IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ellen Faustine
March 24, 2025

Outline

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

Executive Summary

Summary of methodologies

- Data Collection with API
- Data Collection with WebScraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Visualization
- Interactive Visual Analytics with Plotly Dash
- Interactive Visual Analytics with Folium
- Predictive Analytics with Machine Learning

Summary of results

- Launch success consistently show an upward trend since 2013
- Launch sites are located on southern coastline near the equator and transportation systems
- A probable correlation exists between launch success and payload mass (kg)
- Certain orbit types such as SSO and VLEO outperform others
- The classifier with highest accuracy is decision tree

Introduction

- SpaceX is an American aerospace company founded by Elon Musk that has revolutionized the commercial space industry with its Falcon 9 rocket, a partially reusable, two-stage-to-orbit, medium-lift launch vehicle.
- SpaceX's Falcon 9 launches are advertised at a cost of $62 million, significantly lower than the $165 million or more by other providers.
- The primary reason for this cost difference is SpaceX's ability to recover and reuse the first stage of the Falcon 9 rocket after launch, which is achieved through a controlled vertical landing.
- The project goal is to explore the commercial space industry, particularly focusing on SpaceX's Falcon 9 and its cost-saving reuse of the first stage to be able to predict launch costs by determining if the first stage will land.

Introduction

Scenario

Estimate the first stage flight landing outcomes of Falcon 9 for competitors to bid against SpaceX. To determine the launch cost by analyzing first stage landing, we will use predictive analytics.

Process

To achieve this model, we will collect the SpaceX launch data with API and Web scraping.
Then, to prepare data for analysis, we apply data cleaning and wrangling to gather insights and finally combine multiple features to build a predictive model.

Questions

- What is the historical success of Falcon 9 landings?
- What factors lead to success or failure of a launch?
- Can a machine learning model forecast launch costs?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Requests from RESTFUL API endpoints
  - Web scraping tabular data from Wikipedia
- Perform data wrangling
  - Review attributes, remove columns, impute missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Explore and visualize data to discover relationships between variables
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mark launch sites and successful landings on geospatial map using Folium
  - Build an interactive Plotly Dash dashboard for analytics
- Perform predictive analysis using classification models
  - Build models with several classification methods and determine accuracy

Data Collection

1. SpaceX launch data with HTTP requests on API endpoints using:

   v4/launches/past

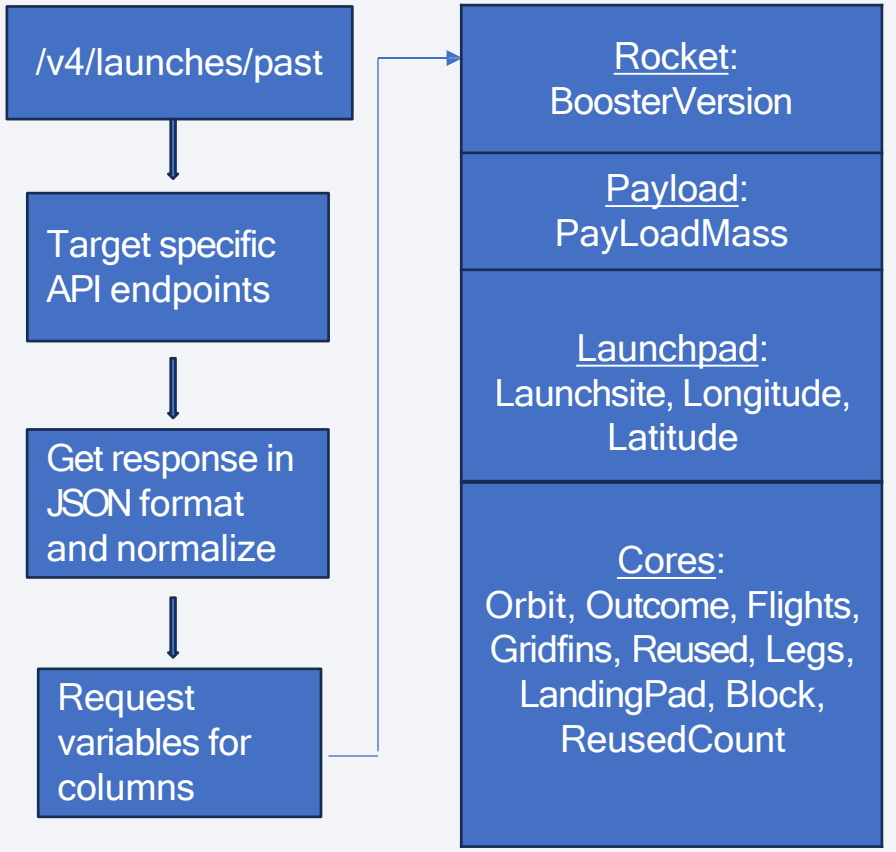
   v4/rocket, v4/payloads, v4/launchpad, v4/cores

2. Tabular data from List of Falcon 9 and Falcon Heavy launches on Wikipedia

Data Collection - SpaceX API

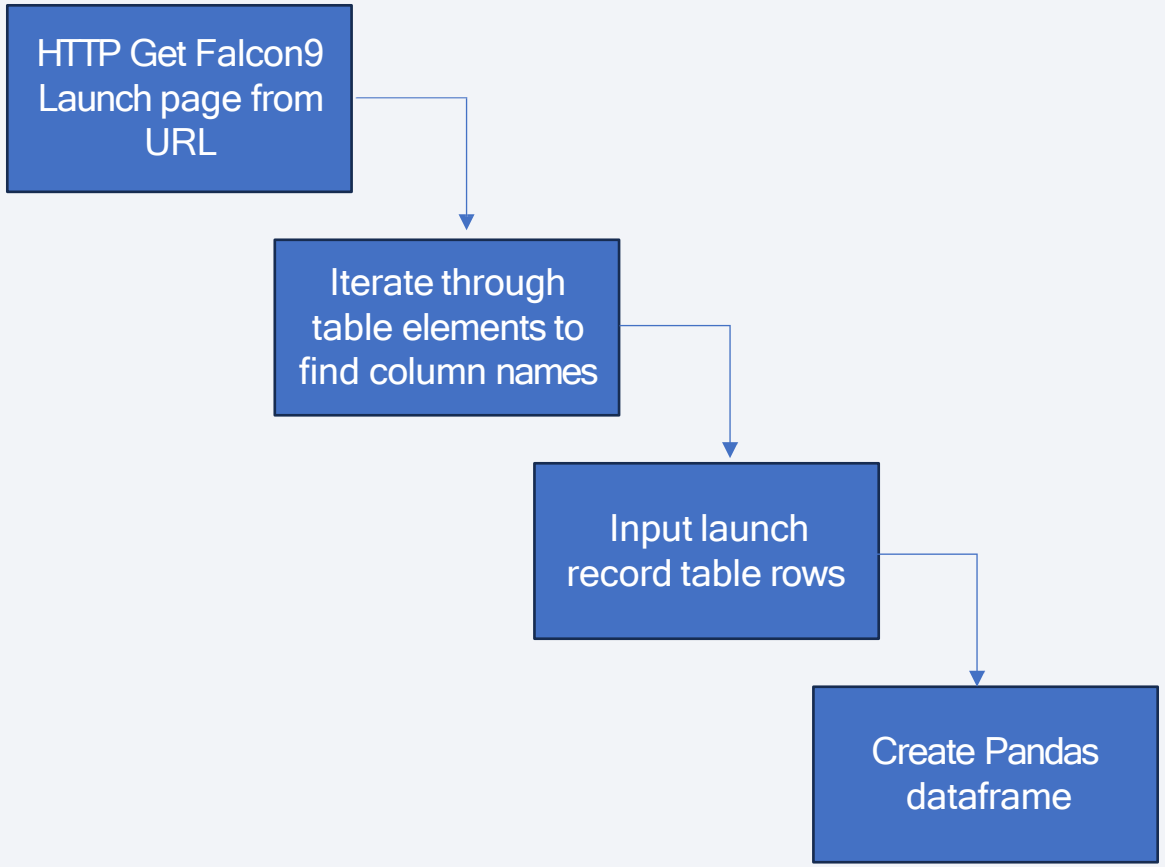
- Initial data collected from v4/launches/past
- Additional request data from v4/rocket, v4/payload, v4/launchpad, v4/ cores

- GitHub URL



Data Collection - Scraping

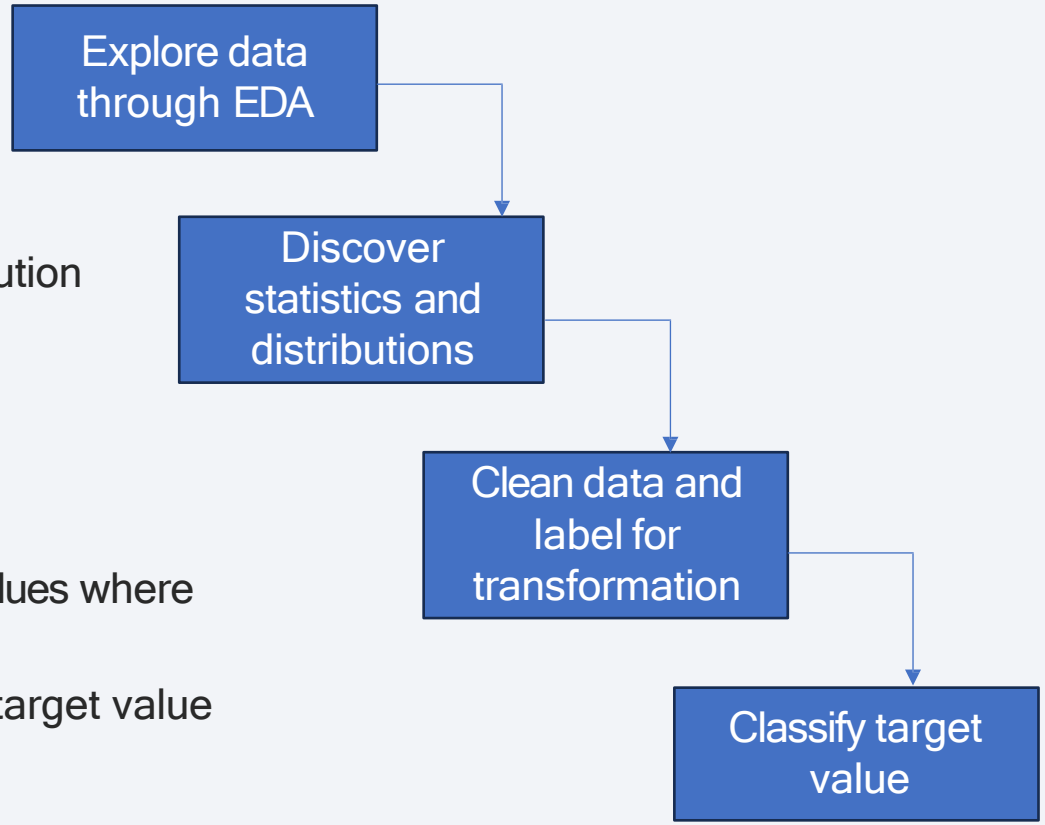
- Get request on HTML page as HTTP response
- Create BeautifulSoup object from response
- Parse through tables and convert into a Pandas dataframe
- Export to csv file
- GitHub URL



Data Wrangling

Review column attributes

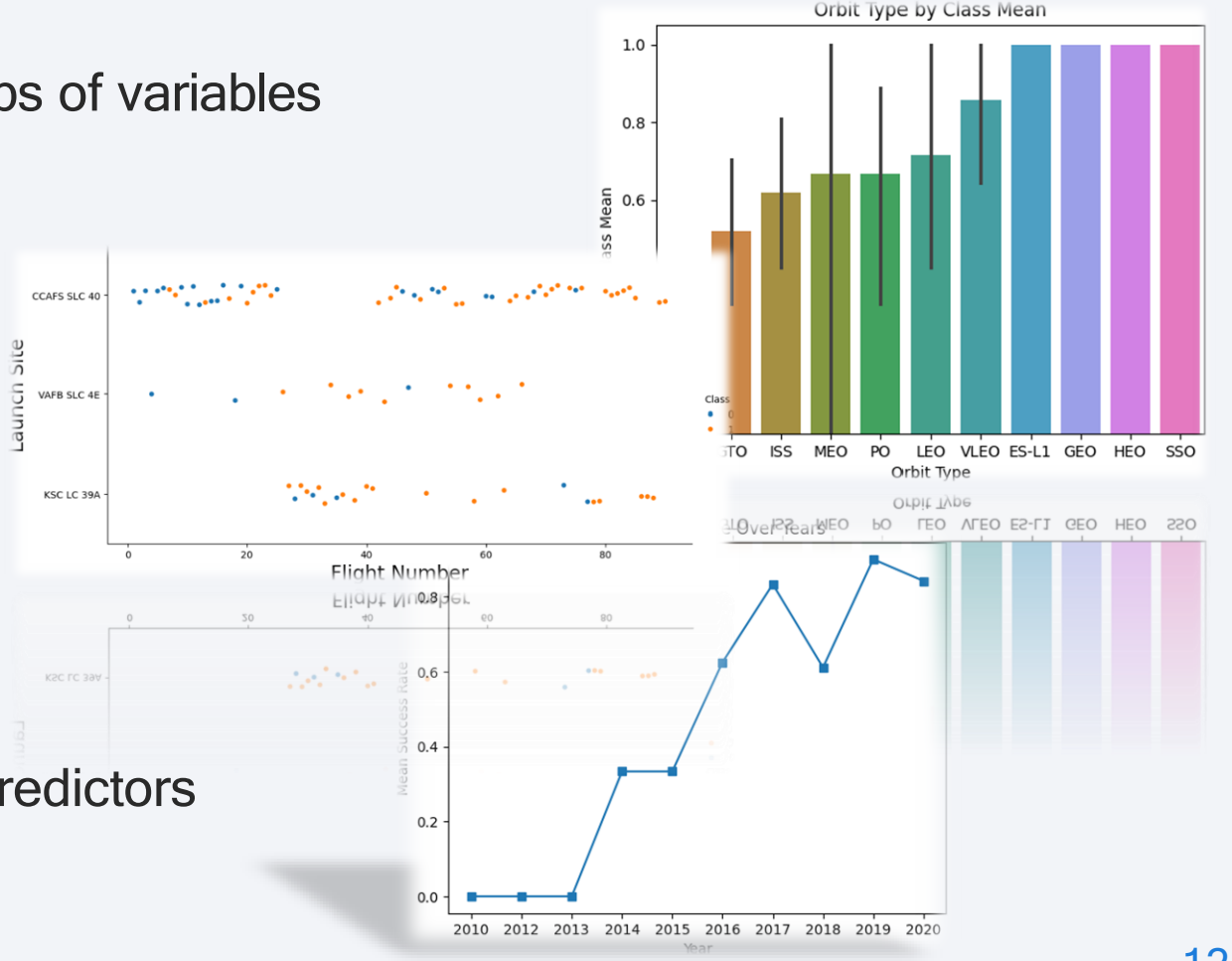
- Find percentage of missing values
- Identify numerical and categorical columns
- Discover number of launches per site
- Explore the orbit type and mission outcomes distribution
- Remove columns and input mean in missing values



- Convert categorical variables
  - Create landing_outcome label
  - Transform outcome into dummy variables (binary values where 1=successful, 0= failed)
  - Assign values to Class label to be used for training target value
- GitHub URL

EDA with Data Visualization

- To prepare data for predictive analytics:
- Scatterplots used to visualize relationships of variables within color encoded ‘Class’ feature
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
- Bar chart comparison
  - Success Rate of Each Orbit Type
- Line chart comparison over time
  - Mean Success Rate Over Years
- Create categorical dummy variables on predictors
  - Cast all numeric to float64
- GitHub URL



EDA with SQL

SQL Queries

- Display the unique names of launch sites
- Display the total payload mass carried by boosters launched by NASA
- Display the average payload mass carried by booster version F9 v1.1
- Date of first successful ground pad landing
- List names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Display the total number of successful and failure mission outcomes
- List names of booster versions which have carried the maximum payload mass
- Display the month names, landing outcomes, booster versions, launch site for failure landing outcomes in drone ship the months in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

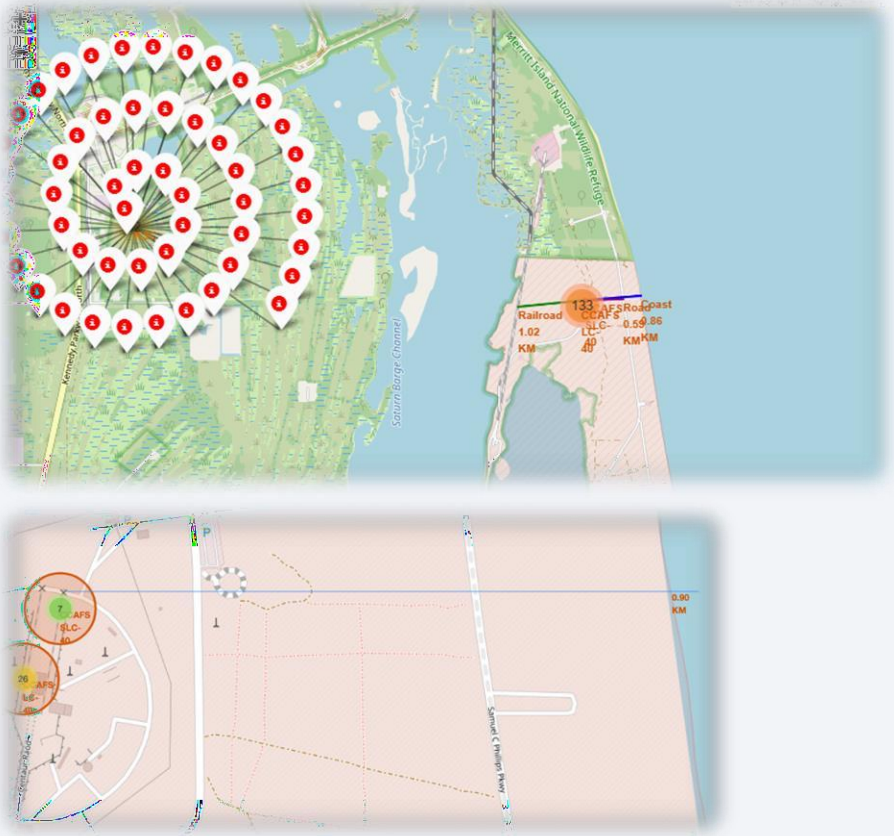
GitHub URL

Build an Interactive Map with Folium

- Locate launch sites in geospatial map
- Color encoded folium.Marker() to mark ‘Class’ (success/failure) of each launch site is added to marker cluster
  - If Class=1, marker color is green, if Class=0 marker color is red
- Use folium.Circle() to highlight launch site coordinates
- Calculate distance and proximity using folium.PolyLine()
  - Creates a distance line to the closest city, railway, and highway

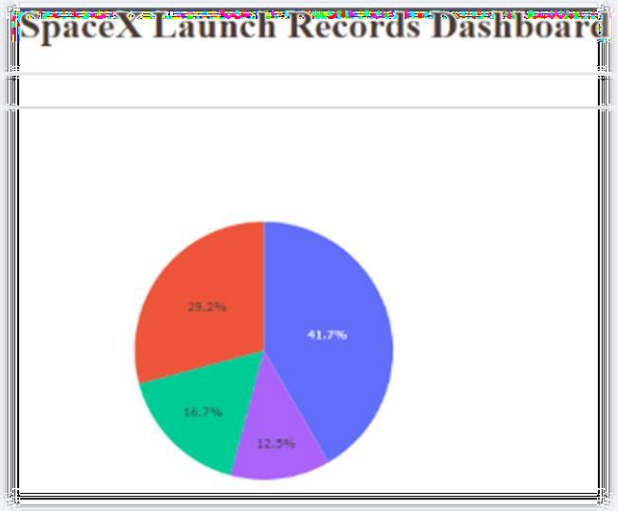
➢ Answer questions (e.g., Do launch sites keep a distance from certain landmarks and cities?)

- GitHub URL



Build a Dashboard with Plotly Dash

- Application includes input components
  - Launch site drop-down
  - Range slider to select payload
  - Callback function on selected inputs to render visualization
- Interactive visuals include charts
  - Pie chart showing success launches by launch sites (individual and total)
  - Scatter chart showing successful launches by payload mass with ‘booster version category’ color-coded



- GitHub URL

Predictive Analysis (Classification)

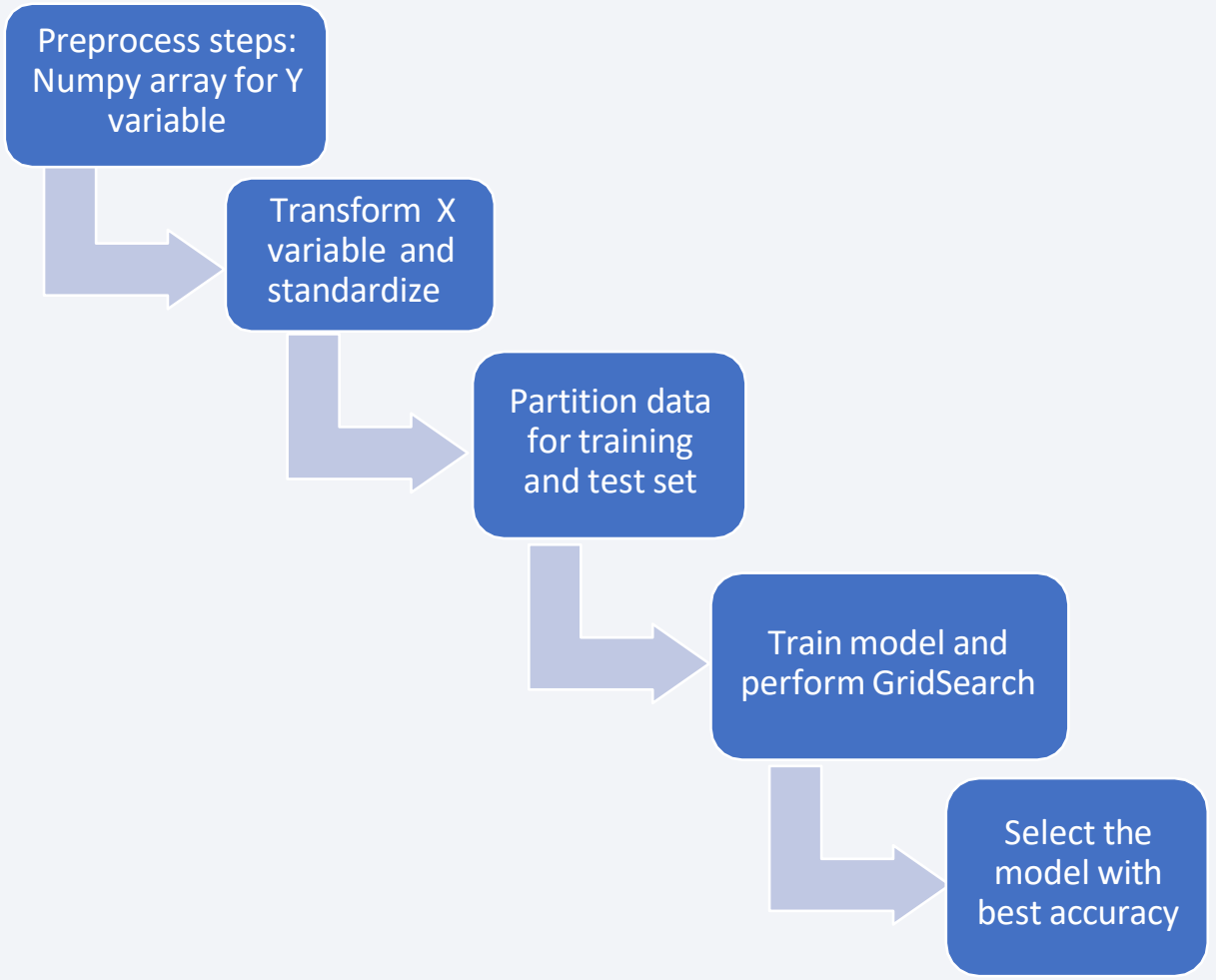
Classifiers: Logistic Regression, Classification Trees, K-Nearest Neighbor, Support Vector Machine

Steps To Building a Machine Learning Model

- Preprocessing: Transform, label, and partition data
  - Assign Class to Y using NumPy array
  - Apply StandardScalar(), fit_transform() to X
  - Split into training/test sets

```
X train/test set: (72, 83) (18, 83)
Y train/test set: (72,) (18,)
```

- Training: Tuning Hyperparameters/Cross-Validation
  - Use GridSearchCV evaluate the optimal parameters
  - Choose the model with best accuracy
- GitHub URL



Results

Exploratory data analysis *finds hidden patterns with data affecting success rate*

- CCAFS LC-40 has the greatest number of flights and a success rate of 60% and others is 77%
- However, for payload mass above 10,000 kg, CCAFS LC-40 has a success rate of 100%

Interactive analytics *such as Dash and Folium maps*

- Launch sites usually fall within close proximity to coastlines and highways
- Different boosters perform better with different payload mass

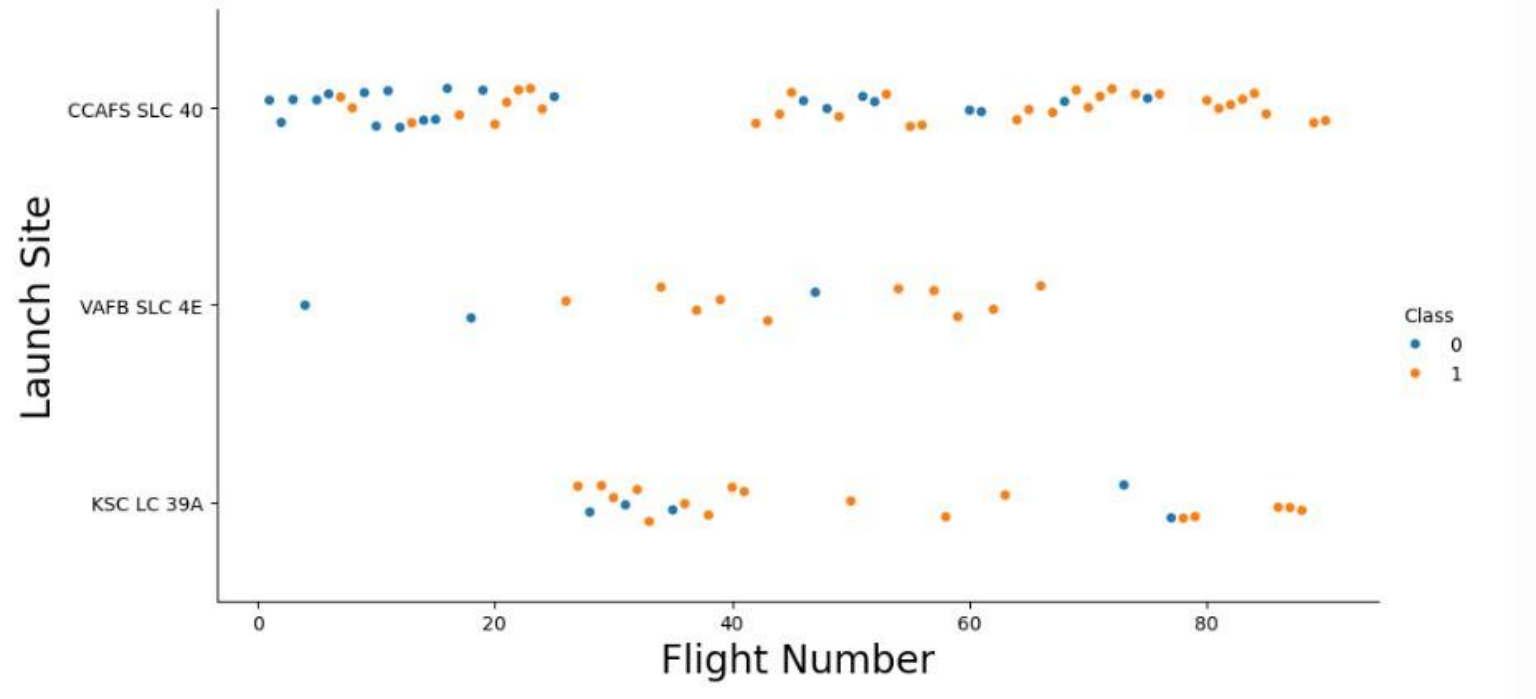
Predictive analysis *uses machine learning classifiers and distinguishes the best fitting model*

- All classifiers performed similarly to the test dataset at 0.83 accuracy but decision tree training accuracy was 0.94, the highest of all the methods

Section 2

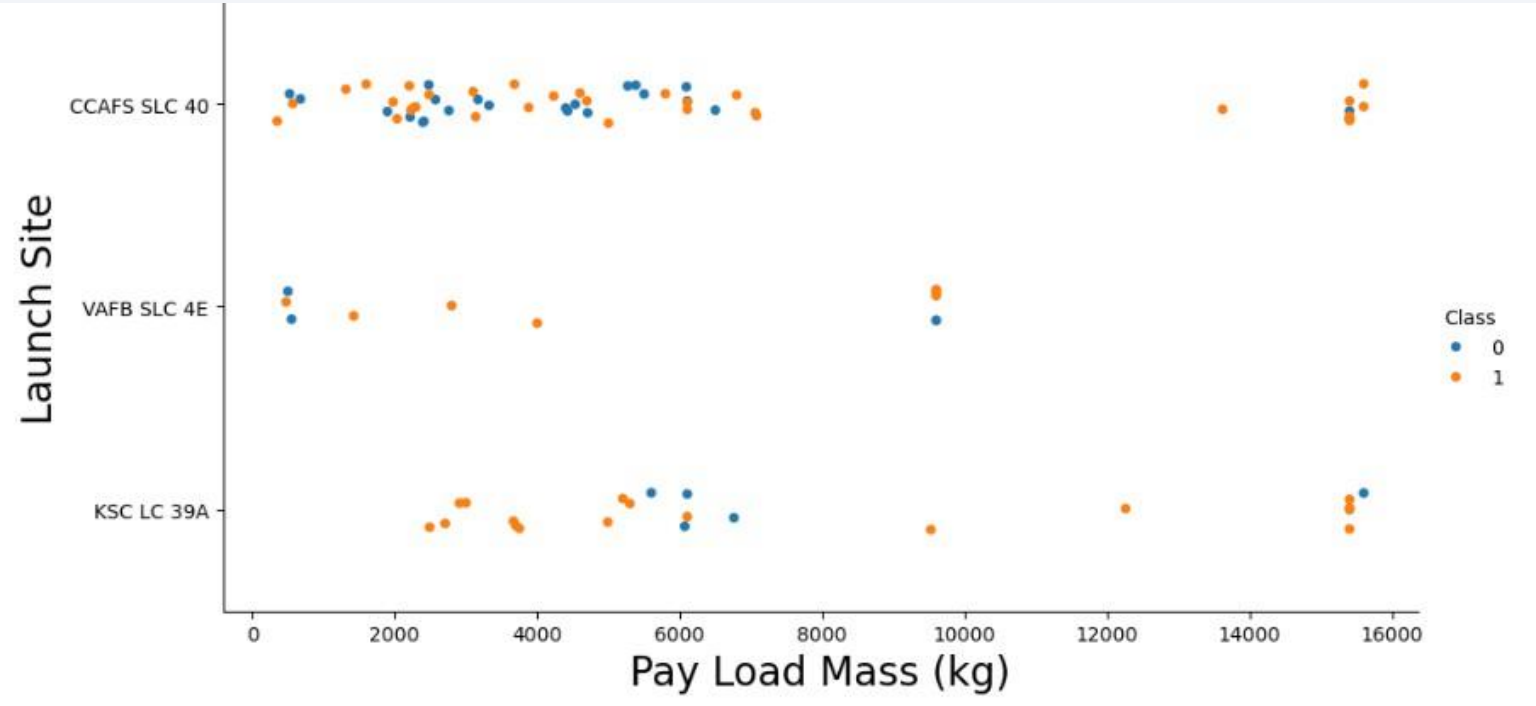
Insights drawn from EDA

Flight Number vs. Launch Site



- There seems to be a correlation of landing success and flight number as indicated by orange dots.
- SLC 40 shows the highest number of flights, while SLC 4E shows the least.
- All launch sites showed higher failure rates at the beginning but improved over time.

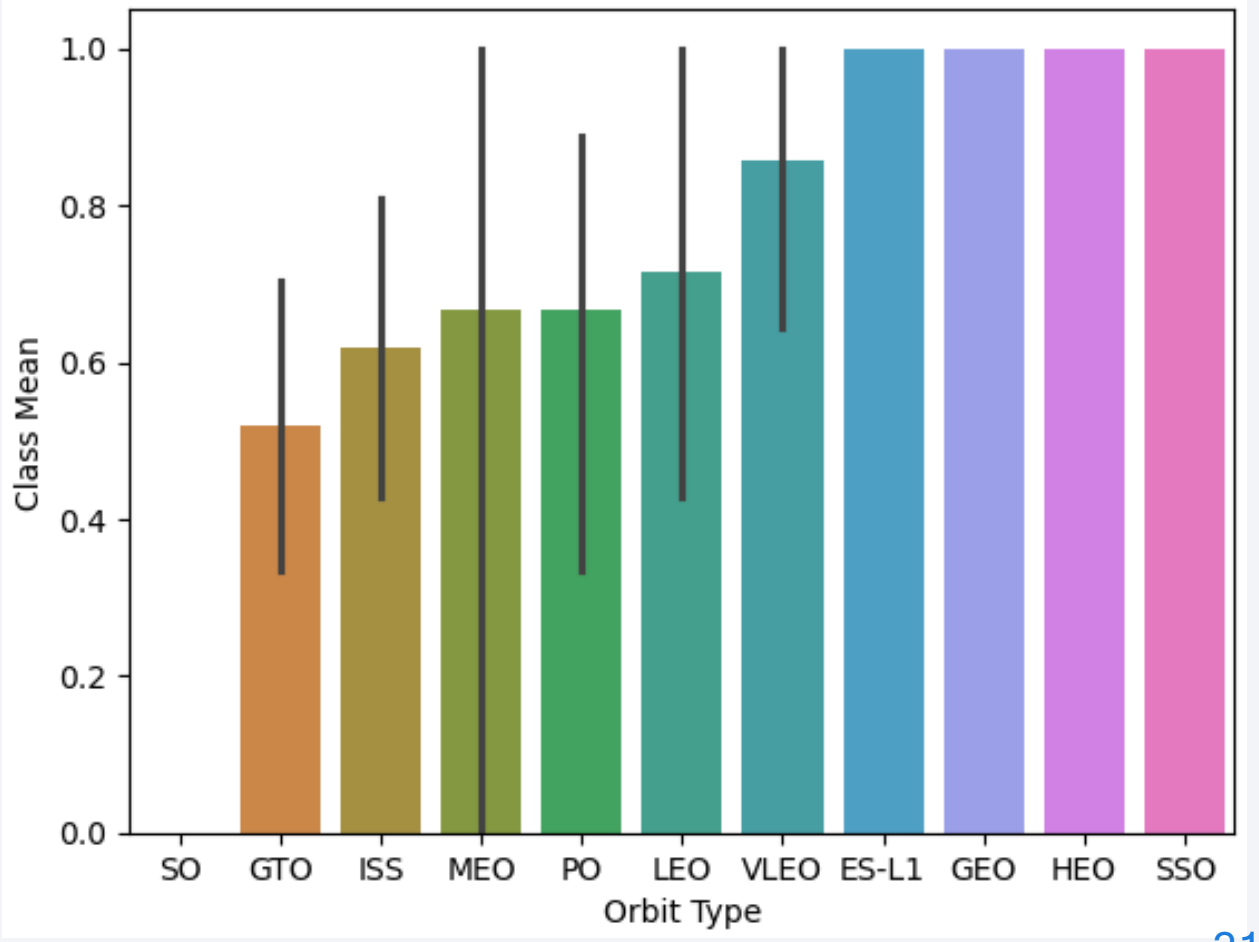
Payload vs. Launch Site



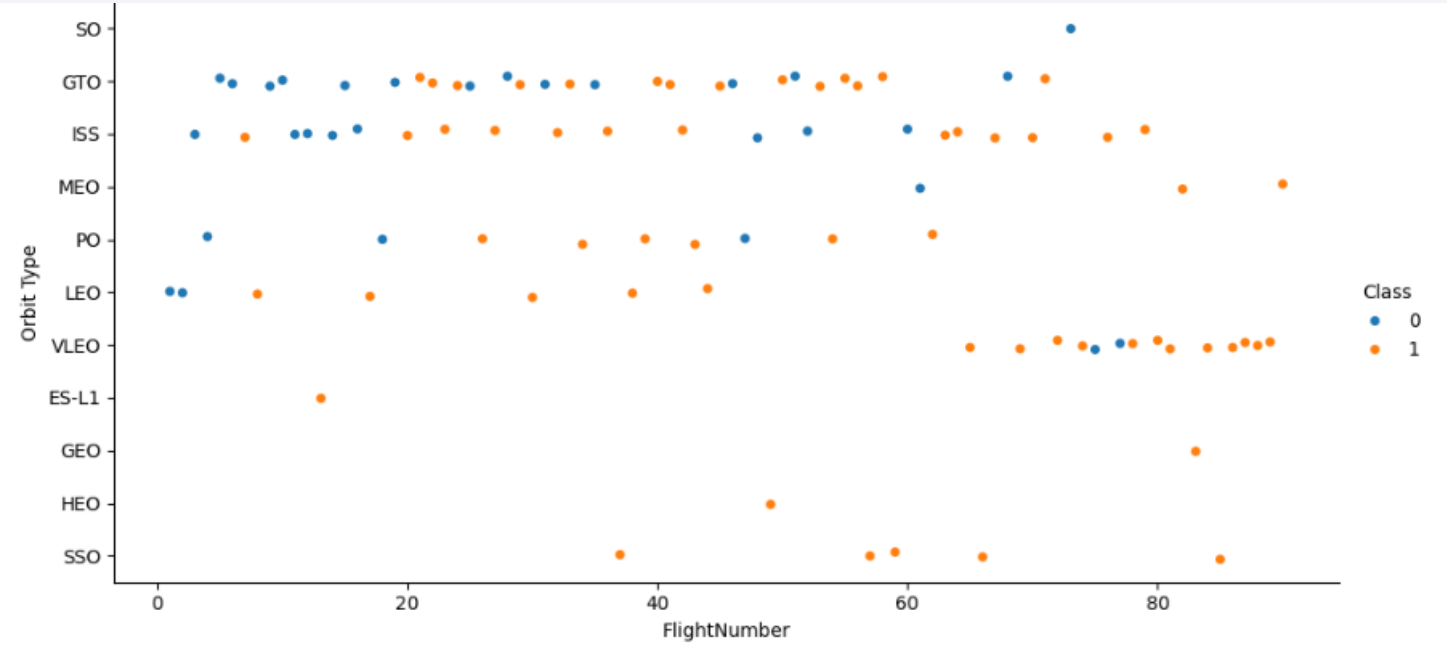
- The majority of payload attempts by SLC 40 fall under 7000 kg with a few around 16000 kg.
- The success rate for SLC 40 is evenly mixed under 7000 kg but for16000 kg it overcomes failure by 5x.
- The data slightly correlates higher payload mass to success which may be due to improvements from earlier.

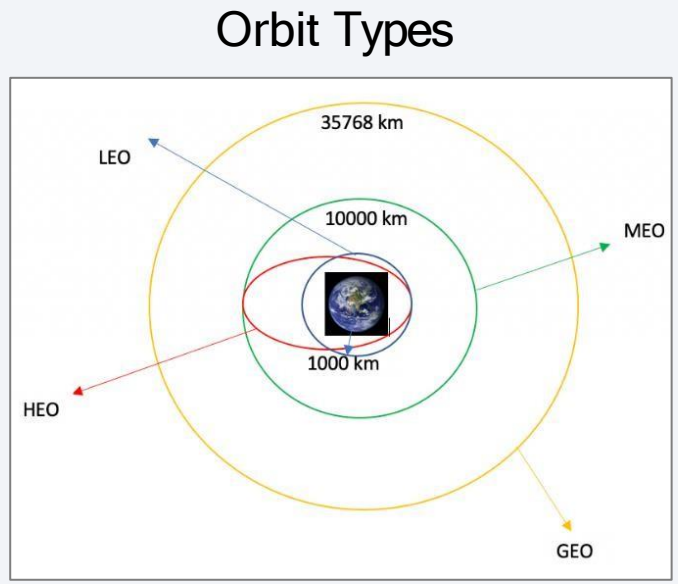
Success Rate vs. Orbit Type

- VLEO has the best success rate having a high sample size and 0.85 mean.
- SSO had only 5 flights with 100% success rate.
- GTO and ISS had the lowest success mean yet the highest number of flights.
- HEO, GEO, and ES-L1 had 1 flight to account for the success rate.
- SO had 1 flight and lacks data.
- MEO had only 3 flights and average success.
- PO had LEO had moderate sample of 7 and 9 flights.



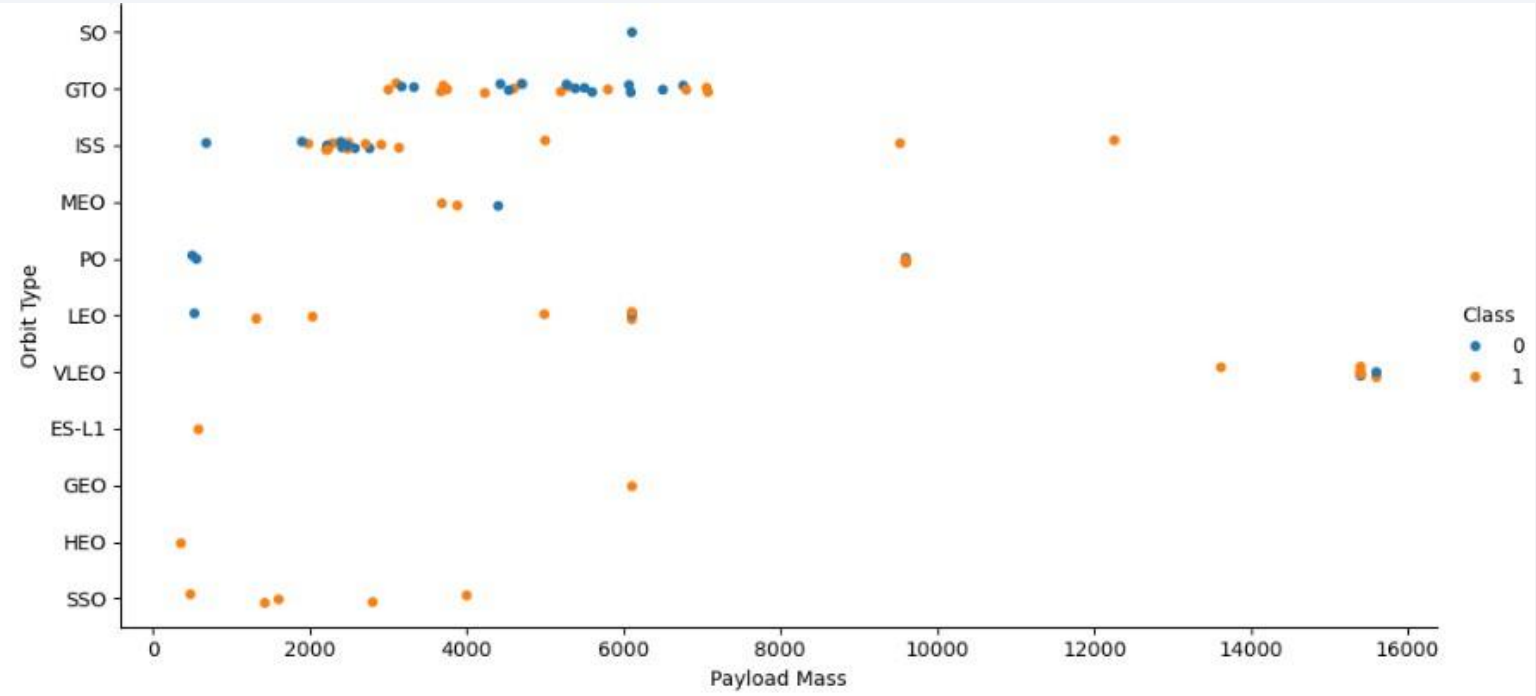
Flight Number vs. Orbit Type





- In the LEO orbit, success is substantially increased by number of flights whereas for GTO it is not.
- Some orbits have 100% success such as SSO, but begin at a later flight number.
- VLEO started launch past 60 flights and was successful the majority of flights except in the middle.

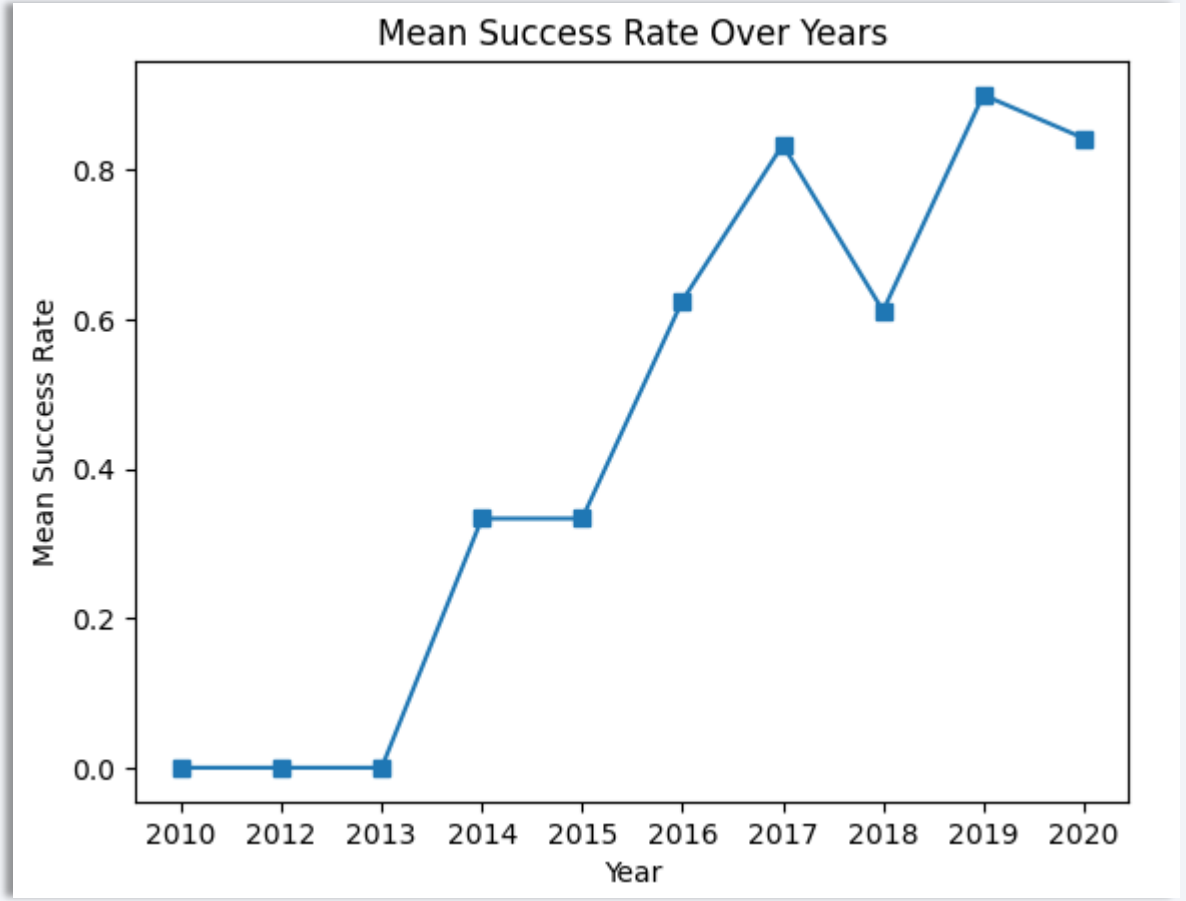
Payload vs. Orbit Type



- Similar to launch site, the majority of orbits had less than 7000 kg payload.
- For GTO and ISS there is equal success and failure by payload range.
- SSO payload mass is exclusively on the lower end, under 4000 kg, with 100% success
- VLEO seem exclusively higher payload at above 14000 kg with equal distribution of success.

Launch Success Yearly Trend

- SpaceX shows an overall trend of rising success rates beginning in 2013.
- The steepest growth were from 2015 – 2017 improving twice as much as before.
- There was a dip in 2018, but another improvement for next year 2019 marked its highest.
- The line chart shows consistent improvement of success rate over time.



All Launch Site Names

- Names of unique launch sites

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

%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;

Launch Site Names Begin with 'CCA'

- List of 5 records where launch sites begin with `CCA`

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

%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

Total Payload Mass

- The total payload (kg) carried by boosters from NASA is 45,596

SUM(PAYLOAD_MASS__KG_)
45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
   FROM SPACEXTBL \
   WHERE CUSTOMER = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2,928.4 kg

AVG(PAYLOAD_MASS__KG_)
2928.4

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

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on December 22, 2015

MIN(DATE)
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of boosters that landed on drone ship successfully and had payload mass greater than 4000 but less than 6000

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

```
%sql SELECT * FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' \
   AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

```
%sql SELECT Mission_Outcome, COUNT(*) AS 'Count' FROM SPACEXTBL \
   GROUP BY Mission_Outcome;
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT BOOSTER_VERSION \
   FROM SPACEXTBL \
   WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
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

2015 Launch Records

- List of failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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

```
%sql SELECT substr(Date, 6,2) AS 'Month', Date, Booster_Version, Launch_Site,
Landing_Outcome  FROM SPACEXTBL \
   WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

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

```
%sql SELECT [Landing_Outcome], COUNT(*) AS 'Count' FROM
SPACEXTBL \
   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 (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

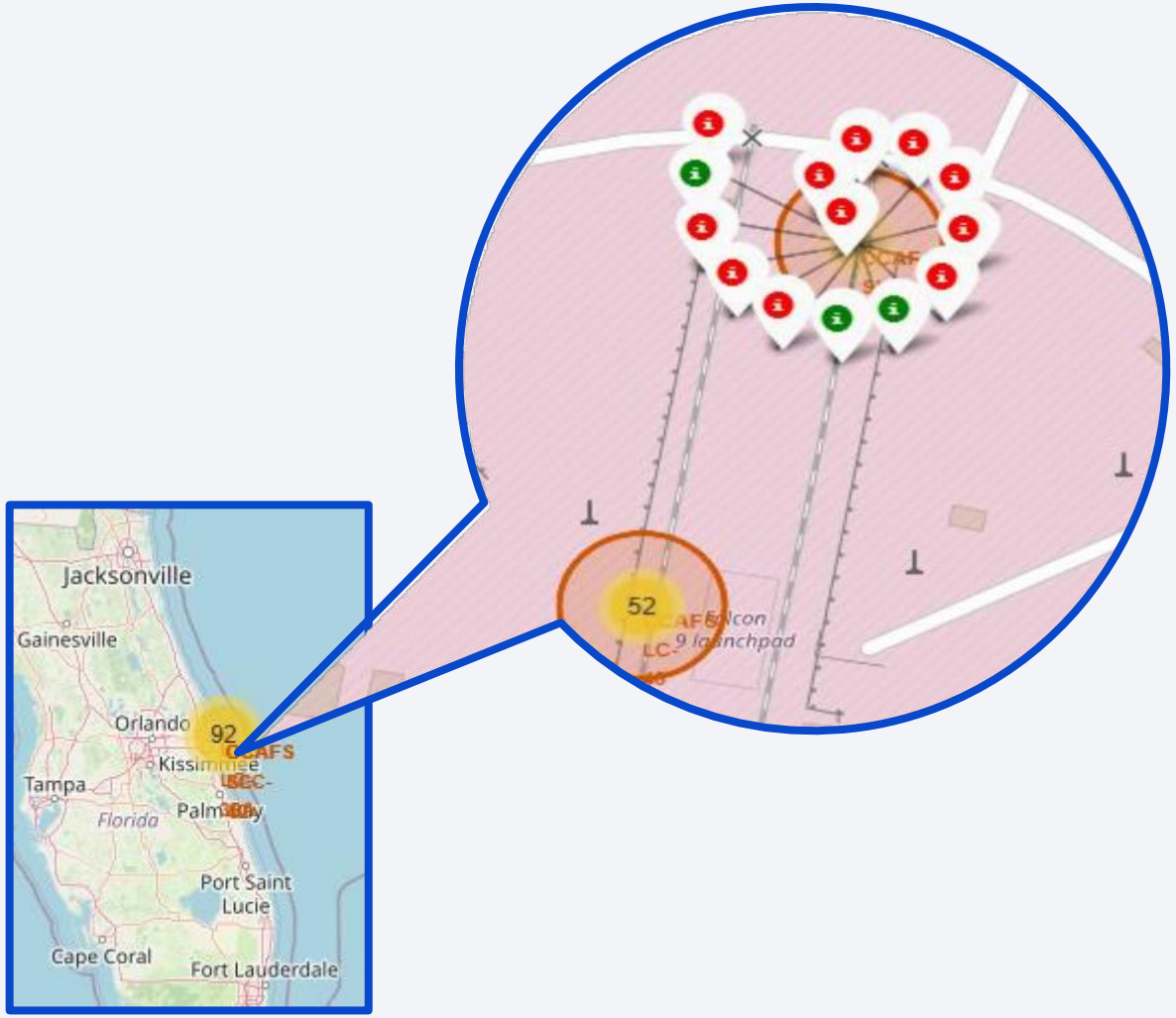
Launch Sites Proximities Analysis

Launch Sites Locations

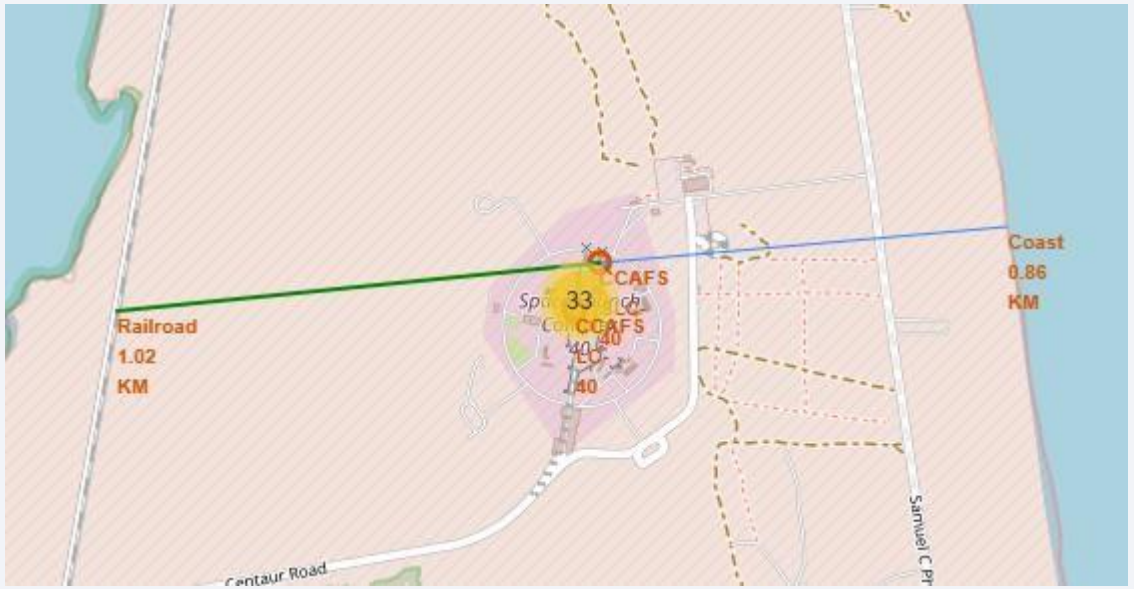
- Launch sites are located on the coastal regions of SW California and SE Florida
- Coastal sites prepare for failure debris to crash into the ocean and mitigate civilian risk
- Vicinity to the equator takes advantage of the earth's rotational speed for launch boost and communication satellites

Launch Sites Outcomes

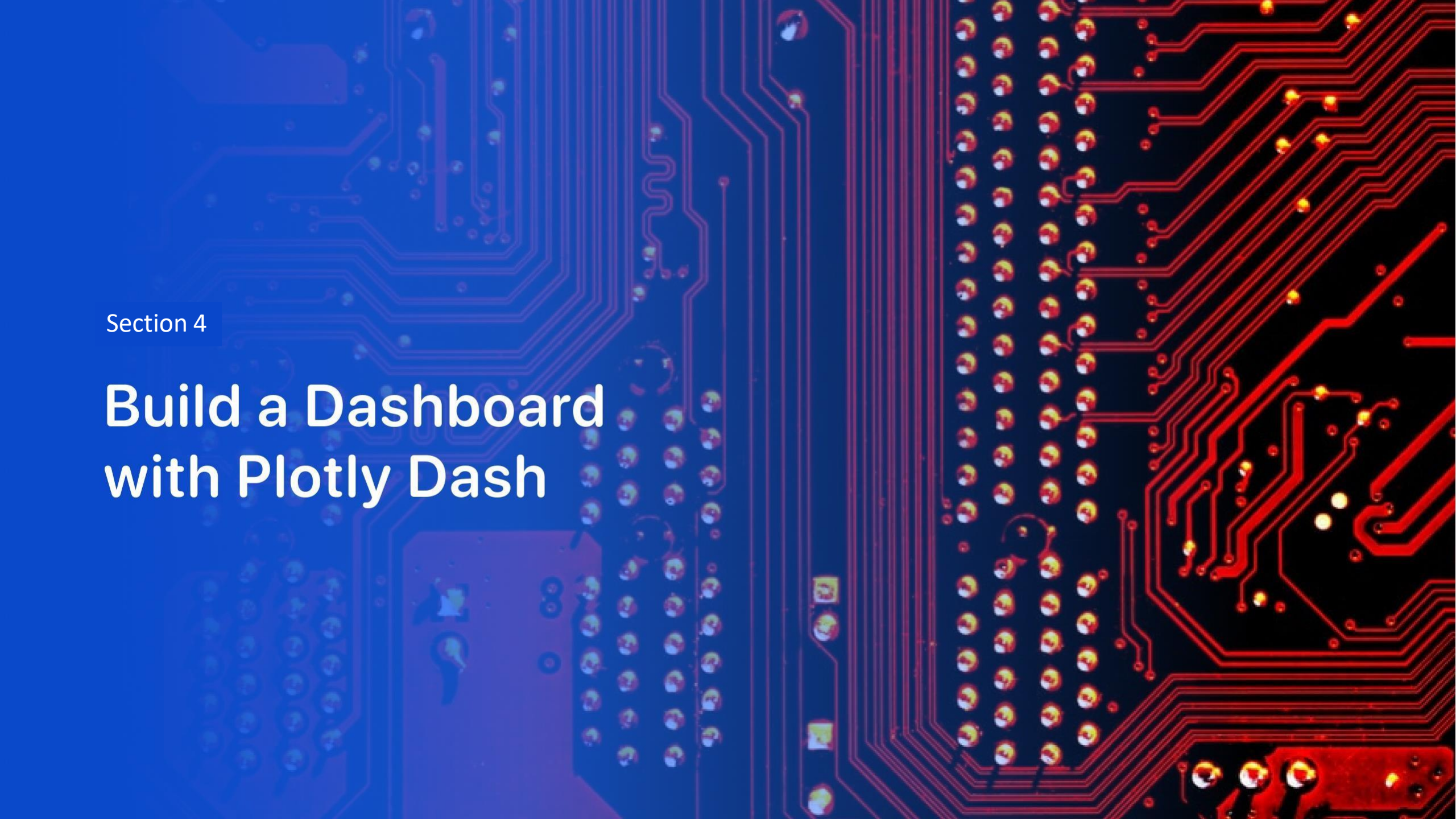
- Each launch site is marked by success/failure
- Green markers denote success and red markers denote failure
- An example is shown for CCAFS-SLC 40
- Allow for identification of launch outcomes by launch site and geolocation



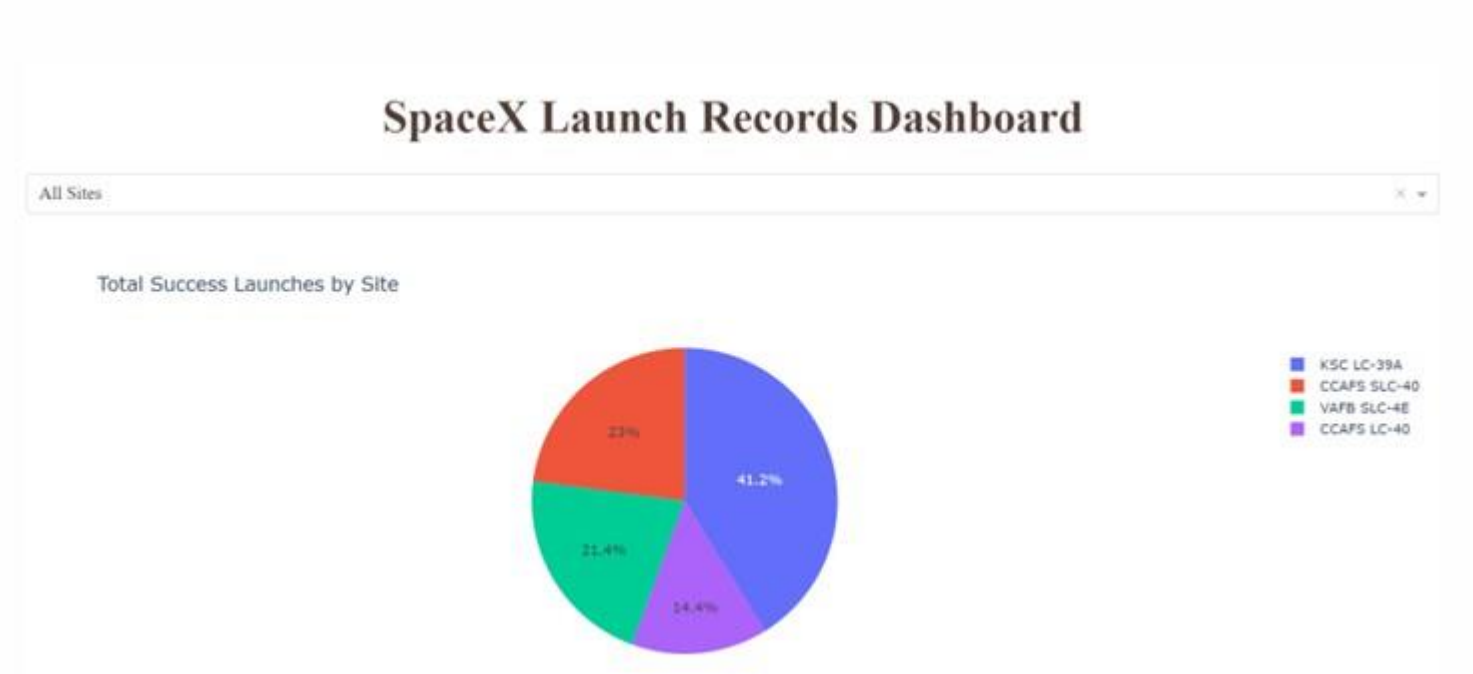
Launch Sites Proximities



- Launch site proximities to nearby railways, highways, and landmarks help with transportation
- Example CCAFS SLC-40 with distance of 0.86 km to the coast and1.02 km to the railroad
- Launch site location must calculate distances for most efficiency cost and convenience

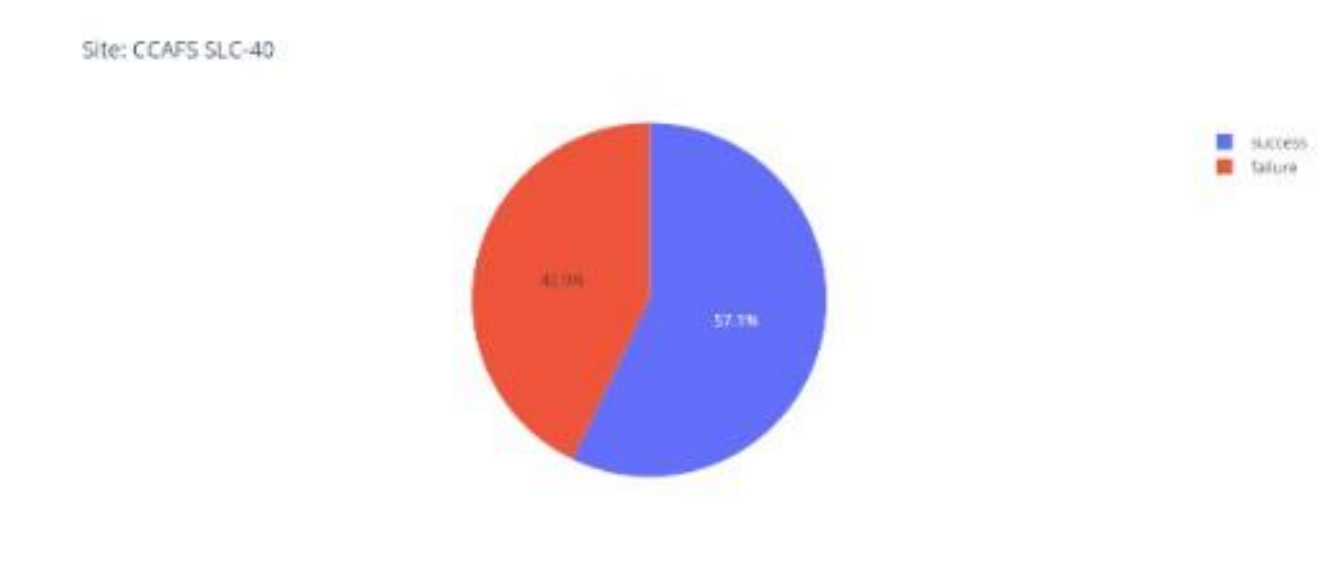

Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Sites



- Chart shows breakdown of success rate by launch site however no information on number of flights

Launch Site SLC-40



- Chart shows breakdown of success rate of launch site SLC-40

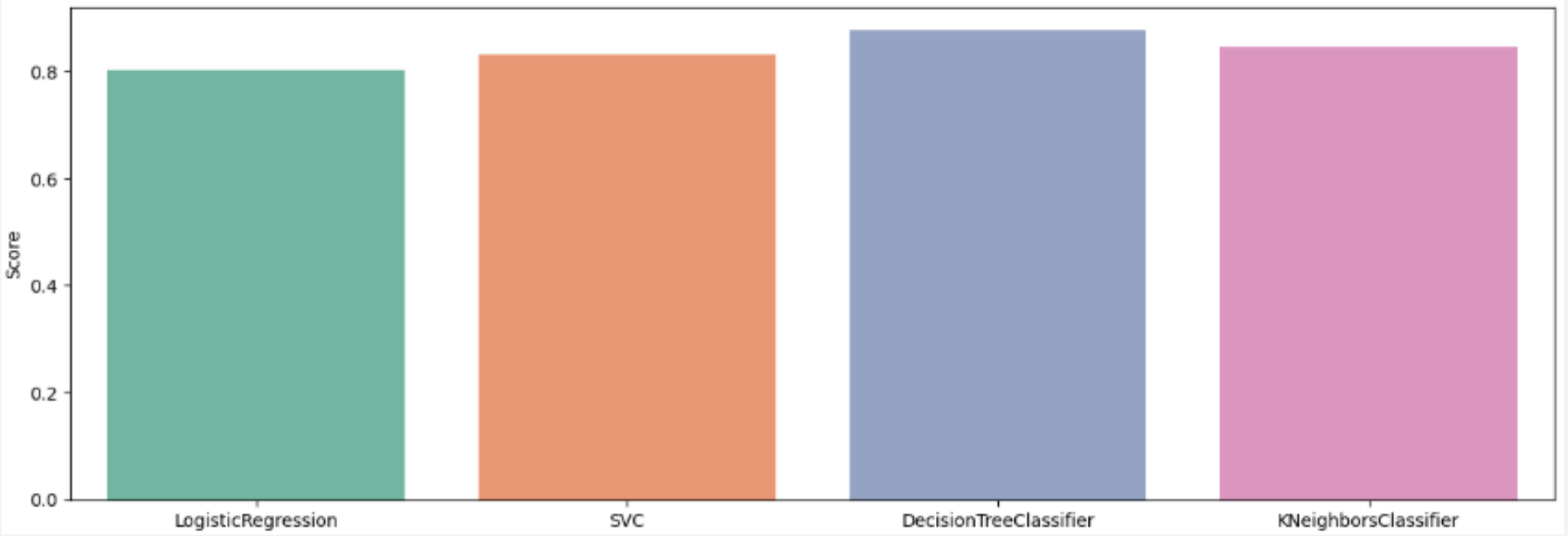
Payload Mass and Booster Version

- Booster version category (F1) are more correlated with success and (v1.1) has more failures
- 2k – 4k payload mass range has an equal distribution of success/failure while 4k-6k shows more failure volume
- The failure volume is highest under 2k payload mass (kg)

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- Bar chart shows overall score of each classifier type
- Every classifier was tested for accuracy, recall, precision, f1, and ROC AUC
- Decision Tree was shown as the best performing one

Confusion Matrix

Decision Tree Classifier

- Correctly identified17/18
- 94% accuracy
- Only 1 false positive

Breakdown:

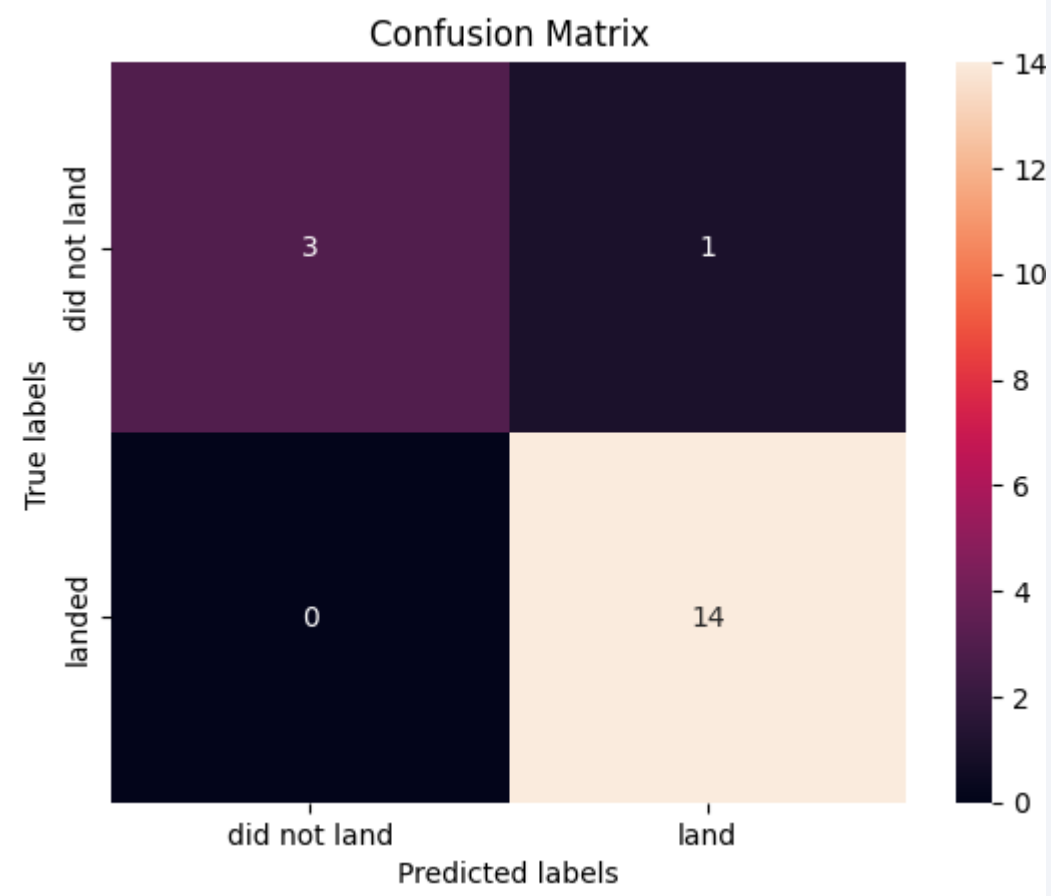
True positives (14)

False positives (1)

True negatives (0)

False negatives (0)

- Future improvement needed on Type 1 Error affecting precision of model



Conclusions

- Total landing outcomes favor success over failure/none by approximately 2:1 with success rates increasing over time
- Payload mass on the upper range seems to perform better than lower payload mass, but it could be due to improving operational experience starting at a higher payload mass
- SLC-40 launch location has a near perfect success rate after 80 number of attempts
- Statistically, VLEO (very low earth orbit) with high flight volume was more successful than the others
- SSO and VLEO both had high success rates whereas SSO is exclusively low payload and VLEO is high payload mass (kg)
- While all the classification types' predicted accuracy is high at 0.83 accuracy, decision tree results in the highest accuracy
- With the Decision Tree classifier, we can predict with 94% accuracy the landing outcome of SpaceX Falcon 9

Appendix

- Data collection dataset.csv
- List of Falcon 9 and Falcon Heavy launches Wikipedia
- Cleaned dataset datawrangling.csv
- Data visualization output dataviz.csv
- Space X site location geolaunch.csv

Thank you!