IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paulo Barbosa
2022-02-27

Outline

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

Executive Summary

The goal of this project is to estimate if falcon 9 first stage will land successfully, as it plays a major role in predicting the price of its relaunch.

The data for this project used different source of information from SpaceX, 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, pie, bar 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, booter version and past experience results, among various other technical factors.

Introduction

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.

The problem that we are to answer is how can we predict the launch price of falcon 9, so this information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Predicting that whether first stage will land successfully, plays a crucial role in predicting the launch price. As that stage can be reused again with different payloads, thus greatly reducing the cost of each rocket launch.

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 BeautifulSoup.
- 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
  - Were used models SVM, logistic regression, tree classifier and k nearest neighbors.

Data Collection

The data were collected in two different ways:

1. Collected from the SpaceX rest API with endpoint: Click me , some data of certain attributes was encoded like its name so it was necessary to decoded with the help of rest api connection to the specific attributes detail's endpoint.

    - From endpoint rocket's name we get the booster name
    - From launchpad we get the launch site being used, the logitude, and the latitude
    - From payload we get the Payload mass and the orbit
    - From cores we get the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core

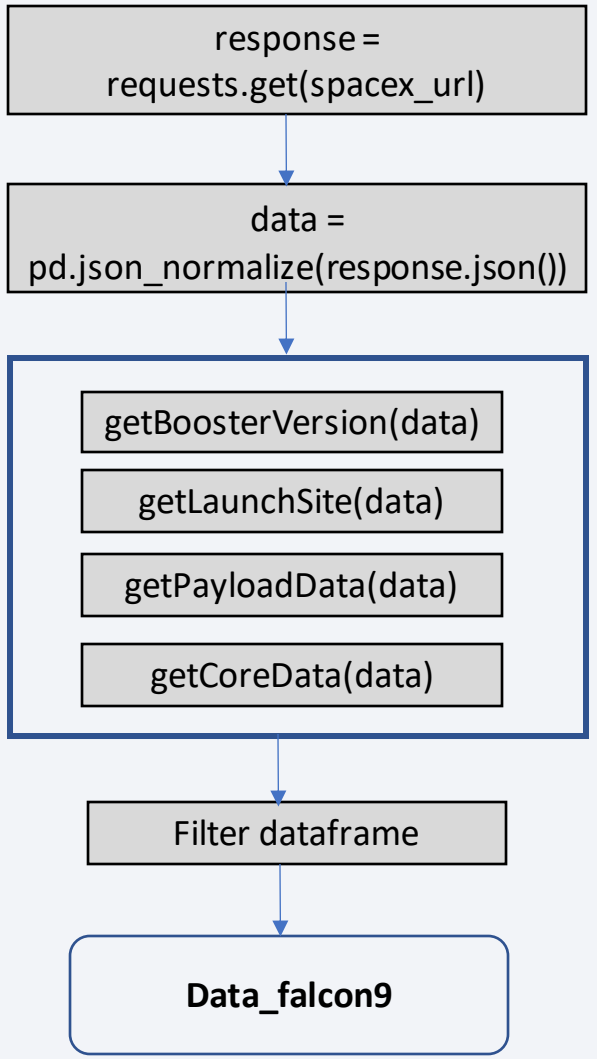
2. Collected from the webpage List of Falcon 9 and Falcon Heavy launches - Wikipedia using web scraping, with the help of BeautifulSoup framework.

Data Collection – SpaceX API

A request object was created using a rocket launch data from SpaceX API with the following URL Click me , the response content was decode a Json using .json() and turn it into a Pandas dataframe using .json_normalize().

Then From the rocket we get the booster name, from the payload we get the mass of the payload and the orbit that it is going to, from the launchpad we get the name of the launch site being used, the longitude, and the latitude, from cores we get the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used.

Github link: Data Collection API



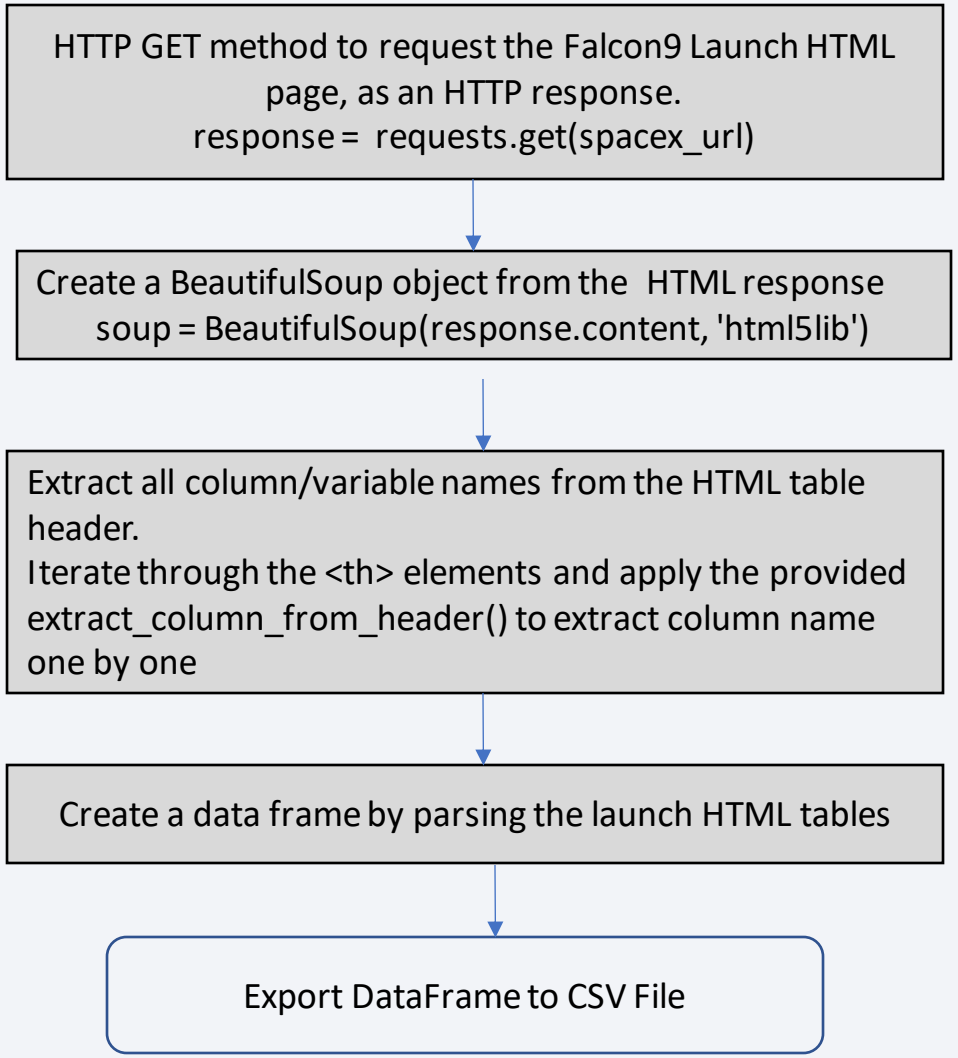
Data Collection - Scraping

Web scraping with help of BeautifulSoup was used to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches

Used Web scrap Falcon 9 launch records with BeautifulSoup:

- To extract a Falcon 9 launch records HTML table from Wikipedia
- To Parse the table and convert it into a Pandas data frame

Github link: Data Collection with Web Scraping



Data Wrangling

- After the data were filtered to contain only the records of falcon 9, it was found that there were some null values in the dataset. Before we can continue we must deal with these missing values. The LandingPad column will retain None values to represent when landing pads were not used. Null values in the column PayloadMass will be replaced by the mean value. See section Data Wrangling on GitHub link Collecting Data

- One hot encoding was done for some categorical variables 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 link: Data-Wrangling

EDA with Data Visualization

Scatter plots, bar chart and line chart was created to visualize several relationships.

- Was visualized the relationship between Flight Number and PayloadMass, between Flight Number and Launch Site, between Payload and Launch Site, between success rate of each orbit type (Bar chart), between FlightNumber and Orbit type, between Payload and Orbit type and visualized the launch success yearly trend (Line Chart)

Some conclusion of the visualization were:

- As the flight number increases, the first stage is more likely to land successfully. It seems the more massive the payload, the less likely the first stage will return.
- Was observed Payload Vs. Launch Site scatter point chart and find that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- We 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.
- 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.
- We observe that the sucess rate since 2013 kept increasing till 2020

Github link: EDA with Visualization

EDA with SQL

Several queries to SQL SpaceX dataset was realized to explore and understand the data.

The dataset includes a record for each payload carried during a SpaceX mission into outer space, was made several queries.

The Dataset queries allow us to understand for instance that:

- *Was used 4 launch sites*
- *The total payload mass carried by boosters launched*
- *The average payload mass carried*
- *The date where the first succesful landing outcome in drone ship was achieved*
- *List the total number of successful and failure mission outcomes*
- *List the names of the booster_versions which have carried the maximum payload mass.*

GitHub link: EDA With SQL

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 markers for all launch records. If a launch was successful (`class=1`), then we use a green marker and if a launch was failed, we use a red marker (`class=0`)

At the end a polyline object was added to the map to show the distance of the launch site from its proximities such as closest city, railway, highway, etc.

GitHub link: Launch Sites Locations Analysis with Folium

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 allow the user to visualize the data with more flexibility.

GitHub link: SpaceX Plotly Dash

Predictive Analysis (Classification)

Various types of ML model was used with in addition with GridSearchSV. The Machine Learning algorithms used were:

- Logistic regression
- SVM
- Decision tree
- KNN

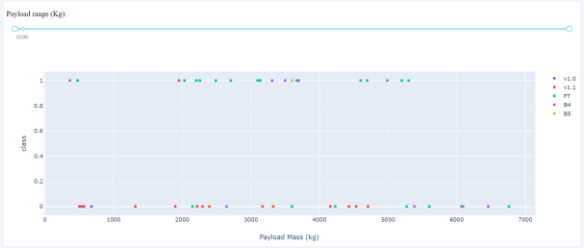
	AUC	F1-Score	Precision	Recall	Accuracy
SVM	0.958	0.889	0.8	1.00	0.833
KNN	0.896	0.889	0.8	1.00	0.833
Logistic Regression	0.889	0.889	0.8	1.00	0.833
Decision Tree	0.792	0.818	0.9	0.75	0.778

The summary result of the different ML model, allow to conclude that the model with best out of sample accuracy were KNN, logistic regression and SVM. The R2 score was around .83

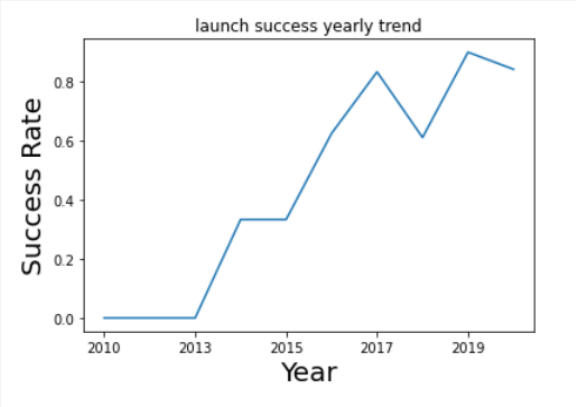
GitHub link: Machine Learning Prediction

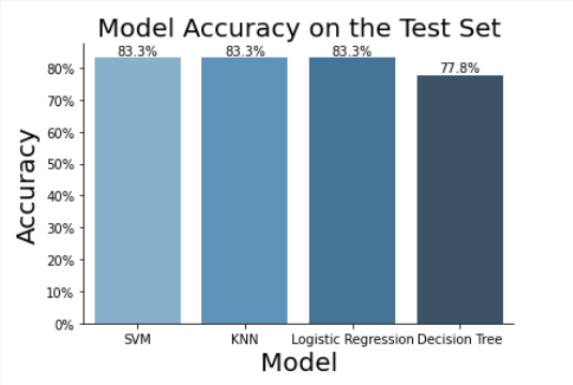
Results

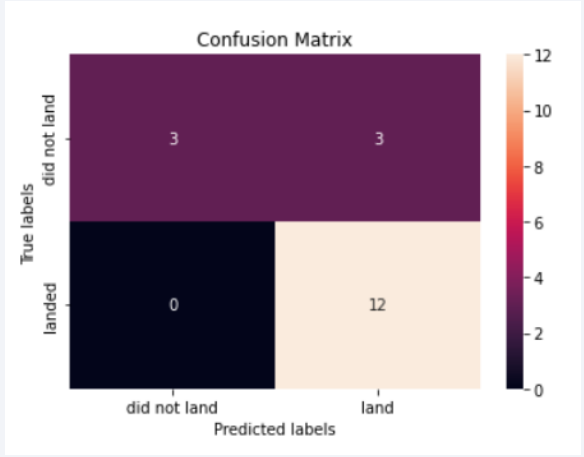
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



	AUC	F1-Score	Precision	Recall	Accuracy
SVM	0.958	0.889	0.8	1.00	0.833
KNN	0.896	0.889	0.8	1.00	0.833
Logistic Regression	0.889	0.889	0.8	1.00	0.833
Decision Tree	0.792	0.818	0.9	0.75	0.778

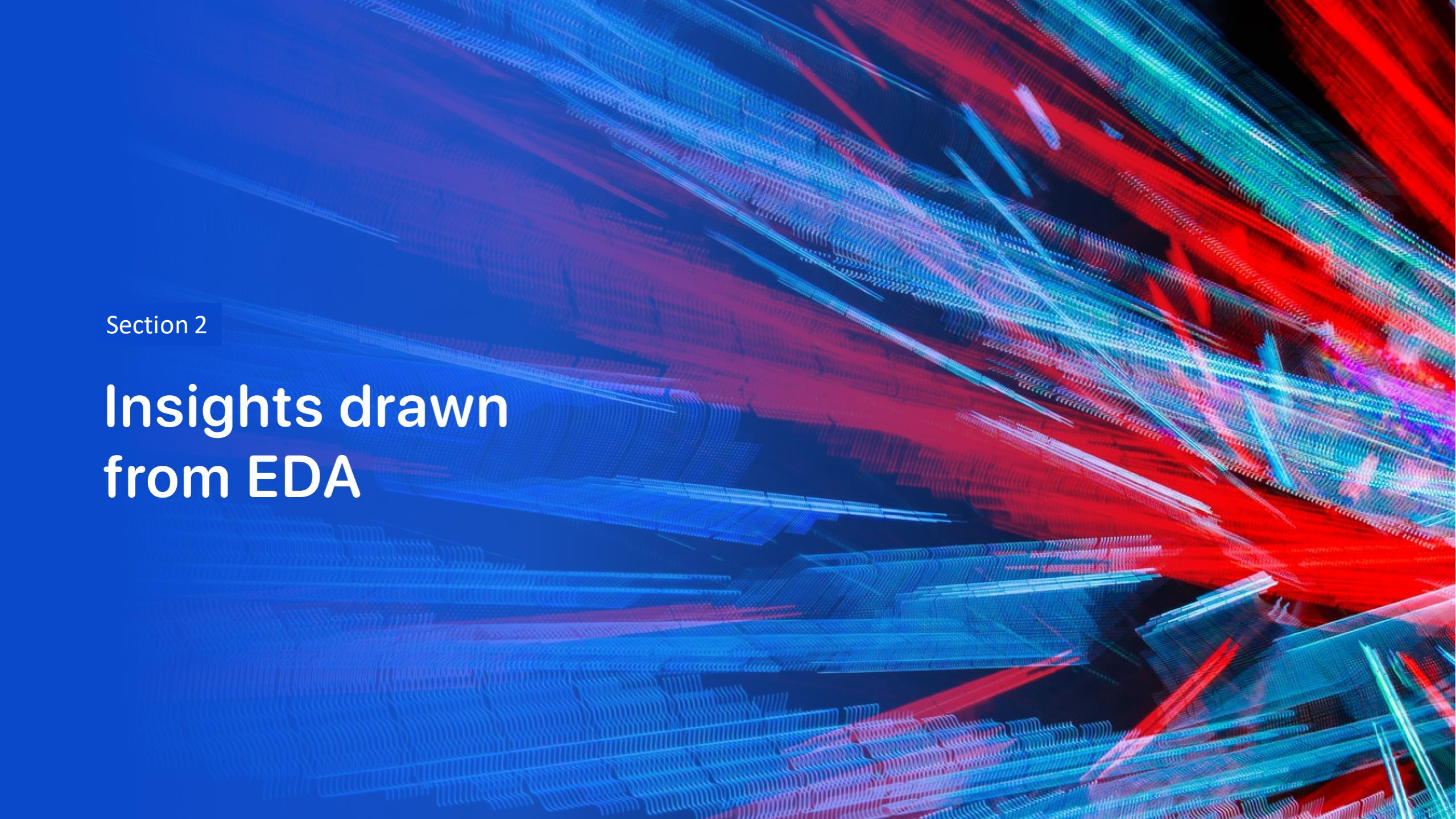




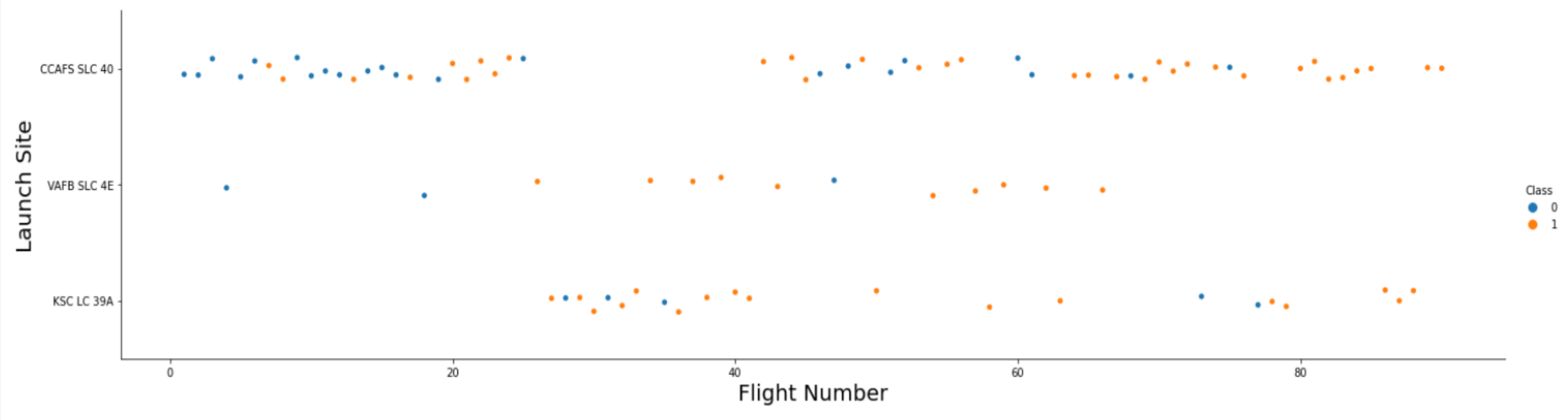


Section 2

Insights drawn from EDA

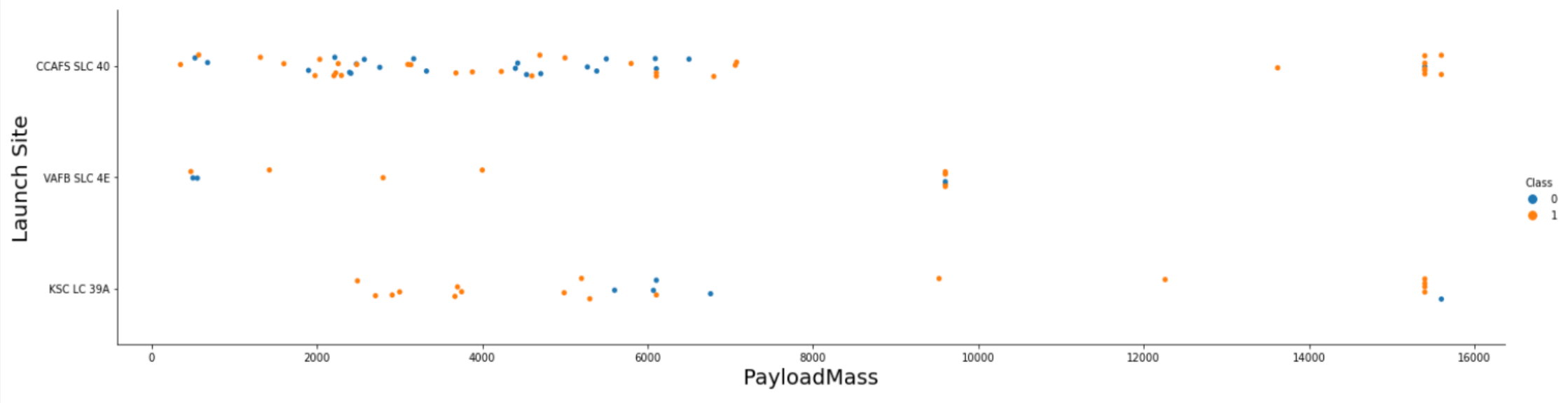


Flight Number vs. Launch Site



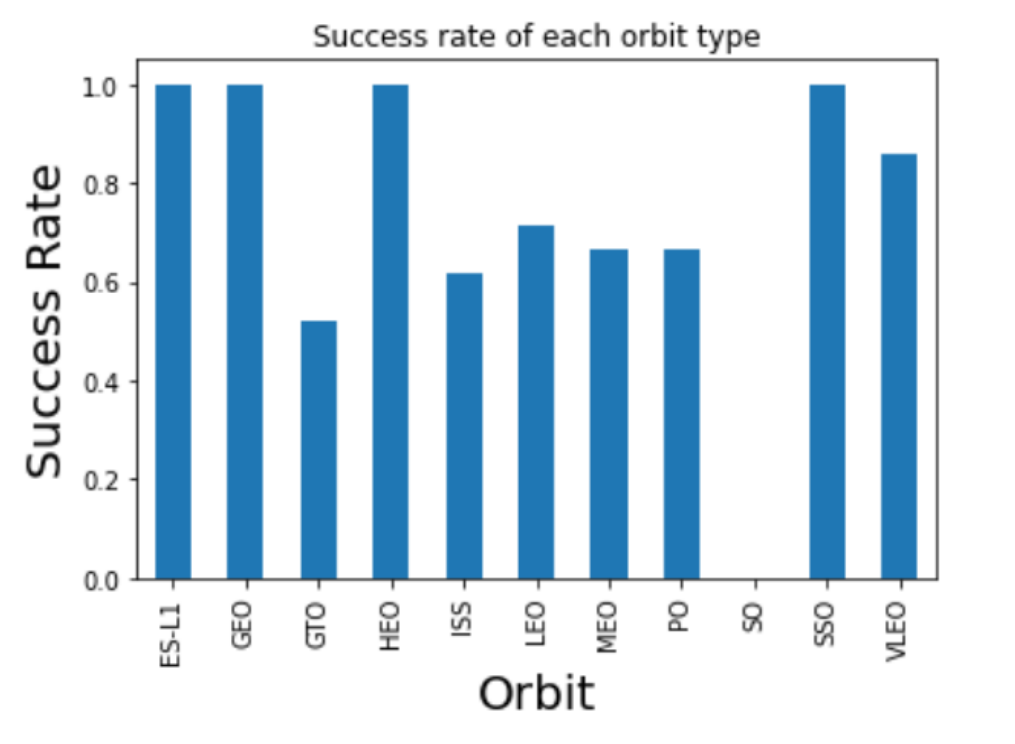
We see that different launch sites have different success rates.
CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
If we split in two intervals the flight Numbers [0-40] and [41-100] for CCAFS LC-40, we verify that in the second interval the success rate is almost the same as the other Launch sites

Payload vs. Launch Site



We observe that for the VAFB-SLC launch site there are no rockets launched for heavypayload mass (greater than 10000).

Success Rate vs. Orbit Type

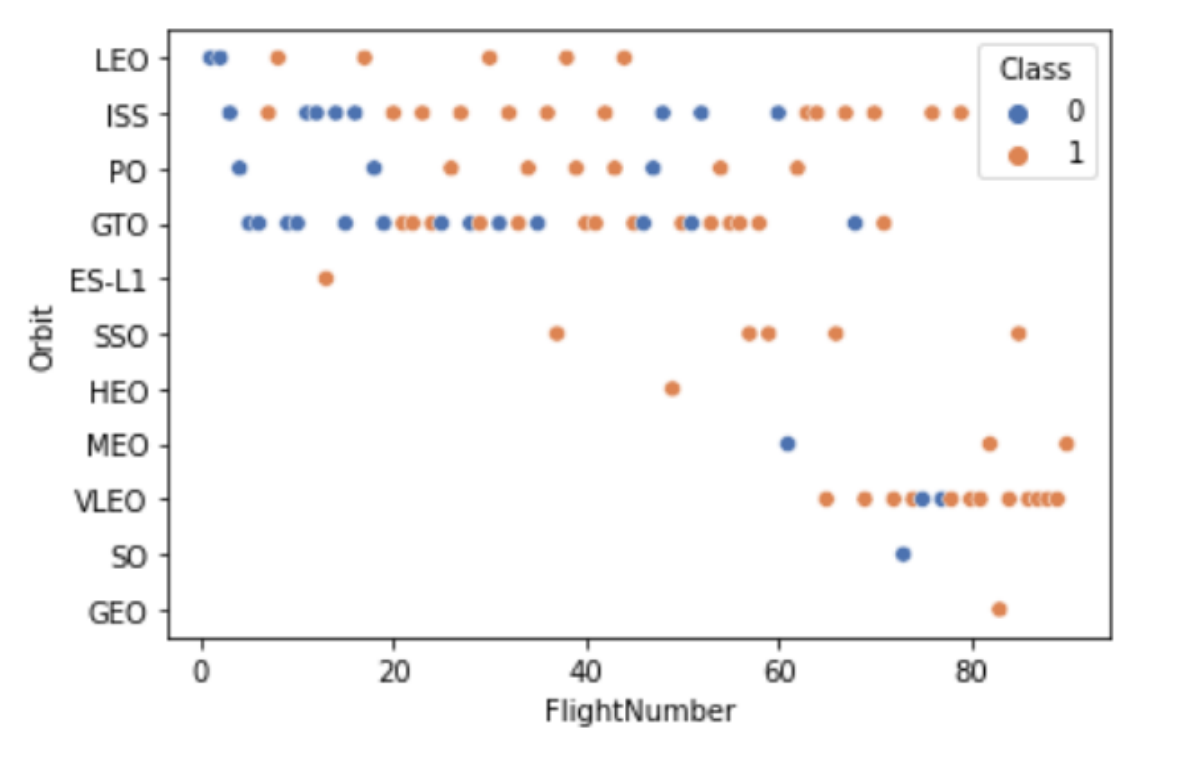


We observe that the most successful Orbits are the ES-L1, GEO, HEO and SSO with 100% Success rate. The less successful Orbits are SO (0%) and GTO (60%).

Meanwhile we should carefully analyze this values, because the number of some orbits as low value (ex. ES-l1, GEO, HEO and SO as 1) and therefore no statistical meaning.

Having in consideration that information, the most successful Orbit is SSO with 100% Success rate

Flight Number vs. Orbit Type

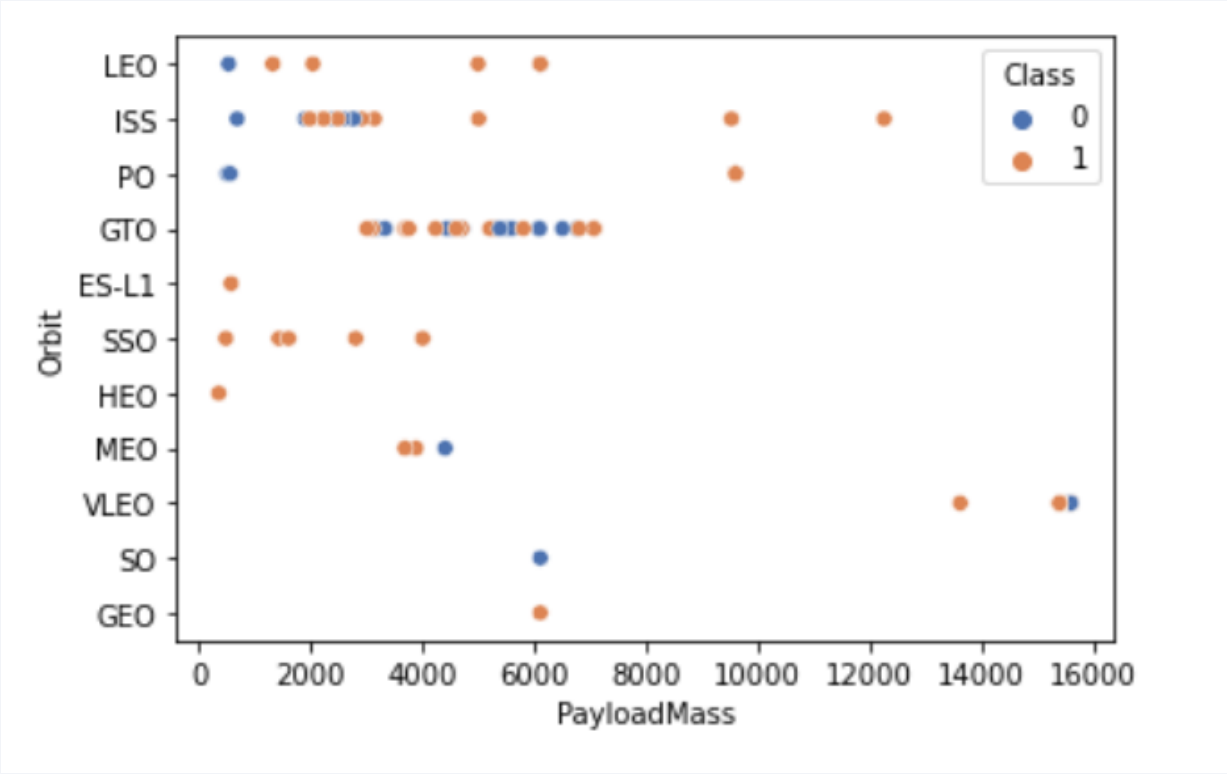


We see that in the LEO orbit the Success appears related to the number of flights.

On the other hand, there seems to be no relationship between flight number when in GTO orbit

We also verify that SSO Orbit the Success is 100%

Payload vs. Orbit Type

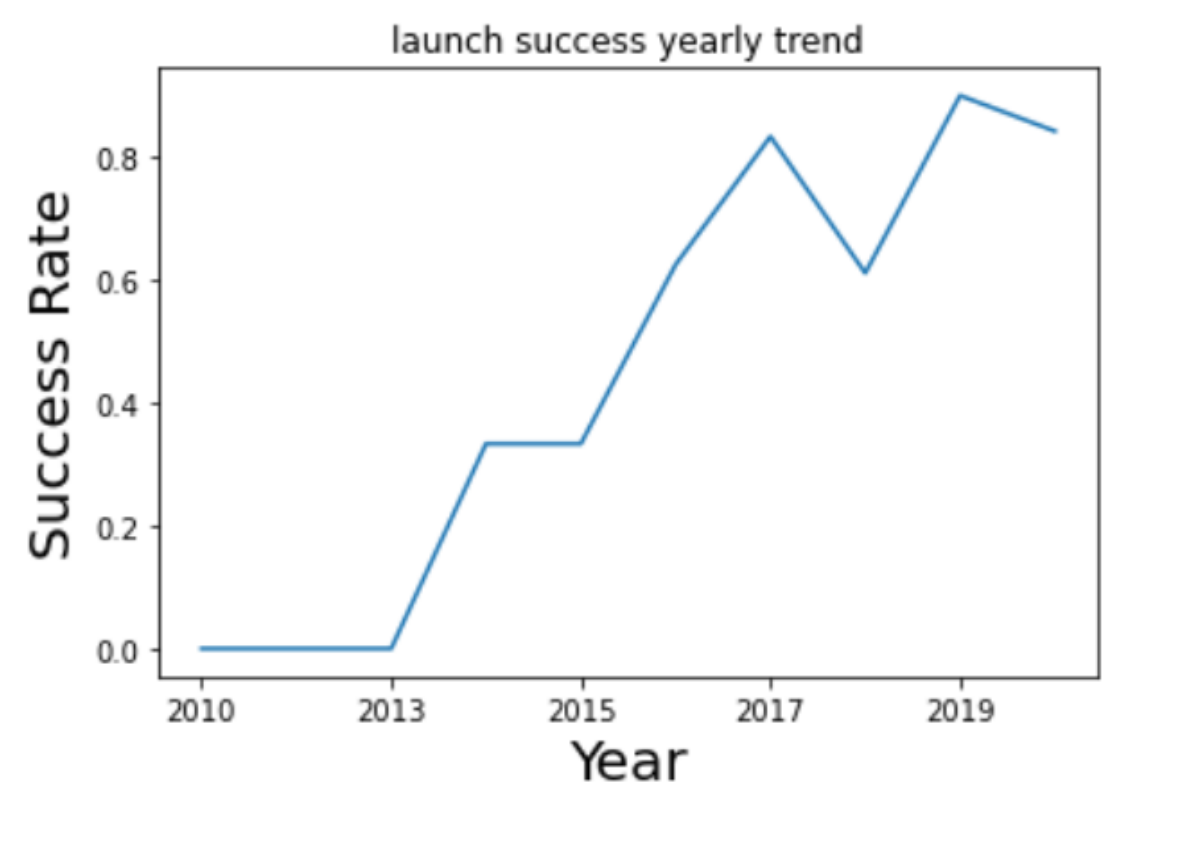


With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

With not heavy payload the successful landing or positive landing rate are for SSO

Launch Success Yearly Trend



We observe that the sucess rate since 2013 kept increasing till 2020, reaching values of success rate over 80%

All Launch Site Names

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

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

Select all distinct values from column LAUNCH_SITE from database SPACEXTBL

Launch Site Names Begin with 'KSC'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

```
%sql SELECT * FROM SPACEXTBL WHERE upper(LAUNCH_SITE) LIKE 'KSC%' LIMIT 5
```

Select all records from database SPACEXTBL where values from column LAUNCH_SITE start with 'KSC' and limit the result to 5 records (LIMIT 5)

Total Payload Mass

payload_mass_by_nasa
45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Payload_Mass_By_Nasa FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

Select and sum all values from column PAYLOAD_MASS_KG_from database SPACEXTBL where values from column CUSTOMER is equal to 'NASA (CRS)'

Average Payload Mass by F9 v1.1

avg_payload
2534

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE upper(BOOSTER_VERSION) LIKE 'F9 V1.1%'
```

Select and calculate the average of all values from column PAYLOAD_MASS_KG_from database SPACEXTBL where values from column BOOSTER_VERSION start with 'F9 V1.1'

First Successful Ground Landing Date

mindate
2016-04-08

```
%sql SELECT min(DATE) AS MinDATE FROM SPACEXTBL WHERE upper(LANDING__OUTCOME) LIKE 'SUCCESS (DRONE SHIP)%'
```

Select Minimum date from column DATE from database SPACEXTBL where values from column LANDING_OUTCOME start with 'SUCCES (DRONE SHIP)'

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE upper(LANDING__OUTCOME) LIKE 'SUCCESS (GROUND PAD)%' \
AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

Select values from column BOOSTER_VERSION from database SPACEXTBL where values from column LANDING_OUTCOME start with 'SUCCES (GROUND PAD)' and values from column PAYLOAD_MASS_KG_ are between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

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

```
%sql SELECT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) AS TOT FROM SPACEXTBL GROUP By MISSION_OUTCOME;
```

Select values from column MISSION_OUTCOME from database SPACEXTBL, Arrange the rows of the query in groups of the column MISSION_OUTCOME and count them per each group

Boosters Carried Maximum Payload

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

Select values from column BOOSTER_VERSION from database SPACEXTBL where PAYLOAD_MASS_KG_ is equal to maximum value retrieve from subquery which select the maximum payload from column PAYLOAD_MASS_KG_

2017 Launch Records

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_YEAR_2017, \
LANDING__OUTCOME, \
BOOSTER_VERSION, \
LAUNCH_SITE \
FROM SPACEXTBL WHERE upper(LANDING__OUTCOME) LIKE 'SUCCESS (GROUND PAD)%' AND YEAR(DATE)='2017';
```

month_year_2017	landing__outcome	booster_version	launch_site
FEBRUARY	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
MAY	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
JUNE	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
AUGUST	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
SEPTEMBER	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
DECEMBER	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

Select values from column DATE (converted to month name), LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from database SPACEXTBL where LANDING_OUTCOME starts with 'SUCCESS (GROUND PAD)' and YEAR is equal to '2017'

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

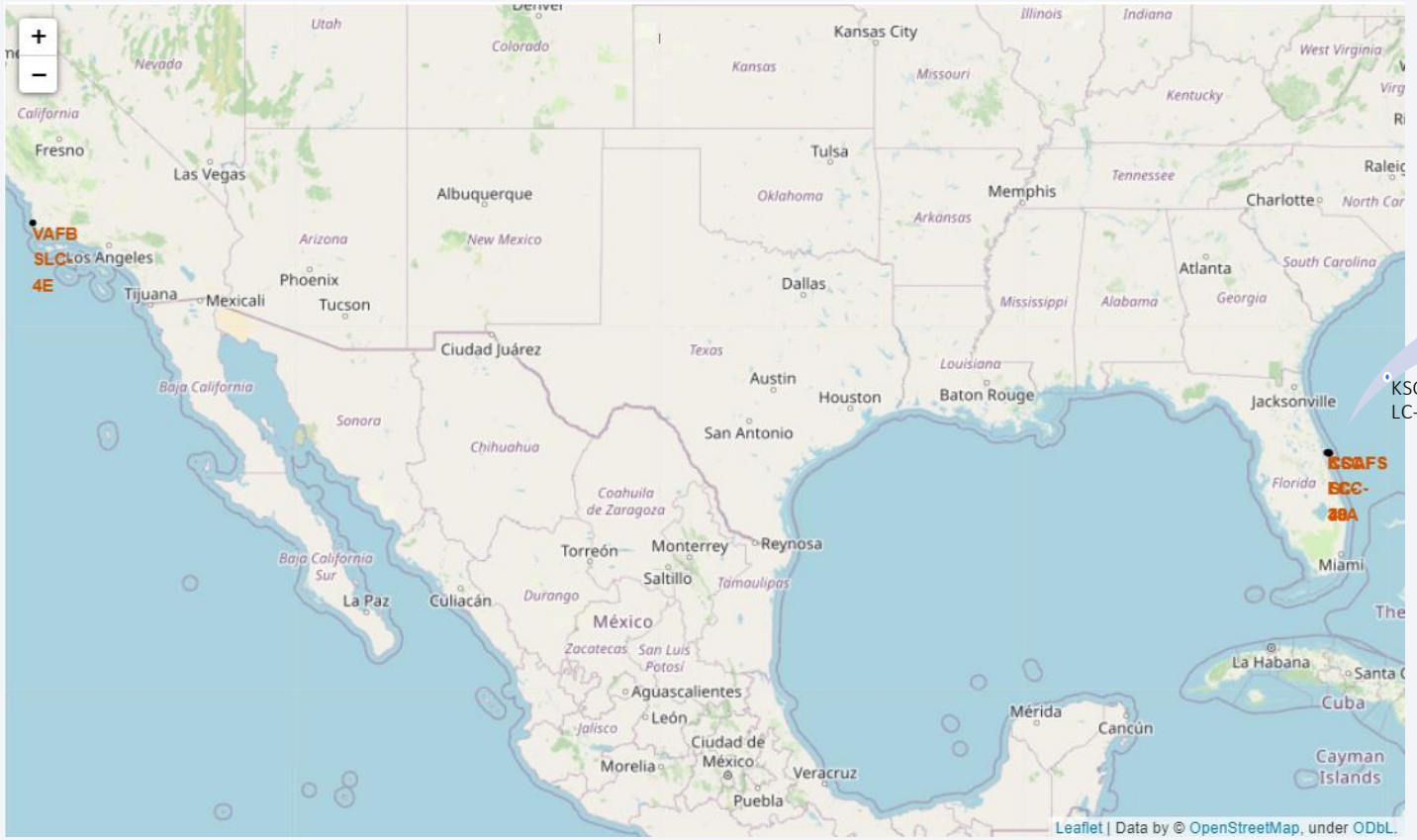
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

```
%sql SELECT * from SPACEXTBL \
WHERE upper(LANDING__OUTCOME) LIKE 'SUCCESS%' AND (DATE BETWEEN '2010-06-04' AND '2017-03-20') \
ORDER BY DATE DESC
```

Select all records from database SPACEXTBL where LANDING_OUTCOME starts with 'SUCCESS' and DATE between 2010-06-04 and 2017-03-20 In descending order DATE


Section 3
Launch Sites
Proximities Analysis

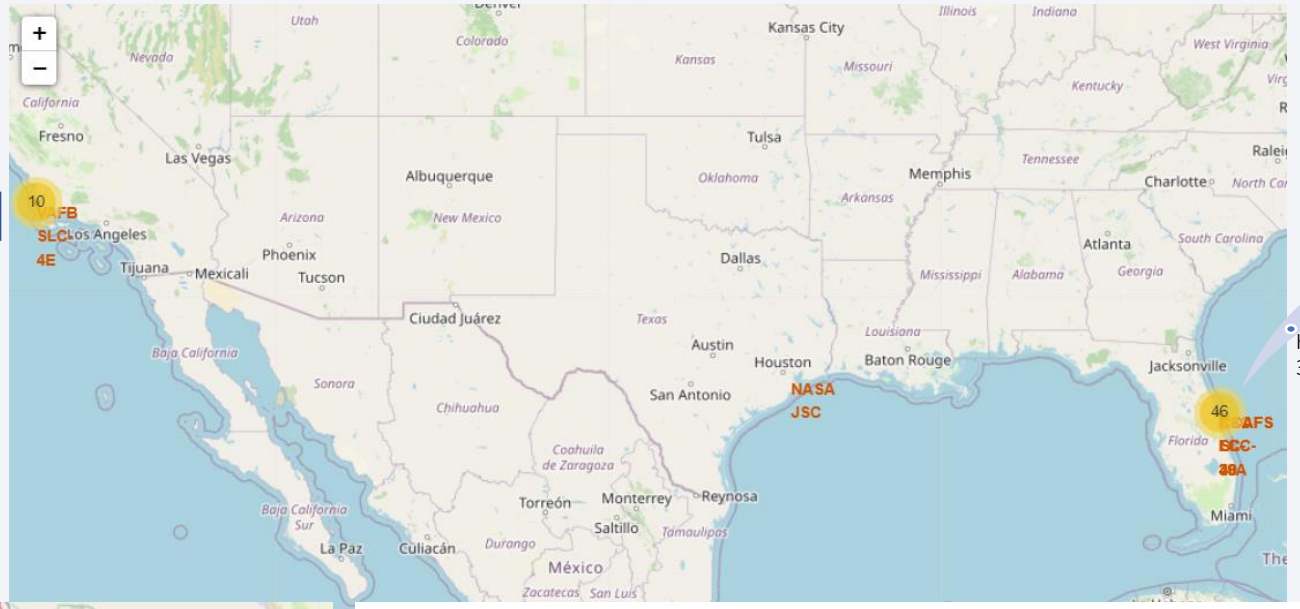
Launch Sites Locations On Global MAP

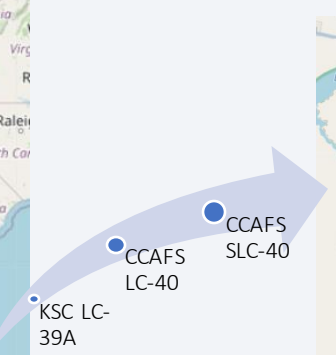


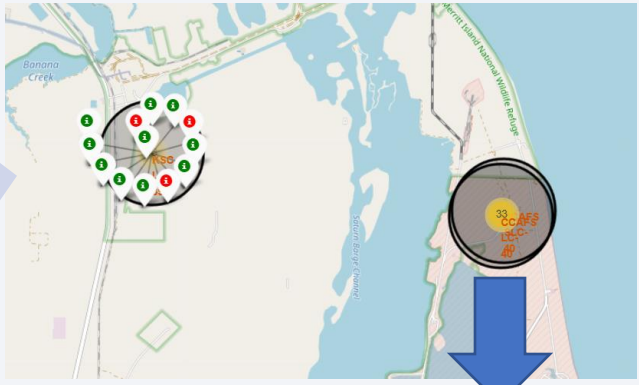
SPACEX use 4 Launch Sites, the VAFB SLC-4E is located near Santa Barbara (West Coast of USA) the others 3 (KSC LC-39A, CCAFS LC-40, CCAFS SLC-40) are very closed to each other, located on Cape Canaveral (Florida)

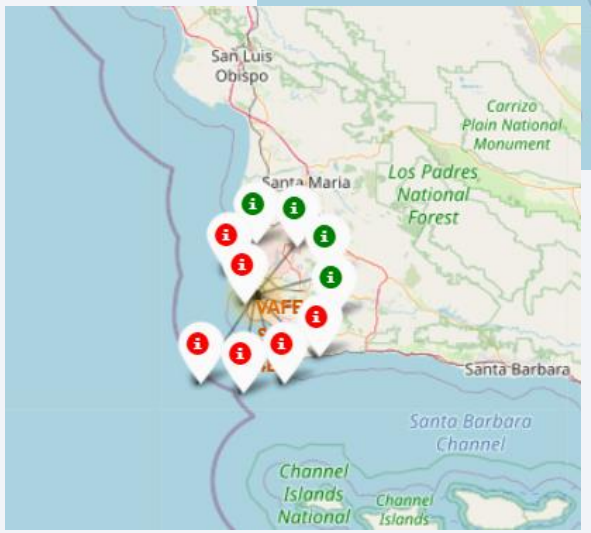
Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

Success/Failed Launches For Each Site



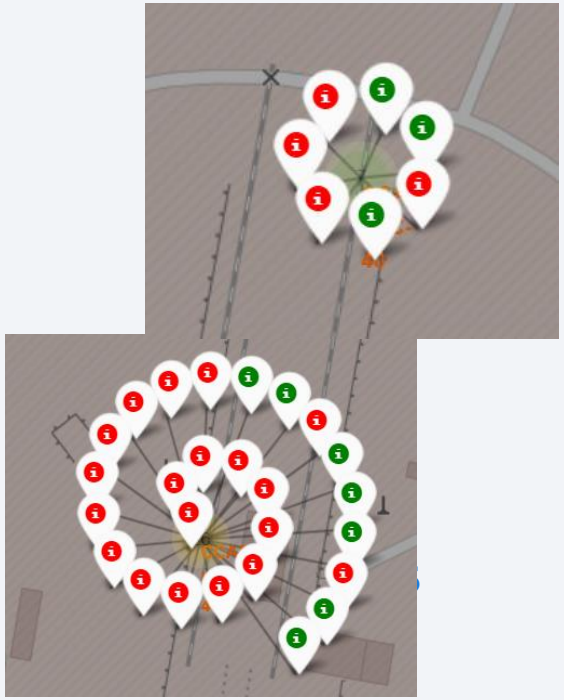




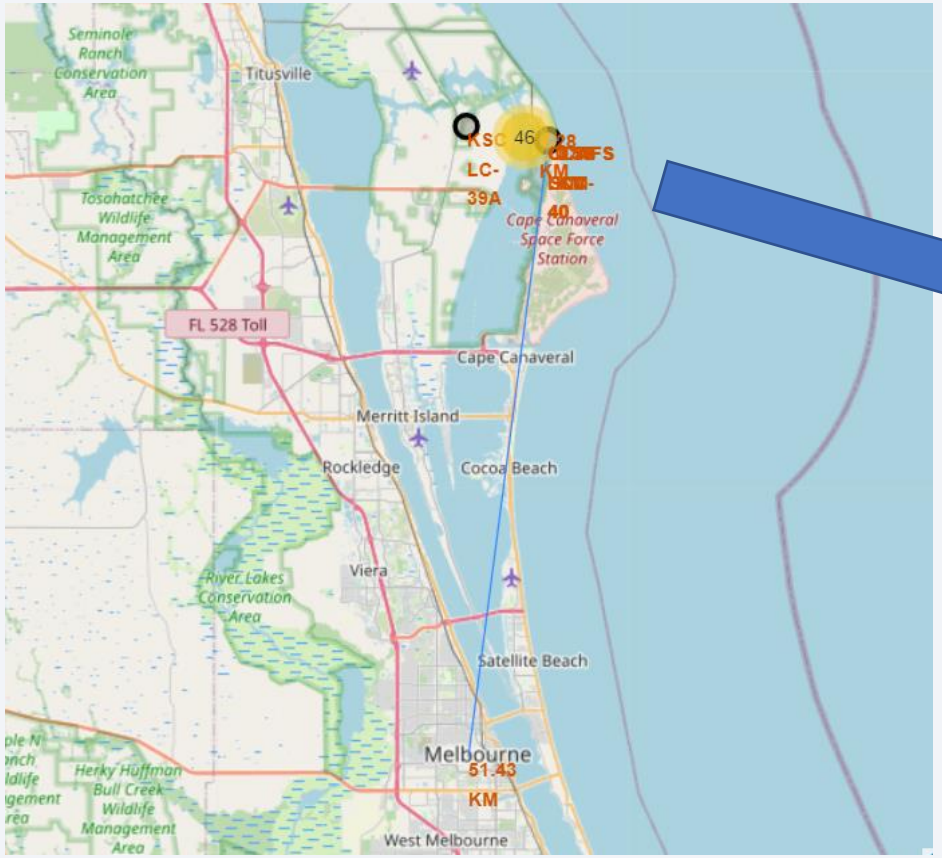


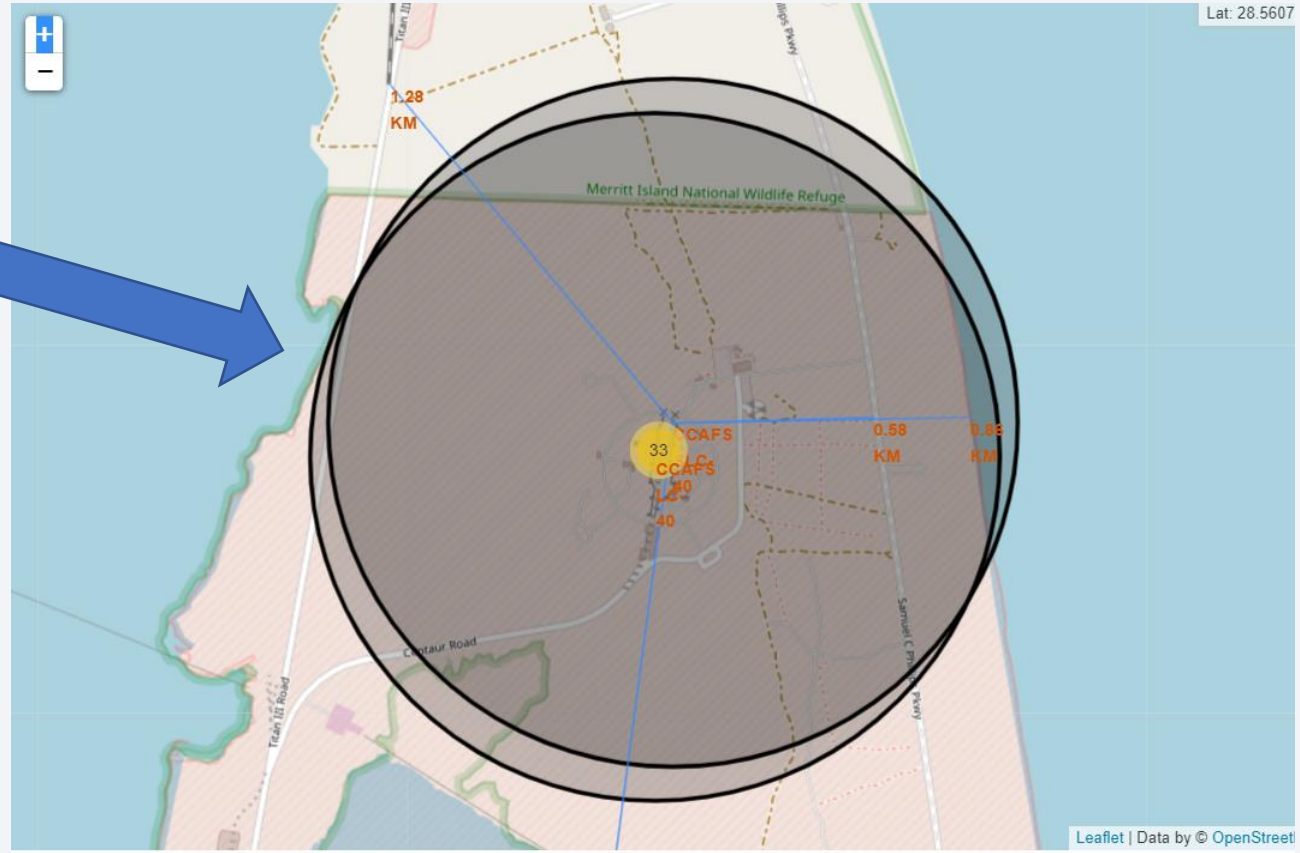
- VAFB SLC-4E 10 Lauches (4 Success and 6 Insuccess)
- KSC LC-39A 13 launches (10 Success and 3 Insucess)
- CCAFS LC-40 7 Launches (3 Success and 4 Insuccess)
- CCAFS SLC-40 26 launches (7 Success and 19 Insuccess)

The best perfomace launch site is KSC LC-39A with 77% Success
The worts perfomance launch site is CCAFS SLC-40 with only 27% Success

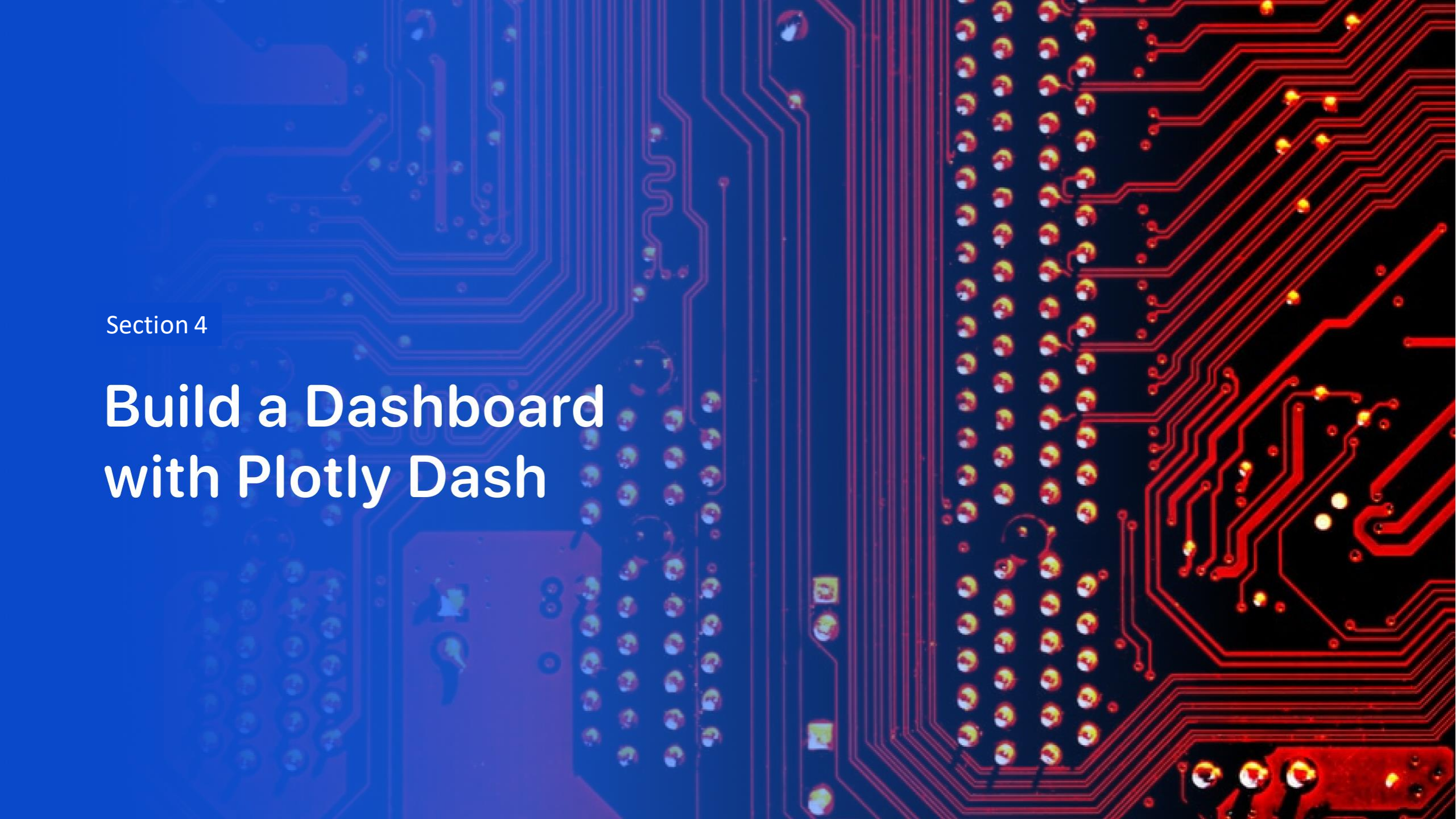


Distances Between a Launch Site To Its Proximities

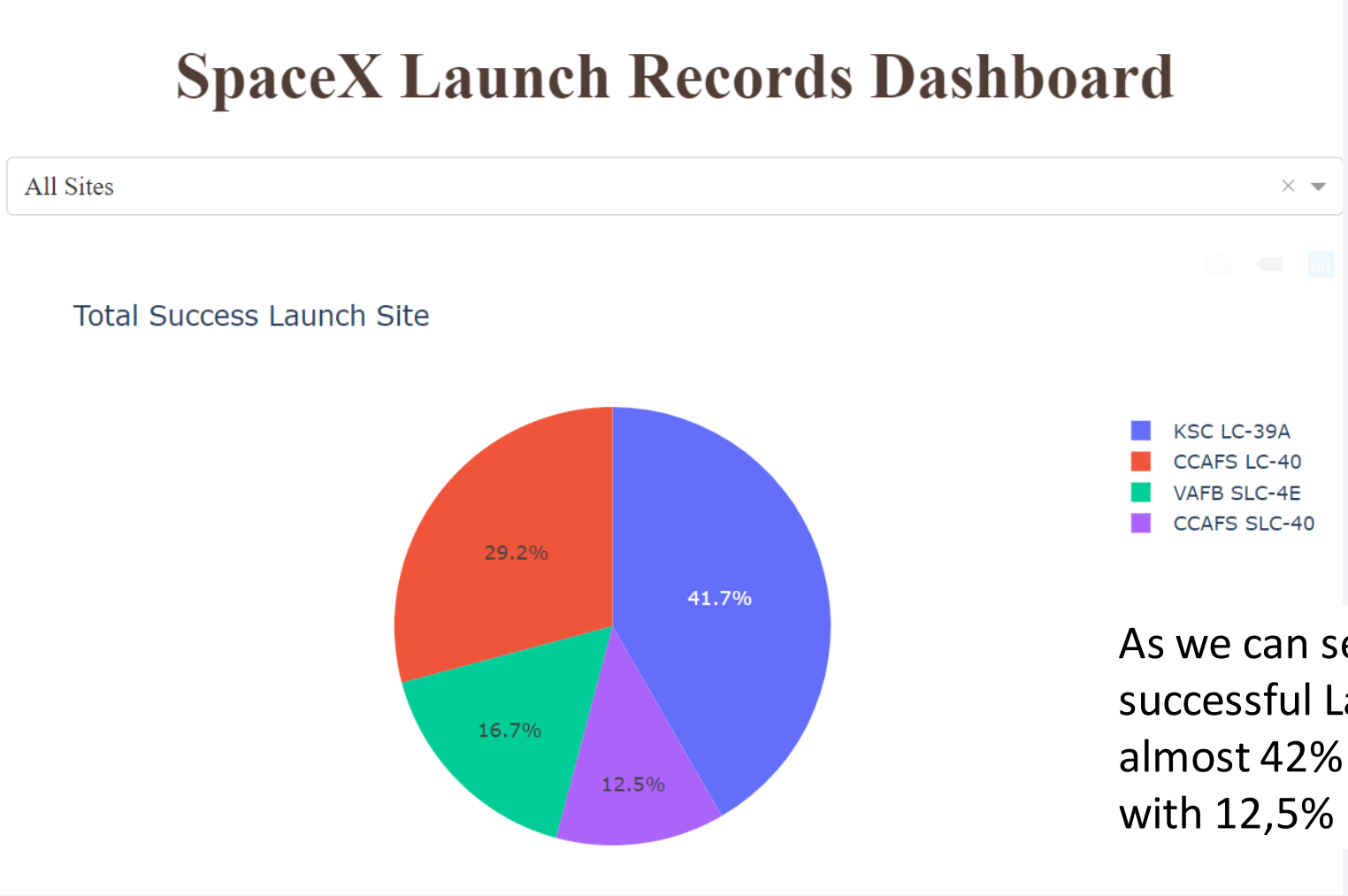




Distance closest coastline = 0.8627671182499878 km
Distance_highway = 0.5834695366934144 km
Distance_railroad = 1.2845344718142522 km
Distance_city = 51.43416999517233 km

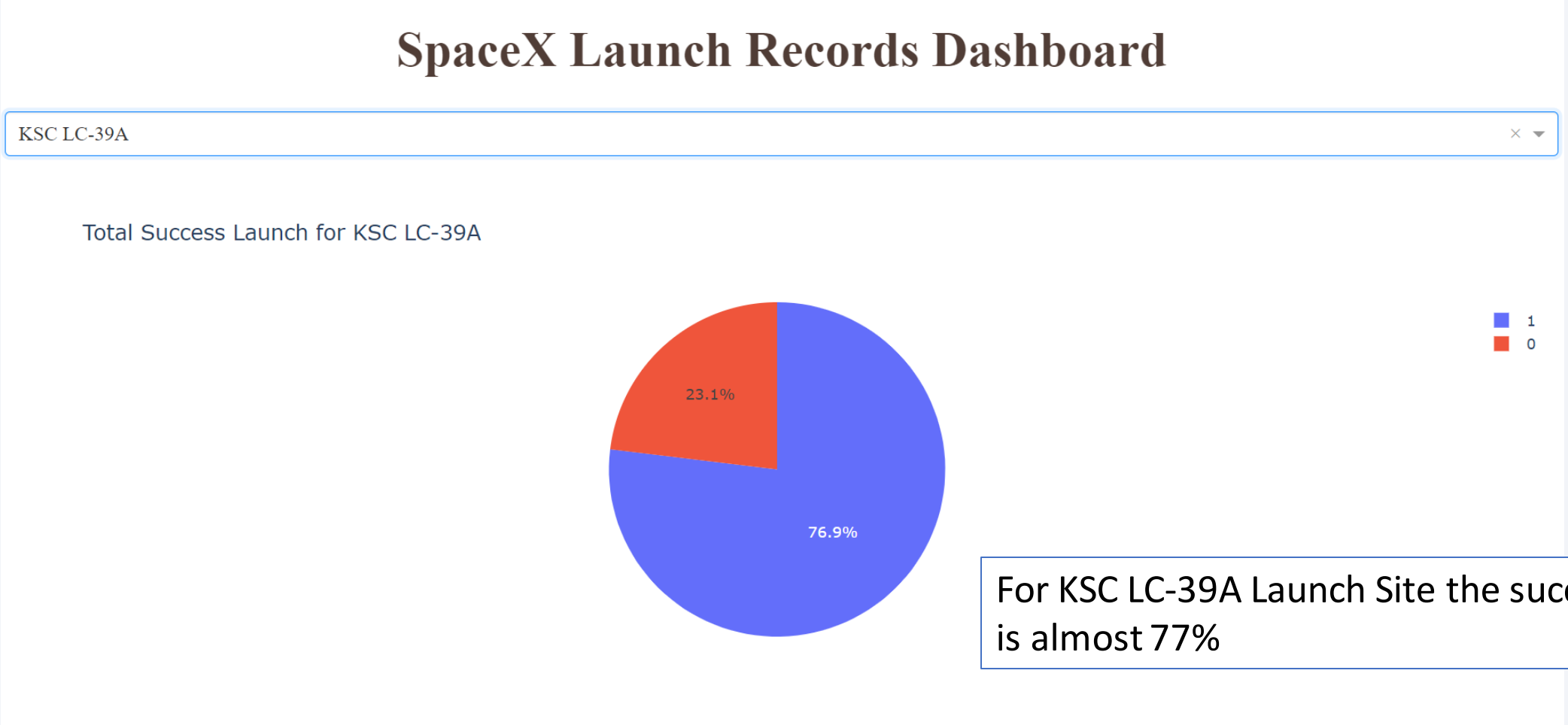

Section 4
Build a Dashboard with Plotly Dash

Success Launch Per Site



As we can see in the Pie Chart, the most successful Launch Site is KSC LC-39A with almost 42% and the worst is the CCAFS SLC-40 with 12,5%

Success Launch for KSC LC-39A



For KSC LC-39A Launch Site the success rate is almost 77%

Payload Vs Lauch Outcome Per Booster Version

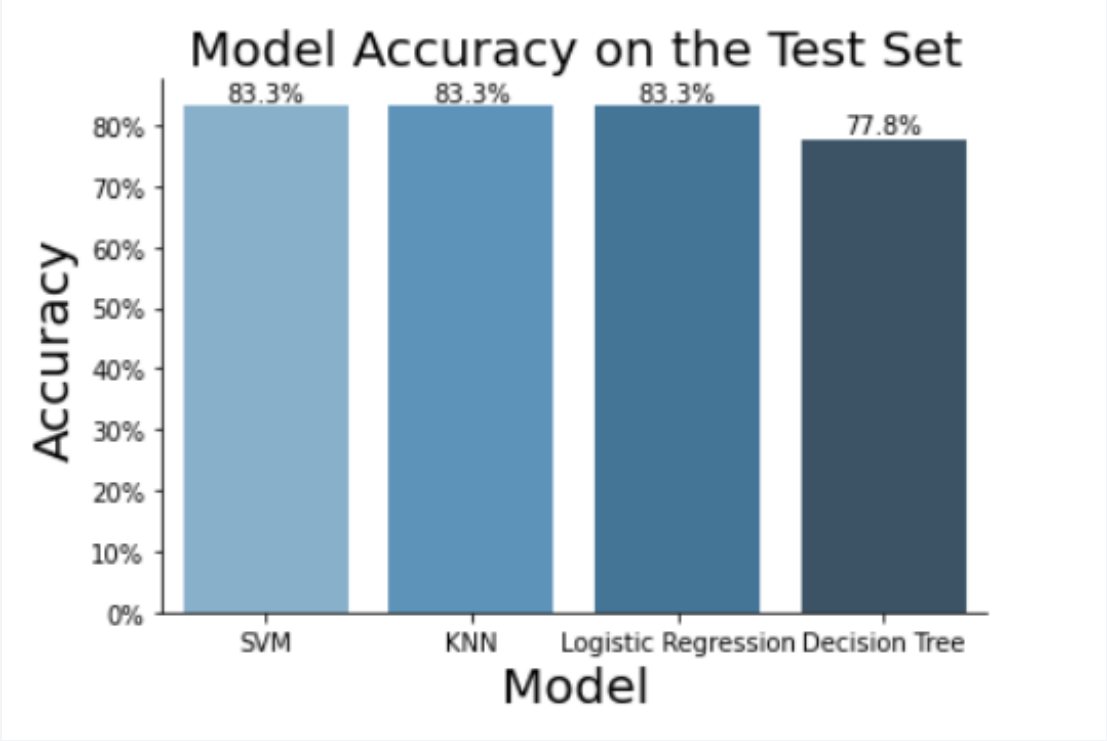
Payload minor than 5500 are much more success than higher Payload

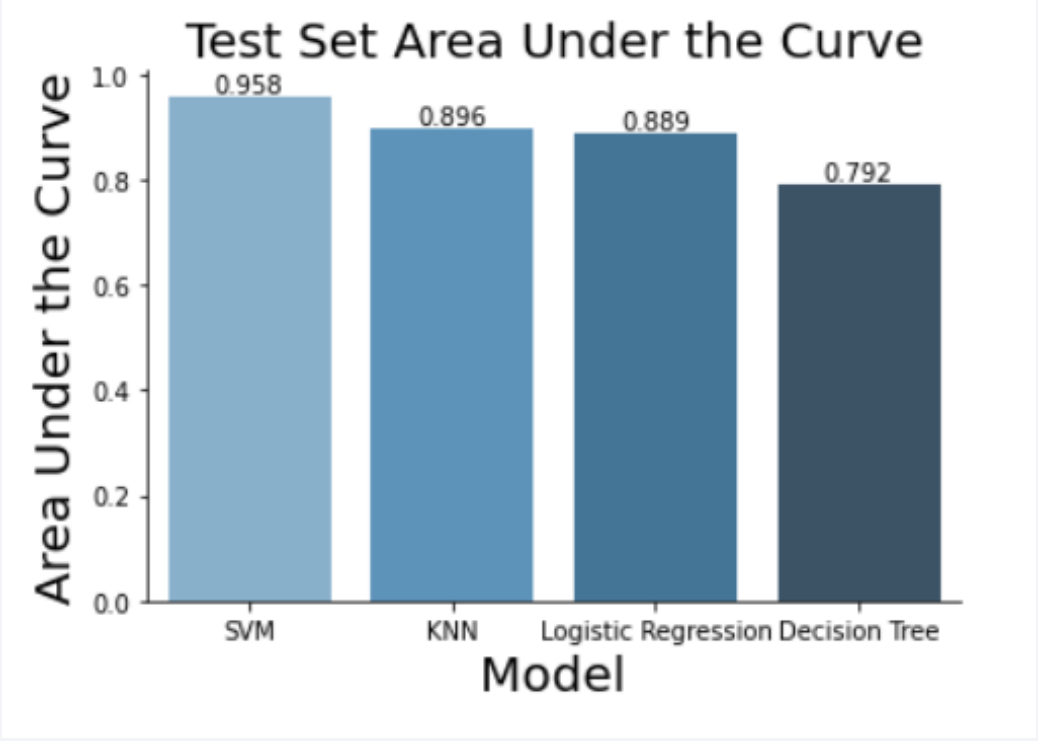
Booster Version FT are more successful and the V1.1 are worst success performance

Section 5

Predictive Analysis (Classification)

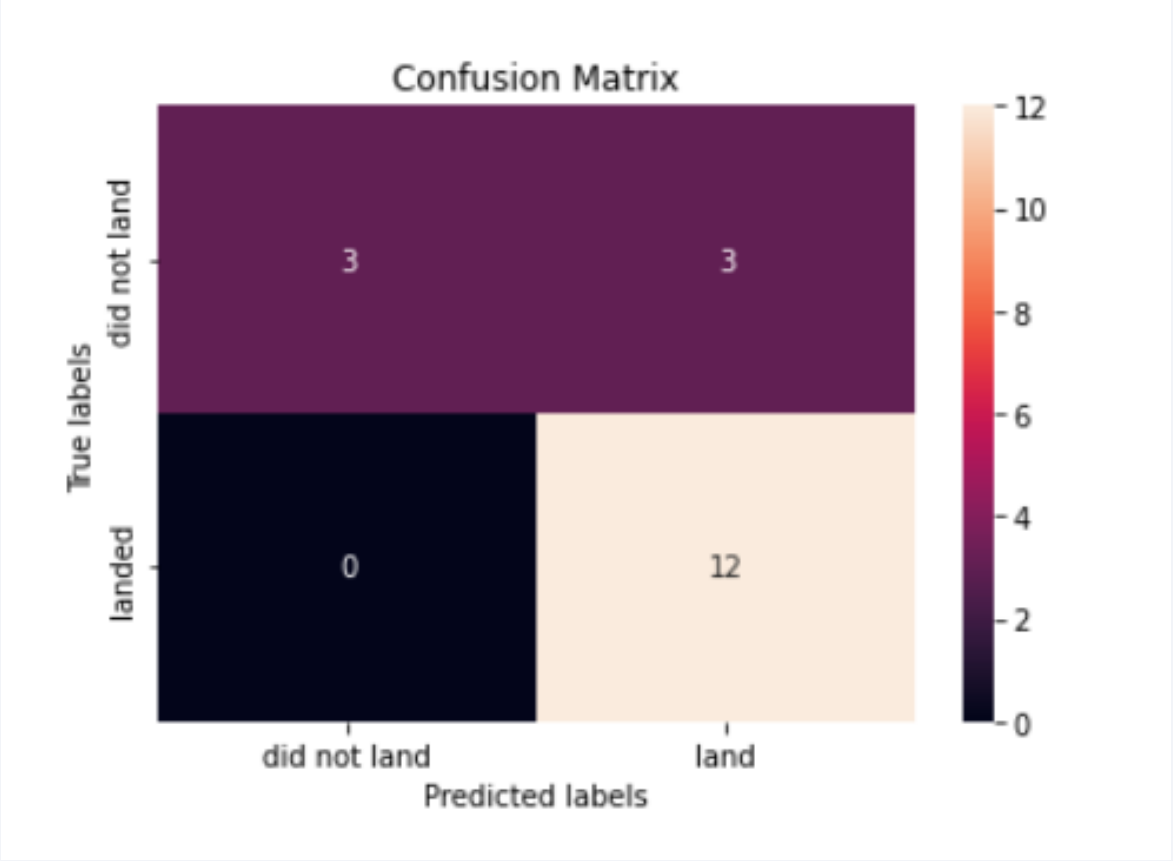
Classification Accuracy





The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958.

Confusion Matrix



•The best performing models were logistic regression, KNN and SVM that achieve the highest accuracy at 83,3%.

•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. Besides that SVM performs the best in terms of Area Under the Curve at 0,958

Conclusions

1. There is a good amount of progress in launch outcomes, we observe that the sucess launch rate since 2013 kept increasing till 2020, reaching values of success rate over 80%

2. Some sites such as KSC LC-39A had the highest success rate in comparison to other launch sites.

3. The success rate was also dependent on the orbit, payload mass and Booster Version.
   - Payload minor than 5500 are much more success than higher Payload
   - With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
   - With not heavy payload the successful landing or positive landing rate are for SSO
   - Booster Version FT are more successful and the V1.1 are worst success performance

4. Support Vector Machine was a suitable model to predict if the stage one would land with success, it had an accuracy of 83% and performs the best in terms of Area Under the Curve at 0,958

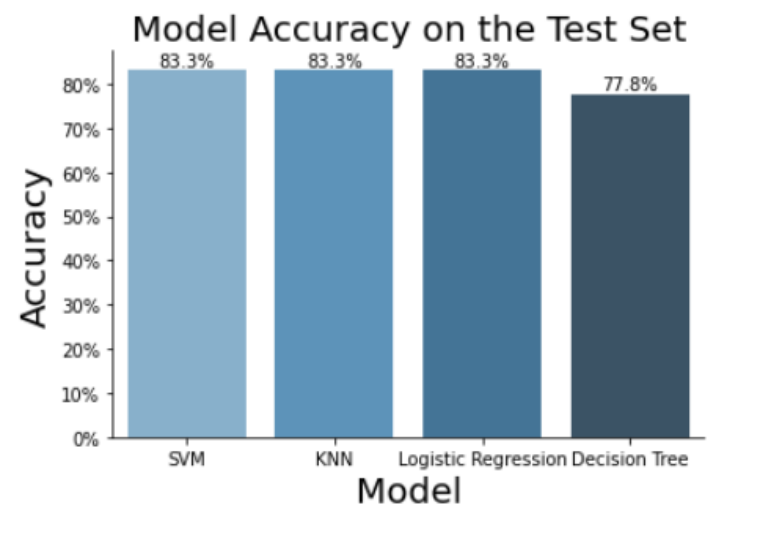
Appendix

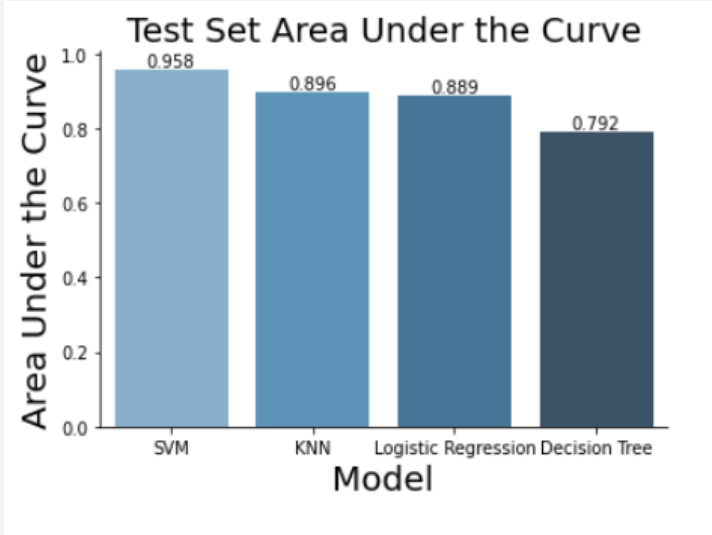
1) - Table with performance of all ML models used

	AUC	F1-Score	Precision	Recall	Accuracy
SVM	0.958	0.889	0.8	1.00	0.833
KNN	0.896	0.889	0.8	1.00	0.833
Logistic Regression	0.889	0.889	0.8	1.00	0.833
Decision Tree	0.792	0.818	0.9	0.75	0.778

The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958.

2) - Bar Chart with Model Accuracy And Test Set Area Under the Curve





Thank you!