IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ken Zhong
04/18/2024

Outline

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

Executive Summary

- Use data science approach to assess Falcon 9 rocket success rate
  - Falcon 9 rocket launches at a cost of $62 million dollars vs others at $165 million dollars each
  - Much of the savings is due to SpaceX can reuse the first stage for landing
  - I use the API to collect historical data, conduct web scraping, data wrangling, compile SQL queries, develop EDA visualization, generate plotly dashboard, leverage 4 machine learning models (Logistic Regression, SVM, Regression Tree, KNN), and evaluate the chance of such success
  - The objective is to provide business insights for future bids on such project against SpaceX
- Summary of all results
  - Overall Falcon 9 success rate is about 67%, with KSC LC-39A and VAFB SLC 4E sites at 77%
  - With heavy payloads, Polar, LEO and ISS Orbits have higher successful landing rate
  - Among 4 models, Regression Tree model delivers the best accuracy at 89% (based on test data)

Introduction

- Project background and context
  - SpaceX Falcon 9 has achieved great success in launching rockets
  - In order to compete with SpaceX effectively, detailed analysis of its past launches are needed
  - With the readily available public info, applicable data analysis modules, and relevant machine models on hand can be leveraged
- Problems you want to find answers
  - How successful is SpaceX for its past rocket launches?
  - Are there any patterns, learnings lessons, key takeaways based on public available info?
  - How to use applicable machine learning model to predict its future success, thus provide valuable insights for future bids against SpaceX?

Section 1

Methodology

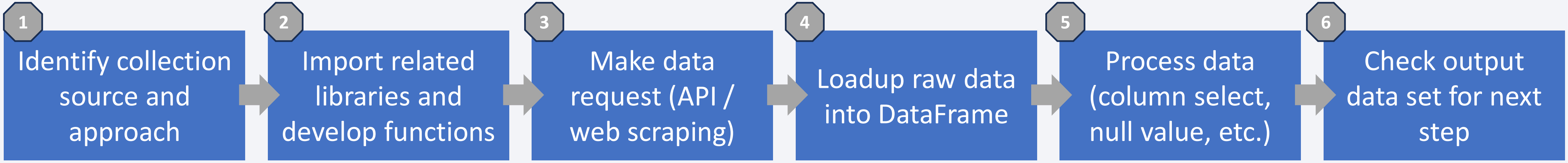
Methodology

Executive Summary

- Data collection methodology:
  - API retrieval and web scraping are used to retrieve data set
- Perform data wrangling
  - Exploratory Data Analysis (EDA) involves data info, feature check, statistic analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Identify ML models, set hyperparameters to optimize, use training set to fit and testing set to predict and compare accuracy score to decide the ideal classification model to use

Data Collection

- 2 methods are used for data collection – API retrieval and web scraping
- API retrieval is effective to retrieve specific data set
  - Past data is available by making an online request through the SpaceX API
  - Raw data from the site need to have additional wrangling / reformatting
  - Use the “Falcon 9” launching data and compile DataFrame object for assessment
- Web scraping covers wide range of data points
  - Apply BeautifulSoup library to extract data for web scrapping
  - Parse HTML file content and table info
  - Compile DataFrame object and cleanup data set for assessment



1 Identify collection source and approach

2 Import related libraries and develop functions

3 Make data request (API / web scraping)

4 Loadup raw data into DataFrame

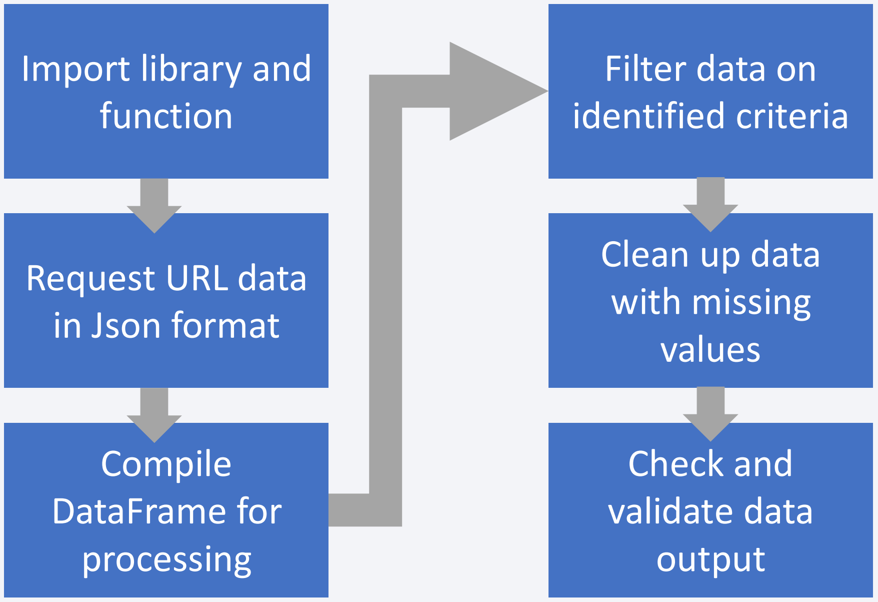
5 Process data (column select, null value, etc.)

6 Check output data set for next step

Data Collection – SpaceX API

- The completed SpaceX API calls notebook on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/jupyter-labs-spacex-data-collection-api.ipynb)
- The output data file on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/data/dataset_part_1.csv)

Flowchart of SpaceX API calls



Import library and function

Request URL data in Json format

Compile DataFrame for processing

Filter data on identified criteria

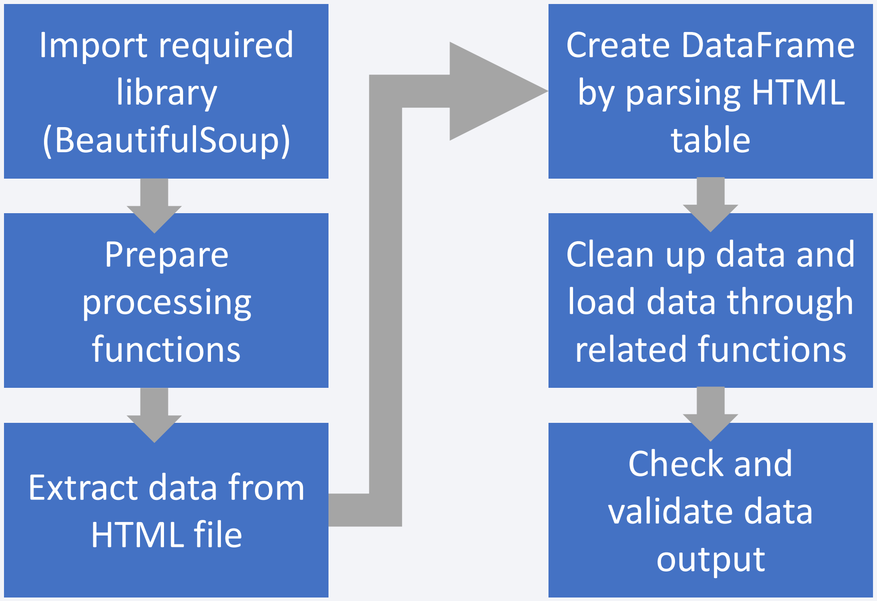
Clean up data with missing values

Check and validate data output

Data Collection - Scraping

- The completed web scraping notebook on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/jupyter-labs-webscraping.ipynb)
- The output data file on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/data/spacex_web_scraped.csv)

Flowchart of Web Scraping



Import required library (BeautifulSoup)

Prepare processing functions

Extract data from HTML file

Create DataFrame by parsing HTML table

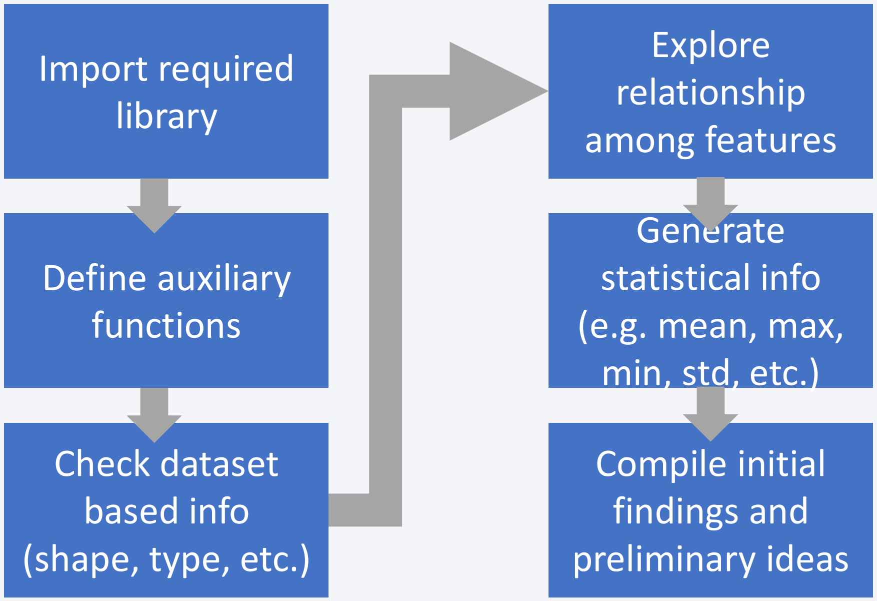
Clean up data and load data through related functions

Check and validate data output

Data Wrangling

- Involves Exploratory Data Analysis (EDA)
- The completed data wrangling notebook on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/labs-jupyter-spacex-Data%20wrangling.ipynb)
- The output data file on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/data/dataset_part_2.csv)

Flowchart of Data Wrangling



Import required library
Define auxiliary functions
Check dataset based info (shape, type, etc.)
Explore relationship among features
Generate statistical info (e.g. mean, max, min, std, etc.)
Compile initial findings and preliminary ideas

EDA with Data Visualization

- Charts were compiled for visualization and assessment
  - Scatter plot on “FlightNumber” (indicating the continuous launch attempts.) vs “Payload” to see impacts on launch outcome
  - Scatter plot on “FlightNumber” vs “LaunchSite” to see different sites on launch outcome
  - Scatter plot on “LaunchSite” vs “PayloadMass” to see payload mass and sites on outcome
  - Bar plot on “Orbit” vs “Class” to see success rate on each orbit type
  - Scatter plot on “FlightNumber” vs “Orbit” to see different orbit type used over time
  - Scatter plot on “PayloadMass” vs “Orbit” to see orbit type used for different payload mass
  - Line chart on “Year” vs “Class” to see the success rate over time
- Apply other data processing – OneHotEncoder, datatype casting, etc.

- The completed EDA with data visualization notebook on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/edadataviz.ipynb)
- The output data file on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/data/dataset_part__3.csv)

EDA with SQL

- Conduct substantial amount of SQL queries to assess data
  - 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 successful landing outcome in ground pad was achieved
  - 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 (use 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 landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

- The completed EDA with SQL notebook on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/jupyter-labs-eda-sql-coursera_sqllite.ipynb)

- The output SQL data file on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/data/my_data1.db)

Build an Interactive Map with Folium

- Several map objects are created and added to a folium map
  - Circle object – mark all launch sites on a map
  - Marker / MarkerCluster object – provide additional info and mark the success/failed launches for each site on the map
  - MousePosition – identify coordinates info for mouse pointer
  - PolyLine – calculate the distances between a launch site to its proximities and mark it on the map
- The completed interactive map with Folium map is available on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/lab_jupyter_launch_site_location.ipynb)

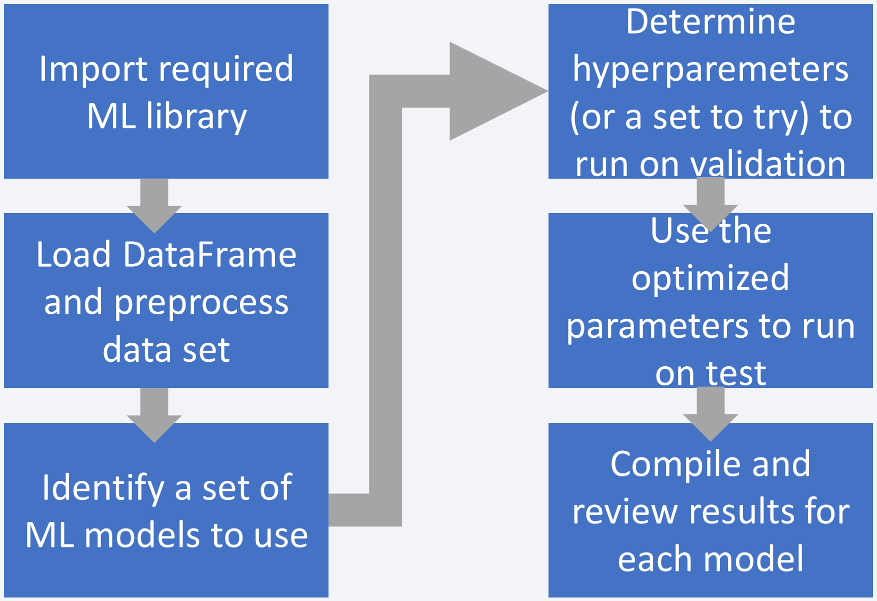
Build a Dashboard with Plotly Dash

- 2 dropdown lists and 1 slide bar are used to compile plots and graphs
  - SpaceX Launch Records Dashboard – based on All sites or each selected site to show success rate
  - Scatter chart to show the correlation between payload and launch success – based on payload mass range selected to show success / no success info for each booster version category
- The completed Plotly Dash lab is available on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/spacex_dash_app.py)

Predictive Analysis (Classification)

- Involves Machine Learning models for classification predictive analysis - Logistic Regression, SVM, Decision Tree, KNN
- The completed machine learning model on Github (https://github.com/nekcool/ibm_ds_capstone/blob/main/code/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

Flowchart of Machine Learning Model



Import required ML library

Load DataFrame and preprocess data set

Identify a set of ML models to use

Determine hyperparemeters (or a set to try) to run on validation

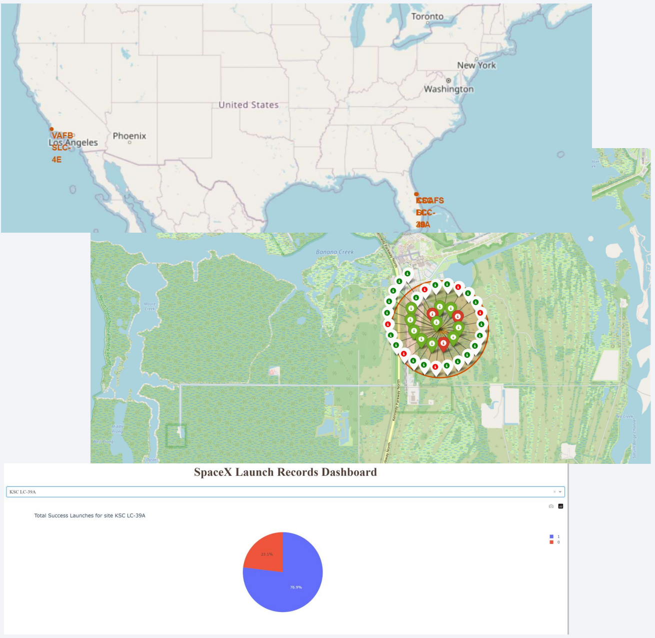
Use the optimized parameters to run on test

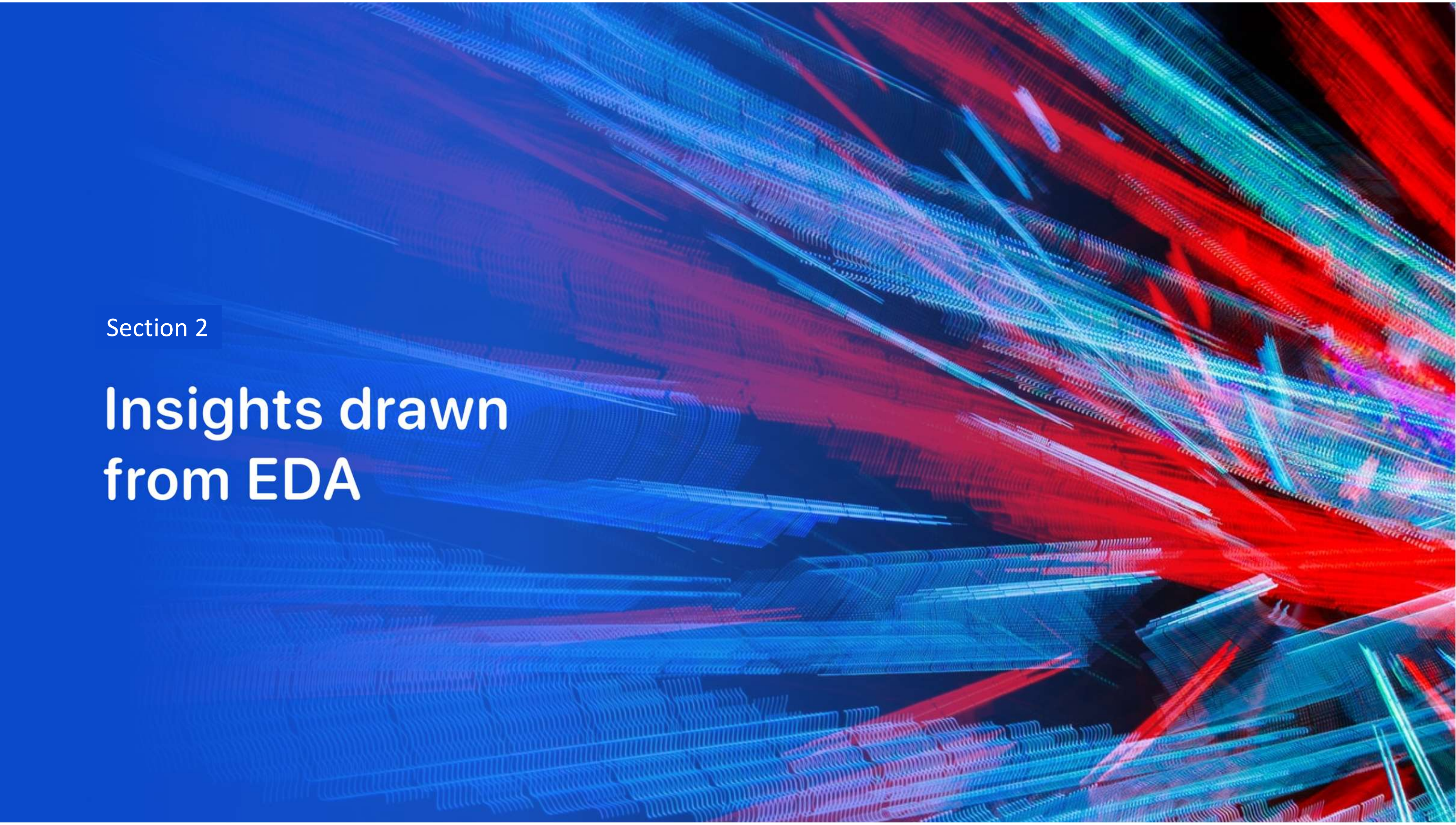
Compile and review results for each model

Results

- Exploratory data analysis results
  - Most of the launches at VAFB SLC 4E site are successful – 10 successes vs 3 failures
  - VAFB-SLC site did not launch heavy payload mass (greater than 10000)
  - ES-L1, SSO, HEO, and GEO orbit types have high success rate
  - Overall success rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots
  - While all 4 launching sites are close to coastal areas, CCAFS LC-40 is close to the coastal line with only 0.93 km away
  - Out of the total 24 successful launches, KSC LC-39A site has the largest successful launches at 10
- Predictive analysis results
  - Logistics Tree model has the highest classification accuracy at 89%

Selected Screen Shots



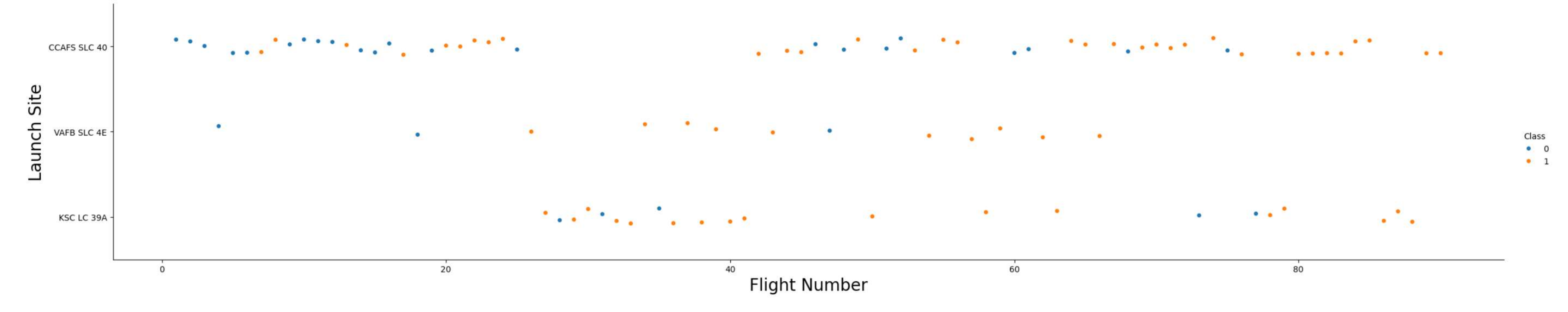


Section 2

Insights drawn from EDA

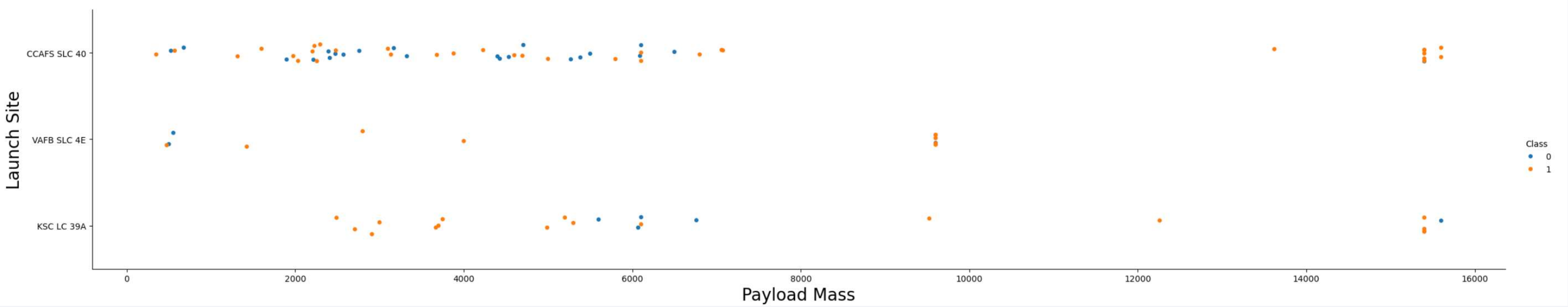
Flight Number vs. Launch Site

- CCAFS SLC 40 site has most of the launches in early days (small flight number), many failed
- Most of the launches at VAFB SLC 4E site are successful – 10 successes vs 3 failures
- KSC LC 39A site started to use after around 25th flight, and have delivered solid success track records since then



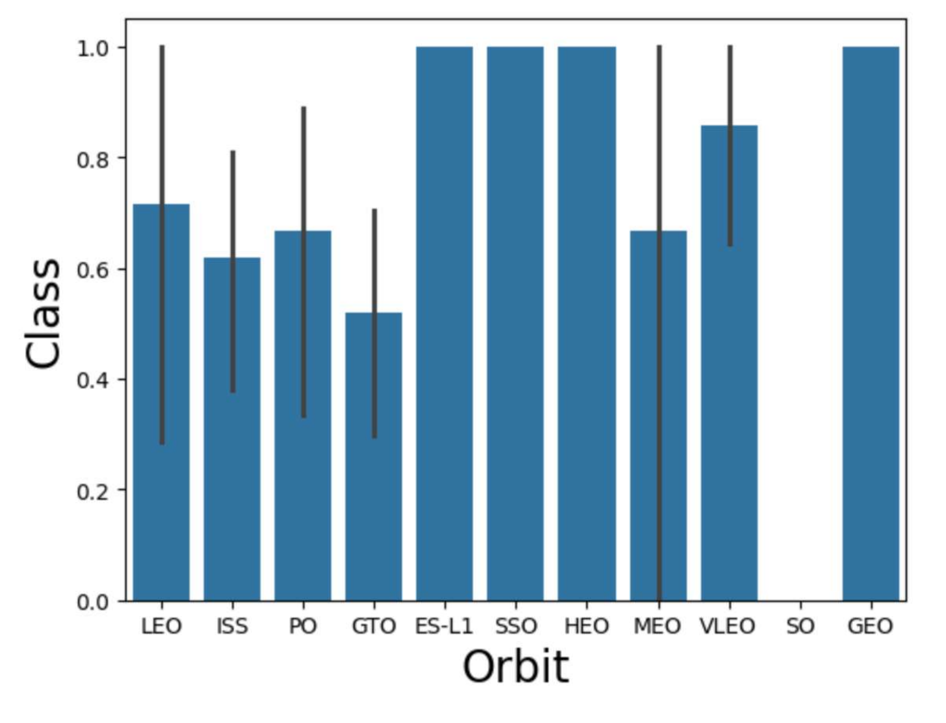
Payload vs. Launch Site

- CCAFS SLC 40 site has a wide range of payload mass, mostly below 7000 kg, and above 15000 kg
- VAFB-SLC site did not launch heavy payload mass (greater than 10000)
- KSC LC 39A site has both light and heavy payload mass launches



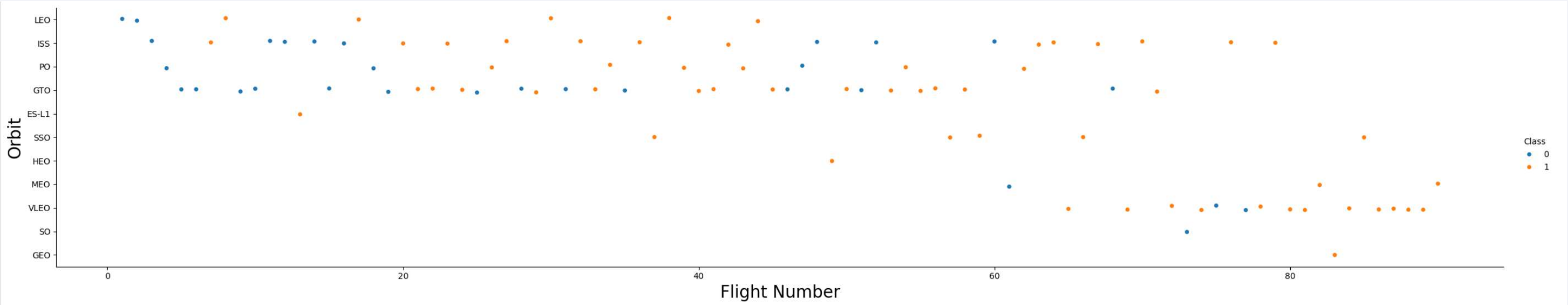
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO, and GEO have high success rate
- GTO and ISS have relatively low success rate
- MEO and LEO variance levels are fairly high



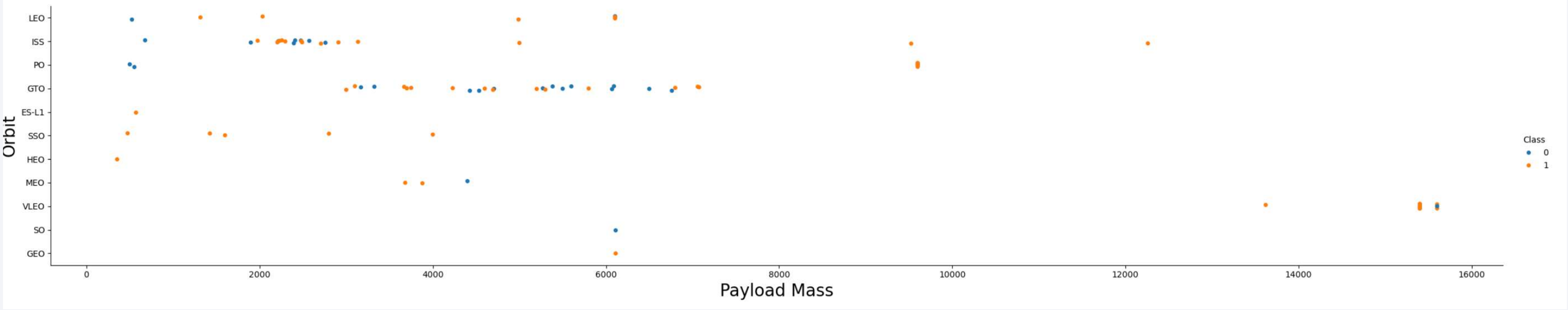
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit



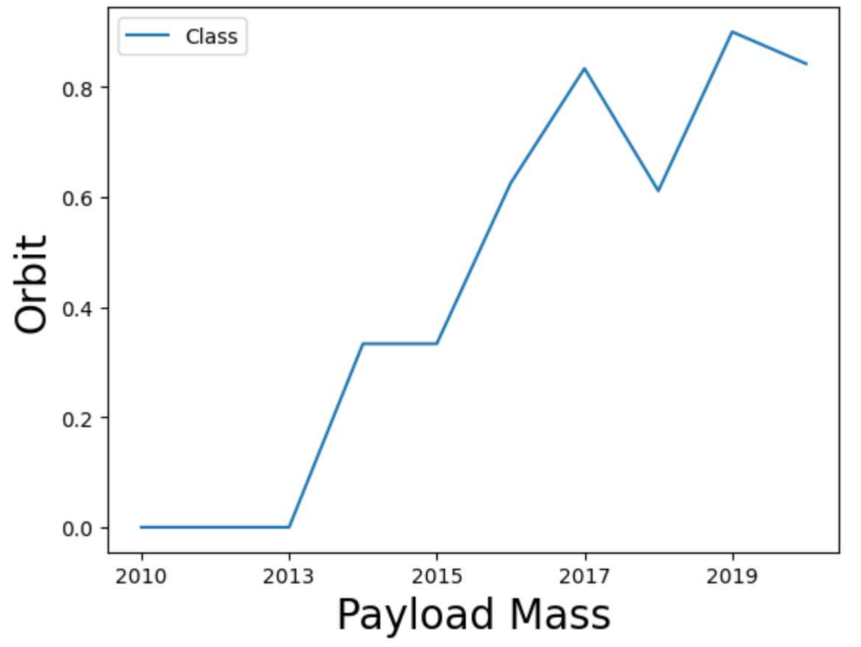
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- GTO performance is quite mixed, both positive landing rate and negative landing (unsuccessful mission) are observed



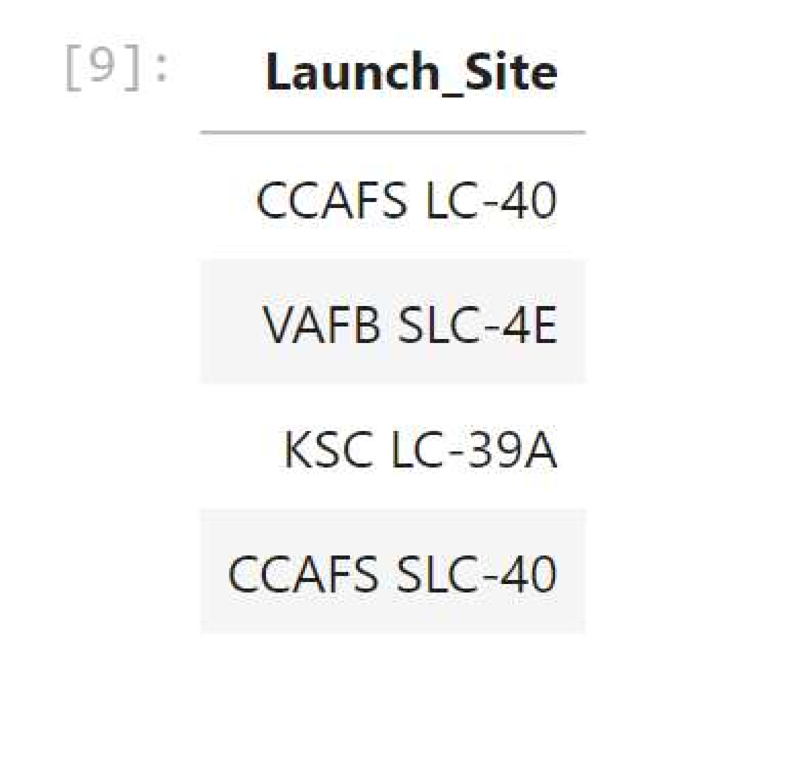
Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020
- Performance in 2018 has a small dip



All Launch Site Names

- There are 4 unique launch sites

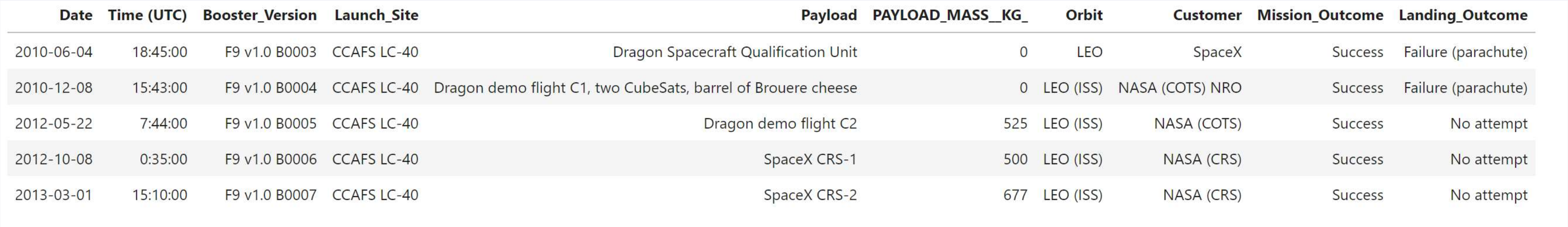


[9]:

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

Launch Site Names Begin with 'CCA'

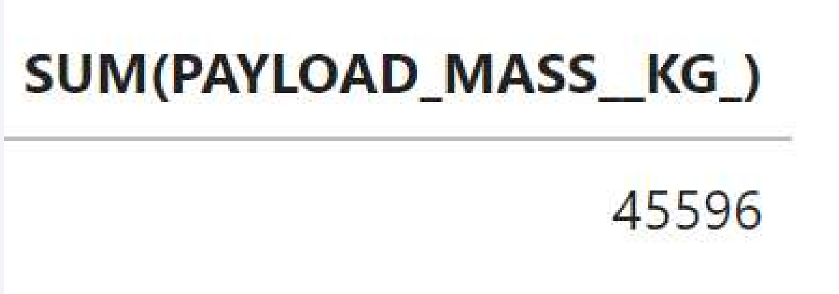
- The 5 earliest launches at launch sites begin with `CCA`
- Actually all of these 5 are at CCAFS LC-40
- Unfortunately 2 are failures (parachute) and 3 are no attempts for landing



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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

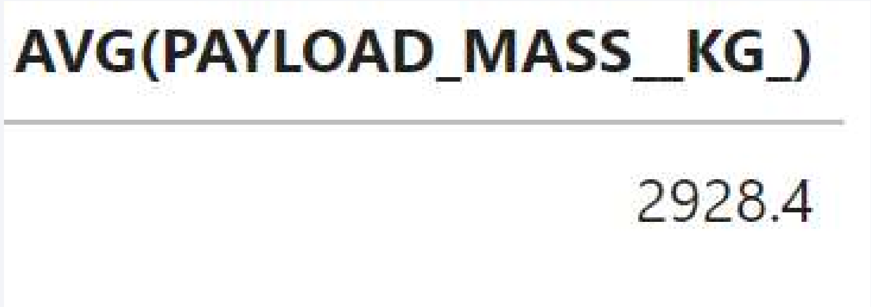
- Total payload carried by boosters from NASA (CRS) is 45,596 kg



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

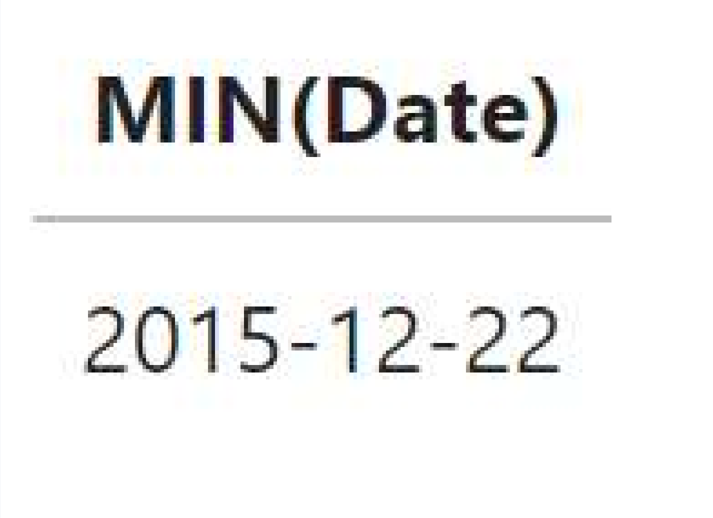
- Average payload mass carried by booster version F9 v1.1 is 2,928.4 kg



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

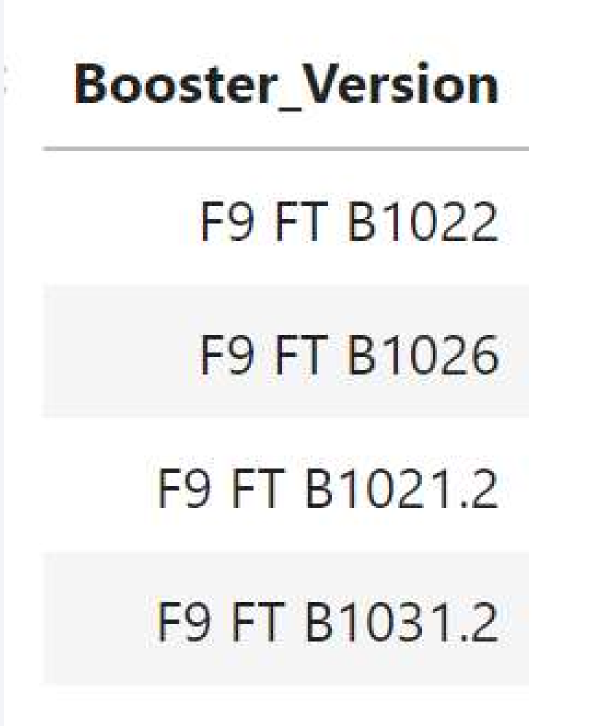
- The first successful landing on ground pad dated on 2015-12-22
- It is a big success for SpaceX, saving the company from going bankrupt



MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

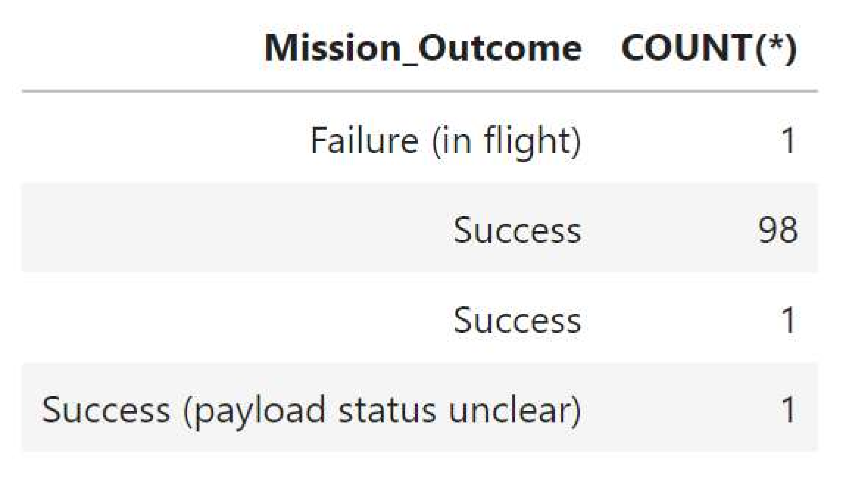
- There are 4 names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg



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

Total Number of Successful and Failure Mission Outcomes

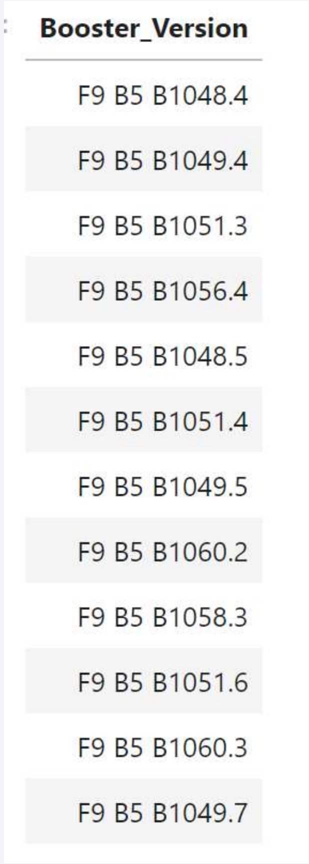
- Out of the 101 launches, based on mission outcome perspective, 99 are success, 1 success with payload status unclear and 1 failure in flight



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

Boosters Carried Maximum Payload

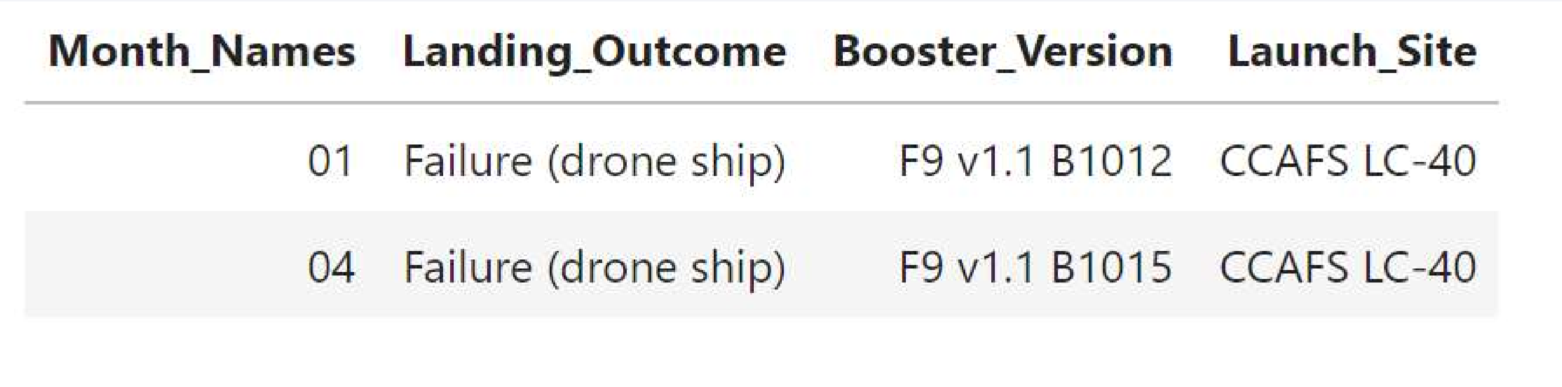
- There are 12 names of the booster which have carried the maximum payload mass



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

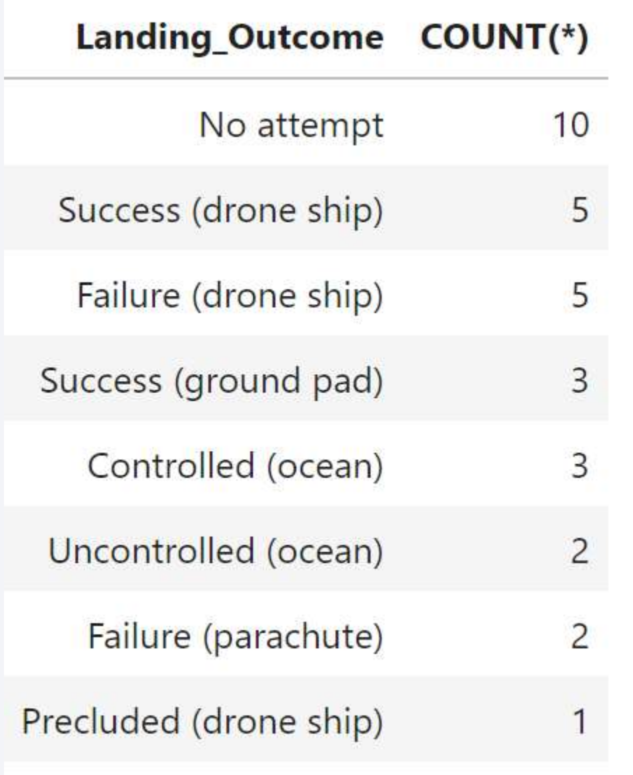
- In 2015, there are 2 records for the failed landing outcomes in drone ship
- Their booster versions are F9 v1.1 B1012 and F9 v1.1 B1015
- Both were launched at CCAFS LC-40 site



Month_Names	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Between the date 2010-06-04 and 2017-03-20, there are 8 types of landing outcomes
- No attempts have the highest count at 10, followed by Success (drone ship) and Failure (drone ship) – both at 5
- The least 3 counts are Precluded (drone ship) at 1, Failure (parachute) at 2, and Uncontrolled (ocean) at 2

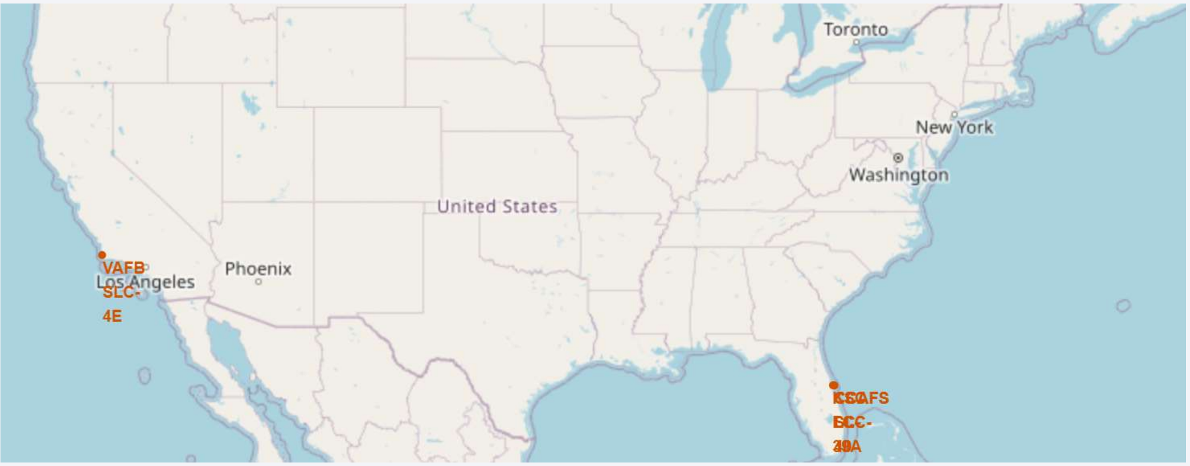


Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

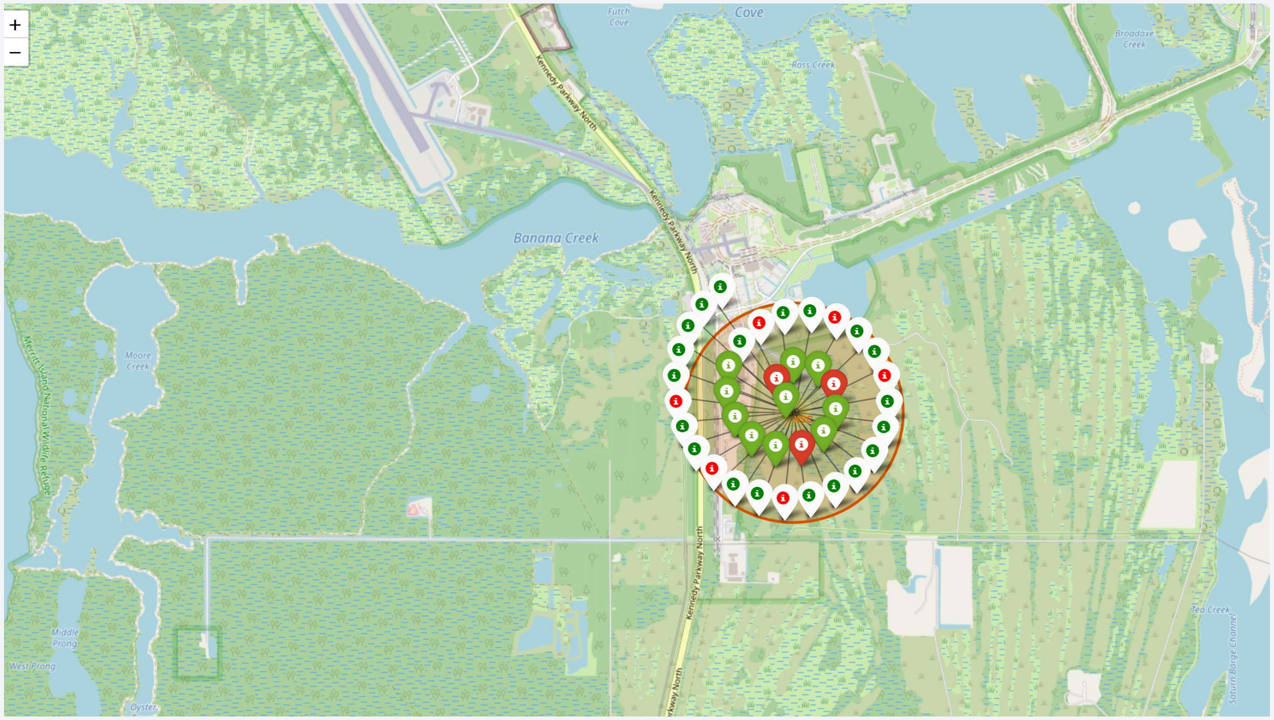
Launch Sites Proximities Analysis

There are 4 launching sites in US



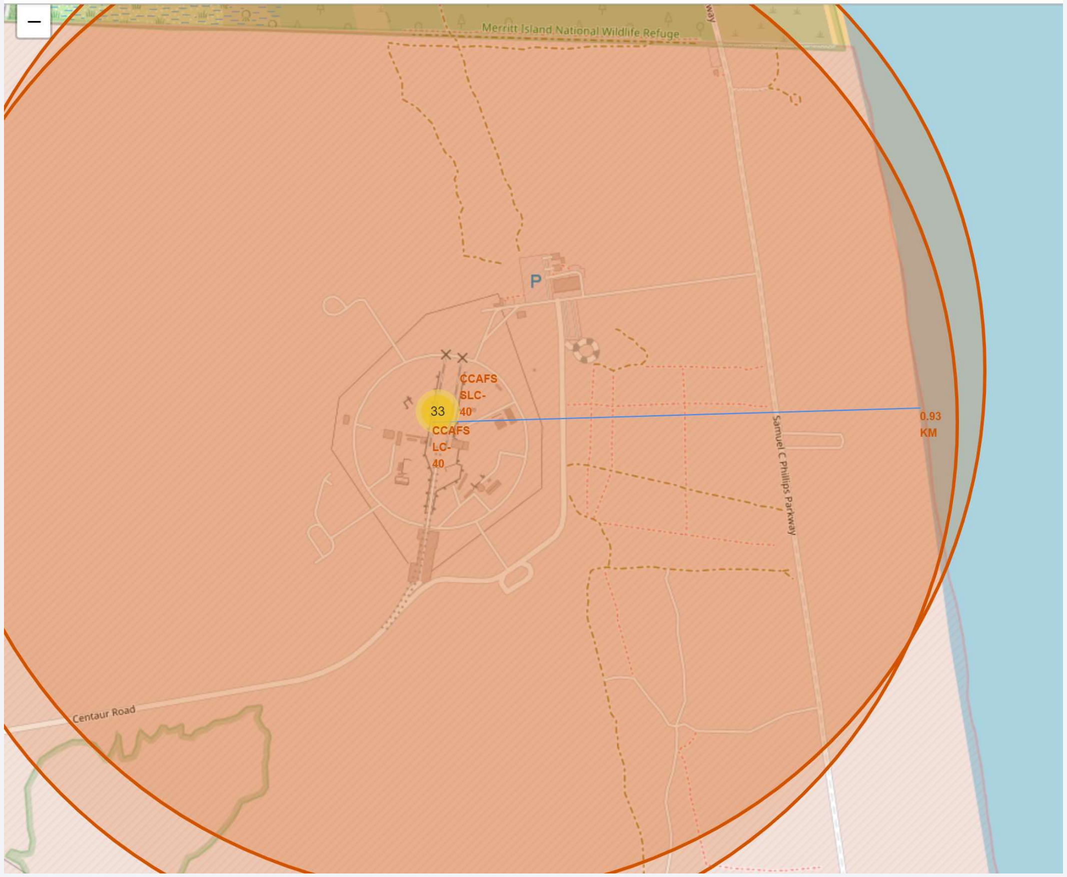
- Florida has 3 sites, including CCAFS LC-40 / CCAFS SLC-40 / KSC LC-39A
- California has 1 site as VAFB SLC-4E
- All launch sites in very close proximity to the coast but NOT to the Equator line

KSC LC-39A site has fairly high chance of success

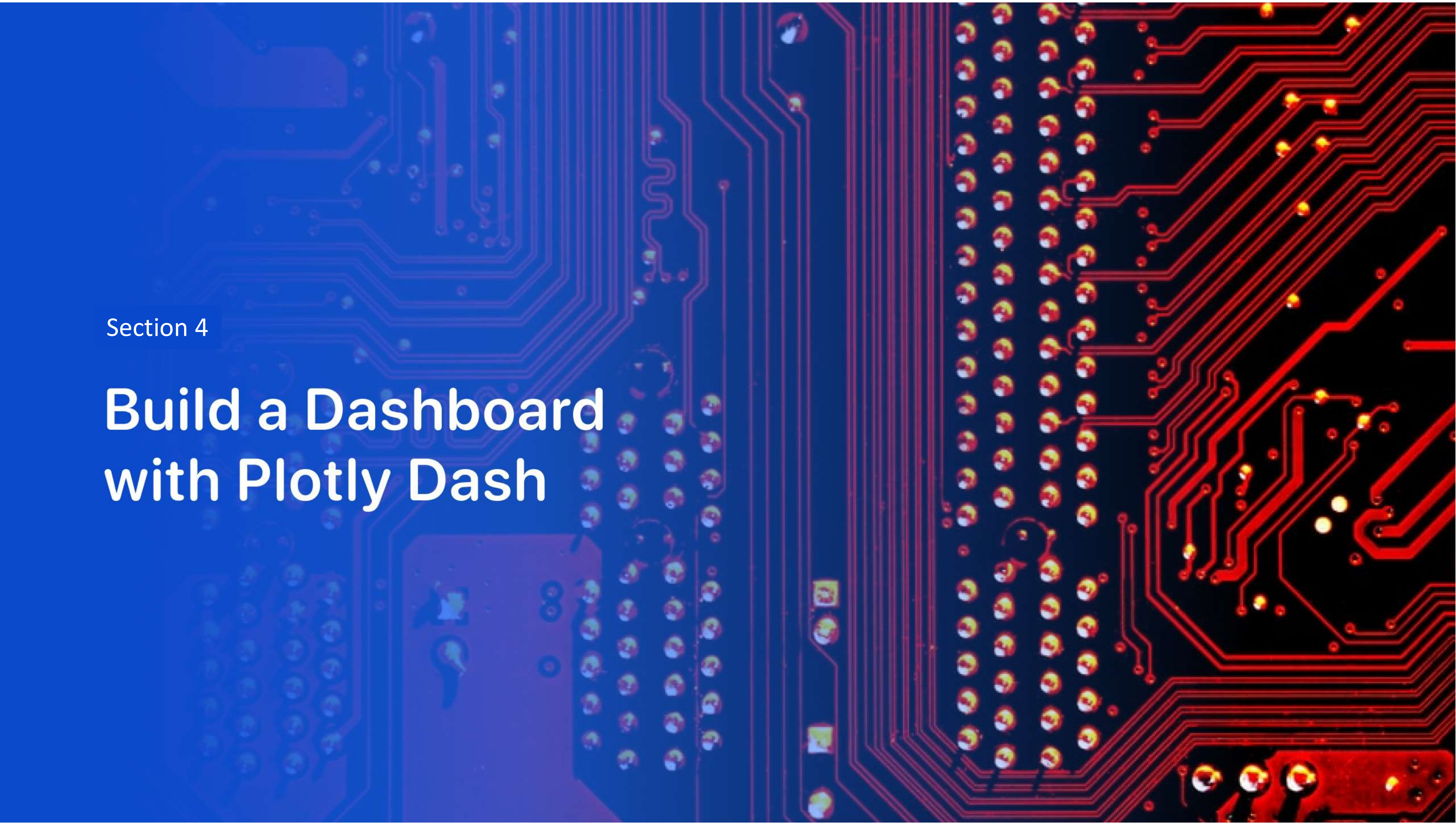


- Out of 39 launches at KSC LC-39A, 30 are successful
- While other sites do not yield such high success rate

CCAFS LC-40 site is only 0.93 km away from the coast


−
Merritt Island National Wildlife Refuge
P
CCAFS SLC-40
33
CCAFS LC-40
0.93 KM
Samuel C Phillips Parkway
Centaur Road

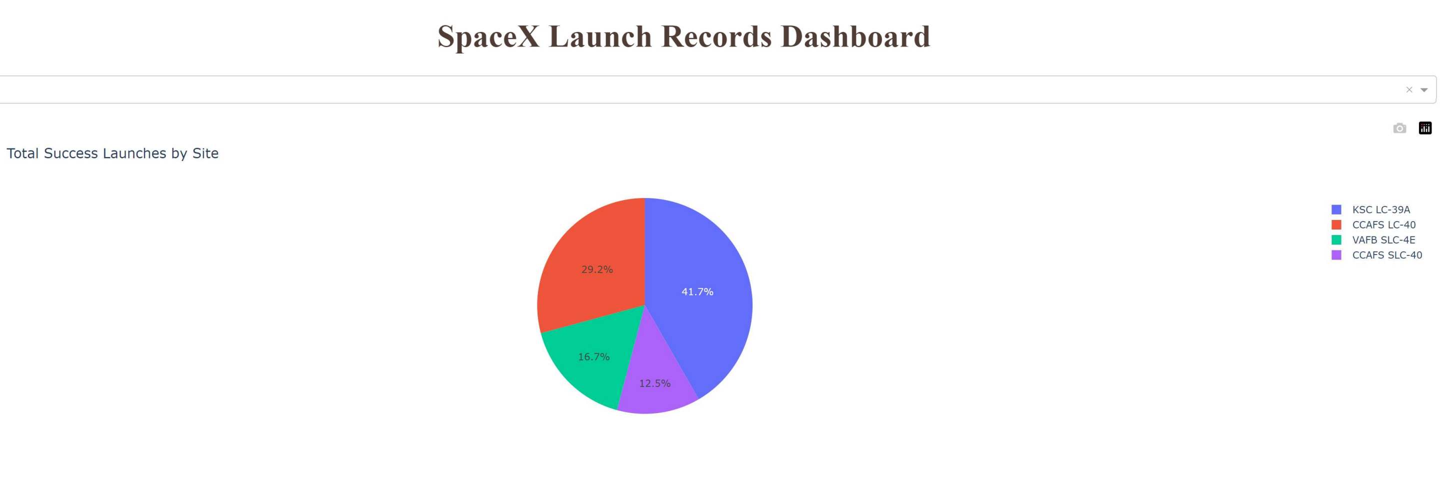
- While all 4 launching sites are close to coastal areas, CCAFS LC-40 is close to the coastal line with only 0.93 km away



Section 4

Build a Dashboard with Plotly Dash

KSC LC-39A site has the largest successful launches

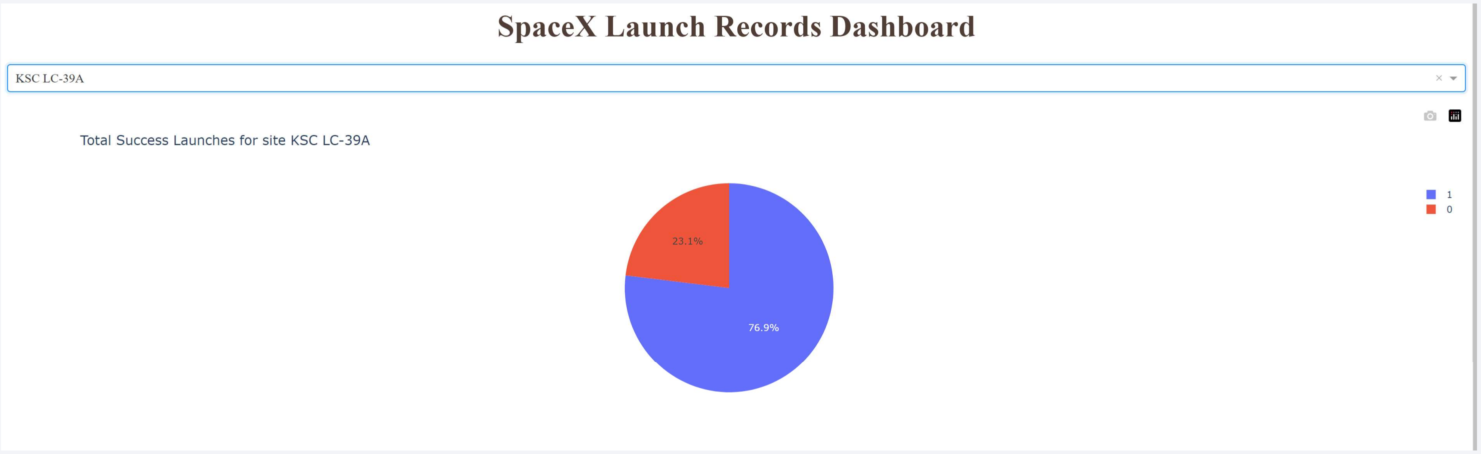


SpaceX Launch Records Dashboard

Total Success Launches by Site

- Out of the total 24 successful launches, KSC LC-39A site has the largest successful launches at 10
- CCAFS LC-40 has 7 successful launches, followed by VAFB SLC-4E (4 successes) and CCAFS SLC-40 (3 successes)

KSC LC-39A site also has the highest success rate

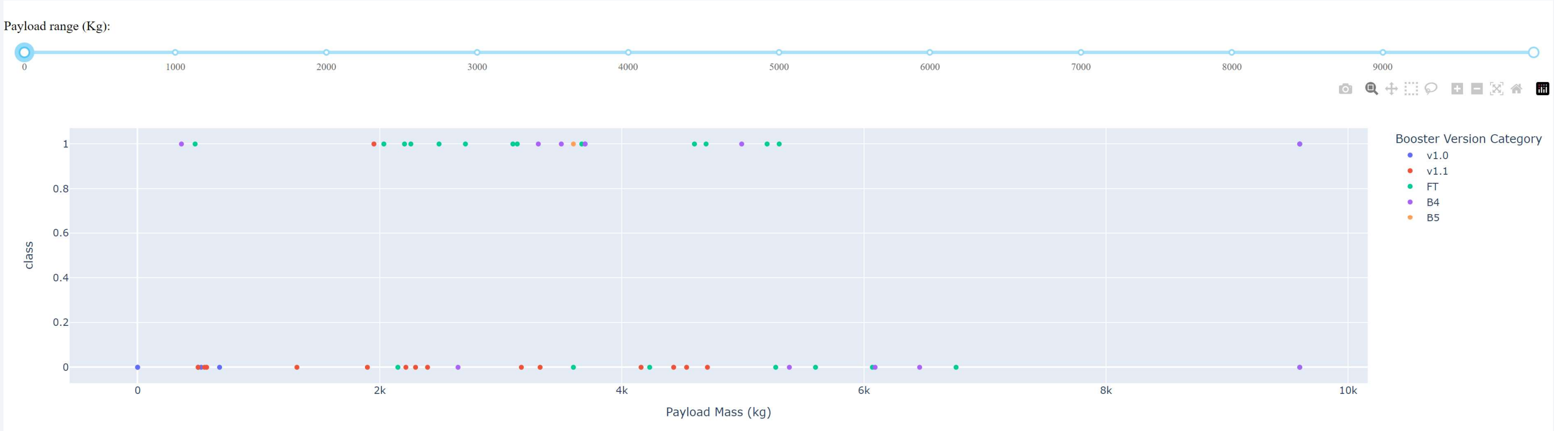


SpaceX Launch Records Dashboard

KSC LC-39A

- KSC LC-39A site has close to 77% of the success rate of all launches at its site

2000-4000 kg payload and B5 booster has high chance to succeed



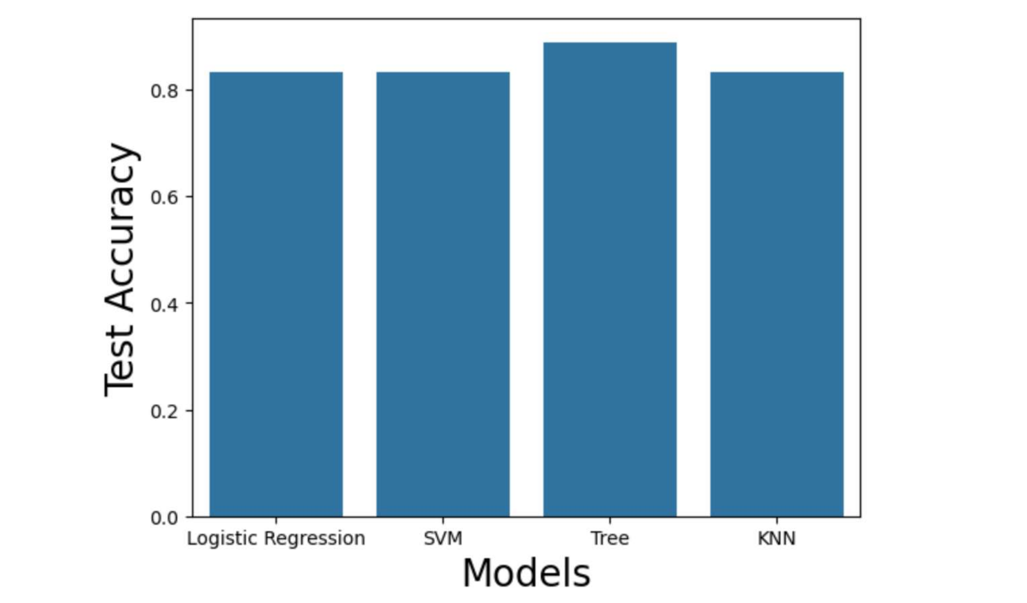
- 2000 – 4000 kg payload range has the highest launch success rate
- 2000 – 4000 kg payload range has the lowest launch success rate
- Among all F9 Booster versions, B5 has the highest launch success rate, followed by FT

Section 5

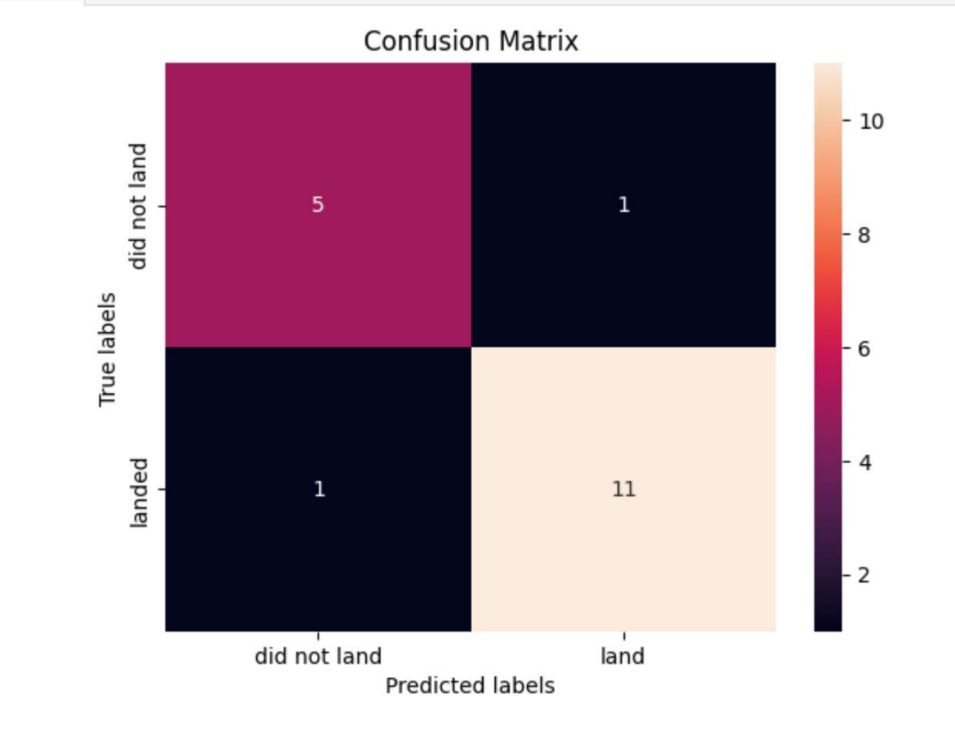
Predictive Analysis (Classification)

Classification Accuracy

- 4 machine learning models are used, with the optimal hyperparameters being tuned
    - Logistics Regression {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs' }
    - SVM {'C': 1.0, 'gamma': 0.03162277660168, 'kernel': 'sigmoid’ }
    - Logistics Tree {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}
    - KNN {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
- Sample size at 90, with 80% as training set, 20% as testing set, cross validation as 10
- Logistics Tree model has the highest classification accuracy at 89%



Confusion Matrix



- Linear Regression model has achieved the overall best score at 89% (test data)
- For landing scenario, prediction shows
  - accuracy at 92% (1 false positive case)
  - recall also at 92% (1 false negative)

Conclusions

- SpaceX has developed a strong foothold in rocket launches
  - Most of the launches at VAFB SLC 4E site are successful – 10 successes vs 3 failures
  - VAFB-SLC site did not launch heavy payload mass (greater than 10000)
  - ES-L1, SSO, HEO, and GEO orbit types have high success rate
  - Overall success rate since 2013 kept increasing till 2020
- Major launch sites are capable to carry out mission critical projects
  - While all 4 launching sites are close to coastal areas, CCAFS LC-40 is close to the coastal line with only 0.93 km away
  - Out of the total 24 successful launches, KSC LC-39A site has the largest successful launches at 10
- Predictive model is developed to assess future SpaceX launches – can be used for future bid
  - Logistics Tree model has the highest classification accuracy at 89%

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that were created during this project are available at Github (https://github.com/nekcool/ibm_ds_capstone)

Thank you!