IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Jonathan Kim>
<Sep 6, 2022>

Outline

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

Executive Summary

- Summary of methodologies

Data collected from public SpaceX API and Wikipedia and explored with SQL, visualization, folium mapping and dash board. Narrow down columns for features.

Data normalized and standardized to run machine learning models.

And visualized accuracy scores of model.

- Summary of all results

Applied four machine learning model;

Logistic regression, Support vector machine, Decision tree classifier, and N nearest neighbors.

All models produced similar results around 83.3 %

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is due to the fact that SpaceX can reuse the first stage

Develop model which can perform high success rate as SpaceY to compete with SpaceX

- Problems you want to find answers

Train machine learning model to predict successful stage 1 recovery rate

Section 1

Methodology

Methodology

Executive Summary

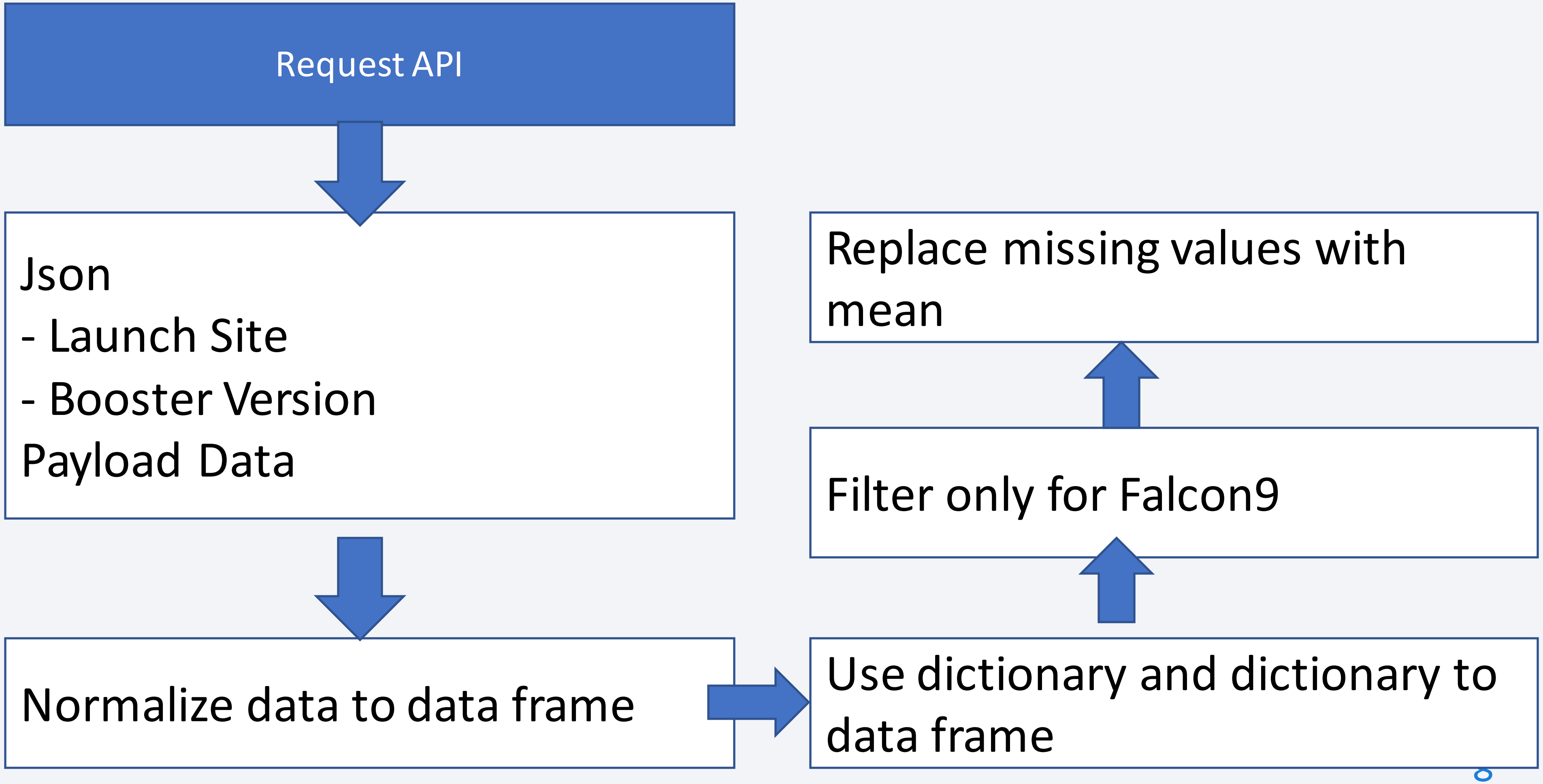
- Data collection methodology:
  - Data from public SpaceX API and Wikipidia
- Perform data wrangling
  - Select relevant columns, normalize, filter for falcon9, identify missing values, classifying landings with successful and unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
  - analyze and investigate data sets and summarize their main characteristics
- Perform interactive visual analytics using Folium and Plotly Dash
  - Visualize data on the map with other characteristics
- Perform predictive analysis using classification models

Data Collection

- Multiple data sets from SpaceX API and table from SpaceX Wikipedia web site
    - BoosterVersion
    - LaunchSite
    - PayloadData
    - CoreData
- You need to present your data collection process use key phrases and flowcharts

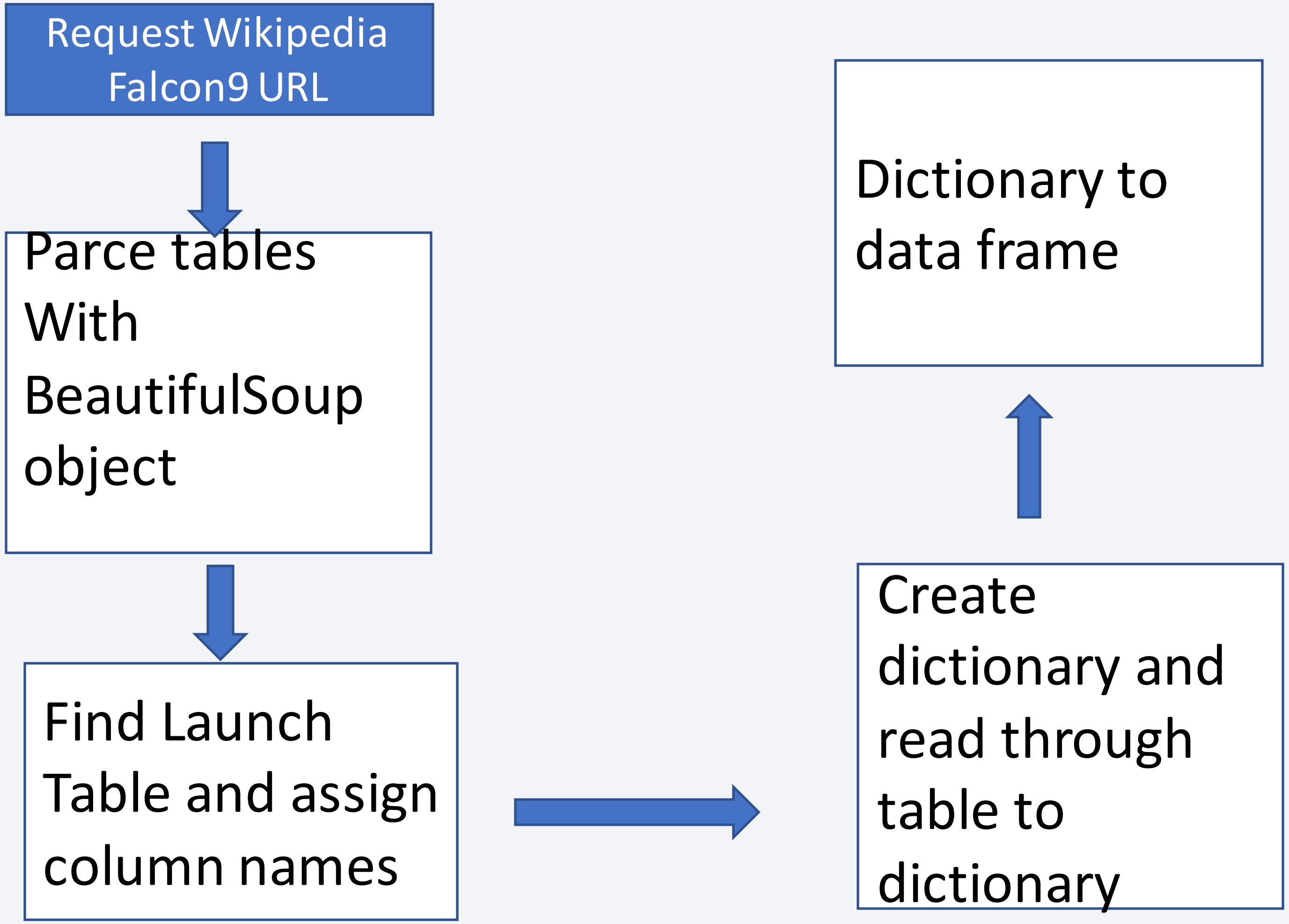
Data Collection and preparation – SpaceX API

- https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/FCourse10-spacex-data-collection-api.ipynb



Data Collection – Web Scraping

https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/Fcourse10-webscrapping.ipynb



9

Data Wrangling

- Fill missing and null values
- Create Mission outcome class with 0=Failure, 1=Success
- Outcome two keys : Mission Outcome and Landing Location
- Success rate with mean of class

- Github URL

https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/FCourse10%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Scattered Plot with Pay load Mass and Flight Number, Launch Site and Flight Number, Launch Site and Pay load Mass to find linear relation
- Bar chart group by Orbit and Class to find the success rate per Orbit
- Line plot with year and success rate to find relation between them

- Github URL

https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/FCourse10-eda-dataviz.ipynb

Build an Interactive Map with Folium

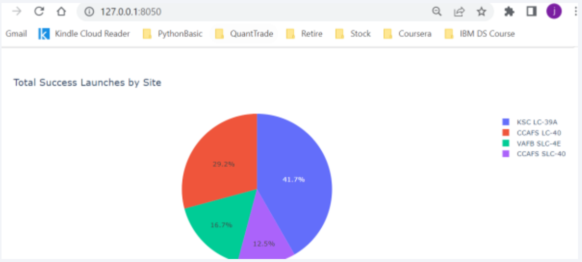
- Mark Launch Site, Success rate, proximity example on the map to visualize the success landings related to locations

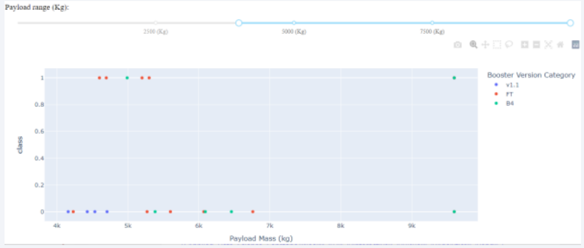
- GitHub URL

http://localhost:8888/notebooks/FCourse10-launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart with launch site and success rate
- Scattered plot with PayLoad Mass and success rate and hue with booster version under slide bar of Payload Mass





- GitHub URL

https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/FCourse_Dashboard.ipynb

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

https://github.com/jonathanykim665/Coursera-Course-10-Final/blob/main/FCourse10-SpaceX-ML-Prediction.ipynb

Load data from CVS to data frame

Class to series
Transform data frame with standardscaler

Split train and test set

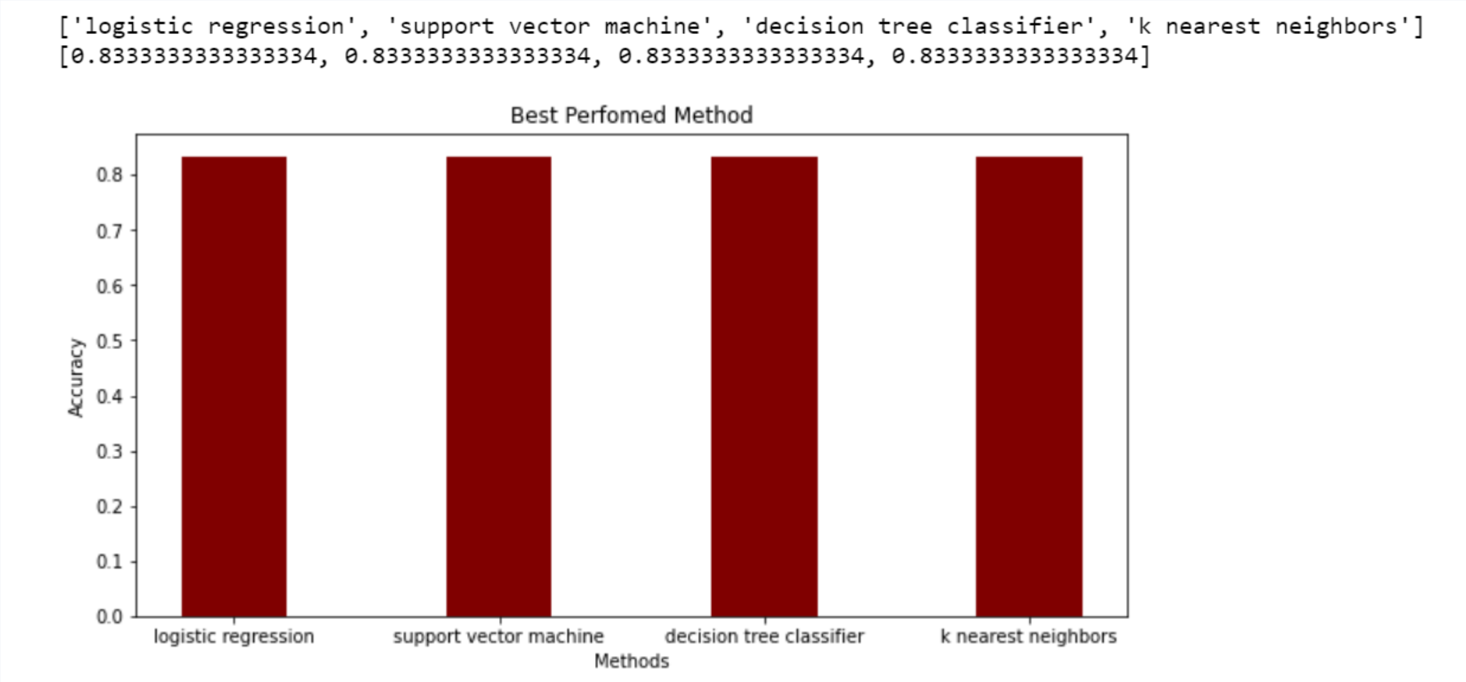
Grid search with cv=10
Calculate accuracy with score

Grid search for logregression, SVM, decision tree, KNN model

Score Model with test set
Bar chart to compare scores of models

Results

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

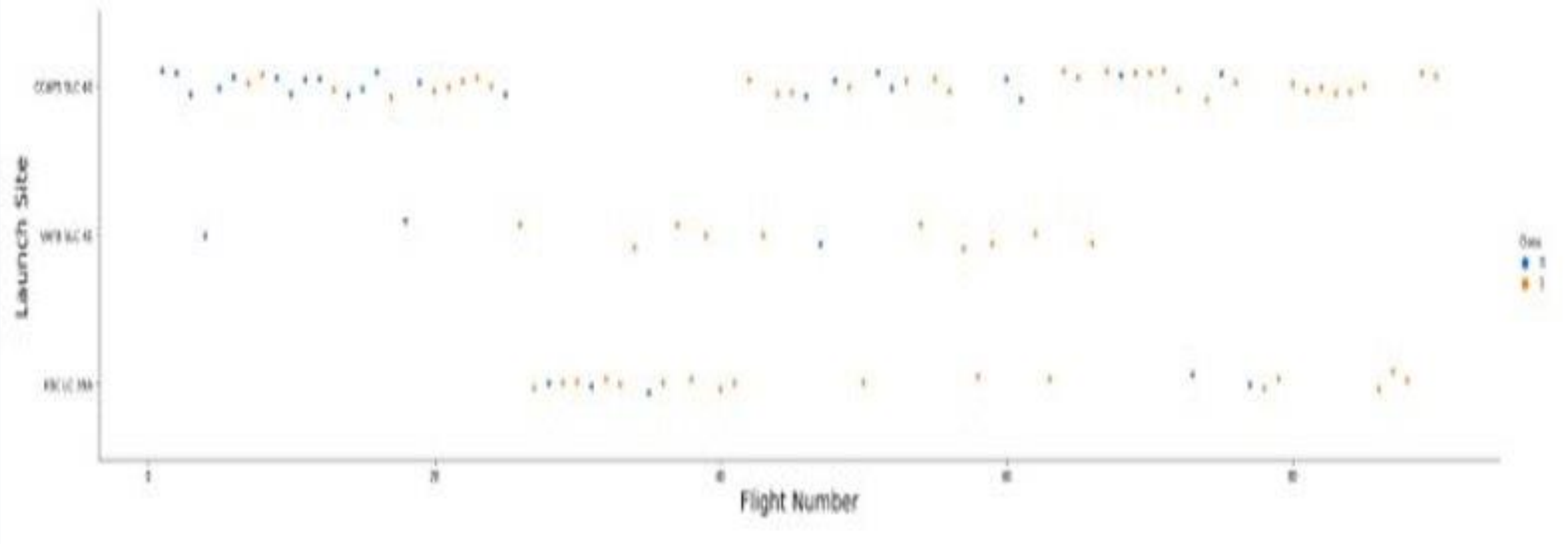


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

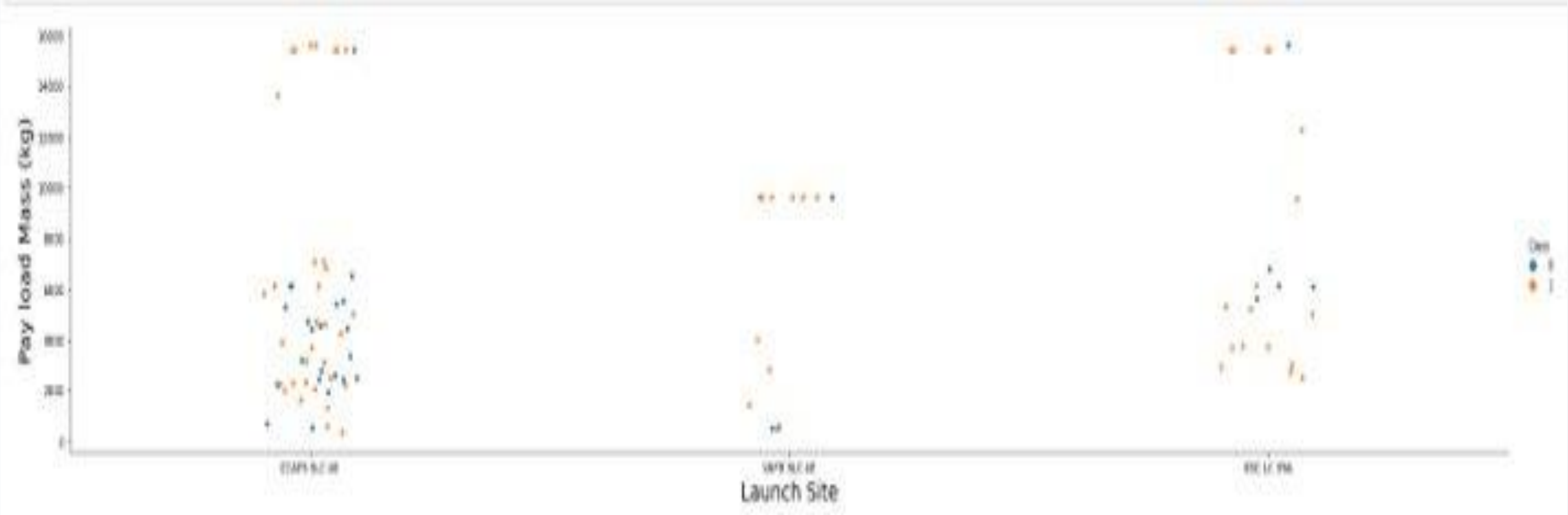
- Scatter plot of Flight Number vs. Launch Site



Increased success rate over time, CCAFS shows the most number of launches

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

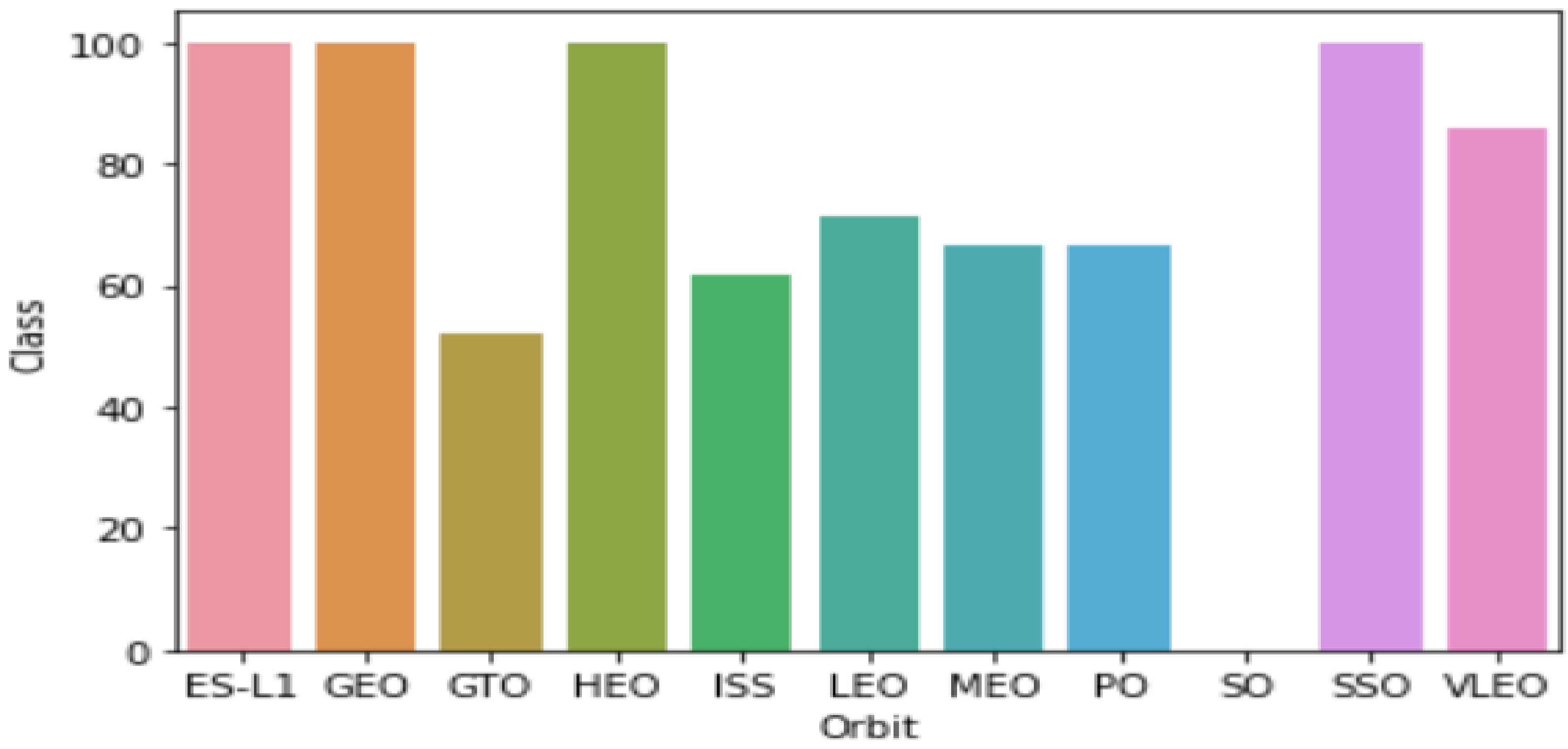


Most of Payload is less than 6000, Each launch site show different Payload

Success Rate vs. Orbit Type

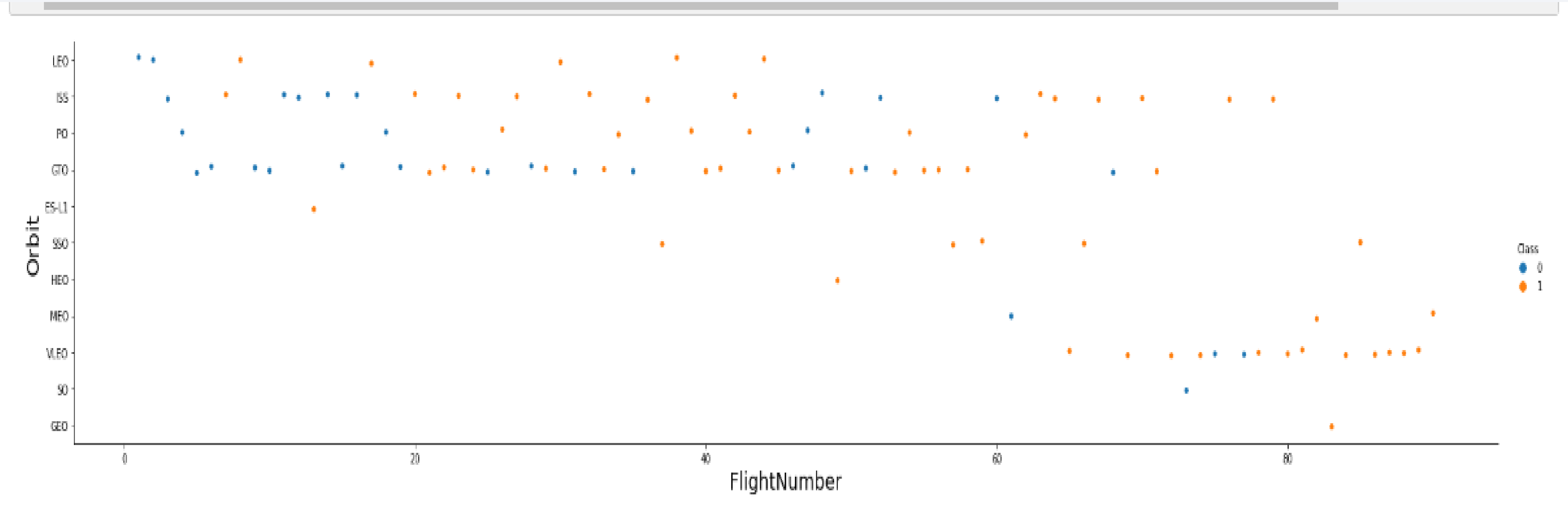
- Bar chart for the success rate of each orbit type

<AxesSubplot:xlabel='Orbit', ylabel='Class'>



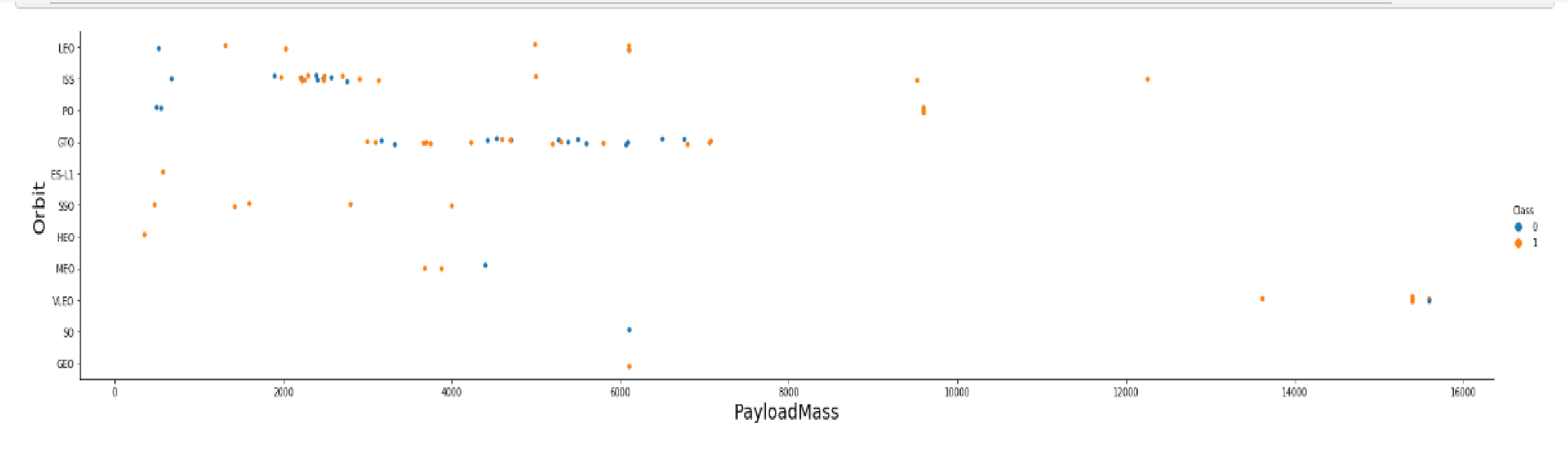
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type : LEO orbit success relate to the number of flight



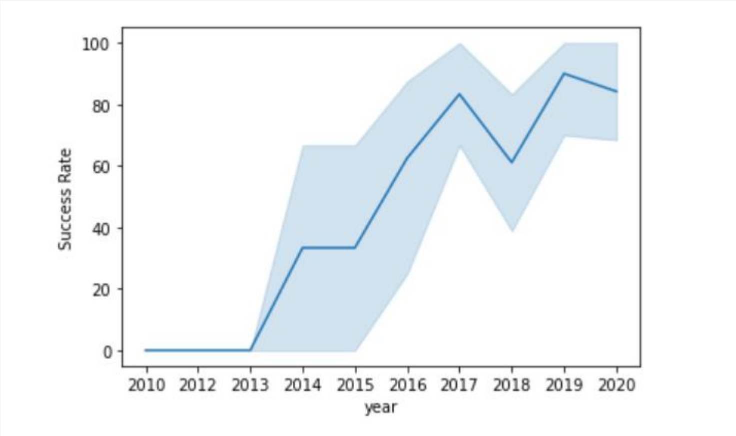
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type
- - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS

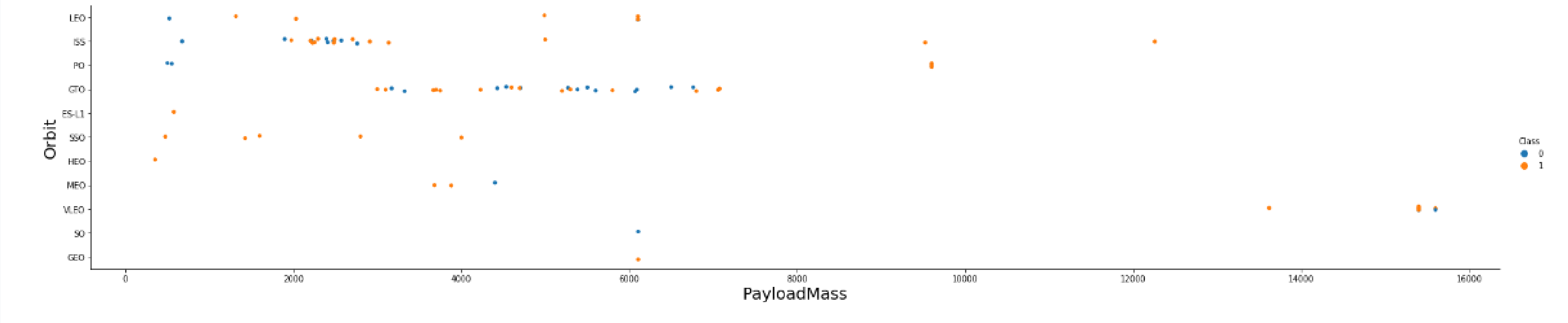


Launch Success Yearly Trend

- Line chart of yearly average success rate



- Screenshot of the scatter plot with explanations – From 2013 success rate increase.



All Launch Site Names

- Find the names of the unique launch sites

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f
        Done.
Out[4]:
```

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

- Present your query result with a short explanation here

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

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_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Landing Date

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

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014.

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 SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload

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

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
[illegible]	[illegible]

This query returns the booster versions that carried the highest payload mass of 15600 kg.

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

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(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

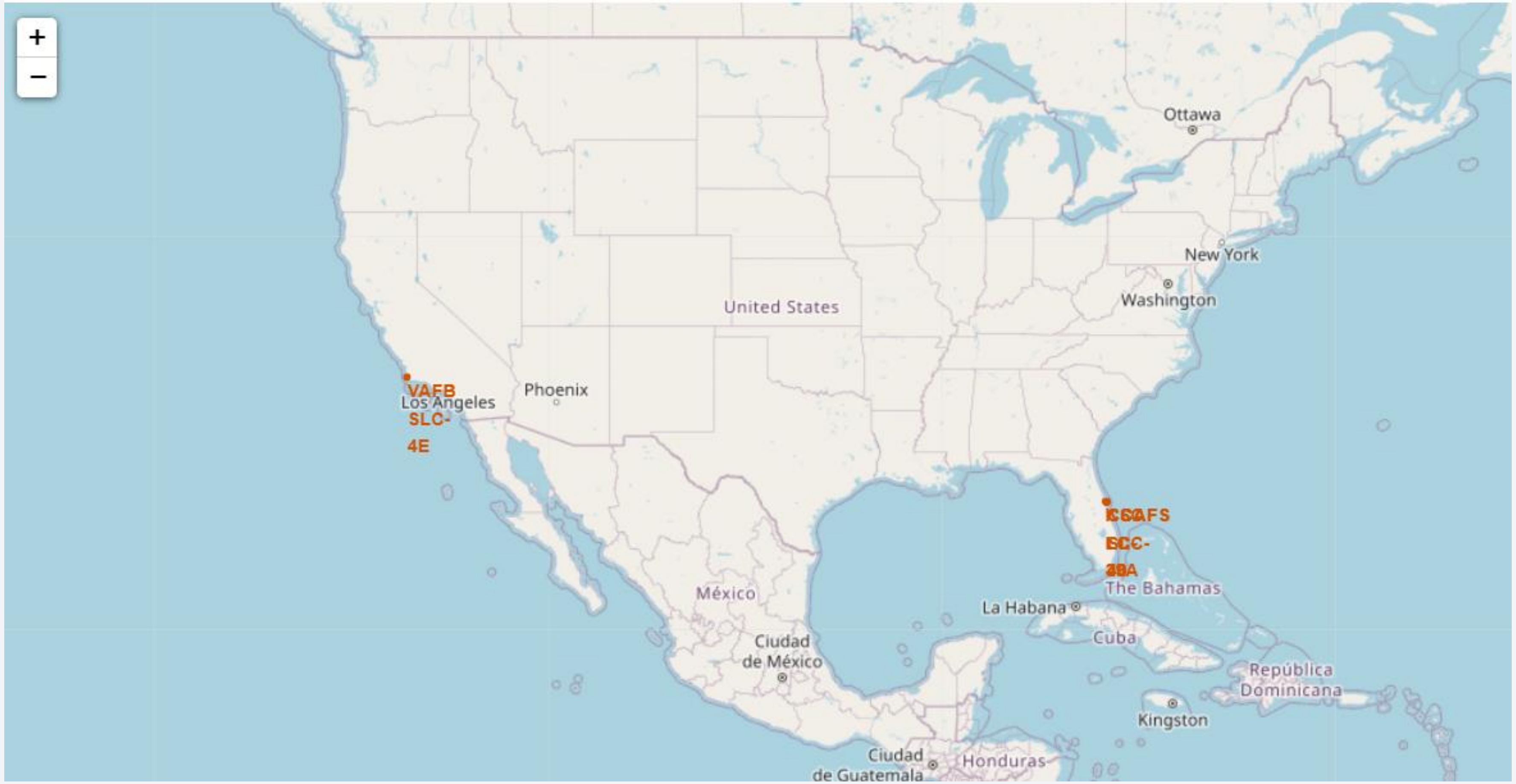
There are two types of successful landing outcomes: drone ship and ground pad landings.

There were 8 successful landings in total during this time period

Section 3

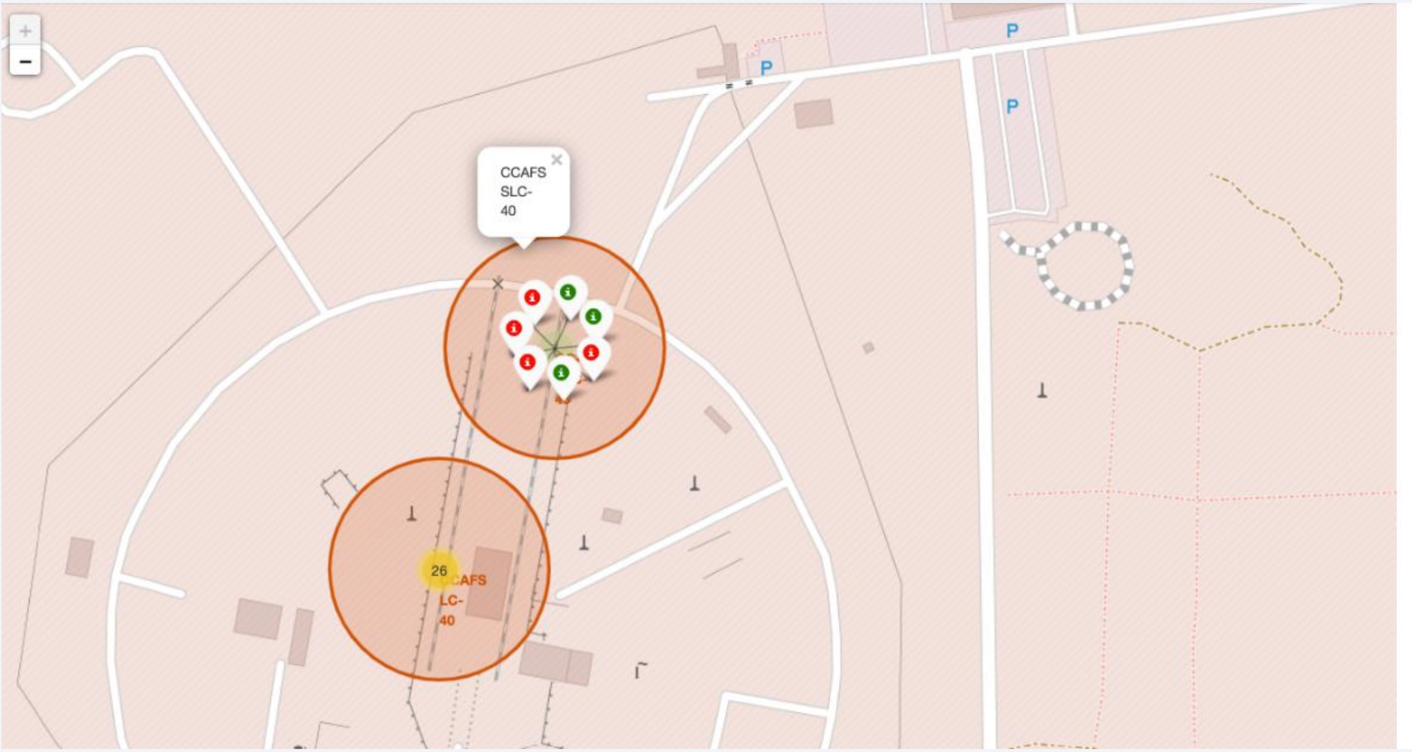
Launch Sites Proximities Analysis

Launch Site Locations



- Launch Site close to Ocean

Color coded marker at Launch site



- Map shows three success and four failure landing at CCAFS SLC-40 site

Key Location Proximities

ected
h

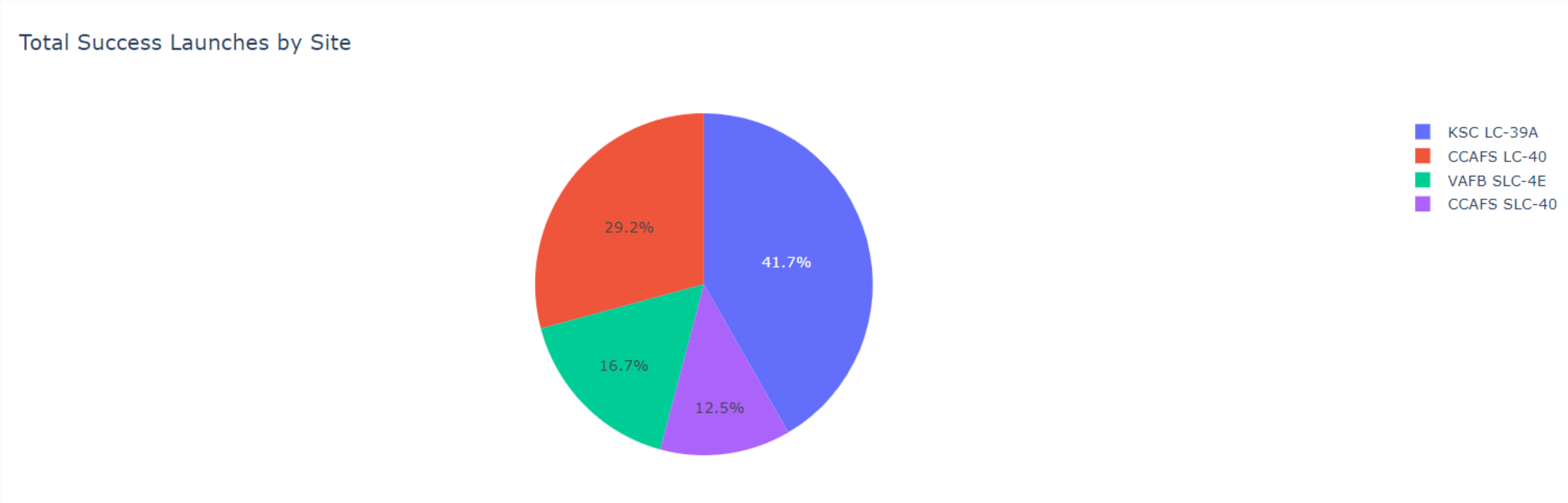
- CCAFS SLC-40 site and AFS LC-40 shows close to supply transport

Section 4

Build a Dashboard with Plotly Dash

<launch success count for all sites>

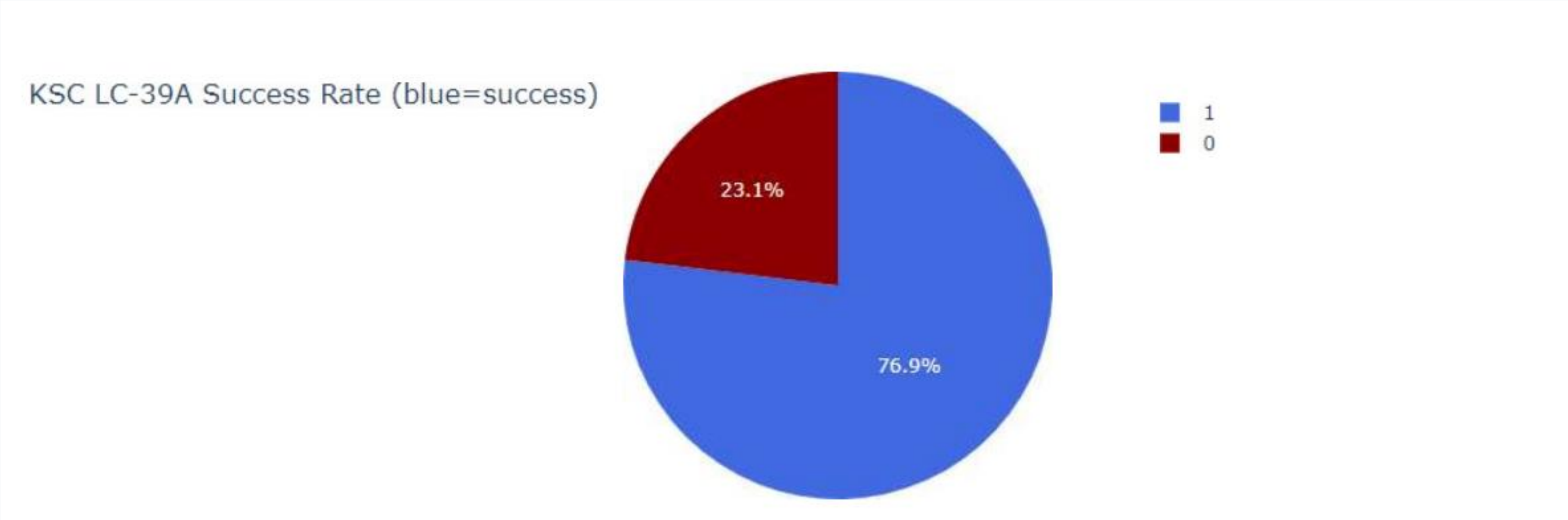
- Screenshot of launch success count for all sites, in a piechart



- Presents distribution of successful landings across all launch sites.
- KSC LC-40 shows the best successful landing.

<launch site with highest launch success ratio>

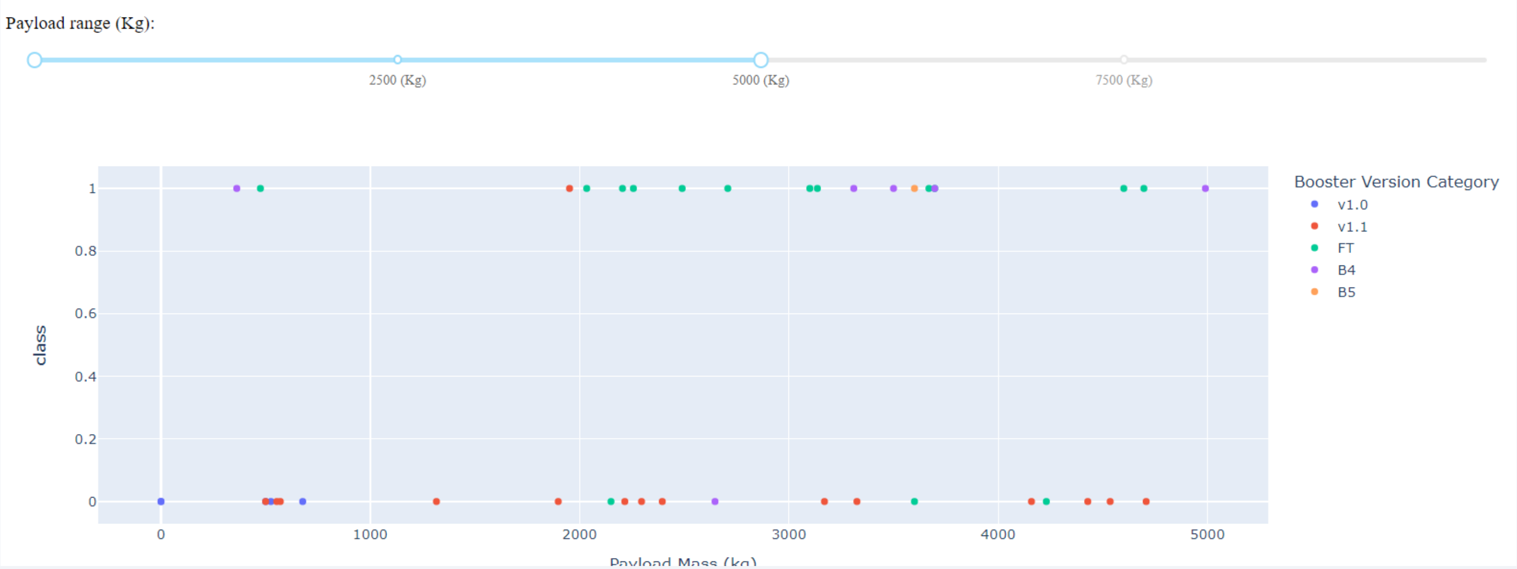
- Screenshot of the piechart for the launch site with highest launch success ratio



- KSC LC-39A has the highest success rate with 10 successful and 3 failed landings.

<Payload vs. Launch Outcome>

- Screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



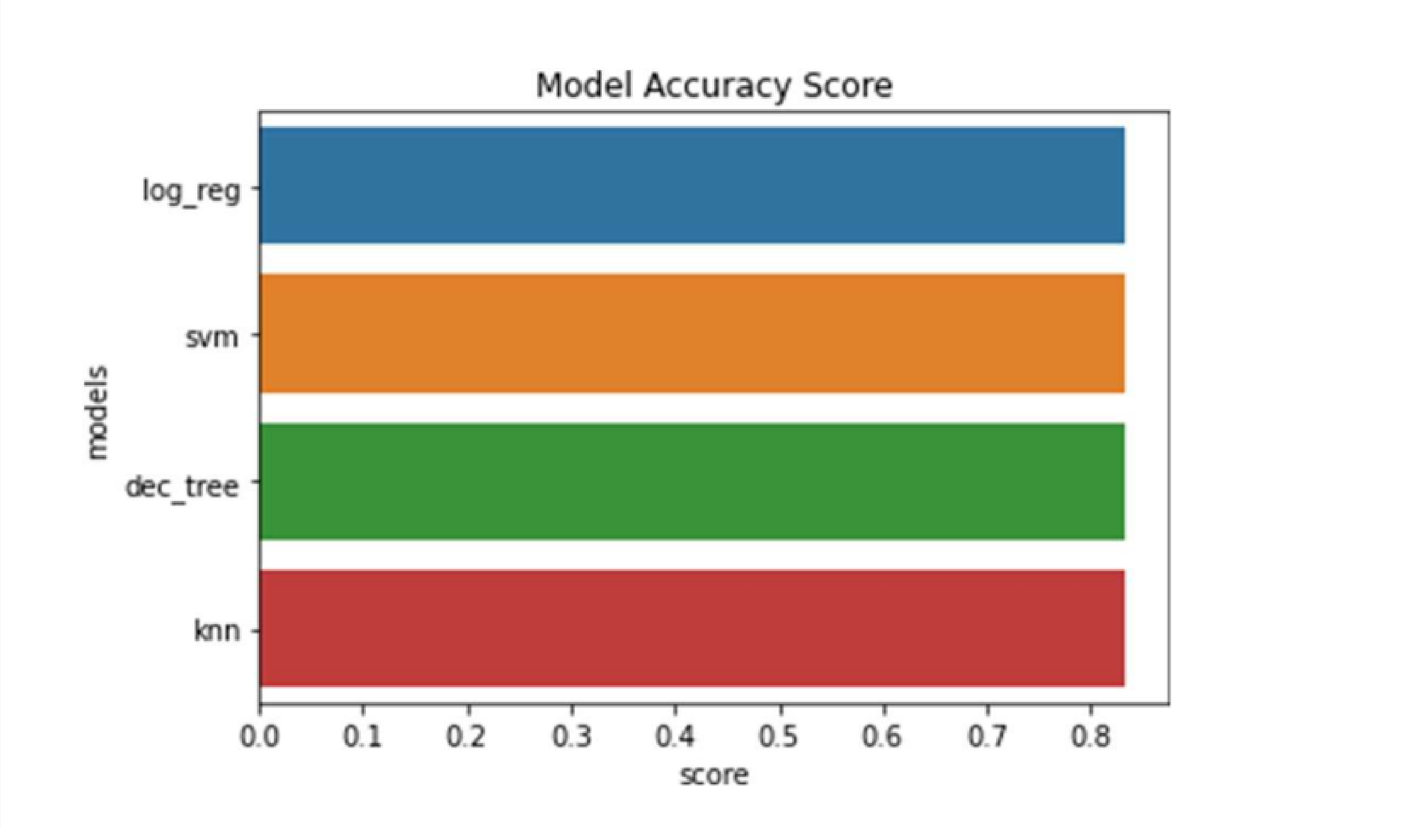
- Class indicates 1 for successful landing and 0 for failure. Scatter plot also

accounts for booster version category in color and number of launches in point size

Section 5

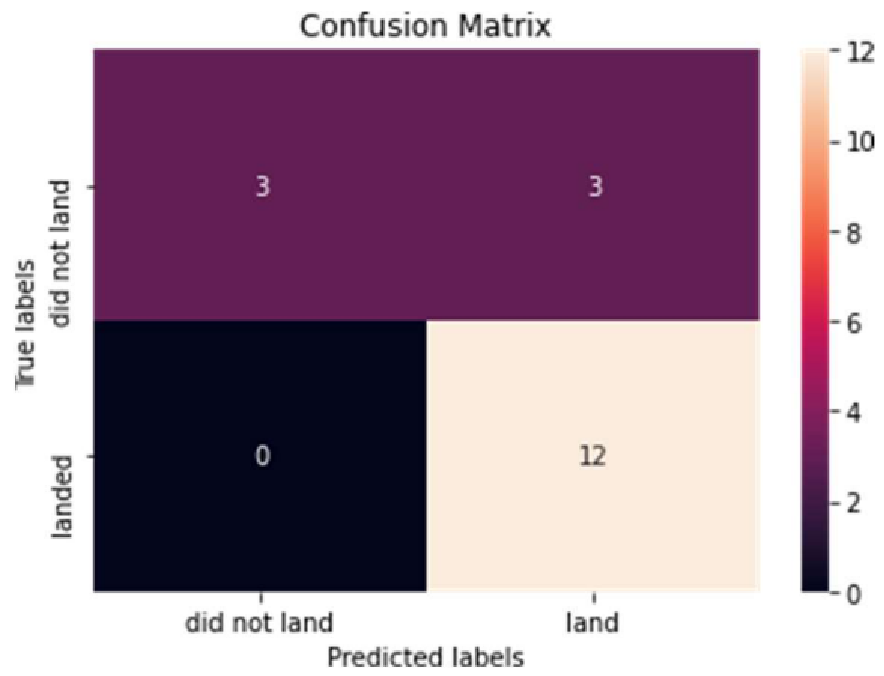
Predictive Analysis (Classification)

Classification Accuracy



- All models had virtually the same accuracy on the test set at 83.33% accuracy.

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

- the confusion matrix is the same across all models, since all models performed the same for the test set.

Conclusions

- Goal is to develop a machine learning model for Space Y who wants to bid agains

- to lower the cost by reusing stage 1 from successful landing.
- - collect data from public SpaceX API and web scrapping SpaceX from Wikipedia
- - create data labels and store into DB2 database
- - create dashboard to visualize data
- - create machine learning model
- - use the model to predict success rate with high accuracy

Thank you!