IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prasanna D Hegde

01-01-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization and SQL
  - Visualizing through interactive folium map
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

  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.

- Problems you want to find answers
  - What are the factors that influence the successful landing ?
  - Does successful landing depend on any particular launch site or orbit?
  - Does the mass of the rocket play a role?

Section 1

Methodology

Methodology

Executive Summary

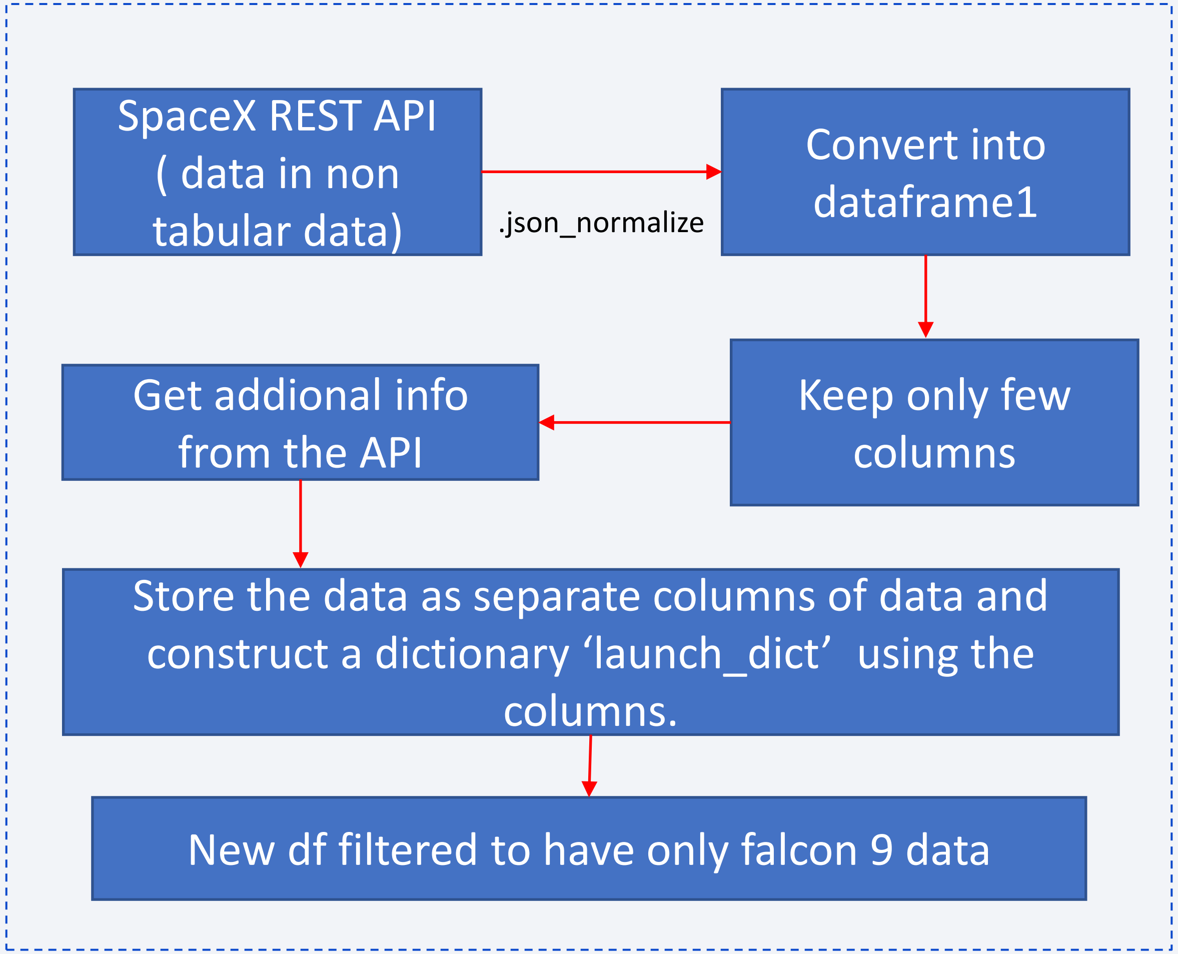
- Data collection methodology:
    - The Data was gathered using SpaceX REST API and from HTML web scrapping of a wiki page.
- Perform data wrangling
    - The json data collected is converted to a data frame. The data is filtered for just the ‘falcon 9’ data. Some data cleaning like replacing the null values with mean values is done.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The data was normalized, split into train and test data and 4 different models were built with the best parameters found from GridSerachCV.

Data Collection

- Datasets were generated from the spaceX REST API and HTML web scraping of a wiki page :
- API: https://api.spacexdata.com/v4/launches/past
- Wiki: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The following slides explains these steps in more detail.

Data Collection – SpaceX API

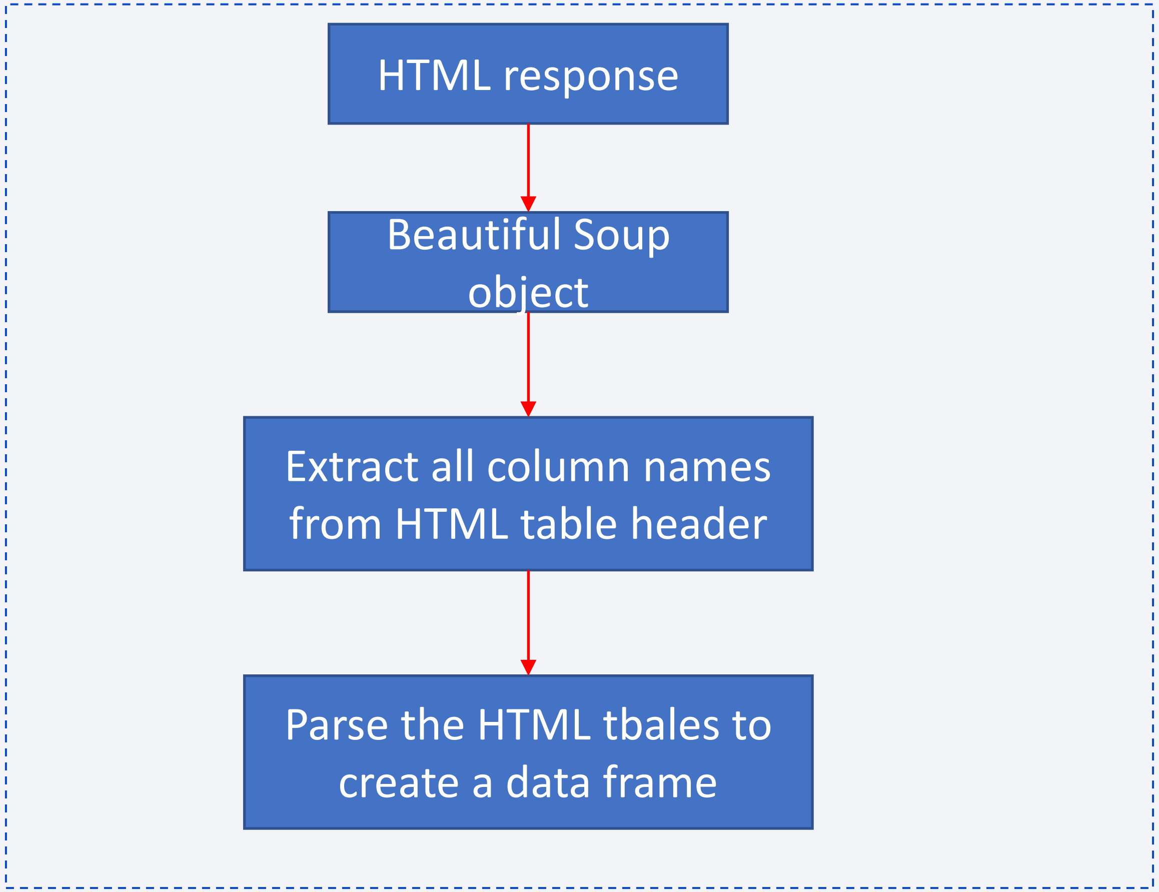
- Get the data from SpaceX API using requests.get(url) and store it as response.
- Encode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
- Use helper functions to get additional data (example: Booster name from the rocket column, from payload get mass of payload and the orbit)


SpaceX REST API (data in non tabular data)
.json_normalize
Convert into dataframe1
Keep only few columns
Get addional info from the API
Store the data as separate columns of data and construct a dictionary 'launch_dict' using the columns.
New df filtered to have only falcon 9 data

Api notebook GitHub

Data Collection - Scraping

- Get data from wiki page of falcon 9 using get_requests and store it as beautiful soup object.
- For step 3, the following code is used : html_tables= soup.find_all('table')
- Then columns are extracted by iterating through the <th> elements.


HTML response

Beautiful Soup object

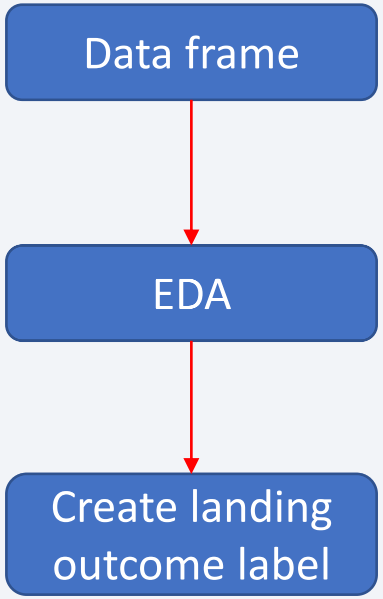
Extract all column names from HTML table header

Parse the HTML tbales to create a data frame

Scraping notebook GitHub

Data Wrangling

- EDA
  - Identify missing values in each column
  - Identify data types of columns (numerical/categorical)
- Calculations on data
  - Count of launches per launchsite
  - Count of each orbit,
  - Count of mission outcome of the orbits
- Create landing outcome label from outcome column, so as to create a classification problem



Data Wrangling GitHub

EDA with Data Visualization

- Payload mass Vs Flight number scatter chart:
  - shows that the success has been improving compared to earlier.
- Launch Site Vs Flight Number scatter chart:
  - To see the success over on different launch sites
- Launch Site Vs Pay Load Mass scatter chart:
  - To see which launch site is suitable for a given payload.
- Bar chart of Success rate by Orbits:
  - To check if there are any relationship between success rate and orbit type.
- Orbit Vs Flight Number scatter chart:
  - To check if there are any relationship between flight number and orbit type.
- Orbit vs Payload mass scatter chart:
  - To check if there are any relationship between flight number and orbit type
- Line chart of Success over the years

EDA Data Visualization GitHub

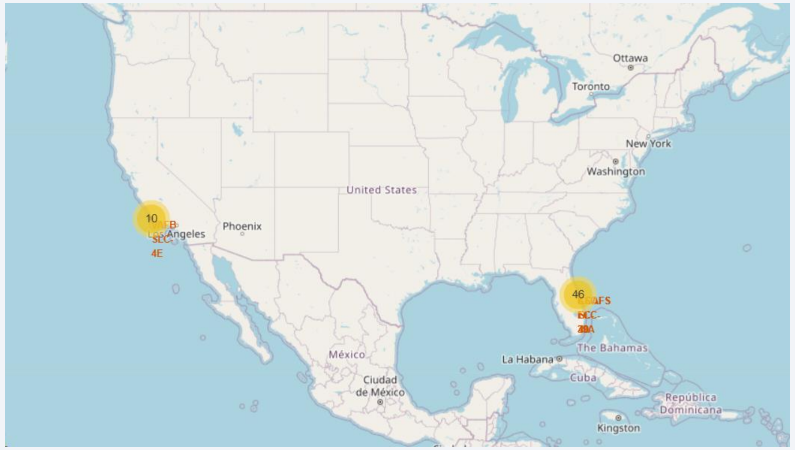
EDA with SQL

- Distinct launch sites [*SELECT DISTINCT*]
- Display where launch sites begin with selected string *["Launch_Site" like 'CCA%' limit 5*]
- Total payload mass carried by boosters launched by NASA (CRS) [*SELECT SUM*]
- Avg. payload mass carried by booster version F9 v1.1 [*where "Booster_Version"='F9 v1.1'*]
- SELECT "Booster_Version" from TABLE where "Landing_Outcome"='Success (drone ship)' AND 4000<"PAYLOAD_MASS__KG_" < 6000
- Date when first succesful landing outcome in ground pad was achieved [*SELECT MIN(Date) , where "Landing_Outcome"='Success (ground pad)']*
- Total number of successful and failure missions (*SELECT "Mission_Outcome", COUNT(*) as total_count FROM SPACEXTABLE **GROUP BY** "Mission_Outcome";*)
- names of the booster_versions which have carried the maximum payload mass (sub query *WHERE "PAYLOAD_MASS__KG_" = [SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);*]
- Failures for drone ship in the months of 2015 [*SELECT SUBSTR(Date, 6, 2) AS Month, booster_version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5) = '2015';*]
- Rank the count of landing outcomes between the dates [*SELECT "Landing_Outcome", COUNT(*) as total_count FROM SPACEXTABLE Where Date BETWEEN '2010-06-04' AND '2017-03-20' Group by "Landing_Outcome" order by total_count DESC ;*]

EDA with SQL GitHub

Build an Interactive Map with Folium

- Map objects added:
    - Circle – to display a small highlighted area around the launch site coordinates
    - Marker clusters – to display success and failure counts around each launch site.
    - polyline – to display a line to the selected nearest coast, railway, city center etc. from a given launch site. This helps to visualize if the launch sites have close proximities to railways, highways and coastline

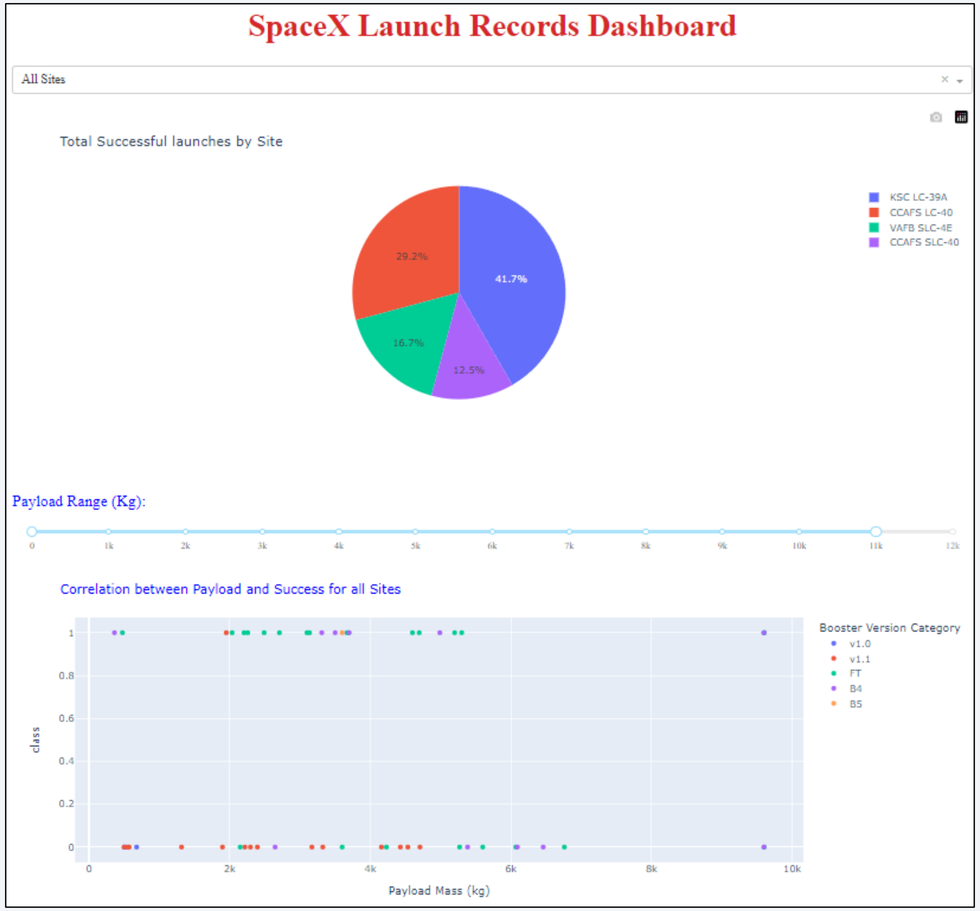


Interactive Map GitHub

Build a Dashboard with Plotly Dash

Dashboard components

- **Dropdown** to select – 'All launch sites' or a 'single launch site'
- Based on the selection above a **pie chart** displays total successful launches for the launch sites
- A **slider** to select the payload mass range (default can be set to max and min payload from the data)
- A **scatter chart** shows correlation between payload and launch success for the selected range of payload.



SpaceX Launch Records Dashboard

All Sites

Total Successful launches by Site

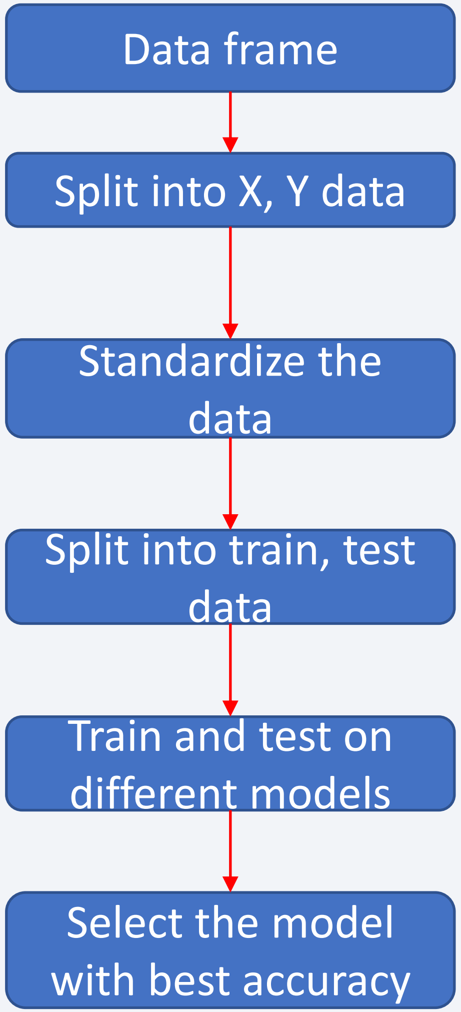
Payload Range (Kg):

Correlation between Payload and Success for all Sites

Dashboard GitHub

Predictive Analysis (Classification)

- Standardize the split X and Y data using sklearn StandardScaler(). This removes the mean and scales the data to unit variance.
- The X and Y data are split using train_test_split with test size of 0.2.
- The models built are: logistic regression, SVM, decision tree, KNN
- All are the models are built with the best parameter search from GridSearchCV
- All the models had the same test accuracy of 0.833. Decision tree had the best accuracy of 0.889 on the training data.



Data frame

Split into X, Y data

Standardize the data

Split into train, test data

Train and test on different models

Select the model with best accuracy

Predictive Analysis GitHub

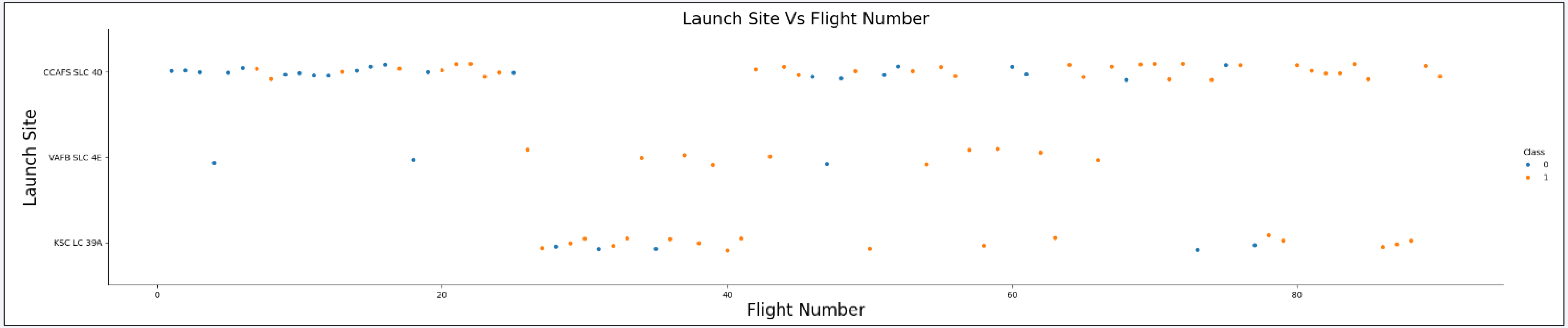
Results

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

Section 2

Insights drawn from EDA

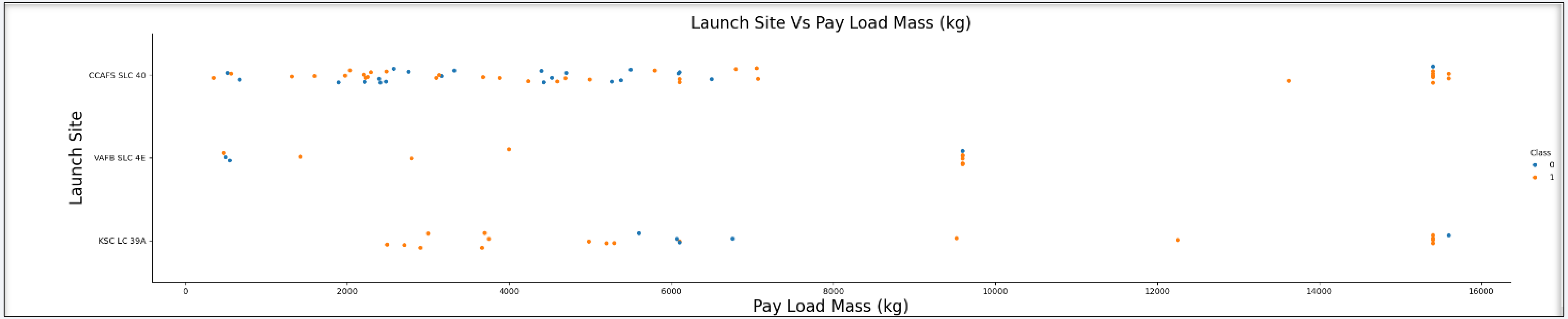
Flight Number vs. Launch Site



Observations

- The launch sites KSC LC-39A and VAFB SLC 4E have a better success rate, but the total number of launches compared to CCAFS LC-40 is less.
- CCAFS LC-40 had a lot of failures initially, but after sometime, it looks like the success rate has improved

Payload vs. Launch Site

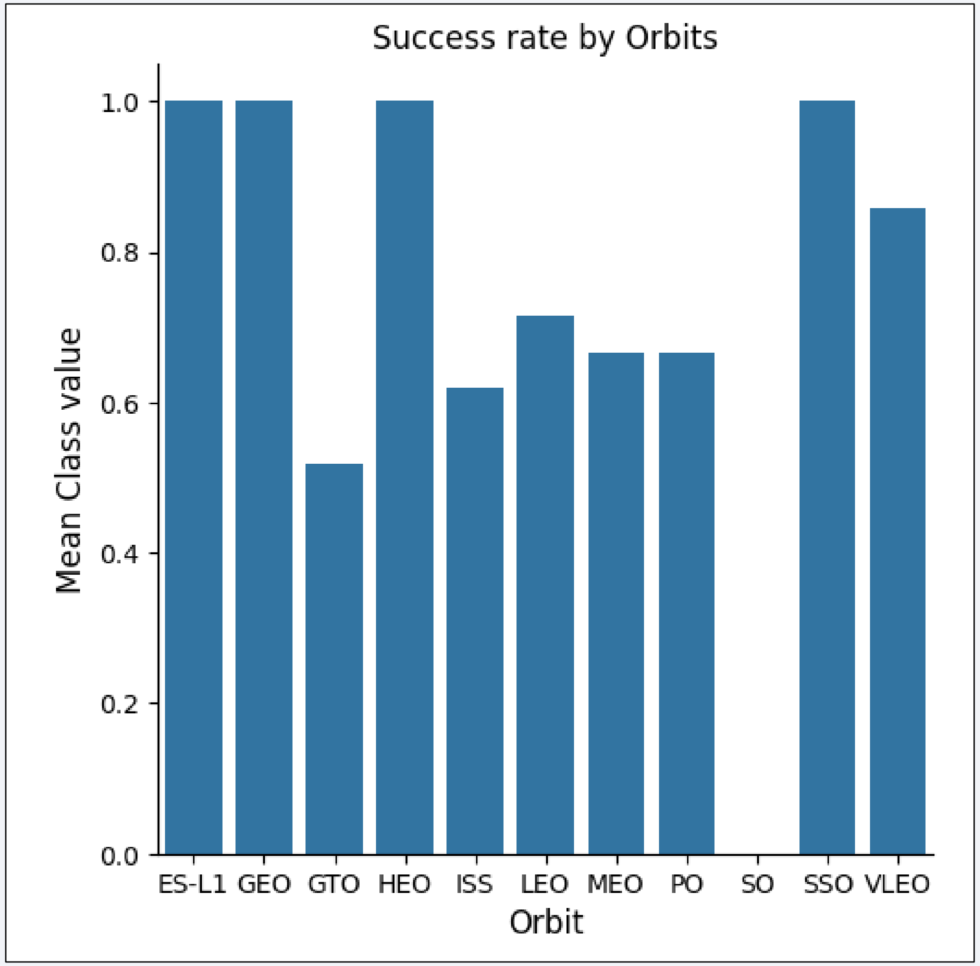


Observations

- KSC LC-39A is a good option for payload between 2000 and 5000 kg.
- VAFB SLC 4E is an alternate option for payload between 1500 and 4000 kg.
- Overall, there are fewer heavy payload (>10000 kg) launches. CCAFS LC-40 and KSC LC-39A are used here.

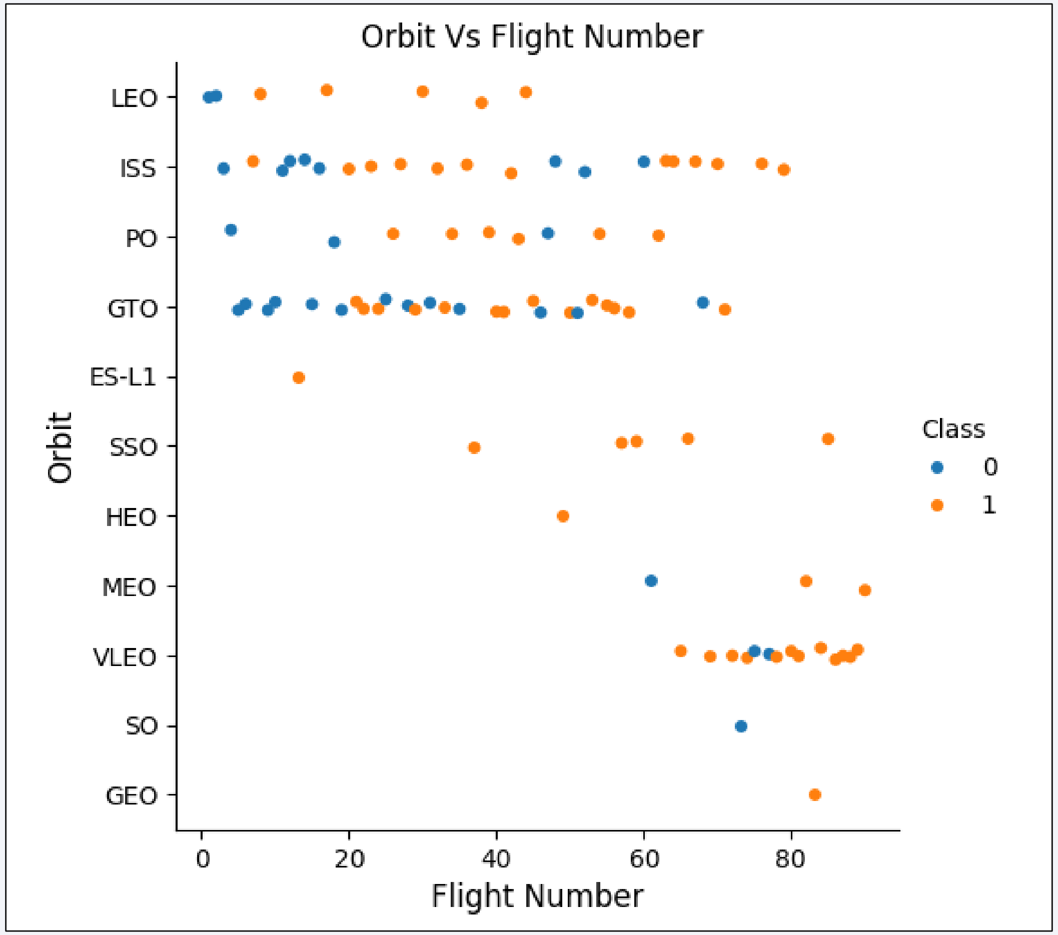
Success Rate vs. Orbit Type

- Orbits GEO, ES-L1, SSO, HEO have high success rate
- orbit GTO has the least success rate followed by ISS, orbit SO had just 1 launch with resulted in failure



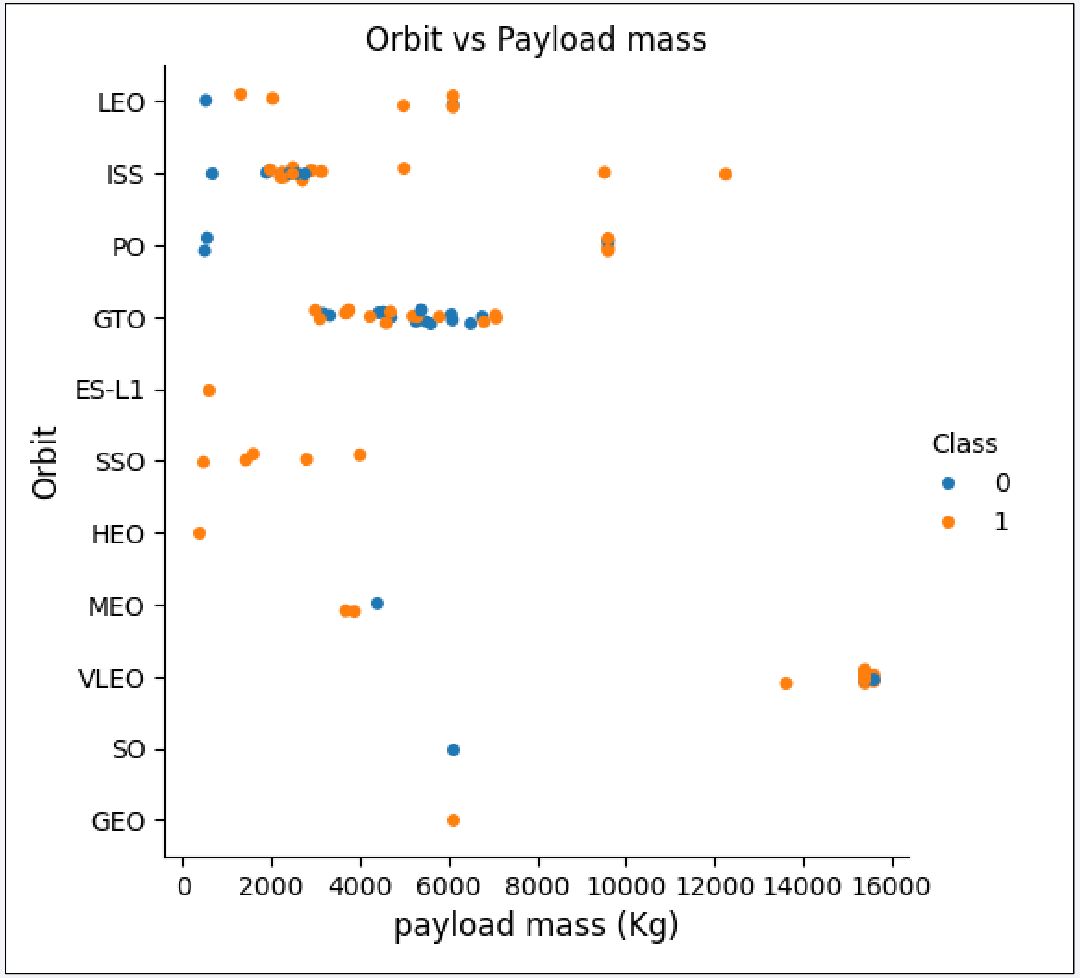
Flight Number vs. Orbit Type

- The success seems to be improving over time for all orbits
- SSO, VLEO and ISS followed by LEO seem to be interesting orbit options



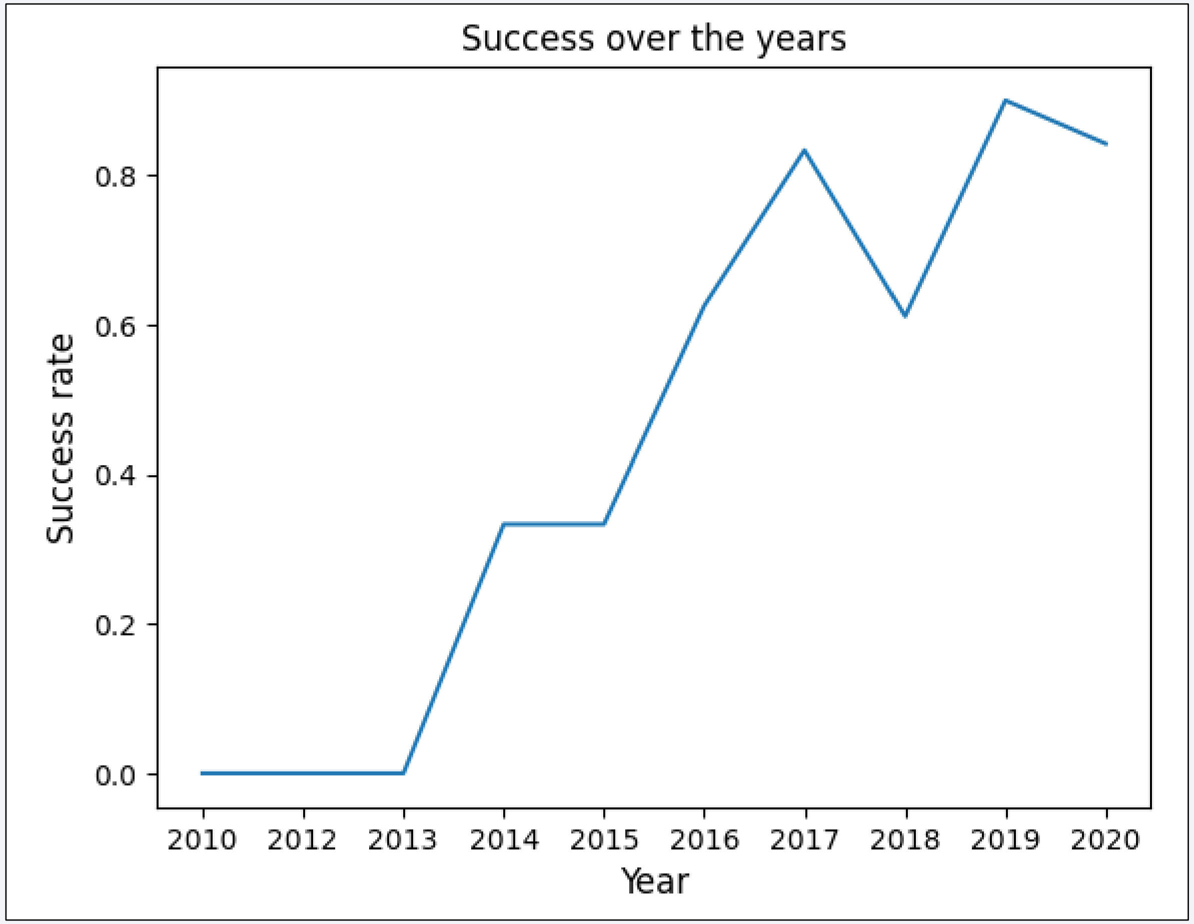
Payload vs. Orbit Type

- Orbit SSO seems to be suitable for payload mass < 5000 kg. For highpayload >= 10000kg, orbits ISS, PO and VLEO seem good
- For payload mass between 5000 and 10000 kg, orbits ISS and LEO could be tried, though there are few data available here



Launch Success Yearly Trend

- The success has been improving since 2013.
- In 2018, the success rate dropped and it increased again in 2019
- The years between 2015-17 seems to be the time where some significant improvements were made.



All Launch Site Names

- Query used:
  - SELECT DISTINCT "Launch_Site" from SPACEXTABLE
  - Selects unique values from the ‘Launch_Site’ column.
- The unique launch sites are:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Query used:
  - SELECT * FROM SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;
  - Selects all data from the table where the ‘Launch_Site’ column has the keyword ‘CCA’, but show only 5 values.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Query used:
  - SELECT SUM(PAYLOAD_MASS__KG_) as total_payload_mass_NASA_CRS from SPACEXTABLE where "Customer"='NASA (CRS)'
  - Selects sum of payload column with the filter of customer= 'NASA (CRS)'
  - The result is 45596 kg.

Average Payload Mass by F9 v1.1

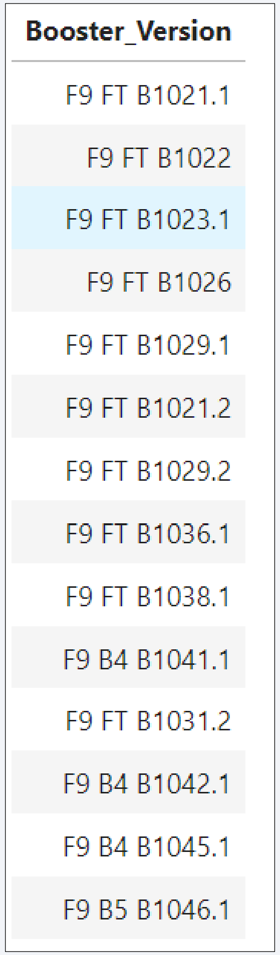
- Calculate the average payload mass carried by booster version F9 v1.1
- Query used:
  - SELECT AVG(PAYLOAD_MASS__KG_) as avg_payload_mass_F9v11 from SPACEXTABLE where "Booster_Version"='F9 v1.1’
  - Selects average of payload column with the filter of the given booster version.
  - The result is 2928.4 kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Query used:
  - SELECT MIN(Date) as first_successful_landing_groundpad from SPACEXTABLE where "Landing_Outcome"='Success (ground pad)’
  - Selects minimum of date column with the filter on landing outcome column for Success (ground pad).
  - The first successful landing ground pad was on 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

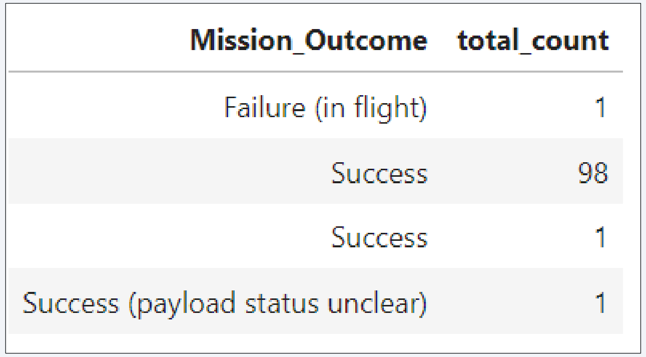
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query used:
  - SELECT "Booster_Version" from SPACEXTABLE where "Landing_Outcome"='Success (drone ship)' AND 4000<"PAYLOAD_MASS__KG_" < 6000
  - Applies filter on payload column for the given weight range, success (drone ship) on the landing outcome column. Then, select the filtered data and shows the booster versions
  - The image shows the result



Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

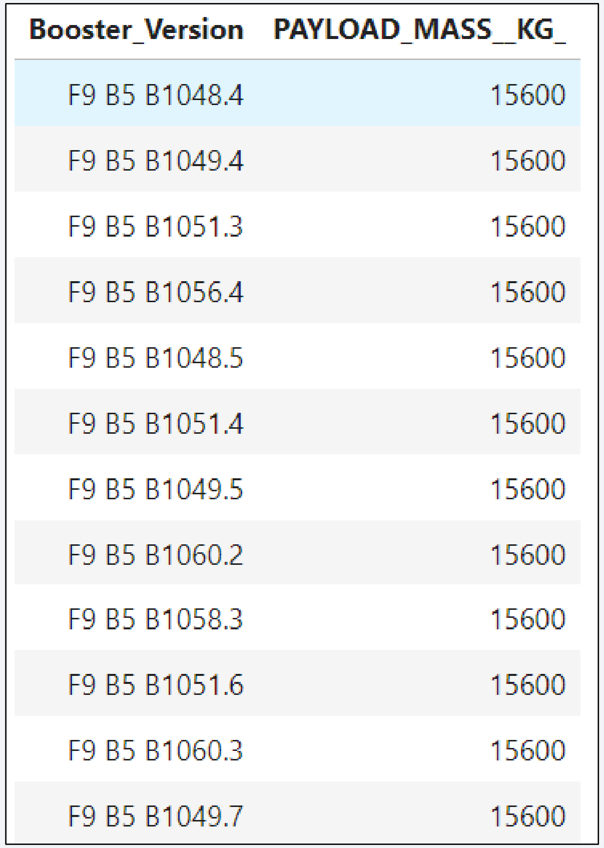
- Calculate the total number of successful and failure mission outcomes
- Query used:
  - SELECT "Mission_Outcome", COUNT(*) as total_count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
  - Gives the count for unique values of mission_outcome column as total count.
  - The image shows the result



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

Boosters Carried Maximum Payload

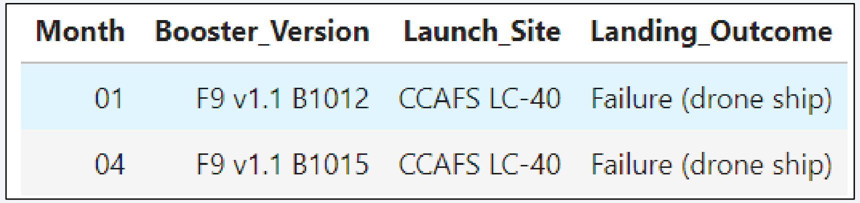
- List the names of the booster which have carried the maximum payload mass
- Query used:
  - SELECT booster_version, "PAYLOAD_MASS__KG_" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
  - We select booster version column and We use a subquery to select max payload mass filter on the payload mass column.
  - The image shows the result



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

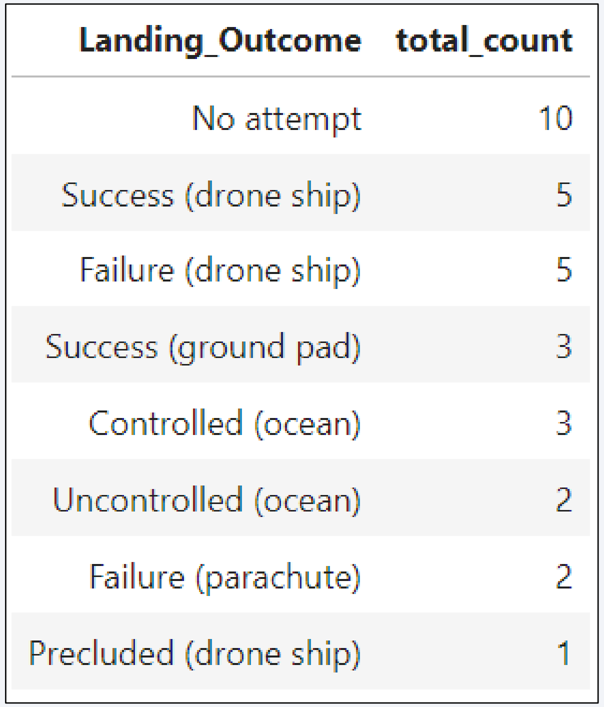
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query used:
  - SELECT SUBSTR(Date, 6, 2) AS Month, booster_version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5) = '2015';
  - SQLLite does not support monthnames. So we need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.
  - The image shows the result



Month	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
- Query used:
  - SELECT "Landing_Outcome", COUNT(*) as total_count FROM SPACEXTABLE Where Date BETWEEN '2010-06-04' AND '2017-03-20' Group by "Landing_Outcome" order by total_count DESC ;
  - It is interesting to see the 'no attempt' in landing.

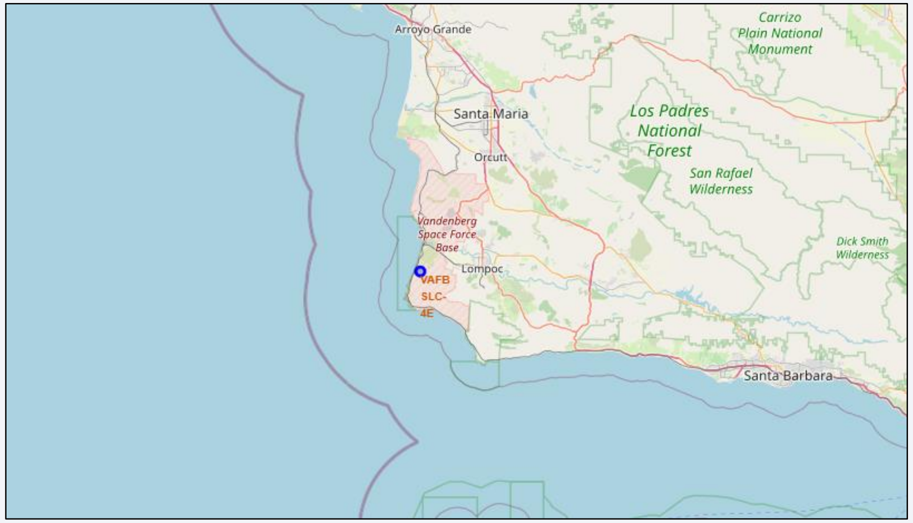


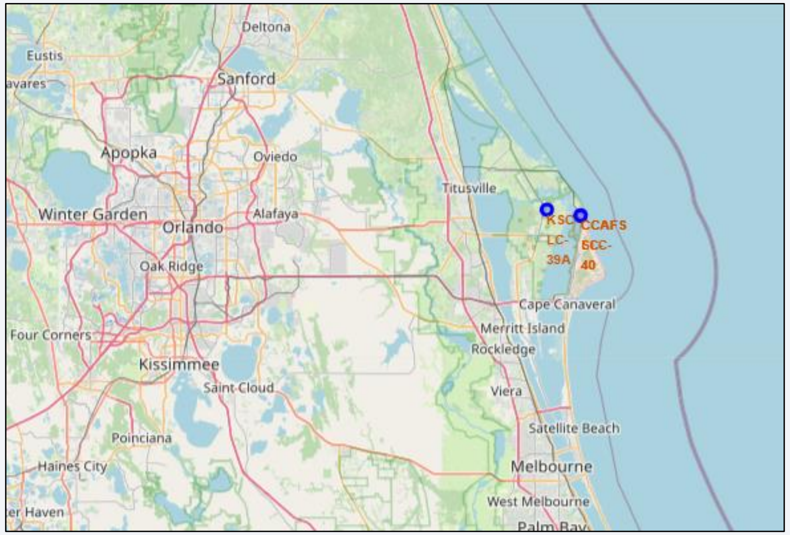
Landing_Outcome	total_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

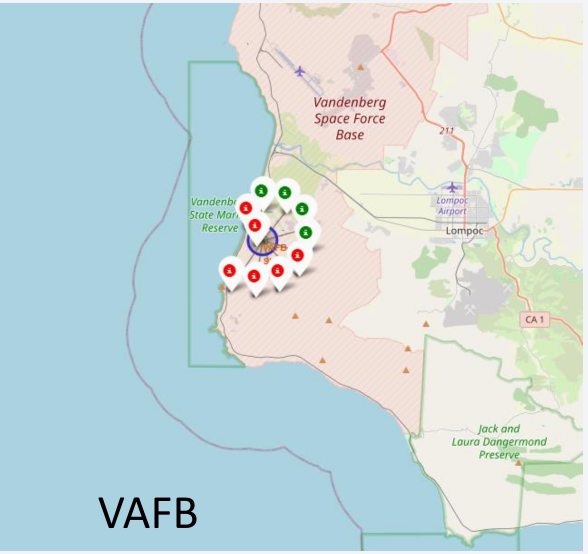
Launch Site locations





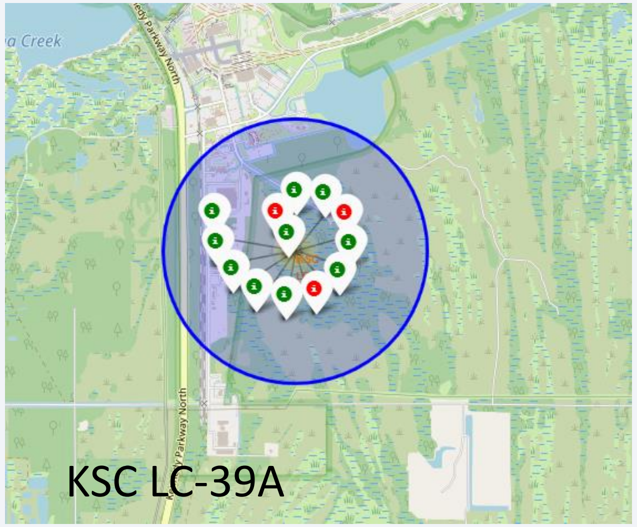
- All the launch sites seem to be located in the outskirts of the city and near to a ocean

Launch Sites with outcomes

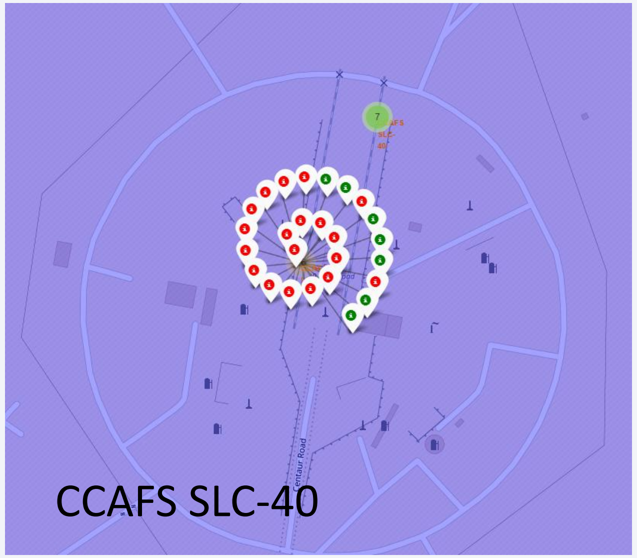


Vandenberg Space Force Base
Lompoc Airport
Lompoc
Jack and Laura Dangermond Preserve

VAFB



KSC LC-39A

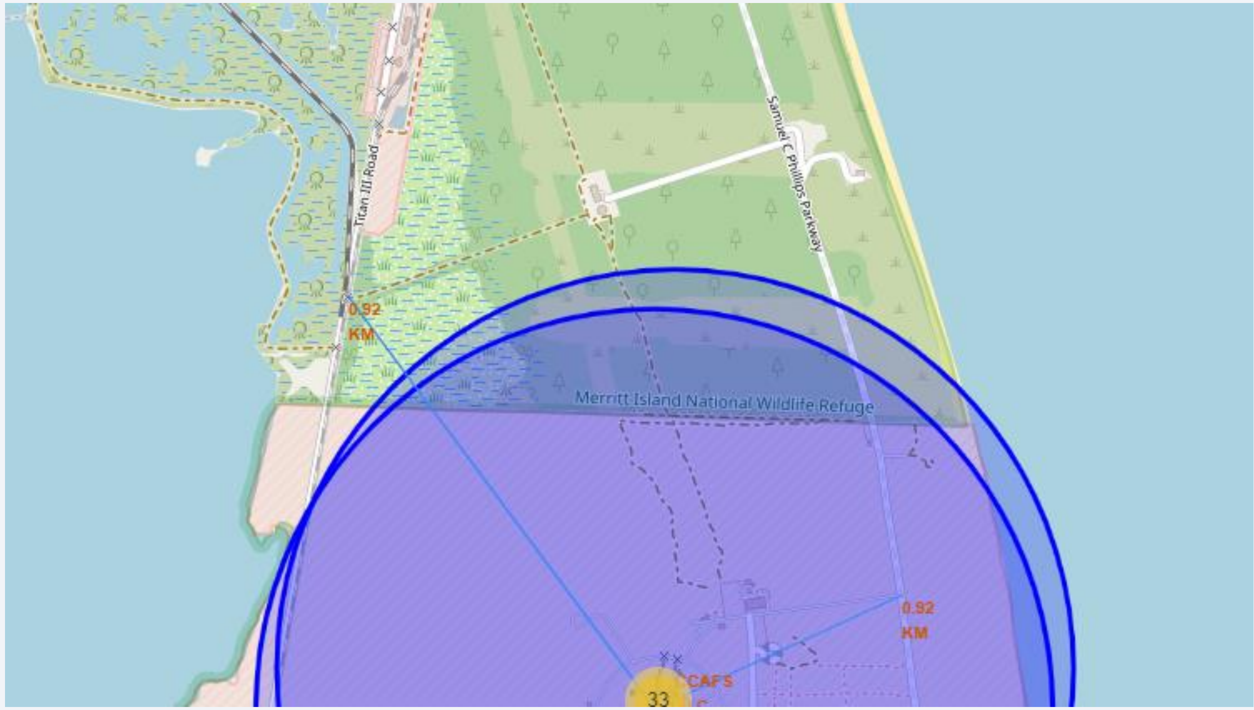


CCAFS SLC-40

- Green markers indicates success and red indicates failure
- KSC LC-39A seem to have more successful launches

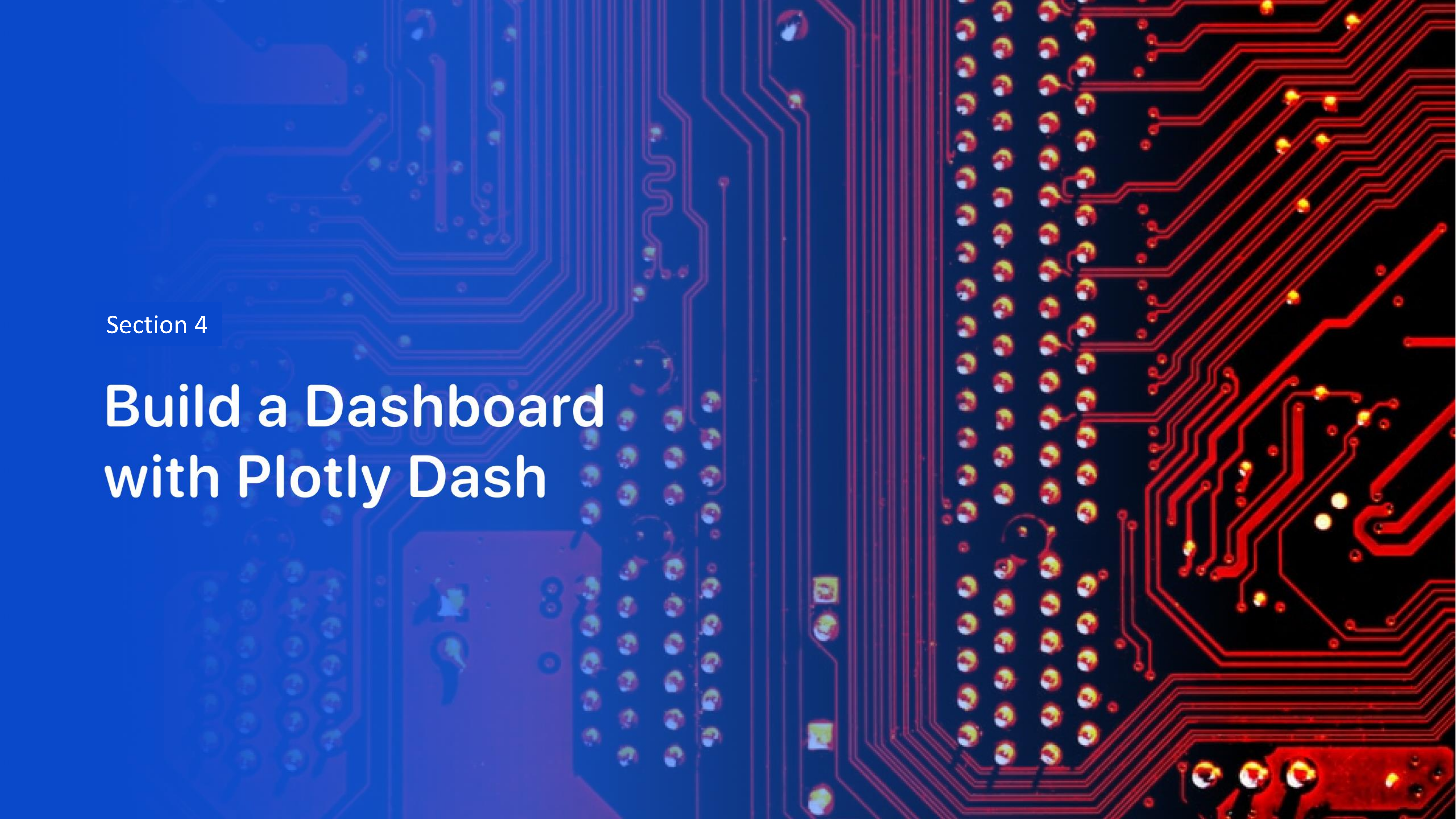
CCAFS SLC-40 proximities

- CCAFS SLC 40 has good proximities to railway, highway and coast.
- All are approximately within a km distance from the launchsite


Titan III Road
Samuel C Phillips Parkway
0.92 KM
Merritt Island National Wildlife Refuge
0.92 KM
CAFS
33

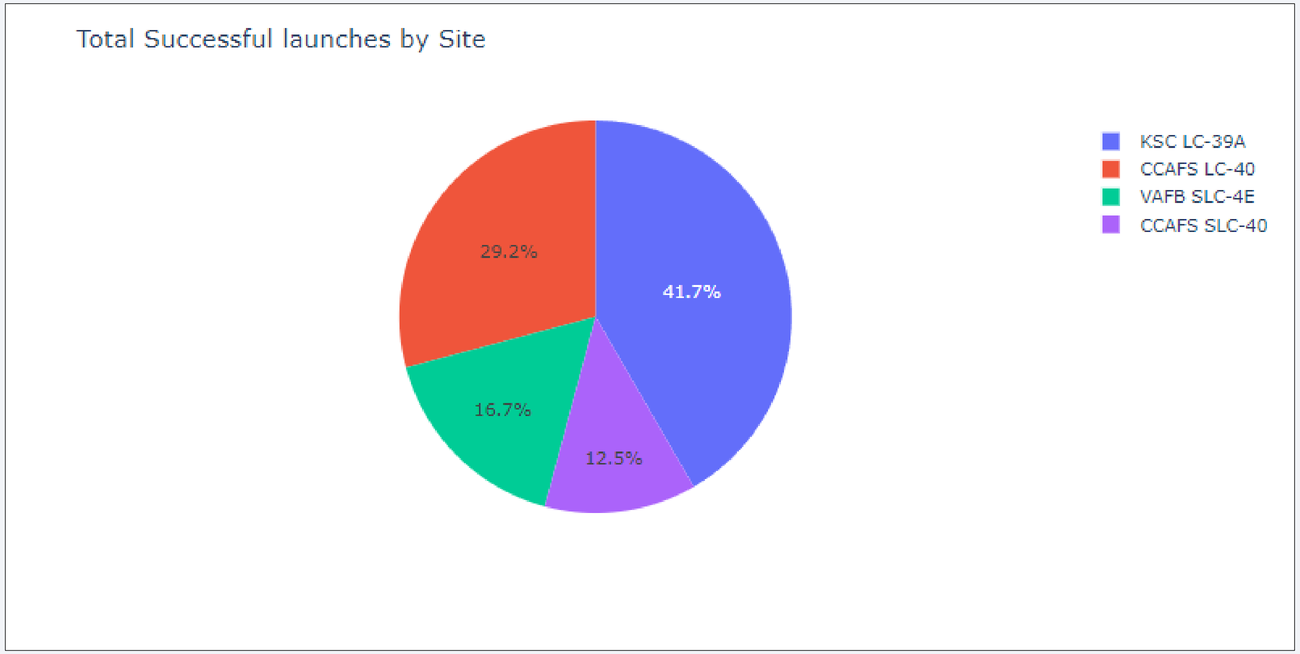
Section 4

Build a Dashboard with Plotly Dash



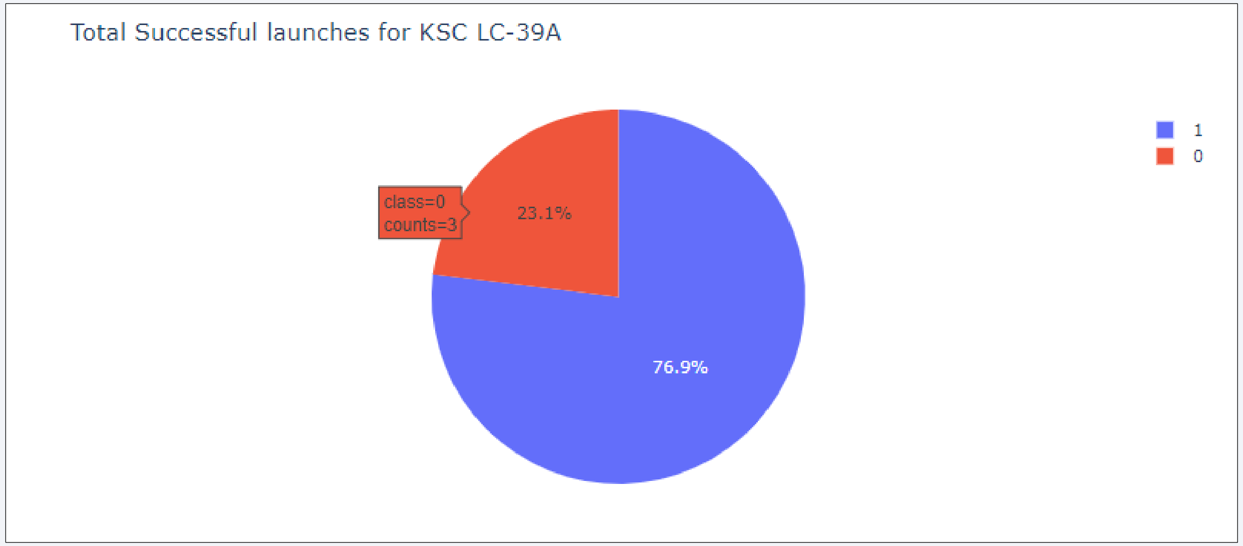
Total Successful launches for all launch sites

- The launch site KSC LC-39A has the highest launch success with 41.7%
- The lowest success rate is from launch site CCAFS SLC-40 (12.5%).



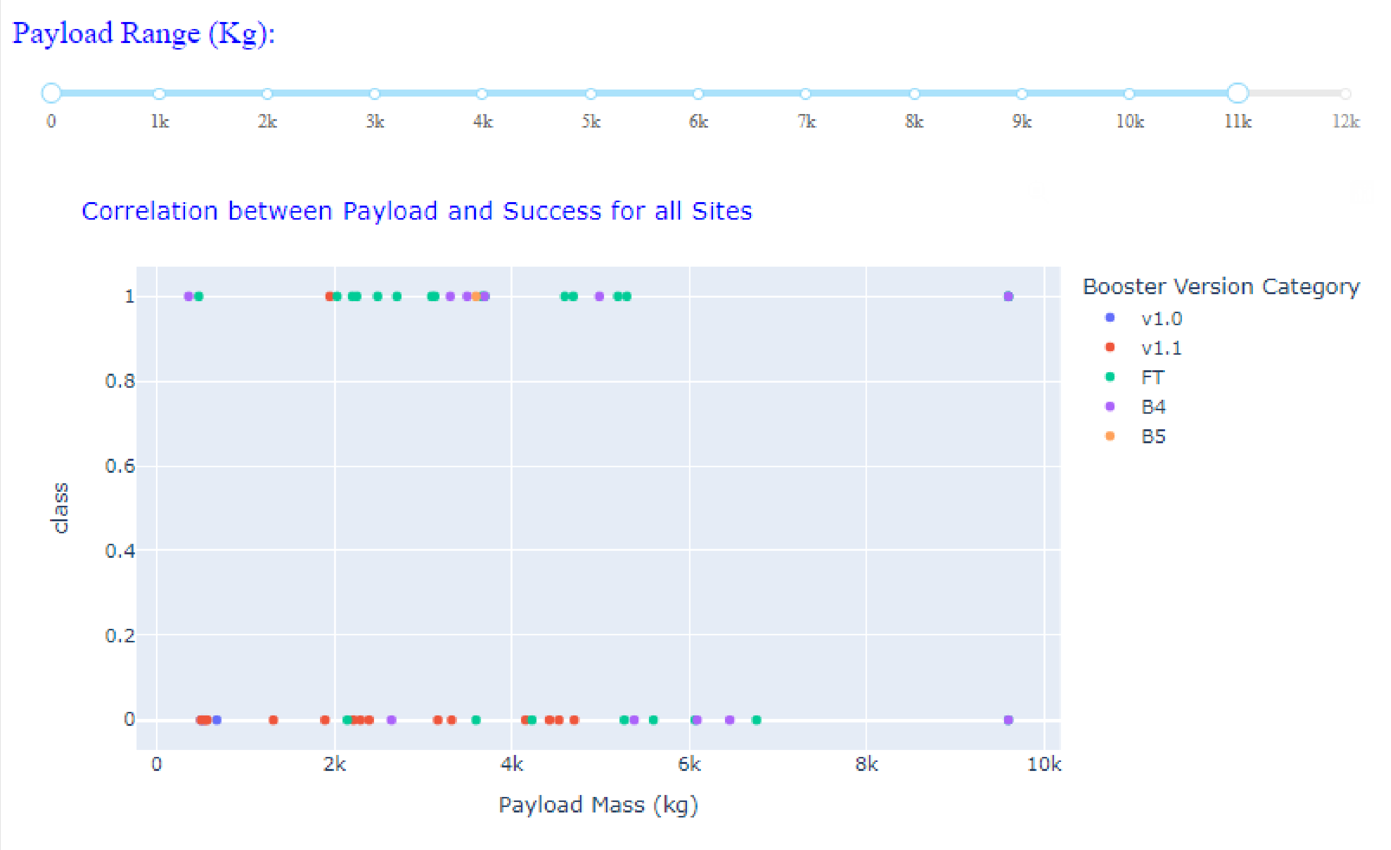
Success ratio for KSC LC-39A

- The given launch site has a success ratio of 76.9 % (highest among all the launch sites)
- Out of 13 total launches, 10 were successful launches and 3 were failures from this site.



Payload vs. Launch Outcome for all launchsites

- The scatter chart shows the success for all sites based on the selected payload (here, 0-11,000 kg is selected)
- The booster version FT has the best success rate for loads between 2000-6000 kg.
- Above 6000 kg, the success is very rare.
- There are very few launches with payload above 8000.

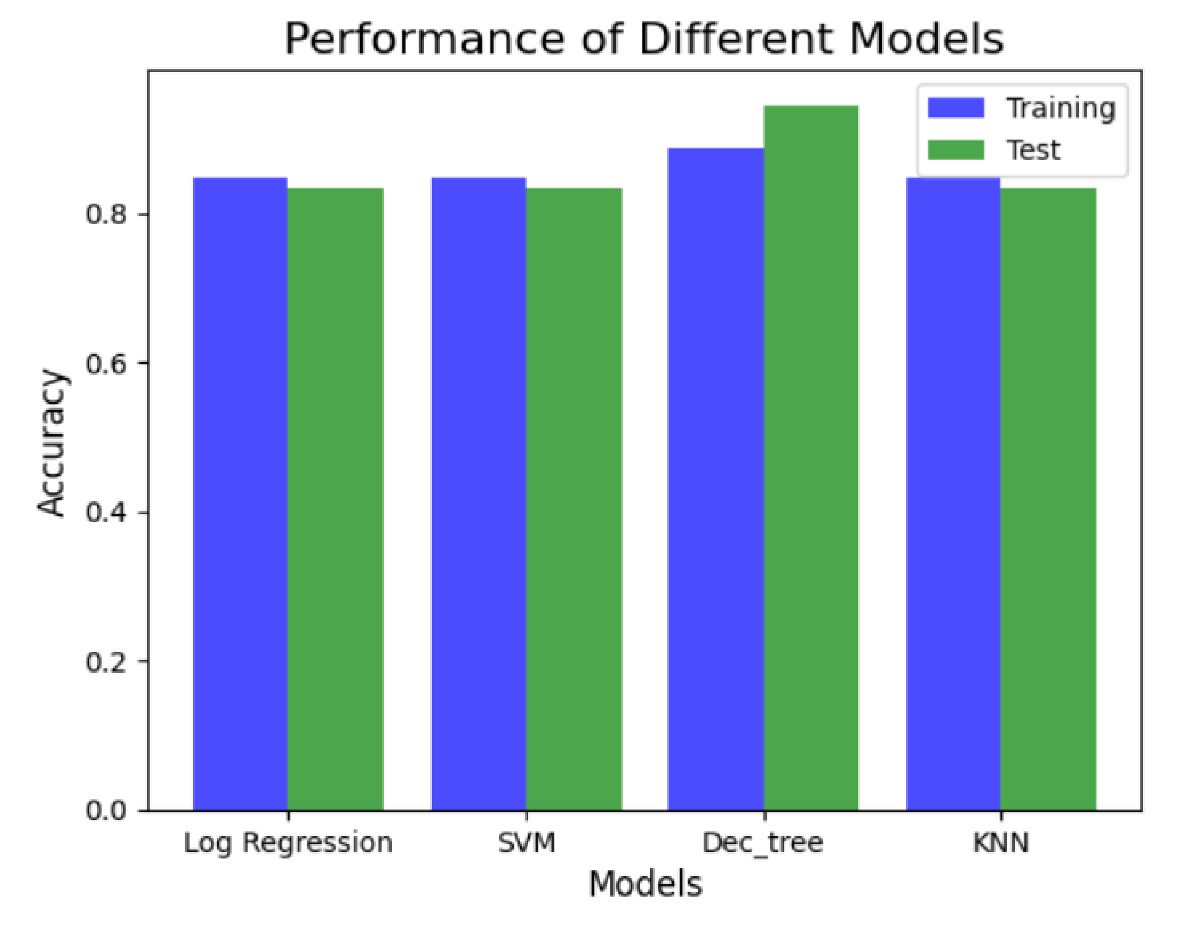


Section 5

Predictive Analysis (Classification)

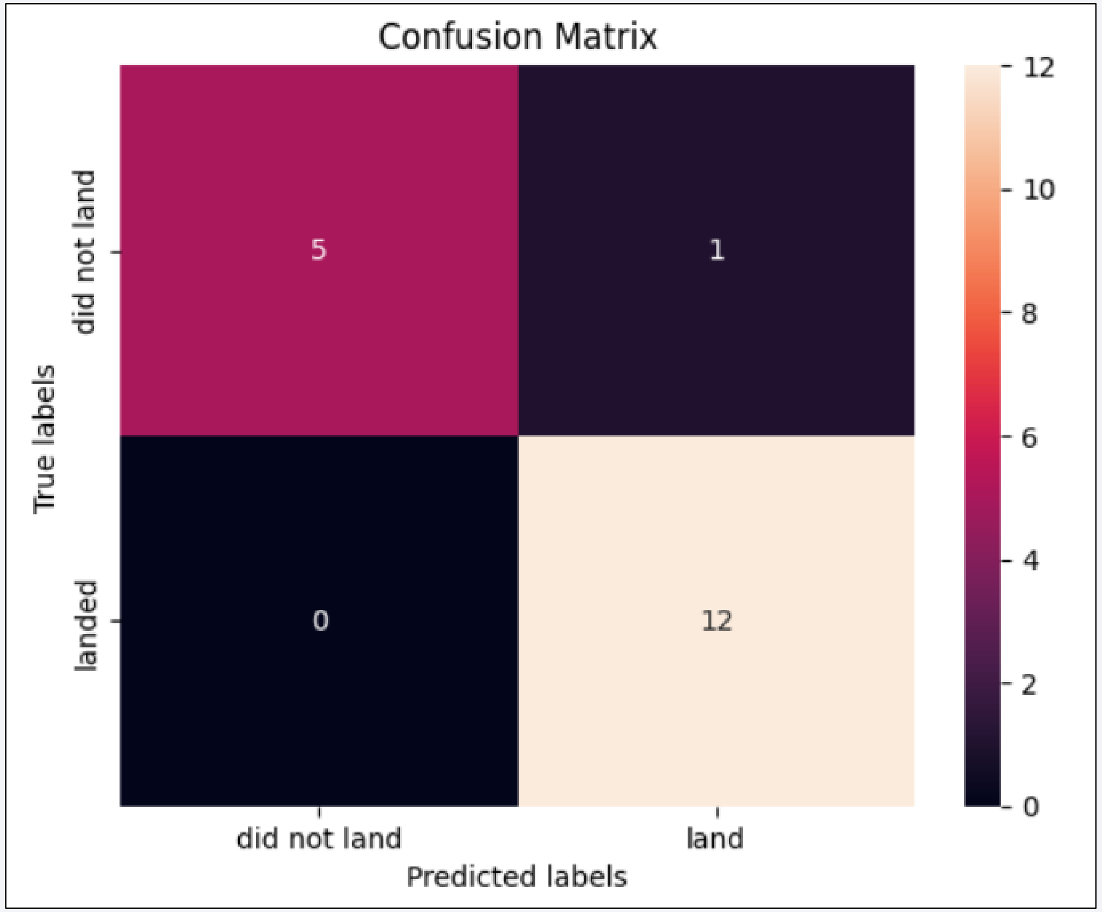
Classification Accuracy

- The decision tree classification model seems to be the best based on its accuracy on both training and test data.



Confusion Matrix

- Confusion matrix of the best performing model (on test data) – Decision tree
- The model does well in most of the conditions except in 1 case, where it was a false positive (model predicted landed, where in actually it did not land)
- Accuracy= 0.944
- Precision= 0.923



Conclusions

- KSC LC-39A is a good option for payload between 2000 and 5000 kg.
- VAFB SLC 4E is an alternate option for payload between 1500 and 4000 kg.
- SSO, VLEO and ISS followed by LEO seem to be interesting orbit options.
- Orbit SSO seems to be suitable for payload mass < 5000 kg. For high payload >= 10000kg, orbits ISS, PO and VLEO seem good
- Although, the project started in 2011, the first successful landing on ground pad was on 22nd December 2015.
- All the launch sites are located in the outskirts of the city, but have close proximities to highways, railways and sea. This helps in transportation and logistics. This also helps in landing on the ground or sea (near to the launch site itself).
- Based on the suitable input data, a decision tree model could be used to predict whether the first stage will land successfully or not. This can come in handy to estimate the project cost and take suitable business decisions.

Appendix

- The dashboard code included in the GitHub needs to be run on a local machine to see the result.

Thank you!