Winning Space Race with Data Science

Mario Alpízar
January 28th, 2024

Outline

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

Executive Summary

- Summary of methodologies:

Diverse methodologies were used, including: web scraping, api usage, exploratory data analysis, data wrangling, visualization generation, feature engineering and prediction with classification algorithms

- Summary of all results

Important information for the success of SpaceY was obtained, including the possibility of predicting wether a rocket landing will be successful and, therefore, being able to predict its cost

Introduction

- This projects main focus is to determine wether SpaceX's rocket landings can be predicted, so that SpaceY can obtain information for the cost of the launches and use that information to create a budget for its own launches and compete with SpaceX.

Section 1

Methodology

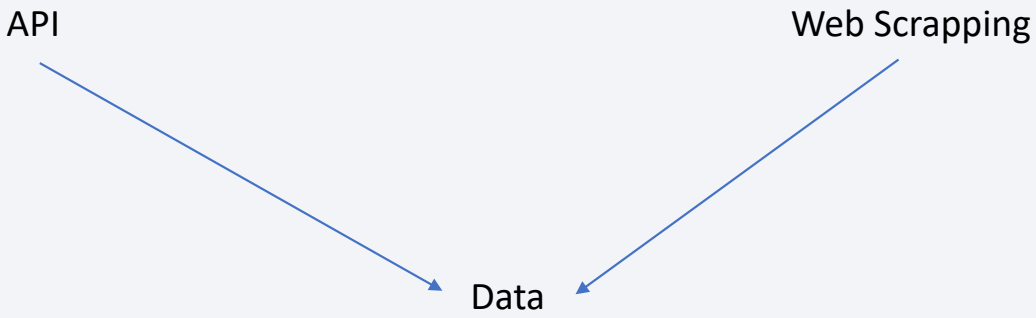
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected with 2 main methods: using SpaceX's API and web scrapping from publicly available information for SpaceX's launches
- Perform data wrangling
  - Data was wrangled by replacing nan or missing values were necessary, as well as performing feature engineering with tools like one hot encoding
- Exploratory data analysis (EDA) was performed using visualization and SQL
- Additional visual analytics using Folium were Included in the Analysis
- Predictive analysis was performed using classification models
  - Various models were built, tuned, and compared, to get the best results

Data Collection

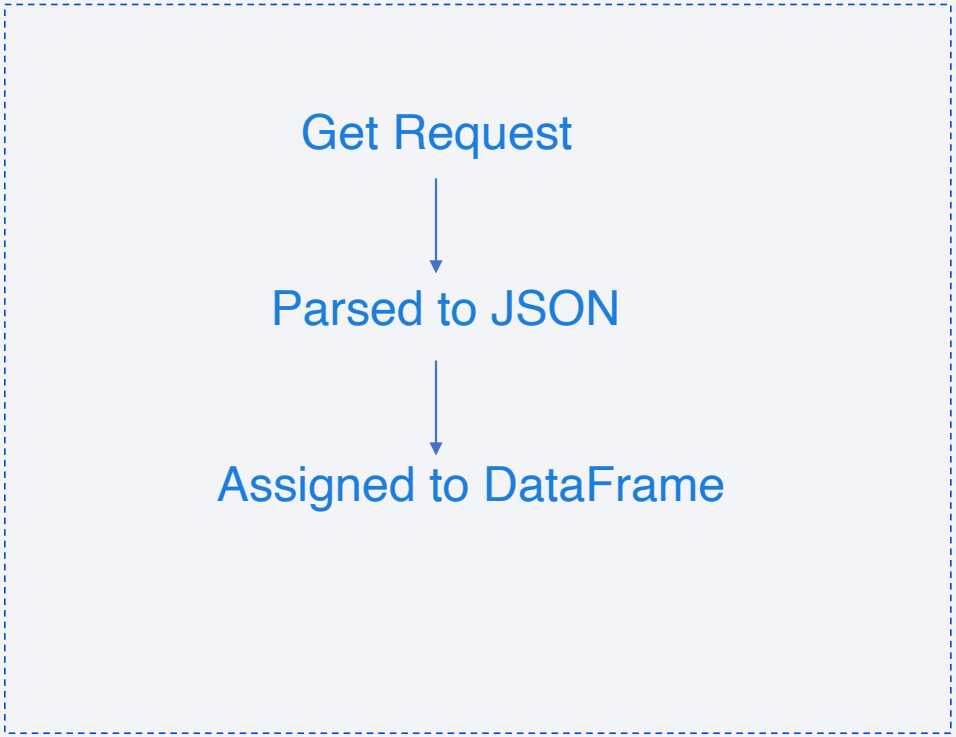
- Data was collected via 2 main methods:
- 1) Data extraction using SpaceX's data available trough their API
- 2) Web scraping from a Wikipedia page that included important details about SpaceX's launches not available trough their API



Data Collection – SpaceX API

- The data was initially collected using a get request
- Then it was parsed to a json file
- Finally it was assigned to a pandas DataFrame

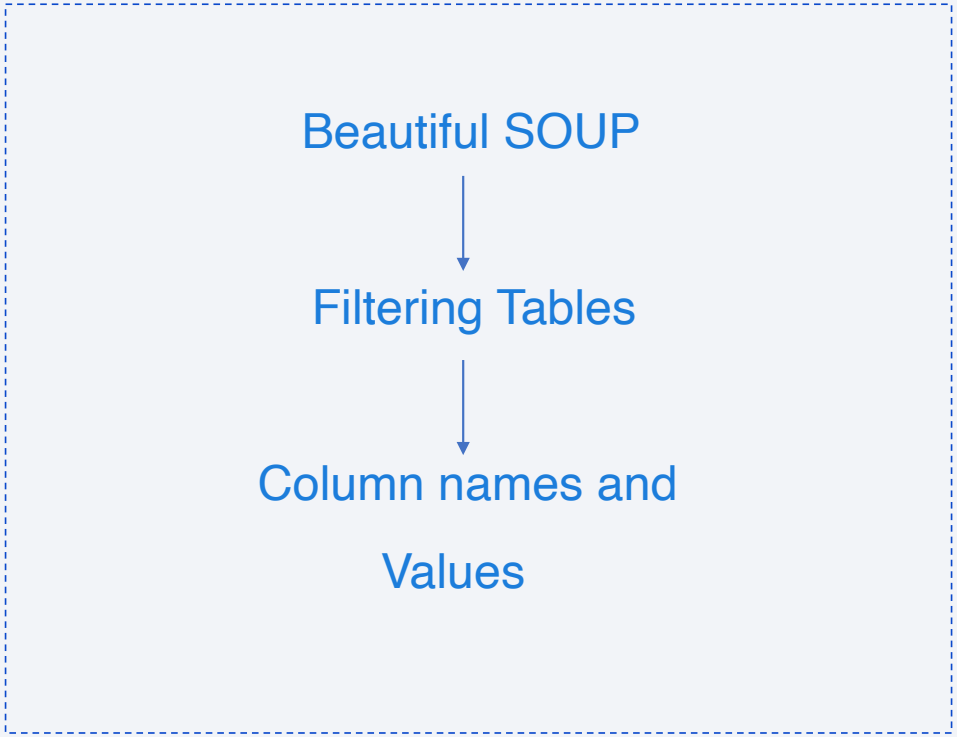
- Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · marioalpizarv/Capstone (github.com)



Data Collection - Scraping

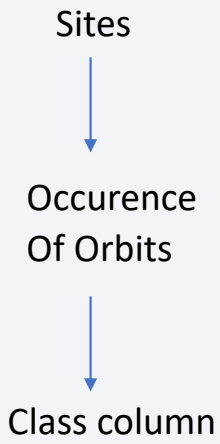
- Data was parsed from the URL with a Beautiful SOUP object
- Tables were filtered with find_all
- Column names and values were obtained from the tables

- Capstone/jupyter-labs-webscraping.ipynb at main · marioalpizarv/Capstone (github.com)



Data Wrangling

- The number of launches from each site was calculated
- Also the number of missions to each different orbit
- An binary class column was created with the outcomes of the missions
- [Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · marioalpizarv/Capstone (github.com)]



EDA with Data Visualization

- The following charts were plotted for EDA:
  - Flight number vs Payload mass, launch site vs flight number, launch site vs payload mass, bar chart of the mean of class for each orbit, orbit vs flight number and orbit vs payload mass
  - The purpose of these plots is to identify trends and correlations between the variables
- Capstone/jupyter-labs-eda-dataviz.ipynb at main · marioalpizarv/Capstone (github.com)

EDA with SQL

- Queries performed:
- Unique occurrences of launch sites
- Payload mass carried by boosters launched by NASA CRS
- Average payload mass carried by Falcon 9 v1.1 boosters
- List the names of boosters with number of successful landings in drone ship greater than 4000 but less than 6000

- Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · marioalpizarv/Capstone (github.com)

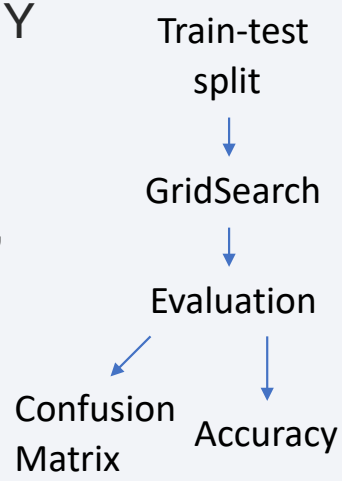
Build an Interactive Map with Folium

- Markers were added at the locations of each launch site
- Color coded circles were added at the launch site of each mission: green for successful and red for unsuccessful
- This was done to identify patterns and trends in successful and unsuccessful missions related to the location they were launched from, and to identify useful information about the landing sites themselves, such as proximity to the equator, the coast, railroads, highways, and the safety distance from populated areas

- Capstone/lab_jupyter_launch_site_location.ipynb at main · marioalpizarv/Capstone (github.com)

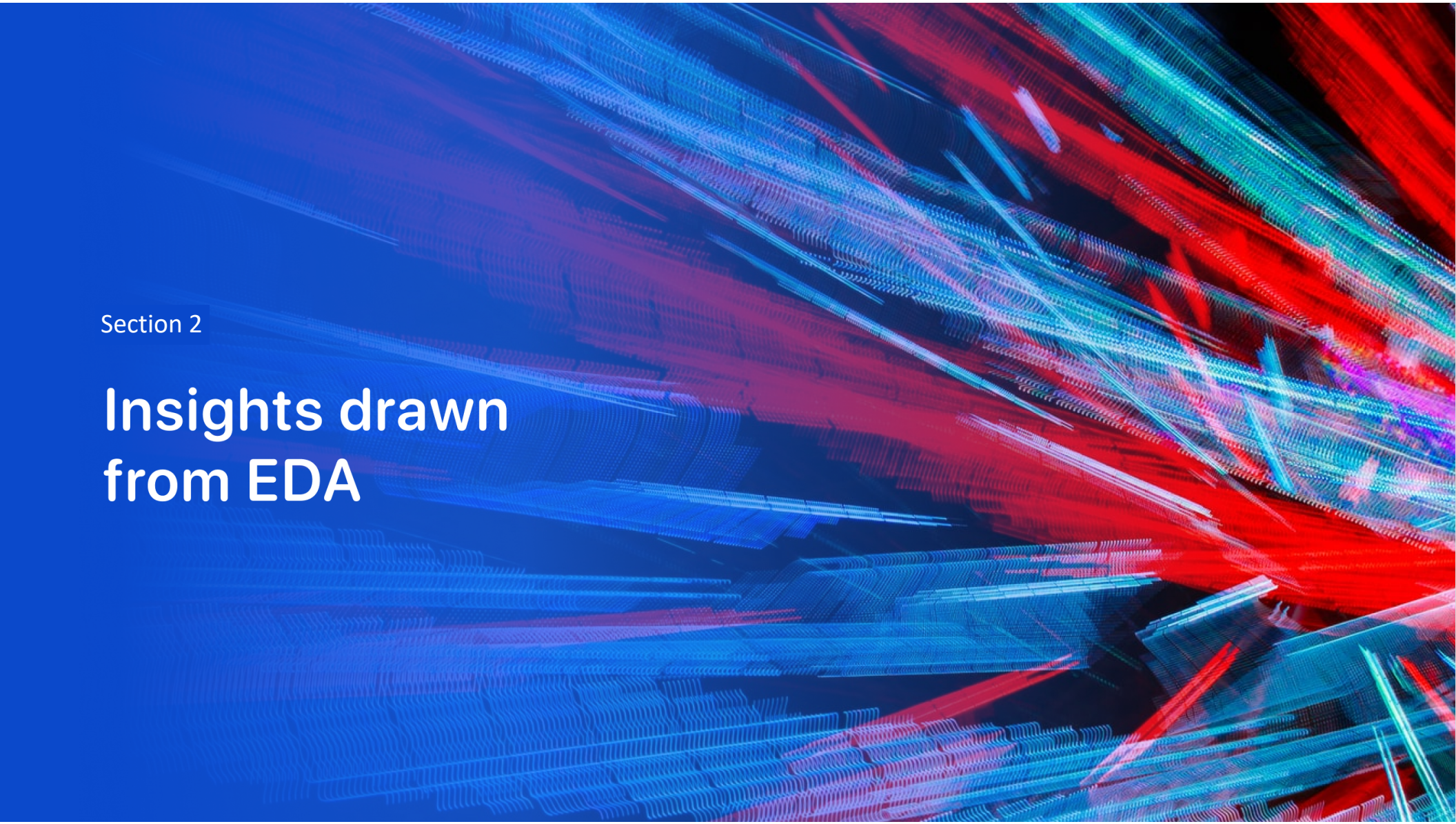
Predictive Analysis (Classification)

- The data was split in independent variables X and the dependent variable Y which was a binary classification binary, 1 for successful landing and 0 for unsuccessful landings. Then it was split into training and test sets.
- Four different models were build, tuned and compared: logistic regression, support vector classifier, decision tree and k-nearest neighbors.
- They were tuned using grid search to find the best set of parameters
- The models were evaluated and compared using the accuracy metric, as well as the confusion matrix to visualize the true positives, true negatives, false positives and false negatives
- Capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · marioalpizarv/Capstone (github.com)



Results

- The most solicited laugh orbit was geostationary transfer (GTO)
- For almost all launch sites the success rate progressively increased with the flight number, meaning that the launch and landing methodology was perfected with experience.
- The most used launch site was CCAFS LC-40, while the least used was CCAFS SLC-40
- The variables with the highest correlation to the success of the launch are: wether there are legs on the rocket, if grid fins were used and the count of reused times.
- The model that was ultimately selected for prediction was the support vector classifier

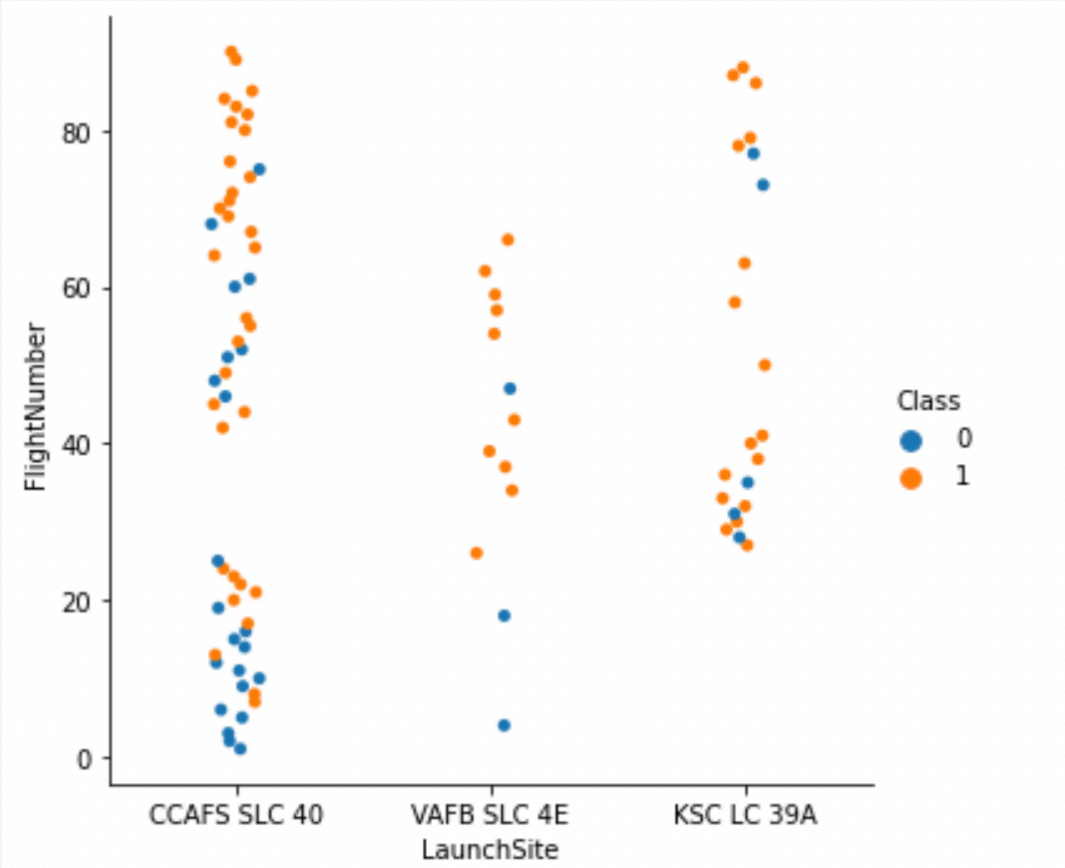


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

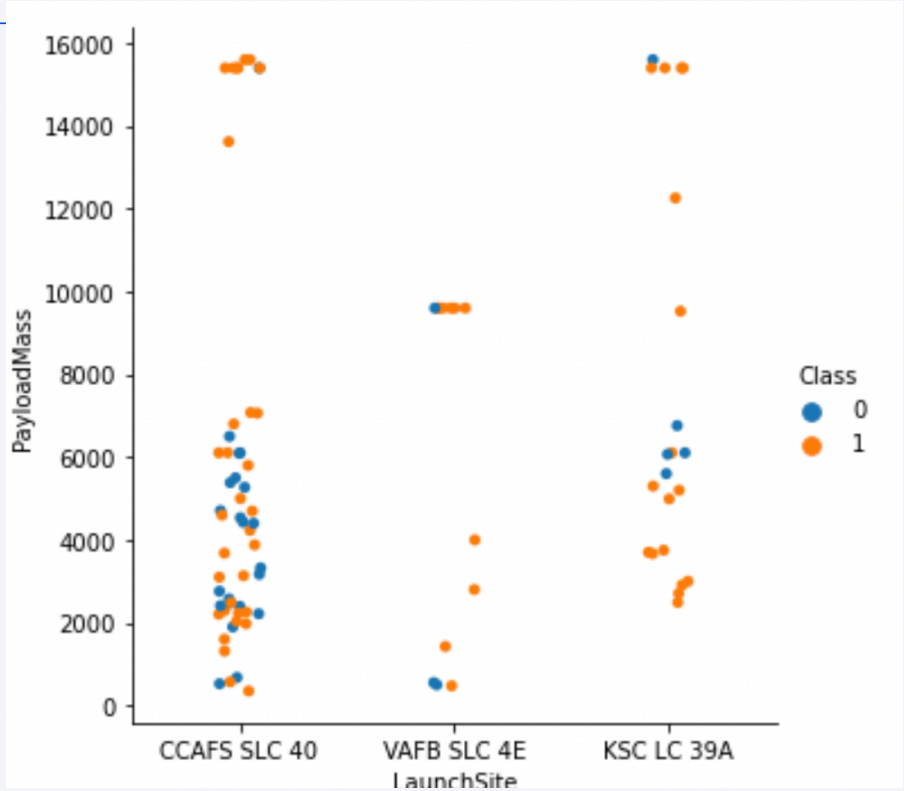
- For almost all launch sites the success rate increases with launch number, meaning that the increased experience worked for all launch sites
- The launch site CCAFS SLC 40 does show a higher number of unsuccessful launches at higher flight numbers



Flight Number vs Launch Site. 1: successful, 0: unsuccessful

Payload vs. Launch Site

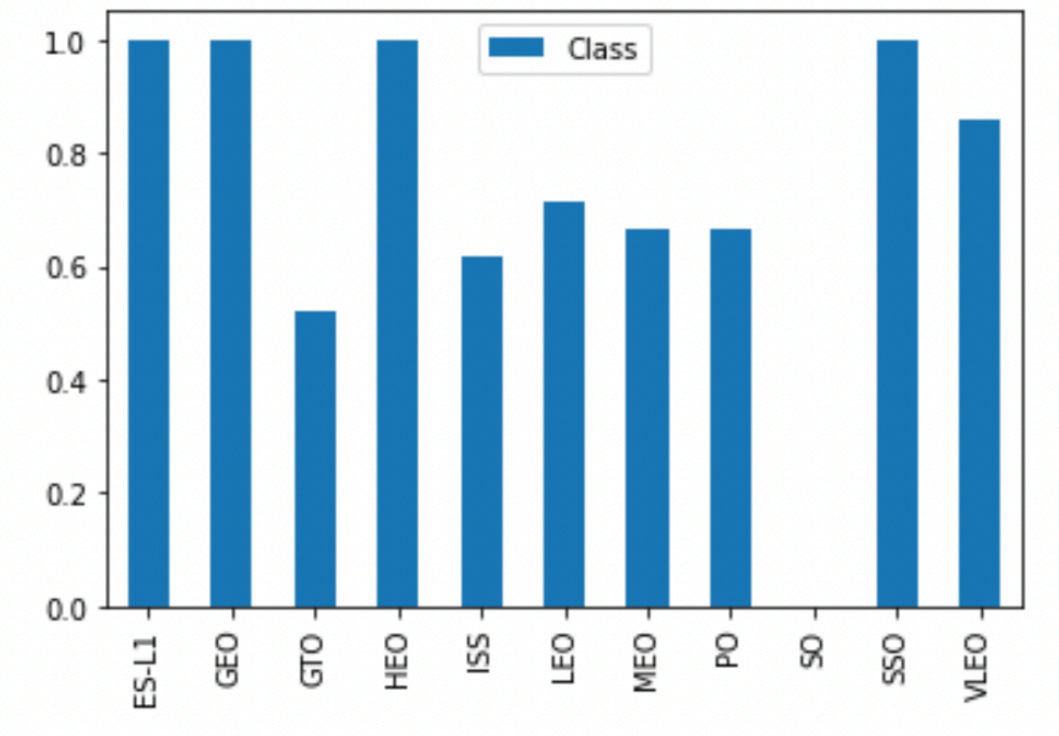
- The launch site that takes care of the highest payload mass launches is CCAFS SLC 40
- The launch site VAFB SLC 4E was not used for launches with a payload mass higher than 10000 kg



Payload vs Launch Site. 1: successful, 0: unsuccessful

Success Rate vs. Orbit Type

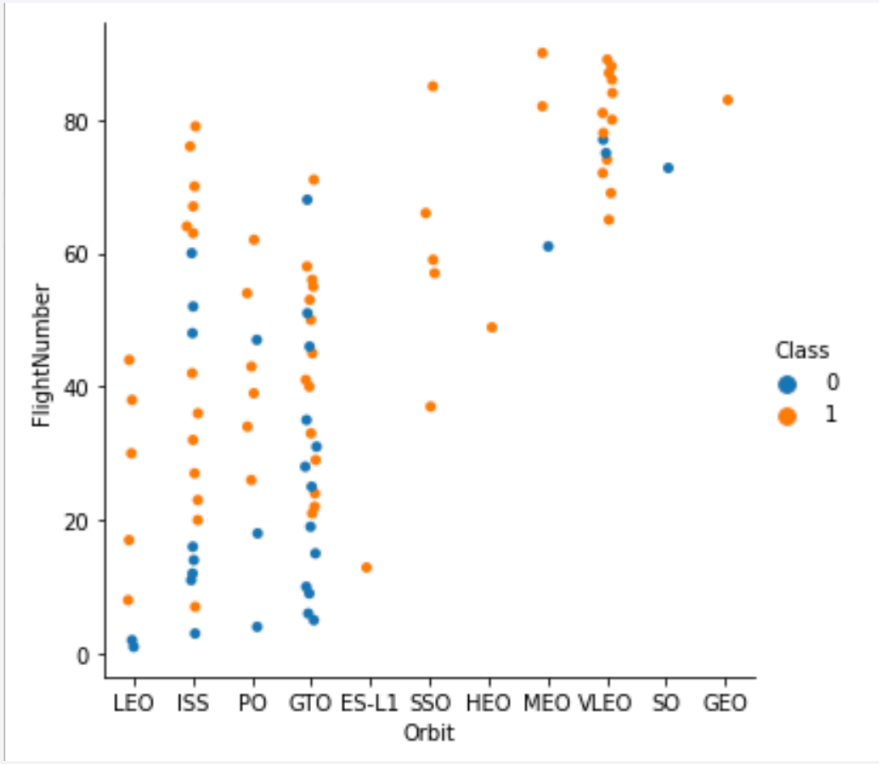
- The orbit types with the highest success rates are: ES-L1, GEO, HEO and SSO
- The orbits with the lowest success rates are GTO and SO
- Orbits ISS, LEO, MEO, VLEO and PO all have success rates higher than 60%



Success rate per launching site

Flight Number vs. Orbit Type

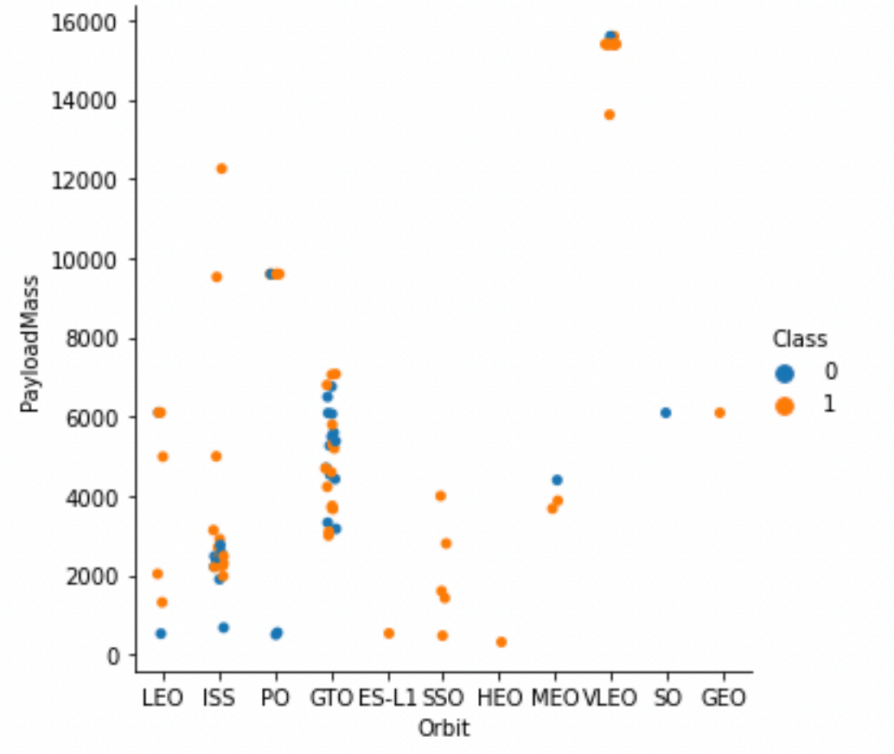
- For all orbit types, the successes increase with flight number.
- For the orbit SSO there are only successful launches.
- The success rate for the orbit GTO does not improve significantly with flight number



Orbit vs Flight number. 1: successful, 0: unsuccessful

Payload vs. Orbit Type

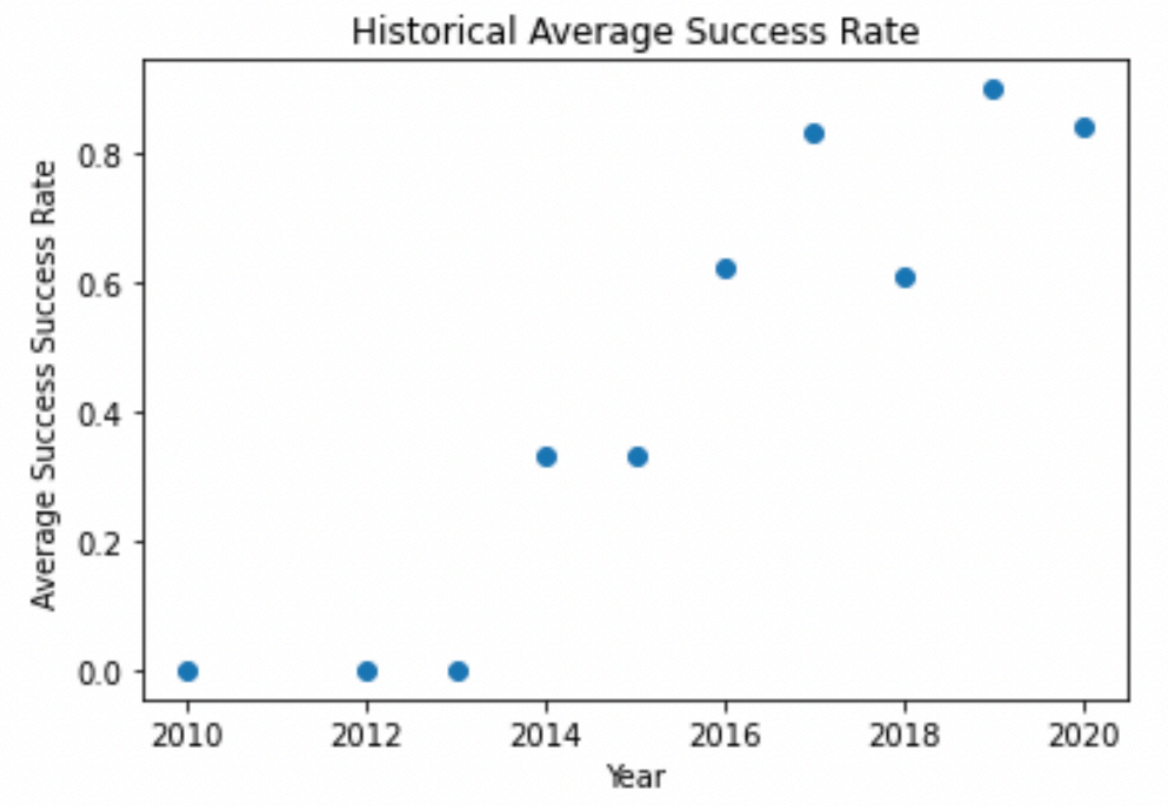
- The orbit were the highest payload masses missions are carried is VLEO
- The missions shipped to GTO have payload masses ranging from 4000 kg to 8000 kg



Orbit type vs Payload mass. 1: successful, 0: unsuccessful

Launch Success Yearly Trend

- The average success rate has clearly increased overtime, meaning that experience was drawn from previous missions.
- Initially the average success rate was close to 0, but after approximately 10 years it was increased to 80%



Average success rate over time.

All Launch Site Names

- The ‘distinct’ discriminator is used to show only the different launch sites
- There were 4 different launch sites in this data set

```
%%sql

select distinct Launch_Site from SPACEXTBL
```

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

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

Query for unique launch sites

Launch Site Names Begin with 'CCA'

- The % symbol was used to indicate that the name of the launch site must start with ‘CCA’
- The limit function was used to show only 5 results

```
1 %sql select * from SPACEXTBL 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	M
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Query for launch site names that begin with ‘CCA’

Total Payload Mass

- Customers by total payload mass ordered alphabetically

```
1 %sql select Customer, sum("PAYLOAD_MASS__KG_") as "Total Payload (kg)" from SPACEXTBL
```

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

Customer	Total Payload (kg)
ABS Eutelsat	7759
AsiaSat	8963
Bulsatcom	3669
CONAE	3000
CONAE, PlanetIQ, SpaceX	3130

Customers by total payload mass

Average Payload Mass by F9 v1.1

- The result for the average payload mass carried by the Falcon 9 version 1.1 booster is 2928 kg

```
1 %sql select "Booster_Version", avg("PAYLOAD_MASS__KG_") as "Average Payload Mass (kg)"
2 from SPACEXTBL where "Booster_Version" = "F9 v1.1";
```

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

Booster_Version	Average Payload Mass (kg)
F9 v1.1	2928.4

Average payload mass carried by booster Falcon 9 version 1.1

First Successful Ground Landing Date

- The first successful ground pad landing was achieved on December 22nd 2015

```
%sql select Date, Landing_Outcome from SPACEXTBL where Landing_Outcome =
"Success (ground pad)" order by Date
```

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

Date	Landing_Outcome
2015-12-22	Success (ground pad)
2016-07-18	Success (ground pad)
2017-02-19	Success (ground pad)
2017-05-01	Success (ground pad)
2017-06-03	Success (ground pad)
2017-08-14	Success (ground pad)
2017-09-07	Success (ground pad)
2017-12-15	Success (ground pad)
2018-01-08	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The list of names of boosters from successful landings with a payload mass between 4000 kg and 6000 kg was obtained using the where clause with a conditional on the payload mass

```
%%sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTBL where
Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ > 4000
and PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Succesful landings with payload between 4000 kg and 6000 kg

Total Number of Successful and Failure Mission Outcomes

- The total number of successful missions was 100, but 1 out of those has a status of unclear on its payload

```
%sql select count(Mission_Outcome), Mission_Outcome from SPACEXTBL group by
Mission_Outcome
```

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

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

Successes and not successes

Boosters Carried Maximum Payload

- The boosters that have carried the max payload mass are all versions of the Falcon 9 B5
- The maximum mass carried by Space X is 15600 kg

```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Boosters that have carried the maximum payload mass

2015 Failed Launch Records on Drone Ship

- There were two failed landings attempted on a drone ship on 2015
- Both failed attempts occurred with a F9 v1.1 B1012 booster

```
%%sql

select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL where substr(Date, 0, 5) = '2015' and
Landing_Outcome = 'Failure (drone ship)';
```

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

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

Failed Launches from 2015 attempted on a drone ship

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

- The most common outcomes from that date range were: no attempt, success on drone ship and failure on drone ship

```
%%sql select Landing_Outcome, count(*) as count, Date from SPACEXTBL where Date betwee
'2017-03-20' group by Landing_Outcome order by count desc;
```

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

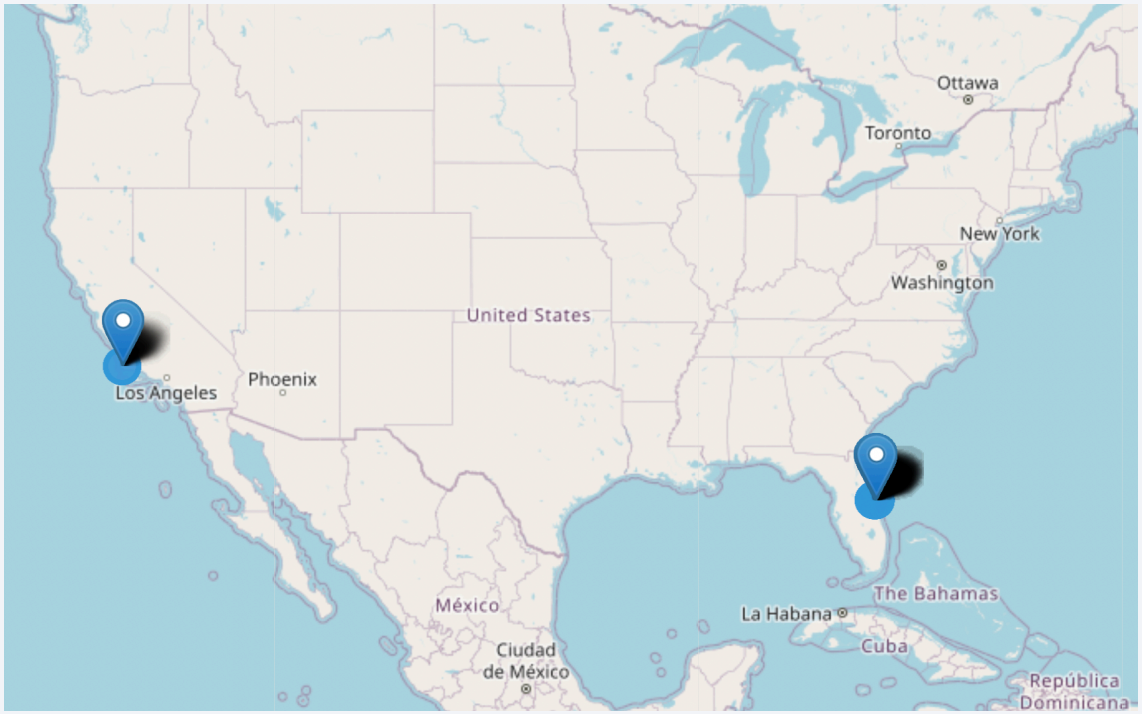
Landing_Outcome	count	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

Section 3

Launch Sites Proximities Analysis

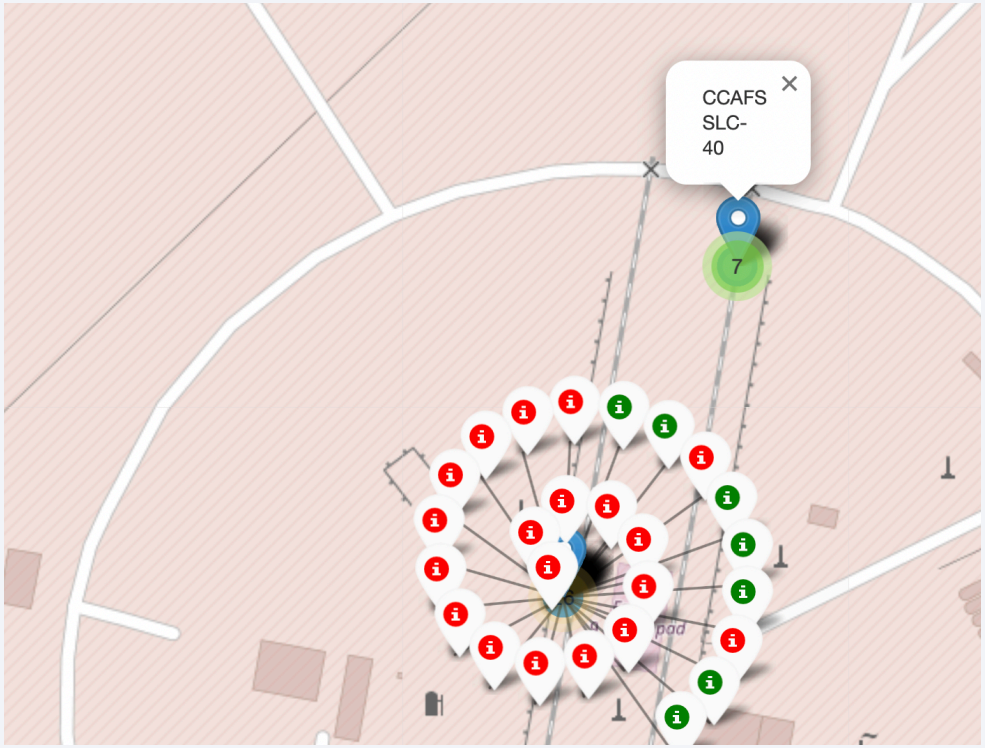
Location of Launch Sites on a National Scale

- The launch sites are located as near to the equator line as possible, as can be seen here that all of them are in the southern regions of the continental USA.
- This is done to take advantage of the rotational speed given to the rockets by Earth's rotation.



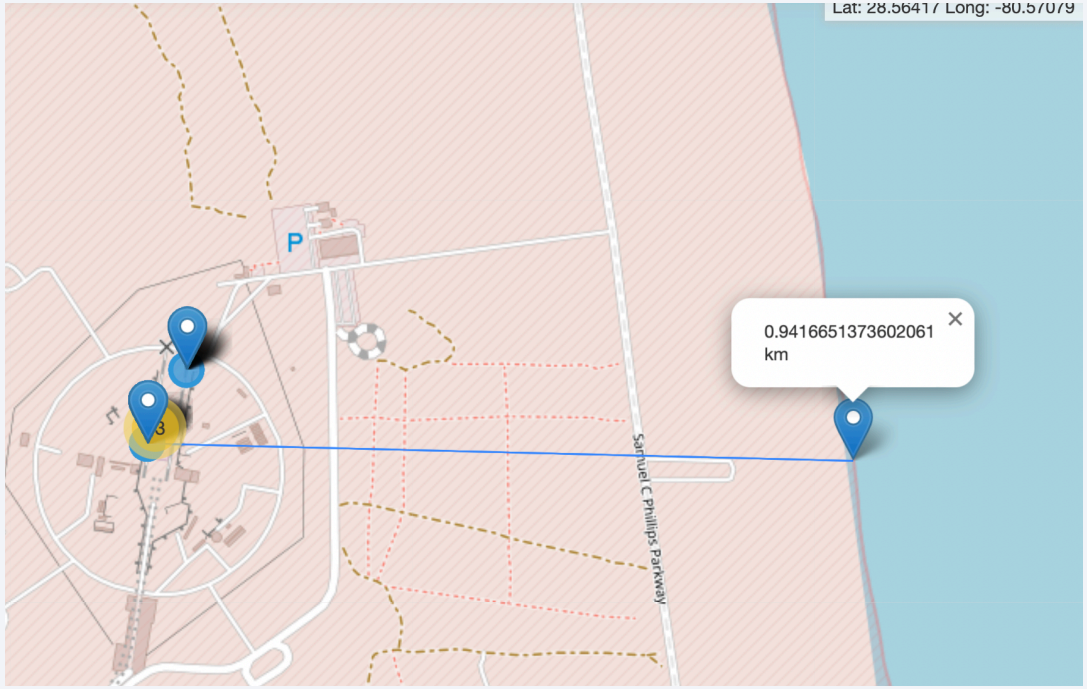
Outcome of Launches from CCAFS Launch Sites

- The launch site CCAFS SLC-40 shows 7 green colored launches, indicating that they were all successful
- The launch site CCAFS LC-40 shows more launches, with the majority of them being unsuccessful. The success rate for this site is 27%



Appropriate Location of a Launch Site

- The launch sites used by Space X all have a coast line nearby
- They also are near railroads or paved roads that facilitate the transport of the rockets and their payload
- The sites are also far away from cities and other inhabited areas for safety reasons



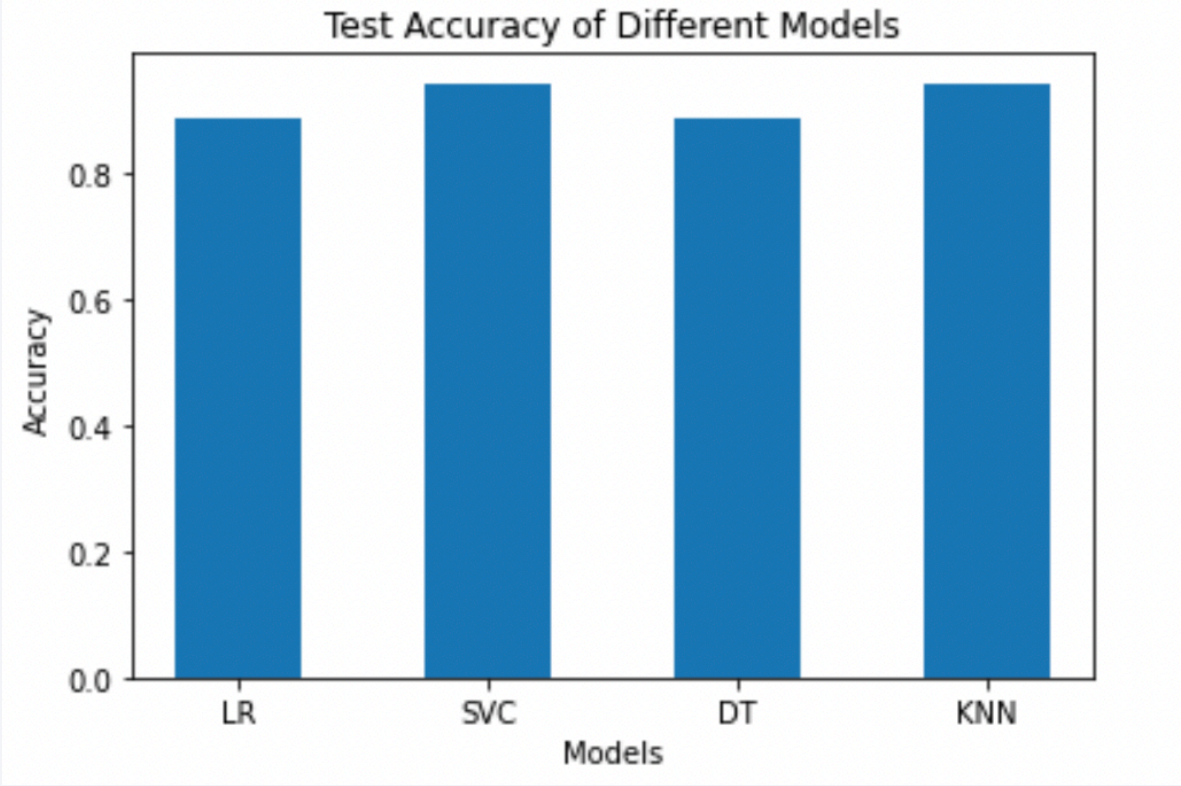
Distance from the launch site CCAFS LC 40 to the coastline.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

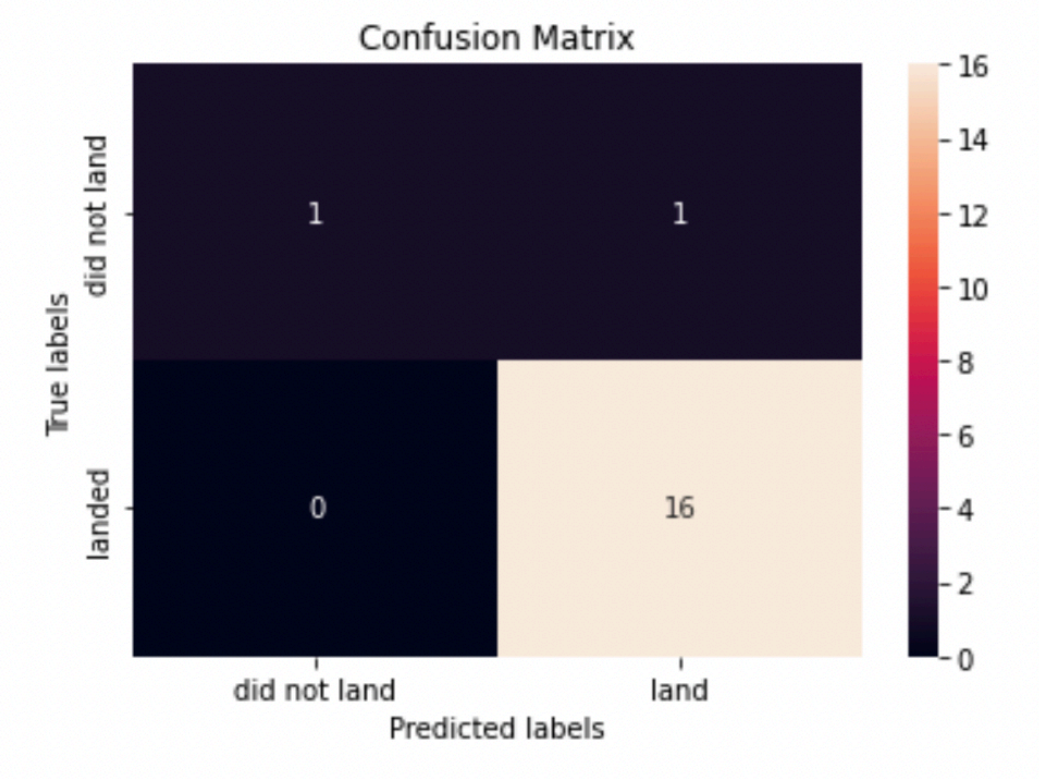
- The accuracy of all models evaluated on the test data set was high
- The models with the least accuracy were logistic regression and decision tree
- The highest test accuracy was by the support vector classifier and the k-nearest neighbors



Accuracy for each model evaluated on the test data set

Confusion Matrix for the Support Vector Classifier

- The best performing model was the SVC with a test accuracy of 94.4%
- It correctly classified 16 successful landings and 1 unsuccessful landing.
- It incorrectly classified only 1 landing, it was unsuccessful and was predicted as successful.



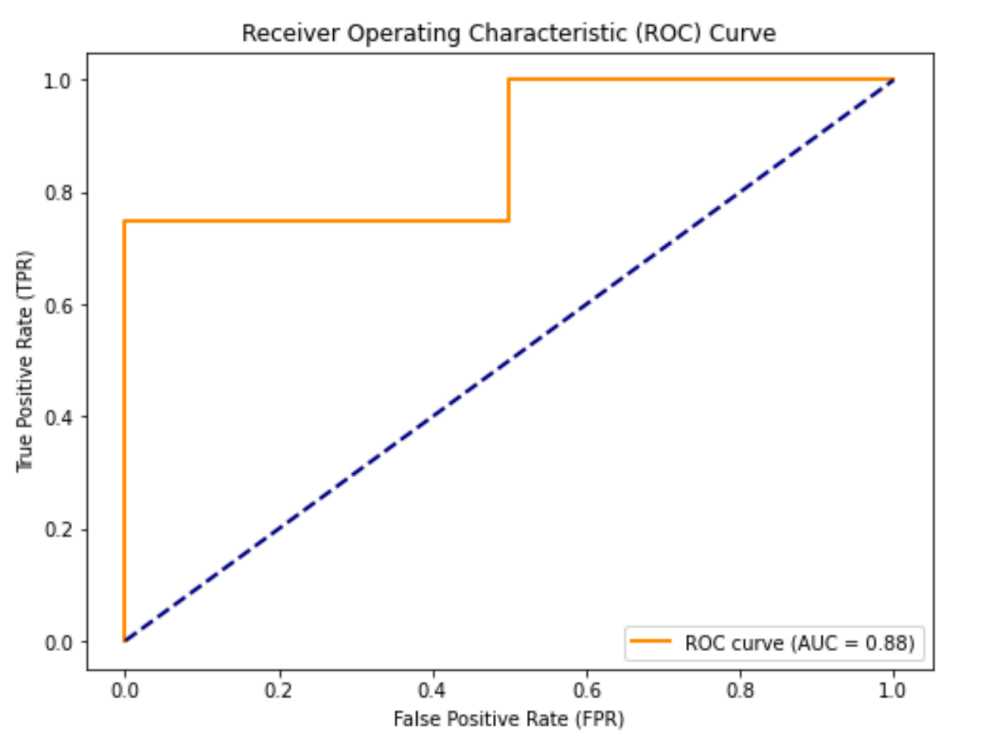
Confusion matrix for the SVC model

Conclusions

- The outcome of the landings can be successfully predicted with the publicly available data
- The best performing prediction model is the Support Vector Classifier
- It can be used to predict landings with an accuracy of 94.4%
- This can help predict the cost of the launches since successfully landing the booster can greatly reduce the cost

Ariane 5 rocket that launched the JWST

Appendix: ROC Curve for SVC



ROC curve for SVC with area under the curve (AUC)

```python
from sklearn.metrics import roc_curve, auc

y_probs = svm_cv.predict_proba(X_test)[:, 1]

# Compute ROC curve and AUC
fpr, tpr, thresholds = roc_curve(Y_test, y_probs)
roc_auc = auc(fpr, tpr)

# Plot ROC curve
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2,
         label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate (FPR)')
plt.ylabel('True Positive Rate (TPR)')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc='lower right')
plt.show()
```

Code used

Thank you!