Winning Space Race with Data Science

Teryll Felix

26th October 2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping
- Perform data wrangling
  - Generate landing Class from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV to find best fit model

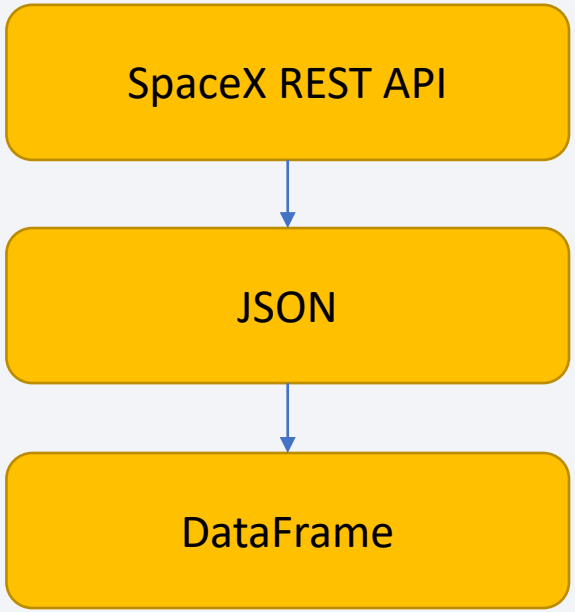
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

Notebook Link

REST API



In [11]:
```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

In [12]:
```
# Get the head of the dataframe
data.head(5)
```

Out[12]:

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3b
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]	[5eb0e4b6b6c3b
2	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 140, 'altitude': 35, 'reason': 'residual stage-1 thrust led to collision between stage 1 and stage 2'}]	Residual stage 1 thrust led to collision between stage 1 and stage 2	[]	[]	[]	[5eb0e4b6b6c3b 5eb0e4b6b6c3b
3	2008-09-20T00:00:00.000Z	1.221869e+09	False	0.0	5e9d0d95eda69955f709d1eb	True	[]	Ratsat was carried to orbit on the first successful orbital launch of any privately funded and developed, liquid-propelled carrier rocket, the SpaceX Falcon 1	[]	[]	[]	[5eb0e4b7b6c3b
4	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	True	[]	None	[]	[]	[]	[5eb0e4b7b6c3b

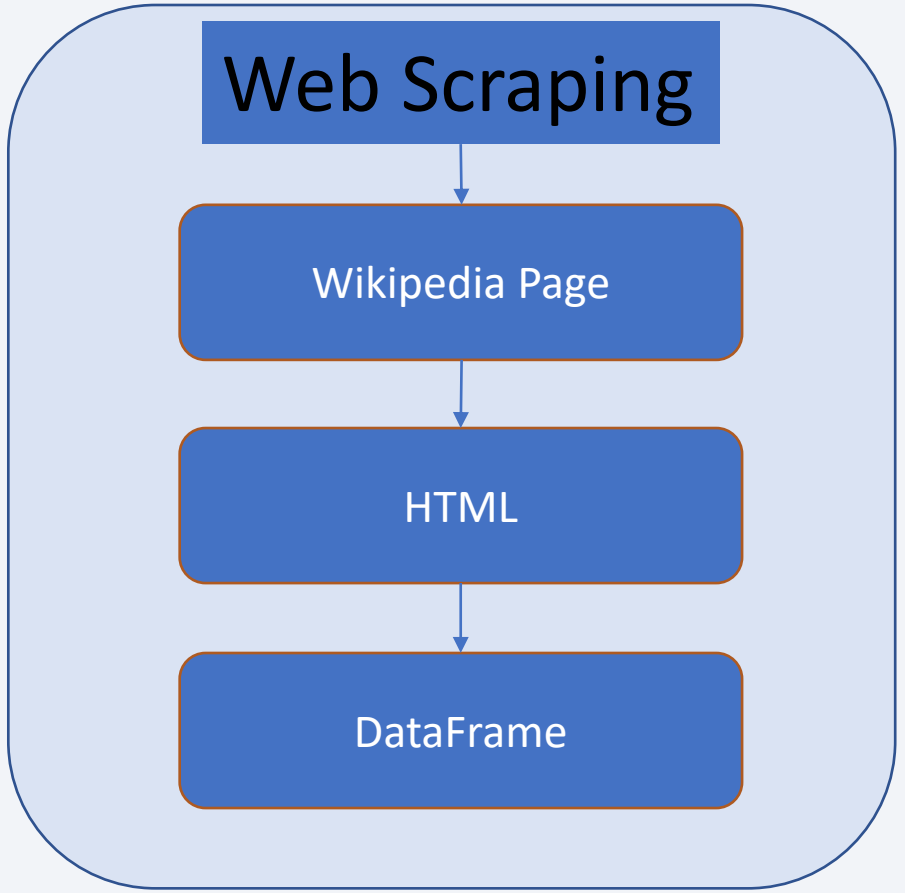
Screenshot of SpaceX API Call

Screenshot of SpaceX API Call Outcome

Data Collection - Scraping

Notebook Link

https://github.com/TF758/IBM-DS-Capstone-Project/blob/7e255d037daf6a336e26c665c149caf7fcf08831/Week%201/jupyter-labs-webscraping.ipynb



Data Collection – Scraping Results

```
import requests
from bs4 import BeautifulSoup
url = 'https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches'
response = requests.get(url)
html_data = response.text
soup = BeautifulSoup(html_data)
```

```
<tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a
>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Vers
ion,<br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-Dragon-12">[c]</a></sup>
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
<th scope="col">Customer
</th>
<th scope="col">Launch<br/>outcome
</th>
<th scope="col"><a href="/wiki/Falcon_9_first-stage_landing_tests" title="Falcon 9 first-stage landing tests">Booster<b
r/>landing</a>
</th></tr>
```

Data Wrangling

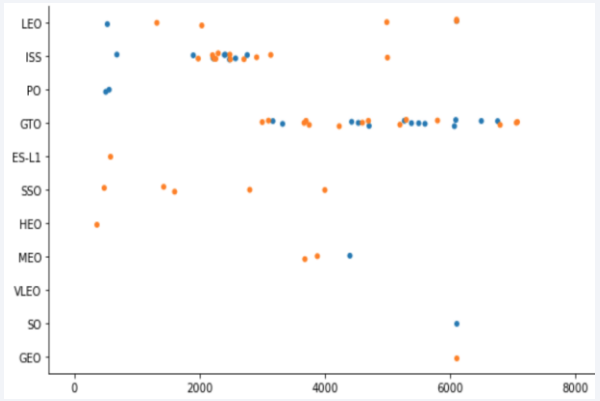
- Notebook Link

https://github.com/TF758/IBM-DS-Capstone-Project/blob/7e255d037daf6a336e26c665c149caf7fcf08831/Week%201/labs-jupyter-spacex-Data%20wrangling.ipynb

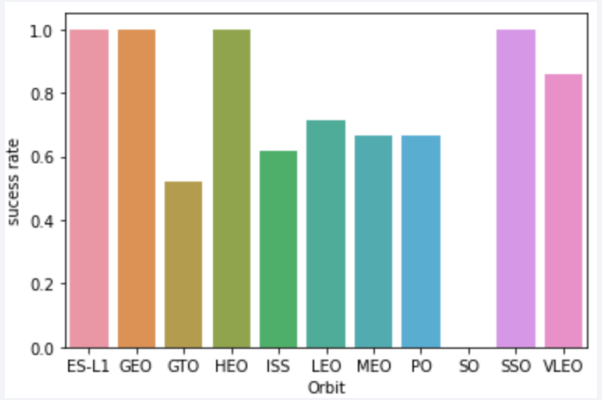
Transform raw data to a useful data format. For example, convert original outcome labels into landing class that represent landing classification which will be our new landing prediction target.

EDA with Data Visualization

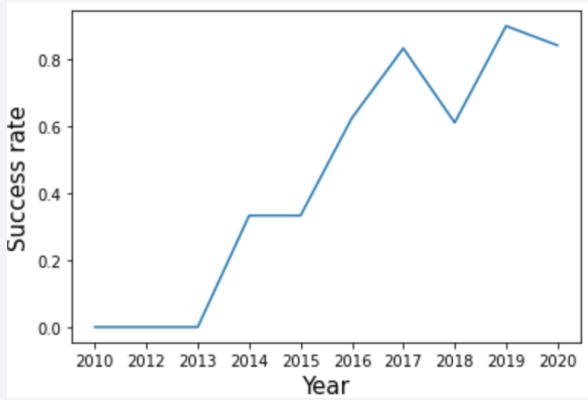
- Notebook
- https://github.com/TF758/IBM-DS-Capstone-Project/blob/6632d90fc2f7419f253dae6e67e046561e464224/Week%202/jupyter-labs-eda-dataviz.ipynb



The **scatterplot** graph was used to find the relationship between 1 or more variables



A **Bargraph** was use to compare the success rates of each orbit



A **line graph** was used to get the yearly average launch success trend

EDA with SQL

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

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

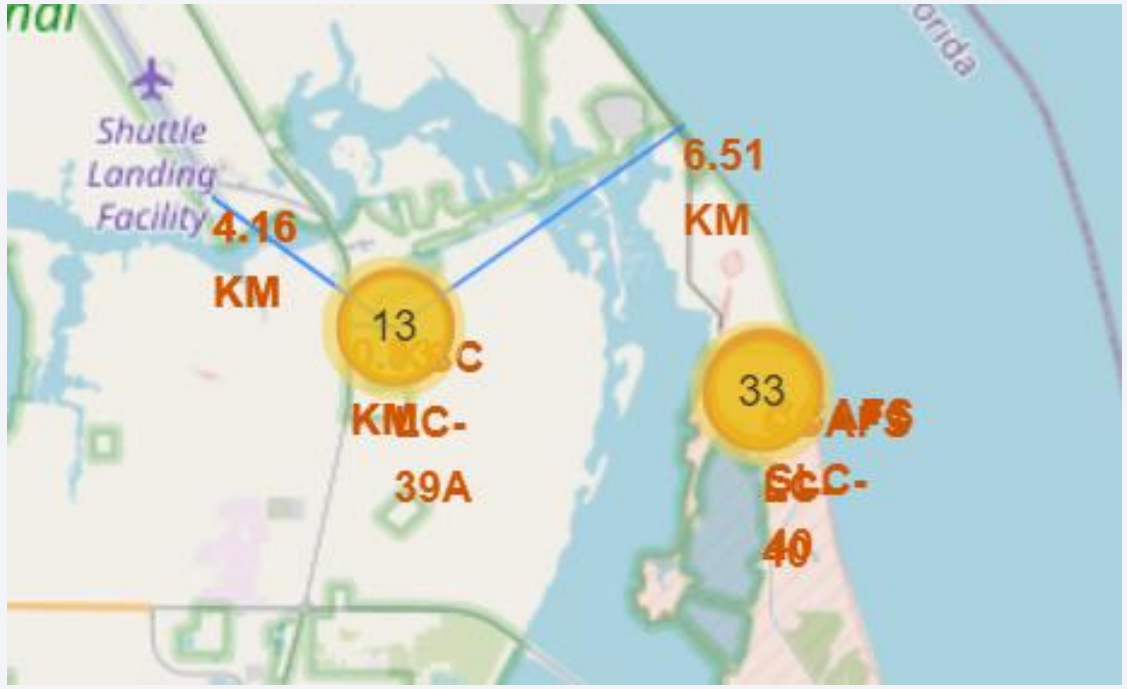
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

- Query the names of the **unique launch sites** in the space mission
- Query the names of the **booster_versions** which have carried the maximum payload mass.
- List the total number of **successful** and **failure** mission outcomes
- List the names of the boosters which have **success in drone ship** and have **payload mass** in some range
- Rank the count of successful **landing_outcomes** in date range in descending order.

Build an Interactive Map with Folium

- Add **Circles** for Launch sites and **Markers** for labels
- Add **Lines** for calculate distance between launch sites and their proximities

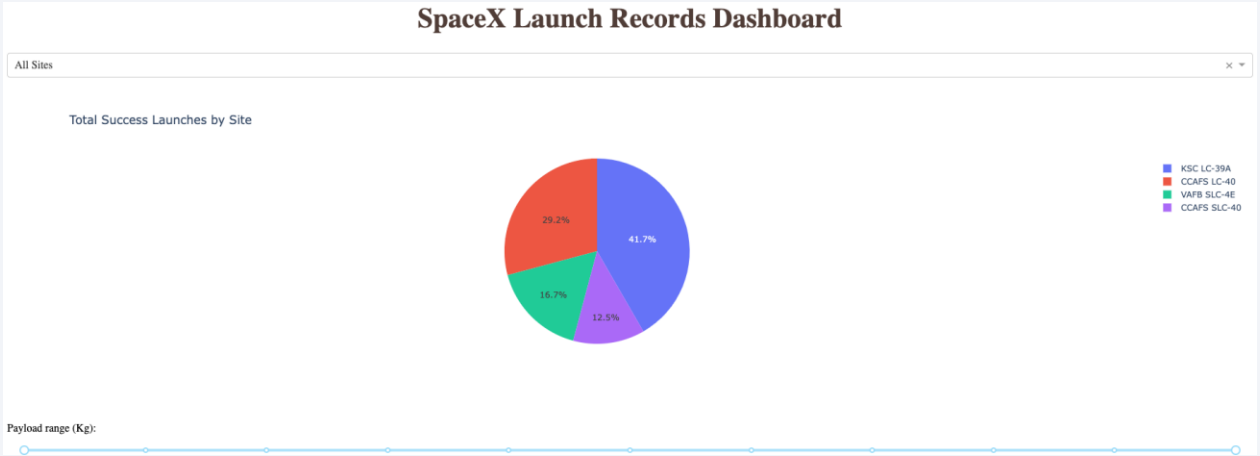


Notebook

https://github.com/TF758/IBM-DS-Capstone-Project/blob/c9b32a1e6005ada3c26b1ee1ff320eaa15fcb165/Week%203/jupyter_launch_site_location_folium.ipynb

Build a Dashboard with Plotly Dash

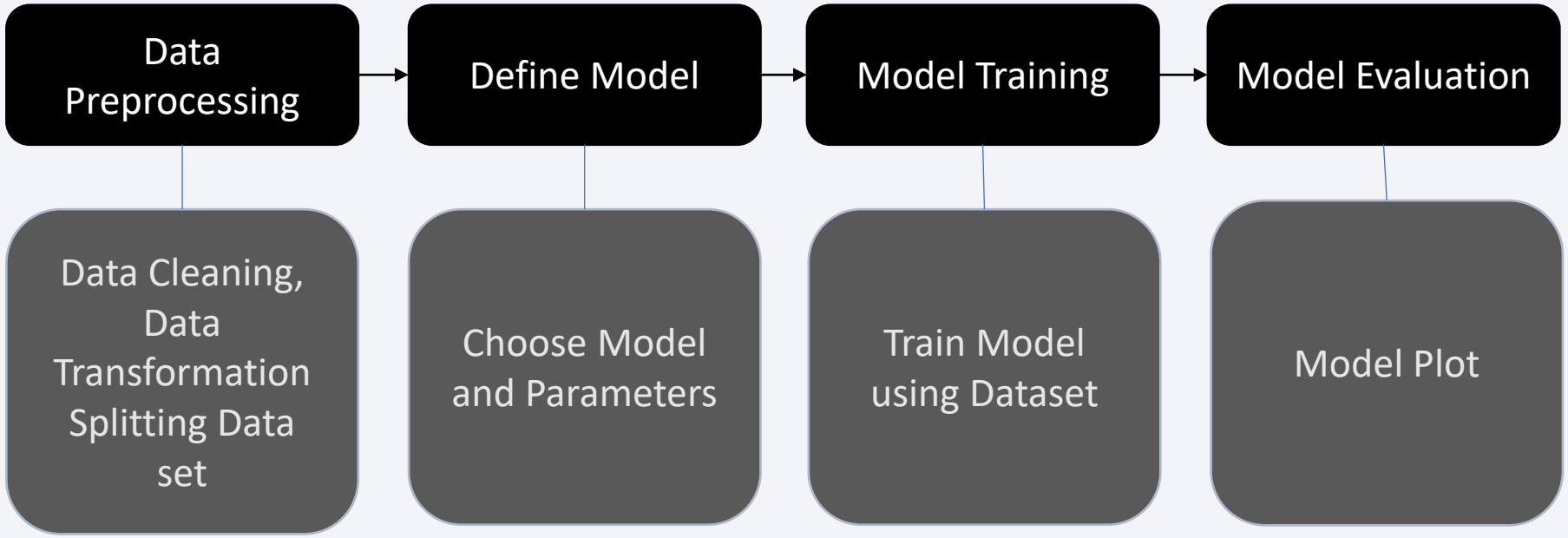
- With a **Dropdown menu** and a **Pie Chart**, we can get success launches distribution by launch site
- With a **Range Slider** and a **Scatter Plot**, we can analyze the correlation between Payload and Success for different launch sites



Github Link

https://github.com/TF758/IBM-DS-Capstone-Project/blob/01b435ca23db6449822d8c78891423500303a226/Week%203/spacex_dash_app.py

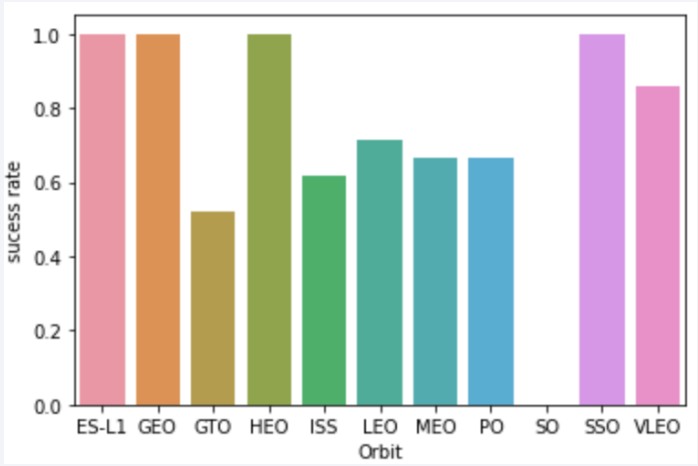
Predictive Analysis (Classification)

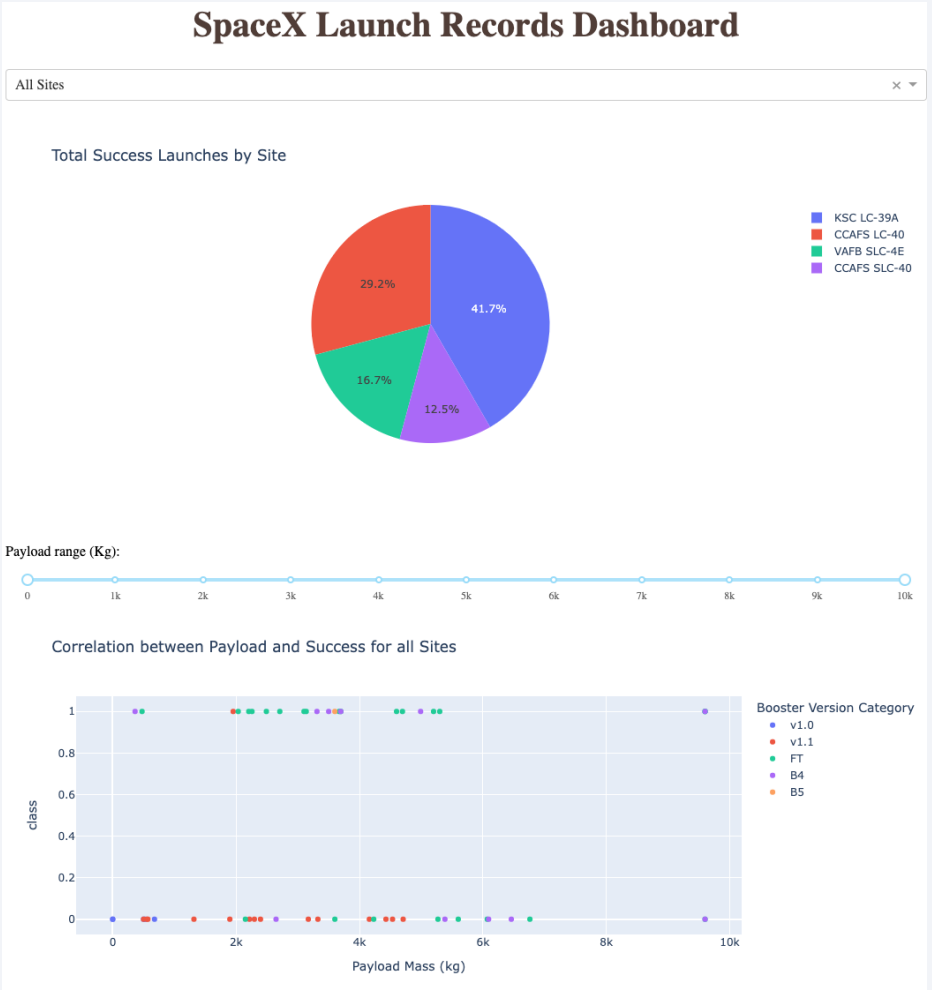


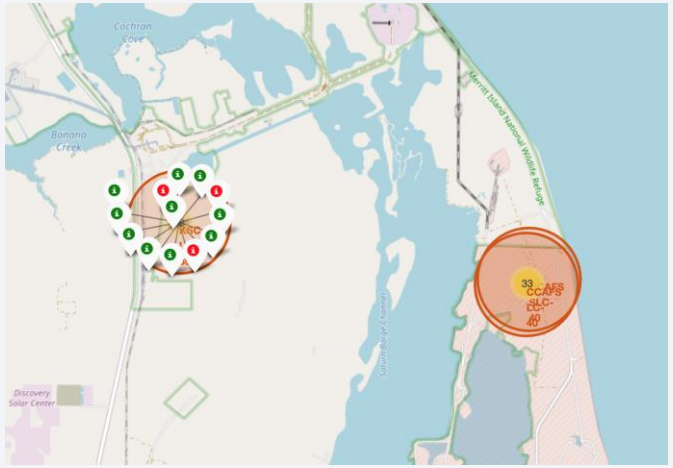
Notebook

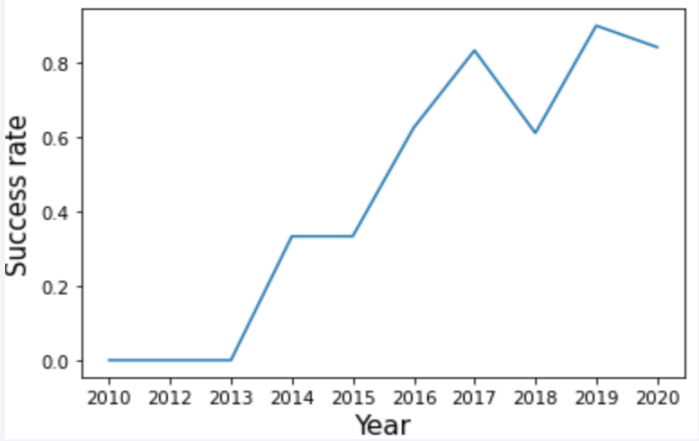
https://github.com/TF758/IBM-DS-Capstone-Project/blob/01b435ca23db6449822d8c78891423500303a226/Week%204/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

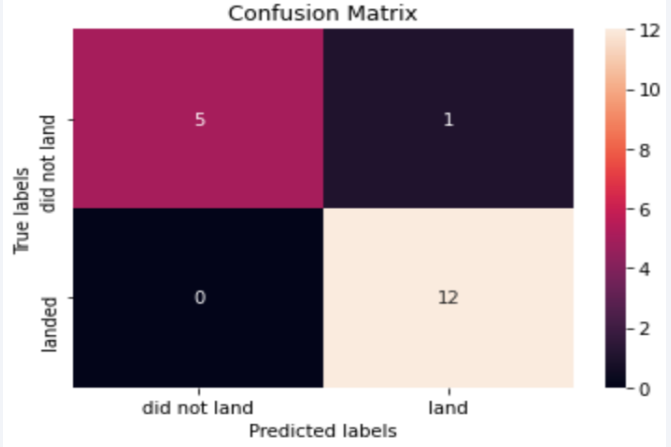
Results

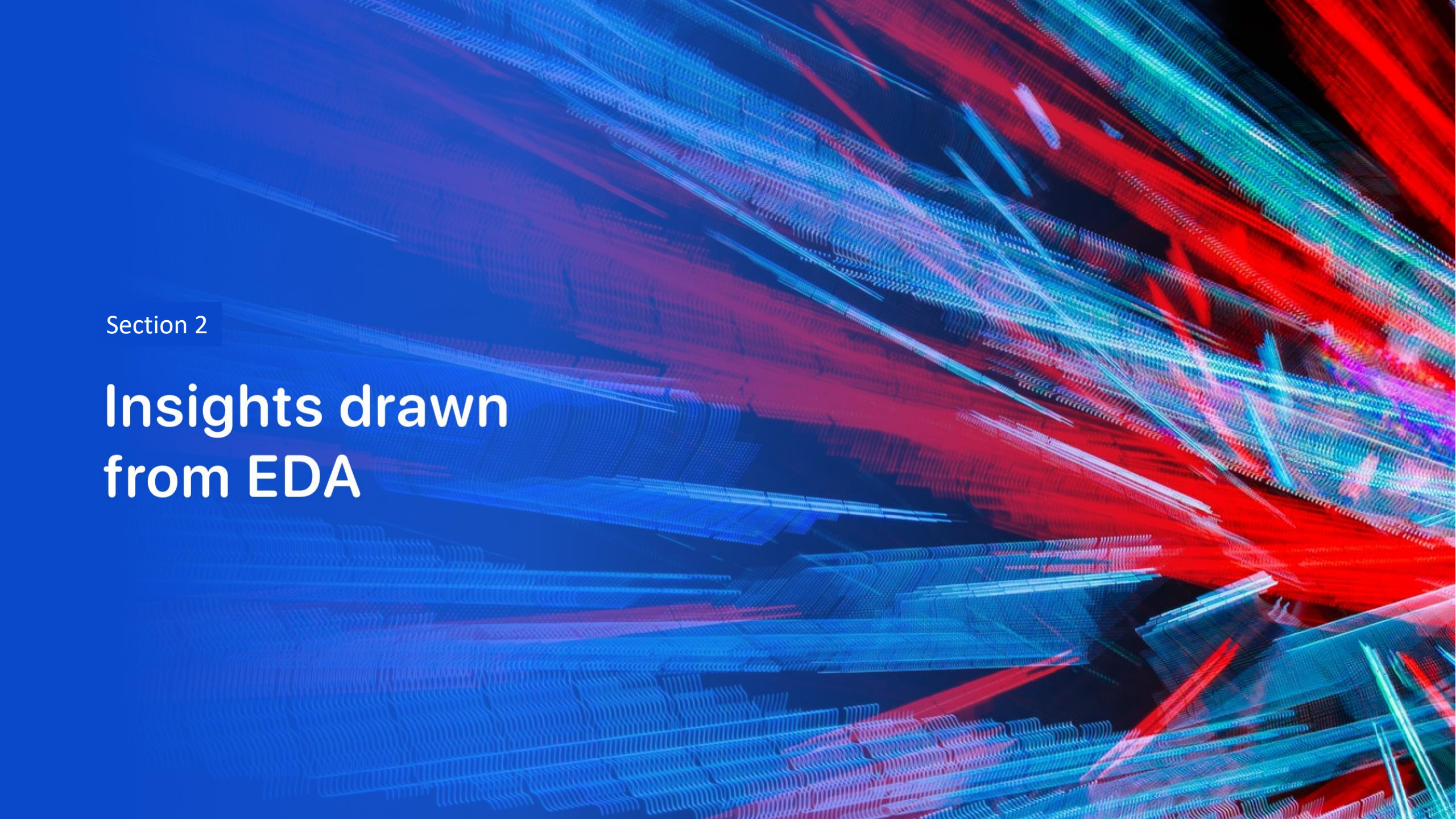

1.0
0.8
0.6
0.4
0.2
0.0
sucess rate
ES-L1
GEO
GTO
HEO
ISS
LEO
MEO
PO
SO
SSO
VLEO
Orbit


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k
Correlation between Payload and Success for all Sites
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5


Cochran Cove
Banana Creek
Merritt Island National Wildlife Refuge
KSC
CCAFS SLC-40
Discovery Solar Center

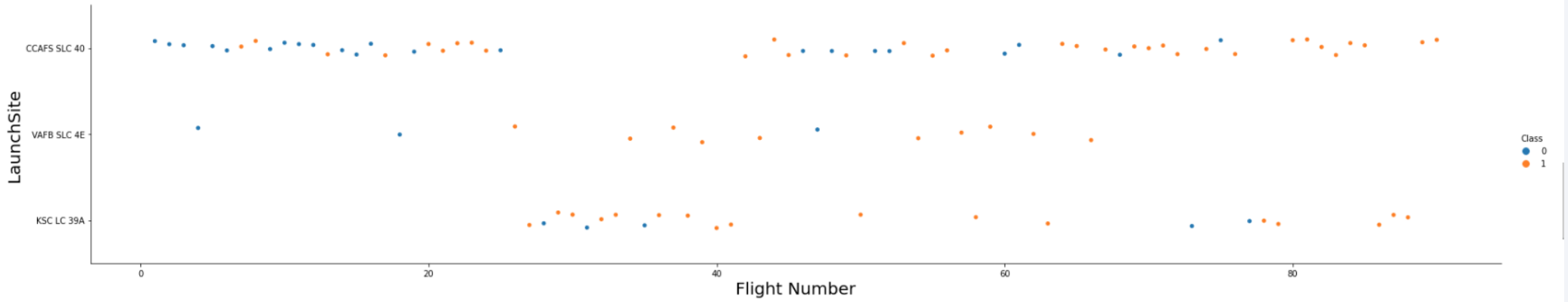

0.8
0.6
0.4
0.2
0.0
Success rate
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year


Confusion Matrix
5
1
0
12
True labels
did not land
landed
did not land
land
Predicted labels
12
10
8
6
4
2
0


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

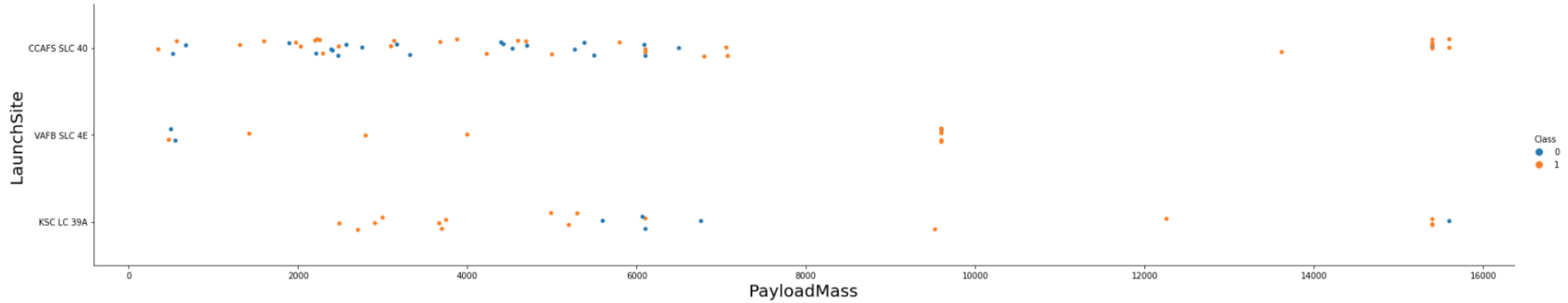
```
[4]  # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hu
     sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
     plt.xlabel("Flight Number",fontsize=20)
     plt.ylabel("LaunchSite",fontsize=20)
     plt.show()
```



The general trend observed is as the total number of flights increases, so does the amount of first stage landings.

Payload vs. Launch Site

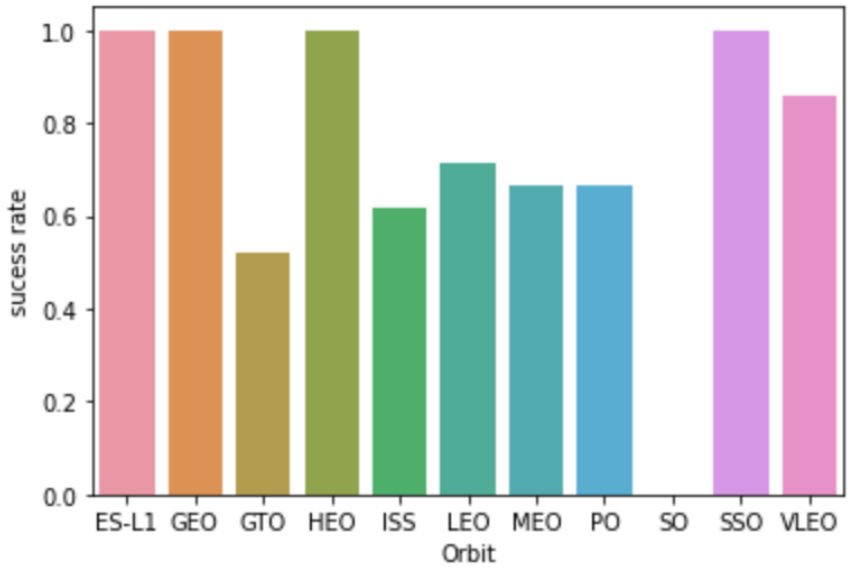
```
[5]  # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, a
     sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
     plt.xlabel("PayloadMass",fontsize=20)
     plt.ylabel("LaunchSite",fontsize=20)
     plt.show()
```



As payload mass increases, so does the success rate. However in he KSC LC391 Launch site is observed there is a much higher success rate with a low payload compared to CCAFS SLC

Success Rate vs. Orbit Type

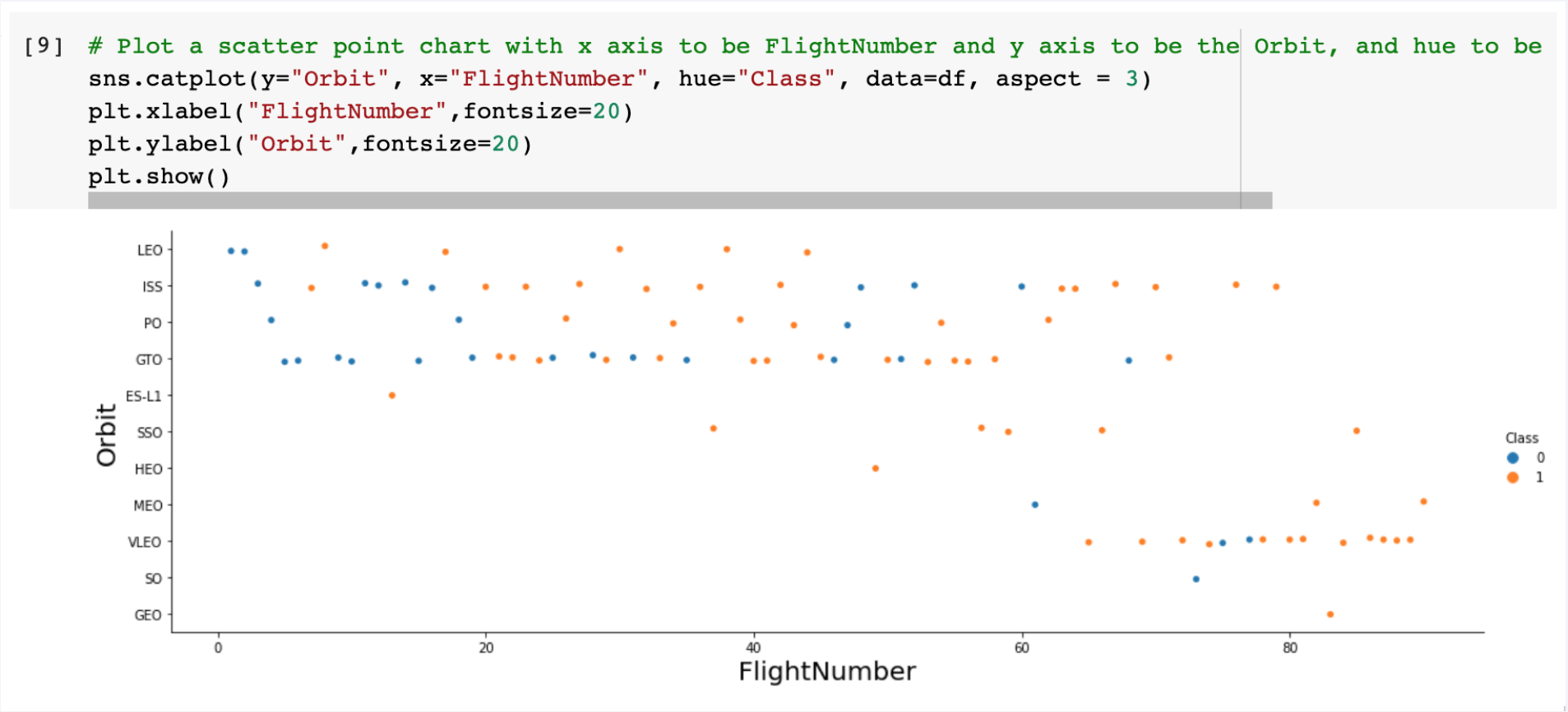
```
[ ]  sns.barplot(y='Class', x='Orbit', data=df_success_rate)
     plt.xlabel("Orbit",fontsize=10)
     plt.ylabel("sucess rate",fontsize=10)
     plt.show()
```



ES-L1, GEO, HEO & SSO all have the highest success rate of 100% however SO has the lowest success rate with 0%

Flight Number vs. Orbit Type

```
[9]  # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be
     sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 3)
     plt.xlabel("FlightNumber",fontsize=20)
     plt.ylabel("Orbit",fontsize=20)
     plt.show()
```

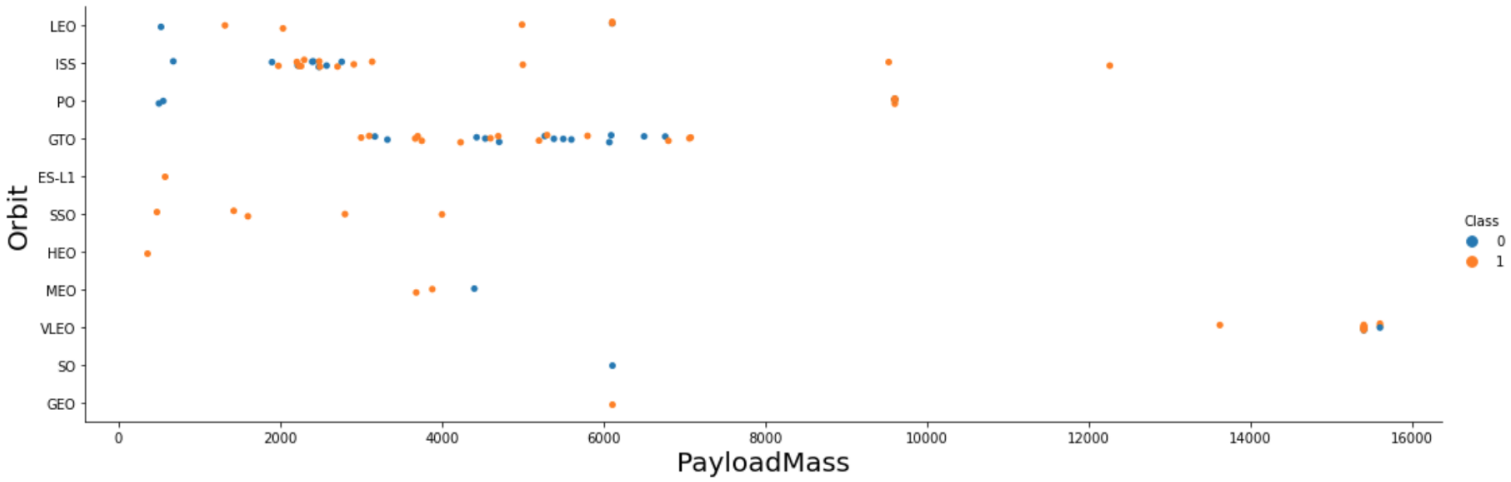


In ES-L1, GEO, HEO & SS all launches are successful

As flight number increases of LEO so does the success rating indicating there is a relationship. This however can't be seen in GTO

Payload vs. Orbit Type

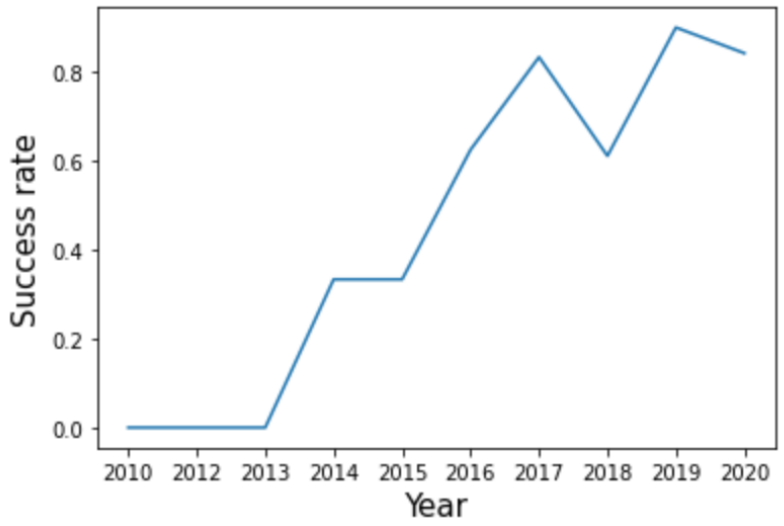
```
[ ]  # Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class v
     sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 3)
     plt.xlabel("PayloadMass",fontsize=20)
     plt.ylabel("Orbit",fontsize=20)
     plt.show()
```



With heavy payloads the successful landing increases for Polar, LEO and ISS. However for this is not observed for GTO

Launch Success Yearly Trend

```
[14] sns.lineplot(y='Class', x='Year', data=df_year_success)
     plt.xlabel("Year",fontsize=15)
     plt.ylabel("Success rate",fontsize=15)
     plt.show()
```



It can be seen that the success rate increases from 2013 with a momentary dip in 2018 before continuing to increase until it dropped again in 2020.

All Launch Site Names

Four Launch Sites:

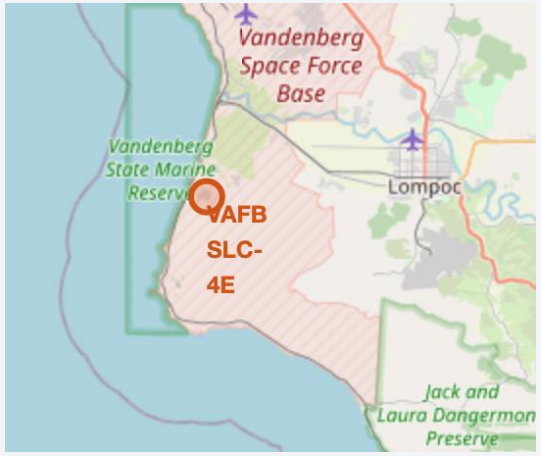
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

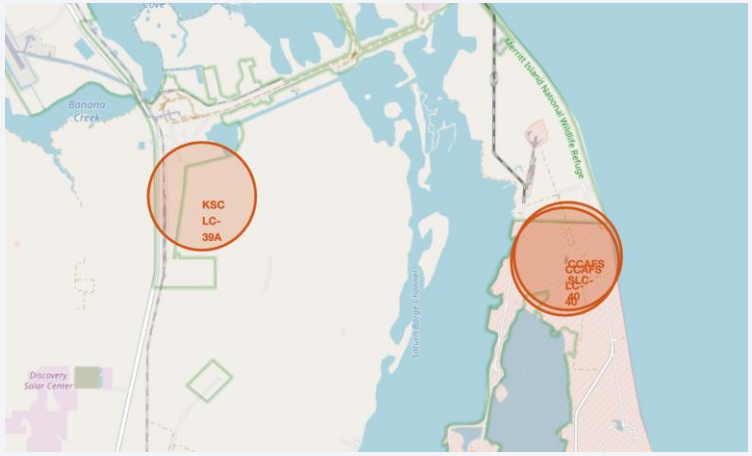
```
%sql select distinct Launch_Site from SPACEXTBL
```

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

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

1 in western coast
VAFB SLC-4E
3 in eastern coast
KSC LC-39A
CCAFS SLC-40
CCAFS LC-40





Launch Site Names Begin with 'CCA'

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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

5 launches happened in LEO orbit, and four of them were from customer NASA.

Total Payload Mass

```
[9]  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA%'
```

```
 * sqlite:///my_data1.db
Done.
sum(PAYLOAD_MASS__KG_)
99980
```

The total payload carried by boosters from NASA is **99980**.

Average Payload Mass by F9 v1.1

```
[ ]  %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
avg(PAYLOAD_MASS__KG_)
2534.6666666666665
```

the average payload mass carried by booster version F9 v1.1 is **2534.67**.

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)"
```

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

min(Date)
01-05-2017

the first successful landing outcome on ground pad is **01-05-2017**.

Successful Drone Ship Landing with Payload between 4000 and 6000

names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

```
%%sql

select Booster_Version from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)"
    and PAYLOAD_MASS__KG_>4000
    and PAYLOAD_MASS__KG_ < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

the total number of successful mission outcomes is 100

the total number of failure mission outcomes is 1

```
[10] %%sql

     select count(*) from SPACEXTBL
     where "Mission_Outcome" like "Success%"
```

```
 * sqlite:///my_data1.db
Done.
 count(*)
 100
```

```
[11] %%sql

     select count(*) from SPACEXTBL
     where "Mission_Outcome" like "Failure%"
```

```
 * sqlite:///my_data1.db
Done.
 count(*)
 1
```

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass:

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

```
%%sql

select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

```
[14] %%sql

     select substr(Date, 4, 2) as Month, Booster_Version, Launch_Site from SPACEXTBL
     where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)"
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20:

Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

```
%%sql

select "Landing _Outcome",
    count("Landing _Outcome") as landings
from SPACEXTBL
where Date >= "04-06-2010" and Date <= "20-03-2017"
group by "Landing _Outcome"
order by landings desc
```

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

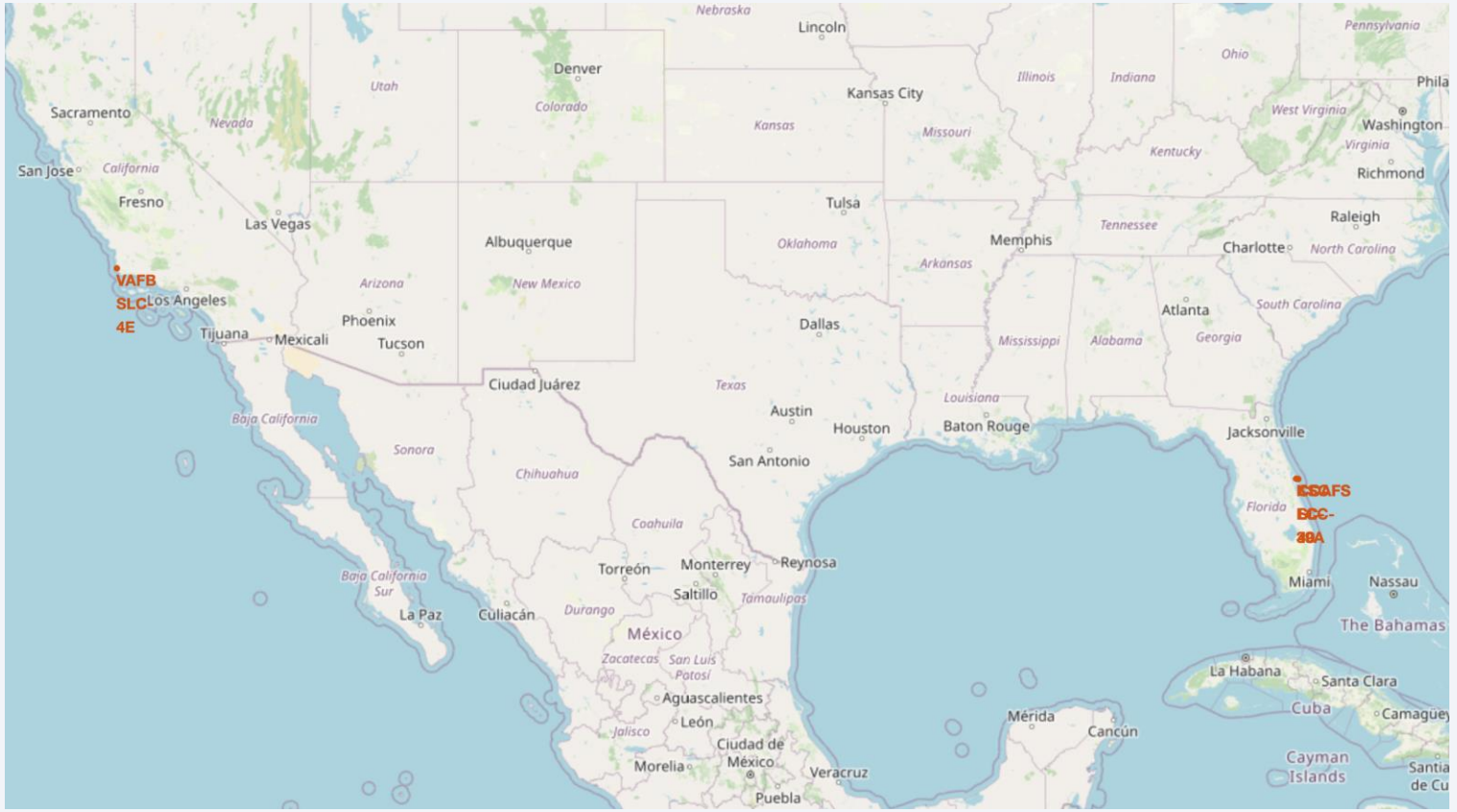
Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

Locations of Launch Sites on Maps

Three in the east
One in the west
All in the south

Launch Site	Lat	Long
CCAFS LC-40	28.56230197	-80.57735648
CCAFS SLC-40	28.56319718	-80.57682003
KSC LC-39A	28.57325457	-80.64689529
VAFB SLC-4E	34.63283416	-120.6107455

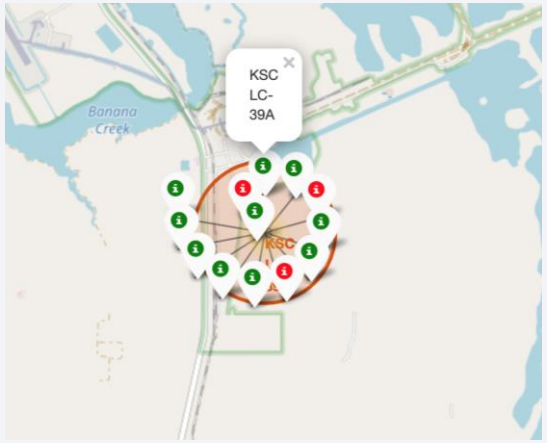


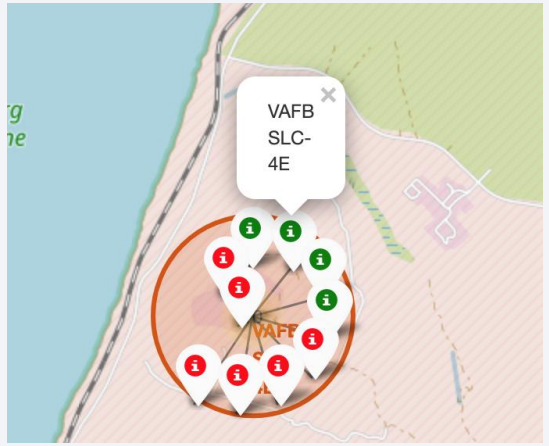
Display Launch Outcome by Color

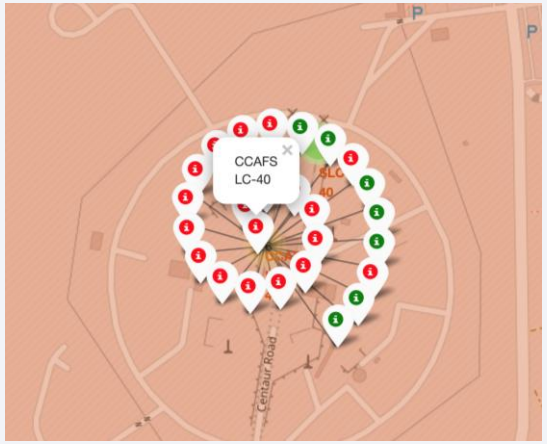
From the color labels, we can easily see

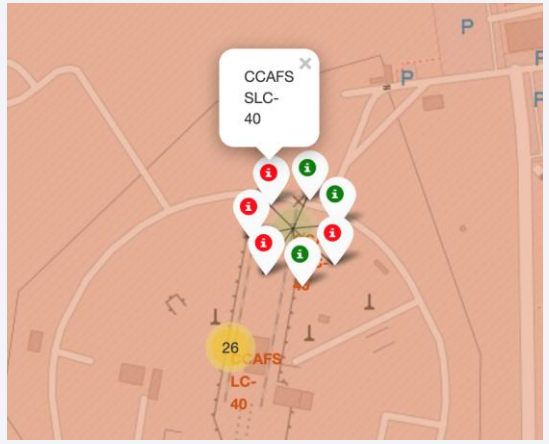
KSC LC-39A has a rather higher success rate

Whereas CCAFS LC-40 and CCAFS SLC-40 have much lower rate







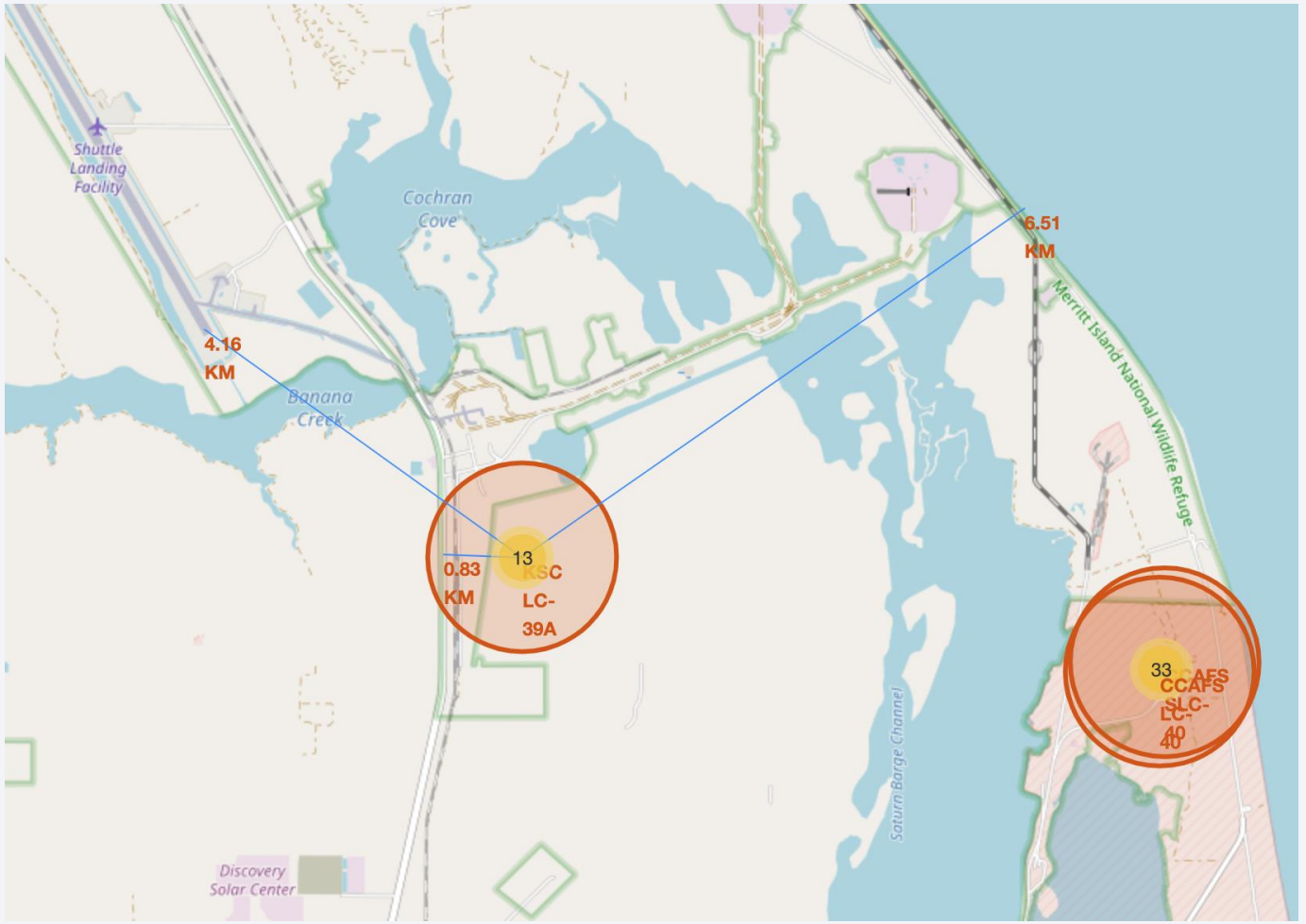


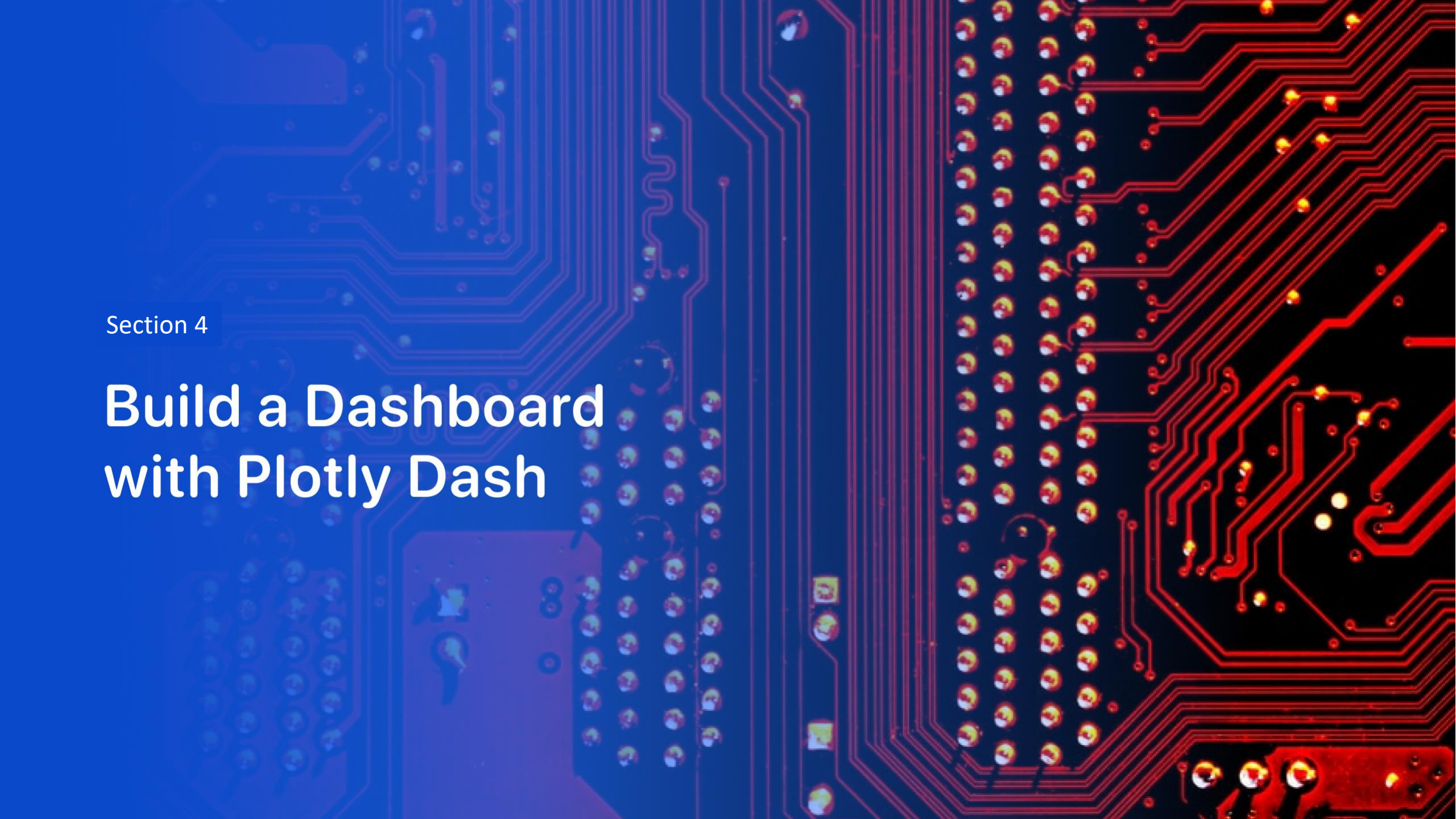
Show Distance to Proximities

The distance from KSC LC-39A to the nearest shuttle landing facility is about 4.16 km.

The distance from KSC LC-39A to the nearest highway is less than 1 km.

The distance from KSC LC-39A to the coastline is around 6.5 km.

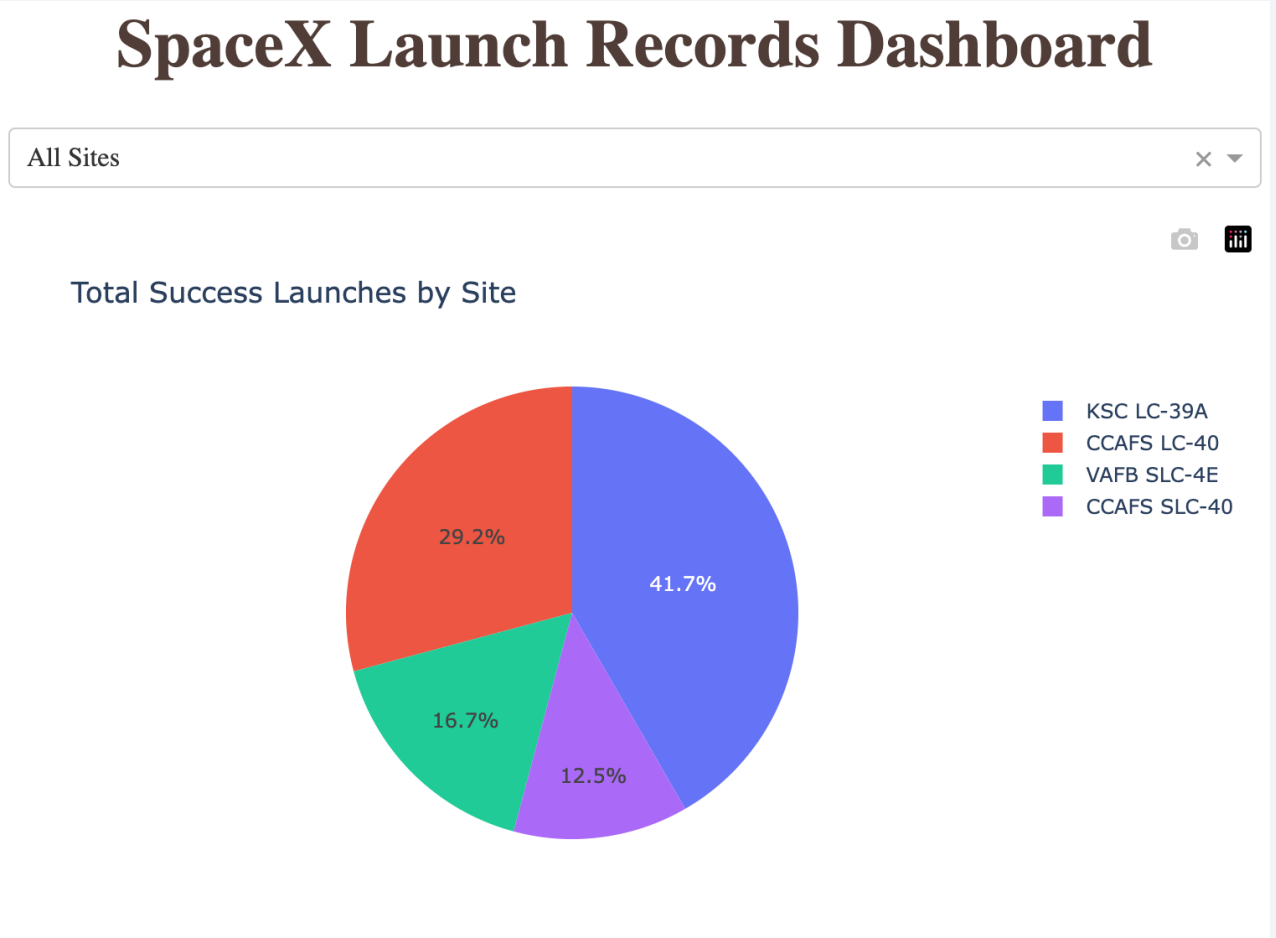



Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites

Total Success Launches for All Sites is

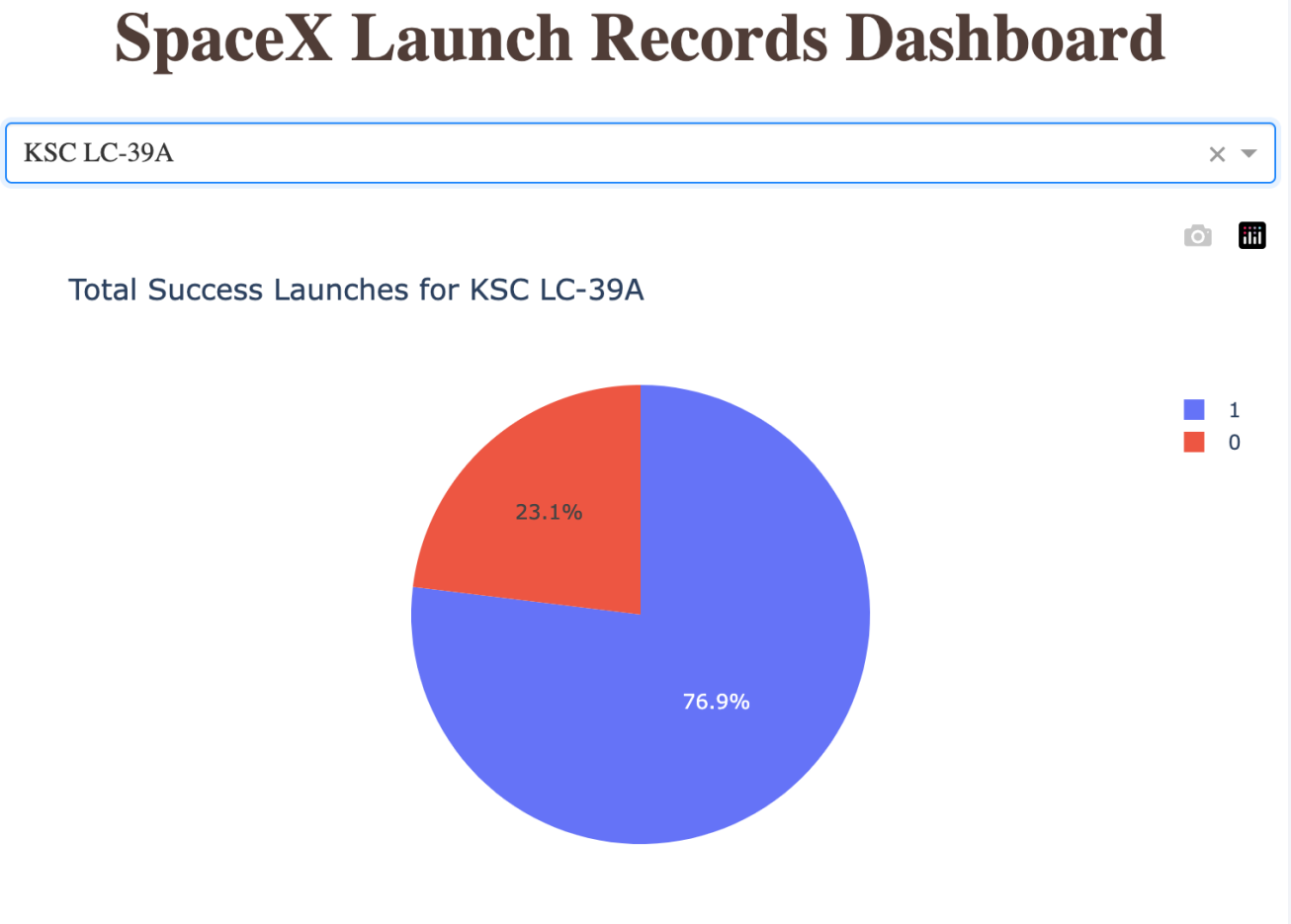
- CCAFS LC-40: 29.2%
- VAFB SLC-4E: 16.7%
- KSC LC-39A: 41.7%
- CCAFS SLC-40: 12.5%



Success Ratio for KSC LC-39A

The launch site with highest launch success ratio is KSC LC-39A.

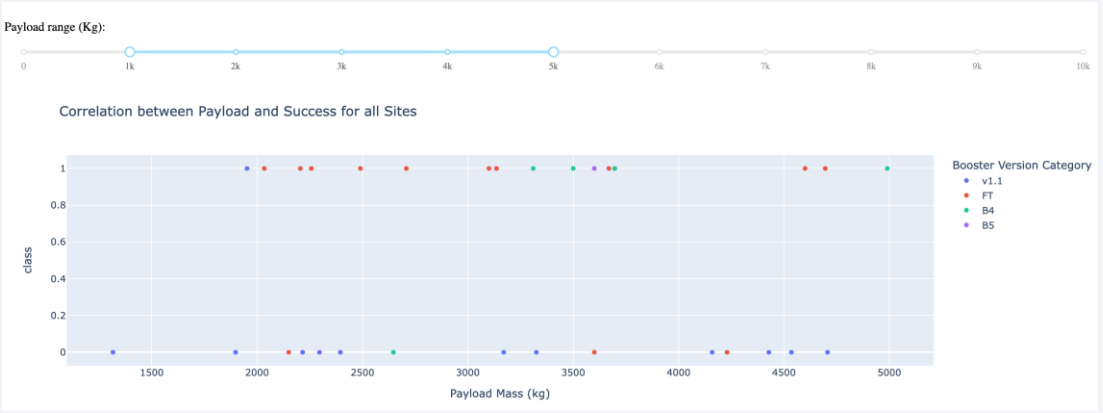
It has a success rate of 76.9%.



Correlation Between Payload and Success

Payload range in [3000, 4000] has the largest success rate.

Booster version of FT has the largest success rate.



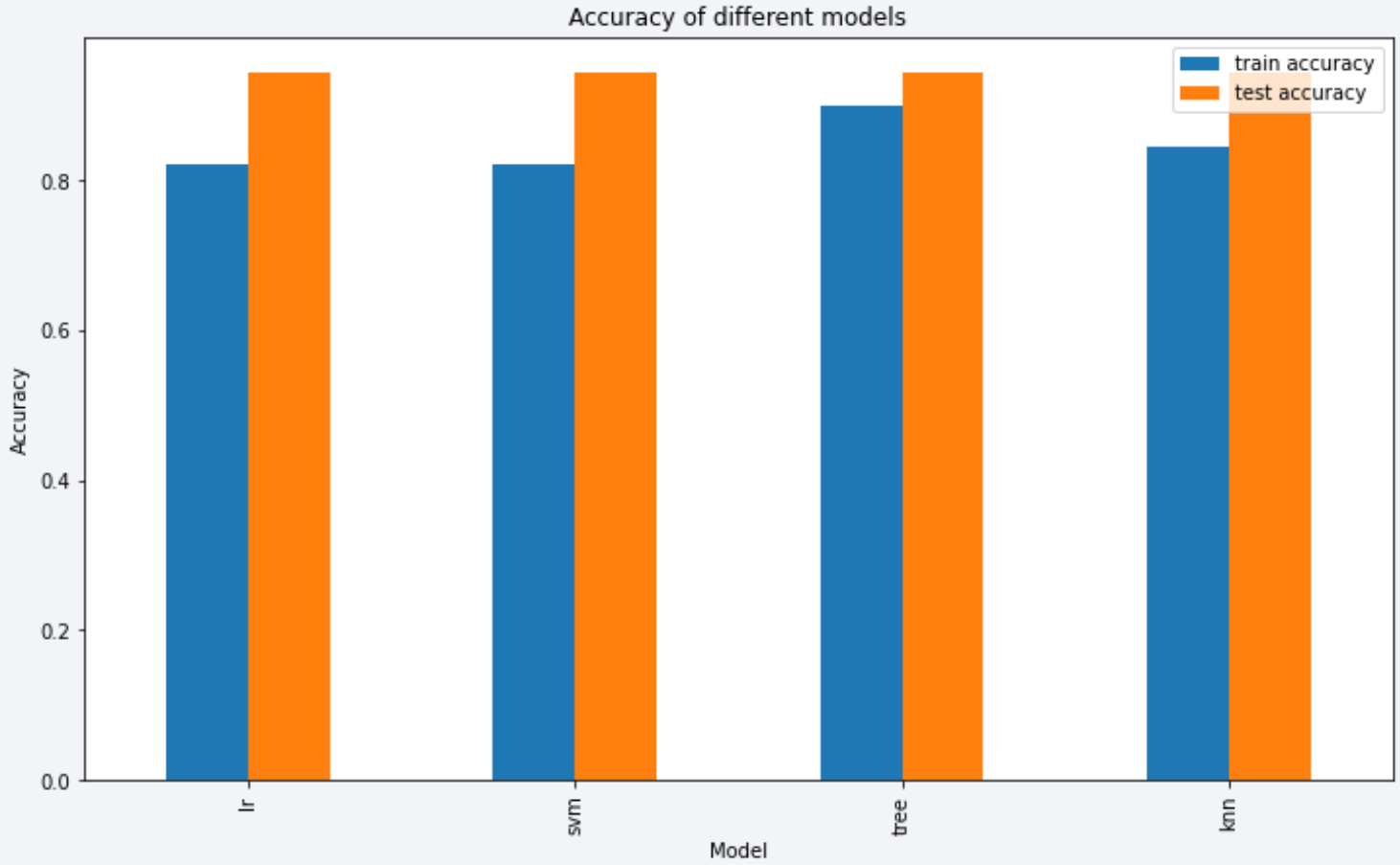
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Decision Tree model has the highest classification accuracy

training accuracy 0.9, testing accuracy 0.94



Confusion Matrix

Decision Tree model can distinguish between the different classes.

The major problem is **false positives**.

Conclusions

- The shape of the dataset is 90 x 83. With a 80/20 split this gives 72 rows for training data and 12 rows for testing data.
- And enhanced by GridSearchCV, we trained four models which have all best performance on test data set.
- Of these models, we can choose Decision Tree as our best model for predicting landing outcome of rocket. However there is an issue with false positives which probably will impact our estimation of next bid for rocket launch.

Thank you!