Winning Space Race with Data Science

Christian Guerra
January 19 2022

Outline

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

Executive Summary

This project has the purpose of predicting landing outcomes from Space X launches. Data is collected via an API and transformed for proper use.

An EDA is done to understand the data, select the relevant variables and have better insight on what information might be useful for the model. The table is transformed for a final model.

For Classification Models are used to predict the outcome: Logistic Regression, SVM, Decision Tree and K-Nearest Neighbors. All models have an accuracy of 83.33%, concluding that all are successful prediction methods.

The confusion matrix shows that the model inaccuracies are related to predicting false positives.

Introduction

- The commercial age is here, each day space travel is becoming more affordable. The most accomplished company to this date has been SpaceX.
- SpaceX has achieved lower costs due to the reuse of the first stage with the Falcon 9 rocket. If we could determine if the first stage will land we are able to determine the cost of the launch.
- This project consists in determining the cost of a launch by using machine learning models to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 build, tune, evaluate classification models

Data Collection

- The first part consists in collecting data through the SpaceX API. Data is obtained through a JSON which is converted to a pandas Data Frame and cleaned afterwards so this information can be used for this project.
- For this project data will also be collected through Web Scraping the SpaceX Wikipedia page to obtain the same Data Frame than the one obtained through the API.
- This information will include information about which rocket was used, payload delivered, launch specification, landing specifications and landing outcome.

Data Collection – SpaceX API

The Following link could be used to see the complete notebook: Lab 1: Data Collection with API.

Import Libraries and Define Auxiliary Functions	We import requests to make HTTP requests to obtain data. We define a series of functions to extract information from the API using ID numbers.
Obtaining Data Frame	Using a GET request and decoding the JSON response we convert the data to a pandas Data Frame. We format columns correctly.
Make additional Requests and Parse Data	We use the getBoosterVersion() function to obtain booster versión from the rocket ID. getLaunchSite(), getPayloadData() and getCoreData() are also used to obtain the launch site, payload and core data. The Final Data Set is created.

Data Collection - Scraping

The Following link could be used to see the complete notebook: Lab 2: Data Collection with Web Scraping.

Import Libraries and Define Functions	We use BeautifulSoup to extract Falcon 9 launches from wikipedia via Web Scraping. Helper functions are defined to facilitate the parsing process.
Request Data From URL	Using the html GET method we request the Falcon9 launch HTML page as an HTTP Response and create a BeautifulSoup object.
Extract Column Names	We find all tables on the wiki page using BeautifulSoup, we select the third table, we iterate through the <th> elements that are the column names and append to a list.
Create Data Frame via Parsing	We create a dictionary with extracted column names. We create a pandas Data Frame. We loop through the tables and obtain each row from the html tables for all of the columns previously extracted.

Data Wrangling

Here the data wrangling process is described, the complete notebook can be seen in the following link: Lab 3: Data Wrangling.

Initial EDA	This step was partially done in Lab 1 but is part of the Data Wrangling process. The Data Frame is Filtered to include only Falcon 9. We calculate the percentage of missing values per attribute. 40% of Landing Pad is missing. Payload Mass's 5 missing instances were replaced to their mean in lab 1. We analize attribute data types.
Number of Launches and Occurence	We obtain the number of launches per launch site using the value_counts() method from the Launch Site attribute. There are three different locations: 'CCAFS SLC 40' (55), 'KSC LC 39 A' (22) and 'VAFB SLC 4E' (13). We obtain the Number of occurences per orbit, each launch aims to a dedicated orbit.
Determine Landing Outcomes	We obtain the different landing outcomes, and create a set of the five that did not land succesfully.
Creation of Classification Variable	Since we wish to predict landing outcomes we create a new 'Class' variable that has a value of 0 if the outcome was unsuccesful and 1 if succesful. This is done by using the information from the set of unsuccesful outcomes. We also now determine the success rate (66%).

EDA with Data Visualization

We explore, visualize and prepare data, the complete notebook can be seen in the following link: Lab 4: EDA with Visualization. Here are the list of charts displayed:

- Flight Number: These charts provide information on different attributes and flight number. They also display the outcome:
  - Payload Mass: we wish to understand how the mass might affect the success outcome.
  - Launch Site: we wish to understand how different launch sights affect the success outcome.
- Launch Site / Payload Mass: We observe if there is a relationship between launch sites and their payload mass.

EDA with Data Visualization

- Orbit Type: we wish to analyze the effect of orbit type and the relation with other attributes:
  - Success Rate: how different orbit types have an effect on the success rate.
  - Flight Number: we see how orbit type and flight number are related.
  - Payload: we visualize how Payload mass is related to orbit type.
- Launch Success Yearly Trend: we see how the success rate has varied throughout time.

The variables that might affect the success rate are selected and dummy variables are created for categorical columns. Now all relevant variables are numerical.

EDA with SQL

The Following Queries Where Performed:

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

The full notebook could be seen at: Lab 5: EDA With SQL.

Build an Interactive Map with Folium

Different objects are added to a map of the launch locations, the objects are the following:

- Circles: To show the different launch locations.
- Markers: To name launch locations, show successful and unsuccessful launches in each location with a cluster, write the distance from an object and a launch location.
- Lines: To show a straight line from an object and its launch location (Ex. Highways, Railways, etc…).
- Cursor: To display the latitude and longitude of the mouse in the map.

The complete notebook could be seen at: Lab 6: Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- Plotly Dash is used to create an interactive dashboard for data visualization. The use has given the option to select the location sites and the payload range that will be displayed.
- A Pie Chart was added to understand the percentage of successful launches and a Scatter Plot was added to see the correlation between payload mass and success.
- These options and charts will give the user a better understanding of behaviors of different locations and the influence that the booster version and payload mass affect the success of a launch.
- The interactive dashboard can be seen here: Lab 7: Plotly Dashboard

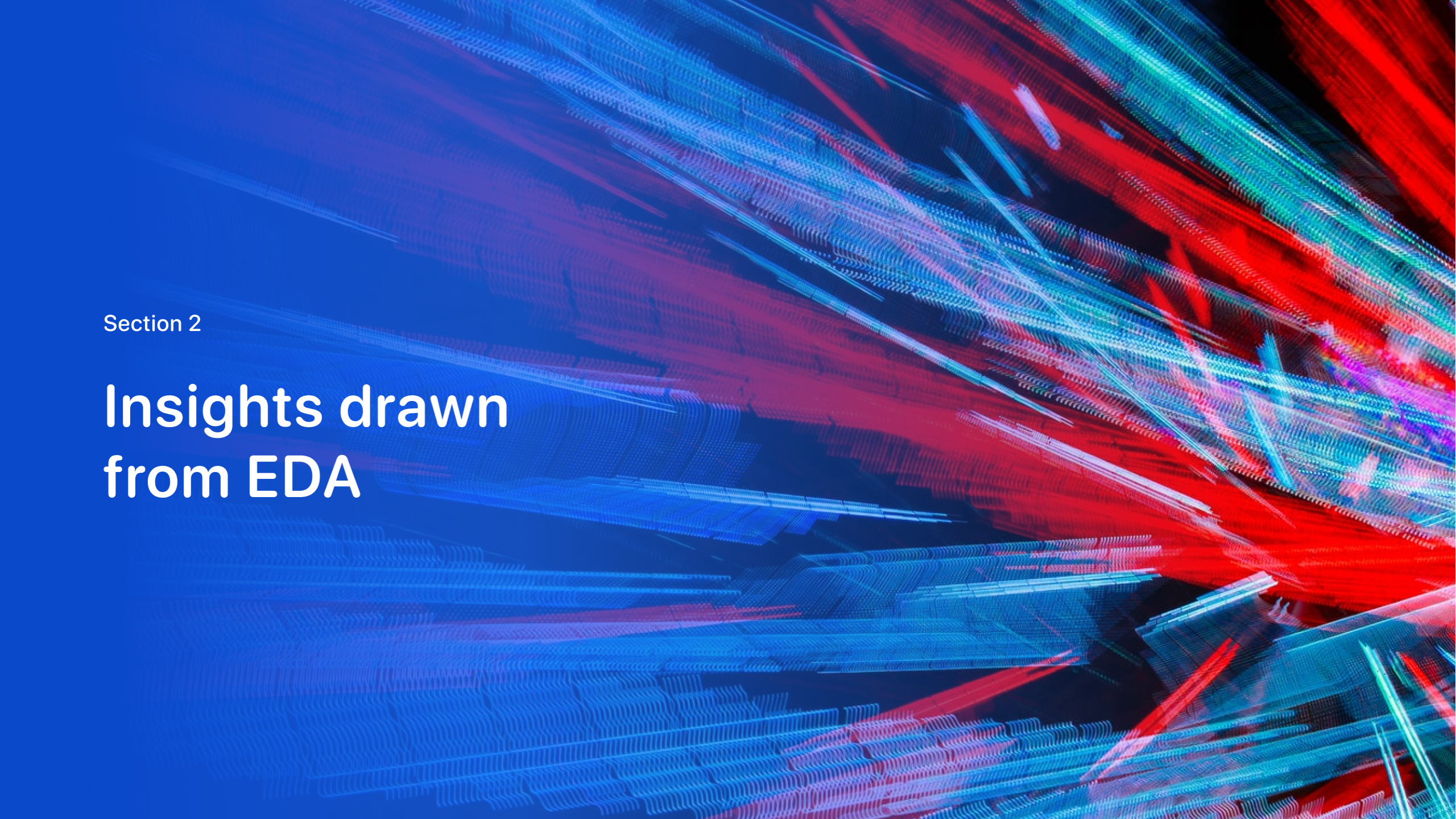
Predictive Analysis (Classification)

Here the data Predictive Analysis process is described, the complete notebook can be seen in the following link: Lab 8: Predictive Analysis

Data Setup and Split	The data is separated in the dependent variable (y) and the independent variables (X). Afterwards X is normalized. Finally the set is split randomly for both X and y in a train and test set. The test set has a size of 20%.
Executing Models	4 models are tested to predict a success: Logistic Regression, Support Vector Machine (SVM), Decision Tree & K Nearest Neighbours. On all models a grid search is used to tune the hyperparameters and obtain the best accuracy for each model.
Comparing Models	All results are studied by reviewing accuracy and a confusion matrix. The best model is selected.

Results

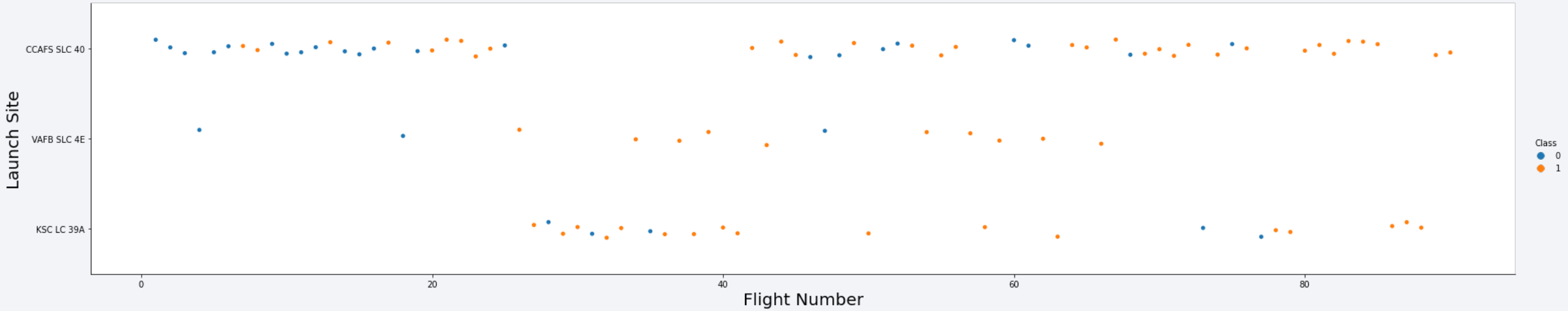
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

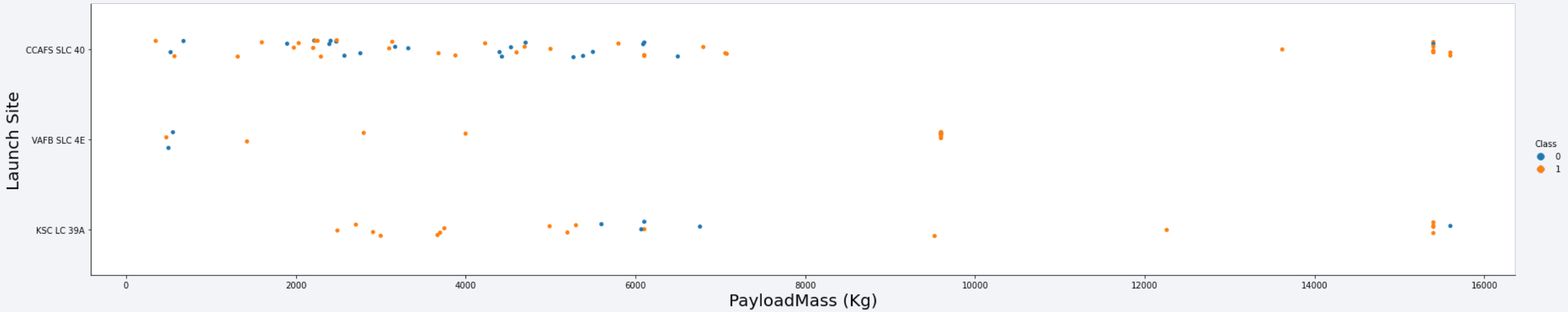
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 seems to be the main launch site used
- KSC LC 39 A seems to be used as a replacement for CCAFS SLC 40 launches
- VAFB SLC 4E is rarely used
- The success rate increases at higher flight numbers, after launch sixty the success rate grows nearer to 100%

Payload vs. Launch Site

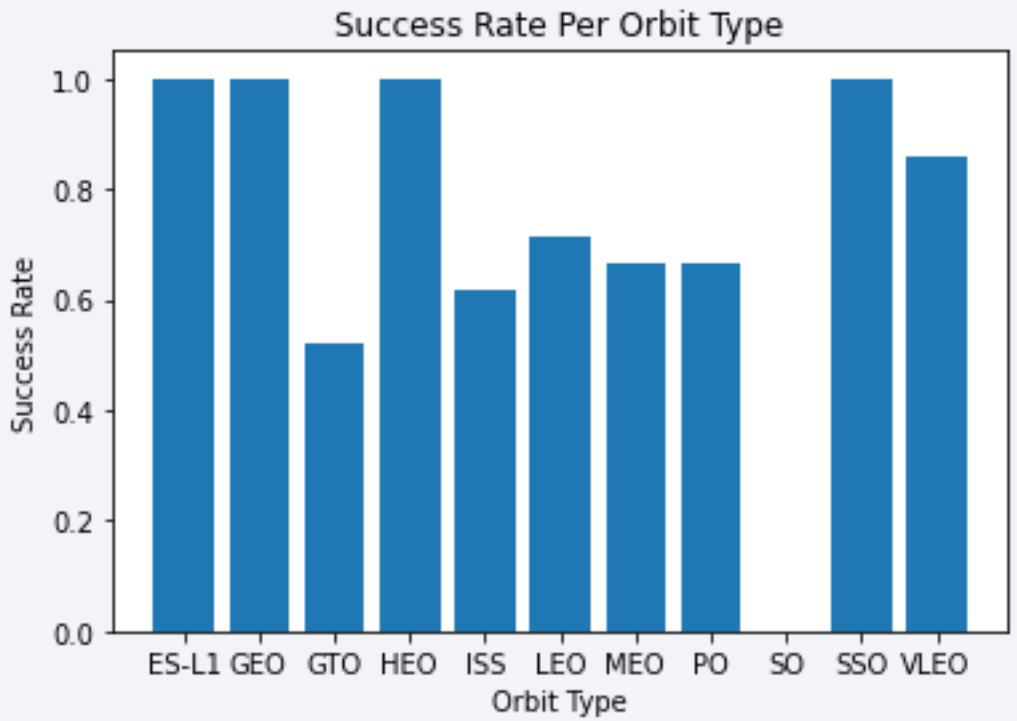


If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).

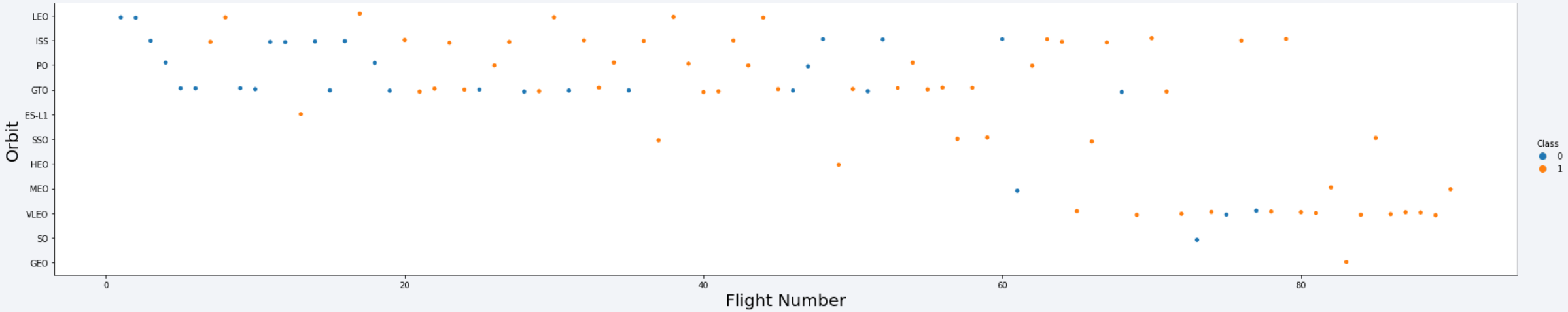
Success Rate vs. Orbit Type

The Success rate clearly varies per orbit type.

The most successful types are ES-L1, GEO, HEO and SSO with a success rate of 100%.



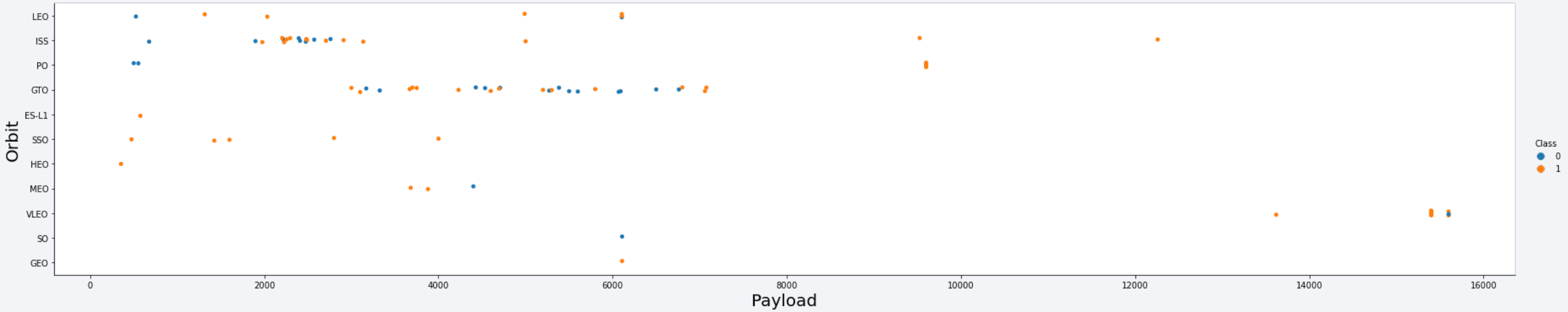
Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

The 4 orbits with 100% success have low number of flights that explain their success, yet VLEO that had a success rate above 80% has a higher amount of flights. It was also implemented after flight 60 and seems to be the prefered method from then on.

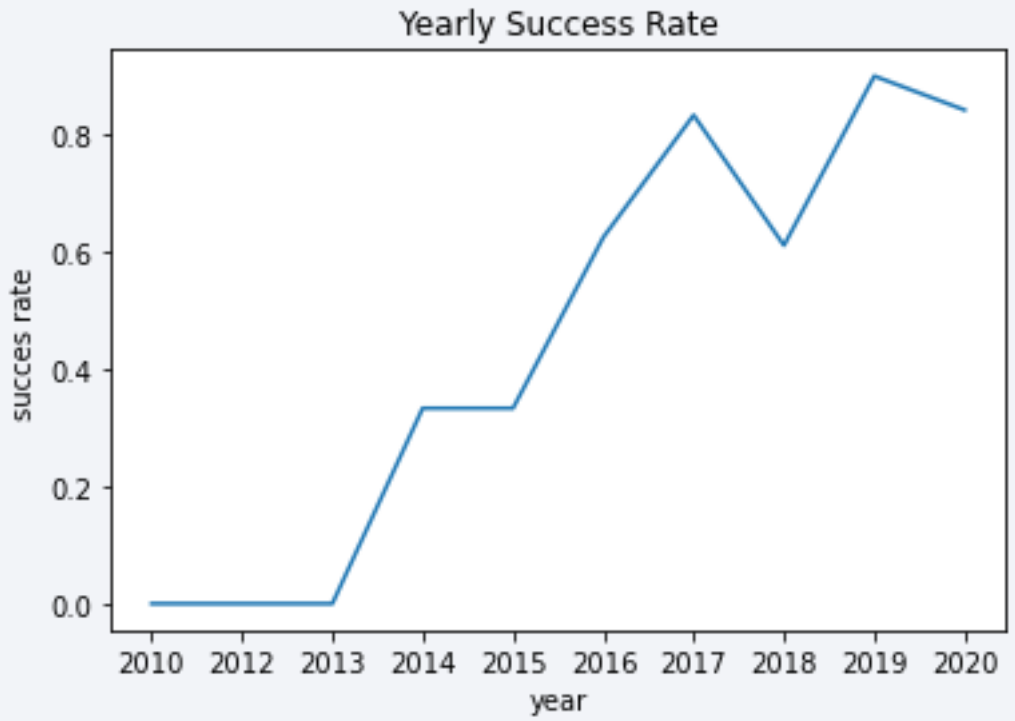
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

You can observe that the sucess rate since 2013 kept increasing till 2020.



All Launch Site Names

- The Following are all the Launch Sites and its query, we calculate it by grouping by launch site.

```
%sql SELECT launch_site FROM SPACEXTBL GROUP BY launch_site;
```

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

Launch Site Names Begin with 'CCA'

- The Following are the first five Launch sites that begin with 'CCA'

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

Total Payload Mass

The total payload mass by boosters launched by NASA is 111.268 kg.

total_payload_mass
111268

```
%%sql
SELECT sum(payload_mass__kg_) AS total_payload_mass
FROM SPACEXTBL
WHERE payload LIKE '%CRS%';
```

Average Payload Mass by F9 v1.1

We calculate the average payload mass carried by booster version F9 v1.1

average_payload_mass
2534

```
%%sql
SELECT AVG(payload_mass__kg_) AS average_payload_mass
FROM SPACEXTBL
WHERE booster_version LIKE '%F9 v1.1%';
```

First Successful Ground Landing Date

This query is the date of the first successful landing outcome on the ground pad.

The first success for this type of landing was December 22, 2015.

first_success
2015-12-22

```
%%sql
SELECT MIN(DATE) AS first_success
FROM SPACEXTBL
WHERE landing__outcome LIKE 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

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

```
%%sql
SELECT booster_version
FROM SPACEXTBL
WHERE
landing__outcome LIKE '%Success (drone ship)%' AND
payload_mass__kg_ BETWEEN 4000 AND 6000;
```

Total Number of Successful and Failure Mission Outcomes

- This is the total number of successful and failure mission outcomes.

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

```
%%sql
SELECT mission_outcome, COUNT(mission_outcome) AS COUNT
FROM SPACEXTBL
GROUP BY mission_outcome;
```

Boosters Carried Maximum Payload

- Here are the names of the booster which have carried the maximum payload mass
- A query is done inside of another query to obtain the maximum payload mass.

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

```
%%sql
SELECT booster_version
FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
```

2015 Launch Records

- These are de landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
%%sql
SELECT landing__outcome, booster_version, launch_site
FROM SPACEXTBL
WHERE
YEAR(DATE) = 2015 AND
landing__outcome LIKE '%Failure%';
```

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

- Here we rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

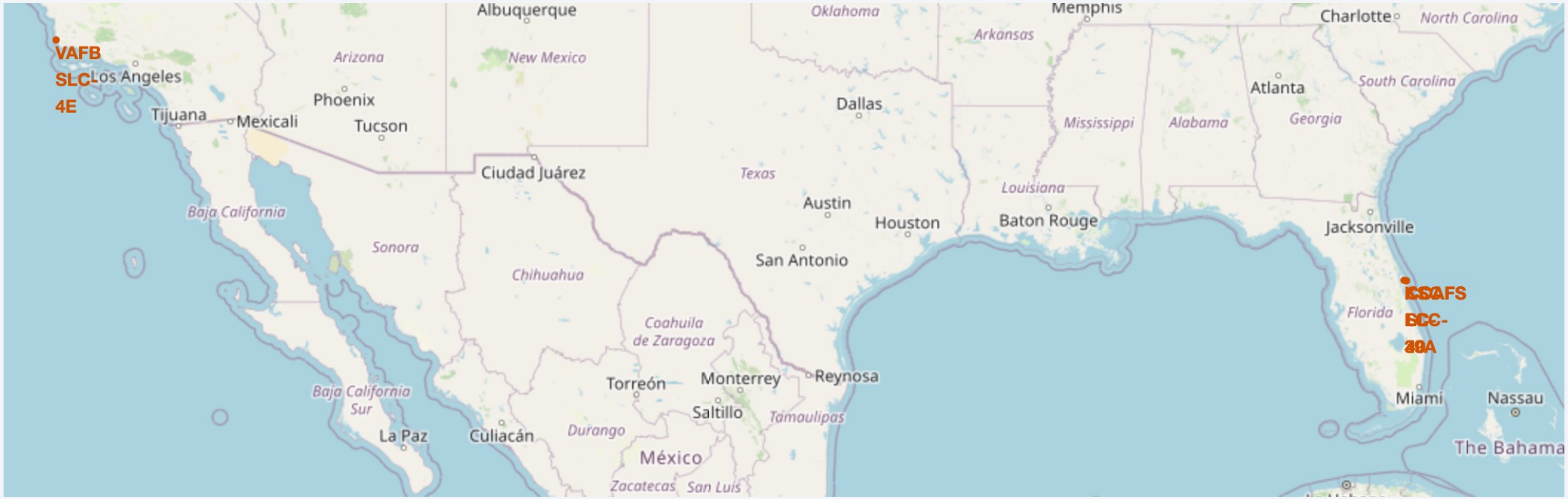
landing__outcome	counter
No attempt	9
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) AS counter
FROM SPACEXTBL
WHERE
DATE BETWEEN '2010-06-04' AND '2017-03-02'
GROUP BY landing__outcome
ORDER BY counter DESC;
```

Section 4

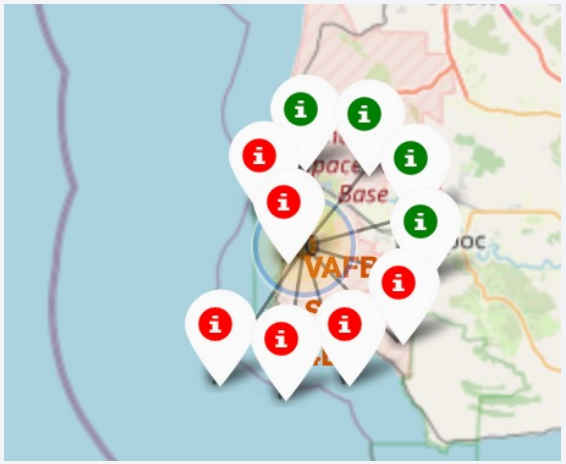
Launch Sites Proximities Analysis

Space X Launch Sites

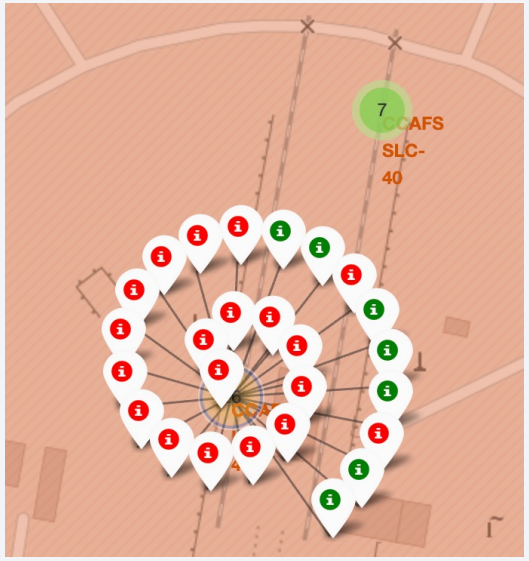


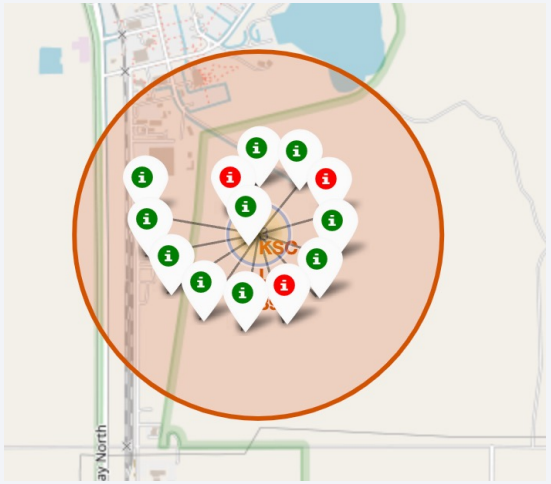
All Launch sites seem to be by the coast and somewhat close to the equator. Based on the previous EDA, Florida seems to be the location of choice for launch sites (CCAFS LC-40, CCAFS SLC-40 and LC-39A).

Success Rate Map per Location



We see that location KSC LC-39A, in Florida, has the highest success rate (lower right). The other two locations in Florida have lower success rates yet more launches (lower left) and the location in California also has a low success rate (upper rights).



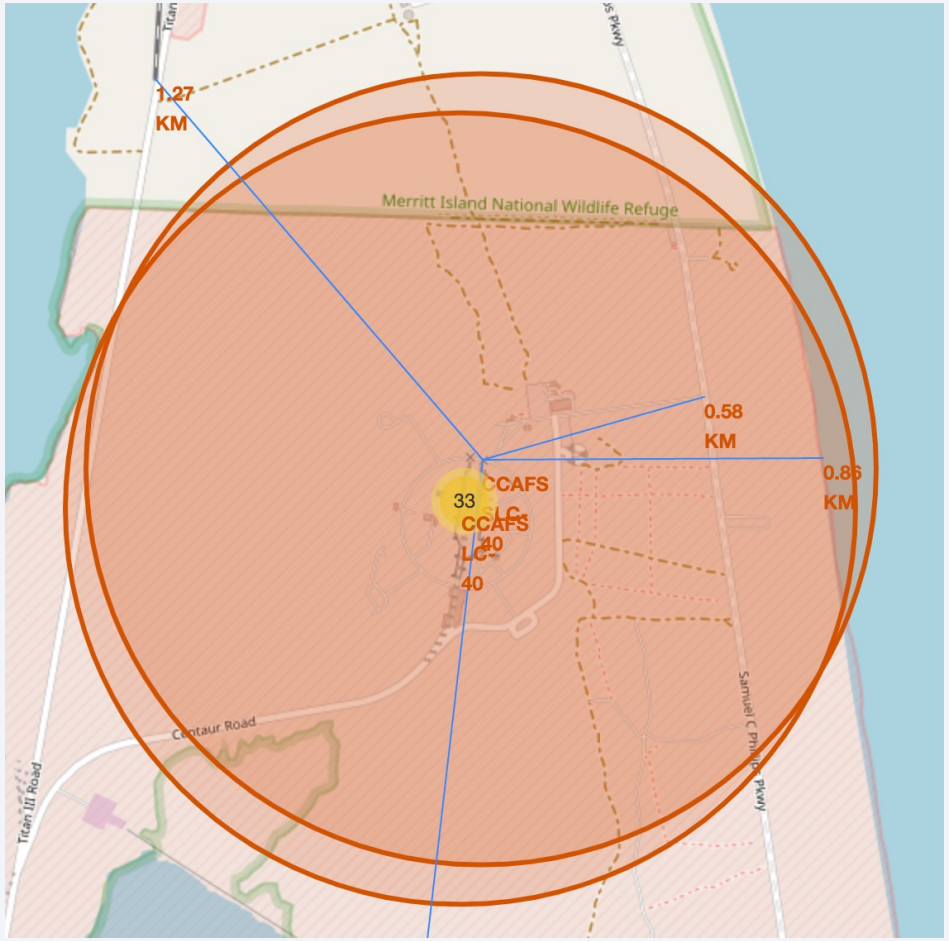


Distance Between Launch Site and Proximities

Launch Site CCAFS LC-40 and CCAFS SLC-40:

Both of these launch sites are at the same location and less than 1.5 km away from a railroad, highway and coastline.

The closest city is Melbourne at a higher distance of 51.2 km.

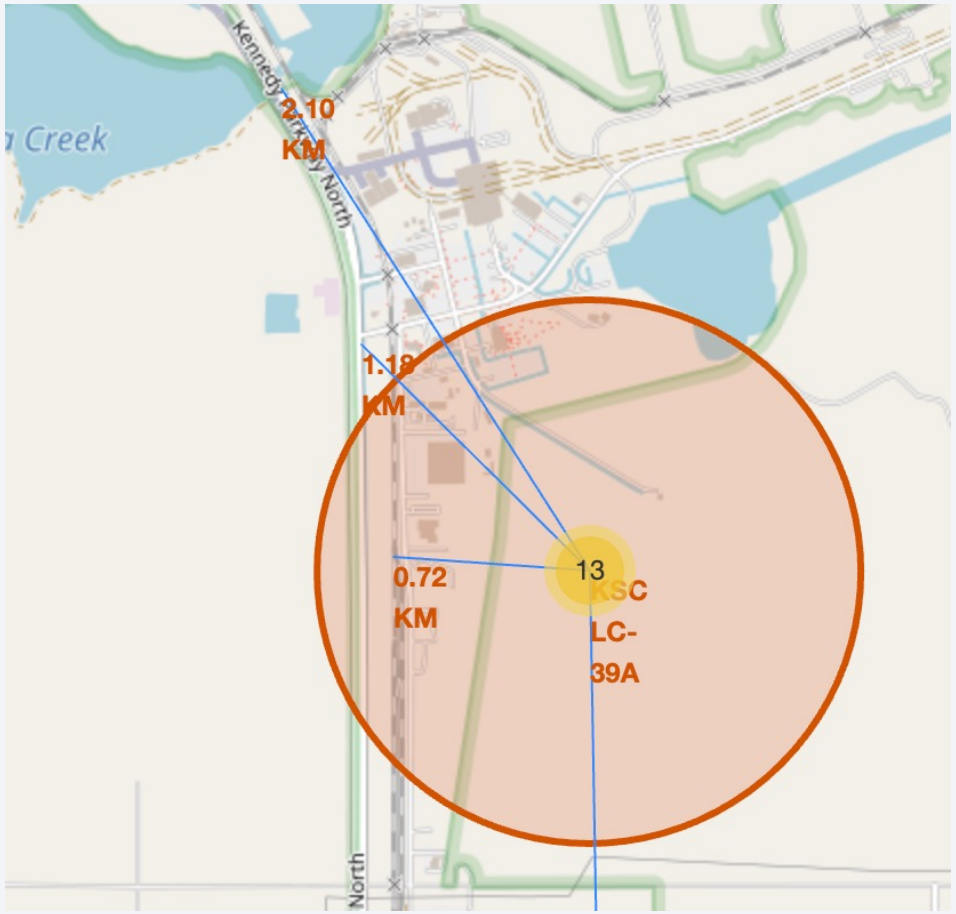


Distance Between Launch Site and Proximities

Launch Site KSC LC-39A:

This launch site is also in Florida and has less than 2.5 km from a railroad, highway and coastline, maintaining the previous slides pattern.

The closest city is also Melbourne at a higher distance of 51.96 km.



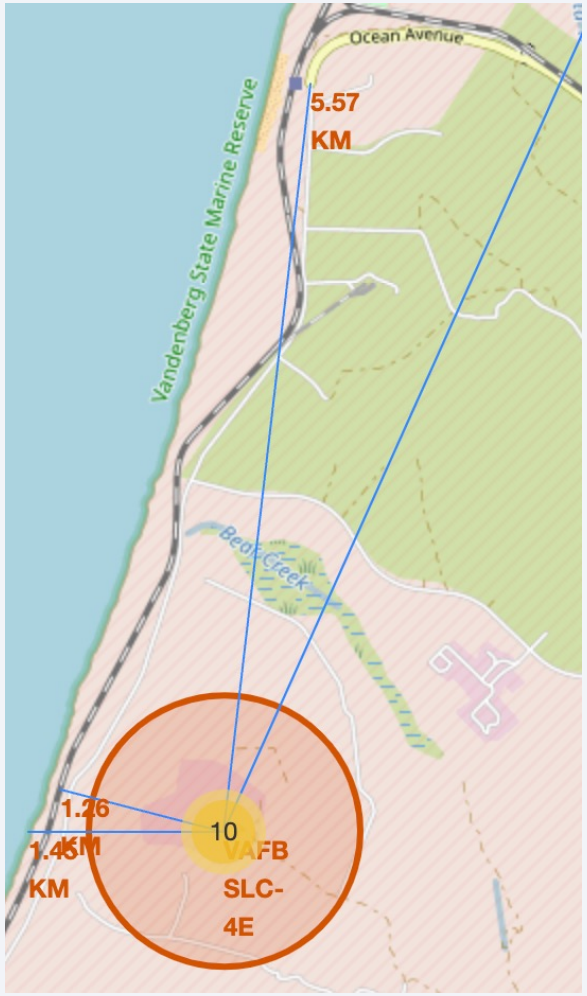
Distance Between Launch Site and Proximities

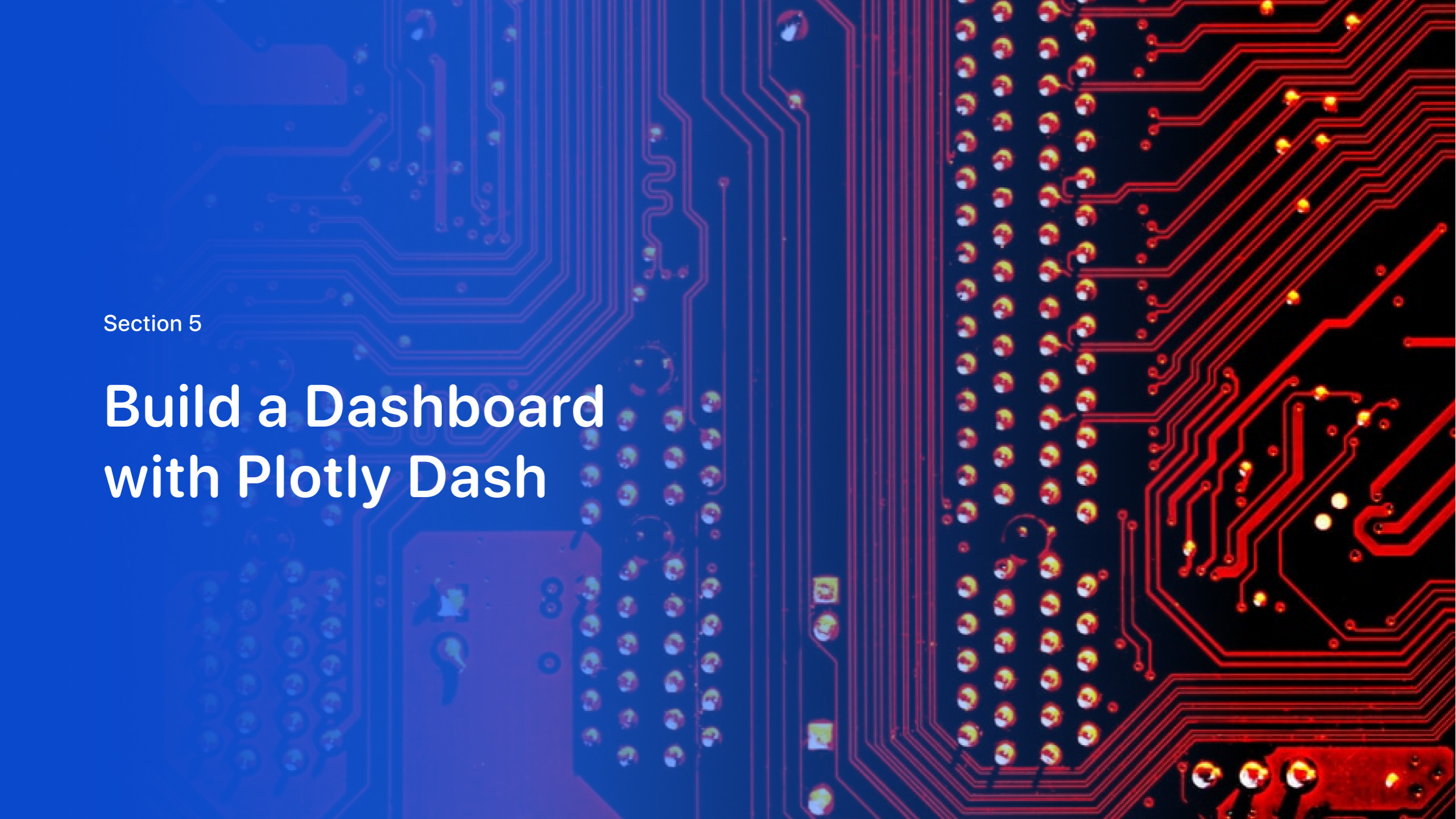
Launch Site VAFB SLC-4E in California:

This launch site has less than 5.6 km from a railroad, highway and coastline.

The nearest city is Santa Maria at 38.94 km.

We see a repeatable pattern in all launch sites: they are close to railroads, highways and coastlines, and further away from big cities.





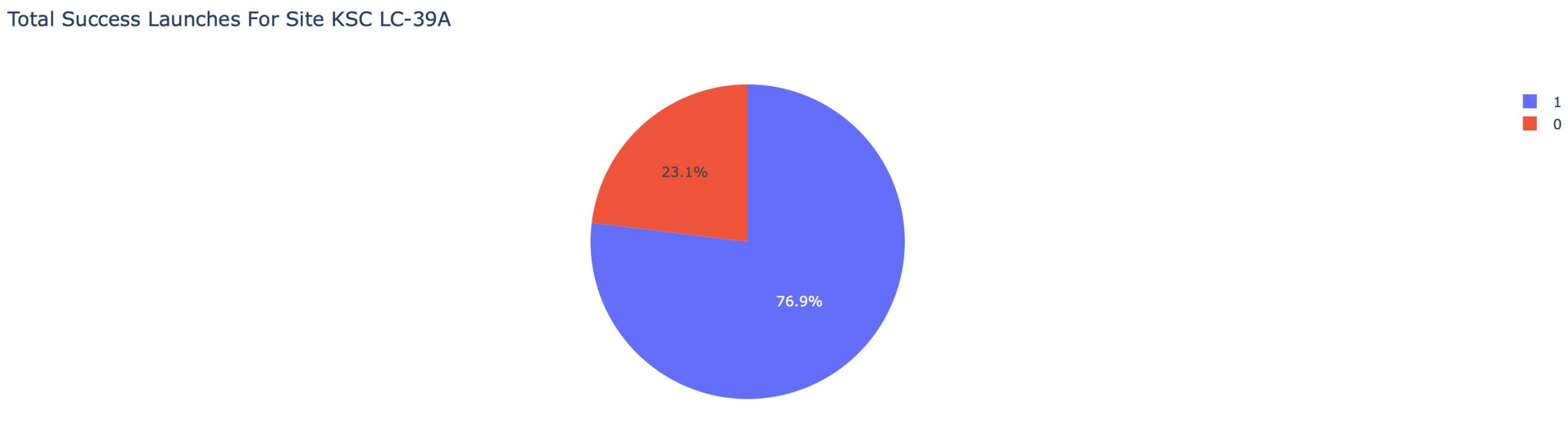
Section 5

Build a Dashboard with Plotly Dash

Successful Launches per location

We see how KSC LC-39A has the most successful Launches.

Launch Success Ratio of KSC LC-39A



We see how KSC LC-39A has the most successful launch rate, with a success rate of 76.9%, since it also has the most launches it means that this site probably has a higher chance of success.

Payload and Success Scatter Plot

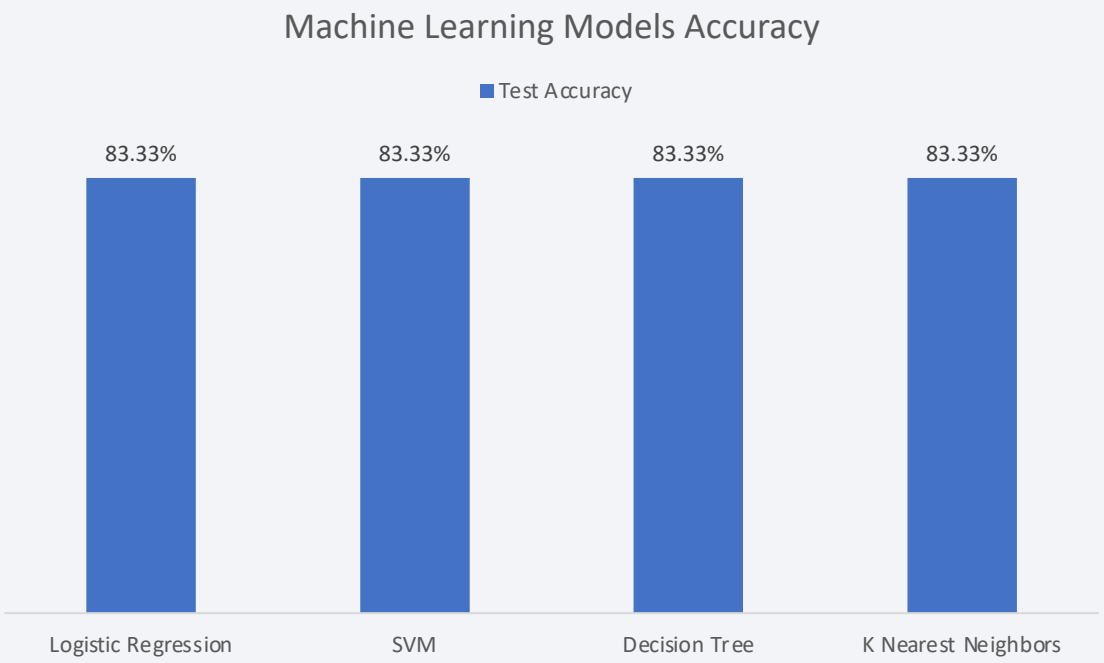
The range was reduced to 7500 kg for better visualization. We see v1.1 has seldom worked. FT is the most successful Booster Version Category.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

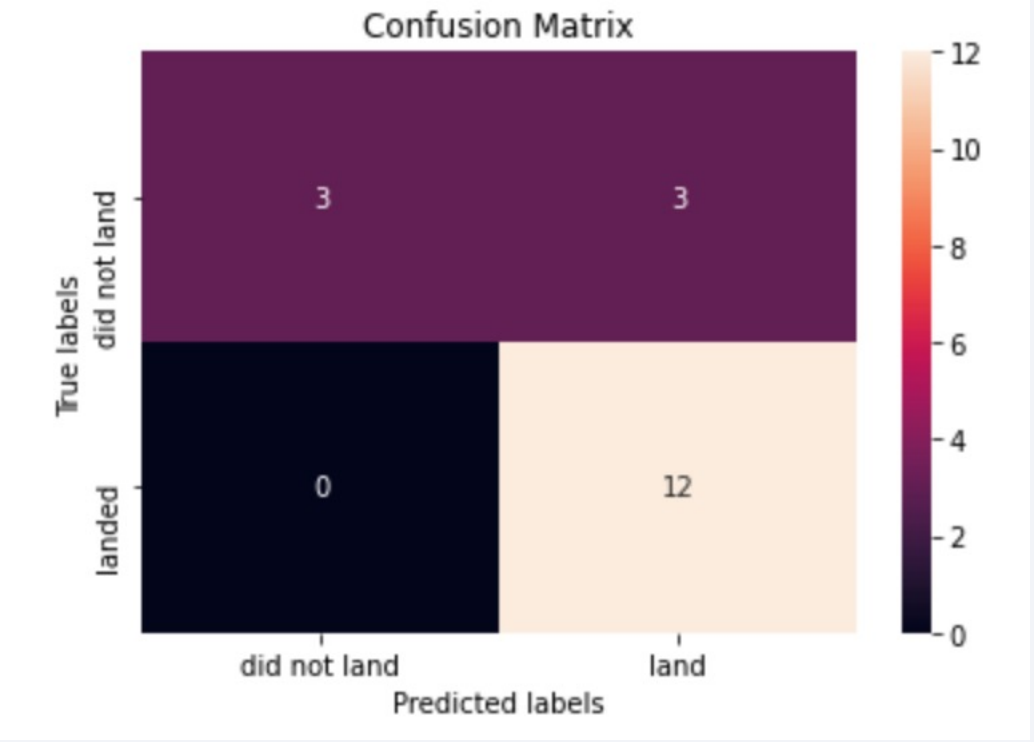
All models had different performances when selecting hyperparameters via a Grid Search, nevertheless on the test set we see that all models with their best hyperparameters had the same accuracy, meaning no model performed better than the other.



Confusion Matrix

Since all models have equal performance, the confusion matrix is the same for all.

We see that there is a problem with false positives, meaning that the model predicts that it will land and it doesn't do so.



Conclusions

- EDA was a useful tool to recognize the important variables for the model.
- An accuracy of 83.33% proves the success of classification models to predict success outcome.
- The confusion matrix shows that prediction errors fall on the False Positive prediction, this will have an effect of overestimating the cost of future launches.
- Since all models have the same accuracy, the best model to choose would be the most efficient one, making Logistic Regression the most convenient model to select.

Appendix

- Initial data obtained through the SpaceX API

	rocket	payloads	launchpad	cores	flight_number	date
0	5e9d0d95eda69955f709d1eb	5eb0e4b5b6c3bb0006eeb1e1	5e9e4502f5090995de566f86	{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	1	2006-03-24T22:30:00.
1	5e9d0d95eda69955f709d1eb	5eb0e4b6b6c3bb0006eeb1e2	5e9e4502f5090995de566f86	{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	2	2007-03-21T01:10:00.
3	5e9d0d95eda69955f709d1eb	5eb0e4b7b6c3bb0006eeb1e5	5e9e4502f5090995de566f86	{'core': '5e9e289ef3591855dc3b2626', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	4	2008-09-28T23:15:00.

Appendix

- Final Dataframe of Independent Variables before normalization

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058	Serial_B1059	Serial_B1060	Serial_B1062	GridFins_False
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	86.0	15400.000000	2.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0
86	87.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
88	89.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0
89	90.0	3681.000000	1.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0

90 rows × 83 columns

Appendix

- Folium coordinates used of nearest points to launch sites:

```
# Create a marker with distance to a closest city, railway, highway, etc.
# Draw a line between the marker to the launch site
cca_site = [28.56342, -80.57678]
city = [28.10621, -80.63716]
coastline = [28.56345, -80.56793]
highway = [28.56485, -80.57104]
railroad = [28.57208, -80.58527]
```

```
#we repeat the process for the next launch site 'KSC LC-39A'
cca_site = [28.573255,-80.646895]
city = [28.10621, -80.63716]
coastline = [28.58919, -80.65845]
highway = [28.58075, -80.65548]
railroad = [28.57374, -80.6542]
```

```
#we repeat the process for the next launch site 'VAFB SLC-4E'
cca_site = [34.632834,-120.610746]
city = [34.95209, -120.43591]
coastline = [34.63275, -120.62654]
highway = [34.68261, -120.60377]
railroad = [34.63556, -120.62411]
```

Thank you!