Findings for SpaceY

Nikisha Mistry

13-10-2022

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models.
- Four models were made using: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

INTRODUCTION

- SpaceX is claiming to make rockets at a cost of $62Million and other companies cost is around $165Million.
- SpaceX is able to use its first stage. We have to analyze if first stage is successful or not
- SpaceY wants to compete with SpaceX
- Our task is to predict successful stage1 by creating models.

Data collection and data wrangling

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.
- Space X API Data Columns:
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
- Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Wrangling:

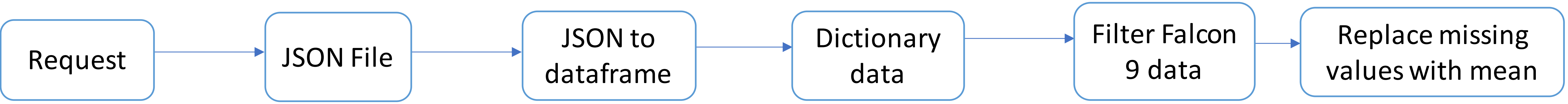
Create a training label with landing outcomes where successful = 1 & failure = 0.

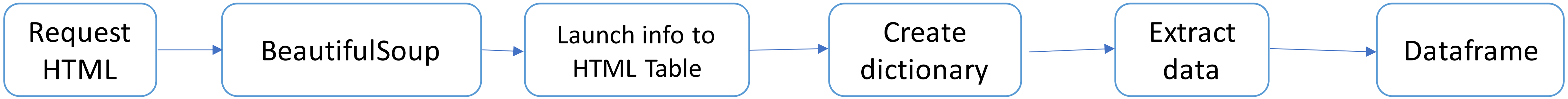
Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. <u>Value Mapping:</u>

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

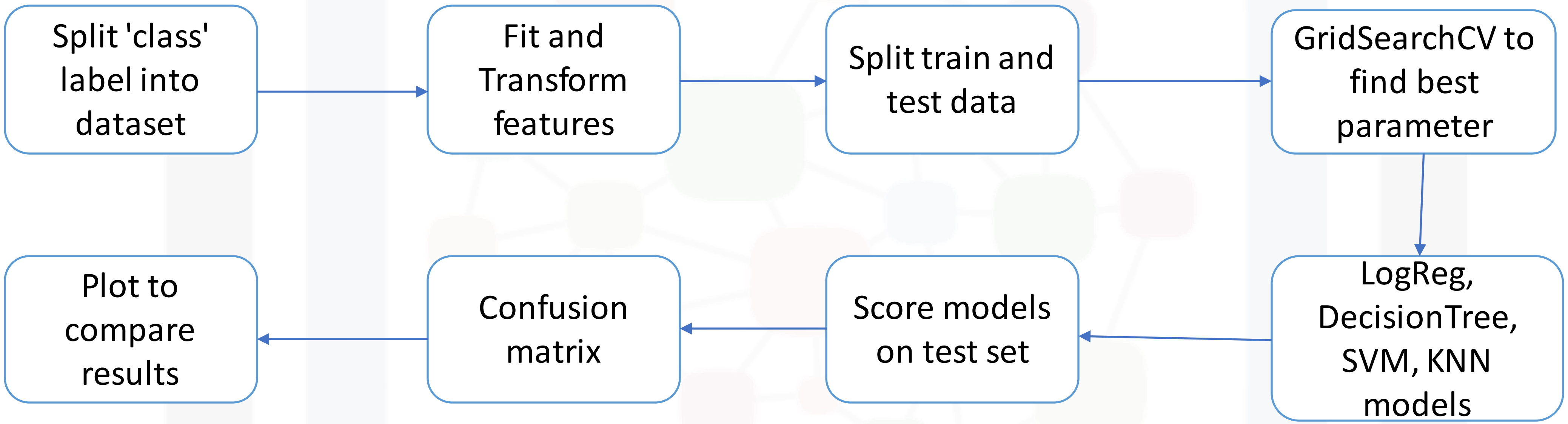




EDA and interactive visual analytics methodology

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- <u>Plots Used:</u>
- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to
- decide if a relationship exists so that they could be used in training the machine learning model

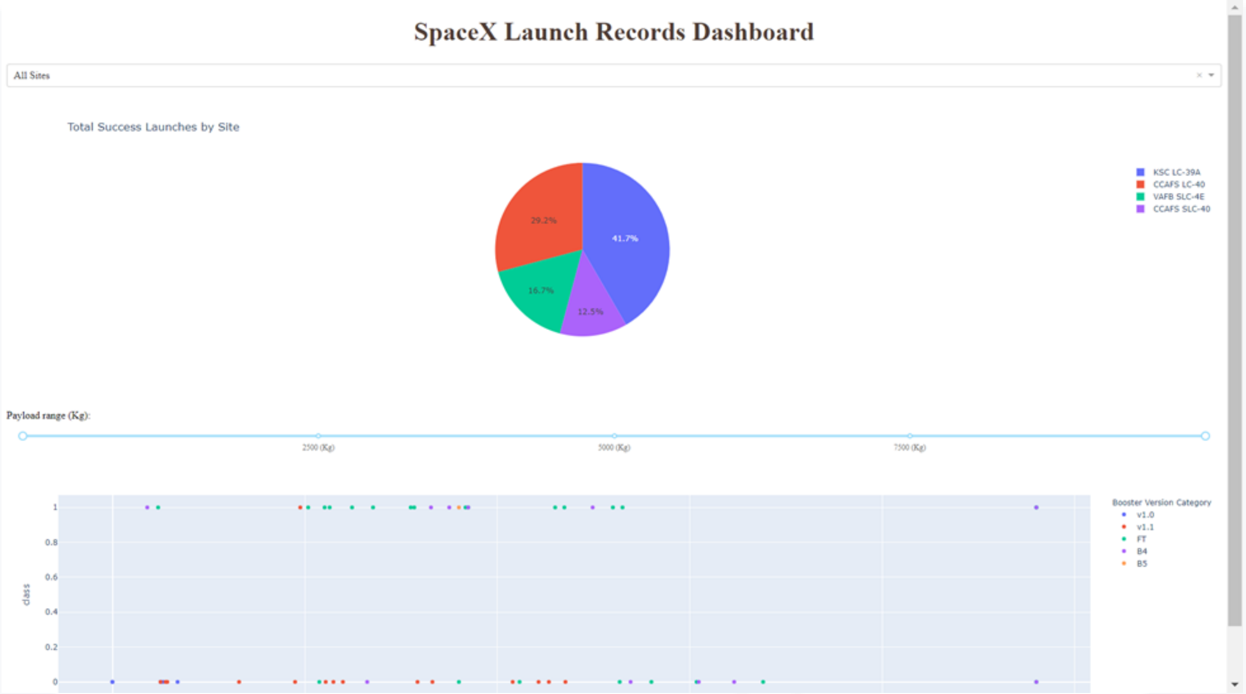
Predictive analysis methodology



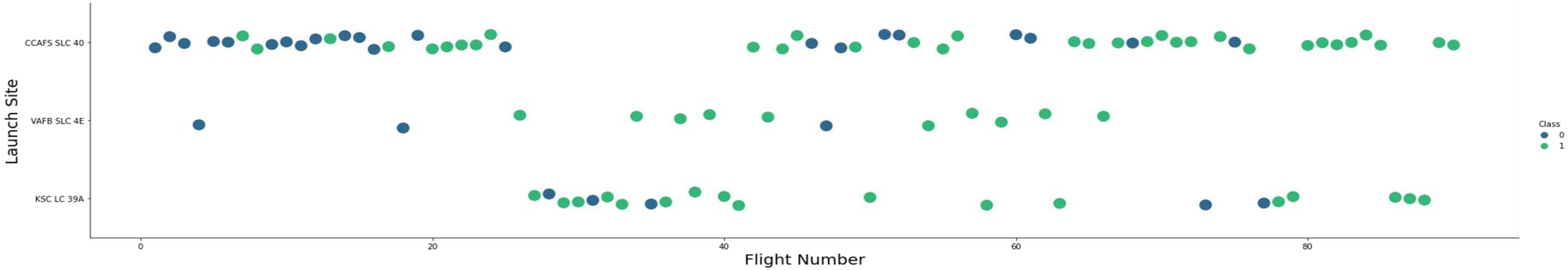
METHODOLOGY

- Data collection methodology:
- Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling: Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: Tuned models using GridSearchCV

EDA with visualization results

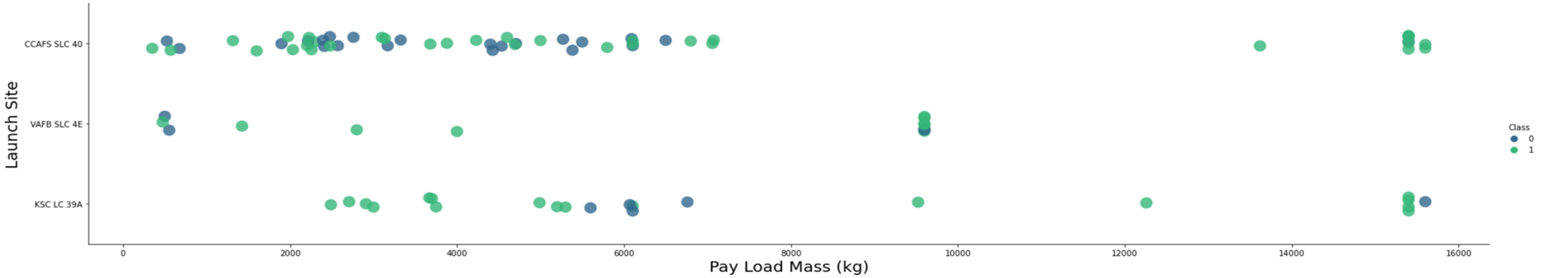


The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.



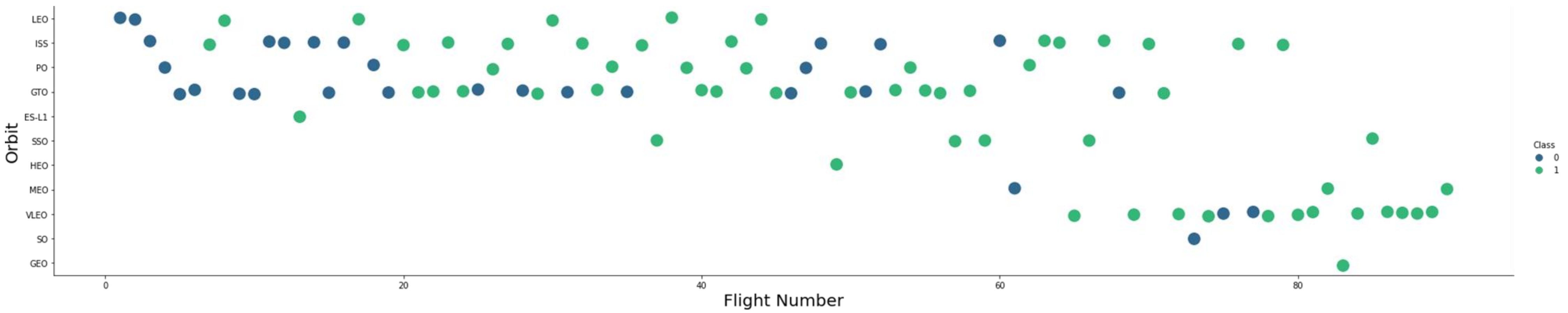
Green dots indicate successful launches and purple represent unsuccessful launches.

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

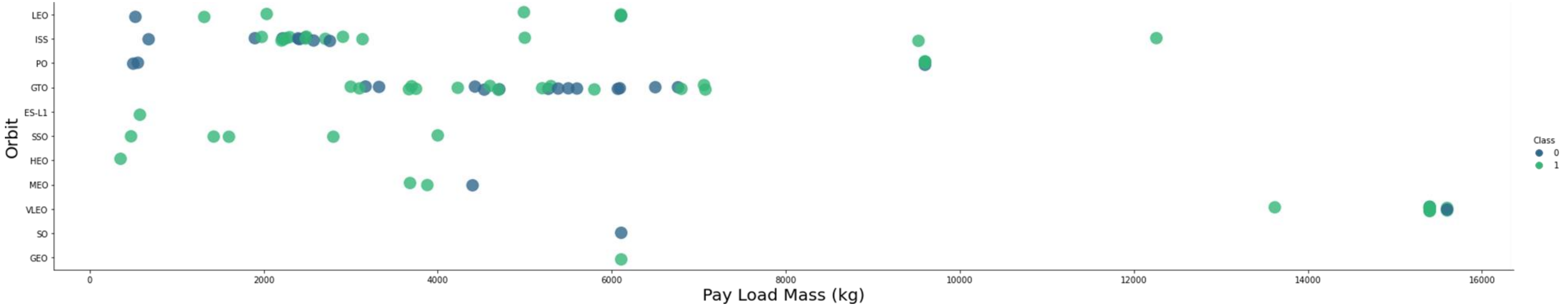


Green dots indicate successful launch and purple represent unsuccessful launch.

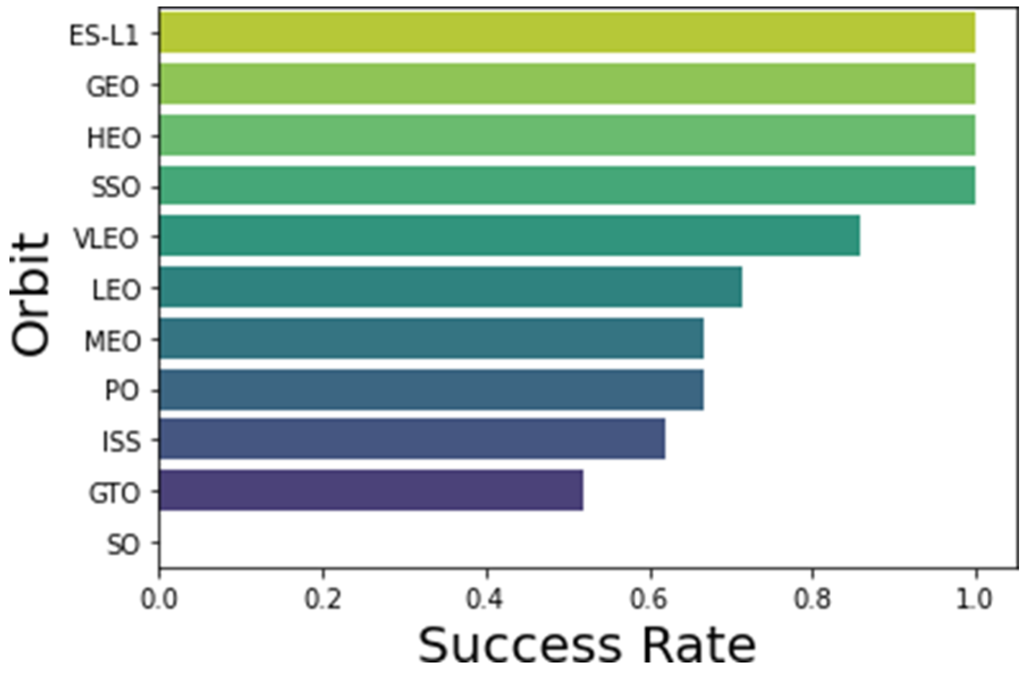
Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.



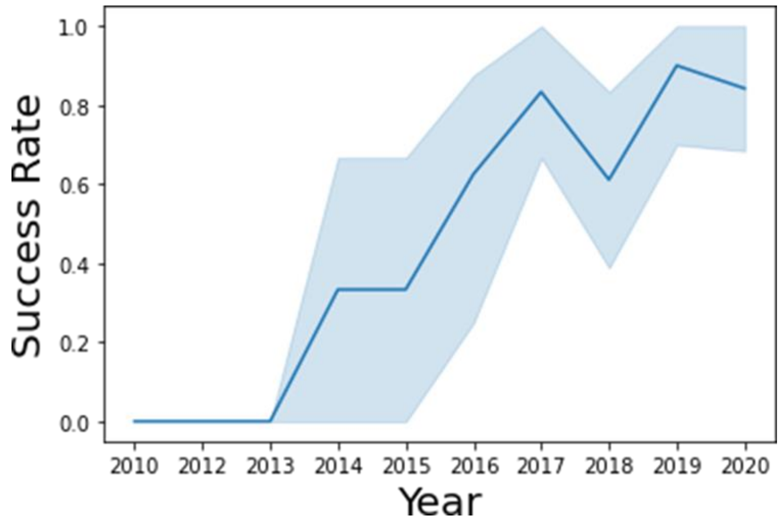
Green indicates successful launch; Purple indicates unsuccessful launch.
Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference.
SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits



Green indicates successful launch; Purple indicates unsuccessful launch.
Payload mass seems to correlate with orbit. LEO and SSO seem to have relatively low payload mass
The other most successful orbit VLEO only has payload mass values in the higher end of the range



- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample



Success generally increases over time since 2013 with a slight dip in 2018
Success in recent years at around 80%

EDA with SQL results

List of all launch sites

In [9]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

Out[9]:

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

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

In [10]:
```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5;
```

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

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

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

In [13]:
```
%sql select sum(payload_mass__kg_) from spacextbl where customer='NASA (CRS)' group by customer;
```

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

Out[13]:

sum(payload_mass__kg_)
45596

average payload mass carried by booster version F9 v1.1

In [14]:
```
%sql select avg(payload_mass__kg_) from spacextbl where booster_version='F9 v1.1' group by booster_version;
```

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

Out[14]:

avg(payload_mass__kg_)
2928.4

the date when the first successful landing outcome in ground pad was acheived.

In [62]:
```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

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

Out[62]:

Date
01-03-2013

IBM Developer SKILLS NETWORK

names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [60]:
```
%sql select booster_version from spacextbl where "Landing _Outcome" like 'Success (drone ship)' and payload_mass__kg_ between 4001 and 6000 and missio
```

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

Out[60]:

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

total number of successful and failure mission outcomes

In [65]:
```
%sql SELECT mission_outcome, count(*) as Count FROM spacextbl GROUP by mission_outcome
```

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

Out[65]:

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

names of the booster_versions which have carried the maximum payload mass.

```
In [66]: %sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ in (select max(payload_mass__kg_) from spacextbl);
```

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

Out[66]:

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
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
In [86]: %sql select substr(Date, 4, 2) as month, "Landing _Outcome", booster_version, launch_site from spacextbl where "Landing _Outcome"='Failure (drone ship
```

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

Out[86]:

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

the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [95]:
```
%sql select "Landing _Outcome", count(*) as count from SPACEXTBL where date between '04-06-2010' AND'20-03-2017'  GROUP by "Landing _Outcome" ORDER BY
```

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

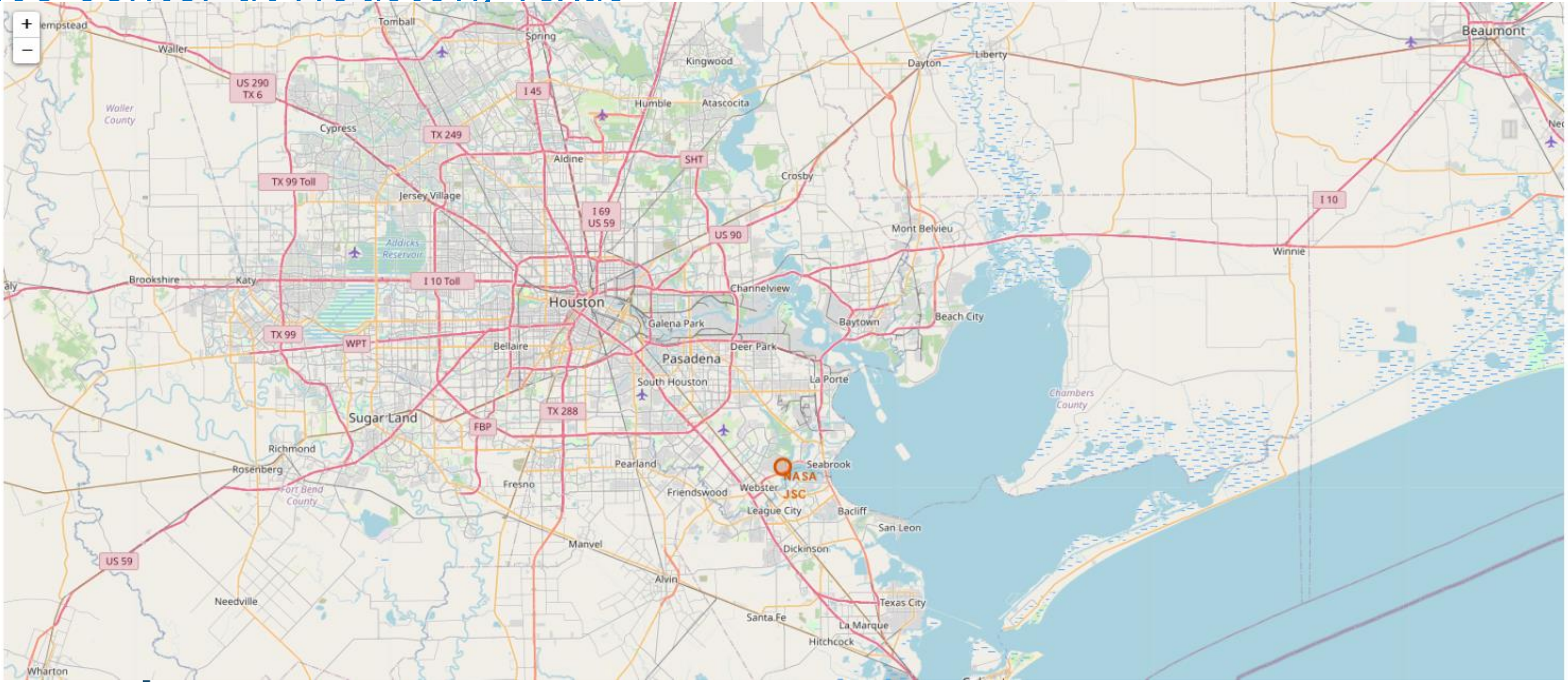
Out[95]:

Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Interactive map with Folium results

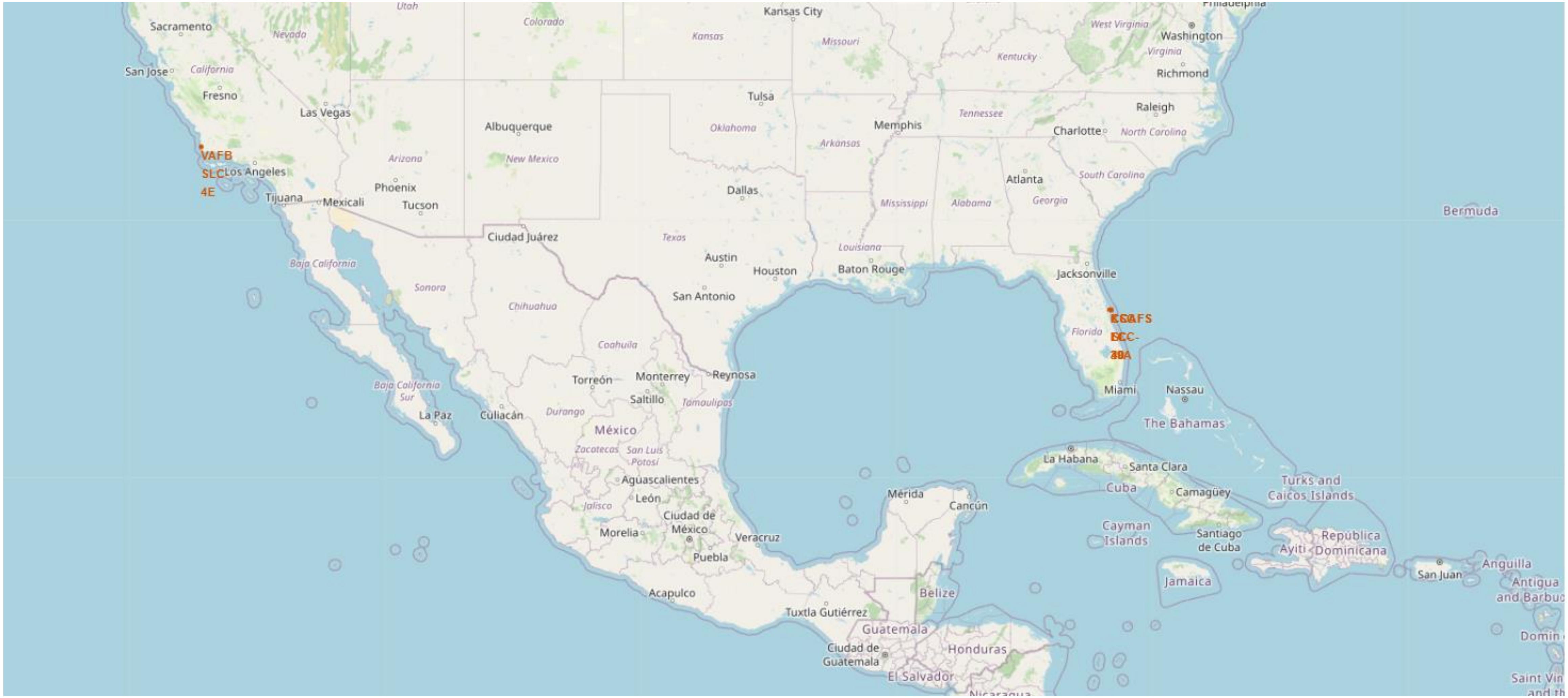
- create a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas

[7]:



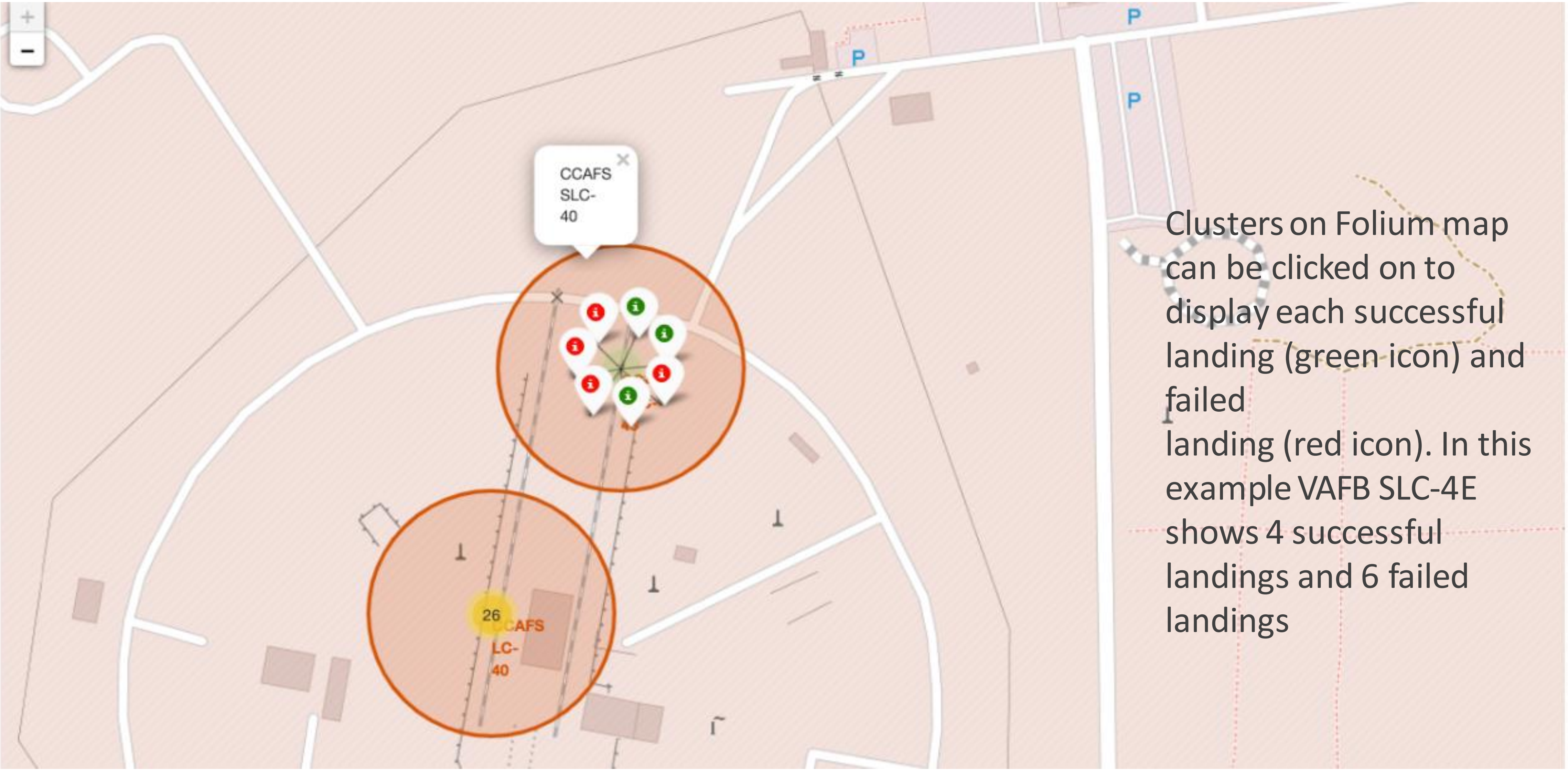
IBM Developer

SKILLS NETWORK

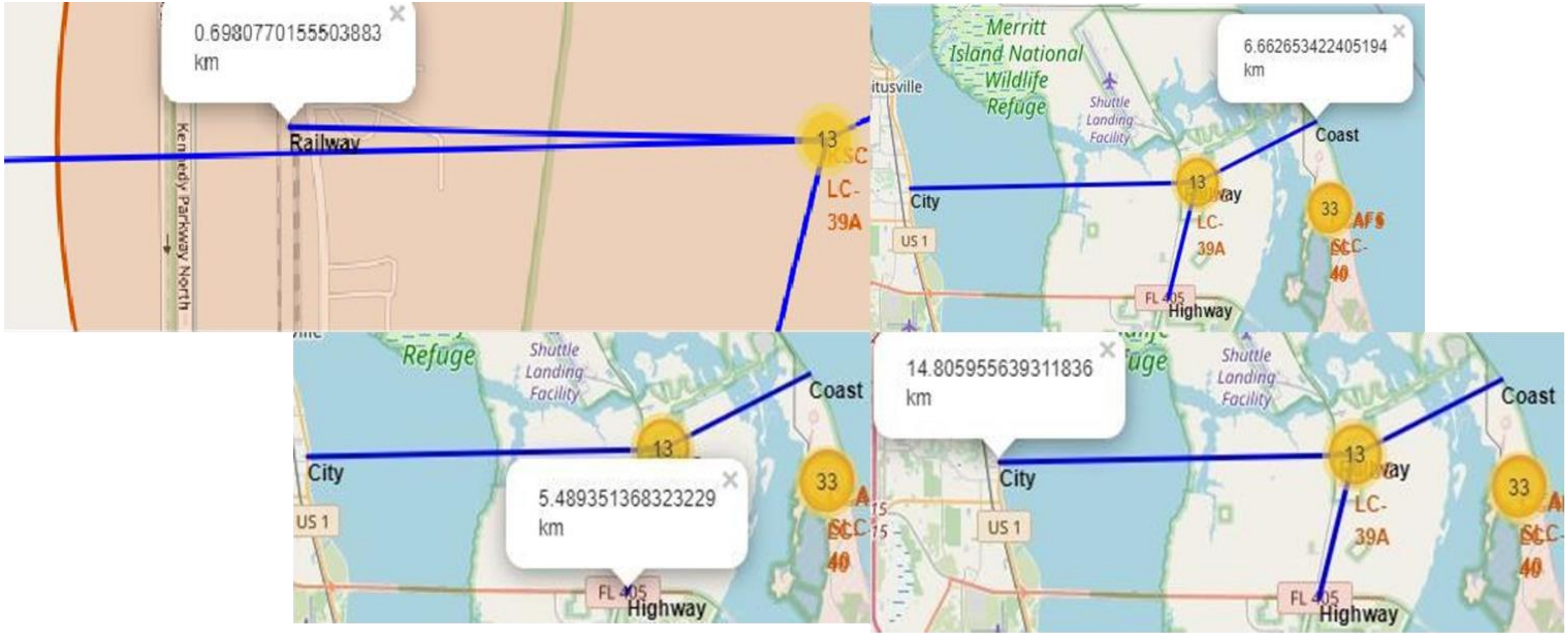


IBM Developer

SKILLS NETWORK

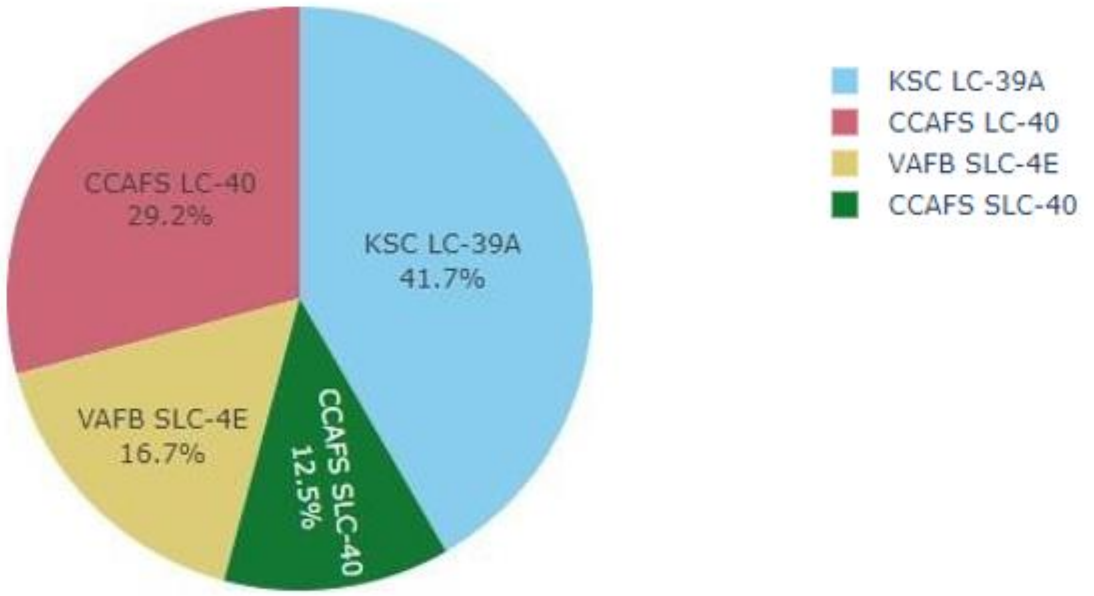


Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings



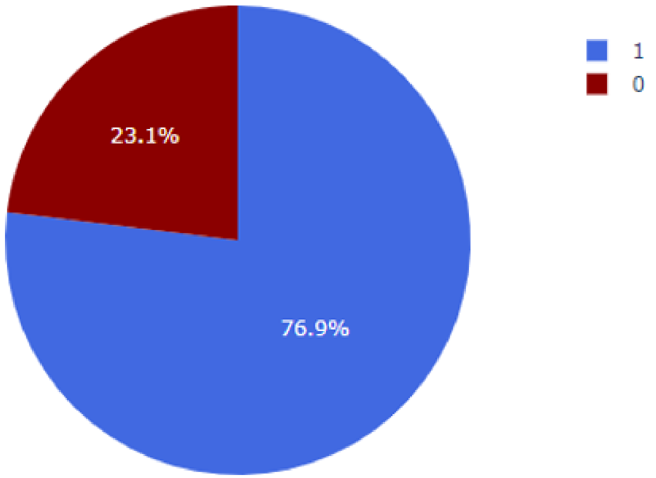
Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

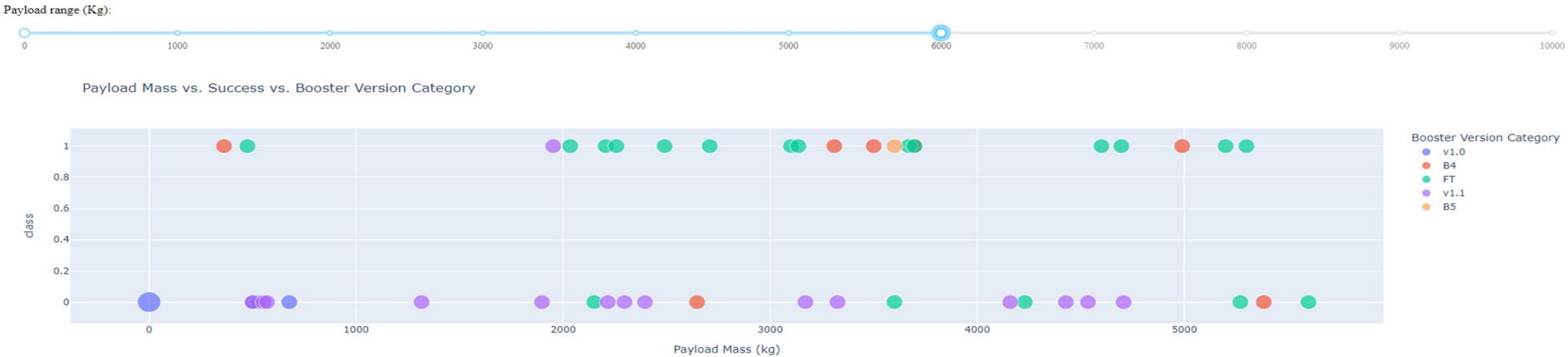
Plotly Dash dashboard results



This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

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

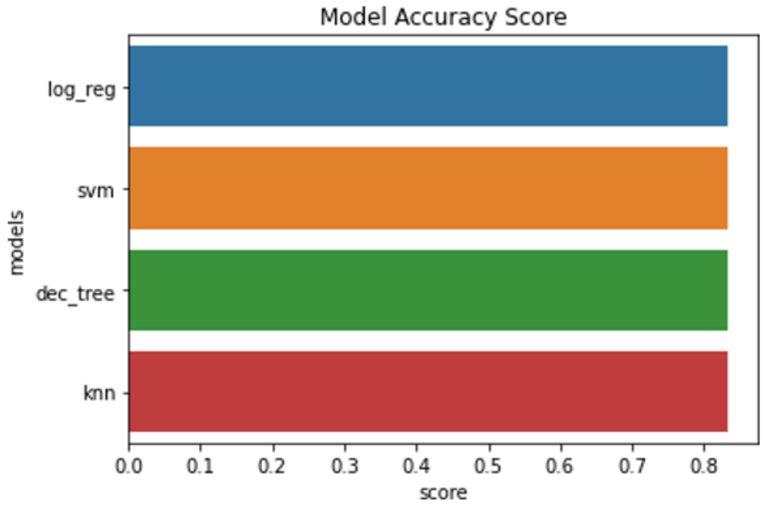




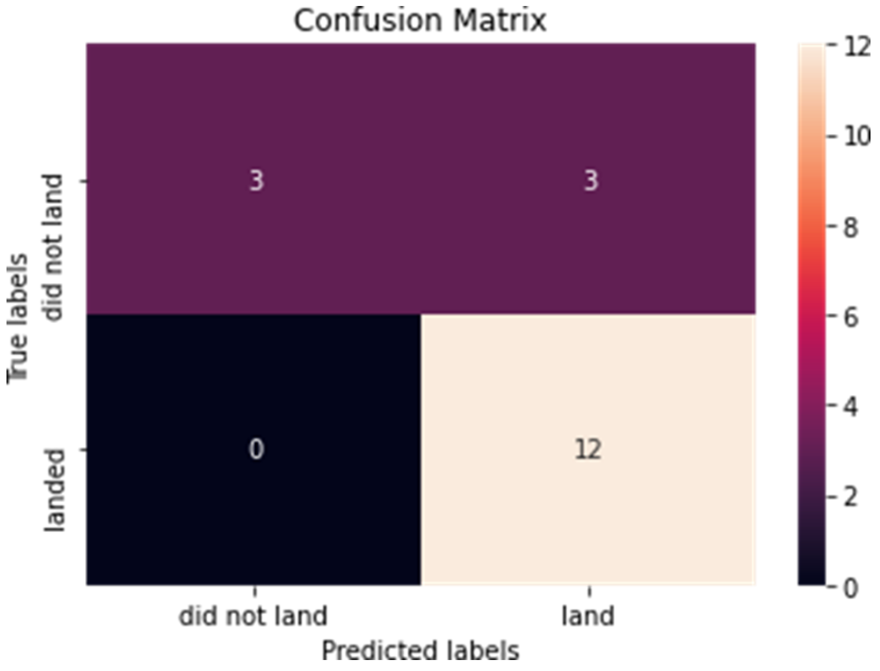
Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Predictive analysis (classification) results

- Classification Accuracy



All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.
This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
We likely need more data to determine the best model.



- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

CONCLUSION

- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy