IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- **Objective is to gain a competitive advantage against SpaceX**
- **Extracting insights**
  - Different flight conditions experienced by the competitor and their impact on success
- **Game Changer**
  - Come up with a model that predicts success rate by learning from all investigated parameters
  - Cost reduction and success maximization can be achieved by taking correct actions according to predictions
- **Summary of methodologies**
  - Data Collection and Cleaning
  - Exploratory Data Analysis (EDA)
  - Data Visualization with Interactive Dashboard
  - Developing a Machine Learning Model
- **Summary of all results**
  - EDA and Visualizations enabled determining best features to predict success of flights
  - Machine Learning model showed that we can predict the success rate of our flight in advance and take actions accordingly

Introduction

- Project background and context
  - FlightNumber, LandingPad, LaunchSite, Payload and Orbit
    - Cross correlation
    - Effects on success
  - Yeary success rates of rival company
  - Fail Scenarios
- Answers found
  - Best place for launch
  - Machine Learning Techniques can be used to predict success rate in advanced
  - Cost reduction and success maximization can be achieved

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating label based on landing outcome after analyzing data
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

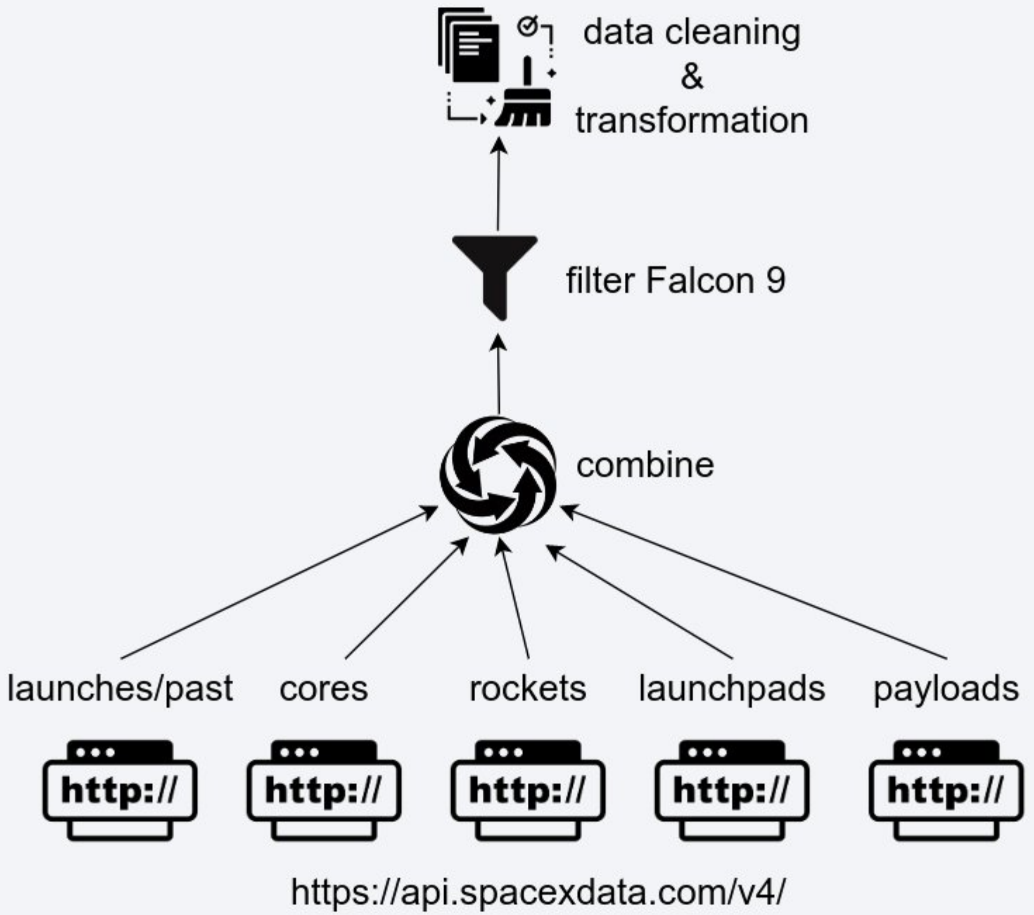
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalize cleaned and transformed data
  - Build different classification models (Logistic Regression, SVM, Decision Tree and kNN)
  - Tune all of them with Grid Search method and find the best ones
  - Evaluate the Train and Test performances
  - Compare the results and select the best one

Data Collection

- SpaceX API
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
  - https://api.spacexdata.com/v4/launches/past
- Web Scraping
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Process
  - Get data from different sources by request
  - Parse responses
  - Clean & combine data

Data Collection – SpaceX API

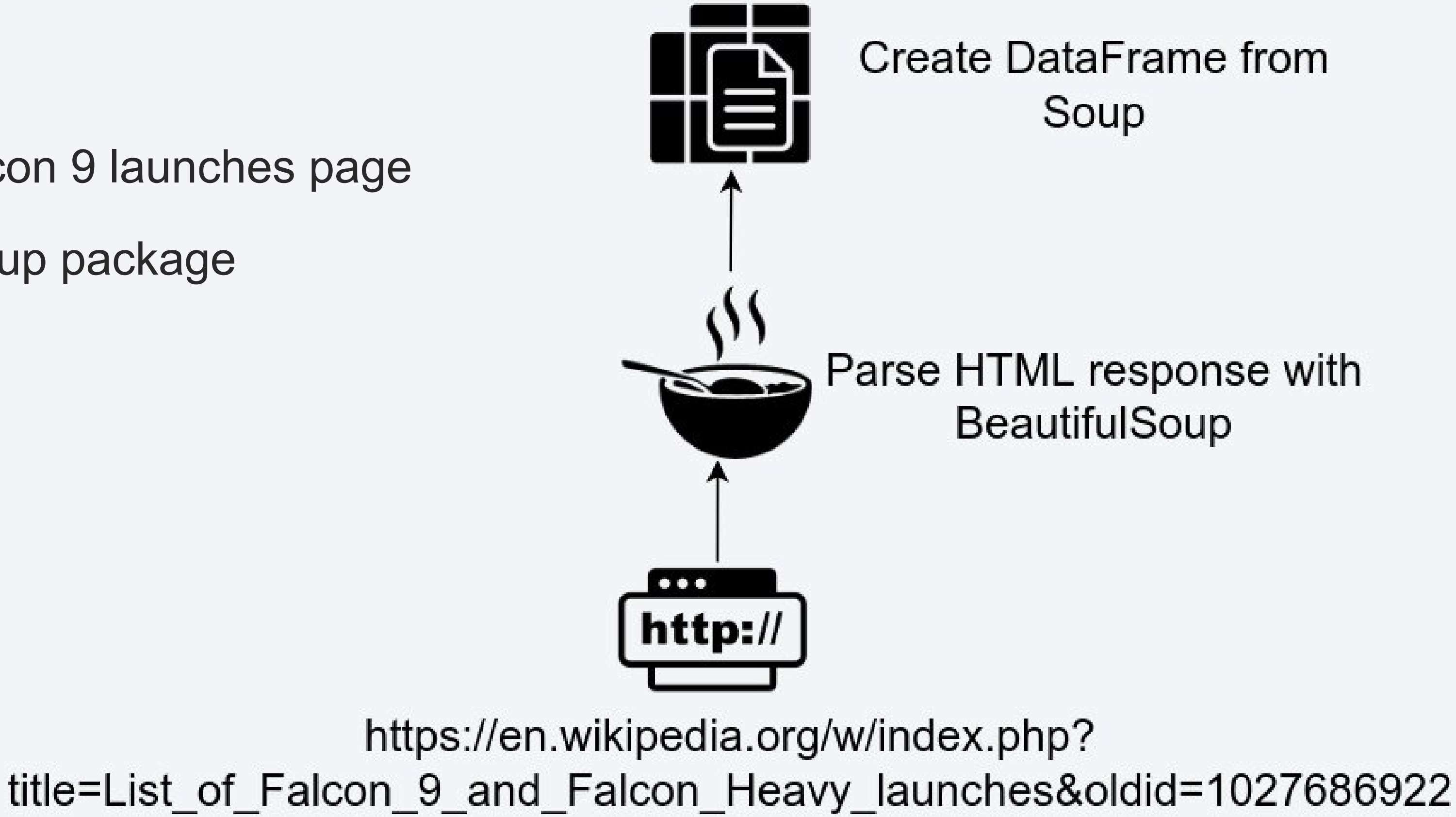
- Steps
  - Gather different data from SpaceX API
  - Combine these data into data_falcon9 DataFrame
  - Filter Falcon 9 data
  - Fill missing values with mean for PayloadMass



GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

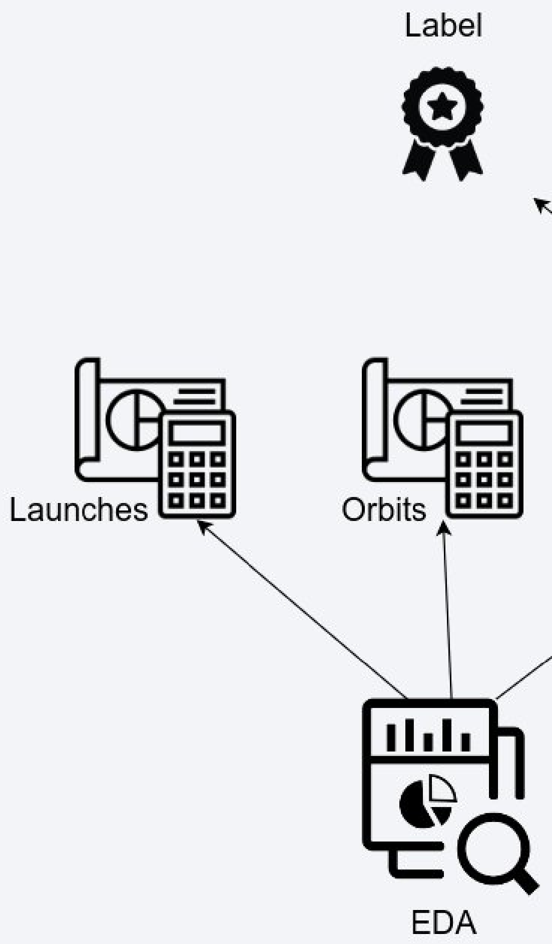
- Steps
  - Gather data from wikipedia Falcon 9 launches page
  - Parse response with BeatifulSoup package
  - Create DataFrame from soup
  - Save as CSV



GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

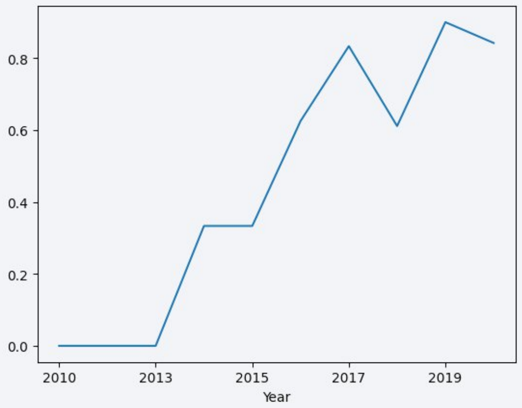
- Performed some EDA on dataset
- Calculations
  - Launches per site
  - Occurrences of each Orbit
  - Mission Outcome per Orbit
  - Create **label** from Mission Outcome

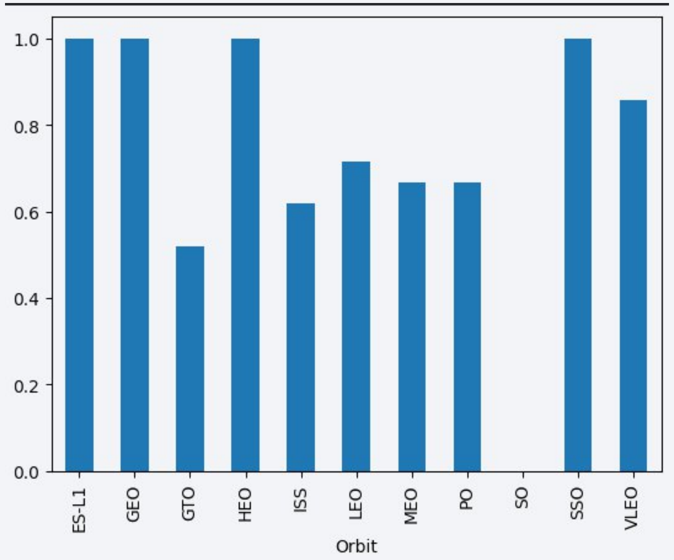


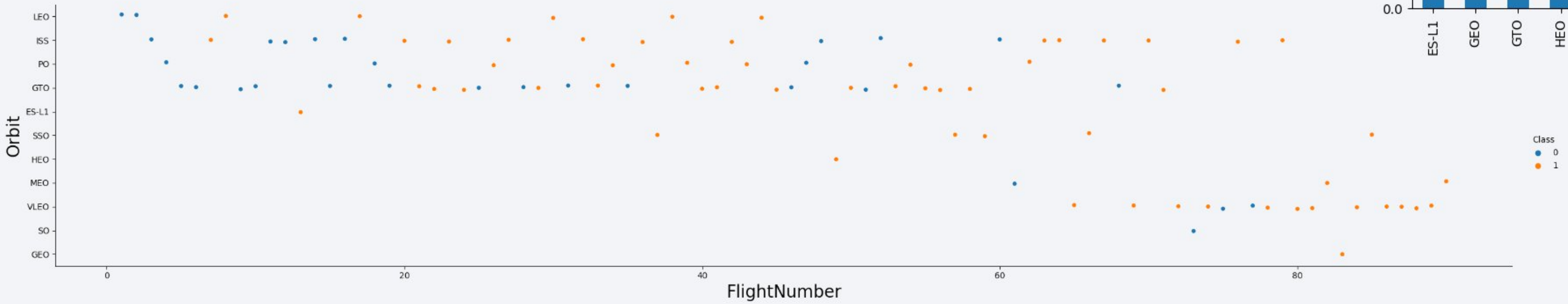
GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter, bar and line plots are used for visualization of relationship between
  - Payload Mass <-> Flight Number
  - Launch Site <-> Flight Number
  - Launch Site <-> Payload Mass
  - Orbit <-> Success Rate
  - Orbit <-> Flight Number
  - Success Rate <-> Year







GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission
  - First 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 of the first successful landing outcome in ground pad
  - 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.

GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera.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 indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates

GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

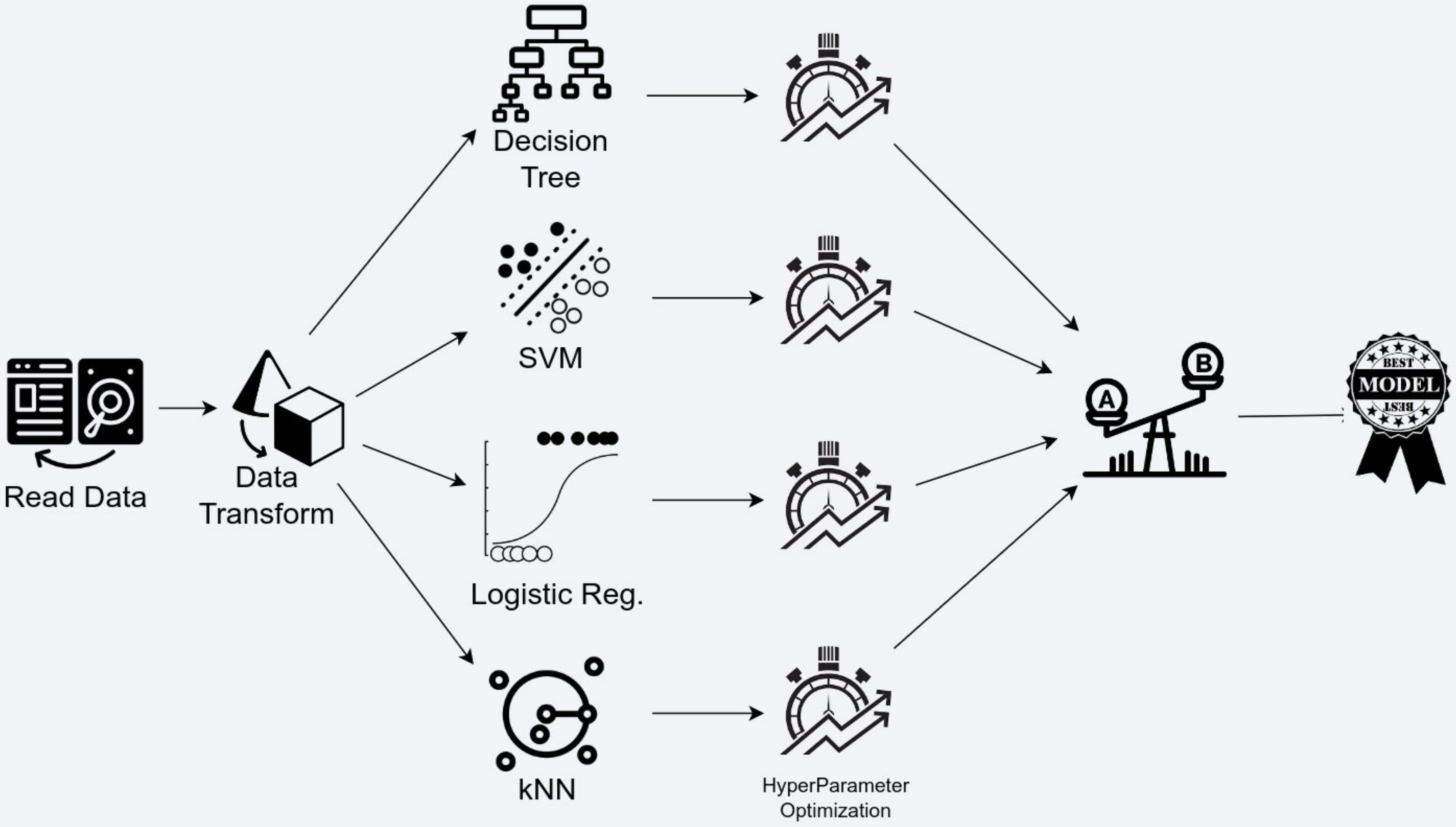
Build a Dashboard with Plotly Dash

- The following charts and plots were used to visualize data
  - Pie Chart: Percentage of launches by site / Success Failure ratios per site
  - Scatter Plot: Payload vs Success (with Payload Range Slider)
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads

GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Steps
  - Data read
  - Scale
  - Build alternative models
  - Hyperparameter Optimization
  - Compare Results
  - Select Best Model

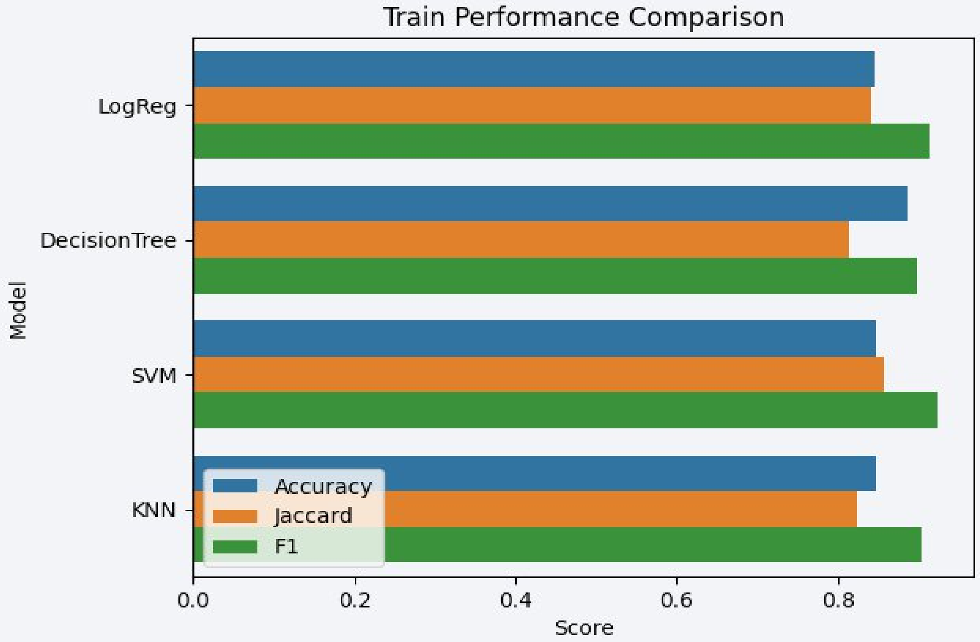


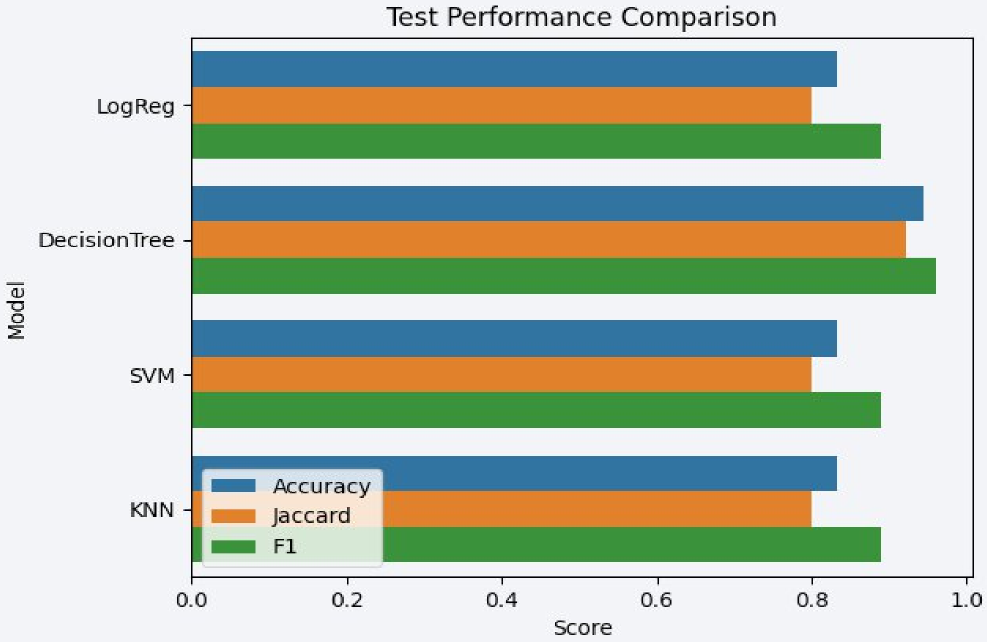
GitHub Link: https://github.com/oktay-sahinoglu/Coursera---IBM---Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites
  - The average payload of F9 v1.1 booster is 2,928 Kg
  - Heaviest cargo sent to VLEO orbit
  - GTO orbit and ISS are most visited destinations
  - Significant increase at success rate
    - Started with 2015 till 2017
    - Continued after 2018 till 2019
  - Different Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015

Results



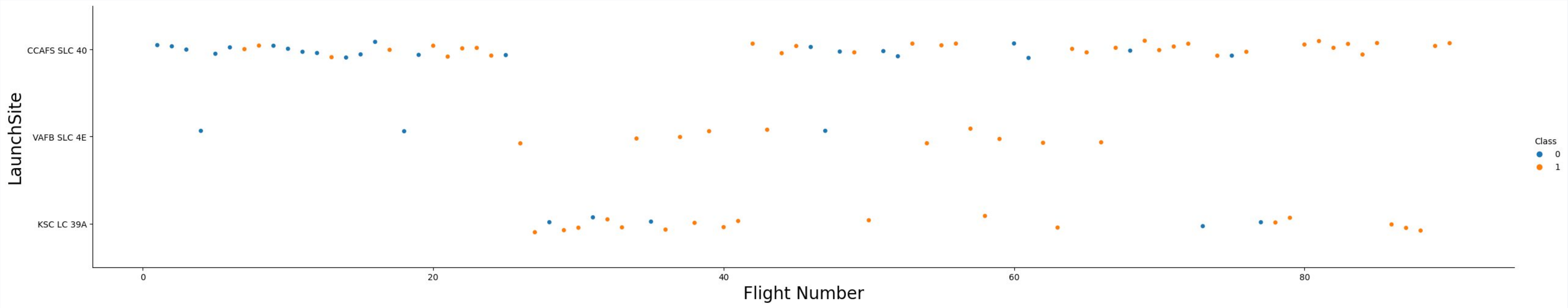


- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings

Section 2

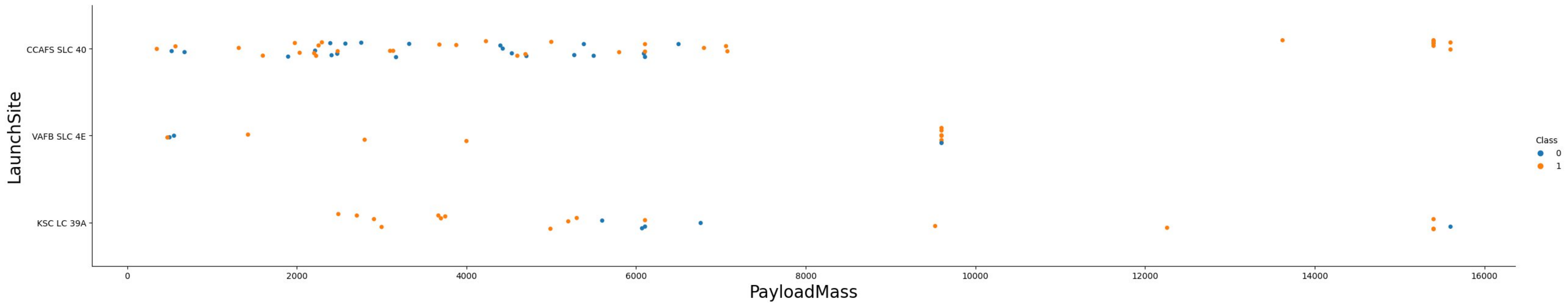
Insights drawn from EDA

Flight Number vs. Launch Site



- Most successful launch sites of recent times is CCAF5 SLC 40 and KSC LC 39A
- It’s also possible to see that the general success rate improved over time

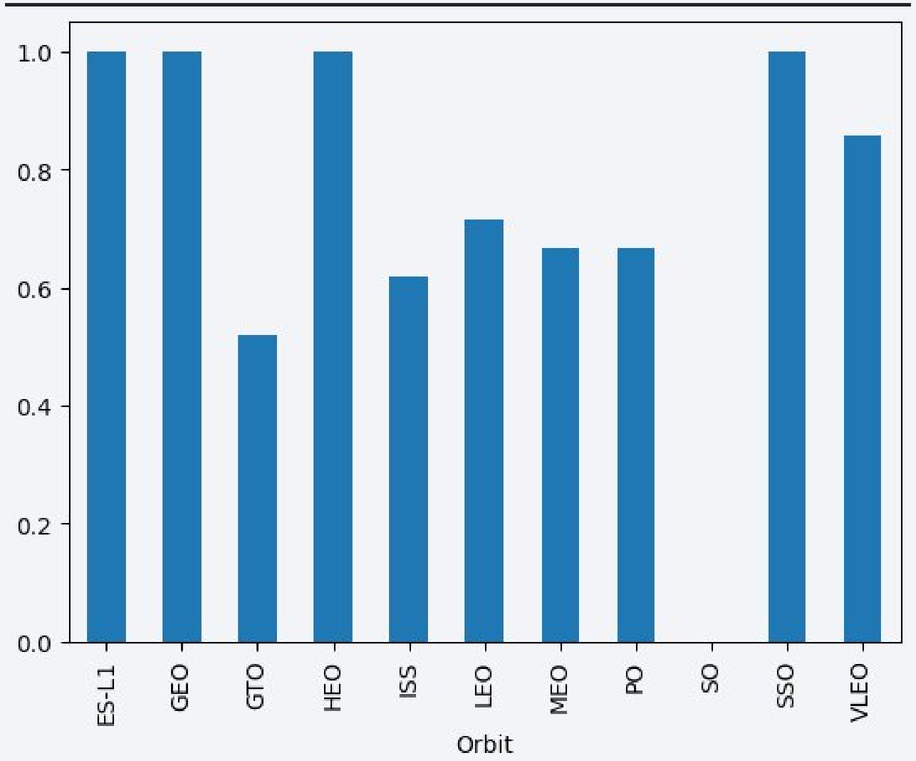
Payload vs. Launch Site



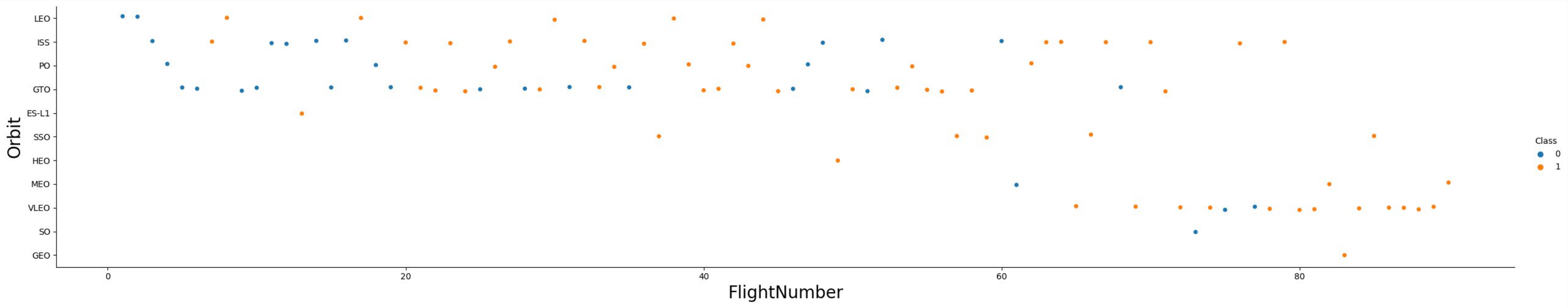
- Payloads over 9,000Kg have very high success rate
- Payloads over 12,000Kg seems to be possible only on
  - CCAFS SLC 40 and KSC LC39A launch sites.
- KSC LC39A has a 100% success rate for payloads under ~5,500Kg

Success Rate vs. Orbit Type

- 100% success rate orbits
  - ES-L1, GEO, HEO and SSO
- VLEO second most successful orbit



Flight Number vs. Orbit Type



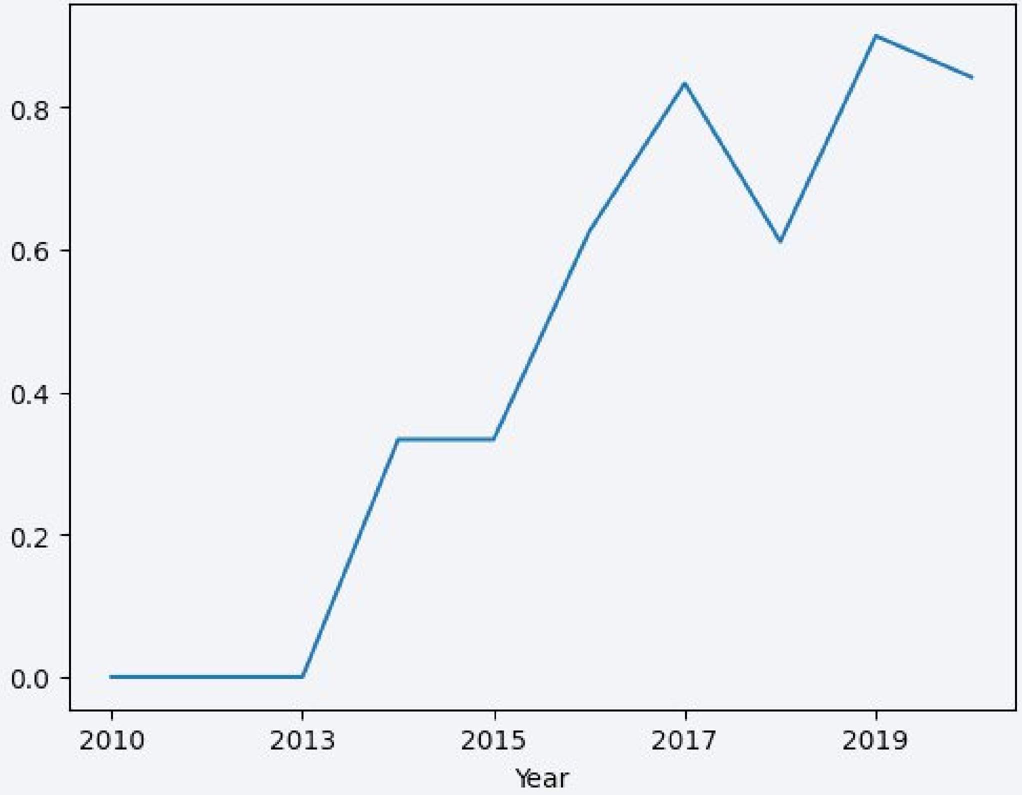
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency
- SO and GEO seem trial orbits with one flight

Payload vs. Orbit Type

- ISS and GTO
  - Most visited orbits
  - No relationship between payloads and success
- SSO has payloads <= 4000Kg, and 100% percent success rate

Launch Success Yearly Trend

- Significant increase at success rate
  - Started with 2015 till 2017
  - Continued after 2018 till 2019
- First three years seem technology maturity period



All Launch Site Names

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

- > `SELECT DISTINCT LAUNCH_SITE FROM SPACEX`
- Obtained by unique occurrences (DISTINCT) of LAUNCH_SITE in dataset

Launch Site Names 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- > `SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE ‘CCA%’ LIMIT 5`

Total Payload Mass

- TOTAL_PAYLOAD = 45,596 Kg   for NASA (CRS)

- > `SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEX WHERE CUSTOMER='NASA (CRS)'`

Average Payload Mass by F9 v1.1

- AVG_PAYLOAD = 2,928 Kg   for F9 v1.1 version boosters

- > `SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEX WHERE BOOSTER_VERSION='F9 v1.1'`

First Successful Ground Landing Date

- FIRST_SUCCESS_GP = 2015-12-22
- > `SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEX WHERE LANDING__OUTCOME='Success (ground pad)'`

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date result as the first occurrence, that happened on 12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- > `SELECT DISTINCT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME='Success (drone ship)'`

- Need to get unique booster versions so DISTINCT is used
- Filtering data by constraints

Total Number of Successful and Failure Mission Outcomes

MISSION_OUTCOME	TOTAL_NUMBER
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- > `SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUMBER FROM SPACEX GROUP BY MISSION_OUTCOME`
- Grouping mission outcomes and counting records for each group led us to the summary above

Boosters Carried Maximum Payload

- These are the boosters which have carried the maximum payload mass registered in the dataset
- >
```
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

BOOSTER_VERSION	LAUNCH_SITE
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- >
```
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX
    WHERE LANDING__OUTCOME = ‘Failure (drone ship)’ AND
    DATE BETWEEN TO_DATE(‘2015-01-01’, ‘YYYY-MM-DD’) AND
                 TO_DATE(‘2015-12-31’, ‘YYYY-MM-DD’)
```

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

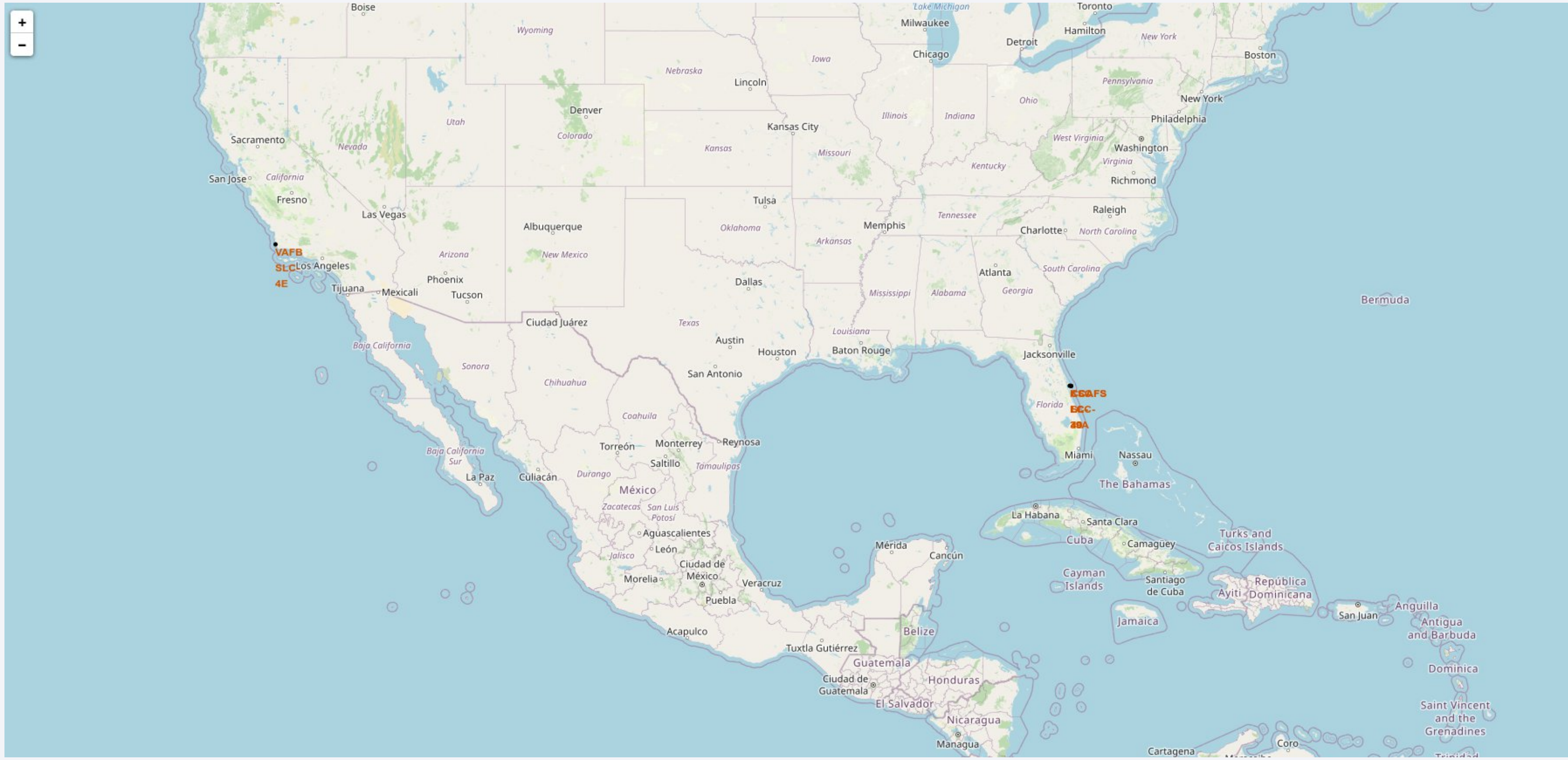
- >
```
SELECT LANDING__OUTCOME, COUNT(*) AS OUTCOME_COUNTS FROM SPACEX
    WHERE DATE BETWEEN TO_DATE('2010-06-04', 'YYYY-MM-DD') AND
                       TO_DATE('2017-03-20', 'YYYY-MM-DD')
    GROUP BY LANDING__OUTCOME ORDER BY OUTCOME_COUNTS DESC
```

LANDING__OUTCOME	OUTCOME_COUNTS
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

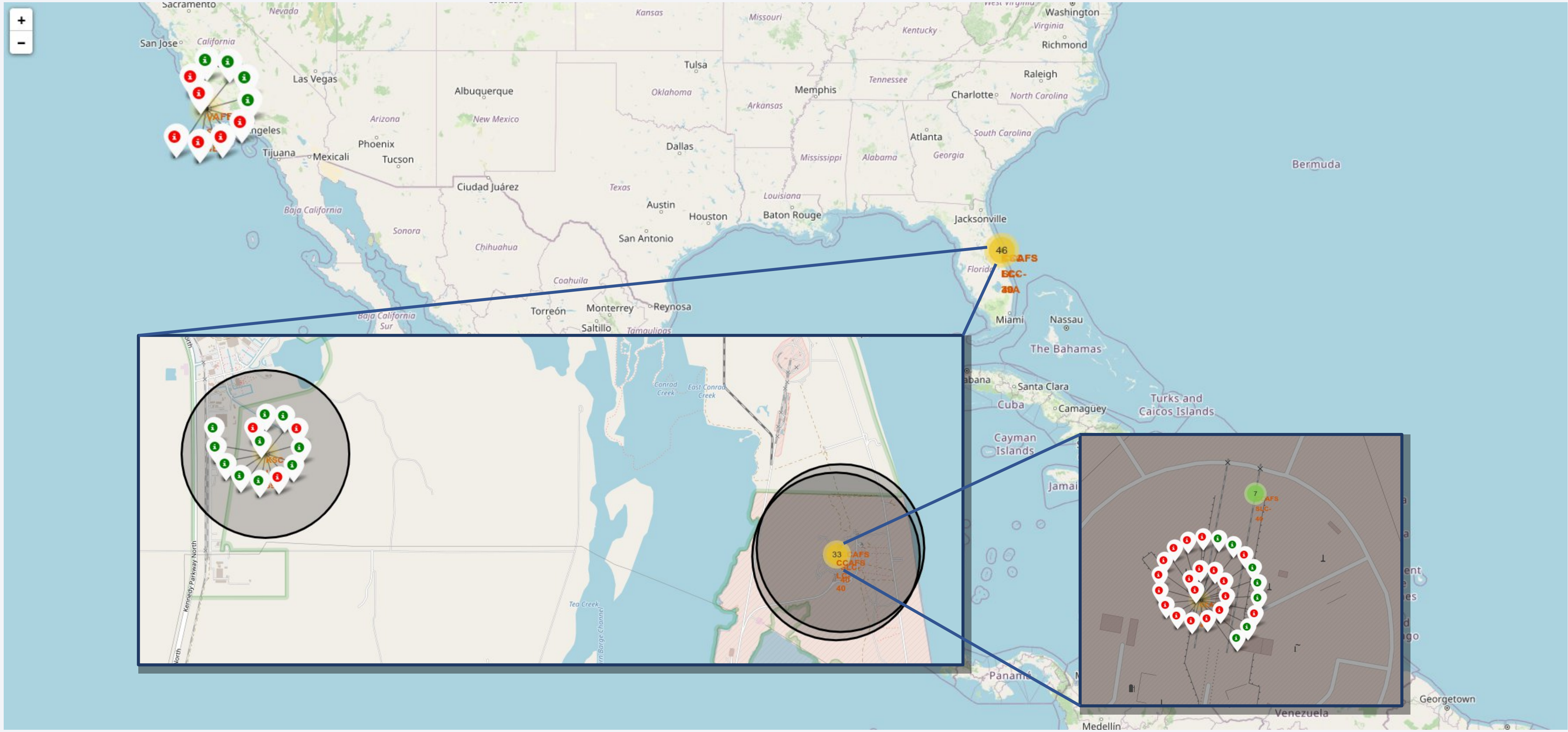
Launch Sites Proximities Analysis

Launch Sites



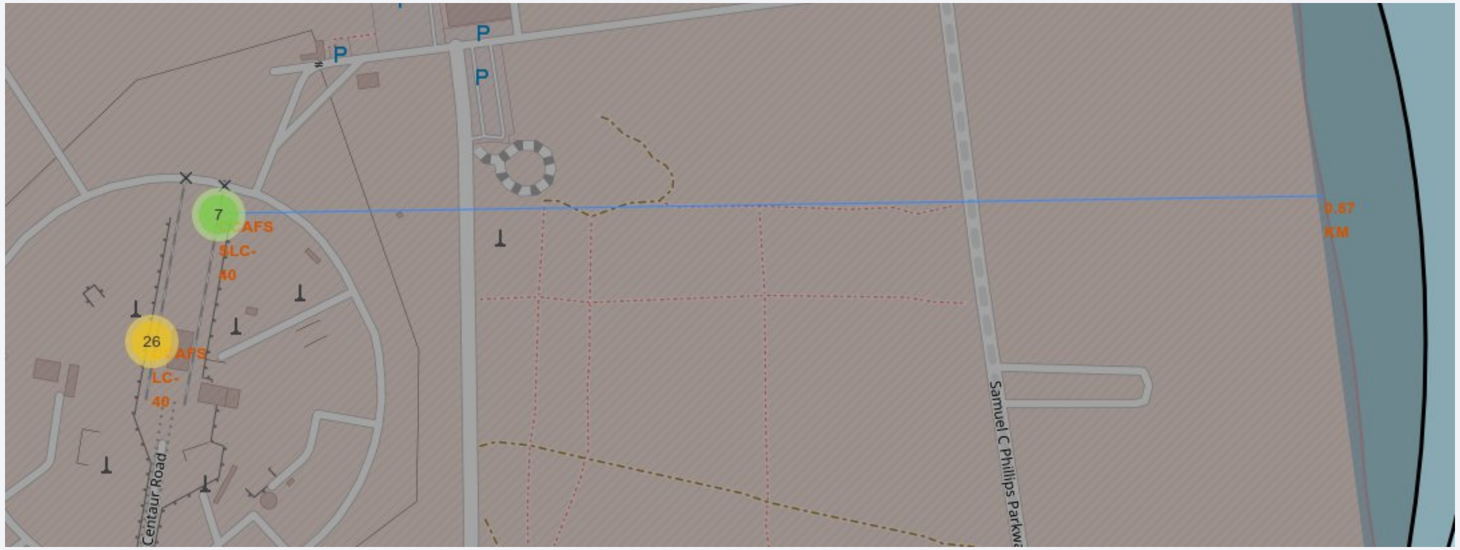
- Launch sites are near sea, probably by safety, but not too far from roads and railroads

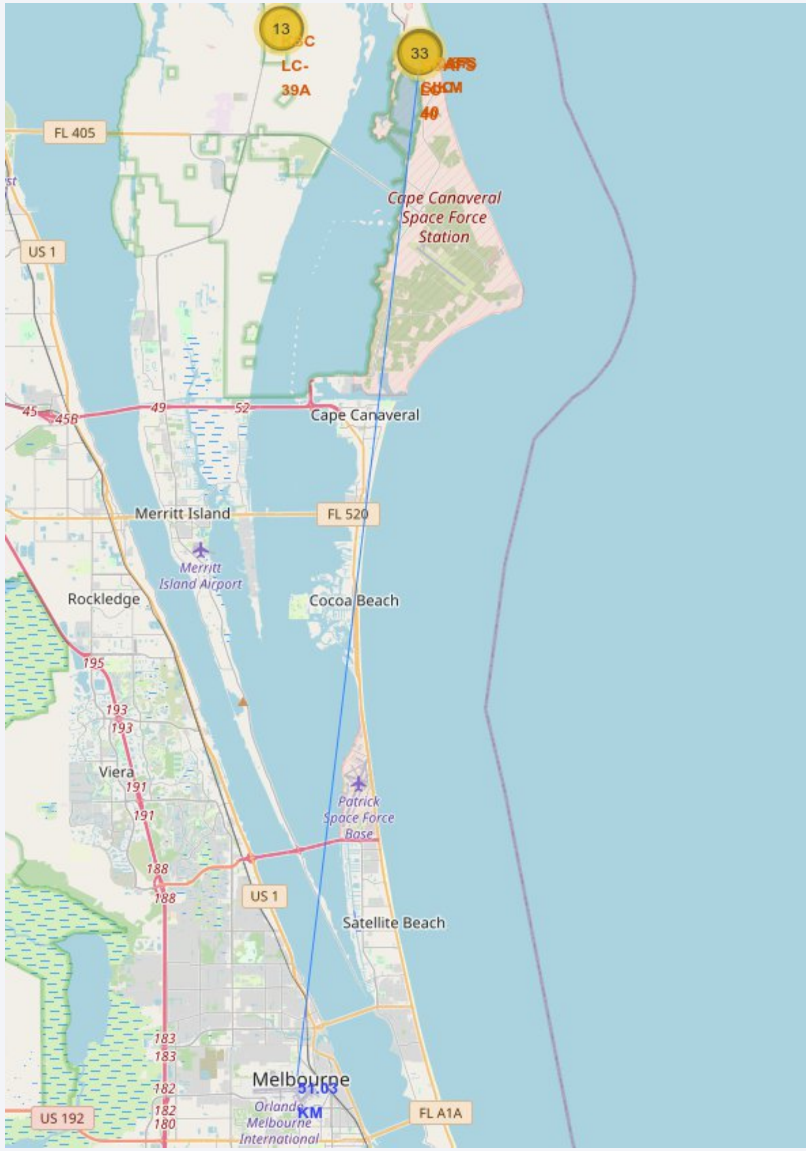
Launch Outcomes



- Green markers indicate successful and red ones indicate failure

Launch Site Proximities



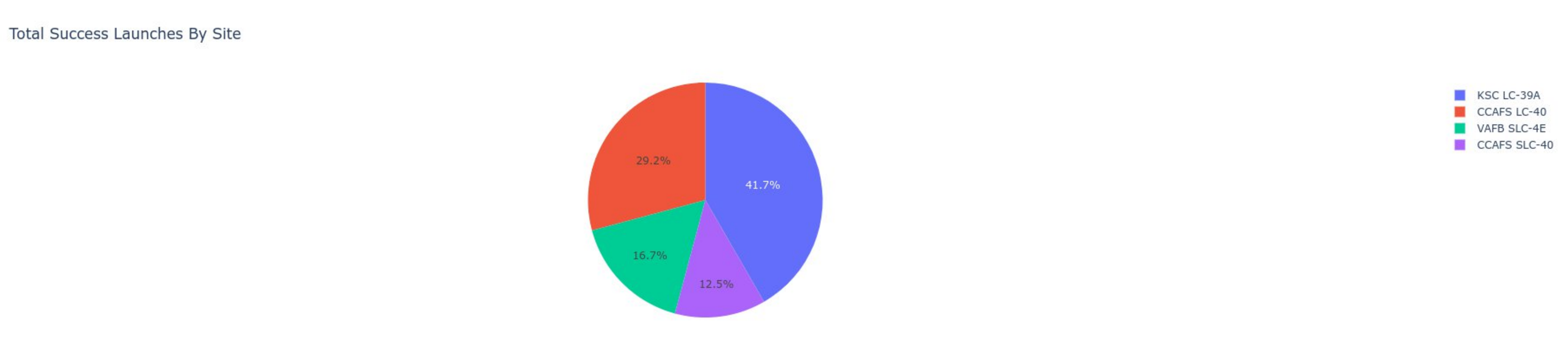


- CAFS SLC-40 has a 0.87 Km distance to coastline
- For security reasons launch site is 51 Km far from nearest city Melbourne

Section 4

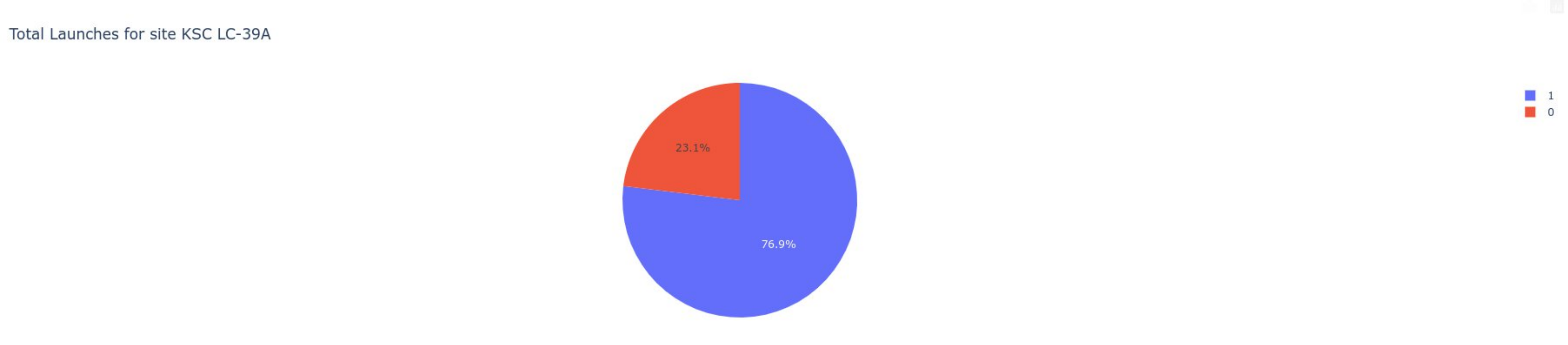
Build a Dashboard with Plotly Dash

Success Counts for All Sites



- The chart clearly shows that KSC LC-39A has the most successful launches

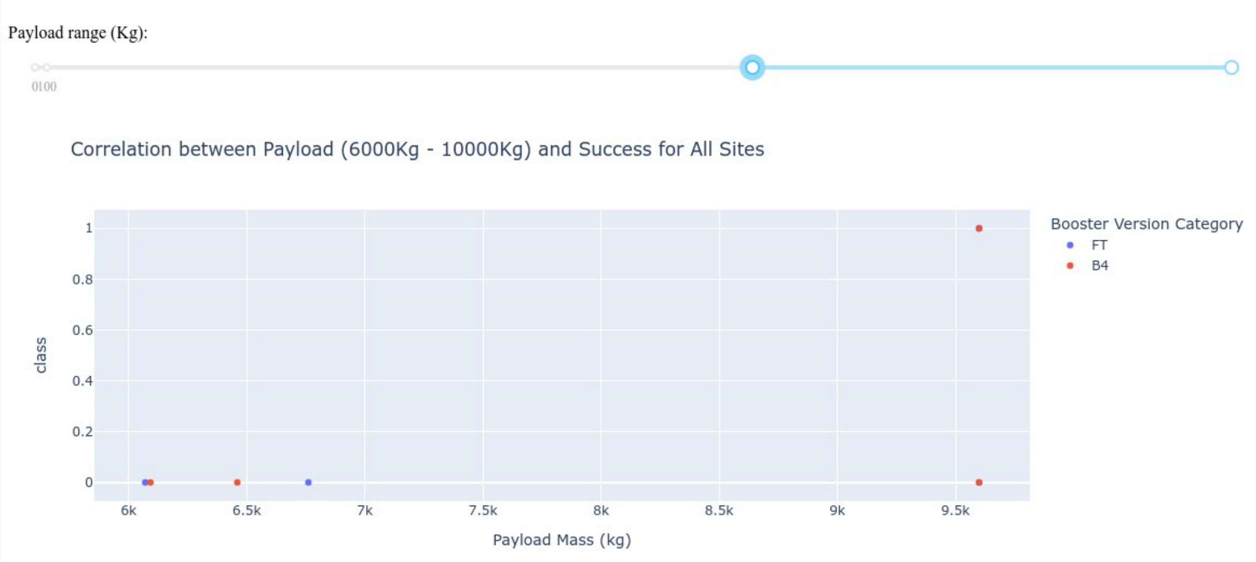
KSC LC-39A The Most Successful One



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

Payload Effects on Success

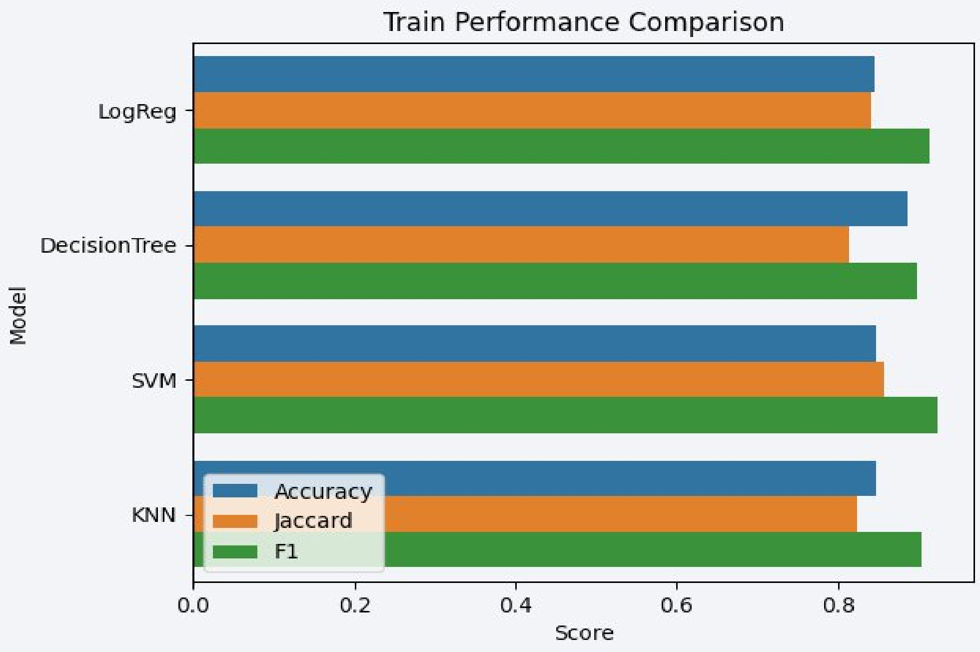
- The charts show that payloads up-to 6000 Kg have better success rate
- FT Boosters are the most successful
- Not enough data above 6000 Kg

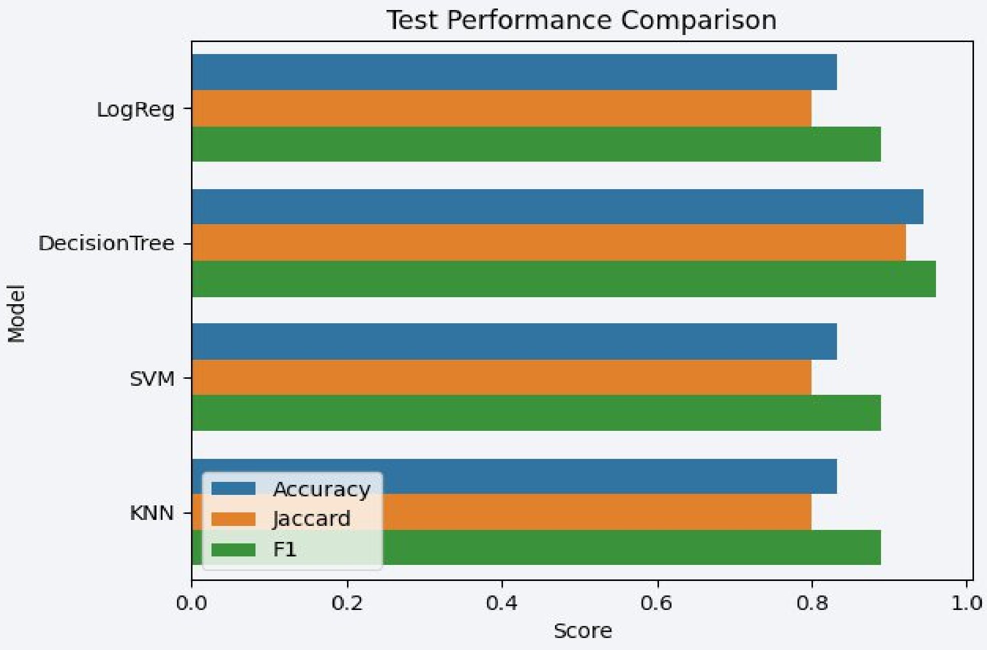


Section 5

Predictive Analysis (Classification)

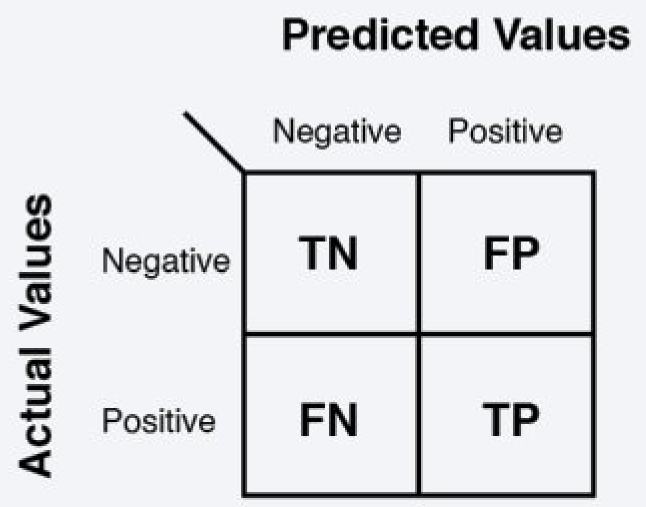
Classification Accuracy

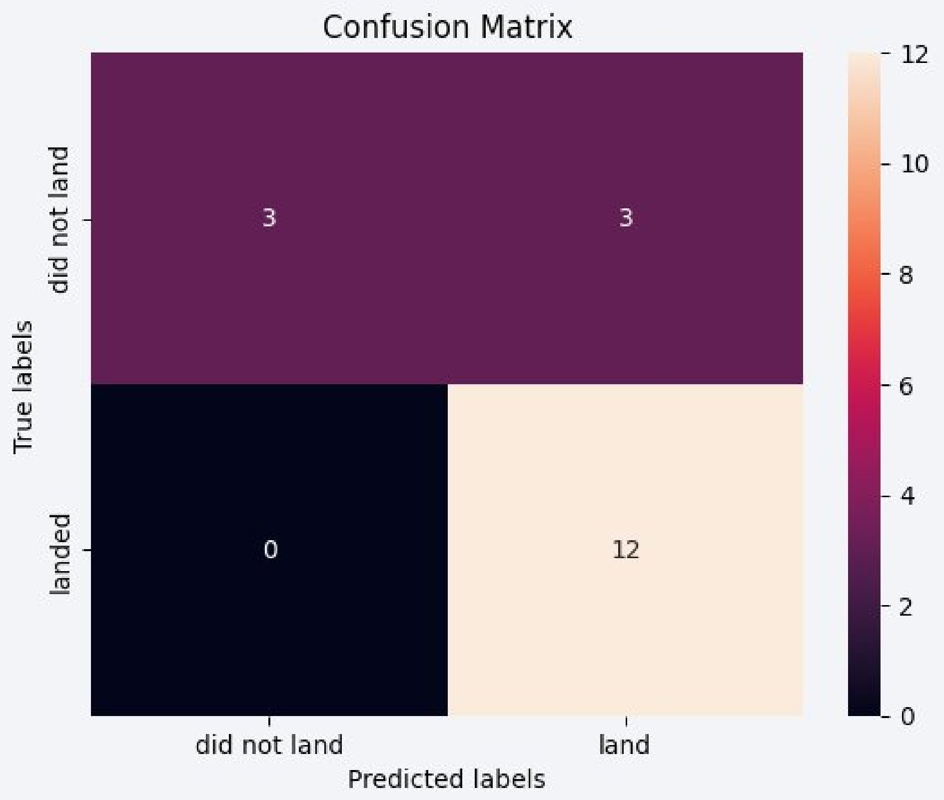




- All 4 Models are close to each other on different type of metrics for Training
- However, Decision Tree got the highest scores on Test Set
- Decision Tree can be chosen as our predictor

Confusion Matrix





- Confusion matrix is used in classification problems to assess where errors in the model were made.
- Some metrics like F1, Precision and Recall are calculated from confusion matrix

Conclusions

- The best launch site is KSC LC-39A
- If payload is under 6000 Kg, select KSC LC-39A as Launch Site
- If payload is above 9000 Kg, select CCAFS SLC-40 as Launch Site
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Although 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

- Thanks to
  - Instructors
  - IBM
  - Coursera

Thank you!