IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fauzan Mohammed

3/20/2022

Outline

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

Executive Summary

Methodology

- Collect SpaceX launch data via SpaceX API and web-scraping Falcon 9 wiki
- Perform EDA using Pandas, Folium, and SQL
- Train predictive models on processed data

Results

- SpaceX facilities located in Florida (primary) and California
- First-stage landings higher success rate and payload mass over time
- Recent launches are primarily into very-low-earth orbit (VLEO)

Introduction

- SpaceY would like to compete with SpaceX, and needs to determine the cost of a SpaceX launch.
- When does a SpaceX first stage land successfully? What factors help predict the outcome of a landing?

Section 1

Methodology

Methodology

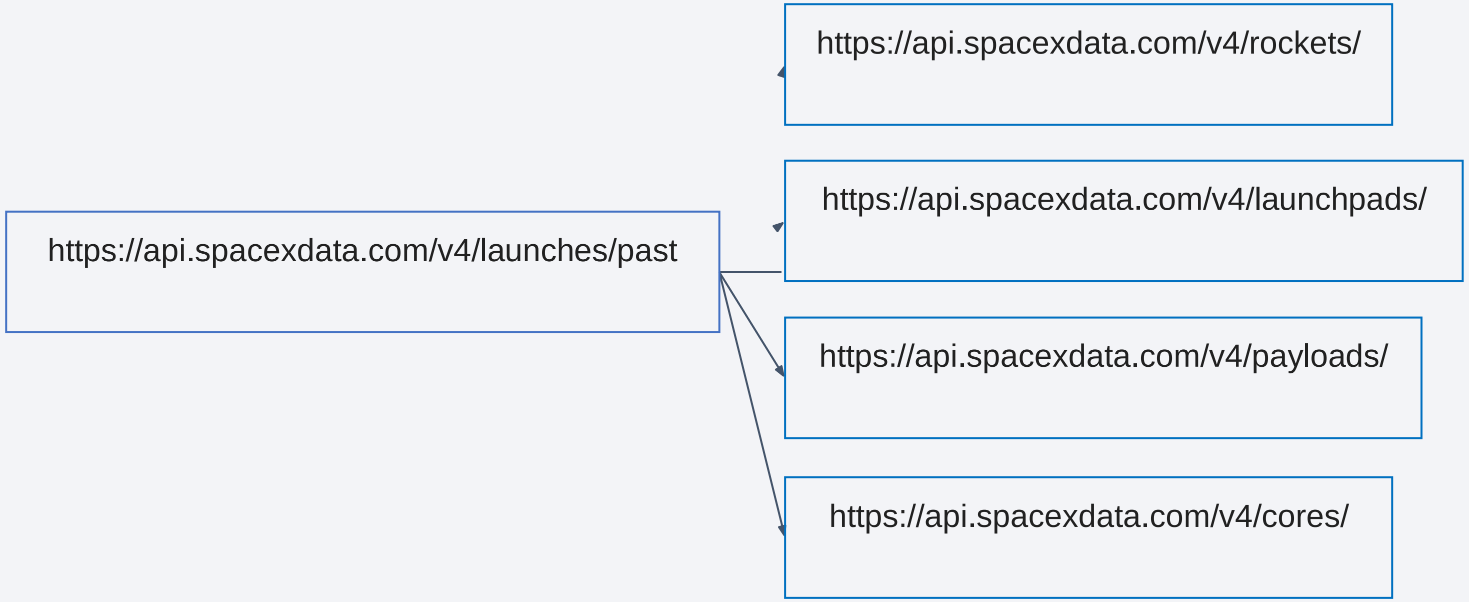
Executive Summary

- Data collection methodology:
  - Collect data with a combination of calls to SpaceX API and web-scraping
- Perform data wrangling
  - Perform EDA with Pandas and create landing outcome label column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data, create train-test split, use grid search to find optimal hyperparameters
  - Experiment with logistic regression, regression trees, and support vector machines

Data Collection

- Collect data with a combination of calls to SpaceX API and web-scraping

Data Collection – SpaceX API



1. Request and parse SpaceX launch data using GET request
2. Filter data to include only Falcon 9 launches
3. Impute missing values with mean

Data Collection - Scraping

Request Falcon9 launch wiki page from its url

↓

Extract column/variable names from html table header

↓

Create data frame by parsing the launch HTML tables

Data Wrangling

Calculate the number of launches on each site

↓

Calculate the number and occurrence of each orbit

↓

Create a landing outcome label from outcome column

EDA with Data Visualization

- Scatter plot of **Flight Number vs Launch Site**, **Payload vs Launch Site**, and **Flight Number vs Orbit**, **Payload vs Orbit**
  - Scatter plots illustrate the relationship between two variables (with an additional third variable via point colors)
- Bar chart of **success rate of each orbit**
  - Bar charts illustrate values across several categories
- Line chart of **success rate vs year**
  - Line chart can illustrate trends over time

EDA with SQL

- Unique launch sites in the space mission
- 5 records where launch sites begin with the string ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of boosters which successfully landed payloads between 4000 and 6000 kg in drone ship
- Total number of successful and failed landing missions
- Names of booster versions which carried the maximum payload mass
- Booster versions and launch sites of failed drone ship landings in 2015
- Counts of landing outcomes between 2010-06-04 and 2017-03-20 in desc order

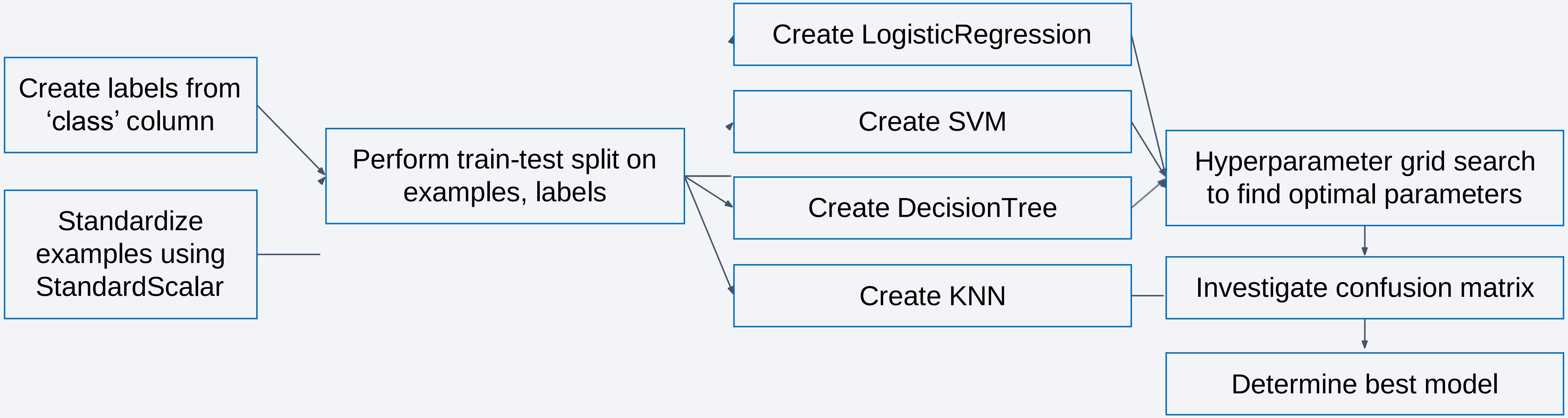
Build an Interactive Map with Folium

- Markers + Circles with Popup that identify launch site by name
  - Shows where launch sites are located
- MarkerCluster of Markers color coordinated by launch status (success/failure)
  - Shows where launches are concentrated
- PolyLine between launch site to selected coastline point
  - Shows distance between launch site and nearest coastline

Build a Dashboard with Plotly Dash

- Launch Site Dropdown input component
  - Allows filtering landings by launch site
- Payload RangeSlider input component
  - Allows filtering landings by payload mass
- Landing success rate pie chart
  - Shows the percentage of successful landings by launch site
- Landing outcome vs payload mass scatter plot
  - Shows relationship between payload mass and landing outcome

Predictive Analysis (Classification)



Results

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

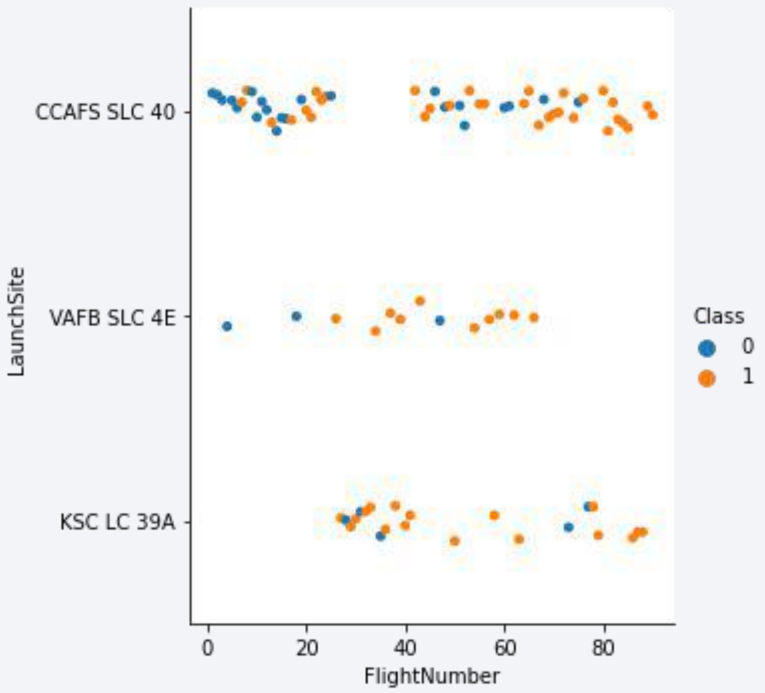
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

There seems to be a positive correlation between flight number and class. Many early launches (and failures) occurred at CCAFS SLC 40, whereas later launches (primarily successes) occurred at KSC LC 39A.

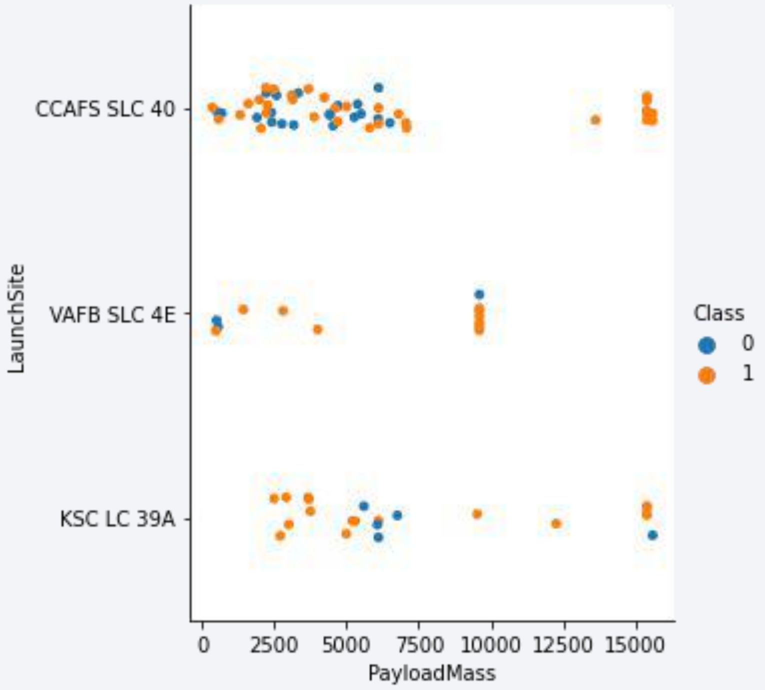
1 = successful launch, 0 = failed launch



Payload vs. Launch Site

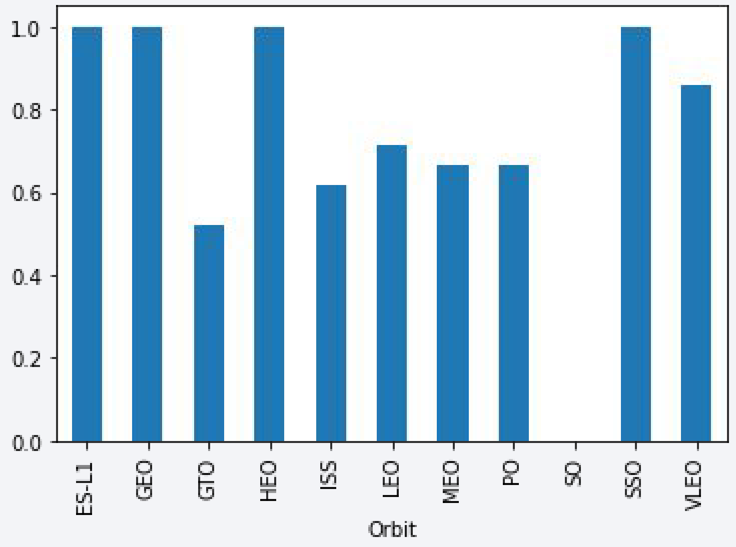
CCAFS SLC 40 and KSC LC 39A have launched the maximum payload mass of around 15000 kg. VAFB SLC 4E seems to support a max payload of 10000 kg. Launches of rockets with the heaviest payloads seem to tend to succeed.

1 = successful launch, 0 = failed launch



Success Rate vs. Orbit Type

ES-L1, GEO, HEO, and SSO launches have a 100% success rate. SO launches have a 0% success rate. All other launches by orbit have between a 50% and 90% success rate.



LEO: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].

VLEO: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].

GTO A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .

SSO (or SO): It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .

ES-L1 :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .

HEO A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].

ISS A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]

MEO Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]

HEO Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]

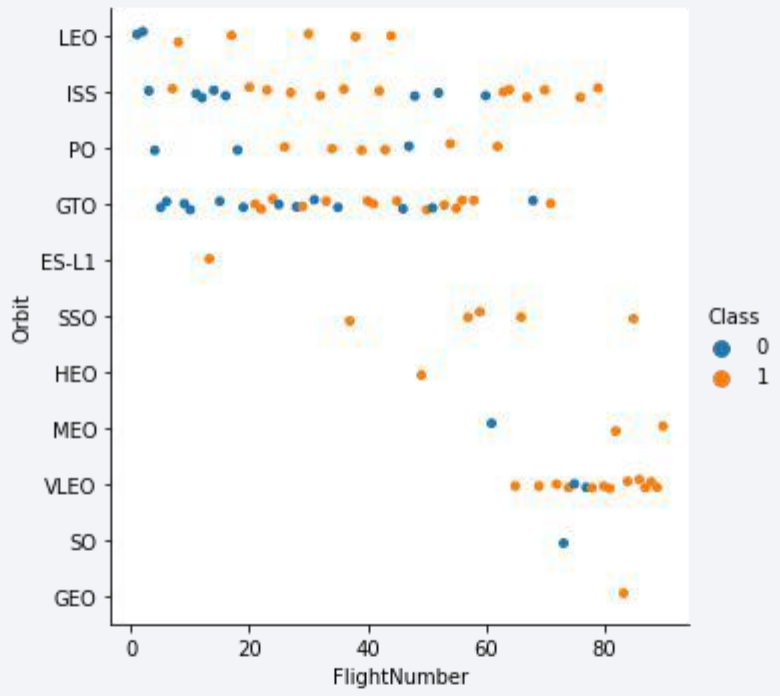
GEO It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]

PO It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11]

Flight Number vs. Orbit Type

More recent launches are concentrated in VLEO and ISS. LEO launches seem to have been discontinued in favor of otherorbits. GTO launches seem to have the highest number of failures. Only one each of SO, GEO, HEO, ES-L1 launch were ever attempted, explaining the 0%/100% success rates.

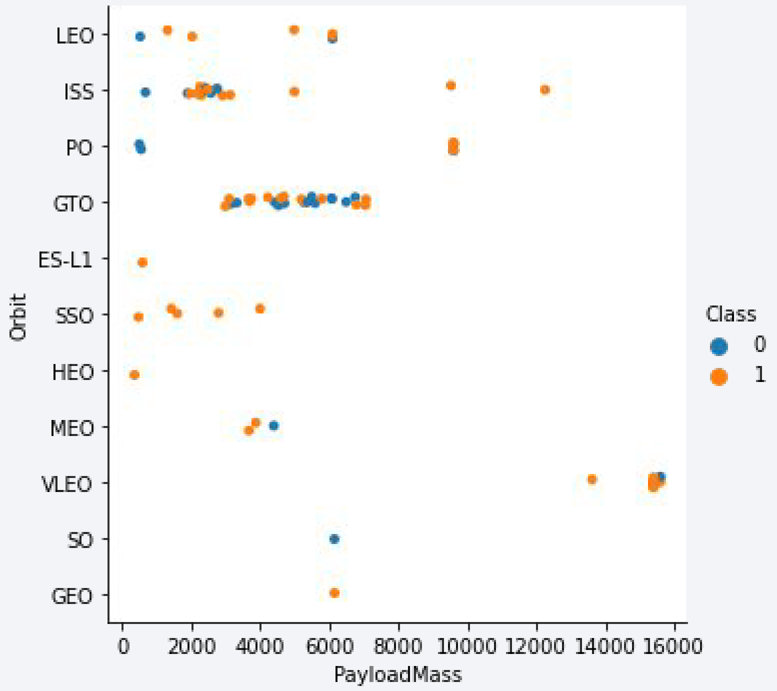
1 = successful launch, 0 = failed launch



Payload vs. Orbit Type

VLEO launches tend to have the highest payload mass of around 16000 kg while ES-L1, SSO, and HEO launches havethe lowest payload.

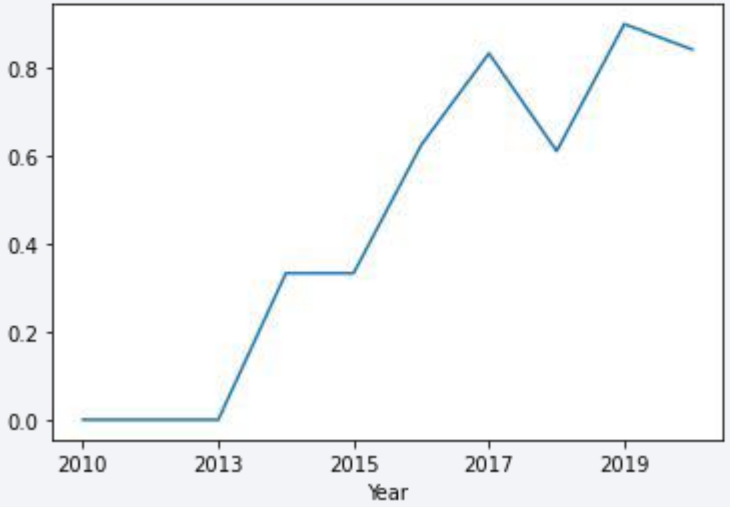
1 = successful launch, 0 = failed launch



Launch Success Yearly Trend

The launch success rate has generally been increasing year-over-year.

y-axis is launch success rate



All Launch Site Names

- **Query**: `select distinct(launch_site) from spacextbl;`
    - `distinct` is SQL function that selects unique values in provided column
    - `launch_site` is the column of interest
    - `spacextbl` is the table of interest
- **Results**:

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

Launch Site Names Begin with 'CCA'

- **Query**: `select * from spacextbl where substring(launch_site, 1, 3) = 'CCA' limit 5;`
  - `substring` function called with above arguments selects first three characters of `launch_site`
  - `limit` keyword limits results to 5 rows
- **Results**:

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS__KG_	ORBIT	CUSTOMER	MISSION_OUTCOME
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- **Query**: `select sum(payload_mass__kg_) from spacextbl where customer='NASA (CRS)';`
  - `sum` function sums values of column
- **Results**:

1
45596

SpaceX launched a total of 45596 kg of payload for `NASA (CRS)`. If we were interested in all NASA customers, then we would need to use

`where substring(customer, 1, 4)=’NASA’.`

Average Payload Mass by F9 v1.1

- **Query**: `select avg(payload_mass__kg_) from spacextbl where booster_version='F9 v1.1';`
- **Results**:
    - The average payload mass of rockets with F9 v1.1 booster is 2928 kg.

1
2928

First Successful Ground Landing Date

- **Query**: `select min(date) from spacextbl where landing__outcome='Success (ground pad)';`
- **Results**:

  The first successful landing on a ground pad occurred on December 22, 2015.

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Query**: `select distinct(booster_version) from spacextbl where landing__outcome='Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000;`

- **Result**:

  The four F9 FT boosters below successfully landed payloads of between 4000 and 6000 kg via drone ship.

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

Total Number of Successful and Failure Mission Outcomes

- **Queries**:
  - `select count(*) from spacextbl where substring(landing__outcome, 1, 7)='Failure';`
  - `select count(*) from spacextbl where substring(landing__outcome, 1, 7)='Success';`
- **Results**:

  Overall, there were **10** failed landings and **61** successful landings.

1
10

1
61

Boosters Carried Maximum Payload

- **Query**: `select distinct(booster_version) from spacextbl where payload_mass__kg_=(select max(payload_mass__kg_) from spacextbl);`
- **Results**:

The following five F9 B5 boosters have carried the maximum payload.

BOOSTER_VERSION
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7

2015 Launch Records

- **Query**: `select booster_version, launch_site from spacextbl where landing__outcome='Failure (drone ship)' and year(date) = '2015';`
- **Results**:

The only failed drone ship landings in 2015 occurred at `CCAFS LC-40` with the following two F9 version 1.1 boosters.

BOOSTER_VERSION	LAUNCH_SITE
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- **Query**: `select landing__outcome, count(landing__outcome) as loc from spacextbl where date >= '2010-06-04' and date <= '2017-03-20' group by landing__outcome order by loc desc;`
- **Results**:

Between June 4, 2010 and March 20, 2017, the majority of landings were either failed or not attempted.

LANDING__OUTCOME	LOC
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3

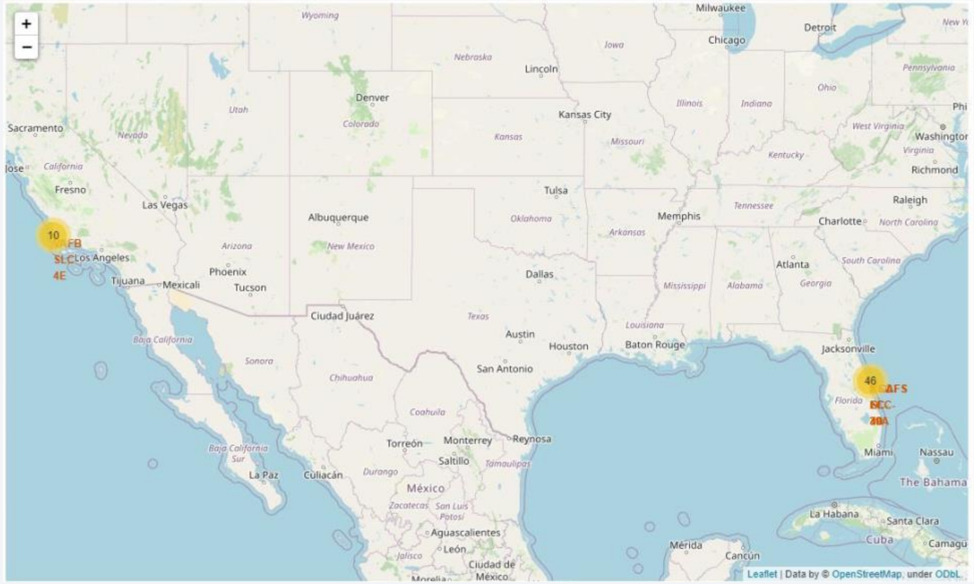
Section 3

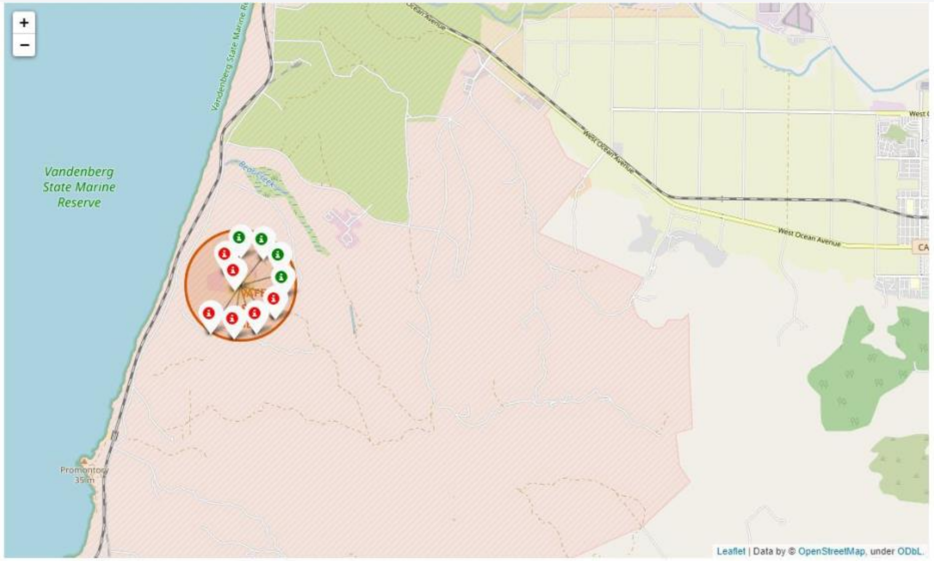
Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites

SpaceX launch sites are on either the West or East coast of the continental United States.

Map of SpaceX Launch Outcomes





The Florida launch sites have supported the most launches (46) compared to the California site (10)

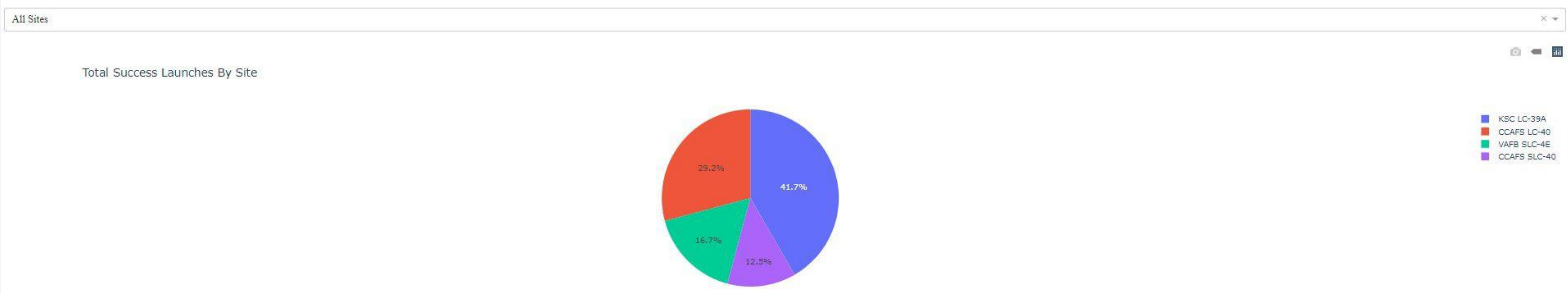
Distance between CCAFS SLC-40 and Coast

Launch sites are generally very close to the coast. For instance, CCAFS SLC-40 is not even 1 km from the Atlantic coast.

Section 4

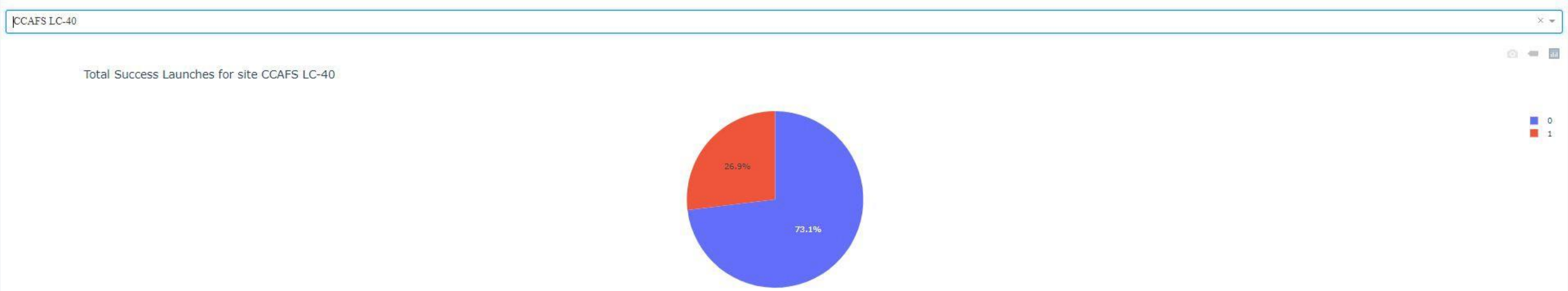
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



- KSC LC-39A responsible for the most successful launches, followed by CCAFS LC-40, then VAFB SLC-4E, then CCAFS SLC-40.

Most successful launch site pie chart



- CCAFS LC-40 is the most successful launch site (73.1% success rate)

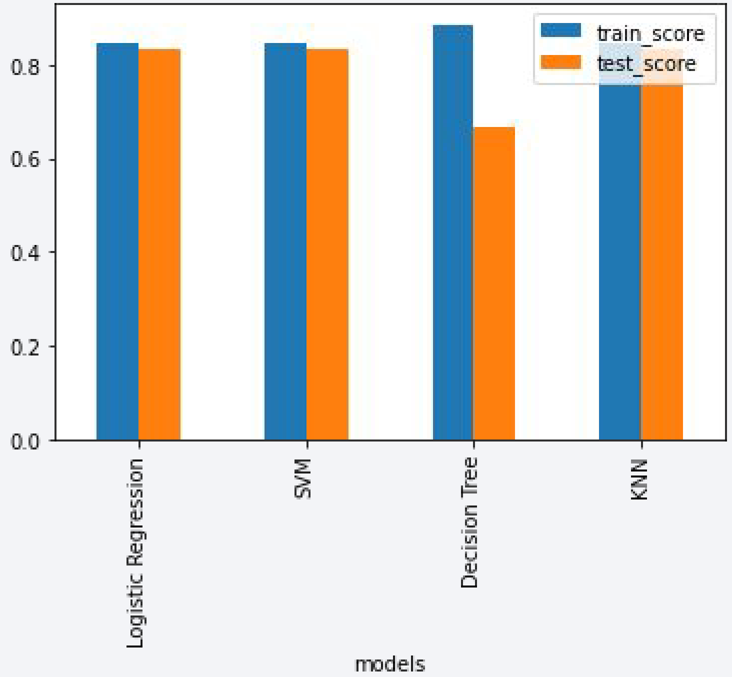
Payload vs. Launch Outcome Scatter Plot

- v1.0 and v1.1 boosters have the lowest success rates
- B4 and FT boosters have the highest success rates

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- All of the models except the Decision Tree have the same test score (~0.83).

Confusion Matrix

- Major-diagonal cells are correctly classified. The only classification error is **three false positives**, where the model incorrectly predicted that the model would land.

Conclusions

- Newer SpaceX boosters (e.g. FT, B4, B5) greatly improved launch success rate
- Linear models trained on API/webscraped data not sufficient to completely predict launch outcome
- Newer SpaceX boosters both carry more payload and have higher landing success rate

Appendix

- Sample model training code:

```
parameters  {'kernel':('linear', 'rbf','poly','rbf' 'sigmoid')
             'C' : np.logspace (-3, 3, 5),
             'gamma':np.logspace (-3, 3, 5)}
, vm   SVC()

svm_cv = GridSearchCV(svm, parameters) cv=10 )

svm_cv.fit(X_train, Y_train)

GridSearchCV(cv=10, estimator=SVC (),
             param_grid={'C': array({1.0-0000000 e-03, 3.16227766e-02 1.00000000e+00, 3.16227766e+01,
       1.00000000e w 3]),
                         gamma  array([1.00000000 e-03 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
       1.00000000ew3]),
                         'kernel': ('linear', 'rbf', 'poly', 'rbf', 'sigmoid')} )

print("'tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)

tuned hpyerparameters :(best parame, s) {"C' : 1. 0, gamma 0. 0316227766e168379, · kernel · : 'sigmoid'}
accuracy : 0.8482142857142856
```

- Sample model test code:

```
svm_cv.scare(X_test, Y_test)

0 .8333333333333334
```


Thank you!