Winning Space Race with Data Science

Shibin Philip
17 March 2023

Outline

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

Executive Summary

Summary of methodologies

As part of this research, I have used multiple methodologies to prepare data and analyze the previous launches data from SpaceX. Data collected using API and web scraping methods. Then started the analysis by conducting exploratory data analysis to gain a better understanding of our data and identify any patterns or outliers. Next, we used regression analysis to explore the relationship between launch success rate and related variables. Created interactive maps using folium to visualize the the geographic elements. And finally used predictive analysis to prepare multiple classification models.

Summary of all results

After evaluating the performance of each model, found that Logistic Regression method and Support Vector Machine method gives the best accuracy of 83%. Another key finding was that over the years success rate is increasing marginally and gives best results with high payload mass.

Introduction

Project background and context

This project's primary objective is to find out the success rate of a Falcon 9 launches based on various parameters used in previous rocket launch experiments. It will help to determine best possible factors for a successful mission. This will also help to minimize the cost for future launches.

Problems we want to find answers

- Identify the factors associated with rocket launch
- Impact of each factors for a successful launch
- Based on the parameters, what will be the success rate for a launch attempt.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data collected from SpaceX website and Wikipedia pages.

  - REST API connections used to get data from SpaceX website
  - Web scraping technique used to fetch data from Wikipedia page

- Perform data wrangling

  - Irrelevant columns are removed from dataset
  - Used One Hot Encoding to convert data for machine learning.

- Perform exploratory data analysis (EDA) using visualization and SQL

  - Bar Graphs and Scatter plots are used for identifying relation between various factors
  - Various SQL queries performed for finding Data points

Methodology

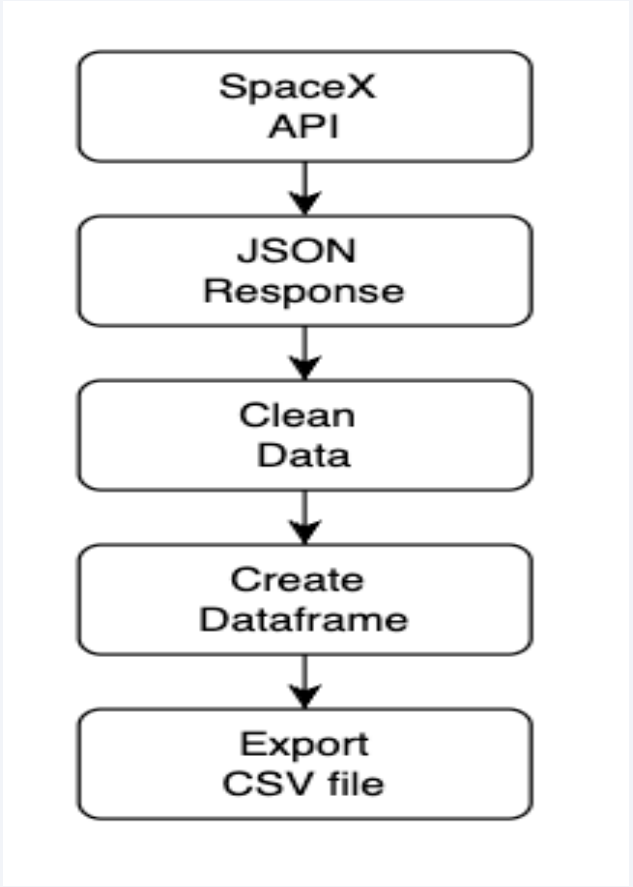
- Perform interactive visual analytics using Folium and Plotly Dash
  - Used Folium for visualizing interactive Geographical factors
  - Used Plotly Dash to visualize interactive charts for analytics
- Perform predictive analysis using classification models
  - These classification models are used for the predictive analysis
    - Logistics Regression
    - Support Vector Machine
    - Decision tree
    - K nearest neighbours

Data Collection

Data Collection is the preliminary step in any data science project. For this project, identified that required data is available in SpaceX website and Wikipedia pages. Used below methods to collect the required data.

- REST API connections used to get data from SpaceX website
- Web scraping technique used to fetch data from Wikipedia page

Data Collection – SpaceX API



SpaceX API

End points

https://api.spacexdata.com/v4/rockets/

https://api.spacexdata.com/v4/launchpads

https://api.spacexdata.com/v4/payloads

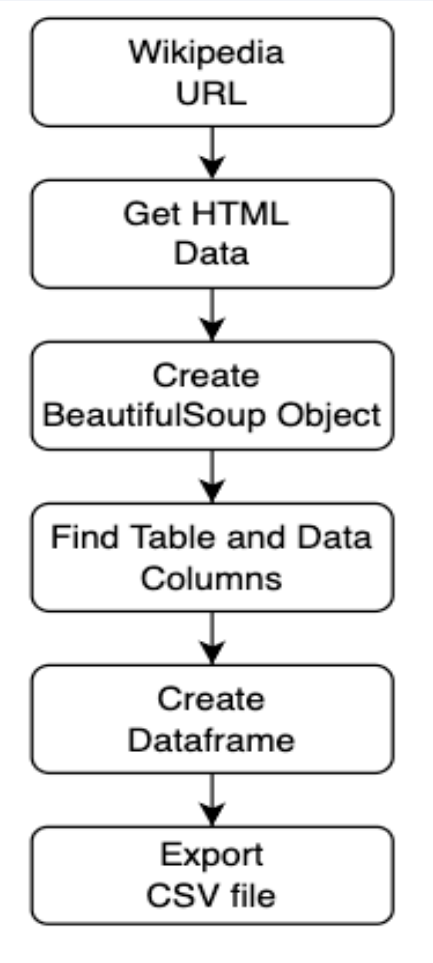
https://api.spacexdata.com/v4/cores/

Data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

GitHub URL : https://github.com/shbn/DS-project/blob/master/spacex-data-collection.ipynb

Data Collection - Scraping



Wikipedia URL

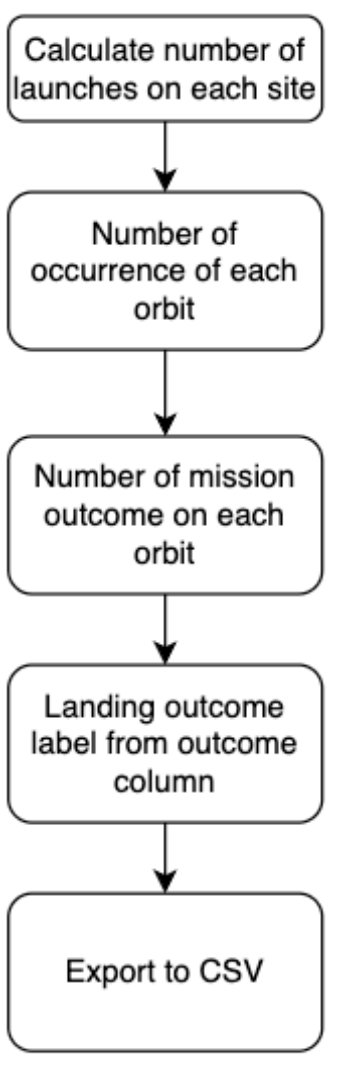
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	[[SpaceX], \n]	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	[[.mw-parser-output .plainlist ol,.mw-parser-o...	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	[[NASA], (, [COTS],)\n]	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

GitHub URL : https://github.com/shbn/DS-project/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

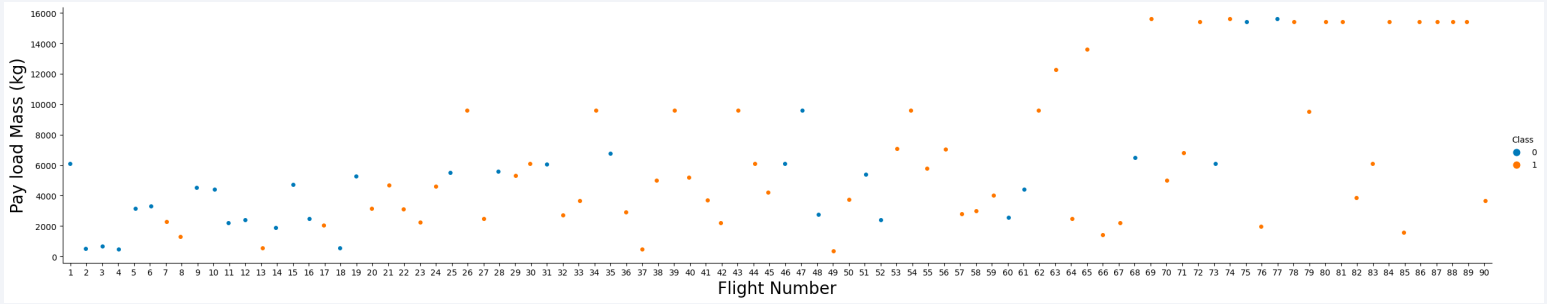


Final Data

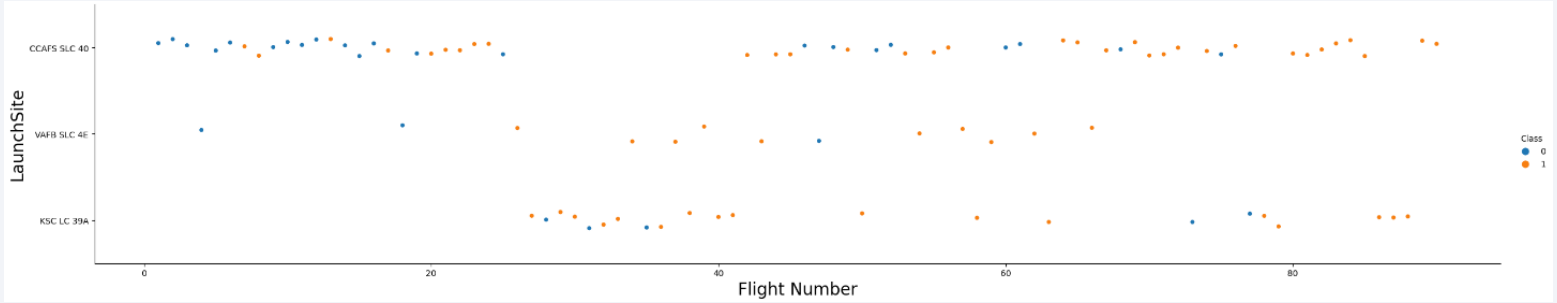
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

GitHub URL : https://github.com/shbn/DS-project/blob/master/dataWranging.ipynb

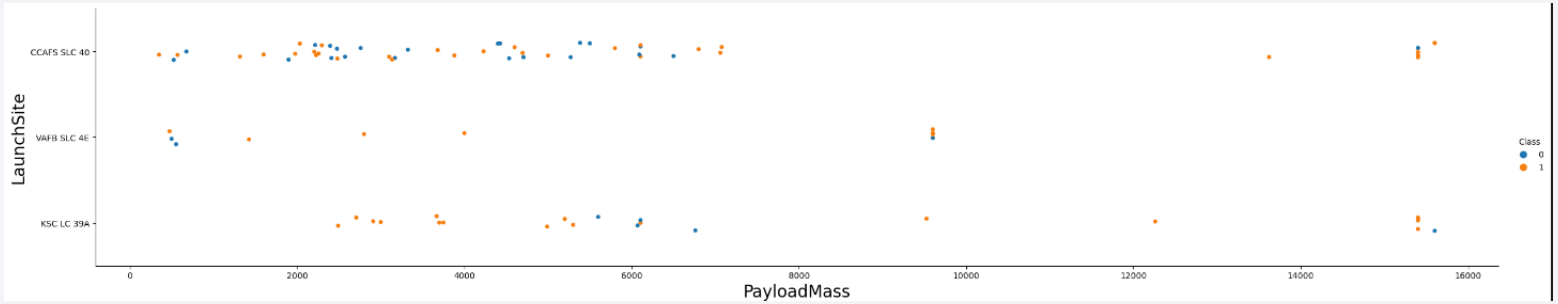
EDA with Data Visualization



Scatter plot for finding relation between Flight number and Payload



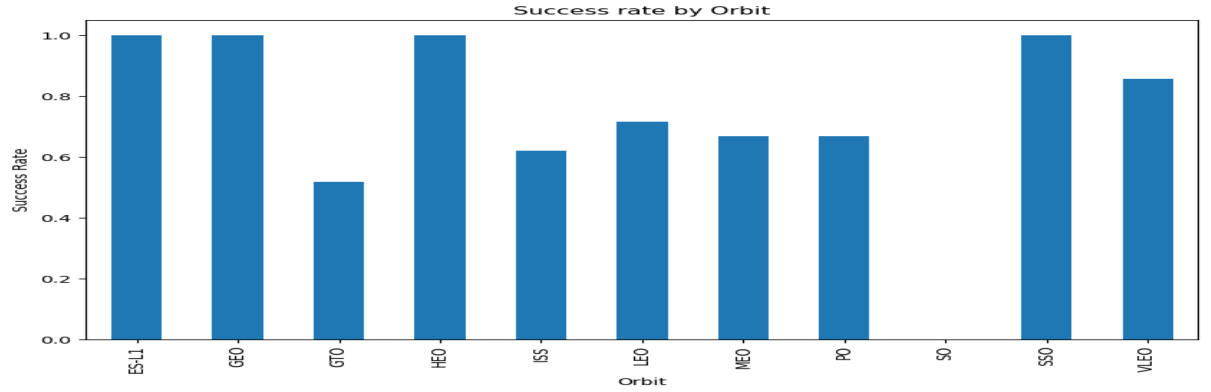
Scatter plot for finding relation between Flight number and Launch site



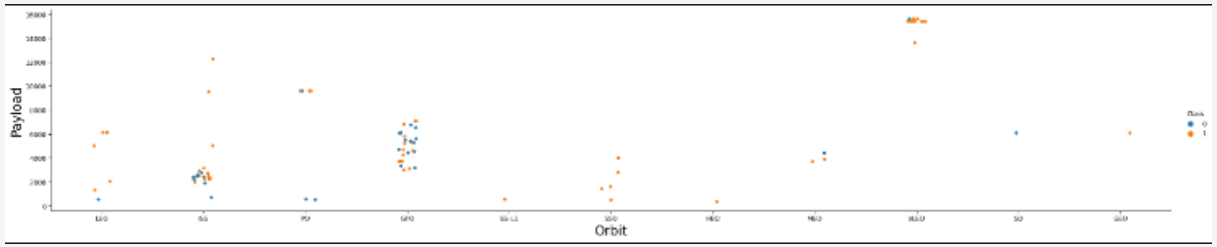
Scatter plot for finding relation between Payload mass and Launch site

GitHub URL : https://github.com/shbn/DS-project/blob/master/EDA-dataviz.ipynb

EDA with Data Visualization



Bar graph for finding success rate by orbit



Scatter plot for finding relation between Payload and Orbit

Line graph to find success rate by year

GitHub URL : https://github.com/shbn/DS-project/blob/master/EDA-dataviz.ipynb

EDA with SQL

Below SQL queries you performed on the database

- 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL : https://github.com/shbn/DS-project/blob/master/EDA-SQL.ipynb

Interactive Map with Folium

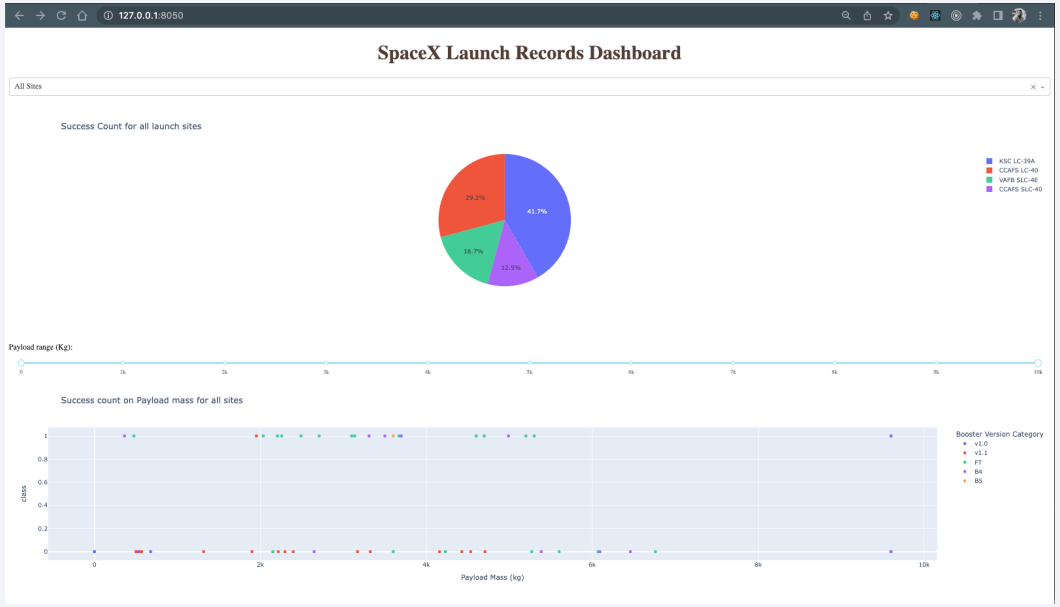
Created these map objects to visualize the launch sites using folium maps

- Map marker
- Icon Marker
- Circle Marker
- Polyline
- Marker Cluster
- AntPath

These are helpful to see the site locations and interact with the map to find various geographical relations.

GitHub URL : https://github.com/shbn/DS-project/blob/master/lab_jupyter_launch_site_location.ipynb

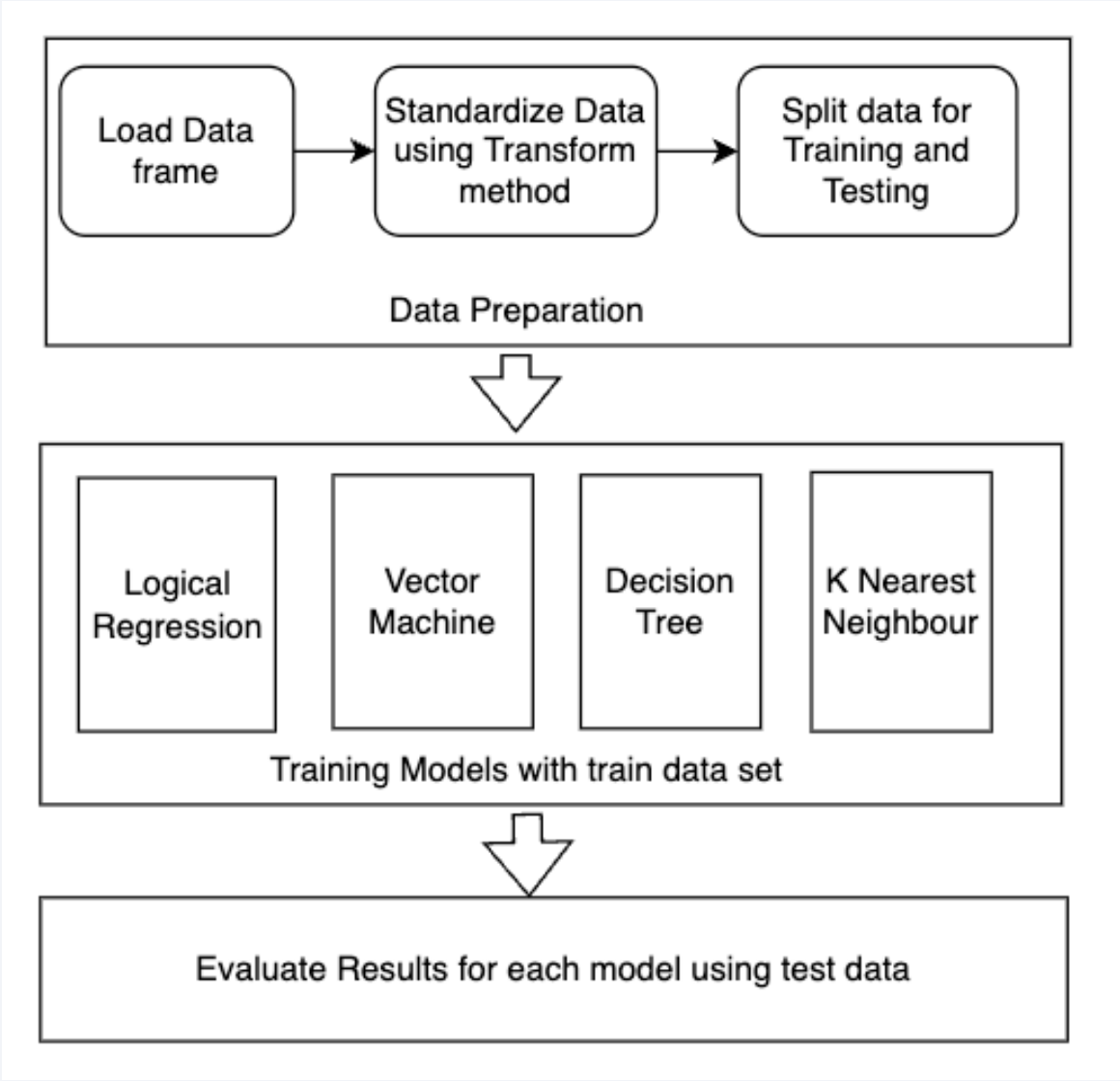
Build a Dashboard with Plotly Dash



An interactive dashboard created using Plotly for showing the Success rate for each launch site. Used Pie chart and Scatter plots to display it effectively. User can select different sites to see result for the selected site

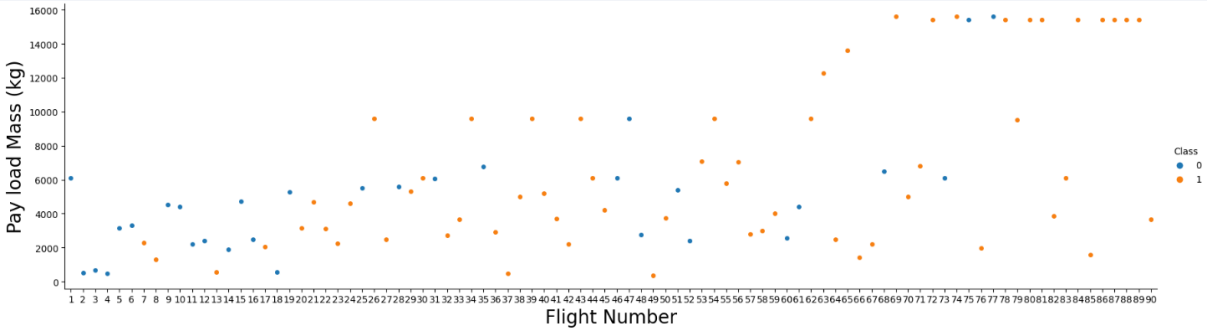
GitHub URL : https://github.com/shbn/DS-project/blob/master/plotlyDash.py

Predictive Analysis (Classification)



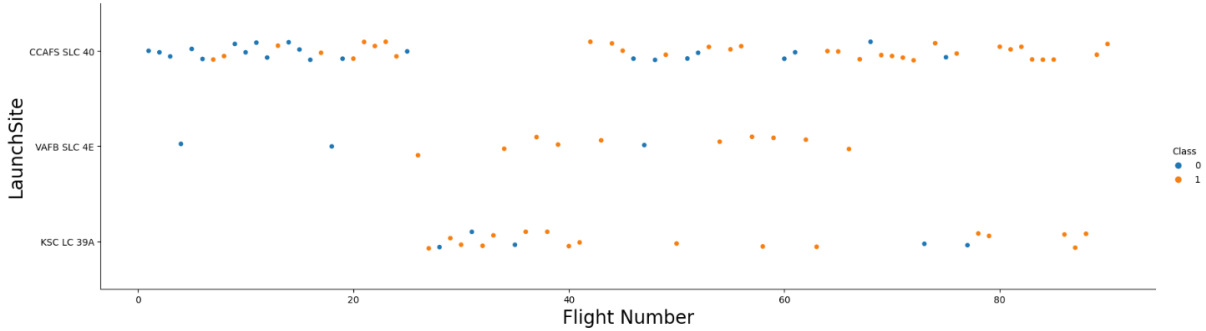
GitHub URL:
https://github.com/shbn/DS-project/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results: Exploratory data analysis



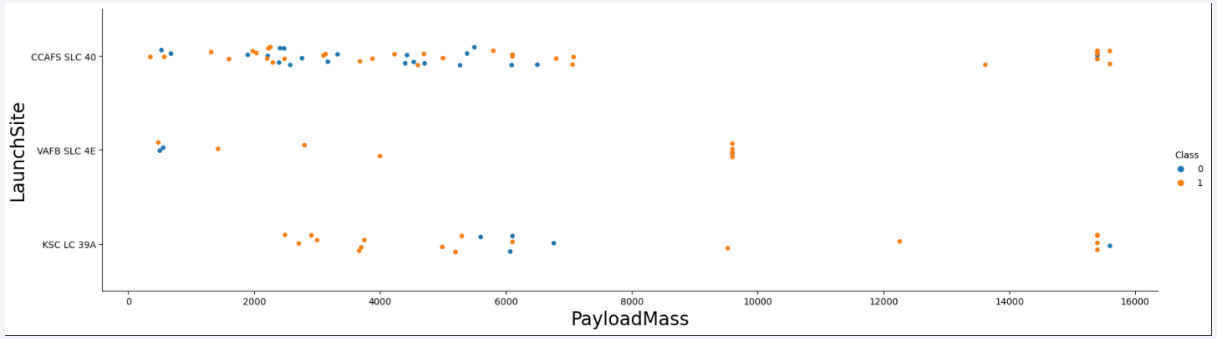
Flight Number Vs Payload Mass

- Success rates are high when Flight numbers are more than 75.
- Payload mass also gives better results with high payload mass



Flight Number Vs Launch site

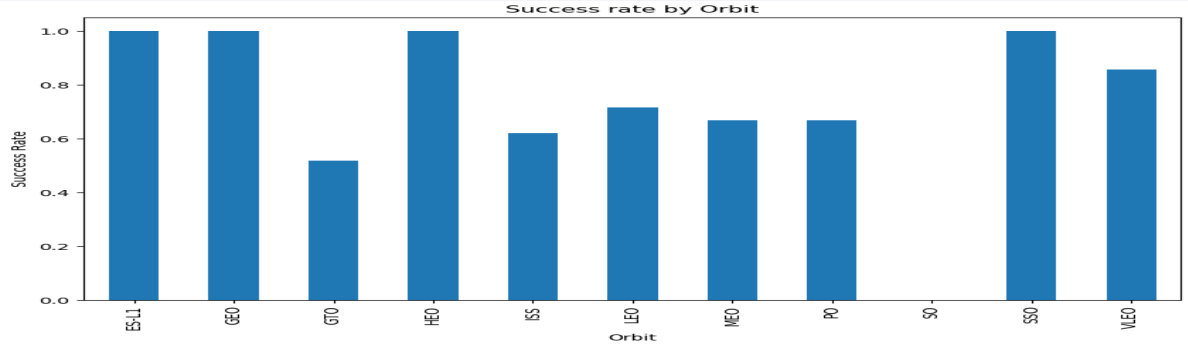
- Success rates are high when Flight numbers are more than 75.
- Results are mixed for Launch sites with lesser flight number

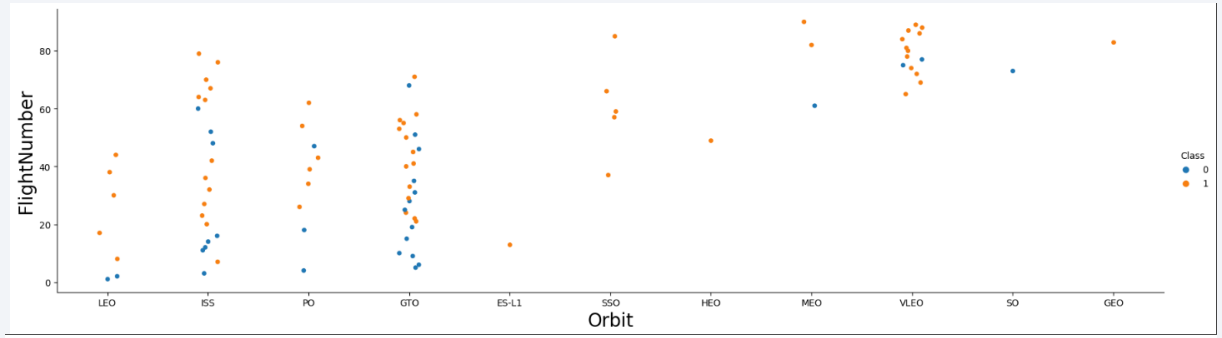


Payload mass Vs Launch site

- High success rate for payload mass between 7000 and 15000. But number of trials are comparatively less in this range

Results: Exploratory data analysis





Success Rate by Orbit

- Rocket launches to ES-L1, GEO, HEO, SSO orbits have a high success rate

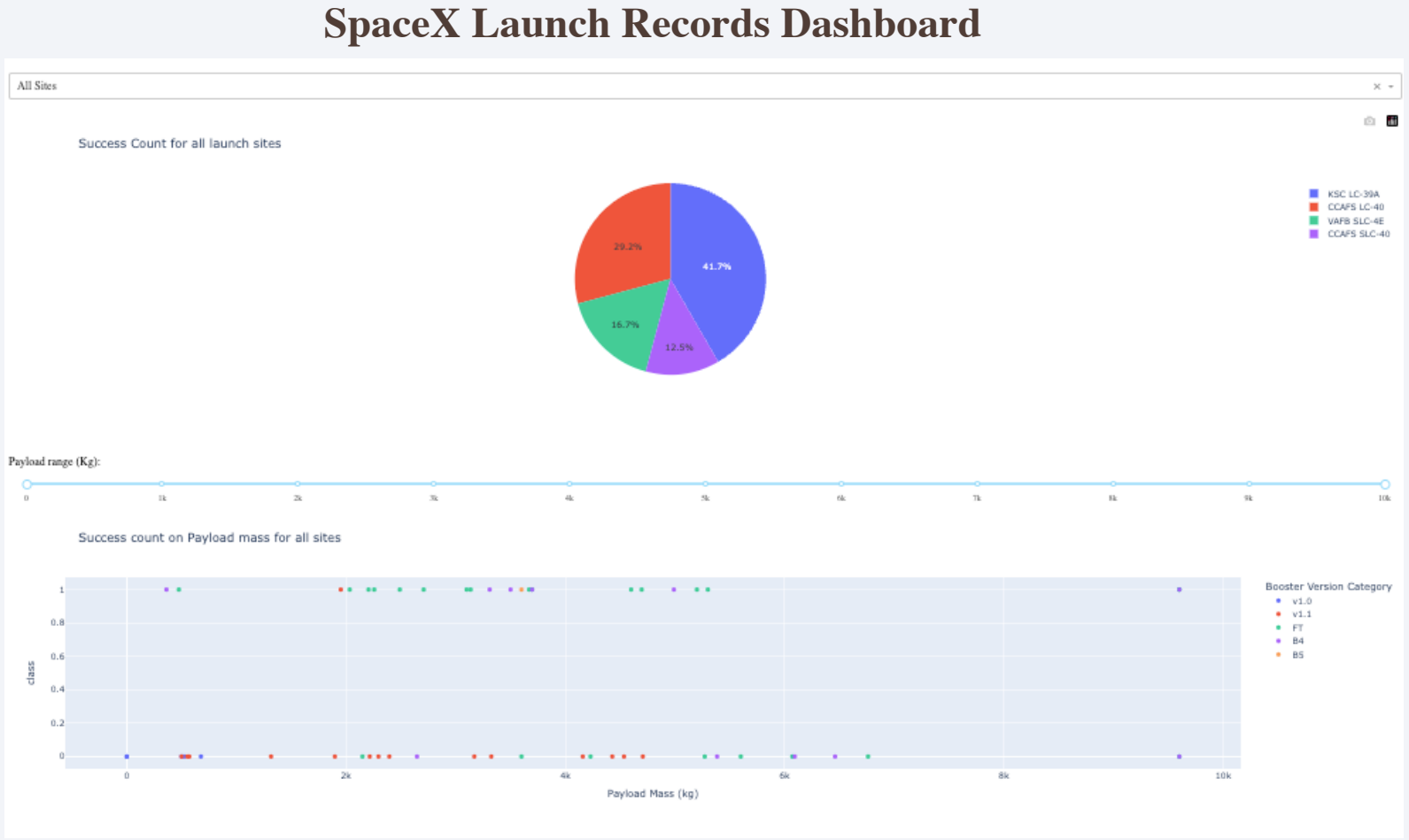
Flight Number Vs Orbit

- Success rates are high when Flight numbers are more than 75 especially for ISS, VLEO, MEO Orbits.
- Success rates are low with less flight numbers for LEO, ISS, PO and GTO orbits.

Success Rate by Year

- Success rate since 2013 kept increasing till 2020

Results: Interactive analytics

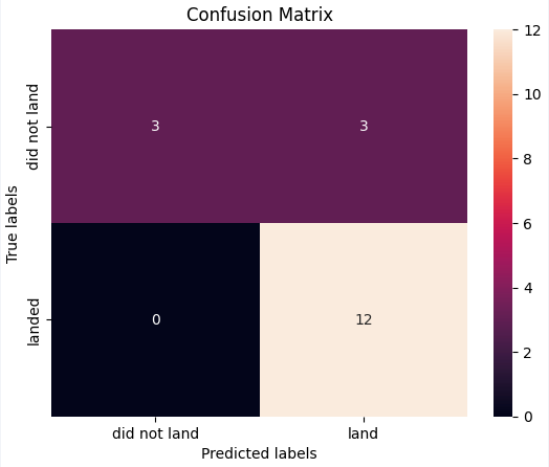


Success Rate for all Sites

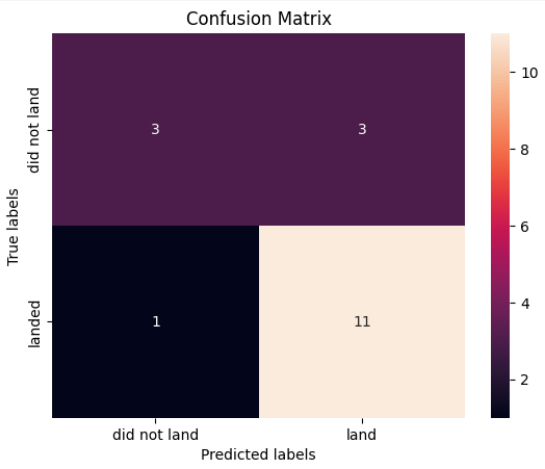
Results: Predictive analysis results

Confusion metrics for each classification models

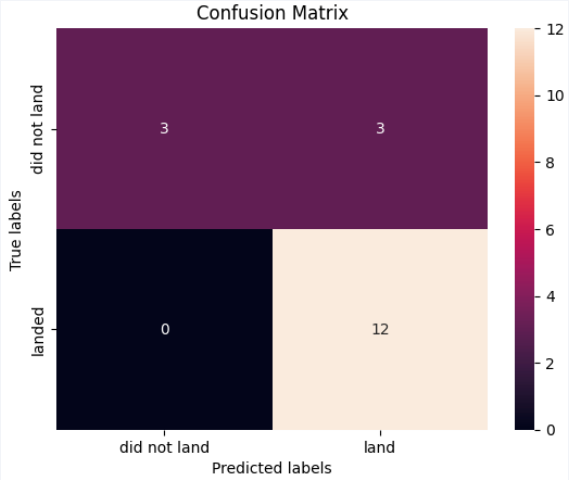
Logistic Regression



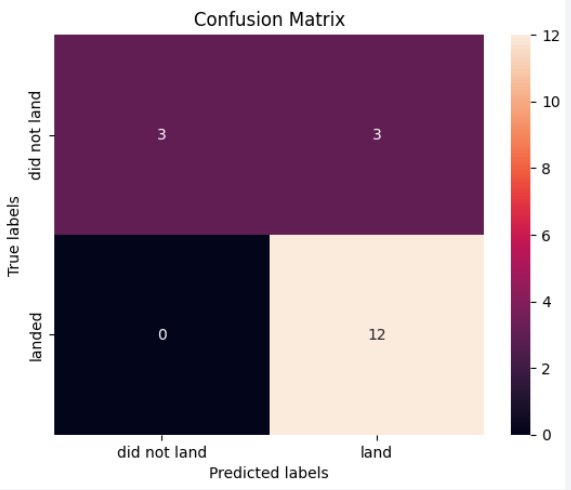
Decision Tree Classifier

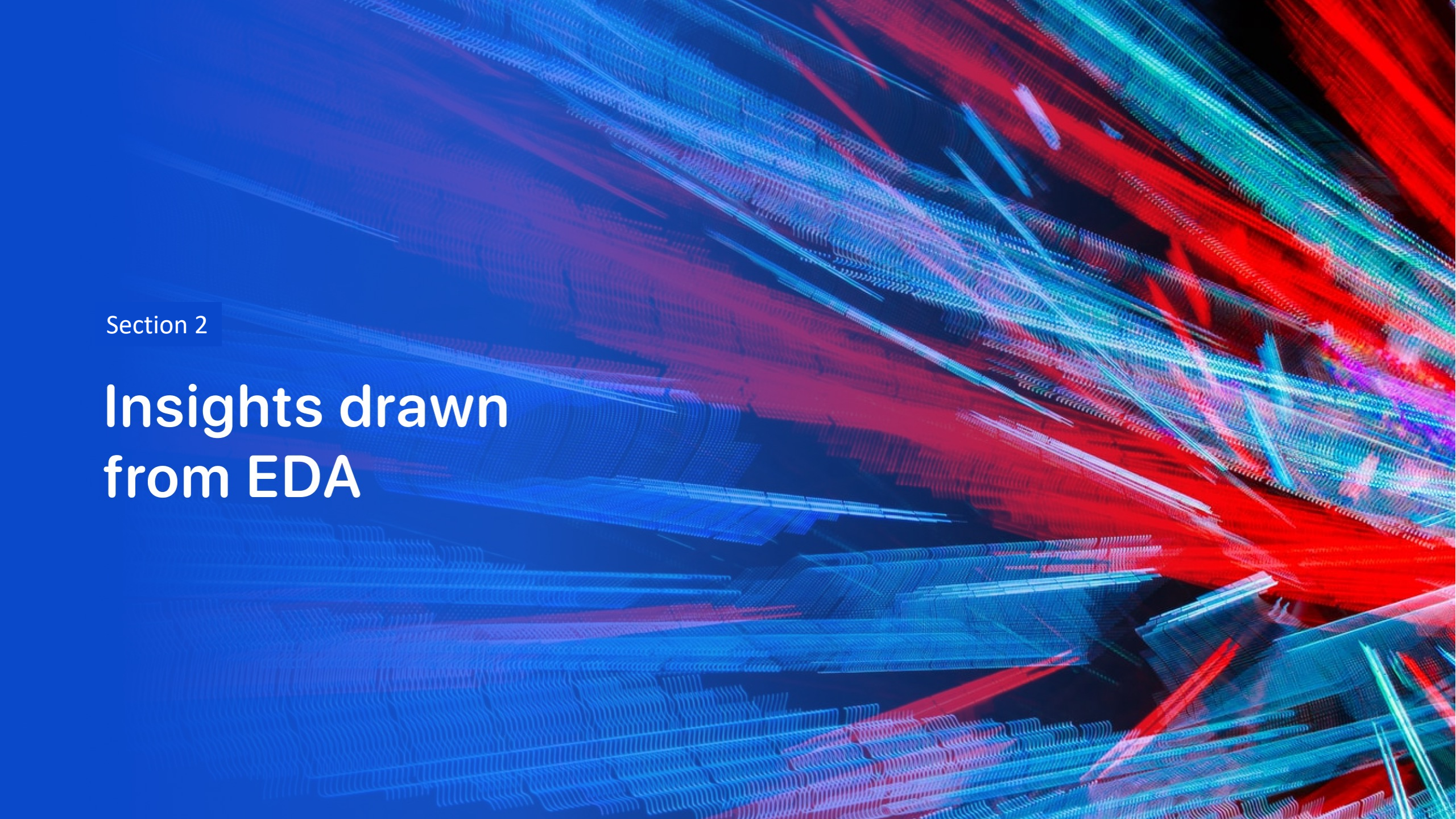


Support vector Machine

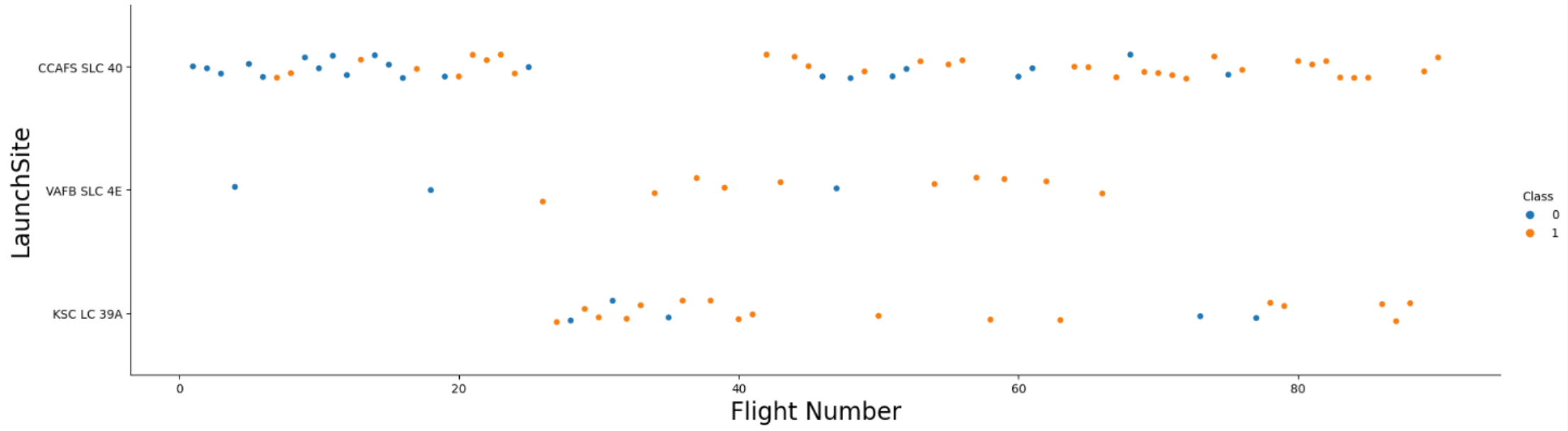


k nearest neighbours




Section 2
Insights drawn from EDA

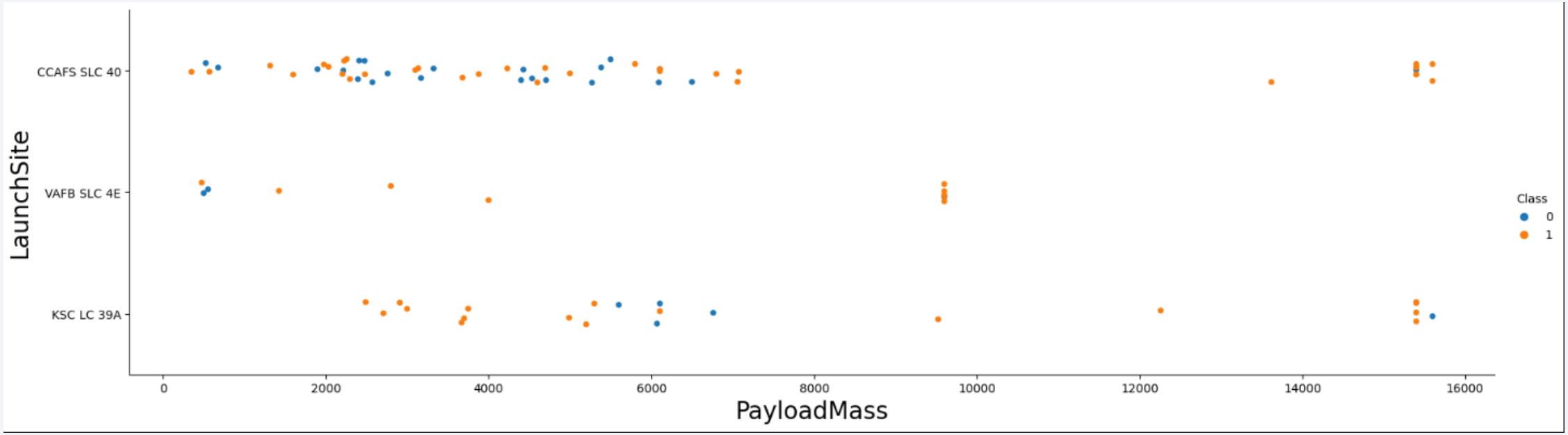
Flight Number vs. Launch Site



Flight Number Vs Launch site

- Success rates are high when Flight numbers are more than 75.
- Results are mixed for Launch sites with lesser flight number

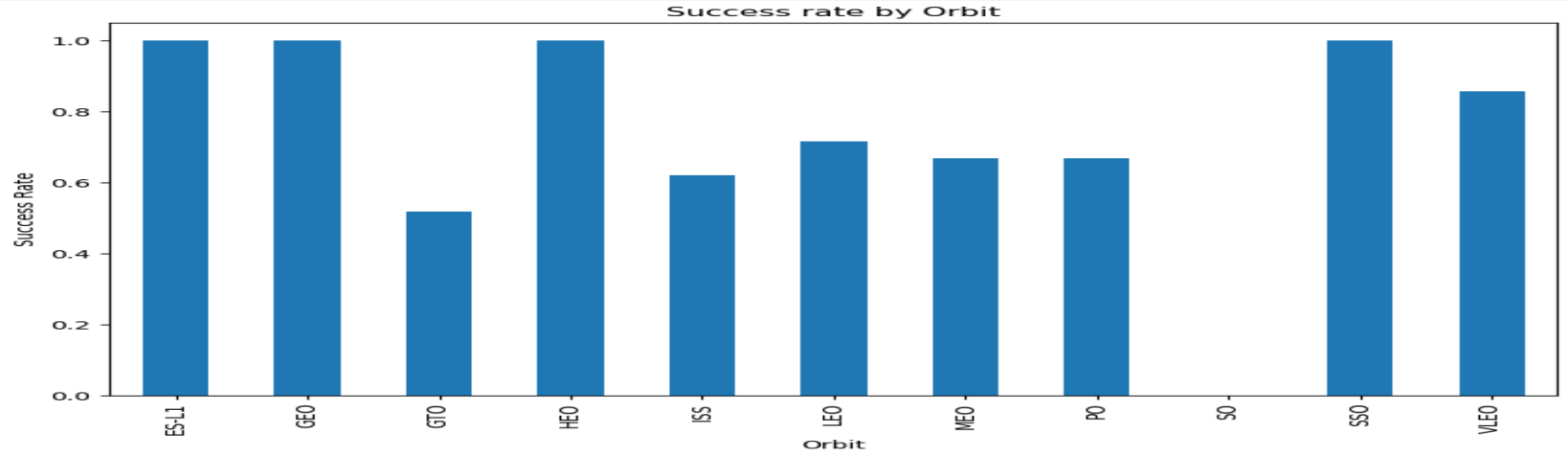
Payload vs. Launch Site



Payload mass Vs Launch site

- High success rate for payload mass between 7000 and 15000. But number of trials are comparatively less in this range

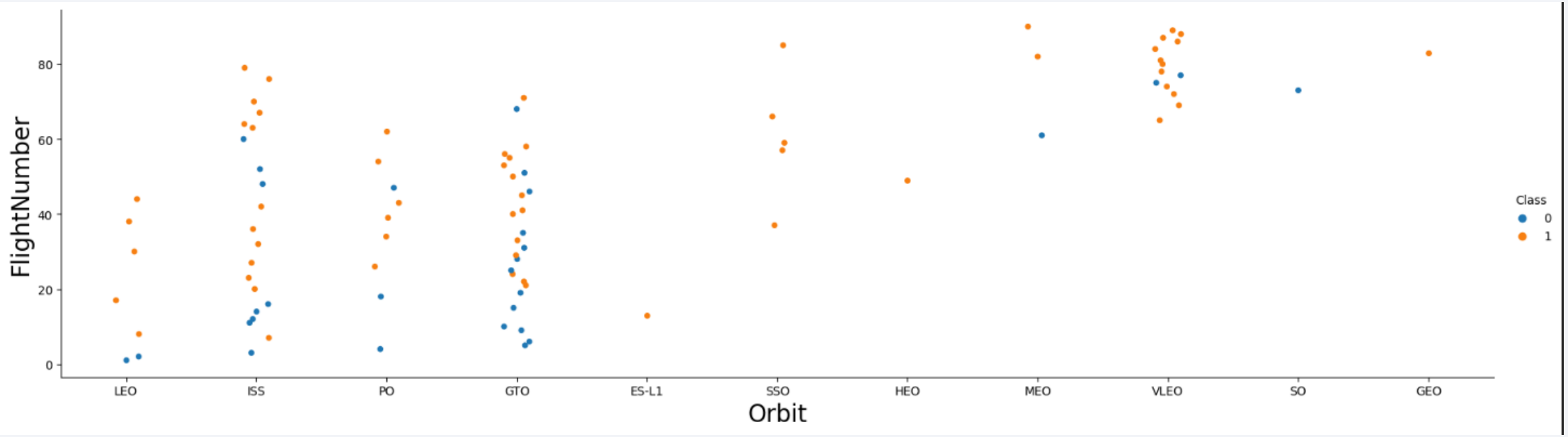
Success Rate vs. Orbit Type



Success Rate by Orbit

- Rocket launches to ES-L1, GEO, HEO, SSO orbits have a high success rate

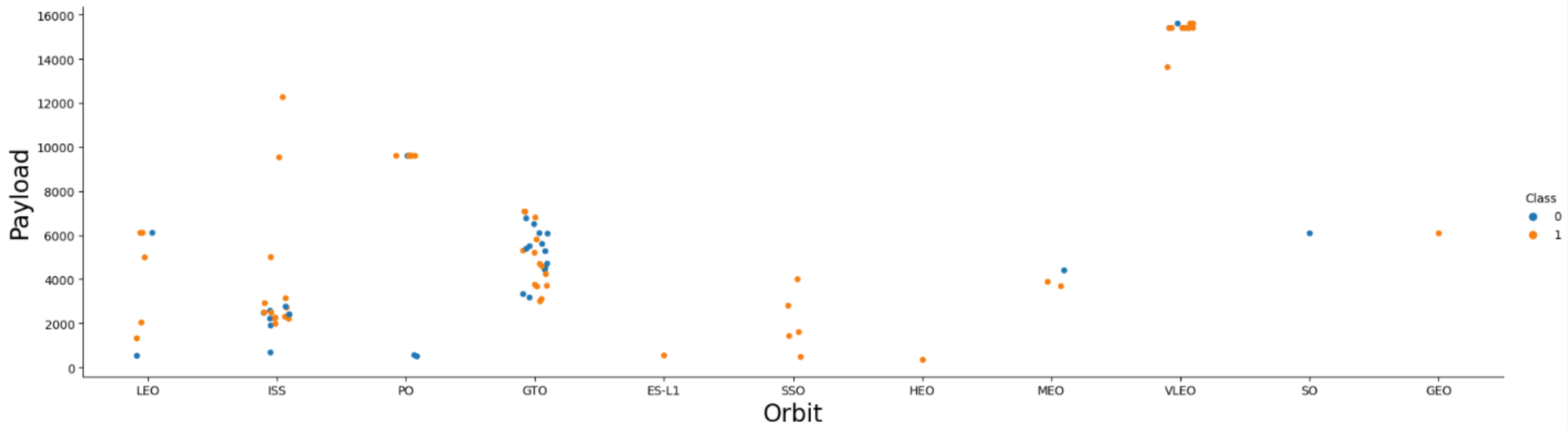
Flight Number vs. Orbit Type



Flight Number vs Orbit Type

- LEO orbit the Success appears related to the number of flights
- No visual relationship between flight number when in GTO orbit.

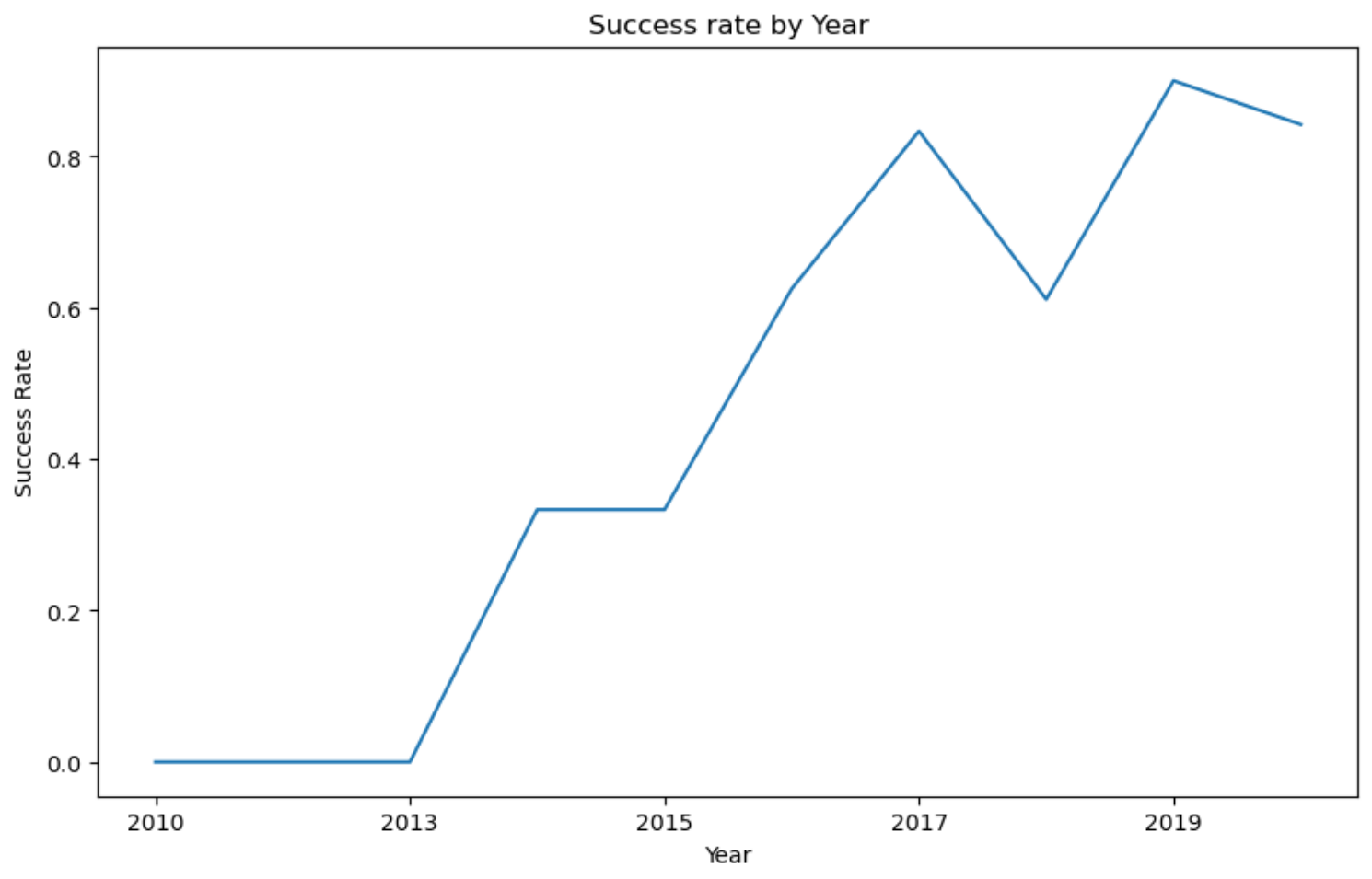
Payload vs. Orbit Type



Payload vs Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- GTO is an exception as the results are mixed for this orbit

Launch Success Yearly Trend



Launch Success Yearly

- The success rate since 2013 kept increasing till 2020

All Launch Site Names

Unique launch sites

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

SQL query used for getting Unique Launch Sites

SELECT Distinct(Launch_Site) FROM SPACEXTBL

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

SQL query used for getting Unique Launch Sites

SELECT * FROM SPACEXTBL where Launch_Site like 'CCA%' limit 5

Total Payload Mass

Total payload carried by boosters from NASA

payload_nasa
45596

Query

```
select sum(PAYLOAD_MASS__KG_) as payload_nasa from SPACEXTBL where Customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

avg_paylod_by_f9
2928.4

Query

```
select avg(PAYLOAD_MASS__KG_) as avg_paylod_by_f9 from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad

Date
22-12-2015

Query

```
select max(Date) as firstSuccessfulGroundpadLanding  from SPACEXTBL
where "Landing _Outcome"= "Success (ground pad)"
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Query

```
select Booster_Version  from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)"
and  PAYLOAD_MASS__KG_ between 4000  and 6000
```

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

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

Query

select count(*),Mission_Outcome from SPACEXTBL group by Mission_Outcome

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

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

Query

```
select distinct(Booster_Version) from SPACEXTBL
where PAYLOAD_MASS__KG_=(
select max(PAYLOAD_MASS__KG_)  from  SPACEXTBL
 )
```

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	year	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query

```
select substr(Date, 4, 2) as month, substr(Date,7,4) as year, "Landing _Outcome", Booster_Version, Launch_Site
from SPACEXTBL where "Landing _Outcome" = "Failure (drone ship)" and substr(Date,7,4)='2015'
```

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

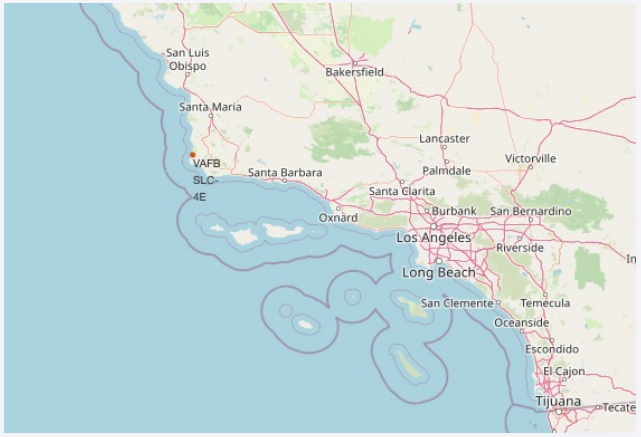
Landing _Outcome	countOfSuccess
Success	20
Success (drone ship)	8
Success (ground pad)	6

Query

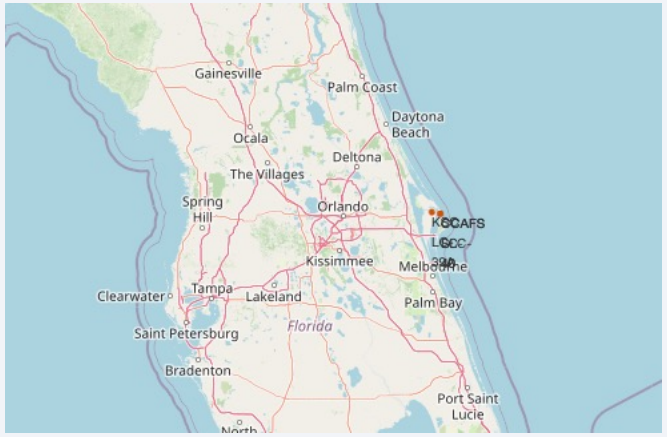
```
select "Landing _Outcome", count(*) as countOfSuccess from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017' and "Landing _Outcome" like '%Success%'
group by "Landing _Outcome" order by countOfSuccess DESC
```


Section 3
Launch Sites
Proximities Analysis

All launch sites on Map

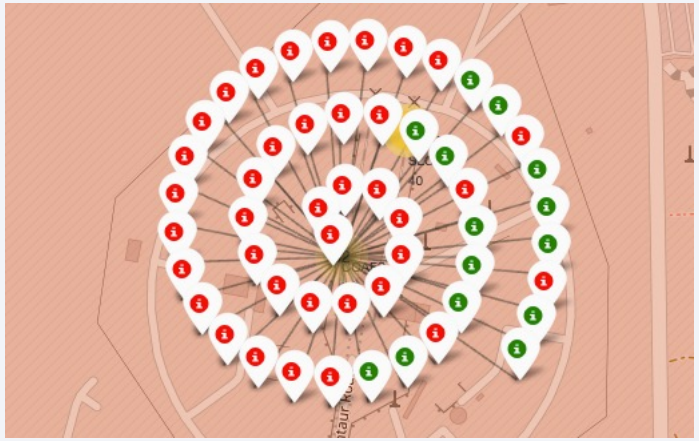


1 Site in California

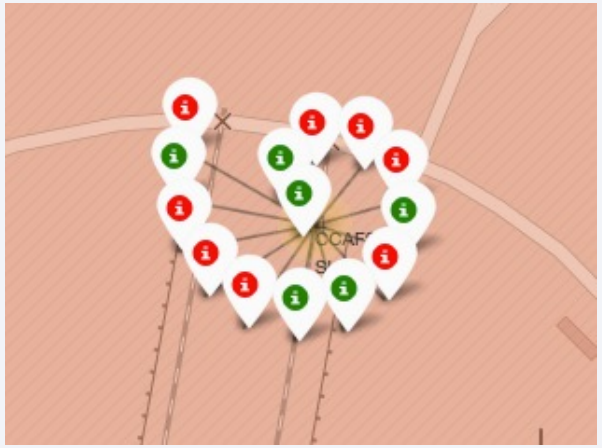


3 Sites are in Florida

Color-labeled launch outcomes on the map for each site



Site – CCAFS LC-40



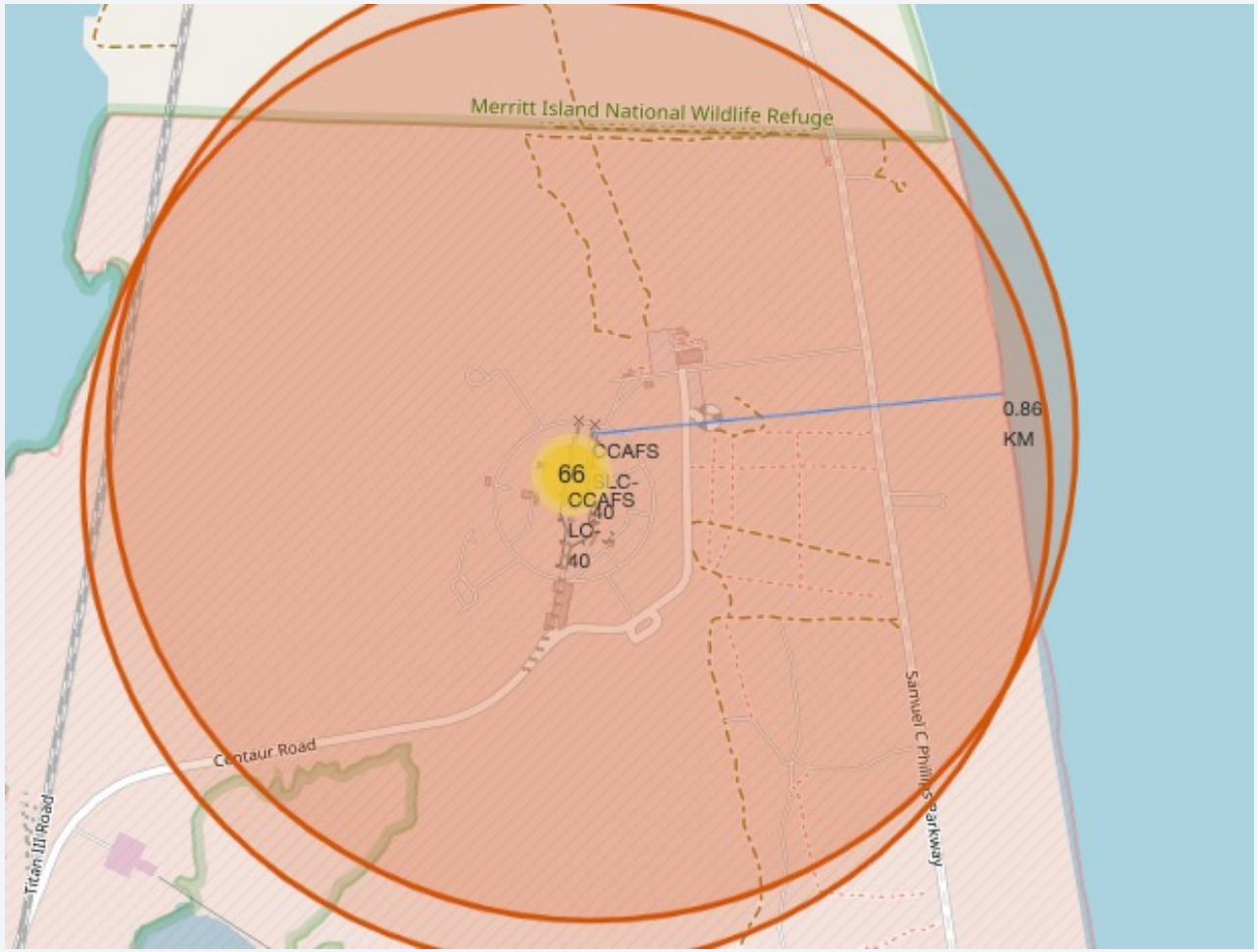
Site – CCAFS SLC-40

Site- KSC LC- 39A

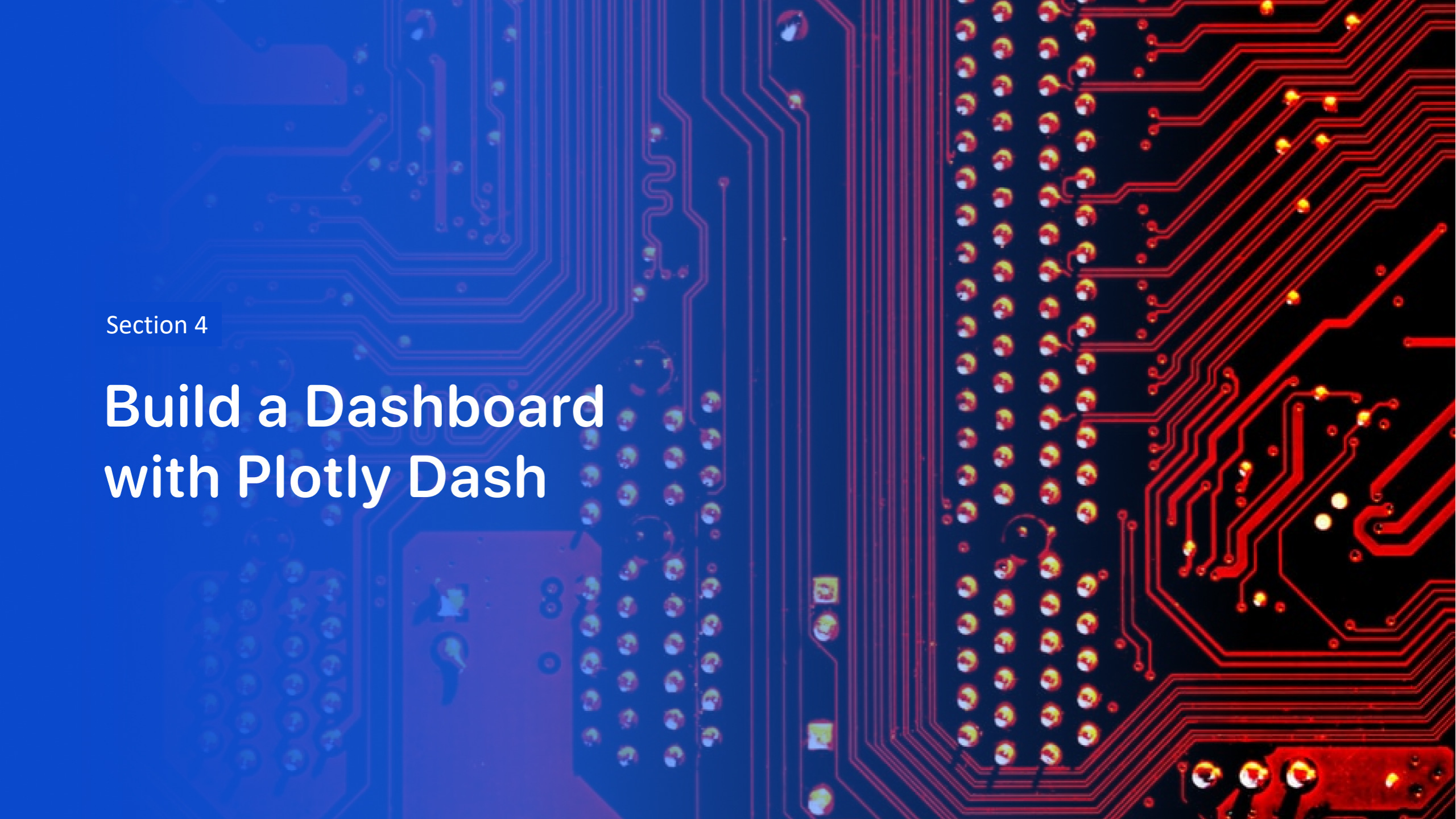
Site – VAFB SLC-4E

- **Site ‘KSC LC 39’ has the highest success rate**
- **Site CCAFC-LC-40 has the the lowest success rate**

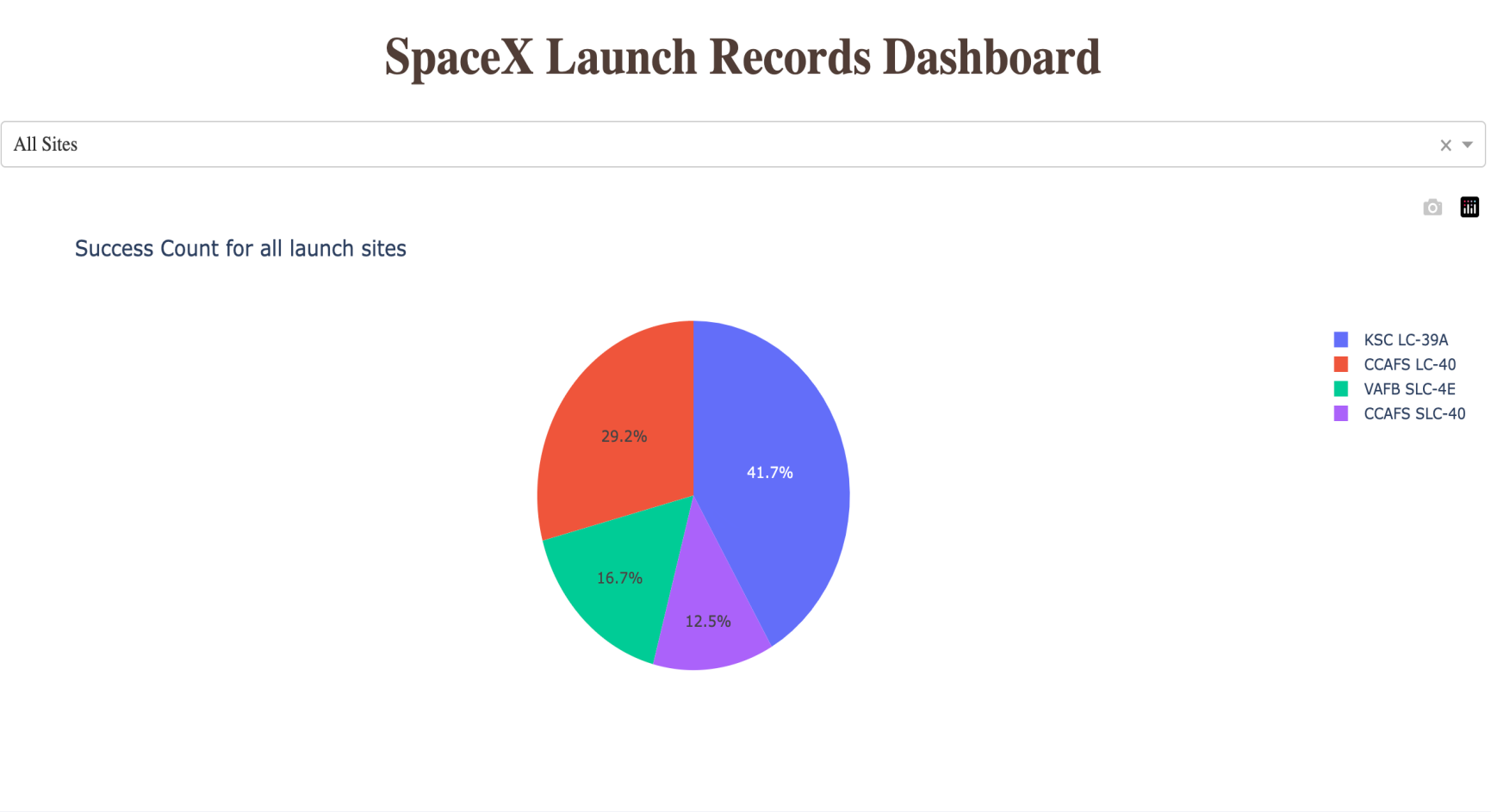
Proximity Map- Nearest Coastline



Site CCAFS SLC-40 is just 0.86 Km near to coast

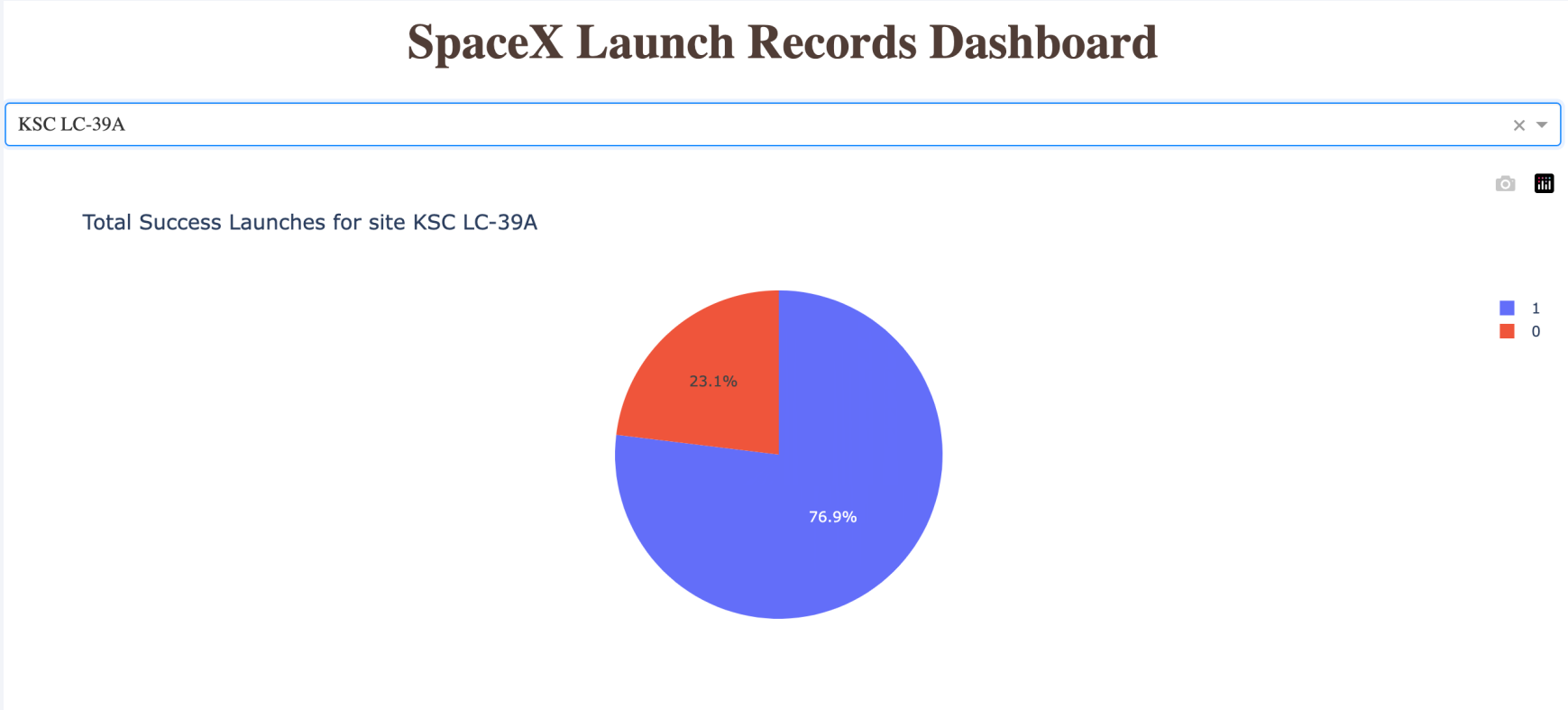

Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Record for All Sites



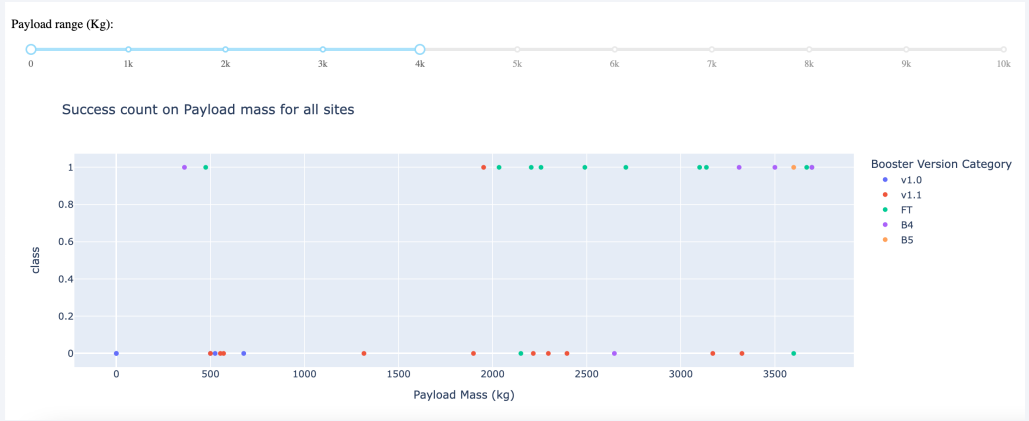
- Site KSC LC-39A has the most successful launches of 41.7%
- Site CCAFS SLC-40 has the least successful launches of 12.5%

Success/ Failure Ratio for Site KSC LC-39A

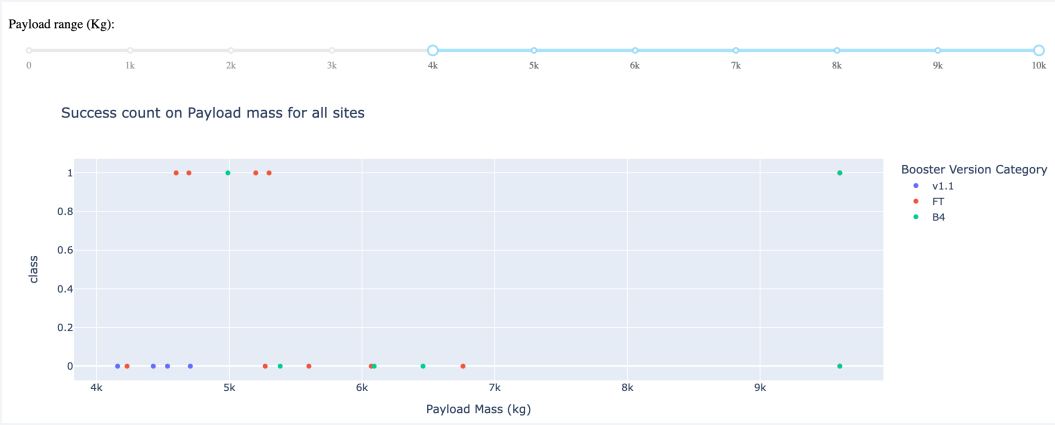


76.9% of all launches from site KSC LC-39A was successful while 23.1% launches failed

Payload vs. Launch Outcome



Success Rate for low Payload between 0 to 4000KG



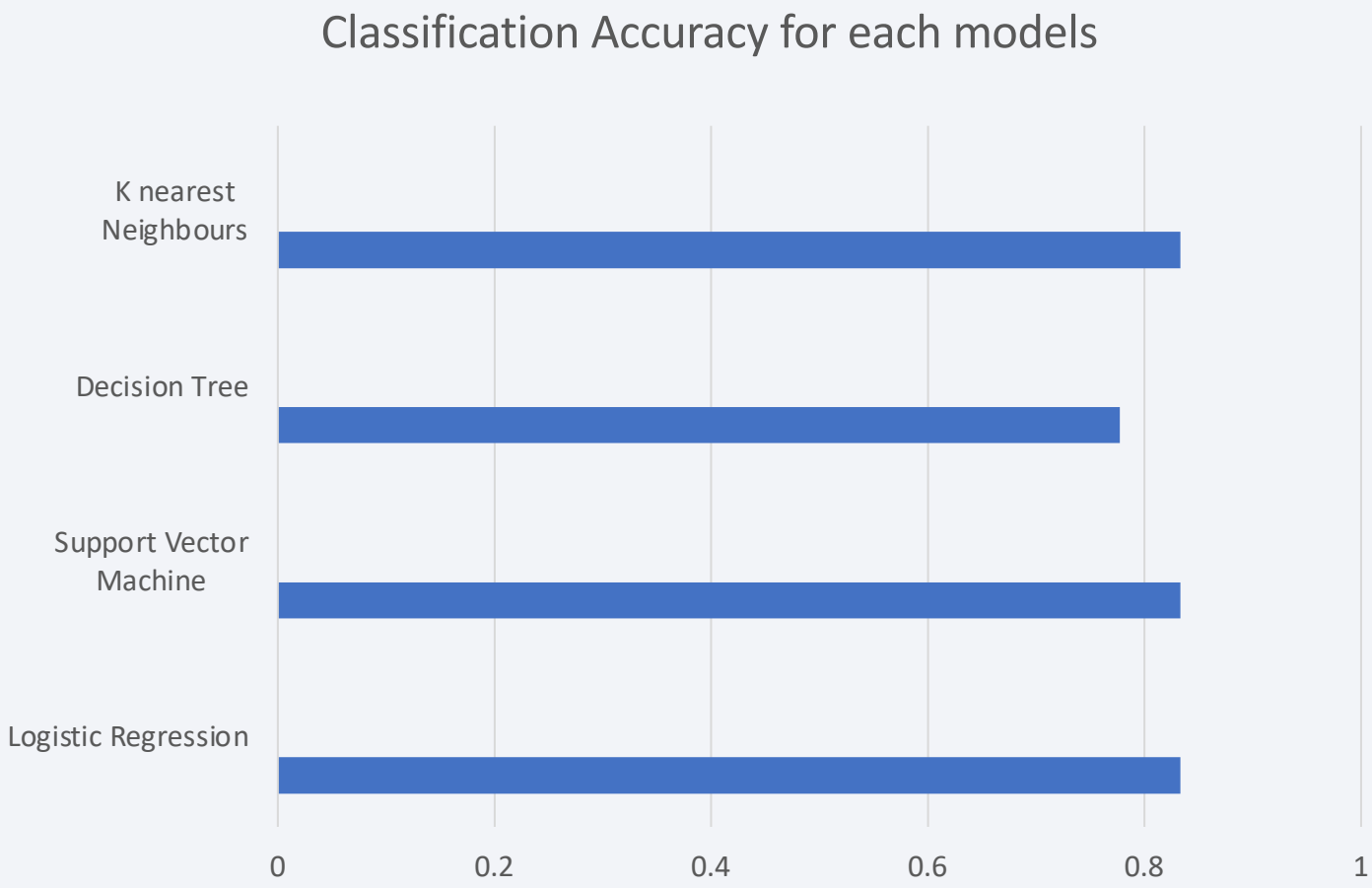
Success Rate for high Payload between 4000KG to 10000KG

Success Rate for launches with lower payloads(< 4000KG) are higher than with high payload of 4000Kg to 10000KG

Section 5

Predictive Analysis (Classification)

Classification Accuracy

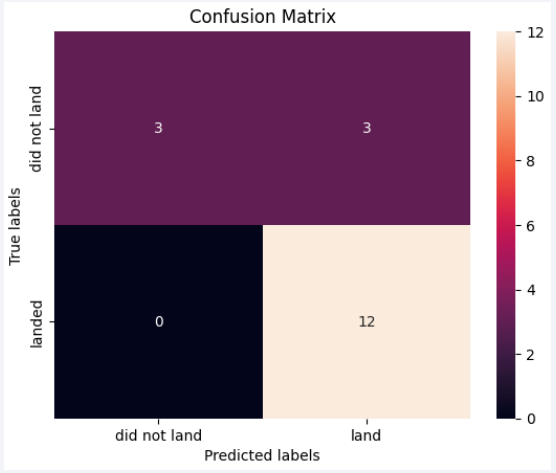


Below 3 models gave the best accuracy of 83.3%

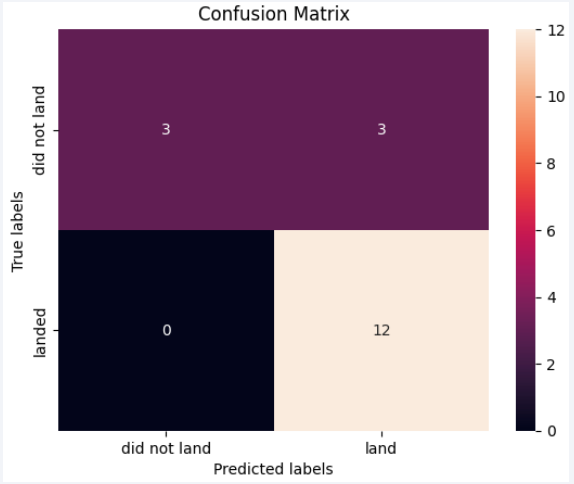
- Logistic Regression
- Support Vector Machine
- K nearest Neighbours

Decision Tree gave a less accuracy of 77.7%

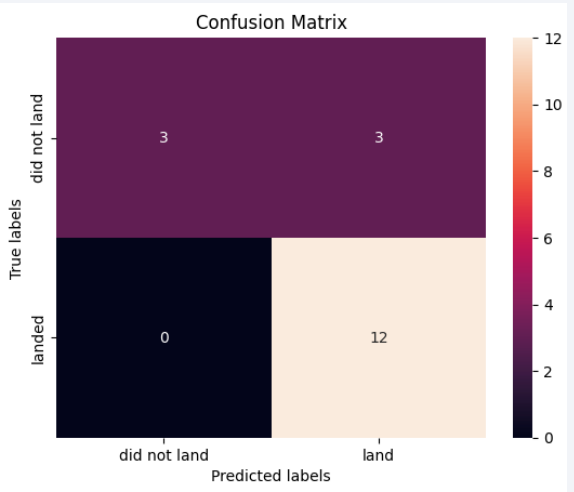
Confusion Matrix – High Accuracy models



Logistic Regression



K nearest neighbours



Support vector Machine

All 3 above models gave the best accuracy of 83.3%

Conclusions

- Success Rate for SpaceX launches increasing relatively every year
- High Payloads gives better success rates
- 3 Models gave a high Success rate in predicting Launch success
- Site KSC LC-39A has the most number of successful launches
- Rocket launches to Orbits ES-L1, GEO, HEO, SSO have a high success rate

Appendix

- Model accuracy evaluation for all 4 models

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Thank you!