IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

William Anderson

June 07, 2023

Outline

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

Executive Summary

- In this project, we used several methods for gathering, cleaning, preparing, and analyzing the spacex data. We gather data form api.spacexdata.com/v4 and we scrapped data from the SpaceX wiki page. Then, we parsed the data from various columns to look at launches and the Falcon 9 rocket. After normalizing the data, we visualized the data using seaborn. we compared the fight number, payload mass, orbit, and launch site. After finding our insights, we used one-hot-encoding on 'Orbit','LaunchSite', 'LandingPad', and 'Serial'. this gives us 80 features to work with during the predictive process. We then used machine learning algorithms to predict success missions. we used a number of classification algorithms due to the categorical nature of the data. In addition, we looked a geomapping to compare launch site success.

- The results indicate a trend toward success as launches progressed. The Orbit of a particular mission does seem to play a role in mission success: GTO, ISS, LEO, MEO, PO, SO have lower success rates. Additionally, we see a number of success with regard to payload, between 0 and 4000. With this data, we see that a decision tree machine learning model gives us the best predictors of success.

Introduction

Space flight started in in the 1920 and gain momentum in the 1960. After the 1960, space flight cold down, until SpaceX in 2020. Other companies have worked on space flight. Blue Origin, Virgin Galatic, and ULA(United Launch Alliance) launch space flights. However SpaceX has more publicly available data. As of June 2023, Space X has a total of 229 total launches. With seventy percent of those being Relaunches, SpaceX has kept the cost down. According ot Jason Davis, a news reporter, "Using its 230-foot-tall Falcon 9, SpaceX charges $62 million to send into orbit commercial satellites weighing up to 50,000 pounds. The closest American competitor is the United Launch Alliance Atlas V, which starts at $73 million for a 41,000-pound payload."

As a new startup space flight company, We are interested in knowing if the first stage landing of a launch will be successful.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

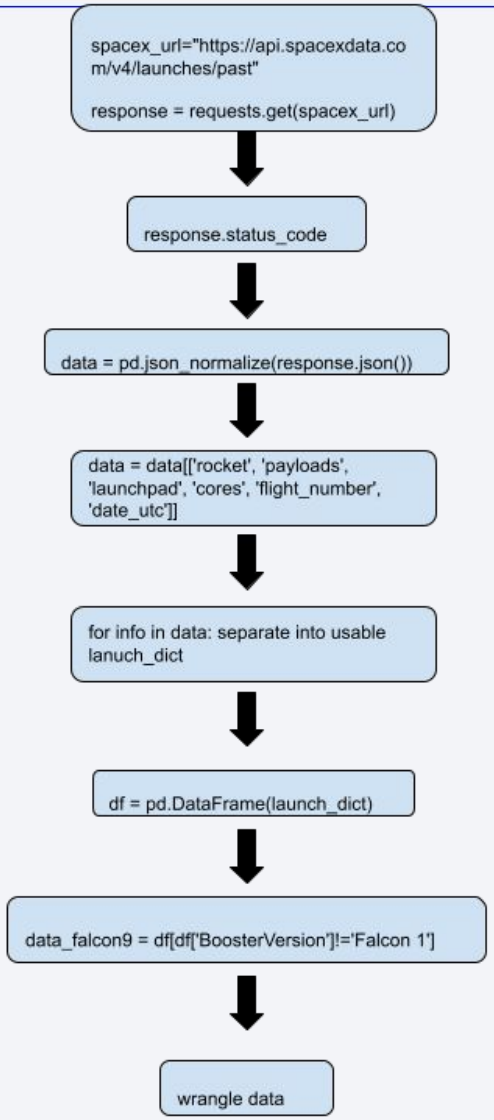
Data Collection

Data was collected using two methods a direct call to a compiled data from r-spacex and Web scraping of the wiki of SpaceX data. The first used the requests module from python to call "https://api.spacexdata.com/v4/launches/past". This returns a json file of a multitude of data. Then pandas used to normalize the json file into a pandas dataframe. once in a data frame, key features were pulled out into a dictionary: BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. The second method used the JS module provide by python to call "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922". this provide a web page that was parsed using beautiful soup. the tables from the url were parsed for the key terms 'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'. after collecting the data into the appropriate features, data wrangling could begin.

Data Collection – r-SpaceX API

- calls for data sent to: "https://api.spacexdata.com/v4/launches/past"

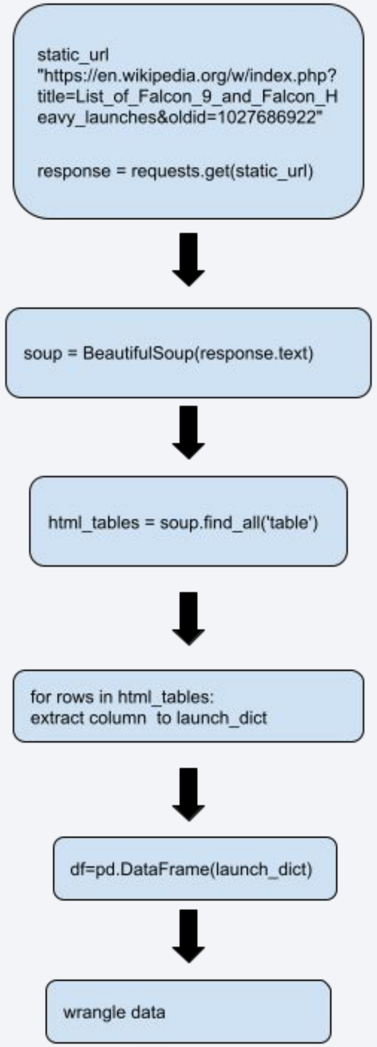
- link to jupyter notebook: https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/14862a441df8046720bb69174134787b5ba73200/step1-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping process using key phrases and flowcharts

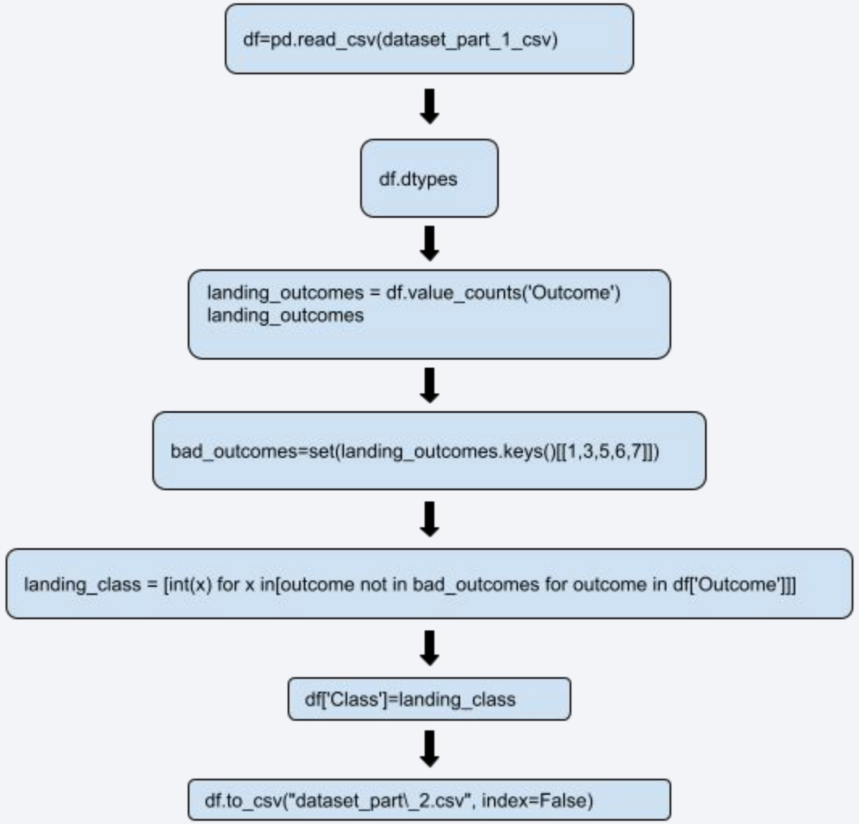
- Link to Jupyter Notebook:https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/14862a441df8046720bb69174134787b5ba73200/step2-webscraping.ipynb



Data Wrangling

- The data was processed as a pandas dataframe. Previously the data was limited to the falcon 9 rocket. This limits the number of variables that could influence the data. Additionally, the missing data for the payload mass was replaced with the mean and NaN for the landing pad was left meaning no landing pad. For purposes of this project the data wrangling was focused on the outcome column. turning it into a class of 0 as a failure or 1 as a success.

- link to Jupyter Notebook: https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/14862a441df8046720bb69174134787b5ba73200/step4-spacex-data_wrangling.ipynb



EDA with Data Visualization

- Strip plot
  - or scatter plot is the simplest way to visualize data and look for trends. This is the quickest way to compare various features. Here the payload mass was compared to flight number. Likewise, launch site was compared to flight number.
- Point plot
  - this compares classes with trends from one data point to the next allowing for visualization of the rate of change with error bars. Point Plots were used to look at launch site/ payload and launch site/flight number.
- Bar plot
  - Bar plots are good for a quick summary of totals. Orbit and class were compared.
- Line plot
  - the line plot is good for showing trends. Average success by year was shown in this case.

- https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/4b5ea670eb30fcaf8d65f34a4c029595b1b747a2/step5-eda-data-viz.ipynb

EDA with SQL

- Using SQL to explore properties of the data
    - looking a unique launch sites
    - total mass rockets carried by company
    - the average payload the booster carried
    - finding the first successful landing
    - looking at which booster had successful landing on drone ships
    - summarizing mission outcomes
    - listing the boosters that have carried the max load
    - listing month, outcome, booster, launch sites for failure on drone ship
    - ranked successful landing outcomes

- https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/dcc8e9dc41f565e857a941d32d3af14ff95092d2/step3-eda-spacex-data-sqllite.ipynb

Build an Interactive Map with Folium

- Using folium maps, geographical visualizations can be made
  - using circle markers, launch sites were added to the map.
  - then using cluster markers success and failures were added to each site
  - additional markers were added to Cape canaveral site show distances to nearby locations

- These markers were added as tool to help understand the data in a real world context. The distance markers were to help visualize the surrounding and personalize the data. The maps also allow for interactions with the data outside of data tables.

- jupyter notebook https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/371d350fed783bef308a70228bb6aa823c2e49b1/step6-launch_site_location.ipynb

- html formate for working with the folium maps/ you will have to download the raw file and open in a browser. https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/858a6f1ff39058c16d7428efd8e6e21df5811dfe/step6-launch_site_location.html

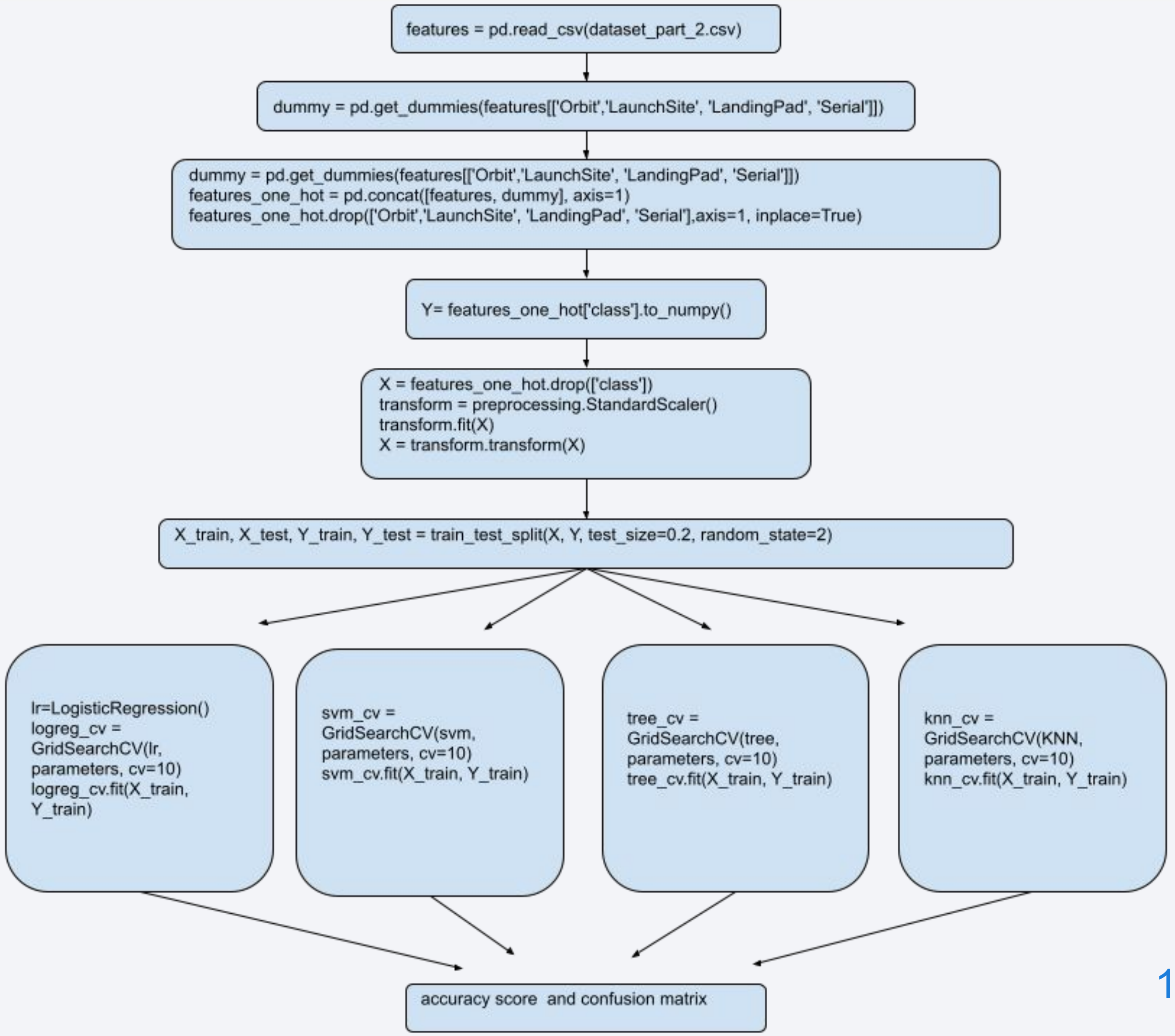
Build a Dashboard with Plotly Dash

- For the dashboard two interactive plots were chosen
  - Pie chart
  - Scatter plot
- These two plots graphs were chose to interact with the key features.
  - the launch site
  - the class
  - the payload mass
- https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/371d350fed783bef308a70228bb6aa823c2e49b1/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was loaded into a pandas dataframe. having all the data converted into numerical data using one hot encoding. the data was then separated from the target values, and turned into numpy arrays. the data was split into training and testing sets. Once that was accomplished, the data was plugged into several classification models using a 10 fold cross-validation set up. Then, Accuracy scores were compared.

- https://github.com/W-Anderson/IBM-DS-Professional-Certificate/blob/4b5ea670eb30fcaf8d65f34a4c029595b1b747a2/step7-SpaceX_Machine_Learning_Prediction.ipynb



Results

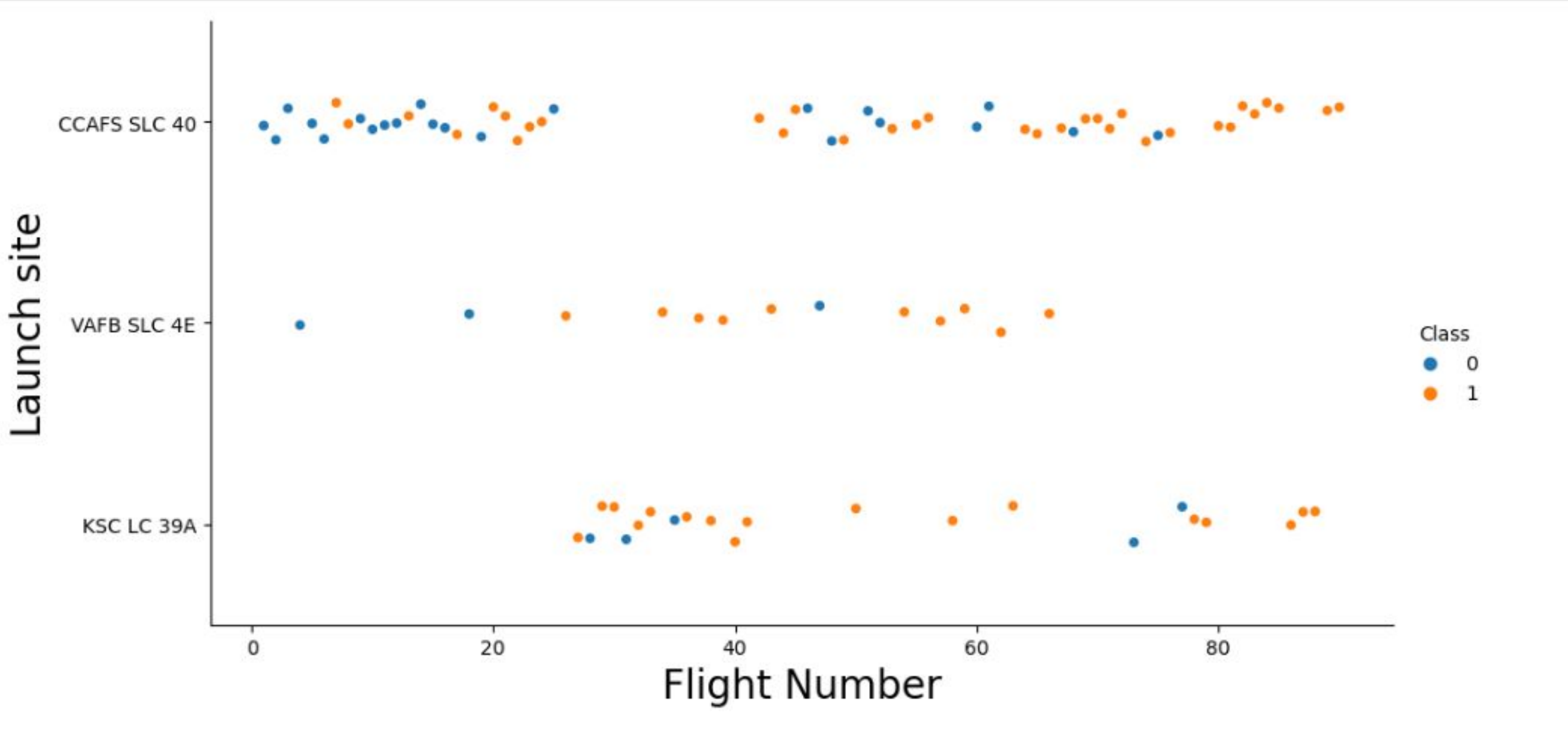
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

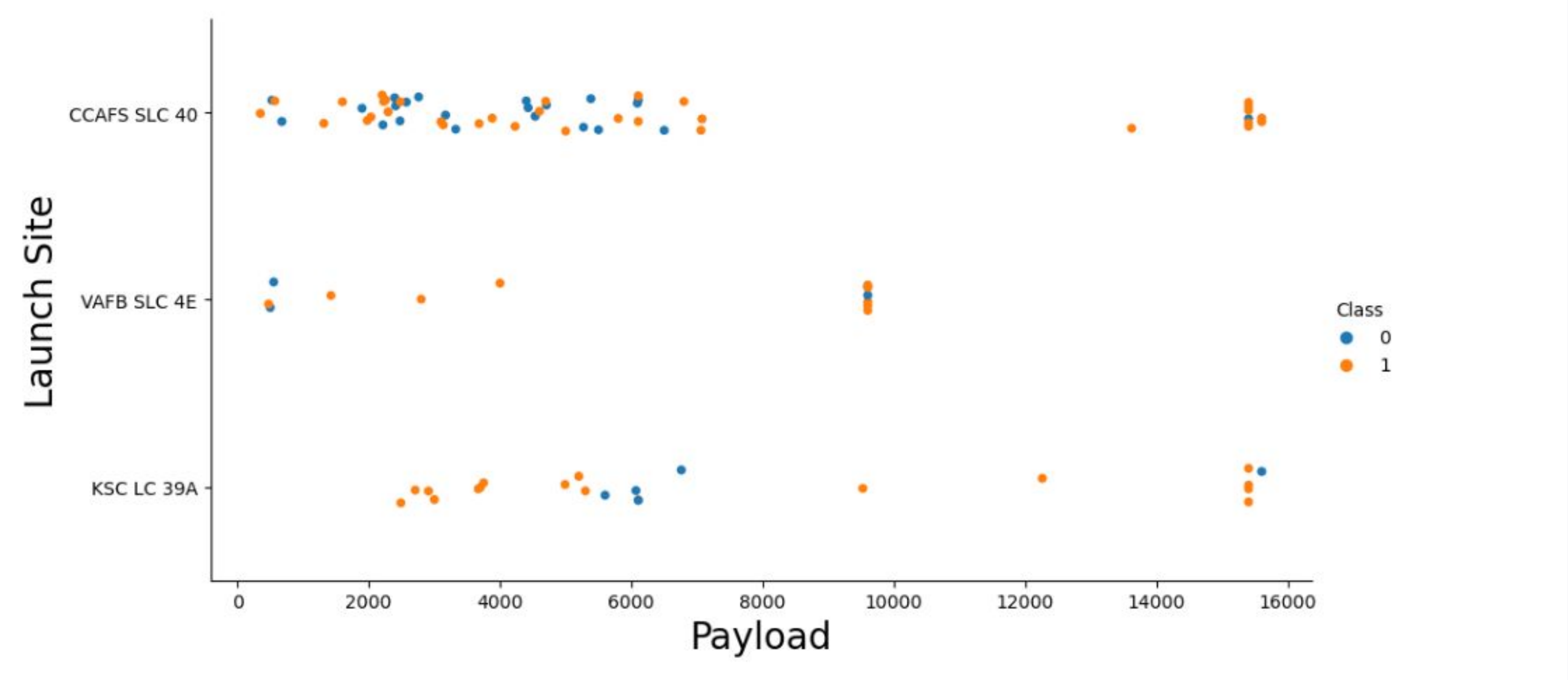
Flight Number vs. Launch Site

Here we see a scatter plot of the launch sites versus the flight number. While VAFB has a few launches, they take place early and in the middle. Additionally, they are more successful than not. Likewise, KSC has a few more than VAFB with a greater spread of success. and happening from mid to later flights. CCAFS looks to have the most flight spanning the entire flight history with the most sordid number of success.



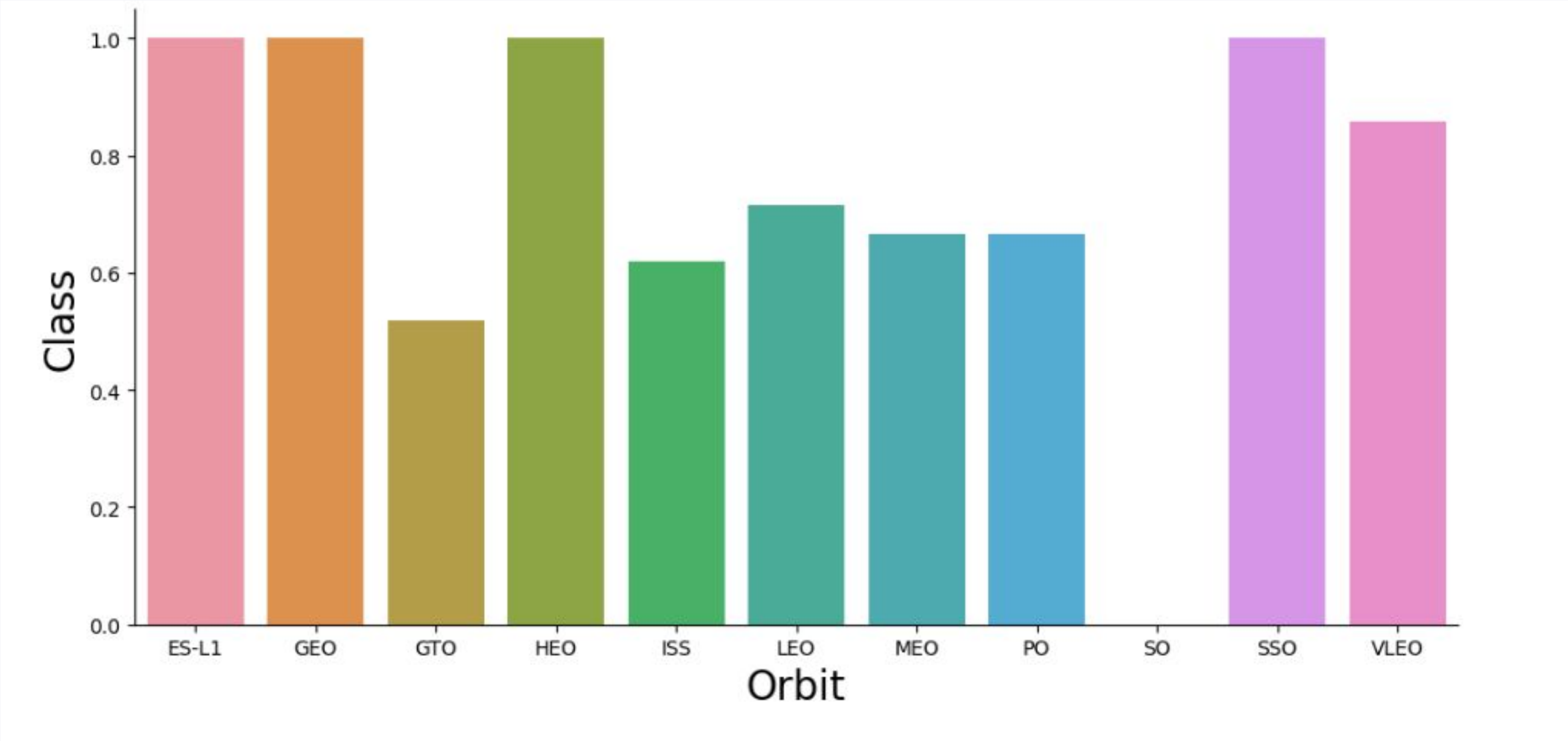
Payload vs. Launch Site

This scatter plot shows the launch sites versus the payload mass in KG. As before we see CCAFS has the most flights. However here we see they have most of there flights with the payload below 8000, and a few over 14000. VAFB has a split between low mass and medium mass. KSC has payload spread between low and high.



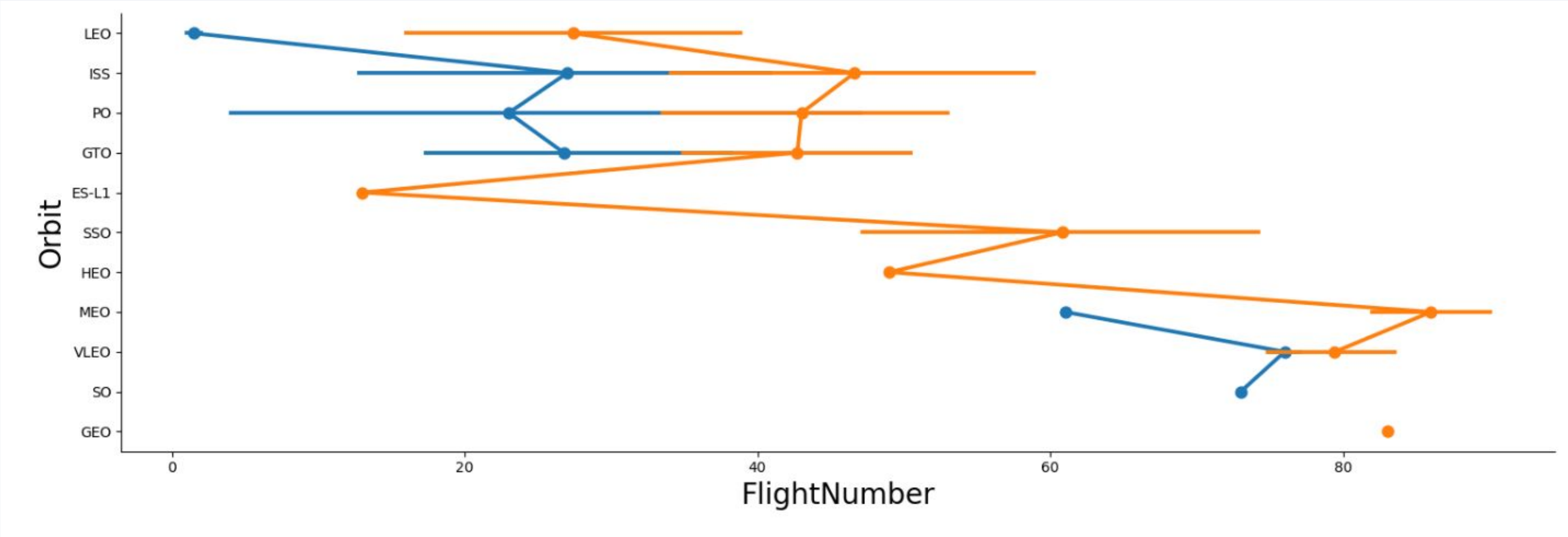
Success Rate vs. Orbit Type

In this bar chart, we see the orbit versus class. this shows us the number of average success of each orbit flight. it is important to note the total number of flights for each orbit. GTO: 27, ISS :21, VLEO: 14, PO: 9 , LEO: 7, SSO: 5, MEO: 3, ES-L1: 1, GEO: 1, HEO: 1, SO: 1

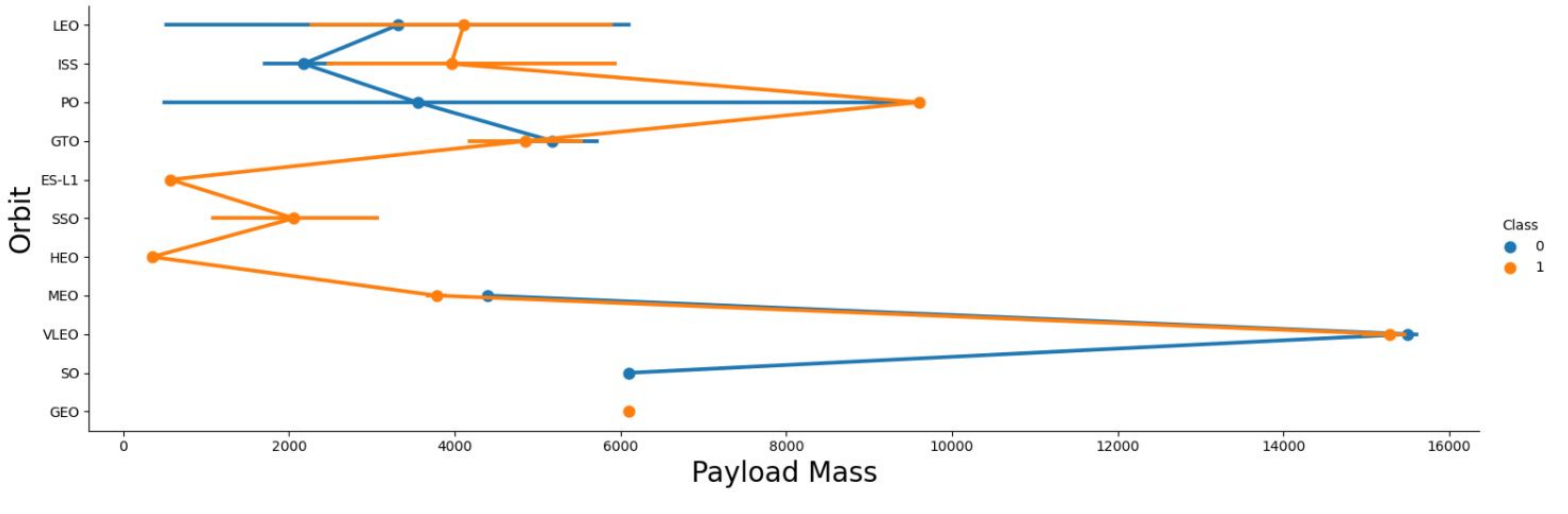


Flight Number vs. Orbit Type

Here we have a scatter point plot showing orbit versus flight number. This shows us an overall trend of increase success for later flights. Where LEO, ISS, PO, GTO have the most counterpoints with success, but still show success with later flights.



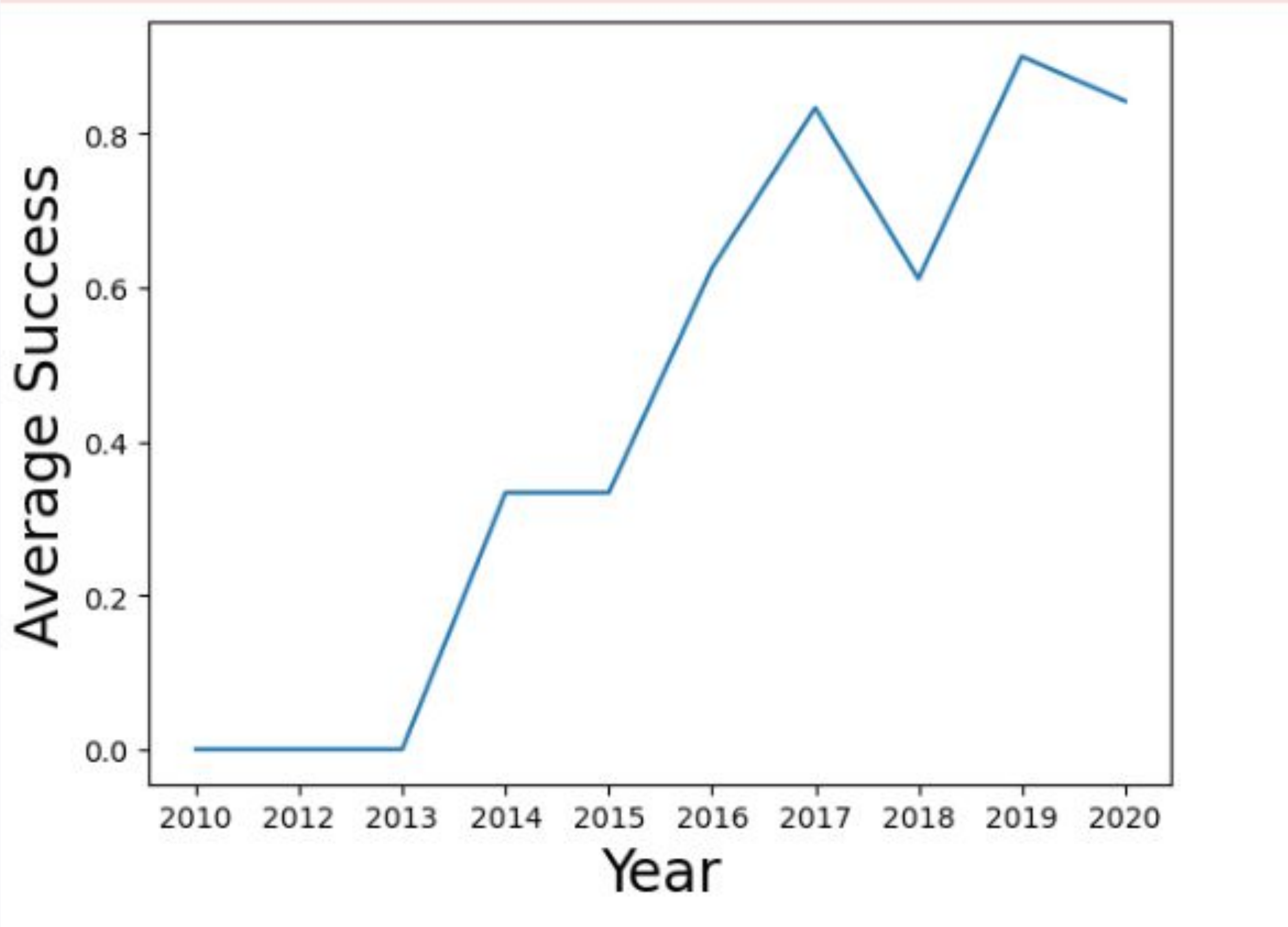
Payload vs. Orbit Type



This scatter point shows Orbit vers payload mass. We see an isolation of failures between 2000 and 4000 kg in the LEO, ISS, PO, and GTO.

Launch Success Yearly Trend

This line plot shows the average success between 2010 to 2020. what we see here is a general trend to improve landings after 2013.



All Launch Site Names

Using SQL we can search the data for all the unique launch sites. We see four uniques launch sites. Launch sites CCAFS LC-40 and CCAFS SLC-40 are in close proximity. However, they are distinct.

```
[17]: # %sql select distinct (Landing_Outcome) from SPACEXTBL
      %sql select distinct (Launch_Site) from SPACEXTBL

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

[17]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

Here are 5 records limited to the search CCA. This important because two launch sites have nearly identical names. If both are need for analysis, this search will include both.

```
[6]: %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
[6]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Here we have calculate the total payload mass that NASA has sent, 45,596 kg.

```
# %sql select * from SPACEXTBL limit 3
%sql select Customer, sum(PAYLOAD_MASS__KG_)  from SPACEXTBL where Customer == 'NASA (CRS)'
```

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

Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596.0

Average Payload Mass by F9 v1.1

Here we see the average payload mass launched by the F9v1.1. The average mass is 2534.67 kg.

```
# %sql select * from SPACEXTBL limit 3
%sql select Booster_Version, avg(PAYLOAD_MASS__KG_)  from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

Here we see that January 8, 2018 was the first successful ground pad landing.

```
%sql select min(Date), Landing_Outcome from SPACEXTBL where Landing_Outcome =='Success (ground pad)'
```

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

min(Date)	Landing_Outcome
01/08/2018	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

Here we see the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
# %sql select distinct(Landing_Outcome) from SPACEXTBL
%sql select Booster_Version,PAYLOAD_MASS__KG_, Landing_Outcome from SPACEXTBL\
where Landing_Outcome == 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696.0	Success (drone ship)
F9 FT B1026	4600.0	Success (drone ship)
F9 FT B1021.2	5300.0	Success (drone ship)
F9 FT B1031.2	5200.0	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

total number of mission outcomes are listed below. we see a startling number of success. However, we must remember they are different than landing outcomes.

```
%sql select Mission_Outcome, count(Mission_Outcome) as Total_Outcome from SPACEXTBL group by Mission_Outcome
```

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

Mission_Outcome	Total_Outcome
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Here we see the boosters which have carried the maximum payload mass.

```
# %sql select * from SPACEXTBL limit 3
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_== (select Max(PAYLOAD_MASS__KG_) from SPACEXTBL) group by Booster_Version
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1049.7	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1060.3	15600.0

2015 Launch Records

Here we see the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(date,4,2) as Month,Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL\
where Landing_Outcome='Failure (drone ship)' and substr(Date, 7,4)='2015'
#Date,Landing_Outcome, Booster_Version, Launch_Site, Failure (drone ship)
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Here we see the ranked count of landing outcome success between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select  Landing_Outcome, count(Landing_outcome) as Number   from\
(select Date, Landing_Outcome from SPACEXTBL where Date between '04-06-2010' and '20-03-2017')\
where Landing_Outcome like 'Success%'\
group by Landing_Outcome\
order by Number desc\
```

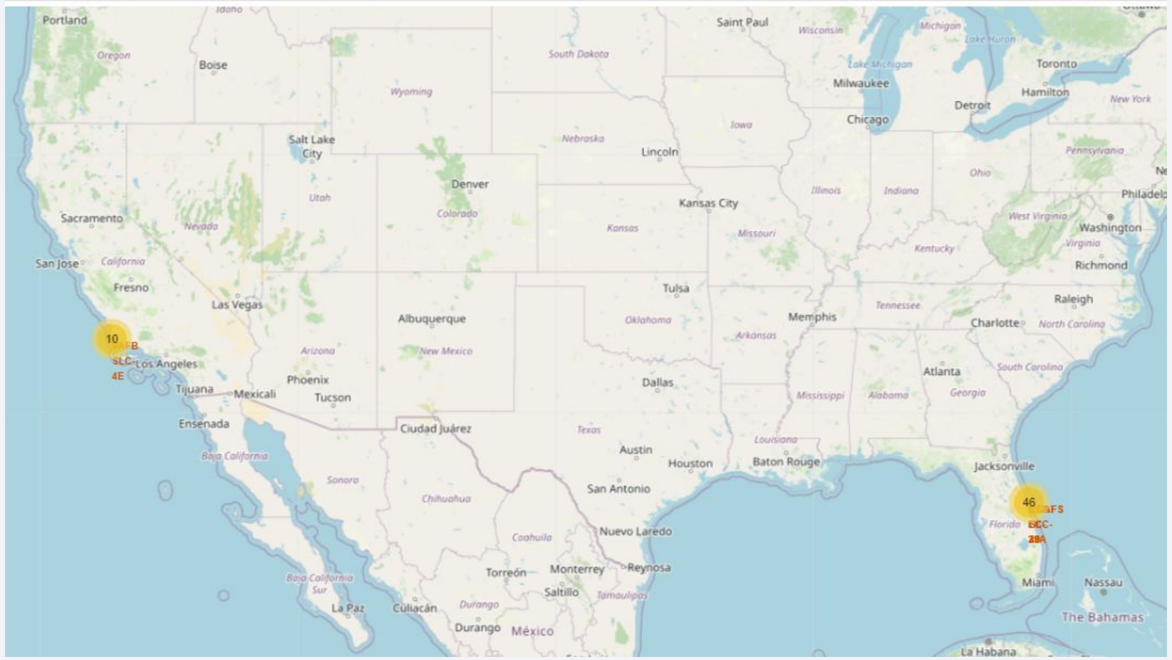
```
 * sqlite:///my_data1.db
Done.
```

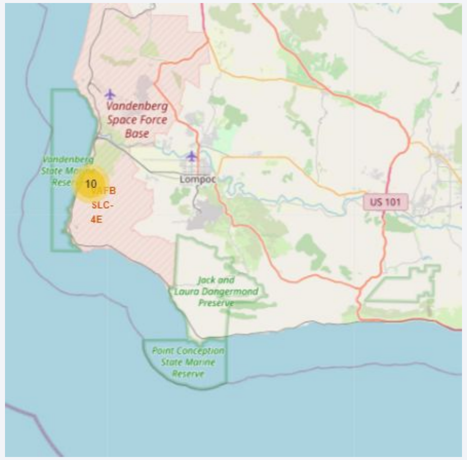
Landing_Outcome	Number
Success	20
Success (drone ship)	8
Success (ground pad)	7

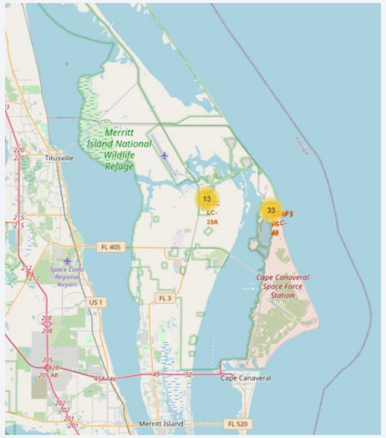
Section 3

Launch Sites Proximities Analysis

Location of SpaceX Launch Sites



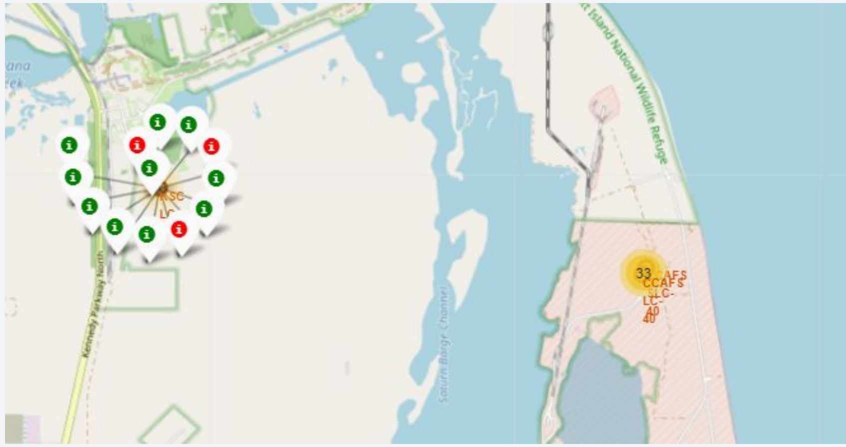


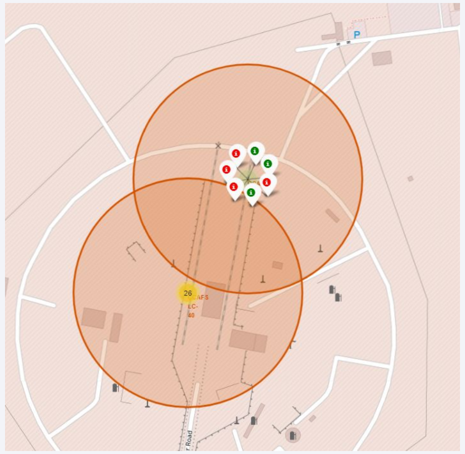


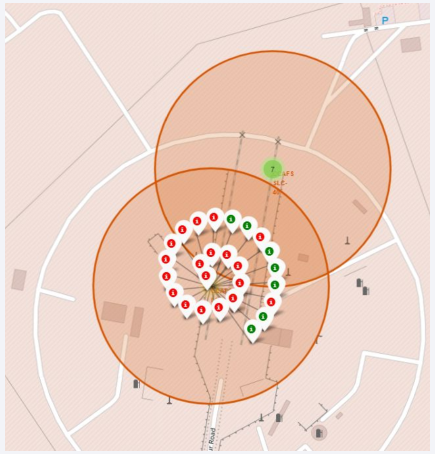
Here we see the launch sites locations. three location are on the Florida coast. and one location is on the California coast. the two smaller maps included show a zoomed in view of the locations.

Launch Outcomes at Each Launch Site

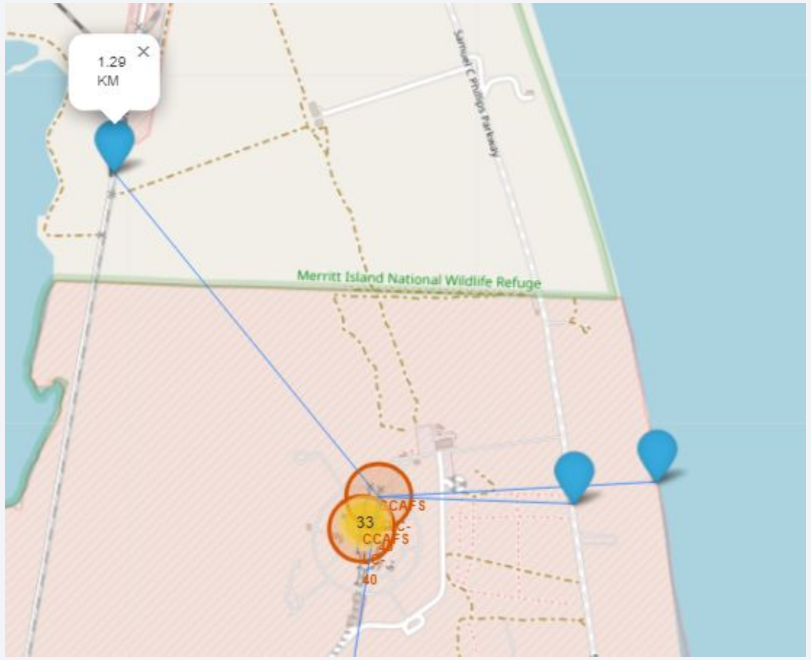
Here we see the outcomes from the different launch sites. Where CCAFS-LC40 has the most outcomes at 26. The outcomes labels colored red are failures while the labels colored green are the successes.

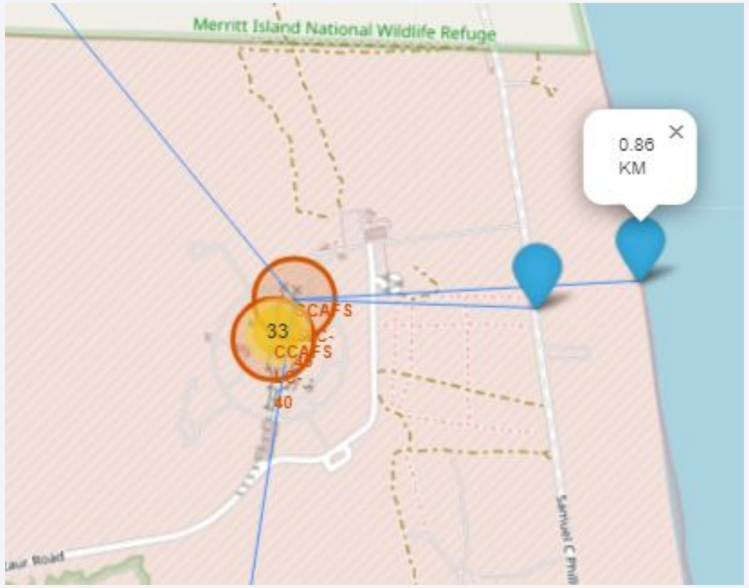






Proximity to a Launch Site



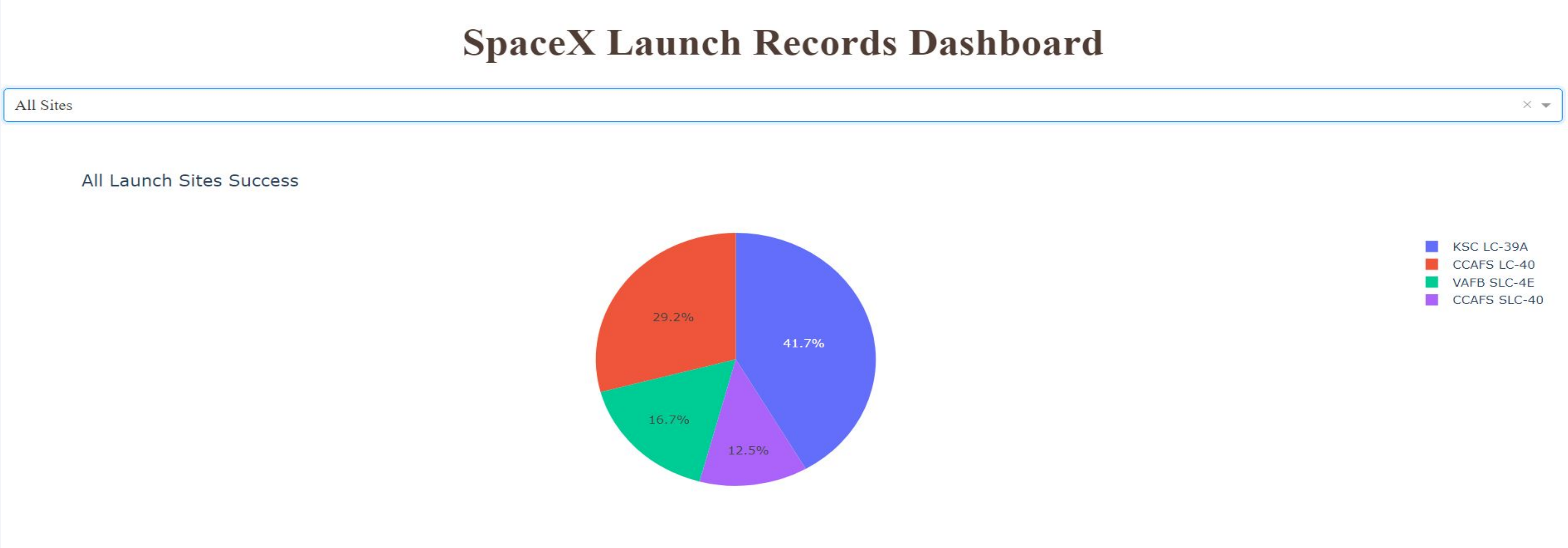


Here we see the proximity of the coast, a highway, a railway and major city to CCASF-SLC40. This demonstrates the power of the maps in gathering additional information making the data relatable.

Section 4

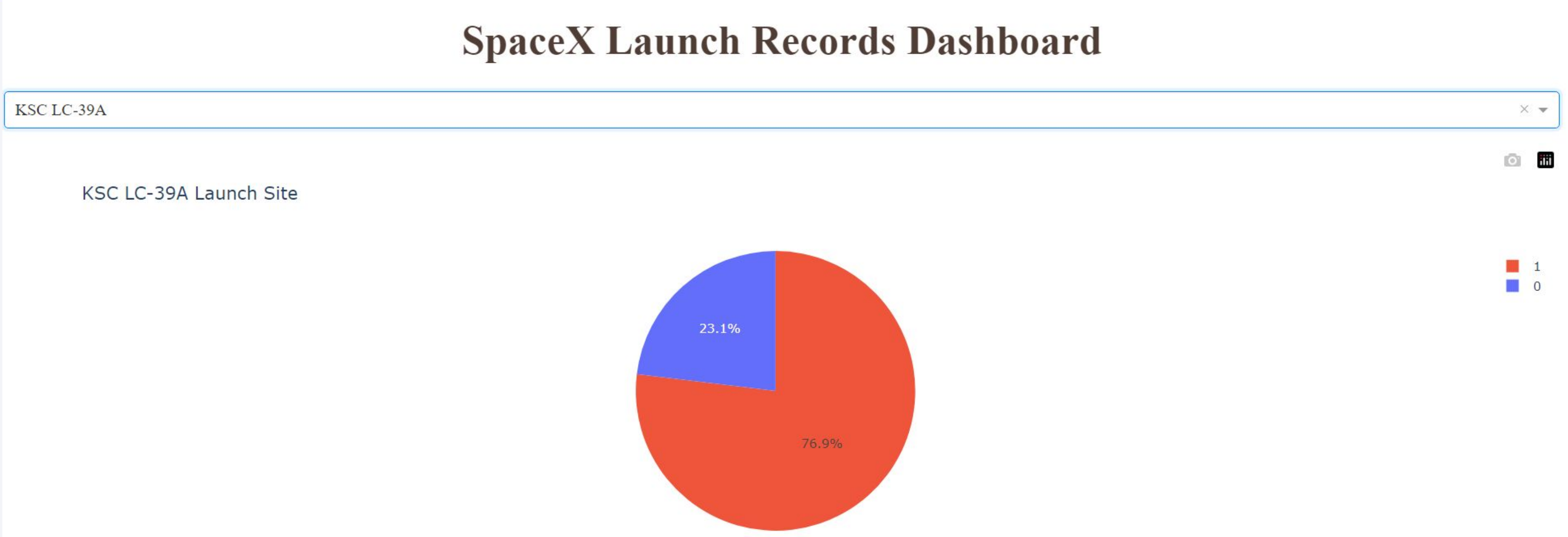
Build a Dashboard with Plotly Dash

Dashboard All sites



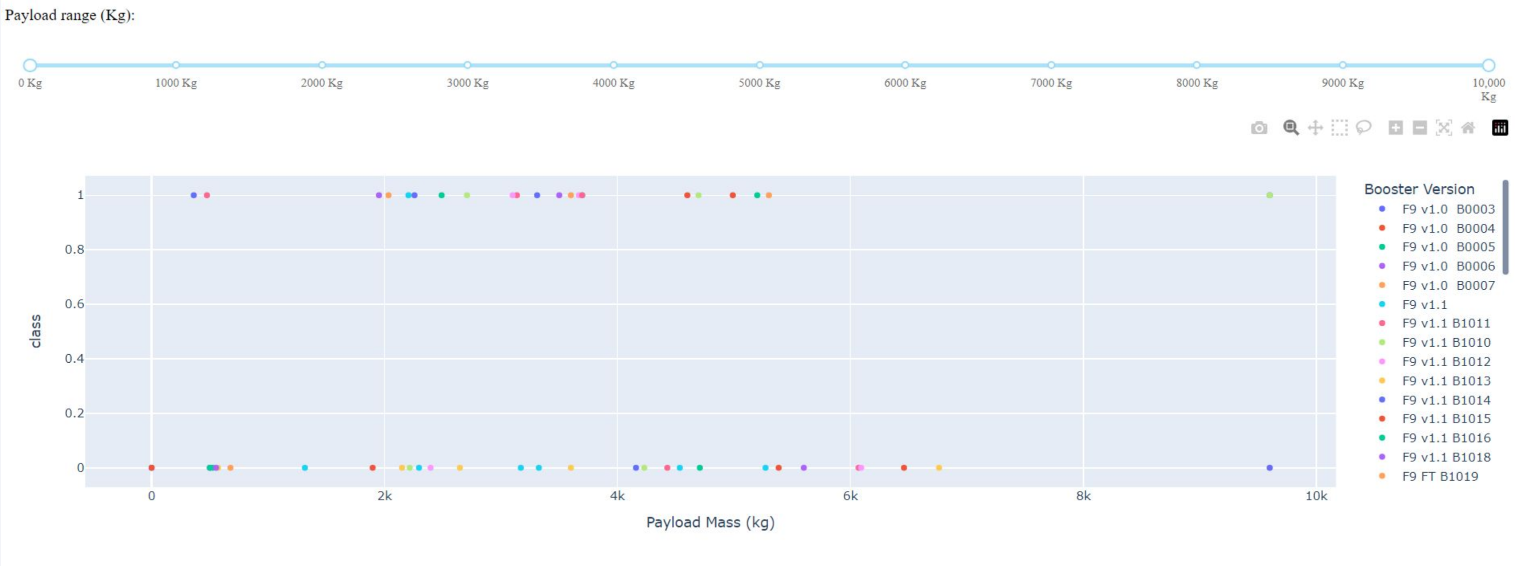
Here we see a screen capture of the dashboard highlighting which location had the highest percentage of successful flights. KSC LC-39A has the highest with CCAFS LC-40 coming in second.

A Dive Into The Site With The Highest Success



Here we are looking at the KSC LC-39A site with the most success. When interacting with the dash board we can see this site had a total of 13 launches with ten being successful. The next highest success, CCAFS LC-40, had a total of 26 flights.

Payload Mass Success Showing associated Booster



Here we are looking at a scatter plot of all Launch site Success and Failures. With watch type of booster identified. In the top chart we see the min payload to max payload. In the bottom chart we see 1 kg to 4 kg payload. We have 21 different boosters. In this interactive, chart we can restrict the payload to gain insight to which payload are significant.

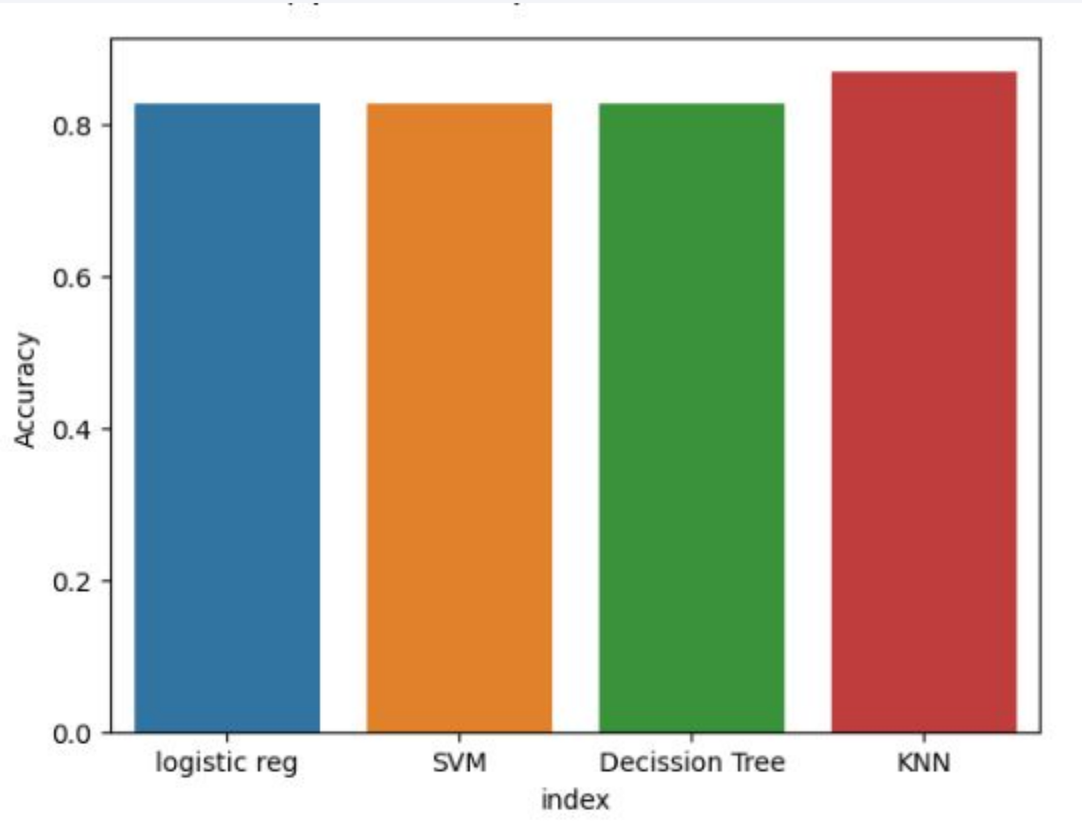
Section 5

Predictive Analysis (Classification)

Classification Accuracy

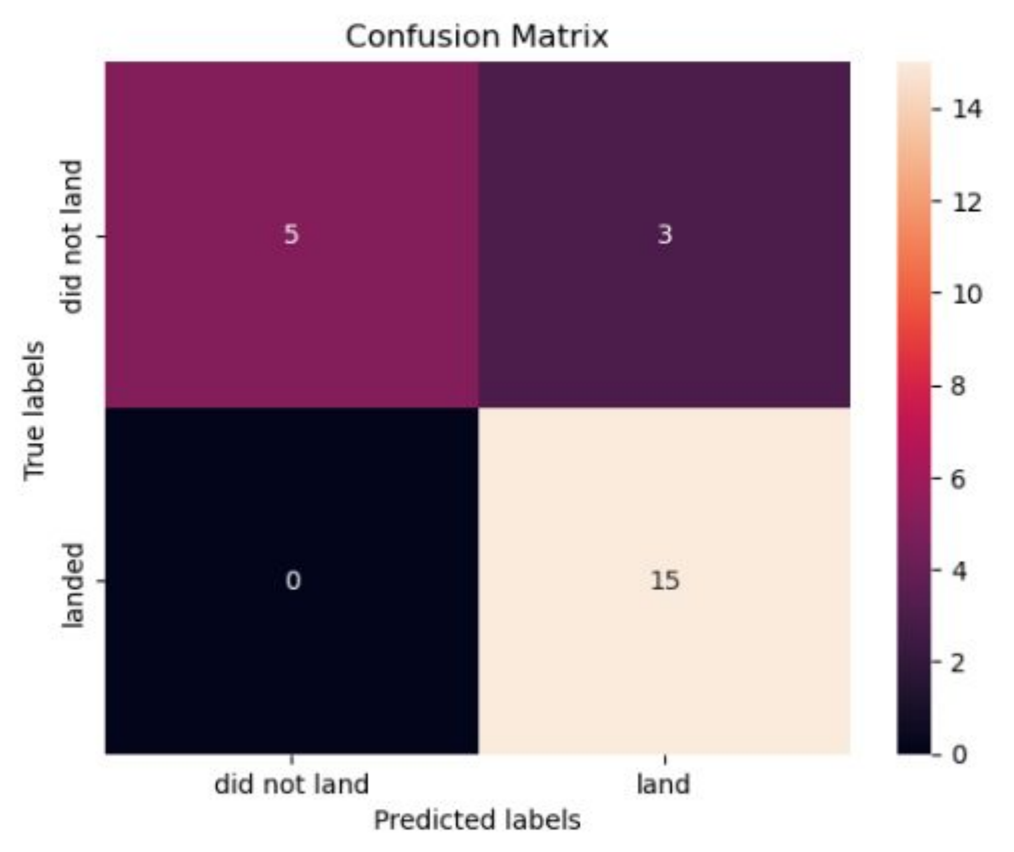
- Here we have a bar chart showing the model accuracy of the different classification algorithms

- it look like the KNN has the Highest model accuracy at 0.87



Confusion Matrix

Here we have the confusion matrix for the KNN model. We see that it predicted 15 landed accurately and predicted 5 did not land. This model did predict that 3 would land that actually didn't land.



Conclusions

- Our data shows that overtime and flights we saw improvements to success over time and the increased number of flights.
- We do see some evidence that orbit and payload have a roll in flight success.
- Likewise the VAFB SLC 4E site does appear to have greater success.
- With our KNN model we do see a fairly high predictive model
- However, Numbers are sparse. we need more data. probably double the data.
- The VAFB SLC 4E site had 13 of the 90 launches.
- Four of the 11 orbital mission on had one data point.
- This is a great start. and some insight has been gained, but more data.
- additional data providing weather patterns could prove helpful

Appendix

- https://www.spacex.com/vehicles/falcon-9/
- https://www.nbcnews.com/mach/science/how-much-does-space-travel-cost-ncna919011
- https://en.wikipedia.org/wiki/History_of_spaceflight

- Launch Site count

```
# Apply value_counts() on column LaunchSite
df.value_counts('LaunchSite')
```

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
dtype: int64
```

Appendix

- Orbit count

```
# Apply value_counts on Orbit column
df.value_counts('Orbit')
```

```
Orbit
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
GEO       1
HEO       1
SO        1
dtype: int64
```

- Landing outcomes success and failures ranked

```
%sql select  Landing_Outcome, count(Landing_outcome) as Number   from\
(select Date, Landing_Outcome from SPACEXTBL where Date between '04-06-2010' and '20-03-2017')\
group by Landing_Outcome\
order by Number desc\
```

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

Landing_Outcome	Number
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Thank you!