IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gradon Kam
17 February 2022

Outline

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

Executive Summary

- This project uses a model to determine likelihood that SpaceX lands a Falcon 9 first-stage to decide whether or not to bid against SpaceX for rocket contracts.
- Data is collected from the SpaceX API as well as via web scraping.
- Exploratory data analysis shows that SpaceX landing performance has improved with time. Several other variables are analyzed and included in the model.
- A model was developed using a Support Vector Machine with tuned hyperparameters. This model is estimated to have 83.3% accuracy when predicting Falcon 9 first-stage landings. The errors on the test dataset were all false positive errors.
- The model error types imply that decisions based on this model will lead to cases where we may decline to bid on a mission that we could win, but would not lead to cases where we enter a bid that is likely to lose against SpaceX.

Introduction

- This project is being developed by a prospective space launch company that intends to compete against SpaceX for rocket launches.
- We use data from past launches to determine the likelihood that SpaceX will be able to land and reuse the first stage from one of its Falcon 9 rockets. This allows SpaceX to reduce the costs of a rocket launch.
- Our company would like to decide whether or not to bid against SpaceX for a rocket launch. The ability to reuse the first stage will drive the costs of SpaceX as our company's main competitor. A competitive bid from our company may be assembled based on the ability to predict if SpaceX can reuse the first stage.
- We would like to know the important variables that control the ability to land the first stage successfully and develop a model to determine the probability of a successful landing. This model will be used to decide if a bid should be made.

Section 1

Methodology

Methodology

Executive Summary

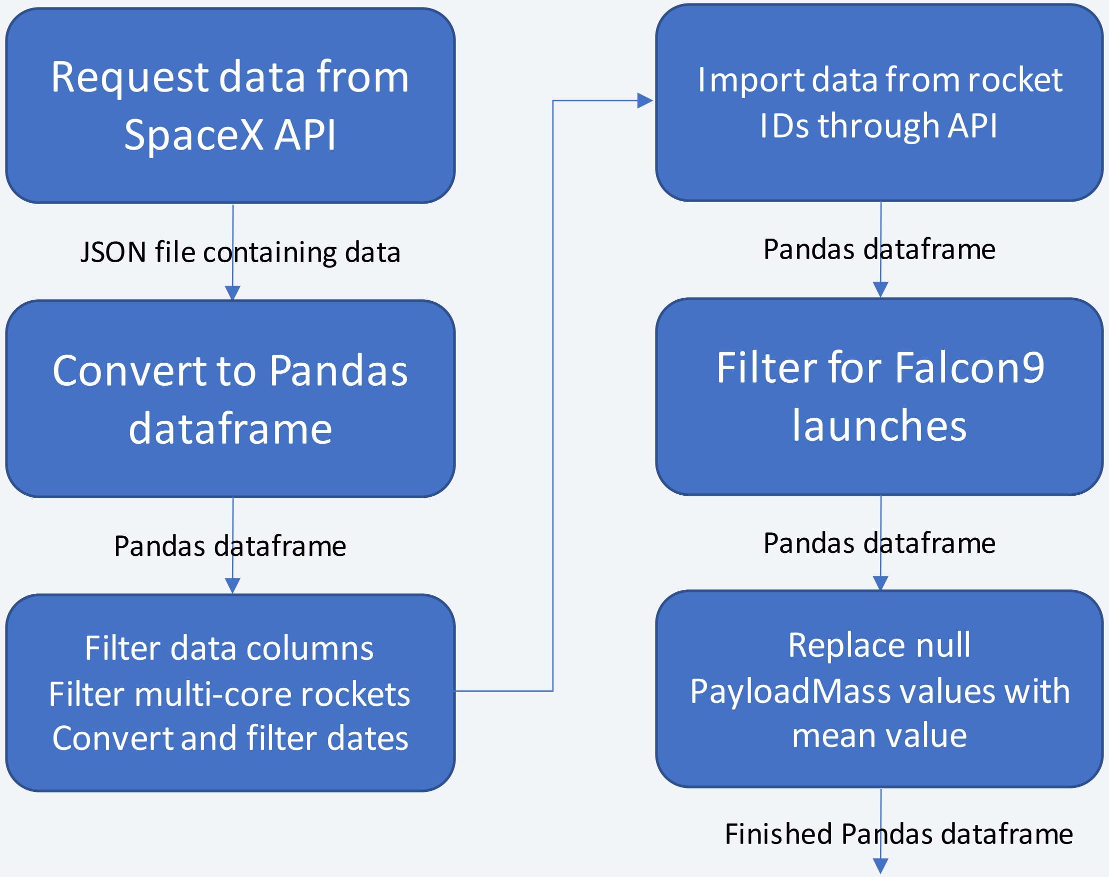
- Data collection methodology:
  - Data collected via SpaceX API and web scraping
- Perform data pre-processing
  - Data processed to filter variables - remove non-Falcon9 boosters as an example
  - Impute missing data where necessary (PayloadMass)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built and tuned LogReg, SVM, Decision Tree, KNN models and analyzed performance

Data Collection

- Data was collected through two methods – via the SpaceX API and web scraping.
- Data imported from the SpaceX API at https://api.spacexdata.com/v4/launches/past as a JSON file; the data was filtered down to include pertinent variables and rocket types, then cleaned and brought into a Pandas dataframe.
- Data imported from the site https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches through web scraping. The data from the launch table was parsed using BeautifulSoup and stored in a Pandas dataframe.
- The next two slides show via flowchart the processes described above.

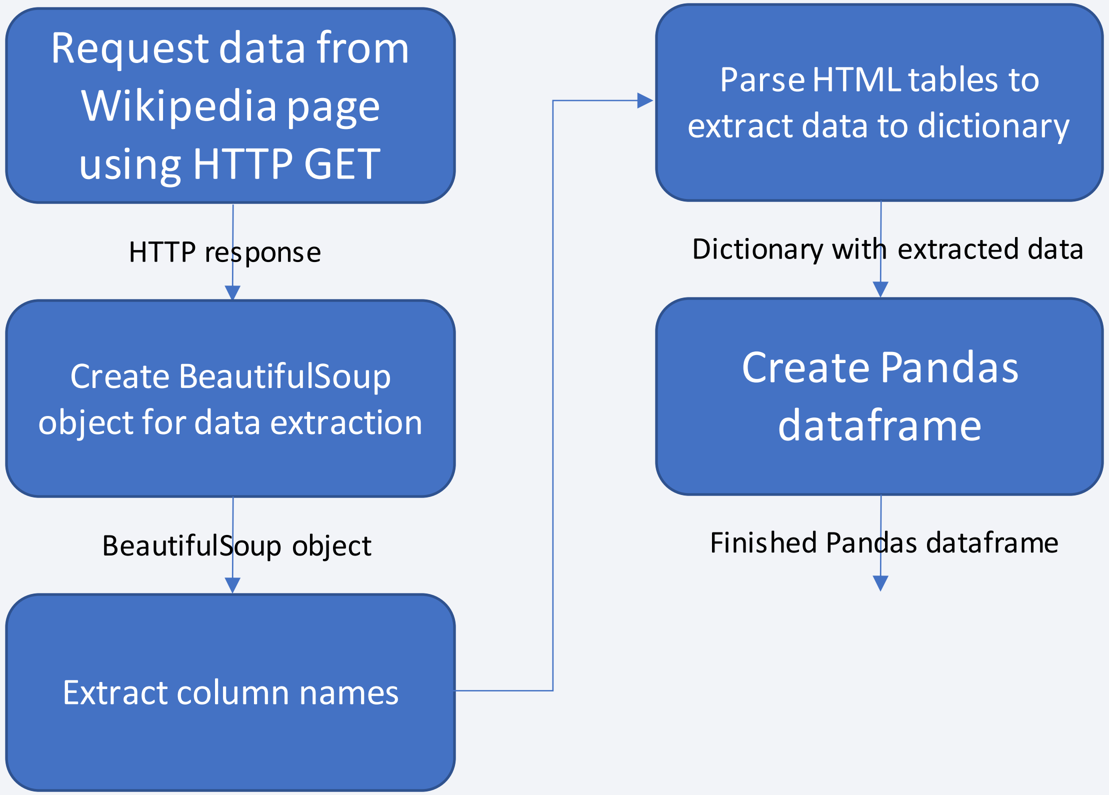
Data Collection – SpaceX API

- Requested data from SpaceX API at https://api.spacexdata.com/v4/launches/past
- Converted JSON file to Pandas dataframe
- Filtered columns to include rocket, payloads, launchpad, cores, flight number, and date; removed multi-core and multi-payload rockets; converted to datetime format and restricted dates of launches.
- Imported data from the rocket IDs through the SpaceX API – imported BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude variables
- Filtered data to include only Falcon 9 launches
- Replaced null PayloadMass values with mean value
- Finished product – Pandas dataframe to store as CSV
- GitHub link to Data Collection Notebook


Request data from SpaceX API
JSON file containing data
Convert to Pandas dataframe
Pandas dataframe
Filter data columns
Filter multi-core rockets
Convert and filter dates
Import data from rocket IDs through API
Pandas dataframe
Filter for Falcon9 launches
Pandas dataframe
Replace null PayloadMass values with mean value
Finished Pandas dataframe

Data Collection – Web Scraping

- Request data from Wikipedia page at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Create BeautifulSoup object for data extraction
- Extract column names from the HTML table on Falcon 9 launches
- Parse HTML tables for data on Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, Time
- Finished product – Pandas dataframe to be stored as CSV file
- GitHub link to Data Collection via Web Scraping


Request data from Wikipedia page using HTTP GET
HTTP response
Create BeautifulSoup object for data extraction
BeautifulSoup object
Extract column names
Parse HTML tables to extract data to dictionary
Dictionary with extracted data
Create Pandas dataframe
Finished Pandas dataframe

Data Wrangling

- Investigated data to check for missing values on each attribute – LandingPad has missing values where no landing pad is used
- Check data to determine column datatypes
- Calculate number of launches for each launch site and each orbit
- Calculate number and occurrences for each outcome per orbit type
- Label outcomes for landing classes – 0 for an unsuccessful landing, 1 for a successful landing

- GitHub link to Data Processing notebook

EDA with Data Visualization

- During Exploratory Data Analysis, charts were developed to investigate the data and determine which variables would be useful in building a model. Several variables were plotted against each other to find relationships.
- FlightNumber/PayloadMass was plotted to determine if there is a relationship between mass of the payload and later flights.
- FlightNumber/LaunchSite was plotted to determine if launch sites were used more often in certain situations. LaunchSite/PayloadMass were plotted to determine if heavier payloads were more likely from a particular site.
- FlightNumber and Orbit type, as well as Payload vs Orbit were plotted to see if the destination orbit had a strong effect on the landing success.
- Finally, Year vs average success rate was plotted to give a baseline for the launch success trend versus which the other variables could have effects.
- GitHub link to EDA with Data Visualization

EDA with SQL

- Exploratory data analysis was also performed by running SQL queries from a database containing launch data.
  - Display unique launch sites and investigate CCA launches
  - Display total payload mass carried by boosters launched by NASA; display average payload mass carried by F9 v1.1 boosters
  - List the boosters successful in drone ship landings with payload mass between 4000 and 6000 kg; list boosters that have carried maximum payload mass
  - List total number of successful and failed mission outcomes
  - List failed landing outcomes for drone ships and their booster versions with launch site names in 2015
  - Rank the landing outcomes in descending order
- GitHub link to EDA with SQL notebook

Build an Interactive Map with Folium

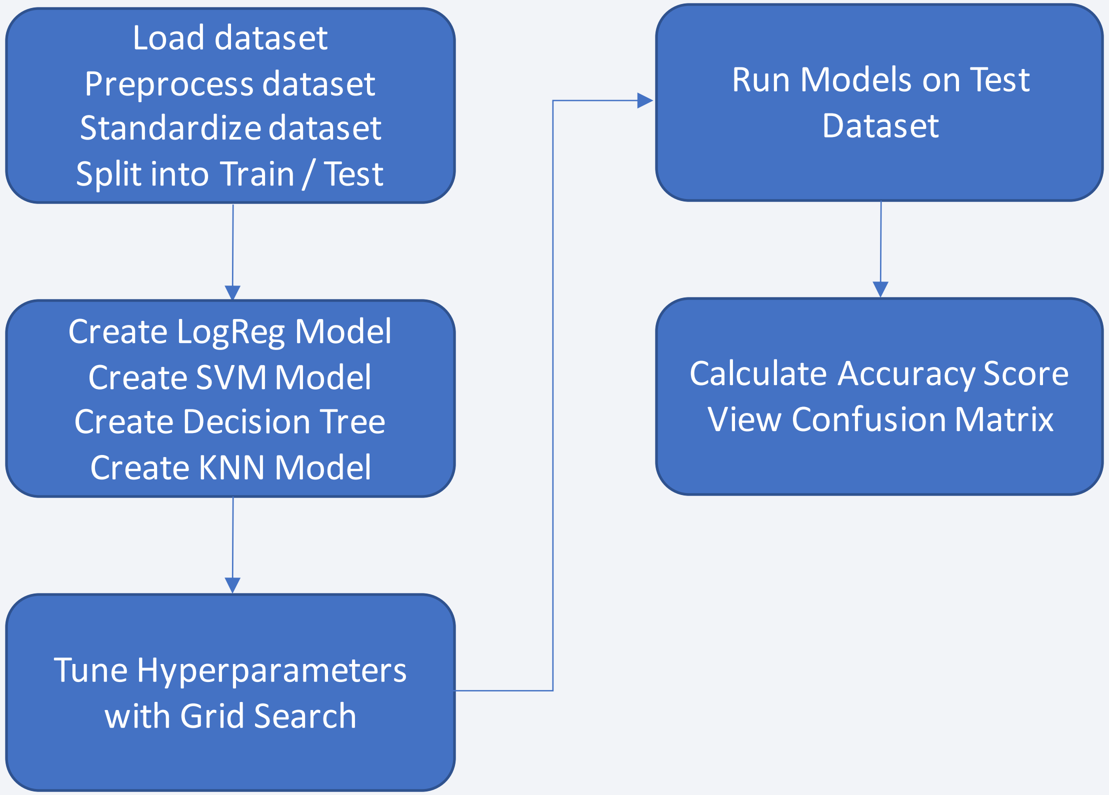
- Folium maps were generated to visualize launch sites and successful/failed launches.
- Circles and markers were placed on each of the launch sites and labeled.
- Successes and failures for each launch site were marked on the map using marker clusters.
- Distances to nearby features such as coastlines, railways, roads were created and plotted via lines.
- These features were identified to help visualize the factors that lead to successful recoveries at particular launch sites.
- GitHub link to Folium notebook

Build a Dashboard with Plotly Dash

- A Plotly dashboard was created to aid in visualizing landing success for various launch sites.
- A pie chart with selectable launch site was built to visualize landing successes and percentages for each individual launch site, with a default option also included to visualize the percentage of all successes contributed by each site.
- A scatter chart with range slider was generated to visualize the correlation between payload mass and landing success, with the range slider used to control the minimum and maximum payload mass displayed.
- GitHub link to Plotly dashboard code

Predictive Analysis (Classification)

- Data for the model variables was loaded, preprocessed and standardized, and split into training/test datasets.
- Models for logistic regression, support vector machine, decision tree, and k-nearest neighbors were created on the training dataset and tuned for their best set of hyperparameters using grid search.
- Each model was run on the test dataset to define performance.
- Model performance was compared using the confusion matrix for each model's analysis of the test dataset.
- GitHub link to Machine Learning Prediction notebook



Load dataset
Preprocess dataset
Standardize dataset
Split into Train / Test

Create LogReg Model
Create SVM Model
Create Decision Tree
Create KNN Model

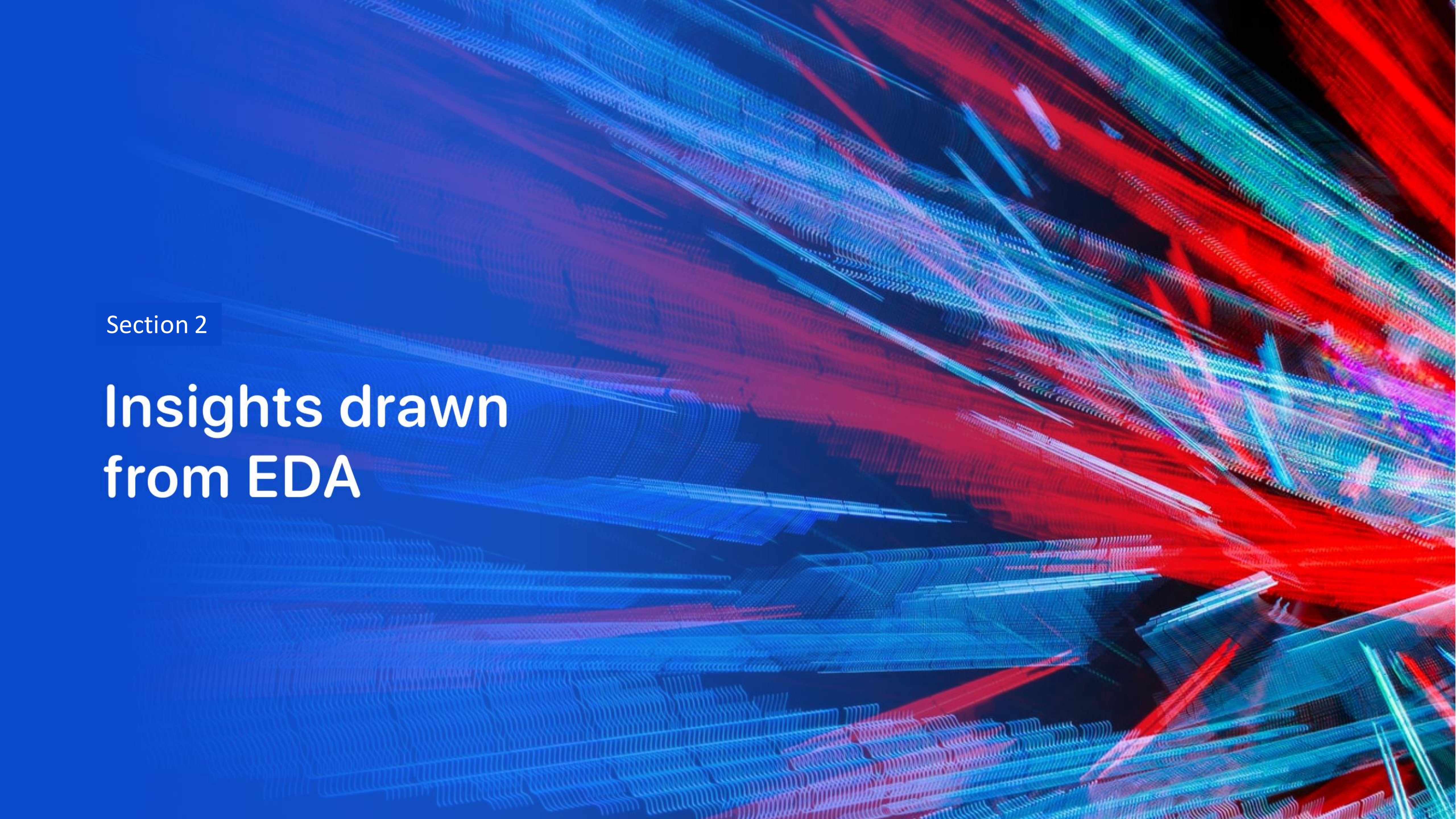
Tune Hyperparameters with Grid Search

Run Models on Test Dataset

Calculate Accuracy Score
View Confusion Matrix

Results

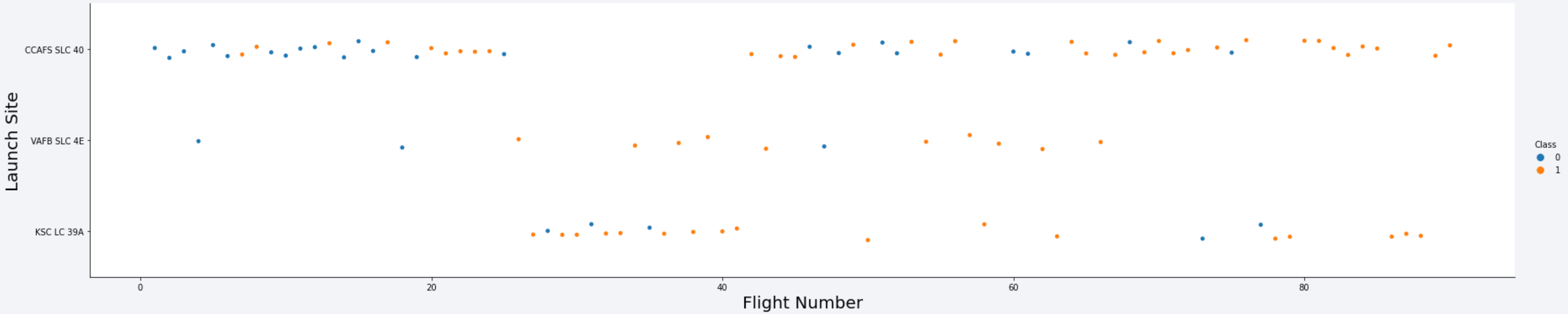
- Exploratory data analysis showed several variables were important in modeling
  - FlightNumber / Year useful in showing more recent launch success
  - LaunchSite useful in showing relationship between site and success (modified by later flights)
  - Orbit shows missions that are more likely to be successful
  - PayloadMass shows that for certain classes of payload, success has a relationship to mass
- Interactive analytics demo to be shown in screenshots in next slides.
- Predictive analysis for the various models shows that many models show similar performance on the test dataset.
  - LogReg, SVM, KNN models all showed same performance on the test dataset
  - Decision Tree model showed more false negatives than other models



Section 2

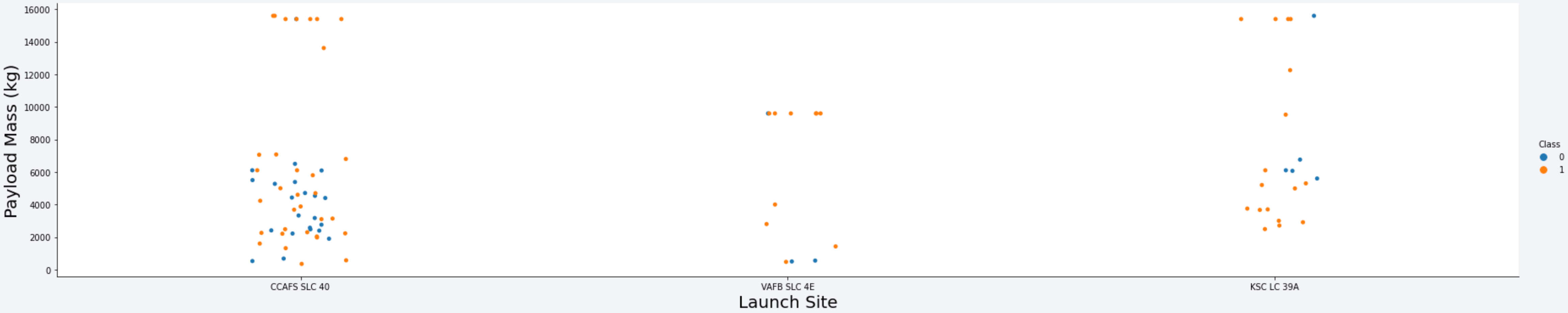
Insights drawn from EDA

Flight Number vs. Launch Site



- This plot shows the successes and failures for launch sites CCAFS SLC-40, VAFB SLC-4E, and KSC LC-39A.
- The chart shows that later launches (higher flight numbers) have a greater chance of landing success.
- The chart also shows that many of the early failures occurred at CCAFS, which could imply that the launch site CCAFS may not be directly linked to landing failure but that the statistics are affected by the early landing failures; statistics after launch 25 may be more representative of current and future capability.

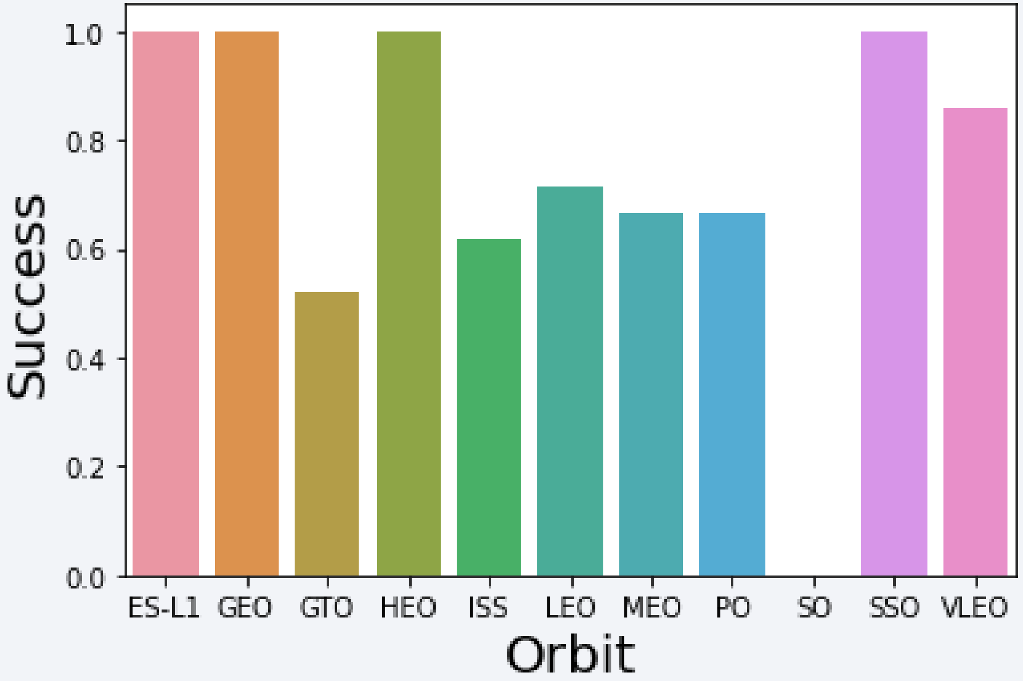
Payload vs. Launch Site



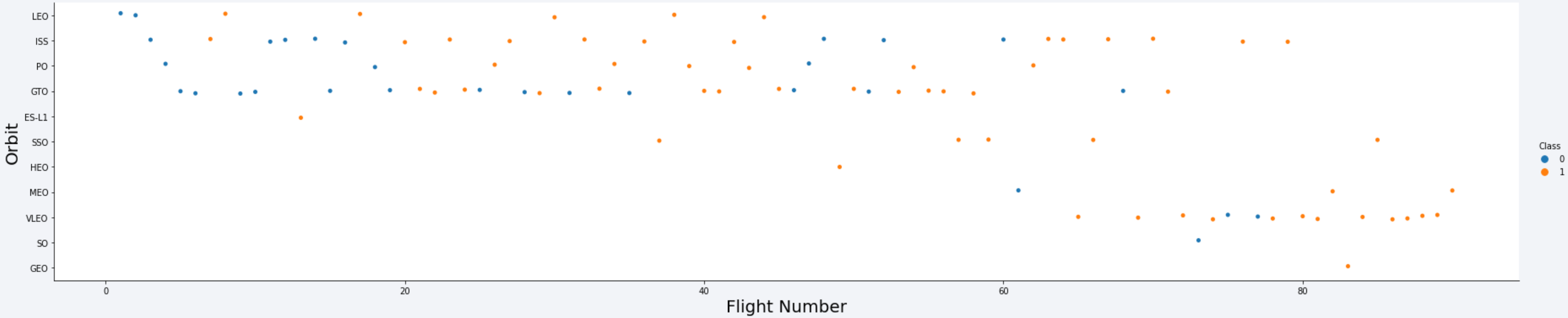
- This plot shows the successes and failures by payload mass and launch site.
- The chart shows that payloads above a certain mass (10000kg and up) appear to have a better chance of landing than payloads <8000kg.
- Among the payloads <8000 kg, KSC LC 39-A shows all its failures in the higher payloads. The failures appear to be distributed evenly among payloads for CCAFS SLC-40 and VAFB SLC-4E does not have enough data to show a trend.
- Additional investigation is still needed – perhaps higher-mass payloads are destined for different orbits or higher-mass payloads were more common for later launches.

Success Rate vs. Orbit Type

- This chart shows the success rate by orbit type.
- From the chart it can be seen that certain orbits have 100% success rate – ES-L1, GEO, HEO, and SSO orbits.
- Other orbits are less successful ranging from ~83% success in VLEO to ~50% for GTO and 0% success for SO orbits.

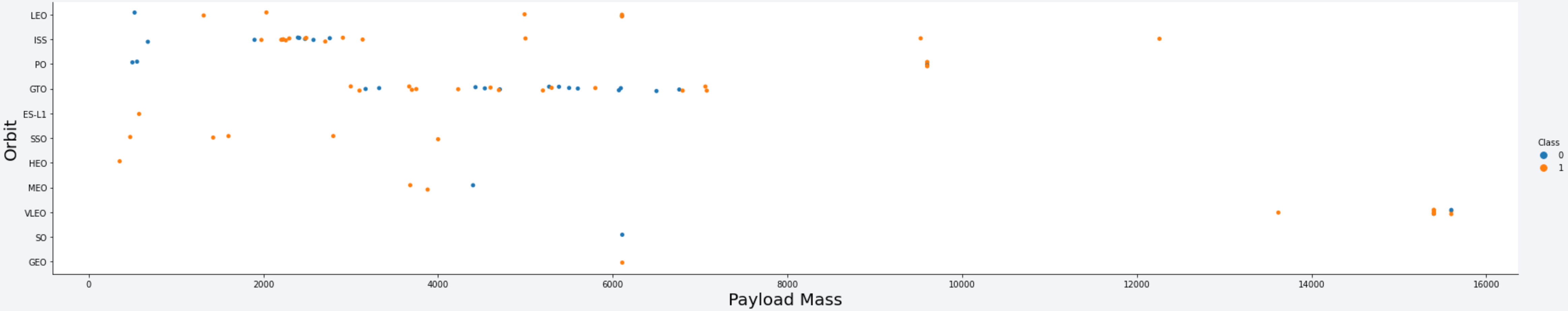


Flight Number vs. Orbit Type



- This chart shows flight number versus orbit type.
- It shows that all of the early flights were to LEO / ISS / PO / GTO orbits with a few other orbits sprinkled in.
- The majority of the later flights (after flight 60) were to VLEO.
- The 100% success rate of HEO, GEO and ES-L1 orbits can be explained by the success of the single launch. SO and SSO orbits also had a very small number of launches.

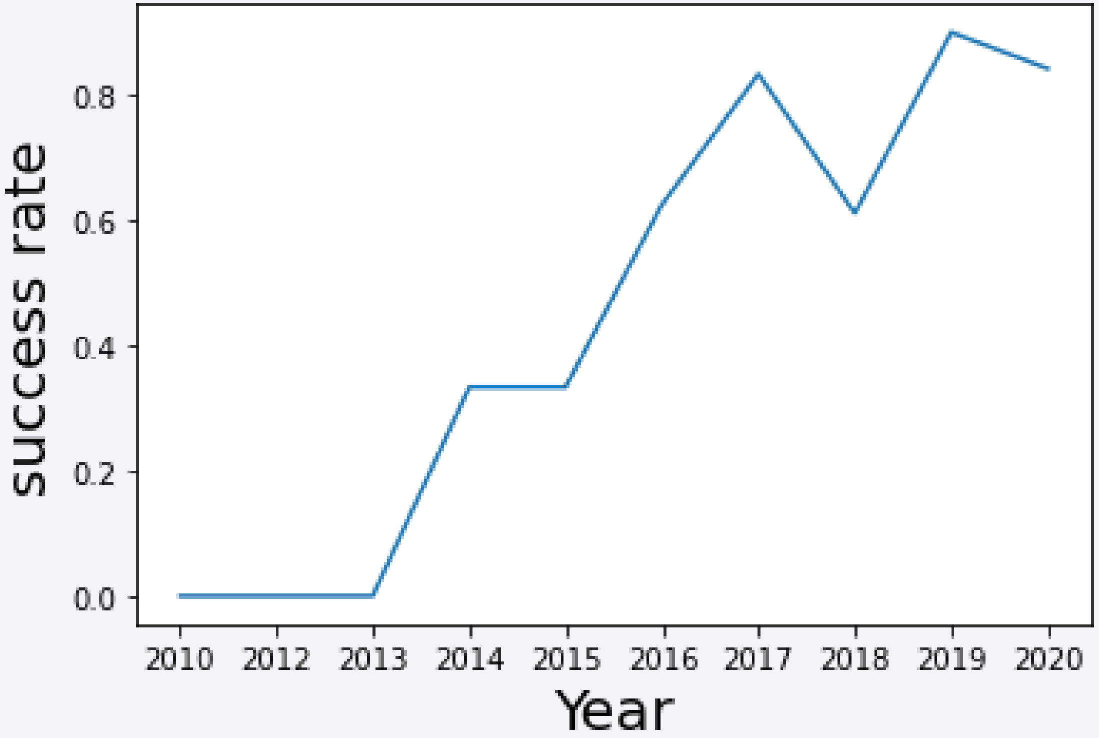
Payload vs. Orbit Type



- This chart shows payload mass versus orbit type.
- It continues to show that large payloads have greater landing success rates, especially for LEO, ISS, and PO orbit types.
- GTO data is unclear with successful and unsuccessful missions at similar payload masses.
- It appears that all of the VLEO orbit payloads were high-mass.

Launch Success Yearly Trend

- This chart shows average success rate by year.
- It shows that successes started occurring in 2014 and continued to increase toward 2020.
- Given that the average success rate from 2017-2020 appears to be near 80%, it is likely that future launches will have a similar success rate with its associated lower cost of launch.



All Launch Site Names

- Launch site names are in the table to the right
- Use query

```
%%sql select distinct LAUNCH_SITE from SPACEXDATASET
```

- These four launch sites have different probabilities of a successful landing

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

Launch Site Names Begin with 'CCA'

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 with launch site name starting with 'CCA' are shown here.
- Use query

```
%%sql select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' LIMIT 5
```

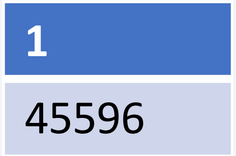
- This shows the first few launches from CCAFS LC-40 and their landing outcomes (failures), with their target orbits, payloads, mission outcomes, etc.

Total Payload Mass

- This query shows the total payload mass in kilograms for the customer NASA (CRS). NASA is a fairly large customer.

- Use query

```
%%sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA
(CRS)'
```



1
45596

Average Payload Mass by F9 v1.1

1
2534

- This query calculates the average payload mass carried by booster version F9 v1.1.
- Use query

```
%%sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where booster_version like 'F9
v1.1%'
```

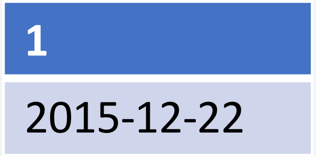
- The average payload mass is about 2500kg. F9 v1.1 did not carry a large payload on average, compared to the size of a typical launch.

First Successful Ground Landing Date

- This query shows the first successful ground pad landing date.
- Use query

```
%%sql select min(date) from SPACEXDATASET where landing__outcome like 'Success%'
```

- The first successful landing did not occur until 2015, although the first SpaceX launch occurred in 2010.



1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query shows the booster versions that have successfully landed on a drone ship with payload mass between 4000 and 6000 kg.
- Use query

```
%%sql select booster_version from SPACEXDATASET where landing__outcome like 'Success (drone%' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

- This query shows that only four booster versions have successfully landed on a drone ship with payload between 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- This query shows the total number of successful and failure mission outcomes.

- Use query

```
%%sql select landing__outcome, count(*) from SPACEXDATASET
group by landing__outcome
```

- The query shows that many of the landings were successes, but also that many of the failures were cases where no attempt was made to land.

landing__outcome	2
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- This query shows the boosters that have carried the maximum payload mass in the dataset.
- Use query

```
%%sql select booster_version from SPACEXDATASET where payload_mass__kg_ =
(select max(payload_mass__kg_) from SPACEXDATASET)
```

- The query shows that the variants of the Falcon9 Block 5 booster are the only types to carry the maximum payload mass.

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

- This query shows the failed landing outcomes on drone ships, with their booster versions and launch site names for launches in 2015.
- Use query

```
%%sql select landing__outcome, booster_version,
launch_site from SPACEXDATASET where date like
'2015%' and landing__outcome like 'Failure (drone%'
```

- The query shows that there were 2 failures to drone ships in 2015, both launching from CCAFS LC-40.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query ranks the landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

- Use query

```
%%sql select landing__outcome, count(landing__outcome) from
SPACEXDATASET where date > '2010-06-04' and date < '2017-03-20'
group by landing__outcome order by count(landing__outcome) DESC
```

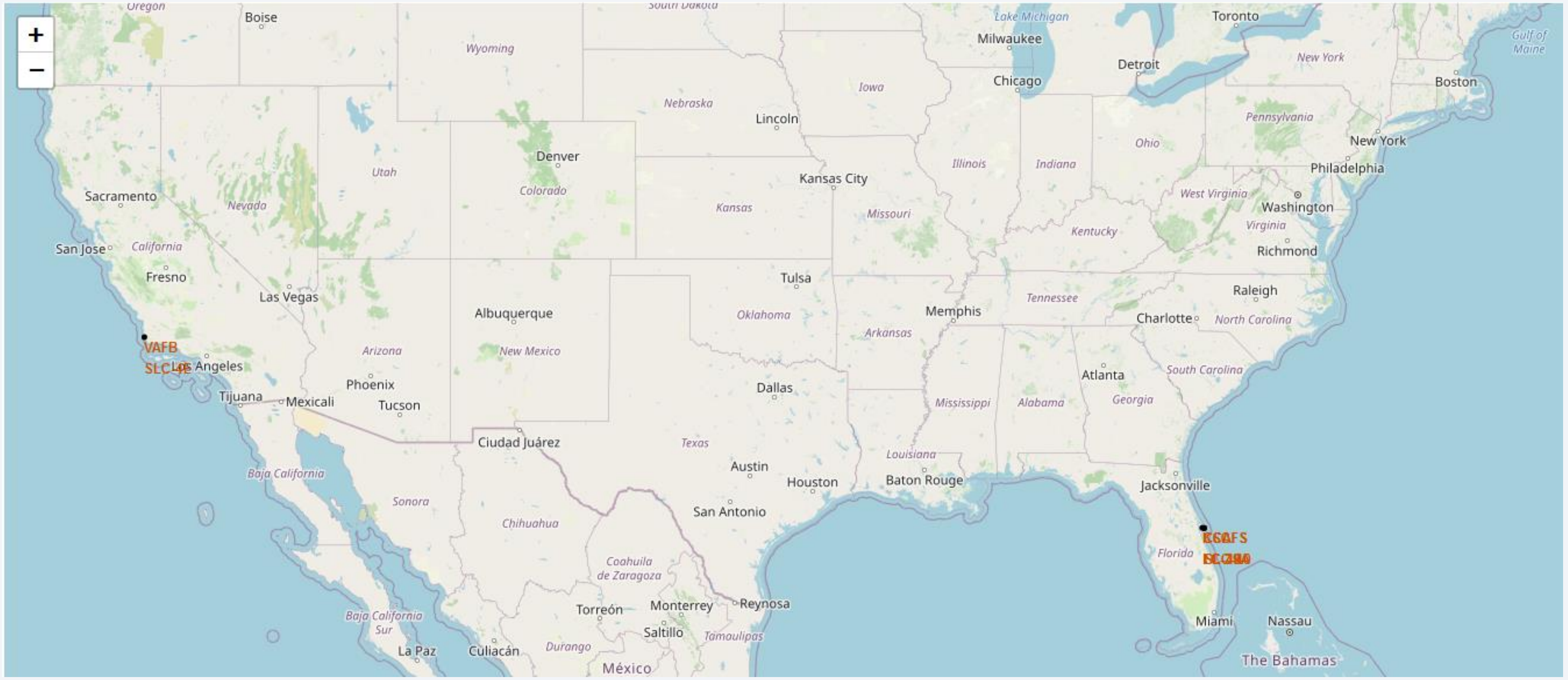
- The query shows that many of the launch attempts before early 2017 did not attempt a landing, and successes were close to evenly balanced with failures when landings were attempted.

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

Section 3

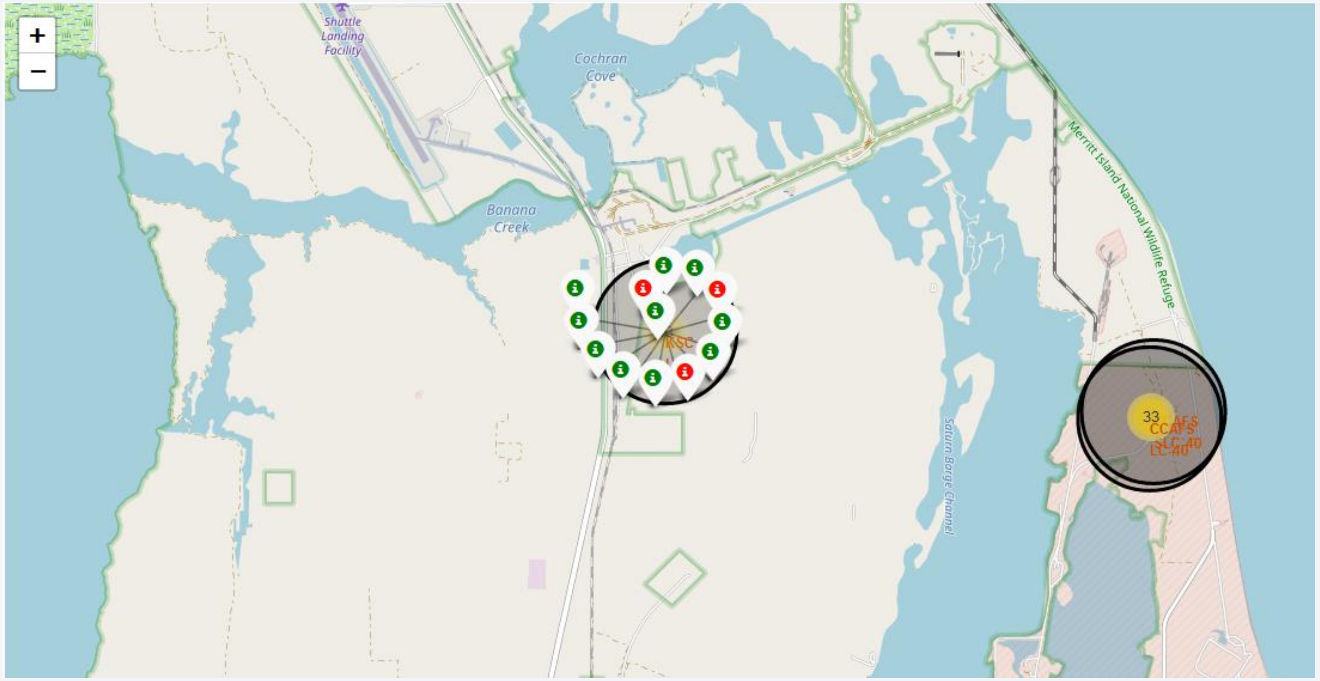
Launch Sites Proximities Analysis

Launch Site Locations



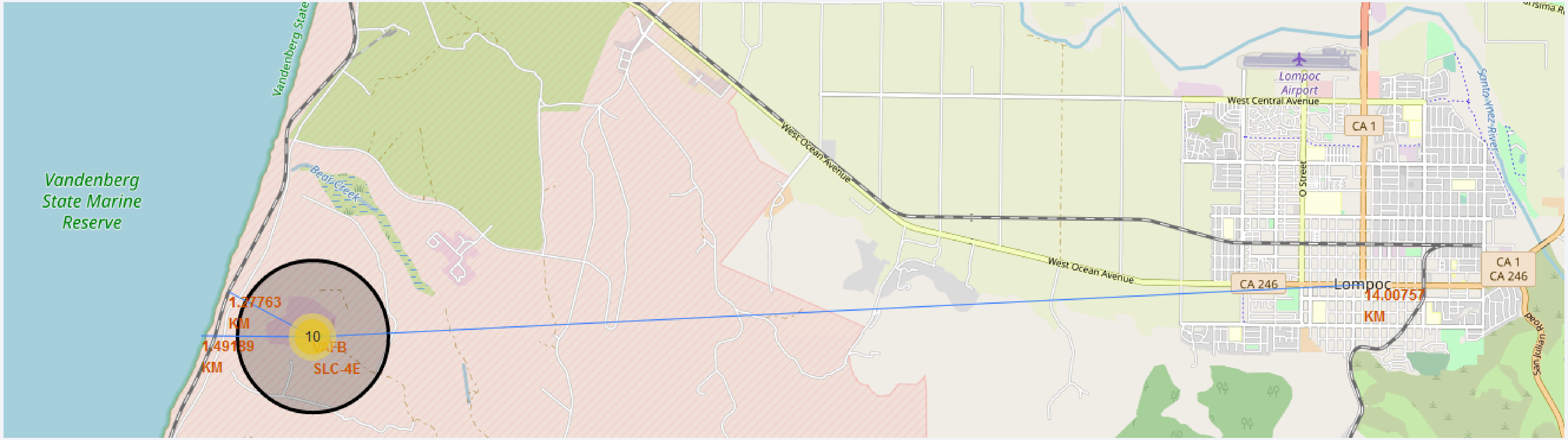
- This map shows the locations of the launch sites used by SpaceX.
- Vandenberg AFB is on the West Coast of the United States, in southern California.
- Cape Canaveral and Kennedy Space Center are located on the East Coast of Florida.
- These sites are closer to the equator than most of the United States and close to coasts, aiding in launch velocity (rotational speed of the Earth) and avoiding major population centers.

Launch Outcome Clusters



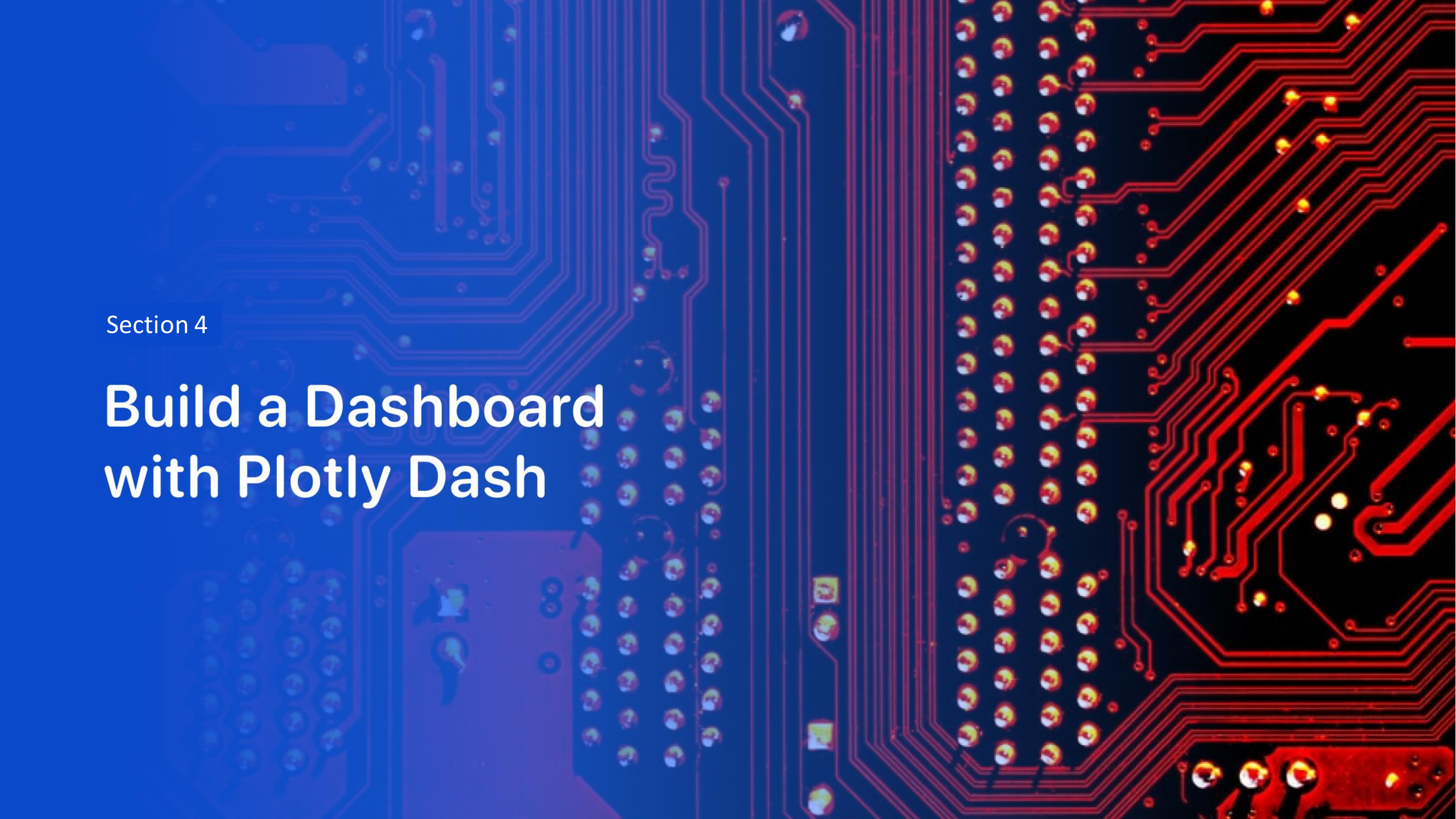
- This map shows the successes (green) and failures (red) for launch site KSC LC-39A.
- It rapidly shows the relative success rate for launches from KSC; the high percentage of green versus red markers shows a good success rate for landings at this site.
- Similar data is available for visualization at the CCAFS and VAFB sites.

Distances to Cities, Railroads, Coastlines



- This map shows the relative location of the VAFB SLC-4E launch site to the nearby town of Lompoc, as well as the nearest railroad and coastline.

- Note that the VAFB launch site is very close (< 2 km) to the coast but also close to railroads and roads. However, Vandenburg AFB is on the West Coast of the United States, which makes it a less ideal launch location as most rocket launches travel to the east to take advantage of the Earth's rotation.



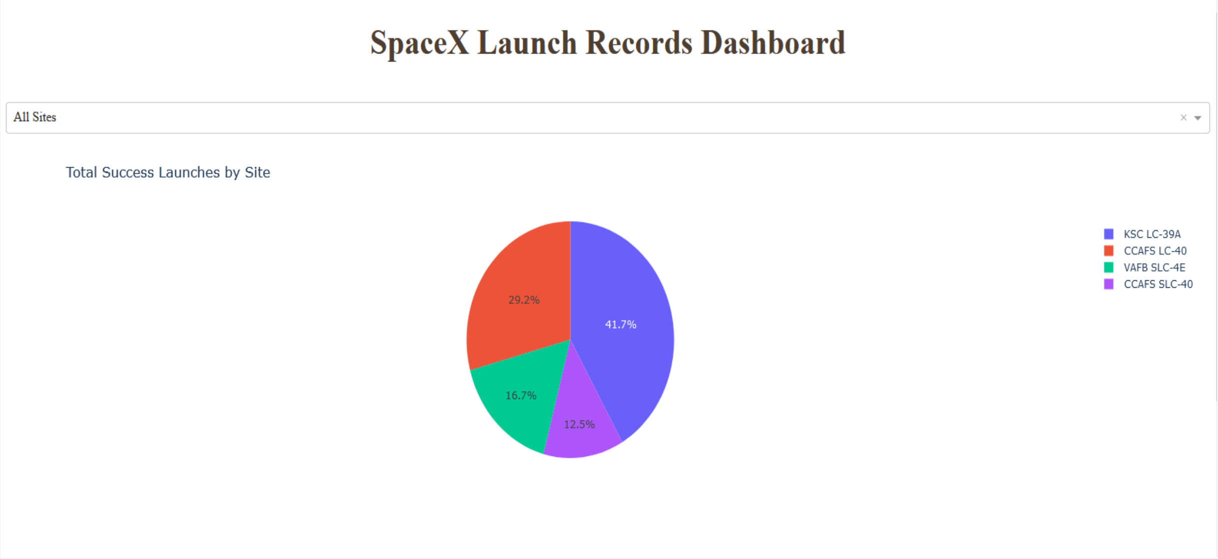
Section 4

Build a Dashboard with Plotly Dash

All Sites – Launch Success

- This dashboard pie chart shows the successes across all launch sites, split by percentage of total successes from each site.

- The pie chart indicates that KSC LC-39A is the site that has the largest number of successes (41.7% of all successes), followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40.

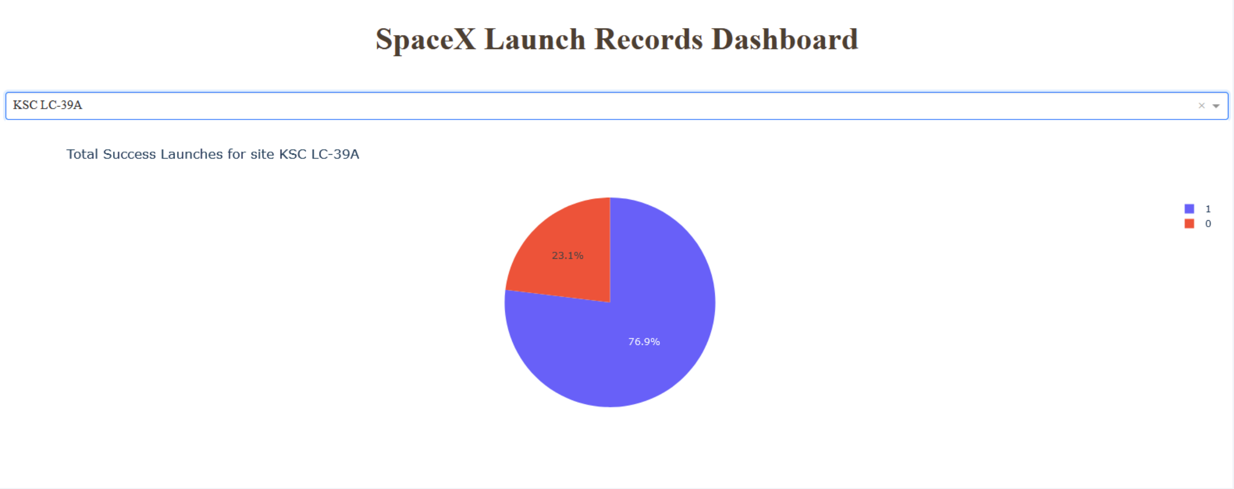


SpaceX Launch Records Dashboard

All Sites

Success Rate for KSC LC-39A

- The dashboard pie chart is now showing the launch success rate for KSC LC-39A.
- This site has the highest success ratio among the four launch sites.
- Its success ratio is 76.9% (10/13 successful launches recovered).

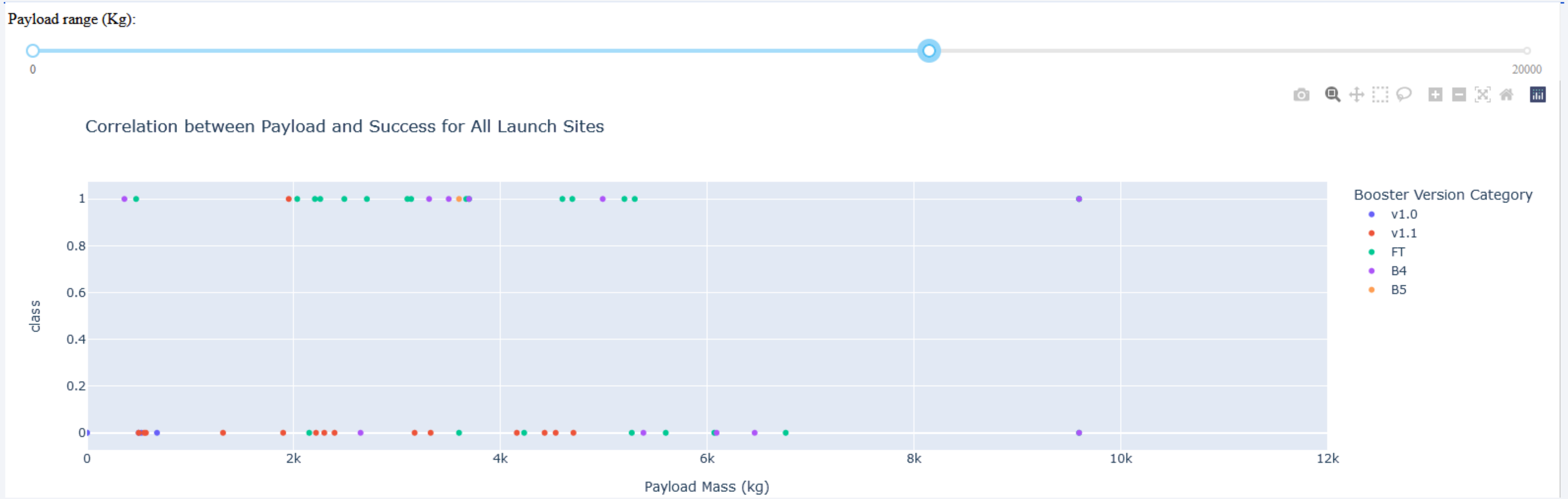


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload vs Launch Outcome Dashboard



- This portion of the dashboard contains a scatter chart showing the correlation between payload and landing success for all launch sites.

- It shows that in this data set, the v1.0 and v1.1 boosters were less successful than later boosters such as the Full Thrust and Block 4/5 boosters.

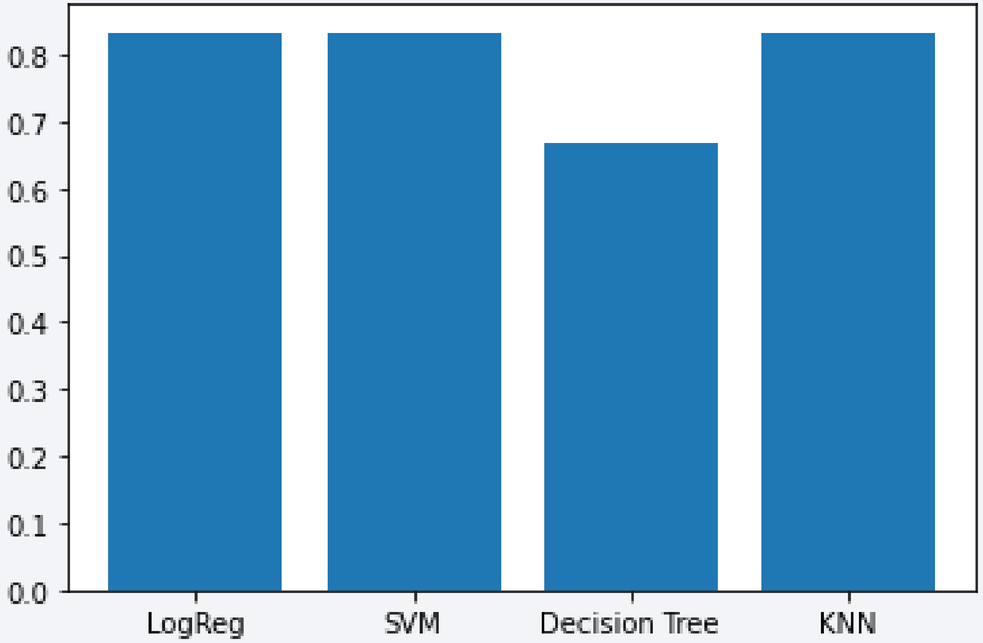
- Payload ranges in the 2000 - 4000 kg category appear to be more successful. It should be noted that several payloads overlap on this chart so may not be as obvious at first glance.

Section 5

Predictive Analysis (Classification)

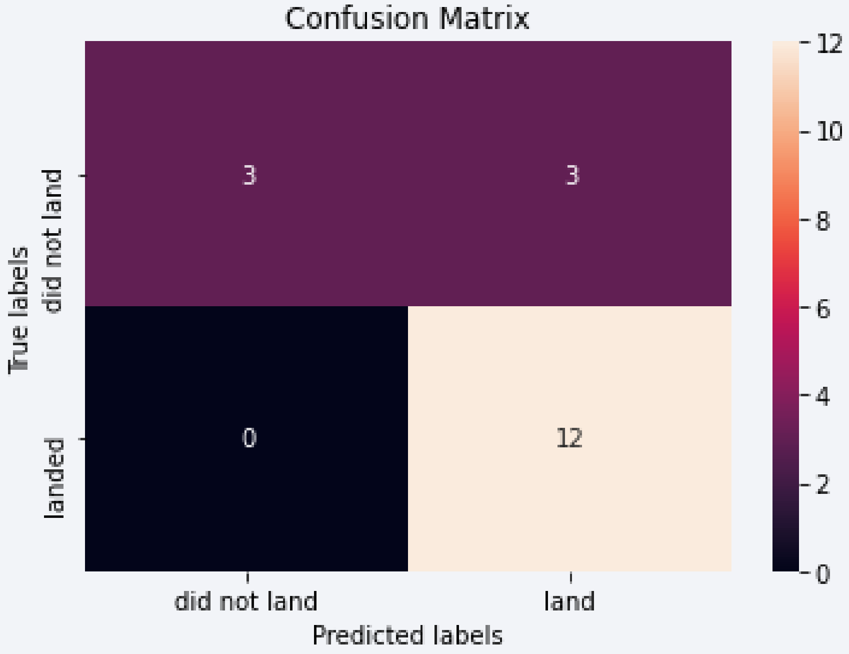
Classification Accuracy

- The bar chart here shows the comparative classification accuracy on the test dataset for four models with tuned hyperparameters; one logistic regression, one support vector machine, one decision tree and one k-nearest neighbors model.
- On the test dataset, the LogReg, SVM, and KNN models all performed at the same level of accuracy (83.3%, or 15/18 cases classified correctly).



Confusion Matrix

- All three of the models with high test dataset accuracy (LogReg, SVM, KNN) classified the cases in the same way, with the same confusion matrix as shown here.
- The models correctly classified three landing failures and twelve landing successes, but predicted that three of the test cases would land when they were not in fact successful. There were no cases where the models predicted a failure when the rocket did in fact land successfully.



Conclusions

- Using the dataset of SpaceX landings to date, I would recommend using a tuned SVM model to predict the ability of SpaceX to land the first stage of their F9 rockets. This model is estimated to be 83.3% accurate based on test data.
- The errors in this model on the test dataset all predicted a successful landing when a landing was not in fact successful. This means that the model errors may lead to an under-prediction of costs for SpaceX and may mean that our company misses out on a bid where SpaceX costs would be higher than we predict.
- Future refinements to this model could include pruning data to predict current and future performance (as SpaceX landing capability has grown with time, making earlier data less relevant).

Thank you!