IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Koltai
08/12/2022

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

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 determine the cost of a Falcon 9 rocket launch, which is subject to the success of landing from stage 1. Therefore, we want to predict if the Falcon 9 stage 1 lands successfully

- Problems to find answers to
  - Relation of features such as launch site, payload mass, orbits, number of flights to landing success rate
  - Finding the parameters with the highest landing success rate
  - Finding highest landing success rate

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Webscraping from Wikipedia
- Perform data wrangling
  - Filtering the data, handling missing values, One Hot Encoding to prepare data for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluating classification models, selecting for highest accuracy model

Data Collection

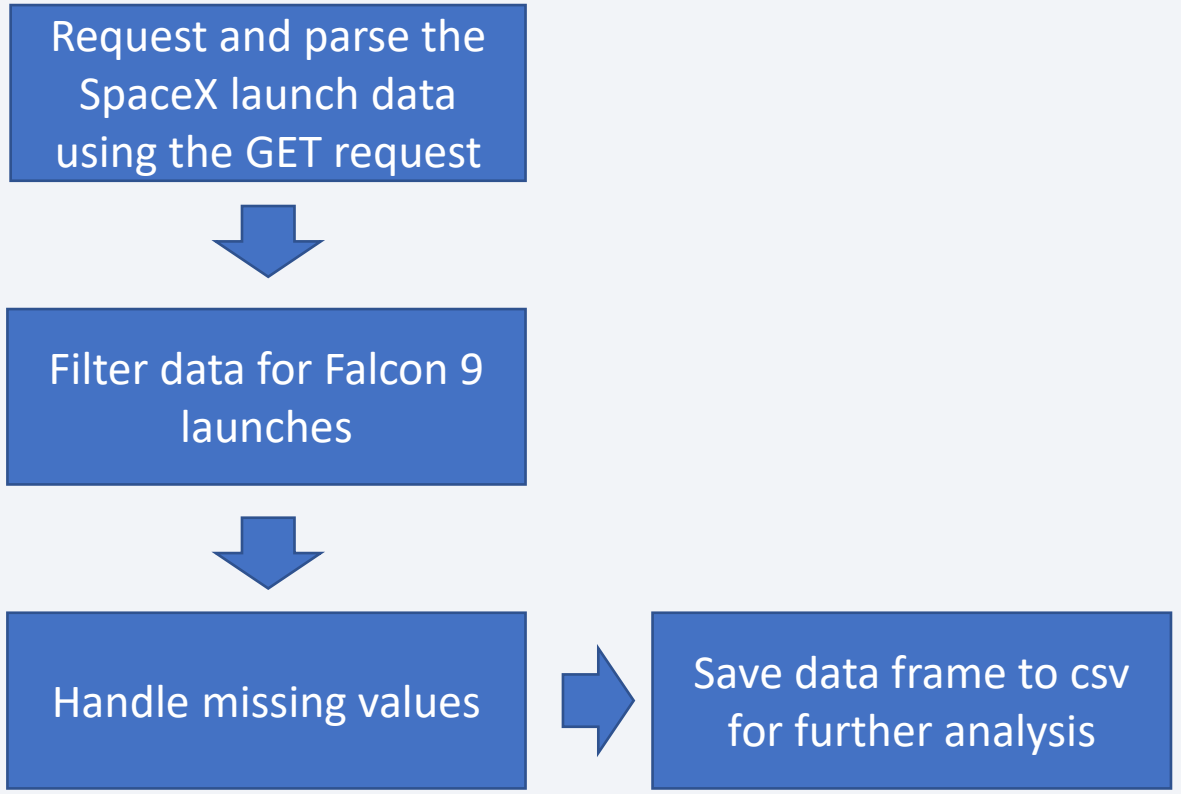
- Data collection was done using get request to the SpaceX API.
- Decoded the response content as a Json using .json() function call,
- Turned Json into a pandas dataframe using .json_normalize().
- cleaned the data, checked for missing values and fill in missing values where necessary

- Performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup
- Extracted the launch records as HTML table, parsed the table and converted it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, cleaned the requested data and did some basic data wrangling and formatting.

- jupyter-labs-spacex-data-collection-api.ipynb

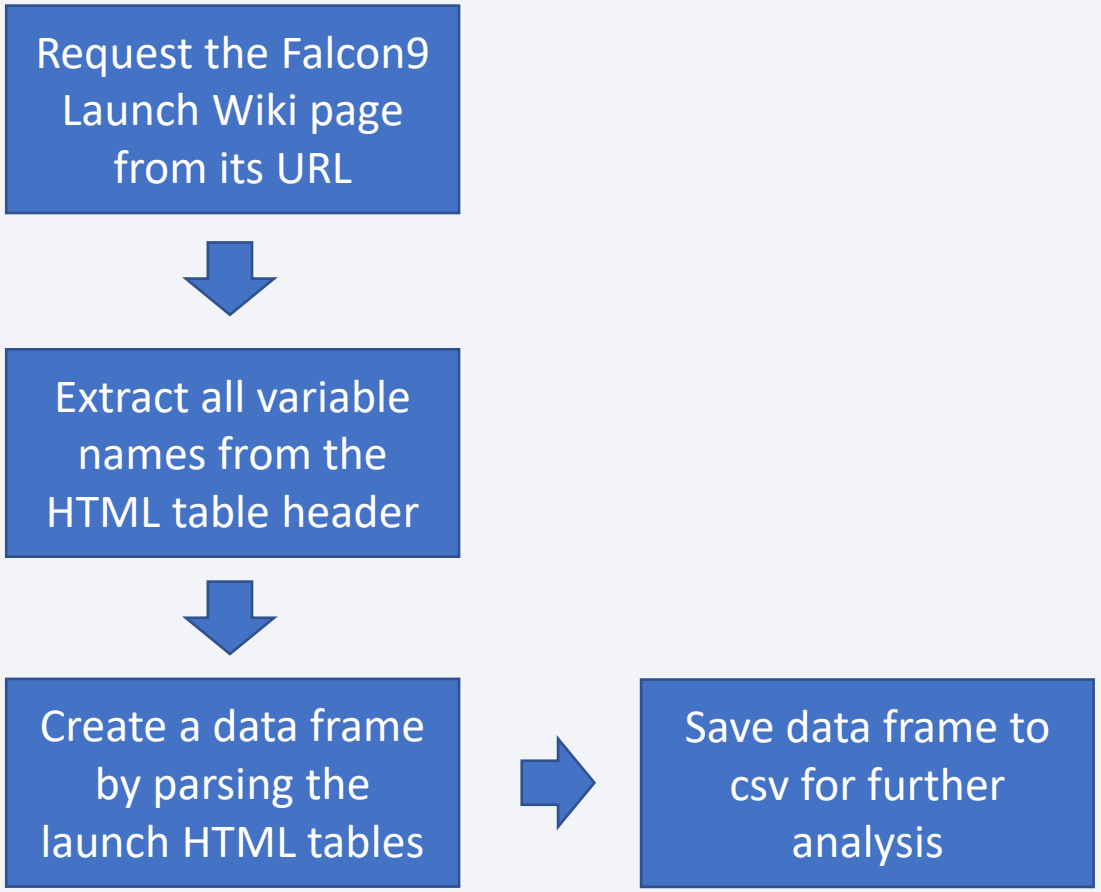


Data Collection - Scraping

Extracted the Falcon 9 launch records HTML table from Wikipedia using Web scraping with BeautifulSoup

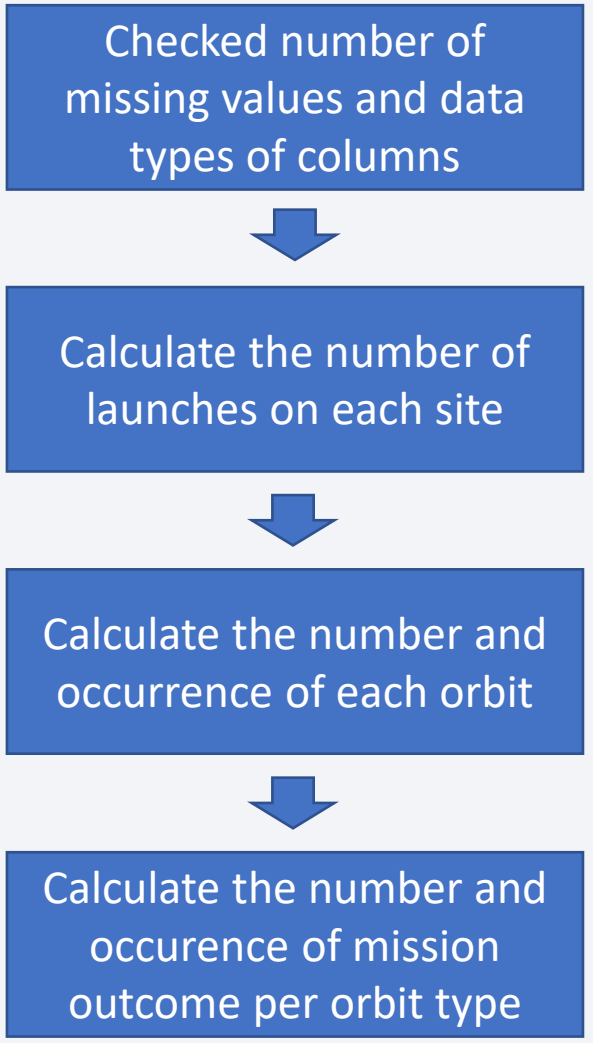
Parsed the table and converted it into a Pandas data frame

jupyter-labs-webscraping.ipynb



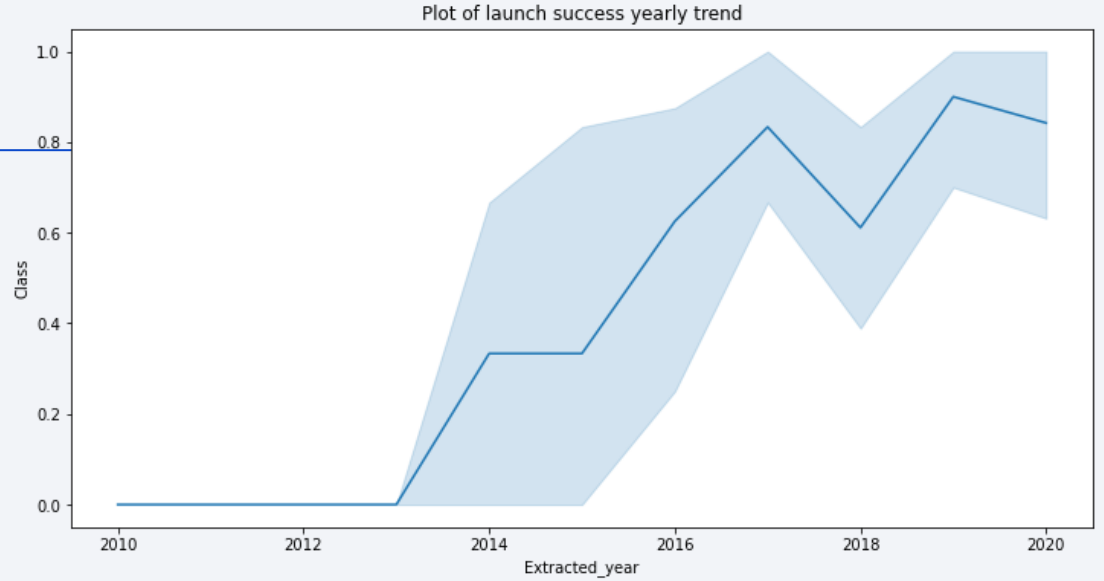
Data Wrangling

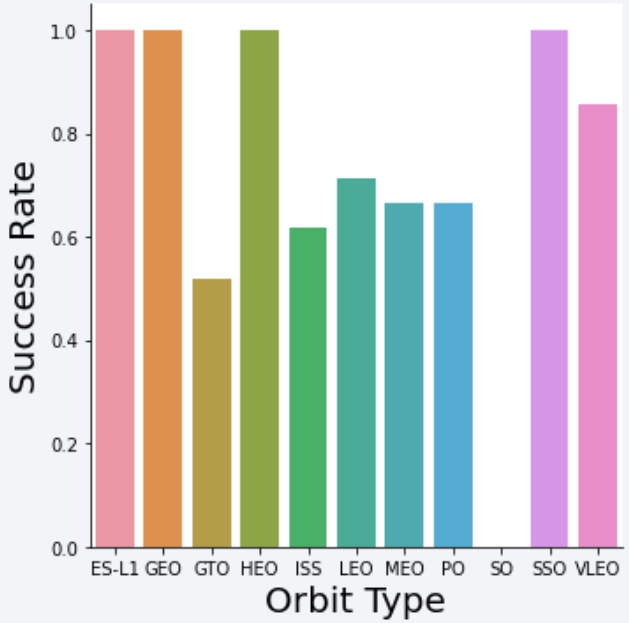
- Analyzed columns for missing values and data types
- Calculate the number of launches on each site, the number and occurrence of each orbit, the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column exported data to csv for further analysis
- [labs-jupyter-spacex-Data-wrangling.ipynb]



EDA with Data Visualization

- Explored the relationship of flight number, payload mass, launch site, orbit type to launch outcome via the following scatter plots
  - Flight Number vs. Payload Mass
  - Flight Number vs Launch Site
  - Launch Site vs Payload Mass
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Also plotted →
- [jupyter-labs-eda-dataviz.ipynb]





EDA with SQL

- 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 acheived.
- 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Marked all launch sites on the map with a circle and a marker containing the site name
  - Added a marker of green/red color to the site of launches indicating its outcome
  - Added a line to the shoreline, railway, highway, city closest to CCAFS SLC-40 launch site

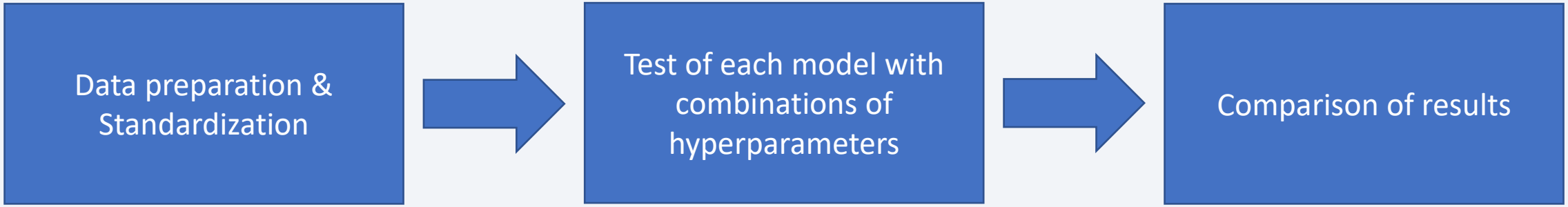
- These objects were added for the purpose of data visualization and analysis of factors related to launch sites affecting the outcome of launches
- lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Pie chart of total successful launches by site
  - Pie charts of launch outcomes for each site
  - Scatter plot of Launch outcome vs Payload mass for all site
  - Scatter plot of Launch outcome vs Payload mass for each site
- To identify optimal launch site for successfully launching a Falcon-9 rocket with a given Payload mass
- spacex_dash_app.py

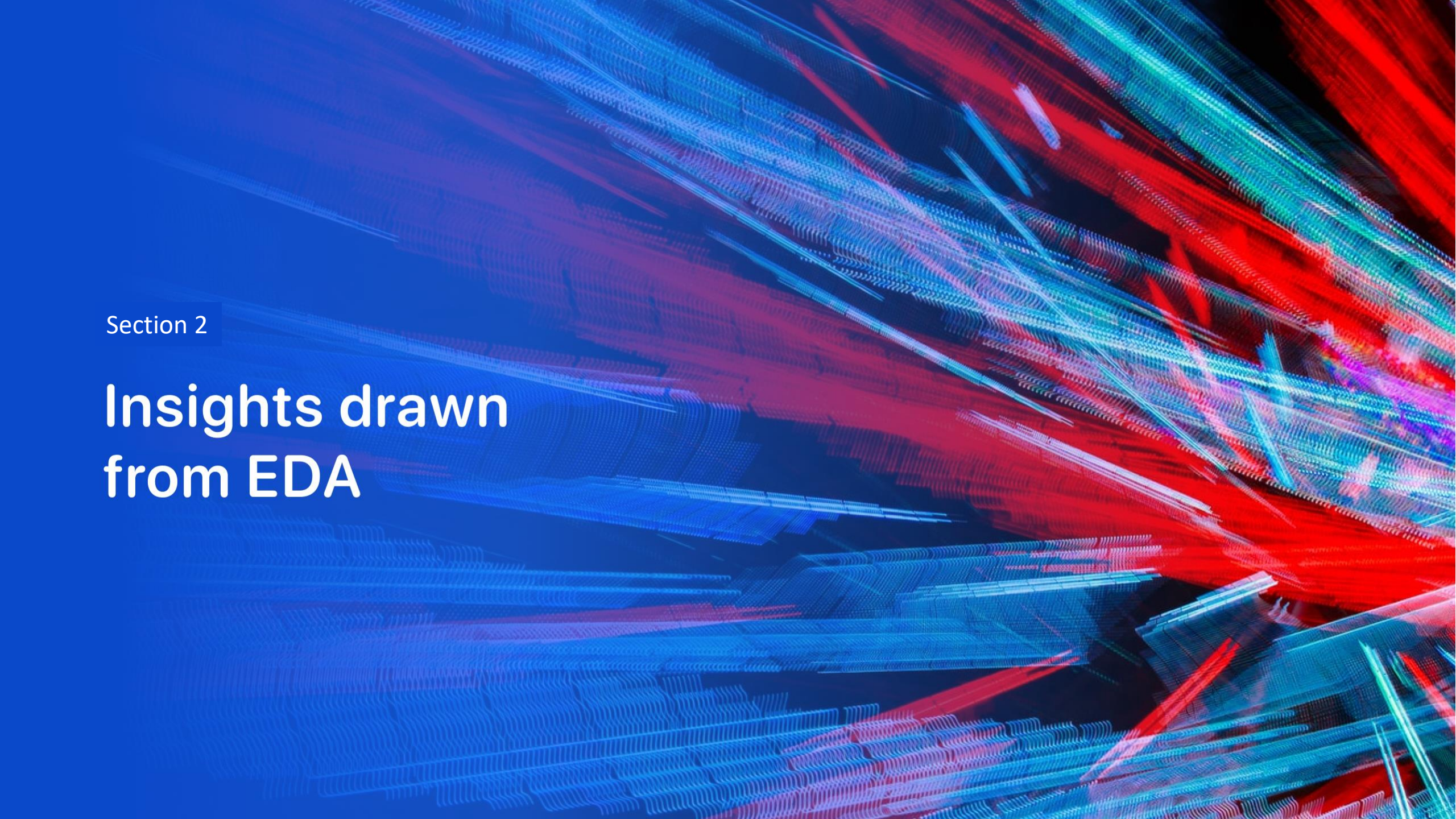
Predictive Analysis (Classification)

- Four classification models with multiple hyperparameters; logistic regression, support vectom machine, decision tree and k nearest neighbors were compared to find the most accurate one
- SpaceX_MachineLearningPrediction_Part_5.ipynb



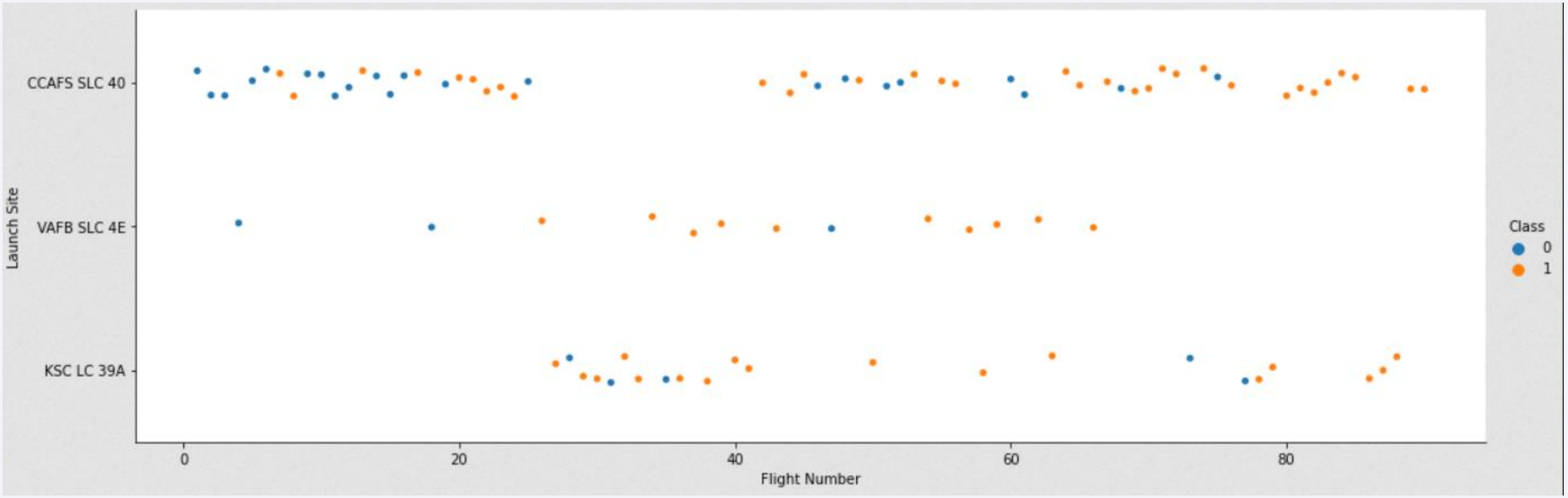
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

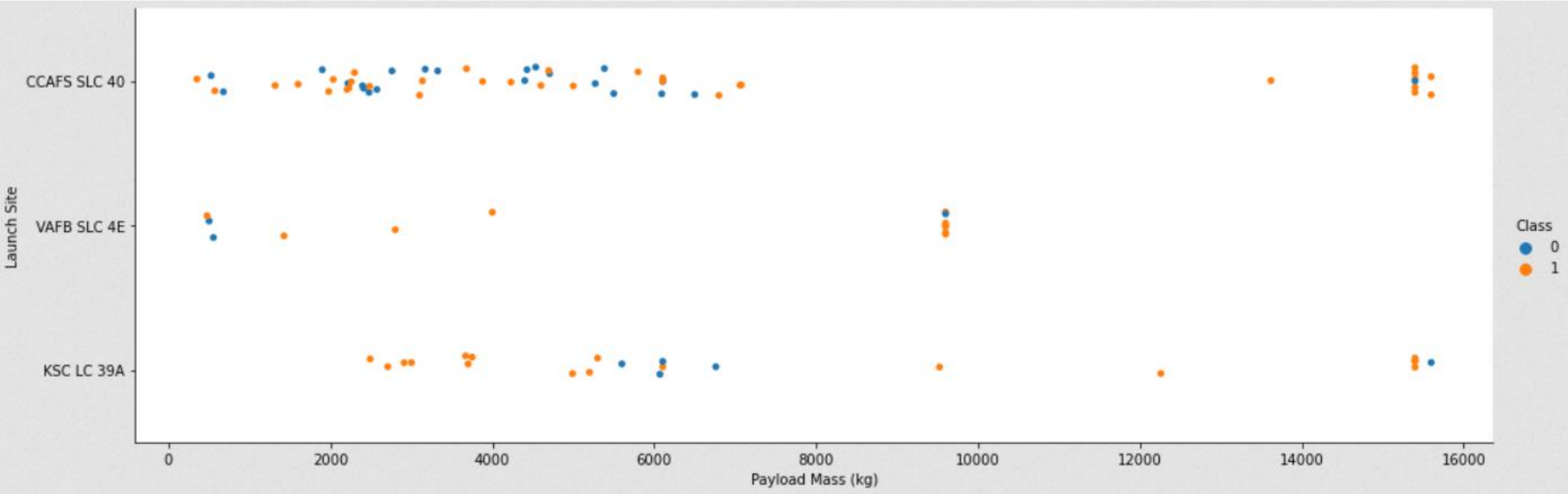
- Scatter plot of Flight Number vs. Launch Site



- As flight numbers increased the number of successful launches increased while the number of unsuccessful launches decreased at all the launch sites
- Launches with a flight number above 80 were all succesful

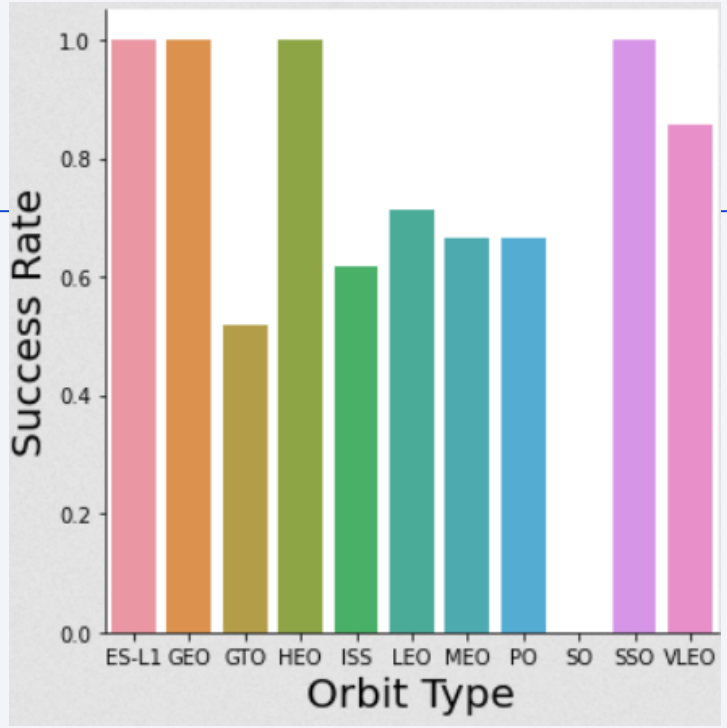
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



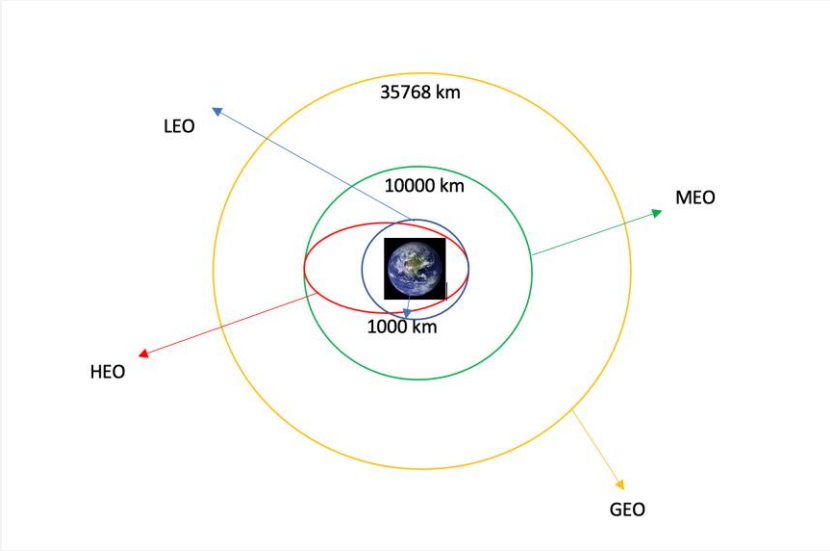
- The success rate increases with increase in payload mass
- CCAFS SLC 40 has the highest success rate for Payload Mass higher than 10000 kg

Success Rate vs. Orbit Type



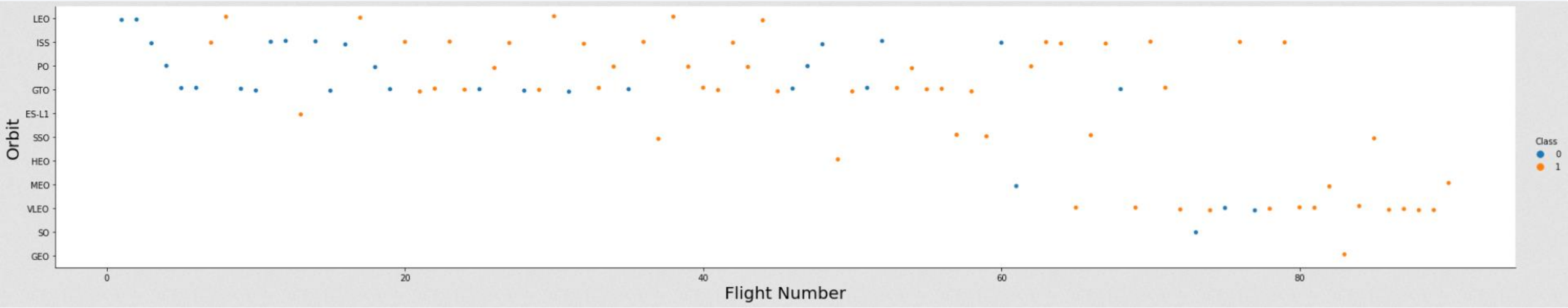
- Bar chart for the success rate of each orbit type

- ES-L1, GEO, HEO, SSO orbits have a success rate of 1
- GTO, ISS, LEO, MEO, PO and VLEO have success rates between 0.4 and 0.8
- SO has a success rate of 0



Flight Number vs. Orbit Type

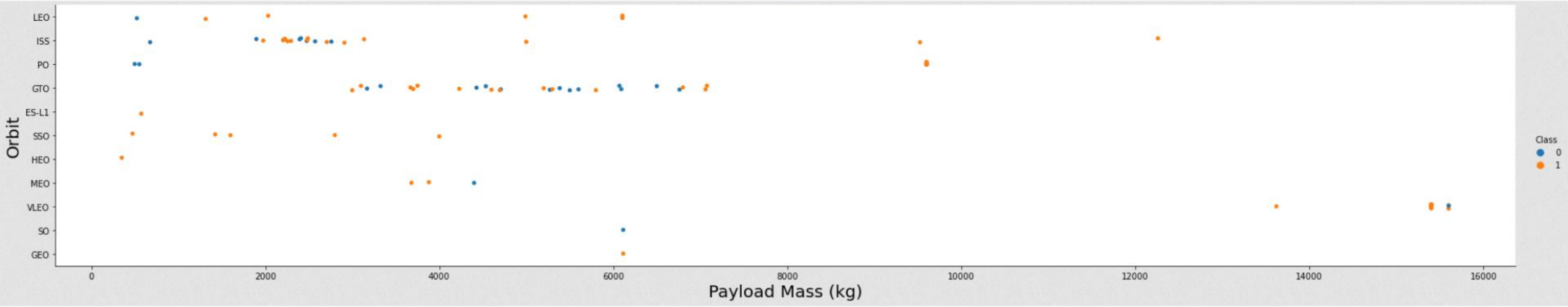
- Scatter point of Flight number vs. Orbit type



- Success rates increase for all Orbit types as Flight numbers increase
- For all orbits in the data, except SO, the highest recorded flight number corresponds to a successful launch

Payload vs. Orbit Type

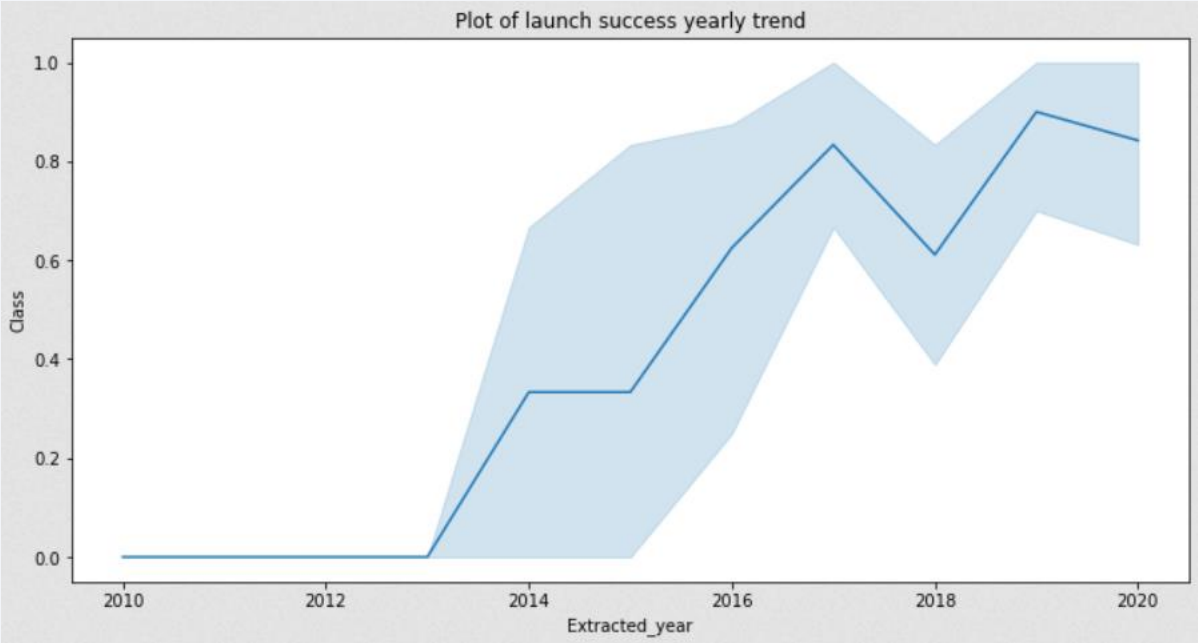
- Scatter plot of payload vs. orbit type



- For all orbits other than GTO and SO success rate increases with the increase of payload mass
- GTO and SO has an unclear relationship with Payload Mass

Launch Success Yearly Trend

- Line chart of yearly average success rate



- Yearly average success rate shows an increasing trend from 2013
- The increase seem to slow down near stagnation from 2017

All Launch Site Names

- The names of the unique launch sites:

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

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset
- %sql select distinct launch_site from SPACEX;

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
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- They are obtained by selecting launch_site values the begin with CCA using the % wildcard for characters of string data

- %sql select * from SPACEX where launch_site like 'CCA%' limit 5on here

Total Payload Mass

- Total payload carried by boosters from NASA

tot_payload_mass
45596

- It was obtained by selecting the sum of payload mass kg where customer values is ‘NASA (CRS)’
- %sql select sum(payload_mass__kg_) as tot_payload_mass from SPACEX where customer = 'NASA (CRS)';

Average Payload Mass by F9 v1.1

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

avg_payload_mass
2928

- It was obtained by selecting average of payload_mass_kg where booster_version starts with ‘F9 v1.1’ using the % wildcard for characters of string data
- %sql select avg(payload_mass__kg_) as avg_payload_mass from SPACEX where booster_version like '%F9 v1.1';

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

first_successful_landing
2015-12-22

- It was obtained by selecting the minimum of date where landing_outcome is ‘Success (ground pad)’
- %sql select min(date) as first_successful_landing from SPACEX where landing__outcome = 'Success (ground pad)';

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

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

- It is obtained by selecting booster_version where landing_outcome is ‘Success (drone ship)’ and payload_mass_kg_ is between 4000 and 6000
- %sql select booster_version from SPACEX where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

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	99
Success (payload status unclear)	1

- Obtained by selecting mission outcome and count of all data grouped by mission_outcome
- %sql select mission_outcome, count(*) as total_number from SPACEX group by mission_outcome;

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- Obtained by selecting booster_version where payload_mass_kg_ equals the value of the subquery of selecting maximum payload_mass_kg
- %sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX);

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

- List the 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
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Obtained by selecting month of date, data, booster_version, launch_site, landing_outcome where landing_outcome is ‘Failure (drone ship)’ and the year of date is 2015
- %sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEX where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

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

- 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
- Obtained by selecting landing_outcome and count of all record where the date is between 2010-06-04 and 2017-03-20 grouped by landing_outcome and ordered descending by the count of outcomes
- %sql select landing__outcome, count(*) as count_outcomes from SPACEX where date between '2010-06-04' and '2017-03-20’ group by landing__outcome order by count_outcomes desc;

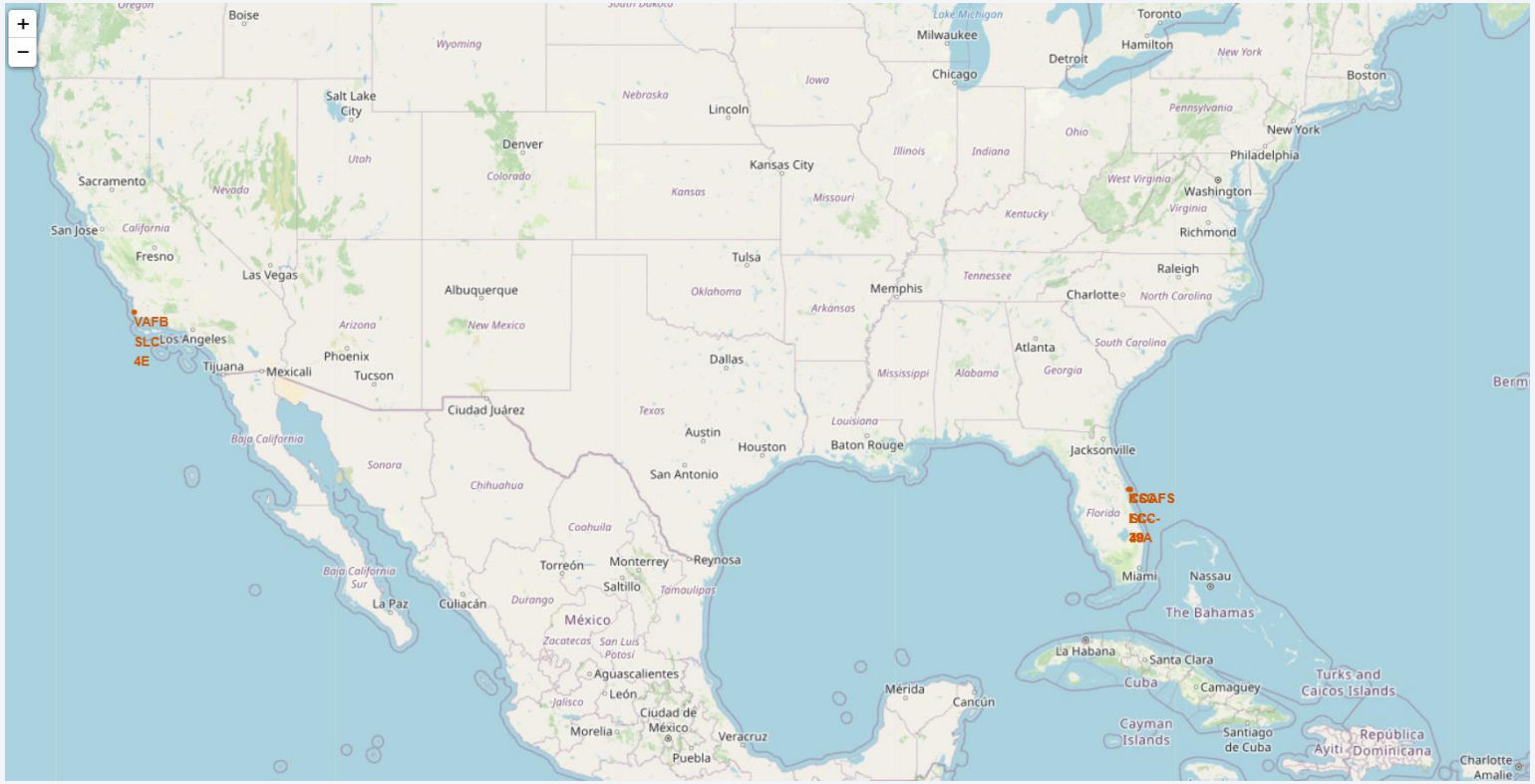
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

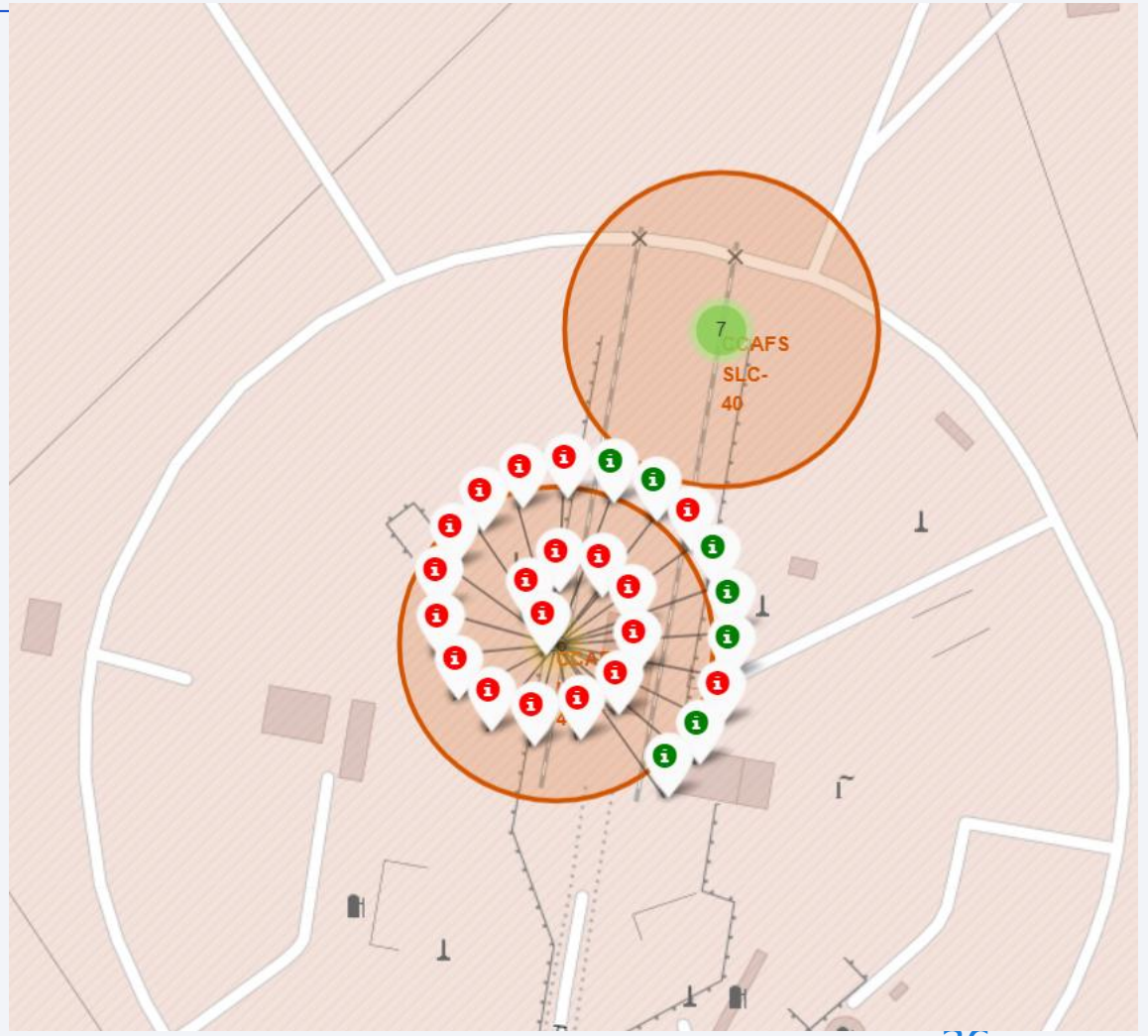
Launch Site Locations

- There are 3 launch Sites (KSC LC-39A, CCAFS SLC-40, CCAFS LC-40) in close proximity to each other near Cape Canaveral, Florida, US
- And another VAFB SLC-4E in Vandenberg Space Force Base in California



Launch sites marked with outcome of launches

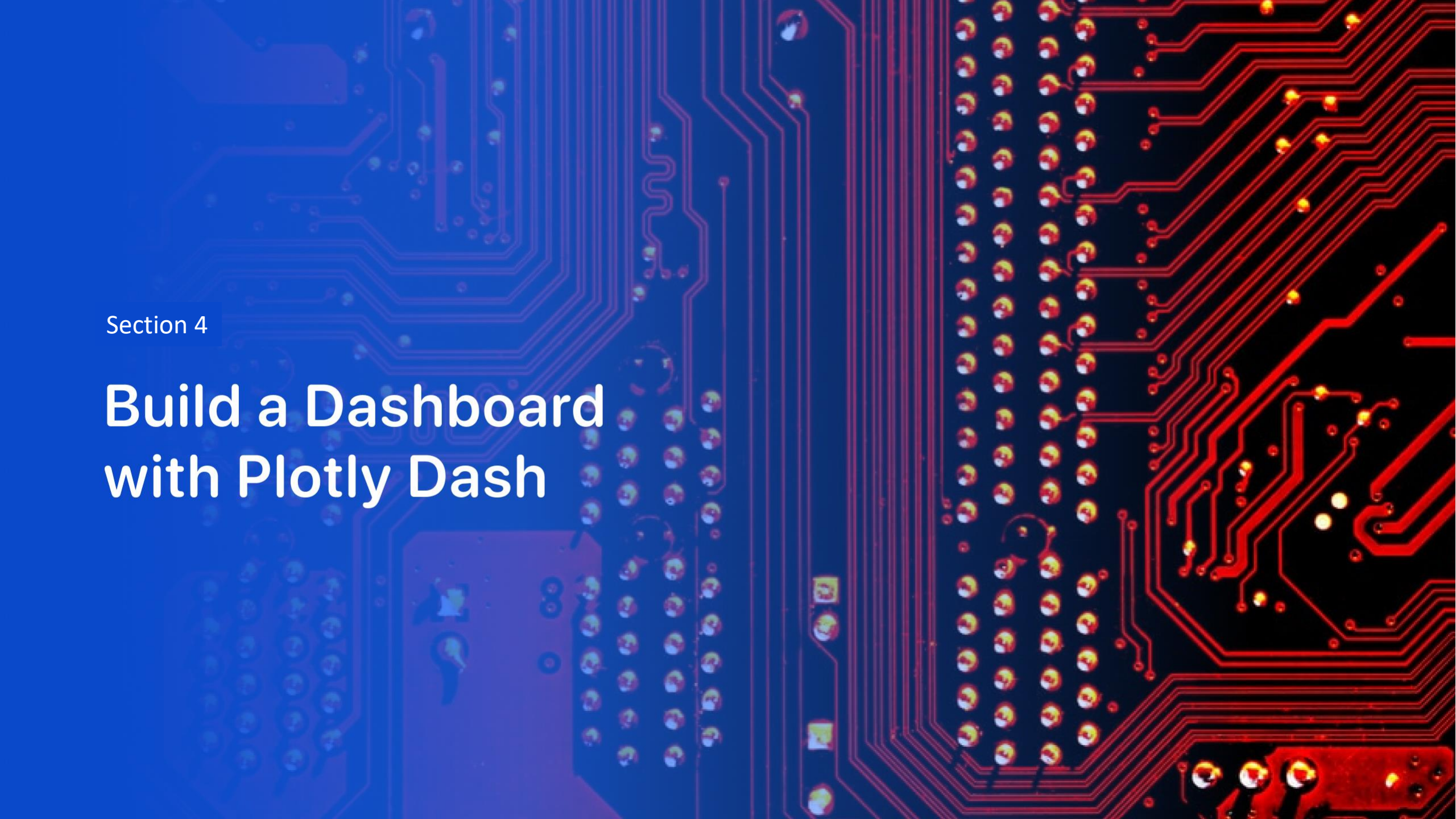
- Green markers mean a record of a successful launch
- Red markers mean a record of an unsuccessful launch
- CCAFS LC-40 launch site has a low success rate



Distance of Launch Site and its Proximites

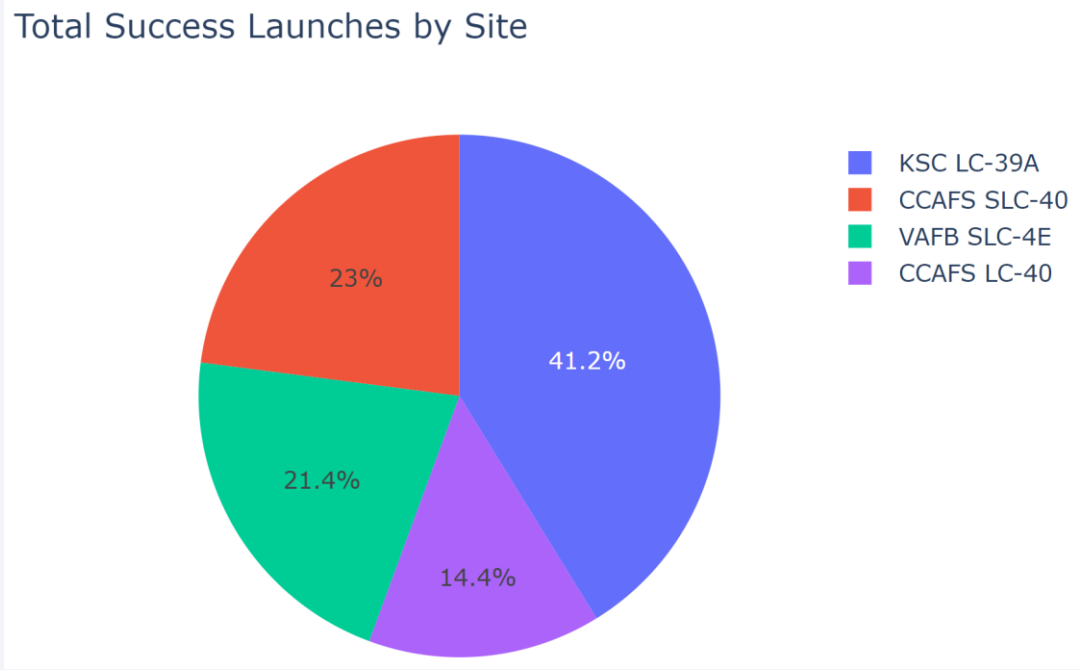
CCAFS SLC-40 launch site is in close proximity of

- railway (1.28 km)
- highway (0.59 km)
- coastline (0.86 km)
- while it is not so close to any city (18.21 km)


Section 4
Build a Dashboard with Plotly Dash

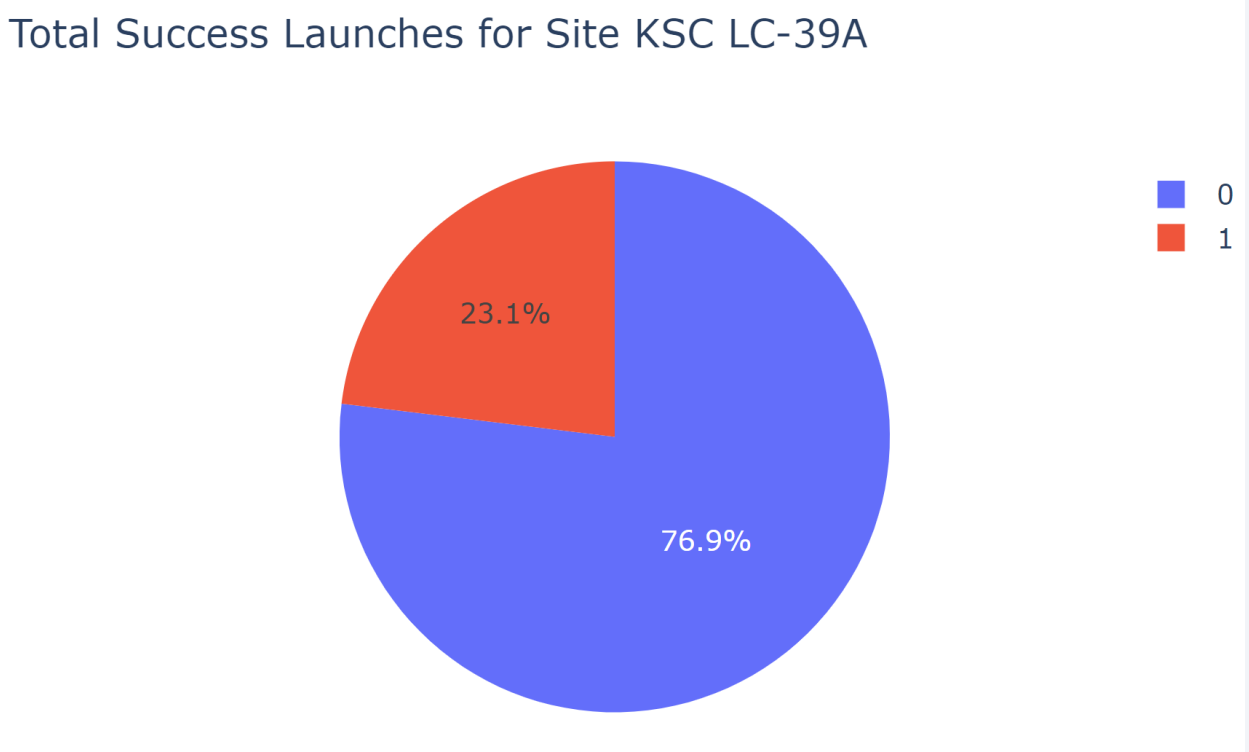
Launch Success Count for All Sites

- KSC LC-39A had the most successful launches followed by CCAFS SLC-40, VAFB SLC-4E and CCAFS LC-40 in this order
- The highest proportion of successful launches from all successful launches is 41.2% at KSC LC-39A, the lowest is 14.4% at CCAFS LC-40



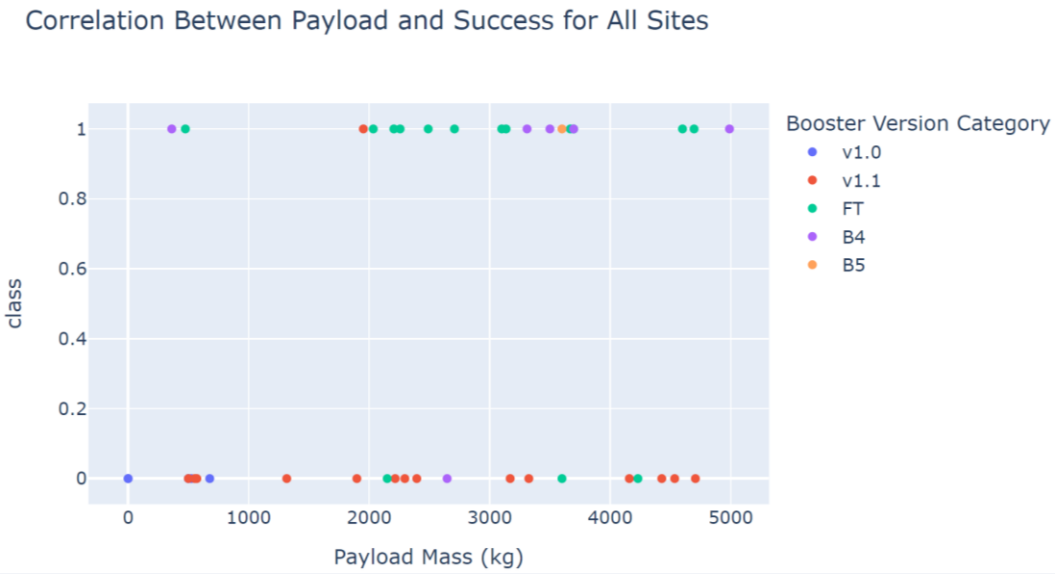
Launch site with highest launch success ratio

- KSC LC-39A launch site had the highest launch success rate (76.9%) with 10 successful and 3 failed launches



Payload Mass vs. Launch Outcome for all sites

- For payload mass above 5000 kg the B4 booster had the highest launch success rate
- For payload mass below 5000 kg the v1.1 booster had the highest launch success rate



Payload mass range between 0 kg and 5000 kg

Payload mass range between 5000 kg and 10000 kg

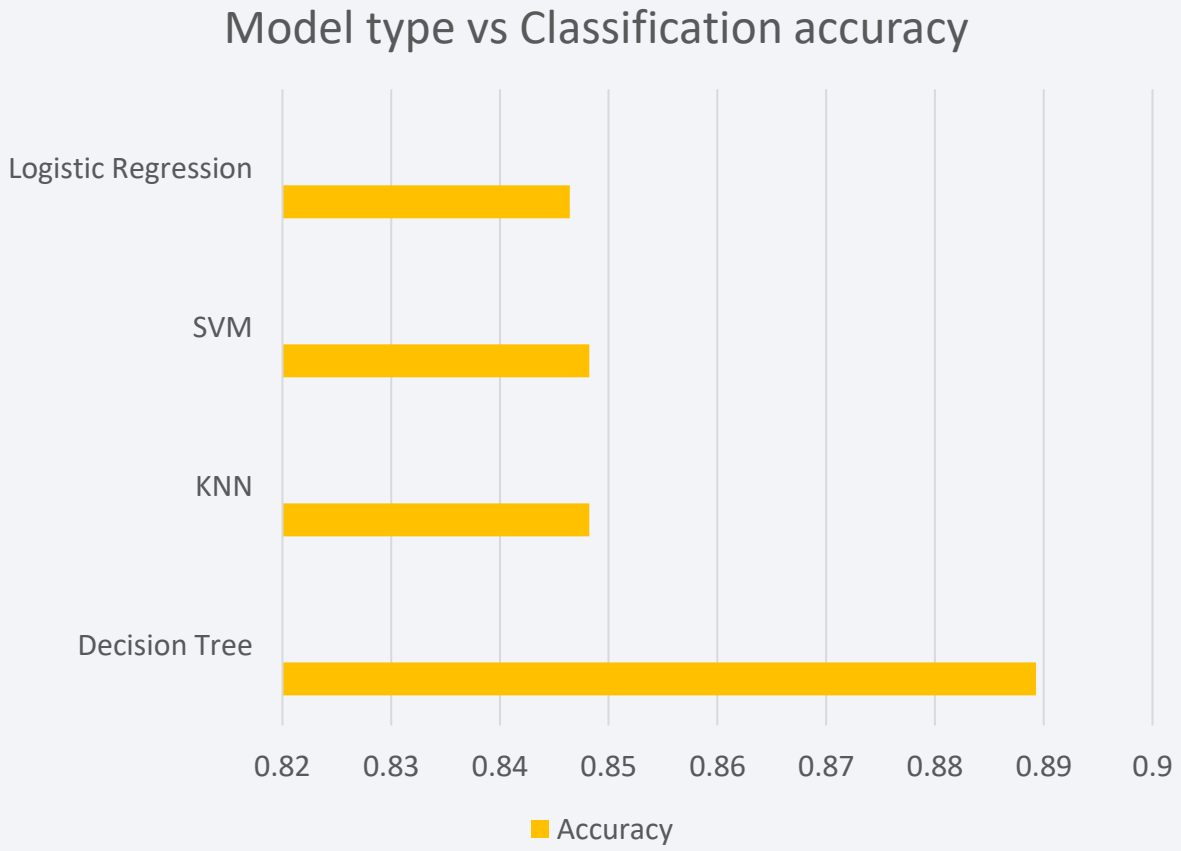
Section 5

Predictive Analysis (Classification)

Decision Tree 0.889286 KNN 0.848214 SVM 0.848214 Logistic Regression 0.846429

Classification Accuracy

- The **Decision Tree Model** has **89%** classification accuracy for Falcon 9 launch outcome prediction, that is 5% higher than SVM, KNN, and Logistic Regression models, which are have an 85% accuracy



Confusion Matrix

- Confusion matrix of the Decision Tree Model (best performing model)

- For launches that landed the model is 100% accurate (12/12)
- For launches that did not land the model is 50% accurate (3/6)

Conclusions

- Higher payload mass results in higher launch success rate
- Launch success rates increased from 2013 to 2020. From 2017 to 2020 the increase was near stagnation.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate
- KSC LC-39A had the most successful launches of any sites
- The Decision tree classifier is the best machine learning algorithm for launch success rate prediction

Thank you!