IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roberto Sanz Garcia
February 24th, 2022

Outline

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

Executive Summary

- Data collection has been performed using 2 methods: API and Webscrapping
- Data wrangling mostly done using Pandas dataframe flexibility
- Ordinary Exploratory Data Analysis (EDA) was made using Matplotlib and Seaborn libraries and SQL
- More advanced interactive EDA was made with Folium and Plotly and Dash

- EDA revealed what features were important for the outcome
- Interactive EDA stressed the importance of 'Payload' and 'Launch Site' as variables for a landing success outcome
- Prediction models were trained and tuned with the relevant data to be able to make a prediction with an expected accuracy of 94%

Introduction

- The problem to solve is to determine whether the first stage of Space X Falcon 9 rockets will land successfully
- A successful landing will enable to reuse the first stage
- The savings of reusing the first stage will determine the cost
- An accurate cost is an essential information to calculate the price of a launch when bidding and this is our TARGET

Past launches data will be used to predict if the first stage will land successfully in a new launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using 2 methods: Space-X API (REST API) and Webscrapping
- Perform data wrangling
  - Normalizing json format data collected, and puttting into a Pandas dataframe for processing: filling up missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different models from Sci-kit learn library were used: created trained and evaluated

Data Collection

- Data has been collected using 2 methods:
  - Space X API (REST API)
    - Data comes from Space X server via API
    - Github: https://github.com/RobertoSanzCapstone/testrepo/blob/main/Capstone%20Data%20Collection%20with%20REST%20API%20Lab.ipynb
  - Webscrapping
    - Data comes from a Wikipedia Web page
    - Github https://github.com/RobertoSanzCapstone/testrepo/blob/main/Capstone%20Data%20Collection%20with%20Webscraping.ipynb
- Same Library and function is used: requests library and get function

Data Collection - `Using Space X API (REST API)`

Import requests
requests.get(spacex_url)

Python library to get data from an API

'get' function

Python Library Pandas has a data structure called Dataframe very convenient to extract information easlily

Info comes from Space X API

response.json()
pd.json_normalize(data2)

b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/3c/0e/T8iJcSN3_o.png","large":"https://images2.imgbox.com/40/e3/GypSkayF_o.png"},"reddit":{"campaign":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":[],"original":[]},"presskit":null,"webca

Normalizing json format retrieved
And into 'dataframe' data structure

ayloads	launchpad	cores	flight_number	date_utc	date
eeb1e1	5e9e4502f5090995de566f86	{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	1	2006-03-24T22:30:00.000Z	2006-03-24
eeb1e2	5e9e4502f5090995de566f86	{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	2	2007-03-21T01:10:00.000Z	2007-03-21
eeb1e5	5e9e4502f5090995de566f86	{'core': '5e9e289ef3591855dc3b2626', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	4	2008-09-28T23:15:00.000Z	2008-09-28

Extracting information with Python code functions from Pandas dataframe

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Re
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	

Data Collection - Using Webscrapping

Import requests
requests.get(wiki_url)

Also 'requests' library to get data

Also 'get' function used

Source is a wikipedia Web page !!!

From json format with an html parser

We also use Pandas Dataframe
Previously we use BeautifulSoup library

2010 to 2013

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Re
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	

Data Wrangling

- Firstly, filling up missing values (with the mean)
  - to use all data and avoid discarding rows

https://github.com/RobertoSanzCapstone/testrepo/blob/main/Capstone%20EDA%20Lab%20Data%20Wrangling.ipynb

- Then creating the Outcome variable (new column 'Class') relevant for my study

Outcome

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
None ASDS	2
False Ocean	2
False RTLS	1

Filtering with set of bad_outcomes

'False ASDS',
'False Ocean',
'False RTLS',
'None ASDS',
'None None'

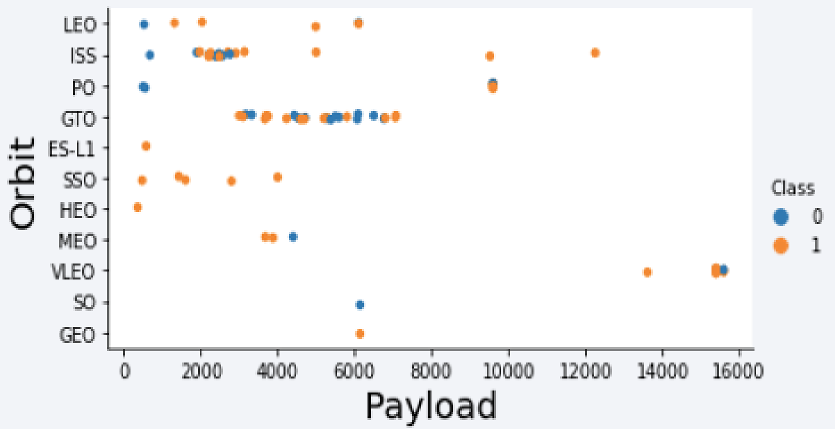
Rest

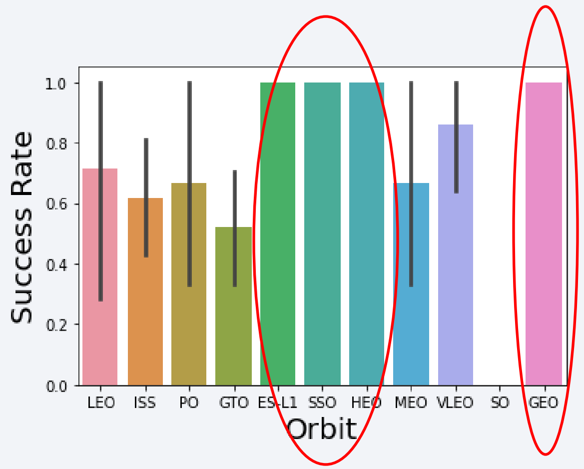
Class

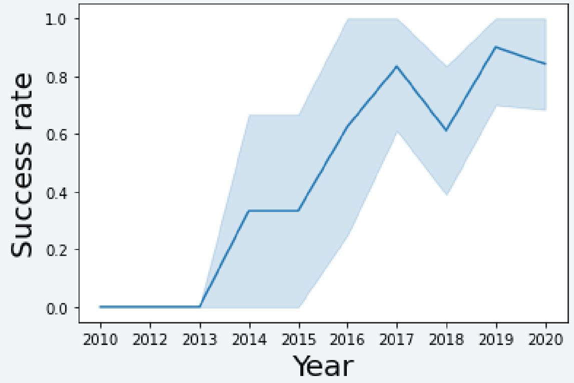
0	30
1	60

EDA with Data Visualization

- Plots of 2 variables each, to see their relationship
  - Flight number vs Launch Site
  - Payload vs Launch Site
  - Success Rate vs Orbit type
  - Flight number vs Orbit type
  - Payload vs Orbit type
  - Success rate yearly trend







- Features with an influence on outcome are kept for the model

EDA with SQL

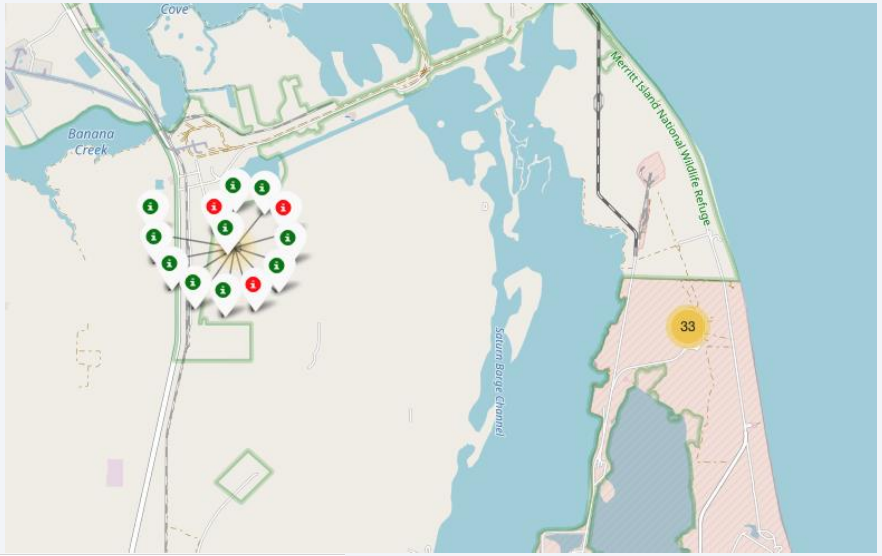
- SQL queries
  - Names of unique launch sites
  - Launch sites beginning with the string 'KSC'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date where the first succesful landing outcome in drone ship was achieved
  - names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - total number of successful and failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass
  - List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

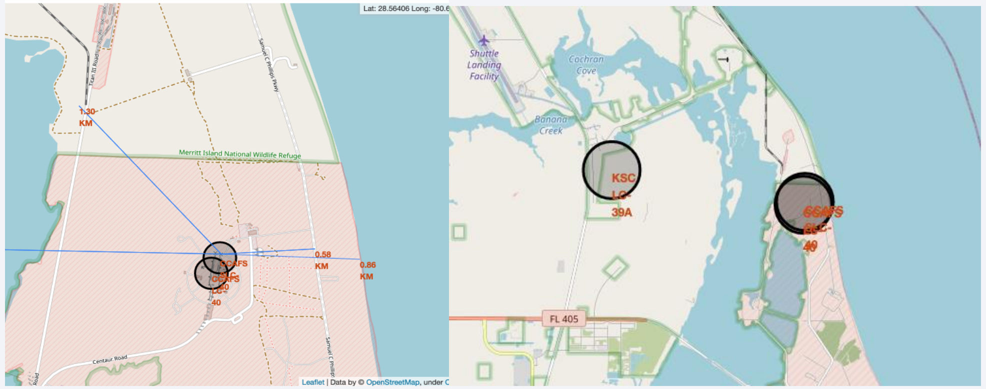
Build an Interactive Map with Folium

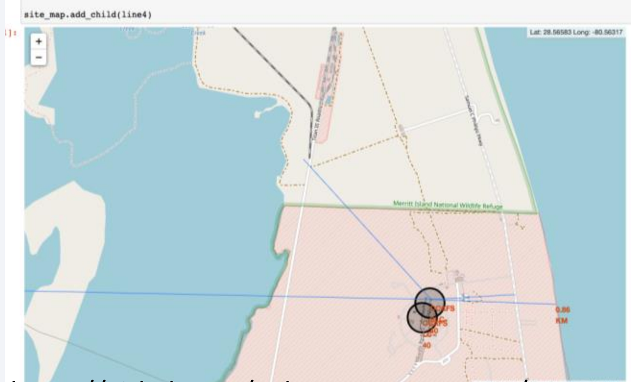
- Marked all launch sites on a map
  - Markers and circles were used
- Marked the success/failed launches for each site on the map
  - Markers and marker_cluster used as they are many in the same place

between a
nities

know POIs





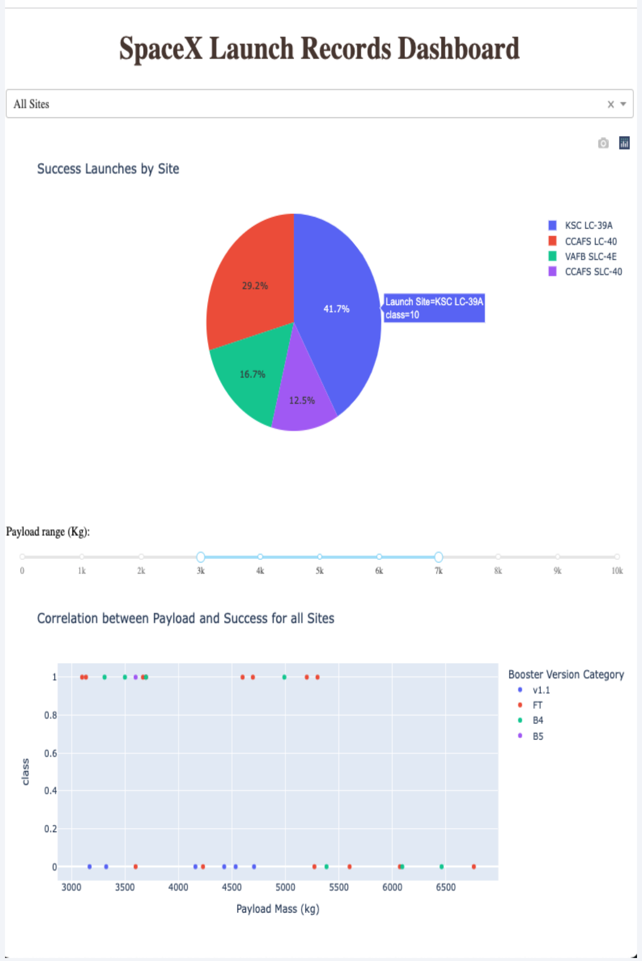


https://github.com/RobertoSanzCapstone/testrepo/blob/main/Capstone%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Landing Success pie chart is plotted choosing a specific site or all of them from a selectable dropdown list
- Landing Success scatter plot is drawn with selected site and filtered by those launches within a certain payload range selected from a interactive slide bar
- This interactive charts reveal the importance of Payload and launch site for a landing success
- GitHub URL

https://github.com/RobertoSanzCapstone/testrepo/blob/main/Capstone%20SpaceX%20Dash.ipynb

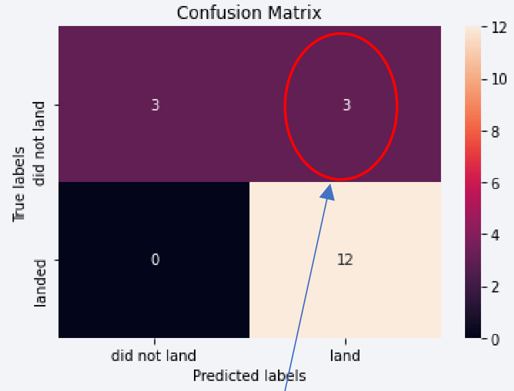


Predictive Analysis (Classification)

- The target (independent) variable to be predicted needs to be extracted
- The features used to predict (dependent variables) need to be split in 2 sets: training set and testing set, to get real accuracy of models
- Several models are trained and their parameters (hyperparameters) tuned to get the best accuracy possible and thus the best model to be used for prediction
- With the chosen model and parameters, a prediction for a new launch can be done
- GitHub URL

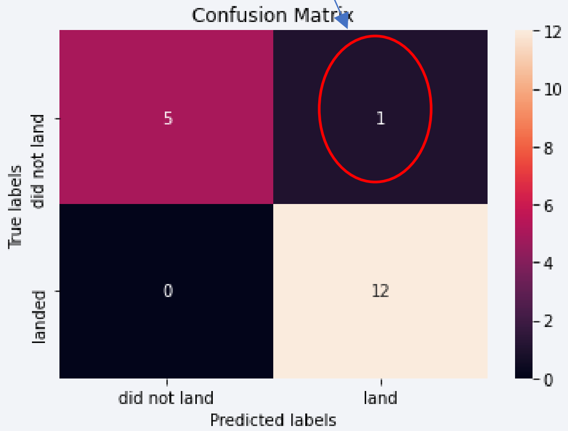
Results

- EDA revealed what features were important for the outcome
- Interactive EDA stressed the importance of 'Payload' and 'Launch Site' as variables for a landing success outcome
- Prediction models were trained and tuned with data from relevant features to be able to make a prediction with an expected accuracy above 83% all of them
- All models showed an expected accuracy of 83%, except 'Decision Tree' with 94%



Rest of models

False positives

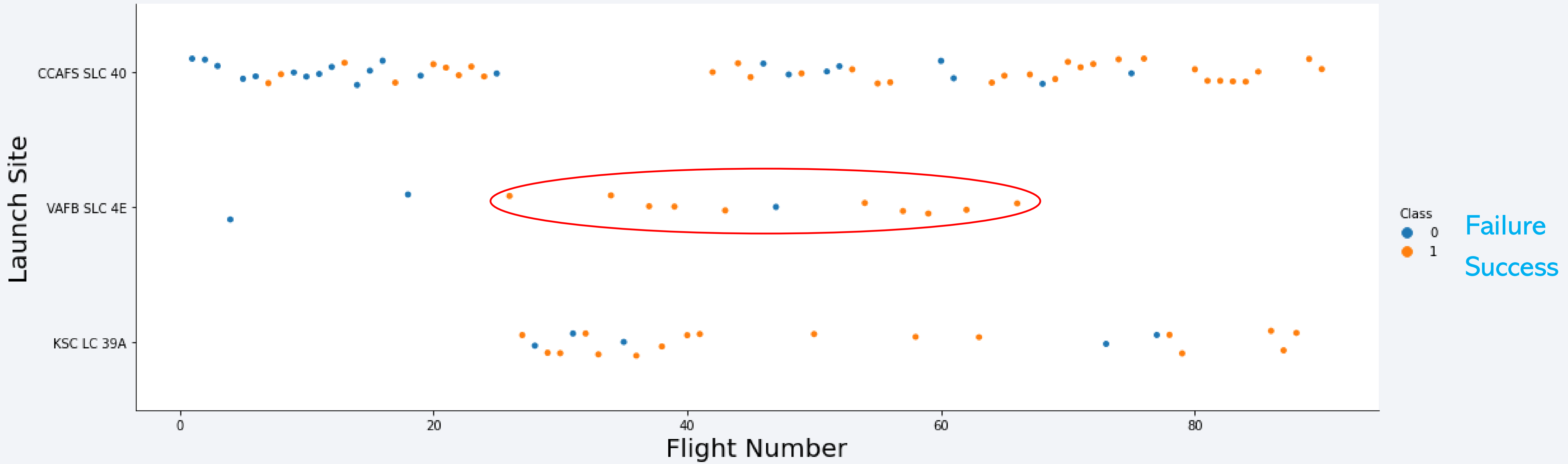


Decision Tree

Section 2

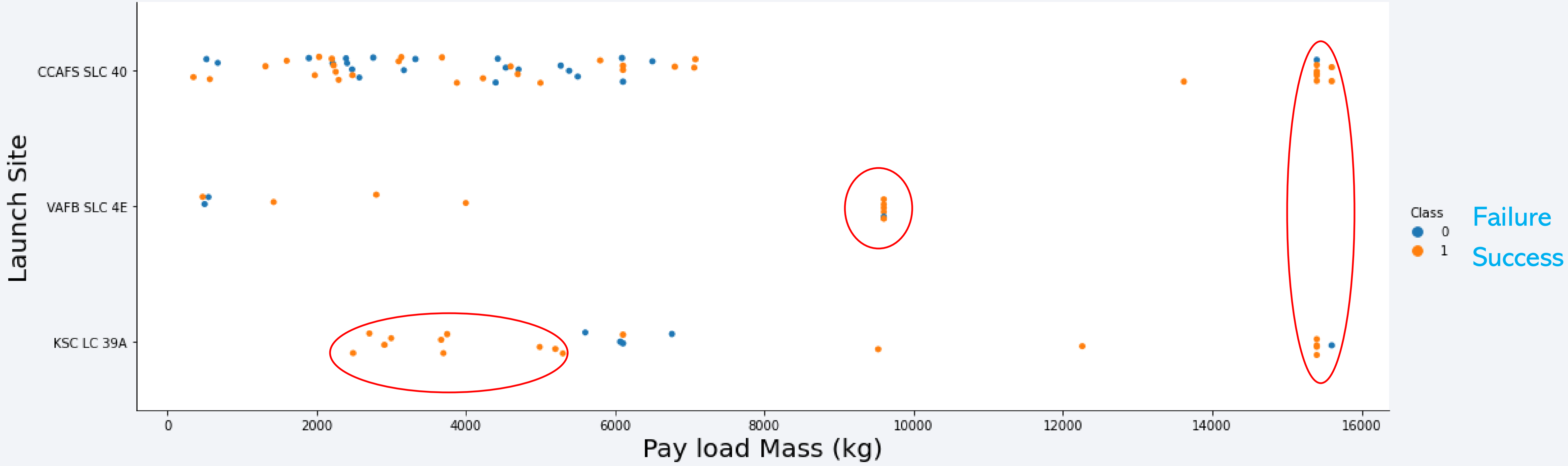
Insights drawn from EDA

Flight Number vs. Launch Site



- VAFB SLC 4E showing less number of flights but quite successully

Payload vs. Launch Site

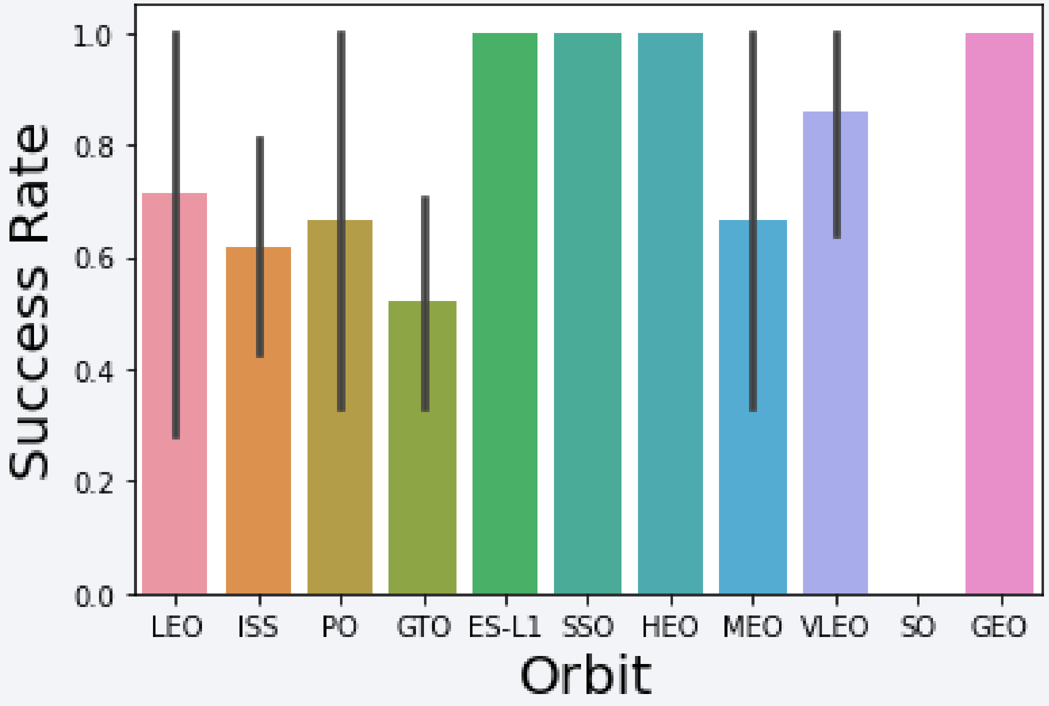


- KSC LC 39A is the best site for smaller payloads and VAFB SLC 4E good in mid range
- CCAFS SLC 40 and KSC LC 39A, both are good for the biggest payloads

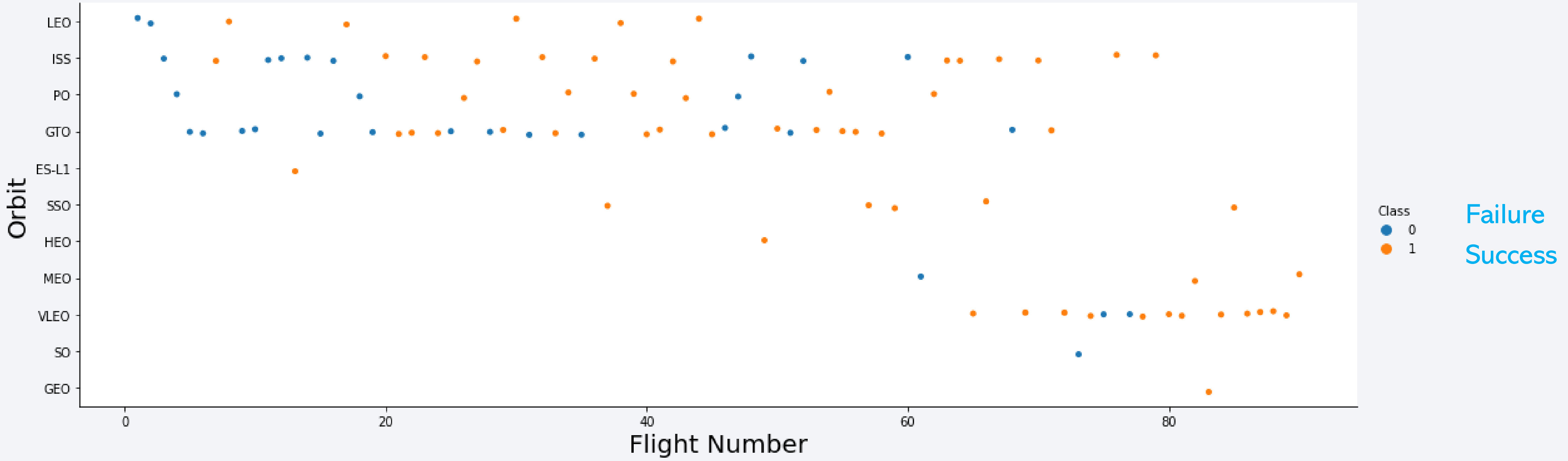
Success Rate vs. Orbit Type

- 4 Orbits have had a perfect performance, no failures

- The other orbits show great variability

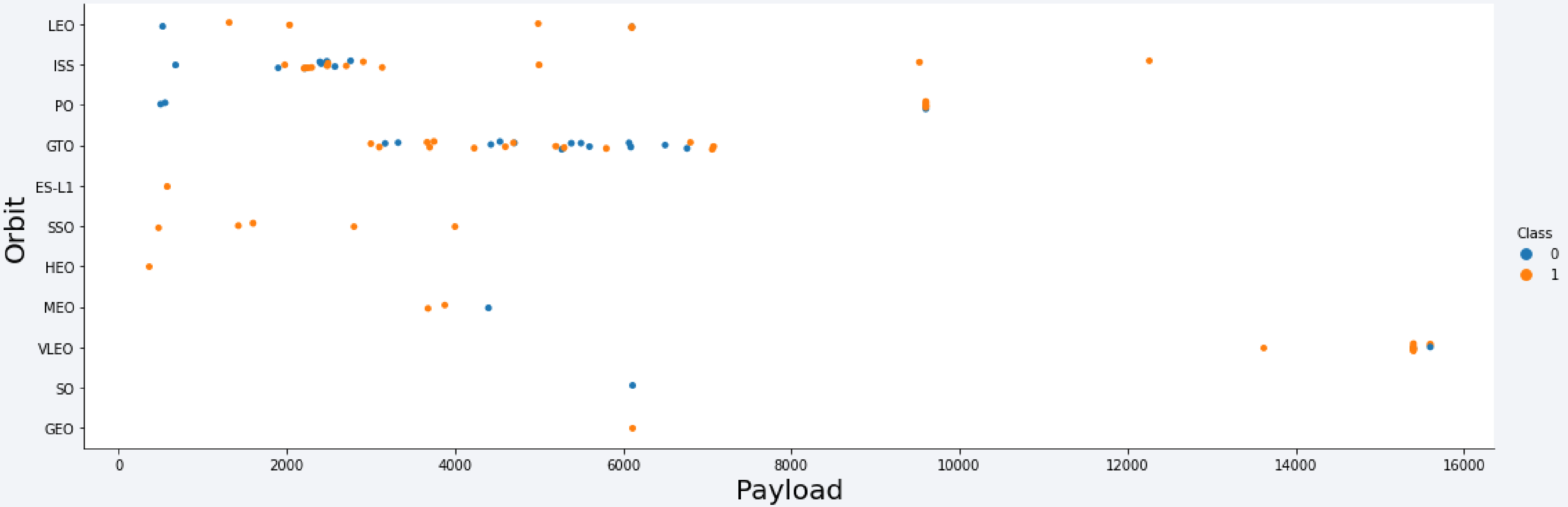


Flight Number vs. Orbit Type



- In 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

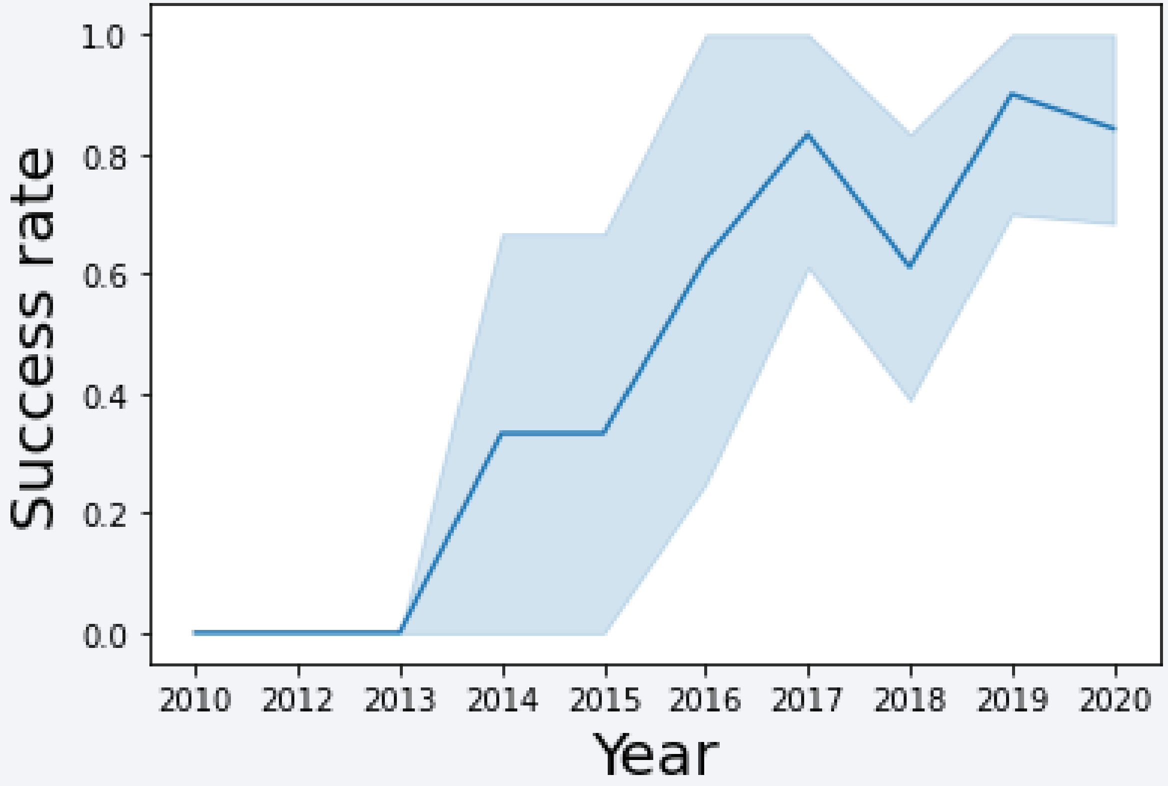


- VLEO, ISS and PO handle the heaviest payloads with good success
- Lightest payloads are best handled in SSO and LEO orbits

Launch Success Yearly Trend

- Landings have been increasing with the years

- Gradual improvement with time is clear



All Launch Site Names

- Only 4 sites used

['CCAFS LC-40',
'VAFB SLC-4E',
'KSC LC-39A',
'CCAFS SLC-40']

Launch Site Names Begin with 'KSC'

- Records where launch sites’ names start with `KSC`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	
29	19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
30	16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
31	30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
32	01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
33	15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA

45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

2928.4

First Successful Ground Landing Date

- First successful landing outcome on ground pad

08-04-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

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

'F9 FT B1032.1',
'F9 B4 B1040.1',
'F9 B4 B1043.1'

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

- Successful 100

- Failure 1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

'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

- Records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Months	Landing _Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

Date	Time (UTC)	Booster_ Version	Launch_S ite	Payload	PAYLO AD_M ASS__ KG_	Orbit	Customer	Mission_ Outcome	Landing _Outcome	new_date
03-06-2017	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)	2017-03-06
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)	2017-02-19
14-01-2017	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)	2017-01-14
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)	2017-01-05
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)	2016-08-14
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)	2016-08-04
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)	2016-07-18
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)	2016-06-05
27-05-2016	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)	2016-05-27
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11	2034	LEO	Orbcomm	Success	Success (ground pad)	2015-12-22

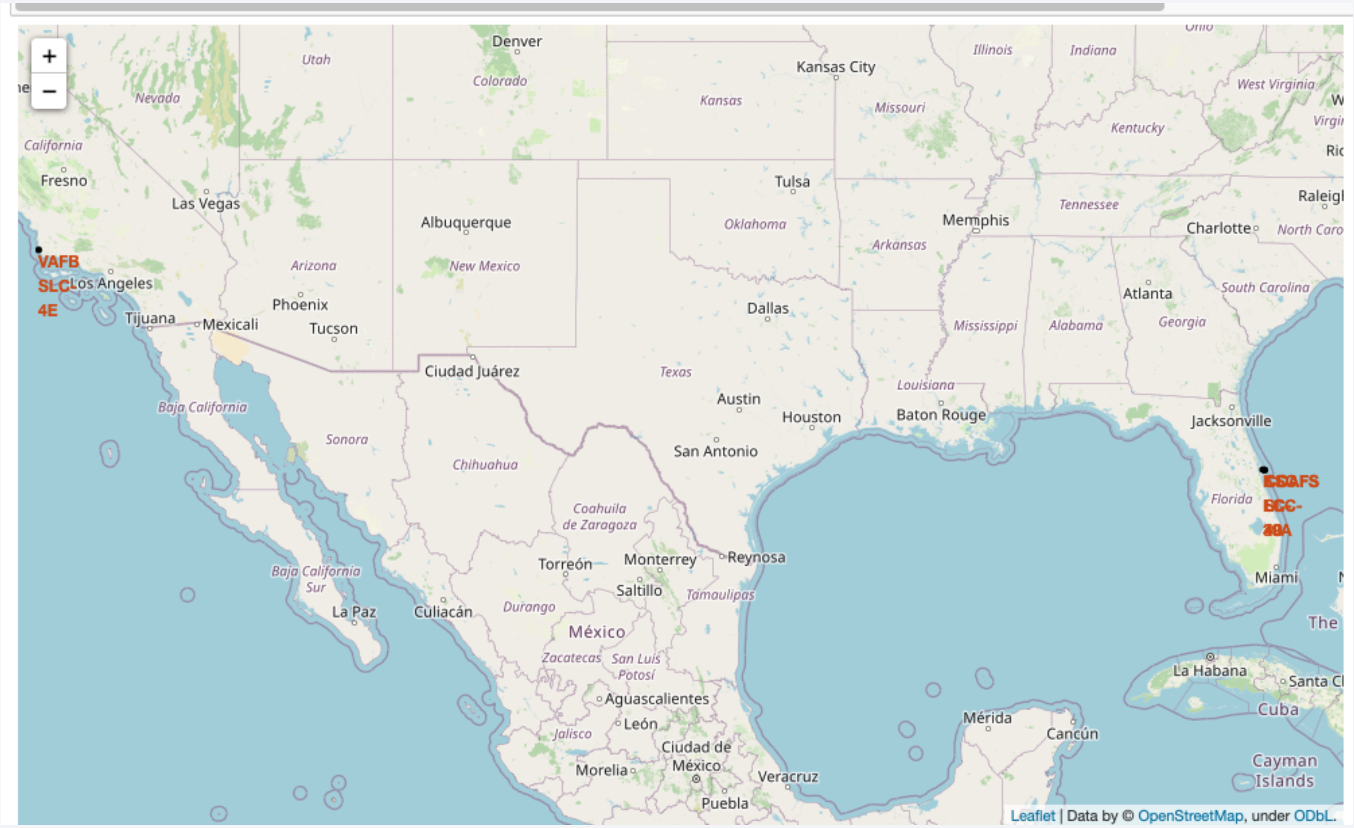
- Ranking of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

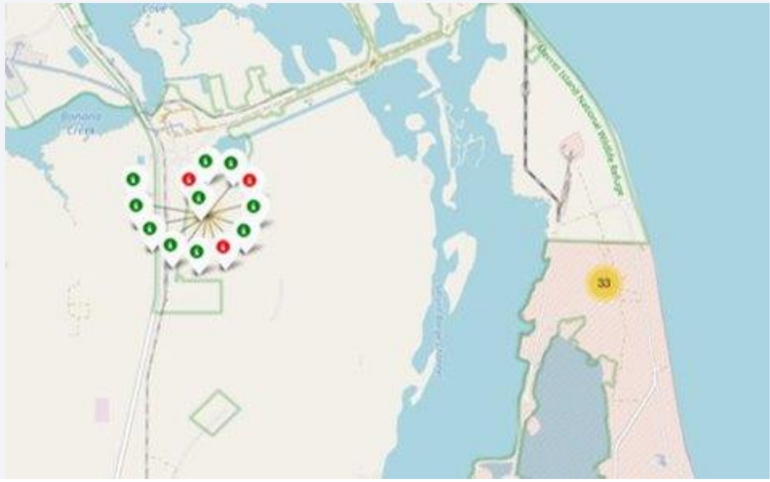
All Launch Sites on a Map

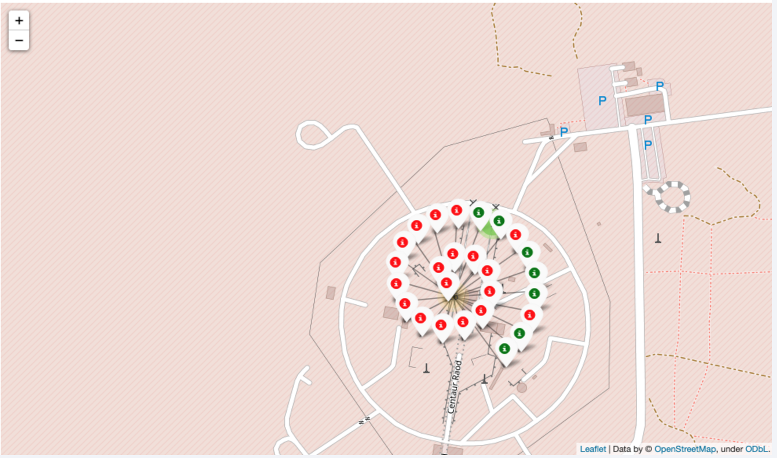
- Launch Sites are located at the most eastern and western locations in the South of USA
- 3 of them in the East and 1 of them in the west
- People Safety must be the reason for seaside locations for launch



Launch Outcomes per Site

- In green the success the successful outcomes
- In red the failures
- KSC on the top picture shows good success ratio whereas CCAFS on the bottom fails most of the time

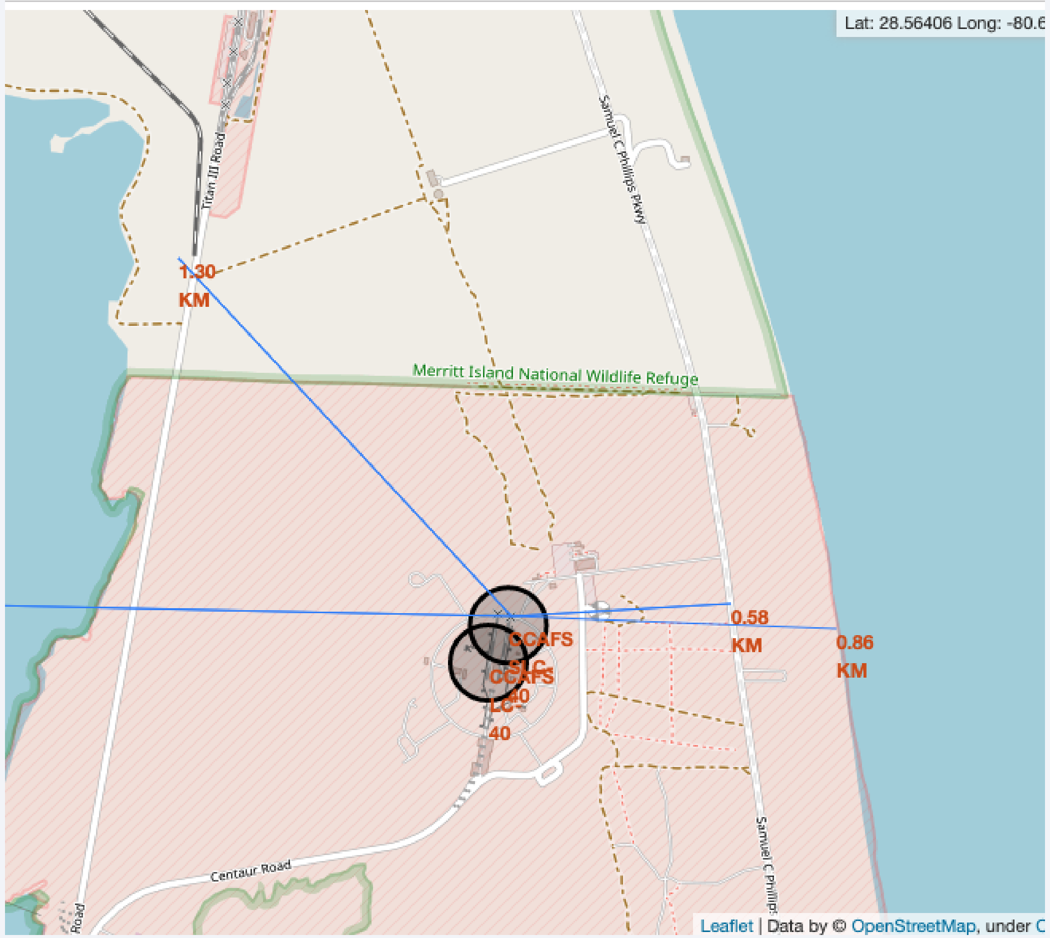




Distance to the closest Points of Interest

- 0.58 Km to the closest highway
- 0.86 Km to the sea
- 1.30 Km to the closest railway
- 21.78 Km to the closest city

All marked on the map

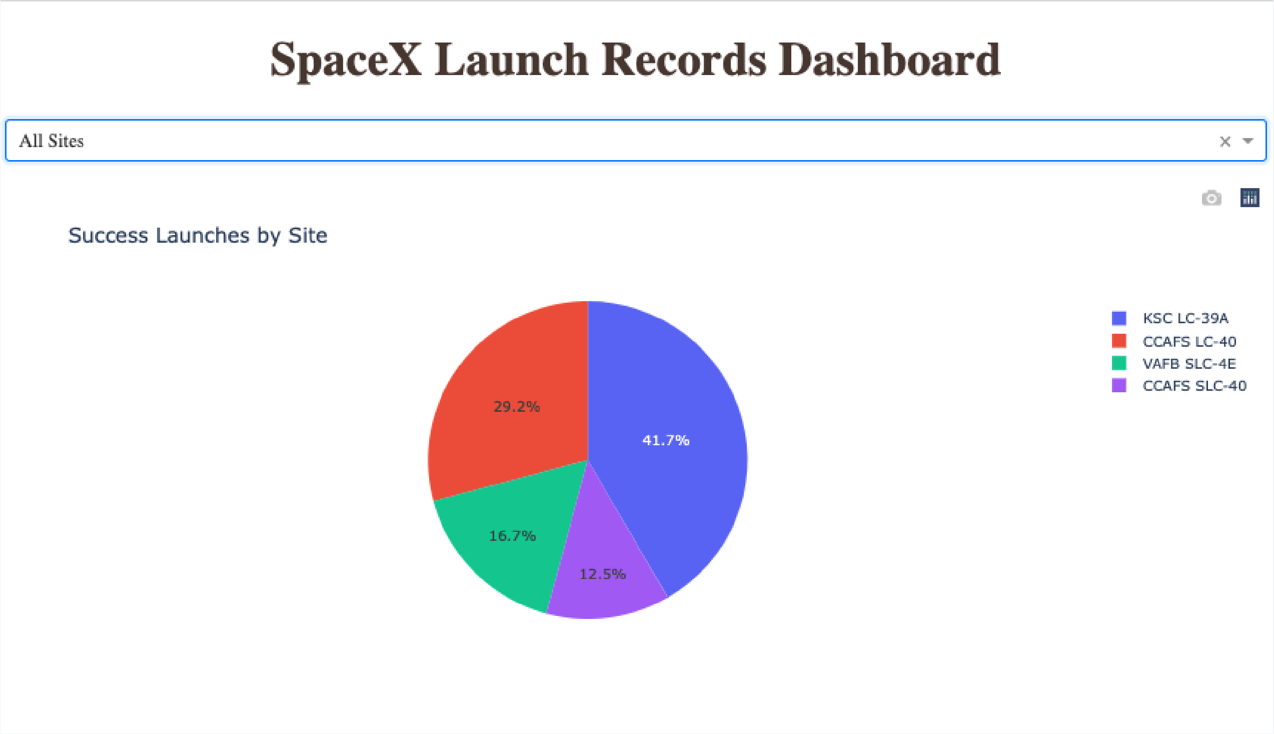


Section 4

Build a Dashboard with Plotly Dash

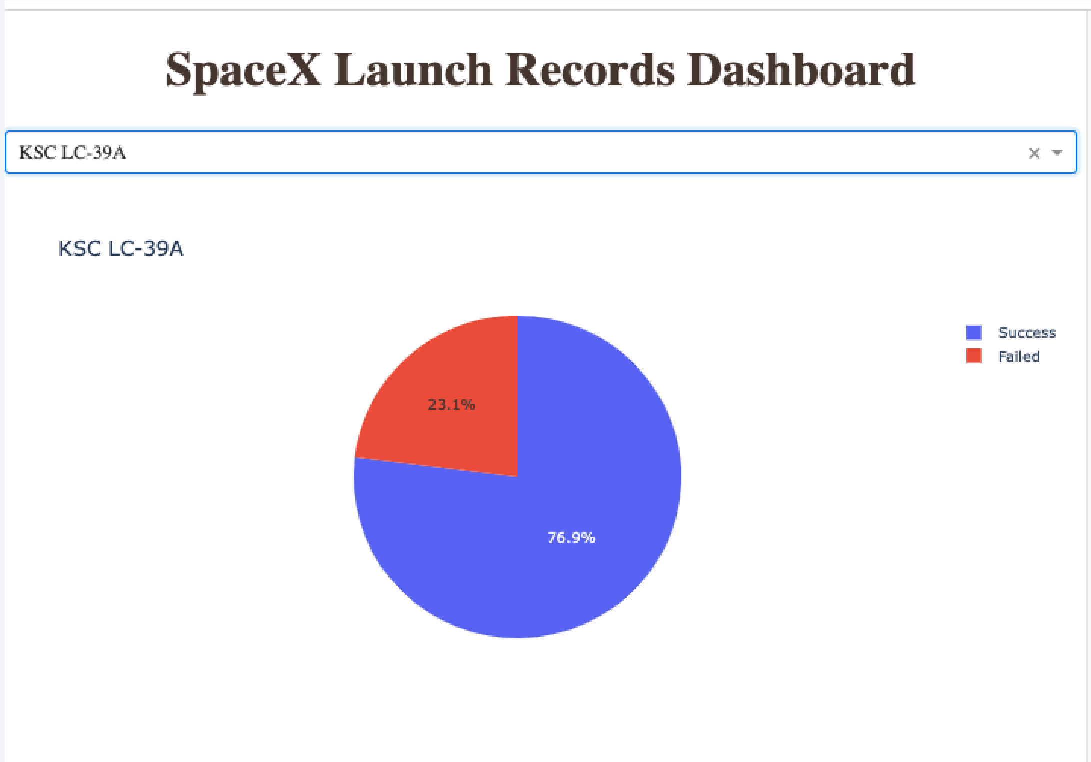
Launch Success for all sites

- Most of the successful launches have been made from KSC site
- The second most successful site was CCAFS LC-40



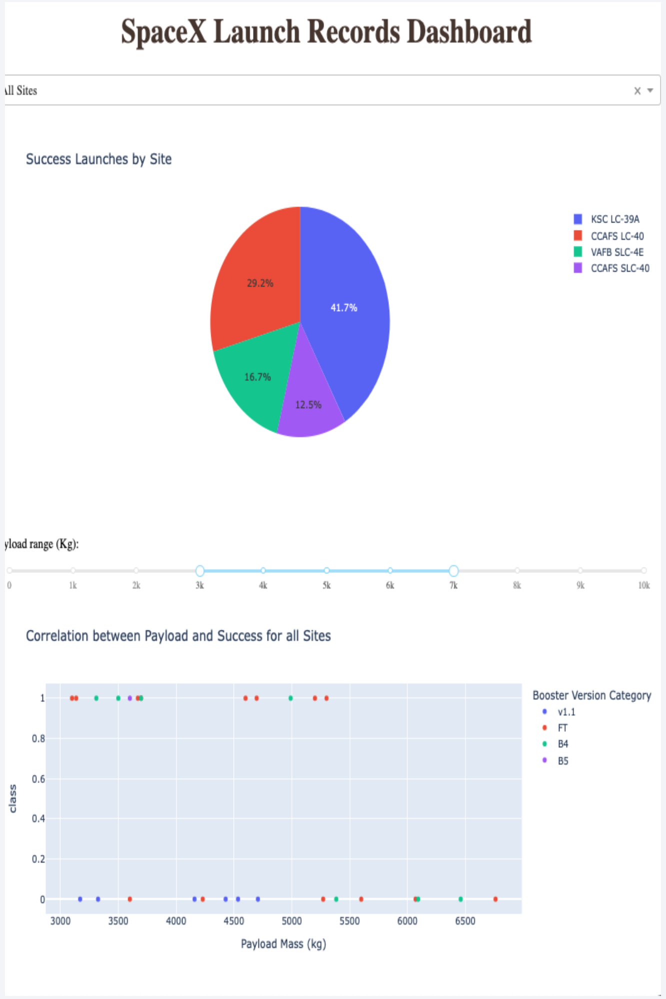
Success pie chart of the best site

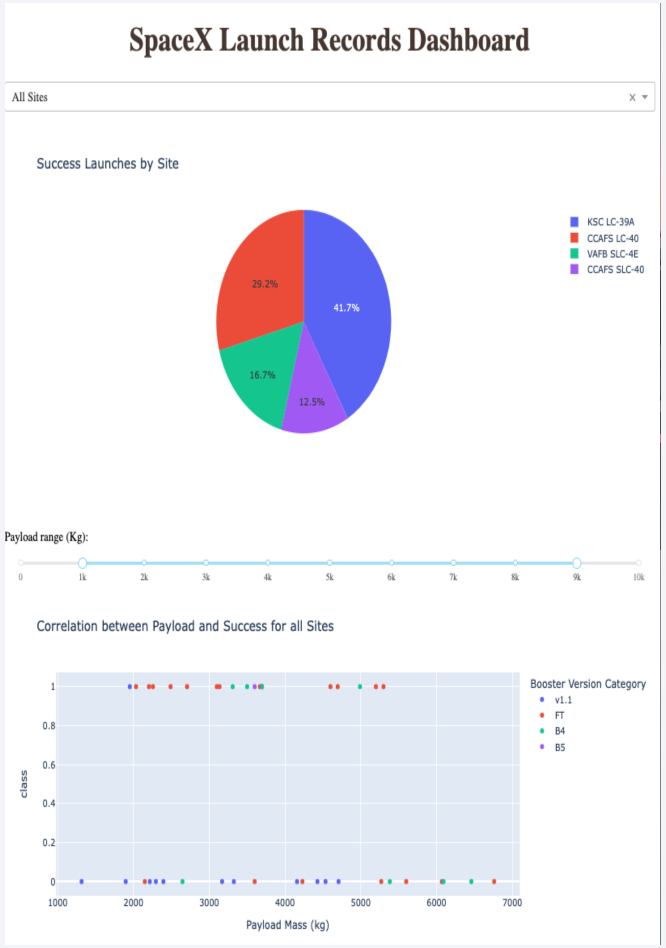
- Launch Success of 76.9% for the best performing site which is KSC



Success per Site and Payload Range selection

- Payload from 1k to 9k on the right picture
- Payload from 3k to 7k on the left picture
- Color of dots means Booster type
- Boosyter v1.1 is failing regardless the payload
- Booster FT and B4 are successful with small payloads and the opposite with big ones



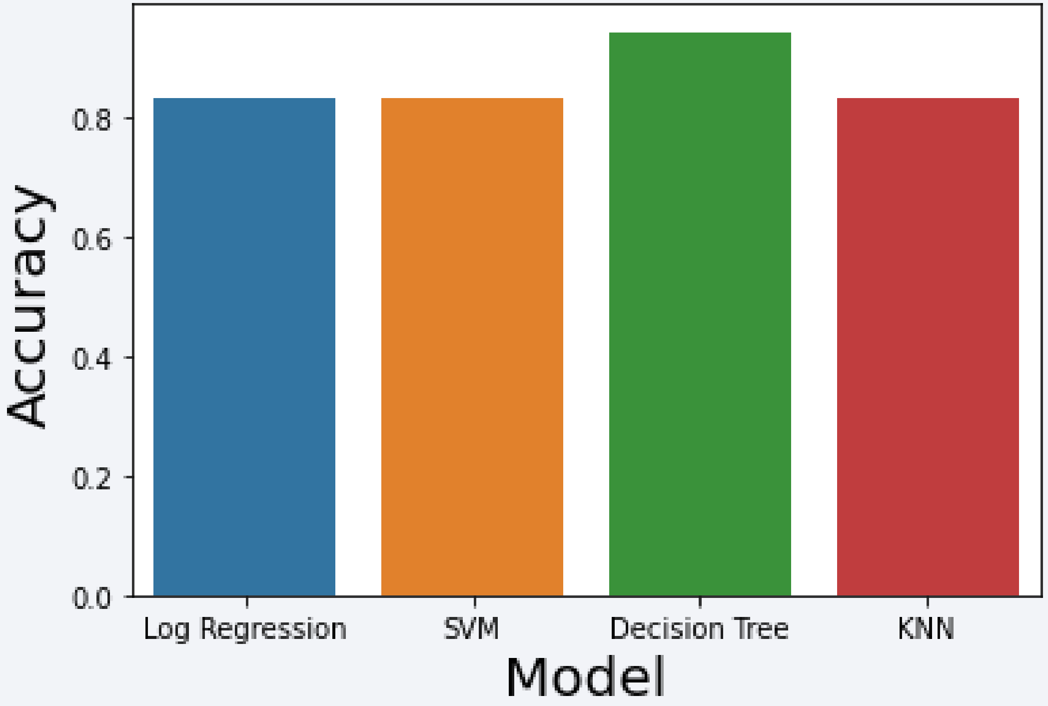


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Accuracy for all prediction models
- Decision Tree is the most accurate with 94%



Confusion Matrix

- 17 in 18 test cases were successfully predicted – 94%
  - 12 predicted as landing and landed
  - 5 predicted as not landing and did not land
- 1 in 18 was predicted as a landing and did not land – this is a False positive

Conclusions

- We can easily collect data from Space X past launches either from Wikipedia or Space X server itself using an API
- Analysis and selection of relevant information can be done using Python libraries and Pandas data structure 'Dataframe'
- Machine Learning models are available in Sci-kit Library to predict the outcome of a future launch, either with a successful landing or a failure
- That would enable the company to calculate the bidding price and to be competitive
- A model was found and tuned to achieve a 94% accuracy in a future prediction

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!