IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Lundberg
2023-04-10

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
    - API
    - Web Scraping
  - EDA
    - SQL
    - Viz
- Summary of all results

Introduction

- Project background and context
  - Space travel is expensive
  - Can SpaceX predict costs using likelihood of successful launches
- Problems you want to find answers
  - impact of independent variables
  - probability of success and expected value

Section 1

Methodology

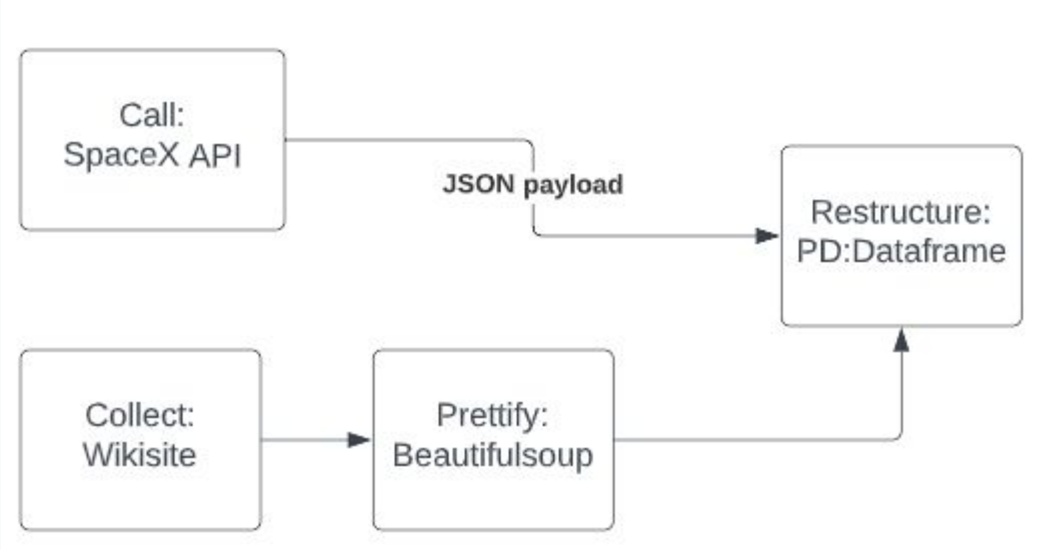
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API & Web scraping: Wikipedia
- Perform data wrangling
  - Aggregations, probabilities, classifications and labeling
- 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 using SpaceX API. The payload was then restructured into a Dataframe
- Data was also collected by scraping Wikipedia, this data was treated using Beautifulsoup and later restructured into a Dataframe



Data Collection – SpaceX API

```
[6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

[7]: response = requests.get(spacex_url)
```

https://github.com/Thoracalis/Blah/blob/master/jupyter-labs-spacex-data-collection-api%20(2).ipynb

Data Collection - Scraping

- Name URL
- Request URL
- Soupifyparse
- Place in DF
- Handle missing values

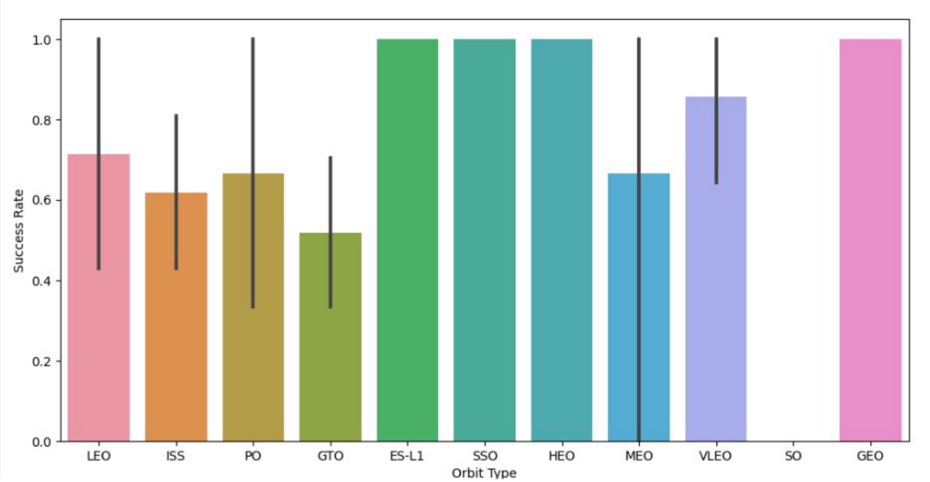
https://github.com/Thoracalis/Blah/blob/master/jupyter-labs-webscraping.ipynb

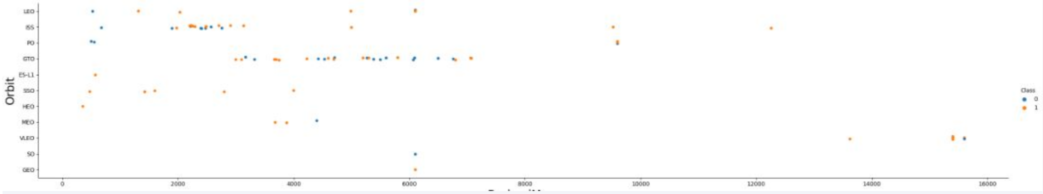
Data Wrangling

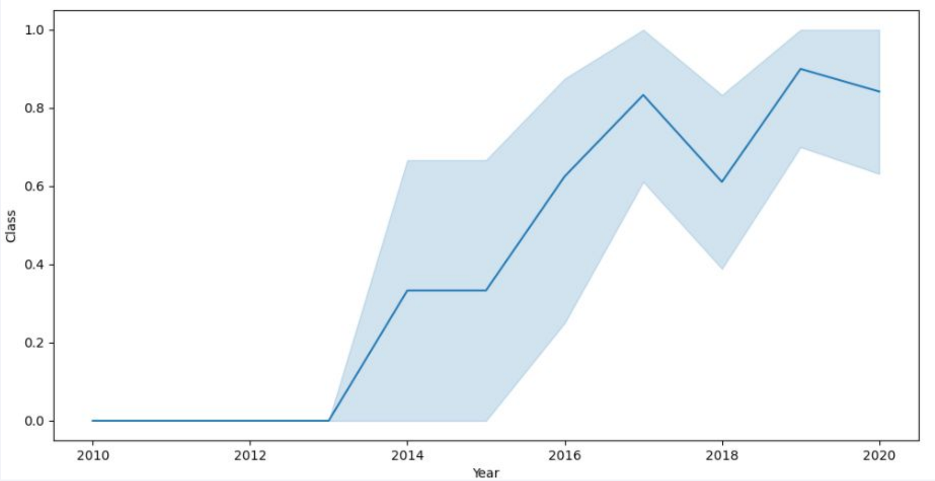
- Aggregation
- Labeling
- Classification

https://github.com/Thoracalis/Blah/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization







Scatter : x, y relationship

Bar : group comparison

Line - with interval : trend

https://github.com/Thoracalis/Blah/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Display Uniqure Launch Sites
- Sample size 5 where at Site %CCA
- Total Payload Carried by NASA CRS
- Avg payload carried by booster version F9 v1.1
- First successful landing on ground pad
- Boosters with successful droneship at payload between 4 K and 6 K
- Total number of Successful and Failed missions
- Boosters which have carried max Payload
- Failed Drone ship landings in year 2015
- Rank of successful landings between 2010-06-04 and 2017-03-20

https://github.com/Thoracalis/Blah/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added sites, distances and clusters with details
- This was done to gain spatial awareness
-

https://github.com/Thoracalis/Blah/blob/master/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

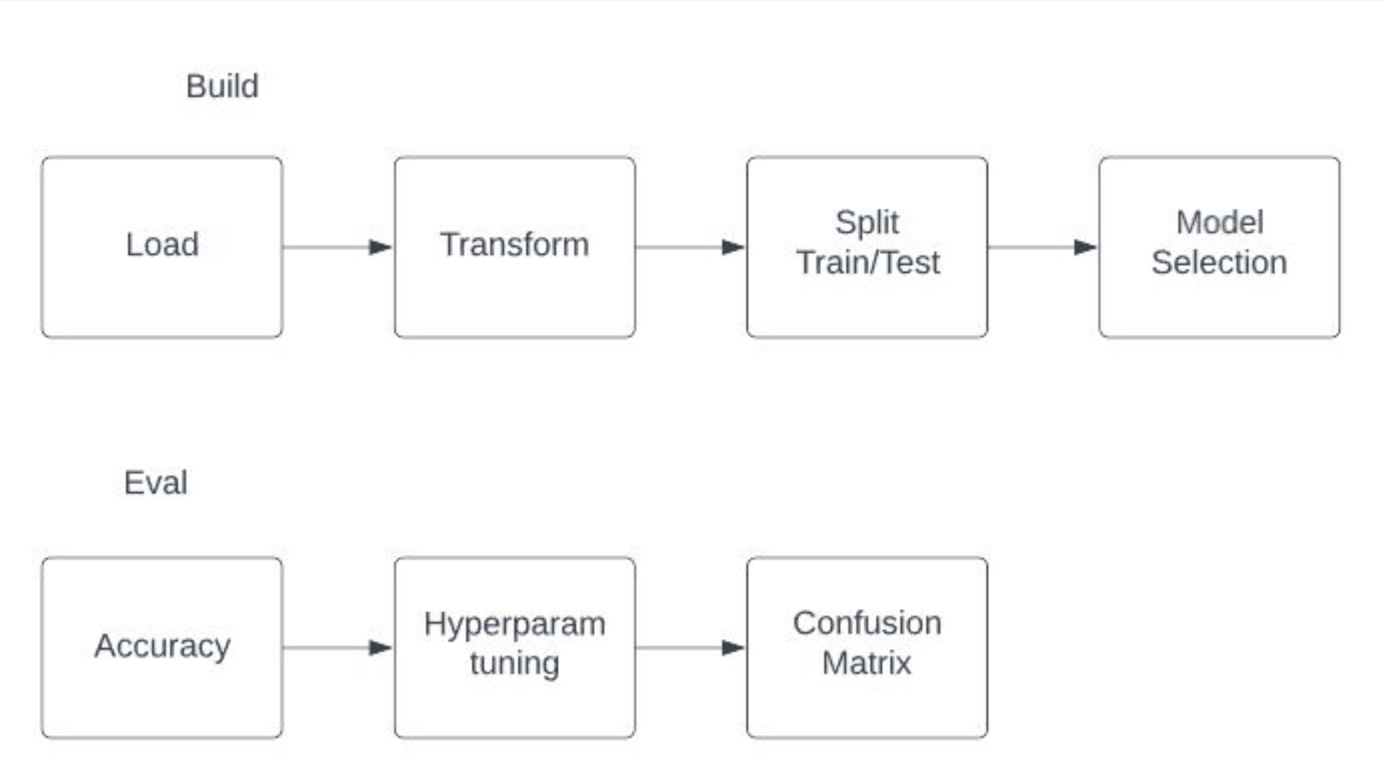
- Pie charts: show launches by site(s)
- Scatter: Relationship between variables

https://github.com/Thoracalis/Blah/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

Got data, normalized and transformed.

Ran through models, evals with comparison and confusion matrix



https://github.com/Thoracalis/Blah/blob/master/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results

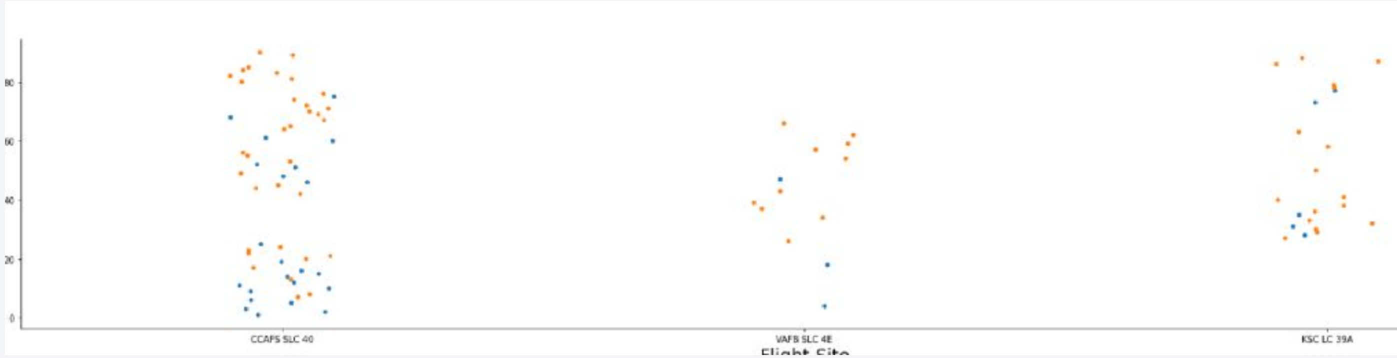
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

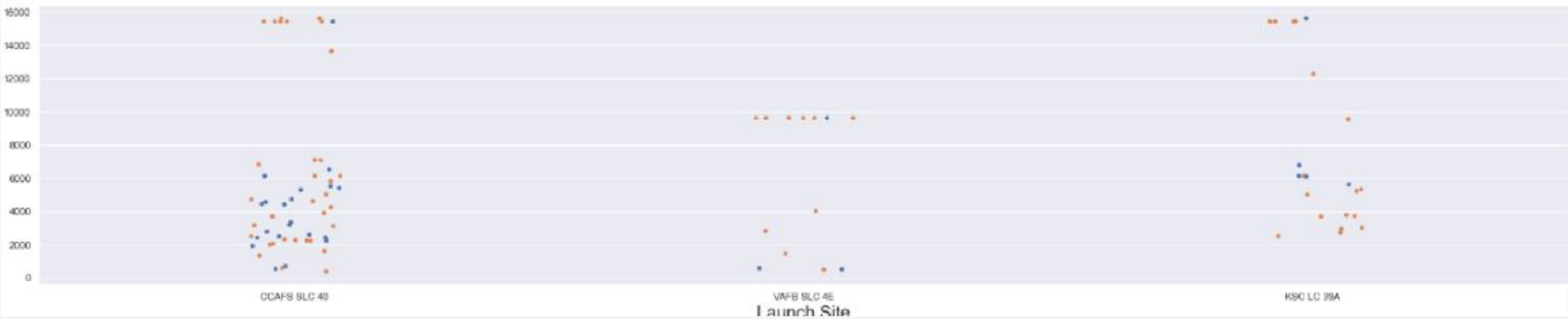
Flight Number vs. Launch Site

- More flights correlates with higher success



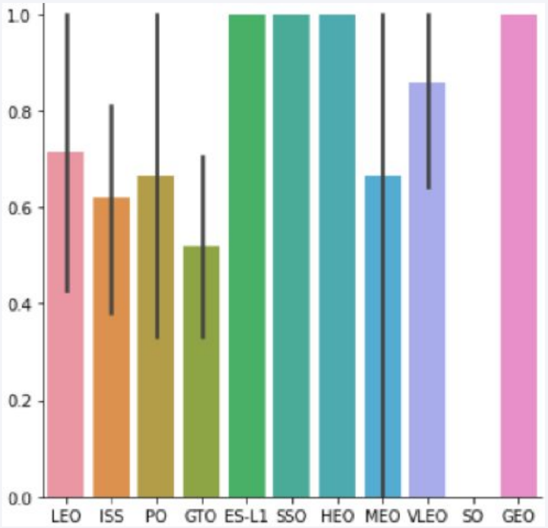
Payload vs. Launch Site

- Higher payload correlates with higher success



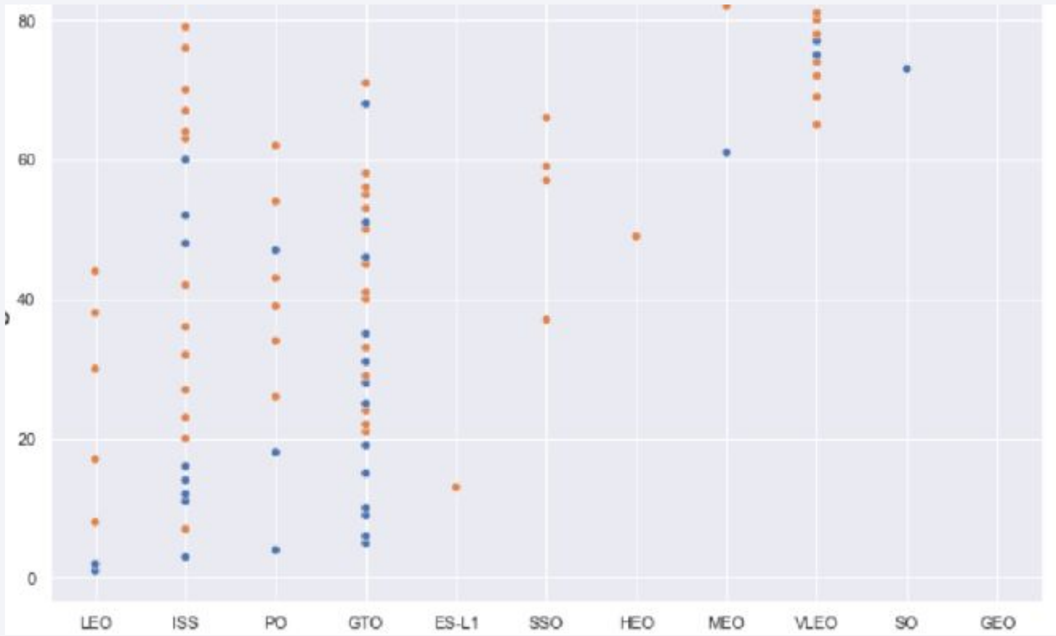
Success Rate vs. Orbit Type

- Geo, heo, sso and es-l1 have highest success rate



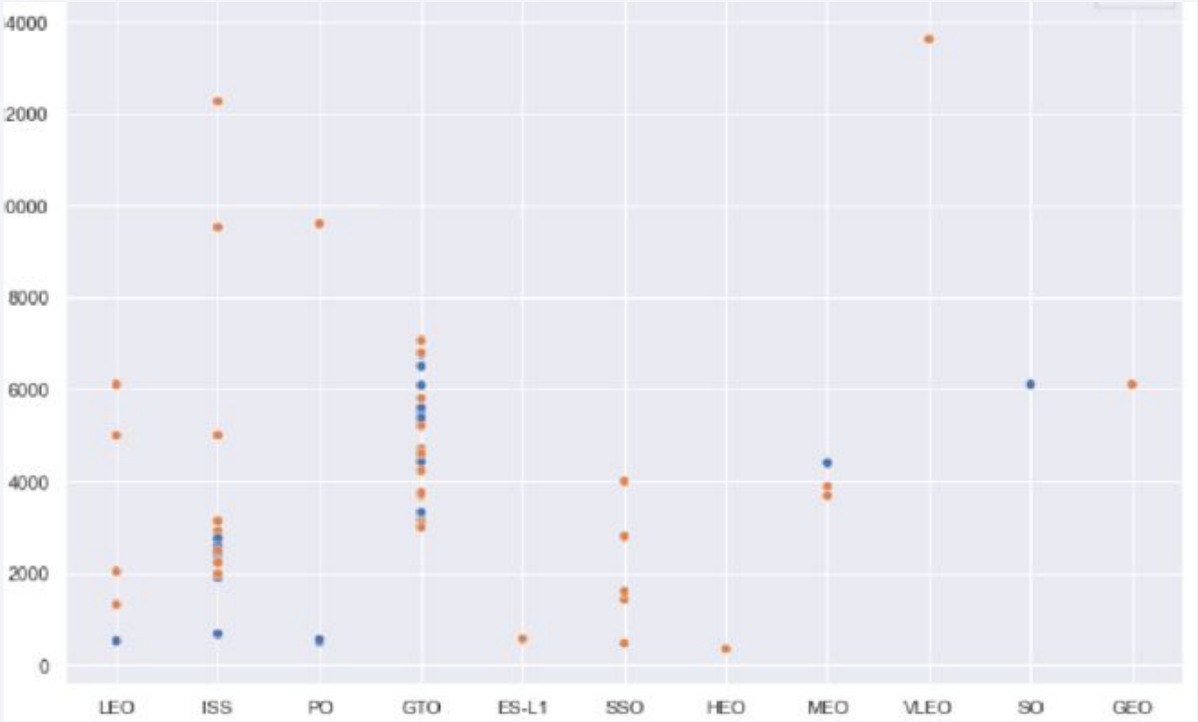
Flight Number vs. Orbit Type

- SSO is always successful, , leo seems to improve by number of flights



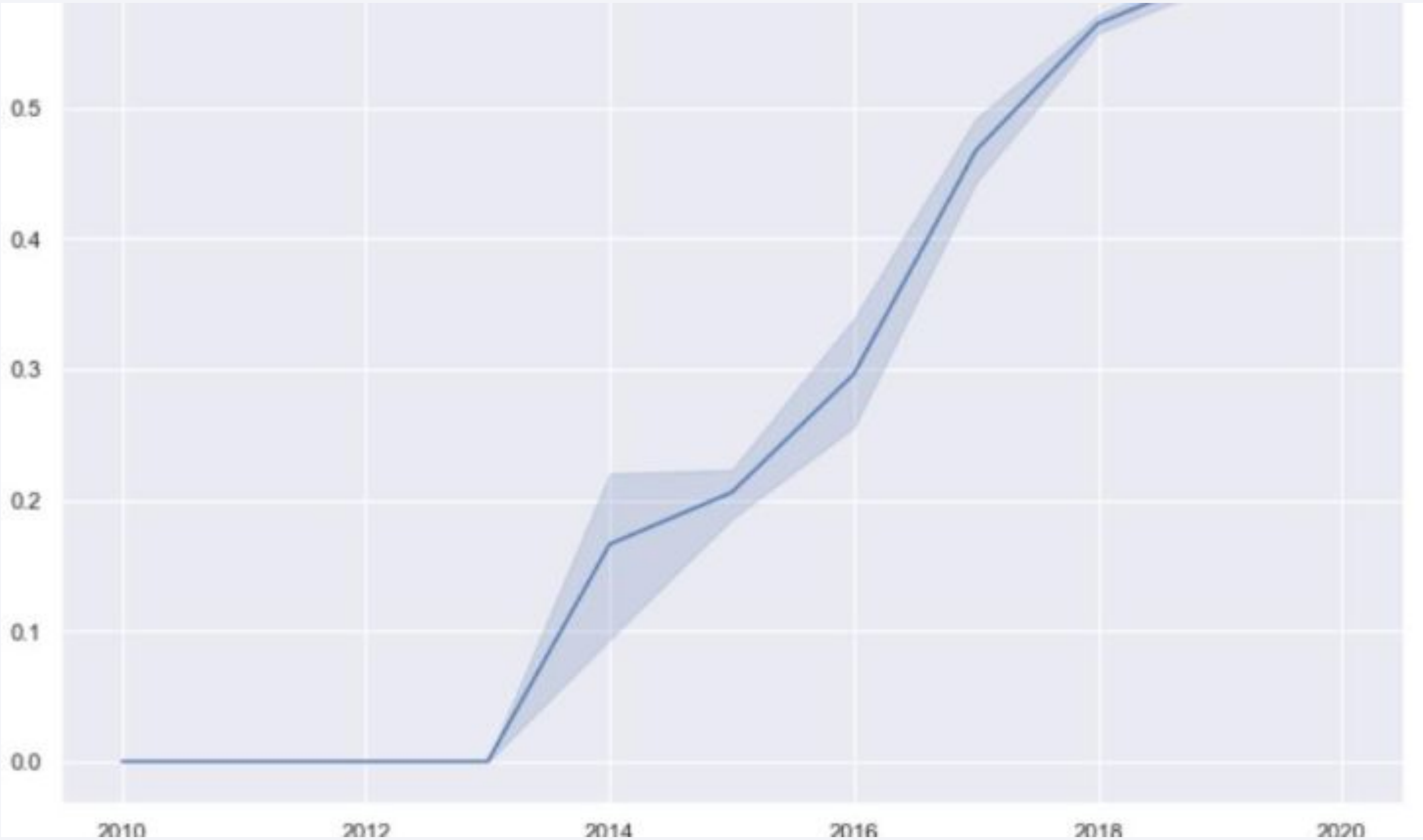
Payload vs. Orbit Type

- Payload does not seem to have significant impact on success



Launch Success Yearly Trend

- success rate has been constantly improving



All Launch Site Names

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

- select distinct launchsite from spacextbl

Launch Site Names Begin with 'CCA'

select * from spacextbl where launchsite like ‘CCA%’

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- total 45596

select sum(payloadmasskg) as total from spacextbl where customer = ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

- avg: 2928,4

select avg(payloadmasskg) as avg from spacetblx where boosterversion = ‘F9 v1.1’

First Successful Ground Landing Date

- 2015-12-22

select min(date) as first from spacextbl where landingoutcome = ‘Success (ground pad)’

Successful Drone Ship Landing with Payload between 4000 and 6000

- select boosterversion from spacextbl where landingoutcome = ‘Success (drone ship)’ and payloadmasskg between 4000 and 6000
  - F9 FT
    - B1022
    - B1026
    - B1021.2
    - B1031.2

Total Number of Successful and Failure Mission Outcomes

- select missionoutcome, count(missionoutcome) as cnt from spacextbl

  where missionoutcome in (‘Success’, ‘Failure’)

  group by missionoutcome

  99 Success

  1 Failure

Boosters Carried Maximum Payload

- SELECT DISTINCT Booster_Version
- FROM SPACEXTBL
  - WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL
- ```
  ('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

- SELECT Date, Booster_Version, Launch_Site, substr(Date, 7, 4) as year
- FROM SPACEXTBL
- WHERE "Landing _Outcome" = ('Failure (drone ship)') and substr(Date, 7, 4) = '2015' """)

```
('10-01-2015', 'F9 v1.1 B1012', 'CCAFS LC-40', '2015')
('14-04-2015', 'F9 v1.1 B1015', 'CCAFS LC-40', '2015')
```

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

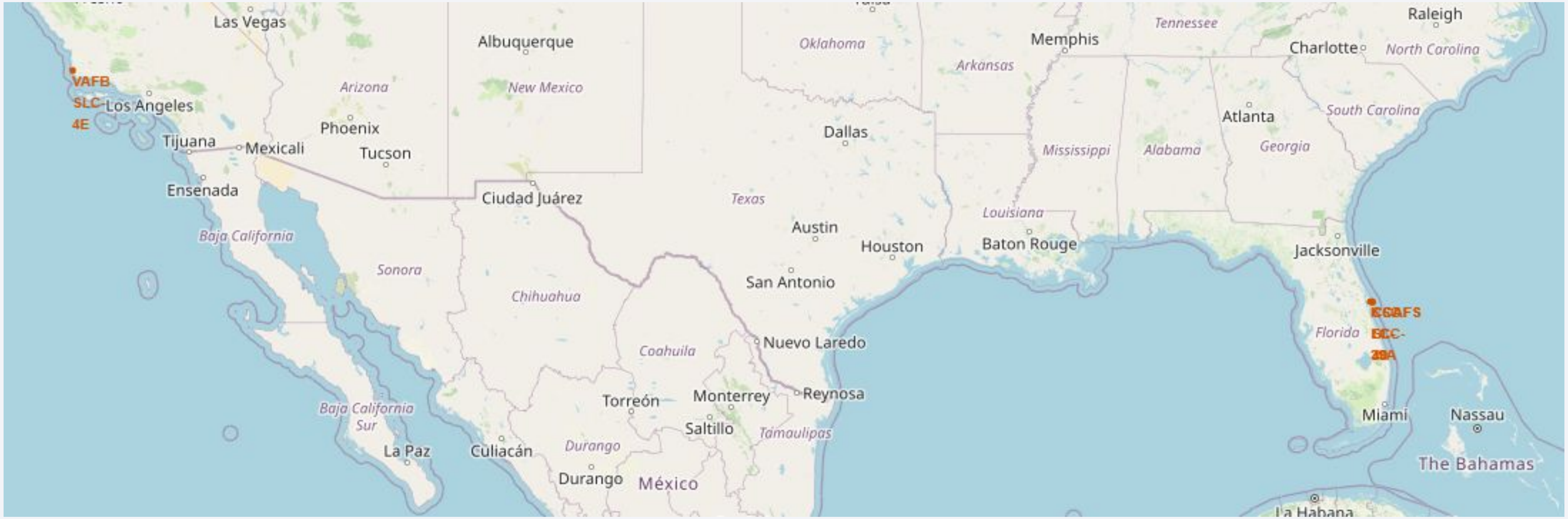
```
SELECT count("Landing _Outcome"), "Landing _Outcome", Date
        FROM SPACEXTBL
        where
        (substr(Date, 7,4) >= '2010' AND substr(Date, 4,2) >= '06' AND substr(Date, 1,2) >= 04)
        AND
        (substr(Date, 7,4) <= '2017' )
        group by "Landing _Outcome"
        order by count("Landing _Outcome") desc
```

```
(7, 'Success (drone ship)', '14-08-2016')
(6, 'Success (ground pad)', '22-12-2015')
(5, 'No attempt', '08-10-2012')
(2, 'Uncontrolled (ocean)', '29-09-2013')
(2, 'Failure (parachute)', '04-06-2010')
(2, 'Controlled (ocean)', '14-07-2014')
(1, 'Precluded (drone ship)', '28-06-2015')
(1, 'Failure (drone ship)', '15-06-2016')
```

Section 3

Launch Sites Proximities Analysis

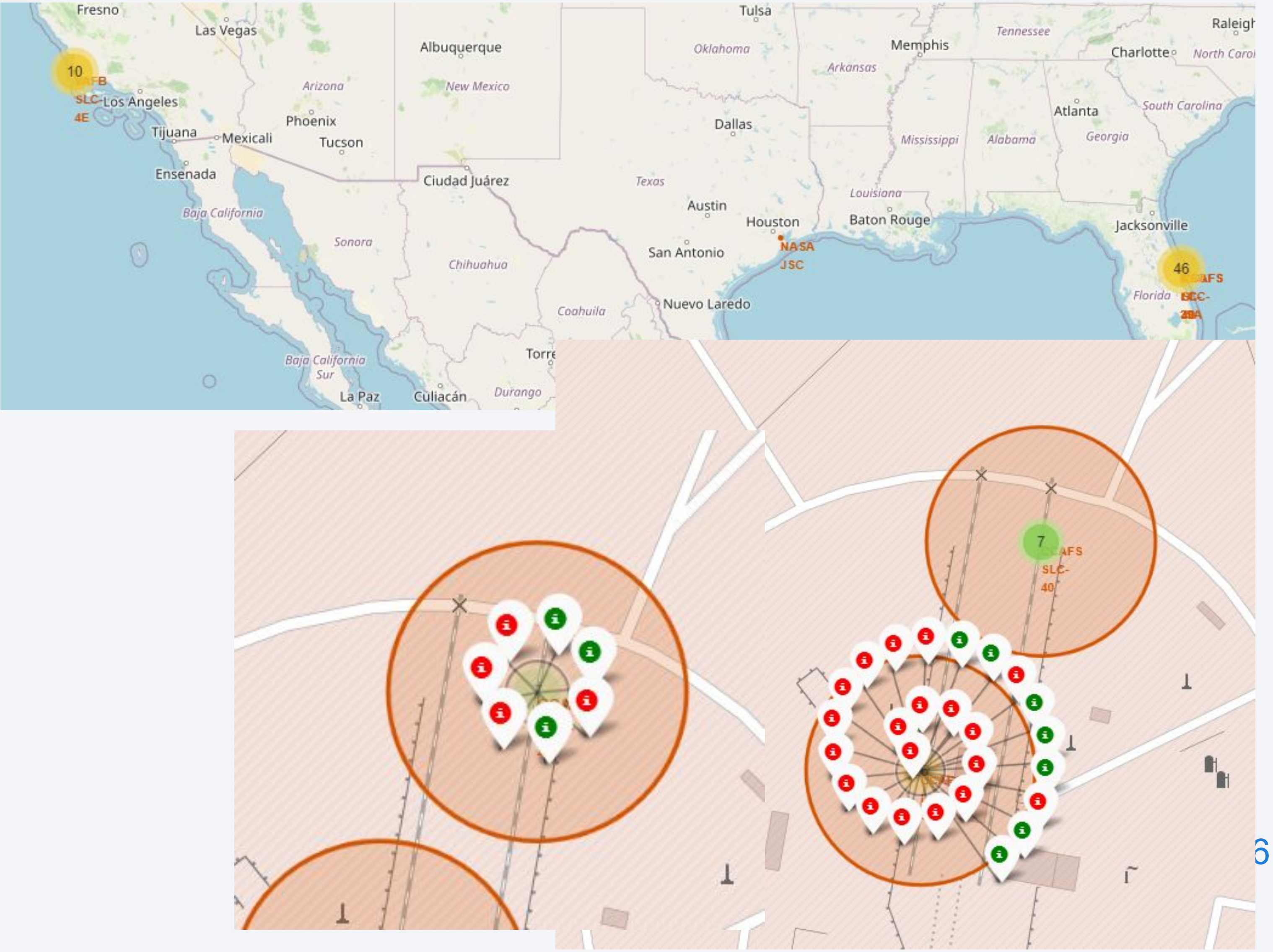
Launch Sites



All sites are in the us and on the coast

Launch Fail / Success

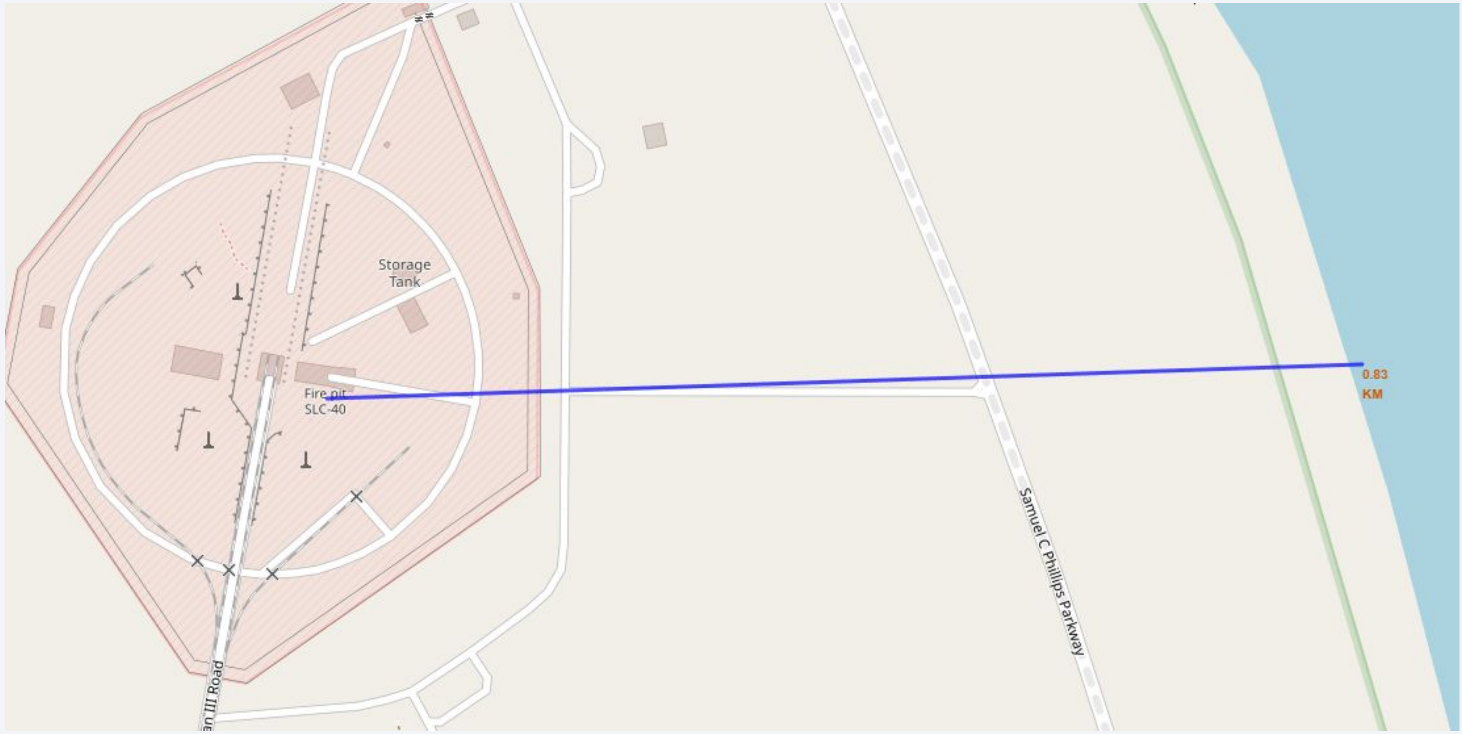
Red for fail, green for success



Proximity

Launch Sites tend to be close to transport : rail

But far from people: avoiding collateral damage

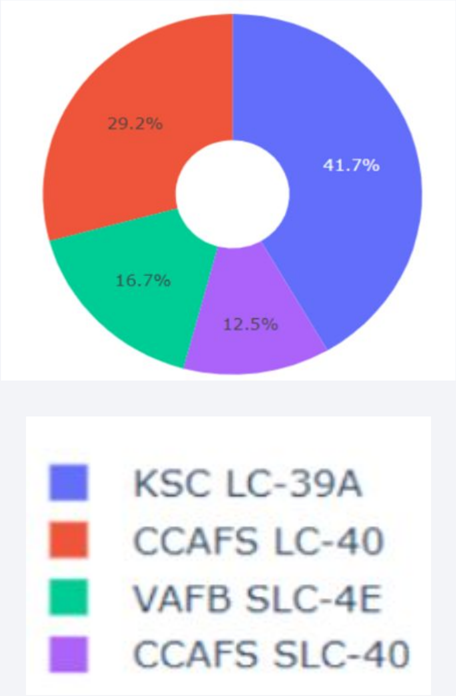


Section 4

Build a Dashboard with Plotly Dash

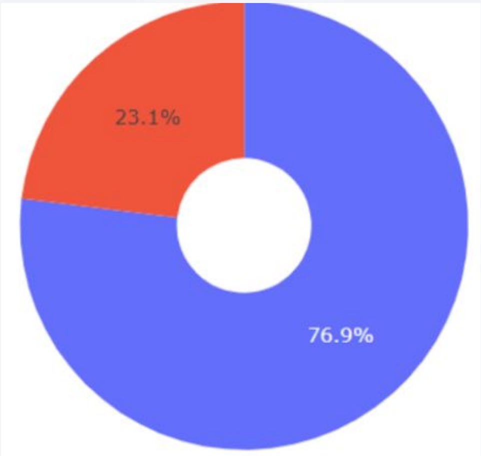
Success Rate by Site

KSC LC-39A has the highest percentage of successful launches



Top Launch Site

KCS leads at 77/23 ratio success

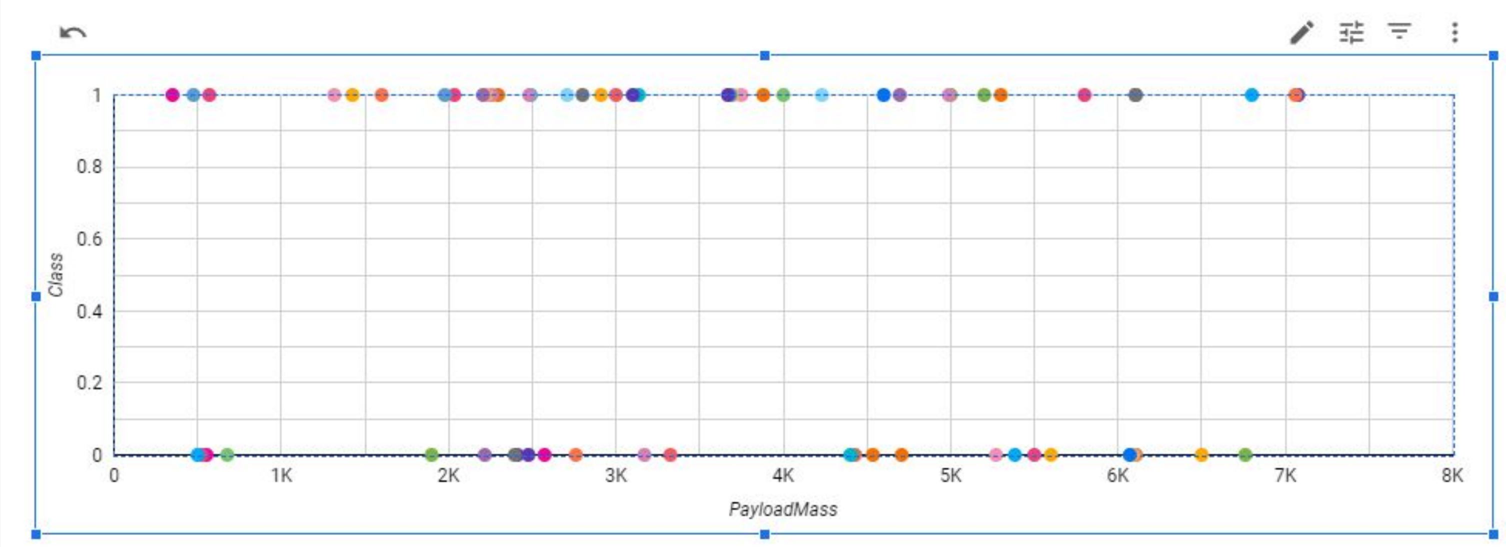


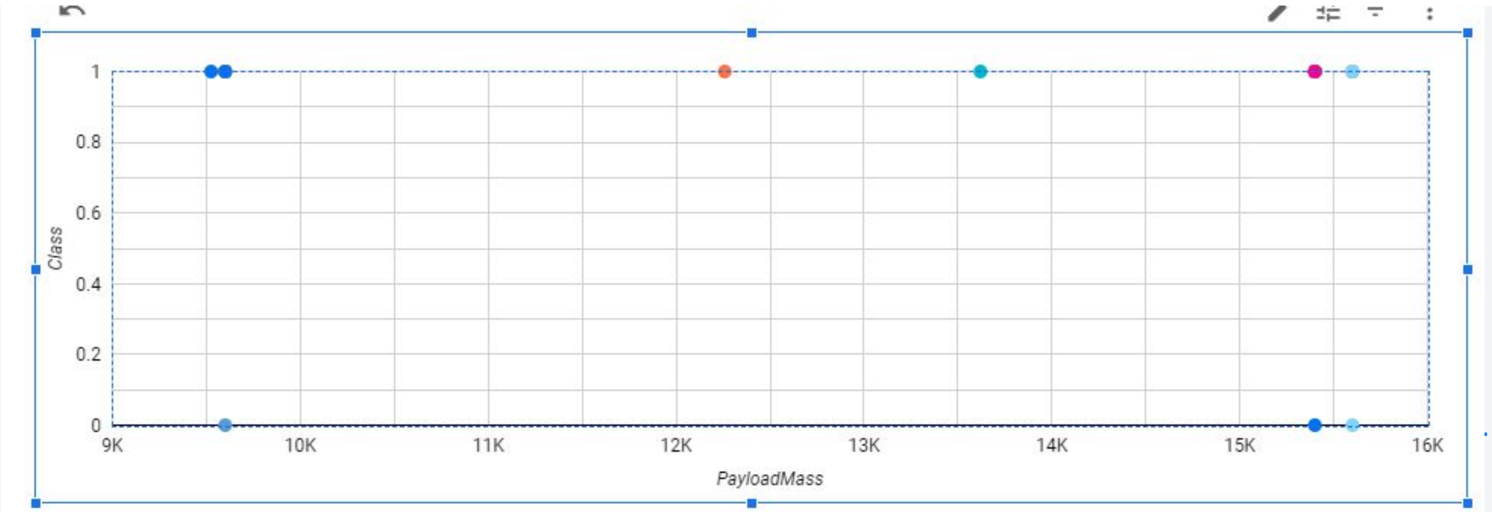
Payload Comparison

Top: sub 8k

Bottom +8k

Higher variance of success among lighter weights



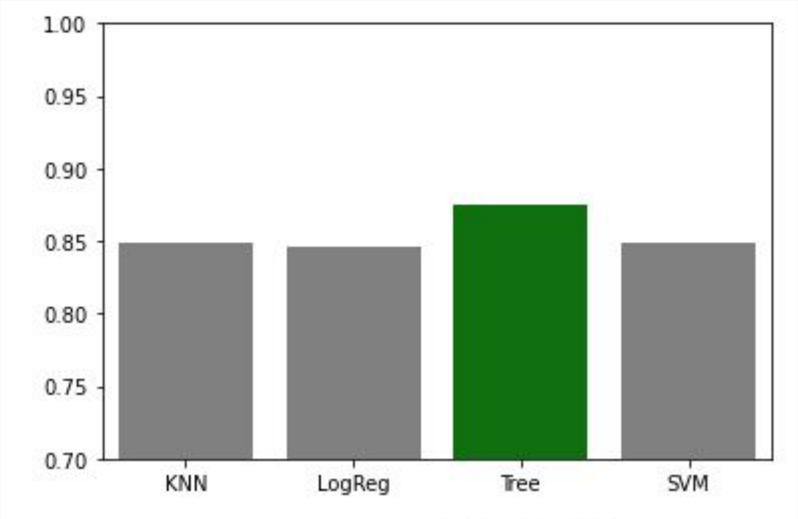


Section 5

Predictive Analysis (Classification)

Classification Accuracy

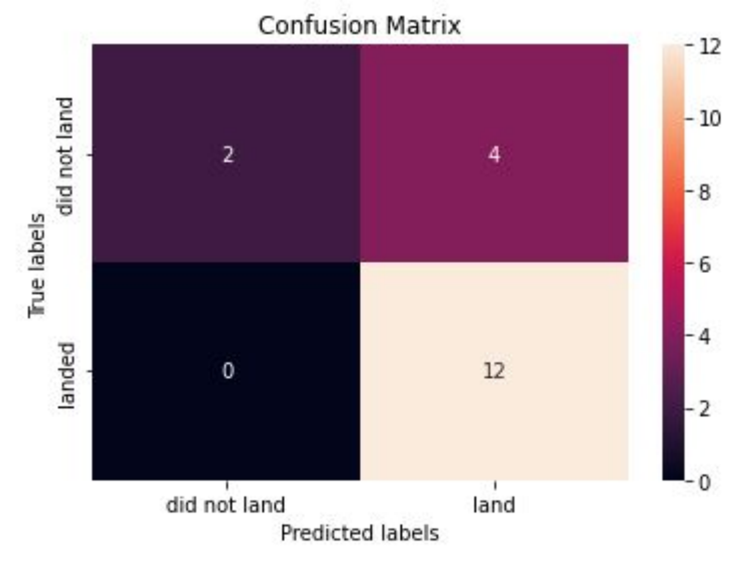
- Tree is the best model



Model	Accuracy
KNN	0.8482
LogReg	0.8464
Tree	0.8750
SVM	0.8482

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Correctly predicted 12 successful landings and 2 fails

Incorrectly labeled 4 events as successful

Conclusions

- Tree algorithm has best performance

  KSC LC 39A is most successful

  Over time success rate has improved

  sites are located far from people and near transportation

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!