IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jiro Inoue
Feb. 6, 2022

Outline

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

Executive Summary

- Data from the SpaceX API and scraped off Wikipedia were used to model Falcon 9 rocket launches and conduct exploratory analysis
- Interactive visualizations were used to understand the data
- Results show that successful landing of the stage 1 rocket was not random, and was correlated to certain parameters of the launch
- Using a machine learning model, we can accurately predict successful landing of the stage 1 rocket based on the parameters and use this knowledge to improve the success rate of landing the stage 1 rocket

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.*
- We seek to answer the following questions:
  - What are some important factors associated with Falcon 9 launches?
  - Which factors are most associated with successful stage 1 landings?
  - Can we design a machine learning algorithm to predict if the first stage will land successfully?

* https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/AddingNotebook_toWatson.md.html

Section 1

Methodology

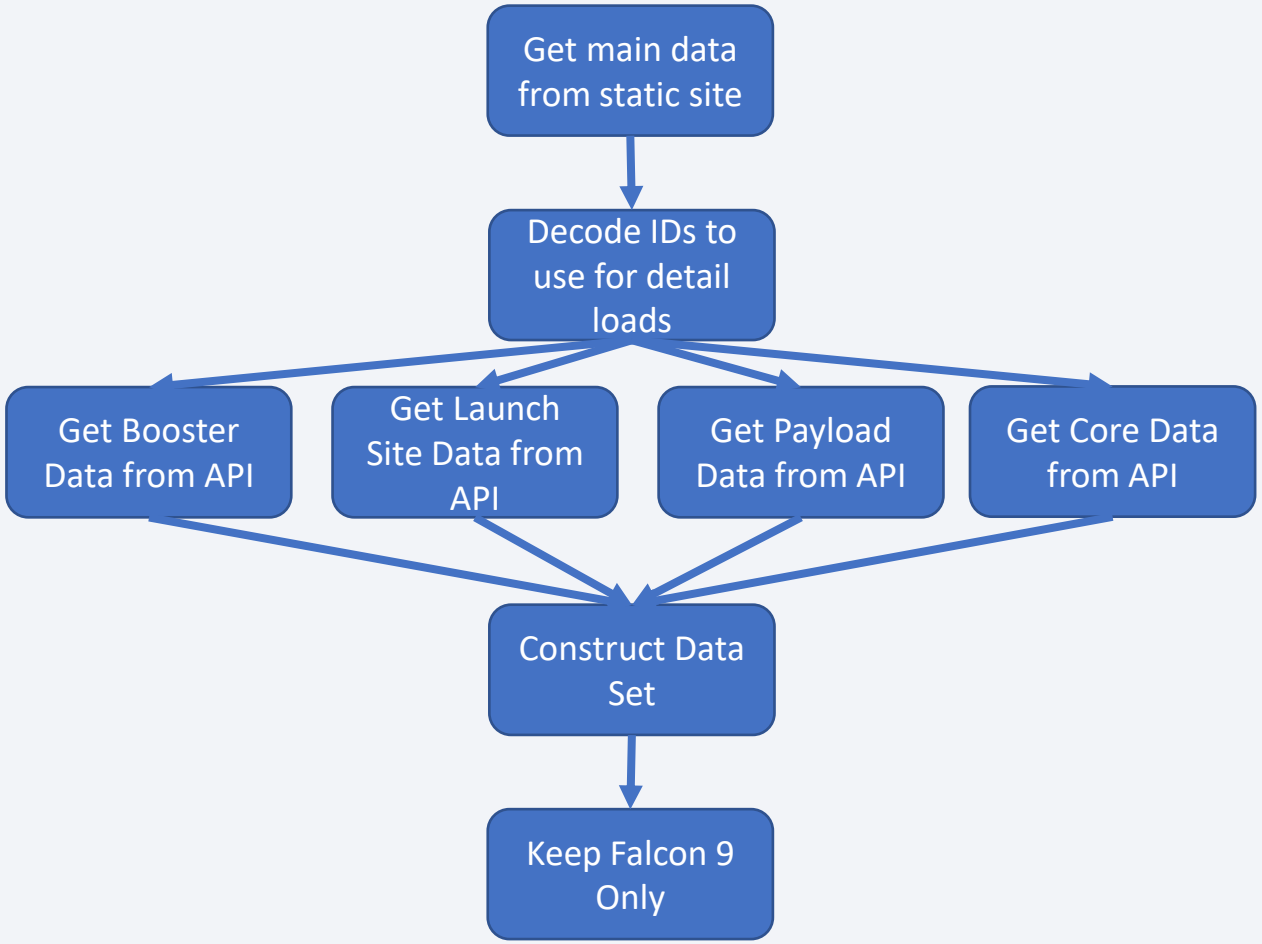
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from https://api.spacexdata.com using a REST interface, and scraping https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 with BeautifulSoup
- Perform data wrangling
  - Data was cleaned by replacing missing payload mass values with the mean payload mass
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were built using scikit-learn and tuned using GridSearchCV

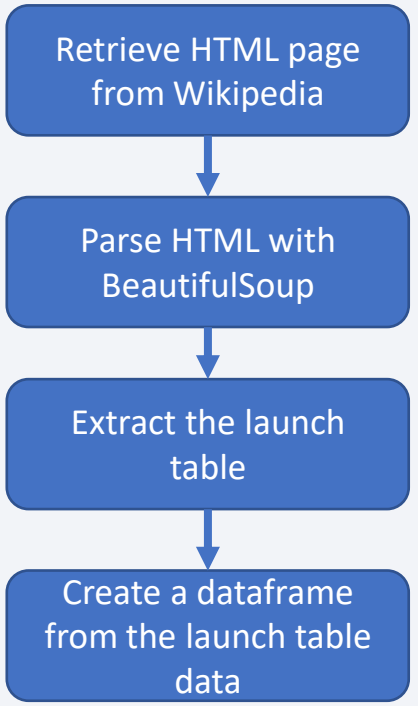
Data Collection – SpaceX API

- The requests library was used to make SpaceX REST calls to the site https://api.spacexdata.com.
- Request.get was used to retrieve data from a static site
- The data was decodes and separated to get IDs for detail data
- Detail data was retrieved from https://api.spacexdata.com using REST
- All data was combined and only Falcon 9 data was kept
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20Data%20Collection%20API%20Lab.ipynb



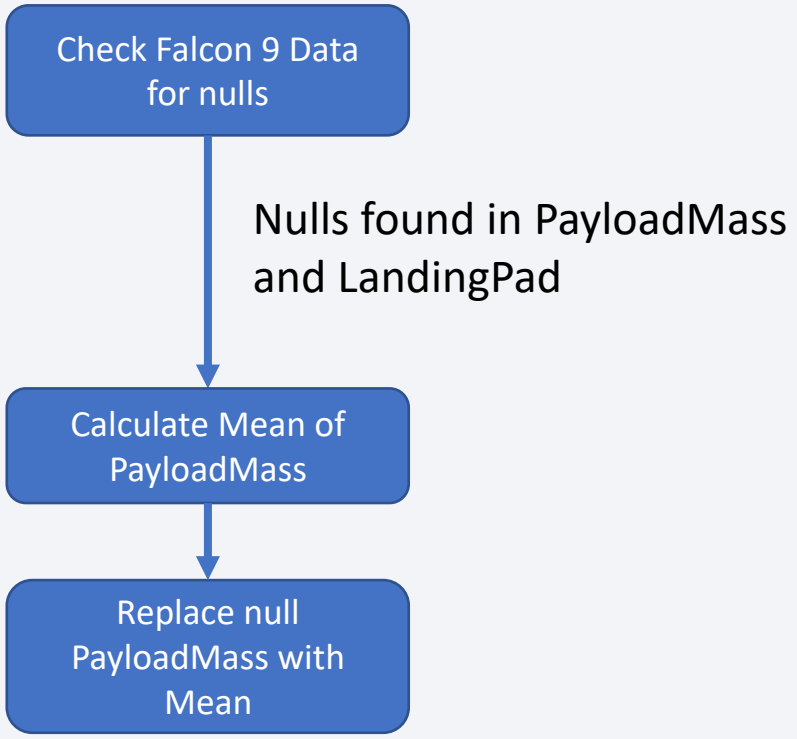
Data Collection - Scraping

- Falcon 9 historic records were scraped from Wikipedia
- Data was scaped from June 9th, 2021 version of the page
- HTTP GET was used to retrieve the page
- Beautiful soup was used to extract the launch data table and a Pandas dataframe was created for the data
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

- Some rows from the SpaceX API had missing values
- 5 missing values of PayloadMass were replaced with the mean value of the PayloadMass column
- 26 missing values of LandingPad were left
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20EDA%20lab.ipynb



EDA with Data Visualization

- Exploratory Data Analysis was performed with the following charts
  - Five scatter charts (Flight Number vs. Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Flight Number vs Orbit, Payload Mass vs Orbit) with Classes 0 and 1 in different colors to see the relationship between two characteristics and effect on success rate
  - Bar chart of Orbit vs. Class to find which orbits have high success rates
  - Line chart of Year vs. Class to show that success rates kept increasing from 2013 to 2020
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- The following SQL queries were used:
  - select distinct LAUNCH_SITE from SPACEXTBL
  - select * from SpaceXTBL where LAUNCH_SITE like 'CCA%' limit 5
  - select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
  - select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
  - select DATE from SPACEXTBL where landing__outcome = 'Success (ground pad)' order by DATE asc limit 1
  - select distinct booster_version from SPACEXTBL where Landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 3999 and 6001
  - select sum(case when Mission_outcome like '%Success%' then 1 end) as Successes, sum(case when Mission_outcome like '%Fail%' then 1 end) as Fails from SPACEXTBL

EDA with SQL

- The following SQL queries were used (continued):
  - select distinct booster_version from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
  - select landing__outcome, booster_version, launch_site from spacextbl where year(date) = 2015 and landing__outcome = 'Failure (drone ship)'
  - select landing__outcome, count(*) as l_o_counts from spacextbl where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(*) desc
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

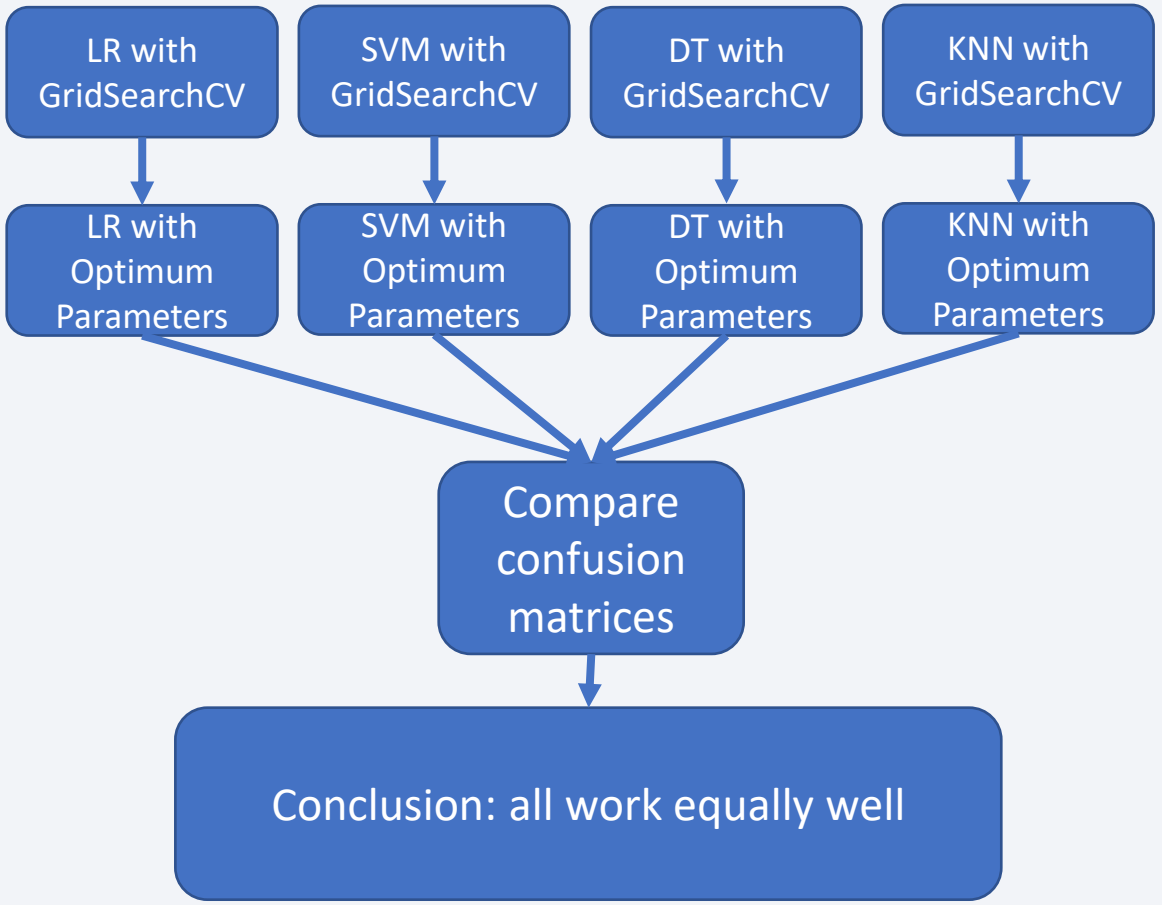
- The following map objects were created:
  - To show key locations:
    - Circles at key locations– NASA Johnson Space Center and Launch Sites
    - Markers with icons as text labels to describe the key locations
  - To show individual launches:
    - Marker clusters that are zoomable and at the high detail level, show individual launches color coded by success/failure
  - To show distances to map features (coastline, railroad, highway, city) from launch site (VAFB SLC-4E)
    - Markers with icons as text labels to describe the distance
    - Polylines to show the measured line
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20Interactive%20Visual%20Analytics%20with%20Fol.ipynb

Build a Dashboard with Plotly Dash

- A dashboard was created with Plotly Dash
- A dropdown menu was used to allow selection of a single or all launch sites
- A pie chart was used to show launch success based on the dropdown
  - If All Sites were selected, then the breakdown of successful launches by site was shown
  - If a single site was selected, then the success/fail rate for that site was shown
  - Pie charts were used because they easily show which sites were more/less successful
- A slider bar was used to allow selection of a payload mass range to inform a scatterplot of success vs payload mass
- A scatterplot was used to show launch success/fail. Colors were used for selected site(s) and the x-axis was based on the payload mass from the slider bar
  - Scatterplot was used because they show two factors influencing success
- See .py file at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py

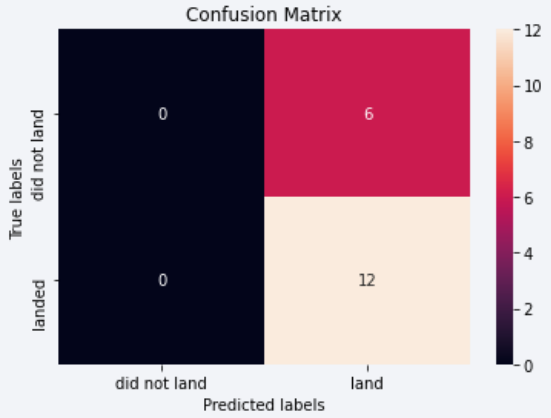
Predictive Analysis (Classification)

- Four Algorithms: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbour were tested for predictive analysis
- The data was split into test-train sets
- GridSearchCV was used to optimize parameters for all four methods and find the best version of each method
- Results on the test set were examined with a confusion matrix and they were all found to be equally good in the best case
- See notebook at https://github.com/githubjiro/Applied-Data-Science-Capstone/blob/master/Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb



Predictive Analysis (Decision Trees)

- The decision tree algorithm produces different results on different runs
- Here we show two different results generated by the same decision tree optimization algorithm on the same data.
- The best possible score after multiple tries appears to be 0.833



SCORE = 0.667

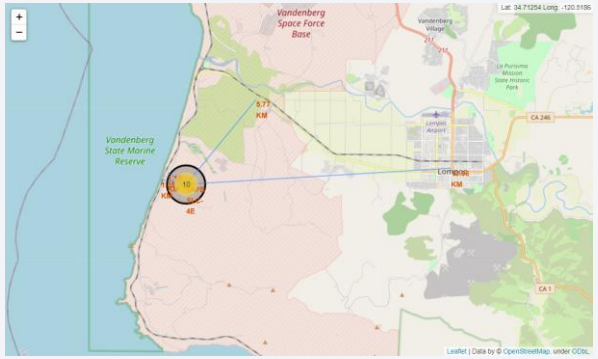
SCORE = 0.833

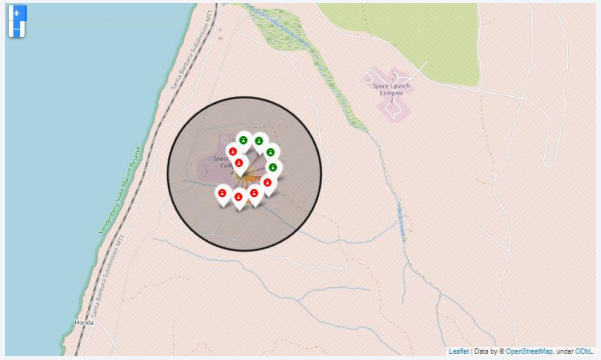
Results (1)

- Exploratory analysis showed that some combinations of factors resulted in greater success, but sample size was small
- Over time, improved knowledge must have been built into the system as parameters such as orbit type used changed, and overall success rates improved to over 80%
- All predictive analysis methods worked equally well after tuning, and accurately predicted success, with the exception of some runs of the decision tree

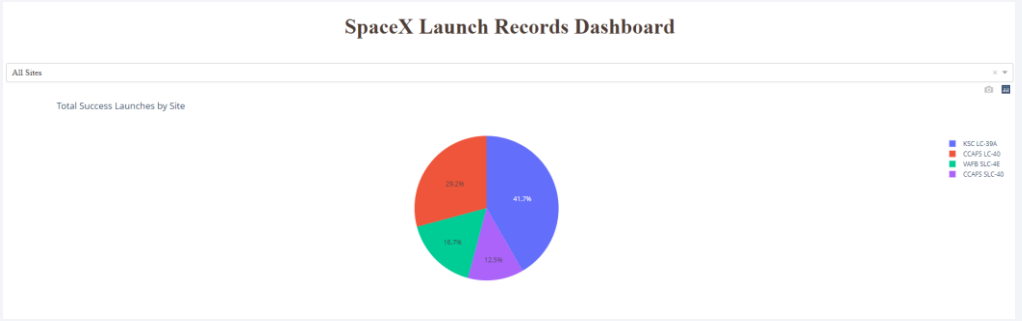
Results (2)

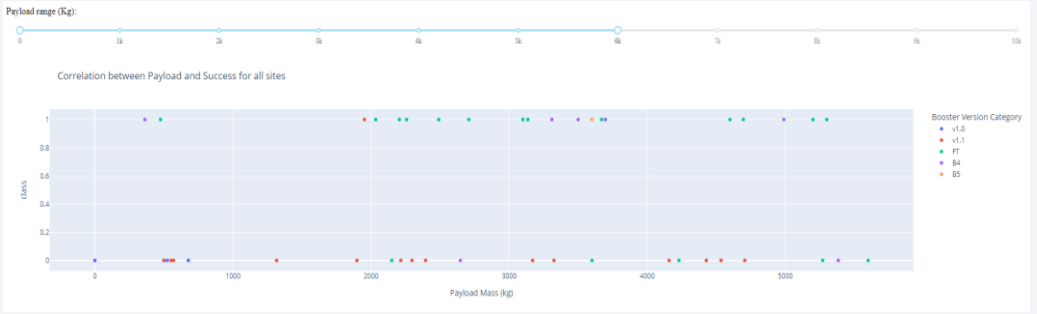
- While performing this analysis, two interactive maps in Folium and a plotly dash dashboard were created.



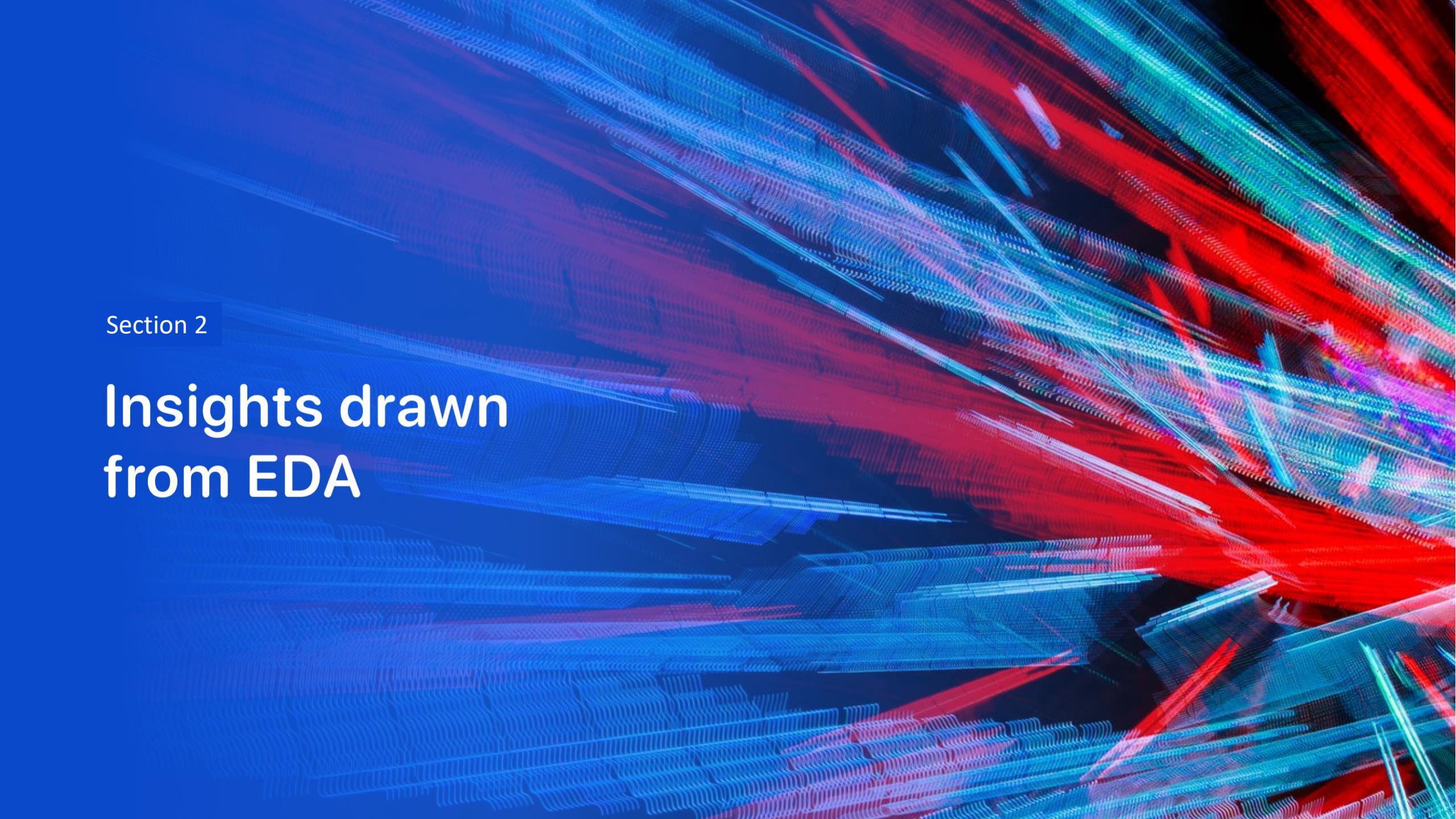


Folium maps

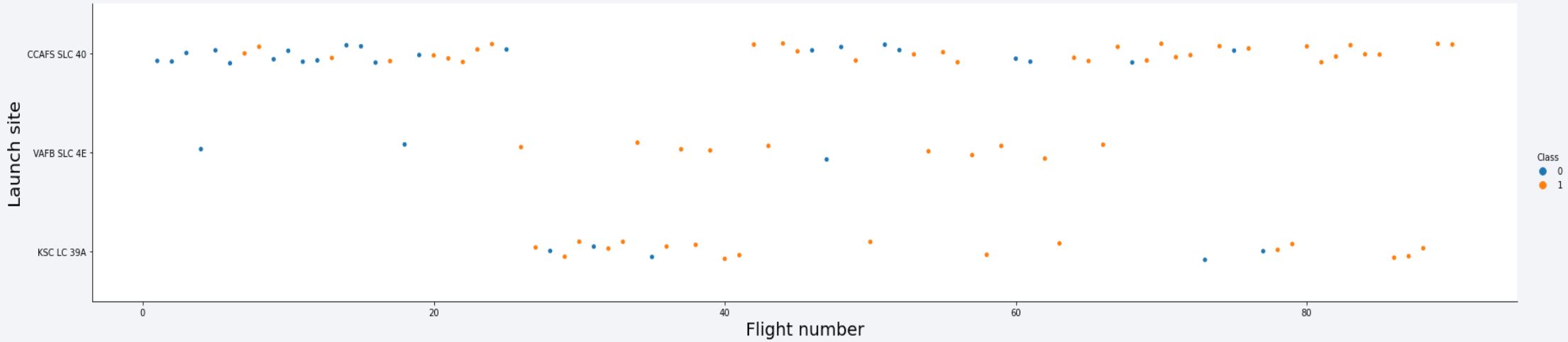




Dash Dashboard

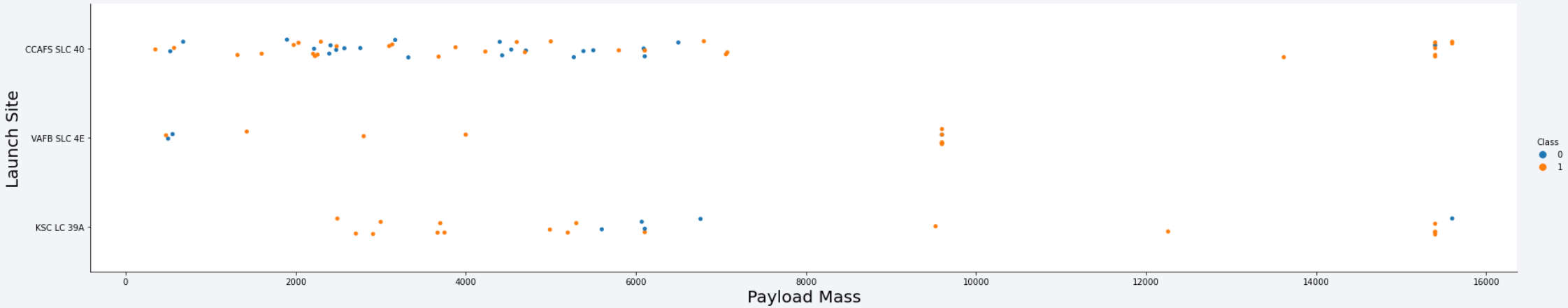

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The chart shows the relationship between flight number and launch site.
- Launch sites are not used equally, and there are gaps in where flights do not occur or occur rarely at some launch sites. This may correspond to maintenance or other “downtime” events at a launch site.

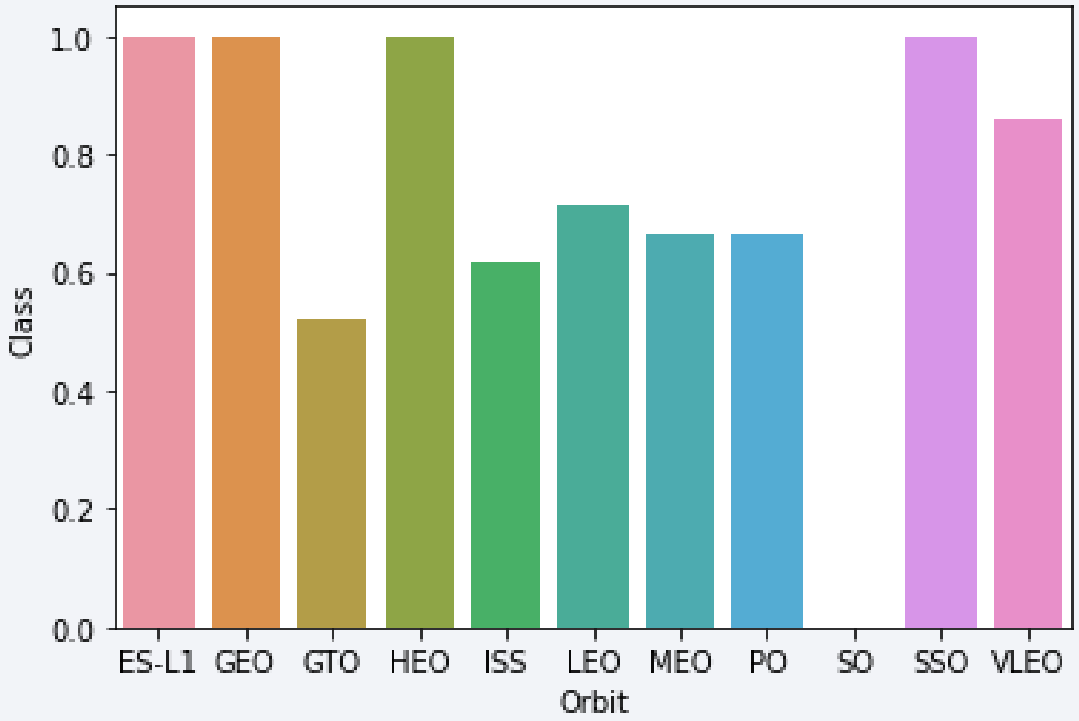
Payload vs. Launch Site



- This chart shows the relationship between payload and launch site.
- CCAFS SLC 40 and KSD LC 30A9 are able to launch all payload masses, while VAFB SLC 4E does not launch payloads over 10,000 kg.

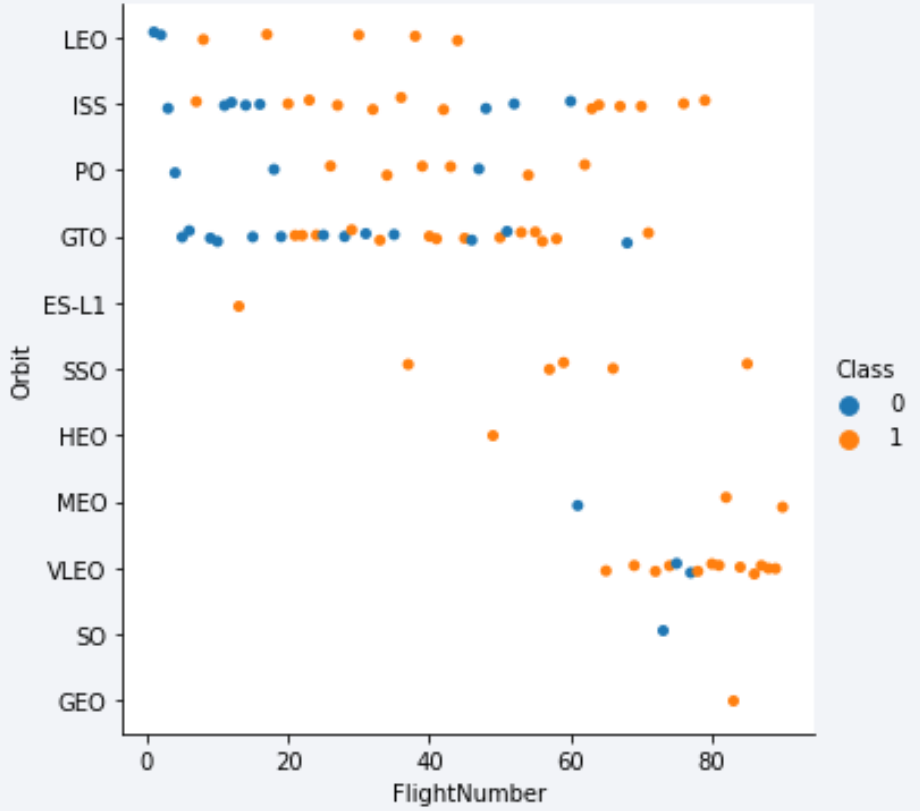
Success Rate vs. Orbit Type

- This chart shows the relationship between Orbit and Success rate.
- ES-LI, GEO, HEO and SSO have high success rates, while SO has never succeeded.



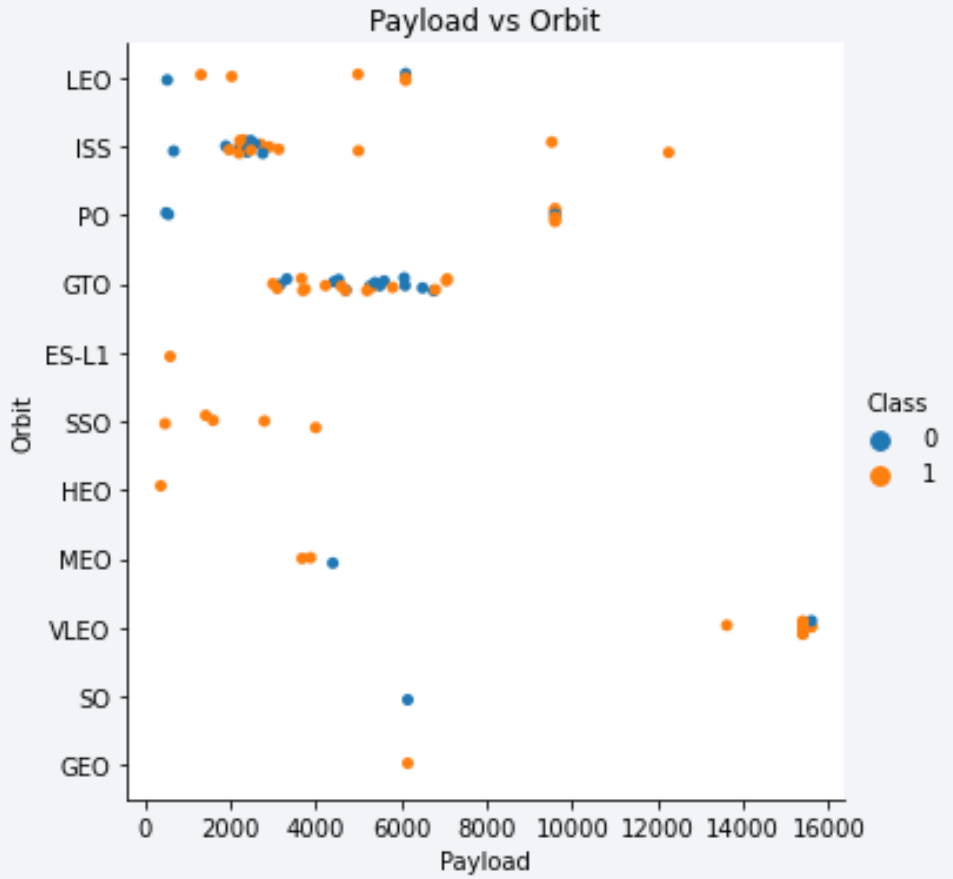
Flight Number vs. Orbit Type

- This chart shows the relationship between Flight Number and Orbit
- Overall, earlier flights have a low success rate regardless of orbit.
- After flight 40, there was a change in the orbits used, with new ones being phased in and old ones being phase out.



Payload vs. Orbit Type

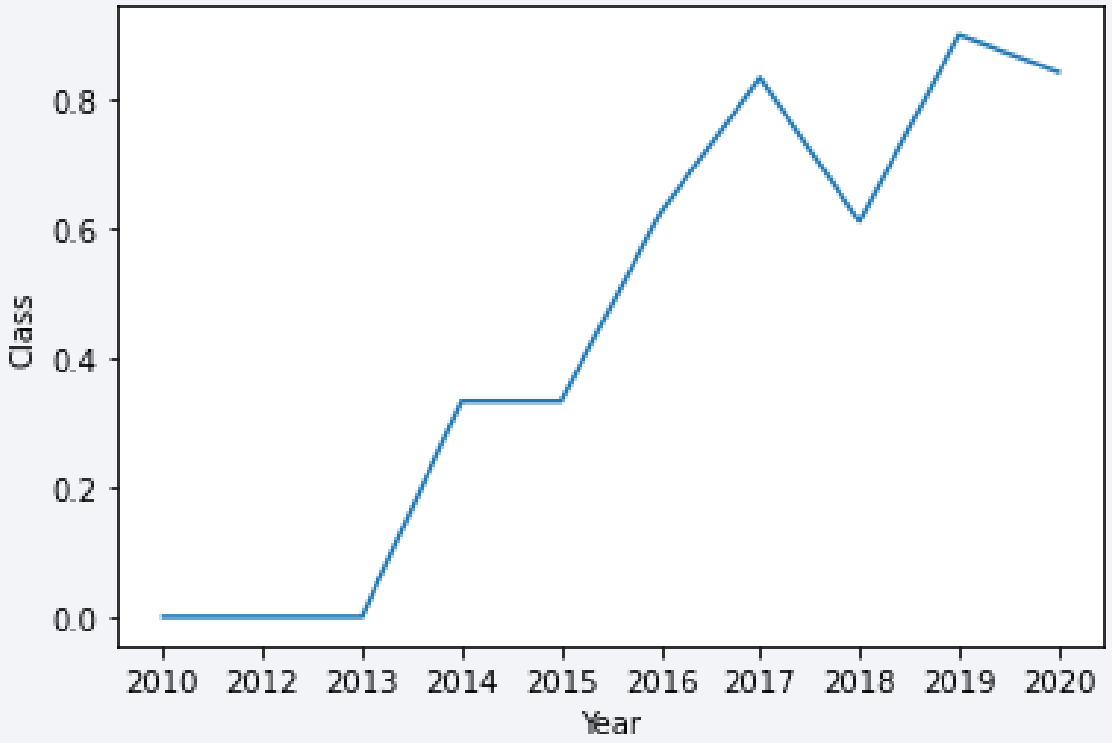
- This chart shows the relationship between Payload and Orbit
- ISS and PO succeed more often for heavier payloads
- For other orbits, payload does not seem to make a measurable difference



Launch Success Yearly Trend

- This chart shows the change in success rate over time

- This chart shows that over time, success rates improve, reaching around 80% by 2020



All Launch Site Names

- This table shows the unique launch sites

- This dataset has four unique launch sites. This table was created with the following SQL query:

```
select distinct
    LAUNCH_SITE
from
    SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- This table shows 5 records where launch sites begin with `CCA`
- The following SQL query was used to create this table:

```
Select
    *
from
    SPACEXTBL
where
    LAUNCH_SITE like 'CCA%' limit 5;
```

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

Total Payload Mass

- The total payload carried by boosters from NASA was 45596 kg

- The following SQL query was used to calculate this sum:

```
Select
    sum(payload_mass__kg_)
from
    SPACEXTBL
Where
    customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928 kg

- The following SQL query was used to calculate this average:

```
Select
        avg(payload_mass__kg_)
from
        SPACEXTBL
where
        booster_version =  'F9 v1.1'
```

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on 2015-12-22

- The following SQL query was used to calculate this average:

```
Select
        DATE
from
        SPACEXTBL
where
        landing__outcome = 'Success (ground pad)' order by DATE asc limit 1
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- This table list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- The following SQL query was used to calculate this table:

```
select distinct
        booster_version
From
        SPACEXTBL
where
        Landing__outcome = 'Success (drone ship)' and
                payload_mass__kg_  between 4001 and 5999
```

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

Total Number of Successful and Failure Mission Outcomes

- There were a total of 100 successful missions and 1 failure mission

- The following SQL query was used to calculate this result:

```
Select
    sum(case when Mission_outcome like '%Success%' then 1 end) as Successes,
    sum(case when Mission_outcome like '%Fail%' then 1 end) as Fails
From
    SPACEXTBL
```

Boosters Carried Maximum Payload

- This table lists the names of the booster which have carried the maximum payload mass
- The following SQL query was used to calculate this table:

```
Select distinct
        booster_version
from
    SPACEXTBL
where
    payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

2015 Launch Records

- This table lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015
- The following SQL query was used to calculate this table:

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

```
select
    landing__outcome, booster_version, launch_site
from
    spacextbl
where
    year(date) = 2015 and landing__outcome = 'Failure (drone ship)'
```

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

- This table ranks 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
- The following SQL query was used to calculate this table:

```
select landing__outcome, count(*) as l_o_counts
from spacextbl
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by count(*) desc
```

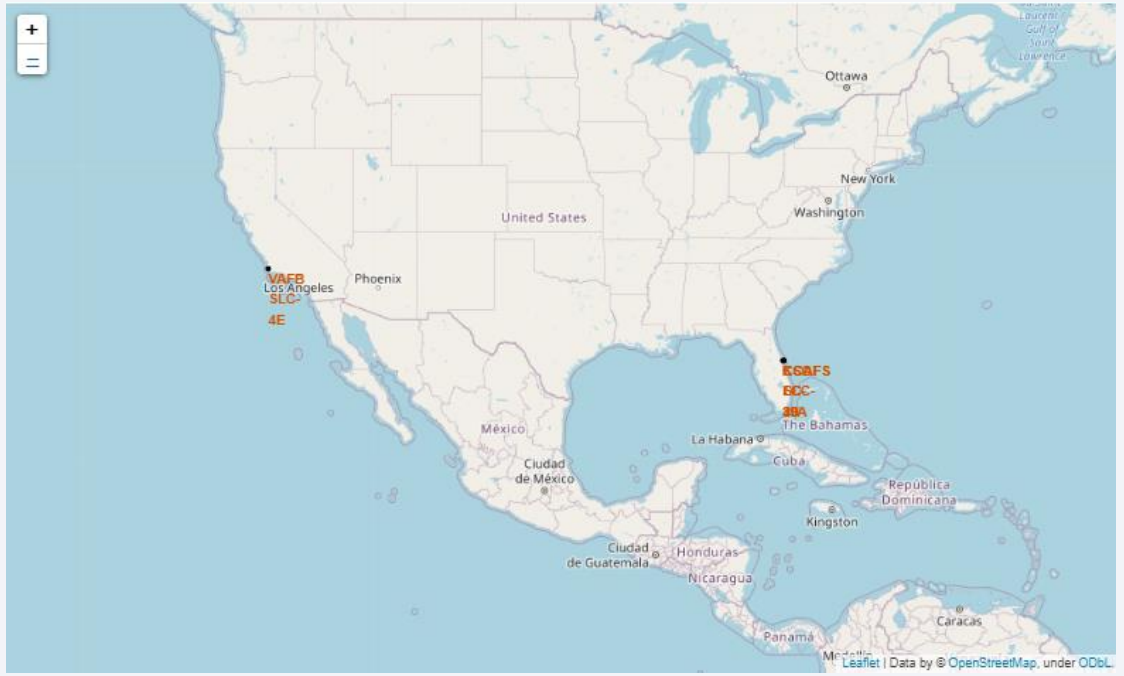
landing__outcome	l_o_counts
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

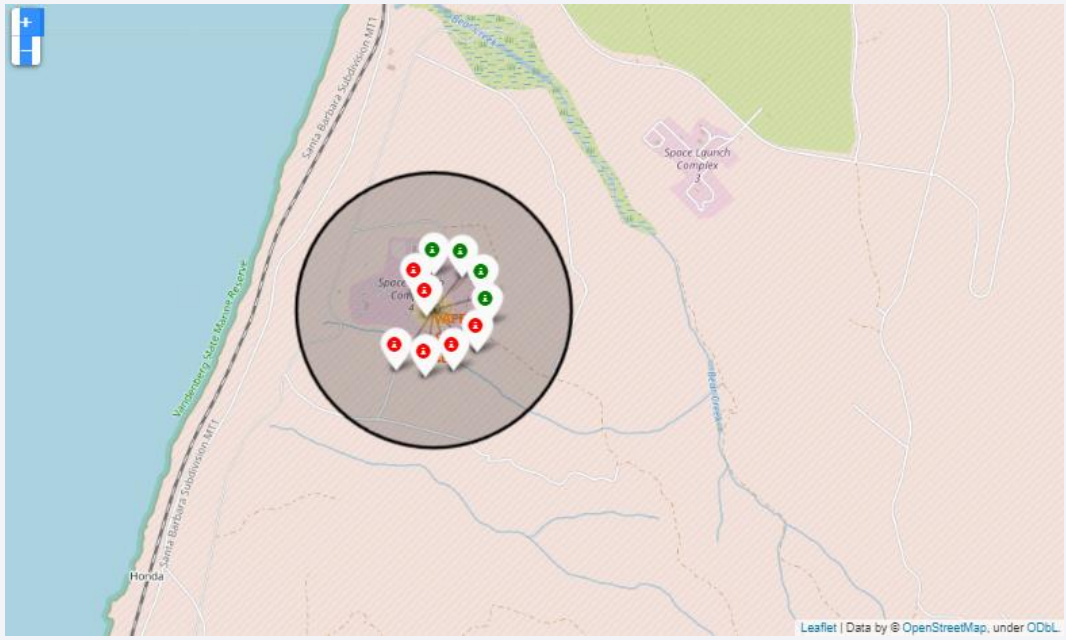
SpaceX Launch Sites

- This map shows all SpaceX launch sites.
  - All sites are in the US, but close to equator on either coast
- One site is on the west coast and three are on the east coast



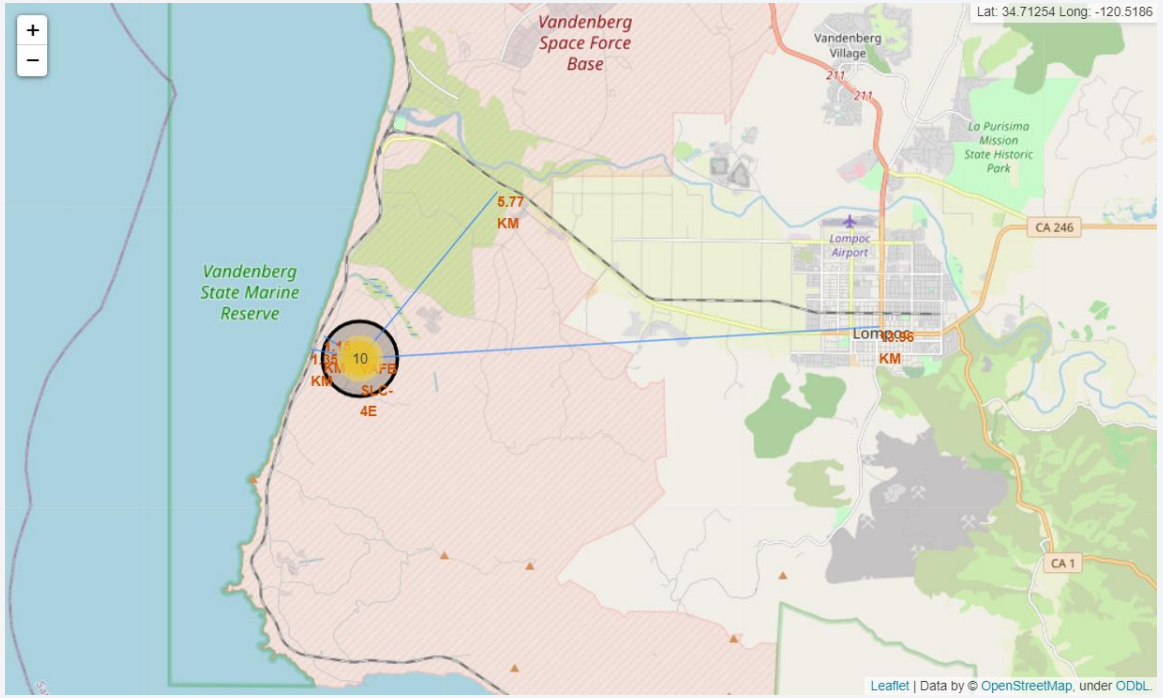
Launch Outcomes at Site VAFB SLC-4E

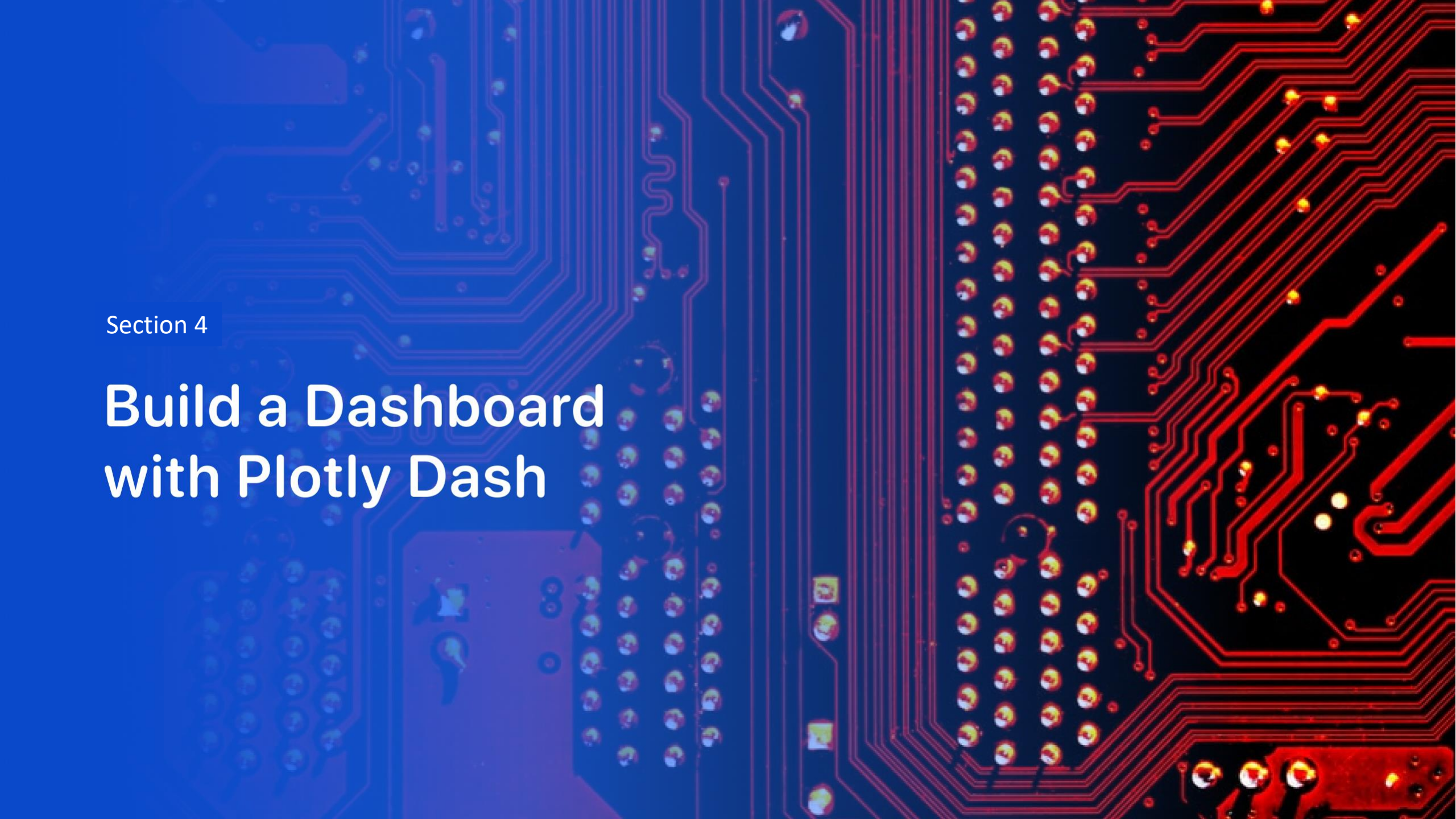
- This map shows the launch outcomes for VAFB SLC-4E
- The shaded-in circle is centered on the launch site
- There is icon test with the site name partially hidden by the red and green flags
- The green flags represent successful outcomes and red flags represent unsuccessful outcomes



Proximities of Map Features from VAFB SLC-4E

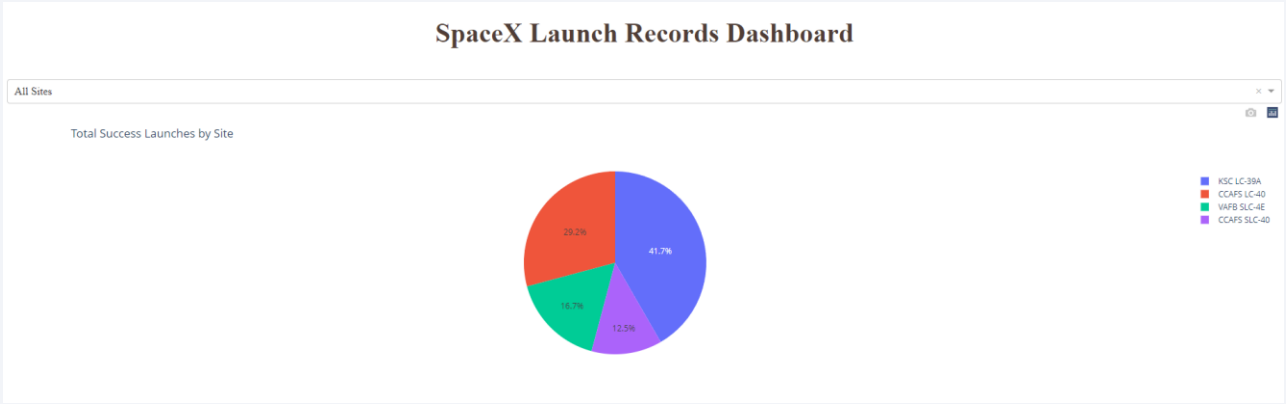
- This map shows the launch outcomes for VAFB SLC-4E with it's nearby map features
- The shaded-in circle is centered on the launch site
- The yellow circle is a marker cluster representing 10 launches that can be zoomed in, to see success/failure
- There is icon test with the site name partially hidden by the yellow circle
- There are blue lines from the launch site to the nearest railway, highway, coastline and city and icon test showing distance




Section 4
Build a Dashboard with Plotly Dash

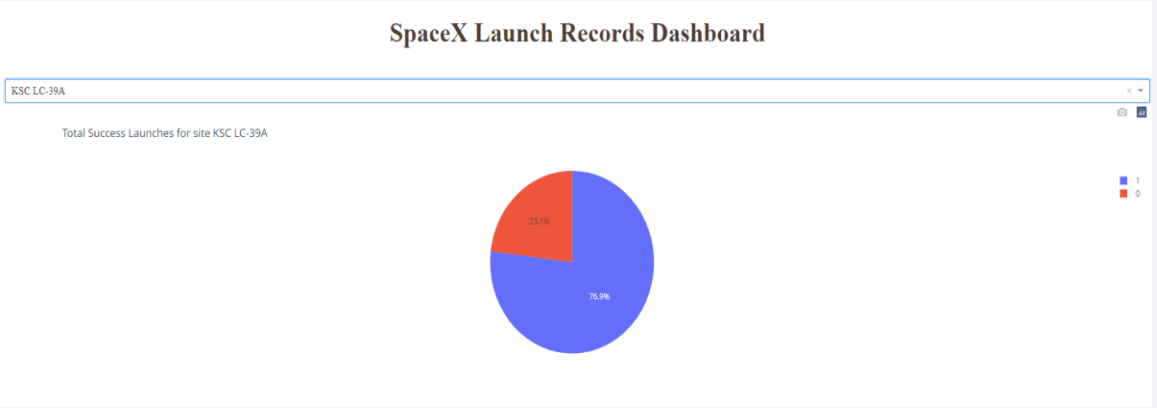
Launch Success for All Sites

- This dashboard shows total successful launches for all sites
- The drop-down menu can be used to select a single site as well
- The pie chart shows the breakdown of successful launches by site
- The legend is on the right side



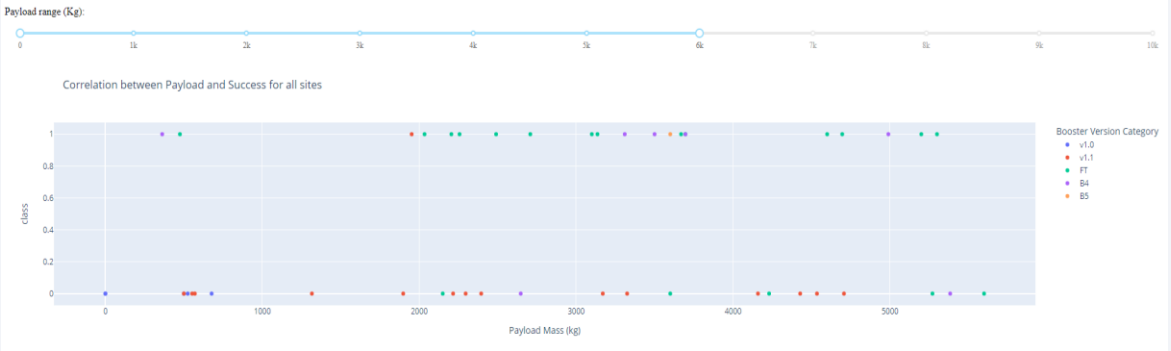
Launch Site with the Highest Success Ratio

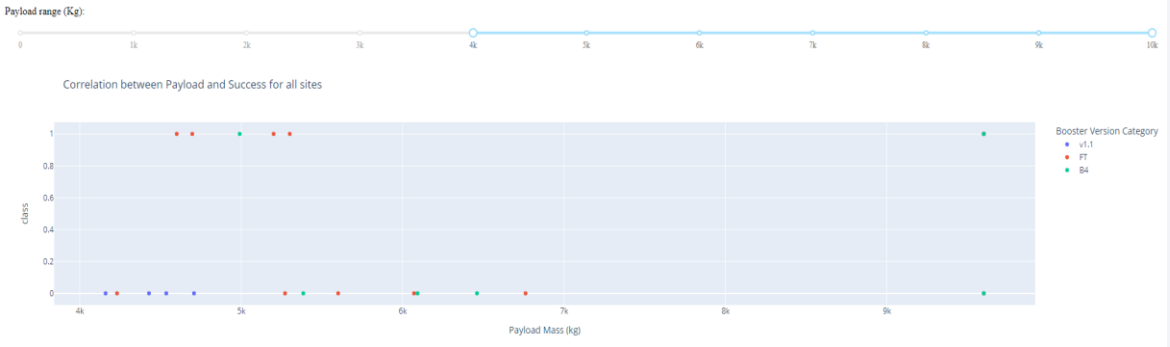
- This chart shows the success ratio of the best performing launch site: KSC LC-39A
- The drop-down menu can be used to select other sites or all sites
- The pie chart shows the breakdown of flight success for this site
- The legend is on the right side – 1 is successful and 0 is failure



Payload vs. Launch Outcome scatter plot for all sites

- These charts show Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider
- There is overlap shown in the 4000 kg – 6000 kg range
- The slider allows selection of a Payload range
- The site selection dropdown is off-screen and cannot be seen
- The scatterplot shows if launches succeeded or failed, by payload mass. The top series of points are success and bottom series are failures



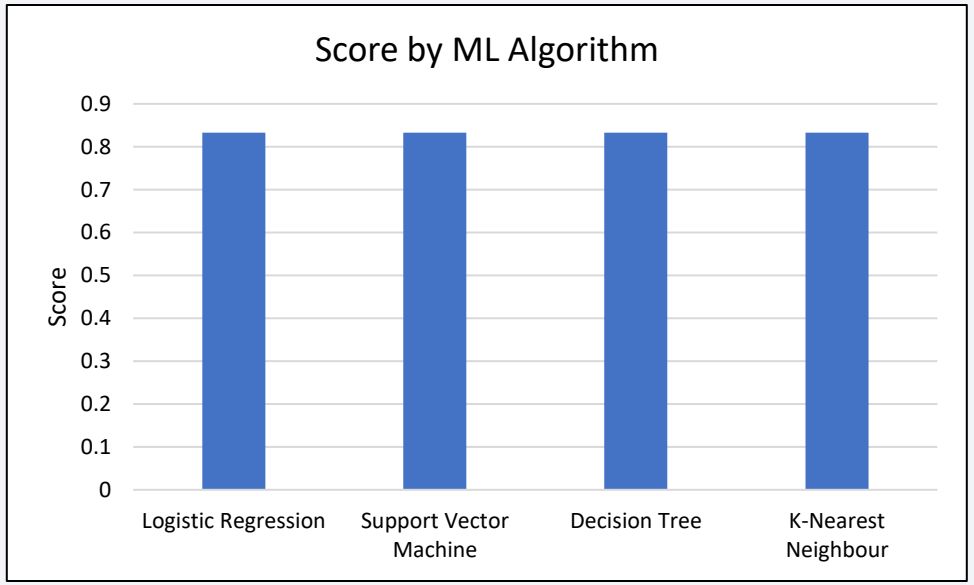


Section 5

Predictive Analysis (Classification)

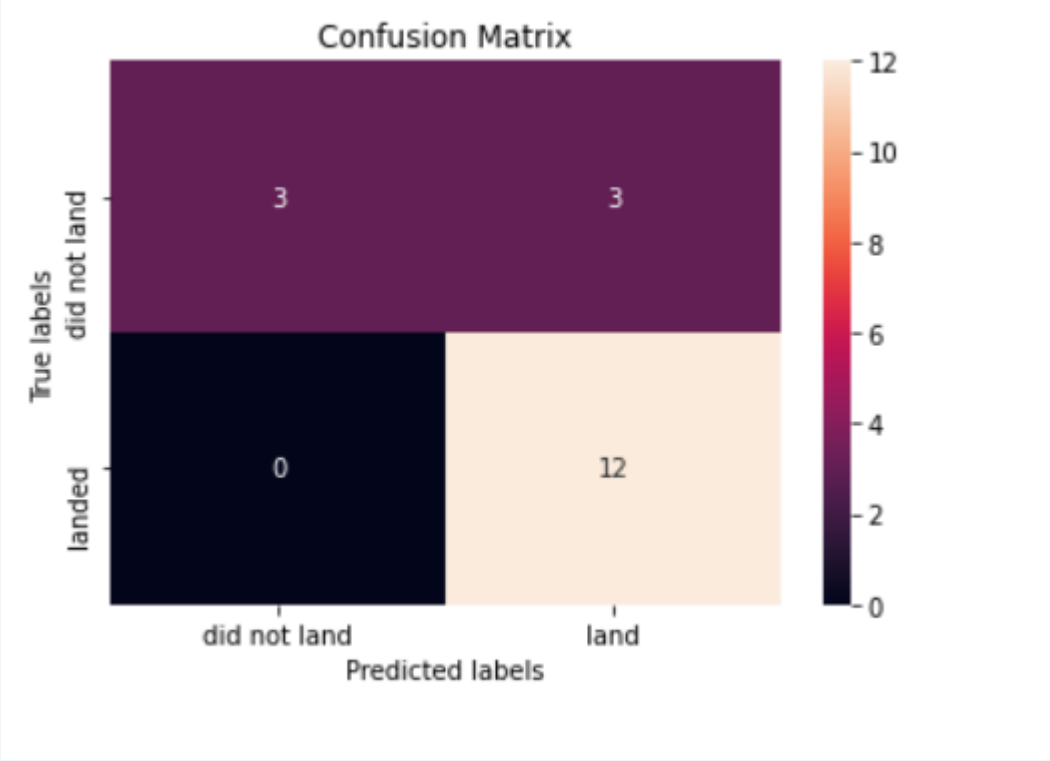
Classification Accuracy

- All methods were equally accurate with the same accuracy score of 0.833
- Decision Tree can be random and perform worse on some occasions



Confusion Matrix

- This confusion matrix shows how well the classifier performed on the test data set – we can see that 15 correcet predictions were made and 3 false positives and no false negatives were predicted.
- All classifiers performed equally well on the test data set and share the same confusion matrix, but decision tree can vary at times.



Conclusions

- The ability re-use the first stage of a Falcon 9 rocket is extremely valuable
- Some combinations of launch parameters were likely to result in better outcomes. For example, ISS and higher payloads were a good combination
- We created a classifier that could accurately predict if a Falcon 9 rocket launch would result in a successful landing of the stage 1 portion
- This knowledge could be used to adjust certain parameters in order to optimize the chance of retrieving the stage 1 rocket

Appendix

- Notebooks at .py files for this project
  - Pulling SpaceX API Data
  - Web Scraping
  - Exploratory Data Analysis
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualizations
  - Dashboard in Dash
  - Interactive Visualization with Folium
  - Machine Learning

Thank you!