IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Md. Samrat-Uz-Zaman
13/04/2024

Outline

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

Executive Summary

- Summary of methodologies
  - The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:
  - Collect data using SpaceX REST API and web scraping techniques
  - Wrangle data to create success/fail outcome variable
  - Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
  - Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
  - Explore launch site success rates and proximity to geographical markers
  - Visualize the launch sites with the most success and successful payload ranges
  - Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)
- Summary of all results

Exploratory Data Analysis:

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:

- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:

- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Project background and context

SpaceY, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceY can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Problems you want to find answers

How payload mass, launch site, number of flights, and orbits affect first-stage landing success
What is the rate of successful landings over time
What is the best predictive model for successful landing (binary classification)

Section 1

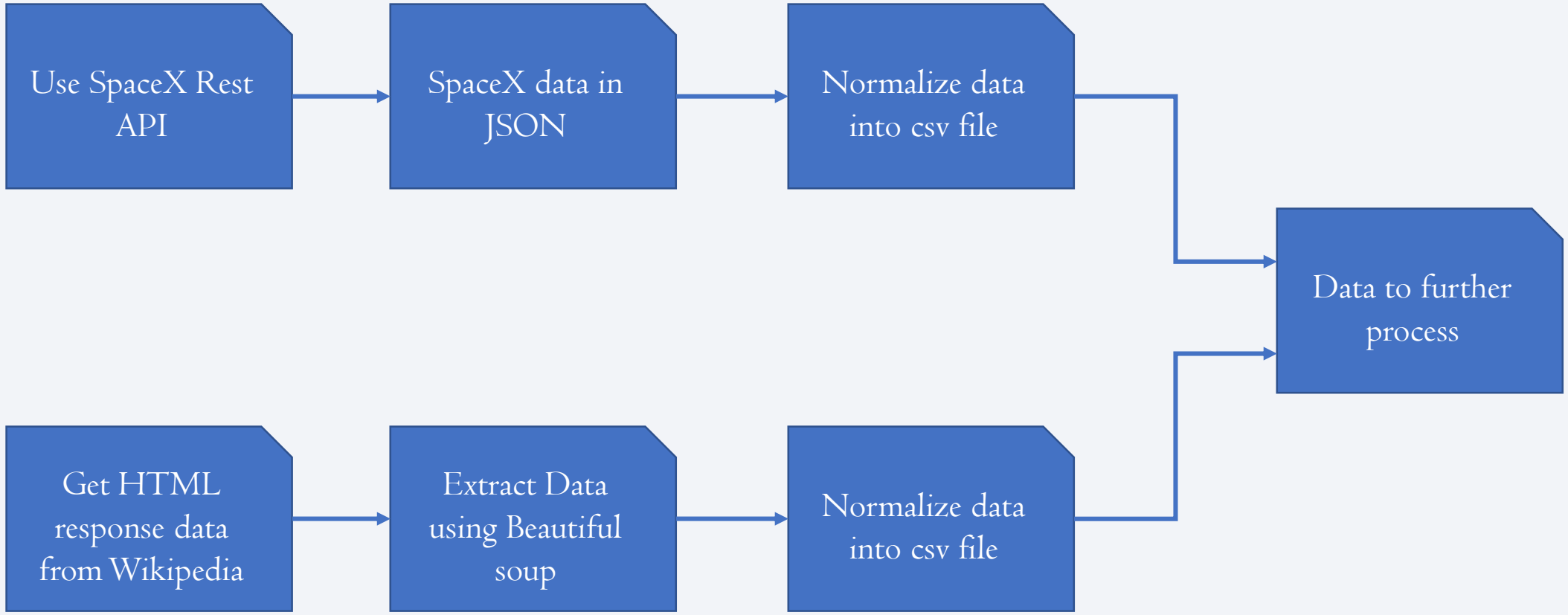
Methodology

Methodology

Executive Summary

- **Data collection methodology**
  - SpaceX REST API
  - web scraping techniques
- **Perform data wrangling**
  - filtering the data
  - handling missing values
  - applying one hot encoding
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - EDA with SQL
  - Data visualization techniques
- **Perform interactive visual analytics using Folium and Plotly Dash**
  - Interactive map with Folium
  - Interactive dashboard with Plotly Dash
- **Perform predictive analysis using classification models**
  - Build Models to predict landing outcomes using classification models.
  - Tune and evaluate models to find best model and parameters

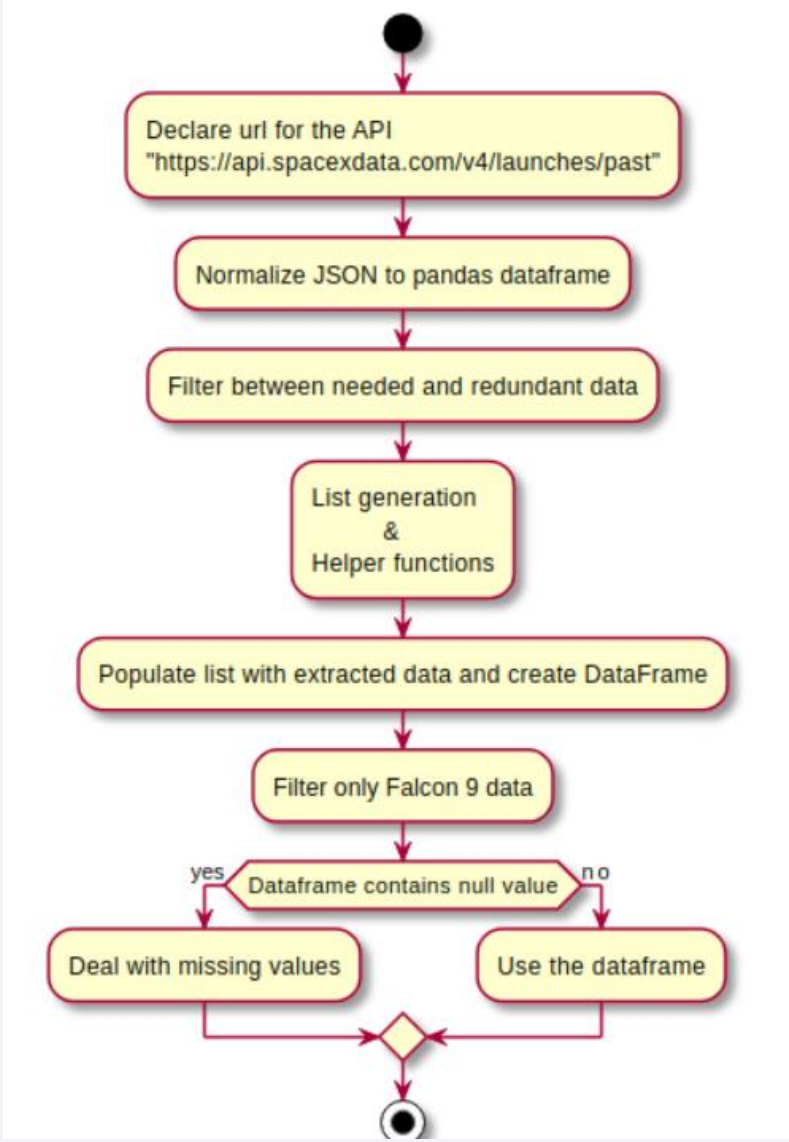
Data Collection



Data Collection – SpaceX API

- Data collection with SpaceX REST calls.
  1. https://github.com/r-spacex/SpaceX-API
  2. https://api.spacexdata.com/v4/rockets/
  3. https://api.spacexdata.com/v4/launchpads/
  4. https://api.spacexdata.com/v4/payloads/
  5. https://api.spacexdata.com/v4/cores/
- Github link

  https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W1/solved-jupyter-labs-spacex-data-collection-api.ipynb



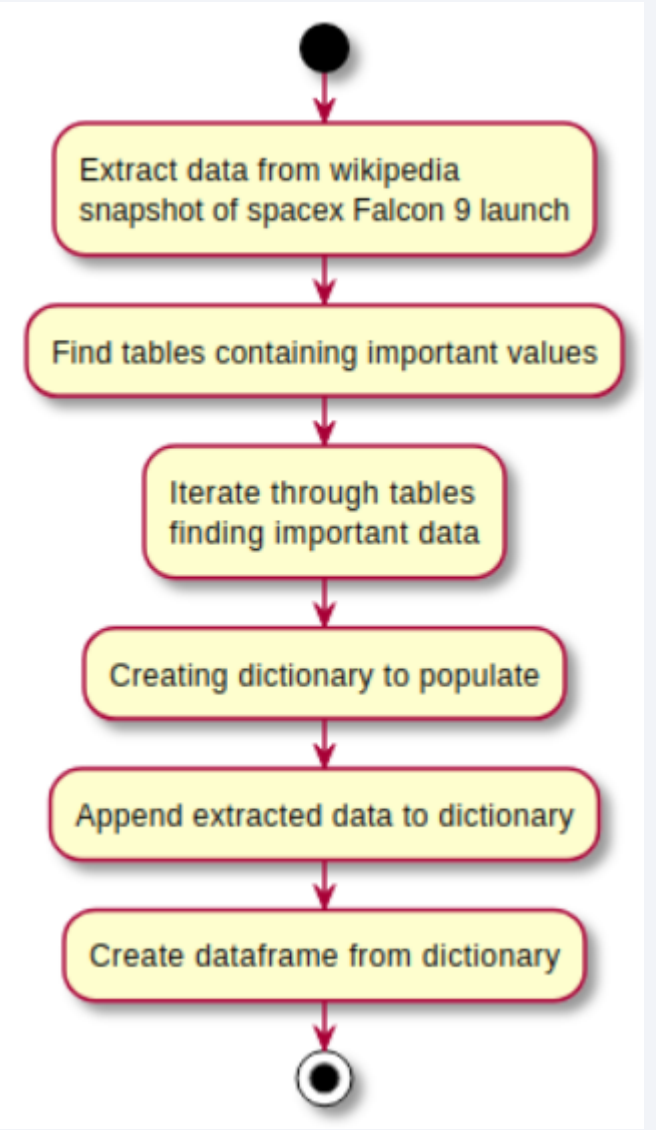
Data Collection - Scraping

- Web Scrapping URL

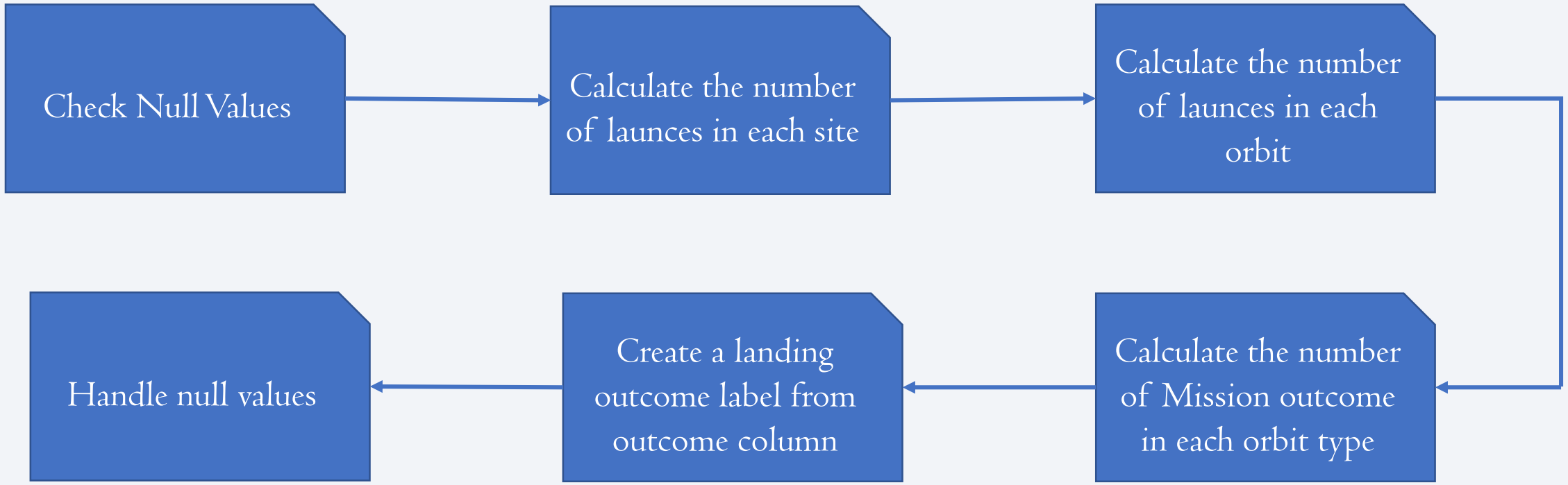
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Github URL:

  https://github.com/saturn-sam/IBM_Data_Science/blob/b0a384I7a03b0e22d0f37db279deacf6b7c46b6b/C10/W1/solved-jupyter-labs-webscraping.ipynb



Data Wrangling



Github URL:
https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W1/solved-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts Used:
  - Flight Number vs. Payload
  - Flight Number vs. Launch Site
  - Payload Mass (kg) vs. Launch Site
  - Payload Mass (kg) vs. Orbit type
- EDA with Visualization Analysis
  - View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
  - Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.
- Github URL:

  https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W2/solved-edadataviz.ipynb

EDA with SQL

SQL Query Performed

1. Displaying the names of the unique launch sites in the space mission
2. Displaying 5 records where launch sites begin with the string 'KSC'
3. Displaying the total payload mass carried by boosters launched by NASA (CRS)
4. Displaying average payload mass carried by booster version F9 v1.1
5. Listing the date where the successful landing outcome in drone ship was achieved.
6. Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
7. Listing the total number of successful and failure mission outcomes
8. Listing the names of the booster_versions which have carried the maximum payload mass.
9. Listing the records which will display the month names, successful landing_outcomes in ground pad, booster
10. versions, launch_site for the months in year 2017
11. Ranking the count of successful landing_outcomes between the date 2010 06 04 and 2017 03 20 in descending order.

Github URL:

https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W2/solved-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates Map with Folium

Colored Markers of Launch Outcomes

- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC- 40 and its proximity to the nearest coastline, railway, highway, and city

Github URL:

https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W3/solved-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Slider of Payload Mass Range

- Allow user to select payload mass range

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

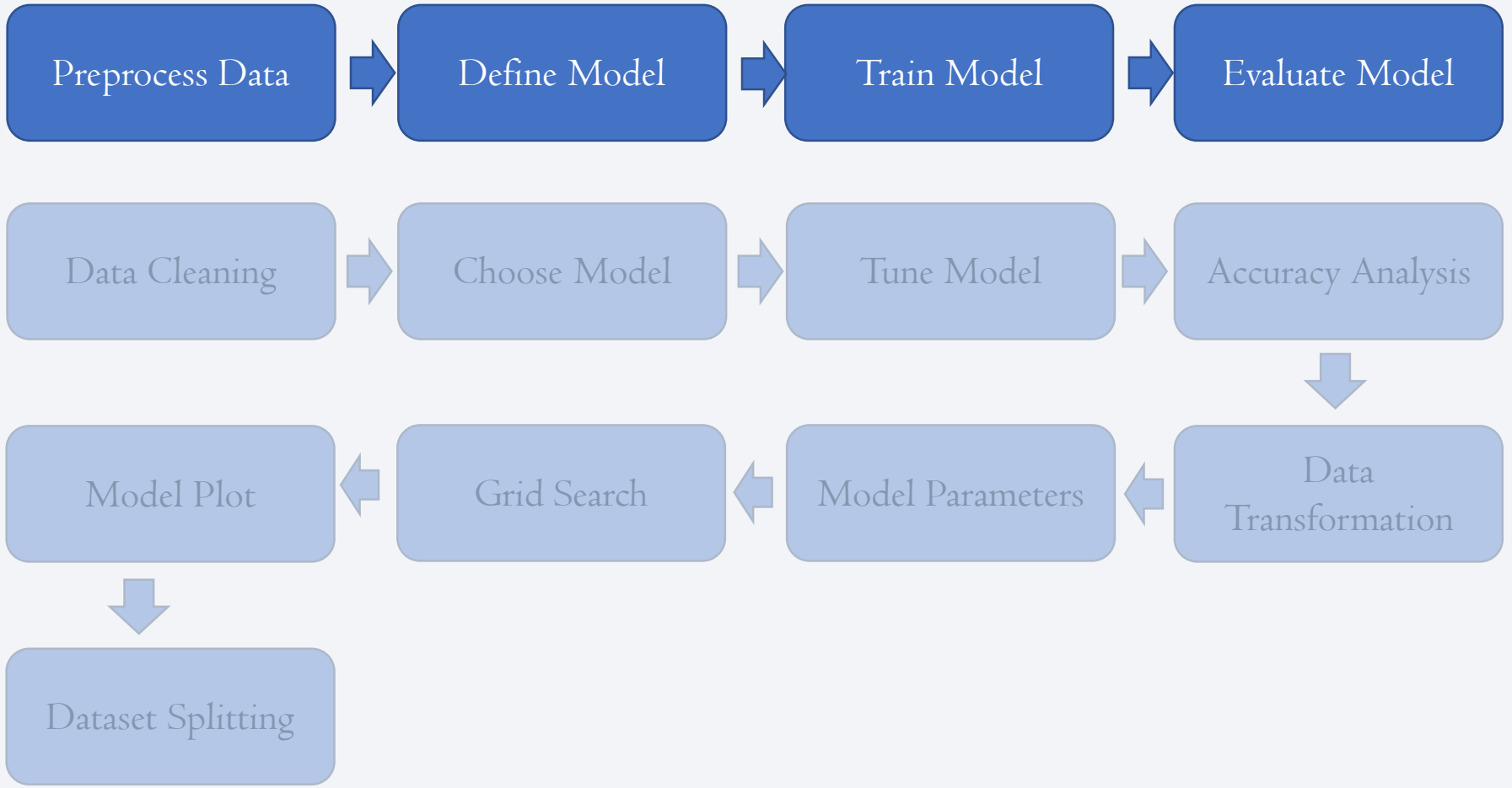
- Allow user to see the correlation between Payload and Launch Success

Github URL:

https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W3/spacex_dash_app.py

Predictive Analysis (Classification)

- ❑ **Prepare** data
- ❑ **Create** a column for the class
- ❑ **Standardize** the data
- ❑ **Split** into training data and test data
- ❑ **Define** model and parameters
- ❑ **Train** and **Grid Search** for best parameters
- ❑ **Evaluation**



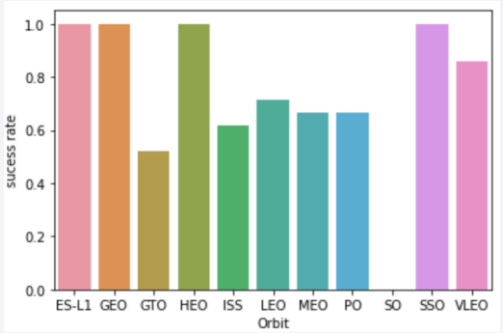
Github Link:

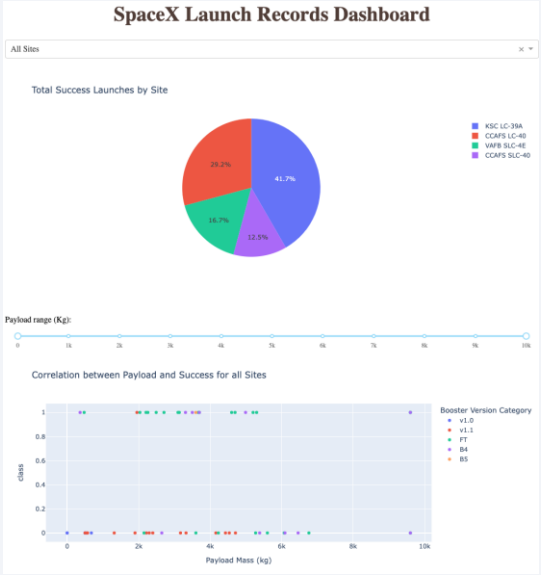
https://github.com/saturn-sam/IBM_Data_Science/blob/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10/W4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory Data Analysis

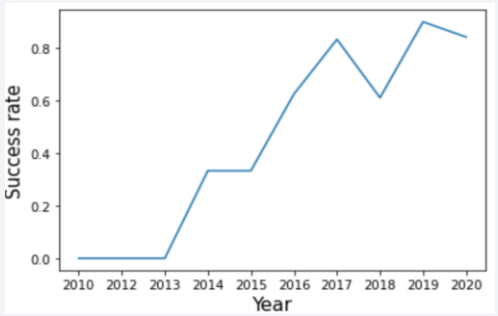
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

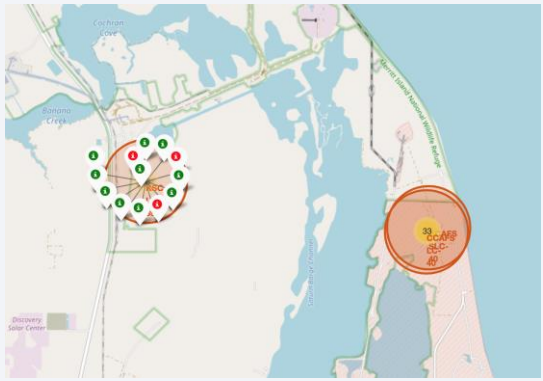




Visual Analytics

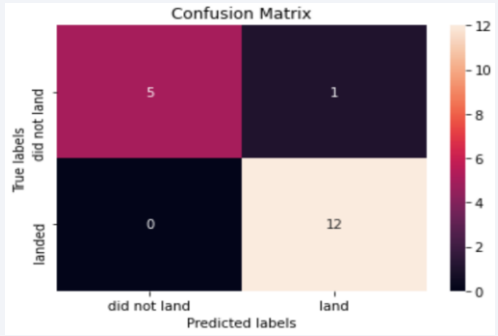
- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

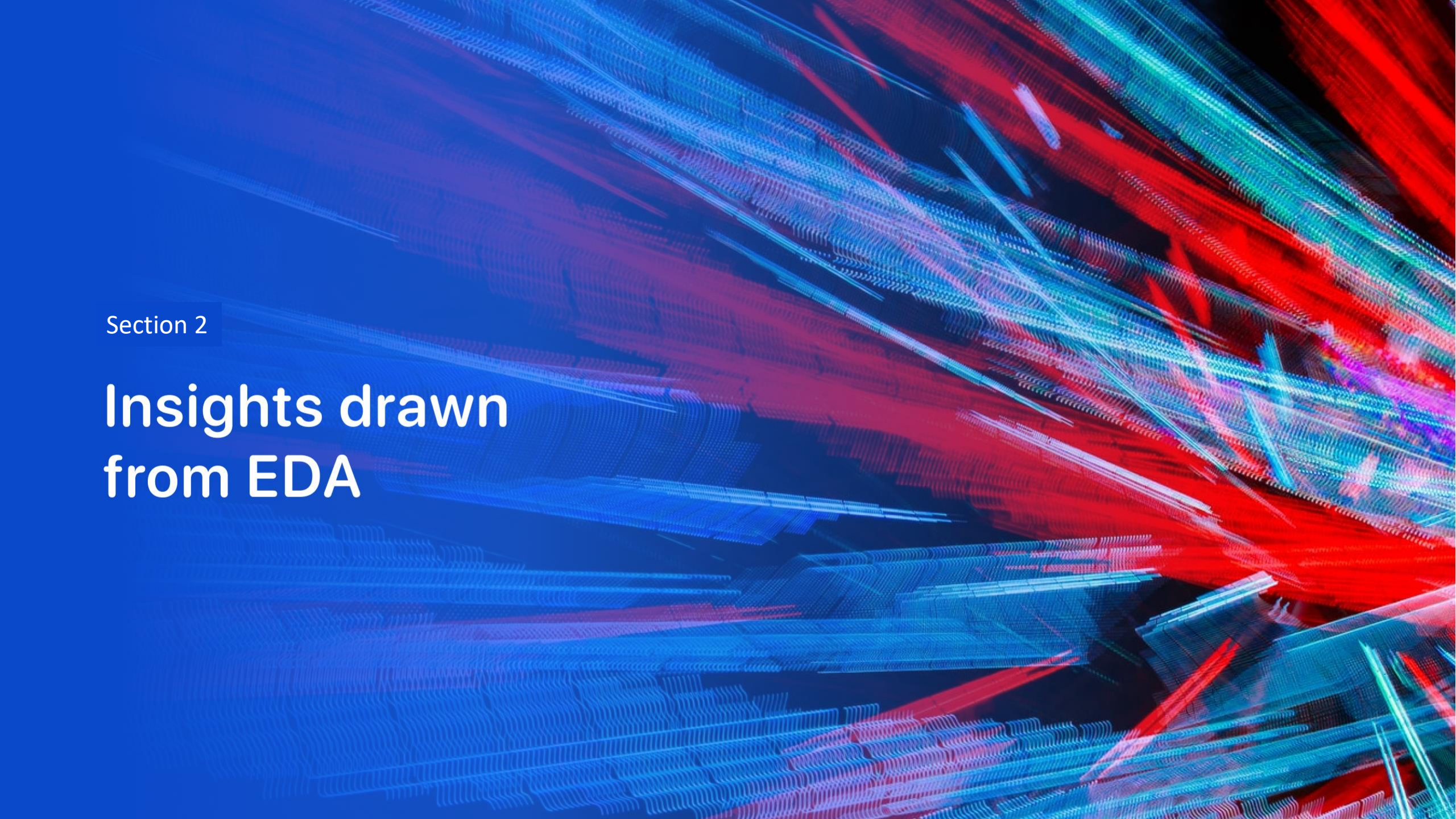




Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

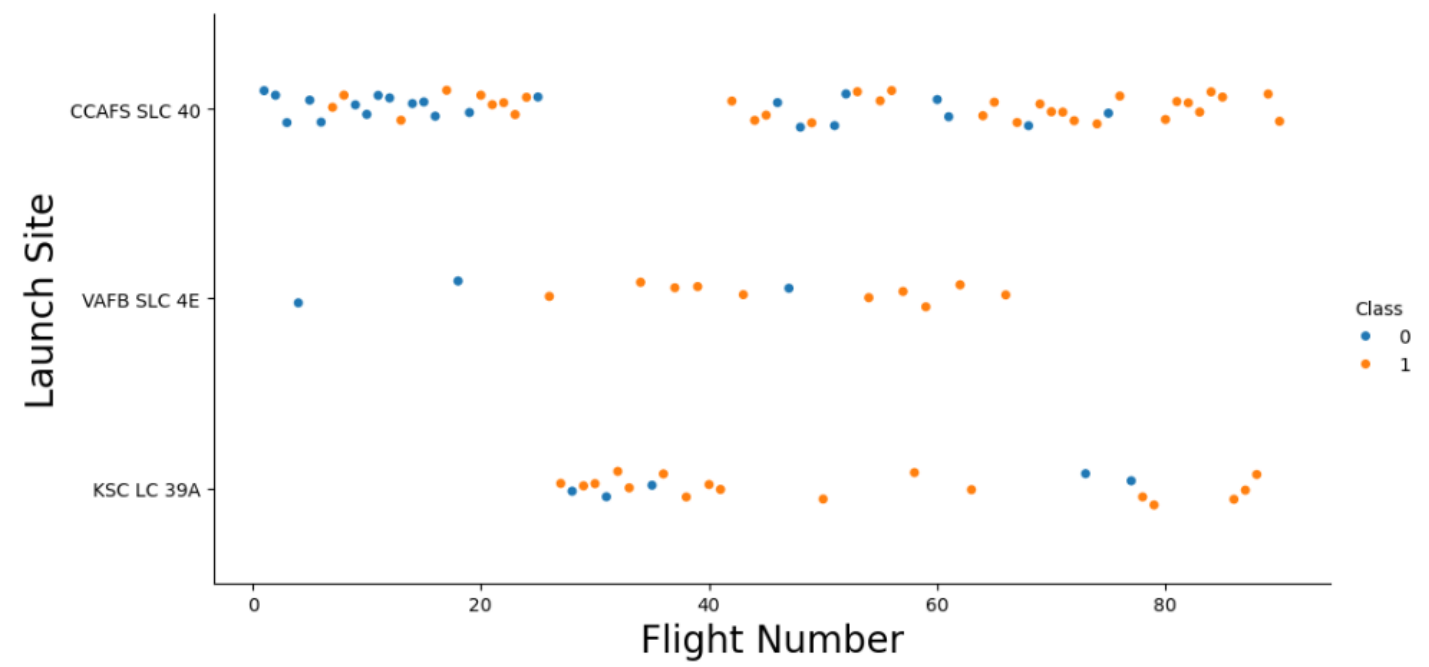



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis:

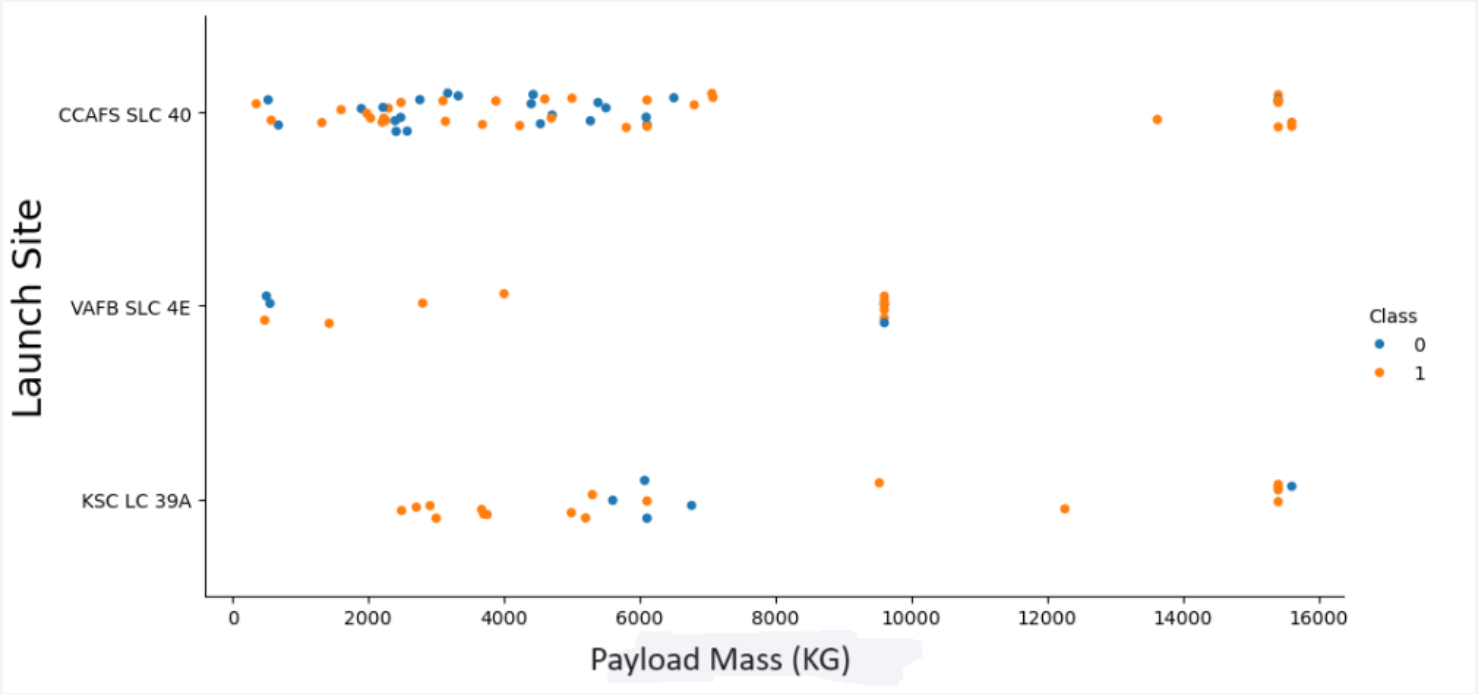
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

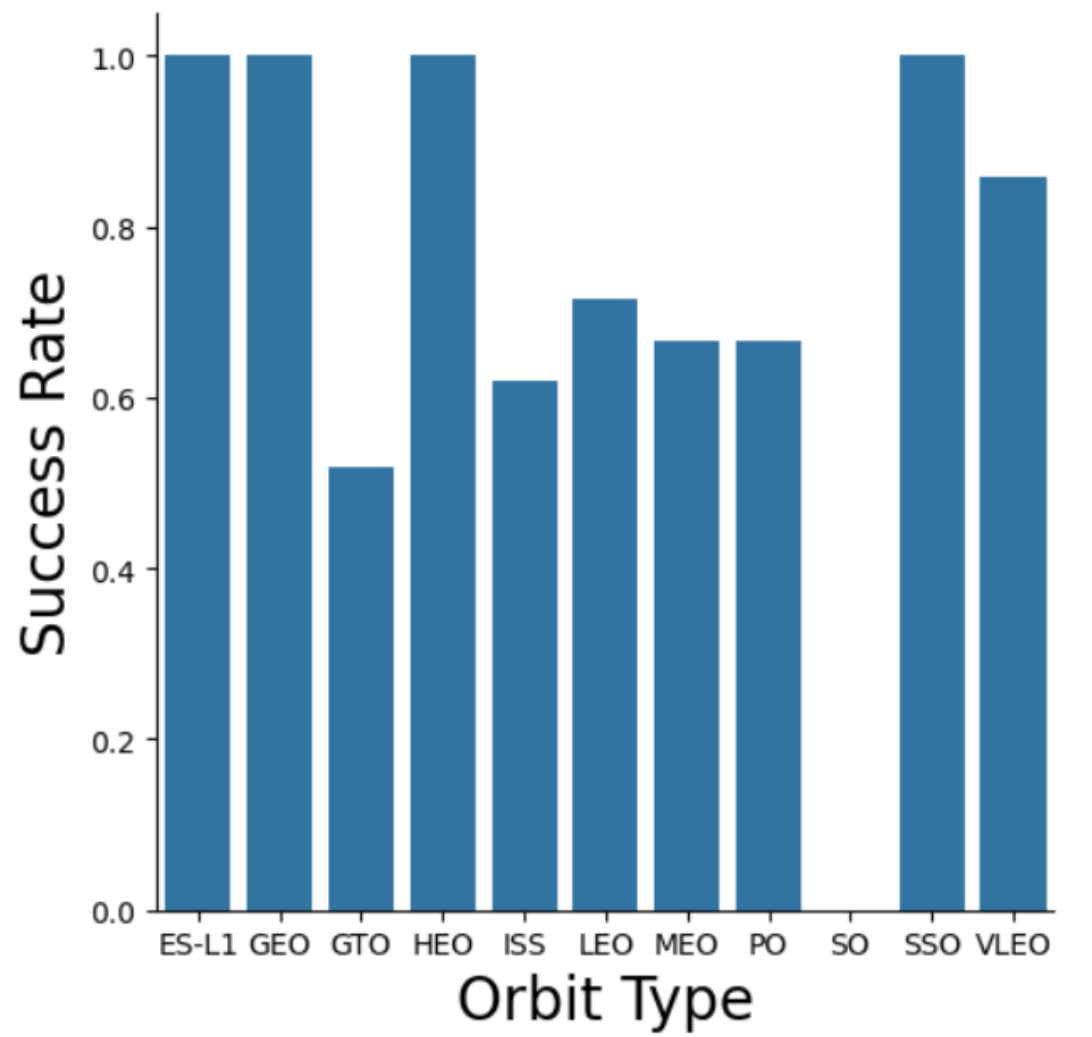
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

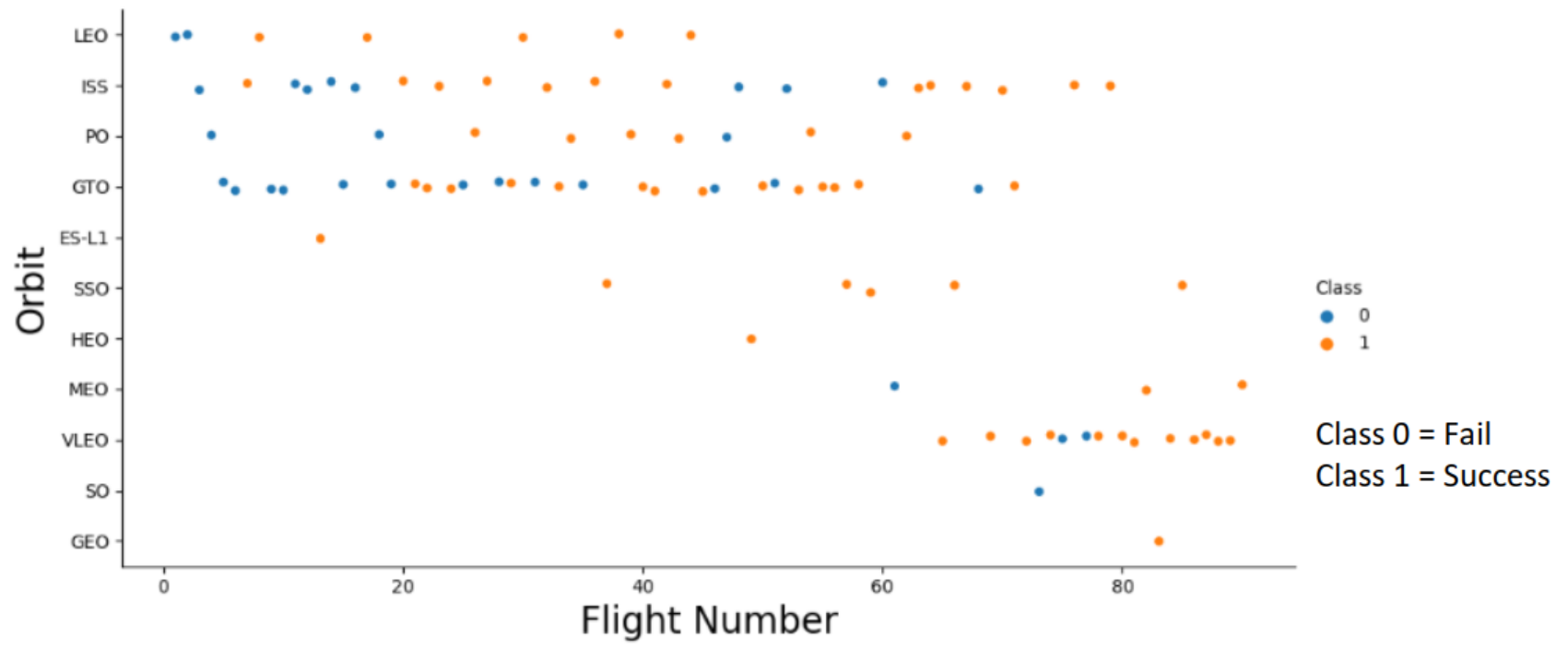
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

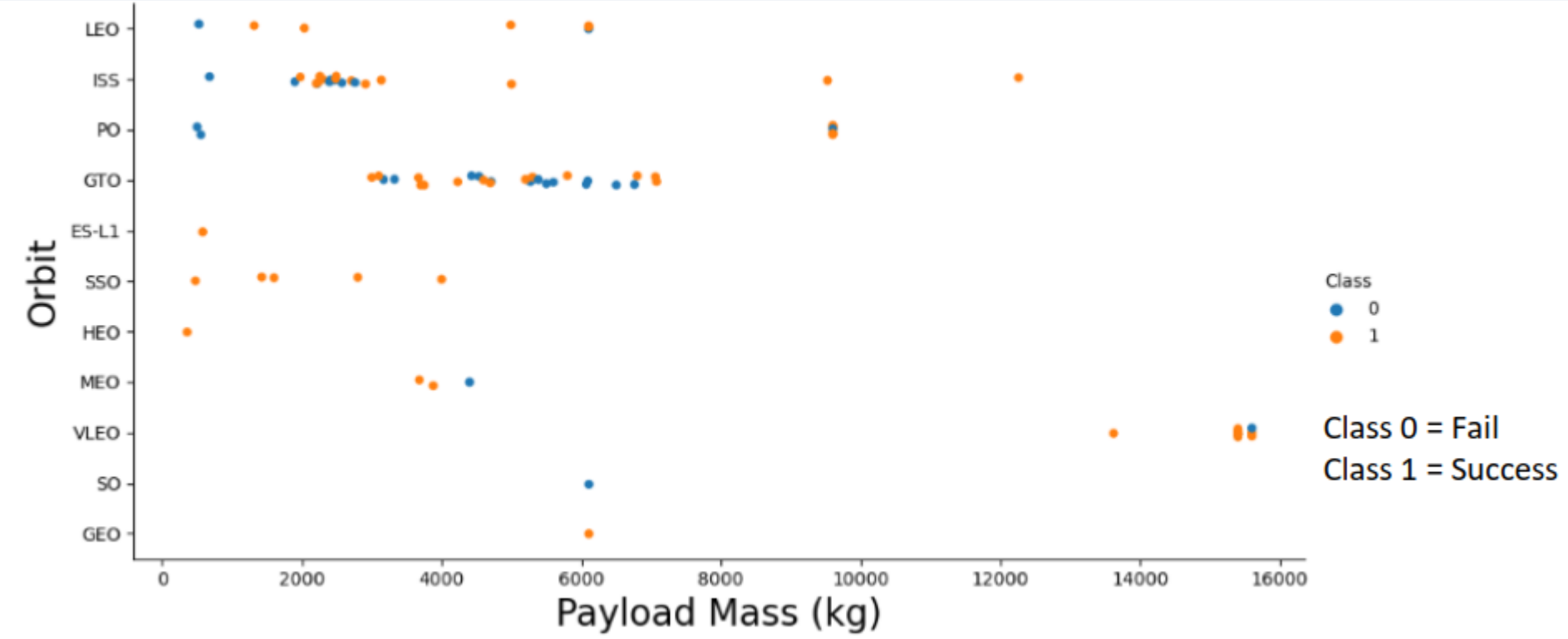
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

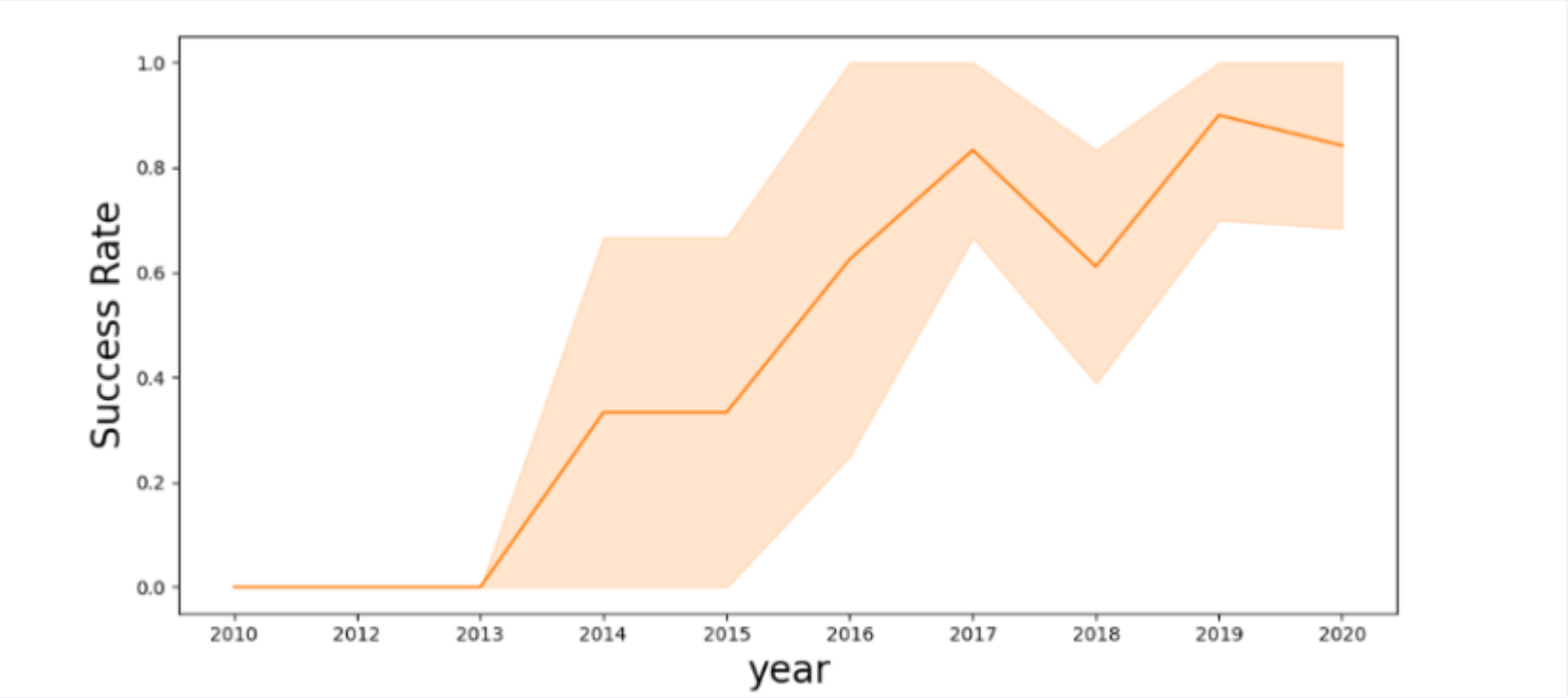
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

Launch Site Names

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

```
In [13]: %sql select distinct Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[13]:

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

Launch Site Names Begin with 'CCA'

Launch Site Names begin with 'CCA'

- CCAFS LC-40
- CCAFS SLC-40

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

In [18]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Out[18]:

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-	0:35:00	F9 v1.0 B0006	CCAFS LC-	SpaceX	500	LEO	NASA	Success	No attempt

Total Payload Mass

Total Payload Mass

- 45,596 kg (total) carried by busters launched by NASA (CRS)

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

```
In [23]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

Out[23]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average Payload Mass

- 2,928 kg (average) carried by booster version F9 v1.1

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

In [24]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

Out[24]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

1st Successful Landing in Ground Pad

- 12/22/2015

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

Hint:Use min function

In [26]:
```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

Out[26]:

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Explanation:** 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

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 SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000
```

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

Out[28]:

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

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

List the total number of successful and failure mission outcomes

In [29]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Out[29]:

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

Boosters Carried Maximum Payload

Carrying Max Payload

- 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

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

```
%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

2015 Launch Records

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

```
[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

Rank Landing Outcomes Between 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.
```

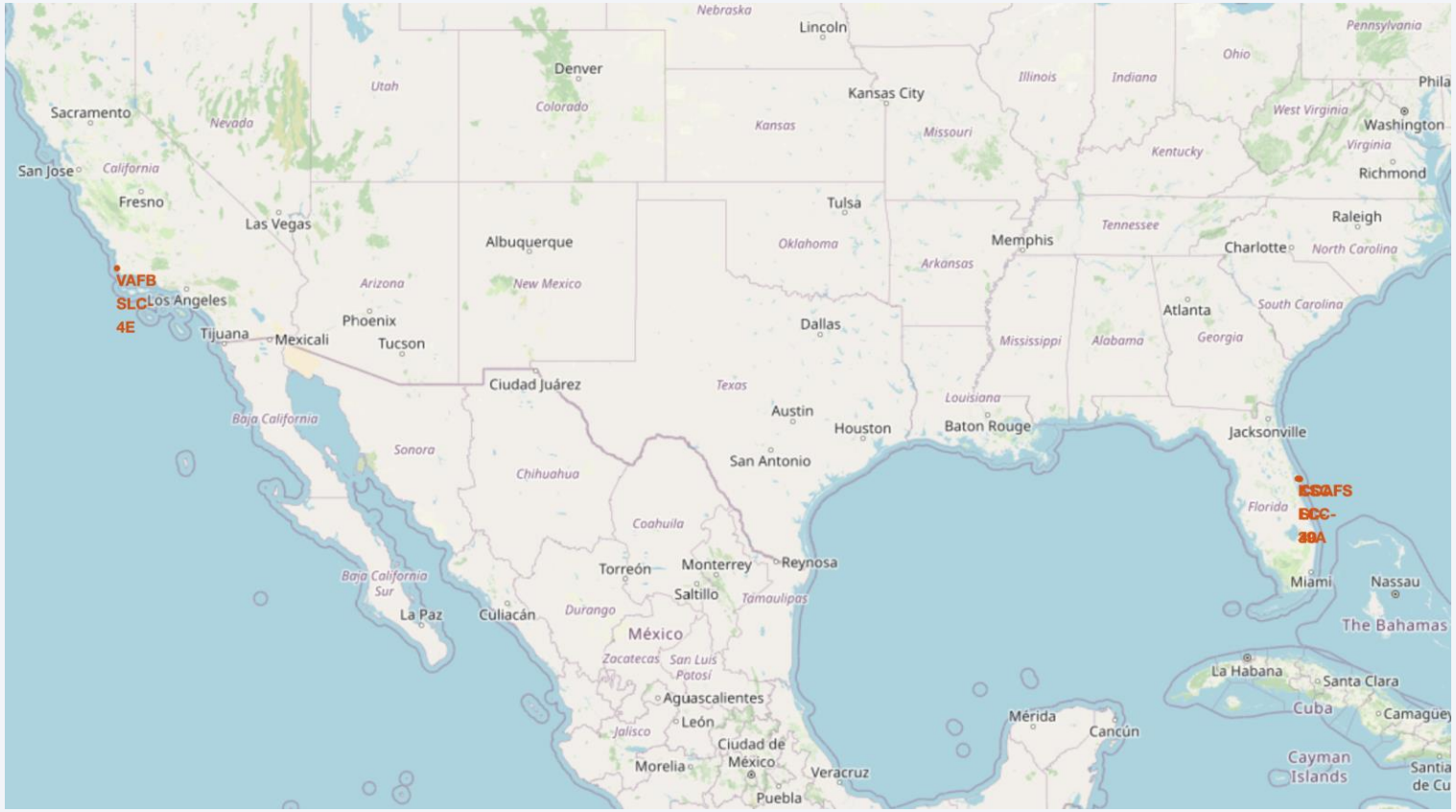
Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Locations of Launch Sites on Maps

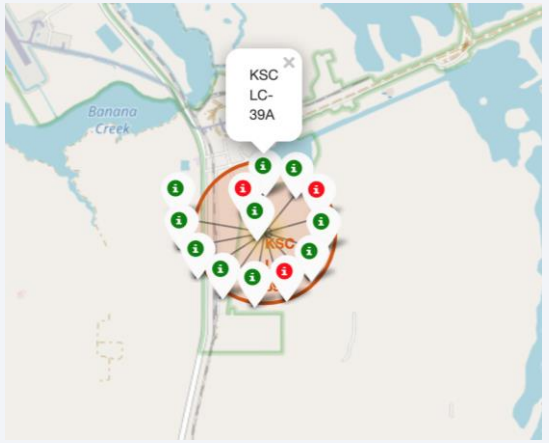
Launch Site	Lat	Long
CCAFS LC-40	28.56230197	-80.57735648
CCAFS SLC-40	28.56319718	-80.57682003
KSC LC-39A	28.57325457	-80.64689529
VAFB SLC-4E	34.63283416	-120.6107455

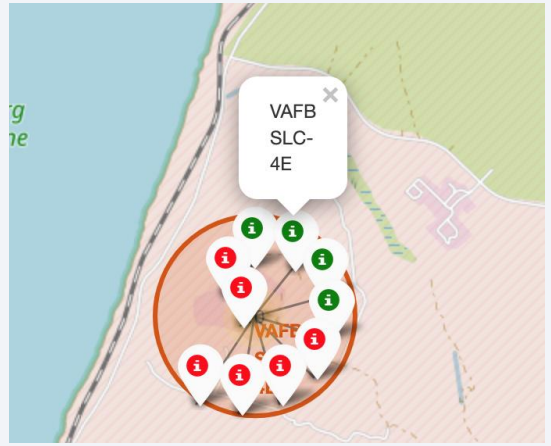


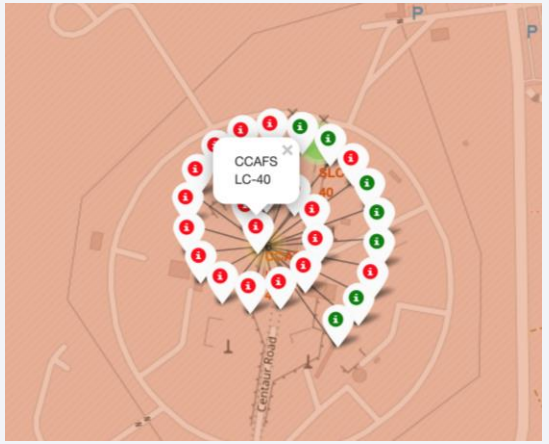
Display Launch Outcome by Color

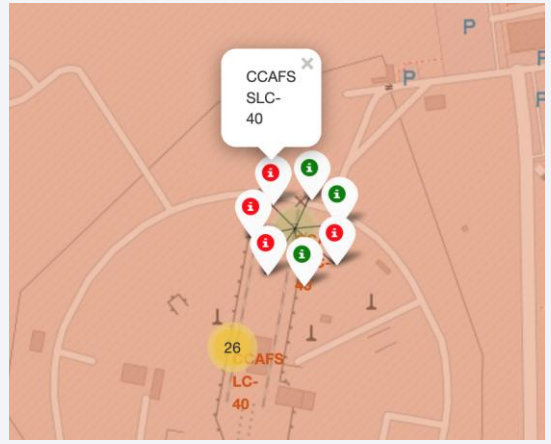
From the color labels, we can easily see

- KSC LC-39A has a rather higher success rate
- Whereas CCAFS LC-40 and CCAFS SLC-40 have much lower rate



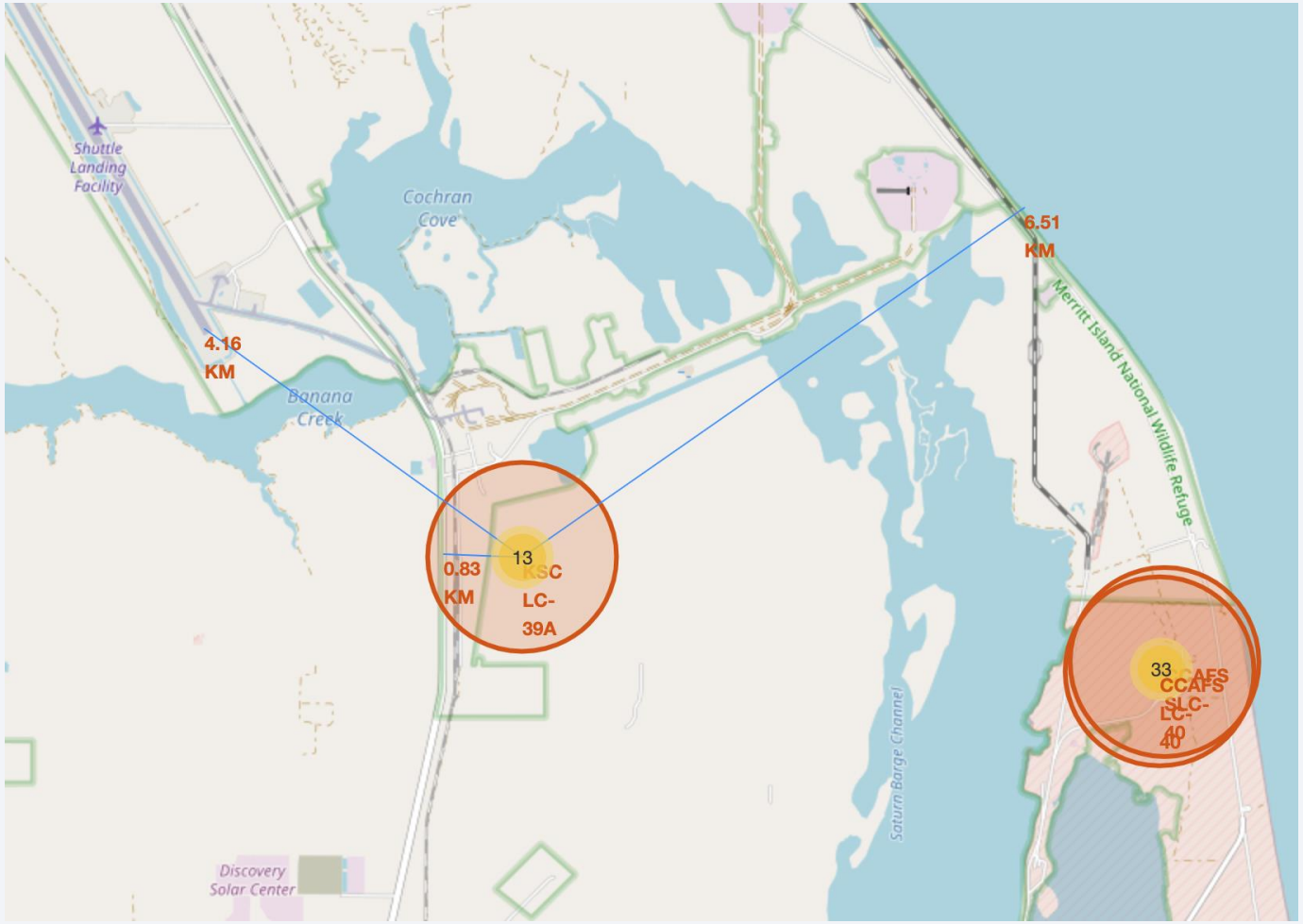


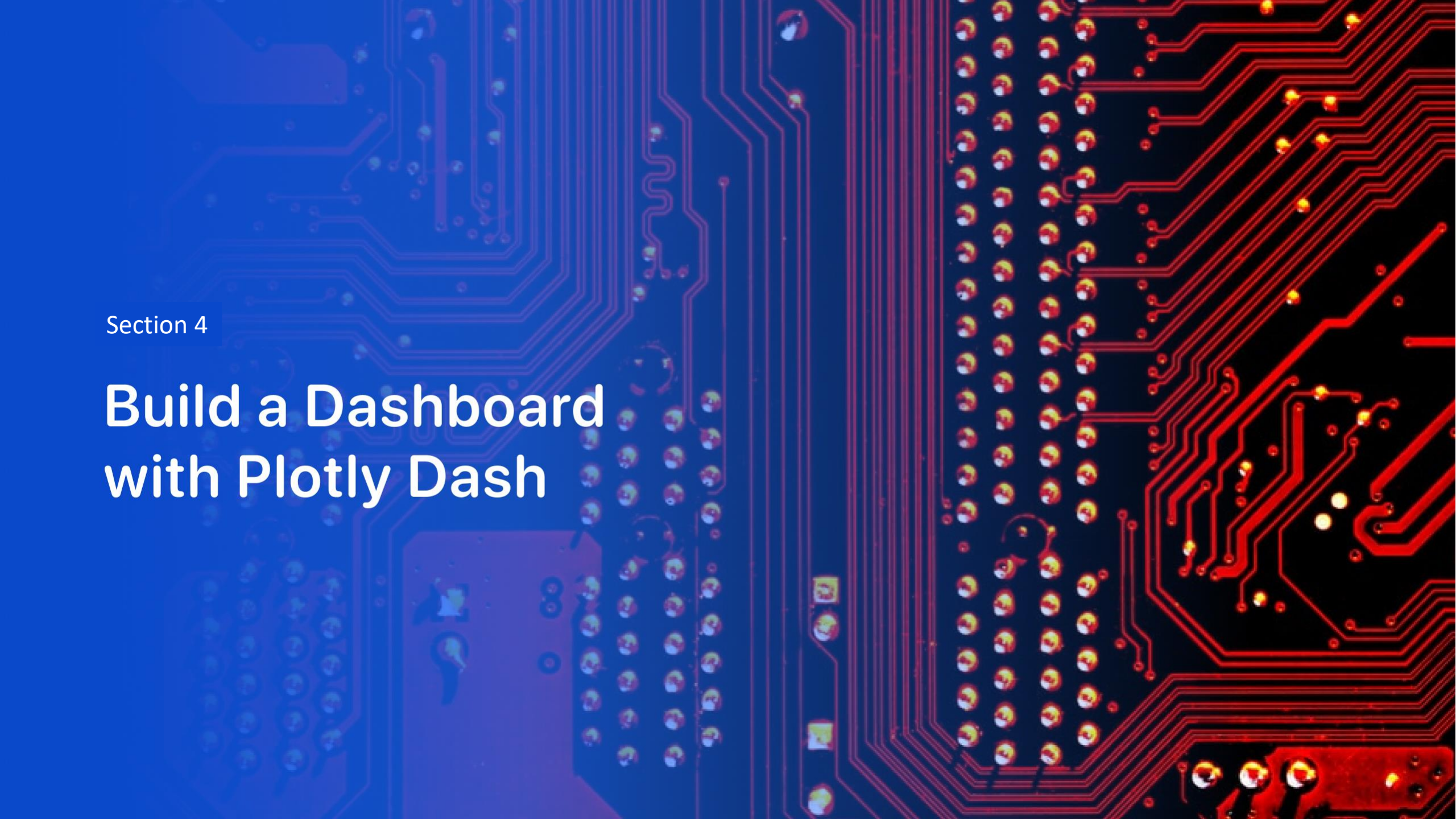




Show Distance to Proximities

- 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.

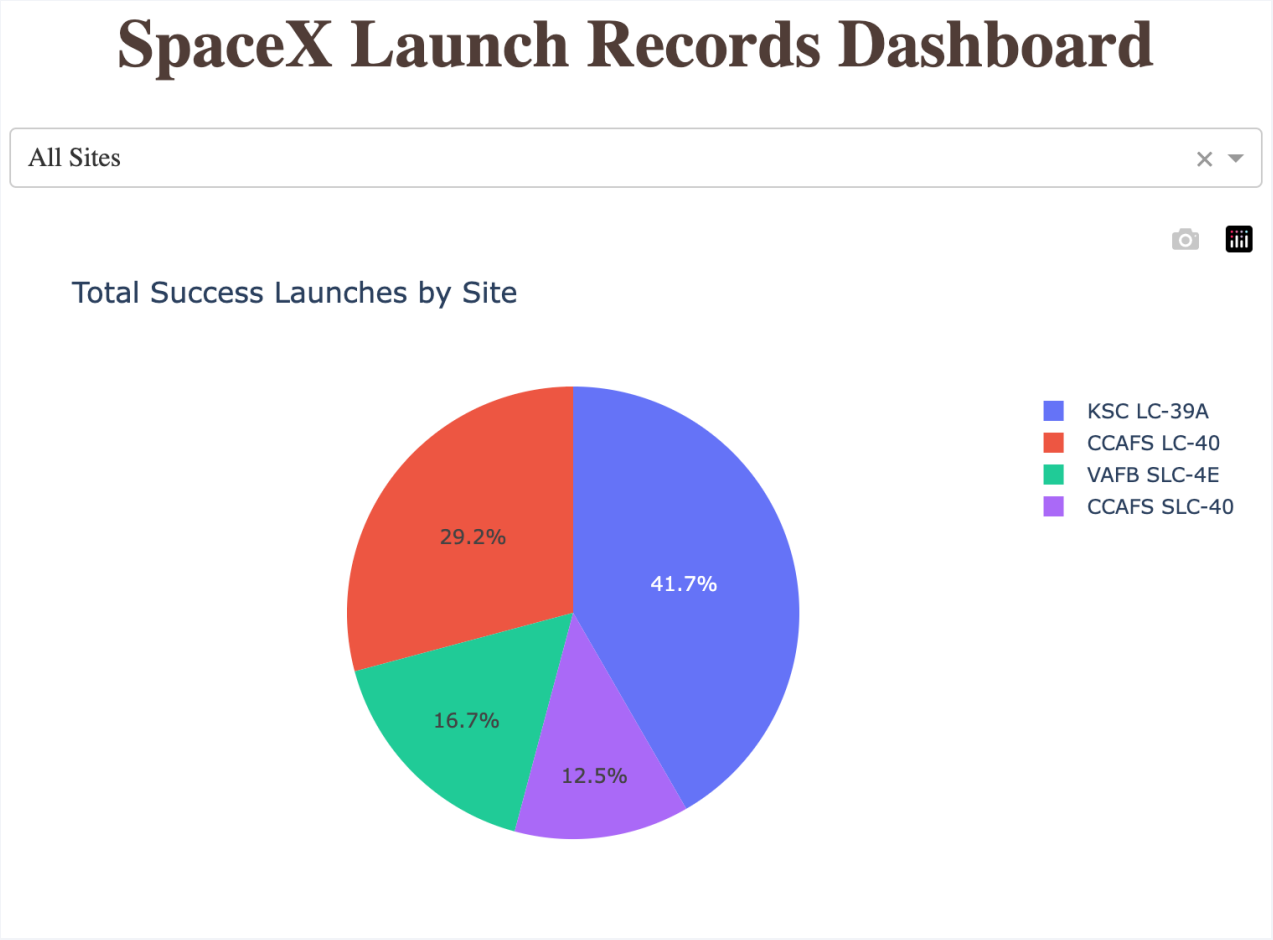



Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites

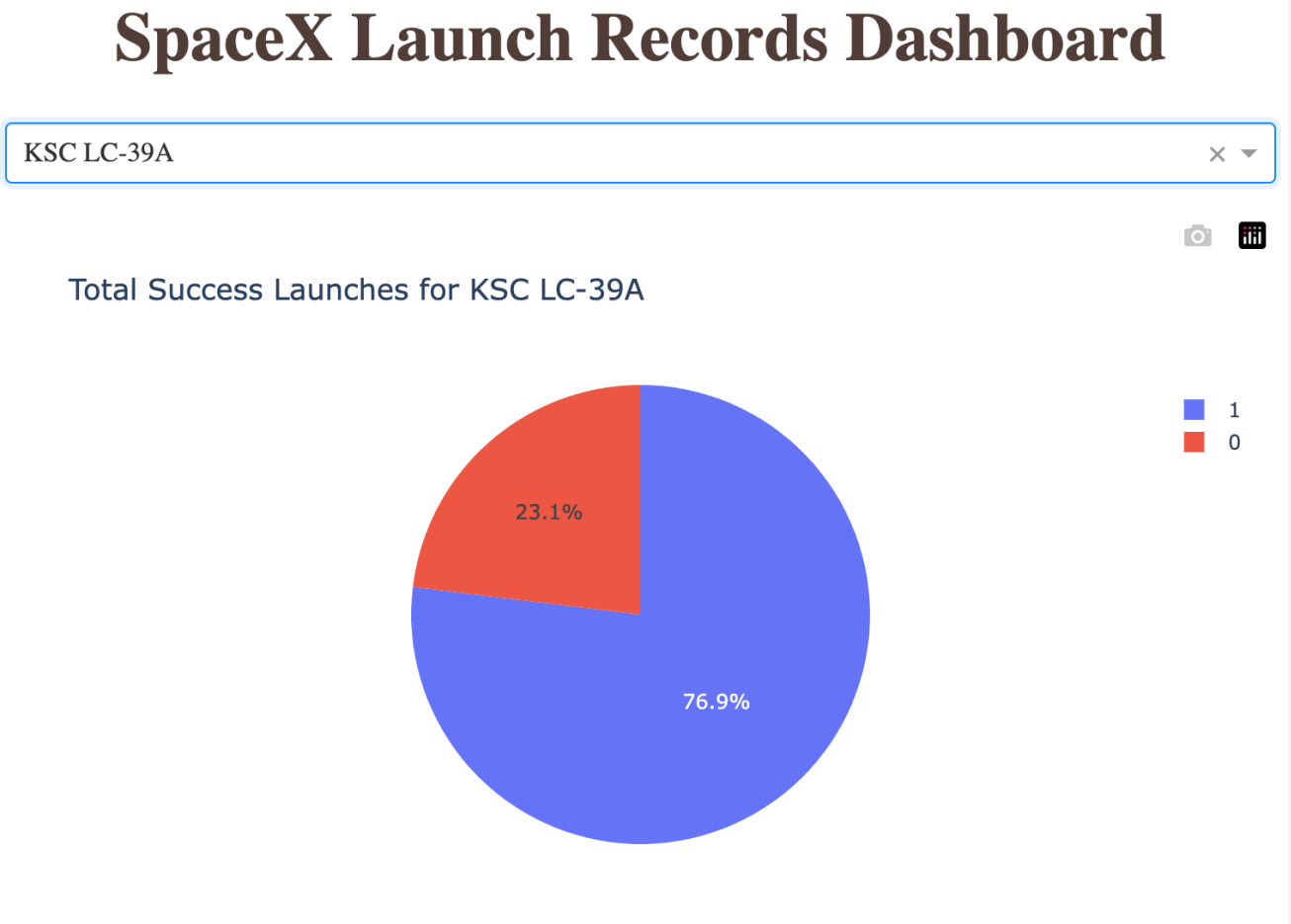
Total Success Launches for All Sites is

- 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

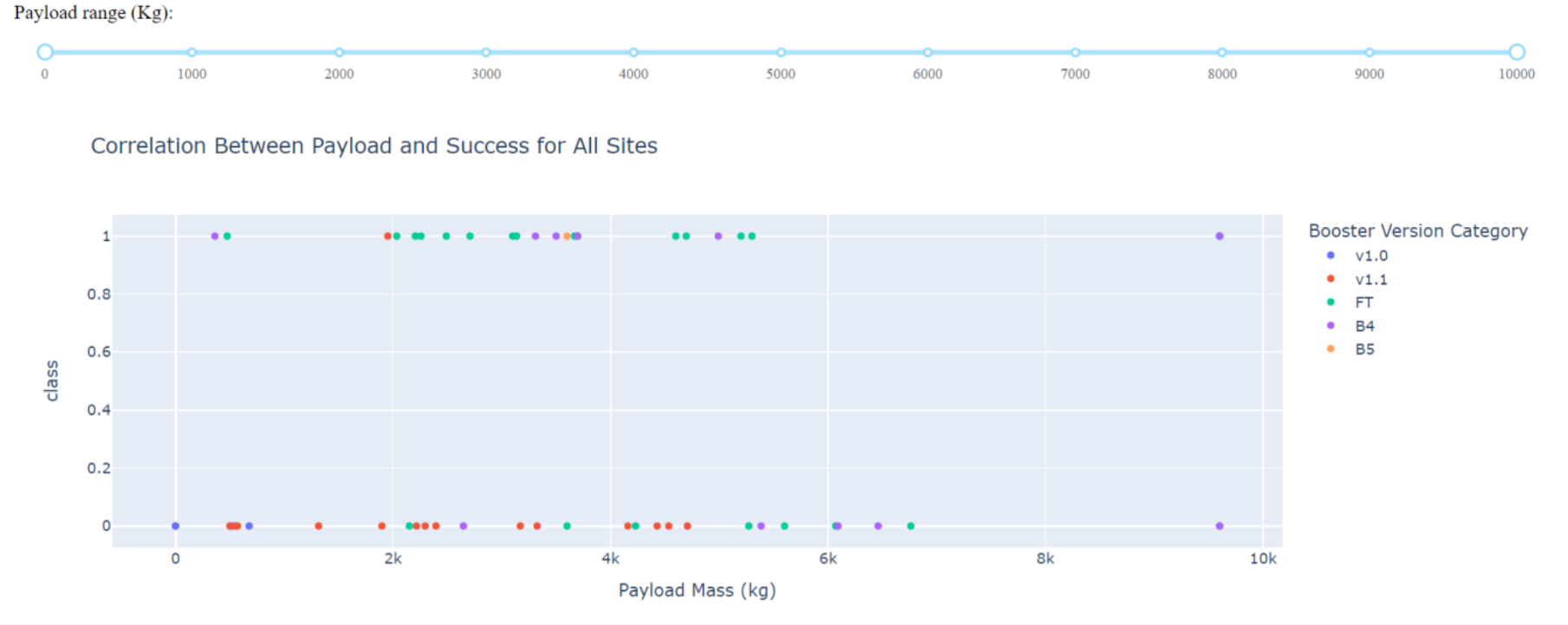
The launch site with highest launch success ratio is **KSC LC-39A**.
It has a success rate of **76.9%**.



Payload Mass and Success

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

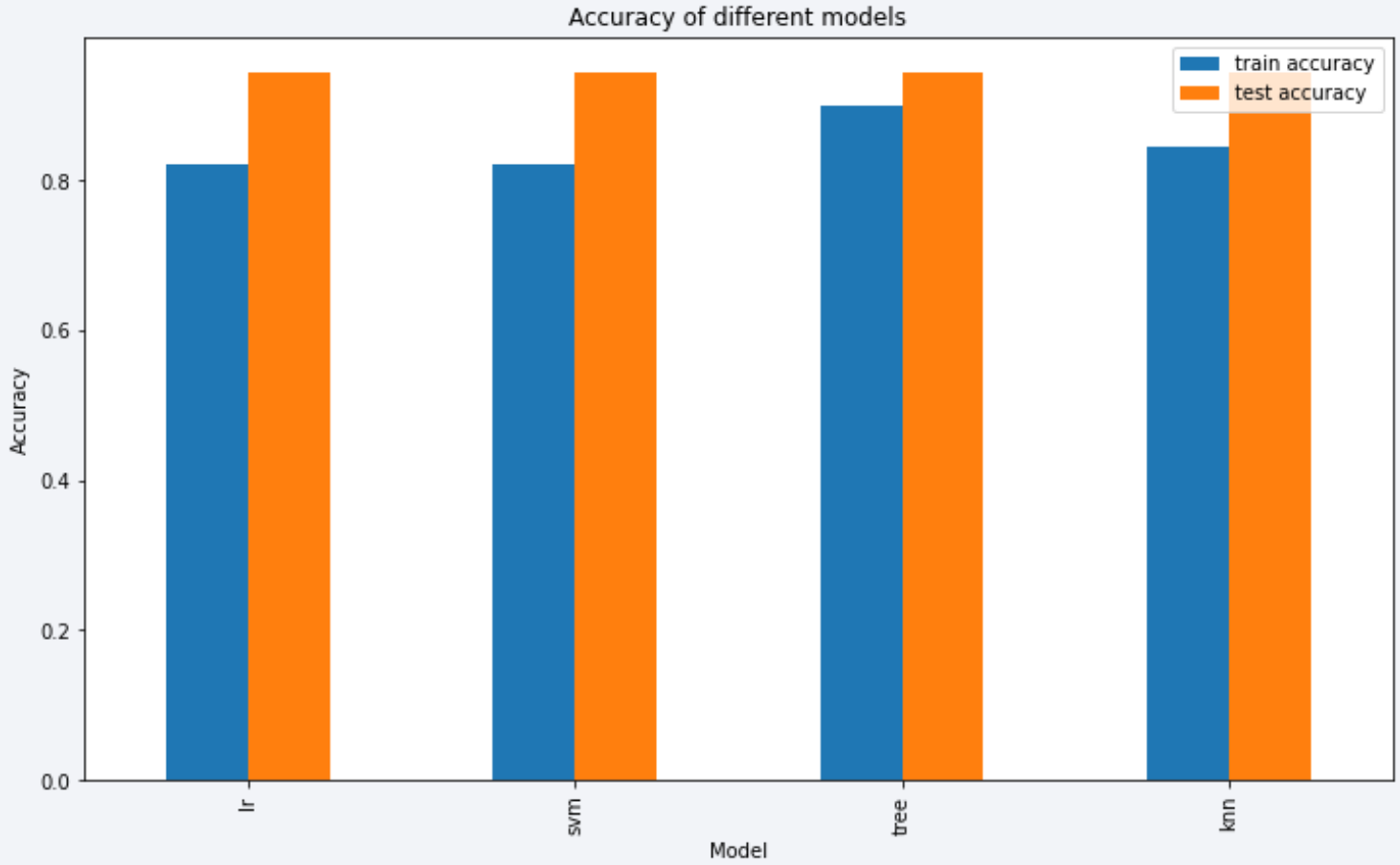


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Decision Tree model** has the highest classification accuracy
- Training accuracy 0.9, **testing accuracy 0.94**
- Parameter: {'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}



Confusion Matrix

- Decision Tree model can distinguish between the different classes.
- The major problem is **false positives**.

Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

My Github repo for the project

https://github.com/saturn-sam/IBM_Data_Science/tree/b0a38417a03b0e22d0f37db279deacf6b7c46b6b/C10

Thank you!