IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: through both APIs and web-scraping techniques
  - Data Wrangling
  - Exploratory Data Analysis: through SQL & data visualization
  - Interactive geospatial analytics using Folium
  - Machine learning for predictive classification
- Summary of all results
  - Exploratory data analysis insights & screenshots
  - Geospatial analysis insights & screenshots
  - Predictive classification results & conclusions

Introduction

- Project background and context
  - While competing companies advertise Falcon 9 rocket launches at a price point upwards of $160 million, Space X offers a far cheaper option costing only 62 million. They are able to do this as the first stage of their builds are reusable from launch to launch, cutting down operation costs dramatically. However, this reusability relies on extra factors, and is not a guarantee. As a result, if the return and thus the reusability of the first stage can be predicted accurately, Space X may alter factors in their prep to guarantee more future success, and other companies may also use these techniques to aid them in bidding wars over rocket launches.
- Problems you want to find answers
  - What are the most important variables when determining the success or failure of an individual rocket launch?
  - Are there geospatial conditions that may affect the success or failure of a rocket launch?
  - What is the best and most practical machine learning technique to make these predictions?

Section 1

Methodology

Methodology

Executive Summary

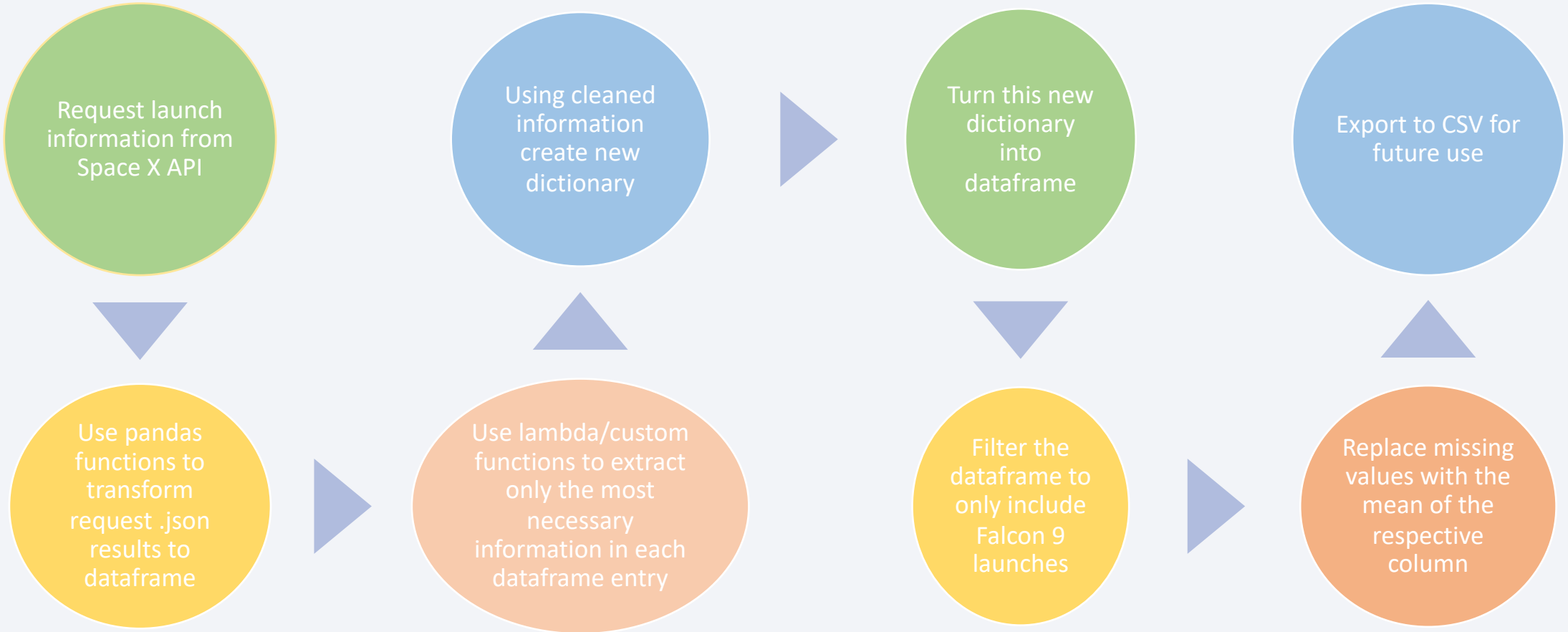
- Data collection methodology:
  - The data was collected first using the Space X API, and then additional information was gathered through traditional web scraping techniques via Wikipedia
- Perform data wrangling
  - One hot encoding: categorical landing outcome variable was transformed into 1,0 values to prepare for future machine learning and analytics
- 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

Data Collection

- The process of collecting information for this project involved the use of both Space X's REST API as well as traditional web scraping techniques that utilized relevant Wikipedia pages. Without the use of both sources, the data used would have been insufficient for the tasks at hand.
- The resulting tables were able to be combined using the reference key of flight number
  - Information Collected via API: flight number, date, booster version, payload mass, latitude and longitude of the launch site, the launch site name, serial numbers, the reuse count for the first stage of the launch, the number of flights, gridfins, legs, landing pad, block status, and the outcome of each individual launch.
  - Information collected via web scraping: flight number, date / time, payload, payload mass, orbit, customer, version booster, booster landing, and launch outcome.

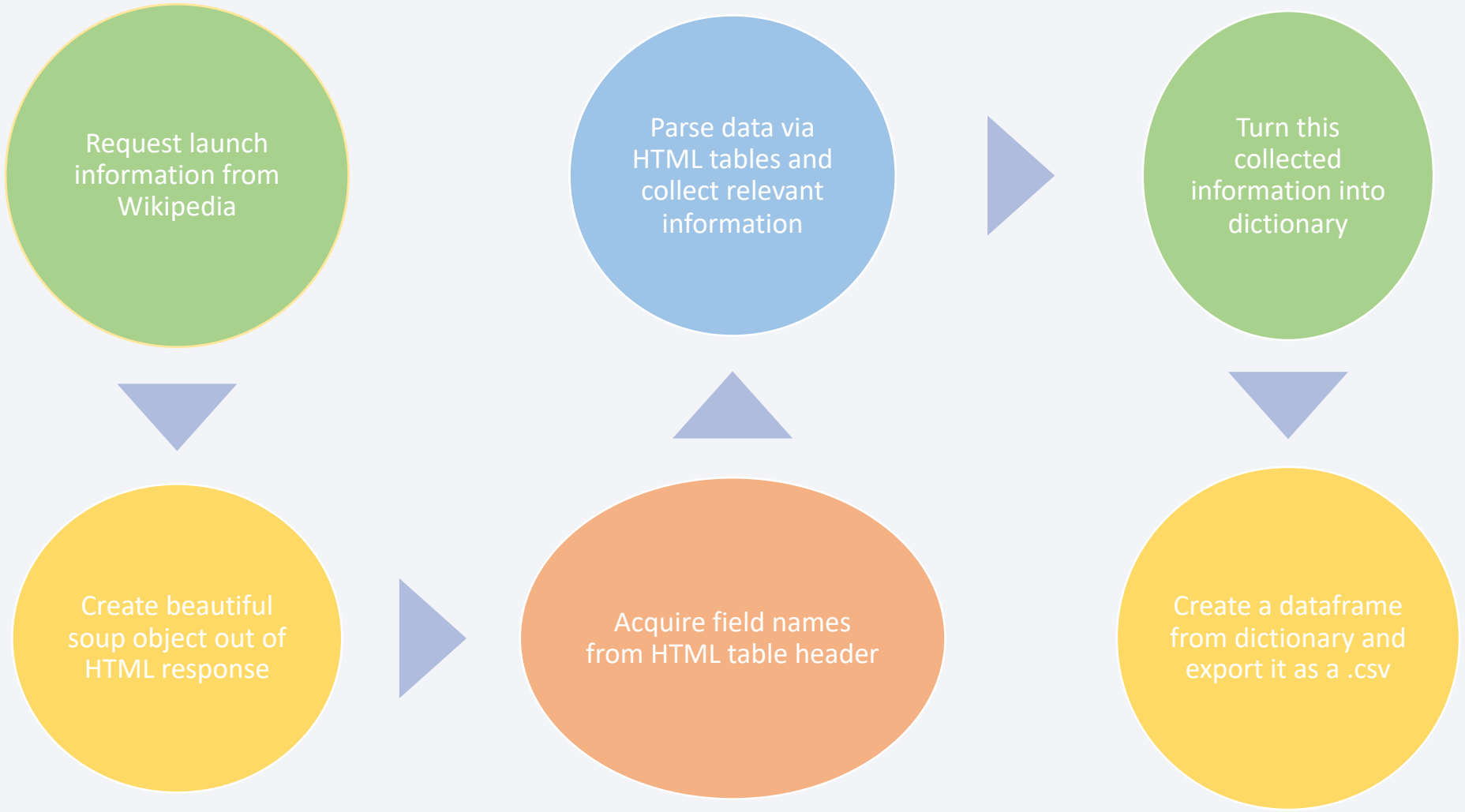
Data Collection – SpaceX API

Link to Notebook



Data Collection - Scraping

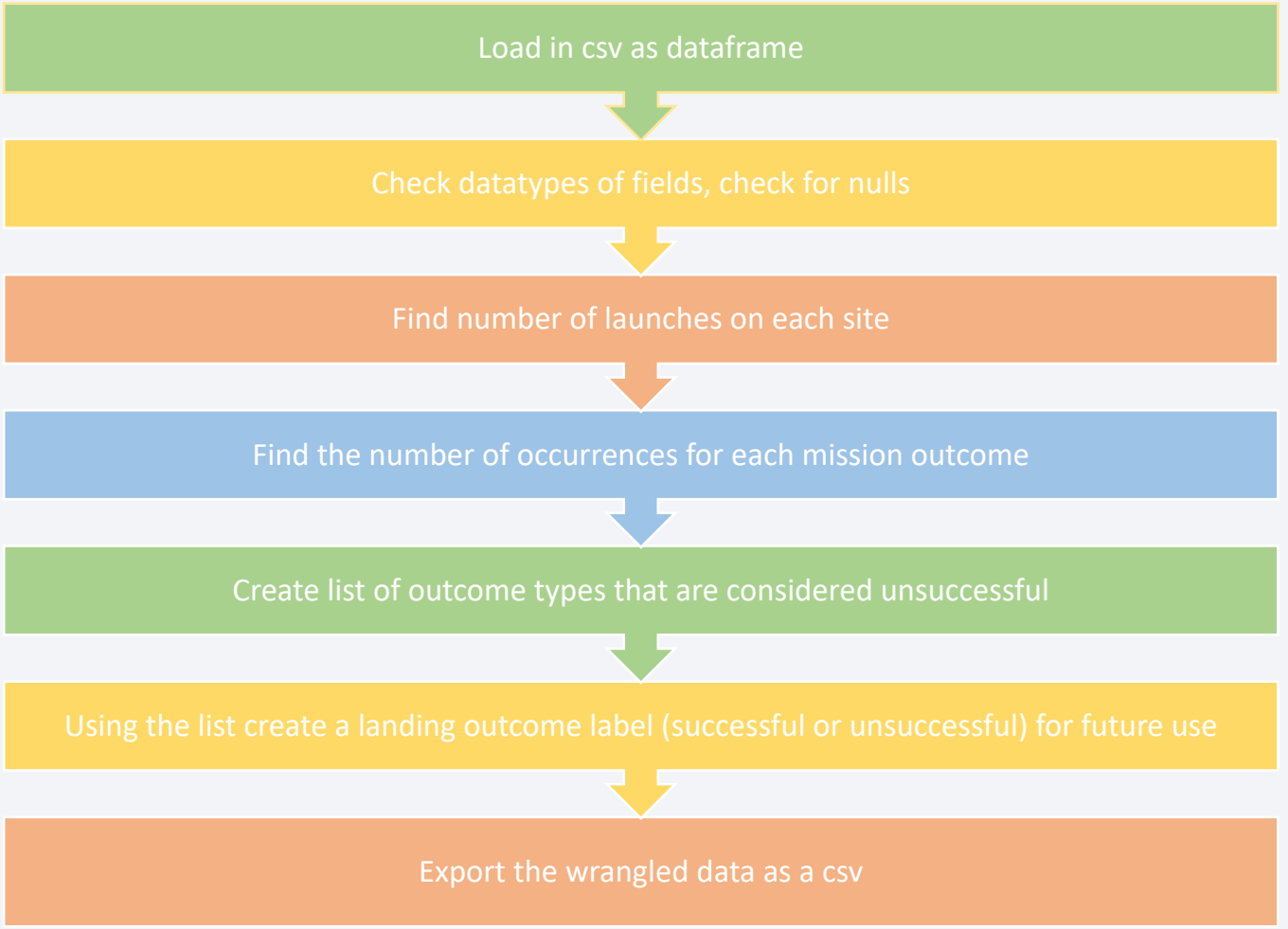
Link to Notebook



Data Wrangling

Link to Notebook

To properly prepare for both analysis and machine learning, some changes needed to be made to the datasets. A key alteration was seen in the mission outcome field. Here there were multiple different types of mission outcomes pertaining to whether it was successful or not as a whole, followed by the manner this occurred (ocean return for instance). To simplify this, a training label was created to show simply if it succeeded or not, which would allow a machine learning process to determine between only one and zero, instead of seven or eight possible outcomes.



EDA with Data Visualization

Notebook Link

- Scatter plots were used to determine if there was any general relationship between variables
- Bar charts were used to determine if any factor on average showed a correlation to the target variable
- Line charts were used to show the relationship between time passing and the target variable

- Flight Number vs. Payload - Scatter
- Flight Number vs. Launch Site - Scatter
- Payload Mass (kg) vs. Launch Site - Scatter
- Orbit vs. Mean Success – Bar Chart
- Flight Number vs. Orbit & Success – Scatter (Hued)
- Payload vs. Orbit and Success – Scatter (Hued)
- Date vs. Success Rate - Line

EDA with SQL

Notebook Link

The SQL Queries Performed:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranked 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.

Build an Interactive Map with Folium

Notebook Link

Markers were placed using launch sites latitude and longitude data, the following specifics allow for further understanding to be developed as to how or why certain areas may have more success

- Used Circle, Popup, and Text Labels to mark all launch sites, also including the NASA Johnson Space Center
- Each launch site had a cluster of smaller markers to visualize their success rates as individual locations
- Colored lines were drawn from locations to notable points of interest – coasts and local infrastructure that could aid/impact the success of a launch site

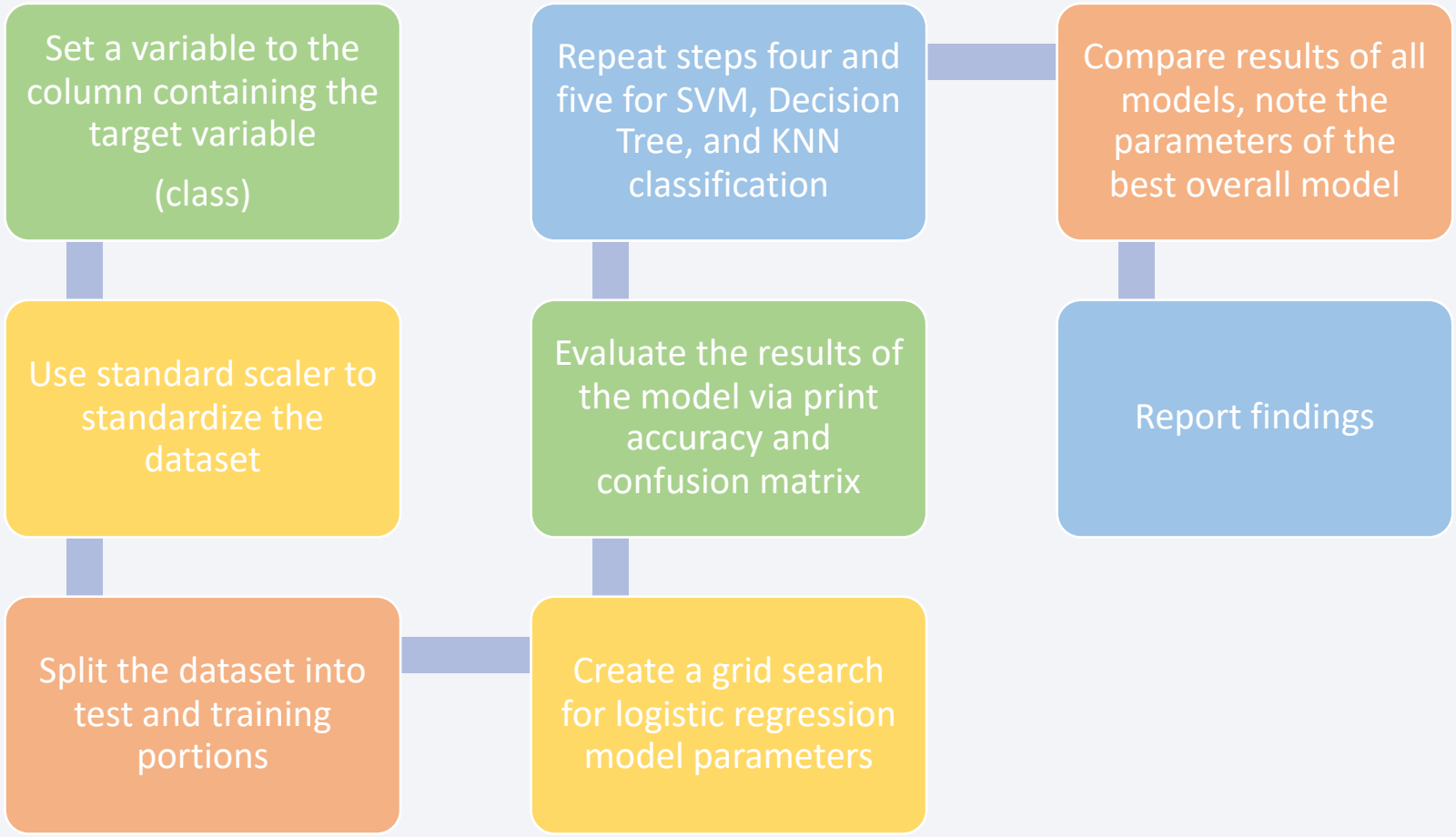
Build a Dashboard with Plotly Dash

Notebook Link

- Interaction Provided:
  - Dropdown menu allowing the user to select a launch site
  - Slider allowing the user to select the payload range for the launch sites they selected
- Plots Provided:
  - Pie chart that shows the number of total launches that were successful for the selected launch site
  - Scatter chart that shows the relationship between payload mass vs. success rate for different booster versions

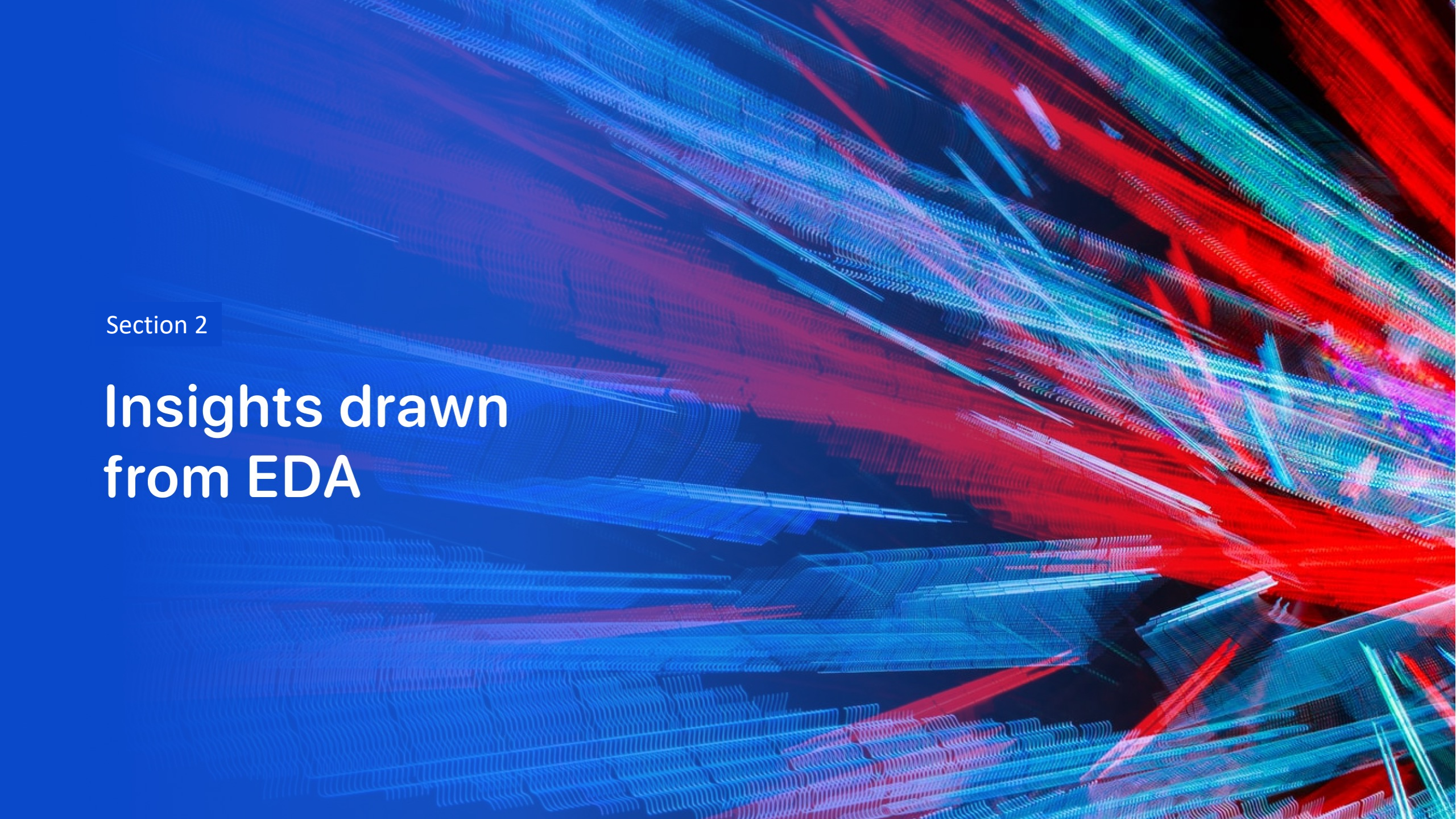
Predictive Analysis (Classification)

Notebook Link

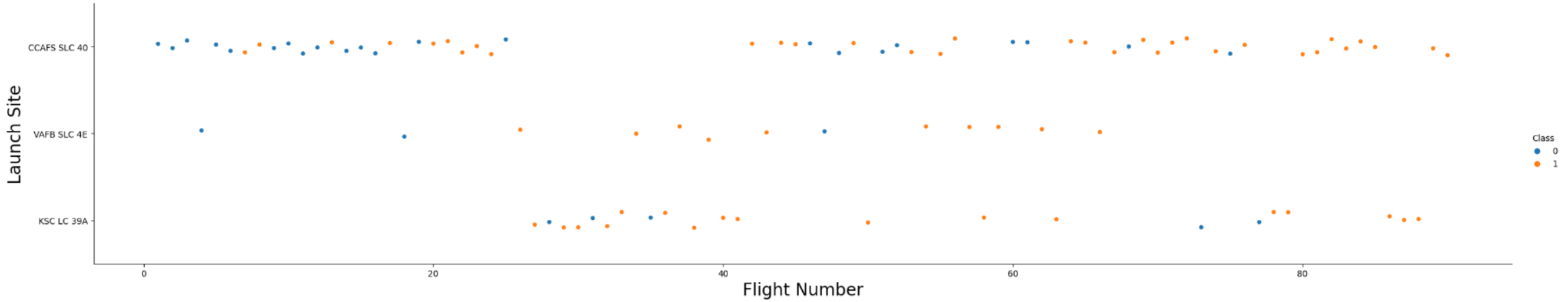


Results

- The results of the exploradtory data analysis will be displayed
- Interactive analytics will be demoed in screenshots
- Predictive analysis results will be displayed

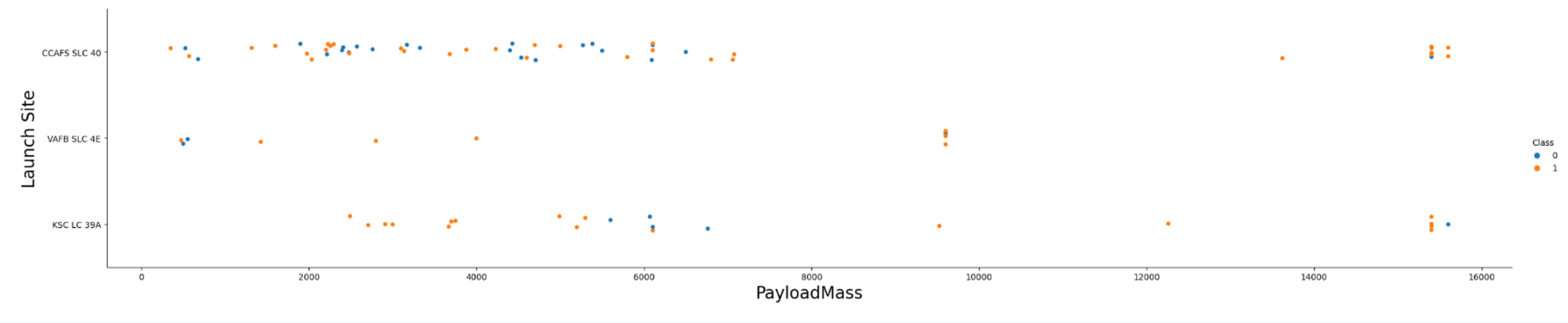

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Given that flight number represents how late in the launch period the flight was, we can see that later flights tend to be more successful
- CCAFS launch site seems to have the most launches, while operating primarily in the later portion of the process
- KSC LC 39A appears to have the highest success rate – it also starts the latest

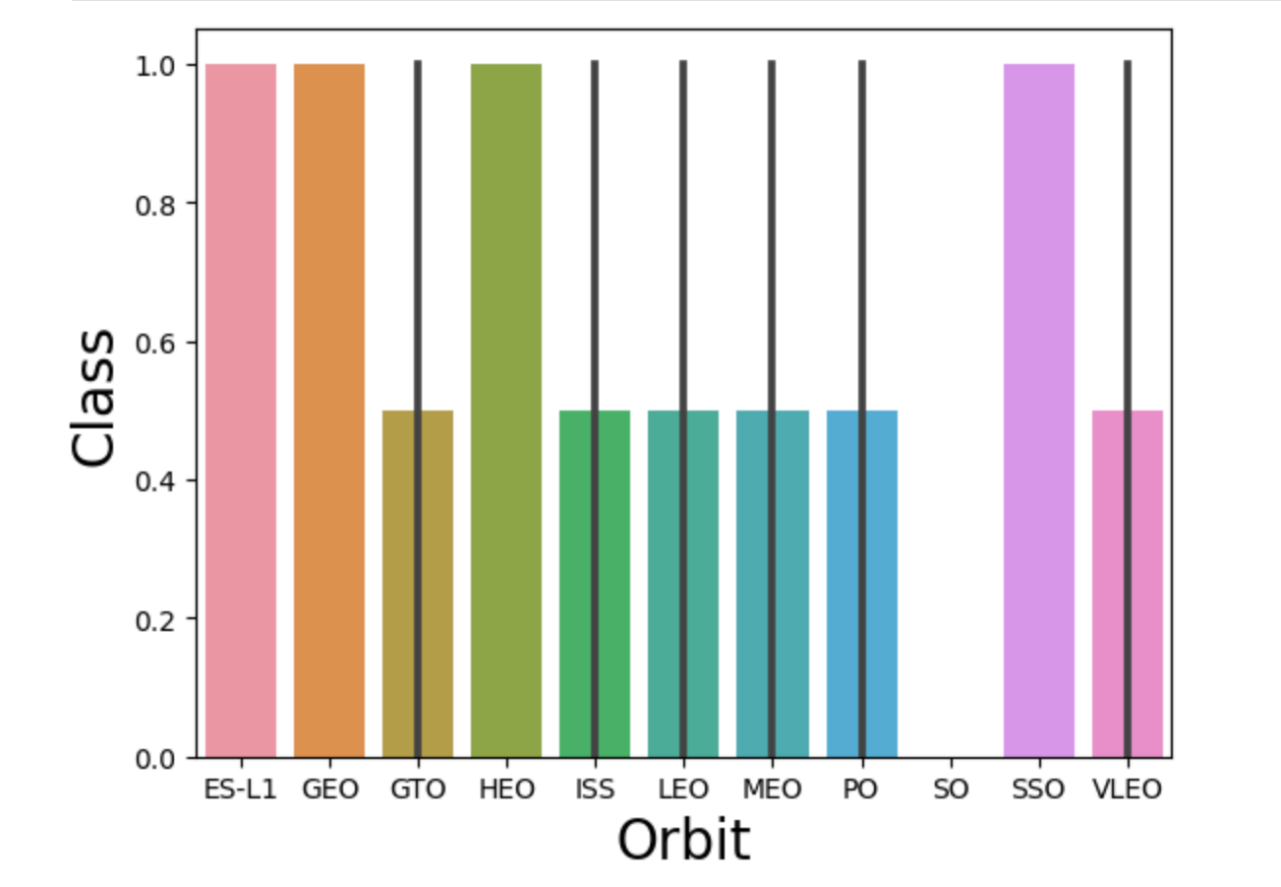
Payload vs. Launch Site



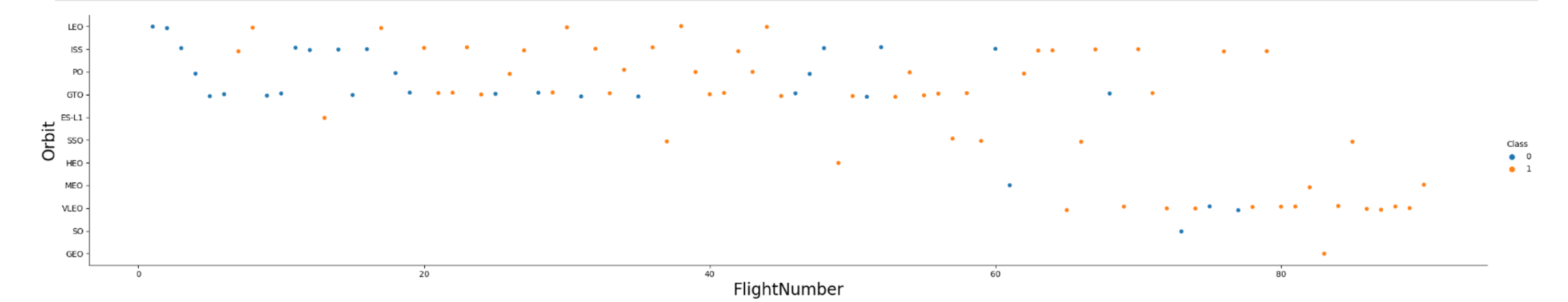
- Higher failure seem to occur with lower payload masses, and the opposite with much higher payloads
- CCAFS and KSC launch sites can take on much greater payload masses
- KSC has a perfect success rate under 5500ish kg

Success Rate vs. Orbit Type

- Four orbits have a 100% success rate
  - ES-L1, GEO, HEO, SSO
- Five orbits have a 50% success rate
  - GTO, ISS, LEO, MEO, PO VLEO
- SO orbit had a 0% success rate

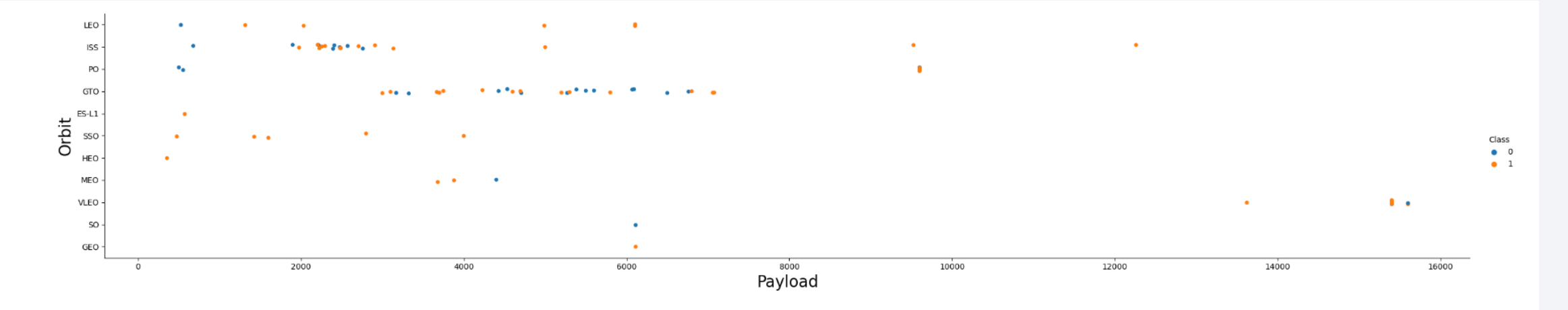


Flight Number vs. Orbit Type



- LEO, ISS, PO, GTO, have very high failure rates earlier on
- HEO, MEO, VLEO, SO, GEO flights all are not launched until the second half of the launch period

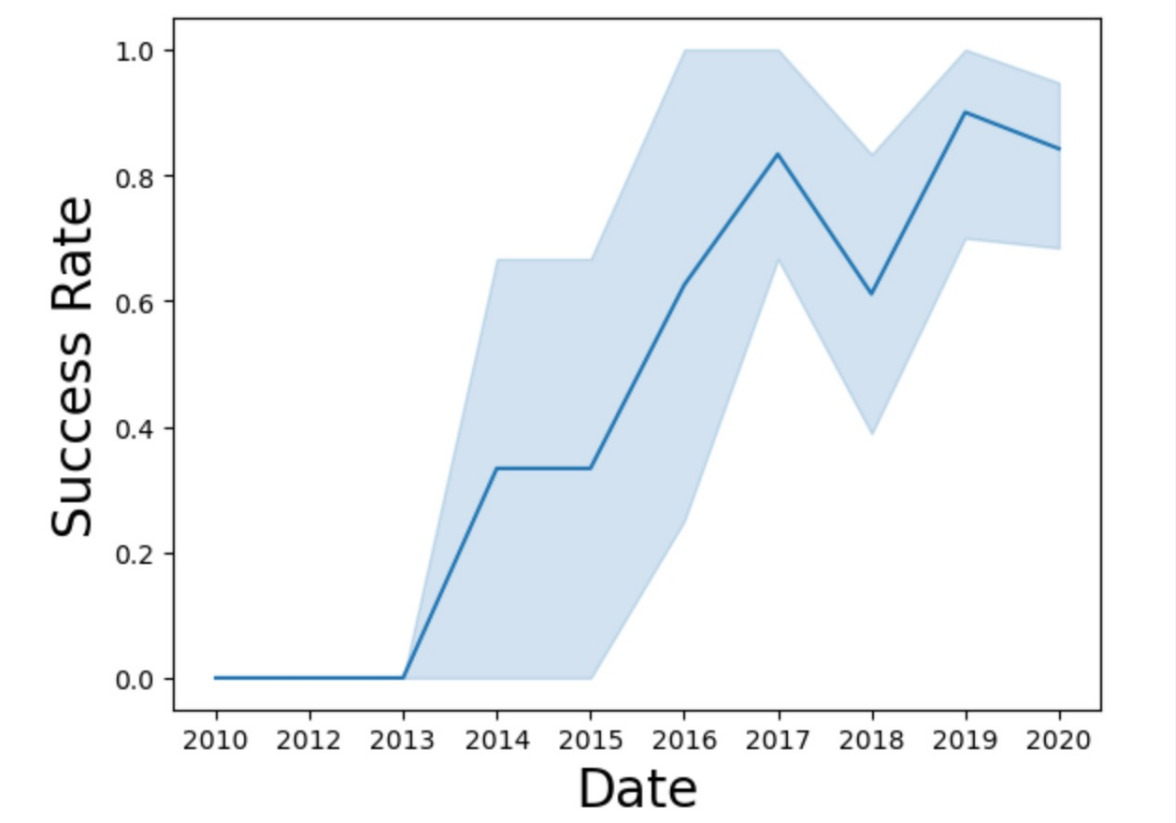
Payload vs. Orbit Type



- THE SSO, despite its 100% success rate, never takes on a payload above 4000 kg
- Although GTO orbits have a higher failure rate, they also appear to have one the highest median payload sizes

Launch Success Yearly Trend

- Its clear that launch success improves dramatically with time, although a strange dip occurs in the earlier part of 2018



All Launch Site Names

- All launch site names present within the SPACEXTABLE dataset

Task 1

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

In [11]:
```
%sql select DISTINCT(Launch_Site) from SPACEXTABLE
```

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

Out[11]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

Task 2

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

In [13]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (I
2010-08-12	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 (I
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Calculated total payload carried by boosters from NASA

Task 3

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

In [24]:
```
%sql select SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass_CRS from SPACEXTABLE where Customer like 'Nasa (CRS)'
```

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

Out[24]:

Total_Payload_Mass_CRS
45596

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1

Task 4

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

```
In [23]: %sql select AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass_F9v11 from SPACEXTABLE where Booster_Version like 'F9
 * sqlite:///my_data1.db
Done.
```

Out[23]:

Average_Payload_Mass_F9v11
2534.6666666666665

First Successful Ground Landing Date

- Found the dates of the first successful landing outcome on ground pad

Task 5

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

Hint:Use min function

In [19]:
```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)'
```

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

Out[19]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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

In [28]:
```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome like 'Success (drone ship)' and \
4000 < PAYLOAD_MASS__KG_ < 6000
```

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

Out[28]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

In [32]:
```
%sql select Mission_Outcome, COUNT(Mission_Outcome) as Occurrences \
from SPACEXTABLE group by Mission_Outcome;
```

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

Out[32]:

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

Boosters Carried Maximum Payload

- Listed the names of the booster which have carried the maximum payload mass

Task 8

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

In [41]:
```
%sql select DISTINCT Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) fr
```

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

Out[41]:

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

2015 Launch Records

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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.

In [45]:
```
%%sql select Landing_Outcome, Booster_Version, Launch_Site, substr(Date, 6,2) as Month
from SPACEXTABLE
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015'
```

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

Out[45]:

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

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

- Ranked 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

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.

In [48]:
```
%%sql
Select Landing_Outcome, count(Landing_Outcome) as Occurrences from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by landing_outcome order by Occurrences desc
```

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

Out[48]:

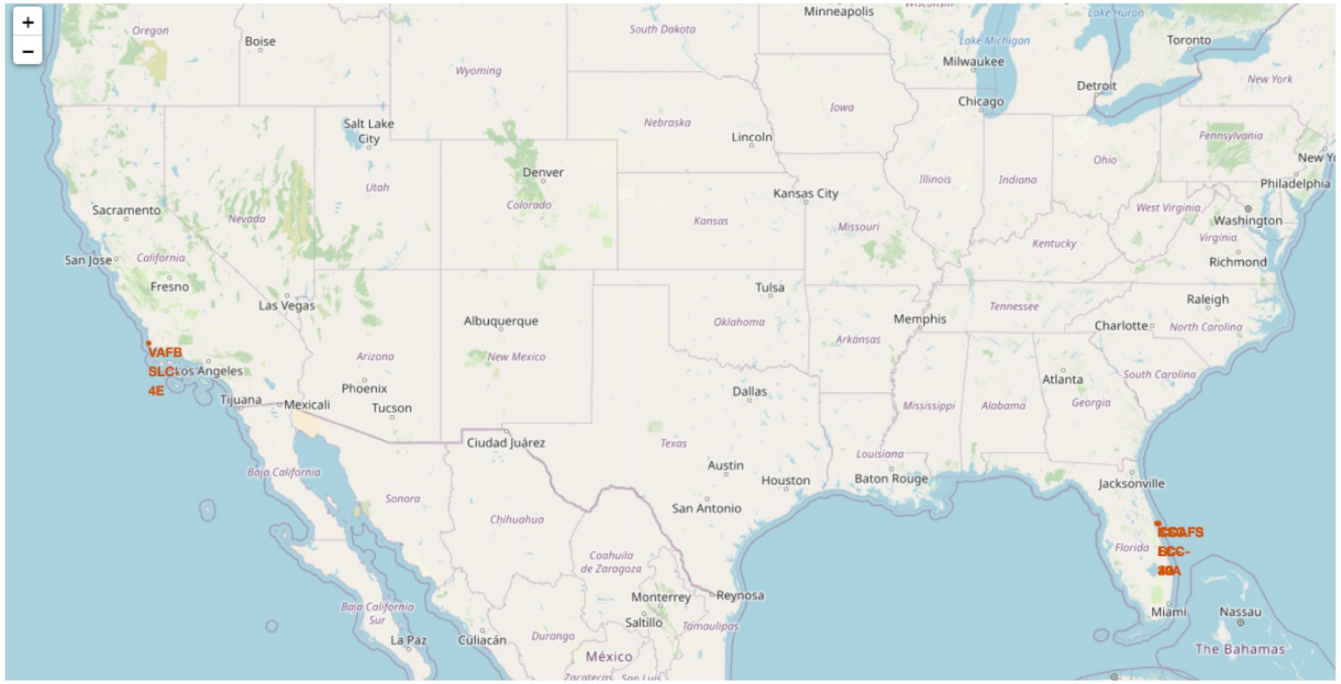
Landing_Outcome	count(Landing_Outcome)
Failure (parachute)	32


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

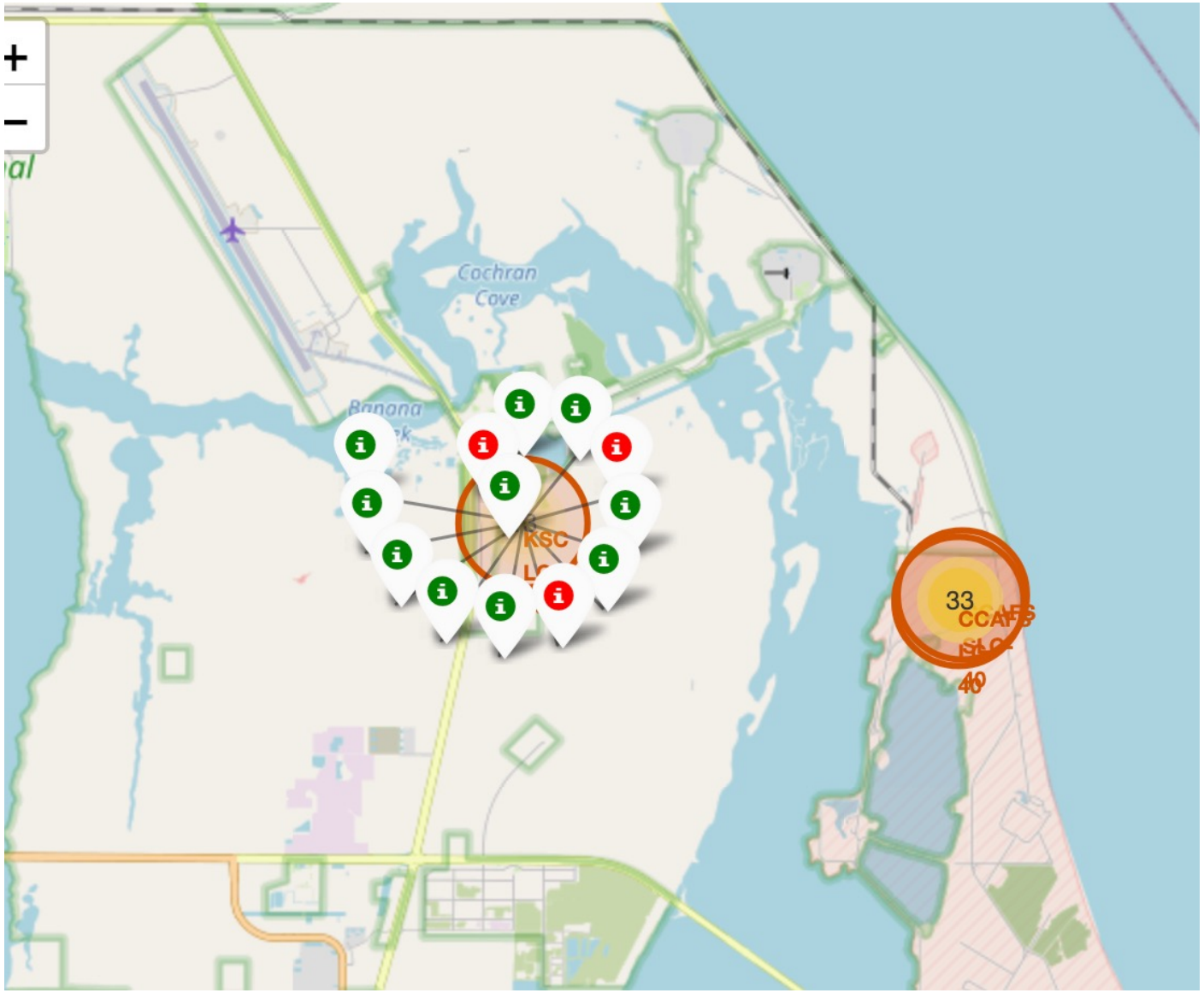
- Marked are the map are the locations of launch sites Space X has utilized
- They are mostly on the southern coast of California, and the Atlantic Coast of Florida

The generated map with marked launch sites should look similar to the following:



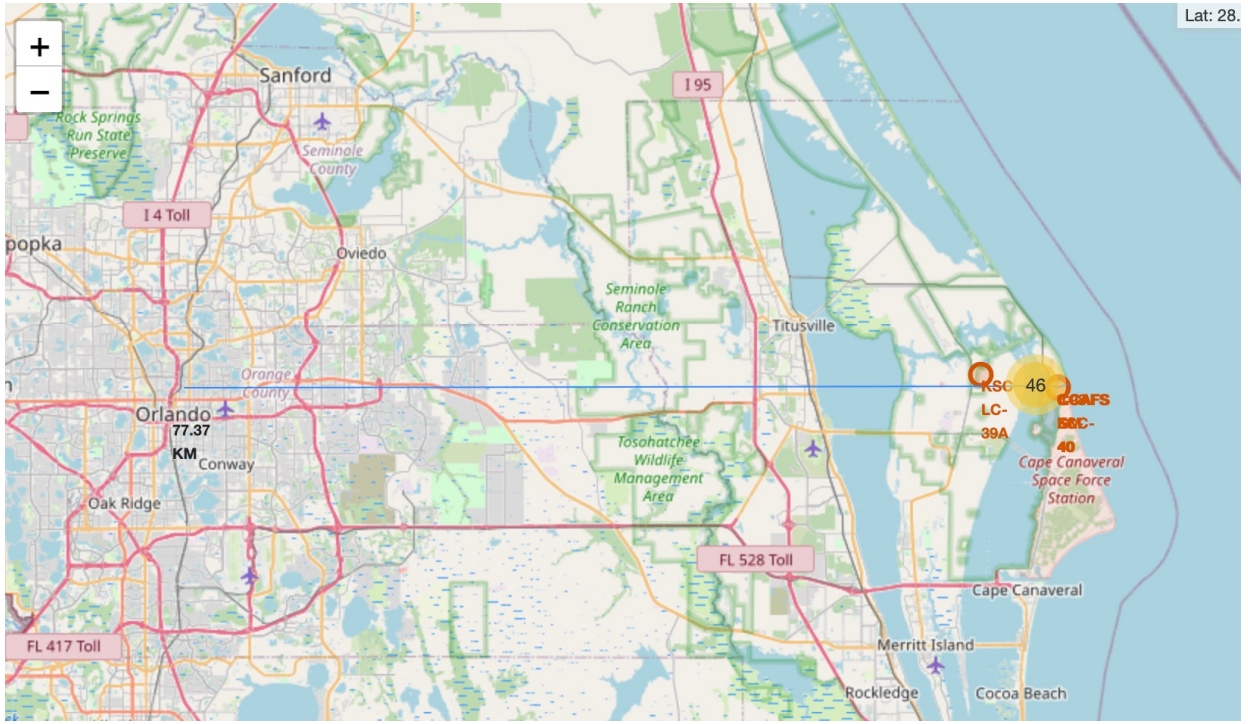
Cluster Displays

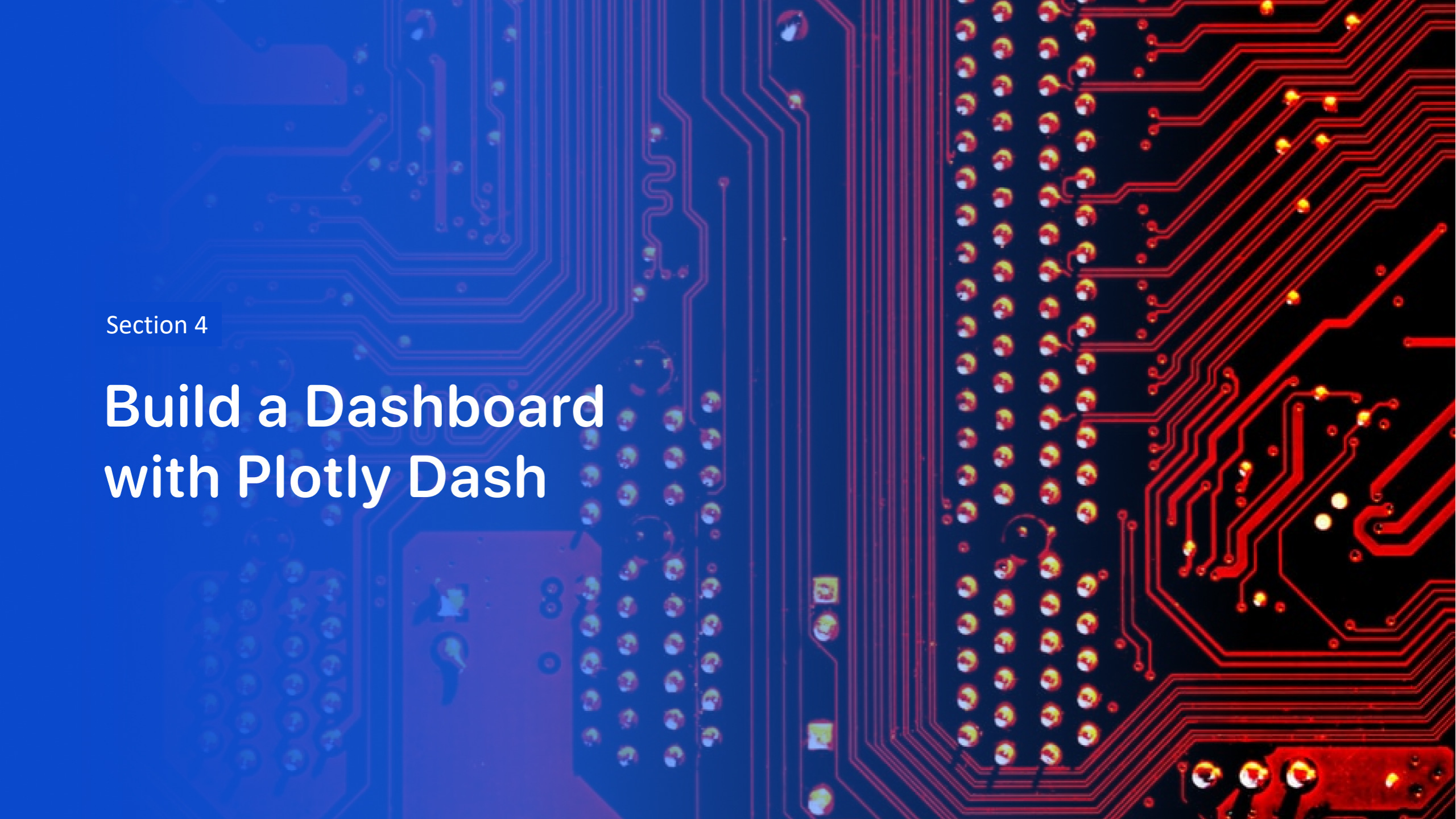
At each launch location, by clicking the user can see a display of all launches that occurred there and if they were successful or not (based on color)



Showing Proximity to Nearby Infrastructure

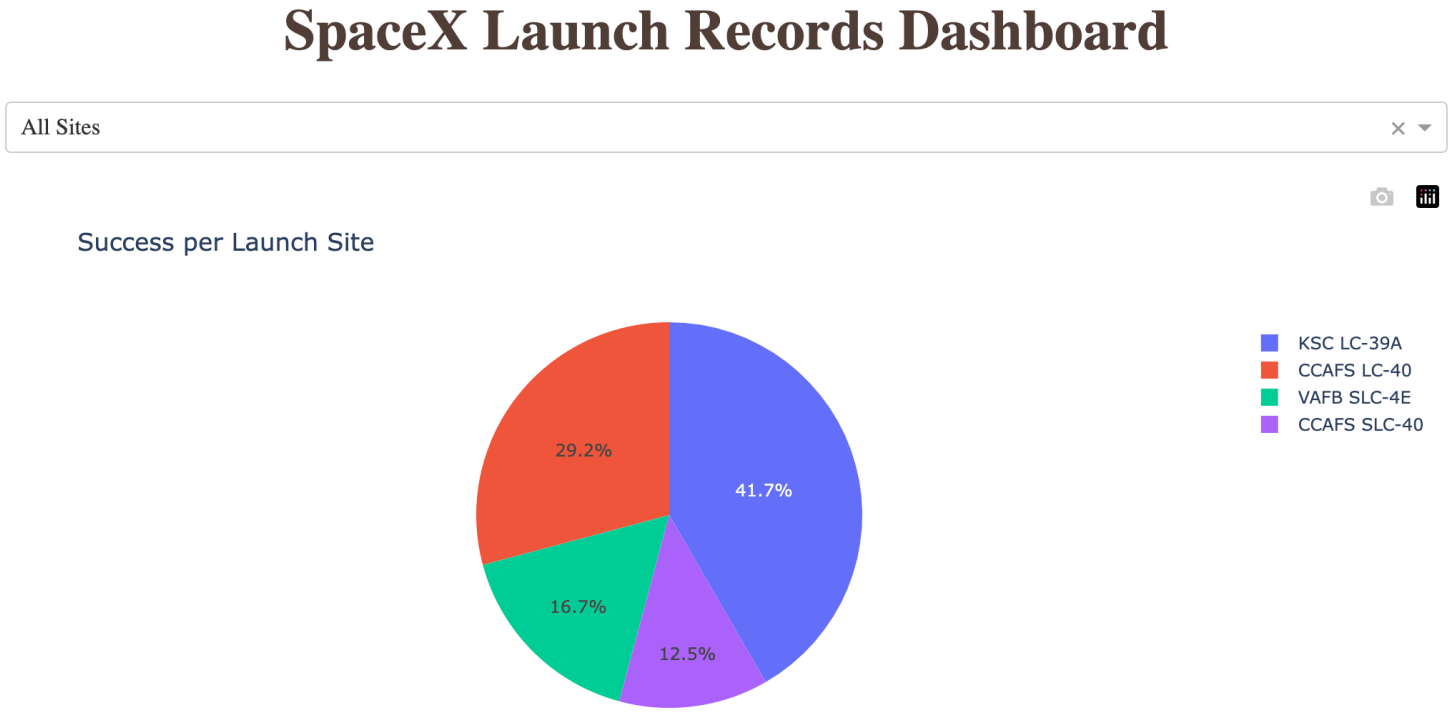
- Throughout making the folium map it was also important to show what infrastructure each location was near to – in this case Orlando and everything important there.




Section 4
Build a Dashboard with Plotly Dash

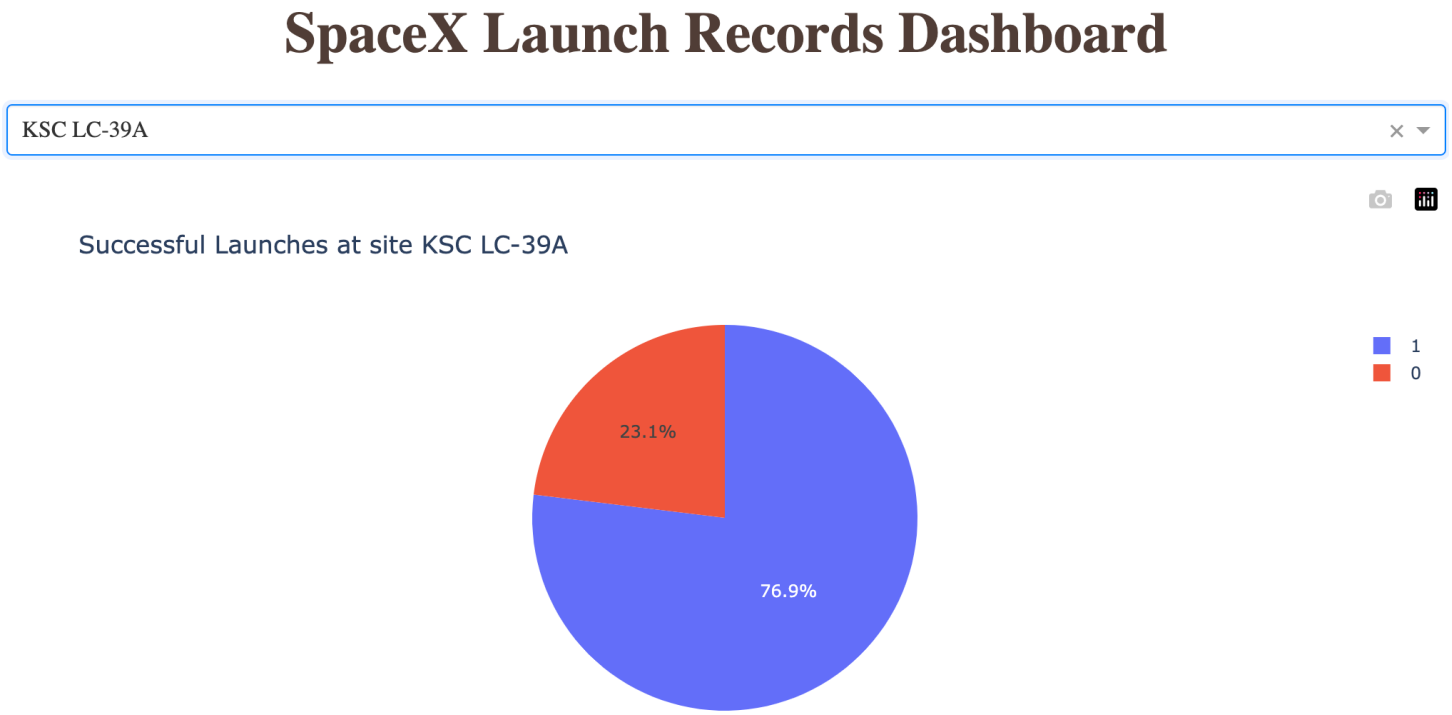
Share of total success per launch site

- A pie chart displaying the share of successful launches each site is responsible for
- KSC LC 39A being responsible for the most
- CCAFS SLC 40 being responsible for the least



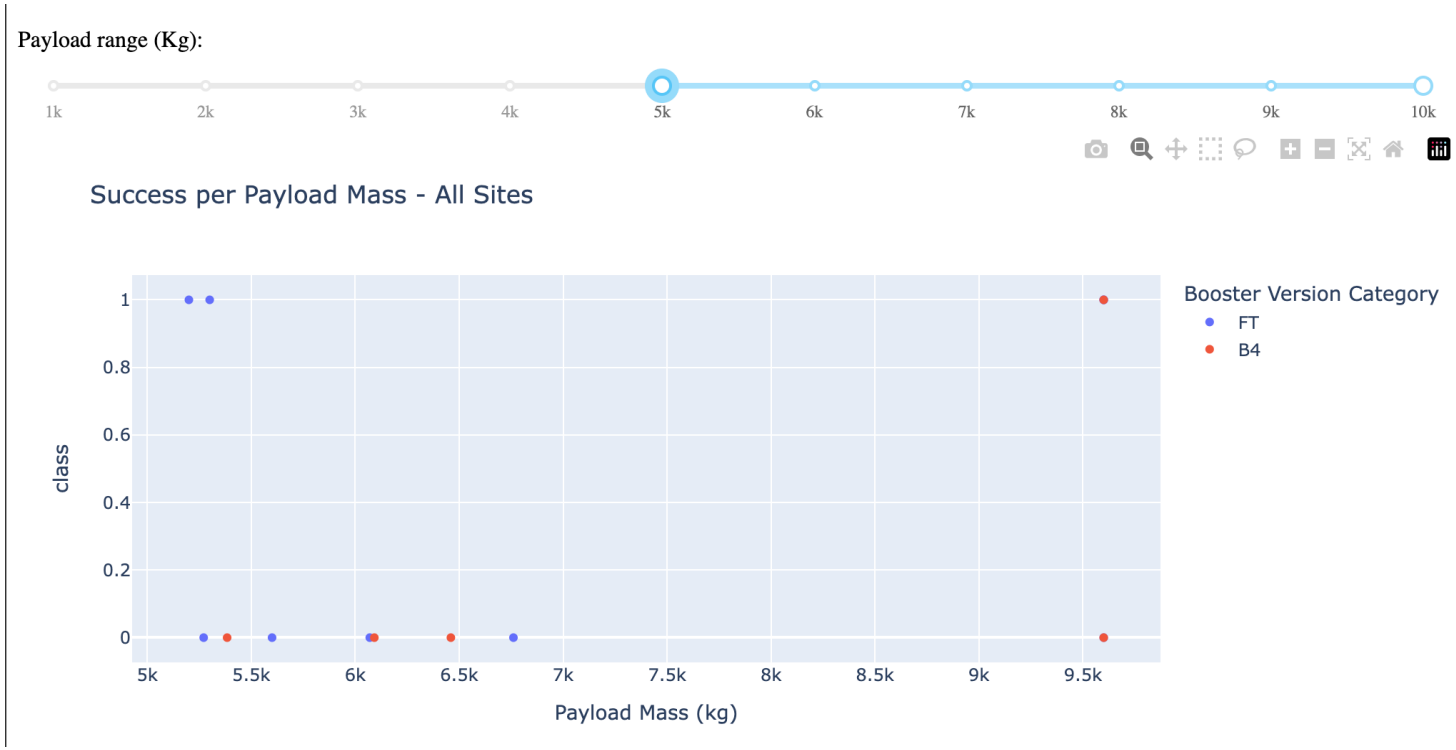
Piechart for the launch site with highest launch success ratio

- The site responsible for the highest number of successful launches, was successful 76.9% of the time, which is also the largest percentage of favored outcomes among all sites



Payload vs. Launch Outcome scatter plot for all sites – 5k-10k payload range

- Only the B4 booster was used after a payload size of 7k
- FT has the slightly larger failure rate, although it is also used more frequently

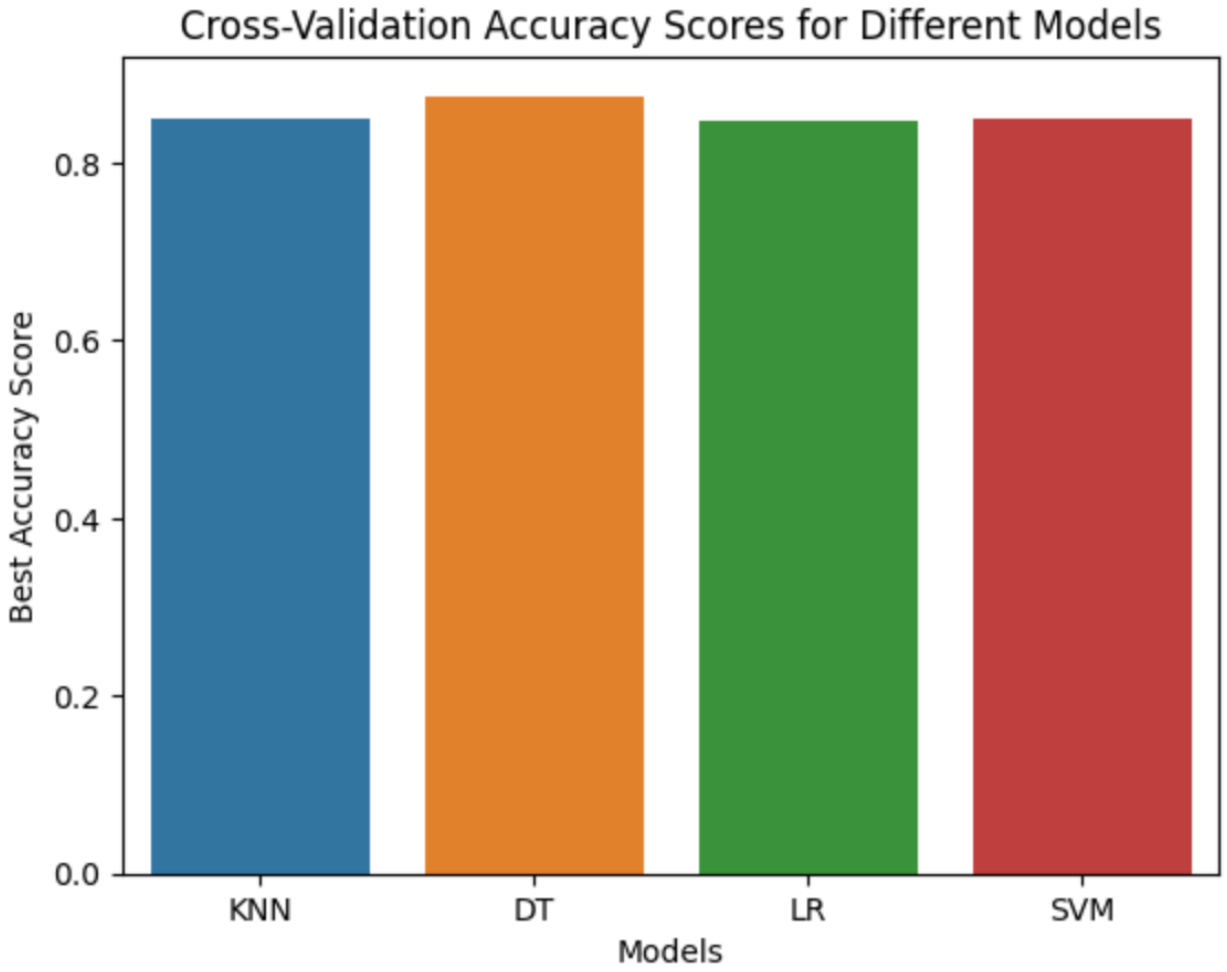


Section 5

Predictive Analysis (Classification)

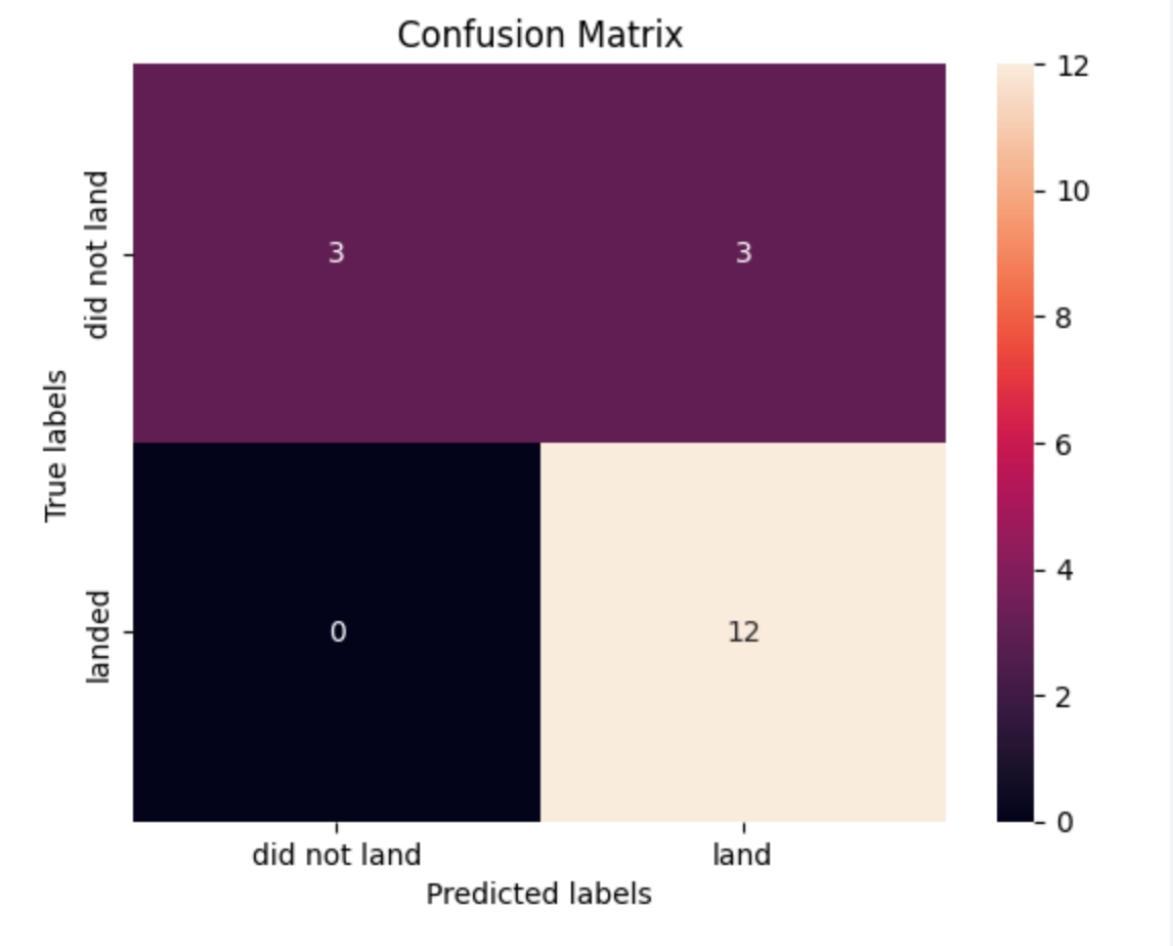
Classification Accuracy

- Although all models' best parameter settings display respectable accuracy scores, decision tree clearly takes the cake with an accuracy of .875



Confusion Matrix

- The decision matrix had the highest accuracy while predicting a land 15/18 times. The three times it predicted incorrectly occurred when it predicted a land and that did not occur. Although it predicted a failed landing only three times, it was correct in each prediction.



Conclusions

- The first major conclusion to be drawn is that practice is perfect and time is the most important factor in a launch. It was repeatedly seen that at each launch site, the more time that had passed the more likely the flight would be succeed.
- The second conclusion would be that some orbits might just be not worth the attempt (financially at least), as five orbits had a success rate of 100%, and the others were all below 50%.
- It was also clear that the higher the payload, the far more difficult it would be to complete a successful launch.
- Finally, while moving forward it seems to reason to prepare decision tree models for future use. Their accuracy is dependable, and although they are slower to run than other models, in the instance of bidding for +100M dollar projects, the time does not seem to be a major factor.

Appendix

- During the project, custom functions provided by the IBM / COURSERA team were used to help its completion: https://www.coursera.org/learn/applied-data-science-capstone/home/info

Thank you!