IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paul Archer-Smith
Oct 16, 2021

Outline

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

Executive Summary

- A critical competitive advantage of SpaceX is the ability to re-land and subsequently reuse first stage rocket boosters
- Relaunching a rocket saves a significant amount of money, as evidenced by a base launch cost of $62 million for SpaceX vs $165 million+ for other providers
- This project consisted of taking information about previous SpaceX launches (e.g. Booster Version and Orbit) and using it to determine which factors lead to a successful rocket landing
- Various data analysis techniques were used, including standard visualization tools, geographic analysis, and machine learning
- It appears that the strongest predictor of landing success is the booster version (e.g. Falcon 9 Block 5), indicating significant improvements have been made over the years by the SpaceX engineering & design teams

Introduction

- As mentioned on the prior slide, a critical competitive advantage of SpaceX is the ability to re-land and subsequently reuse first stage rocket boosters
- Reusing a first stage saves on the significant production costs associated with fabricating a rocket booster
- Understanding the relationship between launch variables and a successful landing allows both for evaluation of SpaceX's progress and a map of which areas should be more closely examined
- These factors lead directly to the central question this work tries to address:

What launch variables are the most indicative of a successful landing?

Section 1

Methodology

Methodology

Executive Summary

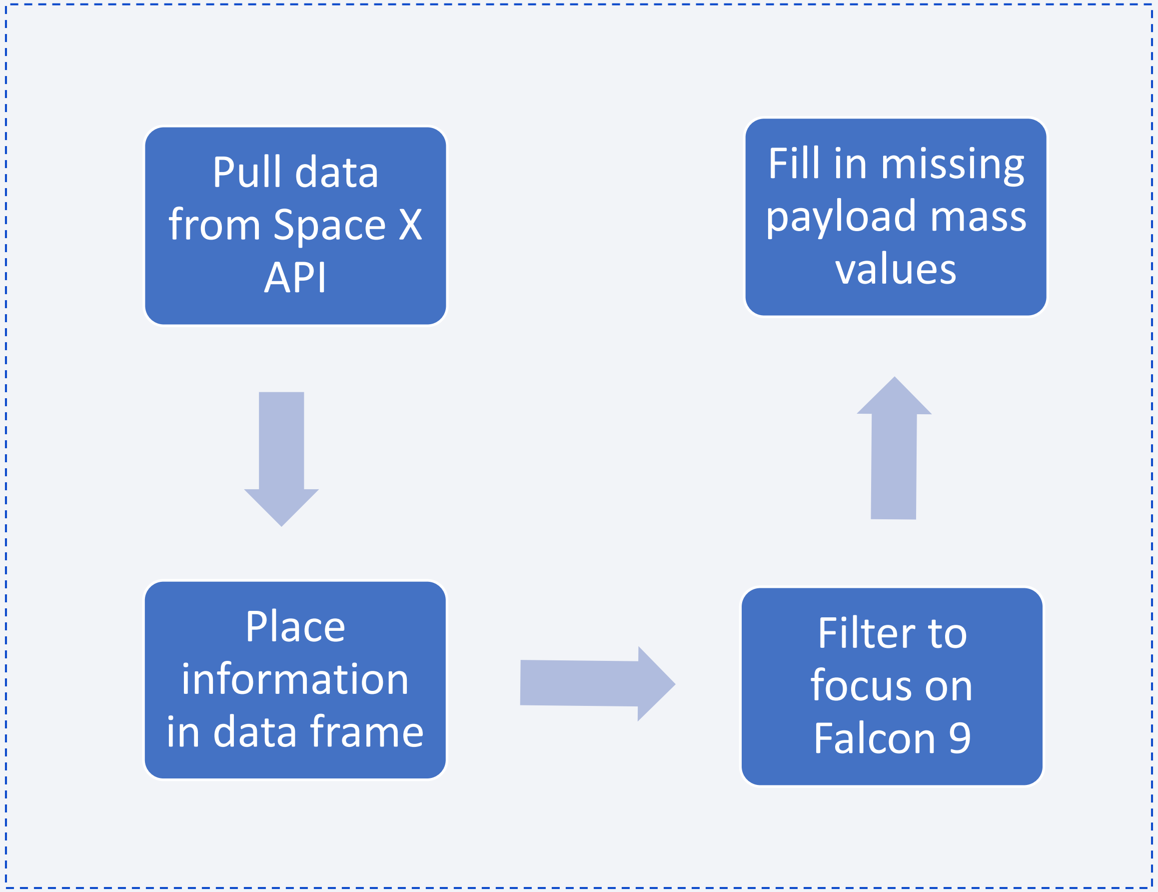
- Data collection methodology
  - Data collected via both the SpaceX API and web-scraping
- Perform data wrangling
  - Using Pandas, data was processes to removed anomalies and select relevant variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Launch data was analyzed using KNN, SVM, Logistical Regression, and Decision Trees

Data Collection

- The data sets used in this project were publicly available Space X launch information
- This includes features like booster version, payload mass, orbit type, date, launch site, and so on
- Data for this project was obtained through two primary methods:
  - 1) The Space X API: Data pulled from https://api.spacexdata.com/v4/launches/past
  - 2) Web-scraping: Beautiful Soup was used to gather data from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- In both cases the data was processed into data frames and cleaned for future use
- The next slides outline the main steps within these processes

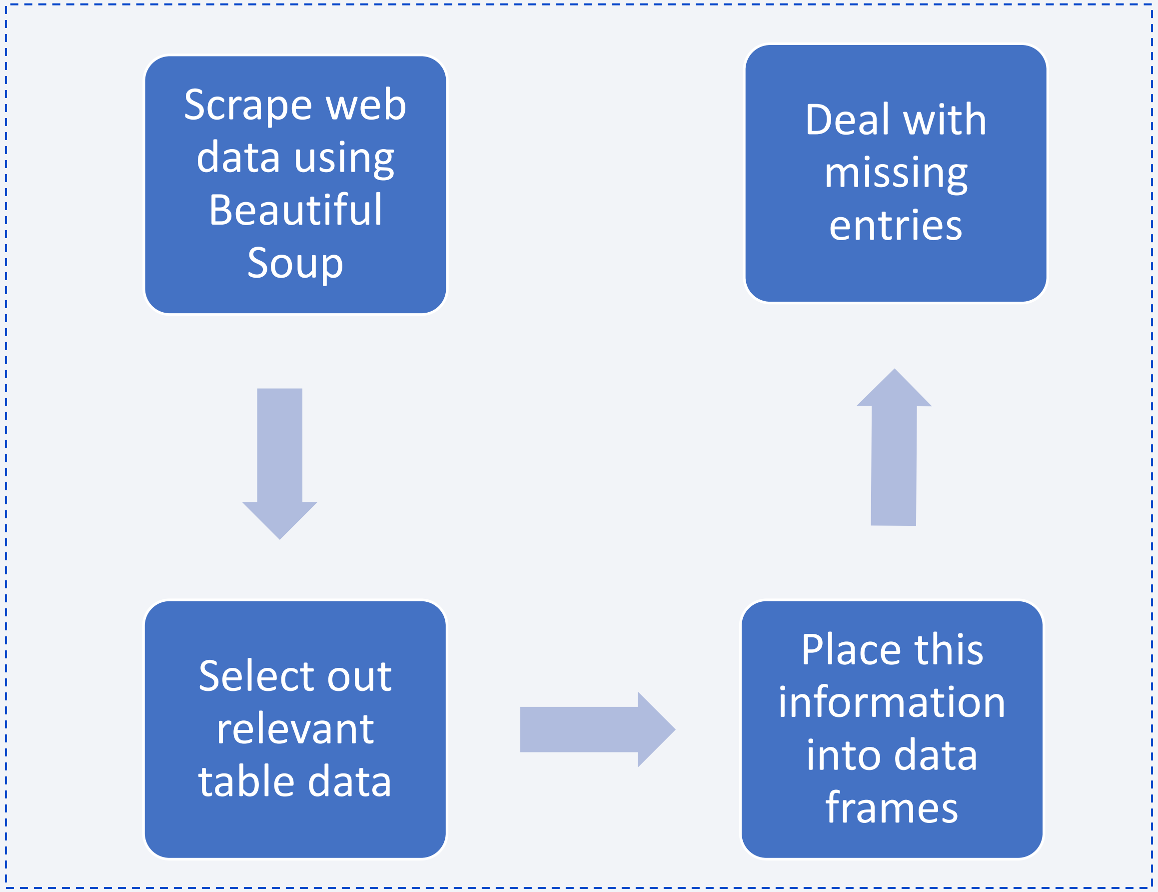
Data Collection – SpaceX API

- Using the SpaceX REST API, .json launch data was obtained
- This data was put into a Pandas data frame
- Falcon 9 data was selected
- Missing payload mass values (6) were replaced with mean payload mass
- GitHub URL: [IBMCapstone/Data Collection Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)]()



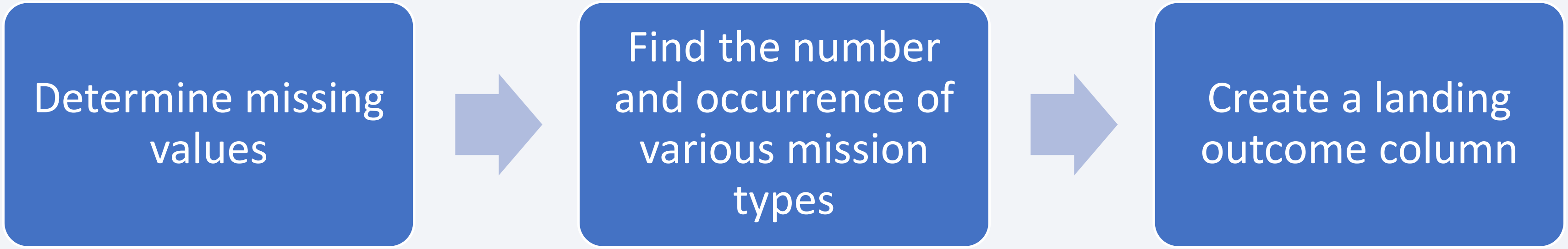
Data Collection - Scraping

- Beautiful Soup was used to gather data from: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Relevant launch data was stripped from tables listing past flights
- Prepared data frames in a similar fashion to the API approach
- GitHub URL: IBMCapstone/Data Collection from Web Scraping Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)



Data Wrangling

- The Pandas data frames obtained from the previous steps were checked for missing values and data attribute types
- The data was then grouped by various attributes (launch site, orbit, and landing outcome) so that the number and occurrence of each case could be viewed
- A new, binary-valued landing outcome column was added to the data frame.
- GitHub URL: IBMCapstone/Data Wrangling Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)



EDA with Data Visualization

- Several exploratory plots were constructed using the seaborn package
- Five scatter plots were created that examined successful landings as functions of payload and flight number, launch site and flight number, payload and launch site, flight number and orbit, and payload and orbit
- A bar chart showing the success rates for different orbits was constructed
- A line plot showing success rate as a function of time was also constructed
- Finally, categorical columns were converted to numerical dummy variables
- Add the GitHub URL: IBMCapstone/EDA and Visualization Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)

EDA with SQL

- The SpaceX launch information was uploaded to IBM-DB2 as a database table
- A list of unique launch sites was produced
- A variety of queries were made, obtaining the total payload mass carried for NASA, average payload mass for a F9 V1.1, the date of first successful ground landing, and the names of boosters with successful drone ship landings with mid-sized payloads
- The total number of successes and failures were determined, and two additional lists were generated: one of the boosters that carried a maximal payload and the other of boosters that failed to land on the drone ship in 2015
- Finally, a list ranking landing outcomes by the number of occurrences was produced
- GitHub URL: IBMCapstone/SQL Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)

Build an Interactive Map with Folium

- In order to better visualize the launch data, as well as look for any potentially relevant geographical features, an interactive map was built using Folium
- All utilized launch sites were marked and titled on a map
- All flights had a marker placed at their launch site with their name: these markers were coloured green if the landing was successful, red if not
- Key nearby geographic features to CCAFS SLC-40 were noted and the distance to these features was calculated
- GitHub URL: IBMCapstone/Interactive Visual Analytics with Folium Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)

Build a Dashboard with Plotly Dash

- Using Plotly, an internet dashboard was constructed to allow for easy exploration of launch data
- Two main plots were constructed: 1) A pie chart about successful launches and 2) a scatter plot showing landing success as a function of booster type and payload mass
- A drop-down menu allowed for the selection of flights from any individual launch site or all launch sites for the plot data (e.g. CCAFS SLC-40)
- A slider allowed for the selection of the permitted payload range for the scatter plot (e.g. 4000-7500 kg)
- GitHub URL: IBMCapstone/SpaceXDashboard.py at master · EasyAsQCD/IBMCapstone (github.com)

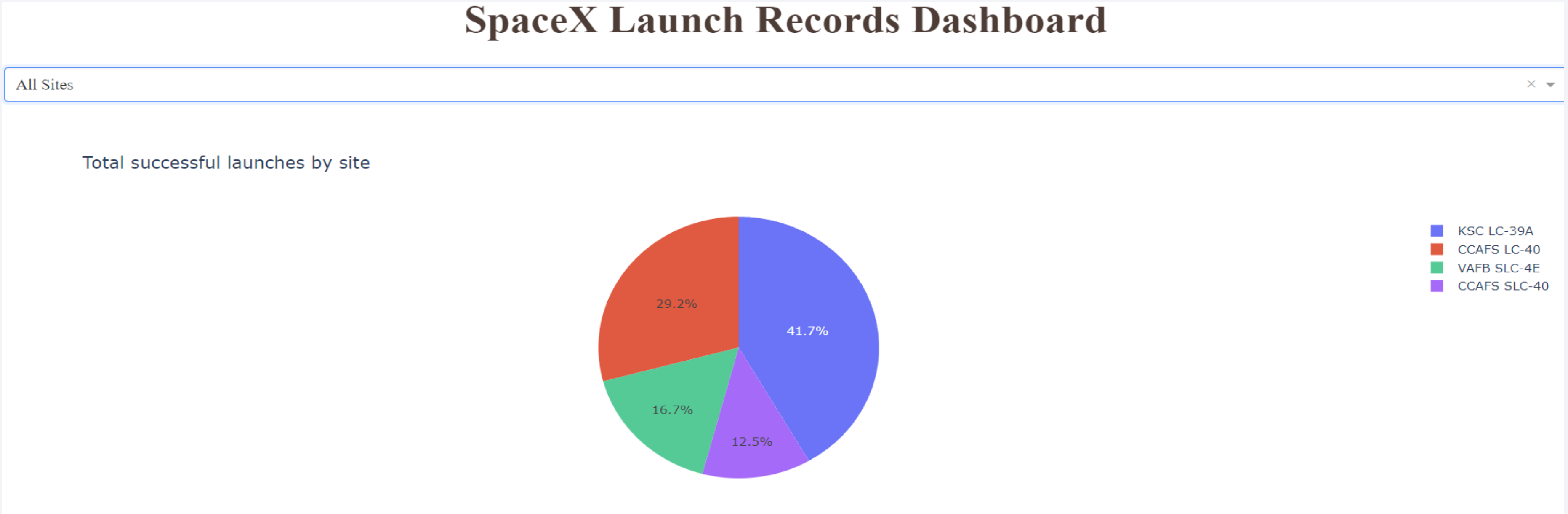
Predictive Analysis (Classification)

- Using the Scikit Learn package, the wrangled launch data was standardized, split into training and testing sets, and used to train various classifier models
- A grid search was done to find the optimal parameters for four different classifier models: logistical regression, SVM, decision tree, and K nearest neighbours
- The classification accuracy on the test set was calculated for each fitted classifier
- GitHub URL: [IBMCapstone/Machine Learning Lab.ipynb at master · EasyAsQCD/IBMCapstone (github.com)]

Standardize and split data into training and test sets → Using grid search, determined the optimal parameters for the classifier models → Fit the classifier using these parameters → Check the classifier accuracy on the test set

Results

- The exploratory data showed that the landing success rate has increased substantially over time --- especially as newer booster version were introduced
- This data also showed a correlation between landing success and payload mass; however, this is most likely because the later iterations of the Falcon 9 can carry larger payloads
- The interactive dashboard shows that the launch site KSC LC-39A has the highest number of successful launches (shown below)
- The predictive analysis showed that all the classifiers perform roughly the same on the test set and can produce an accurate prediction ~83% of the time.

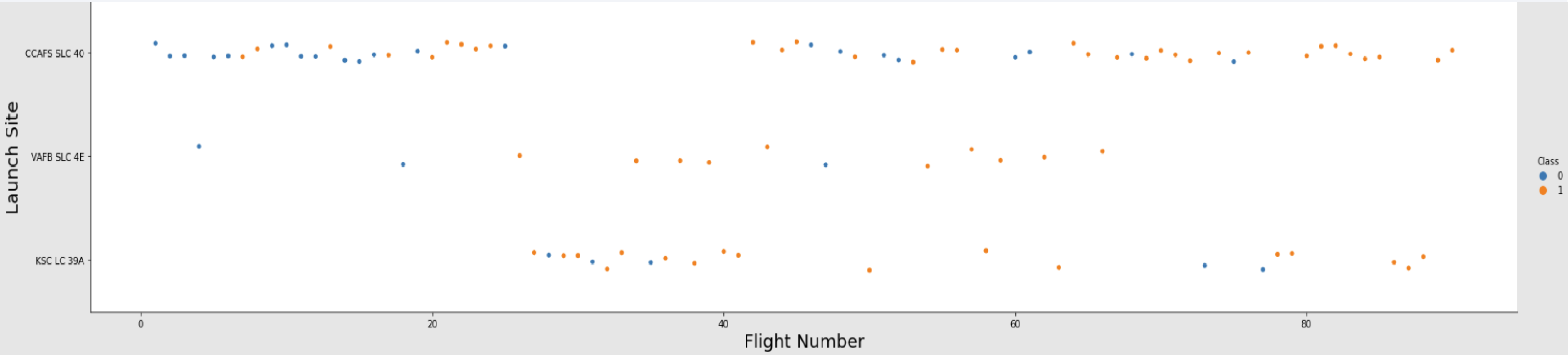


Section 2

Insights drawn from EDA

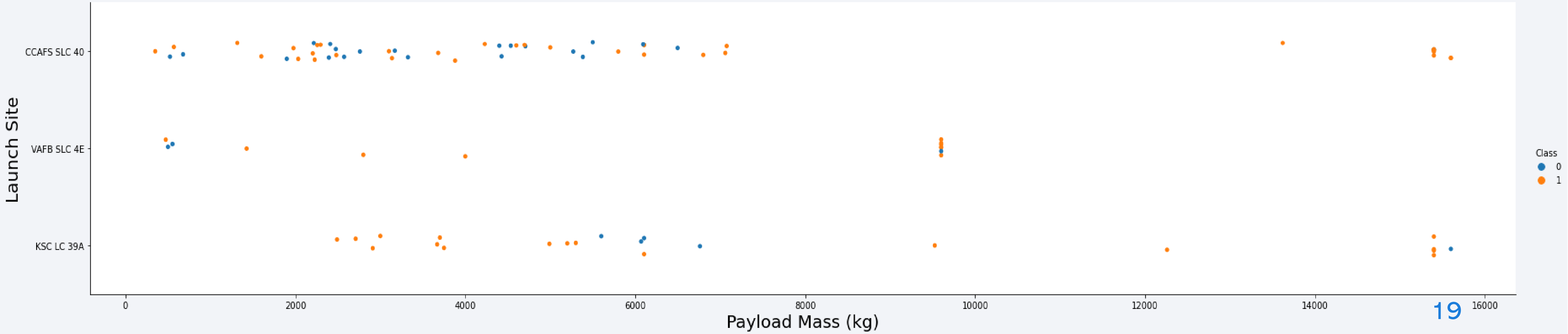
Flight Number vs. Launch Site

- Seaborn scatter plot of Flight Number vs. Launch Site
- Successful launches are shown in yellow, failed are shown in blue
- We can see that as time progresses (flight number increases) the relative number of successes increases
- It also should be noted that site VAFB SLC-4E features no recent launches



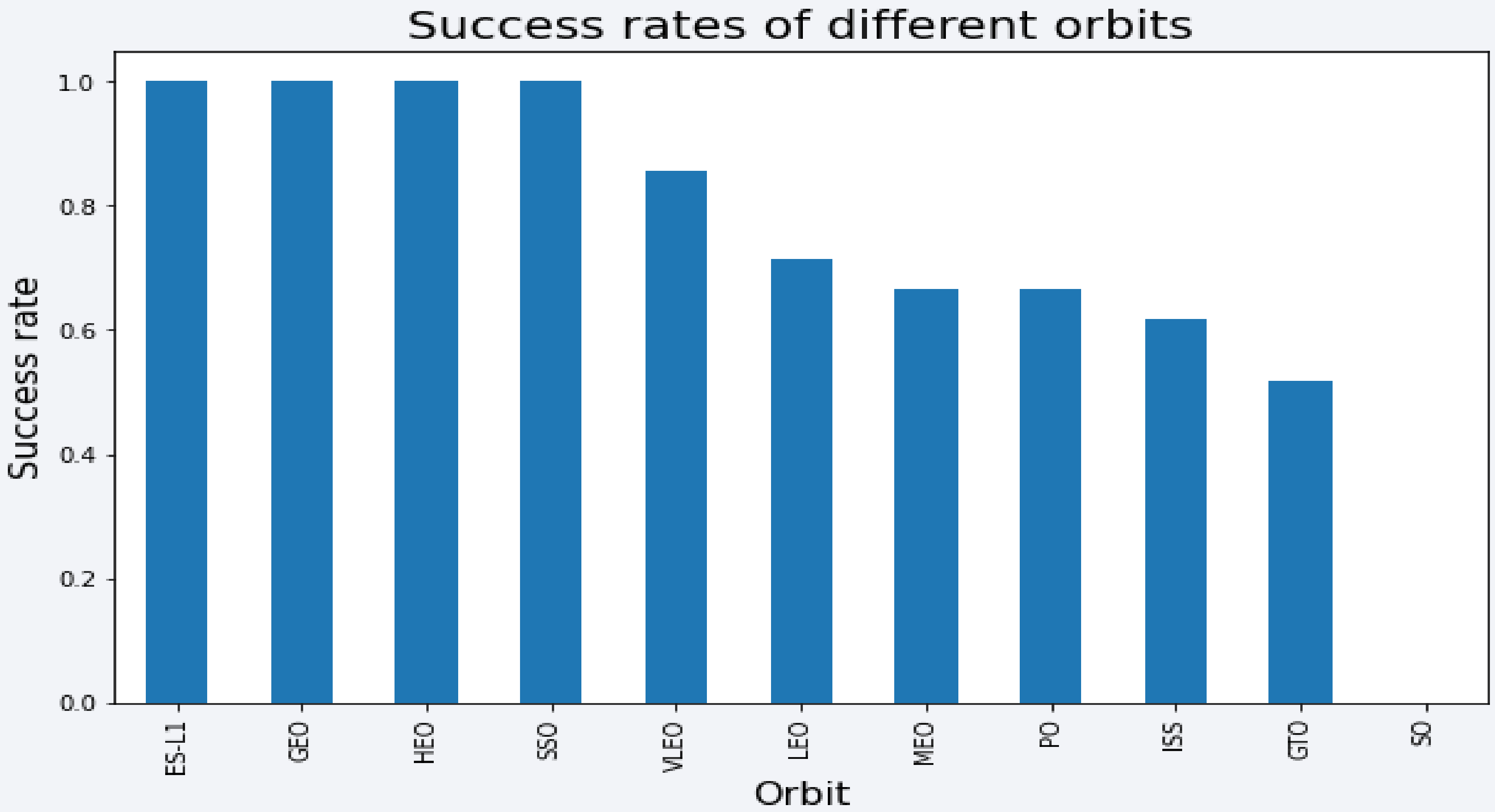
Payload vs. Launch Site

- Seaborn scatter plot of Payload vs. Launch Site
- Successful launches are shown in yellow, failed are shown in blue
- We can see that the relative success rate increases with payload mass
- This is likely due to later version of the Falcon 9 (e.g. Block 5) having greater lift capacity



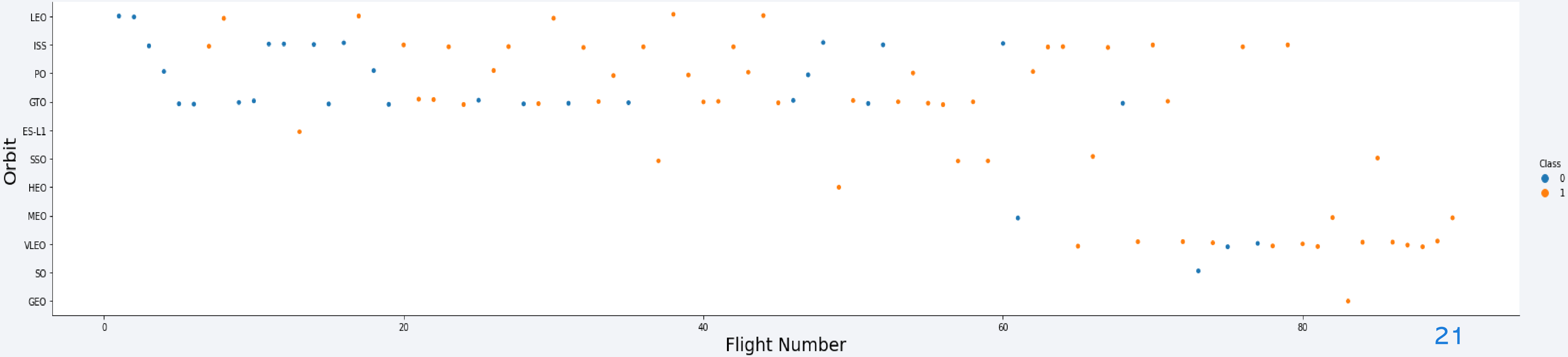
Success Rate vs. Orbit Type

- Seaborn bar chart for success rate as a function of orbit type
- Several orbits have perfect success rate; however, it should be noted that these orbits have significantly fewer launches than the more popular LEO and GTO



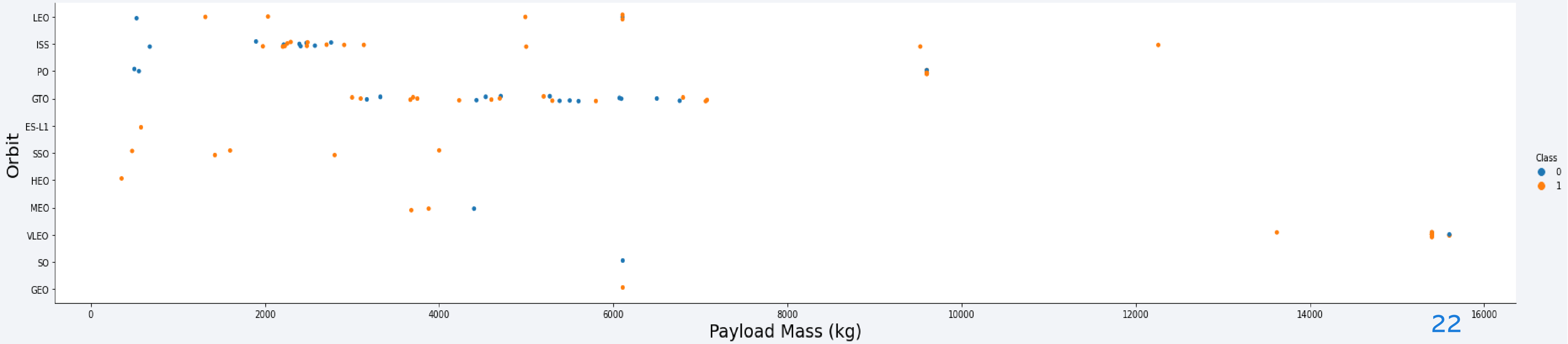
Flight Number vs. Orbit Type

- Seaborn scatter plot of Flight number vs. Orbit type
- Successful launches are shown in yellow, failed are shown in blue
- We can once again see that relative success increase with flight number
- This data also suggests that certain orbits are more attainable with newer versions of the Falcon 9



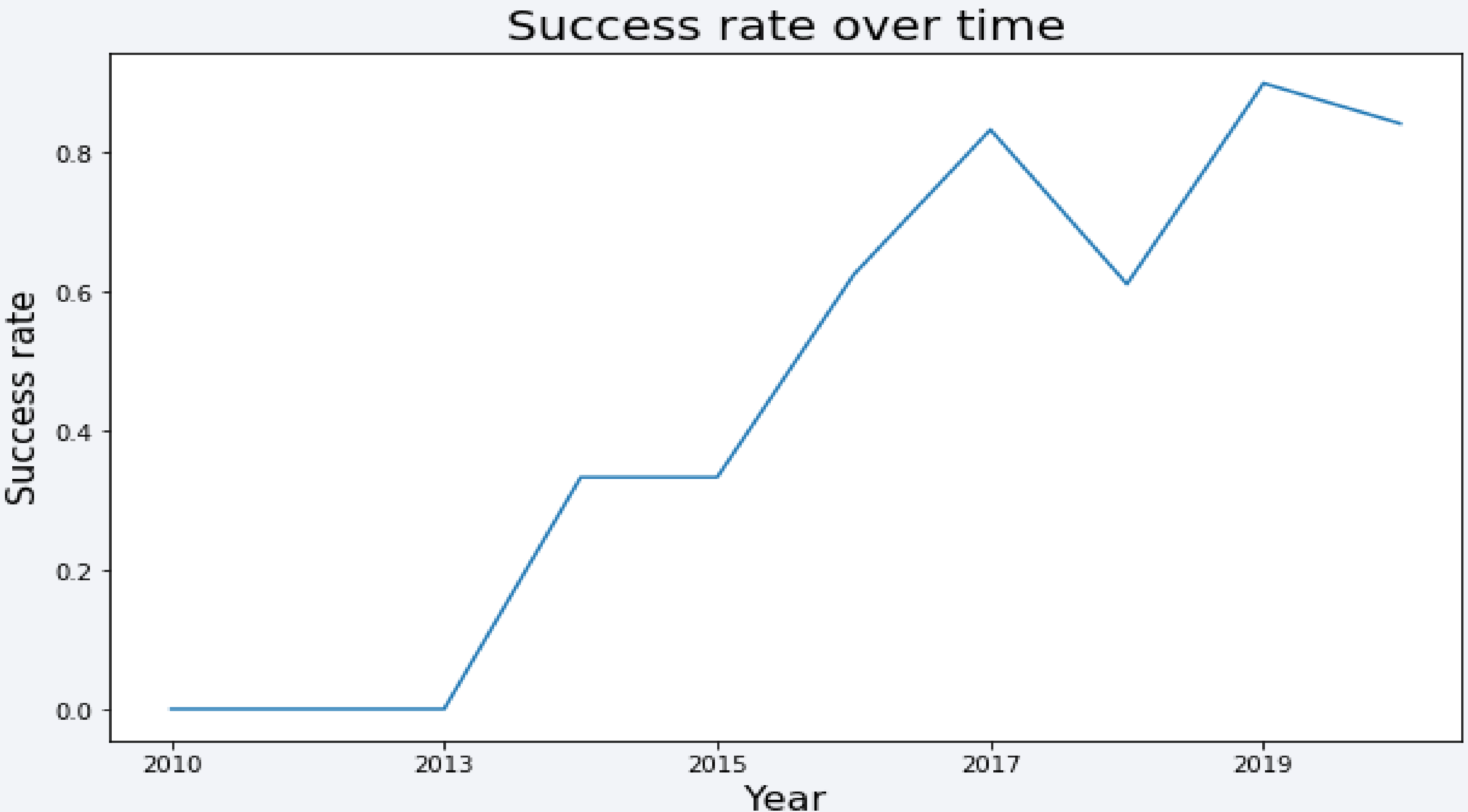
Payload vs. Orbit Type

- Seaborn scatter plot of Payload vs. Orbit type
- Successful launches are shown in yellow, failed are shown in blue
- Low orbits (VLEO and LEO) are shown to have very high success rates
- On the other hand, ISS and GTO orbits have much more mixed records
- In some cases, this may be due to the rocket being launched in expendable configuration



Launch Success Yearly Trend

- Seaborn line chart of the landing success rate over time
- Over time the landing success rate has dramatically improved to over 80%



All Launch Site Names

- From the database, a list of distinct launch sites was generated

```
In [7]: %sql select distinct(launch_site) from spacexdataset

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[7]:

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

Launch Site Names Begin with 'CCA'

- The first 5 records where the launch site began with `CCA` were pulled

In [8]:
```
%sql select * from spacexdataset where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

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

Total Payload Mass

- The total payload for NASA carried by SpaceX boosters was calculated

```
In [15]: %sql select sum(payload_mass__kg_) as total_payload_mass from spacexdataset where customer = 'NASA (CRS)';

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[15]:
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was determined

In [16]:
```
%sql select avg(payload_mass__kg_) as average_payload from spacexdataset where booster_version like 'F9 v1.1%';
```

```
 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[16]:

average_payload
2534

First Successful Ground Landing Date

- The first successful landing outcome on a ground pad occurred on:

```
In [20]: %sql select min(date) as first_success from spacexdataset where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[20]:
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- A list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 was generated

```
In [21]: %sql select distinct(booster_version) as ds_success from spacexdataset where landing__outcome = 'Success (drone ship)' and paylo
ad_mass__kg_ between 4000 and 6000;

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[21]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of mission successes and failures were calculated

```
In [22]: %%sql select sum(case when mission_outcome like 'Success%' THEN 1 ELSE 0 END) as successes,
             sum(case when mission_outcome like 'Failure%' then 1 else 0 end) as failures from spacexdataset;

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[22]:

successes	failures
100	1

Boosters Carried Maximum Payload

- A list of the names of the boosters which have carried the maximum payload mass was generated

```
In [31]: %sql select booster_version, payload_mass__kg_ from spacexdataset where payload_mass__kg_ in (select max(payload_mass__kg_) from spacexdataset);

 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[31]:

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

2015 Launch Records

- A list of the failed landing outcomes on a drone ship, their booster versions, and launch site names for 2015 was generated

In [36]:
```
%sql select landing__outcome, booster_version, launch_site from spacexdataset where year(date) ='2015' and landing__outcome = 'F
ailure (drone ship)';
```

```
 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[36]:

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

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

- A ranking of the number of occurrences of various landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates of 2010-06-04 and 2017-03-20 was constructed

In [50]:
```
%sql select landing__outcome, count(*) as count from spacexdataset where date between '2010-06-04' and '2017-03-20' group by lan
ding__outcome order by count desc;
```

```
 * ibm_db_sa://kkf83834:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[50]:

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

Section 4

Launch Sites Proximities Analysis

Folium Map --- Continental Scale

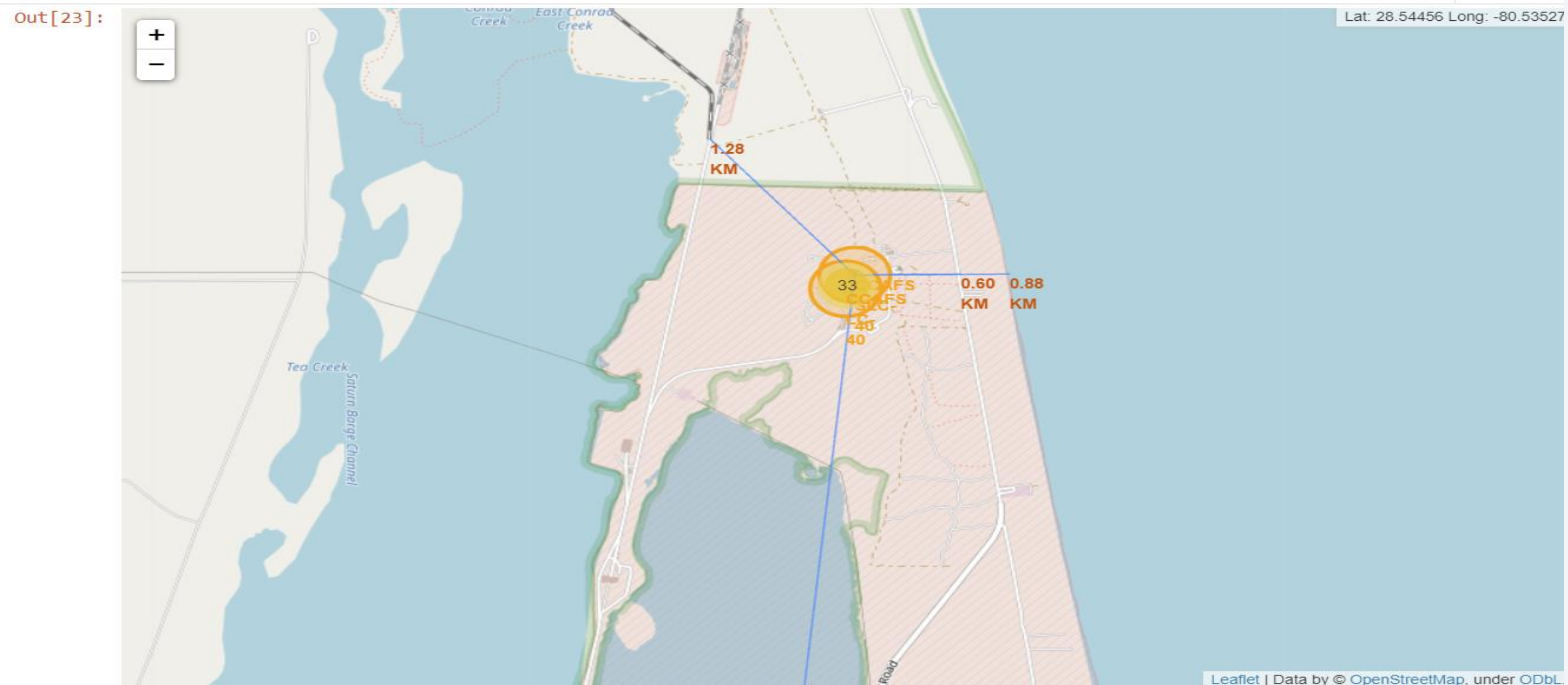
- Folium was used to construct a map of the launch sites and their corresponding launches
- Cluster markers were used to reduce map clutter

Folium Map --- Launch Site Scale

- This is a zoomed in shot of the folium map so we can see the CCAFS launch sites
- Green markers indicate a successful landing, red a failure

Folium Map --- Launch Site Proximity

- Folium map with calculated distance to nearby features
- The nearest: highway (0.6 km), coast (0.8 km), railroad (1.28 km), and city (17.47 km, out of shot) are marked by the far end points of the blue lines

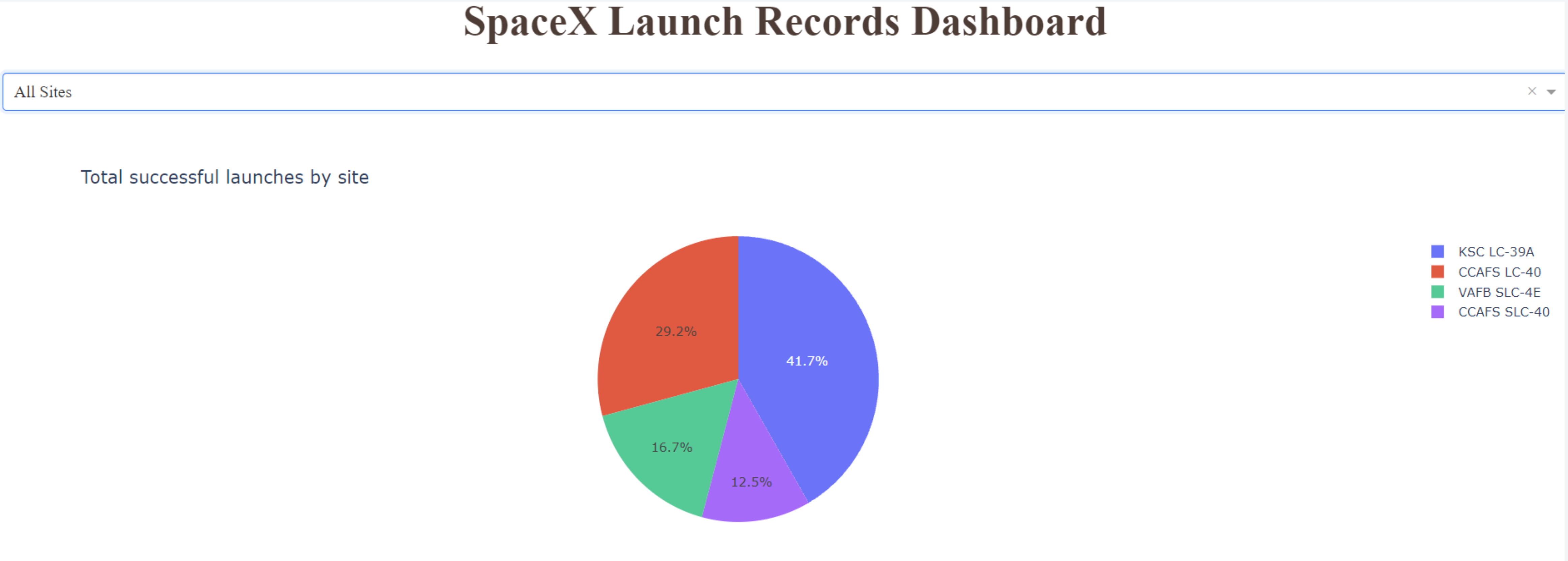


Section 5

Build a Dashboard with Plotly Dash

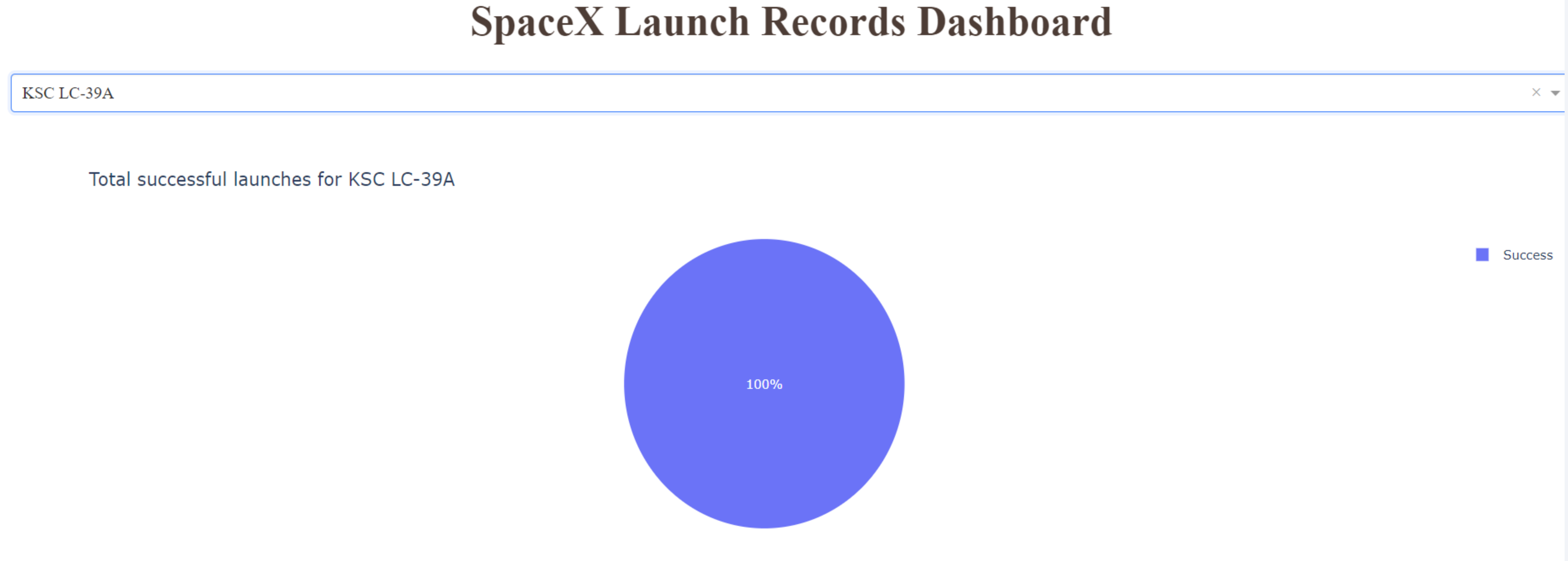
Plotly Dashboard --- Launch Successes

- A Plotly internet dashboard was constructed in order to create an interactive way to select and visualize the SpaceX data
- The first result, shown below, is a pie chart breaking down which launch site successful flights originated at



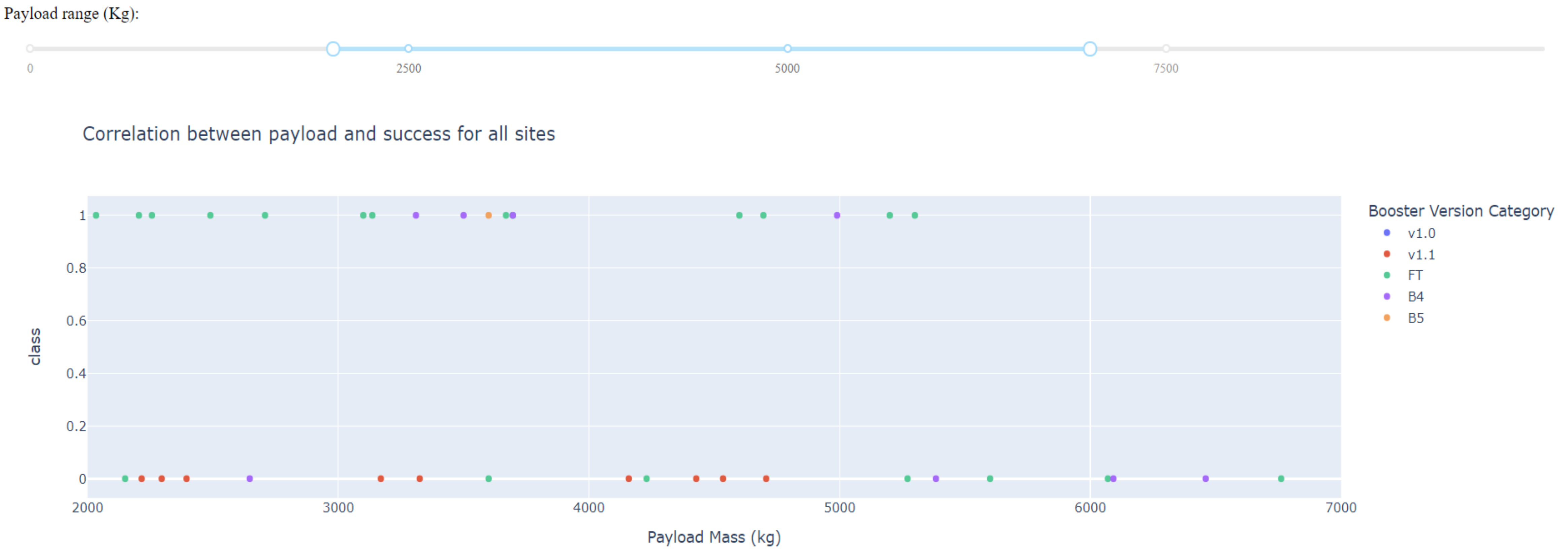
Plotly Dashboard --- Highest Launch Success

- The sites with the highest launch success rates are KSC LC-39A and VAFB SLC-4E with 100%
- This can be seen in the below pie chart



Plotly Dashboard --- Payload vs Launch Outcome

- The final part of the Plotly internet dashboard is a scatter plot that plots launches according to class and payload mass
- The colour of the points indicates the booster category and the slider at the top of the image can be used to control the range of the x-axis

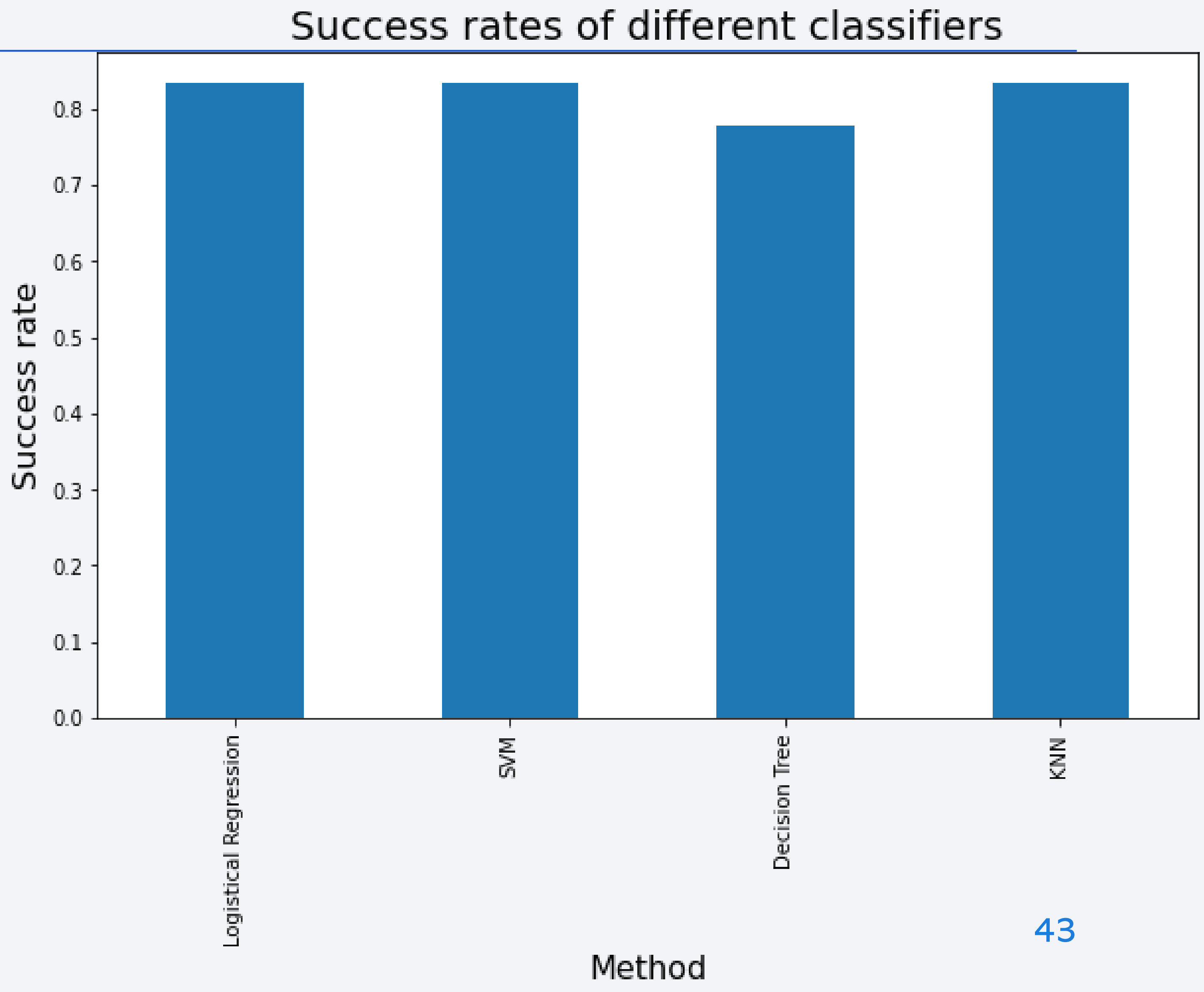


Section 6

Predictive Analysis (Classification)

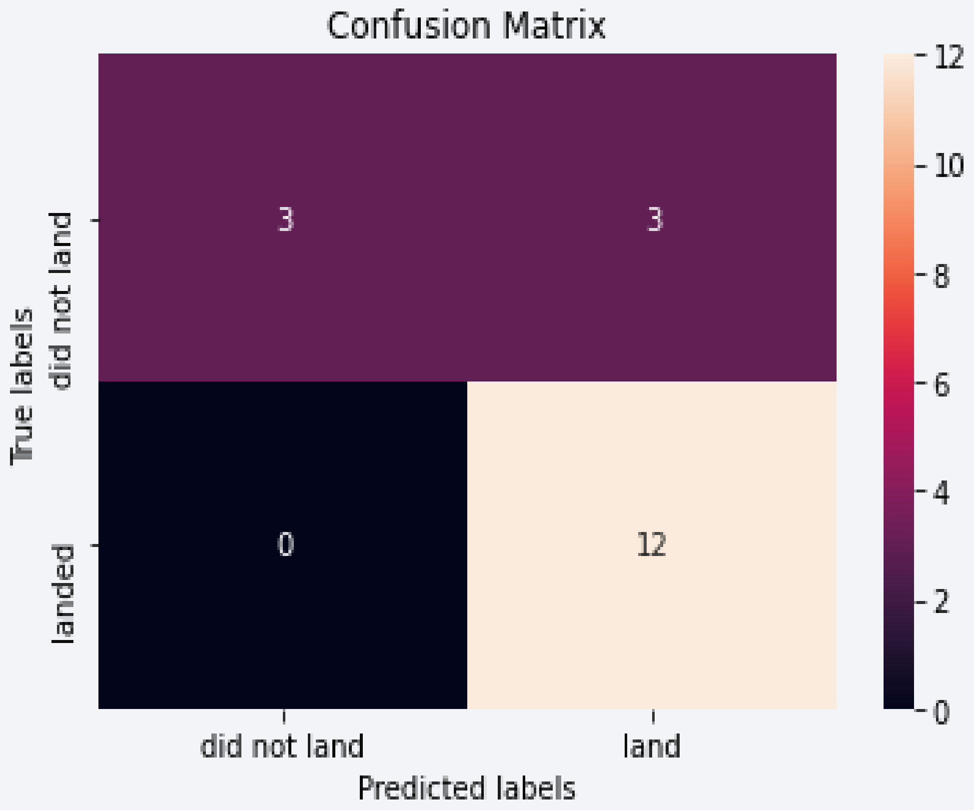
Classification Accuracy

- Four different types of classification models were tested on the SpaceX data: logistical regression, SVM, decision tree, and K nearest neighbour
- After doing a grid search of the hyperparameters of each model, the optimal configurations led to the following test set success rates
- The methods are roughly equivalent; although the decision tree algorithm performs worse than the others, it is by a very small margin



Confusion Matrix

- KNN, SVM, and logistical regression all work equally well on our SpaceX test set data
- The have an 83% accuracy with the only mistakes coming from false positives
- Due to the similarities in accuracy for the different classification methods, it seems reasonable to use whichever method that runs the quickest/is the most resource efficient



Conclusions

- The most crucial point that we can take away from this work is that the development and engineering teams have made massive strides over the last several years as the newer versions of the Falcon 9 have much higher success rates
- This is shown explicitly in the improvement in performance year over year and in the scatter plots involving flight number
- GTO missions still seem to have a relatively high number of missions that do not re-land; this may be due to this orbit requiring more frequent “expendable” configuration use
- Boosters are carrying larger payloads more often, showing the increased power of newer rocket configurations

Appendix

- In case people are wondering:

Successful landing

Failed landing

Thank you!