IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lee C Ly
6/05/2024

Outline

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

Executive Summary

Space X leads the world in lowest price-per-launch due to reusability of resources from successful phase 1 launches. Leveraging the most accurate machine learning model, we were able to predict outcomes of successful launches to determine cost per launch.

Introduction

While rocket launches across industry leading space agencies are similar, Space X leads the world in lowest price-per-launch due to reusability of resources from successful phase 1 launches. This allows Space X to dominate the space travel market.

The goal of this project is to determine the price of each Space X Falcon 9 launch by predicting the success rates of phase 1 launches using advanced Machine Learning models and leveraging historical public data.

The knowledge uncovered here will give SpaceY a competitive advantage in the space travel business by offering competitive partnership agreements.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was obtained directly through SpaceX using the SpaceX API and web scrapping of Falcon 9 and Falcon Heavy launch records from Wikipedia
- Exploratory data analysis (EDA) using Jupyter Notebook with Python and data science libraries including pandas and SQL
- Folium and Plotly Dash was used to visualize the data
- The best performing classification Machine Learning models selected for comparison were
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbors (KNN)

Data Collection - API

- Data collection and cleaning using Space X API
  - Requested and parsed the SpaceX launch data using the GET request
    - Payload data: https://api.spacexdata.com/v4/payloads/
    - Launch site data: https://api.spacexdata.com/v4/launchpads/
    - Booster version data: https://api.spacexdata.com/v4/rockets/
  - Filtered the dataframe to only include Falcon 9 launches
  - Cleaned missing values

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True

90 rows × 17 columns

Data Collection – Web Scrapping

- Data Collection using Web scrapping with BeautifulSoup
  - Requested Falcon 9 Launch Wiki page from its URL
  - Extracted all column/variable names from the HTML table header
  - Created dataframe by parsing the launch HTML tables

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	[4 June 2010,, 18:45]	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	[8 December 2010,, 15:43]	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	[22 May 2012,, 07:44]	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	[8 October 2012,, 00:35]	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	[1 March 2013,, 15:10]	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
844	[13 December 2020, 17:30:00]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
845	[19 December 2020, 14:00:00]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
846	[8 January 2021, 02:15]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
847	[20 January 2021, 13:02]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
848	[24 January 2021, 15:00]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

849 rows × 11 columns

EDA with Data Visualization

Scatter plots were used to show relationship between 2 variables. The charts were customized with special hue to color code classes

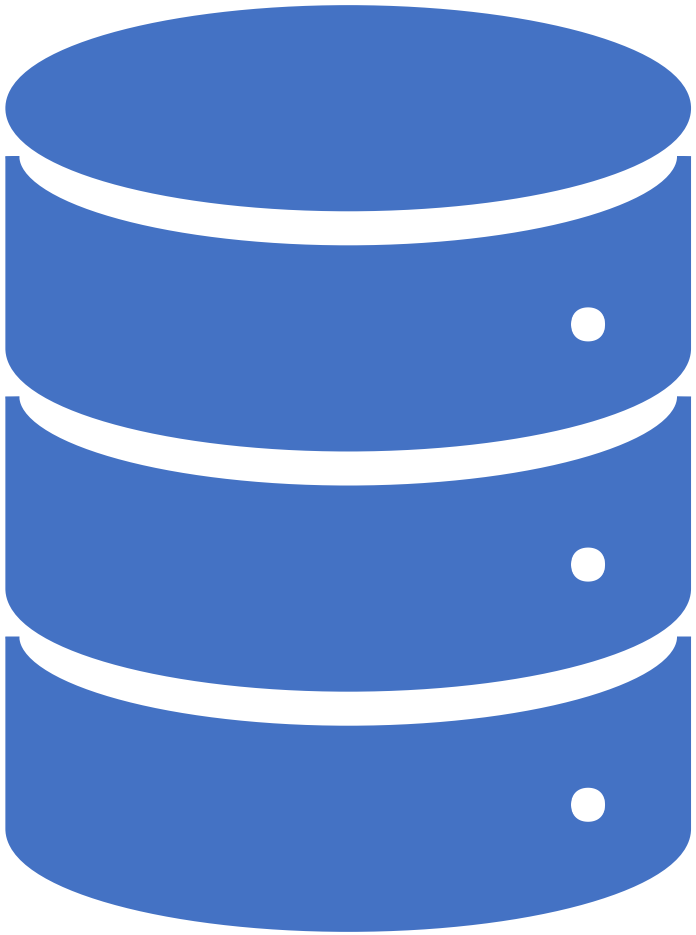
Bar plot was used to observe success rate of each orbit

Line plot was used to visualize success rate of launches through the years

EDA with SQL

SQL queries performed

- Display the names of the unique launch sites in the space mission
    - %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
- Display 5 records where launch sites begin with the string 'CCA'
    - %sql SELECT Launch_Site from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- Display the total payload mass carried by boosters launched by NASA (CRS)
    - %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';
- Display average payload mass carried by booster version F9 v1.1
    - %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
- List the date when the first succesful landing outcome in ground pad was acheived.
    - %sql SELECT MIN(date) FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)';



Build an Interactive Map with Folium

- Tasks
  - Marked all launch sites on the map
  - Marked the success/failed launches for each site on the map
  - Calculated the distances between a launch site to its proximities

Build a Dashboard with Plotly Dash

Tasks

- Created Launch Site drop-down input component
- Added a callback function to render chart based on selected launch site
- Included Range Slide to select different payloads
- Added a callback function to render the scatter plot

Predictive Analysis (Classification)

- Created Y and X columns for machine learning models
- Standardized the data using Sklearn StandardScaler
- Split the data into training and testing using Sklearn train_test_split
- Utilized GridSearchCV to determine best performing model for the following:
  - Logistic Regression
  - SVM
  - Decision Tree

- Selected the best performing model: Decision Tree

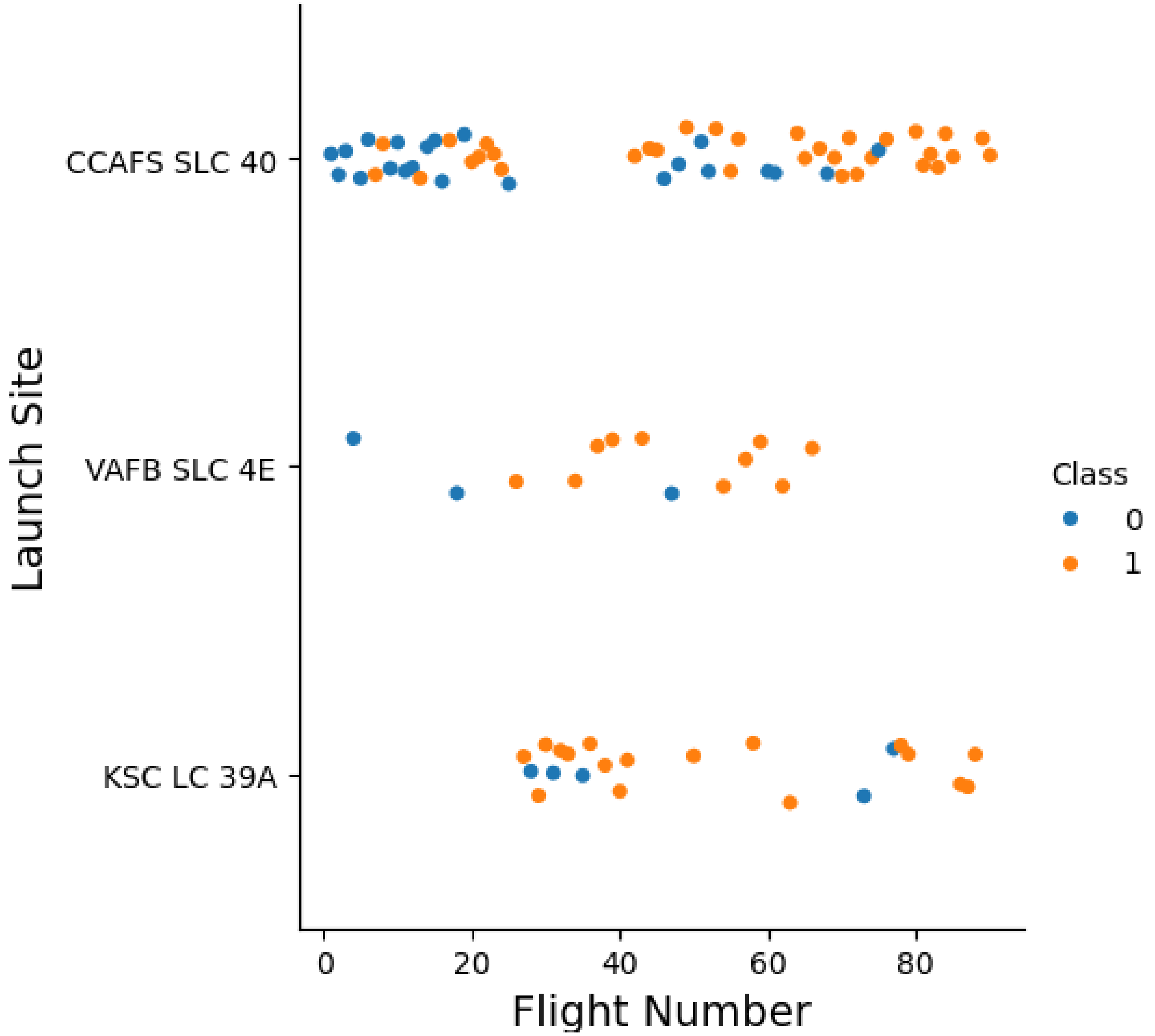
Section 2

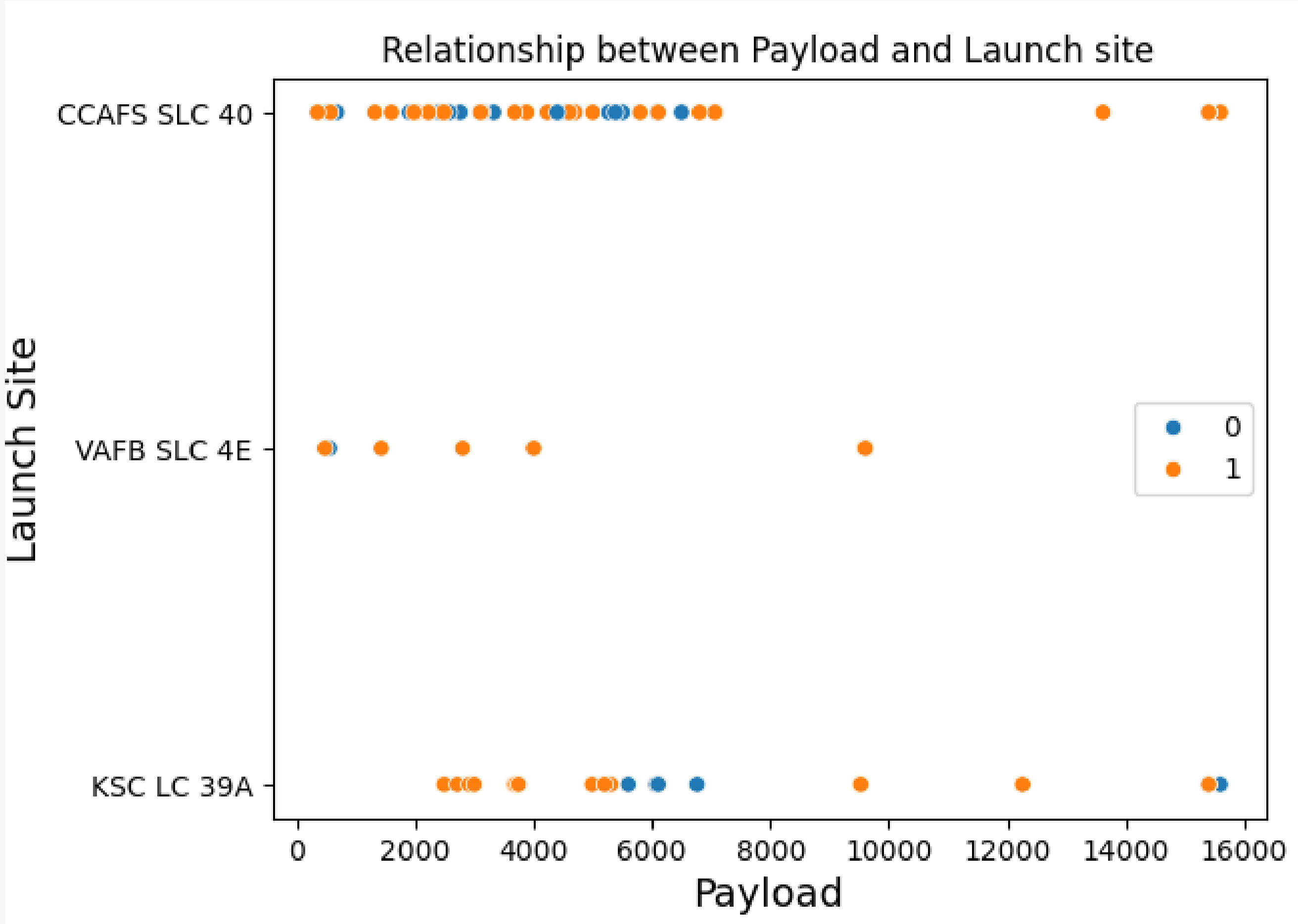
Insights drawn from EDA

Flight Number vs Launch Site

Launch Site Success Rate

- CCAFS LC-40 has a success rate of 60%
- KSC LC-39A has a success rate of 77%
- VAFB SLC 4E has a success rate of 77%

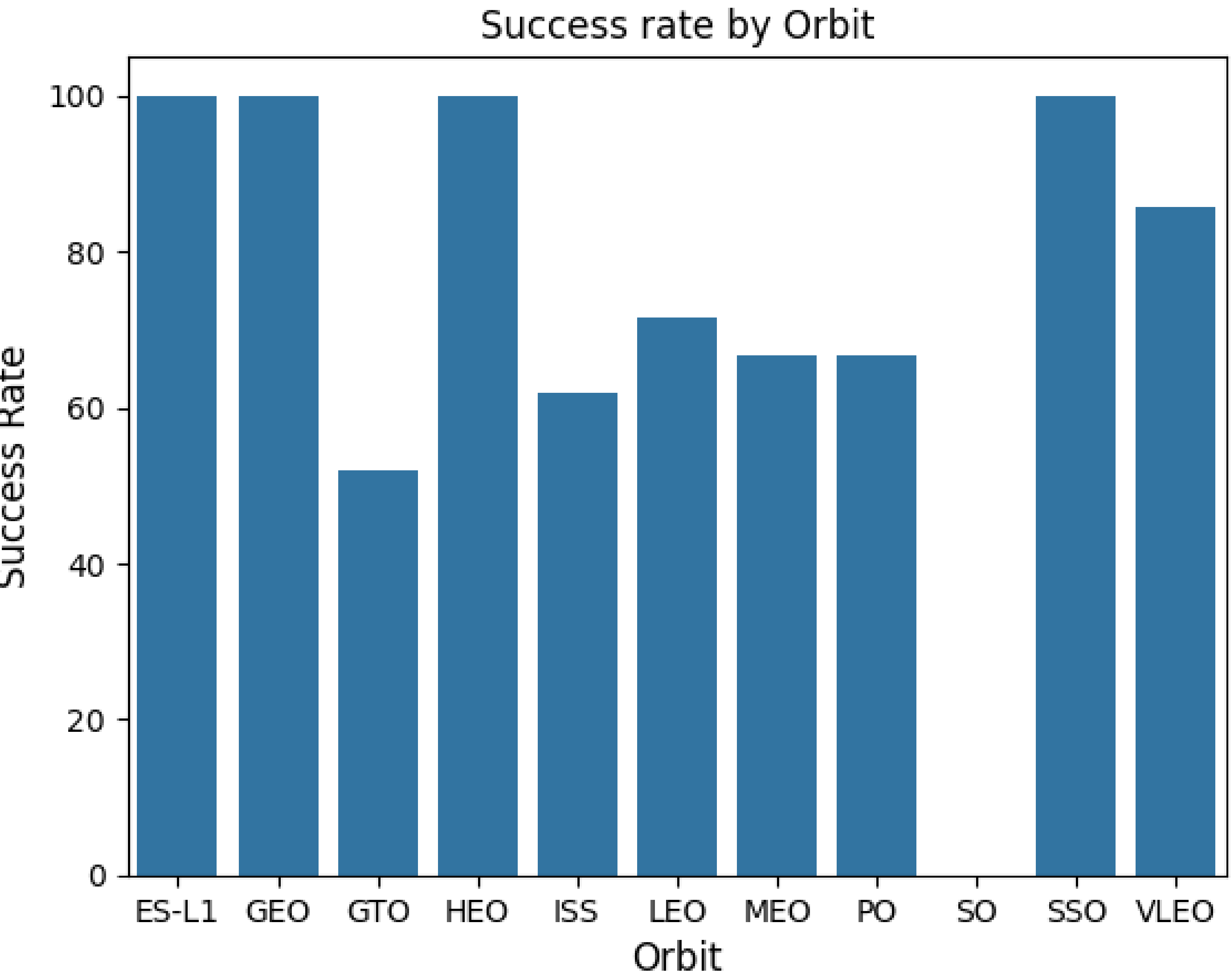




Payload vs Launch Site

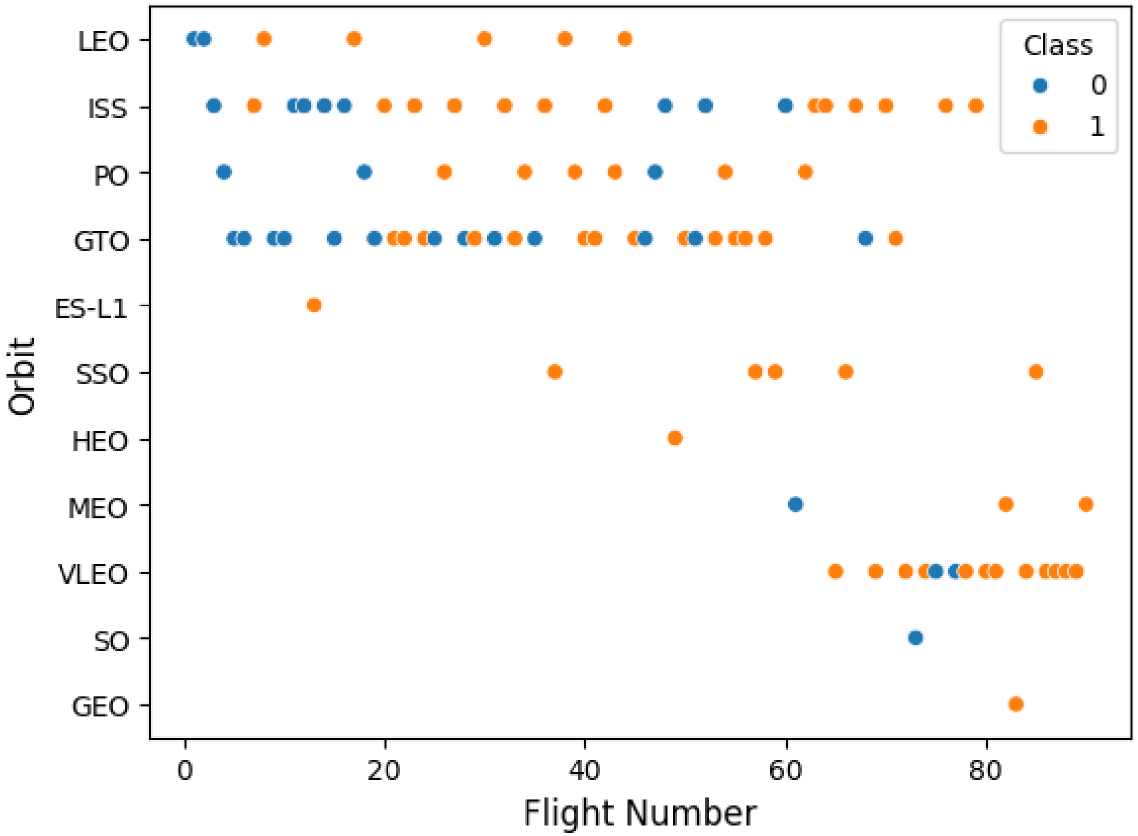
Relationship between Payload and Launch Site

- No rockets launched for heavy payload mass(greater than 10,000)



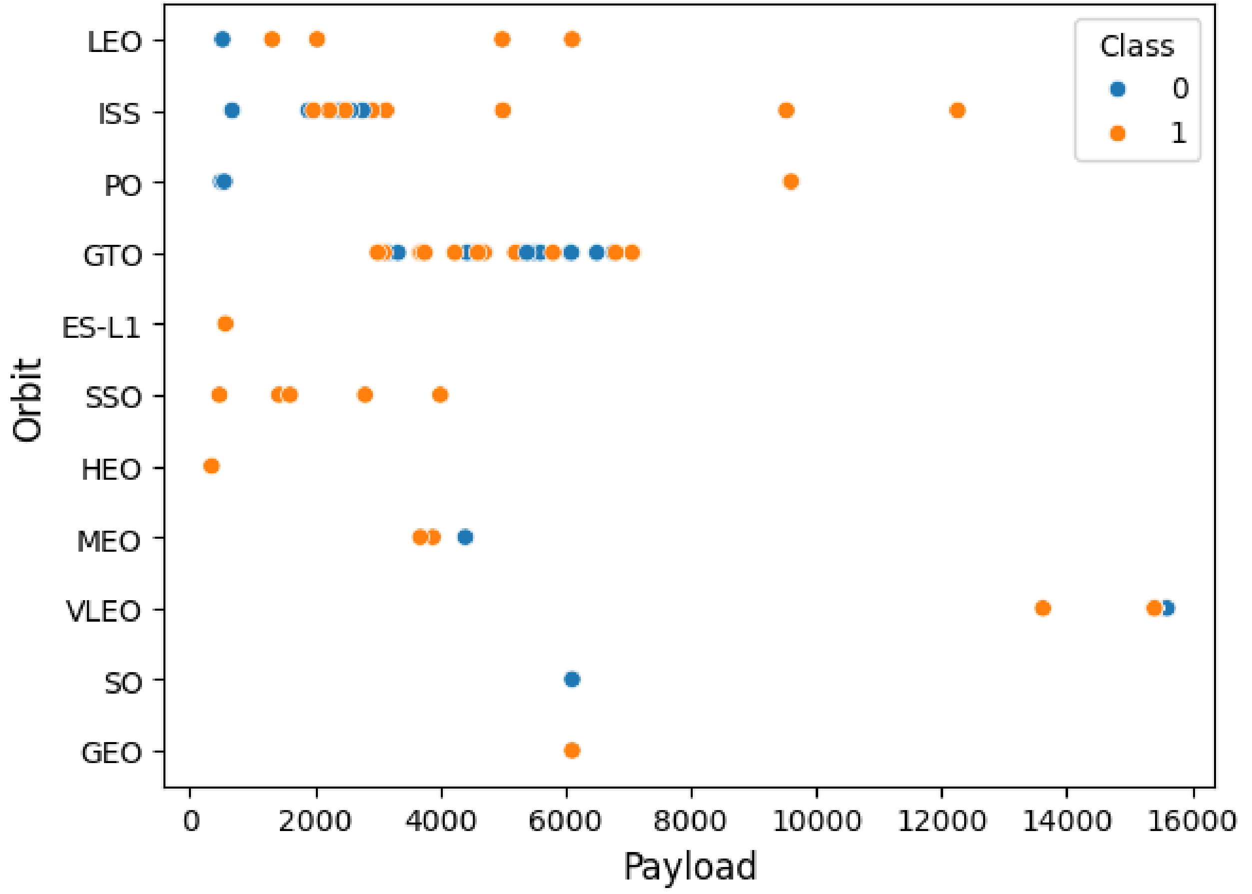
- Bar Chart for the success rate of each orbit

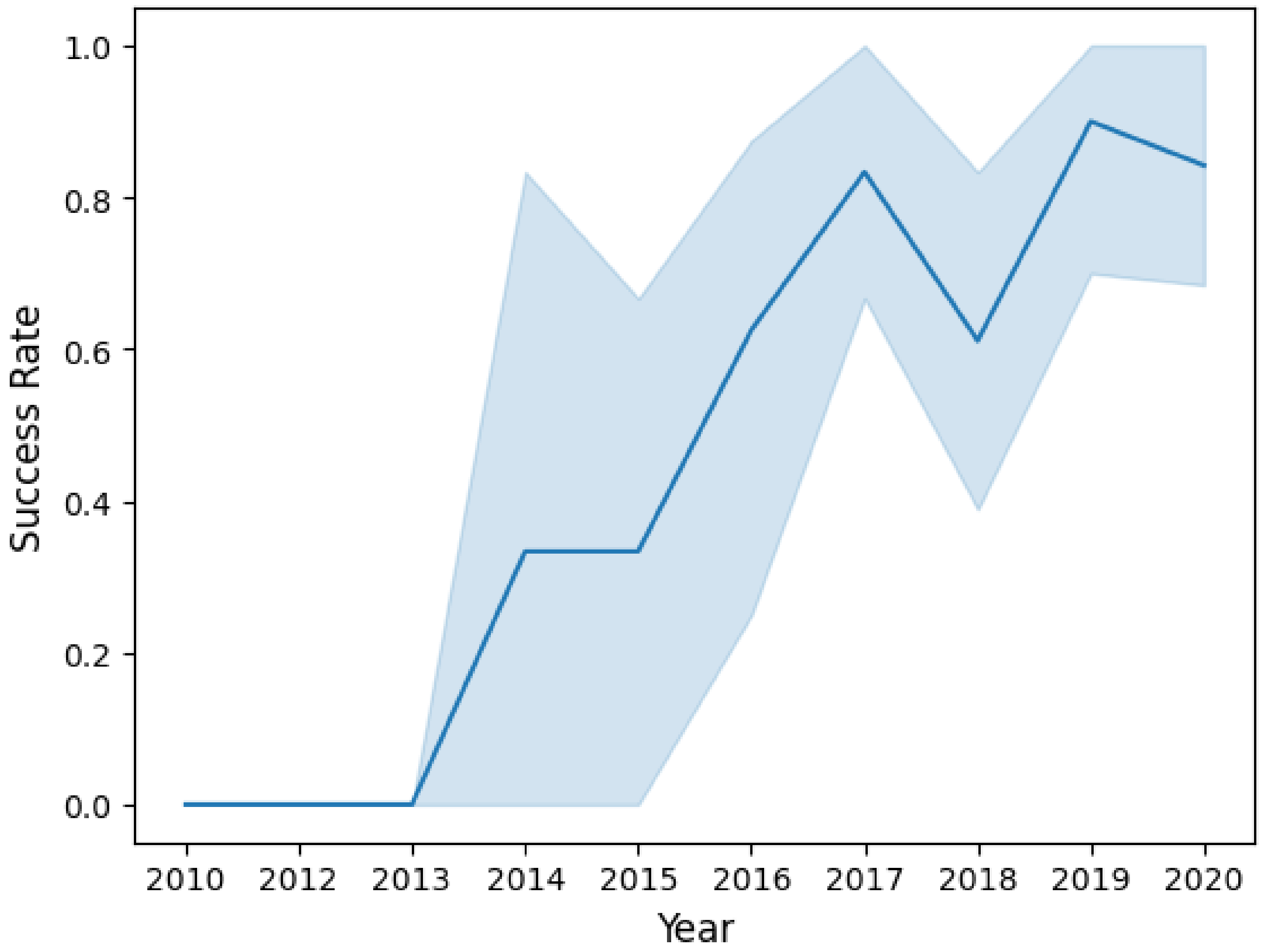
Flight Number vs Orbit Type



- Relationship between Orbit Type and Flight Number

Payload vs. Orbit Type





EDA with Data Visualization

Annual Success Rate

- As you can see, since 2013 the success rates have increased

All Launch Site Names & Average Payload Mass by F9 v1.1

SQL queries used to perform exploratory data analysis (EDA)

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

- Display the names of unique launch sites

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

```
%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

Total payload carried by boosters from NASA

- 45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

```
%sql SELECT MIN(date) FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)';
```

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

MIN(date)
2015-12-22

List the date with first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

- 100 total successful outcomes
- 1 unsuccessful outcome

```
#Successful
```

```
%sql SELECT COUNT(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%';
```

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

COUNT(Mission_Outcome)
100

```
#Failures
```

```
%sql SELECT COUNT(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%';
```

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

COUNT(Mission_Outcome)
1

Boosters Carried Maximum Payload

- Booster Version: F9 v1.1 B1018

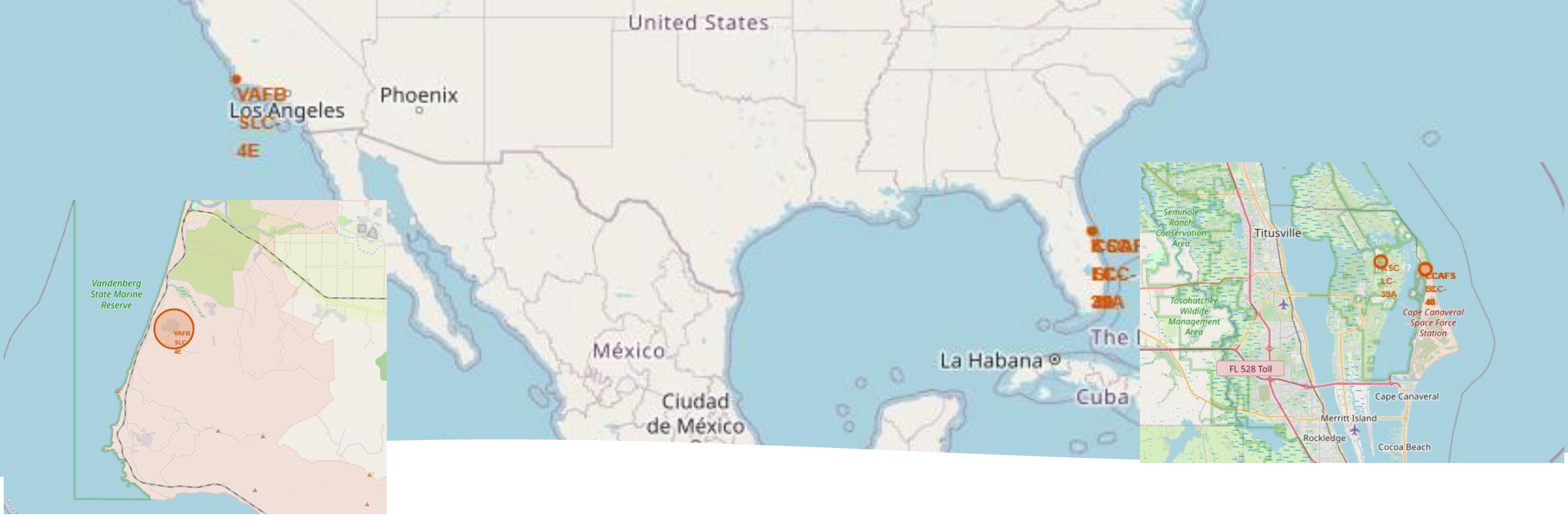
```
%sql SELECT DISTINCT Booster_version FROM SPACEXTBL WHERE Booster_version = (SELECT MAX(Booster_version) FROM SPACEXTBL);
```

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

Booster_Version
F9 v1.1 B1018

Section 3

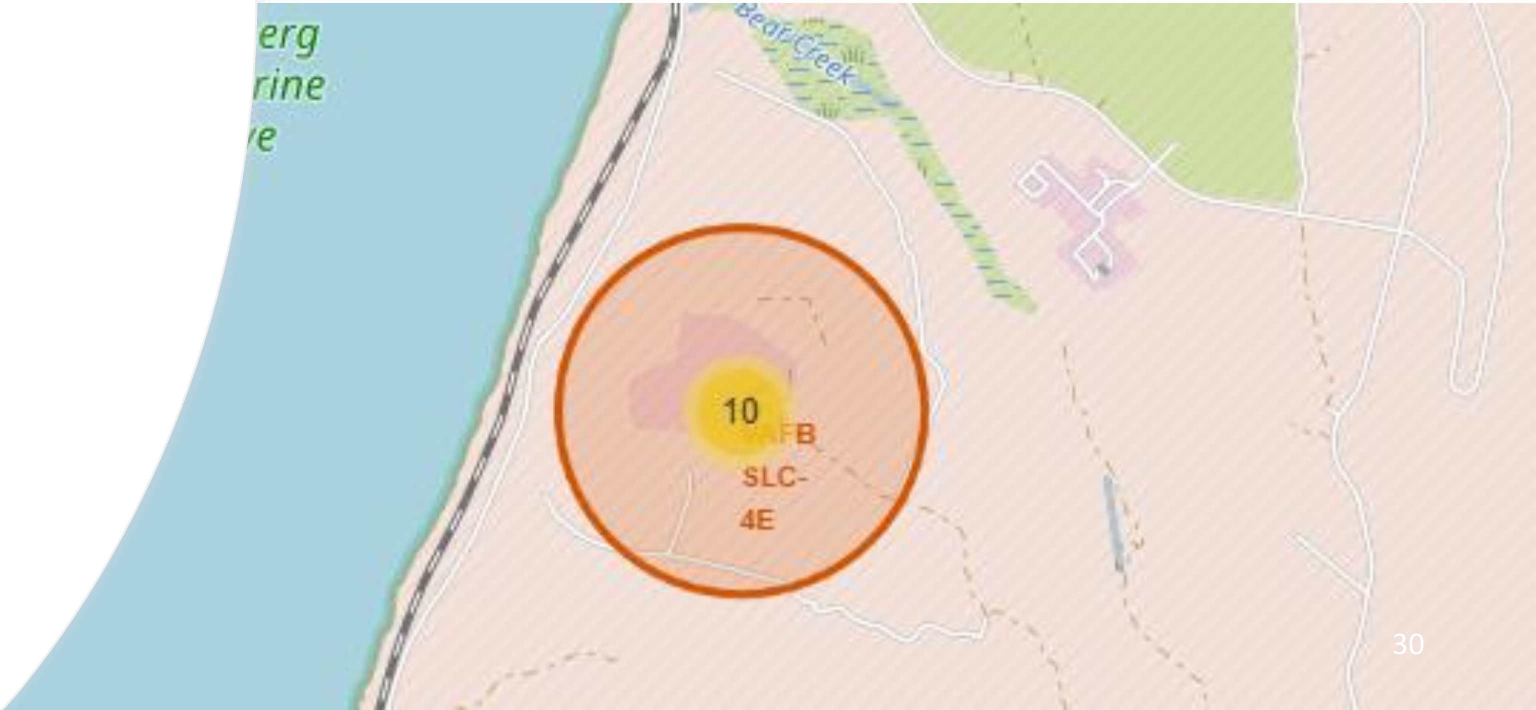
Launch Sites Proximities Analysis

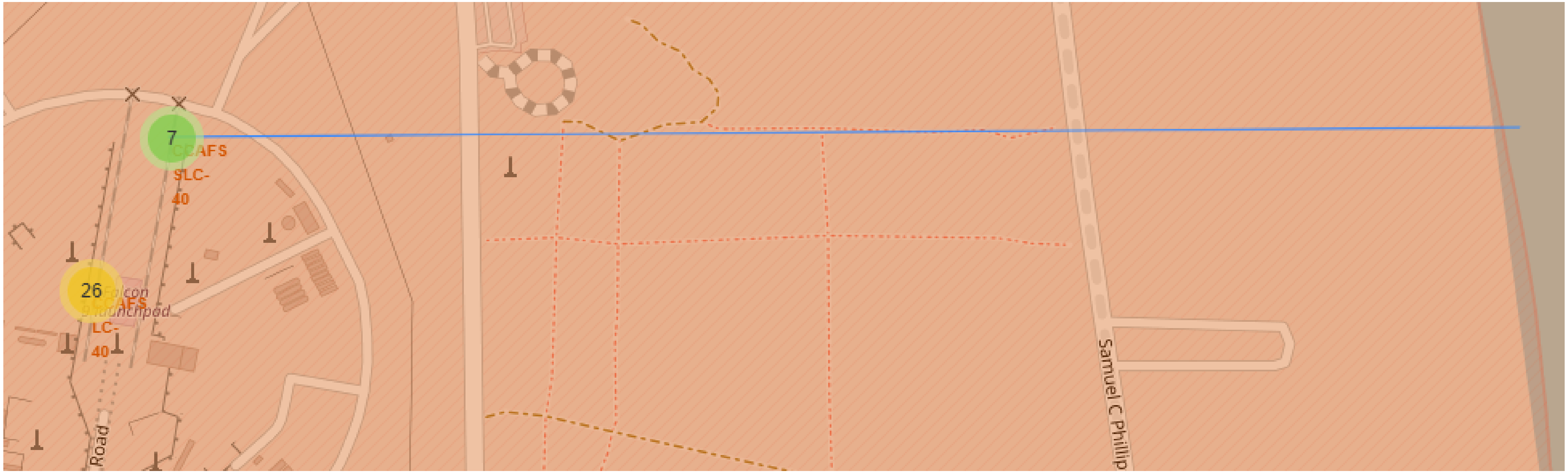


Location of all Launch Sites

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

Success/failed launches for each site





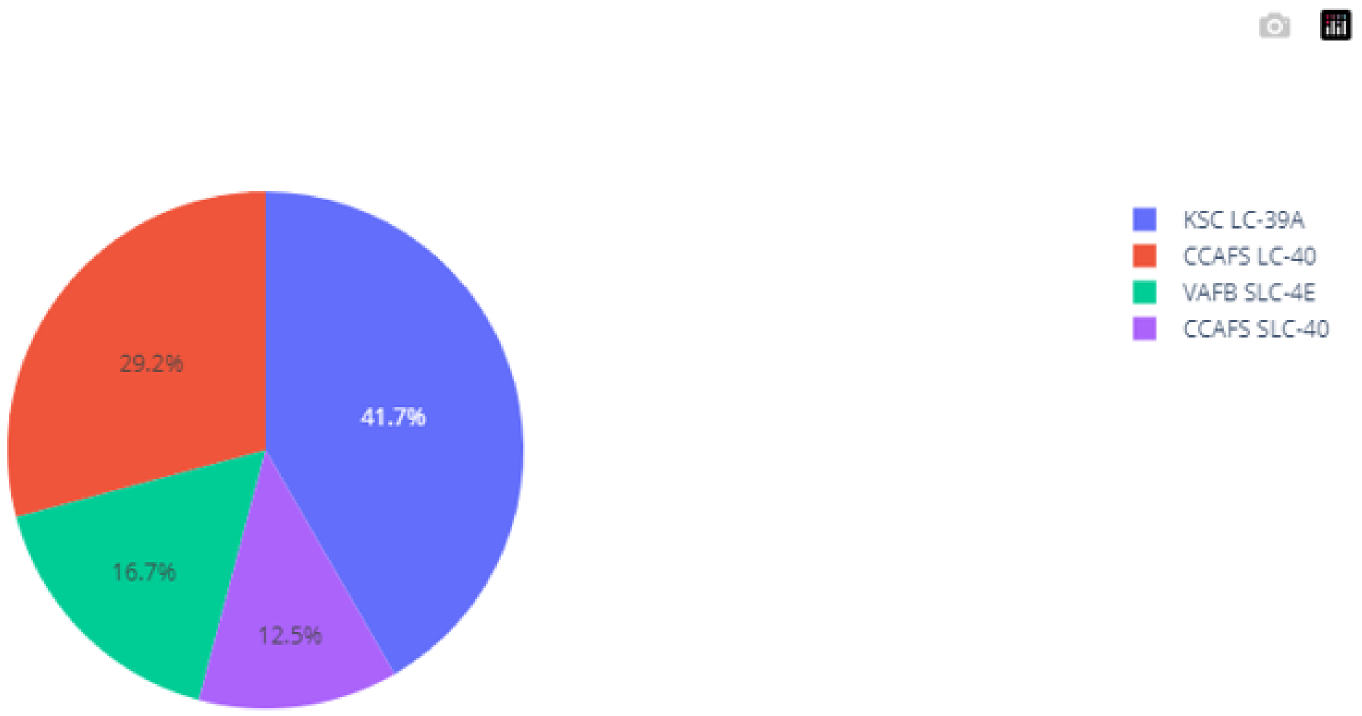
Distances between launch site to:

distance_highway = 0.585382261976943 km
distance_railroad = 1.2821722811561929 km
distance_city = 51.43534836105862 km

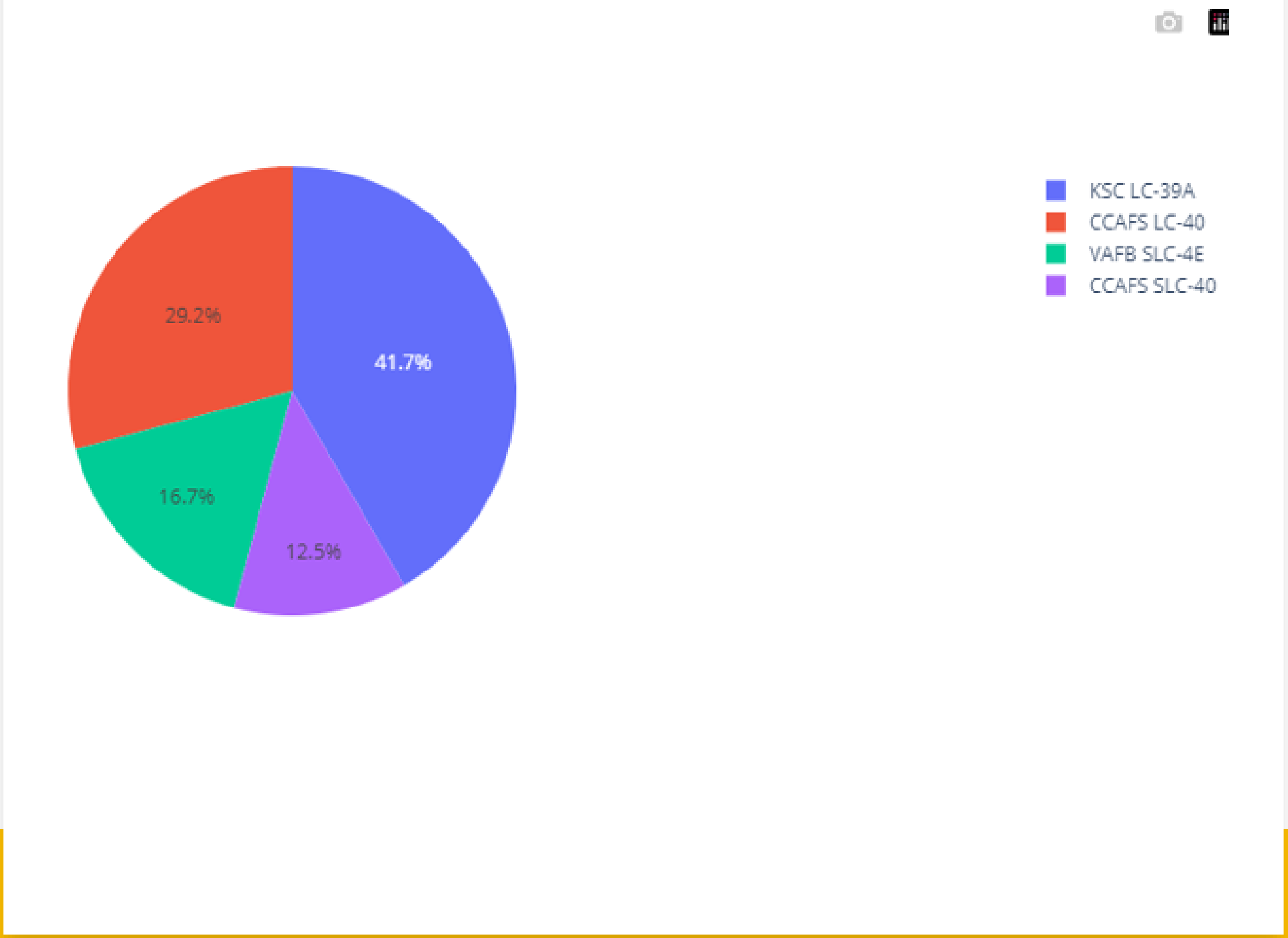
Section 4

Build a Dashboard with Plotly Dash

Successful Launches from all sites



Highest launch success ratio



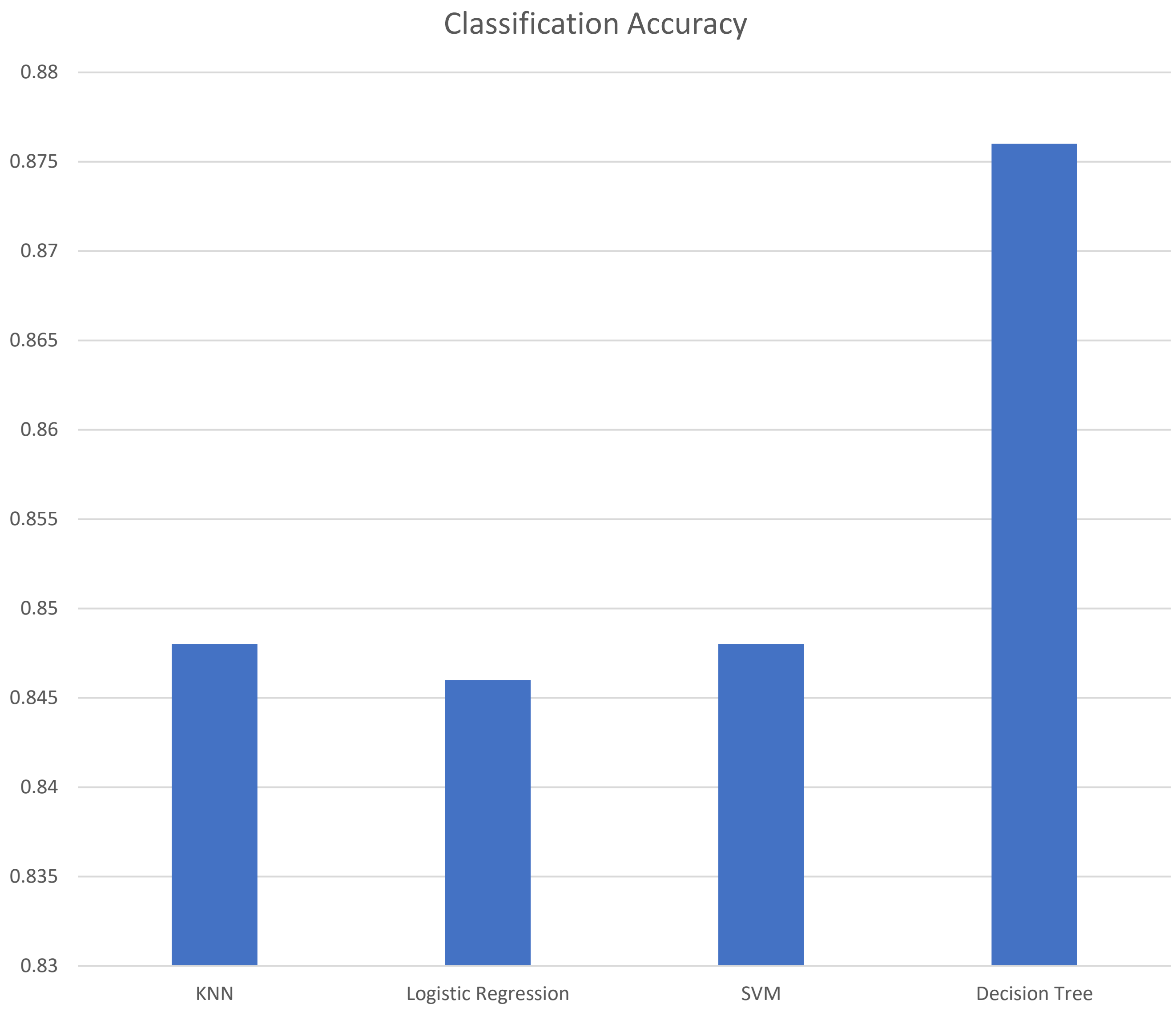
Section 5

Predictive Analysis (Classification)

Classification Accuracy

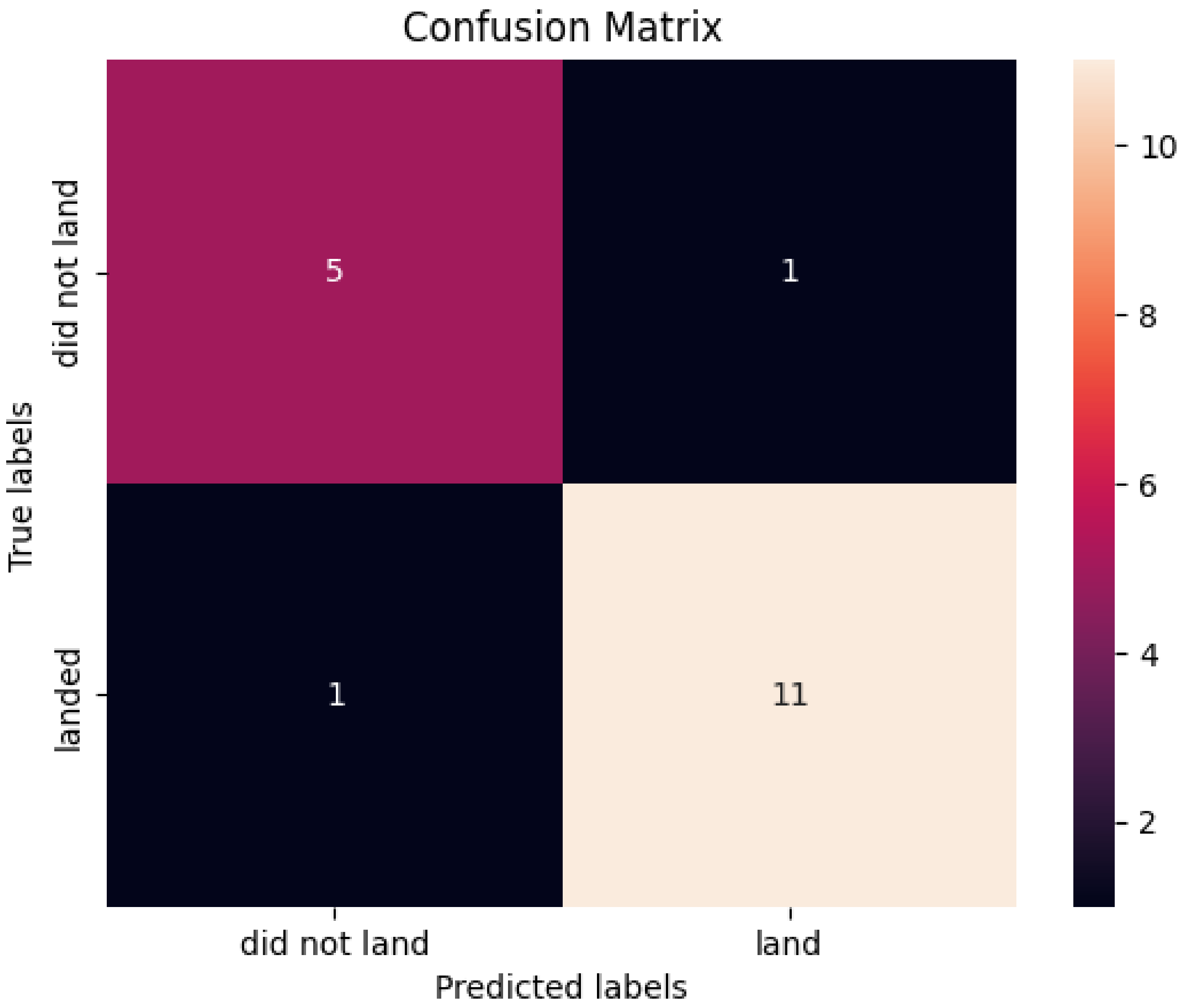
- GridSearchCV was used to determine the best performing model
- Decision Tree is the most accurate of all classification models tested

```
KNN accuracy : 0.8482142857142858
Logistic Regression accuracy : 0.8464285714285713
SVM accuracy : 0.8482142857142856
Decision Tree accuracy : 0.8767857142857143
```



Confusion Matrix

Here is the confusion matrix of the best performing: Decision Tree



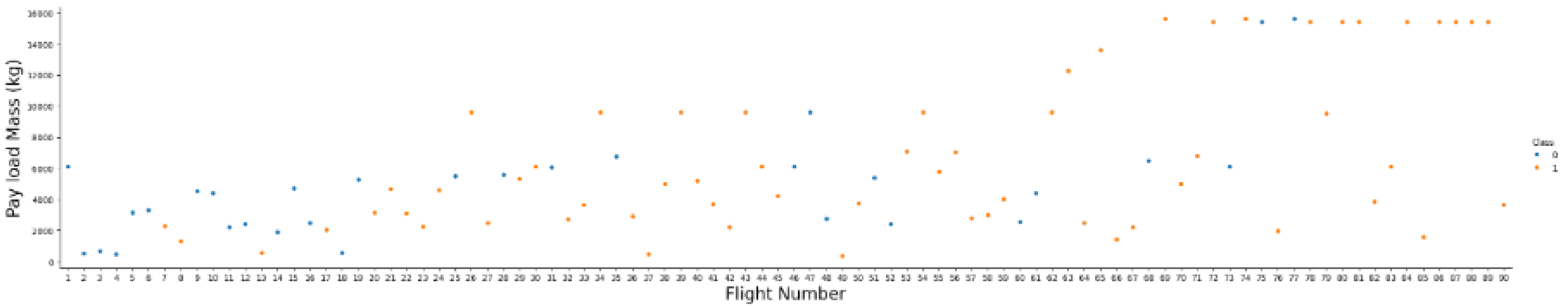
Conclusions

- During our project we were able to predict success rates of phase 1 rocket launches to determine price per cost
- Our success was dependent on the highest performing machine learning models utilizating advanced hyperparameter tuning strategies
- Our findings resulted in identifying decision tree classification model for our predictive analysis

Appendix

EDA with Data Visualization

- The relationship between Flight Number and Payload mass with overlay of outcomes
- As flight number increases, the first stage is more likely to land successfully



Thank you!