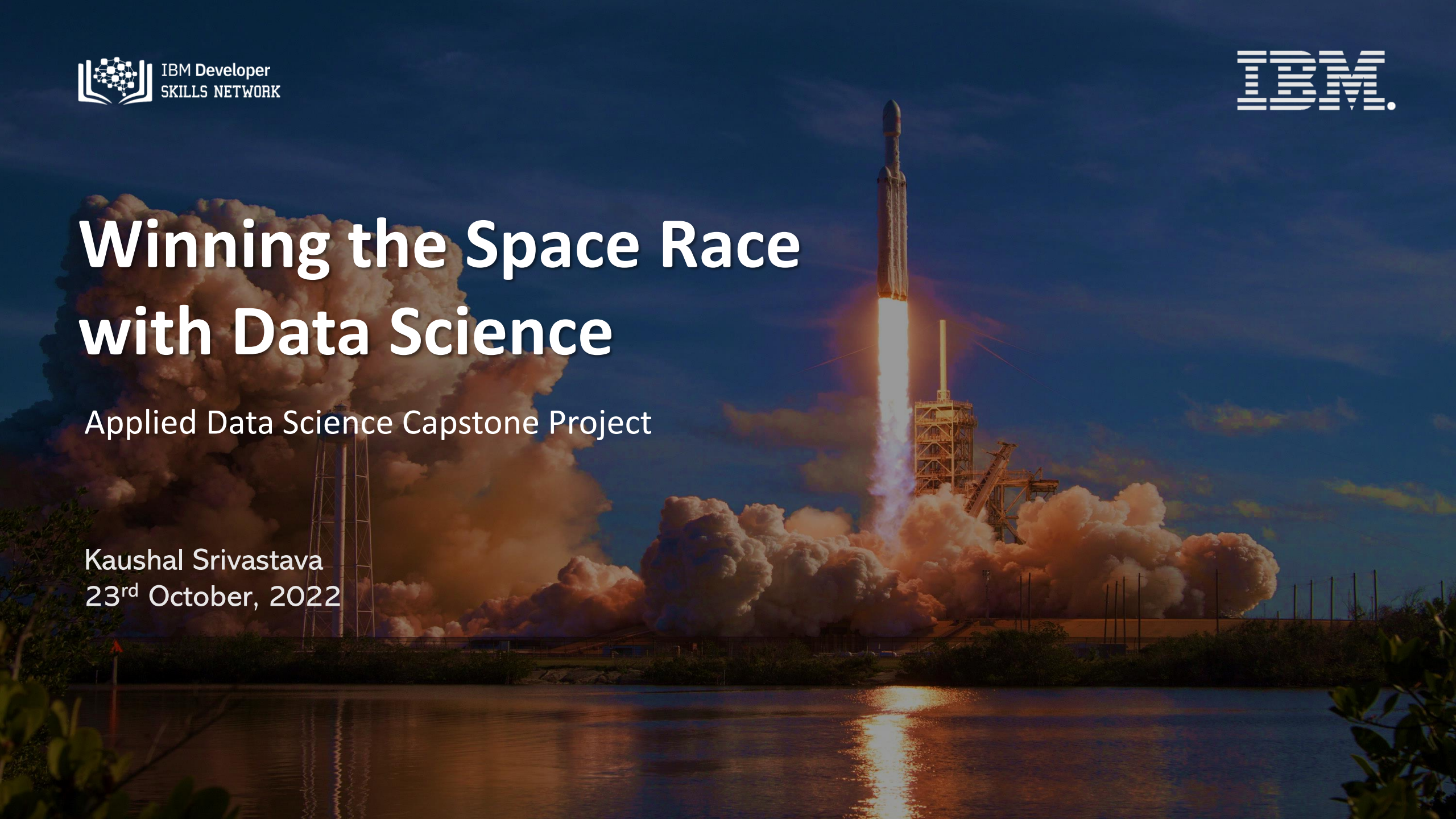


Winning the Space Race with Data Science

Applied Data Science Capstone Project

Kaushal Srivastava
23rd October, 2022



Outline

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

Executive Summary

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
 - Data collection, wrangling, and formatting
 - Exploratory data analysis
 - Interactive data visualization
 - Machine learning prediction
- Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.
- It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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.
- Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

A photograph of a Space Shuttle launching from a launchpad. The shuttle is ascending vertically, leaving a large, bright white plume of smoke and fire. The launchpad structure is visible to the right of the shuttle. In the foreground, there is a body of water reflecting the shuttle and the launchpad. To the left of the shuttle, there is a large, billowing cloud of white smoke. The sky is a deep blue with some wispy clouds. The overall scene is dramatic and captures the power of the shuttle launch.

Section 1

Methodology

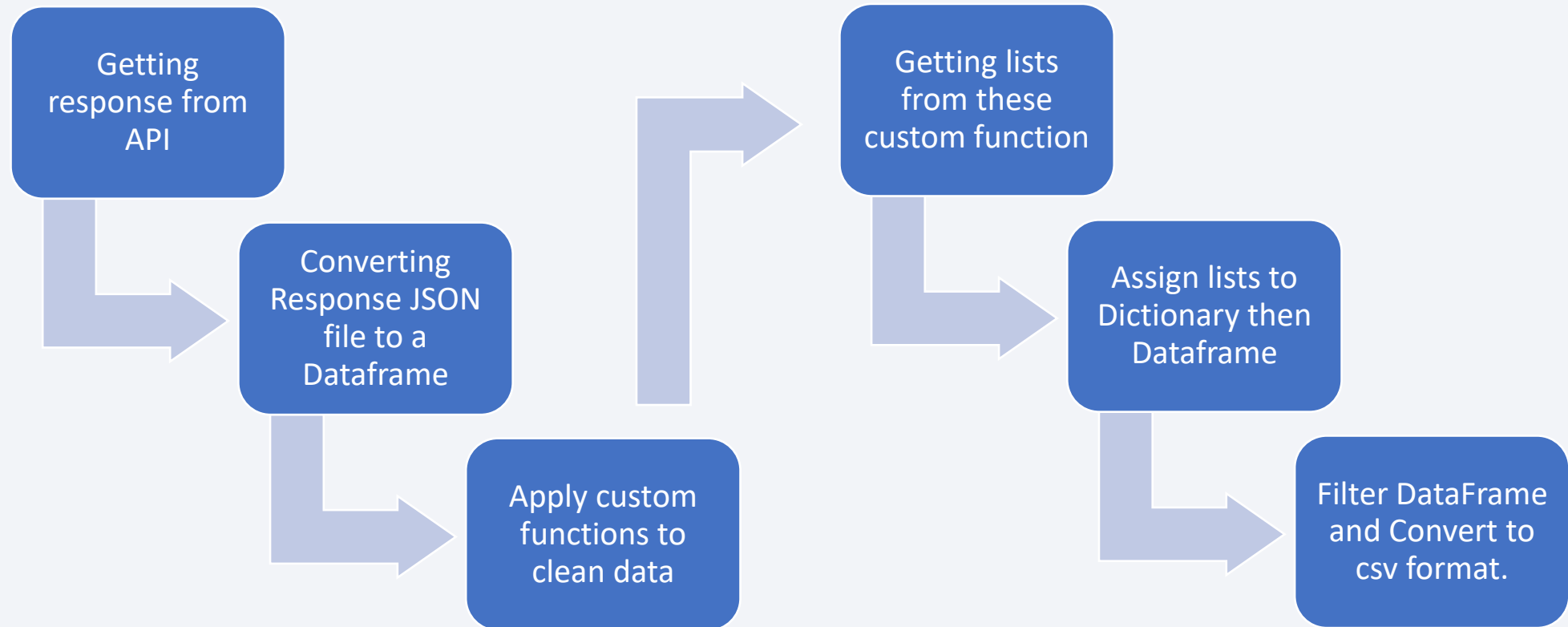
Methodology

- Data collection methodology:
 - SpaceX API
 - Web scraping SpaceX Wikipedia
- Perform data wrangling
 - Exploratory Data Analysis using Pandas
 - Determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Logistic regression
 - Support vector machine (SVM)
 - Decision tree
 - K-nearest neighbors (KNN)

Data Collection

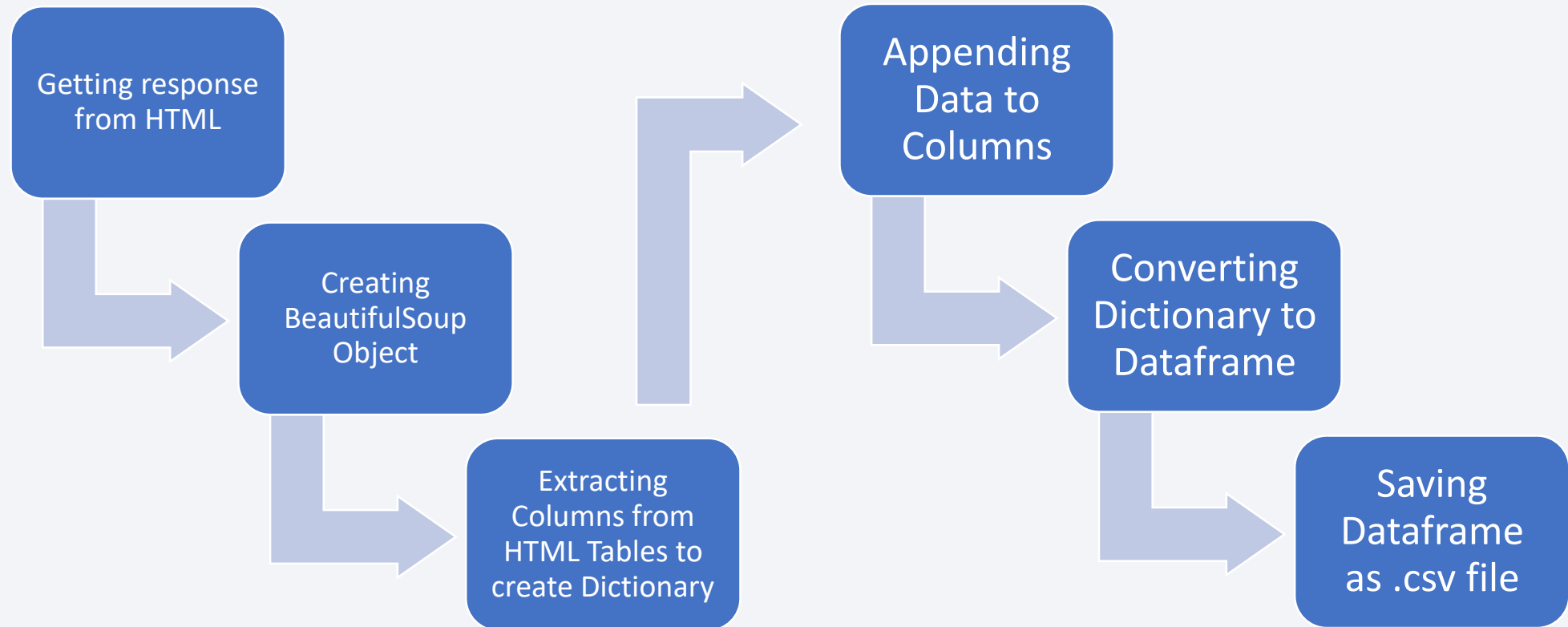
- Required data was fetched from SpaceX API by making a GET request to the SpaceX API.
- Data was also collected using web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled 'List of Falcon 9 and Falcon heavy launches' using python BeautifulSoup library.

Data Collection – SpaceX API



[Data Collection Using API GitHub Link : Click Here](#)

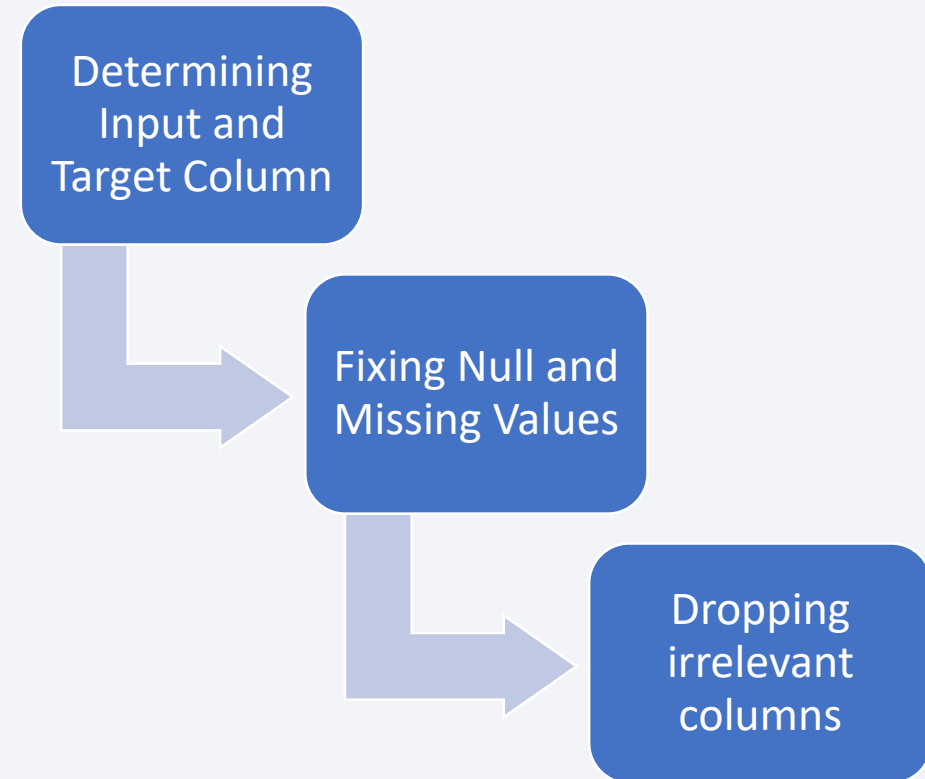
Data Collection – Web Scrapping



[Data Collection Using WebScraping GitHub Link : Click Here](#)

Data Wrangling

- Retrieved data was cleaned and transformed so that it can be used for modeling.
- Landing outcome column was modified to have only two values 1 for all kinds of success and 0 for all kinds of failure.
- Landing outcome was set as the Target feature.
- Missing values in Payload column were replaced with the mean value.
- Irrelevant columns were dropped.
- Columns having categorical values were one hot encoded and column was created for each categorical value dropping the original column.



[Data Wrangling GitHub Link : Click Here](#)

EDA with Data Visualization

Scatter Graphs being drawn:

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass

Bar Graph being drawn:

- Mean VS. Orbit

Line Graph being drawn:

- Success Rate Vs Year

[Data Visualization GitHub Link : Click Here](#)

EDA with SQL

1. Displayed the names of the unique launch sites in the space mission using DISTINCT command.
2. Displayed 5 records where launch sites begin with the string 'CCA' using WHERE and LIKE command.
3. Displayed the total payload mass carried by boosters launched by NASA (CRS) using SUM and WHERE command.
4. Displayed average payload mass carried by booster version F9 v1.1 using AVG command.
5. Listed the date when first successful landing outcome in ground pad was achieved using MIN command.
6. Listed the names of the boosters which have success in drone ship and have payload mass greater than
7. 4000 but less than 6000 using WHERE and BETWEEN command.
8. Listed the total number of successful and failure mission outcomes using COUNT command.
9. Listed the names of the booster versions which have carried the maximum payload mass using MAX command in a subquery.
10. Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 using WHERE and Web Scraping command.
11. Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order using GROUPBY and ORDERBY command.

[EDA with SQL GitHub Link : Click Here](#)

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster() to show successful/failure launches for each launch site.
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks

[Interactive Map with Folium GitHub Link : Click Here](#)

Build a Dashboard with Plotly Dash

- Pie chart of Launch Success counts for all sites and success rate of launches for each site was created in the dashboard.
- Dropdown was added to the dashboard to select launch site.
- Scatter Plot of Payload Mass vs Landing outcome was plotted with a slider for payload given to select payload range. Colors to data points with different booster versions.
- Pie Chart for all sites showed launch site with highest count of launch success.
- Pie chart for each launch site showed respective success rate with highest success rate of site KSC LC-39A.

[Dashboard with Plotly GitHub Link : Click Here](#)

Predictive Analysis (Classification)

BUILDING MODEL

- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

BEST MODEL SEARCH

- The best model is the model with the best accuracy score on the test data.
- Accuracy score of all models was same on test data hence model with highest accuracy score on training data was selected.

EVALUATING MODEL

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

[Predictive Analysis: Click Here](#)

Results

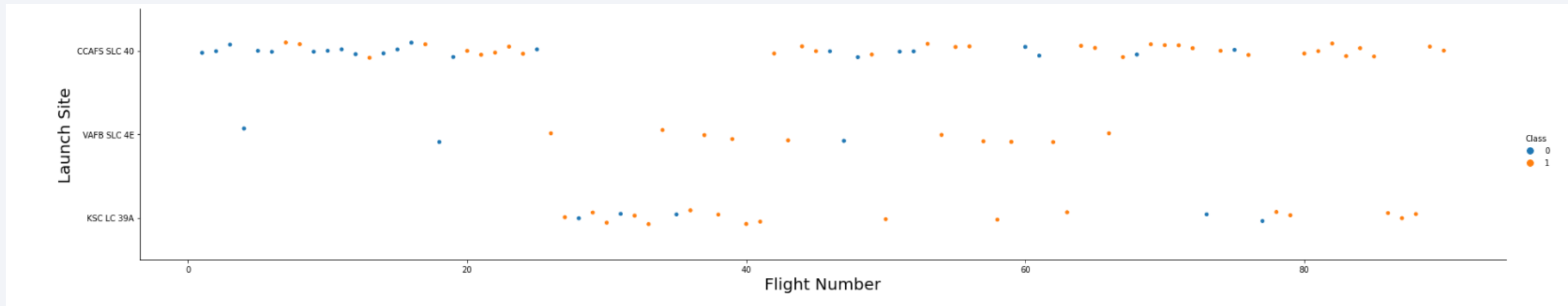
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

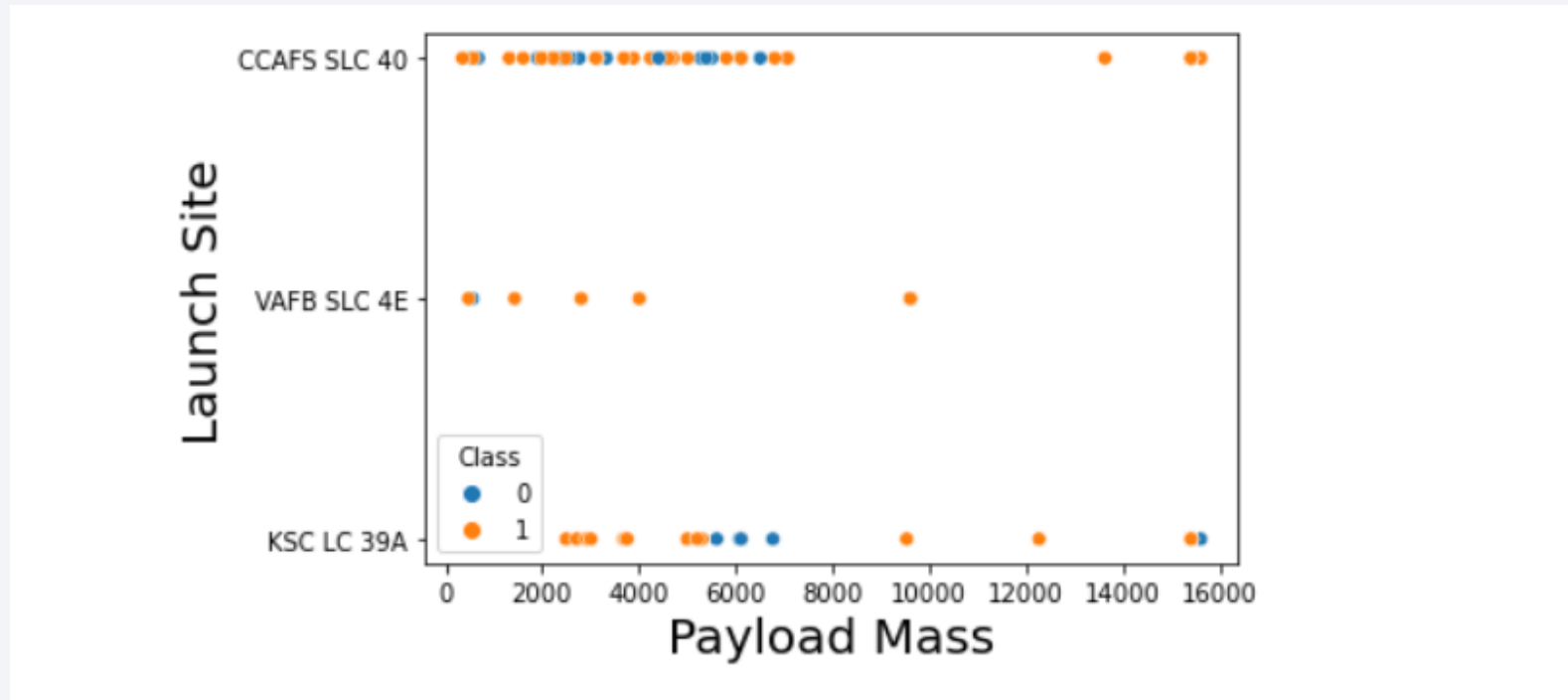


Flight Number vs. Launch Site



- Launch success rate has increased for all the launch sites from the first to the last launch.
- Most part of the launches were performed from CCAFS and KSC launch sites.

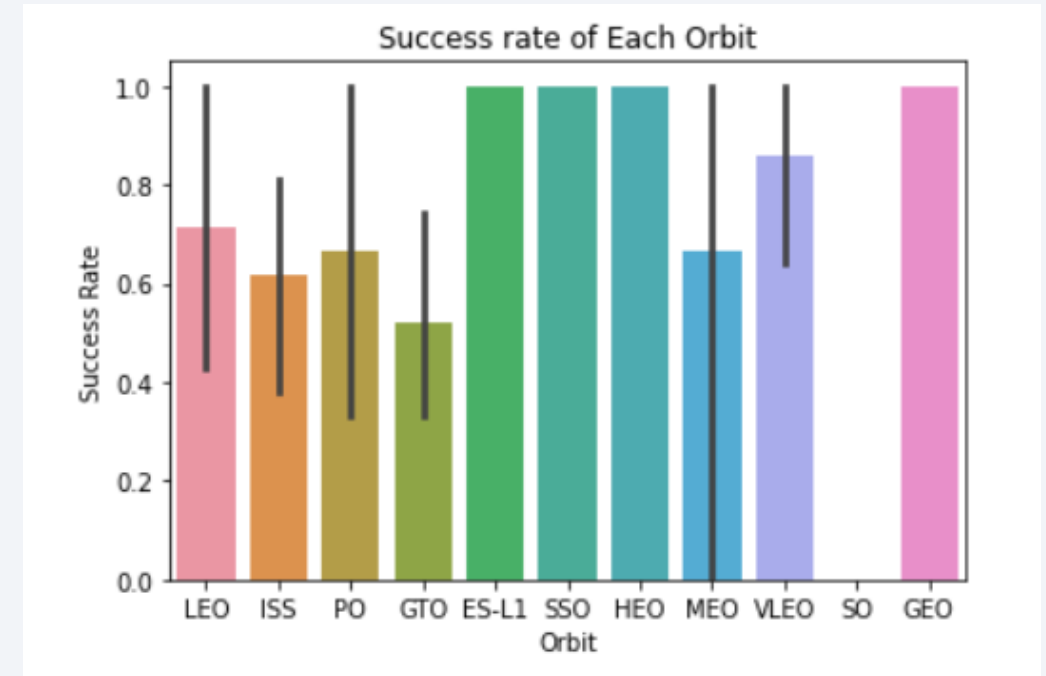
Payload vs. Launch Site



- Launch Success Rate is higher for payload range greater than 8000 kg.
- In VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

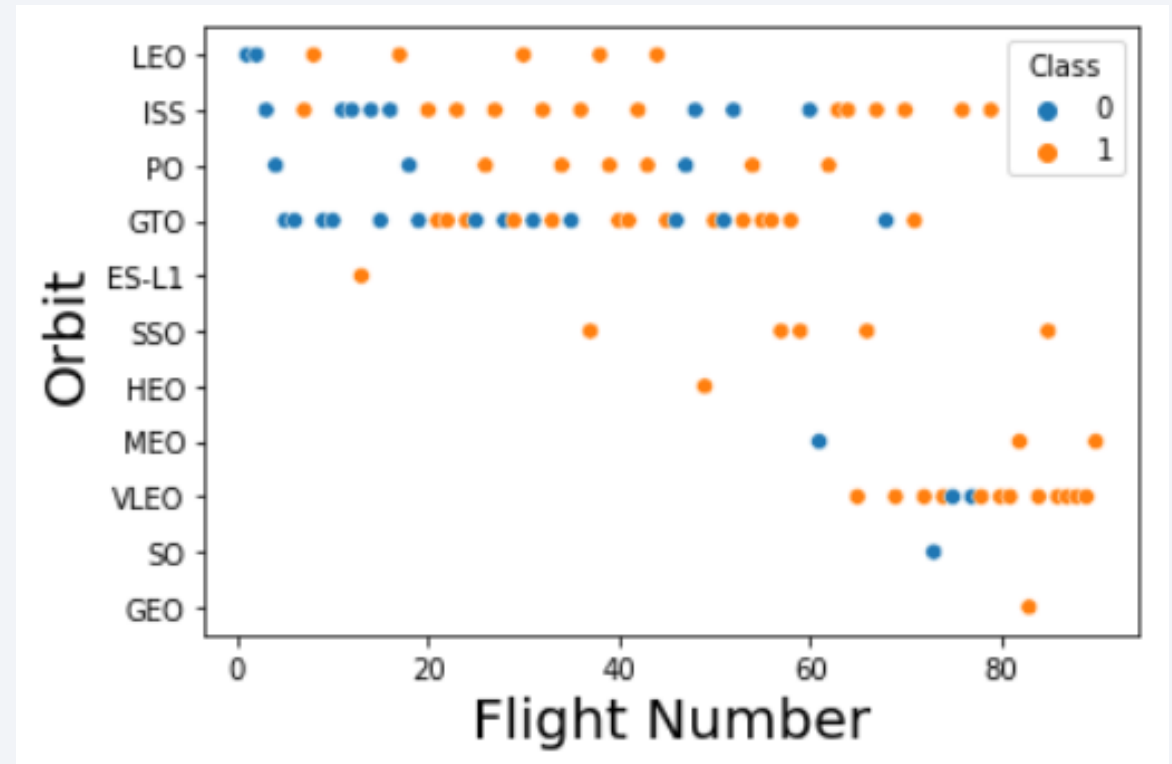
Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO and SSO have highest 100% success rate.
- No successful launch found for orbit type SO.
- The rest of orbit types have a success rate of around 60%.



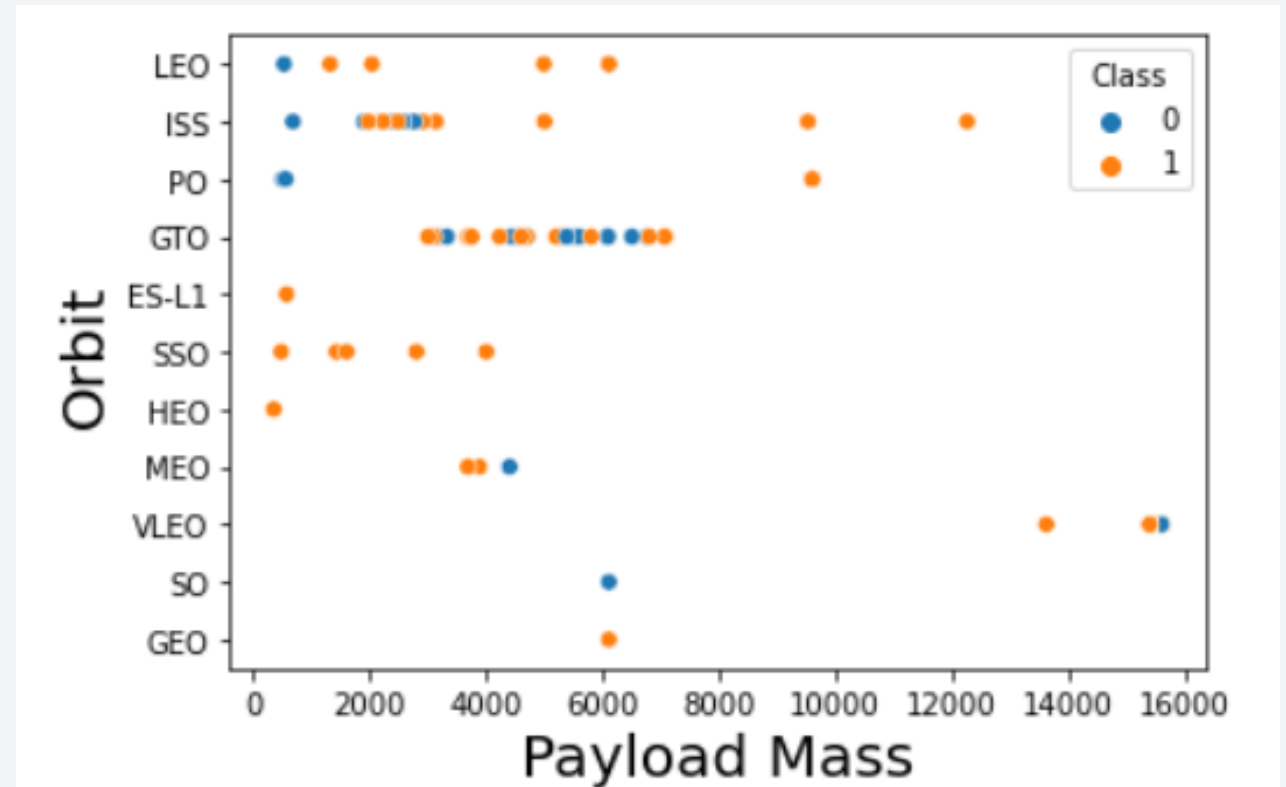
Flight Number vs. Orbit Type

- In LEO orbit the success appears related to the number of flights.
- There seems to be no relationship between orbit type and flight number for GTO and ISS orbits.



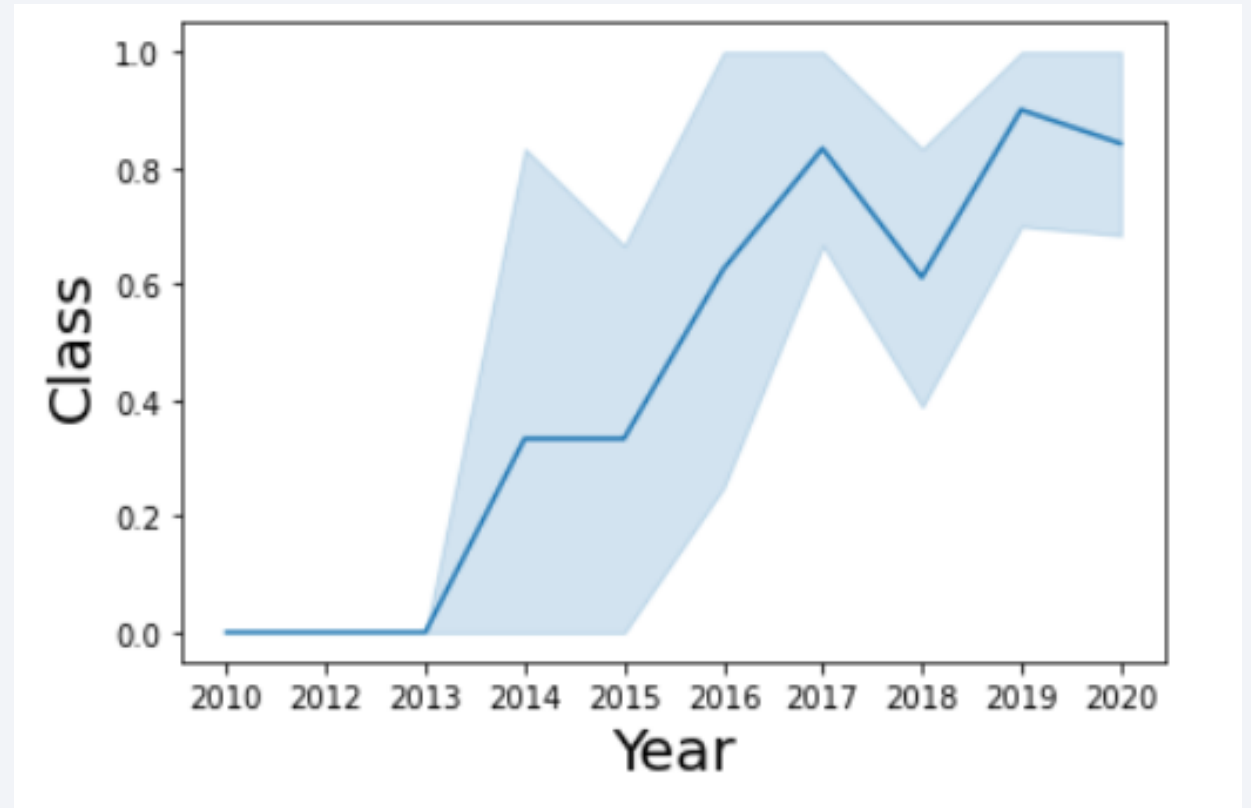
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccesful mission) are both there here.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

```
In [10]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

```
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb  
Done.
```

```
Out[10]: launch_site
```

```
CCAFS LC-40
```

```
CCAFS SLC-40
```

```
KSC LC-39A
```

```
VAFB SLC-4E
```

- We are using SQL DISTINCT query to fetch unique launch sites names

Launch Site Names Begin with 'CCA'

```
In [11]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

```
Out[11]:
```

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

- We are using LIKE SQL Query to find Launch site names that start with `CCA`
- % is used as wildcard character

Total Payload Mass

```
In [12]: %sql SELECT CUSTOMER, SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)' GROUP BY CUSTOMER
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

Out[12]:

customer	total_payload_mass
NASA (CRS)	45596

- We are grouping the rows by customer name “NASA CRS” and using the SUM SQL Query on the Payload Mass

Average Payload Mass by F9 v1.1

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

```
In [13]: %sql SELECT BOOSTER_VERSION, AVG(PAYLOAD_MASS_KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1' GROUP BY BOOSTER_VERSION
```

```
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
```

Done.

```
Out[13]: booster_version  avg_payload_mass
```

F9 v1.1	2928
---------	------

- We are grouping by booster version “F9 v1:1” and then using AVG to find the Average payload mass

First Successful Ground Landing Date

```
In [14]: %sql SELECT MIN(DATE) AS FIRST_SUCCESFUL_LAUNCH_DATE FROM SPACEXTBL WHERE MISSION_OUTCOME = 'Success'
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
Out[14]: first_succesful_launch_date
2010-06-04
```

- Using the MIN Command we are fetching the Date where the Mission outcome was successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [17]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8l1cg.databases.appdomain.cloud:31505/bludb
Done.
Out[17]: booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2
```

- We are using the WHERE and Polar, LEO command to filter the payload between 4000 and 6000 which was successful.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [19]: `%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_COUNT FROM SPACEXTBL GROUP BY MISSION_OUTCOME`

* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.

Out[19]:

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

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

- We are grouping by Mission Outcome and then using COUNT command to find the total number of successful and Failure Mission Outcomes.

Boosters Carried Maximum Payload

```
In [23]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

```
Out[23]: 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

- We are running a query within a query.
- We are using MAX command to find the maximum Payload.

2015 Launch Records

```
In [28]: %sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXTBL WHERE DATE LIKE '2015%' AND LANDING__OUTCOME = 'Failure (drone ship)'
```

```
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

```
Out[28]:
```

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)
------------	---------------	-------------	----------------------

- We are using WHERE and LIKE command to filter the records that have 2015 in the Date column and using AND command to also check if the landing outcome is Failure.

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

```
In [35]: %sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS OUTCOME_COUNT FROM SPACEXTBL WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20'  
GROUP BY LANDING__OUTCOME ORDER BY OUTCOME_COUNT DESC
```

```
* ibm_db_sa://wpg14082:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb  
Done.
```

Out[35]:

landing__outcome	outcome_count
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

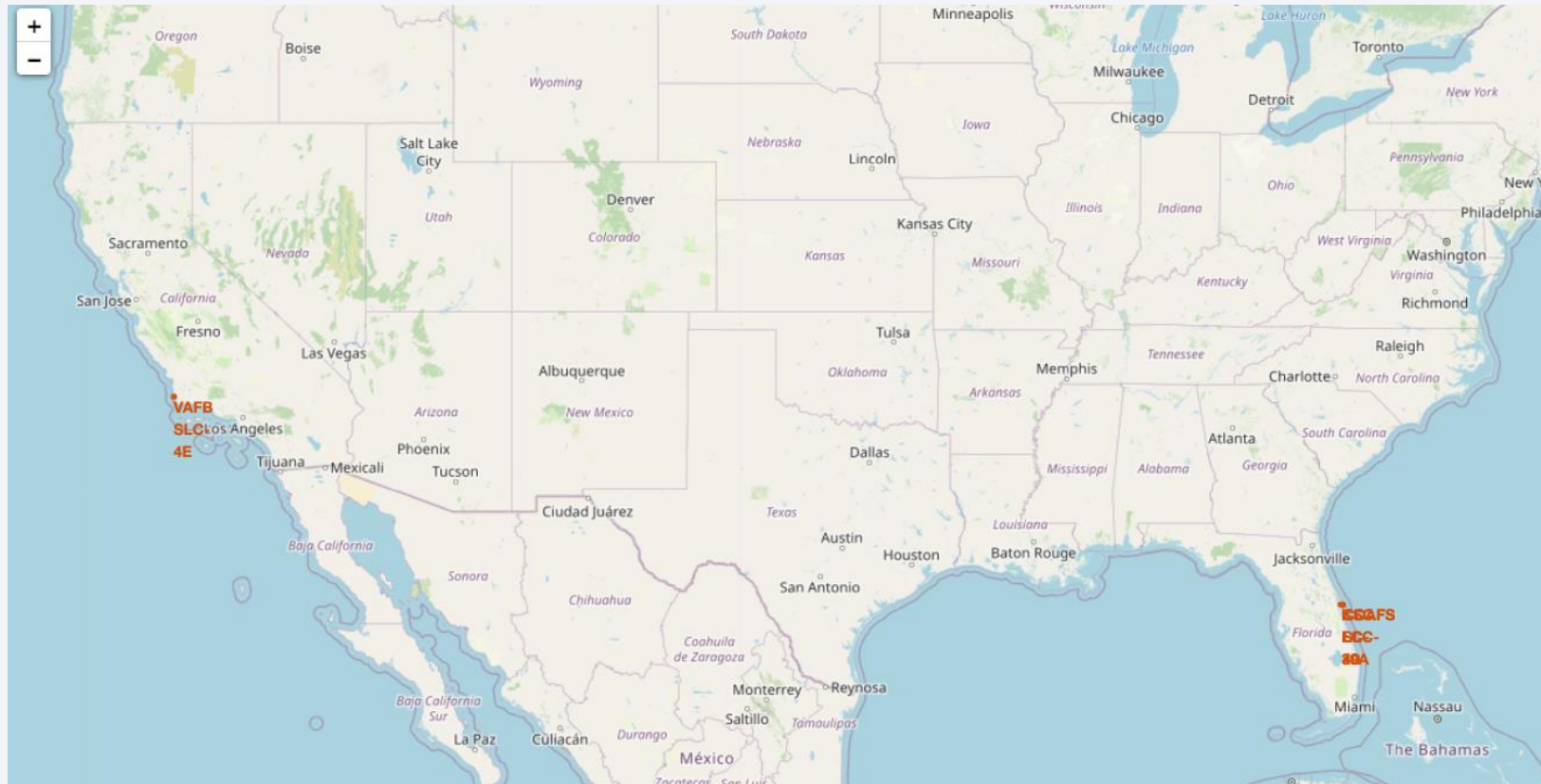
- We are filtering between the date range using WHERE and then grouping by landing outcome and finally using ORDER BY DESC to arrange the result in descending order.

Section 3

Launch Sites Proximities Analysis

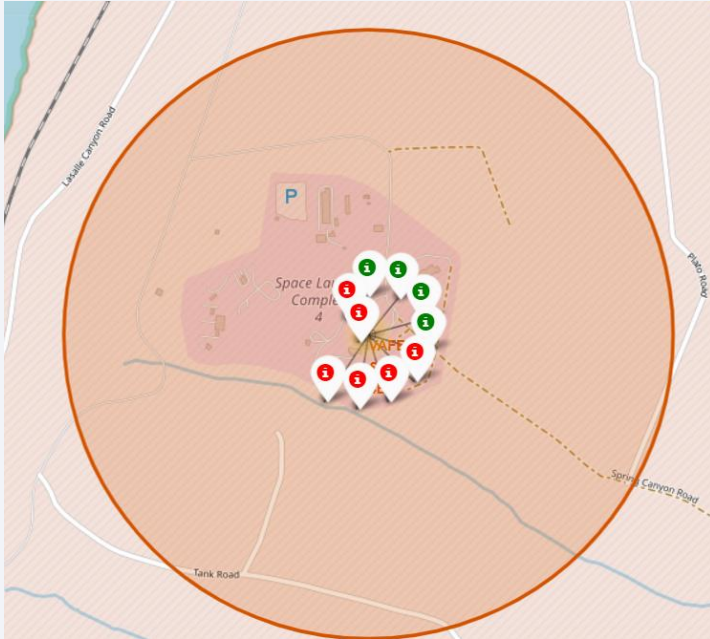


Launch Sites Plotted on the Map

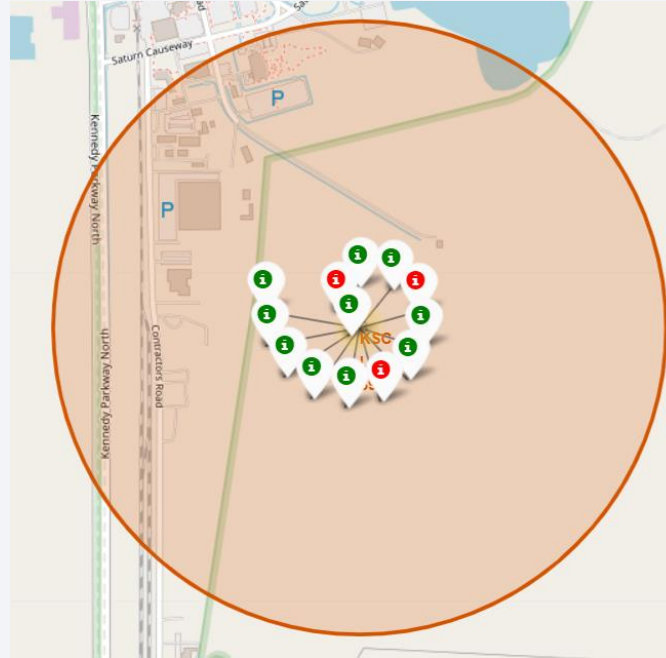


- Three launch sites are close to each other in Orlando, Florida and only one launch site is in California.
- All launch sites are in very close proximity to the coast so that first stage can be thrown to the sea after the take-off.

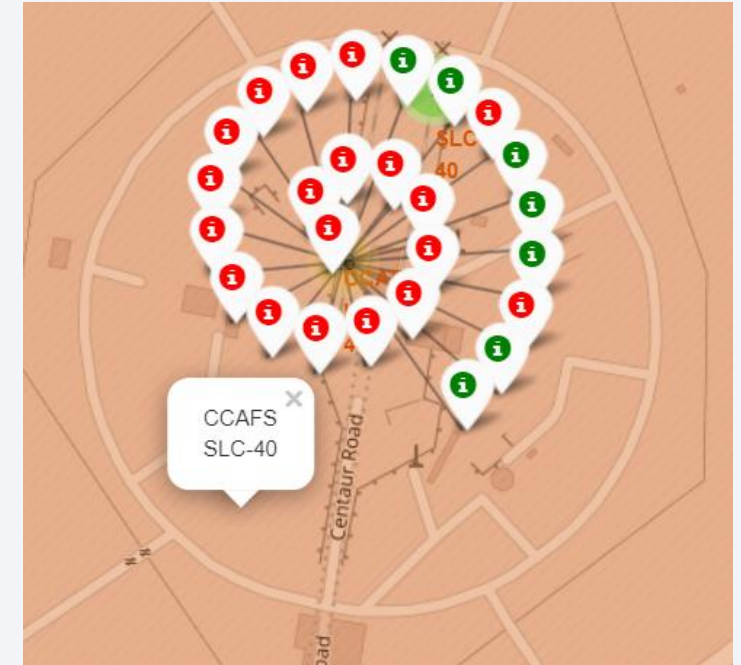
Success and Failure Launches shown in Map



VAFB SLC-4E



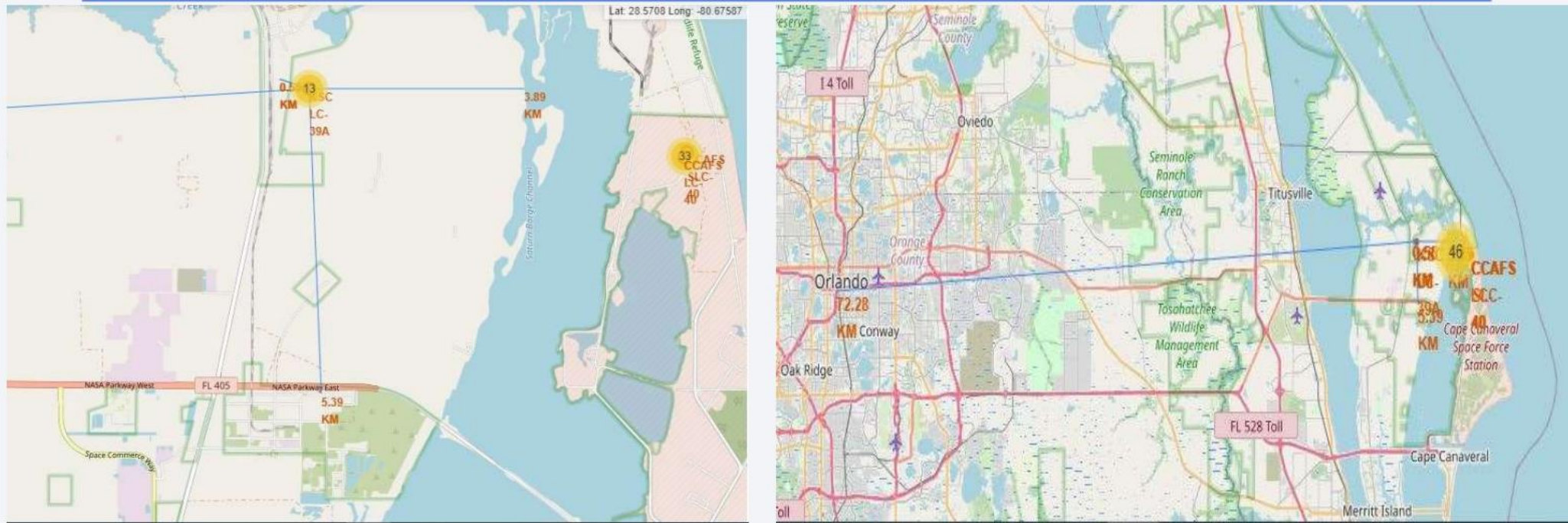
KSC LC-39A




CCAFS SLC-40

Color based Success and Failure markers for launches at sites. Green shows successful launches and Red shows Failed launches.

Distance between Launch sites and proximities like railway, highway, coastline etc.



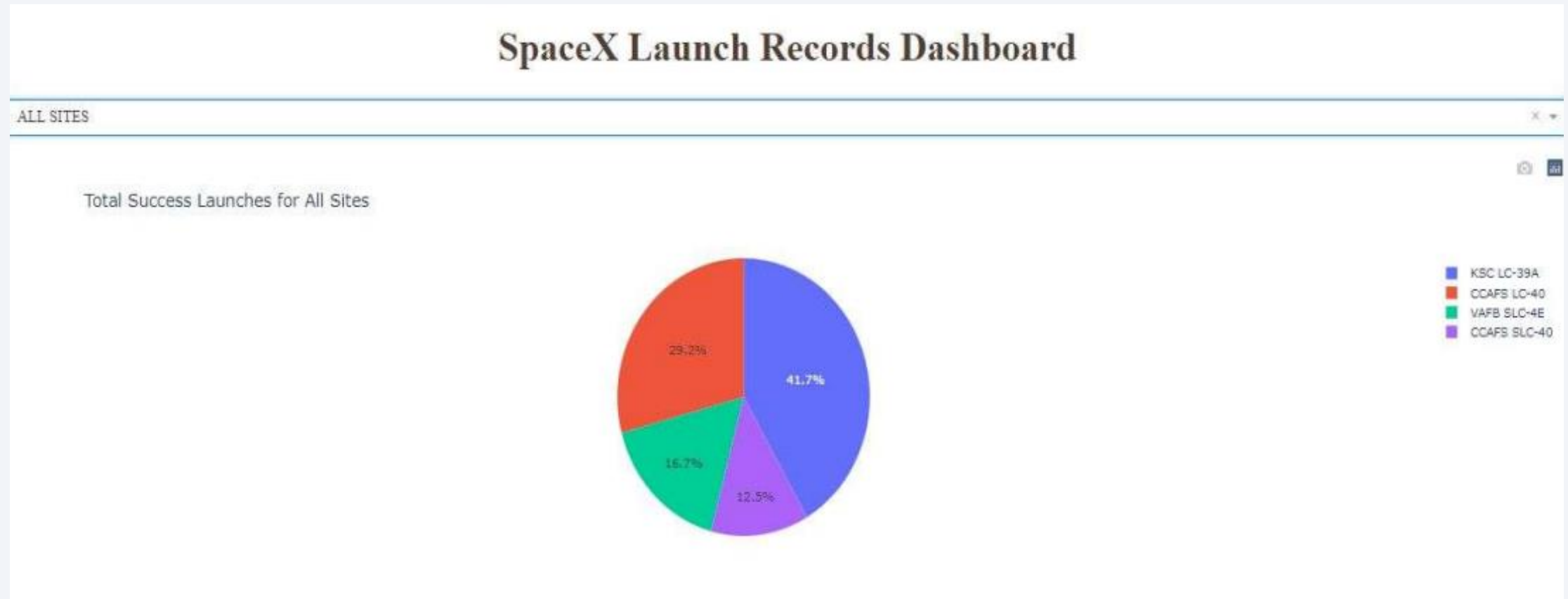
- Launch sites are close to railways and highways for logistic reasons.
- They are also close to the coastline so that first stage can be thrown to the sea after the take-off.
- On the contrary, they keep certain distance away from cities.



Section 4

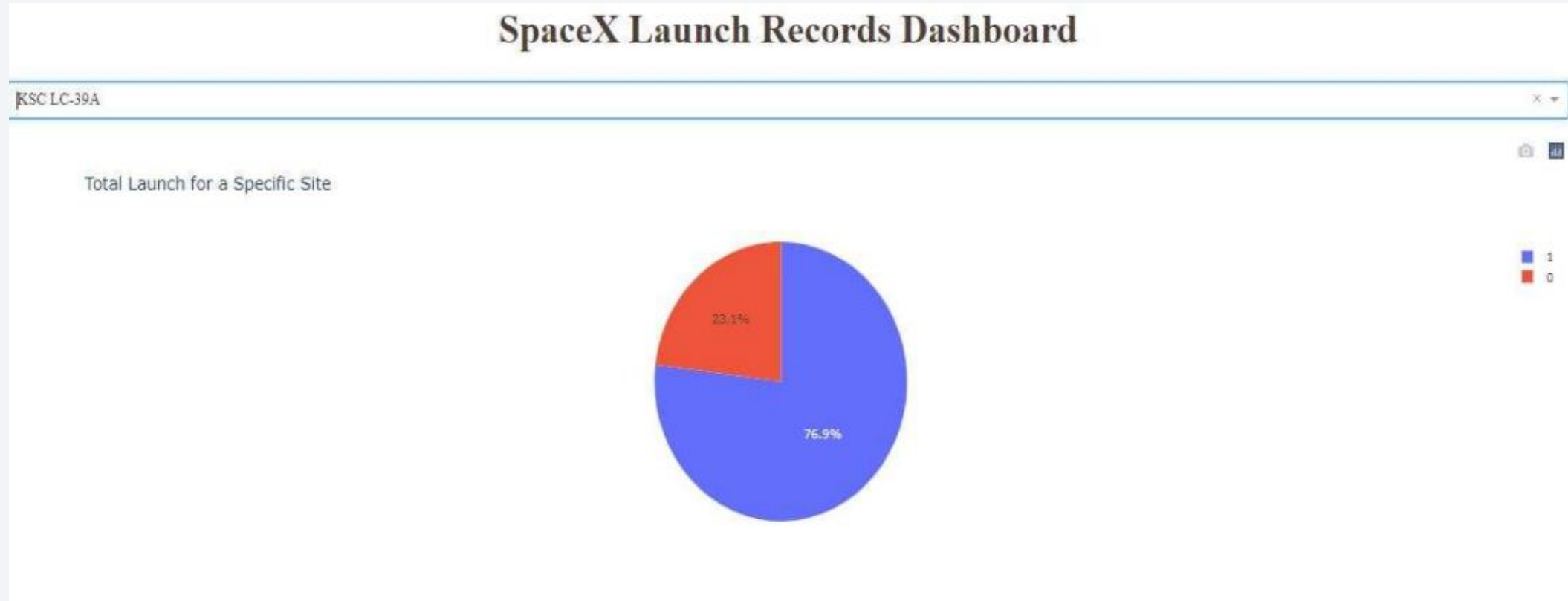
Build a Dashboard With Plotly Dash

Launch Success Count for all sites Pie Chart



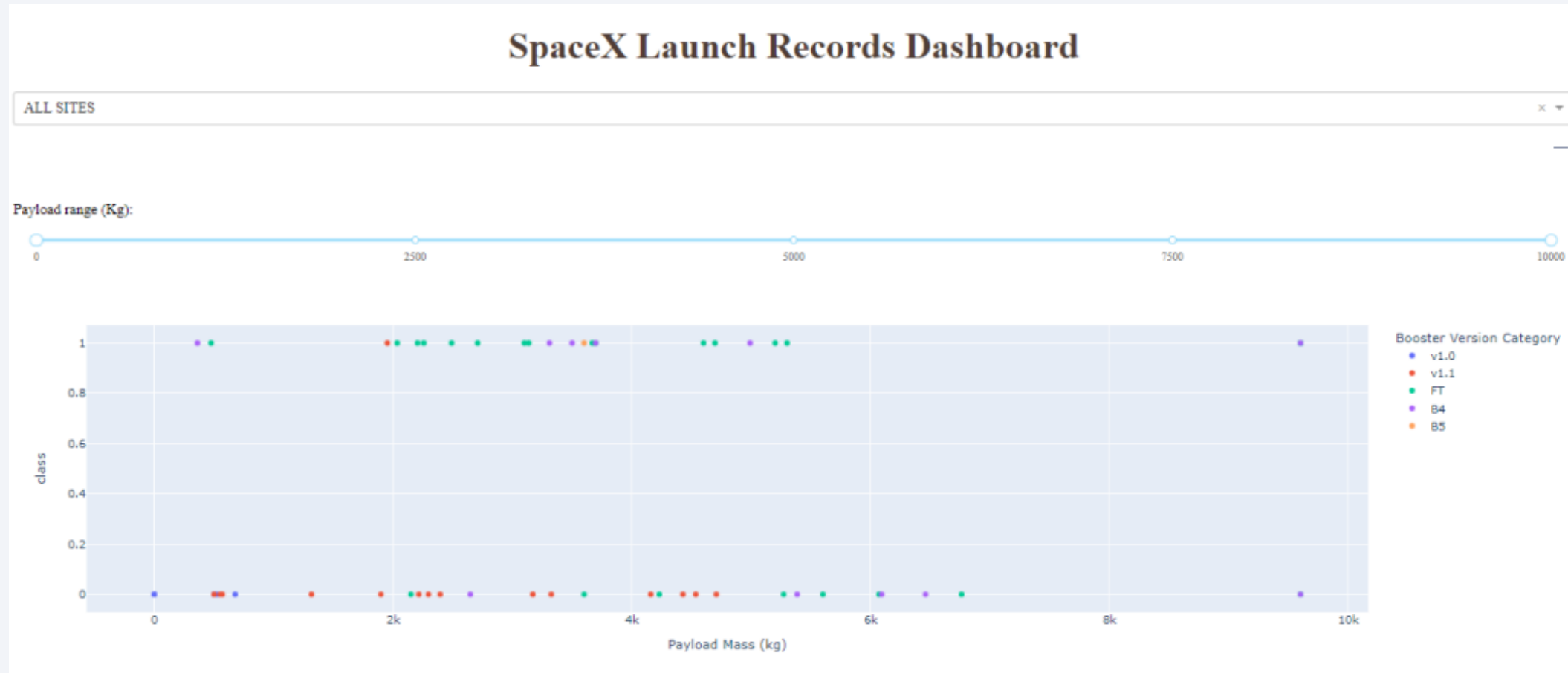
- Launch Site KSC-LC-39A has highest count of successful launches.

The pie chart for the launch site with highest launch success ratio



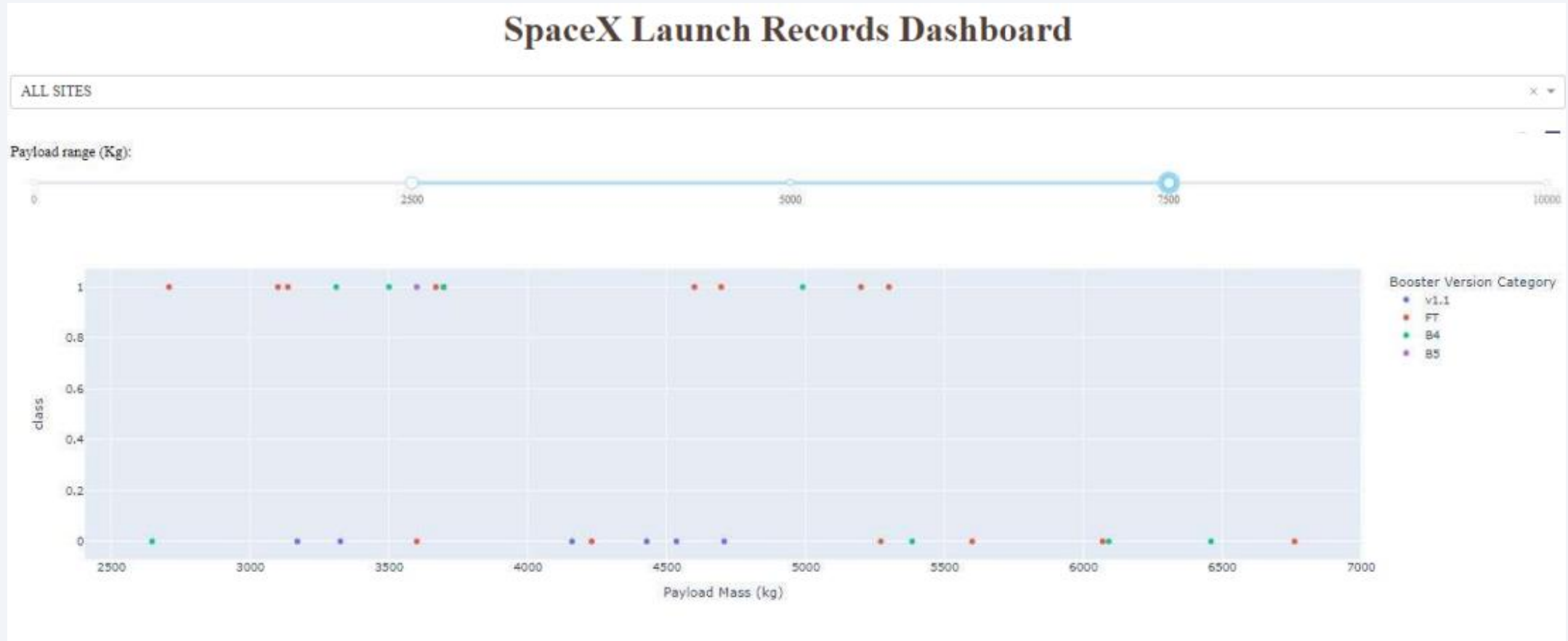
- Launch Site KSC-LC-39A has highest success rate of 76.9%

Payload vs. Launch Outcome scatter plot for all sites



- A success rate of 0% is obtained for payload masses in the range 6000-7000 kg

Payload vs. Launch Outcome scatter plot for all sites (Cont.)



- Most part of the launches are carried out with a payload mass which varies from 2000 to 7000 kg

Section 5

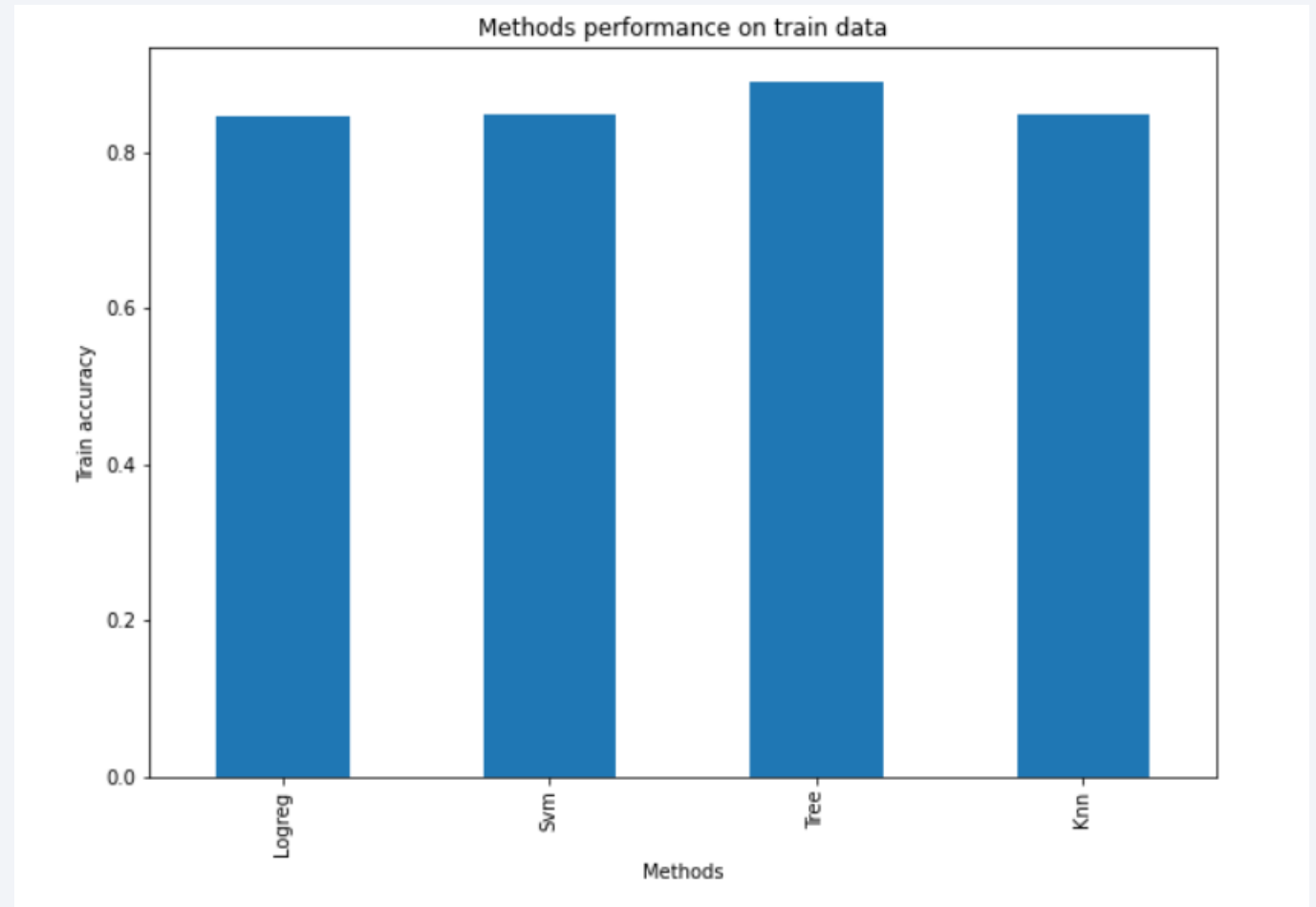
Predictive Analysis (Classification)



Classification Accuracy

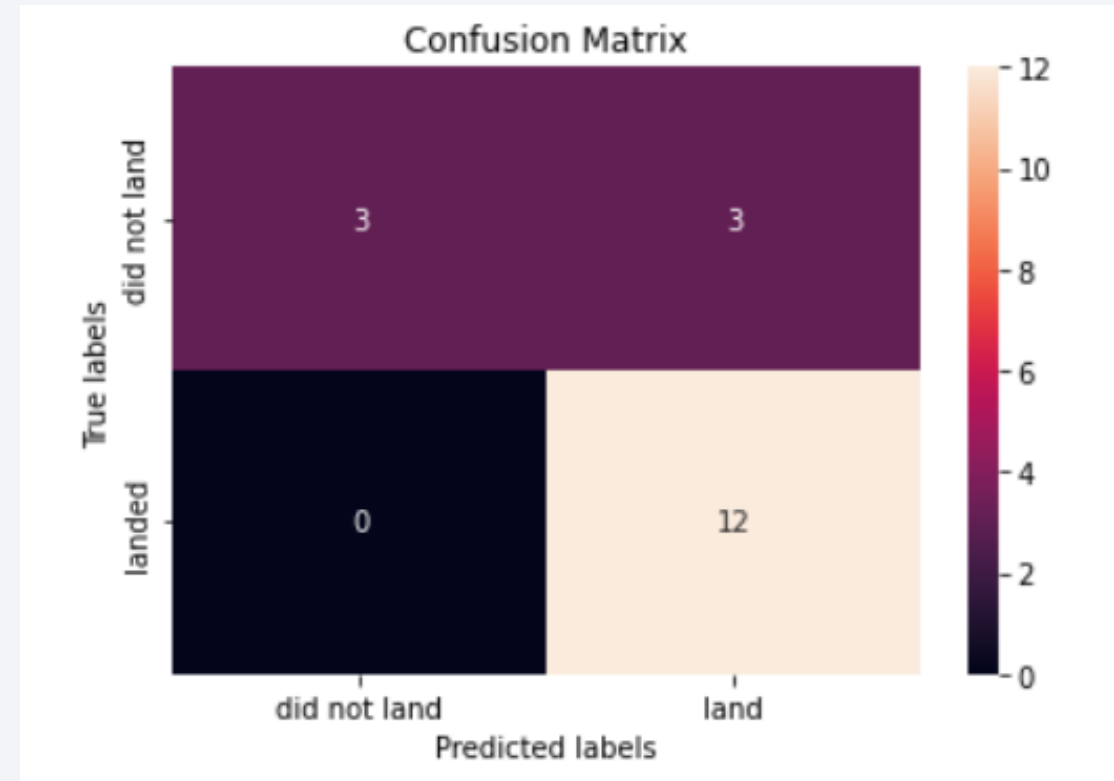
- We observe that the Decision Tree model has the Highest Test Accuracy (around 88.92 %)
- We achieved an accuracy of 83.33% on test data.

	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.889286	0.833333
Knn	0.848214	0.833333



Confusion Matrix of best performing model

- The best performing model was found to be **Decision Tree Model**.
- The confusion matrix of the same shows that the model can distinguish between the classes properly except for false positives cases in which the model says that the launch landed when it did not land.



Conclusions

- The Decision Tree Model performed the best for the dataset.
- Launch success rate since 2013 kept increasing till 2020.
- Success rate improves as number of flights increases.
- Most part of recent launches were made from CCAFS LC-40 with a success rate of 73.1% and from KSC LC-39A with a success rate of 76.9%.
- We can see that KSC LC-39A had the most successful launches from all the sites and also the highest success rate.
- Most part of the launches are carried out with a payload mass which varies from 2000 to 7000 kg with a good success rate. Heavy payload missions (payload mass > 8000 kg) have a higher success rate.
- Launch sites are located in very close proximity to the coast, railway and highway. On the contrary, they maintain a certain distance to the cities.
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate.
- No successful launch found for orbit type SO.

Appendix

Python code for comparing the results of different models and creating a Dataframe :

```
methods = ['Logreg', 'Svm', 'Tree', 'Knn']
accs_train = [logreg_cv.best_score_, svm_cv.best_score_, tree_cv.best_score_, knn_cv.best_score_]
accs_test = [acc_logreg_test_data, svm_test_score, tree_cv_test_score, knn_test_score]

dict_meth_accs = {}

for i in range(len(methods)):
    dict_meth_accs[methods[i]] = [accs_train[i], accs_test[i]]

df = pd.DataFrame.from_dict(dict_meth_accs, orient='index')
df.rename(columns={0: 'Accuracy Train', 1: 'Accuracy Test'}, inplace = True)

df.head()
```

Thank You !