IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vanja Blazinic
March 13th, 2024

Outline

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

Executive Summary

Week 1 one deals with the data collection through SpaceX API and parsing with *requests* and *json* libraries, then data collection from Falcon9 launch page from Wikipedia and parsing with *BeautifulSoup* library. The data is then transferred to *pandas* dataframe for data wrangling stage, namely dealing with the missing data, either by removing entries or replacing entries with mean values (which are determined from column operations in pandas dataframe, e.g. payload mass). Column operations are further used to calculate informative values such as the number and frequency of mission orbits and frequency of mission outcomes, etc. **Week 2, part 1** deals with exploratory data analysis through *sqlite3* library, in order to demonstrate various techniques of data retrieval using SQL syntax in Jupyter environment. **Week 2, part 2** deals with data visualization by exploring relationships between different variables (e.g. mission success rate vs orbit type, payload mass vs orbit type, etc) using *seaborn* library. **Week 3** focuses on creating interactive maps and charts by using Folium and Plotly, respectively. **Week 4** focuses on using four different machine learning algorithms (logistic regression, support vector machine, decision tree and K-nearest neighbours) to determine which one is the optimal one for future model predictions.

Aforementioned machine learning models were trained and tested on the existing data. Of the four, the highest accuracy R2 score (on train dataset) was achieved for the decision tree, but when applying the respective models on the test dataset, the R2 score acquired was the same four all four models. Inability of distincting better results may be attributed to the fact that test dataset consisted of only 20% of available total dataset. In absolute numbers, that is 18 test samples which may have not been sufficient.

Introduction

Motivation

A new age of space exploration is upon us, driven by several companies in the private sector (Virgin Galactic, Rocket Lab, Blue Origin, SpaceX). Of those, SpaceX is the most popular one. The company's *sales pitch* is the reusability of the first stage of its rockets, namely Falcon 9. This reduces the company's costs down to 62 million USD per launch as opposed to roughly 165 million USD per launch with its competitors.

In this scenario, we are contracted by SpaceY (owned by Allon Mask) to evaluate conditions under which Falcon 9 successfully launch and are recovered.

Problem

Based on the publicly available data from SpaceX and Falcon 9 launch wiki information, we are tasked to determine what conditions influence the success of Falcon 9 mission outcome and on which input data to predict it.

The priority is also given to discovering which machine learning algorithm is good for making such predictions.

Section 1

Methodology

Methodology

Executive Summary

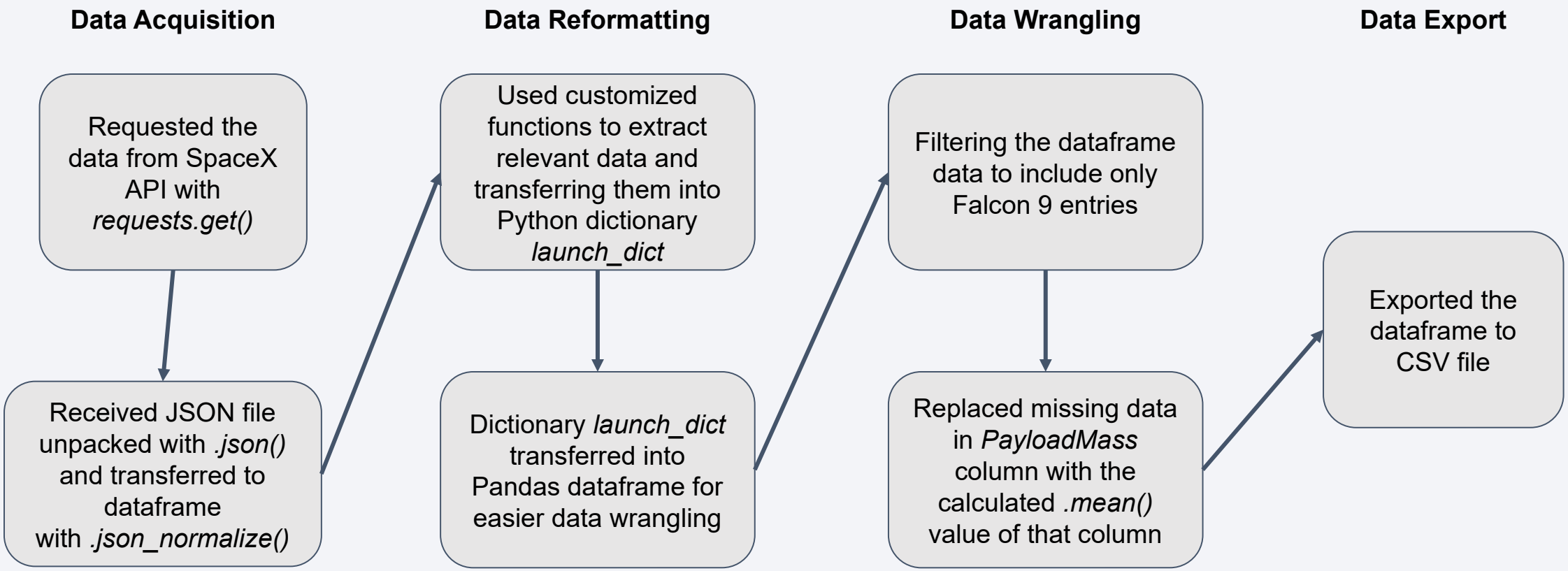
- Data collection methodology:
  - Data collection performed with SpaceX API by web scraping from Falcon 9 launch Wikipedia page via *BeautifulSoup* library in Python.
- Perform data wrangling
  - Data filtering (e.g. removing entries with 2 rockets, date format conversion, etc.)
  - Replacing missing values (e.g. NaN values in payload masses replaced with a mean)
  - One hot encoding of categorical variables with *pandas.get_dummies* function
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building and evaluation of four models on dataset split into train and test subsets.

Data Collection

- With SpaceX API:
  - Acquired raw data with *requests.get()*. Received JSON file decoded with *.json()* and imported into Pandas dataframe with *.json_normalize()*.
  - Extracted a subset of columns, from which we extracted the additional columns via functions customized for the exercise.
  - Data were filtered only to include Falcon 9. Missing data were found in payload mass column and replaced with a mean value.
  - Final dataset included information like launch date, mission outcome, geographical coordinates of launch site, payload mass, serial number of flight and orbit type.
- With *BeautifulSoup* library
  - Inspected the target HTML table on Falcon 9 launch Wikipedia page.
  - Acquired the page content with requests.get() and parsed it with BeautifulSoup html.parser.
  - Extracted the data from parsed HTML table into Pandas dataframe via functions customized for the exercise.

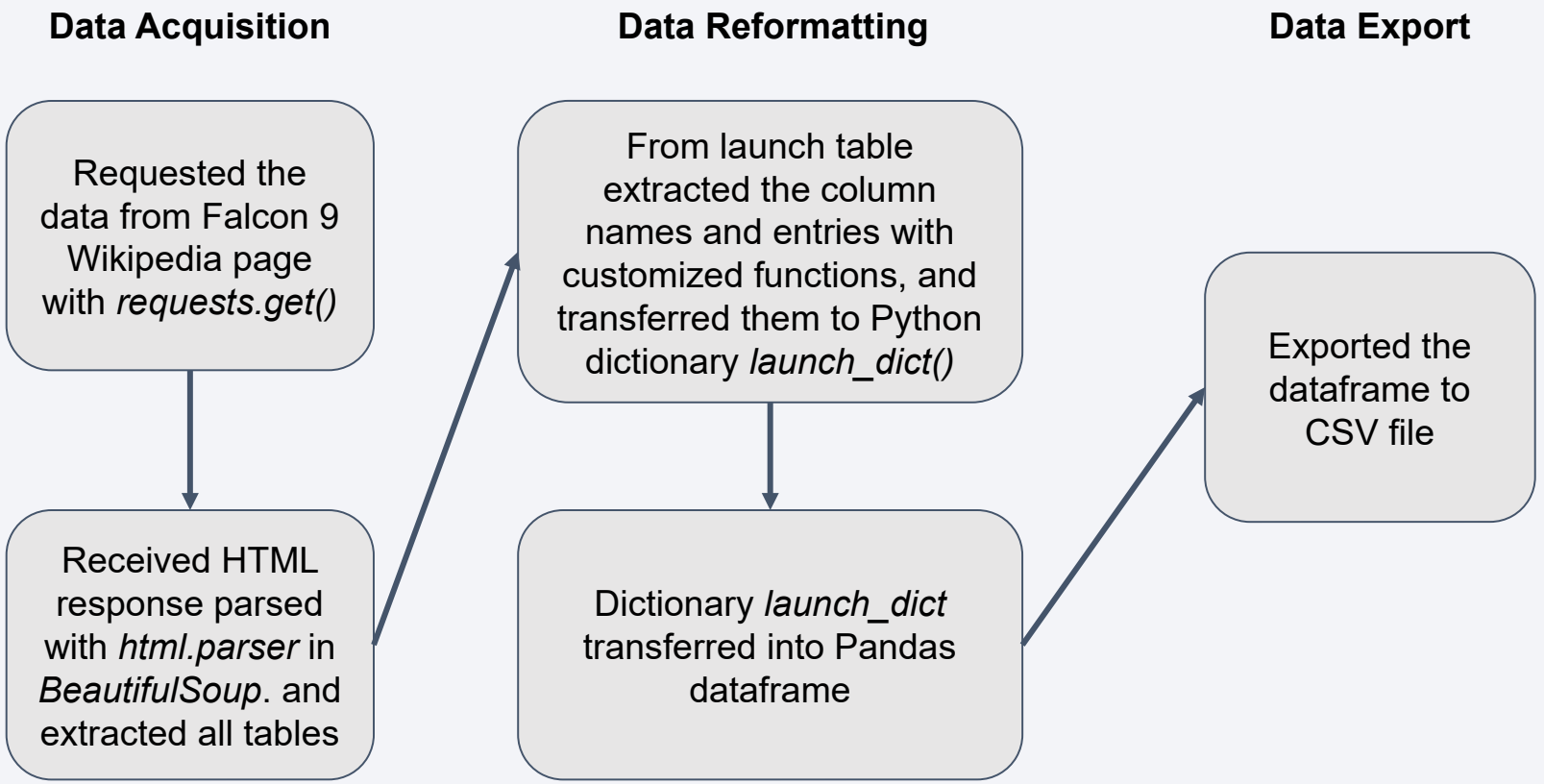
Flowcharts for each process shown in the following slides.

Data Collection – SpaceX API



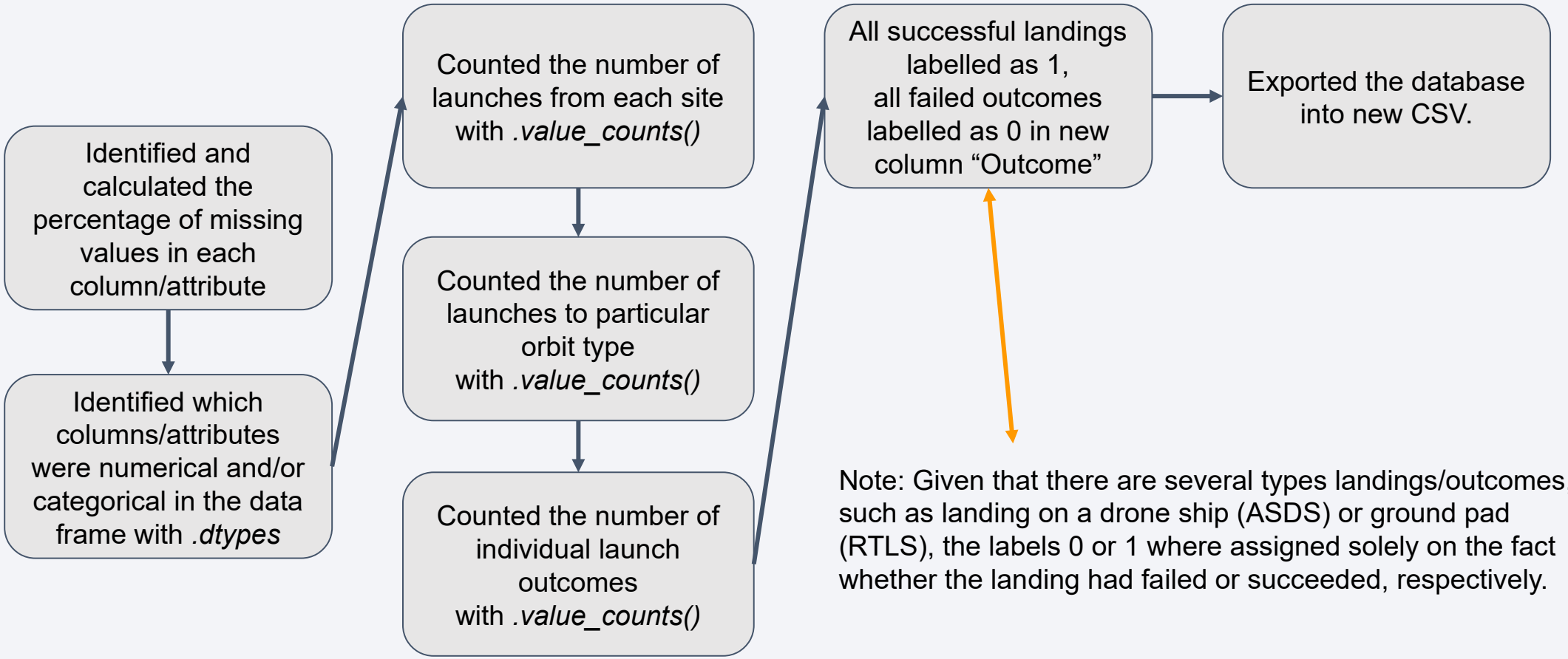
URL: GitHub SpaceX API Data Collection

Data Collection - Scraping



URL: GitHub SpaceX Web Scrapping

Data Wrangling



URL: GitHub SpaceX Data Wrangling

EDA with Data Visualization

- Following scatter charts (x-y) were made:
  - Flight number vs Payload mass, Flight number vs Launch site, Payload mass vs Launch site, Flight number vs Orbit type, Payload mass vs Orbit type

- Following bar charts were made:
  - Orbit type vs Success rate

- Following line charts (x-y) were made:
  - Success rate per year

Scatter plots were made to visually inspect if there is a correlation between respective *x* and *y* variables for each pair. Bar chart was made in order to visually represent the success of missions to specific orbits. Line chart was made to visually represent the trend of the successful missions over the years. Results of individuals visualisations will be commented in the following sections.

URL: GitHub SpaceX Data Visualisation

EDA with SQL

Following SQL queries were made:

- Selecting distinct launch sites
- Selecting and displaying 5 records whose launch sites contain string sequence “CCA”
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass by booster F9 v1.1
- Find and list a date of the first successful landing in ground pad.
- Listing the names of the boosters which have successful outcomes in drone ship, payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failed outcomes
- Listing the names of booster versions which have carried the maximum payload mass by using a subquery
- List months, failed outcomes in drone ship, dates, booster versions and launch sites in year 2015
- Rank the count of landing outcomes between June 4th, 2010 and March 20th, 2017

URL: GitHub SpaceX EDA with SQL

Build an Interactive Map with Folium

- Following was done in Folium:
  - Markers with circles were for all launch sites, by using their longitude and latitude coordinates for positioning, along with total number of launches via Text Label.
  - Color markers were added by attributing red for failed and green for successful launches with MarkerCluster(). This appear as popups when individual launch site is selected.
  - One line from CCAFS launch site to the coast was added, along with the distance of 0.86 kilometres.

URL: GitHub SpaceX Folium

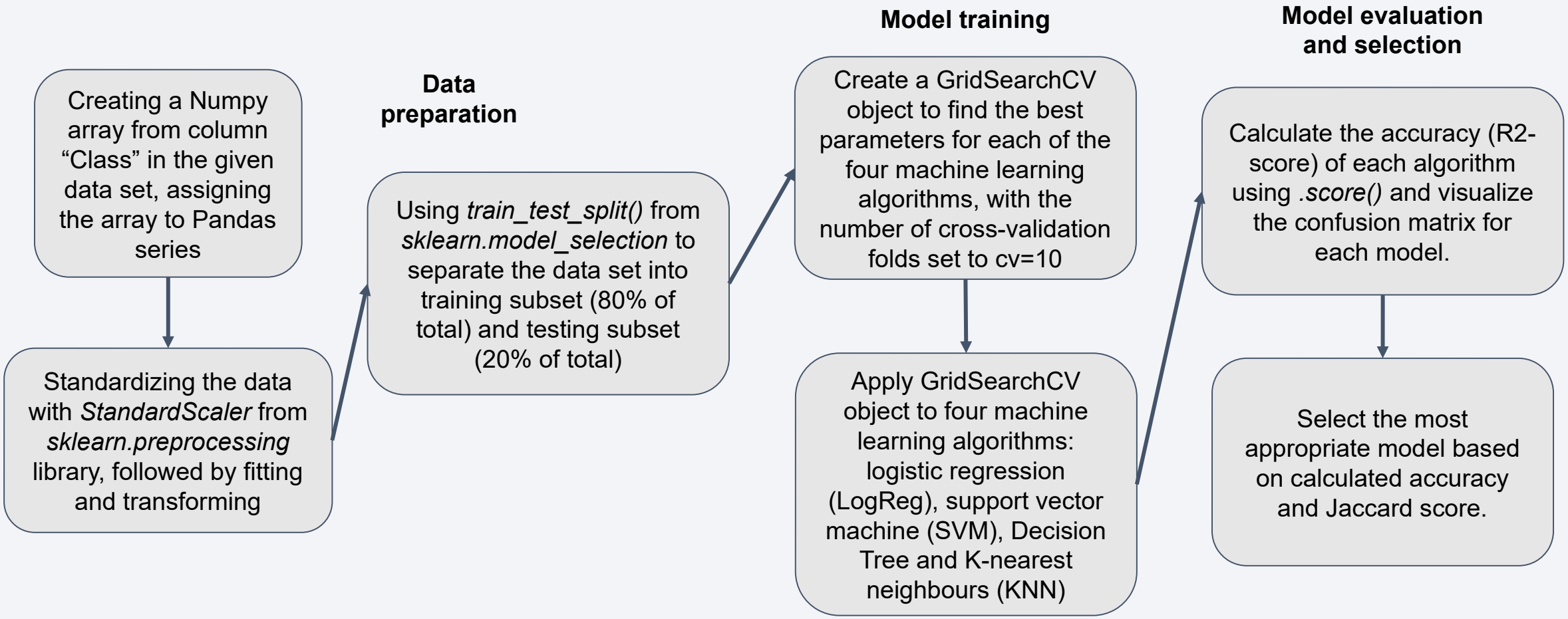
Build a Dashboard with Plotly Dash

To ensure the interactivity in data presentation in Plotly Dash, following was done:

- Addition of dropdown menu to select statistics for individual launch sites or percentage of success rates in total of all launch sides.
- A slidebar for payload mass was made with the ability to select the range of masses.
- Piecharts which display percentages, visually and numerically, of successful and failed outcomes for each launch site individually selected from dropdown menu OR a pie chart which of percentage of successful outcomes of all launch sites, if all are selected in dropdown menu.
- Scatter chart (x-y) of payload mass vs success rate for each individually selected launch site from the dropdown menu, with separate coloring for different booster versions.

URL: GitHub SpaceX Plotly Dash

Predictive Analysis (Classification)



URL: GitHub SpaceX Machine Learning

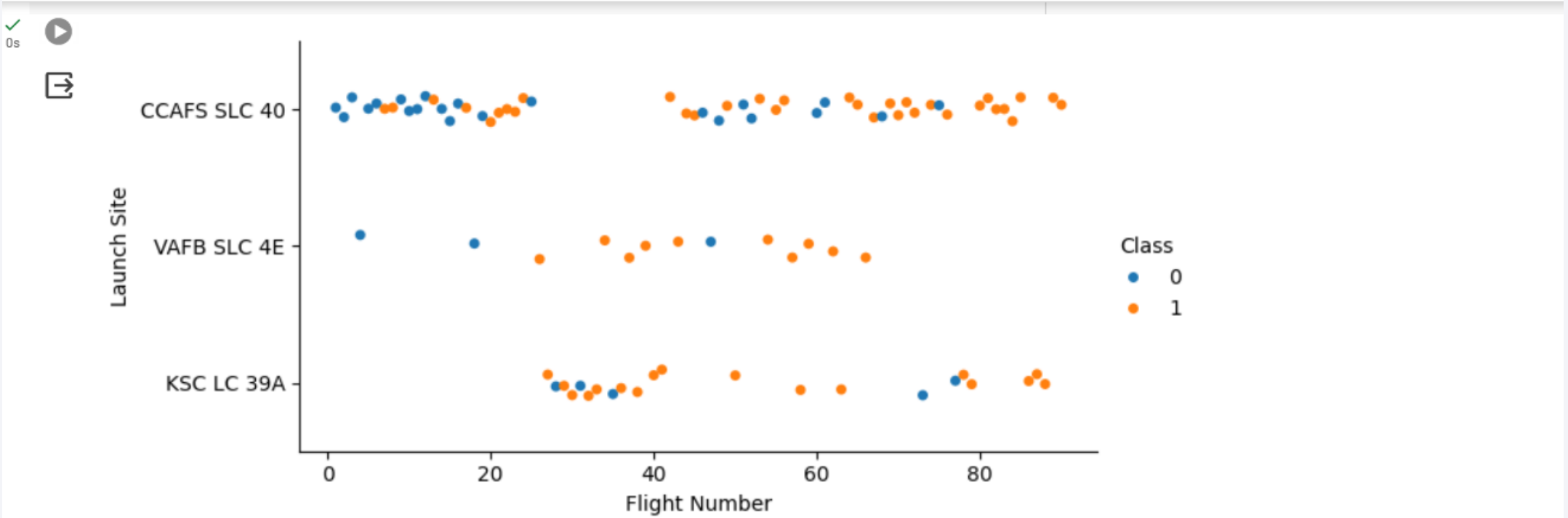
Results

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

Section 2

Insights drawn from EDA

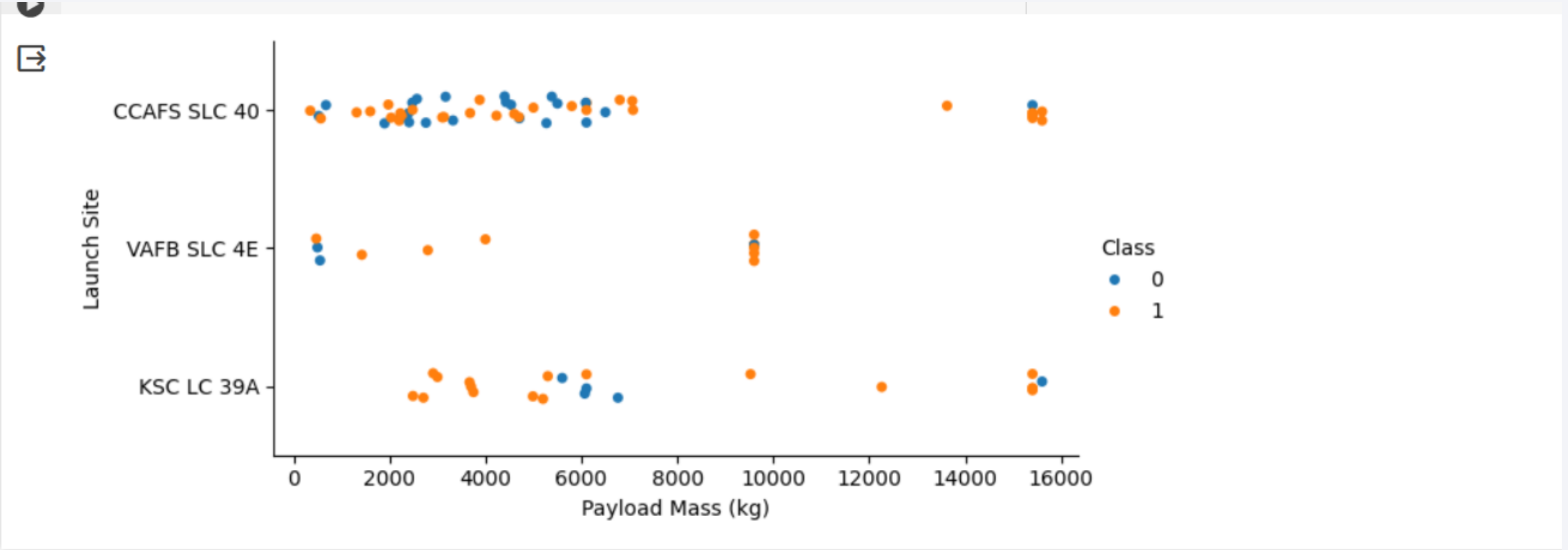
Flight Number vs. Launch Site



Early flights were mostly done at CCAFS site, many of them failed. In time gap that CCAFS site was not used, VAFB and KSC were began to be used and majority were successful at both of these sites. From flight number 40 onwards, all three sites were used, with VAFB no longer used after flight number cca 65.

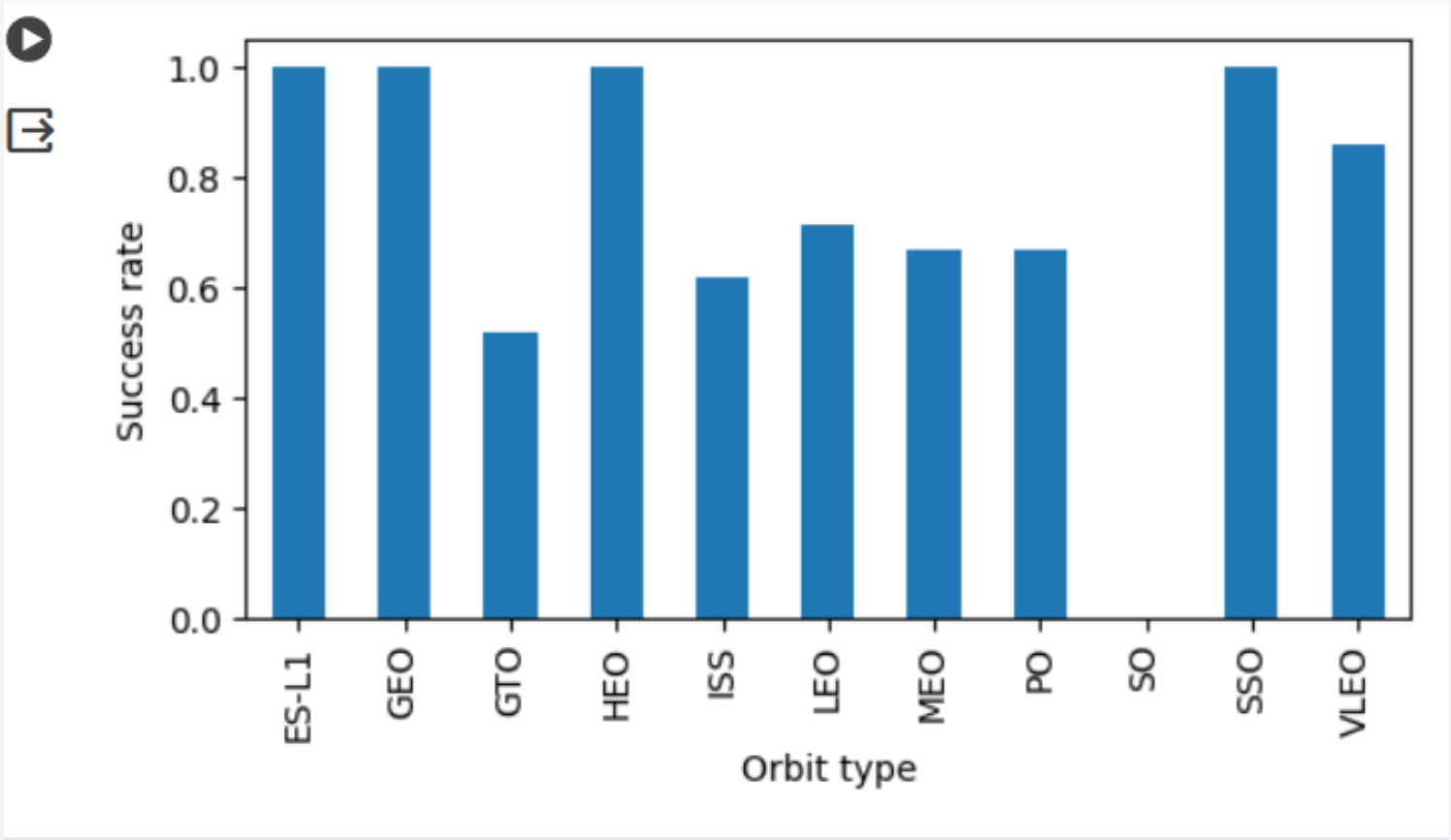
While the number of succsessful launches is increasing with increasing flight number, there is still notable portion of failed ones at CCAFS site. However, the increases in succsess with increasing flight numbers are likely due to increased experience in flight preparation.

Payload vs. Launch Site



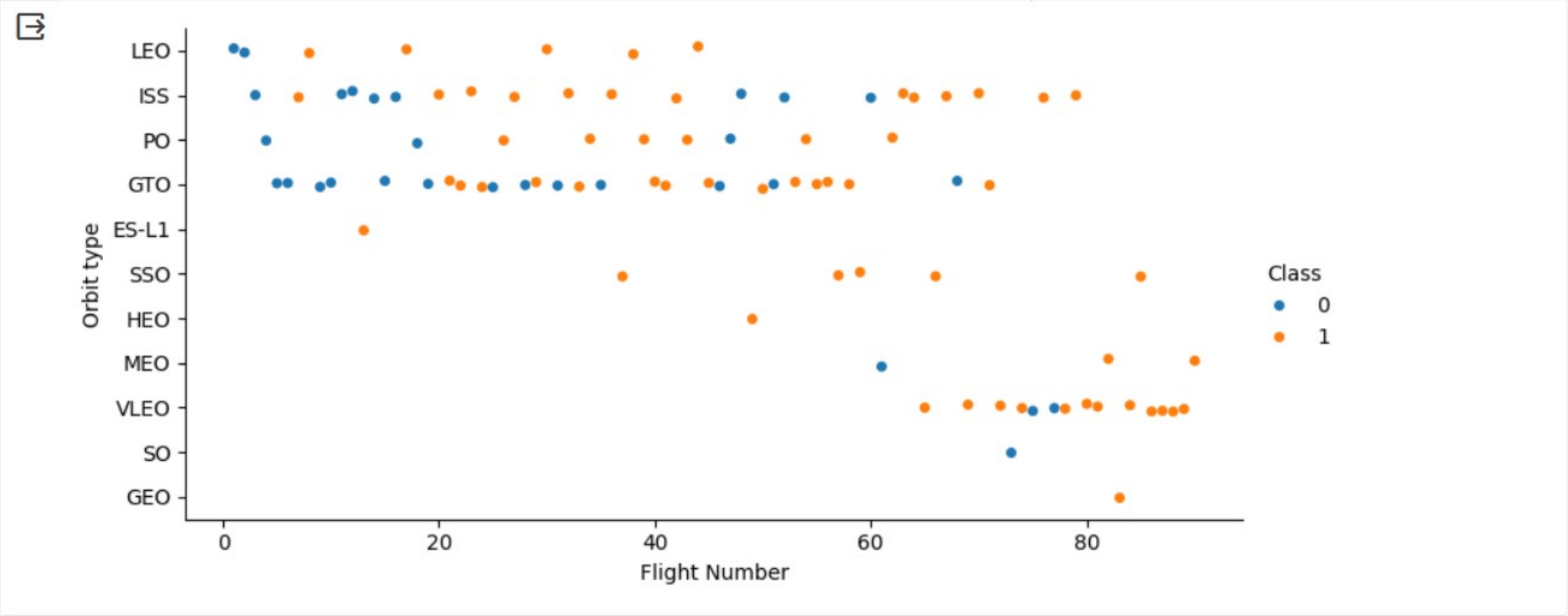
CCAFS site has a variable success with launching of payloads in mass range up to 7000 kg. VAFB site is generally successful with launches but it does so in narrower range, up to 4000 kilograms and, specifically, with payloads just below 10000 kg and it does not go above that limit. KSC has successful launches in payload mass range of 2000 to 5500 kg, but from 5500 to 7000 kg they have failed. CCAFS and KSC sites have been successful launching heavy payloads around 16000 kg.

Success Rate vs. Orbit Type



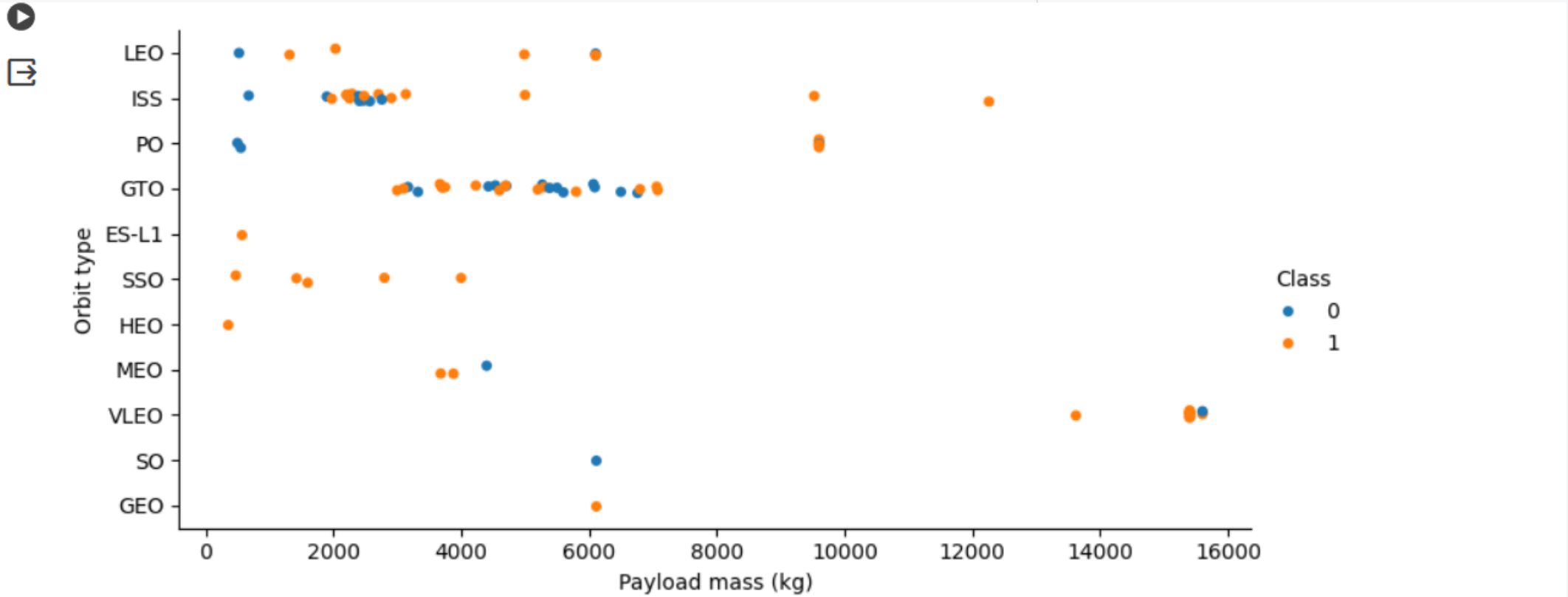
- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with 50%-85% success rate: GTO, ISS, LEO, MEO, PO, VLEO
- Orbits with 0% success rate: SO

Flight Number vs. Orbit Type



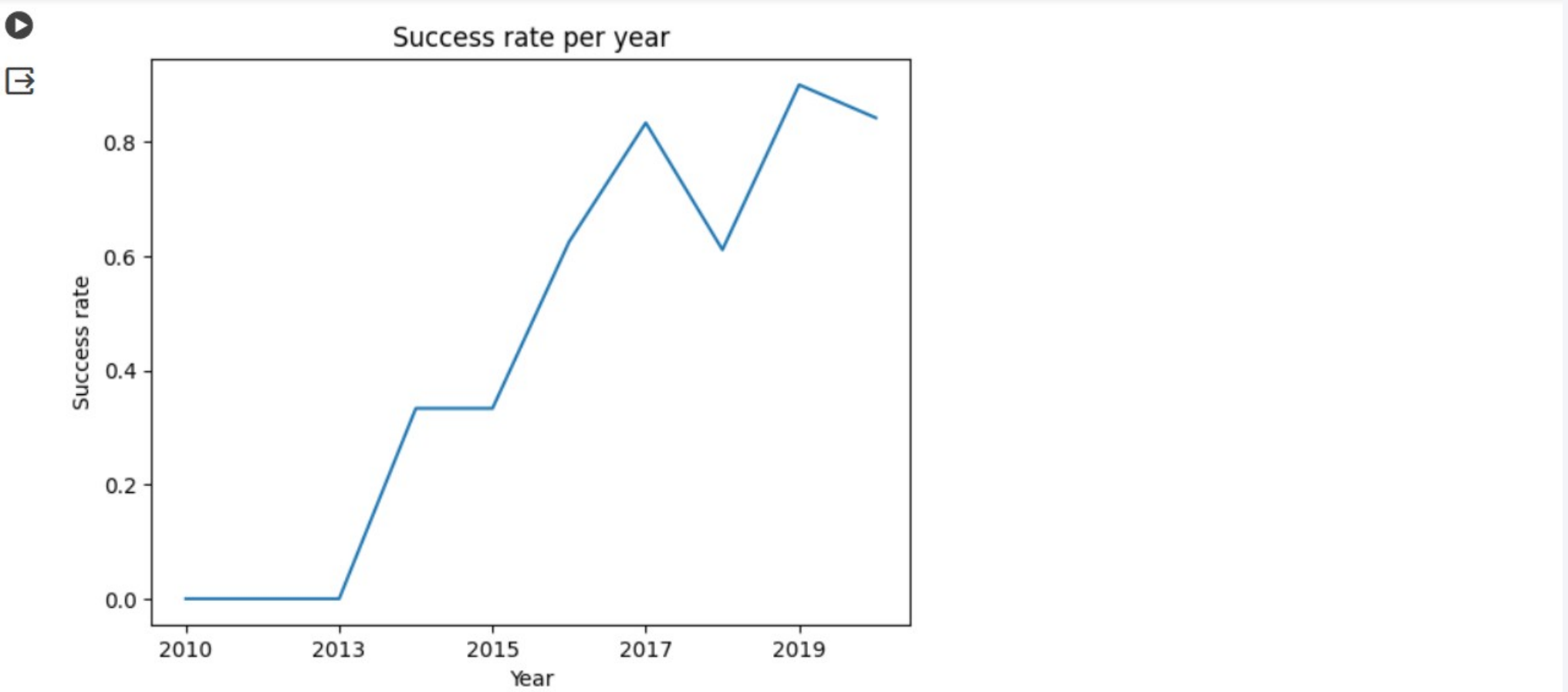
Flights to LEO orbit are almost always successfull, but do not go beyond flight number cca 40. Flights to ISS and PO have generally been successful, but in case of PO they occur less frequently. Flights to GTO occur with variable success/failure with flight number. VLEO flights have been conducted in recents flights only, but majority has been successful. Flight to other orbits occurs infrequently.

Payload vs. Orbit Type



Flights to LEO orbit have had successful landings in payload mass range 1500-6000 kg. Flights to ISS orbit had variable success in payload mass range 2000-4000 kg, just as GTO had in range 2500-7000 kg. SSO has been very successful with flights of payload mass range up to 4000 kg. From around 10000 kg to 16000 kg payload mass range, landings from ISS, PO and and VLEO orbits have been successful, while from other orbits no landings were conducted in that range.

Launch Success Yearly Trend



Average success rate began increasing since 2013 to 2017 with similar trends, with the exception of "stability period" from 2014 to 2015. Aside from a drop in 2018, it seems that general trends on the success rate of mission outcomes is favourable.

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
1 %sql SELECT DISTINCT launch_site FROM SPACEXTBL;
```

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

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

- By introducing DISTINCT clause, SQL is instructed to extract all the launch site entries. They are not listed as they appear, but they are mentioned only once, if different exist.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE "CCA%" LIMIT 5;
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Here, launch sites are selected in the table SPACEXTBL if they satisfy condition to have a sequence of strings similar to (LIKE) “CCA”. The “%” here indicates there might be more string characters following CCA so this should be taken into account. The list is limited to the first 5 entries.

Total Payload Mass

Task 3

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

```
1 %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- Values in PAYLOAD_MASS__KG_ column are summed together, but only those whose another entry in a row (under another column, “Customer”) equals to string sequence “NASA (CRS)”.

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
1 %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = "F9 v1.1";
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Only those entries whose one value in column “Booster_Version” is equal to string sequence “F9 v1.1” are selected, the value from columns PAYLOAD_MASS__KG_ is are taken and the average value is calculated.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
1 %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)";
```

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

MIN(Date)
2015-12-22

- By using MIN function on column “Date”, an entry is selected from if the value of column “Landing_Outcome” is equal to a string sequence “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
1 %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = "Success (drone ship)" AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

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

- Selection of “Booster_Version” entries is dependent on satisfying two conditions:
  1. string sequence in “Landing_Outcome” column must equal “Success (drone ship)”
  2. the numerical value in PAYLOAD_MASS__KG_ has to be in range 4000 and 6000. Same query could have been done by using “greater than” and “less than” operators.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, count(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

- Query is stated in such a way to select all “Mission_Outcome” entries, count them and them group by the same value.

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
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
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Query selects all “Booster_Version” and PAYLOAD_MASS__KG_ entries but conditioned with a subquery which first select maximum values of payload masses from the table.

2015 Launch Records

Task 9

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
SELECT SUBSTR(Date, 6,2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE SUBSTR(Date,0,5)="2015" AND Landing_Outcome = "Failure (drone ship)";
```

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

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

- SELECT query is modified to extract a month from full date entry, with WHERE clause also have a modification to extract entries from year 2015 and with “Landing_Outcome” condition as stated

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

Task 10

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.

```
[ ]  1 %%sql
     2 SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20"
     3 GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC;
```

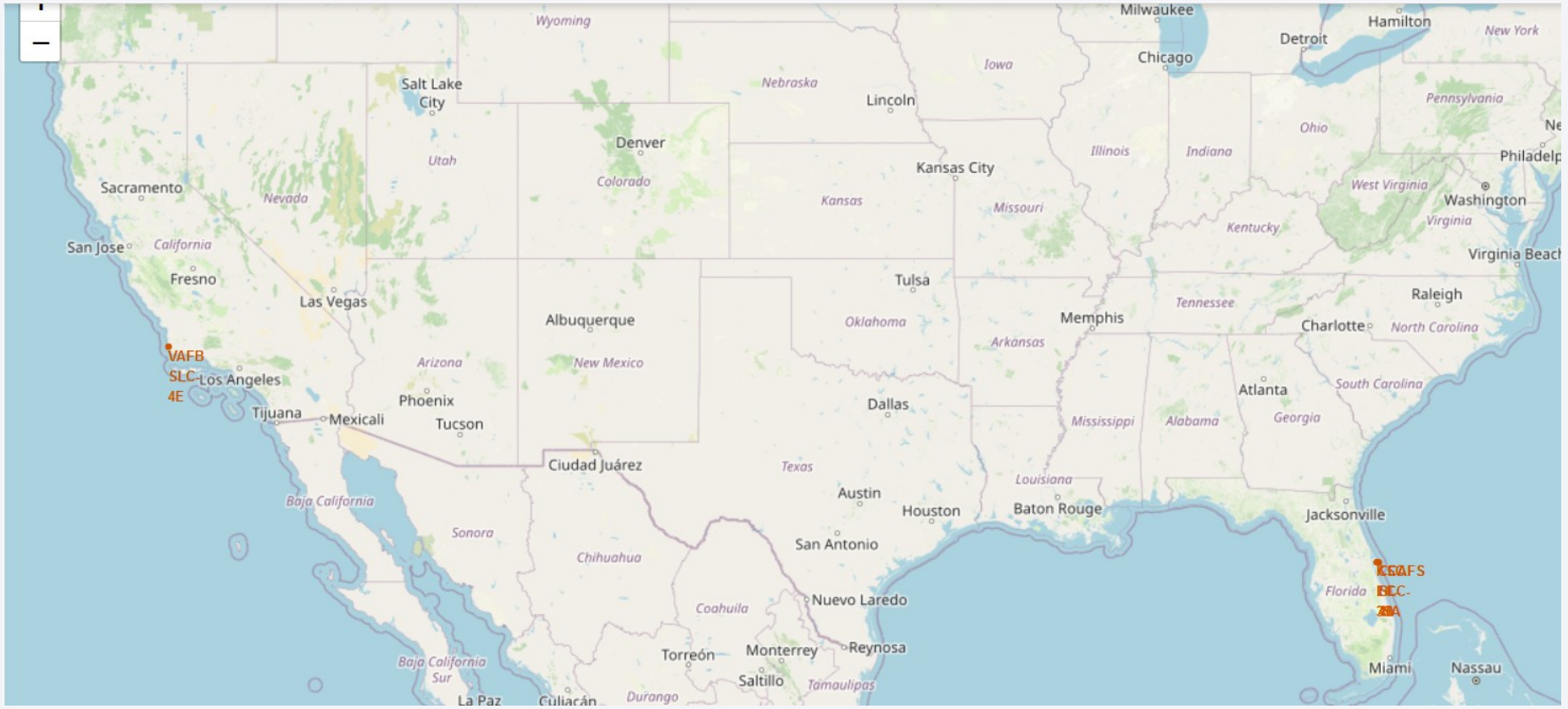
```
 * sqlite:///my_data1.db
Done.
```

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

- Query is conditioned with a date range, where dates are input as string sequences. This is followed by grouping together same landing outcomes, ordered by descending number of occurrences for each. In the give time period, the largest number of landings was not even attempted, as seen from the result


Section 3
Launch Sites
Proximities Analysis

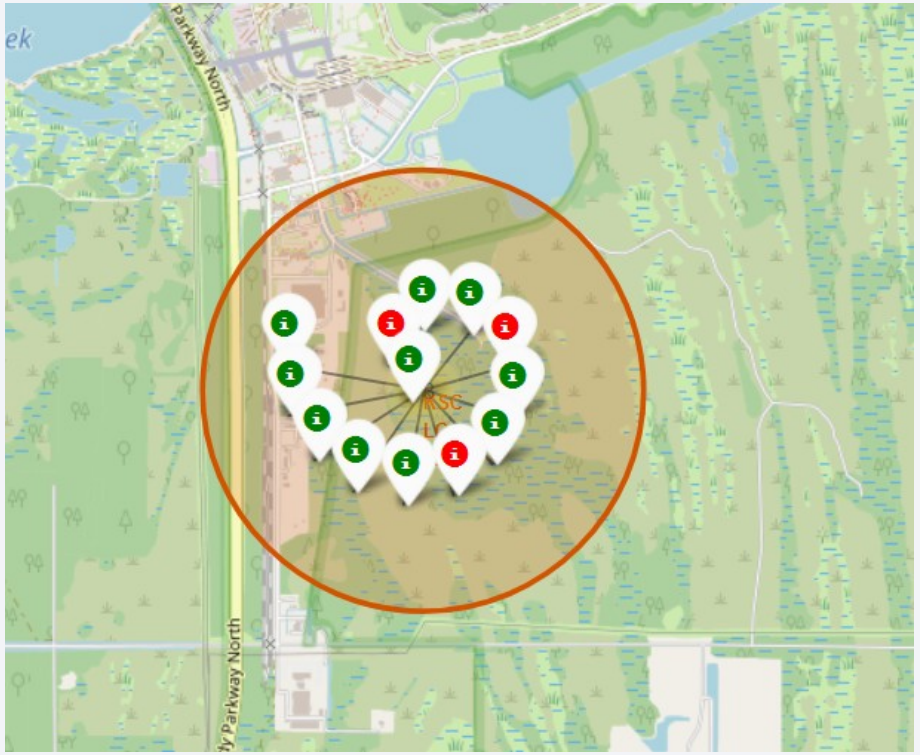
Locations of SpaceX launching sites

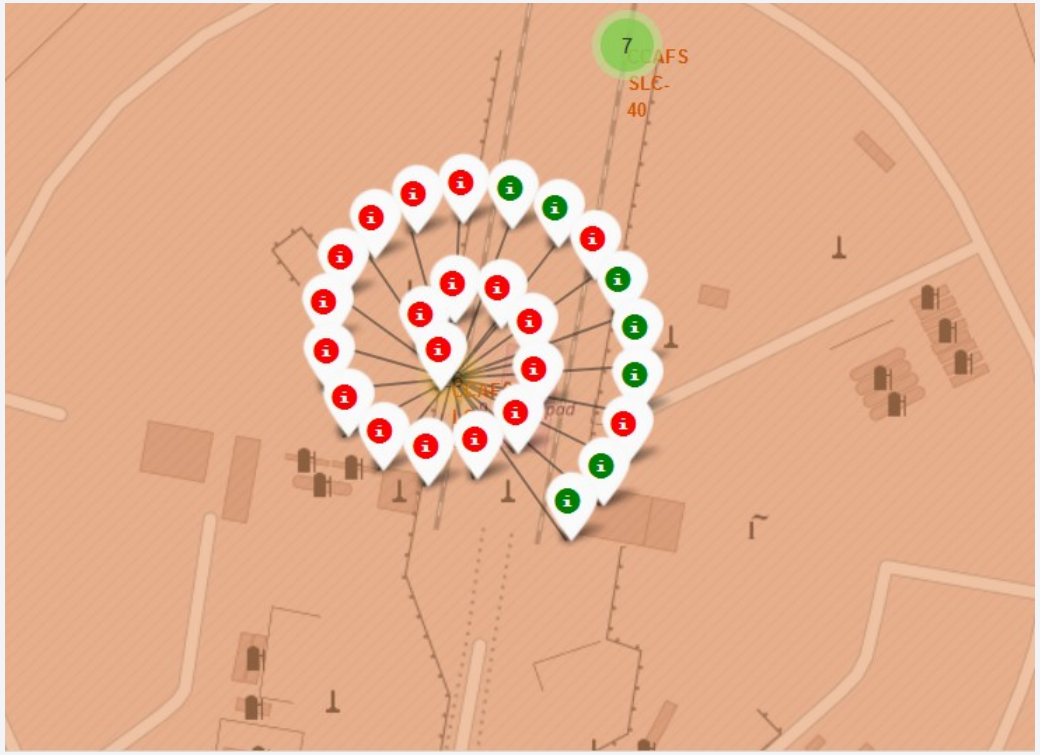


All four launching sites are located at the coastline. This has few benefits: retrieval of the stages is easier via ocean and any disaster happens outside of urban, populated areas.

Three out of four are in south of Florida, closer to the equator. This is due to the fact that all objects closer to the equator rotate around the Earth's center of gravity at higher velocities. This velocity gain is used superimposed (added to) the launch speed of the rocket modules.

Colored launch outcomes

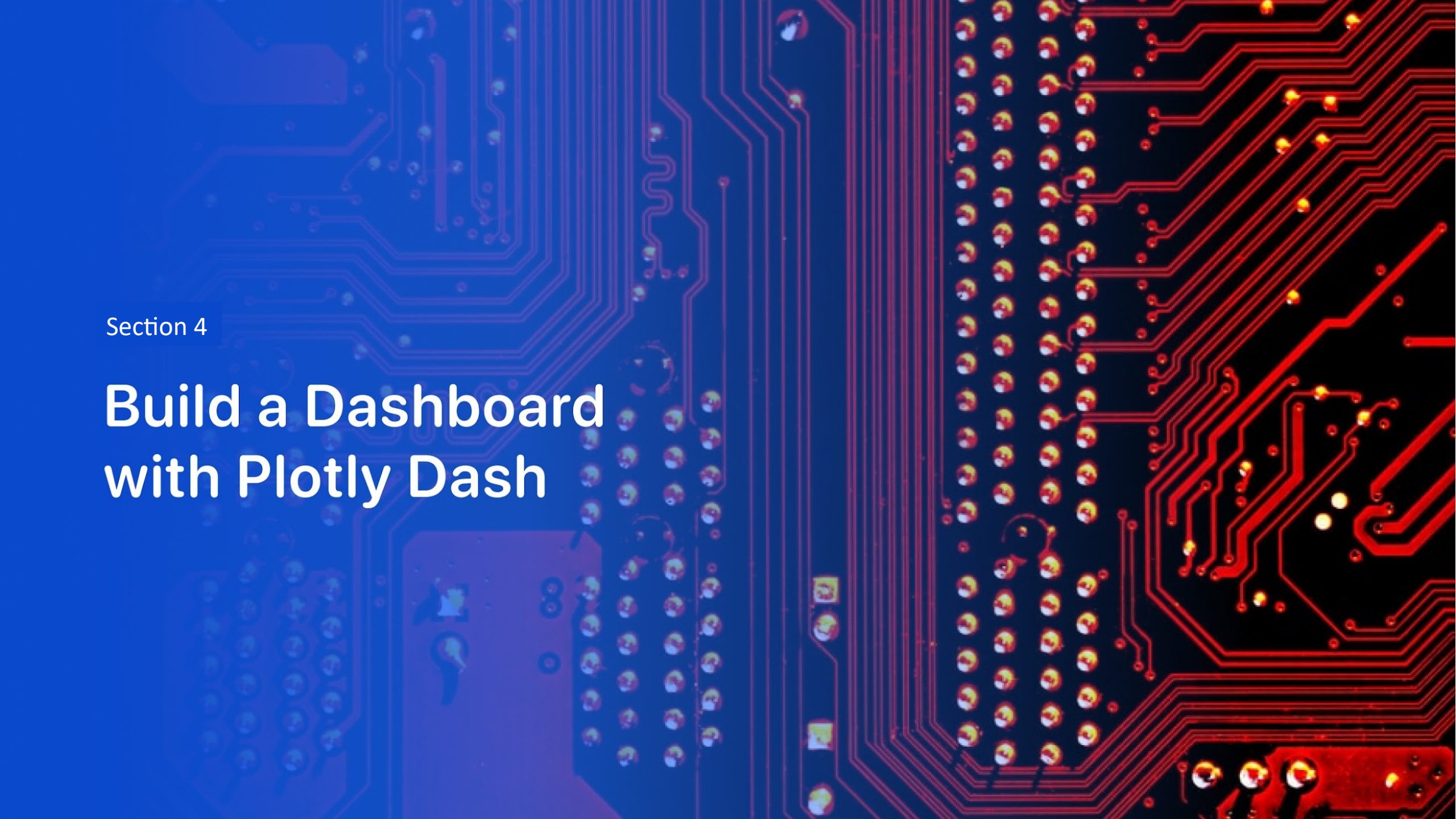




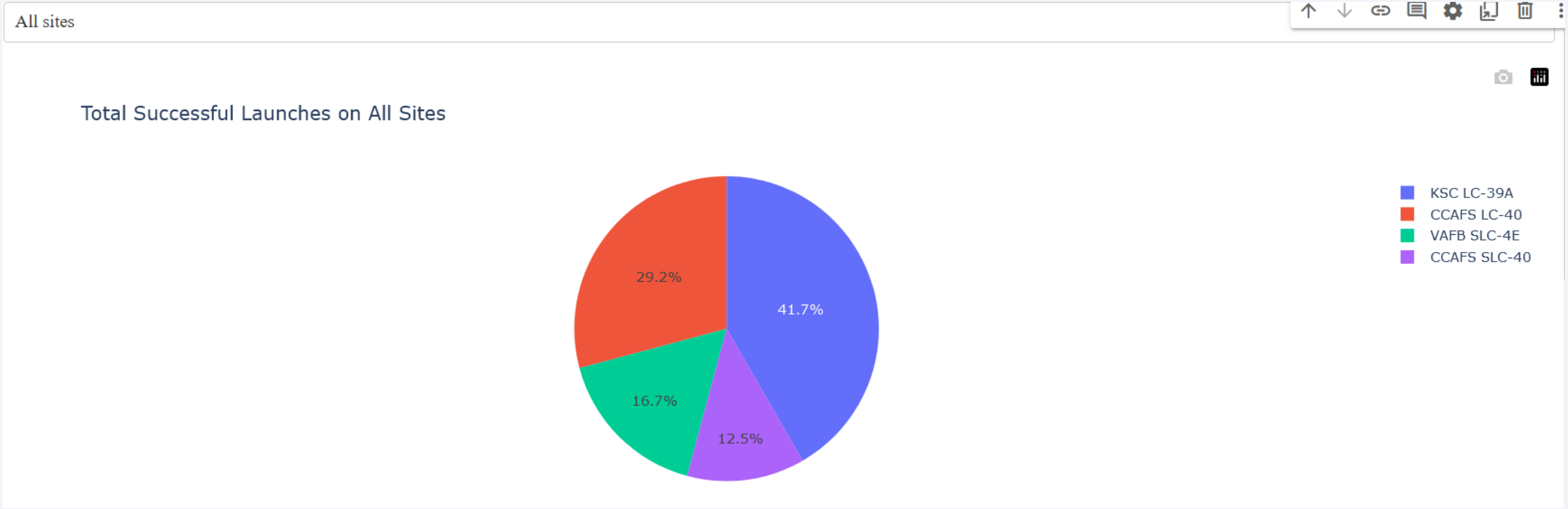
- Green and red markers represent successful and failed mission outcomes respectively.
- KSC LC-39, in the left picture, has the highest success rate of mission outcomes.
- CCAFS LC-40 in the right picture has the worst success rate.

Proximities of CCAFS LC-40

- Blue line shows the proximity from CCAFS LC-40 launch site to the nearest coastal point, distance being 0.86 kilometers.

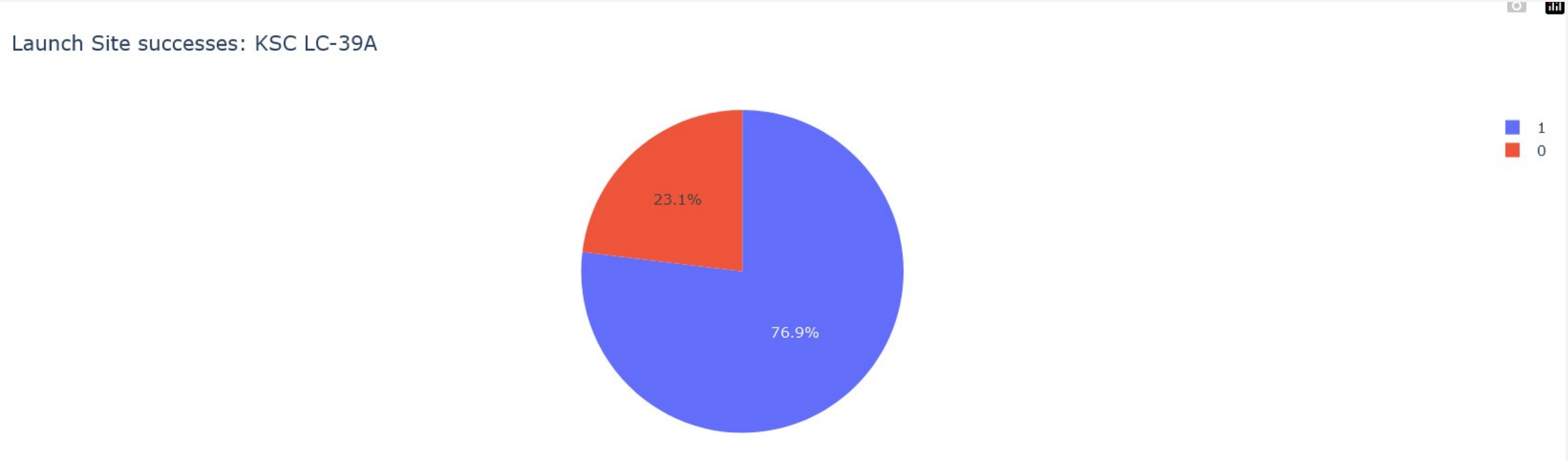

Section 4
Build a Dashboard with Plotly Dash

Success Percentages per Launch Site



The piechart depicts percentage per launch site of total successful launches. As indicated in the slide “Colored launch outcomes” (slide 36), launch site KSC LC-39A is leading with the number of successful launches with 42%, while CCAFS SLC-40 has the lowest percentage of them.

Launch Site Statistics for KSC LC-39A



Although KSC LC-39A has the highest percentage of successful launches among all the launch sites, its individual statistics should be of concern. Of all launches at this site, a little under a quarter of them (23.1%) are not successful. This means that 1 in 4 launches will is likely to fail.

Payload vs Launch Outcome Stats 1 - lower payloads

The scatter plot depicts success/failure of launches in dependence of payload masses in the lower range (up to 4000 kg) for different boosters. We can see that Falcon 9 v1.1 (orange) has high rate of failures, while Falcon 9 v1.2 (FT, green) performs very well. Other boosters are not used as much in this payload range.

Payload vs Launch Outcome Stats 2 - heavy payloads

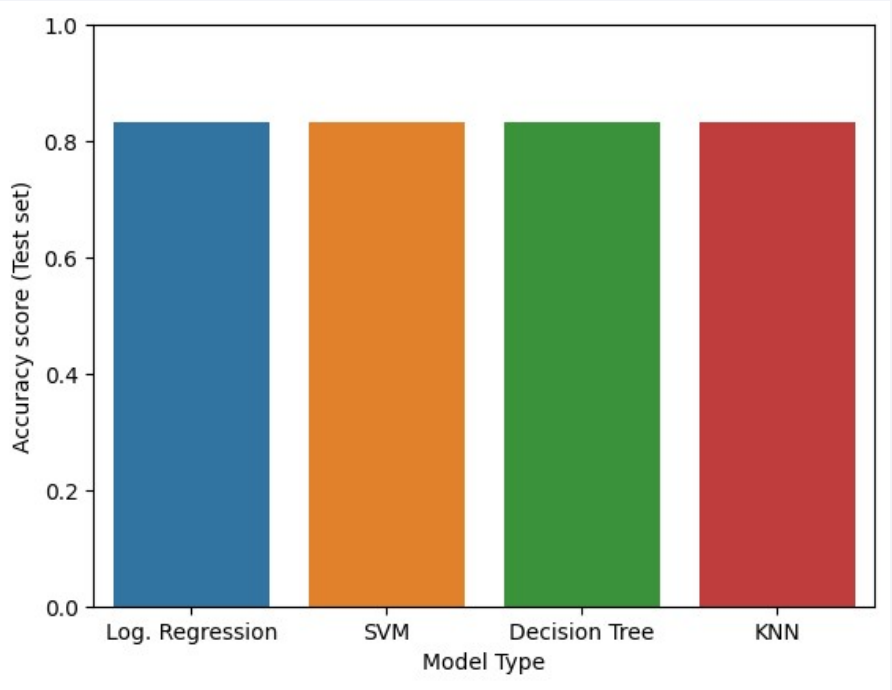
In the heavy range (4000-10000 kg), two boosters, v1.0 and B5, are not used at all. The success rate of FT booster seems to drop roughly above 6500 kg limit, while B4 booster has variably success in the upper limit, at 10000 kg.

Section 5

Predictive Analysis (Classification)

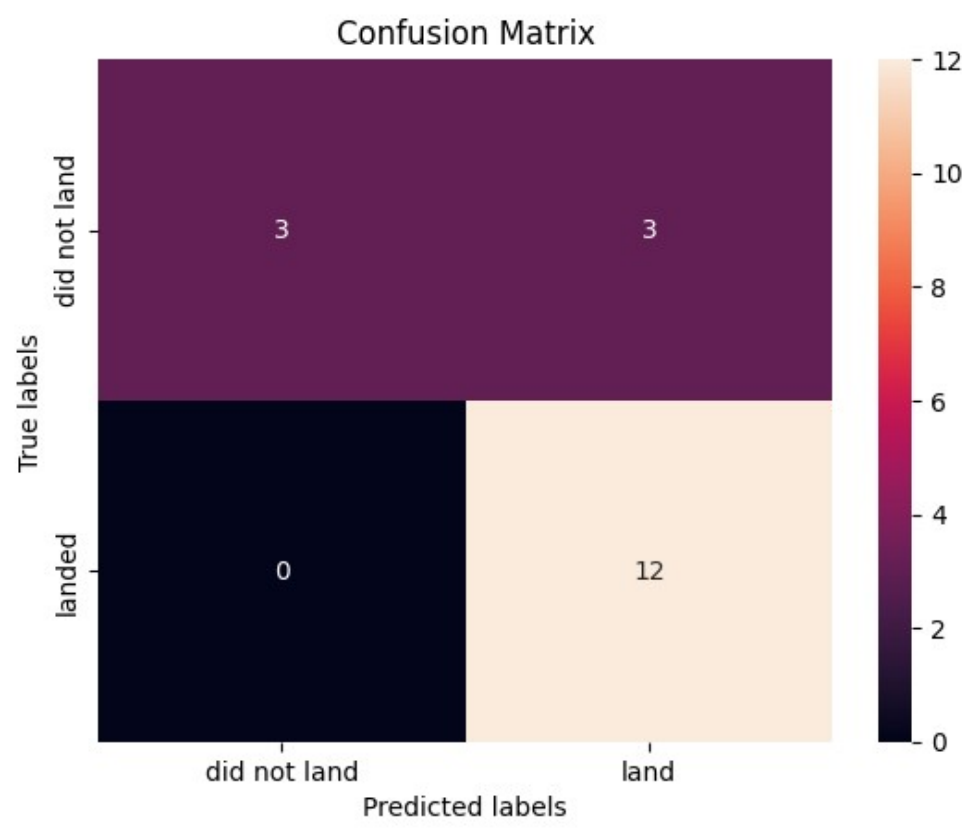
Classification Accuracy

	Model Type	Accuracy score (Train set)	Accuracy score (Test set)	Jaccard score	F1 score
0	Log. Regression	0.846429	0.833333	0.8	0.888889
1	SVM	0.848214	0.833333	0.8	0.888889
2	Decision Tree	0.875000	0.833333	0.8	0.888889
3	KNN	0.848214	0.833333	0.8	0.888889



- As shown in the accuracy (R2 score) and Jaccard score values, all four modules have performed reasonably well when the mode accuracy score was estimated for the train sets. Decision Tree has a slightly higher value. All other metrics are also identical.
- It has to be noted that, due to some code deprecations (as seen in code output), estimations of metrics for Decision Tree model were not always reproducible in terms of accuracy score for test set, Jaccard score and F1 score.
- Thus it remains inconclusive which of the four models would be the best for use. Given that, even with the issue of aforementioned reproducibility, Decision Tree model has slightly higher accuracy score on train set, its confusion matrix will be discussed in the next slide.

Confusion Matrix - Decision Tree



- Confusion matrix for all four methods were the same (see ipynb file). Prediction was made on 18 samples.
- 15 of 18 were correctly predicted: 12 predictions for successful landing when landing *has* occurred, 3 prediction of failed landing when landing *has not* occurred.
- There have been 3 predictions of successful landing when such *did not occur*, leading to a false positive.
- There have been no false negatives.

Conclusions - EDA

- Two methods of data collection presented, via SpaceX API and web scrapping from Wikipedia, after which data wrangling was performed as described in flowcharts.
- Explorative data analysis (EDA) was performed with Python visualization libraries and SQL queries, as described in the flowcharts.
- From EDA with data visualization, the following has been noted:
  - General trend of rising success rate of SpaceX missions with time passing, as seen from “Flight Number vs Launch Site” (sl. 18) and “Annual average success rate) (sl. 23)
  - Higher success rates are inclined towards higher payload masses (“Payload vs Launch Site”, sl. 19)
  - Missions are most successful to four orbits: ES-L1, GEO, HEO, SSO (sl. 20)
- From EDA with SQL, the following has been noted:
  - Four distinct launch sites are used by Space X: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E (sl. 24)
  - Earliest successful landing on the ground occurred as early as 2015 (sl. 28)

Conclusions - geoanalysis and summary thus far

- From geoanalysis in Folium, it has been noted:
  - 3 of 4 launch sites are in Florida (sl. 35), closer to the equator. This is explained by basic physical principles - rotation of bodies on Earth's surface is closer to the equator and this velocity superimposes on the initial speed of launched rockets, providing a greater escape velocity for particular orbit type (sl. 35).
  - Florida sites are all easily reachable by sea, highway and railway traffic systems, providing easier transport of reusable rocket stage upon its retrieval.
- From analysis in Plotly Dash, it has been noted:
  - Launch site KSC LC-39A has the largest proportion of successful launches among the used sites. Still, 1 in 4 launches may fail.
  - Booster Falcon 9 v1.2 FT has very good performance across wide range of payload masses (sl. 39-42)
- **In summary thus far**, most optimal conditions for SpaceX seem to be launching from KSC LC-39A site with booster Falcon 9 FT, with payloads 2000 - 6500 kg (with inclination towards even higher upper limits) into higher orbits.

Conclusions - machine learning

- From predictive analysis, it is indicative that all four ML models (Logistic Regression, SVM, Decision Tree, KNN) are of similar accuracy in prediction.
- Some discrepancies in the reproducibility of results have been seen Decision Tree due to the code deprecation.
- That issue aside, test subset of the data was, in absolute size, only 18 samples which may be insufficient for testing. Two recommendations are proposed:
  - redefine initial train/test ratios
  - acquire a larger initial, total dataset (stronger recommendation)

Appendix

In the very end, I would like to express a big thanks to both Coursera and IBM Skills Network staff for providing educational and technical help throughout the IBM Data Science specialization programme.

Particular thanks goes to the instructors who readily replied to my inquiries on discussion forums. Special thanks goes to Glenn from Coursera Support Team who patiently helped me with SN Labs technical issues during the “Data Visualization with Python” (Course 8 of 10 in the programme).

Thank you!