DATA SCIENCE CAPSTONE PROJECT

Nicholas Rebello

2024/03/03

A G E N D A

- **Executive Summary**
- **Introduction**
- **Methodology**
- **Results**
- **Concluding Remarks**
- **Appendix**

Executive Summary

List of Methodologies Used:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
  - With Data Visualization
  - With SQL
- Exploratory Data Analysis
- Building an Interactive Map Using Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis

Results:

- Exploratory Data Analysis
- Interactive Analytics Demo
- Predictive Analysis

INTRODUCTION

PROJECT BACKGROUND & GENERAL QUESTIONS

SpaceX leads the commercial space industry with affordable Falcon 9 rocket launches, costing $62 million compared to competitors' $165 million. Their key advantage is reusing the first stage, significantly reducing costs. Predicting first stage reuse, using machine learning and public data, will optimize launch cost estimates and enhance decision-making in space travel.

---

1. How do payload mass, launch site, flights, and orbits impact first stage landing success?

2. Does the rate of successful landings rise over time?

3. Which algorithm is optimal for binary classification here?

METHODOLOGY

MEET OUR EXTENDED TEAM

TAKUMA HAYASHI
President

MIRJAM NILSSON
Chief Executive Officer

FLORA BERGGREN
Chief Operations Officer

RAJESH SANTOSHI
VP Marketing

GRAHAM BARNES
VP Product

ROWAN MURPHY
SEO Strategist

ELIZABETH MOORE
Product Designer

ROBIN KLINE
Content Developer

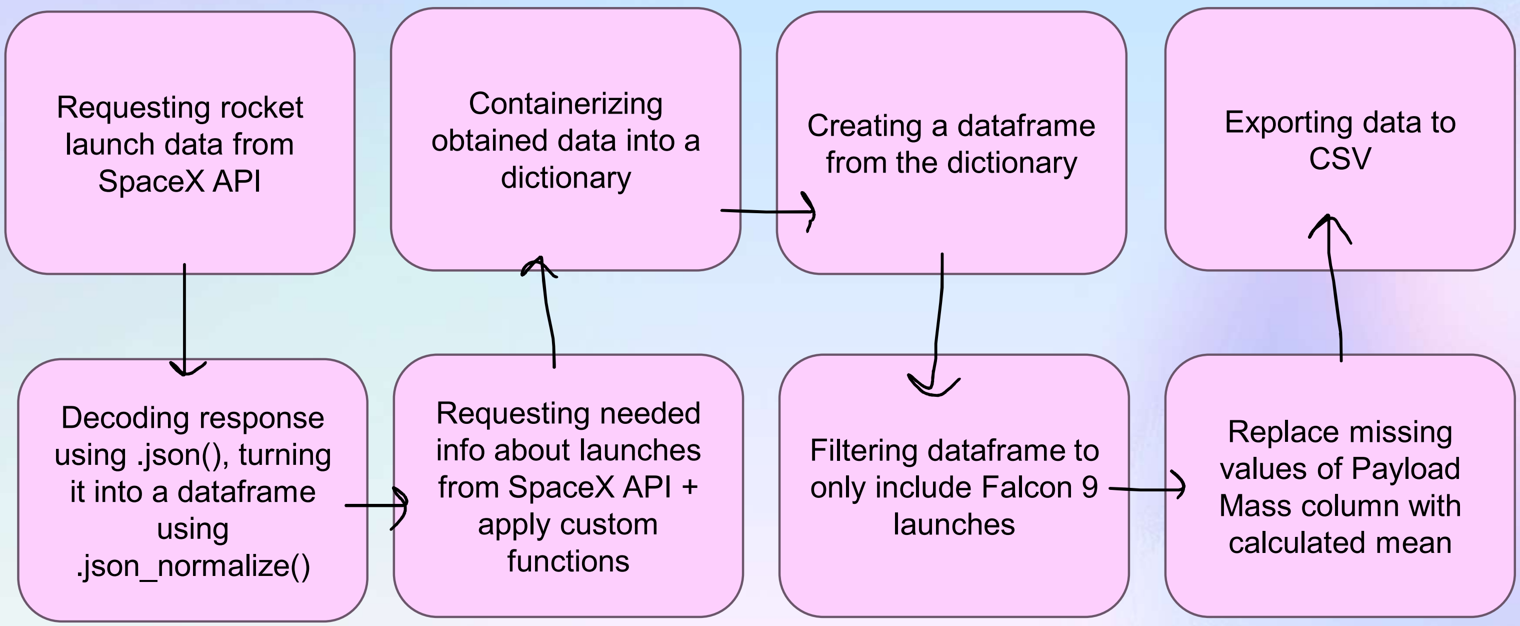
DATA COLLECTION

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.
We had to use both data collection methods to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite,Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

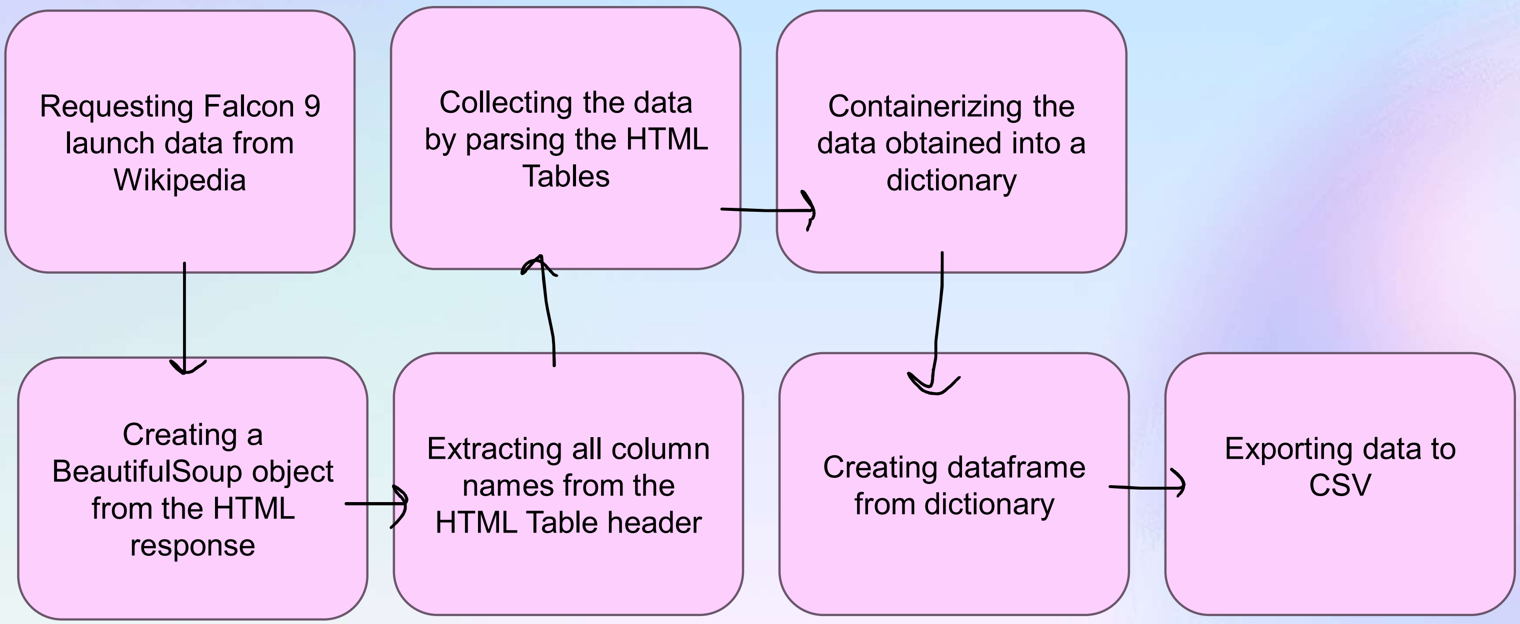
Data Columns are obtained by using Wikipedia Web Scraping:
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

DATA COLLECTION - API



Methodology Flow Chart

DATA COLLECTION – WEB SCRAPING



DATA WRANGLING

The dataset captures diverse scenarios of booster landings, distinguishing between successful and unsuccessful outcomes.

These outcomes are categorized based on specific conditions: whether the landing occurs in designated ocean regions (True/False Ocean), on ground pads (True/False RTLS), or on drone ships (True/False ASDS).

Each condition corresponds to a binary label, with "1" representing a successful landing and "0" indicating an unsuccessful one.

This classification scheme provides a comprehensive framework for analyzing and predicting the success of booster landings in various settings.

EDA WITH DATA VISUALIZATION

Charts were plotted:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots visually depict relationships between variables, aiding in the identification of patterns suitable for machine learning models.

Bar charts, on the other hand, facilitate comparisons between discrete categories, illustrating the relationship between specific categories and corresponding measured values.

Line charts are effective for visualizing trends in data over time, particularly useful for analyzing time series data.

EDA WITH SQL

Several SQL queries were performed on a space mission dataset:

1. Names of unique launch sites were displayed.
2. Five records with launch sites starting with 'CCA' were shown.
3. Total payload mass carried by NASA (CRS) boosters was displayed.
4. Average payload mass carried by booster version F9 v1.1 was calculated.
5. The date of the first successful landing outcome on a ground pad was listed.
6. Boosters with success in a drone ship and payload mass between 4000 and 6000 were listed.
7. The total number of successful and failure mission outcomes was listed.
8. Booster versions that have carried the maximum payload mass were listed.
9. Failed landing outcomes in drone ships, along with their booster versions and launch site names, were listed for the months in the year 2015.
10. The count of landing outcomes between 2010-06-04 and 2017-03-20 was ranked in descending order.

BUILD AN INTERACTIVE MAP WITH FOLIUM

Markers of all Launch Sites:
- Utilized latitude and longitude coordinates to add markers with circles, popup labels, and text labels for each launch site, including NASA Johnson Space Center.
- Displayed geographical locations of all launch sites, highlighting their proximity to the Equator and coastlines through markers with circles, popup labels, and text labels.

Coloured Markers of the launch outcomes for each Launch Site:
- Employed Marker Cluster to add colored markers indicating the success (green) and failure (red) of launches for each launch site.
- This visualization facilitates the identification of launch sites with relatively high success rates.

Distances between a Launch Site to its proximities:
- Incorporated colored lines to depict distances between a launch site (e.g., KSC LC-39A) and its nearby features such as railways, highways, coastlines, and the closest city.
- This visualization enhances understanding of the spatial relationships and proximities of launch sites to various features.

BUILD A DASHBOARD WITH PLOTLY DASH

Launch Sites Dropdown List:
- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):
- Added a pie chart to show the total successful launches count for all sites and the
Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- Added a scatter chart to show the correlation between Payload and Launch Success.

PREDICTIVE ANALYSIS

Creating a NumPy array from the column “class” in data

Standardize data with StandardScaler, fitting the data, transforming data

Splitting Data into testing and training sets with train_test_split function

Creating a GridSearchCV object (cv=10) to find best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN Models

Calculating accurage on test data using .score() on all models

Examining confusion matrix for all models

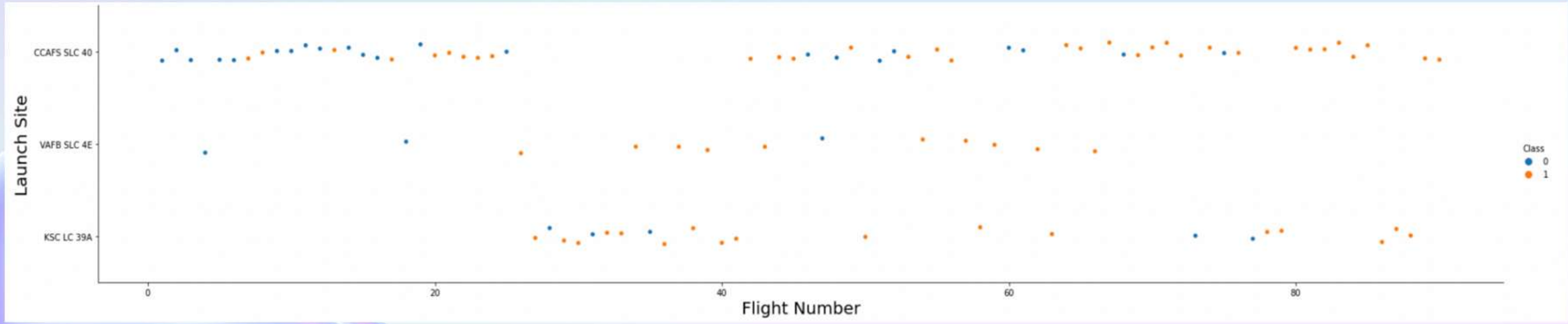
Finding which methods performs best by examining Jaccard_score & F1_score metrics.

RESULTS

FLIGHT NUMBER VS. LAUNCH SITE

Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.



PAYLOAD VS. LAUNCH SITE

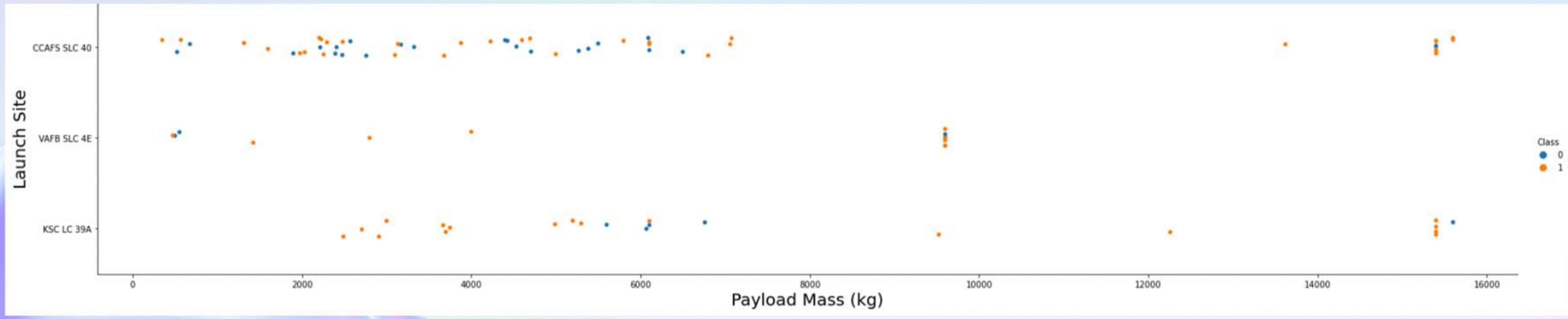
Explanation:

• For every launch site the higher the payload mass, the higher the success

rate.

• Most of the launches with payload mass over 7000 kg were successful.

• KSC LC 39A has a 100% success rate for payload mass under 5500 kg too



SUCCESS RATE VS. ORBIT TYPE

Explanation:

• Orbits with 100% success rate:

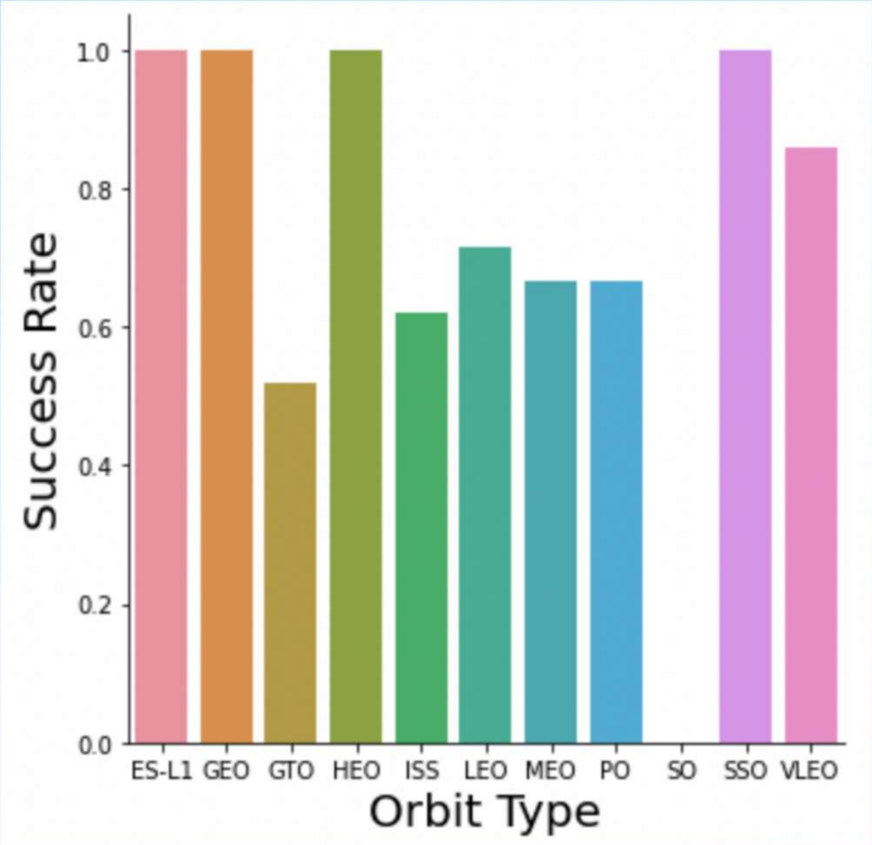
- ES-L1, GEO, HEO, SSO

• Orbits with 0% success rate:

- SO

• Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO

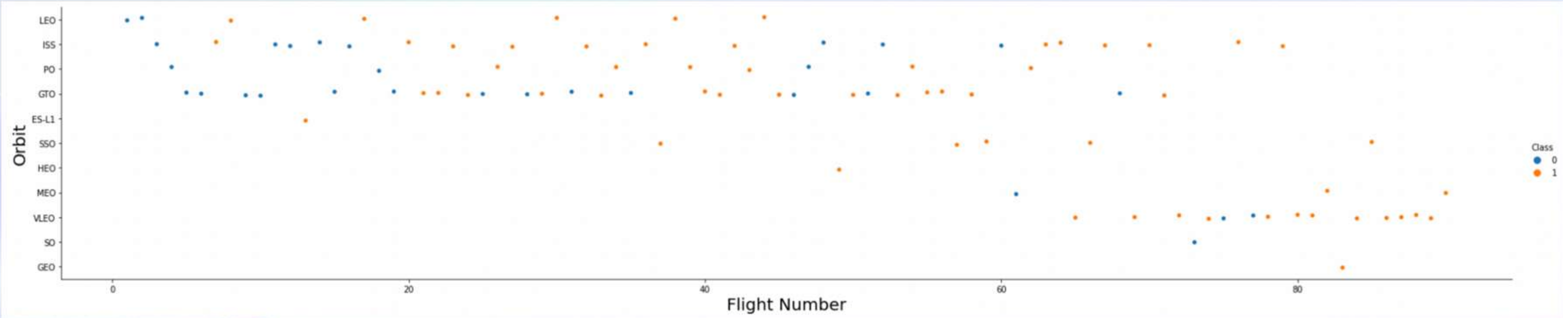


FLIGHT NUMBER VS. ORBIT TYPE

Explanation:

In the LEO orbit the Success appears related to the number of flights.

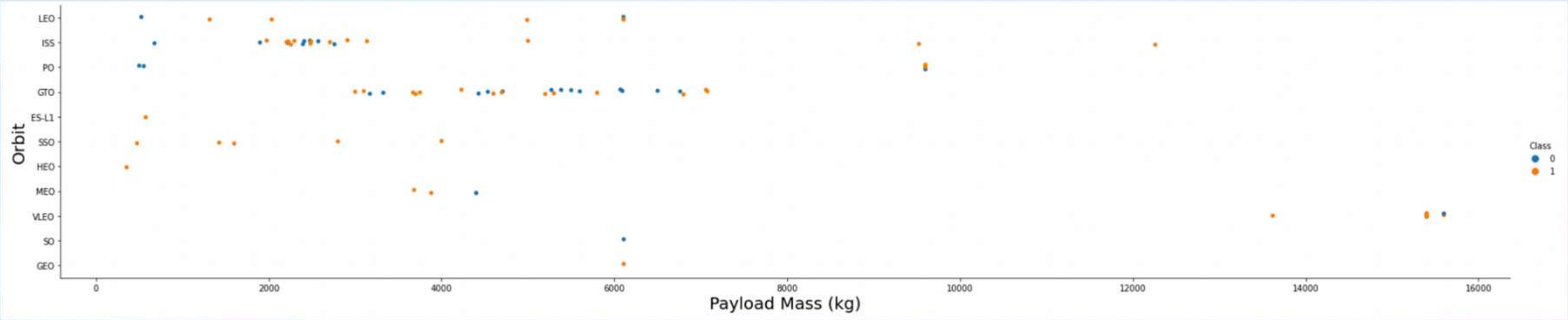
On the other hand, there seems to be no relationship between flight number when in GTO orbit.



PAYLOAD MASS VS. ORBIT TYPE

Explanation:

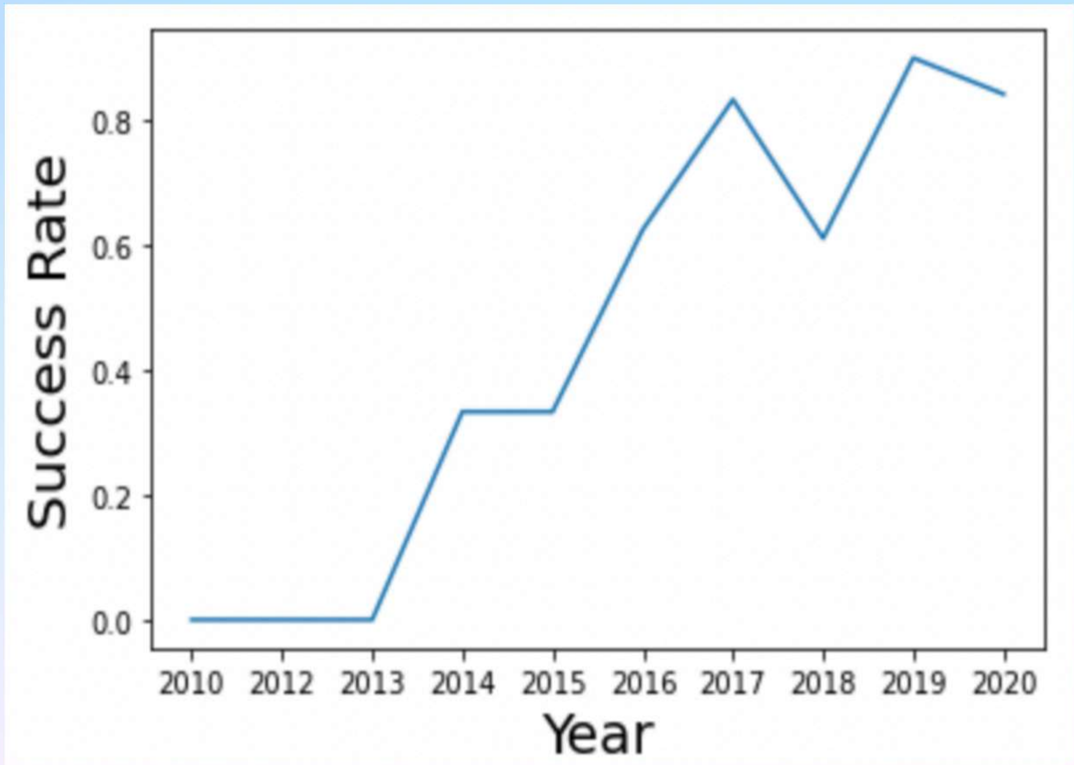
- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



LAUNCH SUCCESS YEARLY TREND

Explanation:

• The success rate since 2013 kept increasing till 2020.



ALL LAUNCH SITE NAMES

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

LAUNCH SITE NAMES BEGIN WITH `CCA`

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[7]:
```

average_payload_mass
2534

FIRST SUCCESSFUL GROUND LANDING DATE + SUCCESSFUL DRONE LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

In [8]:
```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES
+
BOOSTERS CARRIED MAXIMUM PAYLOAD

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

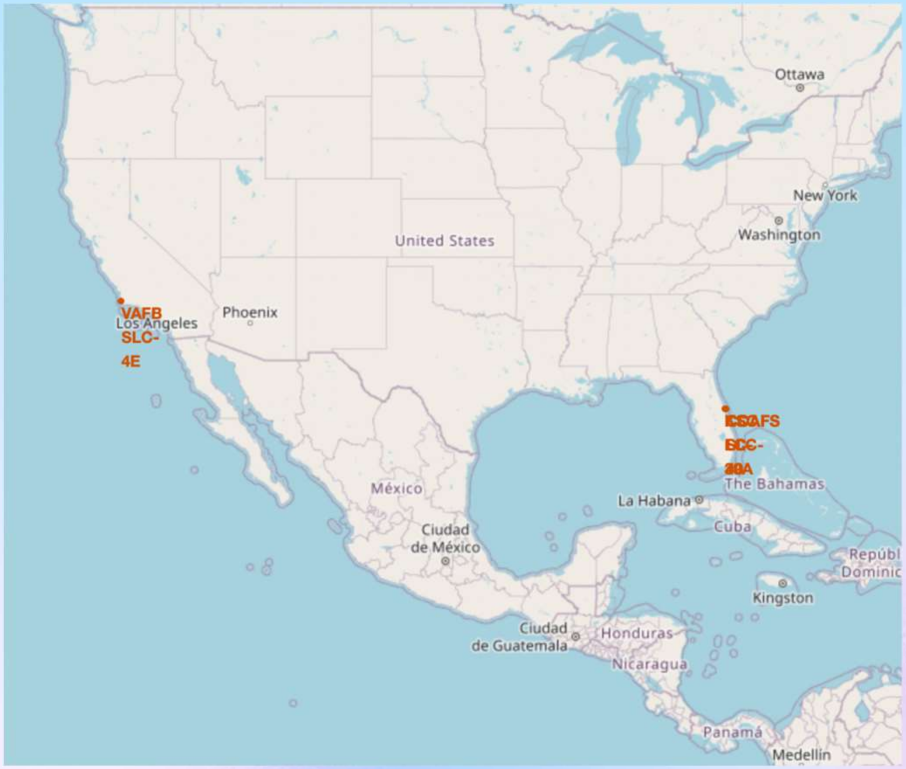
In [12]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

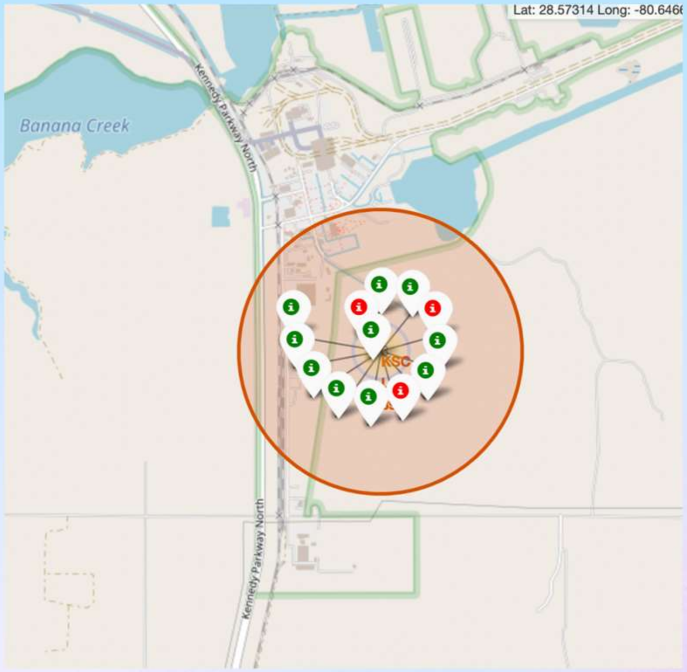
Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

ALL LAUNCH SITES' LOCATION MARKERS ON A GLOBAL MAP



COLOUR-LABELED LAUNCH RECORDS ON THE MAP



Explanation:

Green Marker = Successful Launch

Red Marker = Failed Launch

DISTANCE FROM THE LAUNCH SITE KSC LC-39A TO ITS PROXIMITIES

Explanation:

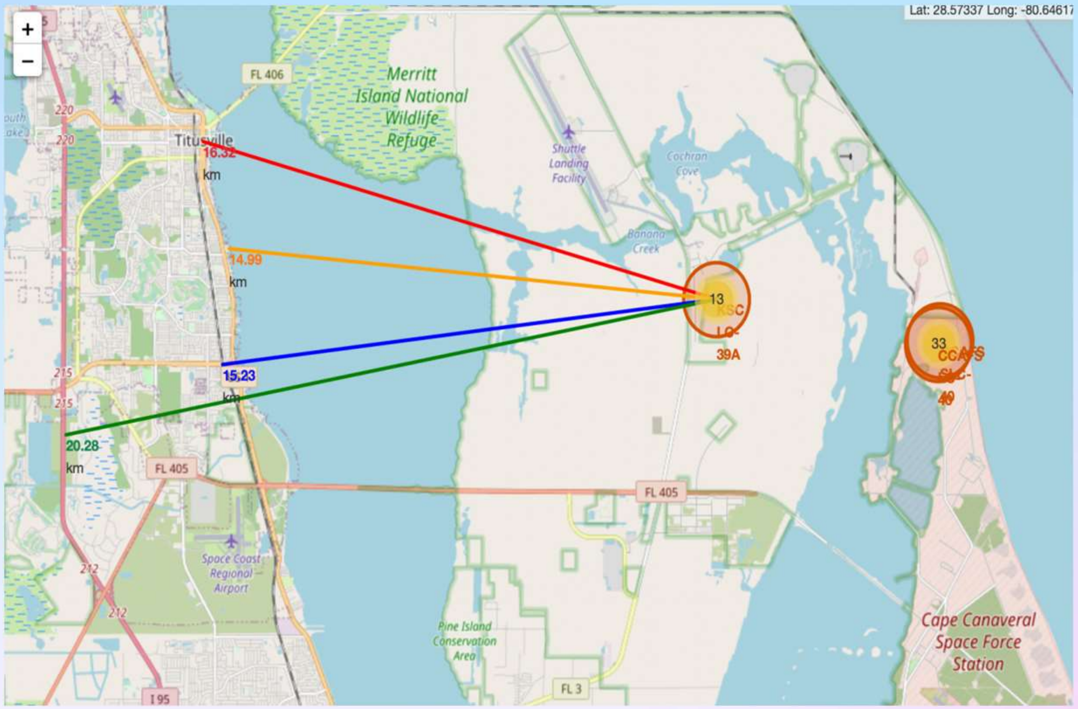
From the visual analysis of the launch

site KSC LC-39A we can clearly see that

it is:

- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

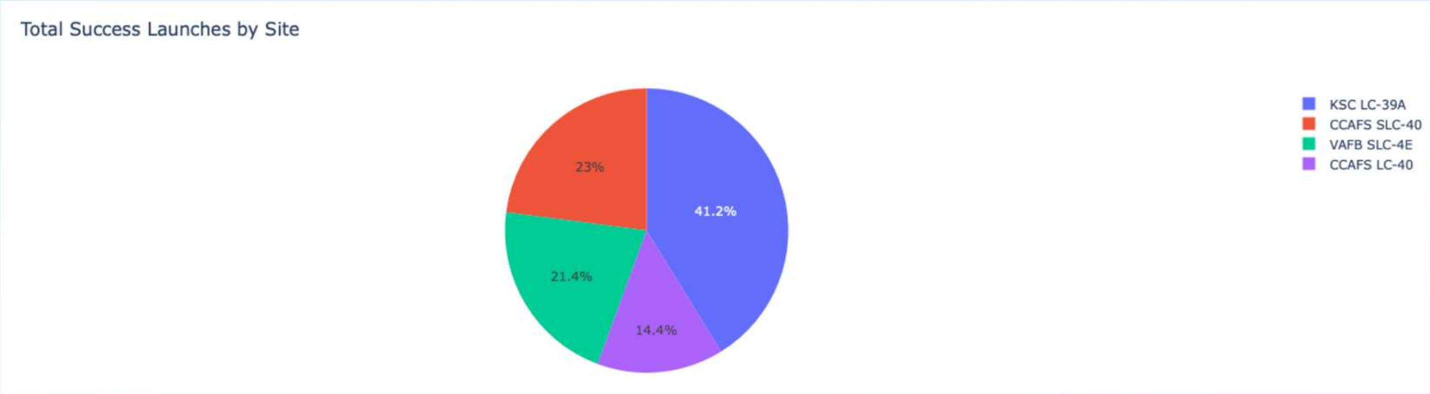
Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).



LAUNCH SUCCESS COUNT FOR ALL SITES

Explanation:

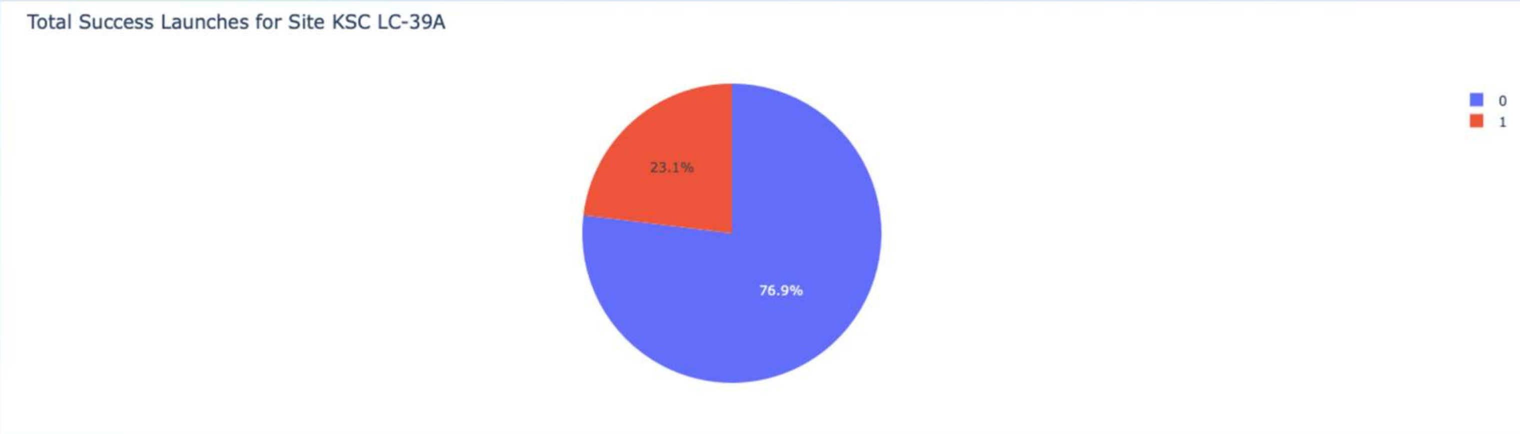
- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches



LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

Explanation:

• KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES

Explanation:

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

CLASSIFICATION ACCURACY

Explanation:

Based on the scores of the Test Set, we can not confirm which method performs best.

Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

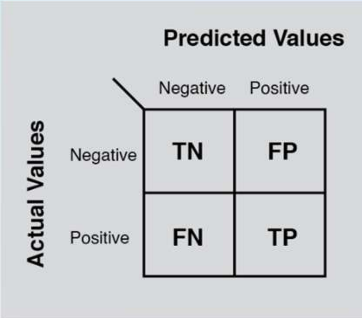
Scores and Accuracy of the Entire Data Set

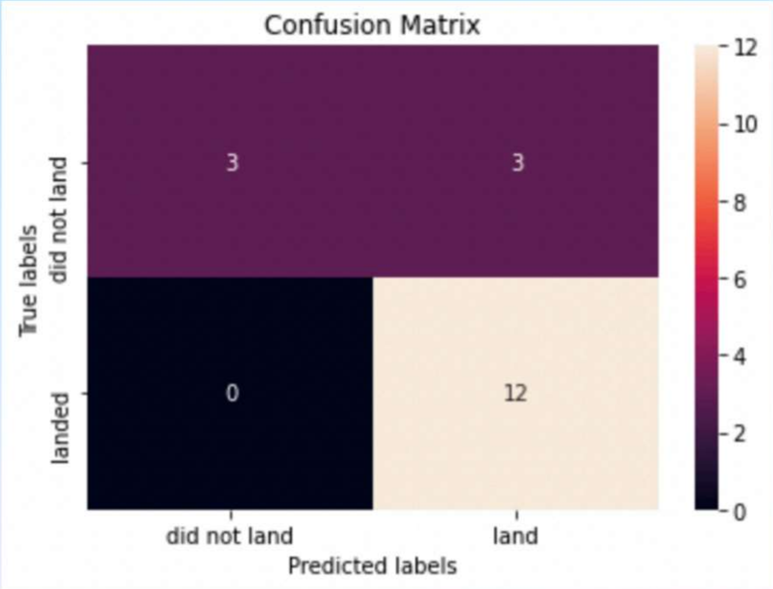
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

CONFUSION MATRIX

Explanation:

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





CONCLUDING REMARKS

- Decision Tree Model is the best algorithm for this dataset.

- Launches with a low payload mass show better results
- than launches with a larger payload mass.

- Most of launch sites are in proximity to the Equator line
- and all the sites are in very close proximity to the coast.

- The success rate of launches increases over the years.

- KSC LC-39A has the highest success rate of the launches
- from all the sites.

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

APPENDIX

Nicholas Rebello