IBM Data Science Capstone Project – SpaceY Price for Each Launch

Denise Jeannie Marti

06/12/2023

OUTLINES

EXECUTIVE SUMMARY

Data Collection using API	We generate a get request to the SpaceX API and a basic data organization and formatting.
Data Collection with Web Scraping	Web data scraping to collect historical Falcon 9 launch records from Wikipedia.
Data Wrangling	Find some patterns in the data and determine what the label would be for training supervised models.
Exploratory Data Analysis with SQL	Analyze first-stage feedback data to determine the cost of a launch.
Exploratory Data Analysis with Data Visualization	Conducting exploratory data analysis and resource engineering to predict the successful return of the first stage.
Interactive Visual Analytics with Folium	Interactive visual analytics using Folium.
Machine Learning Prediction (Predictive analysis - Classification)	Creating a machine learning pipeline to predict whether the former will be successful in return.

INTRODUCTION

As a data scientist working for a new rocket company, my goal is to determine the price for each launch of SPACE X.

• Collecting information about the Space;
• Demonstrating data through dashboards for the team;
• Seeking data on reuse of the first stage of SpaceX rockets.
• Training a machine learning model to determine whether or not the first stage successfully lands using the public information.

METHODOLOGY

METHODOLOGY

- SpaceX launch datasets collected via SpaceX REST API, giving us data such as:
  - Launch, rockets used, useful carga, launch sites, launch and landing specifications and whether or not there was success in returning the first phase of the rocket.
- Using this data we can predict what are the chances of successful return of the first phase of the rocket.

SpaceX REST API endpoints, or URLs, start with api.spacexdata.com/v4/.

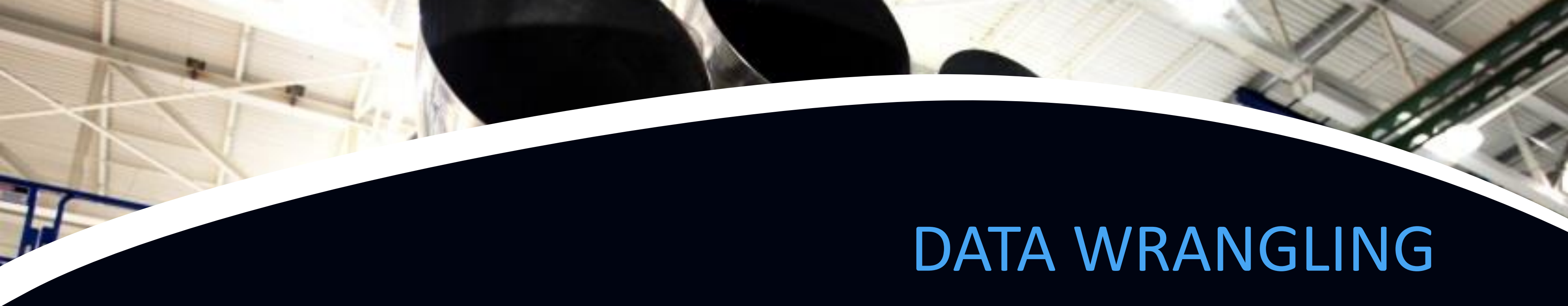
- Another data source to obtain Falcon 9 launch data is Wikipedia using Beauty Soup.

DATA COLLETION – SPACEX API

- Request data from SpaceX API (rocket launch data) ;
- Decode response using .json() and convert to a dataframe using .json_normalize();
- Request information about the launches from SpaceX API using custom functions;
- Create dictionary from the data;
- Create dataframe from the dictionary;
- Filter dataframe to contain only Falcon 9 launches;
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file.

DATA COLLECTION - Web Scraping

- Request data (Falcon 9 launch data) from Wikipedia;
- Create BeautifulSoup object from HTML response;
- Extract column names from HTML table header;
- Collect data from parsing HTML tables ;
- Create dictionary from the data ;
- Create dataframe from the dictionary ;
- Export data to csv file.



DATA WRANGLING

- Perform EDA and determine data labels;
- Calculate number of:
  - launches for each site, occurrence of orbit, and occurrence of mission outcome per orbit type;
- Create binary landing outcome column (dependent variable)
- Export data to csv file

DATA WRANGLING– Landing Outcoming

- Landing was not always successful.
- True Ocean: successful landing to a specific region of the ocean.

Basic Information:

- ✓ False Ocean: an unsuccessful landing to a specific region of ocean;
- ✓ True RTLS: had a successful landing on a ground pad;
- ✓ False RTLS: an unsuccessful landing on a ground pad;
- ✓ True ASDS: had a successful landing on a drone ship;
- ✓ False ASDS: an unsuccessful landing on drone ship;
- ✓ Outcomes converted into 1 for a successful landing and 0 for an unsuccessful landing.

EDA WITH VISUALIZATION - CHARTS

Flight Number vs. Payload

Flight Number vs. Launch Site

Payload Mass (kg) vs. Orbit type

Payload Mass (kg) vs. Launch Site

Orbit VS. Flight Number

Orbit VS. Payload Mass

Payload VS. Orbit Type

EDA WITH SQL

SQL queries were used to get answers on the dataset such as:

- ✓ Display the names of unique launch sites in the space mission;
- ✓ Display 5 launch site records that start with the string 'KSC';
- ✓ Display the total mass of payload carried by NASA-Launched thrusters (CRS);
- ✓ Display the average load mass carried by the F9 v1.1 booster version;
- ✓ Date of first successful landing on ground pad
- ✓ List names of boosters that were successful on the ground pad with charge mass greater than 4000 and less than 6000;
- ✓ Total list of successful and unsuccessful missions;
- ✓ List the names of booster versions that carried the maximum payload mass;
- ✓ Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015;
- ✓ Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc) ;
- ✓ Classification.

MAP WITH FOLIUM

- Markers Indicating Launch Locations
  Added blue circle at NASA's Johnson Space Center coordinate with a pop-up label showing name and its latitude and longitude coordinates.
  Added red circles to all launch location coordinates with a pop-up label showing name and their latitude and longitude coordinates.
- Map with folium
  Colored launch results markers;
  Added colored markers of successful (green) and unsuccessful (red) launches to each launch site to show which launch sites have high success rates;
  Distances between a launch site and surroundings;
  Added colored lines to show the distance between the CCAFS SLC40 launch site and its proximity to nearby coastline, railway, highway and city.

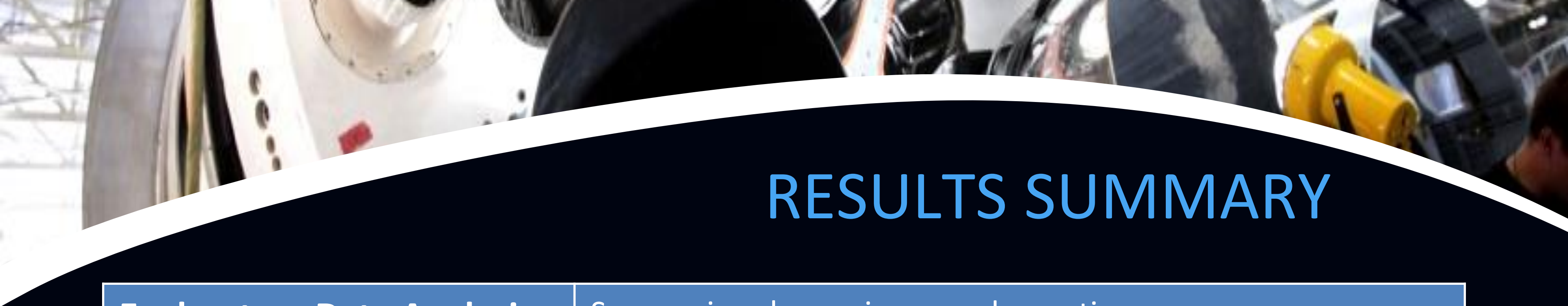
DASHBOARD WITH PLOTY DASH

- Dropdown list: with launch sites to allow user to select all launch sites or a particular launch site.
- Dashboard with Plotly Dash: Payload mass range slider to allow user to select payload mass range.
- Pie Chart: Showing successful launches to allow the user to see successful and unsuccessful launches as a percentage of the total.
- Scatter plot: showing payload mass x success rate by boost release to allow user to see correlation between payload and launch success.

PREDICTIVE ANALYSIS

- ✓ Create a NumPy array with Pandas from the Class column;
- ✓ Standardize data with StandardScaler by fitting and transforming data;
- ✓ Split the data using train_test_split;
- ✓ Create a GridSearchCV object with cv=10 for parameter optimization;
- ✓ Apply GridSearchCV in different algorithms: logistic regression
- ✓ (LogisticRegression()), support vector machine (SVC()), decision tree;
- ✓ (DecisionTreeClassifier()), K-nearest neighbor (KNeighborsClassifier());
- ✓ Calculate the accuracy of the test data using .score() for all models;
- ✓ Evaluate the confusion matrix for all models;
- ✓ Identify the best model using SVM, Classification Trees and Logistic Regression

RESULTS

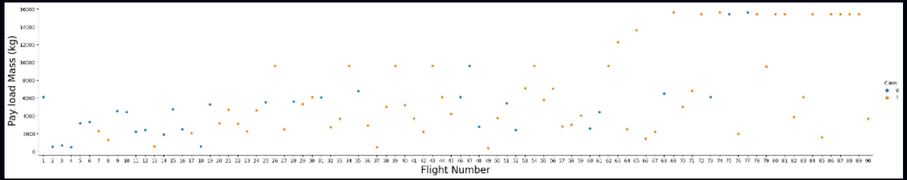


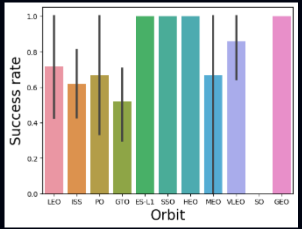
RESULTS SUMMARY

Exploratory Data Analysis	Success in releases improved over time; • KSC LC-39A has the highest success rate among landing sites • ES-L1, GEO, HEO and SSO orbits have a 100% success rate
Visual Analytics	Launch sites are mostly near the equator and near the coast; Launch sites are far enough away to avoid damage (city, highway, rail) should a launch fail, but far enough away to bring people and materials to support launch activities.
Predictive Analytics	Decision Tree model is the best predictive model for the dataset.

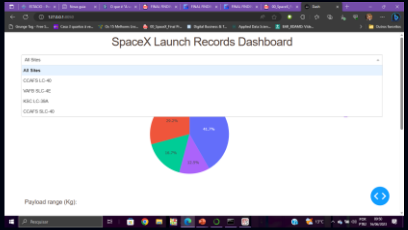
RESULTS SUMMARY cont.

EDA Results





Int. Analytics demo

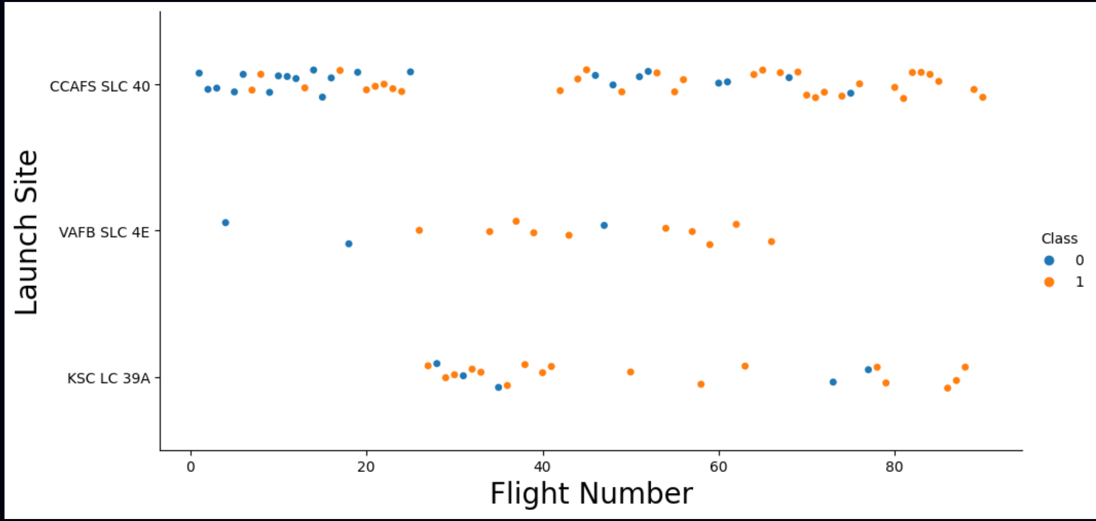


Predictive Results

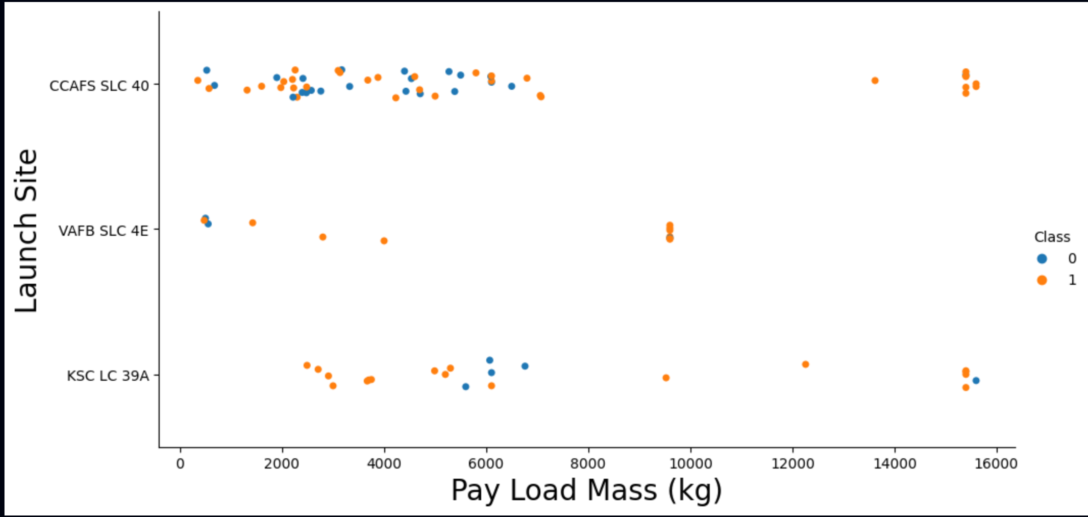
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.83333
KNN	0.84821	0.83333

FLIGHT NUMBER VS. LAUNCH SITE

- Earlier flights had a lower success rate (blue = fail(0))
- Later flights had a higher success rate (orange = success(1))
- Around half of launches were from CCAFS SLC 40 launch site.
- For launch site ‘KSC LC 39A’, it takes at least around 25 launches before a first successful launch.

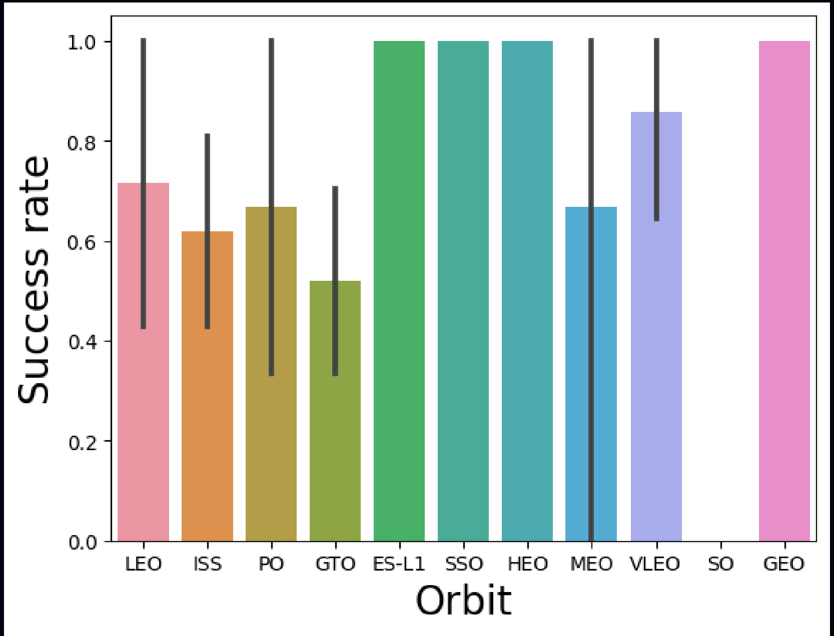


PAYLOAD VS. LAUNCH SITE



- Typically, the higher the payload mass (kg), the higher the success rate .
- Percentage of successful launch increases for launch site ‘VAFB SLC 4E’ as the payload mass increases.
- Most launces with a payload greater than 7,000 kg were successful .
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg .
- For launch site ‘VAFB SLC 4E’, there are no rockets launched for payload greater than 10,000 kg.

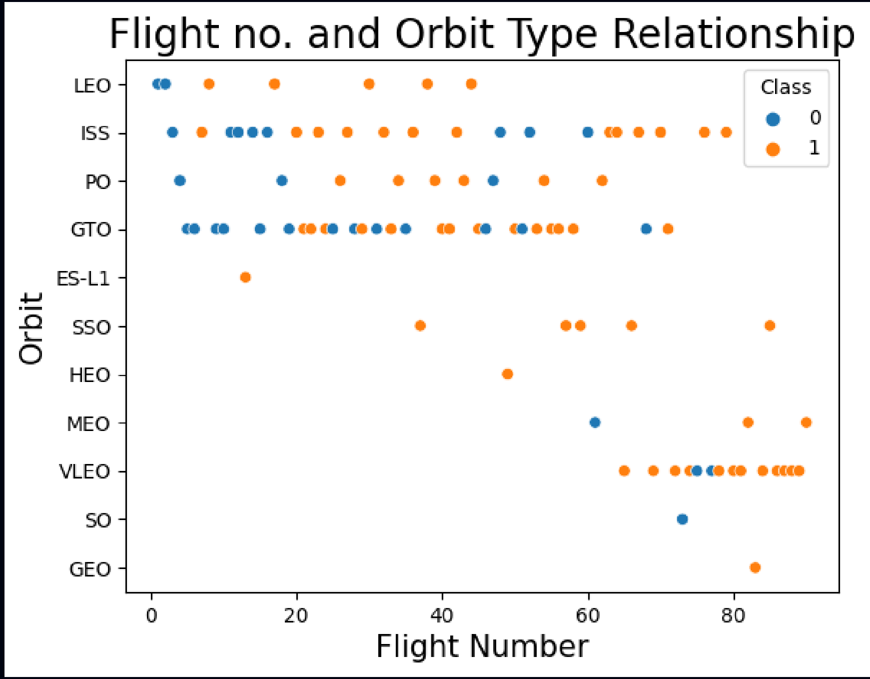
SUCCESS RATE VS. ORBIT TYPE

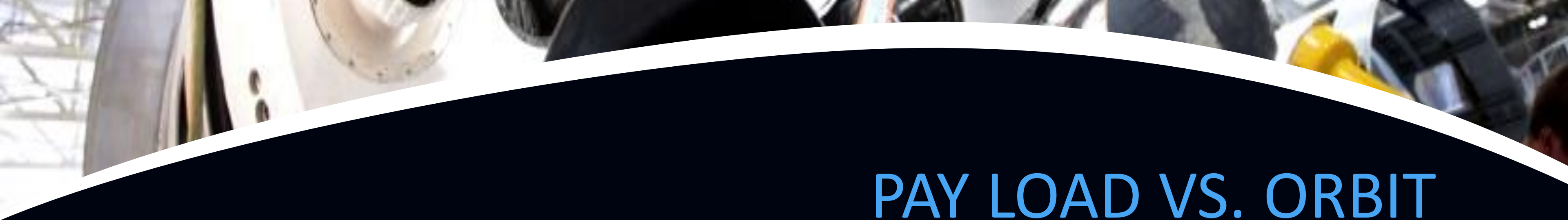


- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO
- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates.

FLIGHT NUMBER VS. ORBIT

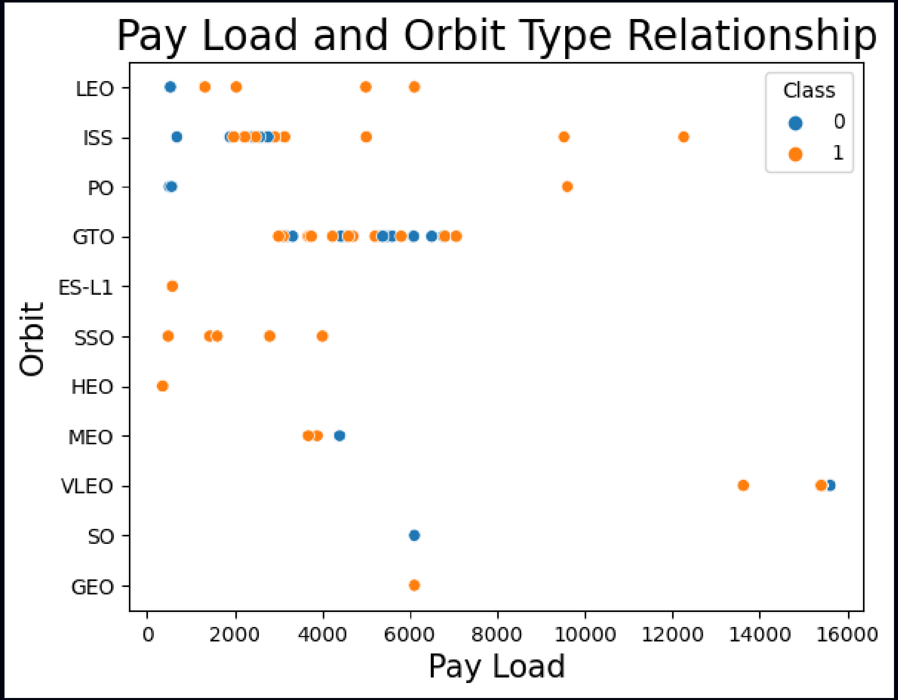
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers
- This relationship is highly apparent for the LEO orbit
- There is no relationship between flight number and orbit for GTO
- For orbit VLEO, first successful landing (doesn't occur until 60+ number of flights



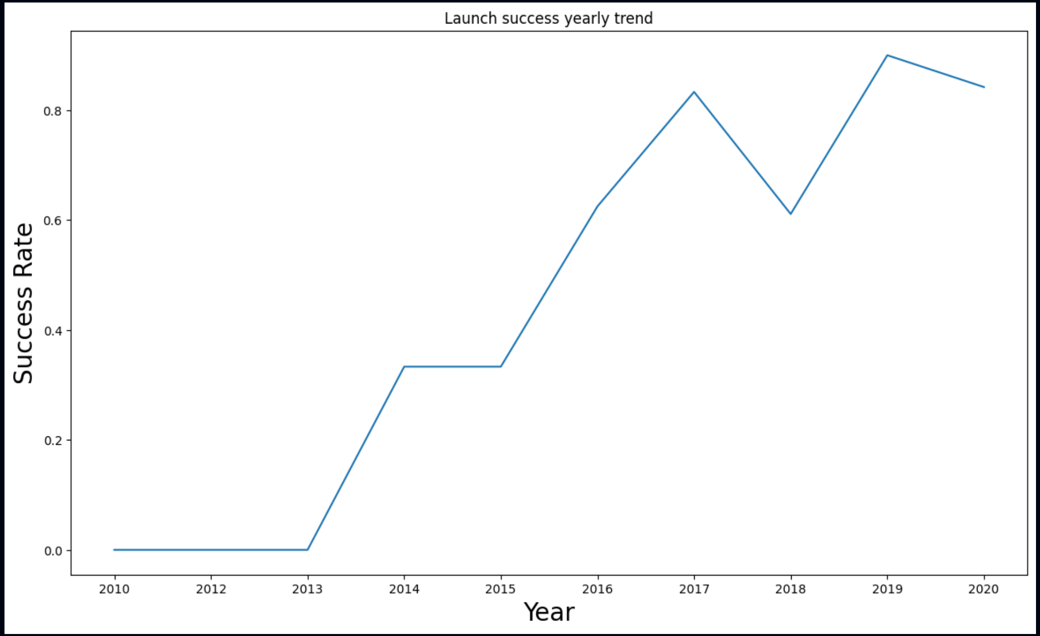


PAY LOAD VS. ORBIT

- ➢ Successful landing rates appear better with pay load for orbits LEO, ISS and PO.
- ➢ • The GTO orbit has mixed success with heavier payloads.



LAUNCH SUCCESS YEARLY TREND



- Success rate increased by about 80% between 2013 and 2020
- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013

LAUNCH SITE INFORMATION

[12]:
```
%sql select Launch_Site from SPACEXTBL GROUP BY Launch_Site;
```

```
 * sqlite:///my_data1.db
Done.
```

[12]:

Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

LAUNCH SITE INF. WITH CCA

```
[15]: %sql SELECT * from SPACEXTBL where LAUNCH_SITE LIKE ('CCA%') LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

PAYLOAD MASS

Total Payload Mass (kg) carried by boosters launched by NASA

```
[20]: %sql SELECT SUM (PAYLOAD_MASS__KG_) as payloadmass FROM SPACEXTBL WHERE customer='NASA (CRS)';

 * sqlite:///my_data1.db
Done.

[20]: payloadmass
          45596.0
```

Average Payload Mass (kg) carried by booster version F9 v1.1

```
[29]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmasskg from SPACEXTBL where booster_version = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

[29]: payloadmasskg
           2928.4
```

FIRST SUCCESSFUL LANDING & INFOS

1st Successful Landing in Ground Pad

```
[39]: %sql select date as first_successful_landing from SPACEXTBL where landing_outcome LIKE 'Success (ground pad)' order by first_successful_landing desc limit 1;
 * sqlite:///my_data1.db
Done.
```

[39]:

first_successful_landing
22/12/2015

Booster Drone Ship Landing

```
[47]: %sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ as PaylossMassKG\
FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE '%drone ship%' \
AND MISSION_OUTCOME = 'Success' AND PaylossMassKG BETWEEN 4000 AND 6000;
 * sqlite:///my_data1.db
Done.
```

[47]:

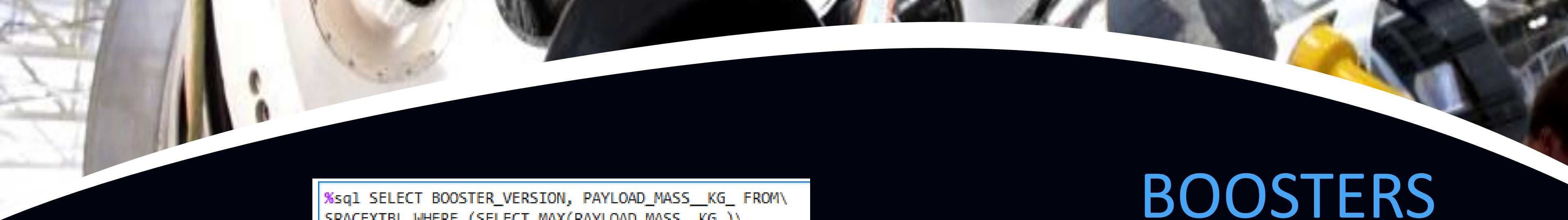
Booster_Version	PaylossMassKG
F9 FT B1020	5271.0
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

TOTAL NUMBER OF SUCCESSFUL AND FAILED MISSION OUTCOMES

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	TOTAL
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1



BOOSTERS

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM\
SPACEXTBL WHERE (SELECT MAX(PAYLOAD_MASS__KG_)\
FROM SPACEXTBL) = PAYLOAD_MASS__KG_;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

- The query returned the maximum mass by versions where the maximum is 15600.

FAILED LANDINGS ON DRONE SHIP

```
%sql SELECT substr(Date,4,2) as month, Date, BOOSTER_VERSION,\
LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE \
[LANDING_OUTCOME] = 'Failure (drone ship)' \
AND substr(Date,7,4)= '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The query lists landing outcome, booster version, and the launch site where landing outcome is failed in drone ship and the year is 2015.

COUNT OF SUCCESSFUL LANDINGS

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes FROM SPACEXTBL\
WHERE DATE between '04-06-2010' and '20-03-2017'\
group by [Landing_Outcome] order by count_outcomes DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- Count in descending order of landing results between 04-06-2010 and 20-03-2017.

LAUNCH SITE ANALYSIS

LAUNCH SITE ANALYSIS

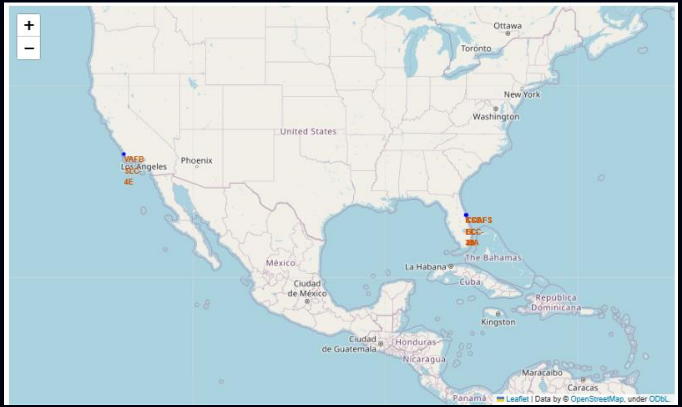


Fig 1 – Global View

Fig 2 – Zoom View: CCAFS LC-40 (FL), KSC LC-39A (FL) e CCAFS SLC-40 (FL).

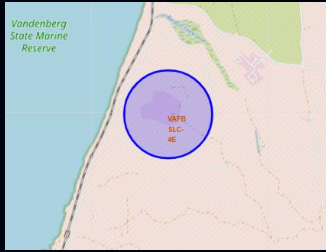
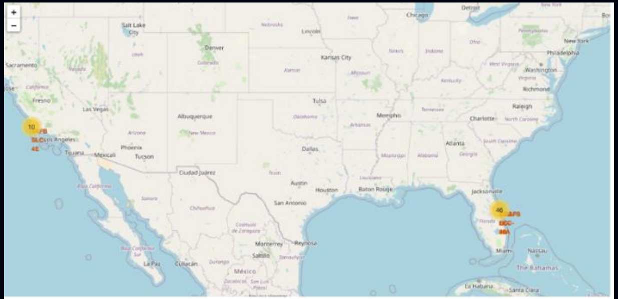


Fig 3 – Zoom VAFB SLC-4E (CA)

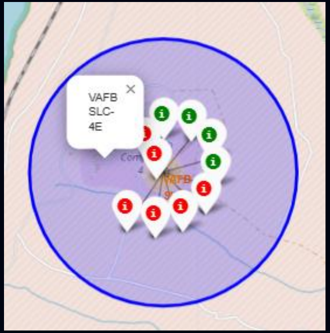
The closer the launch site is to the Equator, the greater the Earth's natural thrust which, due to the rotational speed, helps to save on spending on fuels and boosters for a progressive orbit.

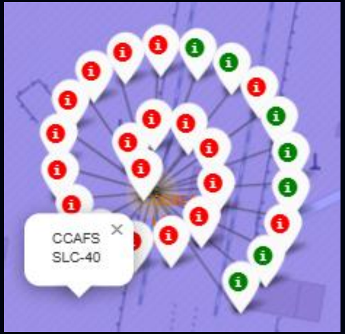
LAUNCH OUTCOMES



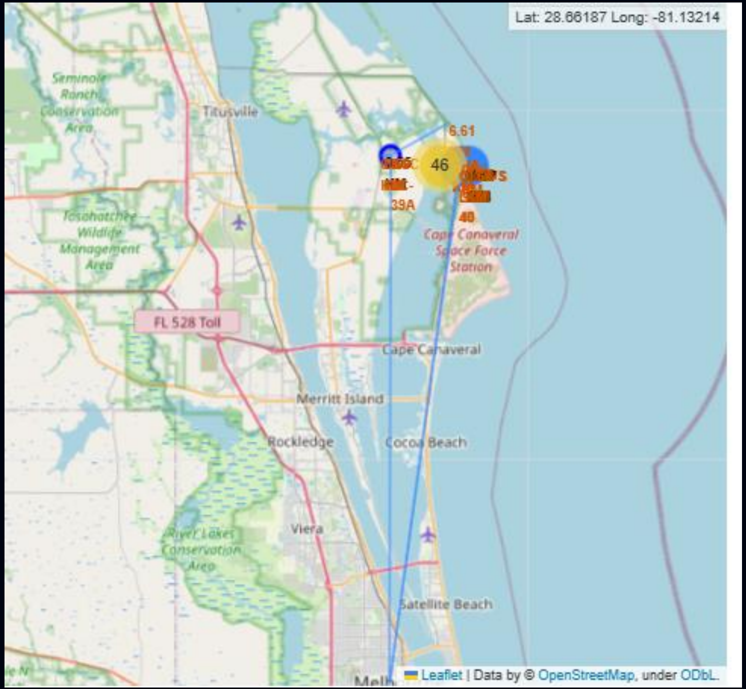
US map with all Launch Sites

- The picture below are zoom each site displaying success (green) and failure (red) markers.
- Looking at each sitemap, the KSC LC-39A Launch Site has the most successful launches.





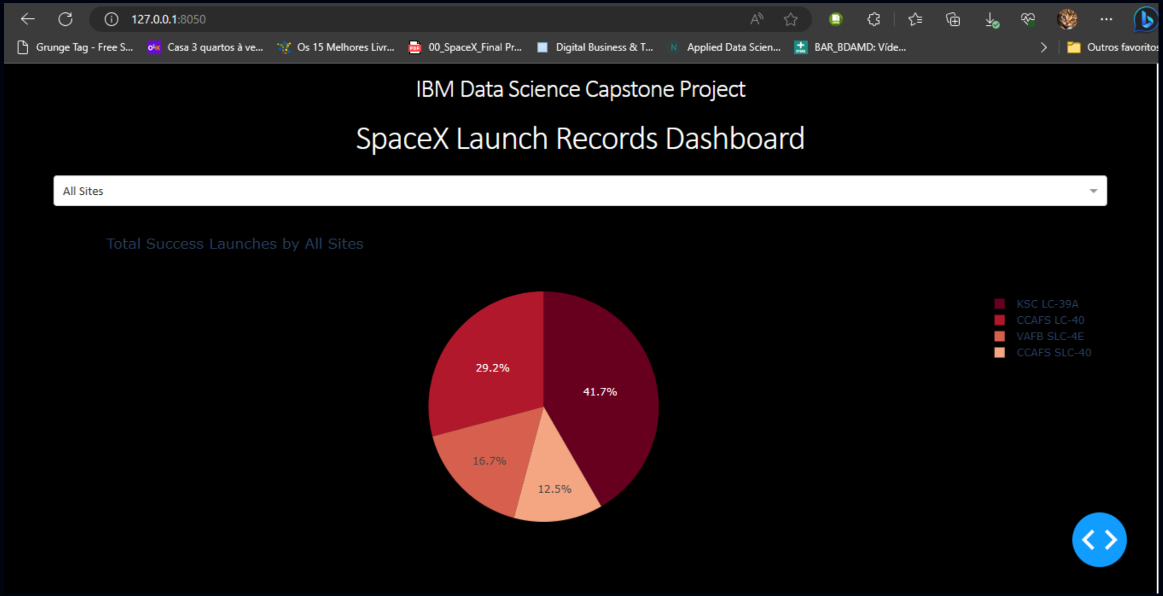
DISTANCE TO PROXIMITIES



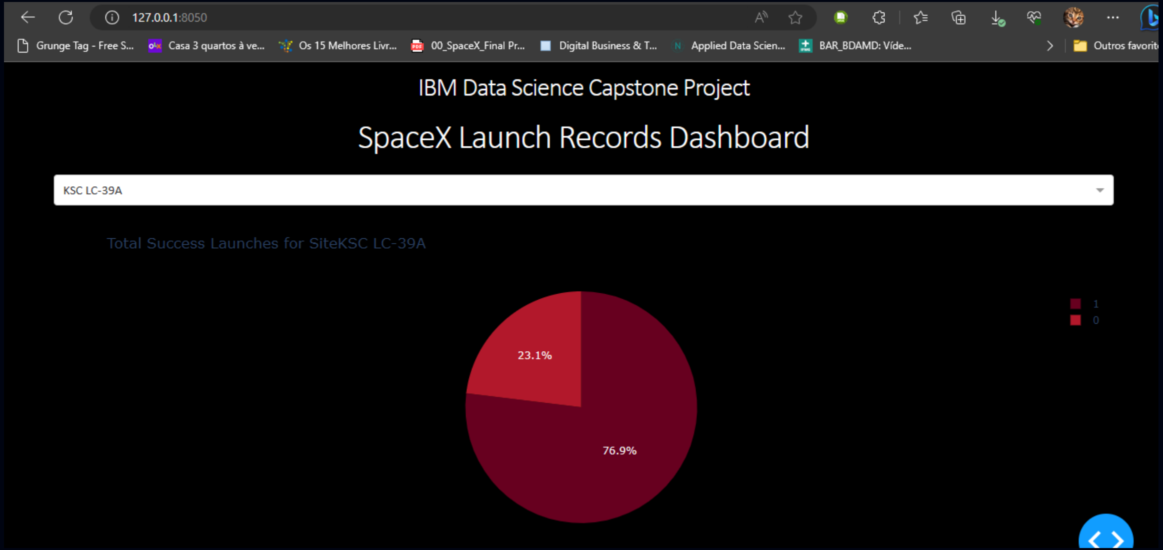
- CCAFS SLC-40
- 0.86 km from nearest coastline
- 1.38 km from nearest railway
- 56.82 km from Melbourne

SPACEX

DASHBOARD



The KSC LC-39A has 41.7% of the most successful launches when compared to the other sites.



KSC LC-39A has the highest success rate among launch sites at 76.9%.

BoosterVersion

Payloads between 2000 kg and 5000 kg have the highest success rate

PREDICTIVE ANALYSIS (CLASSIFICATION)

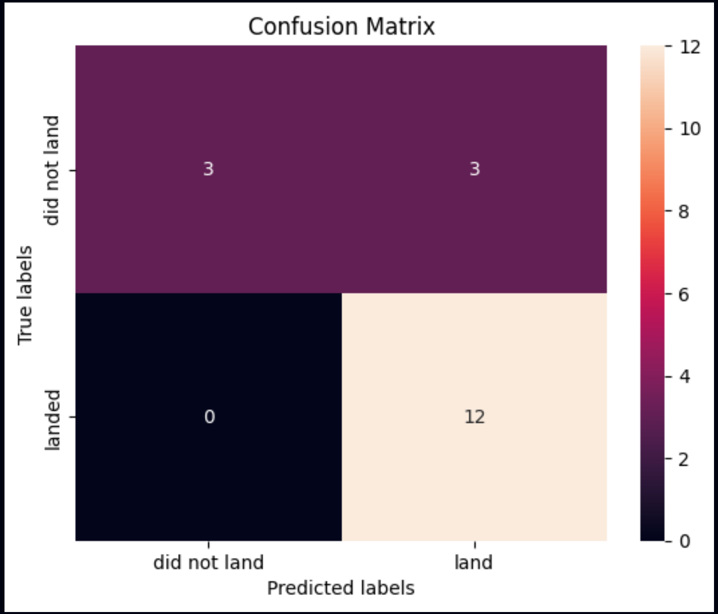
CLASSIFICATION ACCURACY USING TRAINING DATA

```
print("Model\t\tAccuracy\tTestAccuracy")#,logreg_cv.best_score_)
print("LogReg\t\t{}\t\t{}".format((logreg_cv.best_score_).round(5), logreg_cv.score(X_test, Y_test).round(5)))
print("SVM\t\t{}\t\t{}".format((svm_cv.best_score_).round(5), svm_cv.score(X_test, Y_test).round(5)))
print("Tree\t\t{}\t\t{}".format((tree_cv.best_score_).round(5), tree_cv.score(X_test, Y_test).round(5)))
print("KNN\t\t{}\t\t{}".format((knn_cv.best_score_).round(5), knn_cv.score(X_test, Y_test).round(5)))

comparison = {}

comparison['LogReg'] = {'Accuracy': logreg_cv.best_score_.round(5), 'TestAccuracy': logreg_cv.score(X_test, Y_test).round(5)}
comparison['SVM'] = {'Accuracy': svm_cv.best_score_.round(5), 'TestAccuracy': svm_cv.score(X_test, Y_test).round(5)}
comparison['Tree'] = {'Accuracy': tree_cv.best_score_.round(5), 'TestAccuracy': tree_cv.score(X_test, Y_test).round(5)}
comparison['KNN'] = {'Accuracy': knn_cv.best_score_.round(5), 'TestAccuracy': knn_cv.score(X_test, Y_test).round(5)}
```

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.83333
KNN	0.84821	0.83333

The accuracy is extremely close, the models performed approximately the same level of performance and had the same scores and accuracy but, the Decision Tree model outperformed the rest when looking at .best_score_

CONFUSION MATRICES



The confusion matrix, which summarizes the performance of a classification algorithm in this case, had identical results in all matrices. Both had a false positive (type 1 error), which is not good.

Results:

- 12 true positives
- 3 true negatives
- **3 false positives**
- 0 False Negative

CONCLUSION

CONCLUSION

KSC LC-39A:

- Highest success rate among launch sites.
- 100% success rate for launches under 5,500 kg.

EQUATOR:

- Most launch sites are close to the Equator, because the Earth's rotation speed is greater than in any other part of the planet - and this makes the rockets gain an extra boost, saving fuel.

ORBITS:

- ES-L1, GEO, HEO and SSO have a 100% success rate

<u>LAUNCH</u>

- Launch success Increases over time
- All the launch sites are close to the coast
- At all launch sites, the greater the payload mass (kg), the higher the success rate.
- ES-L1, GEO, HEO and SSO have a 100% success rate