Winning Space Race with Data Science

Anna Bortle
October 24, 2023

Outline

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

Executive Summary

- Methodologies
  - Gain insight on previous launches by SpaceX with the Falcon 9 rocket
    - What contributes to a successful launch? What contributes to a successful landing?
  - Utilize publicly available launch data to determine key factors
    - SpaceX API
    - Webscraping
- Results
  - A decision tree classification model will best determine the landing outcome of the launch, with determining factors including launch site, orbit type, payload type, and booster version.

Introduction

- Background
  - This project is aimed to determine what the relative cost of a launch will be by determining if the first stage of the rocket will be reused (successful landing outcome.)
    - Cost of launch is significantly cheaper if the first rocket stage can be successfully landed and reused.
    - Many factors determine if a landing with be successful or even attempted.
    - Goal is to determine when given a set of parameters for launch (payload, launch site, etc.), whether the cost of the launch will be expensive (first stage does not land) or relatively cheap (first stage successfully lands to be reused.)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX previous launch data collected via SpaceX API and web scraping
- Perform data wrangling
  - Refined to only include data related to Falcon 9 launches
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various classification models were trained and tested on the data to determine best model for accuracy

Data Collection

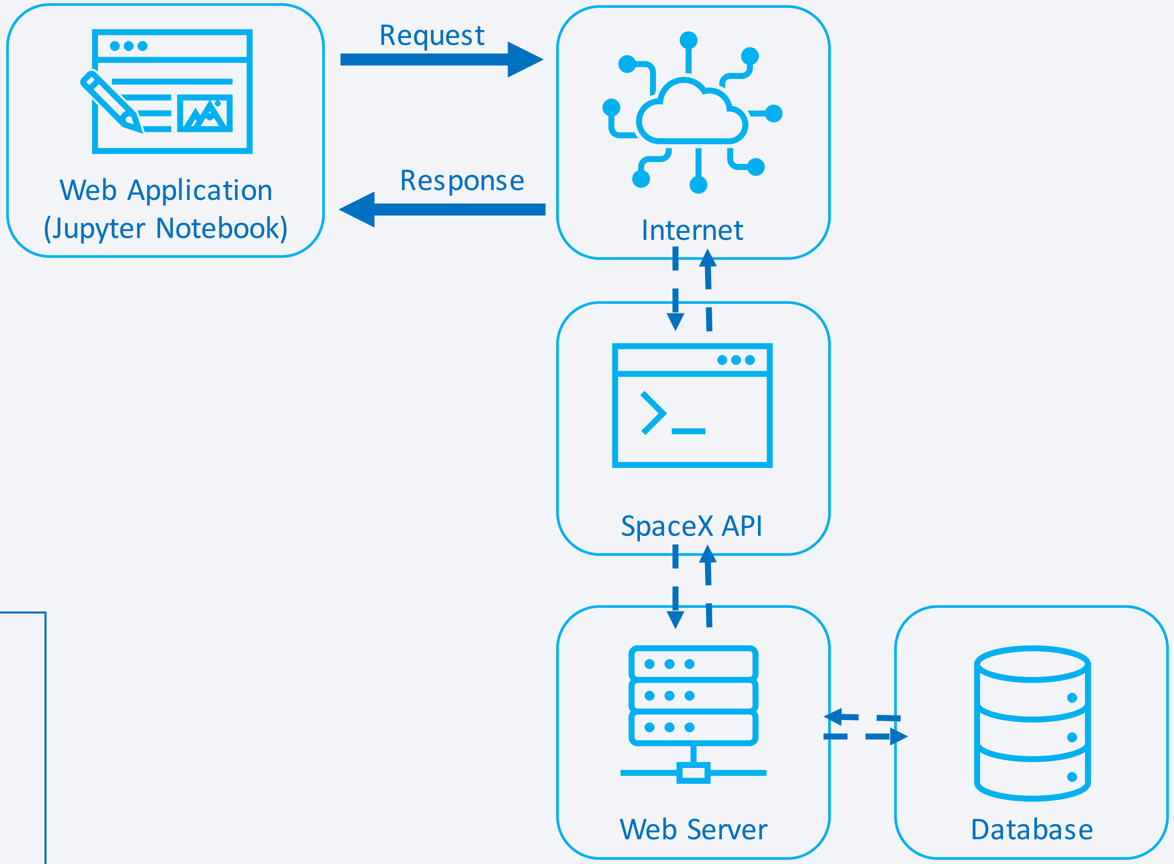
- Data was collected from publicly available launch data from previous SpaceX launches
  - SpaceX API – past launches
  - Webscraping – Falcon 9 launches Wikipedia page
- Data was received in either a .json or .html format, cleaned, and formatted into pandas dataframes to allow for easier analysis
  - Cleaned data to only include information related to Falcon 9 launches

Data Collection – SpaceX API

- Previous SpaceX launch data collected via SpaceX API
    - Received through GET request
    - Stored in pandas dataframe
    - Cleaned and wrangled to extract relevant data for Falcon 9 launches

- SpaceX API Jupyter Notebook

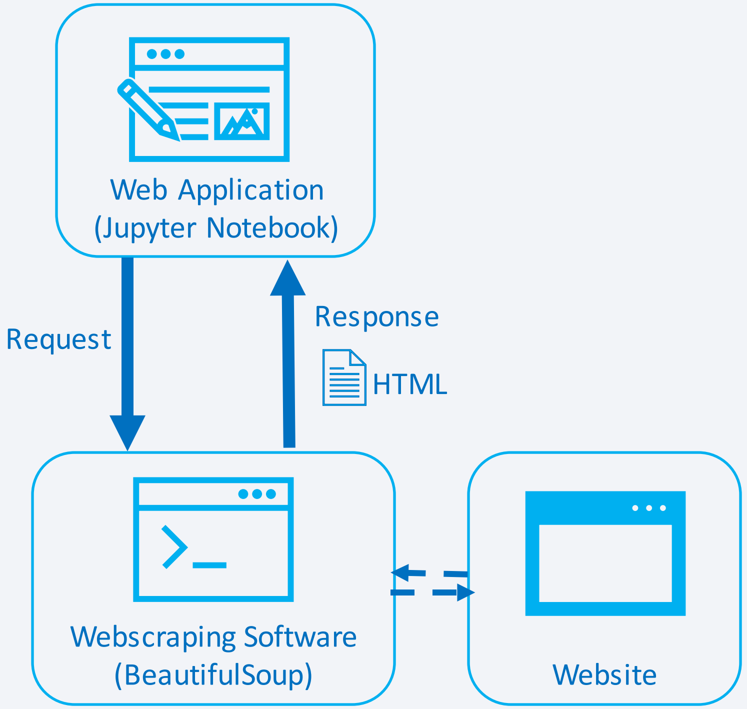
```
import requests
import pandas as pd

spacex_url =
"https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
data = pd.json_normalize(response)
```



Data Collection - Scraping

- Previous launch data collected from Falcon 9 Launches Wikipedia webpage via BeautifulSoup
    - Received through GET request
    - BeautifulSoup object created from response
    - Launch tables parsed from html content and stored in dictionary
    - pandas dataframe created from dictionary

- [Web Scraping Jupyter Notebook]




```
import requests
from bs4 import BeautifulSoup

wiki_url =
"https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches"
response = requests.get(wiki_url).content
soup = BeautifulSoup(response)
```

Data Wrangling

- After data was reformatted and cleaned to contain only that which related to the Falcon 9 launches, additional work was needed to better handle the landing outcomes
  - Outcomes were descriptive labels that indicated the result of the landing and possible additional information on landing platform
    - For example: True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
  - These outcomes were recast as 0 (failed landing) or 1 (successful landing) for the training models

- Data Wrangling Jupyter Notebook

EDA with Data Visualization

- Seaborn plotting library was used to visualize data
  - Able to plot two variables along with the "class" as our hue to see what combination of the two variables have higher or lower success rates
  - Critical variables: orbit type, payload mass, launch site, flight number

- Data Visualization Jupyter Notebook

EDA with SQL

- Queries
  - Unique launch sites
  - Records at launch sites beginning with "CCA"
  - Total payload mass carried by flights for NASA
  - Average payload mass carried by F9 v1.1
  - Date of first successful ground landing
  - Booster types of successful drone landings with payloads between 4k and 6k kg
  - Total numbers for failed and successful mission outcomes
  - Boosters that have carried the maximum payload
  - Failed drone landings in 2015
  - Ranked landing outcomes between 2010-06-04 and 2017-03-20
- SQL Jupyter Notebook

Build an Interactive Map with Folium

- The folium library was utilized to visualize the locations of the launch sites on a world map to draw any insights on these locations
  - Each launch site was marked
  - Green/red marker clusters were added at each site to display successful/failed landings
  - Distances were drawn to local proximites (railways, coastlines, cities, etc.) to gain more insight on the location choices for these launch sites
- [Folium Jupyter Notebook]()

Build a Dashboard with Plotly Dash

- An interactive dashboard was created with a pie chart depicting landing success rates at each and all of the launch sites, as well as a scatter plot with an interactive range slider depicting payload mass vs "class" (successful landing) and booster version.
    - Allowed for easy determination of launch sites with greater landing success rates as well as how the payload and booster version of the launches impacted the success or failure of the landing attempts

- Plotly Dash lab

Predictive Analysis (Classification)

- 4 different classification models (Logistical Regression, Support Vector Machine, K Nearest Neighbors, Decision Tree) were built to test for accuracy with the data
  - Data was split into training and testing subsets
  - Each model was trained on the same training data and tested for accuracy on the testing data
  - Confusion matrices were constructed for each model

- Predictive Analysis Jupyter Notebook

Results

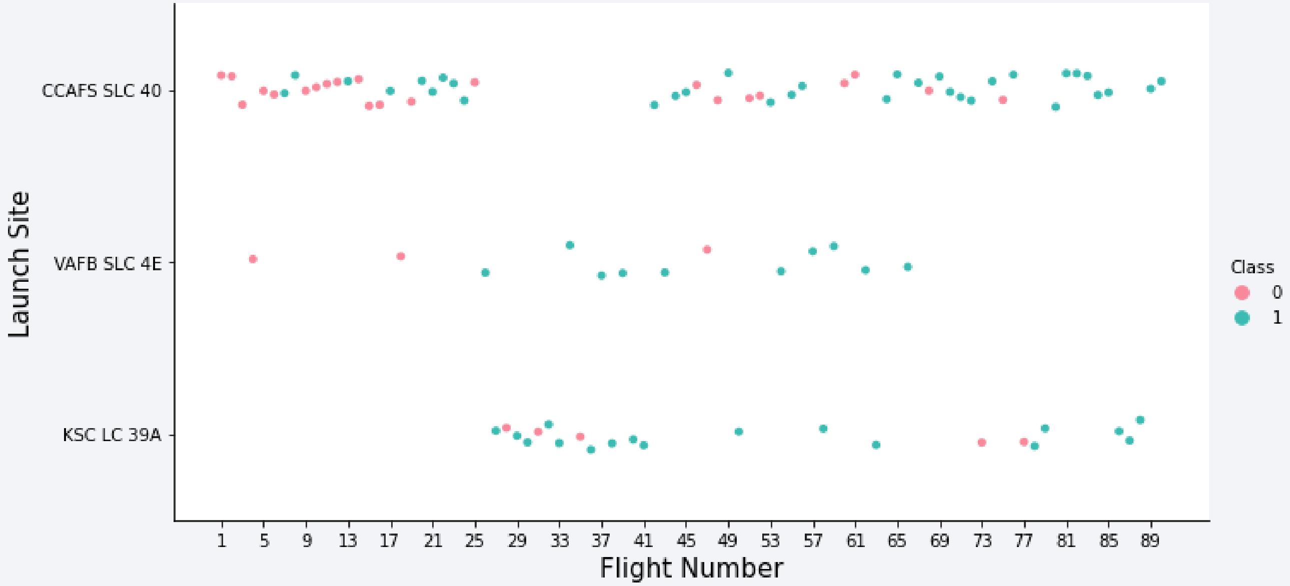
- Launch site, payload mass, orbit type, and booster version have large impacts on whether or not flights will have successful landings
  - Explored in next set of slides
  - Interactive plotly dashboard and folium maps will make it easier to visualize these dependencies
- A decision tree classification model is the most accurate for the determination of whether or not the launch landing will be successful

Section 2

Insights drawn from EDA

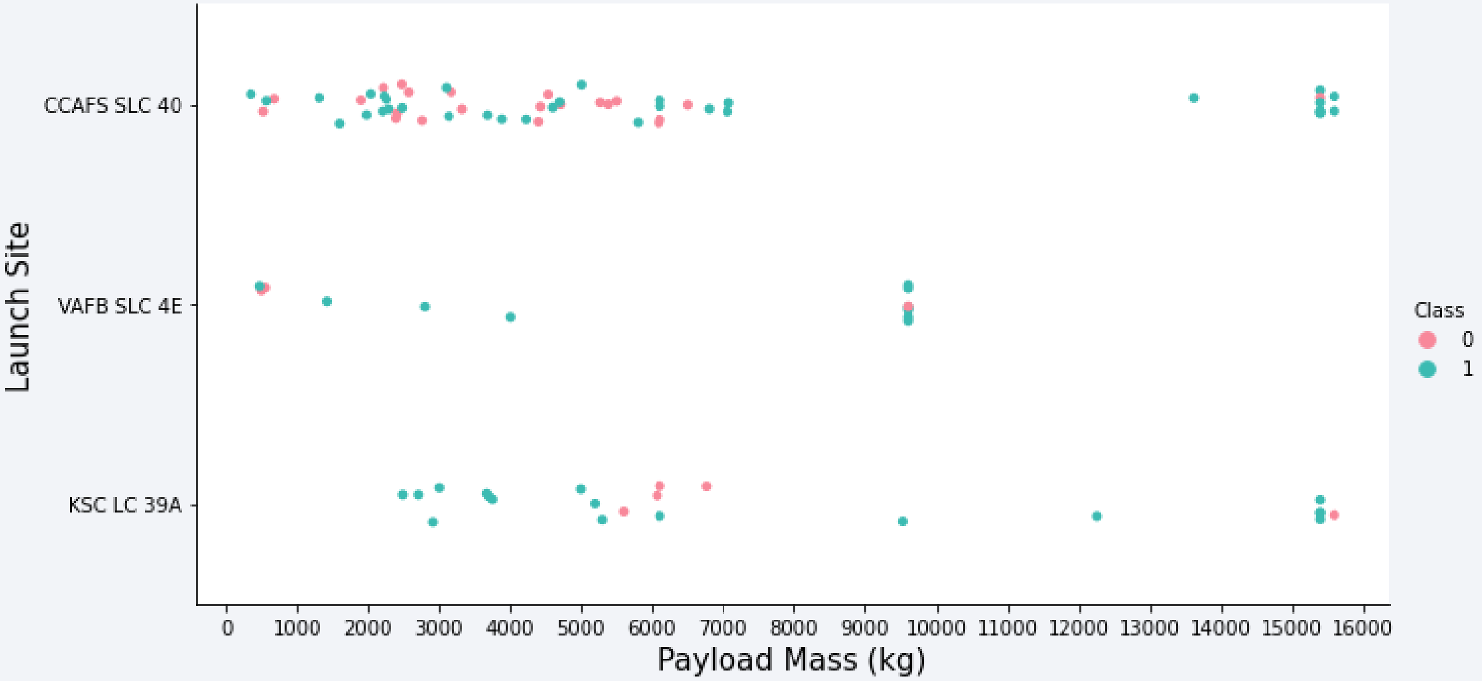
Flight Number vs. Launch Site

- VAFB launch site has been used the least and hasn't been used since Flight Number 67
- A clear break at CCAFS launch site occurred between flight numbers 27-42
    - During which had the first use of KSC site almost exclusively
- CCAFS remains the most common launch site



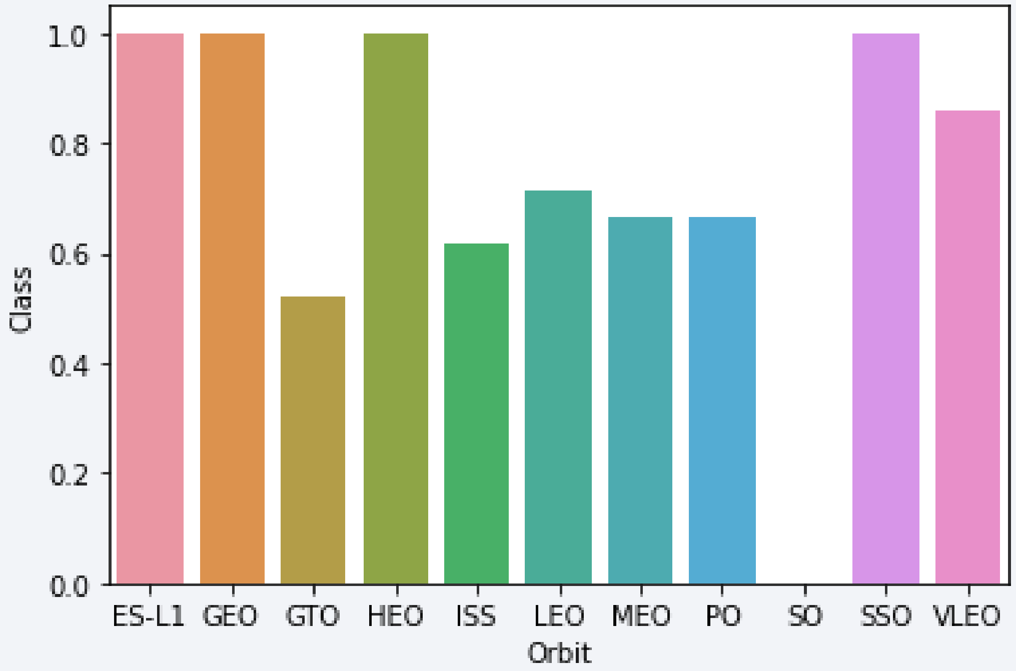
Payload vs. Launch Site

- Flights with payloads >8'000kg have high success rate
- Flights <8'000kg are least successful at CCAFS SLC 40 launch site
- No flights with payloads >10'000kg have launched at VAFB SLC 4E
- No flights with payloads <~2'500kg have launched at KSC LC 39A



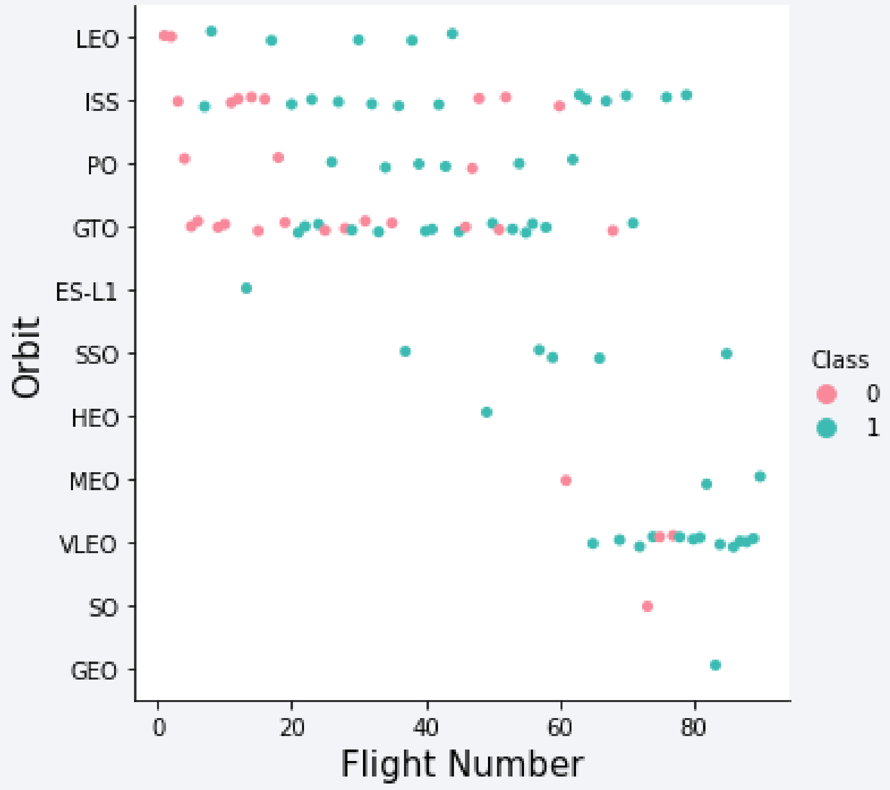
Success Rate vs. Orbit Type

- Orbits with 100% landing success rate: ES-L1, GEO, HEO, SSO
- Success rate descending order:
    - ES-L1, GEO, HEO, SSO
    - VLEO
    - LEO
    - MEO, PO
    - ISS
    - GTO
    - SO



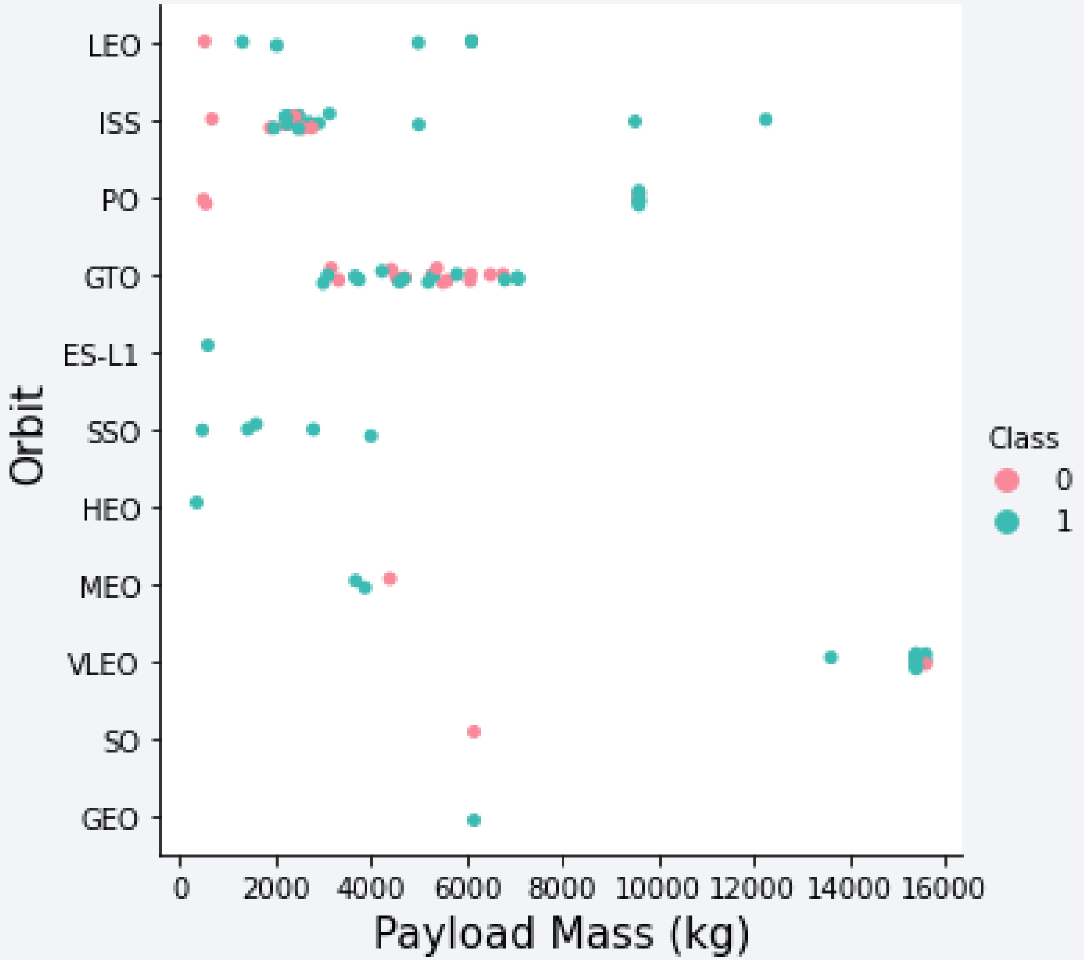
Flight Number vs. Orbit Type

- Orbit types LEO, ISS, PO, and GTO have been the most common types since the beginning of the launches
- As more flights occurred, different orbit types were sporadically added into the mix (bottom 2/3 orbit types on chart)
- The VLEO orbit type has been the most common in recent flights and has a high landing success rate



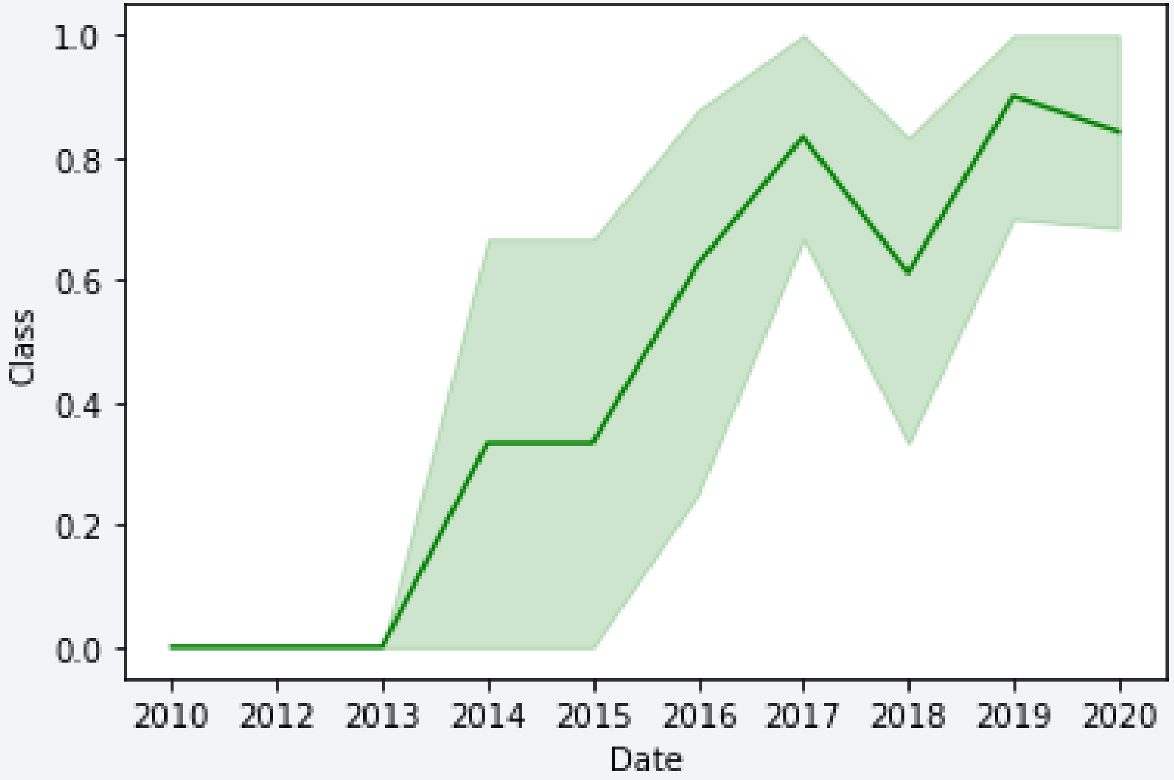
Payload vs. Orbit Type

- Launches with orbit types ES-L1, SSO, GEO, and HEO have seen 100% success rate in landing the first stage booster
- Launches with payloads <1000 kg have been successful with ES-L1, SSO, and HEO orbit types but unsuccessful with orbit types of LEO, ISS, and PO
- Orbit type GTO has ~50% success rate
- No successful landings have been achieved with SO orbit type



Launch Success Yearly Trend

- The average yearly launch success rate for the Falcon 9 launches has had an overall increase
  - Likely due to more experience with launches
  - Dips in the general increase could be due to trying new payloads/booster versions/etc.



All Launch Site Names

- There are **4*** unique launch sites in the working data
  - This was queried in SQL using the distinct() function on the Launch_Site field

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

Launch Site Acronym	Name
CCAFS LC-40	Cape Canaveral *[Space]* Launch Complex 40
CCAFS SLC-40	Cape Canaveral Space Launch Complex 40
VAFB SLC-4E	Vandenberg AFB Space Launch Complex 4
KSC LC-39A	Kennedy Space Center Launch Complex 39A

Launch_Site

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

Query result

Cape Canaveral Launch Complex 40 was renamed to Cape Canaveral* ***Space* *Launch Complex 40. They correspond to the same location.*

Launch Site Names Begin with 'CCA'

- Below is an example of 5 records where launch sites begin with `CCA`
  - Query is found using the SQL like() keyword and % wildcard character to match records with Launch_Site beginning with "CCA..."

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__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

Total Payload Mass

- The total payload mass carried by launches for NASA (CRS) amounts to **45'596 kg** (~100'522 lbs / ~50 US t)
  - Total was found through use of sum() function on records where the Customer field was 'NASA (CRS)'

Total_Payload_Mass_(kg)
45596

Query result

```
%sql select sum(PAYLOAD_MASS__KG_) as 'Total_Payload_Mass_(kg)'
from SPACEXTABLE where (Customer = 'NASA (CRS)')
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 across the entire dataset was **2'928.4 kg** (~ 6456 lbs)
    - Average was found through use of avg() function on records where the Booster_Version field was 'F9 v1.1'

Average_Payload_Mass_(kg)
2928.4

Query result

```
%sql select avg(PAYLOAD_MASS__KG_) as 'Average_Payload_Mass_(kg)' from
SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

First Successful Ground Landing Date

- The first recorded successful landing on a ground pad for the Falcon 9 launches occurred on **December 22, 2015**
  - Queried using the min() function on the Date field of records which had a Landing_Outcome of 'Success (ground pad)'

```
%sql select min(Date) from SPACEXTABLE
where Landing_Outcome = 'Success (ground pad)'
```

min(Date)
2015-12-22

Query result

Successful Drone Ship Landing with Payload between 4000 and 6000

- To the right are the 4 names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg
  - This query utilizes the distinct() function and the 'and' operand to match records that have both a Landing_Outcome of 'Success (drone ship)' and a PAYLOAD_MASS__KG in the range of 4000 and 6000 kg

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

Query result

```
%sql select distinct(Booster_Version) from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
```

Total Number of Successful and Failed Mission Outcomes

- To the right is a table showing the total number of successful and failed mission outcomes
  - The 'Success' on the third row is singular because of a space character after 'Success' in that record
  - The table was created utilizing the 'group by' keyphrase

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

Query result

```
%sql select Mission_Outcome, count(Mission_Outcome) as Count from SPACEXTABLE
group by Mission_Outcome
```

Boosters Carried Maximum Payload

- To the right is a list of 12 booster versions that have carried the maximum payload in the dataset
  - Query utilizes a subquery to obtain the max() record of PAYLOAD_MASS__KG_
  - The distinct() function is used to generate the list of booster versions whose records have a payload equal to the max obtained from the subquery

```
%sql select distinct(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

Query result

2015 Launch Records

- There are 2 records of failed landing outcomes in 2015 with a drone ship
  - Both occurred at the CCAFS LC-40 launch site, in the months of April and October
  - The query utilizes the substr() function to extract the months and years from the Date field in the data set

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

```
%sql select substr(Date,6,2) as
'Month_(2015)',Landing_Outcome,Booster_Version,Launch_Site from SPACEXTABLE
where substr(Date,0,5) = '2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

- To the right is a rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20
    - Query was obtained using a combination of the between, 'group by', and 'order by' keywords and phrases.
    - About a third of the launches during this time period made no attempt to land the first stage
    - There are no failed landings at a ground pad

```
%sql select Landing_Outcome, count(Landing_Outcome) as Count
from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by Count desc
```

Landing_Outcome	Count
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

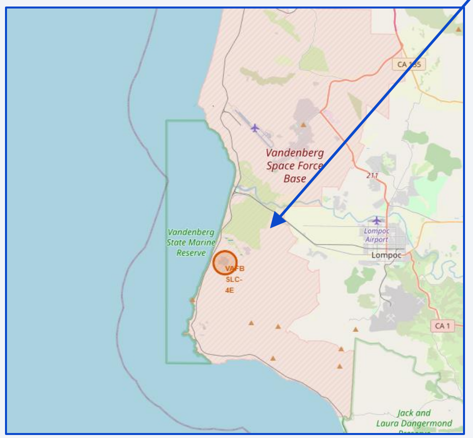
Query result

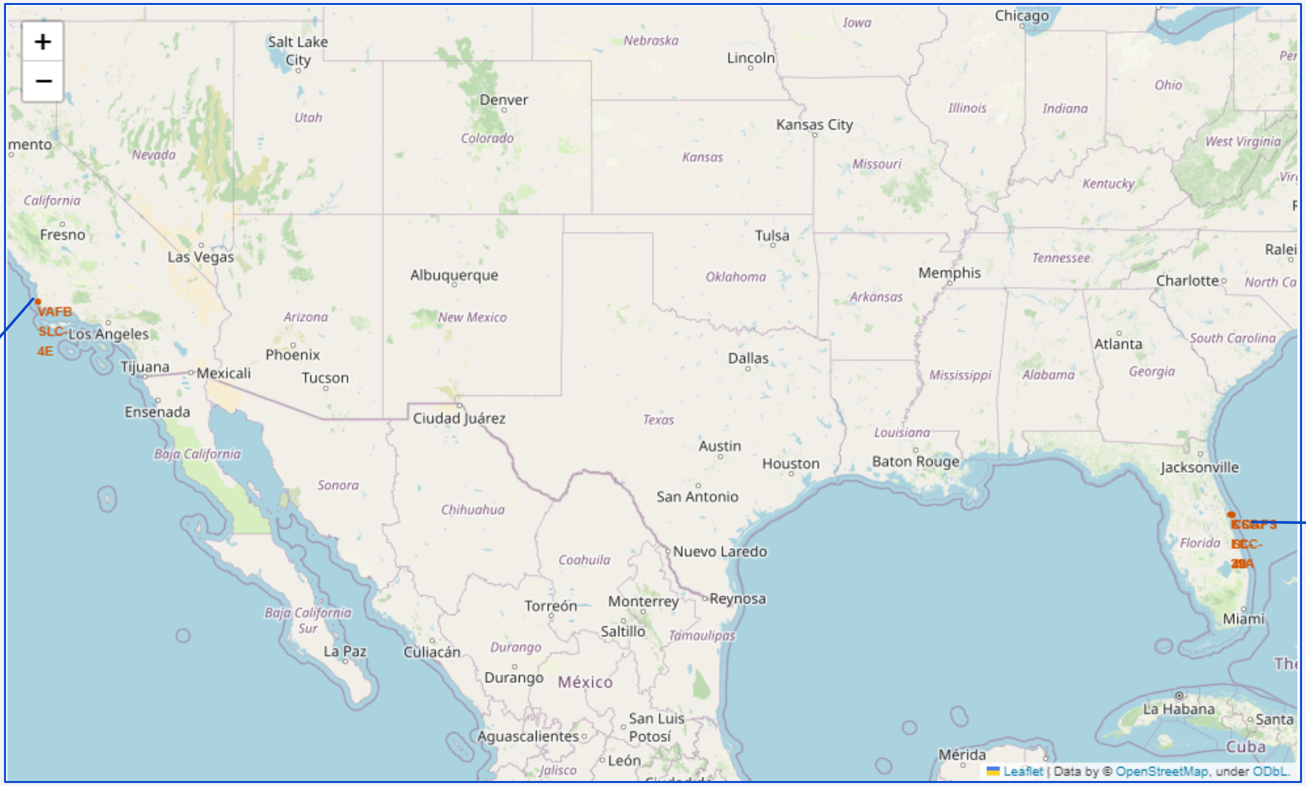
Section 3

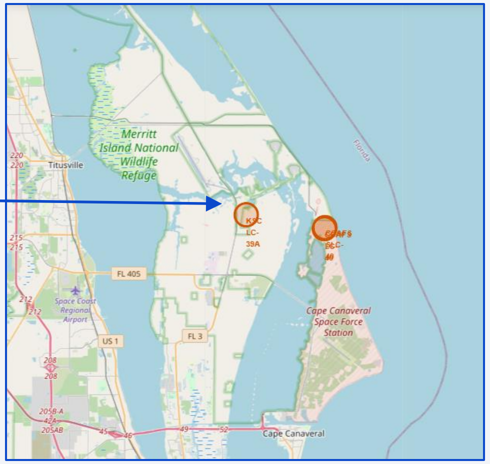
Launch Sites Proximities Analysis

Location of Launch Sites

- Launch sites are located on the coasts of southern states Florida and California
  - As these sites are closer towards the equator, the launches can use the greater speed of the ground's orbit compared to the Earth's center



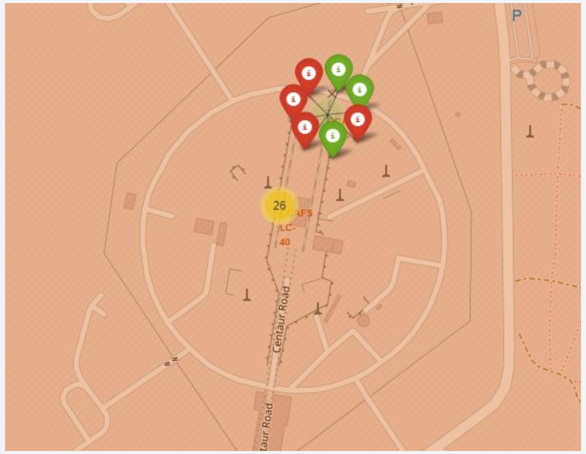




Launch Outcomes on the Map

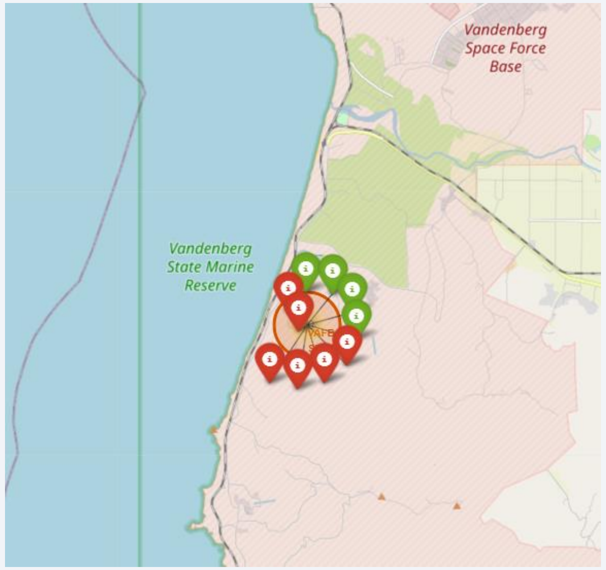
- Putting the launch outcomes on the map helps visualize the success/fail rate of each of the launch sites. For example, We can see that the Kennedy Space Center's launch site (KSC LC-39A) has the highest successful launch outcomes rate

CCAFS LC-40

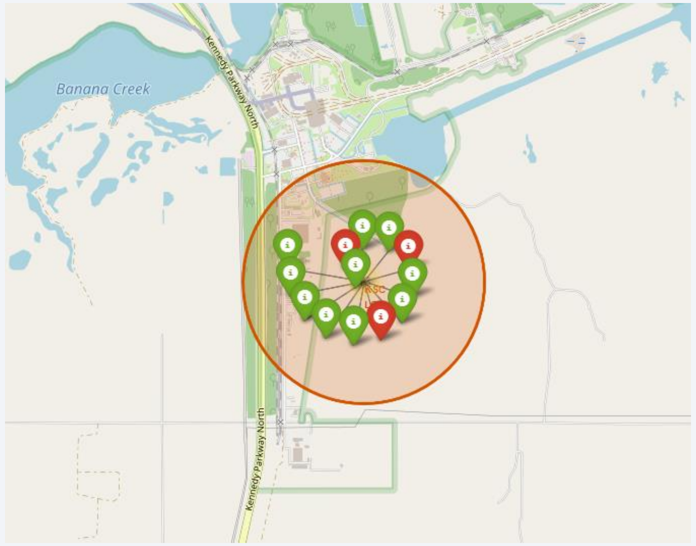


CCAFS SLC-40

VAFB SLC-4E



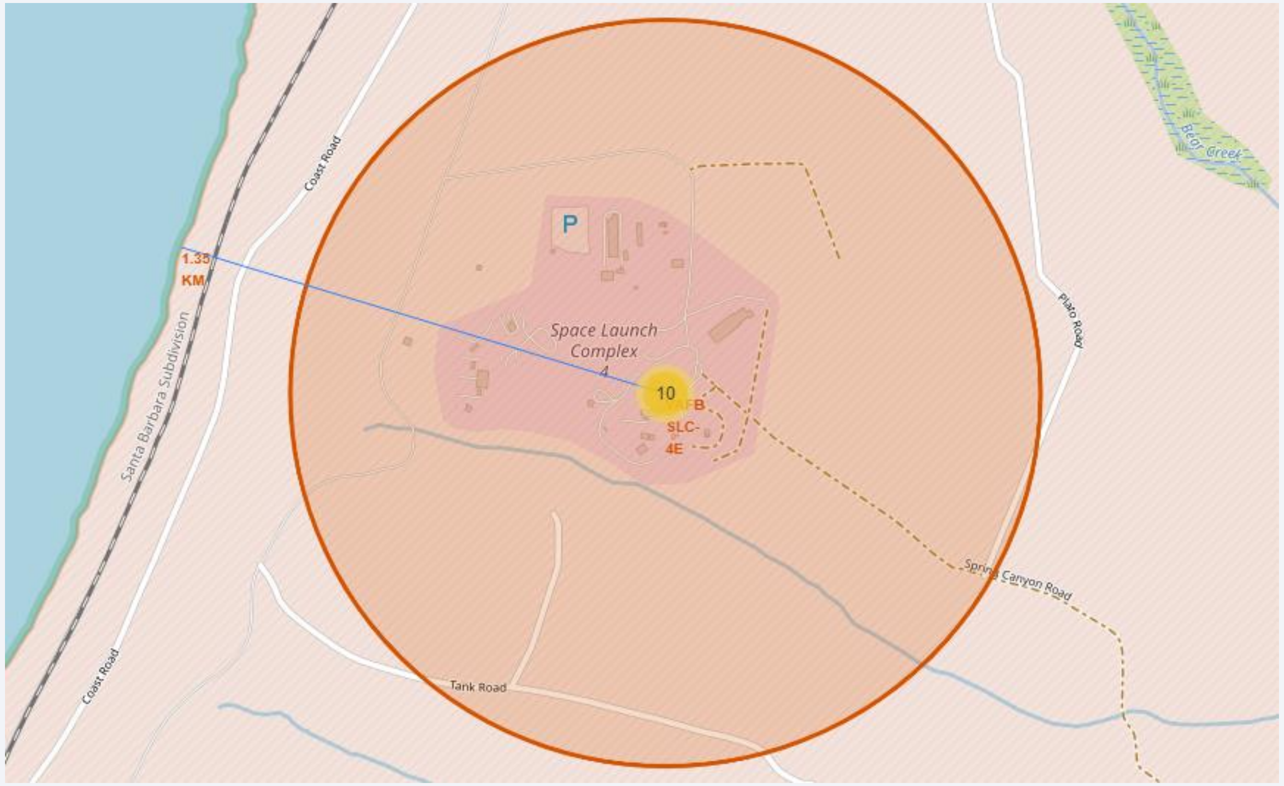
(KSC LC-39A)



Launch Site Proximities

- Launch sites are close to the coast in case of failures and crashes so they can occur in the water.
- Launch sites are close to railways to allow for easy transport of materials and equipment to the launch site
- The Vandenburg launch site is only ~1km from the coast and the nearest railway

VAFB SLC-4E

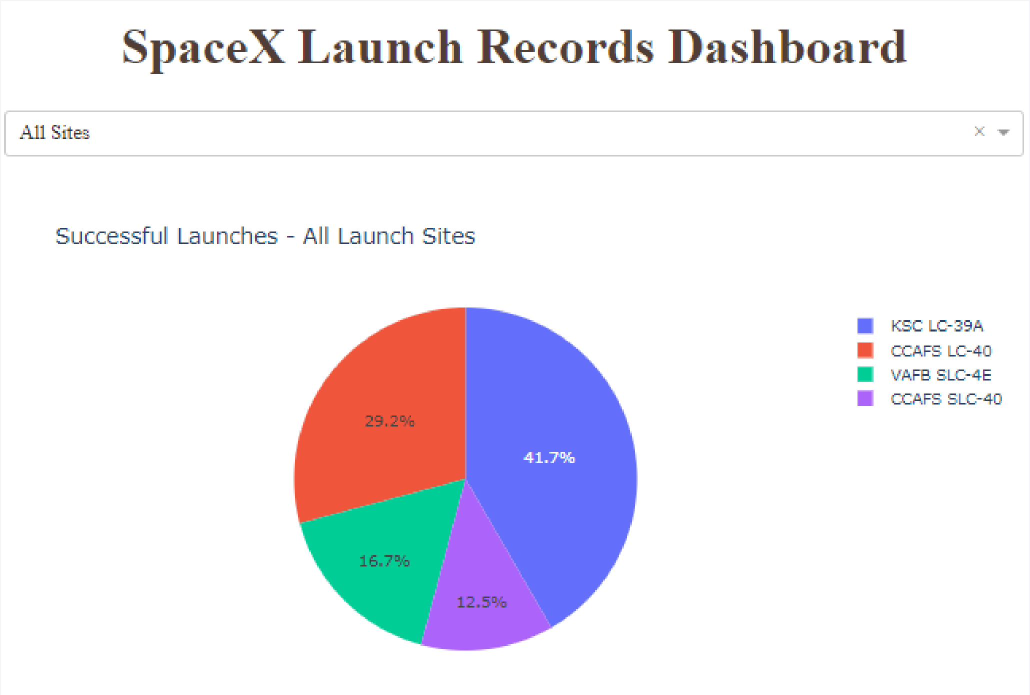


Section 4

Build a Dashboard with Plotly Dash

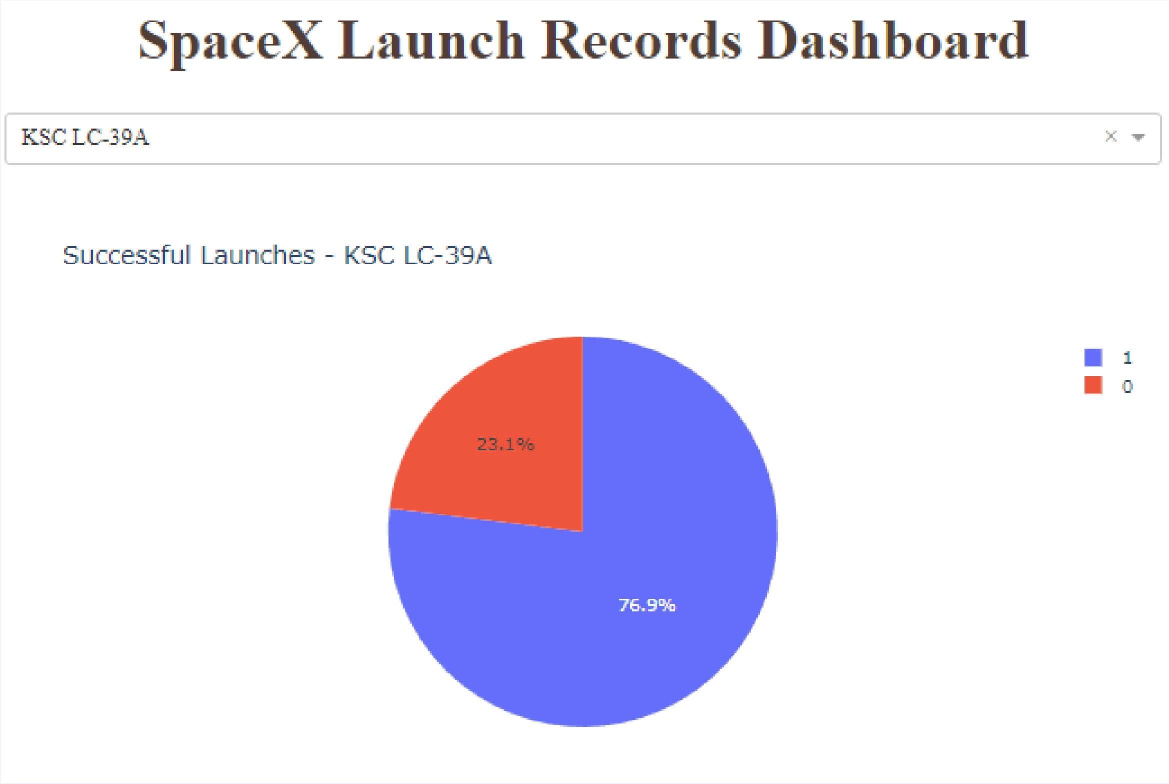
Launch Success Pie Chart

- A pie chart depicting the successful launches of the data set lets us see that the KSC launch site has the greatest number of successful launches, followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40
  - The percentages represent the percent of successful launches that occurred at each of the sites



Launch Success Pie Chart – KSC LC-39A

- The ability to select the individual sites on the dropdown makes it easy to see that the KSC site has the greatest success to failure ratio with over ¾ of the launches being successful



Payload vs. Launch Outcome

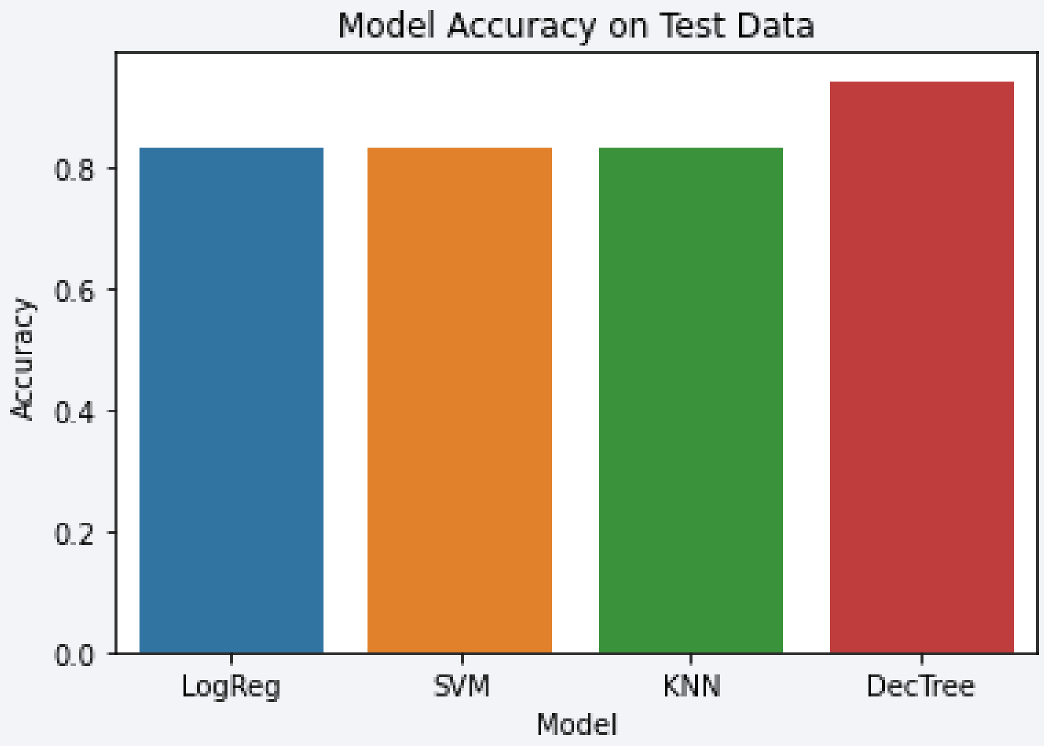
- The interactive payload range slider lets us see that in the payload range of 0-6000kg, booster version category FT has a much higher success rate than that of the v1.1 booster version.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree classification model had the highest accuracy on both the training and test data



Confusion Matrix – Decision Tree Model

- The confusion matrix of the decision tree model depicts the summary of the prediction results on the test data
- The model only misclassified 1 of the 18 records in the test data as a false positive

Conclusions

- A decision tree classification model will be the most accurate in determining the success of the first stage landing
- Average probability of a successful first stage landing has increased since 2013 and will likely continue to increase
- There is a strong dependency between the orbit type and the payload mass in the landing success rate
- Launches from the Kennedy Space Center launch site have had the most success in landing the first stage
- Landing on a ground pad has the highest probability of land success

Appendix

- SpaceX API
- BeautifulSoup
- Seaborn
- Folium
- Orbit Types
    - **LEO**: Low Earth orbit
    - **VLEO**: Very Low Earth Orbits
    - **GTO: G**eosynchronous orbit
    - **SO** Sun-synchronous orbit
    - **ES-L1** :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth
    - **HEO:** A highly elliptical orbit
    - **ISS:** A modular space station (habitable artificial satellite) in low Earth orbit
    - **MEO:** Geocentric orbits ranging in altitude from 2,000 - 35,786 km
    - **HEO:** Geocentric orbits above 35,786 km
    - **GEO:** Circular geosynchronous orbit 35,786 km
    - **PO:** Passes above both poles of the Earth

Thank you!