Winning Space Race with Data Science

Tushar Suvarna

31/01/2022

Outline

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

Executive Summary

- The goal of this project is to determine if Space Y company can compete with SpaceX
- The SpaceX Falcone 9 launch data was collected from the SpaceX REST API and wiki pages
- Exploratory Data Analysis (EDA) were performed to determine labels used for training supervised models.
- A dashboard application was built for users to perform interactive visual analytics on SpaceX launch data in real-time.
- Machine learning models were built to determine if the Falcon 9 launch's first stage is successful and whether the SpaceX will reuse the first stage.
- All models predicted successful landings with an accuracy of about 83%. More datasets are needed for better prediction and higher accuracy.

Introduction

- The Commercial Space Era is here!
- SpaceX has best launch pricing due to ability of recovering and reusing parts of Falcon 9 rocket (first stage)
  - $62 million vs $165 million
- Our aim is to predict if Space Y company, a new rocket company, can compete with SpaceX
  - Determine the price of each launch
  - Predict successful recovery of first stage

Section 1

Methodology

Methodology

- Data collection methodology:
  - ➢ Request data from SpaceX REST API and Data Scraping from Wikipedia pages
- Perform data wrangling
  - ➢ Clean and classify data to represent outcomes as successful and unsuccessful landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - ➢ Determine Training Labels, Train models, Select hyperparameters using GridSearchCV

Data Collection

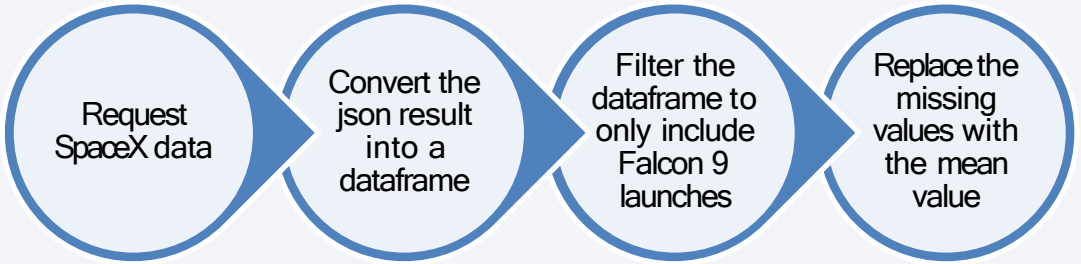
Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

- The SpaceX Falcone 9 launch data was collected from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches":
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data was requested from SpaceX API and basic data wrangling and formatting was performed.

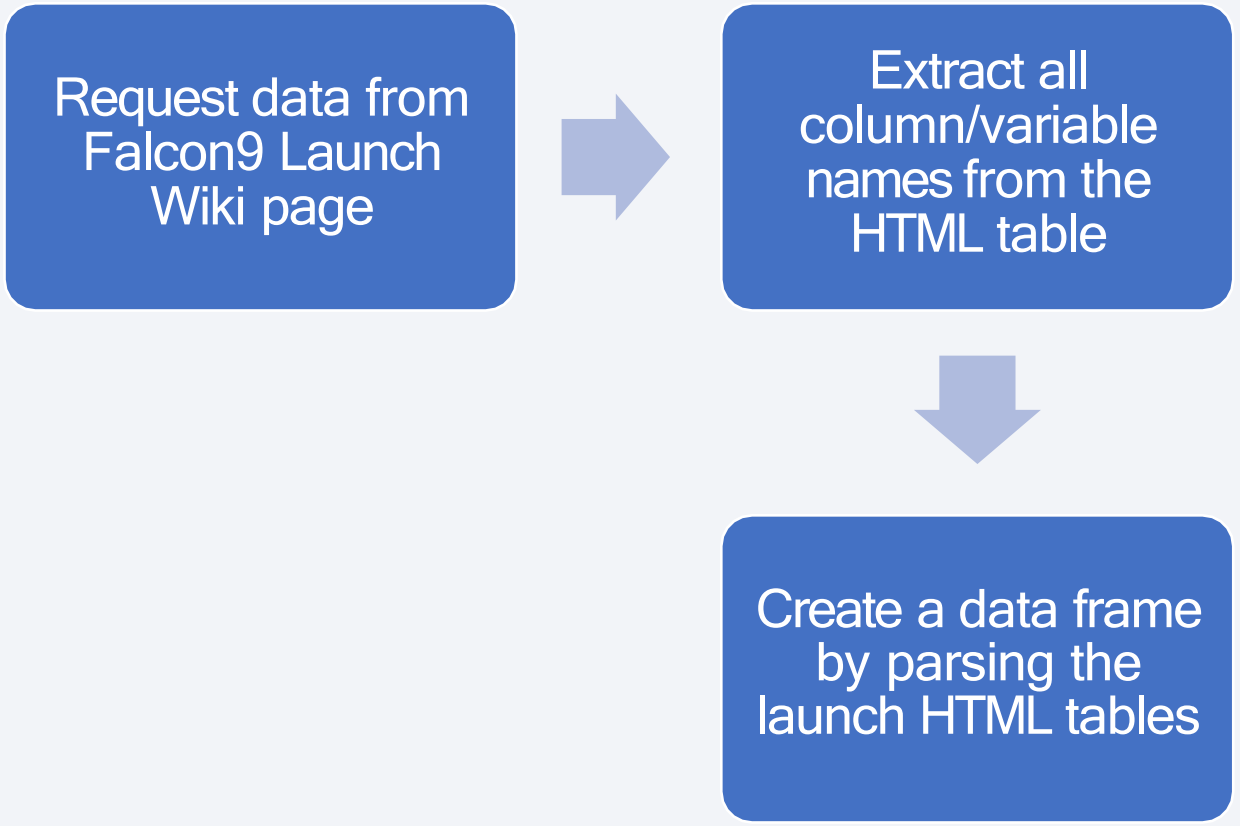
The next two slides contain the flowcharts showing the process of data collection.

Data Collection – SpaceX API

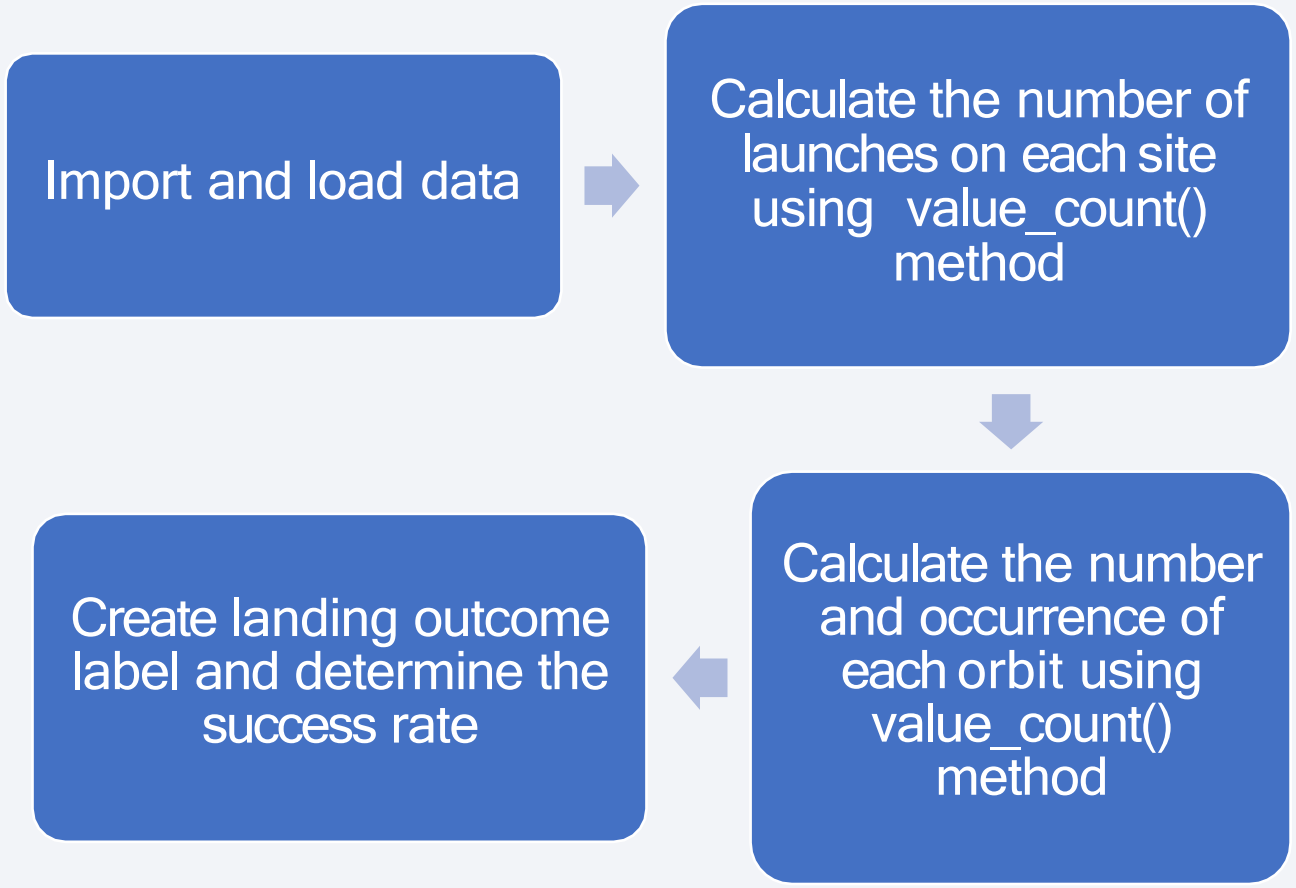
Github:
https://github.com/tusharS29/Data_science_capstone



Data Collection - Scraping



Data Wrangling



EDA with Data Visualization

•Exploratory Data Analysis was performed to explore and analyze the relationship between timeline, flight number, launch site, payload mass, orbit type and the success rate and class.

- The following plots were created: flight number vs. launch site, payload mass vs. launch site, success rate vs orbit type, flight number vs. orbit type, payload vs orbit type, and success yearly trend
- The results were used in determining the training data

EDA with SQL

- Created queries to explore launch site names, payload mass for different missions, booster versions, mission and landing outcomes

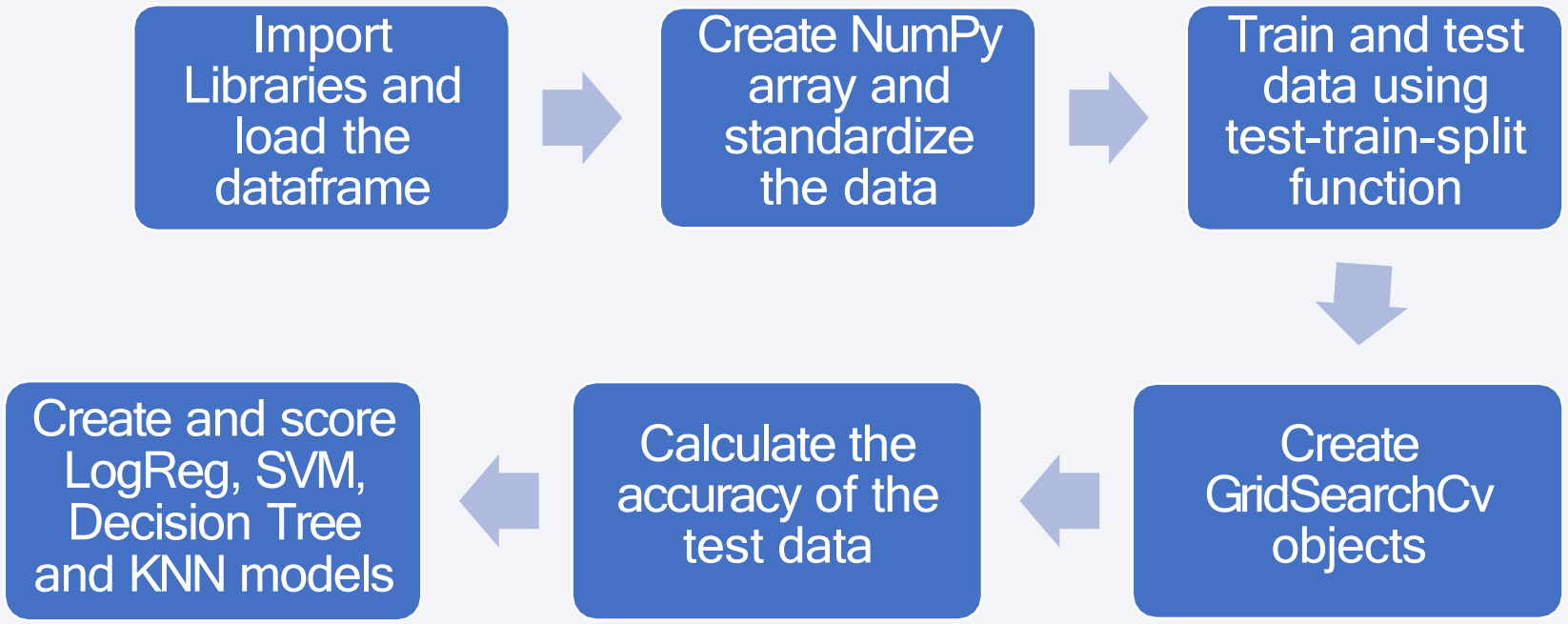
Build an Interactive Map with Folium

- Marked all launch sites on a global map to visualize the location of launch site and in next step their relative success rate.
- Marked successful/failed launches for each sites on the map to determine and which site has a higher success rate.
- Calculated the distance between a launch site to its proximities (such as: railway, highway, city and coastline).

Build a Dashboard with Plotly Dash

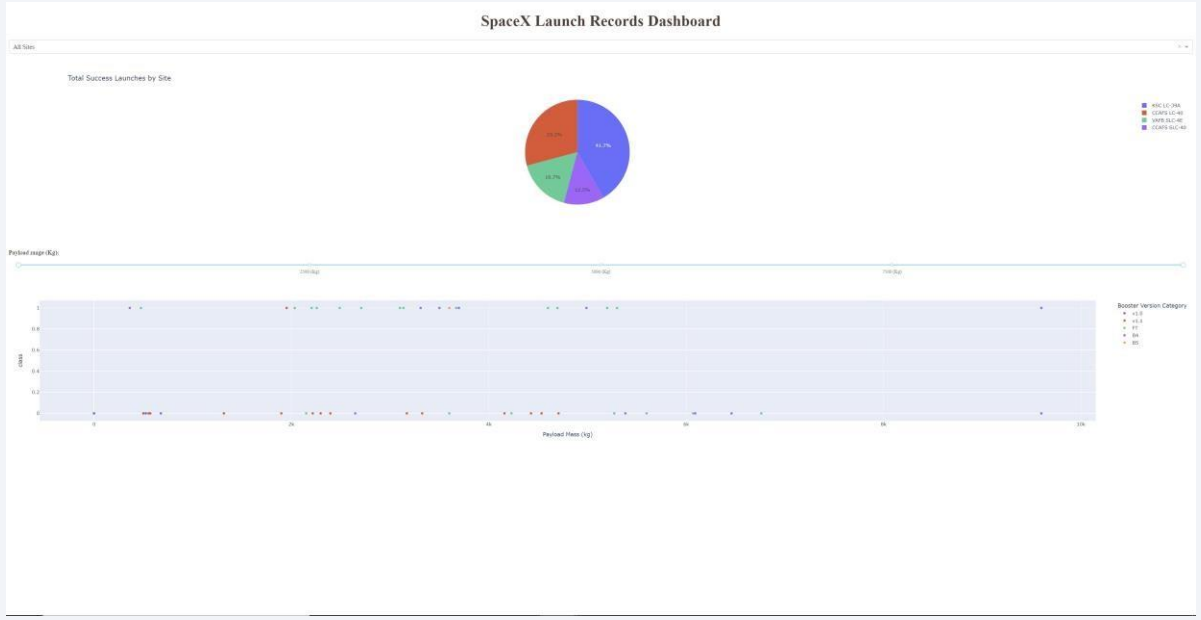
- A Plotly Dash application containing a dropdown list (to interact with a pie chart showing the Total Success Launches by site), and a slider (to interact with a scatter plot chart showing the correlation between payload and success for all sits) was built for users to perform interactive visual analytics on SpaceX launch data in real-time.

Predictive Analysis (Classification)



Results

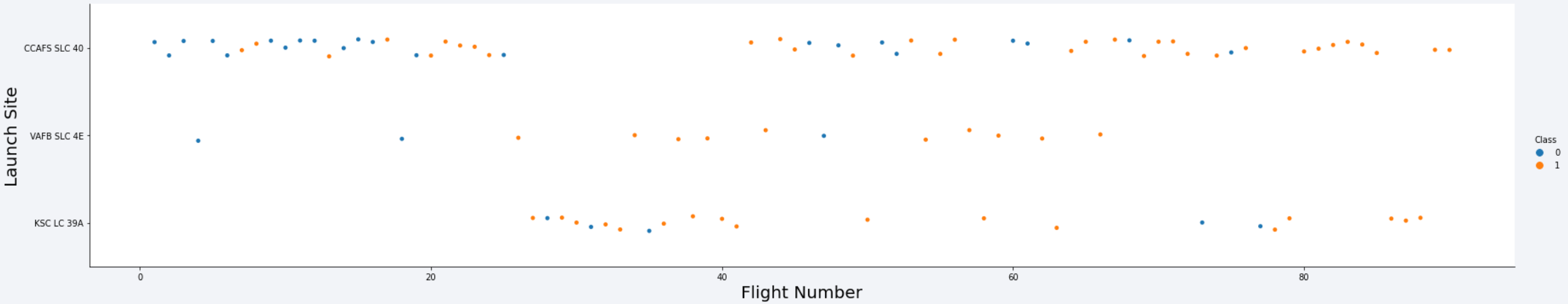
- The relationship between the flight number, launch site, orbit type, payload mass, and the mission's success rate were studied by exploratory data analysis
- Machine learning models predicted successful landings with an accuracy of about 83%.
- The image shows the dashboard created by using Plotly Dash:



Section 2

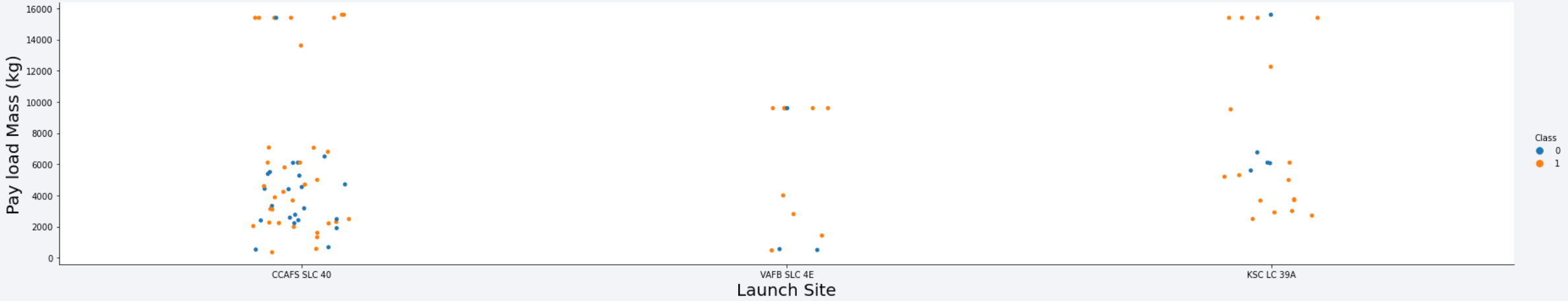
Insights drawn from EDA

Flight Number vs. Launch Site



Successful (orange) and unsuccessful (green) launches are shown for each launch site. The number of successful flights has increased over the time period. Having more flights from CCAFS SLC-40 suggests that Cape Canaveral is the main location chosen for launching the SpaceX rockets.

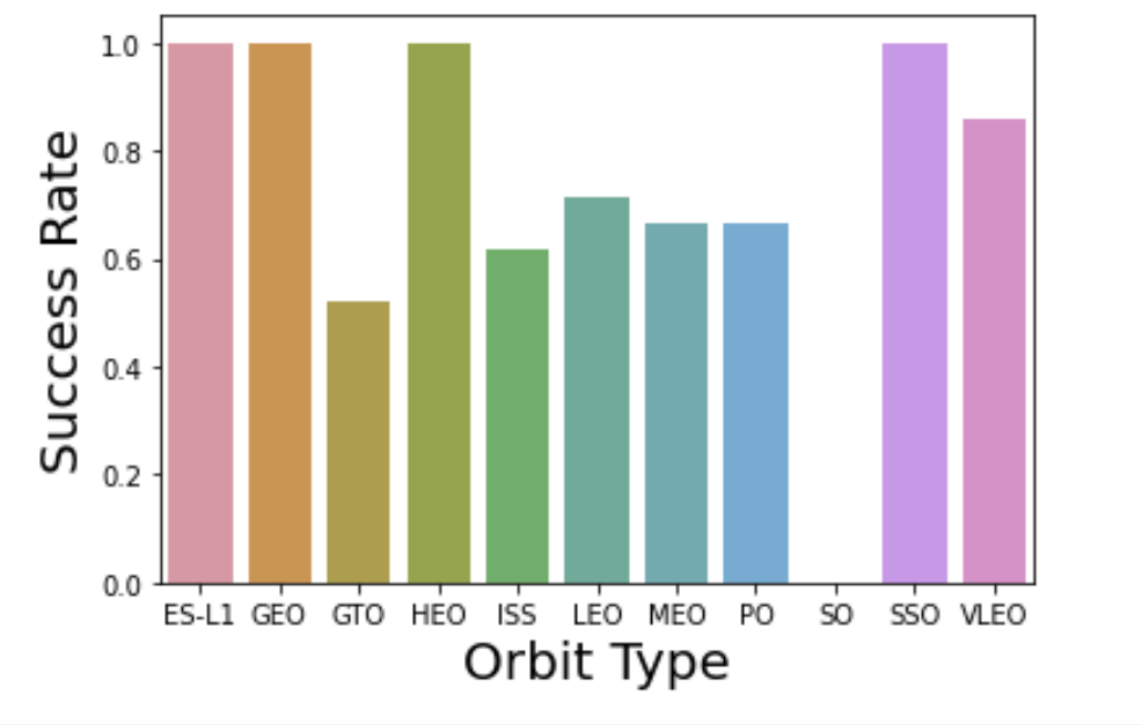
Payload vs. Launch Site



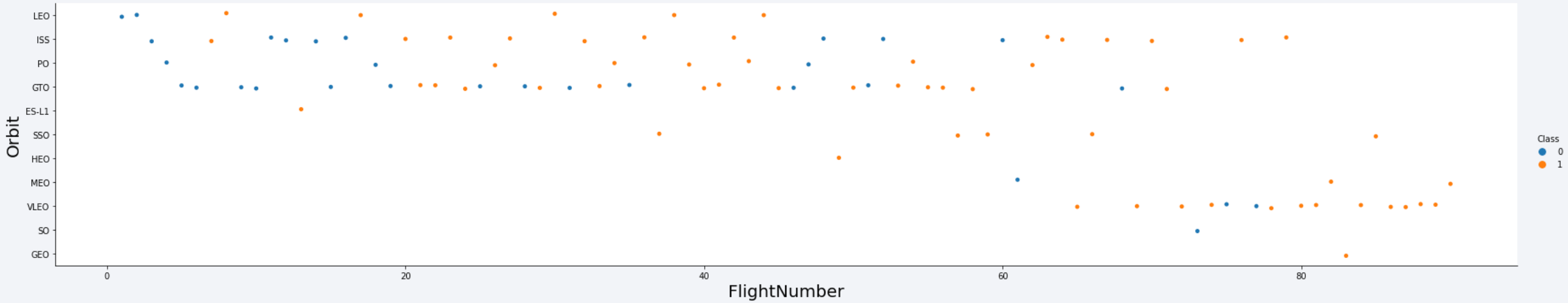
Successful (orange) and unsuccessful (green) launches are displayed for each launch site. Most Launched rockets have a payload mass less than 7000 kg.

Success Rate vs. Orbit Type

ES-L1, GEO, HEO and SSO have 100% success rate. SO with 0% success rate and GTO with about 50% success rate have the lowest success rates among orbit types.

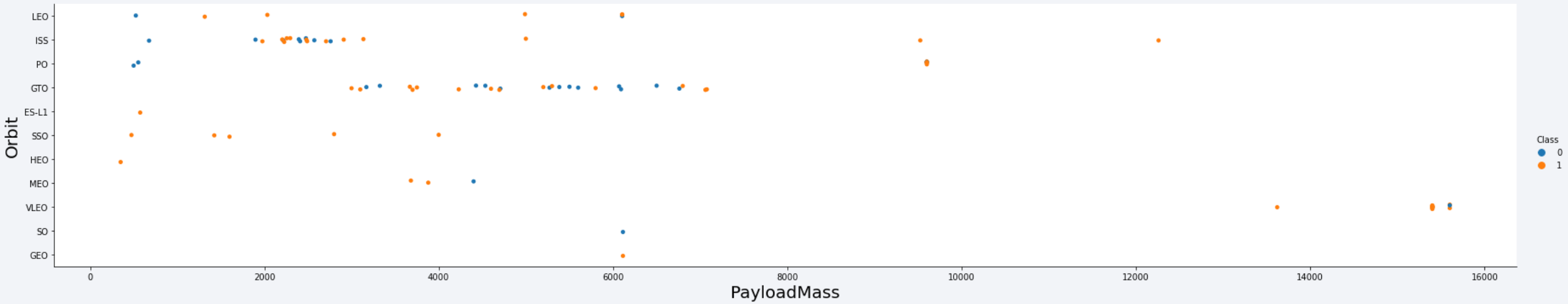


Flight Number vs. Orbit Type



Successful launches are displayed in orange and failed launches are displayed in green. In the LEO orbit the Success appears related to the number of flights; There seems to be no relationship between flight number when in GTO orbit. SpaceX performs better in low orbits.

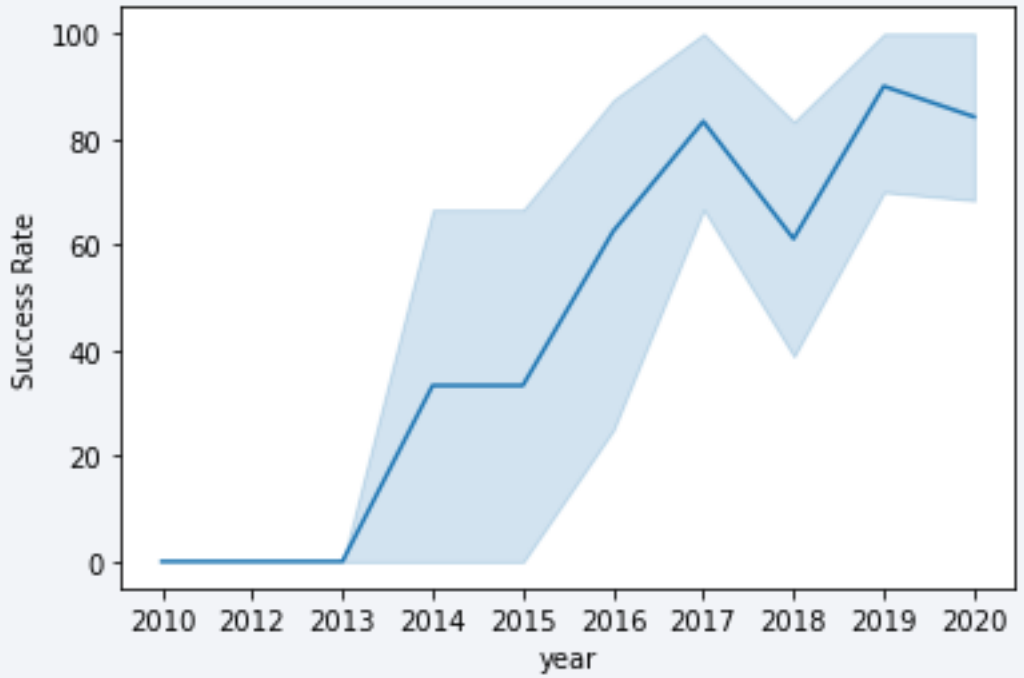
Payload vs. Orbit Type



Successful launches are displayed in orange and failed launches are displayed in green. Most lunches have lower payload mass (less than 7000kg) The SO and VLEO orbits have higher range of payload mass.

Launch Success Yearly Trend

Success rate has increased between 2013 (0%) and 2020 (~80%) with a dip in 2018. Light blue is the confidence interval range.



All Launch Site Names

```
In [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
```

Out[10]:

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

- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40, previously known as CCAFS LC-40) for United States Space Force national security launches.
- Vandenberg Space Force Base Space Launch Complex 4E (VAFB SLC-4E) for polar launches
- Kennedy Space Center Launch Complex 39A (KSC LC-39A) for NASA launches

Launch Site Names Begin with 'CCA'

```
In [11]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
```

Out[11]:

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	Success
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

The first 5 launch site names beginning with ‘CCA’

Total Payload Mass

```
In [8]: %sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
```

Out[8]:

SUM
45596

The total payload mass carried by boosters from NASA is 45596 kg.

Average Payload Mass by F9 v1.1

```
In [9]: %sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version = 'F9 v1.1'
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
Out[9]:
average
2928
```

Average payload mass carried by booster version F9 v1 is 2928 kg.

First Successful Ground Landing Date

```
In [13]: %sql select min(date) as Date from SPACEXDATASET where landing__outcome like 'Success (ground pad)%'
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
ain.cloud:31198/BLUDB
Done.
```

Out[13]:

DATE
2015-12-22

SpaceX first successful ground landing occured on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [37]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ BETWEEN 4000 AND 6000 and landi
         ng__outcome like 'Success (drone ship)'

          * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
         ain.cloud:31198/BLUDB
         Done.
```

Out[37]:

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

List contains 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

```
In [17]: %sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome

 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
```

Out[17]:

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

SpaceX has about 99% success rate in missions. The total number of successful mission outcome is 100 and only one mission was failed.

Boosters Carried Maximum Payload

In [34]:
```
maxm = %sql select max(payload_mass__kg_) from SPACEXDATASET
maxv = maxm[0][0]
%sql select booster_version from SPACEXDATASET where payload_mass__kg_=(select max(payload_mass__kg_)
from SPACEXDATASET)
```

```
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
ain.cloud:31198/BLUDB
Done.
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
ain.cloud:31198/BLUDB
Done.
```

Out[34]:

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

List shows the Falcon 9 booster versions which have carried the highest payload mass (15600 kg).

2015 Launch Records

```
In [33]: %sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site from SPACEXDATASE
         T where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
ain.cloud:31198/BLUDB
Done.
```

Out[33]:

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

List shows the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

```
In [32]: %sql select landing__outcome, count(*) as count from SPACEXDATASET where ((Date >= '2010-06-04') and
         (Date <= '2017-03-20')) GROUP BY landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://mwh44990:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
ain.cloud:31198/BLUDB
Done.
```

Out[32]:

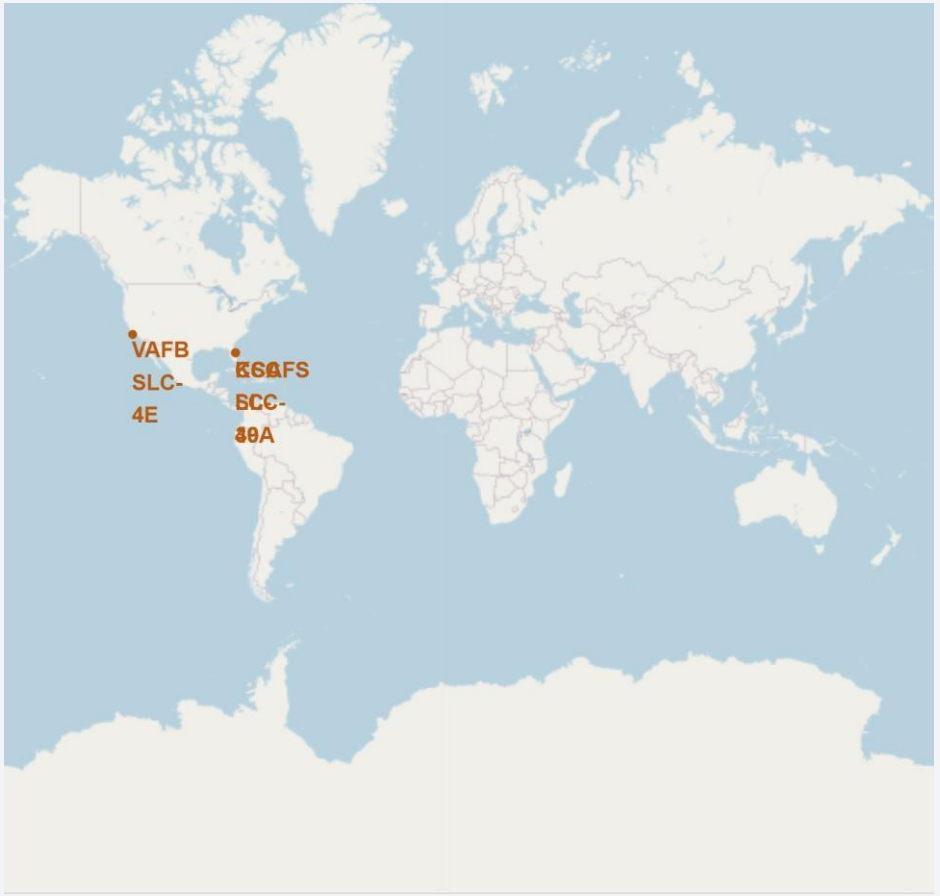
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

The list shows the number 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.

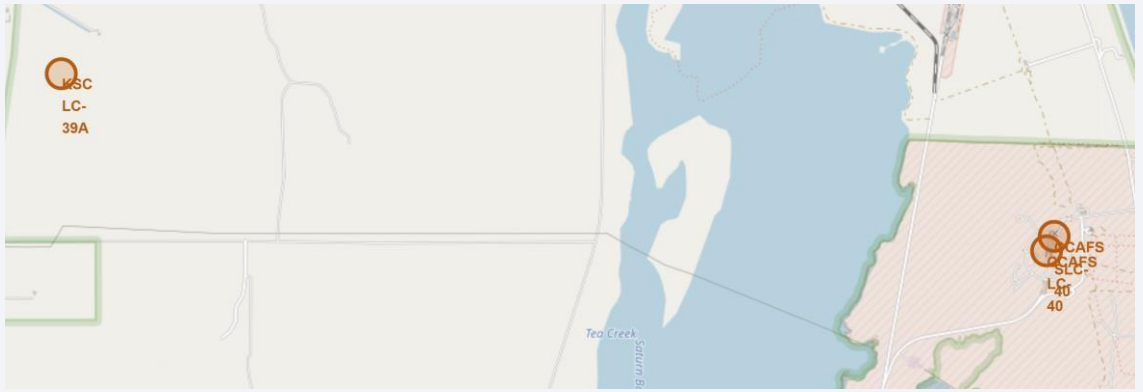
Section 4

Launch Sites Proximities Analysis

Locations of Launch Sites



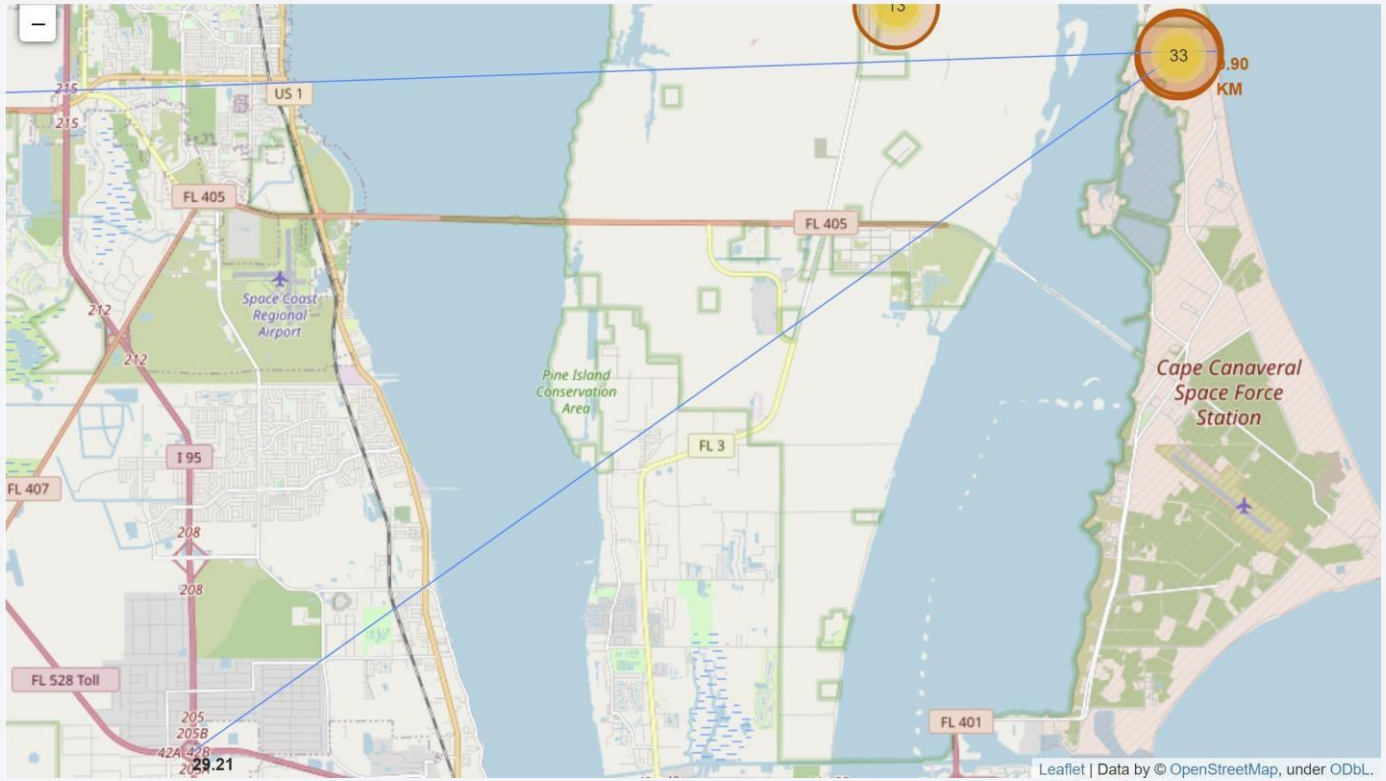
The figure on the left shows all launch sites on a global map. All sites are located in the USA. VAFB SLC-4E is in California. KSC LC-39A , CCAFS LC-40 and CCAFS SLC-40 are within close distance from each other in Florida. The figure on the bottom shows the three launch sites in Florida.



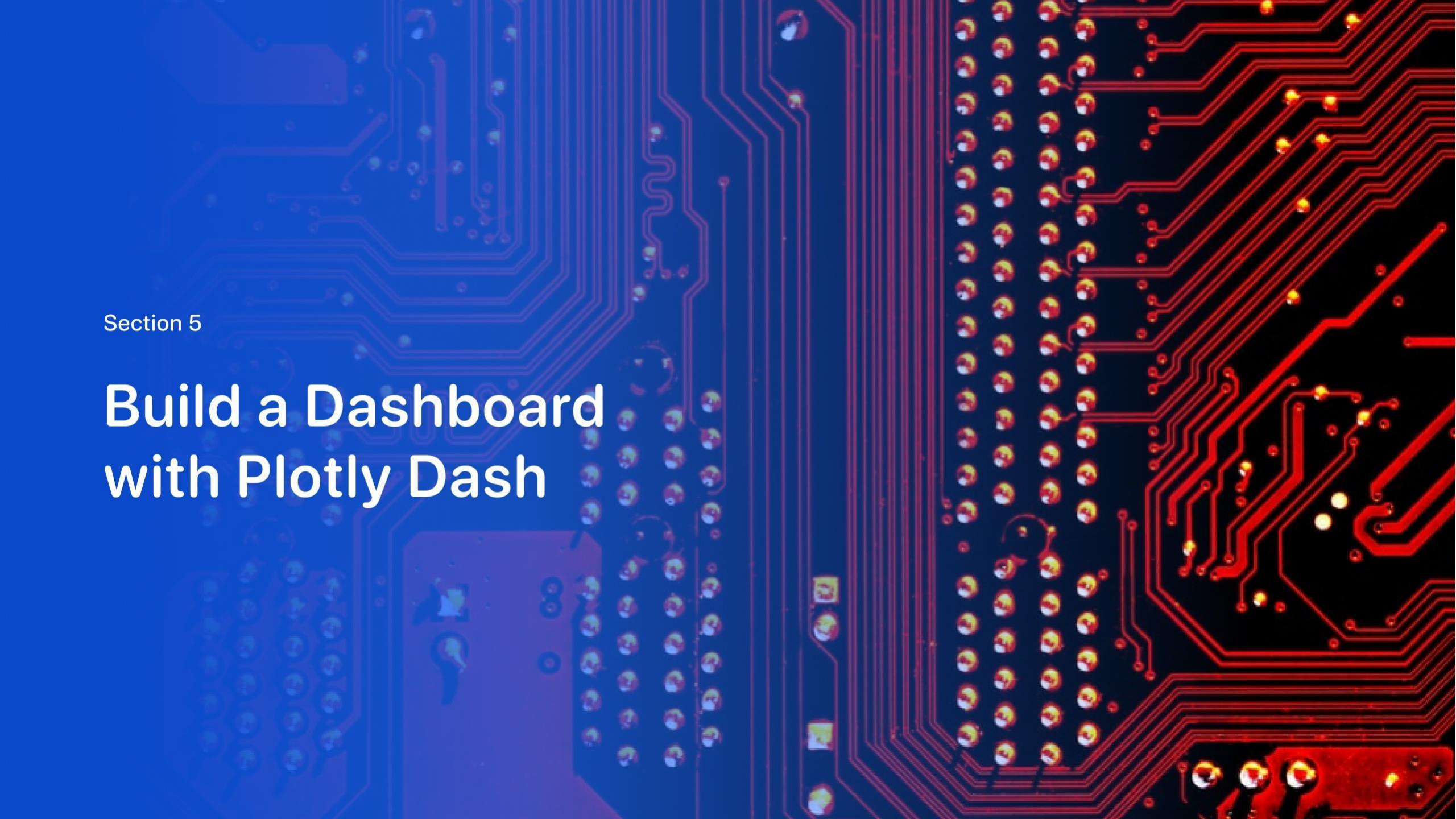
Launch Outcomes

On the left the number of landings for the launch sites is shown. For each cluster, the successful landings are displayed in green and failed landing are displayed in red. Right image shows the successful and failed landings for VAFB SLC-4E launch sites.

Launch Sites Proximities

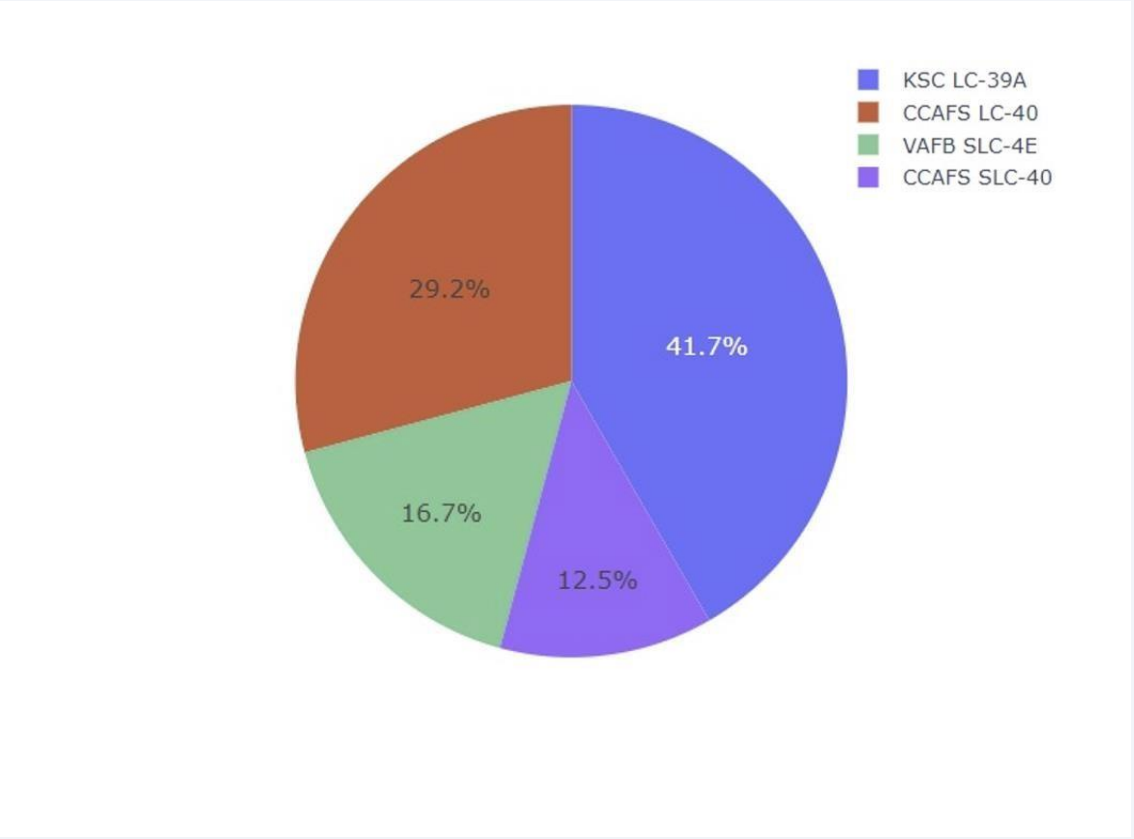


Launch sites are relatively close to railways and highways for transportation convenience and are located far from cities (near the coast) to reduce the risks of launch over the populated areas. The image shows the LCCAFS LC-40 proximities.

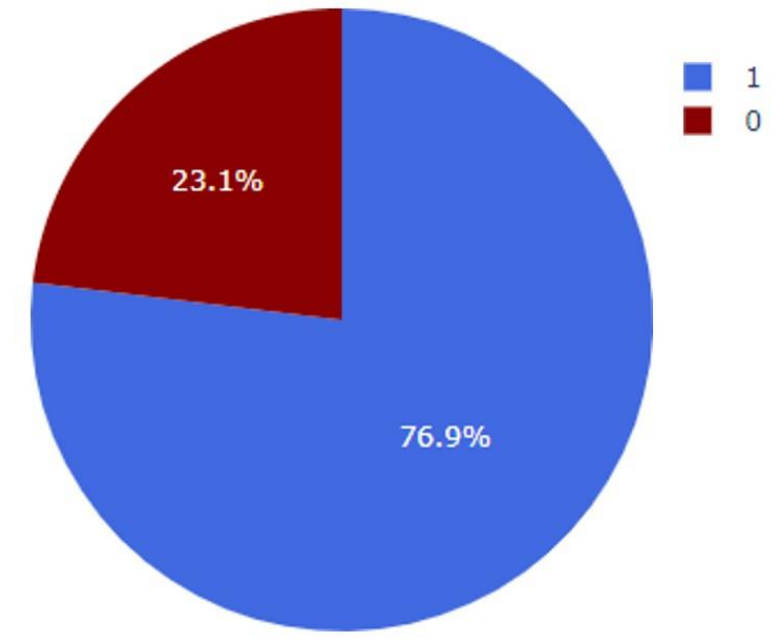

Section 5
Build a Dashboard with Plotly Dash

Total Successful Launches by Site

The graph shows the successful launch rates for four different launch sites. KSC LC-39A has the highest success rate followed by CCAFS LC-40 and VAFB SLC-4E. The smallest rate belongs to CCAFS SLC-40.

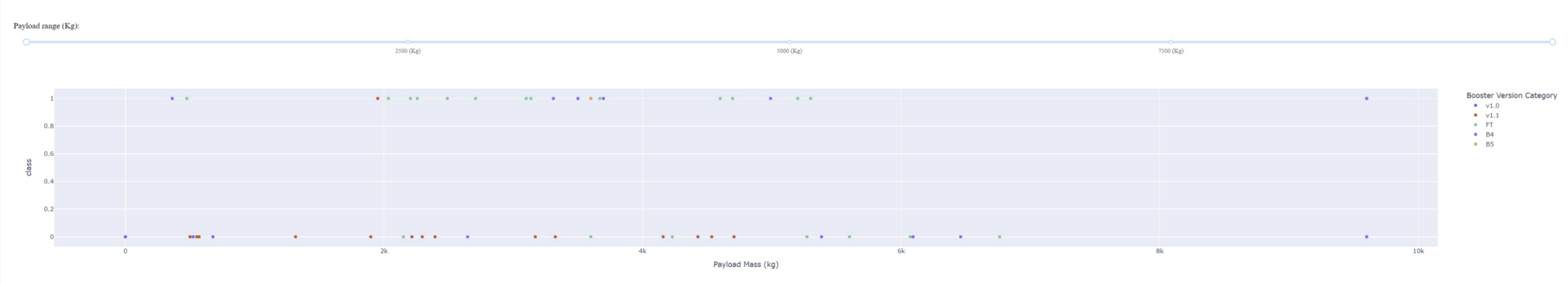


Launch sites with highest launch success ratio



Launch site KSC LC-39A at Kennedy Space Center has the highest success ratio with 76.9% success (color blue) and 23.1% failure (color red).

Booster Version Play Mass



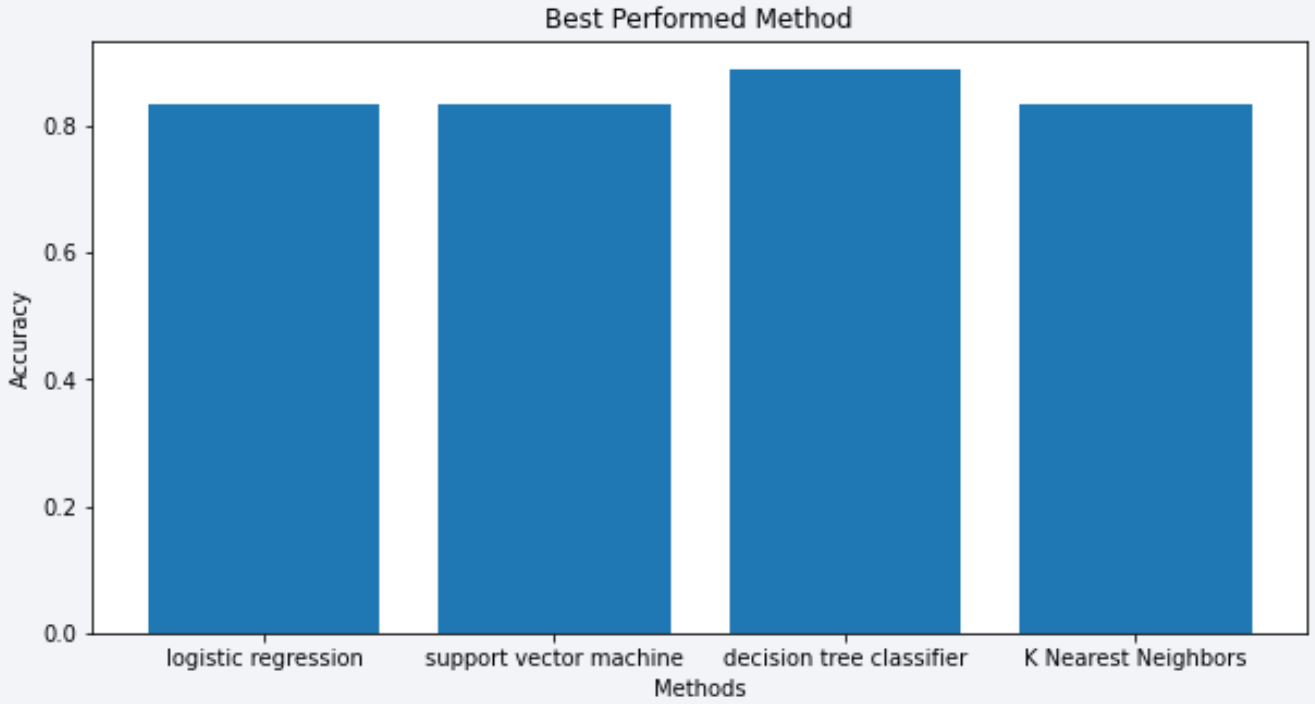
Payload mass vs Success Class vs Booster Version Category

The dashboard has a Payload range selector from 0 to 10000 kg. Class 0 shows the failure and class 1 shows the success. There is also a booster version category with different color points that indicates which booster launch was successful.

Section 6

Predictive Analysis (Classification)

Classification Accuracy



For 18 samples, almost all models have accuracy about 83.33%. Decision tree classifier model has slightly higher accuracy (about 88.88%).

Confusion Matrix

The confusion matrix was the same for all models. The models predicted 3 unsuccessful landings, when the true label was unsuccessful landing, 3 successful landings when the true label was unsuccessful landings (false positives) and 12 successful landings when the true label was successful landing.

Conclusions

- Our goal is to determine if Space Y company can compete with SpaceX by creating a machine learning model that can predict the successful stage 1 landing
- Data was collected from SpaceX API and SpaceX Wiki pages
- Data was analyzed and then visualized using plots, Folium and Plotly Dash
- Machine learning models were created with an accuracy of 83%
- SpaceY can use the created machine learning models to predict whether the SpaceX stage 1 landings are successful and use it to decide about biding against SpaceX

Appendix

For more information about Python codes, SQL queries, charts, Notebook outputs, and data sets please refer to the GitHub address below:

https://github.com/tusharS29/Data_science_capstone

Thank you!