IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sumit Agrawal
16-09-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data has been collected from SpaceX API and Wikipedia website. It has been preprocessed and analyzed.
  - Four machine learning algorithms has been investigated and also their hyper-parameters are tuned for obtaining best performance.

- Summary of all results
  - With different algorithm it is possible to tell that a rocket will land or not with 83% of accuracy.

Introduction

- Space X 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 Space X can reuse the first stage.
- We need to determine if the first stage will land, and so the cost of a launch.
- We also need to identify factors by which the launch can be successful.

Section 1

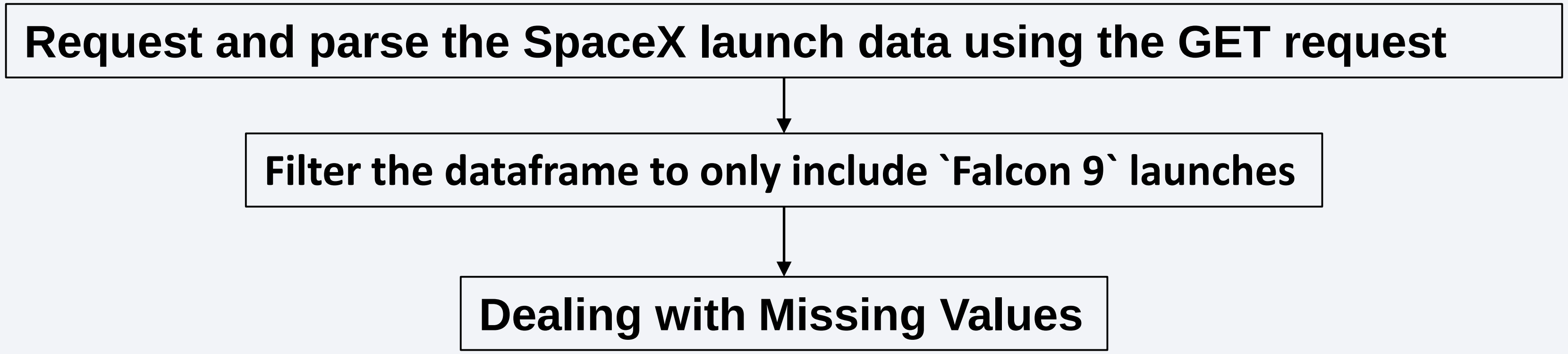
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API and Web Scraping. Important data are launch date, launch site, booster versions, outcome of launch etc.
- Perform data wrangling
  - Set the outcomes into Training Labels with `1` for booster successfully landed `0` for unsuccessful landing.
- 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, SVM, KNN and Decision tree methods are tuned and accuracy of each method has been determined.

Data Collection - SpaceX API



Request and parse the SpaceX launch data using the GET request

Filter the dataframe to only include `Falcon 9` launches

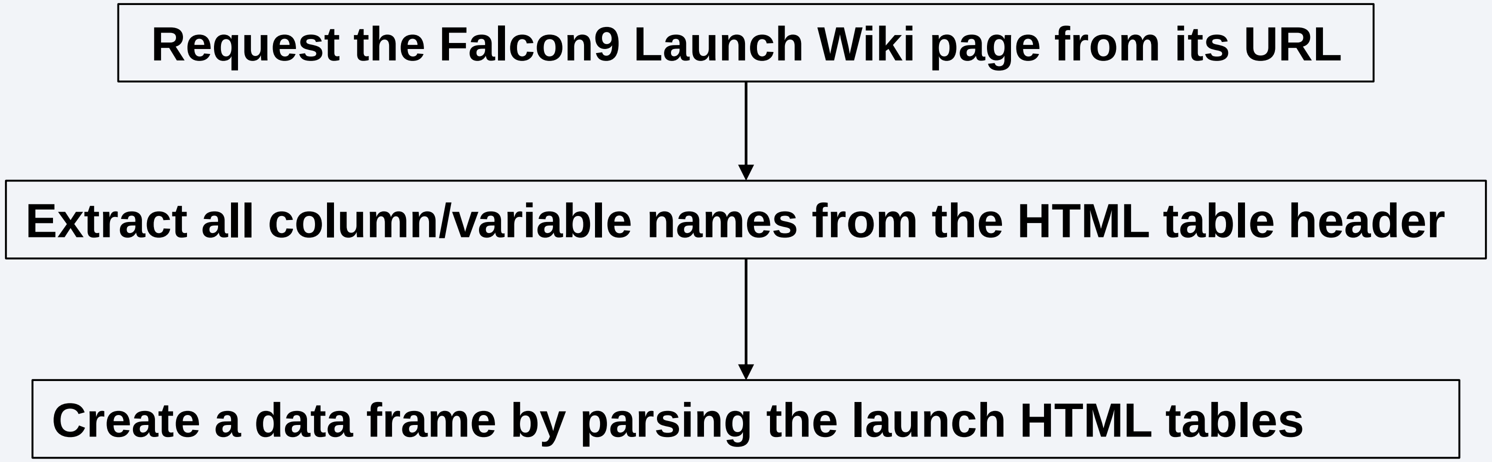
Dealing with Missing Values

Watson studio link:-

https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/b0005603-4003-4312-9985-ed8817ef0fcf/view?access_token=ba51ef4e1041e3022992a3879178625b8b16b8af6fd9a7e068911fe038a46876

Data Collection - Scraping

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



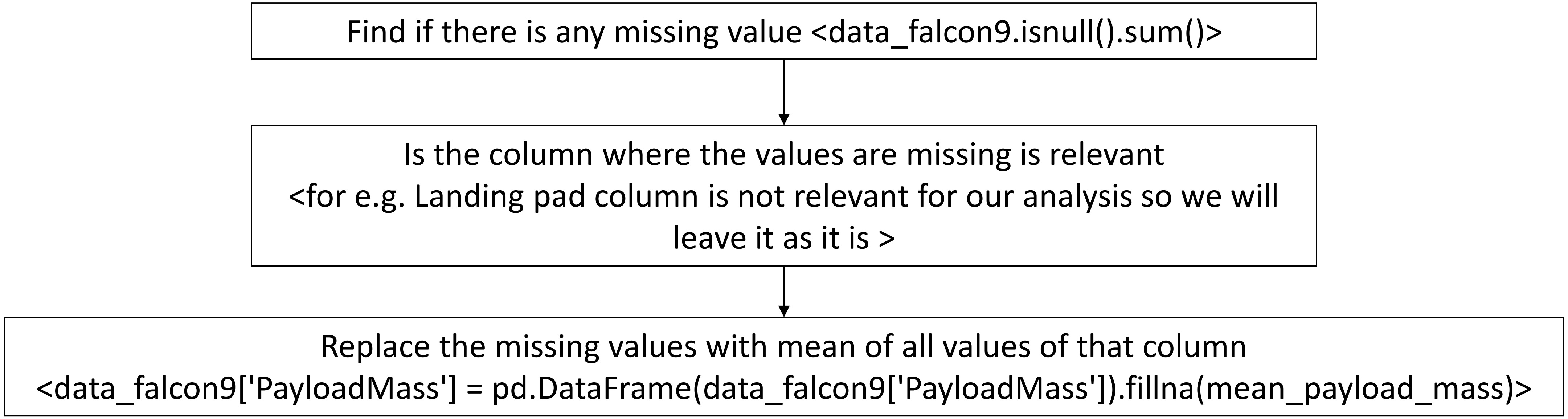
Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Watson studio link:-
https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/654bd4af-be4a-4e52-822d-ecb69501c27d/view?access_token=f969bd3b4c28232e0db8492b0021f8c95b78945db4a9d8b1c1c0d6d10f843f51

Data Wrangling



Find if there is any missing value <data_falcon9.isnull().sum()>

Is the column where the values are missing is relevant
<for e.g. Landing pad column is not relevant for our analysis so we will leave it as it is >

Replace the missing values with mean of all values of that column
<data_falcon9['PayloadMass'] = pd.DataFrame(data_falcon9['PayloadMass']).fillna(mean_payload_mass)>

EDA with Data Visualization

- Flight number v/s launch site (scatter plot)
  - To see any correlation between flight number and launch site. Does choice of launch site affect the outcome of launch?
- Flight number v/s payload mass (scatter plot)
  - How the payload mass has been changed with flights? Does mass of payload matters?
- Payload mass v/s launch site (scatter plot)
  - All kind of payload mass can be flown from each launch site or not? Is there any pattern in success of launch with launch site?
- Success rate of each orbit type (bar chart)
  - Does orbit type determines the rate of success of launches?
- Flight number v/s orbit type (scatter plot)
  - Over the time what orbits has been chosen for launch?
- Payload mass v/s orbit type (scatter plot)
  - Does payload mass and orbit type affect the outcome?
- Yearly trend of launch success (line chart)
  - The outcome of launch has been improved or not as the time progresses?

Watson studio link:-
https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/f4baf2cc-1298-44bf-8670-add8ebd7b7a5/view?access_token=bb88e1a5eb97f3647dd3465dabe2455c30a8ce2d4849d252eaecc3d5a17c21ee

EDA with SQL

- Name of launch sites
- Total payload mass launched for a particular customer
- Total payload mass carried by a particular booster
- Date, when the first successful landing was achieved
- Total number of mission outcomes : success and failure
- Lading outcomes between certain dates

Watson studio link:-
https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/238c20d8-f197-4afa-b33c-144ce9660ab7/view?access_token=c5fed88f6bef93409fa295355712d866d254f74dd517fac62983363be0de3d4a

Build an Interactive Map with Folium

- Marking all launch sites on a map
  - To know where are all the launch sites.
- Marking the success/failed launches for each site on the map
  - Does the outcome of launch depend on launch sites?
- The distances between a launch site to its proximities
  - Does proximity of launch site to railroad, highway or city matters in the outcome of the launch. Also if the proximity is low then cost would also becomes low.

Watson studio link:-

https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/0f4a8d7f-409b-41b8-acd6-d06ad6cc00f9/view?access_token=dc5ffa8a89aa75b78d9642596460eff30b257855933ad5d4df34627e39f14d98

Build a Dashboard with Plotly Dash

- Success rate of all launch sites and of individual launch site
  - It helps in visualizing which launch site provides best outcome.
- Payload and outcome, and booster version
  - It helps in understanding what booster can carry how much of payload mass and what was the outcome of those launches.

Watson studio link:-
https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/a49ff394-8089-4275-a7c7-0d4c3277d3bb/view?access_token=19de3c10d5cfc32820123d85cea56e20329379f3ce6f692d7b0b7a37d274b819

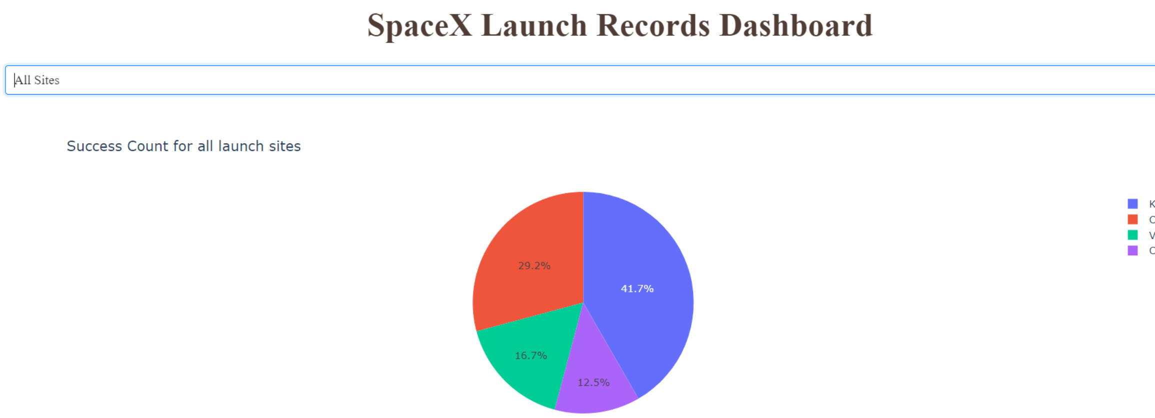
Predictive Analysis (Classification)

- The data has been split into train data and test data.
- Train data is used to build different algorithms such as logistic regression, KNN, SVM and decision tree.
- Many hyper-parameters has been identified for each of the algorithms and best of them has been decided using test data.
- Also the accuracy for each model with best identified hyper-parameters is evaluated on test data.

Watson studio link:-

https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/fa16258e-81f2-43b7-b988-0d5a515bab76/view?access_token=f89ab34d1b50f9d732091f9a13d12af42703101d46e27cdb7d26166b1b68f83d

Results



SpaceX Launch Records Dashboard

All Sites

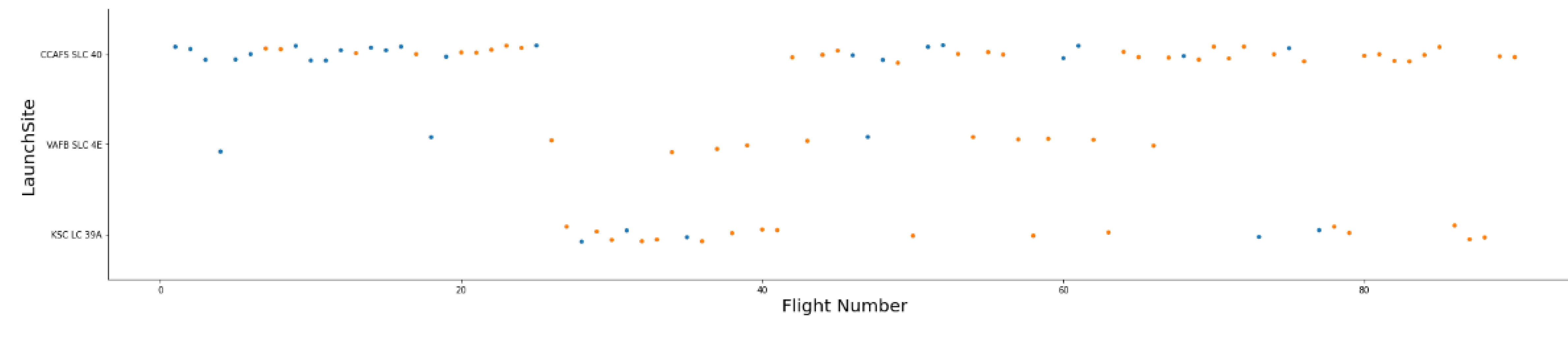
Success Count for all launch sites

- Exploratory data analysis results:
  - As time progresses the success rate increases.
  - Information about launch site, proximities etc. has been identified.
- Interactive analytics demo in screenshots
  - KSC LC 39A seems to be good launch site as per the outcomes.
- Predictive analysis results
  - Decision tree is found to be the best model for predicting the outcome of launches.

Section 2

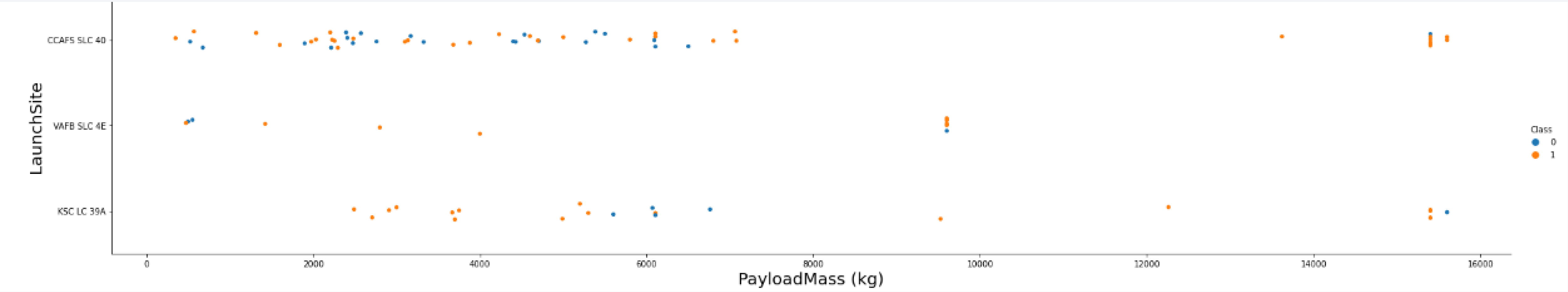
Insights drawn from EDA

Flight Number vs. Launch Site



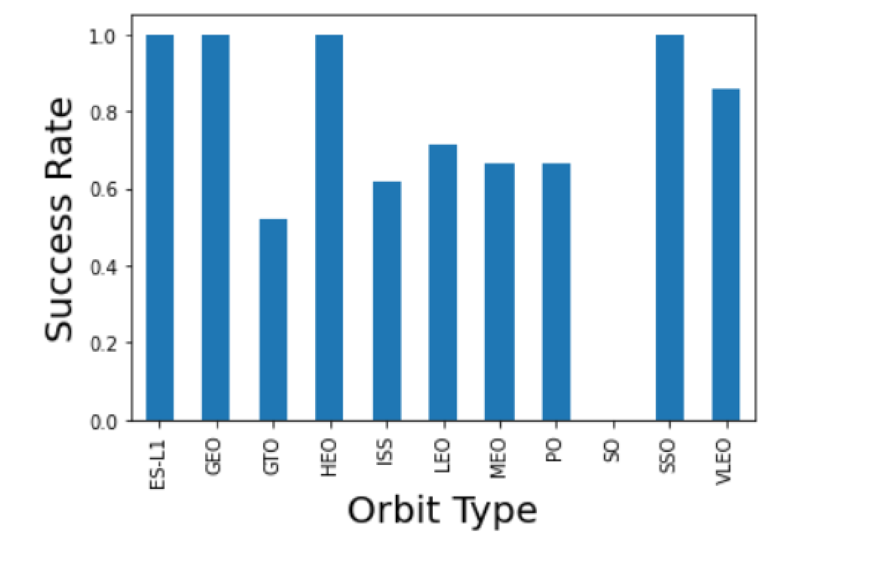
1. As the flight number increases the success of landing from each launching site improves.
2. Most of the rockets has been launched from CCAFS SLC 40.

Payload vs. Launch Site



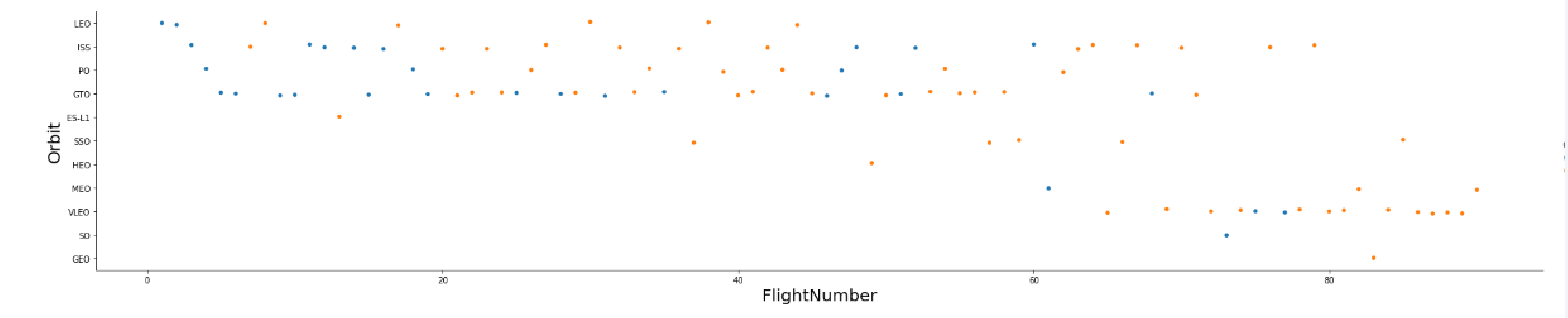
1. For payload higher than 10,000 kg VAFB SLC 4E was not chosen.
2. For payload lower than 2000 kg KSC LC 39A was not chosen.
3. Seems like for payload 2000 kg to 5000 kg KSC LC 39A gives best outcome among other launching sites.

Success Rate vs. Orbit Type



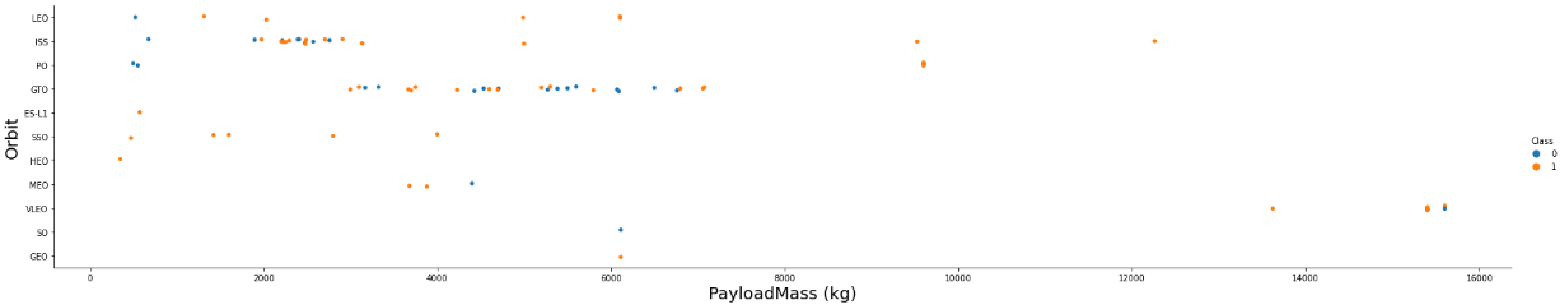
1. Many of the orbits such as ES L1, GEO, HEO, SSO gives 100% success rate.
2. GTO orbit has the lowest success rate.

Flight Number vs. Orbit Type



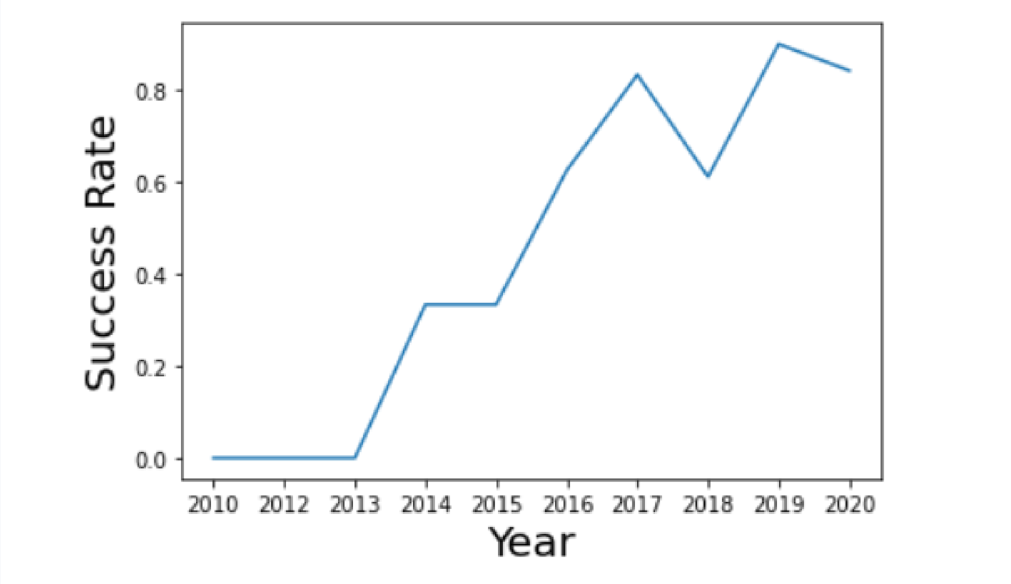
1. As the flight number increase the success rate increases as failures provide some learning.
2. Initial flights were from orbits GTO, LEO, ISS and PO that is why there the overall success rate is not good (as seen from previous slide).
3. VLEO orbit seems like to be preferred now.

Payload vs. Orbit Type



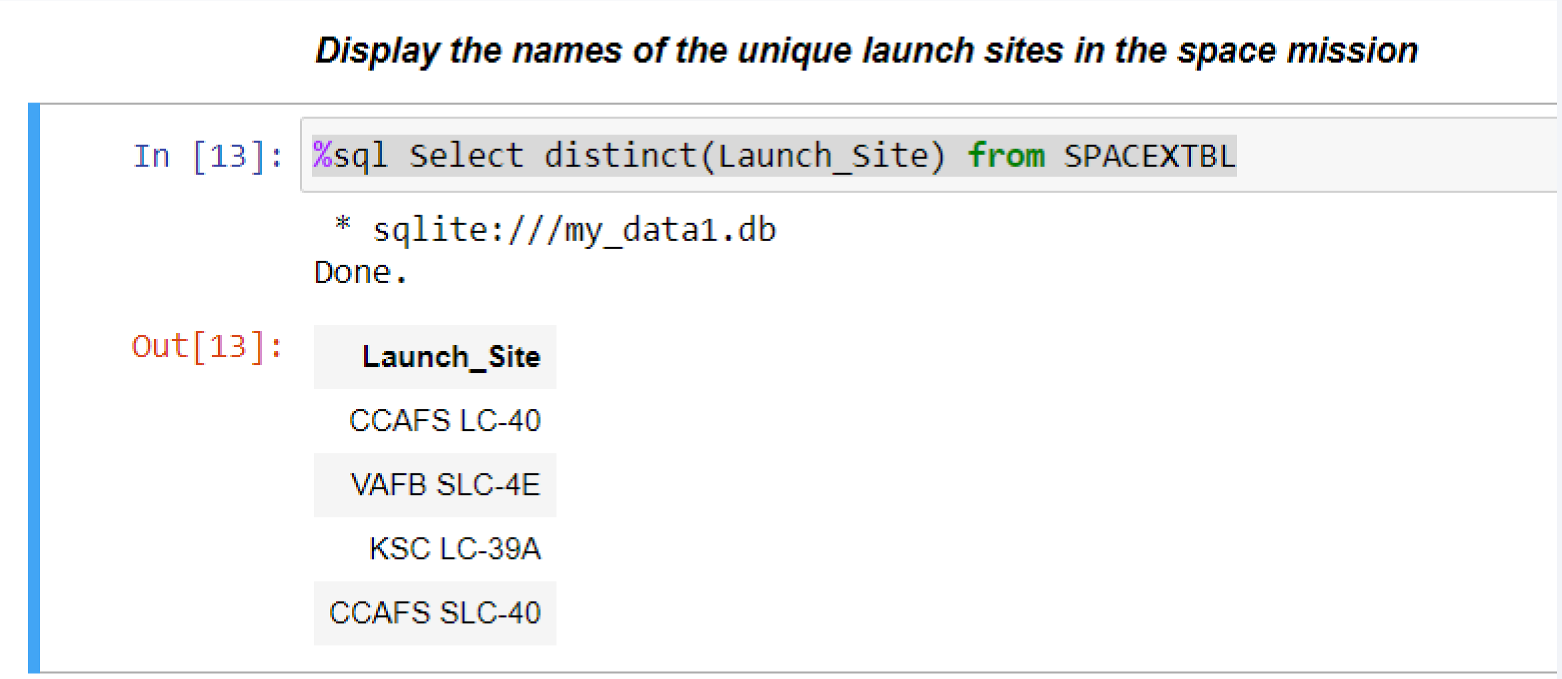
1. For higher payload either VLEO or ISS has been chosen.
2. For lower payload SSO seems to be good orbit, if there is a requirement of SSO orbit.

Launch Success Yearly Trend



1. As the time passes the learning improves and success rate also increases.

All Launch Site Names



Display the names of the unique launch sites in the space mission

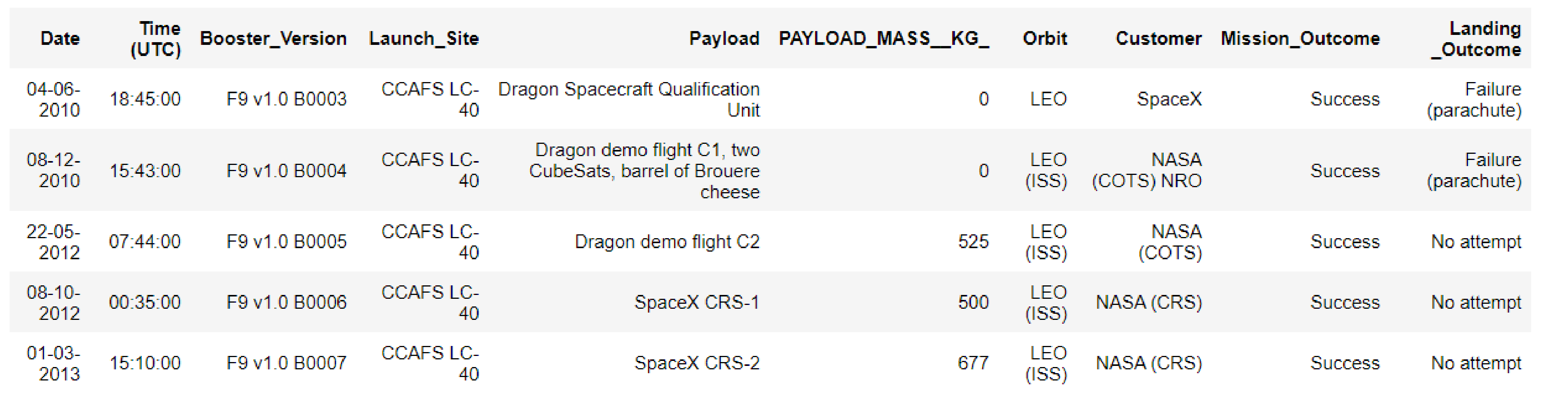
```
In [13]: %sql Select distinct(Launch_Site) from SPACEXTBL
```

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

Out[13]:

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

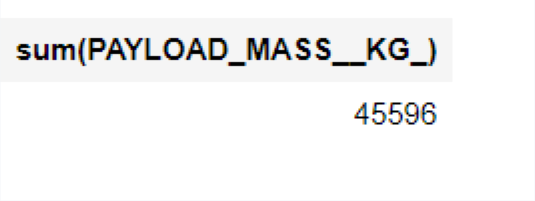
Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

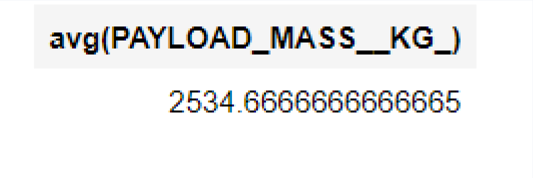
- Calculate the total payload carried by boosters from NASA



sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

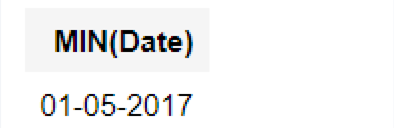
- Calculate the average payload mass carried by booster version F9 v1.1



avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

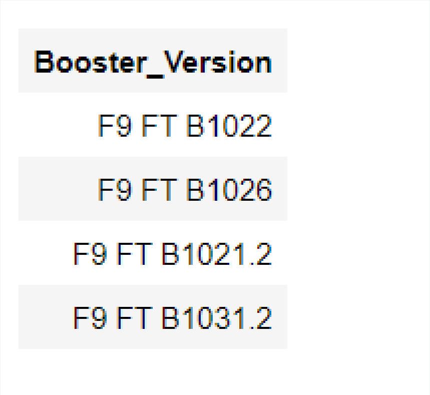
- Find the dates of the first successful landing outcome on ground pad



MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

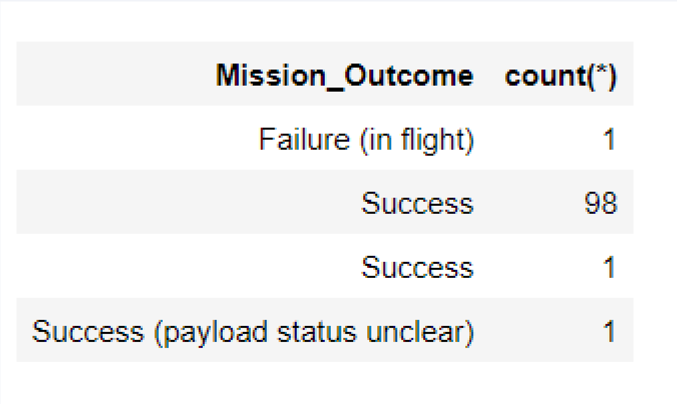
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



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

Total Number of Successful and Failure Mission Outcomes

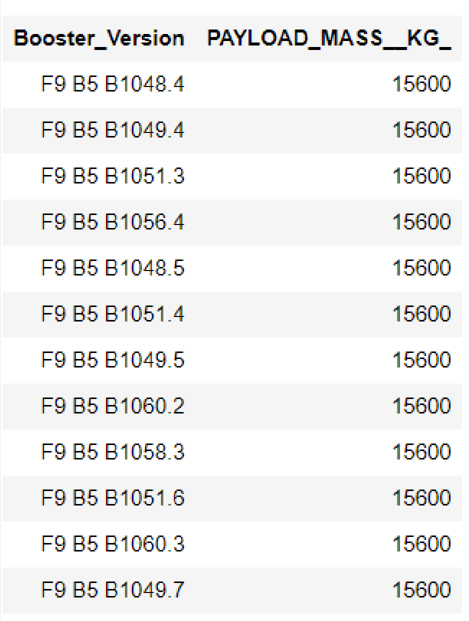
- Calculate the total number of successful and failure mission outcomes



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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

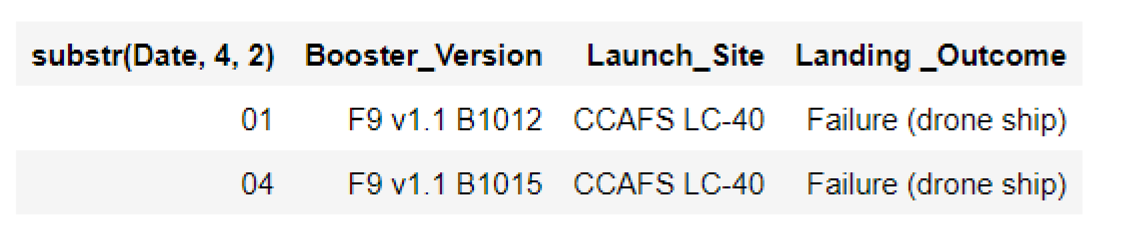


Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

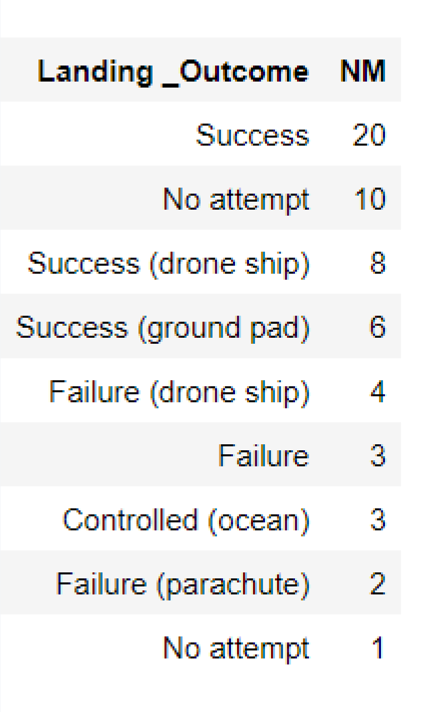
- Present your query result with a short explanation here



substr(Date, 4, 2)	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	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

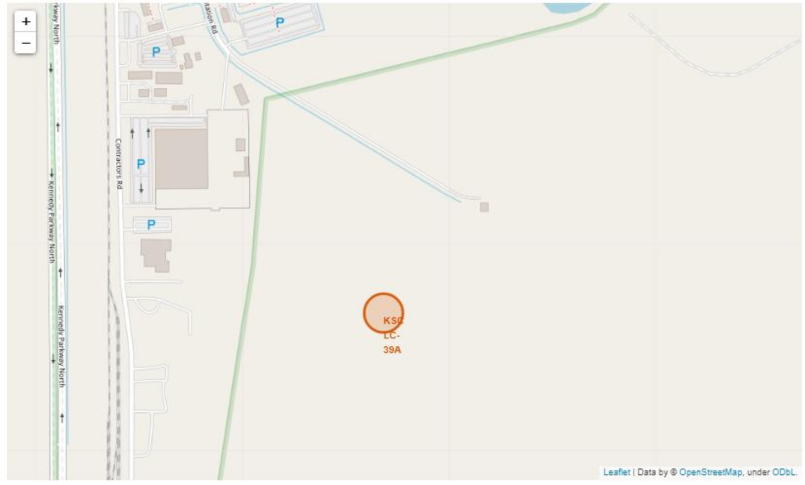


Landing _Outcome	NM
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

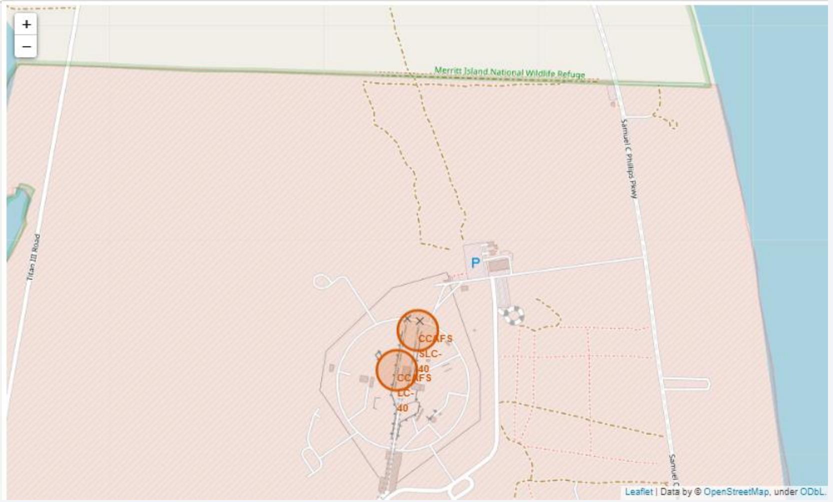
Section 3

Launch Sites Proximities Analysis

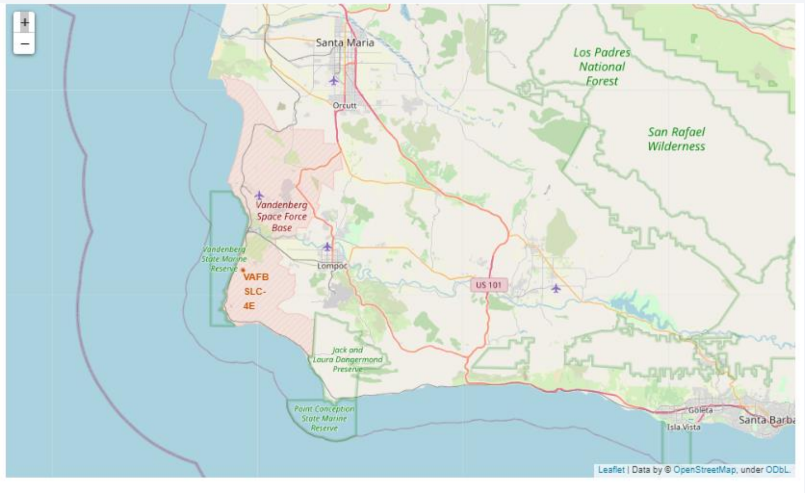
Launch Site Locations



KSC LC-39


CCAFS
SLC-
40
CCAFS
LC-
40
Merritt Island National Wildlife Refuge
Titan III Road
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

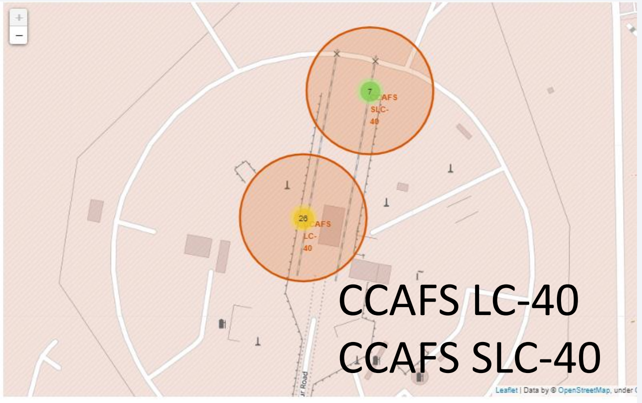
CCAFS LC-40
CCAFS SLC-40



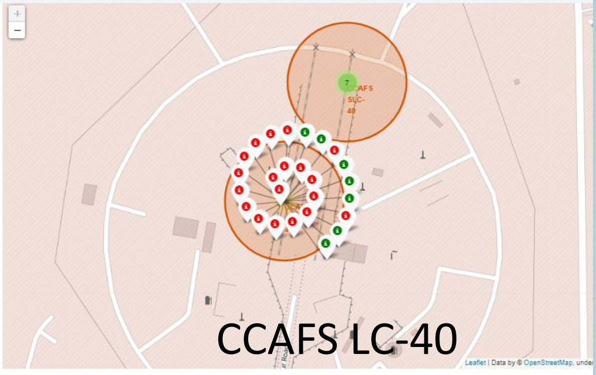
VAFB SLC-4E

1. CCAFS LC-40 and CCAFS SLC-40 are very close to each other.
2. VAFB SLC -4E is far away from other launching sites.
3. All launching sites except from KSC LC-39 are close to sea.

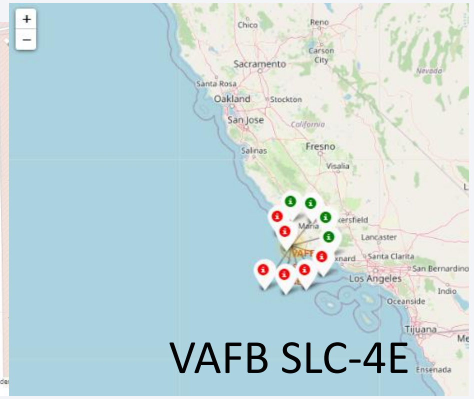
Launch Outcome



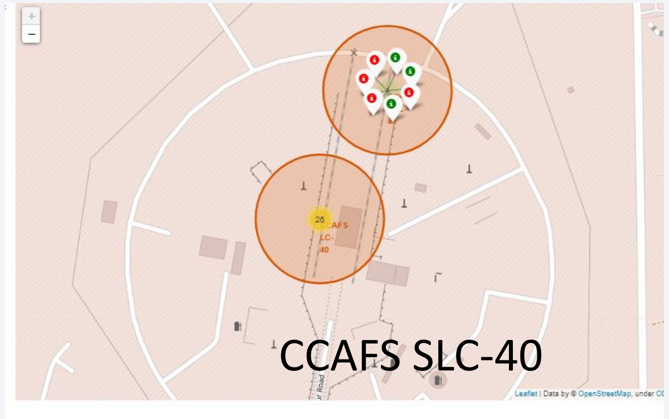
CCAFS LC-40
CCAFS SLC-40



CCAFS LC-40

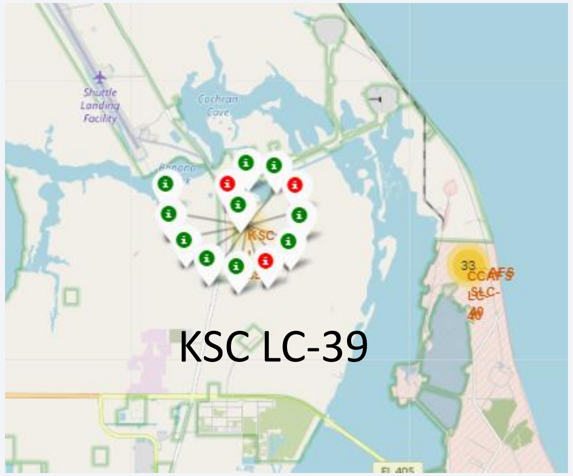


VAFB SLC-4E



CCAFS SLC-40

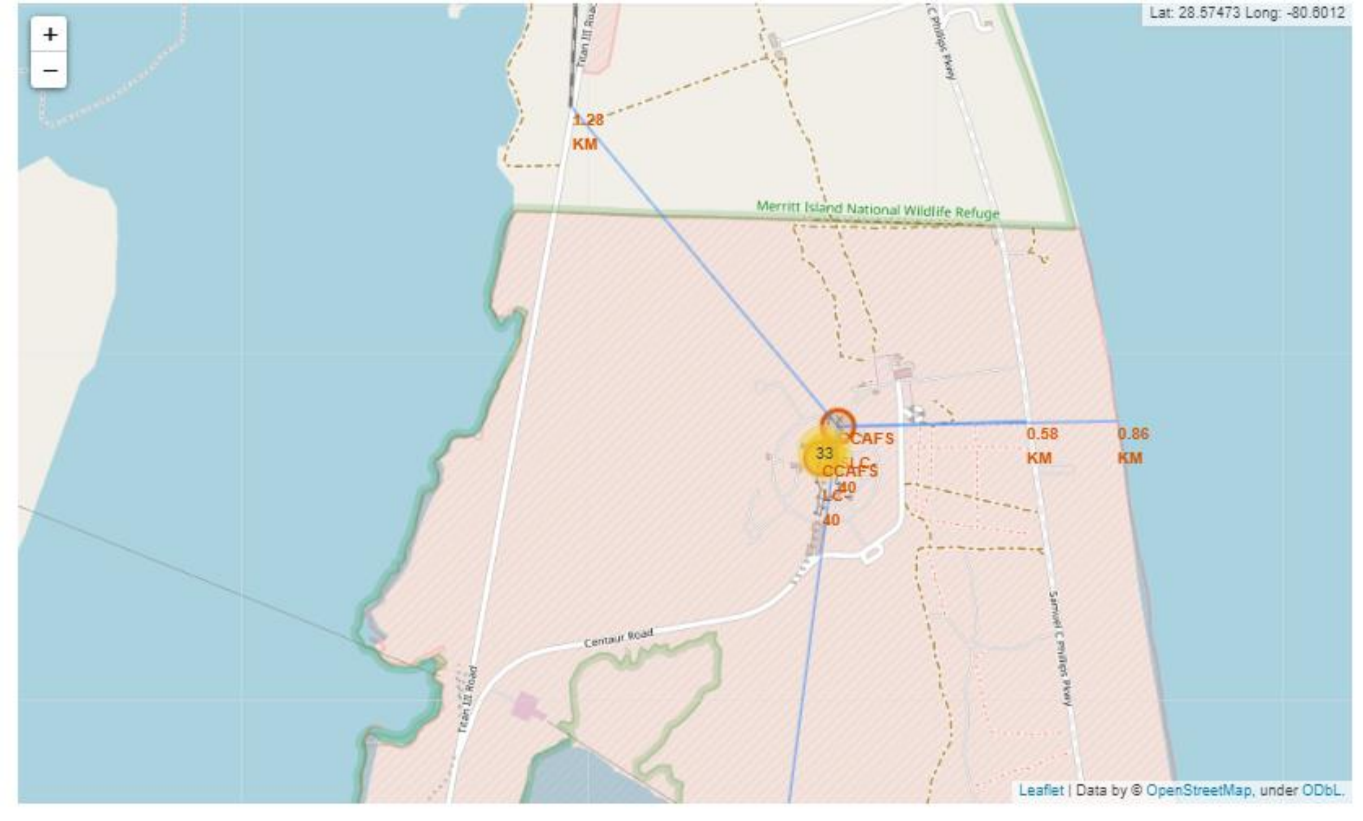
Launch site	Success	Failure	Total	Success rate
CCAFS LC-40	7	19	26	27%
CCAFS SLC-40	3	4	7	42.8%
KSC LC-39	10	3	13	77%
VAFB SLC-4E	4	6	10	40%



KSC LC-39

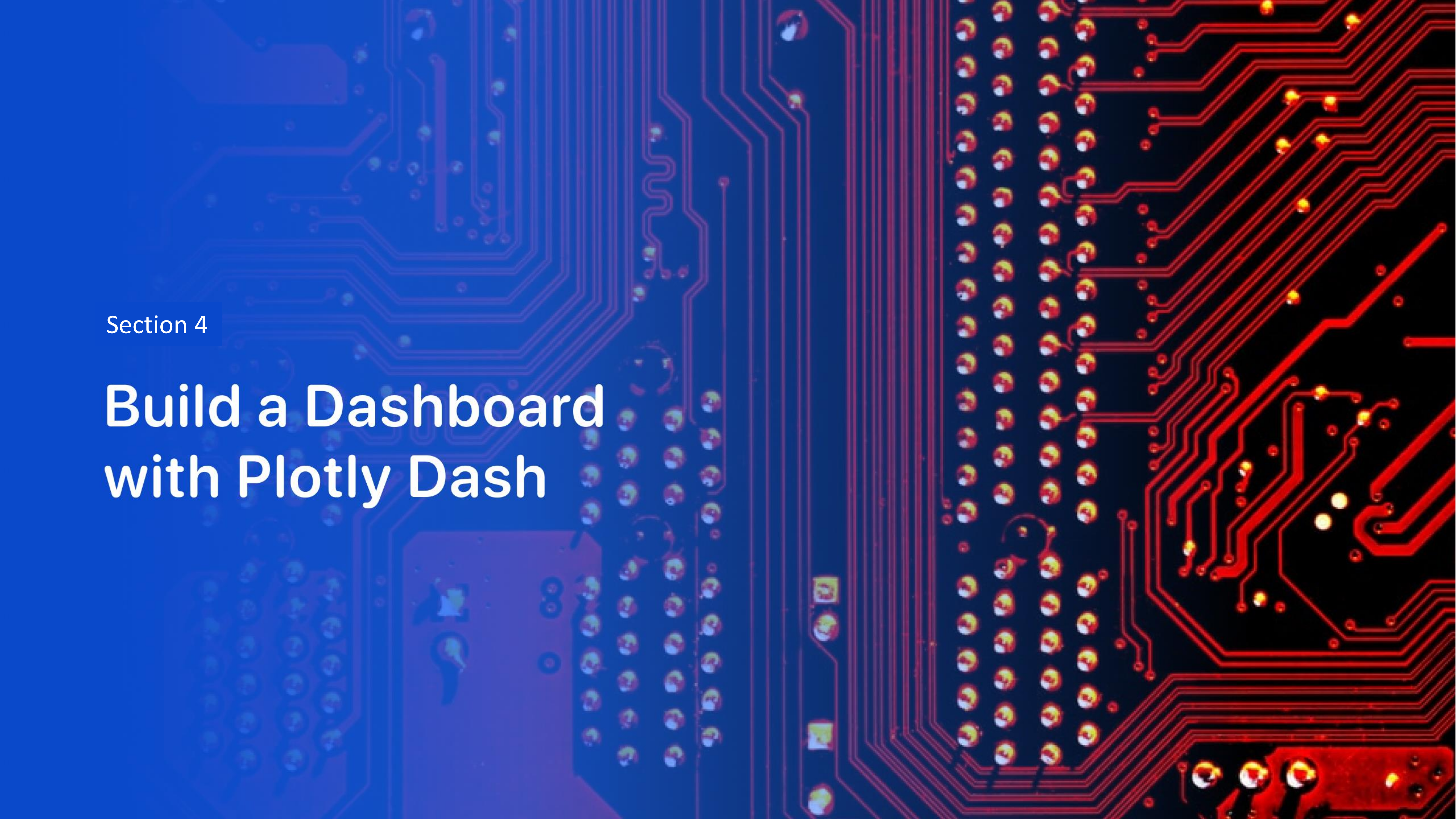
KSC LC-39 has highest success rate of 77%.

Launch site and its proximities to transport locations

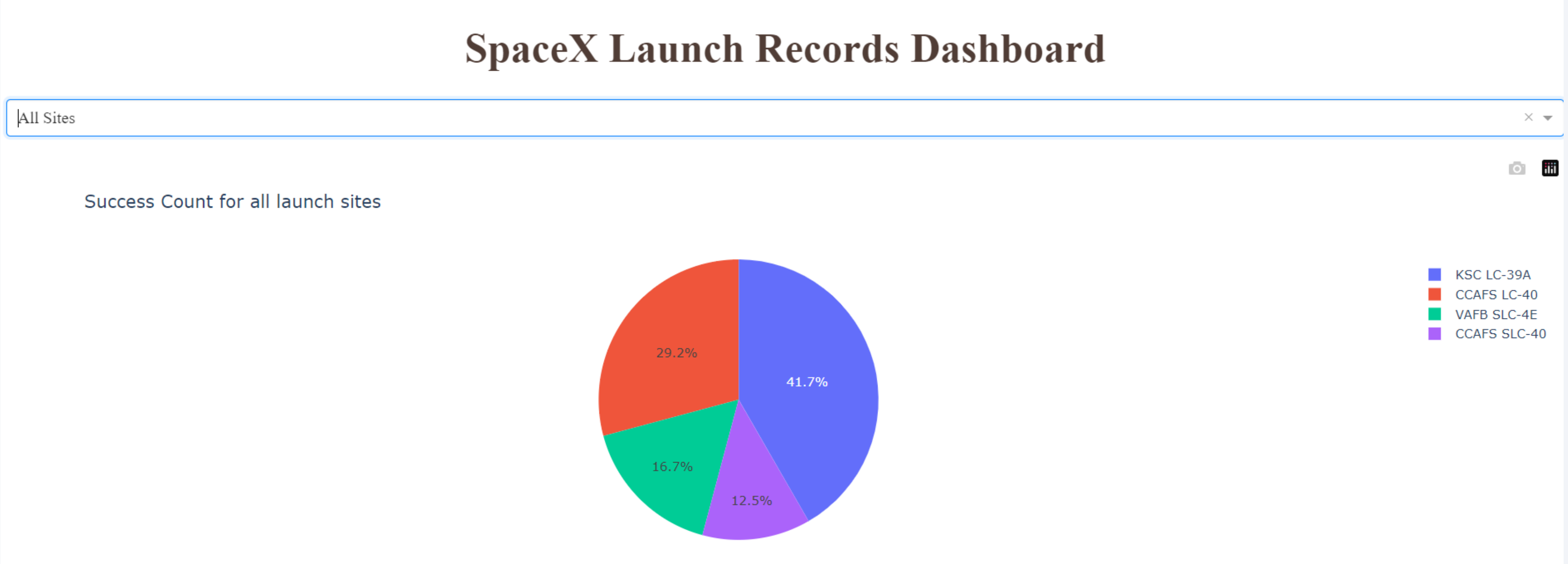


Section 4

Build a Dashboard with Plotly Dash



Launch success count for all sites

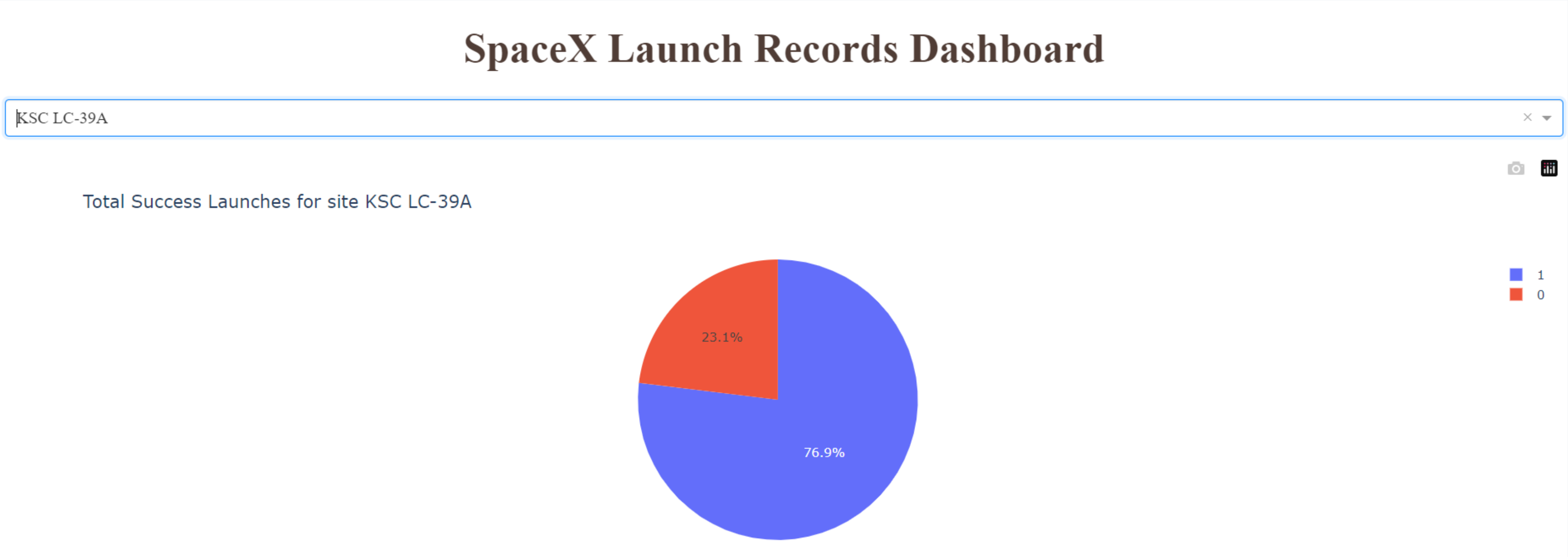


SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites

Launch fro site KSC LC-39A

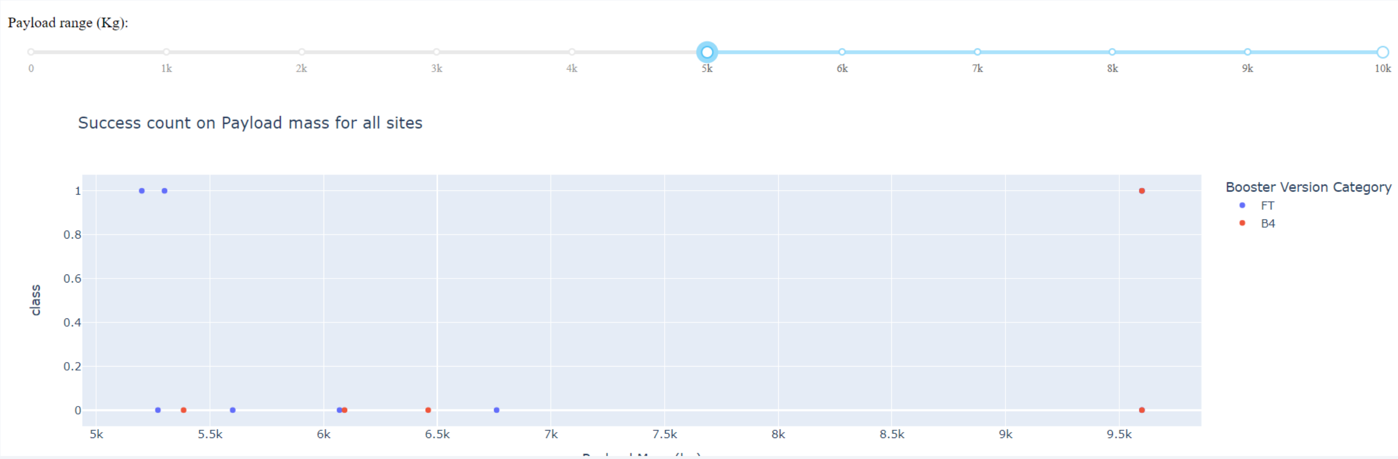


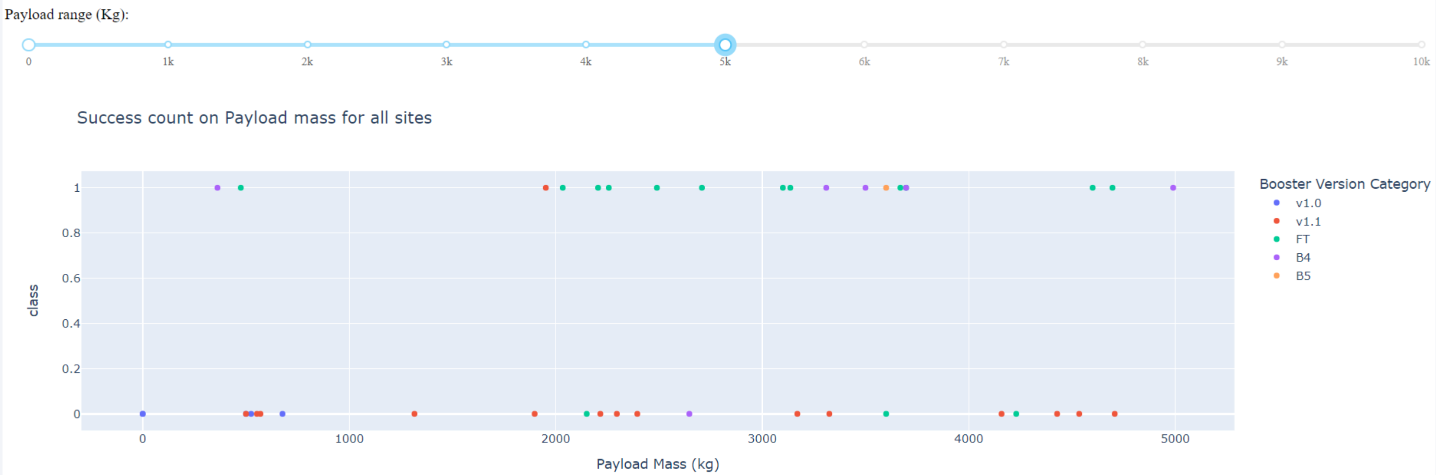
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload and launch outcome

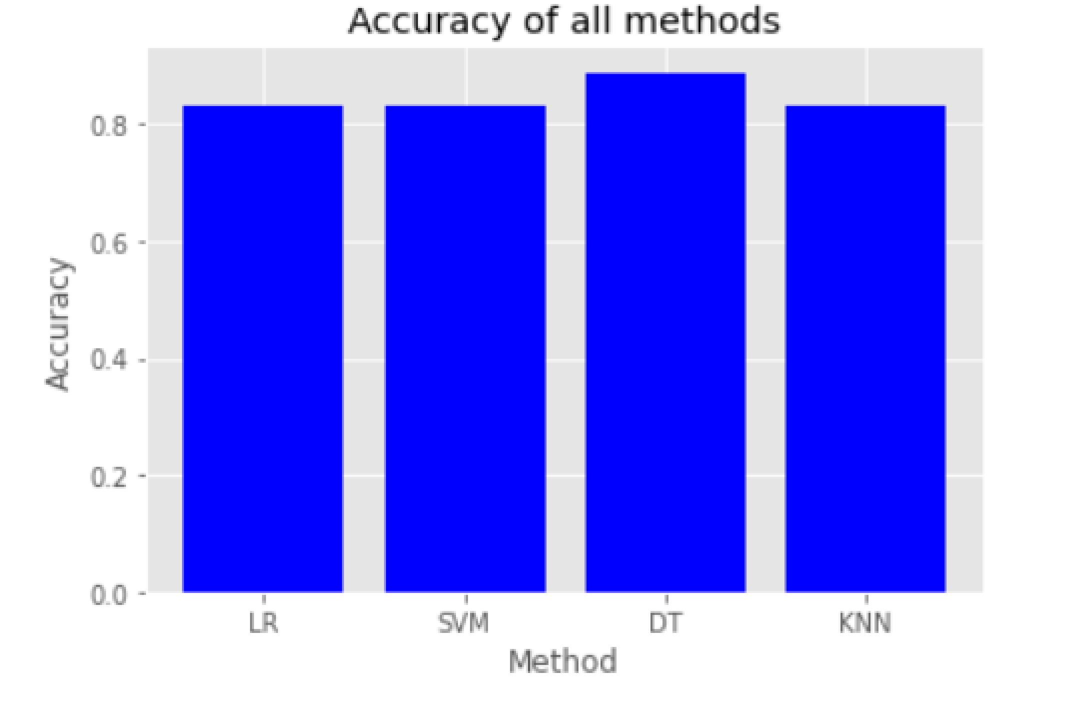




Section 5

Predictive Analysis (Classification)

Classification Accuracy

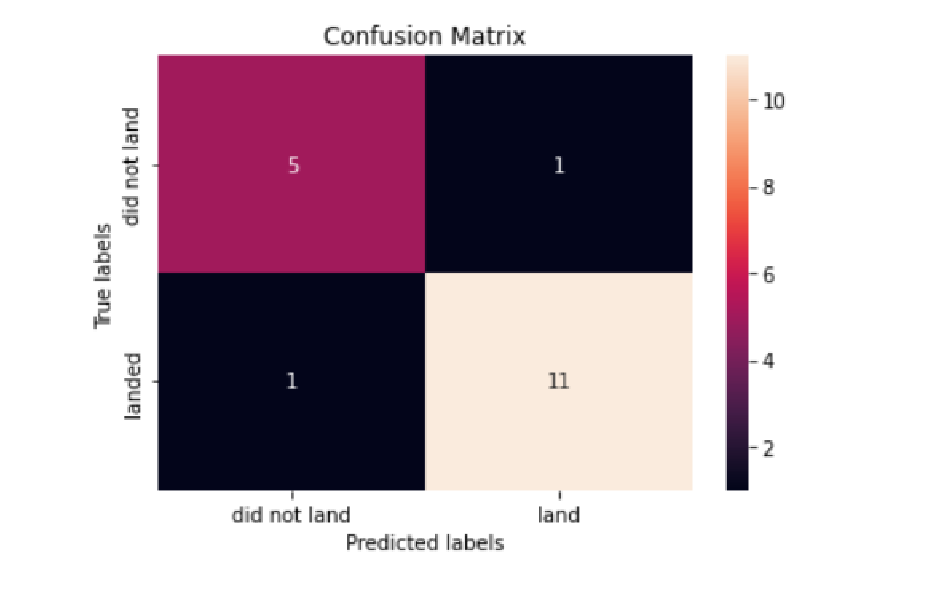


- Decision tree model has the highest classification accuracy with 88.9%.

Confusion Matrix

- Confusion matrix of the decision tree model

- Out of 6 unsuccessful landing, it predicts 5 outcomes correctly; while out of 12 successful landing, it correctly predicts 11 outcomes.



Conclusions

1. As the time passes the learning improves and success rate also increases.
2. In this case, decision tree model has the highest classification accuracy with 88.9%.
3. With this analysis one can predict the outcome of launch and invest/plan accordingly.
4. There is no failure in landing, when landing was done at ground pad. Although some failure is there for landing on drone ship. To have higher chances of success one may choose ground pad landing.
5. KSC LC-39 has highest success rate but after 80th flight number all launch sites are giving good outcomes. So currently, launch site may not be determining factor.

Thank you!