IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael Sims
1/16/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection Using API
  - Web Scraping
  - Exploratory Analysis with SQL
  - Data Visualization and Initial Findings
  - Visual Presentation and Analysis with Folium
  - Machine Learning Modeling and Predictive Analysis
- Summary of all results
  - Initial Exploratory Analysis and Impressions
  - Interactive Data Presentation
  - Machine Learning Model Results

Introduction

- Project background and context
  - The SpaceX Falcon-9 is a reusable, two-stage rocket, whose continued successful operation would allow more cost efficient delivery of persons and material into Earth orbit. Long-term operation also creates cheaper alternative to state-funded Earth orbit delivery systems. However, the cost-efficiency of the Falcon-9 is depending on its reusability, with the major obstacle to reuse being successful landing. There are multiple variables to be examined, such as launch-site, landing site, and landing method.
  - Problems you want to find answers
    - What variables correlate to increased rates of successful landings?
    - Do any of these specific variables hold a correlation statistically significant enough to merit increased focus for future launch and landing operations?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Web Scraping from Wikipedia and using SpaceX API
- Perform data wrangling
  - Applied one-hot encoding to clean categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis conducted with Logistic Regression, SVM, Decision Tree Classifier, KNN methods

Data Collection

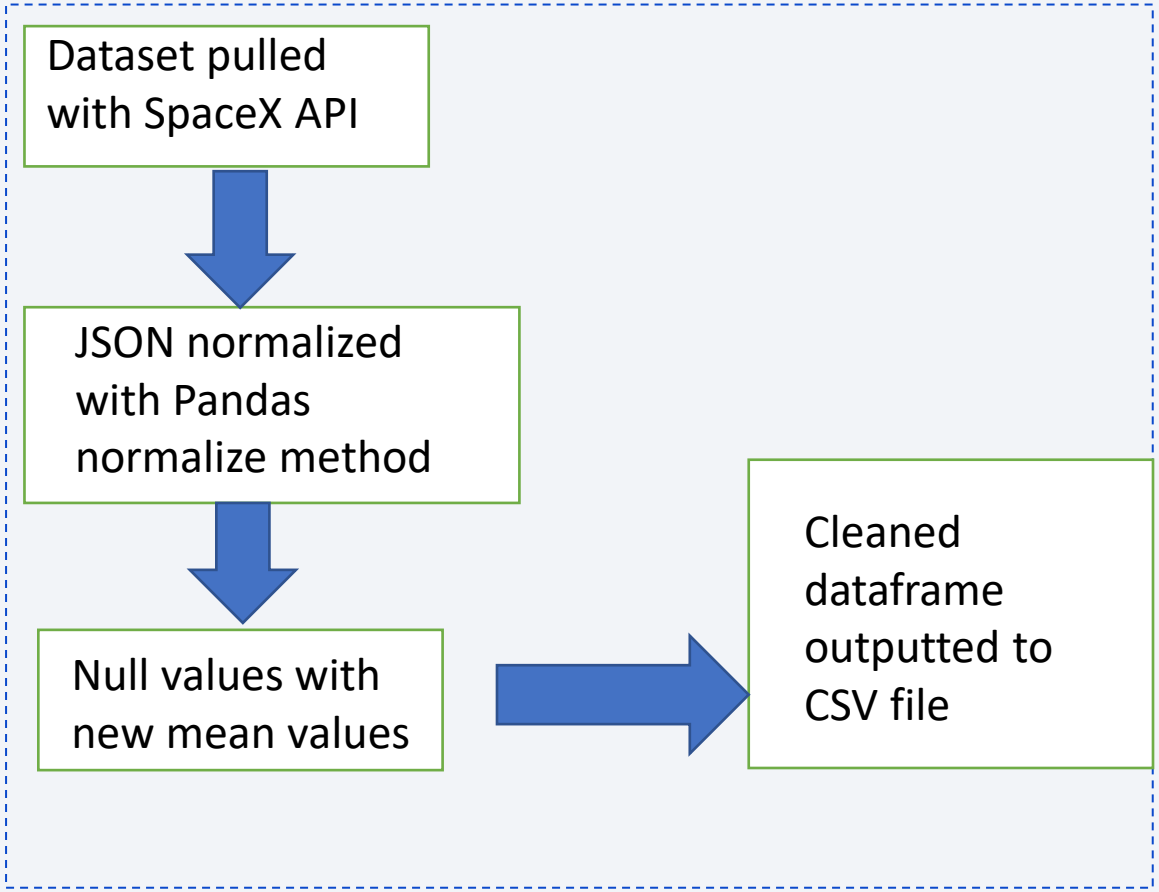
- Data scraped from Falcon9 Wiki page using BeautifulSoup library, organized according to Falcon9 model, Outputted to file for later use

- Data gathered via SpaceX API: Results formatted into .JSON file, normalized, and missing values replaced with suitable averages

Data Collection – SpaceX API

- This represents major steps in the data flow

- Notebook link: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%201/SpaceX API.ipynb

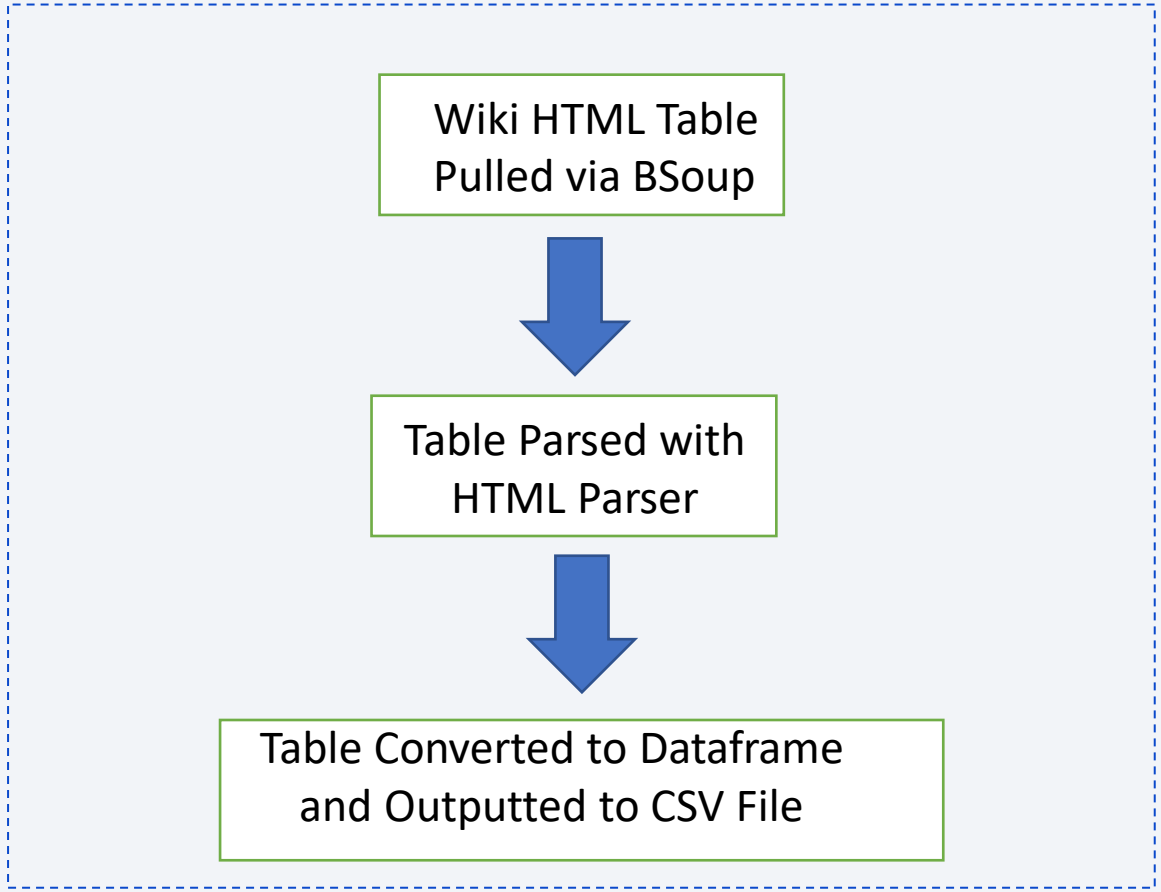


Data Collection - Scraping

- Key Steps in Webscraping Data Flow

- Notebook link:

https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%201/Webscraping%20Lab.ipynb



Data Wrangling

- First, I calculated the number of launches at each launch site, then the number and occurrence of each orbit type, and then mission outcome, creating a binary failure/success landing outcome column for better analysis; this was then outputted to a CSV file
- The notebook for this portion can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%201/Data%20Wranging%20Capstone.ipynb

In [16]:
```
df.head(5)
```

Out[16]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	I
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28

We can use the following line of code to determine the success rate:

EDA with Data Visualization

- Multiple variables were visualized for initial investigation.
  - Flight number vs. Payload mass indicated that the higher a flight number, the more likely a successful landing outcome. However, the higher the payload mass, the less likely.
  - Visualizing success based on launch site also showed that different launch sites held different success rates, with Kennedy Space Center and Vandenberg Air Force Base having a higher success rate than Cape Canaveral Space Force Station
  - A detailed visual breakdown of orbit type also indicated a correlation between orbit type and success, which was shown through bar and catplots
- The notebook can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%202/EDA%20with%20Visualization.ipynb

EDA with SQL

- The following exploratory SQL queries were performed:
  - Distinct launch sites identified
  - Investigative query on Cape Canaveral launch data performed
  - Total payload mass of NASA affiliated launches determined
  - Average payload of booster F9 v1.1 determined
  - First successful ground pad landing determined
  - Successful booster types with drone ship landing and payload between 4000 and 6000 kg determined
  - Total successes and failures determined
  - Boosters that have carried maximum payload mass determined
  - Record search for failed drone ship landings in 2015 based on booster version and launch site
  - Successful landings between June 4, 2010 and March 20, 2017
- Notebook can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%202/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Created map marking all launch sites, then marked success and failure with color labelled markers in marker clusters; then added lines showing proximity to coastline and highways
- Marking clusters based on success or failure enabled rapid visual analysis of which sites have higher success rates
- Adding proximity indicators showed proximity to cities, coastlines, and infrastructure such as highways and railways
- Notebook can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%203/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

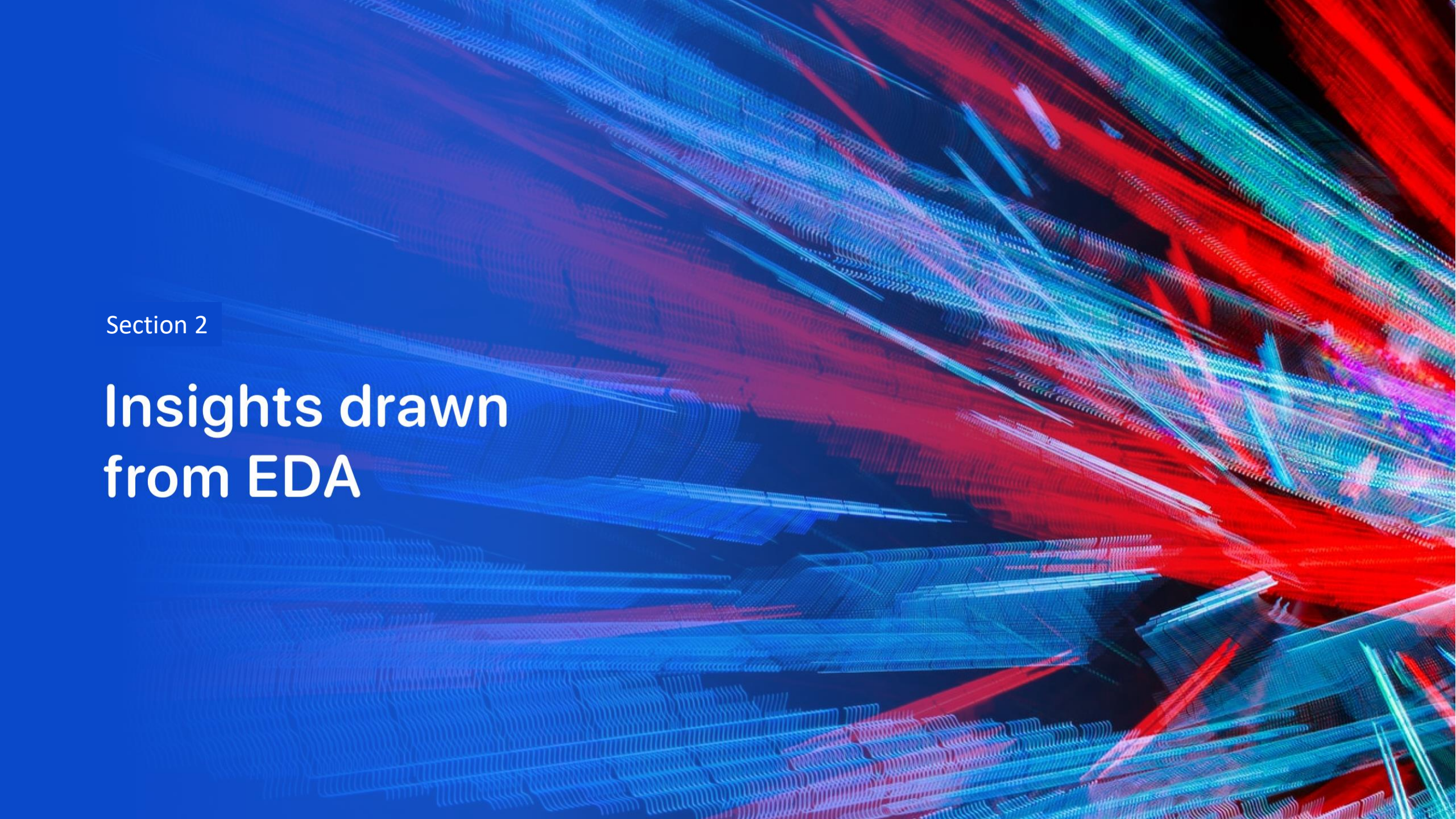
- Created interactive dashboard using Plotly
  - Created dropdown menu for launch site selection
  - Created pie chart to show success rates
  - Added slider to adjust and select payload range
  - Created scatter plot showing relationship between booster type, payload, and outcome
- The notebook containing Python code for the dashboard can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/2a036f6cf544b9a344adabfadae58fdc1cb92f2a/Capstone/Week%204/spacex_dash_app.py

Predictive Analysis (Classification)

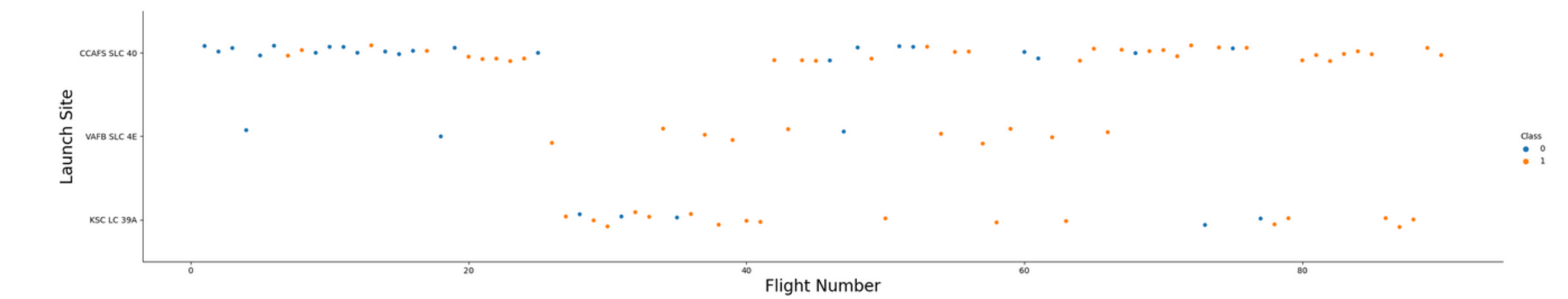
- Classification models built using Numpy, Pandas, Sklearn
- Data standardized, assigned to dataframe, and divided into training and test sets
- Four predictive models tested: Logistic Regression, SVM, Decision Tree, KNN
  - SVM determined to be most accurate
  - Model accuracy tested with Score method, visualized with confusion matrixes
- Notebook can be viewed here: https://github.com/michaelbsims/Data_Science_Tests/blob/main/Capstone/Week%204/Machine%20Learning%20Prediction.ipynb

Results

- Success increased over time; but also correlated to launch site, payload, orbit type
- SVM most useful predictive model

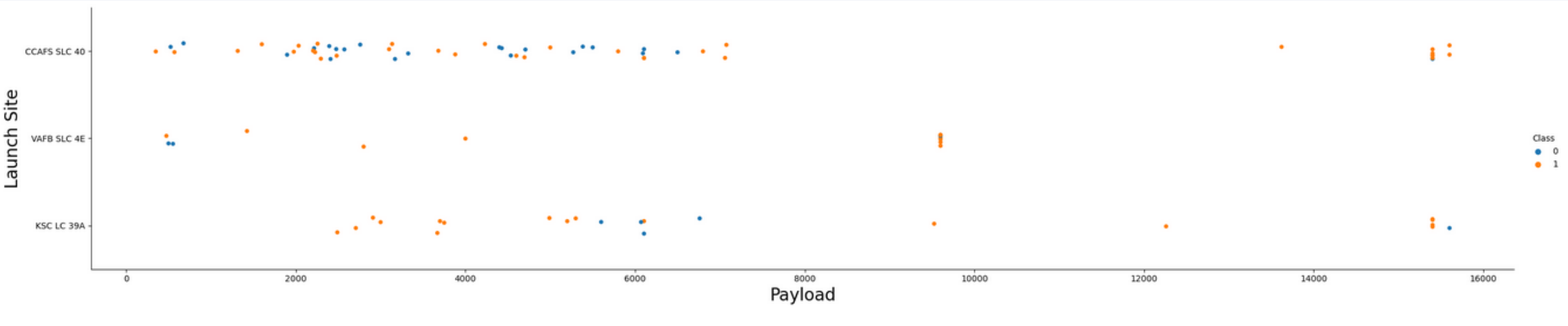

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



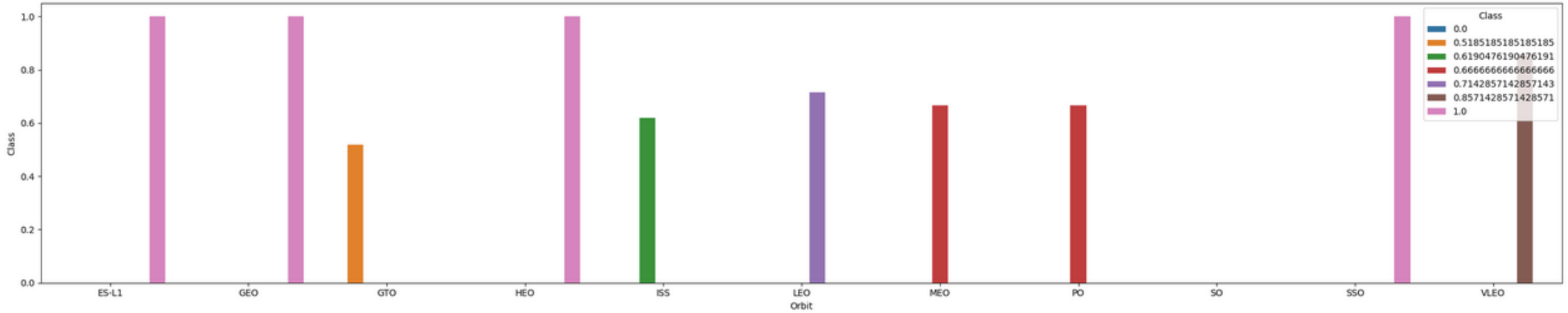
- Scatter plot of Flight Number vs Launch Site:
- Plot: 1/Orange = successful landing, Blue/0 = failure
- This indicated the higher the flight number in sequence, the more successful at all launch cites. Also indicated batch of failures at Cape Canaveral early in testing

Payload vs. Launch Site



- Scatter plot of Payload in KG vs Launch Site:
- Plot: 1/Orange = successful landing, Blue/0 = failure
- Indicates general trend of higher success with higher Payload, but not uniform at different launch sites

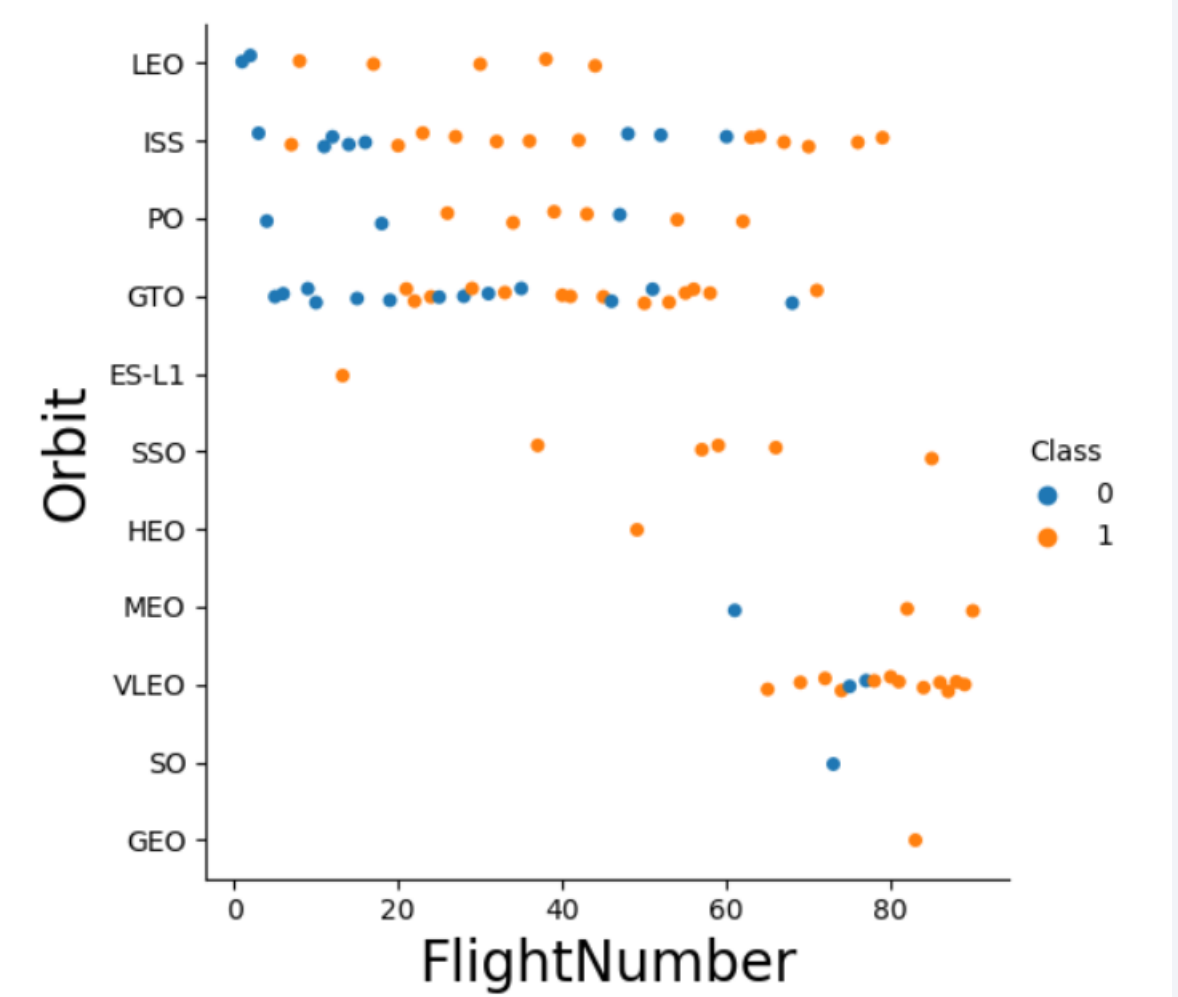
Success Rate vs. Orbit Type



- Indicates highest success rates with ES_L1, GEO, HEO, SSO orbit types

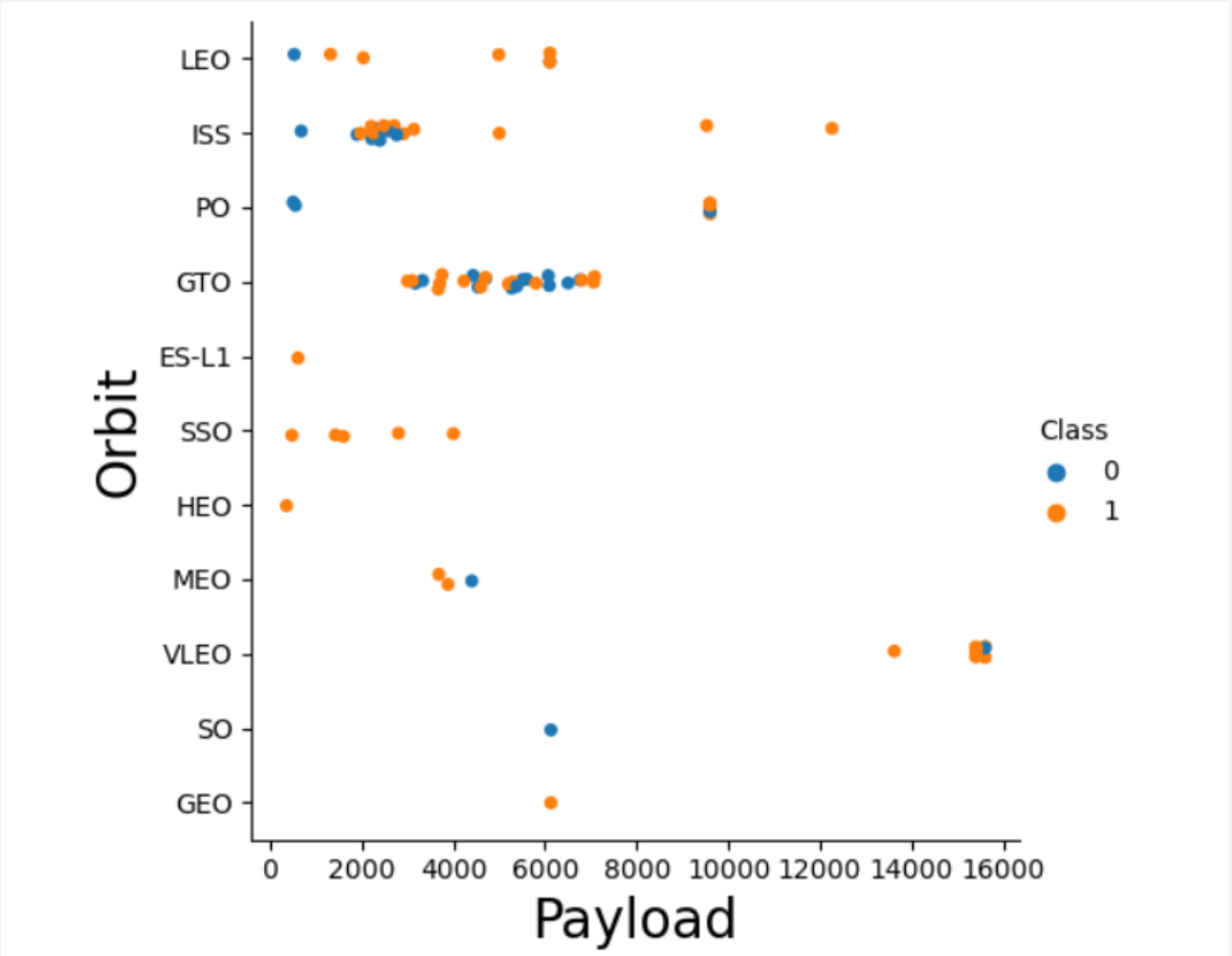
Flight Number vs. Orbit Type

- Indicates higher success rate with VLEO and LEO with increasing flight number, but varied with GTO and ISS orbits



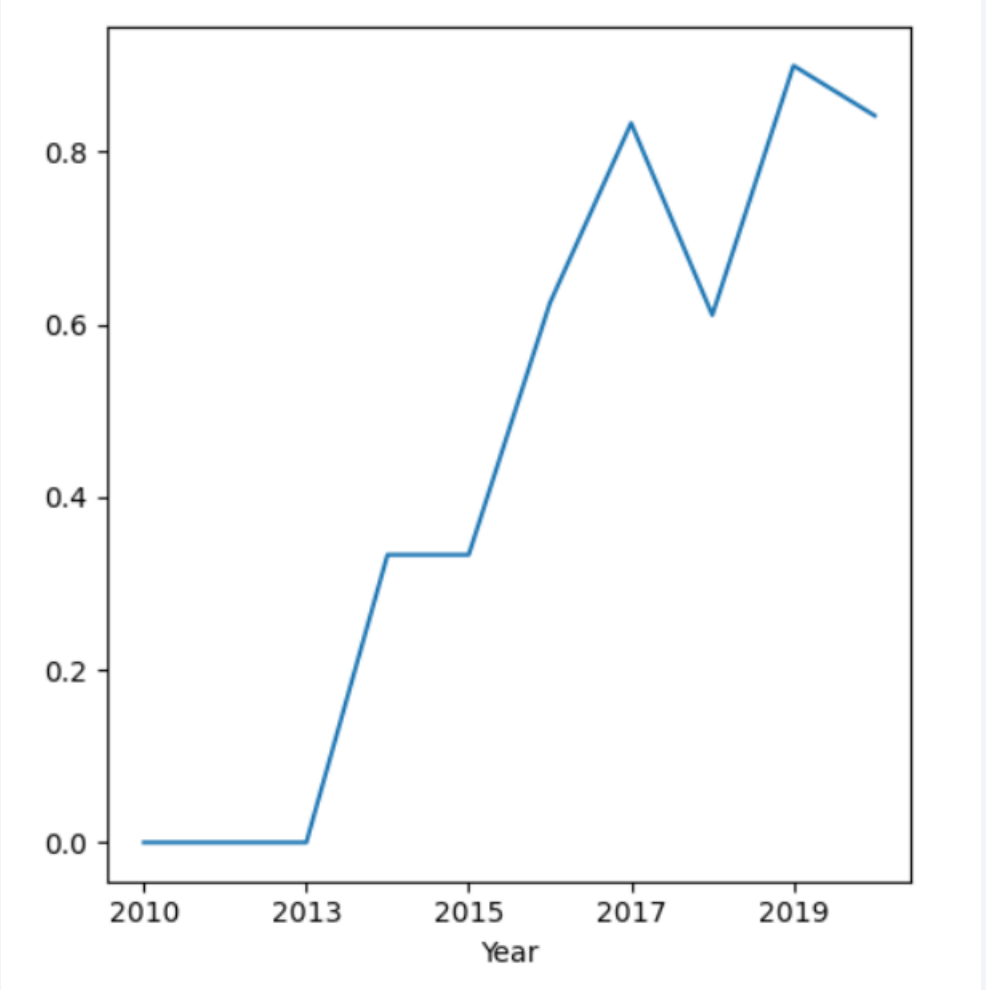
Payload vs. Orbit Type

- Indicates general success for VLEO, positive correlation between payload and success rate for LEO, ISS, but not GTO



Launch Success Yearly Trend

- Clear indication of increased success rate over time, but with drop in 2017-2018



All Launch Site Names

- Distinct launch sites drawn from SpaceXTBL

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

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

- Records beginning with CCA found using wildcard search

Total Payload Mass

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

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

SUM(PAYLOAD_MASS__KG_)
45596

- Query that determines sum of Payload Mass in KG where customer is NASA

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass with booster type F9 v1.1x determined as 2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Success (ground pad)"
```

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

MIN(DATE)
22-12-2015

- First successful ground pad landing determined as December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 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

- Four successful drone ship landings determined with payload range between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

- A total of 100 success and 1 failure mission outcomes found in table

Boosters Carried Maximum Payload

List of boosters by maximum payload

```
%sql SELECT DISTINCT(Booster_Version), PAYLOAD_MASS__KG_ FROM SPACEXTBL where PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

```
%sql SELECT Date, "Landing _Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Failure%" AND DATE LIKE "%20
```

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

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

- Information for 2015 drone ship landing failures gathered using date wildcard query

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

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") as COUNT FROM SPACEXTBL WHERE "LANDING _OUTCOME" LIKE "Success%" GROUP BY "L
```

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

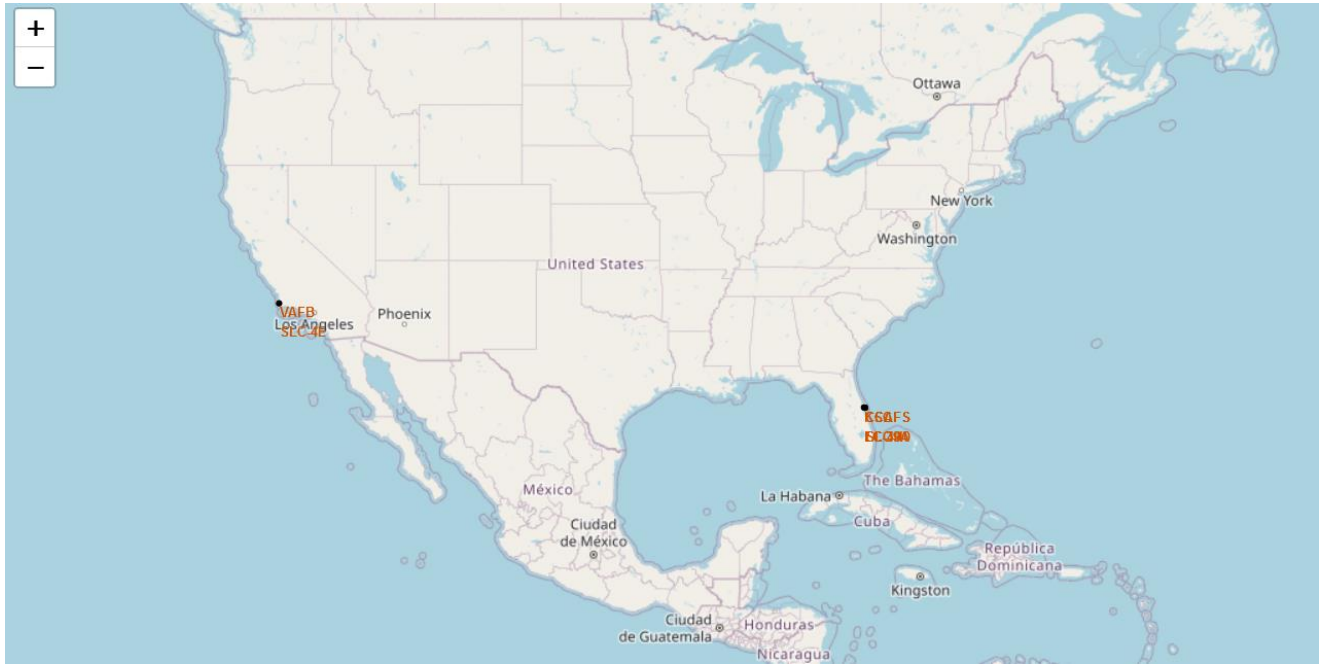
Landing _Outcome	COUNT
Success	38
Success (drone ship)	14
Success (ground pad)	9

- Drone ship landings shown as most successful, when landing type is specified


Section 3
Launch Sites
Proximities Analysis

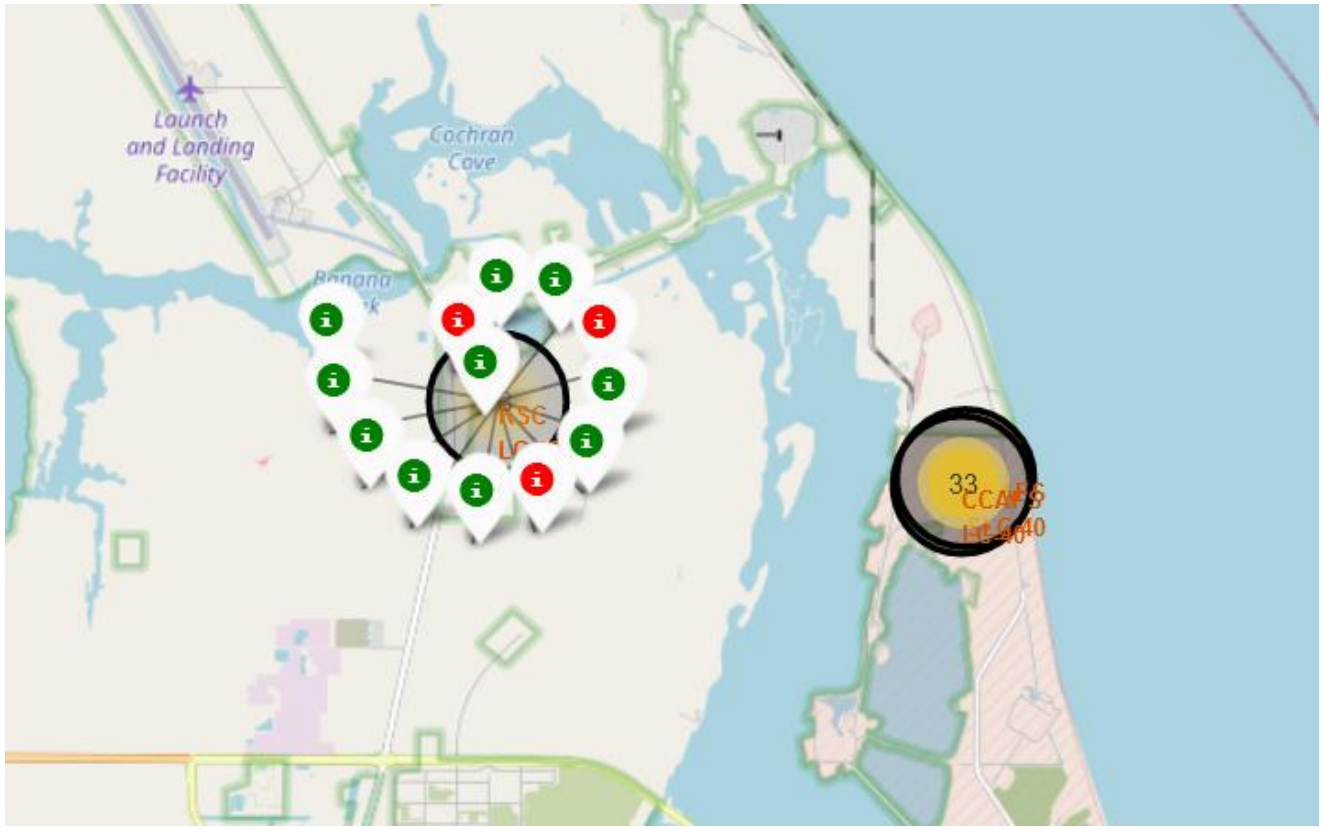
Folium Launch Site Map

- Launch sites located in coastal areas: California, Florida



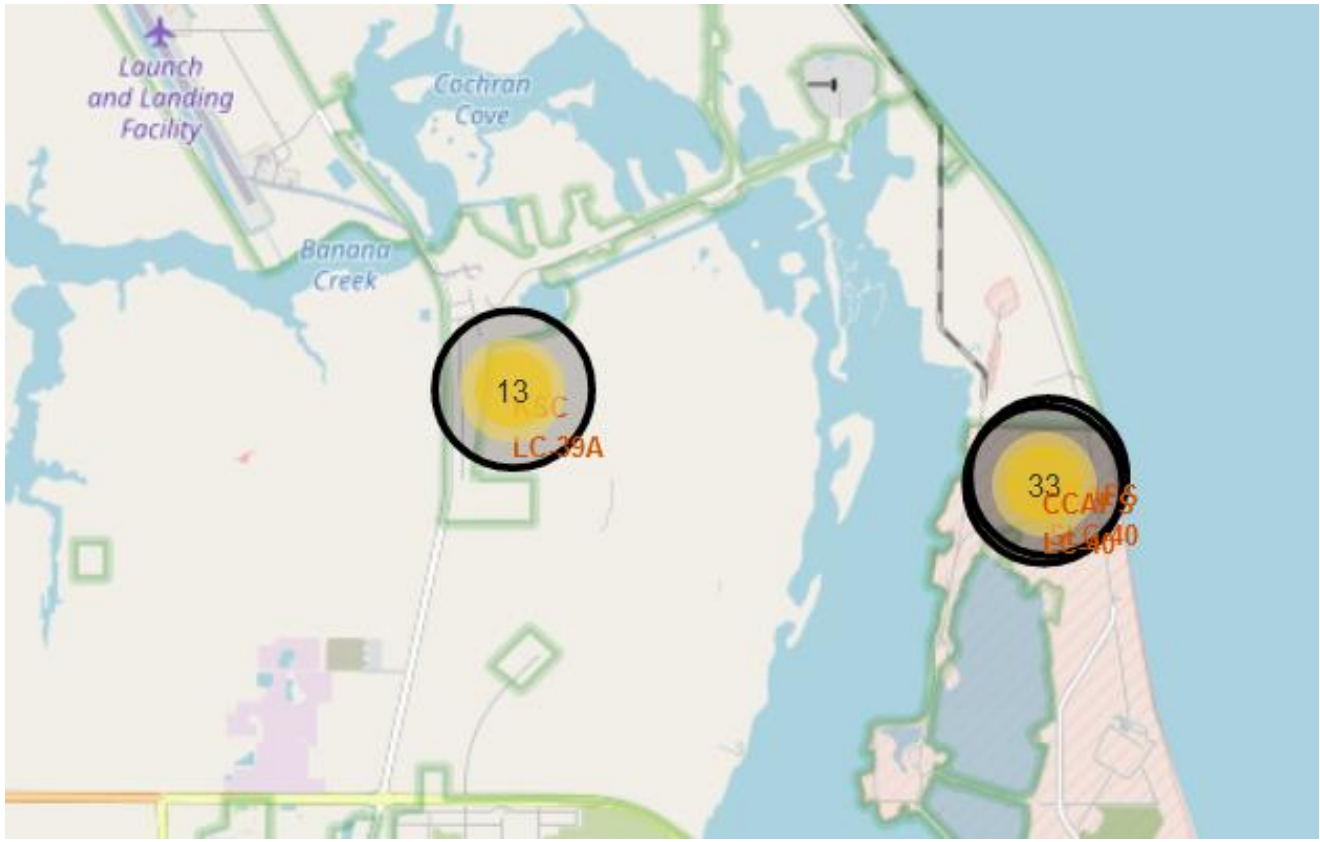
Interactive Outcome Viewer

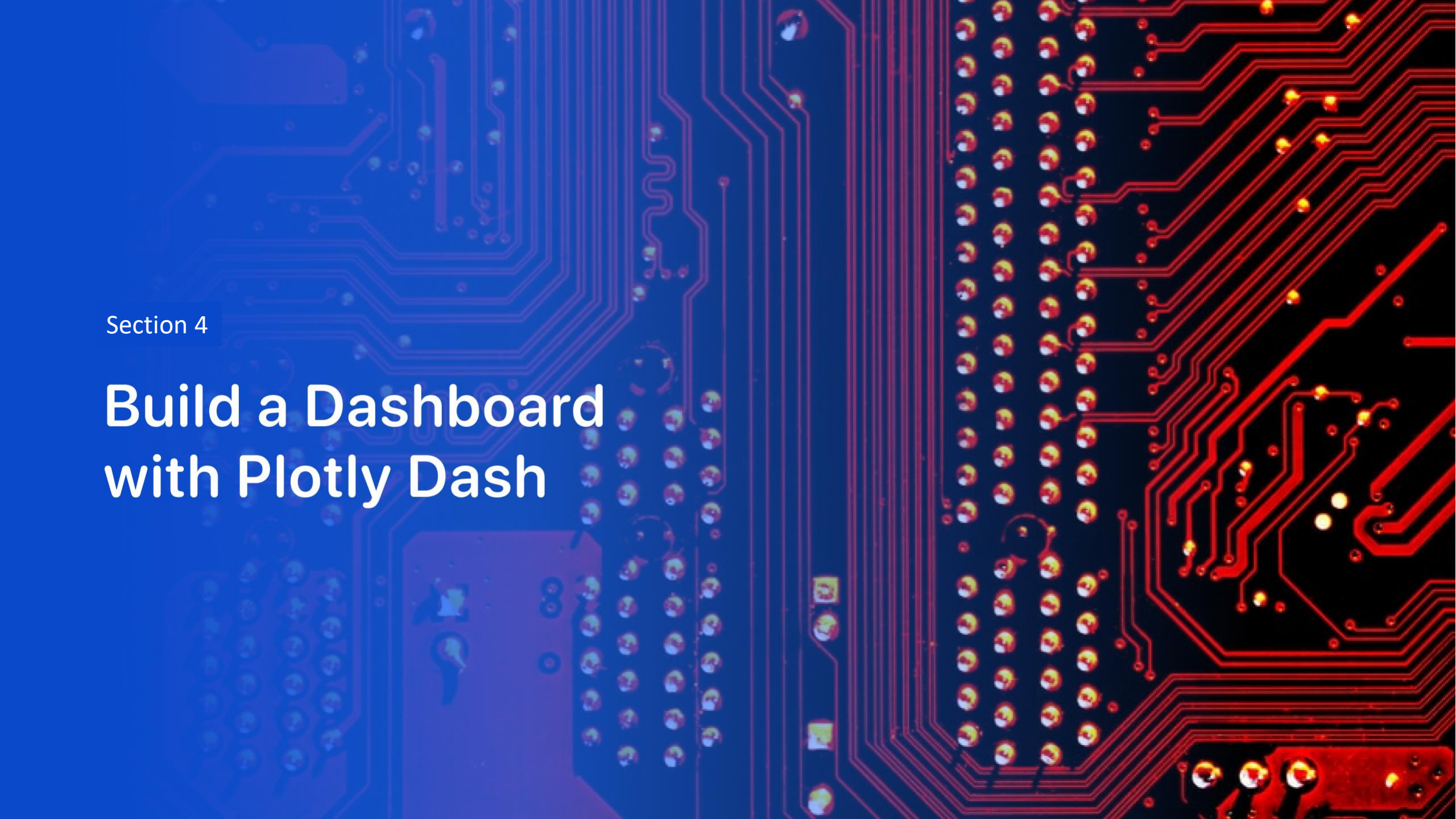
- Map allows zoom in to SpaceX launch sites
- Launch sites clickable, expand to indicate marker clusters
- Green indicates success, red indicates failure



Cost and Infrastructure Proximity

- Interactive map shows general proximity to highways, railways, coastal areas, cities

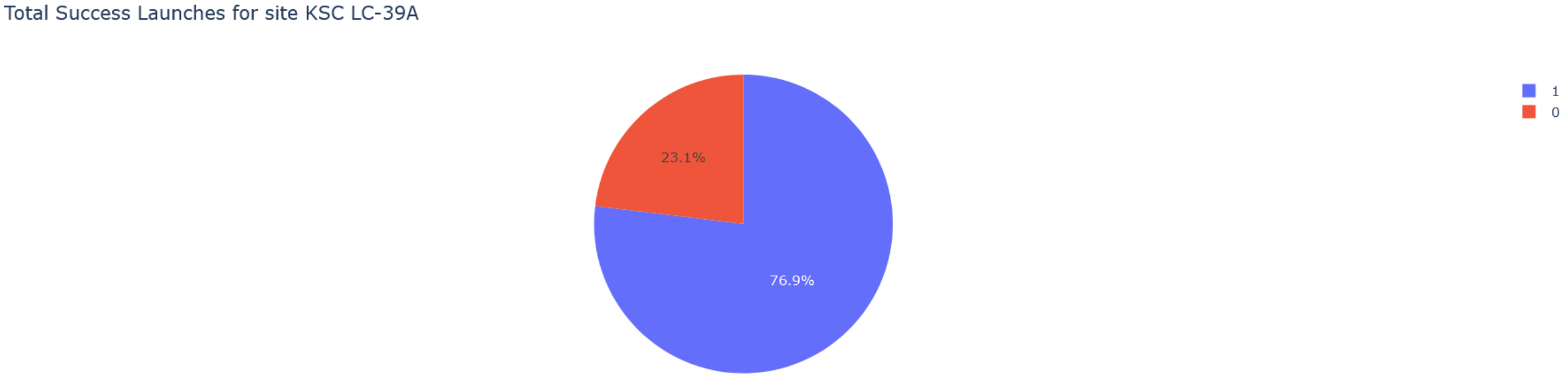



Section 4
Build a Dashboard with Plotly Dash

Success Count for Launch Sites

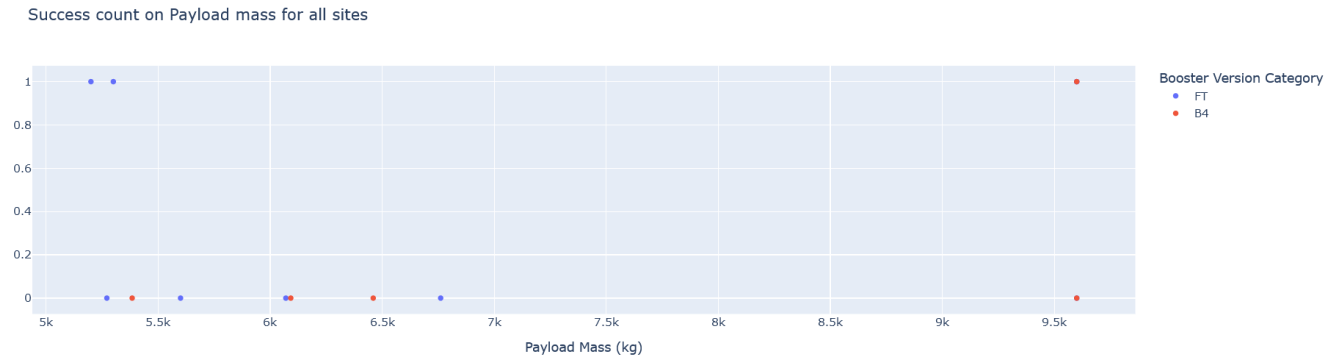
- Graph indicates Kennedy Space Center LC-39A as having largest share of successes

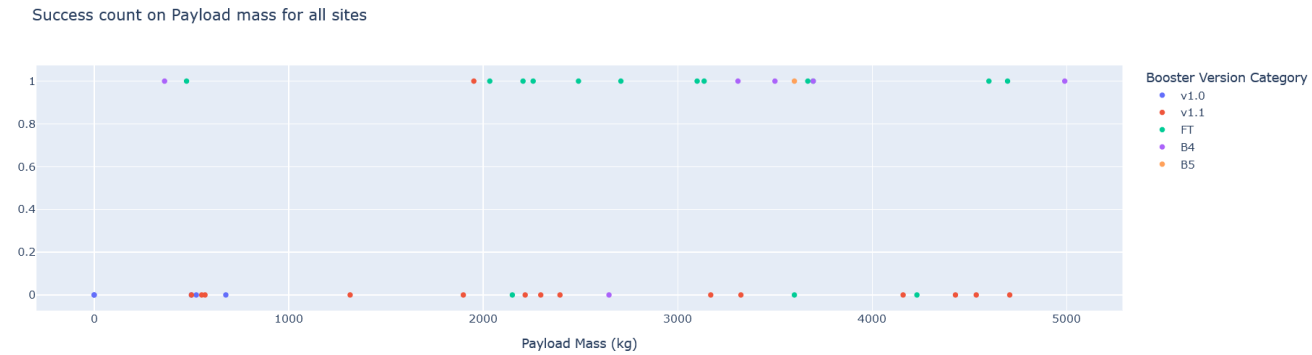
Breakdown Chart of Highest Success Site



Success by Payload Range

- Two charts indicating success by payload range: 0-5000k, 5000kg-10,000kg





Section 5

Predictive Analysis (Classification)

Classification Accuracy

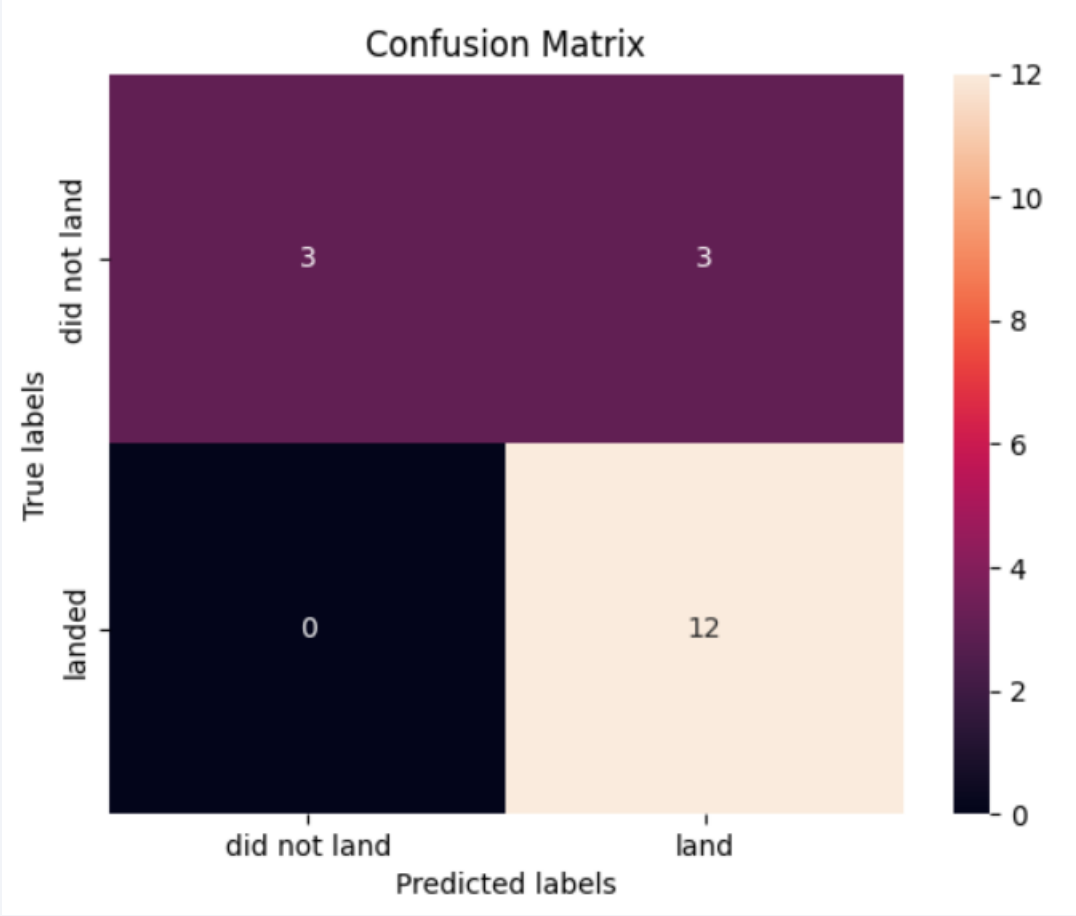
- SVM had the highest accuracy, with .889
- This table represents accuracy for the four tested methods

```
Report = {'method': testmethod, 'accuracy': accuracy}
Report = pd.DataFrame(data=Report)
print(tabulate(Report, headers = 'keys', tablefmt = 'psql'))
```

```
+----+--------------------------+------------+
|    | method                   |   accuracy |
|----+--------------------------+------------|
|  0 | Logistic Regression      |   0.833333 |
|  1 | SVM                      |   0.888889 |
|  2 | Decision Tree Classifier |   0.805556 |
|  3 | K Nearest Neighbor       |   0.861111 |
+----+--------------------------+------------+
```

Confusion Matrix

- SVM confusion matrix indicates some false positives and negatives, but ability to distinguish between classes



Conclusions

- There was an overall increase in success rate over time, with a slight drop in 2017-2018, indicating overall steady improvement of systems
- Based on this, flight number positively correlates to successful outcome
- KSC LC-39A had the largest portion successful launches
- ES_L1, GEO, HEO, SSO orbit types had the highest rates of success

Appendix

- Hyperlinked Notebooks have been included where relevant

Thank you!