IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

RocketMan
01/12/24

Outline

1. Executive Summary
2. Introduction
3. Methodology
4. Results
    1. Insights from EDA
    2. Launch Site Proximity Analysis
    3. Dashboard with Plotly Dash
    4. Predictive Analysis
5. Conclusion
6. Appendix

Executive Summary

Historic SpaceX launch data was collected using both an API and through web scraping a wikipedia page. The data was cleaned and prepared for further analysis. Initial insights were plotted to further direct a deeper dive. Geographic launch site data was shown to understand proximities to infrastructure and finally a suite of predictive analyses were run to show the expected outcome of a launch based on a number of factors.

Results Summary:

1. Landing success is shown to have drastically improved between 2013 and 2020.
2. Launches from KSC LC-39A lead to greatest landing success rate.
3. Launch sites share common proximities: large easterly area of sparse/no population, critical infrastructure (roads & railways), towns nearby.
4. Predictive analysis shows that decision trees might be slightly overfitting. LR, SVM, and KNN all perform similarly on training and testing data with an overall test accuracy of 83.3% and no such indication of model overfitting.

Introduction

SpaceX is a leading supplier of inexpensive and reusable rockets capable of supplying payloads to any earth-centric and many intra-solar system orbits. When looking to open a competing company, predicting a successful landing will strongly drive the cost of launch.

Questions:

1. Explore how launch site, payload mass, destination orbit, and time affect landing success.
2. Evaluate four (LR, SVM, Decision Tree, KNN) models for predicting landing outcome.

Section 1

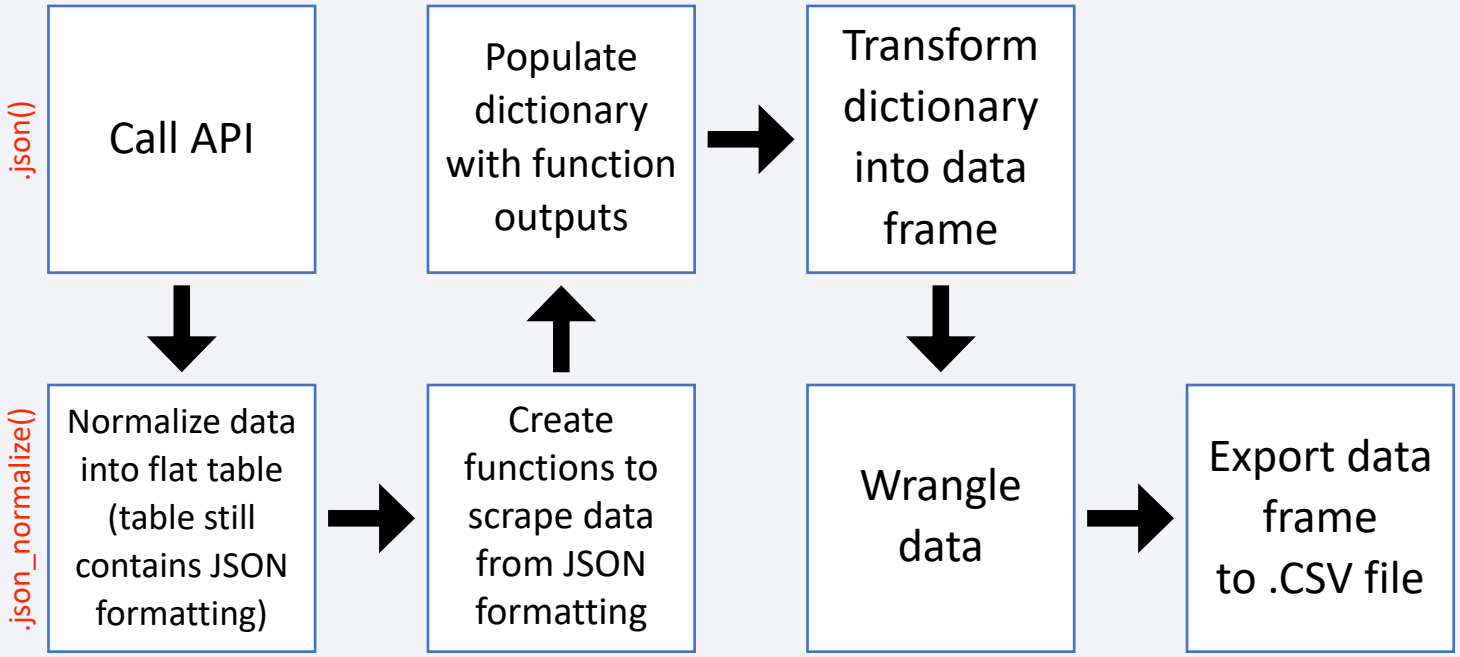
Methodology

Methodology

1. **Collect data**: built web scraper and used API's.
2. **Wrangle data**: replaced missing values and organized the data into structured tables suitable for data frame creation.
3. **Explore data**: used SQL and visualizations in Python.
4. **Visualize data**: used Folium maps and built a Plotly dashboard.
5. **Model data**: used four classification models to predict landing outcome: logistic regression (LR), Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN).

Data Collection – SpaceX API

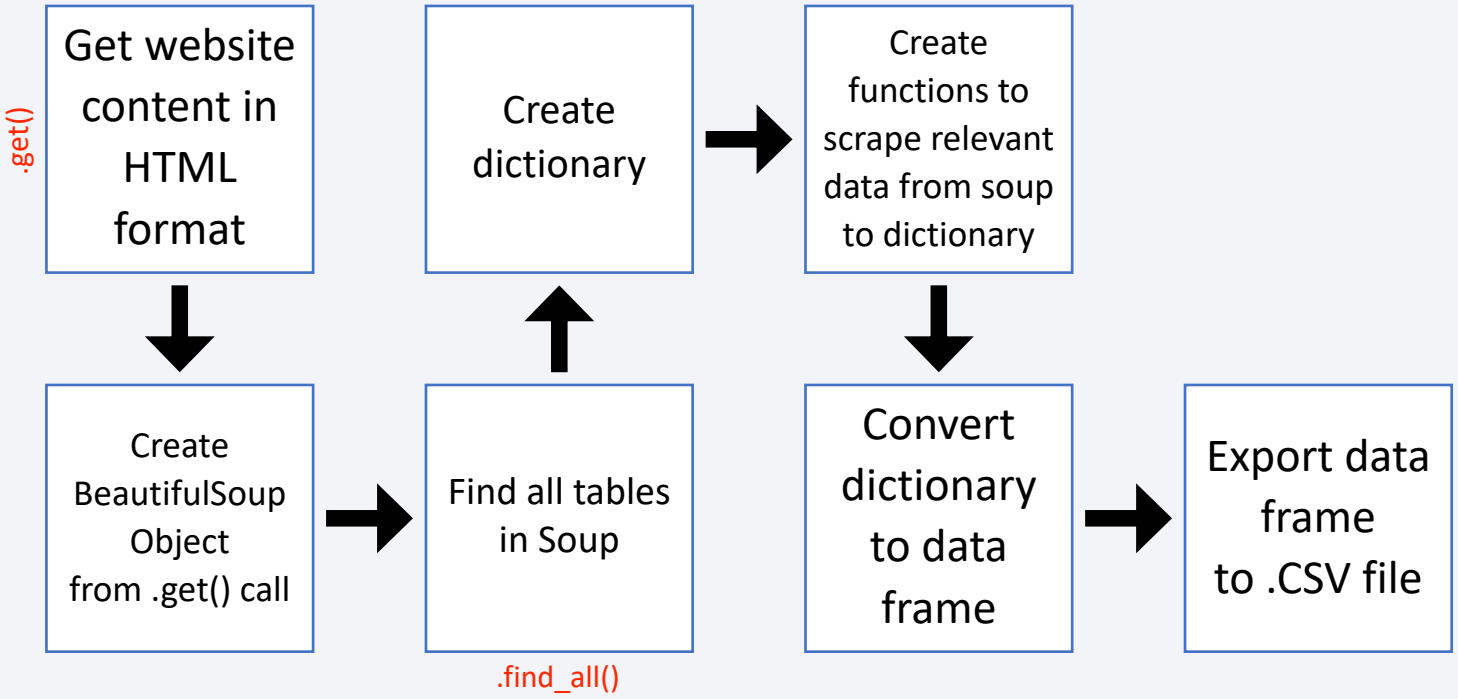
Data found using the SpaceX API https://api.spacexdata.com/v4/launches/ through the *requests* module



https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk1 - Lab1 - Data Collection with API.ipynb

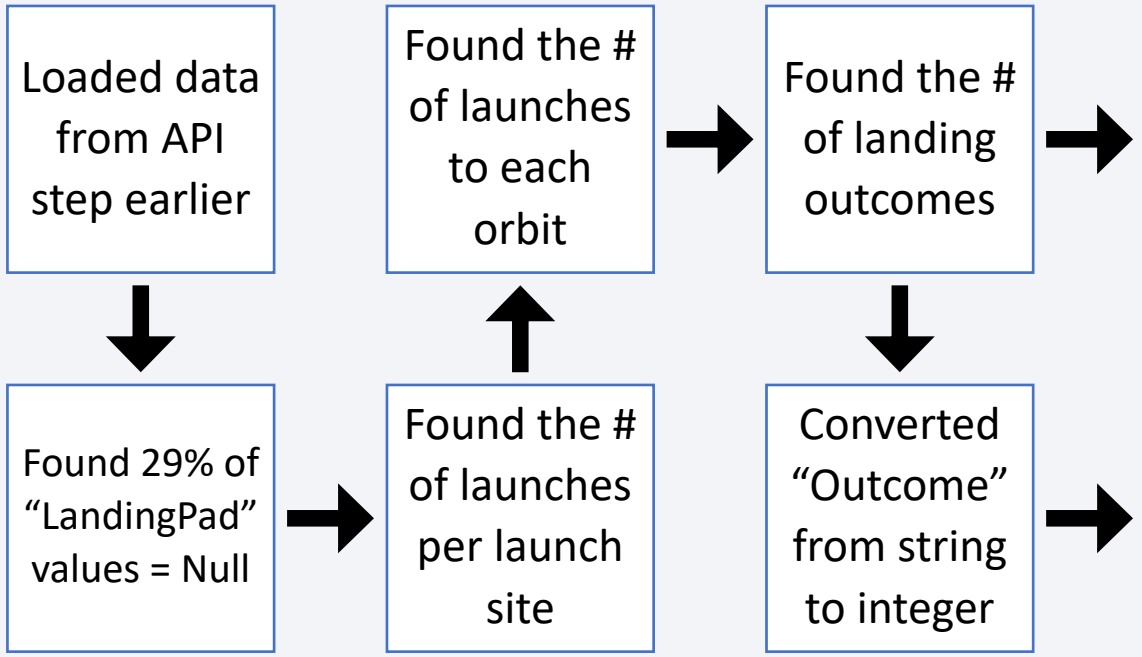
Data Collection – Web Scraping

Data found on the SpaceX launch Wiki https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk1 - Lab2 - Data Collection with Webscraping.ipynb

Data Wrangling



True Ocean: Successful landing in ocean
False Ocean: Unsuccessful landing in ocean
True RTLS: Successful landing to ground pad
False RTLS: Unsuccessful landing to ground pad
True ASDS: Successful landing on drone ship
False ASDS: Unsuccessful landing on drone ship
None ASDS: Failure to land
None None: Failure to land

https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk1 - Lab3 - Data Wrangling.ipynb

EDA with SQL

1. Display names of unique launch sites
2. Display five records where launch site starts with ‘CCA’
3. Display total mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List date of first successful landing outcome on landing pad
6. List booster versions that have successful drone ship landing *and* carried payloads between 4000kg and 6000kg
7. List total number of successful and failed landing outcomes
8. List booster versions that have successfully carried the maximum payload mass
9. List the month, booster version, launch site, and landing outcome for all failed missions in 2015
10. Rank the count of the landing outcomes between 06/04/2010 and 03/20/2017

https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk2 - Lab1 - EDA with SQL.ipynb

EDA with Data Visualization

1. **Scatter plot:** payload mass vs flight number
    1. See progression of launch success and payload mass through time
2. **Scatter plot:** launch site vs flight number
    1. Shows popular launch sites and how successful they are
3. **Scatter plot:** launch site vs payload mass
    1. Shows which sites are best for heavier masses
4. **Bar chart:** success rate vs destination orbit
5. **Scatter plot:** orbit vs flight number
6. **Scatter plot:** orbit vs payload mass
    1. Heavier payloads tend towards greater landing success rate for polar, LEO, ISS missions. No such correlation for GTO orbit.
7. **Line chart:** Success rate vs year
    1. Overall strong growth in success rate with a few small hiccups. Masterful years between 2013 and 2017 with constantly-increasing success rate.

https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk2 - Lab2 - EDA with Visualization.ipynb

Build an Interactive Map with Folium

Map elements created:

1. Marker - demarcates landing attempt success/failure per launch site
2. Circle - highlights launch sites for ease of identification
3. Line - representing distances from launch pads to key infrastructure

Noteworthy Observations:

1. Launch sites are all near coastal waters and close to the equator
    1. Launch sites near equator assist with reaching orbit due to rotation of earth
2. Launch sites are located next to coastal waters for safety reason (debris over population is bad)
3. Critical infrastructure (rail and highways) are found near launch sites for ease of transporting rocket & payload hardware to launch site

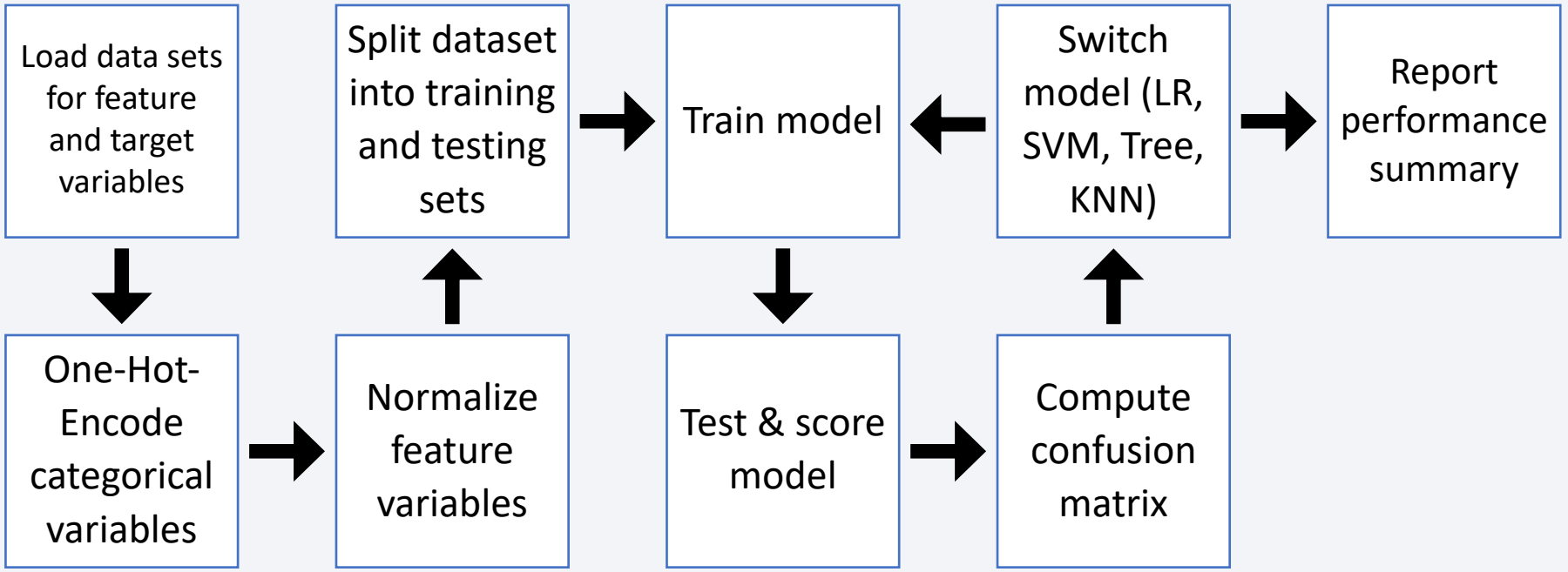
https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk3 - Lab1 - Visualization with Folium Lab.ipynb

Build a Dashboard with Plotly Dash

1. **Pie chart:** showing the success or failure of each landing attempt
    1. Shows landing success rate by launch site and individual landing percentages for specific launch sites
2. **Scatter plot:** showing the payload vs launch site with color indicating the success/failure of the landing.
    1. Red indicating failed landing attempts
    2. Green indicating successful landing attempts
3. Selecting “All Sites” or a specific launch site and the payload mass range of interest and the two charts will update accordingly.
4. It is important to understand if certain launch sites or payload mass ranges have a higher likelihood of a successful landing. If a strong correlation can be found, it could lead to a more accurate prediction model and/or a greater focus on the right features for training the model.

https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk3 - Lab2 - PlotlyDashboard.py

Predictive Analysis (Classification)

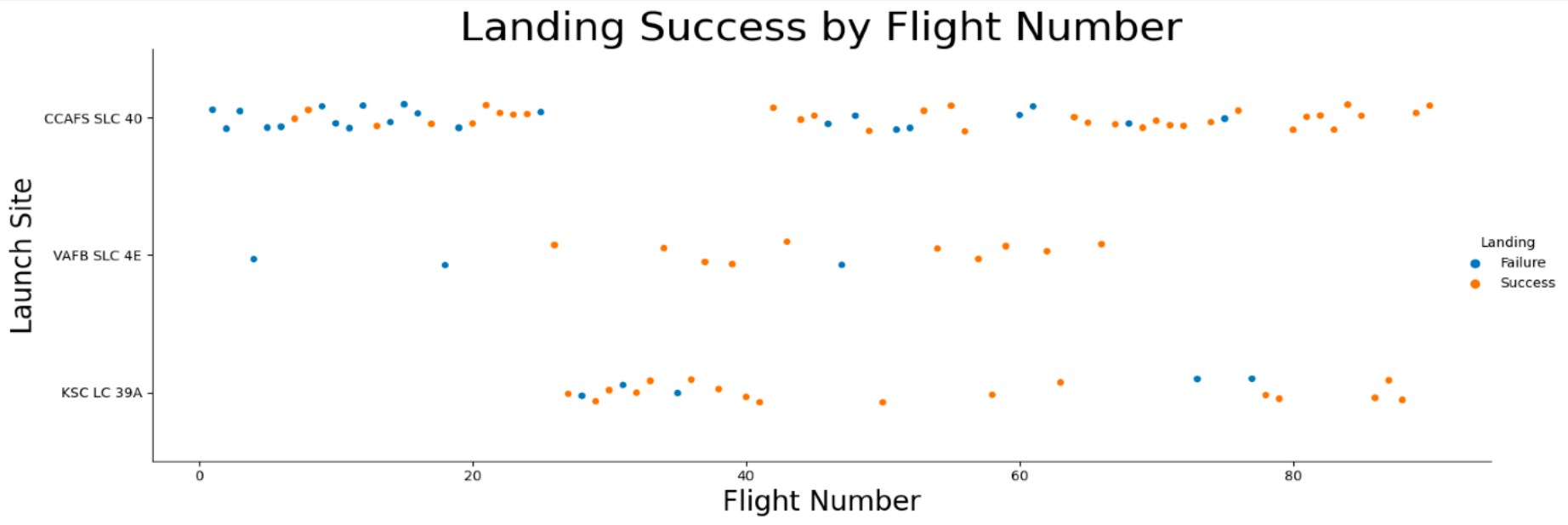


https://github.com/RocketMan-n-Oaks/IBMDataScience-Capstone/blob/main/Course10 - Wk4 - Lab1 - ML Prediction.ipynb

Section 2

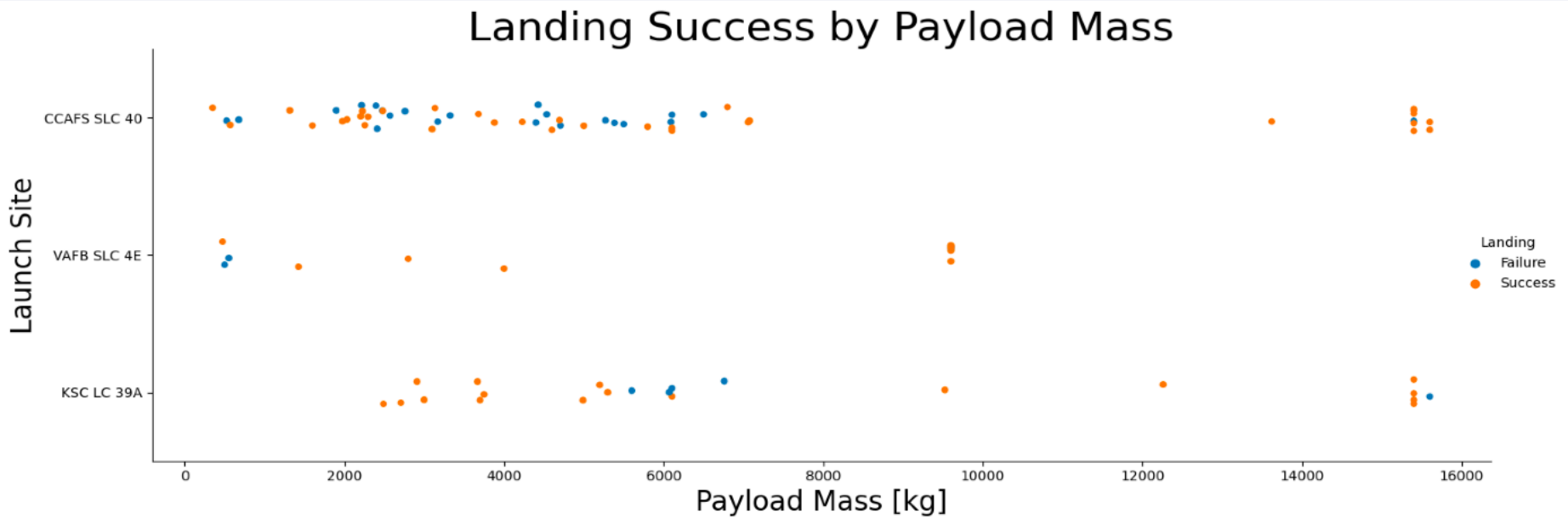
Insights drawn from EDA

Flight Number vs. Launch Site



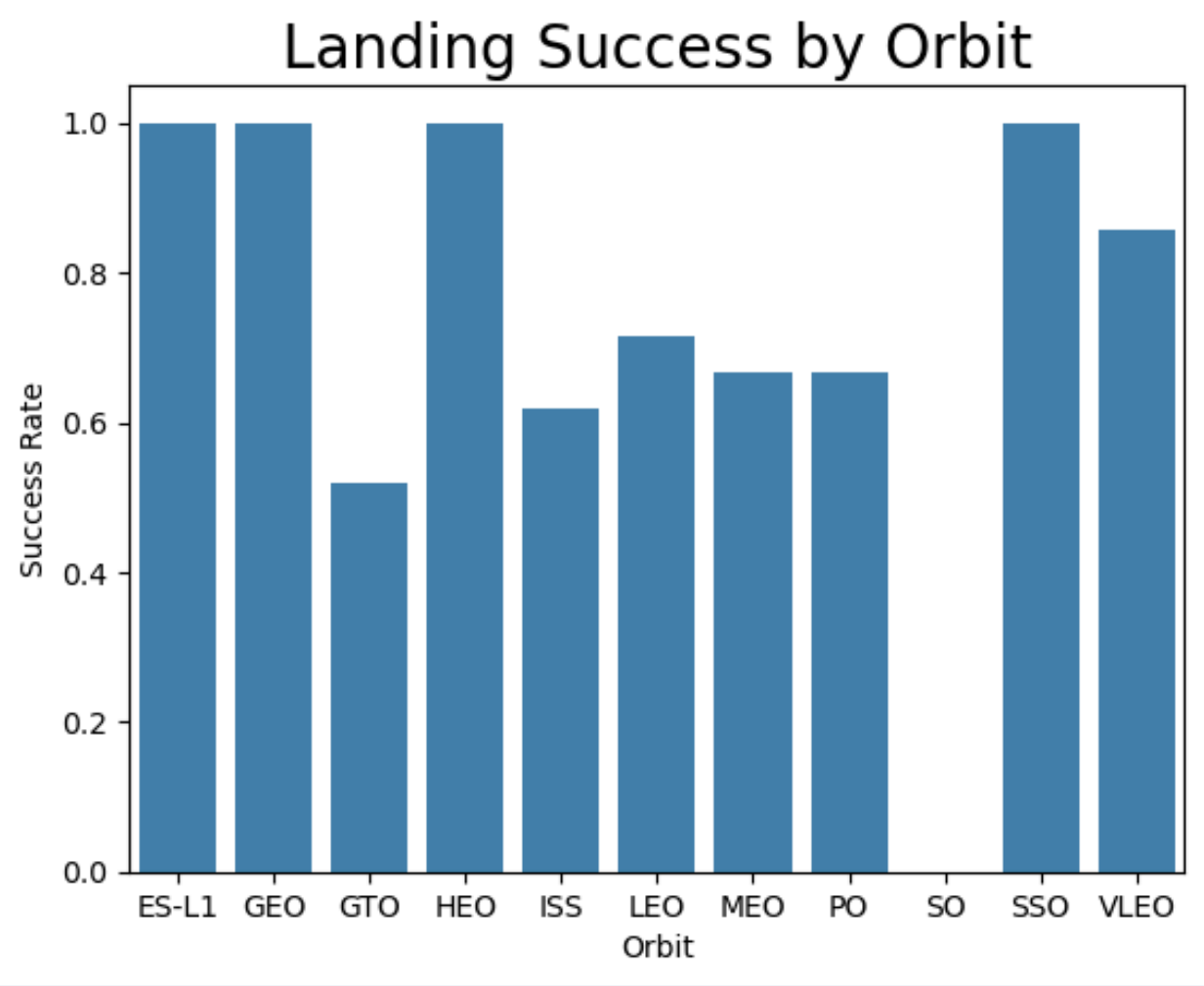
1. Landing attempt successes at CCAFS SLC 40 have increased greatly over time.
2. KSC LC 39A has the best overall success rate.
    1. Likely due to its primary use after landing technology matured.

Payload vs. Launch Site



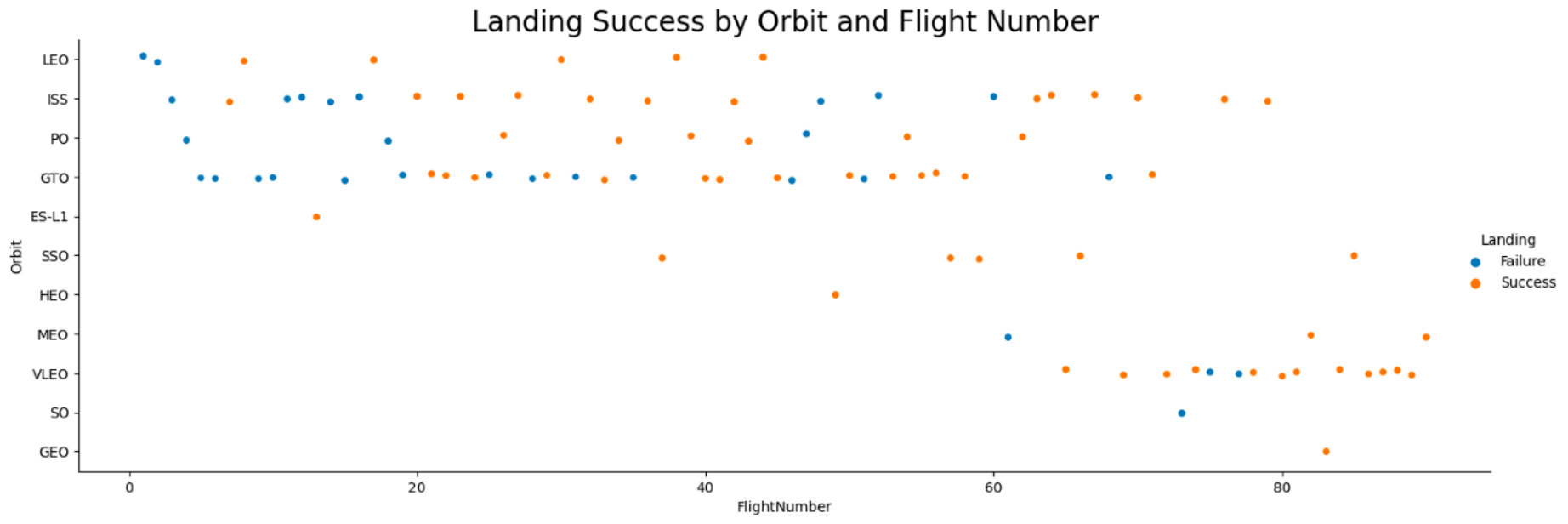
1. Heavier payloads have a greater landing success rate.
2. CCAFS SLC 40 and KSC LC 39A have the largest payload capability.
    1. Likely due to benefit of easterly launch using earth's spin to assist in reaching escape velocity.
3. About half launches with payloads less than 8,000kg have failed landing attempts.

Success Rate vs. Orbit Type



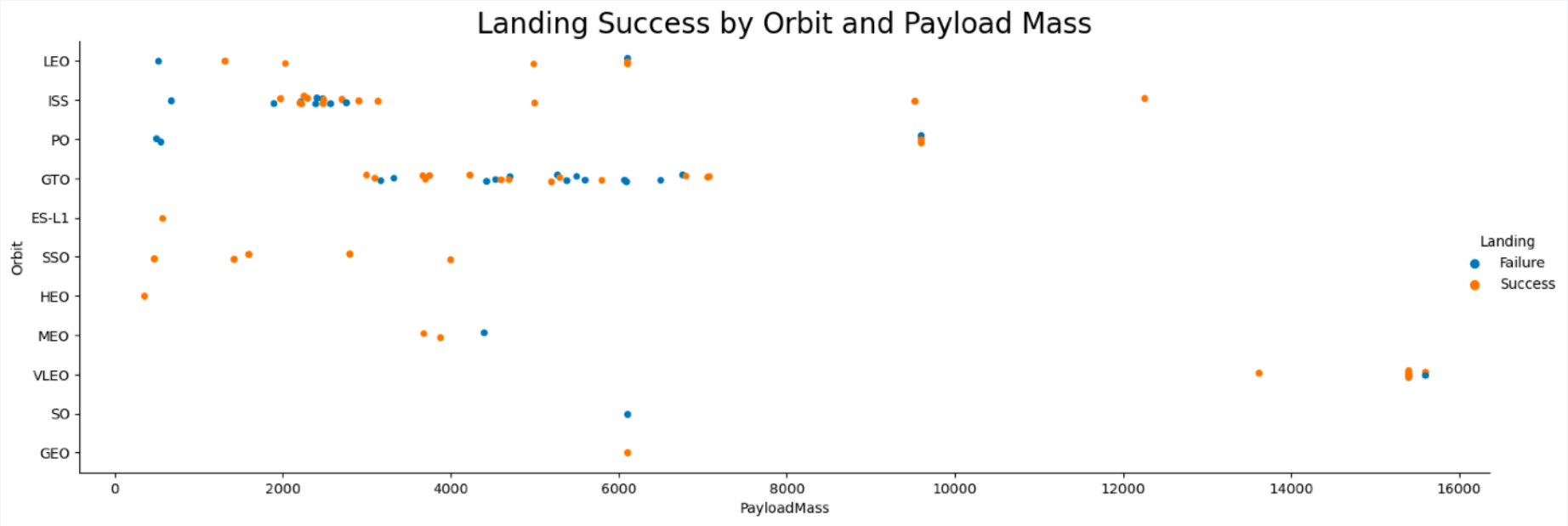
1. L1, GEO, HEO, and SSO have a 100% success rate.
2. More information about dates, failed landing types, total launches per orbit, and payload mass would be required to come to any concrete conclusions about this plot.

Flight Number vs. Orbit Type



1. It is clear that as flight number (time) increases, the landing success rate steadily increases.
2. Some failures shown above are “no [landing] attempt” and thus should not be considered in this study. This fact will be brought up a few more times in this package and again in my conclusions.

Payload vs. Orbit Type



1. Overall trend showing greater success rate as payload increases
    1. GTO missions are exception
2. Upon deeper inspection of the data, a great number of failed landings are actually “no attempt” and should not be considered a failure for the purposes of this plot. I believe this causes inaccurate conclusions to be drawn.

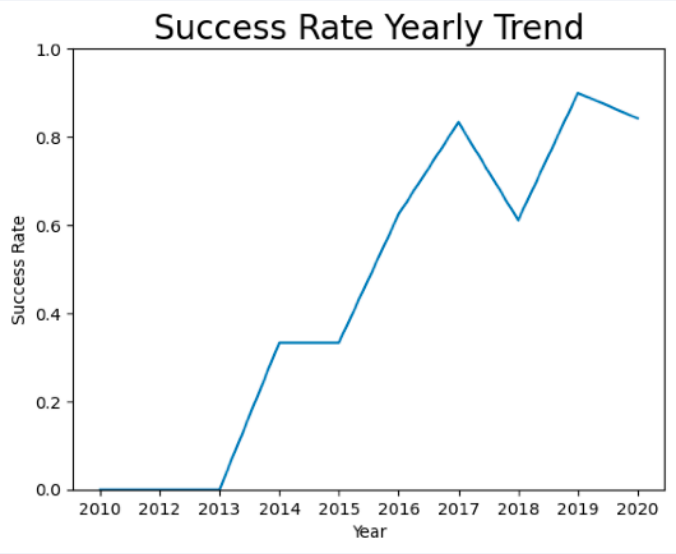
Launch Success Yearly Trend

```
%%sql SELECT substr(Date,0,5) AS Year, Date, Landing_Outcome
FROM SPACEXTABLE
WHERE Year = '2018'
ORDER BY Date ASC
```

✓ 0.0s

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

Year	Date	Landing_Outcome
2018	2018-01-08	Success (ground pad)
2018	2018-01-31	Controlled (ocean)
2018	2018-02-22	No attempt
2018	2018-03-06	No attempt
2018	2018-03-30	No attempt
2018	2018-04-02	No attempt
2018	2018-04-18	Success (drone ship)
2018	2018-05-11	Success (drone ship)
2018	2018-05-22	No attempt
2018	2018-06-04	No attempt
2018	2018-06-29	No attempt
2018	2018-07-22	Success
2018	2018-07-25	Success
2018	2018-08-07	Success
2018	2018-09-10	Success
2018	2018-10-08	Success
2018	2018-11-15	Success
2018	2018-12-03	Success
2018	2018-12-05	Failure
2018	2018-12-23	No attempt



1. Between the years of 2013 and 2017, SpaceX has seen a constantly increasing landing success rate.
2. 2018 decrease artificially low due to disproportionately more “No [landing] Attempts” that were counted as a failed landing in the line plot above.
    1. No other year had as many “No Attempts”

All Launch Site Names

Query:

```
%sql SELECT DISTINCT “Launch_Site”
FROM SPACEXTABLE
```

Result:

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

Launch Site Names Begin with 'CCA'

Query:

```
%sql SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE '%CCA%'
LIMIT 5
```

Result:

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

Total Payload Mass

Query:

```
%sql SELECT SUM(Payload_Mass__KG_) AS [Total Payload Mass]
FROM SPACEXTABLE
WHERE Customer LIKE '%NASA (CRS)%'
```

Result:

Total Payload Mass
48213

Average Payload Mass by F9 v1.1

Query:

```
%sql SELECT AVG(Payload_Mass__KG_) AS [Total Payload Mass]
FROM SPACEXTABLE
WHERE Booster_Version LIKE '%F9 v1.1%'
```

Result:

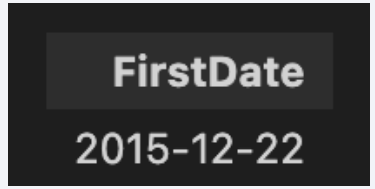
Total Payload Mass
2534.6666666666665

First Successful Ground Landing Date

Query:

```
%sql SELECT min(Date) AS FirstDate
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE '%Grounds%'
```

Result:



Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Payload_Mass__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome LIKE '%Success (d%'
```

Result:

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

Total Number of Successful and Failure Mission Outcomes

Query:

```
%sql SELECT Mission_Outcome, count(*) AS Qty
FROM SPACEXTABLE
GROUP BY Mission_Outcome
ORDER BY Qty DESC
```

Result:

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

Boosters Carried Maximum Payload

Query:

```
%%sql SELECT Booster_Version, Payload_Mass__KG_ AS [Payload Carried]
FROM SPACEXTABLE
WHERE Payload_Mass__KG_ = (SELECT MAX(Payload_Mass__KG_)
                           FROM SPACEXTABLE)
```

Result:

Booster_Version	Payload Carried
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Failed Landing Records

Query:

```
%%sql SELECT substr(Date,6,2) AS 'Month in 2015', Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE substr(Date,0,5) = '2015' AND Landing_Outcome LIKE '%Failure (d%'
```

Result:

Month In 2015	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Query:

```
%%sql SELECT Landing_Outcome, count(*) AS Landing Qty
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY LandingQty DESC
```

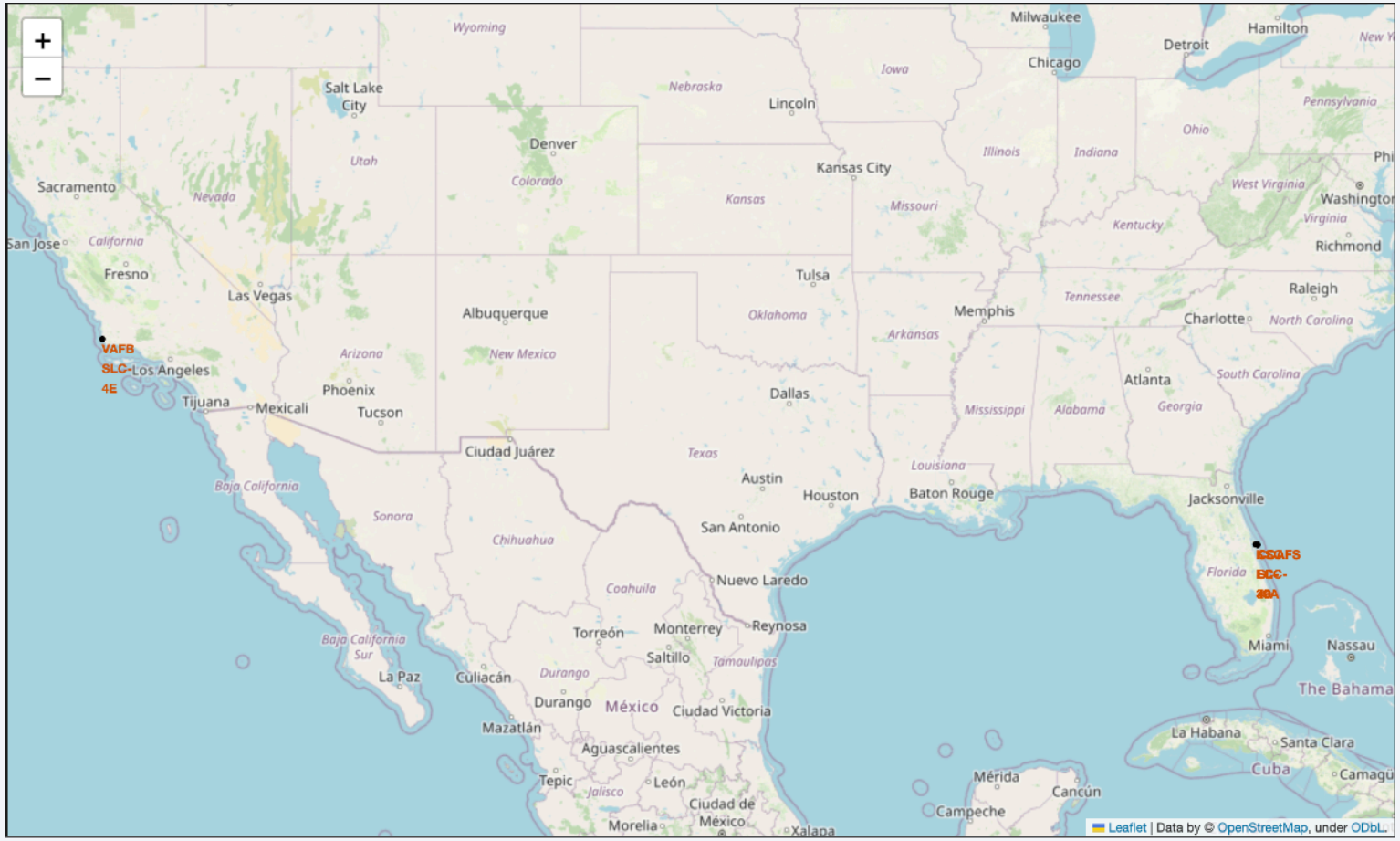
Result:

Landing_Outcome	LandingQty
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

Section 3

Launch Sites Proximities Analysis

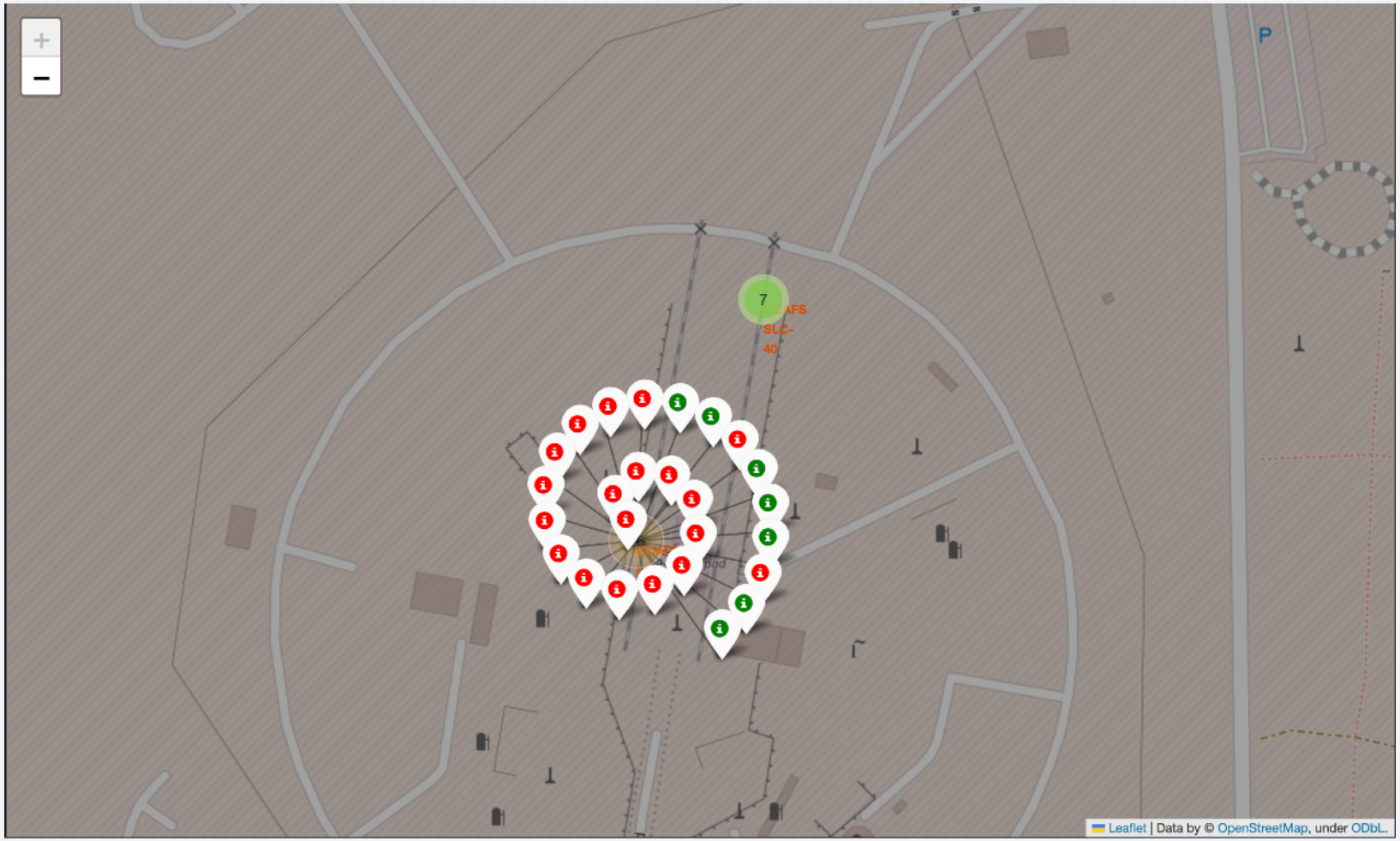
US-Based Launch Facilities



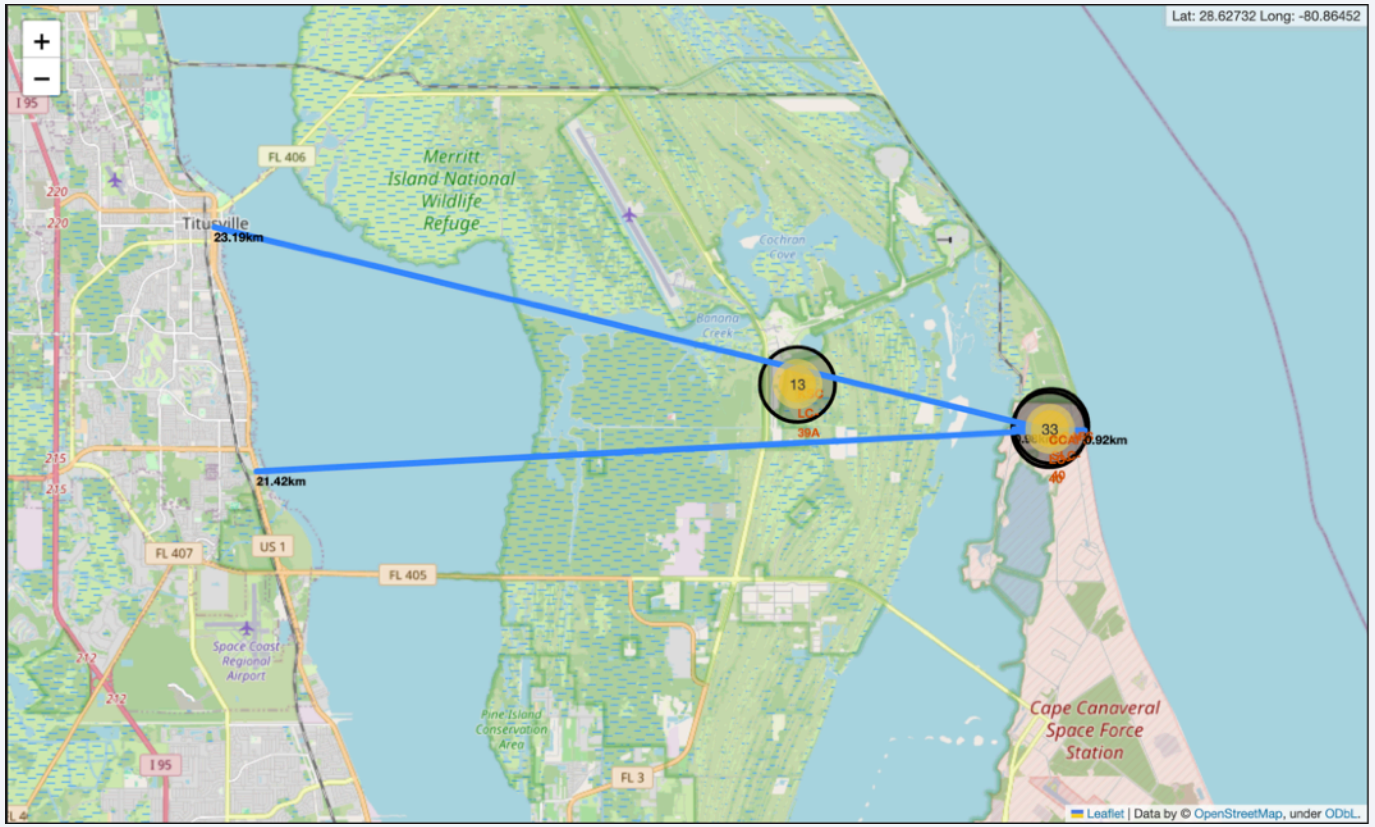
1. Launch facilities are located on coastal waters. They are also located as near the equator as possible, considering other factors such as infrastructure, water access and proximity to cities.
2. The equator is advantageous as rockets can benefit from the increased rotational velocity of earth to assist in reaching orbit.

Landing Successes from KSC Pad 40

1. Here we see the number of successful (green) and unsuccessful (red) landings when launched from CCAFS LC-40.
2. From this data, CCAFS LC-40 has a 7/26 (26.9%) success rate.



Launch Pad Critical Proximities



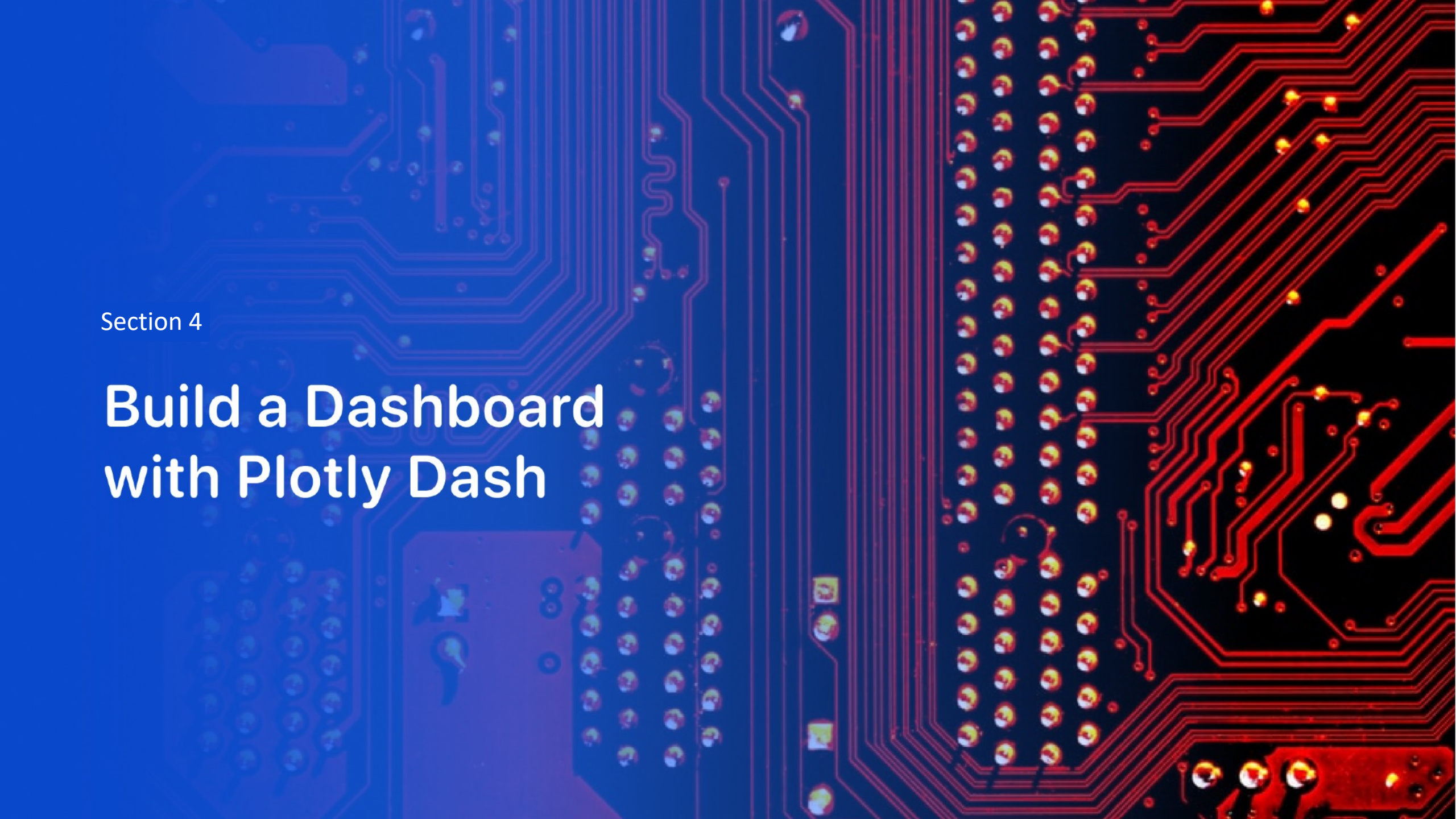
Critical Proximities:

Coast: 0.9km | City: 23.2km | Railway: 1.0km | Highway: 21.4km

1. Kennedy Space Center is conveniently located near key infrastructure such as highways and railways to support launch hardware and payload deliveries.
2. The launch pads are also located to the west of the coast which is advantageous since rocket launches will primarily be toward the east to take full advantage of the planets eastward rotation.
3. Finally, the pads are located close enough to cities to support staff housing, family utilities, and extracurricular activities.

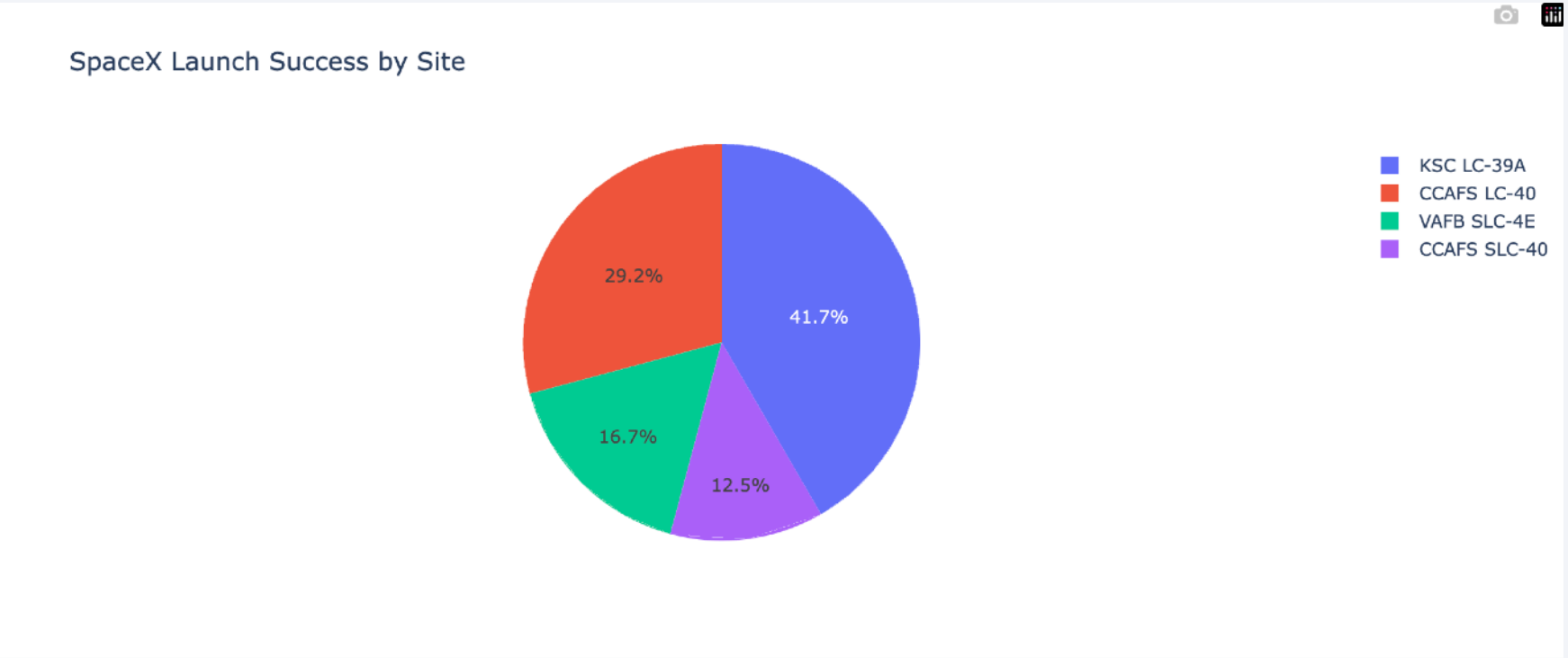
Section 4

Build a Dashboard with Plotly Dash



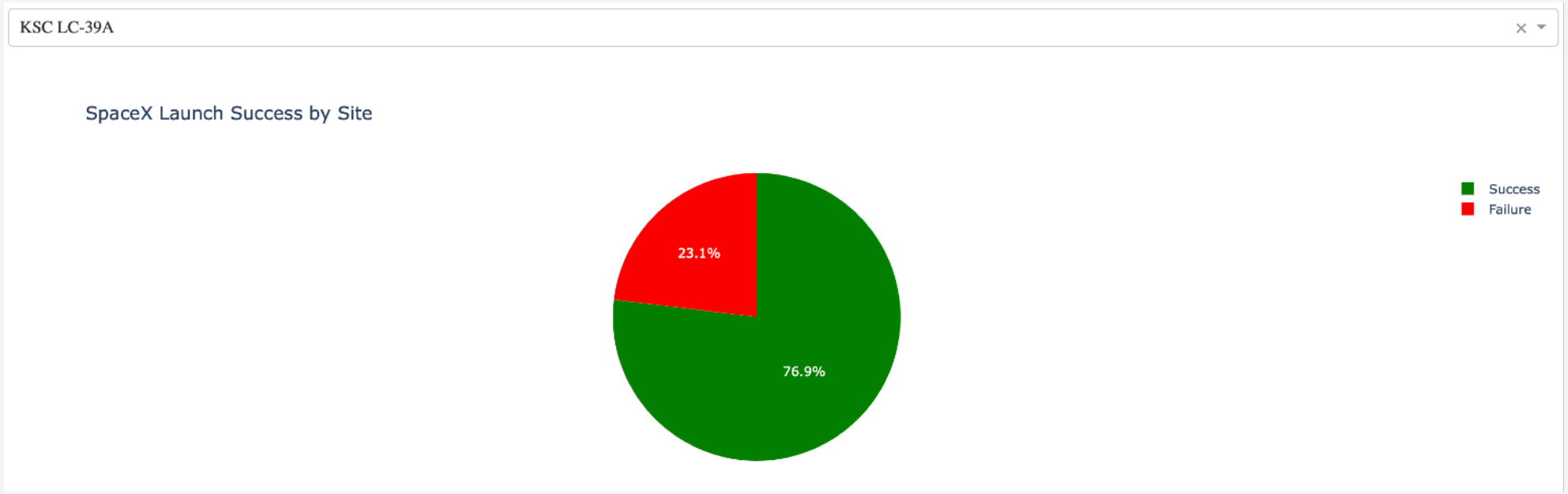
Landing Success Across All Launch Sites

1. KSC LC-39A has the greatest number of successful landings. This is likely due to SpaceX using this launch site for most of its more recent (and not its earliest) launches.
2. According to the data used in previous sections of this project, there is a 7 year history of launches prior to SpaceX using KSC LC-39A as its primary launch site.



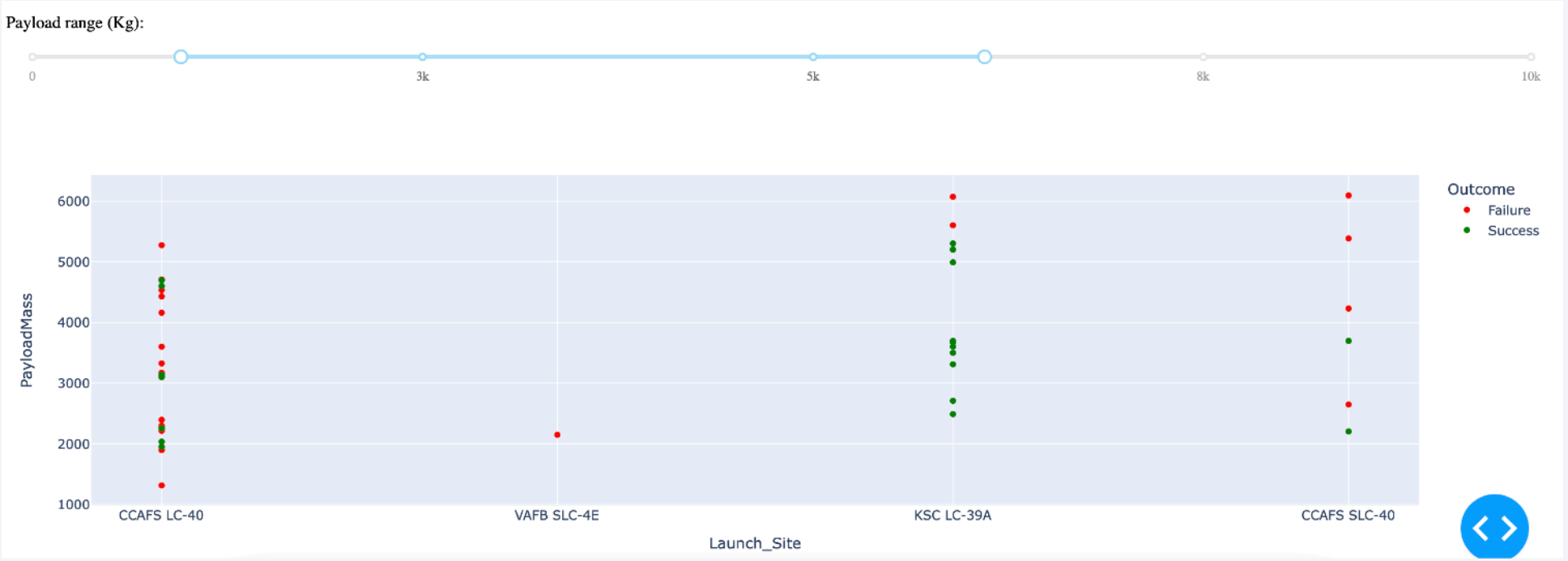
Launch Site with Greatest Landing Success Ratio

1. KSC LC-39A has the greatest landing success ratio of the four launch sites considered in this study. It would be of interest to plot the landing attempts vs time and launch site.
2. CCAFS LC-40 has the worst landing success ratio and the greatest number of landing attempts. Once again, if looking at the data present earlier in this study, it can be seen that CCAFS LC-40 was the primary launch site for most of the early SpaceX launches. It would be expected they would suffer the greatest number of failed landing attempts.



Successful Landings by Payload Mass and Launch Site

1. KSC LC-39A has the greatest overall success rate but does suggest a trend of greater landing failure rate as payload mass increases.
    1. Trend likely due to large payload mass requiring all fuel to be expended during orbit raising.
2. No other such trend is immediately visible.



Section 5

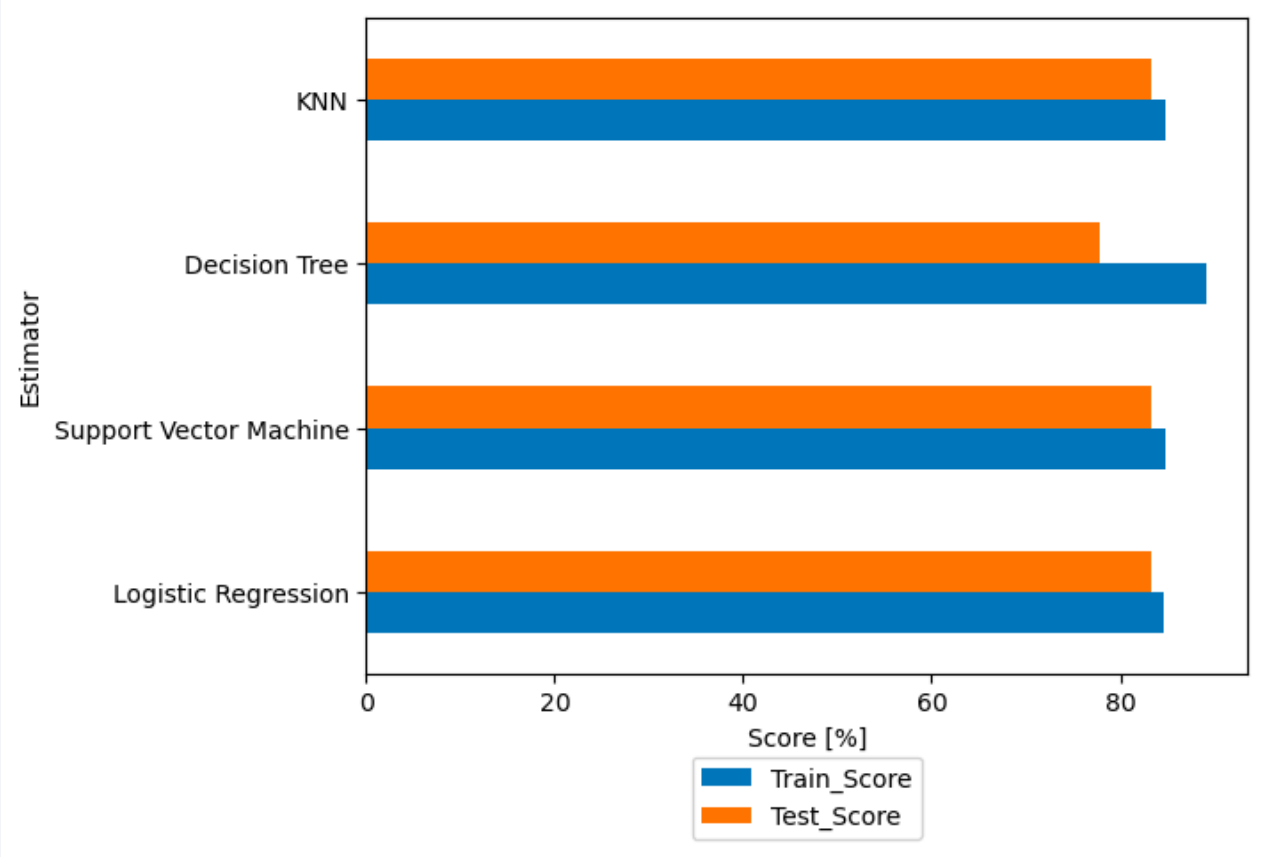
Predictive Analysis (Classification)

Classification Accuracy

1. All four models performed very similarly with the decision tree slightly under-performing the rest on the test data and slightly over-performing the rest on the train data. This could indicate a slight overfitting.

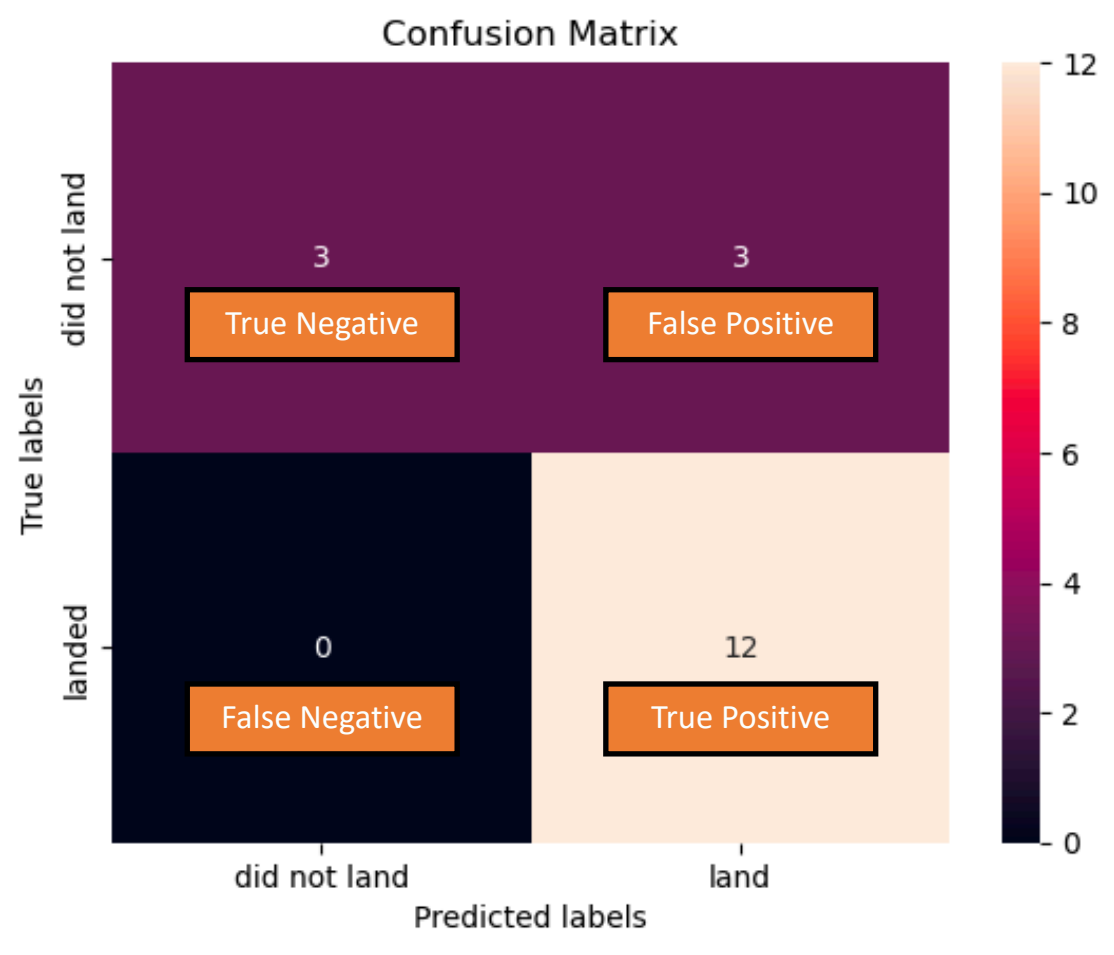
3. The similarity across all models is likely due to the small data set used for this capstone.

5. It would be interesting to see how these vary, on average, with a different train_test_split ratio, random_state value, and a more up-to-date set of rocket launch data.



Confusion Matrix

1. Confusion matrices were the same across all models.
2. False positive results in under-pricing cost of rocket launch.
3. Key metrics:
    1. **Precision**: How often is predicted “yes” correct?
    2. **Recall**: When actually “yes”, how often does it predict “yes”?
    3. **Specificity**: When actually “no”, how often does it predict “no”?
    4. **Misclassification Rate**: Overall, how often is it wrong?
    5. **Accuracy**: Overall, how often is it right?
    6. **F1 Score**: Weighted average of recall and precision.



Confusion Matrix Calculations

	Equation	Our Value
Precision	TP/(TP + FP)	12/(12 + 3) = 0.80
Recall	TP/(TP + FN)	12/(12 + 0) = 1.00
Specificity	TN/(TN + FP)	3/(3 + 3) = 0.50
Misclassification Rate	(FP + FN)/(TP + TN + FP + FN) = 1 - Accuracy	1 - 0.83 = 0.17
Accuracy	(TP + TN)/(TP + TN + FP + FN) = 1 - Misclassification Rate	(12 + 3)/(12 + 3 + 3 + 0) = 0.83
F1 Score	2 * (Precision * Recall)/ (Precision + Recall)	2 * (0.80 * 1.00)/(0.80 + 1.00) = 0.89

Conclusions

1. SpaceX has continuously and greatly improved their landing success rate throughout the years with an overall trend showing an increase in landing success rate with payload mass.
2. Launch sites are located more towards the equator to take full advantage of the earth's angular velocity.
3. All models performed similarly with the decision tree indicating a slight overfit. The remaining three models show an 83.3% success rate on the test data.
4. All four models show a false positive (1-Specificity) rate of 50%. This is will lead to under-pricing the cost of each launch. The false positive rate should be minimized in any future investigations.

Future Considerations

1. The launch rate continues to increase after 2020. Rerunning the analysis with the complete data set would yield a more representative conclusion.
2. Further work to reduce the consistent false positives should be done.
3. Feature variables such as weather, time of year, and core reuse count should be added to further refine the models.
4. Ensemble modeling might yield a more accurate landing prediction.
5. XGBoost is a very popular model and should be considered for further investigation.
6. Reclassify landing outcomes to exclude “no landing attempt” from failures. A no landing attempt should be accounted for in the price of launch and not as a failure for the purposes of this analysis.

Thank you!