IBM Developer SKILLS NETWORK

SpaceY

Winning Space Race with Data Science

Thomas Young

9/19/24

Outline

Executive Summary

- This project attempts to go through the research process of identifying a factors to a successful rocket landing using the following steps:
- Summary of all results
  - EDA (exploratory data analysis)
    - Orbits ES-L1, GEO, HEO, and SSO have highest success rate of 100%
    - Launch rates improve over time
  - Visualization
    - Most successful launch sites are near the ocean / equator
  - Predictive Modeling
    - All models performed well with the best being Decision Tree

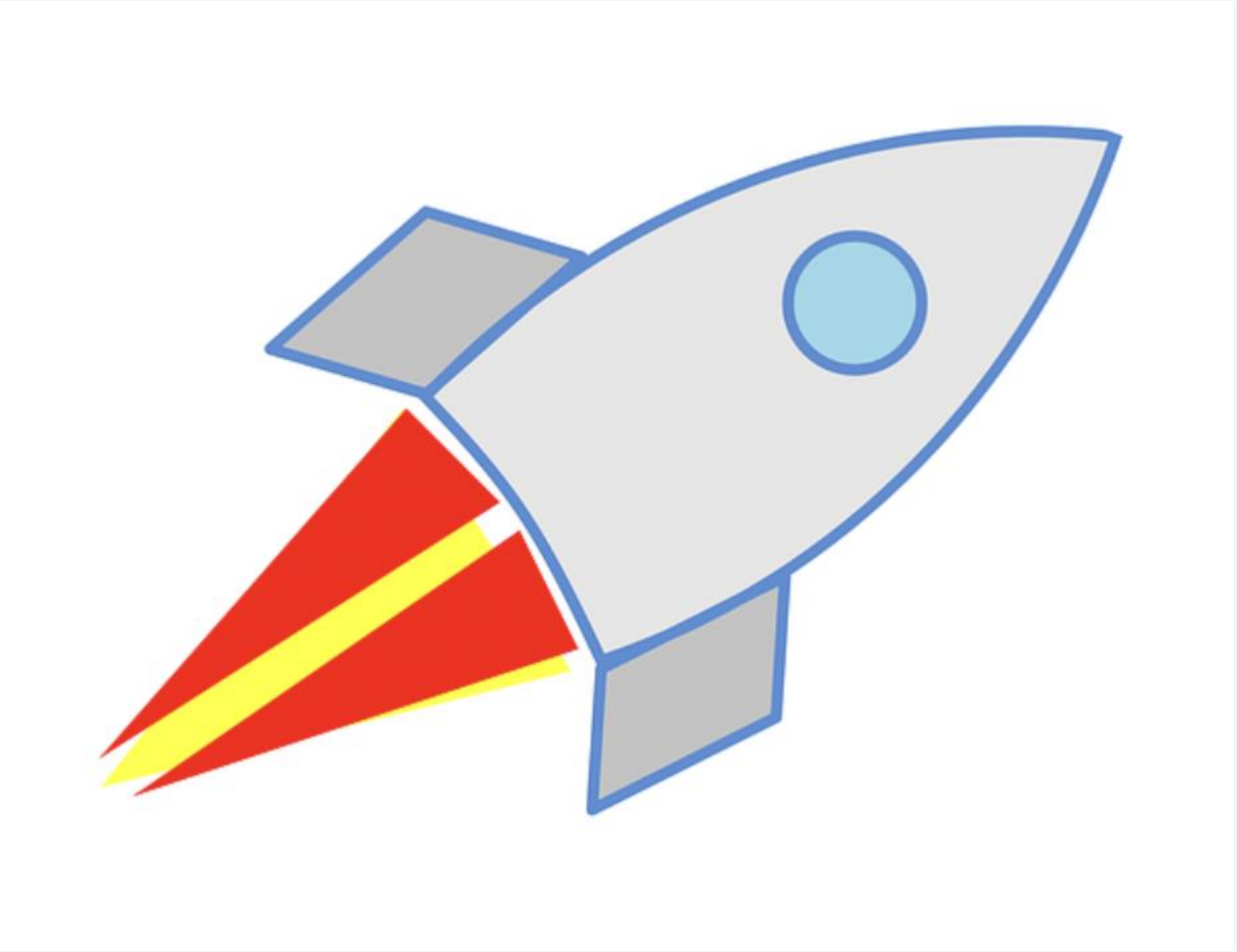
Introduction

Background

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, by determine if the first stage will land, we can determine the cost of a launch. Spaces X's Falcon 9 launch like regular rockets.

Metrics

- How the following affect stage 1 landing: payload mass, launch site, number of flights, and orbits
- Rate of successful landings over time
- Best predictive model for successful landing through sklearn library



Section 1:

Methodology

Methodology

Steps

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using data visualization techniques and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection – SpaceX API

- **Request Data**
  - SpaceX API
  - Rocket Launch Data
- **Decode Response**
  - Using .json() to convert
  - Using .json_normalize()
- **Request Information**
  - Launch Data using custom functions
- **Create / Filter Dictionary**
  - From the data and from dictionary
- **Replace Missing Values**
  - Specifically the payload mass with calculated .mean()
- **Export Data**
  - To CSV file

Data Collection – Scraping

Request Data
- (Falcon 9 launch data) from Wikipedia

Create BeautifulSoup object
- from HTML response

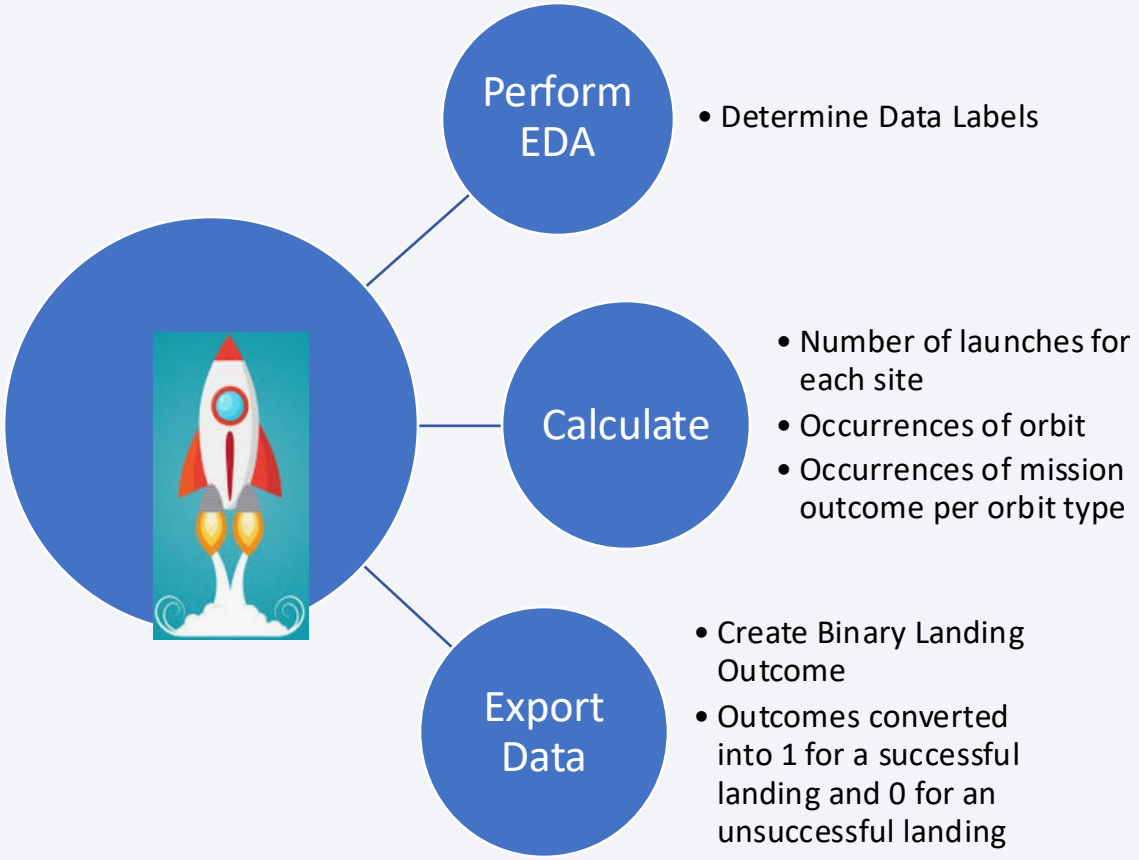
Extract column names
- from HTML table header

Collect Data
- from parsing HTML tables.mean()

Create Dictionary / Dataframe
- From the data / dictionary

Export Data
- To CSV file

Data Wrangling



Landing Outcomes

- False Ocean
  - represented an unsuccessful landing to a specific region of ocean
- True RTLS
  - meant the mission had a successful landing on a ground pad
- False RTLS
  - represented an unsuccessful landing on a ground pad
- True ASDS
  - meant the mission outcome had a successful landing on a drone ship
- False ASDS
  - represented an unsuccessful landing on drone ship

EDA with Data Visualization

- Charts
  - Flight Number vs. Payload
  - Flight Number vs. Launch Site
  - Payload Mass (kg) vs. Launch Site
  - Payload Mass (kg) vs. Orbit type
- Analysis
  - View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
  - Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

EDA with SQL – Queries

- Display:
  - Names of unique launch sites
  - 5 records where launch site begins with ‘CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1.
- List:
  - Date of first successful landing on ground pad
  - Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
  - Total number of successful and failed missions
  - Names of booster versions which have carried the max payload
  - Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Build an Interactive Map with Folium

- Markers Indicating Launch Sites
  - Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
  - Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates
- Colored Markers of Launch Outcomes
  - Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- Distances Between a Launch Site to Proximities
  - Added colored lines to show distance between launch site CCAFS SLC- 40 and its proximity to the nearest coastline, railway, highway, and city

Build a Dashboard with Plotly Dash

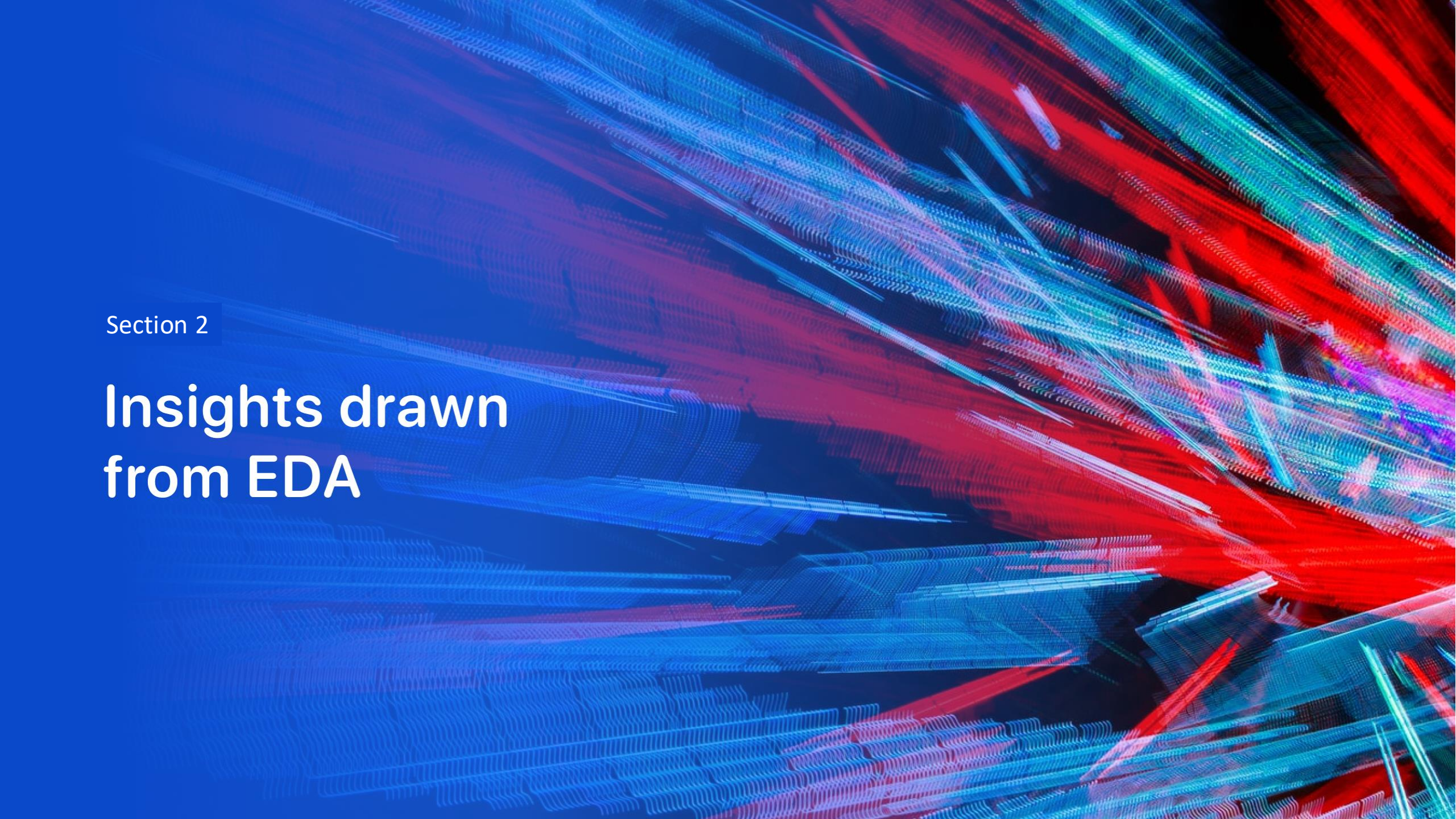
- Dropdown List with Launch Sites
  - Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range
  - Allow user to select payload mass range
- Pie Chart Showing Successful Launches
  - Allow user to see successful and unsuccessful launches as a percent of the total
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
  - Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

Create	• NumPy array from the Class column
Standardize	• the data with StandardScaler. Fit and transform the data.
Split	• the data using train_test_split
Create	• a GridSearchCV object with cv=10 for parameter optimization
Apply	• GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree(DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
Calculate	• accuracy on the test data using .score() for all models
Access	• the confusion matrix for all models
Identify	• the best model using Jaccard_Score, F1_Score and Accuracy

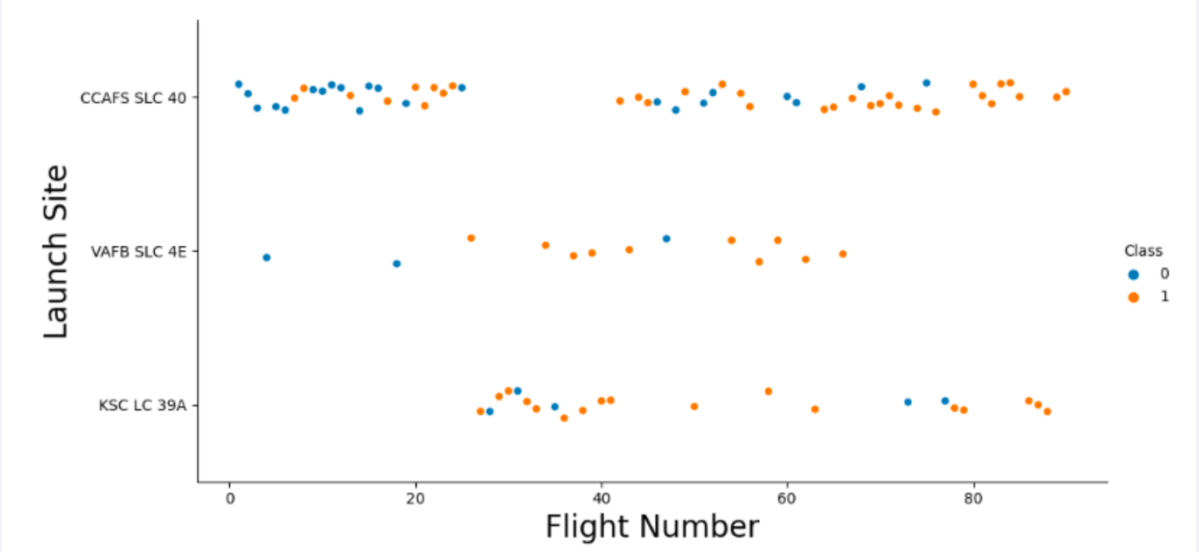
Results

- Exploratory Data Analysis
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual Analytics
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- Predictive Analytics
  - Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

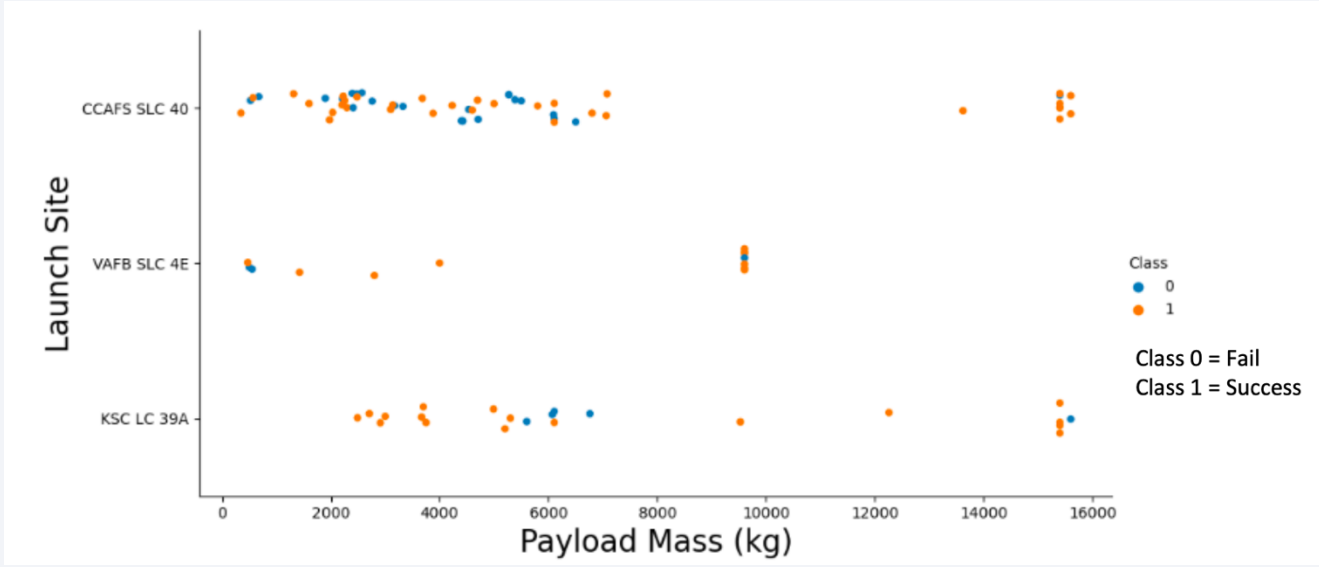
Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



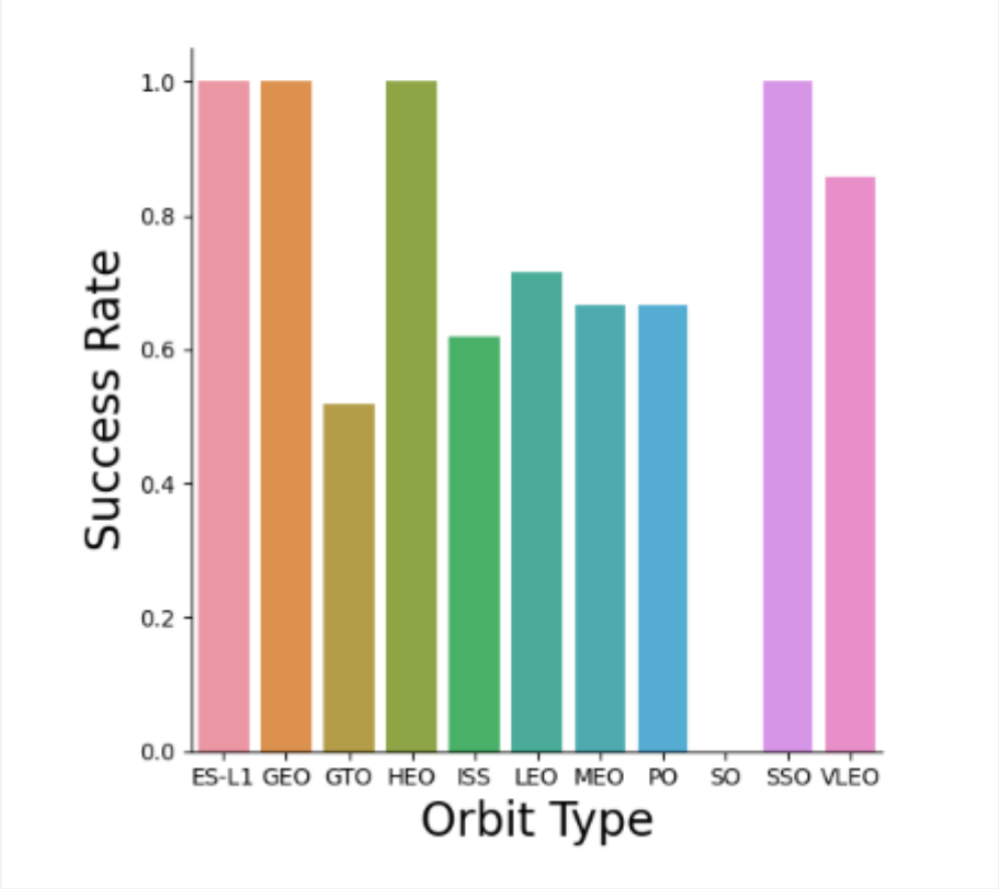
Payload vs. Launch Site

- There exhibits a trend that the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



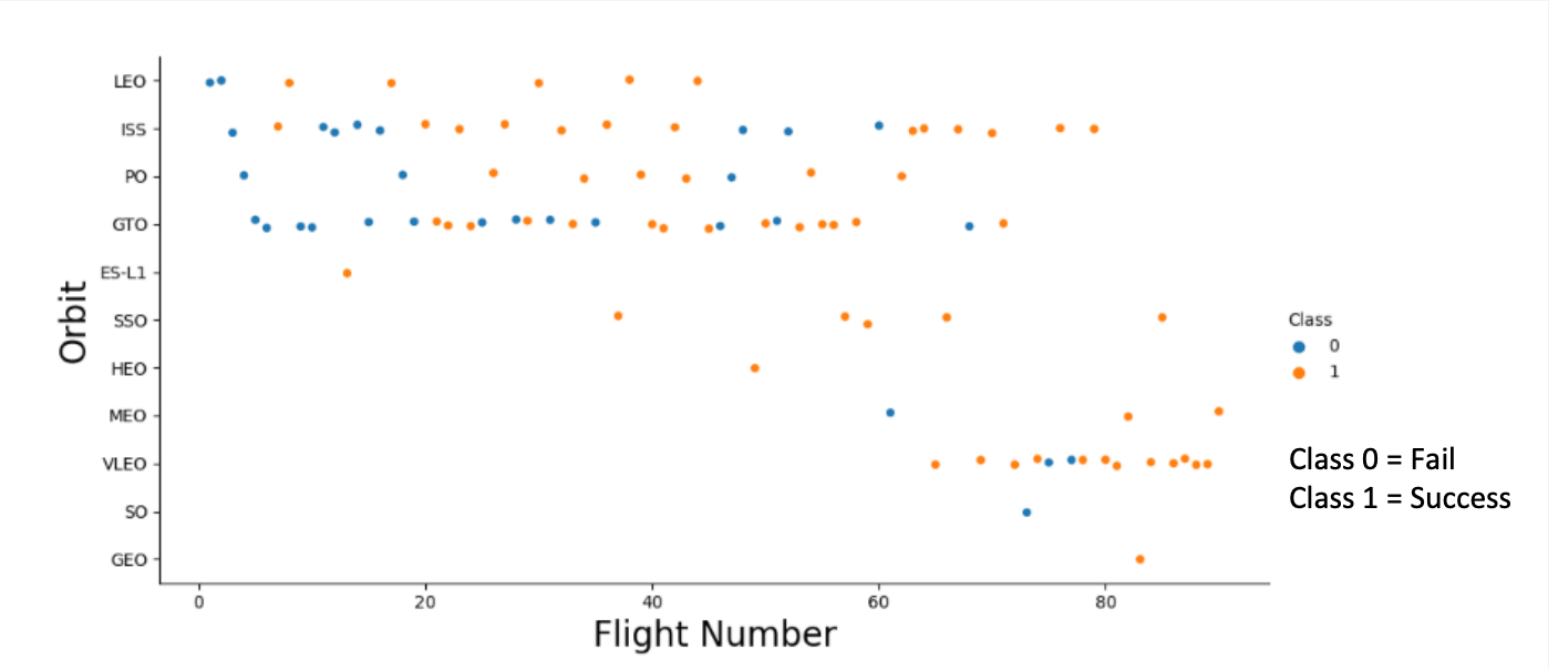
Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



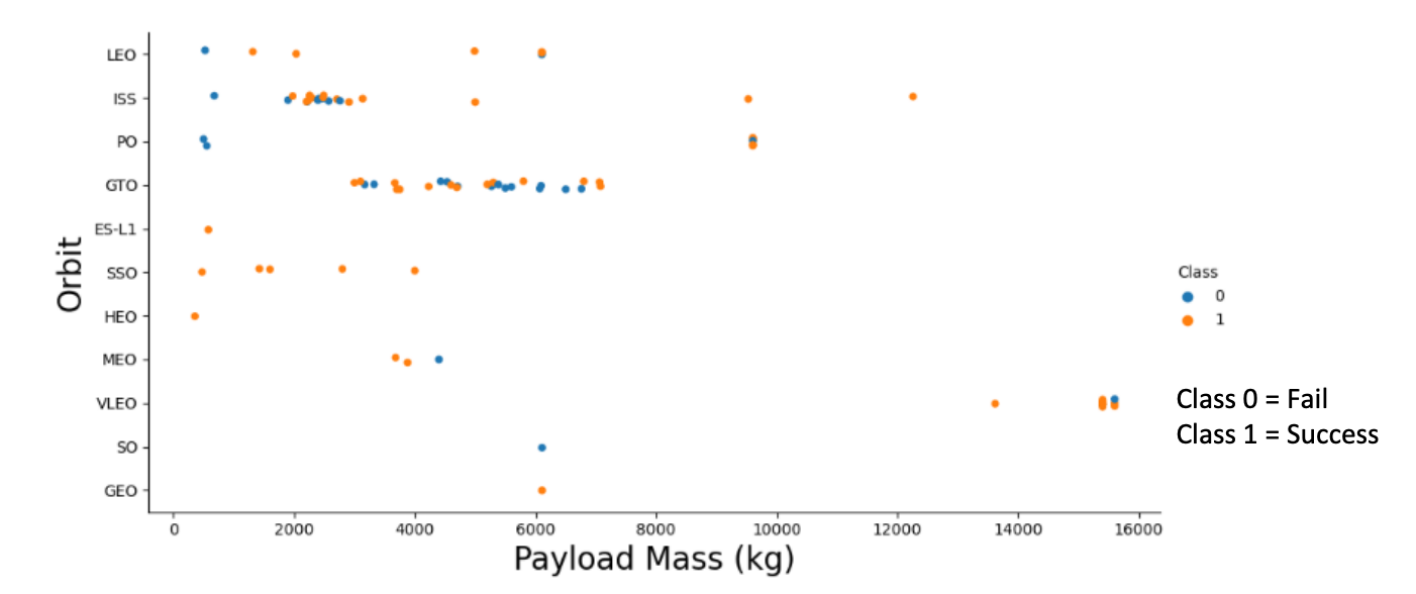
Flight Number vs. Orbit Type

- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



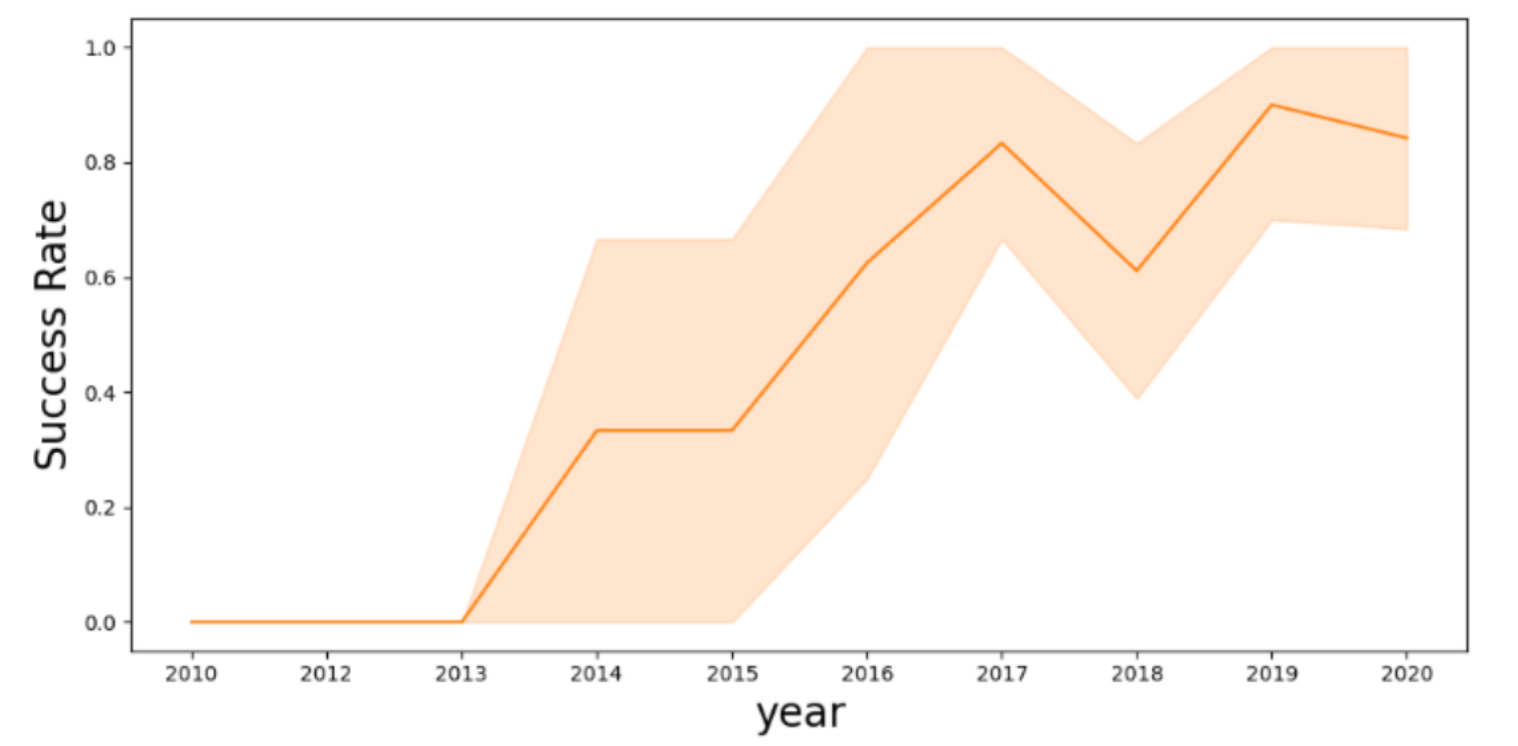
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

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

```
In [10]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[10]:

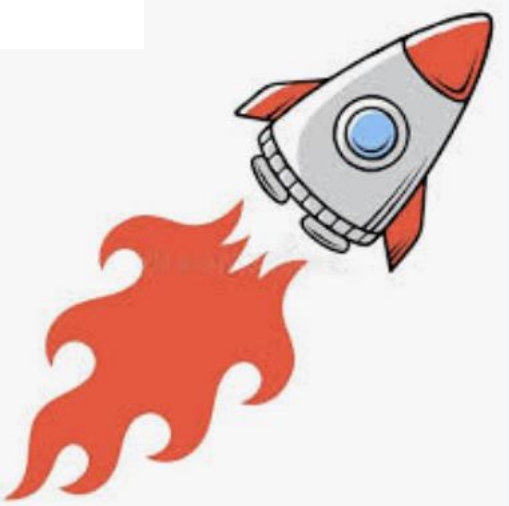
Launch_Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Records with Launch Site Starting with CCA

```
In [11]: %sql SELECT * FROM 'SPACEXTBL' WHERE Launch_Site LIKE 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 | Average Payload Mass by F9 v1.1

Total Payload Mass

- 45,596 kg (total) carried by boosters launched by NASA (CRS)

Average Payload Mass

- 2,534 kg (average) carried by booster version F9 v1.1

```
In [12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as PM_KG_TOTAL, Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[12]:

PM_KG_TOTAL	Customer
45596	NASA (CRS)

```
In [13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as PM_KG_AVG FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
```

Out[13]:

PM_KG_AVG
2534.6666666666665

First Successful Ground Landing Date | Payload between 4000 and 6000

1st Successful Landing in Ground Pad

- 7/22/20118

```
In [27]: %sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success'
 * sqlite:///my_data1.db
Done.
Out[27]:
```

min(DATE)
2018-07-22

Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Mission_Outcome = 'Success' AND PAYLOAD_MASS__KG_ > 4000 A
 * sqlite:///my_data1.db
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
In [16]: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total FROM SPACEXTBL GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

Boosters Carried Maximum Payload

Carrying Max Payload

- 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

```
In [17]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[17]:

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

2015 Failed Landings

- Showing month, date, booster version, launch site and landing outcome

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Ranked Descending

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

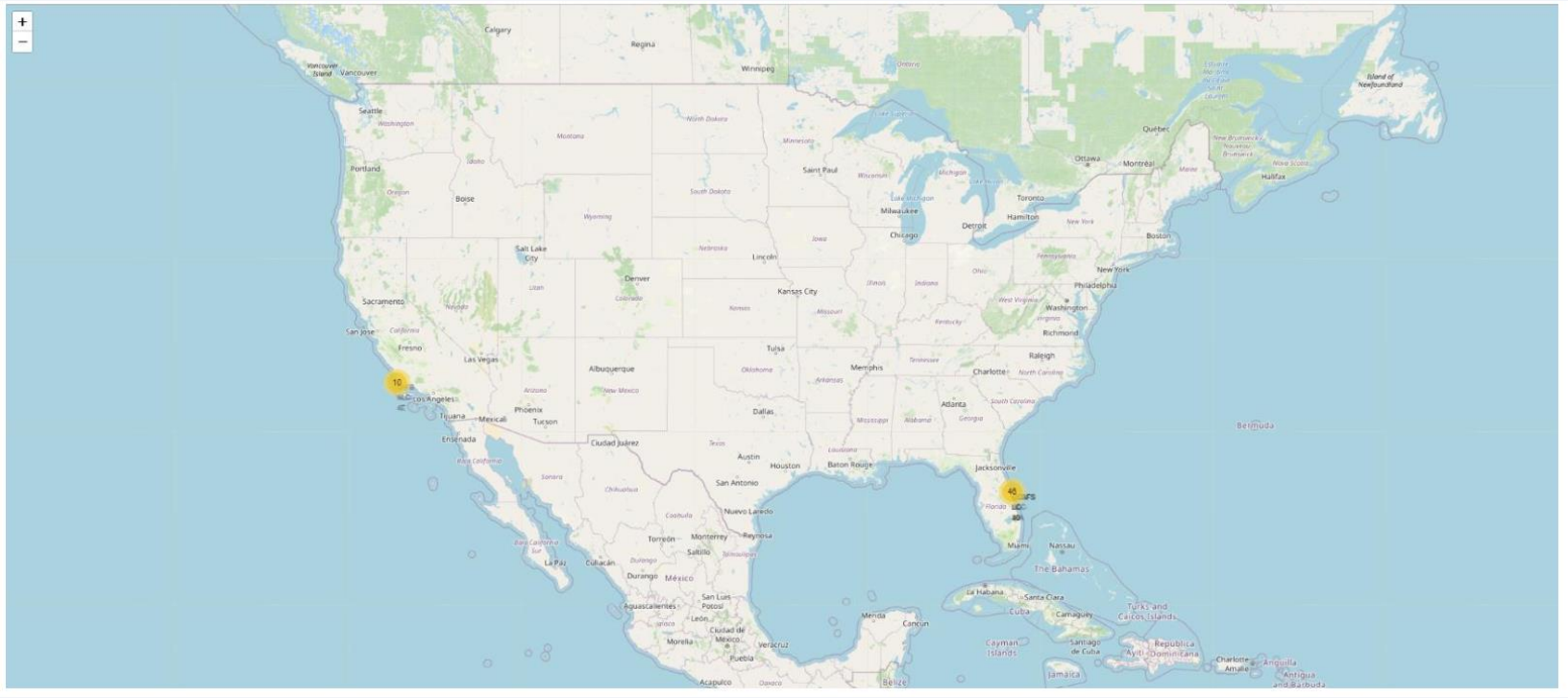
Section 3

Launch Sites Proximities Analysis

Launch Site Map

With Markers

- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost- due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



Launch Outcomes at each site

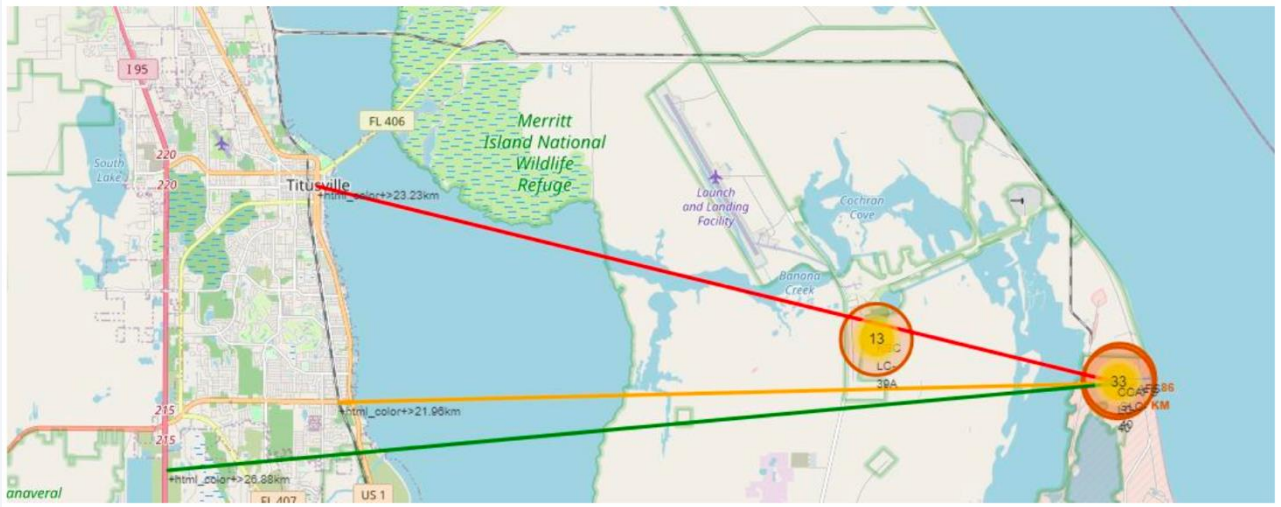
Outcomes:

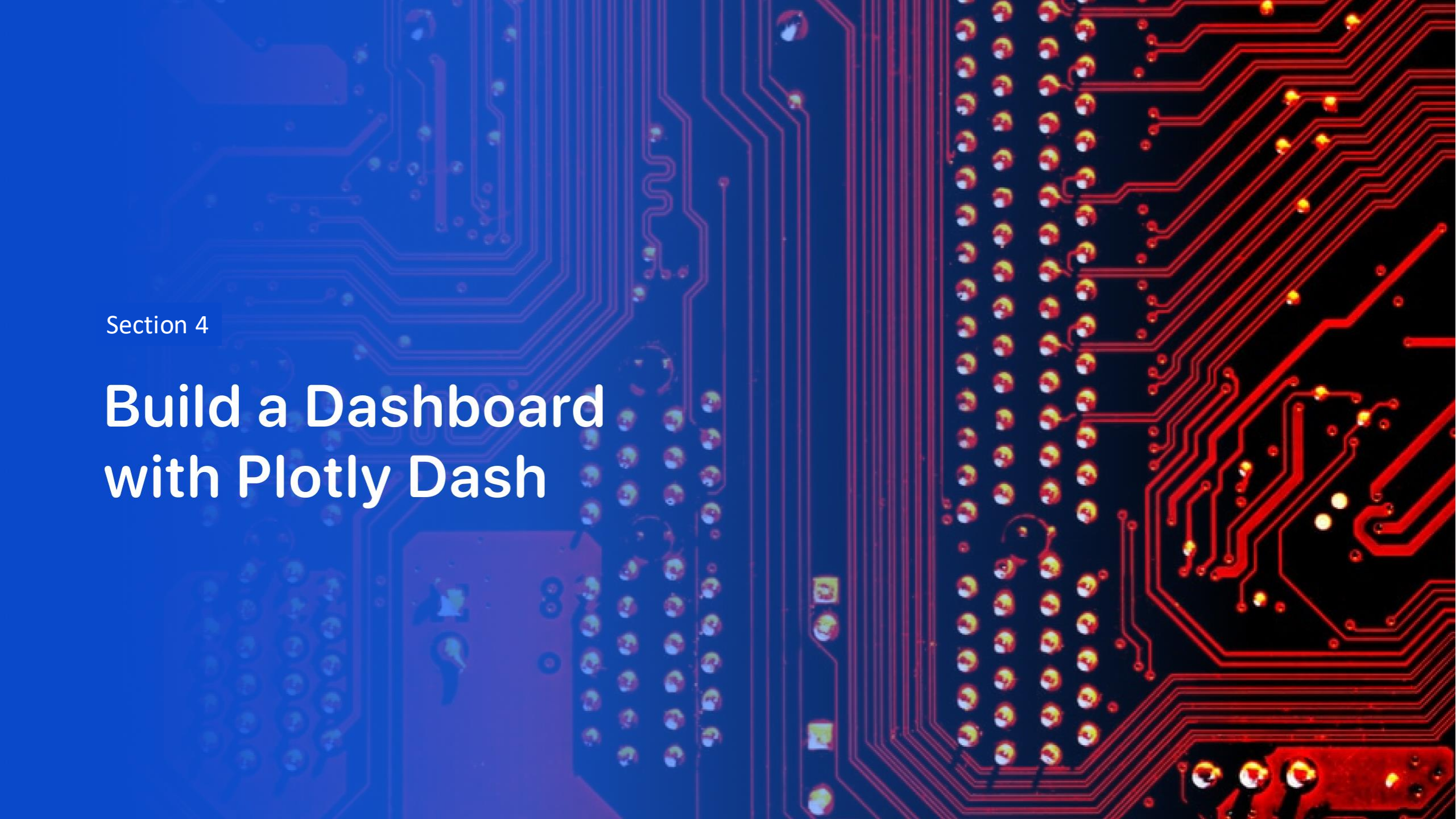
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

Launch Site Surroundings

CCAFS SLC-40

- 0.86 km from nearest coastline
  - helps ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- 21.96 km from nearest railway; 23.23 km from nearest city; 26.88 km from nearest highway
  - Transportation/Infrastructure and Cities: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.

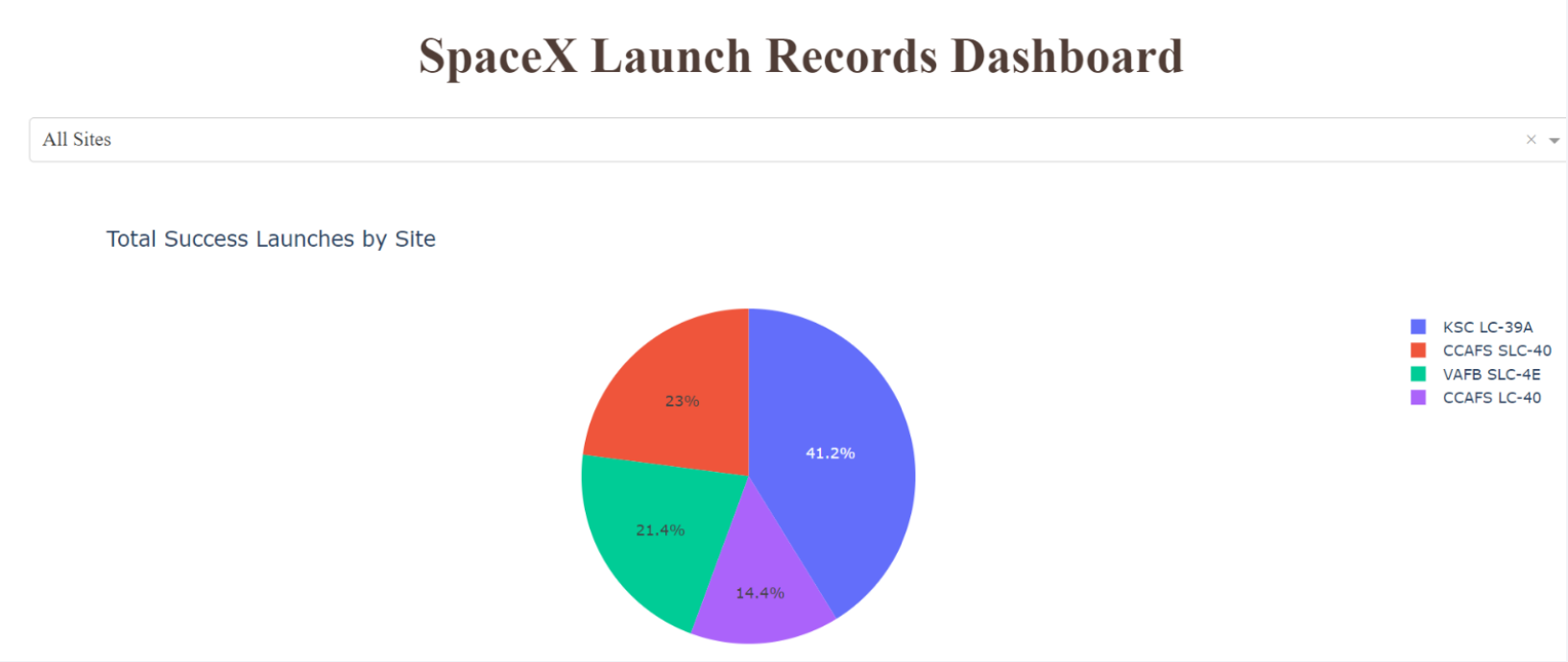



Section 4
Build a Dashboard with Plotly Dash

Site Launch Success

Success as Percent of Total

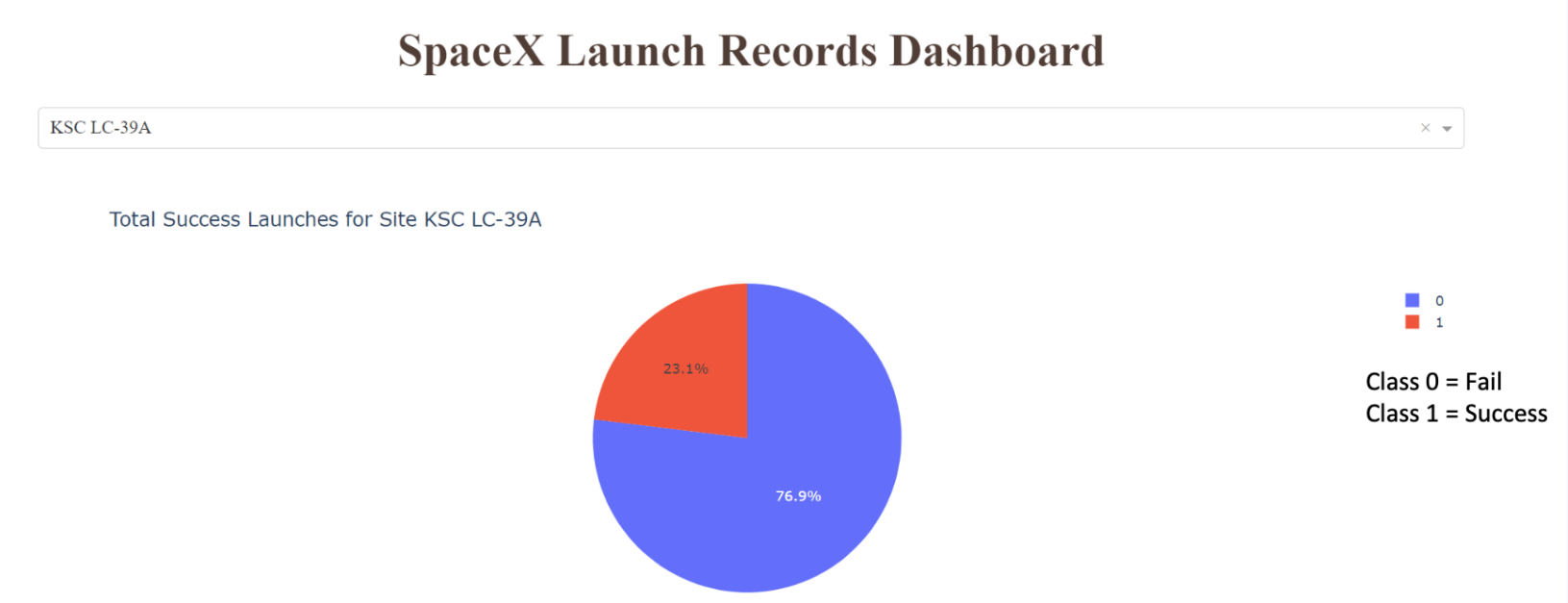
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



KSC LC-29A Launch Success

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



Payload and Mass Success

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Accuracy

- All the models performed at about the same level and had the same scores and
- accuracy. This is likely due to the small dataset. The Decision Tree model slightly
- outperformed the rest when looking at .best_score_
- .best_score_ is the average of all cv folds for a single combination of the parameters

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good

Confusion Matrix Outputs:

- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative

Precision = TP / (TP + FP)

- 12 / 15 = .80

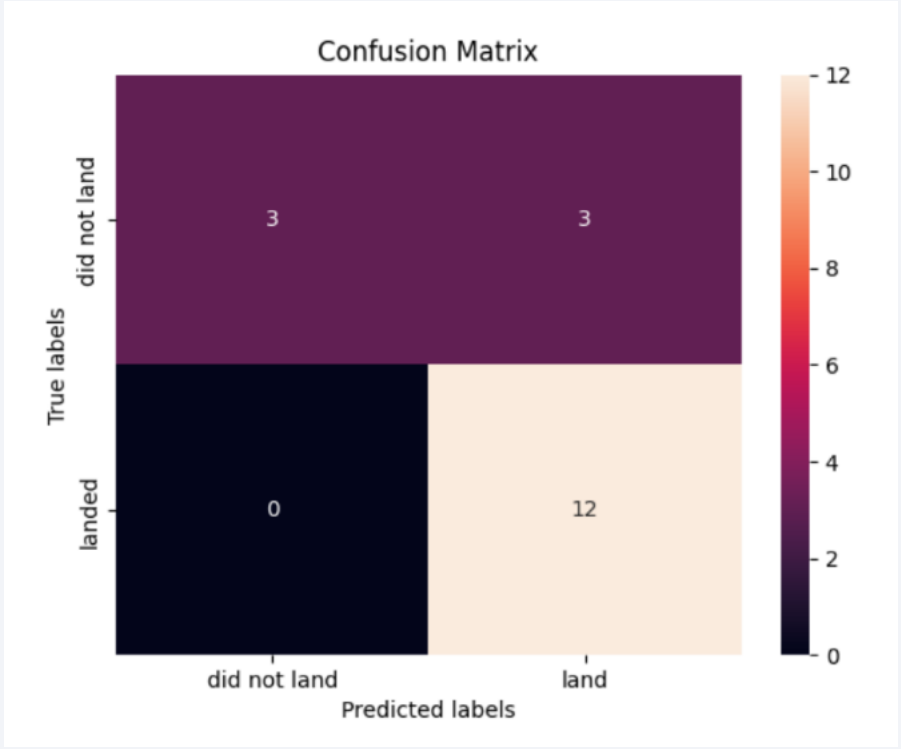
Recall = TP / (TP + FN)

- 12 / 12 = 1

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- 2 * (.8 * 1) / (.8 + 1) = .89

Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming
- Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- Coast: All the launch sites are close to the coast
- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!