IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sofya Guseva
04.04.2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies

We performed a data analysis on the data from SpaceX. We collected the data using REST API, web scrapping. We performed data wrangling: sorting and dealing with missing values in the dataset. We applied exploratory analysis using SQL and visualization using Matplotlib. The result was presented in the form of scatter plots and bar charts. To make interactive visual analytics we used dashboards created with Plotly library. Finally, we applied classification models to our data and determined the best with the highest accuracy.

- Summary of all results

We used the methods above to obtain the insights about some trends and factors determining the success of the landing of the rocket first stage. We found that it depends on such property as payload mass, the amount of launches. Also, the mission outcome success increases over time. In addition it can be predicted with 89% accuracy using the tree classification model.

Introduction

- We entered the commercial space age where several companies can make space travel affordable for everyone
- Among such companies are:

- Virgin Galactic
- Rocket Lab
- Blue Origin
- SpaceX

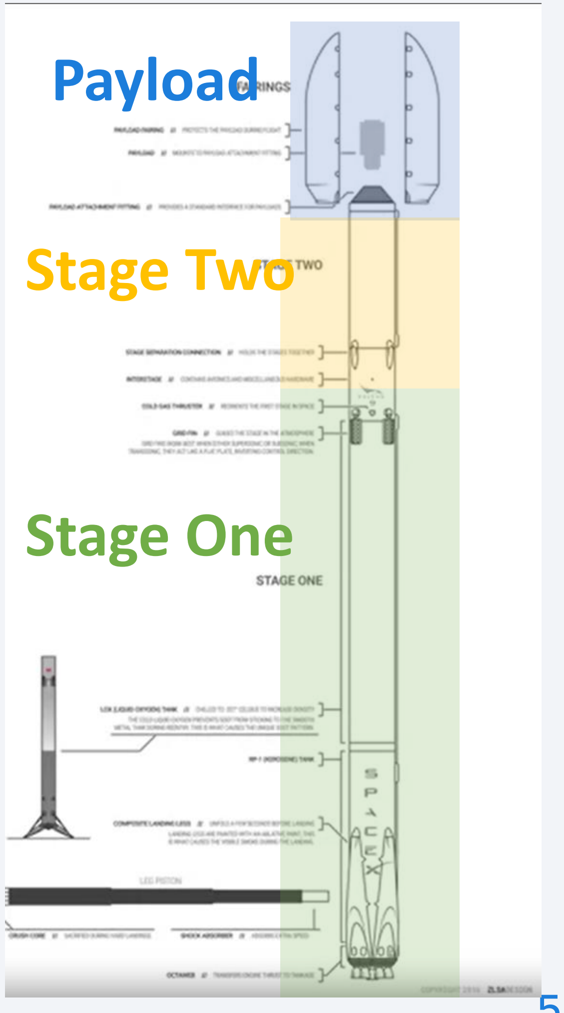
- Perhaps, the most successful company is SpaceX which provides:

- sending spacecraft to the International Space Station
- satellite Internet access (Starlink) using a satellite internet constellation
- sending manned missions to Space

Introduction

- Success of SpaceX can be explained by the fact, that it can launch the rockets at a relatively low cost: e.g., SpaceX advertises Falcon 9 rocket launches with a cost of 62 million $ and other providers cost upwards of 165 million $ each.
- Much of the savings is because SpaceX can reuse the first stage.

SPACE Y - Our company



Questions:

How can we compete with SpaceX?

What are the factors determining the successful landing of the first stage?

Can we predict which landing will be successful? This will influence the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Requesting to the SpaceX API
  - Web scrapping launch records from a Wikipedia
- Perform data wrangling
  - Cleaning the data
  - Preparing the data in a dataframe format
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Determining training labels
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluation of classification models

Data Collection

Obtaining data using SpaceX REST API

- Use URL to target a specific endpoint of the API to get past launch data
- Get request
- Response in a form of a list of JSON objects
- Use of normalize function to convert the data to a Pandas dataframe

Obtaining data using web scrapping Falcon 9 Launch records

- Web scrapping with BeautifulSoup object
- Parse the data from the tables
- Convert into a Pandas dataframe

Data Collection – SpaceX API

- Use URL to target a specific endpoint of the API to get past launch data

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

- Get request

```
response = requests.get(spacex_url)
```

- Response in a form of a list of JSON objects

```
response.json()
```

- Use of normalize function to convert the data to a Pandas dataframe

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

URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

Data Collection - Scraping

- Use static URL
- Request the HTML page from the URL and get response object
- Create BeautifulSoup object from the HTML response
- Parse the table from the object
- Parse columns and data and convert them to a dataframe

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_laun
```

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get("https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon
```

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, "html.parser")
```

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables=soup.find_all('tr')
```

```
column_names = []

th_objects = first_launch_table.find_all('th')
#print(th_objects)
for i,row in enumerate(th_objects):
    print("row",i,"is",row)
    column_name = extract_column_from_header(row)
    if column_name is not None and len(column_name) > 0:
        column_names.append(column_name)
```

```
launch_dict= dict.fromkeys(column_names)
```

```
df=pd.DataFrame(launch_dict)
```

Data Wrangling

- Filtering the data to get information only for Falcon 9 launches

```python
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = df[df['BoosterVersion']!='Falcon 1']
```

- Replace the missing values in the "PayloadMass" column with a mean value of a column

```python
# Calculate the mean value of PayloadMass column
mean_pay = data_falcon9["PayloadMass"].mean()

# Replace the np.nan values with its mean value
data_falcon9.replace(np.nan, mean_pay, inplace=True)
```

- Replace column "Outcome" with 0 or 1 values, corresponding to landing failure and **landing success**

```python
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

```python
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for index, row in df['Outcome'].iteritems():
    #print(index)
    if row in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
landing_class
```

```
[0,
 0,
 0,
 0,
```


URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

EDA with Data Visualization

- We plotted the following charts:

- scatter plots with different colors indicating the successful landing
- bar chart to show success rate of orbits

All of them helped to get insights of correlated and non-correlated variables

URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

EDA with SQL

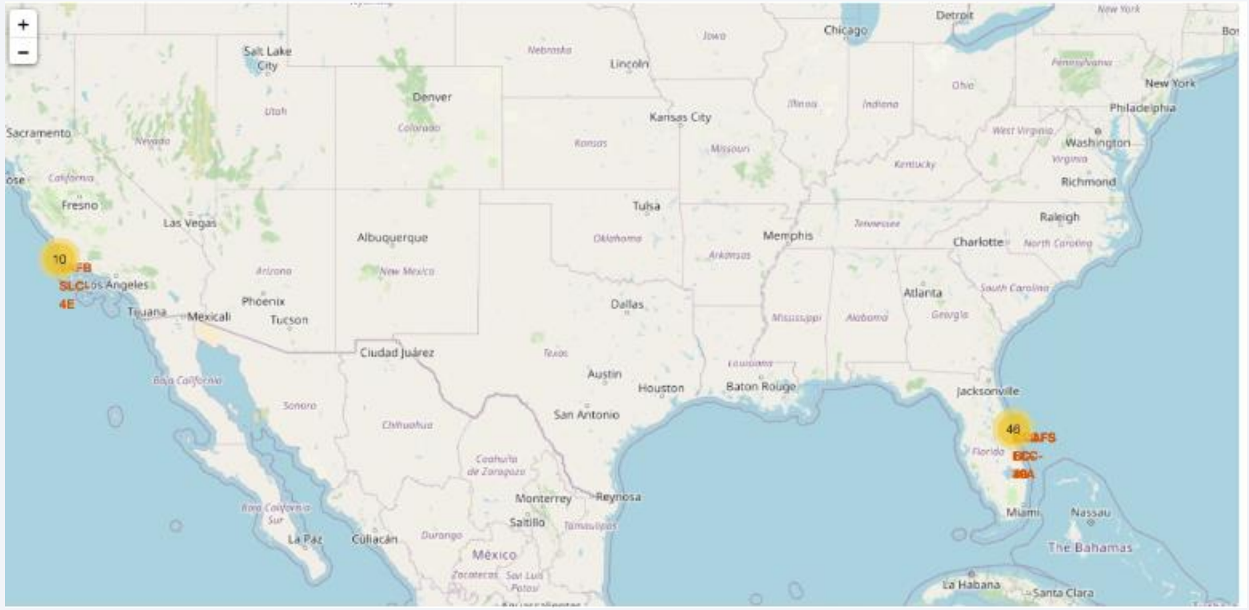
- We performed the following SQL queries:

- SELECT, DISTINCT
- SUM(), AVG(), MAX(), MIN()
- Condition WHERE, LIKE, AND
- Subqueries
- GROUP BY, ORDER BY DESC, COUNT

URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

Build an Interactive Map with Folium

- We added all launch sites on a folium map using such objects as markers, circles
- We marked the success/failed launches for each site on the map using Marker cluster
- We calculated the distances between a launch site to its proximities using lines



URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

Build a Dashboard with Plotly Dash

- To perform interactive visual analytics on SpaceX launch data in real-time we created a dashboard, which included:

- a Launch Site drop-down input with several components: all launch sites and each one individually
- a callback function to render pie chart based on the selected site dropdown; pie chart shows the success rate of all or each launch site
- a range slider to select a payload mass
- a callback function to render the scatter plot for selected payload mass to show how payload may be correlated with mission outcomes for selected sites

URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

Predictive Analysis (Classification)

- We prepared a the column ‘Class’ which indicated successful or non-successful landing. That was our target Y. We prepared parameters X
- We standardize the data and split it to training and test sets
- The models are trained and best hyperparameters are selected using the function GridSearchCV
- We evaluated the models on the test data calculating the score method

Data preprocessing

Splitting dataset to the training and test data

Train model on the training data set

Select best parameters Grid Search

Evaluate the model

URL GitHub: https://github.com/guseva-sofya/Final_Assignment_Applied_Data_Science_Capstone.git

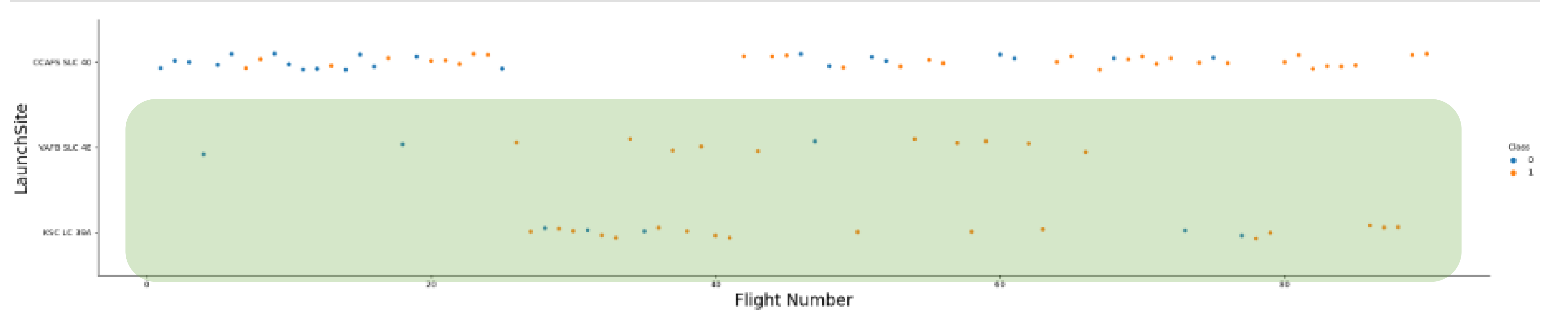
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

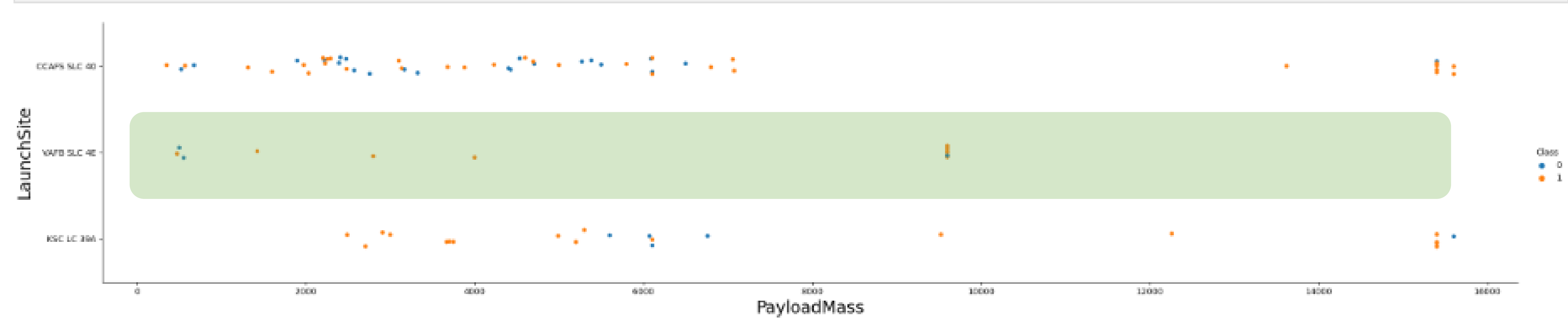
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS LC-40, has a success rate of 60 % while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- This is due to the fact that there are less launches at these places

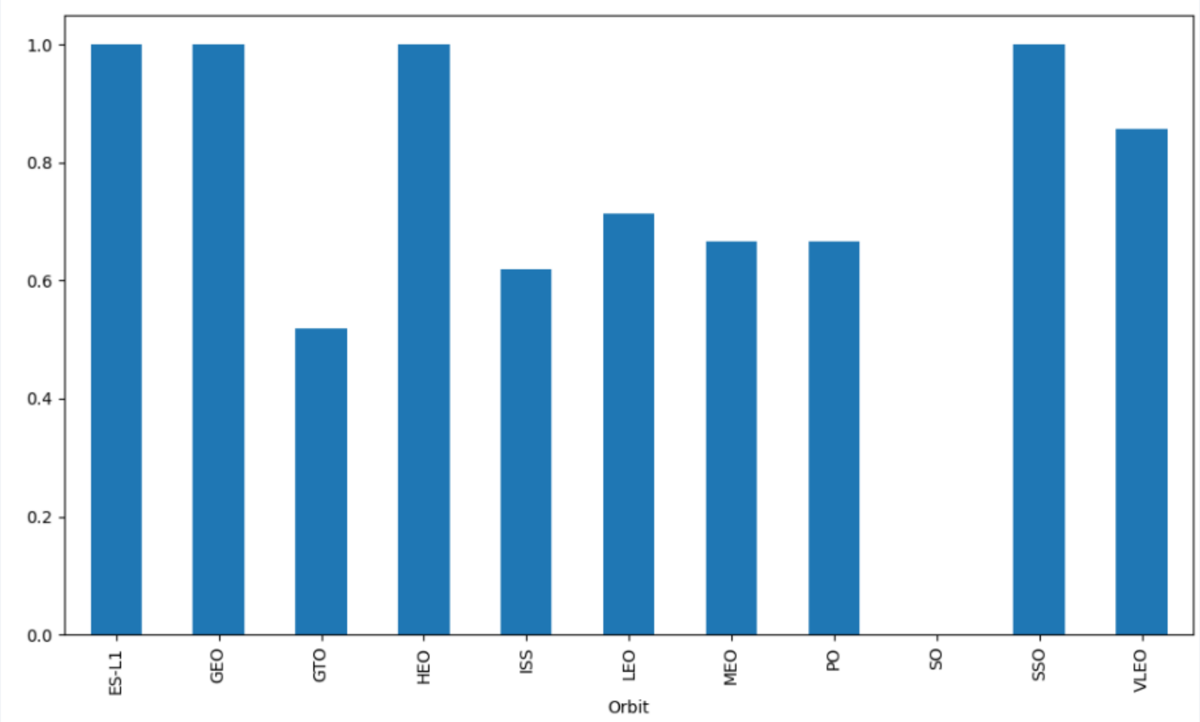
Payload vs. Launch Site



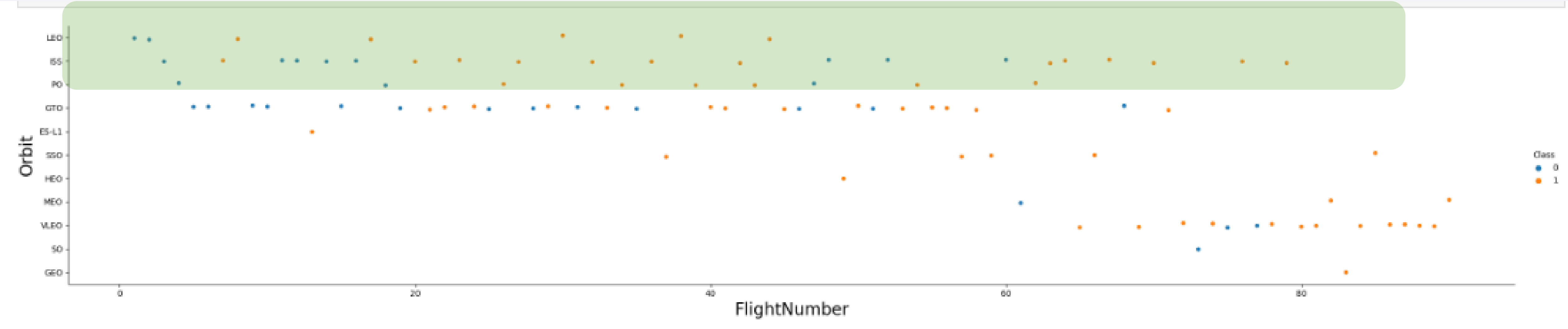
- At VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- The most successful orbits are: ES-L1, GEO, HEO, SSO, VLEO

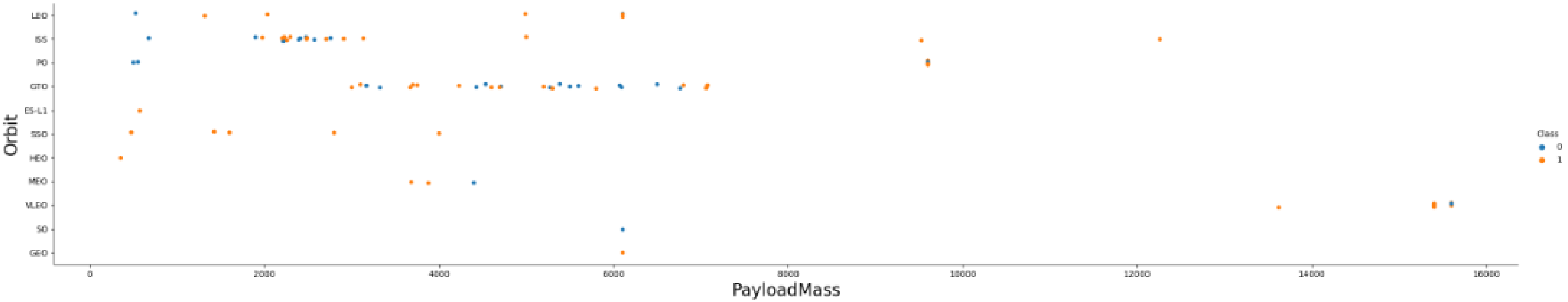


Flight Number vs. Orbit Type



- LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

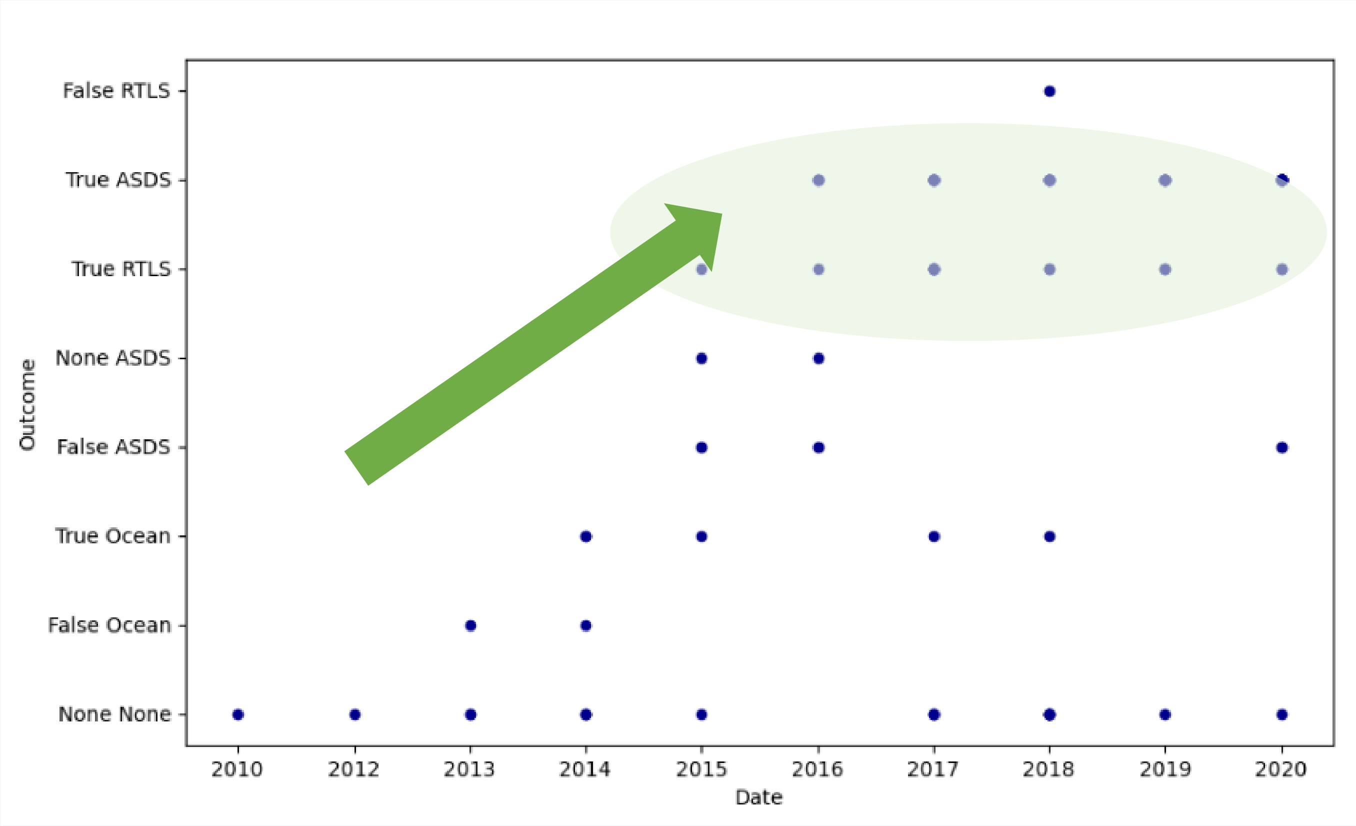
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Sum_payload_mass_kg FROM SPACEXTBL WHERE Customer='NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Sum_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_payload_mass_kg FROM SPACEXTBL WHERE Booster_Version='F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Avg_payload_mass_kg
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN(Date) AS Date_success FROM SPACEXTBL WHERE "Landing _Outcome"='Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

Date_success
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE "Landing _Outcome"='Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) as count FROM SPACEXTBL GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

```
%sql SELECT substr(Date, 4, 2) as months,"Landing _Outcome", Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" LIKE '%drone ship%'
```

```
 * sqlite:///my_data1.db
Done.
```

months	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

- 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

```
%sql SELECT "Landing _Outcome", COUNT(*) as count FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND Date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY COUNT(*) DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

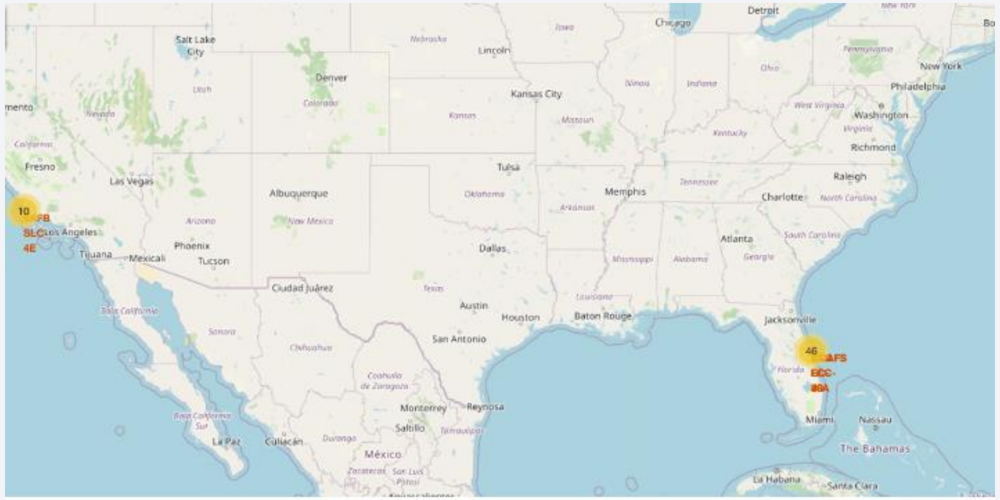
Launch Sites Proximities Analysis

Folium Map Launch Sites

- All launch sites are close to the Equator line
- All launch sites in very close proximity to the coast

Folium Map color-labeled Launch records

KSC LC-39A launch site has relatively high success rate and CCAFS SLC-40 relatively low success rate

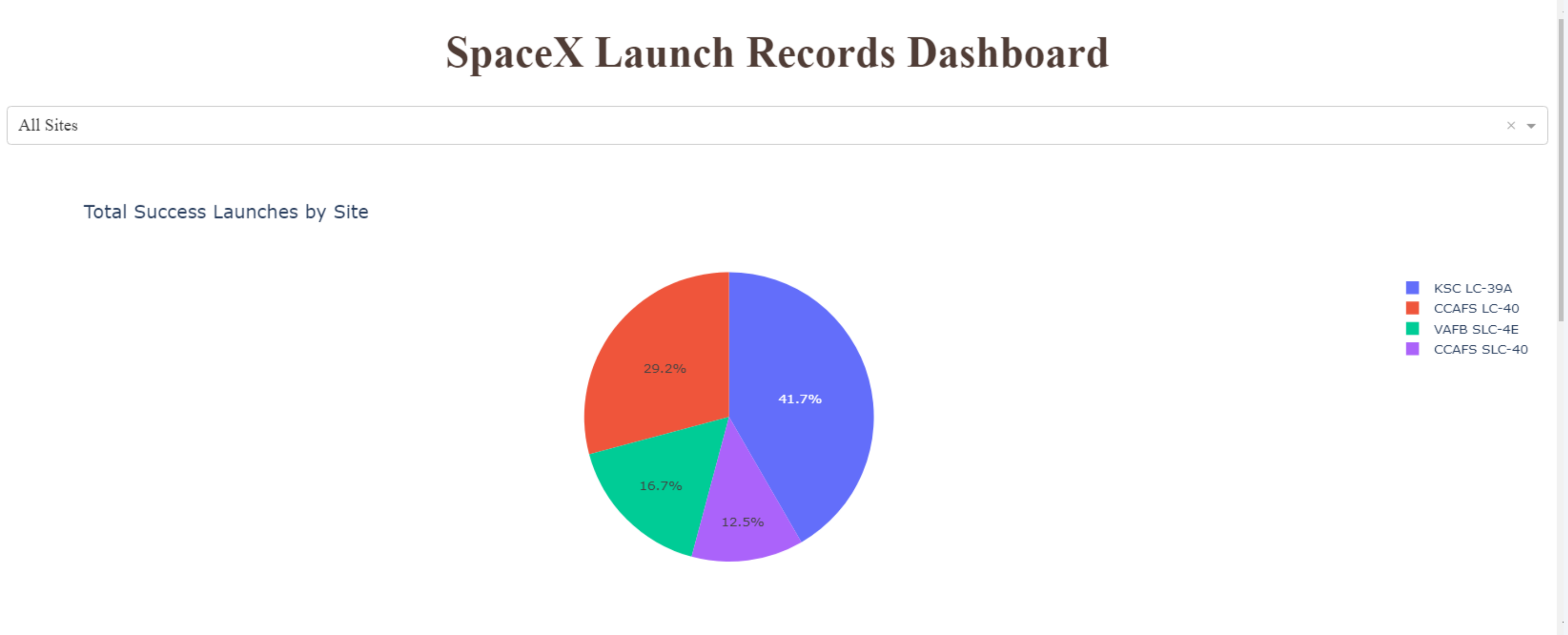


Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

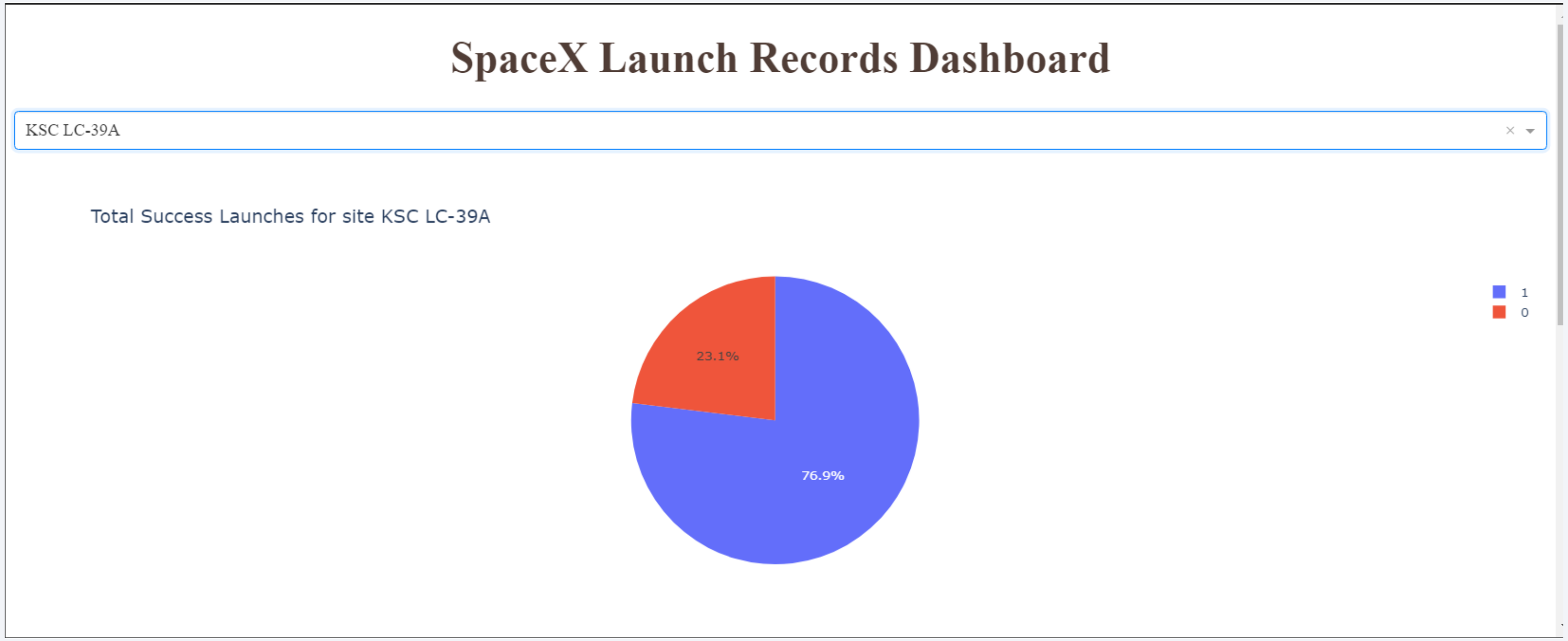
Total Success Launches by Site



- The most successful launch site is KSC LC-39A

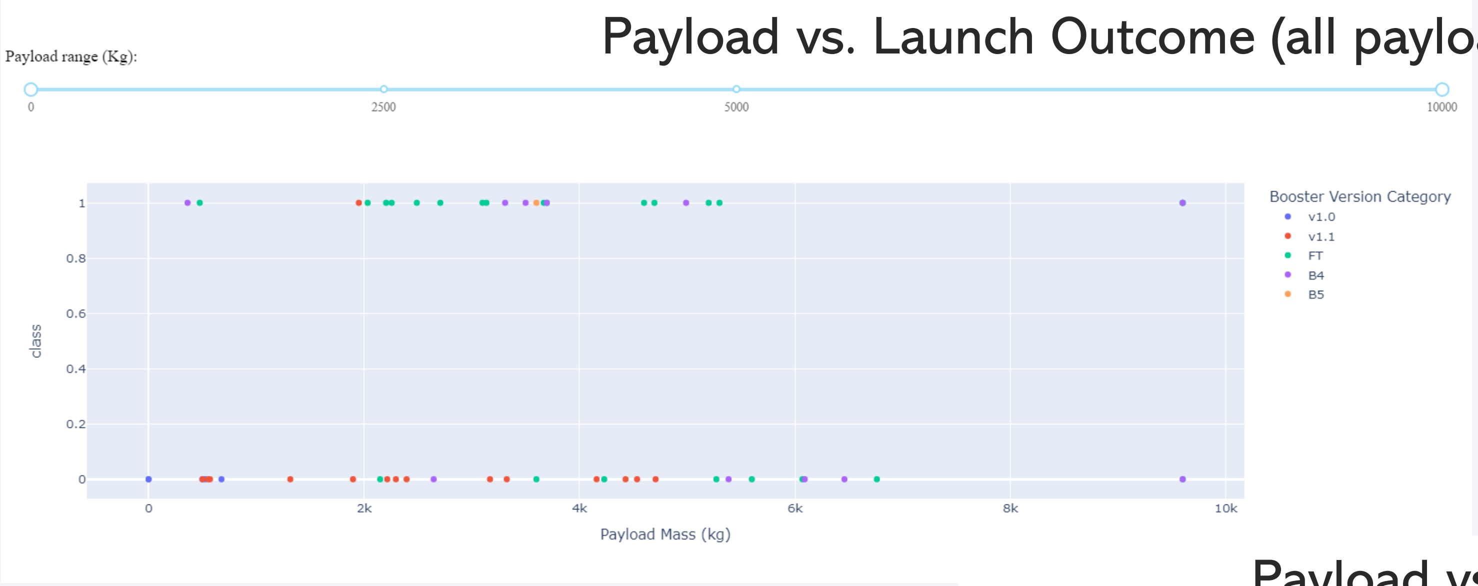
SpaceX Launch Records Dashboard

Total Success Launches for site KSC LC-39A

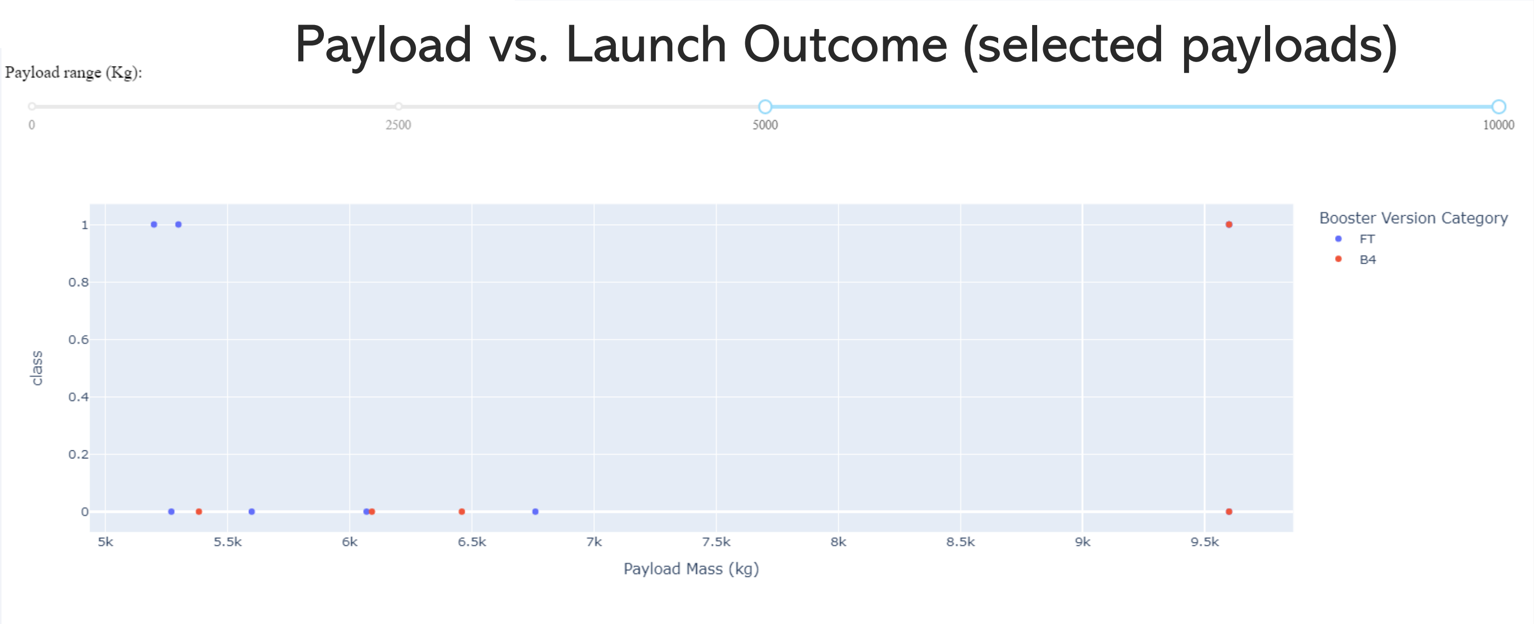


- The launch success rate for KSC LC-39A is around 77%

SpaceX Launch Records Dashboard



- FT booster version has the largest success rate



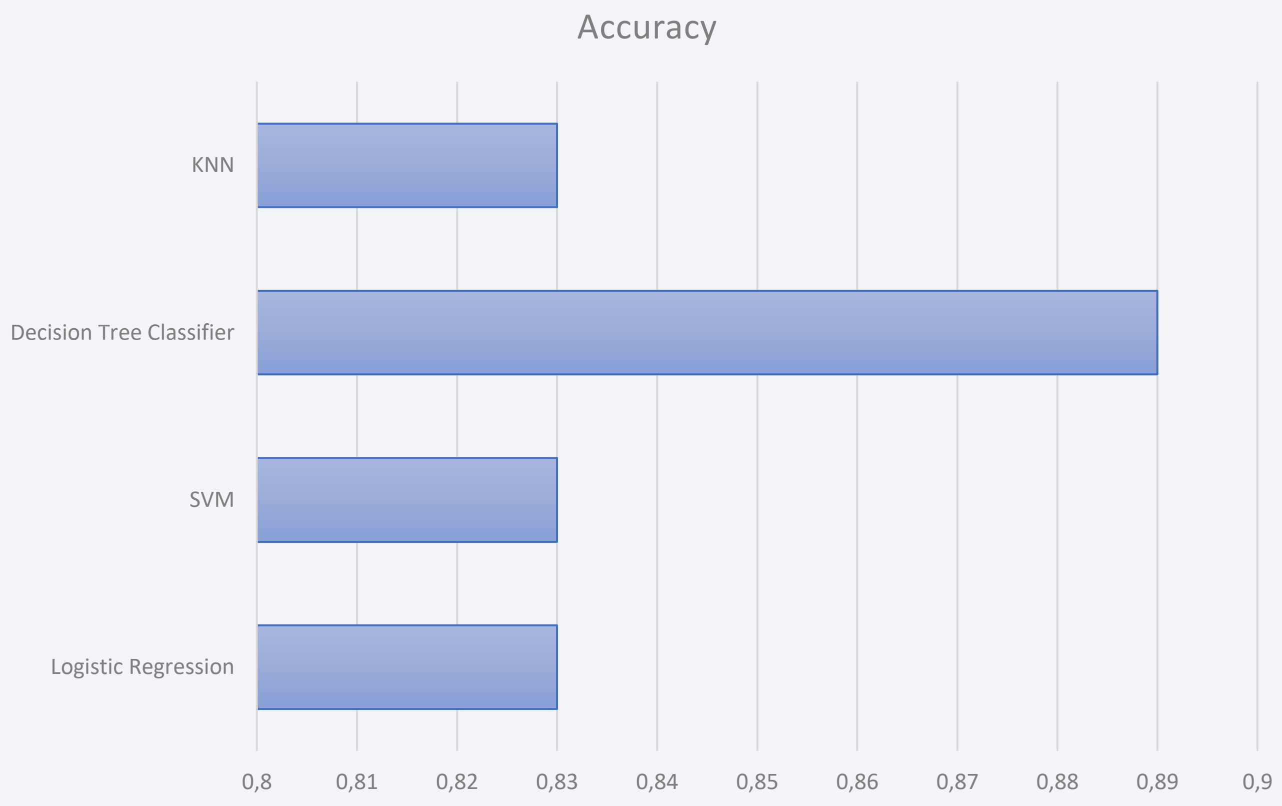
- For higher payload range there are less successful launch outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision tree classifier has the largest accuracy



Confusion Matrix

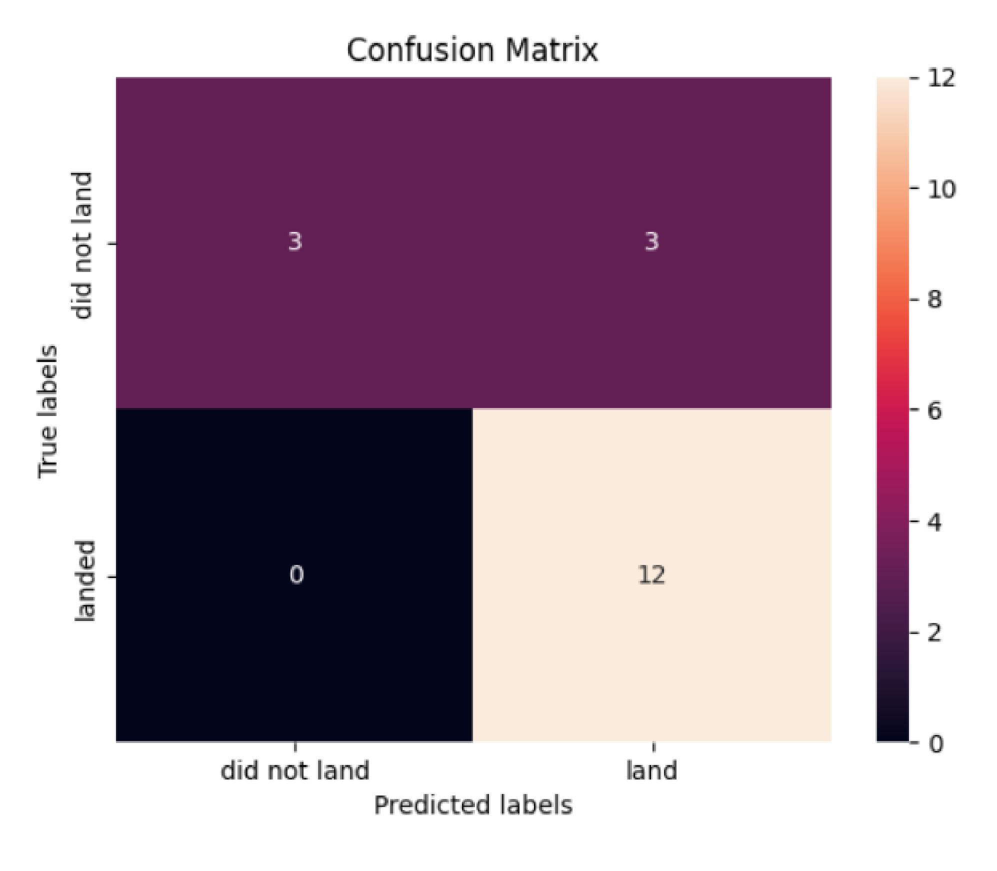
- The classifier correctly predicted 'did not land' for 3 out of 6.
- The classifier correctly predicted 'landed' for 12 cases out of 12 which shows a very good result.

'did not land'

- Precision = 3/(3+0)=1
- Recall = 3/(3+3)=0.5
- F1-score = 2 * Precision * Recall / (Precision + Recall) = 2 * 1.0 * 0.5 / (1.0 + 0.5) = 0.67

'land'

- Precision = TP / (TP + FP) = 12 / (12 + 3) = 0.8
- Recall = TP / (TP + FN) = 12 / (12 + 0) = 1.0
- F1-score = 2 * Precision * Recall / (Precision + Recall) = 2 * 0.8 * 1.0 / (0.8 + 1.0) = 0.89



Conclusions

Determination of the launch cost largely depends on the fact that first stage lands successfully in order to reuse it again. We performed an analysis to identify the factors which can affect the success of the mission. According to our data analysis we can make some conclusions:

- The success rate of landing raised since 2013 and with a number of launches
- It depends on such property as payload mass: the larger payload – the less successful landing is.
- The most successful rate was observed for KSC LC-39A launch site
- Success of the launch outcome can be determined with 89% accuracy based on the tree classification model

Thank you!