IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Purnima Kulkarni
17th February 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected from SPACEX API and Wikipedia pages.
  - We performed wrangling on this data to identify and treat missing cases. We created label column ‘Class’
  - Exploratory Data Analysis (EDA) was done through scatterplots, SQL queries, Folium maps and interactive dashboard. and with one hot encoding created features columns for the Classification model.
  - Finally, classification models were run to predict the launch outcome, ‘Class’.
- Summary of all results –
  - EDA helped us with identifying patterns & important variables in the data. Orbit, Launch Site and Payload Mass seem to have the most impact on the launch outcome.
  - All the classification models, namely Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbours have an accuracy of 83% and can be used reliably for predicting the success or failure of future launches.

Introduction

- Project background and context
  - Humans are known to explore space to learn the unknown and in recent times, even to find habitable homes
  - Therefore, commercial space travel has fast become popular
  - Currently, Spacex has the lowest cost 62 million, mainly due to reusable first stage of their Falcon rockets
  - We want to get into this business
- Problems you want to find answers –
  - We need to be competitive in our pricing if we want to be successful
  - Therefore, it's imperative that, like Spacex, we too reuse the Stage 1 rocket
  - We aim to build a Machine Learning Classification model for predicting successful Stage 1 launch
  - We will use publicly available Spacex data and analytical techniques to build our model

Section 1

Methodology

Methodology

Executive Summary

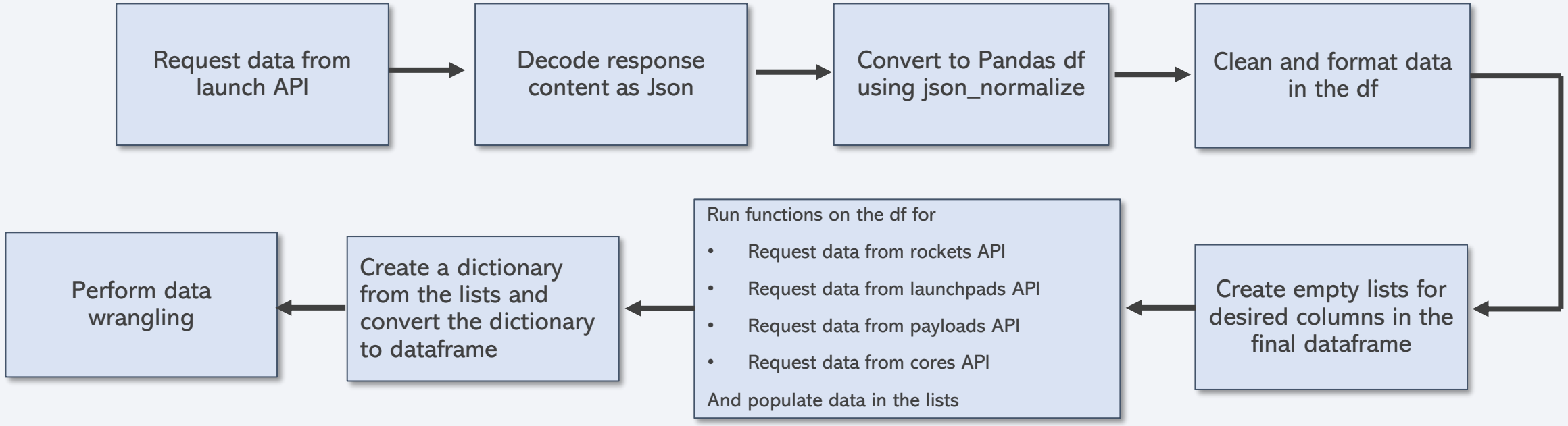
- Data collection methodology: Collect data from SPACEX REST APIs and by Webscraping from Wikipedia
- Perform data wrangling: Check for missing values, counts, create the Landing Outcome variable as 1/0 for Successful/Unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: Find the best predictive model among Logistic Regression, SVM, Decision Tree and KNN models with GridSearchCV

Data Collection

- Data was collected using SPACEX REST APIs and by Webscraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

Data Collection – SpaceX API

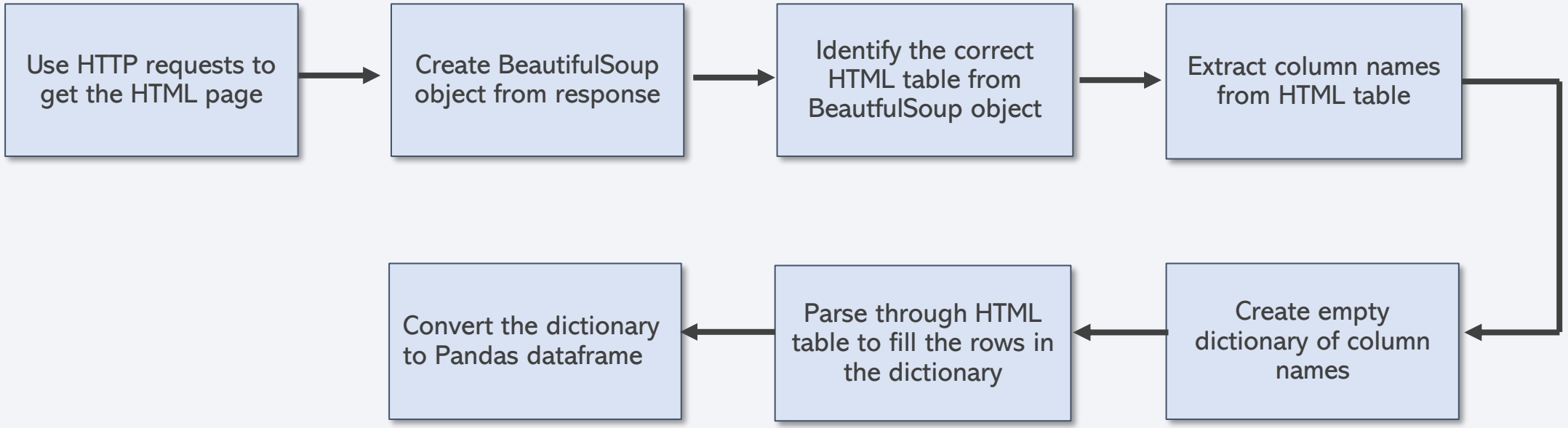
Data was collected by making call to the SpaceX API and then cleaning the data. First a call was made to launch API and data saved into a dataframe. Then, a few helper functions were used which in turn made calls to Rockets, Launchpads, Payloads and Cores APIs. This helped us to extract information using identification numbers in the launch data.



https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

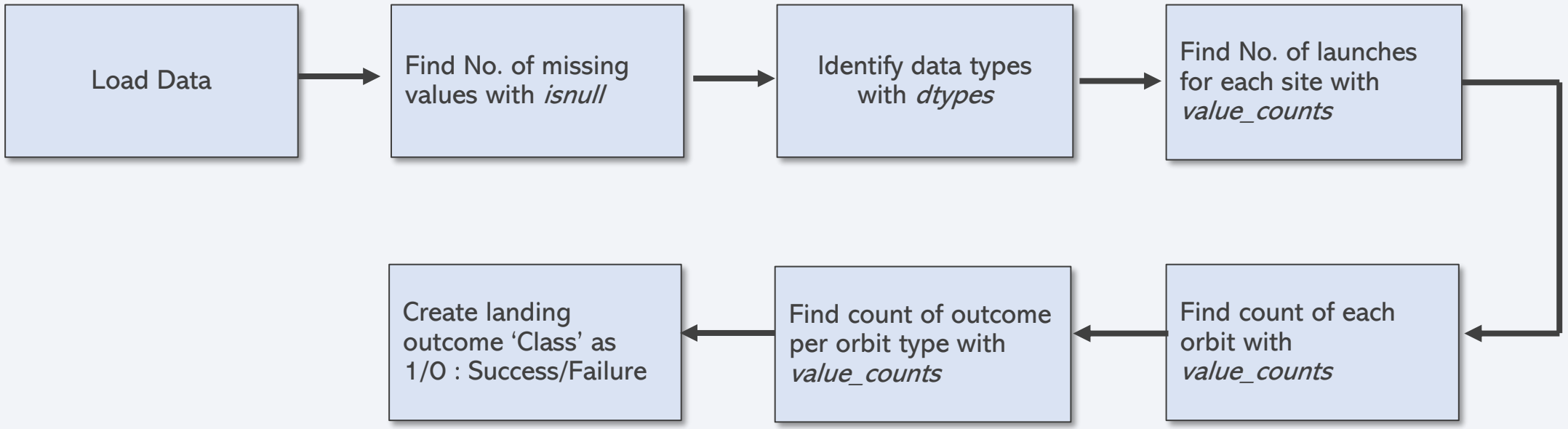
Web scraping was done to collect data from Wikipedia page. Falcon 9 launch records were read from the HTML table from Wikipedia, parsed and converted into a Pandas data frame.



https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/Webscraping.ipynb

Data Wrangling

We exploring data types, contents and transformed it into usable format for further analysis. Created our Training Label, the 'Class' variable in this exercise.



https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/Data%20Wrangling%20.ipynb

EDA with Data Visualization

To understand relationships between the launch outcome and other features, several charts like scatterplots, line chart and bar plots were prepared.

- Scatterplot of Flight Number vs Payload mass
- Scatterplot of Flight Number vs Launch Site
- Scatterplot of Payload mass vs launch site
- Barplot of Success Rate by orbit type
- Scatterplot of Flight Number vs Orbit Type
- Scatterplot of Payload mass vs Orbit Type
- Line chart of yearly trend of Success Rate

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/jupyter_labs_eda_dataviz_final_use_this.ipynb

EDA with SQL

Following SQL queries were run to explore the data further

- 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 succesful landing outcome in ground pad was acheived.
- 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 using 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.

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20final.ipynb

Build an Interactive Map with Folium

In order to do visual analysis of the launch sites like proximities to other areas and success rate, we have created interactive maps through Folium. Following objects are added.

- A folium Map object was created with an initial center location to be NASA Johnson Space Center at Houston, Texas and a Folium.Circle was added to the position.
- 4 more Circle objects were added to the map to show location of 4 sites. Now the map has all the launch sites on it for us to view.
- Next a MarkerCluster object was added to the map to mark the success/failed launches for each site on the map, with each of the launches colour coded as “Green” for success and “Red” for failure. With this on the map, we can now analyse success rate of each of the sites.
- A mouseposition object was added to the map so that we can get Lat and Long of any location on the map. We find out Lat and Long of a location of interest and use Polyline to draw line between the launch site and the point and get distance between the two in kms. With this added, we can do proximities analysis like how far is coastline, how far are the cities, highways or railway lines.

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/lab_jupyter_launch_site_location_final_use_this.ipynb

Build a Dashboard with Plotly Dash

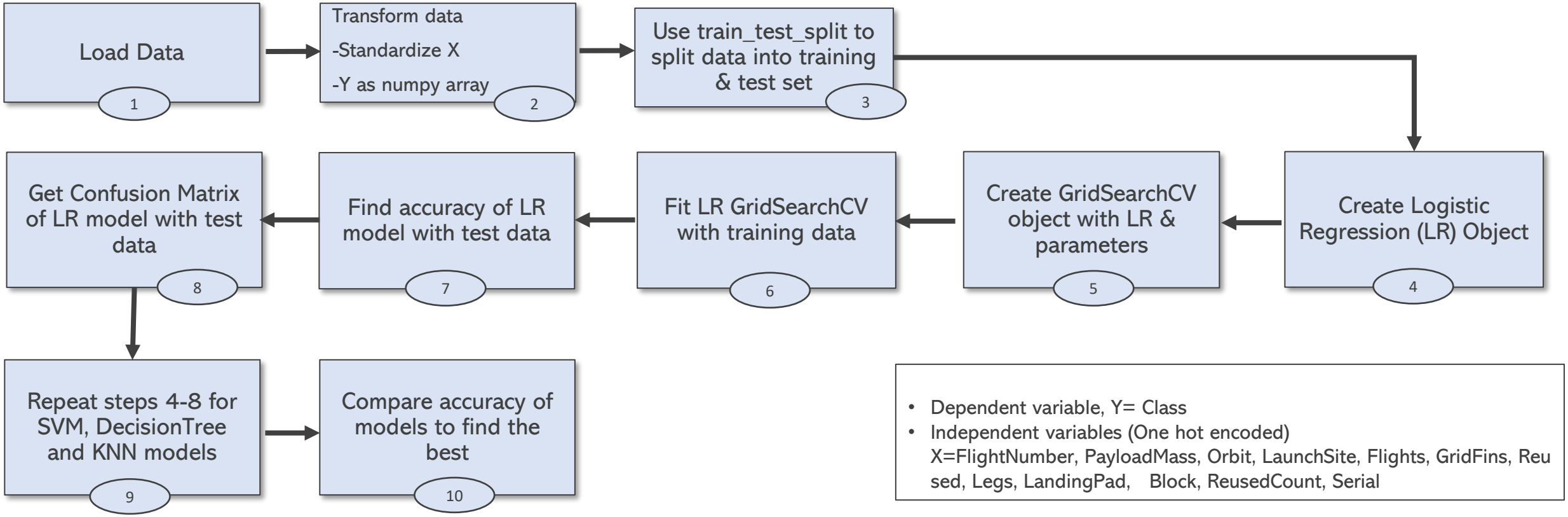
In order to understand the relationship between launch outcome and attributes like Launchsite, Booster version, Payload mass, an interactive dashboard has been built.

- A pie chart showing Successful outcome % for All Sites and individual sites. A drop down has been added to select the site name.
- Scatterplot of Payload mass versus Launch Outcome for the selected launch site with Booster version overlayed. A rangeslider has been added to select the payload mass range.

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/spacex_dash_app%20final.py

Predictive Analysis (Classification)

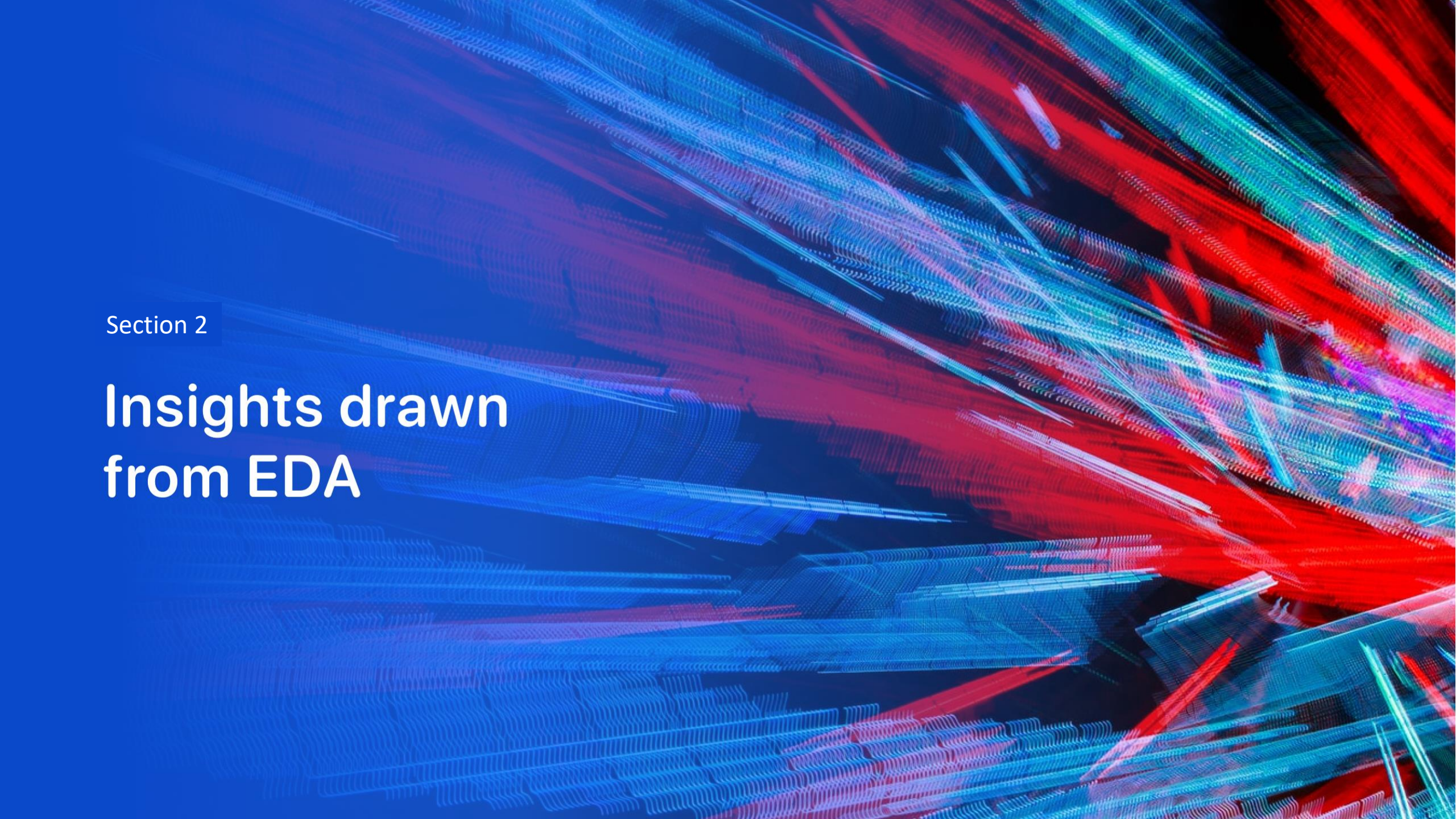
The objective of this research is to predict the outcome of the first stage of the rocket. In order to do this, 4 classification models were run 1)Logistic Regression 2)Support Vector Machine 3)Decision Tree 4)K Nearest Neighbours. GridSearchCV was used to find the optimal hyperparameters. Crossvalidation was set to 10.



https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/Spacex_data_ML_Prediction_final%20use%20this.ipynb

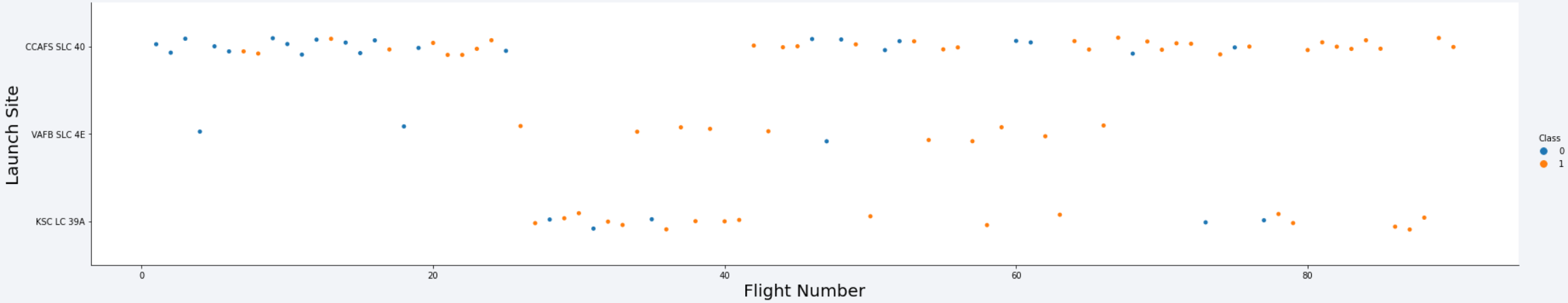
Results

- All the launch sites are near the coastline and far from cities and densely populated areas.
- Success rate has been on the rise since 2013 and peaked at 85% in 2019. First successful ground landing occurred in 2015.
- Site KSL LC-39A has the highest success rate and contributes 42% to the total success of launches.
- Launch site and combination of Orbit type & payload mass has an impact on the outcome.
- Average payload carried per launch is 6105 kgs.
    - Highest number of launches were done in the Payload range of 2500-5500 kgs with many different orbit types.
    - Orbit VLEO is only with heavy payloads of 15000 kgs on an average and is the most successful orbit with >80% success rate.
- All the Classification models have an accuracy of 83%


Section 2
Insights drawn from EDA

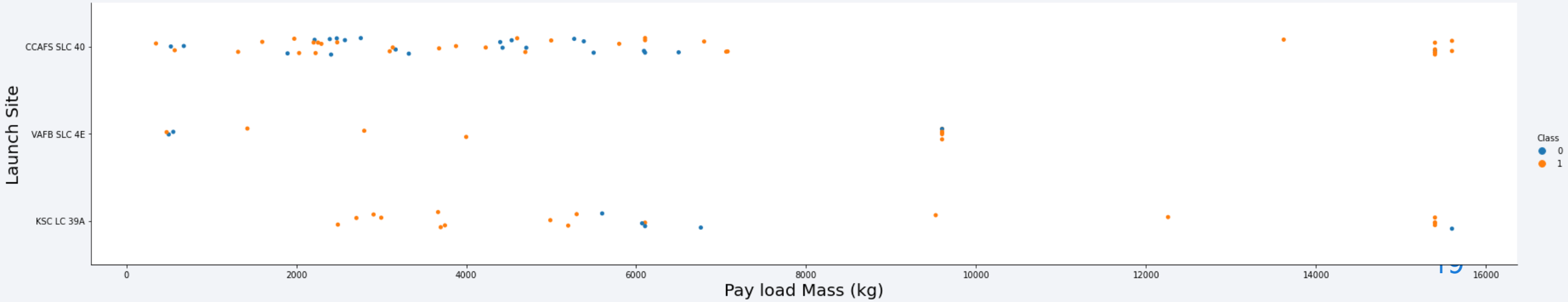
Flight Number vs. Launch Site

- As the flight number increases, the number of successful launch outcomes increase.
- Flight numbers in turn represent time. With time, the success rate of launch has increased.
- For site CCAFS SLC 40, the first 60 flights had more failures and later success rate improved.
- For site KSC LC 39A, Success rate in the later flights dropped.
- Site VAFB SLC 4E did not have any launches recently. Prior to that, it has had good success rate.



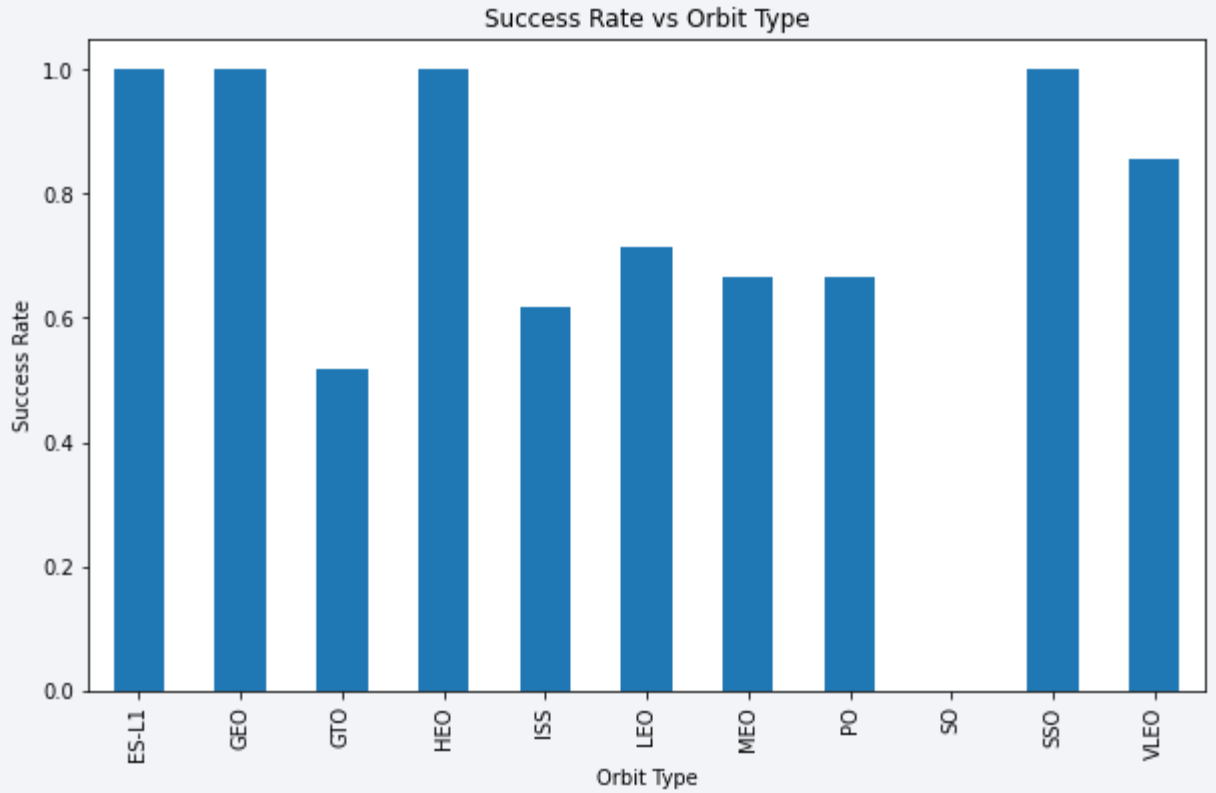
Payload vs. Launch Site

- For site CCAFS SLC 40, outcomes up to payload mass of 8000 kgs are mixed and as the payload mass crosses 8000 kgs, all the launches are successful.
- Site VAFB SLC 4E did not have any rocket launches of >10000 kgs.
- Site KSC LC 39A has high success rate for payload mass of <6000 kgs.



Success Rate vs. Orbit Type

- Orbit type ELS1, GEO, HEO and SSO have 100% success rate. They have low number of launches, only 1 launch each for the first 3 and 5 launches for SSO.
- So, the most successful orbit type is VLEO with a success rate of > 80% among 14 launches.
- The least successful orbit type is GTO with a success rate of ~50% among 27 launches.

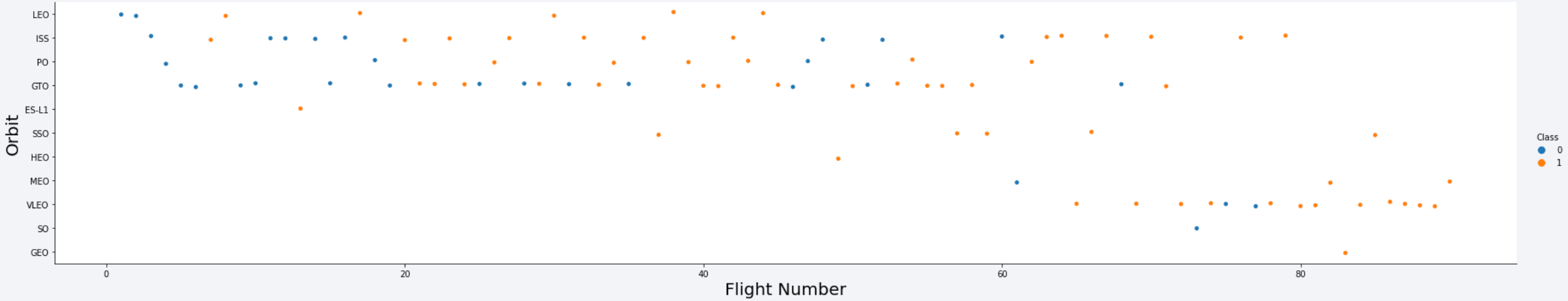


No of launches per site

ES-L1 1
GEO 1
GTO 27
HEO 1
ISS 21
LEO 7
MEO 3
PO 9
SO 1
SSO 5
VLEO 14

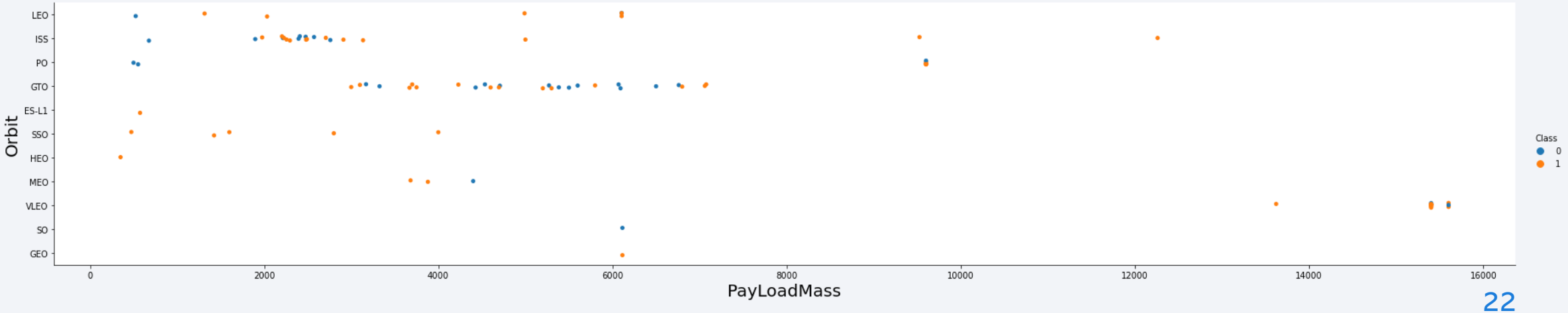
Flight Number vs. Orbit Type

- For Orbit type LEO, most of the launches are successful except the first 2.
- For Orbit GTO and Orbit ISS, no pattern can be found with a mix of successful and unsuccessful launches over the range of flight numbers.
- Orbit SSO have no early launches and all the launches are successful.
- Orbit VLEO started launches much later and have majority of them become successful.



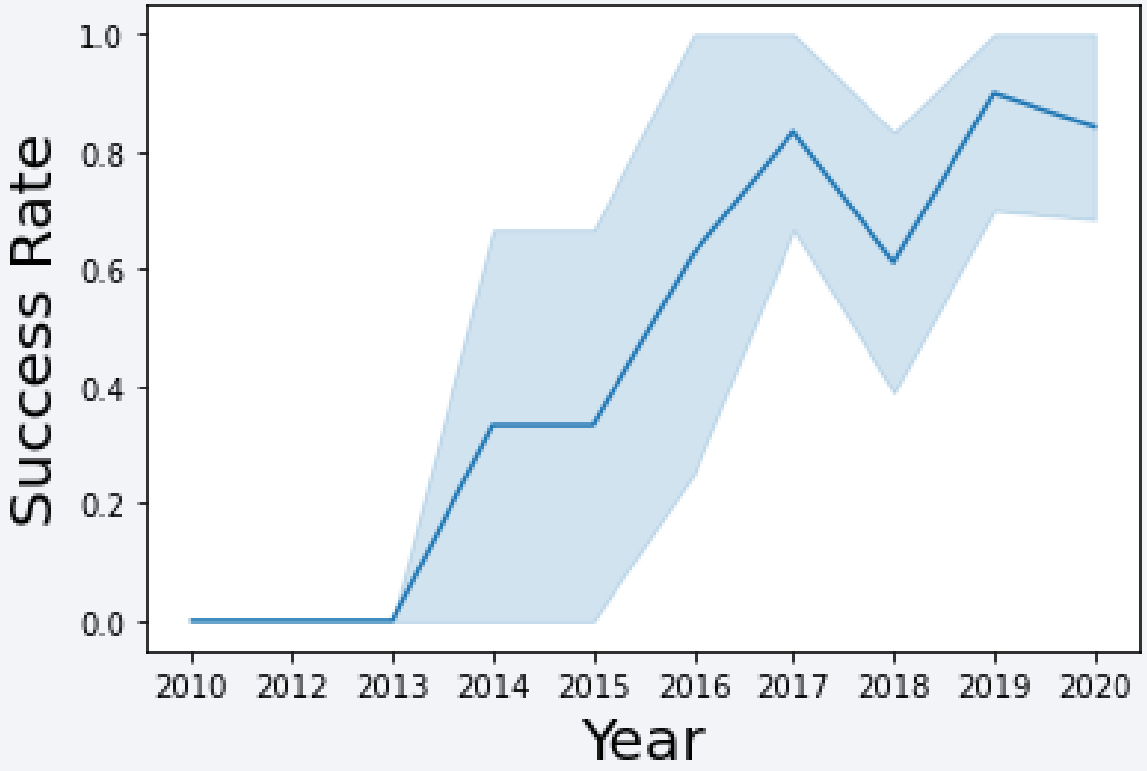
Payload vs. Orbit Type

- For Orbit type LEO and ISS there is a trend of increased success rate with increase in payload mass.
- For GTO orbit type, no correlation between success rate and payload mass can be seen. No launches with heavy payload mass can be seen.
- Orbit type VLEO had only heavy payload mass launches.
- ES-L1, SSO, HEO and MEO had no launches over 5000 kgs.



Launch Success Yearly Trend

- Success Rate started increasing from year 2013.
- It saw a steady increase till year 2017 with a drop in the year 2018. However, it picked up and reached the highest in 2019.



All Launch Site Names

- Now, we will look at some of the data from Spacex.csv which was converted to SPACEXTBL.
- SQL query with “distinct(Launch_site)” was used to get the unique site names.

SrNo	Launch Site
1	CCAFS LC-40
2	VAFB SLC-4E
3	KSC LC-39A
4	CCAFS SLC-40

```
%sql select distinct(Launch_Site) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
  Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- SQL query with “Launch_Site like ‘CCA%’ limit 5” was used to get 5 records where Launch site name starts with ‘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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass carried by boosters from NASA CRS was calculated using a query with “SUM(PAYLOAD_MASS__KG_)”
- Total payload mass for NASA CRS is 45596 kgs

```
%sql select sum(PAYLOAD_MASS__KG_) as "Total payload mass for Customer=NASA CRS" from SPACEXTBL where Customer='NASA (CRS)'
```

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

Total payload mass for Customer=NASA CRS
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using a query with “ avg(PAYLOAD_MASS__KG_) where Booster_Version= 'F9 v1.1’ ”
- Average payload mass carried by booster version F9 v1.1 is 2928.4 kgs.

```
%sql select avG(PAYLOAD_MASS__KG_) as "Average payload mass for Booster_Version= F9 v1.1" from SPACEXTBL where Booster_Version= 'F9 v1.1'
#%sql select * from SPACEXTBL where Booster_Version= 'F9 v1.1'
```

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

Average payload mass for Booster_Version= F9 v1.1
2928.4

First Successful Ground Landing Date

- The date of first successful landing outcome on ground pad was found using a query with “min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) as date_yyyymmdd”.
- The first successful landing outcome on ground pad was on December 22, 2015.

```
%sql select min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) as date_yyyymmdd, 'Date of first successful landing outcome in ground pad'
from SPACEXTBL where "Landing_Outcome"="Success (ground pad)"
```

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

date_yyyymmdd	'Date of first successful landing outcome in ground pad'
20151222	Date of first successful landing outcome in 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 were found using a query with "("Landing _Outcome"="Success (drone ship)" and PAYLOAD_MASS__KG_ between 4001 and 5999)"
- There were 4 such landings mentioned below.

```
%sql select distinct(Booster_Version) from SPACEXTBL where ("Landing _Outcome"="Success (drone ship)" and PAYLOAD_MASS__KG_ between 4001 and 5999)
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes were calculated using a query with “group by Mission_Outcome”
- There were 100 Successful outcomes and 1 Failure

```
[ ]  %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- Names of the booster which carried the maximum payload mass was found by running the query with a subquery as “distinct(Booster_Version) from SPACEXTBL where PAYLOAD_MASS__KG_= (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- The results are presented below

```
%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

2015 Launch Records

- To get the failed landing outcomes in drone ship in year 2015, a query was run with “where (substr(Date,7,4)='2015' and "Landing _Outcome"="Failure (drone ship)")”
- There were two such failed launches, in January 2015 and April 2015 from site CCAFS LC-40

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

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

Date	Landing _Outcome	Booster_Version	Launch_Site	Month
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04

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

- To get the count of landing outcomes between the date 2010-06-04 and 2017-03-20, “count("Landing _Outcome") as Count from SPACEXTBL where date between '04-06-2010' and '20-03-2017' group by "Landing _Outcome" order by count desc”
- Results are shown below

```
_Outcome", count("Landing _Outcome" ) as Count from SPACEXTBL where date between '04-06-2010' and  '20-03-2017' group by "Landing _Outcome" order by count desc
```

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

Landing _Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

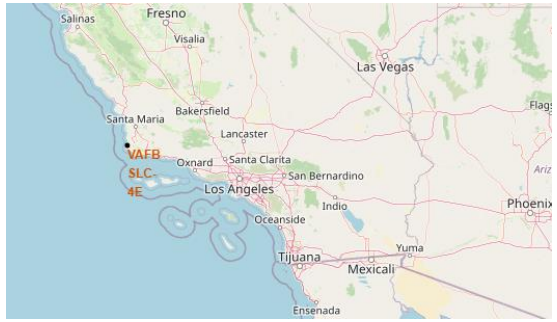
Section 3

Launch Sites Proximities Analysis

Launch Site Locations on the World Map

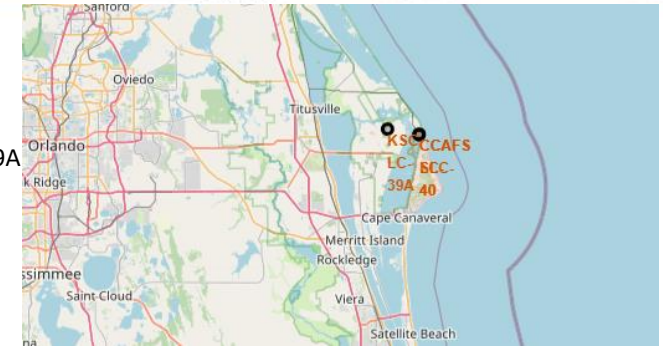
- 4 launch sites have been plotted on the map. All of them are in close proximity to the coastline. VAFB SLC 4FE is on the west coast of the United States. Other three sites, KSC LC 39A CCAFS SLC40 and CCAFS LC40 are on the southeast coast of the Unites States.
- Launch sites are somewhat close to the Equator.

Zoomed location for site VAFB SLC-4FE

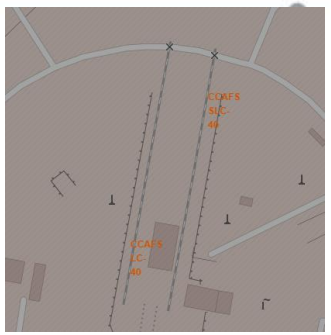


Zoomed location for sites

- KSC LC 39A
- CCAFS SLC40
- CCAFS LC40

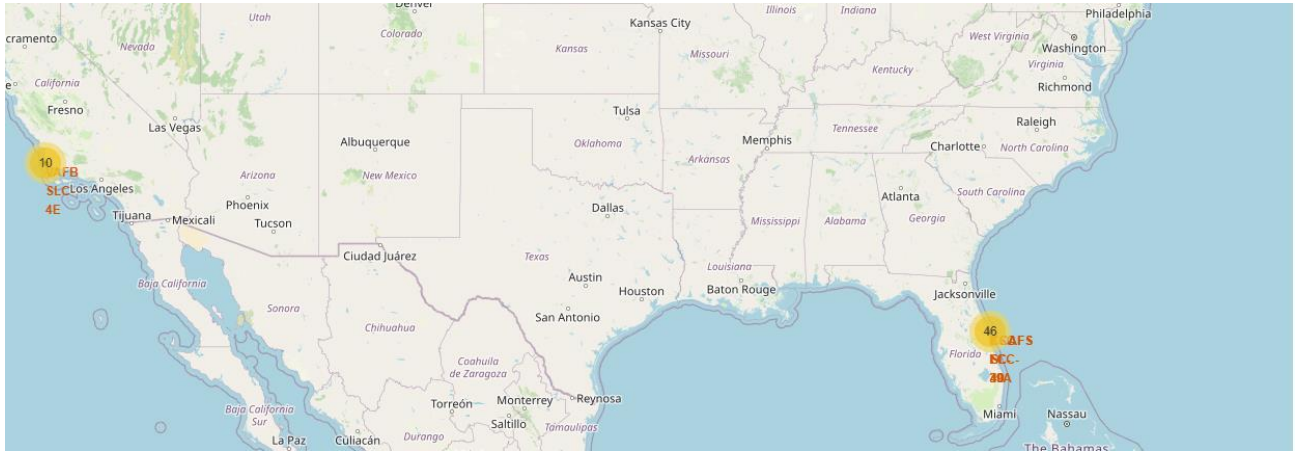


Sites CCAFS SLC40 & CCAFS LC40 are very close to each other



Launch Outcomes by Sites

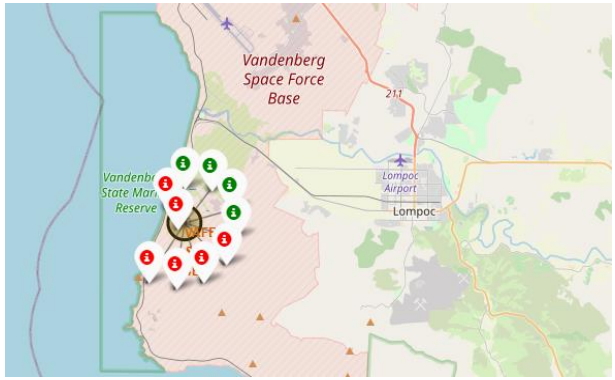
Launch site KSC LC 39A has the highest success rate with 10 out of 13 launches being successful whereas site CCAFS LC40 has lowest success rate.



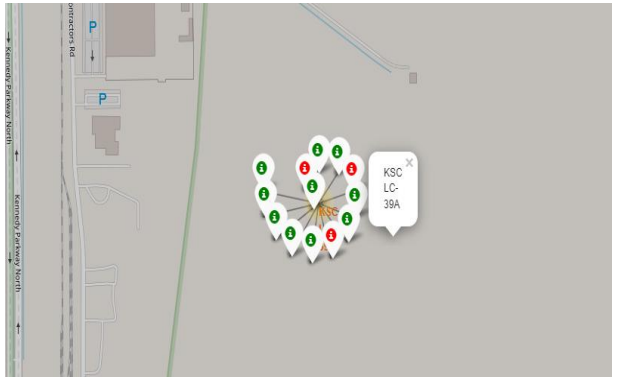
Successful launch

Failed launch

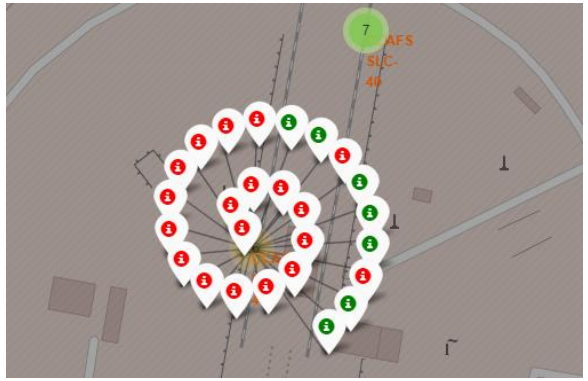
Site VAFB SLC 4FE



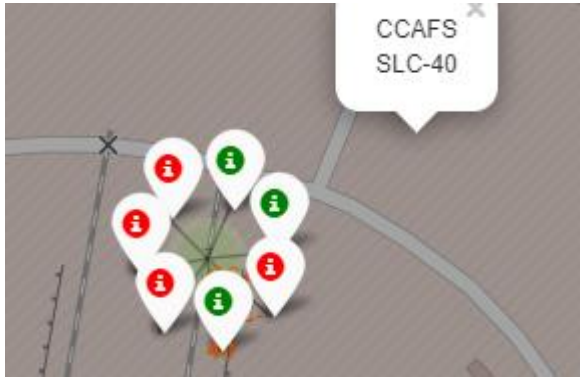
Site KSC LC 39A



Site CCAFS LC40



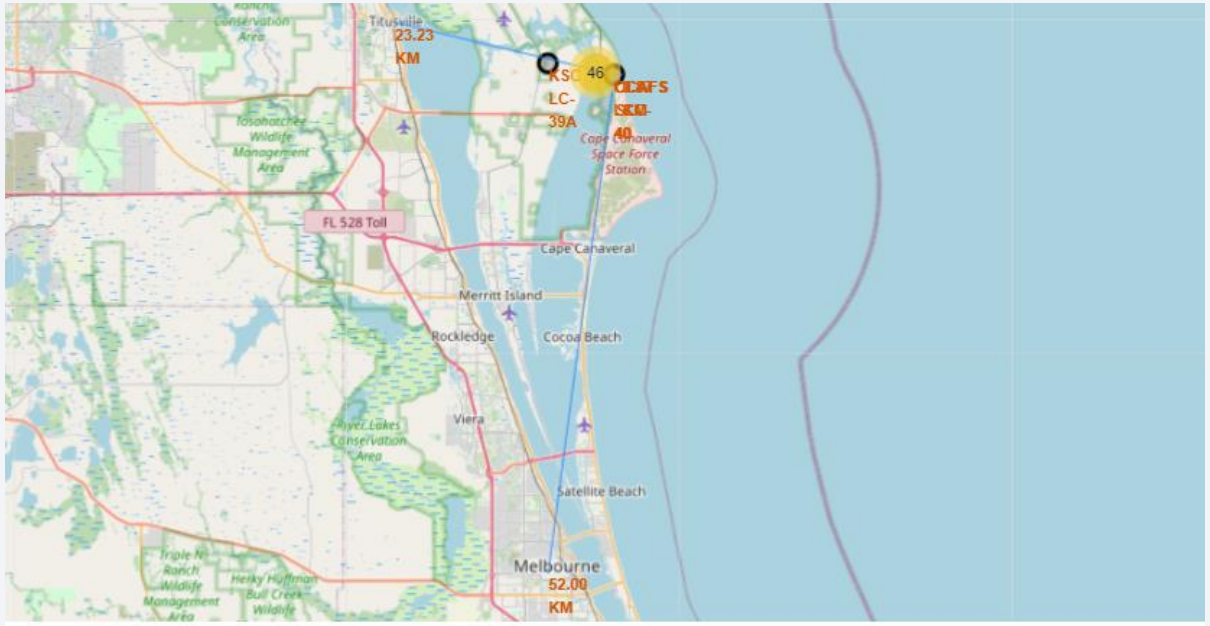
Site CCAFS SLC40



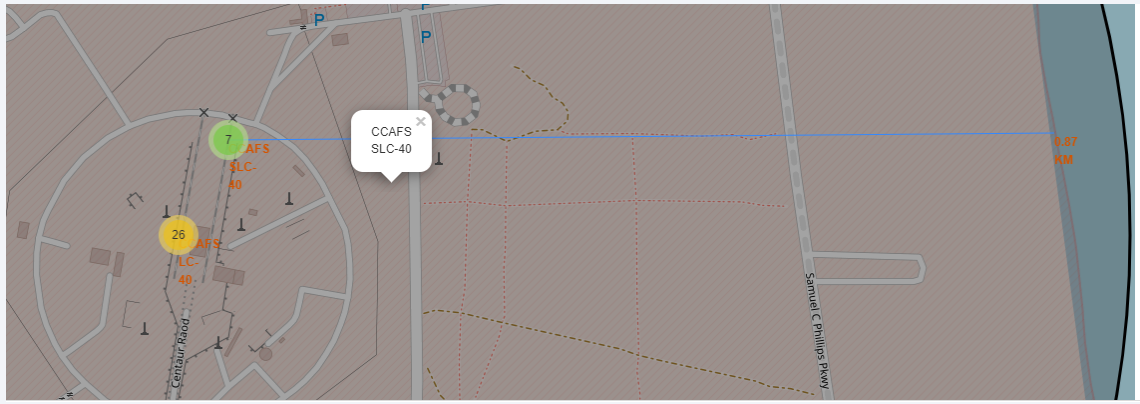
Launch Sites Proximity Areas

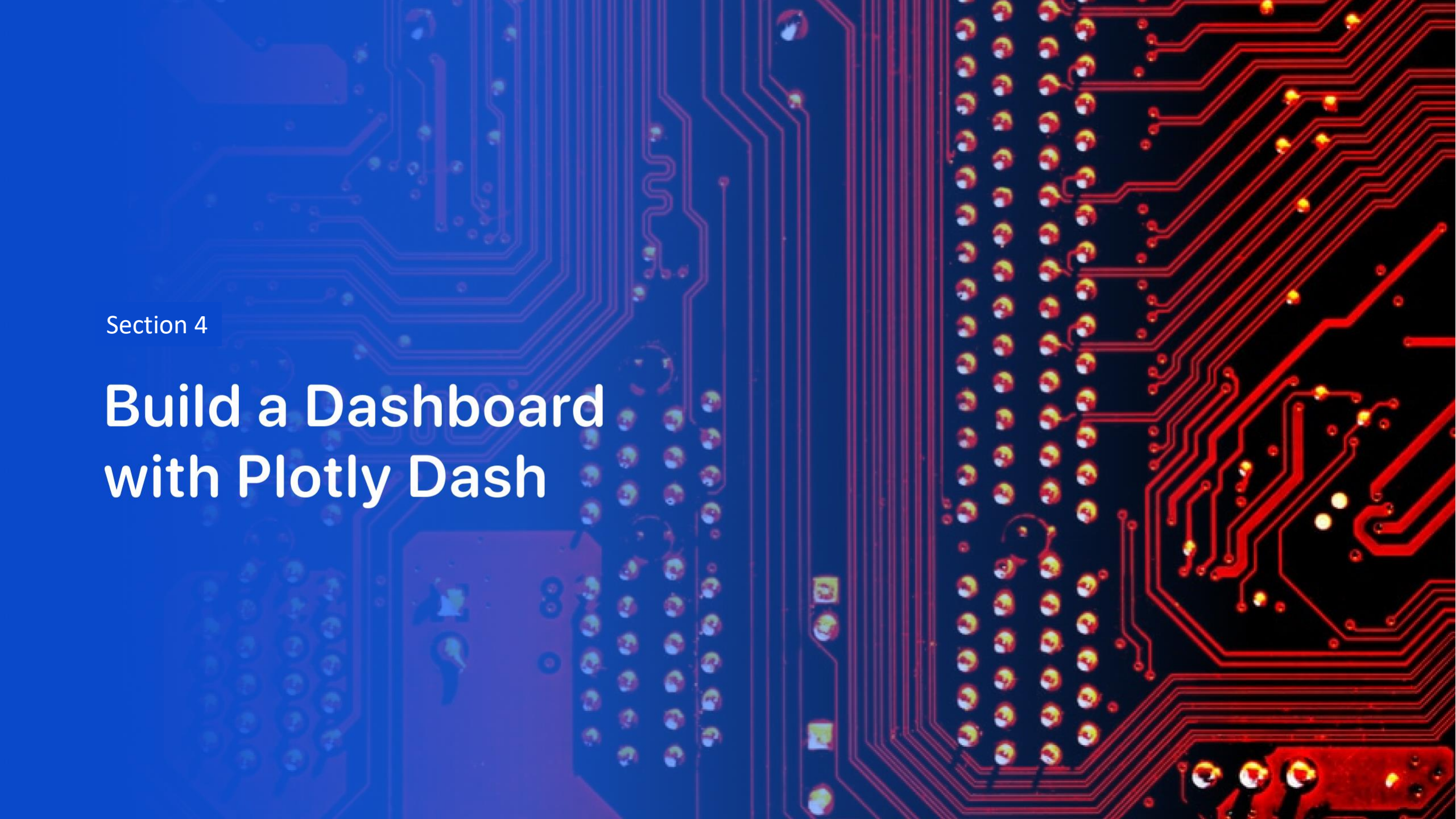
- All the launch sites are close to coastline.
- We reviewed surrounding areas of CCAFS LC40.
  - The coastline is less than 1 km away
  - Nearest town/city are at least 20 kms away.
  - There is only a NASA railway near the site and the site doesn't seem to be close to a highway.

Distance between site CCAFS LC40 and Titusville city = 23 kms



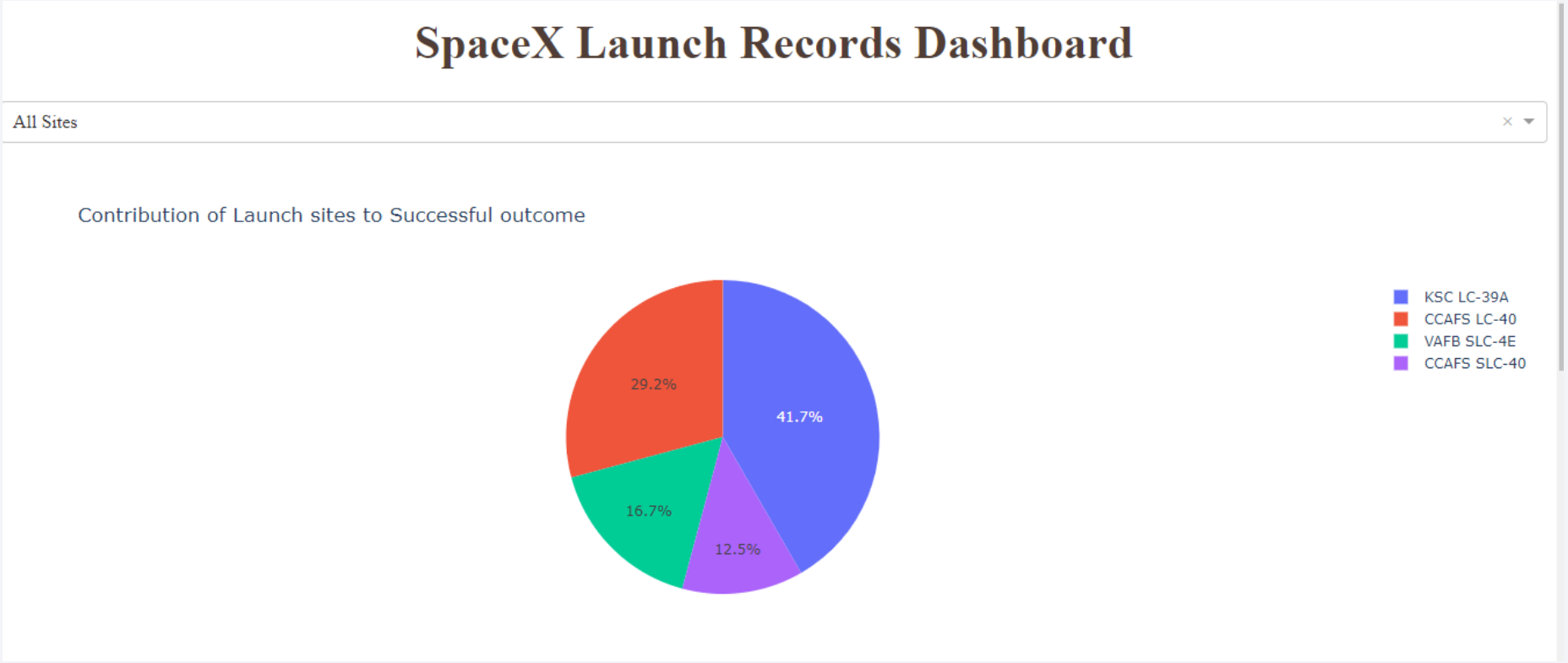
Distance between site CCAFS LC40 and coastline = 0.87km




Section 4
Build a Dashboard with Plotly Dash

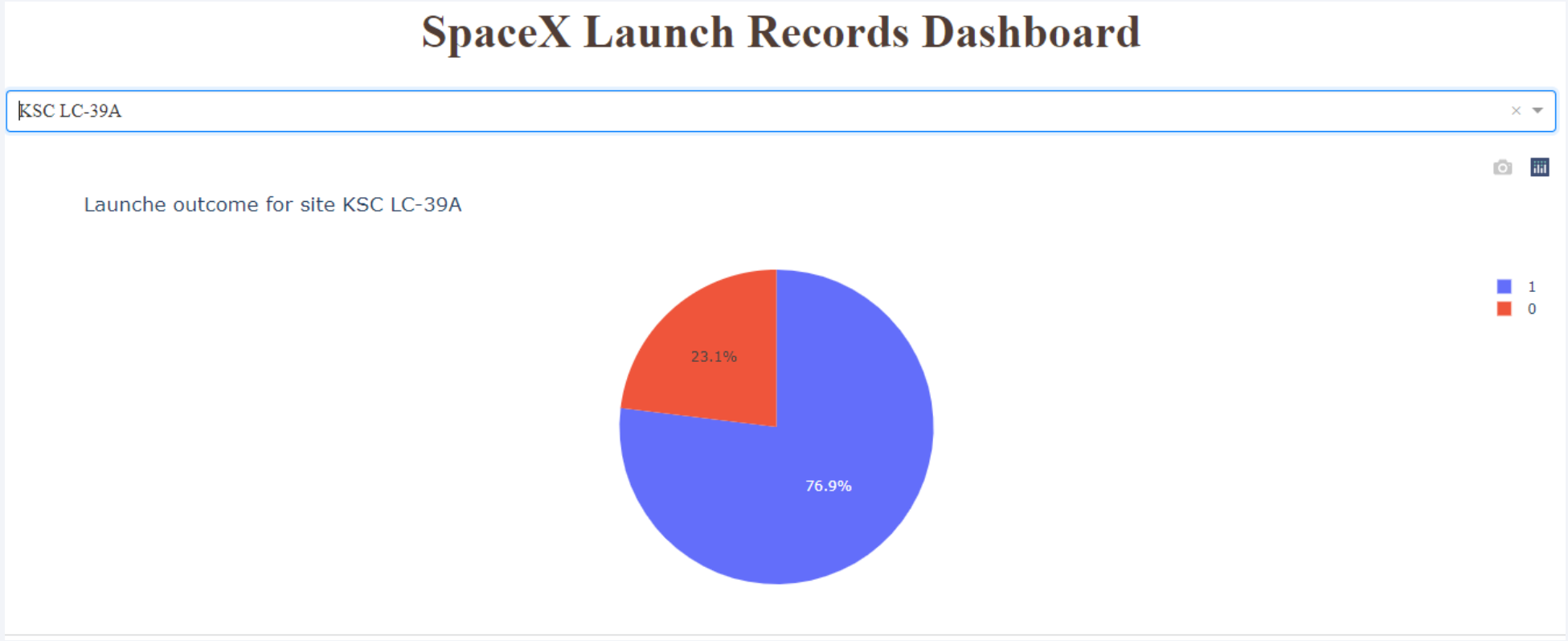
Contribution of Launch Sites to Successful Outcome

- Maximum around 42% of all successful launches are contributed by site KSC LC-39A, followed by site CCAFS LC-40
- Lowest contribution is from site CCAFS SLC-40.



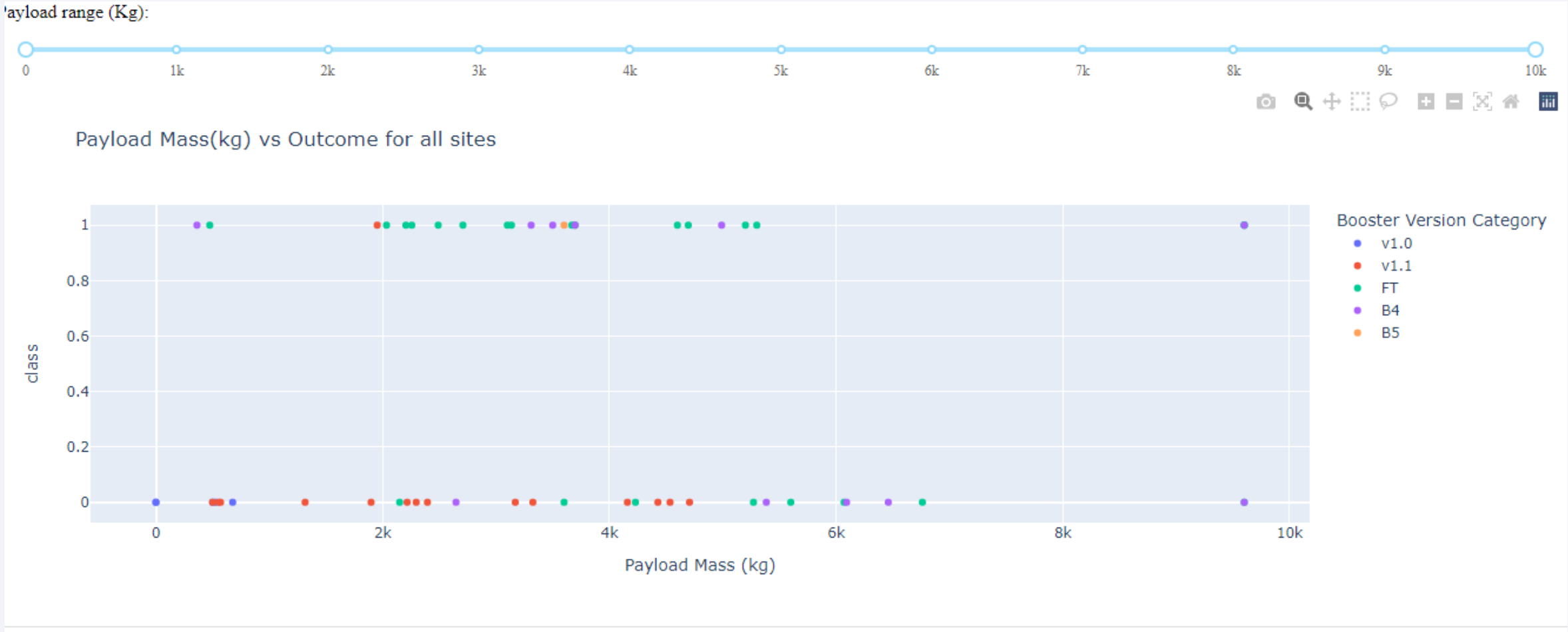
Launch site with Highest Success Rate

- KSC LC-39A has the highest success rate of 77%.



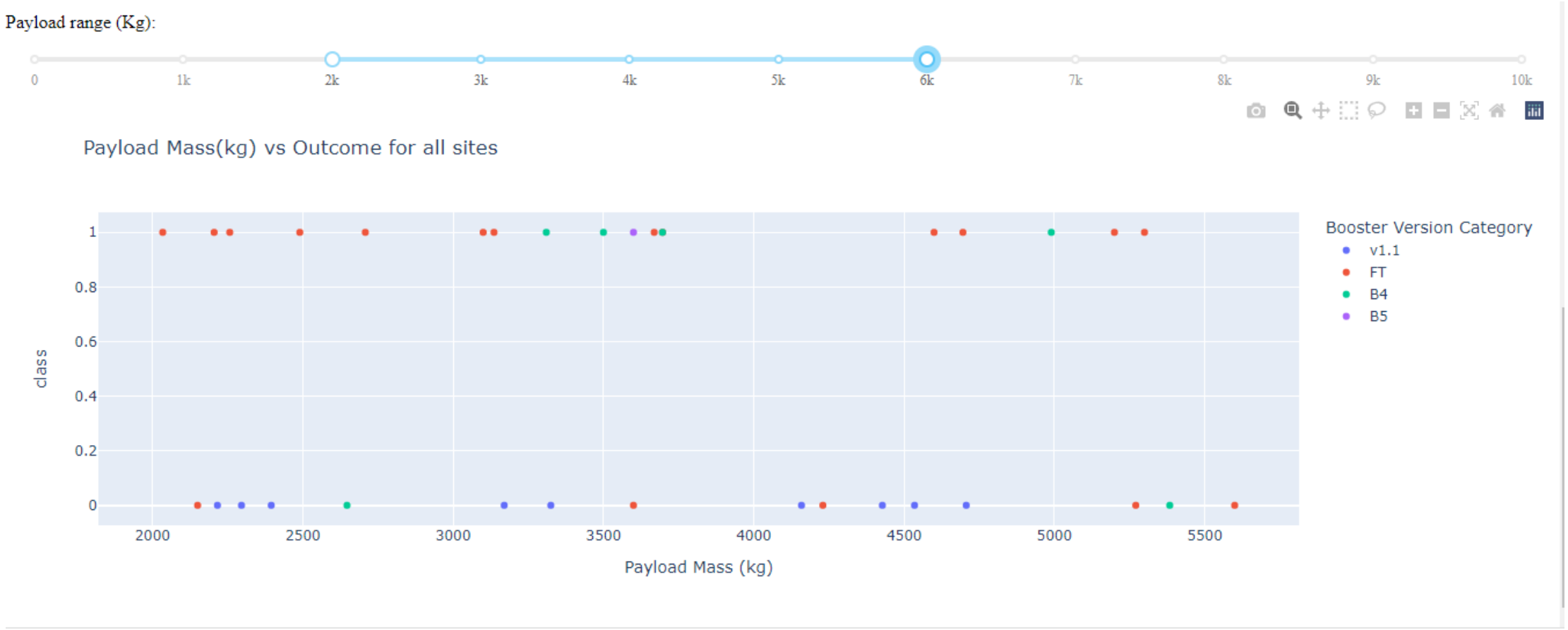
Payload Mass versus Launch Outcome (1)

- Maximum launches were done in the Payload range of 2500-5500 kgs
- It has a mix of successful and unsuccessful launches.



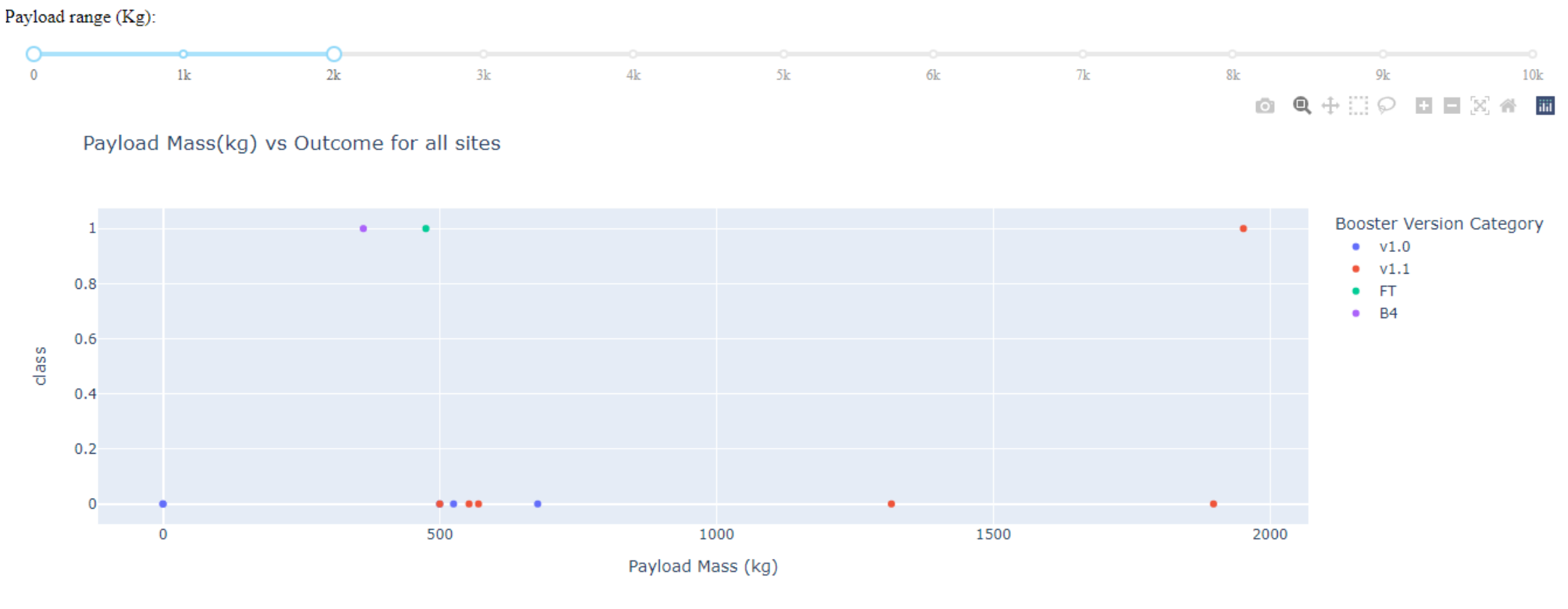
Payload Mass versus Launch Outcome (2)

- Let's look at Payload mass range of 2000 kgs and 6000kgs as most of the data points are within this range.
- Booster version FT appears to be the most successful and Booster version v1.1 has the lowest success rate.
- There is no v1.0 booster rocket between this range



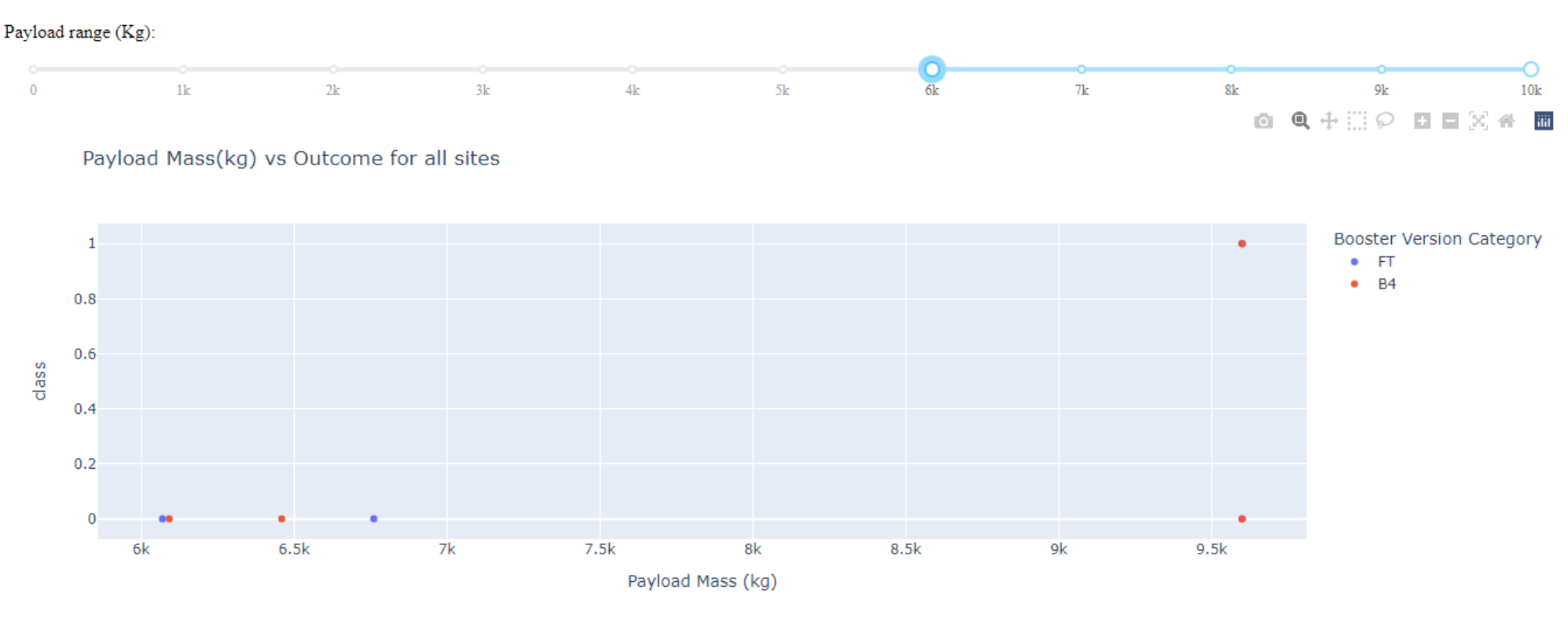
Payload Mass versus Launch Outcome (3)

- Let's look at Payload mass range upto 2000 kgs
- Success rate is low with this payload mass
- Booster version V1.0 were launched only within this payload mass range but none were successful



Payload Mass versus Launch Outcome (4)

- Few count and low success rate for launches with Payload mass >6000kgs



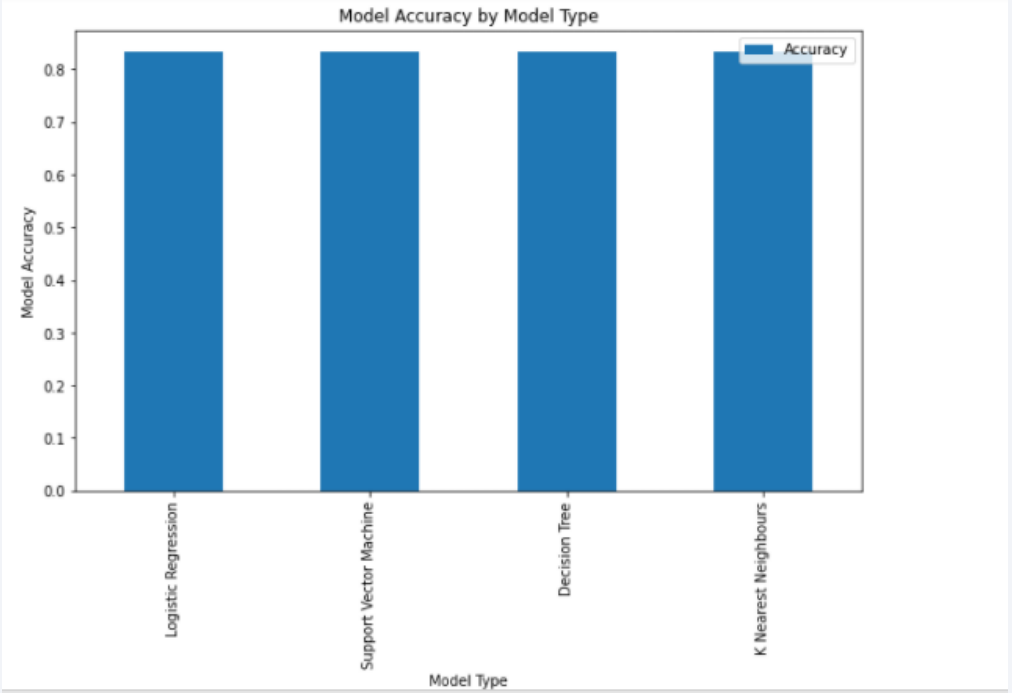
Section 5

Predictive Analysis (Classification)

Classification Accuracy

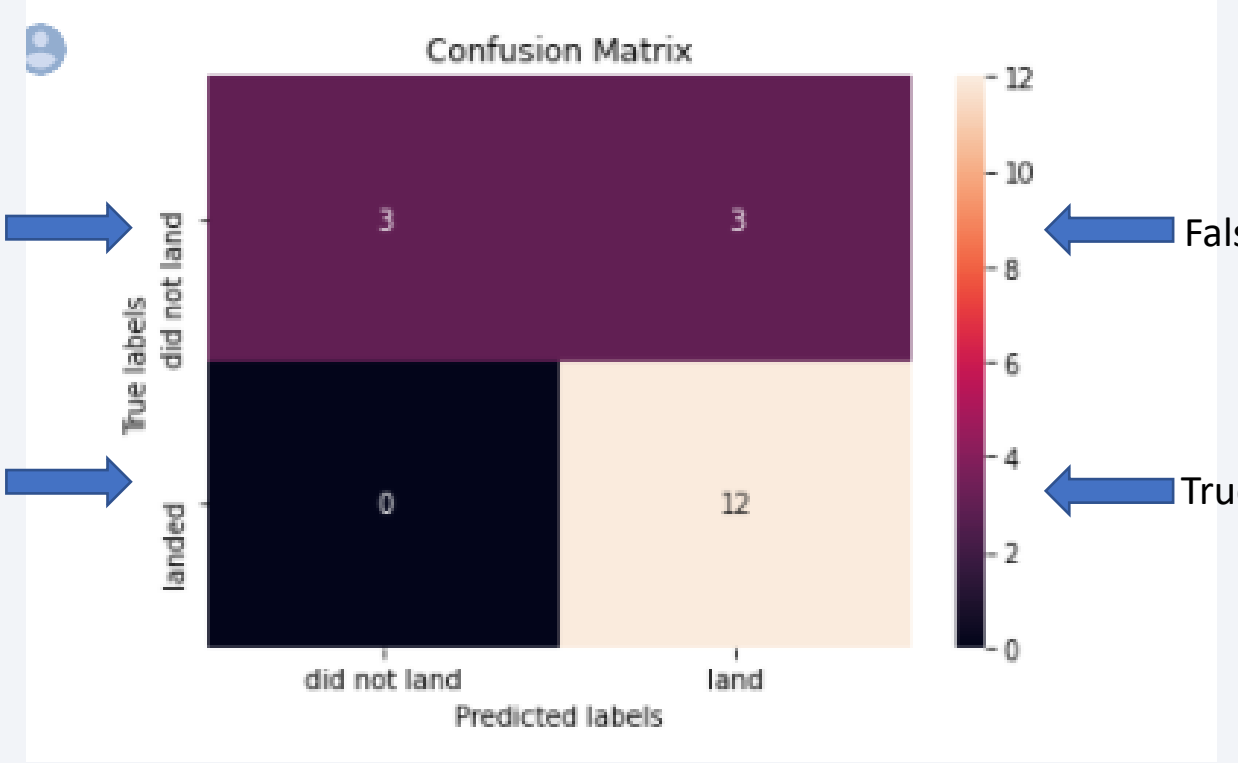
- Accuracy is a measure of model performance, defined as a fraction of correct predictions out of total predictions done.
- Classification accuracy of each of the models is .83 or 83% for the test data. Model accuracy is high (>80%) and hence, all the models are equally good at predicting the outcome of the landing.

- Dependent variable, Y= Class
- Independent variables Y = FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial



Confusion Matrix

- Confusion matrix is another measure of performance of the model. For all our models, Confusion matrix is the same.
- Our model predicted high True Positive which is favourable
- Count of 3 in False Positive cell is a concern but the number is small and model can be used as accuracy is >80%.



Summary Table

	Number of launches			Launch Site			Payload mass in kgs		
	Total	Success	Failure	CCAFS SLC 40	KSC LC 39A	VAFB SLC 4E	Avg	Min	Max
Orbit		=100%		=100%					
Total	90	67%	33%	61%	24%	14%	6105	350	15600
GTO	27	52%	48%	67%	33%	0%	5012	3000	7076
ISS	21	62%	38%	76%	24%	0%	3280	677	12259
VLEO	14	86%	14%	64%	36%	0%	15316	13620	15600
PO	9	67%	33%	0%	0%	100%	7584	500	9600
LEO	7	71%	29%	71%	29%	0%	3883	525	6105
SSO	5	100%	0%	20%	0%	80%	2060	475	4000
MEO	3	67%	33%	100%	0%	0%	3987	3681	4400
ES-L1	1	100%	0%	100%	0%	0%	570	570	570
HEO	1	100%	0%	100%	0%	0%	350	350	350
SO	1	0%	100%	0%	100%	0%	6105	6105	6105
GEO	1	100%	0%	100%	0%	0%	6105	6105	6105

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final/blob/main/jupyter_labs_eda_dataviz_final_use_this.ipynb

Conclusions

- Our machine learning classification model has an accuracy of 83% and can be reliably used for prediction of Stage 1 rocket launches in future.
- Currently, all the four classification models, Logistic Regression, SVM, Decision Tree and KNN give the same accuracy, due to low number of observations in the test data. In future, as more data becomes available, we could revisit the analysis to get the best model out of the four.

Appendix

Link to the Github repository for the project

https://github.com/PurnimaKulkarni/IBM-Data-Science-Capstone-Final

Thank you!