IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fauzi Achmad Ramadhan
December 2021

Outline

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

Executive Summary

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia, Perform Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. Performing more interactive visual analytics. Create a machine learning pipeline to predict if the first stage will land.
- Four machine learning results similar accuracy around 80%, confusion matrix shows that all machine learning models perfectly predict “Land” but poorly predict “did not land” due to very few test dataset

Introduction

- Background:

Space X launch rockets with very cheap cost compare to other company, this is because Space X stage 1 rockets can be reuse if its land correctly

- Problems:

We (Space Y) want to compete with Space X and want to know when Space X stage 1 rockets will land correctly or not using Machine Learning models and data from Wikipedia Space X historical launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Getting data from Space X API and Space X Wikipedia page
- Perform data wrangling
  - Convert outcomes into Training Labels with 1 means the booster successfully landed, 0 means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create ML models (sklearn), tune hyperparameters (gridsearchcv), and evaluate using accuracy (score())

Data Collection

- The data acquired by collecting from 2 sources, the first one is from Space X public API and the second one is from Space X Wikipedia table.
- By using Space X public API we can get data as follow:

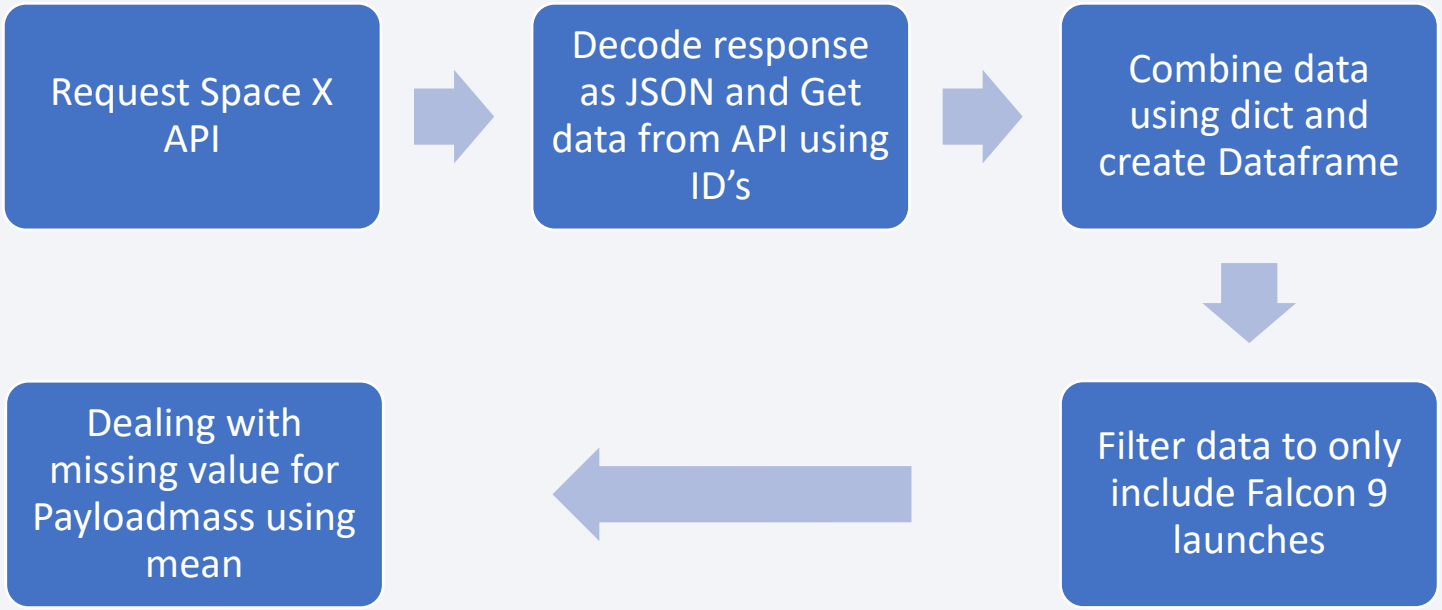
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude

- By webscrapping table from Space X Wikipedia we can get data as follow:

- Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time

Data Collection – SpaceX API

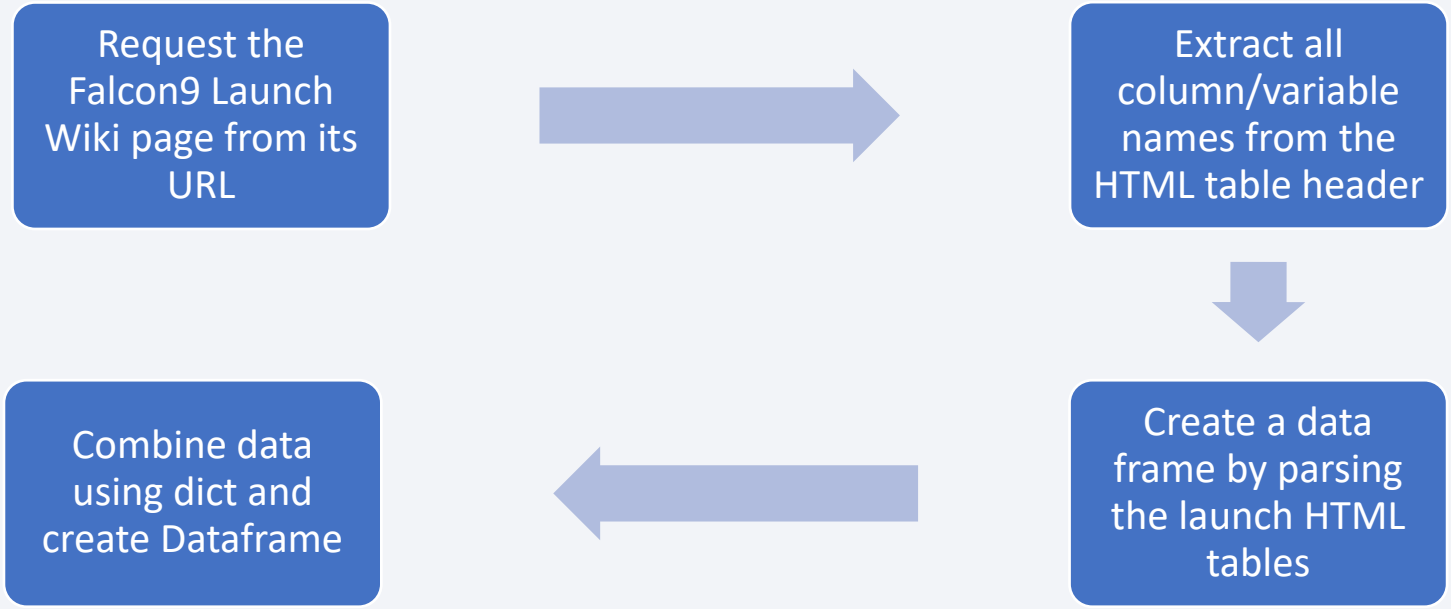
- Data Collection Flowcharts – Space X API



https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Data Collection Flowcharts – Scraping



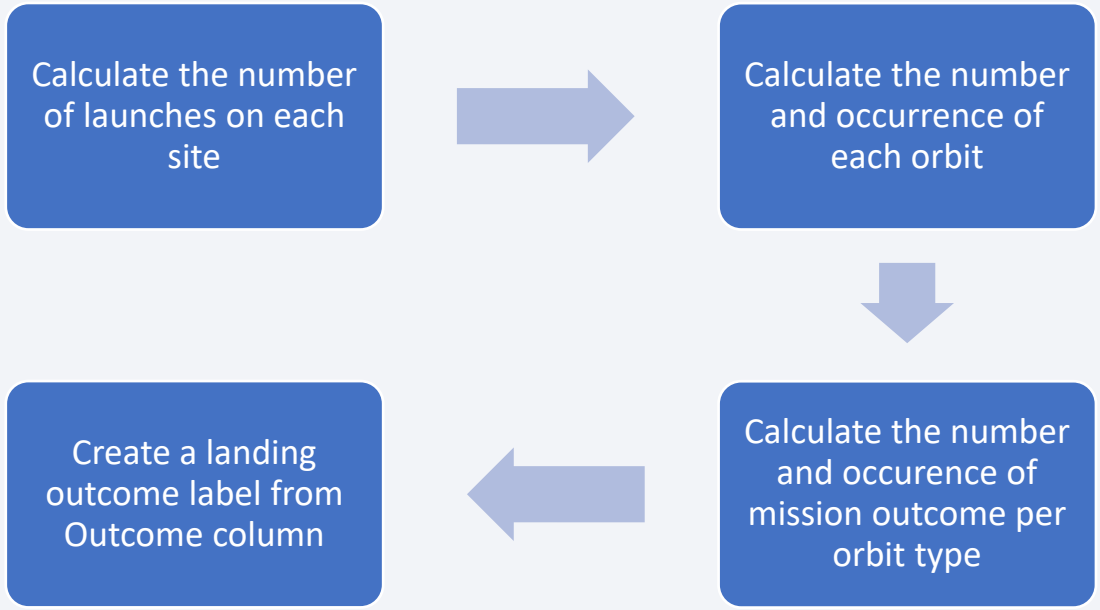
https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Data Wrangling

- As of the last step in the flowchart, we create a list where the Outcome is failure = 0; otherwise, = 1.

- Where failure is 'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None‘, other than that is successful outcome



https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Wrangling.ipynb

EDA with Data Visualization

- In this part we visualize variable vs other variable to see the relationship between the 2 variables
- List of Variables visualize :

- Flight Number and PayloadMass (Scatter Plot)
- Flight Number and Launch Site (Scatter Plot)
- Payload and Launch Site (Scatter Plot)
- Success rate of each orbit type (Bar Plot)
- FlightNumber and Orbit type (Scatter Plot)
- Payload and Orbit type (Scatter Plot)
- The launch success yearly trend (Line Plot)

- Features Engineer that done in this part is create dummy variables for categorical columns (FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial)

https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Load Data to SQL and Connect python to database
- Query information about the data like unique value of launch site, total payload mass carried, avg payload mass carried, first ever successful landing in ground, boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, List the total number of successful and failure mission outcomes, List the names of the booster_versions which have carried the maximum payload mass, List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, and 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

https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map with circles and markers (name)
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

- The purposes of interactive map is to visualize where did the launches take places and where successful landing and failure landing take places

https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

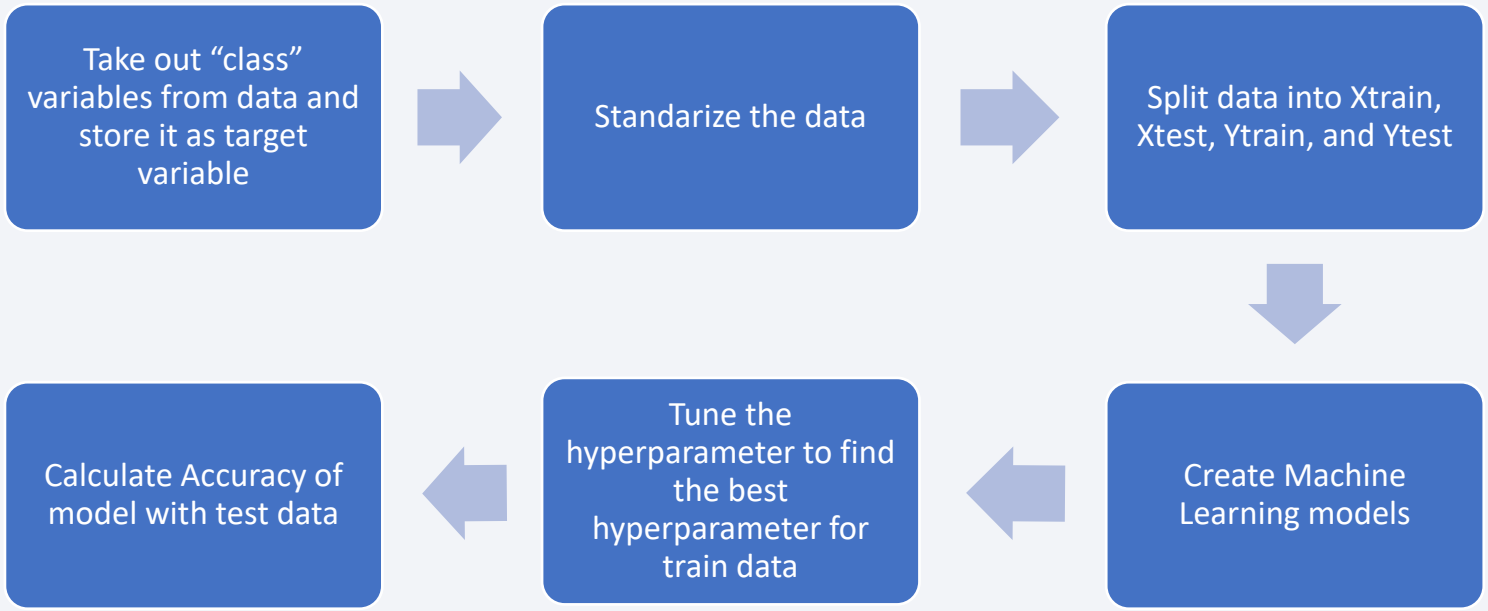
Build a Dashboard with Plotly Dash

- Build dashboard of scatter plot and pie chart
- Scatter plot can customize the input of payload mass from 0 to 10000kg and can choose all launch sites or single site, help us to figure out how relationship between payload mass and successful landing
- Pie chart shows portion of successful and failure landing of all sites or single sites, this chart help us to visualize successful landing rate

https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Dash.py

Predictive Analysis (Classification)

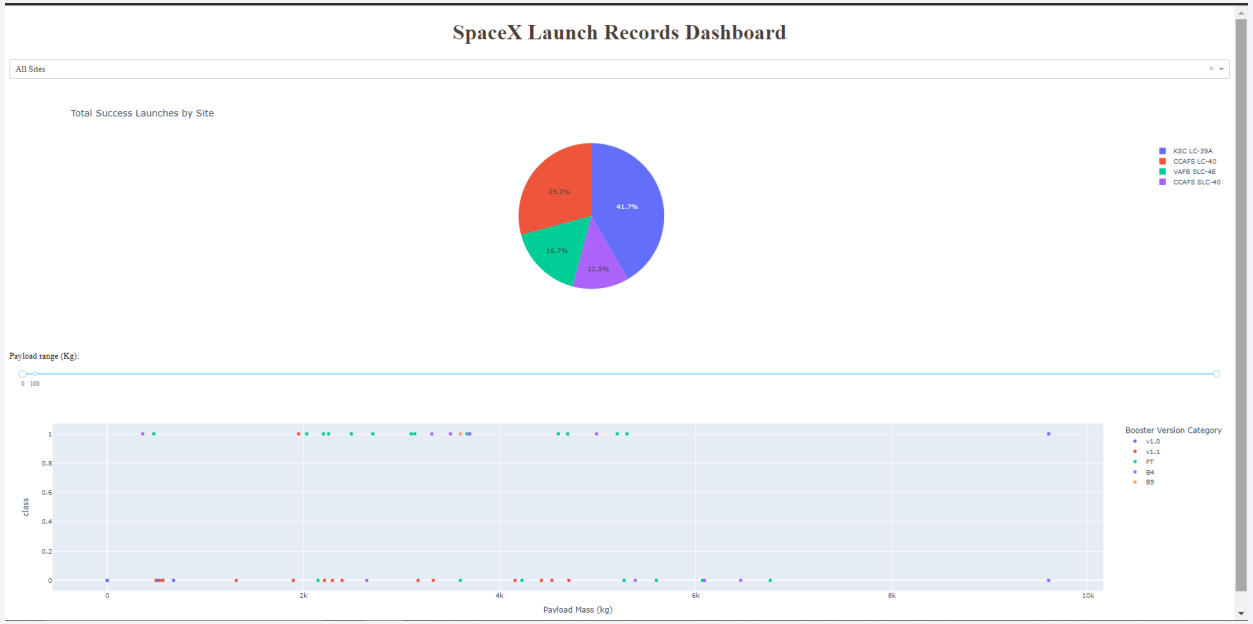
- Predictive Analysis – Classification



https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Results

- With EDA we get some variables vs variables scatter plot give us more information about successful landing and failure landing, also we can can observe that the sucess rate since 2013 kept increasing till 2020, scatter plot and pie chart demo from dash is shown below



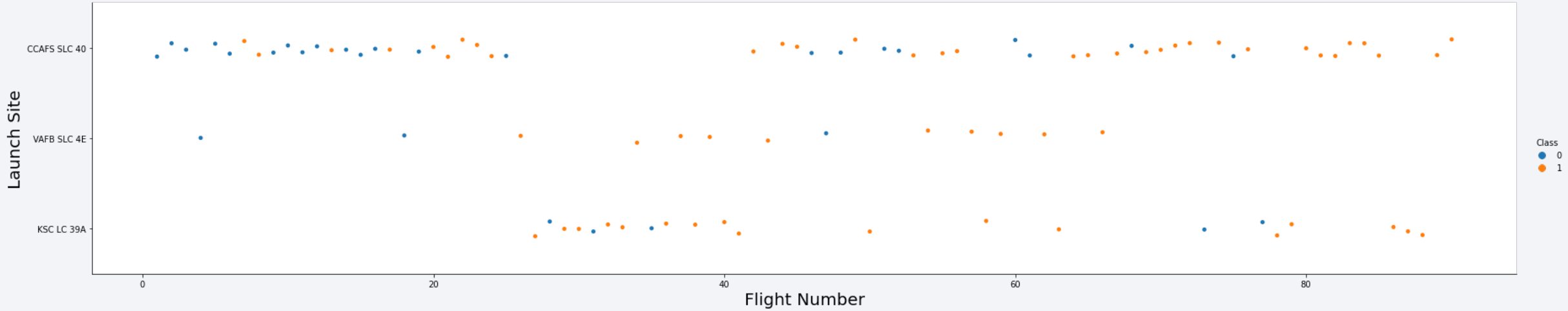
- With Predictive analysis, our Machine Learning models highest accuracy is around 83%

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site

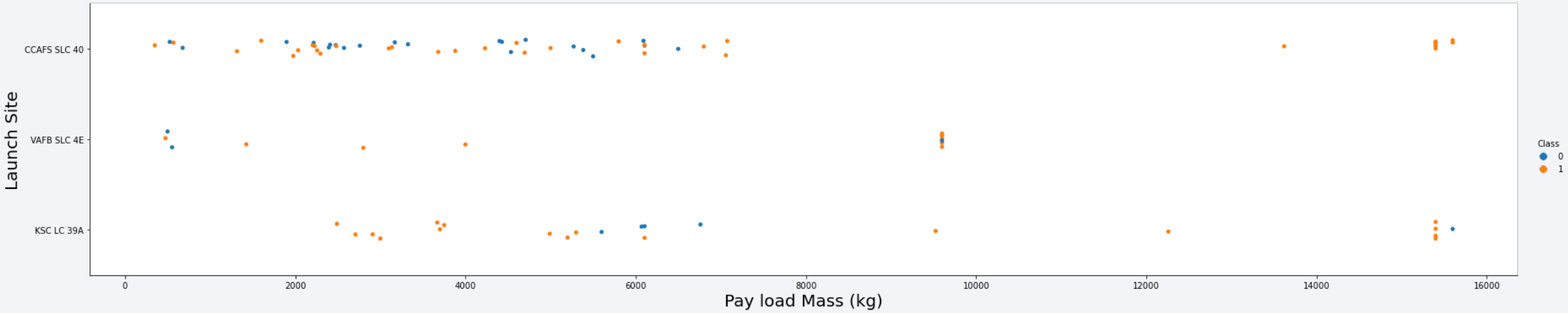


Blue is failure and orange is successful

CCAFS SLC 40 has less success rate than VAFB SLC 43 and KSC LC 39A, also when flight number increase, the first stage more likely to land success

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

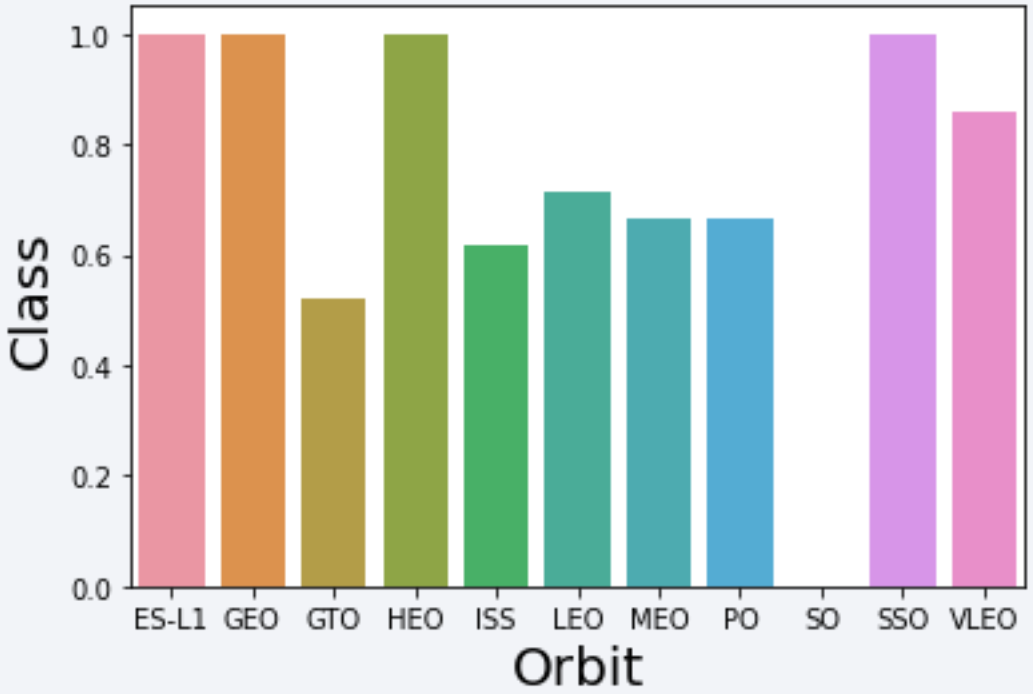


Blue is failure and orange is successful

VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

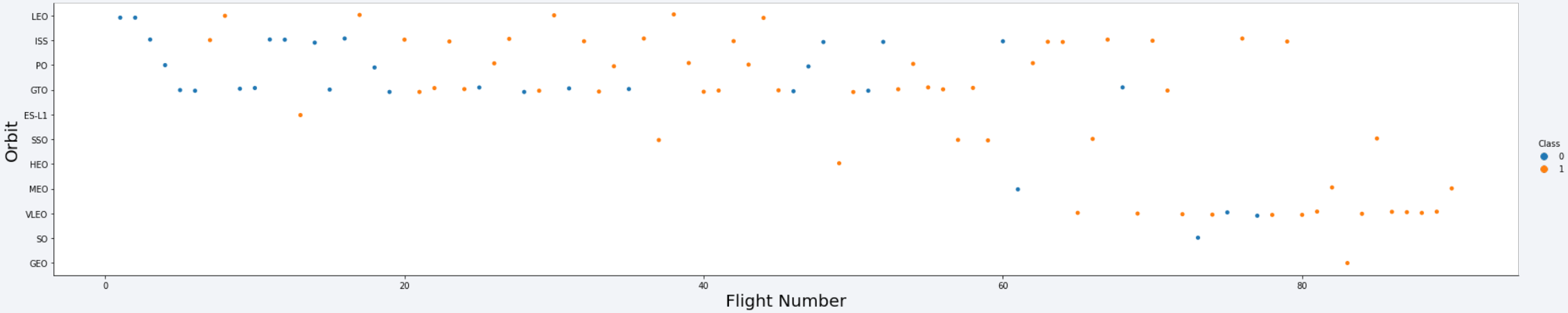
- Success Rate vs. Orbit Type



Orbit ES-L1, GEO, HEO, and SSO has 100% success rate, but their sample size are very low (only 1 sample, for SSO has 5 sample), VLEO also has very high success rate with 14 samples

Flight Number vs. Orbit Type

- Scatter plot of Flight Number vs. Orbit Type

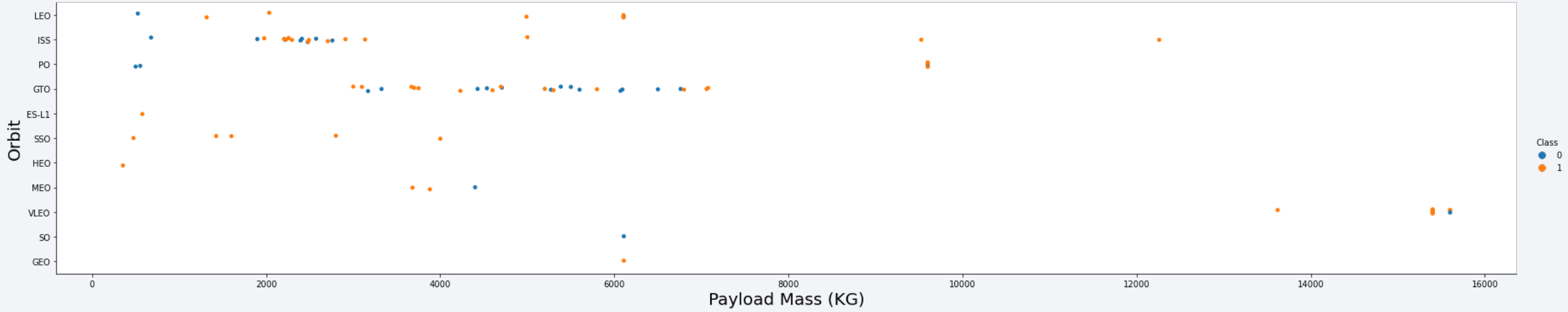


Blue is failure and orange is successful

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.

Payload vs. Orbit Type

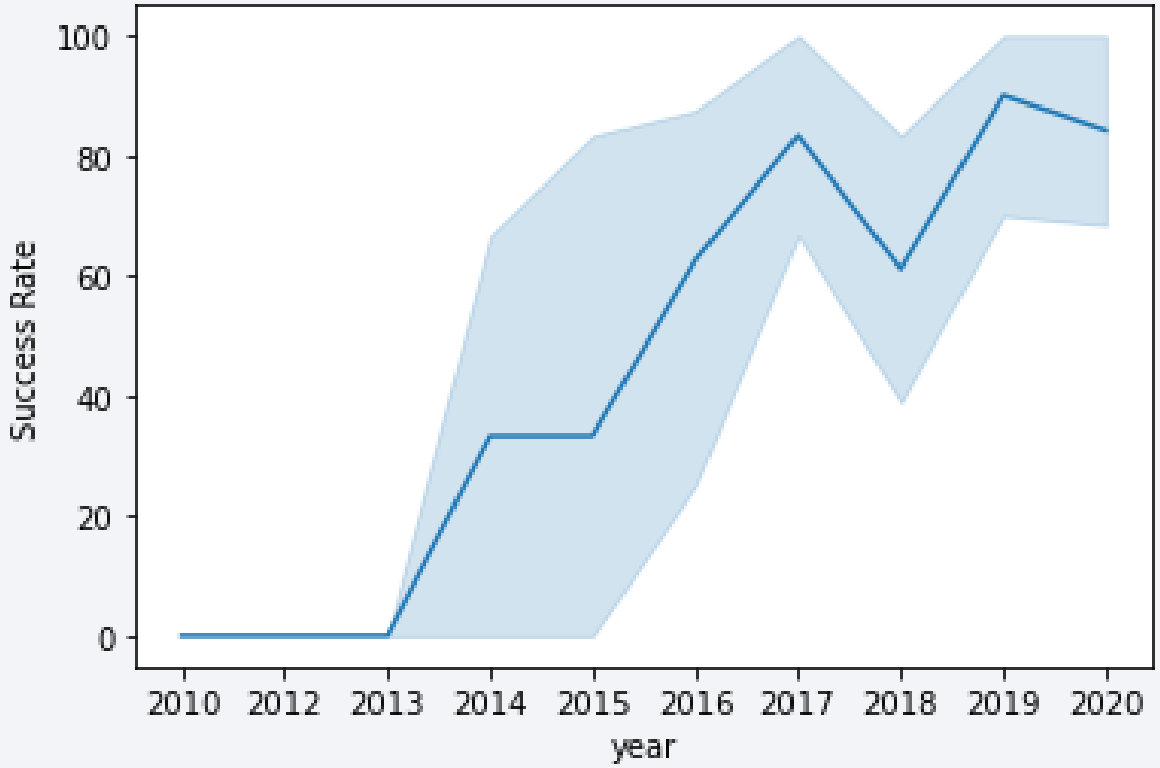
- Scatter plot of Payload vs. Orbit Type



Blue is failure and orange is successful

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

Launch Success Yearly Trend



We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select DISTINCT launch_site from SPACEXDATASET
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a
Done.
```

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

There are 4 unique launch site as shown above

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like '%CCA%' limit 5
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/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

Shown the first 5 data that launch site begin with CCA

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEXDATASET where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.data
Done.
```

1
45596

Total Payload Mass is 45596

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEXDATASET where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.a
Done.
```

1
2534

Average Payload Mass by F9 v1.1 is 2534

First Successful Landing Date

```
%sql select min(date) from SPACEXDATASET where mission_outcome ='Success'
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0
Done.
```

1
2010-06-04

The first ever successful landing is on 2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXDATASET where mission_outcome='Success' and landing__outcome='Success (drone ship)' and payload_mass__kg_ betwee
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Successful Drone ship landing with payload between 4000 and 6000 is shown above

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) from SPACEXDATASET group by mission_outcome
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Success mission is 99 compare to failure mission is 1, this means almost all of failure landing are intended by Space X, and with 1 success with payload status unclear

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
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

These booster is used to carry maximum payload by Space X, these booster start wit F9 B5 B10 which indicate these type of booster likely intended to carry maximum payload

2015 Launch Records

```
%sql select booster_version, launch_site from SPACEXDATASET where landing__outcome='Failure (drone ship)' and date like '%2015%'
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

2015 failed drone ship landing occurs 2 times in the same launch site with similar booster version

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

```
%sql select landing__outcome, count(*) as counts from SPACEXDATASET where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by
```

```
 * ibm_db_sa://gxz03189:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

landing__outcome	counts
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

There is 10 launch that didn’t attempt to landing, but drone ship seems to fail and success in 50:50 with each outcome has 5 outcome, followed by controlled ocean and success ground pad with 3 outcome

Section 4

Launch Sites Proximities Analysis

Launch Site with folium

All Sites are near ocean, with KSC LC and CCAFS landing site are close to each other

Launch Site successful and failure

In folium the marker can be click to show each launch site successful and failure landing, green shows successful and red for failure, here shown for CCAFS SLC - 40

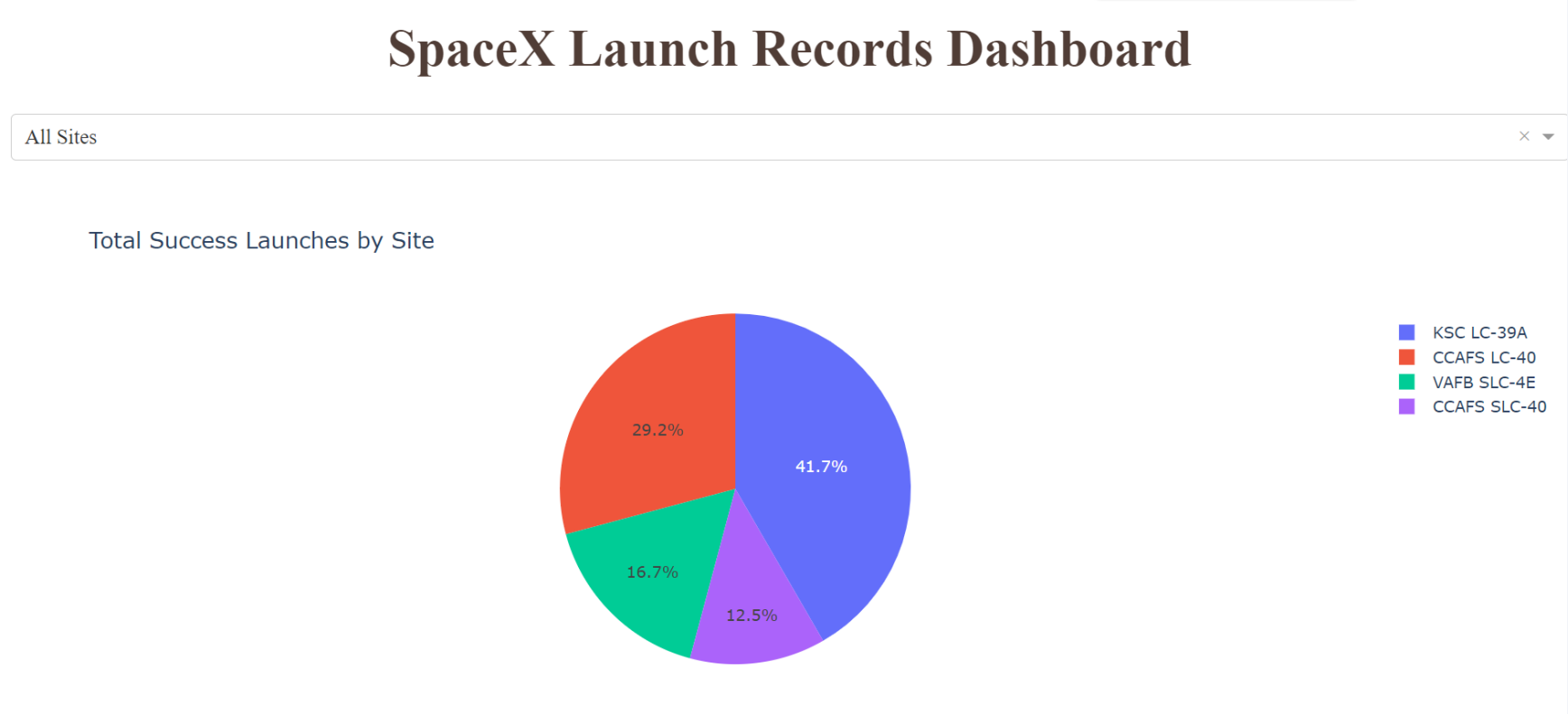
Launch site location

- CCAFS SLC to the clostest coasts is 0.90 km

Section 5

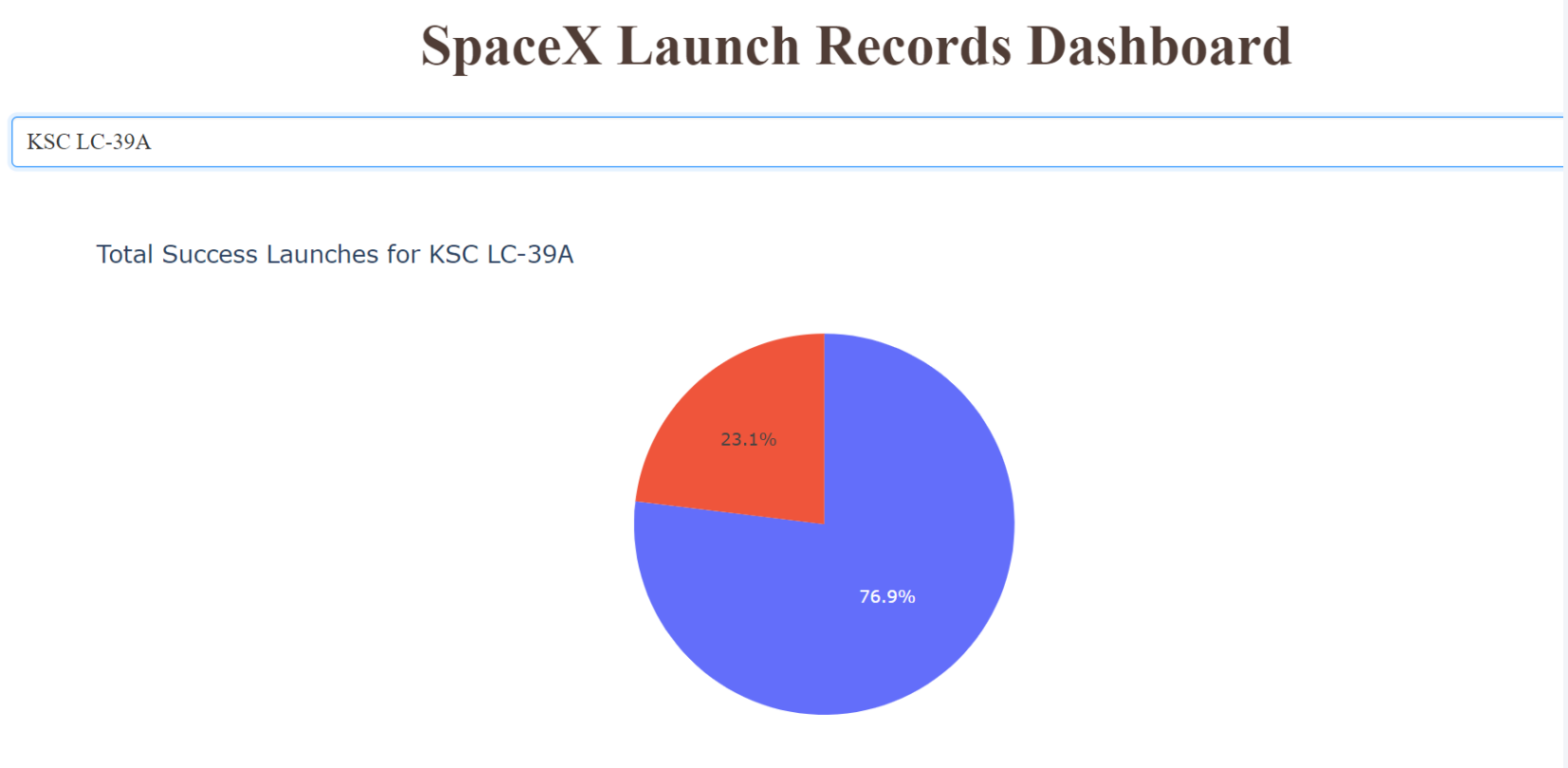
Build a Dashboard with Plotly Dash

Total Success Launch All Site



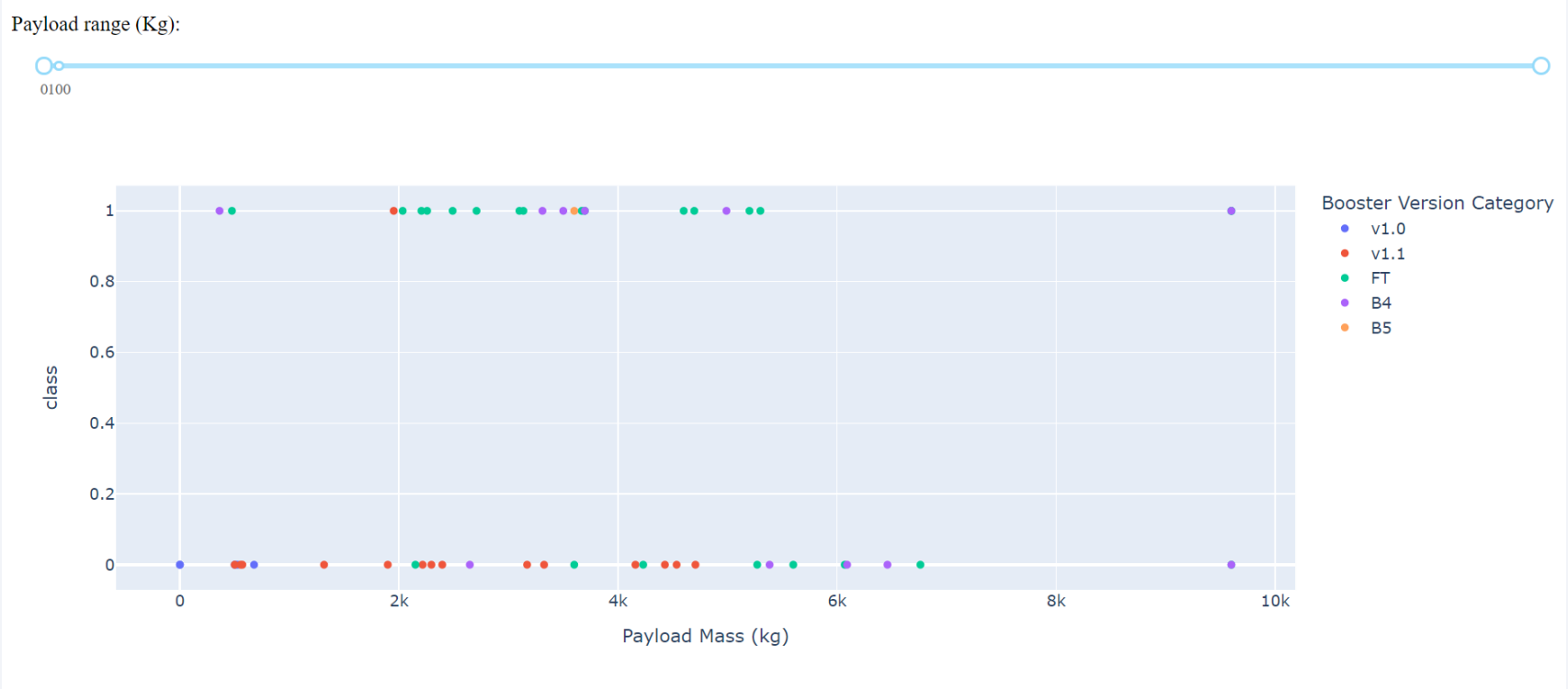
The Majority of success launches is on KSC LC-39A Site, behind that is CCAFS LC-40 and CCAFS SLC-40, VAFB SLC-4E has the least success launches

Highest total success launches



As seen in previous slide, KSC LC-39A is the highest launch site success rate, with 76.9 % success landing and 23.1% failure landing

Payload vs Launch Outcome



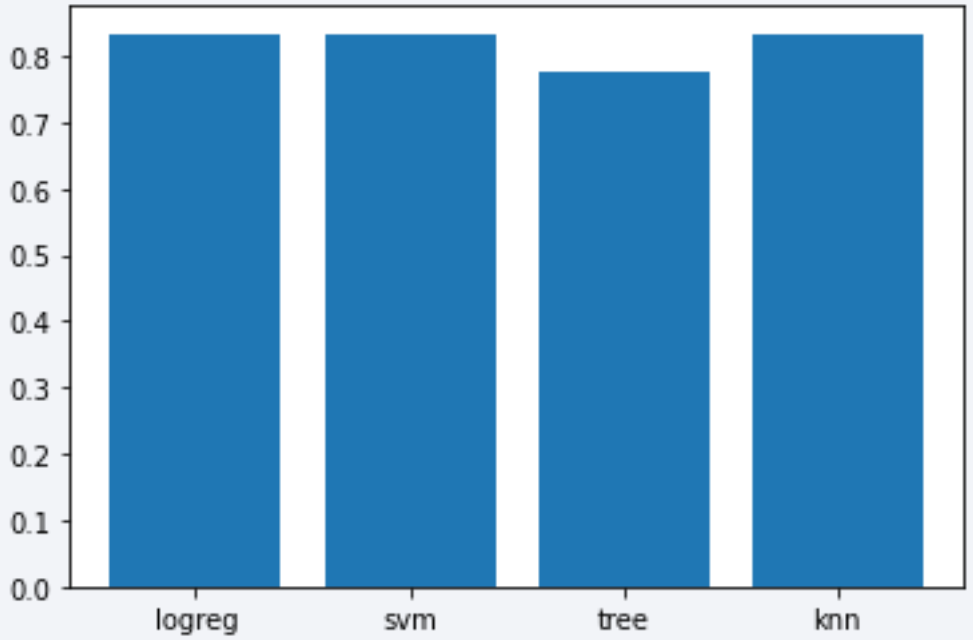
For booster version FT and B4 have high success rate, but after payload more than 5k kg, both version fail more than the succsess, on the other hand V1.0 and 1.1 have very low success rate, meanwhile B5 only have 1 observation so we can’t conclude anything

Section 6

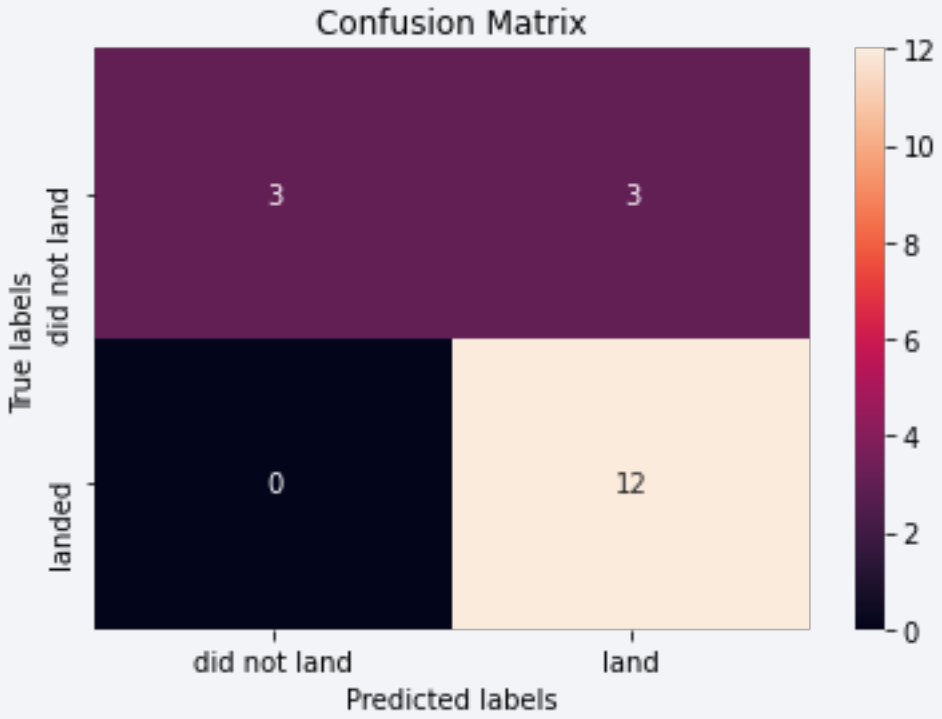
Predictive Analysis (Classification)

Classification Accuracy

From the analysis, we get Log regression, SVM, and KNN has 83.33% accuracy, meanwhile Decision Tree has 77.77% accuracy, note that this accuracy is generate with 18 data test, so more test data required to determine which model is the best among the other



Confusion Matrix



Again as shown is confusion matrix of Log regression, SVM, and KNN that only has 18 test data, they did a really good job on land prediction but not so good on did not land prediction

Conclusions

- Our Goal is to predict whether Stage 1 land successful or not, if land successful then Space X will reuse Stage 1 for future launches and save cost of launches.
- Space X data acquired from public API and from Wikipedia page.
- After EDA, we found there are only 4 launch site (with 2 site approximately the same place) and at site KSC LC is the highest successful landing
- Booster Version FT and B4 has the highest successful rate under 5k kg Payload Mass
- Orbit type VLEO is the highest successful rate with observation more than 10
- Using Machine Learning Model we got 83.33% using Log regression, SVM, and KNN, but these result acquired with minimal test data, more data required to be more precise what model is the best among other.

Appendix

- Github: https://github.com/fauzi417/Final-Chapter-IBM-Data-Science-Space-X

Instructors:

- Joseph Santarcangelo
- Yan Luo
- Rav Ahuja
- Hima Vasudevan
- Aije Egwaikhide
- Polong Lin
- Saishruthi Swaminathan
- Svetlana Levitan
- Romeo Kienzler
- Azim Hirjani
- Saeed Aghabozorgi
- Alex Aklson
- Lakshmi Holla

Thank you!