Winning Space Race with Data Science

SIYAN JOHNY MUNDACKAL

24th JANUARY 2022

OUTLINE

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

Executive Summary

Summary of methodologies

Data collected from SpaceX API and SpaceX Wikipedia page via API and webscraping

Data wrangling

Exploratory analysis with SQL, Visualizations, folium maps and dashboards.

Used GridSearchCV to find the best parameters for machine learning models

Summarized accuracy scores of all four machine learning models

Summary of results

Performed data analysis

Machine Learning predictions.

Introduction

Project background and context

The project is centered on the space industry, and it analyses data from SpaceX's website to see if the Falcon 9 first stage will successfully land. The cost of a Falcon 9 launch, according to Space X, is $62 million dollars, compared to $165 million dollars for its competitors. In this project, Space Y wants to compete with Space X, therefore it employs machine learning algorithms to determine if the Falcon 9 rocket launch first stage will land successfully. This information will aid Space Y's campaign for rocket launch contracts versus Space X.

Insights to be drawn

- A study of Falcon 9 landing success rates is being explored
- To do so, a relationship between specified first-stage parameters is determined to show how each parameter has an impact on the percentage of people that succeed.
- As a result, the lowest cost of a rocket launch may be established.

Section 1

Methodology

Methodology

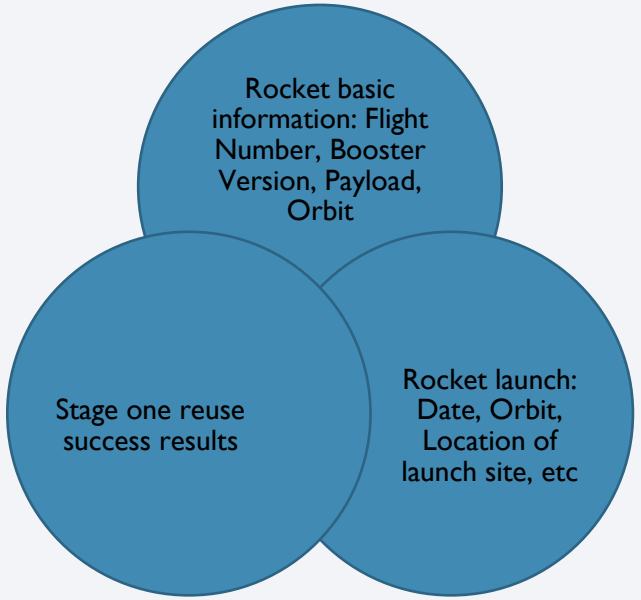
Methodology

- Data collection methodology: Web-scraping via BeautifulSoup & REST API from Falcon 9 were extracted and analyzed from Wikipedia data source
- Perform data wrangling:
- Data were transformed into suitable formats, i.e. one-hot-encoding method , statistical method such as standardize of data were complied to get analytic results
- Perform exploratory data analysis (EDA) using visualization and SQL:
- Analysis was done using SQL and visualization to determine relationship/correlation between variables.
- Perform interactive visual analytics using Folium and Plotly Dash:
- Explore launch data further by using folium maps and dashboard reporting.
- Perform predictive analysis using classification models:
- Performed analytics on Logistic Regression, Classification tree, and SVM and find accuracy of models by verification on test data

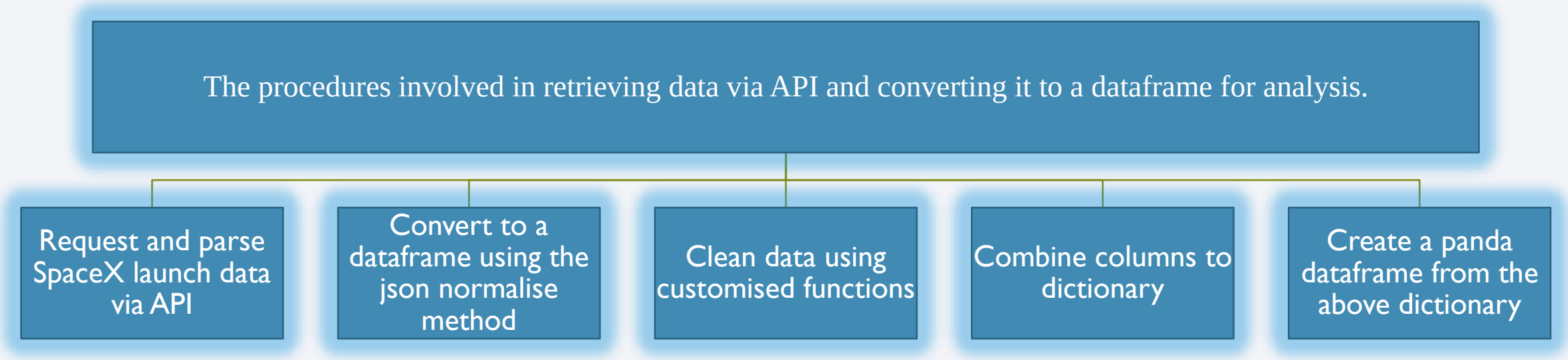
Data Collection

Data was collected from Wikipedia through web scraping (BeautifulSoup) and REST API to predict success/failure of Falcon 9 First Stage landing.

Information from data includes:

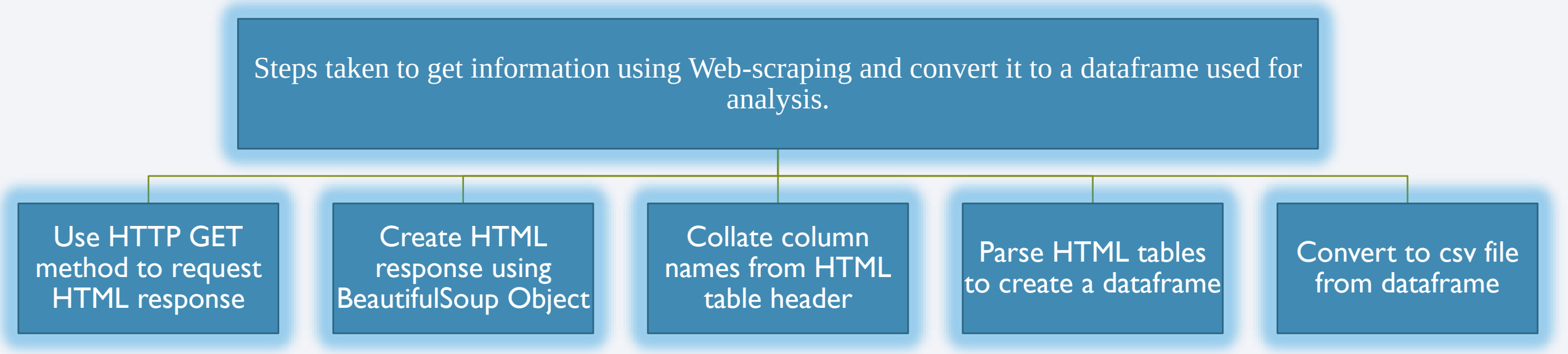


Data Collection – SpaceX API



https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/ba7b655977607687d317e6ae65cc4f3ce5179a51/Capstone%20project%20data%20collection%20API.ipynb

Data Collection - Scraping



https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone%20project%20web%20scrapping.ipynb

Data Wrangling

The data wrangling process aids in the preparation of data for analysis and the extraction of insights from exploratory analysis in order to choose the appropriate labels for training supervised models.This stage groups all successful and unsuccessful launch results under the heading 'Class,' allowing us to get insight into launch site success rates.

1. CALCULATE NUMBER OF LAUNCHES ON EACH SITE
2. CALCULATE THE NUMBER AND OCCURRENCE OF EACH ORBIT
3. CALCULATE THE NUMBER AND OCCURRENCE OF MISSION OUTCOME PER ORBIT TYPE
4. CREATE LANDING OUTCOME LABEL FROM OUTCOME COLUMN
5. DETERMINE SUCCESS RATES

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone%20project%20data%20wrangling.ipynb

EDA with Data Visualization

The insights regarding the Launch data received from the website are explored via visualizations.

Scatter plot: are used to demonstrate the impact of one variable on another. It is appropriate for huge datasets.

Flight number vs. Payload mass
Flight number vs. Launch site
Payload vs. Launch site
Orbit vs. Flight number
Payload mass vs. Orbit type
Orbit vs. Payload mass

Bar chart: Easy to compare values by category or continuous dependent variable

Mean vs. Orbit

Line graph: Visualize trend data and helps in making predictions

Success rate by year

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone%20project%20EDA%20with%20python.%20viz.ipynb

EDA with SQL

- Display 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 achieved
- 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

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/c1acbbf6f924a2dbf8650beb84a475e0cc33bf04/Capstone%20project%20EDA%20with%20SQL%20(2).ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- Folium has been used to create an interactive map for Launch Site analysis.
- A circle marker was built using the dataset's latitude and longitude to graphically show all of the launch points.
- To highlight launch success and failures on the map, a green and red colour marker was constructed.
- Several distance computations were performed against launch locations and other markers such as railways, coastlines, highways, and cities to see if their closeness to launch sites was a decisive factor in their success or failure.

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone%20project%20folium.ipynb

Build a Dashboard with Plotly Dash

Plotly Dash was used to construct an interactive dashboard for analyzing and visualizing data from the launch.

PIE CHART : The pie chart was made to compare total launch success rates to launch locations

SCATTER PLOT: shows if there is a link between 'Payload Mass' and launch success by 'Booster Version Category.' Scatter plots are important because they allow for the comparison of a large number of data points while also indicating whether factors are positively or negatively connected.

Pie chart
Total launches for each site
Relative of multiple classes of data
Quantity shown as size of each circle

Scatter plot: Outcome vs Payload mass by booster version

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone%20project%20dash.ipynb

PREDICTIVE ANALYSIS (CLASSIFICATION)

Model build

Load data frame

Standardize data

Creating training/test datasets

Set up parameters

Use GridSearchCV function to loop through predefined parameters

Evaluation

Check model score

Analyse model using Confusion Matrix

Improvement

Tuning related parameters

Best Fit Model

Assess best Model

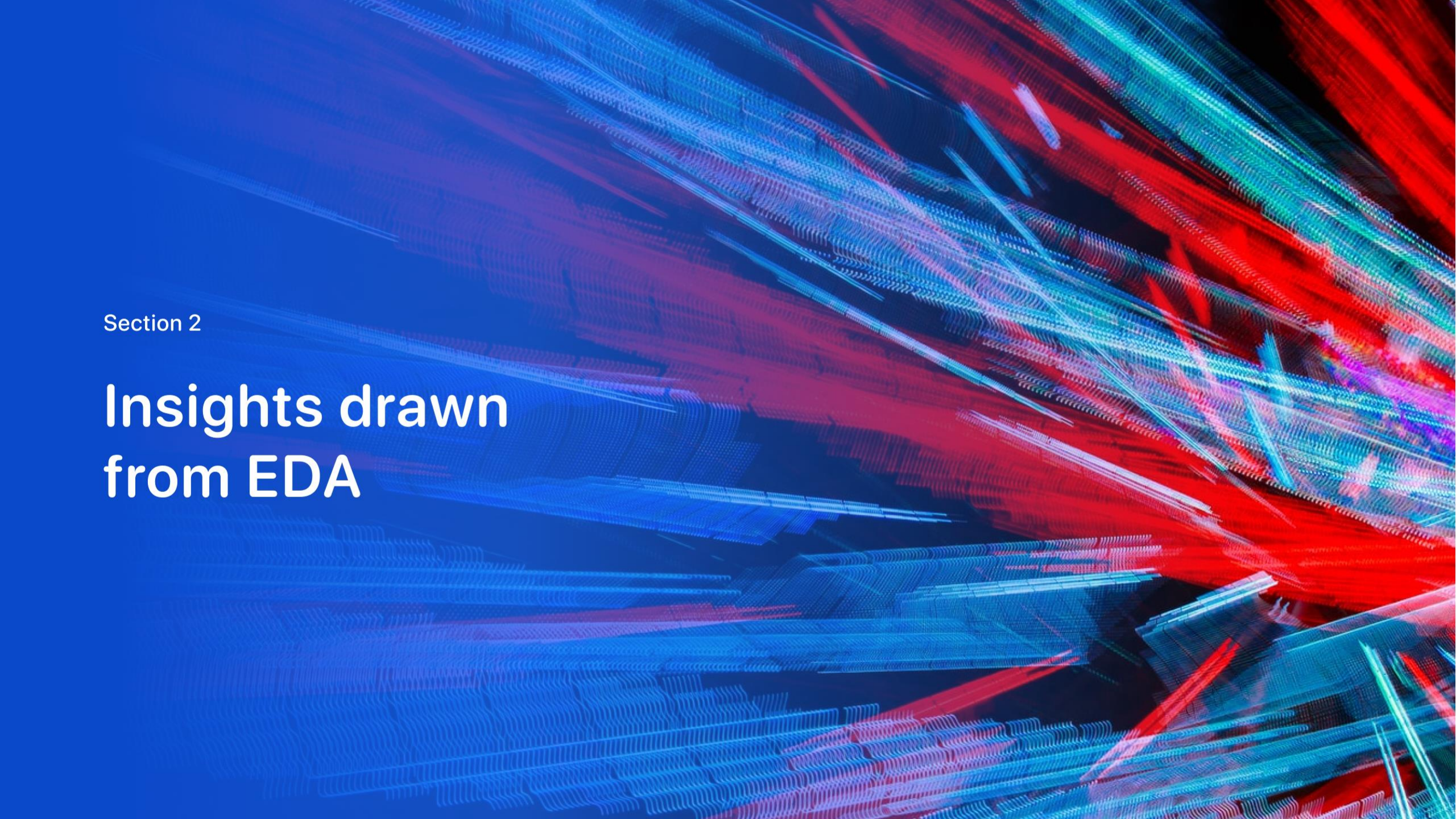
https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project/blob/812b6bf849a7a29aa3d5f9685207720ded4617ef/Capstone_Machine%20Learning%20Prediction.ipynb

Results

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

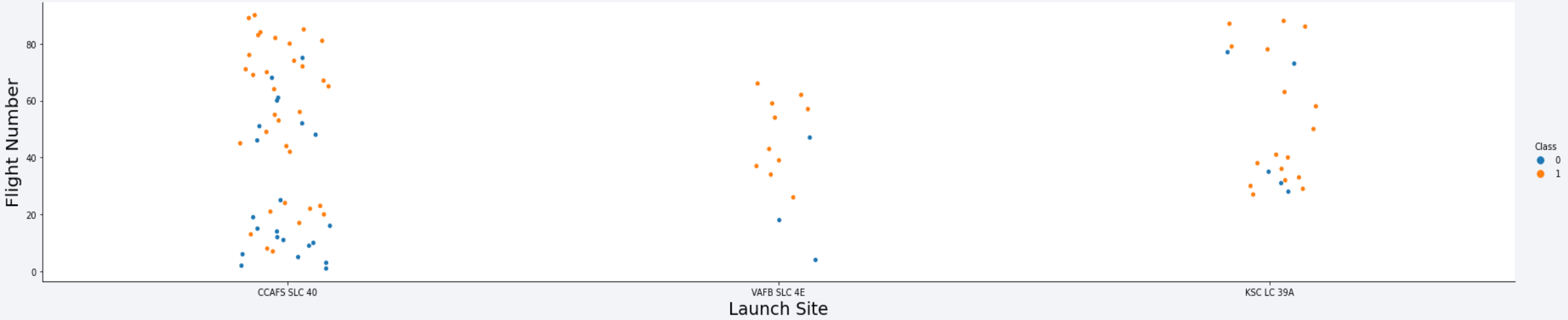
Section 2

Insights drawn from EDA



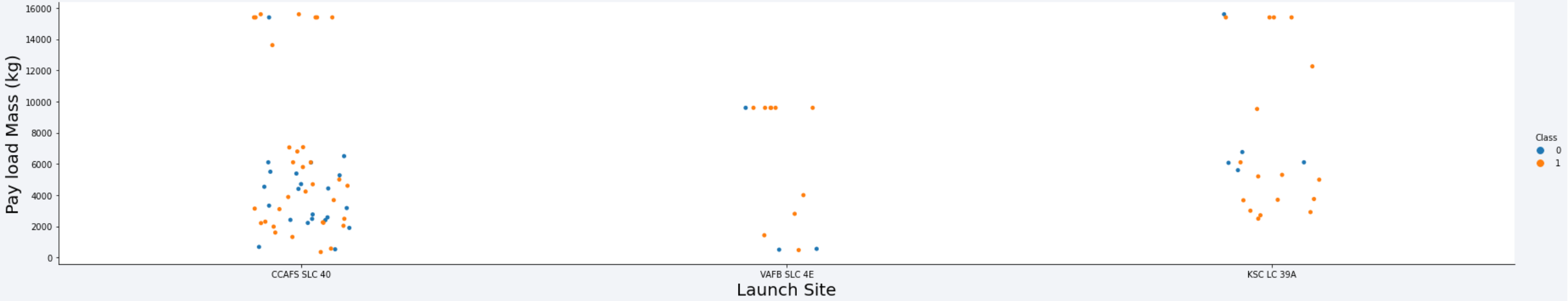
Flight Number vs. Launch Site

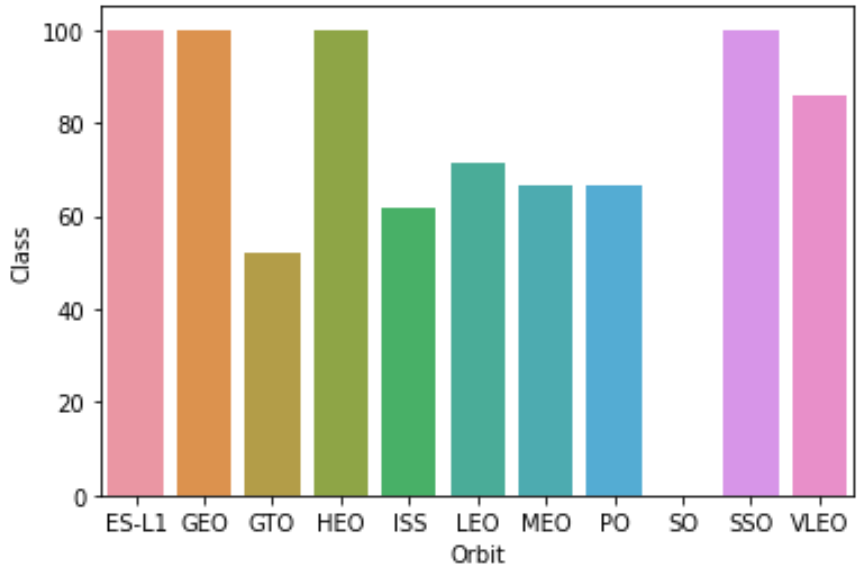
- The scatter plot indicates that the number of flights is connected to the success of the launch. The success rate of the launch location grows as the number of flights increases.



Payload vs. Launch Site

- Higher payload shows the most success rate compared to lower ones.
- Payload Vs. Launch Site scatter point chart you will find for the VAFB SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).





SUCCESS RATE VS. ORBIT TYPE

ES-L1(1),GEO(1), HEO(1)have100% success rate(sample sizes in parenthesis) SSO(5) has 100% success rate

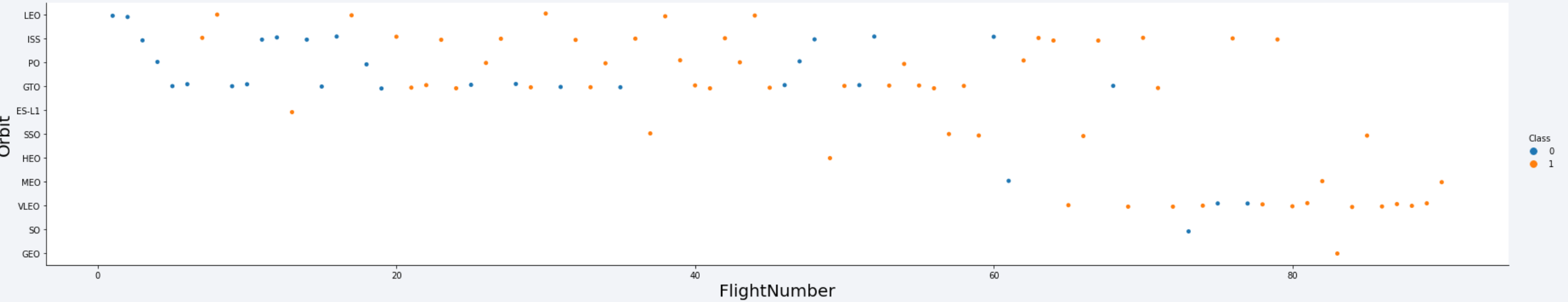
VLEO(14)has decent success rate and attempts

SO(1) has 0% success rate

GTO(27) has the around 50% success rate but largest sample

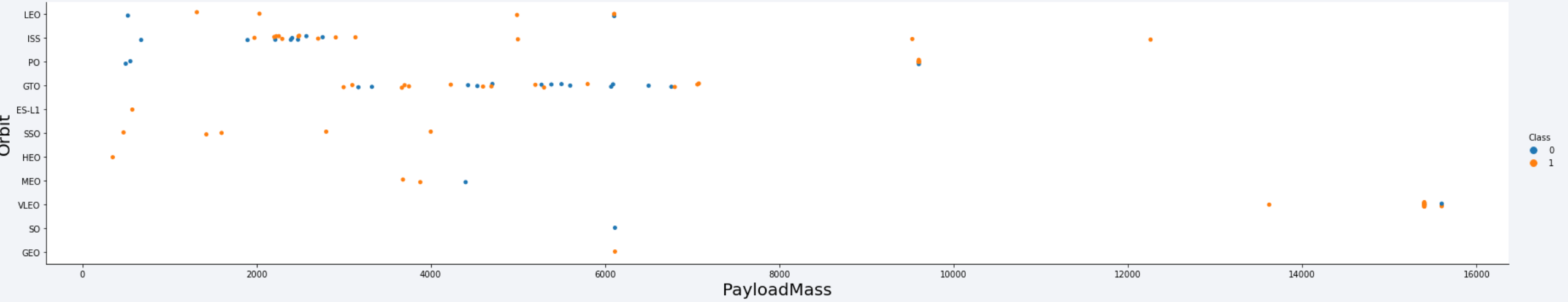
Flight Number vs. Orbit Type

- The frequency of flights is correlated with launch success. As volumes of flight increases so does the likelihood of successful launches.
- There appears to be no relationship between GTO orbit and flight number.



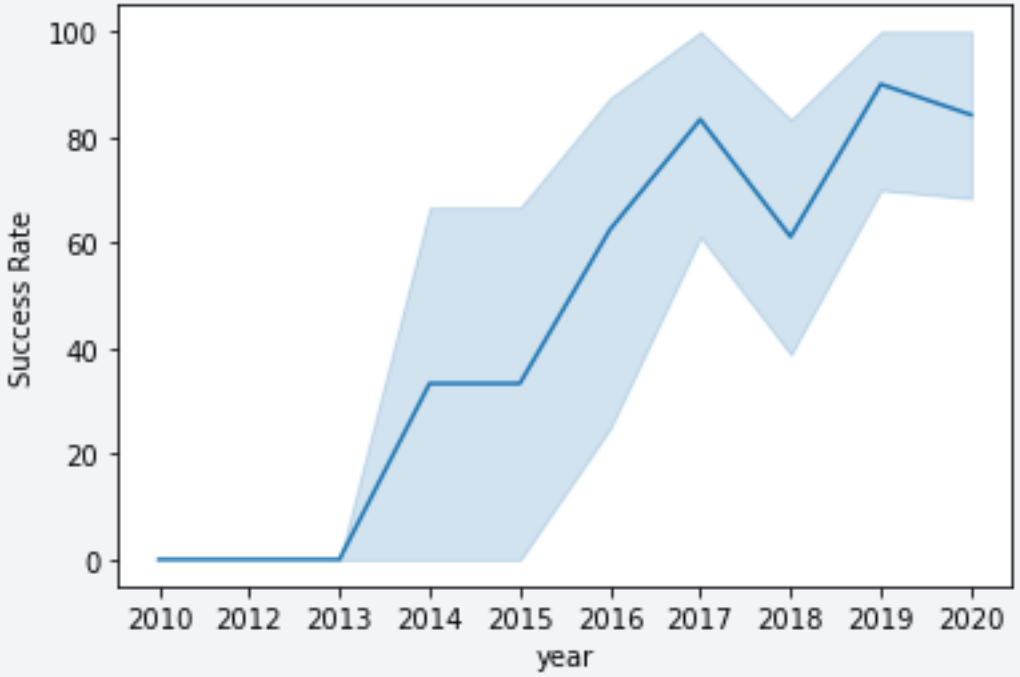
Payload vs. Orbit Type

- 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.



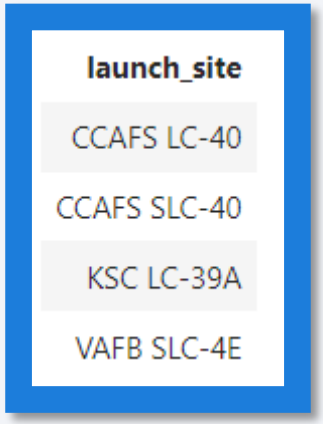
Launch Success Yearly Trend

- We can see as years went by, the higher success rate it would get by average.



All Launch Site Names

Apply “Unique ” command to get distinct list from column “ launch_site ” in table SPACEXTBL2



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

The query returns only unique (‘distinct’) values from the column Launch

Launch Site Names Begin with 'CCA'

First five entries in database with Launch Site name beginning with CCA.

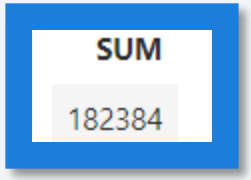
DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Asterics

(*) returns all values from the table and 'Like' is an SQL operator that looks for specified pattern in a column. 'Limit 5' retrieves the top 5 rows of the

Total Payload Mass

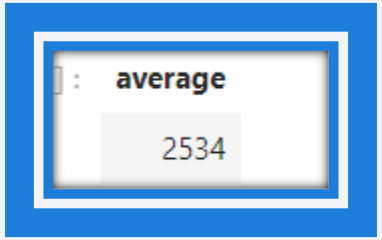
- The ‘sum’ function returns the ‘sum’ of all payload mass kg that falls under customer ‘NASA (CRS)’



Average Payload Mass by F9 v1.1

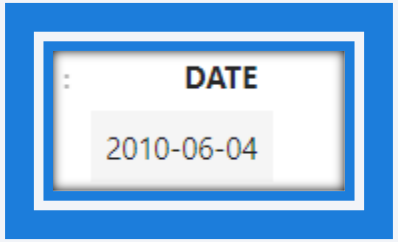
This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 v1.1 is on the low end of our payload mass range



First Successful Ground Landing Date

- ‘Min’ function returns the earliest date of the successful ground pad launch. The ‘where’ selects the specified landing outcome and ‘Like’ operator matches the string and retrieves the



SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 non inclusively.

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

mission_outcome	COUNT
Failure (in flight)	4
Success	396
Success (payload status unclear)	4

- The query calculates the total number of failure and success as defined in mission_outcome. The 'count' function and 'group by' aggregates the values in the mission_outcome and 'order by' organizes it alphabetically. The default is always ascending.

BOOSTERS CARRIED MAXIMUM PAYLOAD

Selects Booster version, payload and payload mass kg, sub query selects max payload mass, order by organizes the payload in ascending order.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 LAUNCH RECORDS

```
%%sql
select landing__outcome, booster_version, launch_site from SPACEXTBL
where landing__outcome like 'Failure%'
and year(DATE) = 2015
```

```
 * ibm_db_sa://szt36742:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0
Done.
```

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

- The ‘where’ selects all failed drone ship launches and year (date) function
- converts the date into year.

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

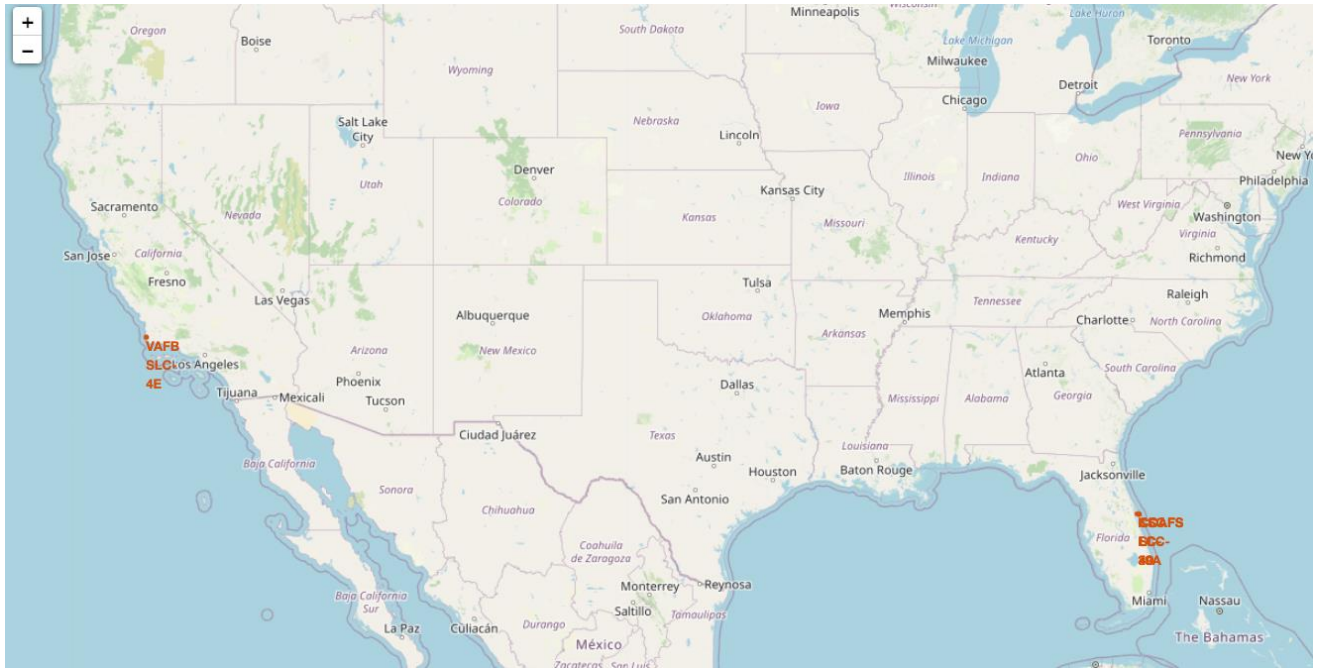
Count function aggregates landing outcome for the selected ones specified in the 'where' condition. 'Date between' ensures landing outcome is only aggregated for those that meet the time frame criteria. The 'and' allows multiple condition to be specified.

landing__outcome	COUNT
No attempt	40
Failure (drone ship)	20
Success (drone ship)	20
Controlled (ocean)	12
Success (ground pad)	12
Failure (parachute)	8
Uncontrolled (ocean)	8
Precluded (drone ship)	4

Section 4

Launch Sites Proximities Analysis

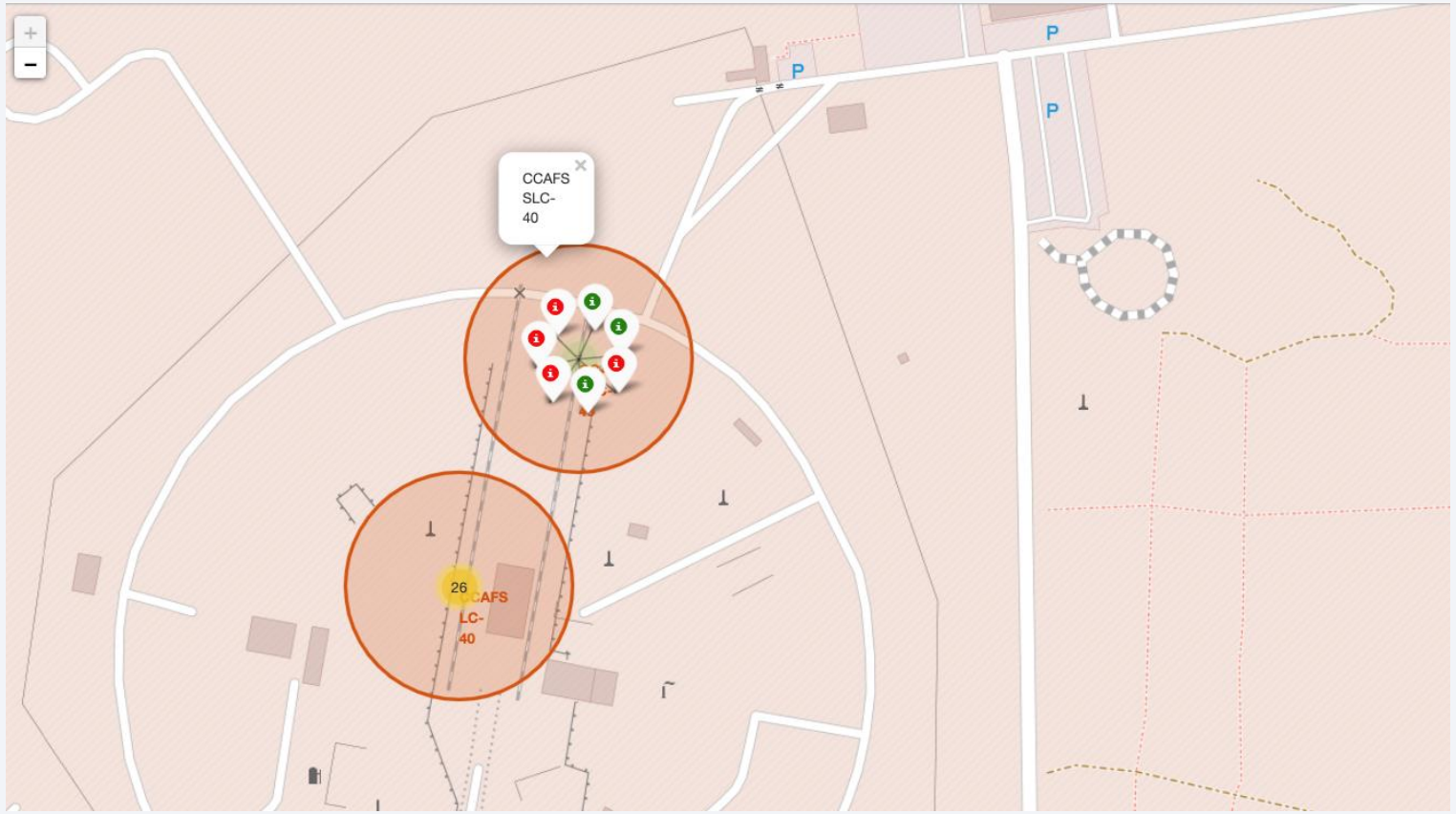
SPACE X LAUNCH SITES

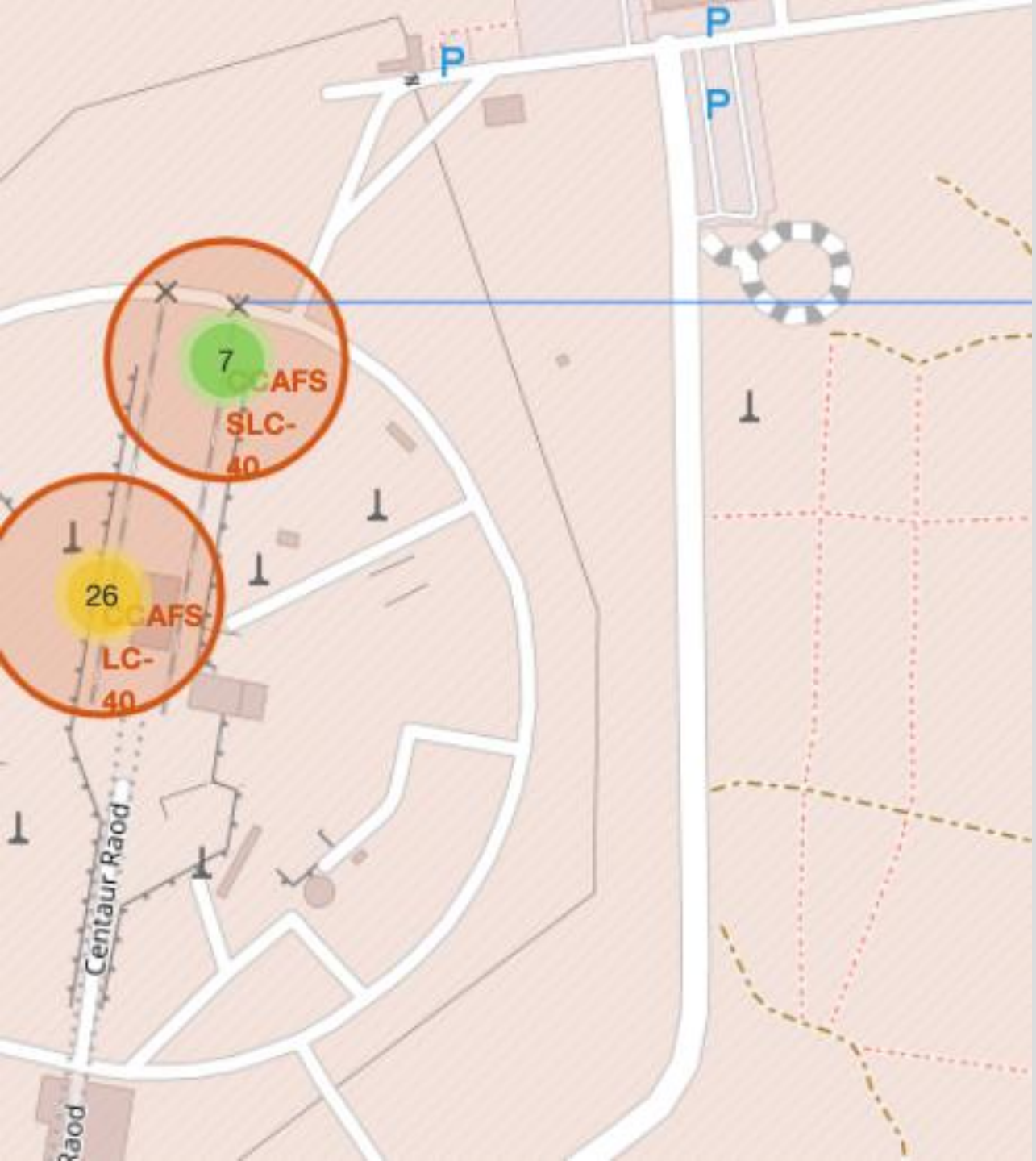


- The map highlights Space X launch sites in America, distributed around the coasts of California and Florida.

Color Markers Displaying Successful/Unsuccessful Launches

- The above map highlights KSC LC39A launch site as the most successful launch site in Florida.



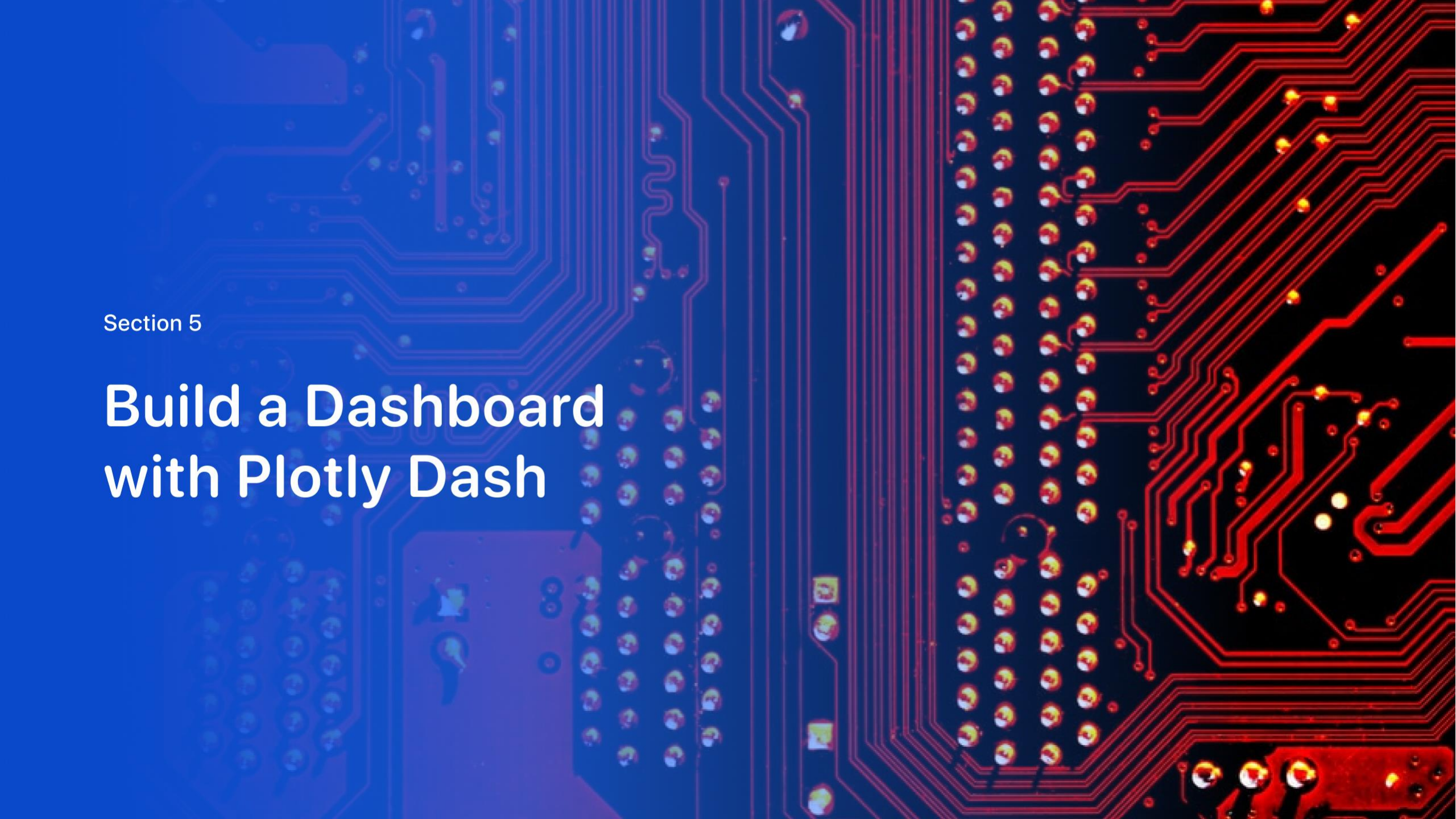


LAUNCH SITES LOCATED NEAR LANDMARKS: QUESTION TO SOLVE

- Launch sites are build in close proximity to coastline and highways as rocket launch is transported to coastline via highways. Distance is also maintained between railway and launch sites but not with as much distance as cities. It is also crucial for launch sites not to be in close proximity of cities as shown above.

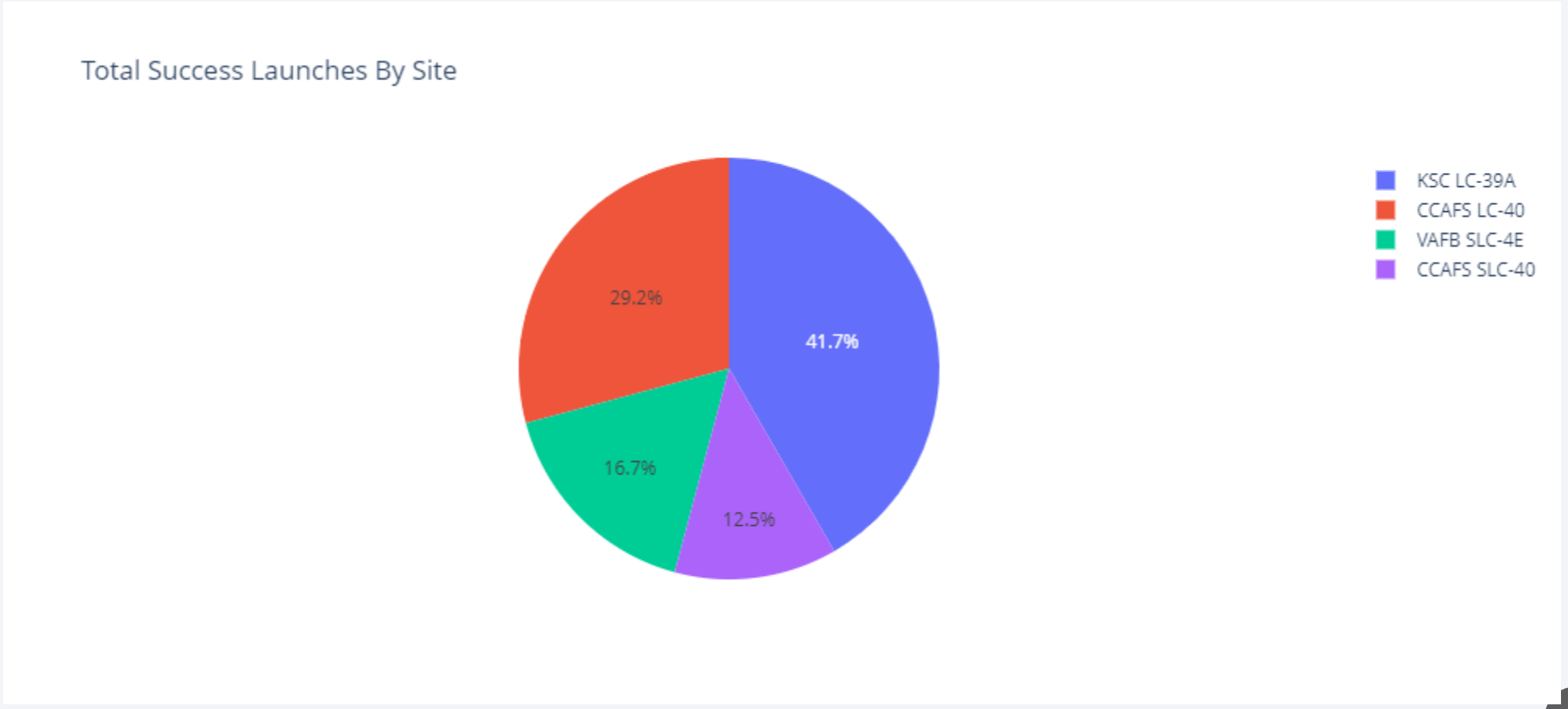
Section 5

Build a Dashboard with Plotly Dash



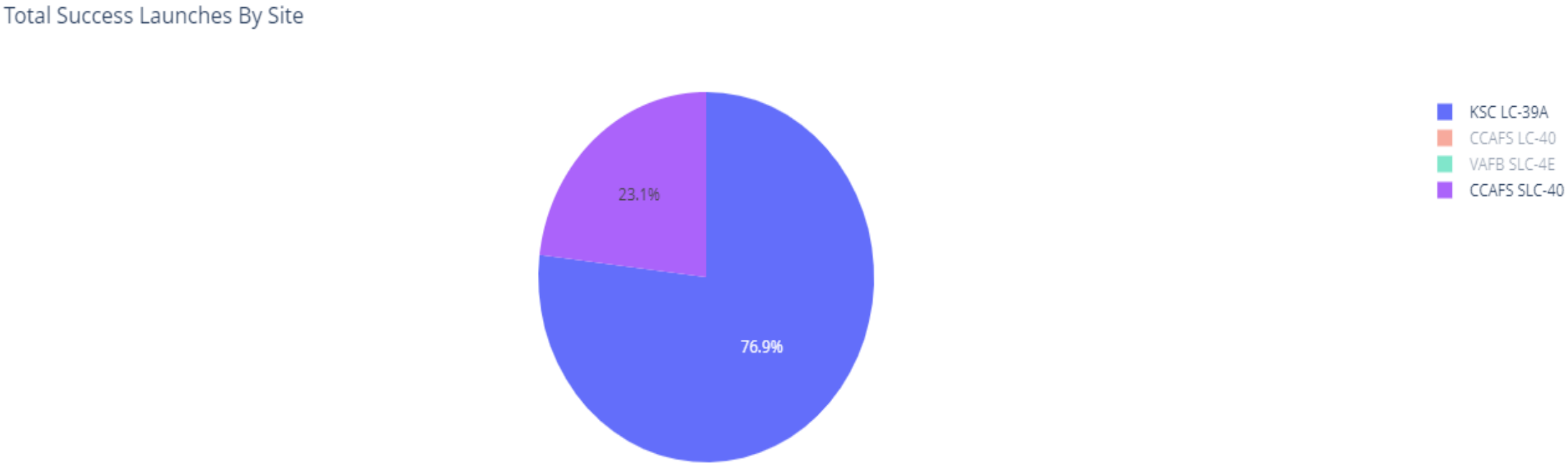
Launch Site Success

- Site KSC LC 39A showed the most success percentage.



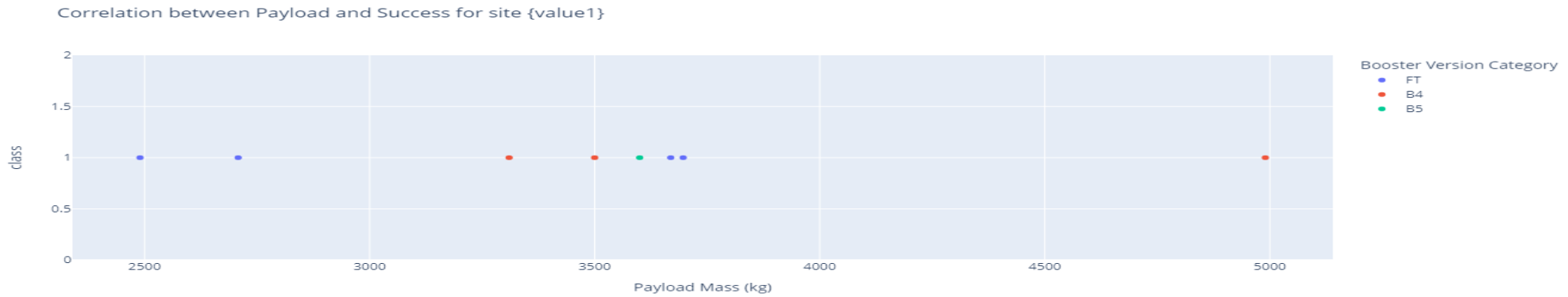
Launch site with highest success

- Site KSC LC 39A showed the most success at 76.9%

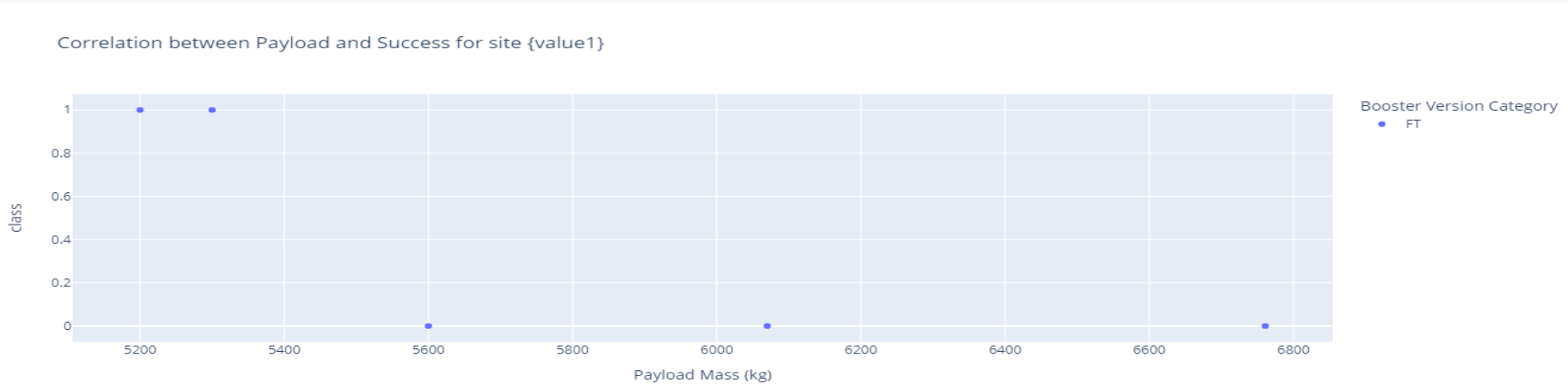


Payload vs. Launch Outcome

- When Payload mass is low, it is likely to have success landing outcomes



Payload between 0&5000



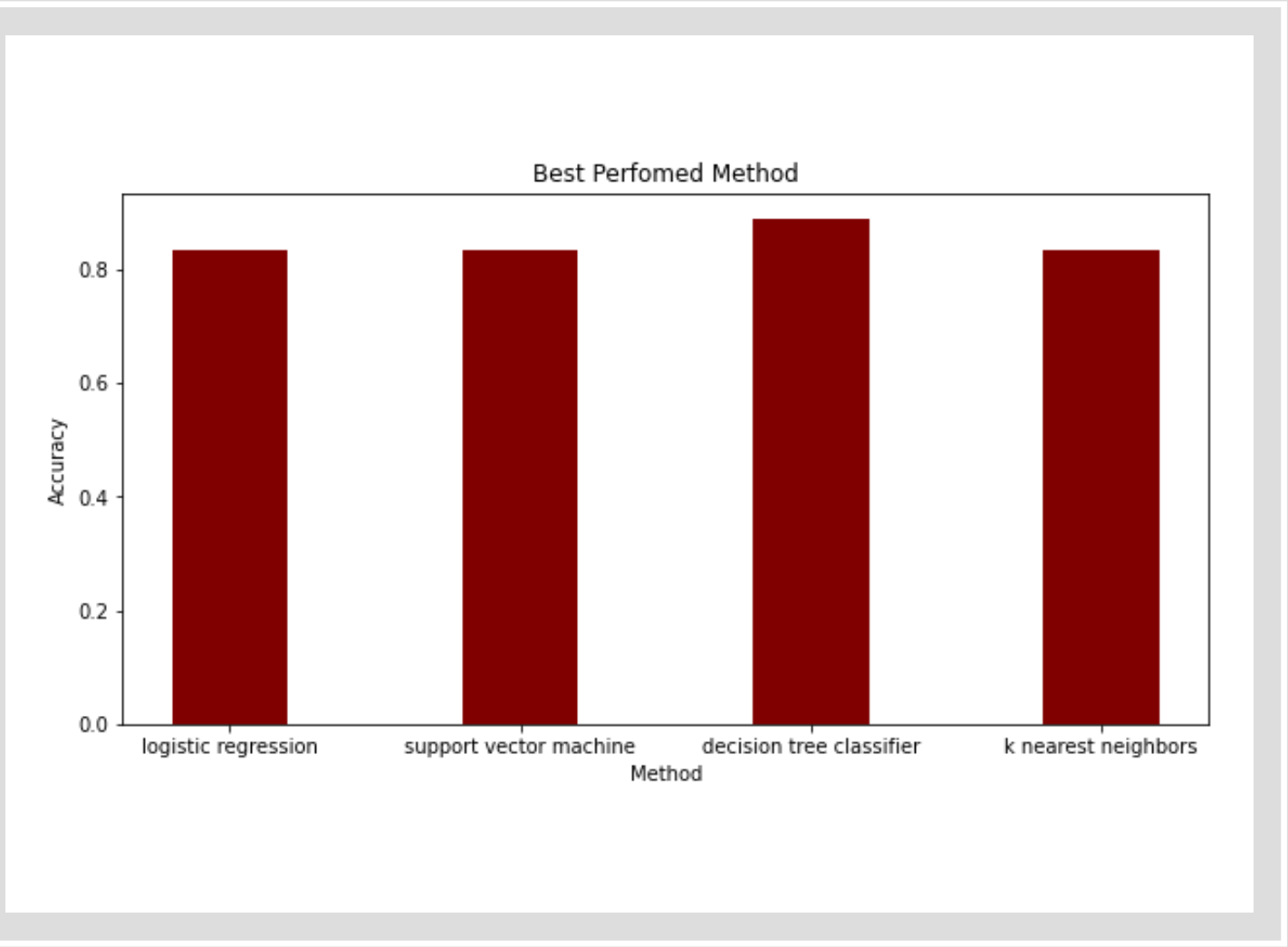
Payload between 5000 & 10000

Section 6

Predictive Analysis (Classification)

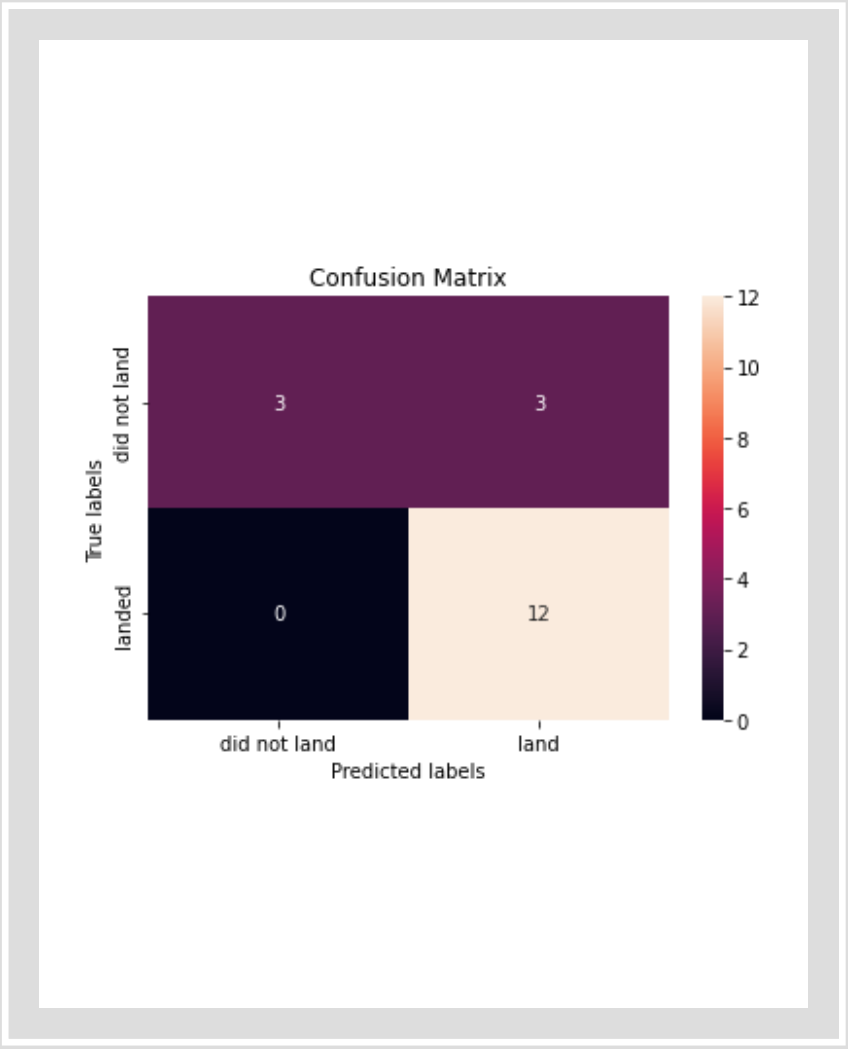
CLASSIFICATION ACCURACY

- KNN, Logistic Regression, Support vector machine and Tree classification model were employed. After tuned, we found the best model to present is
- Three of the models show the same accuracy rate. This could be due to small samples.
- The Tree classification model show 88.88% accuracy rate using test data
- Hence the best model is Tree classification



CONFUSION MATRIX

- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models overpredicts successful landings.



Conclusions

- Logistic regression is the chosen model as its easy to apply and explain across the business as all the models have same accuracy results.
- KSC LC-39A launch site has launched the most successful launches and is near a coast and highway.
- Low weighted payload rocket launches are more successful than heavier payloads.
- There is relationship between flight number and success rate, successful launch increases with number of flights.
- ISS, SSO and LEO orbits have also demonstrated higher success rate with lighter payloads.

Appendix

- GitHub repository url:

https://github.com/Siyan-Johny/Applied-Data--Science--Capstone-Project

- Special Thanks to my fellow peer reviewing my assignment

Thank you!