IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francis de Ladurantaye
February 25th, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - We chose the well-known SpaceX launch provider as the source of our data.
  - This data was used to perform Exploratory Data Analysis (EDA) and to train a model predicting the probability of successful landing for a booster launch given launch parameters.
- Summary of all results
  - We discovered that SpaceX was more successful in recovering its boosters after getting some experience with the landing process.
  - More recent versions of the Falcon 9 boosters are more likely to be recovered with success.
  - For some reason, launches with greater payload mass were more likely to be successfully recovered.

Introduction

- Project background and context
  - The recovery of boosters used to launch cargo to space has dramatically reduced the cost of sending payload to orbit, as was demonstrated by SpaceX's Falcon 9.
  - As a new entrant in the space launch industry, our goal is to analyze SpaceX launched and determine the factors making a successful booster recovery more likely, allowing us to minimize our costs and offer best prices to our customers to send payload to space.

- Problems you want to find answers
  - We'd like to understand which factors are the most indicative in determining if a booster landing will be successful or not.
  - Doing so, our company could specialize in deserving customer whose needs fit with these optimal factors, maximizing our chances of success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data for this project was collected using a combination of web scraping and data fetching from a public API.
- Perform data wrangling
  - Data wrangling was applied to the collected data to deal with missing values and categorical variables, thus making our data ready for the final predictive task.
- 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

- For this project, data was collected by:
  - Querying the SpaceX API:
    - https://api.spacexdata.com/v4/
    - Recovered data features include:
      - Date, booster version, payload mass, target orbit, launch outcome, coordinates, etc.
  - Web scraping the "List of Falcon 9 and Falcon Heavy launches" page on Wikipedia:
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Recovered data features include:
      - Flight number, date, time, booster version, launch site, payload, payload mass, target orbit, customer, launch outcome and booster landing status.

Data Collection – SpaceX API

- Data fetch process from the API:
  - Query the API and parse response using pd.json_normalize()
  - Extract relevant data using custom functions
  - Filter DataFrame to only include Falcon 9 launches
  - Deal with missing values (were replaced by the mean of the appropriate column)
- Data fetching notebook

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Web scraping process:
  - Fetch page using `requests` package
  - Parsing using BeautifulSoup and extracted relevant table
  - Retrieve column names from table header
  - Creating the DataFrame and saving it to CSV
- Data scraping from Wikipedia: link here
- Data scraping notebook

Place your flowchart of web scraping here

Data Wrangling

- Summary of the Data Wrangling process:
  - DataFrame filtering to only keep Falcon 9 launches.
  - Missing values replaced by column mean.
  - Creation of `Class` column equal to 1 if landing was successful and 0 otherwise.
  - One-hot encoding of Categorical columns
- Data Wrangling Notebook

EDA with Data Visualization

- For the Exploratory Data Analysis, those are the charts that were plotted:
  - Scatter plot of landings status, represented as Flight Number vs Payload Mass
  - Scatter plot of landings status, represented as Flight Number vs Launch Site
  - Scatter plot of landings status, represented as Payload Mass vs Launch Site
  - Bar plot of successful landings rate by Orbit Type
  - Scatter plot of landings status, represented as Flight Number vs Orbit Type
  - Scatter plot of landings status, represented as Payload Mass vs Orbit Type
  - Line plot of successful landing rate by Launch Year
- EDA with Visualization Notebook

EDA with SQL

- For the Exploratory Data Analysis with SQL, we discovered (among others):
  - There are 4 different launch sites (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
  - The total mass carried by boosters launched for `NASA (CSR)` is 45596 kg
  - First successful landing of a Falcon 9 booster occurred on 2015-12-22
  - There are much more successful landings than failed landings
  - Falcon 9 Block 5 is the only booster version able to send the maximum payload weight
- EDA with SQL Notebook

Build an Interactive Map with Folium

- Using Folium interactive maps, we generated:
  - A map showing the launch sites to discover that all launch sites are close to the coast and the equator.
  - An interactive map showing for each launch site and launch, if the landing was successful or not.
  - A map showing the distance from a launch site to the nearest coast, indicated by a drawn line.
- We also discovered that no launch site has been built near a city.
- Link to Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- In this interactive dashboard were included:
  - A pie chart section representing either:
    1) Successful launches by launch site.
    2) The proportion of successful launches for a specified launch site.
  - A scatter plot section representing either:
    1) Landing status vs Payload Mass and Booster Version for all launch sites.
    2) Landing status vs Payload Mass and Booster Version for a specified launch site.
- The scatter plot results could also be filtered to only show launches within a specified Payload Mass range.
- These plots were added to give quick insights on launch landing statuses based on payload and launch site. Interactivity was added to further refine the display for specific launch sites.
- Dashboard with Plotly Dash script

Predictive Analysis (Classification)

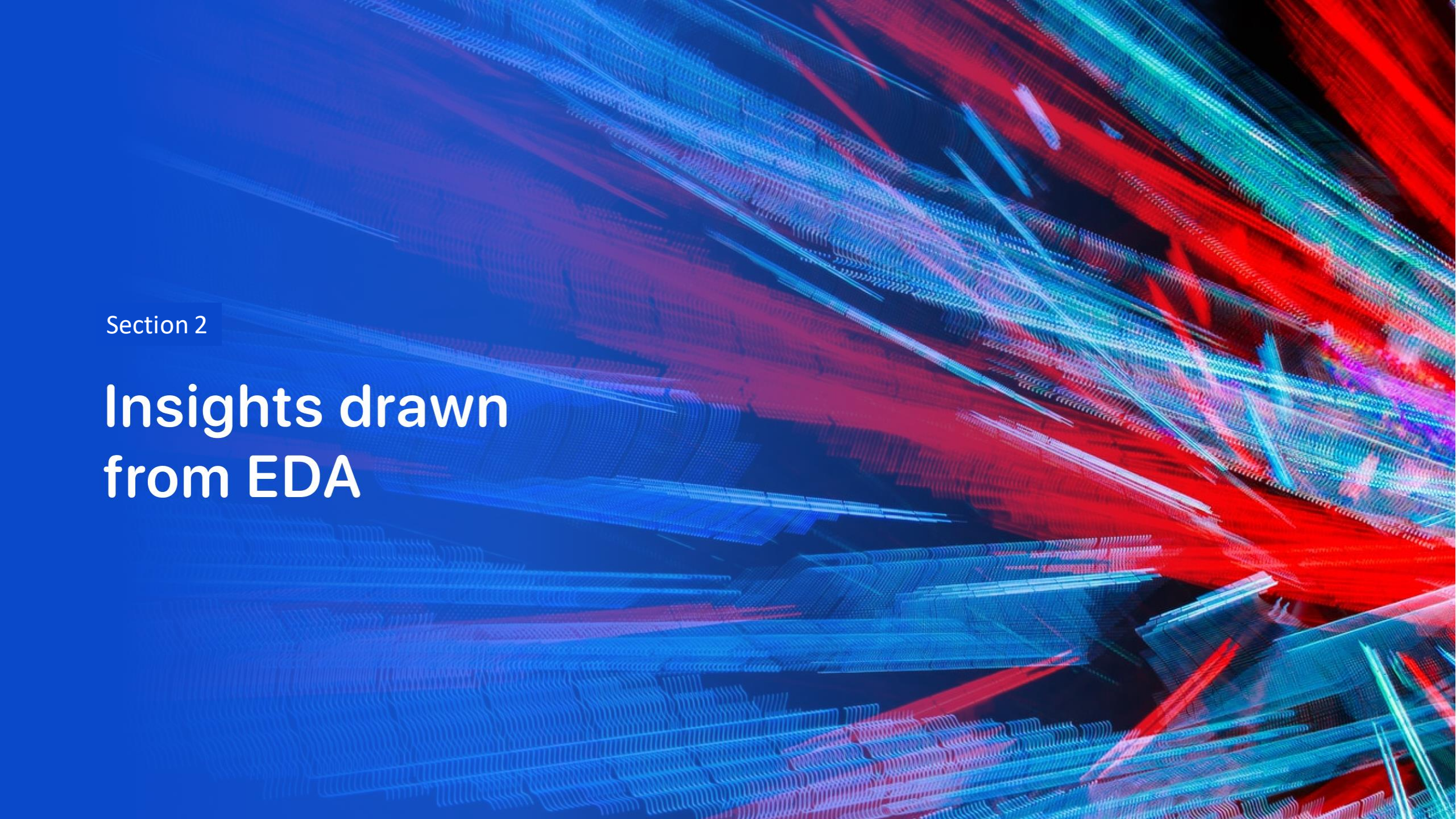
- The predictive analysis process was conducted as follows:
  - Data normalization by rescaling to Gaussian(0, 1).
  - Splitting the dataset into a train and a test set.
  - Test set was held out to validated trained models performance.
  - A variety of models were selected and grid search was used to determine best hyperparameters for these models:
    - Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN)
  - Generalization performance was computed using the GridSearchCV `score` function on the held out test set.
- The best performing model was the Decision Tree Classifier, achieving a 94.4% prediction accuracy on the held out test set.
- Predictive Analysis Notebook

Results

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

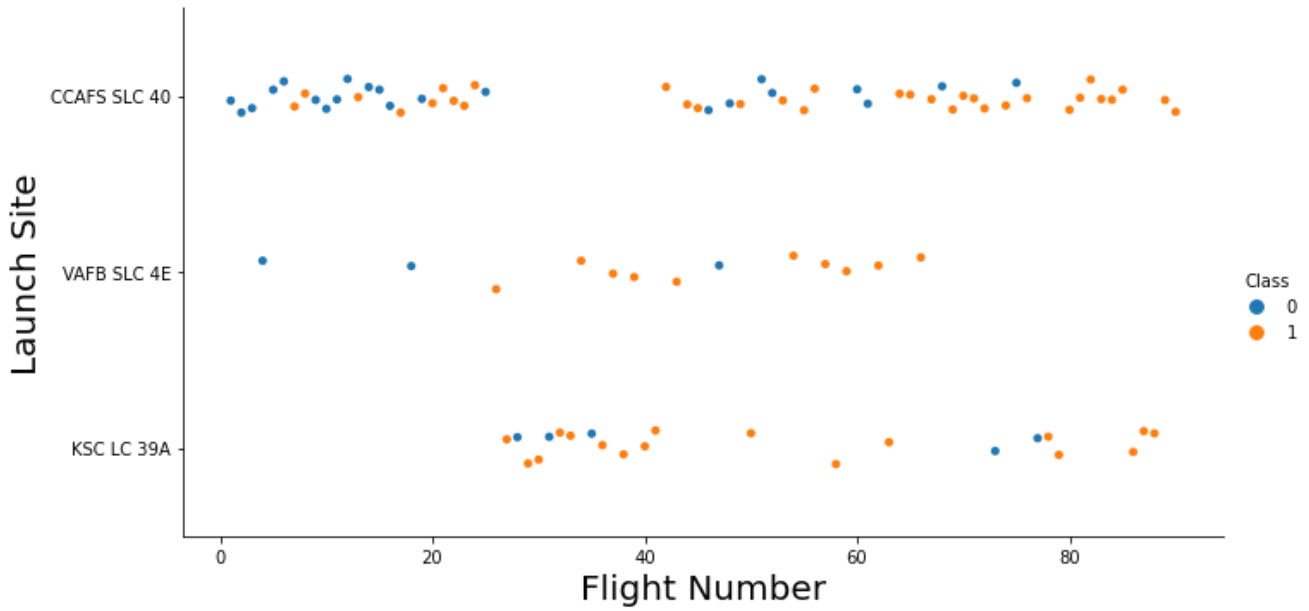
Section 2

Insights drawn from EDA



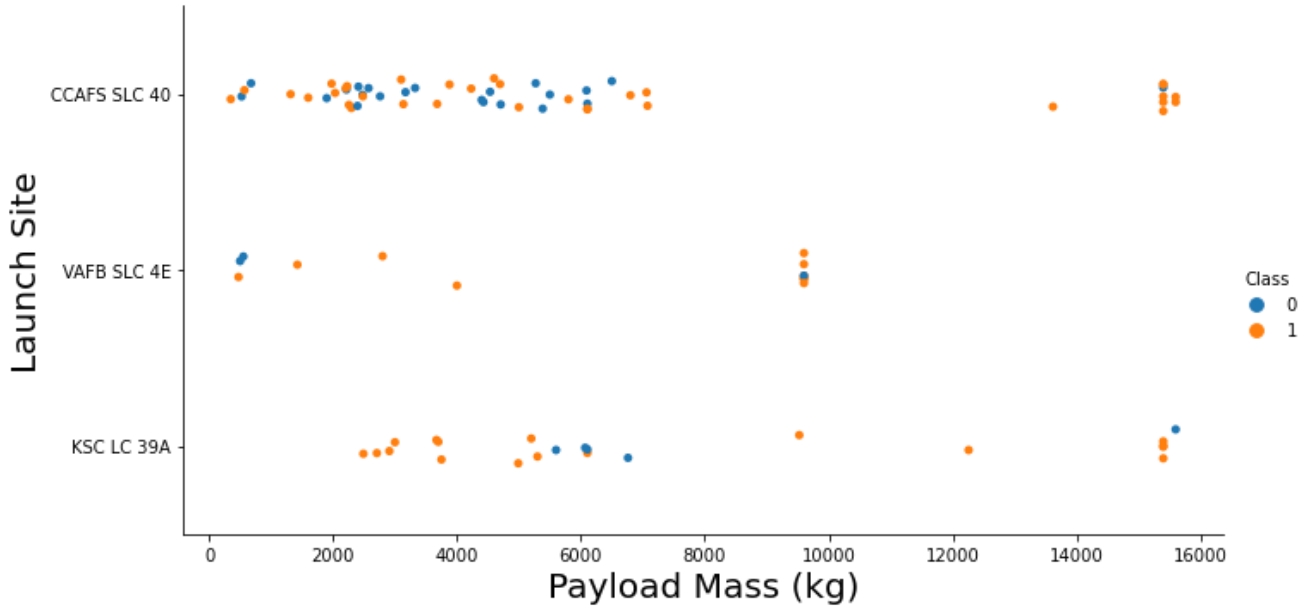
Flight Number vs. Launch Site

- We can see that the success rate of landings increases over time.
- The VAFB SLC 4E launch site doesn't seem to be used anymore.
- The main launch site is now the CCAFS SLC 40 launch site



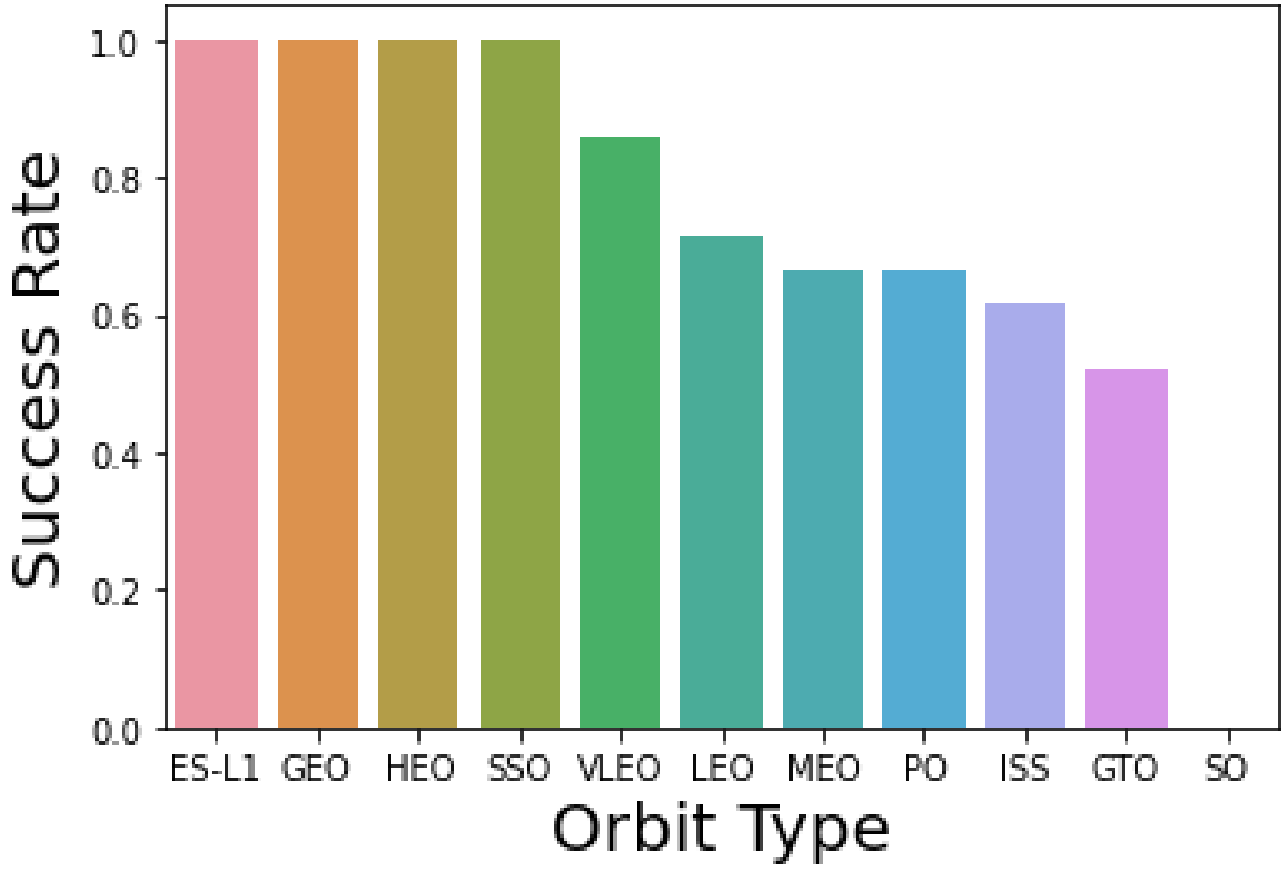
Payload vs. Launch Site

- High mass payloads are only launched from the CCAFS SLC 40 and DSC LC 39A launch sites.
- There has been more launches from the CCAFS SLC 40 launch site.
- Some payload masses have occurred multiple times.



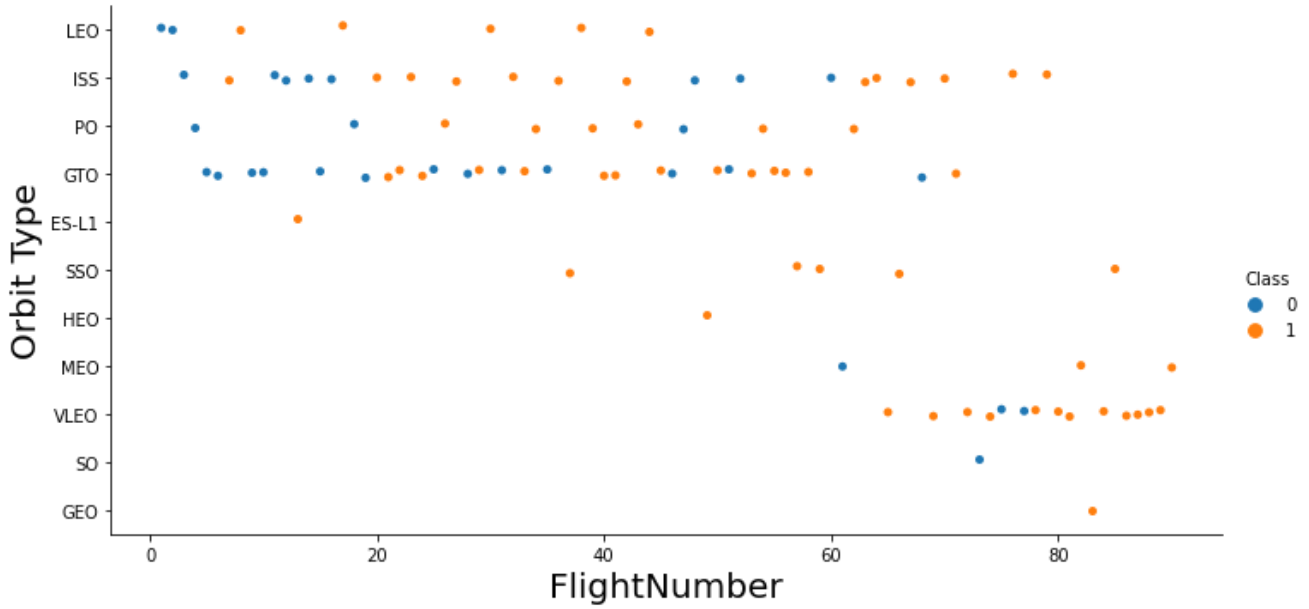
Success Rate vs. Orbit Type

- Some orbits have almost a 100% booster recovery rate:
  - ES-L1: Lagrange point L1
  - GEO: Geosynchronous Earth Orbit
  - HEO: Geocentric Earth Orbit
- Other orbits have varying booster recovery rate between 50-85%.
- SO is an alias for SSO, the Sun-synchronous orbit.



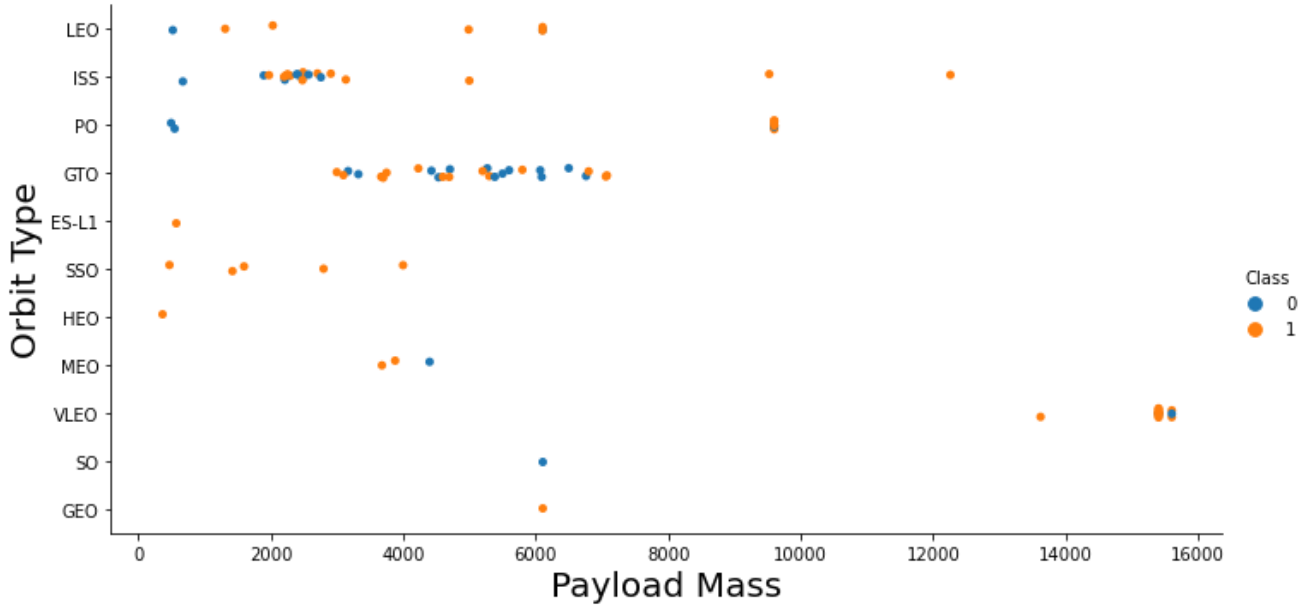
Flight Number vs. Orbit Type

- Most of recent launches have targeted the VLEO orbit (Very Low Earth Orbit).
  - Those are the Starlink satellite launches, aiming to make internet accessible everywhere on earth.
  - These satellites have low altitude to minimize latency of the provided internet service
- LEO, ISS, PO and GTO have historically been the most targeted orbits.



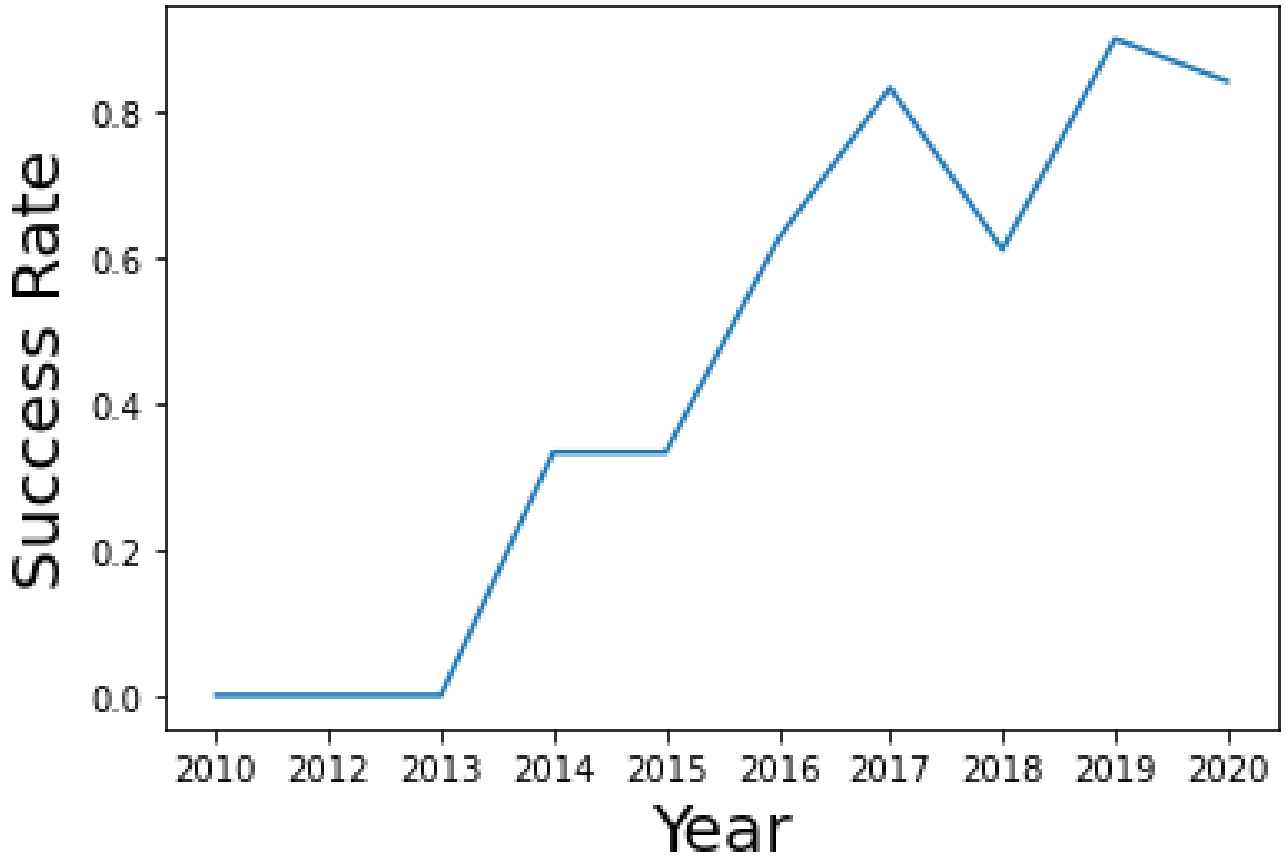
Payload vs. Orbit Type

- Payloads to Very Low Earth Orbit (VLEO) tend to maximize payload mass.
- Payloads to the ISS usually have low payload mass.
- The Geostationary Earth Orbit (GEO) is the orbit having highest payload mass variance.



Launch Success Yearly Trend

- Here we clearly see an increasing trend in the booster recovery success rate over time.
- SpaceX has gotten experience from its previous recovery attempts.
- Competitors attempting to do the same must expect failing their first recovery attempts, so they must ensure they have proper financing during their first years.



All Launch Site Names

- SpaceX uses four different launch sites for its Falcon 9 launches.

```
%%sql
select distinct LAUNCH_SITE from SPACEXTBL;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-464
pdomain.cloud:30699/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- We used the LIMIT keyword to get only the first five results to avoid getting too many results.

```
%%sql
select * from SPACEXTBL
where LAUNCH_SITE like 'CCA%'
limit 5;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/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

- We summed the payload mass to get the total.
- Entries were filtered to only include those having NASA as customer.

```
%%sql
select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL
where CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Entries are filtered to exclude other booster versions.
- Then the average is computed.

```
%%sql
select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL
where BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8
Done.
```

1
2928

First Successful Ground Landing Date

- Results are first filtered for successful landings on a ground pad.
- Results are then sorted by date and the first entry (earliest date) is selected.
- We could also have used min(Date) to obtain the same result.

```
%%sql
select DATE from SPACEXTBL
where LANDING__OUTCOME = 'Success (ground pad)'
order by DATE
limit 1;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8ab
Done.
```

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Entries are filtered to only include desired entries.
- The first condition is to be a successful drone ship landings.
- The second condition is to have a payload mass between 4000 and 6000 kilograms.

```
%%sql
select BOOSTER_VERSION from SPACEXTBL
where LANDING__OUTCOME = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ between 4000 AND 6000;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8aba
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Since there are multiple sites where a successful or failed landing is possible, we first extract the first 7 characters of the landing outcome.
- We then only keep entries where these first 7 characters are equal to `Success` or `Failure`.
- The remaining entries are then grouped and counted by outcome.

```
%%sql
select outcome, count(*) as "count"
from (
    select SUBSTRING(LANDING__OUTCOME, 1, 7) as outcome
    from SPACEXTBL
    where SUBSTRING(LANDING__OUTCOME, 1, 7) in ('Failure', 'Success')
    )
group by outcome;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3
Done.
```

outcome	count
Failure	10
Success	61

Boosters Carried Maximum Payload

- Entries are filtered to exclude all those that don't have a payload mass equal to the maximum payload mass of our dataset.
- As a result, we see that only Falcon 9 Block 5 boosters have launched payloads with the highest mass.

```
%%sql
select BOOSTER_VERSION from SPACEXTBL
where PAYLOAD_MASS__KG_ = (
    select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL
);
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8ab
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

- First we filter entries to retrieve launches that occurred in 2015 with a failed drone ship landing.
- We see that only Falcon 9 v1.1 boosters launched from the CCAFS LC-40 launch site meet these criteria.

```
%%sql
select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
from SPACEXTBL
where YEAR(DATE) = 2015
    AND LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8aba-9dcf
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Entries are filtered to meet the 2010-06-04 to 2017-03-20 time window.
- Resulting entries are then grouped by landing outcome and counted.
- The final result set is obtained by sorting landing outcomes by decreasing counts.

```
%%sql
select LANDING__OUTCOME, count(*) as "count"
from SPACEXTBL
where DATE between '2010-06-04' AND '2017-03-20'
group by LANDING__OUTCOME
order by COUNT DESC;
```

```
 * ibm_db_sa://dyv12776:***@19af6446-6171-4641-8aba
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

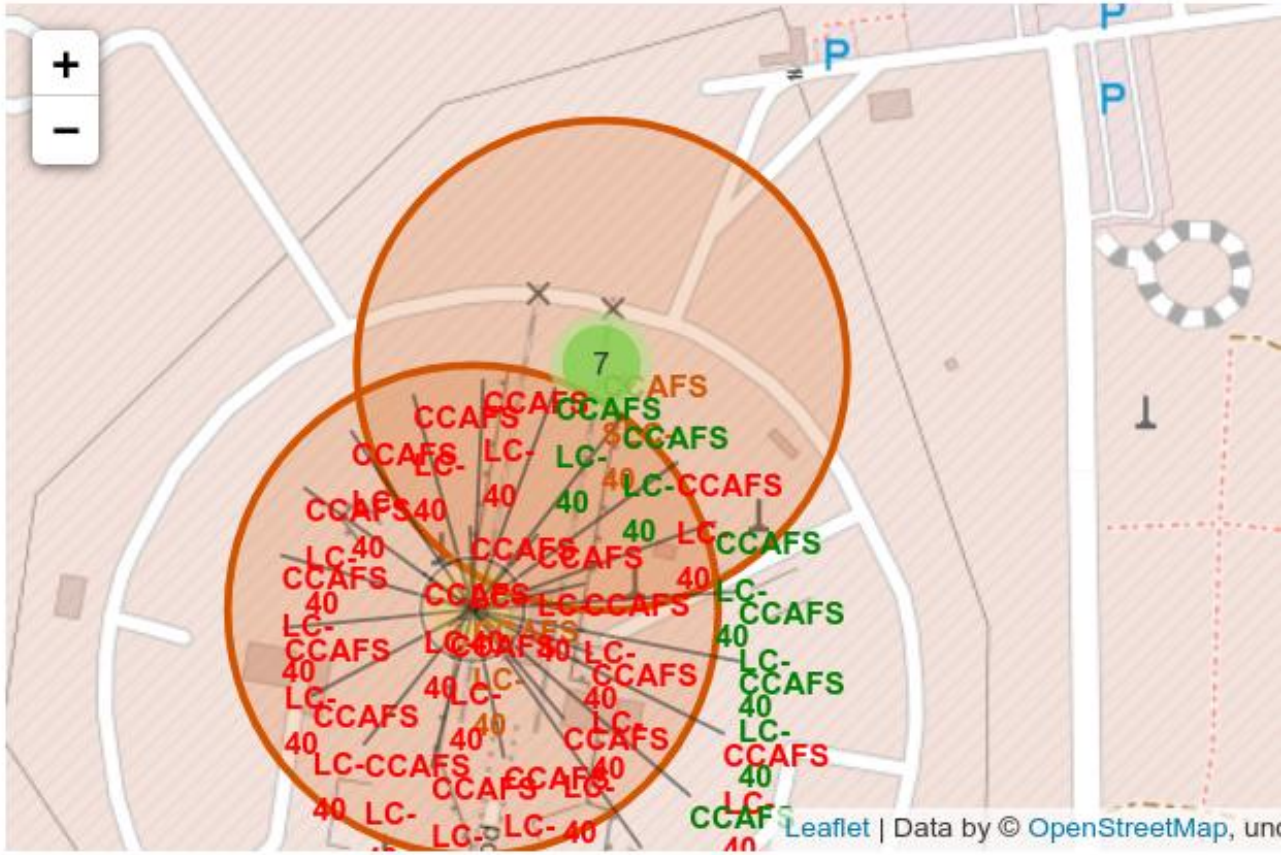

Section 3
Launch Sites
Proximities Analysis

Location of SpaceX launch sites on a global map

- All SpaceX launch sites are located in the United States.
- Launch sites are in coastal area.
- Launch sites are near the equator.

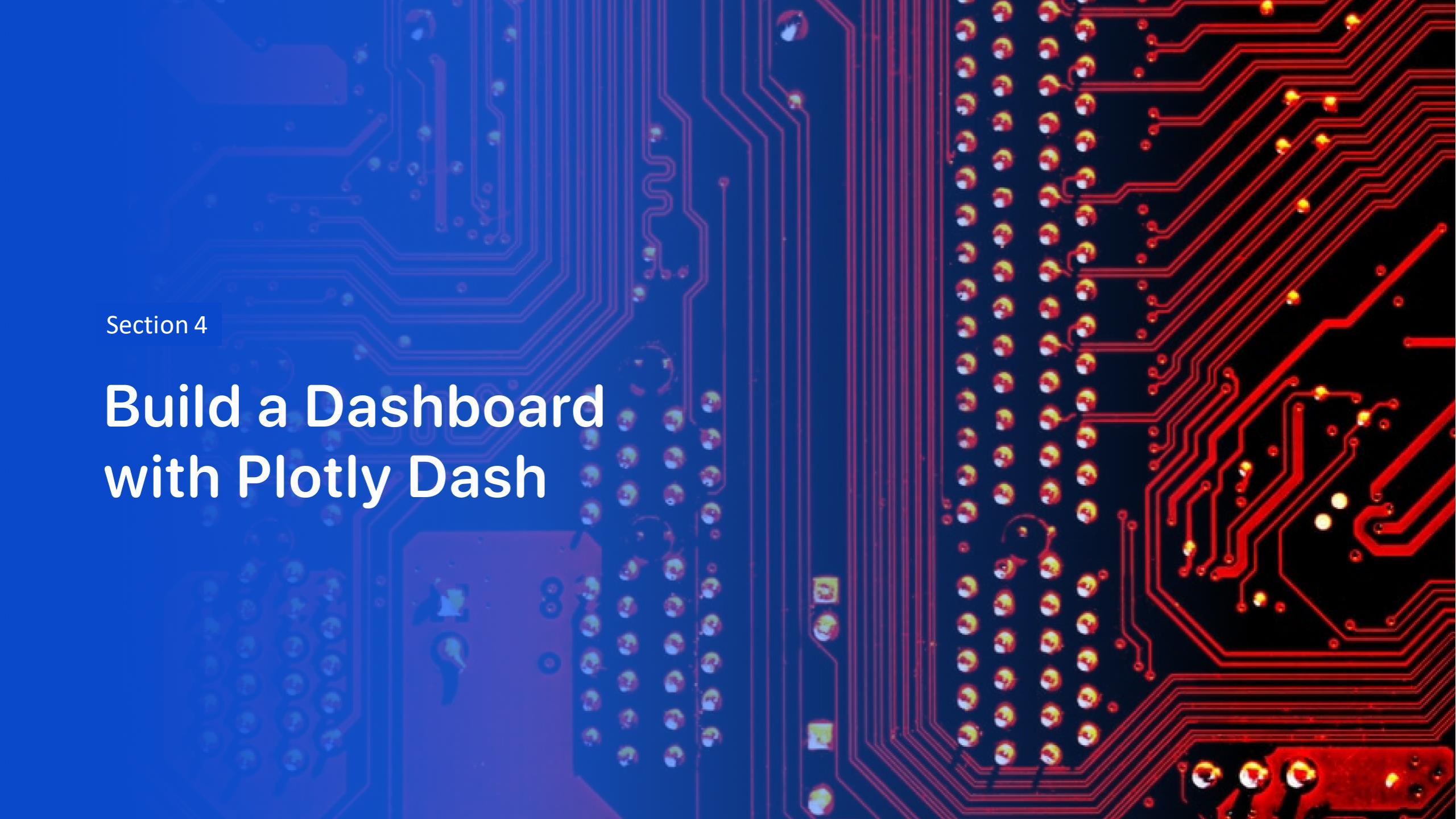
Outcomes of launches by launch site

- Colors make it easy to determine which launches were successes or failures.
- The numbers on the launch sites make it easy to determine how many boosters were launched from each site.
- We quickly see from the green color of the upper launch site that all launches from the site were successful.

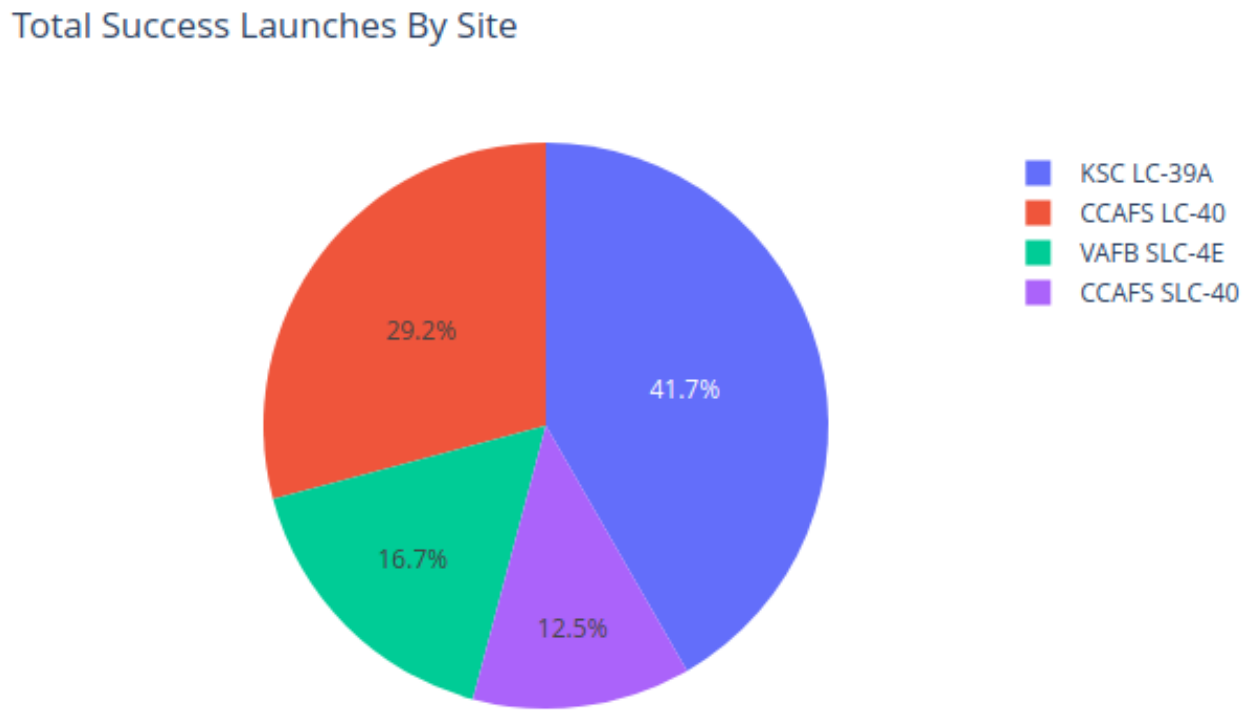


<Folium Map Screenshot 3>

- Coastline and railway aren't far from the launch site.
- The nearest city is much further away (17.9 km).

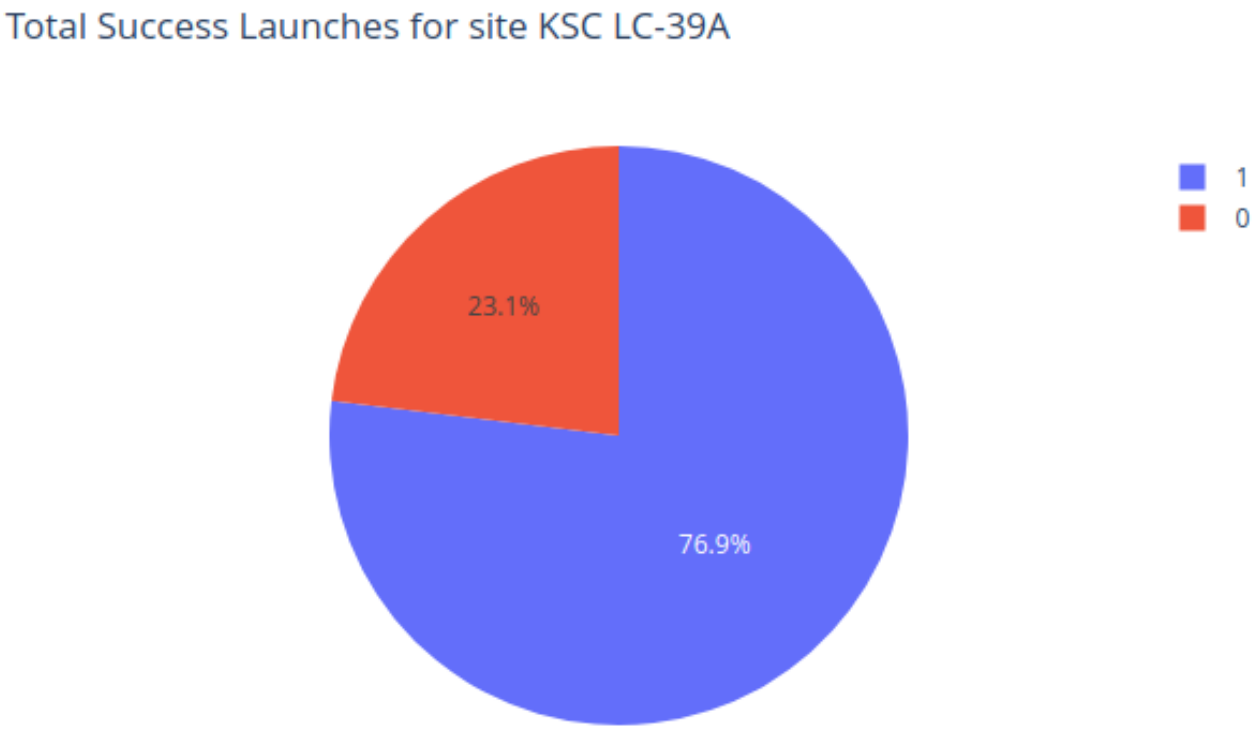

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- The KSC LC-39A launch site has contributed much more than other launch sites to the number of successful booster recovery.
- Then comes the CAFS LC-40 launch site.
- The VAFB SLC-4E and CCAFS SLC-40 launch sites are those contributing the less to the successful booster recovery count.

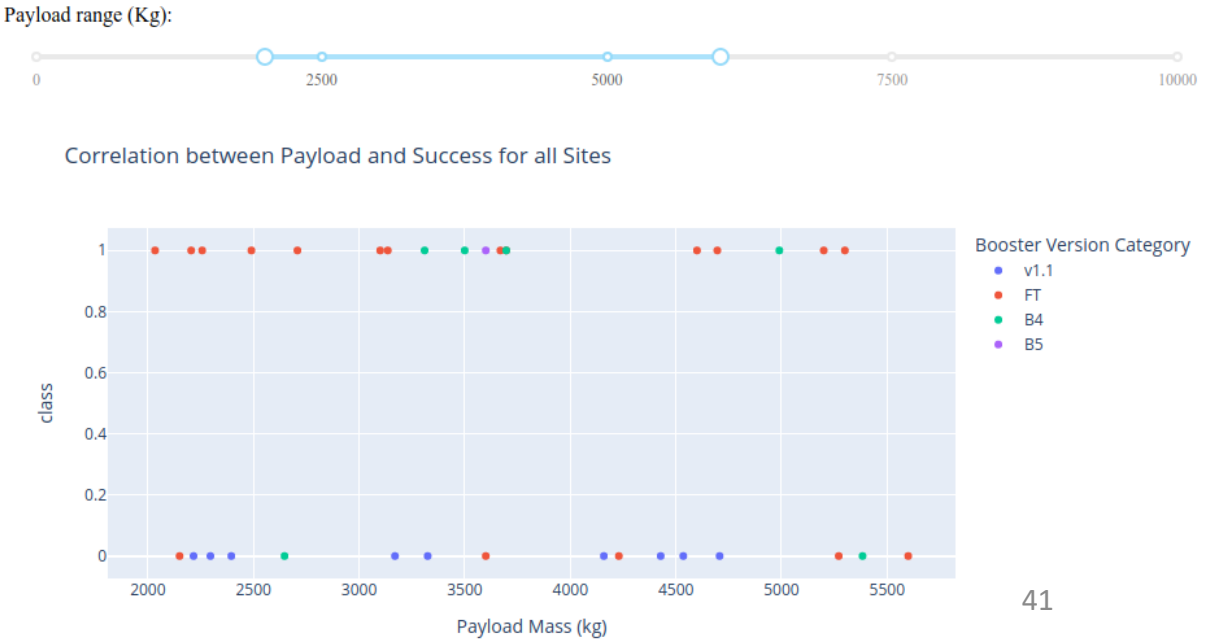
Landings for the most successful launch site



- The KSC LC-39A launch site is the one with the highest booster landing success rate.
- Boosters launched from this site have successfully landed 76.9% of the time.

Landing status by launch site and booster version

- Launches with payload mass in the 2000 to 6000 kg range were the ones with the highest recovery rate.
- Booster version 1.0 and 1.1 were those with most of the failed landings.
- The Full Trust (FT), Block 4 and Block 5 versions were recovered successfully most of the time.
- Block 4 boosters were those with the highest successful landing rate.

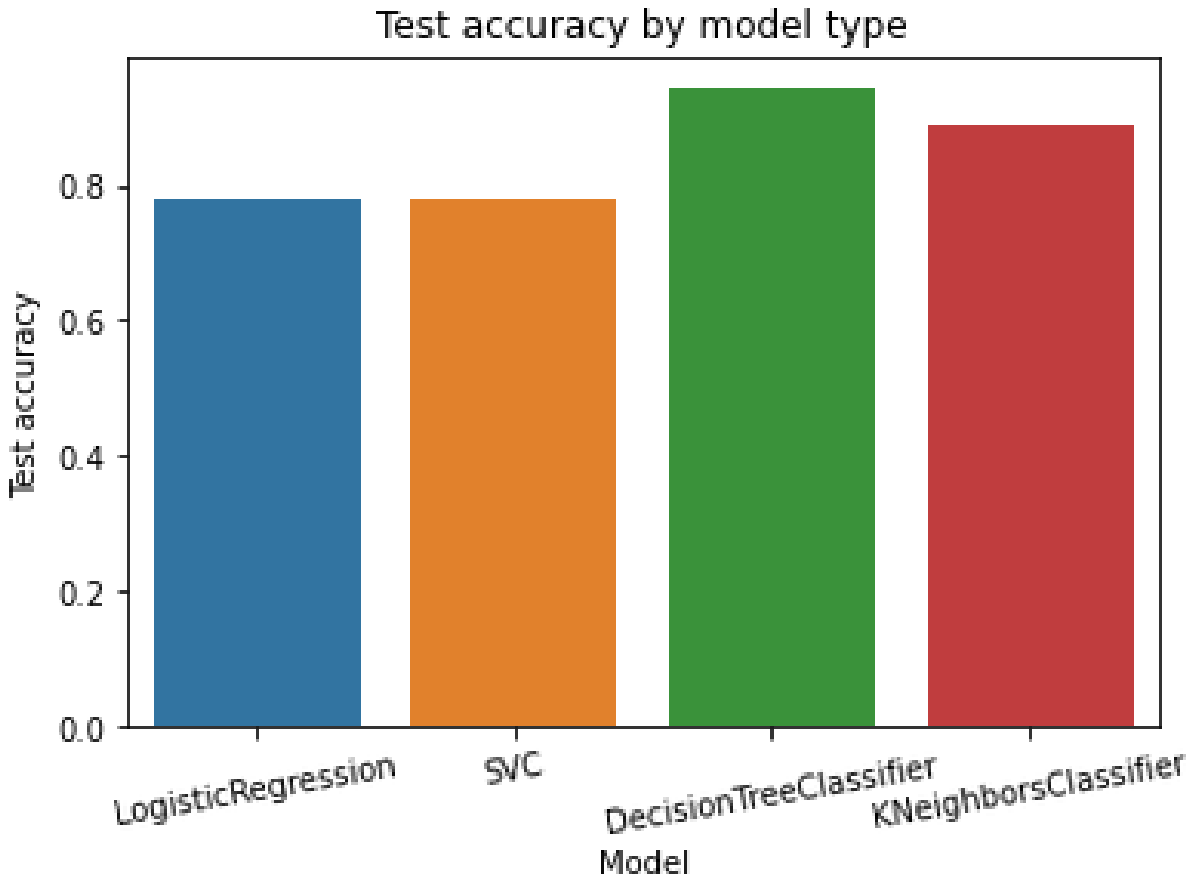


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The best model is the Decision Tree Classifier.
- It is followed by the KNN Classifier.
- Then come the Logistic Regression and the Support Vector Machine.



Confusion Matrix of the Decision Tree Classifier

- The Decision Tree Model performed very well on the held out test set, achieving an accuracy of 94.4%.
- It only made a single wrong prediction, a false negative where the model predicted that the booster did not land successfully while it actually did.

Conclusions

- In this project, we were able to:
  - Create a dataset about SpaceX Falcon 9 launches and landings, using both SpaceX's API and data scraped from Wikipedia.
  - We did Data Wrangling to prepare our data, by filtering for Falcon 9 launches and replacing missing values.
  - We performed Explanatory Data Analysis (EDA) to retrieve valuable insights from our data, by both creating insightful visualizations and performing queries using SQL.
  - Predict the landing outcome of a launch using this data, with a success rate of 94.4% using our best model.
- We are now (almost) ready to compete with SpaceX in sending payload to space at minimal cost.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!