELECTED COLUMN

Boosters Carried Maximum Payload

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

QUERY USED IN SQL:

USING DISTINCT TO GET UNIQUE VALUES FROM THE COLUMN

```
%%sql
select distinct"Booster_Version" from SPACEXTABLE
where "PAYLOAD_MASS__KG_"==(select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

SUBQUERY USED TO RETURN THE MAX VALUE

FILTER USING A BOOLEAN OPERATOR AND THE "RIGHT HAND SIDE" IS GIVEN BY ANOTHER QUERY

2015 Launch Records

substr("Date", 6,2)	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

QUERY USED IN SQL:

FUNCTION USED TO EXTRACT THE MONTH FROM THE DATE COLUMN

```
%%sql
select  substr("Date", 6,2),"Landing_outcome","Booster_Version","Launch_Site" from SPACEXTABLE
where "Landing_outcome"=="Failure (drone ship)" and substr(Date,0,5)="2015"
```

FILTER USING A BOOLEAN OPERATOR

FUNCTION USED TO EXTRACT THE YEAR FROM THE DATE COLUMN

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

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

QUERY USED IN SQL:

RETURN THE COLUMN FROM TABLE AND THE COUNT FOR EACH DISTINCT VALUE

```
%%sql
select  "Landing_outcome",count("Landing_outcome") from SPACEXTABLE
where "Date" between "2010-06-04" and "2017-03-20"
group by "Landing_outcome"
```

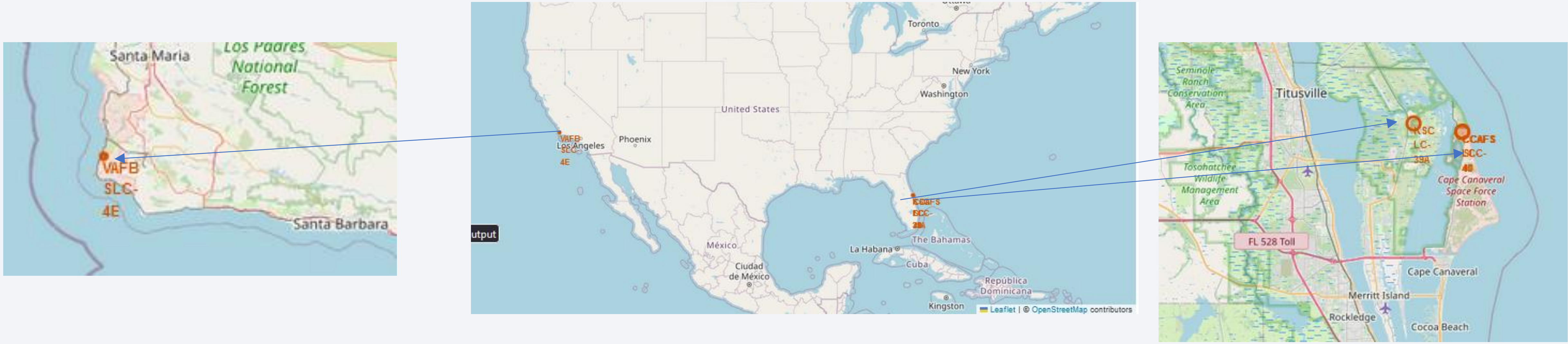
THE BETWEEN-AND OPERATOR TO FILTER RESULTS BASED ON NUMERICAL VALUES

GROUP RESULTS BASED ON UNIQUE VALUES OF THE SELECTED COLUMN

Section 3

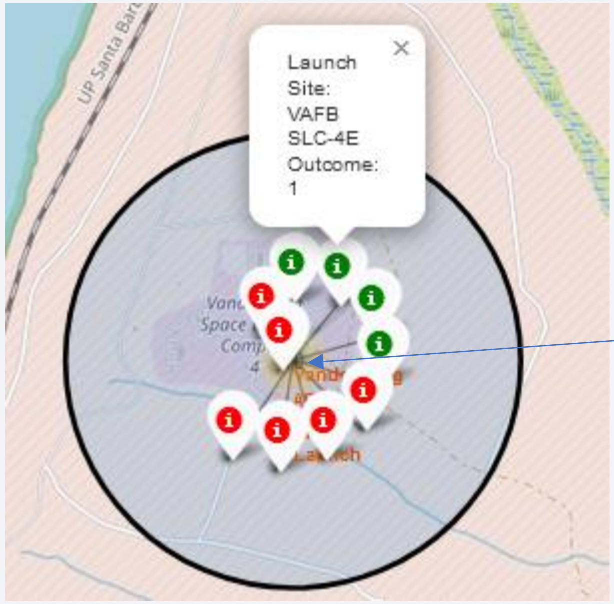
Launch Sites Proximities Analysis

Launch sites in USA



Launch sites appear to be near the coastline of the continental USA and near the equator line.

Visualizing the successes/failures for each launch site



By adding markers for each launch we can easily examine which site has a highest success rate

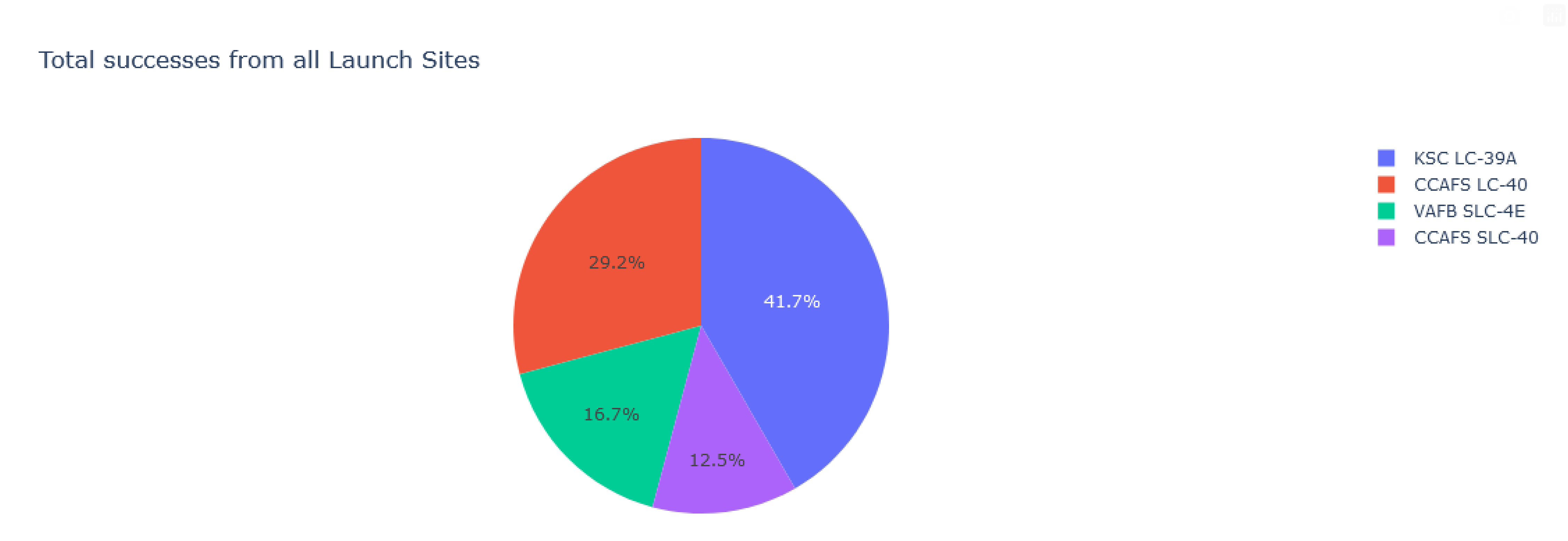
For example: the KSC LC-39A launch site seems to be linked with a higher success rate in comparison to the VAFB SLC-4E launch site

Visualizing distance from coastline

Section 4

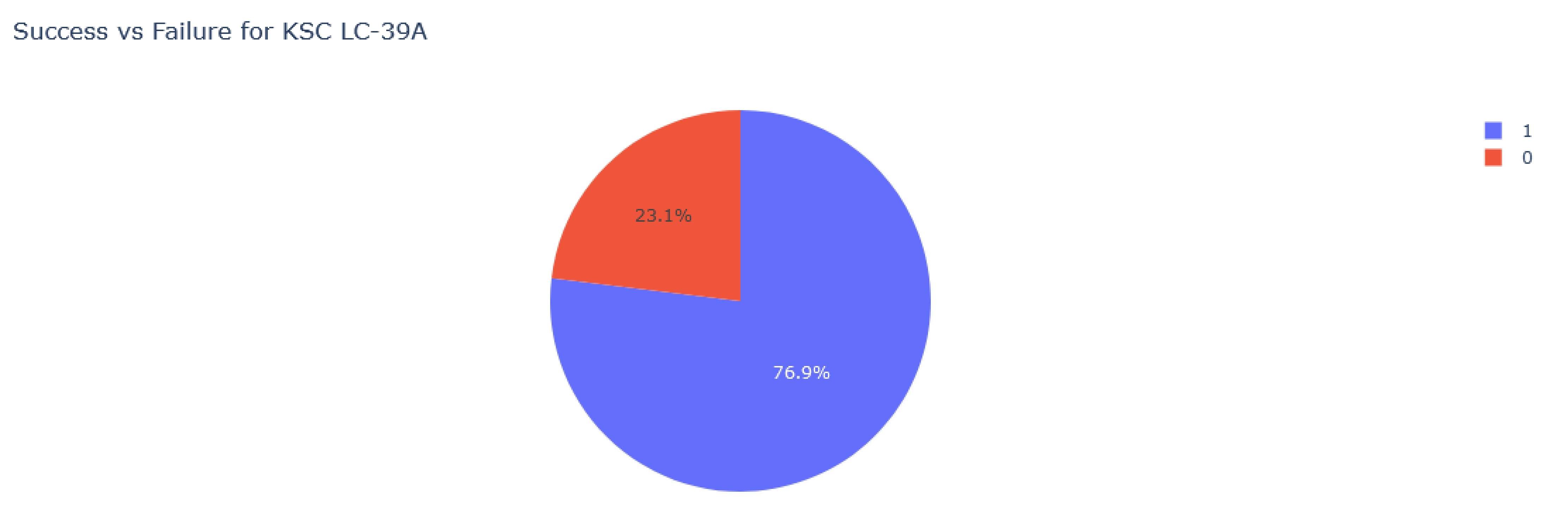
Build a Dashboard with Plotly Dash

Success percentage for each site



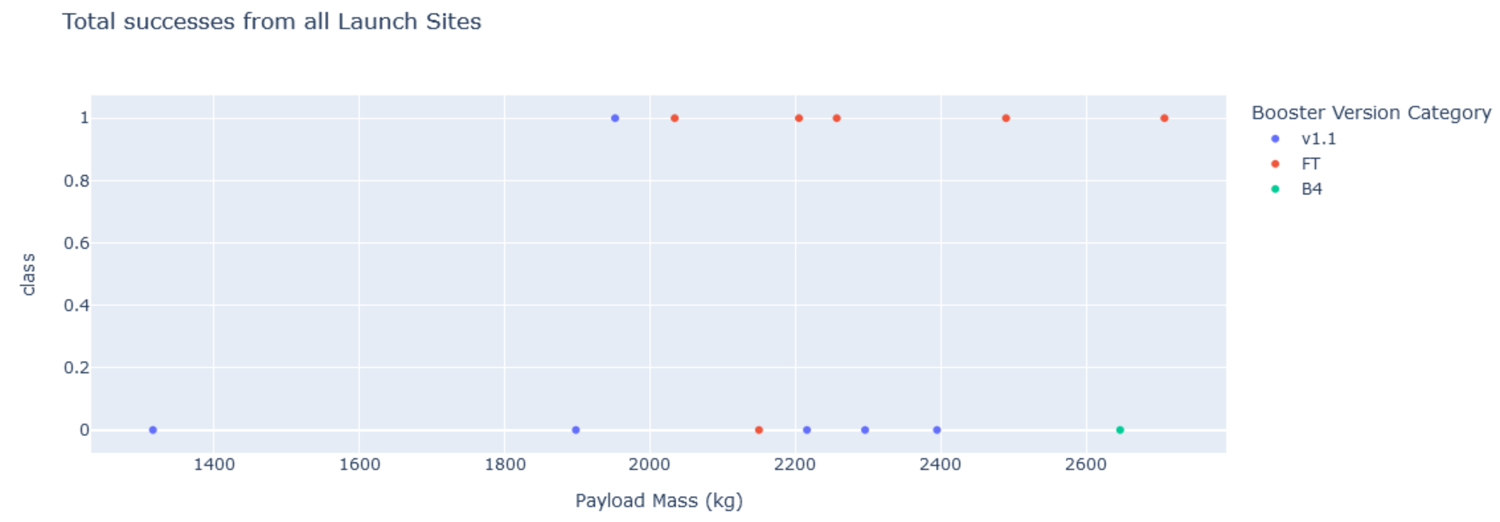
KSC-LC-39A contributes the most to successful launches

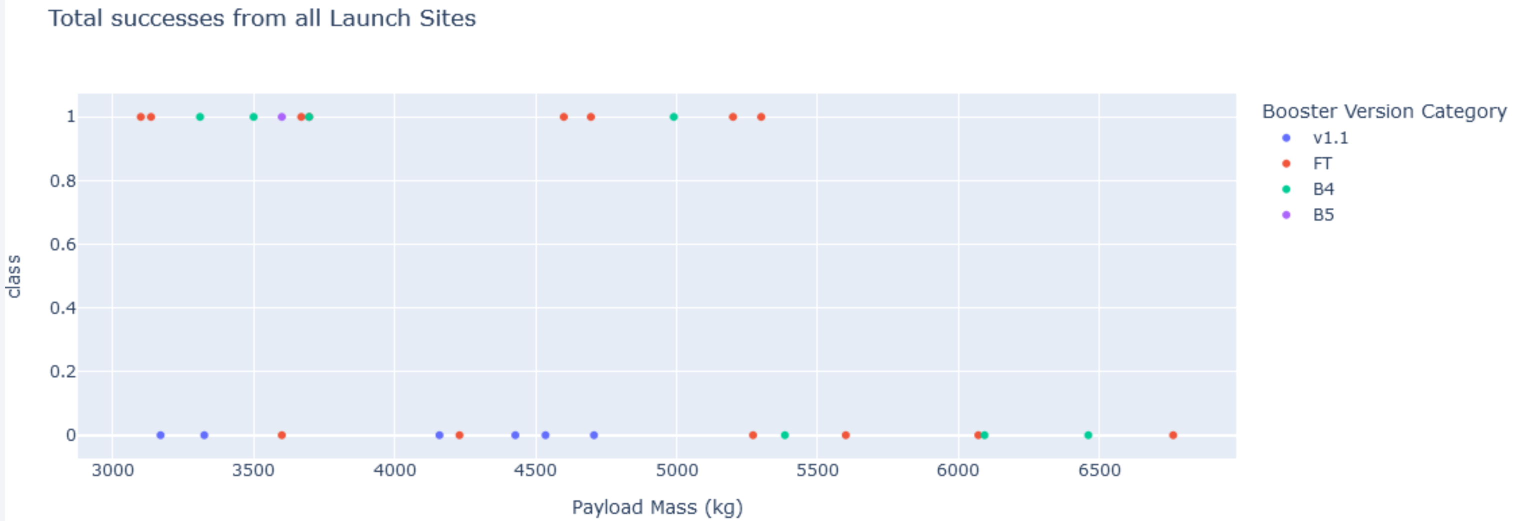
Success vs Failure for KSC LC-39A



Success is 3 times more likely for the launch site KSC LC-39A

Comparing successes for all sites on different weight ranges



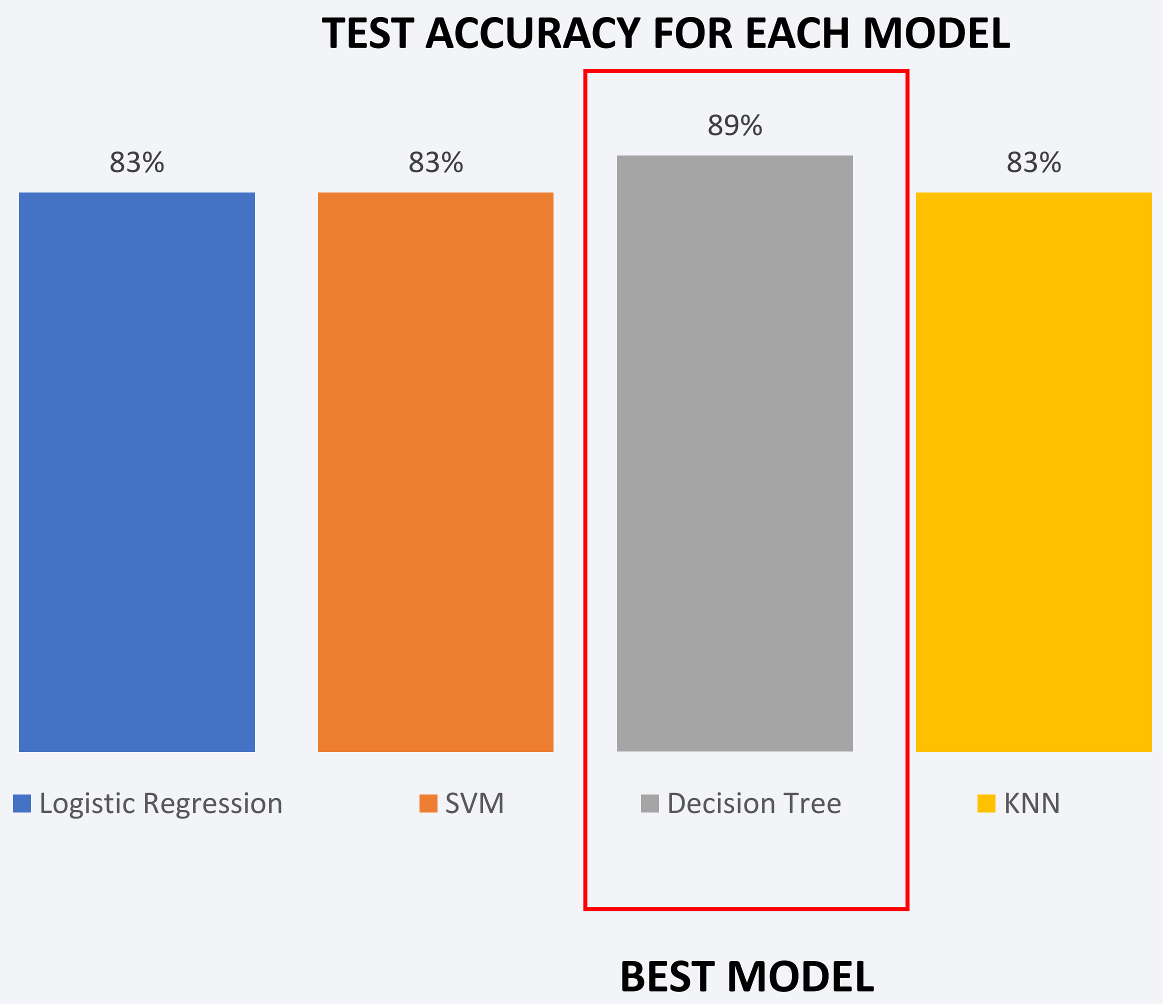


As weight increases the launches become more successful

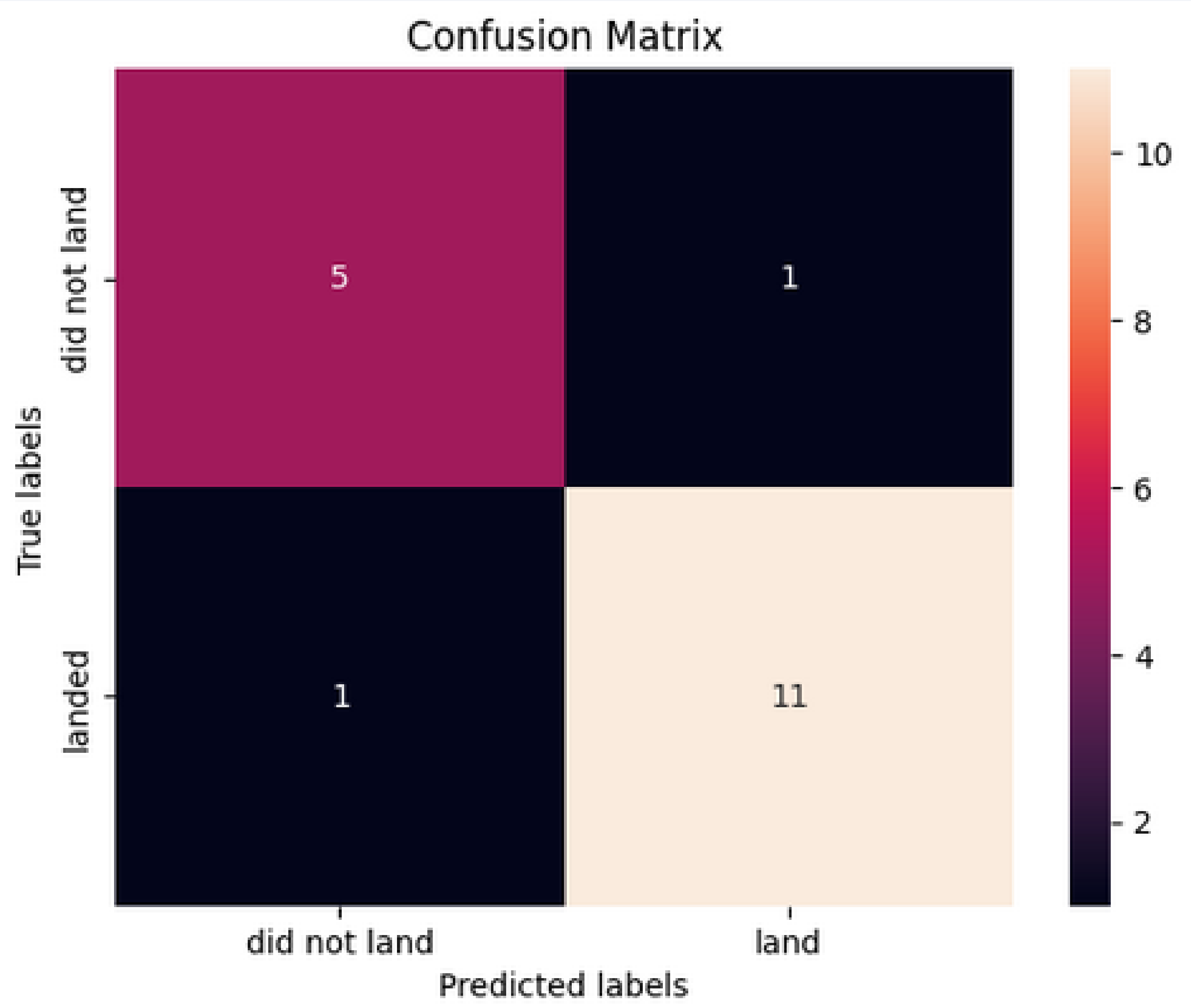
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Out of 18 test cases 16 were correctly identified.
1 case of false positive
1 case of false negative

Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!