IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luv Surve
30th August 2022

Outline

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

Executive Summary

A new rocket company, wants to make space travel affordable for space enthusiasts. This report collects and analyses data from SpaceX rockets in order to develop a predictive model that can predict the reusability of the first stage based on various rocket and launching parameters hence, evaluating the price of each launch. The approach uses APIs, web-scraping from public information to retrieve data, store it in structured formats, which are used for exploratory data analysis, the outcomes are visualised and presented through dashboards. Finally, machine learning algorithms such as SVM, KNN, Decision Tree Classifier, and Logistic Regression are fitted with the data and tuned to optimal parameters, resulting in impressive accuracy.

Introduction

- A new rocket company wants to compete with existing commercial space travel companies.
- The project involves collecting launch data of competitor SpaceX, explore it to find insights that can allow predicting the reusability of the first stage of the rocket and hence, allowing the prediction of prices for each launch.

- Problems:
  - What kind of Booster Versions have highest success rate?
  - Which payload range is most effective ?
  - Can we predict the reusability of the first stage of the SpaceX rockets?

Section 1

Methodology

Methodology

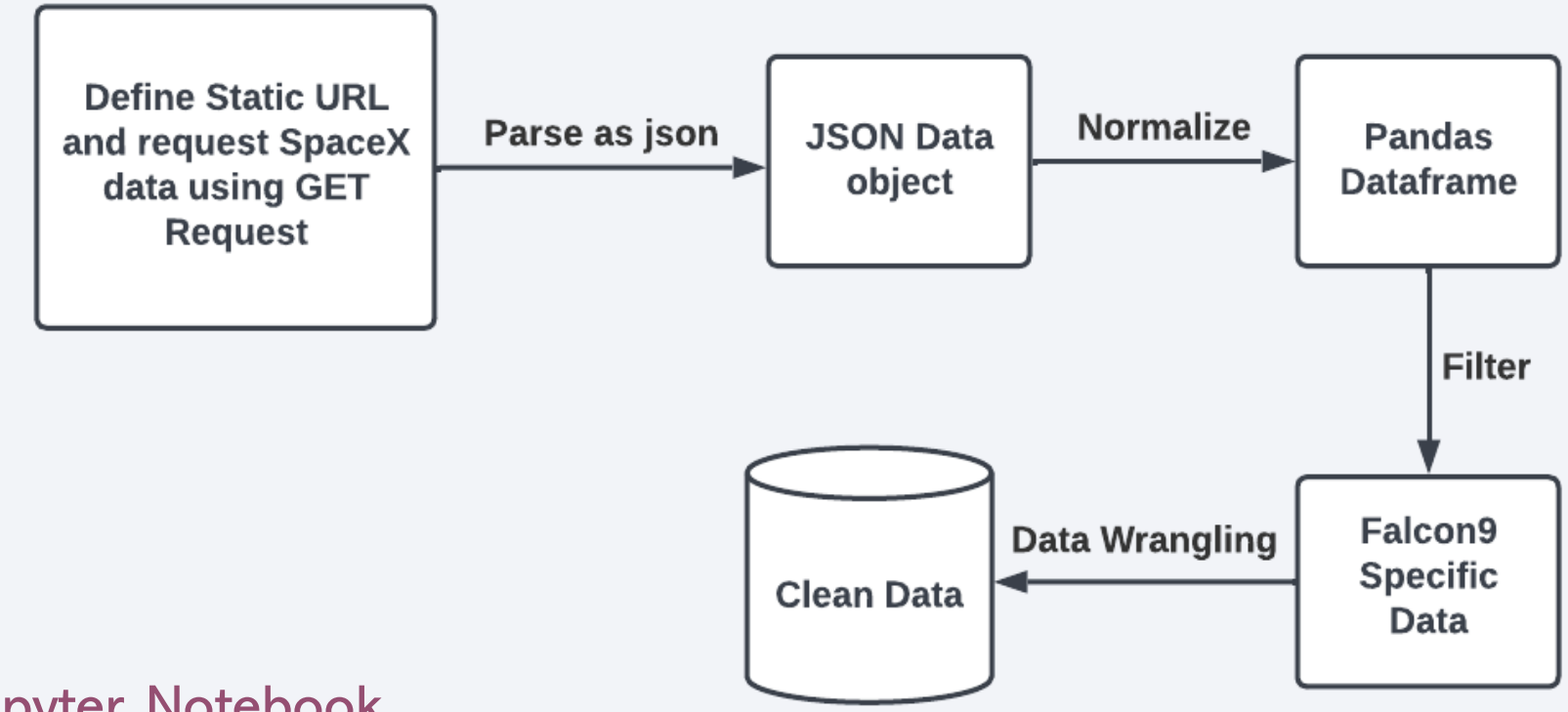
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and through Public data availed from Wikipedia.
- Perform data wrangling
  - EDA was performed on collected data and relationships between various attributes and target variables was analyzed to determine the training variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various classification GridSearchCV models are initialized to find the best set of optimum parameters that result in the highest accuracy.

Data Collection

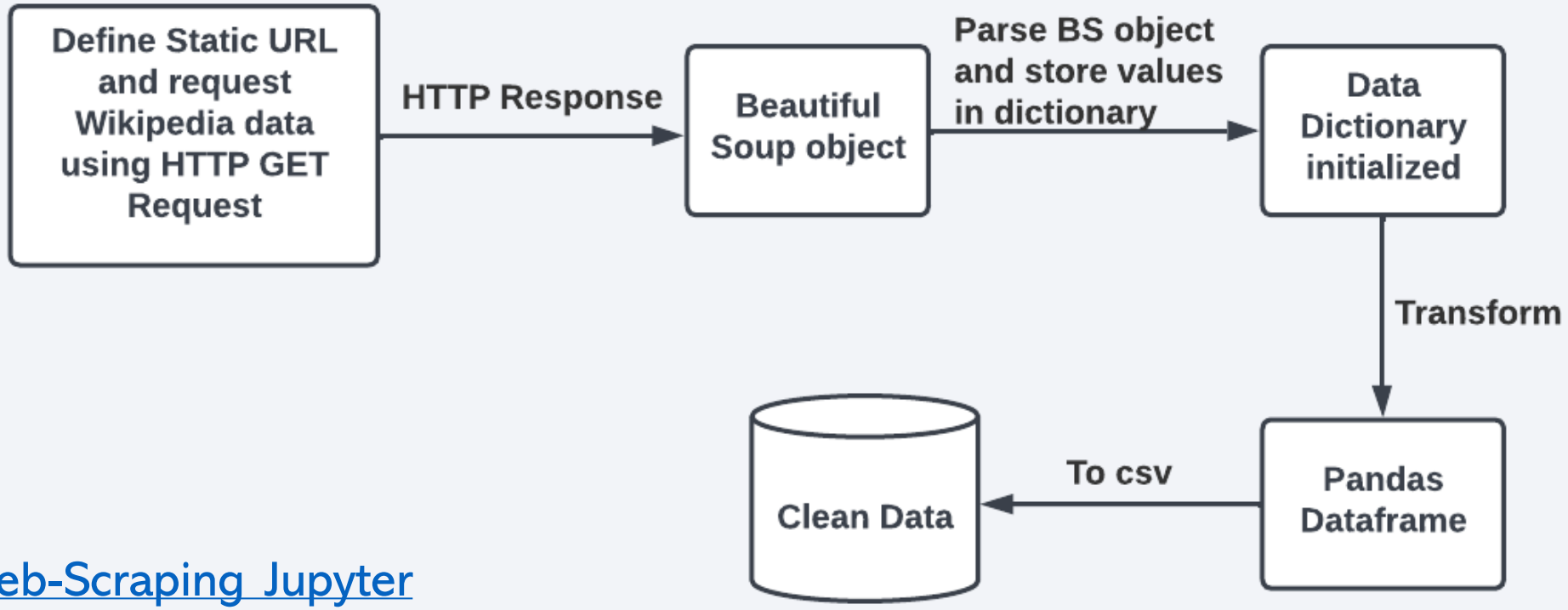
- Data is collected using SpaceX REST API and Web-scraping Wikipedia info.
- API:
  - SpaceX API static url defined.
  - Using GET Request and static url the data is received as response.
  - Response is normalized and stored.
  - Data is filtered to obtain Falcon9 specific data and is cleaned.
- Web-scraping:
  - Define Wikipedia page static url.
  - Using GET request and static url the HTTP response is collected.
  - Response is transformed as BeautifulSoup object.
  - Extracting and Parsing BS object and storing values in dictionary.
  - Converting dictionary to pandas dataframe and exporting as .csv file.

Data Collection – SpaceX API



- API Jupyter Notebook

Data Collection - Scraping

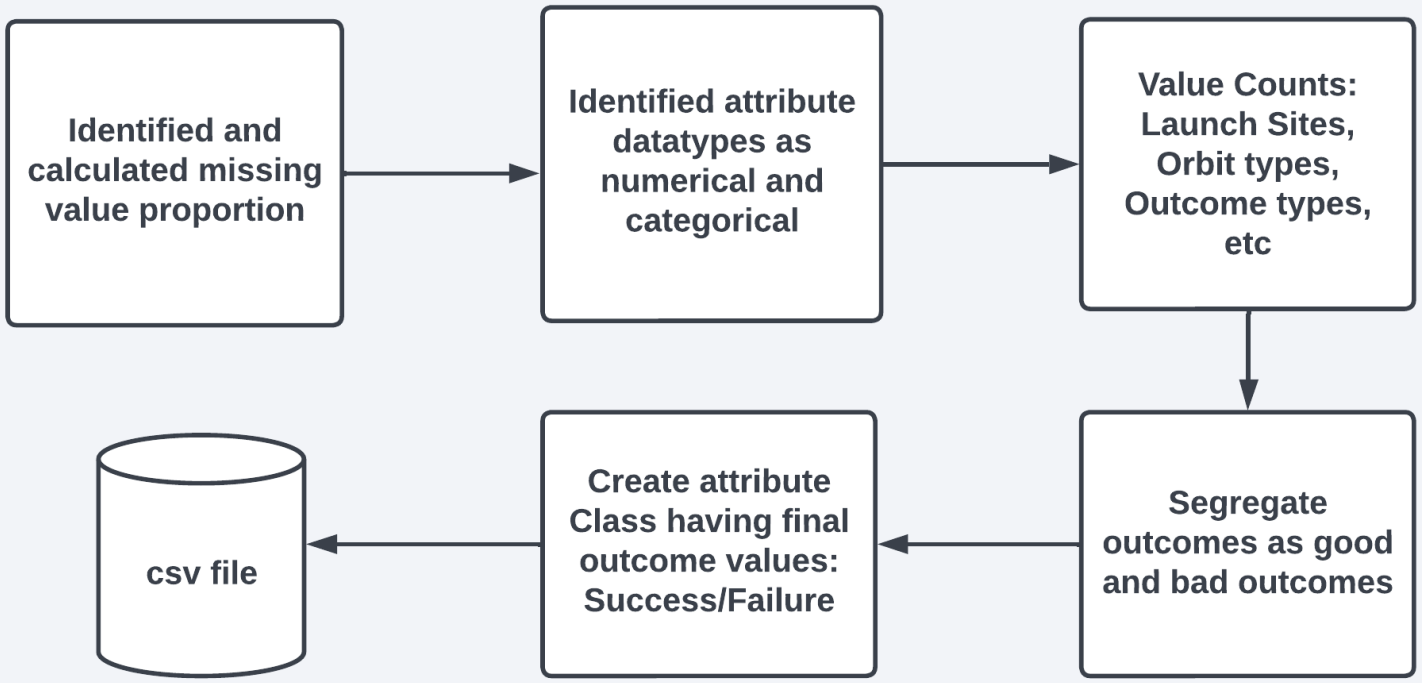


- Web-Scraping Jupyter Notebook

Data Wrangling

- Data Wrangling Jupyter Notebook

- Identify and find missing values
- Identified and found categorical and numerical values
- Value Counts of:
  - Launch sites
  - Orbit types
  - Landing Outcomes
- Landing outcomes classified as successful/Failure
- Exported data file as .csv file



EDA with Data Visualization

- Plotted scatterplots representing data:
  - Flight Number Vs. Payload
  - Flight Number Vs. Launch Site
  - Payload Vs. Launch Site
  - Flight Number Vs. Orbit Type
  - Payload Vs. Orbit Type
- Plotted Horizontal Bar Graph for data:
  - Success Rate Vs. Orbit Type
- Plotted a Line Graph representing data:
  - Success rate over the years-2010 to 2020
- Data Visualisation Jupyter Notebook

EDA with SQL

- SQL Queries:

- SQL EDA Jupyter Notebook

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA'
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date of the first succesful landing outcome in ground pad was acheived.
6. Names of the boosters which have success in drone ship and have payload mass in range [4000,6000) kgs
7. Total number of successful and failure mission outcomes
8. Names of the booster_versions which have carried the maximum payload mass using a subquery
9. Records which display the month, failure landing_outcomes in drone ship ,booster versions, launch_site in year 2015.
10. Ranked count of successful landing outcomes between the date 04-06-2010 and 20-03-2017.

Build an Interactive Map with Folium

- Marked and labelled all launch sites using circle markers with radius of 1000.
- Added launch site clusters with markers representing:
  - Successful launches(green markers)
  - Failed launches(red markers)
- Marked, labelled with distance from nearest locations:
  - High way
  - Railway line
  - City
  - Coastline
- Folium Map Jupyter Notebook

Build a Dashboard with Plotly Dash

- Added dropdown list to dashboard to select details of specific launch sites alongside default all sites option.
- Interactive pie-chart representing successful launch rate.
- Added a scatterplot representing Payload vs Outcome along with a payload mass range-slider.(To analyse outcomes of different payload ranges)
- Plotly Lab Python file(.py)

Predictive Analysis (Classification)

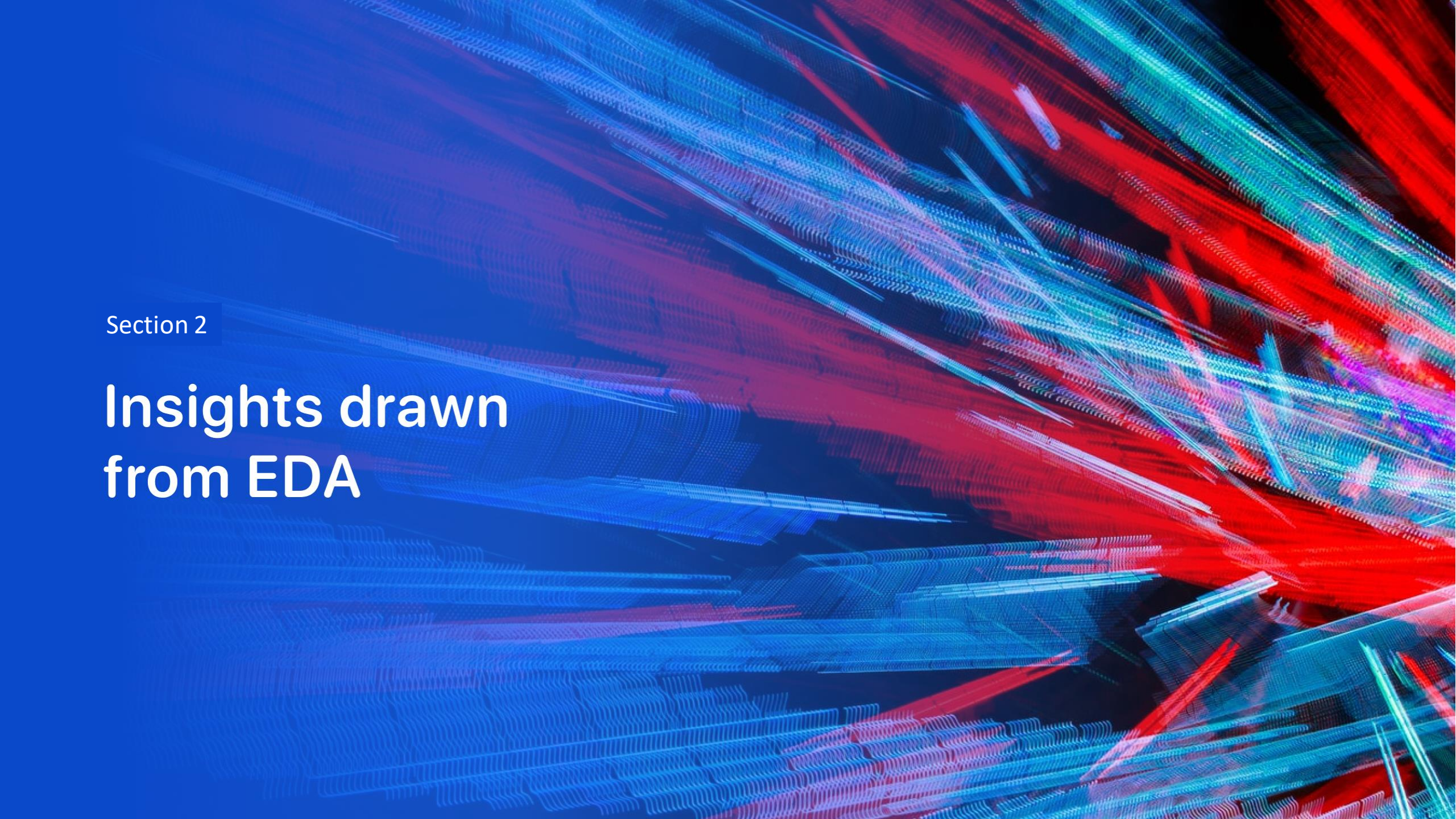
- Standardized inputs to make it suitable for machine learning algorithms(Using StandardScaler())
- Created Train-Test splits to measure the performances of various models
- Created various machine learning model instances:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Hypertuned parameters for each model using GridSearchCV()
- Developed confusion matrices for each model and calculated accuracy scores to analyse performance
- Compared performances of all models(Using a bar-chart) and selected the best model(Decision Tree Classifier)
- ML prediction Jupyter Notebook

Results

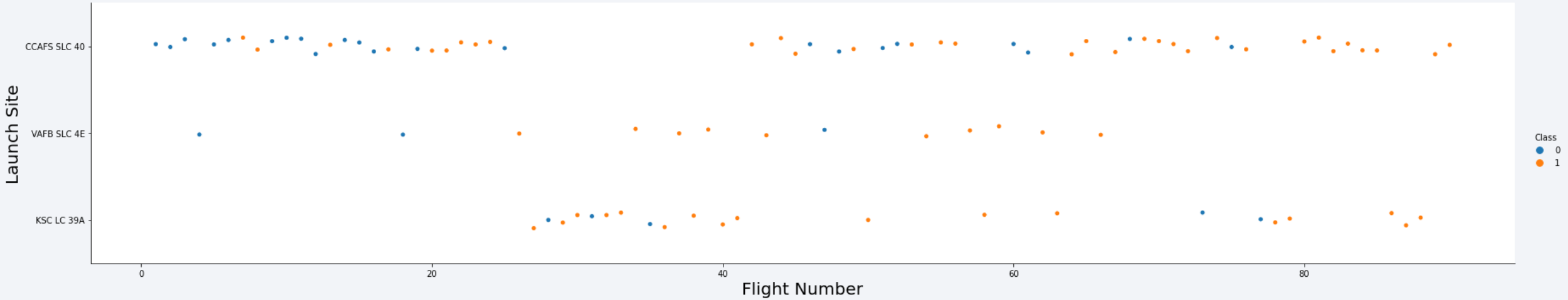
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

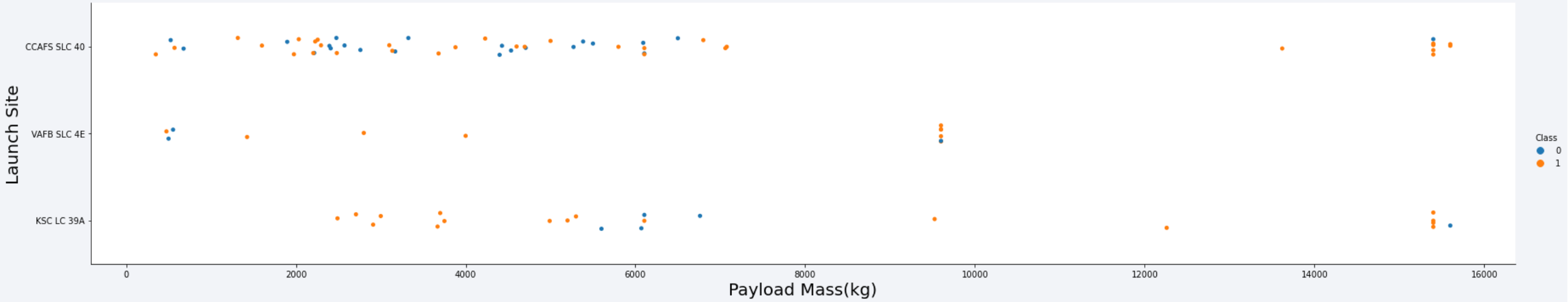


Launch success rates improve as the Flight numbers increase.

VAFB SLC 4E Shows improvement after the Flight number 30 mark.

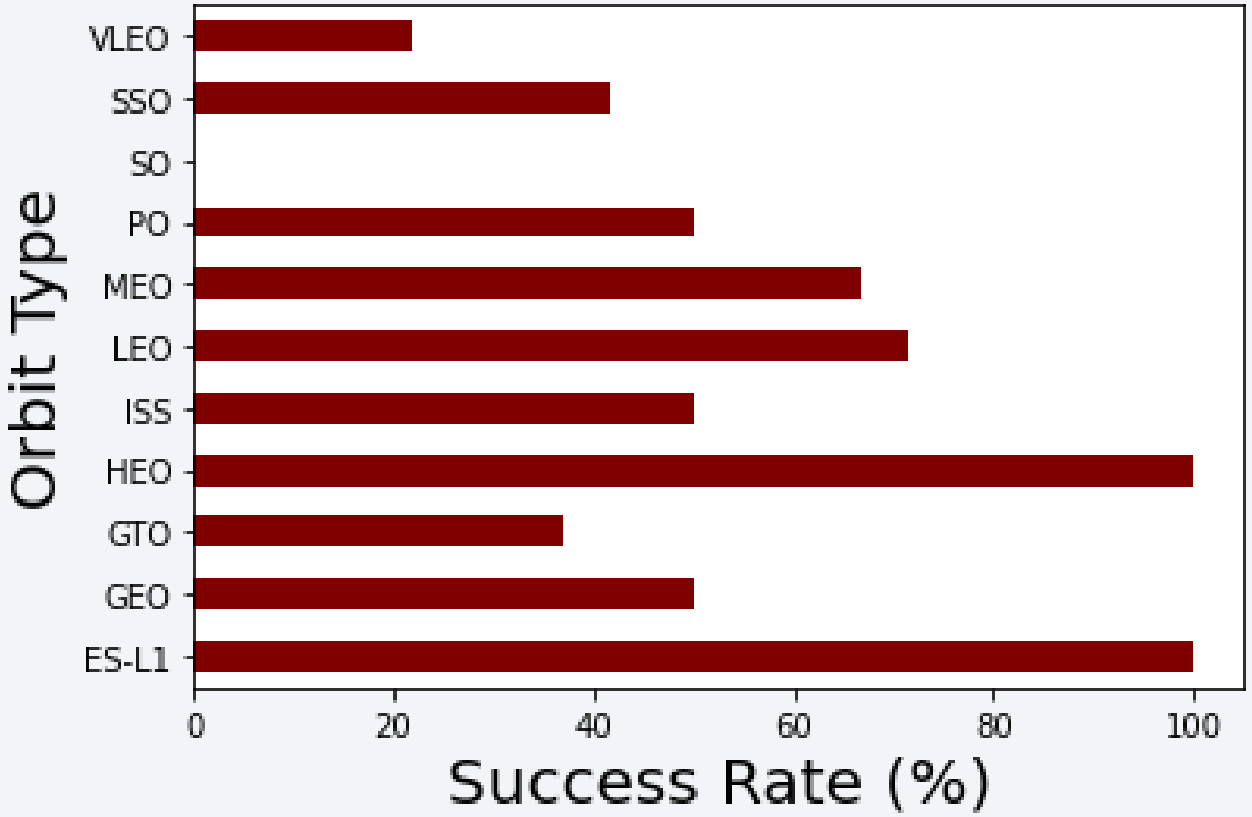
Remaining sites show improvement after the Flight number 70 mark.

Payload vs. Launch Site



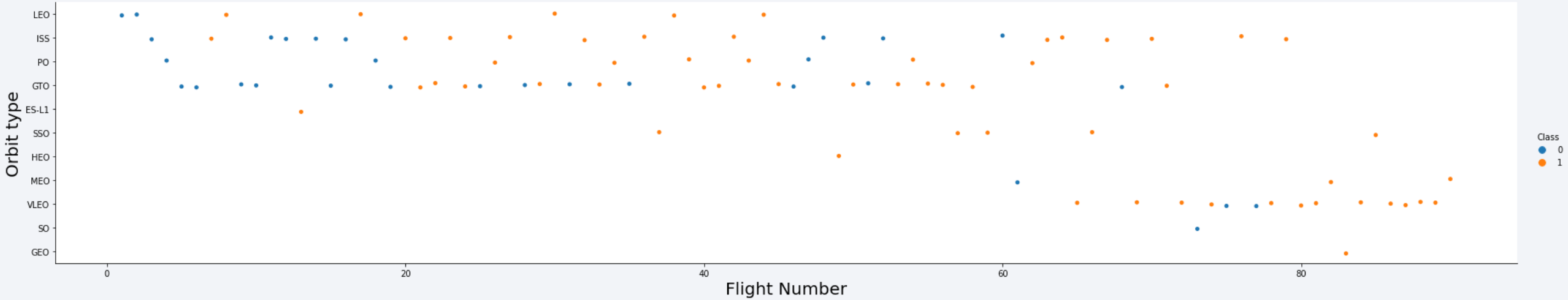
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



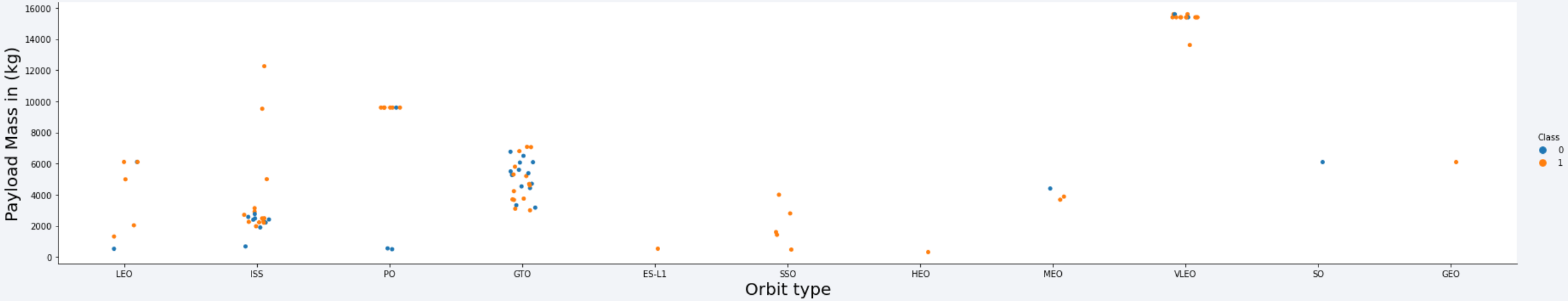
The Orbits HEO and ES-L1 have the highest success rates(100%) however, they together account for only 7 launches and are tailed by LEO with 76% success rate.

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.

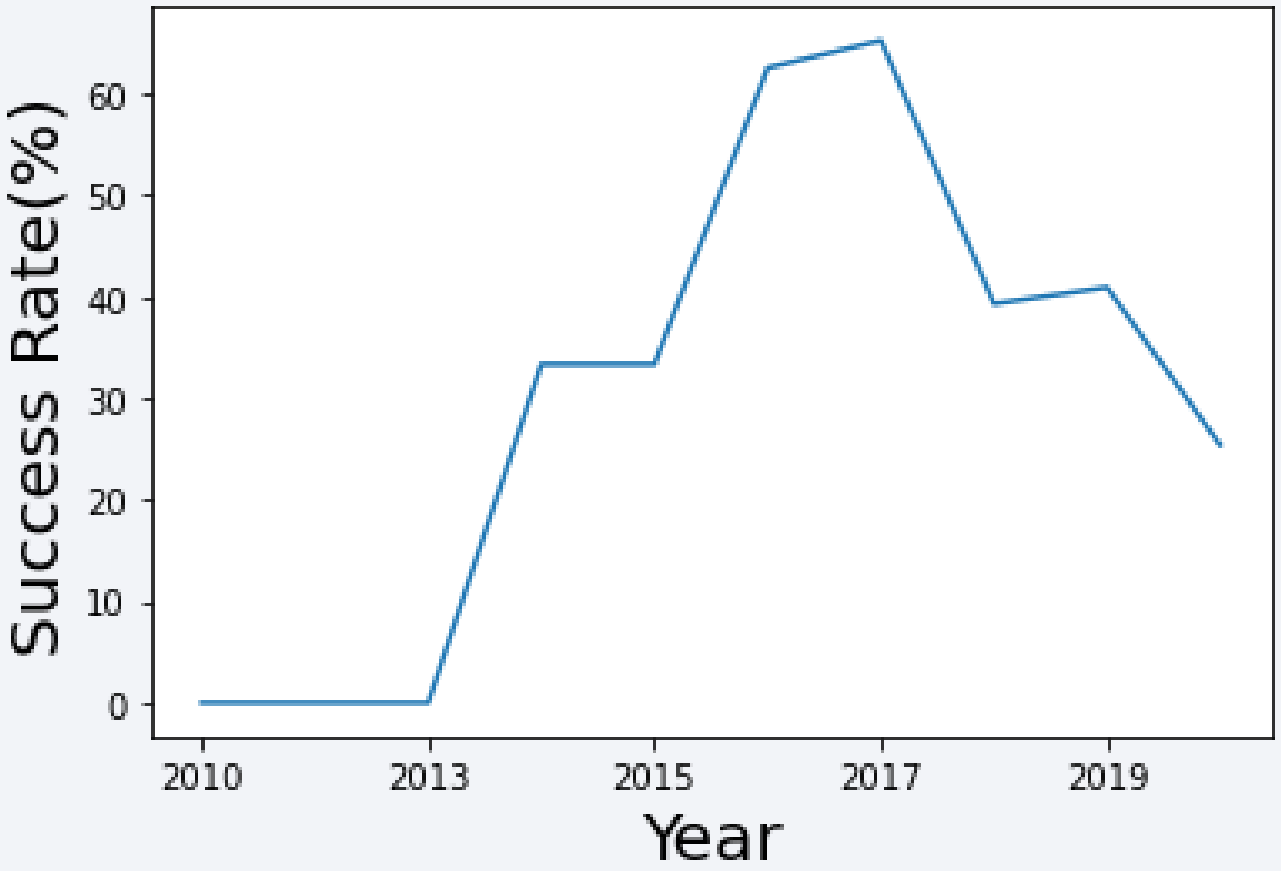
Payload vs. Orbit Type



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

- Unique launch sites:

Display the names of the unique launch sites in the space mission

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

In [8]:
```
%%sql
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

Total Payload Mass

- Total payload carried by boosters from NASA.
- 619967 kg

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [9]:
```
%%sql
select sum(payload_mass__kg_) from spacextbl;
```

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

Out[9]:

sum(payload_mass__kg_)
619967

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- 2928.4 kg

Task 4

Display average payload mass carried by booster version F9 v1.1

In [10]:
```
%%sql
select avg(payload_mass__kg_) from spacextbl
where booster_version = 'F9 v1.1';
```

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

Out[10]:

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad.

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [11]:
```
%%sql
select date from spacextbl
where date = (select min(date) from spacextbl
              where mission_outcome='Success');
```

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

Out[11]:

Date
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [12]:
```
%%sql
select booster_version 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
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes.
- 101 Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [13]:
```
%%sql
select count("Mission_Outcomes") from spacextbl;
```

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

Out[13]:

count("Mission_Outcomes")
101

Boosters Carried Maximum Payload

- Booster versions that have carried maximum payload mass.
- F9 B5 B10xx.x versions are heavy duty boosters.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [14]:
```
%%sql
select "Booster_Version" from spacextbl
where "PAYLOAD_MASS__KG_" = (select MAX(PAYLOAD_MASS__KG_) from spacextbl);
```

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

Out[14]:

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

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Both failures occurred at CCAFS LC-40 site.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [15]:
```
%%sql
select substr(Date,4,2) as Month, "Landing _Outcome", "Booster_Version","Launch_Site" from spacextbl
where substr(Date,7,4) = '2015' and "Landing _Outcome" = 'Failure (drone ship)';
```

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

Out[15]:

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

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

- Successful landings between 2010-06-04 and 2017-03-20.

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [16]:

```
%%sql
select count("Landing _Outcome") as COUNT, "Landing _Outcome" from spacextbl
where ("Date" between '04-06-2010' and '20-03-2017') and "Landing _Outcome" like 'Success%'
group by "Landing _Outcome"
order by "Date" DESC;
```

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

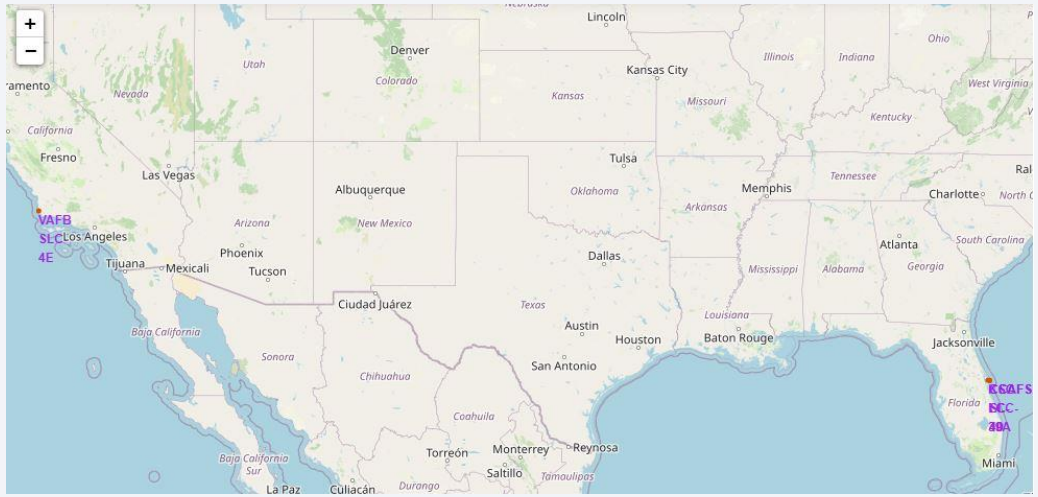
Out[16]:

COUNT	Landing _Outcome
6	Success (ground pad)
8	Success (drone ship)
20	Success

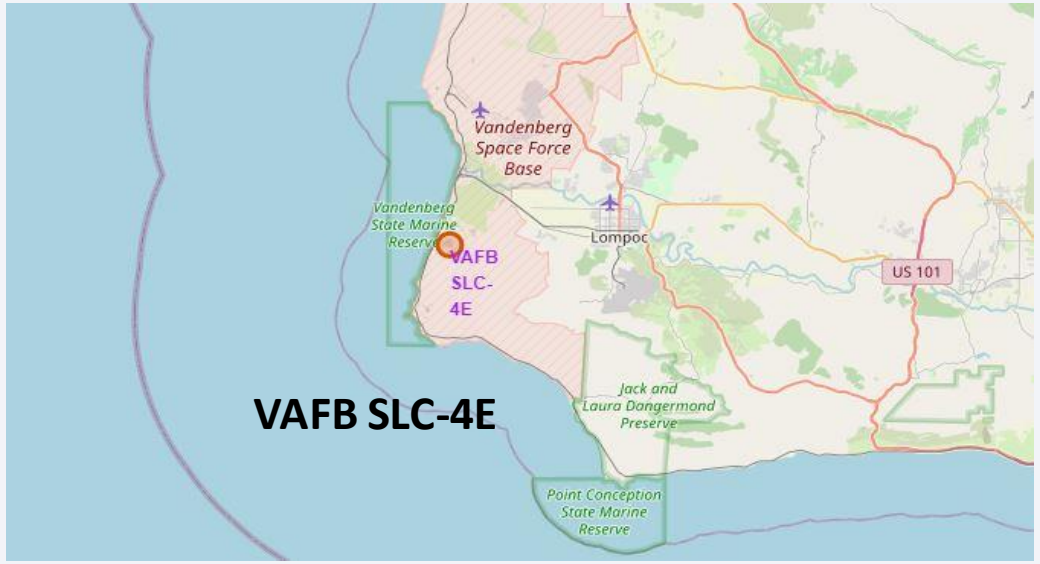
Section 3

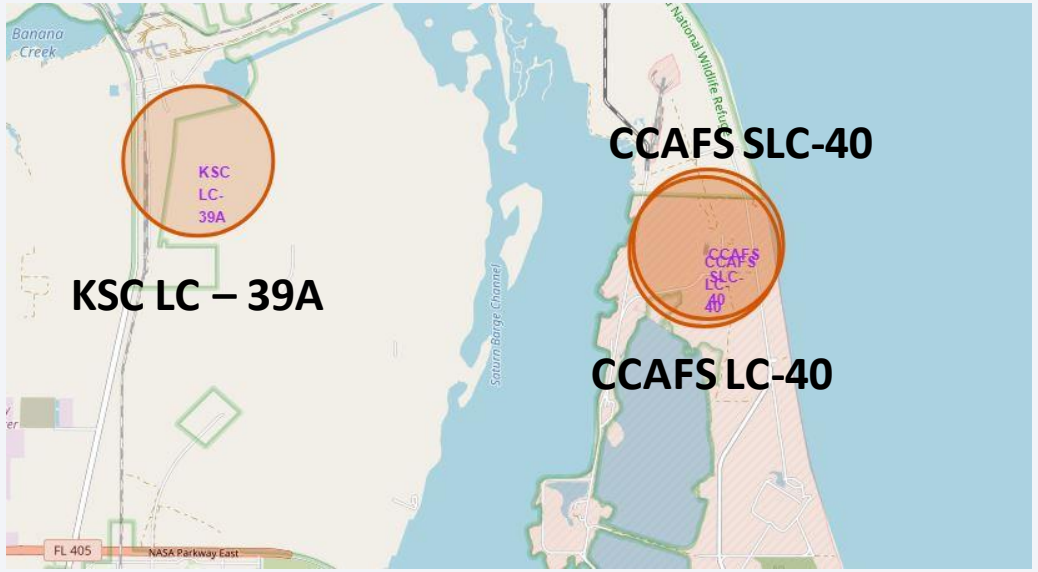
Launch Sites Proximities Analysis

Launch Site locations

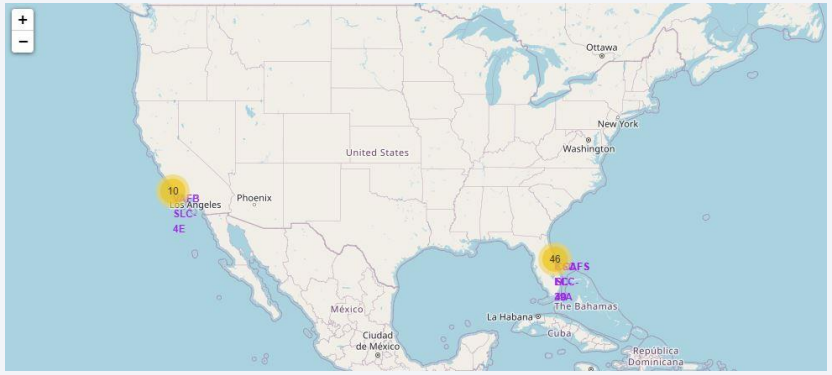


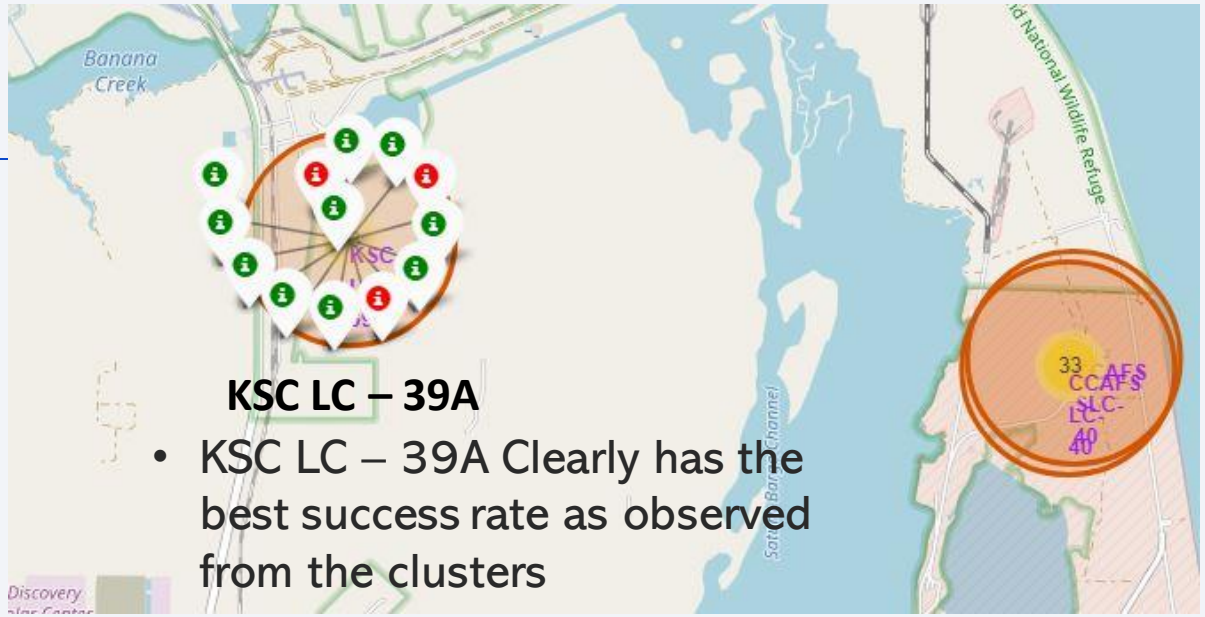
- Launch sites are situated at coastal regions of Los Angeles and Florida





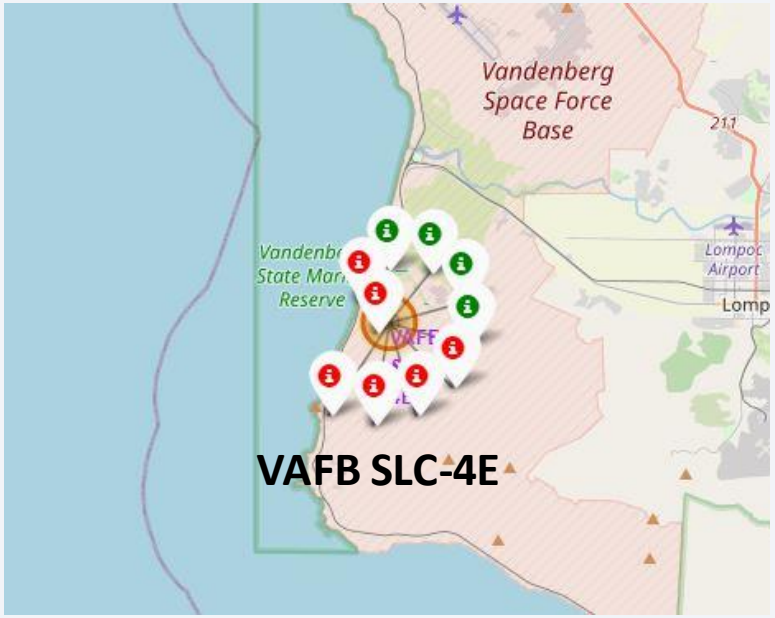
Launch Outcome Clusters



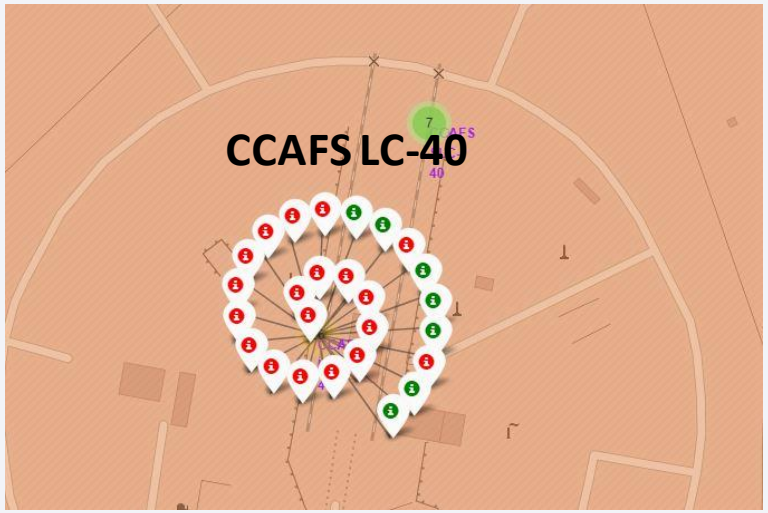


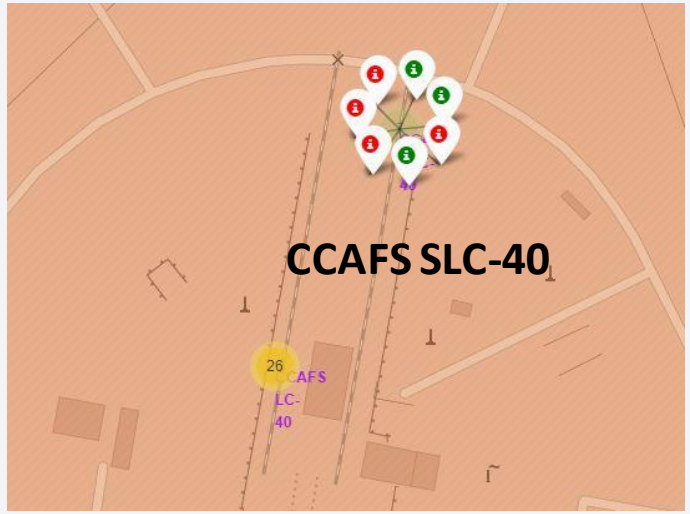
KSC LC – 39A

- KSC LC – 39A Clearly has the best success rate as observed from the clusters

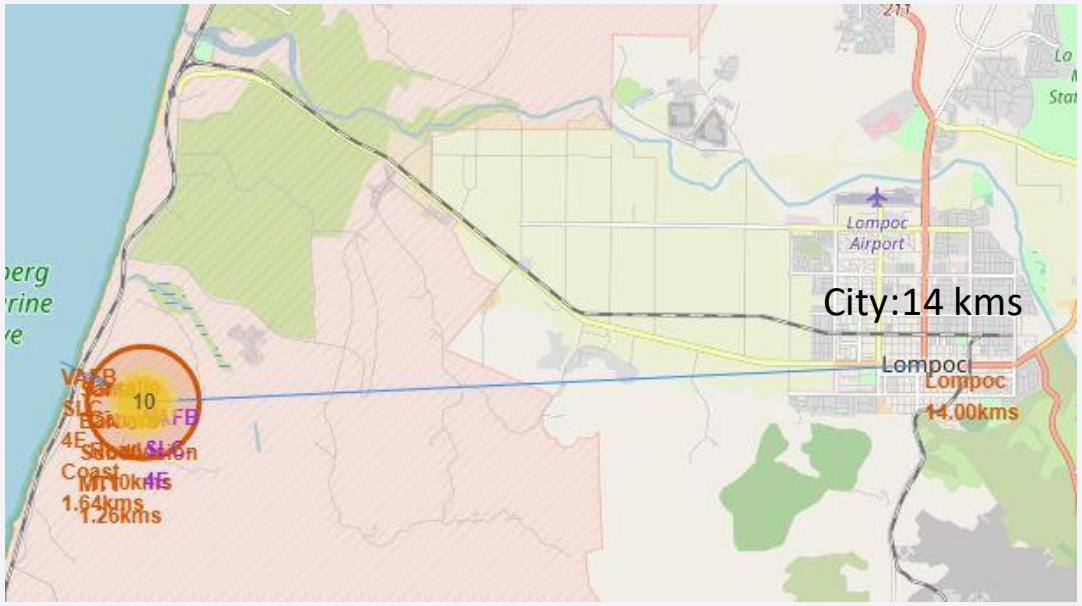


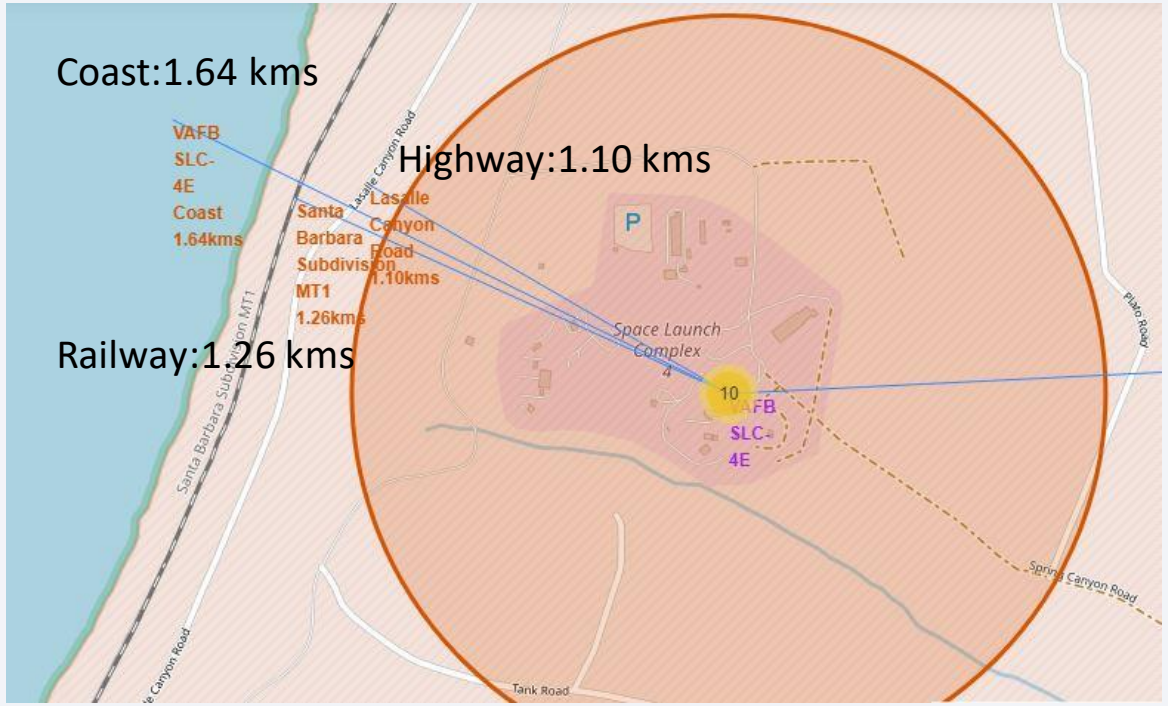
VAFB SLC-4E



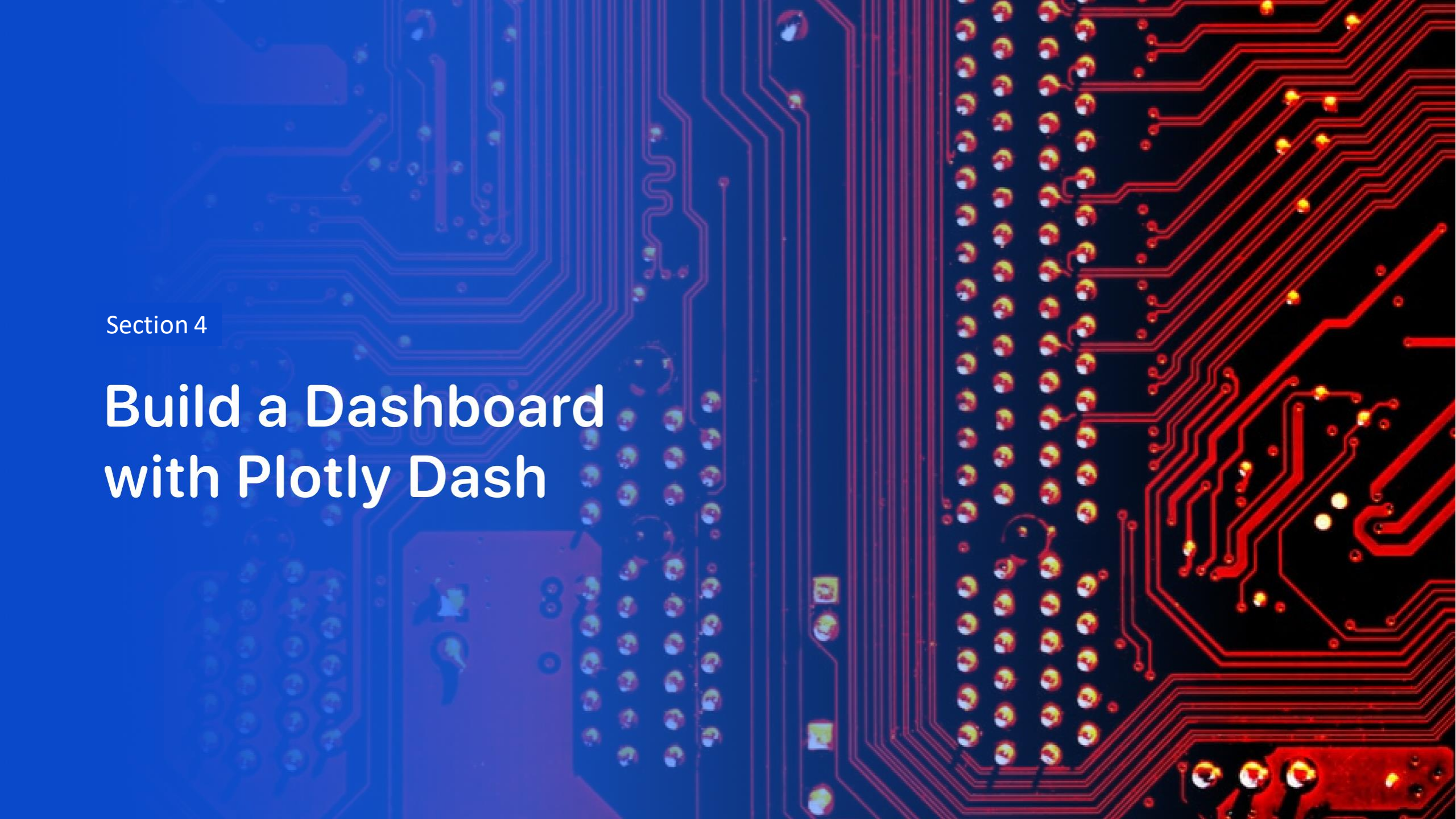


Important Nearest Locations



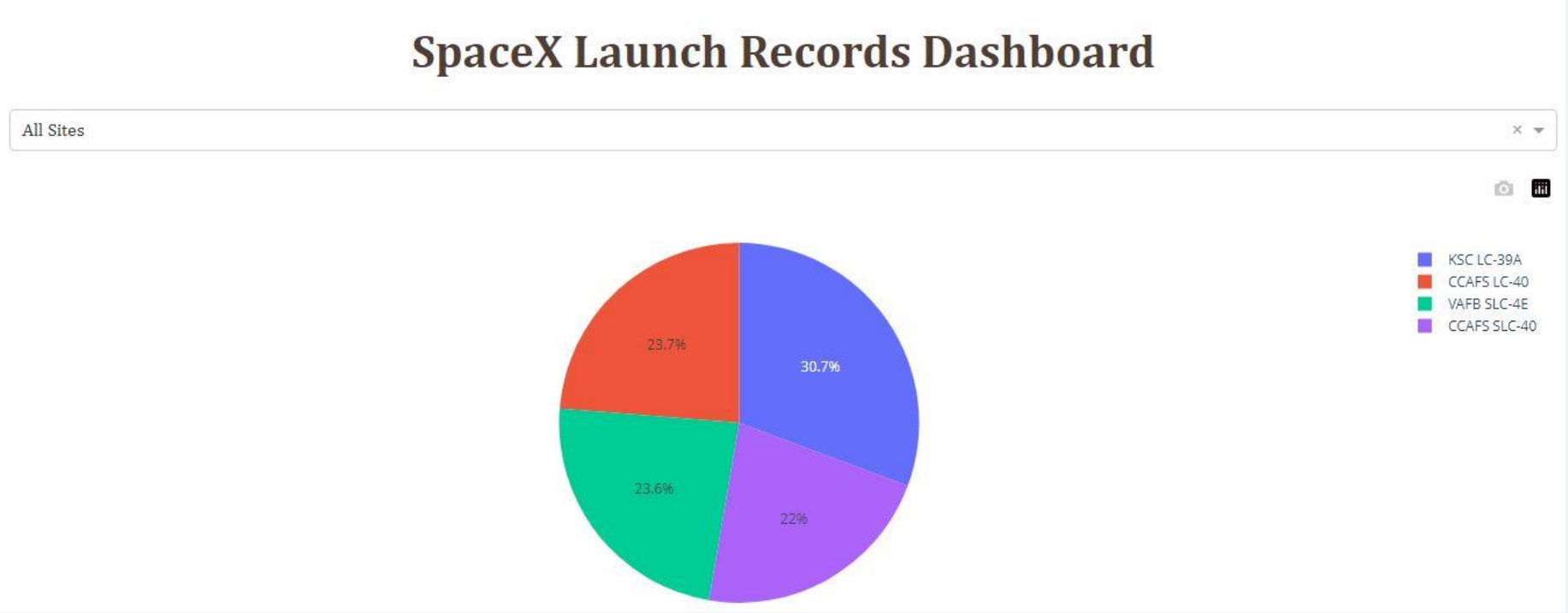


- All major vehicular pathways are situated at least 1km away of the site.
- Site is situated 10+km away from the city


Section 4
Build a Dashboard with Plotly Dash

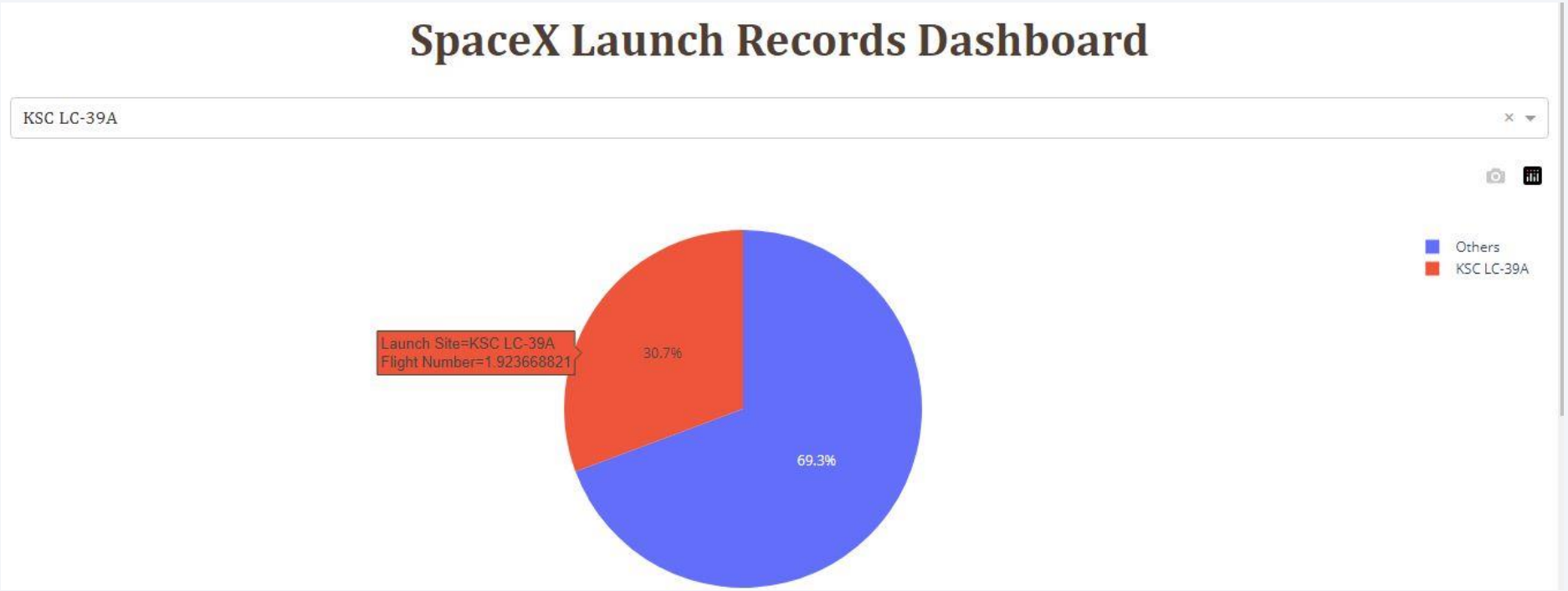
Launch Site Success Rates

- Launch sites VAFB SLC – 4E(Los Angeles) and CCAFS LC-40(Florida) are situated in similar positions on either coasts and perform similar.
- KSC LC 39A of Florida is situated in a more interior region in comparison to other sites and performs significantly better than them.



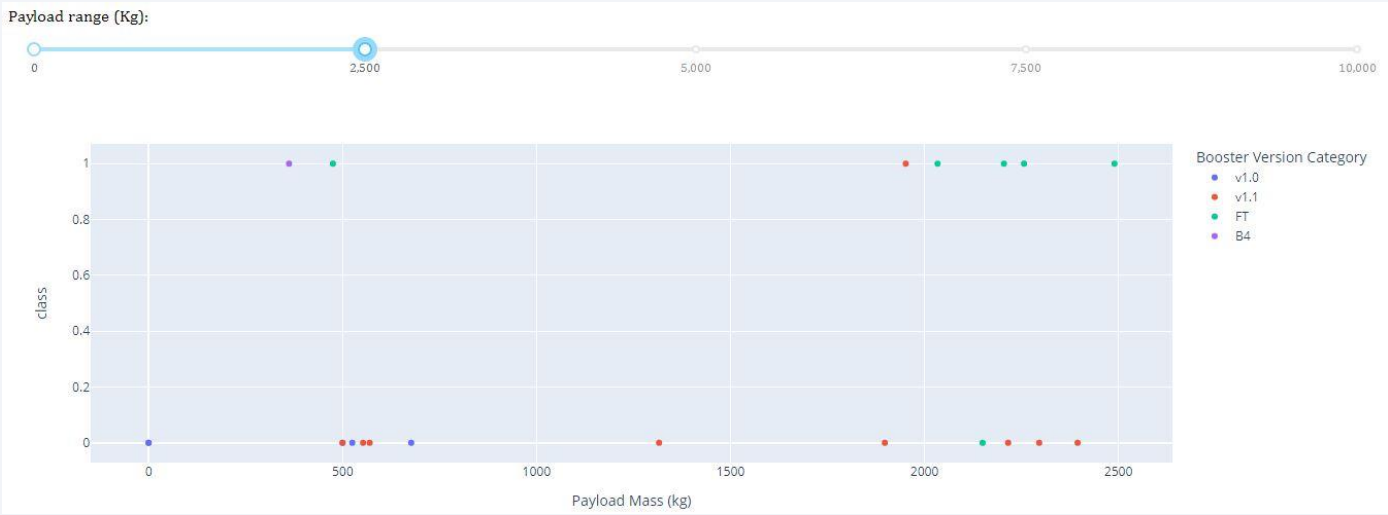
Most Successful Launch Site

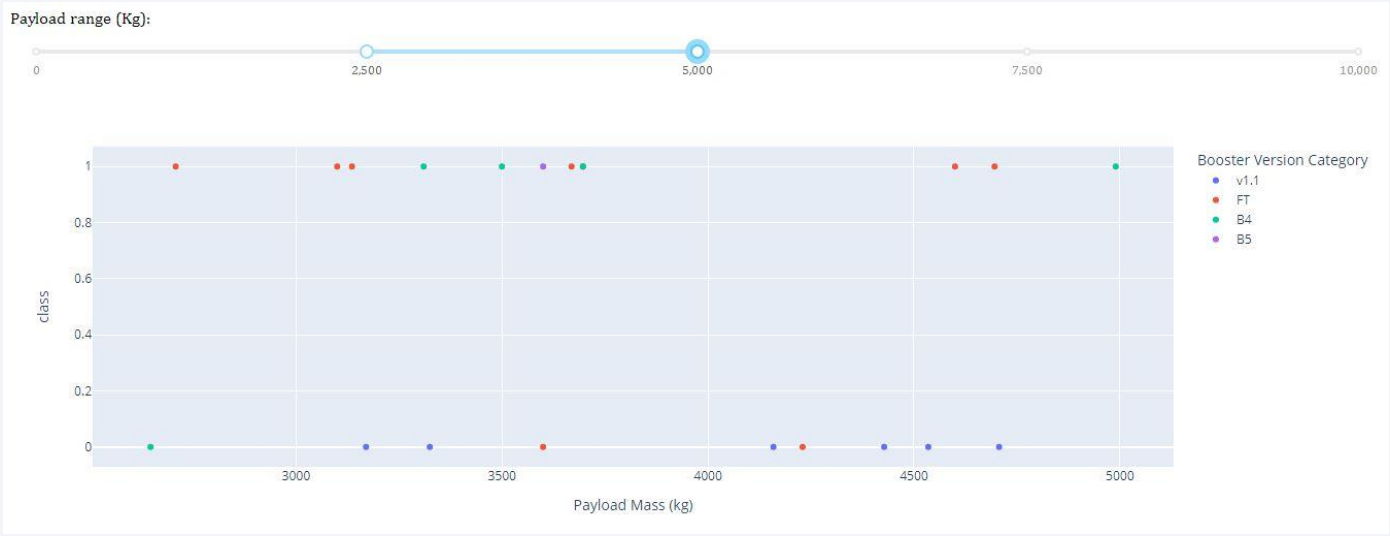
- KSC LC 39A of Florida is situated in a more interior region in comparison to other sites and performs significantly better than them.
- Has a success Rate of 30.7%



Payload Mass Vs Outcome

- Range with the highest launch success rate: 2500-5000
- Range with the lowest launch success rate: 7500-10000





- F9 Booster version with the highest launch success rate: FT

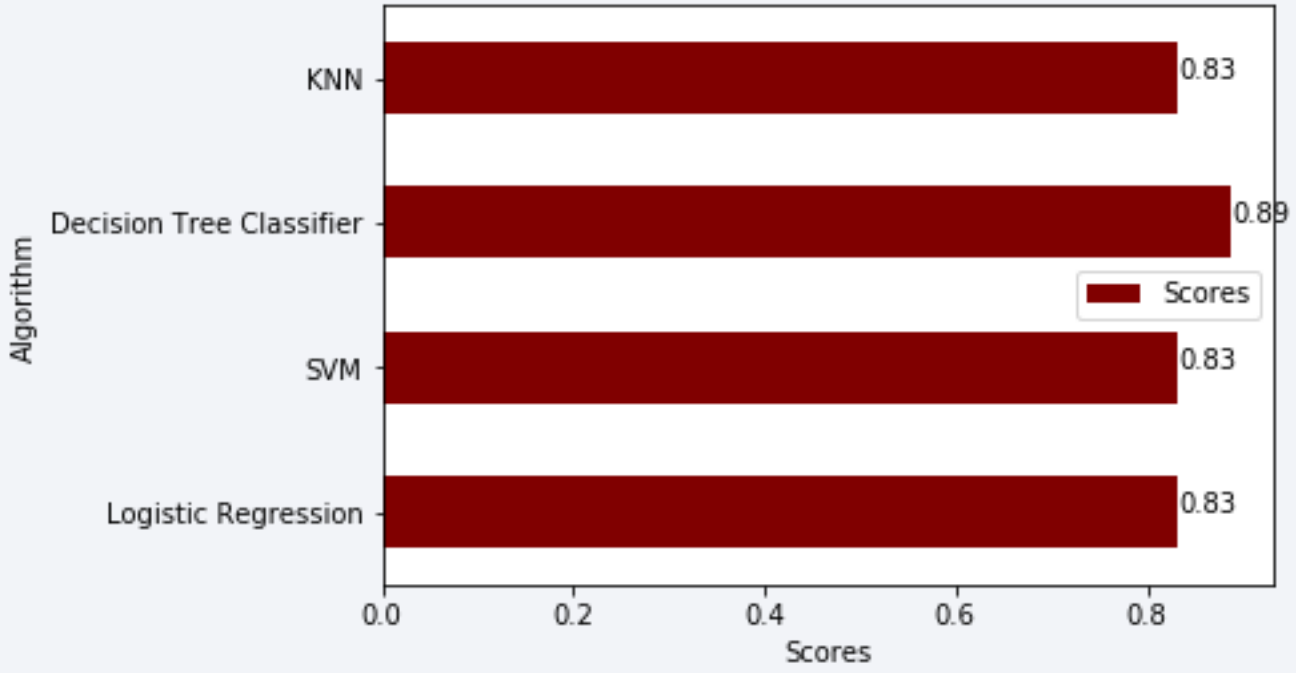
- 5000 - 7500

- Range with the lowest launch success rate: 7500-10000

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree Classifier has the highest accuracy score

Confusion Matrix

- Confusion Matrix of Decision Tree Classifier.
- Predictions:
  - TP count: 5
  - TN count: 11
  - FP and FN: 1 each

Conclusions

- Launch success rates increase with increasing number of flights indicating developing technology and reinforcement approaches taken by the research teams.
- Not all launch sites cater heavy payload launches affecting the success rates of launch sites.
- Different orbit types have different behaviours (e.g. relationship with flight numbers)
- Success rates of launches has increased over the years indicating advancements.
- Launch Sites are located at near coastal regions to limit the effects of launch failures.
- However, KSC LC-39A situated further from coastal regions compared to other sites performs better than others, could indicate to the possibility that launch sites further from coastal regions have a better chances of success.
- Decision Tree Classifier adapts better compared to other models with the provided dataset and is promising for similar datasets.

Appendix

- API Jupyter Notebook
- Web-Scraping Jupyter Notebook
- Data Wrangling Jupyter Notebook
- Data Visualisation Jupyter Notebook
- SQL EDA Jupyter Notebook
- Folium Map Jupyter Notebook
- Plotly Lab Python file(.py)
- ML prediction Jupyter Notebook

Thank you!