SpaceX Falcon 9 First Stage Landing Prediction

JG
IBM Data Science Capstone Project

December 2024

Outline

Executive Summary

- This presentation summarizes results from data collection and analysis of data on SpaceX first stage launches.
- Results demonstrate first stage launches may be predicted from common machine learning classification models.
- The presentation contents includes methods and detailed results to inform predicting SpaceX first stage launches:
  - Data collection, wrangling, and pre-processing
  - Data analysis with SQL and Visualization
  - Interactive visual analytics and dashboarding
  - Predictive Analytics with Machine Learning

Introduction

- Space exploration offers unique opportunities for research & development innovations.
- However, it is prohibitively expensive.
- SpaceX advertises Falcon 9 rocket launches at ~37% of the costs compared to other providers.
- Reduced costs for SpaceX launches are generated from savings gained by reusing the rockets' first stage.

Business Opportunity

- Other providers may offer competitive bids when reusing the first stage.
- To reuse the first stage, there must be a successful first stage landing.
- Therefore, the cost of a launch can be determined based on successful landing of the first stage.

Introduction: Project Objective

Business Opportunity

- Other providers may offer competitive bids when reusing the first stage.
- To reuse the first stage, there must be a successful first stage landing.
- Therefore, the cost of a launch can be determined based on successful landing of the first stage.

Project Objective

This project will analyze publicly available Falcon 9 data to assess characteristics of first stage launches to predict successful landings.

Information gleaned from this analysis and prediction will inform proposals competing with SpaceX bids.

Section 1

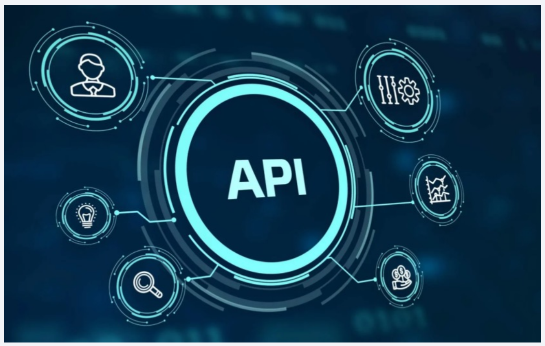
Methodology

Methodology | Executive Summary

Methods will describe:

- Data collection
- Data wrangling processes
- Exploratory data analysis (EDA)
- Interactive visual analytics
- Predictive analysis using classification models

Data Collection



WIKIPEDIA
The Free Encyclopedia
Search Wikipedia
Search

List of Falcon 9 and Falcon Heavy launches

17 languages

Article Talk Read Edit View history Tools

Contents hide

(Top)
Launch statistics
Past launches
Future launches
Notable launches
See also
Notes
References

From Wikipedia, the free encyclopedia

As of 17 December 2024, rockets from the Falcon 9 family have been launched 423 times, with 420 full mission successes, three failures,[a] and one partial failure. Designed and operated by SpaceX, the Falcon 9 family includes the retired versions Falcon 9 v1.0, v1.1, and v1.2 "Full Thrust" (blocks 3 and 4), along with the active Block 5 evolution. Falcon Heavy is a heavy-lift derivative of Falcon 9, combining a strengthened central core with two Falcon 9 first stages as side boosters.[1]

The Falcon design features reusable first-stage boosters, which land either on a ground pad near the launch site or on a drone ship at sea.[2] In December 2015, Falcon 9 became the first rocket to

Left to right: Falcon 9 v1.0, v1.1, v1.2 "Full Thrust", Falcon 9 Block 5, Falcon Heavy, and Falcon Heavy Block 5.

API

https://api.spacexdata.com/v4/launches/past

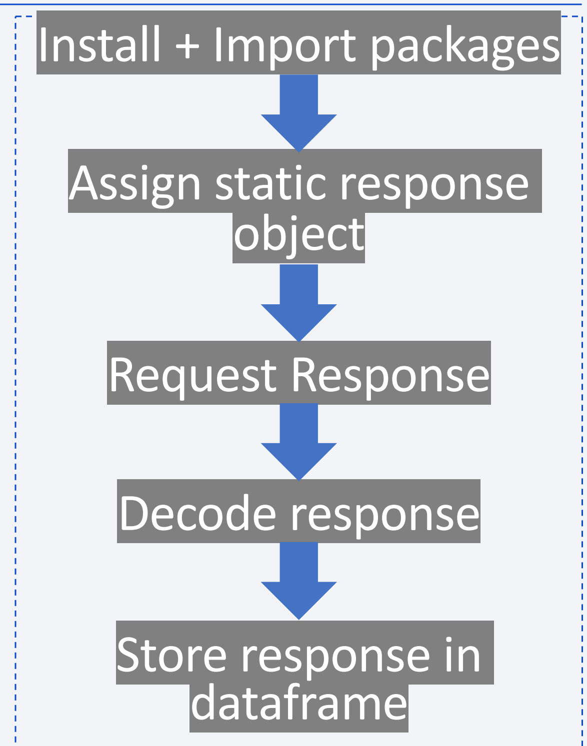
Webscraping

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

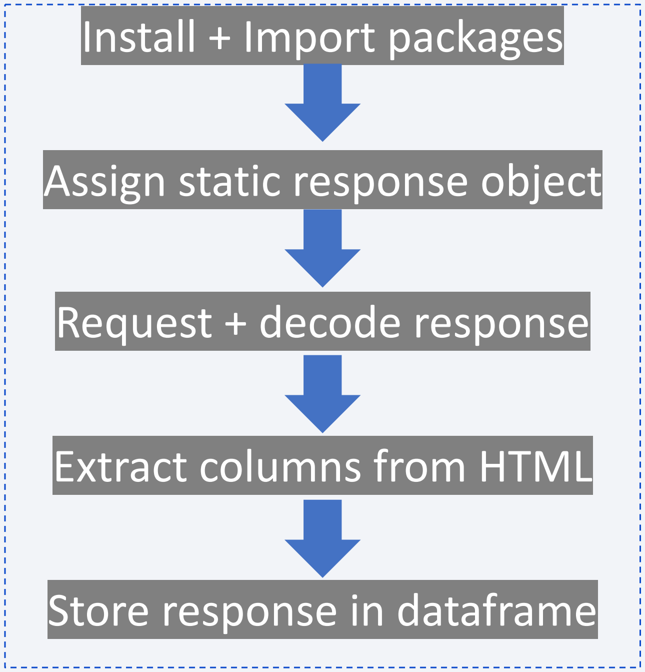
- Import python packages (requests, pandas, numpy, datetime)
- Assign static response object to static_json_url
- Use requests.get() to request API response
- Use .json() to decode the response content
- Use json_normalize() to turn json data into dataframe

- See Github notebook for details



Data Collection - Scraping

- Install required packages
- Assign Falcon 9 website to static_url variable
- Use requests.get() method to request response from Falcon 9 website
- Use BeautifulSoup() to create an object from the response
- Extract column/variable names from Falcon 9 website
- Create data frame by parsing the Falcon 9 website tables
- See Github notebook for details



Data Wrangling

Data processing methods to assure data quality & prepare data for analysis

- Computed the percentage of the missing values in each attribute
  - df.isnull().sum()/len(df)*100
- Calculated the number of launches by launch site
  - launch_site_counts = df['LaunchSite'].value_counts()
- Calculated the number and occurrence of each orbit
  - orbit_counts = df['Orbit'].value_counts()
- Calculated the number & occurrence of mission outcome of the orbits
  - landing_outcomes = df['Outcome'].value_counts()
- Created binary landing outcome label (0/1) from Outcome column
  - landing_class=[0 if outcome in bad_outcome else 1 for outcome in df['Outcome’]]
- See GitHub for notebook

EDA with Data Visualization

The following visualizations were created:

1. Flight Number and Launch Site relationship
2. Payload and Launch Site relationship
3. Success rate of each orbit type
4. Flight Number and Orbit type relationship
5. Payload and Orbit type relationship
6. Launch Success yearly trend

- See Github for details

EDA with SQL

Summary of the SQL queries conducted

- Names of the unique launch sites in the space mission
- Launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship & payload mass greater than 4000 but less than 6000
- Names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- See Github for details

Build an Interactive Map with Folium

- Marked all launch sites on a Folium map
- Marked the success/failed launches for each site
- Calculated the distances between a launch site to its proximities
- Added a launch site drop-down input component
- Added a callback function to render the required success outcome pie chart
- Added a range slider to choose payload
- Added callback function to render the payload-outcome scatter plot
- See GitHub for details

Build a Dashboard with Plotly Dash

Dashboards allows for interactive exploration of data visualizations

- Features added for interactivity:
  1: Launch Site Drop-down Input Component

  2: Callback function to render success-pie-chart based on selected site dropdown

  3: Range Slider to Select Payload

  4. Callback function to render the success-payload-scatter-chart scatter plot

See GitHub for details

Predictive Analysis (Classification): 1. Overview

- Developed four classification models to predict first-stage launch success
  1. Logistic regression
  2. Support vector machine (SVM)
  3. Decision tree classification (decision tree)
  4. k-means nearest neighbor (KNN)
- Used a training data sample for model development
- Predictive accuracy of the classification models computed using test data sample.
- Compared predictive performance across classification models

Predictive Analysis (Classification): 2a. Process Flow

1 **Data pre-processing**

- Import Libraries and Define Auxiliary Functions
- Load the dataframe
- Normalize data with StandardScaler()

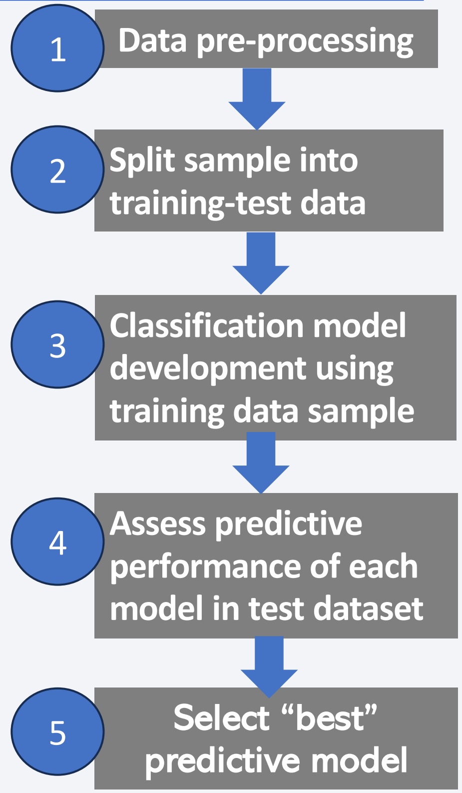
2 **Split sample into training-test data**

Randomly select 80/20 training/test data split

using (train_test_split(X, Y, test_size=0.2, random_state=2)

3 **Classification model development using training data sample**

- LogisticRegression()
- SVC()
- DecisionTreeClassifier()
- KNeighborsClassifier()



Predictive Analysis (Classification): 2b. Process Flow

3 Classification model development using training data sample (cont.)

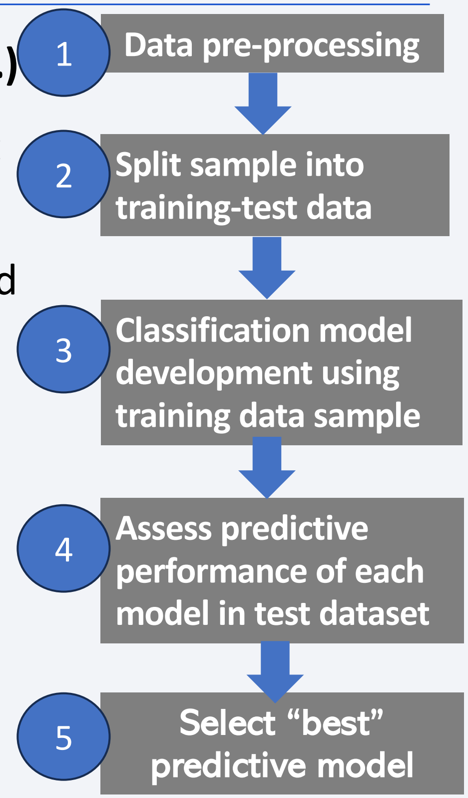
- GridSearchCV object for the model. Fit the object to find the best parameters from the dictionary parameters.
- Select best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score_

4 Assess predictive performance of each model in test dataset

- Calculate the accuracy on the test data using the method score()
- Generate confusion matrix with predict() and plot_confusion_matrix()

5 Select "best" predictive model

See GitHub for details



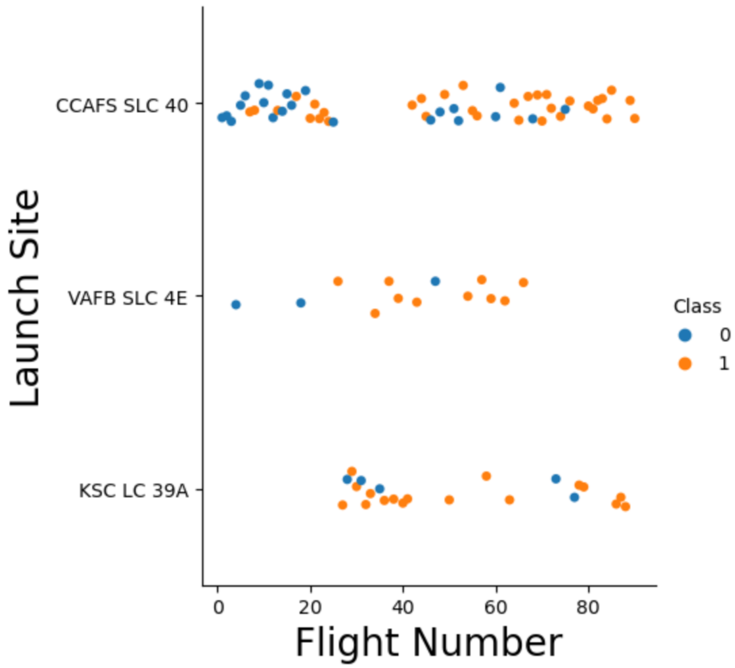
Results Overview: See Sections 2-4 for Data

- Exploratory data analysis results
  - Visualizations
  - SQL
- Interactive analytics demo in screenshots
  - Folium Maps
  - Dashboards
- Predictive analysis results
  - Classification model development
  - Accuracy and predictive performance of each classification model

Section 2

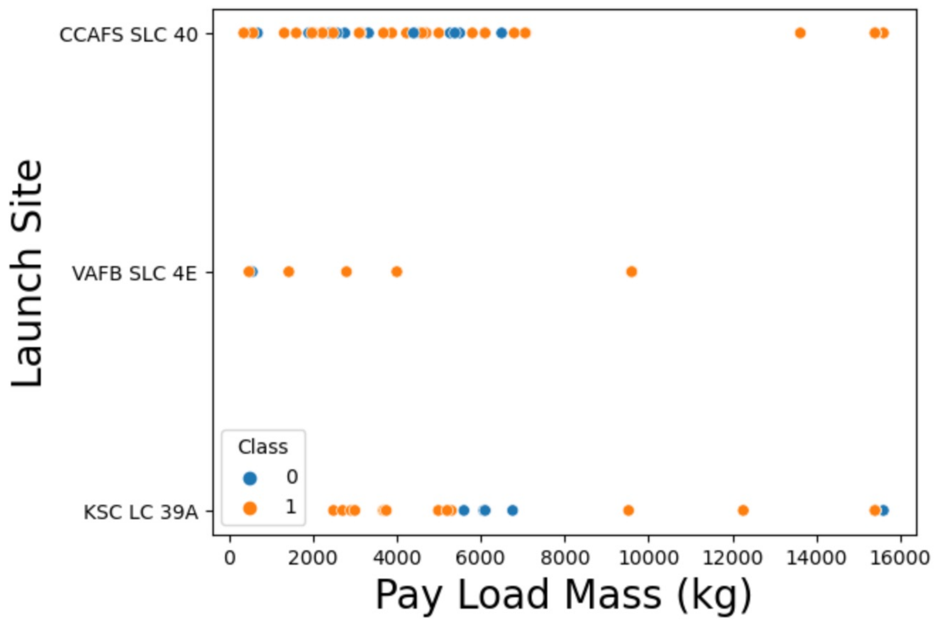
Insights drawn from EDA

Relationship: 1. Flight Number and Launch Site



- Across launch sites
  - Higher flight number, fewer unsuccessful flights

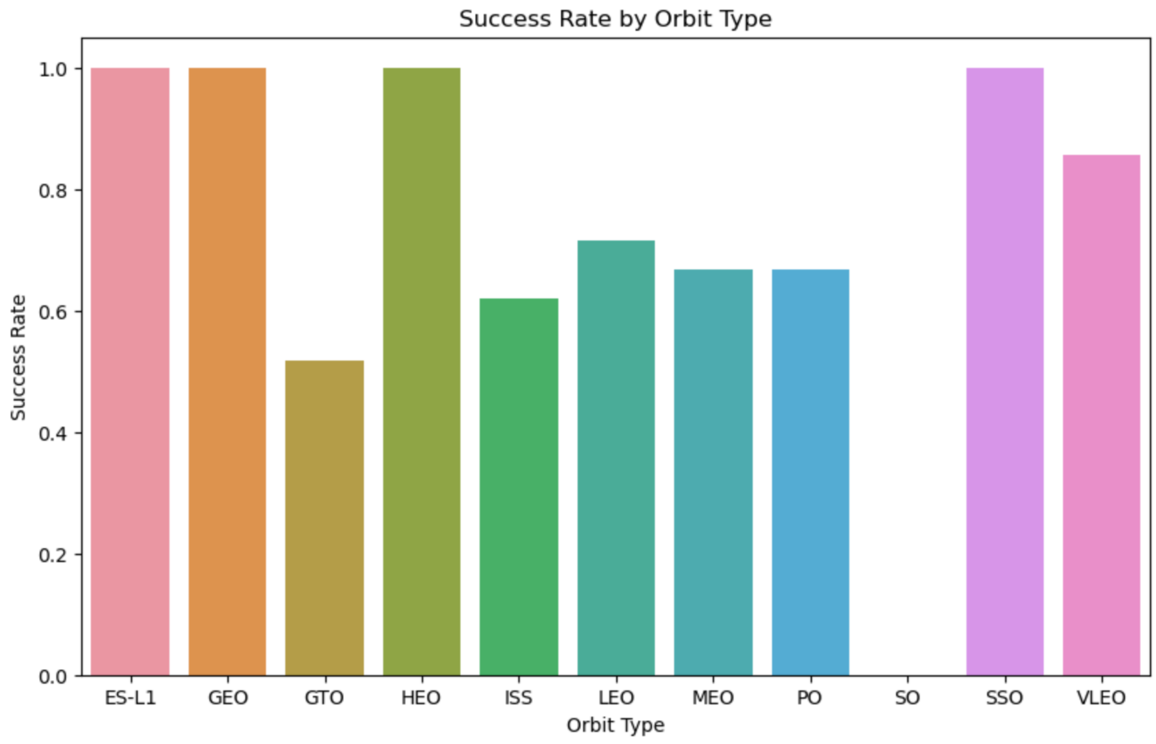
Relationship: 2. Payload and Launch Site



VAFB SLC 4E: payloads < 10,000 Kg launched at this site

Unclear patterns:
CCAFS SLC 40 & KSC LC 39A

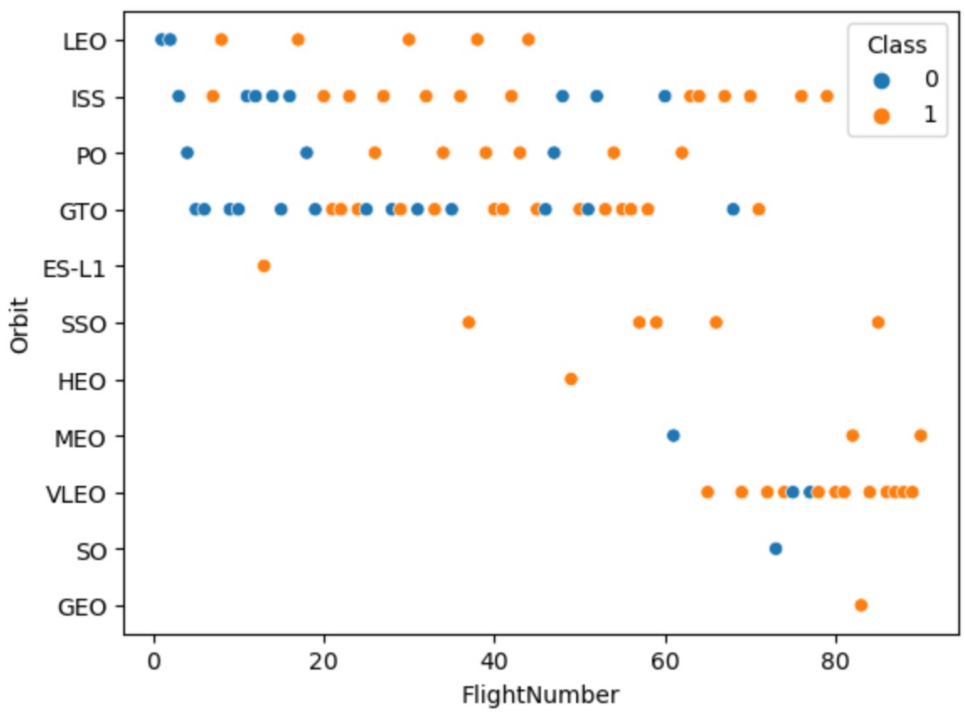
Relationship: 3. Success rate of each orbit type



Higher success rates observed for orbits:

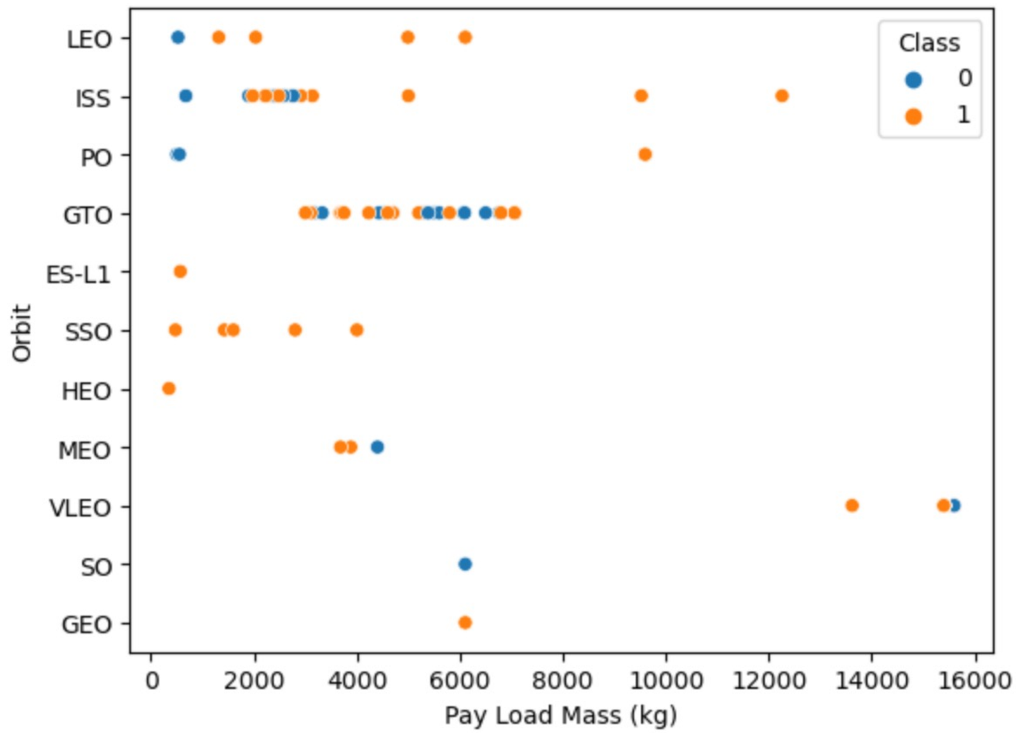
ES-L1
GEO
HEO
SSO
VLEO

Relationship: 4. Flight Number and Orbit type



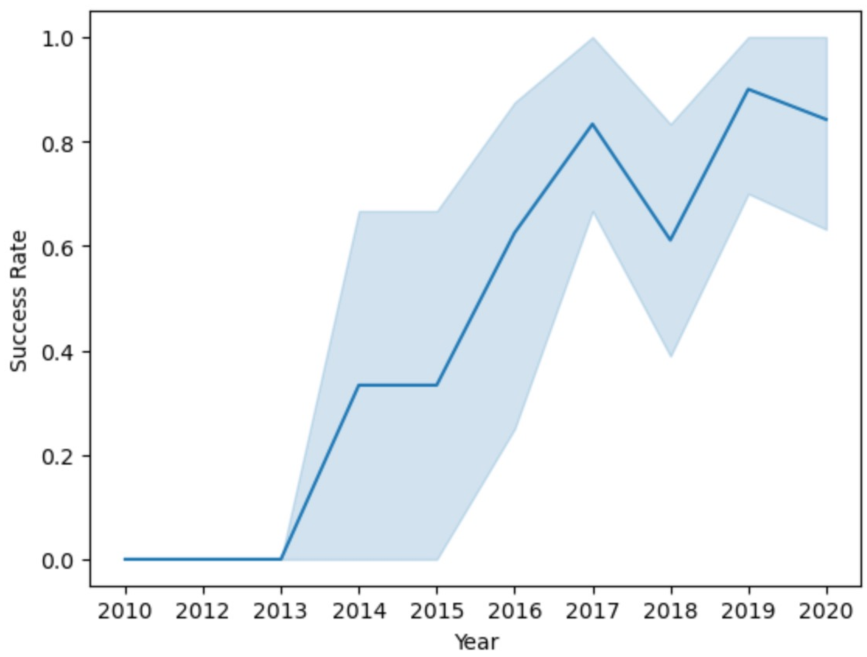
- LEO: Higher flight number, greater likelihood of successful landings
- GTO: indeterminable pattern

Relationship: 5. Payload and Orbit type



- LEO and ISS: have successful landing with heavier payloads.
- GTO & other orbits: Unclear relationship based on plotted date

Relationship: 6. Launch success yearly trend



- After 2015, Generally upward trend of success rate observed

1. All Launch Site Names

- Code + output

```
#SPACEXTABLE is the new table; SPACEXTBL is the original data

%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

There are 4 unique launch sites.

Based on the data, those names are:

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

2. Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEXTBL where LAUNCH_SITE like 'CCA%' LIMIT 5
#filter to count=5
```

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

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

The output of 5 selected launch sites (CCAFS LC-40) produced output with orbit=LEO, different customers (SpaceX, NASA), and successful mission outcomes.

3. Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
]: #customer = NASA
%sql SELECT sum(PAYLOAD_MASS__KG_) AS payloadmass from SPACEXTABLE where Customer = 'NASA (CRS)';

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

]:

payloadmass
45596

```
]: #customer LIKE or related to NASA
%sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL where Customer like 'NASA%'

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

]:

sum(PAYLOAD_MASS__KG_)
99980

Where the customer data was “NASA”, the total payload carried by its boosters totaled 45,596 KG..

There were customers with names similar to NASA, and for those, the total payload carried was 99,980 KG.

4. Average Payload Mass by F9 v1.1

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

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

```
%sql SELECT AVG (PAYLOAD_MASS__KG_) AS payloadmass from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

payloadmass
2928.4

The average payload mass carried by booster version F9 v1.1 was 2,928.4 KG.

5. First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
#correct
%sql SELECT min(Date) as mindate from SPACEXTABLE where "Landing_Outcome" ='Success (ground pad)' ;
```

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

mindate
2015-12-22

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

6.Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
#correct
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome"= 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 AN
```

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

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

- Four boosters successfully met the criteria:

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

7. Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome , count(*) FROM SPACEXTBL group by Mission_Outcome
```

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

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

According to the SpaceX data, there were 100 successful mission outcomes and 1 failed mission outcome.

8. 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)
```

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

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

There were 12 booster names that carried maximum payload mass according to the data. The names are displayed in the output.

9. 2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%%sql SELECT "Booster_Version", "Launch_Site" FROM SPACEXTABLE

WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,1,4) = '2015';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

There were two failed landing outcomes in drone ship in 2015 and both were at the launch site CCAFS LC-40.

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql SELECT "LANDING_OUTCOME", COUNT(*) as 'COUNT' FROM SPACEXTBL

WHERE substr(Date,1,4) || substr(Date,6,2) || substr(Date,9,2)

between '20100604' and '20170320' GROUP BY "Landing_Outcome" ORDER BY "COUNT" DESC;
```

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

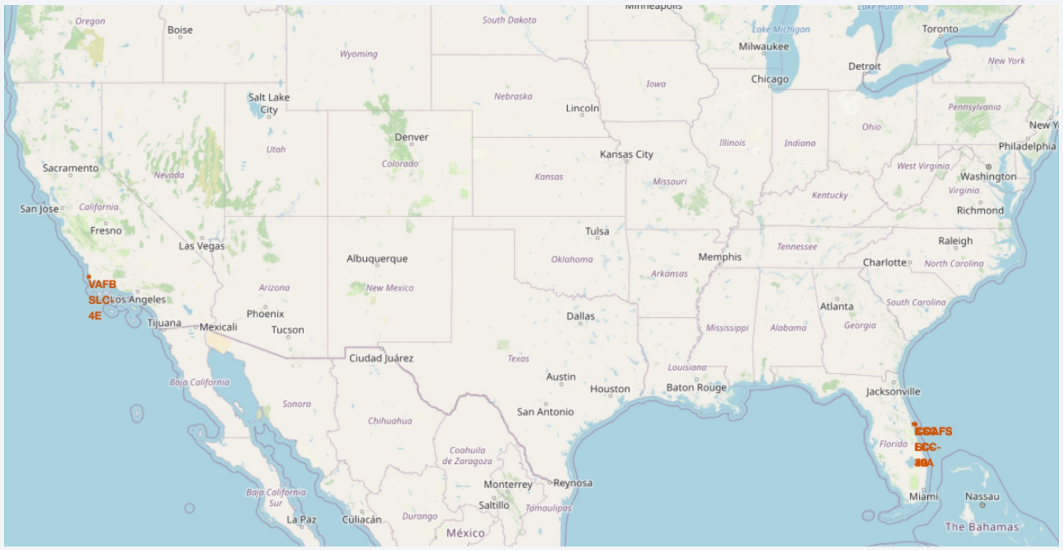
Landing_Outcome	COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

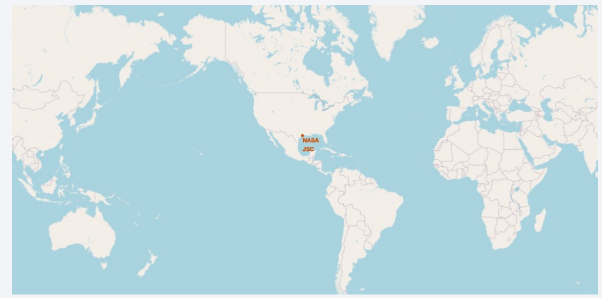
There were 8 types of landing outcomes during June 4, 2010 and March 20, 2017. The most frequent outcome was no attempt (n=10), followed by success or failure on drone ship (n=5).

Of the ranking, 2 were failed parachute outcomes.


Section 3
Launch Sites
Proximities Analysis

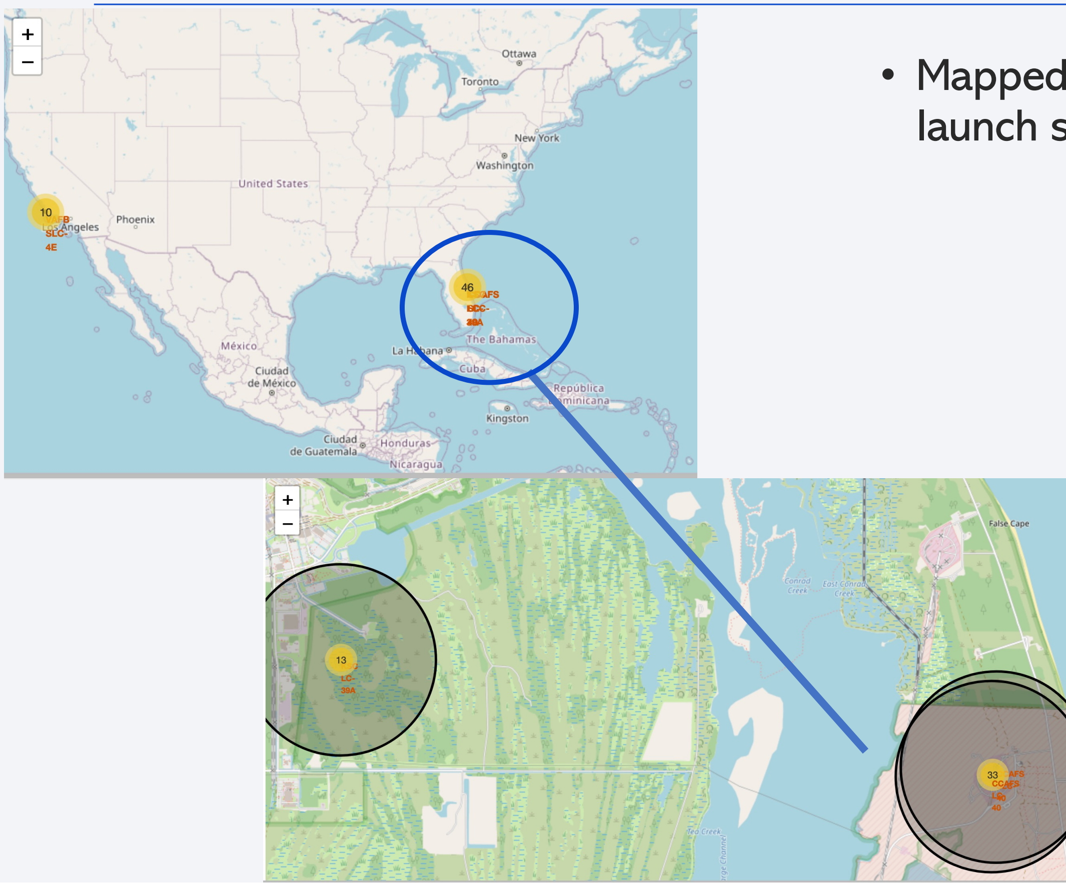
Folium Map: 1. Mark all launch sites





The launch sites are located on coasts with near proximity to the ocean and proximity to the Equator line.

Folium Map: 2a. Success or Failure Launches

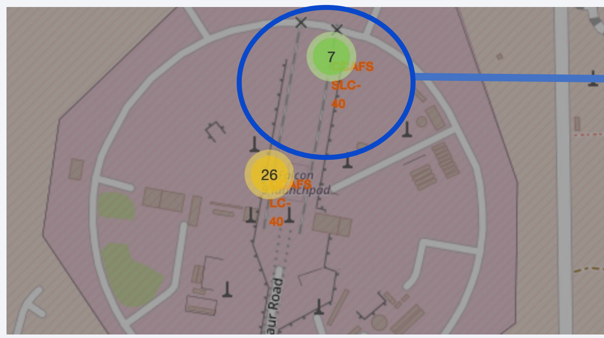


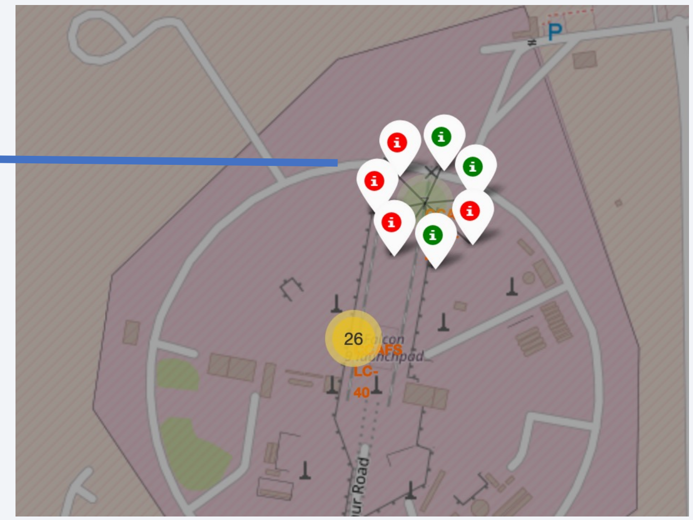
- Mapped launch outcomes to each launch site

Drill down to illustrate map:

Examine outcomes of one launch site: Florida's Cape Canaveral's Air Force Space Station Launch Complex (CCAFS SLC-40)

Folium Map: 2b. Success or Failure Launches





2b-1. At Florida's Cape Canaveral's Air Force Space Station Launch Complex (CCAFS SLC-40), data included n=33 launches.

2b-2.

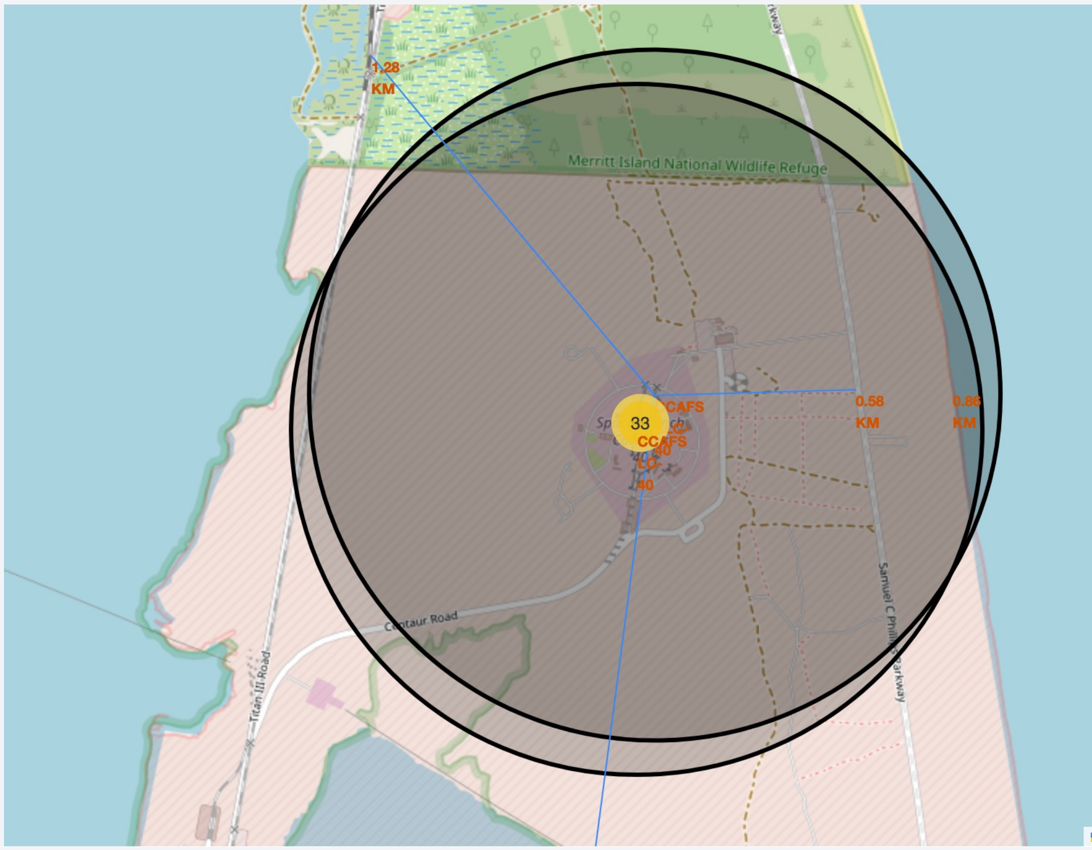
Launches occurred from two locations at the complex (CCAFS SLC-40);

n=7 and n=26.

2b-3.

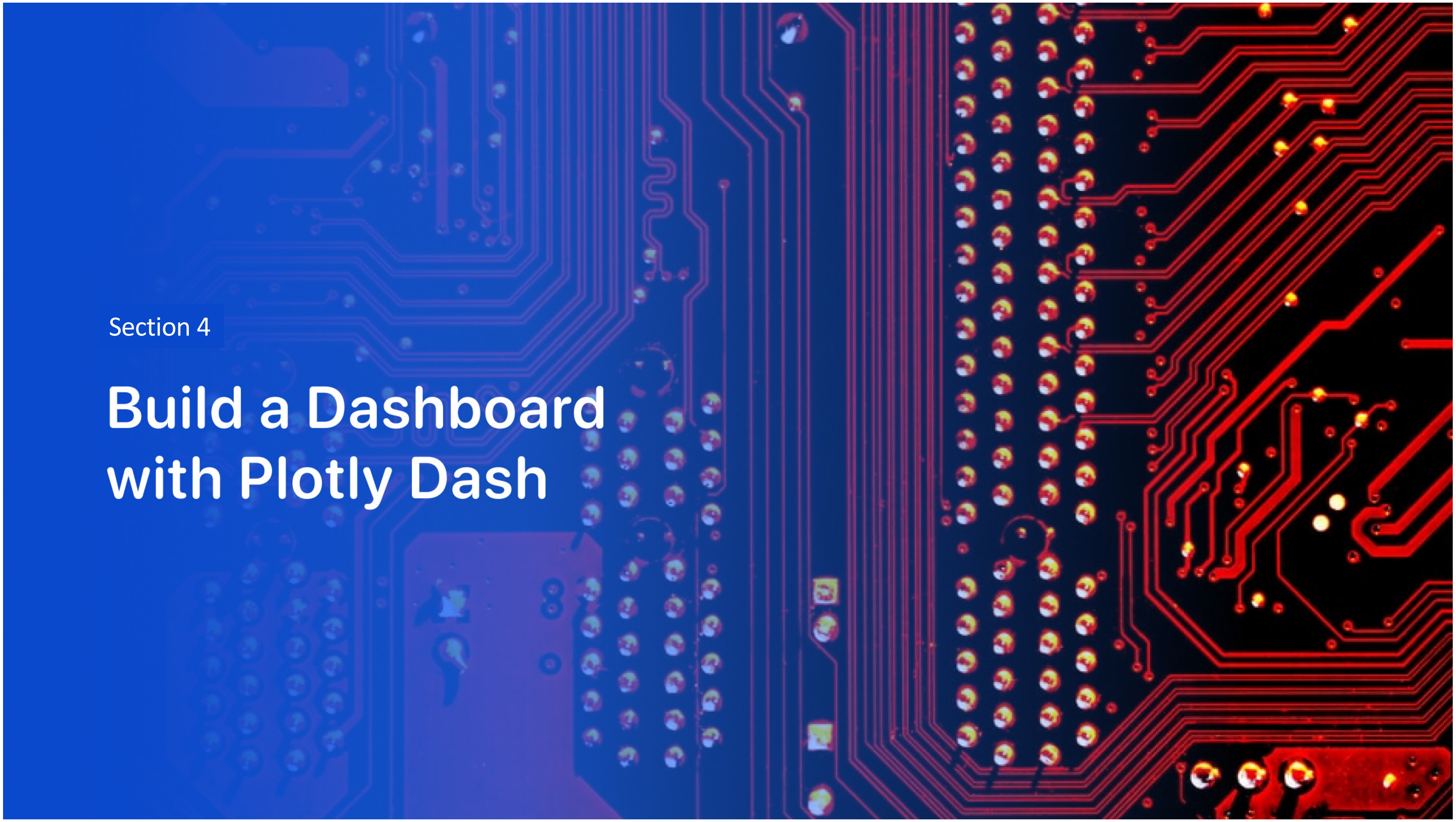
Of the 7 launches conducted at one of the locations, 3 missions were successful (green) and remaining were unsuccessful (red).

Folium: 3. Distance to Points of Interest - Highways



- From Cape Canaveral's Air Force Space Station Launch Complex (CCAFS SLC-40), mapped distance to nearest highways.
  - The blue lines map the distance from the launch site to the nearest highways northwest (1.29 km) and east (.56 km).

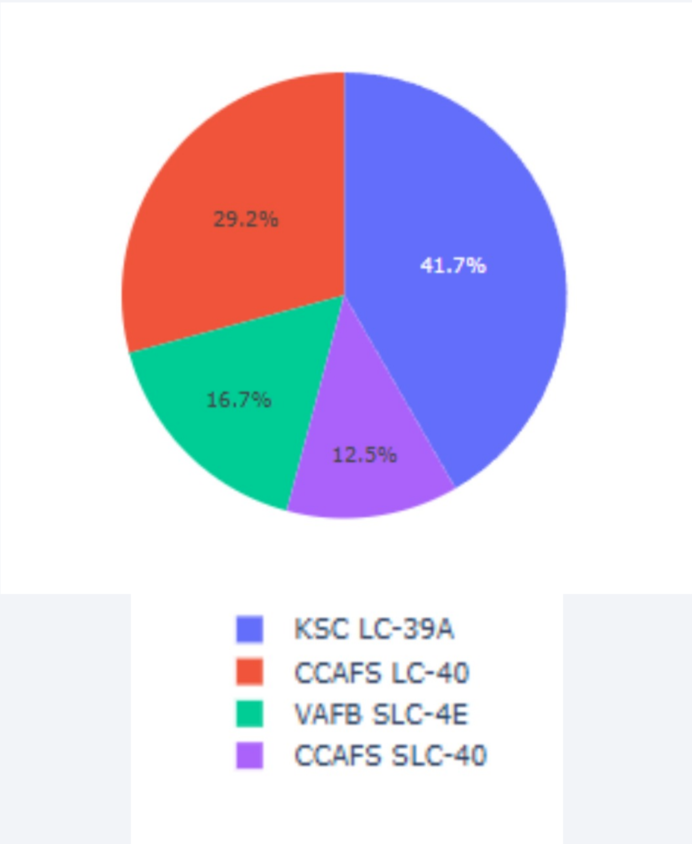
- Overall, the data indicated the shortest distance from launch sites to nearest highway was 0.58 km, railroad was 1.28 km, and to nearest city was 51.43 km.



Section 4

Build a Dashboard with Plotly Dash

Dashboard:1a. Success Count Across Launch Sites



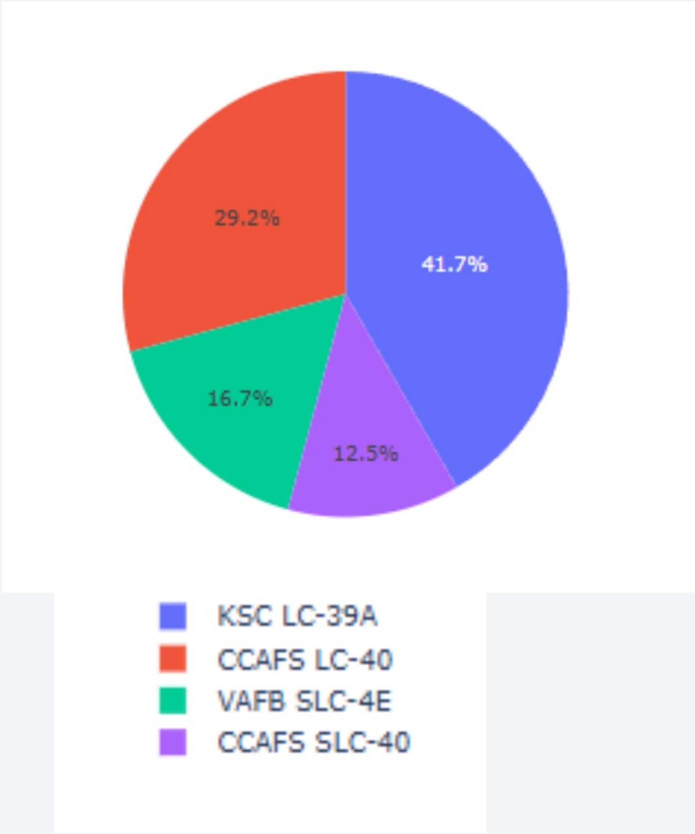
Pie chart displays the proportion of successful launches contributed by each launch site in relation to the total count of successful outcomes achieved by SpaceX overall.

The largest portion of the pie chart represents the launch site with greatest number of successful launches.

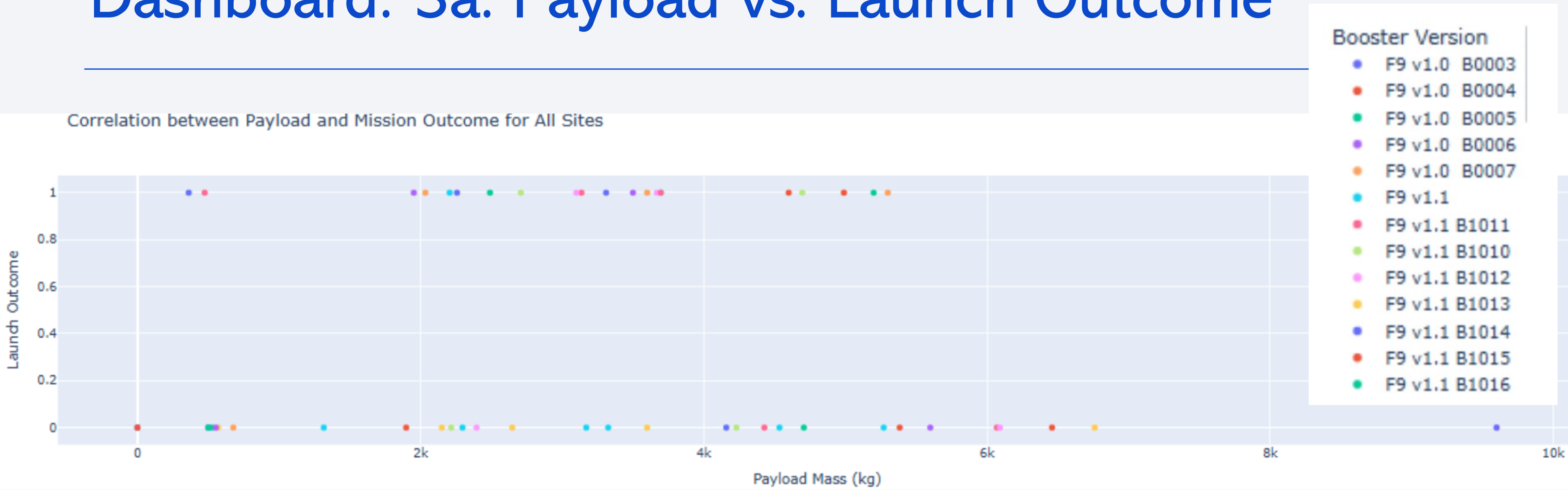
Dashboard: 1b. Success Count Across Launch Sites

41.7% of successful launches occurred at Kennedy Space Center Launch Complex 39A (KSC LC-39A).

The fewest proportion of successful outcomes occurred at Cape Canaveral (CCAFS SLC-40).



Dashboard: 3a. Payload vs. Launch Outcome



- No clear pattern emerges that can be easily visually discerned by the correlation graphed by the scatterplot of payload mass vs launch outcome for booster versions.

Dashboard: 3b. Payload slider vs. Launch Outcome

- **Payload slider allows for selection of kg range of interest that can be demonstrated in the scatterplot of payload vs launch outcome.**
- No clear pattern emerges that can be easily visually discerned by the correlation graphed by the scatterplot of payload mass vs launch outcome for booster versions.

Section 5

Predictive Analysis (Classification)

Classification Accuracy: 1. Visualization

1. Bar chart of classification four machine learning models computed:

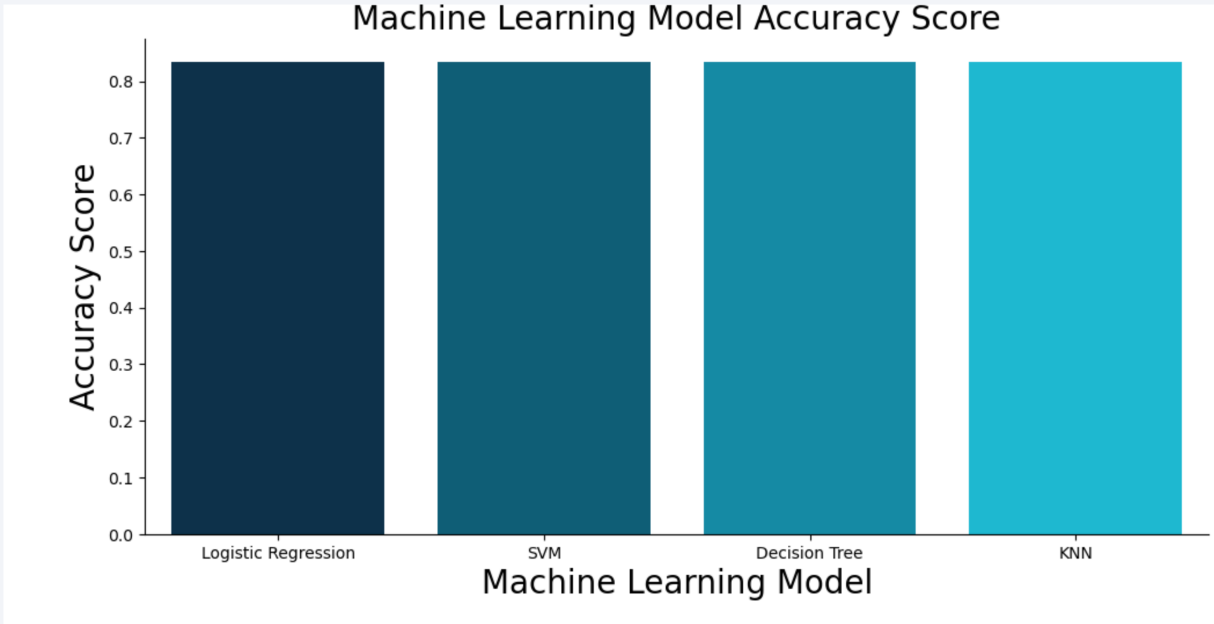
1.logistic regression

2. support vector machine (SVM)

3. decision tree classification (decision tree)

4. k-means nearest neighbor (KNN)

- classification accuracy were consistent across all classification models.



Models trained using sample of n=72.

Model Accuracy calculated using test sample of n=18

Classification Accuracy: 2. Numerical Comparison

2. Classification accuracy was consistent across all classification models:

Accuracy=0.833333
n=18 test samples

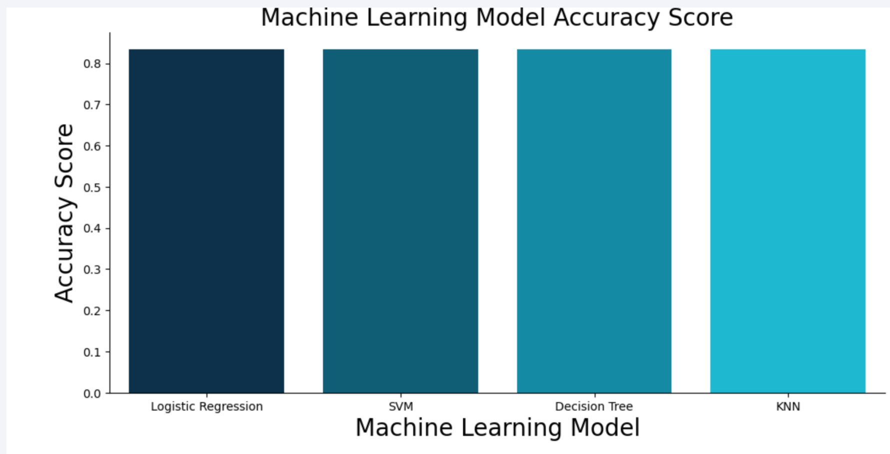
1.logistic regression

2. support vector machine (SVM)

3. decision tree classification (decision tree)

4. k-means nearest neighbor (KNN)

	Model	Score	Max Score
0	Logistic Regression	0.833333	Max Score
1	SVM	0.833333	Max Score
2	Decision Tree	0.833333	Max Score
3	KNN	0.833333	Max Score

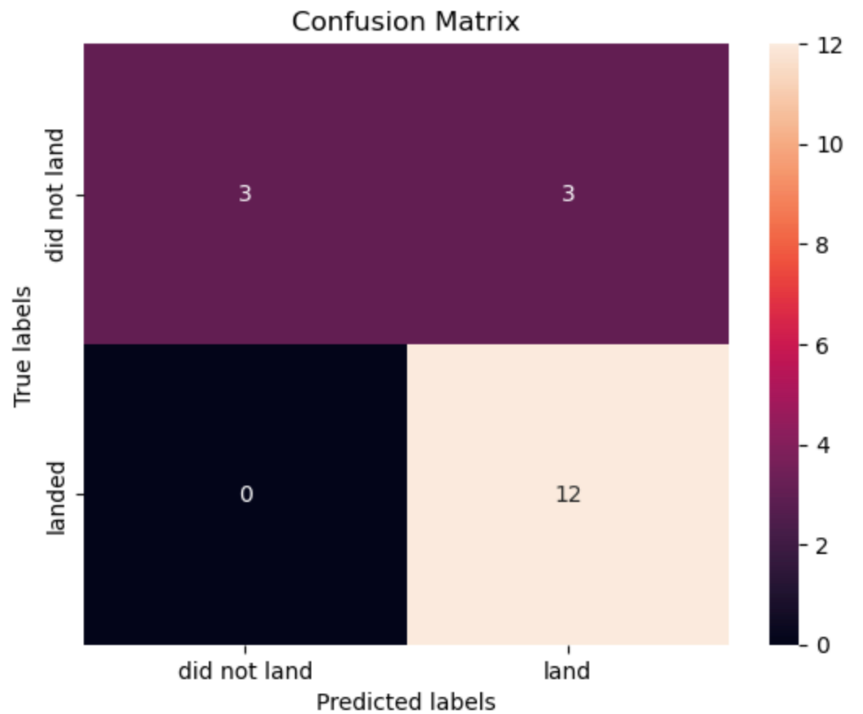


Confusion Matrix | 3. Overall Results

3. Models tested using sample of n=18.

- Confusion matrix results **were consistent (and indicate equivalent performance)** across the 4 classification models:
- Log reg
- SVM
- Decision Tree
- KNN

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



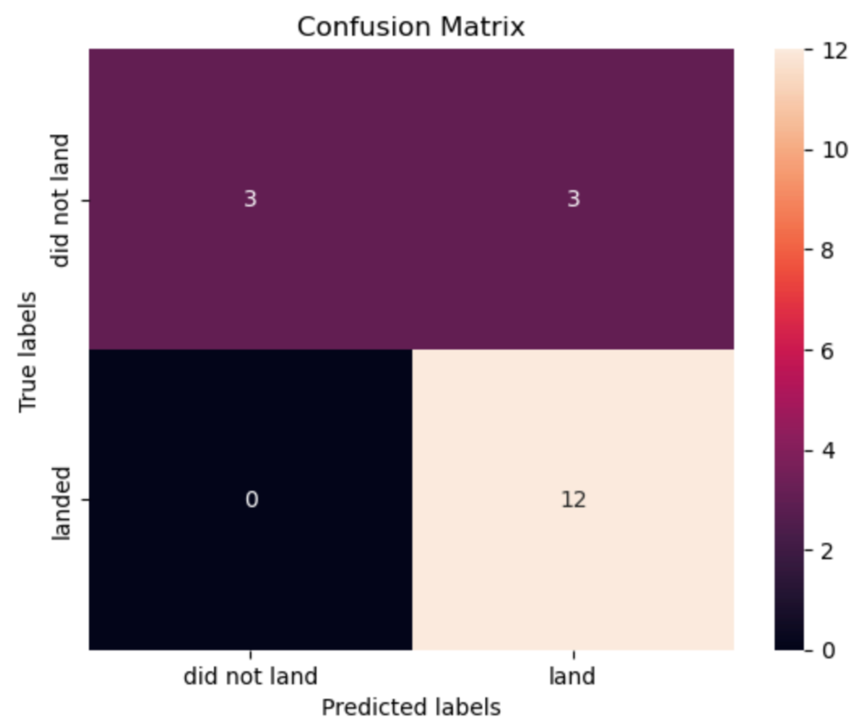
Confusion Matrix | 4. KNN Results

4. Confusion matrix of KNN classification illustrated here and represents the same results generated from the other 3 models.

Of 18 samples tested, 12/12 were accurately labeled as landed (true positives, high sensitivity)

However, 3/6 were incorrectly labeled as landed when they did not land (low specificity=0.50).

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

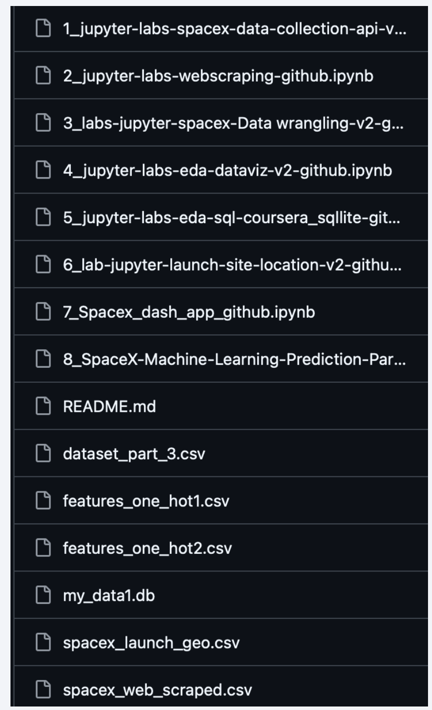


Conclusions

- Success rate for SpaceX launches have consistently improved across time and across launch sites with higher number of flights launched.
- Classification models of logistic regression, support vector machine, k-means nearest neighbor, and decision tree could be used to develop predictive models for SpaceX launch success.
- The specificity rate to determine likelihood of unsuccessful launch is not ideal (50%) based on the classification models tested; therefore, additional refinements (additional data, models, data transformations) would be required for competitor company to accurately predict first-stage launch outcomes.
- From the data available from SpaceX Wikipedia page for Falcon 9 launches, inadequate data or incomplete data models are available to confidently predict likelihood of success of first stage launch based on the current analysis.

Appendix

- Github repository of jupyter Notebooks of code used for labs found located here:
- Slides (pdf)
- Notebooks (8)
- Output from notebooks




Thank you!