IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

YEUNG Tsz Hei
1 November 2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Goal:
  - Predict whether SpaceX will attempt to land a rocket or not.
- Problems:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which Falcon 9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

SpaceX and Falcon 9 Rocket

- Founded by Elon Musk in 2002
- Reduce space transportation costs and
- Enable Mars colonization
- Falcon 9 is a Primary rocket for SpaceX's satellite and cargo missions

Section 1

Methodology

Methodology

Executive Summary

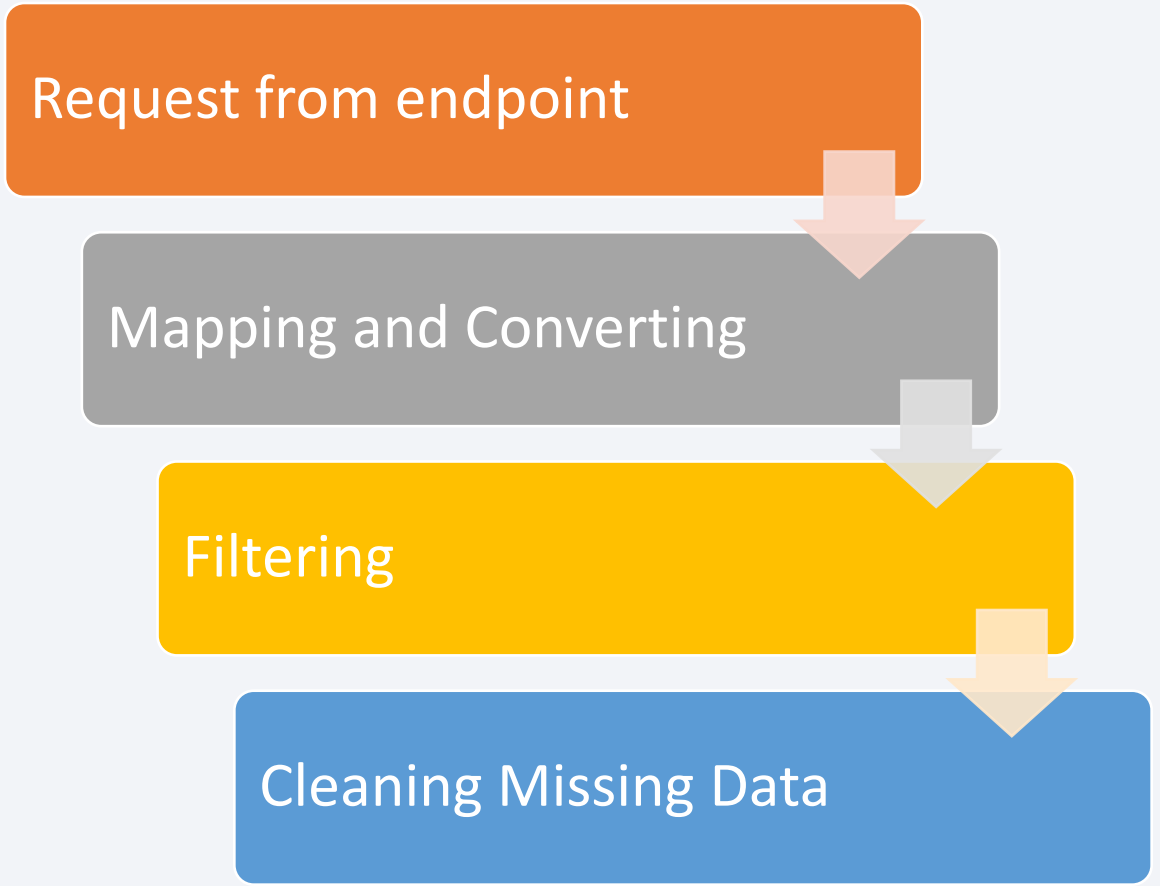
- Data collection methodology:
  - SpaceX Launch data
  - Falcon 9 Launch data
- Perform data wrangling
  - Calculate the frequency and transform datatype
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Preprocessing, Tuning Model and Evaluation

Data Collection

- Collect raw launch data from Space X API endpoints.
- Scraping Falcon 9 launch data from Wiki pages
- Data wrangling for further visualization and analysis

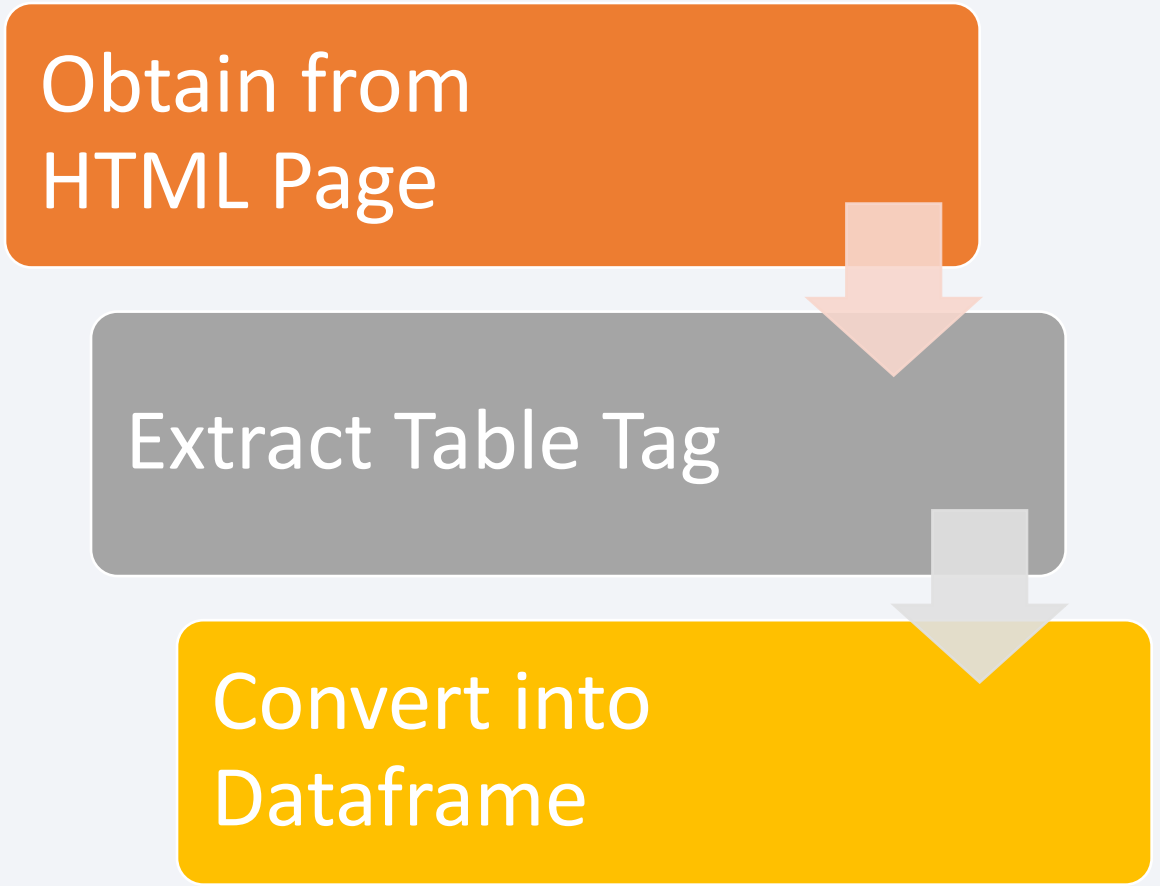
Data Collection – SpaceX API

- Request to the SpaceX API from 4 endpoints: rocket, launchpad, payload, cores
  https://api.spacexdata.com/v4/
- Map and convert into a dataframe .json() and .json_normalize()
- Filter “Falcon 9”
- Clean the missing data
- https://github.com/heihei0314/IBMCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



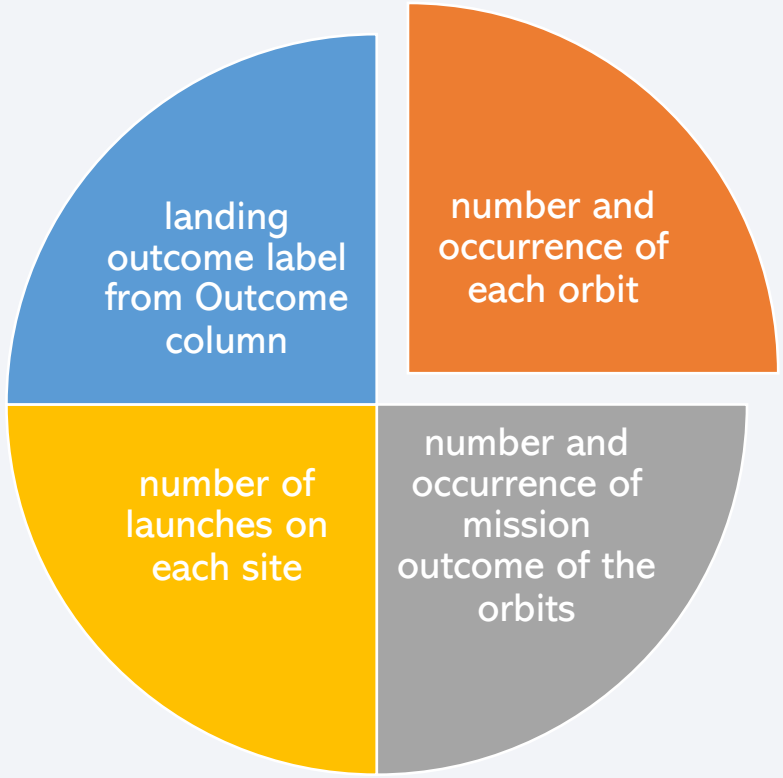
Data Collection - Scraping

- Obtaining Falcon 9 Launch data from Wiki pages
- Parse the table and convert it into a Pandas data frame
- https://github.com/heihei0314/IBMCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- https://github.com/heihei0314/IBMCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter point
  - Payload Mass Vs. Launch Site
- Clearly show characteristic of the VAFB-SLC launch site
  - No rockets launched for heavy payload mass(greater than 10000).
- Line chart
  - Year and Success rate
- Clearly see the improvement of these years
  - No rockets launched for heavy payload mass(greater than 10000).
- https://github.com/heihei0314/IBMCapstone/blob/main/edadataviz.ipynb

EDA with SQL

- Total payload mass carried by boosters launched by NASA (CRS)
  - 45596 Kg
- Average payload mass carried by booster version F9 v1.1
  - 2928.4 Kg
- Total number of successful and failure mission outcomes
  - Most mission are successful
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad))
  - Failure (parachute) is the most failure
- https://github.com/heihei0314/IBMCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

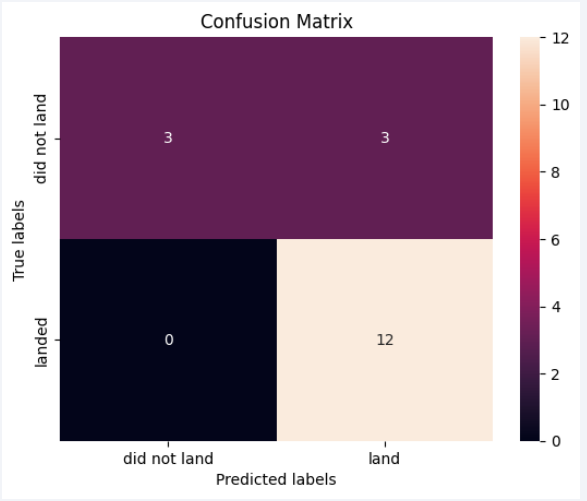
Build an Interactive Map with Folium

- Green marker for successful case and Red markers for failure case
  - It is a good way to spot out the success and failure outcomes in different launch sites
- PolyLine between a launch site to the selected coastline point
  - Observe the distance between the coastline and launch site, further analyze other factors of launch outcomes
- https://github.com/heihei0314/IBMCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

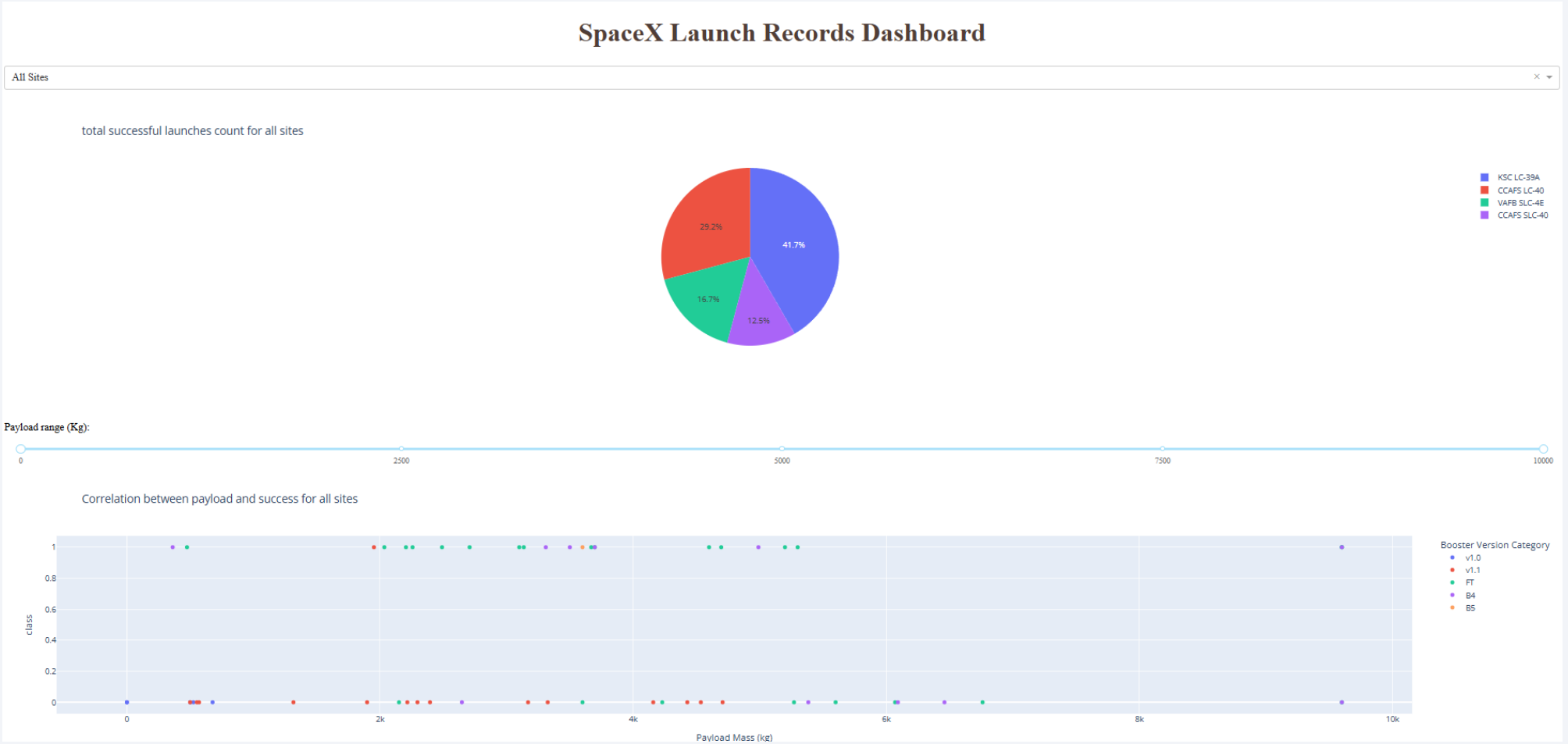
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)



- Split the data into training (80%) and test data (20%).
- GridSearchCV to find the best parameters of each approach
  - Logistic regression
  - support vector machine
  - decision tree classifier
  - k nearest neighbors (KNN)
  - Summarize how you built, evaluated, improved, and found the best performing classification model
- All Score are the same (0.833), but decision tree has the highest accuracy Hence decision tree has the best performance.
- https://github.com/heihei0314/IBMCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

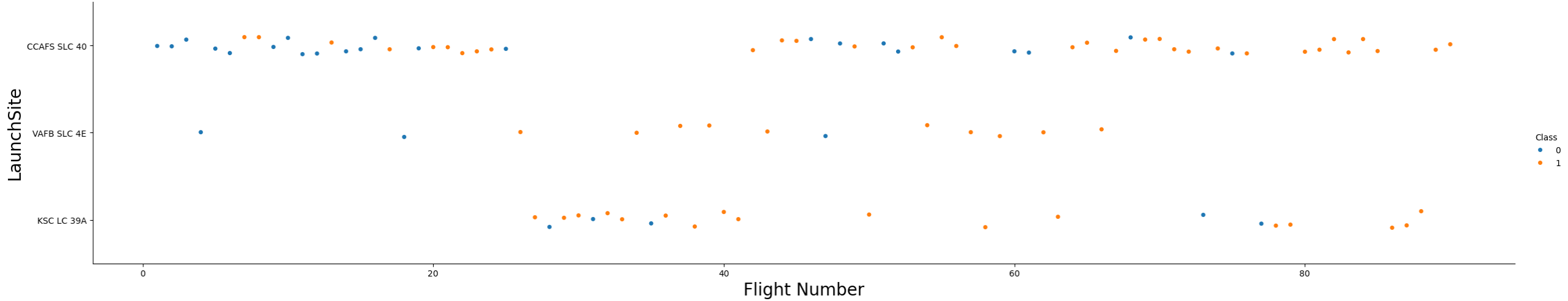

SpaceX Launch Records Dashboard
All Sites
total successful launches count for all sites
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
2500
5000
7500
10000
Correlation between payload and success for all sites
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

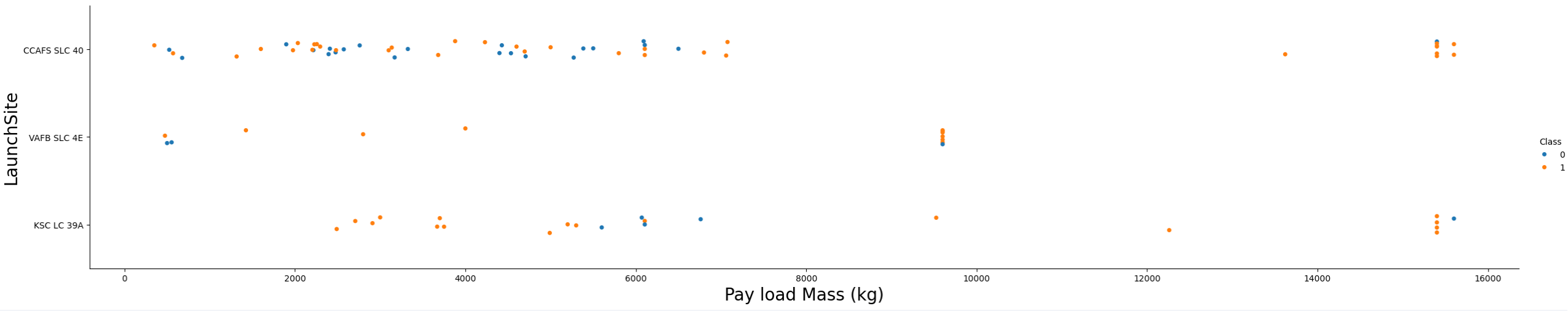
- Scatter plot of Flight Number vs. Launch Site



- Flight Number, indicating the continuous launch attempts
- CCAFS is the most common used launch site from the beginning and currently
- In a middle time, KSC was heavily used
- Now the three launch sites shared the flight, but CCAFS is the major.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

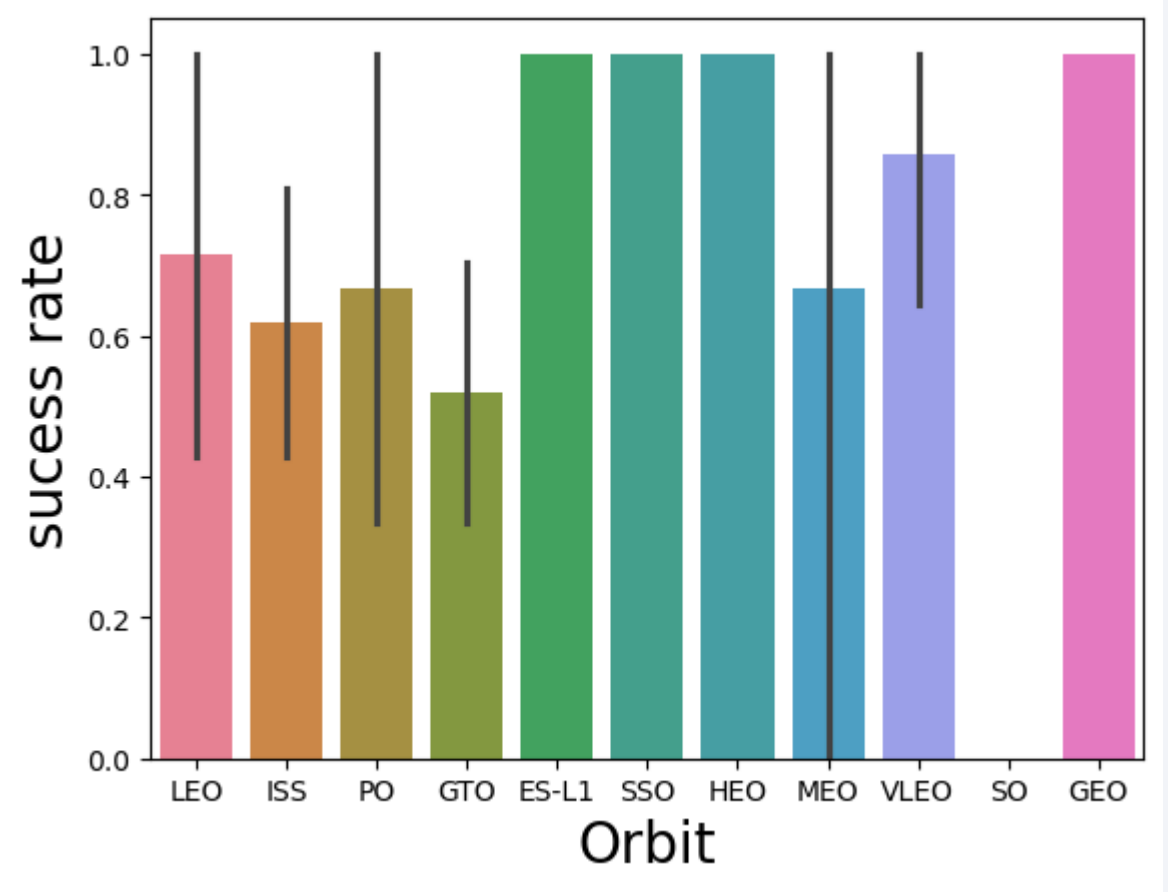


- VAFB-SLC launch site no rockets launched for heavy payload mass(greater than 10000).
- KSC LC launch no rockets launched for heavy payload mass(greater than 2000).
- Very few heavy payload mass launched in other two sites had

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type

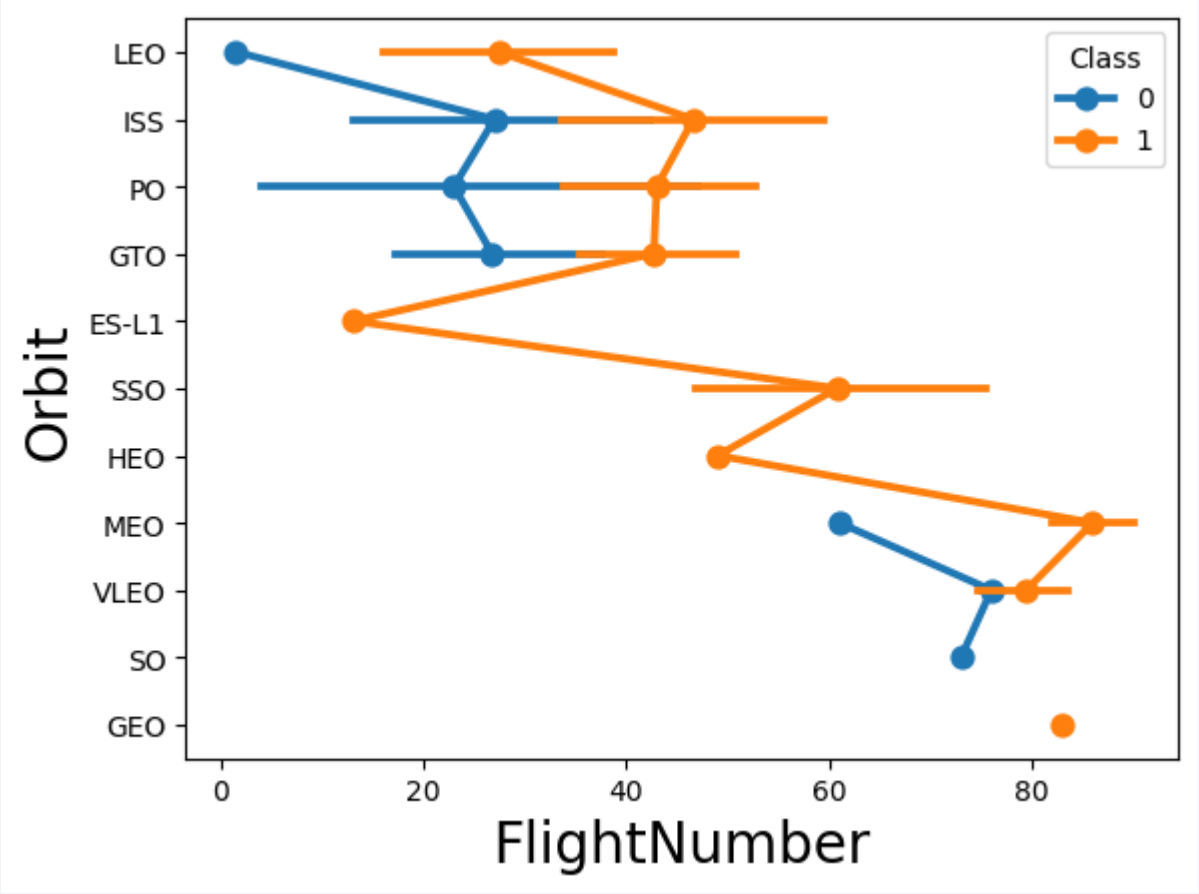
- ES-L1, SSO, HEO and GEO have the highest success rates



Flight Number vs. Orbit Type

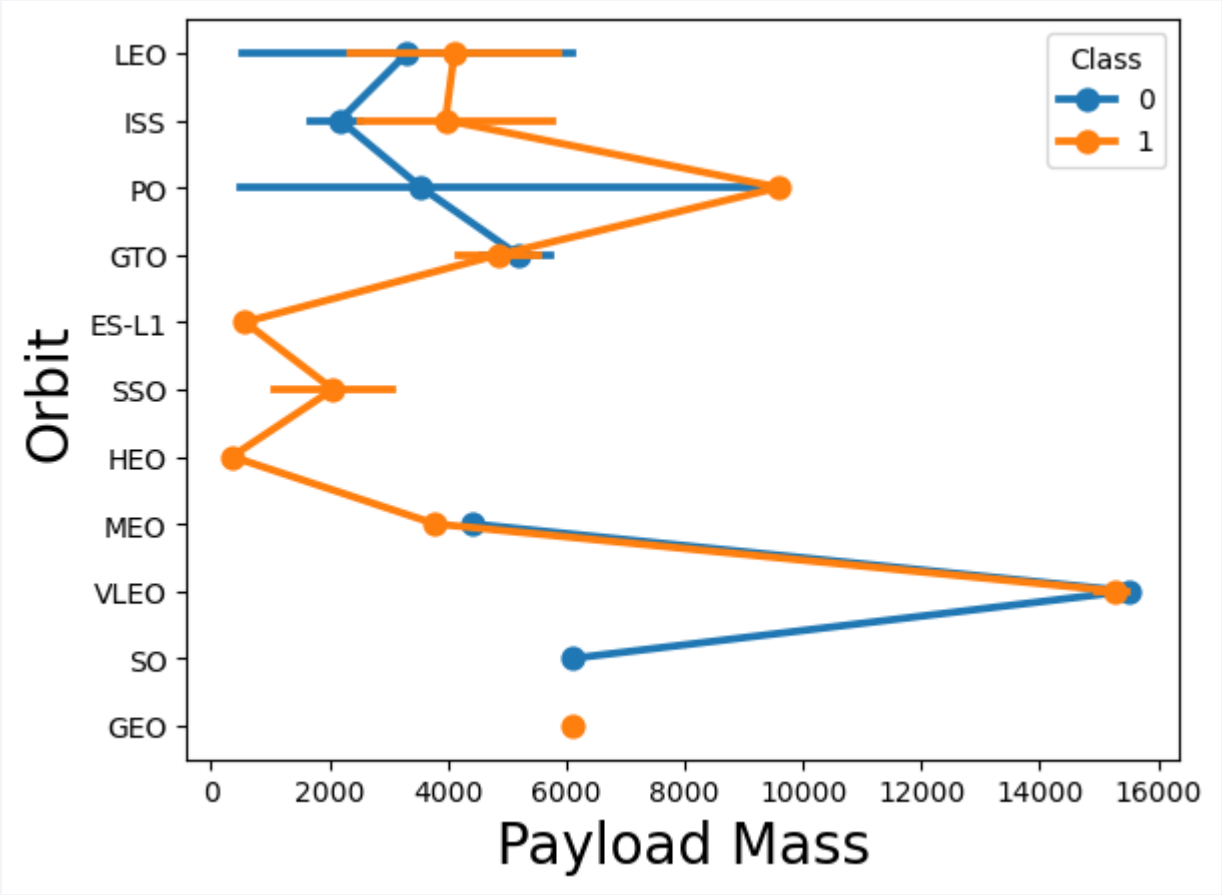
- Scatter point of Flight number vs. Orbit type

- The LEO orbit, success seems to be related to the number of flights.
- The GTO orbit, no relationship between flight number and success.



Payload vs. Orbit Type

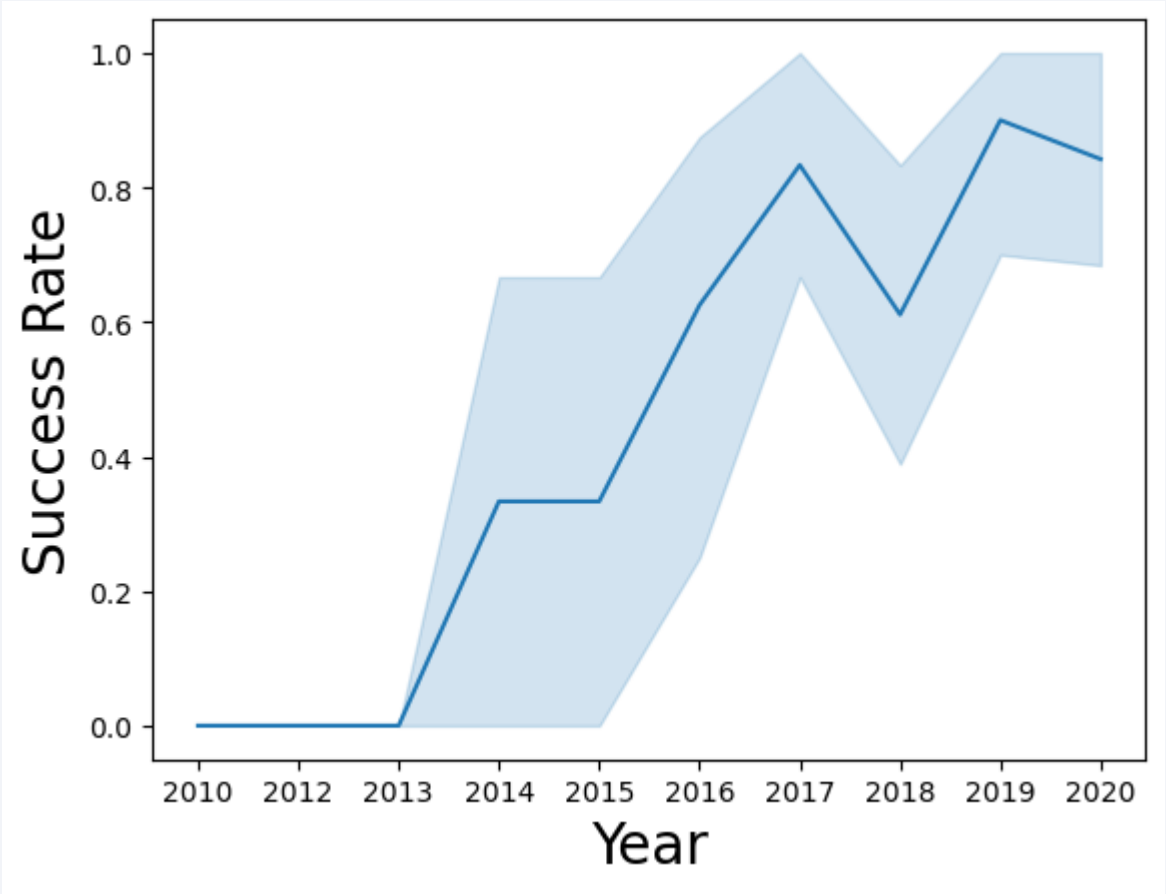
- Scatter point of payload vs. orbit type
- Polar, LEO and ISS, more successful landing with heavy payloads
- The GTO, no relationship between between successful and unsuccessful landings



Launch Success Yearly Trend

- Line chart of yearly average success rate

- Successful rate kept increasing since 2013



All Launch Site Names

- Find the names of the unique launch sites
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Present your query result with a short explanation here
  - %sql SELECT Launch_Site FROM SPACEXTABLE Group by Launch_Site
  - Using Group by to find unique value

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - CCAFS LC-40
  - CCAFS LC-40
  - CCAFS LC-40
  - CCAFS LC-40
  - CCAFS LC-40
- Present your query result with a short explanation here
  - %sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
  - Use like and % to filter string
  - Use LIMIT to find exact number of rows

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - 45596
- Present your query result with a short explanation here
  - %sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'
  - Use SUM() to find total

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - 2928.4
- Present your query result with a short explanation here
  - %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1’
  - Use AVG() to find the total

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - 2015-12-22
- Present your query result with a short explanation here
  - %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)'
  - Use MIN() to find the earliest date

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - JCSAT-14
  - JCSAT-16
  - SES-10
  - SES-11 / EchoStar 105
- Present your query result with a short explanation here
  - %sql SELECT Payload FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- Present your query result with a short explanation here
  - %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
  - Use COUNT() to find the frequency

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here
  - %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
  - Use sub query to find complex competition

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

substr(Date, 6,2)	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

- Present your query result with a short explanation here
  - %sql SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome like 'Failure (drone ship)'
  - Use substr() to extract the year

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

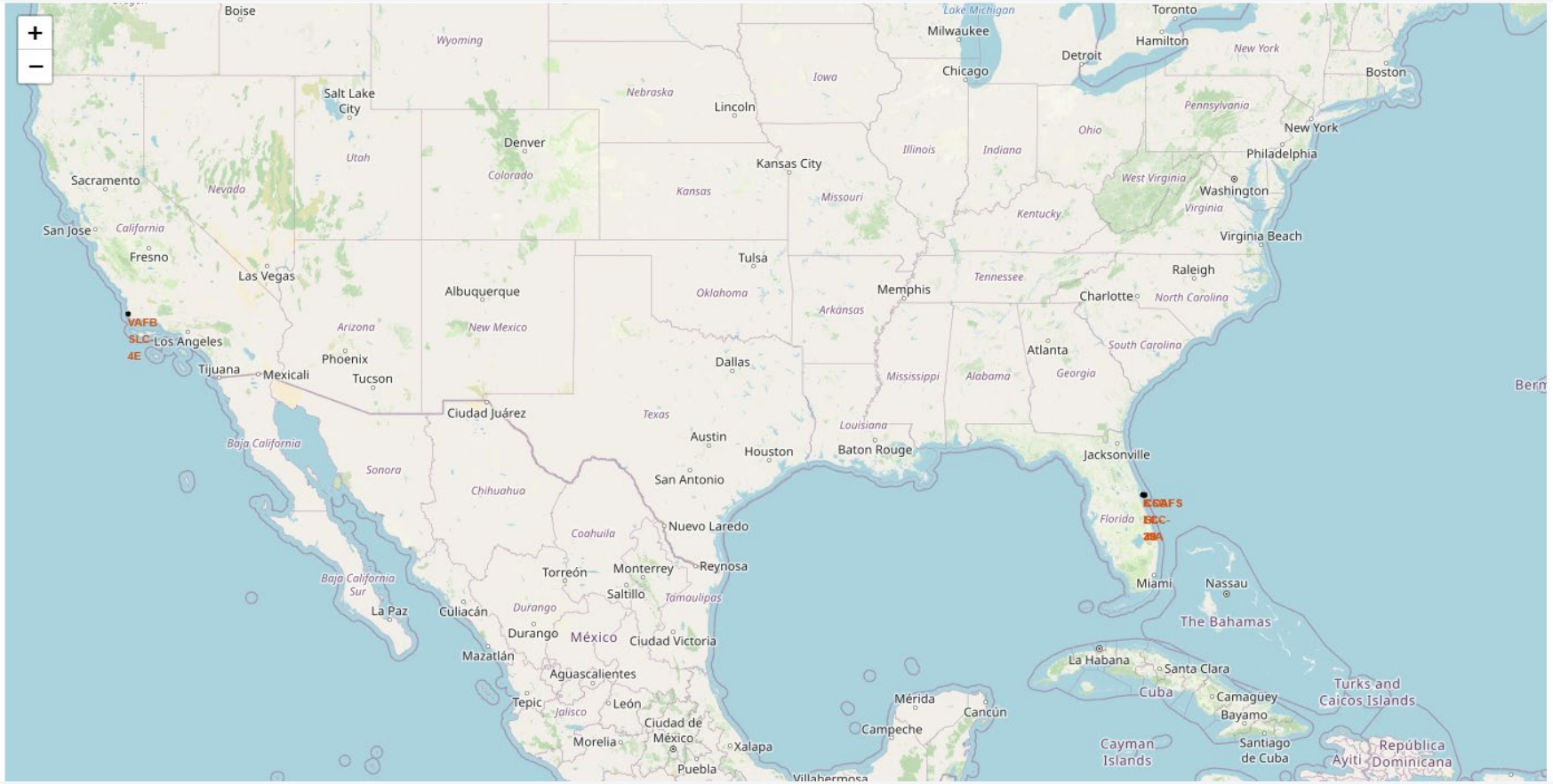
RANK	Landing_Outcome	COUNT(Landing_Outcome)
1	Success (drone ship)	12
1	No attempt	12
3	Success (ground pad)	8
4	Failure (drone ship)	5
5	Controlled (ocean)	4
6	Uncontrolled (ocean)	2
7	Precluded (drone ship)	1

- Present your query result with a short explanation here
  - %sql SELECT RANK () OVER (ORDER BY COUNT(Landing_Outcome) DESC) as RANK, Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date between '2010/06/04' and '2017/03/20' GROUP BY Landing_Outcome
  - Use RANK() to order and extract


Section 3
Launch Sites
Proximities Analysis

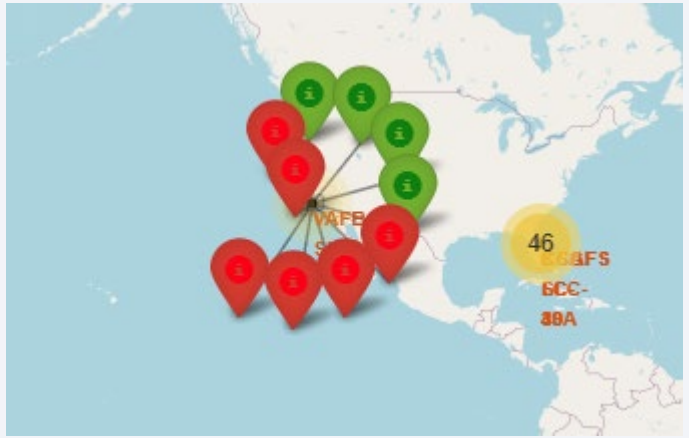
All launch sites location

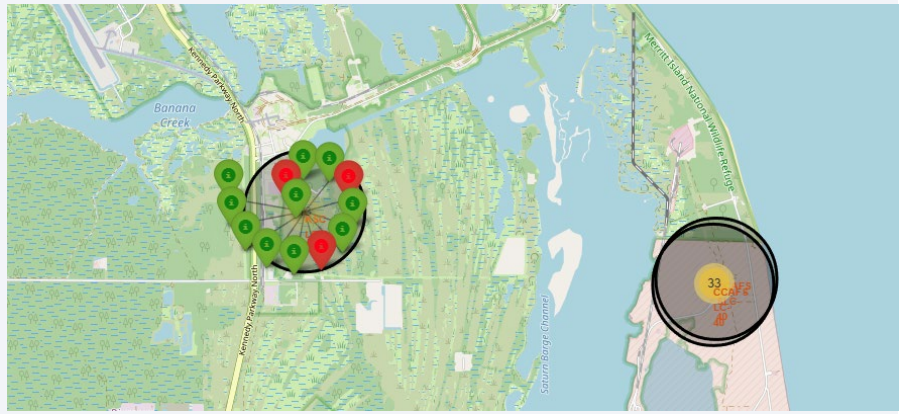
- All launch sites are near to coastal and on the south of US



Success and Failed launches for Each Site

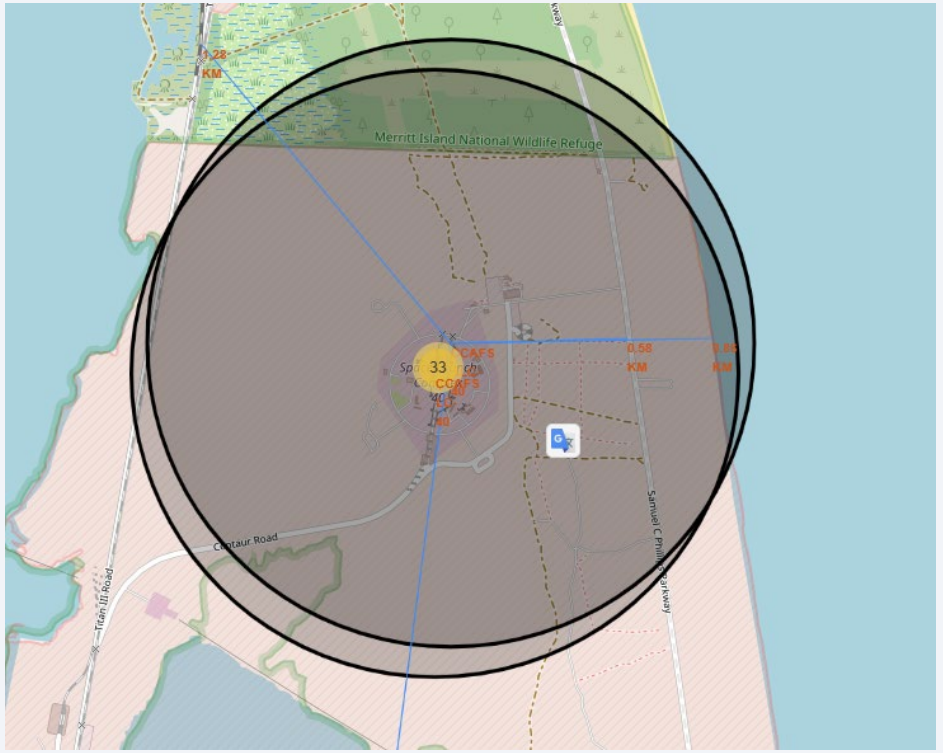
- KSC LC have relatively high success rates

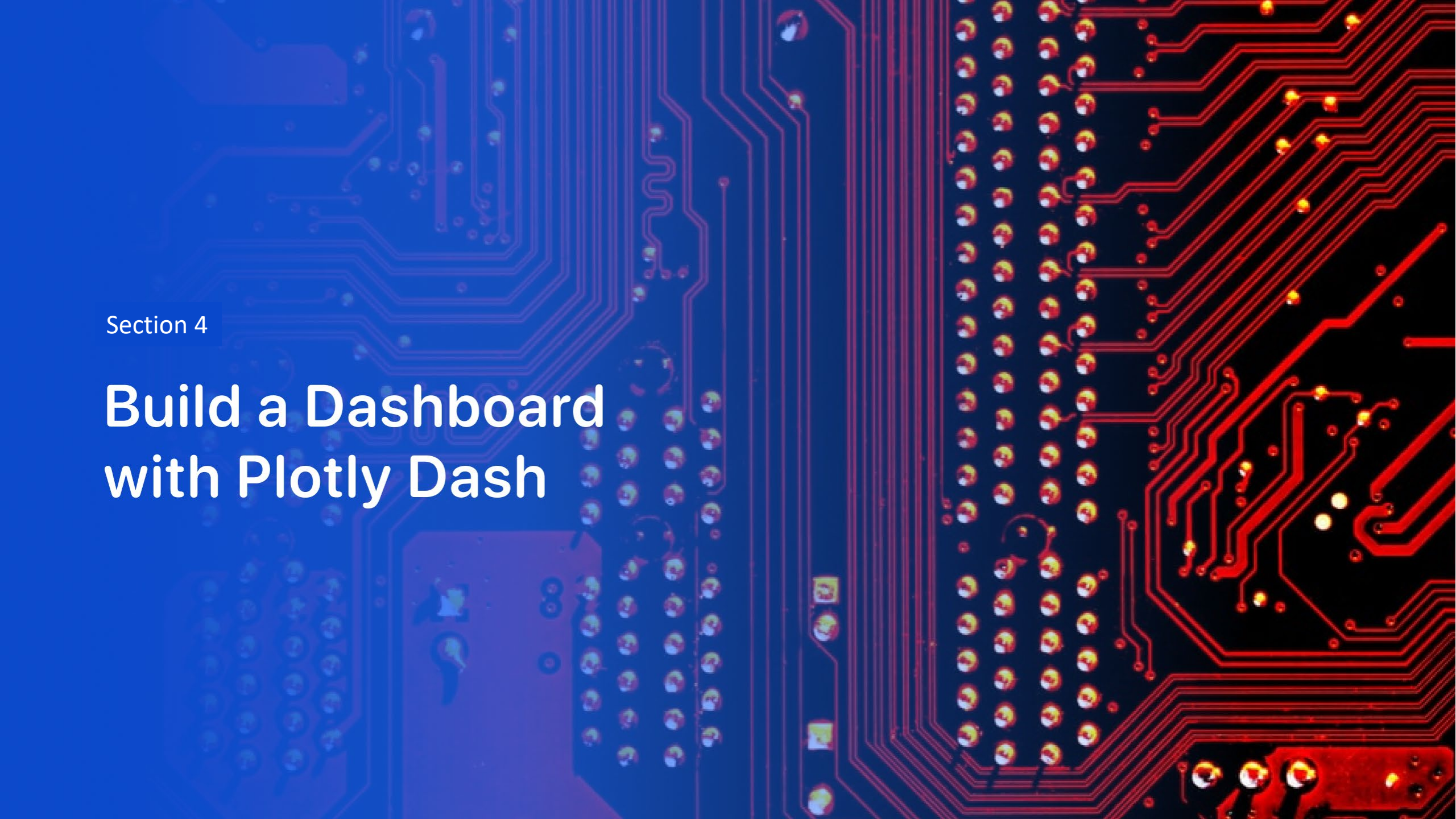




Distances between Launch Sites and Proximities

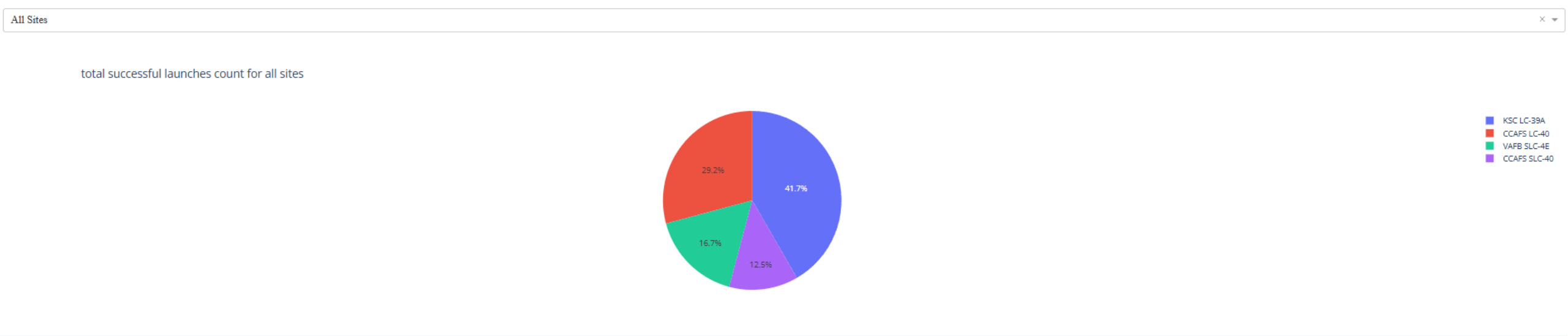
- Launch sites are close to coastline and highway
- But far from airport




Section 4
Build a Dashboard with Plotly Dash

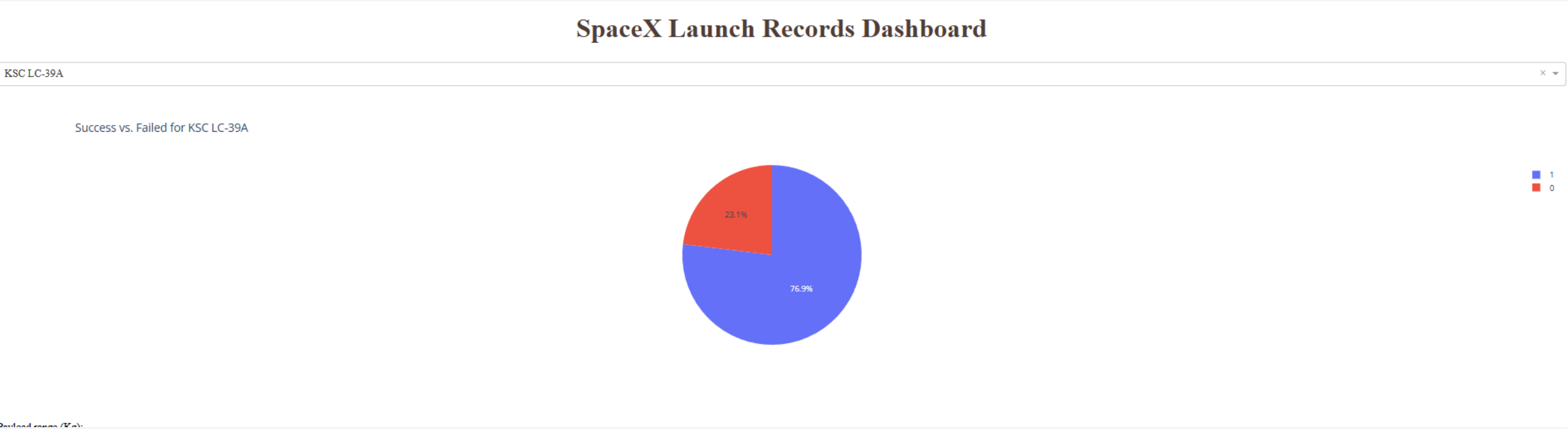
Total Successful Launches Count for All Sites

- Most success is KSC LC



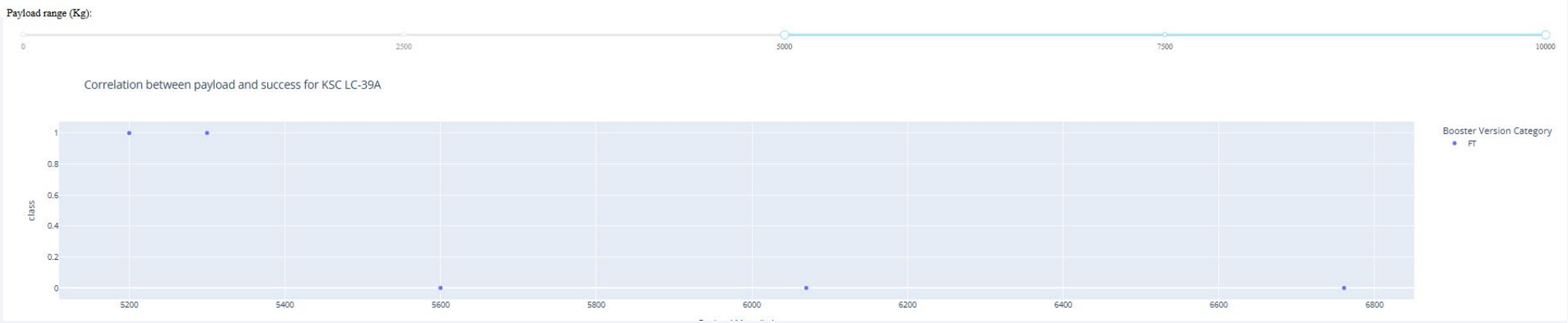
Success Ratio for KSC LC

- Over 75% success rate for KSC LC



Payload vs. Launch Outcome For All Sites

- Light payload has high successful rate
- The heavy payload had low successful rate and only FT has heavy payload

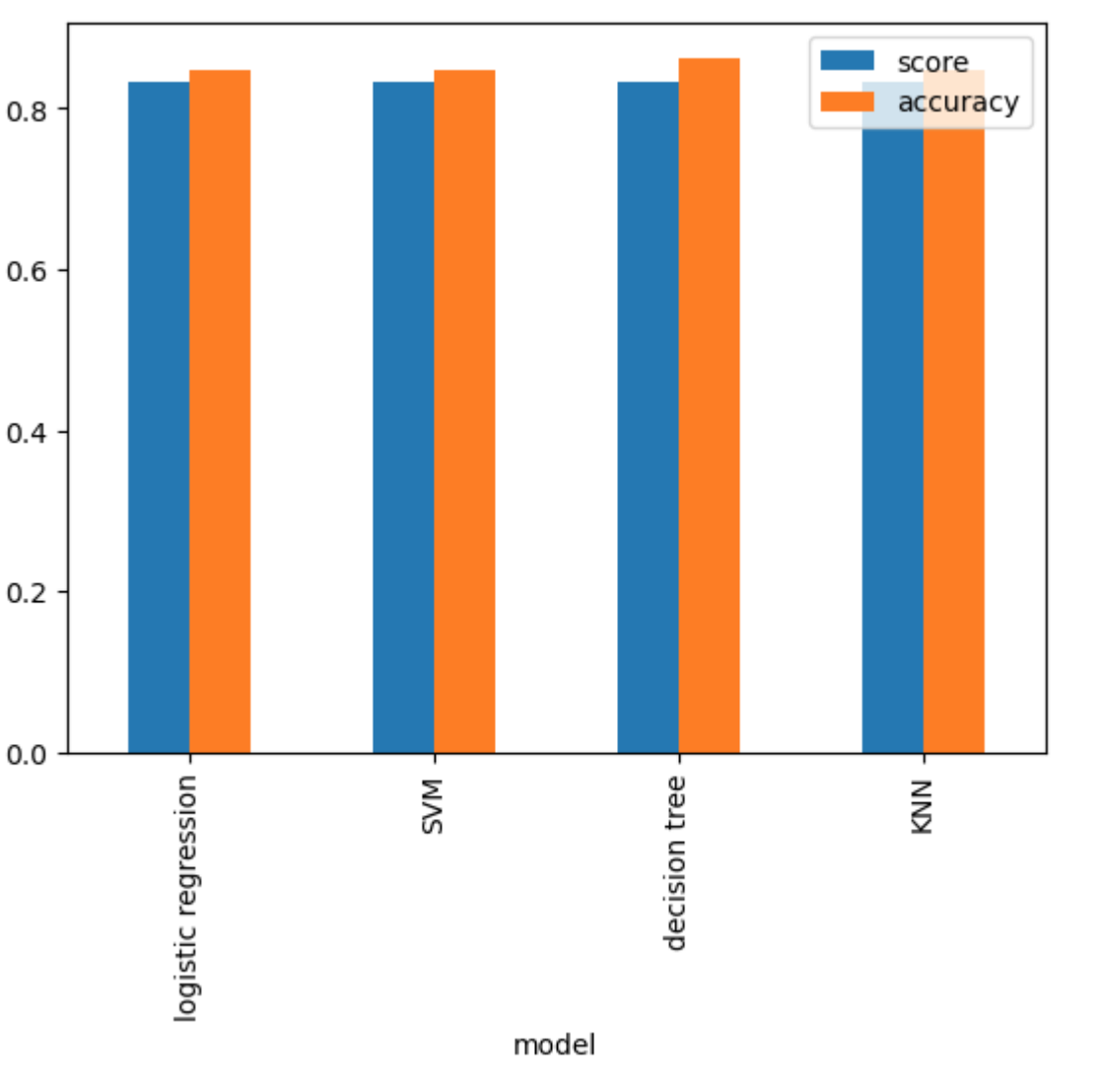


Section 5

Predictive Analysis (Classification)

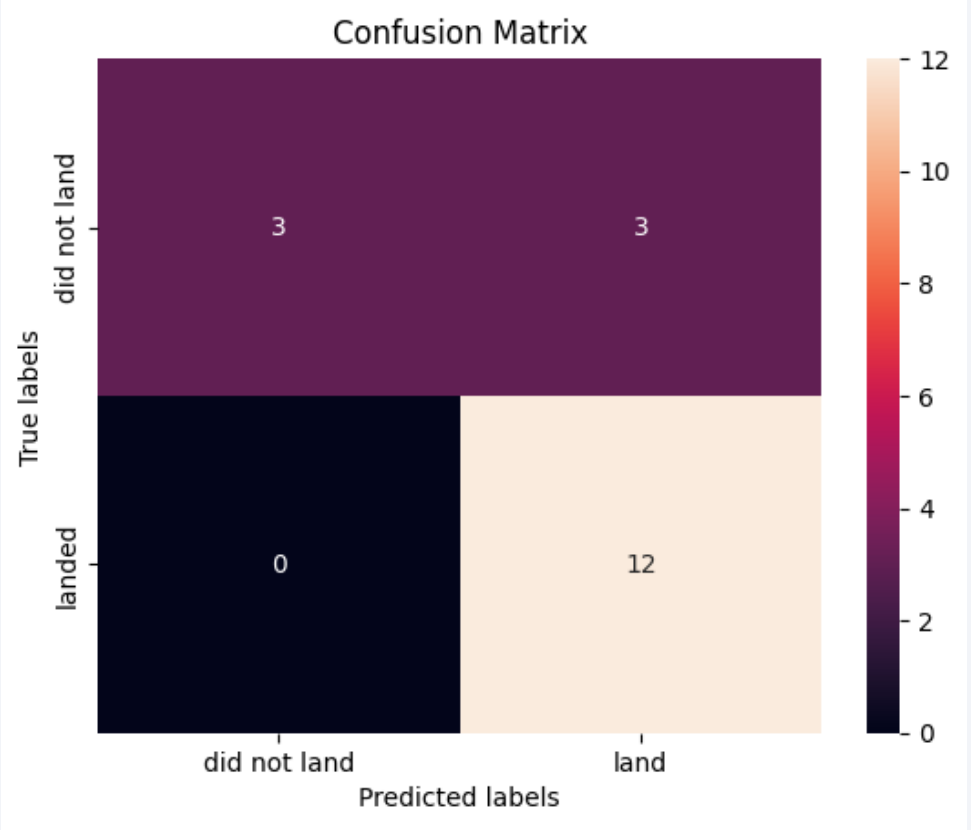
Classification Accuracy

- Decision tree has the highest accuracy



Confusion Matrix

- All model shared the same matrix



Conclusions

- The CCAFS has the largest successful launches
- The KSC LC has the highest launch success rate
- Heavy payload range(s) (5000 or up) has the low launch success rate
- Falcon 9 Booster version FT has the lowest launch success rate

Thank you!