IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vince Lauro
Aug. 30, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo (screenshots)
  - Predictive analysis results

Introduction

- Project background and context

  The era of commercial space has arrived, and there are several companies that are making space travel affordable for everyone. Perhaps the most successful of them all is SpaceX, and one of the reasons is that their rocket launch is relatively inexpensive.

  SpaceX advertises Falcon9 rocket launches on its website with a cost of $62M, while other providers cost upwards of $165M – much of the savings is because SpaceX is able to recover and reuse the 1st stage rocket.

  Therefore, we will predict if the Falcon 9 first stage will land successfully, and similarly we can determine the cost of a launch.

- Topics to seek answers for
  - Correlations between rocket variables and successful landing rate
  - Conditions to get the best results and ensure the best successful landing rate

Section 1

Methodology

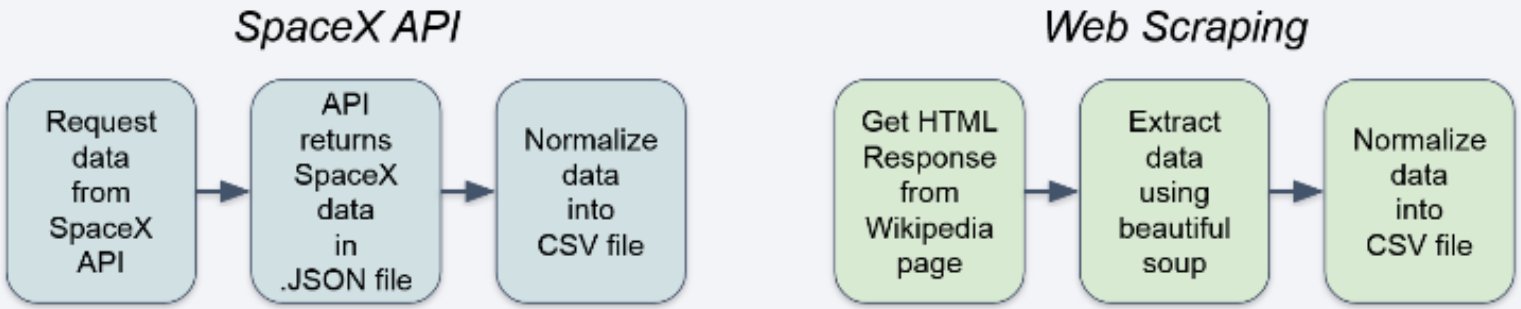
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API & Web Scraping Falcon 9 and Falcon Launches from Wikipedia
- Perform data wrangling
  - Convert outcomes into Training Labels (successful/unsuccessful landings)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees, and Logistic Regression

Data Collection

- The data collection process included a combination of API requests from the SpaceX API and web scraping data from the Falcon 9 and Falcon Heavy Launch Records.
  - SpaceX API Data: *FightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*
  - Web Scrape Data: *Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, time*



Data Collection – SpaceX API

1. Requested rocket data from SpaceX API

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

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

2. Converted response to JSON file

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

3. Used custom functions to clean data

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

4. Combined into a dictionary -> df

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Filtered dataframe and export to CSV

```
data_falcon9 = launch_df[launch_df['BoosterVersion'] == 'Falcon 9']
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

GitHub URL

Data Collection - Scraping

1. Get response from HTML

```
html_data = requests.get(static_url).text
```

2. Created a Beautiful Soup object

```
soup = BeautifulSoup(html_data, 'html.parser')
```

3. Find & add all tables to a list

```
html_tables = soup.find_all('table')
```

4. Extracted each column name

```
column_names = []
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if(name != None and len(name) > 0):
        column_names.append(name)
```

5. Created empty dictionary with keys

```
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Fill up launch_dict with records

7. Added to Dataframe and covert to CSV

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

GitHub URL

Data Wrangling

- There are several cases where the booster failed to successfully land, and sometimes it failed because of some accident.
  - *True Ocean: the mission has successfully landed (ocean)*
  - *False Ocean: the mission has not successfully landed (ocean)*
  - *True RTLS: the mission has successfully landed on the ground pad*
  - *False RTLS: the mission has not successfully landed on the ground pad*
  - *True ASDS: the mission has successfully landed on the drone ship*
  - *False ASDS: the mission has not successfully landed on the drone ship*

- Converting these results into training labels:
  - *successful = 1 , failure = 0*

GitHub URL

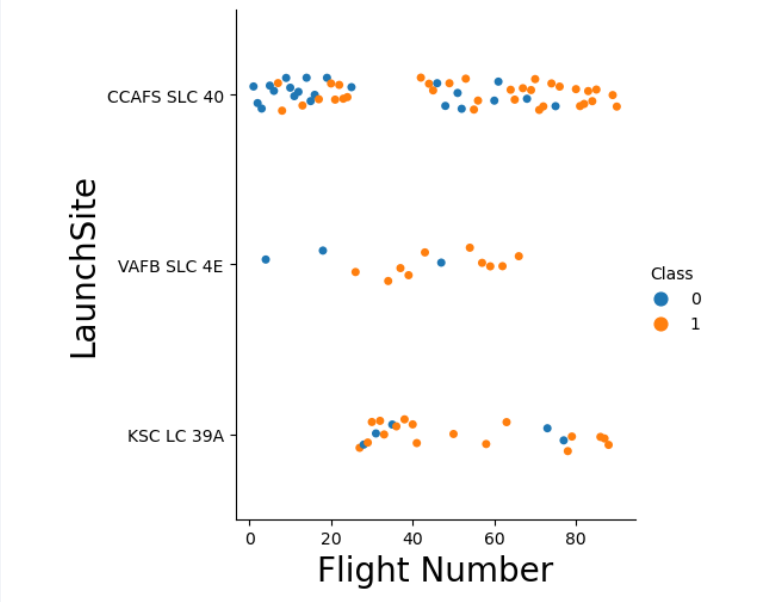
EDA with SQL

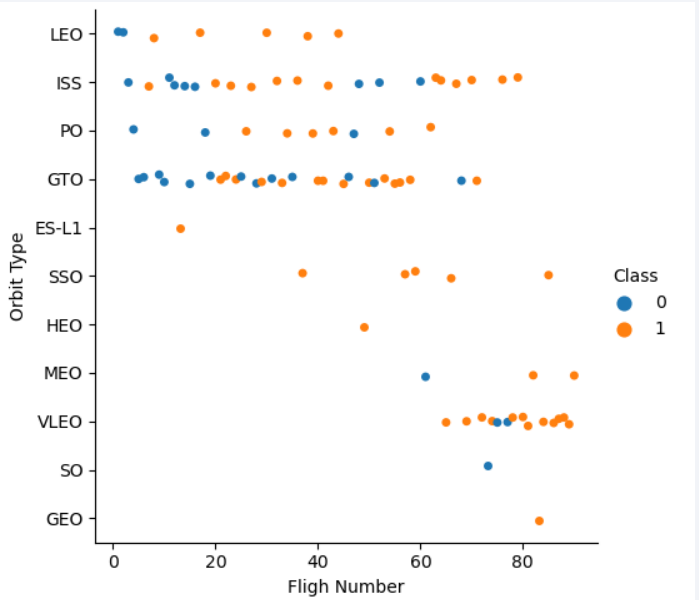
- Loaded the SpaceX database into Db2 database to answer the following queries:
  - Display the unique launch sites in the space mission
  - Display records with launch name string ‘CCA’
  - Display total payload mass launched by NASA (CRS)
  - List the total number of successful/failed mission outcomes
  - List the failed landing outcomes for each ship type in 2015

GitHub URL

EDA with Data Visualization

- Scatter Chart
  - *Flight Number vs. Launch Site*
  - *Payload vs. Launch Site*
  - *Flight Number vs. Orbit Type*
  - *Payload vs. Orbit Type*
- Bar Chart
  - *Orbit Type vs. Success Rate*
- Line Chart
  - *Year vs. Success Rate*





GitHub URL

Interactive Map with Folium

- Objects created and added to a Folium map:
  - Markers that show all launch sites on a map
  - Markers that show the successful/failed launches for each site on the map
  - Lines that measure distance from launch sites to nearby markers
- The following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes

GitHub URL

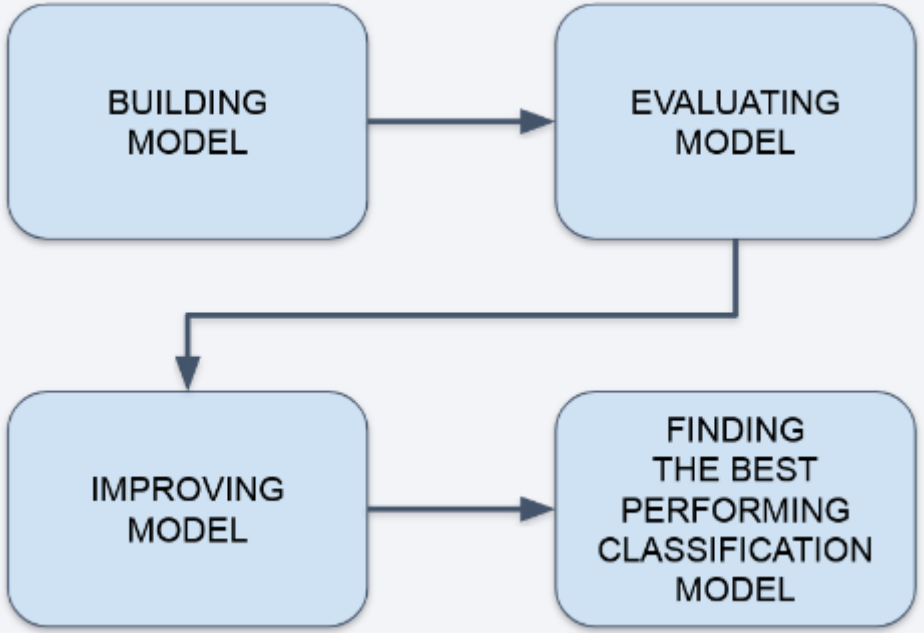
Dashboarding with Plotly Dash

- In Section 4, a dashboard application details successful launches by proportion and location
  - Proportional: total successful launches by site
  - Scatter: detail relationships between *Outcomes* and *Payload mass (Kg)* by booster type
  - Inputs: (1) By launch site, (2) By Payload mass (slider: 0 – 10,000kg)
  - Results: output visual detailing success rates by launch point, payload mass, and booster version categories

GitHub URL

Predictive Analysis (Classification)

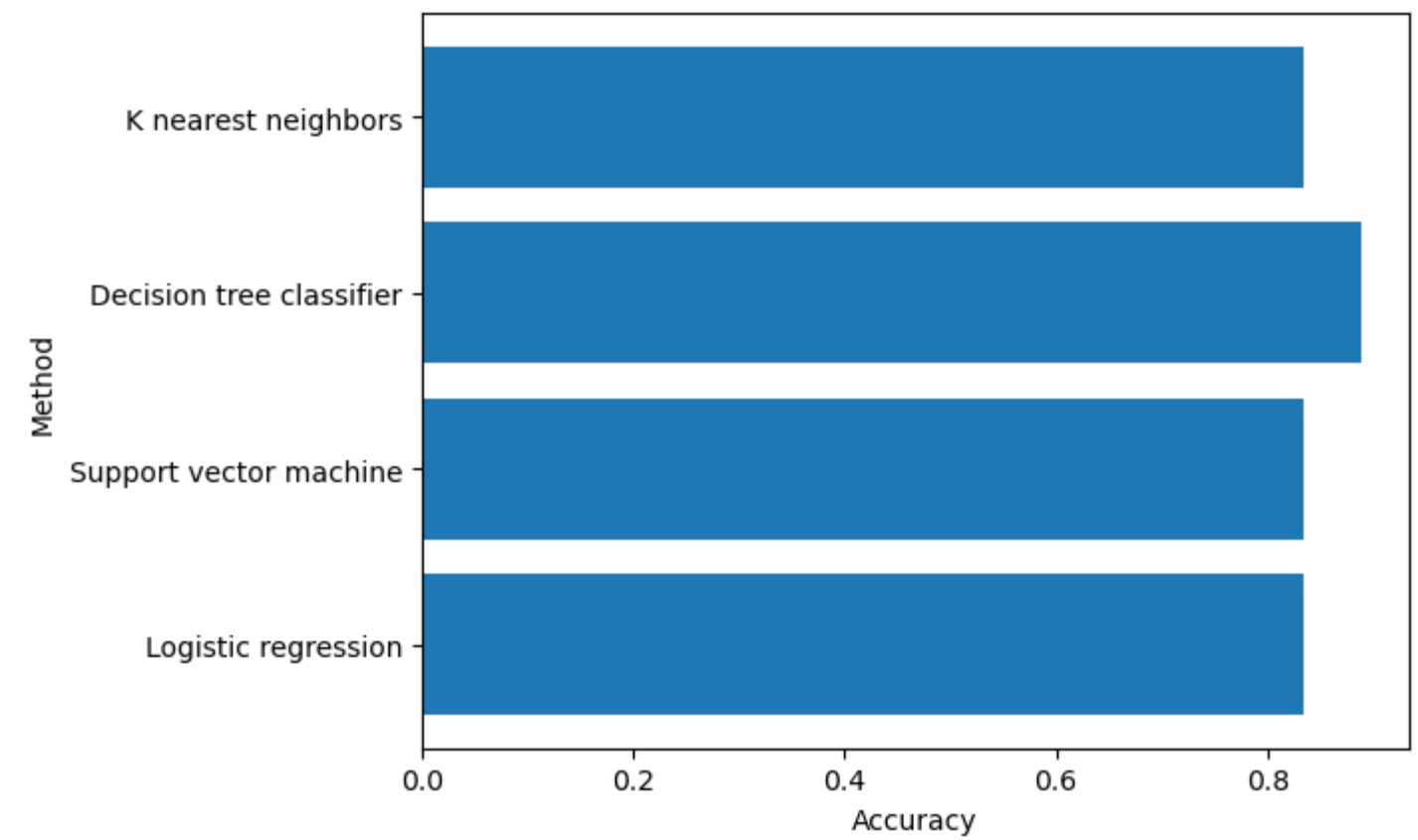
- Perform exploratory Data Analysis and determine Training Labels
  - Create columns for the class
  - Standardize the data
  - Split into Training Data and Test Data
- Find best Hyperparameter for SVM, Classification Trees, and Logistic Regression
  - Find the highest performing method based on testing data



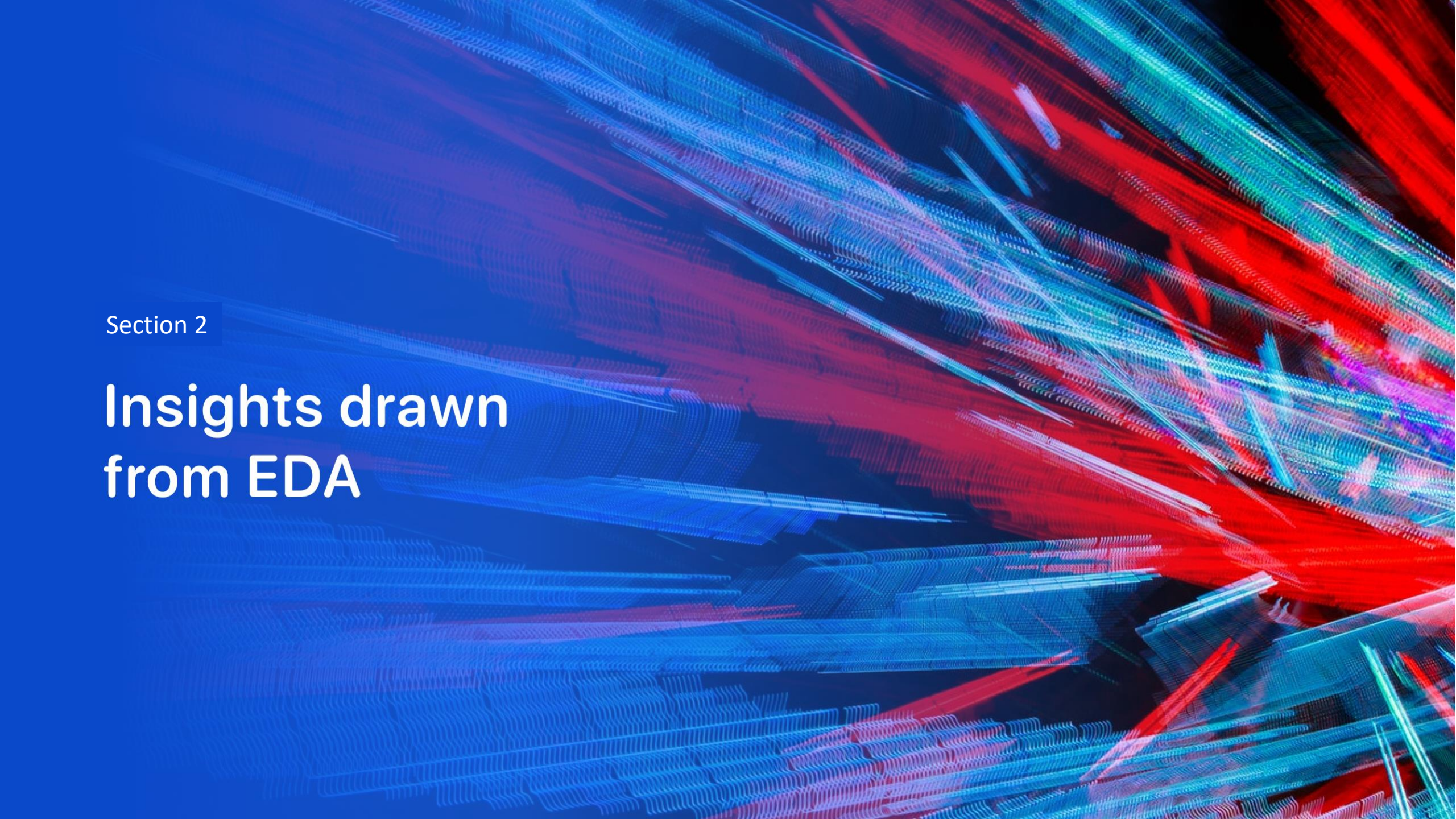
GitHub URL

Results

- Results for all models were found to be above 83%, with the Decision Tree model performing best overall.

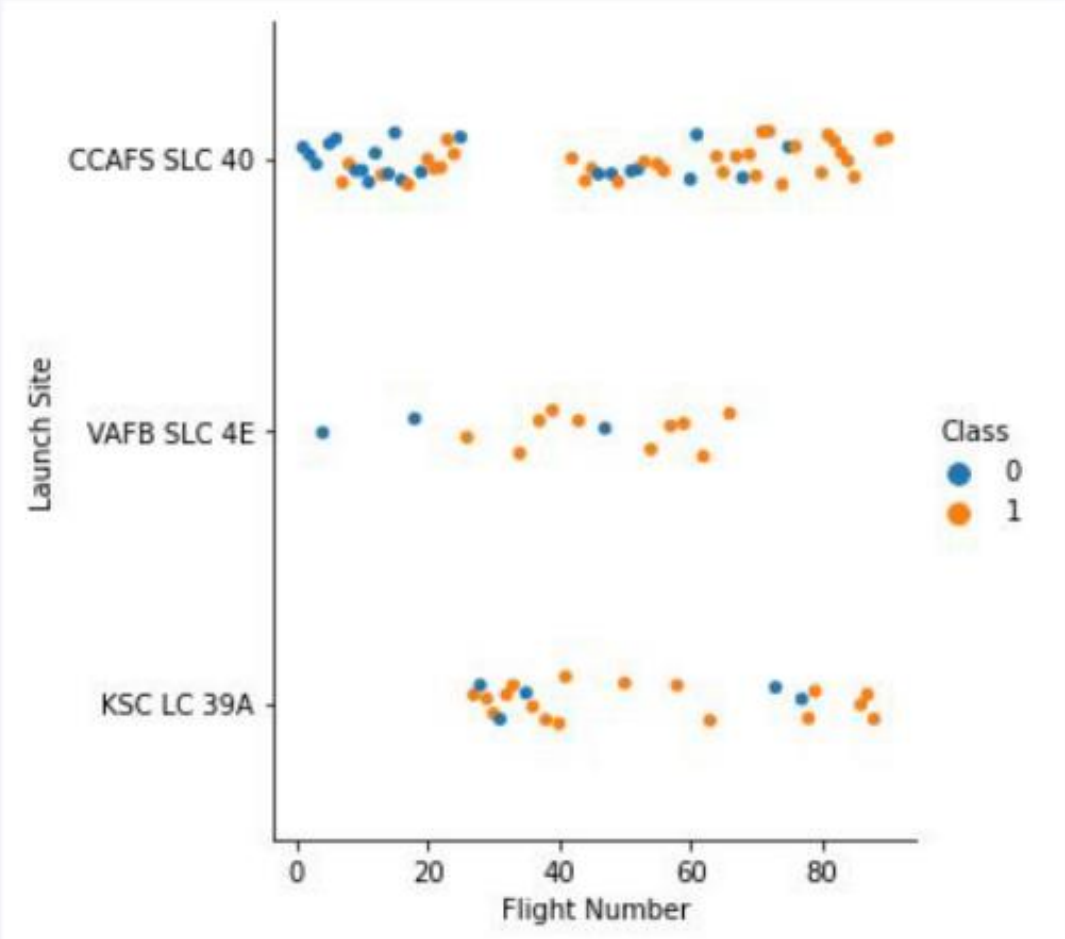


GitHub URL


Section 2
Insights drawn from EDA

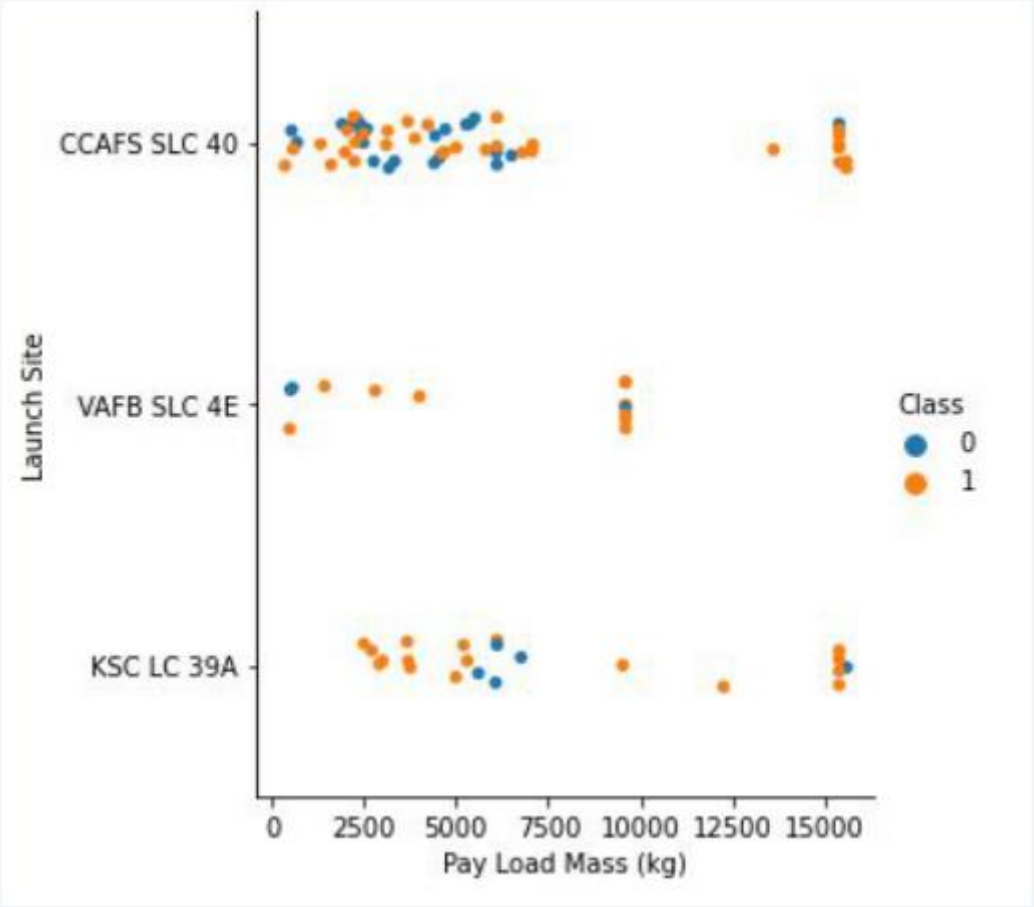
Flight Number vs. Launch Site

- Class 0 (blue) represents an unsuccessful launch, and Class 1 represents a successful launch
- Success Rates increased as the number of Number of Flights increased



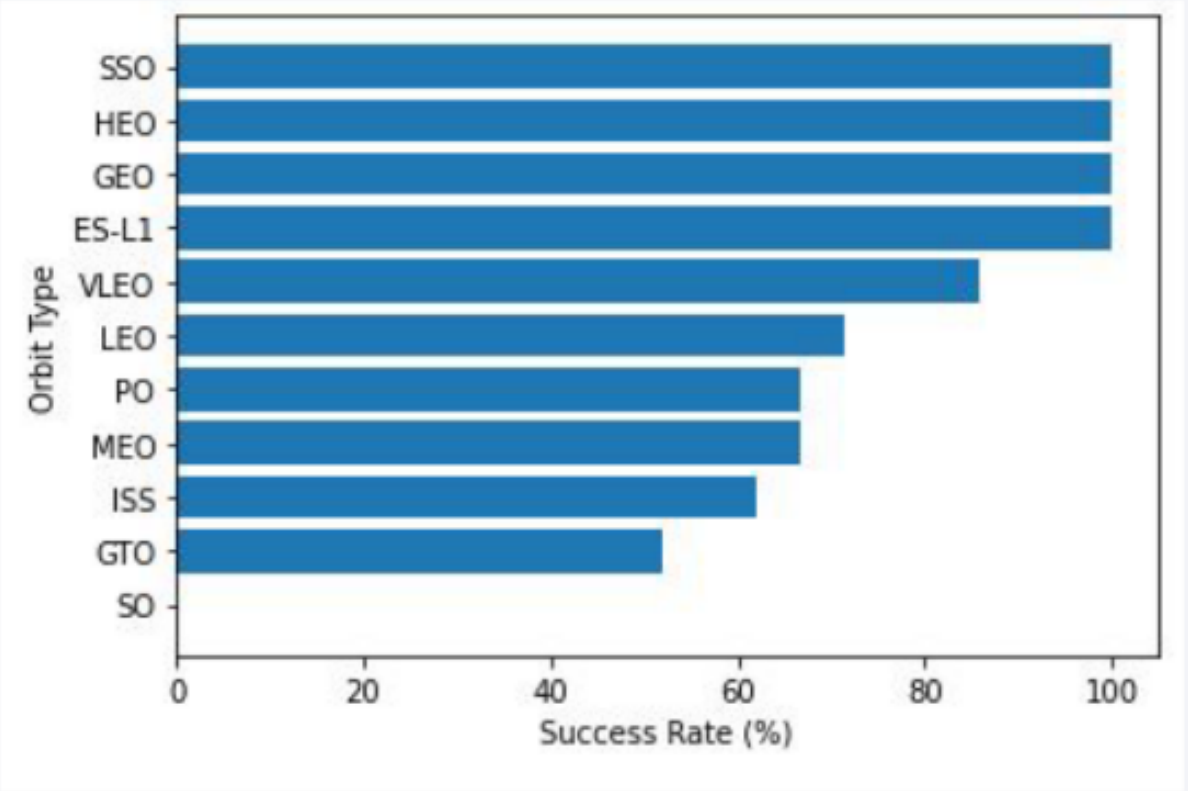
Payload vs. Launch Site

- Class 0 (blue) represents an unsuccessful launch, and Class 1 represents a successful launch
- No clear pattern can be found between Payload Mass and successful launches



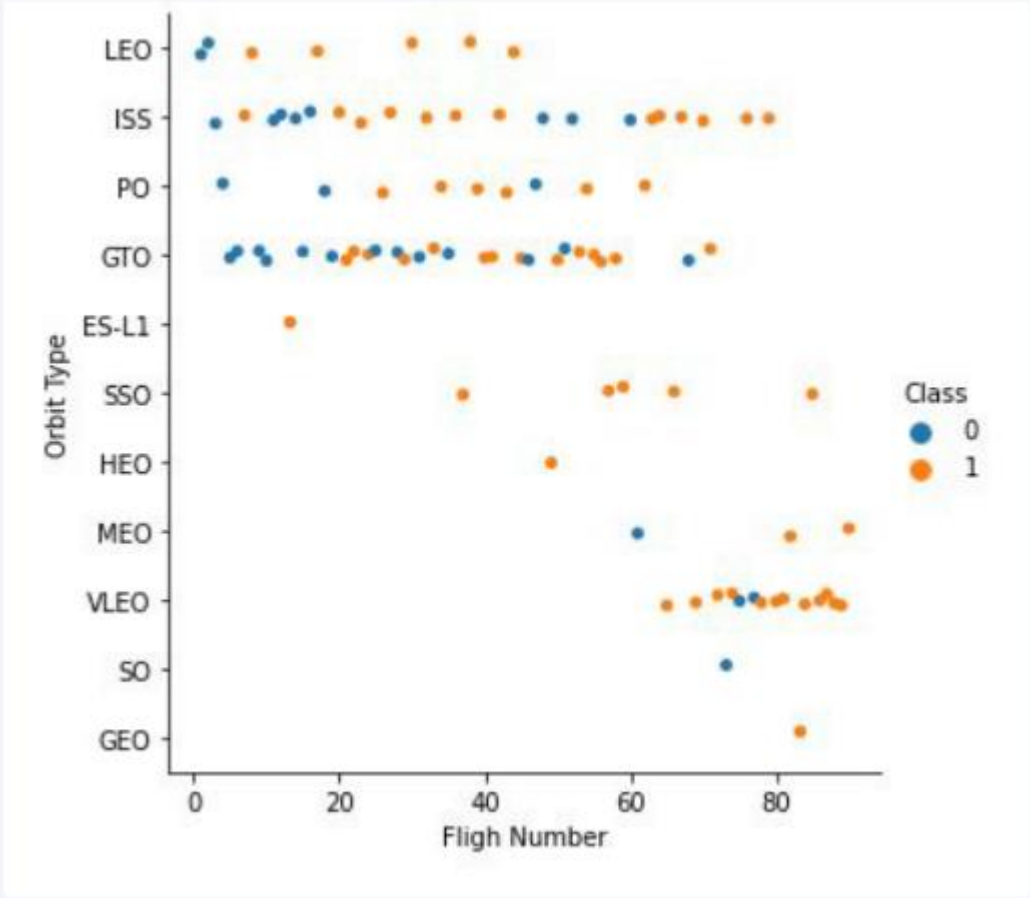
Success Rate vs. Orbit Type

- Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%)
- However, the success rate for Orbit type GTO is only 50%, and it is the lowest except for type SO, which recorded failure in a single attempt



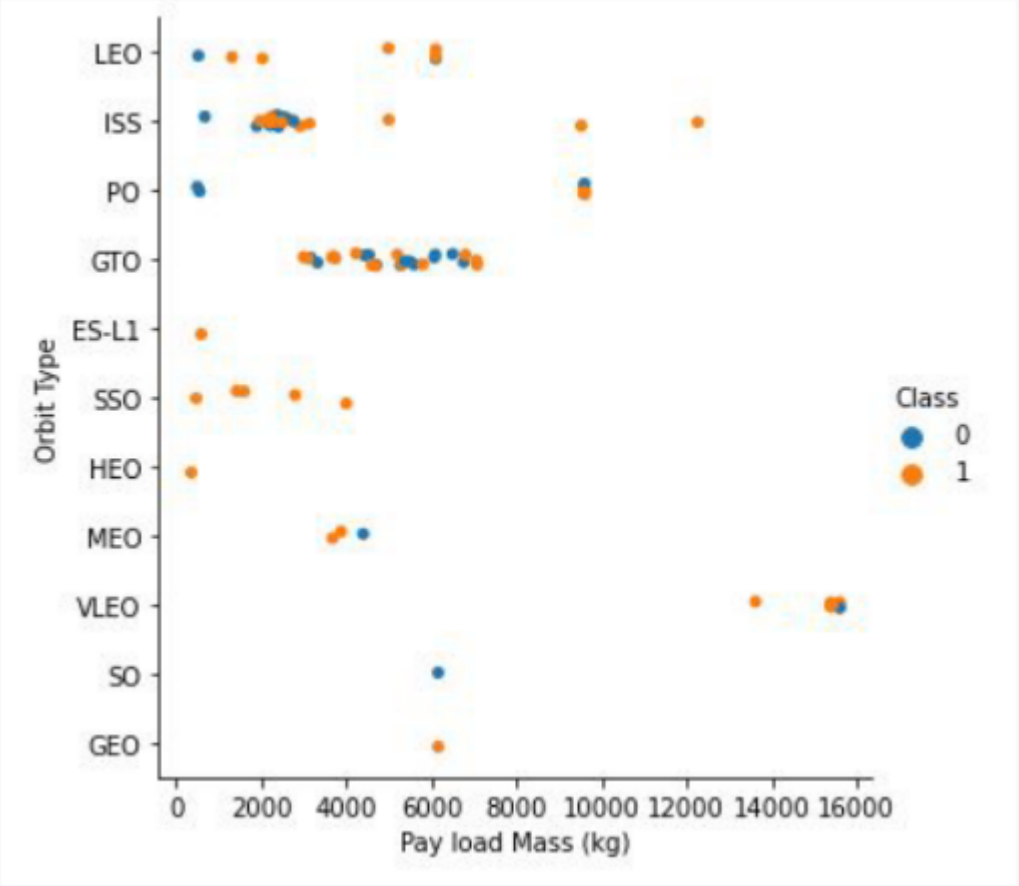
Flight Number vs. Orbit Type

- Class 0 (blue) represents an unsuccessful launch, and Class 1 represents a successful launch
- Launch success rate seems to be improving with flight number (learning curve)
- In Orbit type GTO, there seems to be no relationship between success rate and flight number



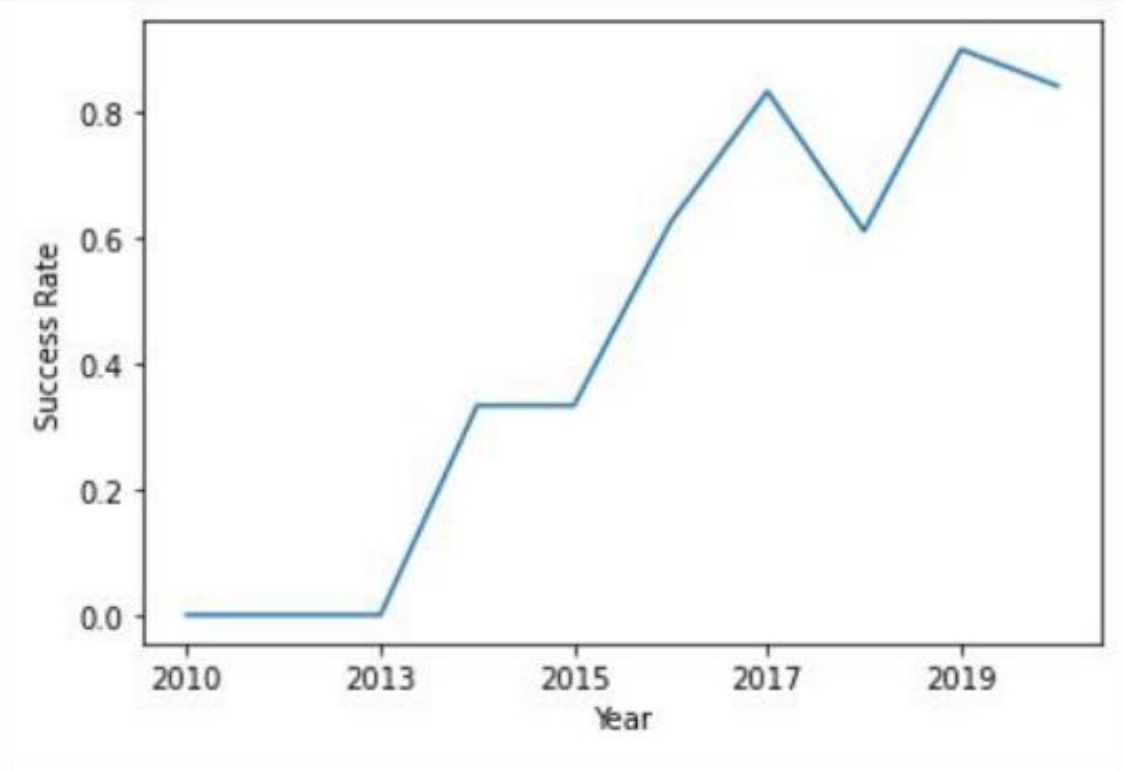
Payload vs. Orbit Type

- Class 0 (blue) represents an unsuccessful launch, and Class 1 represents a successful launch
- Successful landing rates are much improved for Orbit types LEO and ISS when Payload Mass is high
-



Launch Success Yearly Trend

- Since 2013, the successful landing rate has increased through to 2017
- The rate decreased slightly in 2018
- Recent years show a success rate of about 80%



All Launch Site Names

- Query

```
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

- Result

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

- With SQL DISTINCT used in the query, only unique values are displayed in the Launch_Site column from the SpaceX table

- There are 4 unique launch sites as shown

Launch Site Names Begin with 'CCA'

- Query

```
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

- Result

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	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

- Launch Site names starting with CCA are found with the LIKE '%' operators to look for similarities across this location

Total Payload Mass

- Query

```
SELECT SUM(PAYLOAD_MASS__KG_)
       AS total_payload_mass_kg
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

- Result

total_payload_mass_kg
45596

- Total Payload Mass of 45,596 kg is found using the SQL Sum() function for dataset records where the Customer is NASA (CRS)

Average Payload Mass by F9 v1.1

- Query

```
SELECT AVG(PAYLOAD_MASS__KG_)
       AS avg_payload_mass_kg
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

- Result

avg_payload_mass_kg
2928

- The average Payload Mass for Booster type F9 v1.1 is 2,928 kg

First Successful Ground Landing Date

- Query

```
SELECT MIN(DATE)
  AS first_successful_landing_date
FROM SPACEXTBL
WHERE LANDING__OUTCOME
      = 'Success (ground pad)'
```

- Result

first_successful_landing_date
2015-12-22

- Using the Min() function, the earliest date with a successful ground pad landing is found to be 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

- Result

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

- Several F9 booster types have had successful drone ship landings when their payload mass was between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

- Query

```
SELECT MISSION_OUTCOME,
       COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

- Result

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SpaceX has successfully completed nearly 99% of its missions
- Count() and Group By allow output of summary totals for each Mission Outcome

Boosters Carried Maximum Payload

- Query

```
SELECT DISTINCT BOOSTER_VERSION,
       PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
       SELECT MAX(PAYLOAD_MASS__KG_)
       FROM SPACEXTBL)
```

- Version F9 B5 boosters were able to carry the highest payloads

- Result

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- Query

```
SELECT LANDING__OUTCOME,
       BOOSTER_VERSION,
       LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING__OUTCOME
          = 'Failure (drone ship)'
       AND YEAR(DATE) = '2015'
```

- Result

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- In 2015 there were only two landing failures onto drone ship landing pads

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

- Query

```
SELECT LANDING__OUTCOME,
       COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

- According to the results, the number of successes and failures between mid-2010 and Q1-2017 was similar

- Result

landing__outcome	total_number
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

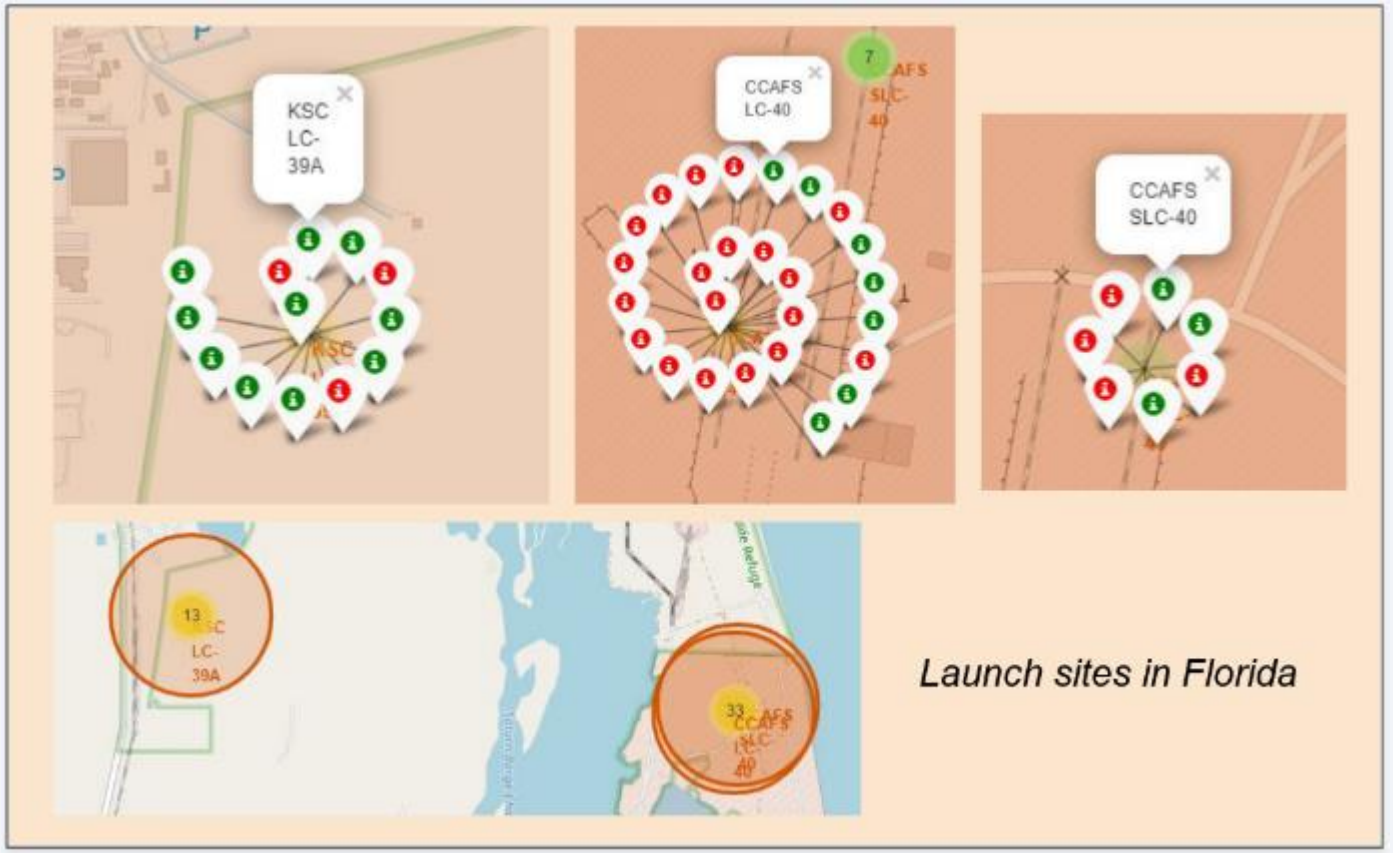
All Launch Site Locations

- All SpaceX launch sites as shown are located in coastal US cities.

Visualizing SpaceX Launch Outcomes

Launch site in California

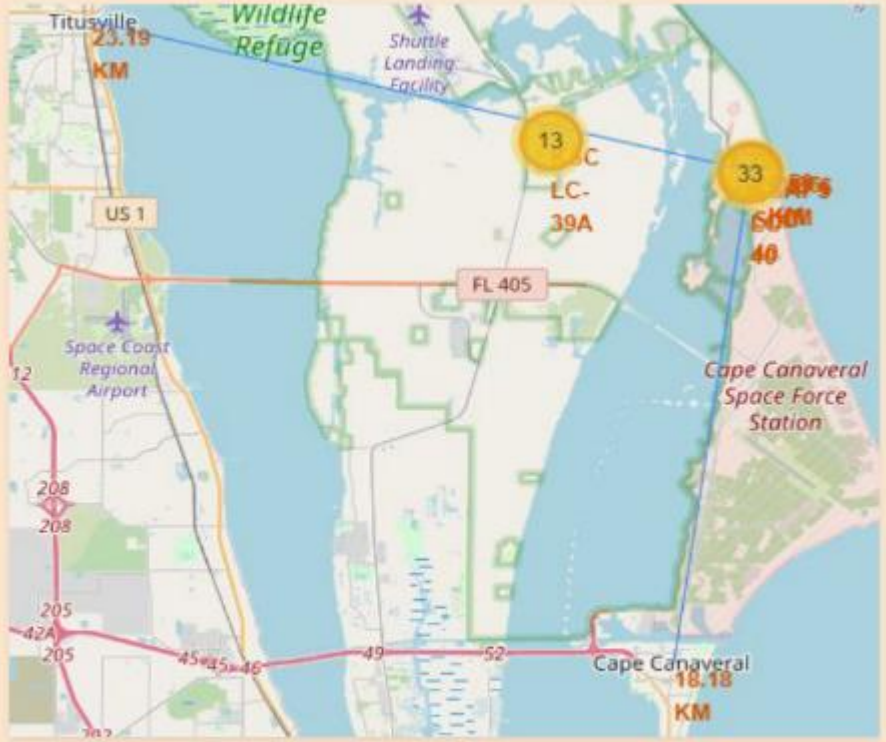
- Launch site clusters in California (left), and Florida (right)
- Successful landings (green) vs Unsuccessful landings (red)



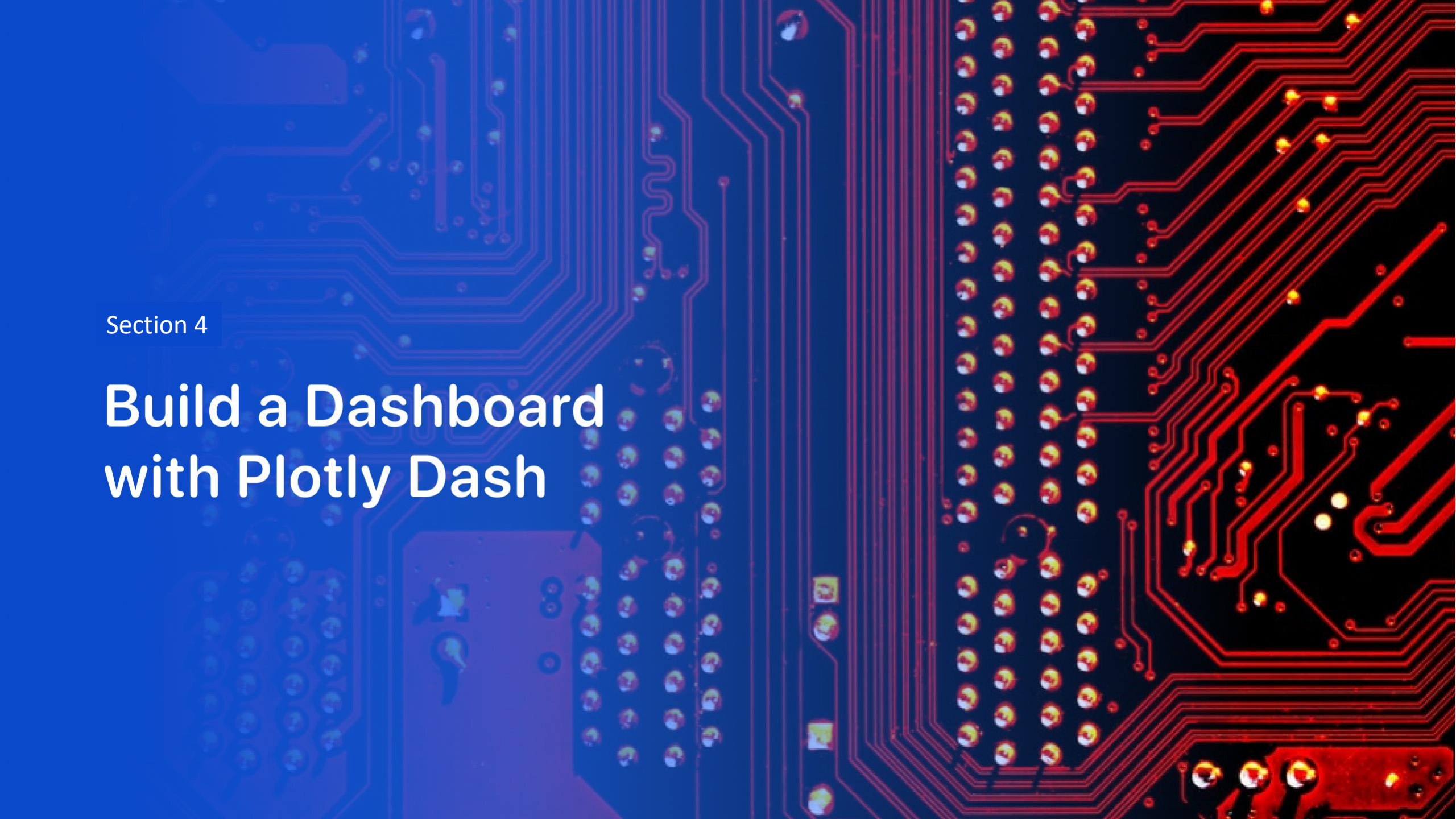
Launch sites in Florida

Proximities of SpaceX Launch Sites

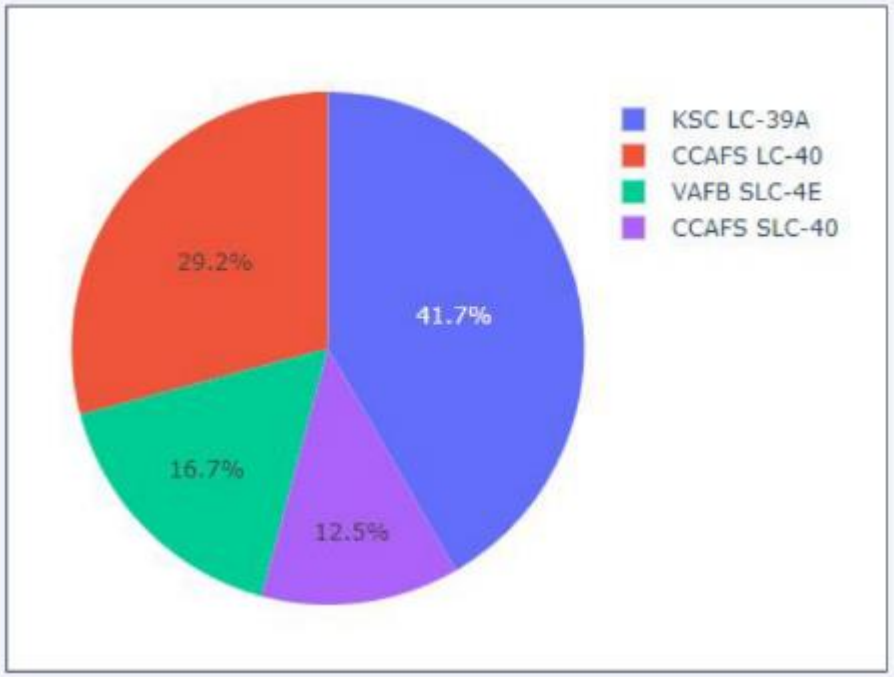
Are launch sites in close proximity to railways? ***Yes***
Are launch sites in close proximity to highways? ***Yes***
Are launch sites in close proximity to coastline? ***Yes***



Do launch sites keep certain distance away from cities? ***Yes***


Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



- KSLC-39A records the most launch successes among all launch sites
- The VAFB SLC-4E has the fewest successes, possibly because:
  - the data sample size for this site is small, or
  - because it is the only site located in California, so the launch difficulty may be higher relative to other locations

Launch Site with Highest Launch Success Ratio

- KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%)

Payload vs Launch Outcome Scatter Plot for all sites

- The launch success rate (class 1) for low weighted payloads (<5000 kg) is higher than that of heavy weighted payloads (5000-10000 kg)

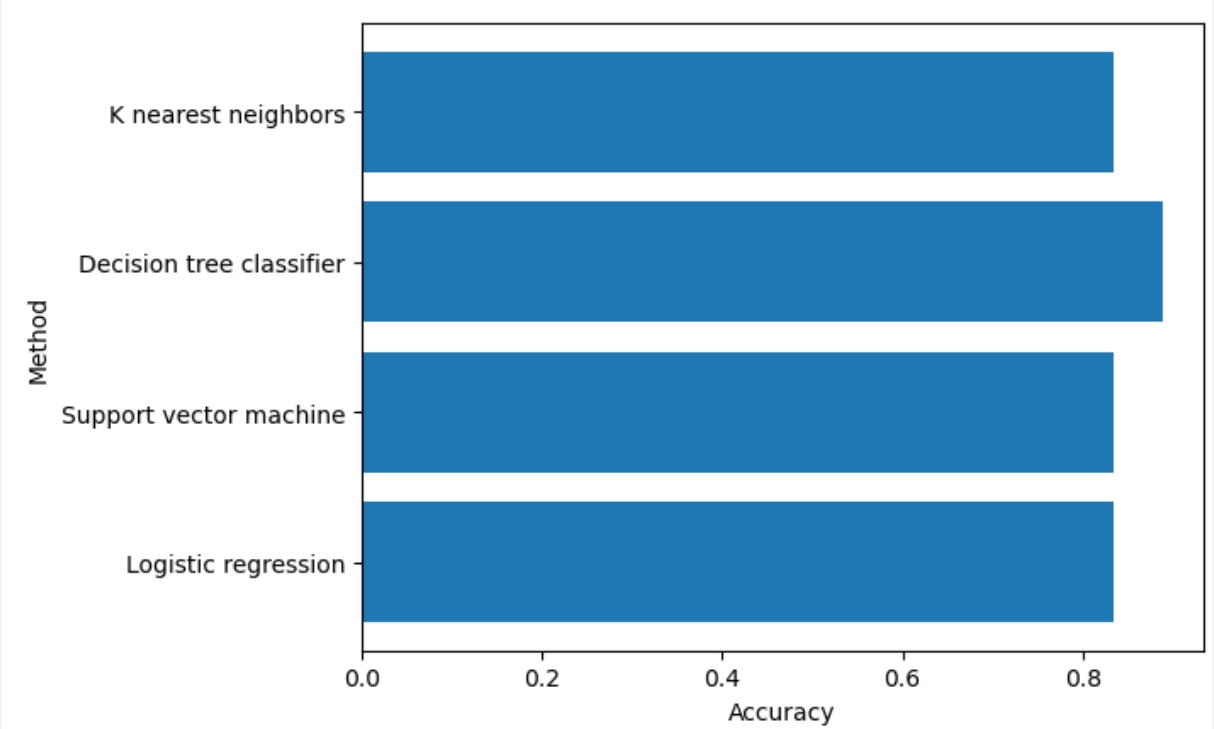
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- From the test dataset, the model with the highest accuracy was the Decision Tree Classifier model.
- To note, the test size was relatively small at 18, and to find an optimal model more data may be needed.

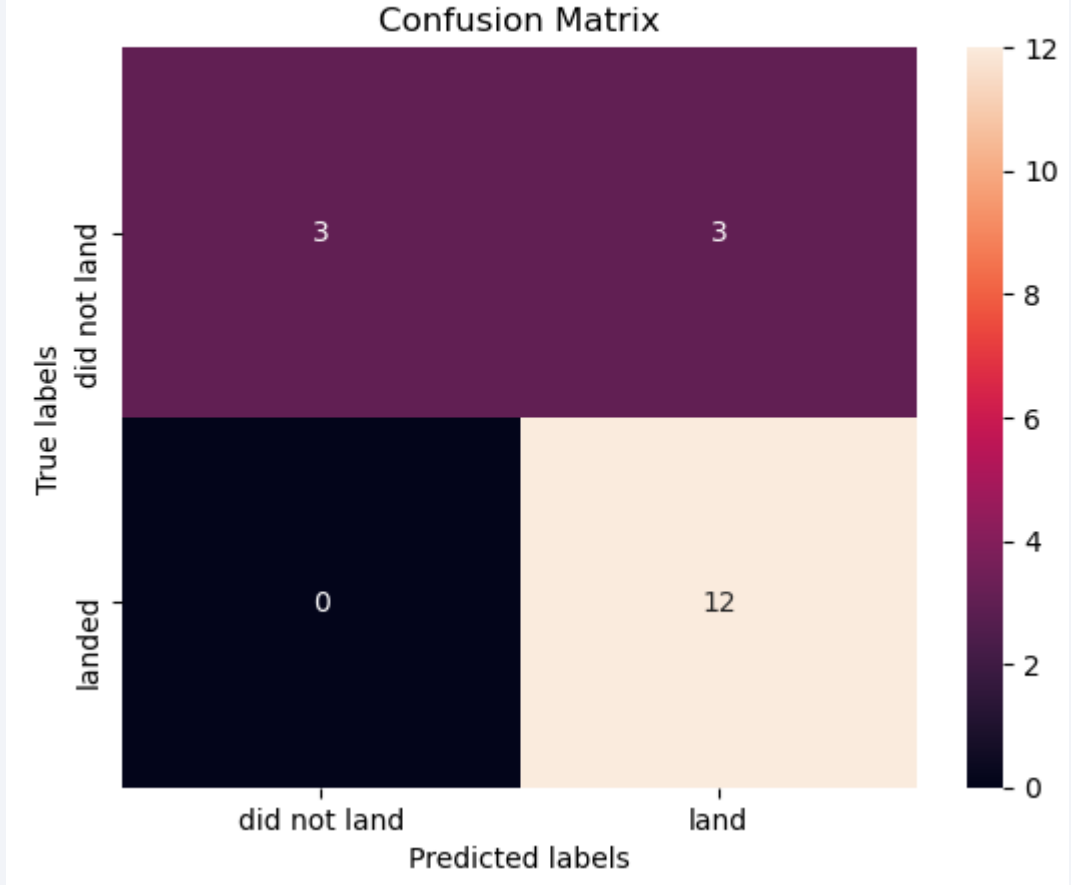
```
plt.barh(methods, accuracy)
plt.xlabel('Accuracy')
plt.ylabel('Method')
plt.show()
```



Confusion Matrix

- The Decision Tree model predicted 12 successful landings when the correct label was successful, and predicted 3 failed landings when the correct label was failure.
- There were also 3 predictions for successful landings where the correct label was failure (false positive).
- Overall, the model predicts successful landings at a rate of 88.8%

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- As the number of flights increased, the success rate increased to a recent high of over 80%
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%)
- The launch site is close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes, and also the highest success rate among all sites.
- The launch success rate of low-weighted payloads is significantly higher than that of heavy weighted payloads.

Appendix

- Coursera Applied Data Science Capstone Course URL

Thank you!