IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dereck Dowler
5/17/2023

Outline

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

Executive Summary

- Data was collected using a rest API from SpaceX and webscrapping was performed from Wikipedia.
- Data was then cleaned and learning classes were added to label each launch as a success or failure for learning reasons.
- EDA was performed using SQL and Scatter Plots and Bar Charts in Python. In particularly payload trends were investigated and frequencies of success/failure types was looked at.
- Dash Board was created comparing launchsite success patterns.
- Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors predictive analysis created and optimized.
- It was found that Falcon 9 launches had a 66.7% rate of landing succesfully.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Heavier Payloads had more successful landings rates with POLOR, LEO, and ISS orbit Types.
- All predictive analysis performed equally well.

Introduction

- This project looked at the question, "What factors could be used to predict whether the landing of a Space X Flacon 9 rocket landing was succesful or not."
- Factors such as payload, launch site, and orbit type where found to be important in whether or not a rocket successfully landed.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Rest API from Space X and webscrapping from Wikipedia
- Perform data wrangling
  - Data from multiple sources was first slimmed down by taking irrelevant factors and records with null values were handled. Data was then combined into one dataframe and classification factor of success/failure was added.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, decision trees, SVM, and K-Nearest Neighbor models were created and optimized using GridSearch. These models were then evaluated using confusion matrices and accuracy scores.

Data Collection – SpaceX API

- GitHub Link to Data Collection notebook:
- https://github.com/ddowler/Capstone/blob/master/data_collection.ipynb

Data Collection API FlowChart:

Make Get Request to SpaceX API ->
Store response ->
Parse response to Pandas Dataframe ->
Data formatted and repeated records removed ->
API called again to extract Booster, Payload, Launchsite, and Core Data ->
Information used to create a dictionary ->
dictionary converted into a Pandas Dataframe.

Data Collection - Scraping

- Github URL of Web Scraping Notebook:
- https://github.com/ddowler/Capstone/blob/master/data_collection_web_scrapping.ipynb

Web Scraping Flow Chart

Create Get Request from Wikipedia ->
Create Beautiful Soup Object from Response ->
Create a list of tables ->
Extract Header names from the table containing information on launches into a list->
Extract Data from the same table creating a dictionary using header names as keys->
Convert Dictionary into a Pandas Dataframe

Data Wrangling

- First, the dataframe was checked for null values. Null values were found in the PayLoadMass value. These values were replaced with the mean of the column. Null values were also found in the LandingPad column. These were left alone as they indicate no landing pad being used.
- Data was then classified into success/failure classification based on results found in the landing_outcomes column of the dataframe.
- The mean of this data was then used to determine the success rate at which a Falcon 9 Rocket landed. This was 66%.

- Github URL for the notebook:
- https://github.com/ddowler/Capstone/blob/master/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- The following scatter plots labeled by success/failure were as follows:
  - flight number vs. payload mass
    - Used to look at if success could be correlated to flight number and the mass of the payload
  - Flight Number vs. Launch Site
    - Used to determine which Launch Sites were more successful
  - Payload Mass vs Launch Site
    - Used to see if payload mass affected success landing from each launch site
  - Flight number vs. Orbit Type
    - Used to see if Success over time could be determined by the Orbit Type of a launch
  - Payload Mass vs. Orbit Type
    - Used to determine how Payload Mass affected success of each Orbit Type
- A Bar chart of Orbit Type vs Success Rates was looked at as well to compare the success rated or the different orbit types.

- GitHub URL for this notebook:
- https://github.com/ddowler/Capstone/blob/master/IBM-DS0321EN-SkillsNetworeda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- First, I queried the distinct launch sites used for visualization and dashboard analysis later
- Total and Average Payload was found to be used in future analysis
- Which boosters carried maximum payloads was queried
- Number of each landing outcome queried to investigate success rates by which type of landing used.

- GitHub URL of Notebook:

- https://github.com/ddowler/Capstone/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using Folium, I first marked the location of the launch sites
- I then marked successful and failure landings that happened from these launch sites and grouped these results into marker clusters around each launch site.
- From there I marked distances to the coast and to a neighboring railway for launch site CCAFS SLC-40. I used the polyline tool to mark these distances
- I marked these launch sites on the map to look at if geography had any correlation with success rates of landings.
- I marked distances between CCAFS SLC-40 as an exercise in using the polyline tool and finding distances given coordinates.
- Github URL of this notebook:
    - https://github.com/ddowler/Capstone/blob/master/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- In my dashboard I created a drop down menu to select launch sites and a slider to limit payload.
- A pie chart was generated by the launch site inputted by the user.
  - If all sites were selected, the pie chart displayed each sites successful landings as a percentage of the total of launch site successes. This allowed user to see which launch site was the most successful.
  - If a particular landing site was selected one could see the successes vs failures of the site displayed in the pie chart.
- A scatter plot of Payload Mass vs success or failure was charted using the launch site declared in the pulldown and limited by the payload range specified by the slider. This allows the user to look to see how payload mass affects the success chances of a flight at each launch site.
- Github URL for this Notebook:
  - https://github.com/ddowler/Capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

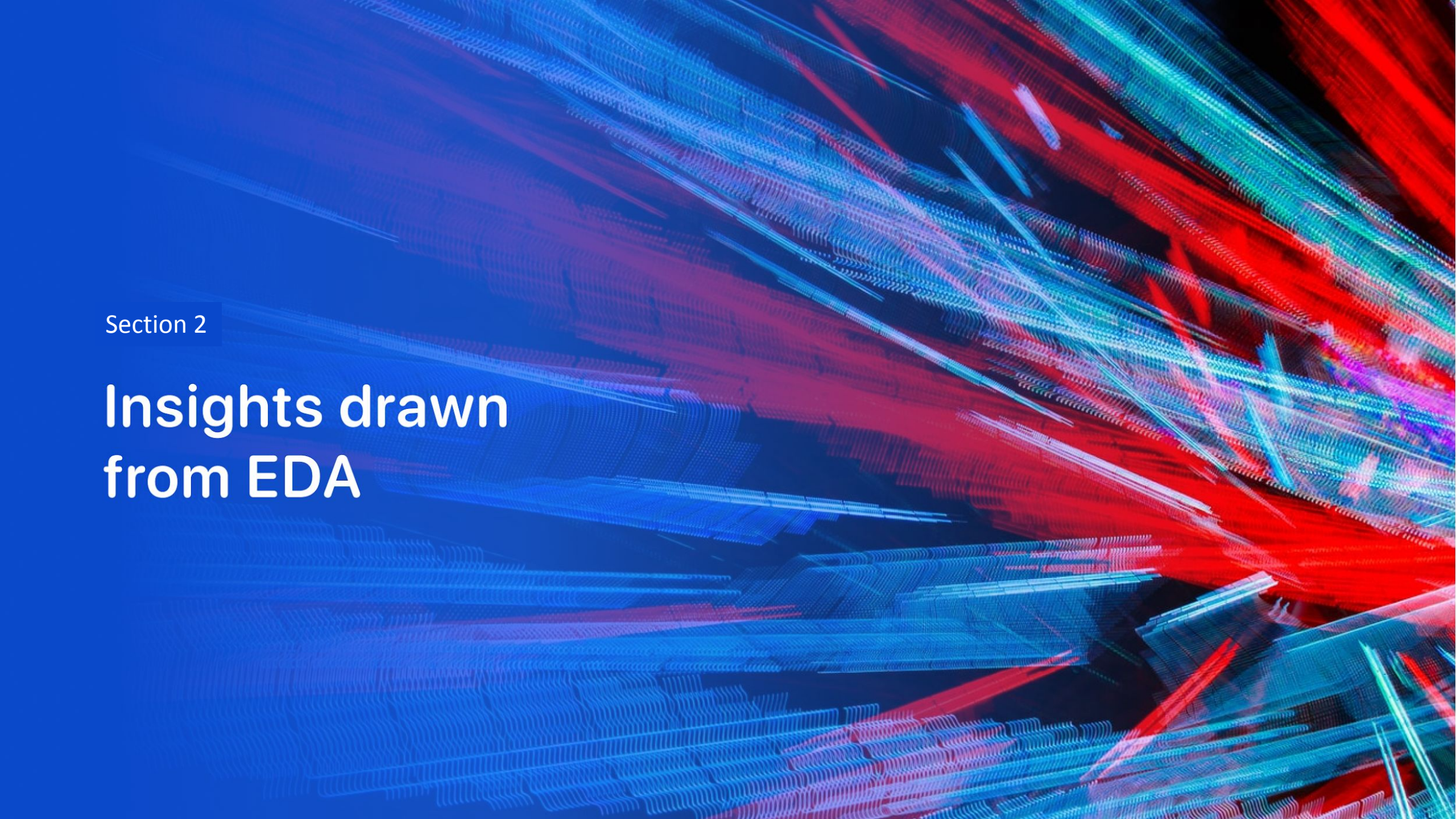
- Created Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors models to predict if a Falcon 9 rocket launch would have a successful landing.
- After optimizing and analysing these models I found they all were equally accurate in their predictions using the test data.
- Github link to the Predictive Analysis notebook:
  - https://github.com/ddowler/Capstone/blob/master/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Model Development Flowchart

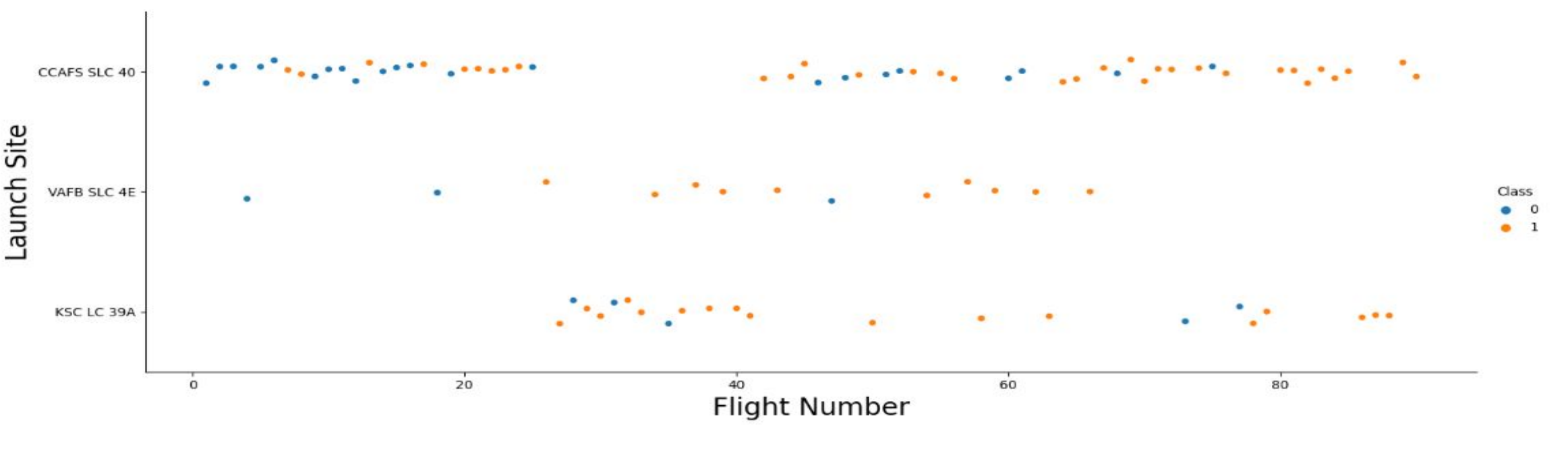
Transformed features using standard scaler ->
Split Data into Train/Test sets ->
Created Logistic Regression Model using GridearchCV to optimize parameters of model ->
Checked the accuracy of this model using test data and the score method ->
Created a confusion matrix ->
Repeated this process while creating and evaluating Support Vector Machine, Decision Tree, K-Nearest Neighbors models

Section 2

Insights drawn from EDA

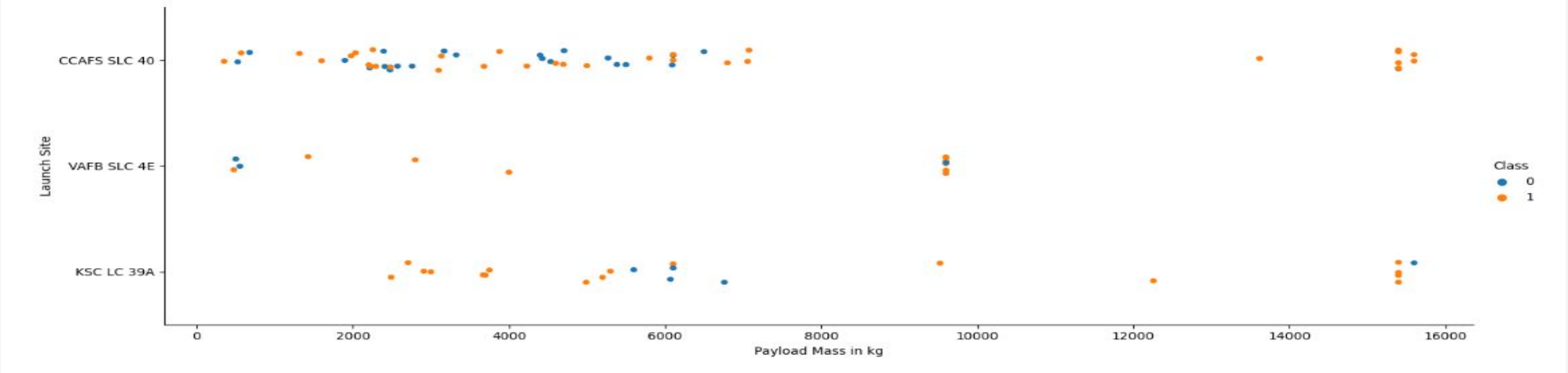


Flight Number vs. Launch Site



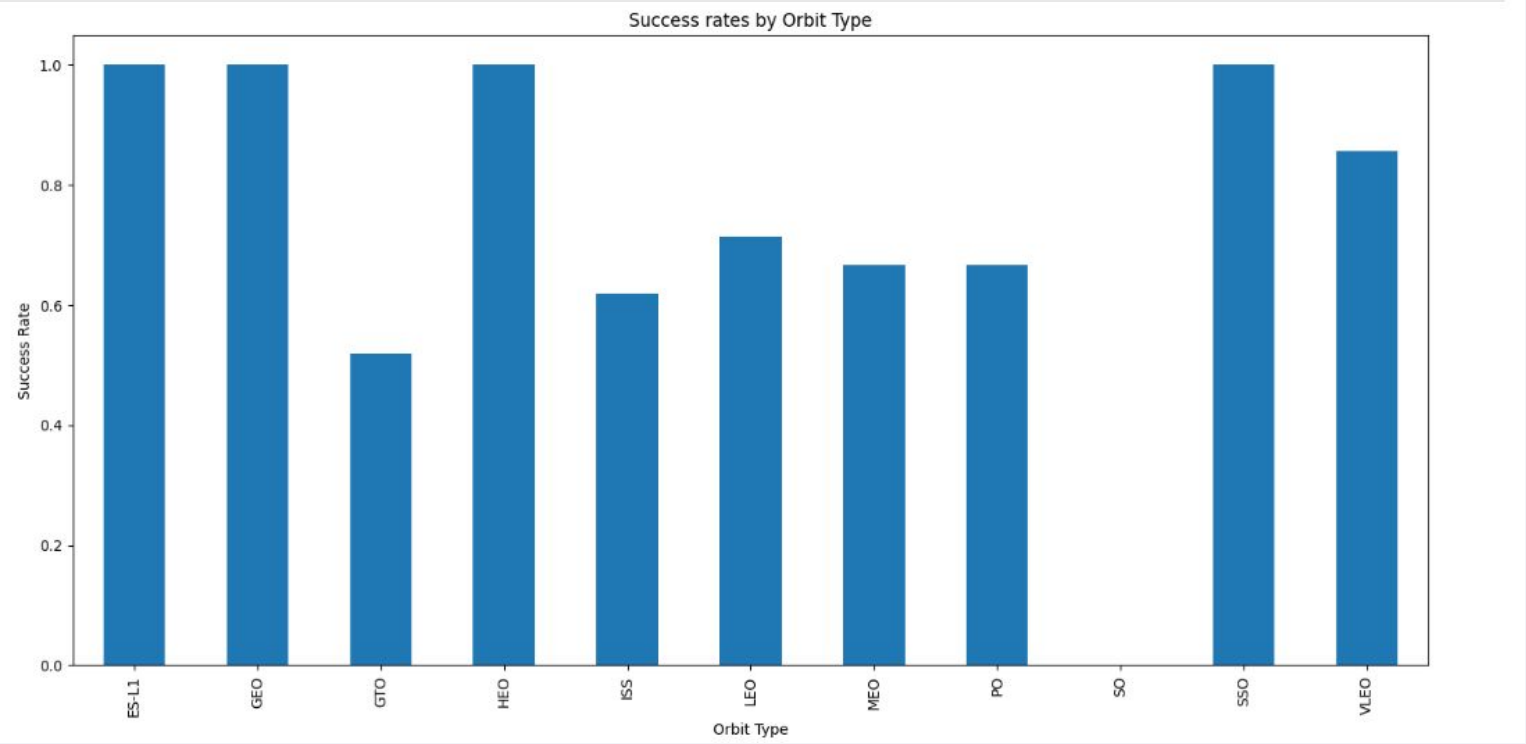
- This is a scatter chart illustrates the successful and unsuccessful launches from each Launch site as time progresses.
- In the earlier flights, there launches are mostly done from the CCAFS launch site with mostly failed landings. After about 25 launches more flights launch from the other two launch sites. After about the 65 launch, launches occur from the CCAFS and KSC launch sites with mostly successful landings.

Payload vs. Launch Site



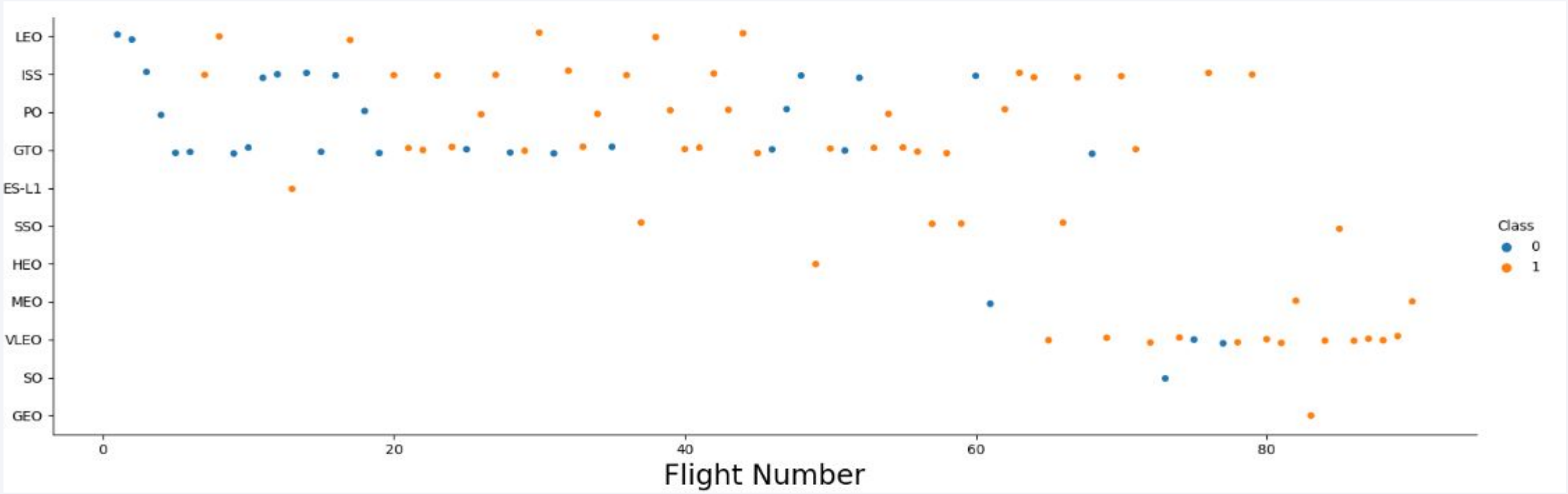
- This scatter chart illustrates the success and failure of a falcon 9 landing based on payload mass and launch site. There were no launches from the VAFB site with a payload of more than 10,000 kg. These higher payload launches had a higher percentage of successful landings than launches using smaller payloads.

Success Rate vs. Orbit Type



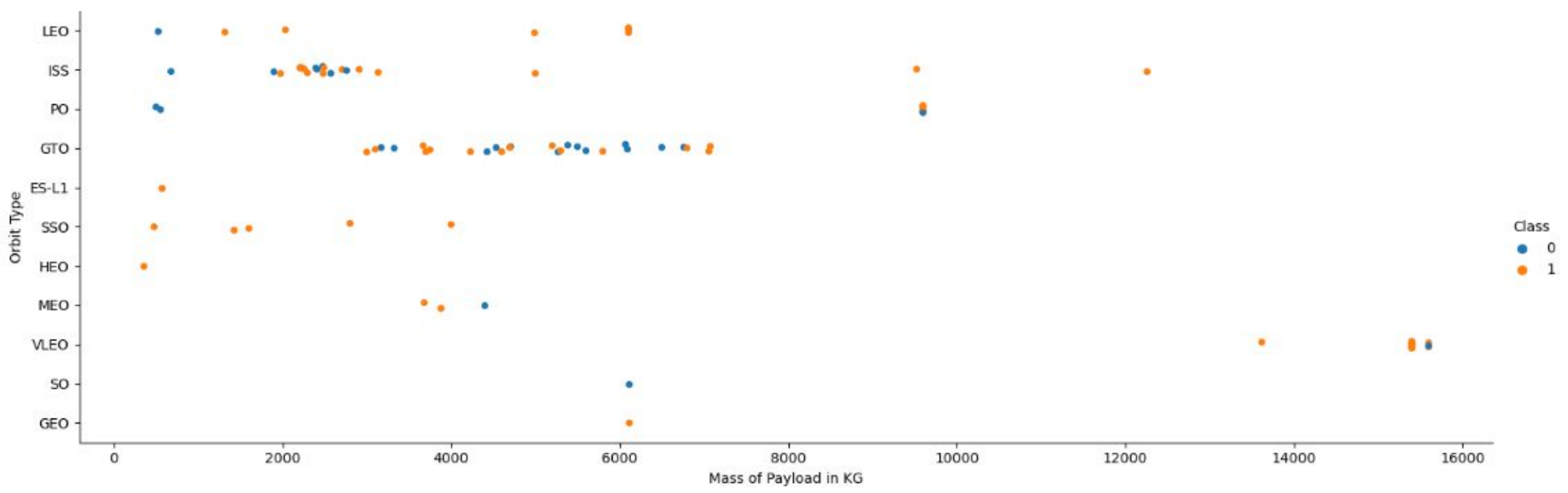
- This chart illustrates the success rate of landing a falcon 9 rocket based on its orbit type. ES-11, GEO, HEO, SSO, orbit types had a 100% landing rate. Only one rocket launched into the SO orbit type and did not land successfully. The rockets entering GTO orbit landed with about a 50% success rate.

Flight Number vs. Orbit Type



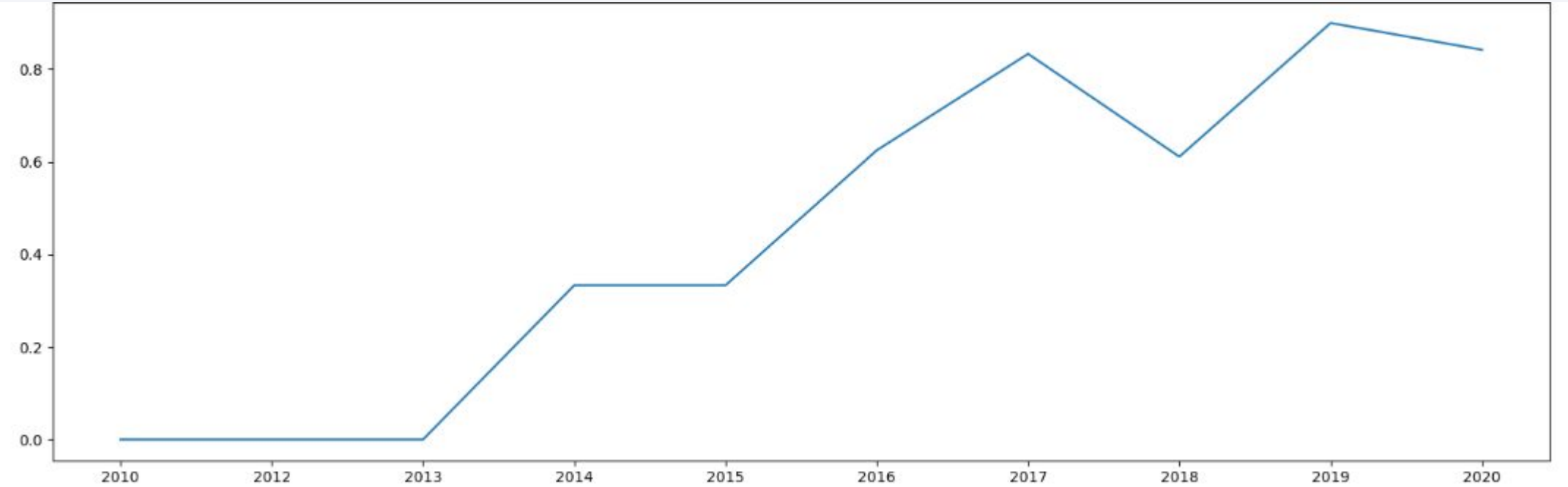
- Scatter plot looks at the successful and unsuccessful landings of falcon 9 rockets based on their orbit types over time.
- There appears to be a relationship between flight number and successful landing of flights in the LEO orbit.

Payload vs. Orbit Type



- This scatter chart looks at the success and failure of falcon 9 landings given the flights payload and orbit type.
- Flights in the LEO, PO, and ISS orbits had higher success rates with larger payloads than lighter payloads.

Launch Success Yearly Trend



- This line chart illustrates the success rate of landing a falcon 9 rocket over time
- After 2013 an overall increase in successful landings can be seen up until 2020.

All Launch Site Names

```
In [7]: %%sql

        SELECT DISTINCT("Launch_Site") FROM SPACEXTBL

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

Out[7]:

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

- This SQL query found four distinct launch sites used to launch falcon 9 rockets.

Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTBL
    WHERE "Launch_Site" LIKE "CCA%"
    LIMIT 5
```

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

Out[8]:

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

- Here are the first 5 records launched from Cape Canaveral.
- This was done with an SQL query using a where and like clause to isolate Cape Canaveral launches.

Total Payload Mass

In [9]:
```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload", Customer FROM SPACEXTBL
    WHERE Customer IS "NASA (CRS)"
```

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

Out[9]:

Total Payload	Customer
45596	NASA (CRS)

- The total payload that was launched from NASA(CRS) rockets was 45596 kg.
- This was created using the sum function while using a where clause to isolate NASA launches in a SQL query.

Average Payload Mass by F9 v1.1

In [10]:
```
%%sql

SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS "Average Payload" FROM SPACEXTBL
    WHERE BOOSTER_VERSION = "F9 v1.1"
```

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

Out[10]:

Booster_Version	Average Payload
F9 v1.1	2928.4

- The average payload of F9 v1.1 rockets was 2928.4 kg
- This SQL query used the Average function and a where clause to isolate v1.1 flights.

First Successful Ground Landing Date

```
In [11]:
%%sql

SELECT MIN(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) As "Min Date" from SPACEXTBL
    WHERE  "Landing _Outcome"='Success (ground pad)'

 * sqlite:///my_data1.db
Done.
Out[11]:
```

Min Date
20151222

- The first time a falcon 9 rocket landed successfully on a ground landing pad was December 22, 2015.
- A SQL query including the min function on the date of launch and and a where clause was used to isolate successful ground landings.
- The date column in the table needed to be reformatted so it could be ordered chronologically.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [12]:
```
%%sql

SELECT DISTINCT(Booster_Version), "Landing _Outcome", PAYLOAD_MASS__KG_ FROM SPACEXTBL
    WHERE ("Landing _Outcome" = "Success (drone ship)") AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

Out[12]:

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- Booster Versions F9 FT B1022, F9 FT B1026, F9 FT b1021_2, and F9 FT B1031_2 had successful drone ship landings and had payload masses between 4000 and 6000 kg.
- This SQL query was created using the distinct clause on booster version to avoid duplicates and a where clause to isolate launches with successful drone shop landings in the specified payload range.

Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL
    GROUP BY Mission_Outcome
```

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

Out[13]:

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

- There were 100 successful missions and 1 failed mission.
- This was found by a SQL query using Group By on the Mission_Outcome feature.

Boosters Carried Maximum Payload

```
SELECT Booster_Version, PAYLOAD_MASS__KG_ AS "Payload Mass" FROM SPACEXTBL
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Out[14]:

Booster_Version	Payload Mass
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

- There where 12 Booster Versions carrying the max payload of 15600 kgs.
- This query was completed using a subquery in a where clause in order to compare the payload mass of each booster version to the maximum payload mass.

2015 Launch Records

```
SELECT substr(Date,4,2) AS "Month", "Landing _Outcome", Booster_Version, Launch_Site FROM SPACEXTBL
    WHERE ("Landing _Outcome" = "Failure (drone ship)") AND (substr(Date, 7, 4) = '2015')
```

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

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

- Booster Versions F9 v1.1 B1012 and F9 v1.1 B1015 had failed landing outcomes onto a drone ship.
- This SQL query was created by using a where clause checking for failures from a drop ship from the year 2015. A substring was used to extract the year 2015 and month from the date feature.

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

```
SELECT "Landing _Outcome", COUNT("Landing _Outcome") FROM SPACEXTBL
    GROUP BY "Landing _Outcome"
    HAVING (substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) BETWEEN "20100406" AND "20170320"
    ORDER BY COUNT("Landing _Outcome") DESC
```

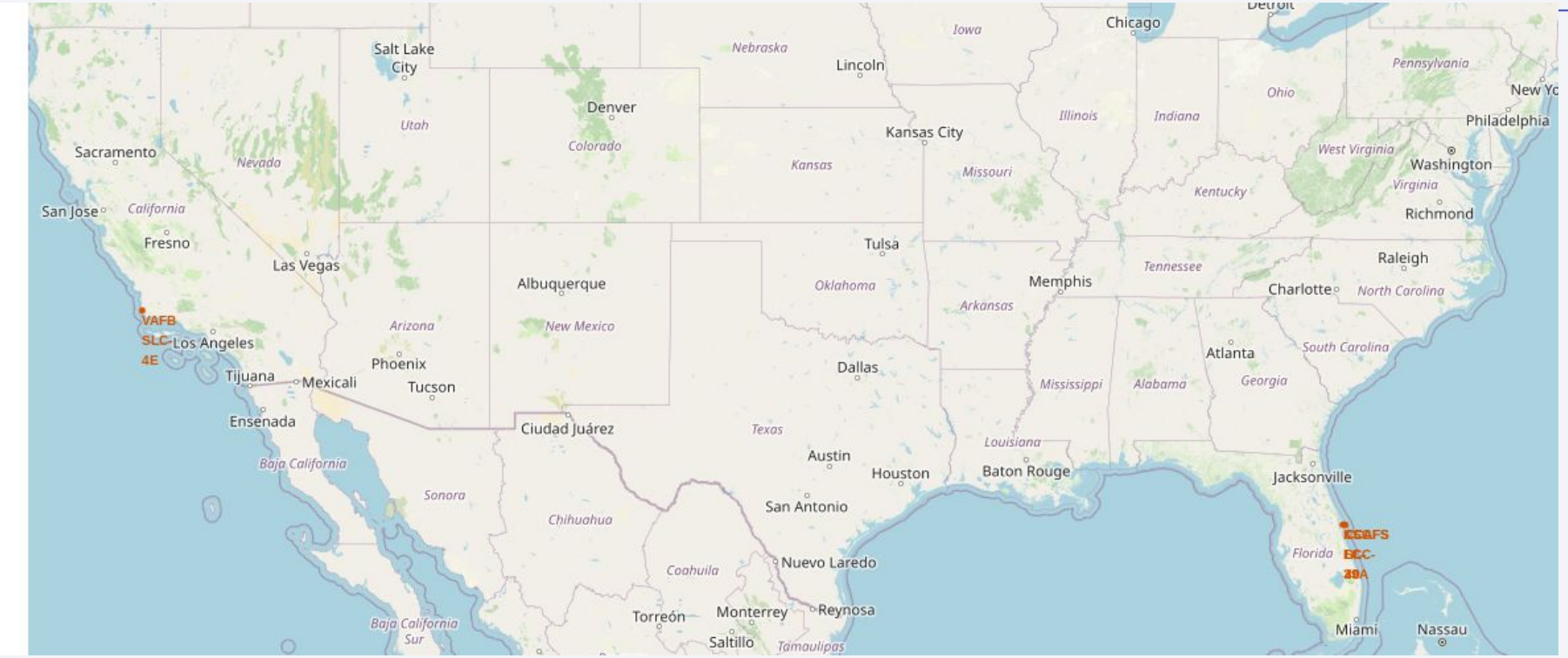
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	COUNT("Landing _Outcome")
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Above is a SQL query listing the number of each type of landing outcome. This list is then ranked by the number of each landing type.
- This was done by using a Group By clause to count the number of each type of landing outcome. A having clause was then needed to isolate the results within the specified dates. Finally, an order by clause was added to sort the results.

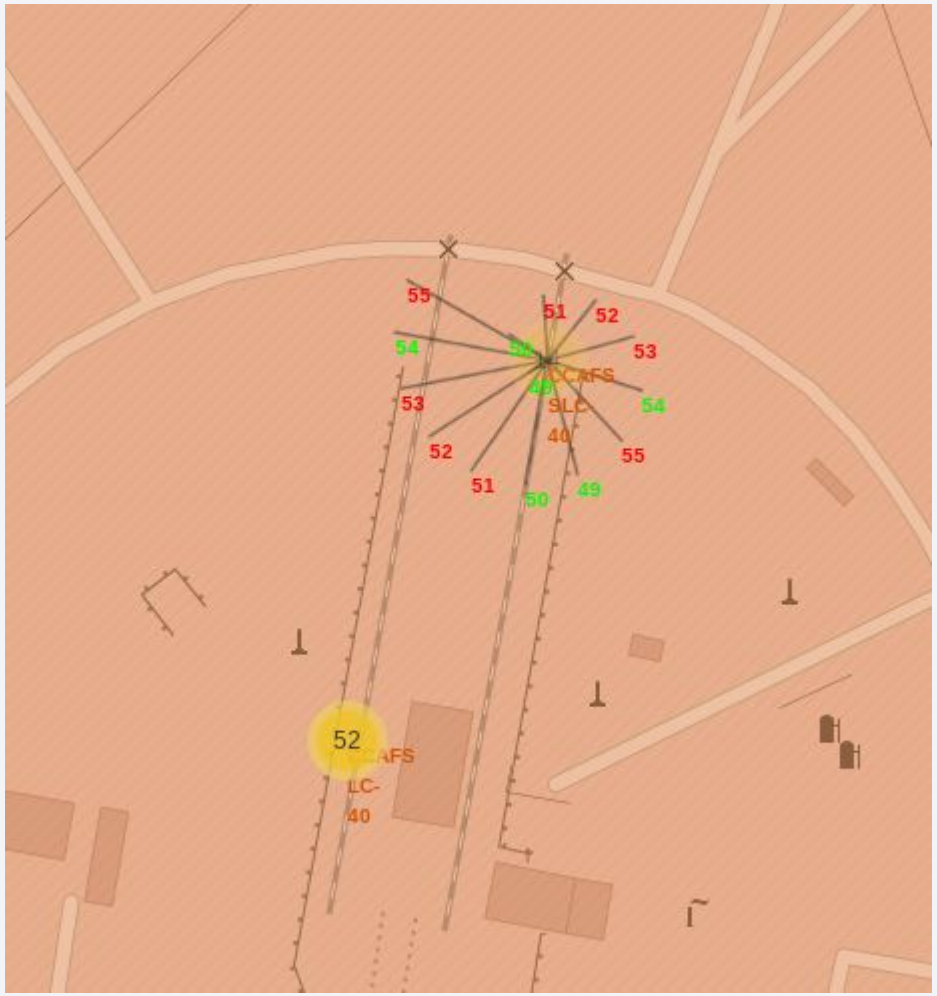

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations



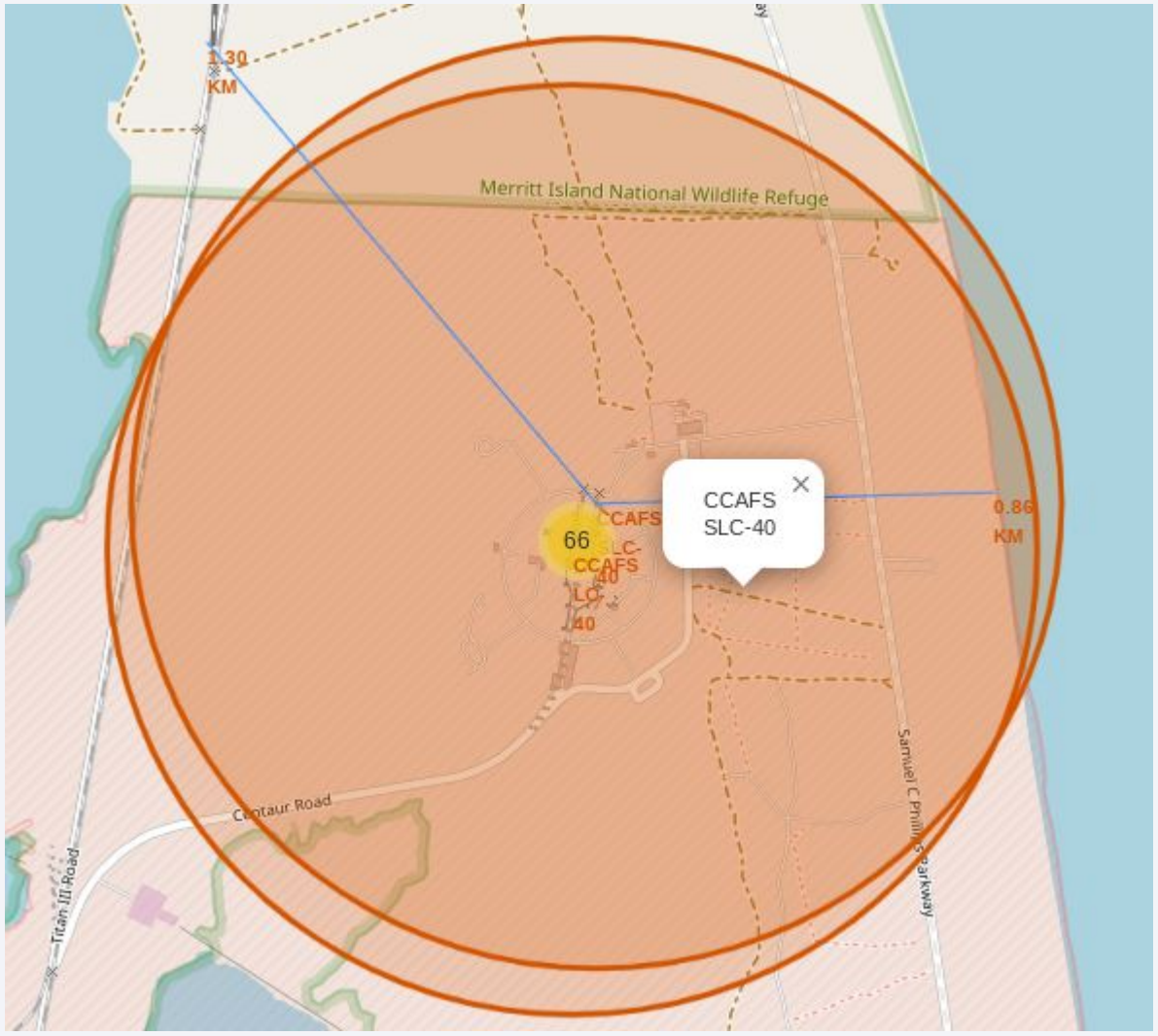
- This map was created using Folium. The launch sites are in red on the coast of Florida and California.

Successful and Failed Landings from CCAFS SLC-40

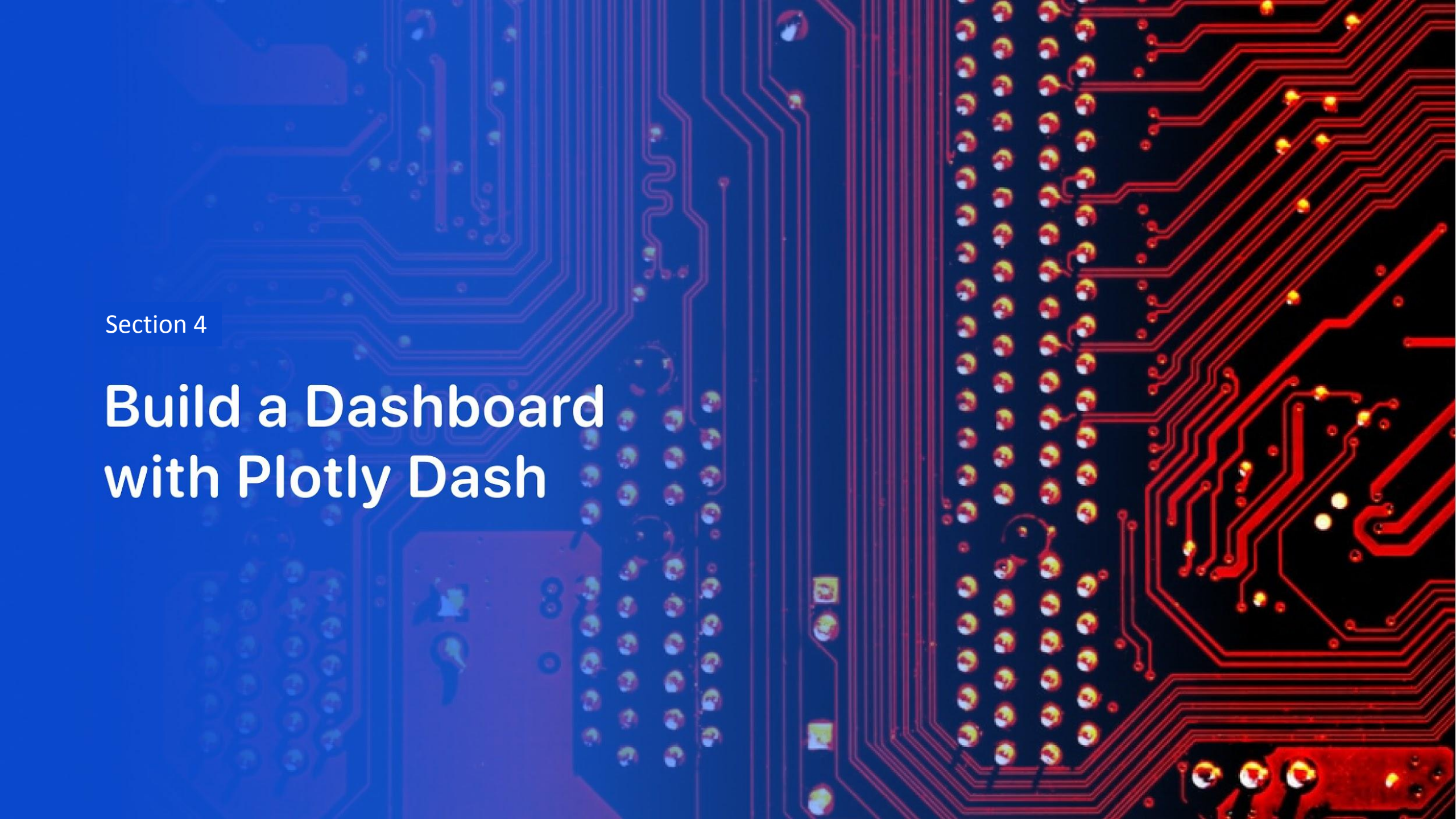


- This map created using the folium library shows successful and failed landings launched from this site.
- There are slightly more failed landings(in red) than successful landings(in green)

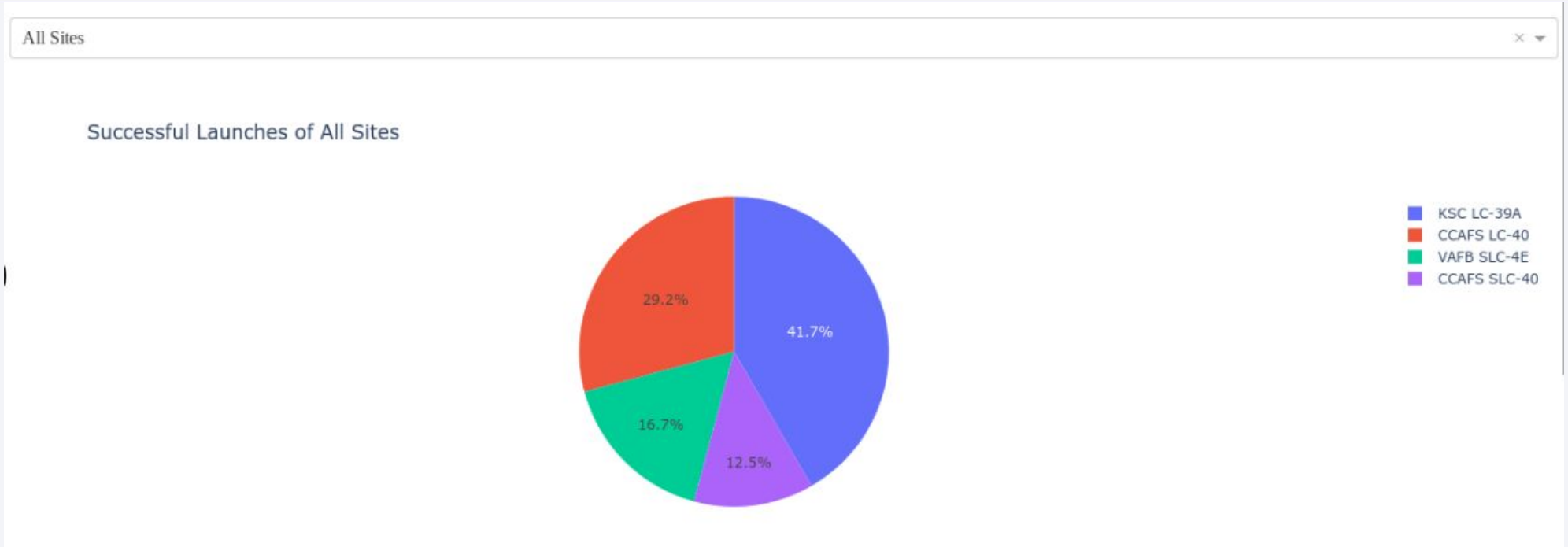
Coastal Distance and Railway Distance from Site CCAFS SLC-40



- This folium map shows the distances from CCAFS SLC-40 launch site to the Atlantic coast and the nearest railway.
- Distance are labeled with the blue Polyline and red text.

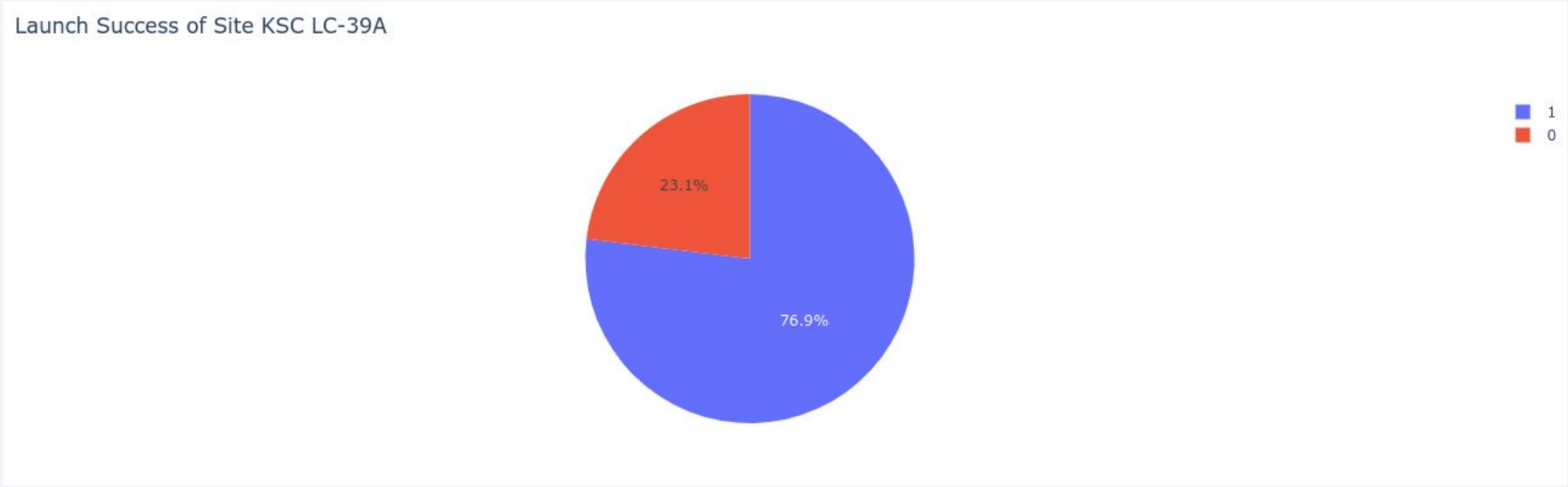

Section 4
Build a Dashboard with Plotly Dash

Successful Landings by Launch Site



- This pie chart shows the successful landings of each launch site as a percentage of all launches. KSC LC-39A has more successful launches than any other site while CCAFS SLC-40 has the least amount of successful landings.

Launch Success and Failures of KSC LC-39A



- KSC LC-39A had the largest success rate of landing the falcon 9 booster than any other launch site. 76.9% of launches from this site had successful landings.

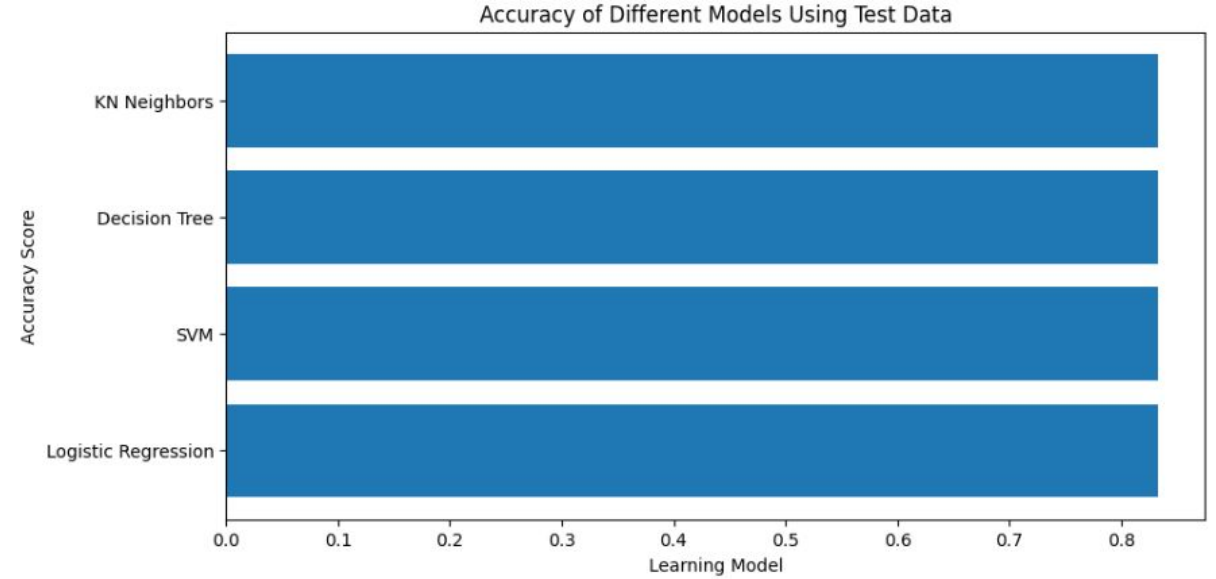
Landing Success Rate Within Specified Range

- This scatter plot shows the landing success and failure of each launch based on launch site and payload mass. Between 3000kg and 4000kg it can been seen that there are more successful landings than failed landings.

Section 5

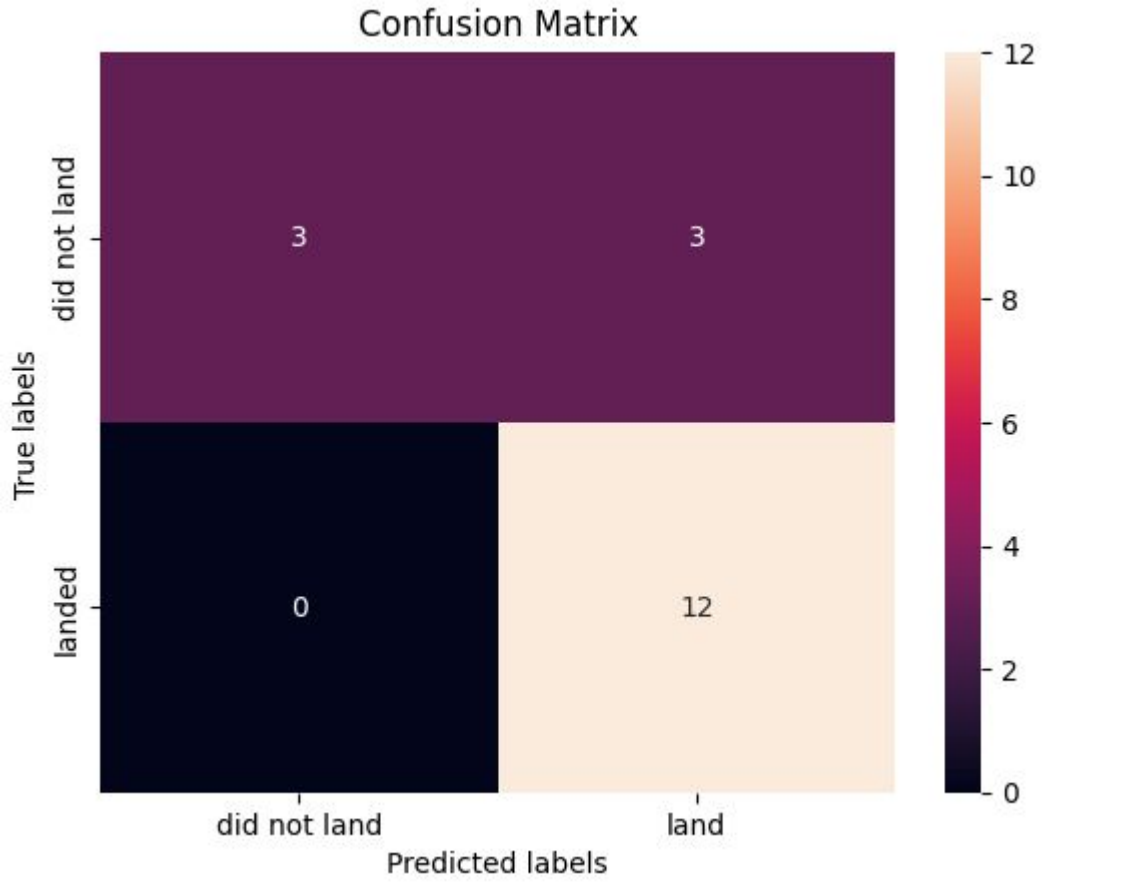
Predictive Analysis (Classification)

Classification Accuracy



- This Bar Chart illustrates the accuracy of each of the classification models used in this project. Each model performed equally well on the test data.

Confusion Matrix



- Each classification model produced the same resulting confusion matrix. The models had an issue with false positive predictions.

Conclusions

- Launch Site KSC-LC 39A had the highest percentage of successful landings with a 76.9% success rate.
- Both KSC-LC 39-A and CCAFS had high success rates of landings after launch 65.
- High landing success rats occurred when the payload reached above 9500kg.
- ES-11, GEO, HEO, SSO, orbit types had a 100% landing rate. But, ES-11, GEO, and HEO orbits were only reached from one launch, SSO orbit was reached with 5 launches.
- Success rate of landing has increased since 2013.
- Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbor models predicted landing success equally well. Though, they did have some problems with false positives.

Appendix

- Falcon 9 Data sets used from IBM skills network labs, SpaceX API, and Wikipedia. URL’s from these data sets can be found in the notebooks linked in this report.

Thank you!