IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nick Driscoll
10/12/2022

Outline

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

Executive Summary

- Summary of methodologies
    - This study was conducted following the methodology outlined in the IBM Data Science Capstone. Most steps of the methodology were also conducted using multiple different techniques, for example:
        - Data collection was performed using Webscraping of Wikipedia and the SpaceX API
        - EDA was performed using SQL, pandas and visualization
        - Visualization was performed using matplotlib plotting of pandas DataFrames, Folium maps, and Plotly Dash dashboards
        - Predictive analysis was conducted using logistic regression, SVM, Decision Tree, and KNN models
- Summary of all results
    - EDA uncovered insights and relationships between features that would be used as predictors: Flight Number, Payload Mass, Orbit, Launch Site
    - Training of multiple classification models produced the same classification accuracy performance

Introduction

- SpaceX advertises that they can launch their Falcon 9 rocket with a cost savings of over 100 million dollars versus other competitors. The key is that they can land and re-use the first stage of the Falcon 9 rocket. Publicly available historical launch data for the Falcon 9 may help competitor companies predict what it takes to build a re-usable first stage rocket.
- Using the historical launch data, which includes rocket parameters and results, a company may be able to build predictive Machine Learning models to identify if the first stage of the rocket will land or not.

Section 1

Methodology

Methodology

Data collection methodology:

- Data on SpaceX Falcon9 Launches is publicly available. It was procured in multiple ways; using the SpaceX API, and WebScraping the Falcon 9 Wikipedia page

Perform data wrangling

- Filtered out rockets that were not Falcon 9, Calculated number of launches, occurrence of each orbit type, created binary landing outcome column

Perform exploratory data analysis (EDA) using visualization and SQL

- Visualized: mission outcome by orbit type, mission outcome by launch location, size, weight, etc.
- Used SQL Queries to identify distinct mission outcomes, booster version types, payload masses, and rank the successful mission outcomes

Perform interactive visual analytics using Folium and Plotly Dash

- Plotted and categorized mission outcomes and site proximity to landmarks

Perform predictive analysis using classification models

- Cleaned and split data into training/test sets, transformed with StandardScaler, trained Logistic Regression, SVM, Tree and KNN Models, and assessed performance using scoring metrics and Confusion Matrix against test data

Data Collection

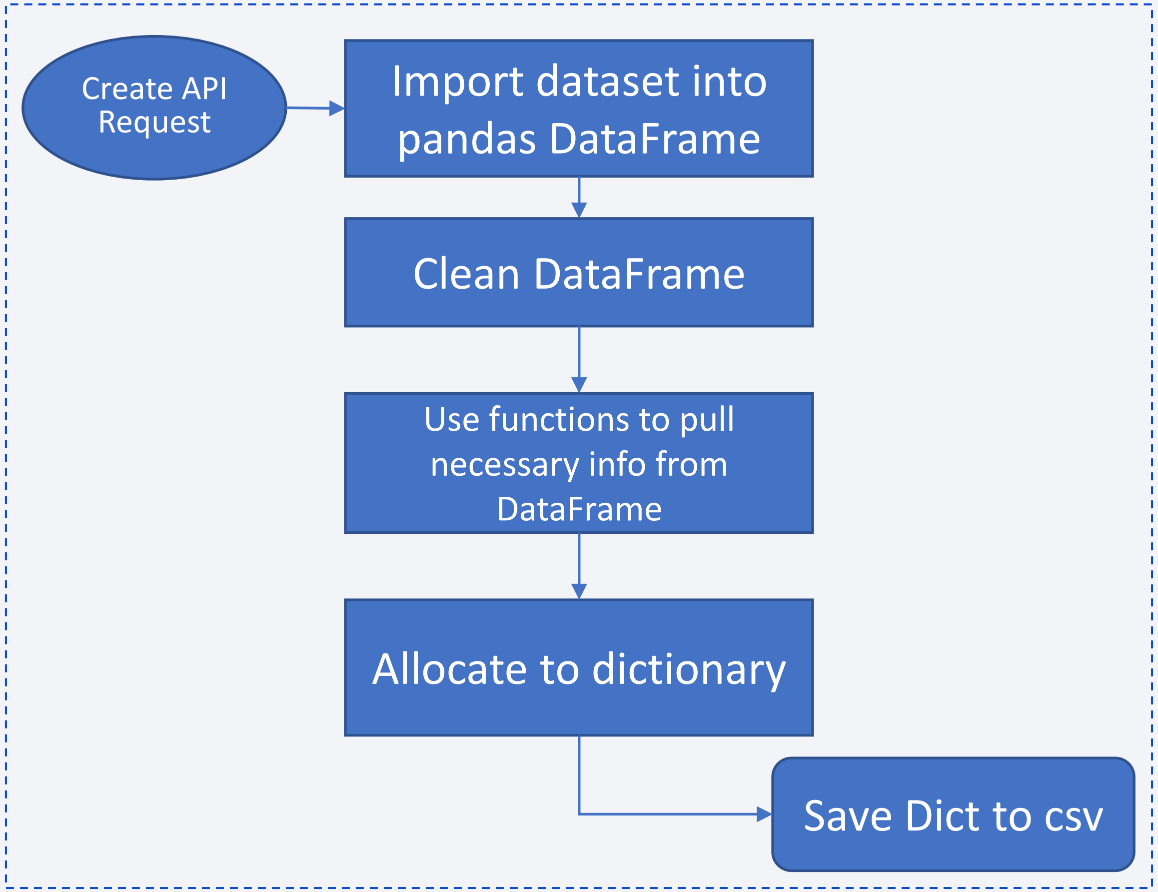
- Data sets were collected via the SpaceX API and the Webscraping
- To access the SpaceX API, functions were defined to pull in data for the Booster Version, Launch Site, Payload information and other core data
- Webscraping was performed on the Falcon 9 Wikipedia page. Similarly, functions were defined to pull date/time, booster version, landing status, mass from the Wikipedia table
- Additional details are presented on the following charts

Data Collection – SpaceX API

- Used the following calls to pull the 'static' json data

```
static_json_url =
'…/datasets/API_call_spacex_api.json'
response = requests.get(static_json_url)
data = pd.json_normalize(response.json())
```

- Additionally, called specific data from the following apis:
- https://api.spacexdata.com/v4/rockets/
- https://api.spacexdata.com/v4/launchpads/
- https://api.spacexdata.com/v4/payloads/
- https://api.spacexdata.com/v4/cores/

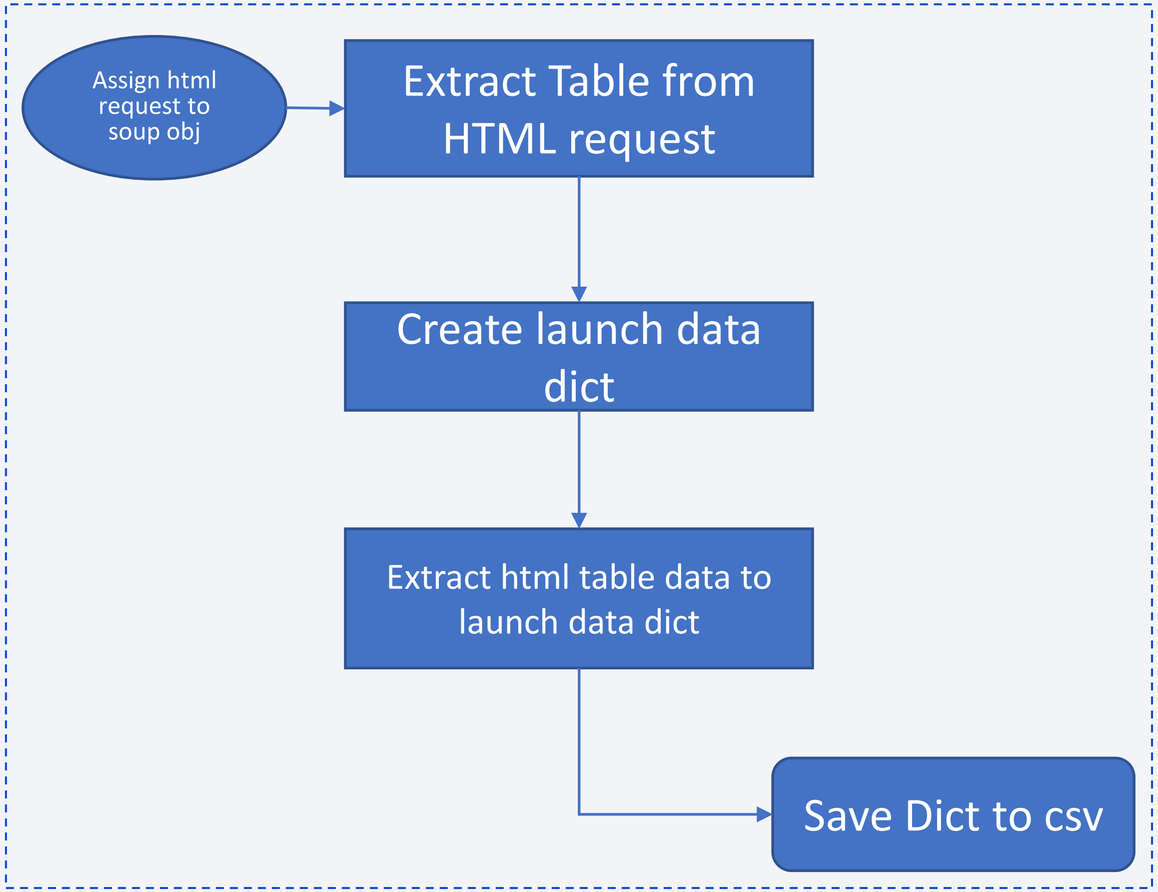


https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_API_Scraping.ipynb

Data Collection - Scraping

- Web Scraped from:

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



https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_Webscraping.ipynb

Data Wrangling

- After data collection, data was cleaned and organized
  - Filter the data only by Falcon 9 data
  - Removing nulls and missing values
  - Find distribution of launches per site, per type of orbit, per booster version, per payload mass
  - Create a ‘landing’ outcome class to make the mission outcome binary

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_Wrangling.ipynb

EDA with Data Visualization

- We visualized the following, in order to gain an understanding of the relationships between features and to start forming a hypothesis of which features contribute to launch success
  - Flight Number vs. Payload Mass Categorized by Landing Success
  - Flight Number vs. Launch Site Categorized by Landing Success
  - Payload Mass vs. Launch Site Categorized by Landing Success
  - Success rate vs. Orbit Type
  - Flight Number vs. Orbit Type, Categorized by Landing Success
  - Payload vs. Orbit Type, Categorized by Landing Success
  - Landing Success over Time

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_EDA_dataviz_pandas.ipynb

EDA with SQL

- The following SQL Queries were executed in order to explore the dataset:
  - Unique Launch Sites
  - Launch Sites beginning with 'CCA'
  - Total Payload Mass for NASA launches
  - Average Payload Mass for Falcon 9 V1.1
  - First Successful Landing
  - Names of Boosters that landed on Drone Ships between a certain weight
  - Total number of Successful and Failure Missions
  - Booster Version with maximum payload
  - Month of Failed Landings in 2015
  - Ranking of Successful Landing Outcomes

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_EDA_w_SQL.ipynb

Build an Interactive Map with Folium

- Folium was used to understand the location of different launches, their success rates, and proximity to landmarks
- We used various folium map objects to identify unique features:
  - Markers were used to identify each unique launch location
  - Marker clusters and coloring were used to categorize launch success at each site
  - Lines were used to point to the closest landmarks such as coastlines, railways, highways and cities
  - Annotations were used to draw the distance to those landmarks

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/site_map.html

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_launch_sites_capstone.ipynb

Build a Dashboard with Plotly Dash

- A Pie Chart was created with a site selection dropdown to show the distribution of successful launches across sites, as well as the ratio of success to failures per site
  - This chart allowed us to identify the sites with the most and least successful launches as well as the ratio of success for each individual site
- A Scatter Chart plotted Payload mass against class (landing success or failure), categorized by booster version. This chart also had a slider bar which allowed us to filter the plot by payload size
  - This chart allowed us to understand how different booster versions were used for different payload sizes, as well as how different payload sizes and the booster version contributed to success

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/python_code/capstonePlotlyDash.py

Predictive Analysis (Classification)

- We used the feature-engineered dataset from the EDA lab to train our model
- Data was split into training and test datasets, and the predictor data was transformed using the StandardScaler
- GridSearchCV was used to optimize the input parameters for each classification model type
  - This method automatically performes optimization for us, so that we did not have to individually select, test, and evaluate the results of each possible input parameter
- Models used for classification: Logistic Regression, SVM, Decision Tree, and KNN
- Each model resulted in the same classification accuracy against the test dataset
  - The only model that varied from one execution of the script to the next was the decision tree. Ultimately, the best performing Decision tree had the same performance of the other models

https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_MachineLearning.ipynb

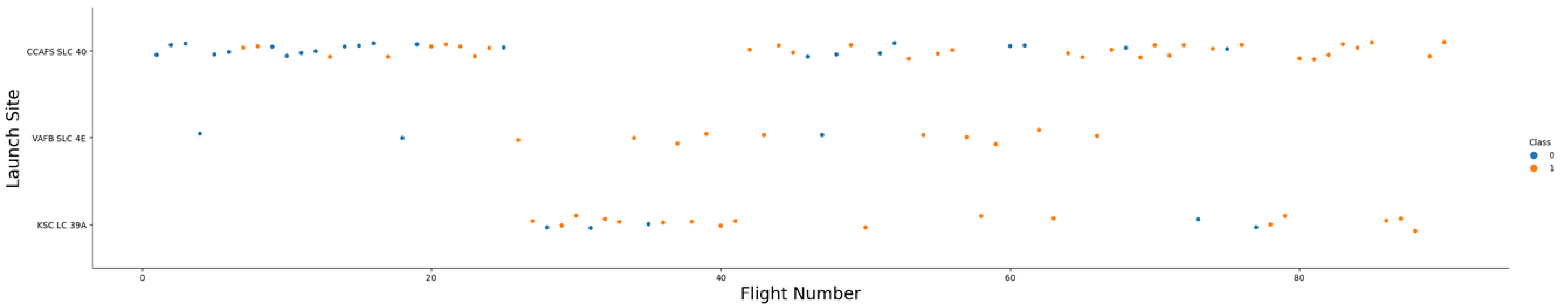
Results

- Exploratory data analysis results
  - During EDA, data was cleaned by removing nulls, and filling in the average for missing data (payload size), filtering by the rocket of interest (Falcon 9)
  - New dummy variables were created for categorical variables and the landing outcome
- Visualization and Interactive analytics demo in screenshots
  - Relationships between Flight Number, Payload Mass, Launch Location, Time, and landing success were uncovered
  - Mapping visualizations gave insights into launch locations, distribution of success across sites, and proximity to landmarks
- Predictive analysis results
  - Each model resulted in the same classification accuracy against the test dataset
  - The consistency across models may be due to the small sample size in the test set (18 samples)

Section 2

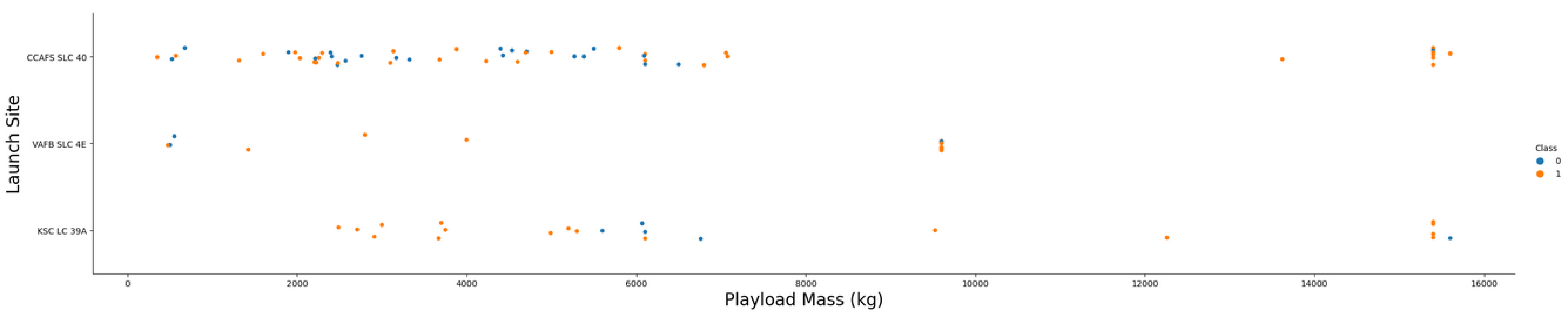
Insights drawn from EDA

Flight Number vs. Launch Site



- Success clearly increases as Flight Number increases
- VAFB has a better success ratio
- CCAFS CLC 40 has the most successes

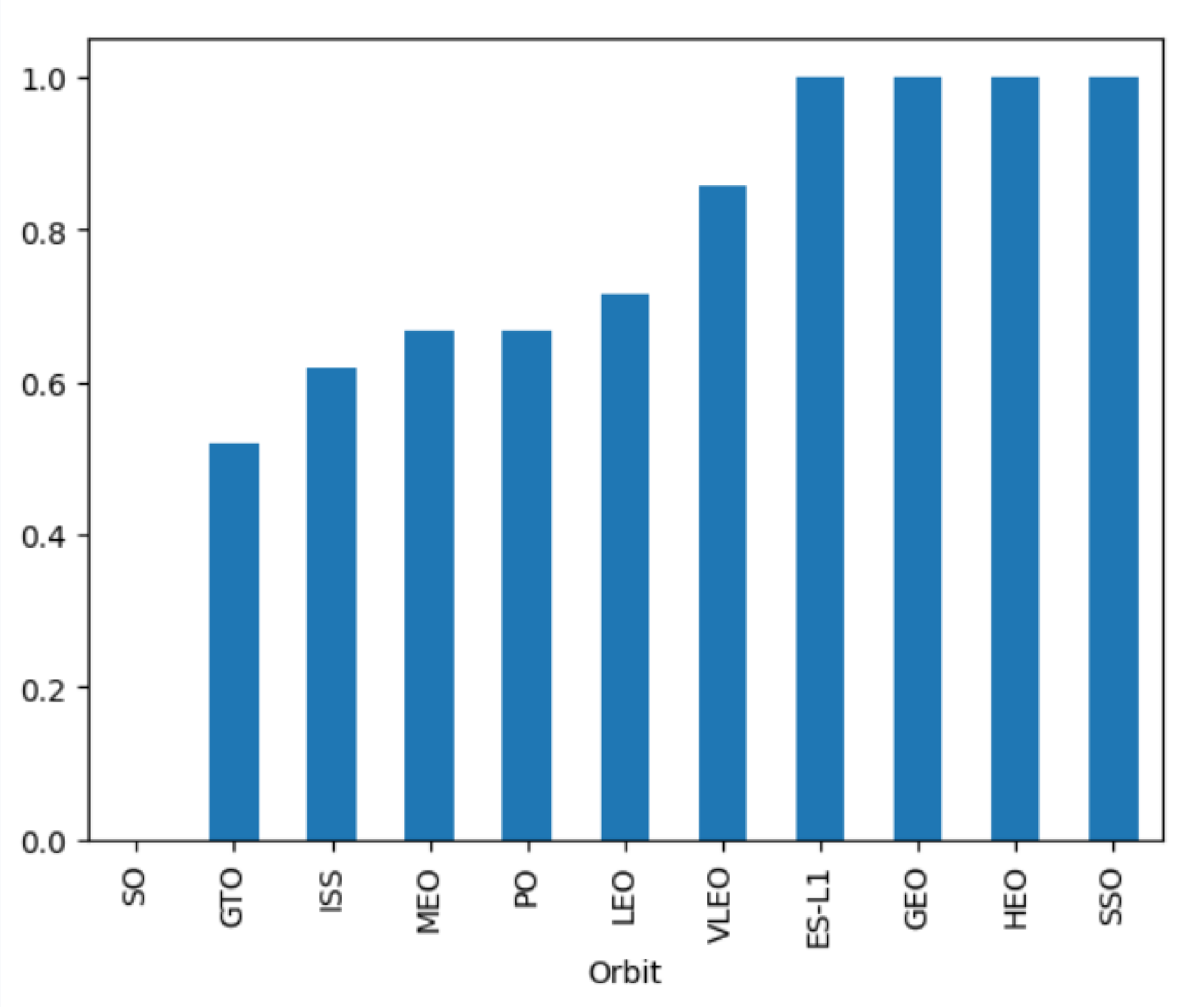
Payload vs. Launch Site



- Success clearly increases as Flight Number increases
- VAFB has a better success ratio
- CCAFS CLC 40 has the most successes

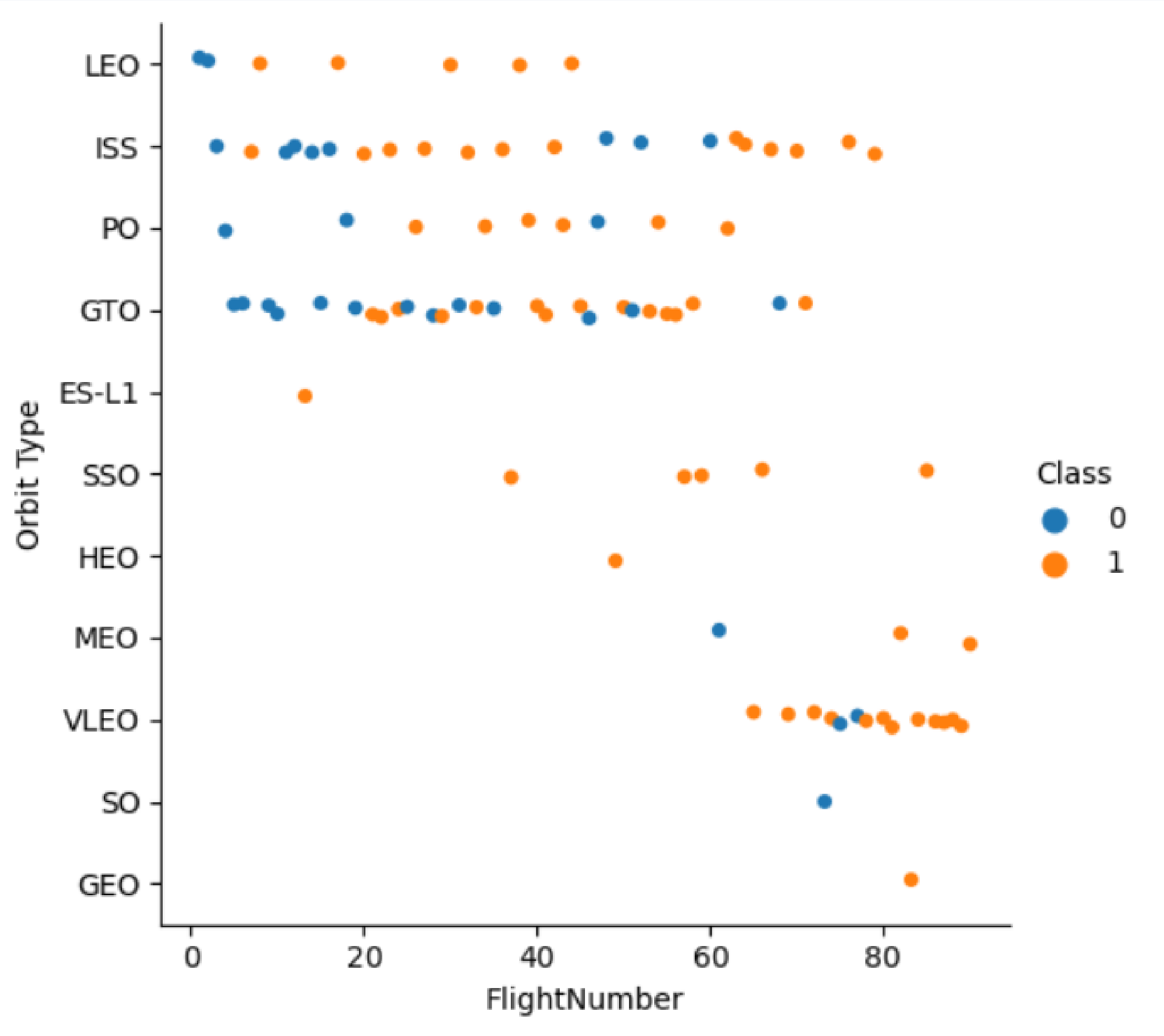
Success Rate vs. Orbit Type

- S-L1, GEO, HEO and SSO all have a high degree of success
- GTO and ISS have the lowest success rates



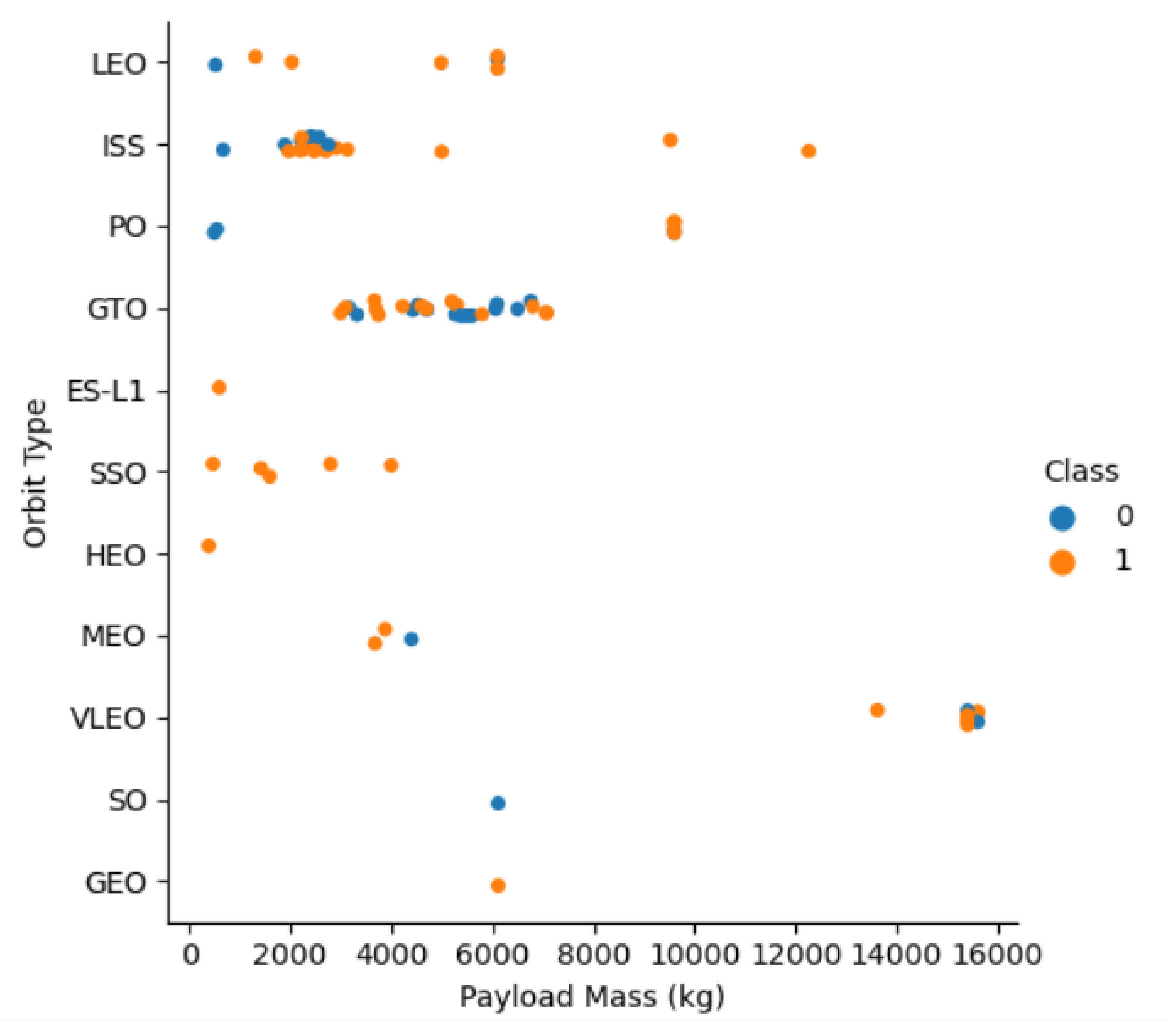
Flight Number vs. Orbit Type

- Higher Flight numbers are typically VLEO orbits
- LEO only had failures for initial flights
- GTO and ISS have the most launches



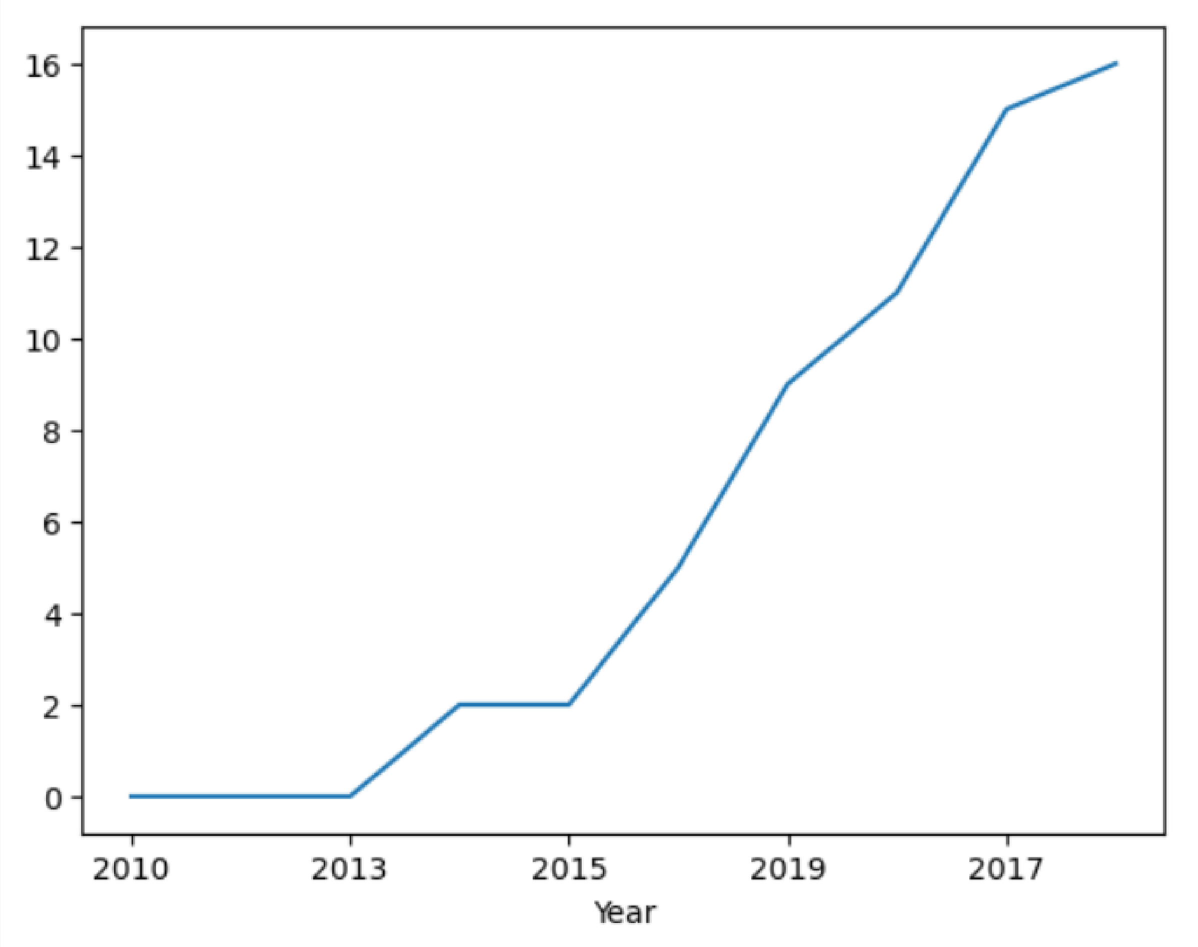
Payload vs. Orbit Type

- Payload sizes are very concentrated to Orbit types
- Higher Payload masses have better success than most



Launch Success Yearly Trend

- This clearly depicts launch success getting better as the years increase



All Launch Site Names

There are 4 unique Launch Sites:

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

All Payload Masses:

PAYLOAD_MASS__KG_	Booster_Version
500	F9 v1.1 B1003
3170	F9 v1.1
3325	F9 v1.1
2296	F9 v1.1
1316	F9 v1.1
4535	F9 v1.1
4428	F9 v1.1 B1011
2216	F9 v1.1 B1010
2395	F9 v1.1 B1012
570	F9 v1.1 B1013
4159	F9 v1.1 B1014
1898	F9 v1.1 B1015
4707	F9 v1.1 B1016
1952	F9 v1.1 B1018
553	F9 v1.1 B1017

Average:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The first Ground Pad Landing Success was

min(Date)
01-05-2017

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

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Total	Success	Failure
101	100	1

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

substr(Date,4,2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success
18-08-2020	14:31:00	F9 B5 B1049.6							
18-07-2016	04:45:00	F9 FT B1025.1	CCAF						
18-04-2018	22:51:00	F9 B4 B1045.1							
17-12-2019	00:10:00	F9 B5 B1056.3							
16-11-2020	00:27:00	F9 B5B1061.1	KSC						
15-12-2017	15:36:00	F9 FT B1035.2							
15-11-2018	20:46:00	F9 B5 B1047.2	KSC						
14-08-2017	16:31:00	F9 B4 B1039.1	KSC						
14-08-2016	05:26:00	F9 FT B1026	CCAF						
14-01-2017	17:54:00	F9 FT B1029.1	VAFB						
13-06-2020	09:21:00	F9 B5 B1059.3							
12-06-2019	14:17:00	F9 B5 B1051.2	VAFB						
11-11-2019	14:56:00	F9 B5 B1048.4							

11-10-2017	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)
11-05-2018	20:14:00	F9 B5 B1046.1	KSC LC-39A	Bangabandhu-1	3600	GTO	Thales-Alenia/BTRC	Success	Success (drone ship)
11-01-2019	15:31:00	F9 B5 B1049.2	VAFB SLC-4E	Iridium NEXT-8	9600	Polar LEO	Iridium Communications	Success	Success
10-09-2018	04:45:00	F9 B5B1049.1	CCAFS SLC-40	Telstar 18V / Apstar-5C	7060	GTO	Telesat	Success	Success
09-10-2017	12:37:00	F9 B4 B1041.1	VAFB SLC-4E	Iridium NEXT 3	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
08-10-2018	02:22:00	F9 B5 B1048.2	VAFB SLC-4E	SAOCOM 1A	3000	SSO	CONAE	Success	Success
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
08-01-2018	01:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)
07-09-2017	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
07-08-2020	05:12:00			Starlink 9 v1.0, SXRS-1,			SpaceX, Spaceflight		
07-08-2018	05:18:00								
07-03-2020	04:50:00								
07-01-2020	02:33:00								
06-12-2020	16:17:08								

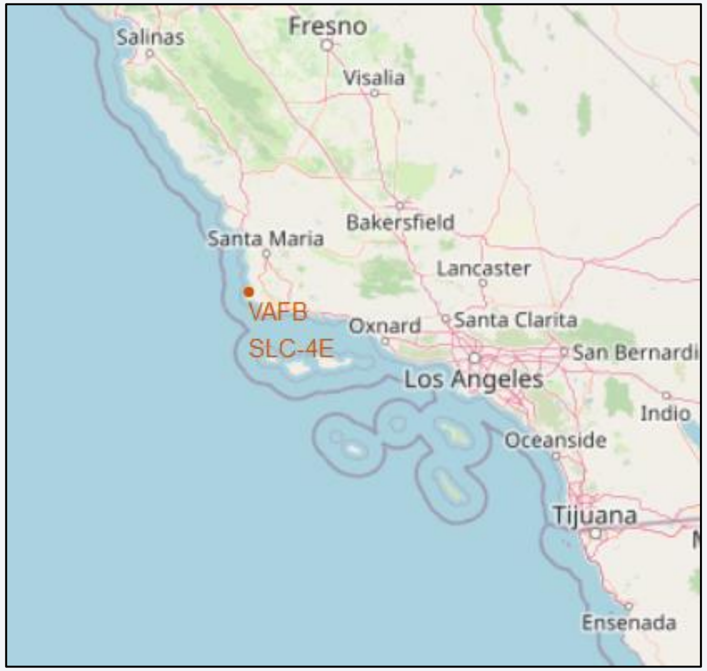
06-10-2020	11:29:34	F9 B5 B1058.3	KSC LC-39A	Starlink 12 v1.0, Starlink 13 v1.0	15600	LEO	SpaceX	Success	Success
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
05-12-2019	17:29:00	F9 B5B1059.1	CCAFS SLC-40	SpaceX CRS-19, JCSat-18 / Kacific 1	2617	LEO (ISS)	NASA (CRS), Kacific 1	Success	Success
05-11-2020	23:24:23	F9 B5B1062.1	CCAFS SLC-40	GPS III-04 , Crew-1	4311	MEO	USSF	Success	Success
04-06-2020	01:25:00	F9 B5 B1049.5	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success

Section 3

Launch Sites Proximities Analysis

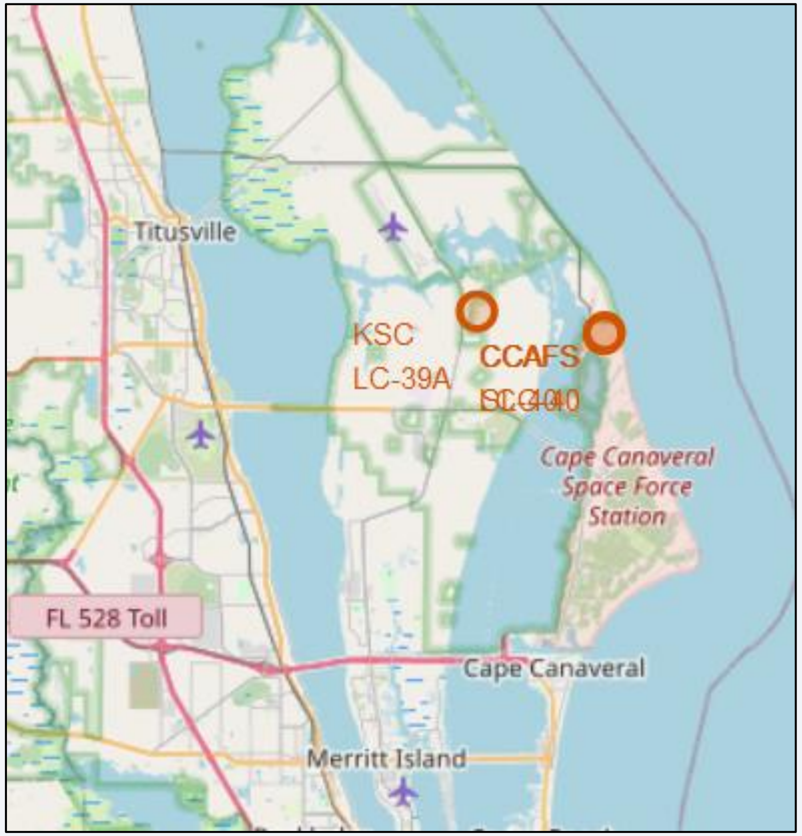
Folium Mapping – Launch Sites

- VAFC: Vandenberg Air Force Center
- KSC: Kennedy Space Center
- CCAFS: Cape Canaveral Air Force Station
- SLC: Space Launch Complex
- LC: Launch Complex

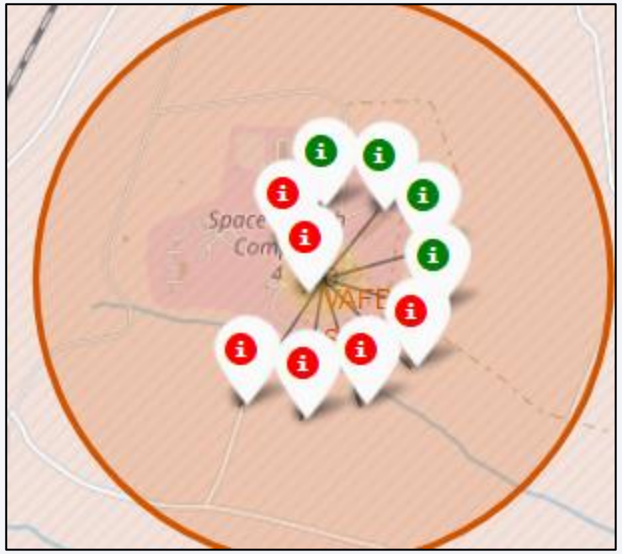


VAFC SLC-4E

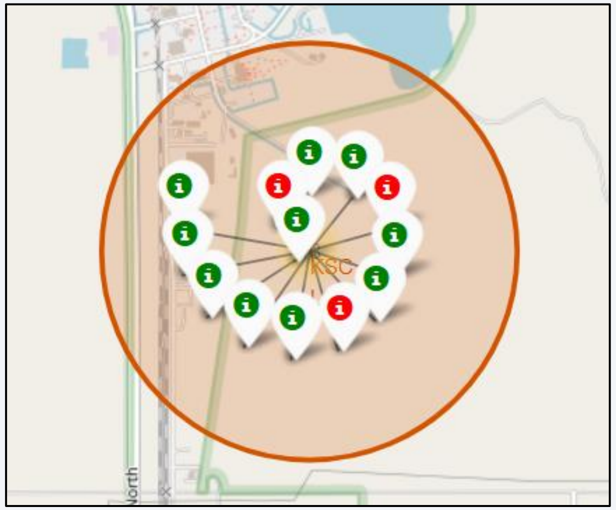
KSC LC-39A
CCAFC SLC-40
CCAFS LC-40



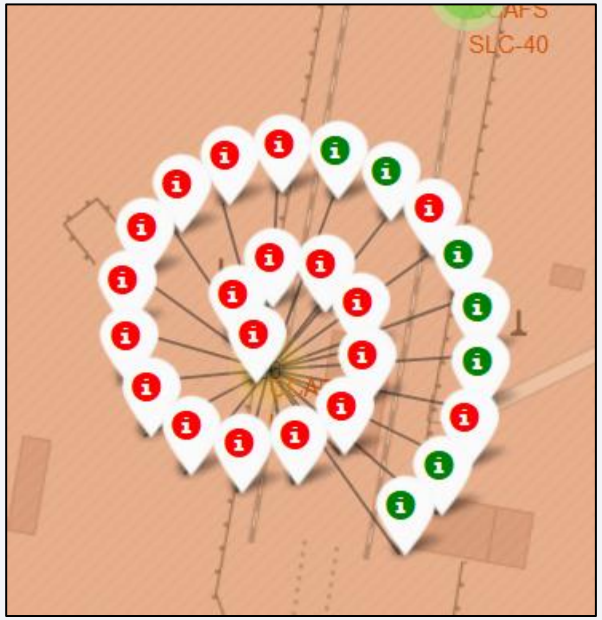
Folium Mapping – Launch Success



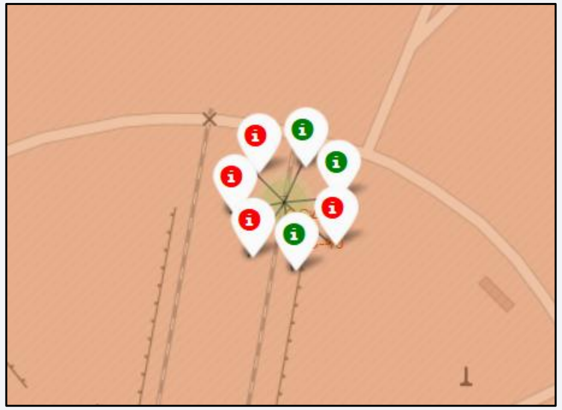
Vandenberg



Kennedy Space Center



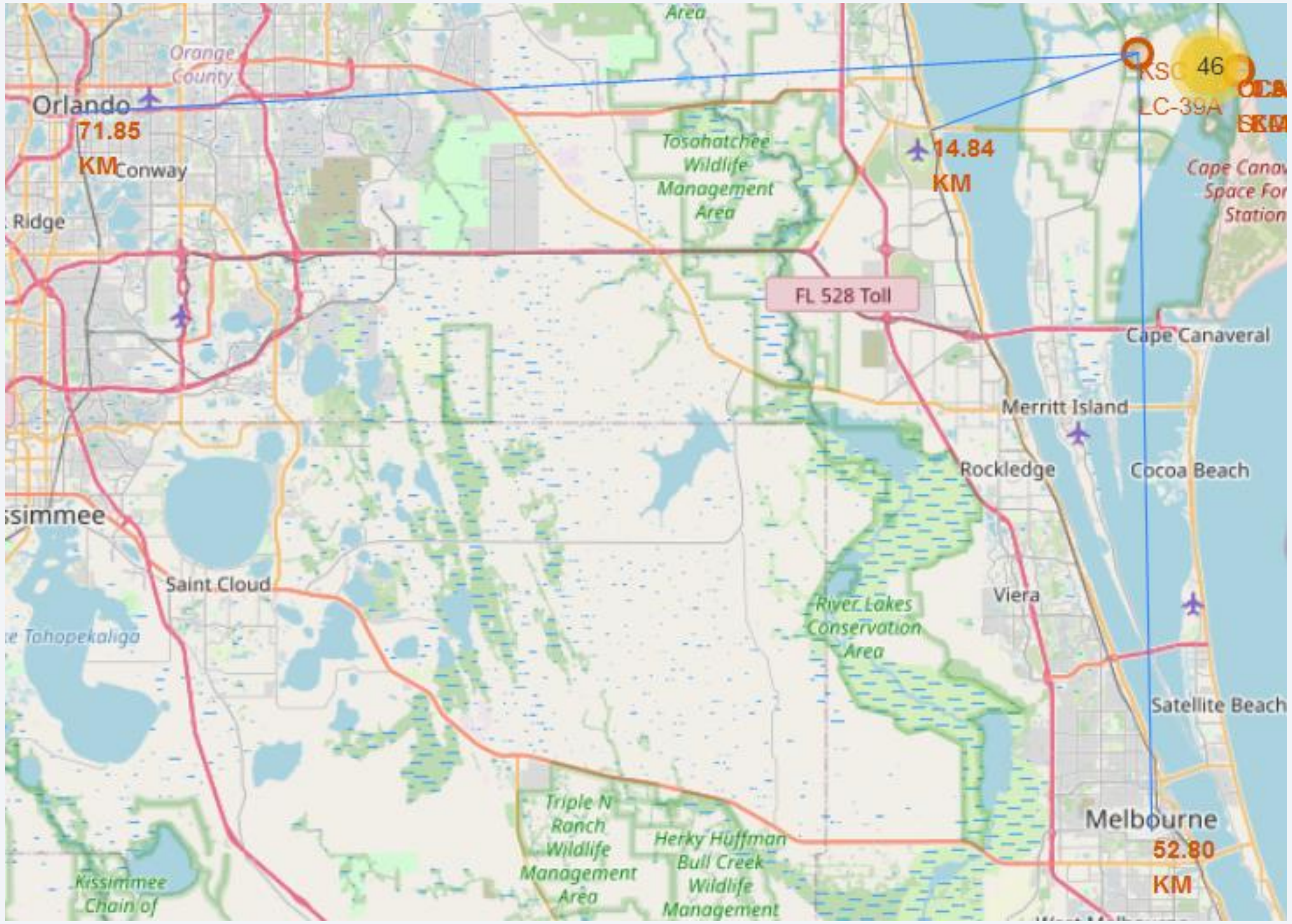
LC-40



SLC-40

- Explain the important elements and findings on the screenshot

Folium Mapping – Distance to Landmarks

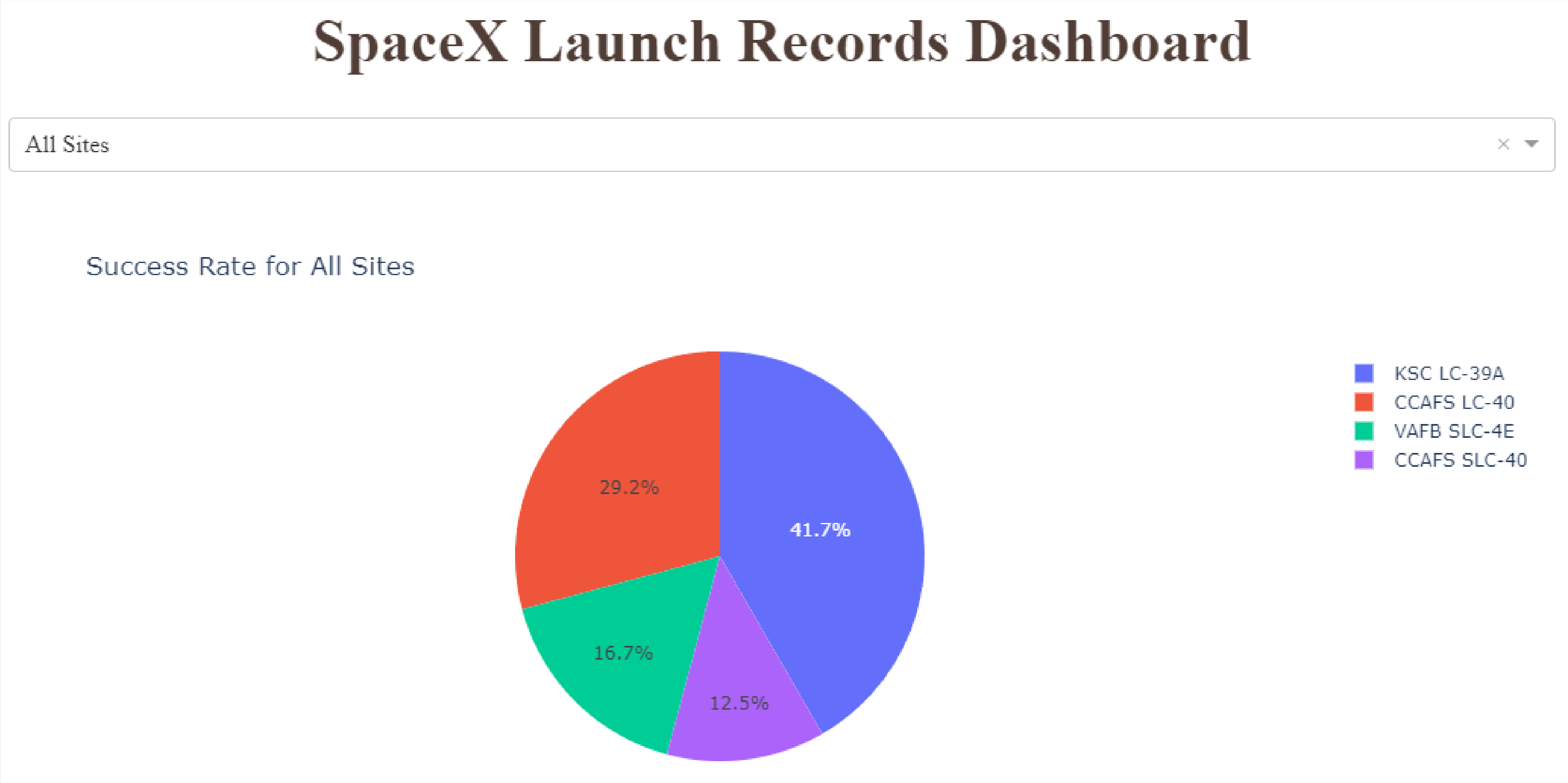


- This Folium map depicts the distance form the Kennedy Space Center to the two closest cities and the closest interstate highway.
- Not pictured: The distance from CCAFS to the coastline
- Distances:
  - Orlando: 71.85km
  - Melbourne: 52.80km
  - US1: 14.84km
  - Coastline: 0.89km

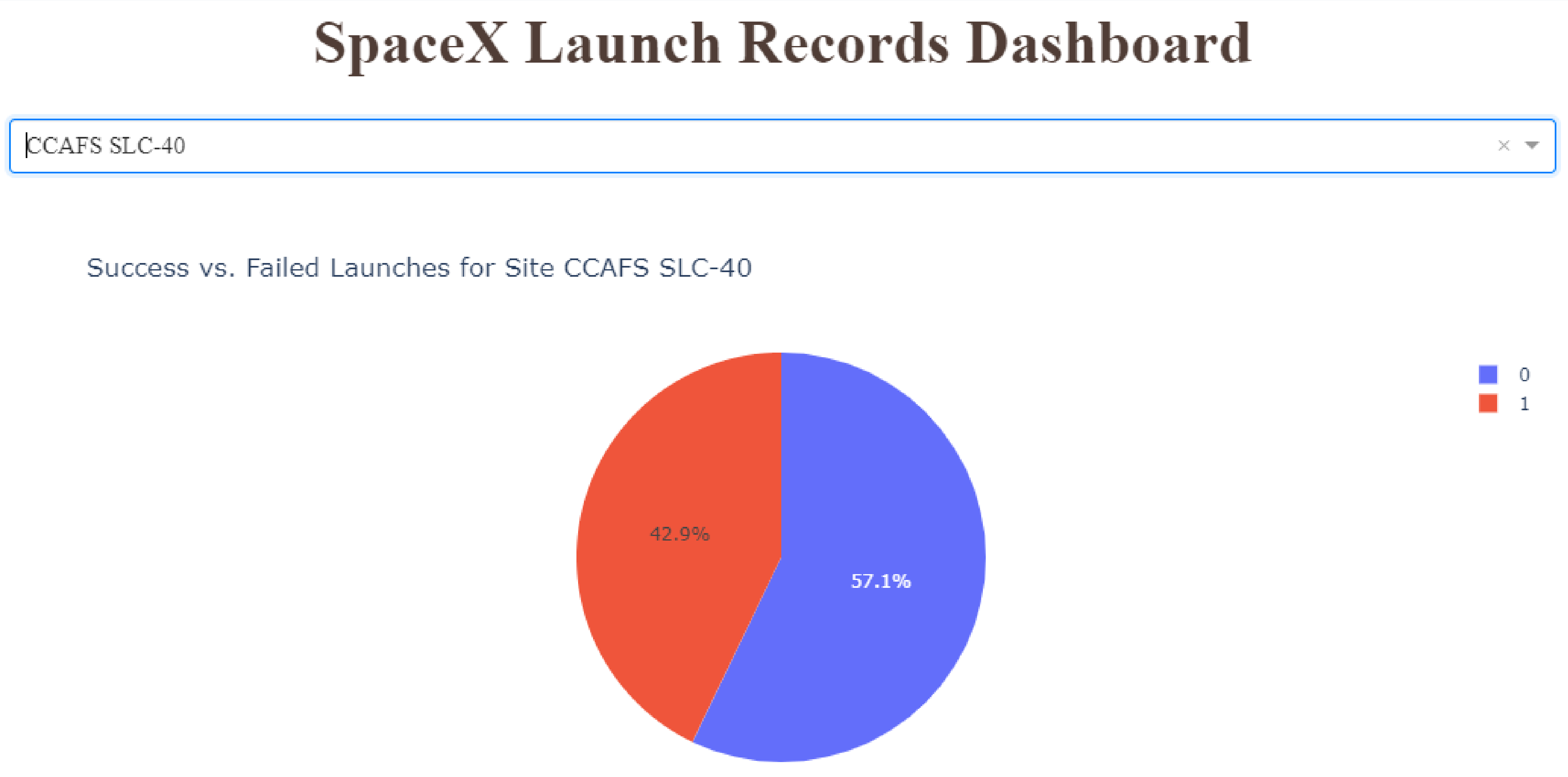
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for all Sites



Launch Site with Highest Launch Success



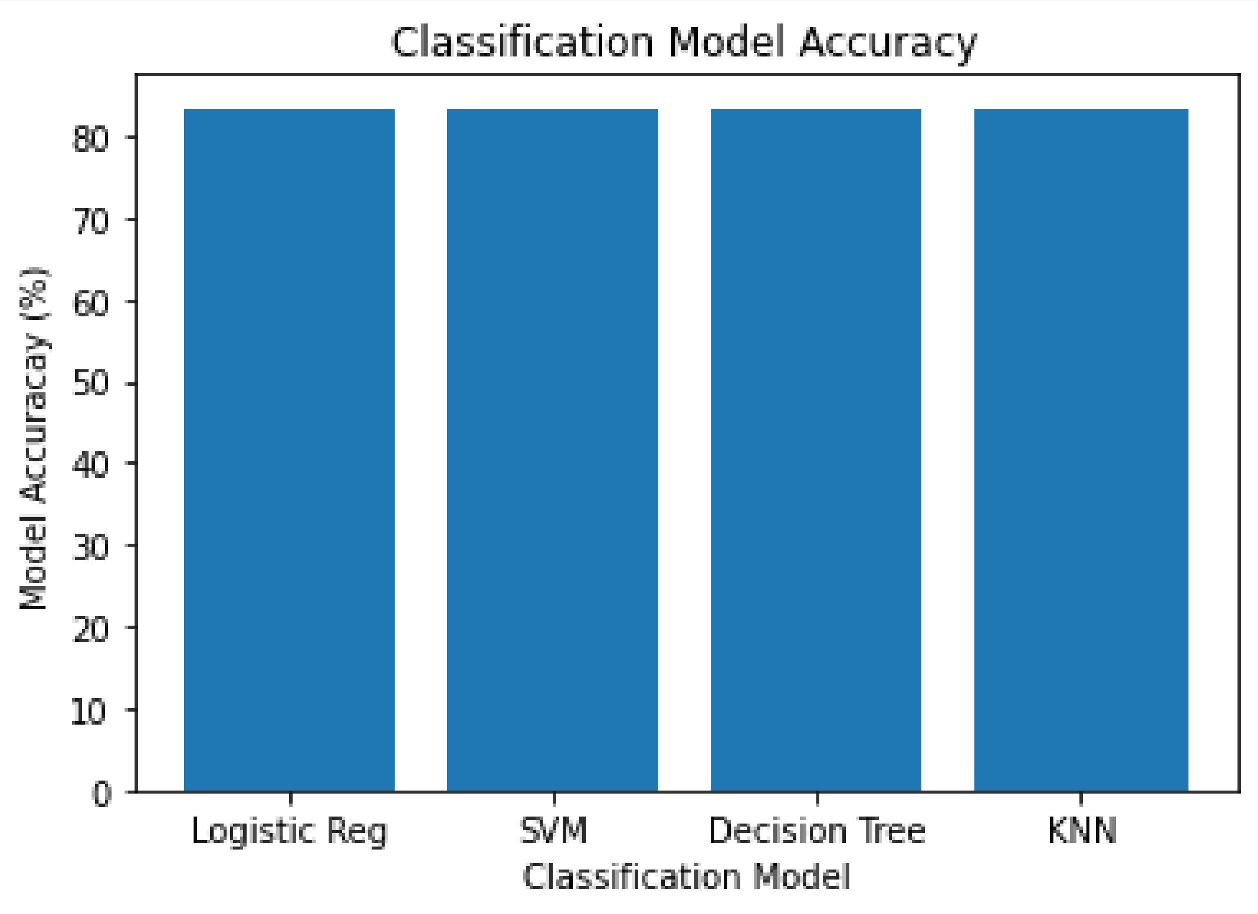
Payload vs. Launch Outcome for All Sites

Section 5

Predictive Analysis (Classification)

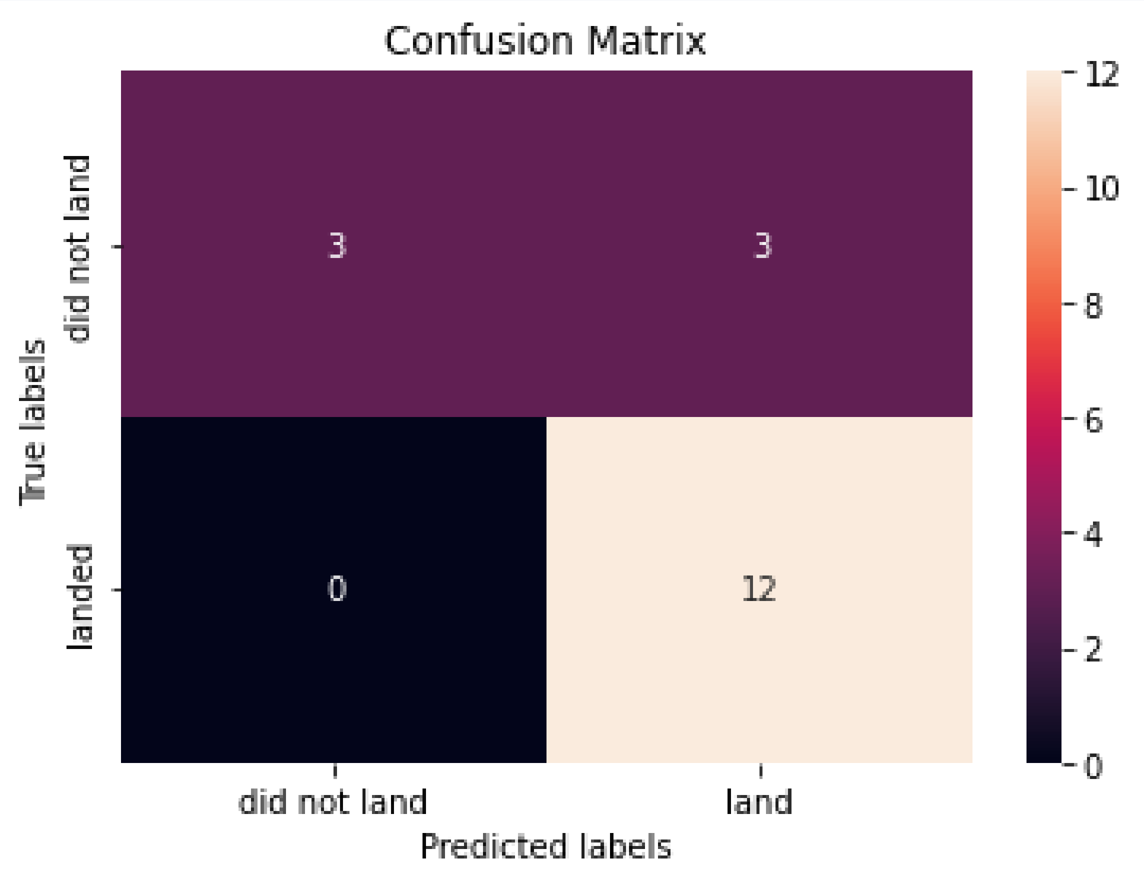
Classification Accuracy

- When used to classify the test data set, each model type has the same performance of 83.33%
- Some specific model parameters:
  - LogReg: C=1, penalty=l2, solver=lbfgs
  - SVM: C=1, gamma=0.0316, kernel=sigmoid
  - Tree: criterion=entropy, depth=4, splitter=random, features=auto, leaf=1, split=2
  - KNN: neighbors=6, algo=auto, p=1



Confusion Matrix

- As all models had the same performance, this confusion matrix represents all four



Conclusions

- The decision tree model seemed to be the only model that varied from one test to another (i.e., using different random splits. This is the best accuracy observed
- It make sense that the Logistic Regression model and SVM model have similar performance, as the SVM best performed using a Sigmoid Kernel – the same basis for a Logistic Regression model
- Despite high accuracy scores, the confusion matrix shows that the errors were in the False Positive category which, in this context are a slightly worse error; we have predicted a landing, when there in fact was a failure

Appendix

- Project Github Page: https://github.com/ndriscoll20/IBMDataScience/tree/master/DataScienceCapstone/jupyter_notebooks
- Plotly Dash python code:
- https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/python_code/capstonePlotlyDash.py
- Data Visualization with Pandas Notebook: https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_EDA_dataviz_pandas.ipynb
- Machine Learning/Landing Prediction Notebook: https://github.com/ndriscoll20/IBMDataScience/blob/master/DataScienceCapstone/jupyter_notebooks/DataScienceCapstone_MachineLearning.ipynb

Thank you!