IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carryll Hsieh
February 28, 2022

Outline

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

Executive Summary

- Goal: predict whether the Falcon 9 first stage will land successfully.
- Collected data from SpaceX API and web scraping
- Organized and cleaned data
- Results: SpaceX's successful launches have the following characteristics:
  - Launched from KSC LC-39A launch site
  - Launched in 2017 or later
  - Light payload (2000-4000kg)
  - Recovered through drone ship
- Our machine learning model predict the outcome of a recovery with around 83% accuracy

Introduction

- While the launches cost other providers upward of $165 million, it only costs SpaceX's Falcon 9 rocket launch $62 million, mainly because SpaceX can *reuse the first stage*.
- Thus, we want to determine whether the first state will land successfully and be reused.
- Specifically, we want to find out
  - Factors that affect the landing outcome
  - Impact on outcome from each factor
  - Best conditions for the SpaceX launches to succeed

Section 1

Methodology

Methodology

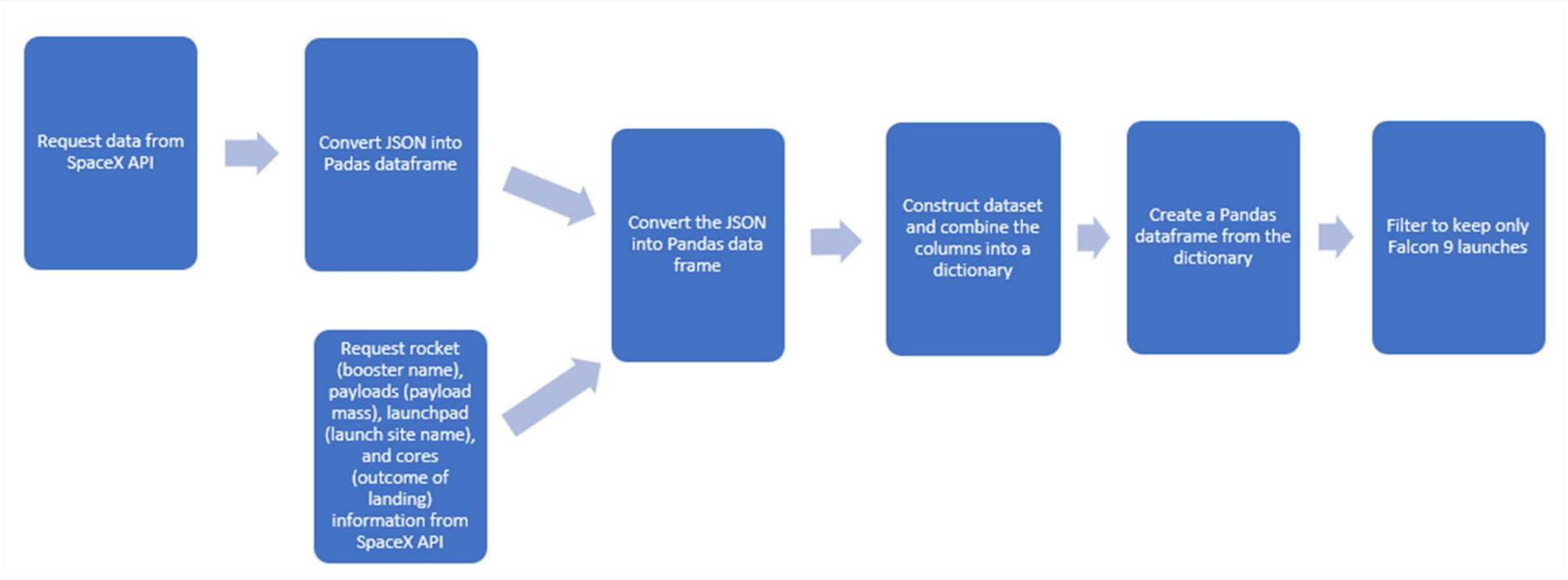
Executive Summary

- Data collection methodology:
  - Request /collected rocket launch data from SpaceX API and
  - Web scraping SpaceX's Wikipedia page using Python BeautifulSoup
- Perform data wrangling
  - Cleaned and organized data / combined columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split dataset into training and testing datasets
  - Used classification models: KNN, Decision Tree, Logistic Regression, and Support Vector Machine
  - Evaluate accuracy scores of the various models and determine the best parameters

Data Collection

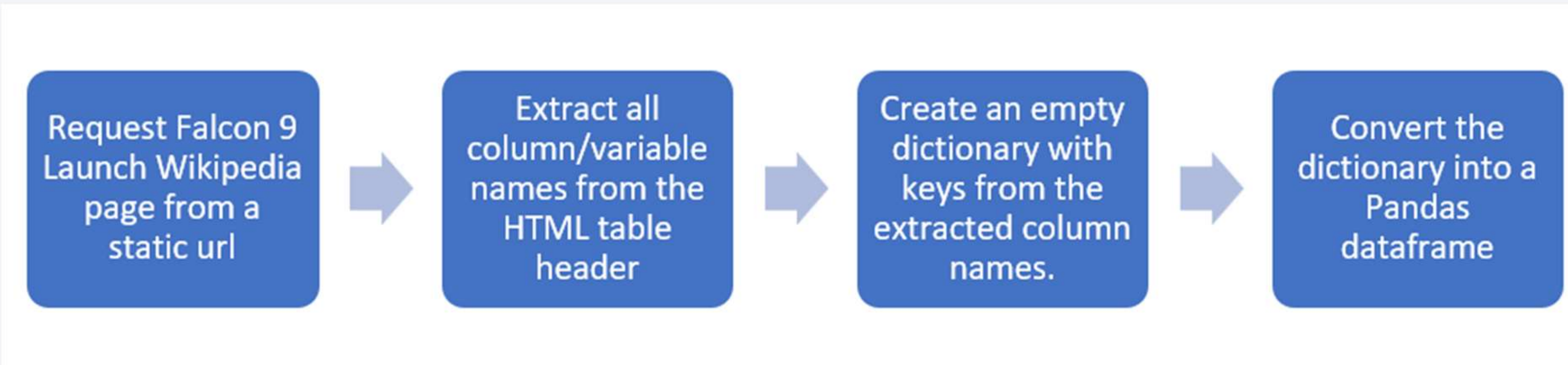
- Data sets were collected using two methods:
  - Request from SpaceX API GitHub Data Collection API
  - Web scraping SpaceX Wikipedia page GitHub Data Collection Web Scraping

Data Collection – SpaceX API



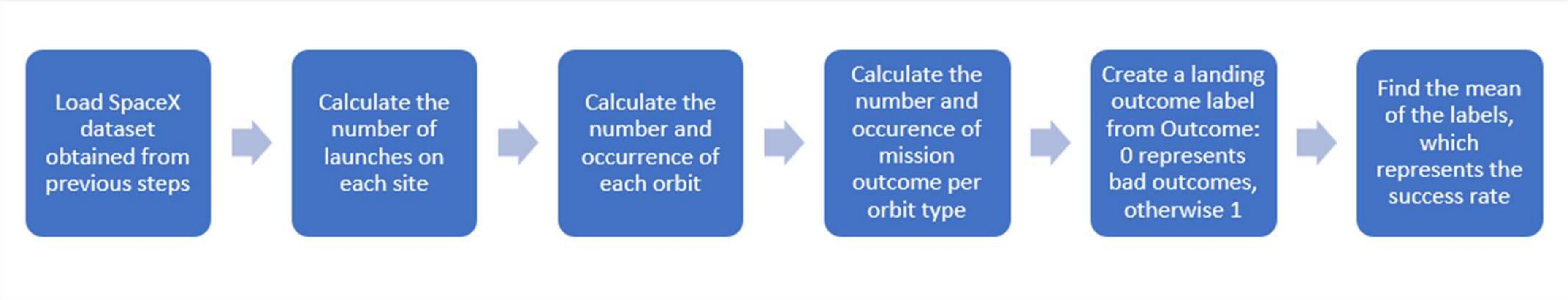
[GitHub Link for Data Collection - API]

Data Collection - Scraping



- Use BeautifulSoup
- GitHub: Data Collection Web Scraping

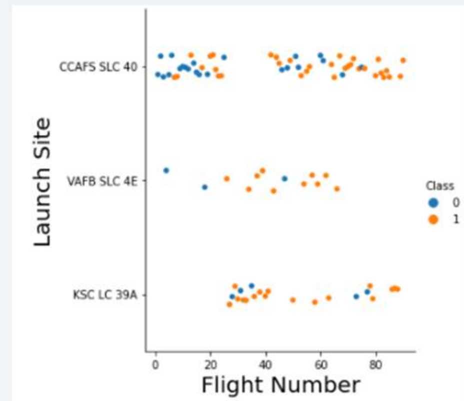
Data Wrangling

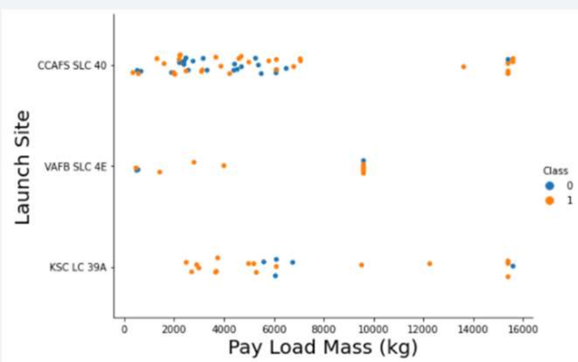


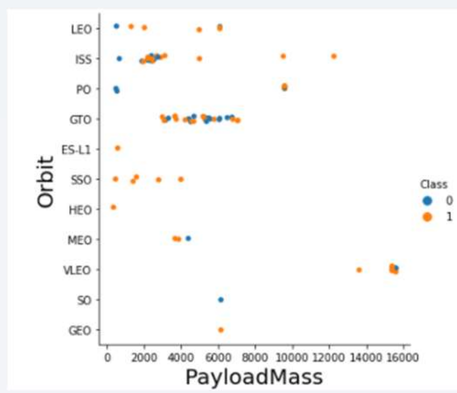
- GitHub Data Wangling

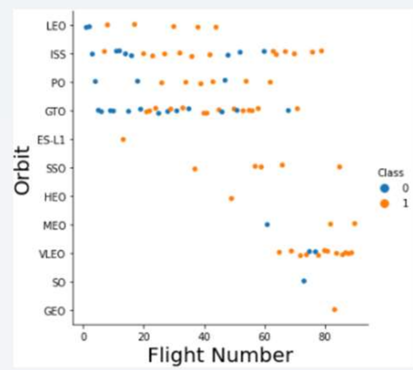
EDA with Data Visualization

- Scatter point chart to see the relationships between two factors:
    - Flight Number and Launch Site
    - Payload and Launch Site
    - Flight Number and Orbit type
    - Payload and Orbit type
- [GitHub EDA with Visualization]()



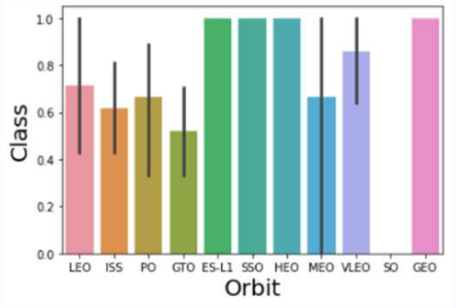






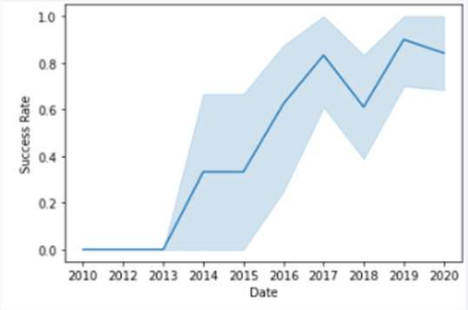
EDA with Data Visualization

- Bar chart to see the each orbit type's success rate



- Line chart to see the launch success trend over the years

- [GitHub EDA with Visualization]



EDA with SQL

- Summary of the SQL queries performed
  - Display the DISTINCT names of the Launch Site from the SpaceX table: there are 4
  - Display 5 records where Launch Site names start with 'CCA'
  - Display total payload mass carried by boosters launched by NASA (CRS) (customer column)
  - Display average payload mass carried by booster version 'F9 v1.1'
  - List the date when the first successful landing outcome in ground pad occurred using the MIN() function: result 12/22/2015
  - List the booster versions that have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the booster versions that have carried the maximum payload mass, using a subquery
  - List the failed landing outcomes in drone ship, their booster versions, and launch sites in 2015
  - Rank the count of landing outcomes between 6/4/2010 and 3/20/2017 in descending order

- GitHub EDA with SQL

Build an Interactive Map with Folium

- Marker: create a mark on map.
- Circle: add a highlighted circle area with a text label on a specific coordinate
- PolyLine: draw a line between two points on a map
- [GitHub Launch Site Folium]

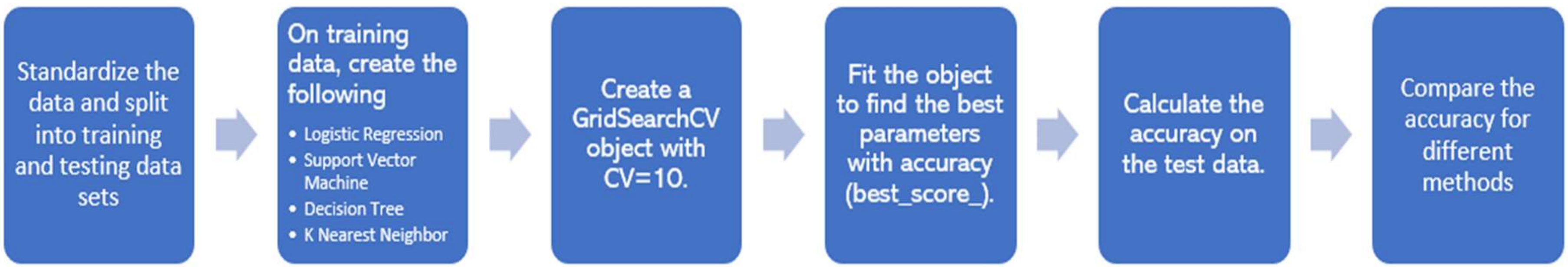
Build a Dashboard with Plotly Dash

- Dropdown List for Launch Site selection
- Pie chart to show total successful launches count for all sites
- Slider to select payload range
- Scatter Chart to show the correlation between payload and launches success
- GitHub Interactive Dashboard

Predictive Analysis (Classification)

- Standardize the data and split into training and testing data sets
- On training data:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K Nearest Neighbors (KNN)
- Create a GridSearchCV object with CV=10.
- Fit the object to find the best parameters with accuracy (best_score_).
- Calculate the accuracy on the test data.
- GitHub Machine Learning Prediction

Predictive Analysis (Classification)-continued



- GitHub Machine Learning Prediction

Results

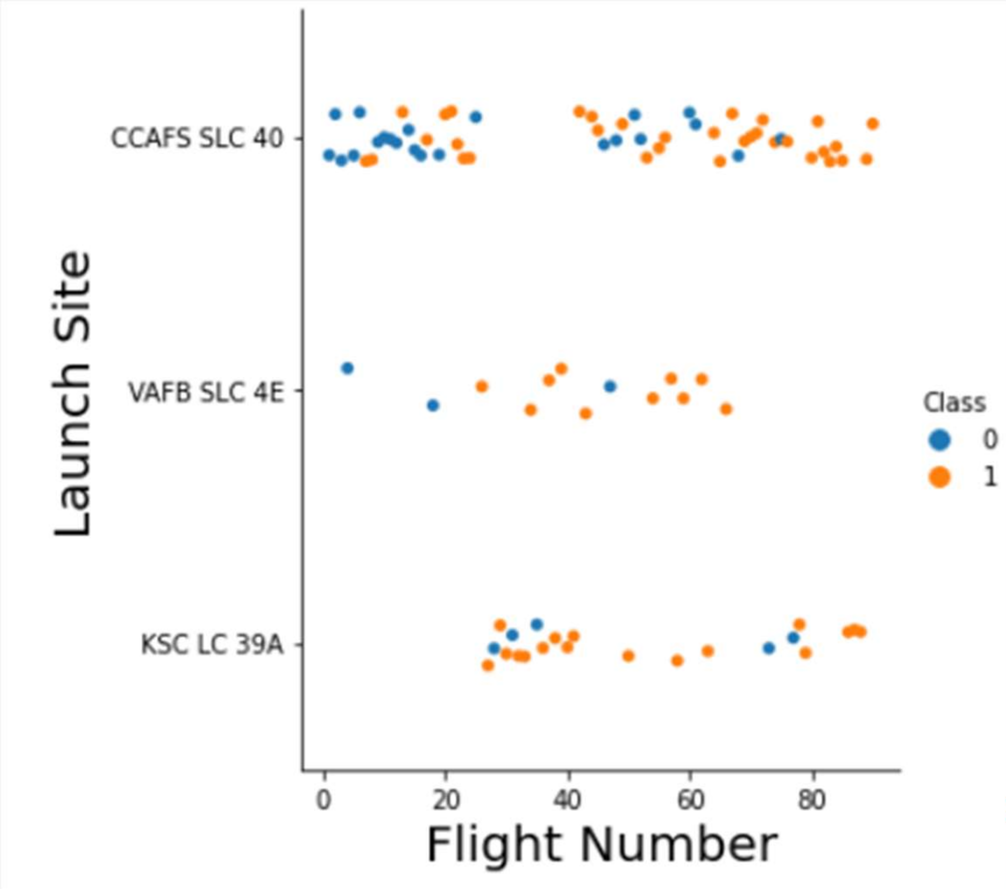
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

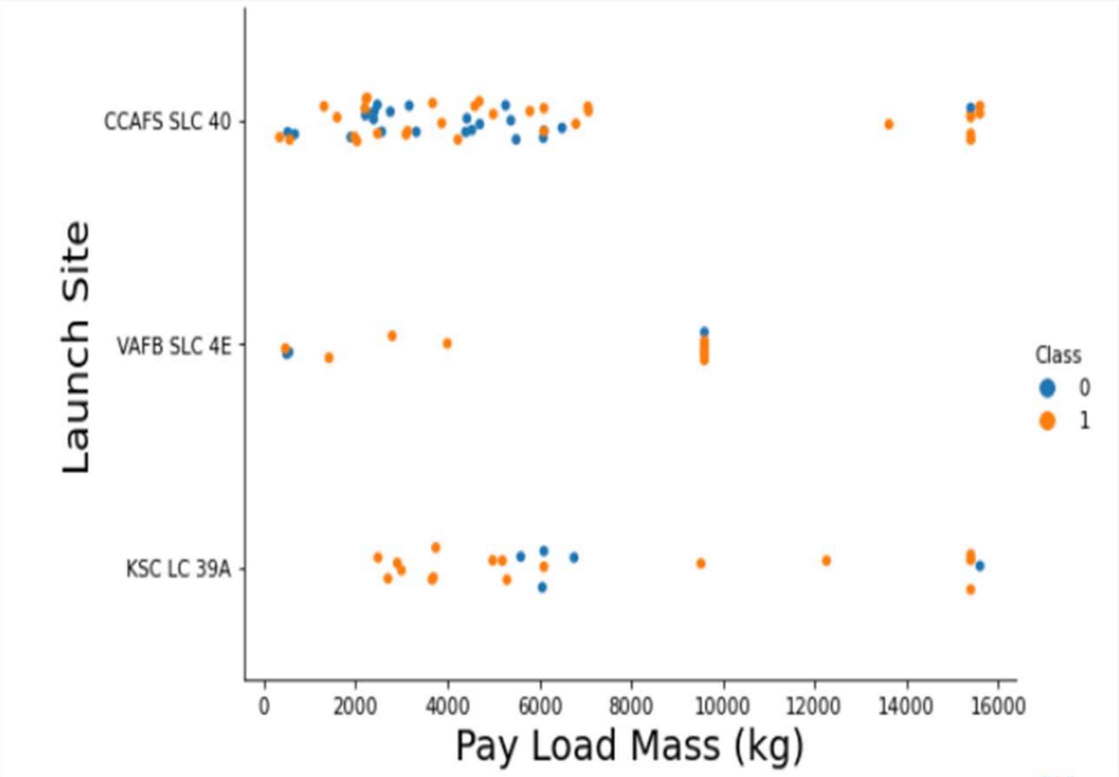
Success rate increases as flight numbers increase (positive relationship)



[GitHub EDA with Visualization]()

Payload vs. Launch Site

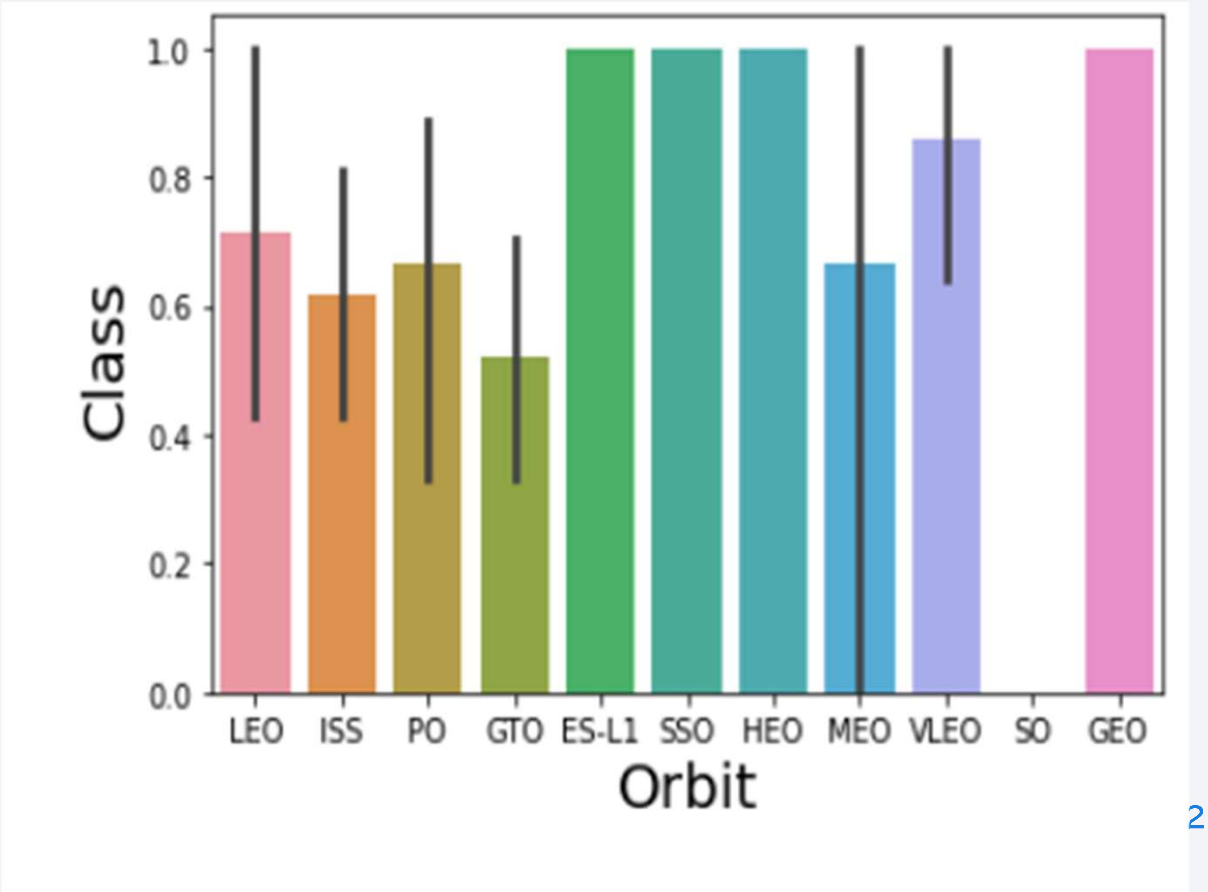
For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000).



[GitHub EDA with Visualization]

Success Rate vs. Orbit Type

GEO, HEO, SSO, and ES-L1 orbit types have the highest success rate

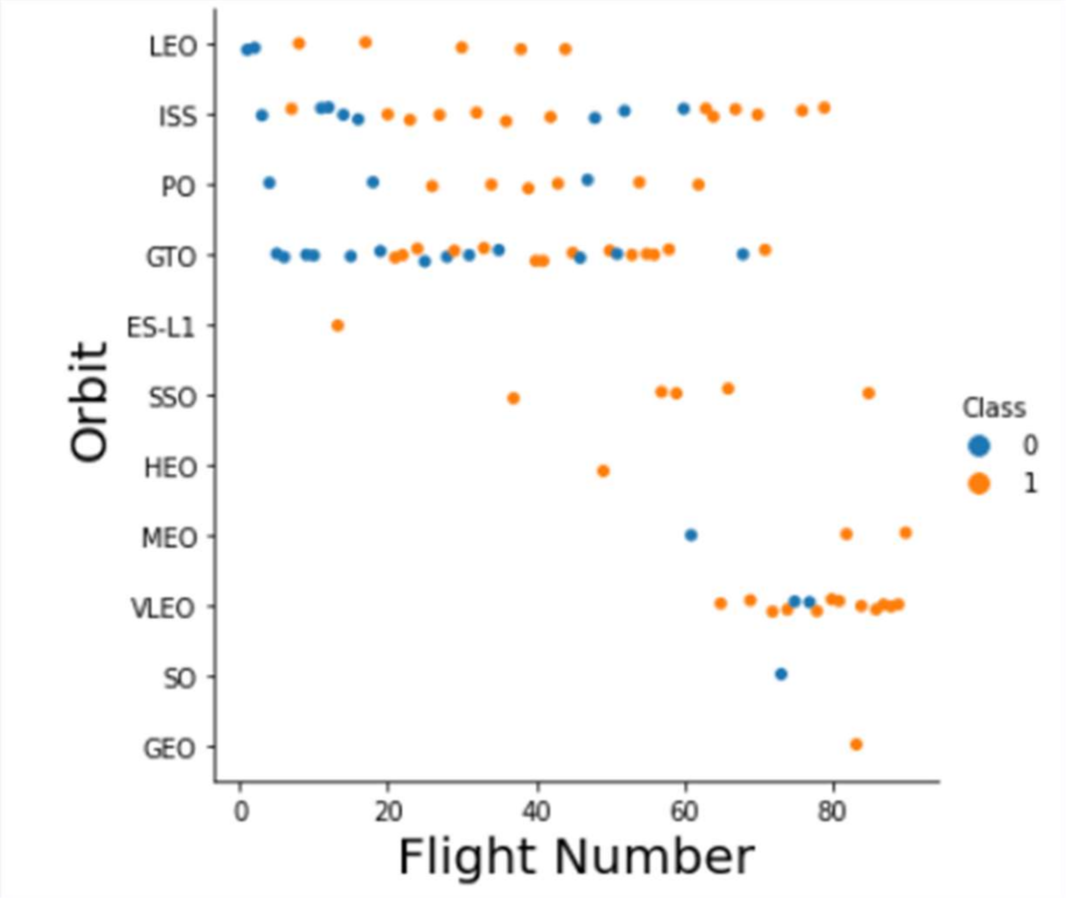


[GitHub EDA with Visualization]

Flight Number vs. Orbit Type

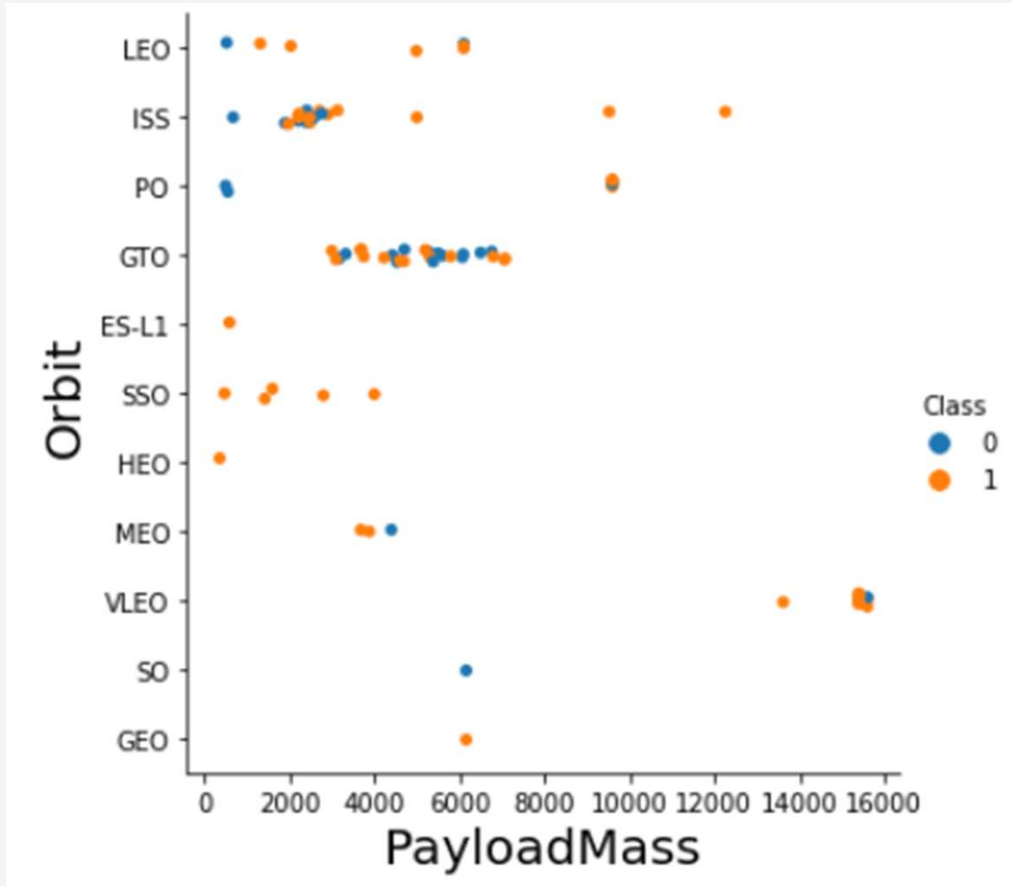
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

[GitHub EDA with Visualization]



Payload vs. Orbit Type

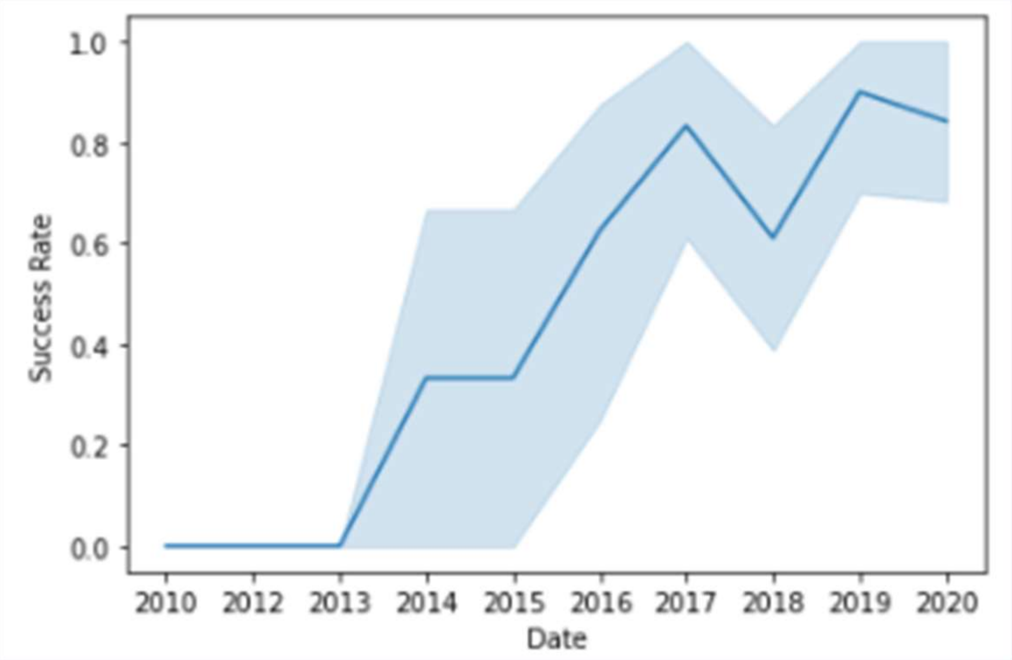
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



GitHub EDA with Visualization

Launch Success Yearly Trend

Success rate has been increasing since 2013, and has been above 0.6 since 2017.



[GitHub EDA with Visualization]

All Launch Site Names

- Using SQL to find the unique (distinct) launch sites:
- %sql select DISTINCT Launch_Site from SPACEXTBL2
- There are 4 distinct launch sites:

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

GitHub EDA with SQL

Launch Site Names Begin with 'CCA'

%sql select * from SPACEXTBL2 WHERE LAUNCH_Site like 'CCA%' limit 5

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

GitHub EDA with SQL

Total Payload Mass

%sql select sum(PAYLOAD_MASS__KG_) AS TotalPayloadMass from SPACEXTBL2 where Customer = 'NASA (CRS)’

totalpayloadmass
45596

GitHub EDA with SQL

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1:

```
%sql select avg(PAYLOAD_MASS__KG_) as AveragePayloadMass from SPACEXTBL2 where
Booster_Version = 'F9 v1.1’
```

averagepayloadmass
2928

GitHub EDA with SQL

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad:

%sql select MIN(Date) AS First_successfull_landing_date from SPACEXTBL2 WHERE
Landing__Outcome = 'Success (ground pad)’

first_successfull_landing_date
2015-12-22

[GitHub EDA with SQL]

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- %sql SELECT Booster_Version from SPACEXTBL2 WHERE Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000

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

- GitHub EDA with SQL

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- %sql SELECT COUNT(Mission_Outcome) from SPACEXTBL2 WHERE Mission_Outcome like 'Success%'

1
100

- [GitHub EDA with SQL]()

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- %sql select distinct Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL2 where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL) order by Booster_Version

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

- GitHub EDA with SQL

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- %sql select Booster_Version, Launch_Site, Landing__Outcome from SPACEXTBL2 where Landing__Outcome LIKE 'Failure%' and Date between '2015-01-01' and '2015-12-31'

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

- GitHub EDA with SQL

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

- Rank 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
- %sql select Landing__Outcome, count(Landing__Outcome) from SPACEXTBL2 where Date between '2010-06-04' and '2017-03-20' group by Landing__Outcome Order by Count(Landing__Outcome)

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

[GitHub EDA with SQL]

Section 3

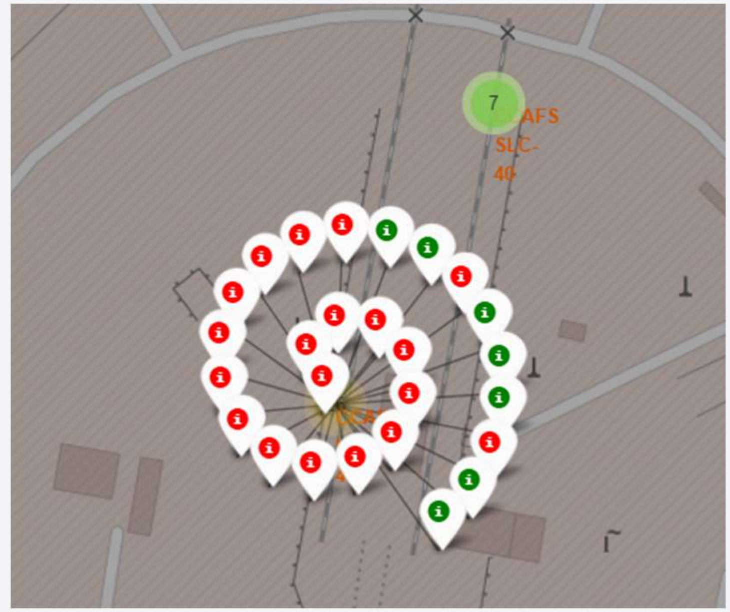
Launch Sites Proximities Analysis

Launch Site Locations

Launch sites are close to coastlines.

[GitHub Launch Site Folium]

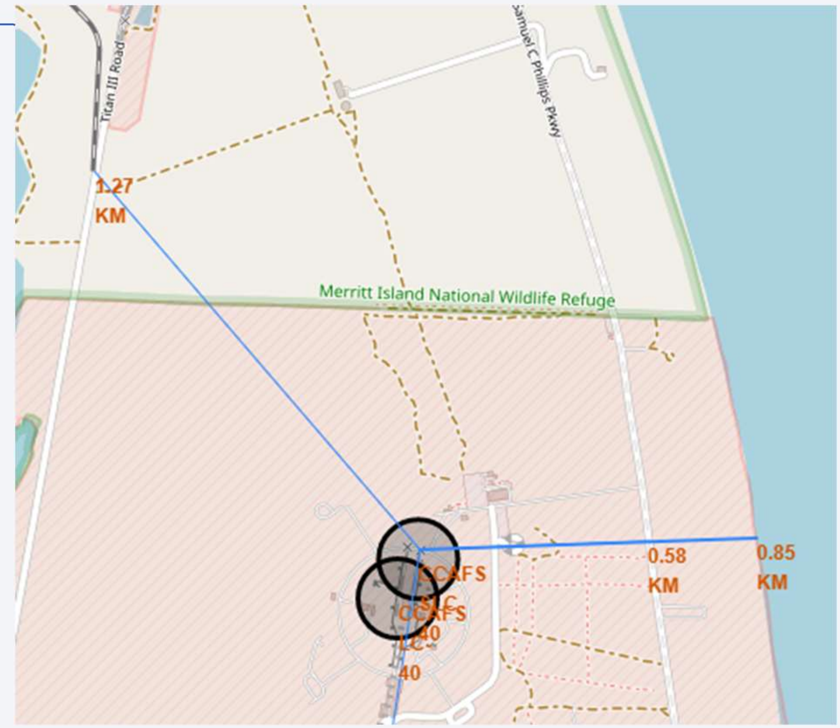
Color-Labeled Launch Outcome



- Green – Successful
- Red – Unsuccessful

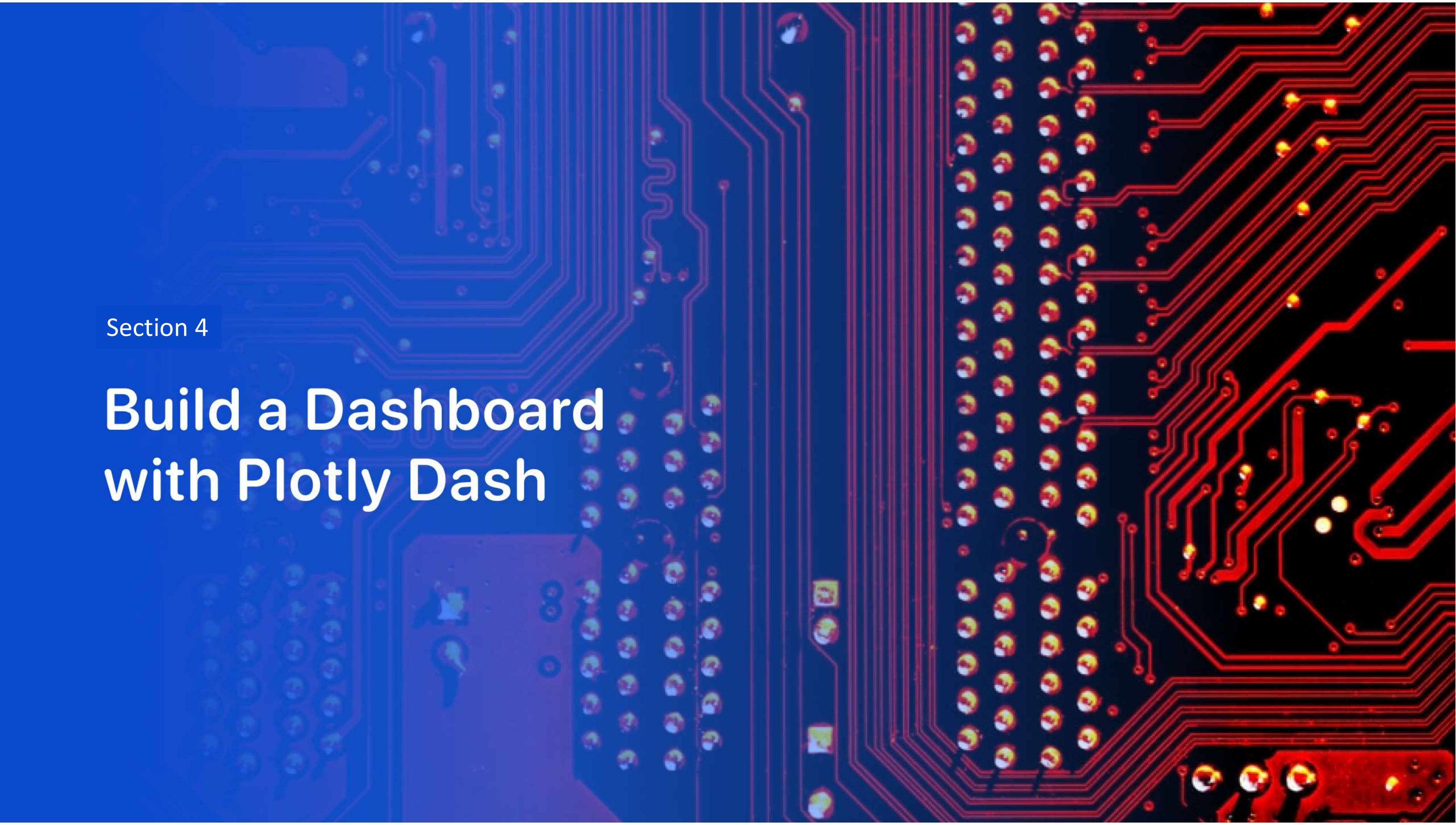
[GitHub Launch Site Folium]

Distances of Launch Sites to Railways, Highway, Coastline



Launch site is
- 1.27KM from railway
- 0.85KM from coastline
- 0.58KM from highway

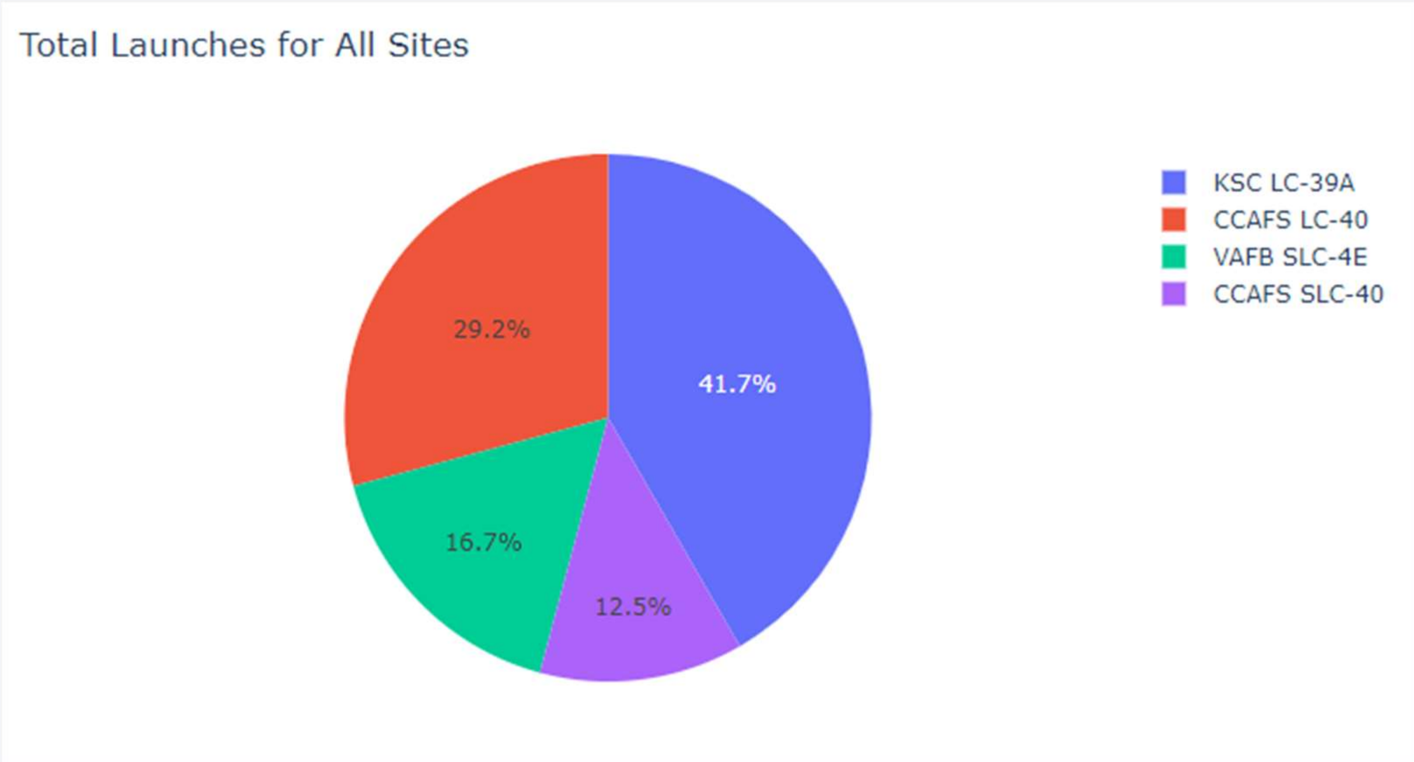
[GitHub Launch Site Folium]



Section 4

Build a Dashboard with Plotly Dash

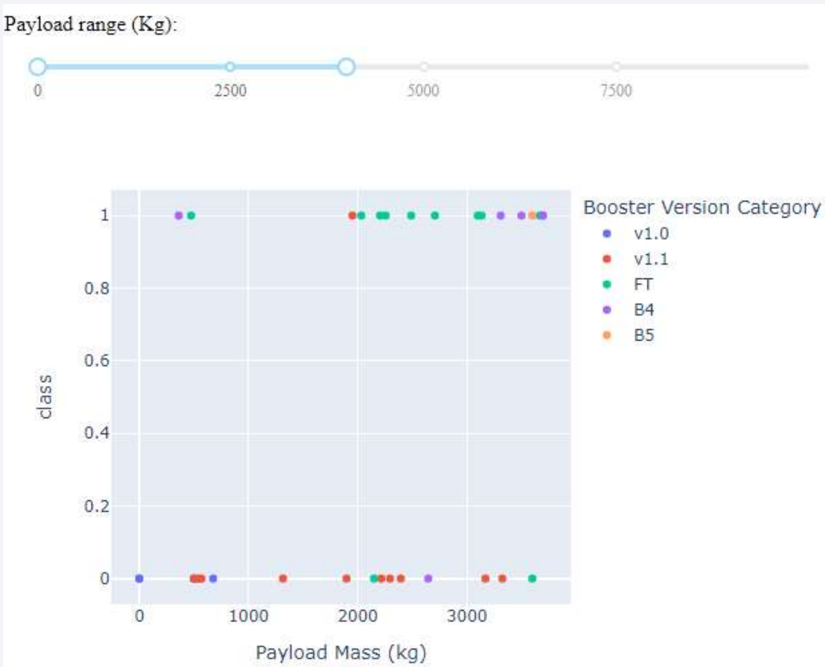
Launch Sites with Successful Launches

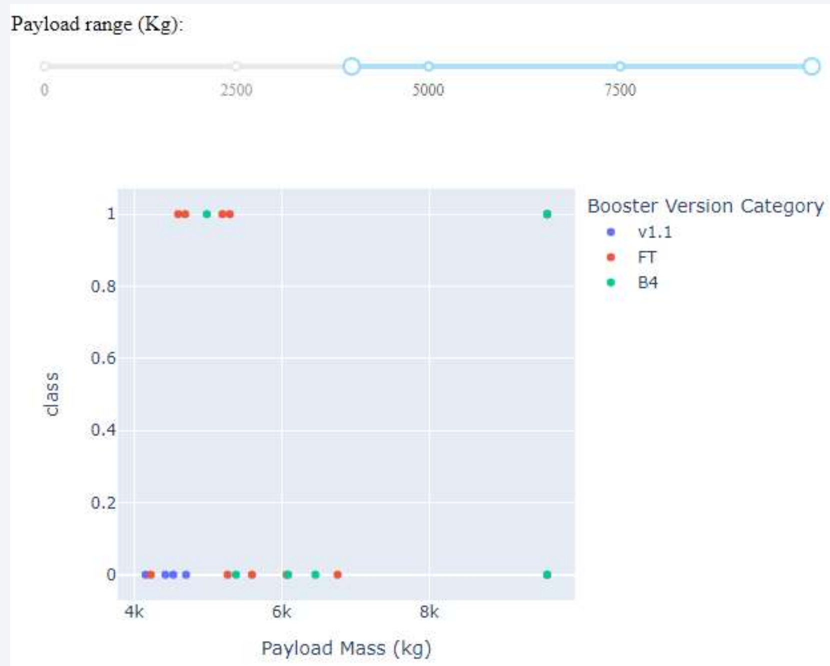


KSC LC-39A has the most successful launches, comparing to other sites.

[GitHub Interactive Dashboard]

How Does Payload Affect Launch Outcome on All Launch Sites





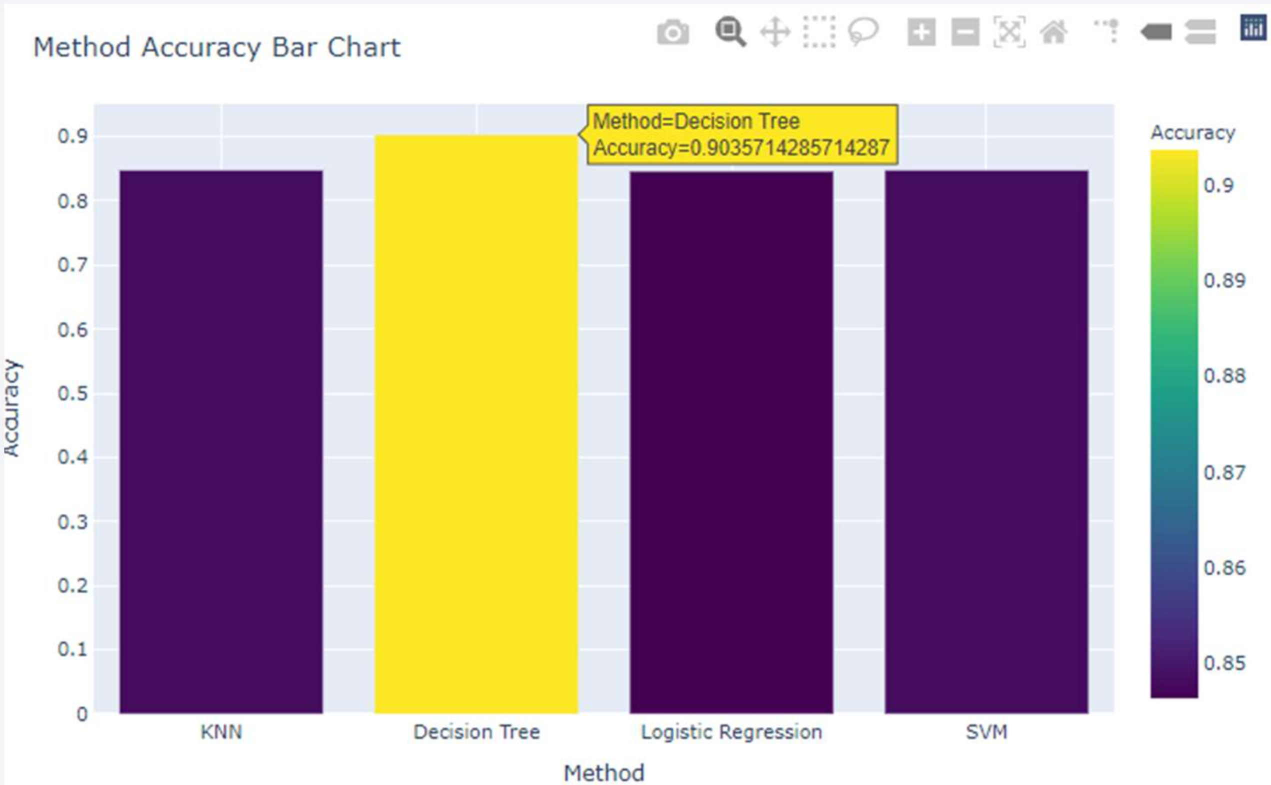
The successful rates for lighter payload is generally higher than heavier payloads.

GitHub Interactive Dashboard

Section 5

Predictive Analysis (Classification)

Classification Accuracy

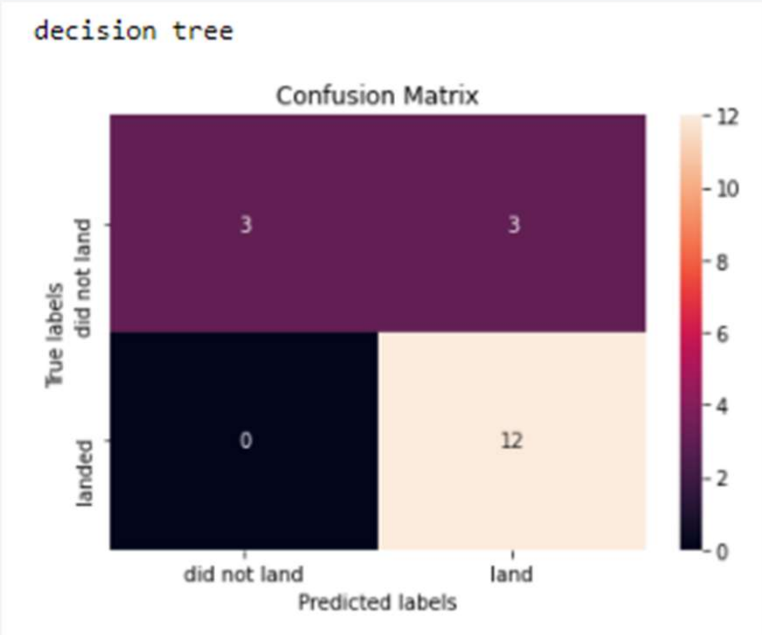


- Decision Tree has the highest accuracy

[GitHub Machine Learning Prediction]

Confusion Matrix

- The model prediction is mostly accurate. However, it predicted three landed while they did not land.



[GitHub Machine Learning Prediction]

Conclusions

Factors that impacted SpaceX's successful recoveries

- Launch date in 2017 or later
- Lighter payload
- Launched from KSC LC-39A
- Successfully recovered via drone ship

Our best model, Decision Tree, predicts the outcome with an accuracy of ~90%.

Appendix

- [GitHub Data Collection API]
- [GitHub Data Collection Web Scraping]
- [GitHub Data Wangling]
- [GitHub EDA with SQL]
- [GitHub EDA with Visualization]
- [GitHub Interactive Dashboard]
- [GitHub Launch Site Folium]
- [GitHub Machine Learning Prediction]


Thank you!