IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arun Deepak Tirkey
December 2, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - SpaceX Data Collection using SpaceX API
  - SpaceX Data Collection with Web Scraping
  - SpaceX Data Wrangling
  - SpaceX Exploratory Data Analysis using SQL
  - Space-X EDA DataViz Using Python Pandas and Matplotlib
  - Space-X Launch Sites Analysis with Folium-Interactive Visual Analytics and Plotly Dash
  - SpaceX Machine Learning Landing Prediction
- Summary of all results
  - EDA results
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis(Classification)

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully using data from Falcon 9 rocket launches advertised on its website.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - How datasets were collected ?
- Perform data wrangling
  - How data were processed ?
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models ?

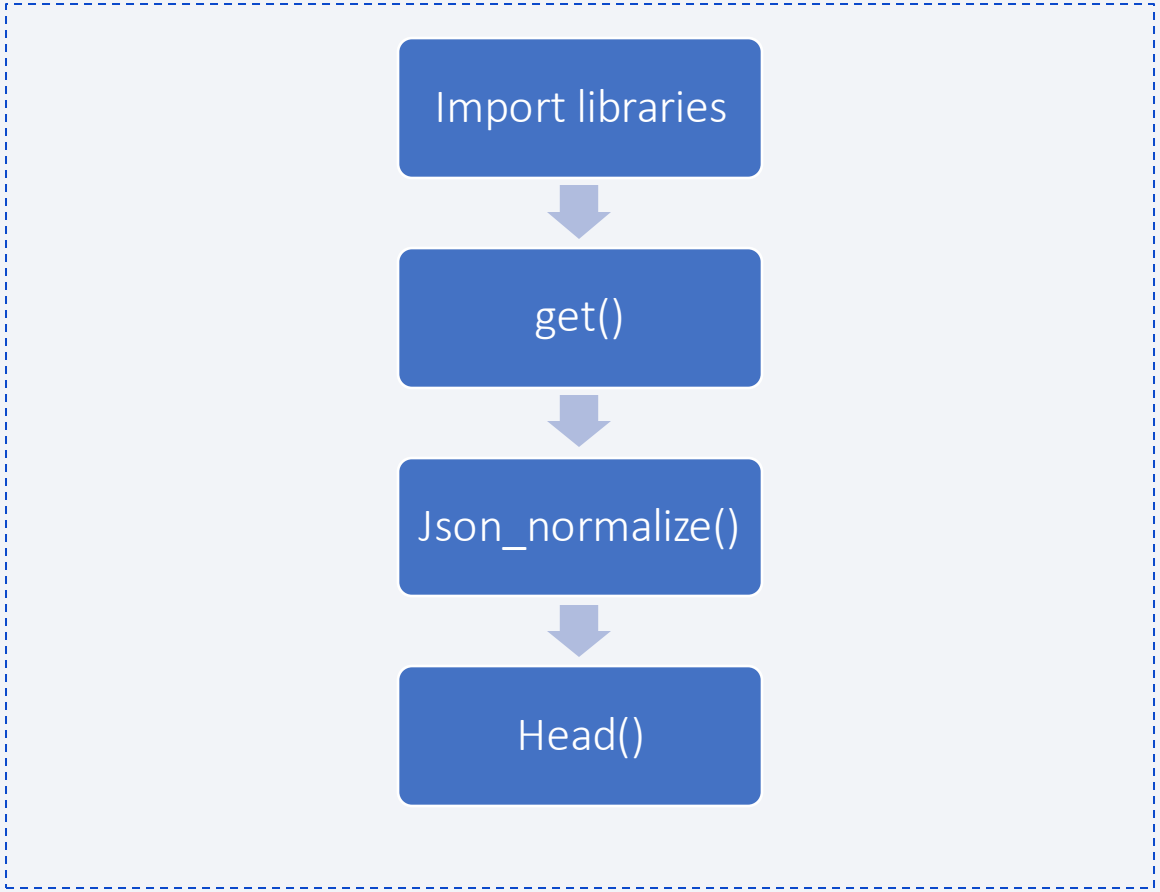
Data Collection

How data sets were collected?

- Data were requested from SpaceX API.
- Web scrapped Falcon9 and Falcon Heavy records from Wikipedia.

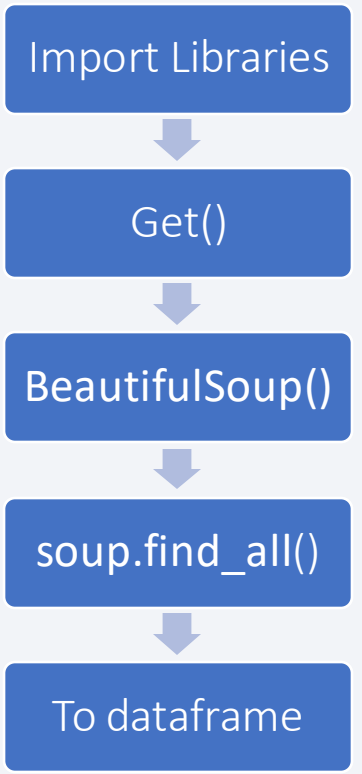
Data Collection – SpaceX API

- Import requests, pandas and numpy libraries.
- Request for data from SpaceX API using Get() request.
- Normalised the json data into tables.
- Output requested data with Head() function.
- URL- Data Collection-SpaceX_API



Data Collection - Scraping

- Import requests, pandas, re and BeautifulSoup libraries.
- Requesting response from wikipedia using Get() request.
- Create a BeautifulSoup object from a response.
- Find relevant entries of table using find_all() function.
- Form dataframe with all entries.
- URL- Data_Collection-WebScrapping

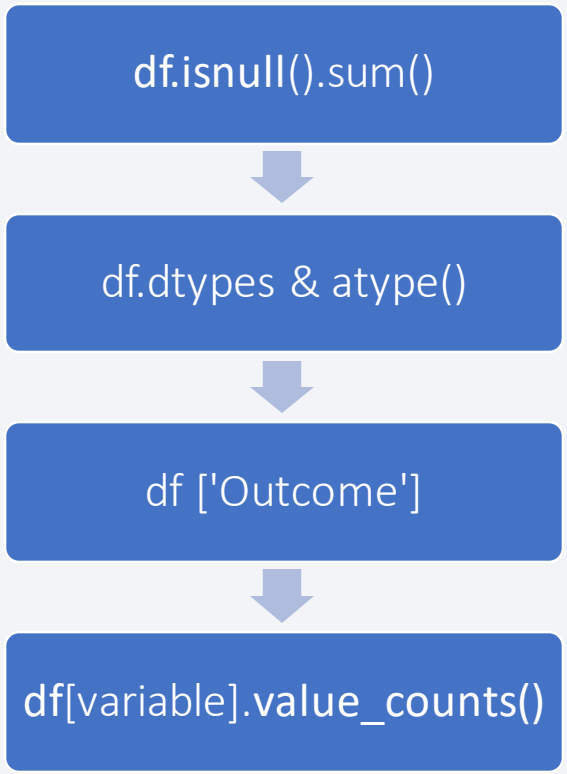


Data Wrangling

How data were processed?

- Check for missing values.
- Correct datatype of variables.
- Identify dependent variable .
- Searching for valuable insights within variables statistically.

- URL- [Data_Wrangling]



EDA with Data Visualization

What charts were plotted and why you used those chart?

- Scatter plot – Scatter plot can have three dimensions(variables) which help to visualize relationship of two variables based on dependent variable. Thus, Flight Number, Launch Site, payload mass and orbit columns were used to get relationship between themselves based on 'Class' variable.
- Bar chart - Bar chart used to check average success rate of each orbit.
- Line chart - Line chart is the best to show trends that's why success outcome was shown over period of years.
- URL- Data_Viz

EDA with SQL

The following SQL queries were performed :-

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

```
%sql select distinct "Launch_Site"  from SPACEXTBL
```

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

```
%sql select *  from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

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

```
%sql select SUM(PAYLOAD_MASS__KG_)  from SPACEXTBL where "Customer" = "NASA (CRS)"
```

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

```
%sql select  Booster_Version from SPACEXTBL WHERE "LANDING_OUTCOME" LIKE "%SUCCESS%" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

EDA with SQL (continued...)

The following SQL queries were performed :-

- List the names of the booster_versions which have carried the maximum payload mass.

```
%sql select DISTINCT Booster_Version  from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_)  from SPACEXTBL)
```

- 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.

```
%%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL where Landing_Outcome = "Failure (drone ship)" and substr(Date,0,5)='2015'
```

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

```
%%sql select Landing_Outcome, count(Landing_Outcome) as Count
from SPACEXTBL where Date between "2010-06-04" and "2017-03-20"
group by Landing_Outcome order by count(Landing_Outcome) desc
```

- URL - Data Viz with SQL

Build an Interactive Map with Folium

- All the launch site were labeled on Folium map and map objects such as markers, circles were used to highlight all launch sites.
- Lines between nearby mode of transport and city were drawn to get the distance.
- These markers were drawn to know success rate of each sites.
- URL - Folium Map

Build a Dashboard with Plotly Dash

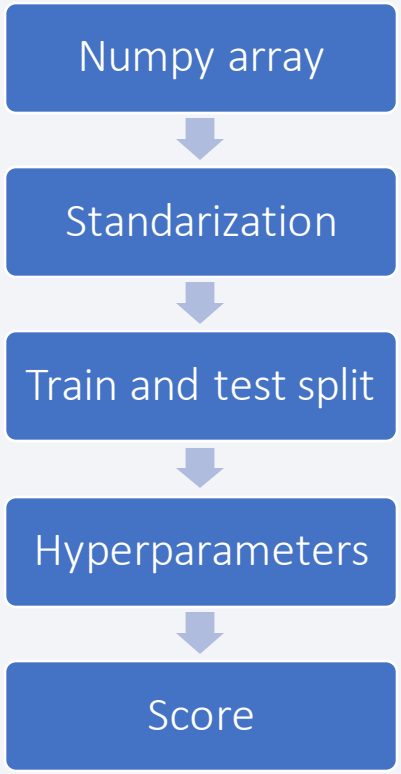
Dashboard application contains:

- Launch Site Drop-down Input Component
- A callback function to render success-pie-chart based on selected site dropdown
- Range Slider to Select Payload
- A callback function to render the success-payload-scatter-chart scatter plot
- These interactions were made to see proportion of success rate of launch sites and relationship between payload and booster version.
- URL- Dash Dashboard

Predictive Analysis (Classification)

how the best classification model was build?

- Convert dependent variable(Class) into Numpy array.
- Standardize data with StandardScaler().
- Split data in ratio 8:2.
- Train model with hyperparameters in GridSearchCV().
- Get R-square score with in-build score parameter.
- URL- ML

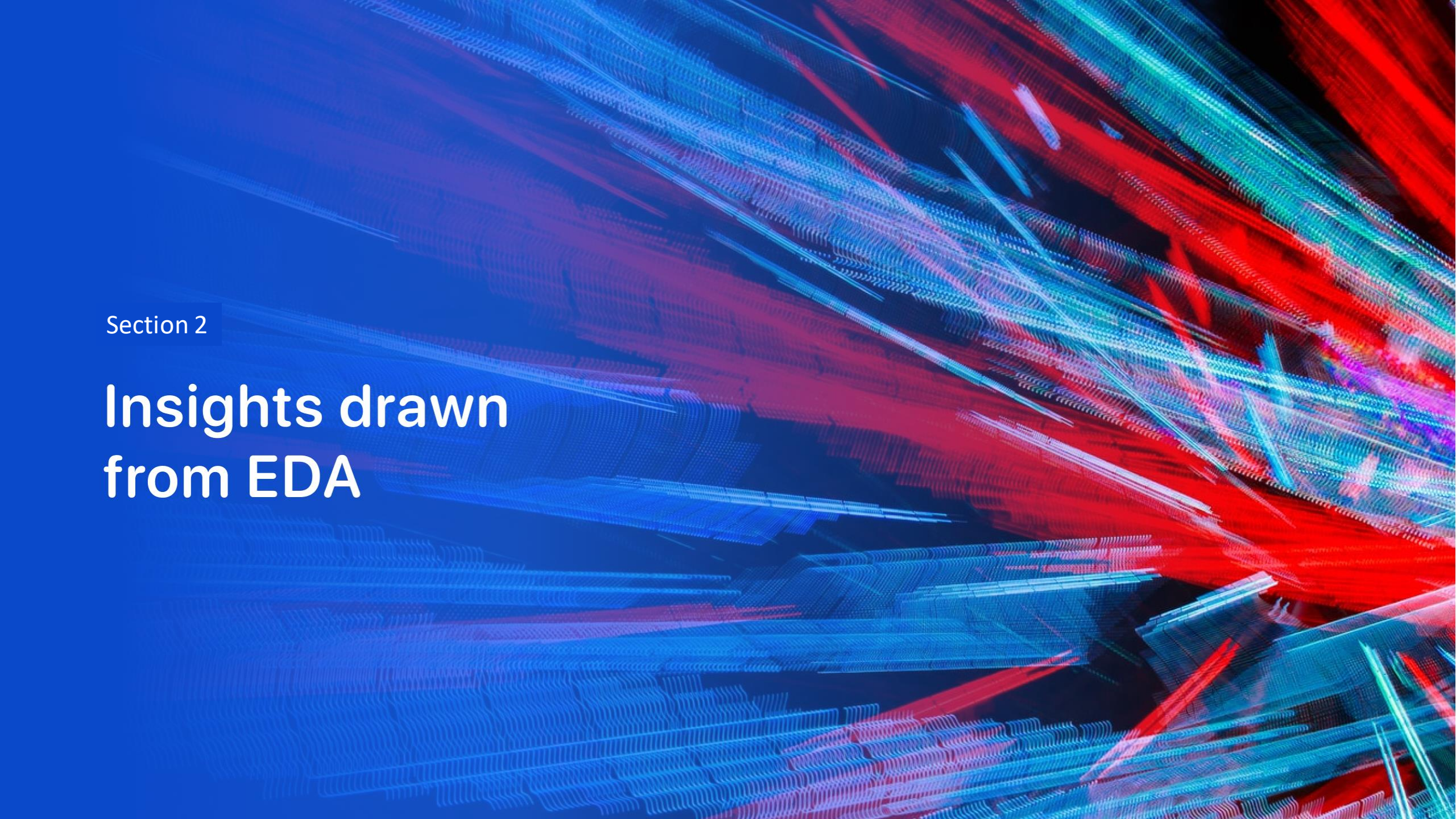


Results

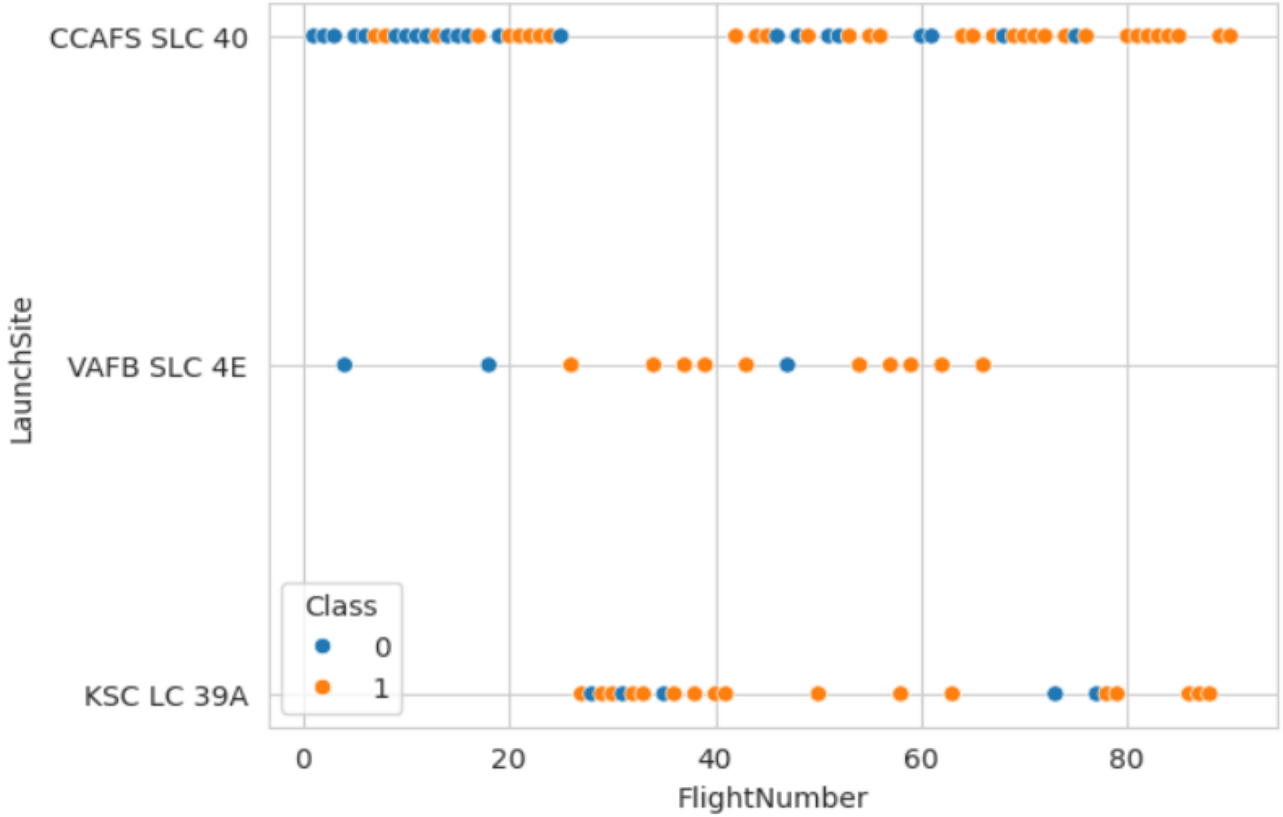
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



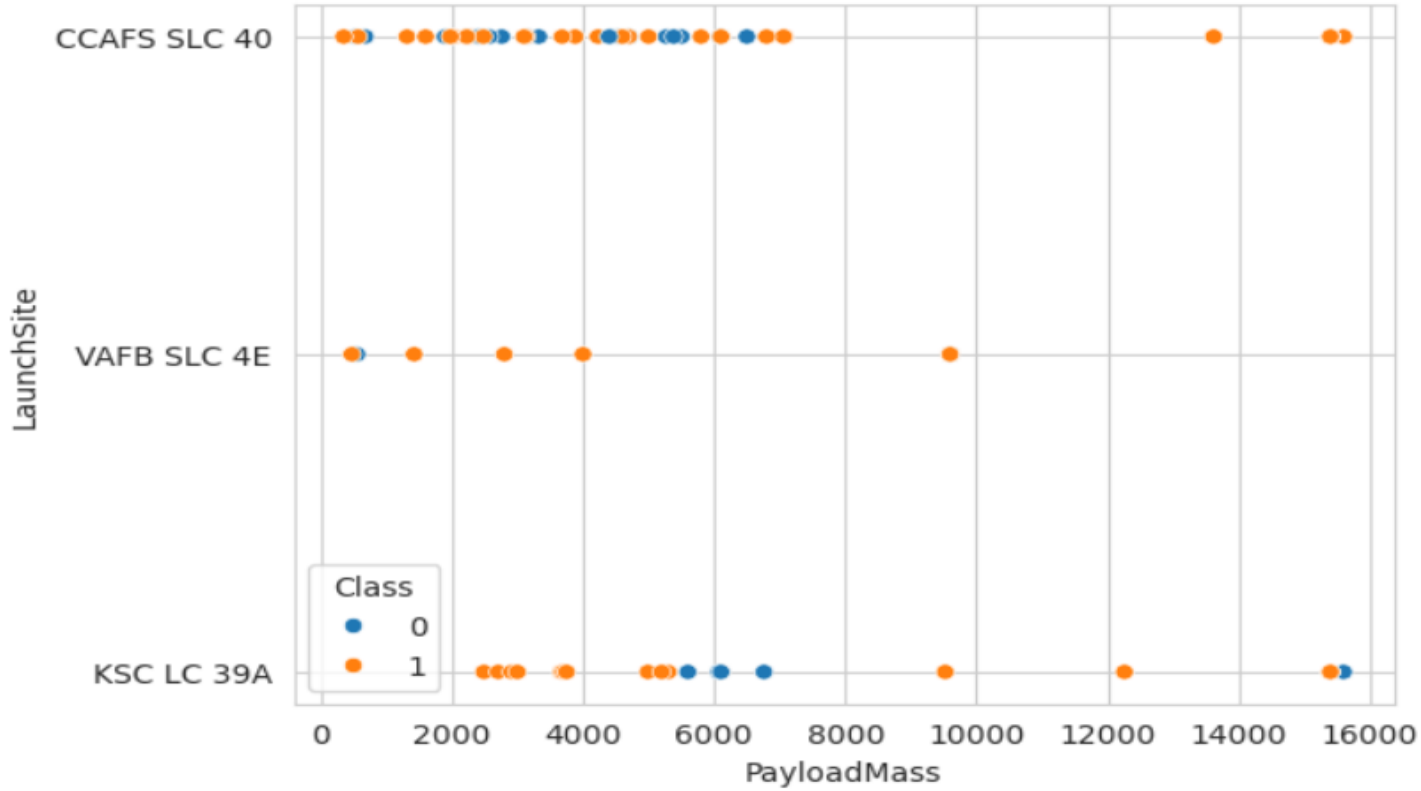
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

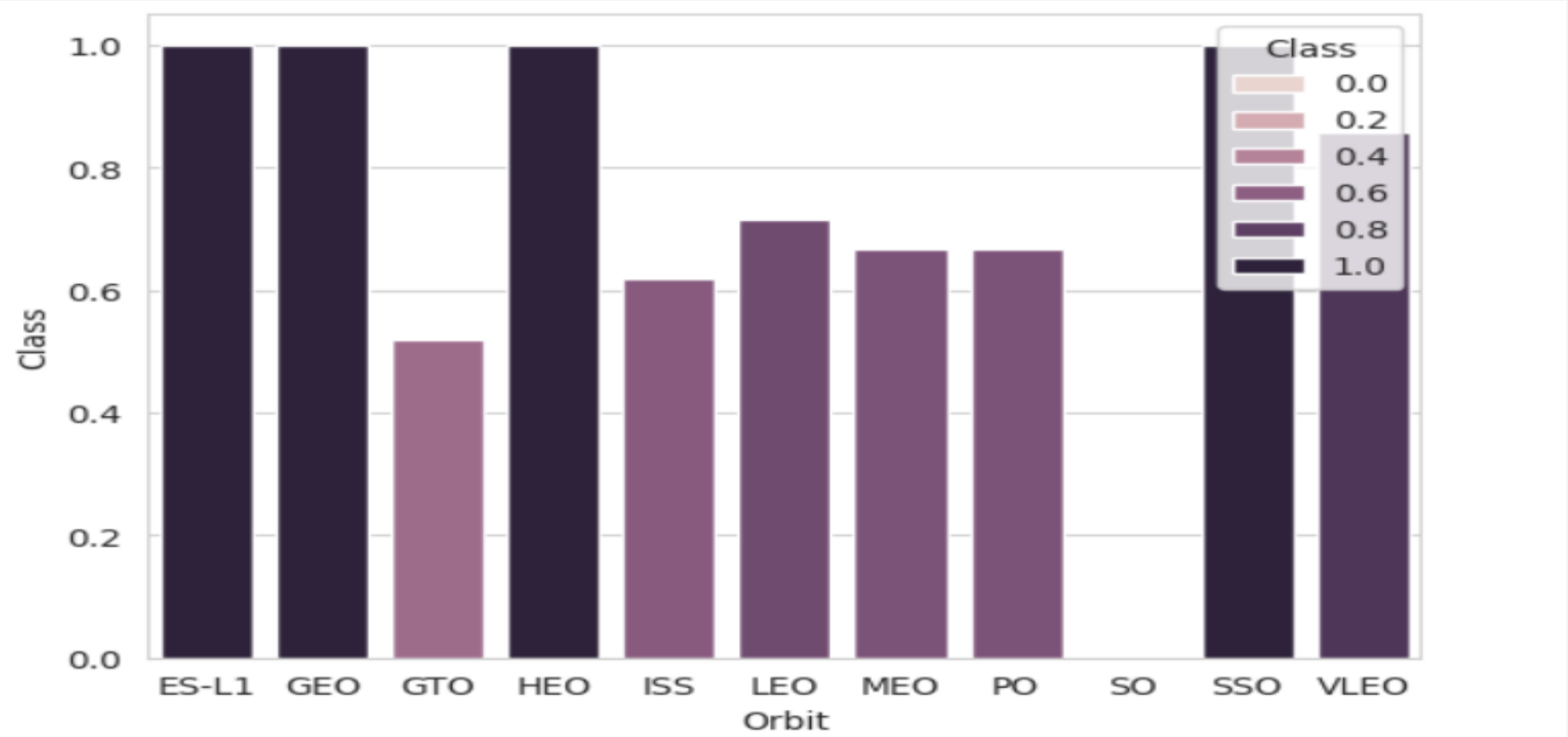
As we can see success rate of flight keeps on increasing after few iteration(Flight number).For "VAFB SLC 4E", success rate jumps to 100% after Flight number 50 and for others after Flight number 80.

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).

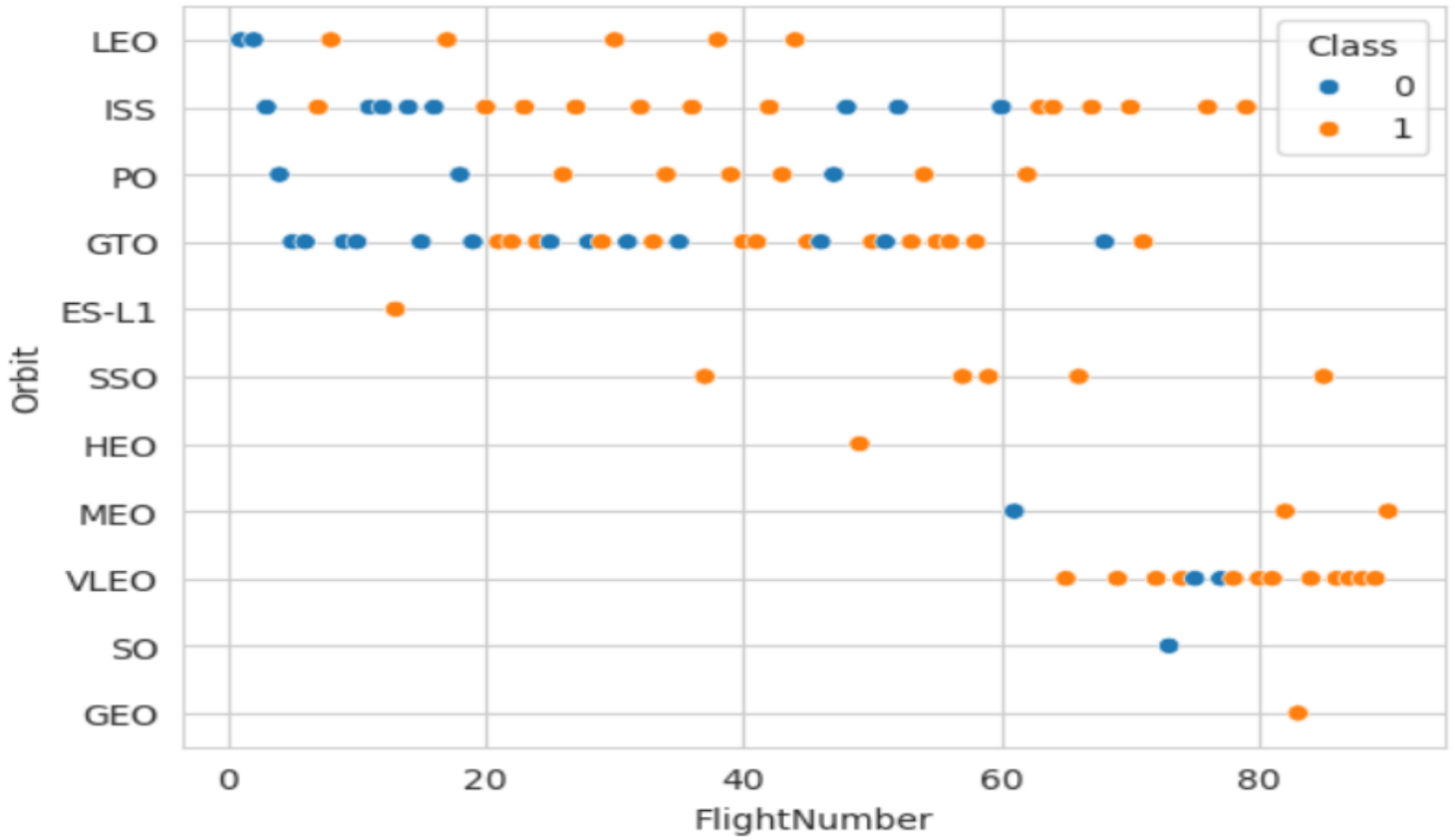
Success Rate vs. Orbit Type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

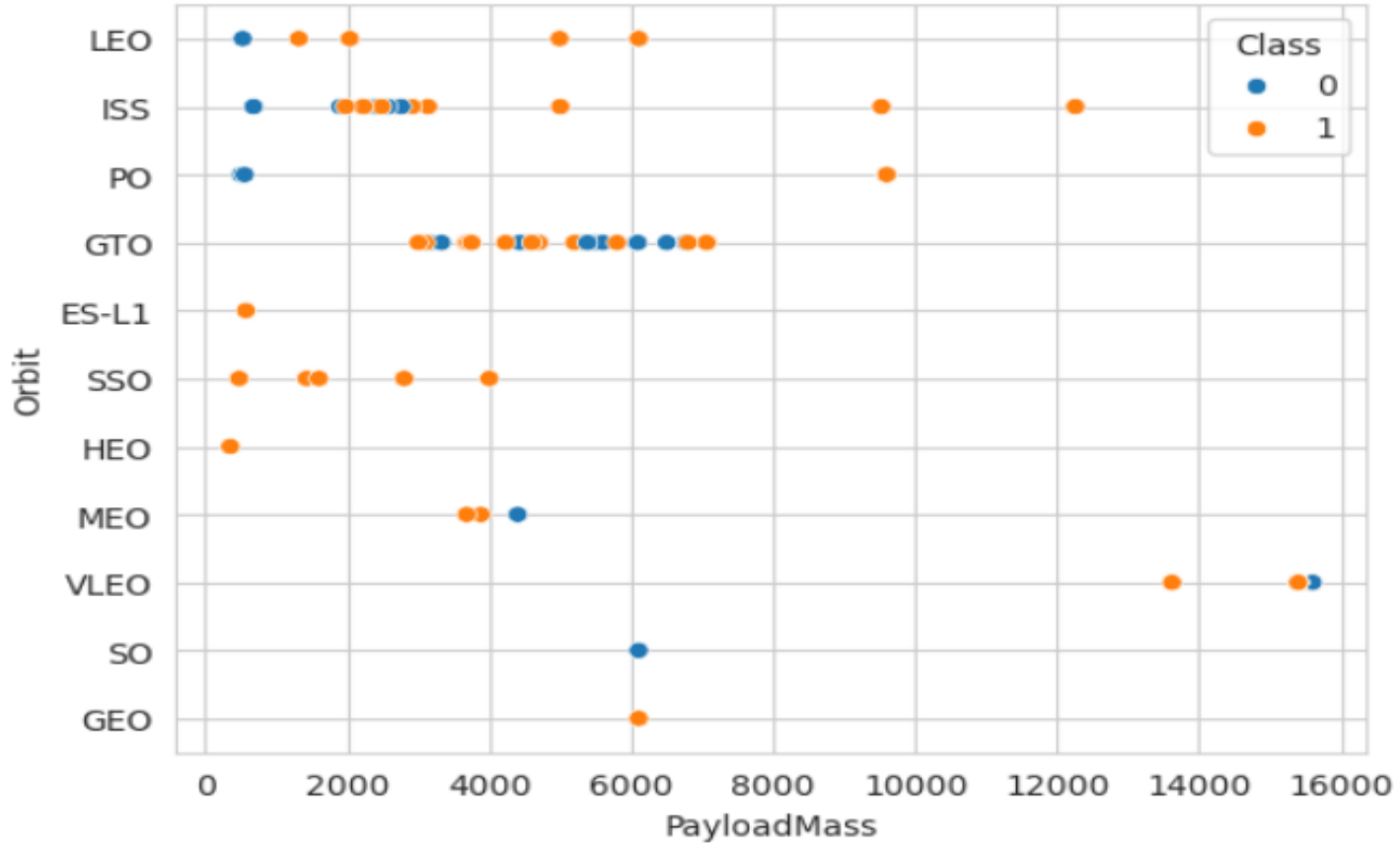
ES-L1, GEO, HEO, SSO, VLEO orbits have 100% success rate whereas SO orbit has 0% success rate.

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.

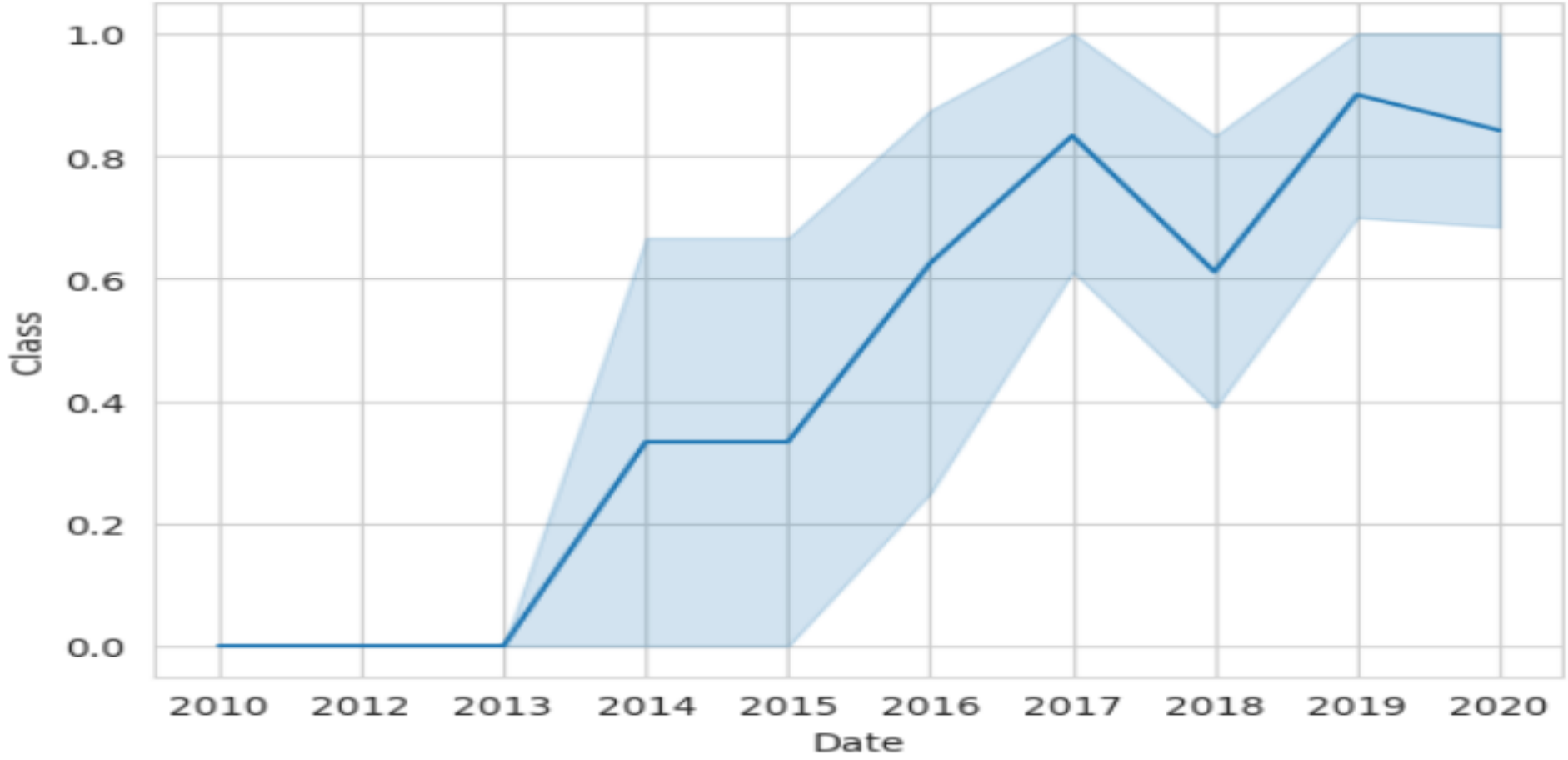
Payload vs. Orbit Type



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

Task 1

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

[11]:
```
%sql select distinct "Launch_Site"  from SPACEXTBL
```

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

[11]:

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

Distinct keyword is used to select only unique values from output

Launch Site Names Begin with 'CCA'

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

```
%sql select *  from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

"LIKE" keyword used for finding regular expression and "CCA%" means finding all matching words starting with CCA.

Total Payload Mass

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

```
%sql select SUM(PAYLOAD_MASS__KG_)  from SPACEXTBL where "Customer" = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

SUM() function sum-up all the values in "Payload_mass_kg" column.

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_)  from SPACEXTBL where "Booster_Version" LIKE "F9 v1.1%"
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

AVG() function gives mean of values in "PAYLOAD_MASS_KG" column

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select MIN(Date)  from SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

MIN(Date)
2015-12-22

MIN() function select minimum date within output

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select  Booster_Version from SPACEXTBL WHERE "LANDING_OUTCOME" LIKE "%SUCCESS%" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

AND keyword used as logical operation which means true only if both side of operator are true.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) as NO_Of_Outcomes FROM SPACEXTBL  GROUP BY "Landing_Outcome"
```

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

Landing_Outcome	NO_Of_Outcomes
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

"Group by" keyword group output according to selected column and "Count" keyword works jointly with groupby.

Boosters Carried Maximum Payload

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

```
%sql select DISTINCT Booster_Version  from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_)  from SPACEXTBL)
```

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

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

First inner query run and output max payload then based on max payload outer query was filtered.

2015 Launch Records

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL
where Landing_Outcome = "Failure (drone ship)" and substr(Date,0,5)='2015'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	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

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.

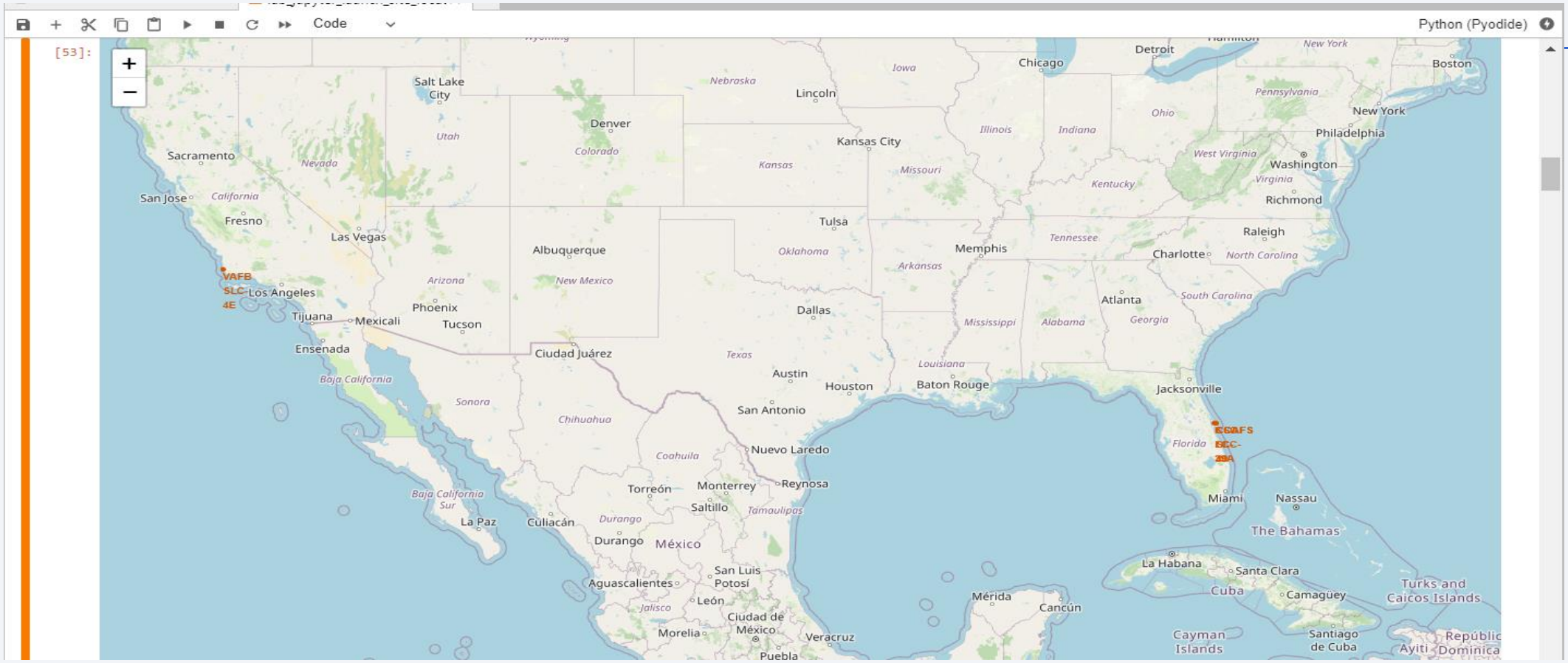
```
%%sql select Landing_Outcome, count(Landing_Outcome) as Count
from SPACEXTBL where Date between "2010-06-04" and "2017-03-20"
group by Landing_Outcome order by count(Landing_Outcome) desc
```

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

Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

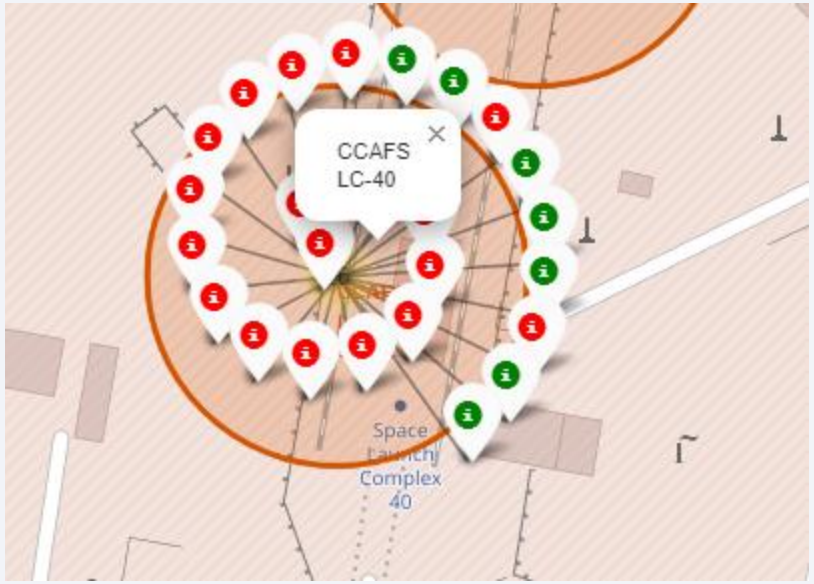
Markers of all launch sites on global map

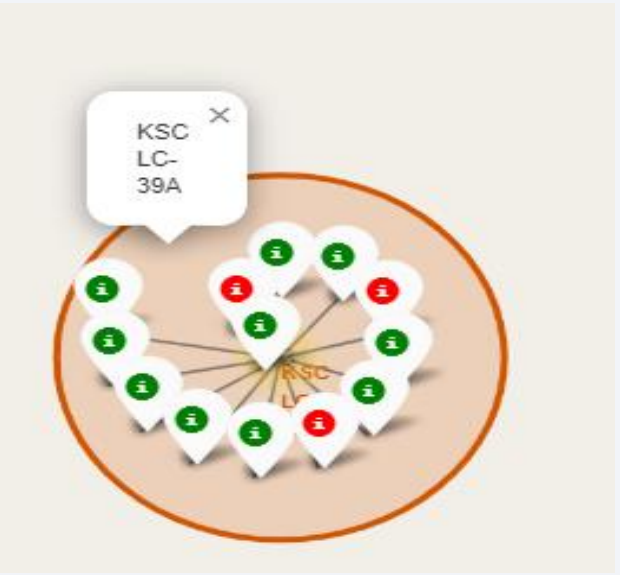


All launch sites are in proximity to the Equator, (located southwards of the US map). Also all the launch sites are in very close proximity to the coast.

Launch outcomes for each site on the map with color markers

Florida Sites

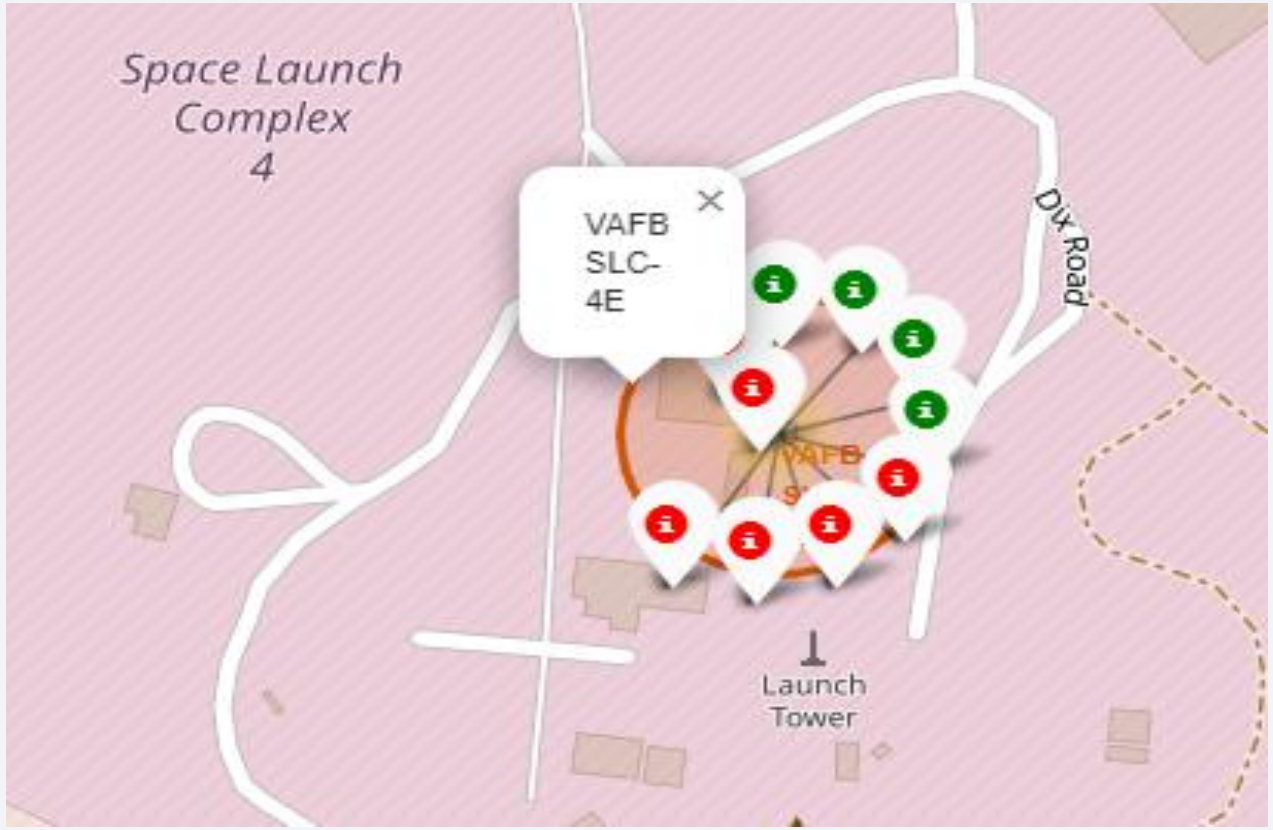




- In the Eastern coast (Florida) Launch site KSC LC-39A has relatively high success rates compared to CCAFS SLC-40 & CCAFS LC-40.

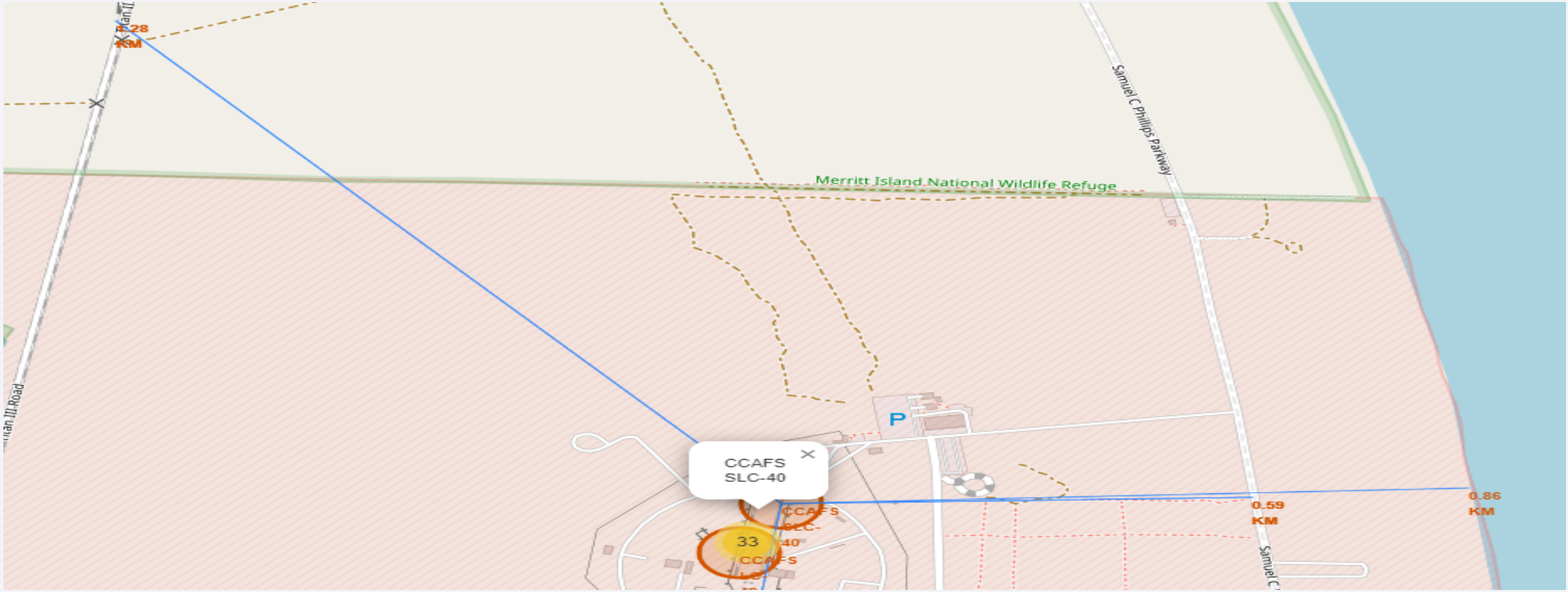
Launch outcomes for each site on the map with color markers

West Coast/ California



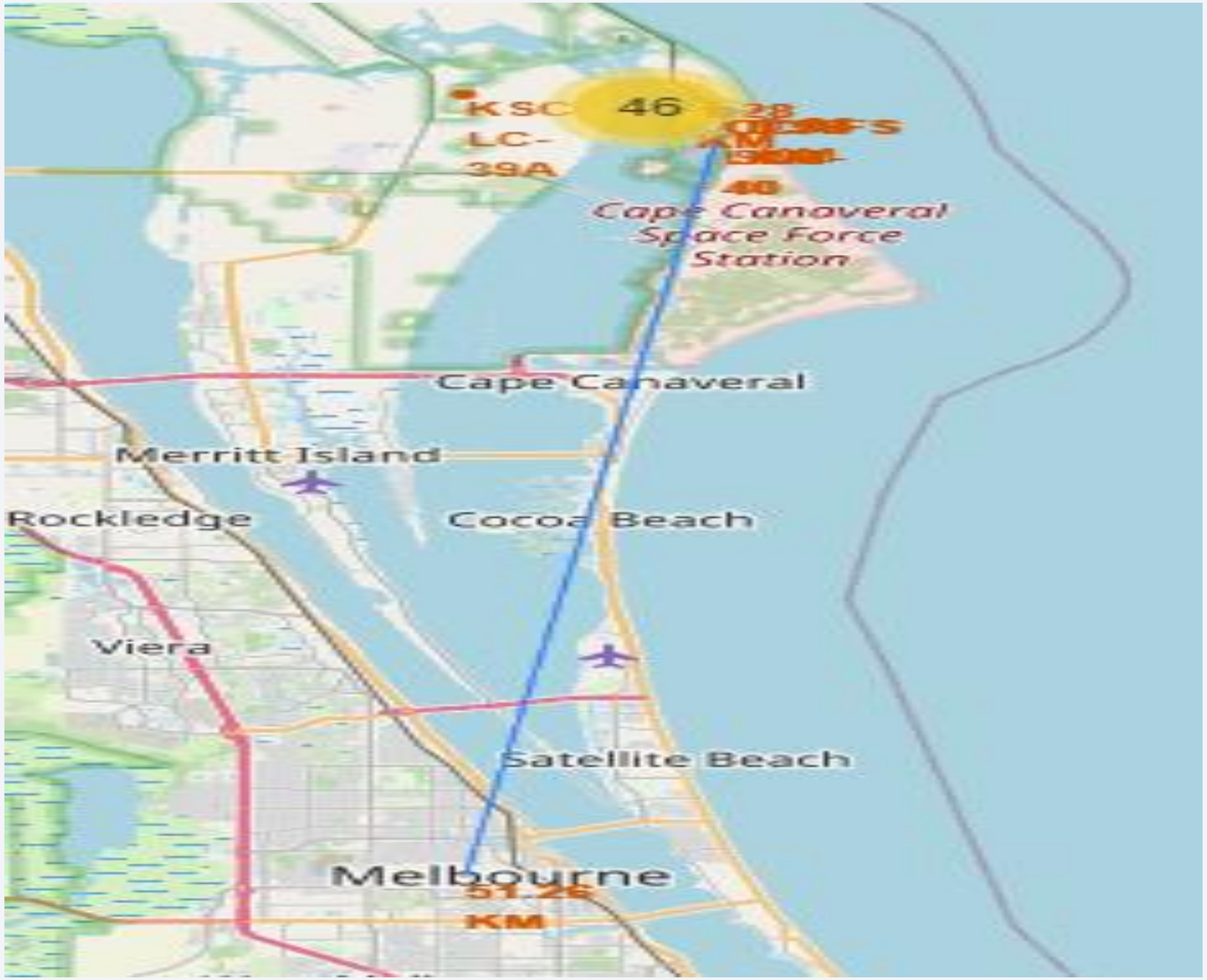
- In the West Coast (California) Launch site VAFB SLC-4E has relatively lower success rates 4/10 compared to KSC LC-39A launch site in the Eastern Coast of Florida.

Distances between a launch site to its proximities

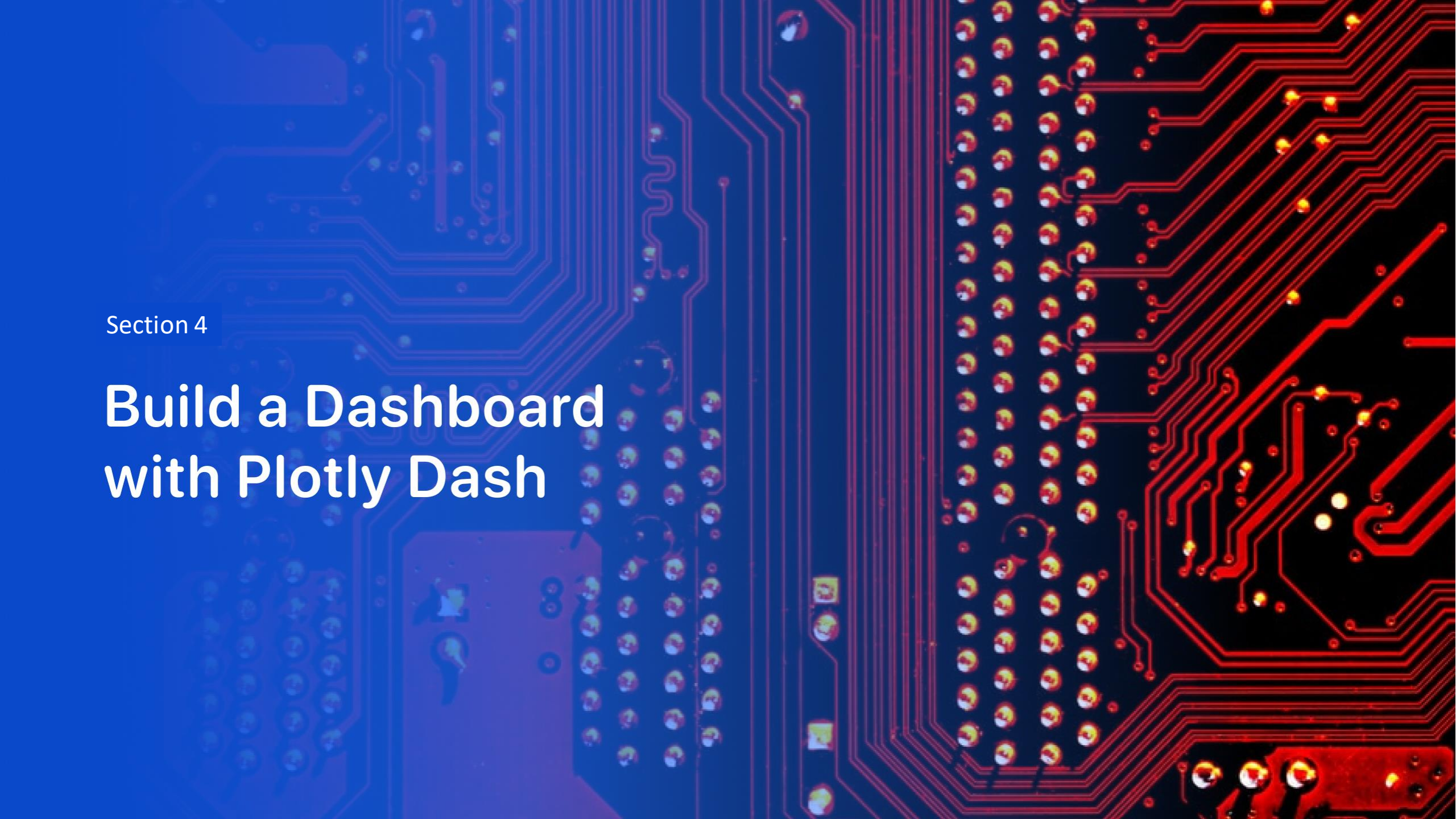


- Launch site CCAFS SLC-40 proximity to coastline is 0.86 km

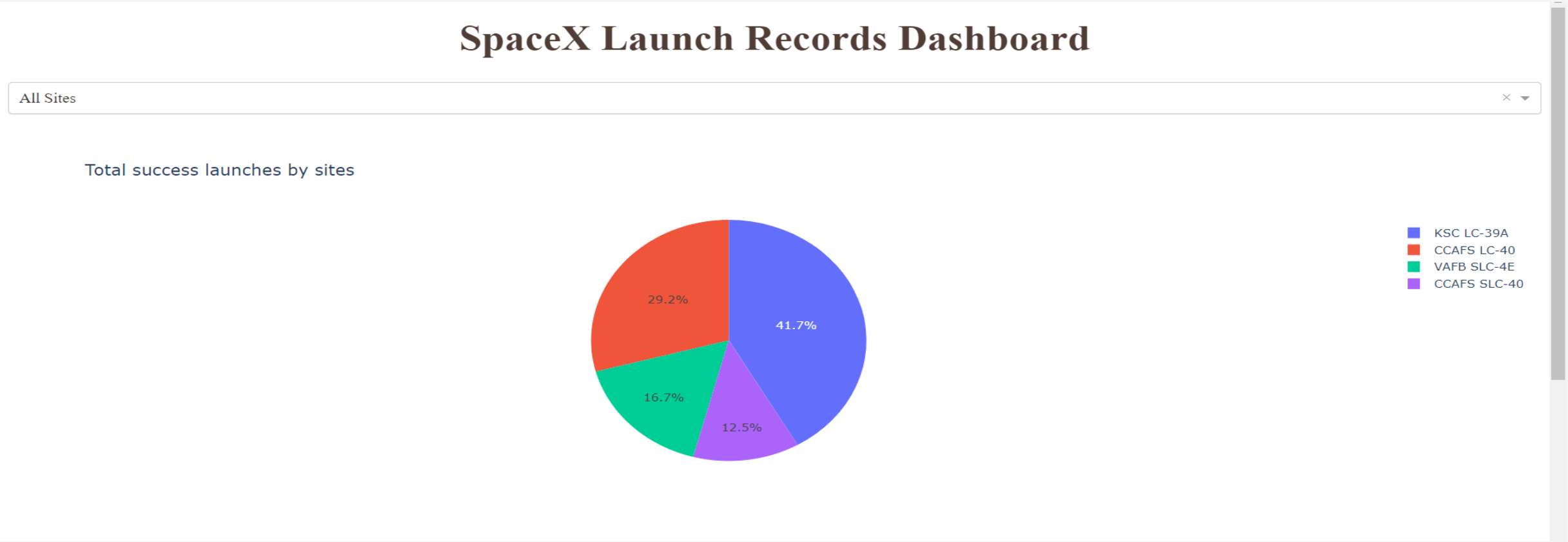
Distances between a launch site to its proximities



- Launch site CCAFS SLC-40 closest to Melbourne City is 51.26 km

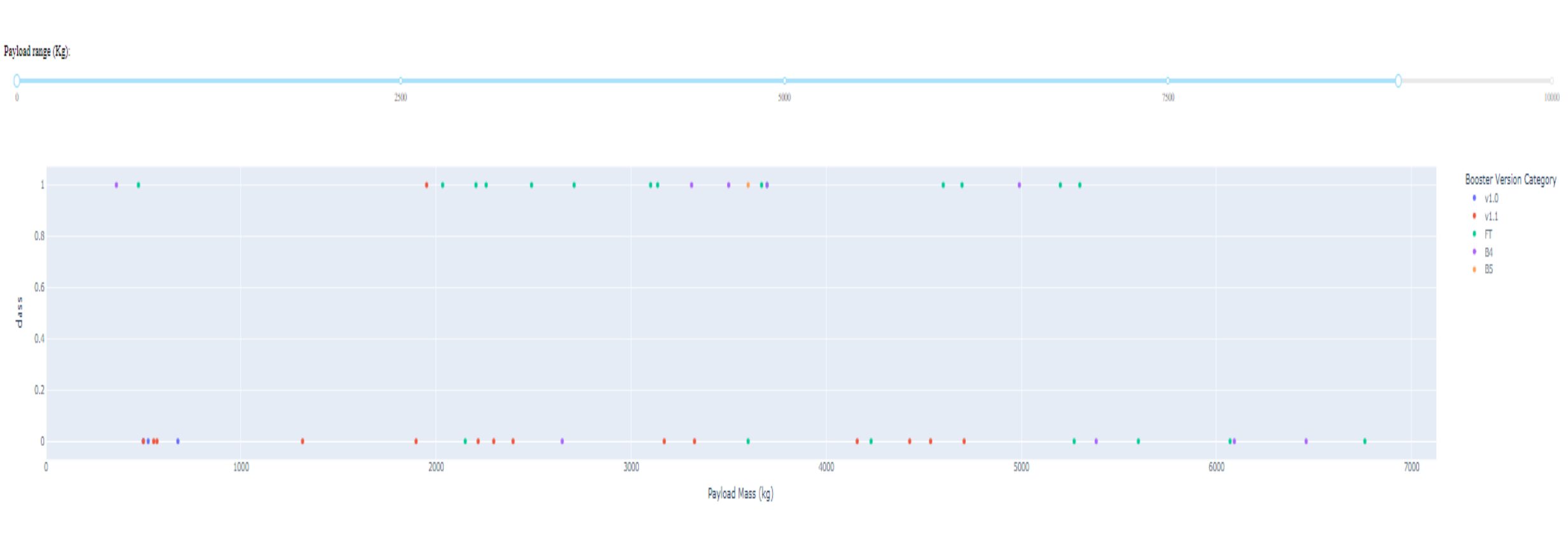

Section 4
Build a Dashboard with Plotly Dash

Pie-Chart for launch success count for all sites



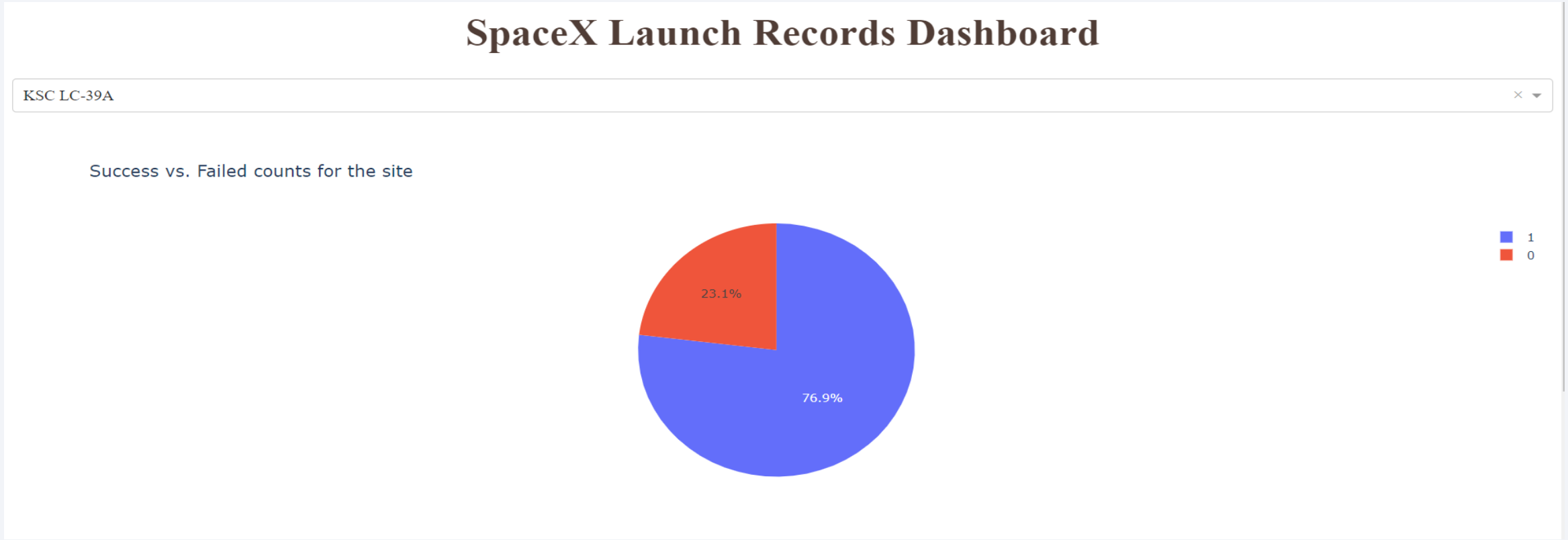
Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%.

Payload vs. Launch Outcome scatter plot for all sites



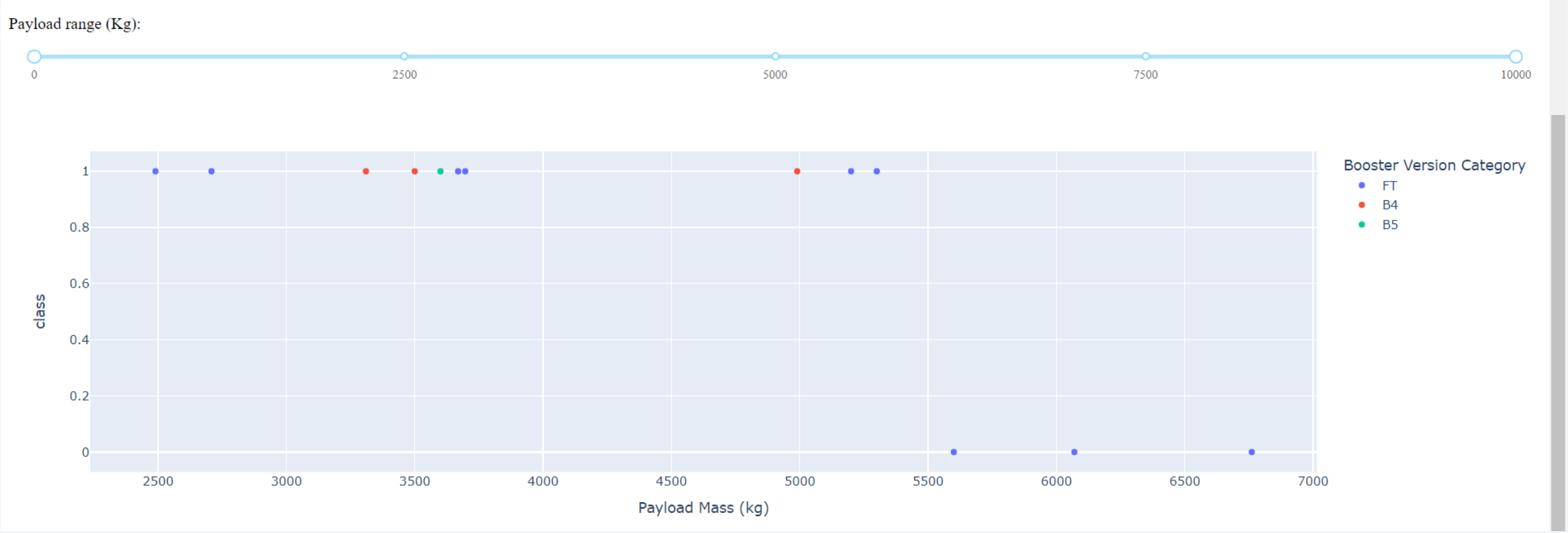
- As we can see success rate drop to zero after 5500Kg payload.

Pie-Chart for the launch site with highest launch site ratio.



Launch site KSC LC-39A had the highest success ratio of 77% success against 23% failed launches.

Scatter plot of the highest success rate launch site



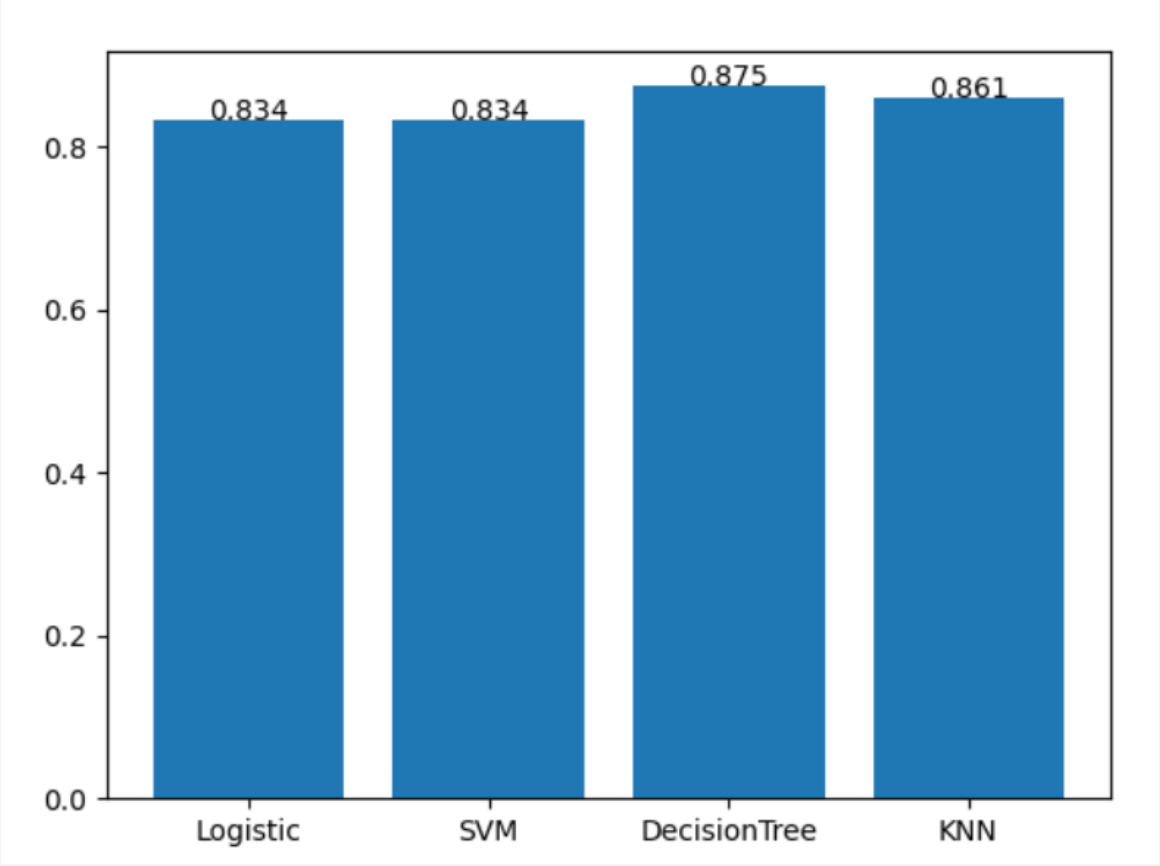
- Ratio of success rate is much higher than failure and all success rate falls below 5500Kg of payload.

Section 5

Predictive Analysis (Classification)

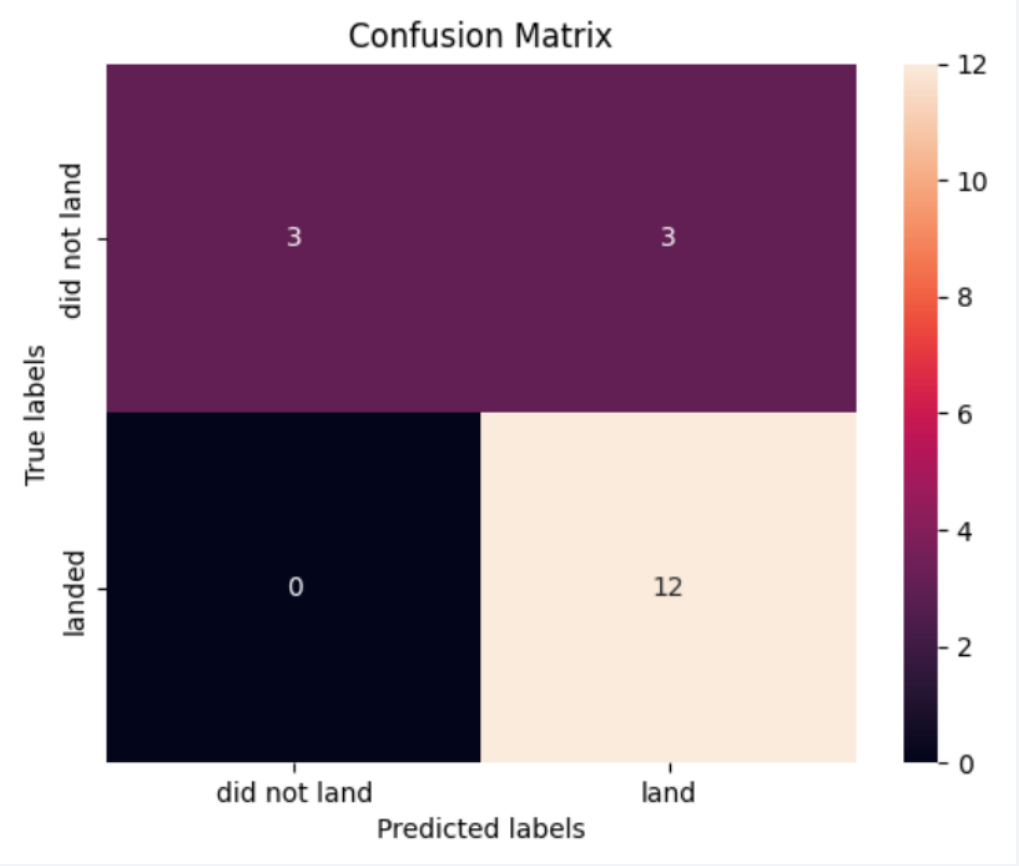
Classification Accuracy

- All the models have almost same accuracy.
- Decision tree has highest among them with 87%.



Confusion Matrix

- All the models have same values for confusion matrix.
- Value of false positive is less but could be reduced for more better landing of booster.



Conclusions

- 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%.
- We can deduce that, as the flight number increases in each of the 3 launch sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight
- If you observe 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).
- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.
- 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

Conclusions cont.

- 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 the same.
- Finally the success rate since 2013 kept increasing till 2020.

Thank you!