IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kavita Tiwari
15 March 2022

Outline

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

Executive Summary

The goal of this project is to analysis SpaceX launches, with focus on stage re-use and landings, to provide our company SpaceY, competitive information on cost of launches and predictions about whether SpaceX will re-use the given first stage.

This report details launch and landing data collection & exploration methods, followed by exploratory & visual analysis results, and finally the results of training classification models on our data to predict the landing success. Therefore, to determine the cost of launch and re-use of first stage.

Introduction

Project background and context.

SpaceY is a new rocket company that would like to compete with SpaceX, with focus on cost of launches and prediction methods whether SpaceX will re-use a given stage.

Common problem that needed solving.

- -What influences if the rocket will land successfully?
- -The effect of each relationship with certain rocket variables will impact in determining the success rate of successful landing.
- -What condition does SpaceX have to achieve to get the best results.

Section 1

Methodology

Methodology

Executive Summary

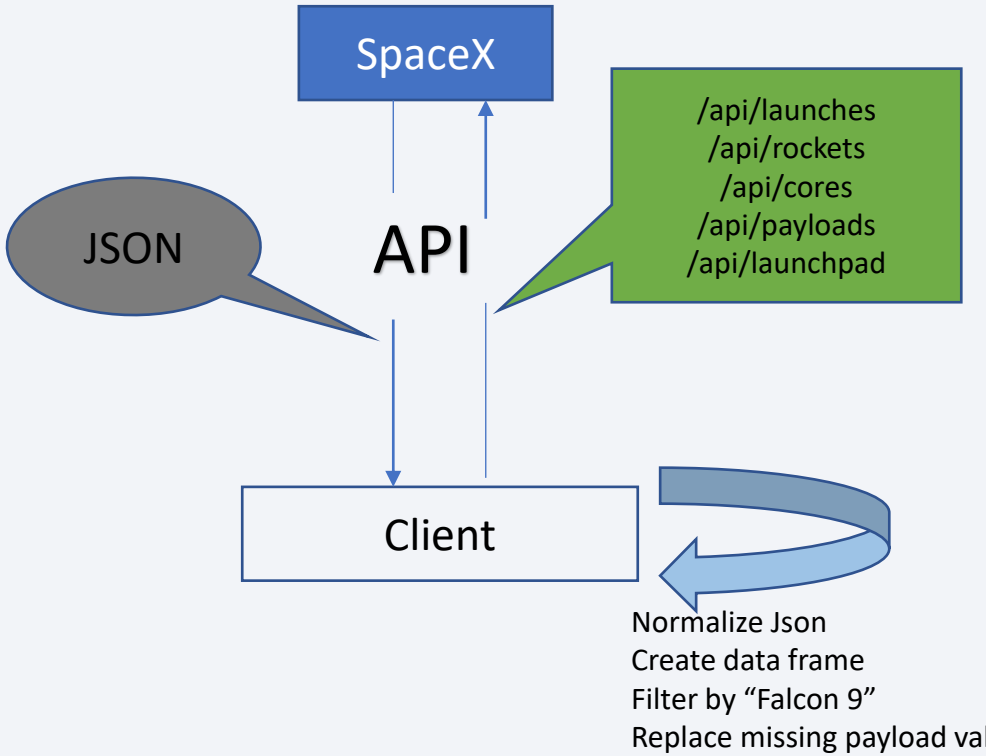
- Data collection methodology:
  - SpaceX Rest API: Past landing data including rocket used, payload deliveries, launch specification, landing specification & outcome.
  - Wikipedia Scrap: Parse HTML tables of past launches into dataframe for further analysis.
- Data wrangling
  - Outcomes(“True ASDA”,”True Ocean”,etc) grouped into two classes – success(1) and failure(0) to be used for model building and prediction.
- Exploratory data analysis (EDA) using visualization and SQL
  - Uploaded data into IBM DB2 instance and executed SQL queries from jupyter notebook.
  - Plotly and Seaborn; categorical plots (scatter, bar, line);
- Interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train and test 4 models: Logistic Regression, Support Vector Machine(SVM), KNN, Decision Tree.

Data Collection

- **REST API**
  - Gathered SpaceX launch data from REST API.
  - Normalized JSON and converted into a pandas Data Frame.
- **Web scraping**
  - Using the python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon9 launch records.
  - Parsed those HTML tables and converted into pandas data frame for further analysis.

Data Collection – SpaceX API

- Data on past launches, launchpads, cores, payloads and rocket collected using API's as JSON, normalized into Data frame.
- Then filtered the data frame to Falcon9 entries only.
- Replaced missing payload values by payload mean.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Data%20Collection.ipynb

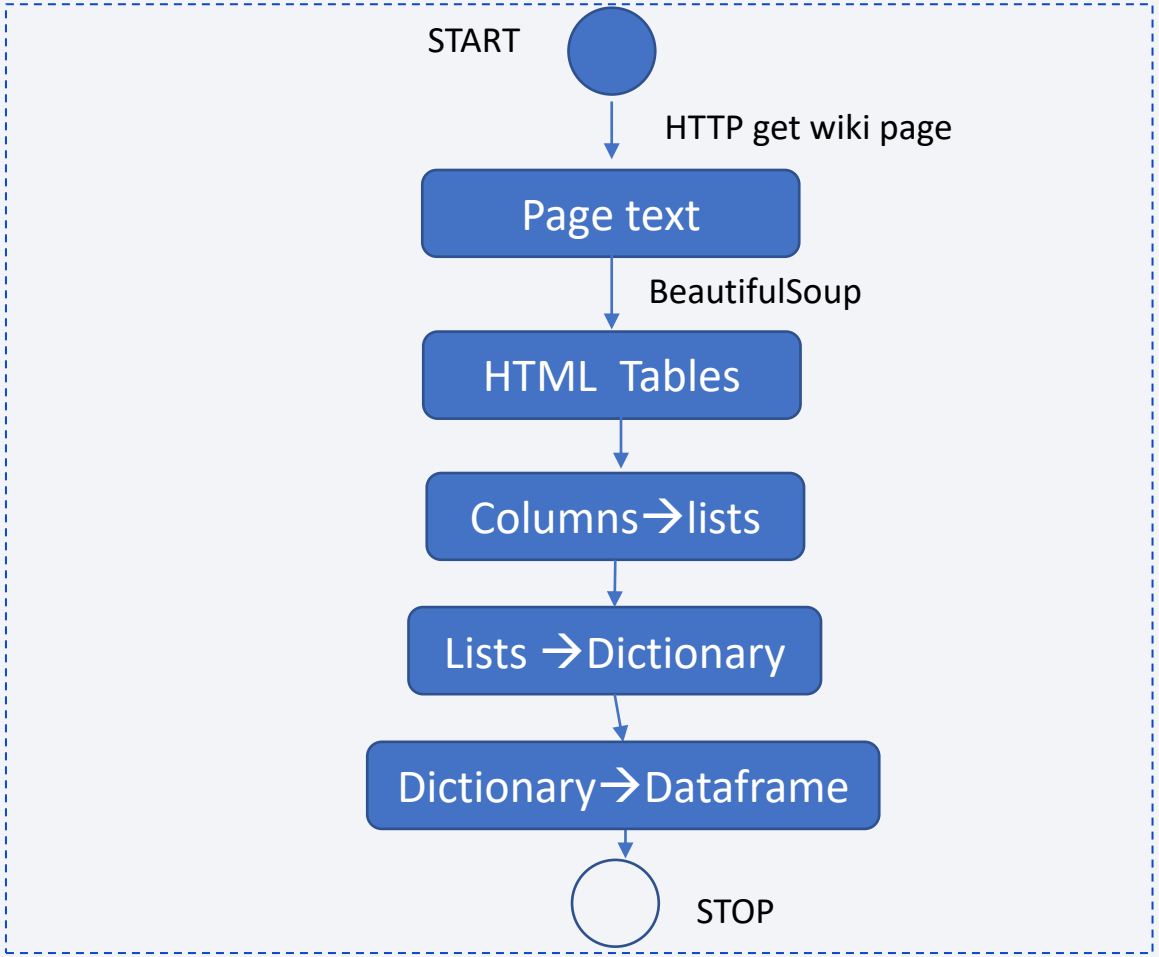


Data Collection - Scraping

- The text of the wikipedia entry “List of Falcon 9 and Falcon Heavy launches” is retrieved and parsed into a searchable HTML structure.

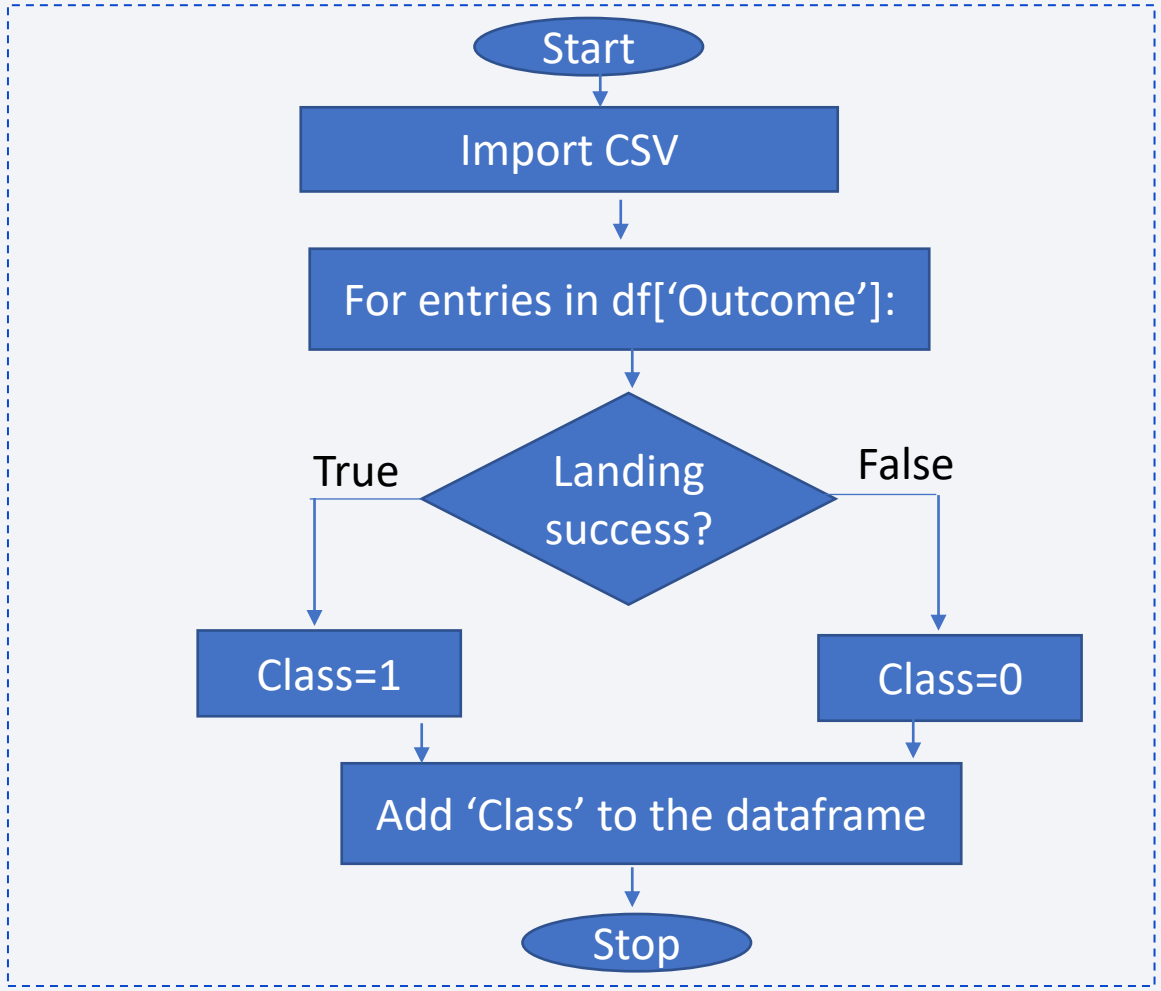
- Columns from each table listing past launches are converted into lists of values, joined into a dictionary, and converted to a DataFrame.

- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Data%20Collection-web%20scraping.ipynb



Data Wrangling

- The source csv data contains 8 different outcomes, out of which 5 were failure like None None, False ASDS.
- Converted the value of Outcome column to success(1) or failure(0) and assigned it to new column called Class.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Data%20Wrangling%20-%20Landing%20Outcome.ipynb



EDA with Data Visualization

Charts Plotted:

- Flight # Vs PayloadMass (Scatter chart)
  - Show payload trend over time (with success trends)
- Flight # Vs Launch sites (Scatter chart)
  - Shows launches at each sites (with success trends).
- Payloads Vs Launch sites (Scatter chart)
  - Shows payload distribution at each sites
- Success rate per Orbit (Bar chart)
  - Able to visualize success trends among different orbit types.
- Flight # Vs Orbit (Scatter chart)
  - Success trends in which orbits are higher with increased number of flights.
- Success rate by year (Line chart)
  - Increased success rate over time.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/EDA%20using%20Pandas%20and%20matplotlib.ipynb

EDA with SQL

- Uploaded the data in DB2
- Below SQL queries were ran:
  - List the name of unique launch sites.
  - *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. Using 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.*
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Exploratory%20Data%20Analysis-1.ipynb

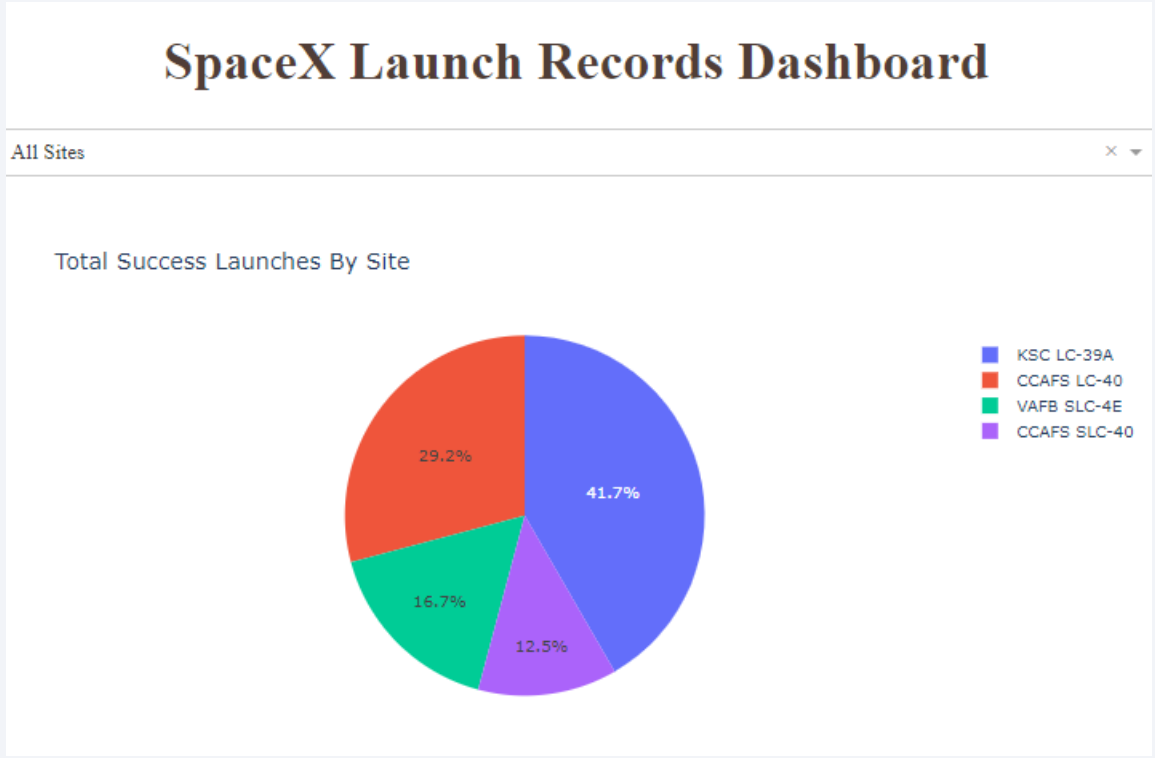
Build an Interactive Map with Folium

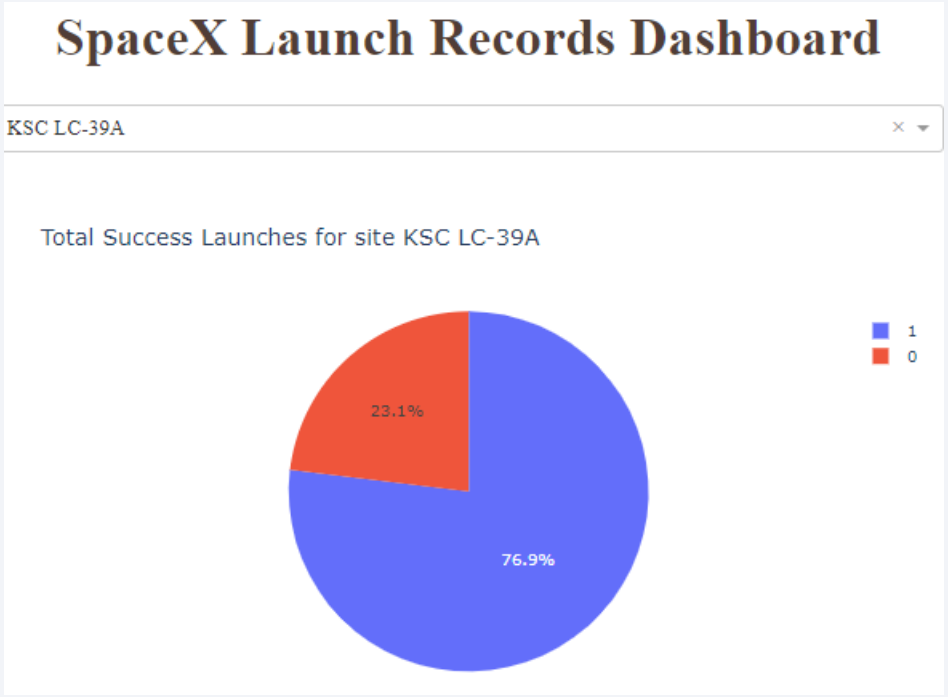
- Map objects added:
  - Circle – location of each launch site.
  - Marker – label each launch site name.
  - Popup - shows more info when mouse click
  - Marker Cluster – depicting all launches successes/failures at each site.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Interactive%20Visual%20Analytics%20using%20Folium.ipynb

Build a Dashboard with Plotly Dash

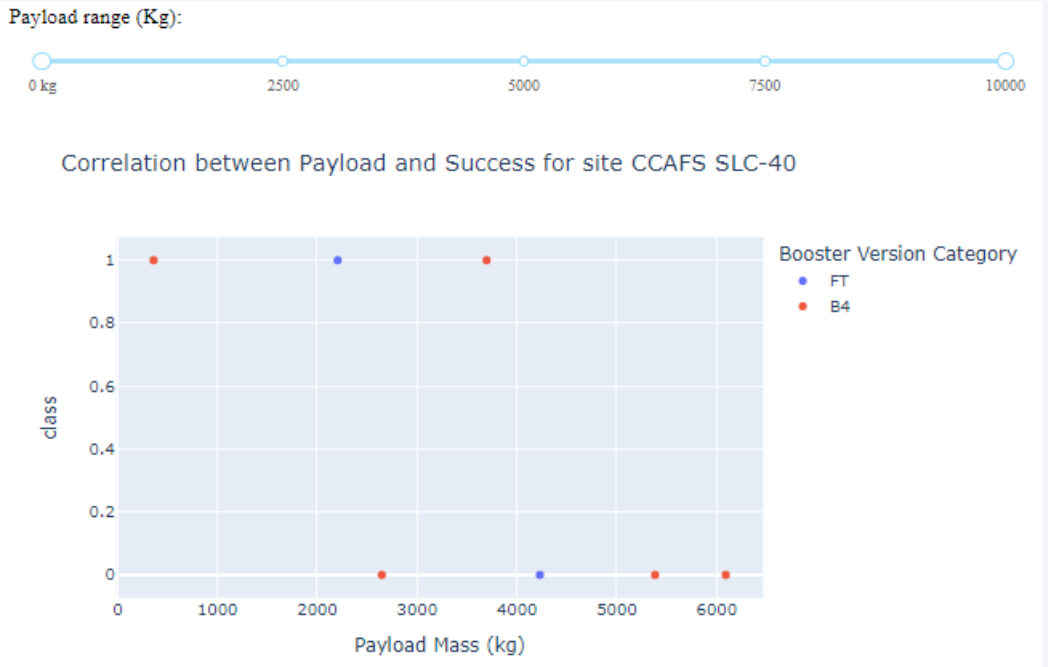
- The Dashboard shows:
  - Interactions : dropdown for launch sites and Payload;
  - Plots :
    - Launch sites (pie charts)
    - Payload Mass (scatter plot)Explain why you added those plots and interactions
- Helped in providing quick overview of relationship between launch sites and payload range success/failures.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/spacex_dash_app.py

Plotly Dashboard screen shots – Launch sites



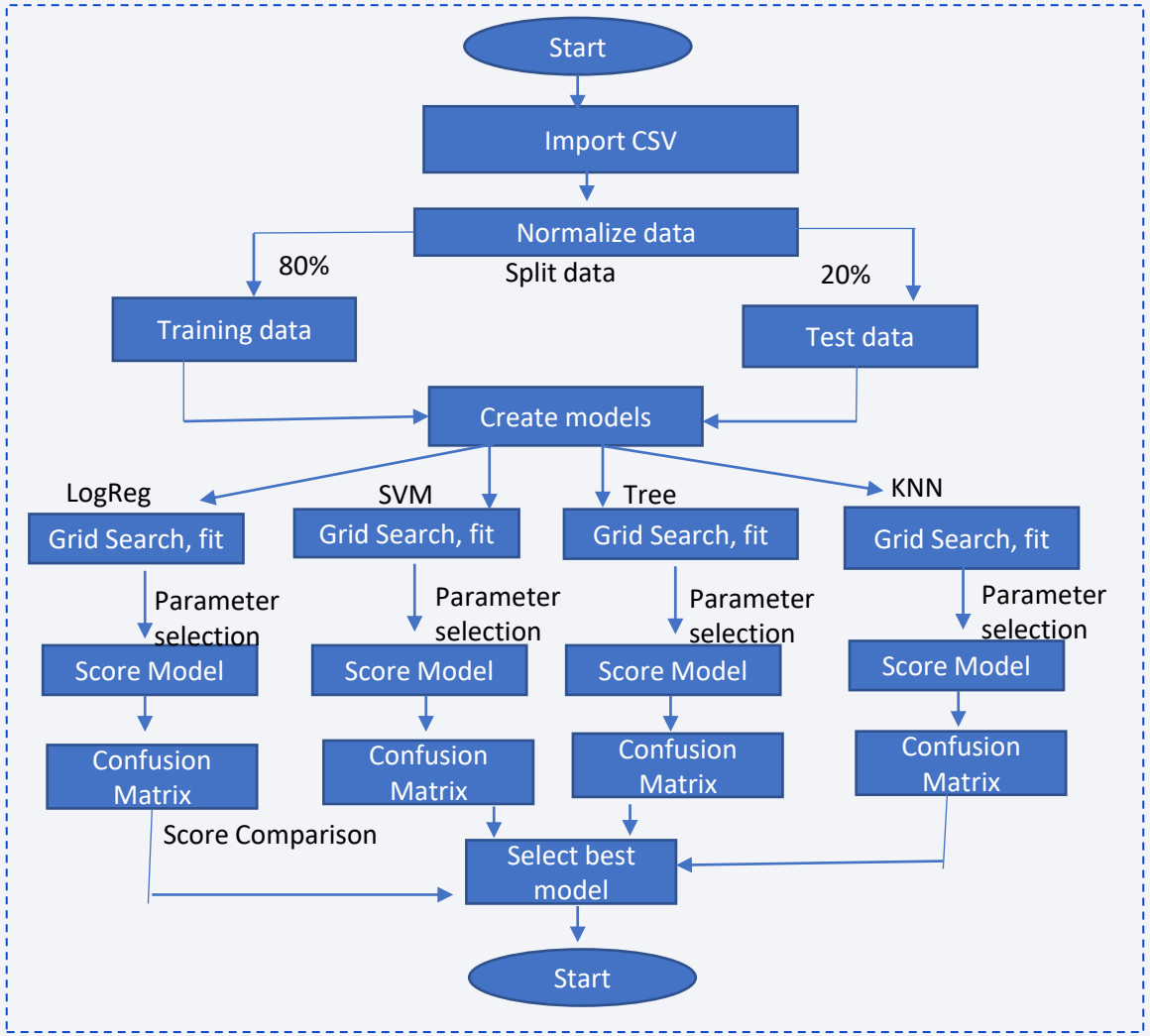


Plotly Dashboard screen shots – Payload Mass



Predictive Analysis (Classification)

- Four models were train and tested on data using the process shown in the flowchart. Comparison was drawn on training accuracy as well as testing data scores.
- https://github.com/tinneminy/DS-CapstoneProject/blob/master/Predictive%20Analysis(Classification).ipynb

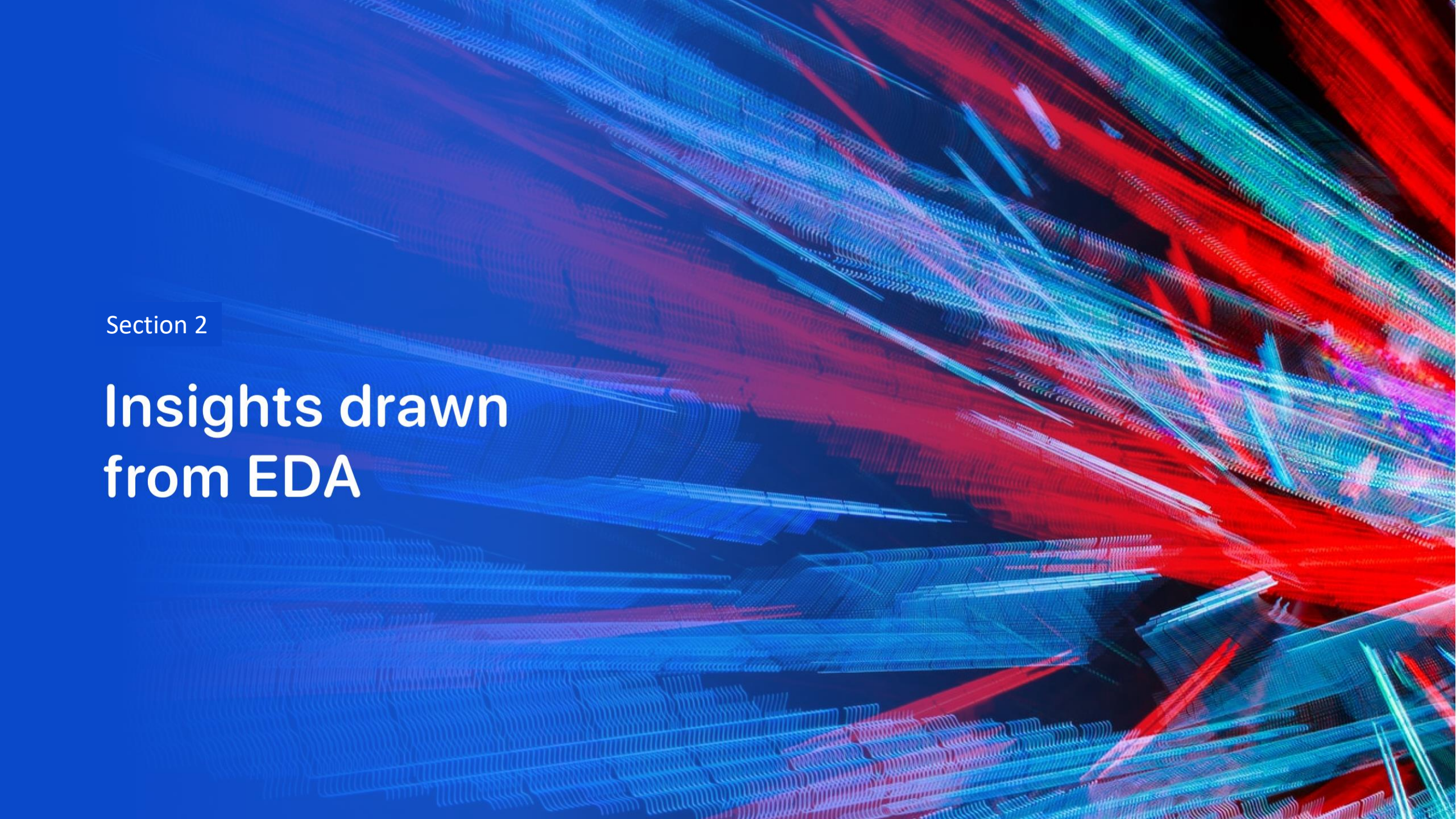


Results – Exploratory Analysis

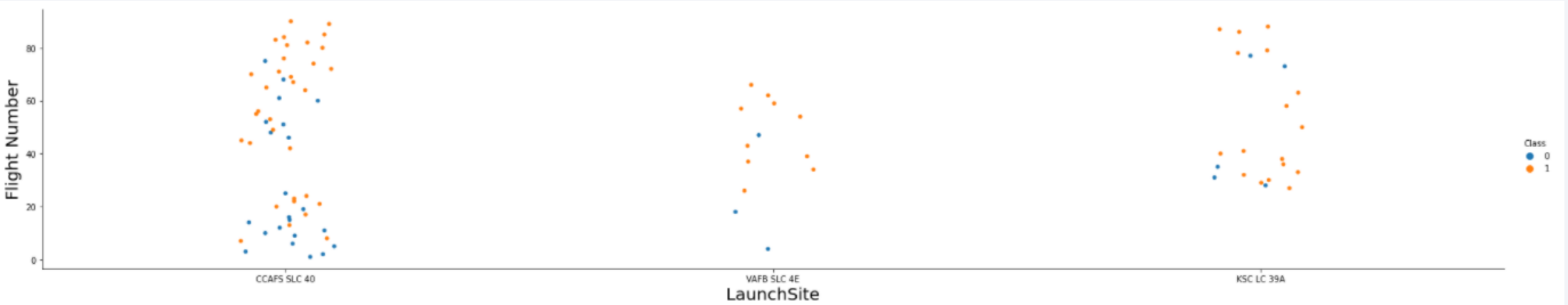
- **SQL**:
  - Majority of loads are to Geosynchronous (GTO) orbit — communications, weather satellites
  - Second most common orbit is Low-Earth orbit (including ISS)
  - Majority of landings are on a drone ship (ASDS)
  - Overall landing success rate is 2/3 (67%)
- **Visualization**:
  - Most initial launches were failures, but the success rate dramatically improved since 2013.
  - Most launches were performed in Florida, with only a handful in California (VAFB).
  - Payload capacity of Falcon 9 has increased, reaching a maximum of 15,400kg starting in 2020.
  - There were two distinct clusters of payload range: ISS the range is 2,000-3,000 kg, and GTO it is higher, at 3,000-7,000 kg.
  - All launch sites are in the proximity to the ocean: this minimizes risk to populated areas in case of rocket failure.
  - All launch sites are also situated in the southern latitudes of the US: The centrifugal force close to the equator provides some assistance with the launch

Results – Predictive Analysis

- Four models were trained and tested on 90 rows of data:
  - LogReg (Logistic Regression)
  - SVM (Support Vector Machine)
  - Tree (Decision Tree)
  - KNN (K-Nearest Neighbour)
- Accuracy on test data for models was 83%.
- Confusion matrix was identical for all models.
- Large training set is needed for further analysis.

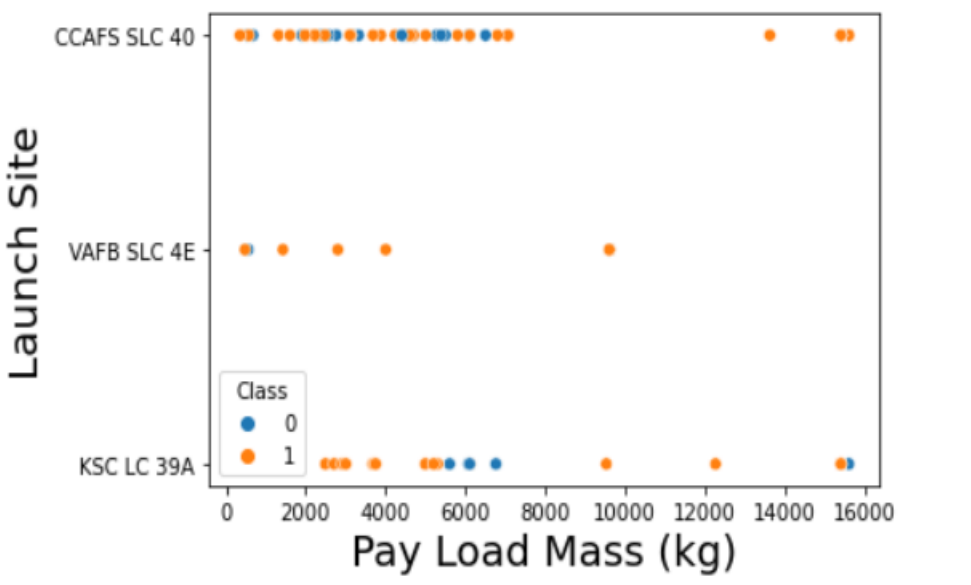

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



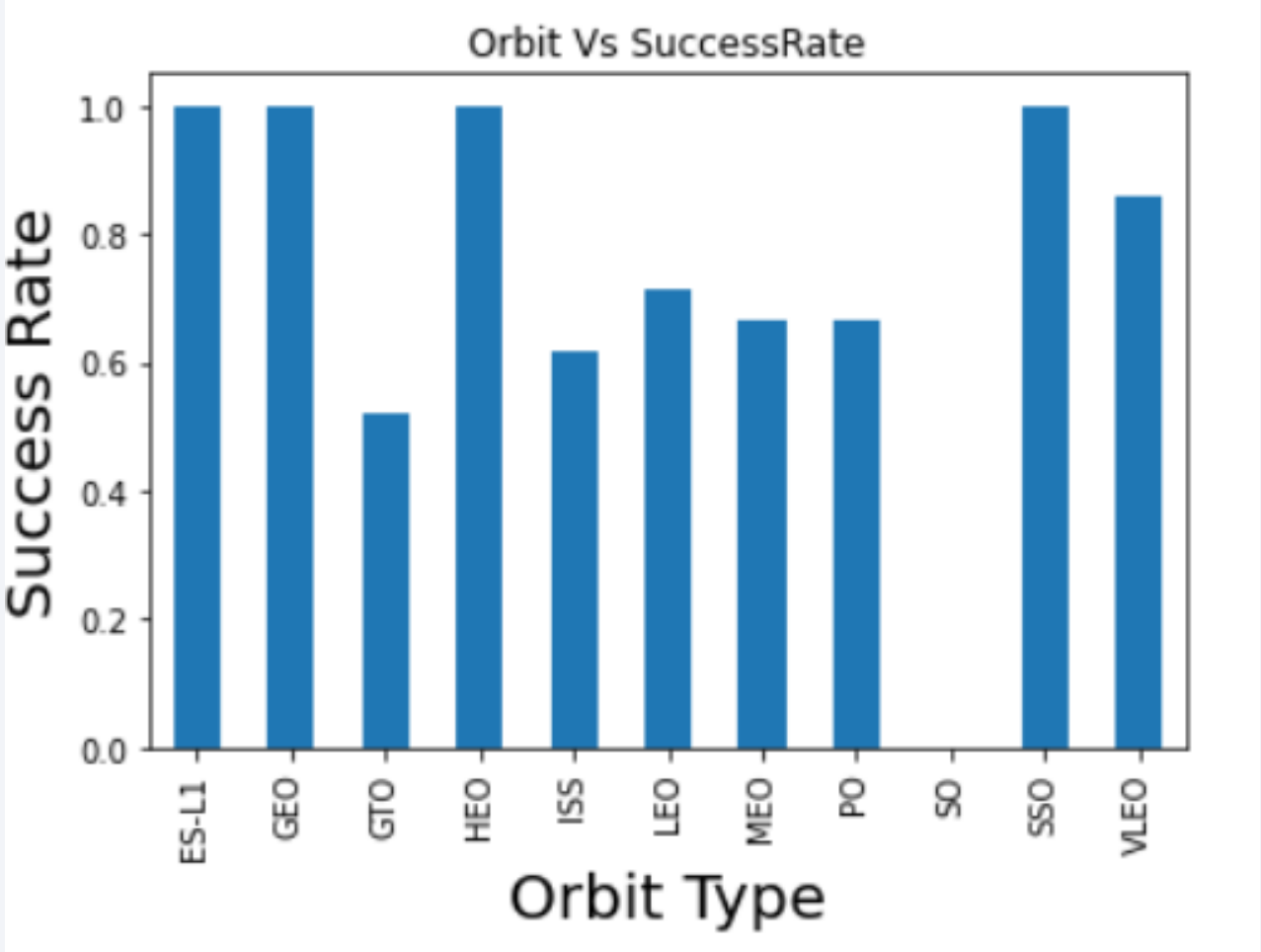
- Flight Number vs. Launch Site
- The greater the numbers of flight, the greater the success rate at a launch site.

Payload vs. Launch Site



- Payload vs. Launch Site
- Higher the payload mass (kg), more successful the launches are.
- There are no heavy payload launches at VAFB SLC-4E launch sites.

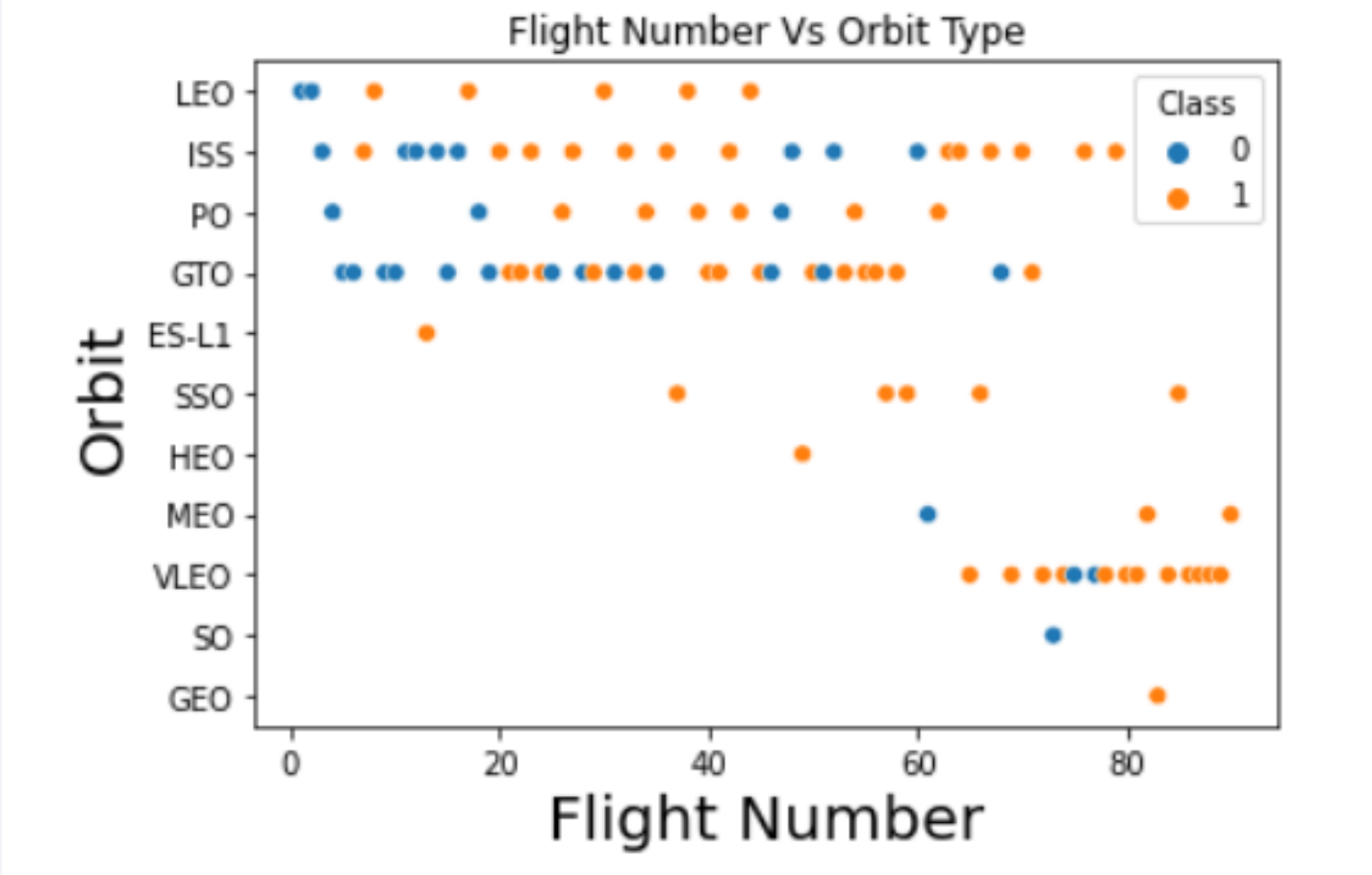
Success Rate vs. Orbit Type



Orbit ES-LI, GEO, HEO, SSO have the highest success rate.

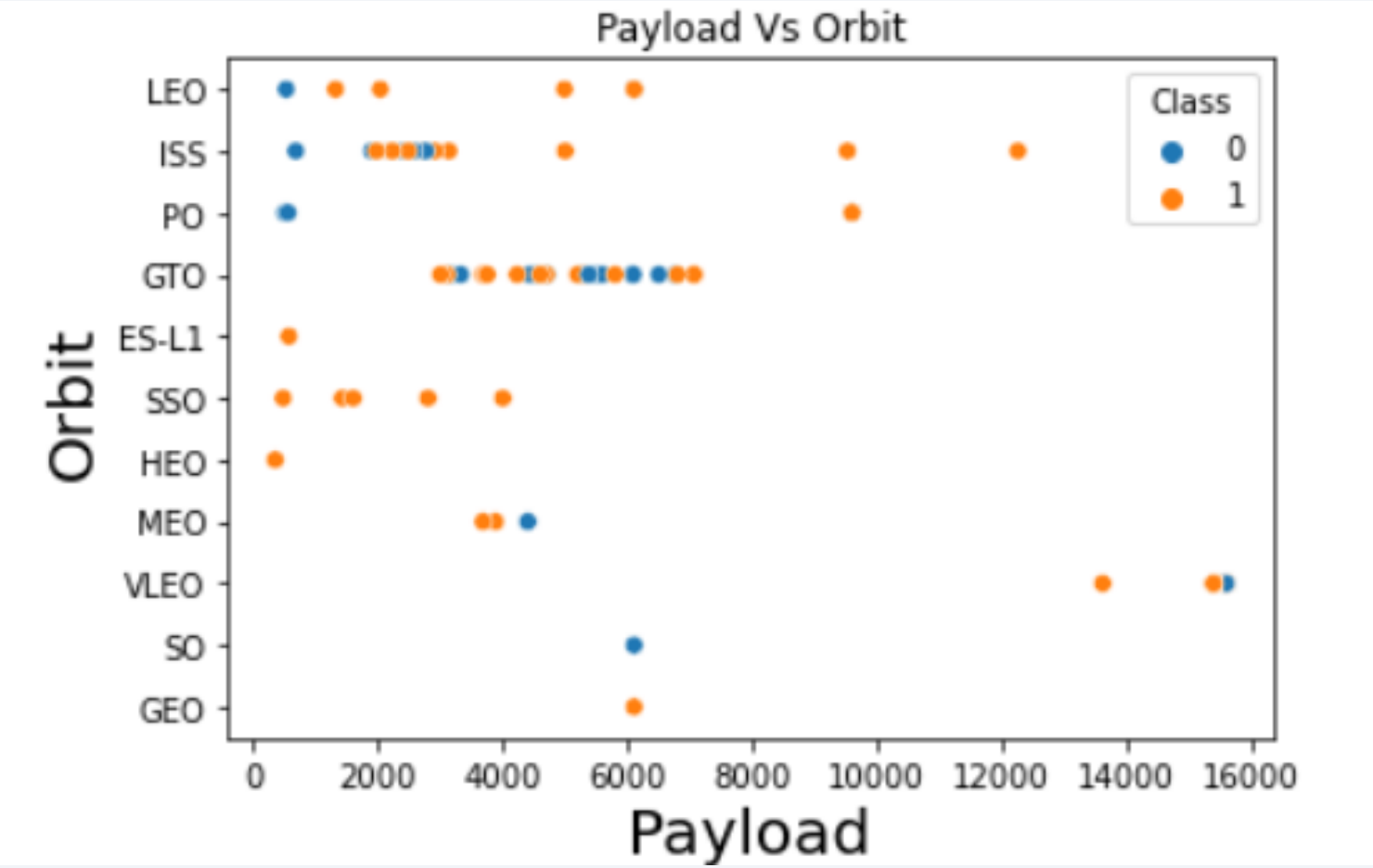
Flight Number vs. Orbit Type

- LEO has greater successful flights.
- ISS and GTO both have low success rates, but more flights were launched.
- ES-LI and HEO both had only one successful flight.
- VLEO orbit has maximum successful flights.



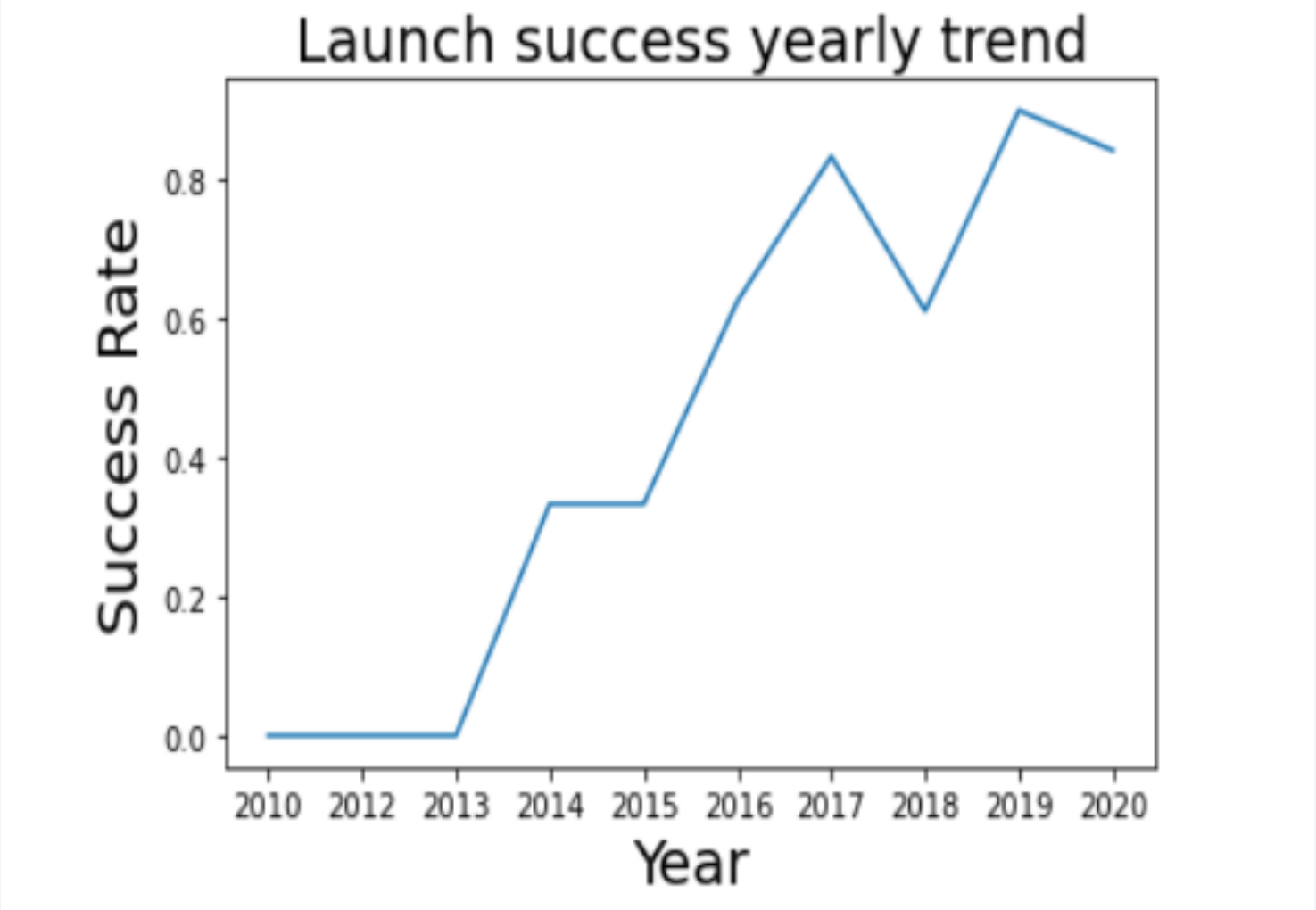
Payload vs. Orbit Type

- Heavy payload are less common in each orbit.
- GTO orbit maintains its payload between 4000-8000 kg range.
- Maximum payload of 16000kg was launched at VLEO orbit with both success and failure rate.



Launch Success Yearly Trend

- The Line graph shows an upward trends in the success rate since 2013.



All Launch Site Names

- SQL query:
  - Select unique(Launch_site) as unq_launchSite from spacextbl;
- Explanation:
  - Using Unique guarantees that only distinct values from the column (Launch_site) will be selected/chosen from spacextbl table.

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

Launch Site Names Begin with 'CCA'

- select * from spacextbl where launch_site like 'CCA%' limit 5;

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

Total Payload Mass

- select sum(payload_mass__kg_) as total_payloadMass from spacextbl where customer = 'NASA (CRS)' ;

total_payloadmass
45596

- Sum() calculates the total payload_mass_kg from table spacextbl where the customer is 'NASA (CRS).

Average Payload Mass by F9 v1.1

- select AVG(PAYLOAD_MASS__KG_) as avg_payloadMass from spacextbl where booster_version = 'F9 v1.1' ;

avg_payloadmass
2928

- AVG() calculates the average payload_mass_kg from table spacextbl where booster_version is 'F9 v1.1'

First Successful Ground Landing Date

- select min(date) as first_successful_Landing_Outcome from spacextbl where Landing__Outcome = 'Success (ground pad)';

first_successful_landing_outcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- select Booster_version from spacextbl where Landing__Outcome = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ between 4000 and 6000);

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

Total Number of Successful and Failure Mission Outcomes

- select count as Success , (select count from spacextbl where Mission_outcome Like '%Failure %') as Failure from spacextbl where Mission_outcome = 'Success';

success	failure
99	1

- 99 successes , 1 failure.

Boosters Carried Maximum Payload

- select Booster_version from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl);

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

- select Date, Booster_version, Launch_site, Landing__outcome from spacextbl where Landing__Outcome ='Failure (drone ship)' and YEAR(Date)=2015;

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

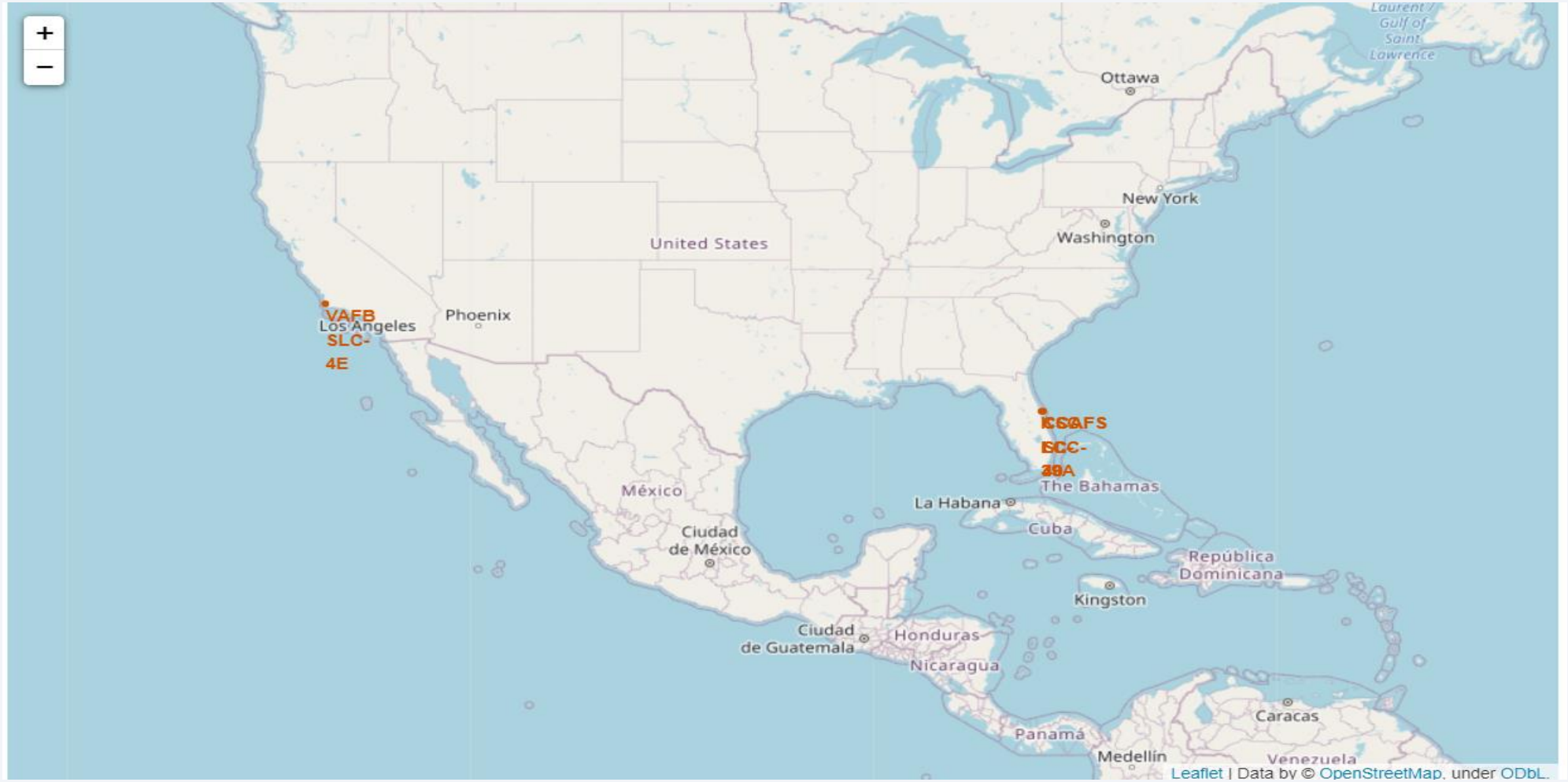
```
%sql select count(Landing__outcome) as count, landing__outcome from spacextbl \
where (date between '2010-06-04' and '2017-03-20') and (Landing__outcome IN ('Failure (drone ship)','Success (ground pad)')) \
group by landing__outcome \
order by count(landing__outcome) desc;
```

COUNT	landing__outcome
5	Failure (drone ship)
3	Success (ground pad)

Section 3

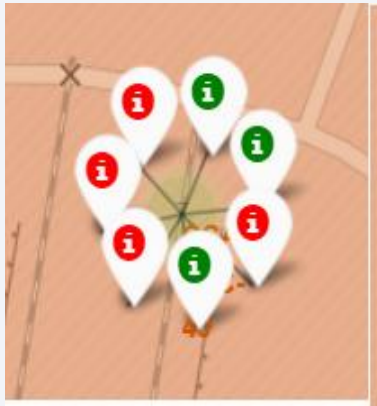
Launch Sites Proximities Analysis

Global SpaceX Launch Sites

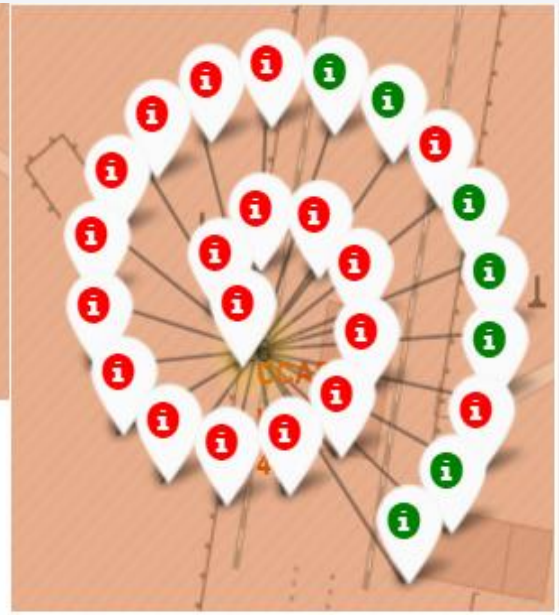


All SpaceX launch sites are in the USA.

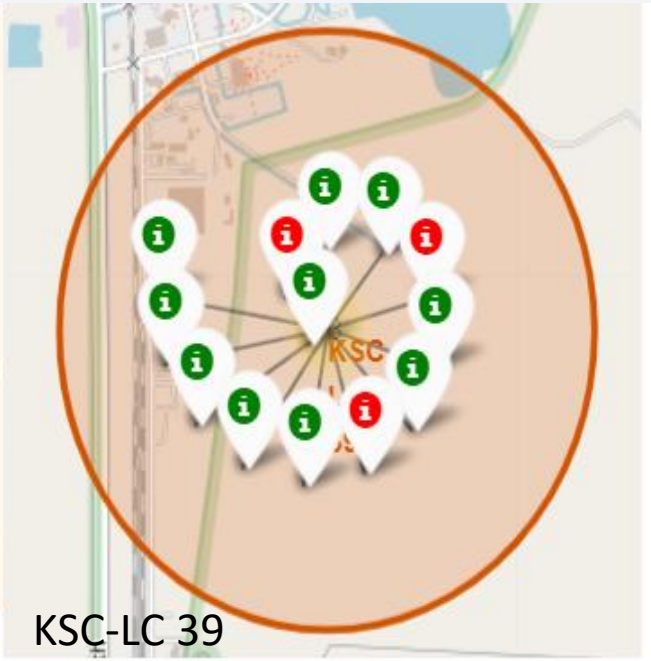
Color Labelled Markers



CCAFS SLC-40



CCAFS LC-40

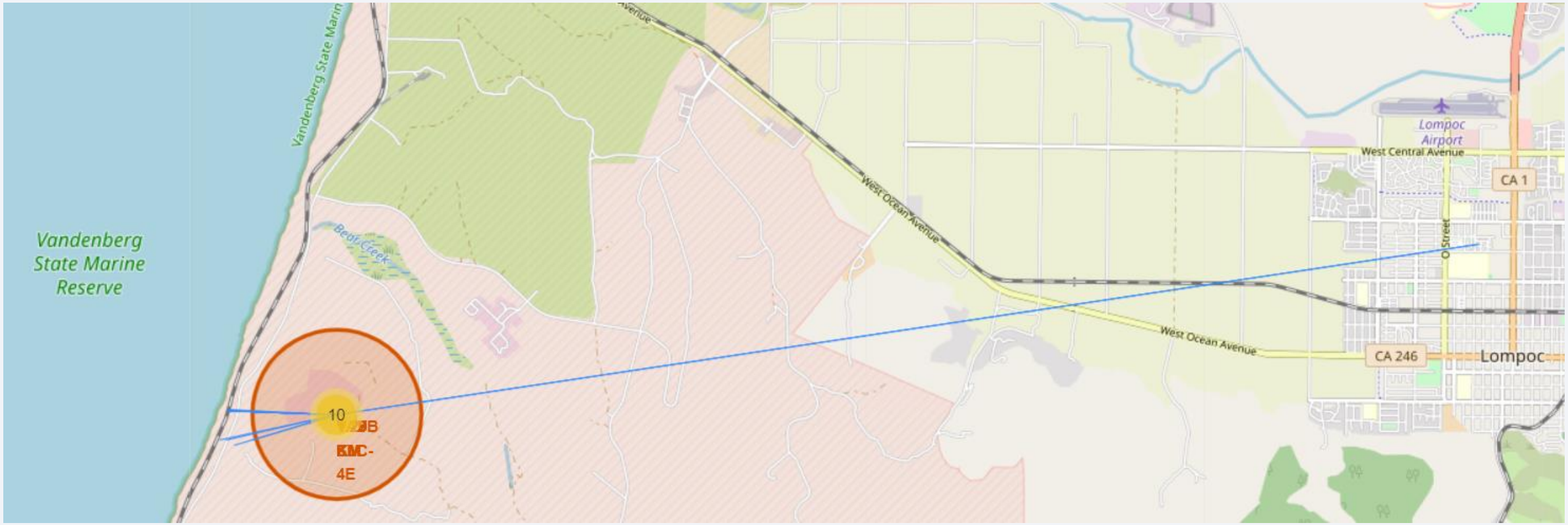


KSC-LC 39

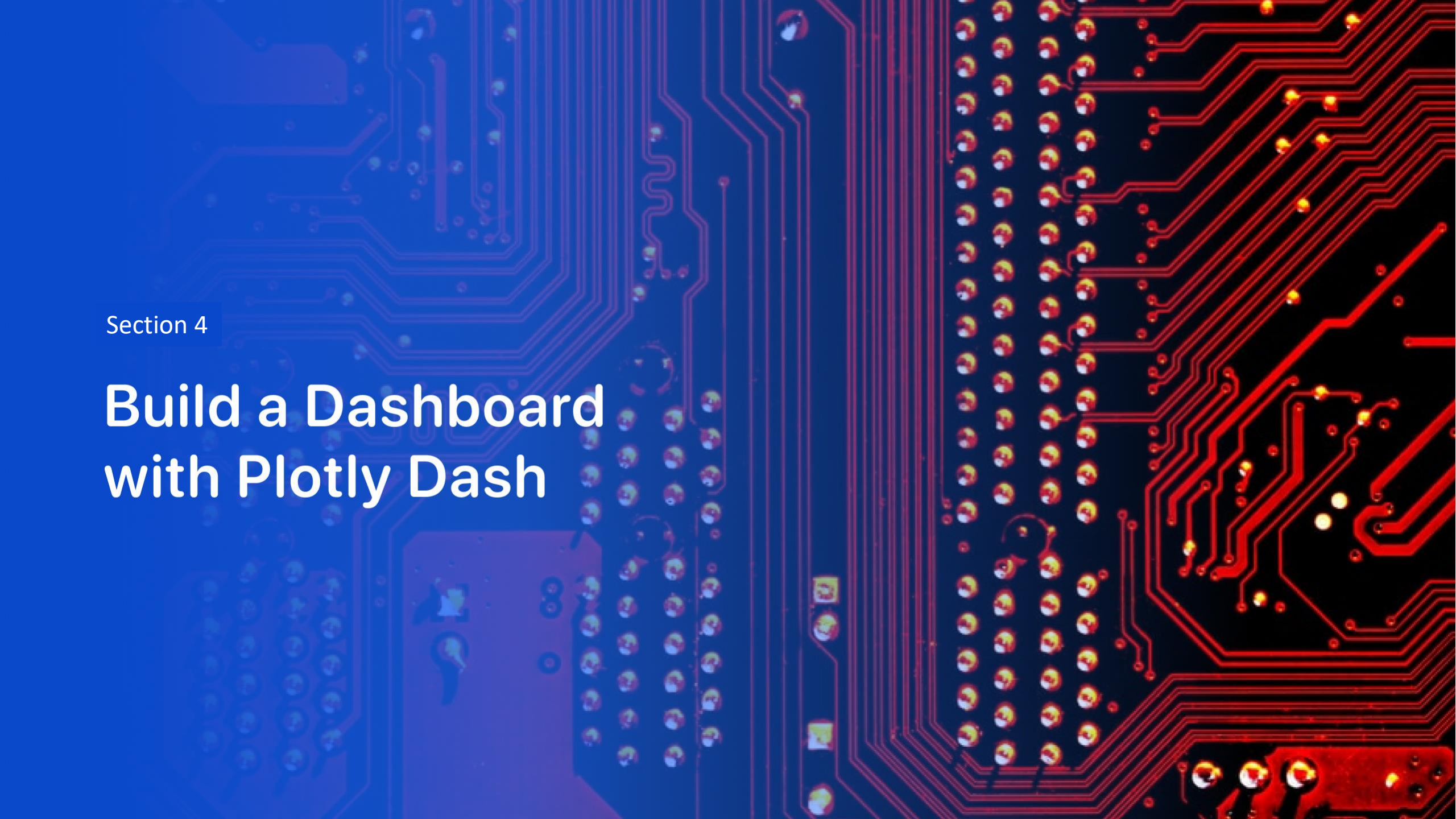
VAFB SLC-4E

- Green Markers depicts success whereas red markers mean failure.
- There are 3 launch sites in Florida and 1 in California.
- KSC-LC 39 have highest successful launches.

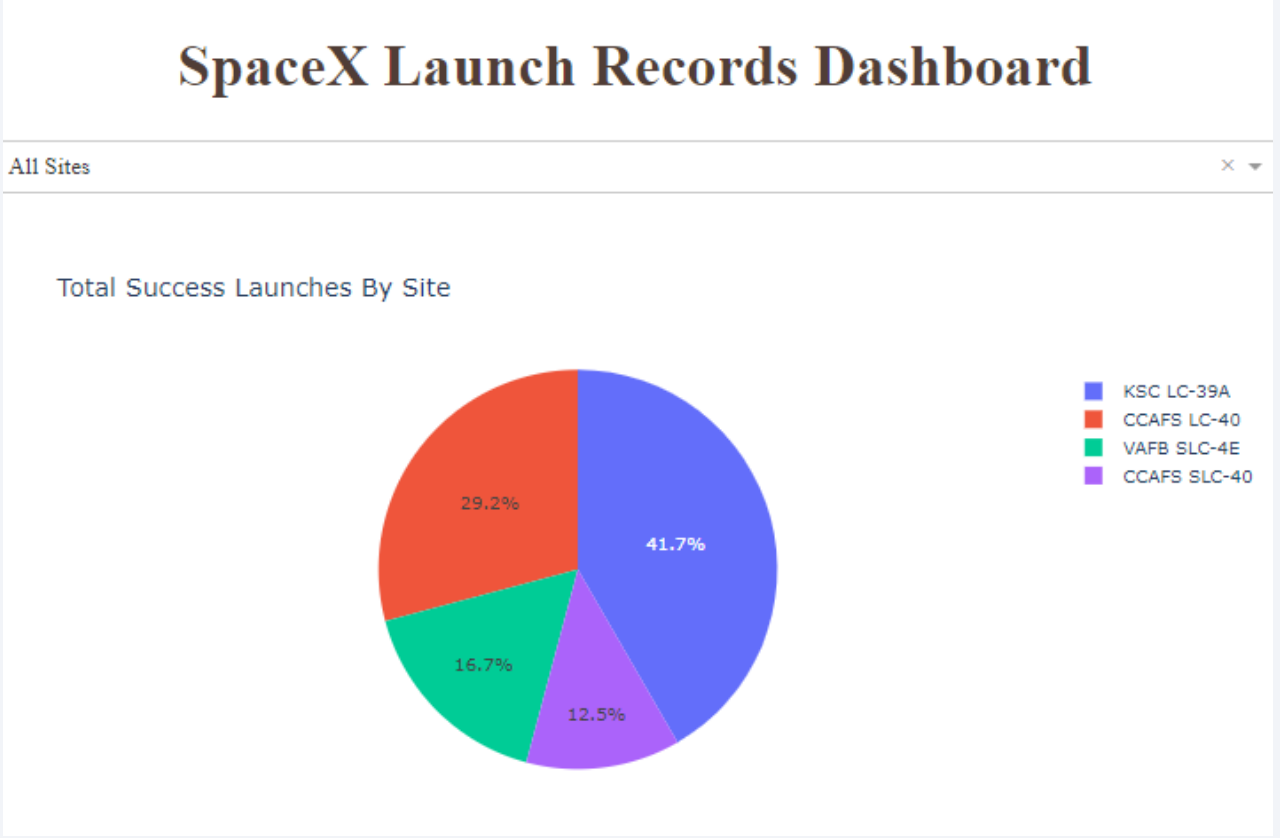
Proximities of launch site VAFB SLC-4E



- The launch sites are in close proximities of railways, highways, coastline but farther away from cities, to avoid any accidents.

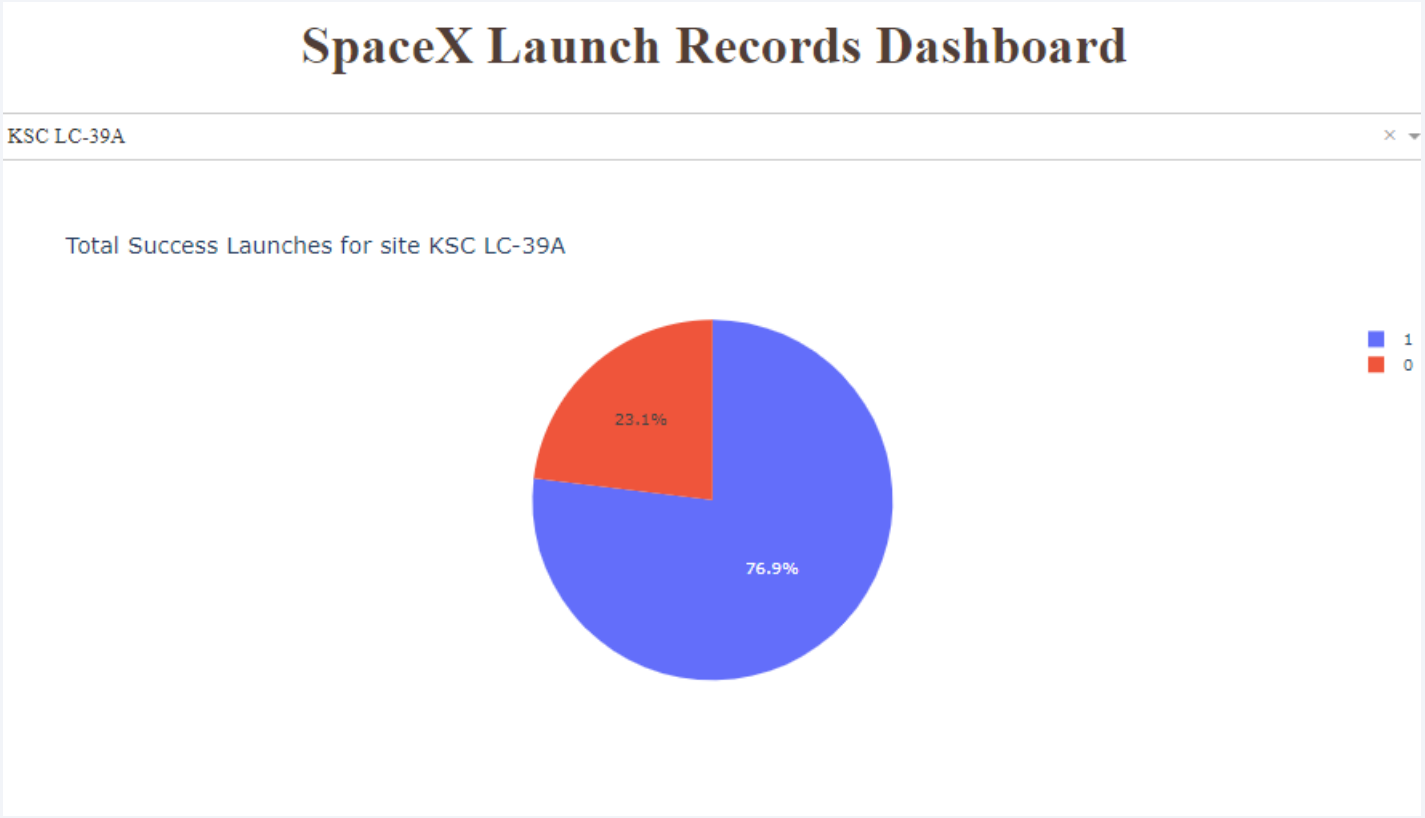

Section 4
Build a Dashboard with Plotly Dash

SpaceX success count for all launch sites



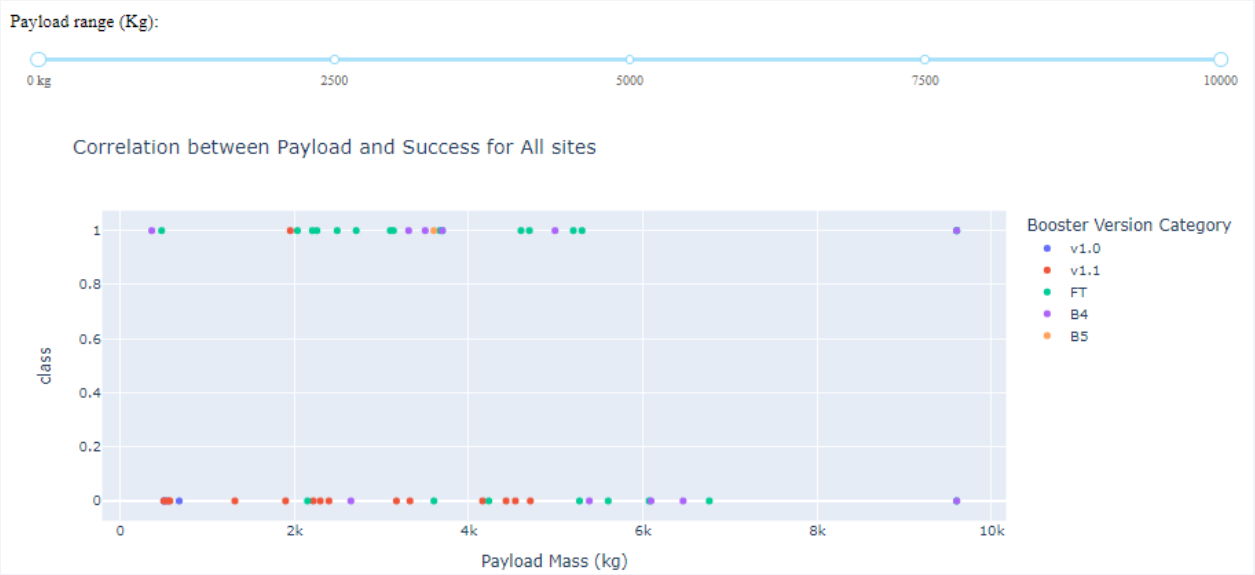
- KSC LC-39A has highest share at 41.7% and CCAFS SLC-40 has lowest share with 12.5%.

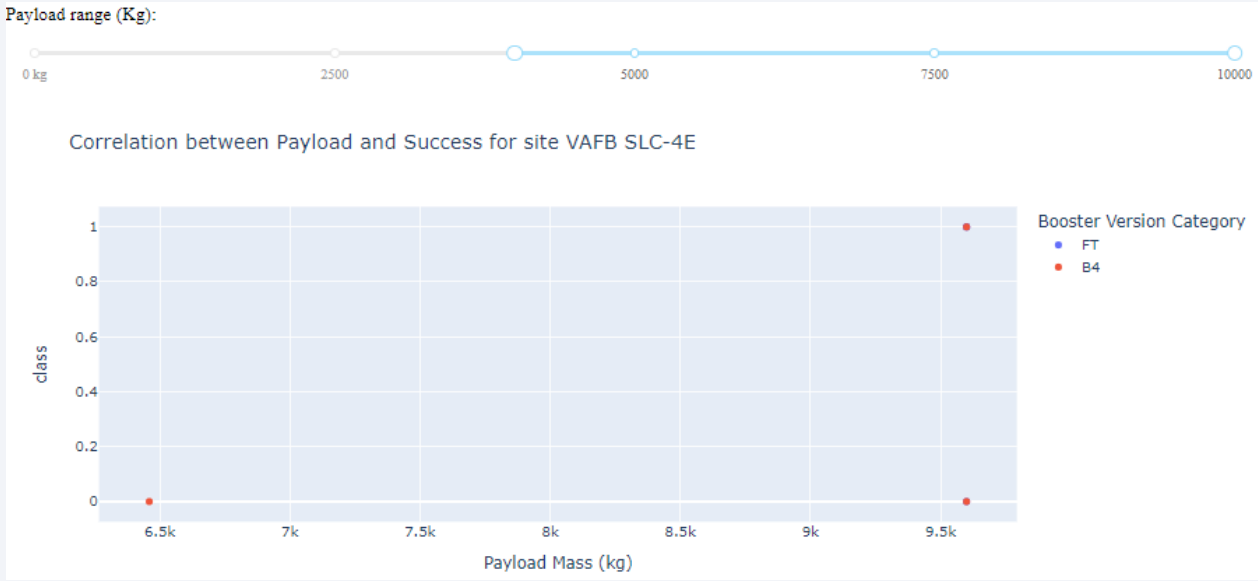
Launch site with highest success ratio



KSC LC-39A has the highest success rate of 76.9% , and only 23.1% failure rate.

Payloads vs outcomes





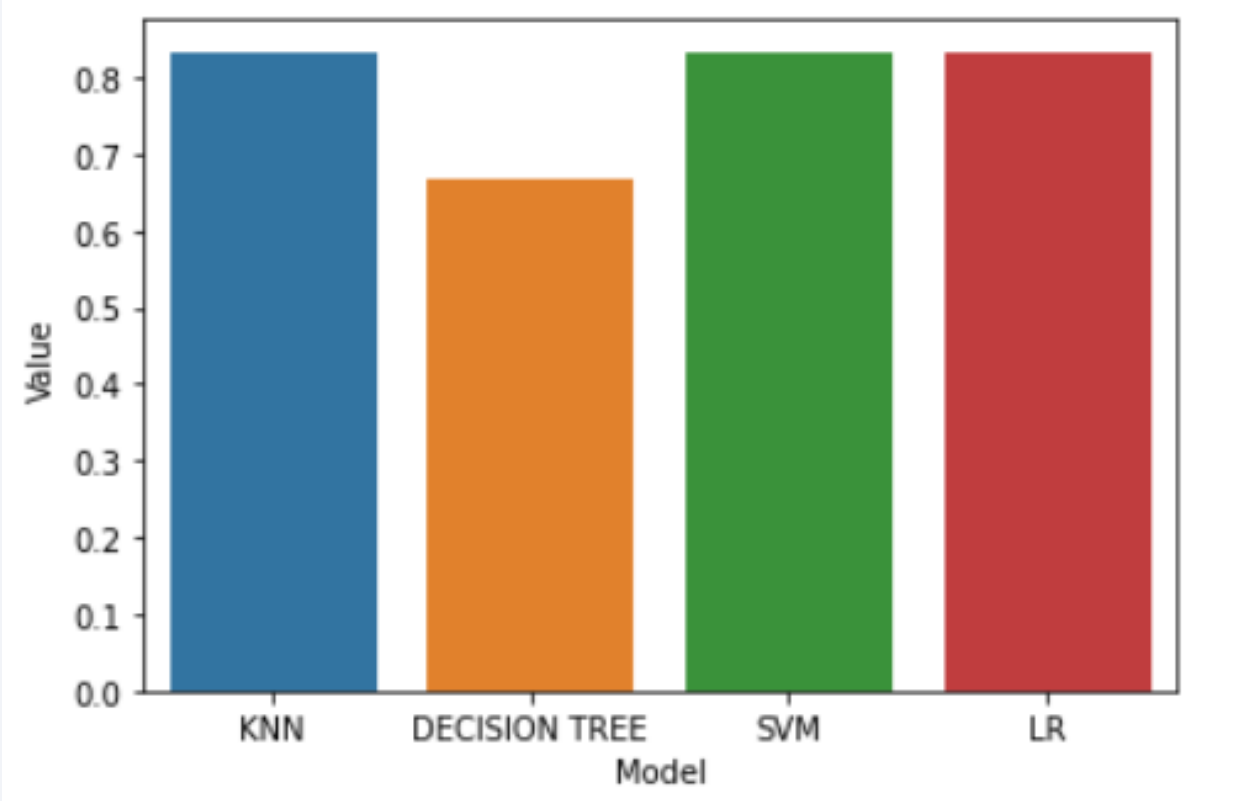
- First screenshots with all payloads
- Second screenshots with heavy payloads only)..
- FT and B4 are also the only ones carrying the highest payloads

Section 5

Predictive Analysis (Classification)

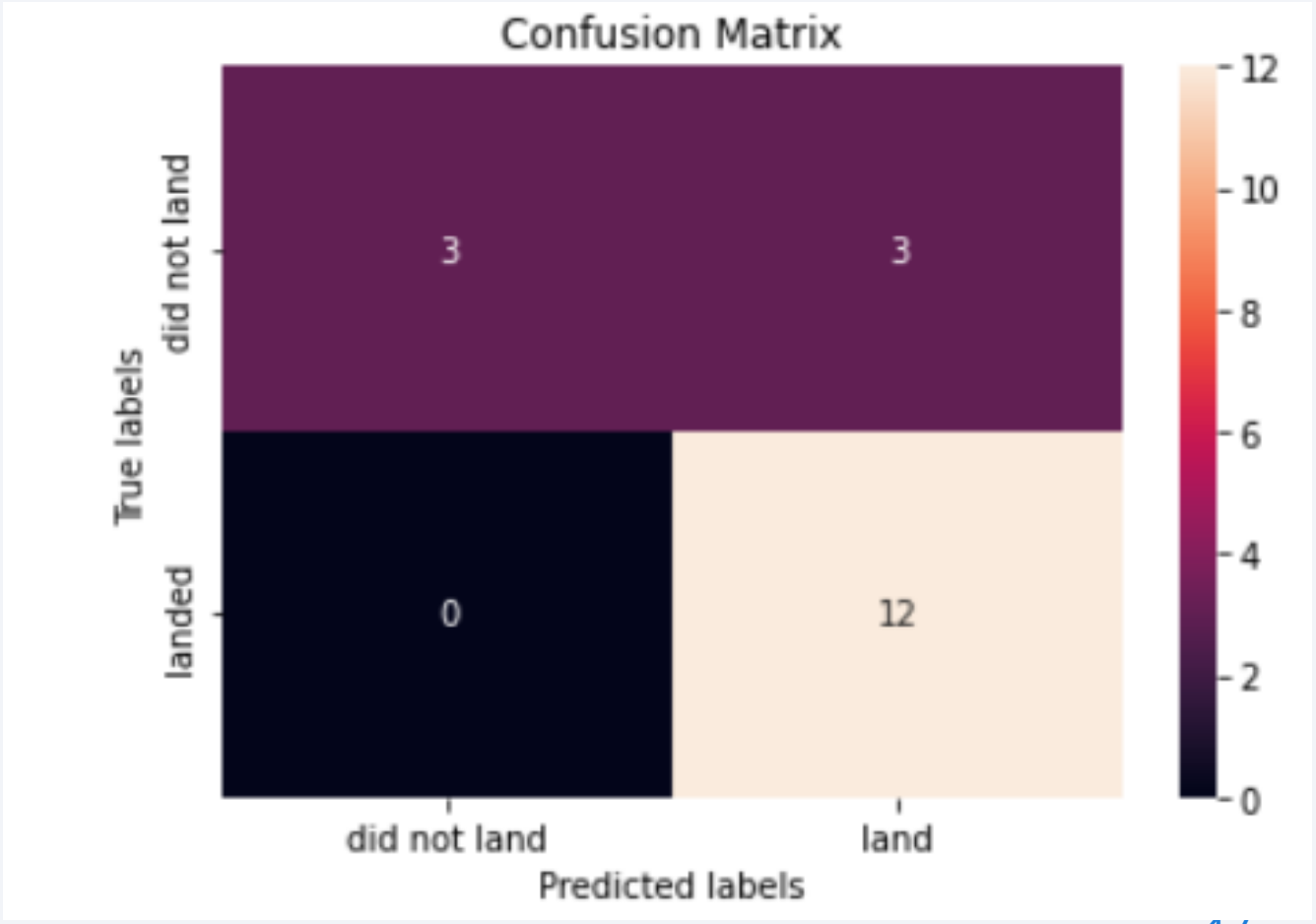
Classification Accuracy

- Model accuracy for all 4 models.
- On test data LR, SVM and KNN all performed similarly, with an R^2 score of 0.83. The Tree model had a lower score of 0.66.



Confusion Matrix

- The Confusion matrix shows 12 true positives , 3 false negatives and 3 negatives.



Conclusions

- 3 models (KNN, LR, SVM) have the same accuracy on test data, it is difficult to say which is best.
- Training Vs. Test data accuracy are not the same.
- Data set is too small (only 90 observations/columns) to distinguish any further between models.
- Success rates for SpaceX launches increased throughout the years
- KSC SLC-39 launch site with the most successful launches.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!