IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Megan Verreynne
30 January 2023

Outline

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

Executive Summary

Research methods and data sources:

- Data collection was done using SpaceX API (RESTful API) and web scraping
- Exploratory Data Analysis (EDA) completed including:
  - Data Wrangling
  - Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics and map with Folium
  - Building a dashboard with Plotly Dash
- Machine learning prediction completed using train/test/split: Predictive analysis.

Summary of all results

- EDA identified which features and variables best to predict successful landings.
- Machine Learning prediction showed the best model accuracy.

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.
- We want to predict if the Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch.
- Valuable data was collected from public sources

Questions to be answered

- What variables affect the success of the first landing stage?
- Can variables be classified according to most to least affect on first landing stage?
- Does rath of successful landings increase over the years?
- What is the best algorithm for classification?
- Where is the best location to launch a rocket?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from 2 sources, namely:
  - SpaceX API
  - WebScraping

  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- Perform data wrangling

  Data was first categorized according to number and occurrence of each orbit and then mission outcome per orbit type.

  Landing outcome label based on outcome data was created

  Data was filtered, missing data was dealt with, data was transposed to binary classification using One hot encoding

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models
  - Data was normalized. Data was divided in training and testing data.
  - Data was evaluated by four classification models to find the best accuracy.

Data Collection

- Describe how data sets were collected.

  Data was collected via a request to the SpaceX API (REST API)

  https://api.spacexdata.com/v4/rockets/ (to get booster name)

  https://api.spacexdata.com/v4/launchpads/ (for launch site)

  https://api.spacexdata.com/v4/payloads/ (payload mass)

  https://api.spacexdata.com/v4/cores/ (for landing outcome)

  https://api.spacexdata.com/v4/launches/past (rocket launch data)

  And from Wikipedia

  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 (for flight no, launch site, payload mass, orbit, customer, launch outcome, booster landing, date and time)

Data Collection – SpaceX API

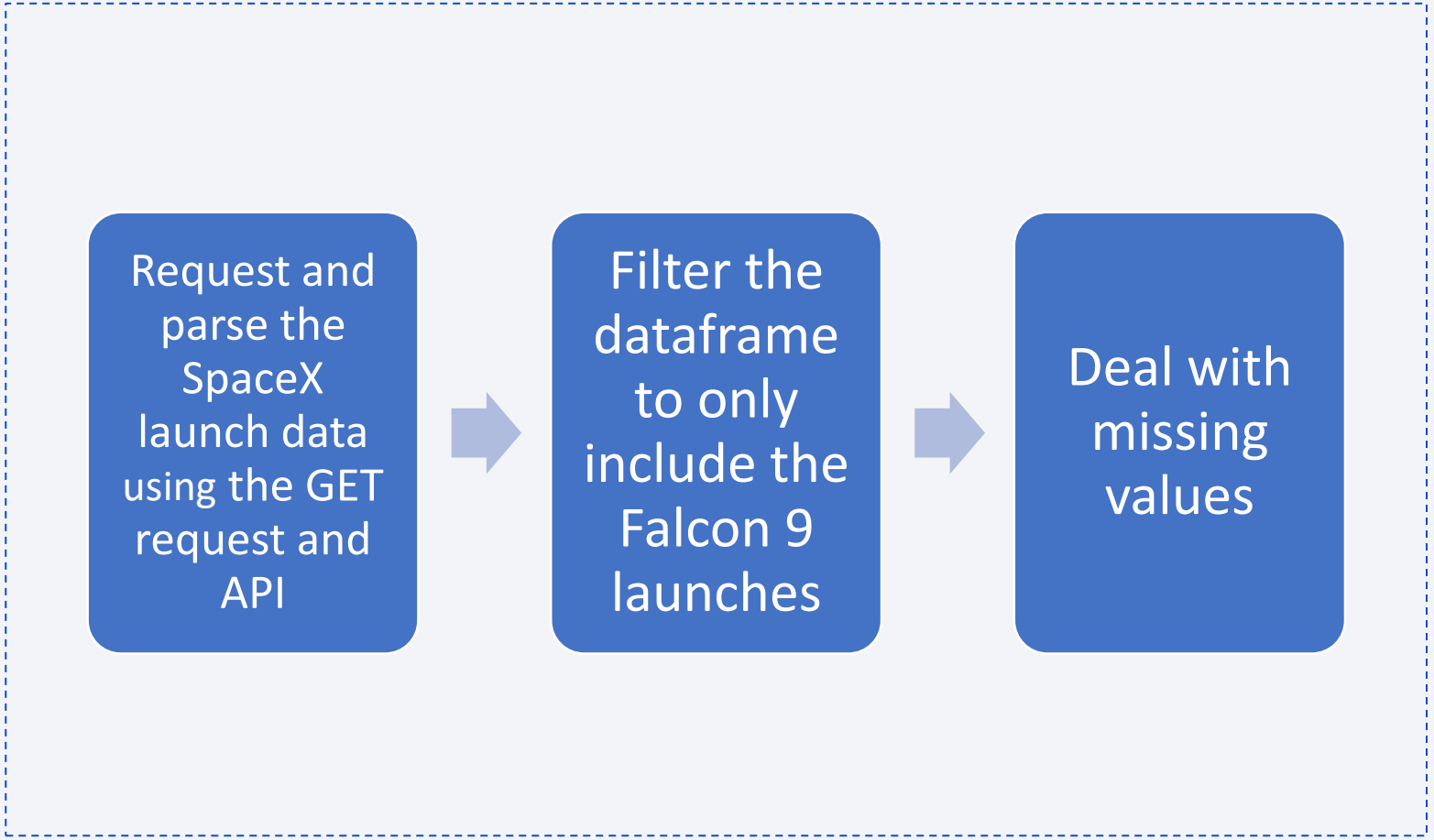
- SpaceX offers a public API from where data can obtained,

Send a request to the API

Clean the requested data

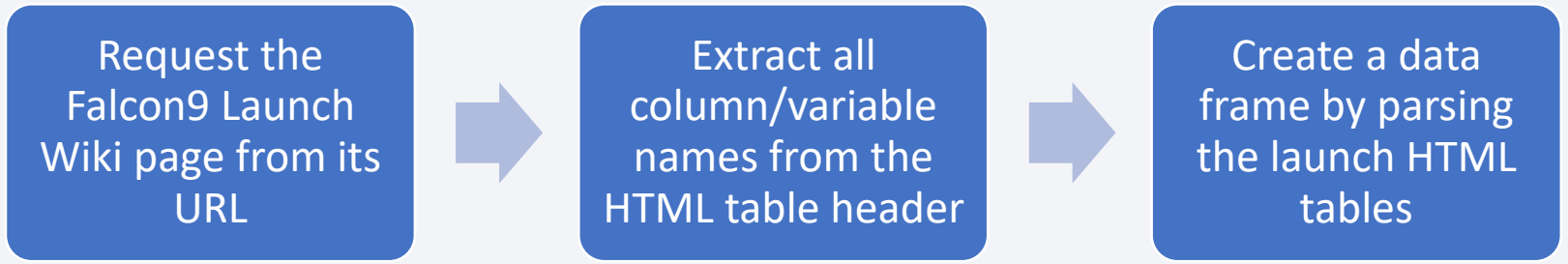
GitHub URL:

- https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

- SpaceX launch data can be obtained from Wikipedia
- Extract a Falcon 9 launch record HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame



Github URL:

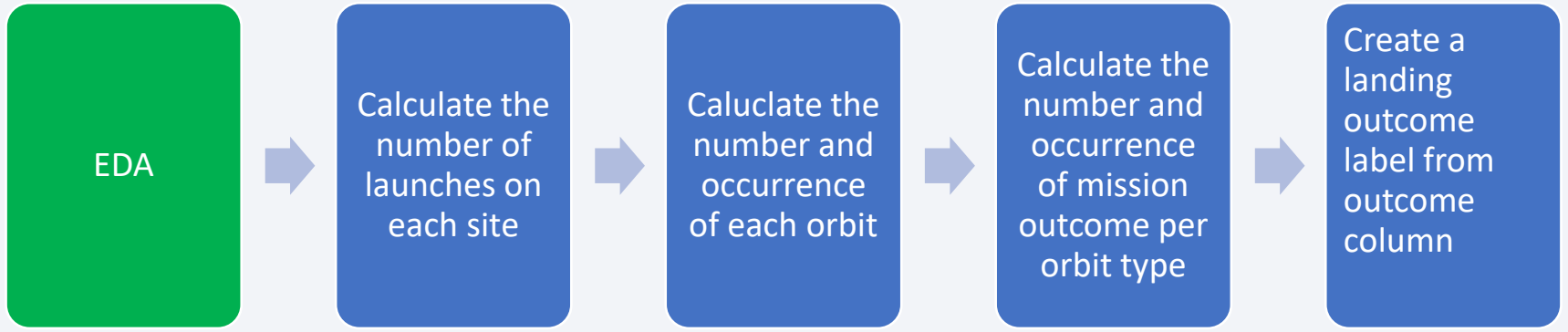
https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/webscraping.ipynb

Data Wrangling

- Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. Convert outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Github URL:

- https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Scatterplots and barplots were used to visualize the relationship between features i.e.

- FlightNumber vs. PayloadMass
- FlightNumber vs LaunchSite
- PayloadMass Vs. Launch Site
- Flight number vs Orbit
- Payload vs. Orbit

GitHub URL

- https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The following SQL queries were performed:

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass.
- 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017.

Github URL

https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers, circles, lines, etc. you created and added to a folium map:

- Markers incl each site's location on a map using site's latitude and longitude coordinates
- Circles are a highlighted circle area with a text label on a specific coordinate.
- Marker clusters are used to simplify a map containing many markers having the same coordinate.
- Lines and polylines are used to draw a line between a launch site to its closest city, railway, highway etc

GitHub URL

- https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

To visualize the data the following graphs and plots were used:

- Percentage of launches by site
- Payload range

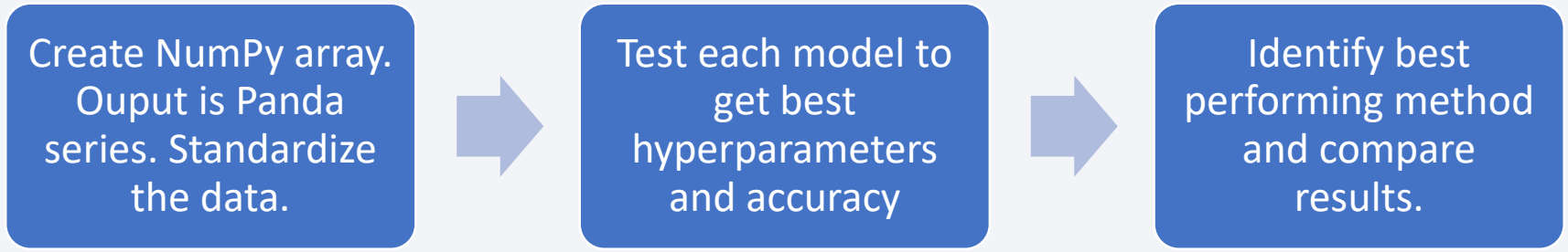
Identified best place to launch according to payloads.

Github URL:

https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/plotly%20dash.txt

Predictive Analysis (Classification)

- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression to find the method that performs best using test data
- Classification models were logistic regression, support vector machine, decision tree and k nearest neighbors.



GitHub URL:

- https://github.com/meganverr/Applied-data-science-capstone/blob/b0b49d74bd289f79fea4827c9a74dfd26330a6ab/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

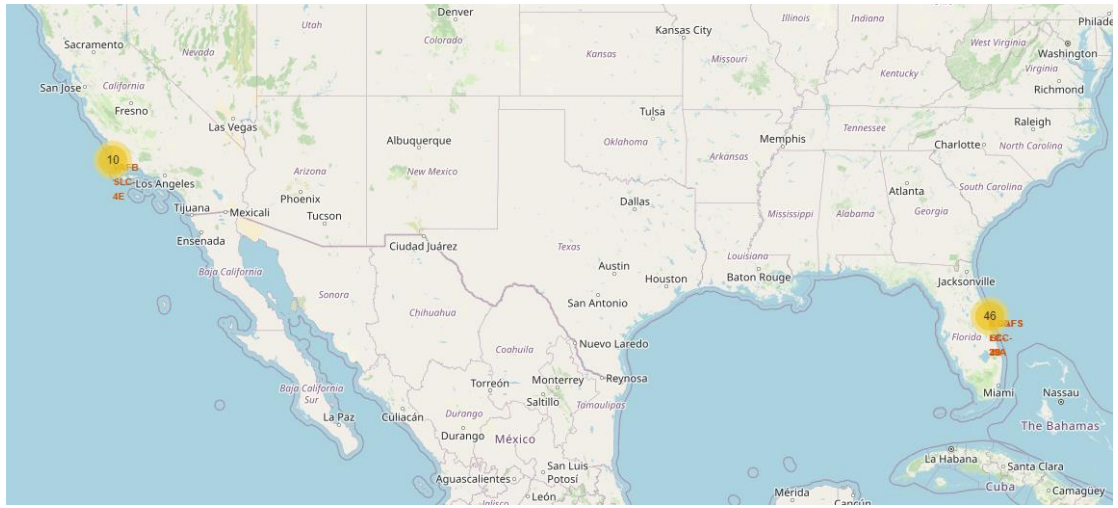
Exploratory Data Analysis

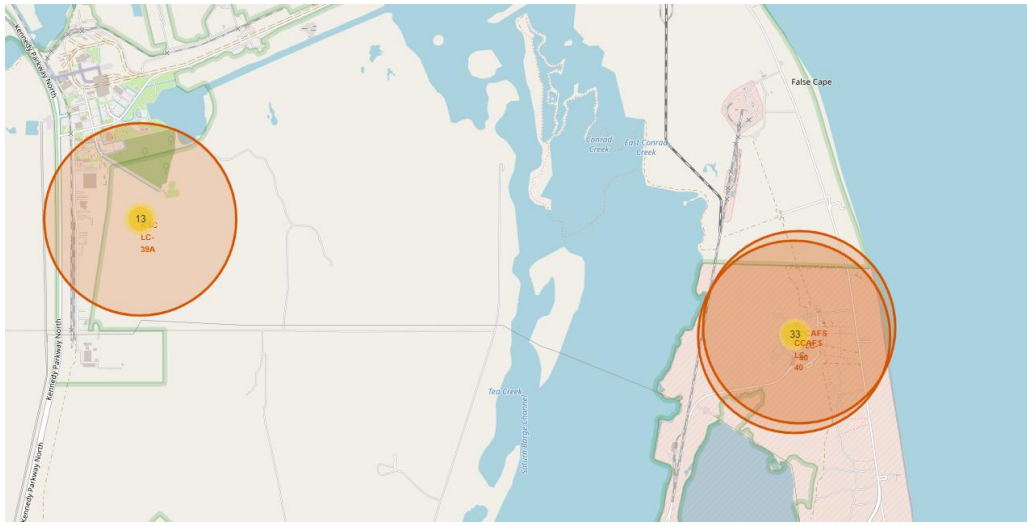
SpaceX has currently has 4 different launch sites

- The number of successful landings increased with time.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- The higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.
- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO

Interactive analytics demo in screenshots

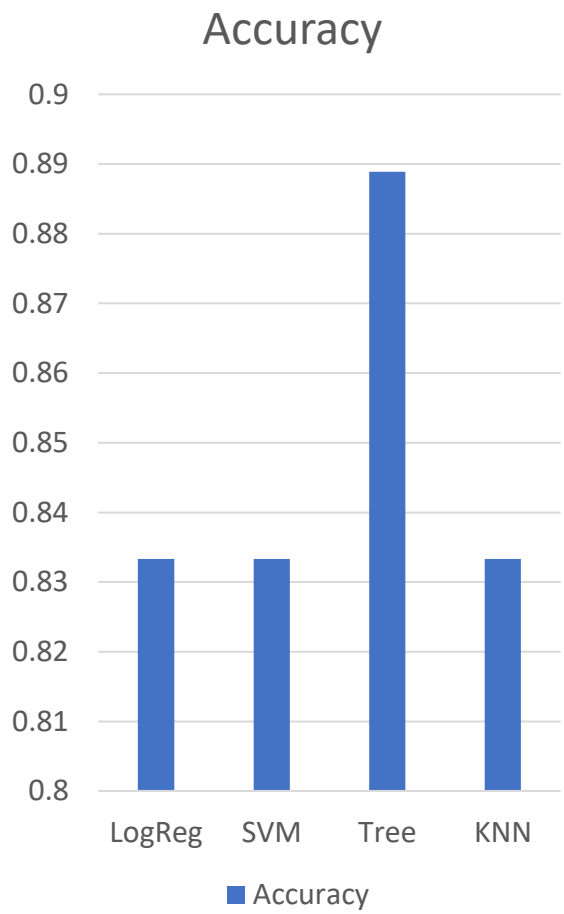
- Interactive analytics identified that launch sites are located near the ocean and have good logistic infrastructure in close proximity but are not too close to inhabited areas for satety precautions.
- Most launches were on the east coast

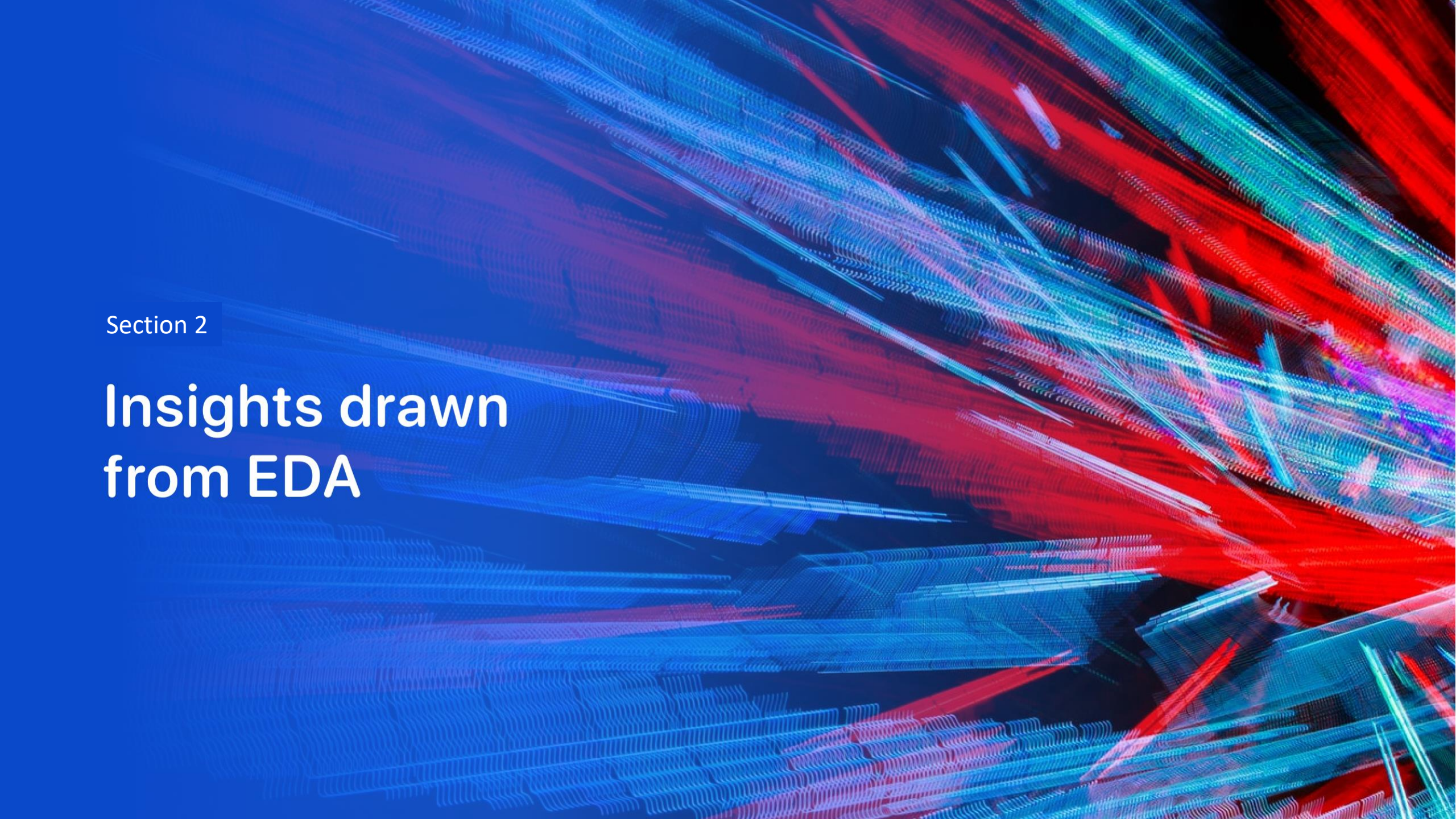




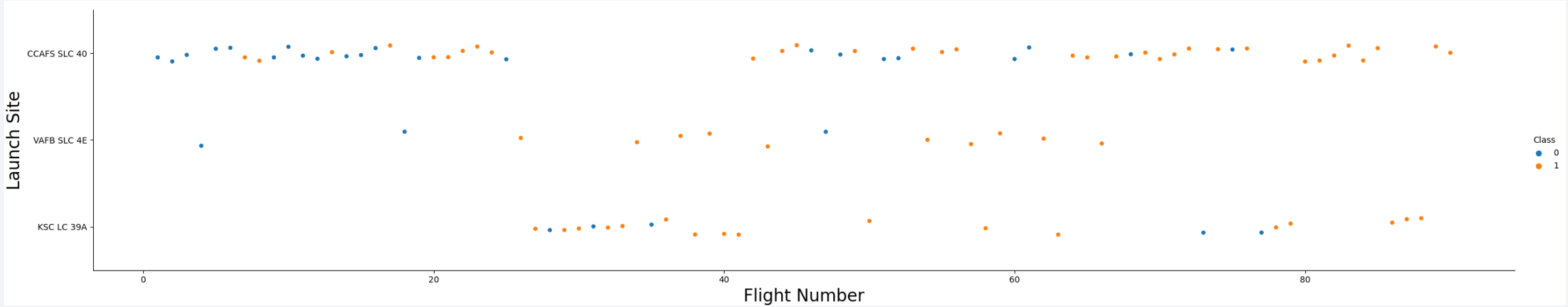
Predictive analysis results

- Predictive analysis showed that the Decision tree classifier is the best model to predict successful landings having an accuracy of 0.88889 (88.9%)



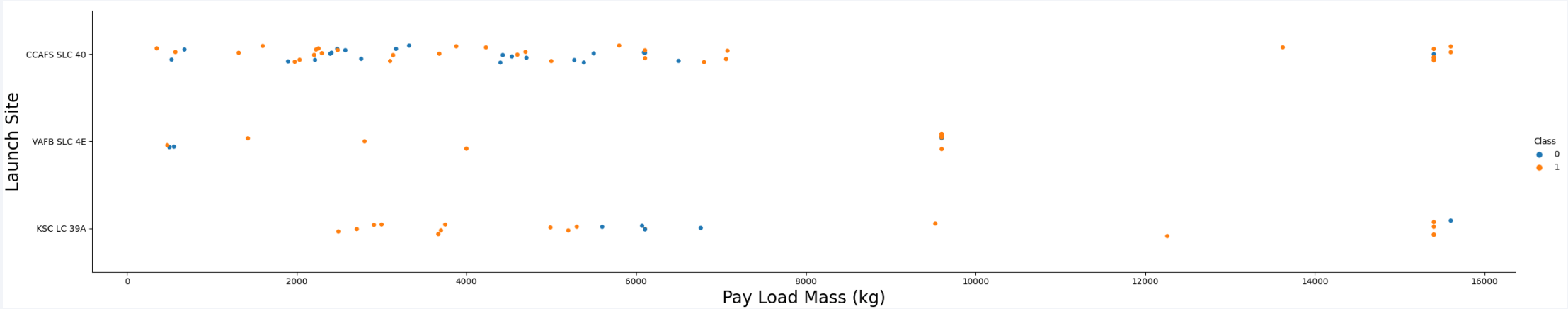

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Space X currently has 4 launch sites
- Earlier flights usually failed while later flights succeeded
- The CCAFS SLC 40 launch site had the most launches
- VAFB SLC 4E and KSC LC 39A have higher overall success rates
- Therefor, it can be surmised that each new launch has a statistically higher possibility of success

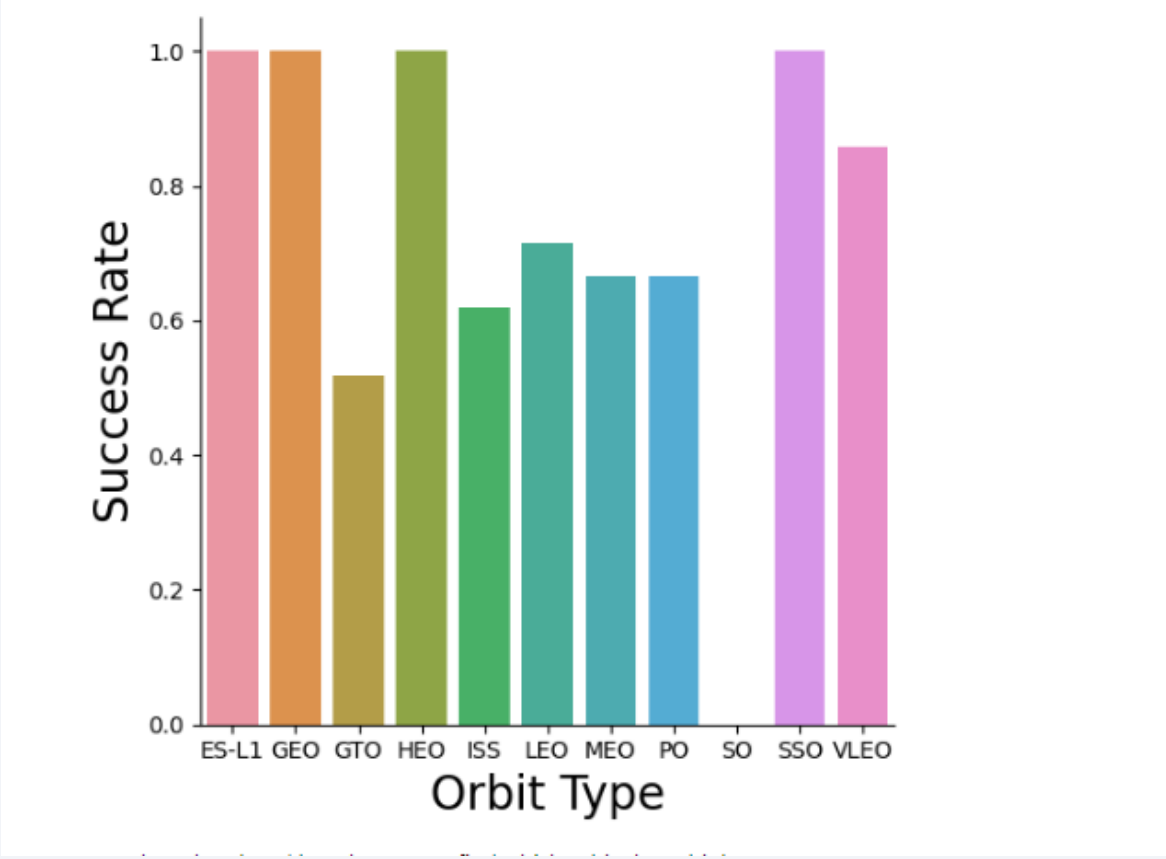
Payload vs. Launch Site



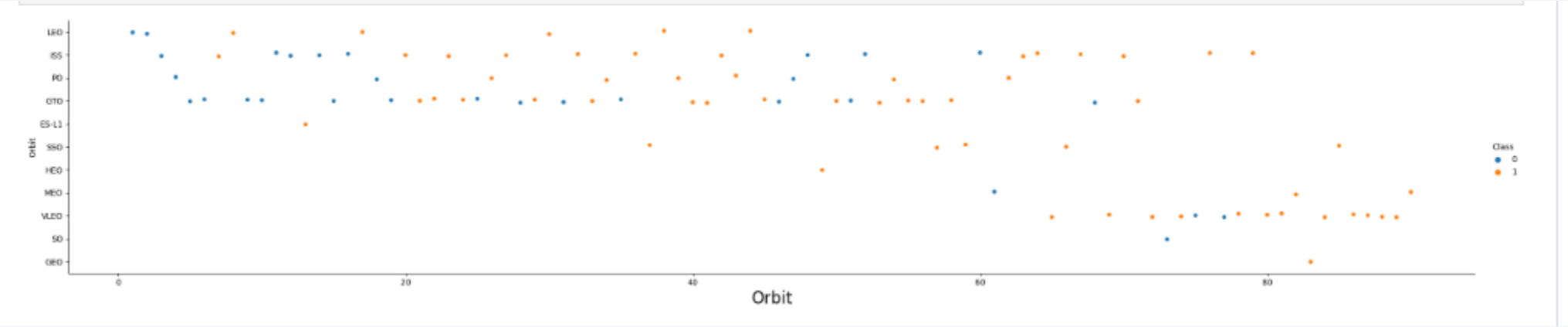
- For every launch site, the higher the payload mass, the higher the chance at a successful outcome
- Most launches with mass over 7000kg were successful
- Although, it is noteworthy that KSC LC 39A has a 100% success rate for payload mass under 5500kg too.
- Payloads over 10000kg have only been tried on CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO had a 100% success rate
- Orbit SO had 0% success rate.

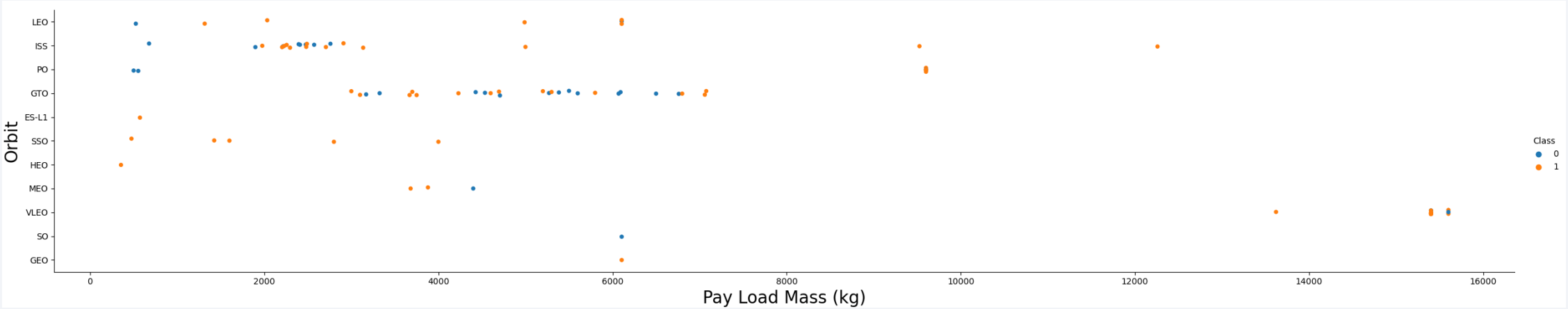


Flight Number vs. Orbit Type



- In the 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.
- There seems to be recent favor in the Orbit: VLEO as the frequency of missions there has increased while other missions decreased over time

Payload vs. Orbit Type

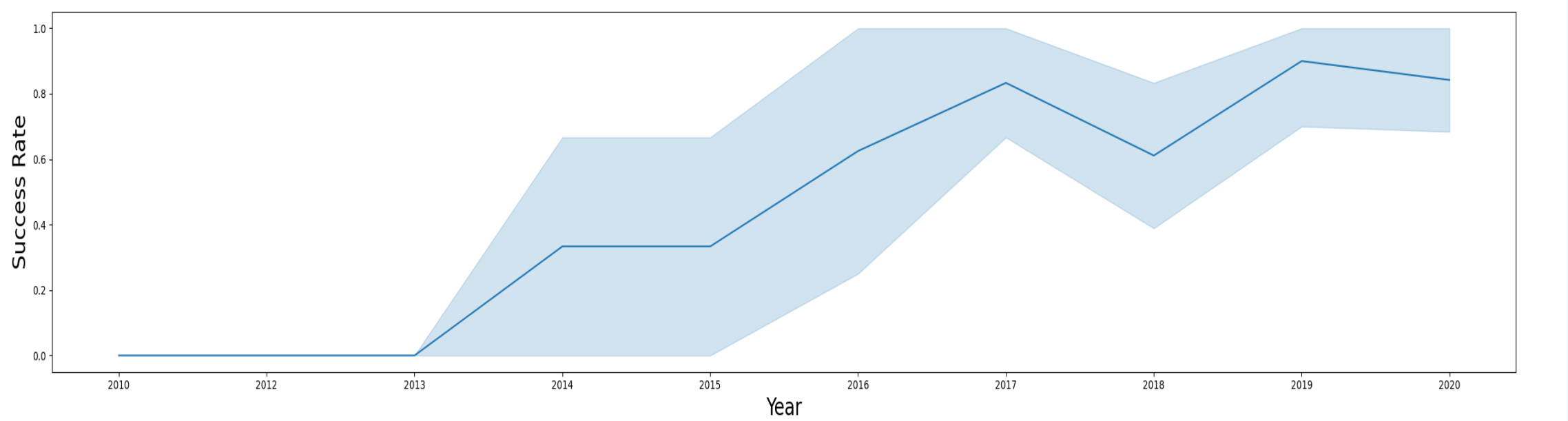


Payload mass vs orbit type:

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

Launch Success Yearly Trend

- Success rate increases between 2013 and 2020 (apart from a dip in 2018 which recovers).
- 2010-2013 was perhaps a period or adjustment and improvement as for most start-ups.



EDA with SQL

All Launch Site Names

- Displayed below are the results of the distinct Launch Site names for the dataset.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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

- Total payload mass carried by boosters from NASA (CRS)

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

avg_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad
- The first successful landing in ground pad from the dataset was achieved on 01-05-2017

first_succesful_landing
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- Total number of successful and failure mission outcomes

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

- 99 were successful, 1 unclear and 1 resulted in failure in flight.

Boosters Carried Maximum Payload

- Booster version carrying the maximum payload mass in the dataset.

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

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Only two occurences in the year 2015

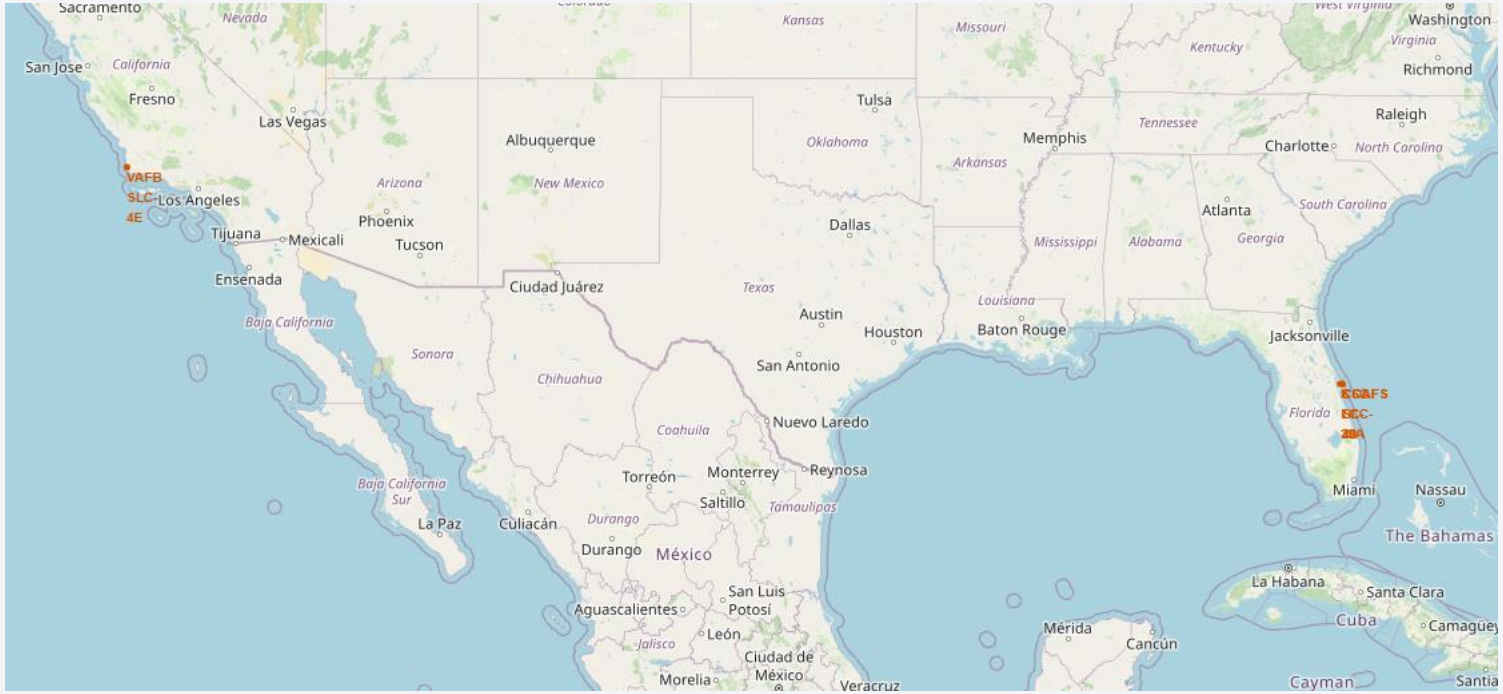
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing _Outcome	count_outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

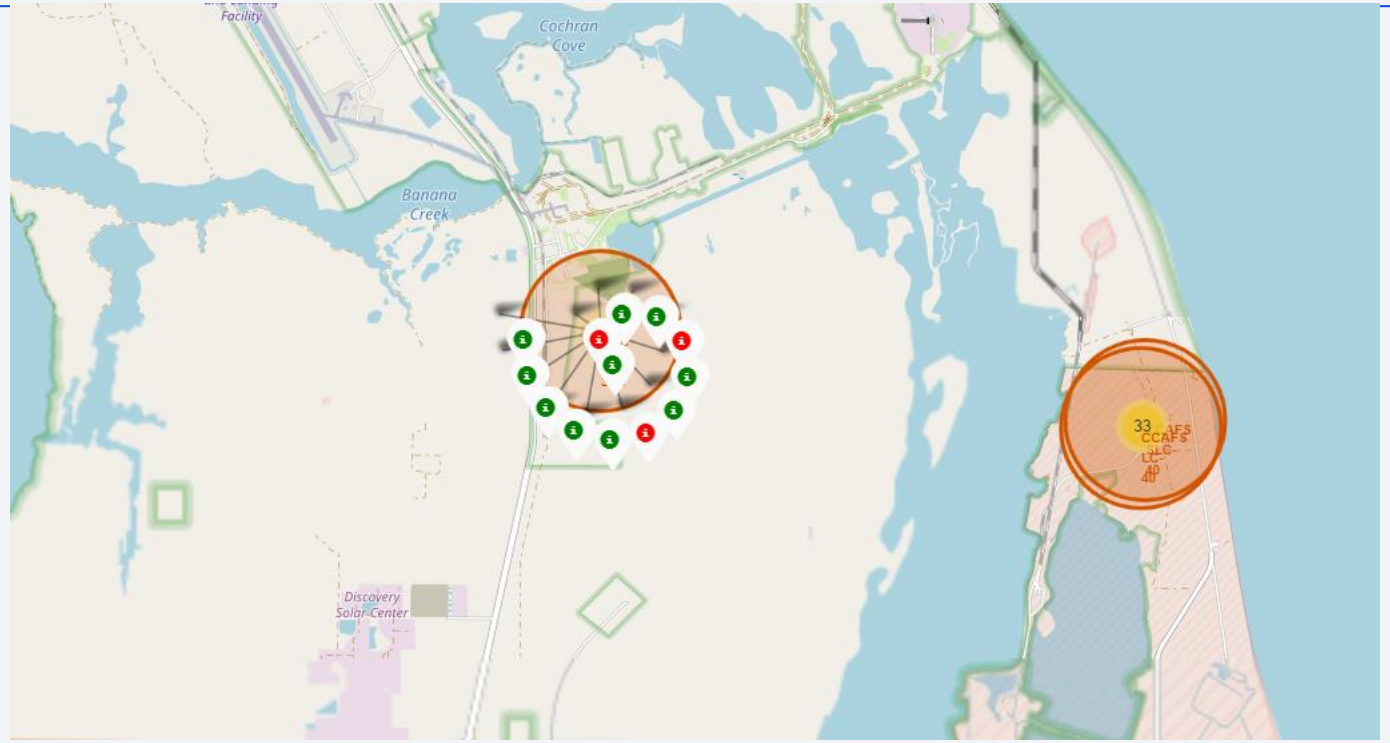
Launch Sites Proximities Analysis

Launch sites



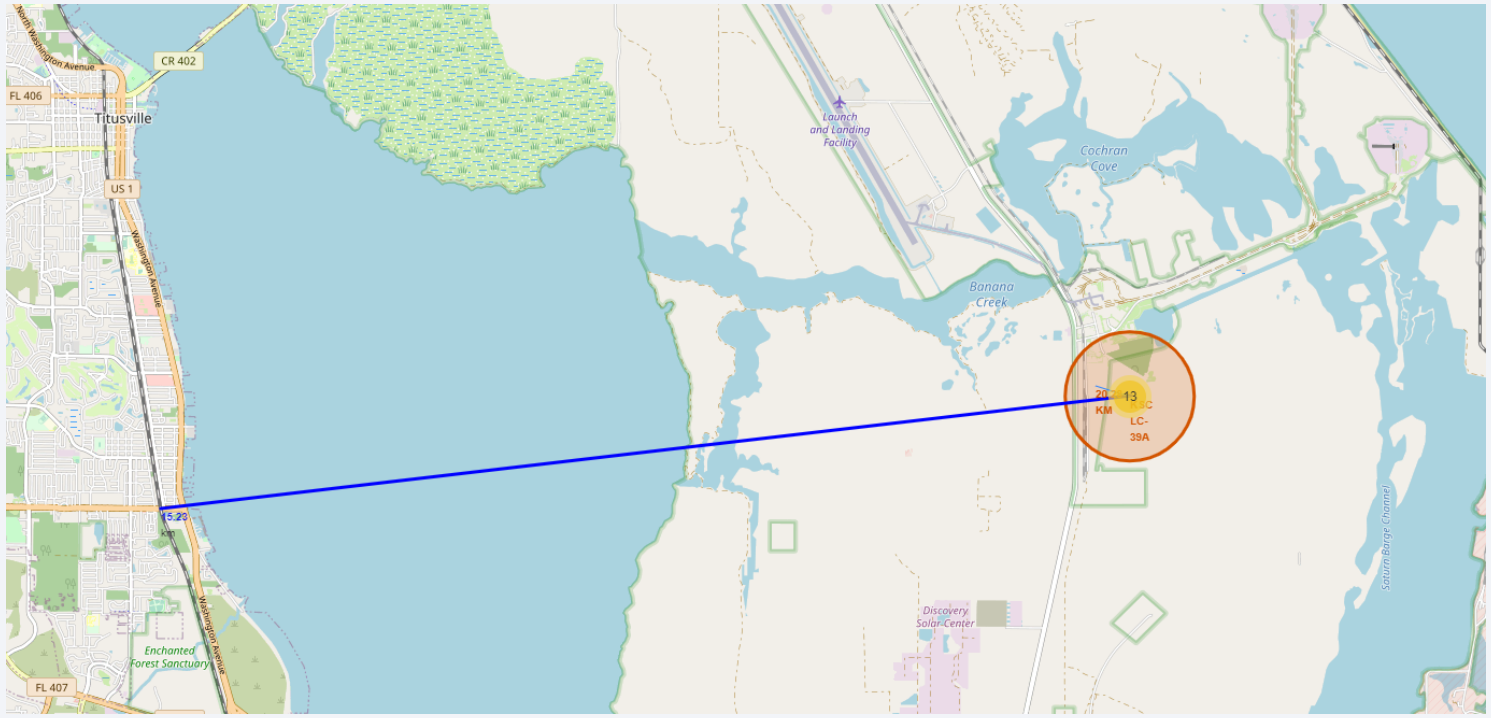
- Launch sites are close to the ocean, in close proximity to roads and railroads and further away from inhabited areas. All for safety precautions.

KSC LC-39A launch site outcomes

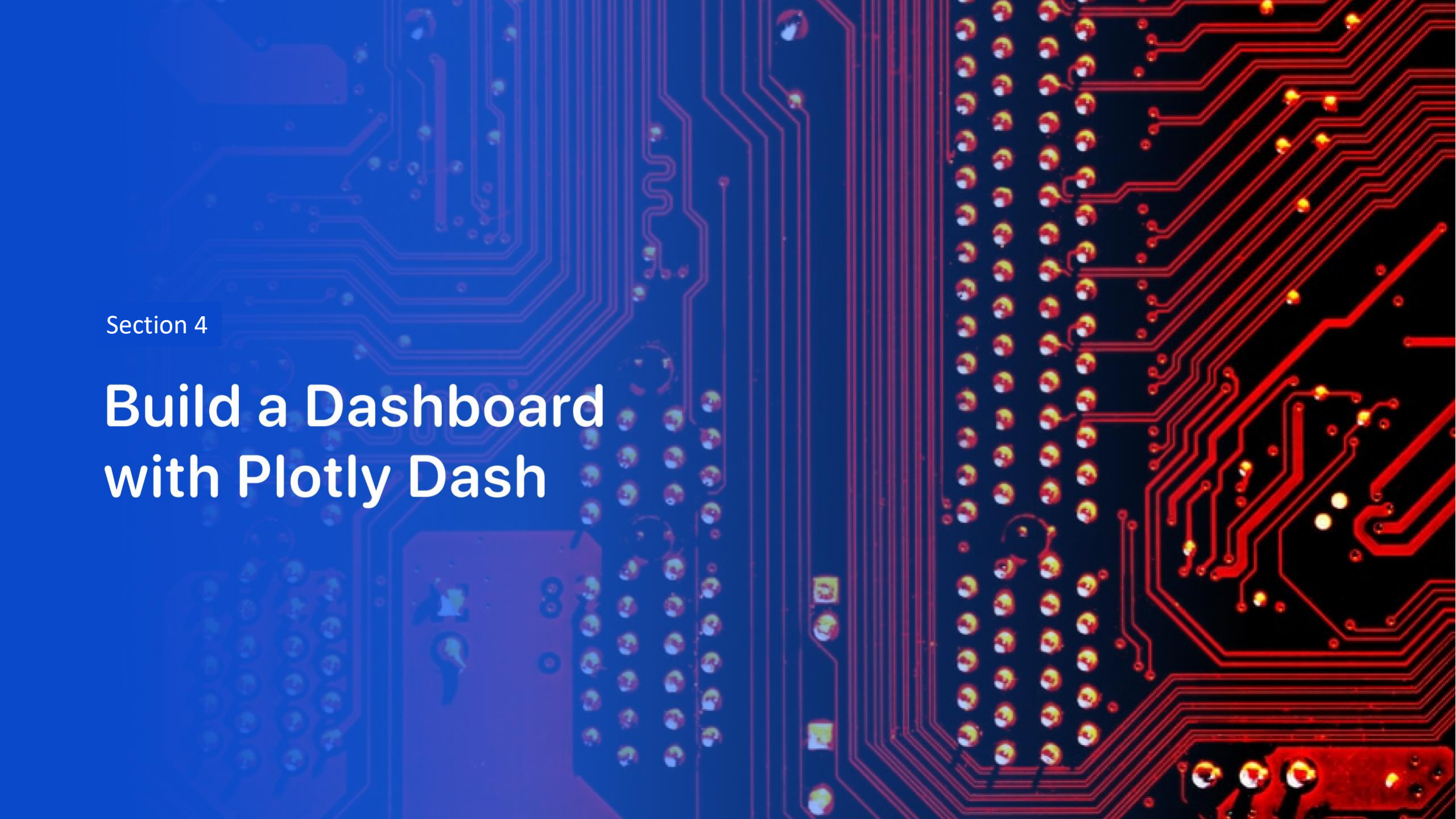


- Colour-labeled markers make it easy to identify success rates.
- Green markers indicate successful and red ones indicate failure

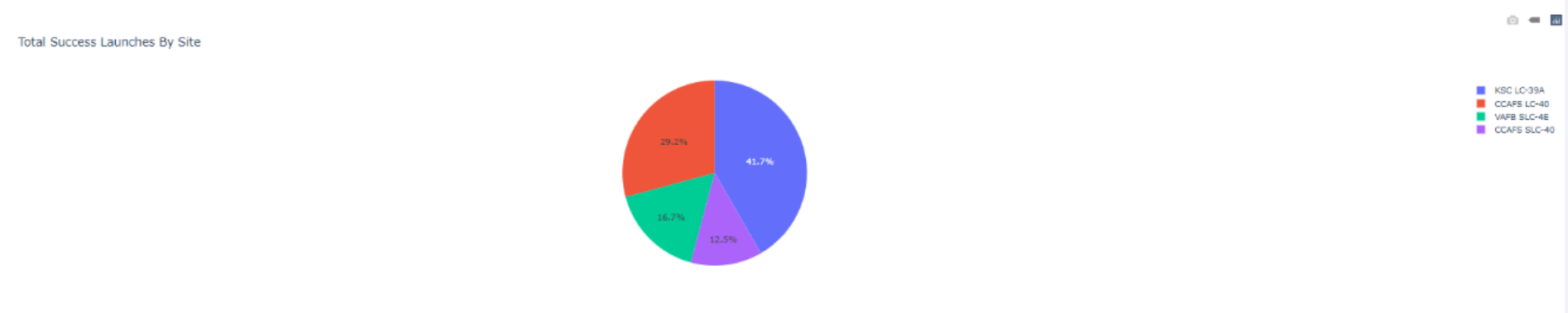
KSC LC-39A Launch site and surroundings



- Relatively close to railways, highways and coastline
- Relatively far from inhabited areas.

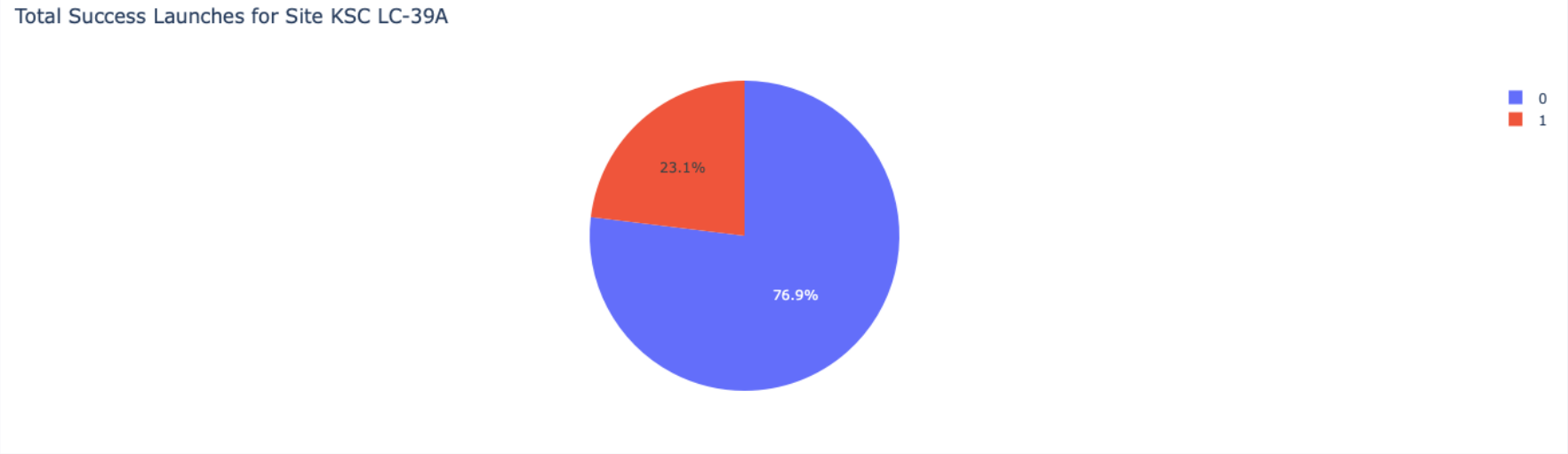

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



Launch site is an important factor of success. KSC LC-39A has the most success.

Launch Success for KSC LC-39A



- 76.9% of launches are successful at this site.

Payload Mass vs Launch Outcome

- Payloads between 2000 and 5500kg have higher rates of success.

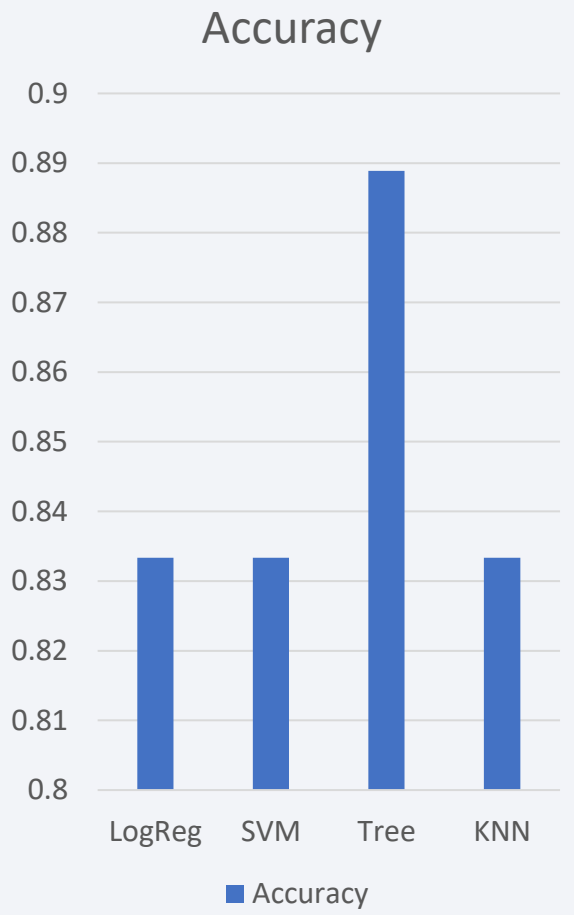
Section 5

Predictive Analysis (Classification)

Classification Accuracy

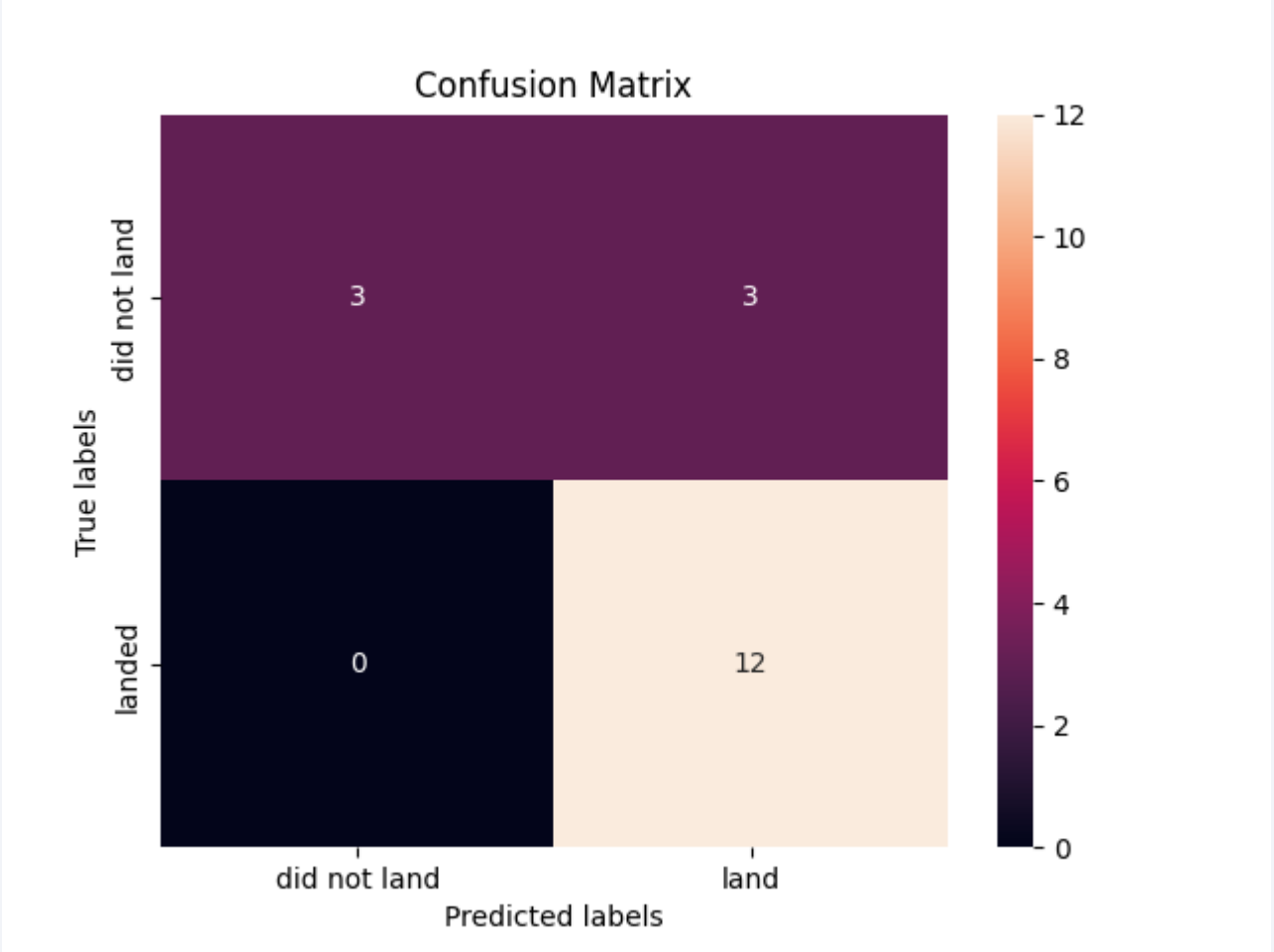
- Decision Tree Classifier had the best accuracy

	LogReg	SVM	Tree	KNN
Accuracy	0.833333	0.833333	0.888889	0.833333



Confusion Matrix

- The decisition tree classifier was the best algorthim.
- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The best site launch is KSC LC-39A
- Launches above 7000kg have better rates of success
- Successful landing outcomes improve over time
- Decision tree classifier can be used to predict successful landings and was the best algorithm in this dataset.
- Launch sites are always in close proximity to the coast.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rates.

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
- Plotly dash saved as txt file

Thank you!