IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael Kleinhans
February 4, 2022

Outline

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

Executive Summary

- The goal of this analysis is to determine whether SpaceX Falcon 9 first stage booster rockets successfully return to Earth's surface, allowing for re-use.
- Data was collected from two sources. The first being directly from SpaceX, via their API (JSON format) and the second, from a Wikipedia page containing SpaceX launch information (via web scraping).
- It was decided to create a new variable labeled 'class' and assign it a value of 1 if the launch was successful, or 0 if the launch was a failure, noting a success is defined as the booster returning to Earth in a re-usable condition.
- Twelve variables were ultimately selected as features (Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count and Serial). This list was expanded to binarize all categorical variables for modeling. The final count for prediction was eighty.
- The modeling phase utilized four different classification algorithms (K-nearest Neighbors, Decision Tree, Support Vector Machine, and Logistic Regression). The data was standardized and split into training and testing subsets prior to model fitting and subsequent testing. Multiple parameter values were run (via Grid Search) in order to obtain the most accurate representation of the ground truth.
- All models performed well and the K-nearest Neighbors algorithm was ultimately selected as the top choice (Model Accuracy of 0.8444 and R-square of 0.9444).

Introduction

- SpaceX is a well-known private company that has been actively involved in rocket launches into space. The purpose of the launches is to get items (payload) and/or people into a particular orbit. The topic of interest for this analysis is whether or not the rocket boosters are able to safely return to the Earth's surface in a condition enabling reuse. If successful, this dramatically reduces the cost of a launch. The booster of interest is the Falcon 9 first stage booster. The goal of this analysis is to use past launch data to predict whether future launches will have success returning the booster.
- Exploratory data analysis (EDA) was performed with a goal of determining two things. The first being what are we trying to predict? And the second, what information can we use to aid in the prediction. The analysis also included visual analytics. A dashboard was created to assist with viewing site specific launch success in addition to viewing the effect of booster version payload amounts on launch success.
- The analysis was considered a success in that an acceptable model was built, tested, and deemed ready for deployment for the purpose of predicting future launch outcomes.

Section 1

Methodology

Methodology

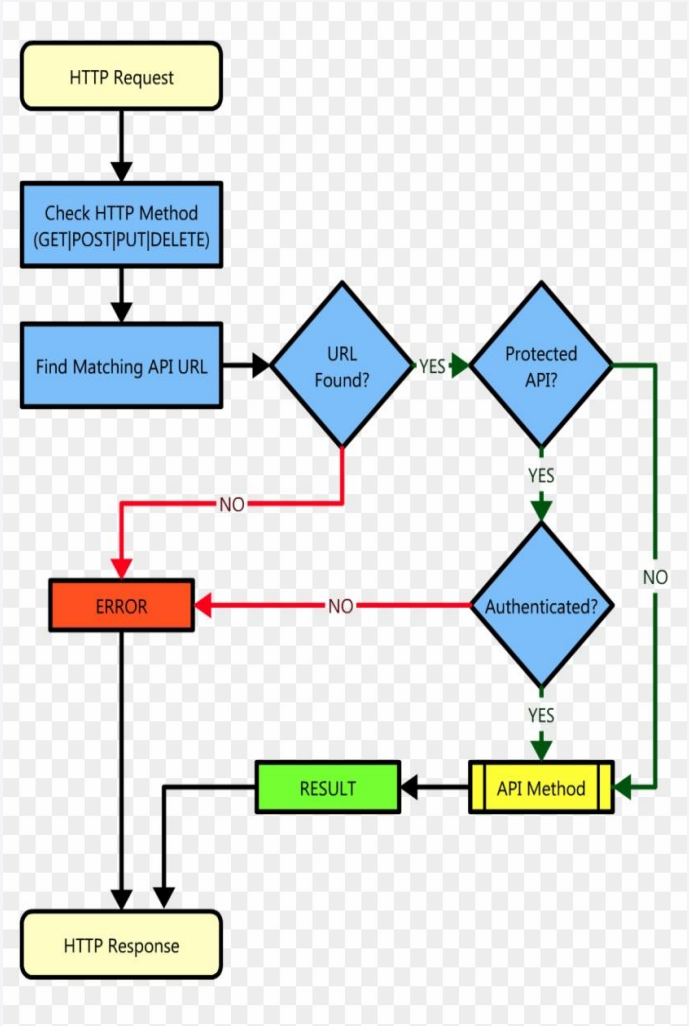
Executive Summary

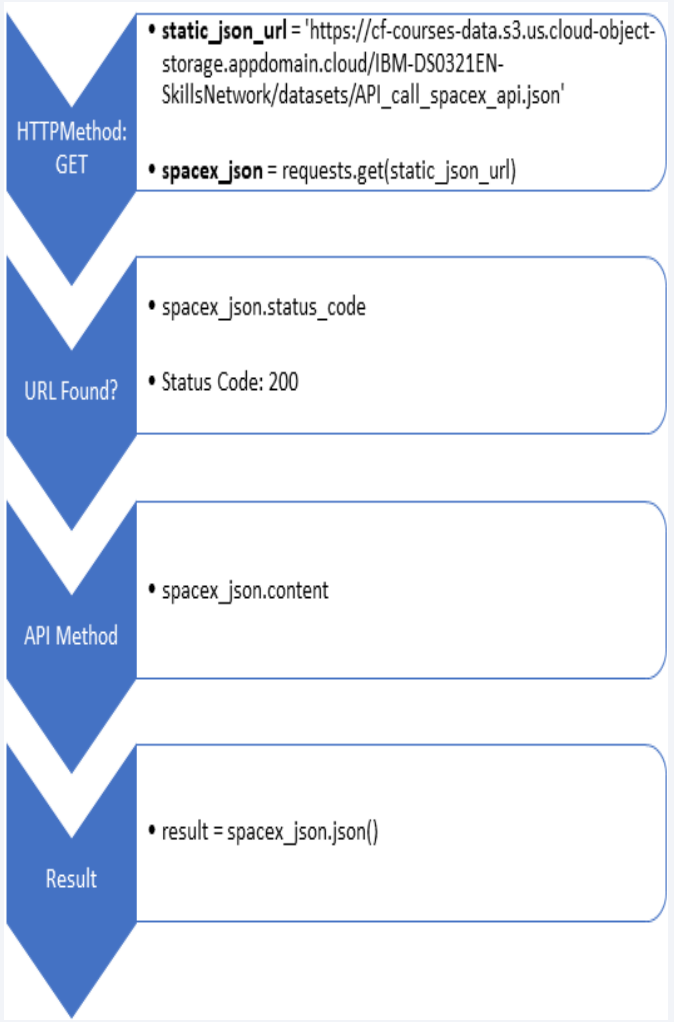
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected two ways:
  - Directly from SpaceX using their API (JSON format)
  - Web scraping a Wikipedia page containing SpaceX launch information

Data Collection – SpaceX API

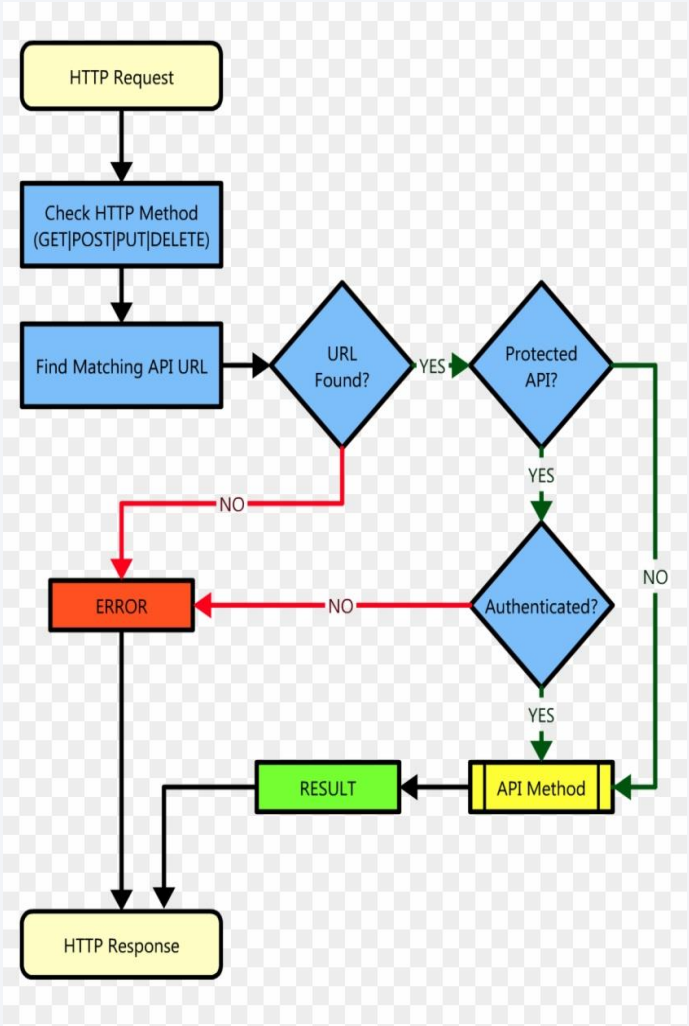




GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/Data%20Collection%20with%20Web%20Scraping%20Final.ipynb

https://www.pngfind.com/mpng/hibmJTJ_full-size-of-wiring-diagram-api-flow-diagram/

Data Collection – Web Scraping



https://www.pngfind.com/mpng/hibmJTJ_full-size_of_wiring-diagram-api-flow-diagram/

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/Data%20Collection%20API%20Final.ipynb

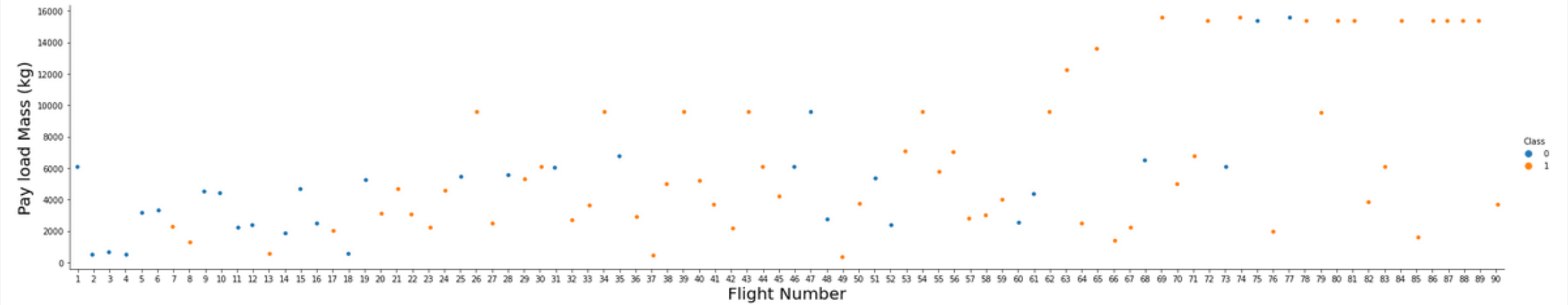
Data Wrangling

- Found missing data for 'Payload Mass' variable
  - Imputed using mean
- Created label variable 'Class'
  - Determined 'successful' vs 'failure' landing outcomes
  - Assigned binary 0 for 'failure'
  - Assigned binary 1 for 'successful'
- Created feature set from significant variables
  - Created design variables for feature set
- Cast all variables to float in preparation for modeling

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/EDA%20Final.ipynb

EDA with Data Visualization

- Category plots were created to highlight any correlations among variables, specifically the label (Class) variable against potential feature variables.
- Below is a visual showing that as the number of flights gradually increase over time, not only are higher payloads possible but the success rate dramatically improves as well.



GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/EDA%20with%20Visualization%20Final.ipynb

EDA with SQL

- SQL queries included:
  - Display the names of the unique launch sites
  - Display 5 records where launch sites begin with 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome (ground pad) was achieved

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/jupyter-labs-eda-sql-coursera%20Final.ipynb

EDA with SQL

- SQL queries included:
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, their booster versions, and the launch site names for 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/jupyter-labs-eda-sql-coursera%20Final.ipynb

Build an Interactive Map with Folium

- Marked launch sites on map (circles)
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Added launch outcome clusters for each site (markers)
  - Failure = 'red'
  - Successful = 'green'
- Calculated distances from launch sites to the coastlines (markers/lines)

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20Final.ipynb

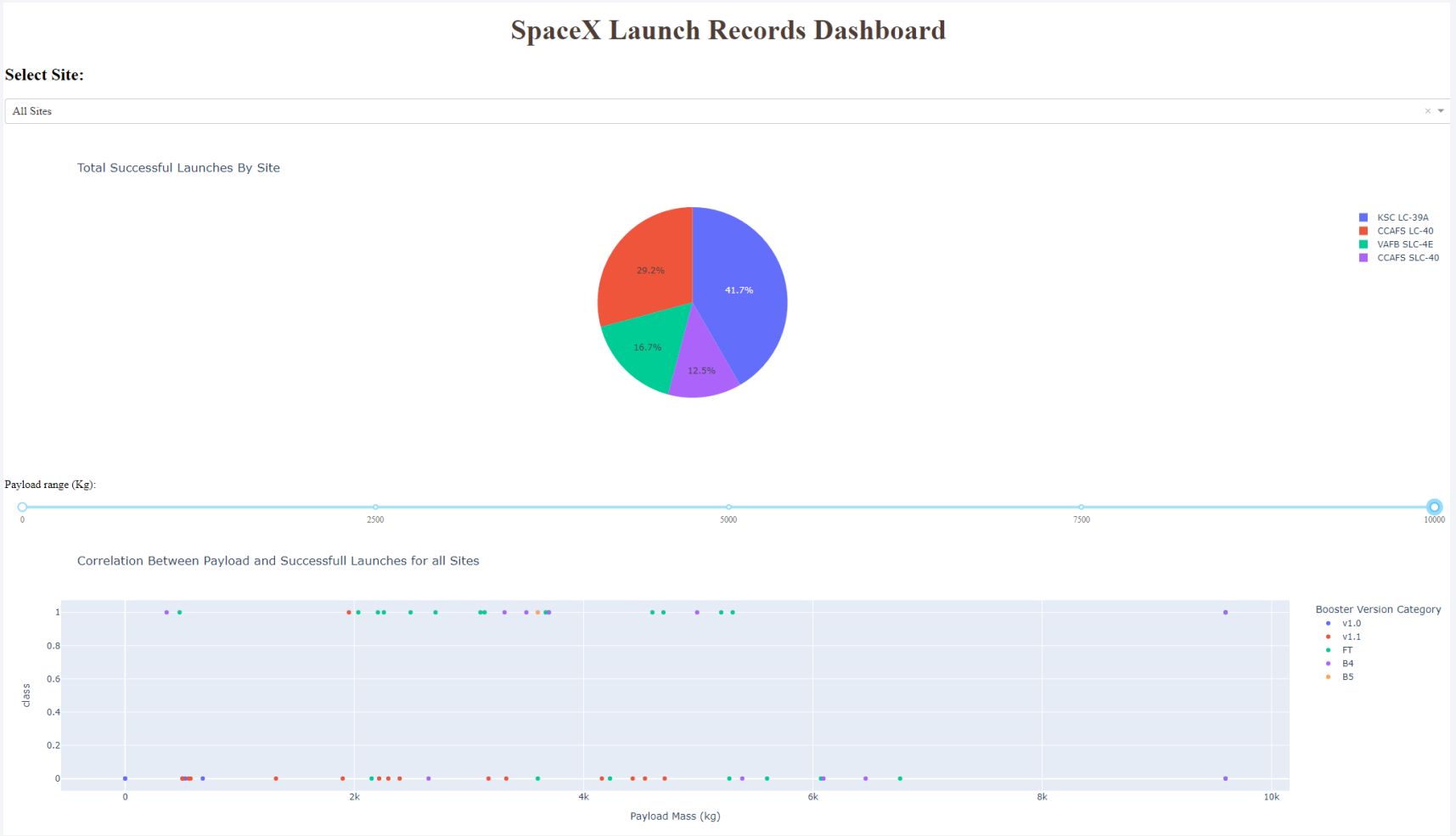
Build a Dashboard with Plotly Dash

- Added 'Launch Site' dropdown menu input component
- Added Pie Chart showing launch success for selected 'Launch Site'
- Added 'Payload' range slider input component
- Added a Scatter Plot showing booster version launch success by 'Payload'

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/SpaceX%20Launch%20Records%20Dashboard

Dashboard enables quick viewing of the following:

- Launch site success rate (pie chart)
- Booster version success rate (scatter plot)
- Effect of Payload on launch success rate (scatter plot)



Predictive Analysis (Classification)

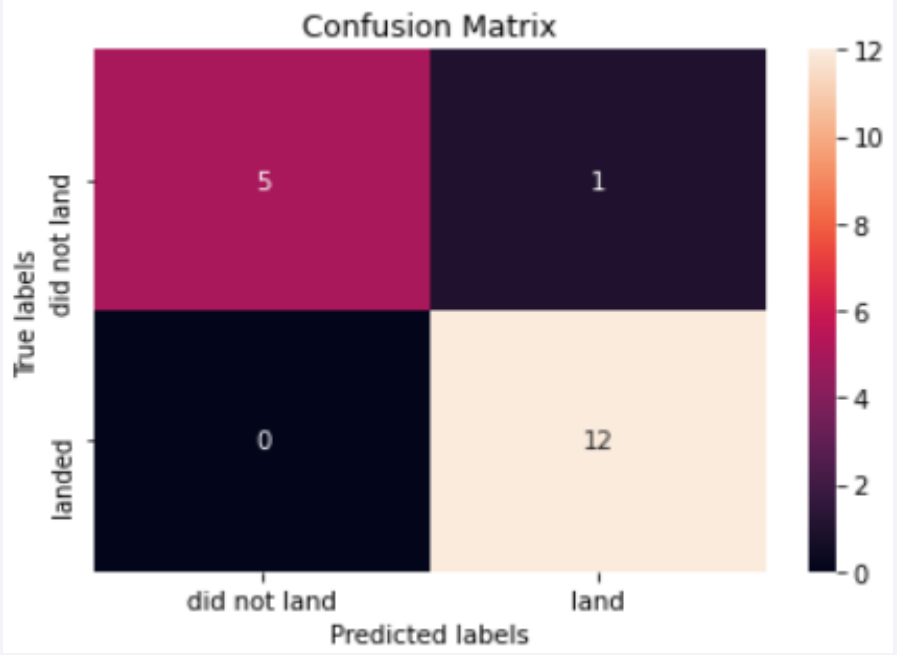
- The flowchart summarizes how the models were built and evaluated
- This is an iterative process of continuous refinement until the best performing model is attained
- Four different classification models were run:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-nearest Neighbors

GitHub URL: https://github.com/mck757/IBM-Capstone-Public/blob/main/Machine%20Learning%20Prediction%20Final.ipynb

Results

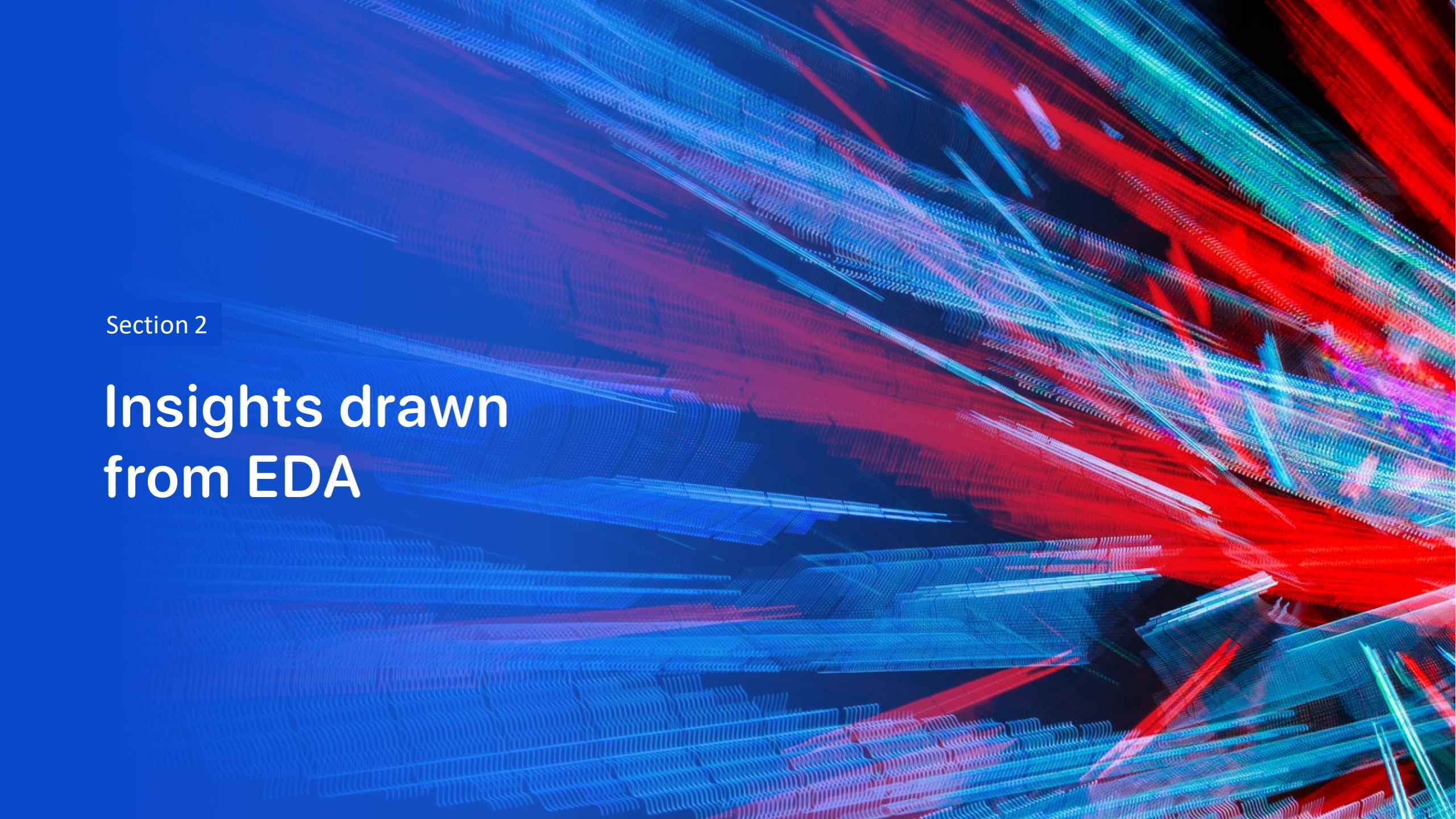
- The chart summarizes the model results
- The k-nearest Neighbor classification model had the highest score (R-Square) and subsequent highest model accuracy.
- Looking at the confusion matrix, the model only predicted one outcome incorrectly

Algorithm	Accuracy	R-Square
KNN	0.8444	0.9444
Decision Tree	0.8778	0.8333
SVM	0.8222	0.9444
LogisticRegression	0.8222	0.9444

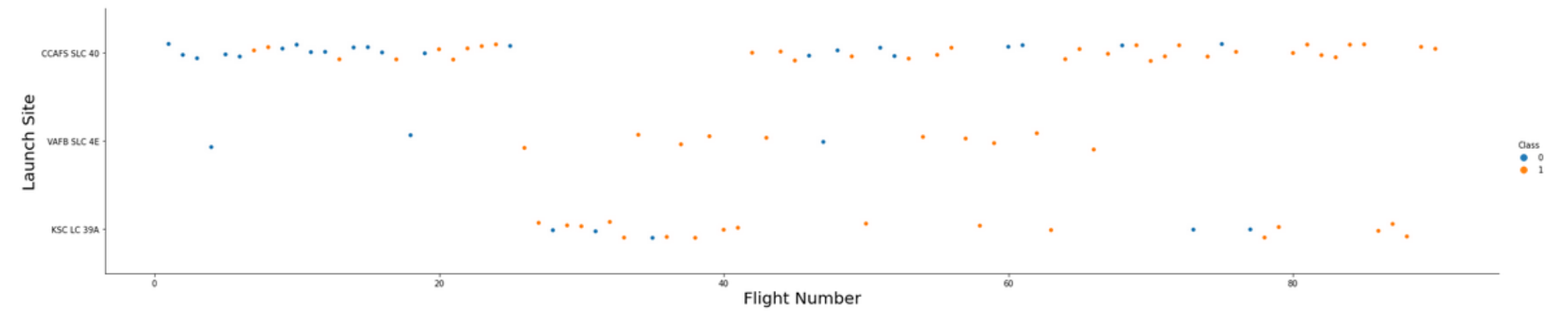


Section 2

Insights drawn from EDA

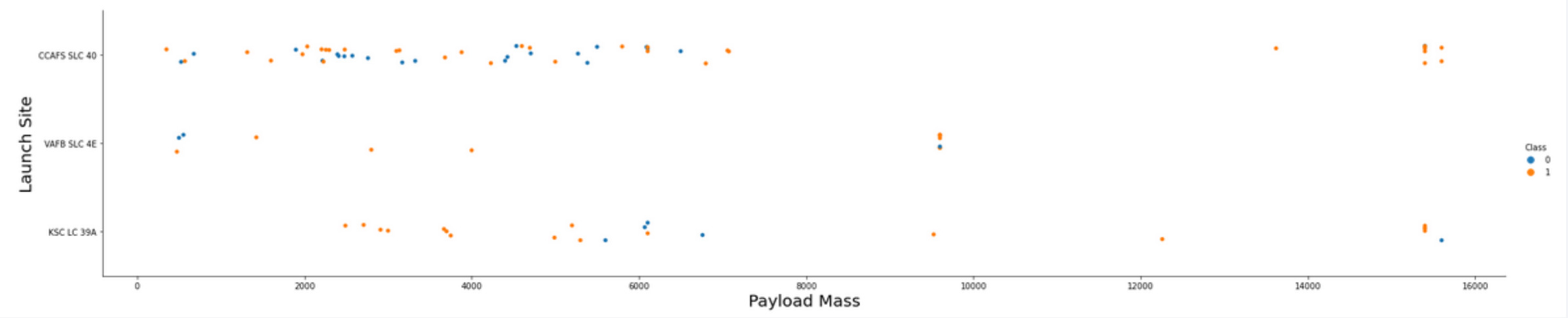


Flight Number vs. Launch Site



- Successful launches are in yellow (1)
- Failed launches are in blue (0)
- Generally, as the number of launches increases, the success of those launches also increases

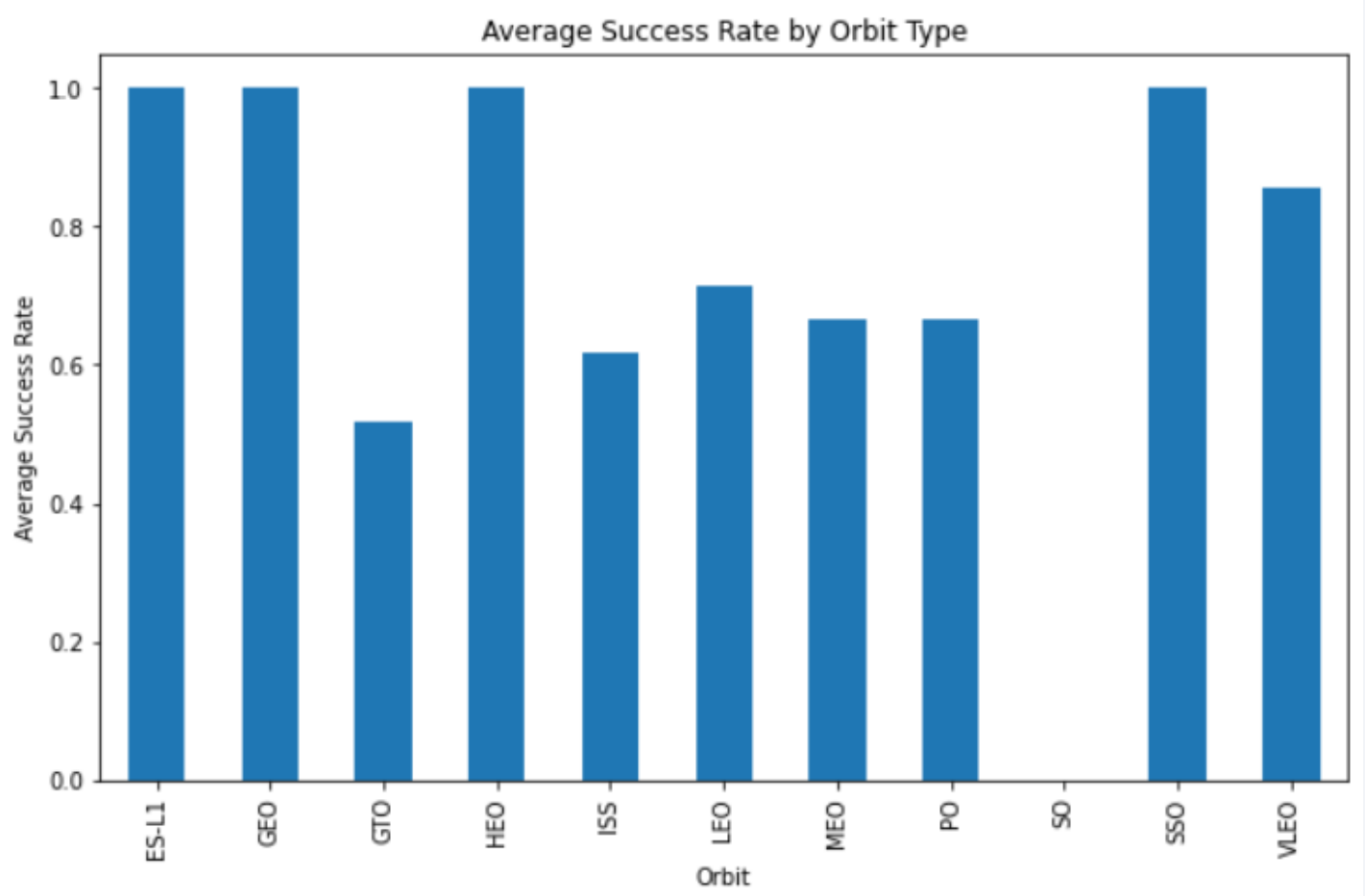
Payload vs. Launch Site



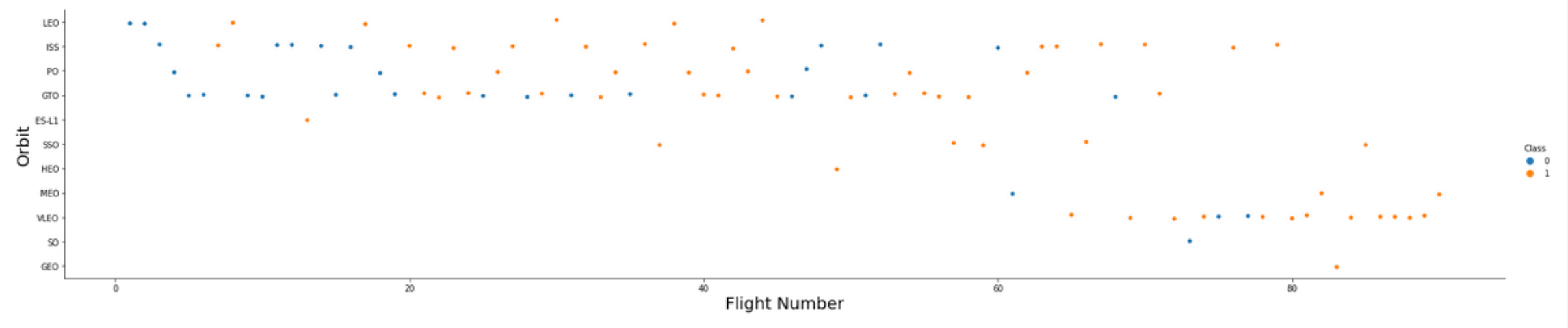
- Successful launches are in yellow (1)
- Failed launches are in blue (0)
- Generally, there does not appear to be any correlation between launch success and payload mass

Success Rate vs. Orbit Type

- Orbits with the highest success rate include:
  - ES-L1
  - GEO
  - HEO
  - SSO

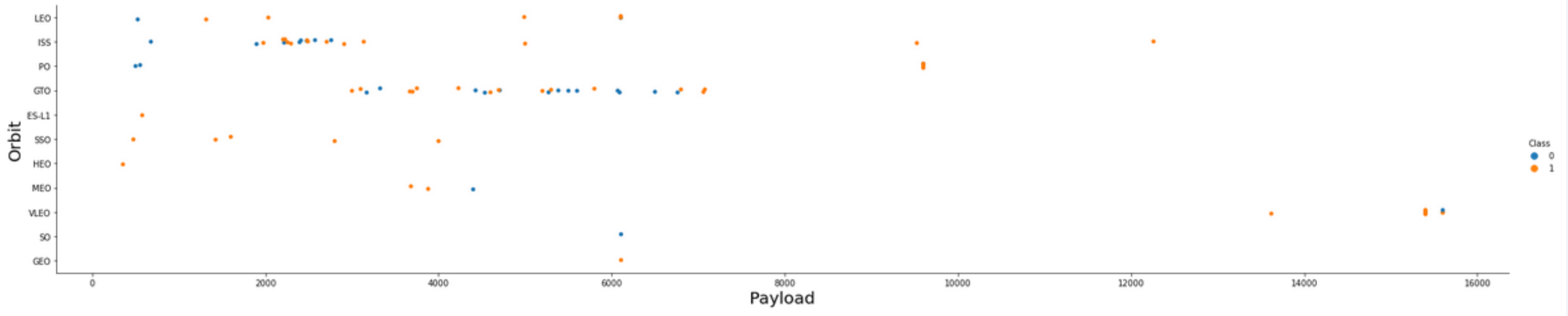


Flight Number vs. Orbit Type



- Successful launches are in yellow (1)
- Failed launches are in blue (0)
- Generally, as the number of launches increases, the success of those launches also increases with respect to orbit

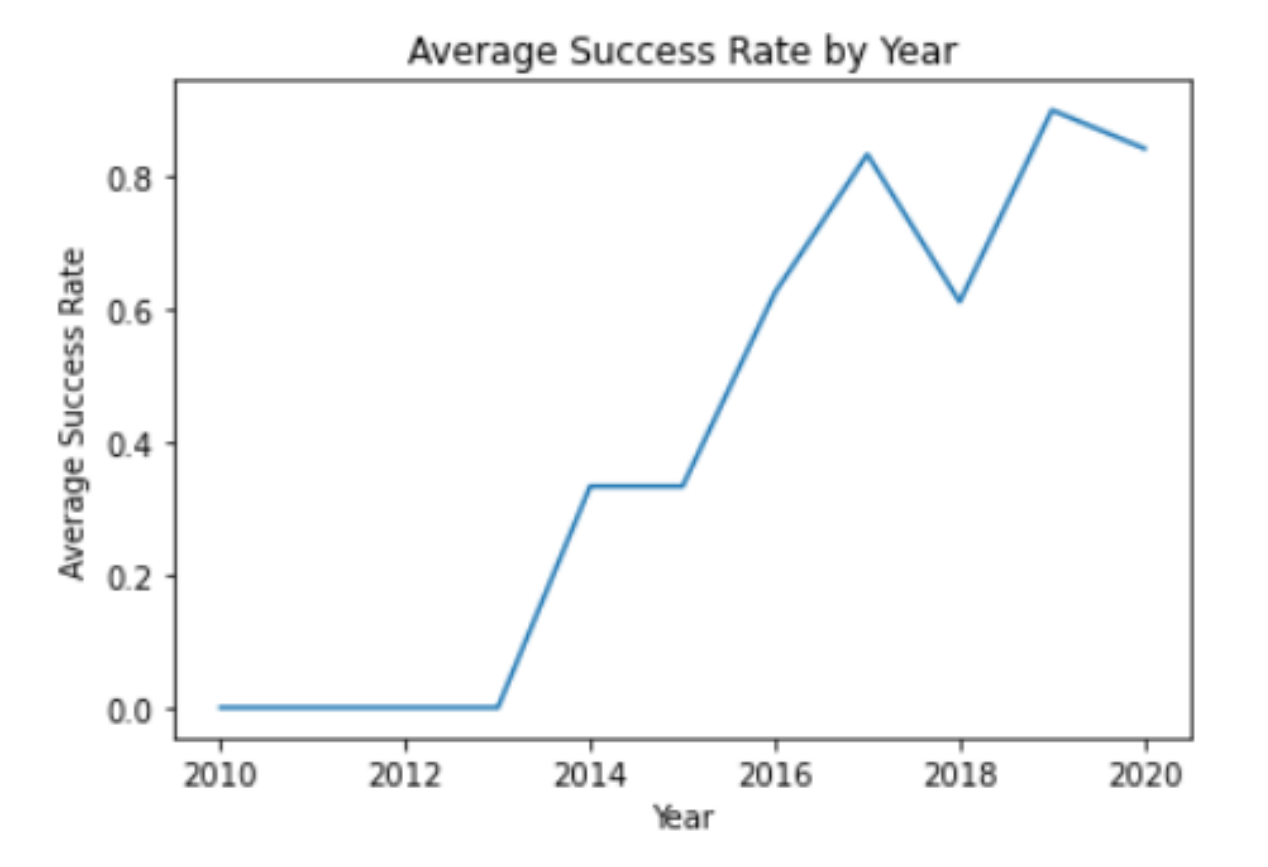
Payload vs. Orbit Type



- Successful launches are in yellow (1)
- Failed launches are in blue (0)
- Generally, there does not appear to be any correlation between launch success and payload mass
- In terms of orbit, LEO, ISS, PO appear to have more success at higher payloads
- Interestingly, SSO was successful for all payloads

Launch Success Yearly Trend

- With the exception of a dip in 2018, launch success has increased over time



All Launch Site Names

Display the names of the unique launch sites in the space mission

In [7]:
```
%sql select distinct(launch_site) from spacexdataset;
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[7]:

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

- There were four launch sites found

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [8]:
```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[8]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Five records that begin with 'CCA'

Total Payload Mass

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

In [9]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[9]:

total_payload_mass
45596

- The total is 45,596

Average Payload Mass by F9 v1.1

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

In [10]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[10]:

average_payload_mass
2928

- The average payload mass is 2,928

First Successful Ground Landing Date

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [12]:
```
%sql select DATE, landing__outcome from SPACEXDATASET where landing__outcome = 'Success (ground pad)' \
        order by DATE limit 1
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[12]:

DATE	landing__outcome
2015-12-22	Success (ground pad)

- The date was December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [13]:
```
%sql select booster_version, payload_mass__kg_ from SPACEXDATASET where landing__outcome = 'Success (drone ship)' \
        and (payload_mass__kg_ between 4000 and 6000)
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[13]:

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- There were four boosters found

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [14]:
```
%sql select mission_outcome, count(*) as total from SPACEXDATASET group by mission_outcome
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[14]:

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- There were 100 successful outcomes and 1 failed outcome

Boosters Carried Maximum Payload

- There were twelve booster versions that carried the max payload of 15,600 kg

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [16]:
```
%sql select max(payload_mass__kg_) from SPACEXDATASET;
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[16]:

1
15600

In [15]:
```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) \
        from SPACEXDATASET)
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[15]:

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

In [18]:
```
%sql select landing__outcome, booster_version, launch_site, date from SPACEXDATASET where year(DATE) = 2015 \
        and landing__outcome = 'Failure (drone ship)'
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[18]:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- There were two drone ship failed landings in 2015

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

In [19]:
```
%sql select landing__outcome, count(*) as total from SPACEXDATASET where DATE between '2010-06-04' \
        and '2017-03-20' group by landing__outcome order by count(*) desc
```

```
 * ibm_db_sa://lps89320:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875
/bludb
Done.
```

Out[19]:

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

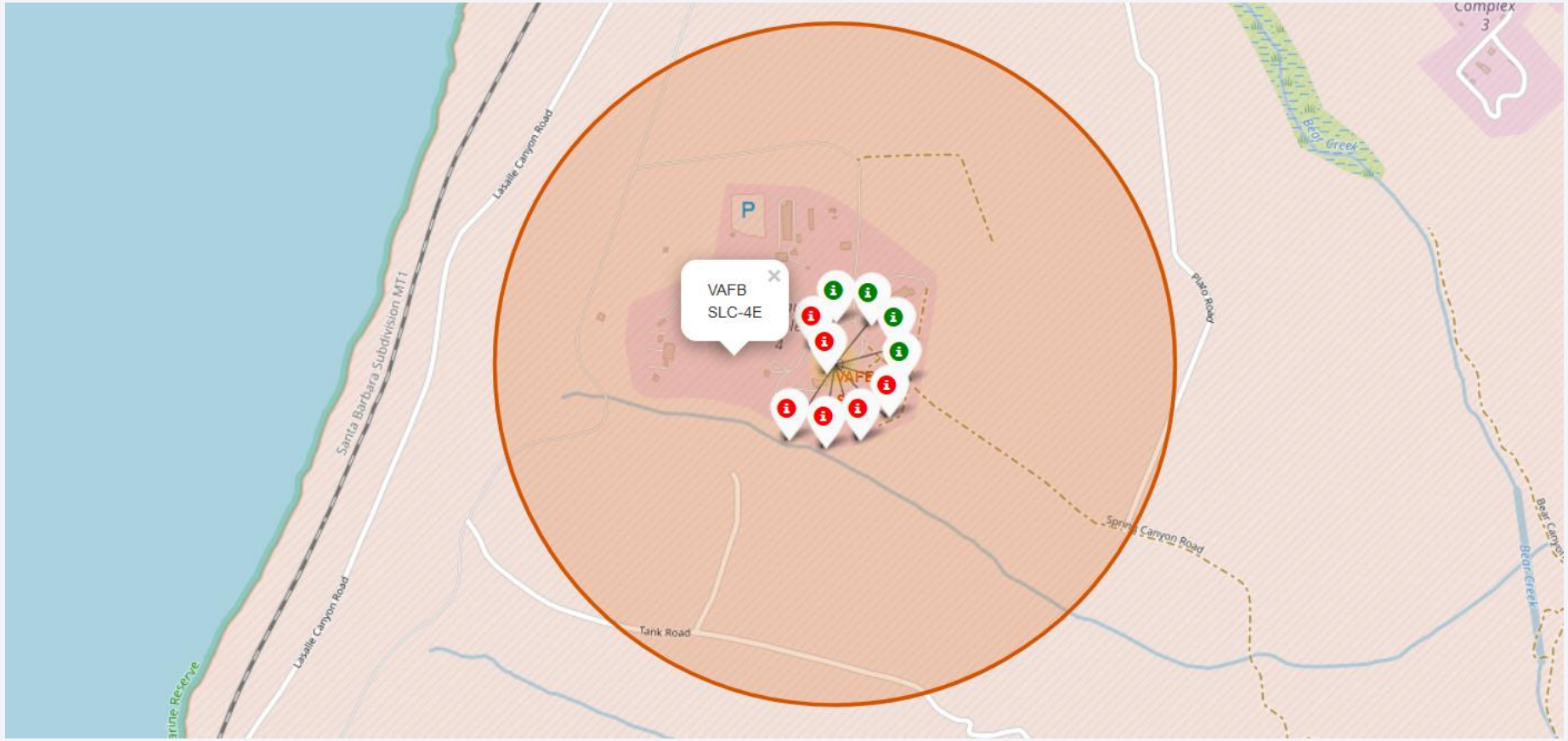
- There were eight different landing outcomes


Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Sites

- All four launch sites are located in the continental United States
- California site VAFB SLC-4E
- Florida sites KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40

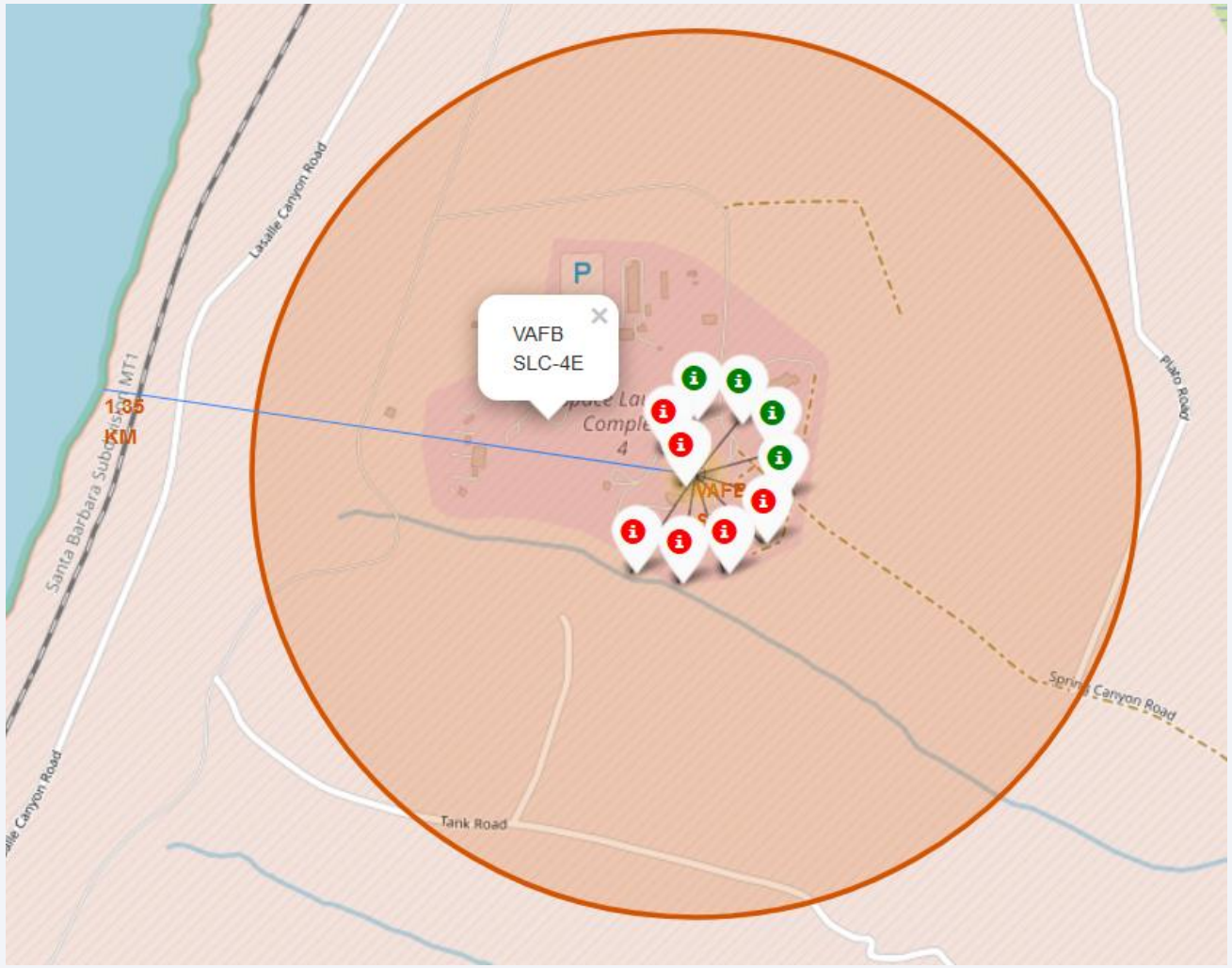
Launch Outcomes for site VAFB SLC-4E

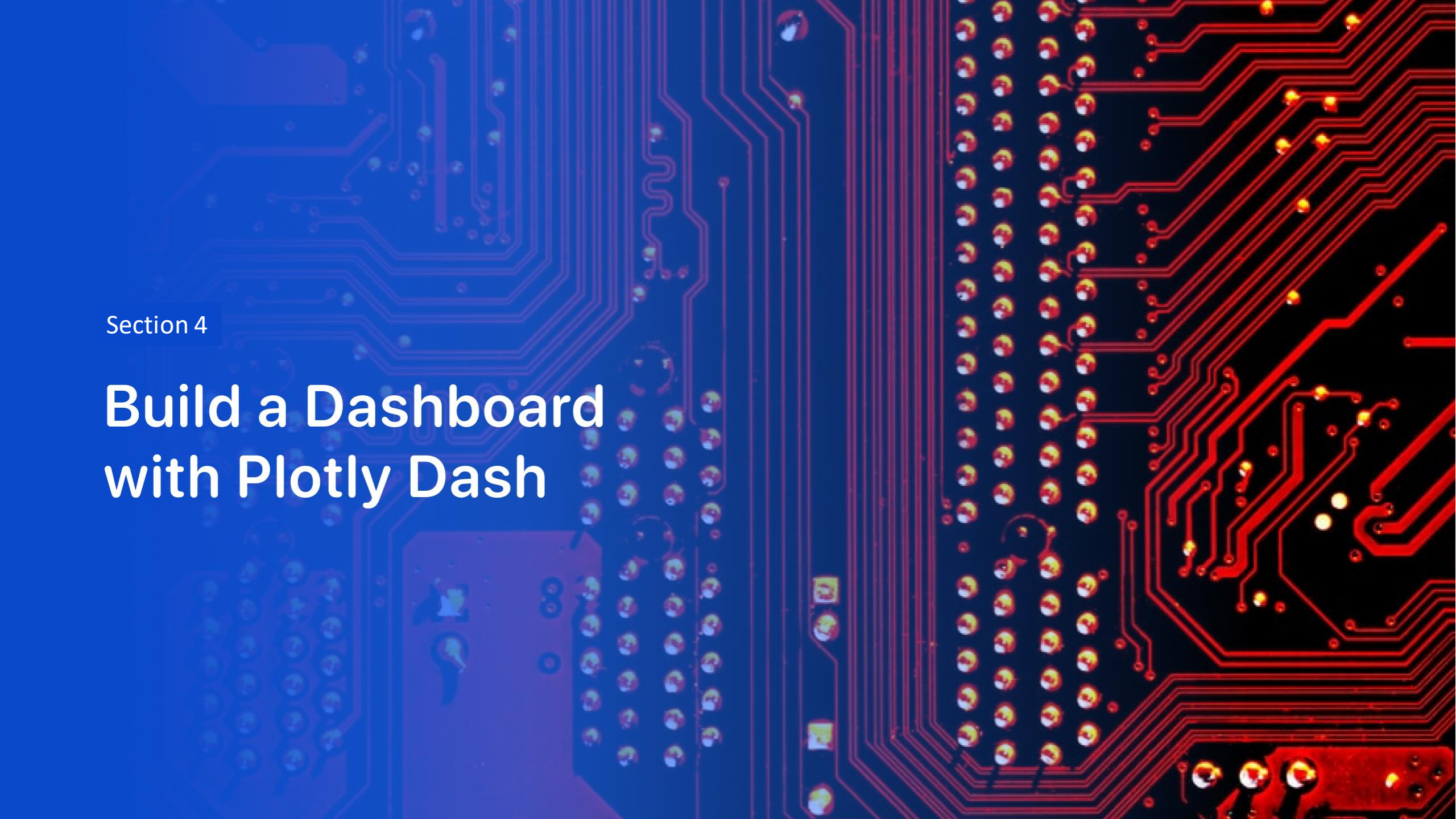


- Launch outcomes are color coded for quick-viewing:
  - Green (successful)
  - Red (failed)

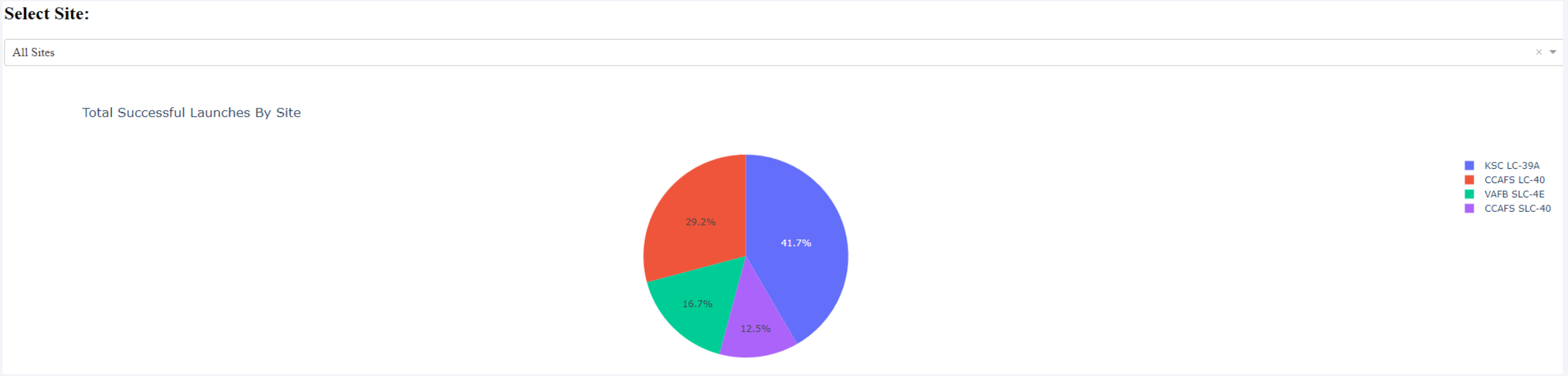
Launch Site to Coastline Distance

- VAFB SLC-4E is approximately 1.35 KM from California's coastline



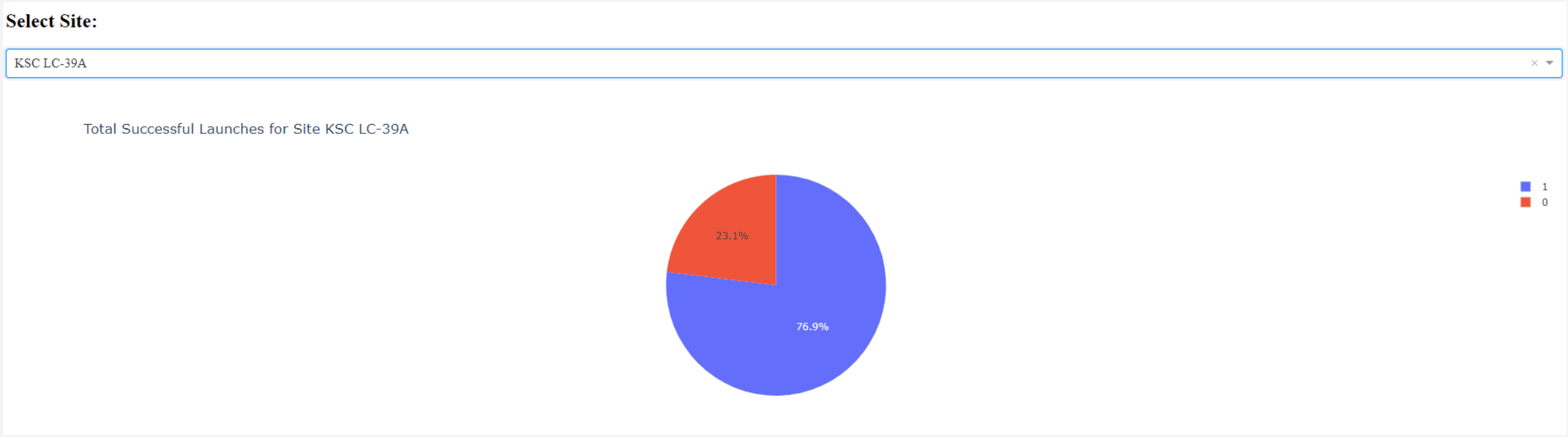

Section 4
Build a Dashboard with Plotly Dash

Launch Success Pie Chart



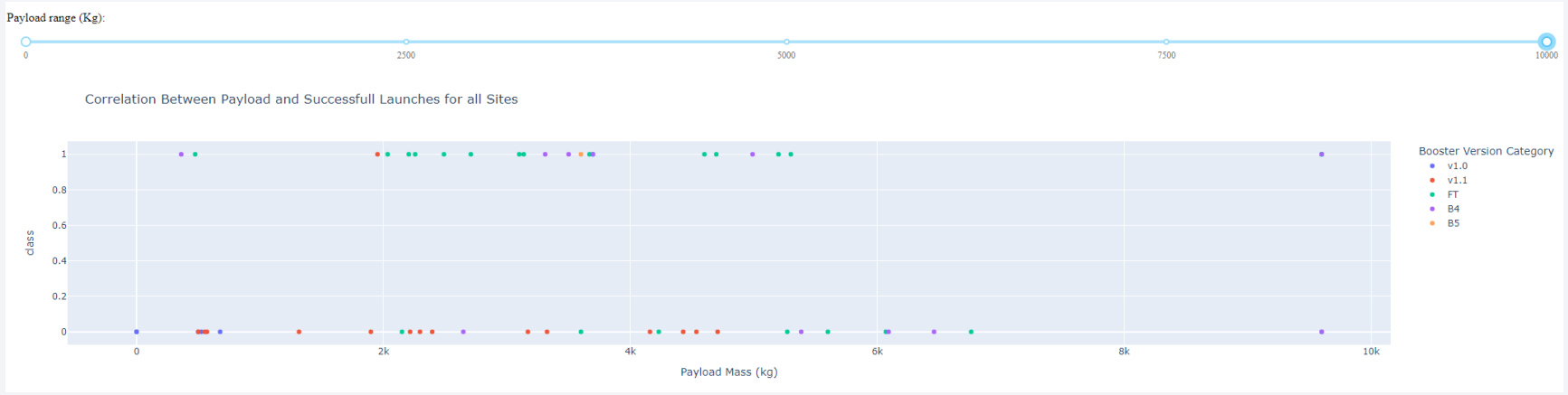
- Dropdown enables all sites or each site individually

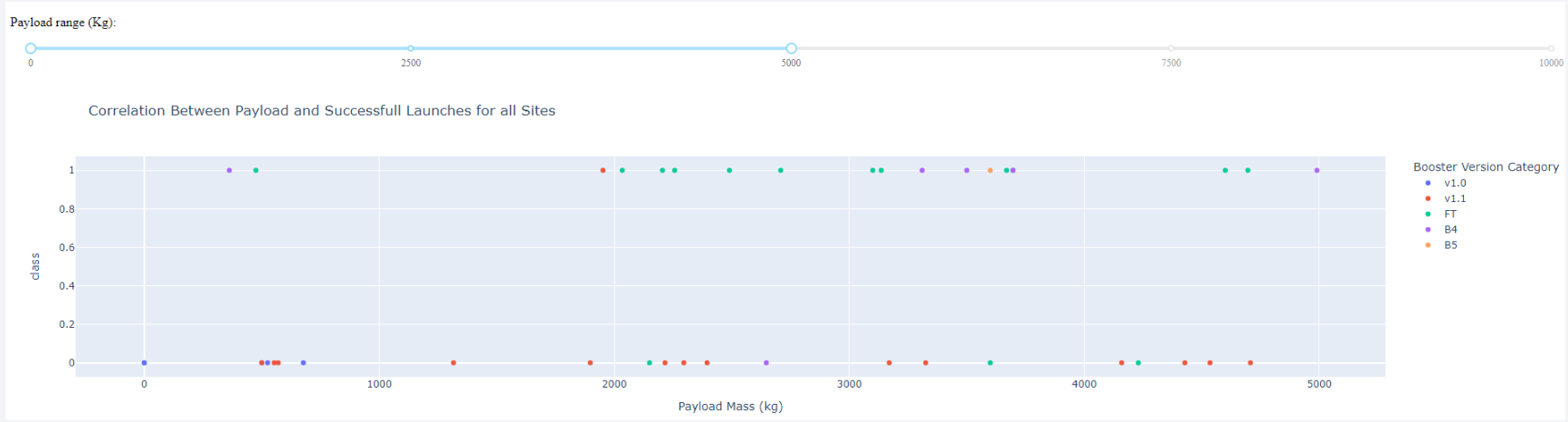
Site KSC LC-39A Launch Success Pie Chart



- Site KSC LC-39A had the highest success rate

Payload vs. Launch Outcome Scatter Plot





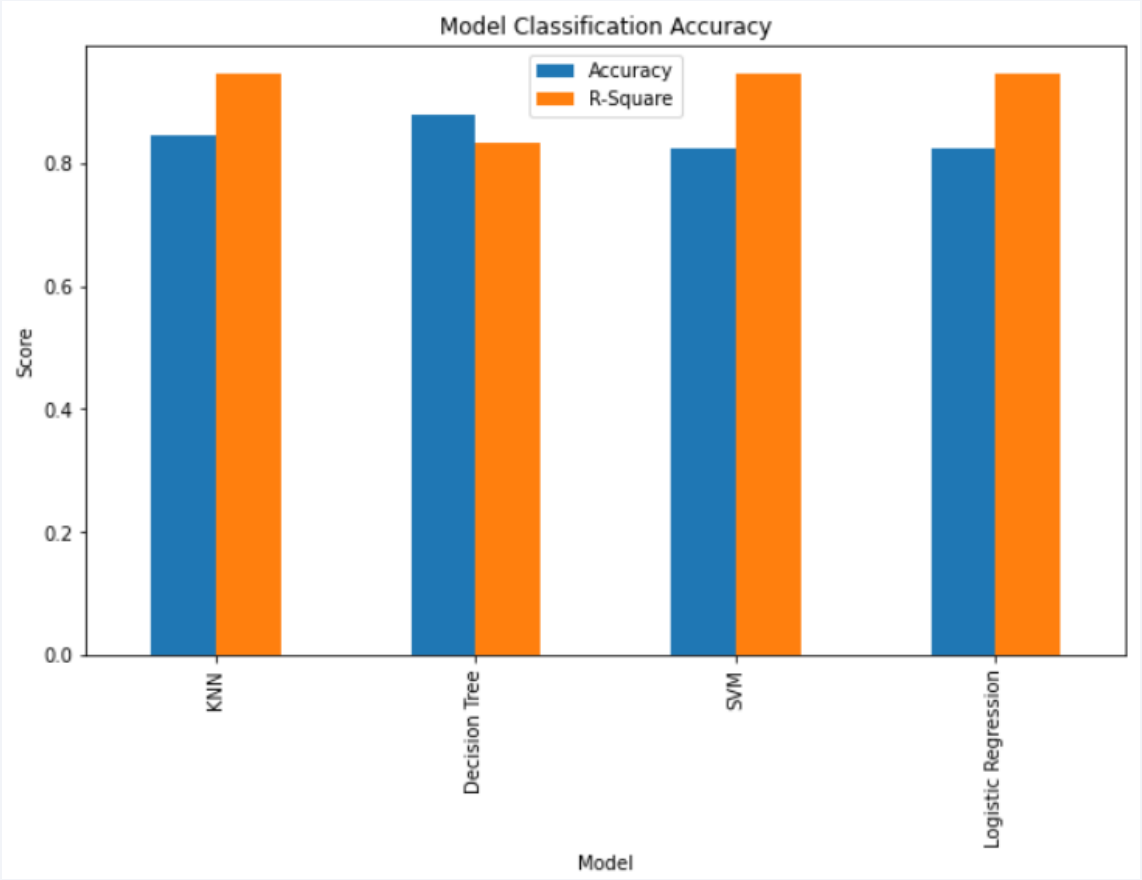
- Booster versions are color-coded
- Class 0 (unsuccessful) and Class 1 (successful)

Section 5

Predictive Analysis (Classification)

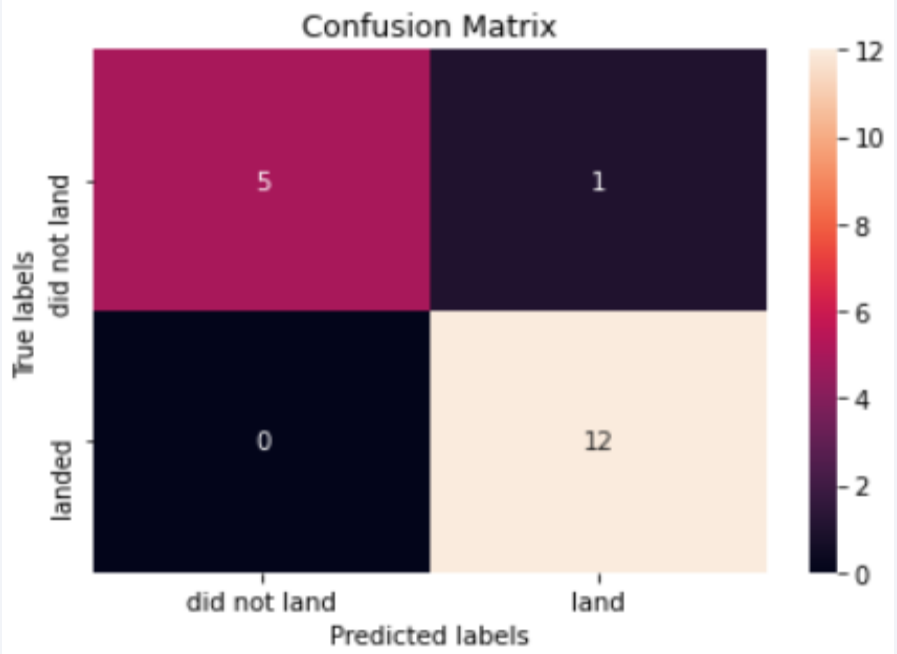
Classification Accuracy

- The k-nearest Neighbor classification model had the highest score (R-Square) and subsequent highest model accuracy.



Confusion Matrix

- Looking at the confusion matrix, the model only predicted one outcome incorrectly



Conclusions

- The goal of this analysis is to determine whether SpaceX Falcon 9 first stage booster rockets successfully return to Earth's surface, allowing for re-use.
- Twelve variables were ultimately selected as features (Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count and Serial).
- The modeling phase utilized four different classification algorithms (K-nearest Neighbors, Decision Tree, Support Vector Machine, and Logistic Regression). The data was standardized and split into training and testing subsets prior to model fitting and subsequent testing. Multiple parameter values were run (via Grid Search) in order to obtain the most accurate representation of the ground truth.
- All models performed well and the K-nearest Neighbors algorithm was ultimately selected as the top choice (Accuracy of 0.8444 and R-square of 0.9444).
- The analysis was considered a success in that an acceptable model was built, tested, and deemed ready for deployment for the purpose of predicting future launch outcomes.

Appendix

- **Datasets for this project were obtained from the following links:**
- https://api.spacexdata.com/v4/launches/past
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Thank you!