Applied Data Science Capstone

Juan Manuel Floría
15/02/2022

https://www.xataka.com/espacio/quien-quien-competidores-spacex-que-empresas-estan-pisando-fuerte-espacio

Outline:

- Executive Summary
- Introduction
- Methodology
- Conclusion
- Appendix
- Results:
  - E.D.A. with Visualizations
  - E.D.A. with SQL
  - Interactive maps with folium
  - Dashboards Plotly Dash
  - Predictive Analysis (Classification)

https://www.spacex.com

Executive Summary:

- Changed all categorical variables to binary using one hot encoding.
- Visualized accuracy of all sorts.
- Standardized data and used GridSearchCV in order to search the best parameters for ML.
- Collected meaningful columns to be used as features
- Data recorded from public SpaceX API and SpaceX Wikipedia.
- SQL used for exploratory data, visualizations, dashboards, and folium maps.
- Models of ML used: Decision Tree Classifier, Logistic Regression, Support Vector Machine, k-nearest neighbours.

Introduction:

Background:

Undoubtedly, this is a brand new first attempt to a fast challenging advance spaceflight for human affordables, when raise funds and comprehension for important reasons are on the rise. Just here and now. The program is made up with a trio of spaceflight missions performed by astronauts. Their core goal is using brand new technologies, performing the first flight of SpaceX's starship with humans on board, and driving a broad and wide research.

Our mission:

As technical crew from Space Y currently, we are fully ready to improve the price for each launching from new missions, designing fresh and meaningful dashboards for stakeholders by recording information on web pages, and training a ML model to predict the profitable and successful launches of Stage1 to commit in approaching terms.

https://www.spacex.com/updates/#starship-update

Methodology:

- Data Wrangling:
  Successful and Unsuccessful landings classified.
- EDA by SQL and visualization tools.
- Data collection sources: by wikipedia SpaceX, and SpaceX web page mainly.
- Interactive Visual Analytics: Plotly dash, and Folium.
- Classification models to perform Predictive Analysis

Data collection overview:

The process of data collection is a mix made up with request from API and of SpaceX in wikipedia. The processing of information is conveyed next from SpaceX and also is viewing the sequence of processing the webscraping in order to fetch the data.
Columns SpaceX: PyloadMass, LaunchSite, Outcome, FlightNumber, Date, GridFinds, BoosterVersion, Longitude, Latitude, ReusedCount, Reused, LandingPad, Serial.
Columns Wikipedia: Orbit, Customer, Flight, Launch site, PayloadMass, Payload, Version Booster, Time, Booster Landing, Date.

Data collection / SpaceX:

- Dictionary meaningful information
- Converting dictionary to a pd.DataFrame
- Request from SpaceX
- JSON files and Lists (Payload Data, Launch site)
- Normalize to DataFrame.
- Filtering according to a discrete flight nr.

https://www.nasa.gov/image-feature/spacex-crew-4-astronauts-participate-in-a-training-session-6

Data collection / Webscraping:

Broadly speaking, the process consisted in creating a training model committing landing outcomes. For it, we arranged ‘1’ for success, and ‘0’ for failure attempt.
Mission outcome, and Landing location make up the outcome column.
New training model label column as ‘class’, with the same results, 1-success, 0-failure.

Value Mapping:
True RTLS-True Ocean, True ASDS -> 1
False ASDS, False Ocean, False RTLS, None Ocean, None ASDS -> 0

https://www.xataka.com/espacio/starship-spacex-lograra-ponerse-orbita-este-ano-ese-paso-crucial-para-llegar-colonizar-marte

EDA with DataVisualization:

This Analysis has been accurately performed on variables class, year, launchsite, orbit, flightnumber, payloadmass.
And it has been done by using the next plots:
scatter plots, line charts, bar plots (taken mainly to know the type of correlation as well as whether some variable might be taken for a ML model). Also, flightnumber/payloadmass, flightnumber/landinglocation, successrate/orbit, and flightnumber/orbit.

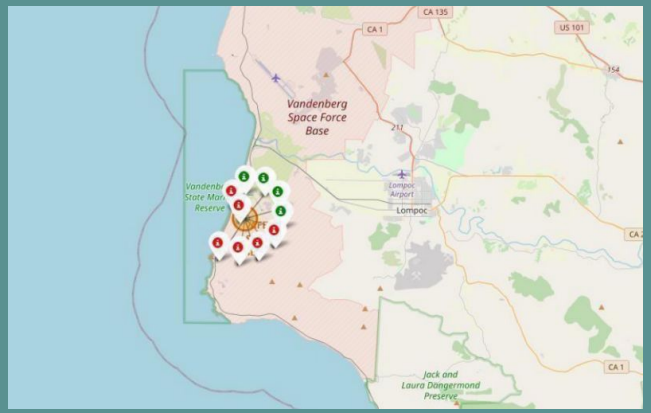
EDA with SQL:

SQL Python integration used for queries.
IBM DB2 as a Database for loading.
Fully explaining added queries shown for better information comprehension and better results deduction.
Mission outcomes, launching locations, dates, times, booster versions, and payload sizes of customers are part of special developed queries on BD.

Folium interactive Map:

Launch sites location analysis with Folium.
Maps marked:

- Successful landings.
- Unsuccessful landings.
- City.
- Coast.
- Railway station.
- Highway.



https://www.atb.com.bo/internacional/despega-cohete-de-spacex-con-cuatro-astronautas-hacia-la-estación-espacial

Predictive Analysis:

- Fit and transform features with standard scaler.
- Train, test and split data.
- Split label column class from dataset.
- GridSearchCV on LogReg, SVM, DecisionTree, k-n models and to find optimal parameters.
- Barplot for score models.
- Confusion Matrix for all models.

Conclusion:

- Created ML model with an accuracy over 83%.
- Used data from a public SpaceX api and web scraping wikipedia web page.
- Created dashboards for visualizations.
- Created data labels and data stored into DB2 SQL
- The goal of model is to predict when Stage1 will successfully land to save 100 M$
- SpaceY can use this model in order to predict if a lunch would be successfully performed before its starting.
- Much more data should be provided so as to gain more predictive accuracy of the launchings, as well as the adoption of Deep Learning techniques to be applied.

https://www.spacex.com/updates/#starship-moon-announcement

Appendix:

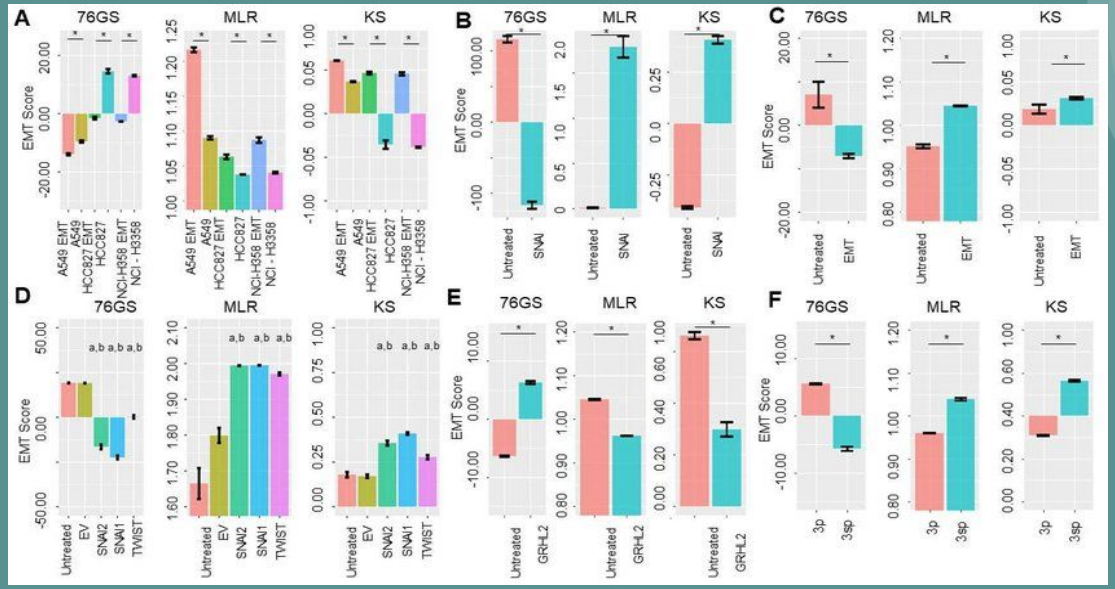
- Space Data / https://www.spacex.com
- Wikipedia / https://es.wikipedia.org/wiki/SpaceX
- Nasa / https://www.nasa.gov/spacex
- YouTube / https://www.youtube.com/spacex

https://www.abc.es/ciencia/abci-spacex-planea-primera-caminata-turistas-espaciales-para-este-202202161323_noticia.html

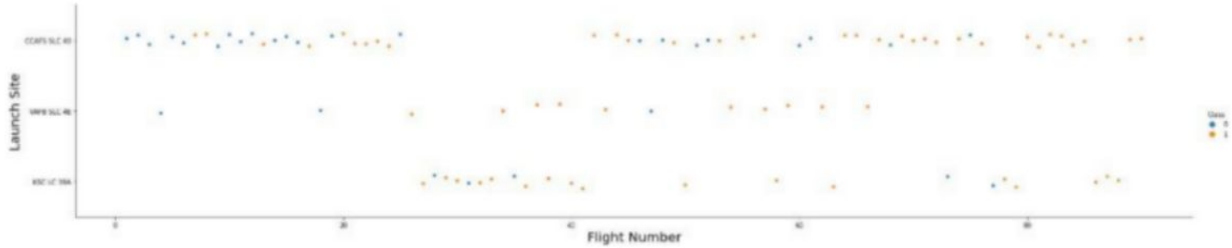
Results:

E.D.A. with Visualizations:

- Exploratory Data Analysis Seaborn plots -



Flight number / Launch site:



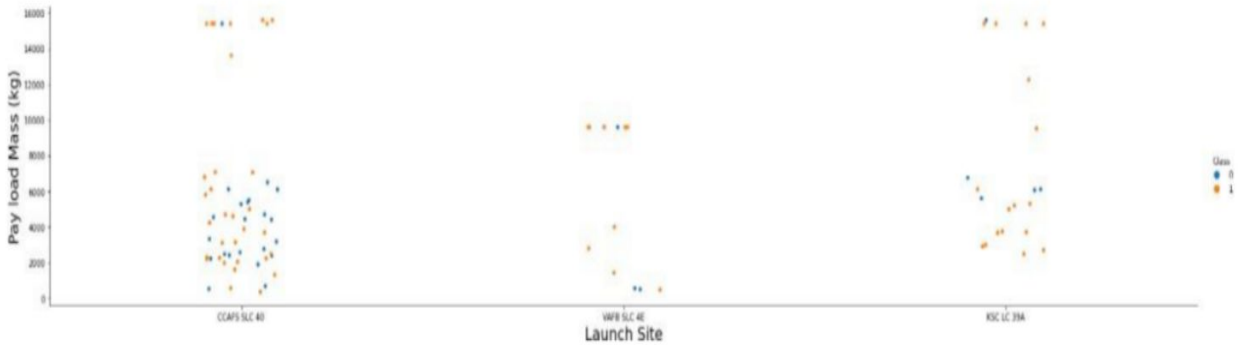
Orange indicates successful launch; Blue indicates unsuccessful launch.

Graphic suggests an increase in success rate over time (indicated in Flight Number).

Likely a big breakthrough around flight 20 which significantly increased success rate.

CCAFS appears to be the main launch site as it has the most volume.

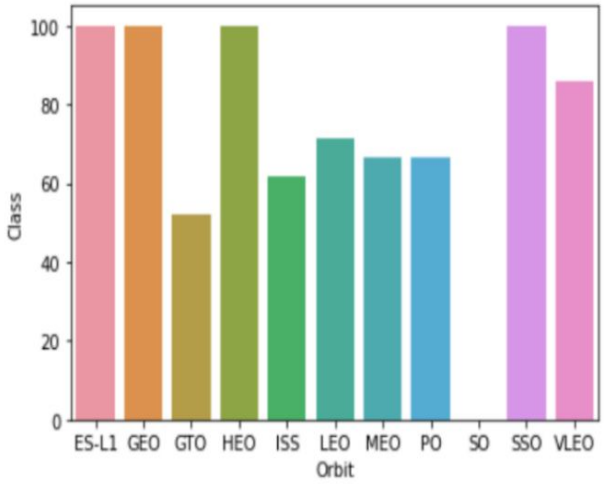
Payload / Launch site:



Orange indicates successful launch; Blue indicates unsuccessful launch.

Payload mass appears to fall mostly between 0-7000 kg.
Different launch sites also seem to use different payload mass.

Success Rate / Orbit type:



Success Rate Scale with %

ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)

SSO (5) has 100% success rate

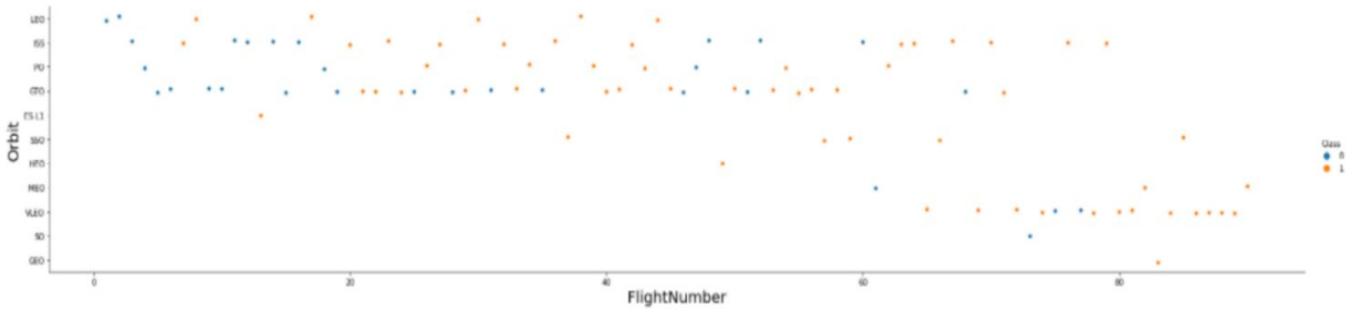
VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

Flight number / Orbit type:

Flight Number vs. Orbit type



Orange indicates successful launch; Purple indicates unsuccessful launch.

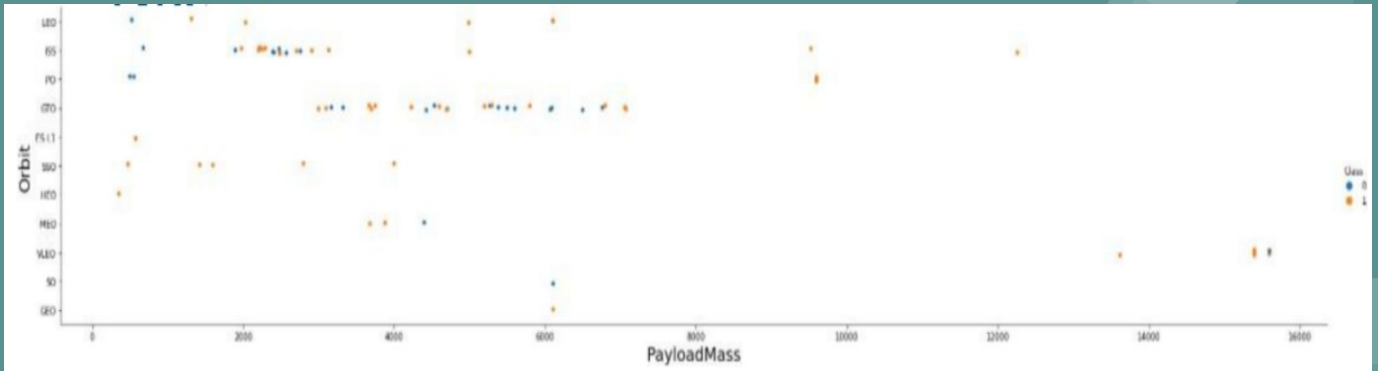
Launch Orbit preferences changed over Flight Number.

Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload / Orbit type:



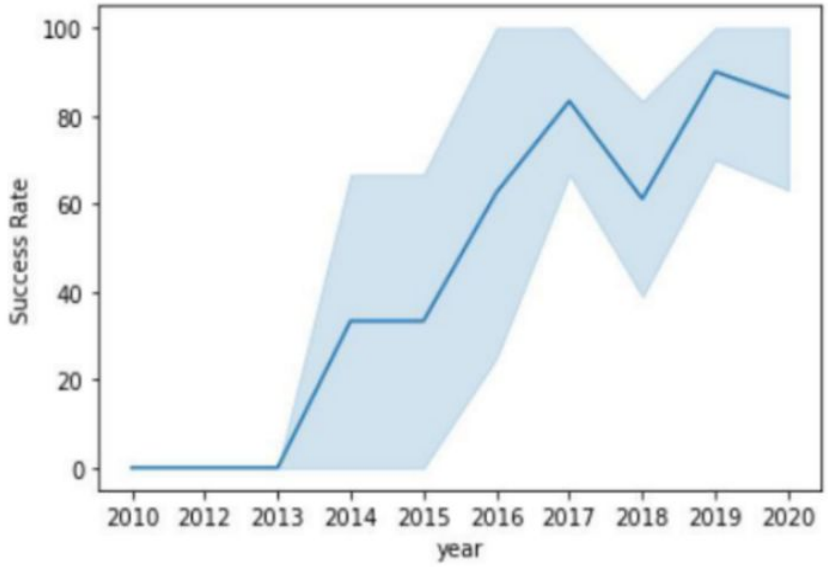
Orange indicates successful launch; Purple indicates unsuccessful launch.

Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch success / Yearly trend :



95% confidence interval
(light blue shading)

Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

E.D.A. with SQL:

Exploratory Data Analysis SQL DB2 Python SQL Alchemy

All launch site names:

Display the names of the unique launch sites in the spac

```
n [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8
d:32733/BLUDB
Done.
```

ut[10]:

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

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name. Likely only 3 unique launch_site

Launch site names beginning with 'CCA':

```
In [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

Out[16]:

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

We can see five entries beginning with 'CCA'

Total Payload Mass from 'NASA':

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

Total Payload Mass (kg) where NASA is a customer.

Average Payload Mass by F9v1.1:

```
In [18]: %sql select avg(payload_mass__kg_) as Average from SPACE
         XTBL where booster_version like 'F9 v1.1%'

          * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
         e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
         32733/BLUDB
         Done.
```

Out[18]:

average
3226

This query calculates the average payload mass which used booster version F9v1.1 Avg.payload mass of F9v1.1 is on the low end of our payload mass range

Successful drone ship landing with payload between 4000 / 6000:

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databas
Done.
```

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

Successful drone ship landings on the quoted range.

Mission outcome total number:

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

Count of each mission outcome.
SpaceX appears to achieve its mission outcome almost 99%.
So, most of landings failures are intended.
Remarkable enough, a launch has an uncertain payload status and one failed in flight course.

Booster that carried maximum payload:

```
]: maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
   maxv = maxm[0][0]

   %sql select booster_version from SPACEXTBL where payload
   _mass__kg_=(select max(payload_mass__kg_) from SPACEXTB
   L)
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

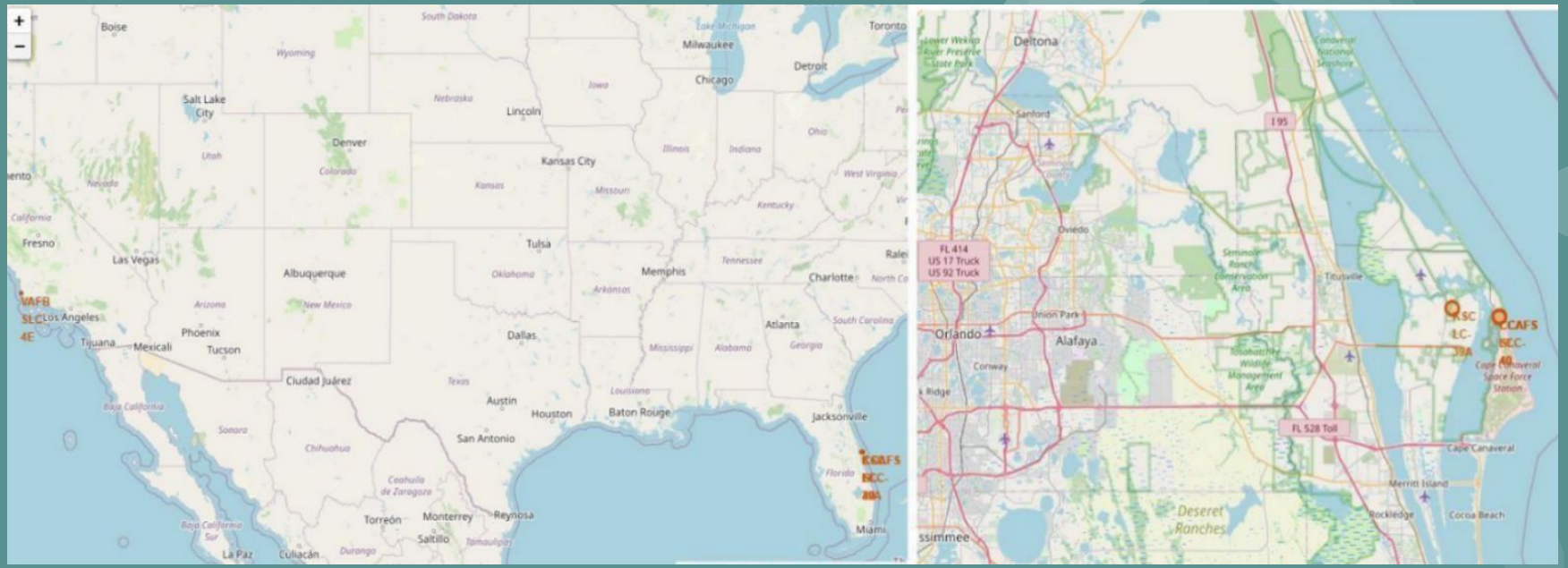
]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

Booster version that carried highest payload mass (15.600 Kg.). It can be noted that these booster versions are quite similar, all of them
F9 B5 B10xx.x kind.
Payload mass correlates with used booster version.

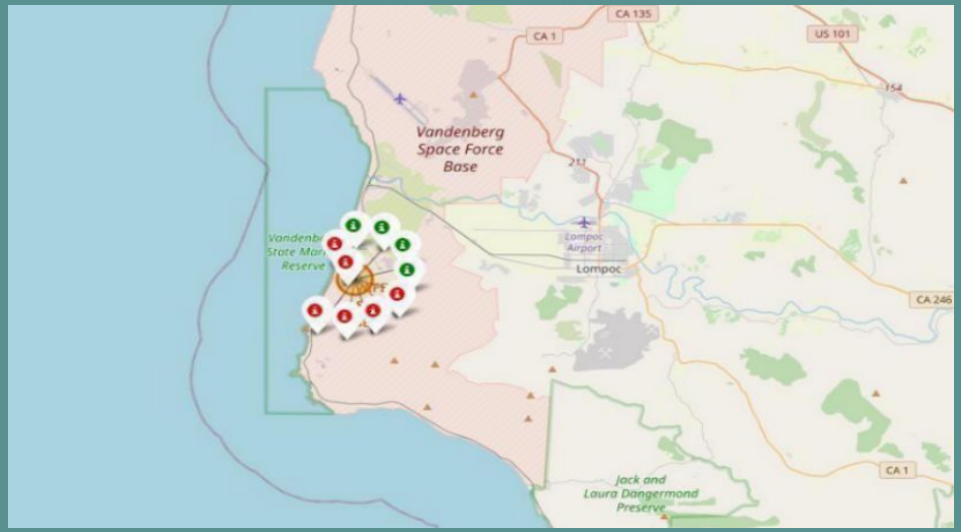
Interactive maps with folium:

Launch site locations:



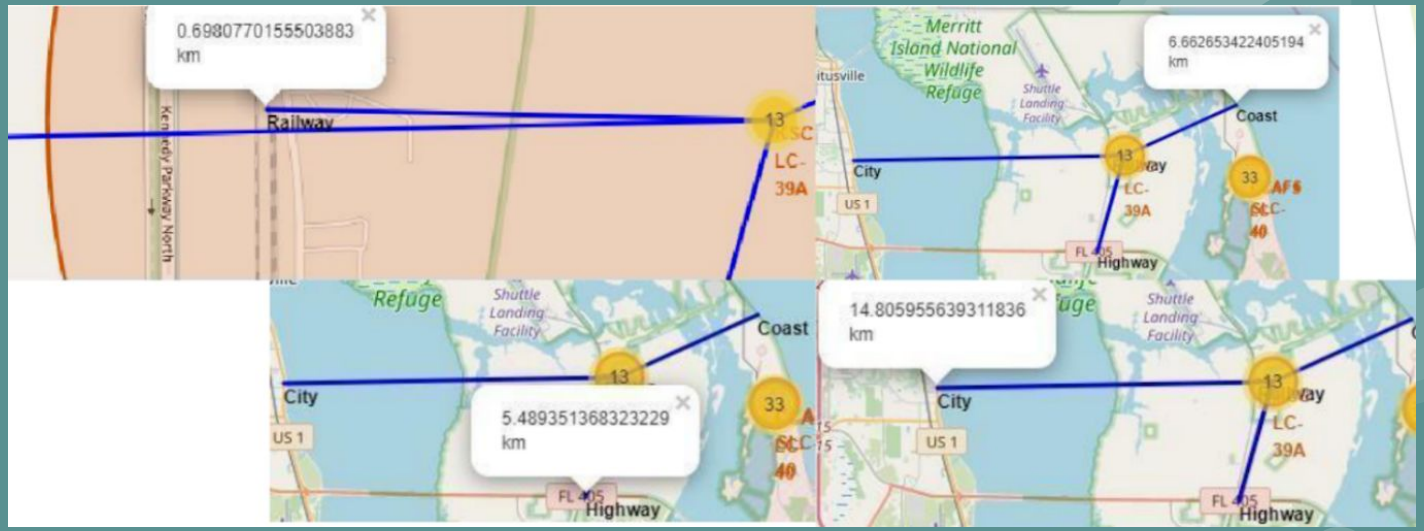
The left map shows all launch sites relative to US map. The right map shows the two Florida launch since they are so close each other. All launching sites are so close to the Ocean.

Color-code launch markers :



Clusters on folium map can be clicked on to display each successful landing (green icon) and failed landing (red). Here, VAFB shows four successful landings and six failed ones.

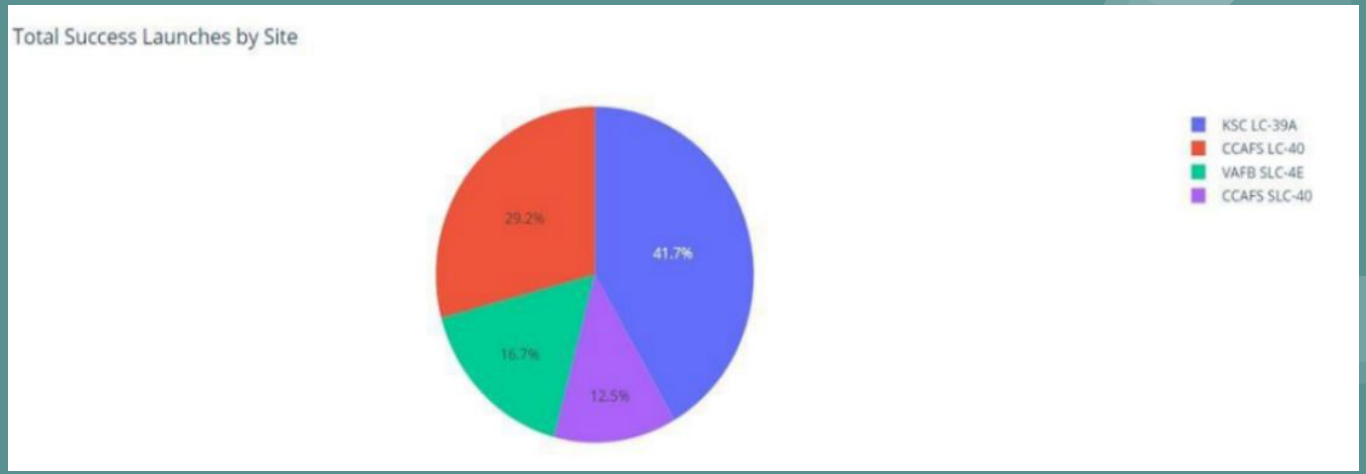
Key location proximities :



By using KSC LC-39A, for instance, launched sites are so close to railways for large part and supply transportation. Launch places located too to coasts and relatively so far from cities in order to avoid great disasters in case of launching failures on populated urban areas.

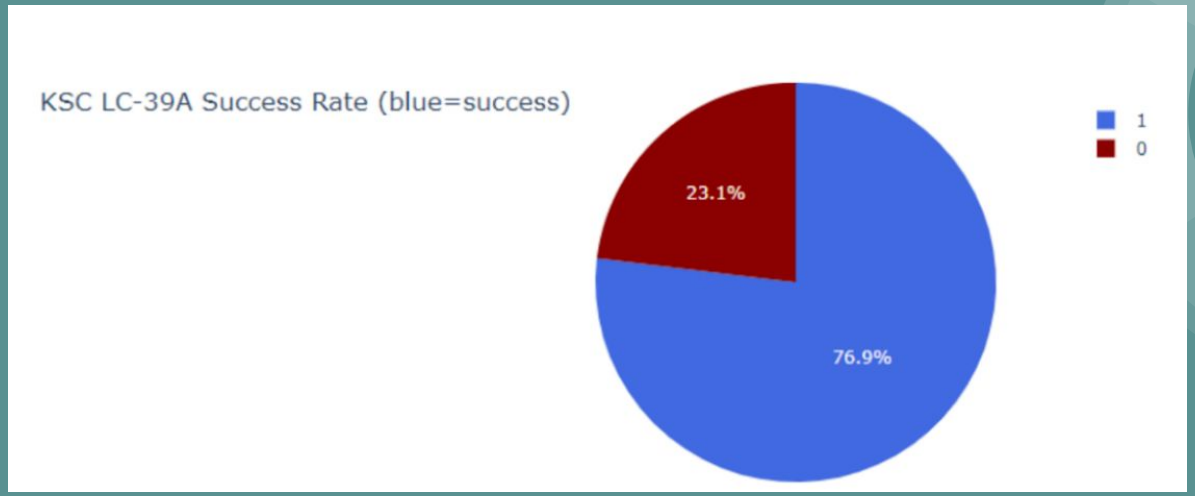
Building Dashboards with Plotly Dash:

Successful launches across launch sites :



Distribution of successful landing across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40. Therefore, CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings have been performed before the name were changed. VAFB has the smallest share of successful landings. This can be based on smaller sample and an increase difficult on launching on the west coast.

Highest success rate of launch site:



KSC LC-39A has the highest rate of success with 10 successful landing and 3 failed attempts. (1: success; 0: failing)

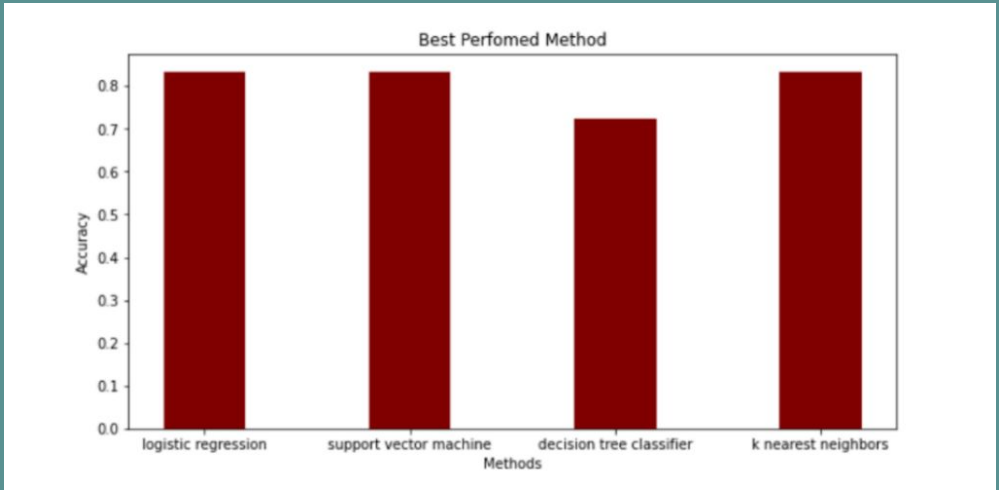
Payload mass / success booster version category:

It can be observed that Payload has a wide range selector. Yet, this is settled from 0 - 10000, not max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot besides accounts for booster version type in color and number of launches in point size. In this particular range of 0 - 7500, interestingly there are two failed landings with payloads of zero kg.

Predictive Analysis (Classification)

-GridsearchCV (CV=10) Logistic Regression SVM Decision Tree KNN-

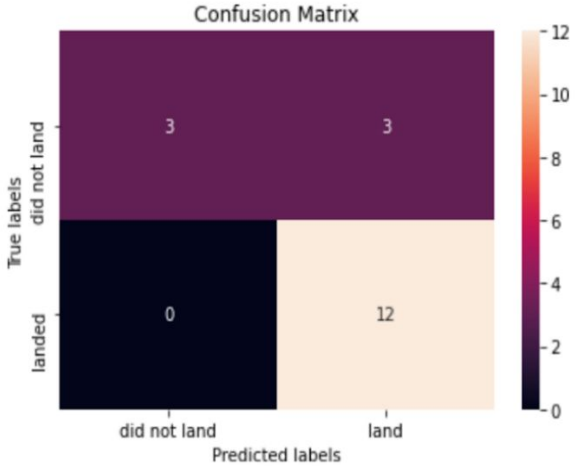
Classification accuracy:



The models have had the same accuracy on the tests set at 83.33% accuracy, but the decision tree classifier (72,23%).
This test size is small, sample size = 18.
This can cause large variance in accuracy results, such as those shown in Decision Tree Classifier model in repeated turns. Obviously, much more data is needed to feed the analysis to determine the best model

Confusion Matrix:

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Since all models performed the same result for the test, confusion matrix is the the same across all models.
The models predicted 12 successful landings when the true label was successful landings.
The models predicted 3 unsuccessful landings if true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).