IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrej
20.08.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting data using web scraping and API
  - Data wrangling
  - Data visualization
  - Iteractive visual analytics
  - Machine learning prediction
- Summary of all results
  - Identification of features that are best to predict succes of laucnhes
  - All models performed similary on the test set

Introduction

- The goal was to predict if the first stage will land successfully
  - What factors determine if the rocket will land succesfully
  - The iteraction amonst various features that determine the success rate of a successful landing
  - What operation conditions needs to be in place to ensure a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from Space X API and web scraping Wikipedia.
- Perform data wrangling
  - Creating 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
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected using
  - Get request to the Space X API and
  - Web scraping from Wikipedia

Data Collection – SpaceX API

- We used SpaceX public API to collect the data
- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/01_Data_Collection_API.ipynb

Now let's start requesting rocket launch data from SpaceX API with the following URL:

```
[8]: spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
[9]: response = requests.get(spacex_url)
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

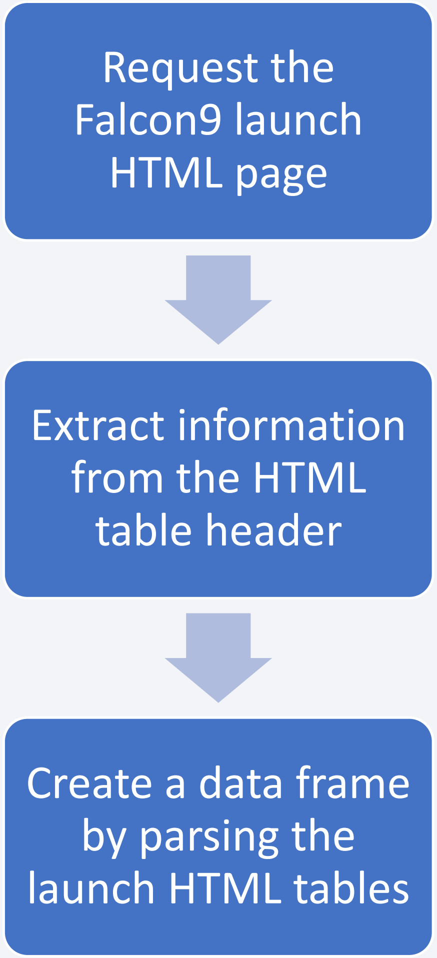
Using the dataframe `data` print the first 5 rows

```
# Get the head of the dataframe
data.head()
```

Data Collection - Scraping

- Data was also collected using web scraping from Wikipedia

- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/02_Data_Collection_Web_Scraping.ipynb



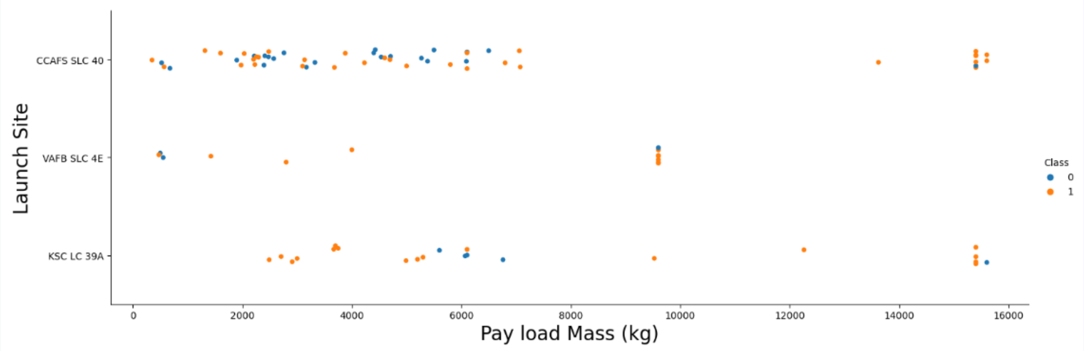
Data Wrangling

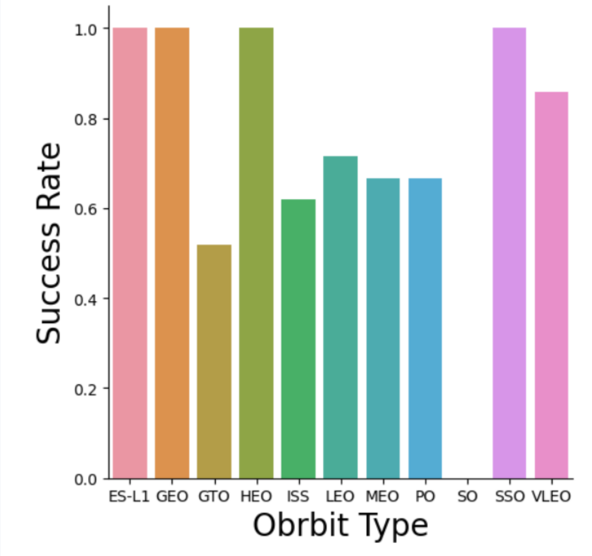
- We performed some Exploratory Data Analysis to find patterns in the data and determine what would be the label for training supervised models

- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/03_Data%20Wrangling.ipynb

EDA with Data Visualization

- Seaborn scatterplots and bar charts were used for the visualization





- Link:https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/05_EDA_with_Visualization_Lab.ipynb

EDA with SQL

- SQL Tasks performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/04_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines were created and added to a folium map
    - Markers show launch sites on a map
    - Circles add a highlighted circle area with text label on a specific condition
    - With lines we showed distances betweeen two points.

- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/06_Interactive_Visual_Analytics_with_Folium_Lab.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- Four classification models were compared
  - Logistic regression,
  - SVM
  - Classification trees
  - K nearest neighbort

- Link: https://github.com/Betijan/Applied_Data_Science_Capstone/blob/main/07_Machine_Learning.ipynb

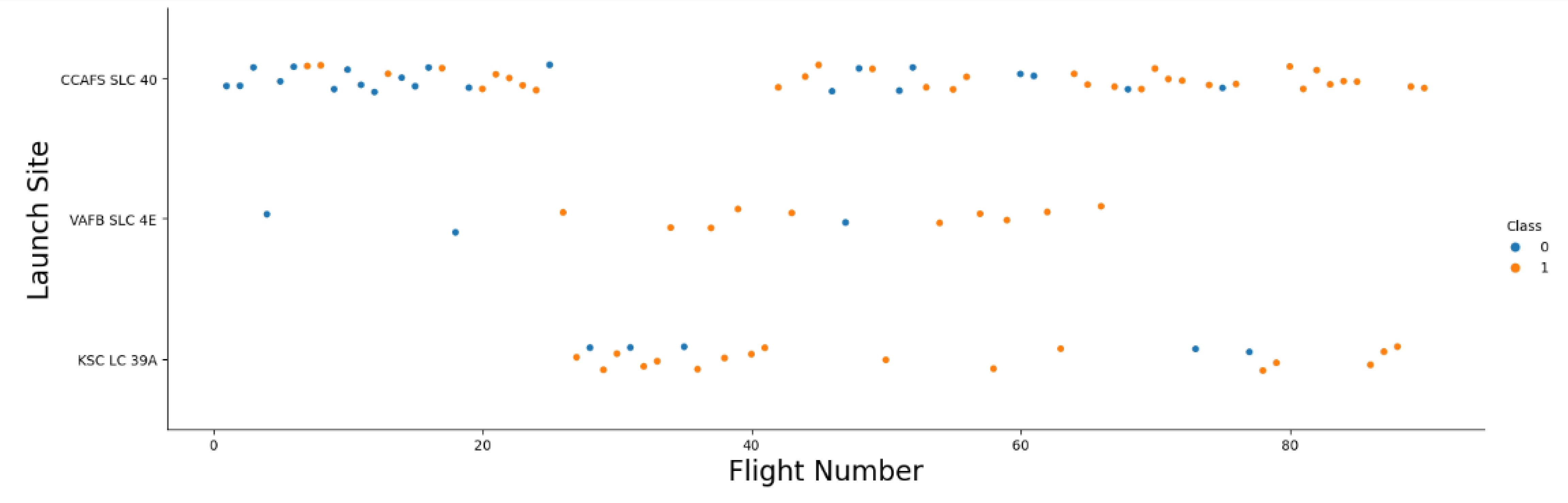
Results

- Exploratory data analysis results
  - SpaceX uses four different launch sites
  - The first success landing outcome happened in 2015, five years after the first launch
  - Almost 100% of mission outcomes were successful

Section 2

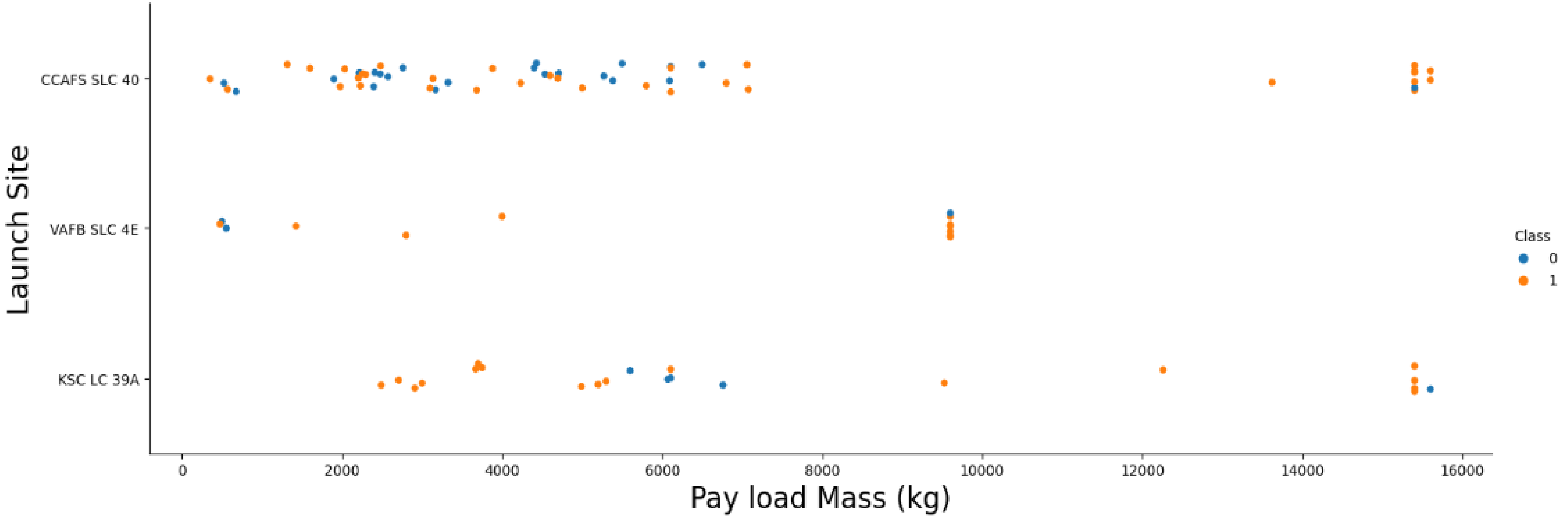
Insights drawn from EDA

Flight Number vs. Launch Site



- the higher the Flight Number the higher the change of a Class 1
- Most of the launched performend on the CCAFS SLC 40 site

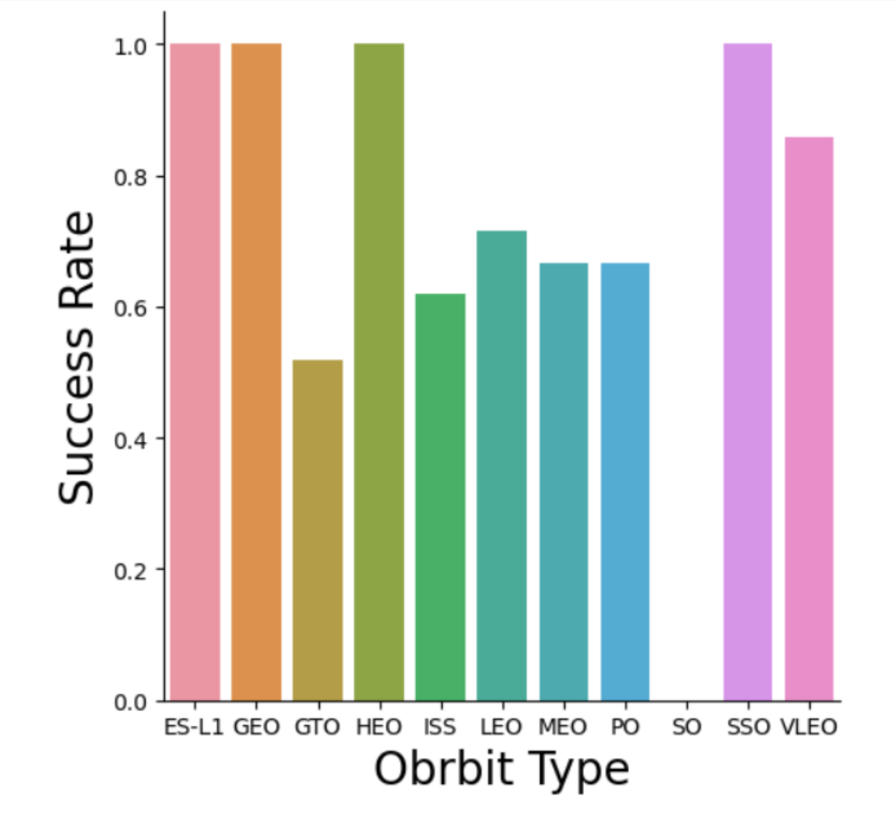
Payload vs. Launch Site



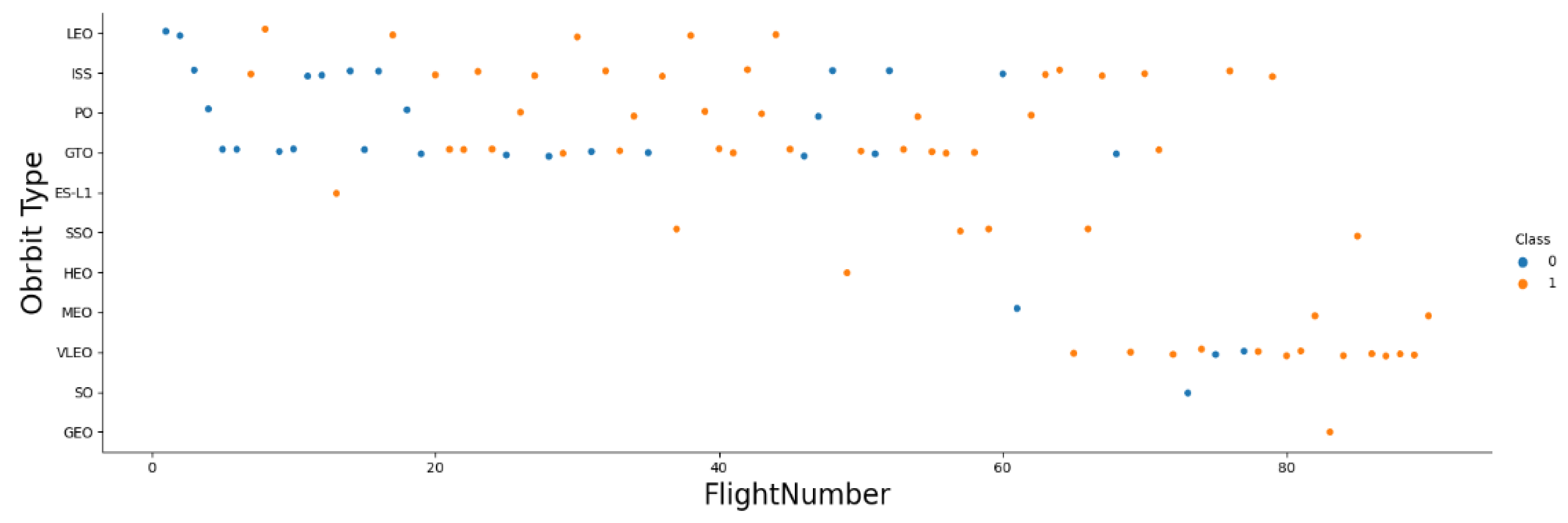
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Biggest success rates for orbit types:
  - ES-L1
  - GEO
  - HEO
  - SSO

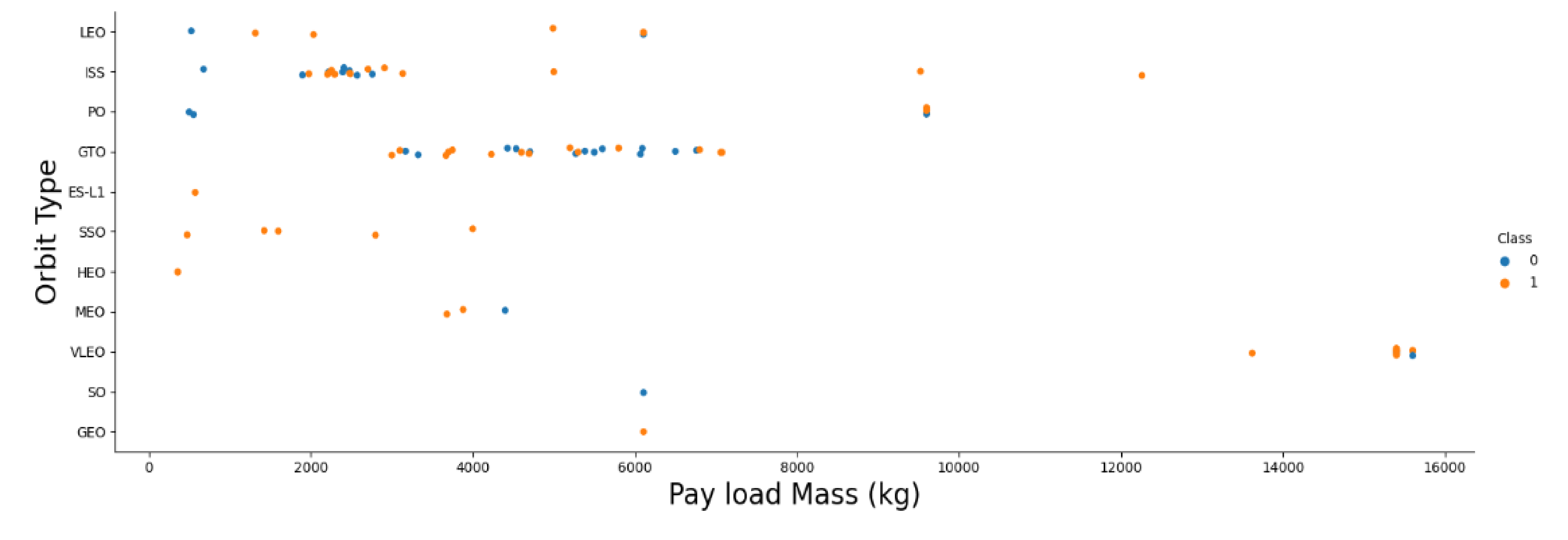


Flight Number vs. Orbit Type



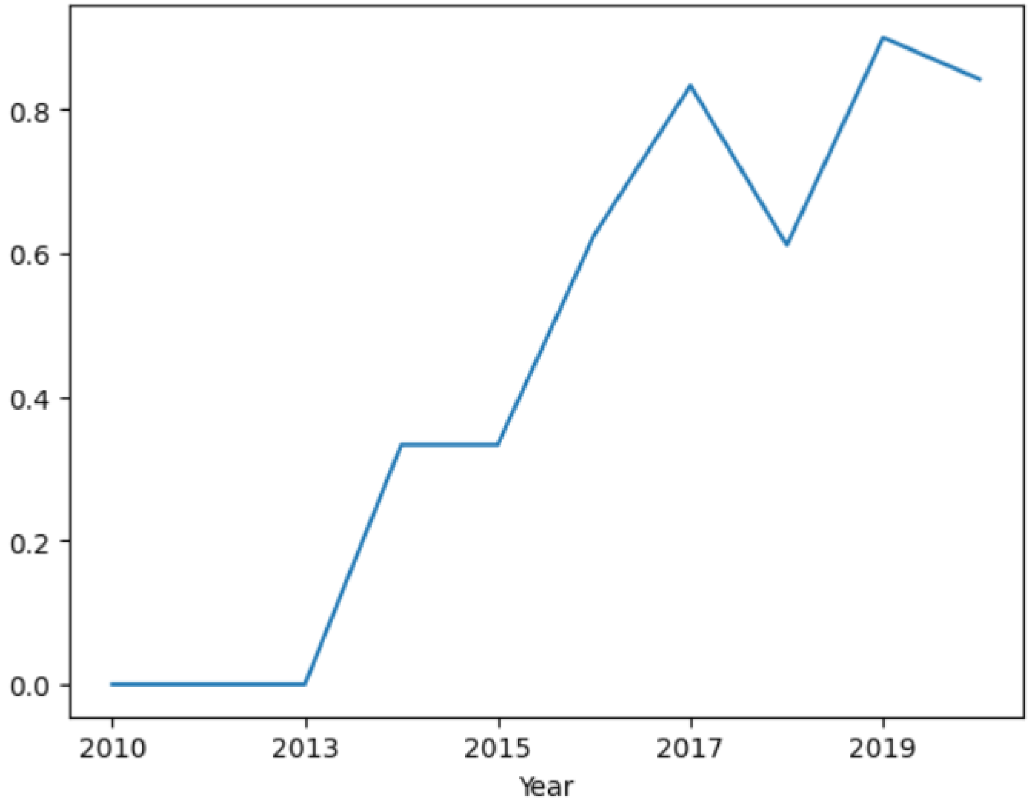
- Success rate improved over time

Payload vs. Orbit Type



There is no relation between payload and success rate to orbit GTO

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Unique launch sites:

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

- Query used: %sql select distinct launch_site from SPACEXTABLE
- By extracting unique (distinct) values for the column launch_site from the table SPACEXTABLE we get the list of unique launch sites

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster Version	Launch Site	Payload	KG	Orbit	Customer	Mission Outcome	Landing Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

We filtered the column launch_site by looking for values starting with CCA and limit the results to five.

Total Payload Mass

- Query used:
  - %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
- We summarize the "Payload Mass KG" column for the customer "Nasa (CRS)".

Total Payload in KG
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

Average Payload in KG
2534.6666666666665

- Query used:
  - %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%‘

- Fitering data by the booster version and calculating the average payload mass.

First Successful Ground Landing Date

- First successful landing outcome on ground pad

Date
2015-12-22

- We filtered the data by successful landing outcome on ground pad and getting the earliest date.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of 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

- We select unique booster versions which successfully landed on drone ship with a payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

- We have grouped records by mission outcome and then count records for each of them.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:

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

Booster_Version
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- SQL:

```
%%sql
select    booster_version
from      SPACEXTABLE
WHERE     Payload_mass__kg_ = (select max(Payload_mass__kg_) from   SPACEXTABLE)
```

2015 Launch Records

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

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
October	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- We filtered the query based on year 2015 and landing outome that resulted in “Failure (drone ship)”

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

- Landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

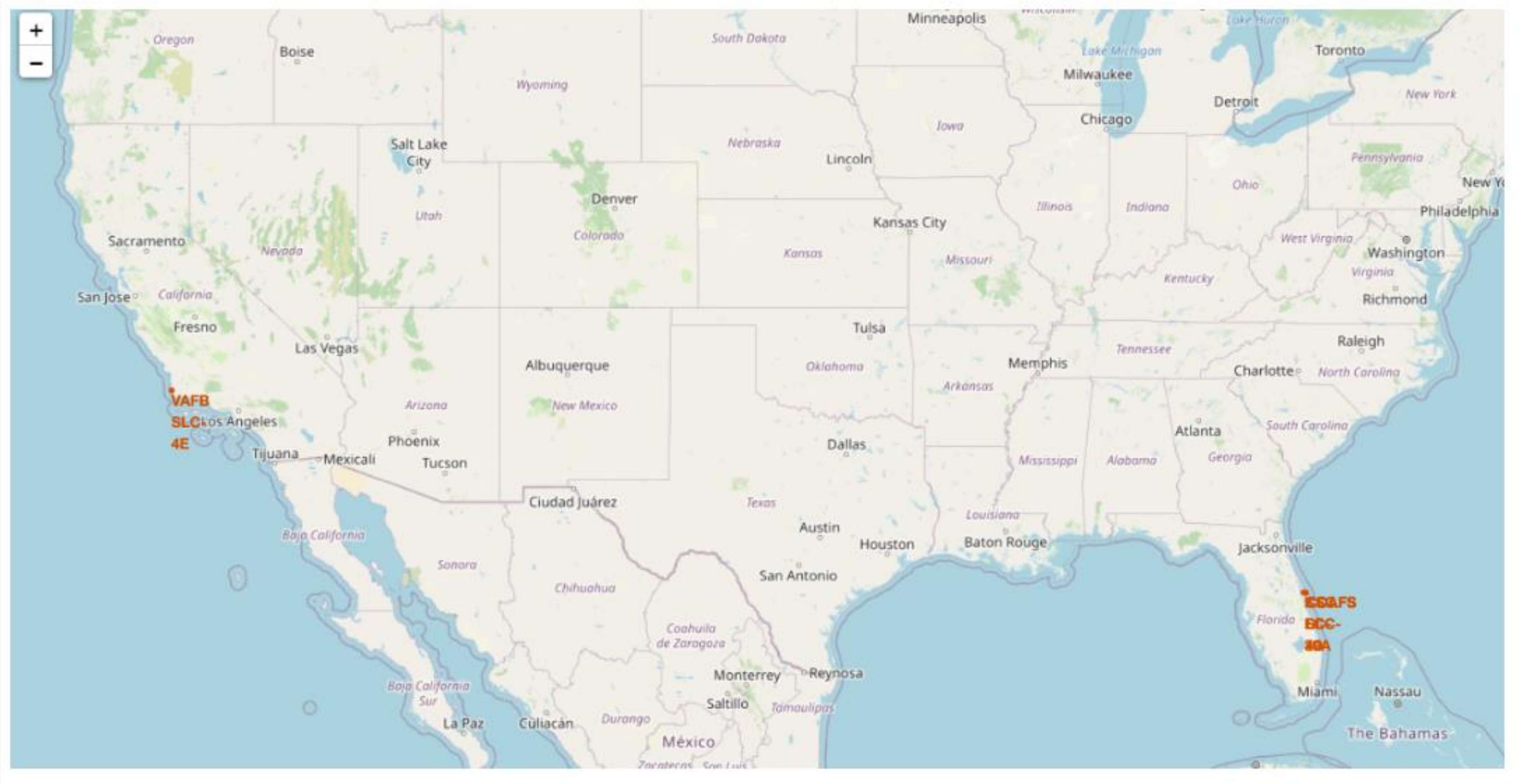
Landing Outcome	Total
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- We should take a better look at the «No attempt».

Section 3

Launch Sites Proximities Analysis

Launch Sites



- Launch sites are close to the coast, to provide safety in case of failure.

Launch Site CCAFS SLC-40

- Green marked represent successful and red marked failed launch outcomes

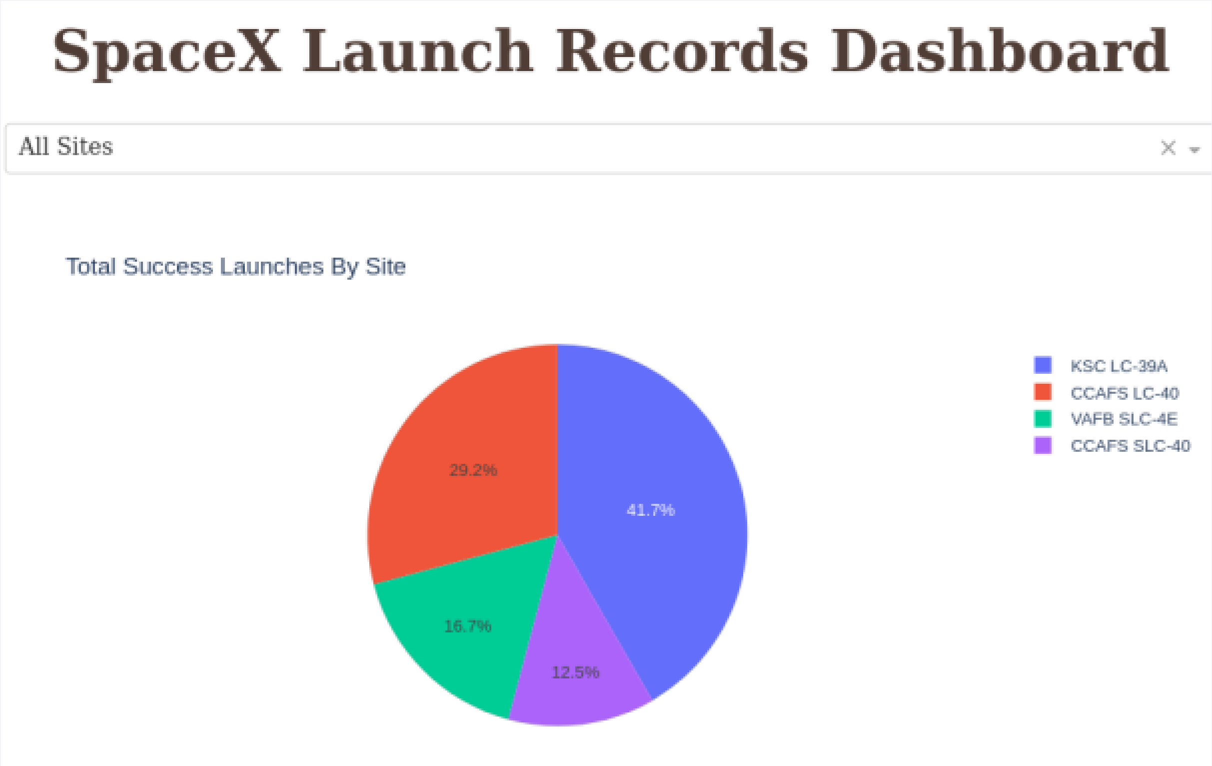
Distance to coast

- The launch site has good distance to the coast.

Section 4

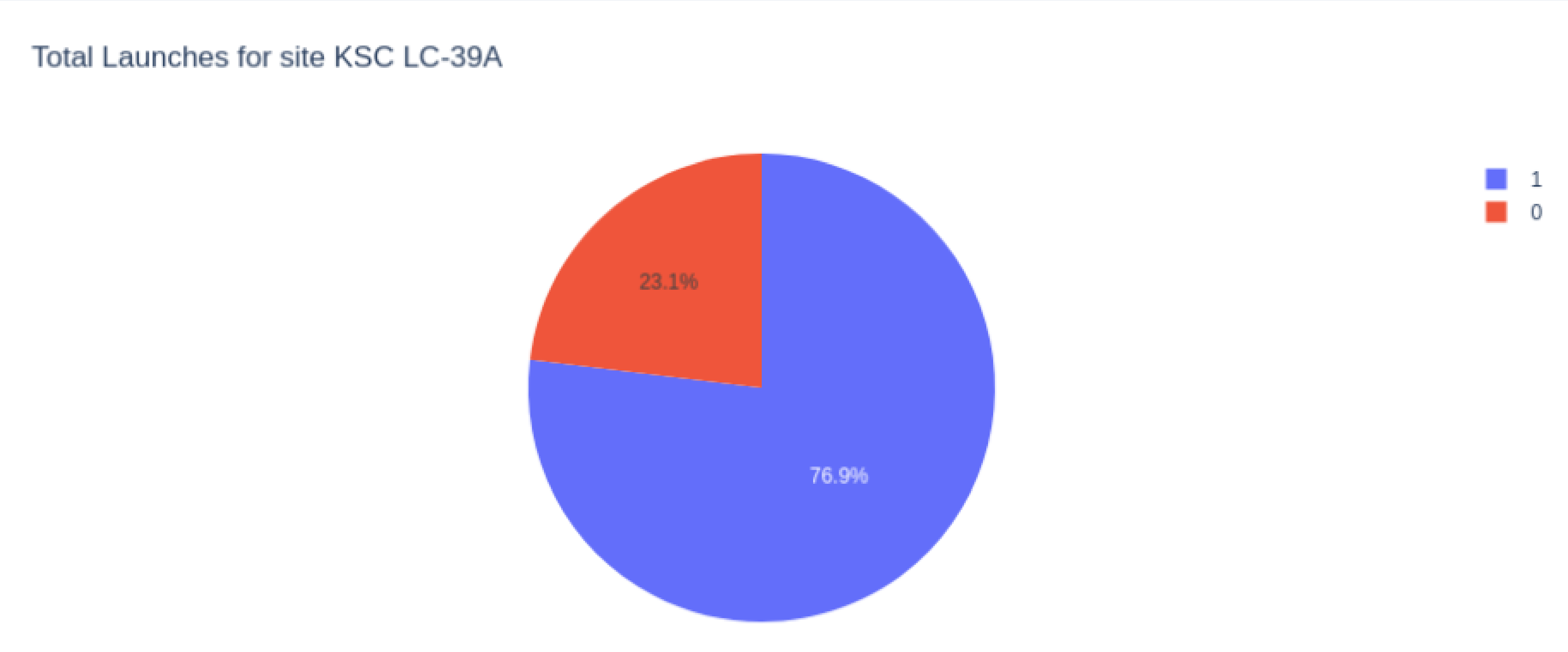
Build a Dashboard with Plotly Dash

Succesful Launches by Site



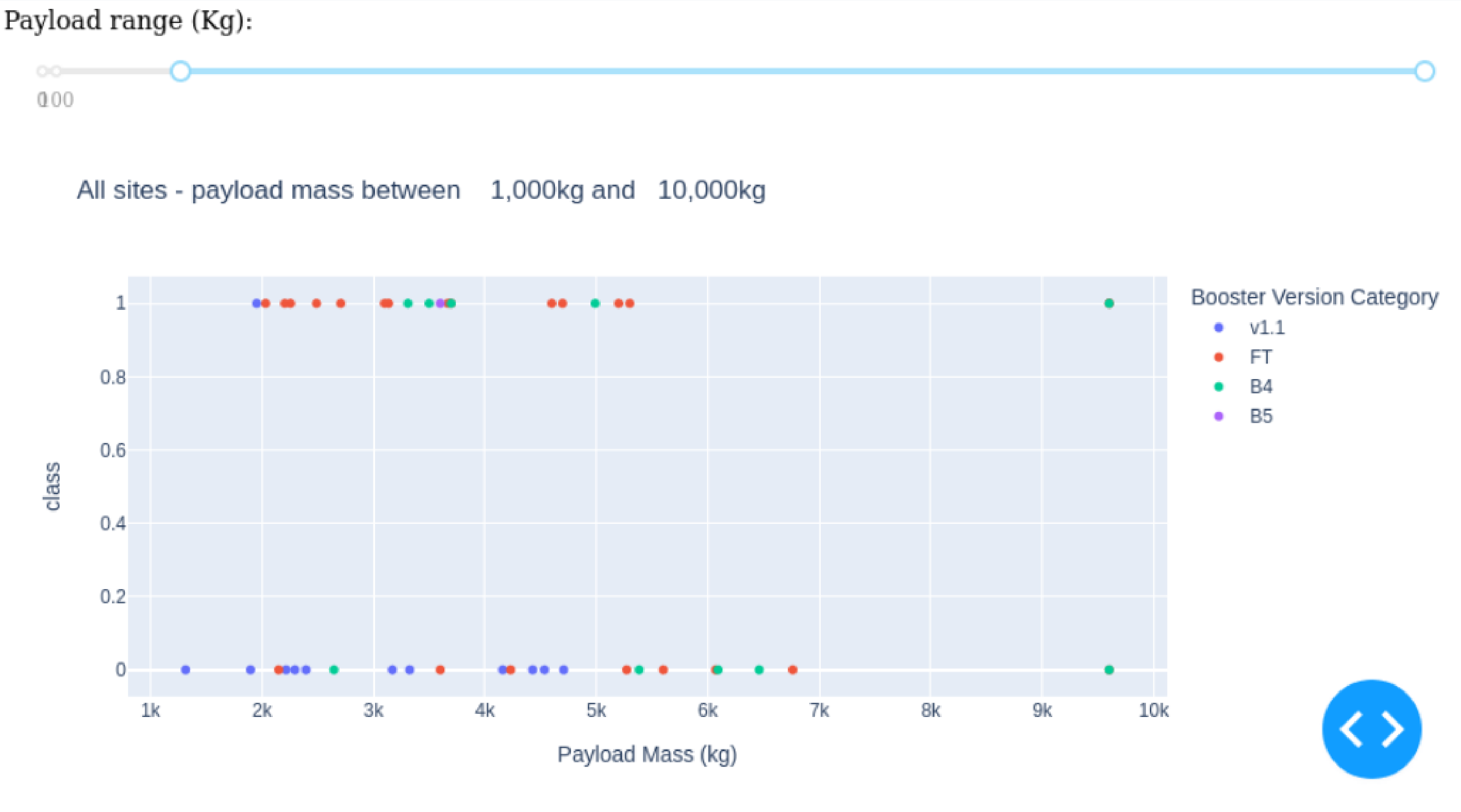
- Launc Site has an important rule on success of the mission.

Launch Success Ratio for KSC LC-39A



- 76.9% of launches were successful on this site

Payload vs Launch Outcome



Section 5

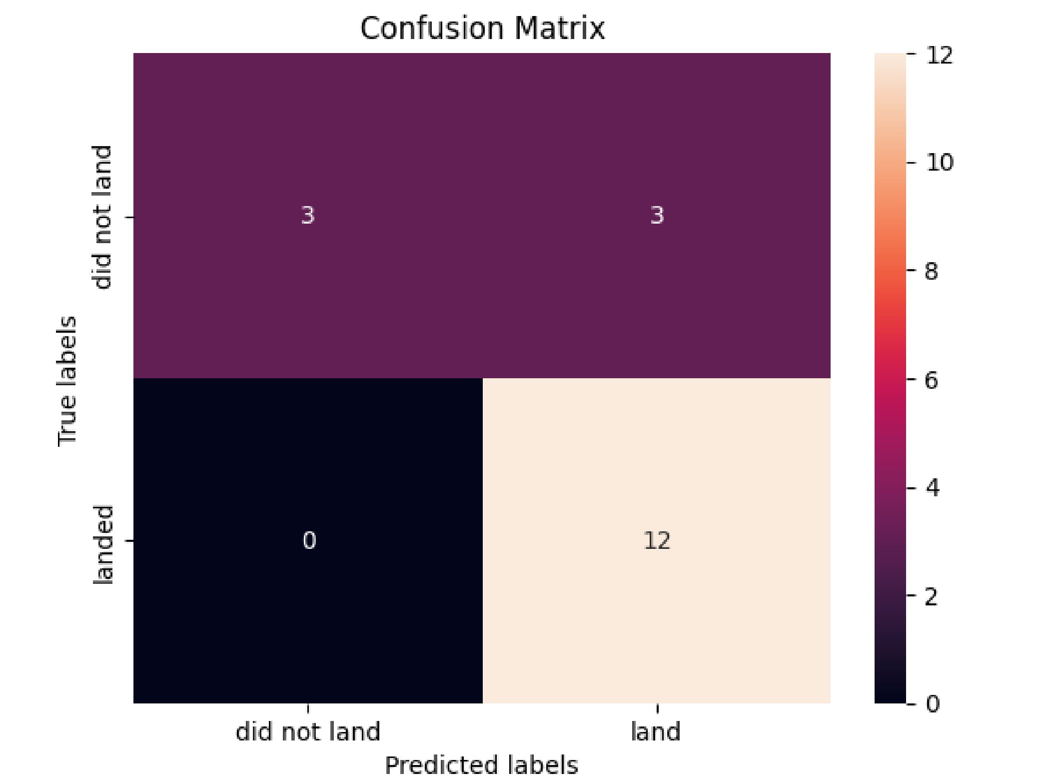
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested

- The model with the highest classification accuracy is decision tree classifier, which has accuracy of over 85%

Confusion Matrix



- Examing the confusion matrix, we see that logistic regression can distinquis between the different classes. We se that hte major problem is false positives.

Conclusions

- The best launch site was KSC LC-39A
- Launches above 7000kg are less risky
- Successful landing outcomes improve over time
- Decision tree classifier can be used to predict successful landings and increase profits

Appendix

Thank you!