Winning Space Race with Data Science

Paulo de Castro Villi
(SpaceY analyst)
pcv@spacey.com

19/mar/2022

Outline

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

Executive Summary

What was done:

- Public data was retrieved from SpaceX and Wikipedia
- It was cleaned, explored for insights, prepared and used for fitting several machine learning models over several possible parameter tunings
- These were all evaluated, to select the one with best performance

Results:

- 83,3% accuracy achieved for all tested models (at their best tunings)
- No false-negative but significant false-positive prediction outcomes
- Several specific insights noted during this process (eg: launch site and payload mass range correlation to successful outcomes)

Introduction

SpaceY likes competing with SpaceX

Predicting SpaceX launch costs can help SpaceY bid against them when viable

SpaceX can reuse stage-1 booster rockets making their launches exceptionally cheap (~ -$100 million) when successful

Crucial Question: when is SpaceX likely to try and succeed in their stage-1 reuse?

Turns out you don’t have to be a rocket scientist…
It can also be done through Data Science!

Section 1

Methodology

Methodology

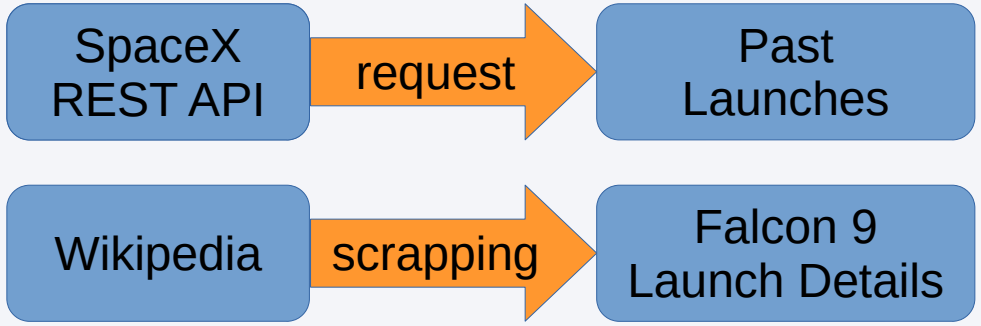
Executive Summary

- Data collection methodology:
  - Public data from SpaceX REST API and Wikipedia
- Perform data wrangling
  - Value-mapped landing outcomes to positive/negative outcome classes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 4 models fitted over several parameter tunings then tested to select the one(s) with best performance

Data Collection

Past launch data retrieved via SpaceX REST API

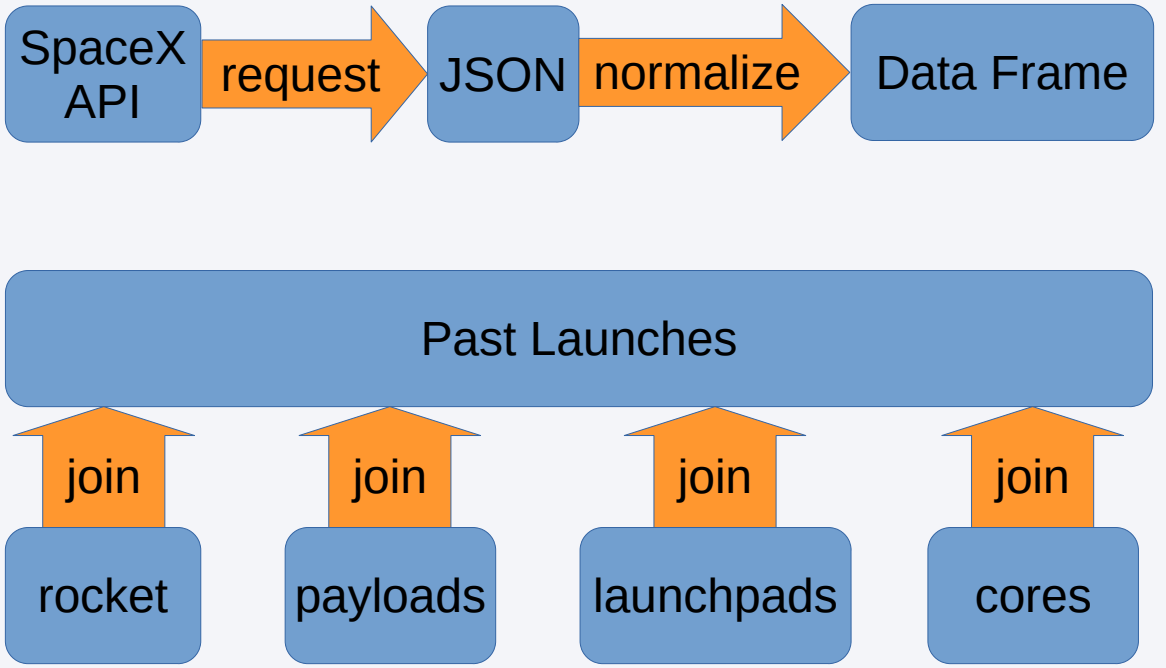
Extra info on Falcon 9 launches retrieved from Wikipedia



Data Collection – SpaceX API

- JSON responses requested from SpaceX REST API, then transformed into dataframes
- Main data collected for past launches
- Complementary data retrieved (rocket, payloads, launchpad, cores), cleaned and joined
- Data filtered for Falcon 9 launches
- Missing payload data filled with average payload

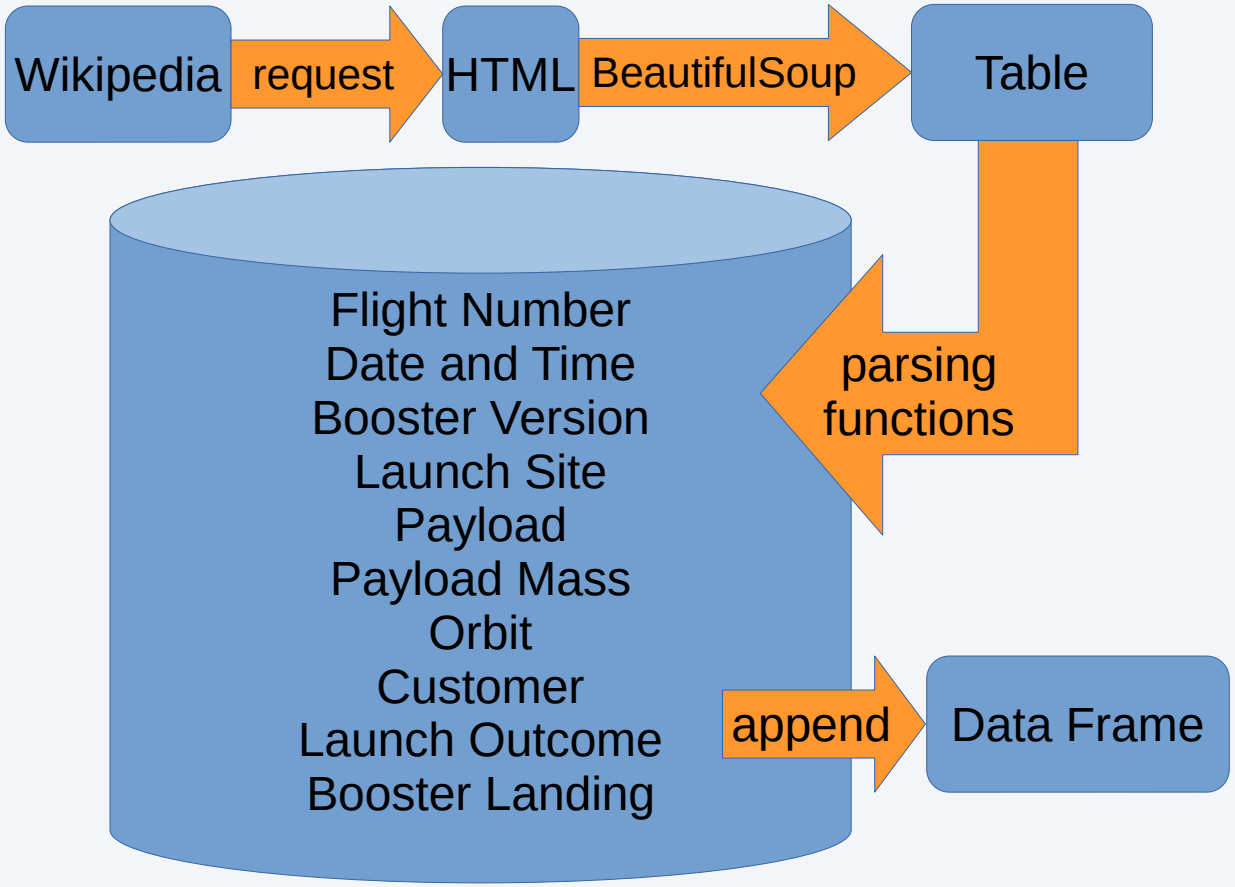
  See the work here: https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20SpaceX%20Data%20Collection%20API.ipynb



Data Collection - Scraping

- Falcon 9 Launch Wiki page requested from Wikipedia
- BeautifulSoup used for locating relevant table (Falcon 9 launch details)
- Parsing functions used for clean retrieval of each data

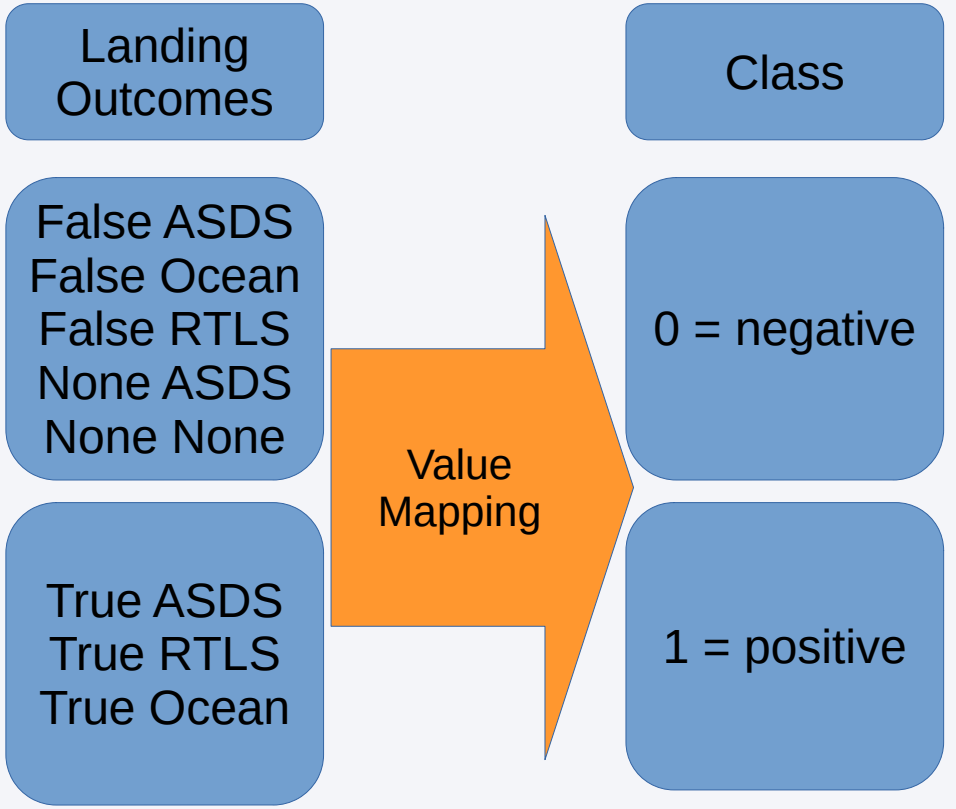
See the work here: https://github.com/paulovilli/coursera/blob/master/DS0321EN - SpaceX Data Collection API.ipynb



Data Wrangling

- Examined launch-site occurences
- Examined target orbit occurences
- Examined landing outcome occurences
- Value-mapped landing outcomes to positive/negative outcome classes

See the work here: https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20EDA%20-%20SpaceX%20Data%20Wrangling.ipynb



EDA with Data Visualization


PayloadMass vs. FlightNumber (vs. Class)


Flight Number vs. Launch Site (vs. Class)


Payload Mass vs. Launch Site (vs. Class)


Success Rate vs. Orbit Type**


Flight Number vs. Orbit Type (vs. Class)


Payload Mass vs. Orbit Type (vs. Class)


Success Rate vs. Year Line Chart


Scatter Plot
Categoric / Categoric correlation


Bar Chart
Categoric / Numeric correlation


Line Chart
Numeric Time Series

See the work here:
https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- ✔ Names of unique launch sites
- ✔ 5 records where launch sites begin with 'CCA'
- ✔ Total payload mass by boosters launched by NASA (CRS)
- ✔ Average payload mass by booster version F9 v1.1
- ✔ Date of first successful landing in ground pad
- ✔ Names of boosters that landed successfully in drone ship after carrying payload mass >4000 but <6000
- ✔ Total successful and failure mission outcomes
- ✔ Names of booster versions which carried max. payload mass
- ✔ Failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
- ✔ Ranked count of landing outcomes between 2010-06-04 and 2017-03-20 (descending order)

See the work here:
https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Explored launch site location criteria and correlation with success ratio:

- ✔ Marked and circled all launch sites
- ✔ Marked success/failed launches for each site
- ✔ Marked nearest relevant features (Coastline, Highway, Railway, City)
- ✔ Drew lines between a launch site and the nearest relevant features
- ✔ Labeled the nearest relevant features with the corresponding distances to the launch site

See the work here:
https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20Interactive%20Visual%20Analytics%20(Folium).ipynb

Build a Dashboard with Plotly Dash

Further exploration (interactive) of Launch Site and Payload Mass with Success Rate:

- ✔ Launch Site (selector)
- ✔ Total Success Launches by Site (pie chart) – if All Sites selected
- ✔ Total Success/Failure Launches by Site (pie chart) - for specific site selected
- ✔ Payload range (kg) (range selector)
- ✔ Payload Mass (kg) vs. Success class (scatter plot) – depending on site and payload range

See the work here:
https://github.com/paulovilli/coursera/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

Target variable (Y): 'Class'

X data standardization: StandardScaler

X and Y data split: 80% train, 20% test

Classification Models:

- LR - Logistic Regression
- SVM – Support Vector Machine
- DT – Decision Tree
- KNN – K Nearest Neighbours

Optimization:

GridSearchCV (parameter sets for each model)

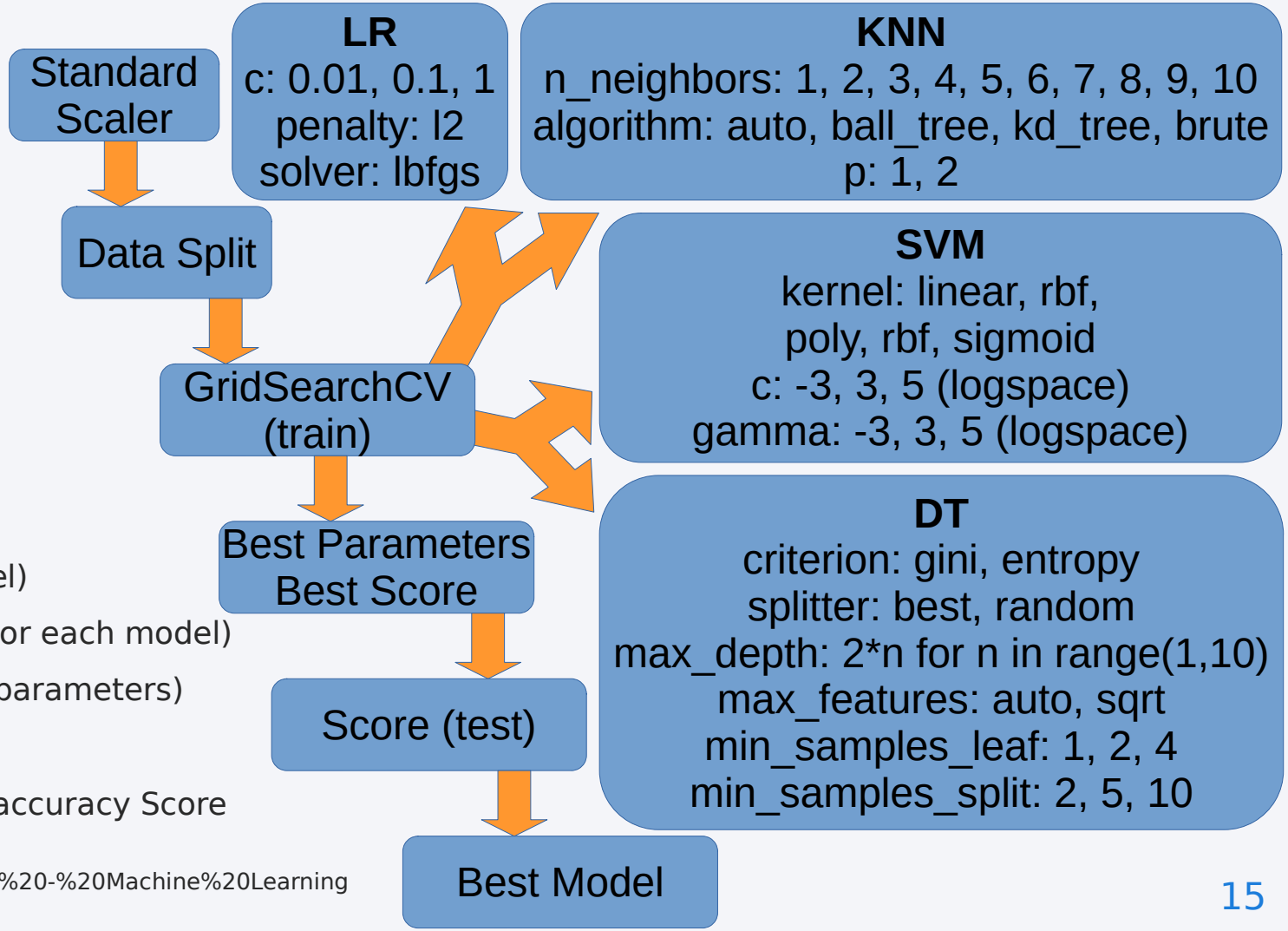
Best Parameters & Best Score (for train set, for each model)

Score (for test set, for each model with best parameters)

Confusion Matrix for each model

Best Model: any model with the maximum accuracy Score

See the work here:
https://github.com/paulovilli/coursera/blob/master/DS0321EN%20-%20Machine%20Learning%20Prediction.ipynb



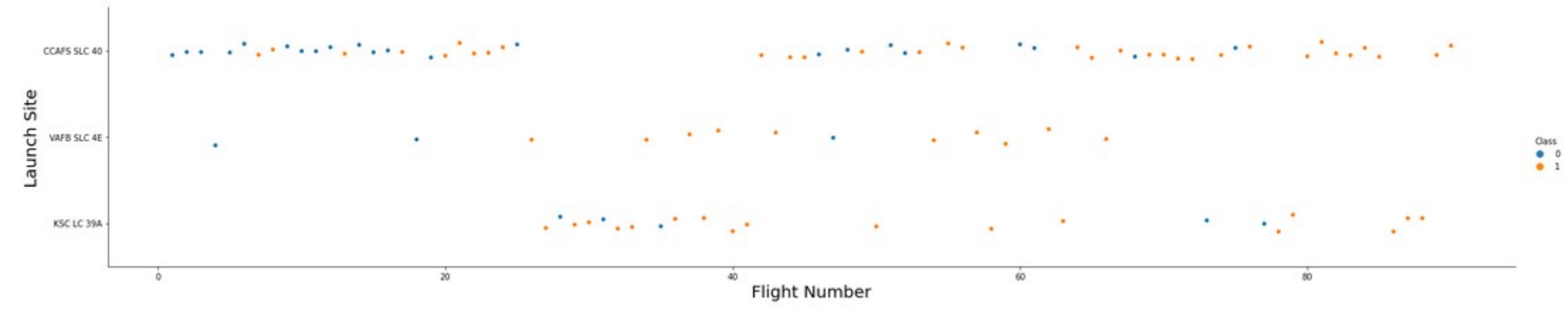
Results

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

Section 2

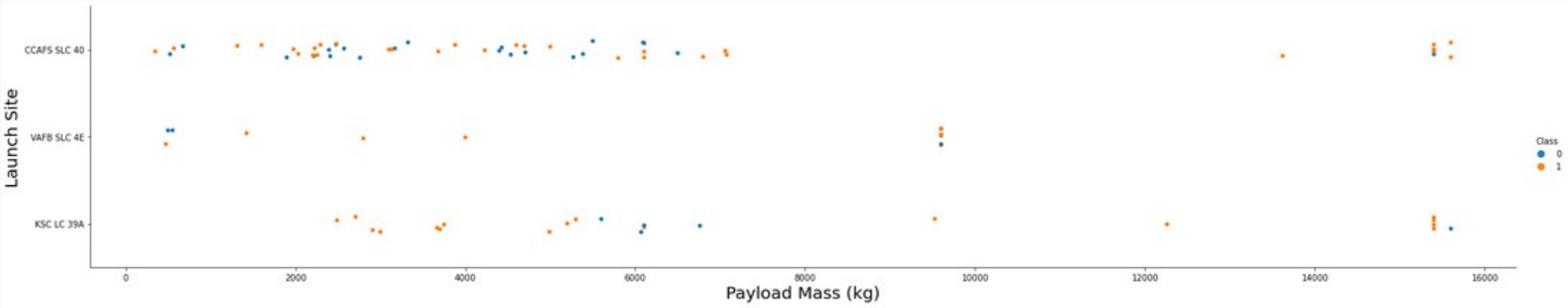
Insights drawn from EDA

Flight Number vs. Launch Site



First several launches at each site had worse reliability than the latest

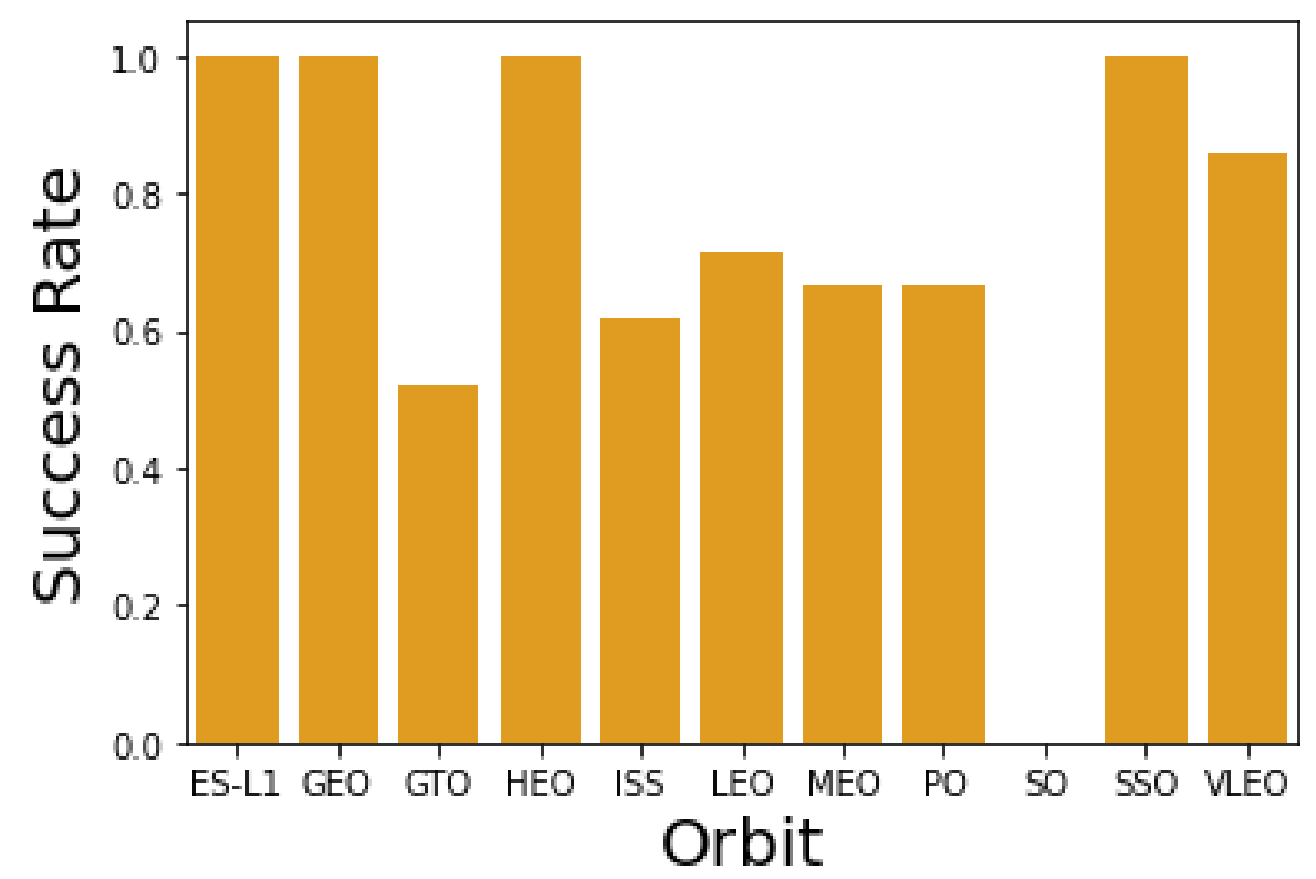
Payload vs. Launch Site



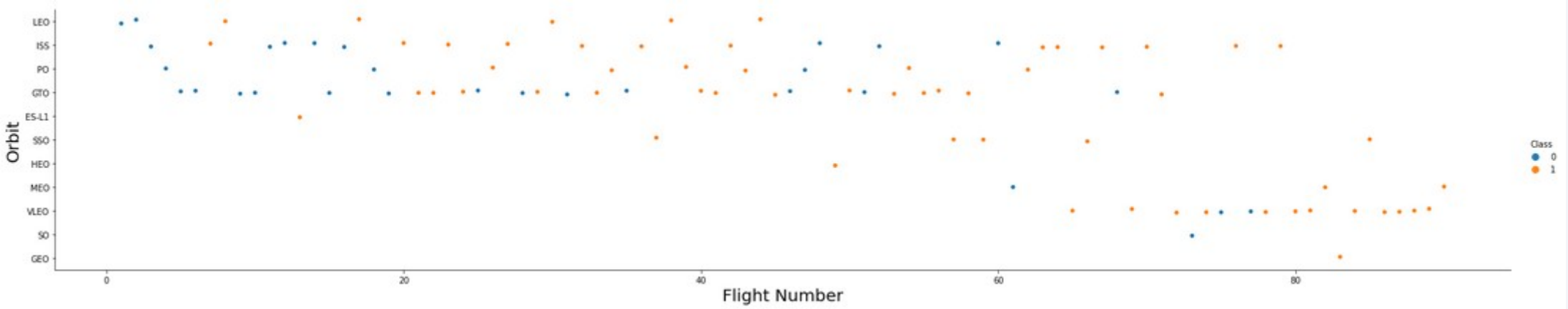
VAFB-SLC launch site: no rockets launched
for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type

Maximal Success Rate:
ES-L1, GEO, HEO and SSO target orbits



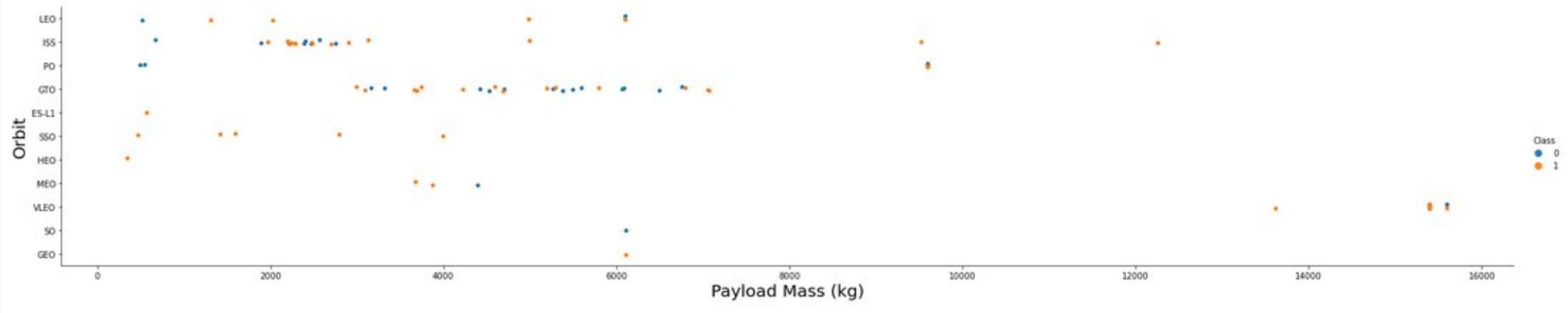
Flight Number vs. Orbit Type



LEO Orbit: success appears related to number of flights

GTO Orbit: success appears unrelated to number of flights

Payload vs. Orbit Type



For Heavy Payloads: more successful landings for Polar, LEO and ISS

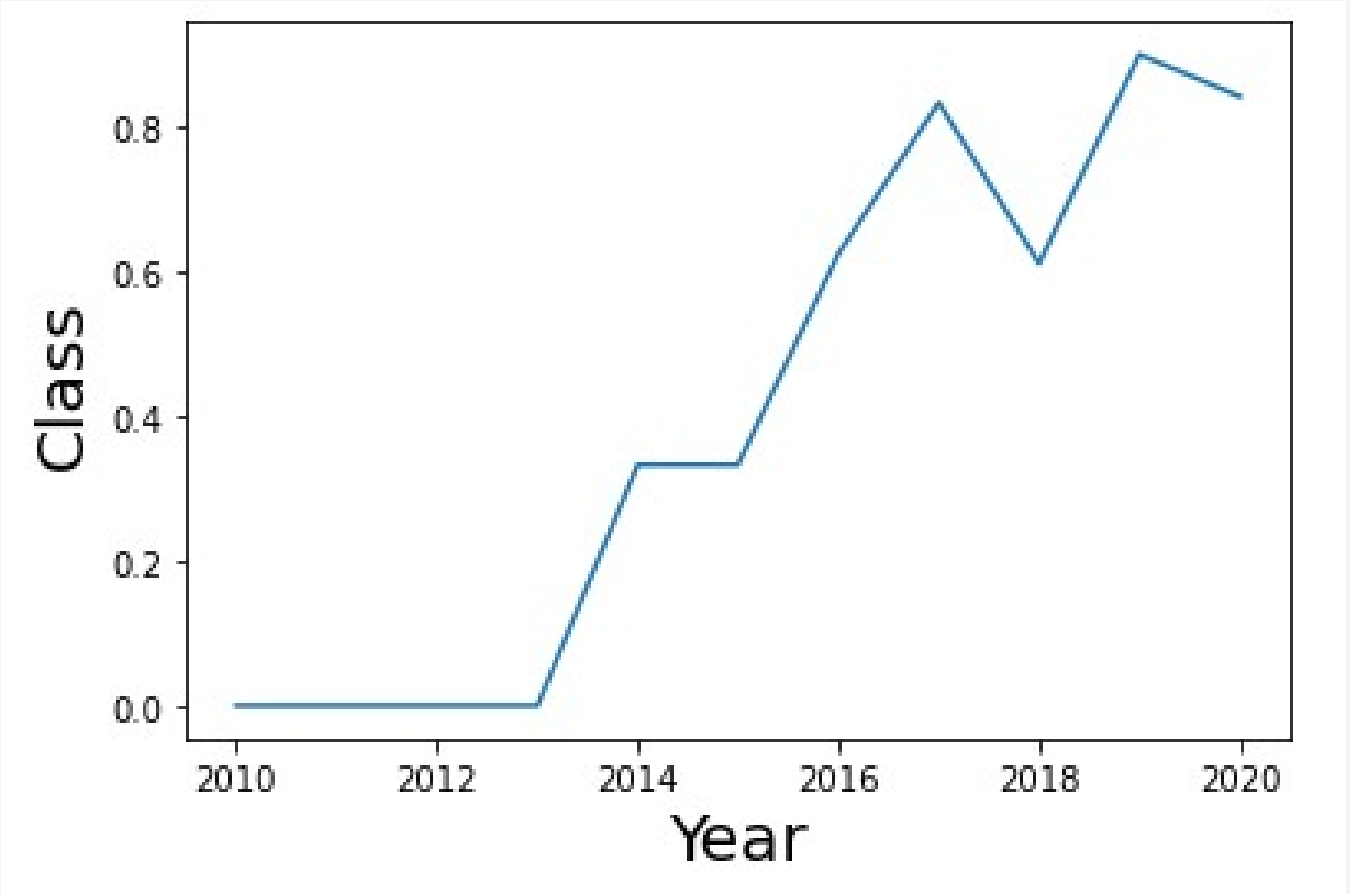
For GTO: both positive and negative landings

Launch Success Yearly Trend

Success Rate:

starts increasing in 2013

kept increasing till 2020



All Launch Site Names

Names of unique launch sites

```
%sql select distinct launch_site from SPACEXDATASET
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9
Done.
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA` (very near sites CCAFS LC-40 and CCAFS SLC-40)

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

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

Total payload carried by boosters from NASA

```
%sql select sum(payload_mass__kg_) from SPACEXDATASET where customer='NASA (CRS)'
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde0
Done.
   1
45596
```

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

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

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.d
Done.
```

1
2928

First Successful Ground Landing Date

Date of first successful landing outcome on ground pad

```
%sql select min(DATE) from SPACEXDATASET where "Landing _Outcome"='Success (ground pad)'
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.data
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters that successfully landed on drone ship after carrying payload mass >4000 but also <6000

```
%sql select distinct(booster_version) as "booster" from SPACEXDATASET where payload_mass__kg_>4000 and payload_mass__kg_<6000
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

booster
F9 B4 B1040.2
F9 B4 B1040.1
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as "count" from SPACEXDATASET group by mission_outcome
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databas
Done.
```

mission_outcome	count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Names of boosters which carried maximum payload mass

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

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

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

Failed landing outcomes in drone ship, with booster versions and launch site names for 2015

```
%sql select booster_version, launch_site, DATE, "Landing _Outcome" from SPACEXDATASET where year(DATE)=2015 and "Landing _Outcome"='Failur
```

```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

booster_version	launch_site	DATE	Landing _Outcome
F9 v1.1 B1012	CCAFS LC-40	2015-01-10	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	2015-04-14	Failure (drone ship)

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

Ranked count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order

```
%sql select "Landing _Outcome" as "landing_outcome", count(*) as "count" from SPACEXDATASET where DATE>='2010-06-04' and DATE<='2017-03-20
```

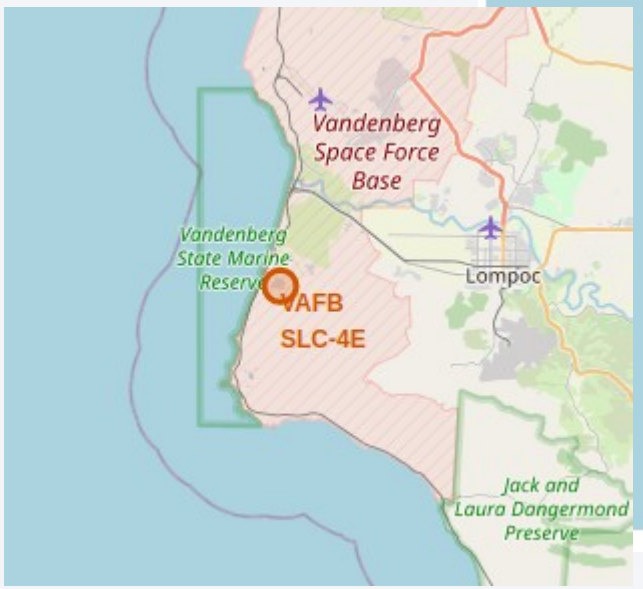
```
 * ibm_db_sa://cgc91496:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

landing_outcome	count
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

Launch Site Locations

All launch sites are located near the coastline, but spanning different coasts/oceans

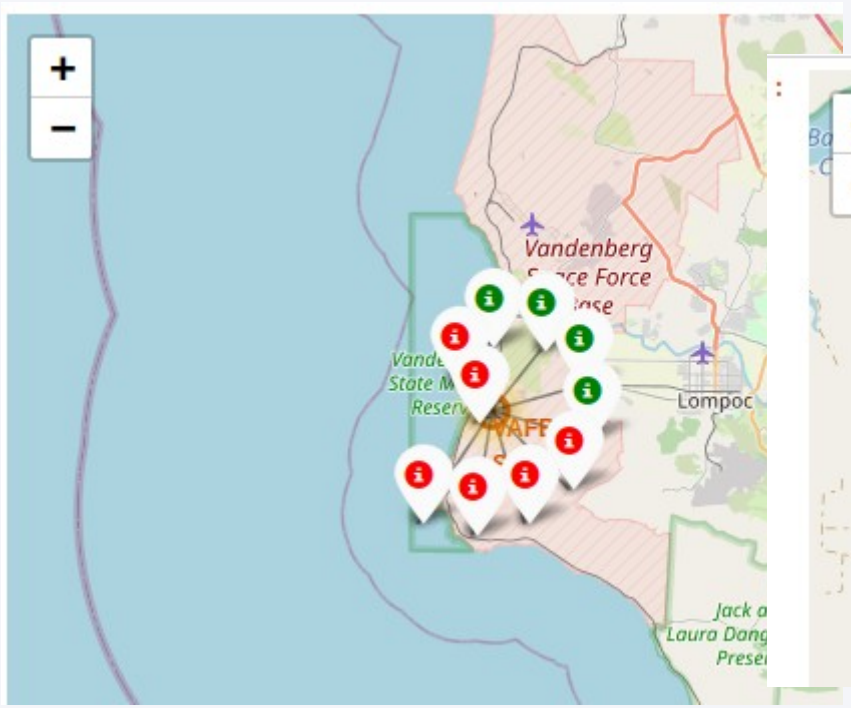


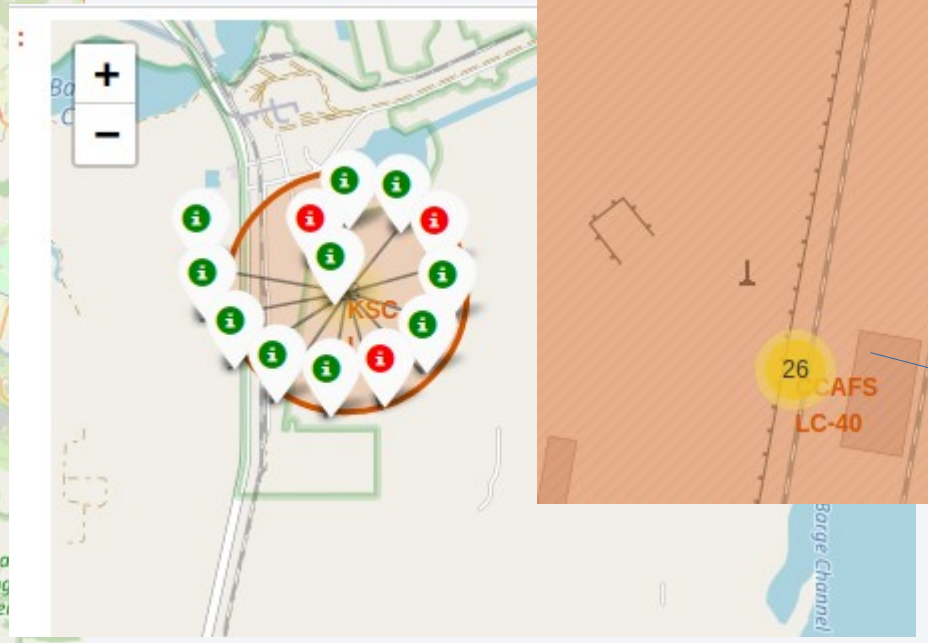
Success/Failure by Launch Site

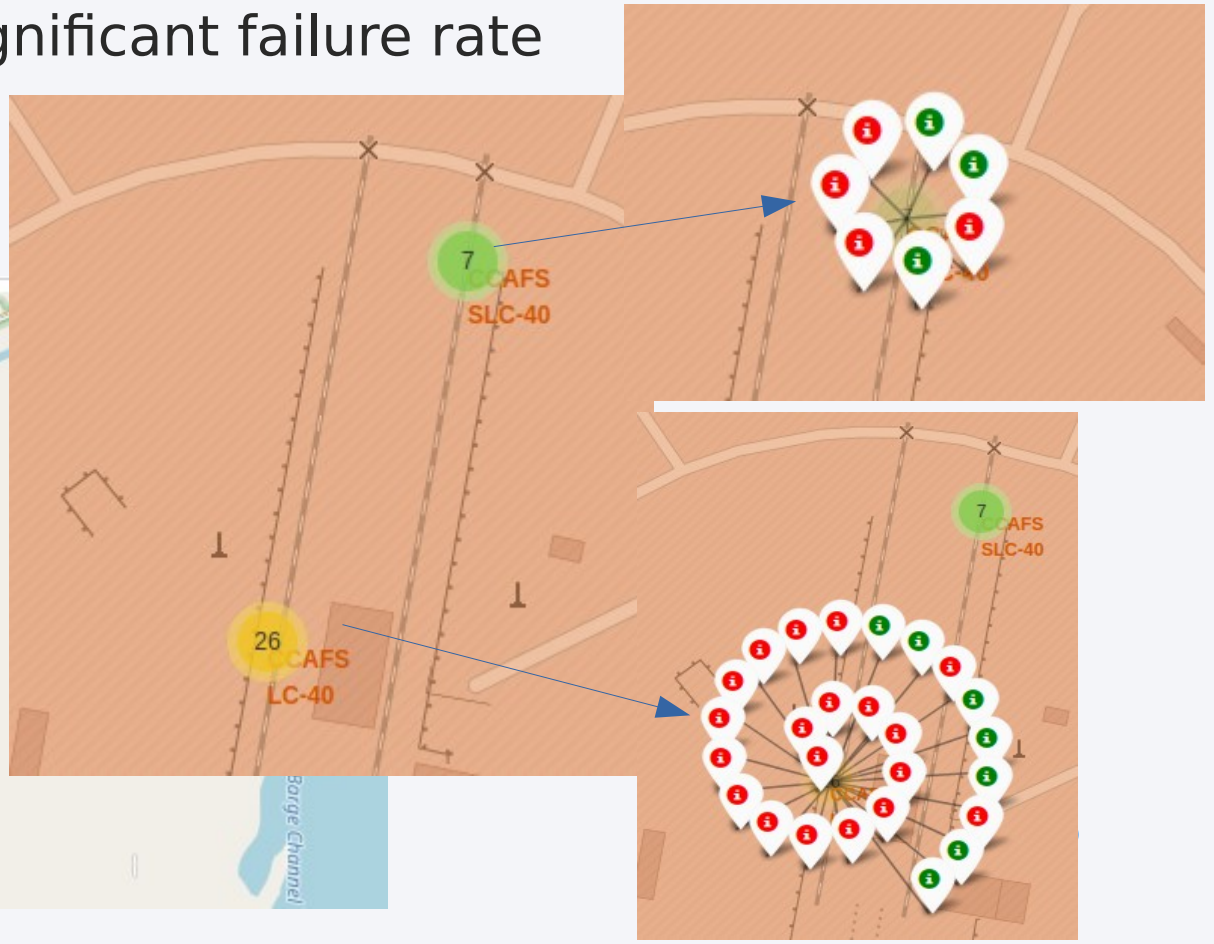
Color-labeled launch outcomes plotted on the map
(red=failure, green=success)

CCAFS LC-40 has largest amount and significant failure rate

KSC LC-39A has best success rate







Nearest Features of Interest

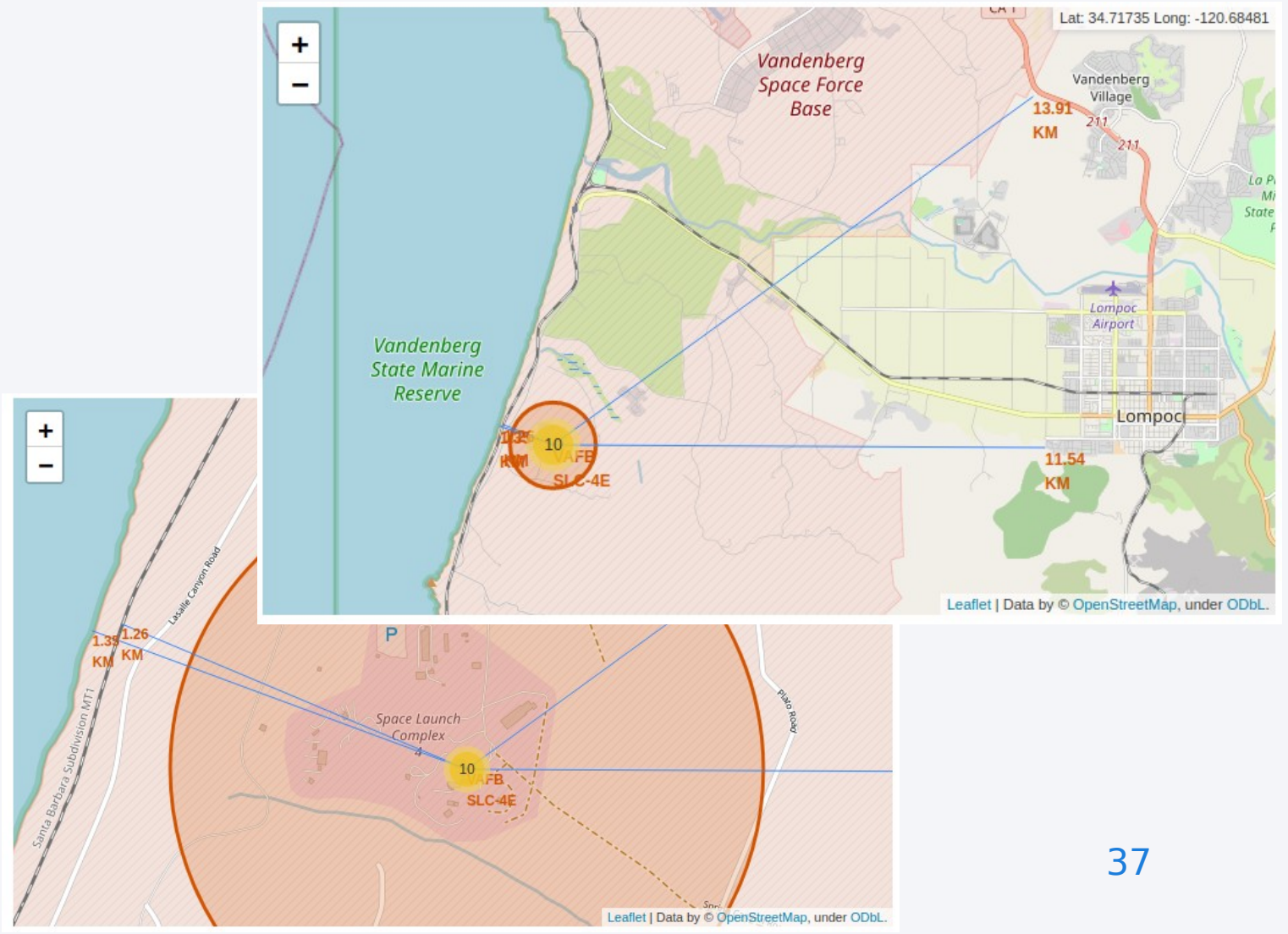
Features of Interest

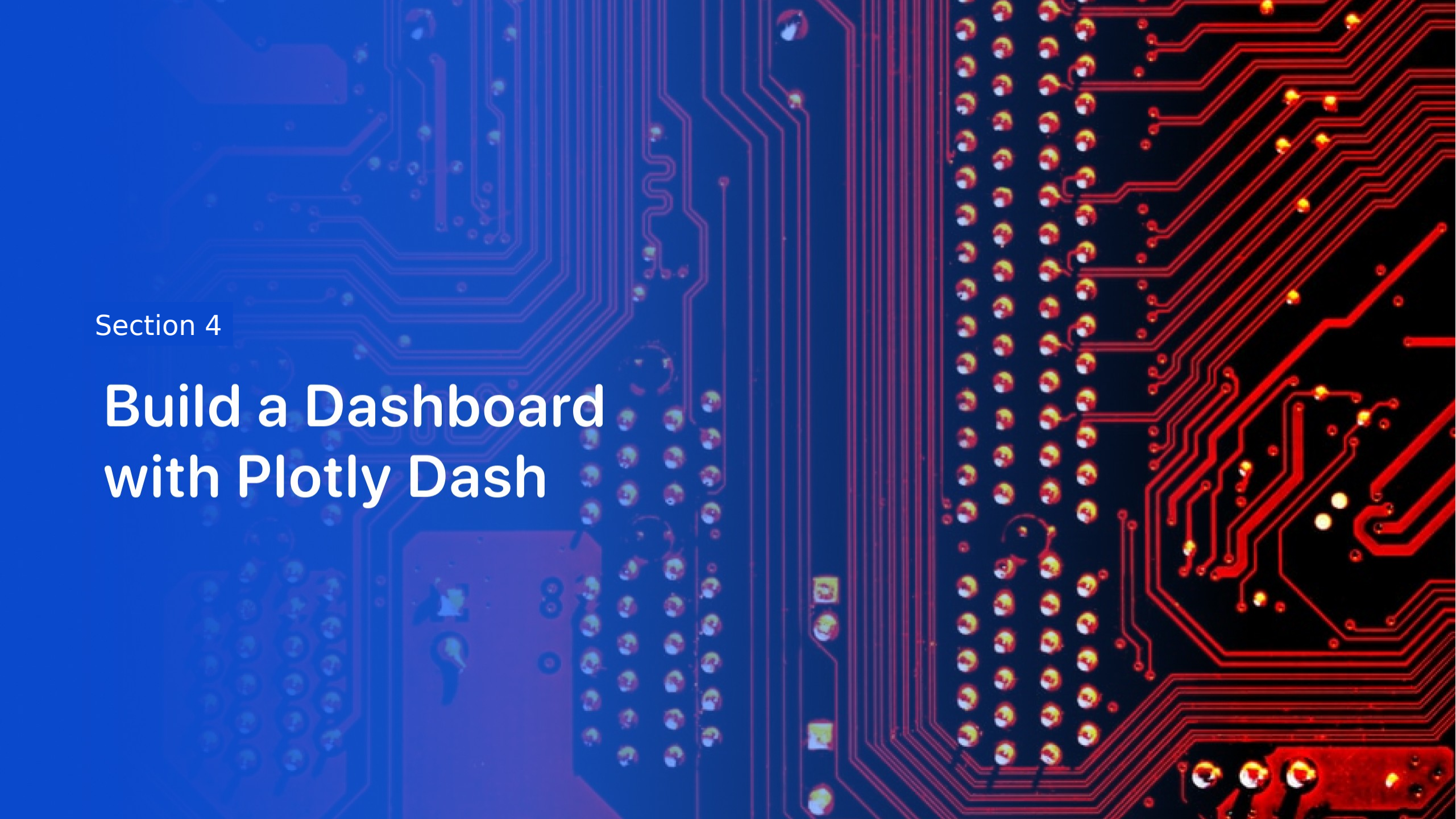
Railways: close proximity
(logistic necessity)

Highways: no proximity
(safety on early-launch stages)

Coastline: close proximity
(safety on later launch stages)

Cities: no close proximity
(safety on early-lauch stages)



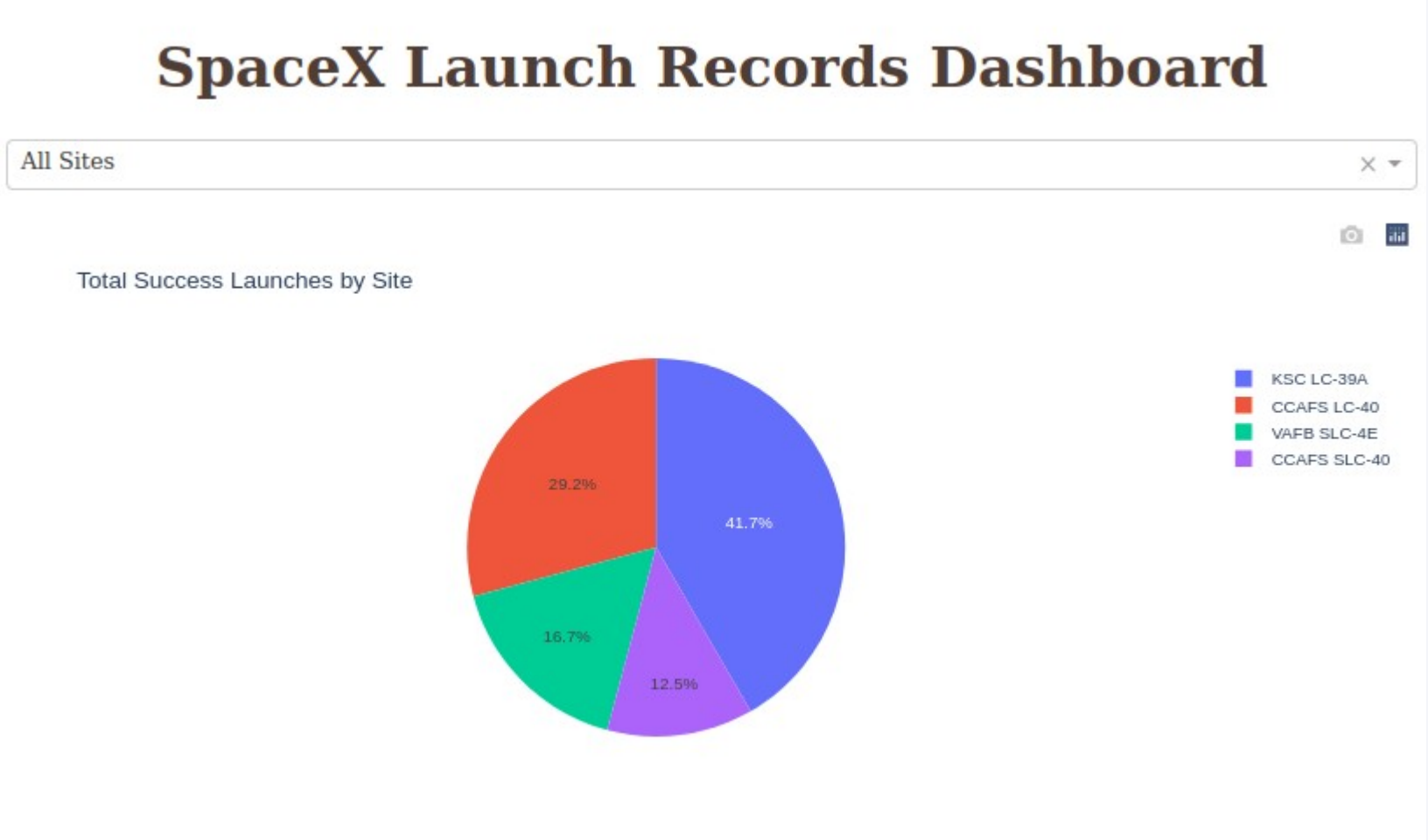

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site

Launch success count for all sites

Launch site KSC LC-39A has the highest count of successful launches

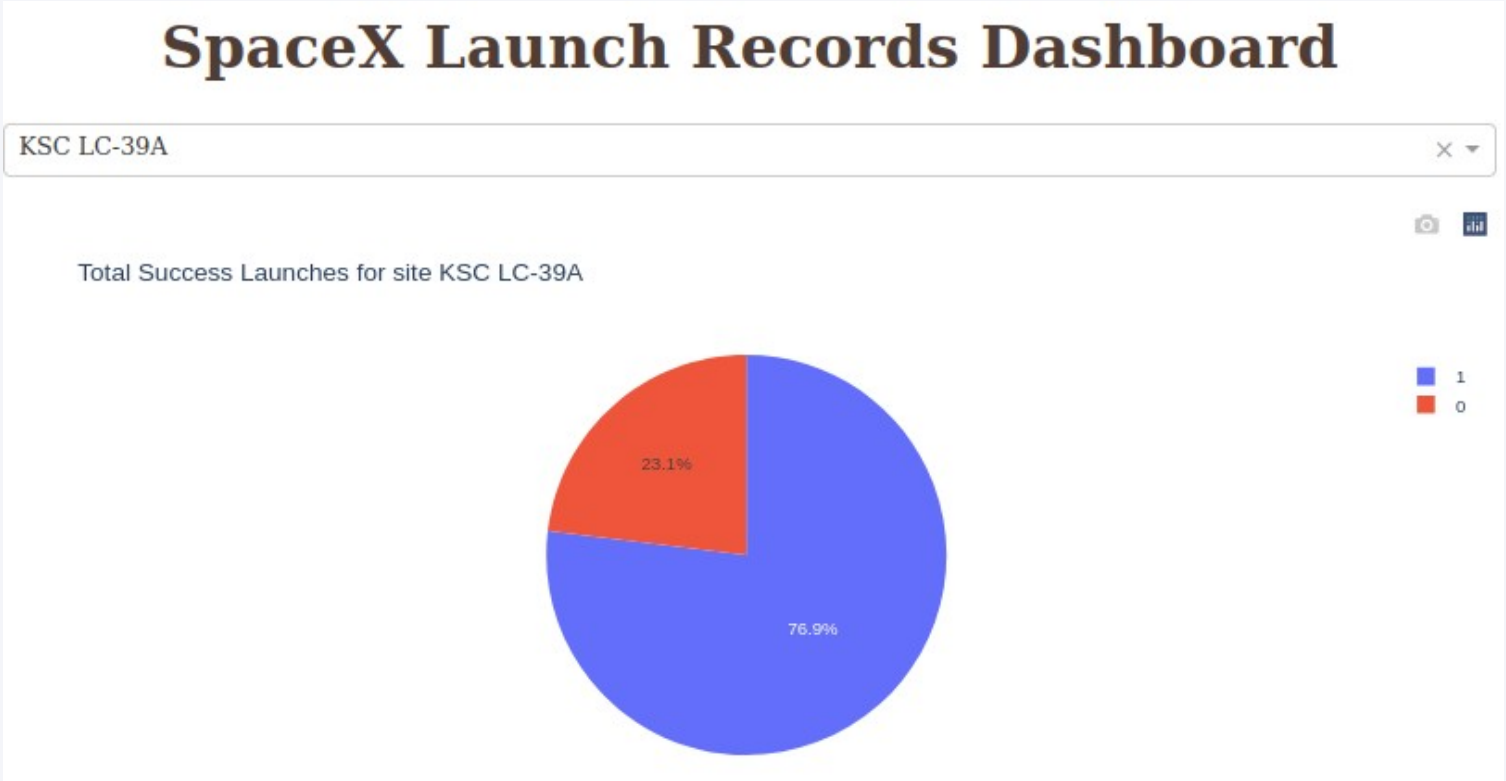
CCAFS SLC-40 has the smallest count of successful launches



Highest Success Ratio for a Launch Site

Launch site with highest launch success ratio: KSC LC-39A

Same site has the highest count of successful launches, so it might be favored over others when possible



Outcome by Payload Mass ranges

Scatter Plot reacts to “Payload Range (kg)” slider and shows outcome class by booster rocket model

Payloads in the 1900-3700 range are more successful

Booster version FT appears to have the largest success rate

B4 and FT are the only booster versions working in the 7500-10000 payload range

Section 5

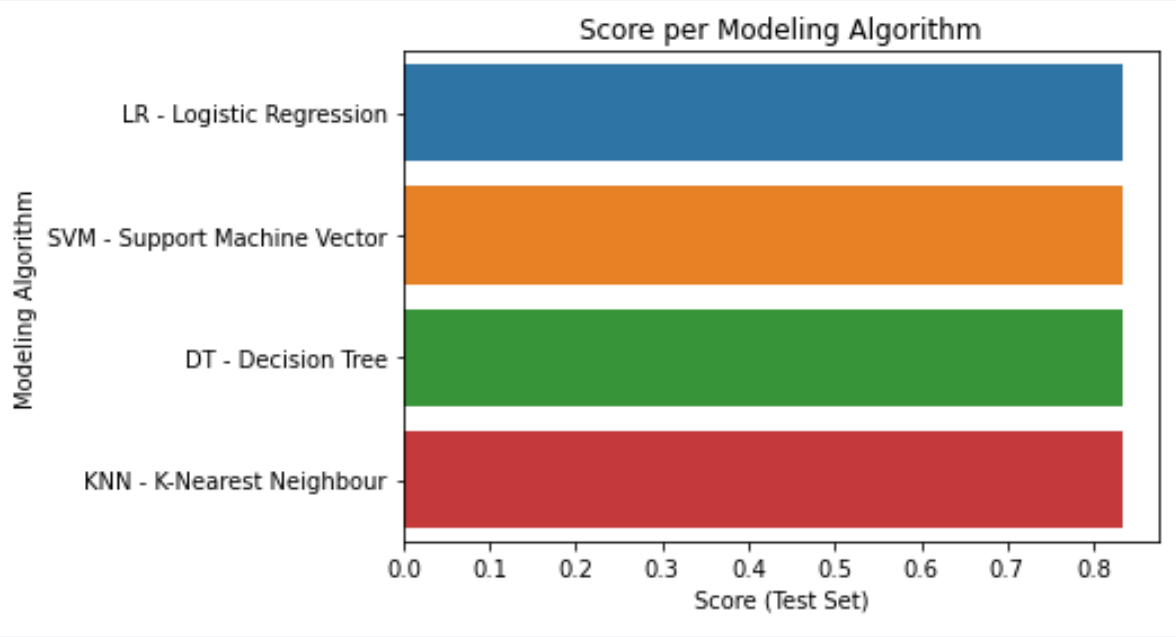
Predictive Analysis (Classification)

Classification Accuracy

Max Accuracy: 0,8333333333334

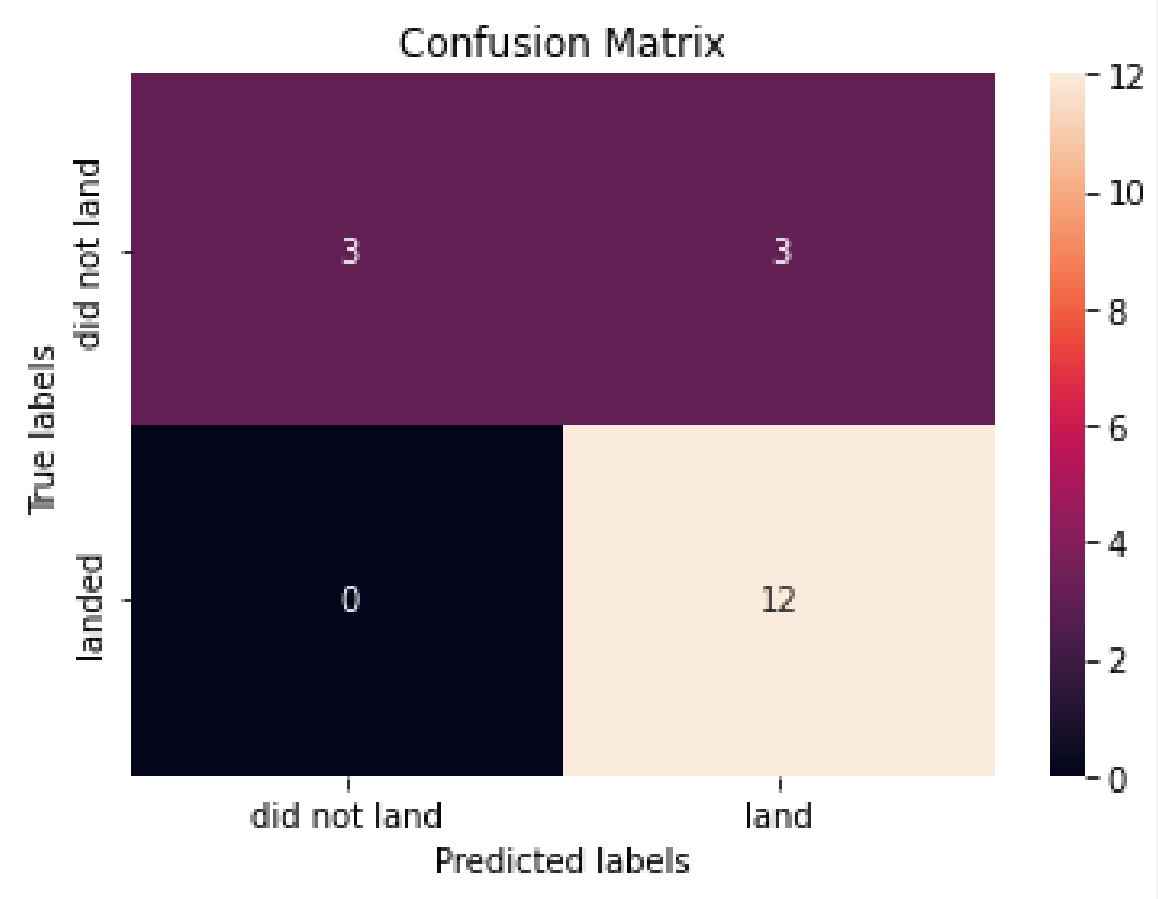
All models show similar accuracy

Decision Tree unreliable between runs with same dataset+parameters (due to very limited dataset)



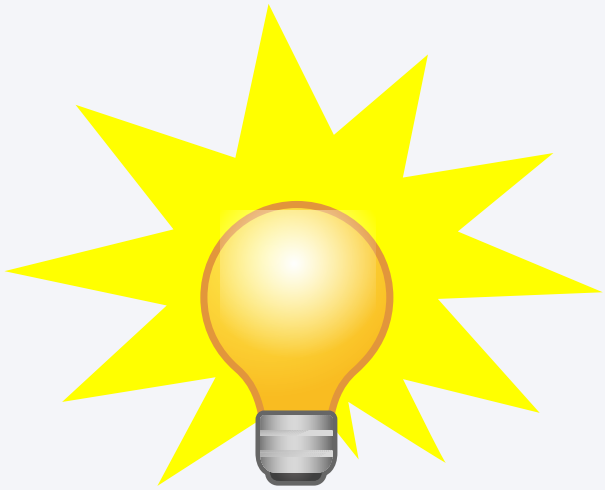
Confusion Matrix

- All models performed similarly and also resulted in a similar confusion matrix
- There were no false negatives
- Half the real negative outcomes in the test set have been predicted as successes
- More records and/or more in-depth information on past launches would be necessary to improve this



Conclusions

- We can predict outcomes with 83,3% accuracy...
- ...but False-Positive predictions are still very significant
- Current working dataset is barely sufficient
- Payload masses in the 1900-3700 range have more reliable launches
- Newer booster models are more reliable (they are getting better at stage 1 design)
- KSC LC-39A has highest successful launches count and highest success rate
- There is a clear logic to launch site location



Appendix

All relevant assets uploaded to GitHub:

https://github.com/paulovilli/coursera/tree/master

SpaceX REST API data source:

https://api.spacexdata.com/v4/

Wikipedia Data Source:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_...

Thank you!