Winning Space Race with Data Science

D. Livnat

Aug 31, 2022

Outline

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

Executive Summary

- Summary of methodologies
- Data Collection (SpaceX Rest-API, Wikipedia Web Scraping)
- Data Wrangling
- Exploratory Data Analysis with Data Visualizations + Results
- Exploratory Data Analysis with SQL + Results
- Interactive Folium Maps + Results
- Interactive Dashboards with Ploytly Dash + Results
- Predictive Analysis (Classification w/ Logistic Regression, SVM, Decision Trees & KNN)
- Classification Results (Decision Trees ranks #1, high variability compared to the other)
- Conclusions

Introduction

The commercial space age is here: companies are making space travel affordable for everyone.
Perhaps the most successful is SpaceX. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land- we can determine the cost of a launch. In this capstone project, we trained a machine learning model and used public information to predict if SpaceX will reuse the first stage.

Questions to be answered:

- How do different variables (e.g., Payload Mass, Launch Site, # of Flights) affect a first-stage landing's success?
- How does landing's success rate change over time?
- What is the most accurate Machine Learning method to predict whether a first-stage will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX Rest-API & SpaceX Wikipedia Web Scraping
- Performed data wrangling
  - Data filtering, filling missing values and one-hot encoding were applied
  - Successful and Unsuccessful landing results were classified respectively
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree & KNN models were trained & tested
  - Decision Tree model provided the best predictive accuracy

Data Collection

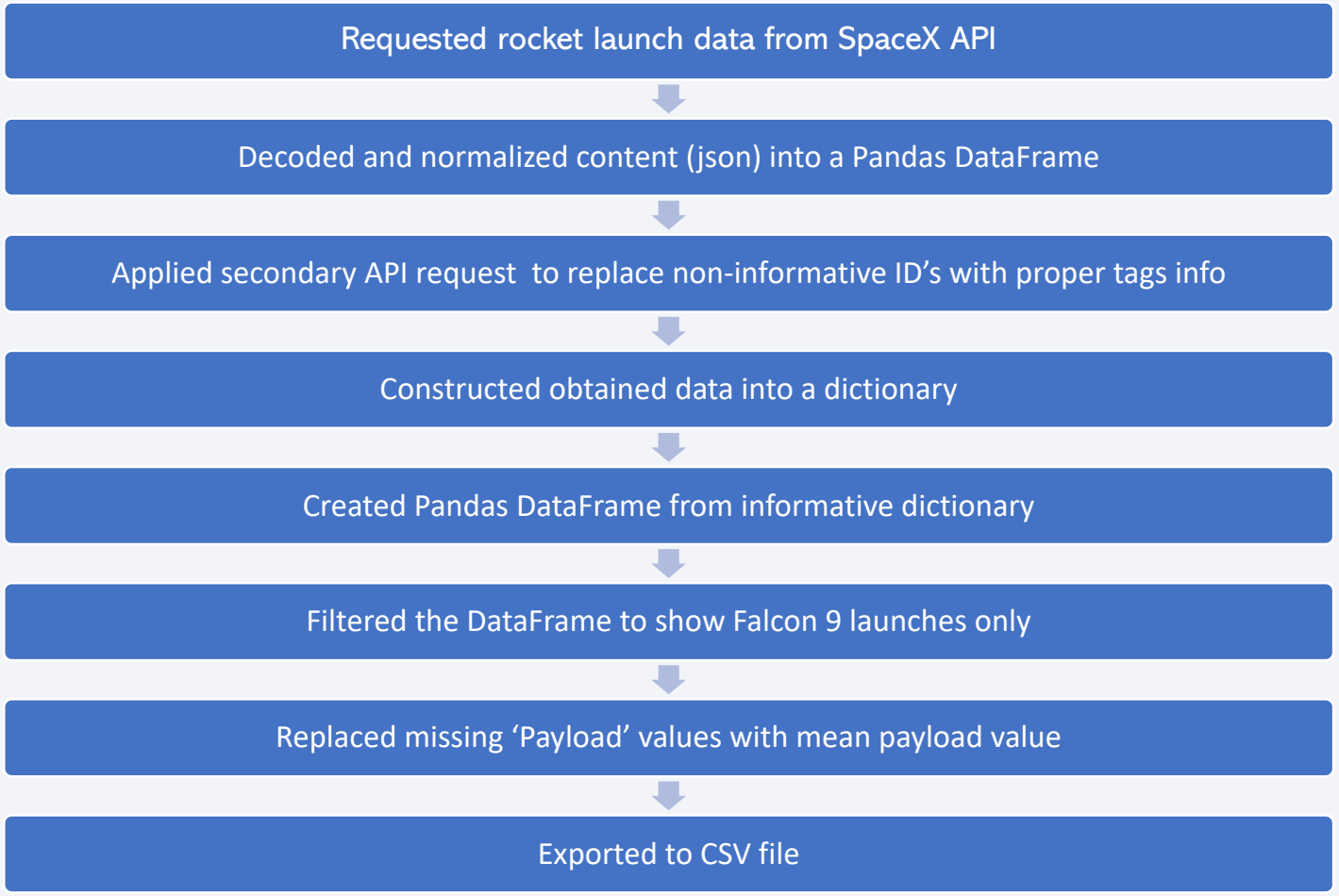
- Data Collection process included combination of SpaceX REST API requests and Web scraping from SpaceX's Wikipedia page with BeautifulSoup
- Data Columns obtained from SpaceX's REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome , Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Data Columns obtained from SpaceX's Wikipedia page:

Flight No., Date and time, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome

Data Collection – SpaceX API



Data Collection - Scraping

Requested the Falcon9 Launch Wiki page from its URL

Created a BeautifulSoup object from HTML response

Extracted all column/variable names from the HTML table header

Created informative dictionary by parsing the launch HTML tables

Created a Pandas DataFrame from dictionary

Exported to CSV file

Data Wrangling

Identified which columns are numerical and categorical, examined null values at each category

Calculated the number of launches on each site

Calculated the number and occurrence of each orbit

Created a landing outcome label (i.e., 'class') from Outcome column for Model Training purposes

Exported to CSV file

EDA with Data Visualization

- Visualized the relationship between Flight Number and Launch Site
- Visualized the relationship between Payload and Launch Site
- Visualized the relationship between success rate of each orbit type
- Visualized the relationship between FlightNumber and Orbit type
- Visualized the relationship between Payload and Orbit type
- Visualized the launch success yearly trend
- Applied Feature Engineering for further examination

EDA with SQL

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first succesful landing outcome in ground pad was acheived.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass
- Listed the records which display the month names, failure landing_outcomes in drone ship,booster versions, launch_site for the months in year 2015.
- Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Built an Interactive Map with Folium

- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map
- Calculated the distances between a launch site to its proximities

All provide highly informative data in a geographically visual way, which makes a lot of sense when evaluating launch sites, success rates and additional consequences

Built a Dashboard with Plotly Dash

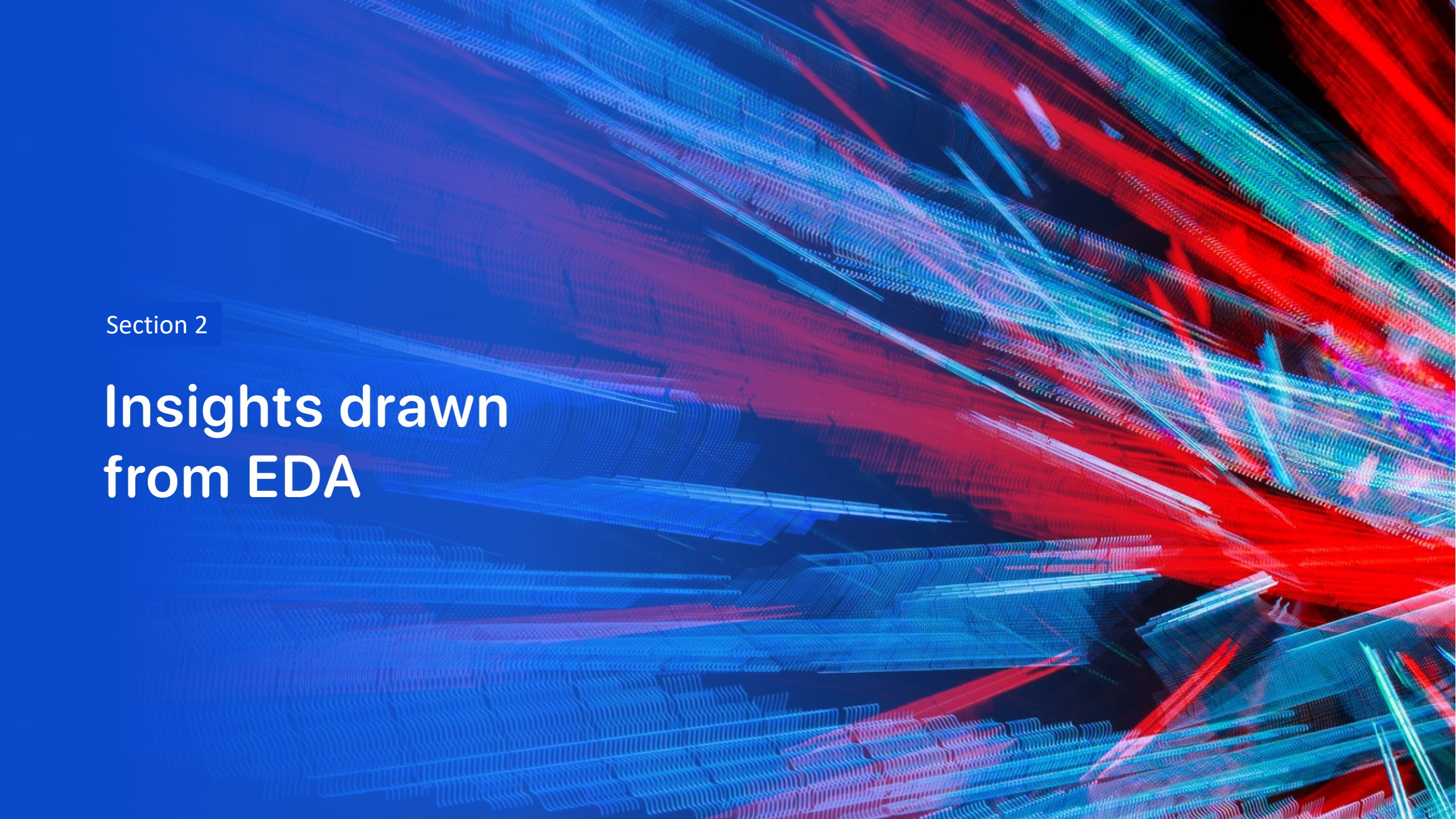
- Dashboard includes:
  - Dropdown list to enable Launch Site selection (including All Sites option)
  - Pie Charts showing success rates and count of successful/unsuccessful landings at each site
  - Payload Range slider to examine differences among Payload Mass ranges
  - Scatter Plot of success rates examining relationships between Payload Mass and Success
  - Booster Version Filtering for Scatter Plot

Predictive Analysis (Classification)

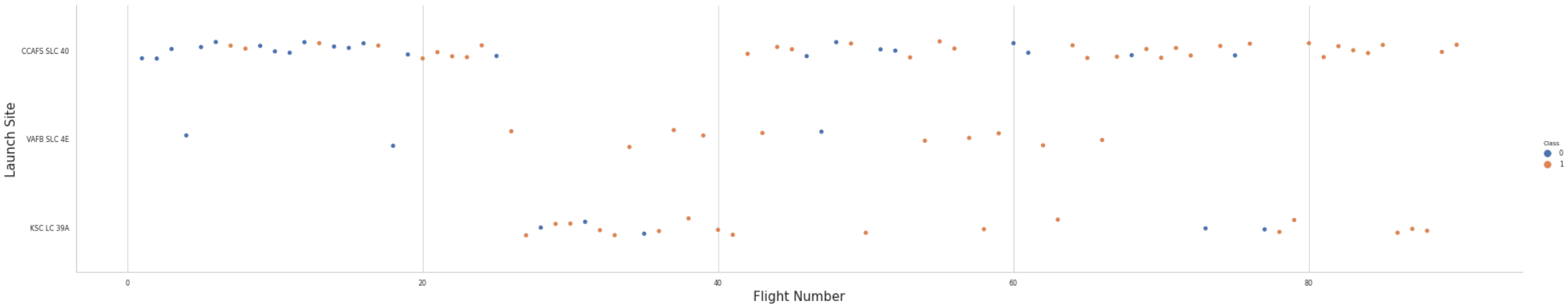
- Created a NumPy Array from ‘class’ value and assigned it to be the dependent variable Y
- Standardized the data of independent variables using StandardScaler
- Split the data into training and testing sets, using 20% for testing (selected randomly)
- Created and used GridSearchCV object to find the best parameters for every model
- Trained & evaluated each Model (Logistic Regression, SVM, Decision Tree, KNN)
- Found the most accurate model by Test Score and Best Score over Training Data

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

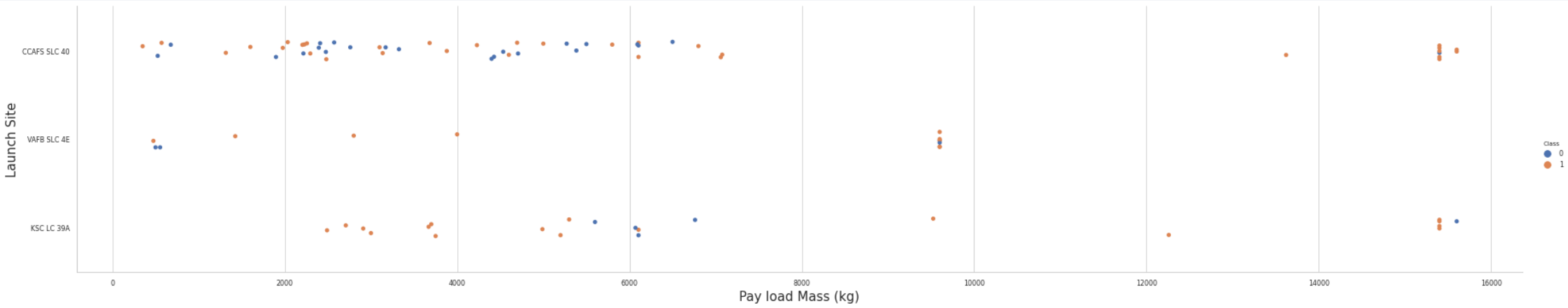

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



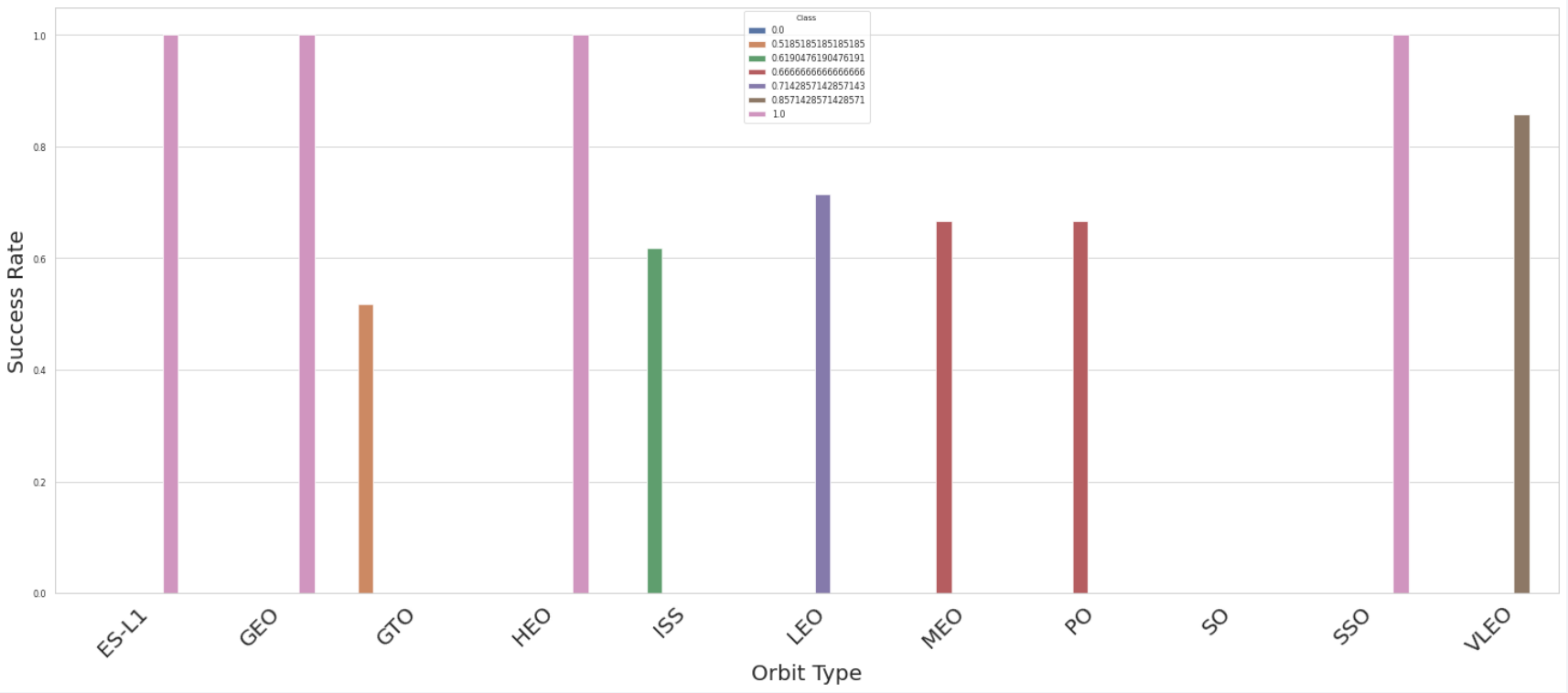
- Vast majority of earlier flights (23 out of first 25) were launched from CCAFS SLC-40
- The launch site is reused for later (success) launches as well; different sites in-between
- Success rates improve as the number of flight increases
- KSC LC 39A & VAFB SLC 4E seem to have higher success rates overall

Payload vs. Launch Site



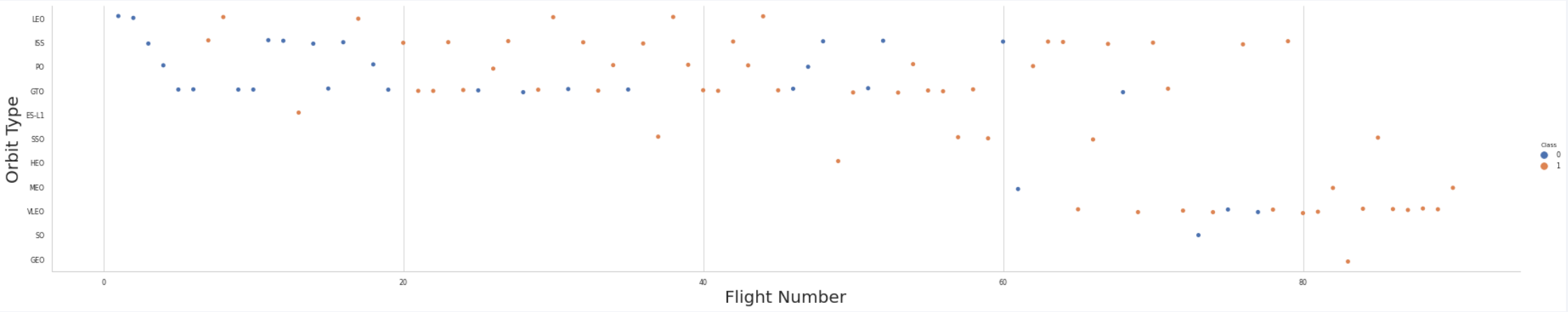
- Heavier payloads ended up in higher success rates at every site
- Highest success rates were around 6500KG+, regardless of Launch Site
- 100% of KSC LC-39A launches under 5500KG ended up in successful landing

Success Rate vs. Orbit Type



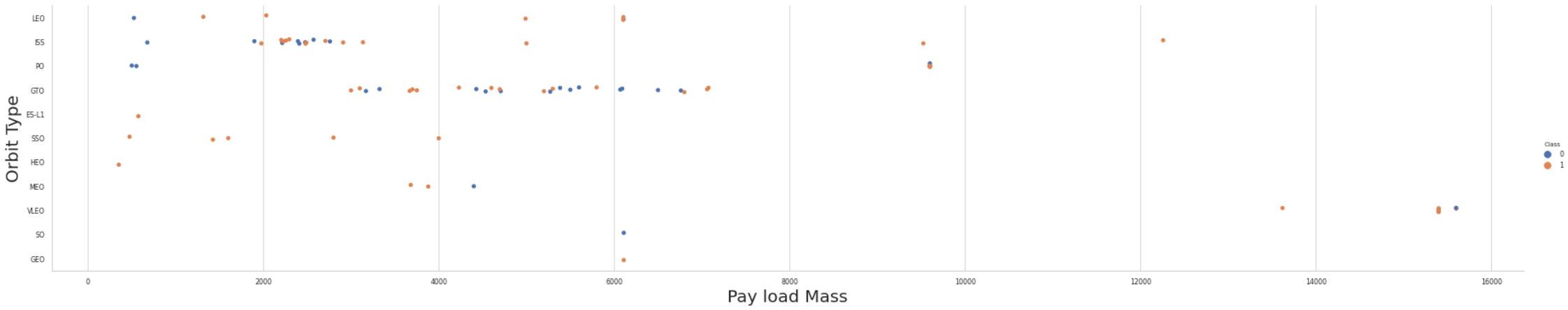
- 100% of flights in Orbits ES-L1, GEO, HEO, SSO successfully landed
- 0% of flights in Orbit SO successfully landed
- VLEO, LEO, MEO & PO, ISS, GTO success rates in ranges of 85%-50%

Flight Number vs. Orbit Type



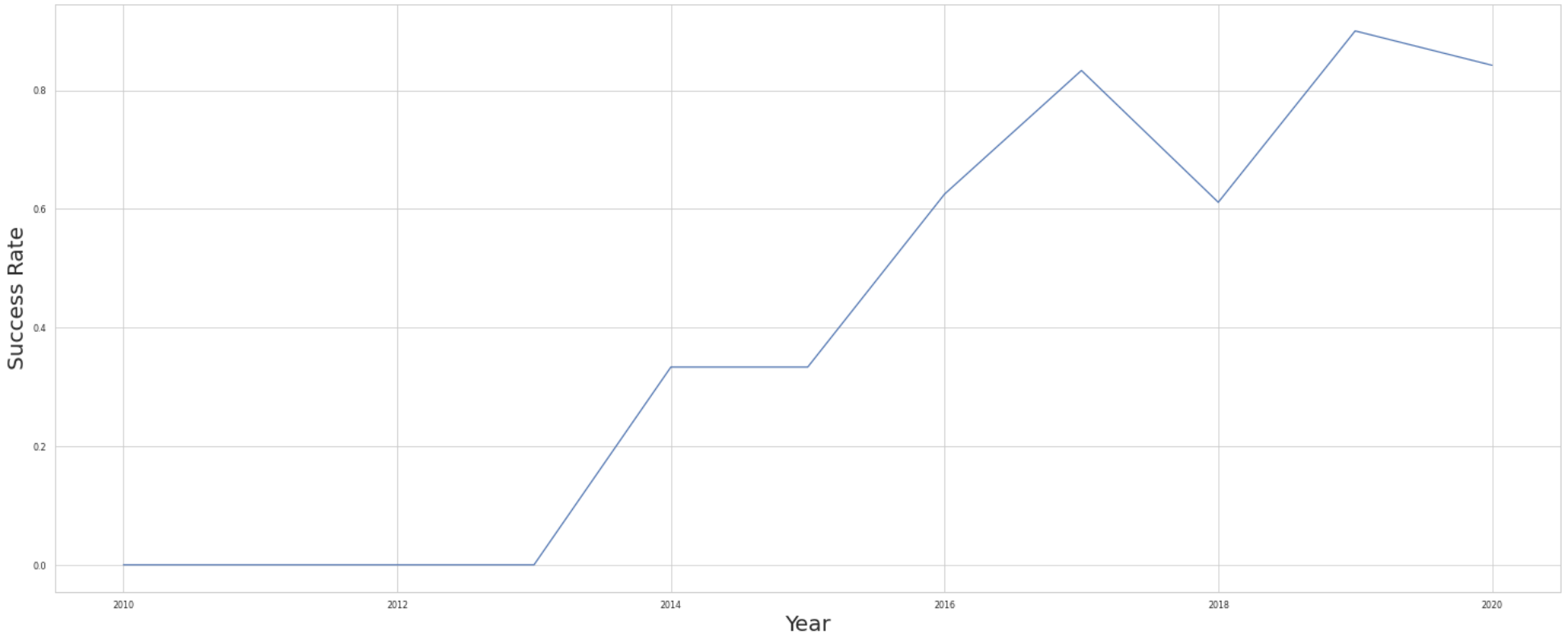
- The success rates in LEO orbit increase as the flight number is higher, in contrast to GTO where the success rates do not improve
- Orbits with 100% success were used in a very small sample size (one or two launches)
- Vast majority of latest flights were launched in the VLEO Orbit (85% success rate)

Payload vs. Orbit Type



- Polar, LEO and ISS Orbits has higher successful landing rates with heavier payloads
- HEO, SSO and MEO Orbits has high success rates with smaller payloads
- In the GTO Orbit we cannot distinguish between successful and failed launches using the Payload Mass metric

Launch Success Yearly Trend



- We can see an overall improvement over the years (from 0% to more than 80%)
- Success rates rising since 13', soaring between 15' and 17', small drop in 18'

All Launch Site Names

- SpaceX used 4 different launch sites
- CCAFS was used in two distinct launch site locations

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

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

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

- First 5 ‘CCA’ records were from ‘CCAFS LC-40’ Launch Site
- All 5 records has not landed

Total Payload Mass

NASA_CRS_Total_Payload_Mass_KG
45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS NASA_CRS_Total_Payload_Mass_KG FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

- NASA Boosters carried 45,596KG of Payload

Average Payload Mass by F9 v1.1

AVG_Payload_Mass
2534.6666666666665

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

- Average Payload Mass for the Falcon-9 v1.1 Booster is 2534.67KG

First Successful Ground Landing Date


```
%sql SELECT Date FROM SPACEXTBL WHERE `Landing _Outcome` = 'Success (ground pad)' ORDER BY Date DESC LIMIT 1
```

- First Successful landing on ground pad happened on December 22, 2015
- It took more than 5 years since beginning

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE `Landing _Outcome` = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

- Boosters which have success in drone ship landing and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Total_Successful_Missions	Total_Failure_Missions
100	1

```
%%sql
SELECT
    (SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%') AS Total_Successful_Missions,
    (SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%') AS Total_Failure_Missions
FROM SPACEXTBL
LIMIT 1
```

- Mission Outcomes were highly successful, with only one case considered failure

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
GROUP BY Booster_Version
HAVING SUM(PAYLOAD_MASS__KG_) = (SELECT SUM(PAYLOAD_MASS__KG_) AS Max_Mass
                                 FROM SPACEXTBL
                                 GROUP BY Booster_Version
                                 ORDER BY Max_Mass DESC
                                 LIMIT 1)
```

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

- List of Boosters that carried the Maximum Payload in KG's (All equal)

2015 Launch Records

Month	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

```
%%sql
SELECT substr(Date, 4, 2) as Month,
       `Landing _Outcome`,
       Booster_Version,
       Launch_Site
FROM SPACEXTBL
WHERE `Landing _Outcome` LIKE 'Failure (drone ship)' AND
       substr(Date,7,4)='2015'
```

- Two cases of failed drone ship landing in 2015 (January, April)

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

Landing _Outcome	Successful_Landings
Success (drone ship)	4
Success (ground pad)	2

```
%%sql
SELECT `Landing _Outcome`,
       COUNT (`Landing _Outcome`) AS Successful_Landings
FROM
       (SELECT `Landing _Outcome`,
               substr(Date, 1, 2) as Day,
               substr(Date, 4, 2) as Month,
               substr(Date,7,4) as Year
        FROM SPACEXTBL
        WHERE `Landing _Outcome` LIKE 'Success%' AND Date BETWEEN '04-06-2010' AND '20-03-2017'
        LIMIT 6)
GROUP BY `Landing _Outcome`
ORDER BY Successful_Landings DESC
```

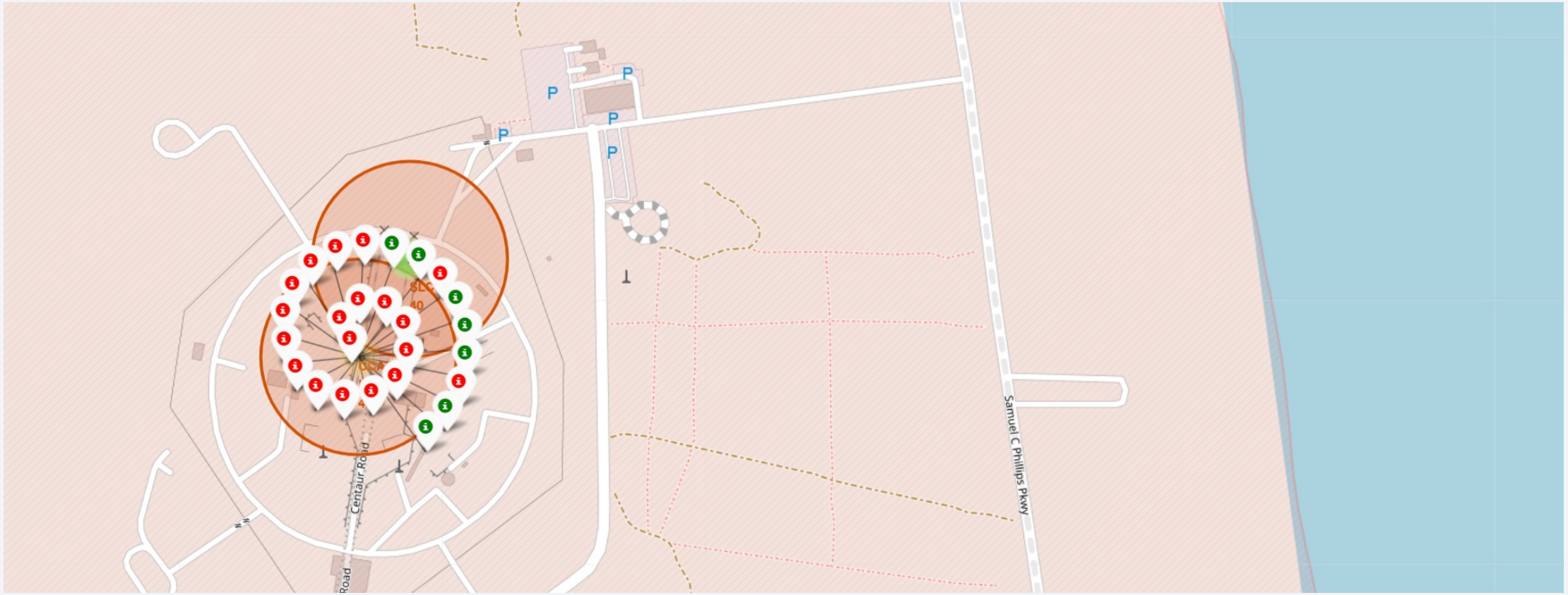
- 6 Successful landings between 2010-06-04 and 2017-03-20
- 4 on drone ship and 2 on ground pad
- SQLite doesn't support monthnames which is 👎


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations (World Map)

All launch sites are in very close proximity to the coastline for safety reasons. Launch Sites located in relative proximity to the equator line as well (to support spacecraft speed).

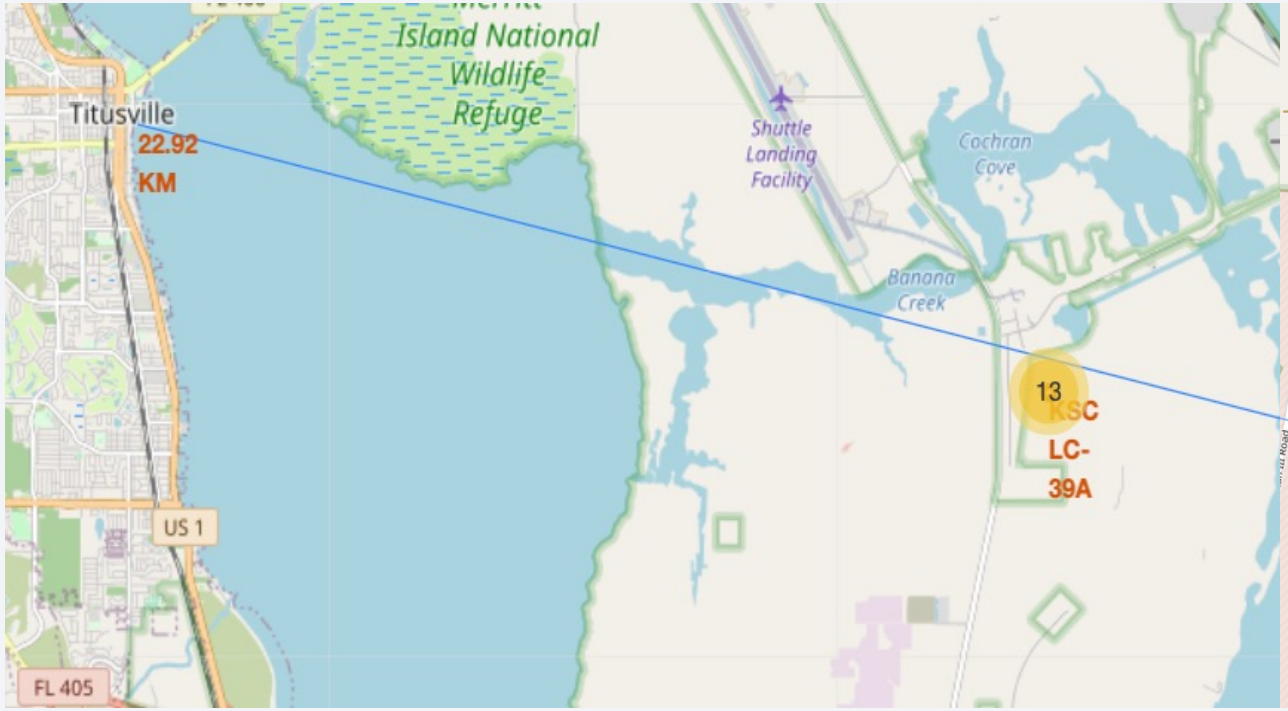
Color-Labeled Markers (Cape Canaveral)

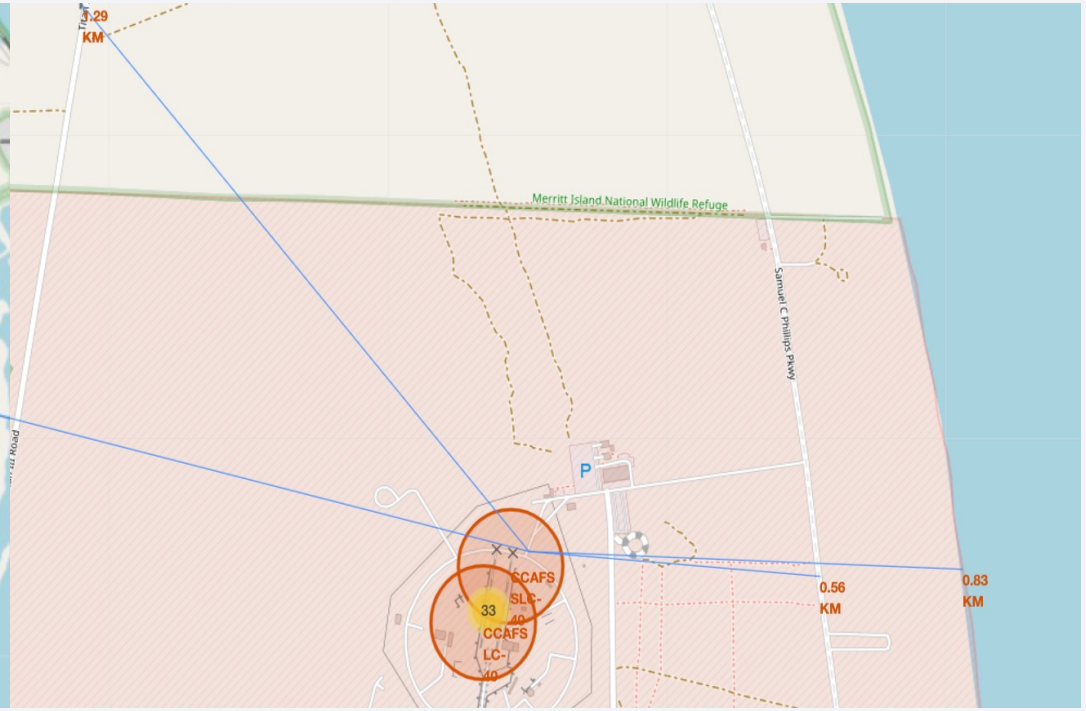


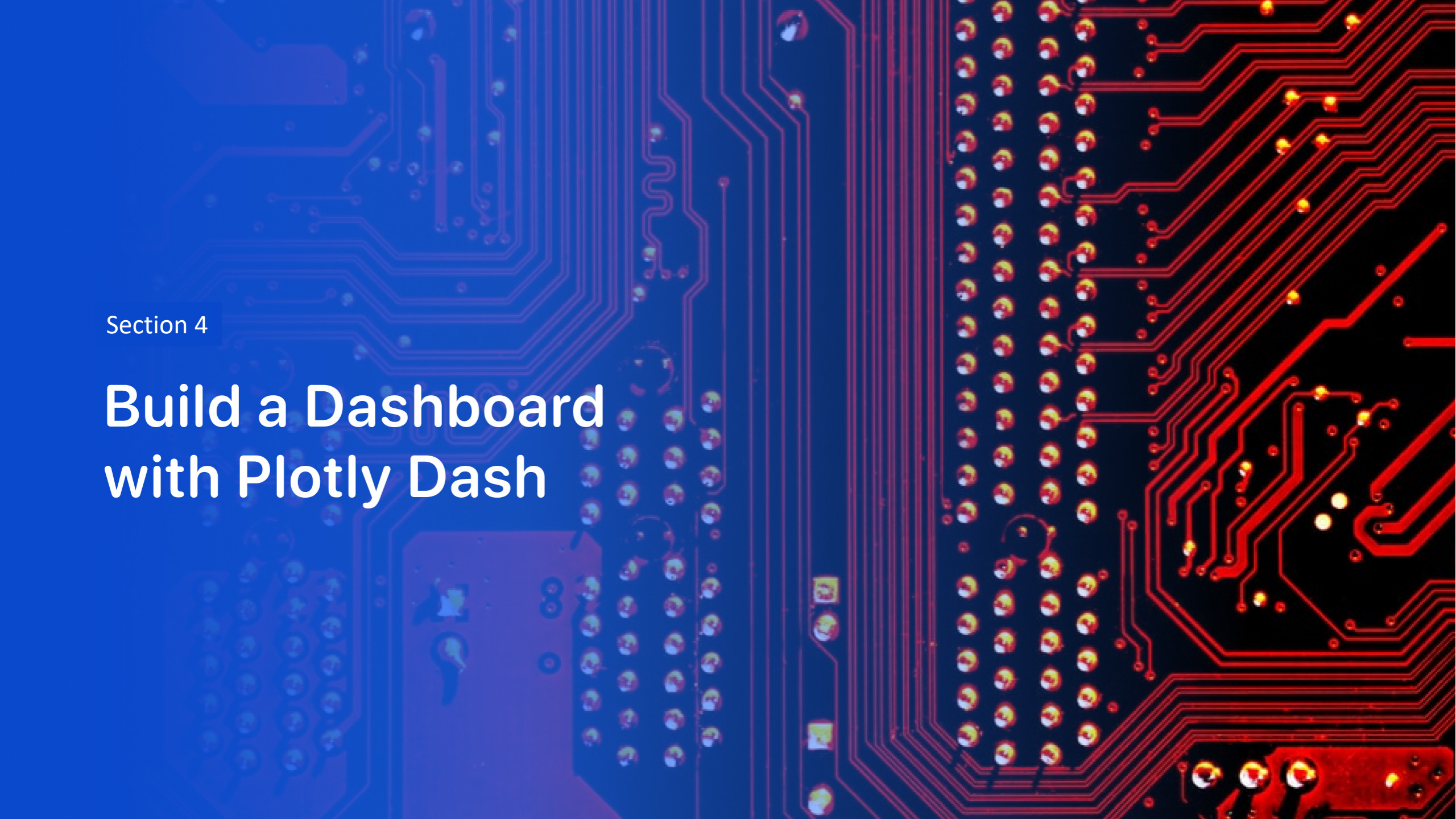
- We used color-labeled markers to represent the outcome of a launch: Green for successful launches and Red for failed ones. (CCAFS LC-40 w/ poor 26.9% success)

Distance from Populated Areas

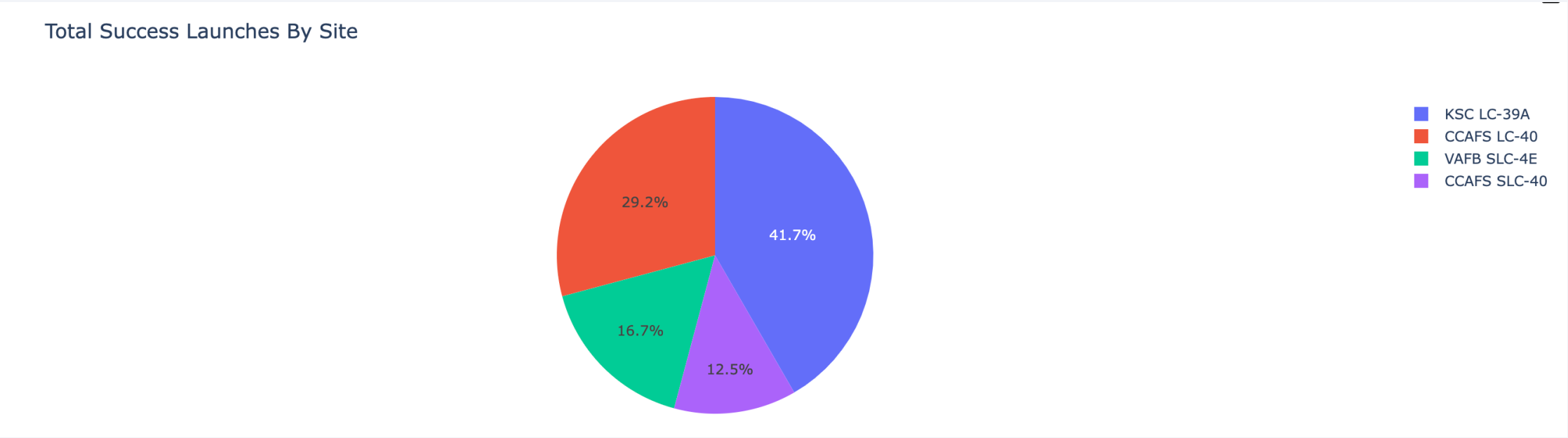
- Although isolated, launch sites are located close to railways, highways and coastlines, and relatively close to cities (Titusville is located 22.92KM from CCAFS & 16.32KM from KSC).
- These locations could be in danger in case of a failed launch which requires safety measures.





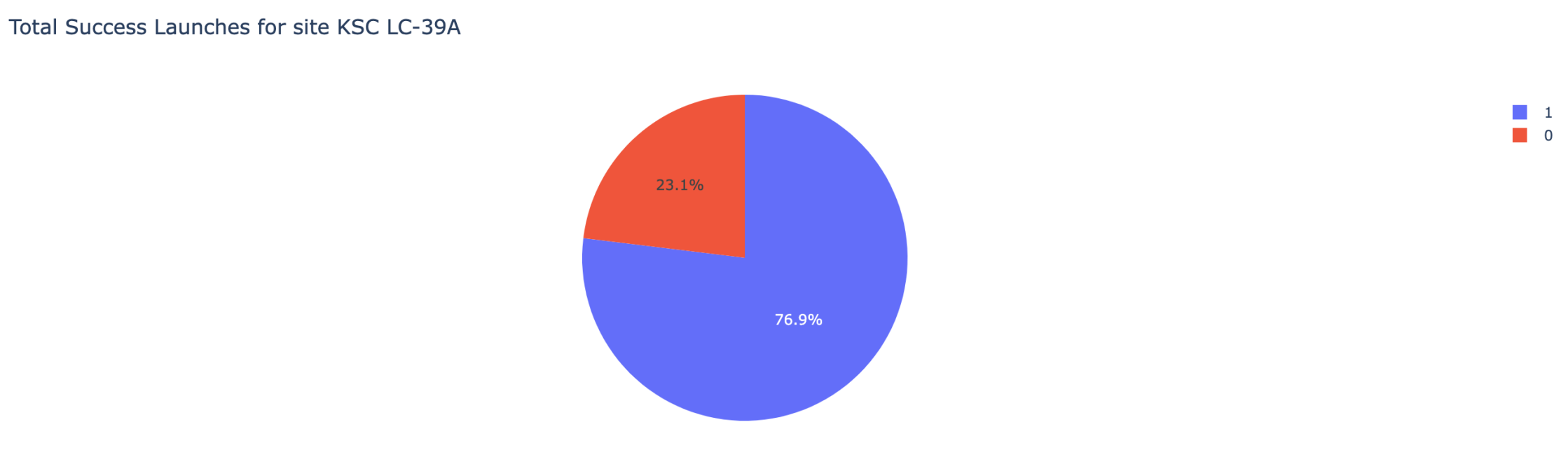

Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site



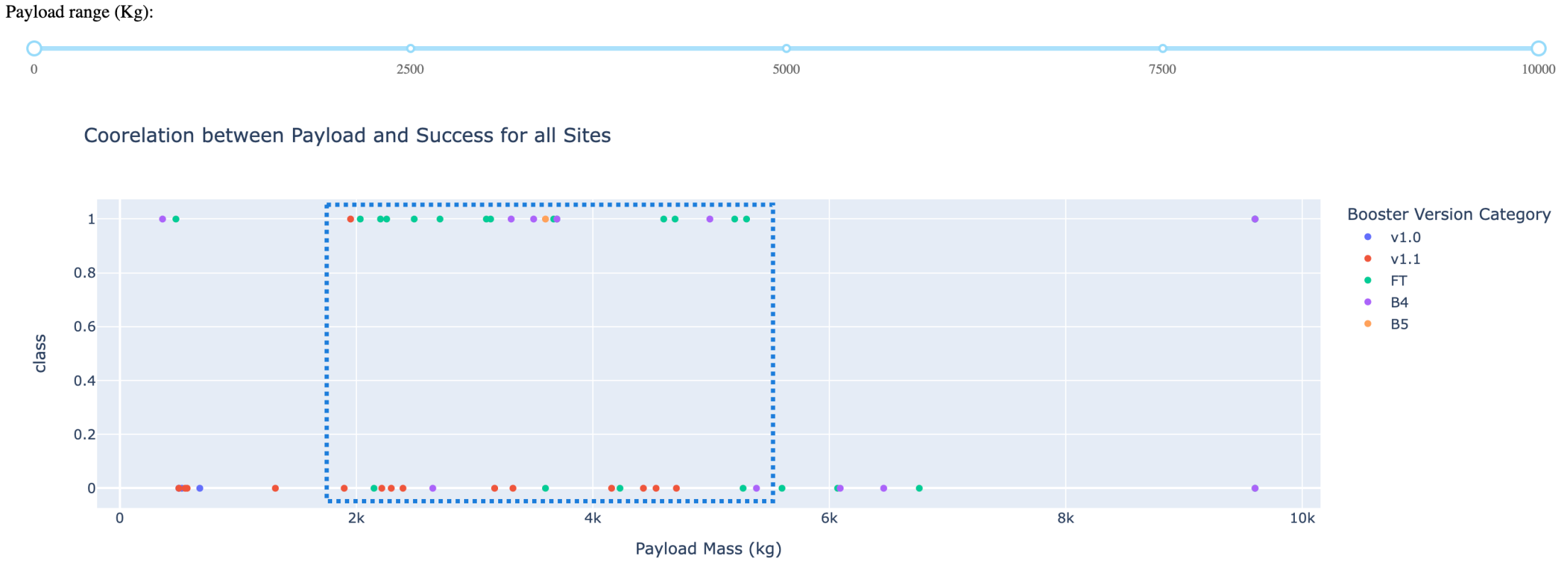
- Of all the launches tested in the sample, the launches from Kennedy Space Center Launch Complex 39A had the highest success rate.

Kennedy Space Center Launch Complex 39A



- 10 out of 13 launches from the Kennedy Space Center Launch Complex 39A landed successfully.

Payload Mass vs Success Rate



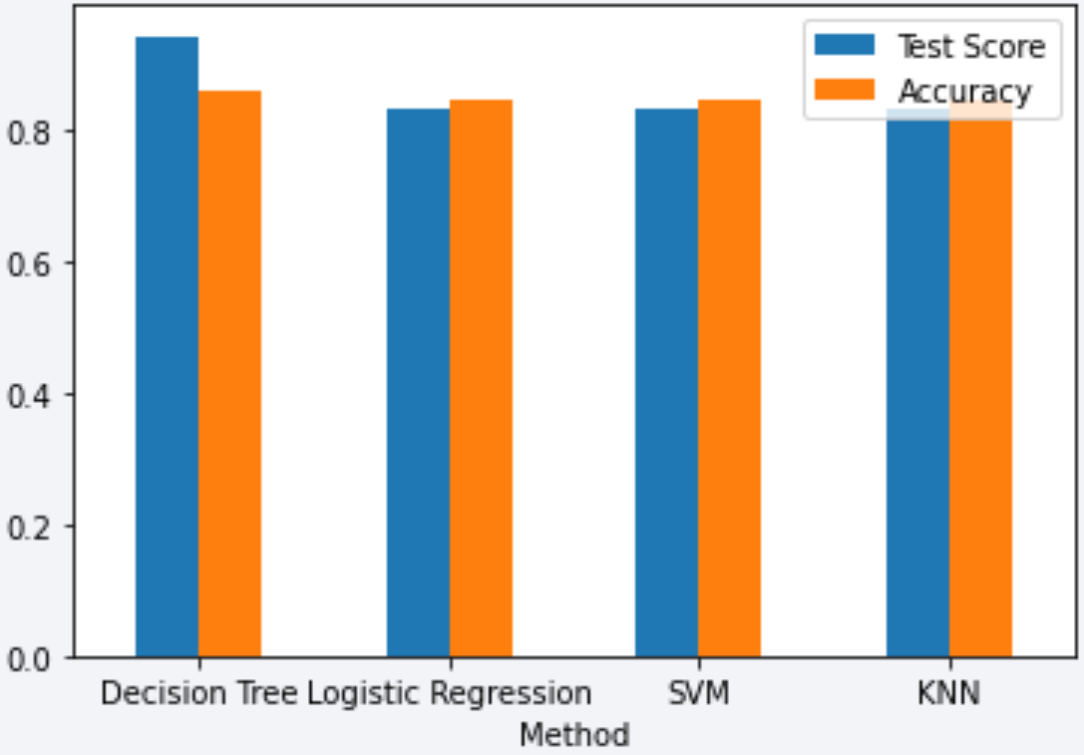
- Payloads in the range of 2K & 5K (kg) provided the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 4 classification models were trained on a randomly-split ‘Training’ dataset.
- Each training set resulted in an ‘Accuracy’ score, i.e- precision score in classifying accurately (0-1)
- Trained models were used to predict a new, different ‘Test’ dataset, resulted in ‘Test Score’.



Confusion Matrix: Decision Tree Classifier

- Examining the confusion matrix, we can see that our Decision Tree model can distinguish between the different classes.
- The model has successfully classified all failed landings in the test-set correctly, while it's overall accuracy score is 94.4% (single false-positive case in this confusion matrix, compared with 3 in alternative models)

Conclusions

- 72 train vs 18 test observations are a very small sample size, which may not be enough to reflect and predict results accurately
- Furthermore, Decision Tree Models vary in results among multiple tests, and while still manage to rank first among classifiers, are far from being dependable. Logistic Regression, SVM and KNN train & tests results were relatively equal with smaller variance – Tree Models might create Overfitting, others are less likely
- Kennedy Space Center Launch Complex 39A has the highest success rates among all launch sites (100% success under 5500KG of payload)
- Payloads in the range of 2K & 5K (kg) provided the highest success rate (B5 100%)
- Vast majority of latest flights were launched in the VLEO Orbit (w 85% success rate)
- It may take some time, but SpaceX's success rates soared after 5 years!

Appendix

- https://github.com/dorlivnat/SpaceX
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- https://www.coursera.org/professional-certificates/ibm-data-science

Acknowledgments

Mr. Allon Mask

Thank you!