IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pavlo Tienin
30/03/2024

Outline

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

Executive Summary

Summary of Methodologies

The research attempts to identify the factors for a successful rocket landing. To make this investigation, the following methodologies where used:

- Collecting data using SpaceX REST API and the web scraping methods
- Wrangling data for creation success/fail outcome variable
- Exploring data with data visualization techniques, considering the following factors: payload, launch site, flight number and the yearly dynamics
- Analyzing data with the help of SQL, calculating the following statistics: total payload, payload range for successful launches, and total number of successful and failed outcomes
- Exploring launch site success rates and proximity to geographical markers
- Visualizing the launch sites with the most success and successful payload ranges
- Building Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Summary of All Results

Exploratory Data Analysis:

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization:

- Most launch sites should be as nearest as possible to the equator, and all are close to the coast

Predictive Analytics:

- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

The project background

SpaceX, the leader in the space industry, has the intentions to make space travelling affordable and easy for everyone. Its accomplishments include sending spacecraft to the ISS, launching the satellite internet connections and sending missions to space. SpaceX can do it because the rocket launches are relatively inexpensive due to its ability to reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost much more. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use available public data and machine learning models to predict whether SpaceX – or any potential competing company – can reuse the first stage.

Problems to solve

- How to predict if the rocket will land successfully
- What parameters (payload mass, launch site, number of flights, orbits etc.) affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

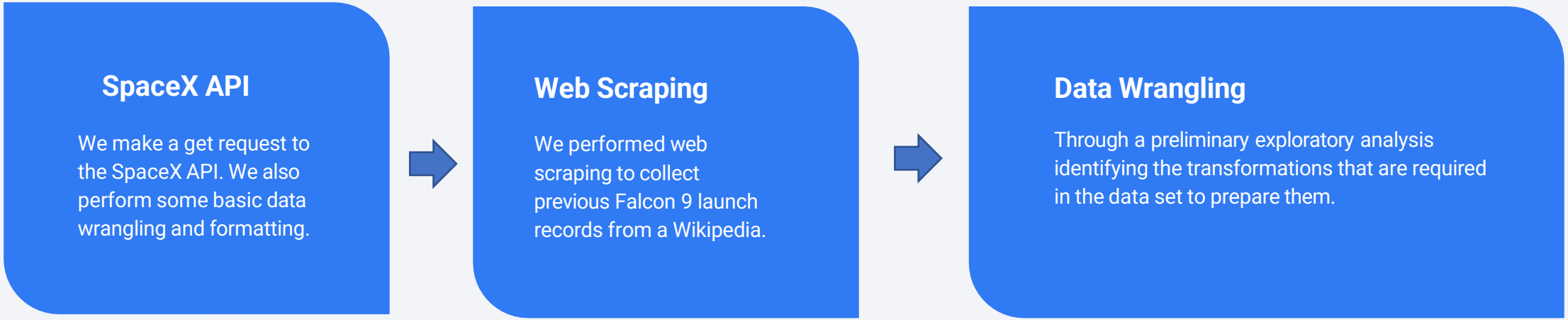
Methodology

Methodology

- Data collection methodology:
  - Data was collecting from previous SpaceX missions. SpaceX API
  - Web scraping from Wikipedia
- Perform data wrangling:
  - Calculated the number of launches on each site
  - Calculated number and occurrences of each orbit
  - Calculated the number and occurrence of mission outcome per orbit type
  - Created a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Dash.
- Perform predictive analysis using classification models.
  - Tuning and evaluating the models to find best decision and parameters

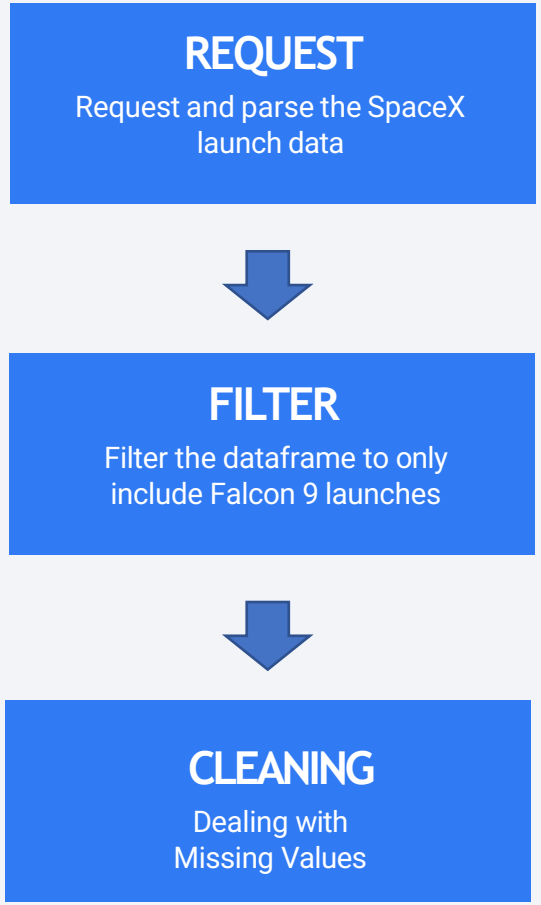
Data Collection

- Data sets were collected from previous SpaceX mission and Wikipedia pages and below processes were obtained to Filter, clean and Transform the data to prepare for Modeling.



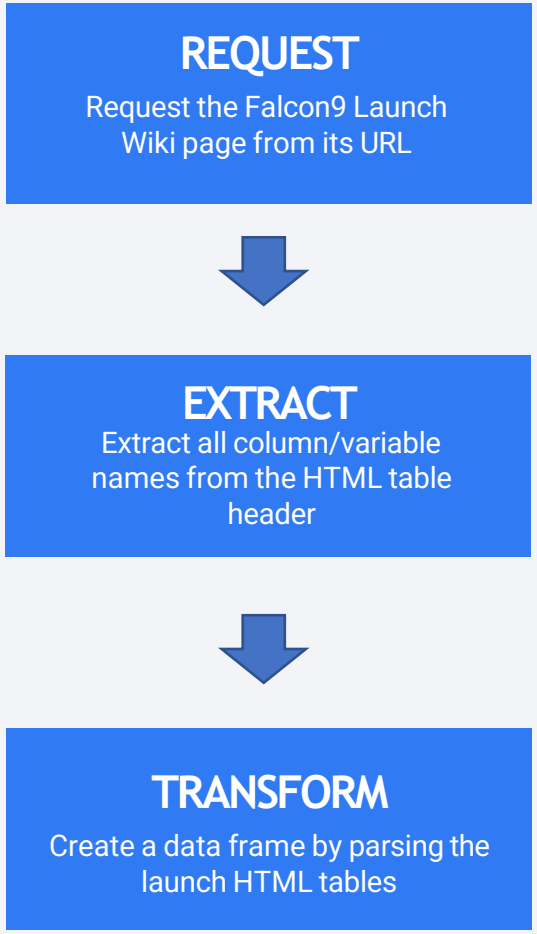
Data Collection – SpaceX API

- We make a get request to the SpaceX API. We also perform some basic data wrangling and formatting.
- We make filtering to leave in the dataset only Falcon 9 cases and make a replacing of the missing values.
- *It can be seen in detail by the following GitHub Link.*



Data Collection - Scraping

- We performed web scraping to collect historical Falcon 9 launch records from a Wikipedia page titled "*List of Falcon 9 and Falcon Heavy launches*".
- We made a request for that web page, then extracted the columns we needed with the help of BeautifulSoup object from HTML response and created a special dataframe from that HTML tables.
- *It can be seen in detail by the following GitHub link.*



Data Wrangling

- Through a preliminary exploratory analysis identifying the transformations that are required in the data set to prepare them.
- Calculating:
  - number of launches for each site
  - number and occurrence of orbit
  - number and occurrence of mission outcome per orbit type
- Creating binary landing outcome column (dependent variable) with "1" will mean the rocket landed successfully, and "0" means it was unsuccessful.

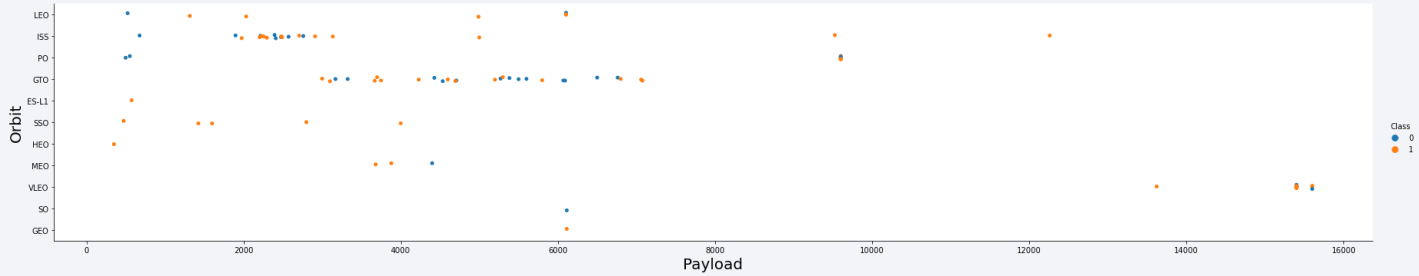
Landing was not always successful:

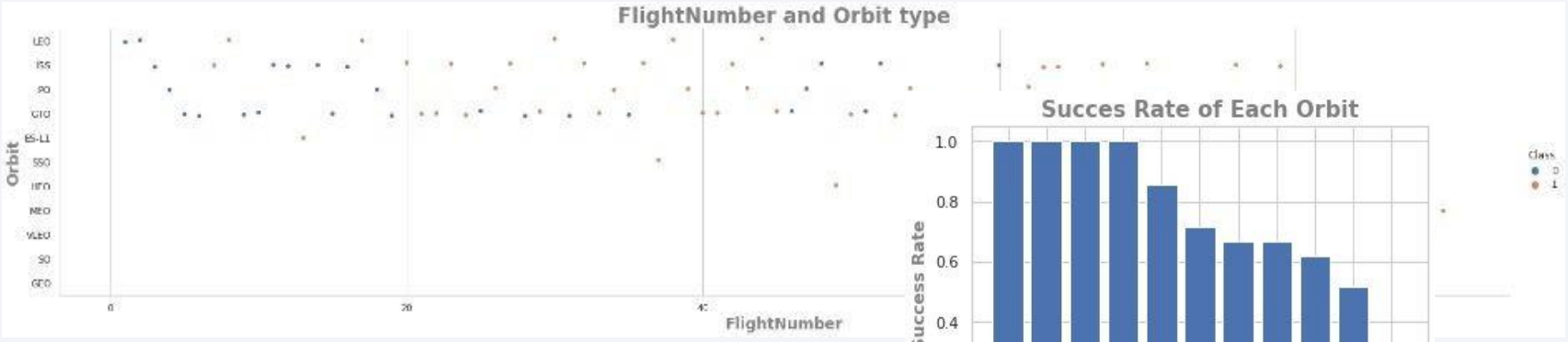
- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean
- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS:** meant the mission had a successful landing on a ground pad
- **False RTLS:** represented an unsuccessful landing on a ground pad
- **True ASDS:** meant the mission outcome had a successful landing on a drone ship
- **False ASDS:** represented an unsuccessful landing on drone ship

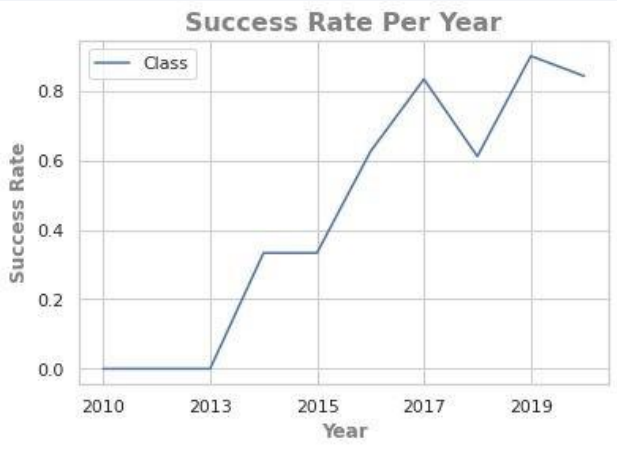
- *It can be seen in detail by the following GitHub link.*

EDA with Data Visualization

- Exploratory Data Analysis to visualize the relationship between:
  - Flight Number and Launch Site.
  - Payload and Launch Site.
  - Success rate of each Orbit Type.
  - Flight Number and Orbit Type.
  - Payload and Orbit Type.
  - Visualizing the launch success yearly dynamics.







- *It can be seen in detail by the following [GitHub link](#).*

EDA with SQL

- SQL queries performed:
    - Names of the unique launch sites in the space mission
    - Top 5 launch sites whose name begin with the string 'CCA'
    - Total payload mass carried by boosters launched by NASA (CRS)
    - Average payload mass carried by booster version F9 v1.1
    - Date when the first successful landing outcome in ground pad was achieved
    - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
    - Total number of successful and failure mission outcomes
    - Names of the booster versions which have carried the maximum payload mass
    - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

- *It can be seen in detail by the following GitHub link.*

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps.
- Indications of each element:

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

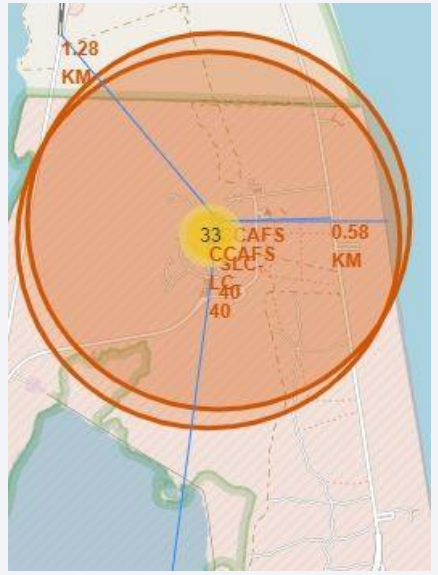
Colored Markers of Launch Outcomes

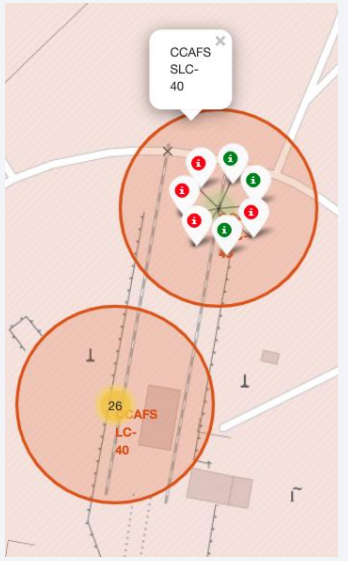
- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

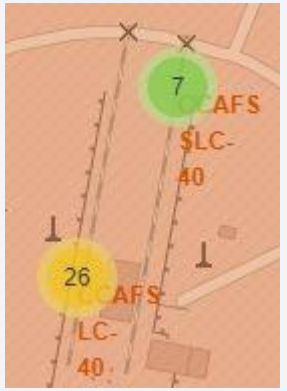
Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

- *It can be seen in detail by the following [GitHub link](#).*

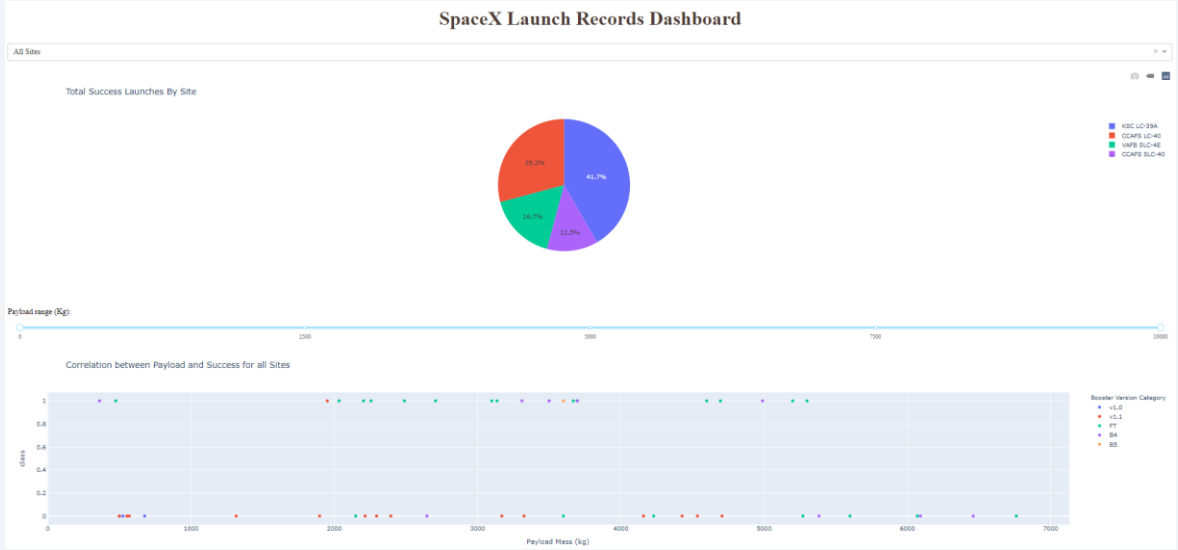






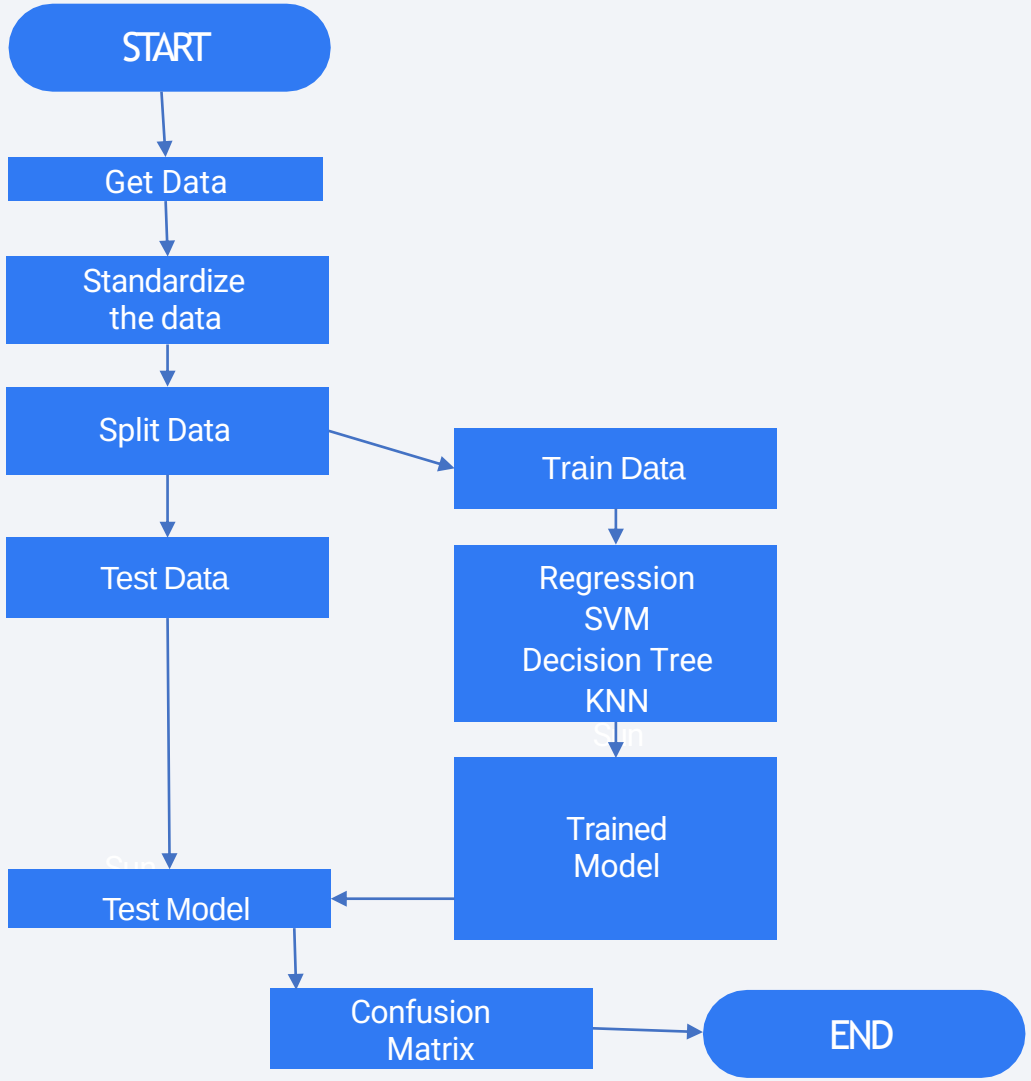
Build a Dashboard with Plotly Dash

- Elements of the dashboard:
  - Dropdown list for the launch site.
  - Slider for selecting the payload mass.
  - PieChart: for showing the success rate of each launch site, or showing the number of successful landing outcomess.
  - Scatterplot: Show success/failure by payload and booster version.
- Key findings from the dashboard:
  - Which site has the largest successful launches? KSC LC-39A.
  - Which site has the highest launch success rate? KSC LC-39A (success rate 76.9%).
  - Which payload range(s) has the highest launch success rate? 2000-4000.
  - Which payload range(s) has the lowest launch success rate? 6000-8000.
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? B5 (only one successful start), apart from that FT (15 successes, 8 failures).



- *It can be seen in detail by the following link.*

Predictive Analysis (Classification)



Steps

- Creating NumPy array from the Class column
- Standardizing the data with StandardScaler. Fit and transform the data.
- Splitting the data using train_test_split
- Creating a GridSearchCV object with cv=10 for parameter optimization
- Applying GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- Calculating accuracy on the test data using .score() for all models
- Assessing the confusion matrix for all models
- Identifying the best model using Jaccard_Score, F1_Score and Accuracy

Results

Exploratory data analysis results

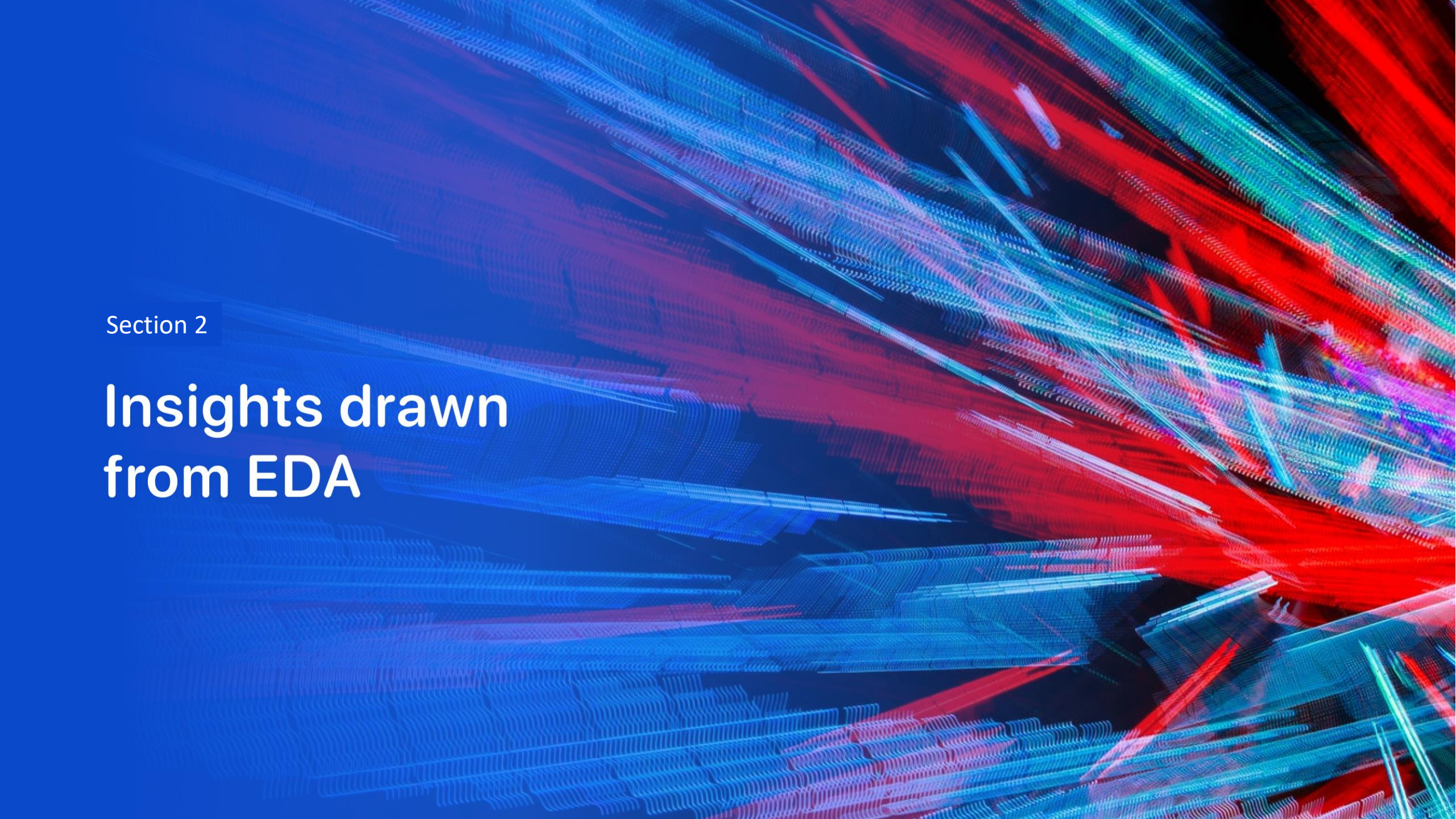
- Launch success rate increases over time
- Higher success rate for higher orbits

Interactive analytics demo in screenshots

- Higher success rate for higher payload mass
- Low success rate for booster versions v1.0, v1.1, high success rate for FT, B4, B5
- Higher success rate for Kennedy Space center and recent starts at Cape Canaveral

Predictive analysis results

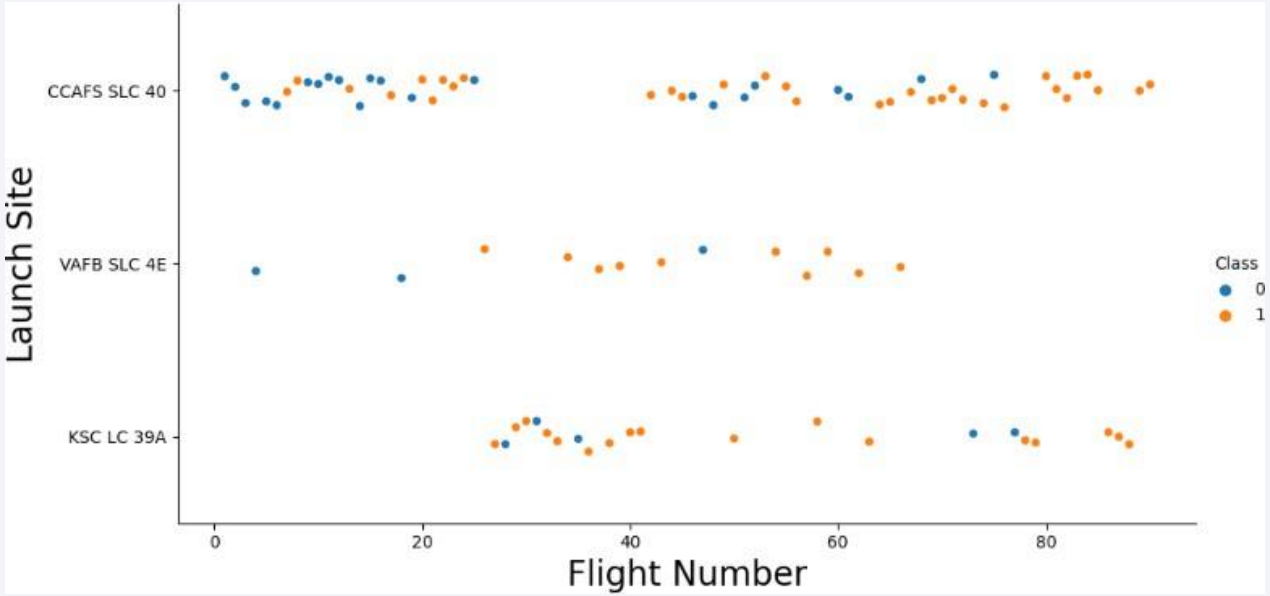
- Best prediction results with Logistic Regression and Support Vector Machine


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Summary:

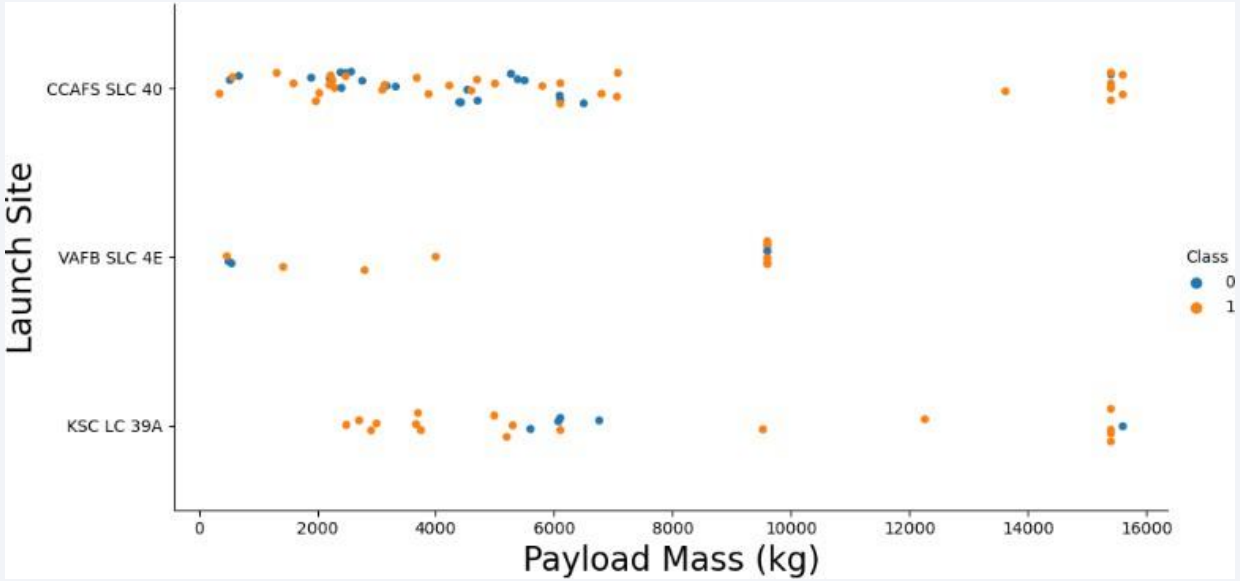
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can say that new launches have a higher success rate



Payload vs. Launch Site

Summary

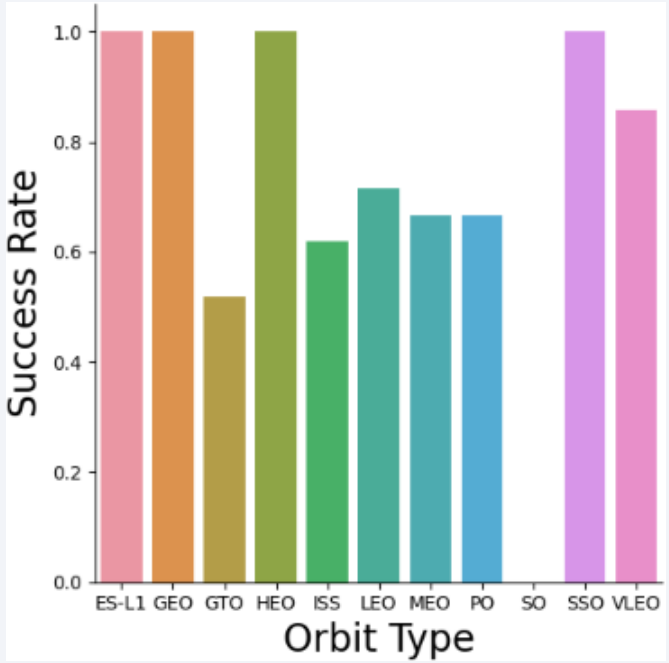
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Summary

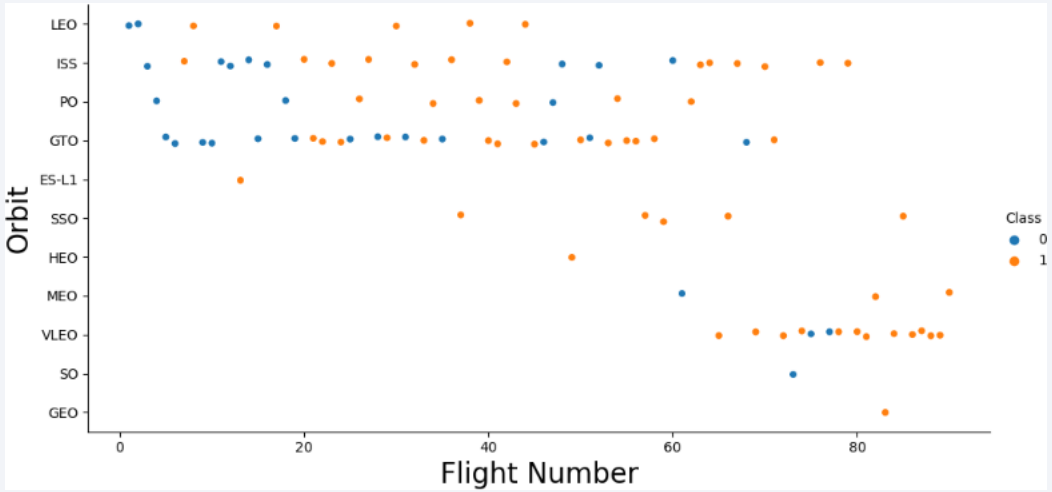
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Summary

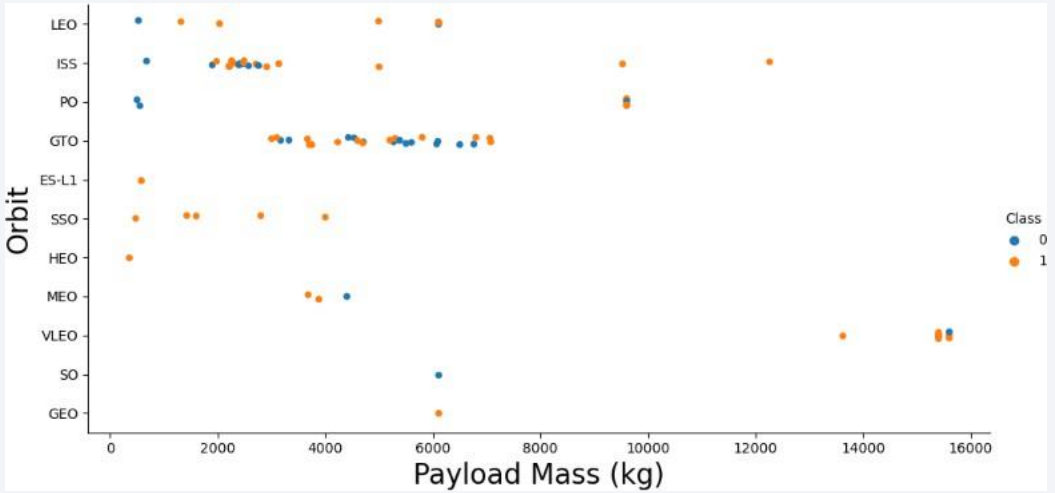
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Summary

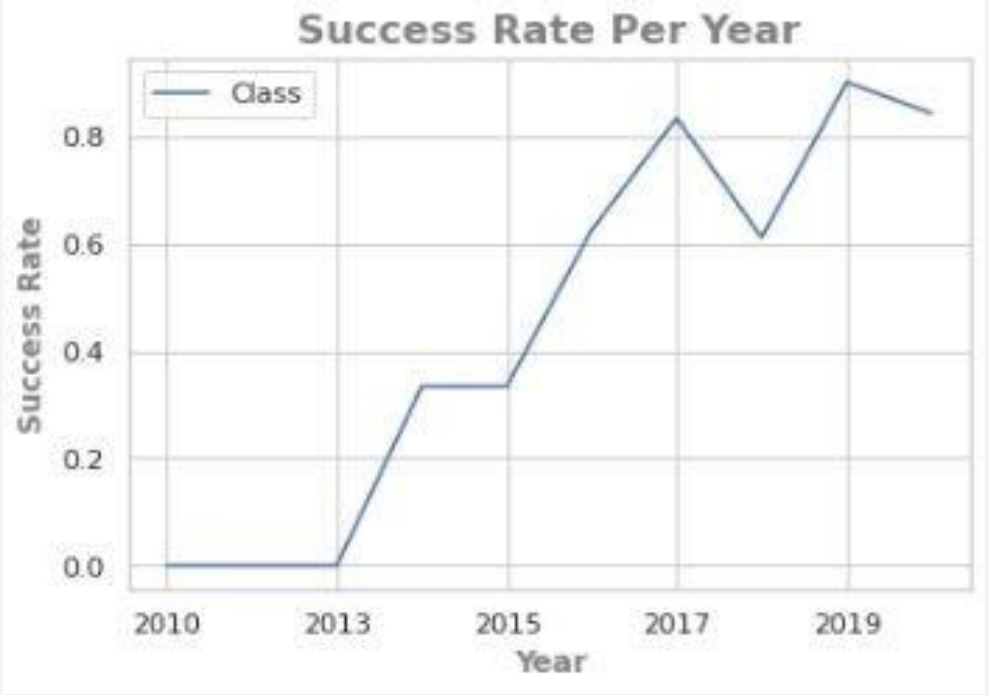
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

Summary

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

SQL query: %sql select UNIQUE(LAUNCH_SITE) from SPACEXTBL;

Launch Sites Overview

[30]:

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

CCAFS Cape Canaveral Space Launch Complex

KSC Kennedy Space Center Launch Complex

VAFB Vandenberg Space Launch Complex

Launch Site Names Begin with 'CCA'

SQL query: %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;

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

Total Payload Mass

SQL query: %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE Customer = 'NASA (CRS)';

Total Payload Mass

- **45 596 kg** (total) carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

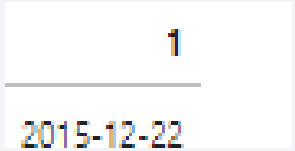
SQL query: %sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.0%;

Average Payload Mass

- **2,928 kg** (average) carried by booster version F9 v1.1

First Successful Ground Landing Date

SQL query: %sql SELECT MIN(Date) FROM SPACEXDATASET WHERE Landing__Outcome = 'Success (ground pad)';



1st Successful Landing in Ground Pad

- 22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL query: %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

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

Successful Booster Drone Ship Landing

- Successfully landed booster versions greater than 4000 but less than 6000: B1021.2, B1031.2, B1022, B1026

Total Number of Successful and Failure Mission Outcomes

SQL query: %sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;

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

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

Boosters Carried Maximum Payload

SQL query: %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

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

Carrying Maximum Payload:

- 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

%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing__Outcome = 'Failure (drone ship)' AND YEAR(DATE) = 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)

2015 Launch Records

- These two launches were unsuccessful in 2015

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

%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY TOTAL_NUMBER DESC

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

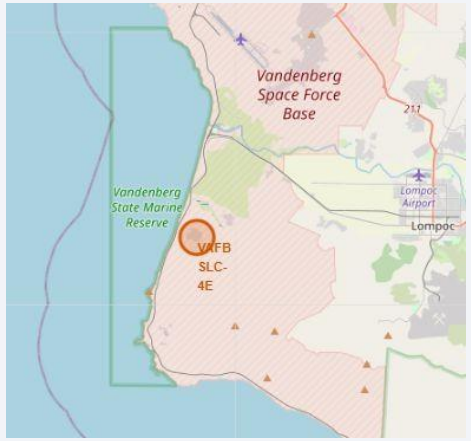
Ranked outcome

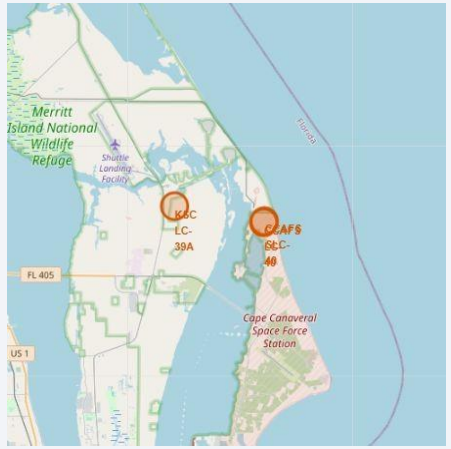
- Most of the landing outcomes were successful

Section 3

Launch Sites Proximities Analysis

Launch Sites





- The launch sites are located at the East and West coast, near the southernmost US mainland area:

CCAFS - Cape Canaveral Space Launch Complex
KSC - Kennedy Space Center Launch Complex
VAFB - Vandenberg Space Launch Complex

Locations analysis

- Nearest points to equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for an orbit.

- Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.

Landing Success by the Launch Site

Vandenberg Space Launch Complex

VAFB SLC-4E

40.00% Success

Kennedy Space Center Launch Complex

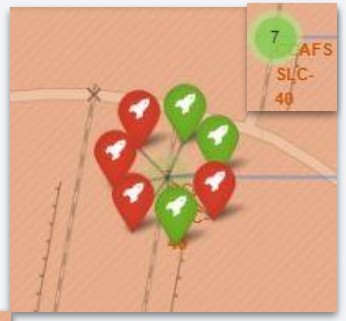
KSC LC-39A

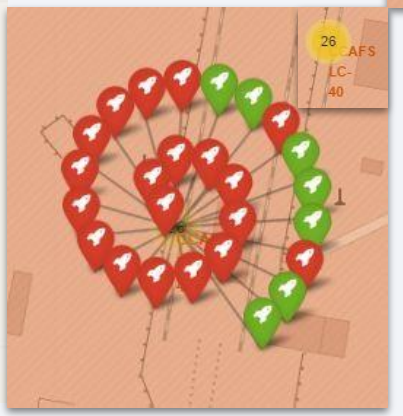
76.92% Success

Cape Canaveral Space Launch Complex

CCAFS SLC-40

42.85% Success



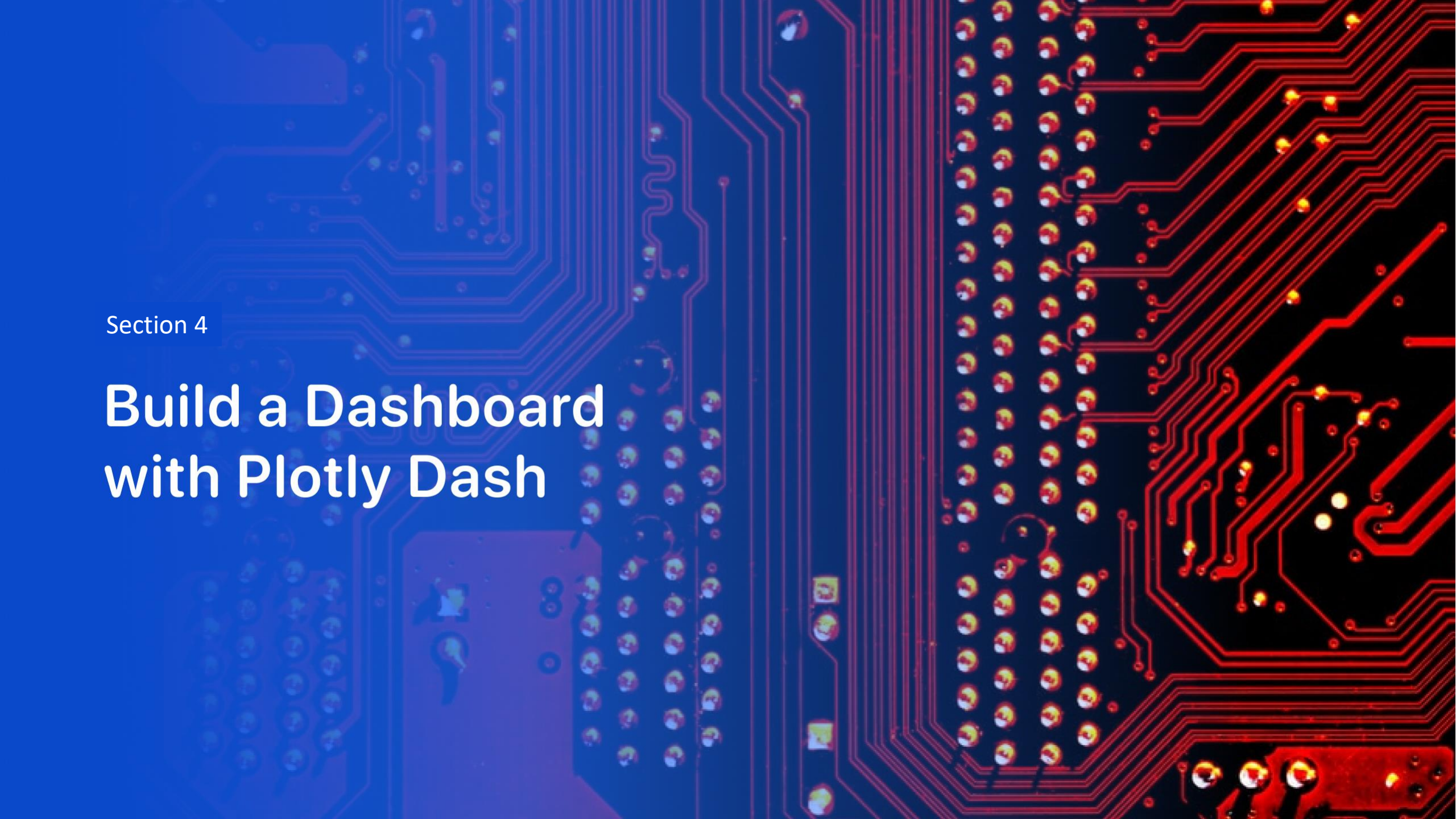


CCAFS LC-40

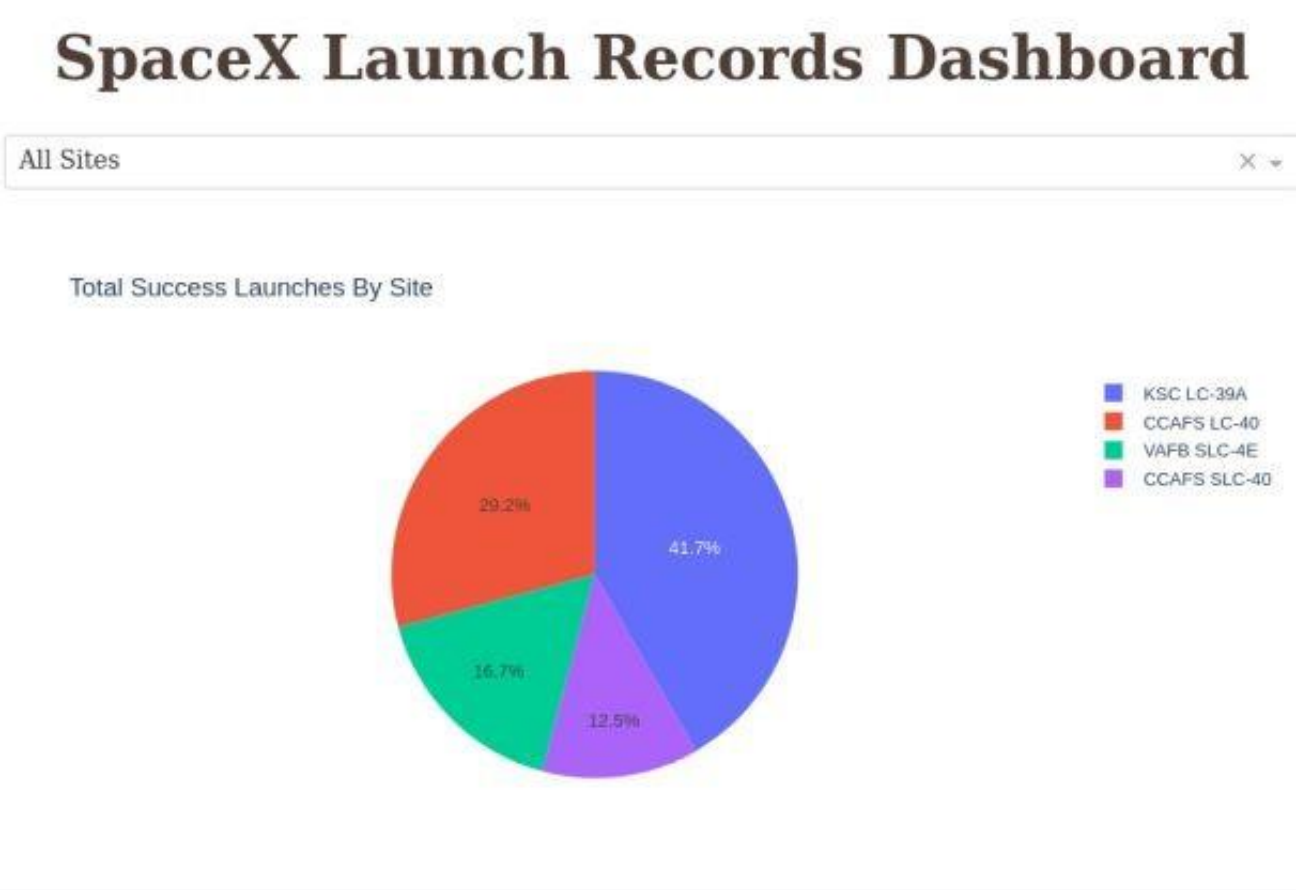
26.92% Success

Distance to Proximities

- The launch site CCAFS SLC-40 has good logistics aspects:
- It is located near railroads/roads/coastline which simplifies all the supplies
- It is relatively far from inhabited areas that minimizes possible risks.

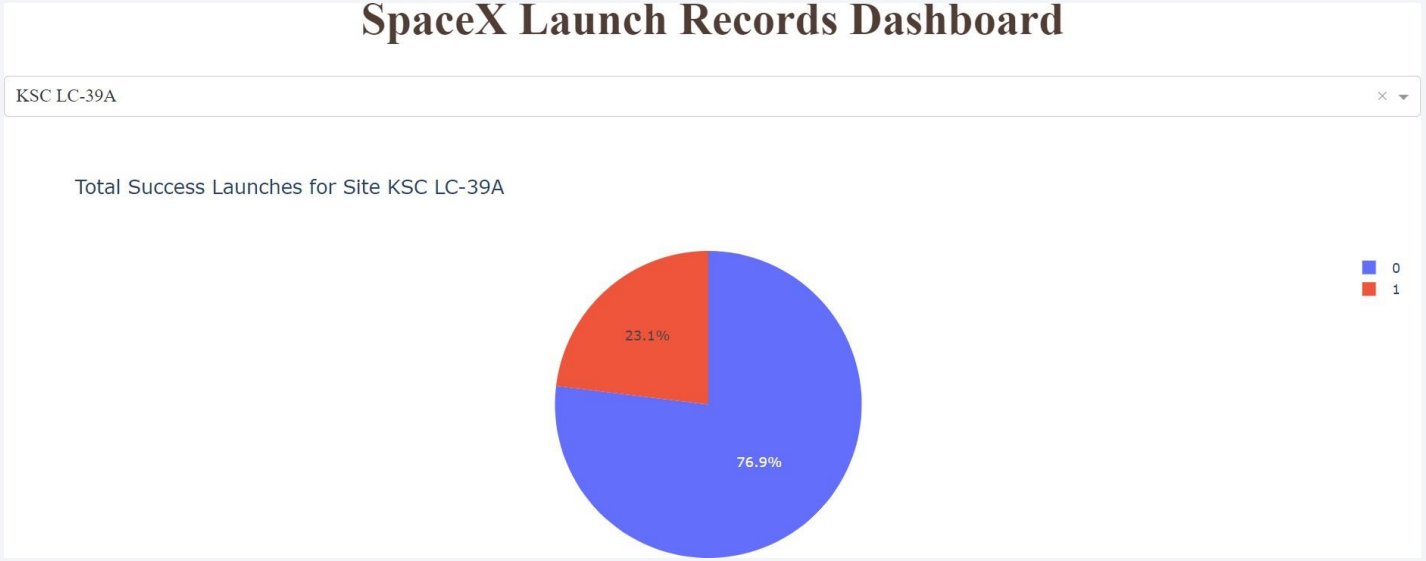

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



- Kennedy Space Center (KSC LC-39A) hast the most successful stage-1 landings (more than 41%)

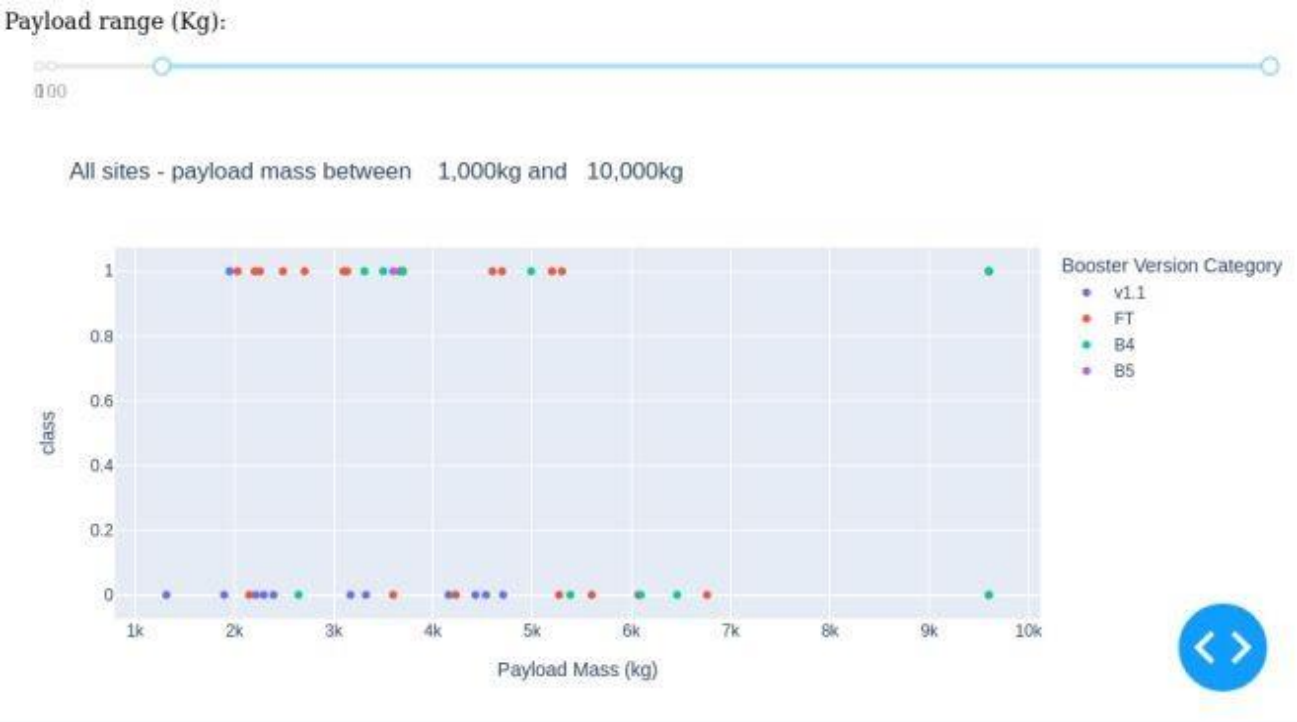
Highest launch success



Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches

Payload vs Launch Outcome

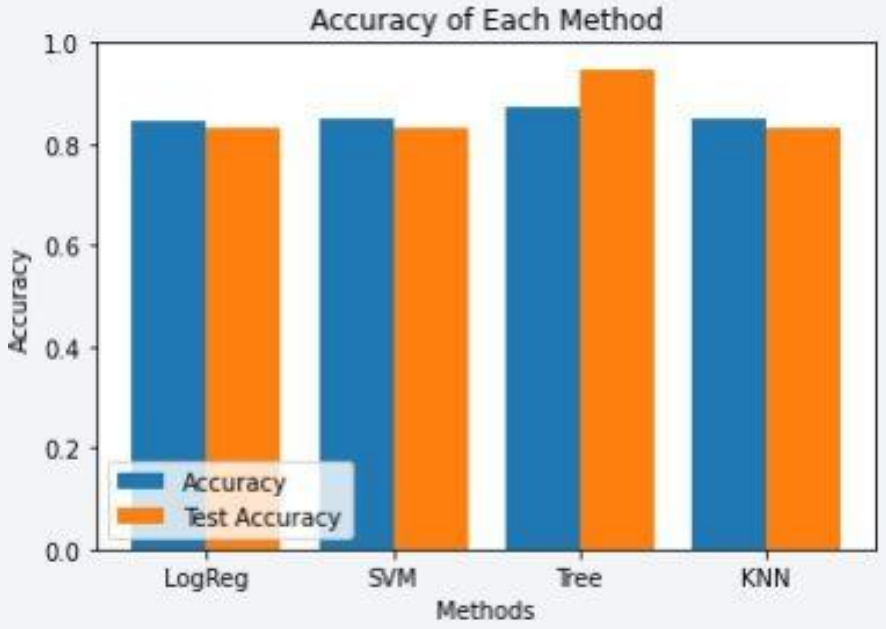


- Payloads between 2000 kg and 6000 kg have the highest success rate
- It's better to use FT boosters for the launch

Section 5

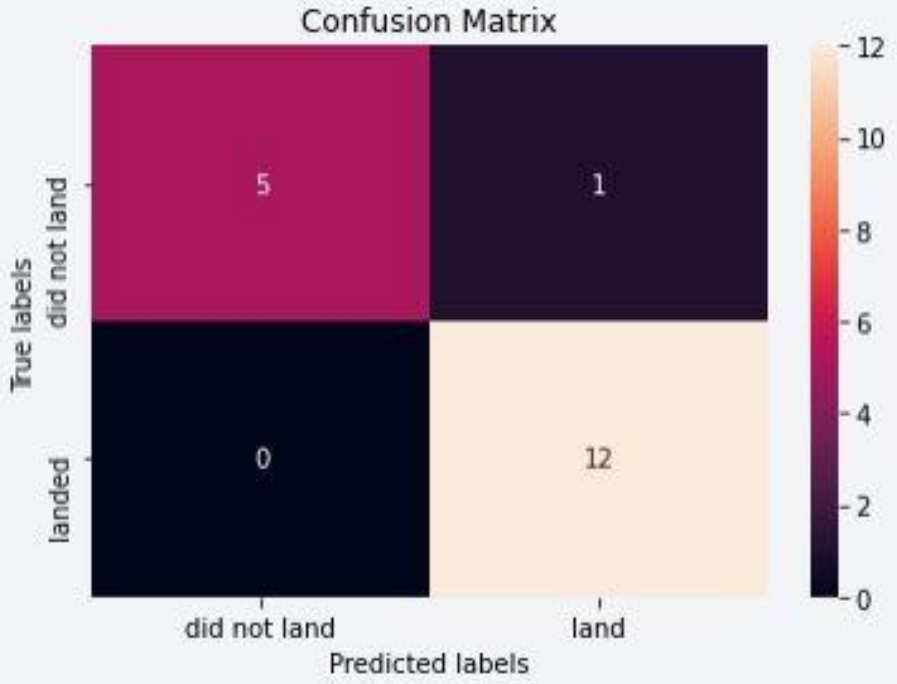
Predictive Analysis (Classification)

Classification Accuracy



- All 4 classification models were tested, and they all have very similar accuracies (please see the chart on the left)
- The model with the highest classification accuracy is Decision Tree.

Confusion Matrix



Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 5 True negative
  - 1 False positive
  - 0 False Negative
- **Precision** = TP / (TP + FP)
  - 12 / 13 = .92
- **Recall** = TP / (TP + FN)
  - 12 / 12 = 1
- **F1 Score** = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.92 * 1) / (.92 + 1) = .958
- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = .944

Conclusions

- Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming
- Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- Coast: All the launch sites are close to the coast
- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

- The folder with all the needed Python code snippets, SQL queries and Notebooks outputs that were created during this project are located here

Thank you!