



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Theodora-Augustina Dragan
3rd of June, 2024



Outline

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

Executive Summary

- Summary of methodologies:
 - Data collection through API
 - Data collection with web scraping
 - Data wrangling
 - Exploratory data analysis with SQL and Folium
- Summary of all results
 - Exploratory data analysis
 - Interactive analytics
 - Predictive analytics

Introduction

- Project background and context
 - Launching Falcon 9 cost Space X 62 million dollars and the costs can rise up to triple this amount
 - How to save these costs: **predict** whether a launch will be successful
 - This project is a prediction pipeline that will **help spare** these costs by predicting the end status of a mission
- Problems you want to find answers to
 - What factors can anticipate the successful landing of a rocket?
 - Relations between deciding (and non-deciding) factors
 - Optimal operating conditions

Section 1

Methodology

Methodology

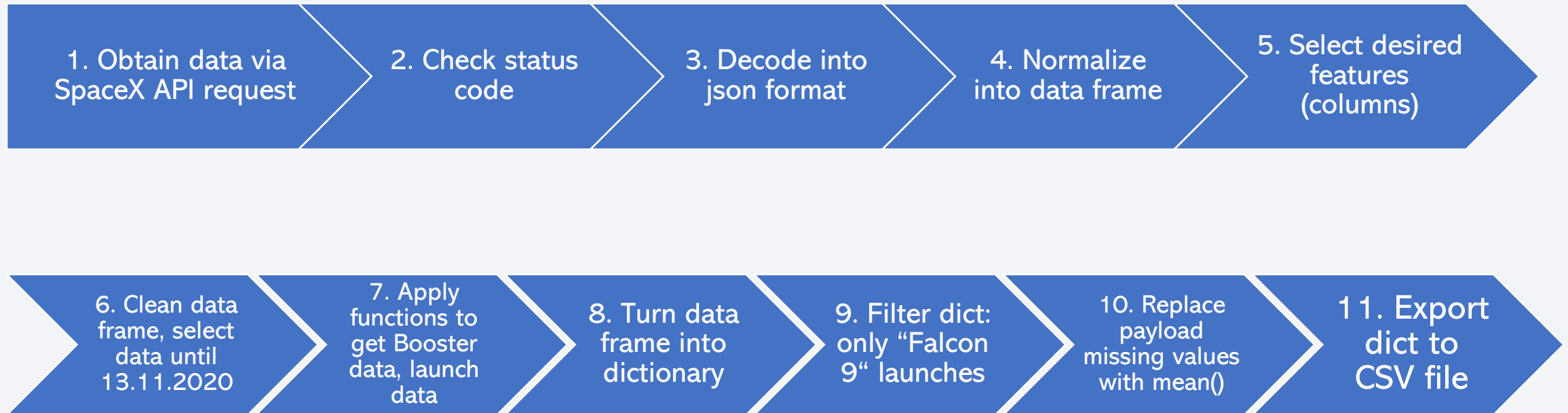
Executive Summary

- Data collection methodology:
 - Data is from Wikipedia and was collected through the SpaceX API and web scraping
- Perform data wrangling
 - One-hot encoding was applied to categorical features
- 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

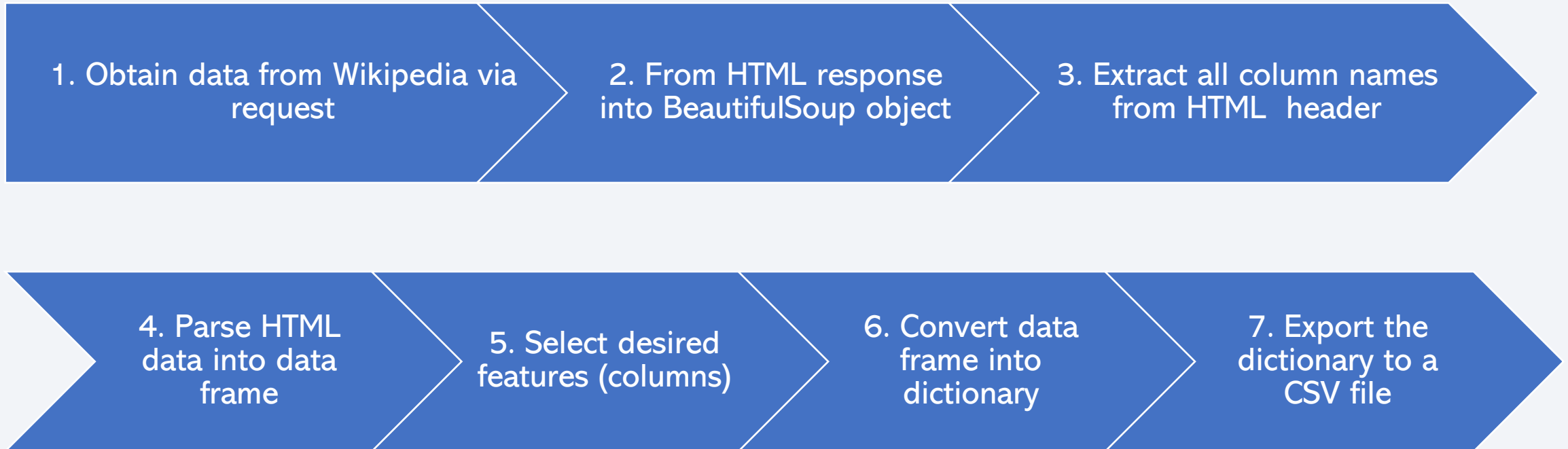
- Data was obtained by using SpaceX REST API and by using Wikipedia Web Scraping
- Concatenating the two allowed us to obtain a broader view on the SpaceX launches
 - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude were obtained from the [SpaceX REST API](#)
 - Flight Number, Launch Site, Payload, Payload Mass, Orbit, [Customer](#), Launch Site, [Version Booster](#), [Booster landing](#) were obtained from [Wikipedia through Web Scraping](#)

Data Collection – SpaceX API



Data Collection Source: [https://github.com/theodoradragan/ds-course/blob/main/Data Collection via API.ipynb](https://github.com/theodoradragan/ds-course/blob/main/Data%20Collection%20via%20API.ipynb)

Data Collection – Web Scraping



Web Scraping Source: https://github.com/theodoradragan/ds-course/blob/main/Data_Scraping_Wiki.ipynb

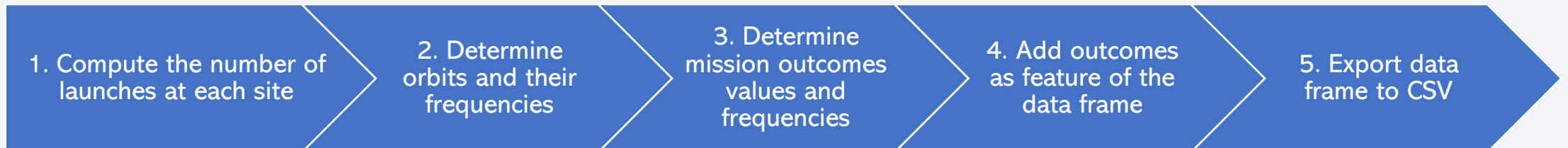
Data Wrangling

Purpose: **Exploratory Data Analysis** to determine the **training labels** of an ulterior machine learning model

Various possible outcomes for the booster in the original data set, for example:

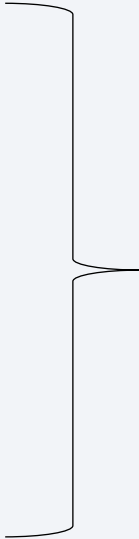
- True Ocean: booster successfully landed in the ocean
- False Ocean: booster crashed in the ocean
- True RTLS: mission was successfully landed to a ground pad

→ We converted these values into “ 1 / 0 “ labels to summarize “successful / unsuccessful” landing.



Data Wrangling Source: https://github.com/theodoradrangan/ds-course/blob/main/Data_Wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
 - 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
 - Success Rate Yearly Trend
- 
- Scatterplots** help investigate a possible correlation between the two variables
- Discrete features can be compared through **bar plots**
- Line plots** to show the evolution of a variable in time

Data Visualisation Source: https://github.com/theodoradragan/ds-course/blob/main/EDA_with_Data_Vizualization.ipynb

EDA with SQL

Performed SQL queries to display:

- the names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- the total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- the date when the first successful landing outcome in ground pad was achieved
- the names of the boosters with success in drone ship and have payload mass is between 4000 and 6000
- the total number of successful and failure mission outcomes
- the names of the booster versions which have carried the maximum payload mass
- the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- a descending count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date s of 2010-06-04

Github link to source file: https://github.com/theodoradragan/ds-course/blob/main/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Added Circles to show all launch sites
 - Reasoning: check proximity to the equator line and to the coast
- Added green/red Clusters to visualize the success/failure of launches:
 - Reasoning: visualize which launch sites are successful
- Computed and showed distances from launch site to proximities (railways, highways, coasts, cities)
 - Reasoning: see how close proximities are and thus how safe the locals are wrt the launches

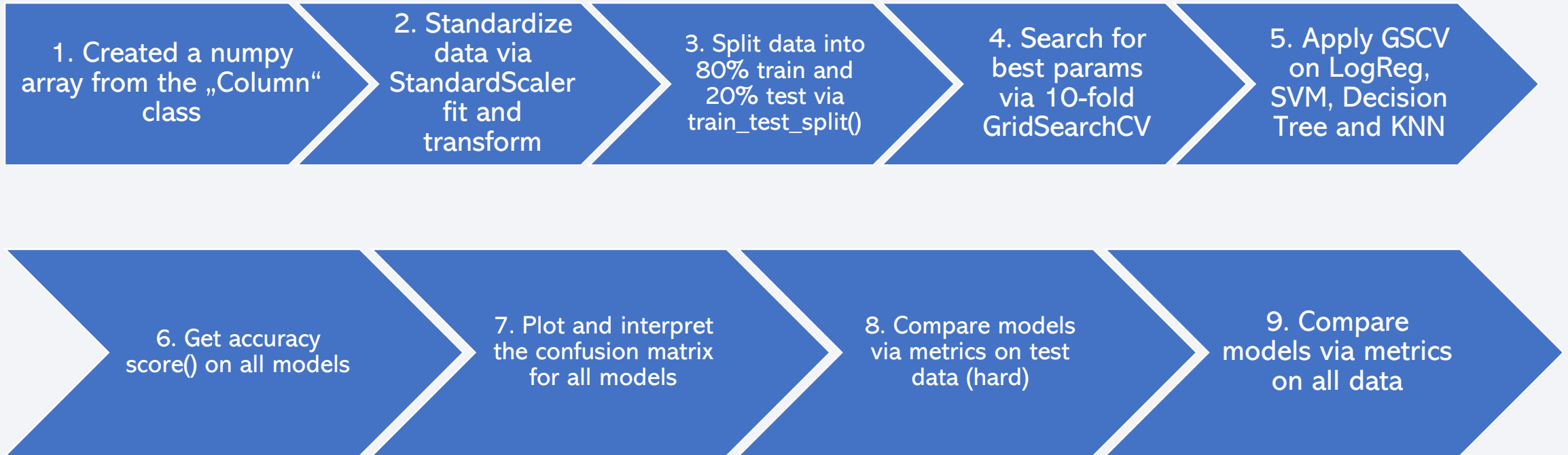
Link: https://github.com/theodoradragan/ds-course/blob/main/Launch_Site_Localization_Viz_Analytics.ipynb

Build a Dashboard with Plotly Dash

- Dropdown to view statistics per launch site or of all launch sites
- Pie chart of success rate (per launch site or of all launch sites)
- Slider of Payload mass range between 0 and 10k kgs
- Scatter Plot: Payload mass vs Success rate, grouped by Booster version, payload mass range can be selected via the aforementioned slider

Link: https://github.com/theodoradragan/ds-course/blob/main/Viz_with_Plotly.py

Predictive Analysis (Classification)



Link: https://github.com/theodoradragan/ds-course/blob/main/Machine_Learning_Prediction.ipynb

Results

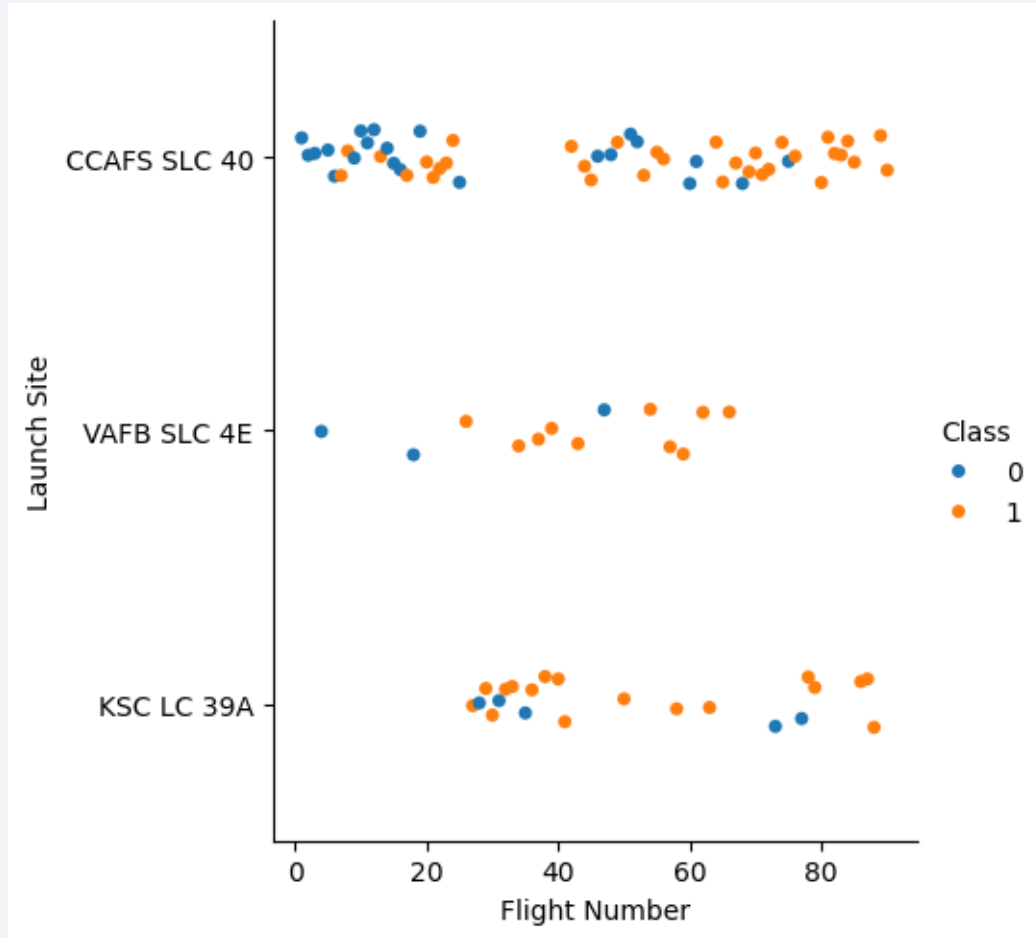
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

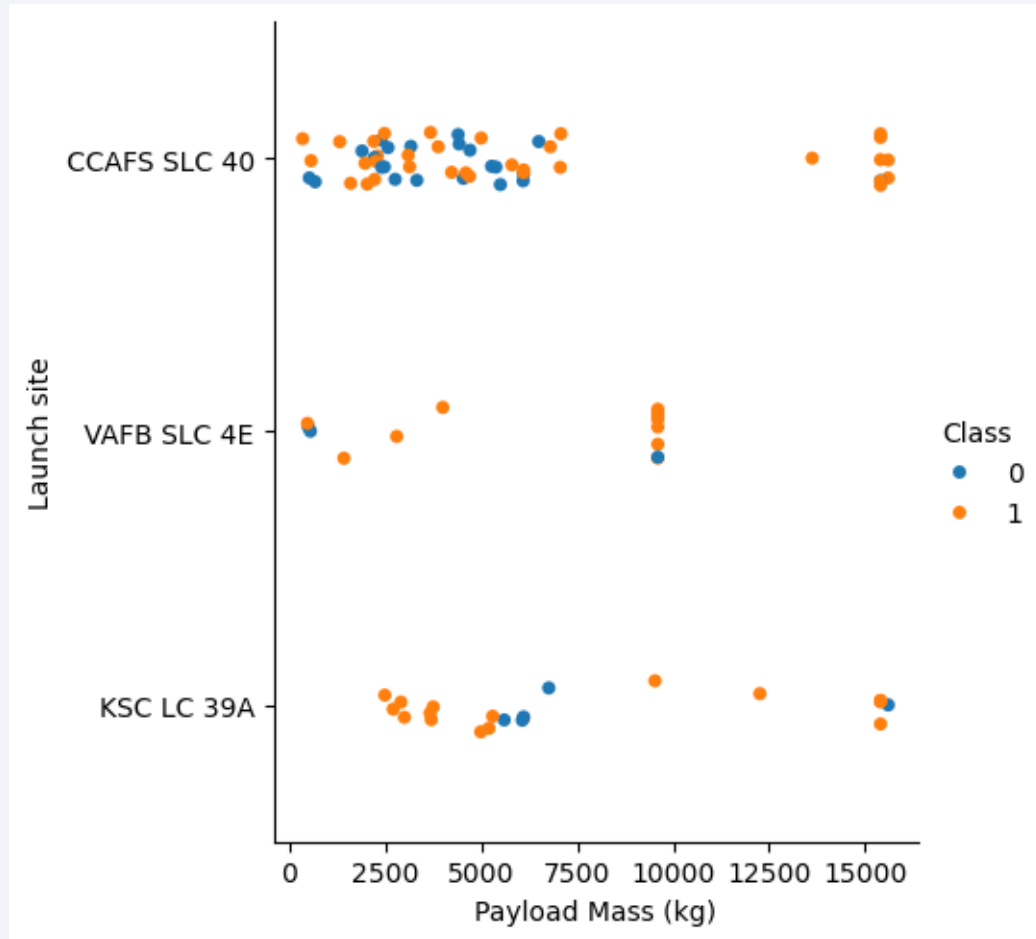
Insights drawn from EDA

Flight Number vs. Launch Site



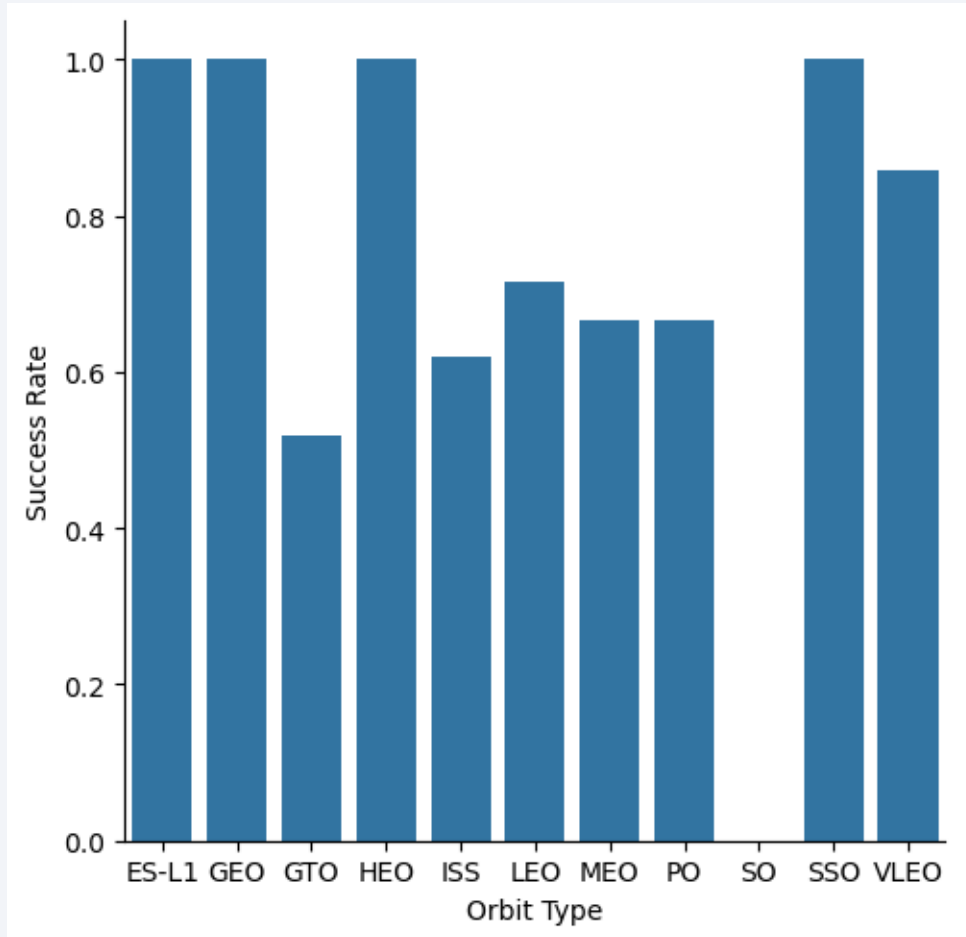
- Earlier flights all failed, late flights almost all succeeded
- Launch site CCAFS SLC 40 has the most launches, but also the most failures
- Launch sites VAFB SLC 4E and KSC LC 39A have higher success rates, but fewer launches in total

Payload vs. Launch Site



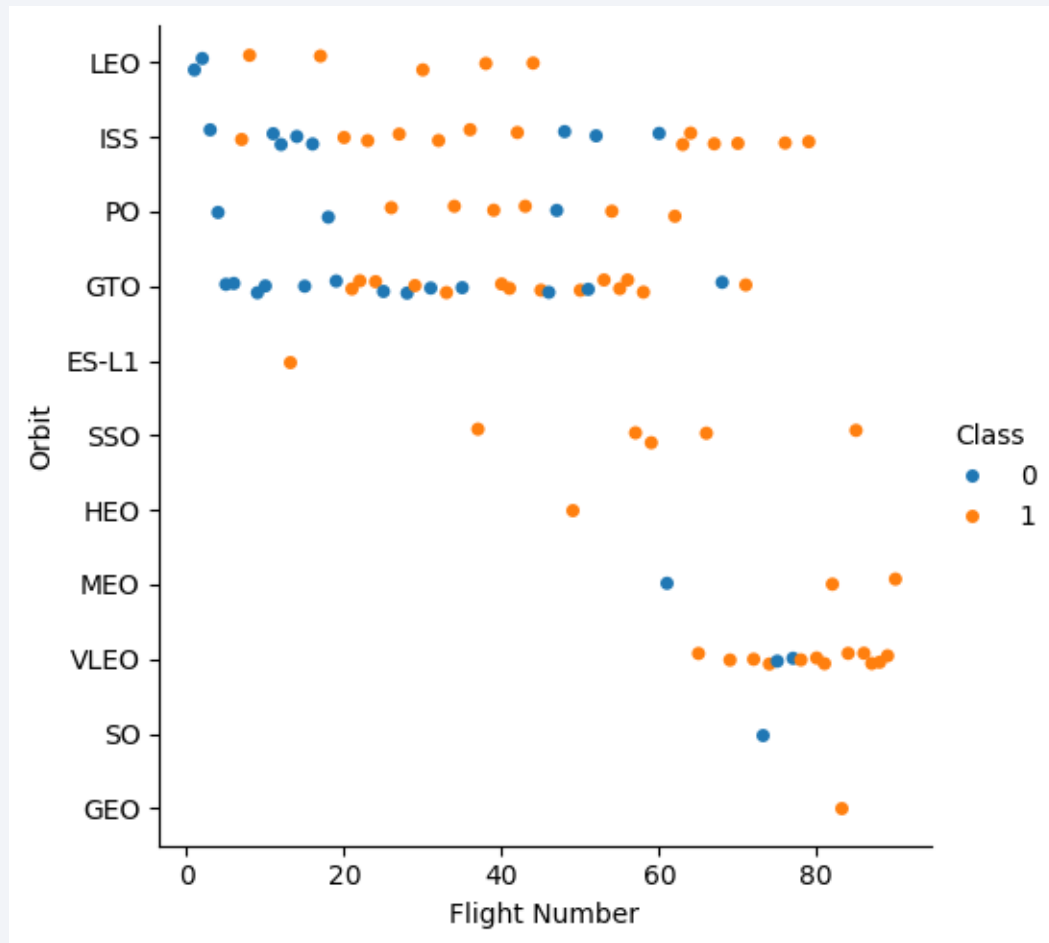
- For all launch sites, higher payloads have higher success rates
- Most payloads above 8k kg were successful
- For sites VAFB SLC 4E and KSC LC 39A there is no clear correlation between the two variables

Success Rate vs. Orbit Type



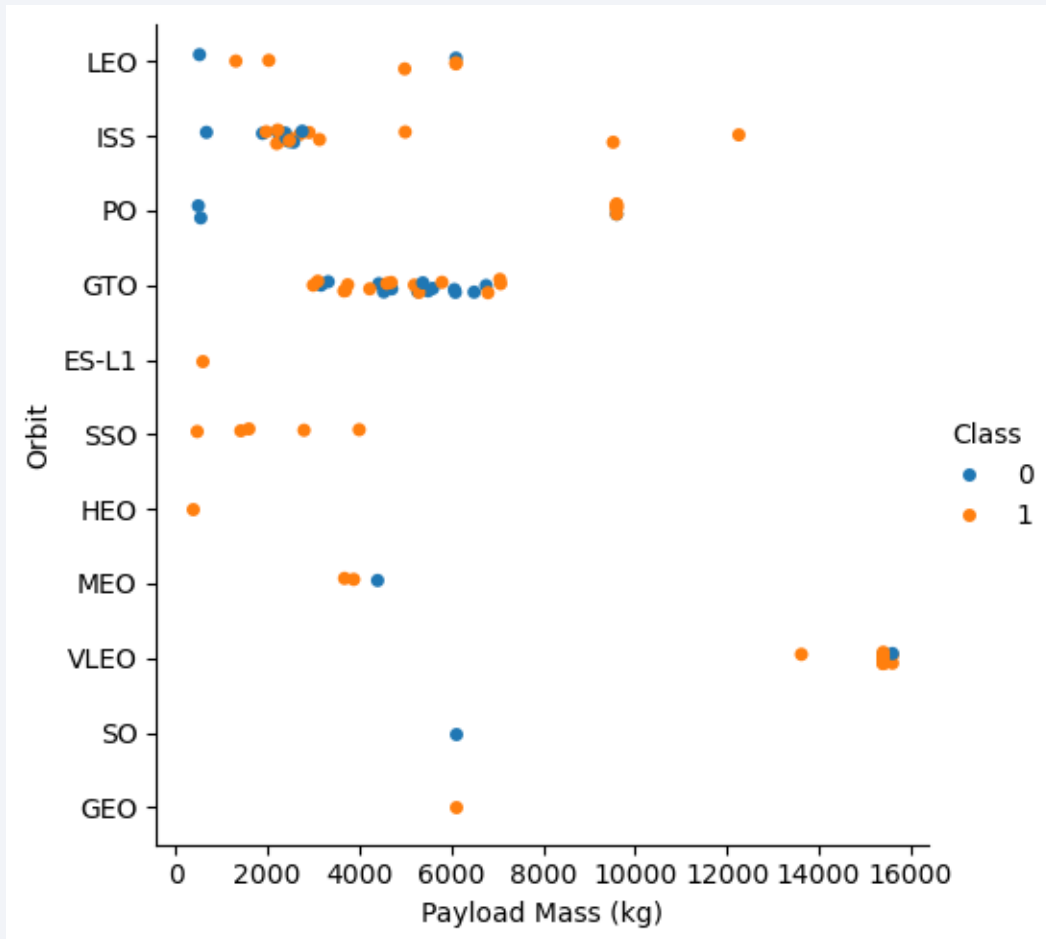
- The orbits with 100% success rate are: ES-L1, GEO, HEO, SSO
- The orbits between 61% and 100% success rate are: LEO, MEO, PO
- The orbits between 1% and 60% success rate are: GTO and ISS.
- The orbit with 0% success rate is: SO

Flight Number vs. Orbit Type



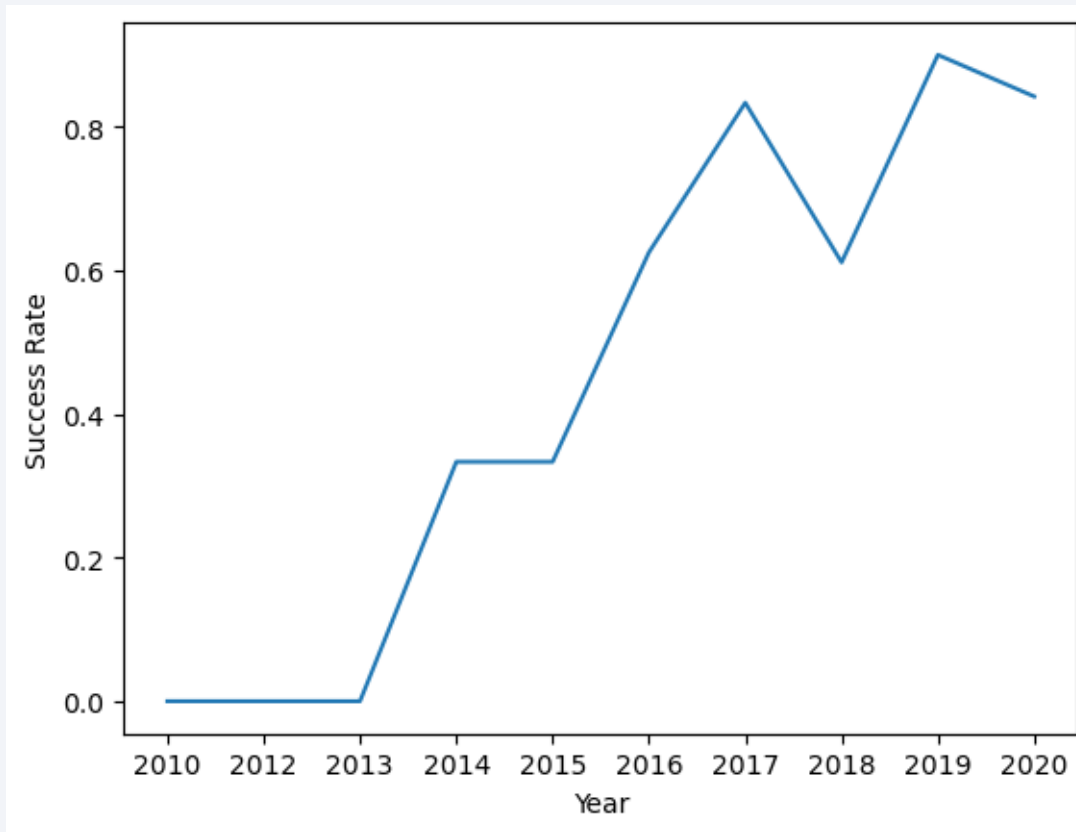
- Correlations seem to be found in LEO and VLEO between flight number and orbit type
- No correlations in GTO, SSO and ISS
- No overall conclusion can be drawn

Payload vs. Orbit Type



- Correlations seem to be found in PO, LEO and ISS, where heavy payloads lead to successful landings
- No correlations in GTO, SSO, or VLEO
- No overall conclusion can be drawn

Launch Success Yearly Trend



- Success rate keeps increasing since 2013, with a short drop in 2018

All Launch Site Names

```
%sql select distinct launch_site from SPACEXTABLE;
```

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

```
Done.
```

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

- Through 'DISTINCT' we selected the unique launch site names from the SpaceX table

Launch Site Names Begin with 'CCA'

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

* sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parac
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 (parac
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

- Through 'like 'CCA%'' we selected the launch site names that begin with the string 'CCA' in the SpaceX table

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

```
Done.
```

total_payload_mass

45596

- Through the 'sum' operator we selected and then added the payloads in the SpaceX table, and through the 'where' operator we specified to only select the launches of the NASA (CRS) customer

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

```
Done.
```

<u>average_payload_mass</u>

2534.6666666666665

- Through the 'avg' operator we selected and then averaged the payloads in the SpaceX table, and through the 'where' and the 'like' operators we specified to only select the launches where the booster contains 'F9 v1.1' in the name

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

```
Done.
```

<u>first_successful_landing</u>

2015-12-22

- Through the 'min' operator we selected the earliest launch (the one that has the lowest date) in the SpaceX table, and through the 'where' operator we specified we are only interested in the successful launches

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000
```

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

Done.

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- Selected the landing outcome to be successful through the ,where' operator and then specified the payload mass to be between 4k and 6k kgs

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

```
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Calculated the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTABLE where PAYLOAD_MASS_KG_ = (select max(PAYLOAD_MASS_KG_) from SPACEXTABLE)
```

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

- Listed the names of the booster which have carried the maximum payload mass through a subquery

2015 Launch Records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
       where landing_outcome = 'Failure (drone ship)' and year(date)=2015;
```

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)

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
where date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by count_outcomes desc;
```

* sqlite:///my_data1.db

Done.

Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

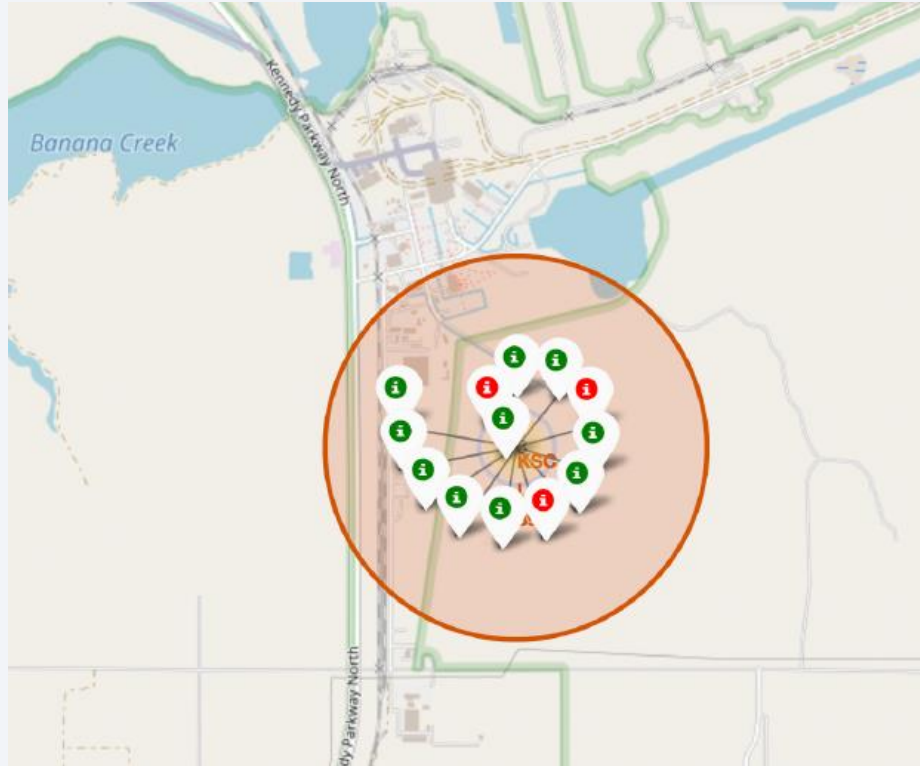
Launch Sites Proximities Analysis

Launch Site Locations



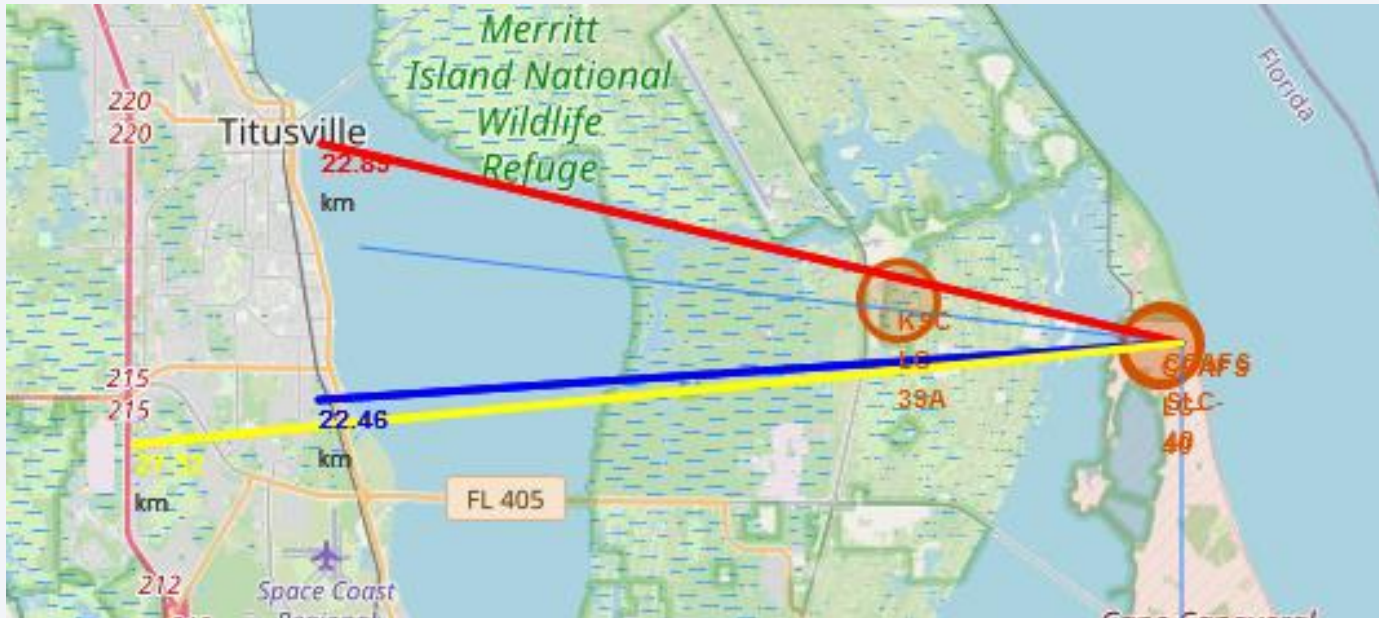
- Most of the launch sites are in proximity to the equator line.
- All launch sites are near the coast. This may be due to the fact that in case of failure, the rocket can be directed to the water and avoid casualties on land.

Success rates of launch sites



- Labelled launches with green/red based on whether they are **successful/unsuccessful**
- Launch site KSC IC-39A has a high success rate

Distance of launch sites to proximities



Looking at the "CCAFS LC-40" launch site we can see:

- It is 23 kms away from railways.
- It is 27 kms away from highways.
- It is 23 kms away from the coastline.
- It is almost 23 kms away from the closest city.

We can observe other sites and come to the same conclusion, that a certain safety distance is maintained.

However, considering the speed of the rockets, it means SpaceX has few seconds to redirect a rocket to avoid a direct impact.

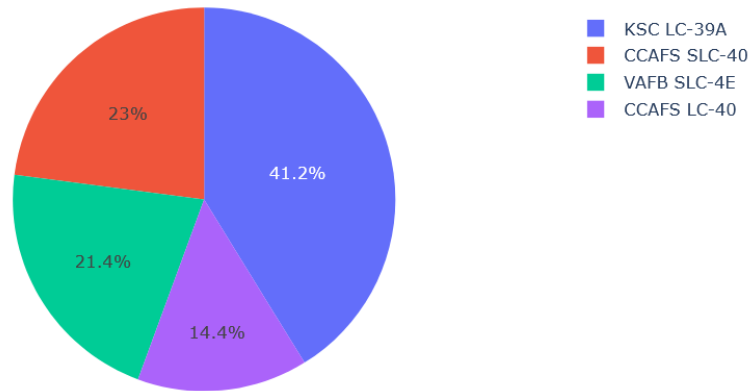


Section 4

Build a Dashboard with Plotly Dash

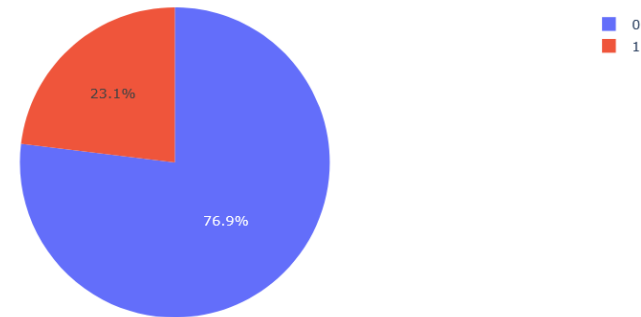
Success Rate All Sites

Success launches for all sites



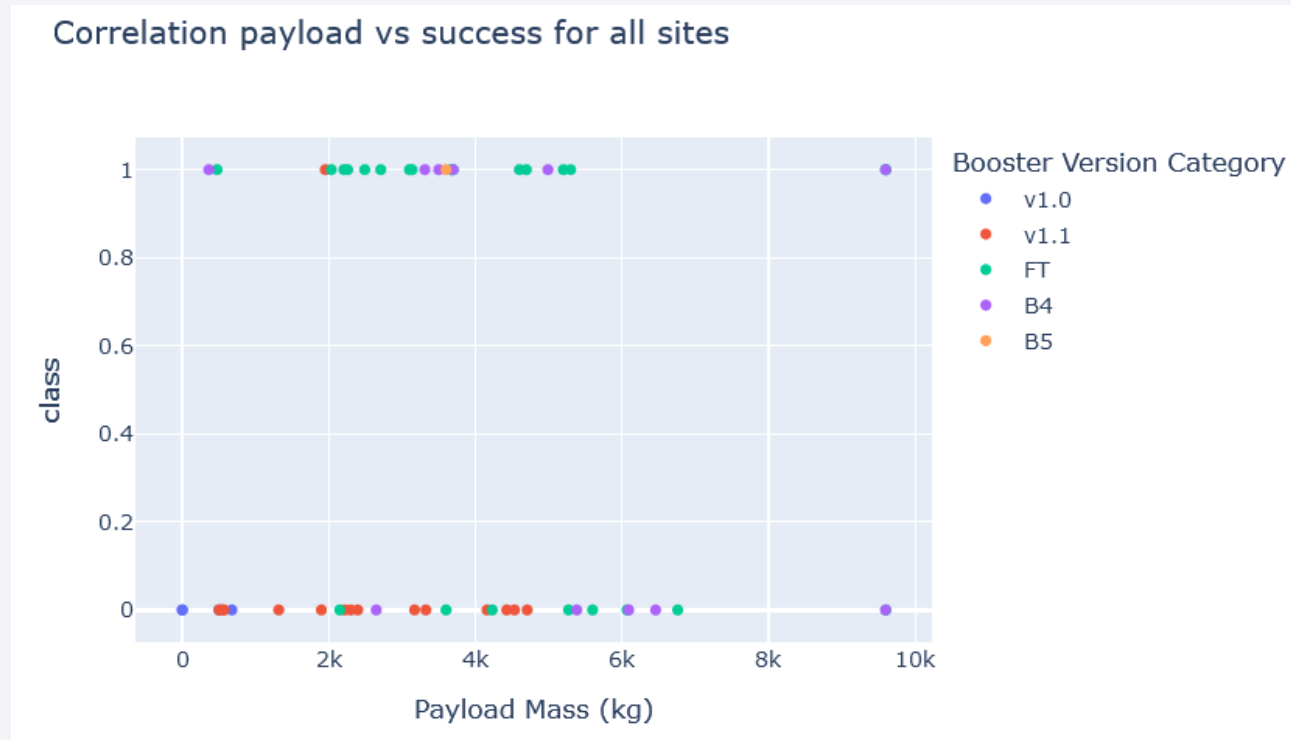
KSC LC-39A

Success launches for sites KSC LC-39A



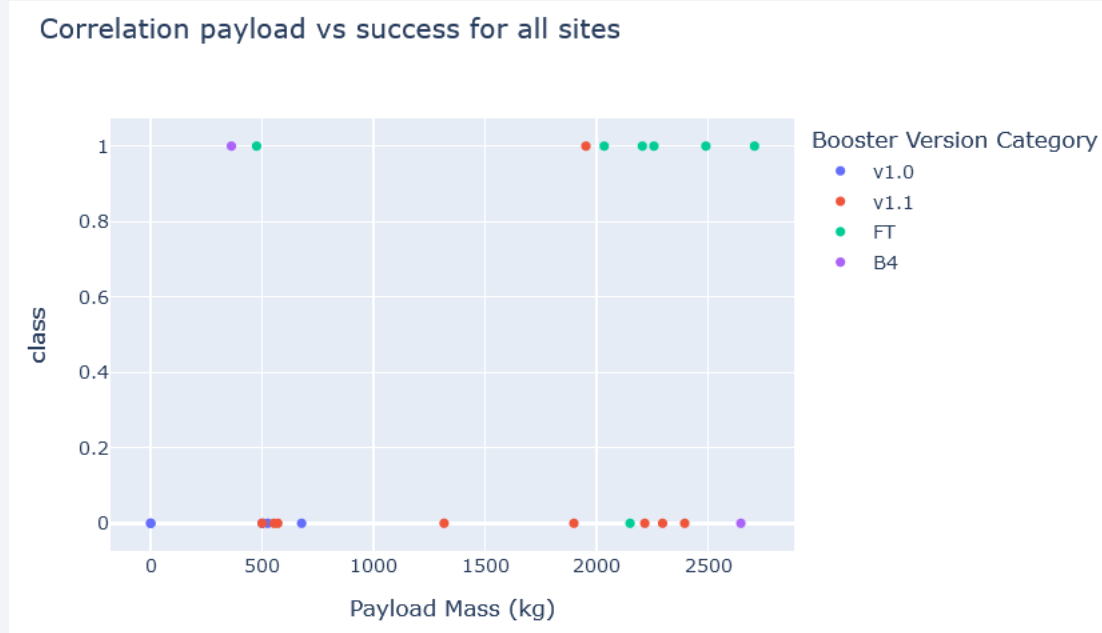
- Site 'KSC LC-39A' has the highest number of success launches, and the highest success rate

Payload mass and Booster Versions

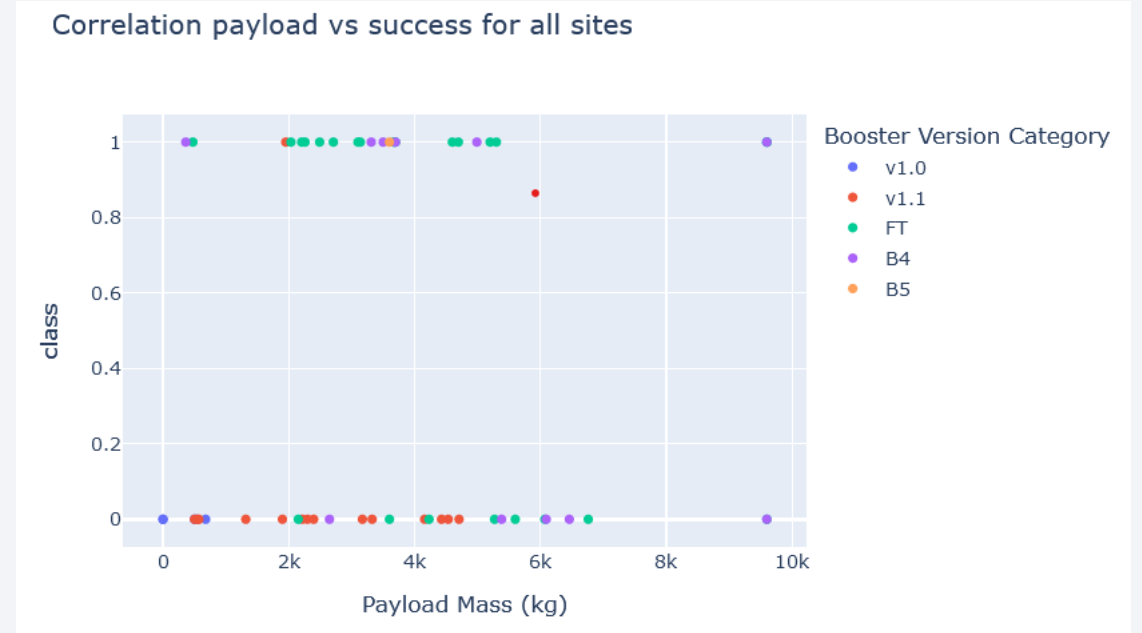


- No obvious relationship between payload mass and class

Payload mass and Booster Versions by Range



Worst success ratio: 0-3k kgs



Best success ratio: 2-4k kgs



Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
from sklearn.metrics import precision_score
accuracies = [logreg_cv.score(X_test, Y_test), svm_cv.score(X_test, Y_test), tree_cv.score(X_test, Y_test), knn_cv.score(X_test, Y_test)]
print(accuracies)

precisions = [precision_score(Y_test, logreg_cv.predict(X_test), average='binary'),
              precision_score(Y_test, svm_cv.predict(X_test), average='binary'),
              precision_score(Y_test, tree_cv.predict(X_test), average='binary'),
              precision_score(Y_test, knn_cv.predict(X_test), average='binary'),
              ]
print(precisions)
```

```
[0.8333333333333334, 0.8333333333333334, 0.8333333333333334, 0.8333333333333334]
[0.8, 0.8, 0.8, 0.8]
```

```
accuracies = [logreg_cv.score(X, Y), svm_cv.score(X, Y), tree_cv.score(X, Y), knn_cv.score(X, Y)]
print(accuracies)

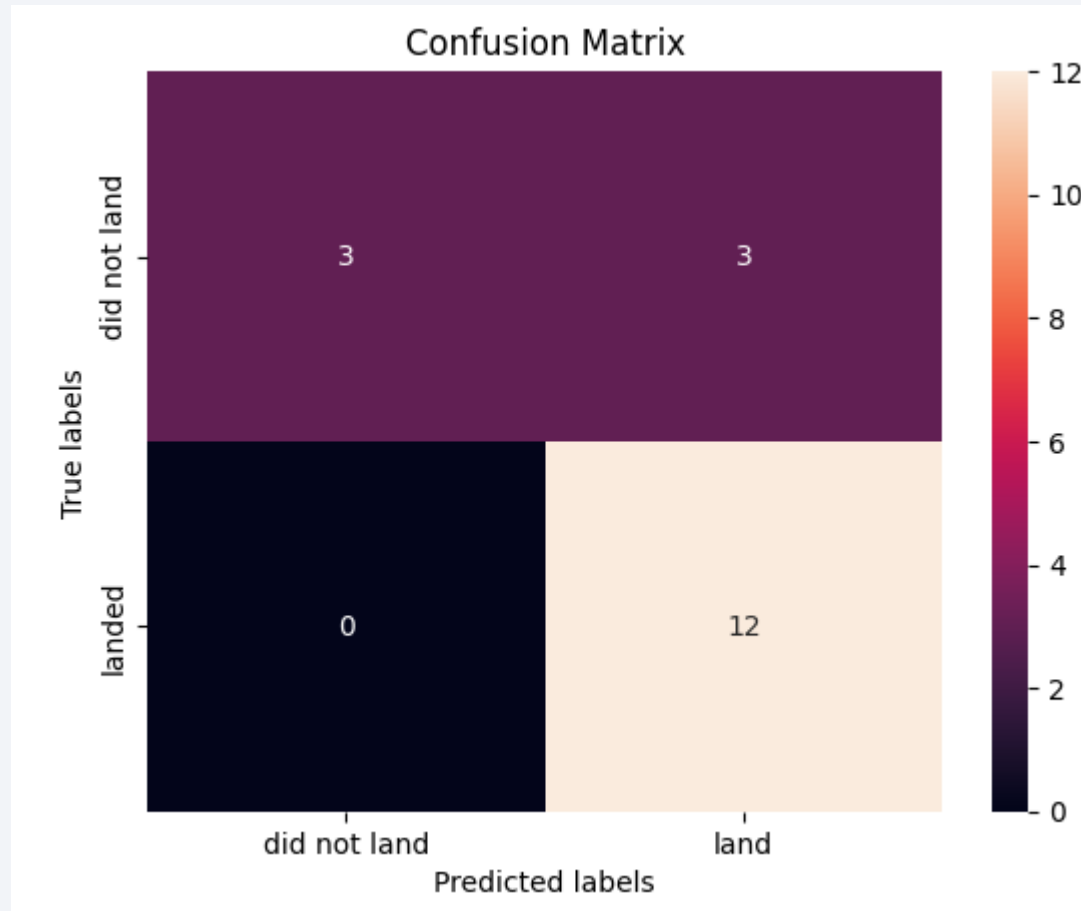
precisions = [precision_score(Y, logreg_cv.predict(X), average='binary'),
              precision_score(Y, svm_cv.predict(X), average='binary'),
              precision_score(Y, tree_cv.predict(X), average='binary'),
              precision_score(Y, knn_cv.predict(X), average='binary'),
              ]
print(precisions)
```

```
[0.8666666666666667, 0.8777777777777778, 0.9, 0.8555555555555555]
[0.8333333333333334, 0.8450704225352113, 0.8695652173913043, 0.8309859154929577]
```

Since we noticed that on the test data sets, the scores are the same (maybe also due to the low number of 18 samples), we ran the metrics on the entire data set.

These metrics show that the **decision tree model** performs best.

Confusion Matrix



The best model, the decision tree, has the following confusion matrix.

We can see that most landed launches are well classified, but half of the non-landed are misclassified.

Conclusions

- Most launches are near the Equator line and the Coastline, but this does not necessarily lead to the safety of the nearby populations.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Launch site KSC LC-39A had the most successful launches
- No obvious relationship between payload mass and class, but certain payload ranges have higher success rates, e.g. btw 2000-4000 kgs
- Decision Tree model is the best algorithm for this data set, but one has to apply the metrics on the entire data set to observe differences in model

Thank you!

