IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gasner J. Barthold
10/01/2021

Outline

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

Executive Summary

Summary of methodologies

- Falcon 9 launch data collected by SpaceX API, and web scrapping
- CSV file of rocket data uploaded to IBM Watson for analysis
- Data cleaning, replacing missing values with either mean of that table or 0.
- Perform EDA and used data visualization tools on data
- Train and evaluate model to assess success rate of future ladings.

Summary of all results

- Rockets with higher payload have a greater landing success rate
- Among 11 orbit types ES L1, GEO, HEO, SSO were 100 successful with less than 6000 kg payload.
- SpaceX has 4 launch sites, one is near California, the other three is near Florida and South Texas. All the sites are in near proximity to ocean and all the sites are bit far away from the city.
- All models achieved similar results in the low to mid 80%.

Introduction

Project background and context

- Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond.
- Falcon 9 is the world's first orbital class reusable rocket. Reusability allows SpaceX to refly the most expensive parts of the rocket, the first stage, which in turn drives down the cost of space access.
- Reusability allows SpaceX to undercut competitors by 100 million dollars or more per launch
- Falcon 9 first stage is capable of re-entering the atmosphere and landing vertically after separating from the second stage, but not all first stage landings are successful.

Problems we want to solve

- We will use data analysis and machine learning techniques to predict whether the first stage will land successfully
- Assess the features that contribute to a successful first stage landing
- Use models to predict success rate of new launches and calculate model's accuracy

Section 1

Methodology

Methodology

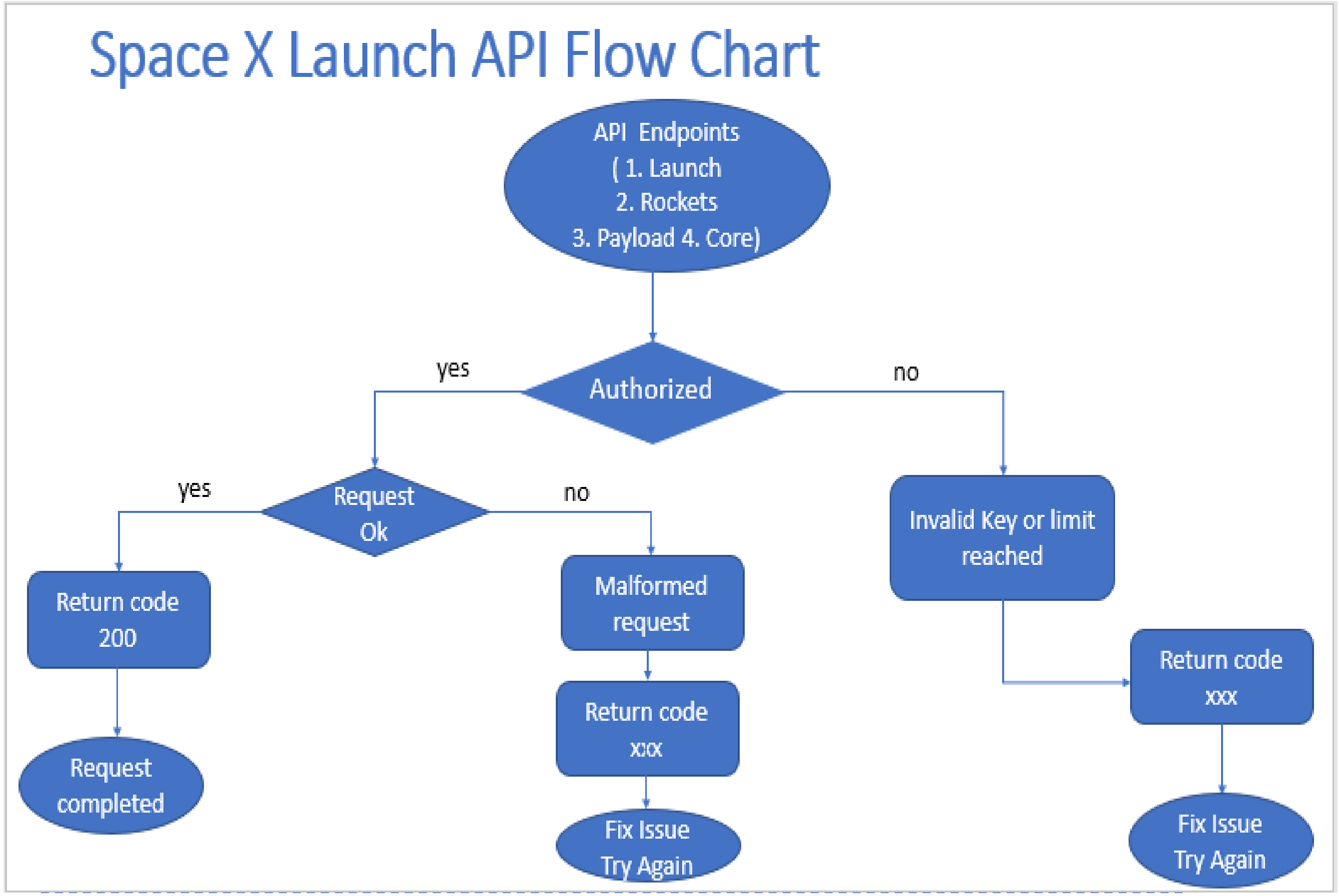
Executive Summary

- Data collection methodology:
  - We used Space X API as well as web scrapping from the Space X Wicki page to gather F9 rocket launch and landing data
- Perform data wrangling
  - Data was initially collected as a JSON then converted to a dictionary which was subsequently transformed into a Pandas dataframe for subsequent conversion to CVS
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models (Experiment usability and compatibility of SVM, Tree maps, KNN, Logistic
  - Regression optimizing parameters were built, evaluated and tuned using sklearn.

Data Collection – SpaceX API

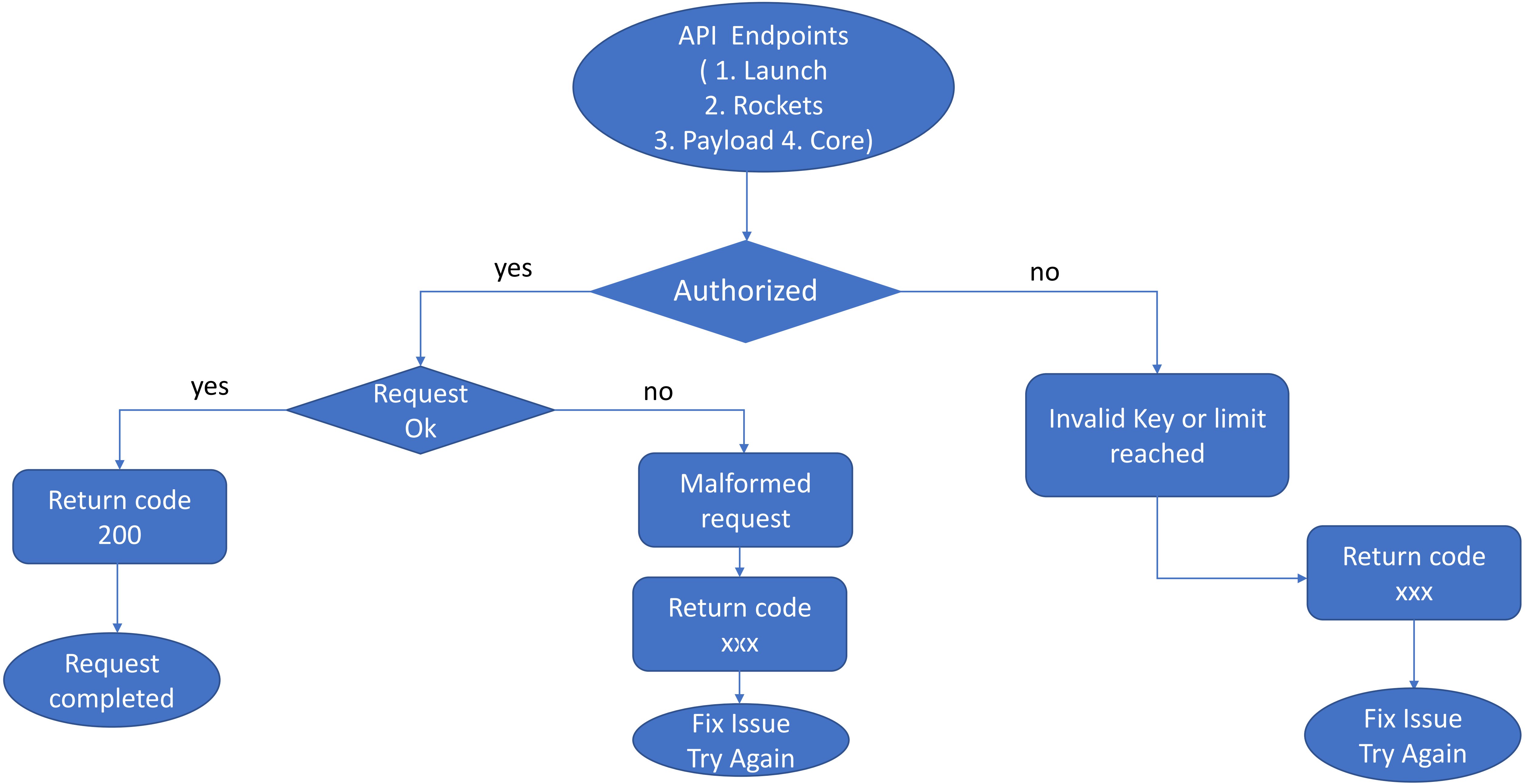
- We used Space X API in combination with several endpoints to acquire data using the following format;

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

response =requests.get(url)

response.json()
```

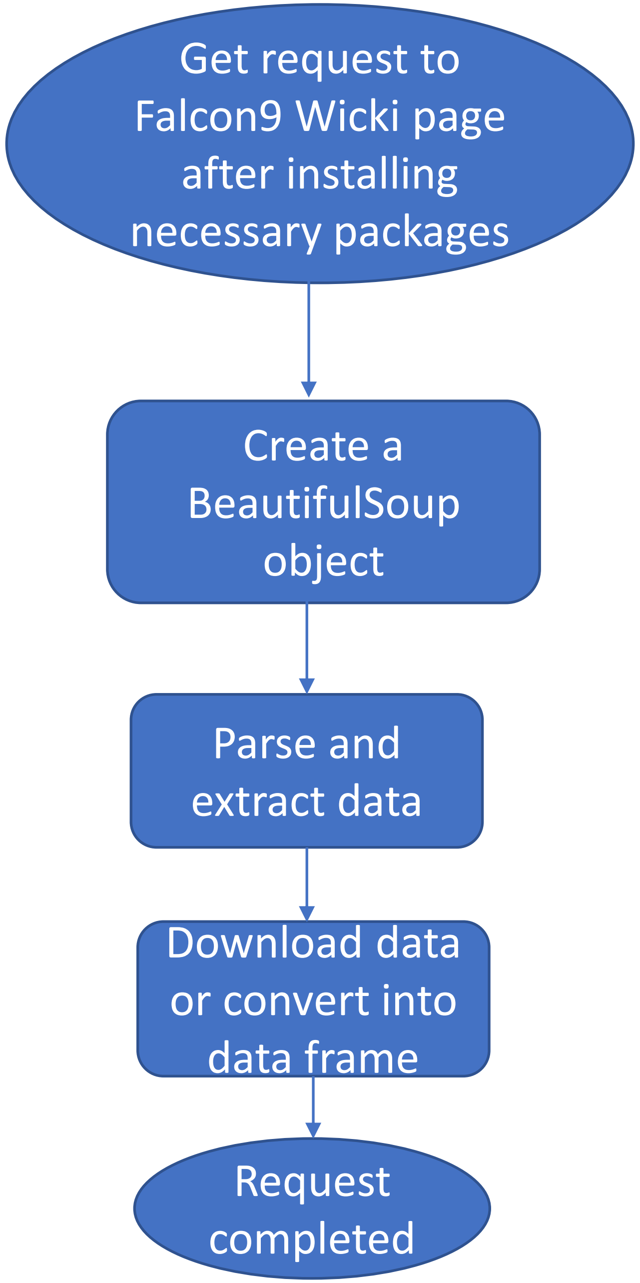


- The response is in the form of a json file which will then need to be transformed
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

Space X Launch API Flow Chart

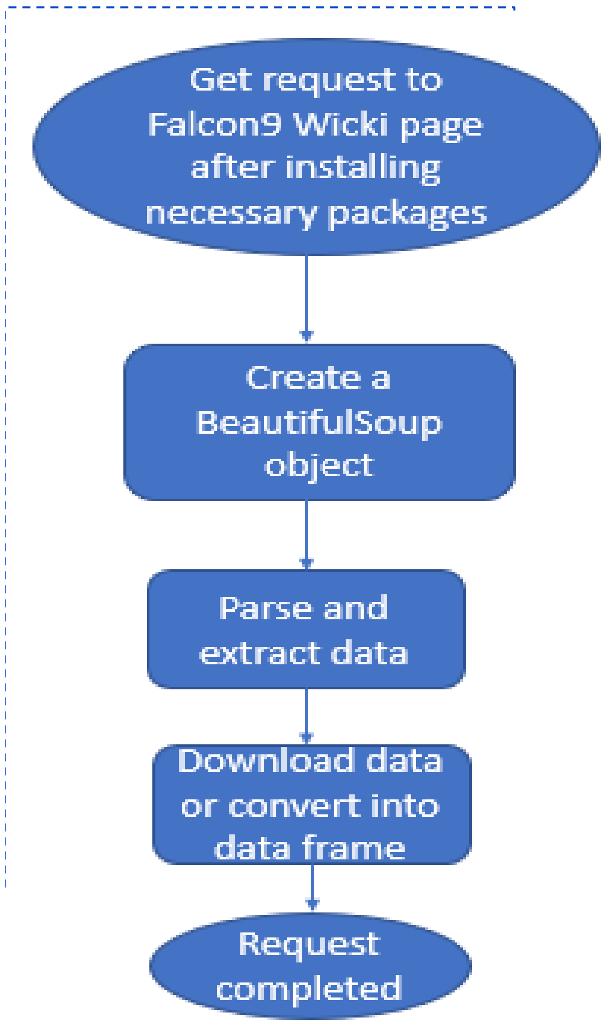


Web Scraping



Data Collection - Scraping

- Used Request Python package to fetch the Falcon9 Wicki HTML page
- Create a BeautifulSoup object
- Used regular expression to extract column headings & variable names
- Create an empty dictionary and populate with the cleaned Falcon 9 data
- Convert dictionary into a Pandas dataframe



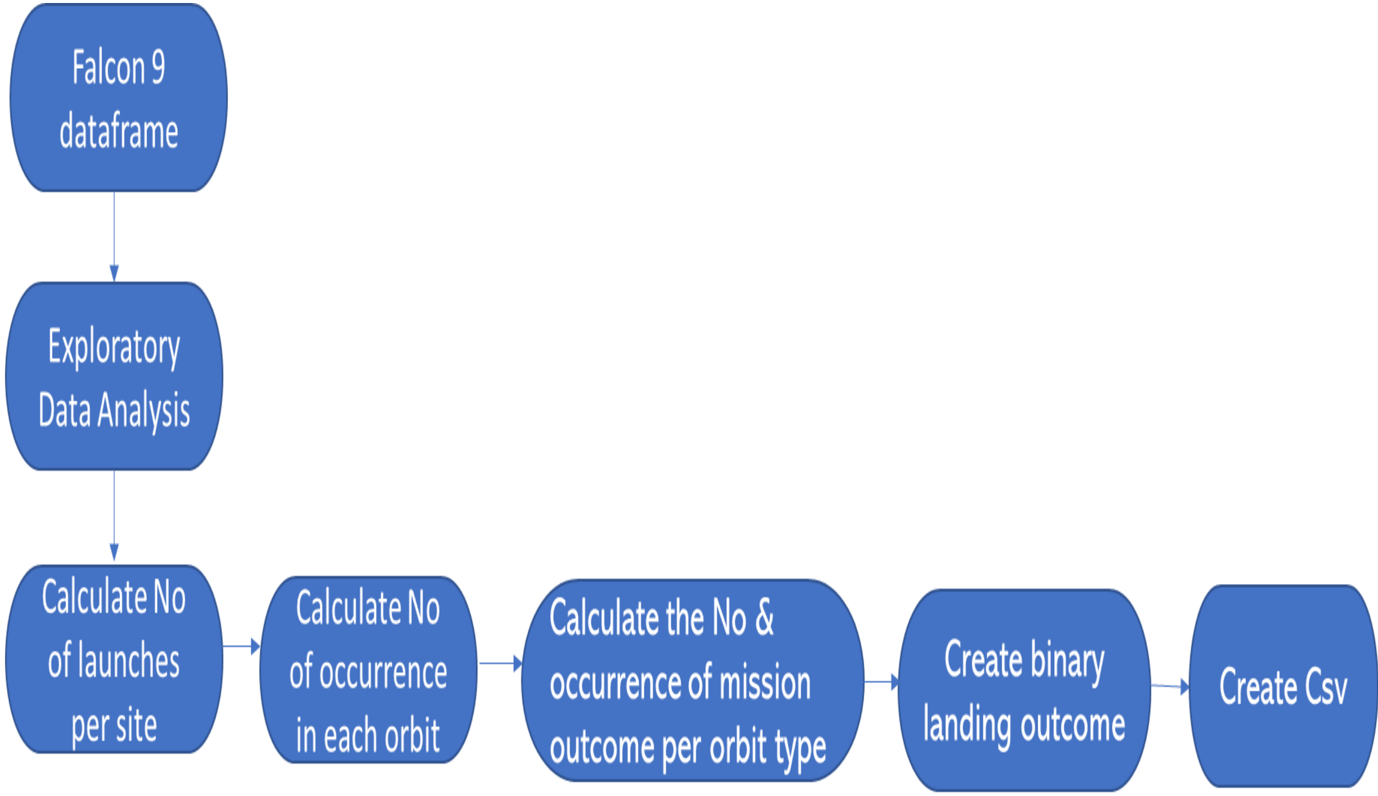
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

Data Wrangling

- Conduct exploratory data analysis
- Calculate the number of launches on 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
- Export data as a csv file

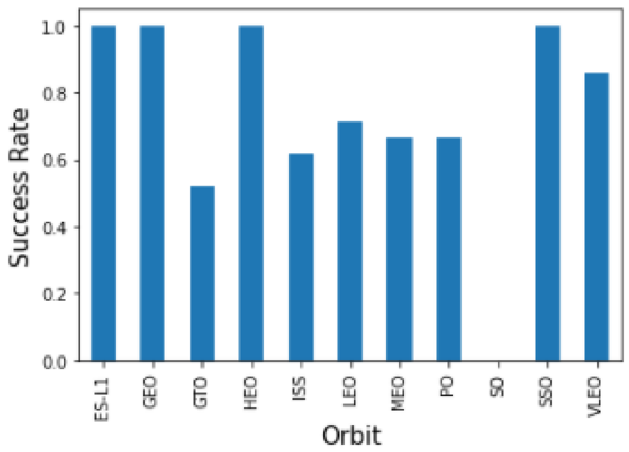


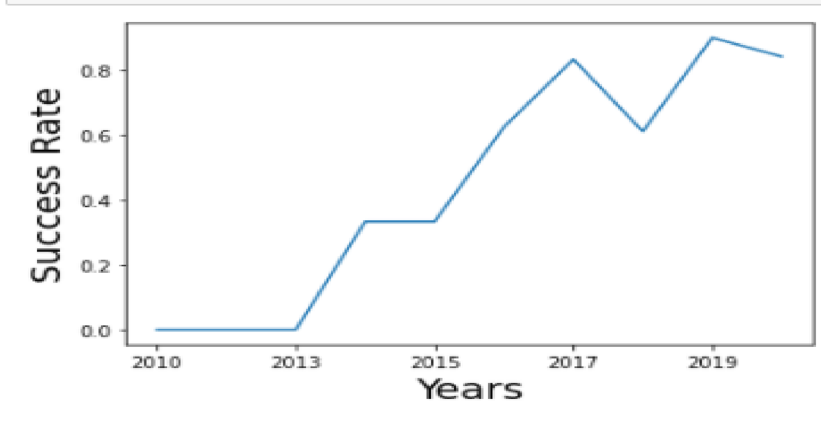
Section 2

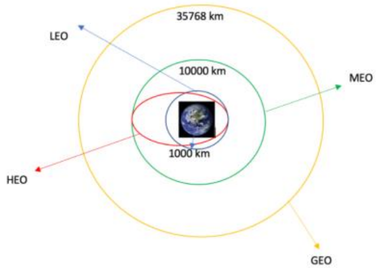
Insights drawn from EDA

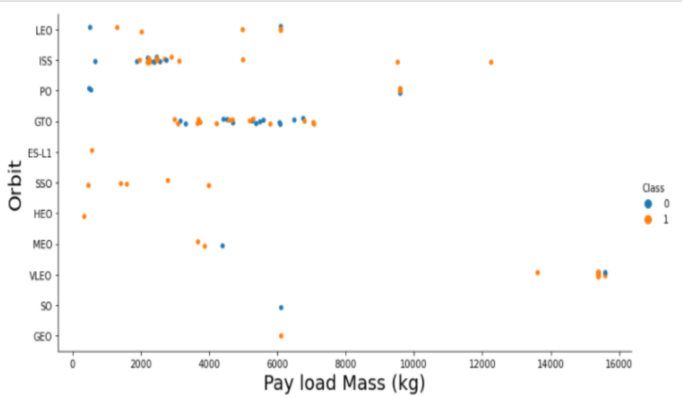
EDA with Data Visualization

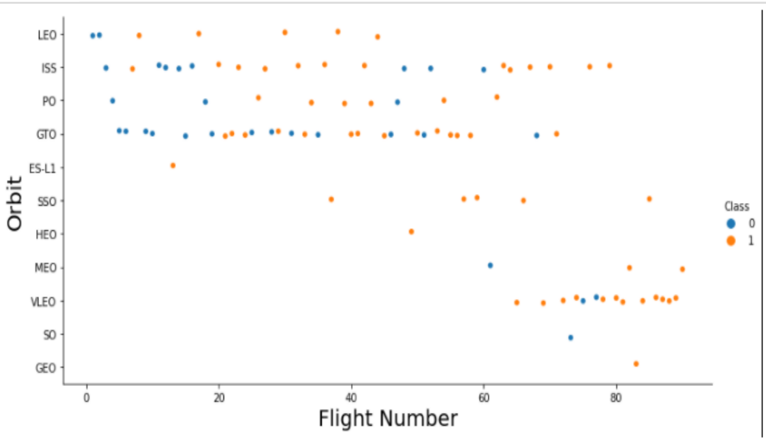
- **Landing success rate vs orbit show varying success rates and orbit distance not an obvious factor**
- **Success rate vs year show rate better with time**
- **Pay load mass vs Orbit not playing a significant impact in success rate**
- **Orbit vs flight number show increased success with number of flights**











EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

EDA with SQL

SQL queries you used

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

```
%sql select DISTINCT ("Launch_Site") AS LaunchSite_Count from SPACEXTABL
```

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

```
%sql SELECT * FROM SPACEXTABL WHERE "Launch_Site" LIKE 'CCA%' ORDER BY "PAYLOAD_MASS__KG_" DESC LIMIT 5
```

- Total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM ("PAYLOAD_MASS__KG_")FROM SPACEXTABL WHERE "Customer" LIKE '%CRS)'
```

- Average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS AVG_PAYLOAD FROM SPACEXTABL WHERE "Booster_Version" LIKE 'F9 v1.1'
```

- Dates when the first successful landing outcome in ground pad was achieved

```
%sql SELECT MIN("Date")FROM SPACEXTABL WHERE "Landing__Outcome" LIKE '%(ground pad)'
```

- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE ("Landing__Outcome" LIKE 'Success (d%') AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)
```

- List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT("Mission_Outcome") AS count FROM SPACEXTABL WHERE("Mission_Outcome" LIKE 'S%') OR ("Mission_Outcome"LIKE 'F%') GROUP BY "Mission_Outcome"
```

- Use a subquery to list the names of the booster_versions which have carried the maximum payload mass.

```
%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE "PAYLOAD_MASS__KG_" IN ( SELECT ("PAYLOAD_MASS__KG_") FROM SPACEXTABL WHERE ("PAYLOAD_MASS__KG_" = 15600))
```

- Months, failure landing outcome in drone ship, booster versions, launch_site for the months in year 2015

```
%sql SELECT * FROM SPACEXTABL WHERE("Date" LIKE '2015%')
```

- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

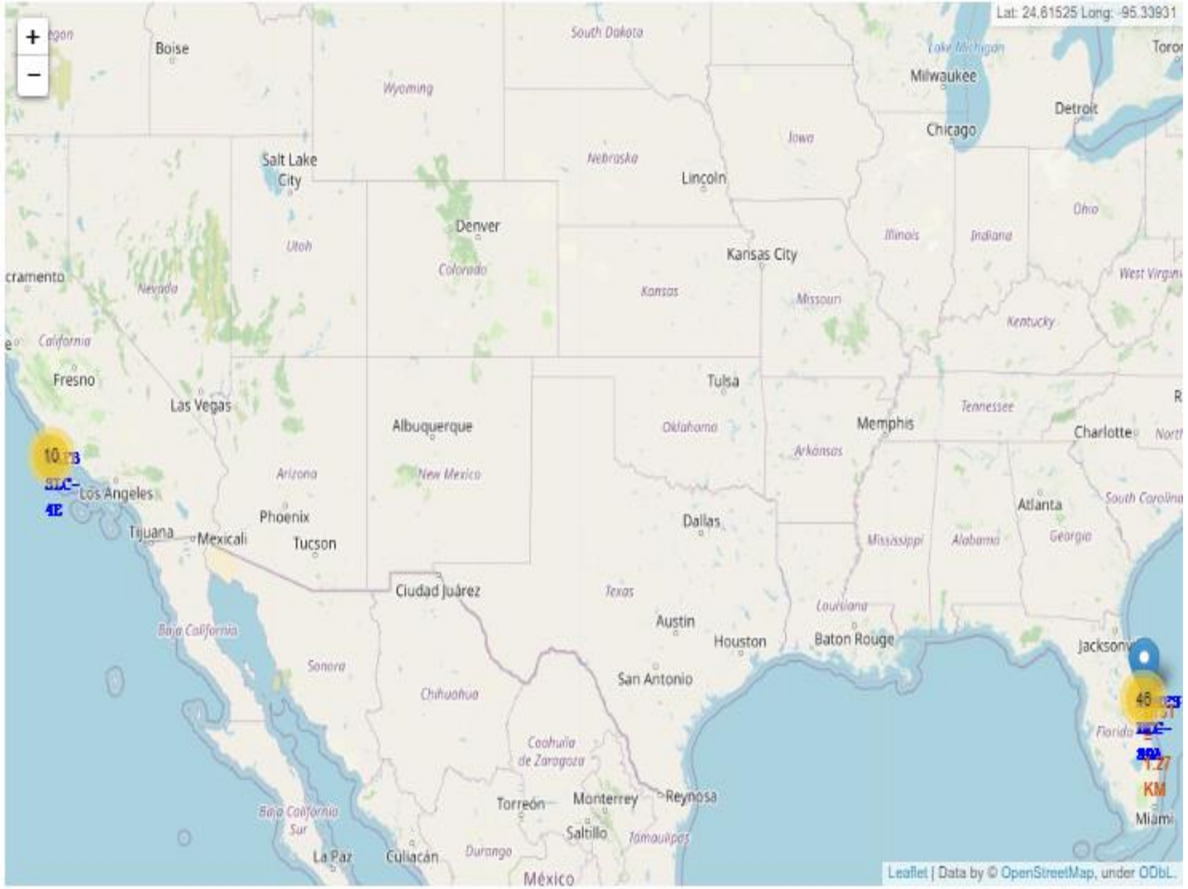
```
%sql SELECT ( "Landing__Outcome"), COUNT("Landing__Outcome") AS Count  FROM SPACEXTABL WHERE ("Date" BETWEEN '2010-06-04' AND '2017-03-20') AND
("Landing__Outcome" LIKE 'Succ%')  GROUP BY ("Landing__Outcome") ORDER BY Count  DESC
```

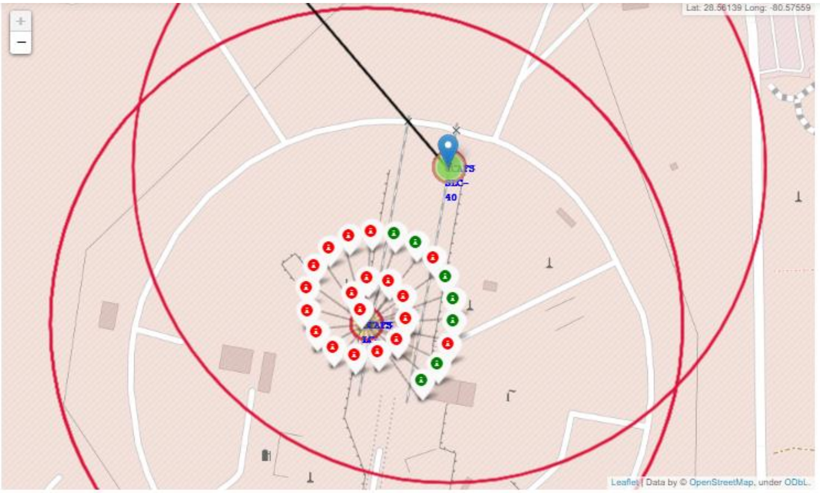
Build an Interactive Map with Folium

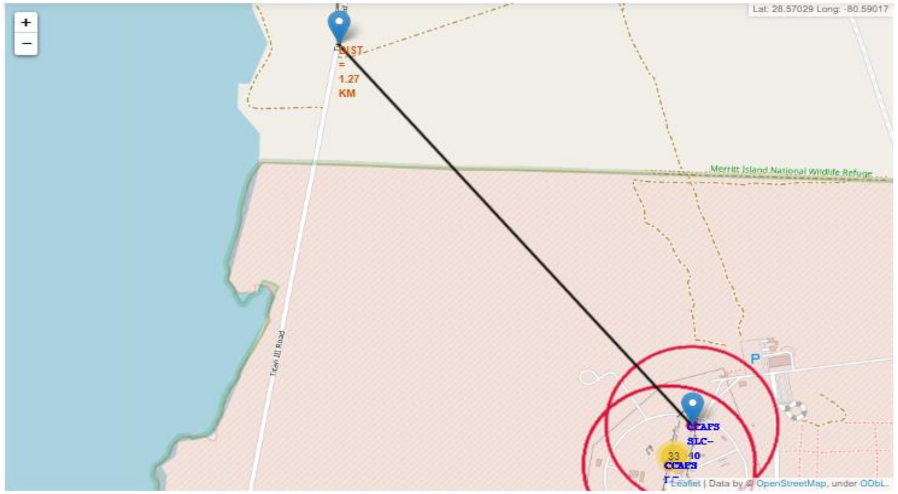
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Interactive Map with Folium

- Markers & circles were used to facilitate identification of the launch sites
- Green and red pop-up marker clusters were used to identify success or failure from the sites over time
- Distance line from one launch site to nearest railroad to indicate proximity to transportation







Build a Dashboard with Plotly Dash

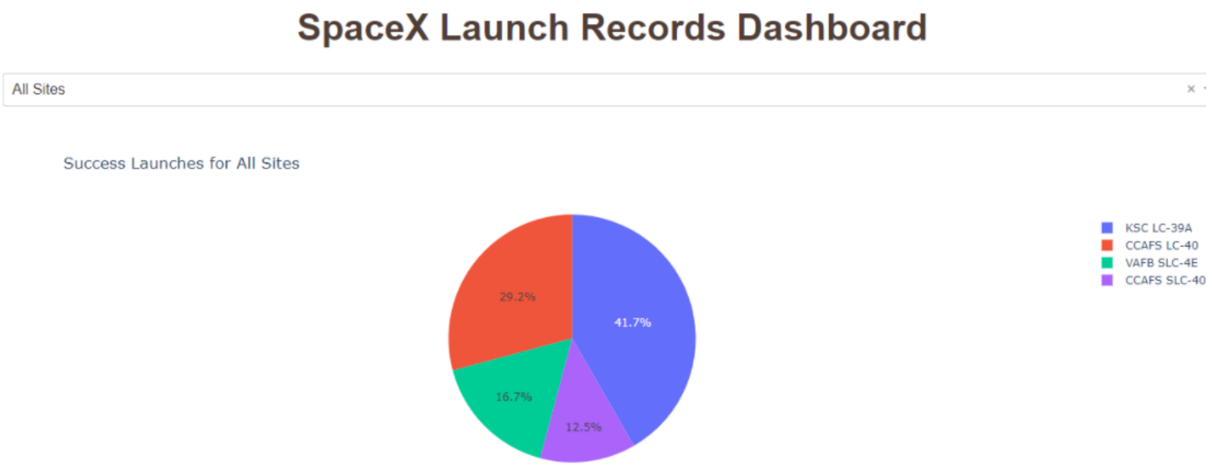
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

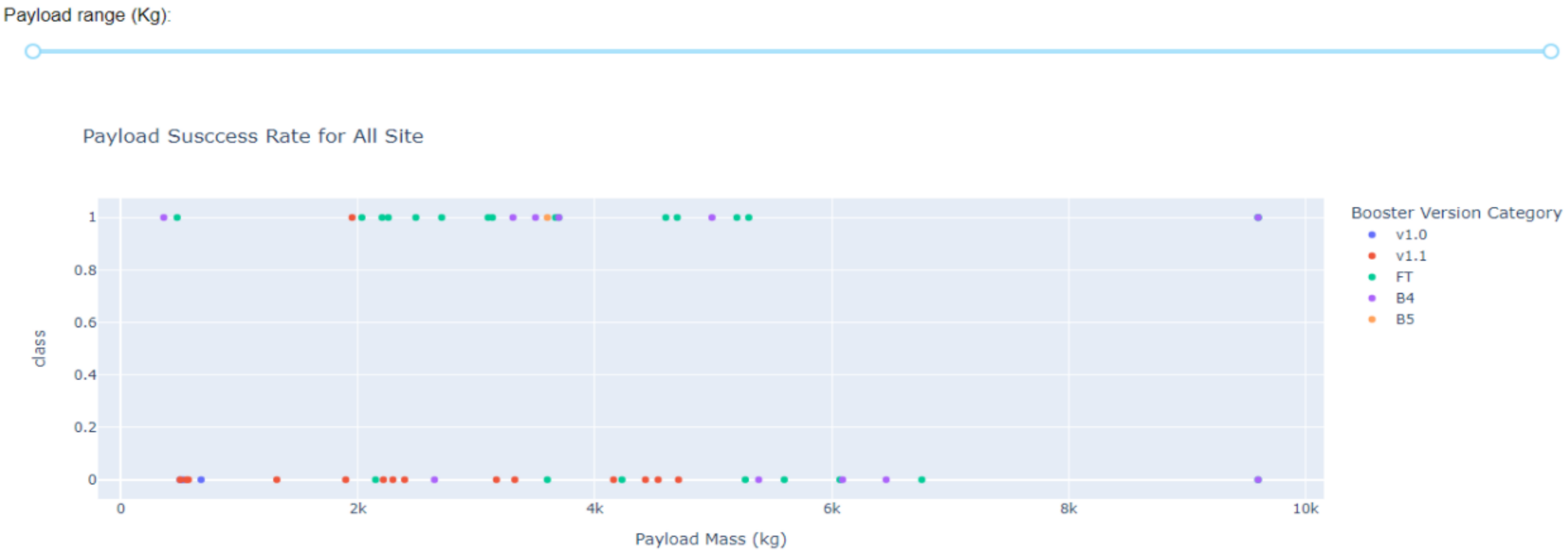
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Dashboard with Plotly Dash

- Using an interactive dashboard consisting of a pie chart, we can see KSC LC 39A has the highest success rate.
- With the interactive dashboard the user can isolate one site to drill down further on the success rate
- With a trend chart of the success rate by payload vs booster version, we see clearly version FT has highest success rate and F9 has a poor rate

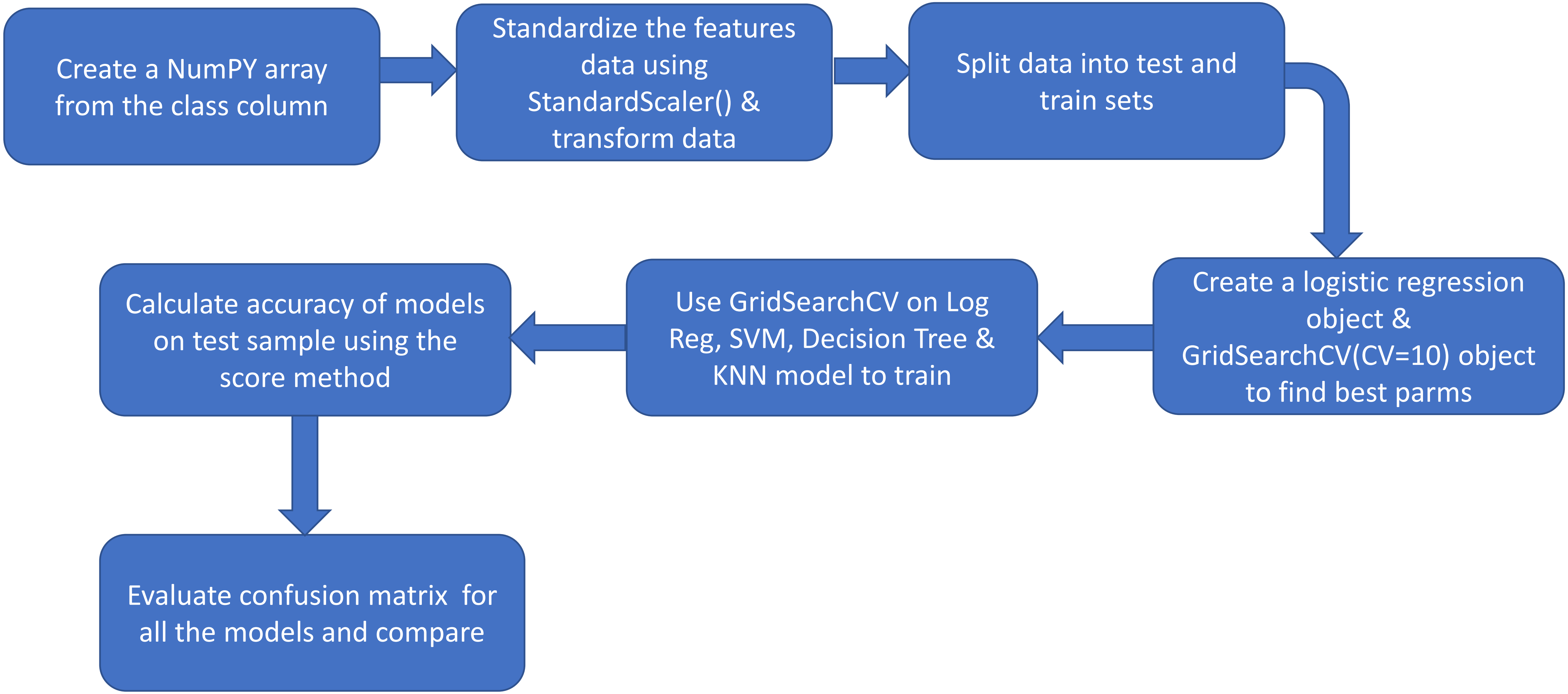




Predictive Analysis (Classification)

- The space X is split into feature and target sets. Standardize the data using StandardScaler() and transform using fit_transform(X)
- Split the data into train and test sets where the test parameter is 20% and random=2
- Create a logistic regression object then create a GridSearchCV object with cv = 10. Train the model with train data to find the best parameters from the dictionary parameters
  - All model types must go thru the creation of a model object and GridSearchCV steps to find the best parameters for each model type
- Calculate accuracy using score method
- Plot confusion matrix

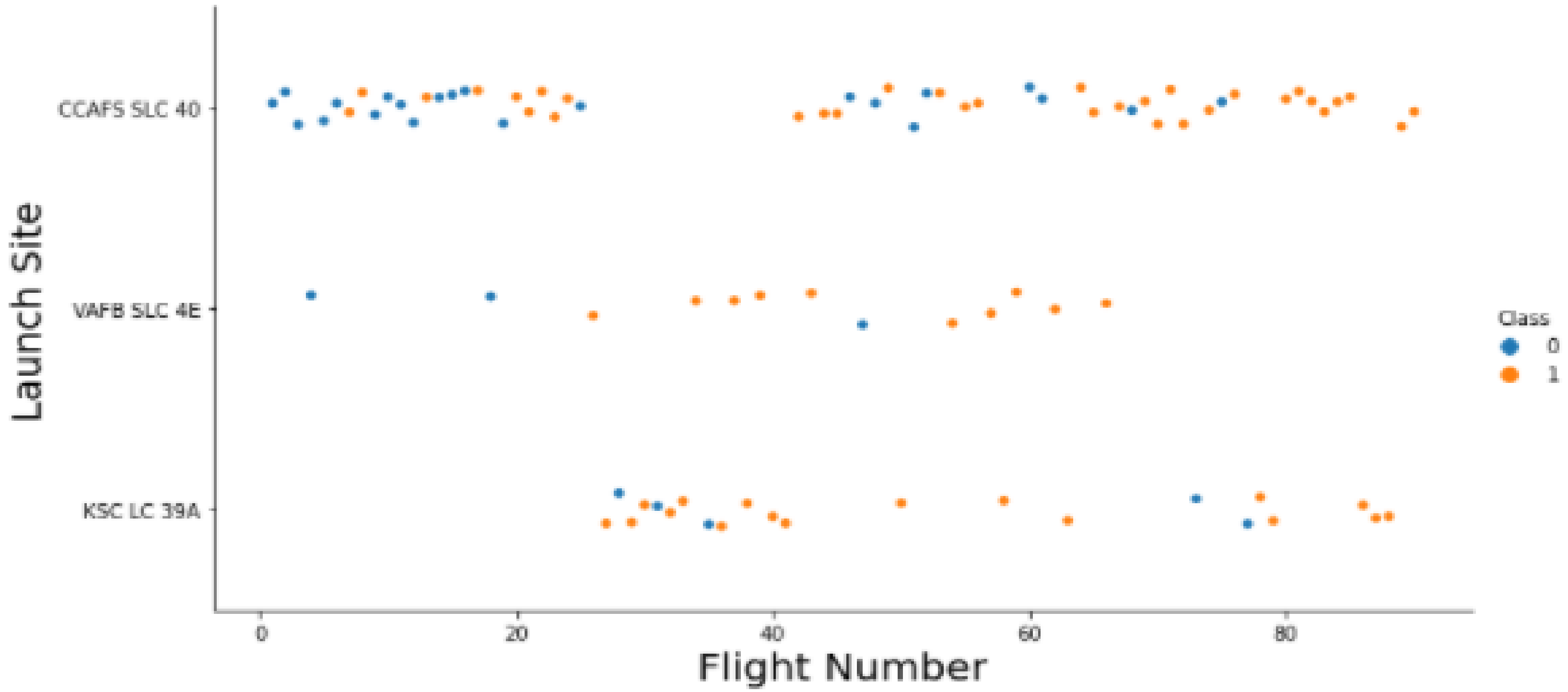
Predictive Analysis (Classification)



Section 2

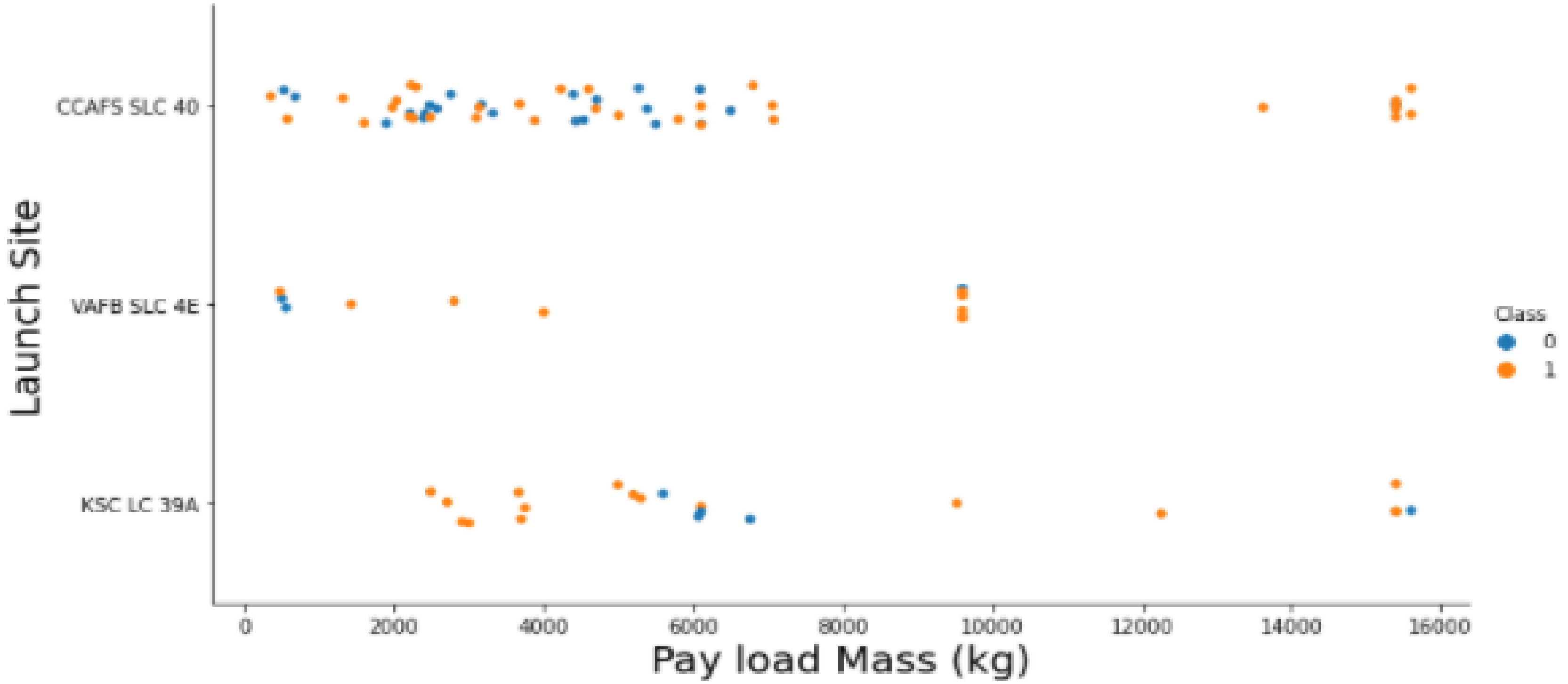
Insights drawn from EDA

Flight Number vs. Launch Site



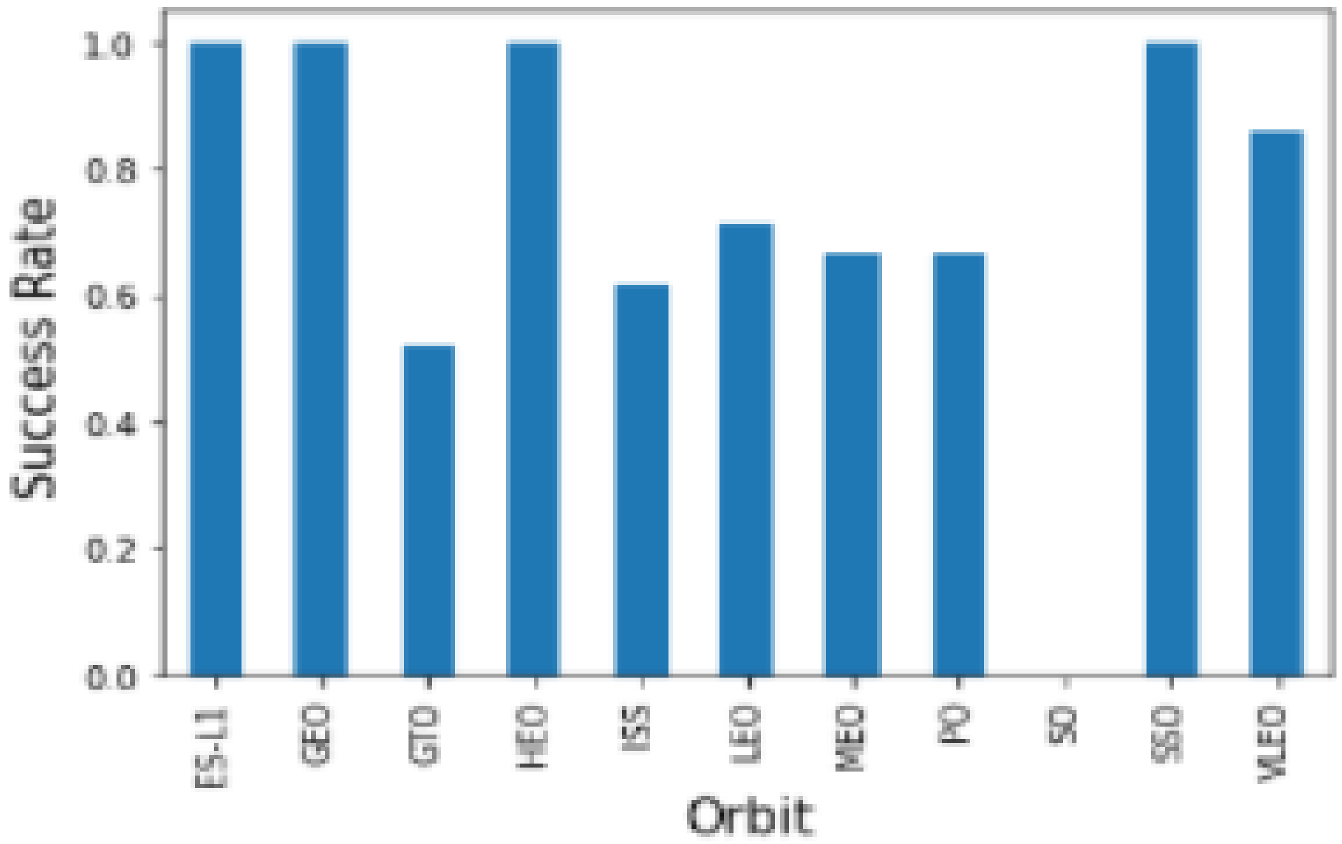
Lower flight numbers had more failures and mostly from one site

Payload vs. Launch Site



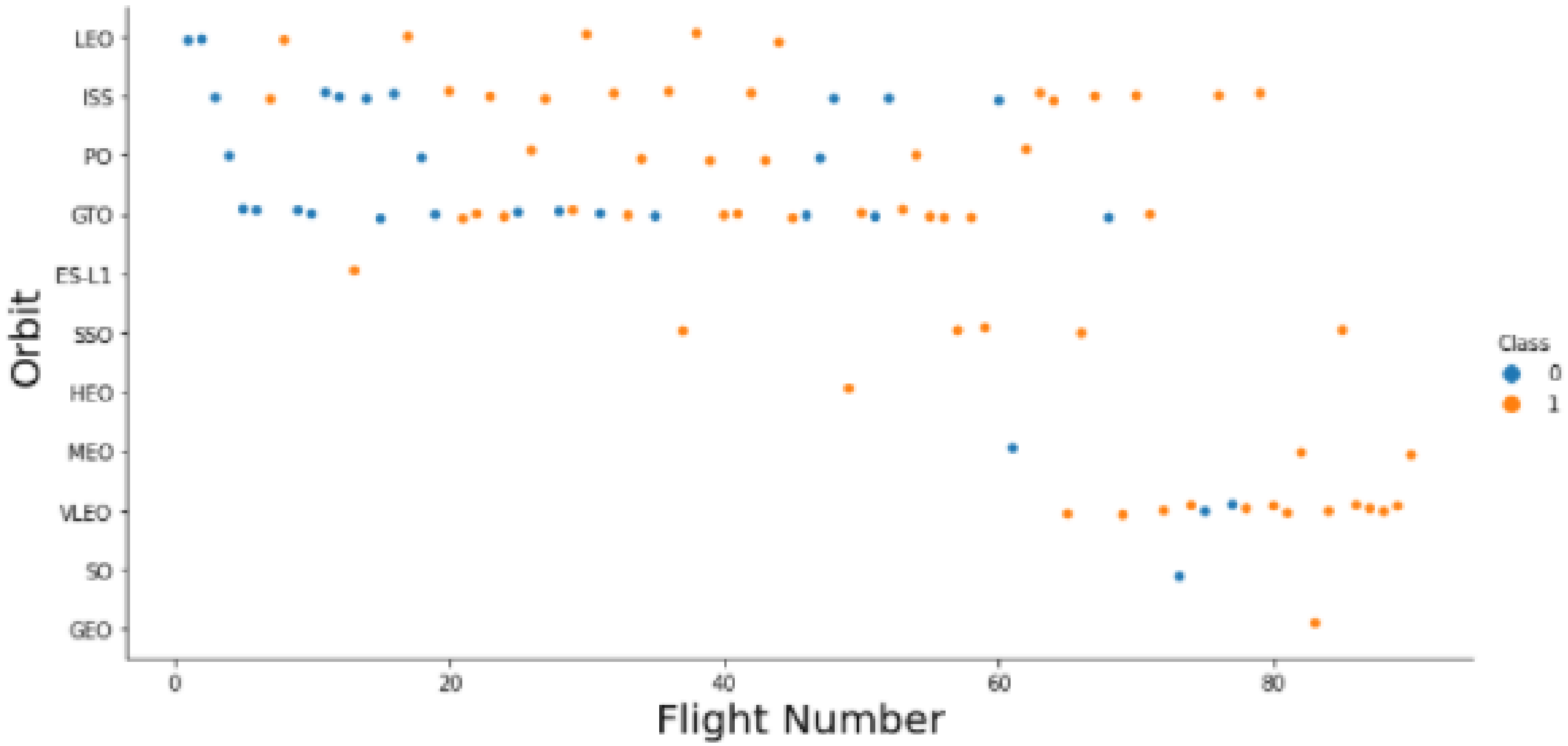
Two site appear to handle the majority of the heavy payload launches, while the mid to low payloads were more spread out
On the whole, CCAFS SLC 40 & KSC LC 39A are the most used

Success Rate vs. Orbit Type



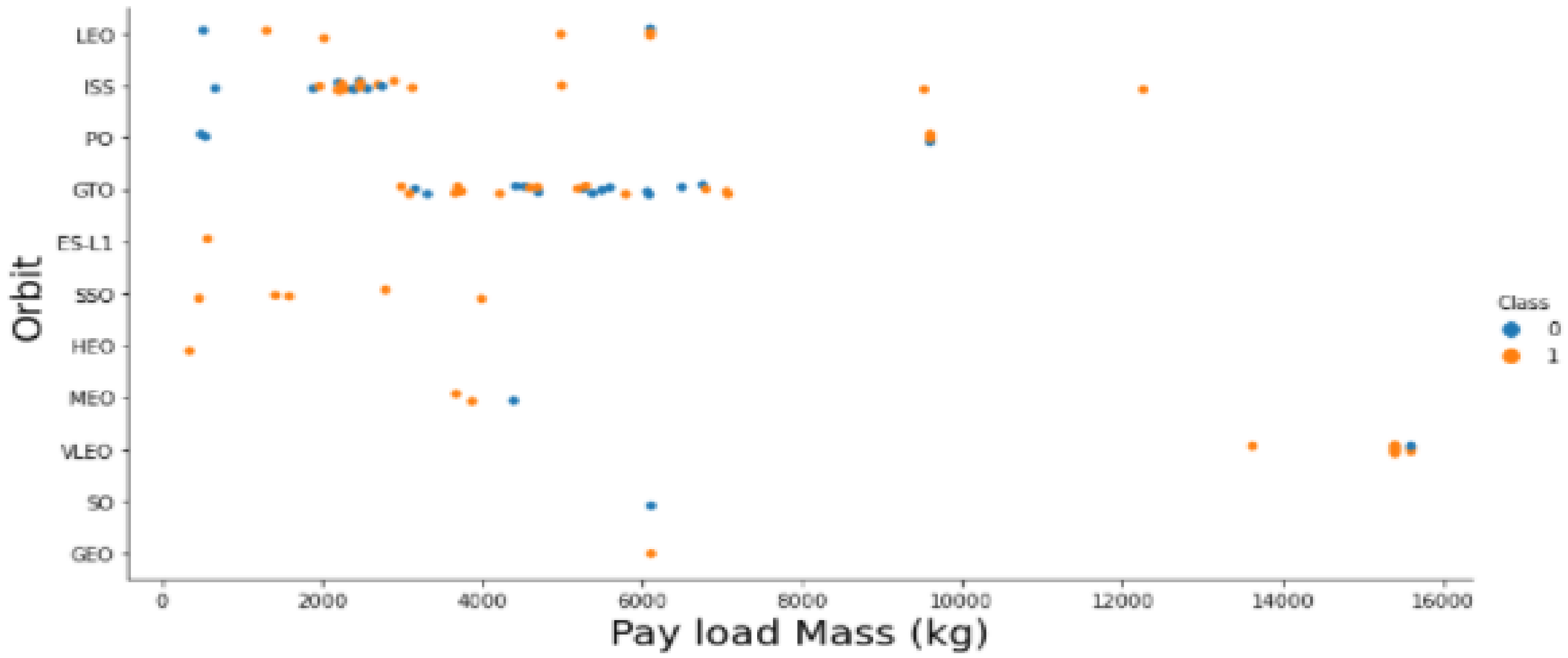
Multiple orbits have a very high success rate
LEO orbit which is important for certain satellites is roughly 62%
No data for Orbit SO

Flight Number vs. Orbit Type



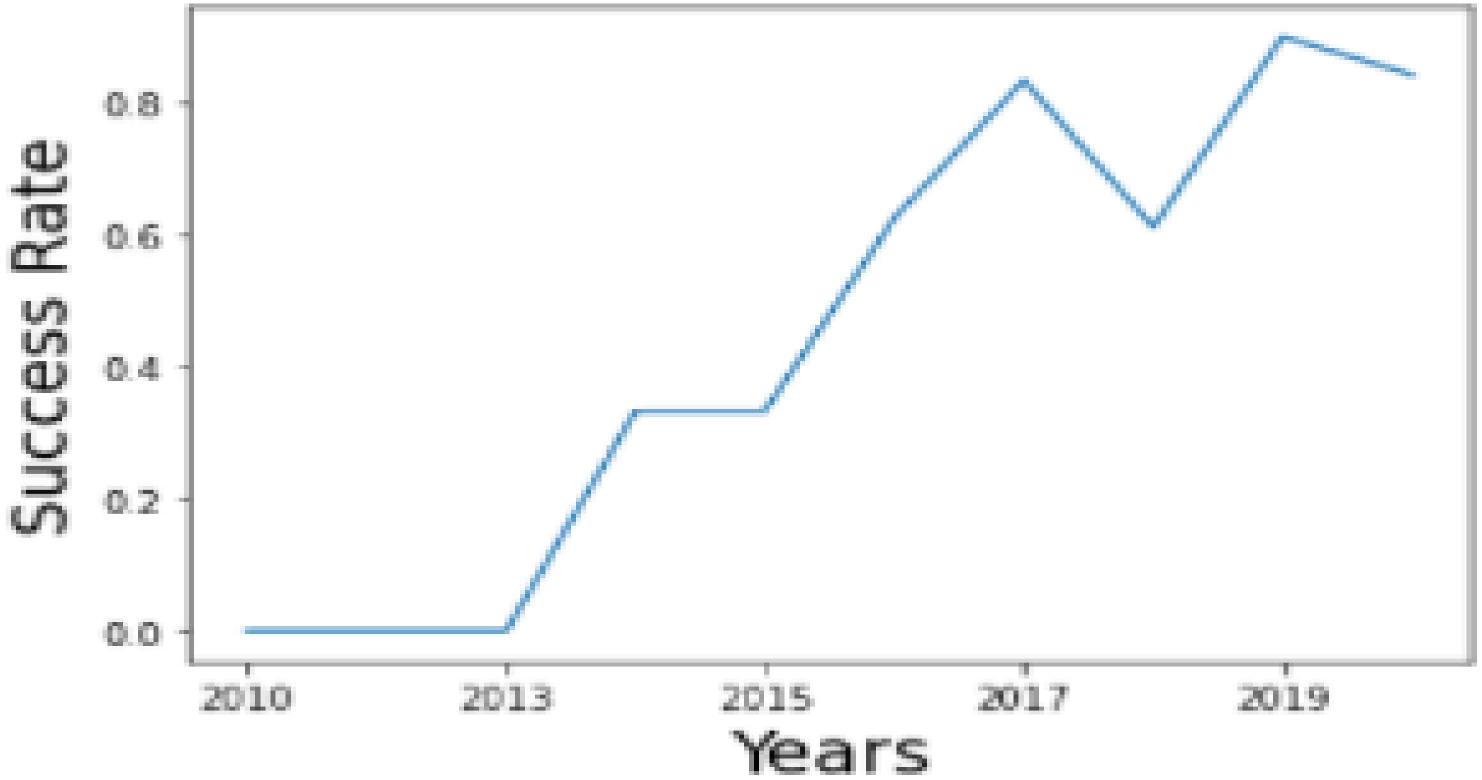
Certain orbit types such as GEO,SO and MEO are sparsely populated
Most latter launches went to VLEO orbit
Early launches or lower flight number went to GTO, PO, ISS & LEO and had a lower success rate compared to the latter launches

Payload vs. Orbit Type



Higher payloads mostly went to VLEO orbit
Most low to medium payloads went to ISS & GTO orbits

Launch Success Yearly Trend



Low success rate early on which significantly improved after 2015

All Launch Site Names

```
%sql select DISTINCT ("Launch_Site") AS LaunchSite_Count from SPACEXTABL
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.blu
Done.
```

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

Results of query showing all launch site in the data set

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABL WHERE "Launch_Site" LIKE 'CCA%' ORDER BY "PAYLOAD_MASS__KG_" DESC LIMIT 5

#%sql SELECT "Launch_Site" FROM SPACEXTABL WHERE "Launch_Site" LIKE 'CCA%' ORDER BY "PAYLOAD_MASS__KG_" ASC
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

Date	Time__UTC_	Booster_Version	Launch_Site	Payload	payload_mass__kg_	Orbit	Customer	Mission_Outcome	Landing__Outcome
2019-11-11	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success
2020-06-04	01:25:00	F9 B5 B1049.5	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success
2020-02-17	15:05:00	F9 B5 B1056.4	CCAFS SLC-40	Starlink 4 v1.0, SpaceX CRS-20	15600	LEO	SpaceX	Success	Failure
2020-01-29	14:07:00	F9 B5 B1051.3	CCAFS SLC-40	Starlink 3 v1.0, Starlink 4 v1.0	15600	LEO	SpaceX	Success	Success
2020-01-07	02:33:00	F9 B5 B1049.4	CCAFS SLC-40	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600	LEO	SpaceX	Success	Success

Results of query of launch site beginning with CCA

Total Payload Mass

```
%sql SELECT SUM ("PAYLOAD_MASS__KG_")FROM SPACEXTABL WHERE "Customer" LIKE '%CRS)'
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50
Done.
```

1
45596

Total payload carried for NASA(CRS) is 45,496 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS AVG_PAYLOAD FROM SPACEXTABL WHERE "Booster_Version" LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

avg_payload
2928.400000

The average payload carried by booster version F9 V1.1 is 2,928.4 kg

First Successful Ground Landing Date

```
%sql SELECT MIN("Date")FROM SPACEXTABL WHERE "Landing__Outcome" LIKE '%(ground pad)'
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
2015-12-22

- First successful landing outcome on ground pad was on Dec 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE ("Landing__Outcome" LIKE 'Success (d%') AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)

```
%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE ("Landing__Outcome" LIKE 'Success (d%') AND ("PAYLOAD_MASS__KG_" BETWEEN 400
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

List the 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

```
%sql SELECT COUNT("Mission_Outcome") AS count FROM SPACEXTABL WHERE("Mission_Outcome" LIKE 'S%') OR ("Mission_Outcome"LIKE 'F%')
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

COUNT
1
99
1

The total number of successful and failure mission outcomes are 99 and 1 respectively

Boosters Carried Maximum Payload

%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE "PAYLOAD_MASS__KG_" IN (SELECT ("PAYLOAD_MASS__KG_") FROM SPACEXTABL WHERE ("PAYLOAD_MASS__KG_" = 15600))

```
%sql SELECT "Booster_Version" FROM SPACEXTABL WHERE "PAYLOAD_MASS__KG_" IN ( SELECT ("PAYLOAD_MASS__KG_") FROM SPACEXTABL WHERE (
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
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

The number and type of booster versions that carried the max payload

2015 Launch Records

```
%sql SELECT * FROM SPACEXTABL WHERE("Date" LIKE '2015%')
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

Date	Time__UTC_	Booster_Version	Launch_Site	Payload	payload_mass__kg_	Orbit	Customer	Mission_Outcome	Landing__Outcome
2015-01-10	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-02-11	23:03:00	F9 v1.1 B1013	CCAFS LC-40	DSCOVR	570	HEO	U.S. Air Force NASA NOAA	Success	Controlled (ocean)
2015-03-02	03:50:00	F9 v1.1 B1014	CCAFS LC-40	ABS-3A Eutelsat 115 West B	4159	GTO	ABS Eutelsat	Success	No attempt
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-27	23:03:00	F9 v1.1 B1016	CCAFS LC-40	Turkmen 52 / MonacoSAT	4707	GTO	Turkmenistan National Space Agency	Success	No attempt
2015-06-28	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

List of landings in 2015

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

%sql SELECT ("Landing__Outcome") , COUNT("Landing__Outcome") AS Count FROM SPACEXTABL WHERE ("Date" BETWEEN '2010-06-04' AND '2017-03-20') AND ("Landing__Outcome" LIKE 'Succ%') GROUP BY ("Landing__Outcome") ORDER BY Count DESC

```
%sql SELECT ( "Landing__Outcome")  , COUNT("Landing__Outcome") AS Count  FROM SPACEXTABL WHERE ("Date" BETWEEN '2010-06-04' AND
```

```
 * ibm_db_sa://hbq74889:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

Landing__Outcome	COUNT
Success (drone ship)	5
Success (ground pad)	3

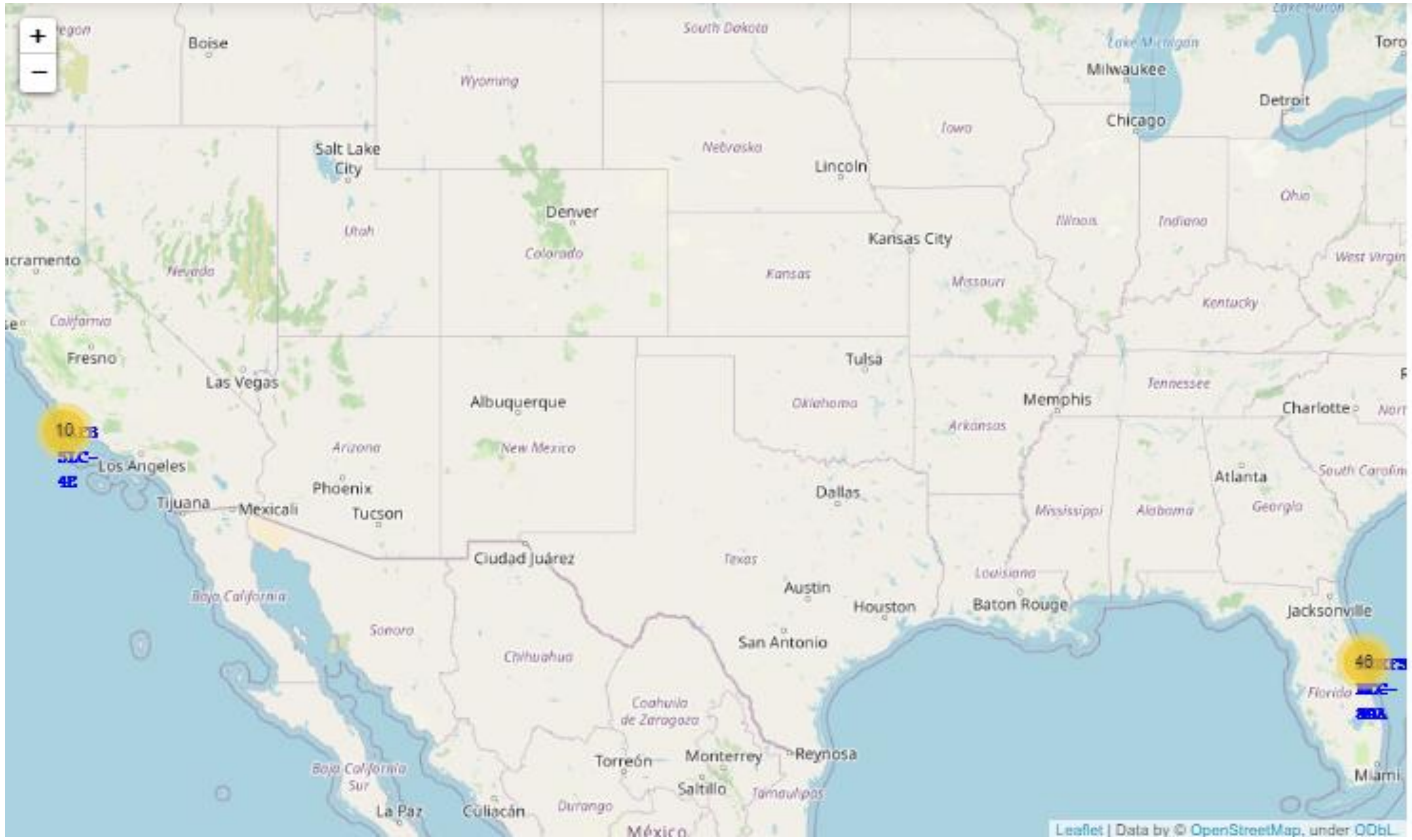
Successful landing outcomes in drone ship between the date 2010-06-04 and 2017-03-20, in descending order

Section 4

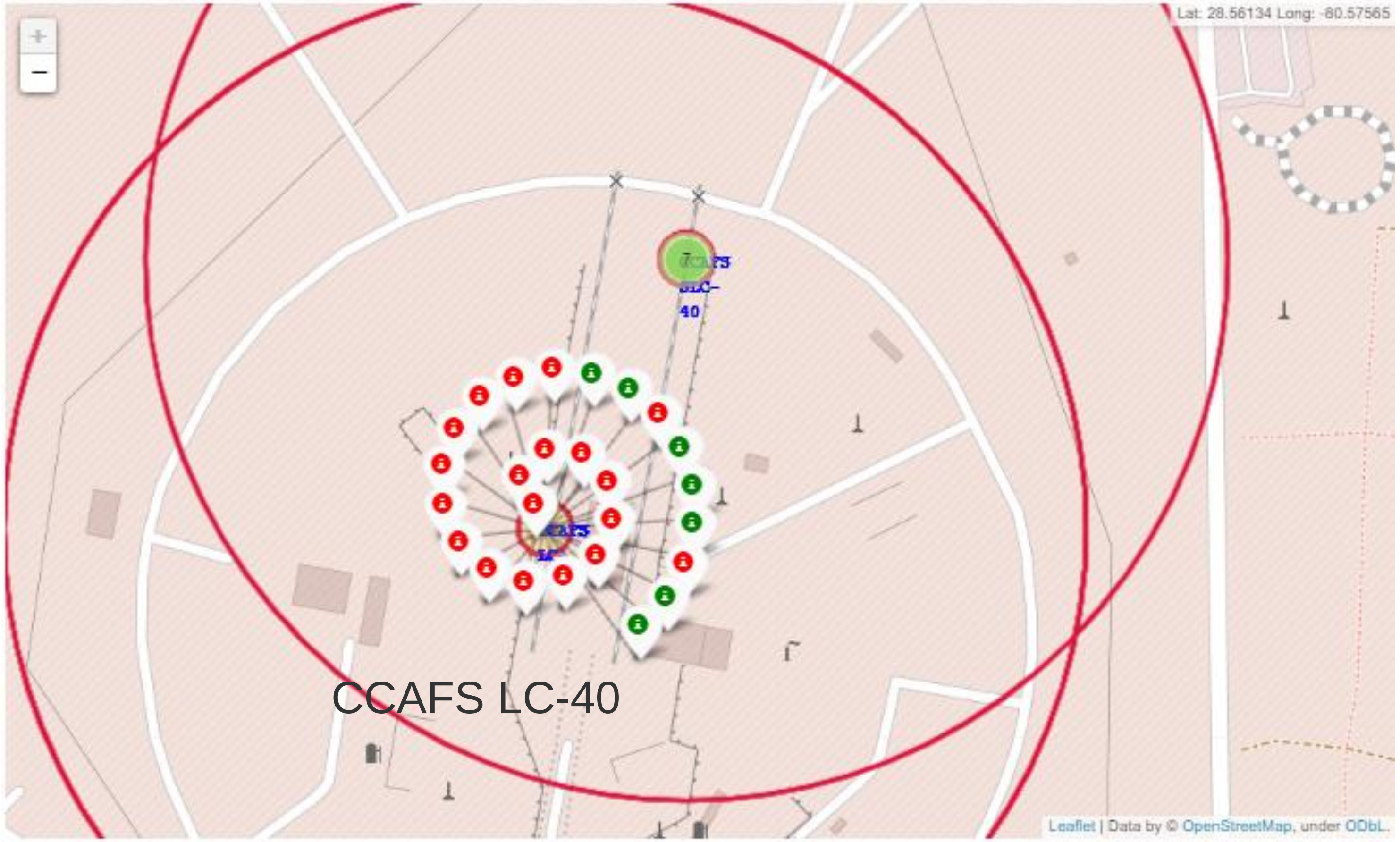
Launch Sites Proximities Analysis

Space X Launch Site In The USA

- Space X launch site as you can see on the map are located on either coast of the US
- The sites are specifically positioned in the southern US and at this high level we can see there are near the coast for safety
- We can see labels designating the launch sites



Regular & Clustered Labels Of Launch Result Records

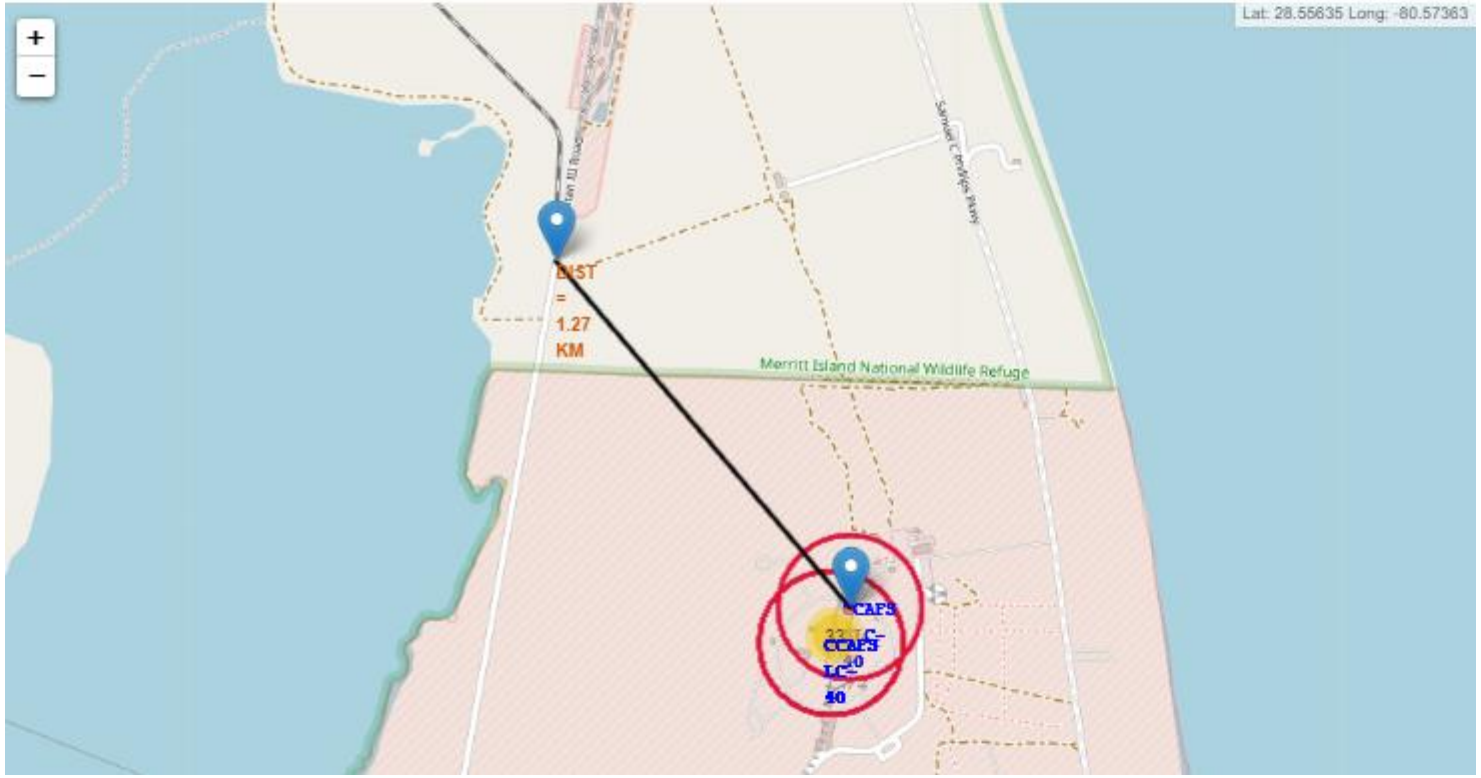


High magnification view of 2 launch sites in Florida

One site showing the clustered labels indicating success or failure over time at that site. Red is good

The other site which has far fewer launches is not shown in detail but can be by a simple click

Proximity To Coast & Railroad for Florida Site

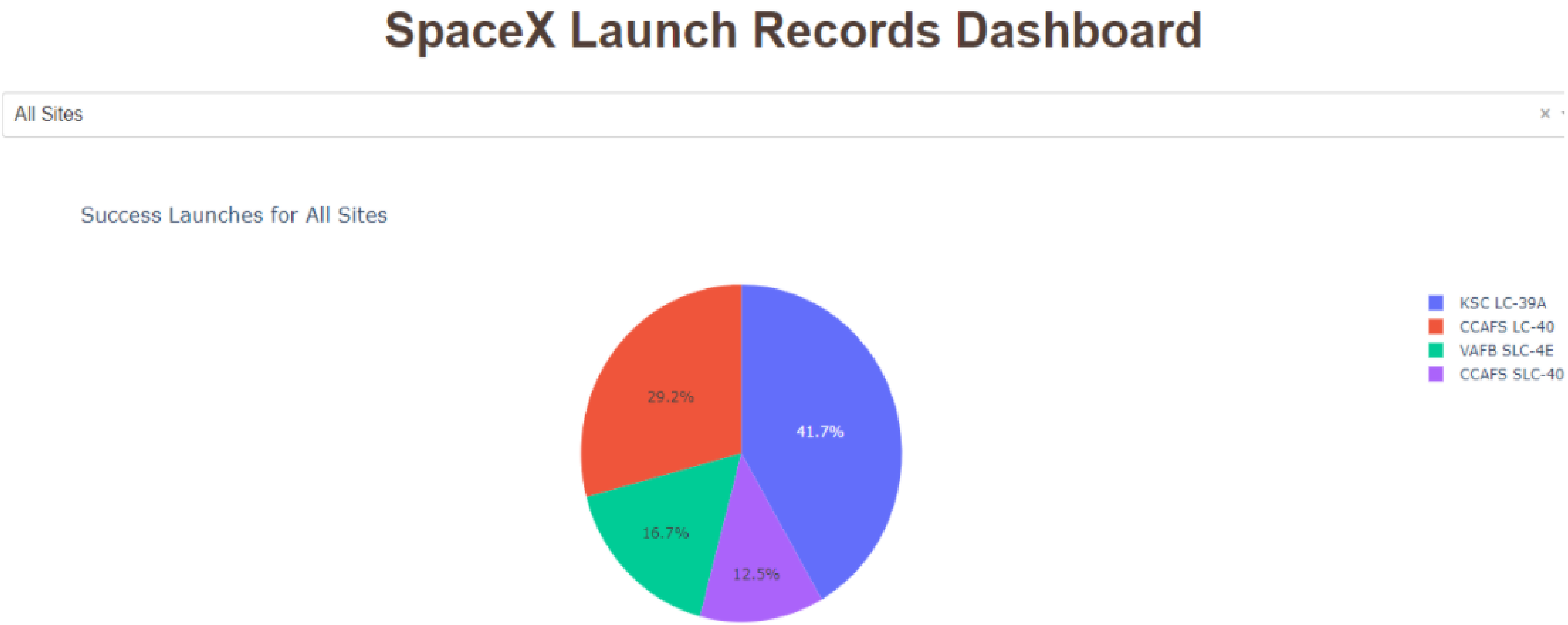


As we can clearly see two marker identify one launch site and a nearby railroad track

Section 5

Build a Dashboard with Plotly Dash

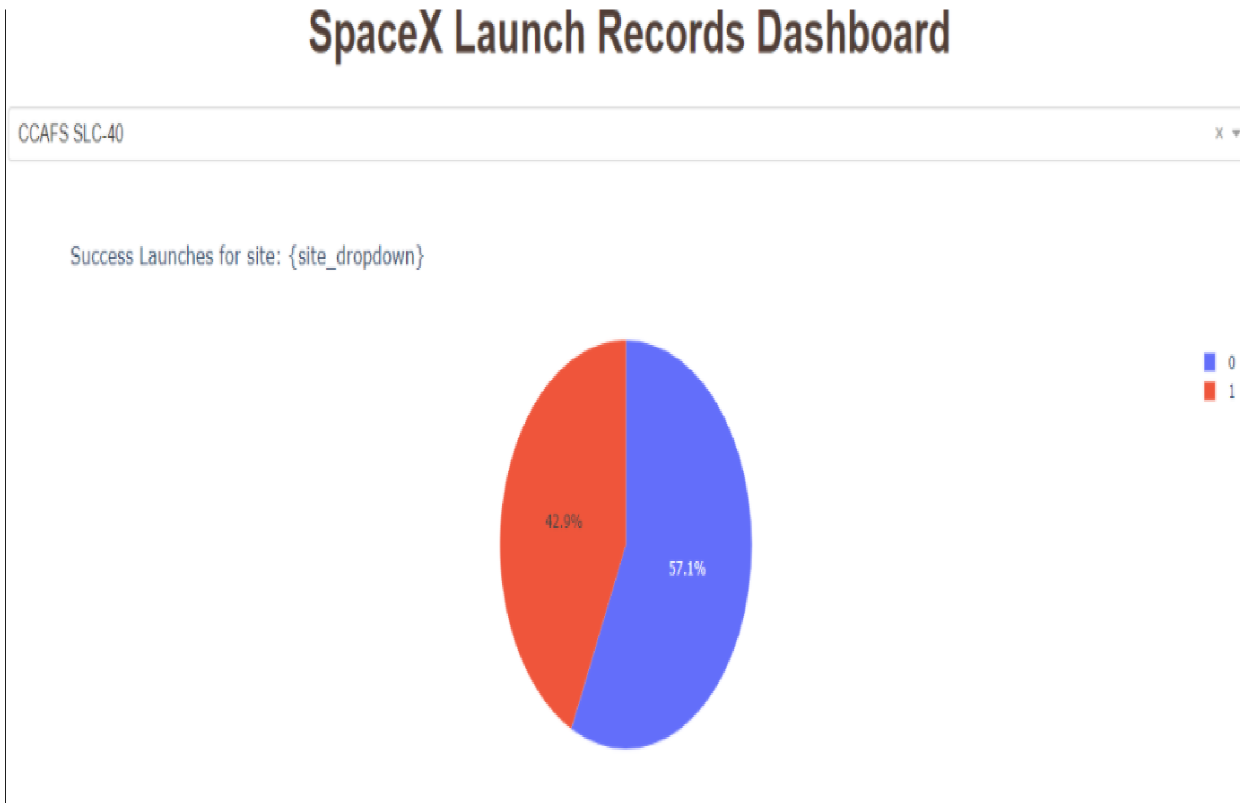
SpaceX Launch Success Rate by Site



KSC LC-39A has the highest success rate

CCAFS SLC-40 has the poorest success rate

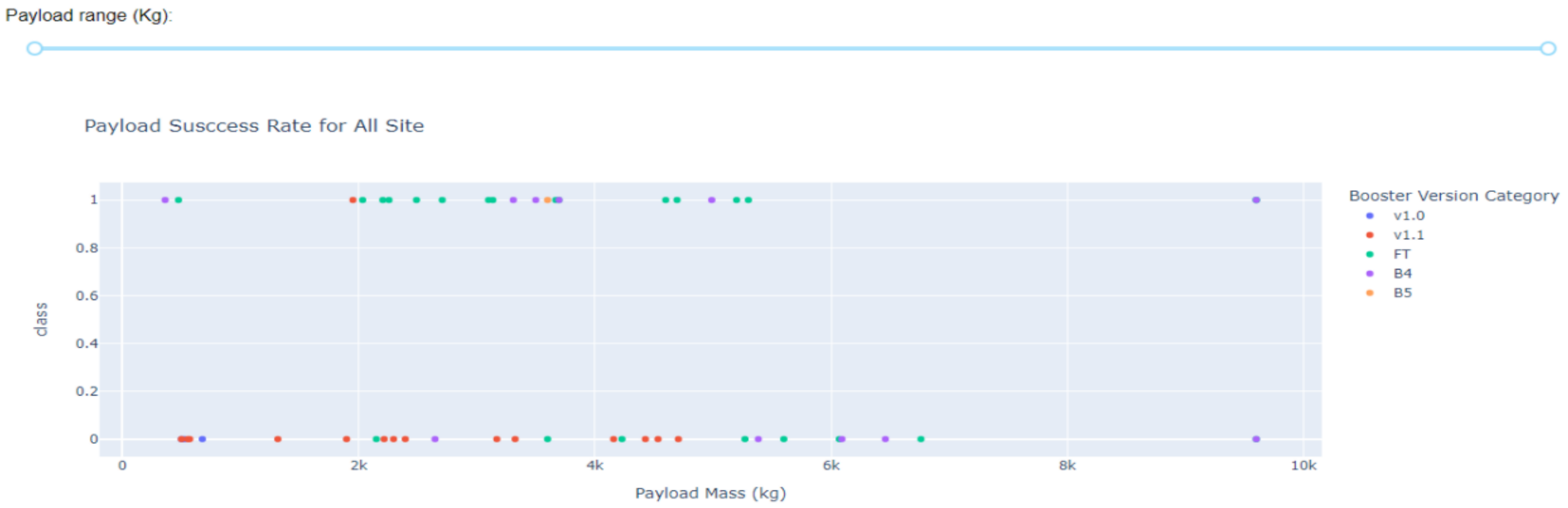
SpaceX Launch Success Rate by Site

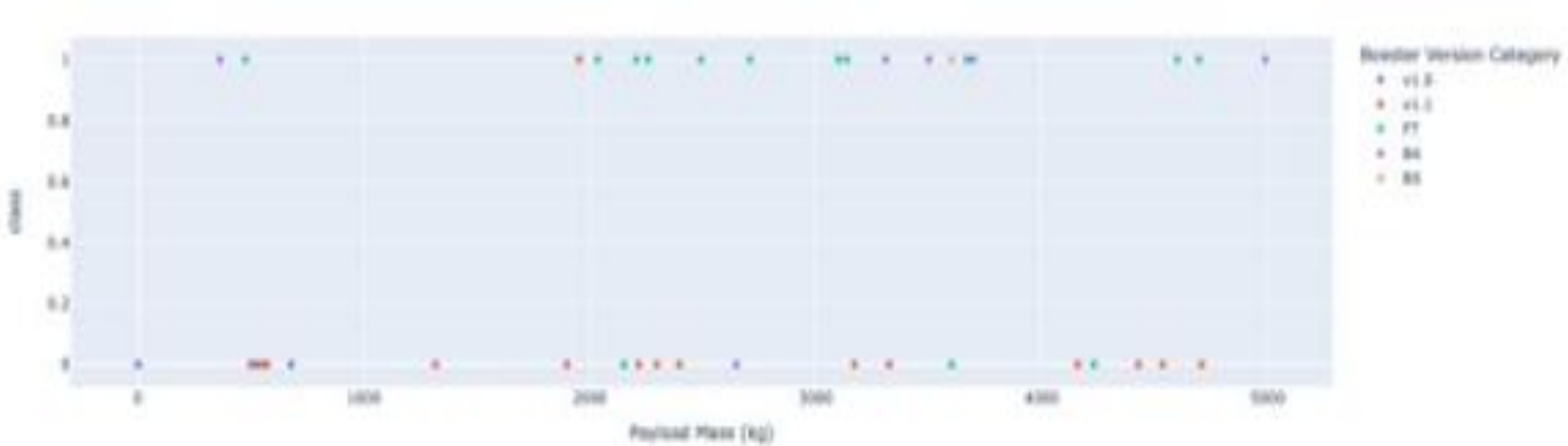


This site has a not so good rocket stage 1 recovery rate

Just over 50% of the launches had a successful recovery outcome

Payload vs Launch Outcome by Booster Category





Two charts with varying payload selected on the adjustable payload range selected

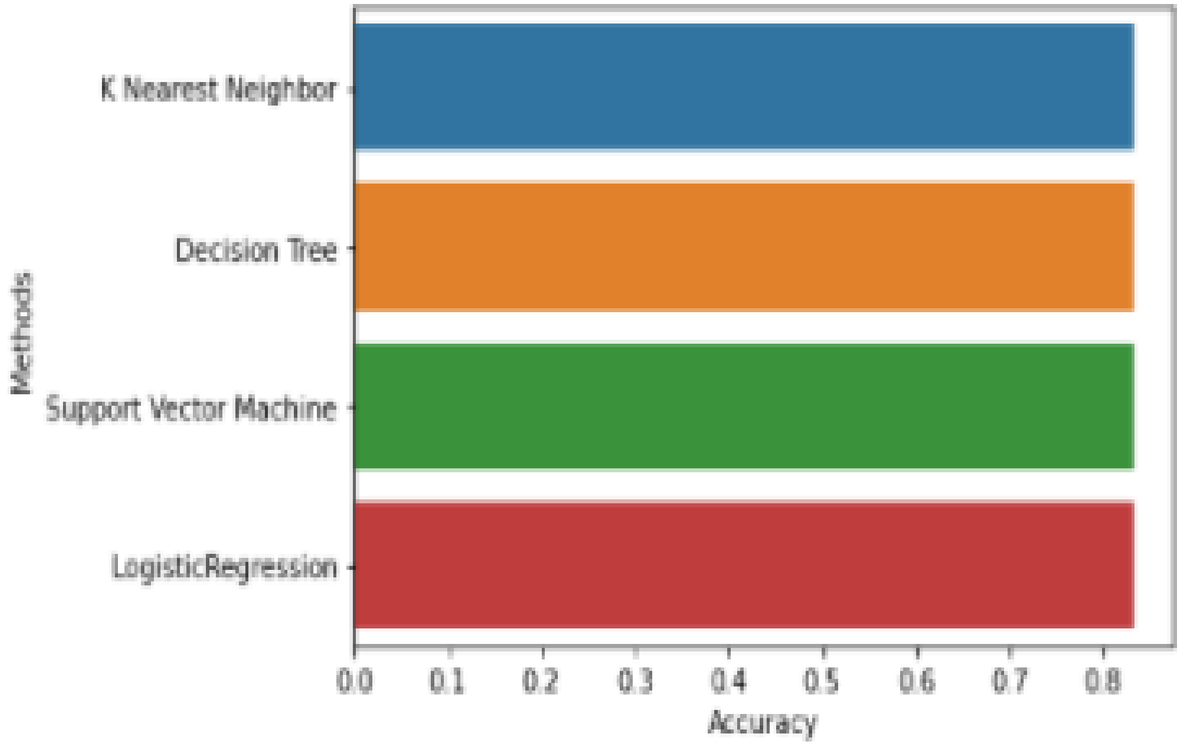
Most failures occur at mid payload range of 2 to 5K

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- All models have similar accuracy in low 80s
- Small sample size is likely the culprit



Confusion Matrix of Decision Tree Classifier Model

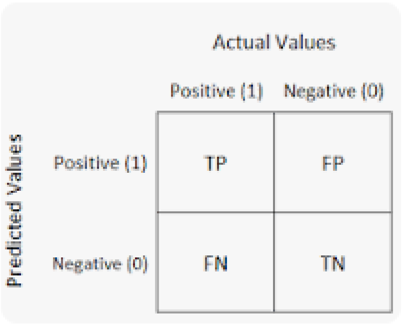
The confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.

True Negative -12
False Negative- 0
True Positive – 3
False Positive -3

The accuracy as calculated from the matrix is 83.3%

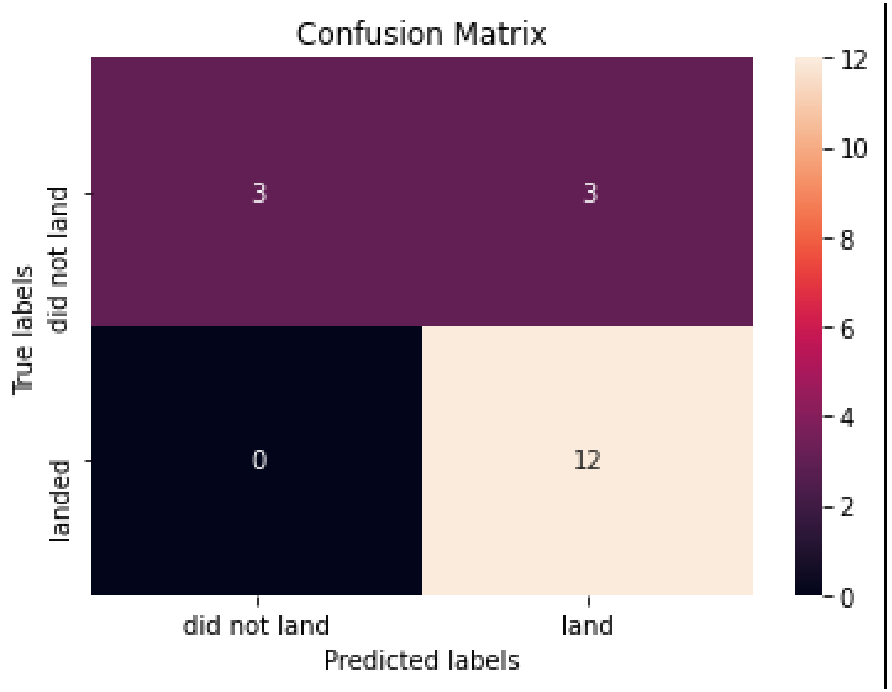
Confusion Metrics

1. Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN.
2. Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN.
3. Precision (true positives / predicted positives) = TP / TP + FP.
4. Sensitivity aka Recall (true positives / all actual positives) = TP / TP + FN.



Confusion Matrix of Decision Tree Classifier Model

- The confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.



- True Negative -12
- False Negative- 0
- True Positive – 3
- False Positive -3

- The accuracy as calculated from the matrix is 83.3%
- We need to increase sample size to see if matrix is similar

Conclusions

- Success rate have significantly increased since 2013
- Models show low 80% accuracy rate on small sample test size
- KSC LC-39 has the best success rate
- Several Orbits have similar high success rate
- Early launches or lower flight number went to GTO, PO, ISS & LEO and had a lower success rate compared to the latter launches which
- Most failures are from the mid size payload category
- Higher payloads mostly went to VLEO orbit

Appendix

Thank you!