Winning Space Race with Data Science

T. Weber
23 October 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results

Executive Summary

- Summary of Methodology
  - API data collection
  - Web scraping data collection
  - Data wrangling collection
  - EDA with visualization tools
  - EDA with SQL
  - Interactive visual analytics with folium
  - Forecasting with machine learning techniques

Introduction

- As data scientist for SpaceY I will show the results of the capstone project with the goal to figure out the relationships and requirements for successfully launch and landing rockets based on SpaceX data
- I trained a ML model to predict the probability of successful landing

Section 1

Methodology

Methodology

Executive Summary

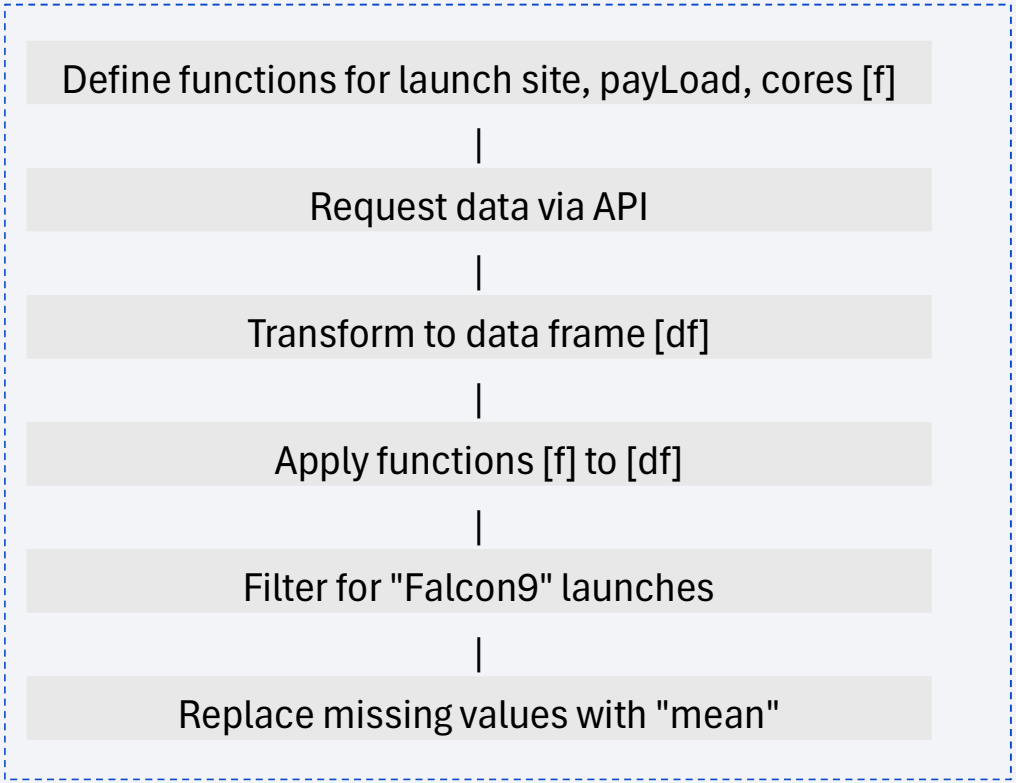
- Data collection methodology:
  - SpaceX API and web scraping from wiki
- Perform data wrangling
  - Data cleaning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare different ML models and choose the best

Data Collection

- Following slides shows the method to collect the required data

Data Collection – SpaceX API

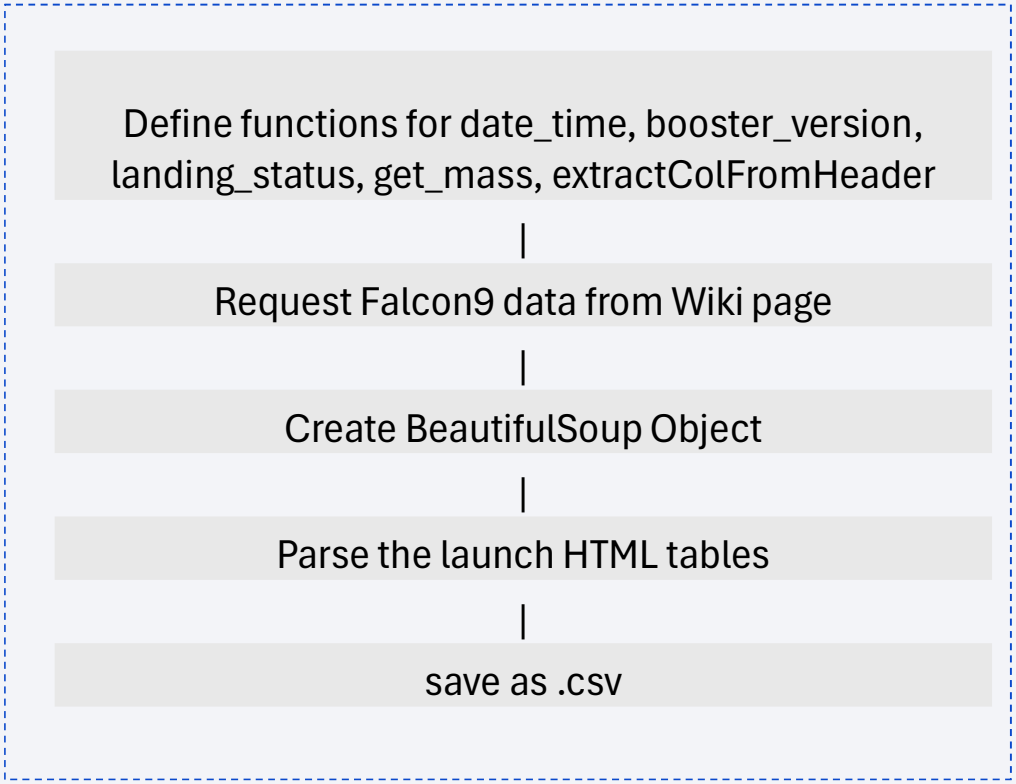
- Used libraries:
  - requests, pandas, numpy and datetime
- Keywords of code:
  - GET requests
  - use of json and transform into pandas data frame
  - column payload has missing values which were replaced by "mean" value using ".replace(np.nan,PayloadMass_mean,inplace=True)"



GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/f54635c1f7c4a15cd81a96e68731c8c5618f5946/Capstone_jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

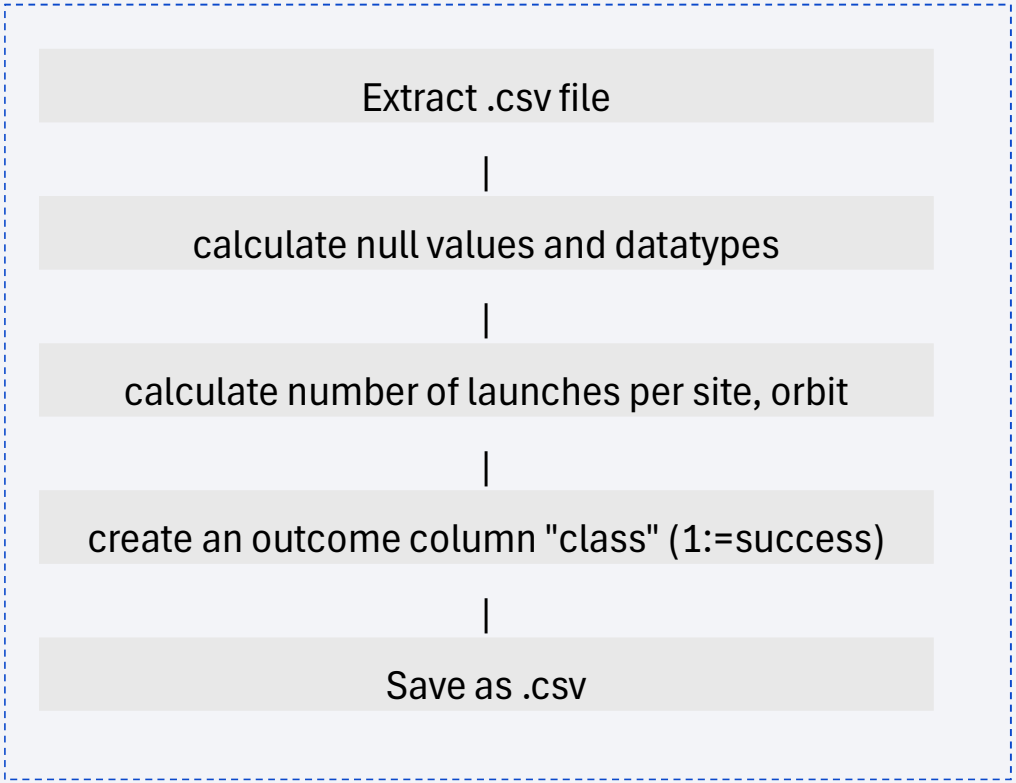
- Used libraries:
  - requests, pandas, beautifulsoup
- Keywords of code:
  - GET requests
  - soup = BeautifulSoup(response.text, "html.parser")



GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/c0e05608bf0ed5ded0a09db446a8047d2f992b1a/Capstone_jupyter-labs-webscraping.ipynb

Data Wrangling

- Used libraries:
  - pandas, numpy
- Keywords of code:
  - Use of Value.counts()
  - df['Class'] = df['Outcome'].isin(['True Ocean', 'True RTLS', 'True ASDS']).astype(int)



GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/94a72ba629f5cb568943e0dec7f4cfa511cd2bc3/Capstone_jupyter-labs-webscraping.ipynb

EDA with Data Visualization

- Used libraries:
  - Pandas, numpy, seaborn, matplotlib
- Plotted charts
  - Scatterplot with class group
    - payload vs flight number to see if there is trend over flightnumber
    - launch site vs flight number
    - Launch site vs payload
    - Orbit vs flight number
    - Orbit vs payload
  - Bar chart
    - Success rate vs orbit
  - Line chart
    - Success rate vs year
- Feature engineering

GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/39afb64df41f07f7302535c0db54eab180fee186/Capstone_jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Used libraries:
  - sqlalchemy, ipython with prettytable, pandas
- Keyword for SQL:
  - … like ‚%%‘
  - … Min() and distinct()
  - … where … between … and …
  - … count()
  - … subquery in where statement
  - … subtsr() to extract year and month from Date

load the .csv and transform into sql structure

|

create and modify using sql statements

GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/b62c406efafc4f6495b077c8313e74551dbc209a/Capstone_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To show the nearest launch sites and nearest important landwarks like railways, highways and cities the folium markers were used
- To measure the distances between spots, polylines were used
- Color green represents successes and red failed launches

GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/b62c406efafc4f6495b077c8313e74551dbc209a/Capstone_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build a Dashboard with Plotly Dash

- Used libraries:
  - import pandas as pd
  - import dash
  - from dash import html, dcc
  - from dash.dependencies import Input, Output
  - import plotly.express as px
- Added to dashboard:
  - Piechart with location selection option
  - Scatterplot with slider for payload variation
  - Excel bar chart for booster analysis
- MISC:
  - I used callback functions for interactive selections

GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/5826923c213c2018089edb9c0c9fa3b74485c7d8/spacex-dash-app(1).py

Predictive Analysis (Classification)

- Used libraries:
  - Numpy, pandas, seaborn, scikit-learn
- Important used functions
  - preprocessing.StandardScaler() from scikit-learn
  - train_test_split() from scikit-learn

Load data from csv as df with pandas

|

Create target dataset y with numpy

|

Standardize feature data set X

|

Split in 80% train and 20% test data

|

apply logistic regression, svm, decision tree and KNN models and find the best fit parameter with GridSearchCV

|

calculate the accuracy of the model and show confusion_matrix

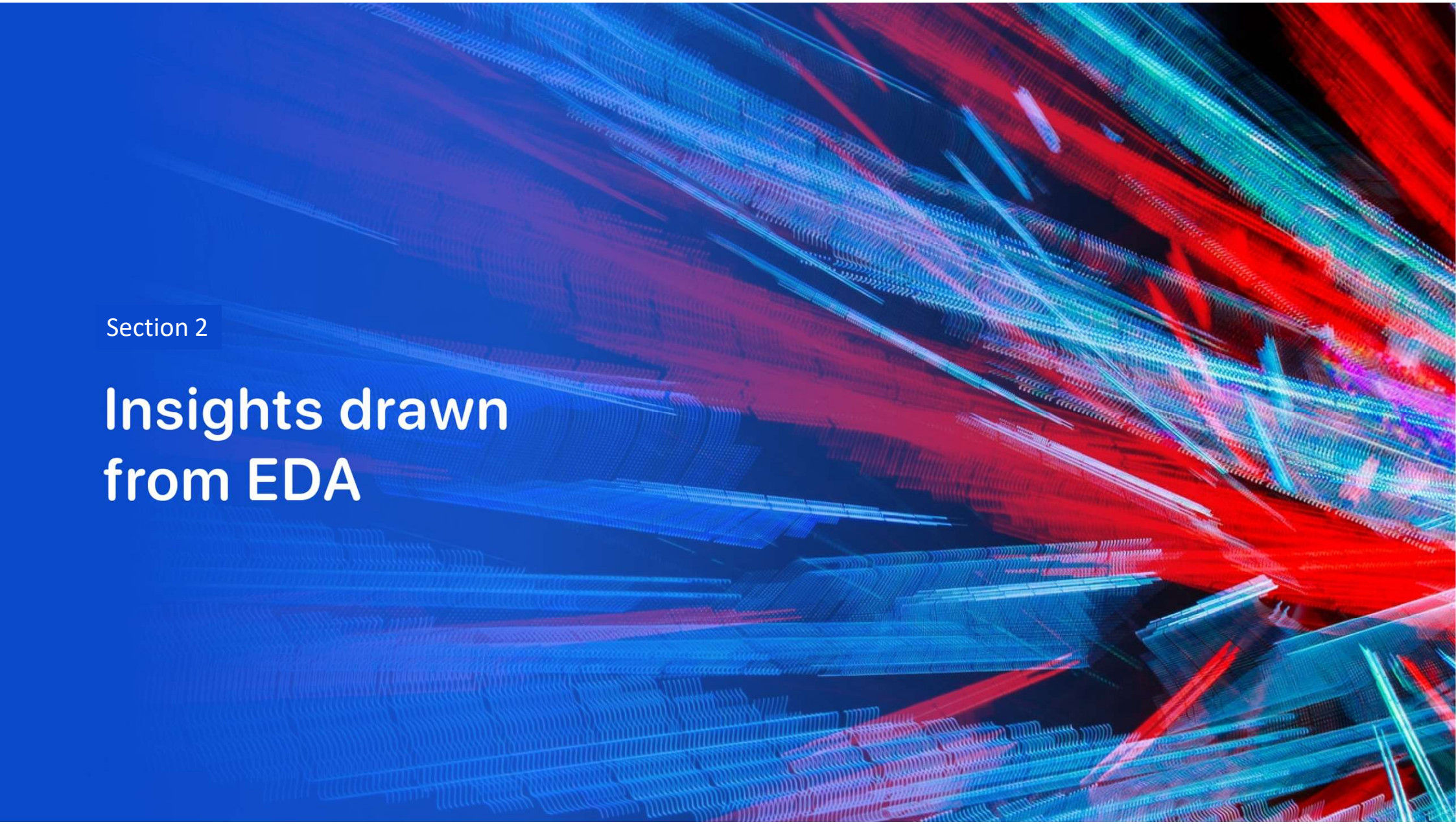
|

compare models

GitHub URL: https://github.com/Schnake13/IBM_DS_Cap/blob/9ee99535c91c1bb9ac9117b7eba34398b73a4cf7/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

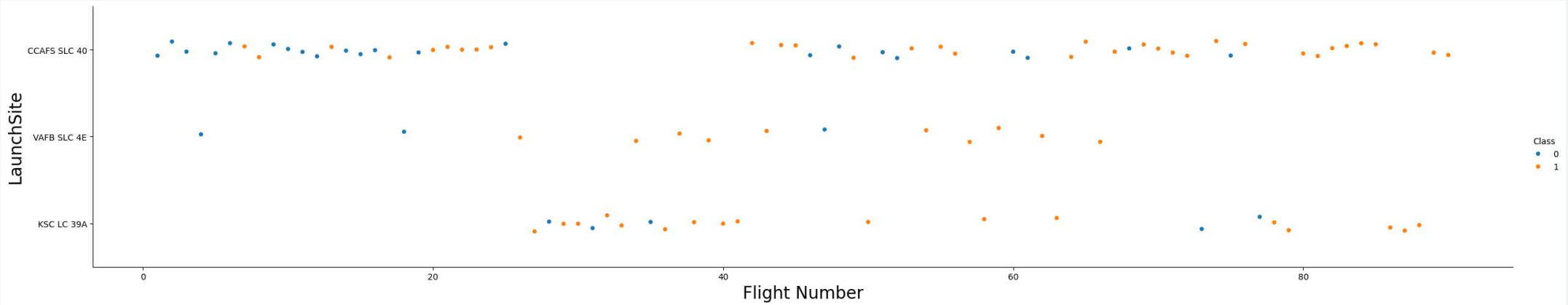
Results

- Exploratory data analysis has shown that
  - the successful landing outcomes strongly increased since 2015
  - CCAFS SLC40 has the greatest number of landings
  - Highest success rate have ES-L1, GEO, HEO and SSO orbits
  - All launch sites are close the coast and couple thousands kilometer away from the equator line
  - Railways are in close proximity because of transportation advantages
  - Coast line is close to the launch site because of possible water landing tests
- Predictive analysis results
  - The ML model shows a prediction probability of 83.3%


Section 2
Insights drawn
from EDA

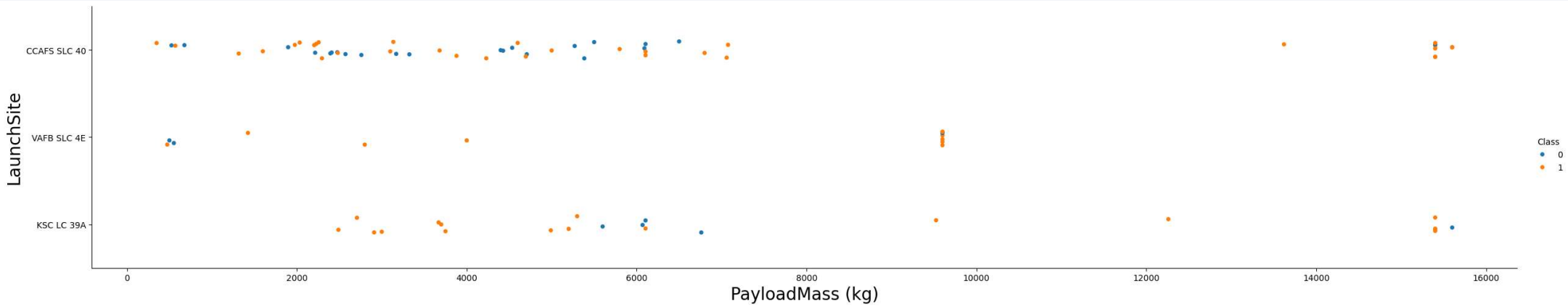
Flight Number vs. Launch Site

- The first ~25 and the launches higher than flight number 40 applied at Cape Canaveral which has also the higest number of landings
- between flight 25 and 40 almost all launches applied at Kennedy Space Center
- Less launches applied at Vandenberg Space Launch Complex but most of them passed
- There is positive trend to higher success of launches and flight number (learning curve)



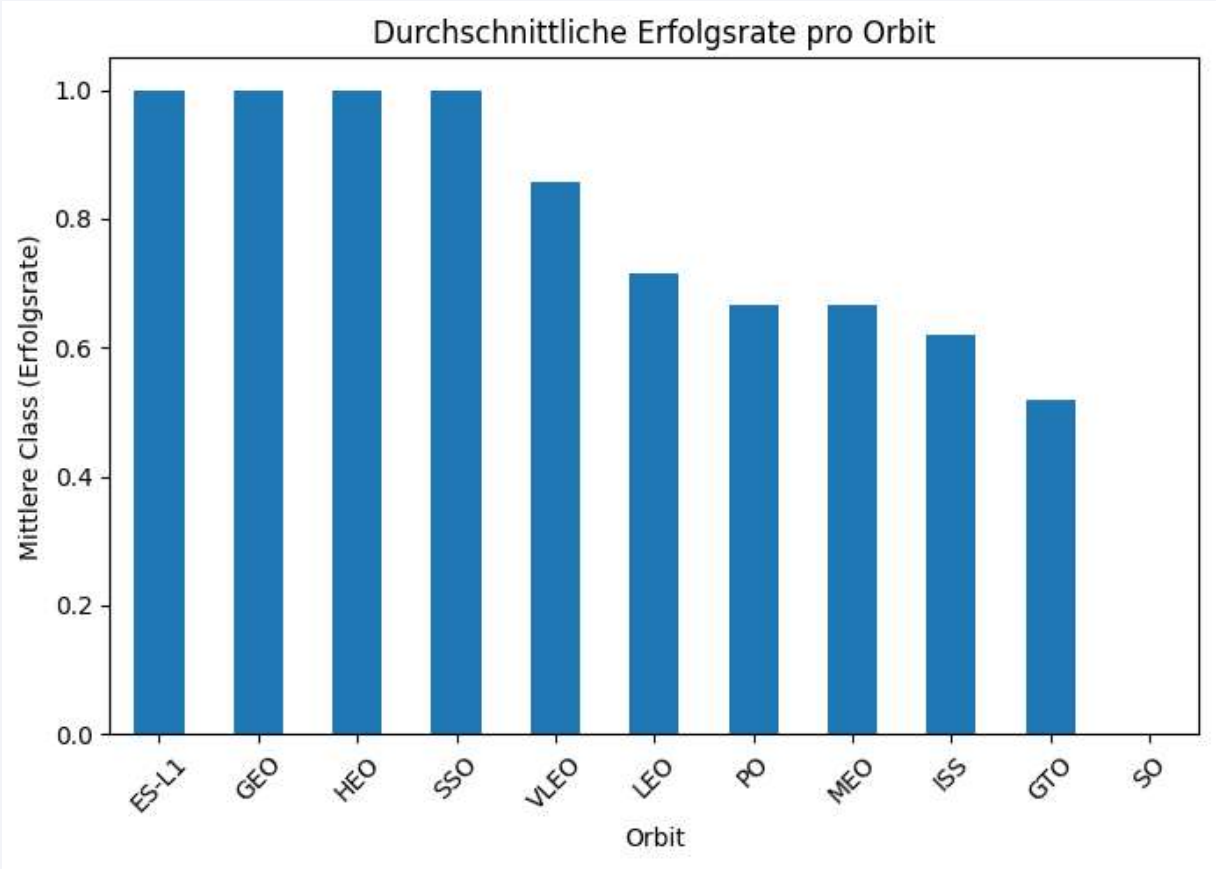
Payload vs. Launch Site

- Most of the payload for CCAFS SLC and KSC LC less than 7t
- no rockets launched for heavypayload mass(greater than 10t) for VAFB-SLC
- It was tested a wide range of payloads < 7t but only 2-3 at higher payloads (~9t and ~15t)



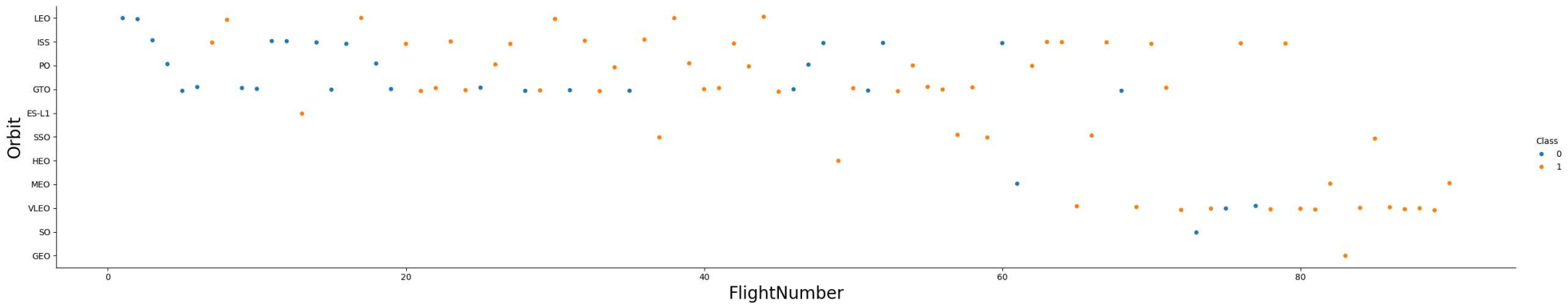
Success Rate vs. Orbit Type

- Highest success rate have ES-L1, GEO, HEO and SSO
- Lowest success rate have the orbit GTO



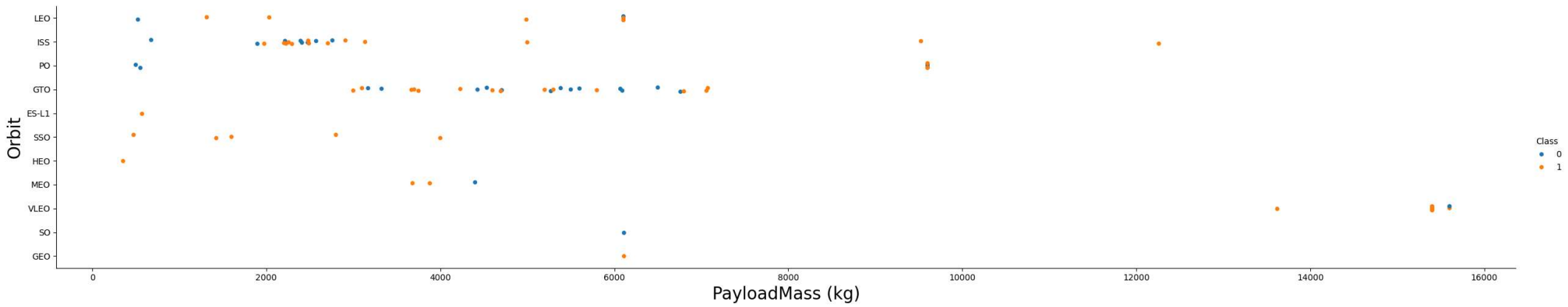
Flight Number vs. Orbit Type

- At LEO the success was stable with flight number
- For GTO there seems to be no relationship between orbit an flight number



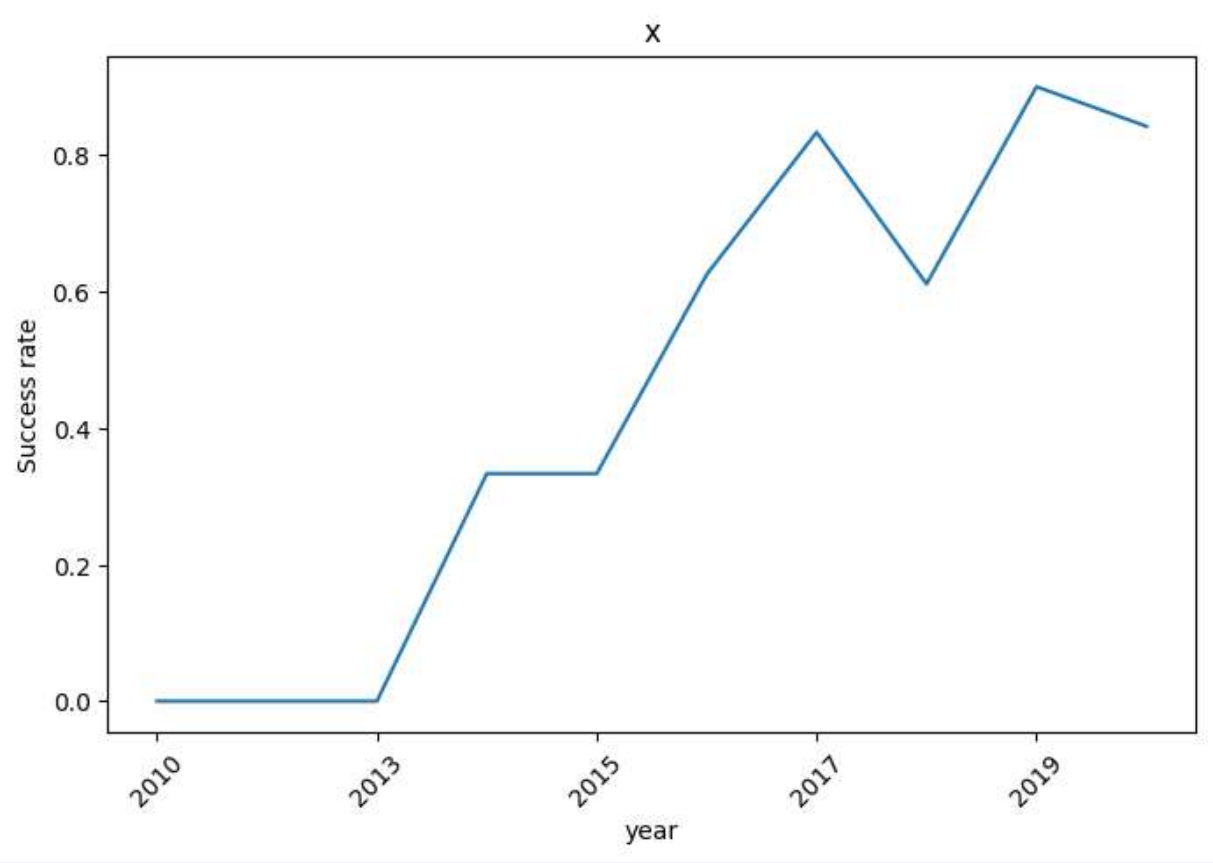
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- 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

- the success rate since 2013 kept increasing till 2017 (stable in 2014) after 2015 it started increasing



All Launch Site Names

- There are four unique launch sites
- %sql select distinct(Launch_Site) from SPACEXTABLE

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

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	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	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

- %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5

Total Payload Mass

- total payload mass carried by boosters launched by NASA (CRS) is 48213kg

- %sql select customer, sum([PAYLOAD_MASS__KG_]) as sumPL from SPACEXTABLE where Customer like'%NASA (CRS)%'

Customer	sumPL
NASA (CRS)	48213

Average Payload Mass by F9 v1.1

- average payload mass carried by booster version F9 v1.1 = 2534kg

- %sql select customer, avg([PAYLOAD_MASS__KG_]) as avgPL from SPACEXTABLE where Booster_Version like 'F9 v1.1%'

Customer	avgPL
MDA	2534.6666666666665

First Successful Ground Landing Date

- %sql select Min(Date) from SPACEXTABLE where Landing_Outcome ='Success (ground pad)’
- First ground landing date was 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select Booster_Version, [PAYLOAD_MASS__KG_] from SPACEXTABLE where Landing_Outcome ='Success (drone ship)' and [PAYLOAD_MASS__KG_] between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Total number = 101
- select count([Mission_Outcome]) as mo_fs from SPACEXTABLE

Boosters Carried Maximum Payload

- %sql select [Booster_Version] from SPACEXTABLE where [PAYLOAD_MASS__KG_] = (select max([PAYLOAD_MASS__KG_]) from SPACEXTABLE)

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

- month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015:

year	month	Booster_Version	Launch_Site	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- %sql select substr(Date,0,5) as year,substr(Date,6,2) as month, [Booster_Version], [Launch_Site] , [Landing_Outcome]from SPACEXTABLE where Date like '%2015%' and [Landing_Outcome] = 'Failure (drone ship)'

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

- Since 2015 the passed landings significantly increased

- %sql select substr(Date,0,5) as year,substr(Date,6,2) as month, [Booster_Version], [Launch_Site] , [Landing_Outcome]from SPACEXTABLE where Date like '%2015%' and [Landing_Outcome] = 'Failure (drone ship)'

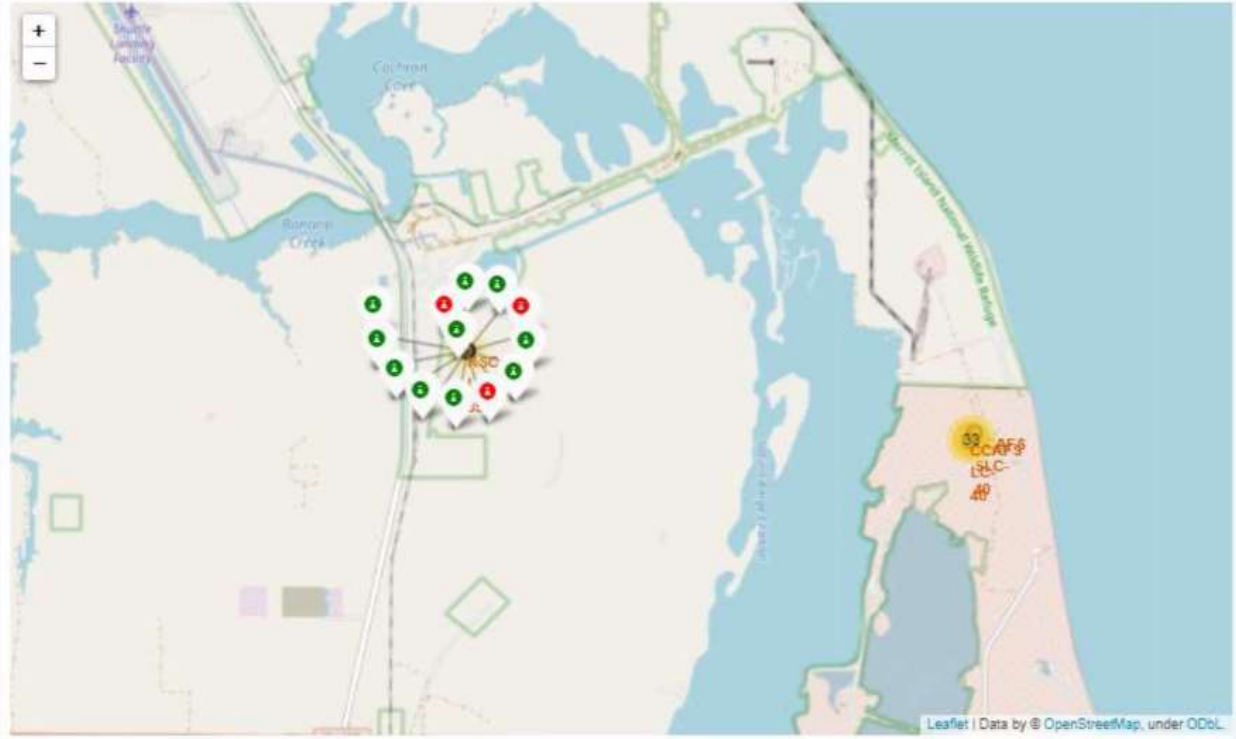

Section 3
Launch Sites
Proximities Analysis

Launch site locations

- All launch sites are close the coast and couple thousands kilometer away from the equator line

Success rate of rocket launches

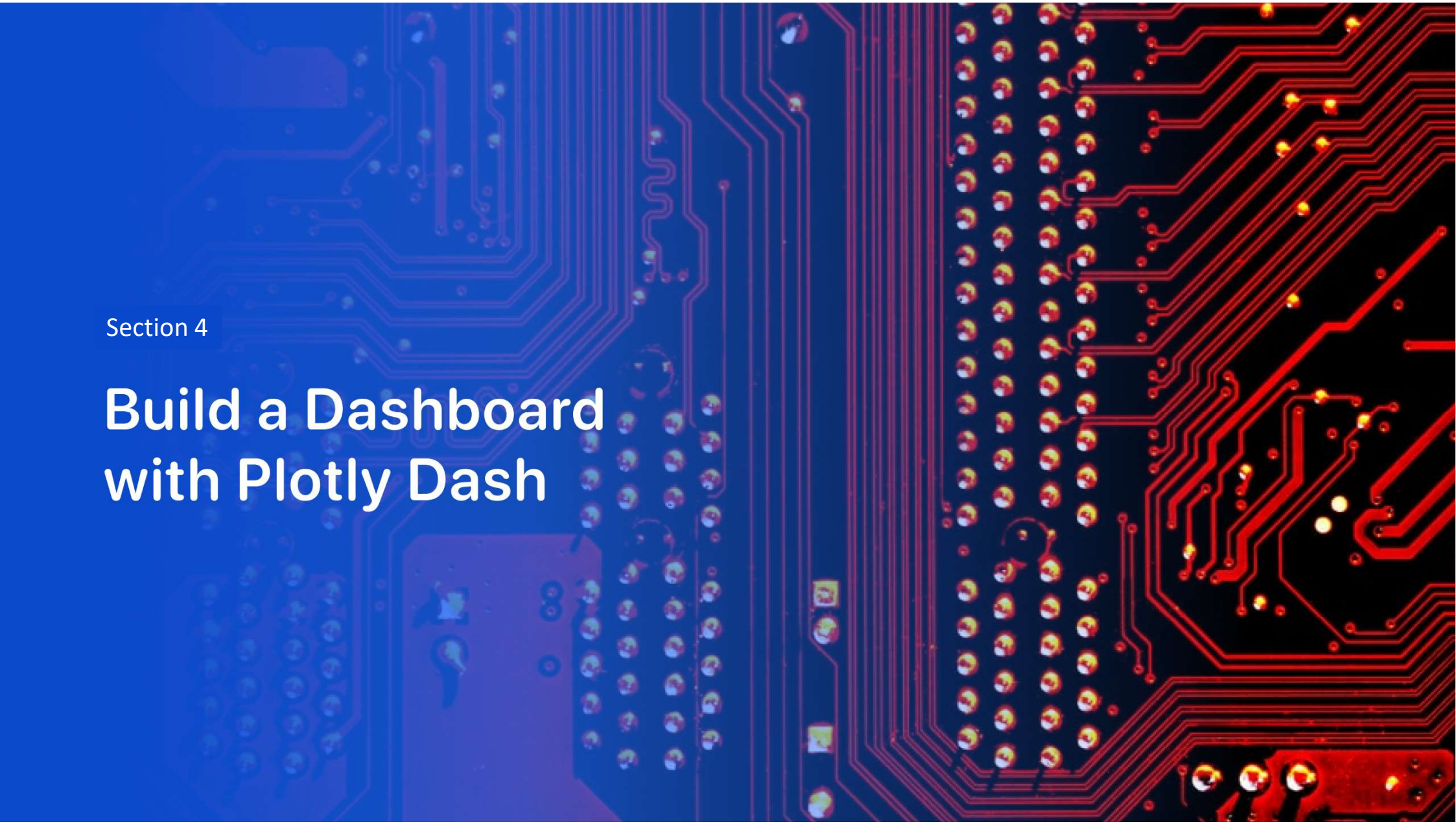
- Passed launched are colored in green and failed launches are colored in red



Surroundings of launch sites

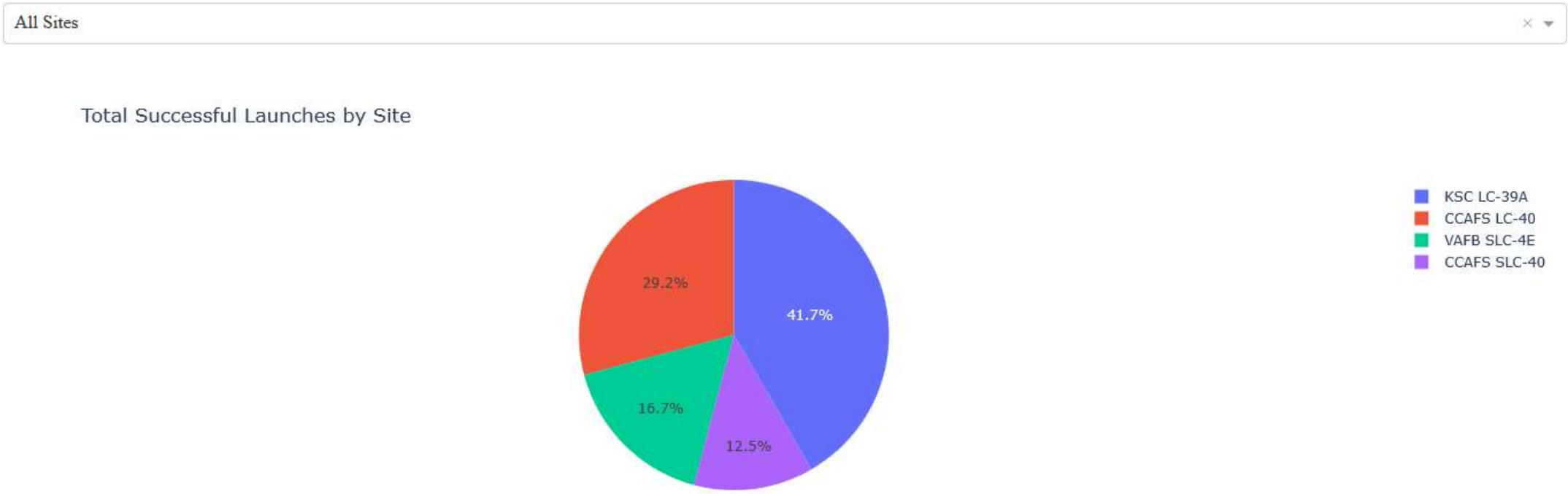
- The launch sites are at least 15km away from cities
- Railways are in close proximity because of transportation advantages
- Coast line is also close to the launch site because of possible water landing tests

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue


Section 4
Build a Dashboard with Plotly Dash

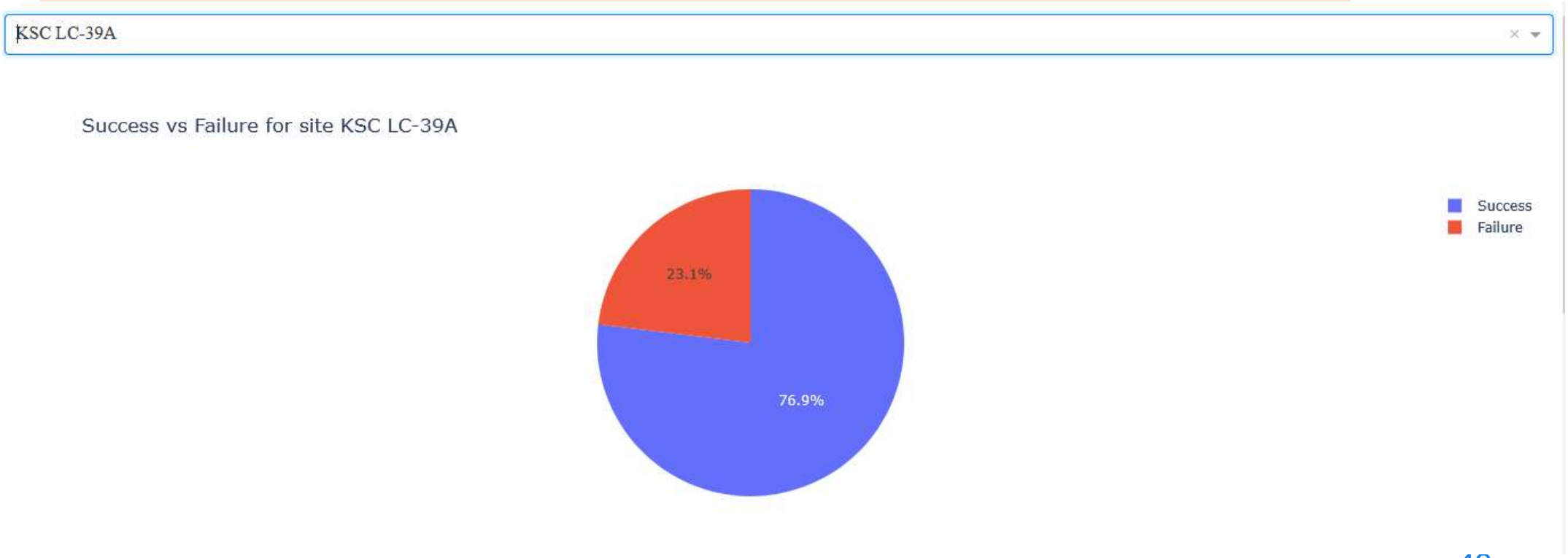
Dashboard Analysis: Locations-1

- Q: Which location has the most successfully launches (with Falcon9)?
- A: KennedySpaceCenter Launch Complex 39A (KSC LC-39A) with 10 successfully launches LINK



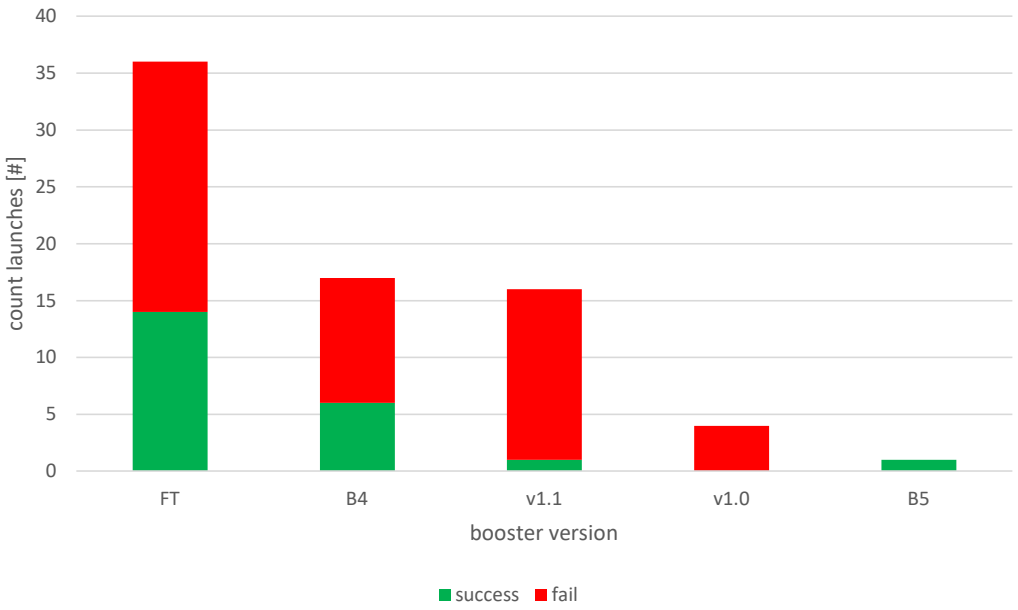
Dashboard Analysis: Locations-2

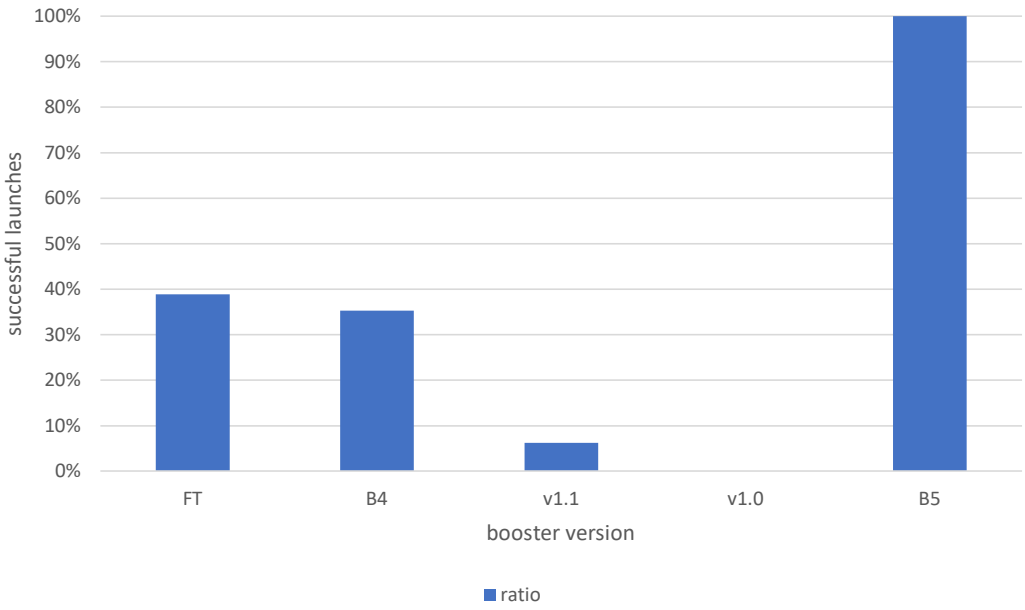
- Q: Which location has the highest launch-success-rate?
- A: The highest success-rate has KSC LC-39A with ~ 77%



Dashboard Analysis: Booster Impact-1

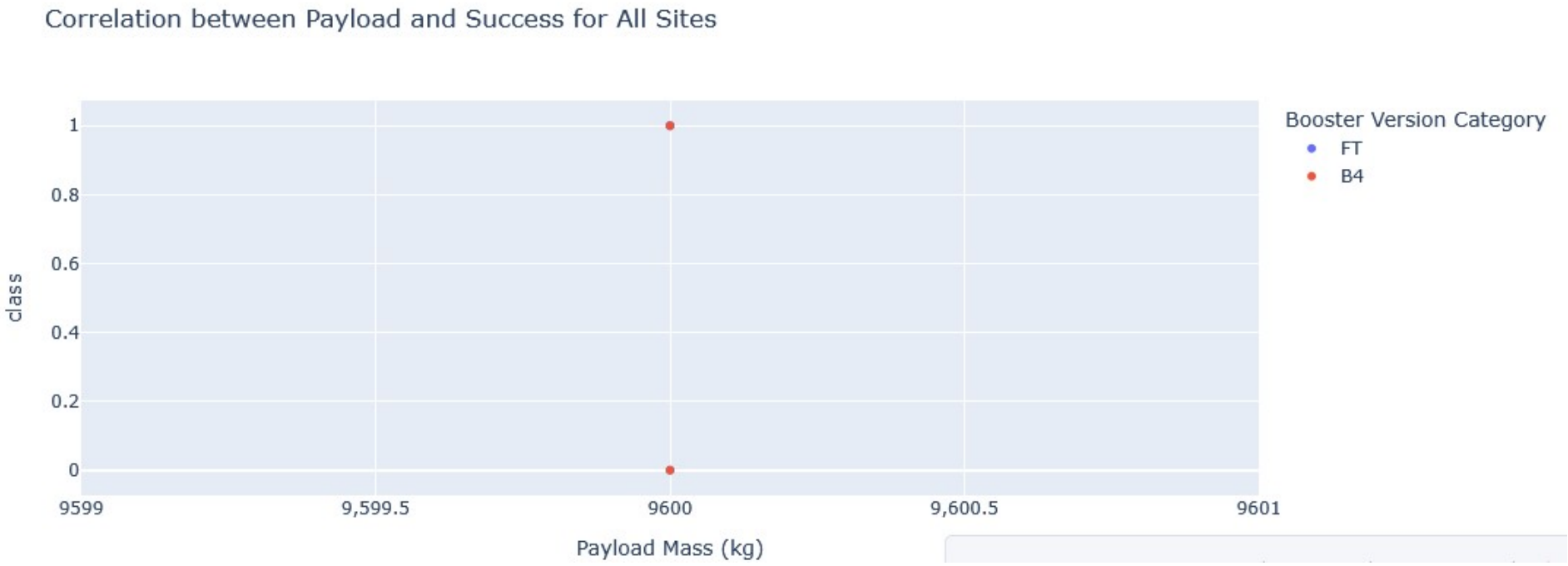
- Outcome:
  - Booster Version FT has the most tested launches and the highest success-rate (here we ignore version B5 because it was tested only once)





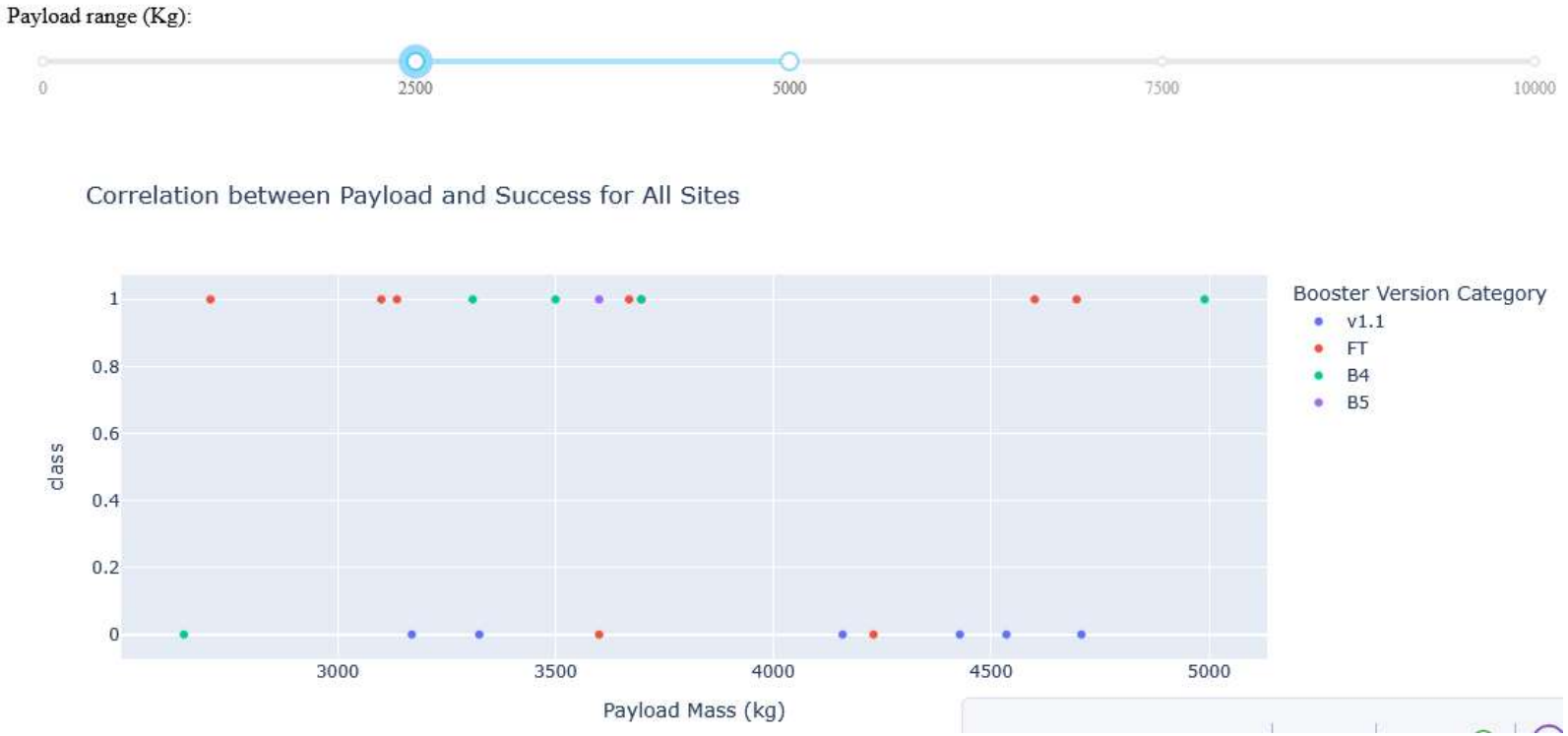
Dashboard Analysis: Booster Impact-2

- Outcome:
  - The only Booster Version which was tested with high payload between 7.5 and 10t was the version B4, one of two launches were successful



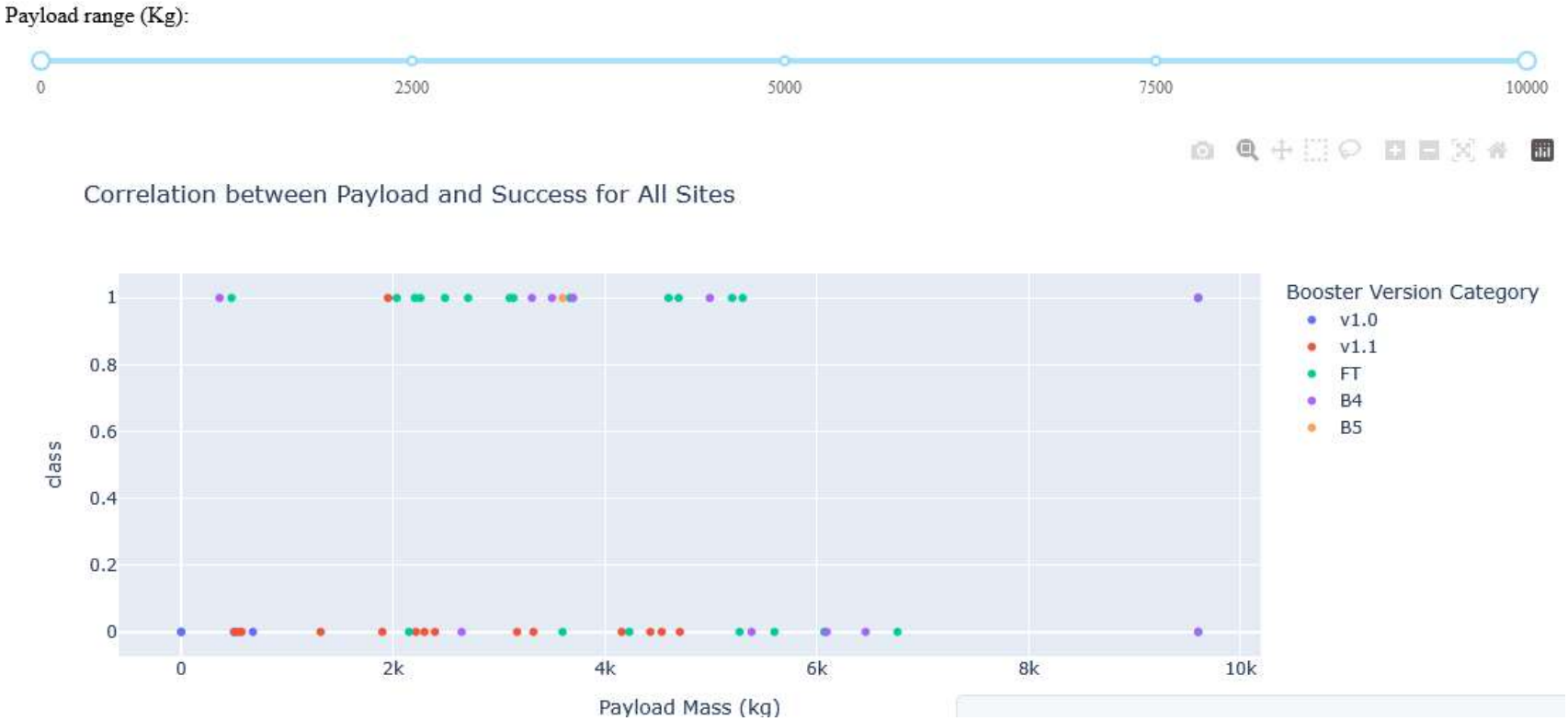
Dashboard Analysis: Booster Impact-3

- Outcome:
  - Payload between 2.5 and 5t shows the highest success-rate of 55% (11/20 launches)



Dashboard Analysis: Booster Impact-4

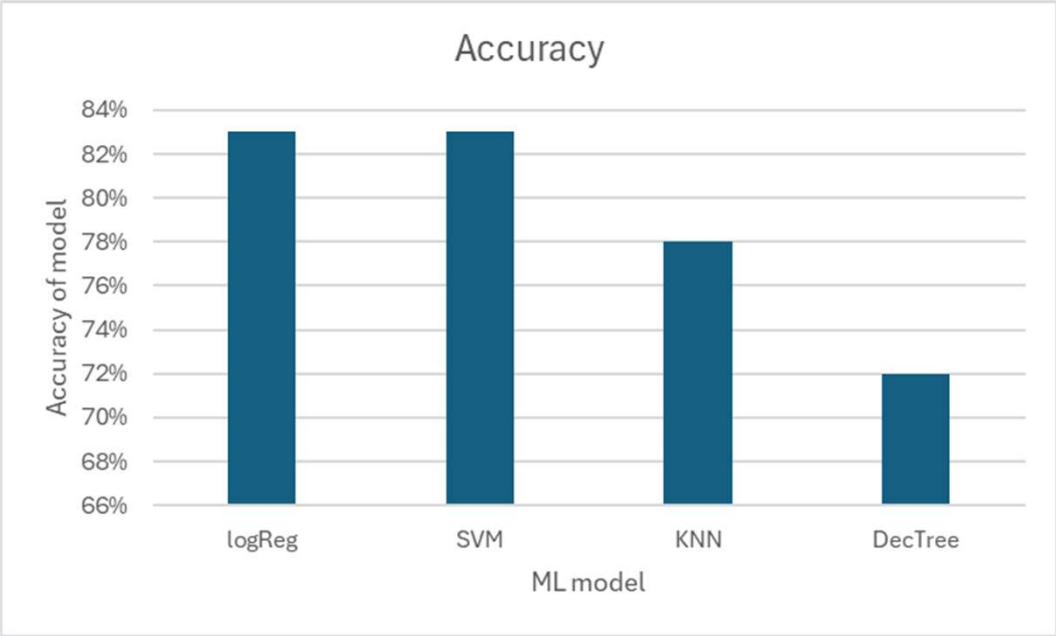
- Total overview



Section 5

Predictive Analysis (Classification)

Classification Accuracy



Find the method performs best:

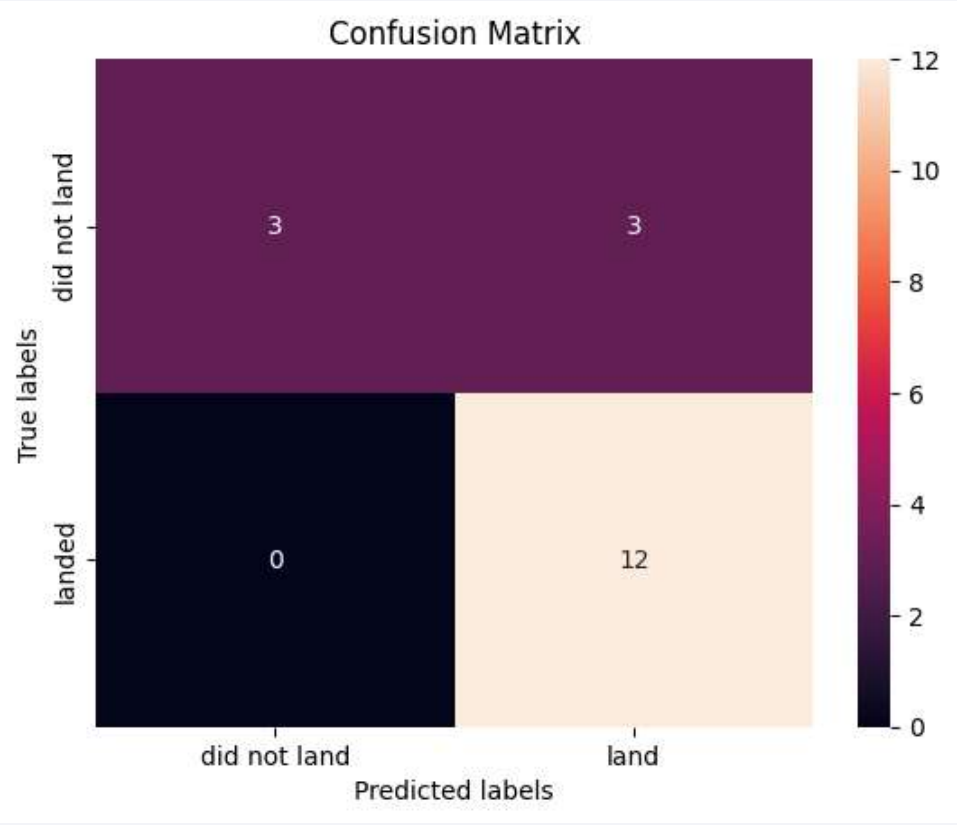
```
accuracy_lr = best_lr.score(X_test, y_test)
accuracy_svm = best_svm.score(X_test, y_test)
accuracy_tree = best_tree.score(X_test, y_test)
accuracy_knn = best_knn.score(X_test, y_test)

import pandas as pd

results = {
    'Model': ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN'],
    'Test Accuracy': [accuracy_lr, accuracy_svm, accuracy_tree, accuracy_knn]
}

results_df = pd.DataFrame(results)
results_df = results_df.sort_values(by='Test Accuracy', ascending=False)
print(results_df)
```

	Model	Test Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
3	KNN	0.777778
2	Decision Tree	0.722222

- Logistic regression and SVM models have the same accuracy 83.3%

Confusion Matrix

- Confusion matrix of logistic regression model



Conclusions

- Exploratory data analysis has shown that
  - the successful landing outcomes strongly increased since 2015
  - CCAFS SLC40 has the greatest number of landings
  - Highest success rate have ES-L1, GEO, HEO and SSO orbits
  - All launch sites are close the coast and couple thousands kilometer away from the equator line
  - Railways are in close proximity because of transportation advantages
  - Coast line is close to the launch site because of possible water landing tests
- Predictive analysis results
  - The ML model shows a prediction probability of 83.3%

Recommendations for SpaceY to become a „real“ competitor are shown above

Thank you!