Winning Space Race with Data Science

Antonio Salgado

2023-06-10

Outline

Executive Summary

- **Available data were analyzed with different methodologies**
  - Web scraping and SpaceX API, in the Data Collection first step;
  - Data wrangling, data visualization and interactive visual analytics, in the EDA second step;
  - Machine Learning, in the Prediction last step. Decision Tree algorithm (83% accuracy) chosen as predictive tool.

- **Summary**
  - Valuable data were collected from different public sources;
  - Best features to predict launchings success were identified with EDA;
  - Factors considered: payload, launch site, orbit type, etc;
  - Model to predict and manage this opportunity was developed using Machine Learning Prediction.

Introduction

- **Objective**

  - Analyze the convenience (or not) of creating the new start up: SpaceY
  - Consider the current SpaceX market dominance as key factor to overcome
  - 62M$ SpaceX launch cost (reusing first stage) vs 165M$ competitors cost

- **Key points to address**

  - Launches cost estimation, based on first stage successful landing
  - Best place for launching
  - Predict launching success to bid against SpaceX

Section 1

Methodology

Summary

- **Data Collection**
  - ML model building through calls to SpaceX API
  - Web Scraping was performed using Wikipedia

- **Data wrangling**
  - Data uniformed tagging with landing results and device summary

- **Exploratory data analysis** (EDA) using visualization and SQL

- **Interactive visual analytic**s using Folium and Plotly Dash

- **Predictive analysis** using classification models
  - Data collected normalized and divided in training and test data sets
  - Evaluated by four different classification models
  - Accuracy evaluated with different parameters combinations

Data Collection – SpaceX API / Web Scraping

- Data sets collected using web scraping technics in SpaceX API and Wikipedia

  - **SpaceX offers an API with data to be used;**
  - API used as per attached flowchart and data is persisted.

Request API and parse SpaceX launch data
- **Filter data to only include Falcon 9 launches**
- **Deal with missing values**

  - **SpaceX launches data obtained from Wikipedia;**
  - Request and BeautifulSoup libraries utilized;
  - Static URL converted to Panda DataFrame;
  - Data downloaded as per attached flowchart

Request the Falcon 9 Launch Wiki page
- **Extract column / variables from HTML table header**
- **Create data frame parsing launch tables**

  - Source Codes: Capstone-Project-Antonio-Salgado/antonio-salgado-jupyter-labs-spacex-data-collection-api.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub
  - Capstone-Project-Antonio-Salgado/antonio-salgado-jupyter-labs-webscraping.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub

Data Wrangling

- **Exploratory Data Analysis** on database;

- **Data gathering**
  - Summary launches per site, occurrences per orbit, mission outcome per orbit;

- **Data checking**
  - Missing values, data type in columns, launches per site, counts per orbit type;

- **Landing outcome label** from outcome column

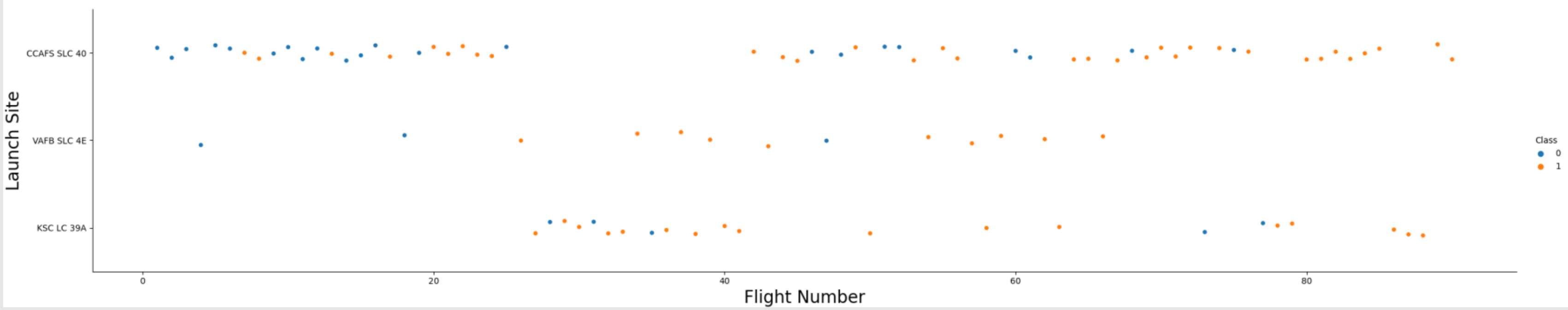
- [Capstone-Project-Antonio-Salgado/antonio-salgado-labs-jupyter-spacex-data_wrangling_jupyterlite.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub]

EDA with Data Visualization

- **First stage**. Exploratory Data Analysis on database

- **Second stage**. Data gathering
  - Summary launches per site
  - Occurrences of each orbit
  - Mission outcome per orbit

- **Last stage**. Landing outcome label from outcome column



- [Capstone-Project-Antonio-Salgado/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub]

EDA with SQL

- **Queries performed**

  - Unique launch sites names;
  - Top 5 launch sites starting with 'CCA';
  - Total payload mass in NASA boosters (CRS);
  - Average payload mass in booster F9 v1.1;
  - First successful landing outcome in ground pad date;
  - Names of successful boosters with payload mass 4000 / 6000 kg;
  - Number of successful and failure mission;
  - Boosters with maximum payload mass names;
  - 2015 failures in drone ship, booster versions, and launch site names;
  - Landing outcomes rank (Failure or Success) between 2010 and 2017.

- Capstone-Project-Antonio-Salgado/antonio-salgado-eda-sql.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub

Interactive Map with Folium

- **Functions utilized in Folium:**
  - **Markers,** indicating map points;
  - **Circles**, indicating specific areas around coordinates;
  - **Marker clusters**, indicating groups of events in each coordinate;
  - **Lines;** indicating distances between coordinates.

- [Capstone-Project-Antonio-Salgado/antonio-salgado-lab_jupyter_launch_site_location.jupyterlite.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub]

Dashboard with Plotly Dash

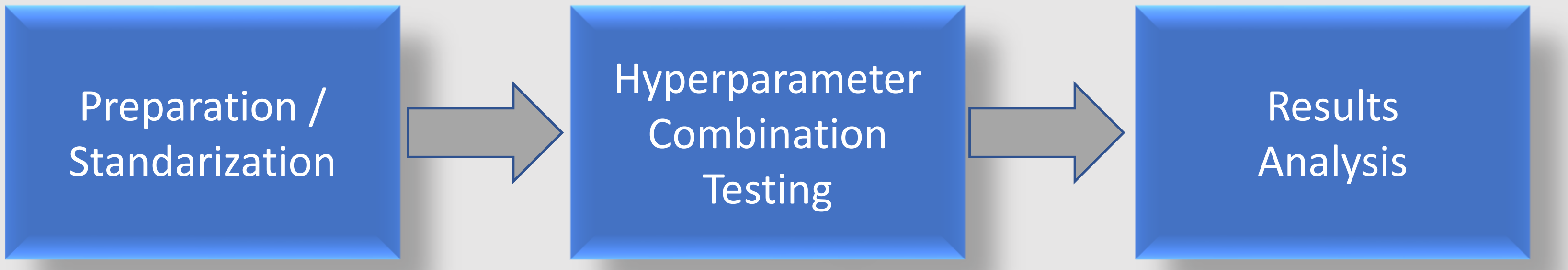
- **Graphs and plots** were used to visualize data
  - **Percentage of launches** by site
  - **Payload range**
- They are key to:
  - **Analyze relation** between payloads and launch sites
  - **Identify best place** to launch, considering payload

[Capstone-Project-Antonio-Salgado/Plotly Dash 2.png at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub]

(7 charts)

Predictive Analysis

- **Classification models** compared:
    - **Logistic regression;**
    - **Support vector machine;**
    - **Decision tree;**
    - **K nearest neighbors**.



[Capstone-Project-Antonio-Salgado/antonio-salgado-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite1.ipynb at main · Antoniosalgado208/Capstone-Project-Antonio-Salgado · GitHub]

Results

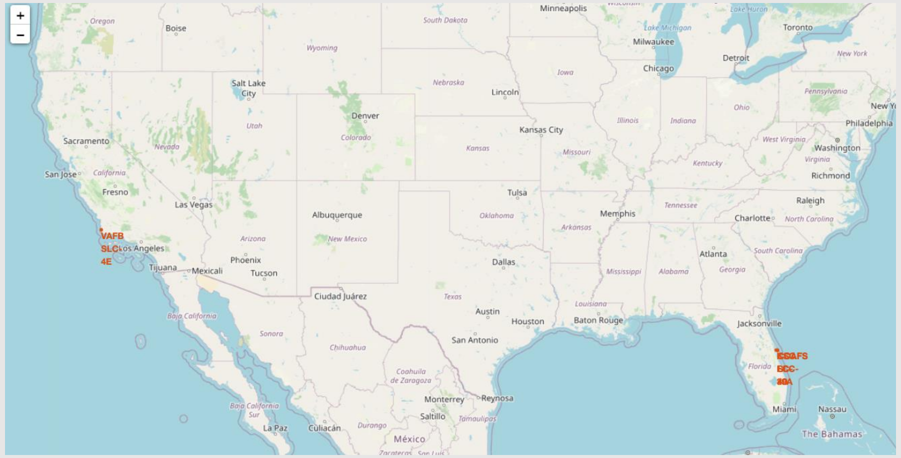
- **EDA Results**
  - Utilized for **different launch sites**;
  - F9 v1.1 booster **average payload**: 2,928kg;
  - **First successful landing** in 2015;
  - **Successful booster landing** in drone ships;
  - **Mission outcomes succ**ess near 100%;
  - **Few booster versions failed**: B1012 and B1015;
  - **Landing outcomes** substantial improvement since then.

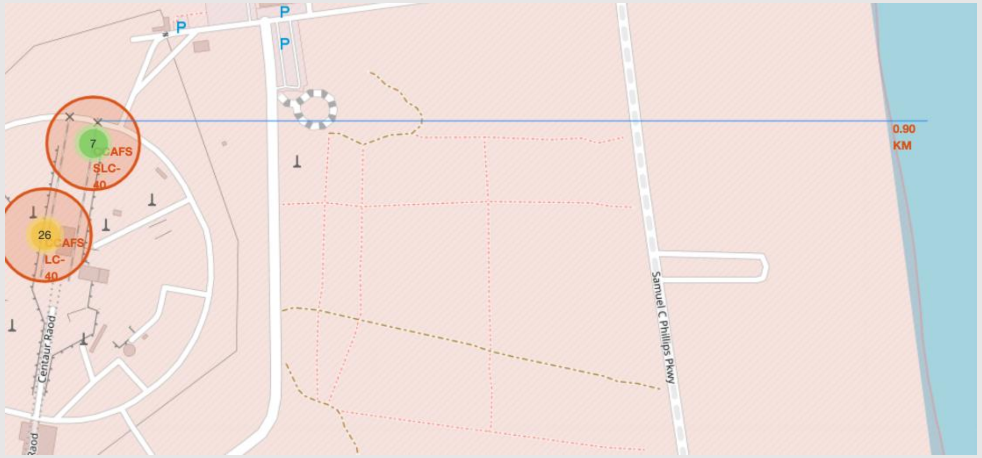
Results

- **EDA Results**

  - Interactive analysis helps **identifying safety launching places**, near sea and with important infrastructure;

  - Most launches took place on **USA east coast.**





Section 2

Insights drawn From EDA

EDA with SQL

Task 1
Display the names of the unique launch sites in the space mission

- sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1

Task 2
Display 5 records where launch sites begin with the string 'CCA'

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

Task 3
Display the total payload mass carried by boosters launched by NASA (CRS)

- sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%'

EDA with SQL

Task 4

Display average payload mass carried by booster version F9 v1.1

- sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

Task 5

List the date when the first successful landing outcome in ground pay was achieved

- sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)'

EDA with SQL

Task 7
List the total number of successful and failure mission outcomes

- sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

Task 8
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

- sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

Task 9
List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

- sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;

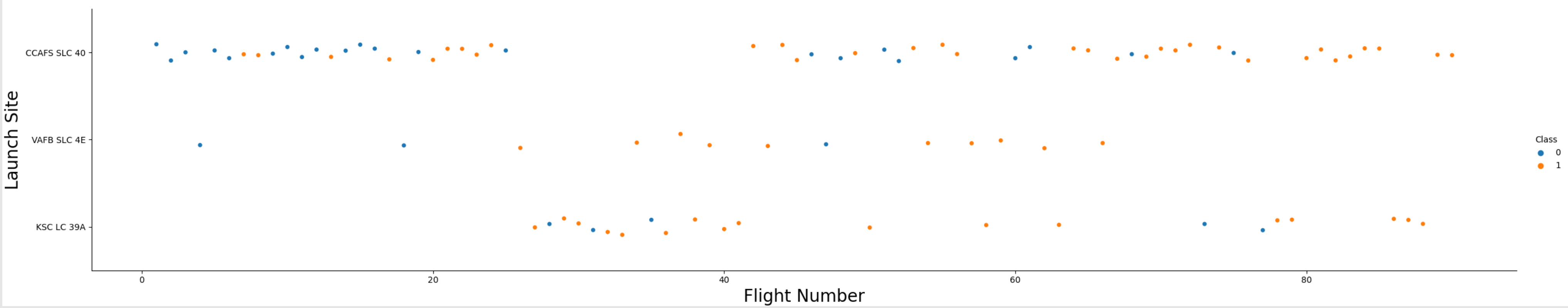
EDA with SQL

Task 10
Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

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

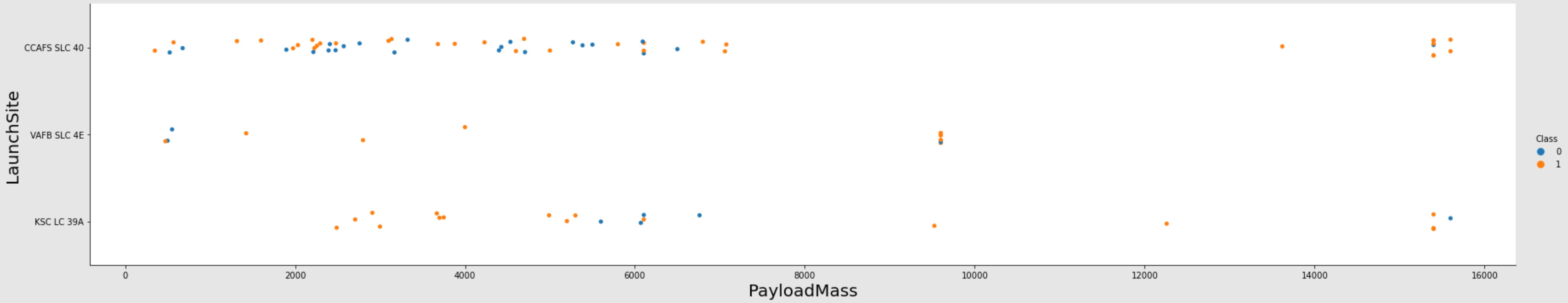
Flight Number vs Launch Site

- **Best launch site**: CCAF5 SLC 4O. Most recent launches successful;
- **Second**: VAFB SLC 4E. Third: KSC LC 39A;
- **Success Rate** improvement over time.



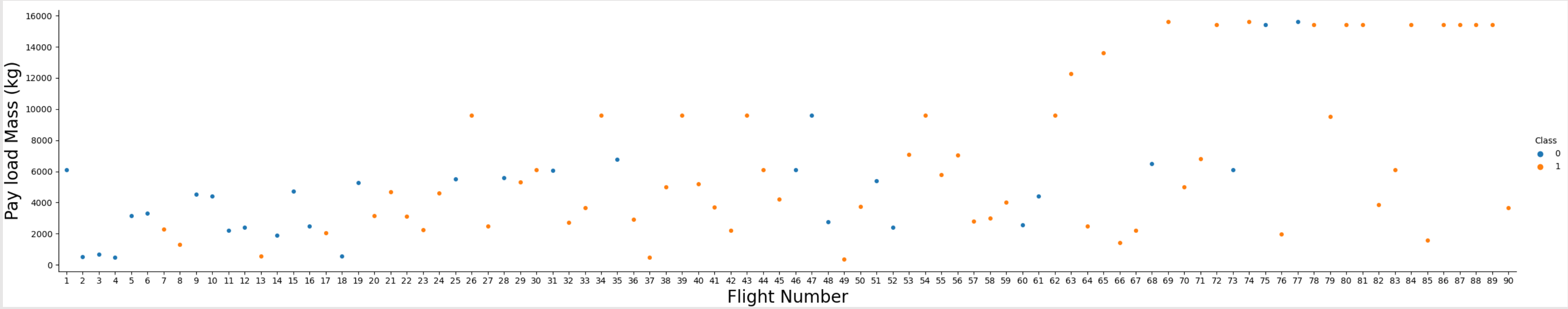
Payload vs Launch Site

- **Excellent success rates** Payloads over 9,000+kg;
- **Payloads over 12,000kg only possible** in CCAFS SLC 40 and KSC LC 39A Launch sites



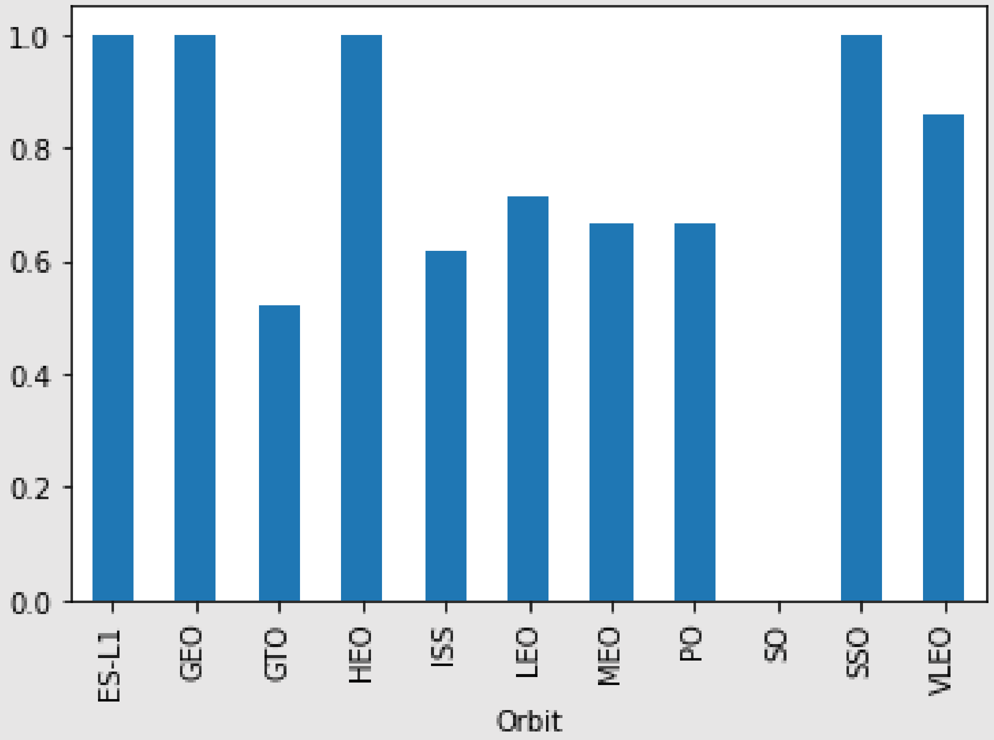
Payload vs Flight Number

- When **flight number increases**, the first stage is more likely to **land successfully**;
- The **more massive the payload, the less likely** the first stage will return.



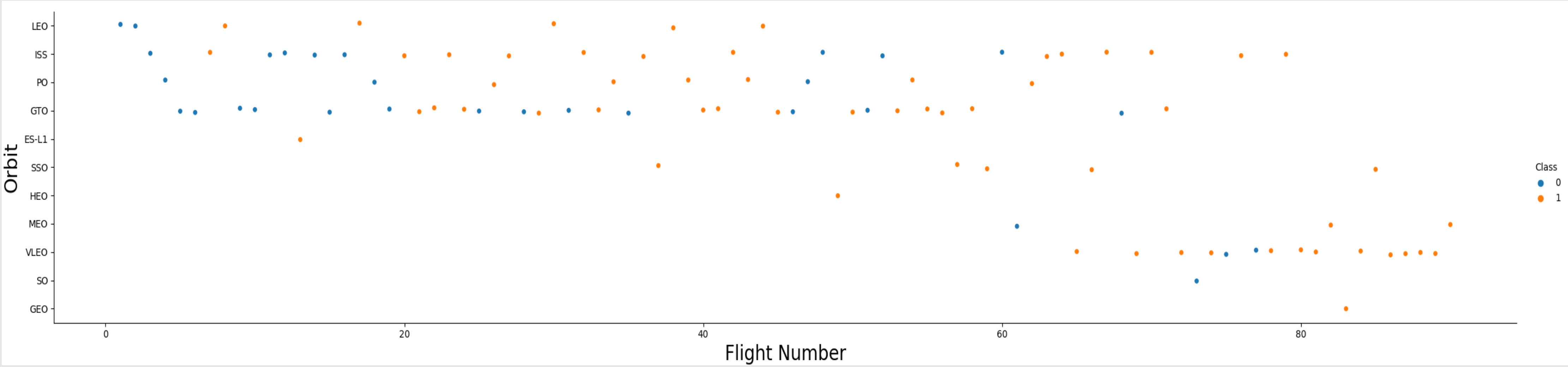
Success Rate vs Orbit Type

- **Very high success rates on orbits**:
  - ES L-1;
  - GEO;
  - HEO;
  - SSO;
  - VLEO.



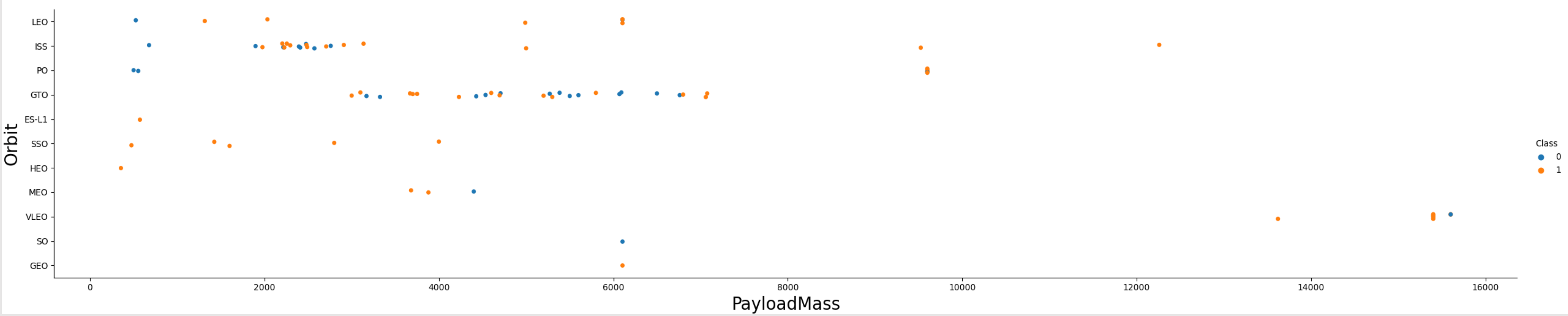
Flight Number vs Orbit Type

- **Success rate improvement** in all orbits over time;
- **Opportunity: VLEO**. Latest frequency increase.



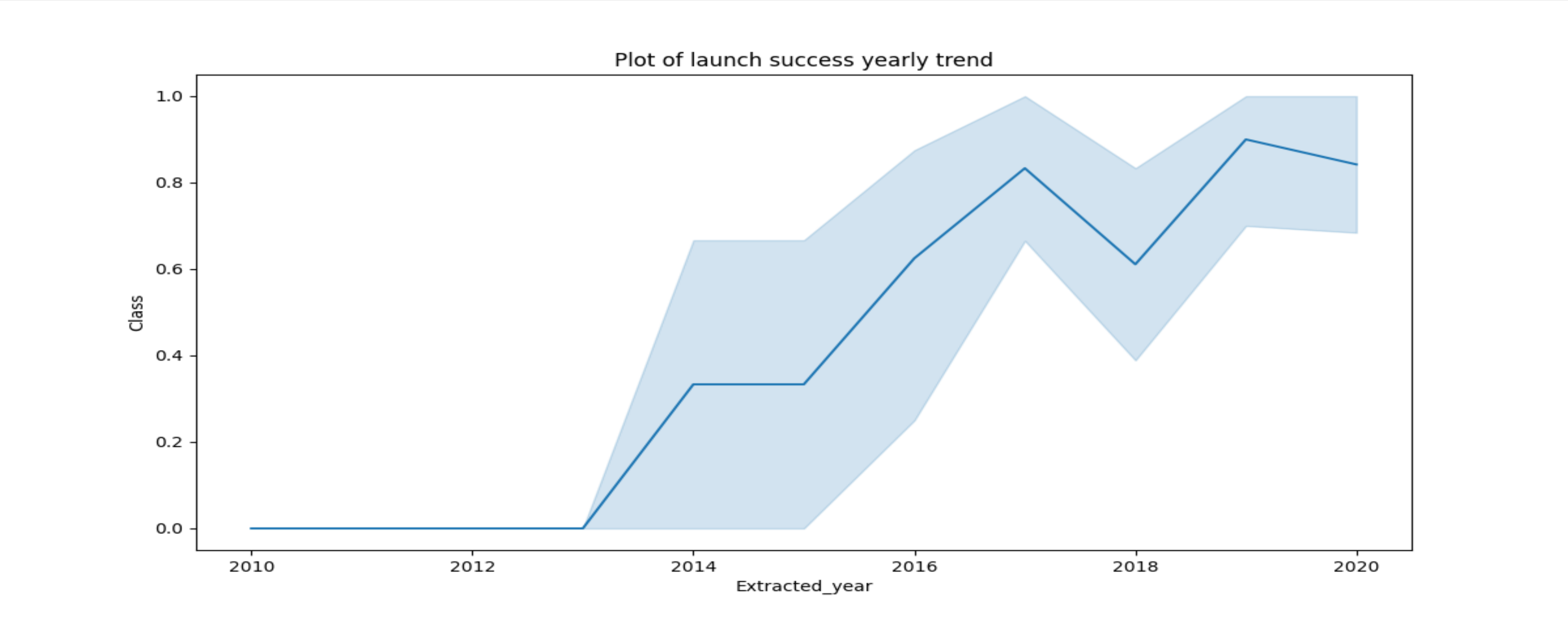
Payload vs Orbit Type

- **No relationship** founded in GTO payload vs success rate;
- **ISS: payload widest rate and success good rate;**
- SO and GEO: few launches.



Launch Success Yearly Trend

- **Remarkable success rate increase** during 2013 /2020 period;
- **Technology improvement** during 2010 /2015 period.



All Launch names

- Process to get them: select unique “launch_site” values from dataset;
- Launch names are as follows:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch sites names starting with “CCA”

- They are five launches from Cape Cañaveral
- Launch sites are the following:

Date	Time	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	0	LEO	SpaceX	Success	Failure (Parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1	0	LEO	NASA (COTS)	Success	Failure (Parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

4.2. Total payload mass

- Total payload mass carried by boosters from NASA;
- Adding all payloads with “CRS” codes

Total Payload (kg)
111.268

Average Payload Mass

- Average payload mass carried by booster version F9 v1.1;
- Filtering and calculating the Avg of above version:

Avg Payload (kg)
2.928

First successful landing date

- Landing outcome on ground pad;

- Filtering and getting the minimum date value:

Minimum Date
2015-12-22

Successful Drone landing. 4000 / 6000 mass

- Successful Booster landing with 4000 / 6000 payload mass;
- Four Booster version meeting above criteria:

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Quantity of Successful / Failure missions

- Number of successful and failure mission outcomes;
- Grouping and counting records for each group:

Mission Outcome	Occurrences
Success	99
Success (unclear status)	1
Failure (in flight)	1

Boosters carried maximum payload

- Booster that carried maximum payload mass, included in database:

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch records

- Drone ship failed landing, booster versions and launch sites:

Booster Version	Launch Site
F9 V1.1 B1012	CCAFS LC-40
F9 V1.1 B1015	CCAFS LC-40

Landing outcomes ranking

- Landing ranking between 2010-06-04 and 2017;
- “No attempt” taken into account

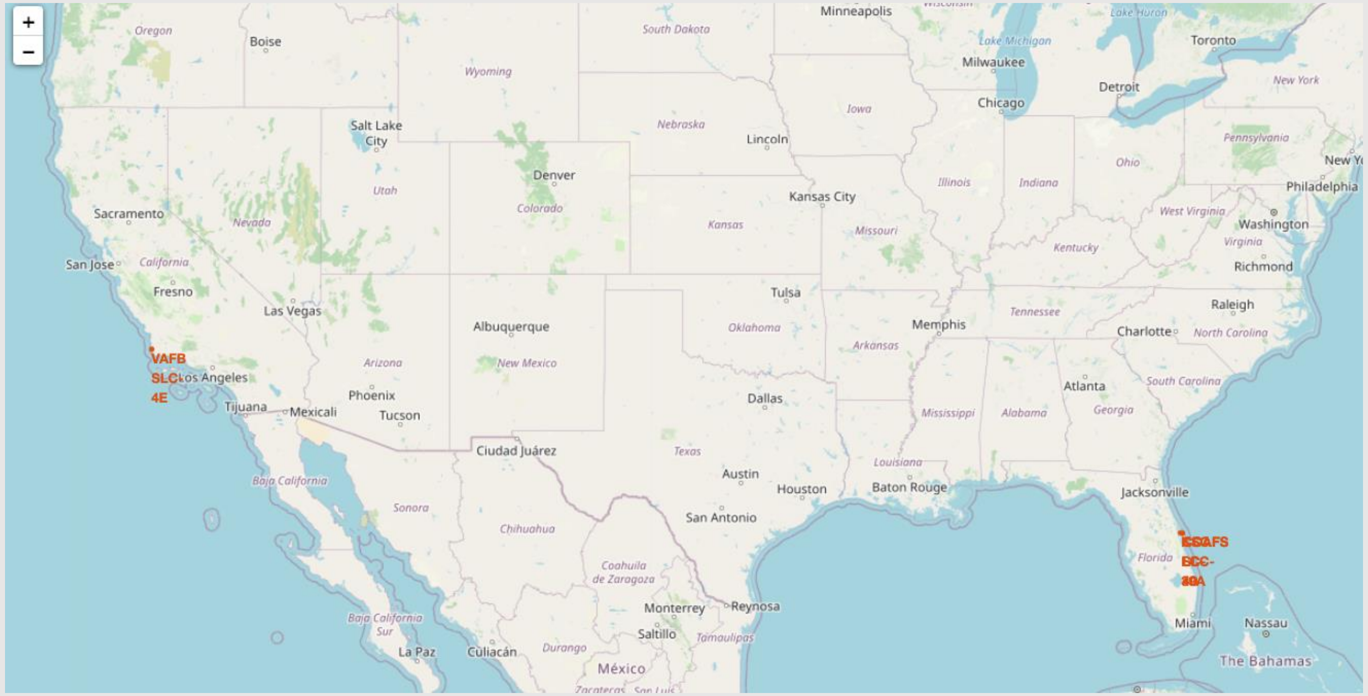
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

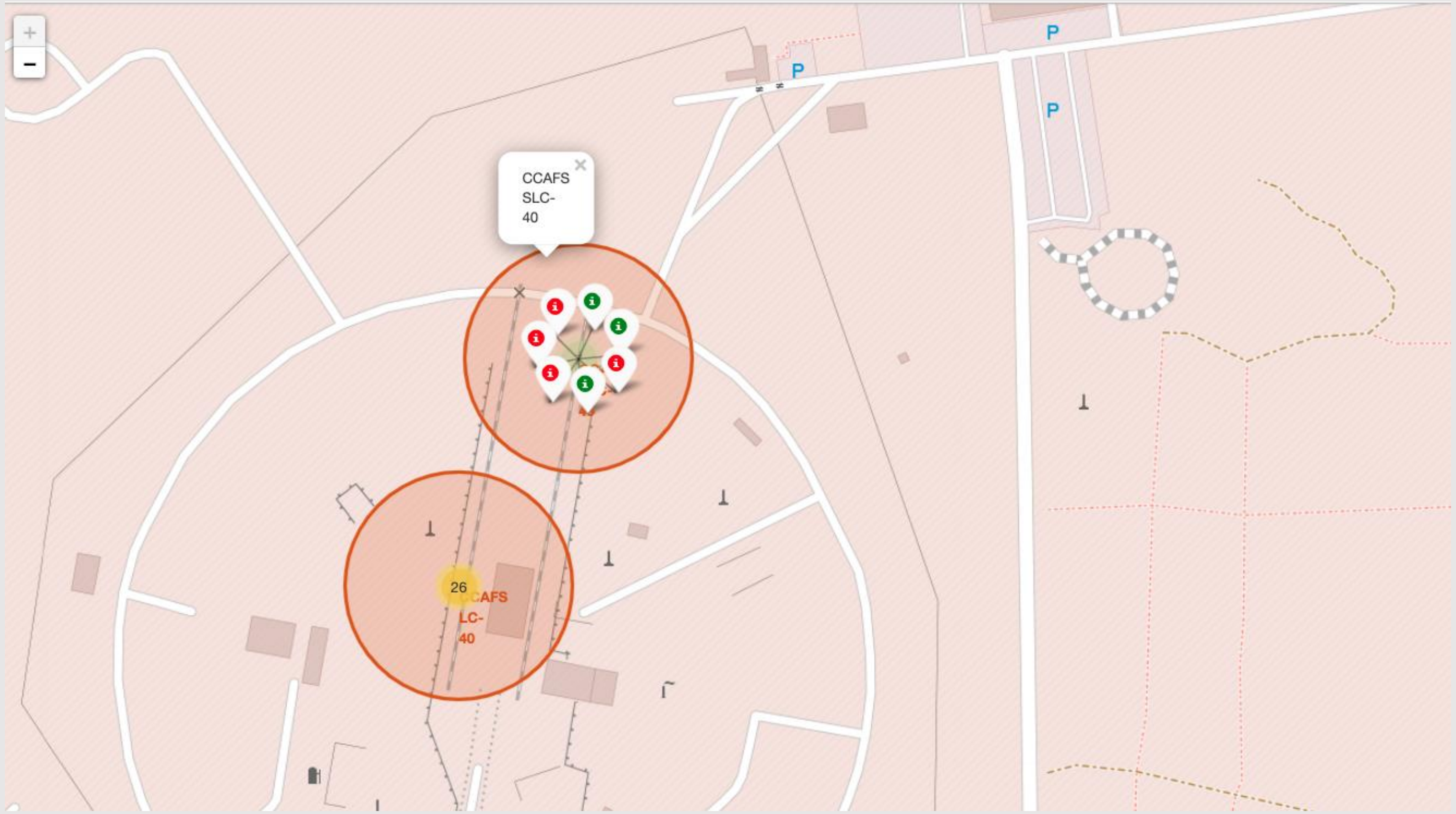
All launch sites proximity analysis

- Sites near sea selected due to safety, not far from roads / railroads



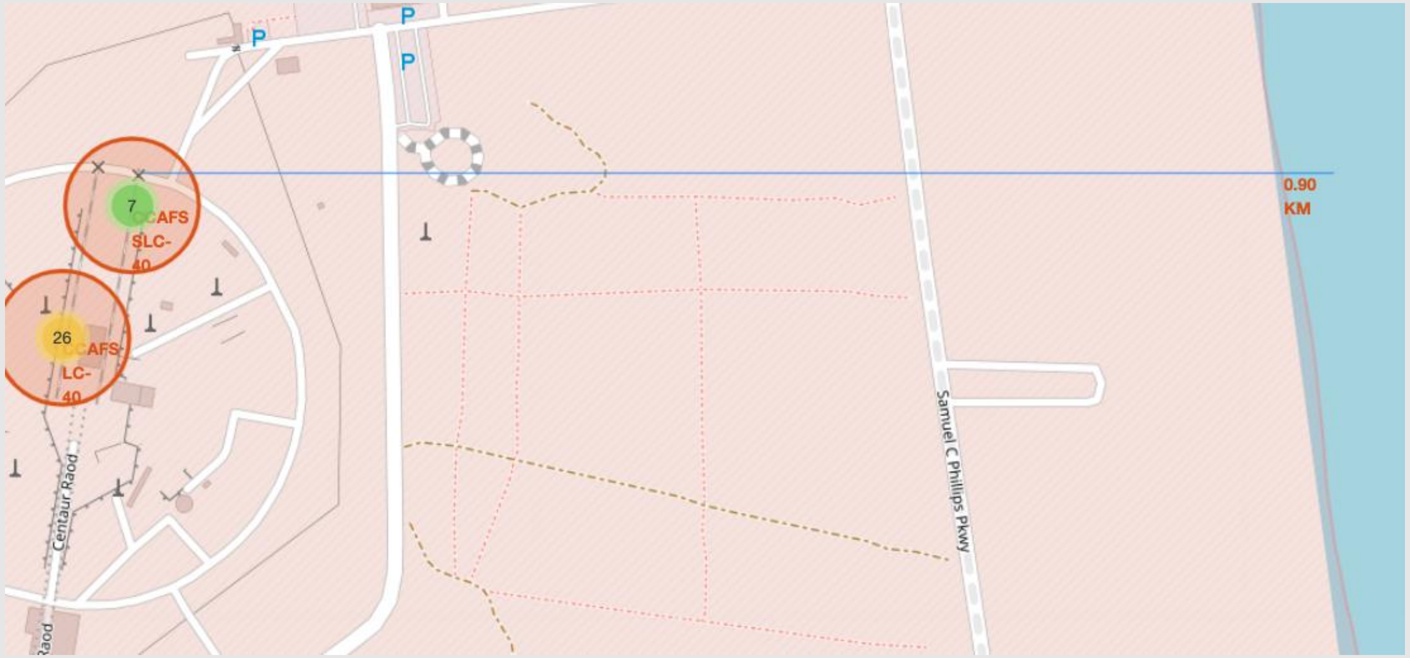
Launch outcomes by site

- Color-labeled markers in clusters identify launch sites with high success rates.



Logistics and safety

- Sites with favorable logistics and far from inhabitaded areas



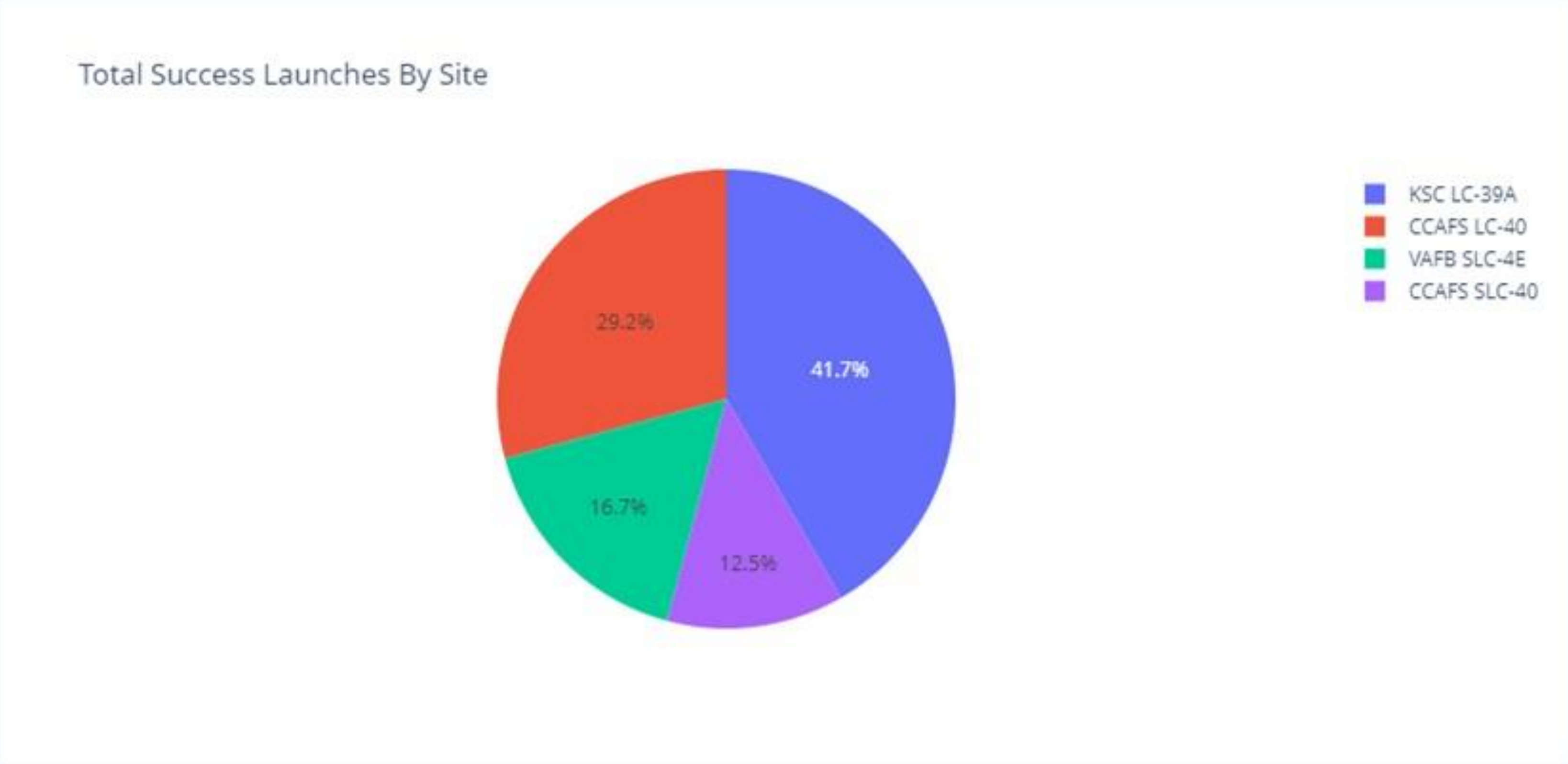
Section 5

Build a Dashboard with Plotly Dash

Total Success Launches By Site

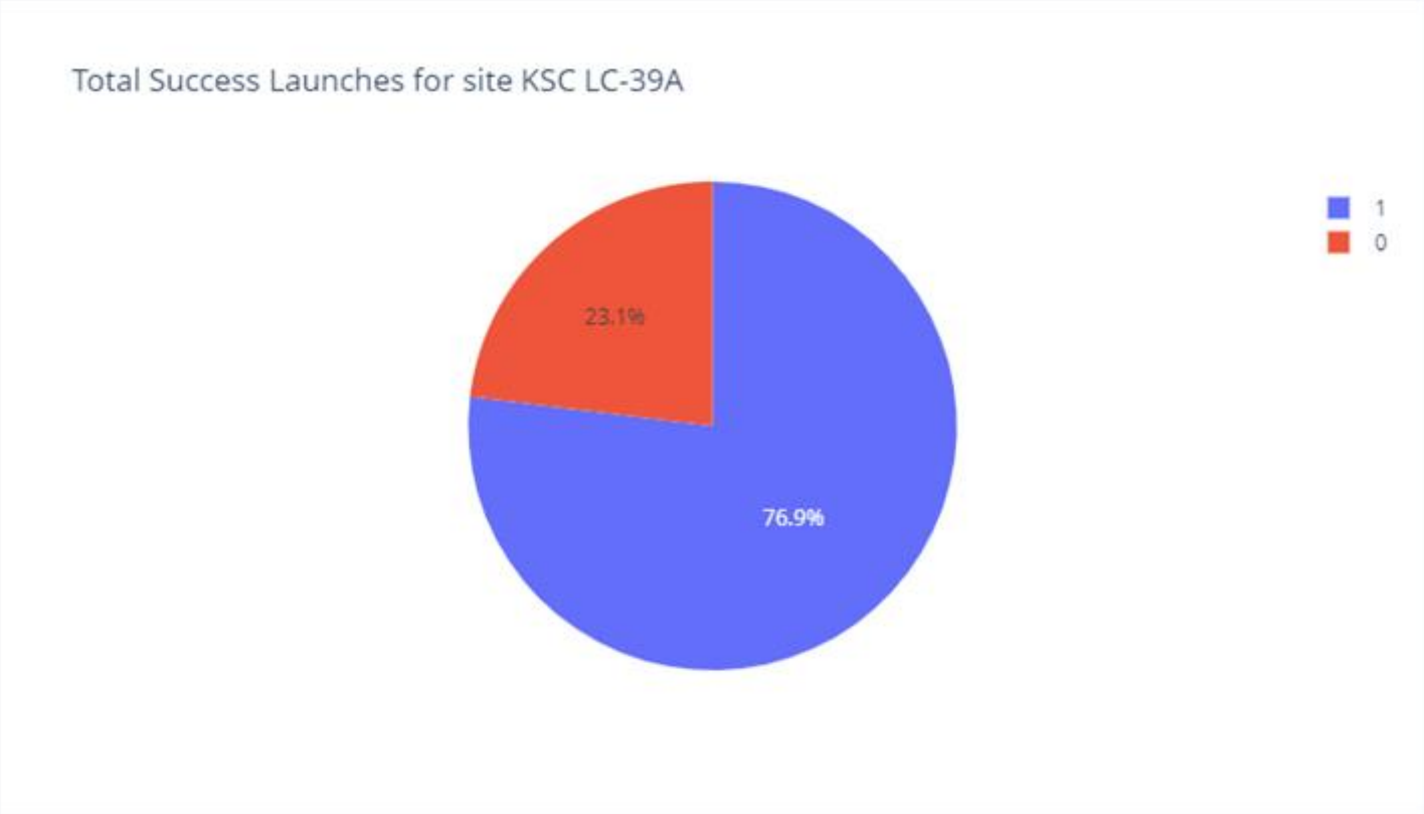
KSC LC-39A is the site with the higher success launches followed by CCAFS LC-40.

- Sites near sea selected due to safety, not far from roads / railroads



KSC LC-39A

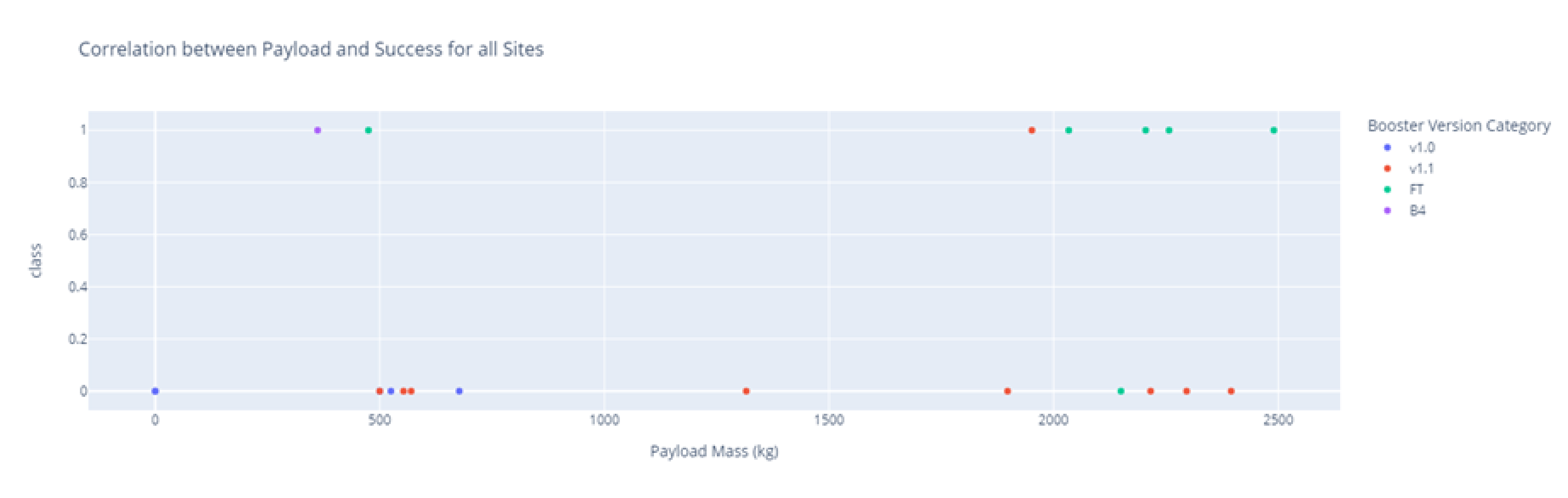
The piechart for the launch site KSC LC-39A shows the site with highest launch success ratio.



Payload vs. Launch Outcome

Scatter plot for all sites with 2500(kg), 5000(kg) and 10000(kg) payload ranges.

The 2500-5000(kg) range concentrate the majority of the successfully launches, the 0-2500(kg) range has most failed launches but all three are similar.

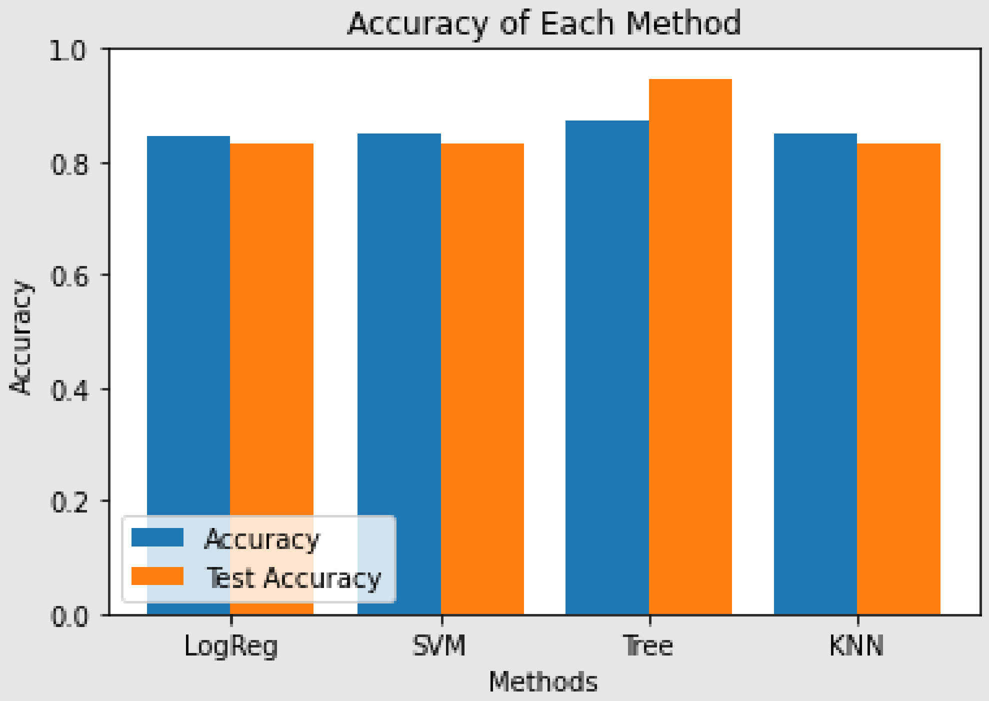


Section 6

Predictive Analysis

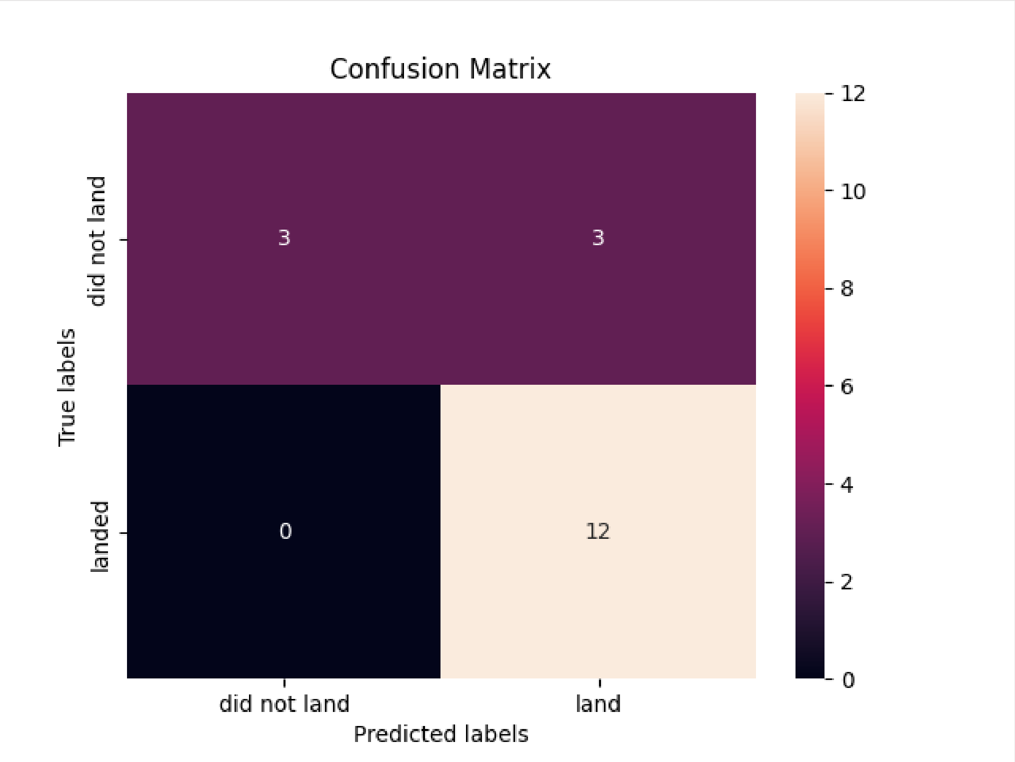
Classification Accuracy

- Best model to predict successful landings: Decision Tree Classifier.
- Accuracy: 87+%; Test Data Accuracy: 94+%.



Predictive Analysis

- Data Confusion Matrix proves accuracy by showing numbers of true positive and true negative vs false.



Section 7

Conclusions

Main Conclusions

- Different data sources were analyzed and conclusions were refined with them;
- The best launch site is KSC LC-39A and launches above 7,000kg are less risky;
- Landing outcomes improve over time due to evolution of technology and processes;
- Best performing models among four algorithms were sized;
- Decision Tree Classifier outperformed others by 5%, with 88% accuracy;
- Accuracy is key to decision making, but must be refined as false positive rate is 0 and negative 0.5;
- With historical data, model predicted successful launches, but half of failed launches were also classified as successes;
- It is key to SpaceY's business success to continuously improve the model by introducing the always evolving DS technologies and methods (see Recommendations)

Recommendations to SpaceY Board – Identifying the GAP

CRITERIA	AS-IS (NOW)	GAP	TO-BE (2050)
GLOBAL POPULATION	7.9 BILLION		9.8 BILLION
INTERNET USERS	5.0 BILLION (63%)		8.9 BILLION (90%)
AVAILABLE DATA	1,000 ZETABYTES		500,000+ ZETABYTES
GLOBAL ROBOT WORKFORCE	3.5 MILLION (0.003 BILLION)		9.4 BILLION
GLOBAL IT INDUSTRY TECHNOL/PROCESS	CLOUD COMPUTING ARTIFICIAL INTELLIGENCE MACHINE LEARNING		QUANTUM COMPUTING NEW NETWORKING TECHNOLOGIES CLEAN ENERGIES
DATA SCIENCE STATE OF ART	SQL REALTIONAL DATABASES NEURAL NETWORKS		ROBUST ANALYTICAL MODELS SECURE PLATFORMS-CYBERATTACKS HUMAN EMOTIONS DETECTION
AEROSPACE INDUSTRY	ACCESS TO PRIVATE SECTORS CONTINUOUS LEARNING 1ST STAGE RECOVERY		1,000 STARSHIPS TO MARS 100+ TIMES CHEAPER VS TODAY ASTEROIDS MINING DEVELOPMENT
CRITICAL SUCCESS FACTORS (CSF)	DATA MINING / INTERPRETATION BUSINESS FORECASTING SMART DECISION MAKING		DS PROFS CONTINUOUS LEARNING EFFICIENCY ADVANCED TOOLS IT ORGS SHARING BUDGETS

SpaceY Board – Addressing 2050 vs 2020 GAP Initiatives

SpaceY needs to provision enough budget for Data Science professionals to:

- Augment business process, amplifying databases for human-machine interactions;
- Incorporate interdisciplinary concepts (sociology / psychology);
- Amplify social media as source of data (Twitter, Facebook, others);
- Develop Team Activity (from creating models to how to use once built them);
- Increase Cybersecurity skills (safeguard business data);
- Be ready for a growing Cloud Computing prevalence;
- Bring apps to capture workflows and train on best practices;
- DS Coding and AI essentials, but develop also more bussines oriented;
- Quantum leap initiation. Algoritms to solve real-time problems.