IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kelley Kent

20-Jun-2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection w/ REST API
  - Data wrangling w/ transformations
  - Exploratory data analysis (visualization)
  - Exploratory data analysis (SQL)
  - Interactive map w/ Folium
  - Predictive analytics w/ Classification
- Summary of all results
  - Exploratory data analysis
  - Predictive analytics

Introduction

- Project background and context
  - SpaceX is one of many companies competing for the commercial space travel market
  - Using Falcon 9 rockets, SpaceX has kept the cost of its launches significantly smaller than its competitors ($62 million vs $165 million)
  - One key factor to SpaceX's lower costs is their ability to reuse the first stage of any Falcon 9 launch
  - When a new launch successfully lands, then the first stage can be reused
  - The better SpaceX is at predicting and completing successful landings, the more money they will save thereby increasing their competitive edge of their competitors
- Problems you want to find answers
  1. What datapoints influence the success of a first stage landing?
  2. What is the optimal predictive model for the given datapoints of first stage landings?

Section 1

Methodology

Methodology (1of 2)

Executive Summary

- Data collection methodology:
  - Rest API from SpaceX website (all launches)
  - Webscrape the Wikipedia SpaceX page (Falcon 9 only)
- Perform data wrangling
  - Only include launch data from Falcon 9 rockets
  - Transform categoricanl variables into factors
  - Replaced missing numerical data with sample mean

Methodology (2 of 2)

Executive Summary (cont.)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Found best model parameters using cross validation
  - Tuned multiple models for predicting successful landing
  - Best model chosen based on highest prediction accuracy of both train/test datasets

Data Collection

- Data on all SpaceX launches obtained @ https://api.spacexdata.com/v4
  - REST API used for data collection
  - Datapoints used: *FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*
- Data on Falcon 9 Space X launches obtained @ https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Webscraping (Beautiful Soup) used for data collection
  - Datapoints used: *Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time*

Data Collection – SpaceX API

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/Data%20Collection%20API.ipynb

1. get all JSON data from SpaceX API

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

```
data = pd.json_normalize(response.json())
```

2. Extract specific datapoints from API data

a) Rockets
```
response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
```

b) Launch Site
```
response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
```

c) Payload
```
response = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
```

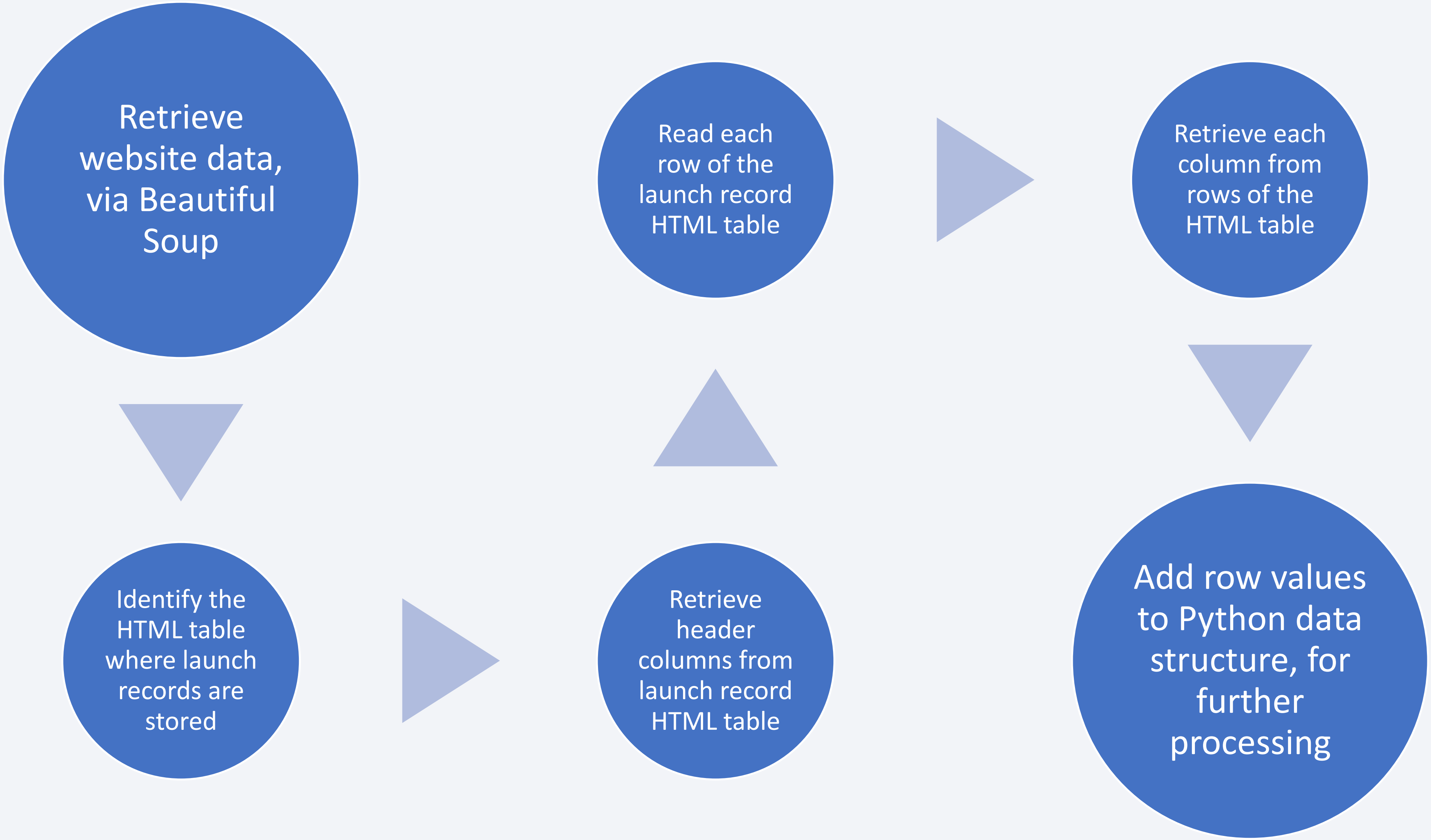
d) Core
```
response = requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
```

3. Data transformation / normalization

a) Remove launches that occurred after November 13, 2020
b) Remove all non-Falcon V9 data rows
c) Remove launches with multiple payloads
d) Replace missing Payload Mass values with column mean

Data Collection - Scraping

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/jupyter-labs-webscraping(1).ipynb



Data Wrangling

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/Data%20wrangling.ipynb

Objective 1: Determine the mission outcome values that indicate successful

- Find the unique labels in the Outcome column
- Enumerate each unique label
- Create a variable that holds all Outcome enumerations for bad outcomes

```
landing_outcomes = df['Outcome'].value_counts()
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
```

Objective 2: create a landing outcome column label from Outcome column

- Assign 0 where original Outcome value is indicative of a bad / failed landing; assign 1 otherwise

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for key, value in df['Outcome'].items():
    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
df['Class']=landing_class
```

Data Wrangling (cont.)

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/Data%20wrangling.ipynb

Objective 3: Calculate the number of launches at each site

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Objective 4: calculate the number and occurrence of each orbit

```
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

EDA with Data Visualization

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

- Visualize orbits of rockets
  - Larger orbits require more fuel which may impact launch landing success
- Bar Chart to show comparison of orbit launch successes
  - Helps to show a relationship between orbit radius and landing success
- Scatter Plots to show the relationships between two variables (see below)

  If a relationship exists, then that datapoints could be useful in machine learning modeling

  1) Flight Number vs. Payload Mass
  2) Flight Number vs. Launch Site
  3) Payload Mass vs Launch Site
  4) Orbit vs Flight Number
  5) Payload Mass vs. Orbit
- Line Graph to show trend of landing success over the years

EDA with SQL

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/EDA%20with%20SQL.ipynb

- Find unique launch site names
- 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 successful landing outcome in ground pad was achieved
- 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
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

Build an Interactive Map with Folium

- Visualized location of launch sites w/ circle markers and display labels
- At each launch site locations, indicated the number of successful outcomes (green markers) and failed outcomes (red markers)
- Lines and distances are shown between each launch site and nearest city, highway, coastline and railway

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/spacex_dash_app.py

- Interactions
  - Drop down list added for selection of launch site
  - Slider to select Payload range

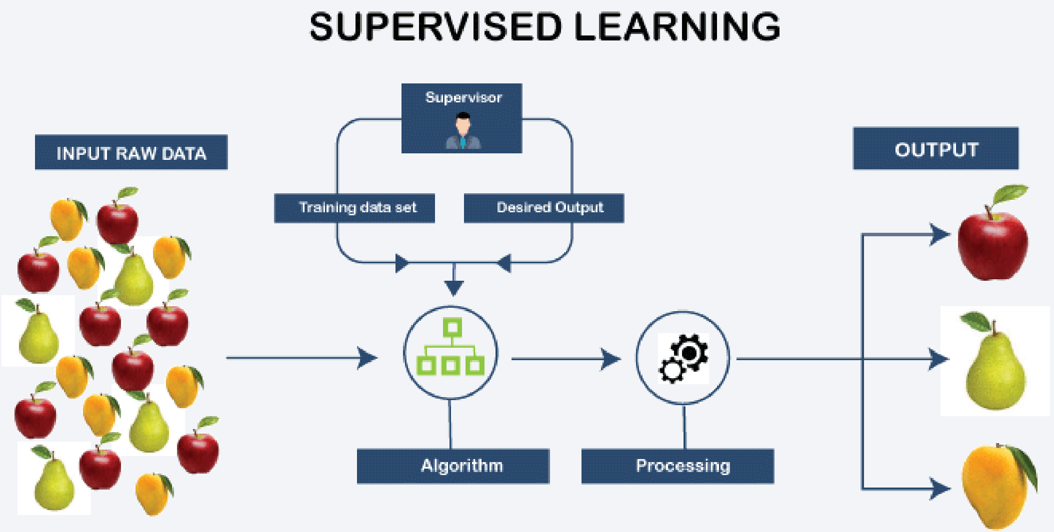
- Visualizations
  - Pie chart added showing breakdown of launch outcome (success, failure)
  - Scatter chart to show correlation between Payload and Launch Success

Predictive Analysis (Classification)

Github URL:
https://github.com/StickyKiwiIBM/IBM_datascience_capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Building

- Transform data via column scaling
- Split data into test and train sets
- Use grid search to identify best parameters
- Run 4 models: LogReg, SVM, Decision Tree and KNN



Evaluation

- Calculate the accuracy of all 4 models
- Plot the confusion matrix for all 4 models

Tuning

- Features engineering

Selection

- Pick model with best accuracy scores

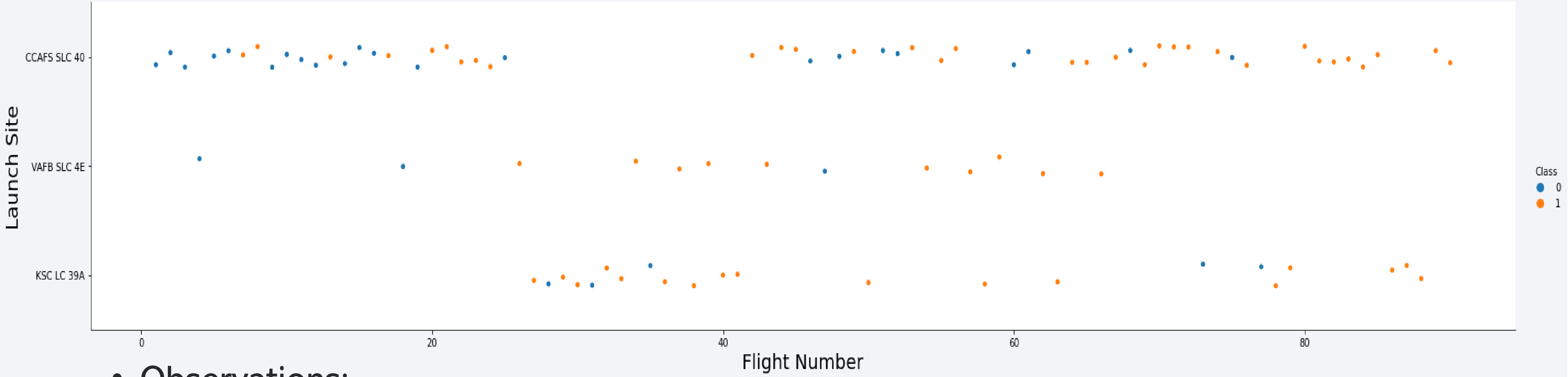
Results

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

Section 2

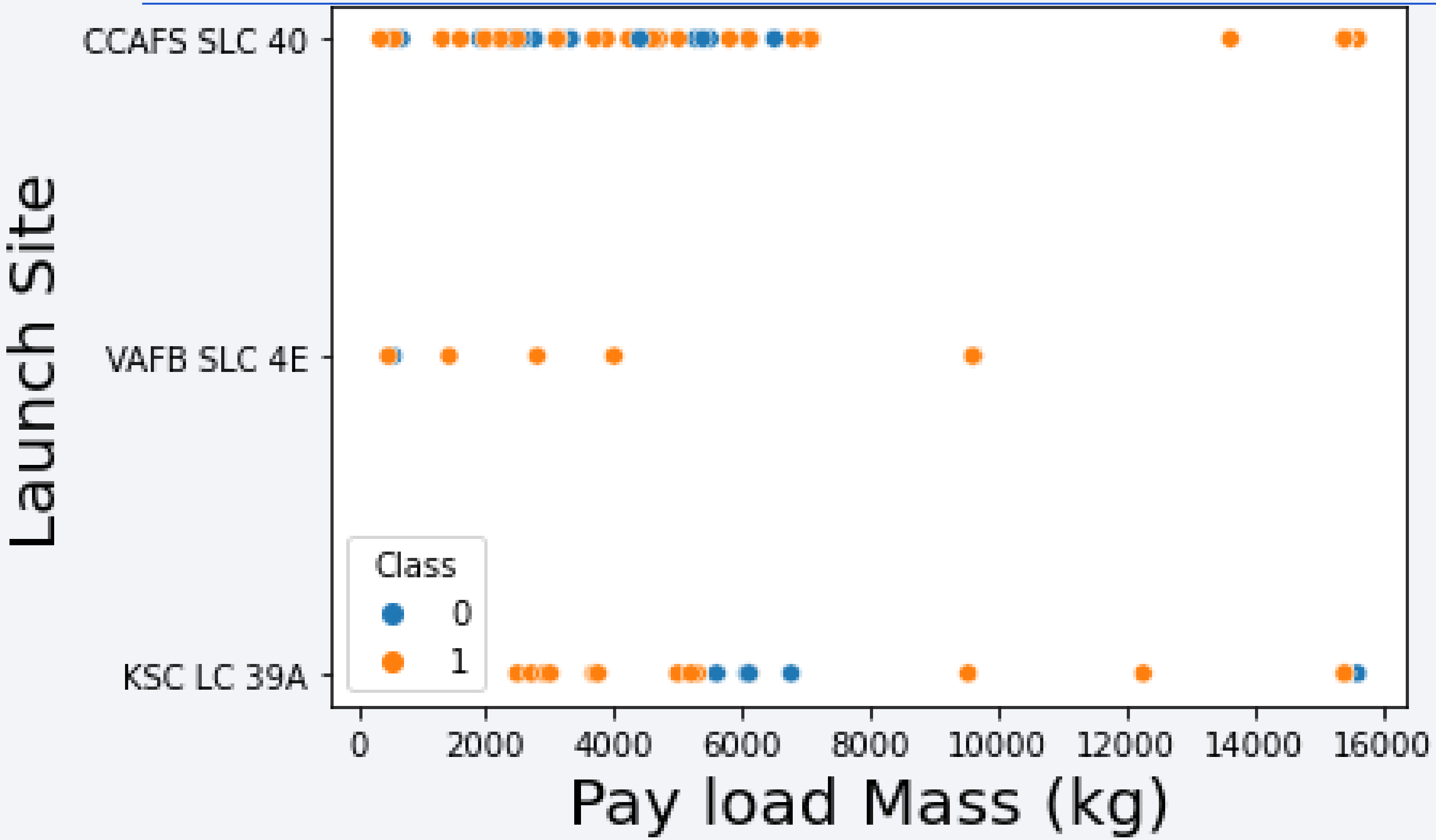
Insights drawn from EDA

Flight Number vs. Launch Site



- Observations:
  - The CCAFS SLC 40 launch site has the most flights of all site locations
  - All of the most recent launches (roughly beginning with flight number 78) have been successful
  - Within each launch site, the most recent 5 launches have all been successful
  - The majority of failures occurred before flight #40

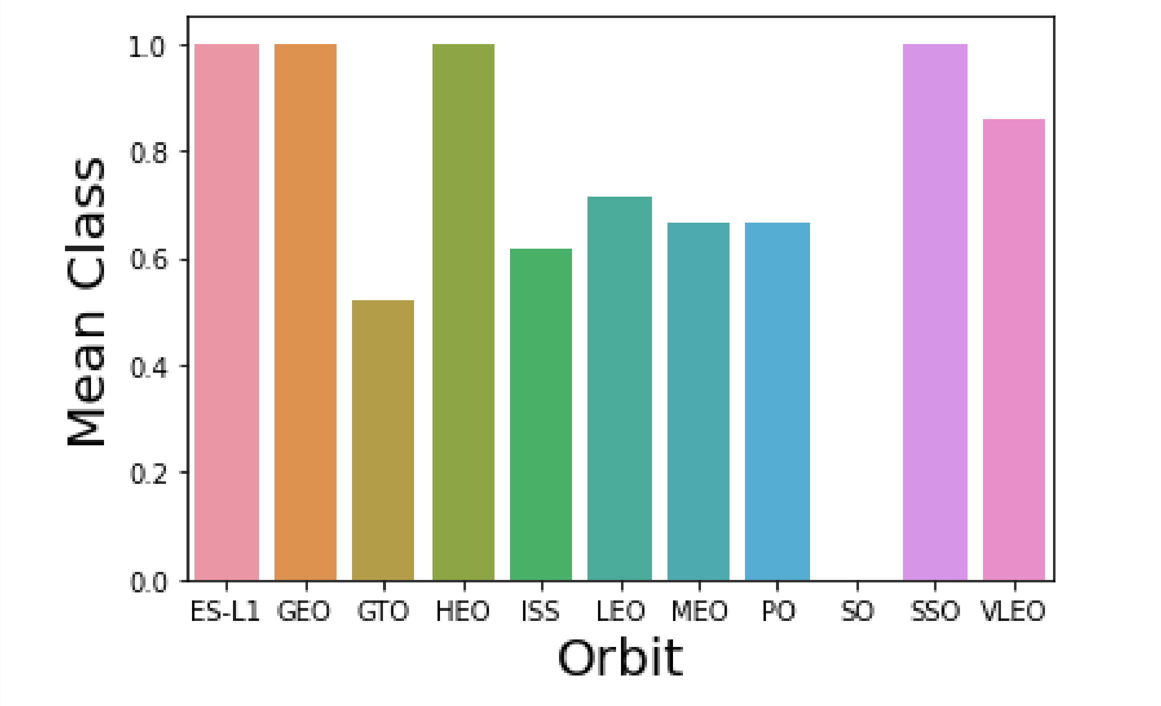
Payload vs. Launch Site



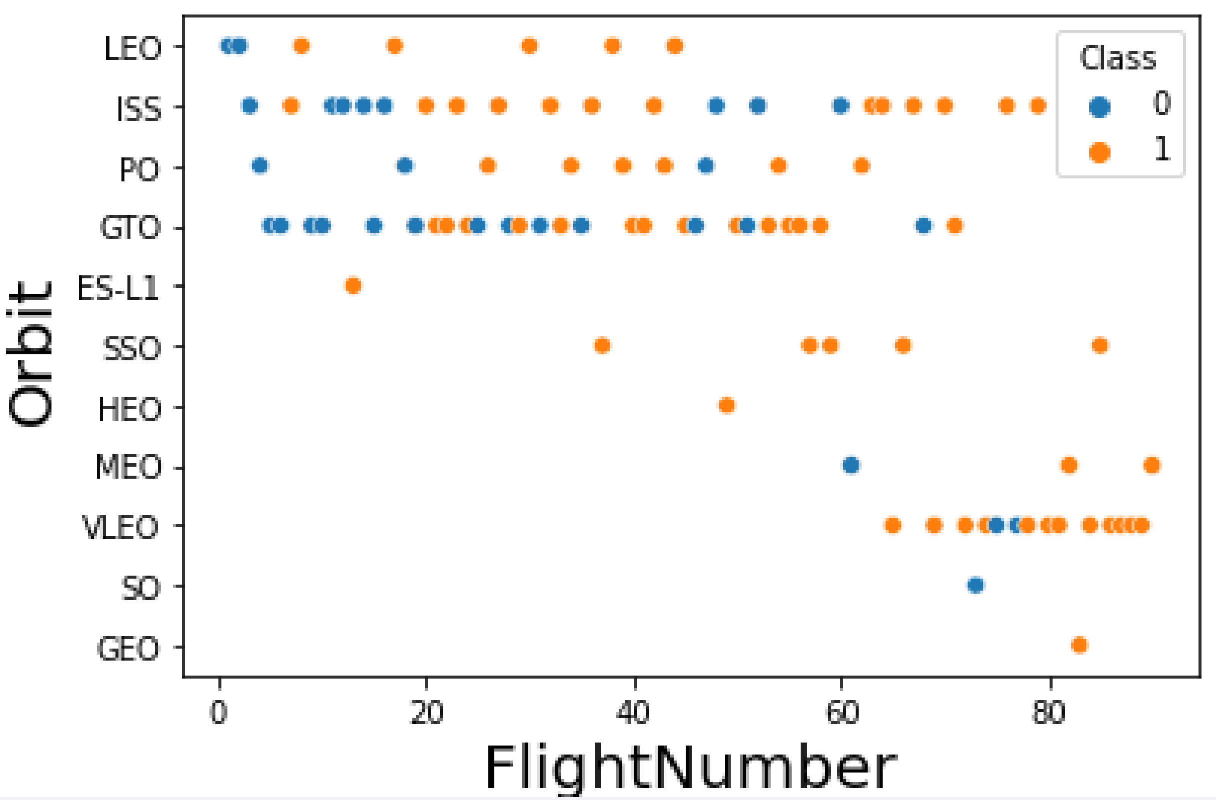
- Observations:
  - The higher the payload mass, the higher the success rater across all sites
  - Nearly all launches with payload greater than 8000kg were successful
  - The VAFB SLC 4E site did not have any launch data with payload greater than 9000kg

Success Rate vs. Orbit Type

- Observations:
    - 4 orbit types have 100% success rate: ES-L1, GEO, HEO, SSO
    - The VLEO orbit has 85% success rate
    - The SO orbit has a 0% success rate
    - All other orbit types have between 50% - 70% success
    - Focus should be on ES-L1, GEO, HEO, SSO and VLEO orbits



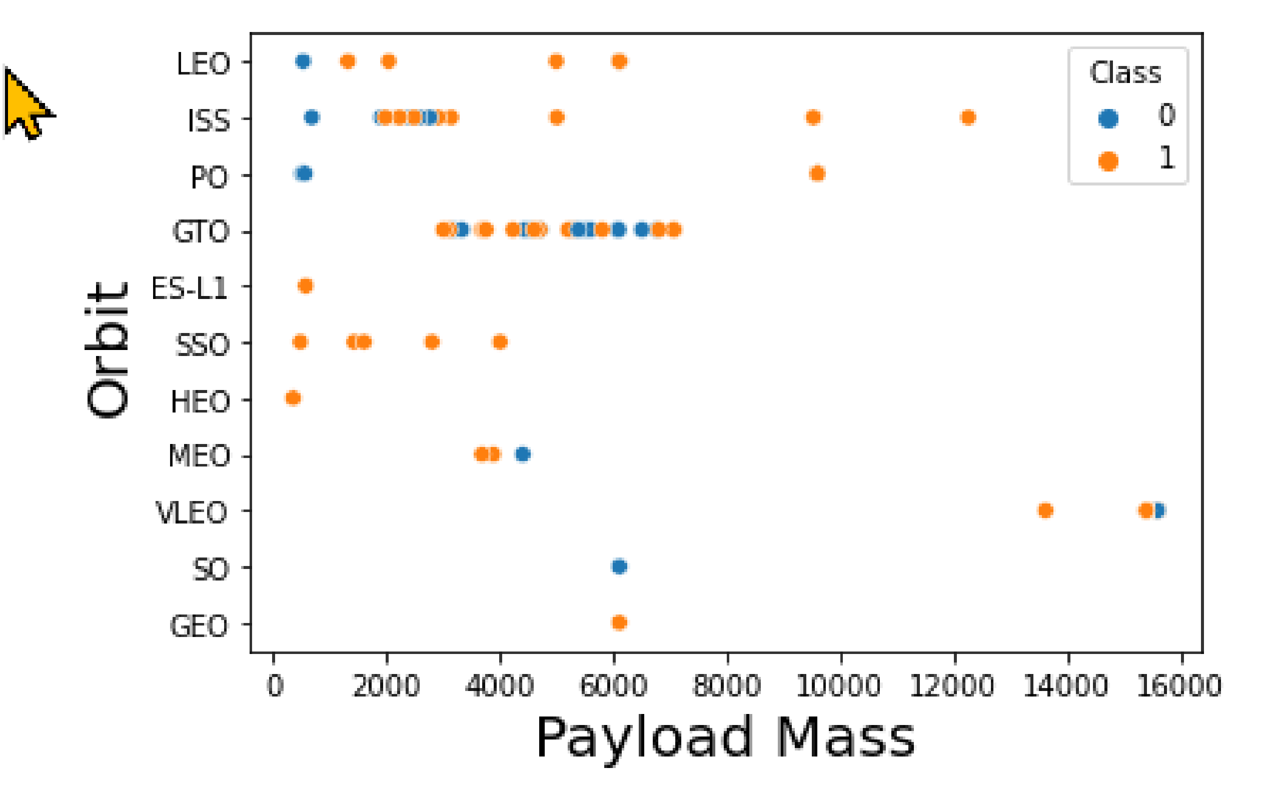
Flight Number vs. Orbit Type



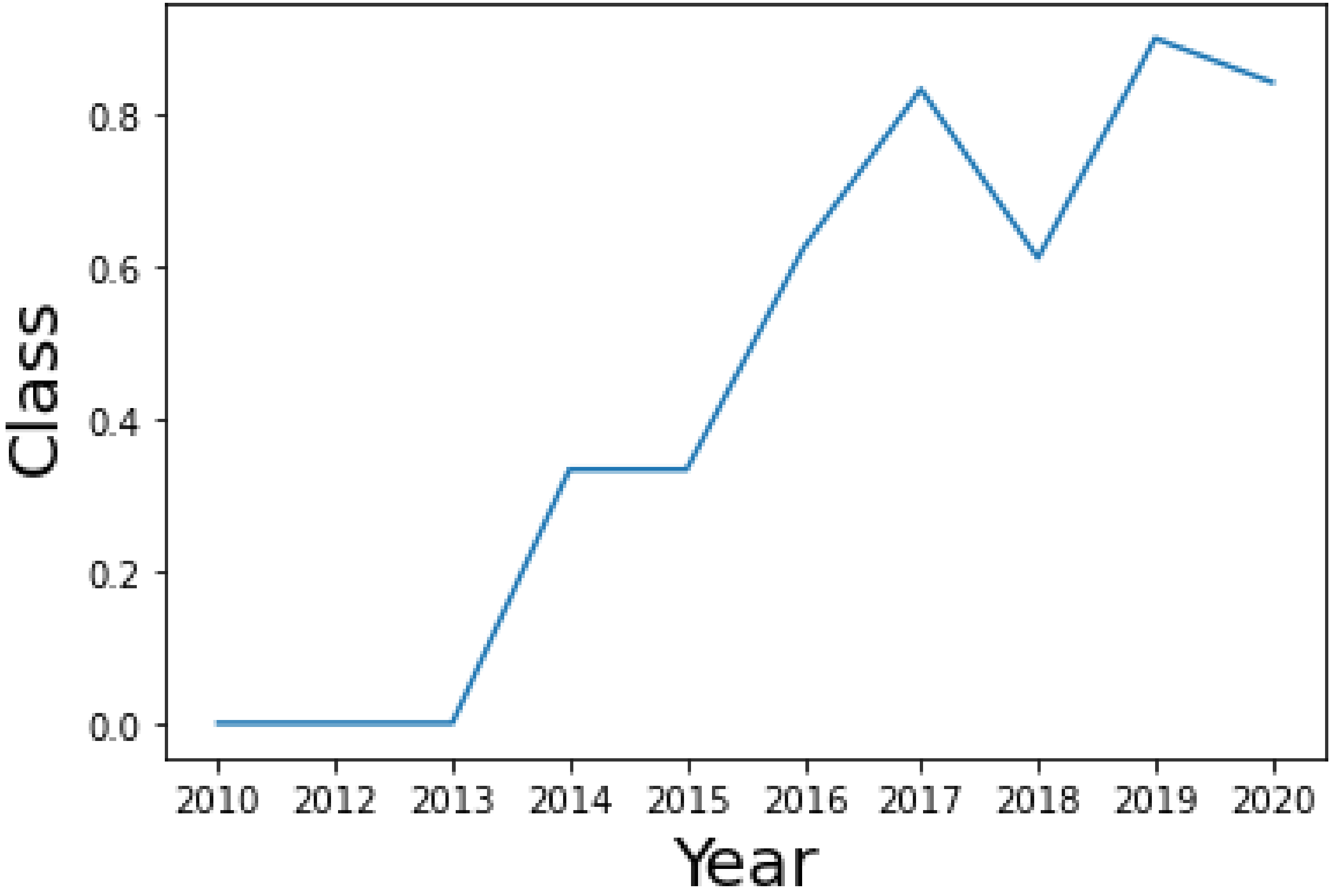
- Observations
  - Only LEO and VLEO appear to show a strong correlation between the number of flights and success
  - ISS, GTO and PO orbits show inconclusive correlations
  - All other orbit types do not have enough data to draw references

Payload vs. Orbit Type

- Observations
  - Heavier loads appear to have a negative impact on GTO orbits
  - Heavier loads appear to have a positive impact on PO, LEO and ISS orbits
  - All other orbit types do not have enough data to draw references



Launch Success Yearly Trend



- Observations
  - All launches prior to 2013 were failures
  - The trend between 2014 and 2020 is that the rate of successes increased
  - 2018 is an anomaly in that the rate decreased; further investigation is required to provide context

All Launch Site Names

```
%sql select distinct LAUNCH_SITE from SPACEXTBL
```

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

- There are 4 launch sites in total

Launch Site Names Begin with 'CCA'

```
%sql select unique LAUNCH_SITE from SPACEXTBL where LAUNCH_SITE like 'CCA%' order by 1
```

launch_site
CCAFS LC-40
CCAFS SLC-40

- There are 2 sites that begin with ‘CCA’
- A sample of 5 data rows for any launch site beginning with ‘CCA’ is shown below

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

DATE	time_utc	booster_version	launch_site	payload	payload_mass	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS) as total_payload_mass from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

total_payload_mass
45596

- Using the customer column, we can see that the total sum of all payloads for NASA was 45,596

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS) as avg_payload_mass from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

avg_payload_mass
2928

- Using the booster version column, we can see that the average payload for the F9 rockets is 2,928

First Successful Ground Landing Date

```
%sql select min(DATE) date_of_1st_gpad from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)'
```

date_of_1st_gpad
2015-12-22

- Using the landing outcome column, we can see that that earliest landing date was December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select unique BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS between 4
001 and 5999
```

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

- Using the columns landing outcome and payload mass, we can see that there are 4 booster versions that had successful landings with payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select count(*) as total, 'success' as category from SPACEXTBL where MISSION_OUTCOME like 'Success%'
union all
select count(*), 'failure' from SPACEXTBL where MISSION_OUTCOME like 'Failure%'
order by 1 desc
```

total	category
100	success
1	failure

- Using the column mission outcome, we can see that there have 100 successful landings and only 1 failure

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS = (select max(PAYLOAD_MASS) from SPACEXTBL)
```

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

- Using the column payload mass, we can see that there are 12 booster versions that have launched with the maximum payload

2015 Launch Records

```
%sql select DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME from SPACEXTBL where LANDING_OUTCOME = 'Failure (drone shi
p)' and year(DATE) = 2015
```

DATE	booster_version	launch_site	landing_outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Using the columns data and landing outcome, we can see that there were 2 failed landings in 2015

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

```
%%sql select LANDING_OUTCOME, count(*) total
from SPACEXTBL
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 2 desc
```

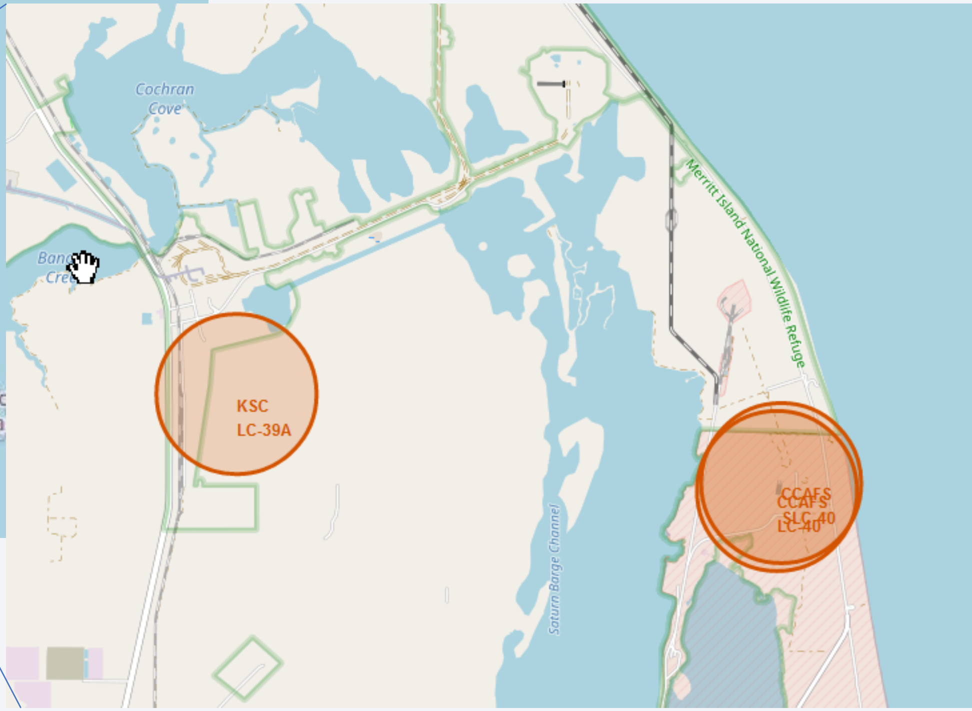
landing_outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Using the date column, we can see that there were 8 distinct landing outcomes between June 4, 2010 and March 20, 2017 as well as how many landings occurred for each outcome
- The most frequent landing outcome during this time period was “no attempt”

Section 3

Launch Sites Proximities Analysis

US Map of SpaceX Launch Sites

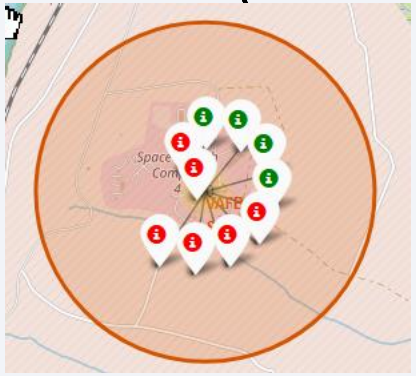


- 4 launch site locations
  - VAFB SLC-4E (California)
  - CCAFS LC-40 (Florida)
  - KSC LC-39A (Florida)
  - CCAFS SLC-40 (Florida)

Launch Site landing outcomes

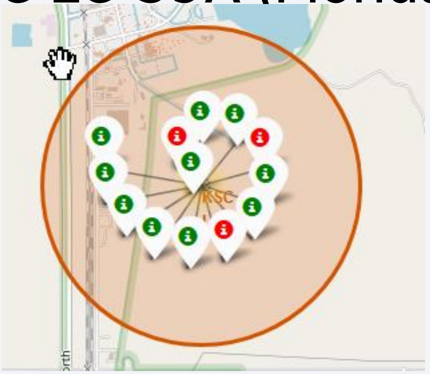
- For each site, green markers were added for each successful landing; red markers for failed landings

VAFB SLC-4E (California)



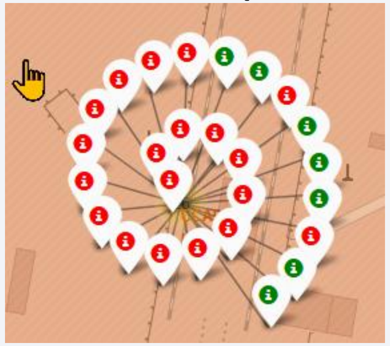
4 and 6

KSC LC-39A (Florida)



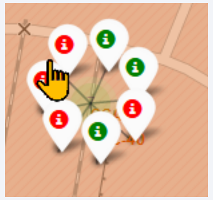
10 and 3

CCAFS LC-40 (Florida)



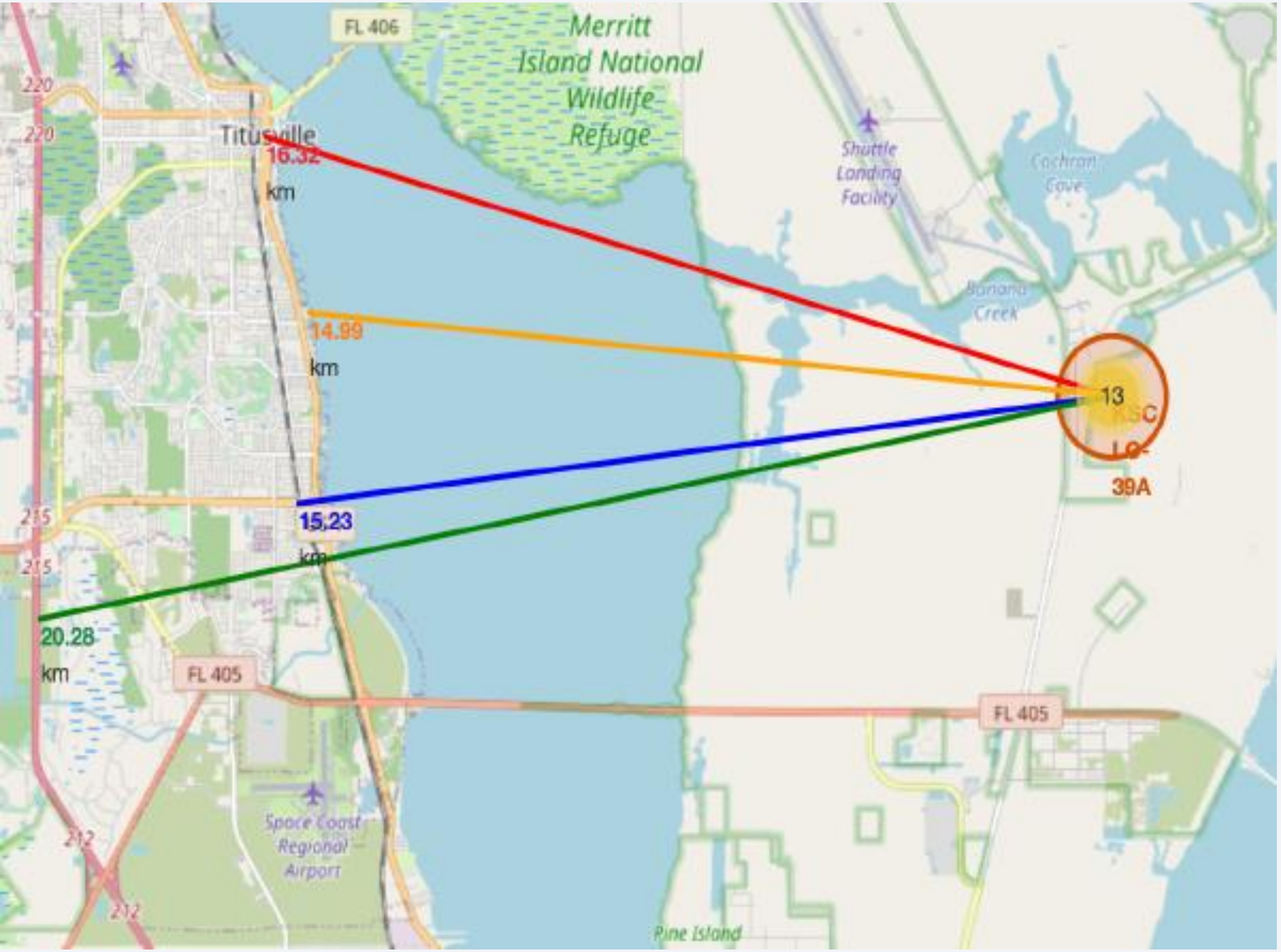
7 and 19

CCAFS SLC-40 (Florida)



3 and 4

Site KSC LC-39A (Florida) proximities

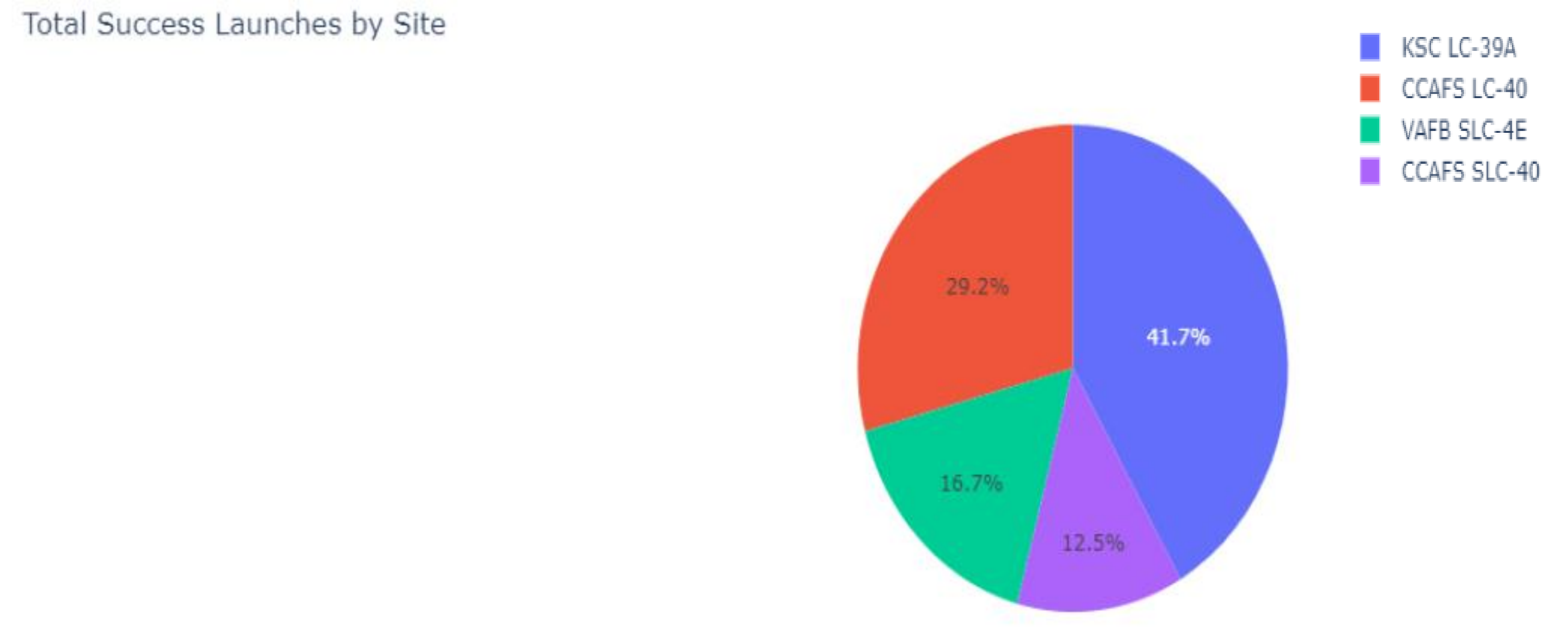


- For site KSC LC-39A:
  - Closest railway = 15.23km
  - Closest highway = 20.28km
  - Closest coastline = 14.99km
  - Closest city = 16.32km (Titusville)

Section 4

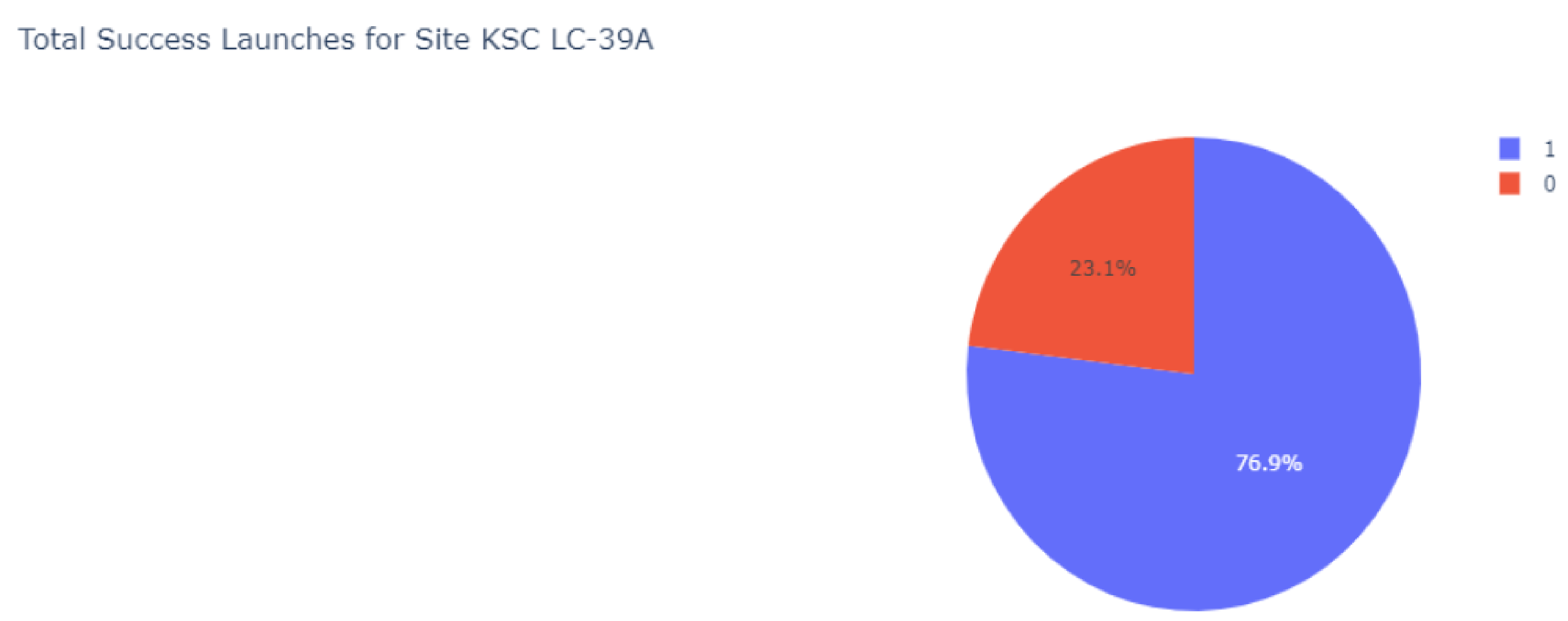
Build a Dashboard with Plotly Dash

Successful launches by site



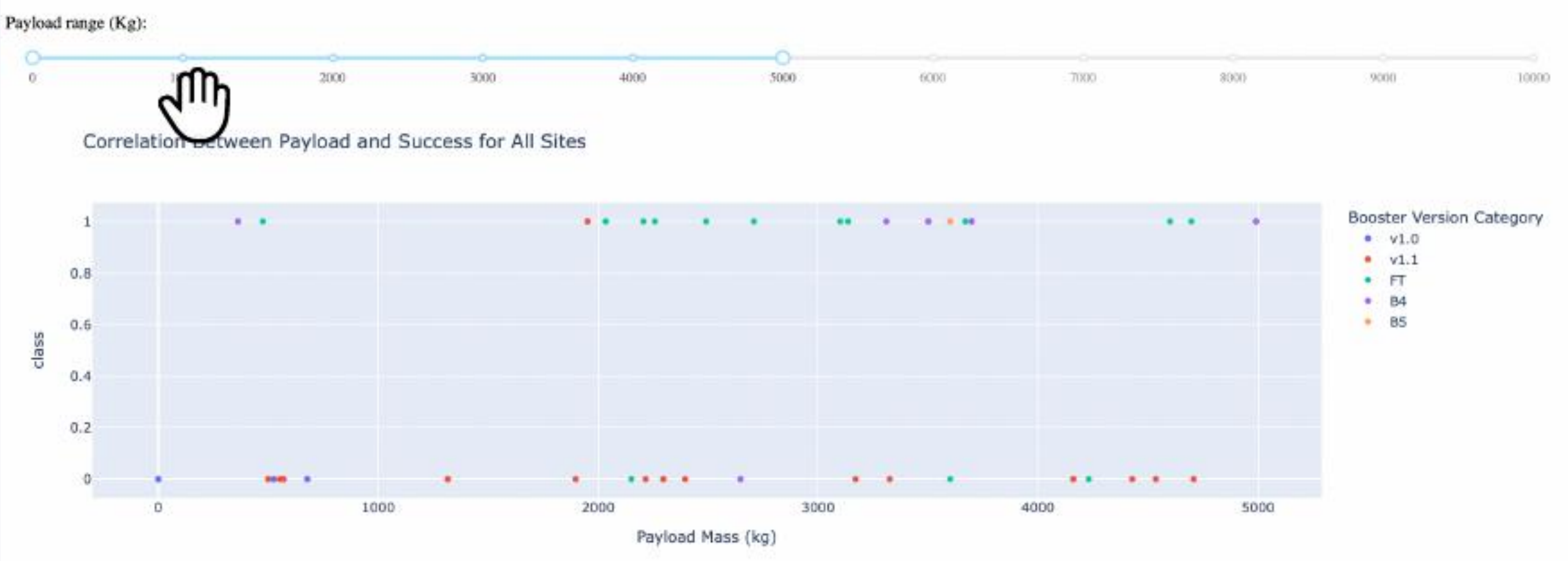
- Site KSC LC-39A (Florida) accounts for the most of all successful launches at 41.7%

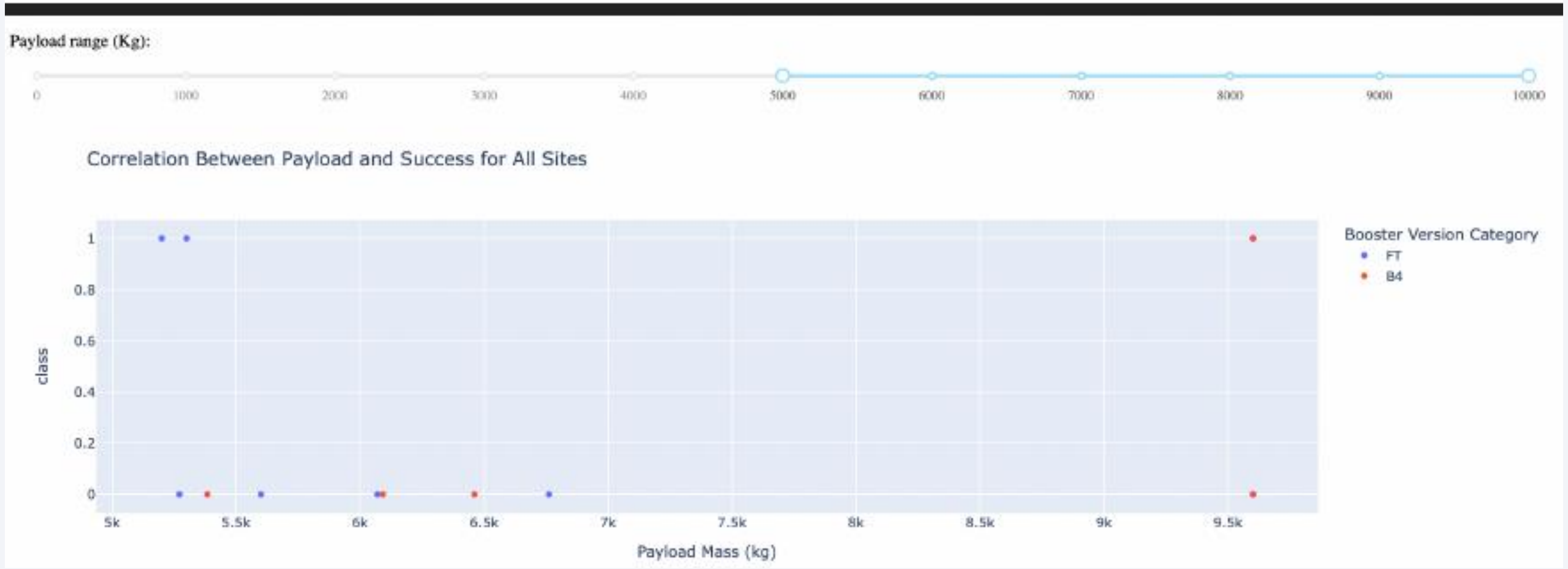
Site KSC LC-39A (Florida) probability of success



- 76.9% of all launches at site KSC LC-39A land successfully

Launch Outcome vs Payload Mass for all sites





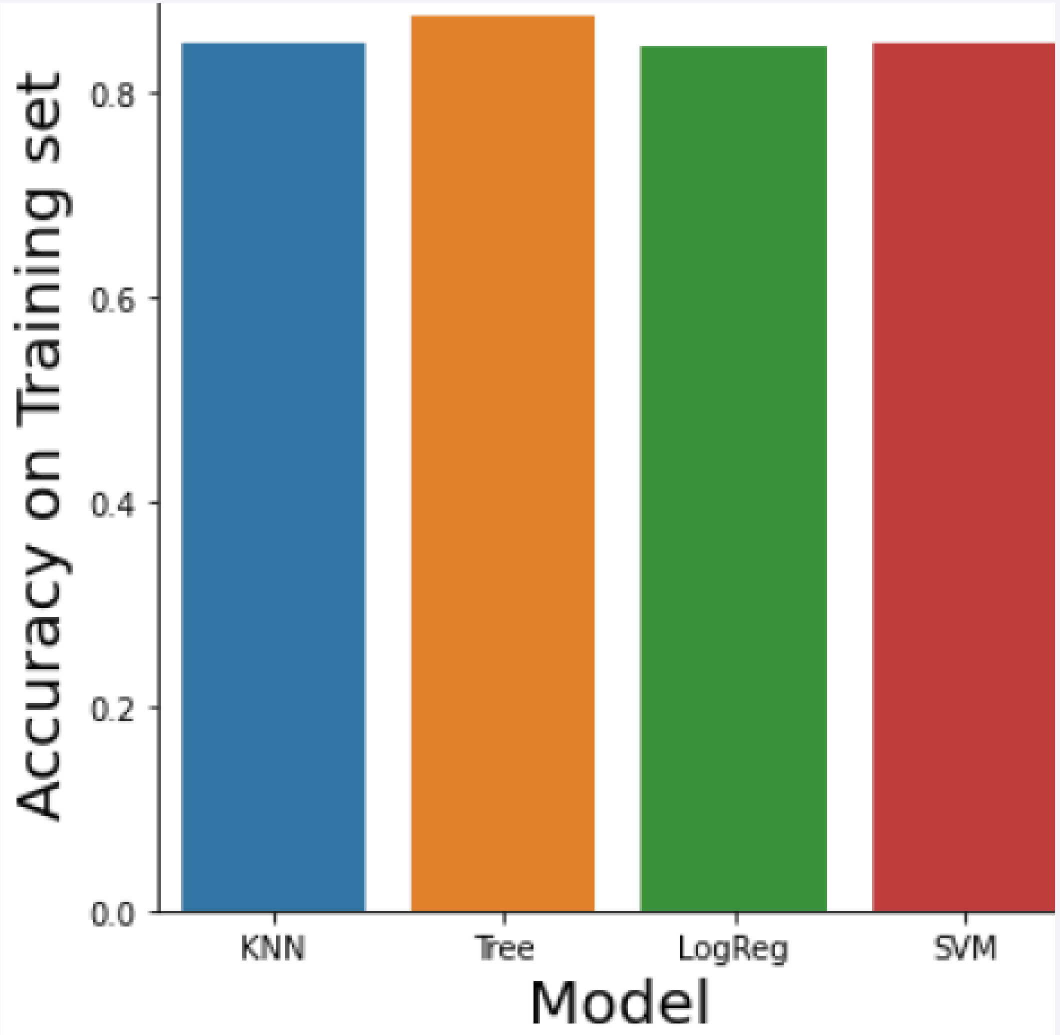
- Payloads under 5000kg have more successful landings

Section 5

Predictive Analysis (Classification)

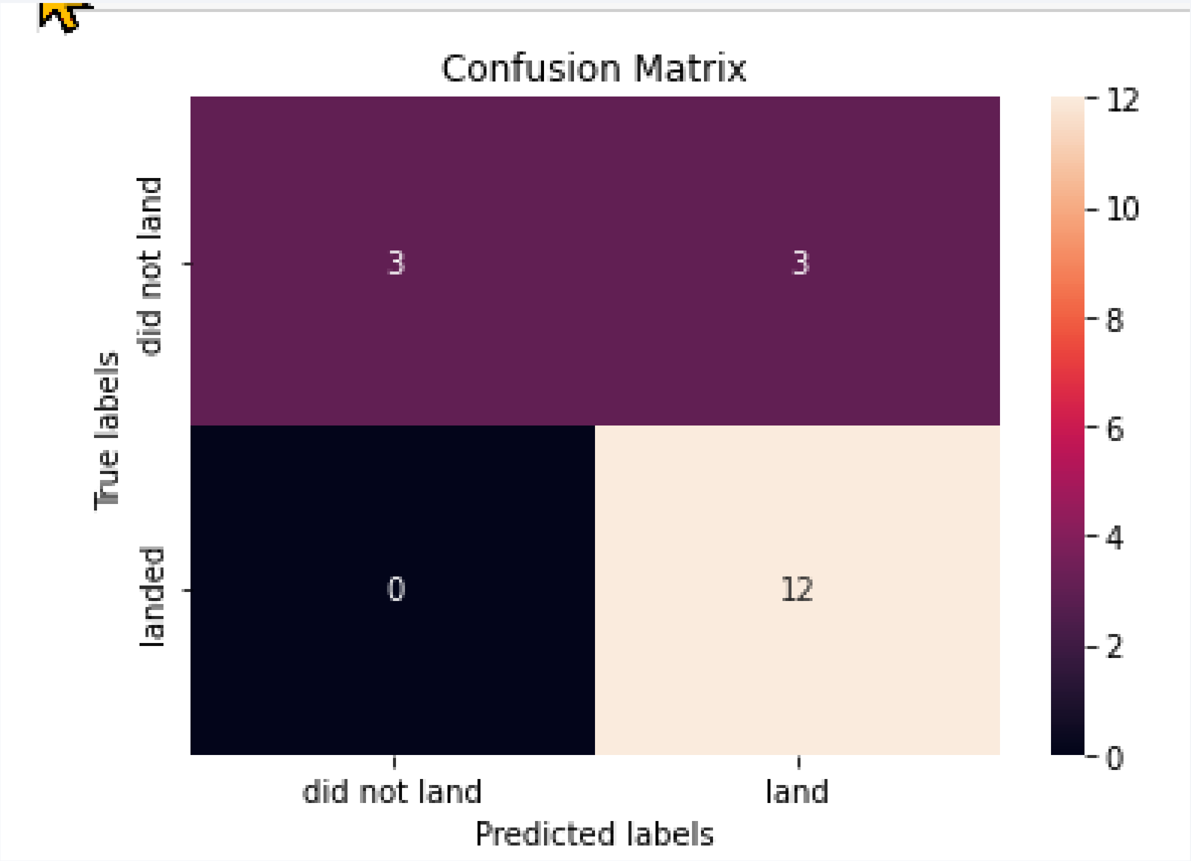
Classification Accuracy

- The tree model has the highest accuracy at 87.5%, but it is close to all other models which had 85%



Confusion Matrix

- True negatives (top left) = 3
- False positives (top right) = 3
- False negatives (bottom left) = 0
- True Positives (bottom right) = 12



- Summary
  - The combination of low (zero) False Negatives and high True Positives, this is a good model – but there is room for improvement
  - Additional feature engineering could prove useful

Conclusions

- Using the KSC LC-39A (Florida) site for the launch gives us the best probability for success
- Low weight payloads (under 5000kg) give a higher probability of success
- Using orbits of ES-L1, GEO, HEO, SSO give a higher probability of success
- Decision Tree was found to be the best performing model for predicting Falcon 9 Rocket landing success

Thank you!