IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robert Szini
2022-10-31

Outline

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

Executive Summary

- The **aim of this project** was to **predict** if the Falcon 9 first stage will **land successfully**. 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**.

- The following **methodologies** were used in this project
  - **Data Collection** using web scraping and SpaceX API
  - **Exploratory Data Analysis** (EDA) applying data wrangling, data visualization and interactive visual analytics
  - **Machine Learning Prediction**.

- Summary of all **results**
  - Valuable **data** has been collected from **public sources**
  - **EDA** allowed to identify which features are the best to **predict success of launchings**;
  - With the application of **Machine Learning** techniques that **model** has been found which predicts the best this opportunity using all collected data.

Introduction

- 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.**
- Therefore if we can **determine** if the **first stage will land**, we can determine the **cost of a launch**. This information can be used if an **alternate company** wants to **bid against SpaceX** for a rocket launch.
- **Questions** to be answered:
  - How the **total cost** for launches can be estimated the best way using the **prediction** of **successful landings** of the **first stage of rocket**
  - Where is the **best place** to make **launches**.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different models (SVM, Classification Trees and Logistic Regression) have been fitted on standardized data in order to predict if the first stage will land
  - Best model has been chosen based on accuracy both on train and test samples

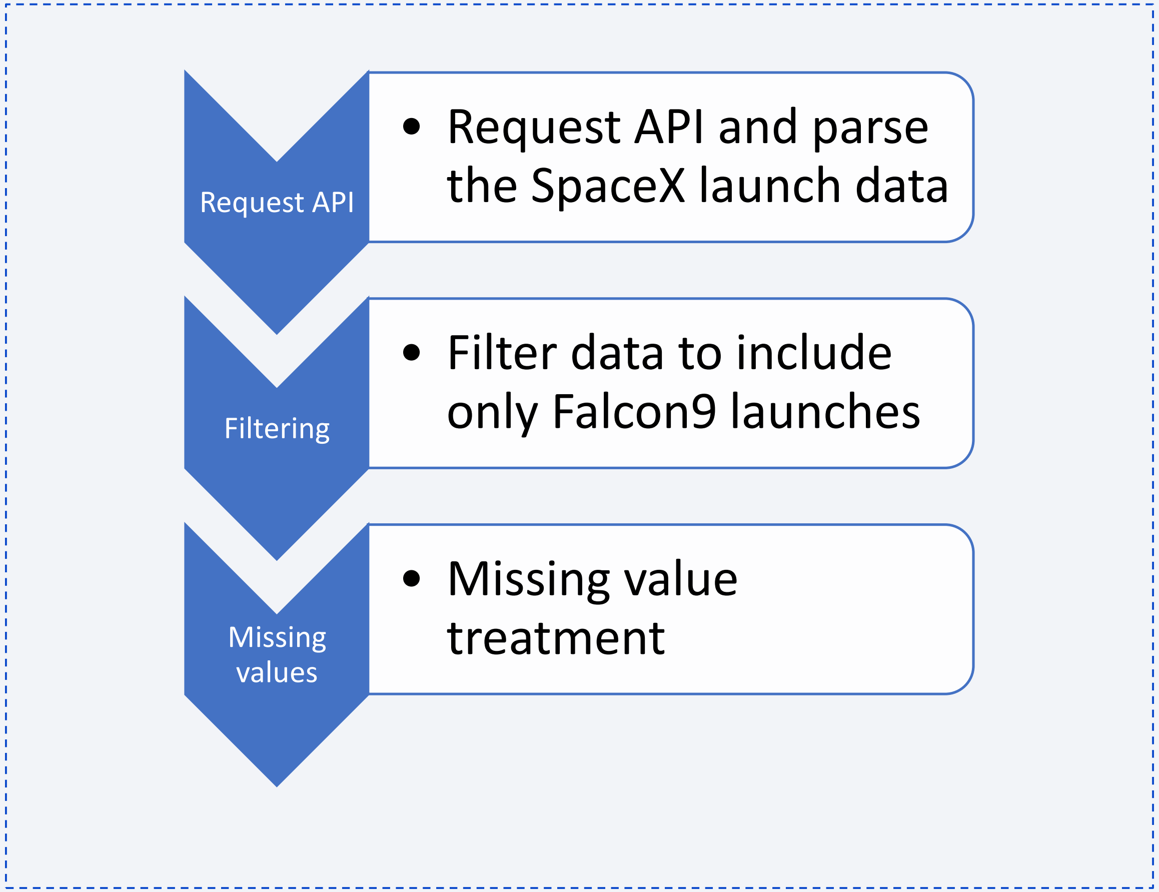
Data Collection

- Data sets were collected in two ways:
  - from SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - from Wikipedia using web scraping technics (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches).

Data Collection – SpaceX API

- SpaceX offers a public API which has been used to collect data
- Flowchart beside shows how this API was used and then data is persisted.

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/749b7249be28228c24f541405a3bcdc4303b9ac4/spacex_data_collection_api.ipynb



Request API

- Request API and parse the SpaceX launch data

Filtering

- Filter data to include only Falcon9 launches

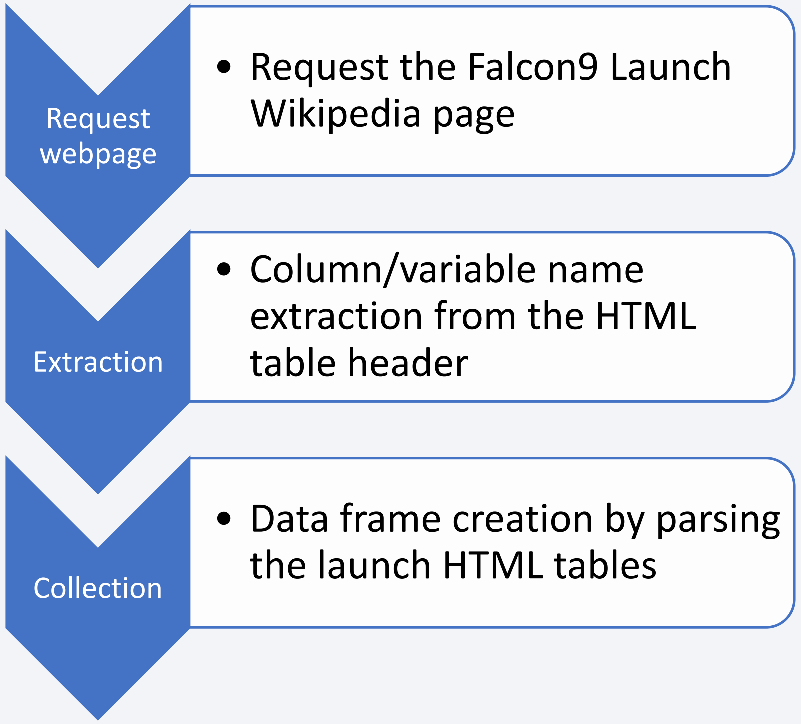
Missing values

- Missing value treatment

Data Collection - Scraping

- The source of the data from SpaceX was Wikipedia
- Flowchart beside shows how web scraping was used to get data from Wikipedia.

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/6af8b928a081524e05d47d7acab0c90479ff6ed2/spacex_webscraping.ipynb



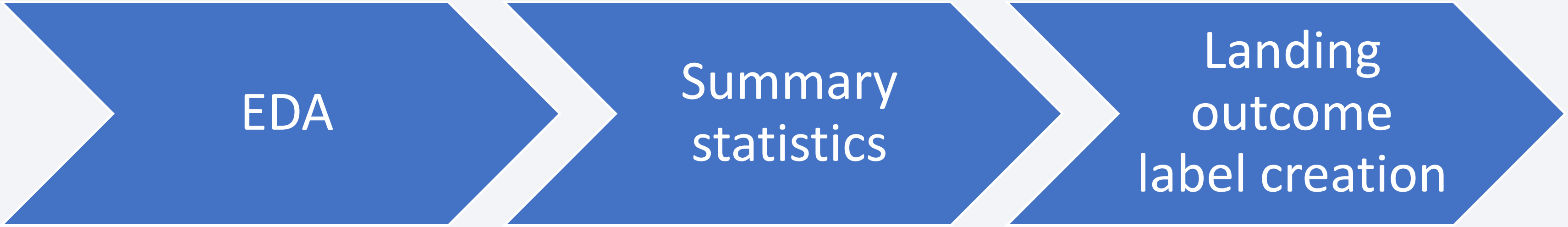
Request webpage
- Request the Falcon9 Launch Wikipedia page

Extraction
- Column/variable name extraction from the HTML table header

Collection
- Data frame creation by parsing the launch HTML tables

Data Wrangling

- **Exploratory Data Analysis** (EDA) was performed on the dataset
- The **following summaries have** been calculated:
  - launches per site
  - occurrences of each orbit
  - occurrences of mission outcome per orbit type
- **Landing outcome label** was created from Outcome column.



EDA

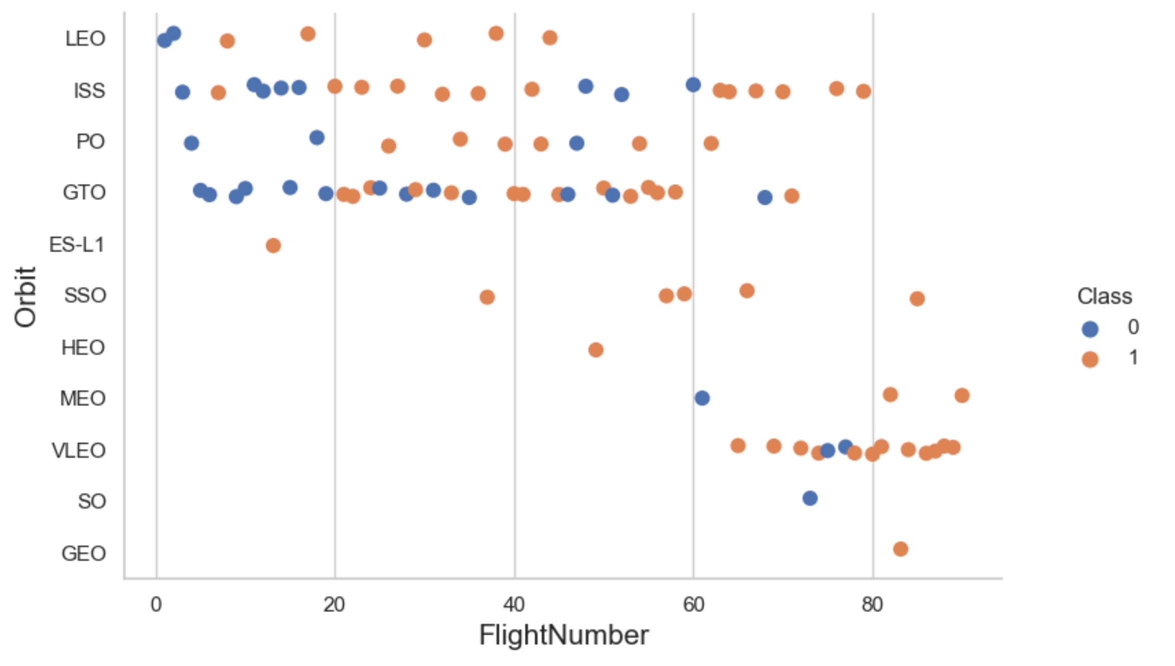
Summary statistics

Landing outcome label creation

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/42901c880066ca97b07f54613cf2446f82935024/spacex_data_wrangling.ipynb

EDA with Data Visualization

- In order to explore data **scatterplots** and **barplots** were used to visualize the relationship between pair of features:
  - Payload Mass and Flight Number
  - Launch Site and Flight Number
  - Launch Site and Payload Mass
  - Orbit and Flight Number
  - Payload and Orbit



- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/0d21adbbdddb285bd039050c0c2a15b31d2dfee4/spacex_eda_dataviz.ipynb

EDA with SQL

The following **SQL queries** were 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.

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/63e71ded856ac6ece2459d426c222e8804feda28/spacex_eda_sql.ipynb

Build an Interactive Map with Folium

- **Markers**, **circles**, **lines** and **marker clusters** were used with **Folium** Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicate groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/b70333355a8e466d3c7565409dc5254cbd5e0609/spacex_folium_launch_site_location.ipynb

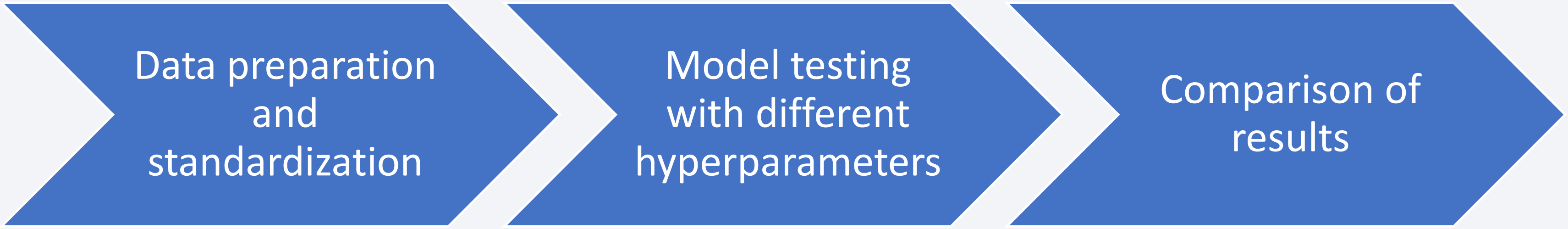
Build a Dashboard with Plotly Dash

- The following **graphs** and **plots** were used to visualize data
  - Percentage of launches by site
  - Success of launches as a function of Payload mass (kg)
- The **dashboard** with the plots and graphs made it possible to quickly analyze the relation between payloads and launch sites, helping to identify what is the best place to launch according to payloads.

- Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/0372b123bd706b3bebc219304153cc811697ddbc/spacex_dash_app.ipynb

Predictive Analysis (Classification)

- Four **classification models** were built and **compared**:
  - logistic regression
  - support vector machine
  - decision tree
  - k nearest neighbors.



Data preparation and standardization

Model testing with different hyperparameters

Comparison of results

Source code: https://github.com/szirob/IBM-Data-Science-Capstone-/blob/2b0a1190356ad0e1d37af12ddbc8517e6545cc0e/spacex_machine_learning_prediction.ipynb

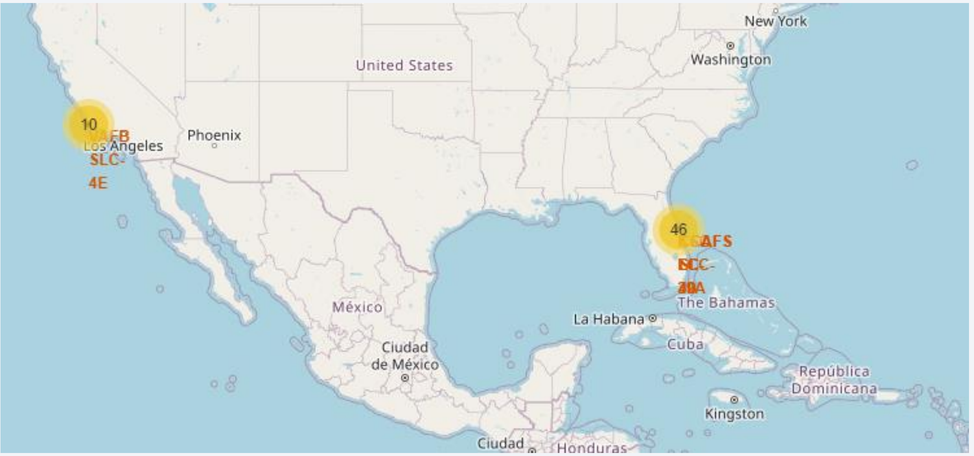
Results

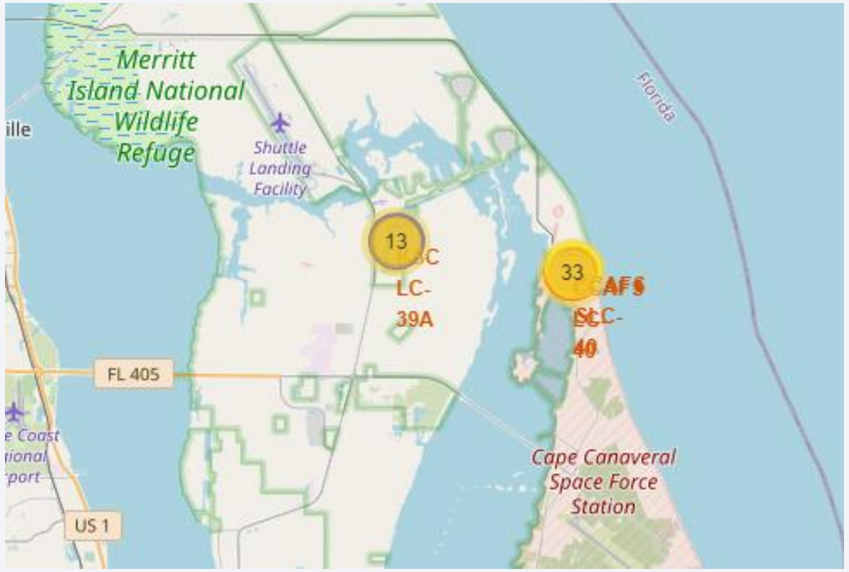
Exploratory data analysis **results**:

- Four different launch sites have been used by SpaceX
- The first launches were done by Space X and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 - five years after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Almost 100% of the missions were successful
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The number of success landings keeps rising as the years go by

Results (2)

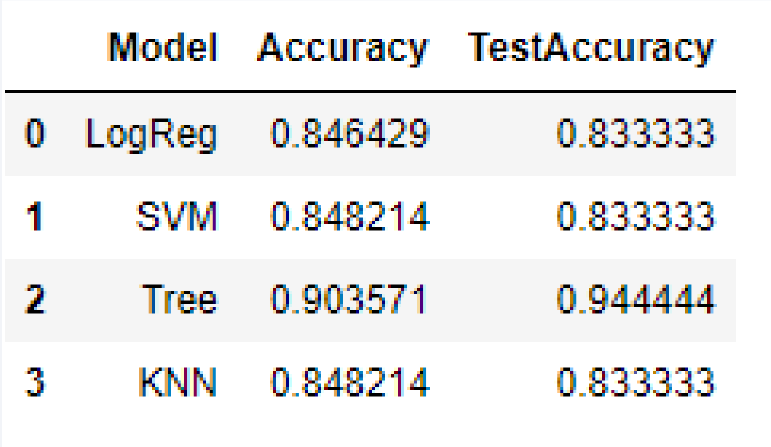
- With **interactive analytics** it was possible to identify that launch sites are used to be near sea with good logistic infrastructure around.
- Most launches happened at east cost launch sites.





Results (3)

- Based on **Accuracy results Decision Tree** Classifier is the best model to predict successful landings having accuracy over 90% and accuracy for test data over 94%.

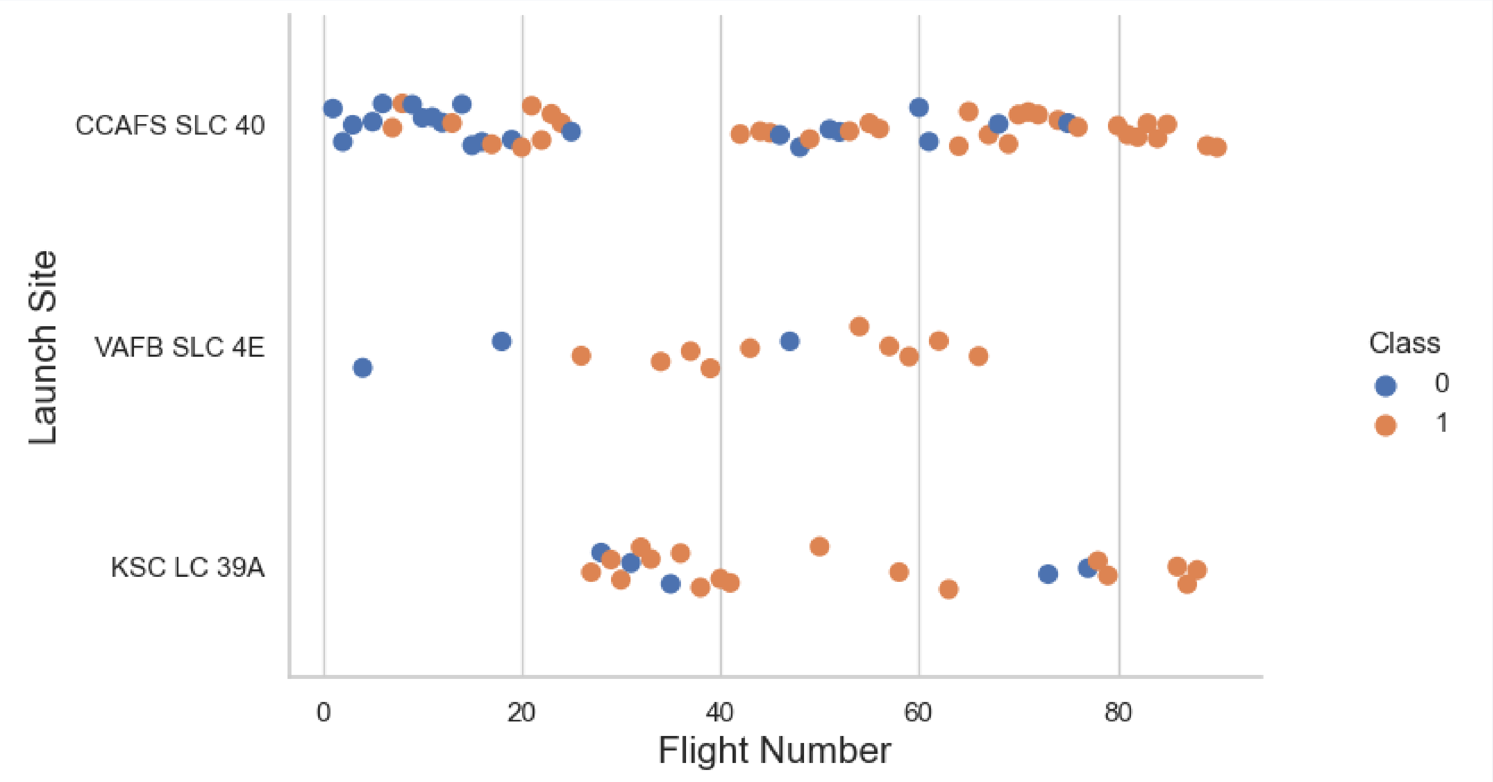


	Model	Accuracy	TestAccuracy
0	LogReg	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Tree	0.903571	0.944444
3	KNN	0.848214	0.833333

Section 2

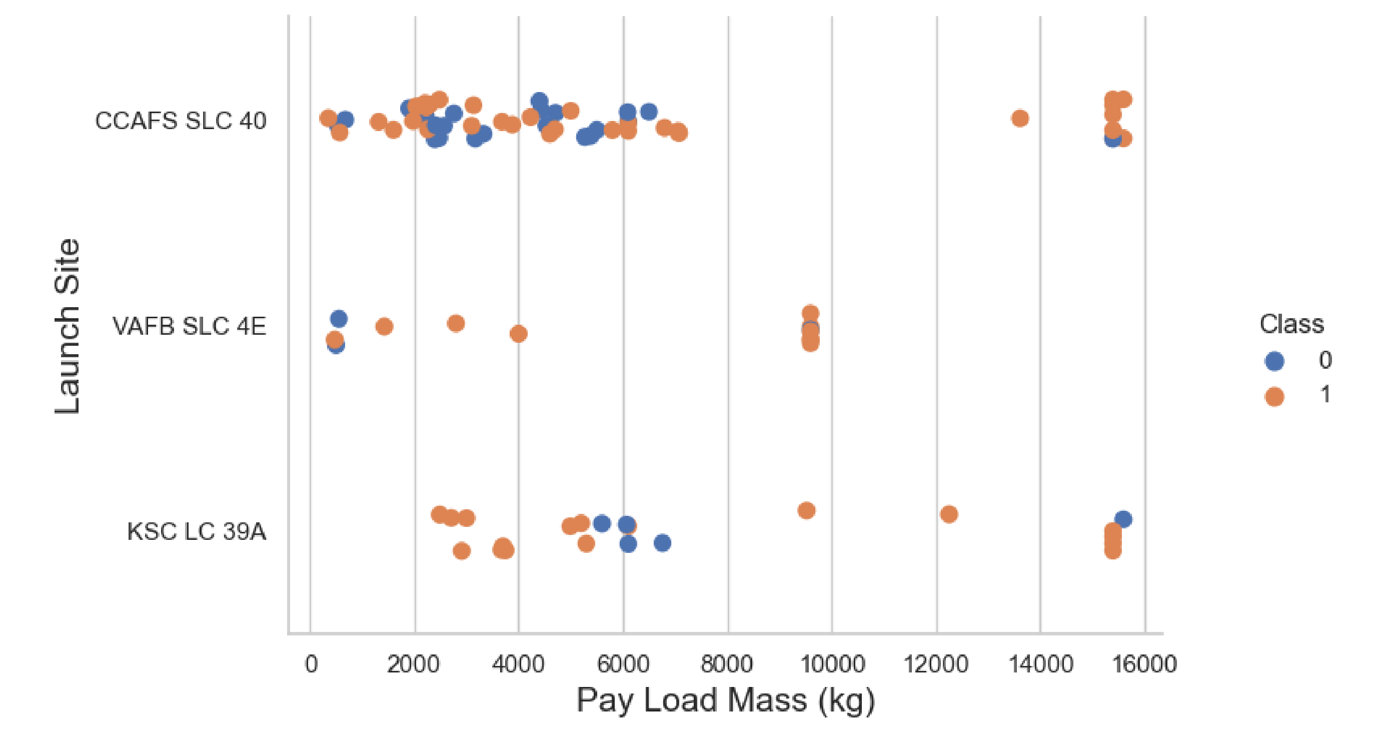
Insights drawn from EDA

Flight Number vs. Launch Site



- Based on the plot above the **best launch site** nowadays is CCAF5 SLC 40 where most of recent launches were successful.
- VAFB SLC 4E takes the second place while KSC LC 39A takes the third.
- It can be seen that general **success rate improved** over time.

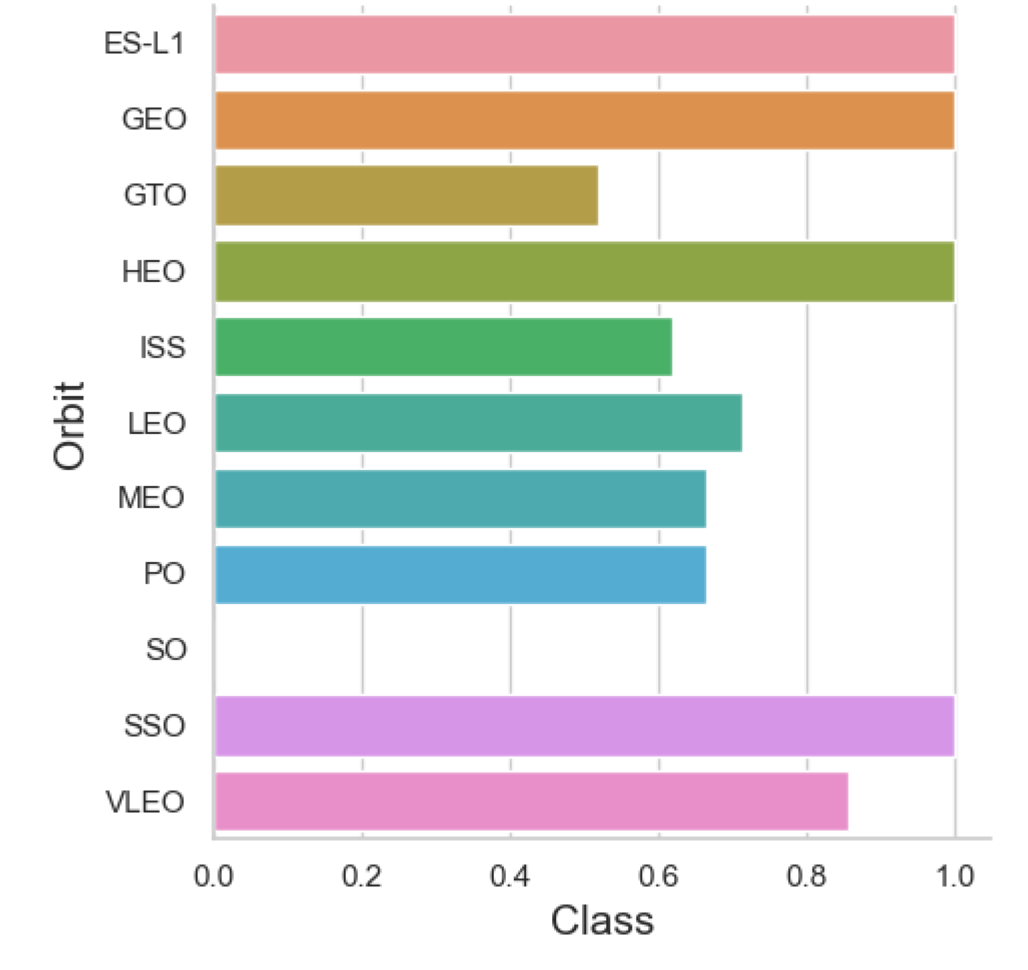
Payload vs. Launch Site



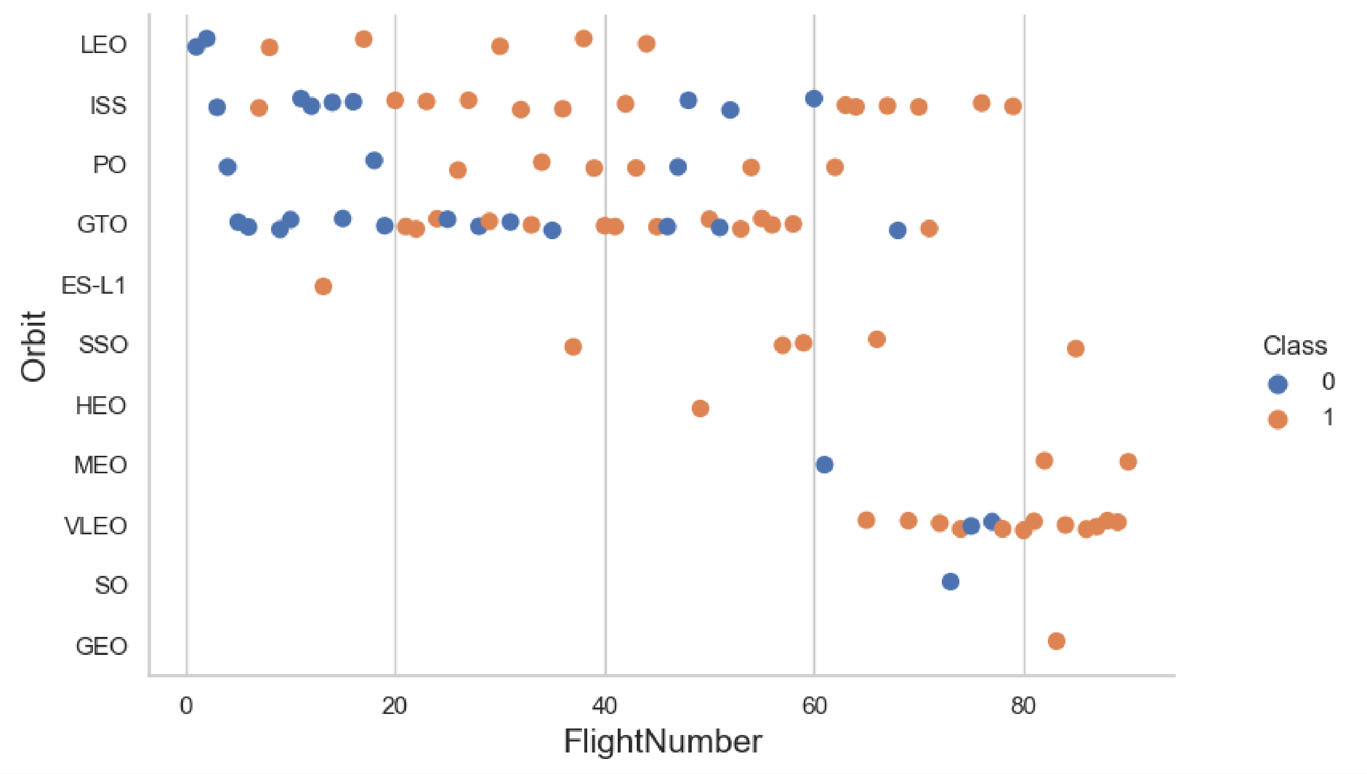
- Payloads over 9,000kg have very good success rate.
- Based on the plot above Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- Based on the plot beside the **highest success rates** belong to the following orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO.
- The **second** and the **third** highest **success rates** belong to the following orbits:
  - VLEO (~80%)
  - LEO (~70%).

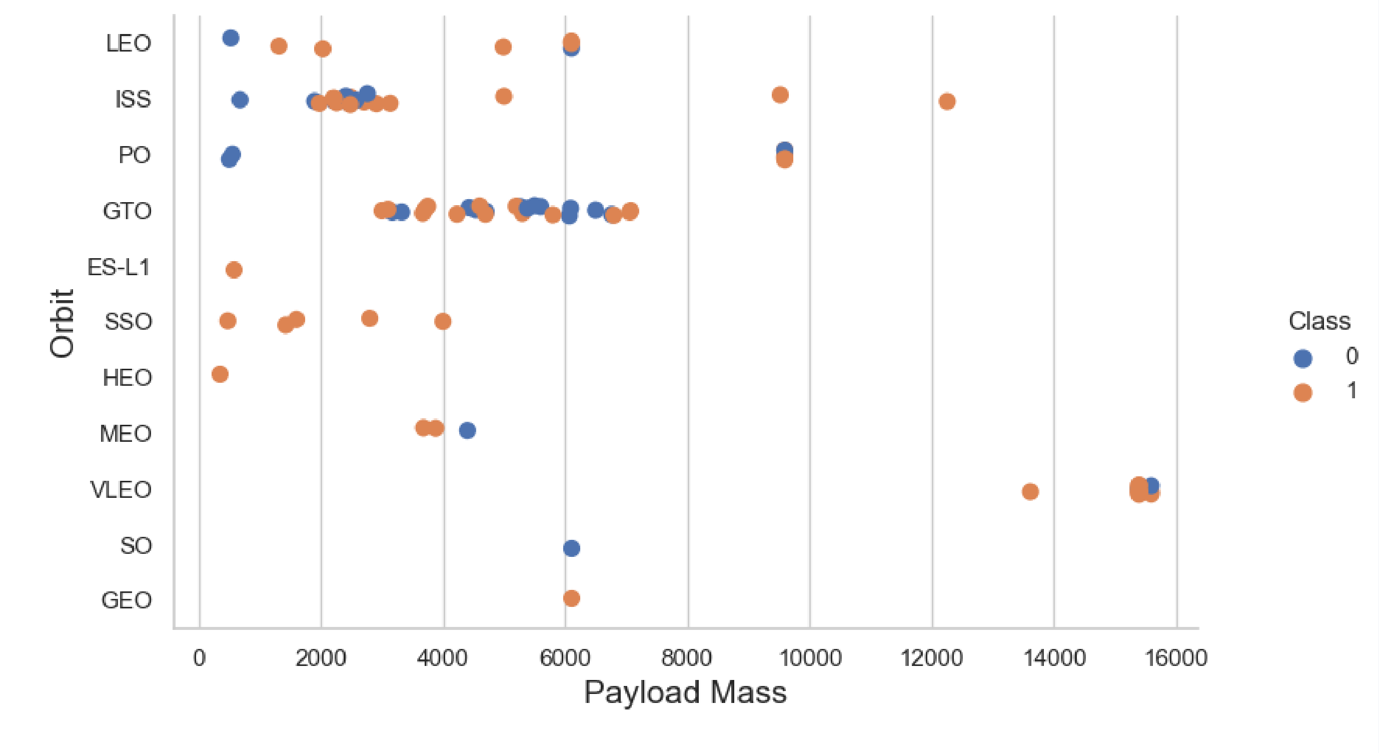


Flight Number vs. Orbit Type



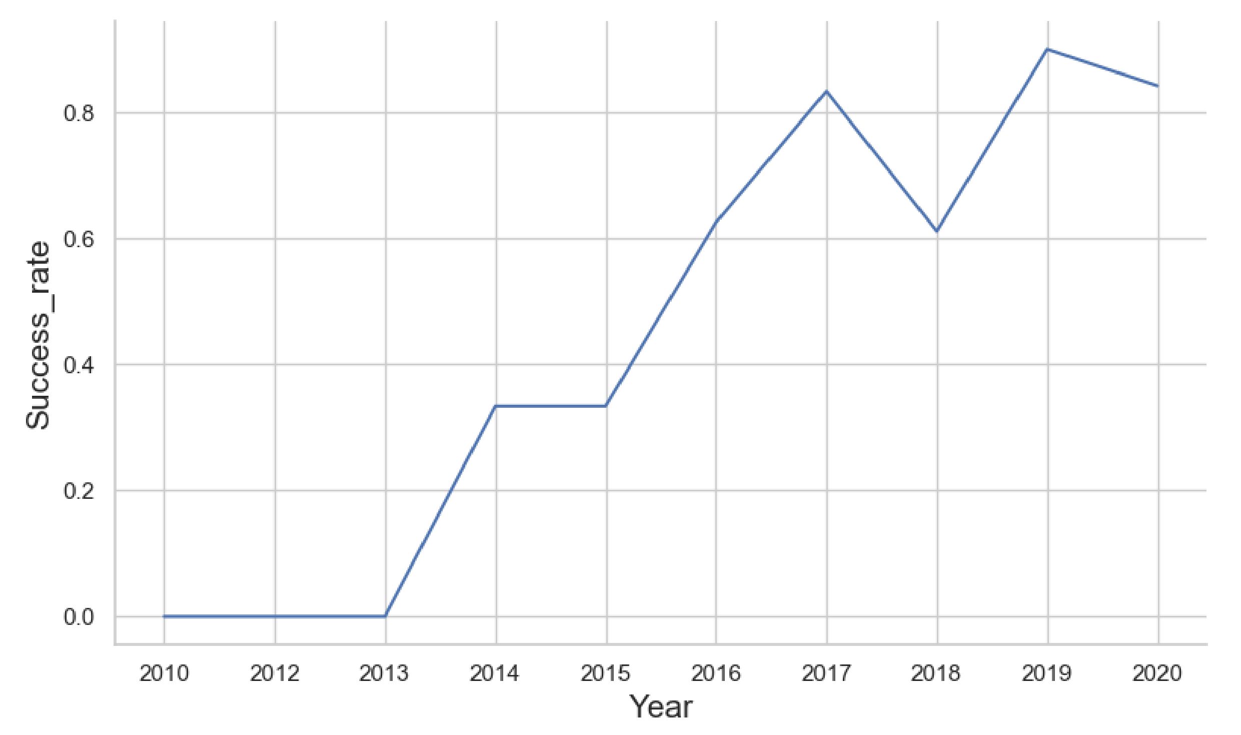
- Success rate improved over time to all orbits.
- In case of VLEO orbit recent increase of its frequency can be seen.

Payload vs. Orbit Type



- It seems that there is no relation between payload and success rate to orbit GTO.
- ISS orbit has the widest range of payload and a good rate of success.
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend



- Success rate started to increase in 2013 and kept until 2020.
- In the first three years the success rate was 0. It seems that this was a period in which the improvement of technology was necessary.

All Launch Site Names

- According to data, there are four launch sites:

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

- Launch sites mentioned above were obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	landing_outcome	year
2010-06-04 00:00:00	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)	2010
2010-12-08 00:00:00	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)	2010
2012-05-22 00:00:00	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt	2012
2012-10-08 00:00:00	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt	2012
2013-03-01 00:00:00	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt	2013

- The Table above shows the 5 records where launch sites begin with `CCA`.

Total Payload Mass

- The **total payload** carried by boosters from **NASA** can be found in the following Table:

Total payload (kg)
45,596

- Total payload has been calculated above by **summing all payloads** whose codes contain '**NASA (CRS)**'.

Average Payload Mass by F9 v1.1

- **Average payload** mass carried by booster version **F9 v1.1** can be found in the following Table:

Booster Version	Average payload (kg)
F9 v1.1	2,928.4

- Average payload has been calculated by filtering data on the specific booster version and calculating the average payload mass.

First Successful Ground Landing Date

- **First successful** landing outcome on ground pad:

Date
2015-12-22

- Date above has been found by **filtering data** on successful landing outcome on ground pad and getting the **minimum value** of the obtained dates.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following Table shows those 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

- Applying the filters above, 4 different booster version can be found in the data.

Total Number of Successful and Failure Mission Outcomes

- The following Table summarizes the number of successful and failure mission outcomes:

Mission outcome	Number of observations
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Table beside shows those **boosters** which have carried the **maximum payload mass**

- These boosters have been selected from the database based on the maximum payload mass they carried.

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 in year 2015:

Booster version	Launch site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

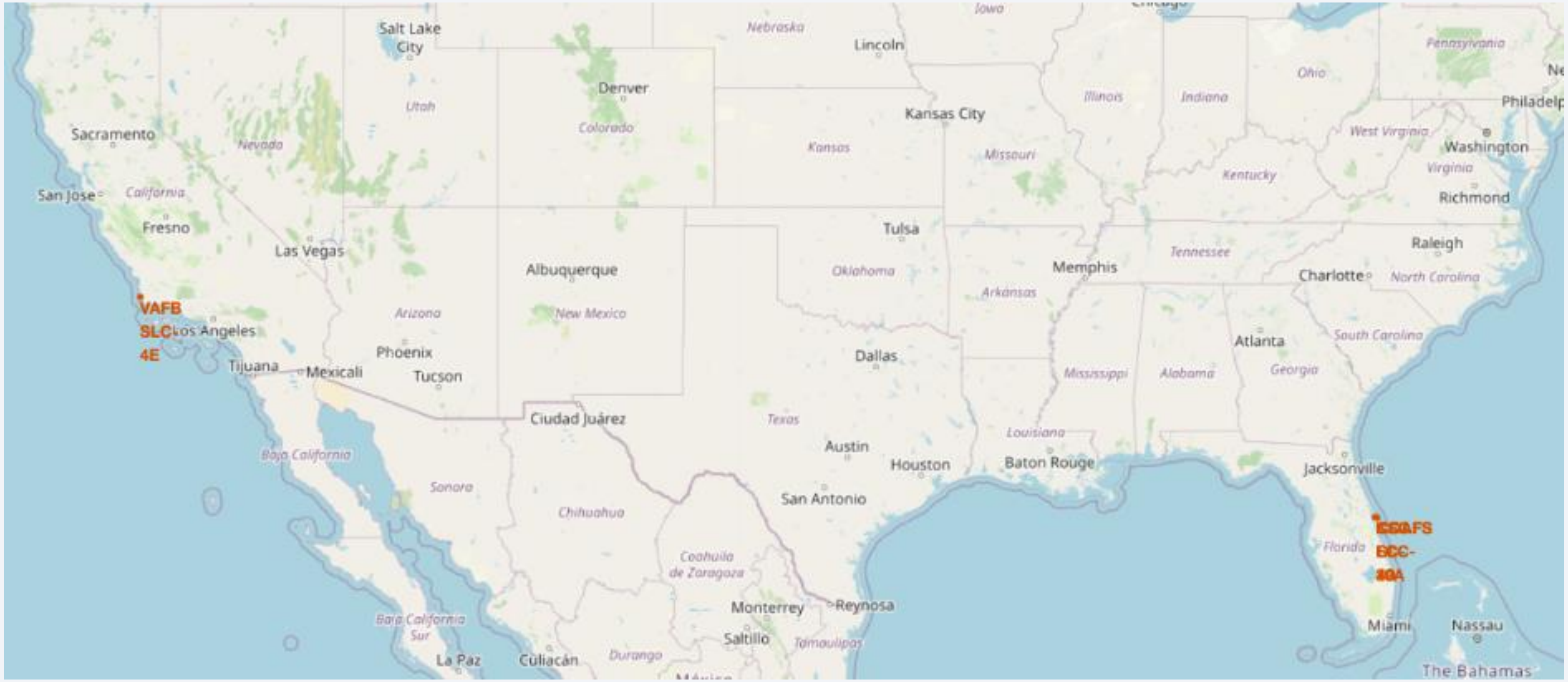
Landing outcome	Number of observations
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Based on the „Number of observations” column „No attempt” must be taken into account.

Section 3

Launch Sites Proximities Analysis

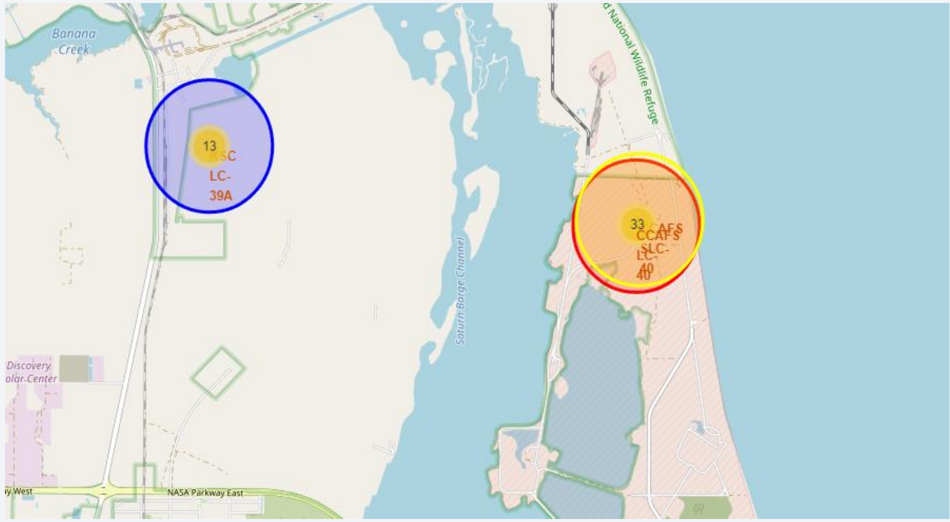
All launch sites

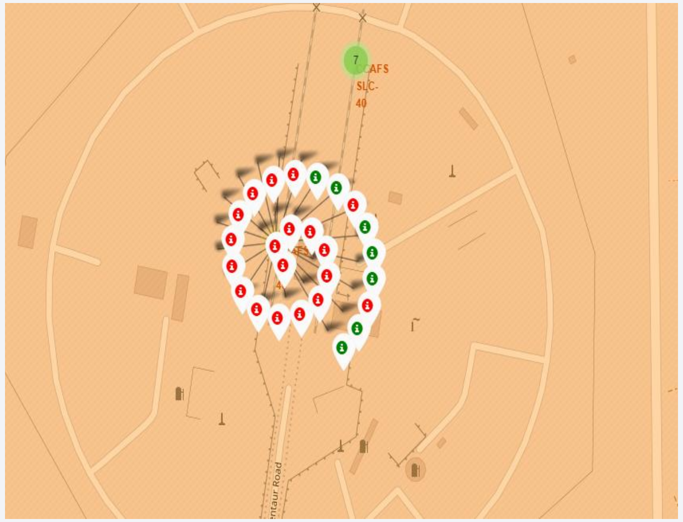


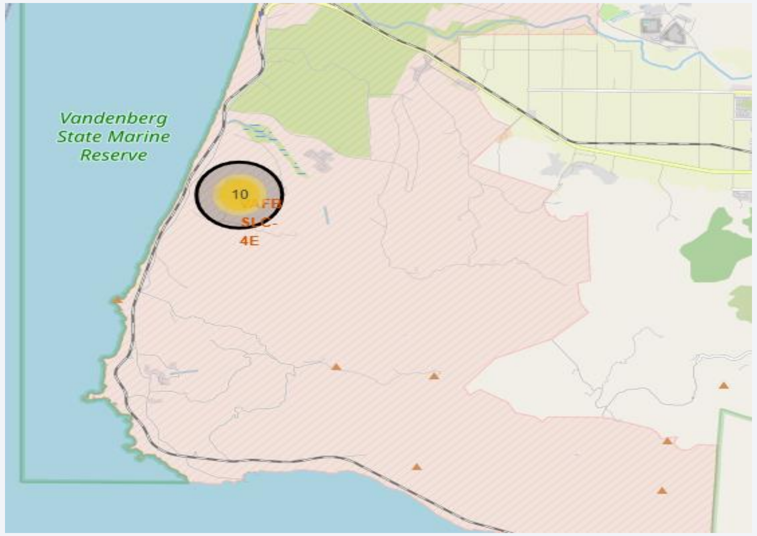
- Launch sites are near sea and not too far from roads and railroads.

Launch outcomes

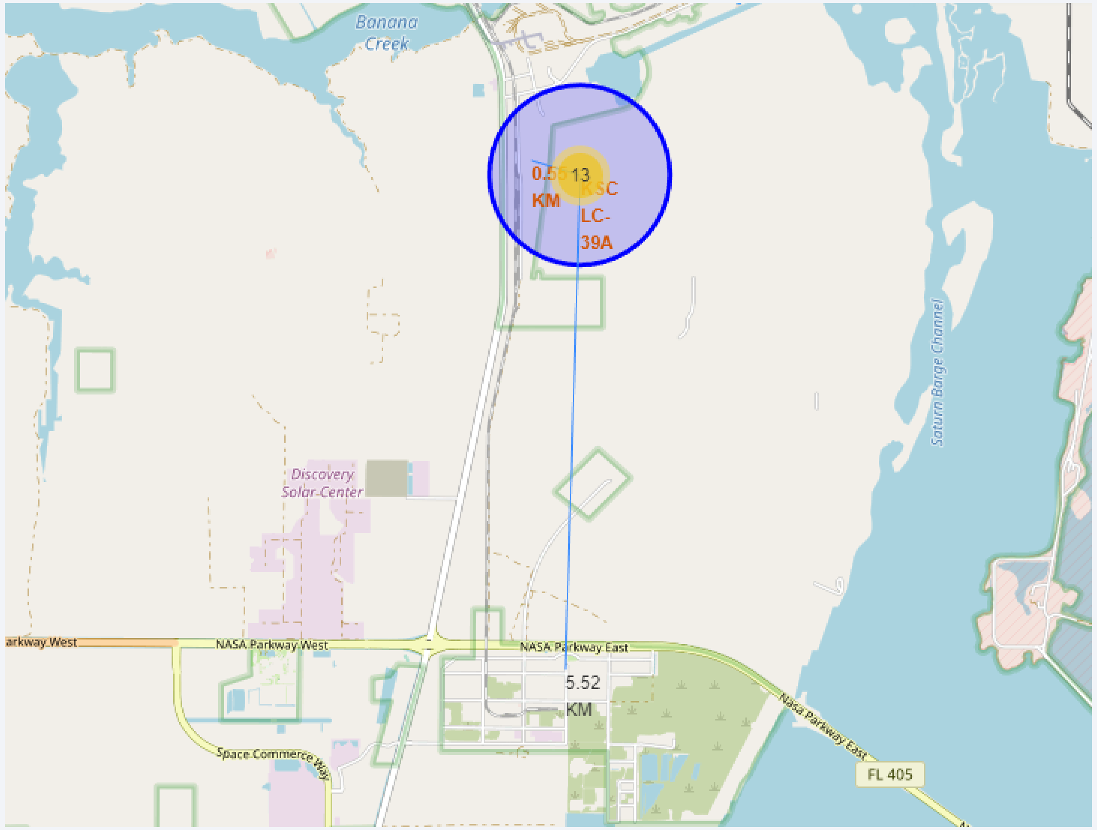
- **First picture** shows the number of launches on CCAFS LC-40, KSC LC-39A and CCAFS SLC-40 launch sites.
- The **second image** shows the launch outcomes on CCAFS LC-40 site. Green markers indicate successful and red ones indicate failure.
- The **bottom image** shows the number of launches on VAFB SLC-4E launch site.



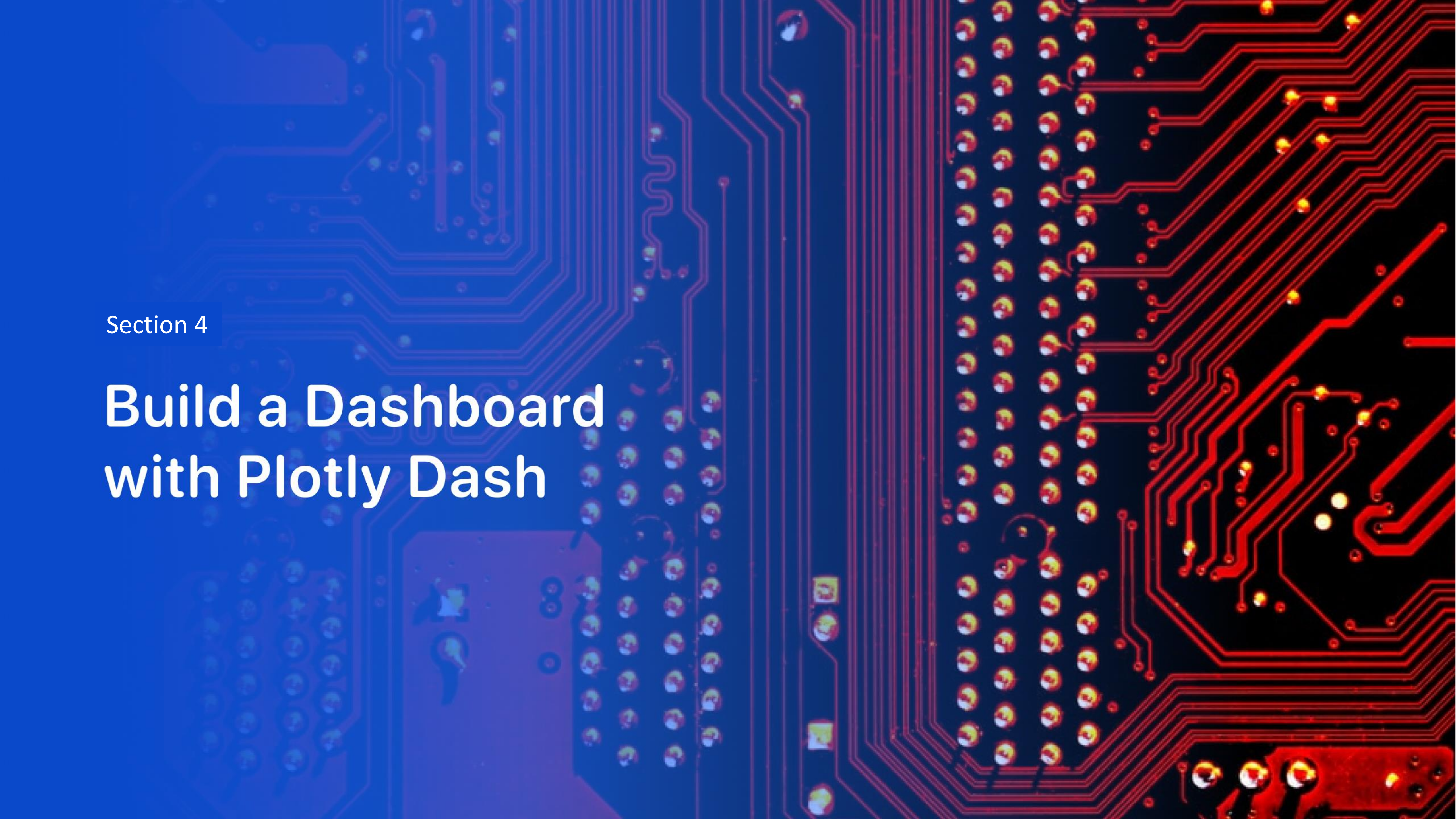




KSC LC-39A logistic properties



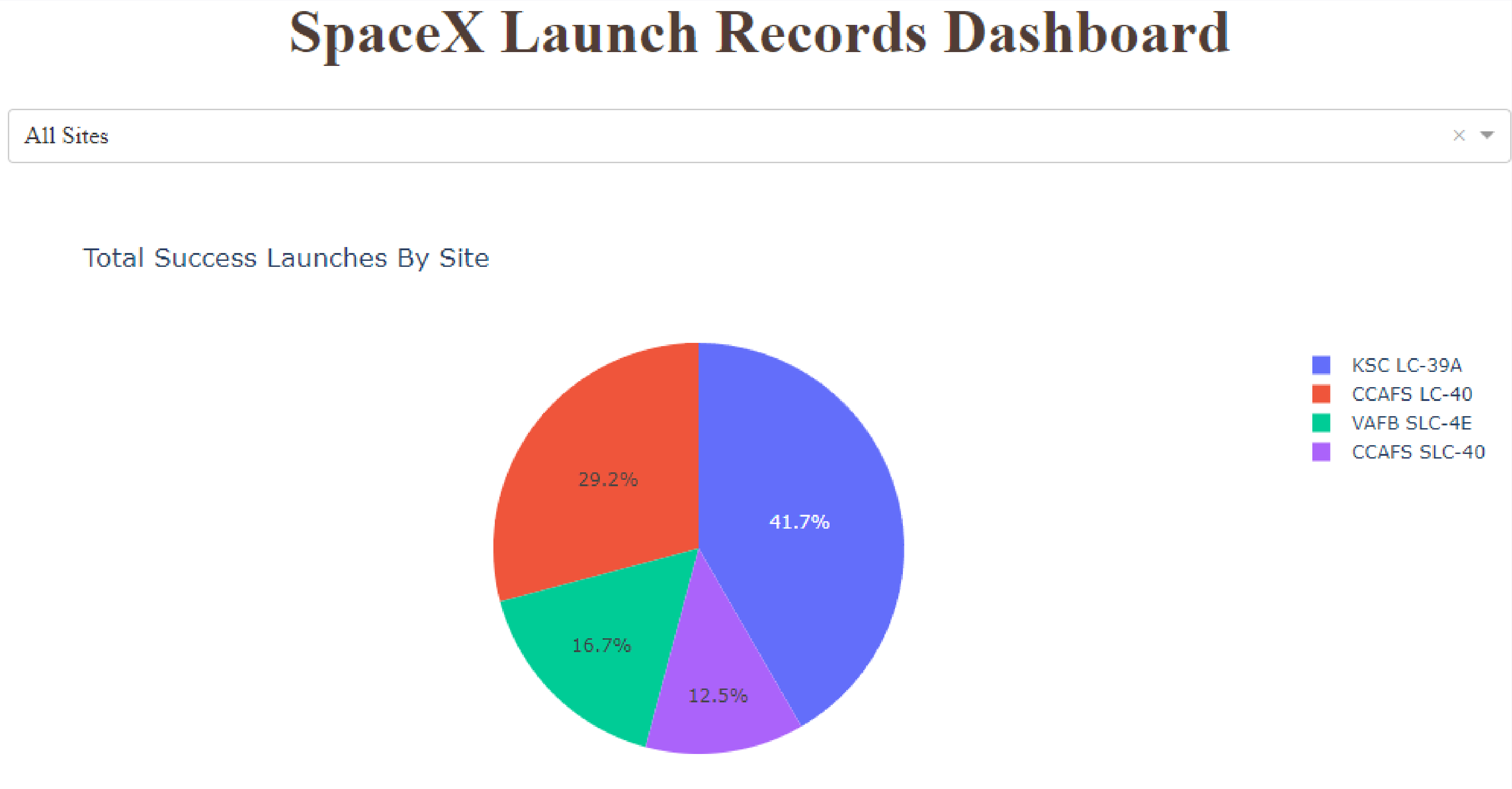
- Launch site **KSC LC-39A** has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

Build a Dashboard with Plotly Dash

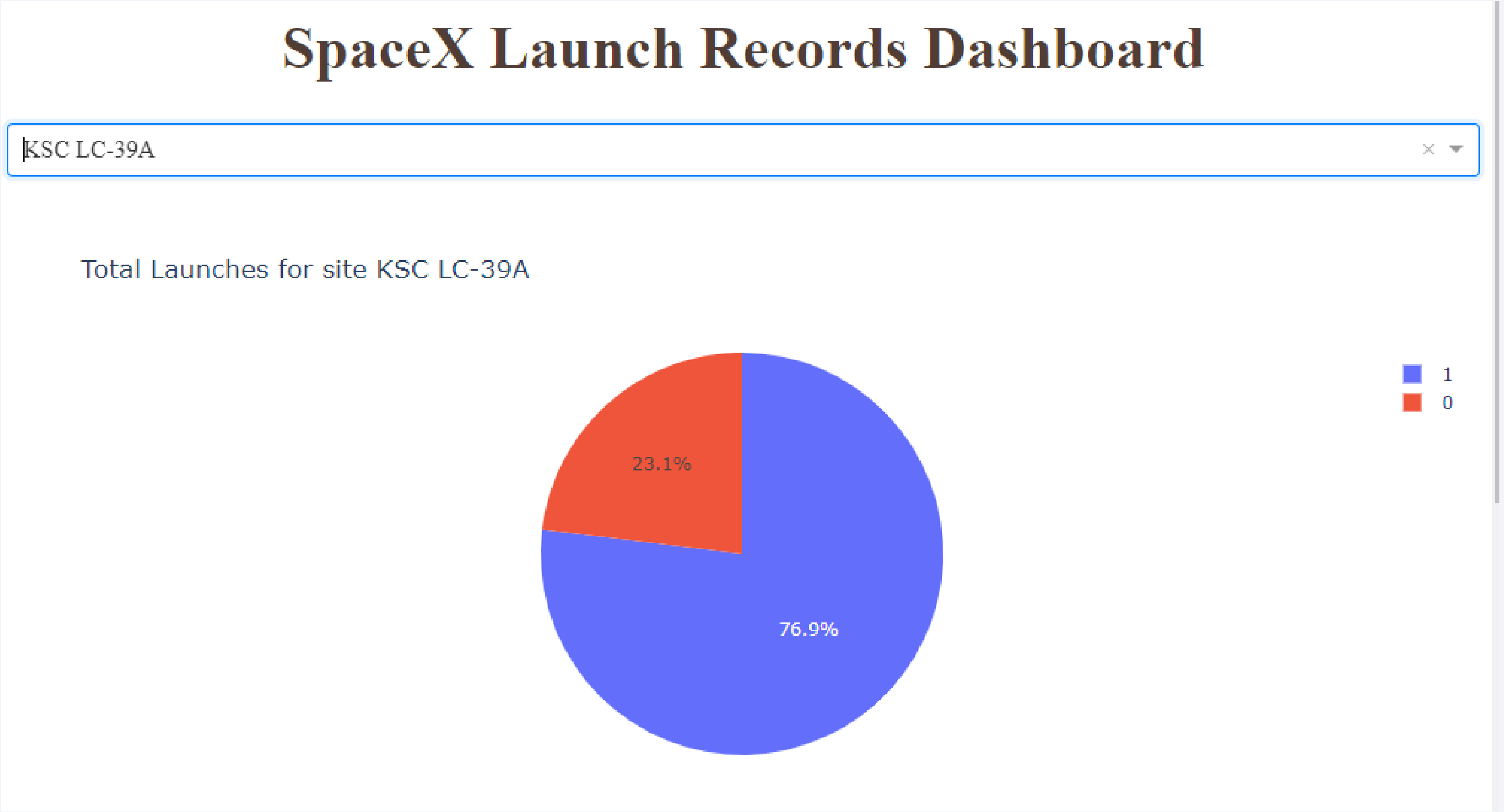
Successful Launches by Site



SpaceX Launch Records Dashboard

- The **place** from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



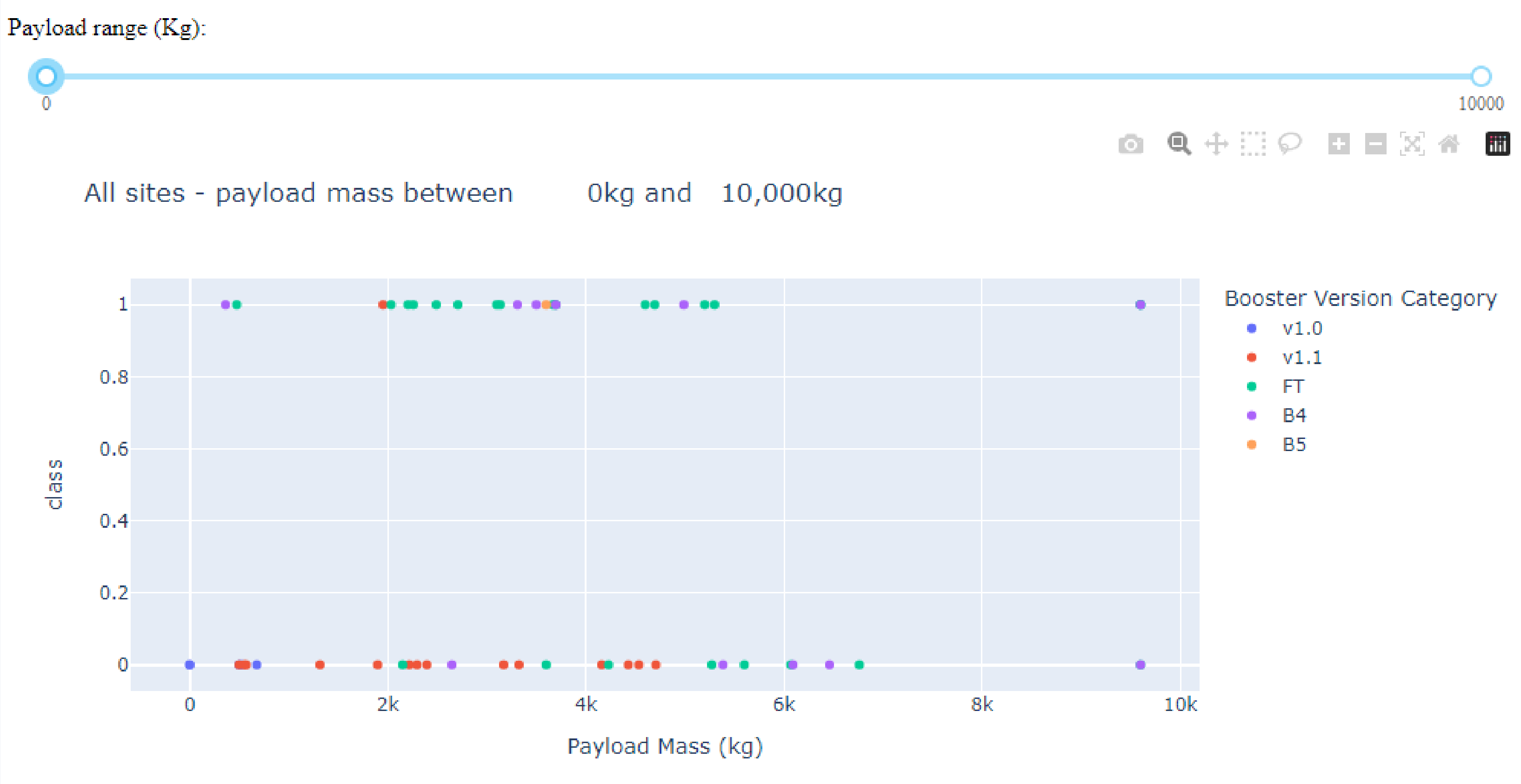
SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for site KSC LC-39A

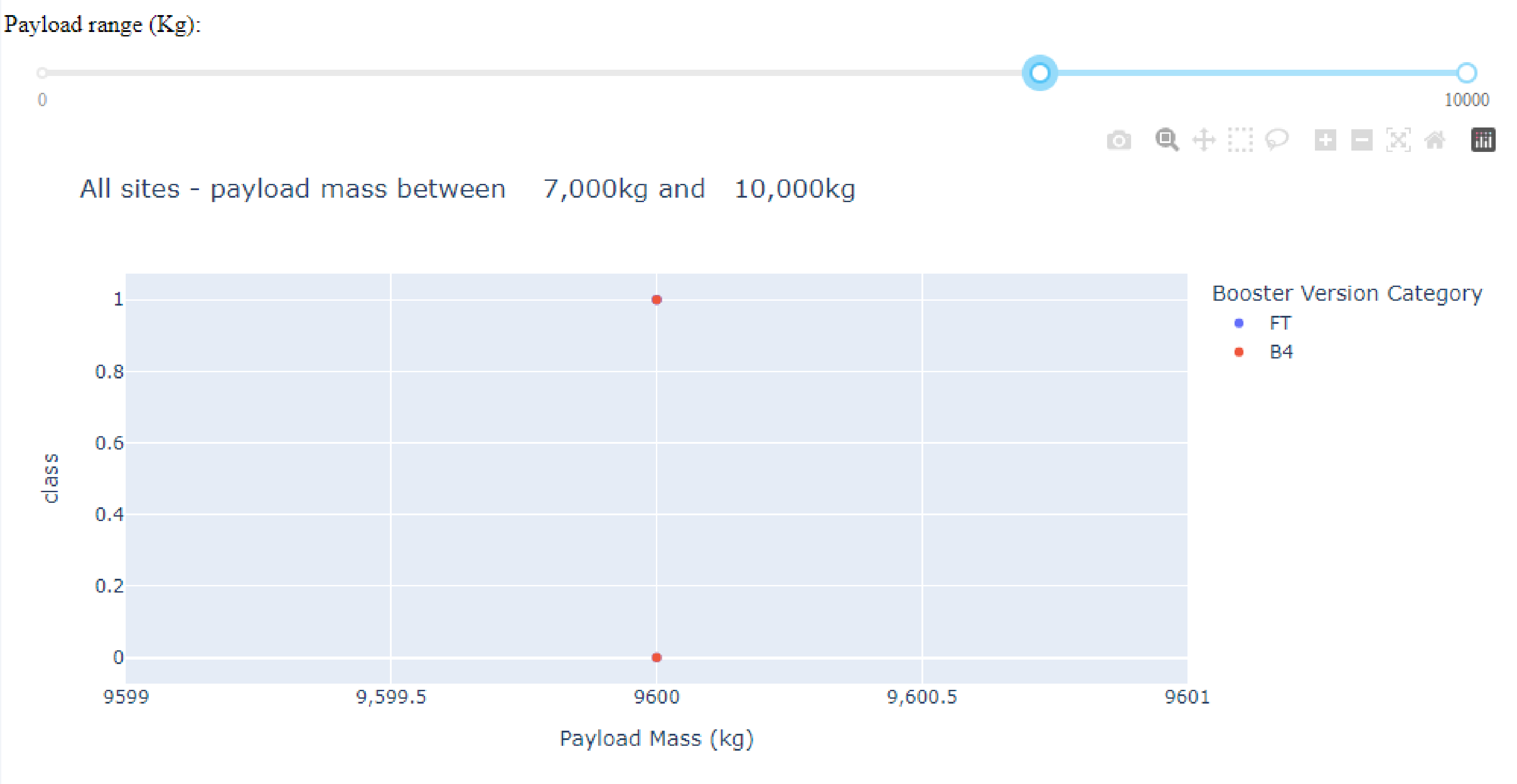
- 76.9% of launches are **successful** from **KSC-LC-39A** site.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.

Payload vs. Launch Outcome (2)



Payload range (Kg):

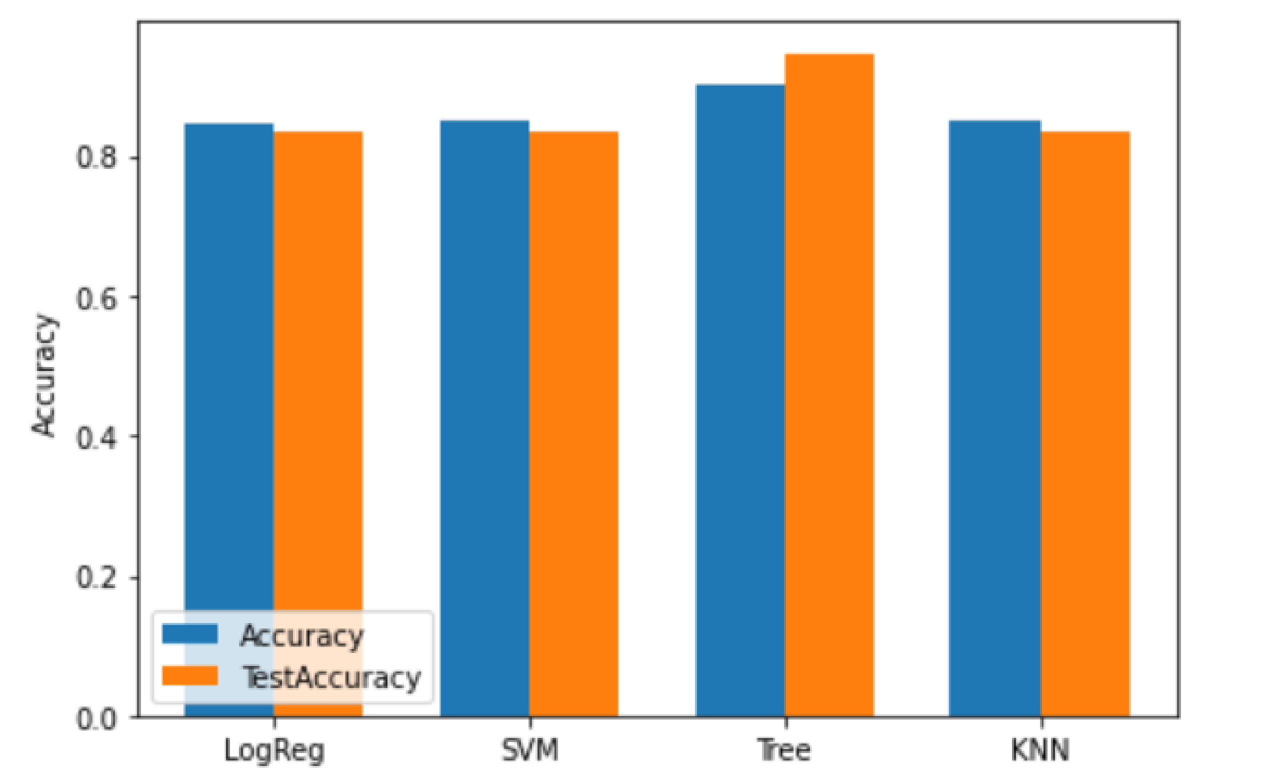
- There’s not enough data to estimate risk of launches over 7,000kg

Section 5

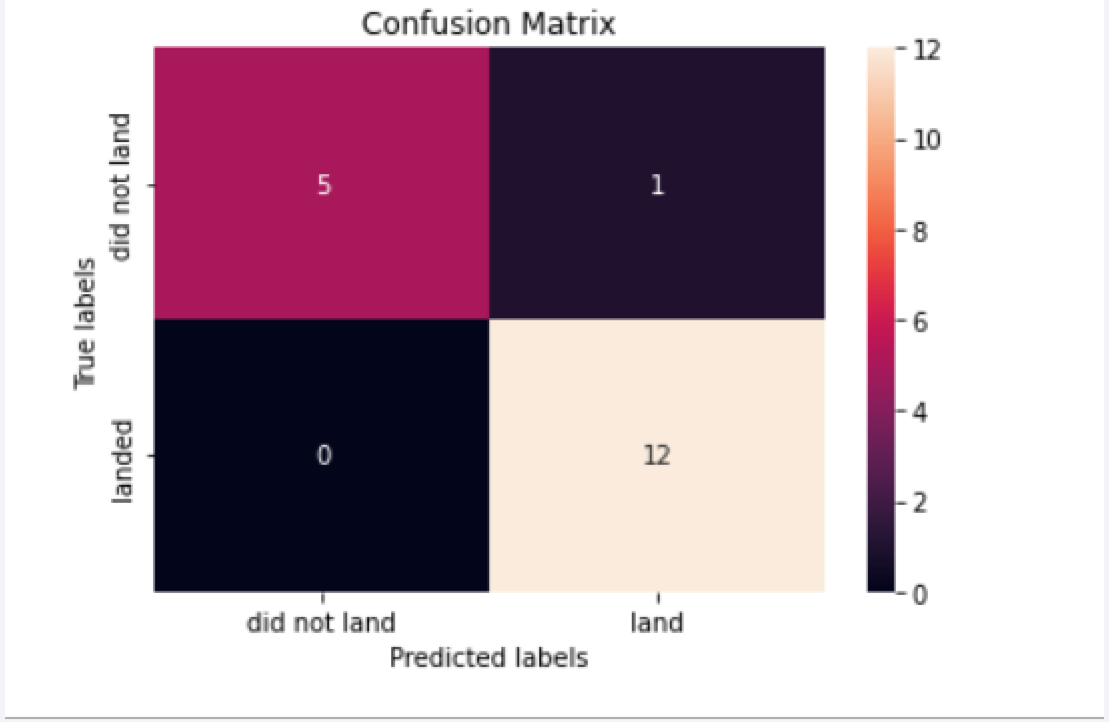
Predictive Analysis (Classification)

Classification Accuracy

- **Four classification models** were tested, and their accuracies can be seen on the beside plot.
- The model with the **highest classification accuracy** is Decision Tree Classifier, which has accuracies over than 94%.



Confusion Matrix



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Different data sources were analyzed, refining conclusions along the process
- The best launch site is KSC LC-39A
- Launches above 7,000kg are less risky
- Most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- All Jupyter Notebooks used during this Capstone can be found here: https://github.com/szirob/IBM-Data-Science-Capstone-

Thank you!