Applied Data Science Capstone

Nishit Kishor Chaudhari

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone

31 August 2021

Outline

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

Executive Summary

- Data was collected from public SpaceX API and SpaceX Wikipedia page.
- Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features.
- Changed all categorical variables to binary using one hot encoding.
- Standardized data and used GridSearchCV to find best parameters for machine learning models.
- Visualized accuracy score of all models.

- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- All ML models produced similar results with accuracy rate of about 83%. All models over predicted successful landings.
- More data is needed for better model determination and accuracy.

Introduction

Background:

- Commercial Space Age is Here
- Space X (Falcon 9) has best pricing ($62 million vs. upwards $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Challenge:

- Space Y asked us to:
  - determine the price of each launch
  - gather public information about Space X and create dashboards for the team
  - train a machine learning model to predict successful Stage 1 recovery

Methodology

OVERVIEW OFDATA COLLECTION, WRANGLING, VISUALIZATION, DASHBOARD, AND MODEL METHODS

Methodology

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection Overview

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

The next slide will show the sequence of processing the data from SpaceX public API and the one after will show sequence of processing the data from webscraping.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/Spacex%20data%20collection%20api.ipynb

1) Request (Space X APIs)

2) .JSON file + Lists(Launch Site, Booster Version, Payload Data)

3) Json_normalize to DataFrame data from JSON

4) Dictionary relevant data

5) Cast dictionary to a DataFrame

6) Filter data to only include Falcon 9 launches

7) Replace missing PayloadMass values with mean

Data Collection – WebScraping

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/Spacex%20Webscraping.ipynb

1)Request Wikipedia html

2)BeautifulSoup html5lib Parser

3)Find launch info html table

4) Cast dictionary to DataFrame

5) Iterate through table cells to extract data to dictionary

6) Create dictionary

Data Wrangling

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb

Create a training label with landing outcomes where successful = 1 & failure = 0.

Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping:

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

EDA with Data Visualization

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/EDA%20with%20python.ipynb

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

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

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

EDA with SQL

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/EDA%20with%20sql.ipynb

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

Build an interactive map with Folium

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/Launch%20site%20location%20with%20folium.ipynb

Launch Sites Locations Analysis with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

Predictive analysis (Classification)

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Split label column / Class' from dataset

Fit and Transform Features using Standard Scaler

Train, Test & Split Data

GridSearchCV (cv=10) to find optimal parameters

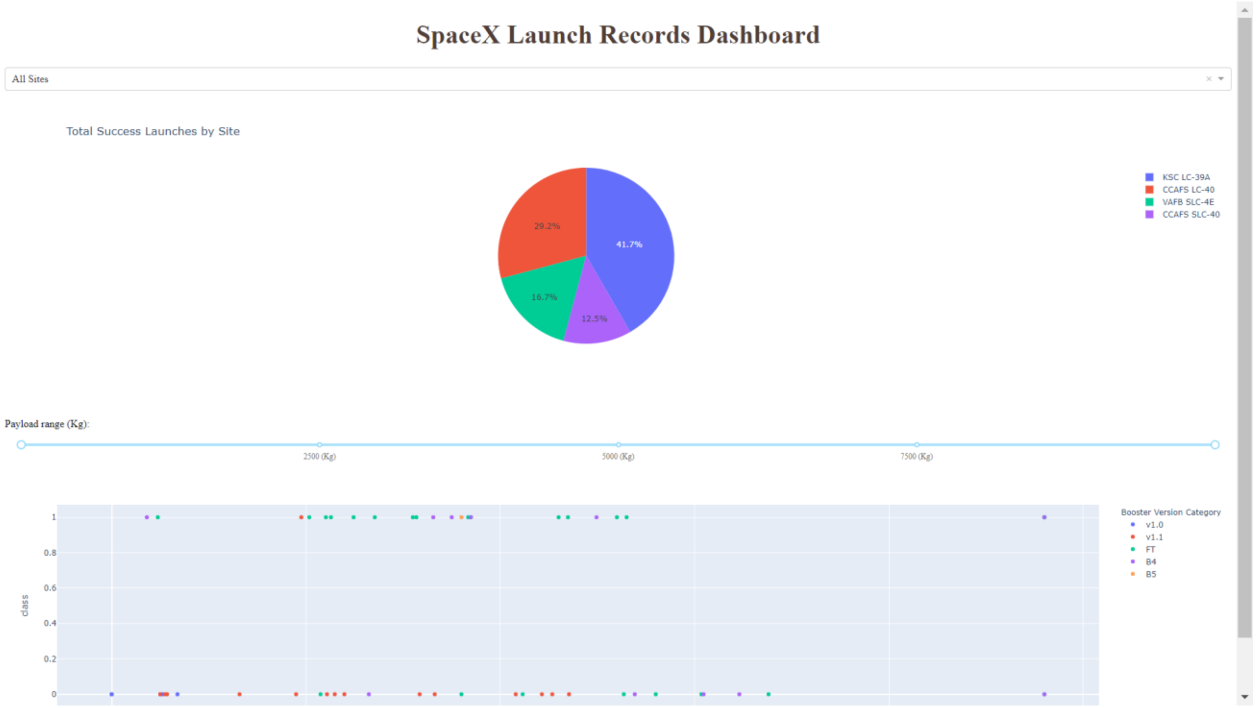
Use GridSearchCV on LogReg, SVM,Decision Tree, and KNN models

Confusion Matrix for all models

Barplot to compare scores of models

Results

Results



SpaceX Launch Records Dashboard

All Sites

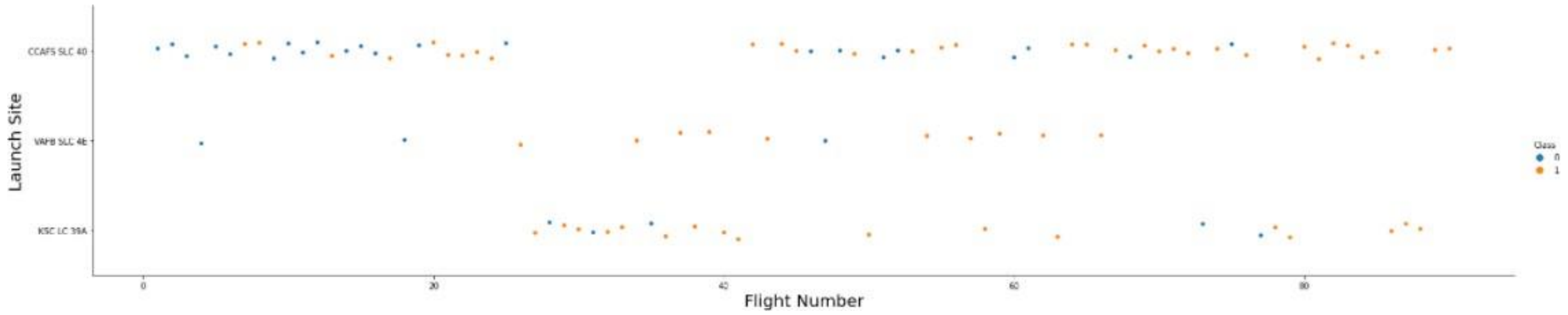
Payload range (Kg):

This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

E.D.A with Visualization

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

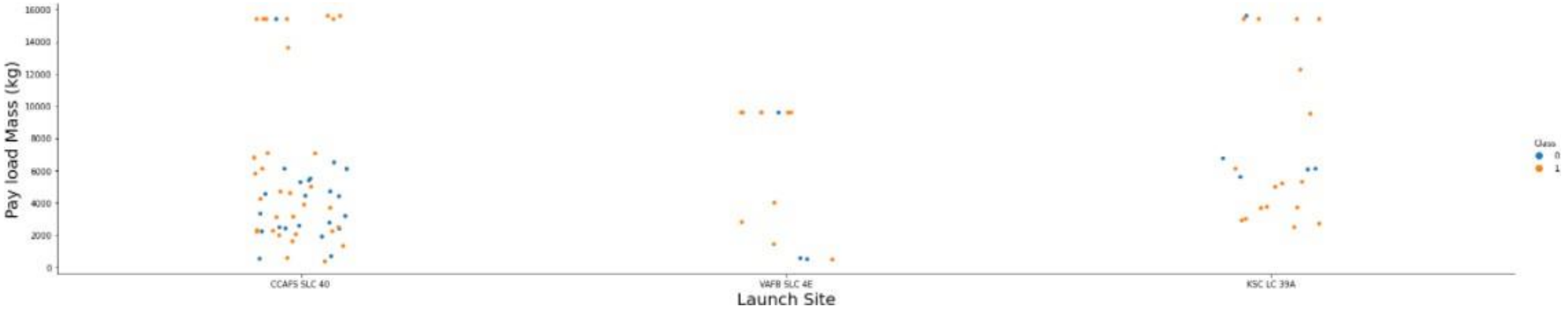
Flight Number vs. LaunchSite



Orange indicates successful launch; Blue indicates unsuccessful launch.

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site

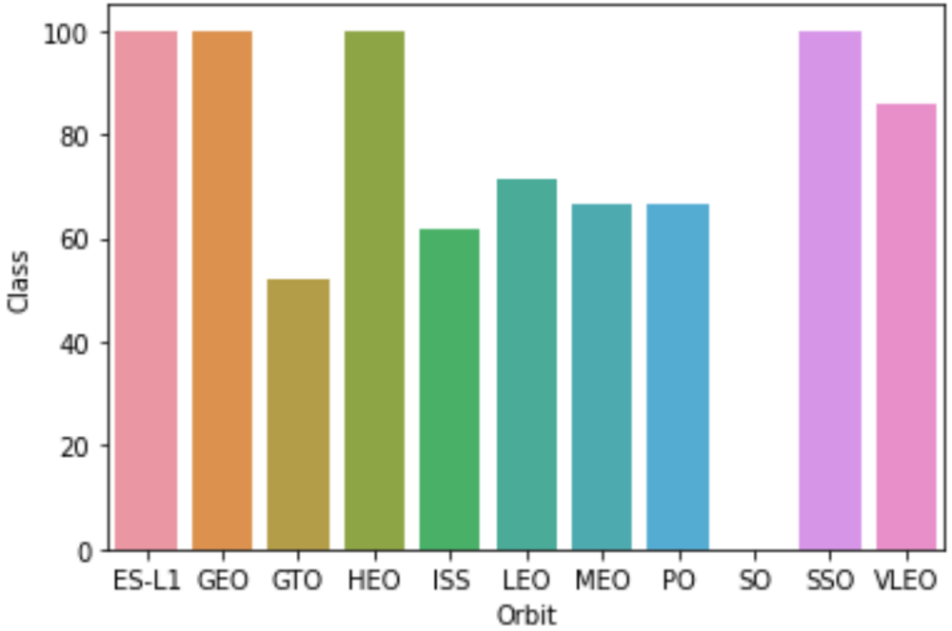


Orange indicates successful launch; Blue indicates unsuccessful launch.

Payload mass appears to fall mostly between 0-7000 kg.

Different launch sites also seem to use different payload mass.

Successrate vs. Orbittype



Success Rate Scale with %

ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)

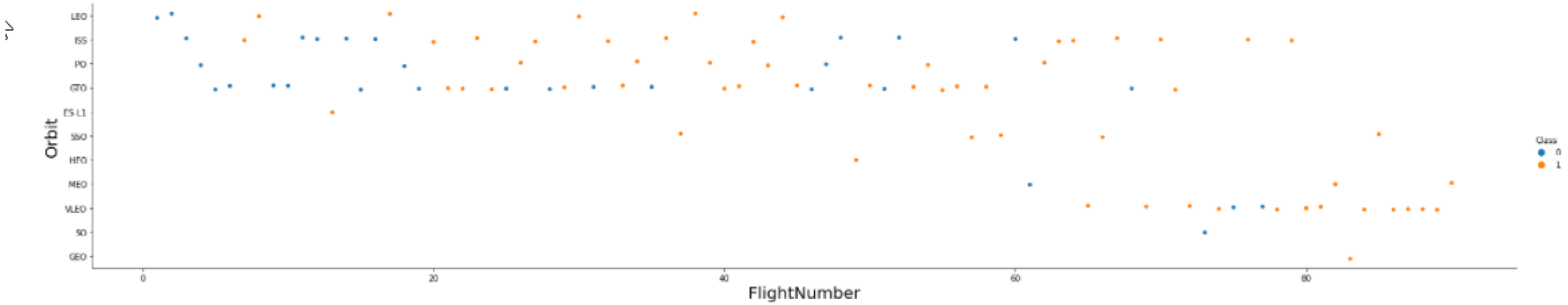
SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbittype



Orange indicates successful launch; Purple indicates unsuccessful launch.

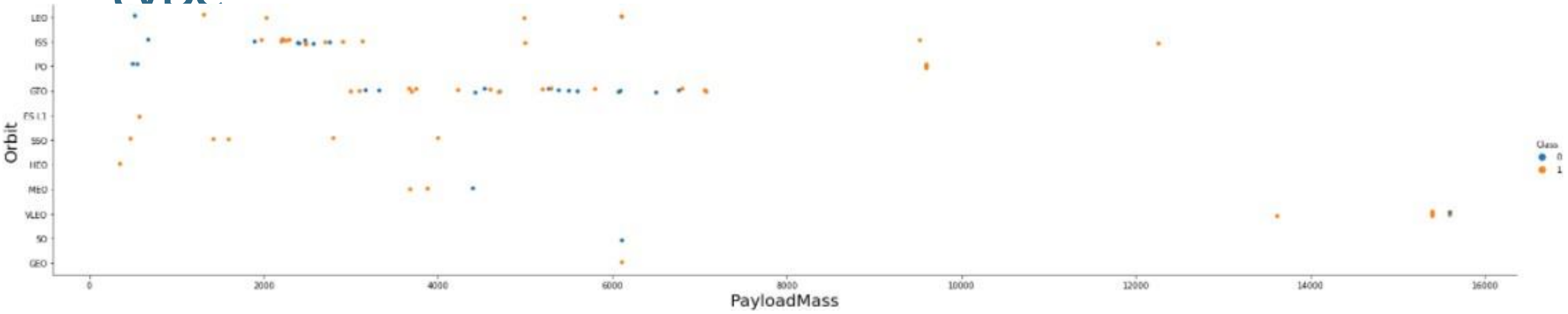
Launch Orbit preferences changed over Flight Number.

Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit type



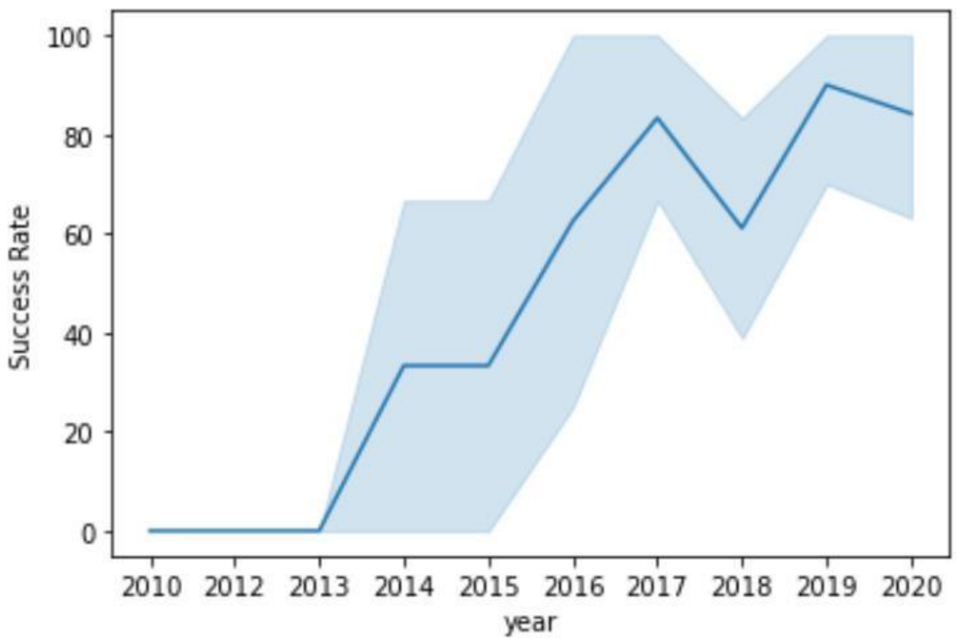
Orange indicates successful launch; Purple indicates unsuccessful launch.

Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

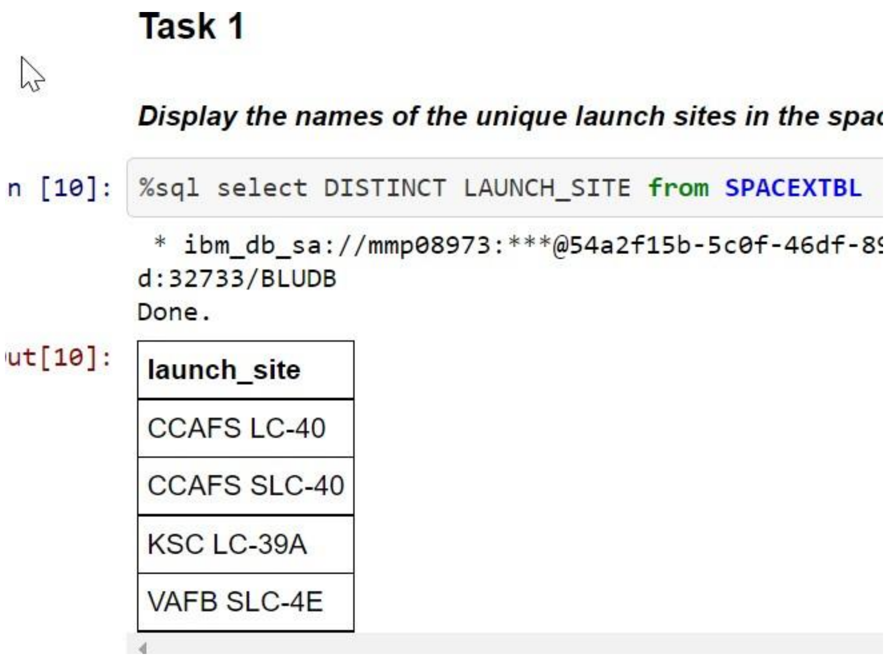
Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

E.D.A. with SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2 INTEGRATED IN PYTHON WITH SQLALCHEMY

All Launch Site Names



Task 1

Display the names of the unique launch sites in the spac

```
n [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-89
d:32733/BLUDB
Done.
```

ut[10]:

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

- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
- launch site with data entry errors.
- CCAFS LC-40 was

  the previous

  name. Likely only

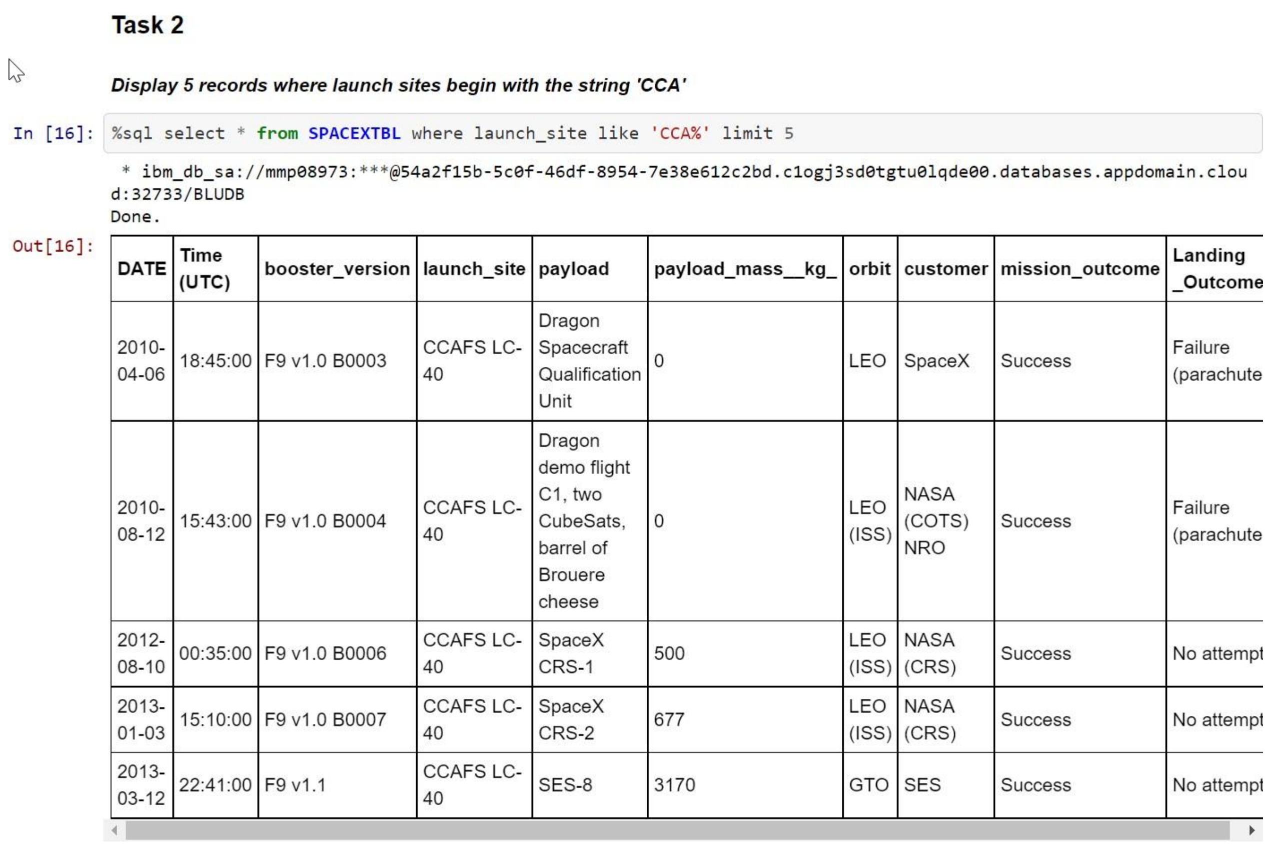
  3 unique

  launch_site

  values: CCAFS

  SLC-40

Launch Site Names Beginning with `CCA`



Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
In [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

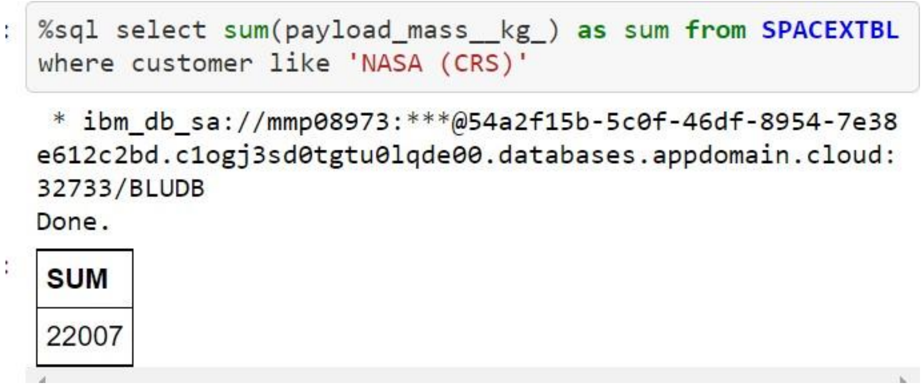
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

Out[16]:

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass from NASA



```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

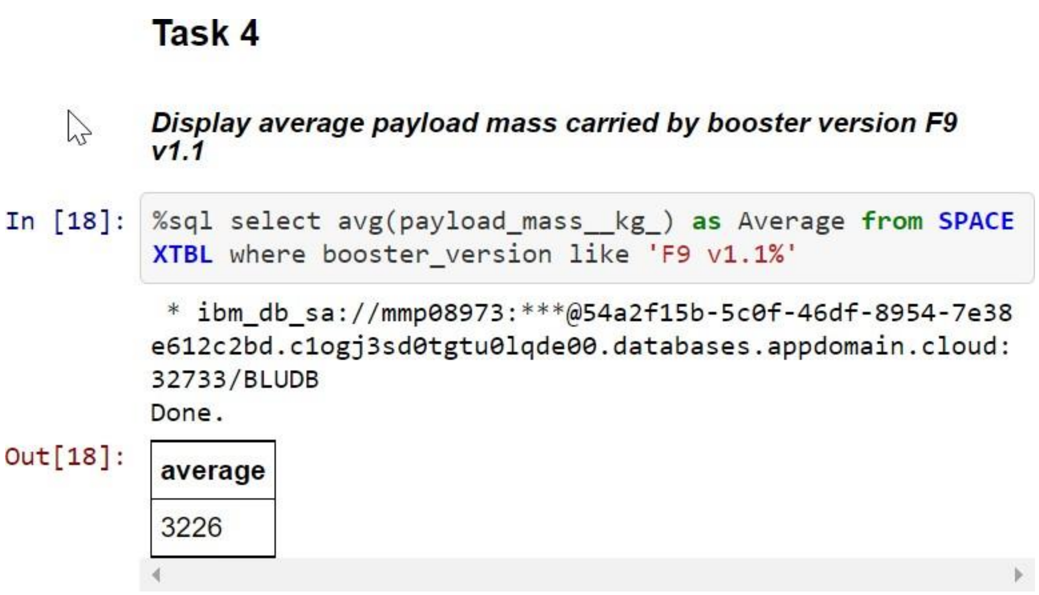
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1



Task 4

Display average payload mass carried by booster version F9 v1.1

In [18]:
```
%sql select avg(payload_mass__kg_) as Average from SPACE
XTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

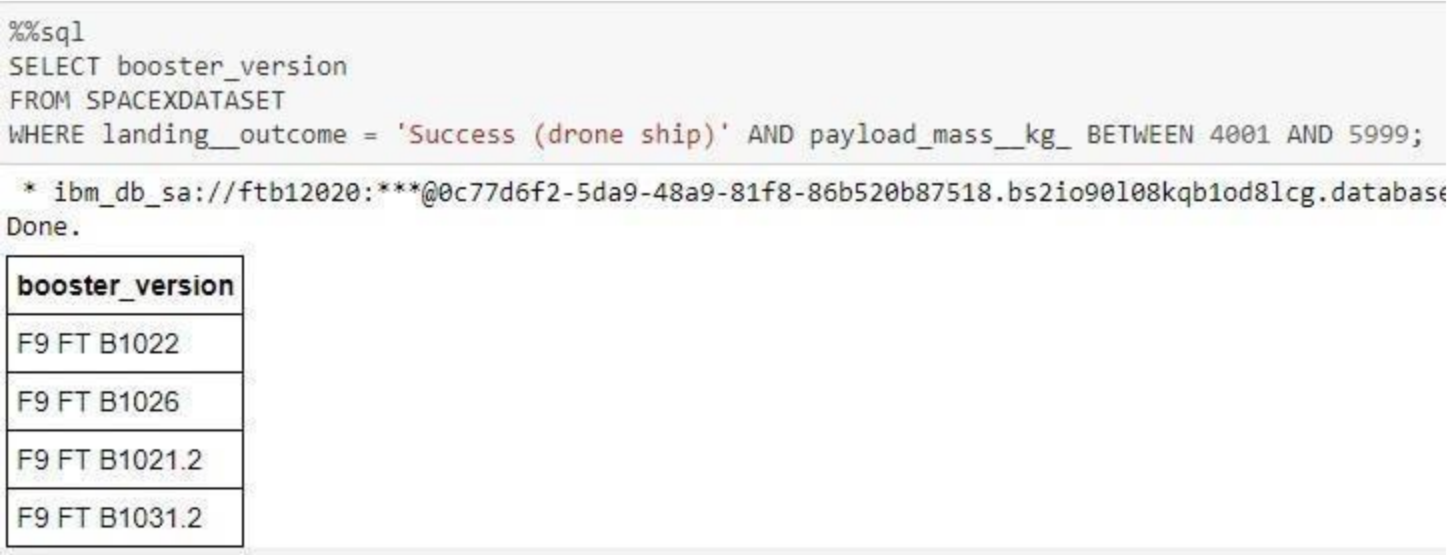
Out[18]:

average
3226

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

Successful Drone Ship Landing with Payload Between 4000 and 6000



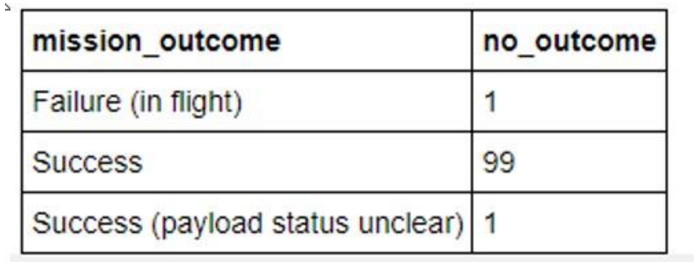
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Each Mission Outcome



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

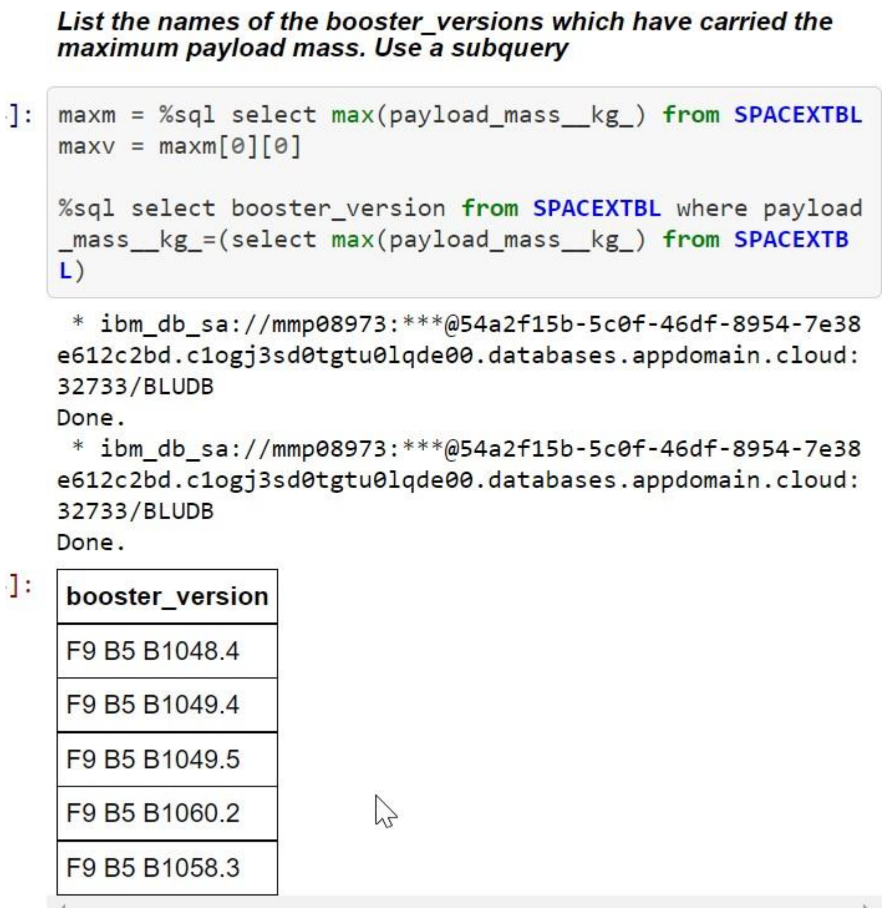
This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters that Carried Maximum Payload



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
]: maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
   maxv = maxm[0][0]

   %sql select booster_version from SPACEXTBL where payload
   _mass__kg_=(select max(payload_mass__kg_) from SPACEXTB
   L)
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

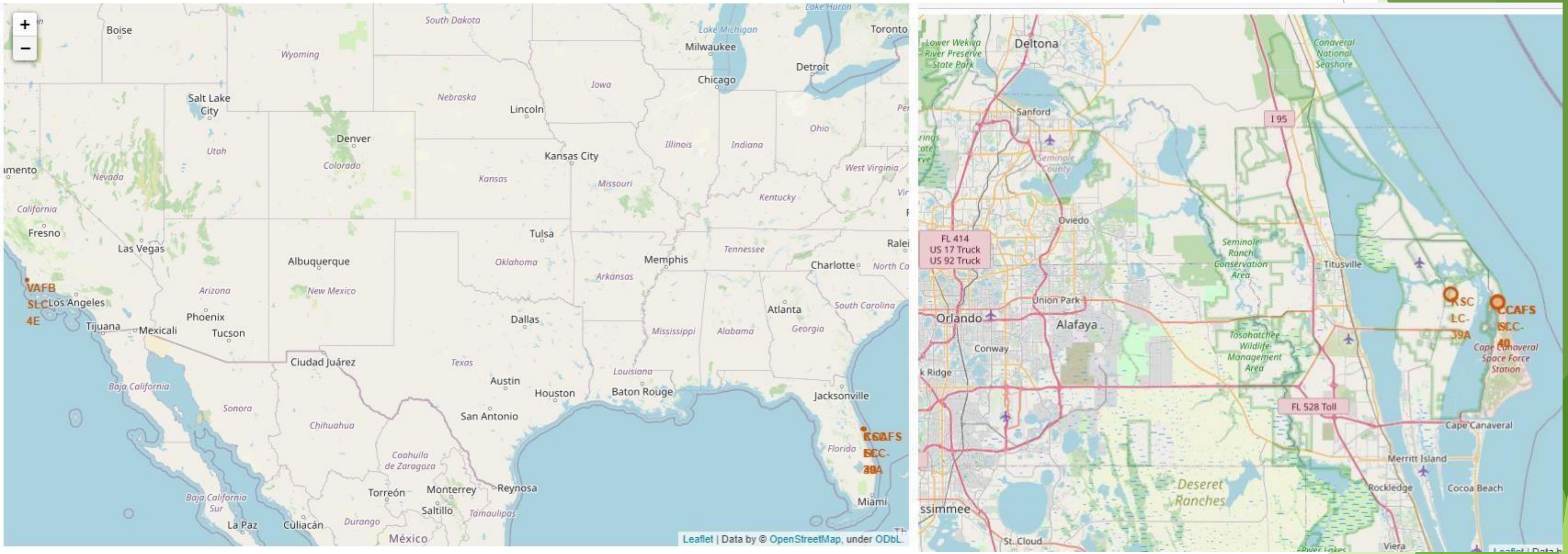
This query returns the booster versions that carried the highest payload mass of 15600 kg.

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

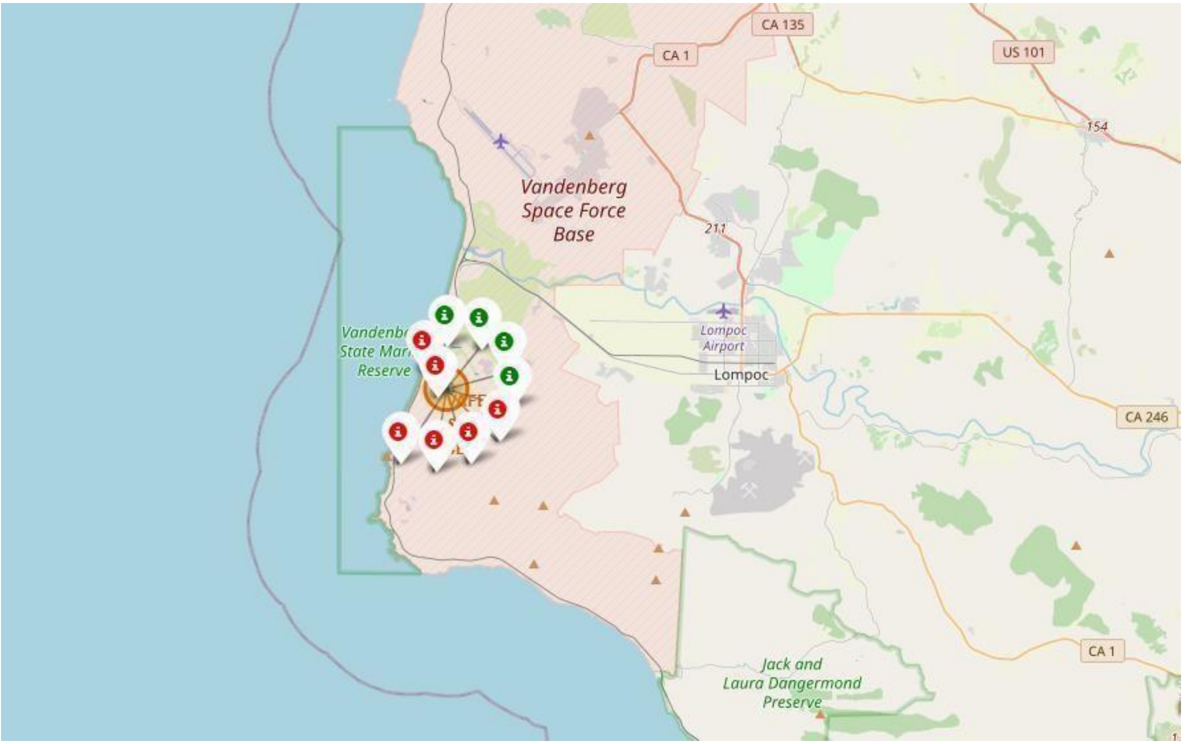
Interactive Map with Folium

Launch Site Locations



The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

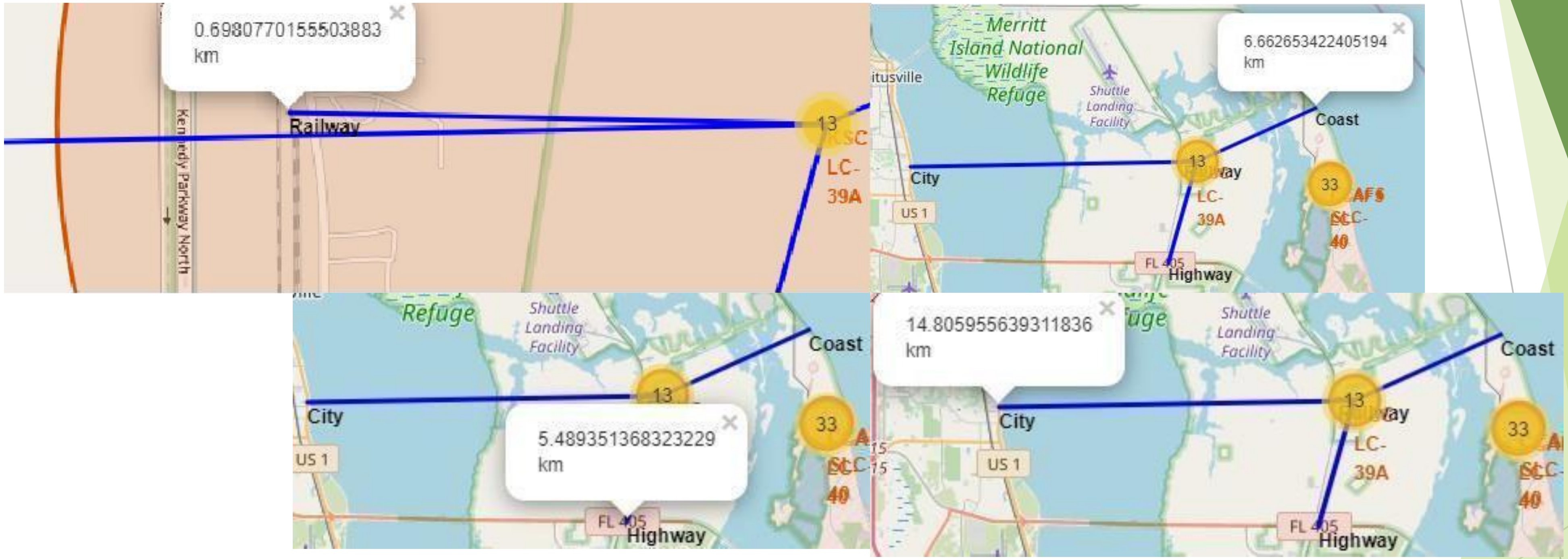
Color-Coded Launch Markers



CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
CA 246
CA 1
Jack and Laura Dangermond Preserve

Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

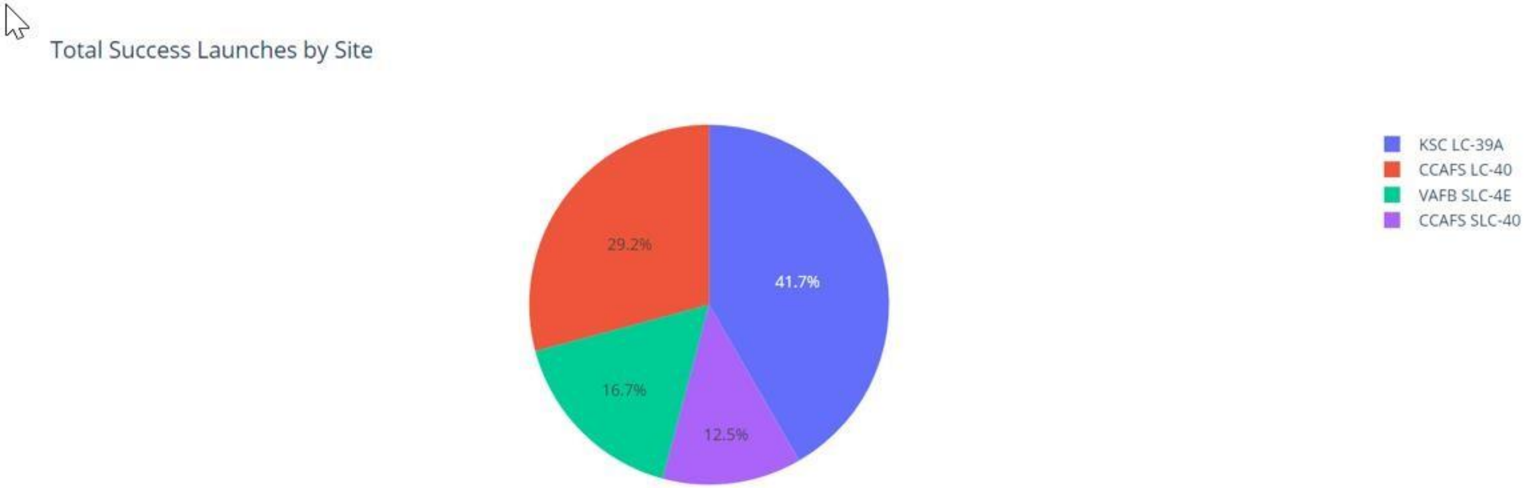
Key Location Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

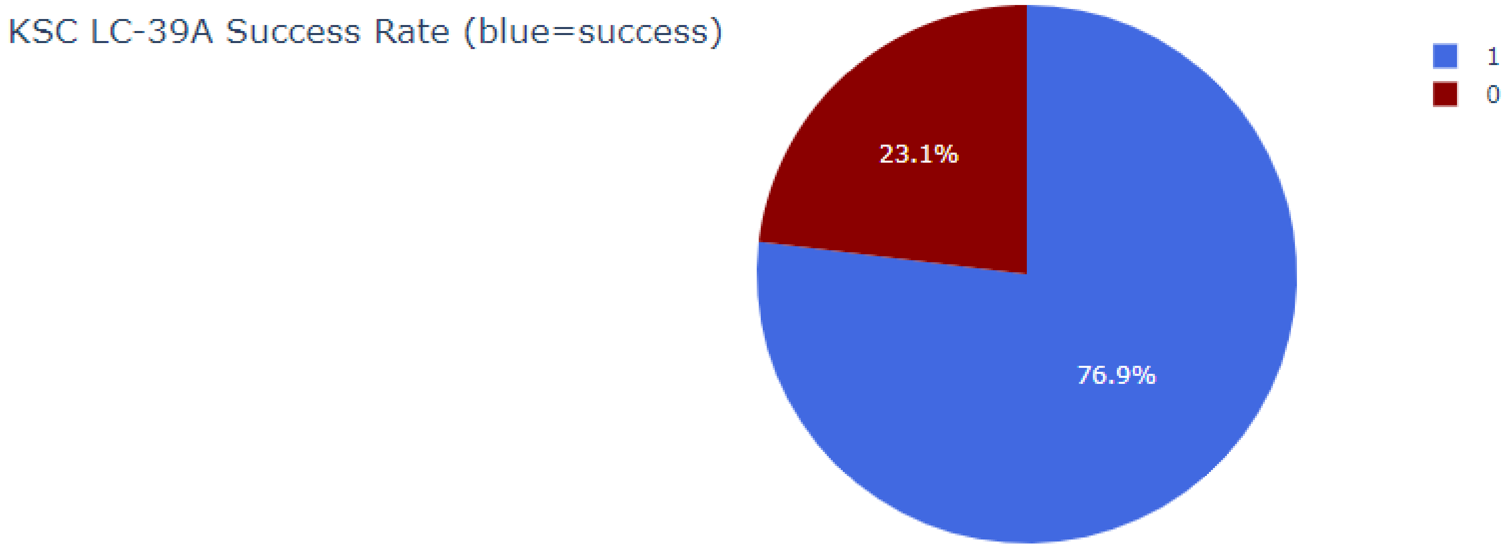
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



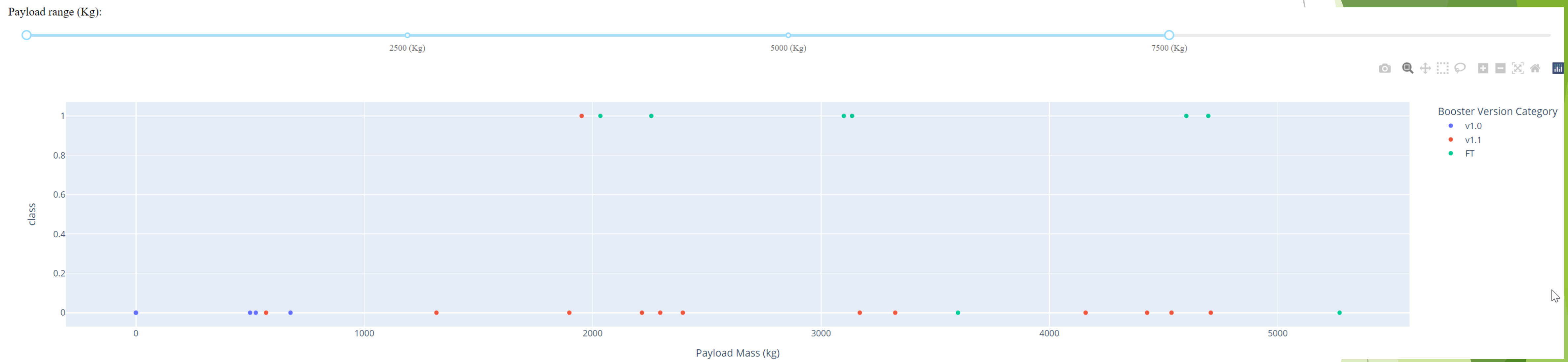
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

Highest SuccessRate Launch Site


KSC LC-39A Success Rate (blue=success)

KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Payload Mass vs. Success vs. Booster Version Category

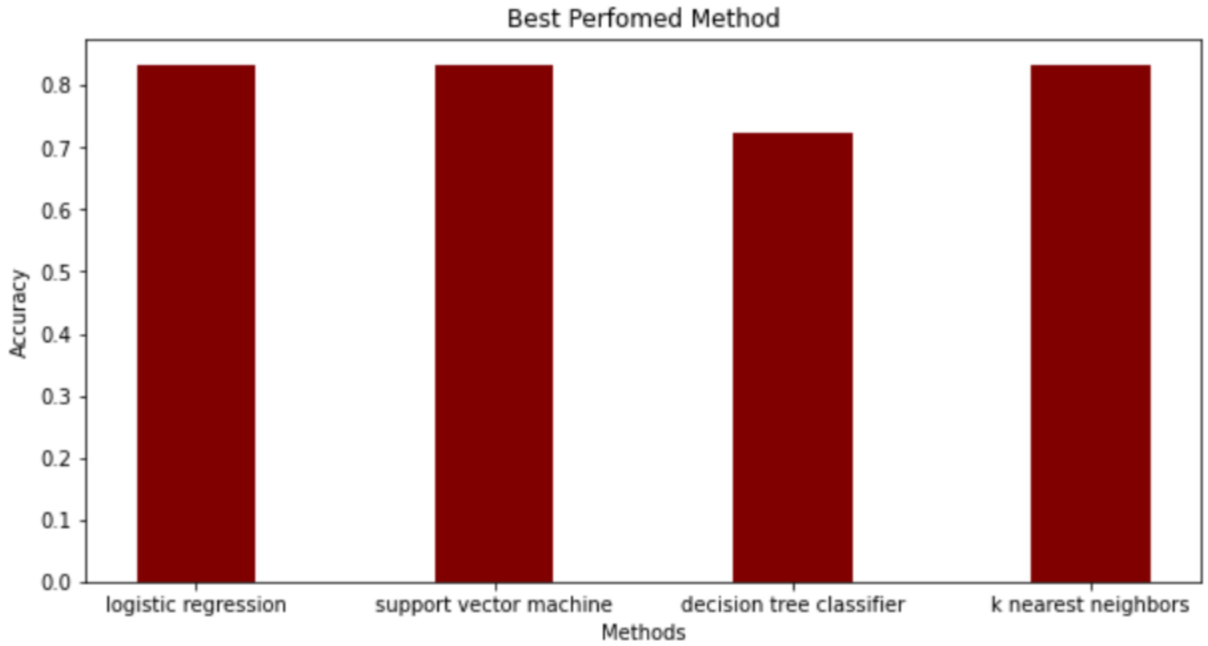


Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-7500, interestingly there are two failed landings with payloads of zero kg.

Predictive Analysis (Classification)

GRIDSEARCHCV(CV=10) ON LOGISTIC REGRESSION, SVM, DECISION TREE, AND KNN

Classification Accuracy



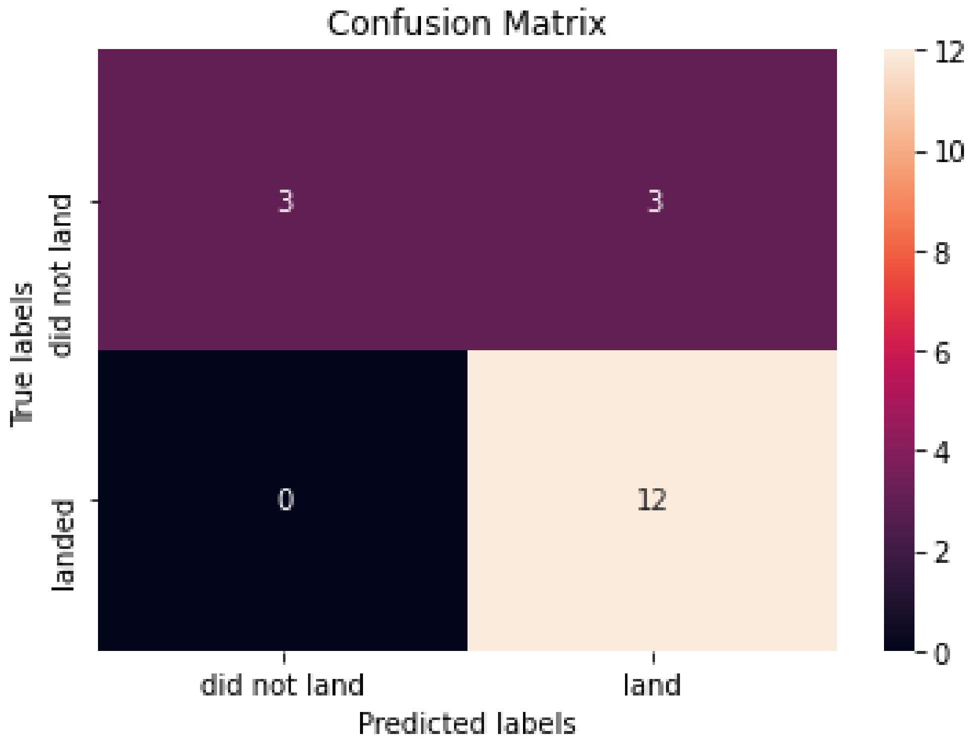
The models had virtually the same accuracy on the test set at 83.33%accuracy, except the decision tree classifier with 72,23 %.

It should be noted that test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

We likely need more data to determine the best model.

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Since all models performed the same for the test set, the confusion matrix is the same across all models.

The models predicted 12 successful landings when the true label was successful landing.

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.

The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).

Our models over predict successful landings.

CONCLUSION

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- More data should be collected to better determine the best machine learning model and improve the accuracy

APPENDIX

GitHub repository url:

https://github.com/NishitChaudhari/Applied-Data-Science-Capstone

Special Thanks to All Instructors:
https://www.coursera.org/professional-certificates/ibm-data-science

SpaceX data

Wikipedia