Winning Space Race with Data Science

Shwetha

2022-02-01

Outline

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

Executive Summary

- The goal of this project is to estimate if falcon 9 stage 1 would land successfully or not, as it plays a major role in predicting the price of its relaunch. The data for this project was sourced from space x REST API, and Wikipedia. After performing some data wrangling In order to determine the best predictors for our outcome, EDA and feature scaling were done with the help of visualization using scatter and line plots. Later some ML models were created to predict future outcomes.
- The results showed that the outcome was dependent on the orbit, mass of payload, launch site, and various other technical factors such as gridfins, cores etc. The results show that there is huge progress in space x regarding this space race.

Introduction

- The evolution of technologies has changed the lives of people a lot, and with the current technologies, we are on the verge of building commercial space flights. Which can make humans multi-planetary species. There are major companies in this space race, namely blue origin, virgin galactic, and space x. The current leader in this race seems to be space x, and the reason behind that is the reusability of their stage 1. Which reduces the cost of launch from a minimum of165 million to around 62 million per launch.
- The problem that we are trying to answer is that how can we predict the launch price of falcon 9, so that we can use this data for companies that want to compete with space x. Predicting that whether stage 1 will land successfully or not, plays a crucial role in predicting the launch price. As that stage can be reused again with different payloads, thus reducing the cost by more than half of original.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from space x using REST API and from Wikipedia using web scrapping frameworks such as Beautiful Soup.
- Perform data wrangling
  - The null values were handled at the time of performing web scrapping, one hot encoding was done on categorical variables such as orbit, launch site, landing pad and serial.
- 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 models like SVM, logistic regression, tree classifier and k nearest neighbors were used.

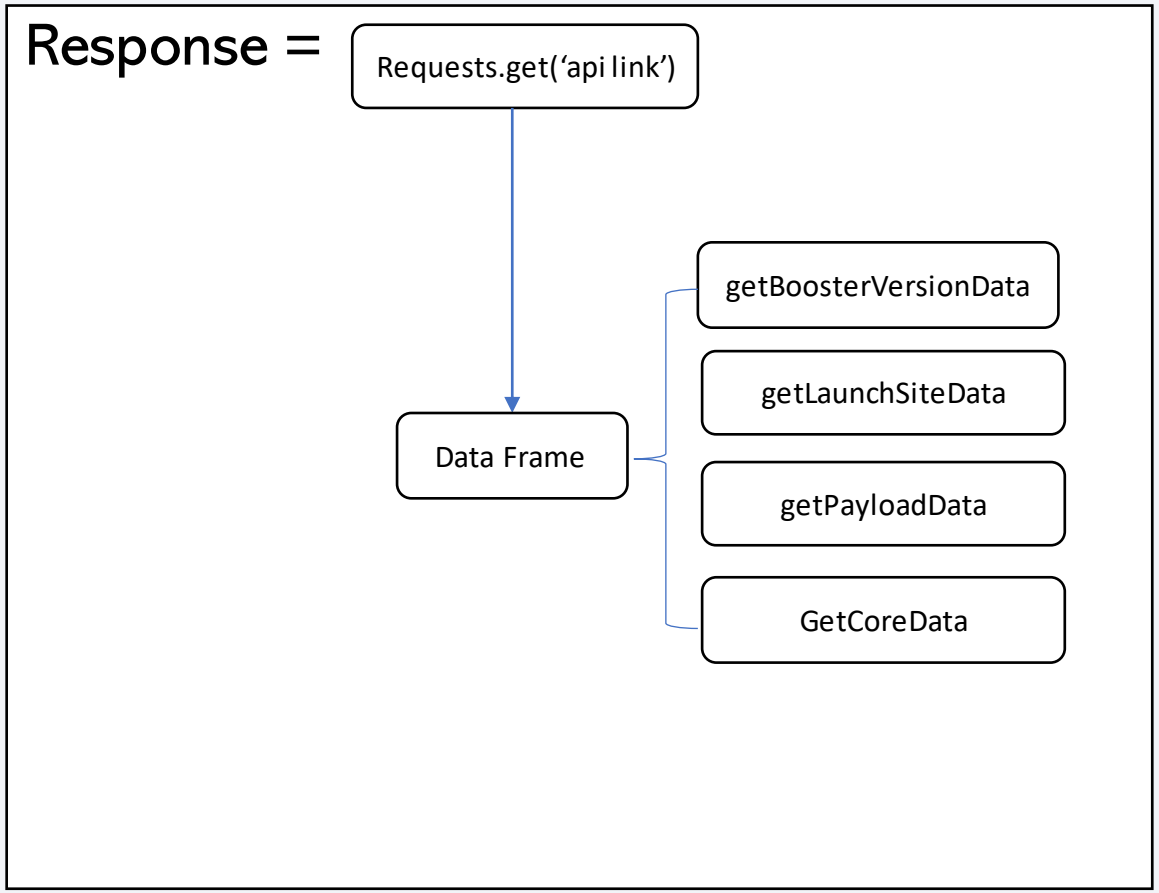
Data Collection

- The data was first collected from the space x rest API with endpoint: Click me most of the data of certain attributes was encoded like its name so it was decoded with the help of rest api connection to the specific attributes detail's endpoint like for example to get the rocket's name a connection to endpoint click me was used. The data was also sourced from the following webpage using web scraping, with the help of BeautifulSoup framework.

Wikipedia: Falcon 9 and falcon heavy list

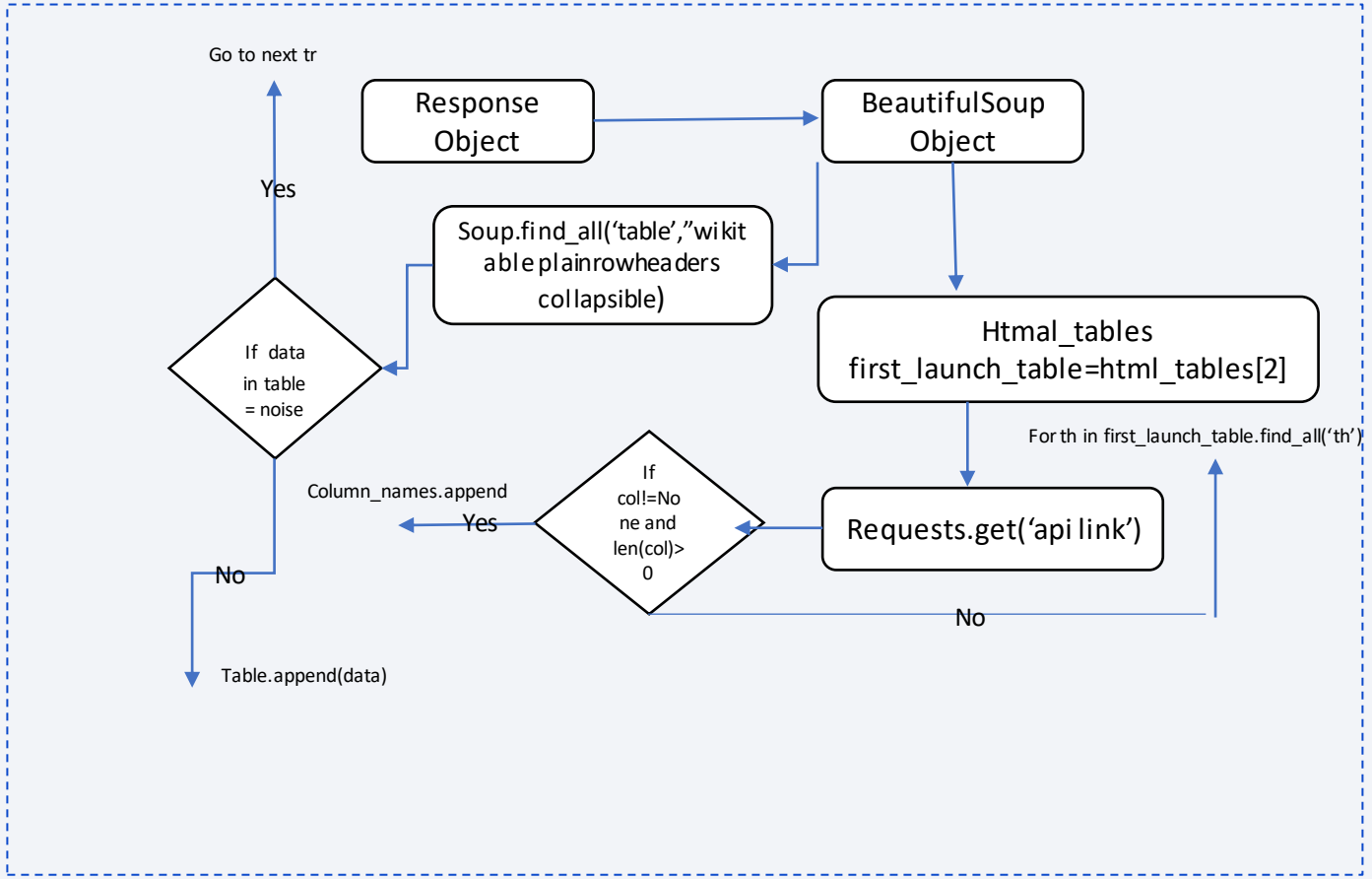
Data Collection – SpaceX API

- A request object was created using the space x api's end point. The response object was converted to a dataframe using: pandas.json_normalize(response .json())
- GitHub URL : Link



Data Collection - Scraping

- Some of the essential data was collected from Wikipedia using web scrapping with the help of beautiful soup framework.
- GitHub URL : [Link]



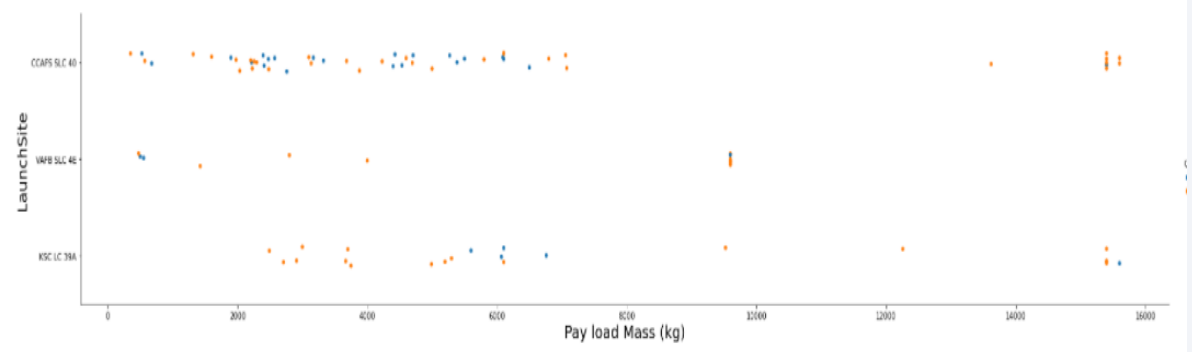
Data Wrangling

- Initially, the data were filtered to contain only the records of falcon 9, later it was found that there were some null values in the data as shown in the figure: Landing pad would remain null as It represents that the landing pad wasn't used. Null values in the column PayloadMass will be replaced by the mean value.
- One hot encoding was done for some categorical varibles in order to feed them to the ML algorithm after performring feature scaling.
- The class variable had the values {'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'} These values were converted to 0 and {'True ASDS', 'True Ocean', 'True RTLS'} to 1 representing successful landing of stage 1. Additionally, success rate was found out to be: 66%
- GitHub URL : Link

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       5
Orbit             0
LaunchSite        0
Outcome           0
Flights           0
GridFins          0
Reused            0
Legs              0
LandingPad       26
Block             0
ReusedCount       0
Serial            0
Longitude         0
Latitude          0
dtype: int64
```

EDA with Data Visualization

- Bar graphs and Scatter plots for launch sites w.r.t number of launches, payload and number of launches w.r.t payload mass was created with hue set to class variables(it represents colors according to the values of class which in our case was 0 for blue and 1 for yellow) as example:



- The visualisation also shows the increasing progress of space x with each year,
- GitHub URL : Link

EDA with SQL

- Performed SQL queries to gather information about the dataset.
- For example of some questions we were asked about the data we needed information about. Which we are using SQL queries to get the answers in the dataset :
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- GitHub URL: Click

Build an Interactive Map with Folium

- An interactive map was created for visualizing the various factors, markers, and highlighted circles with popups were added for different launch sites to easily spot them on the map. Cluster object of markers was created to show lunch outcomes, with red being 0 I.e unsuccessful, and 1 being green I.e., successful landing. At the end a polyline object was added to the map to show the distance of the launch site from its proximites such as nearest railway station.
- GitHub URL : Link

Build a Dashboard with Plotly Dash

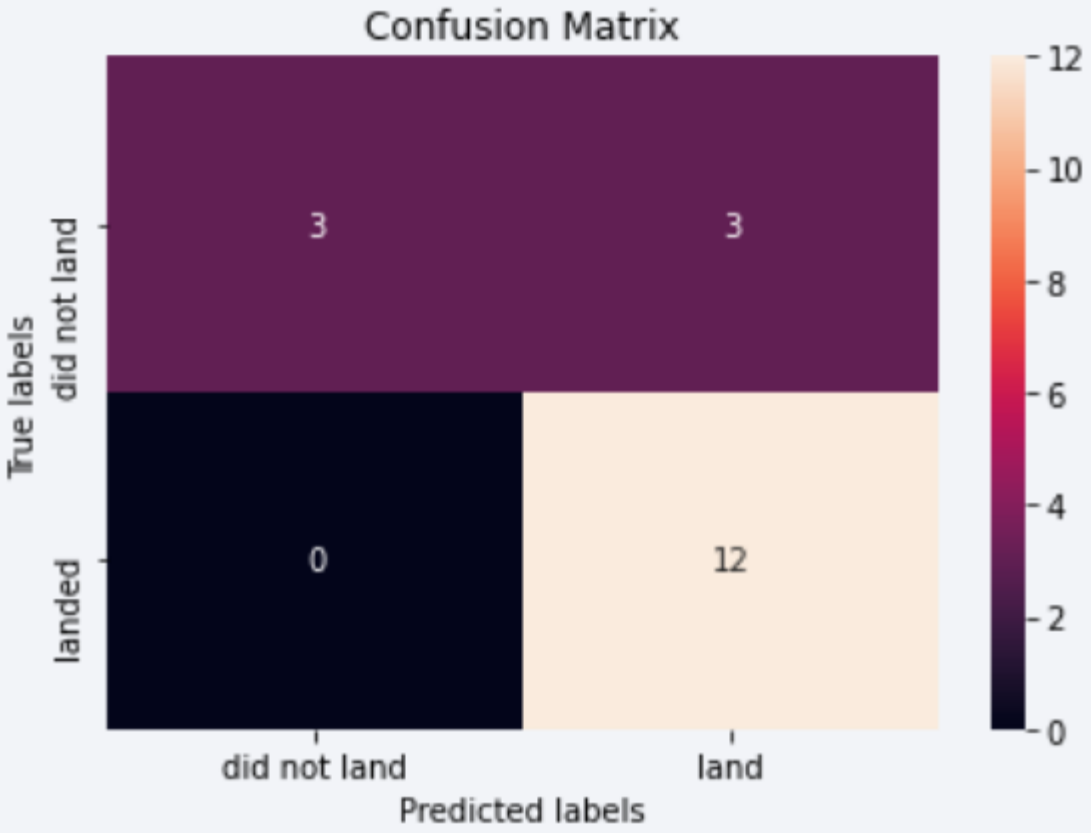
- An interactive web application was created using dash, with a drop-down menu to select the launch site, and range-slider to choose the range of payload. Interactive pie chart showing success rate of all launch sites by default, and a scatter plot showing launch outcomes of all sites according to their payloads in the default range(0-10000) were added.
- Dropdown menu would allow the user to choose the launch site that would alter the figure of pie chart and scatter plot to show outcomes of that launch site, and through the range-slider user can select the range of payload on the x-axis of scatter plot. These interactions would allowthe user to visualise the data more in depth according to his needs.
- GitHub URL : Link

Predictive Analysis (Classification)

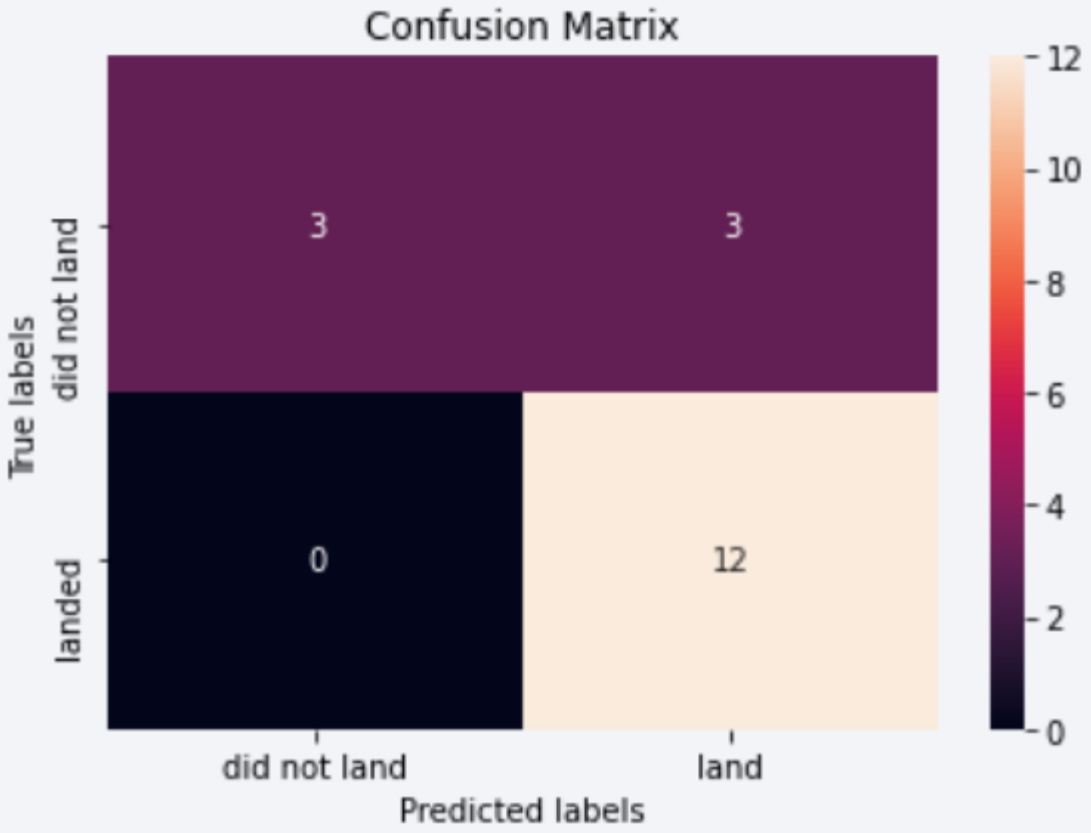
- Many various types of ML model was used with in addition with GridSearchSV. These ML algorithms were logistic regression, SVM, decision tree and KNN. The model with best out of sample accuracy were KNN, logistic regression and SVM. The R2 score was around .83

- GitHub URL : Link

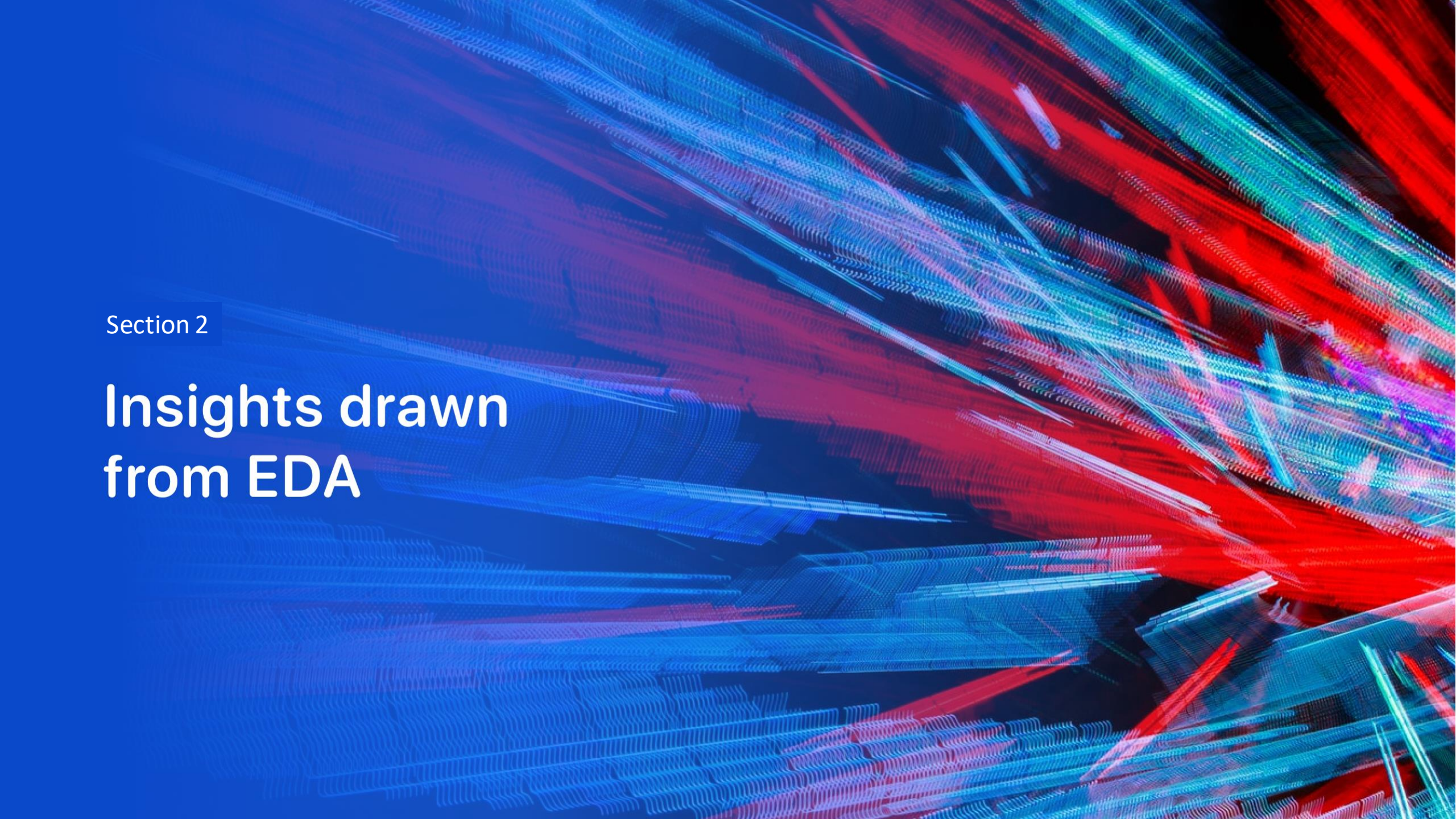
Results



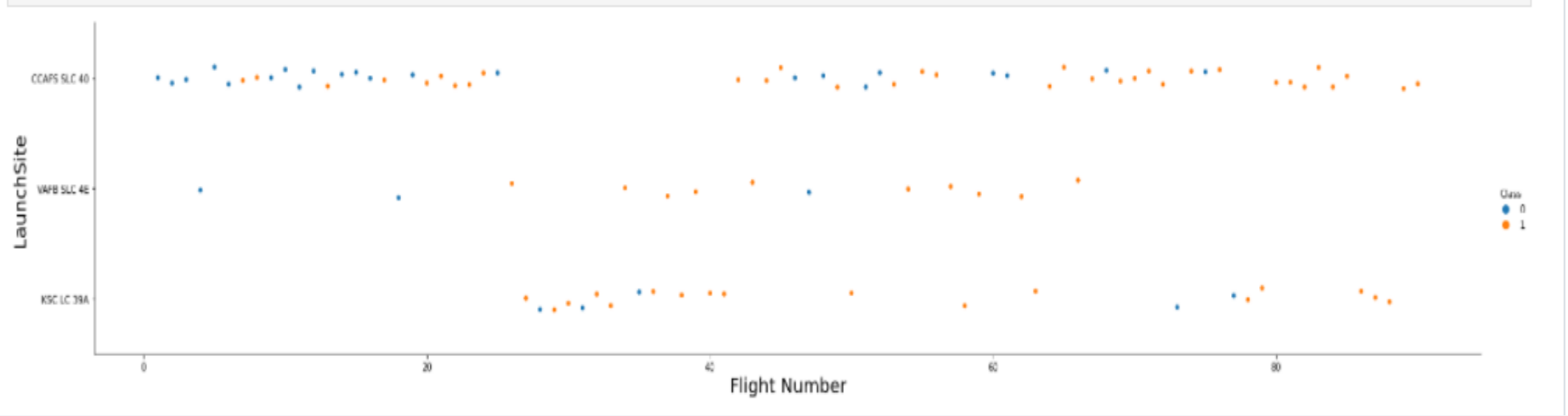
Logistic regression



SVM

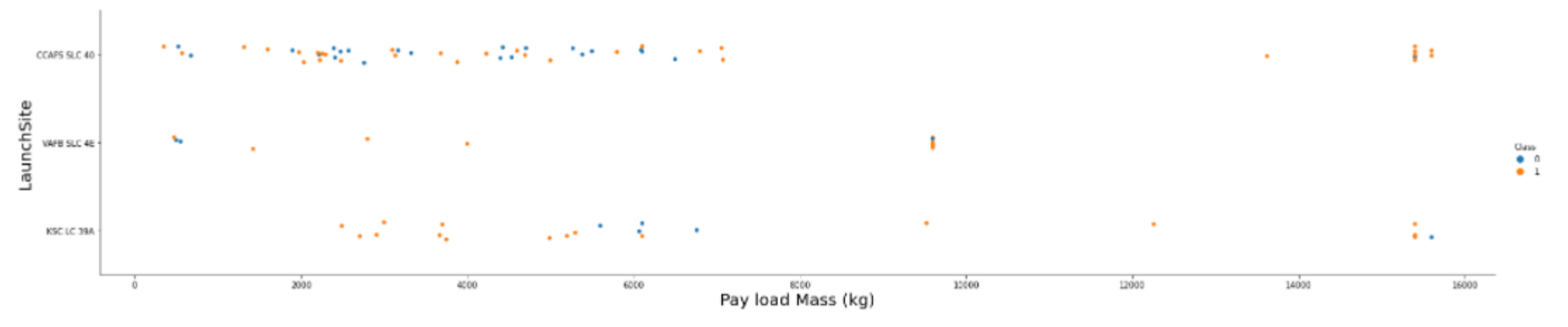

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



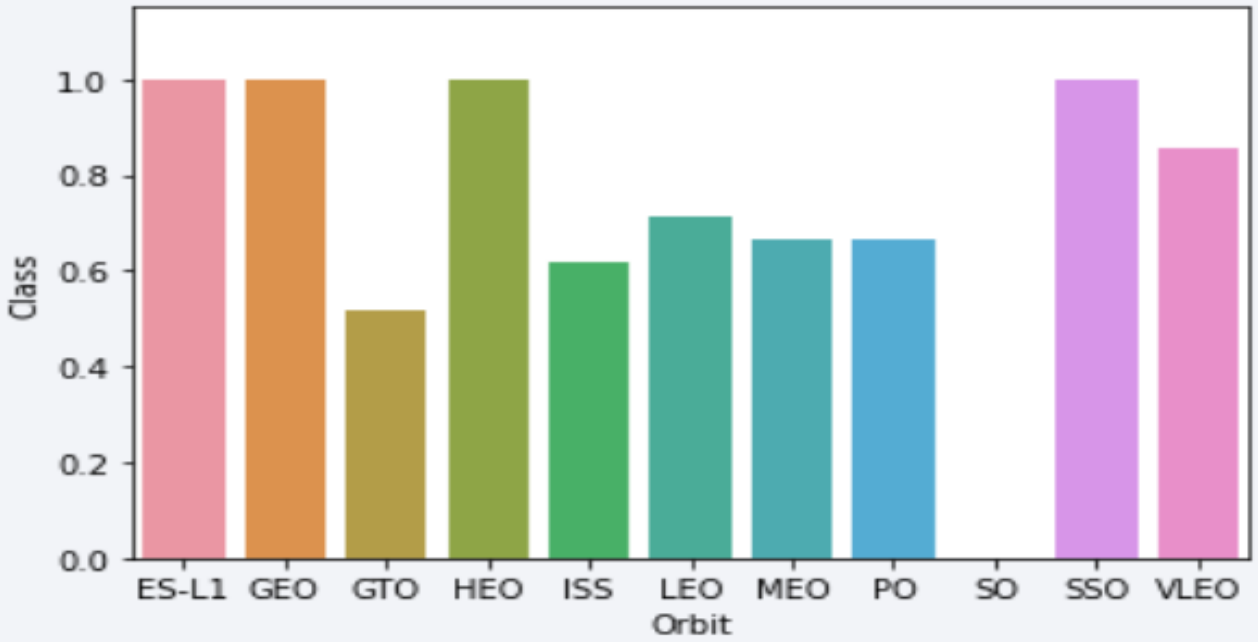
- We can see above that CCAFS SLC 40 have a higher chance of success with increasing number of launches. Same goes for the KSC LC 39A.

Payload vs. Launch Site



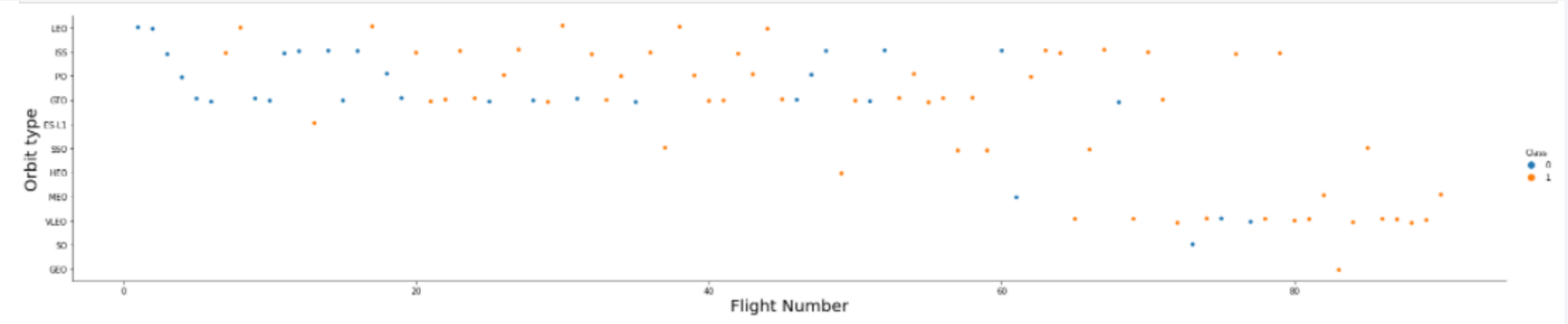
- The launch site CCAFS SLC 40 had a high chances of success, when the payload mass was lower. We can also see that VAFB SLC 4E was not used much in comparison to other launch sites.

Success Rate vs. Orbit Type



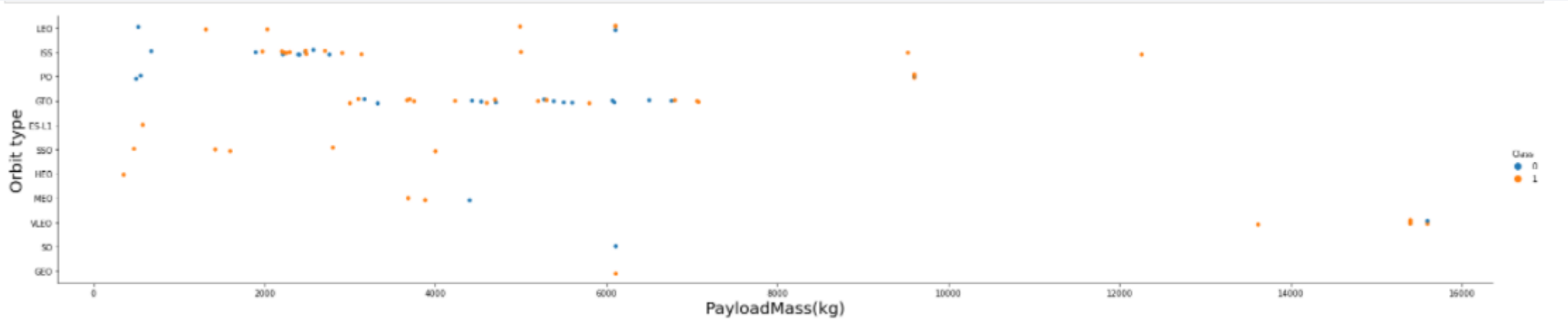
- Here we can notice that launch w.r.t SSO, HEO, ES-L1 and GEO orbits have a higher chances of success

Flight Number vs. Orbit Type



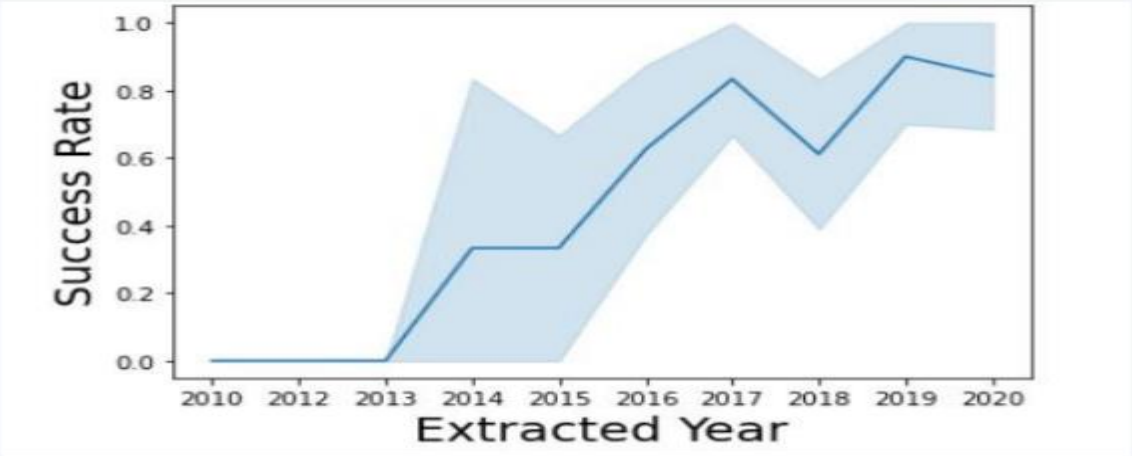
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- We can observe that Heavy payloads have a negative influence on GTO orbits and positive on ISS and LEO orbits.

Launch Success Yearly Trend



- We can observe that the success rate since 2013 kept increasing till 2020. It shows how much progress had space x done in recent times

All Launch Site Names

- The names of the all launch sites are:

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

- We used pandas.groupby() funtion passing 'Launch Sites' as the parameter, and then we printed the first element in those group using .first() function on our groupby object.

```
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
```

%sql select * from SPACEXTBL where launch_site like 'CCA%' Limit 5

Launch Site Names Begin 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

- Used like and limit function to get result.

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' Limit 5
```

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)'
```

```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
45596

- Total payload mass carried by booster launched by NASA(CRS) is 45596, which is filtering customer by NASA(CRS).

Average Payload Mass by F9 v1.1

- 2928KG is average payload mass carried by booster version F9 v1.1, which is obtained by filtering booster_version by ‘F9 v1.1’ and using avg function.

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

```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
2928

First Successful Ground Landing Date

- '2015-12-22' is the first successful landing outcome on ground pad which is obtained by filtering landing_outcome as 'Success(ground pad) and using min function.

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below are the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- To obtain this result, given filter on landing_outcome and between function on payload_mass_kg

```
%sql select BOOSTER_VERSION,PAYLOAD_MASS__KG_,landing__outcome  from SPACEXTBL where landing__outcome  = 'Success (drone ship)'
and PAYLOAD_MASS__KG_  BETWEEN 4000 and 6000
```

Total Number of Successful and Failure Mission Outcomes

- Below are the total number of successful and failure mission outcomes, result obtained by group by clause on mission_outcome and by count function.

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

List the total number of successful and failure mission outcomes

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Total from SPACEXTBL group by MISSION_OUTCOME
```

Boosters Carried Maximum Payload

- Below booster which have carried the maximum payload mass to obtain this result used max function in subquery.

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

```
%sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (Select max(PAYLOAD_MASS__KG_) from SPAC
EXTBL)
```

```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Below are failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select DATE,BOOSTER_VERSION,launch_site,landing__outcome  from SPACEXTBL where landing__outcome ='Failure (drone ship)' and left(DATE,4)=2015
```

```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

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

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

- 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 Total  from SPACEXTBL where DATE between '2010-06-04' and  '2017-03-20' g
roup by landing__outcome order by Total Desc
```

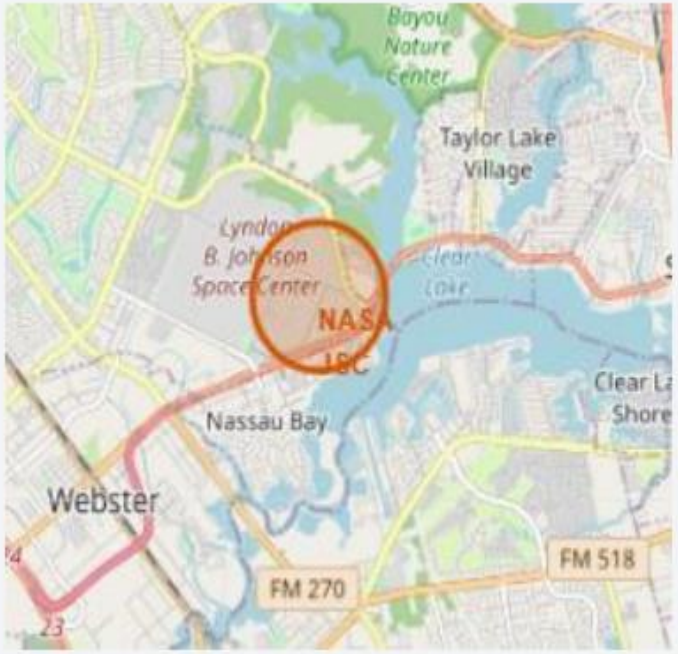
```
 * ibm_db_sa://hfp09989:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

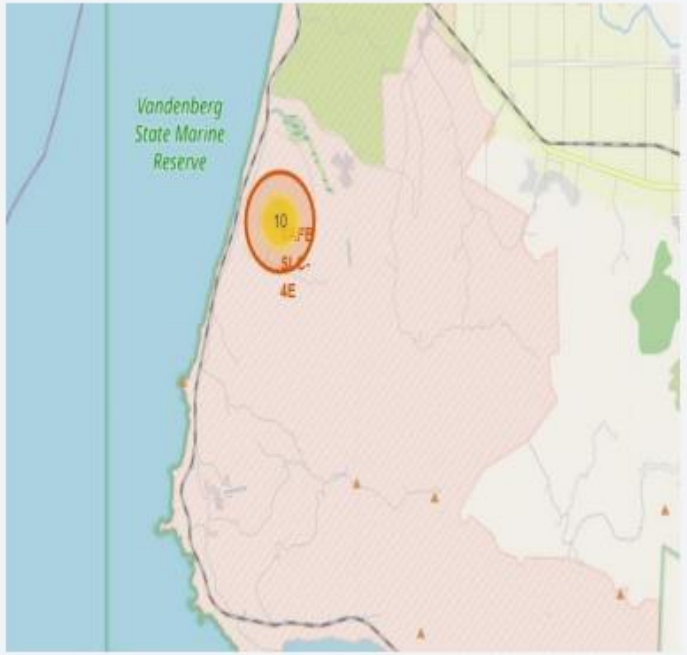
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

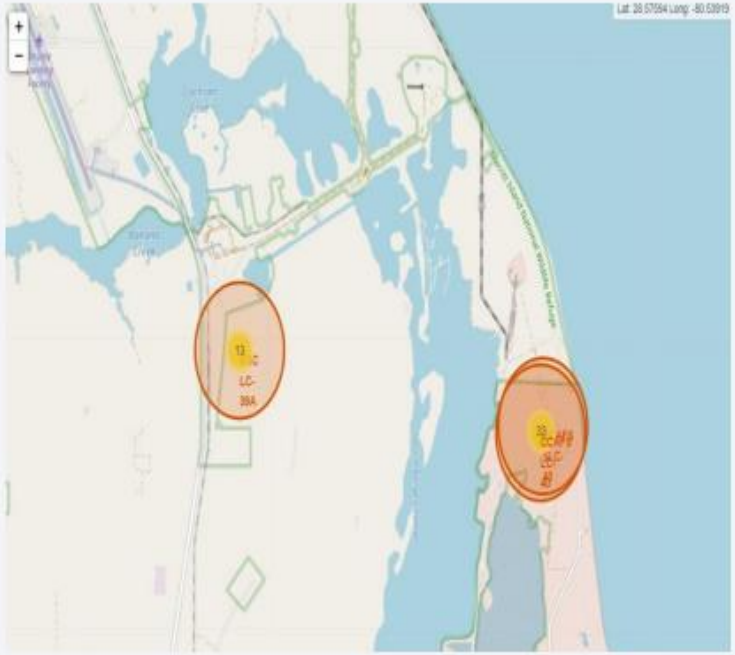

Section 3
Launch Sites
Proximities Analysis

Plotting launch sites on map

- We'll be plotting all the launch sites on the map by using the marker and circler objects of folium map, and then label them too.





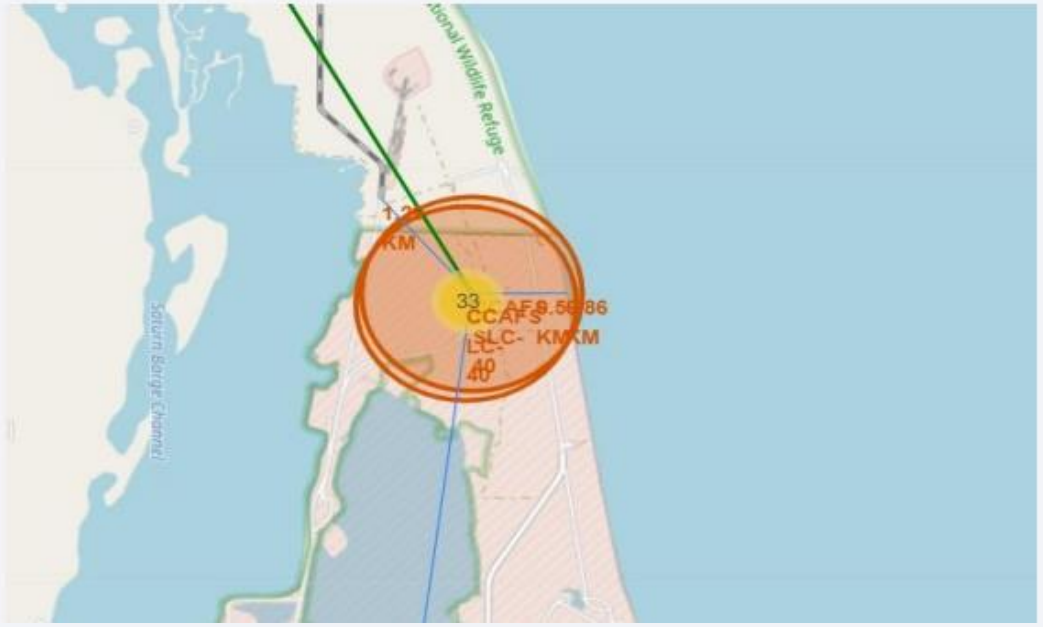


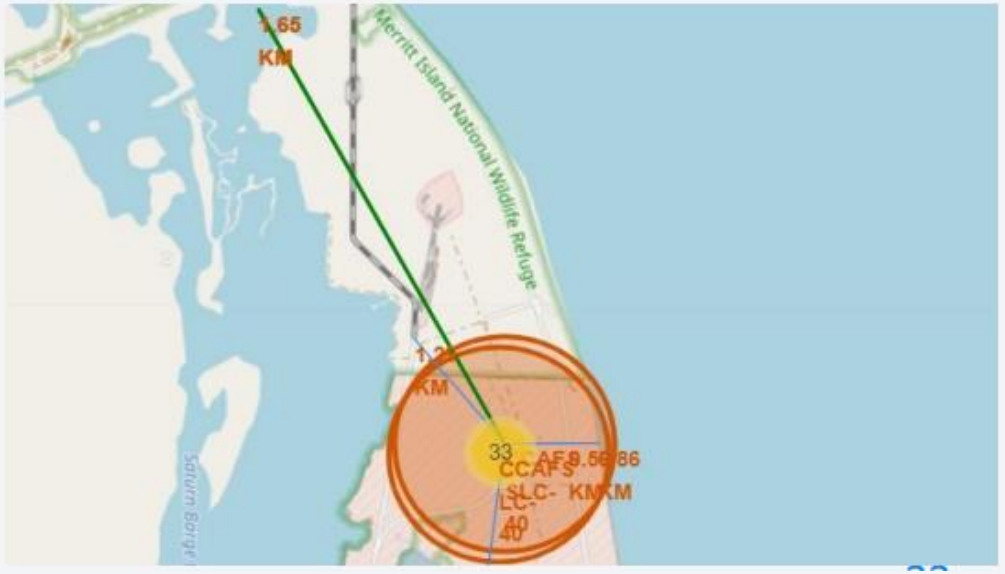
Marking outcomes

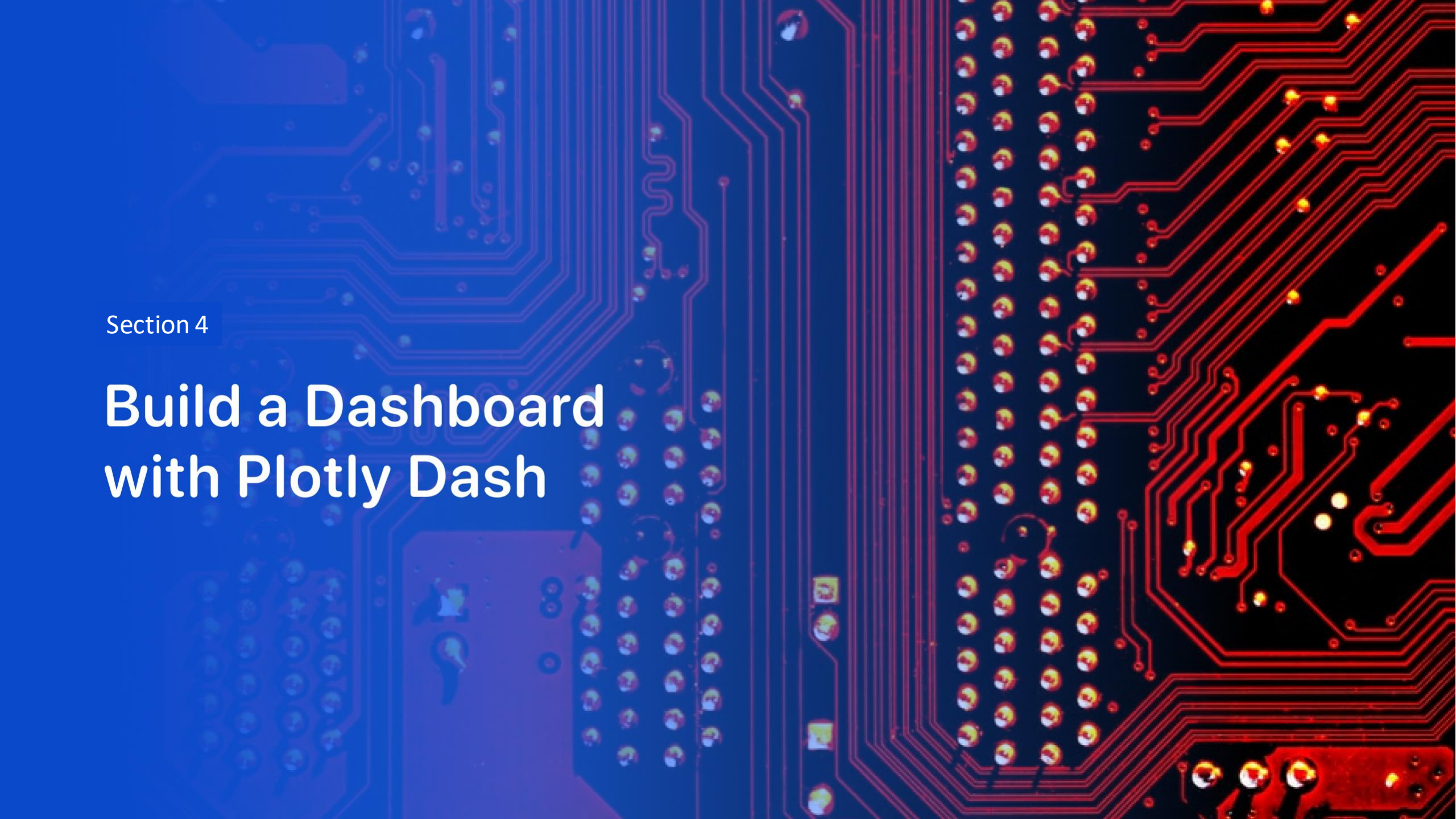
- Next, we will be adding a marker cluster representing total number of launches and their outcome with red representing failure and green success.

Adding distance to proximites

- In the end we'll be adding a polyline to the folium map representing the distance of launch sites from its nearby locations such as ocean and railway station etc.

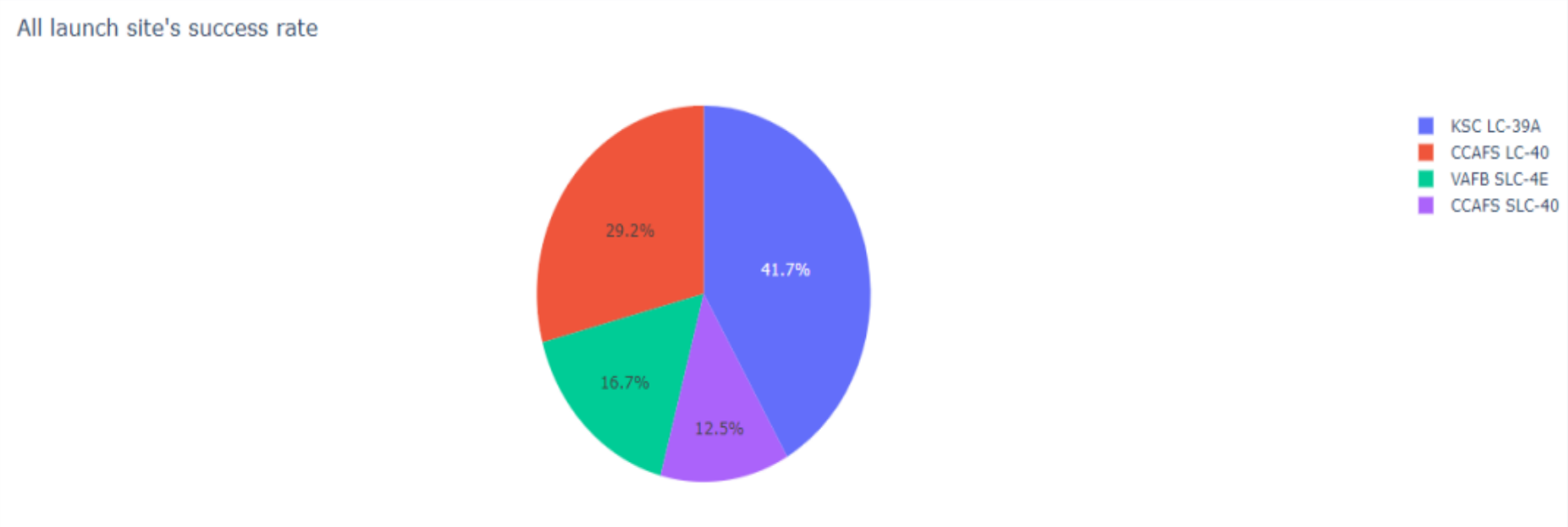





Section 4
Build a Dashboard with Plotly Dash

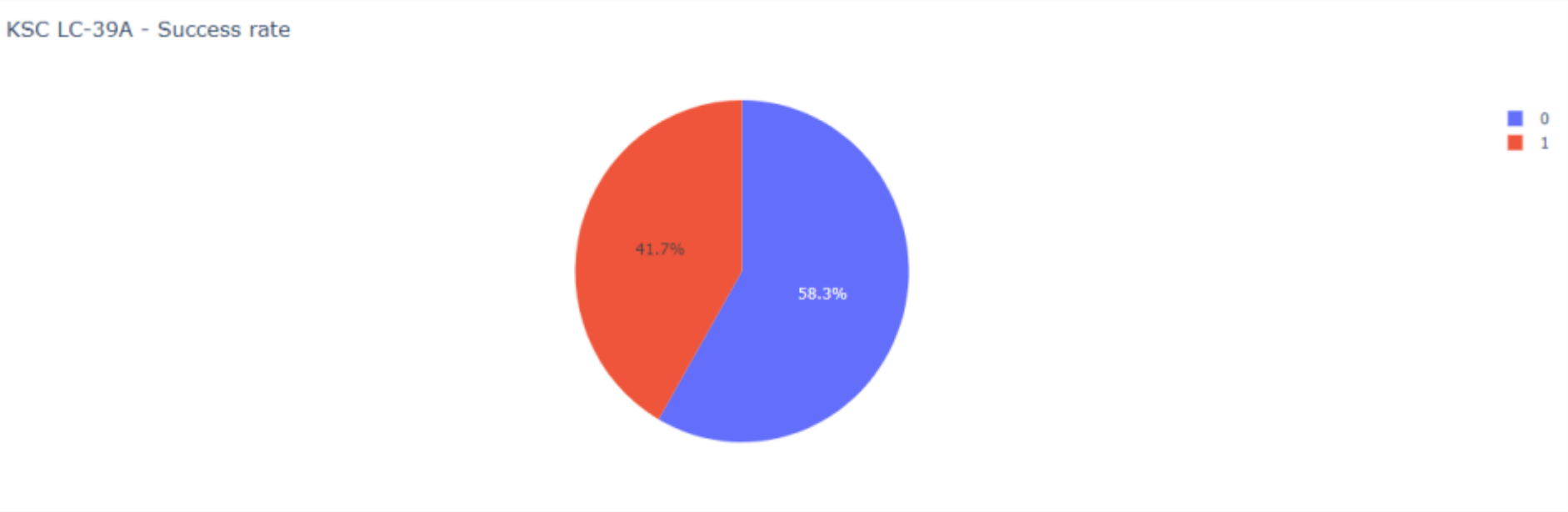
Launch success rate

- The below pie chart shows the success rate of alll the launch sites, we used plotly along with dash to create a web application to show these interactive visual graphs.



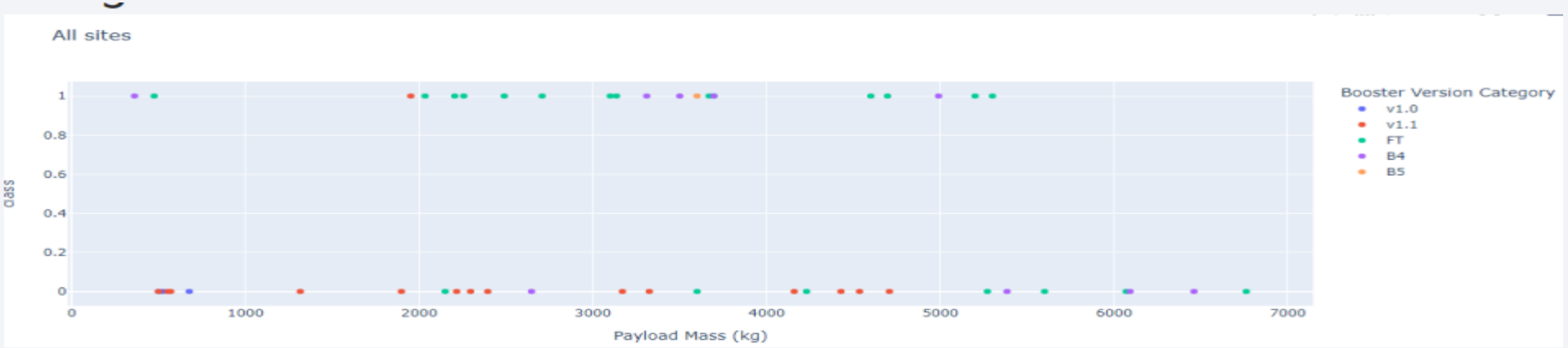
Launch site with highest success

- The below pie chart shows the success rate of the launch site KSC LC-39A which had the highest success rate of around 41.7% in comparison to all the other launch sites.



Payload vs Launch outcome

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



- From the above plot we can see that booster verison FT had the highest success rate in comparison to other booster versions.

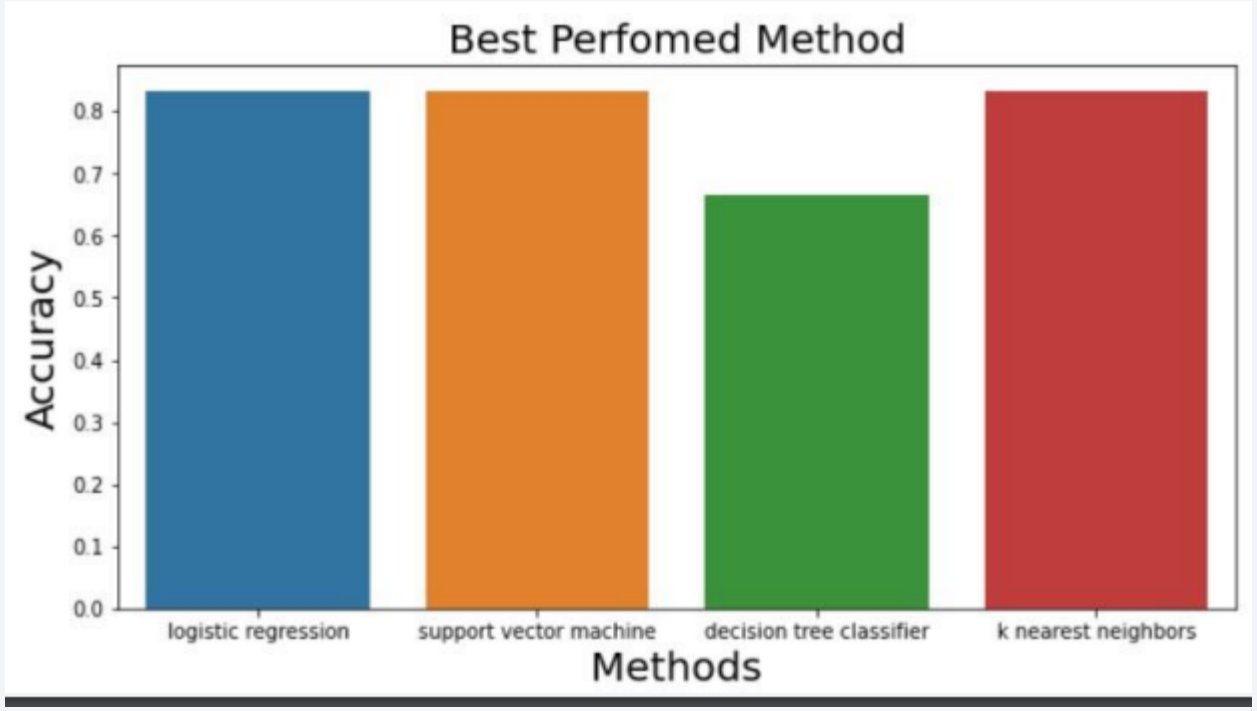
Section 5

Predictive Analysis (Classification)

Classification Accuracy

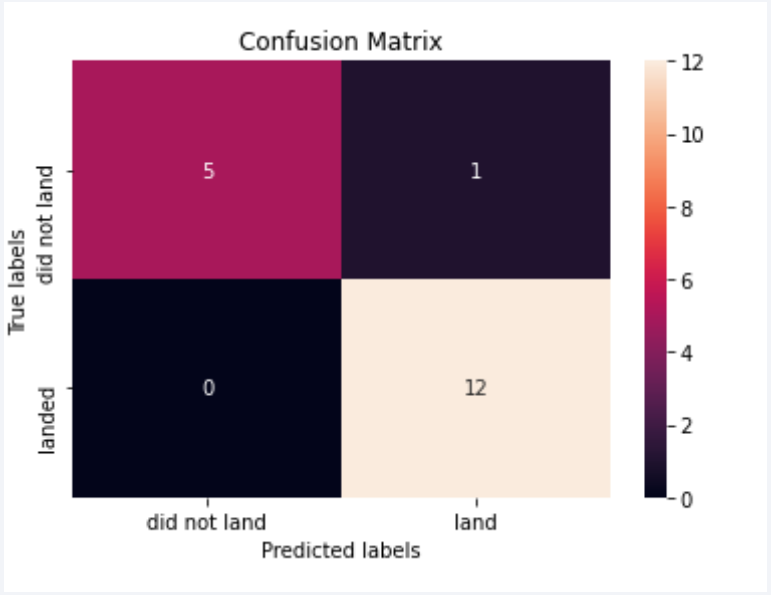
- The best performing models were logistic regression, KNN and SVM. For the outcome prediction SVM would be a good model as it's a decent model for data with high dimensions, like in our case. The model achieved a accuracy of 83% on test data

.



Confusion Matrix

- Here is the confusion matrix for SVM model.



Conclusions

- There is a good amount of progress in launch outcomes w.r.t increasing number of launches each year.
- Some sittes such as KSC LC-39A had the highest success rate in comparision to other launch sites.
- We can clearly see that space x is leading the space race, and there are some major improvments in the rate of success of landing of falcon 9 stage one.
- The success rate was also dependent on the orbit and payload mass, we saw that ISS and VLEO orbits had a good success rate.
- Support Vector Machine was a suitable model to predict if the stage one would land or not, it had an accuracy of 83%

Thank you!