IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ruslan Abdulin
January 16, 2023

Outline

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

Executive Summary

- We collected data from SpaceX API and by scraping SpaceX Wikipedia page with BeautifulSoup package.
- We performed data wrangling and created a landing outcome label 'Class' from Outcome column (classification variable).
- We perform exploratory data analysis (EDA) using visualization and SQL.
- We performed interactive visual analytics using Folium and Plotly Dash.
- We performed predictive analysis using classification models
- Standardized the data, split it into training and testing data, trained different models (logistic regression, SVM, decision tree, K nearest neighbor) and selected hyperparameters.
- We calculated the accuracy on the test data for each of them to find which method performs the best.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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. I will train a machine learning model and use public information to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using SpaceX API and Web scraping
- Perform data wrangling
  - Created a landing outcome label from Outcome column (classification variable)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized the data, split it into training and testing data, trained different models (logistic regression, SVM, decision tree, K nearest neighbor) and selected hyperparameters. Next, calculated the accuracy on the test data for each of them to find which method performs the best

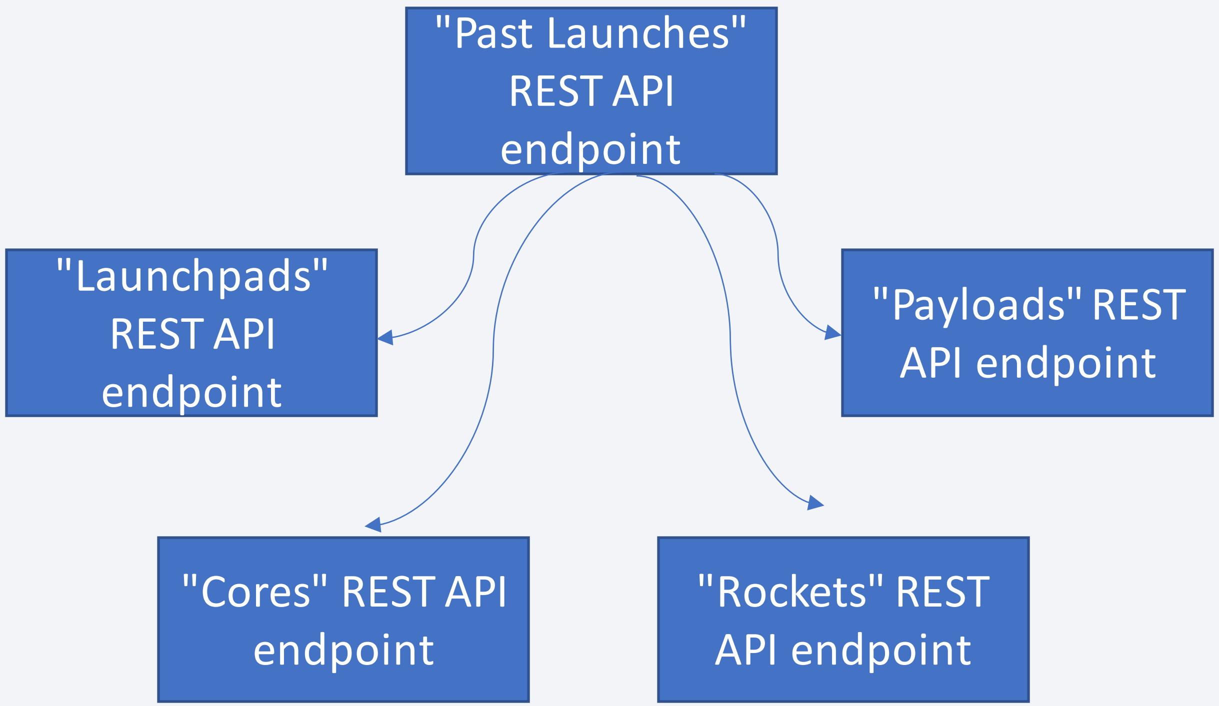
Data Collection

- I've worked with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Another data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. I've used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.

Data Collection – SpaceX API

- I performed a get request using the requests library to obtain the launch data, which is used to get the data from the API. Also, I used other API endpoints to gather specific data for each ID number.

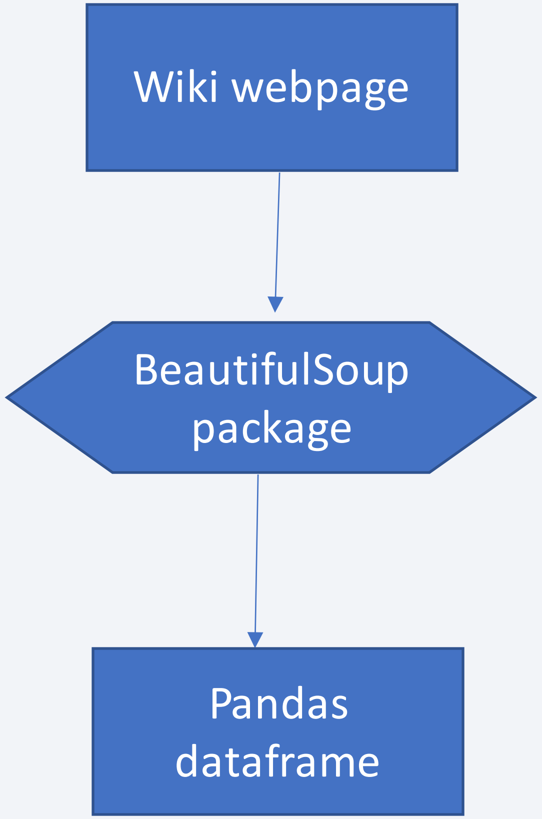
- [Data Collection API lab on GitHub]()



Data Collection - Scraping

- I used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records. Then I parsed the data from those tables and convert them into a Pandas data frame for further visualization and analysis

- [Data Collection with Web Scraping lab on GitHub]



Data Wrangling

- I calculated the number of launches on each site. Then I calculated the number and occurrence of each orbit and mission outcome per orbit type. Finally, I've created a landing outcome label from Outcome column.
- EDA lab on GitHub

EDA with Data Visualization

- I plotted 7 graphs:

1. Scatterplot FlightNumber vs. PayloadMass (to see how Payload mass would affect the launch outcome)
2. Scatterplot FlightNumber vs. Launch Site (to see if Launch site would affect the launch outcome)
3. Scatterplot PayloadMass vs. Launch Site (to see if there is any relationship between launch sites and the payload mass)
4. Bar chart Success Rate for each Orbit
5. Scatterplot FlightNumber vs. Orbit (to see how Orbit type would affect the launch outcome)
6. Scatterplot PayloadMass vs. Orbit (to see if there is any relationship between Orbit type and the payload mass)
7. Line chart Average Launch Success Trend

EDA with Data Visualization lab on GitHub

EDA with SQL

- I performed 10 SQL queries:

1. SELECT DISTINCT to display unique values
2. WHERE clause + LIMIT 5 to display 5 records based on condition
3. SUM to aggregate integer values
4. AVG to calculate an average
5. MIN to find the earliest date
6. SELECT DISTINCT + WHERE clause to display unique records based on condition
7. COUNT + GROUP BY to calculate a total number for each mission outcome
8. Subquery with MAX to filter out only maximum payload mass
9. Substr and WHERE clause to parse the date based on condition
10. RANK OVER PARTITION BY to rank successful landing_outcomes

- [EDA with SQL lab on GitHub]

Build an Interactive Map with Folium

- I've created the following map objects:

1. Circle and Marker to add a highlighted circle area with a text label on a specific coordinate for each Launch site

2. MarkerCluster to simplify a map containing many markers having the same coordinate

3. MousePosition to get coordinate for a mouse over a point on the map

4. Marker to display a distance between 2 coordinates

5. PolyLine to draw a line

- [Interactive Visual Analytics with Folium lab on GitHub]()

Build a Dashboard with Plotly Dash

- I added:

1. Launch Site Drop-down Input Component to let us select different launch sites
2. Callback function to render success-pie-chart based on selected site dropdown
3. Range Slider to select Payload
4. Callback function to render success-payload-scatter-chart scatter plot

- Build a Dashboard with Plotly Dash lab on GitHub

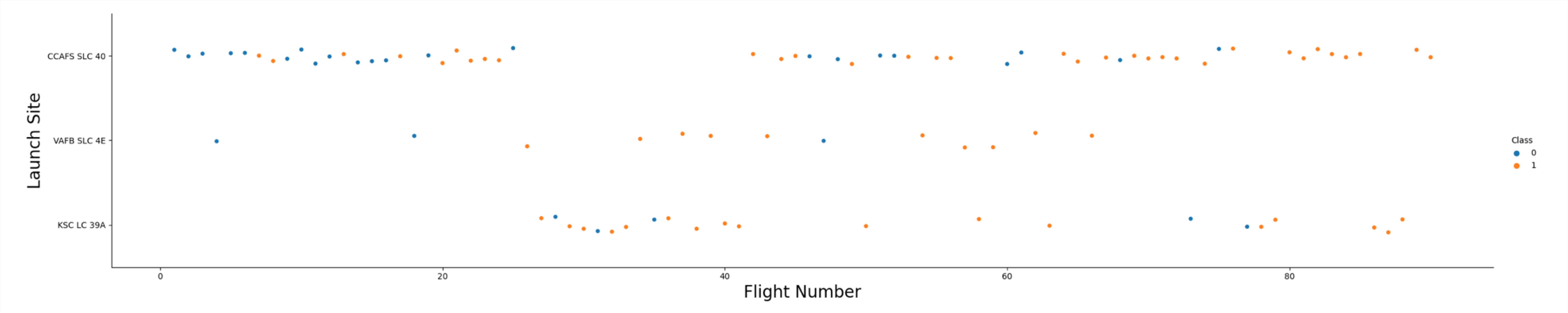
Predictive Analysis (Classification)

- I standardized the data first, then split it into training and testing data. After that, I trained different models (logistic regression, SVM, decision tree, K nearest neighbor) and selected hyperparameters. Next, I calculated the accuracy on the test data for each of them to find which method performs the best.
- Machine Learning Prediction lab on GitHub

Section 2

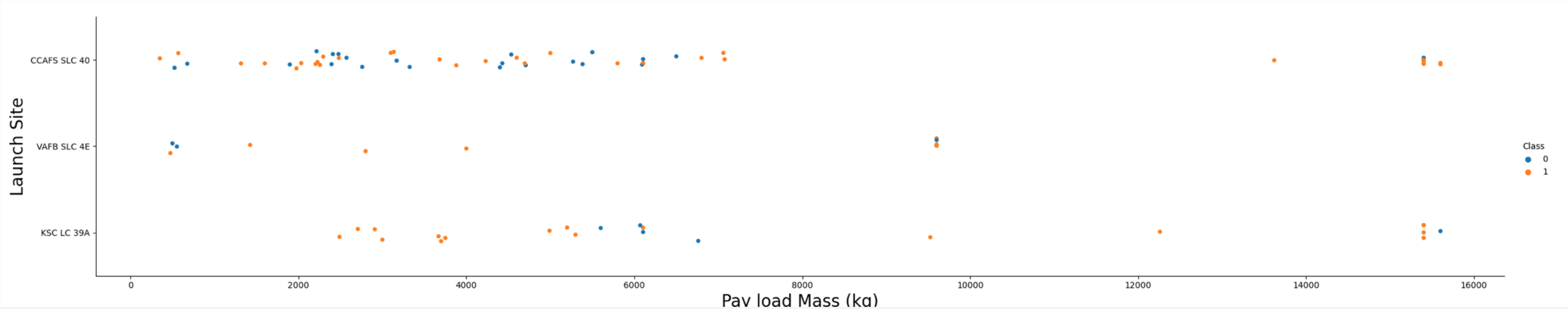
Insights drawn from EDA

Flight Number vs. Launch Site



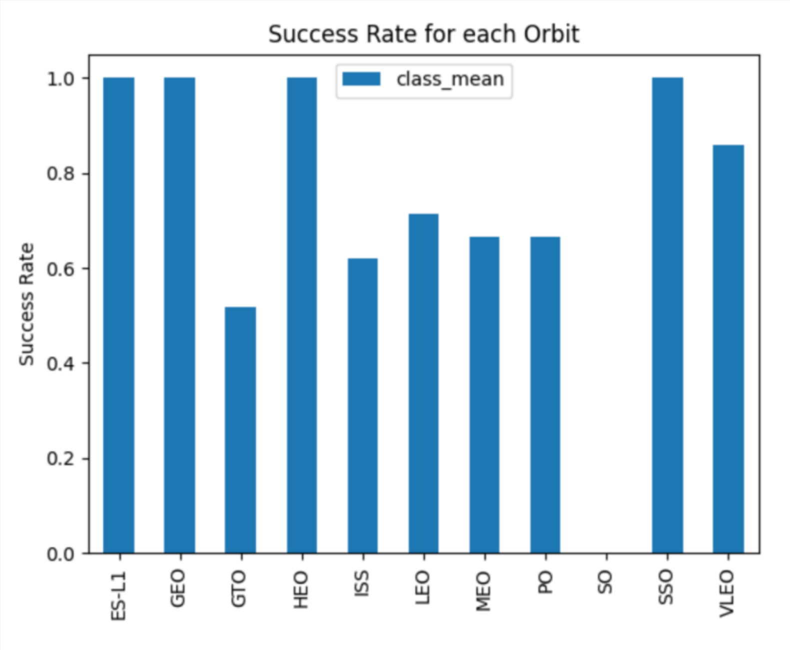
- We can see that the more attempts SpaceX makes the more success launches they get. The largest number of success launches has KSC LC-39A

Payload vs. Launch Site



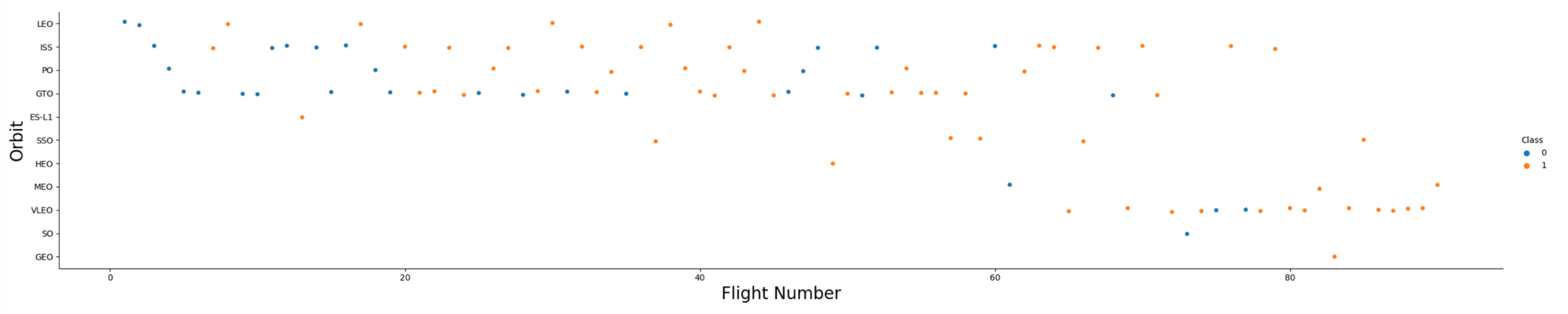
For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



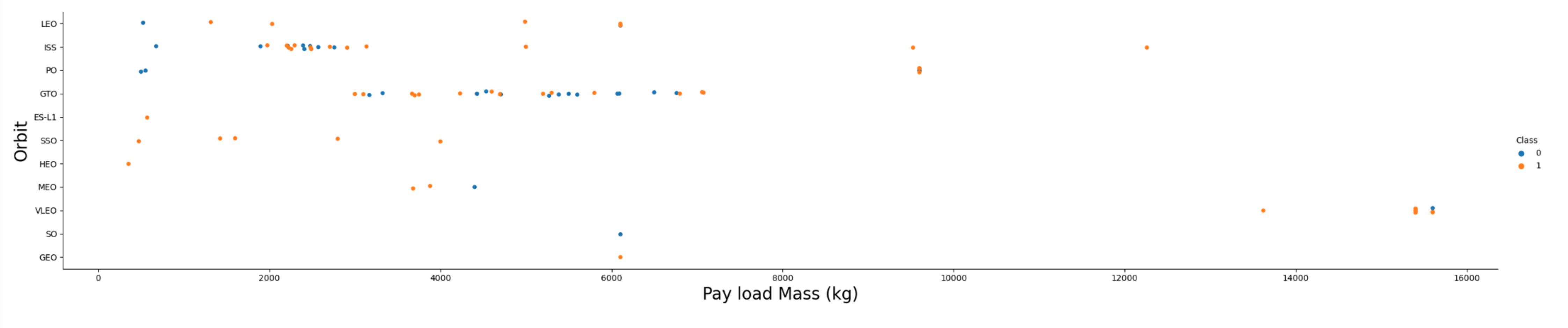
ES-L1, GEO, HEO, SSO are the most successful orbits

Flight Number vs. Orbit Type



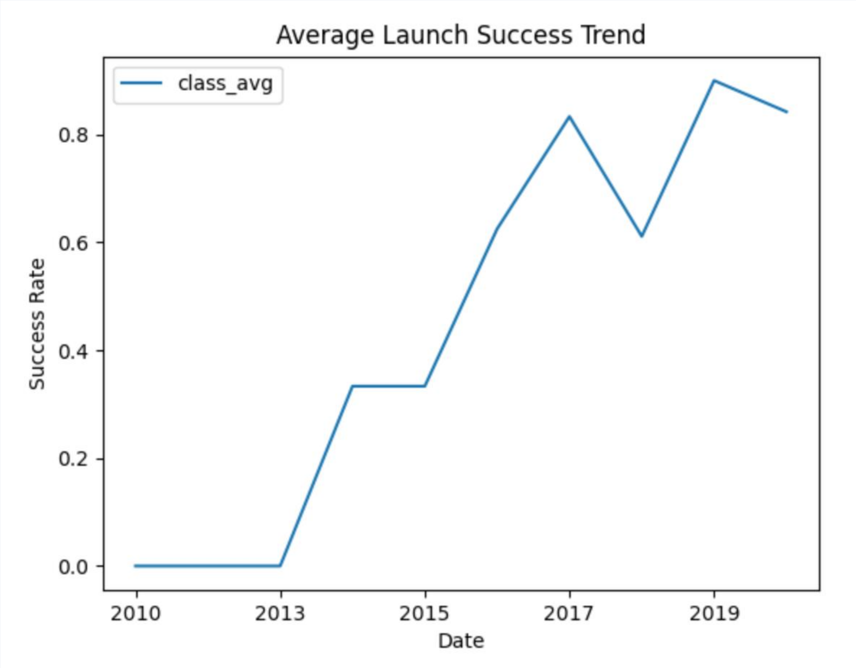
- In LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



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

Launch Success Yearly Trend



We can observe that the success rate kept increasing since 2013 till 2020 with a sharp drop in 2018

All Launch Site Names

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

SELECT DISTINCT (LAUNCH_SITE) FROM SPACEXTBL

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

- SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

Total Payload Mass

TOTAL_PAYLOAD_MASS_KG
45596

SELECT SUM (PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS_KG FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

AVG_PAYLOAD_MASS_KG
2928.4

SELECT AVG (PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1'

First Successful Ground Landing Date

MIN (DATE)
01-05-2017

SELECT MIN (DATE) FROM SPACEXTBL WHERE [LANDING _OUTCOME] = 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

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

SELECT DISTINCT (BOOSTER_VERSION) FROM SPACEXTBL WHERE [LANDING _OUTCOME] = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

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

SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEXTBL GROUP BY MISSION_OUTCOME

Boosters Carried Maximum Payload

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

- SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

2015 Launch Records

month	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

- SELECT substr(Date, 4, 2) as month, [LANDING _OUTCOME], BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE [LANDING _OUTCOME] = 'Failure (drone ship)' AND substr(Date,7,4)='2015'

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

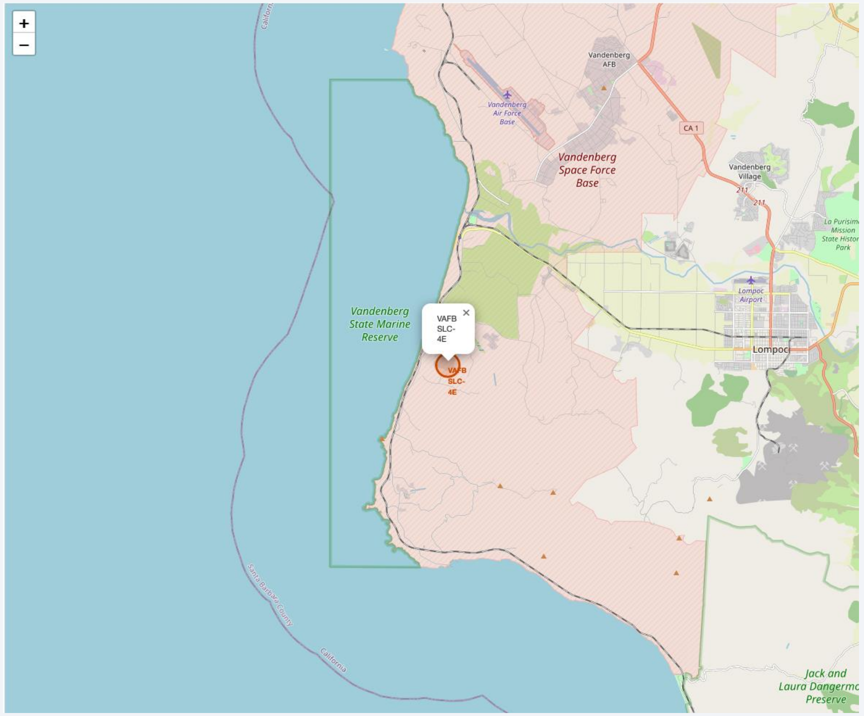
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome	RANK_COUNT
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success	1
18-08-2020	14:31:00	F9 B5 B1049.6	CCAFS SLC-40	Starlink 10 v1.0, SkySat-19, -20, -21, SAOCOM 1B	15440	LEO	SpaceX, Planet Labs, PlanetIQ	Success	Success	2
17-12-2019	00:10:00	F9 B5 B1056.3	CCAFS SLC-40	JCSat-18 / Kacific 1, Starlink 2 v1.0	6956	GTO	Sky Perfect JSAT, Kacific 1	Success	Success	3
16-11-2020	00:27:00	F9 B5B1061.1	KSC LC-39A	Crew-1, Sentinel-6 Michael Freilich	12500	LEO (ISS)	NASA (CCP)	Success	Success	4
15-11-2018	20:46:00	F9 B5 B1047.2	KSC LC-39A	Es hail 2	5300	GTO	Es hailSat	Success	Success	5
13-06-2020	09:21:00	F9 B5 B1059.3	CCAFS SLC-40	Starlink 8 v1.0, SkySats-16, -17, -18, GPS III-03	15410	LEO	SpaceX, Planet Labs	Success	Success	6
12-06-2019	14:17:00	F9 B5 B1051.2	VAFB SLC-4E	RADARSAT Constellation, SpaceX CRS-18	4200	SSO	Canadian Space Agency (CSA)	Success	Success	7
11-11-2019	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success	8
11-01-2019	15:31:00	F9 B5 B1049.2	VAFB SLC-4E	Iridium NEXT-8	9600	Polar LEO	Iridium Communications	Success	Success	9
10-09-2018	04:45:00	F9 B5B1049.1	CCAFS SLC-40	Telstar 18V / Apstar-5C	7060	GTO	Telesat	Success	Success	10
08-10-2018	02:22:00	F9 B5 B1048.2	VAFB SLC-4E	SAOCOM 1A	3000	SSO	CONAE	Success	Success	11
07-08-2020	05:12:00	F9 B5 B1051.5	KSC LC-39A	Starlink 9 v1.0, SXRS-1, Starlink 10 v1.0	14932	LEO	SpaceX, Spaceflight Industries (BlackSky), Planet Labs	Success	Success	12
07-08-2018	05:18:00	F9 B5 B1046.2	CCAFS SLC-40	Merah Putih	5800	GTO	Telkom Indonesia	Success	Success	13
07-03-2020	04:50:00	F9 B5 B1059.2	CCAFS SLC-40	SpaceX CRS-20, Starlink 5 v1.0	1977	LEO (ISS)	NASA (CRS)	Success	Success	14
07-01-2020	02:33:00	F9 B5 B1049.4	CCAFS SLC-40	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600	LEO	SpaceX	Success	Success	15
06-12-2020	16:17:08	F9 B5 B1058.4	KSC LC-39A	SpaceX CRS-21	2972	LEO (ISS)	NASA (CRS)	Success	Success	16
06-10-2020	11:29:34	F9 B5 B1058.3	KSC LC-39A	Starlink 12 v1.0, Starlink 13 v1.0	15600	LEO	SpaceX	Success	Success	17
05-12-2019	17:29:00	F9 B5B1059.1	CCAFS SLC-40	SpaceX CRS-19, JCSat-18 / Kacific 1	2617	LEO (ISS)	NASA (CRS), Kacific 1	Success	Success	18
05-11-2020	23:24:23	F9 B5B1062.1	CCAFS SLC-40	GPS III-04 , Crew-1	4311	MEO	USSF	Success	Success	19
04-06-2020	01:25:00	F9 B5 B1049.5	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success	20

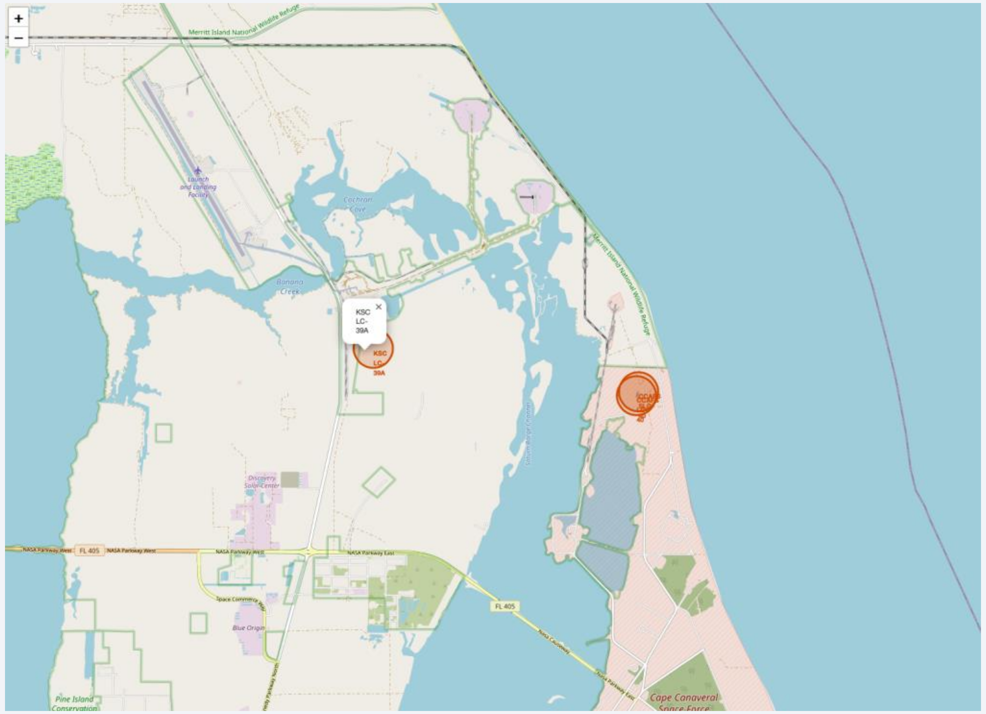
- SELECT *, RANK() OVER (PARTITION BY [LANDING _OUTCOME] ORDER BY DATE DESC) AS RANK_COUNT FROM SPACEXTBL WHERE [LANDING _OUTCOME] LIKE '%success%' AND DATE BETWEEN '04-06-2010' AND '20-03-2017'

Section 3

Launch Sites Proximities Analysis

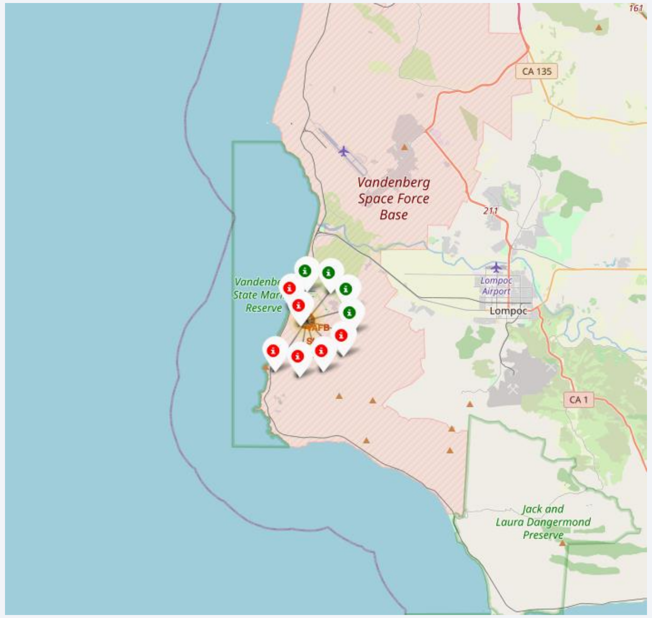
All Launch Sites on a Map

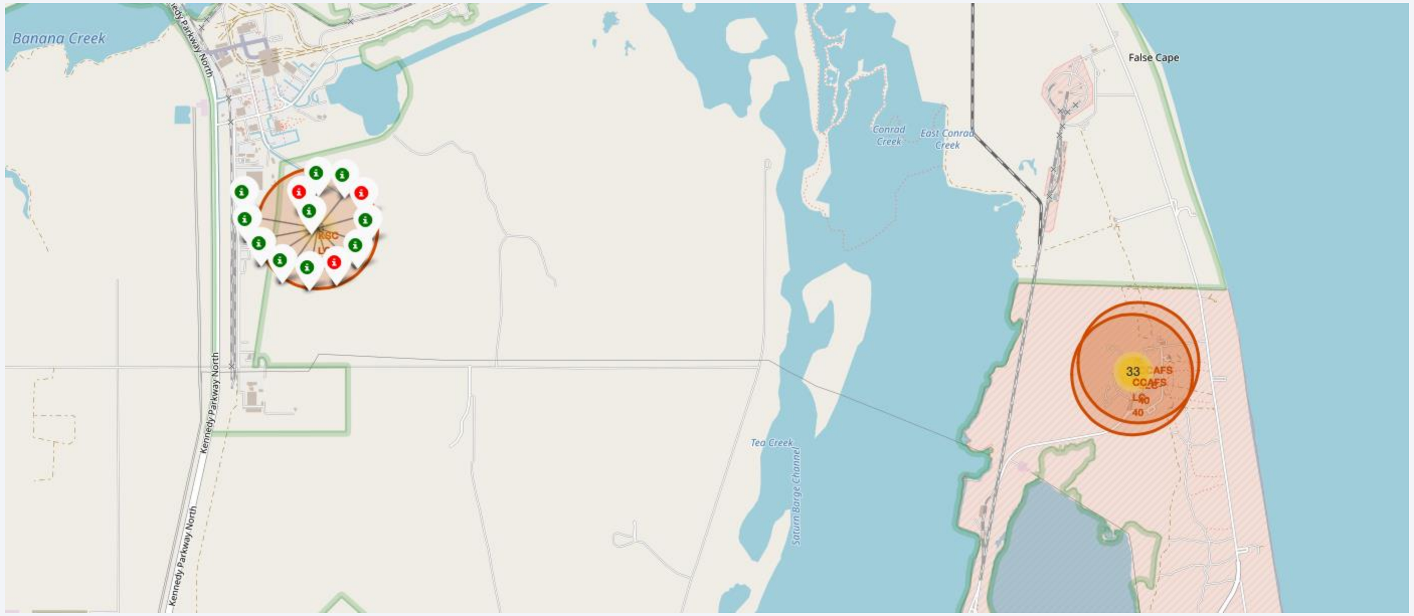




all launch sites in proximity to the Equator line and to the coasts

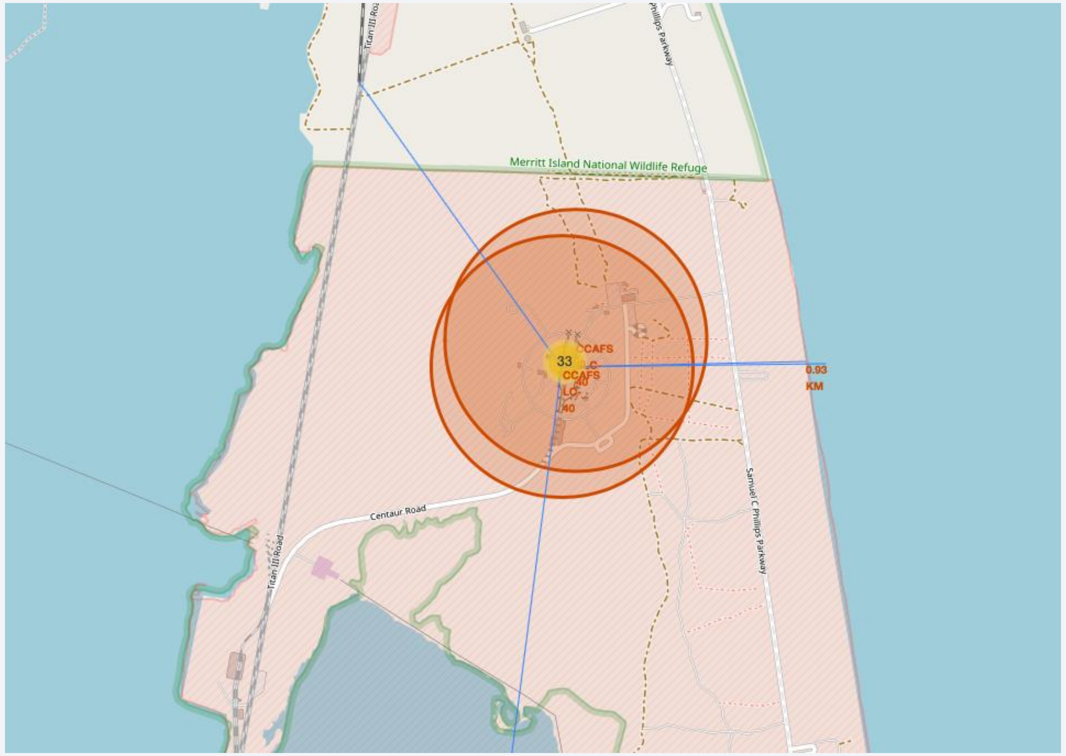
Success/failed launches for each site on the map





We can easily identify that KSCLC-39A has relatively high success rate.

Distances between a launch site and its proximities

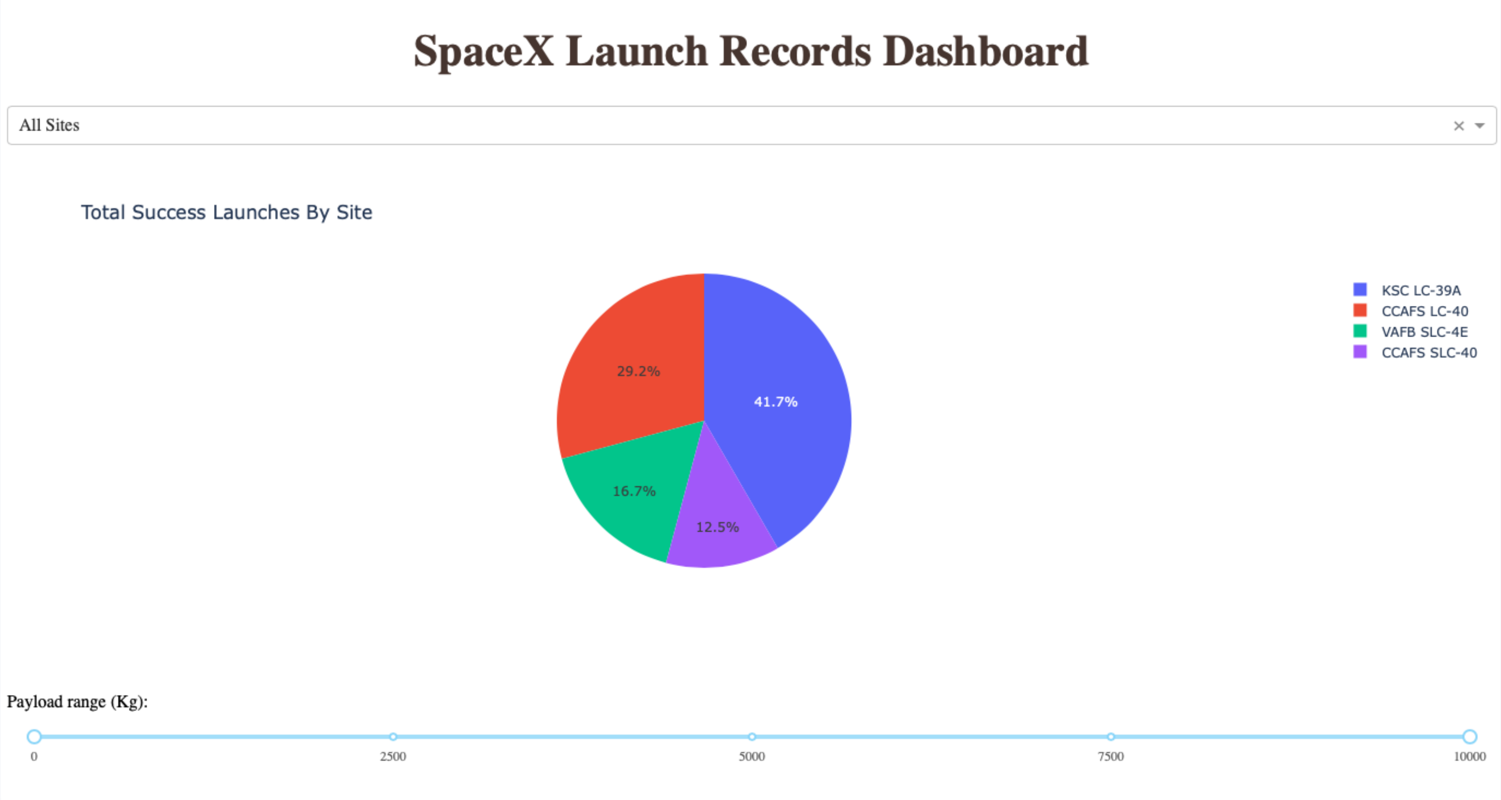


Launch sites are near railways, highways, and a coastline, but far away from cities.

Section 4

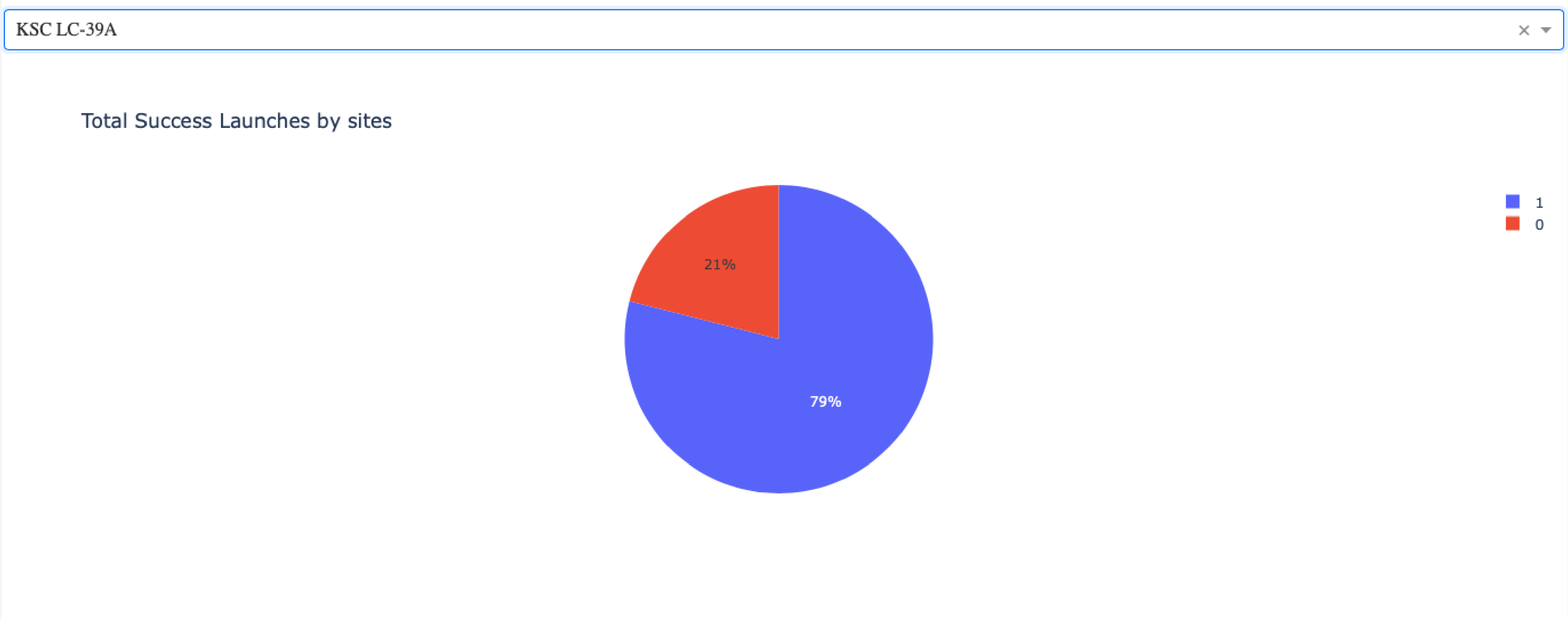
Build a Dashboard with Plotly Dash

Total Success Launches By Site – All Sites



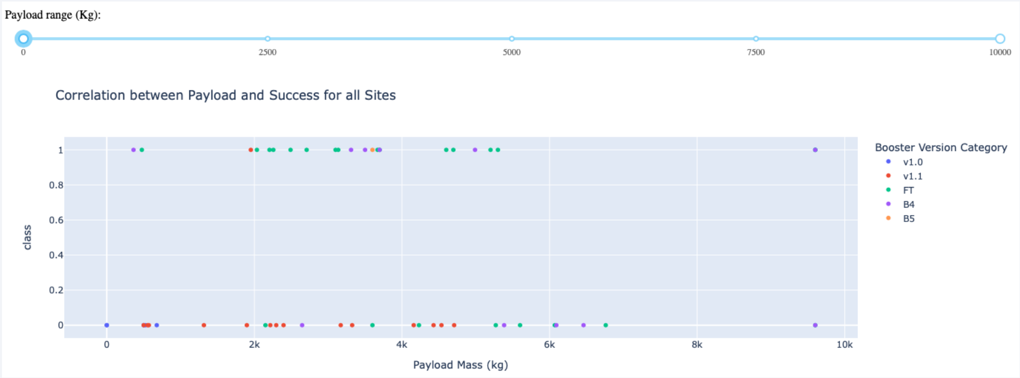
The largest number of success launches has KSC LC-39A

Success Launches ratio for KSC LC-39A



KSC LC-39A has 79% of success launches

Correlation between Payload and Success launch by Booster Version Category



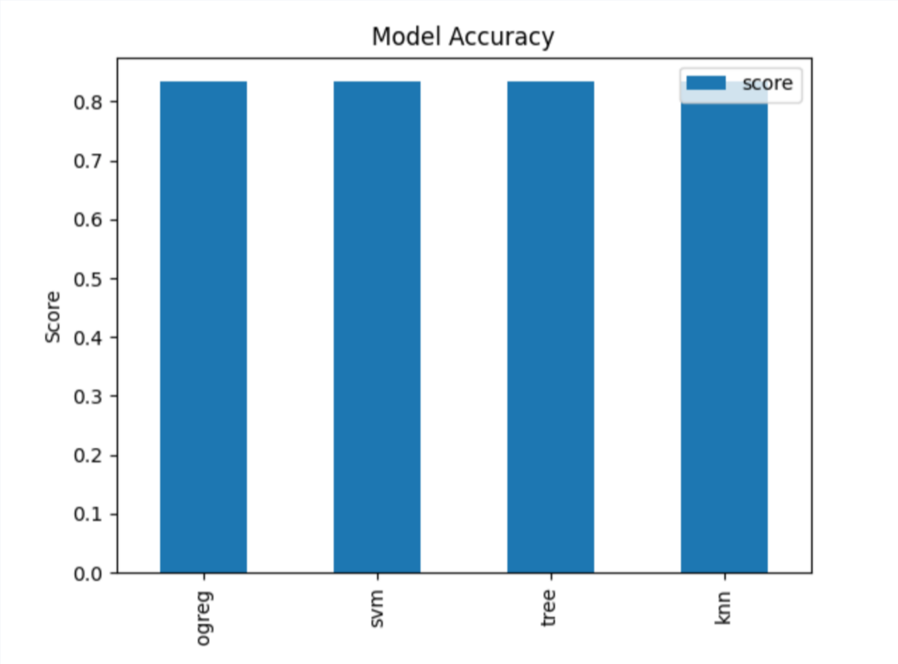
Success Launch varies for different Booster Versions:

1. V1.0 and v1.1 have very low success launches rates
2. FT tends to perform better with lower Payload Mass
3. B4 has more success launches with Payload less than 4000 kg
4. B5 had only 1 launch

Section 5

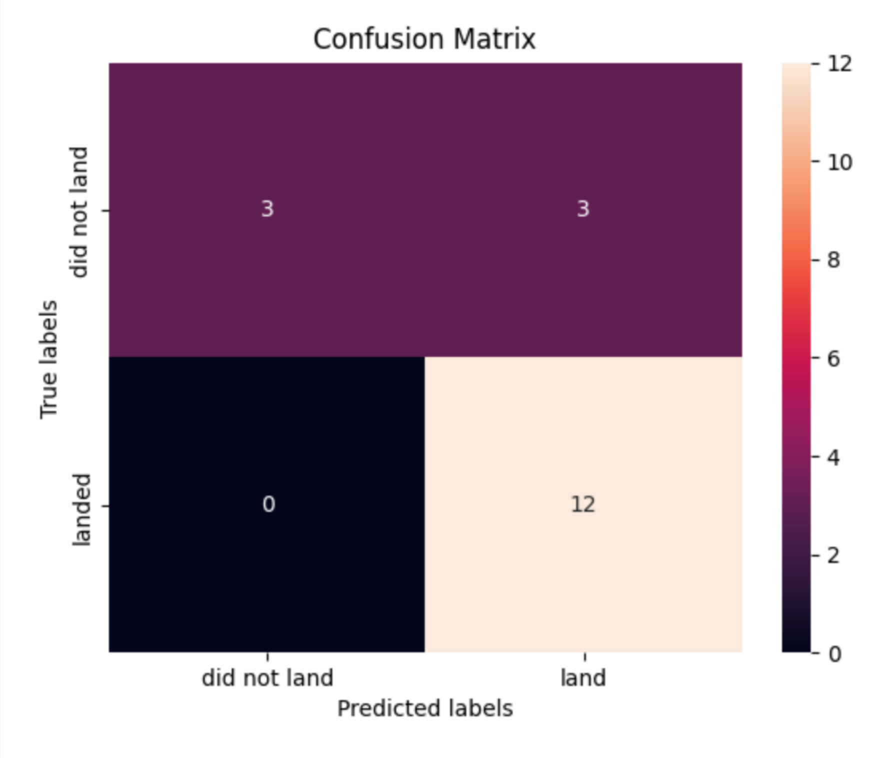
Predictive Analysis (Classification)

Classification Accuracy



- All models have similar accuracy (~83%)

Confusion Matrix



- All models have similar confusion matrices

Conclusions

- The more attempts SpaceX makes the more success launches they get
- ES-L1, GEO, HEO, SSO are the most successful orbits
- We can easily identify that KSCLC-39A has relatively high success rate.
- All ML models showed similar accuracy (~83%) which is sufficient for determining if the first stage will land

Thank you!