IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Xiao LIU>

<27 June 2020>

Outline

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

Executive Summary

Entering the space race against Space X, the key to bring down the cost is to reuse the first stage of the rocket. In this project, the overall aim is to use data science to make predication of the first stage landing outcome.

Project data were collected using SpaceX API and Falcon9 Launch Wiki page. Collected data were explore unique variable/ attribute types and relationship using SQL, pandas and matlibplot. Interactive visual analytic tools are build using Folium and Plotly Dash. Classification models were used for perform predictive analysis.

We identified the attributes *Launch Site, Orbit, Payload Mass, Booster_Version* has impact on the first stage landing outcome, the key findings are listed below:

- The optimal location for rocket launch is at *KSC LC-39A*
- Rockets launched to *SSO* orbit has the highest successful rate
- Highest launch success find for payload mass in the range between *3000 – 4000 kg*
- *F9 FT* Booster version seems have the highest launch success rate

Decision Tree shown the best performance up to 92.2% on the whole dataset

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel **affordable** for everyone
  - SpaceX has demonstrated to reduce the cost **over 60%** for those launched to Low-Earth Orbit space craft, by reuse its' first stage of rocket
  - We aim to make our company SpaceY **competitive** against SpaceX by launch cost
- Project Aims/Scopes
  - We want to determine if the first stage of the spacecraft will land, using only the public historical launch information
  - We want to find what attributes of launches would help determine of the first stage landing outcome
  - We want to build a machine learning model to make the prediction on this

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data were collected using SpaceX API and Falcon9 Launch Wiki page
- Perform data wrangling
  - Explore unique variable/ attribute types
  - Determine desired inputs and output for model prediction
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected models: Logistical Regression, Decision Tree, SVM, K-Nearest Neighbours
  - Models are trained and optimized on 80% training set, evaluate on 20% testing set

Data Collection – SpaceX API

Request and parse the SpaceX launch data using the GET request

→

Filter the dataframe to only include Falcon 9 launches

→

Dealing with Any Missing Values

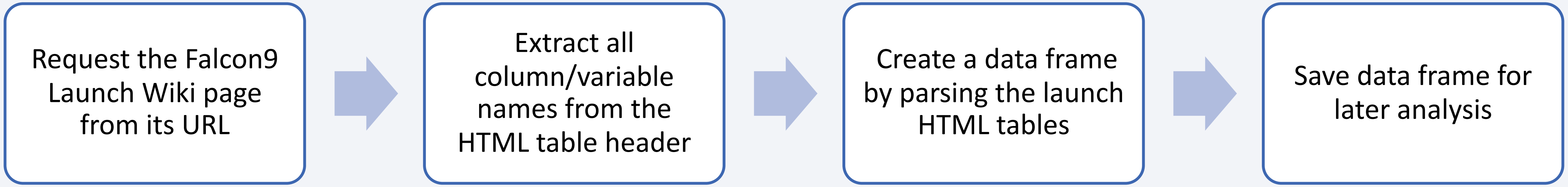
(Fill mean for numeric variables)

→

Save data frame for later processing

- https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



- https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/jupyter-labs-webscraping.ipynb

Data Wrangling

Load Data from .csv

to Pandas DataFrame

→ Find the number of launches on each site

→ Find the number and occurrence of each orbit

→ Find the number and occurrence of mission outcome per orbit type

→ Create a landing outcome label from Outcome column

- https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- THREE types of charts were plotted
  - Scatter plots
    - To explore how the launch outcome were affected by the relationship of selected two variables listed below
      - (Flight Number, Payload Mass, Launch Site and Orbit)
  - Bar plot
    - To check any relationship between success rate and orbit type
  - Line plot
    - To get the average launch success trend through time
- https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd392440274 1b70cb20a65440/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries were used to explore the information related with few launch attributes and the relationship between them

- The attributes are *Launch Sites*, *Launch Outcomes*, *Payload Mass*, *Booster Version*, *Customer*, and Date

- https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd392440274 1b70cb20a65440/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- FOUR types of objects were added on the map
  - Marker – To all launch sites on a map by its coordinates
  - Circle – To highlight the area with a text label on a specific coordinate
  - Icon – To indicate if this launch was successed or failed
  - PolyLine – To make the distances between a launch site to its proximities

- https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd392440274 1b70cb20a65440/lab_jupyter_launch_site_location_folium.ipynb

Build a Dashboard with Plotly Dash

- FOUR items were added on the Dashboard
  - A Launch Site Drop-down Input
    - We have four different launch sites, and we would like to first see which one has the largest success count
  - A success-pie-chart based on selected site dropdown
    - To visualise successful launch distribution of all sites, or launch success rate for the selected launch site
  - A Range Slider to Select Payload
    - To find if variable payload is correlated to mission outcome. From a dashboard point of view, we want to be able to easily select different payload range
  - A success-payload-scatter-chart scatter plot from selected payload mass range
    - To visually observe how payload may be correlated with mission outcomes for selected site(s).
- https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/spacex_dash_app.py

Predictive Analysis (Classification)

Load data to pandas dataframe → Normalise the dataset for each columns using StandardScaler() → Splitting data to training and testing datasets → Create and train models using training dataset with GridSearchCV to find optimal parameters for each selected model

(Selected models are: logistical regression, decision tree, svm, k-nearest neighbours) → Evaluate each trained models on testing set → Compare the evaluation accuracy score to find best trained model

- [https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd392440274 1b70cb20a65440/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb](https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

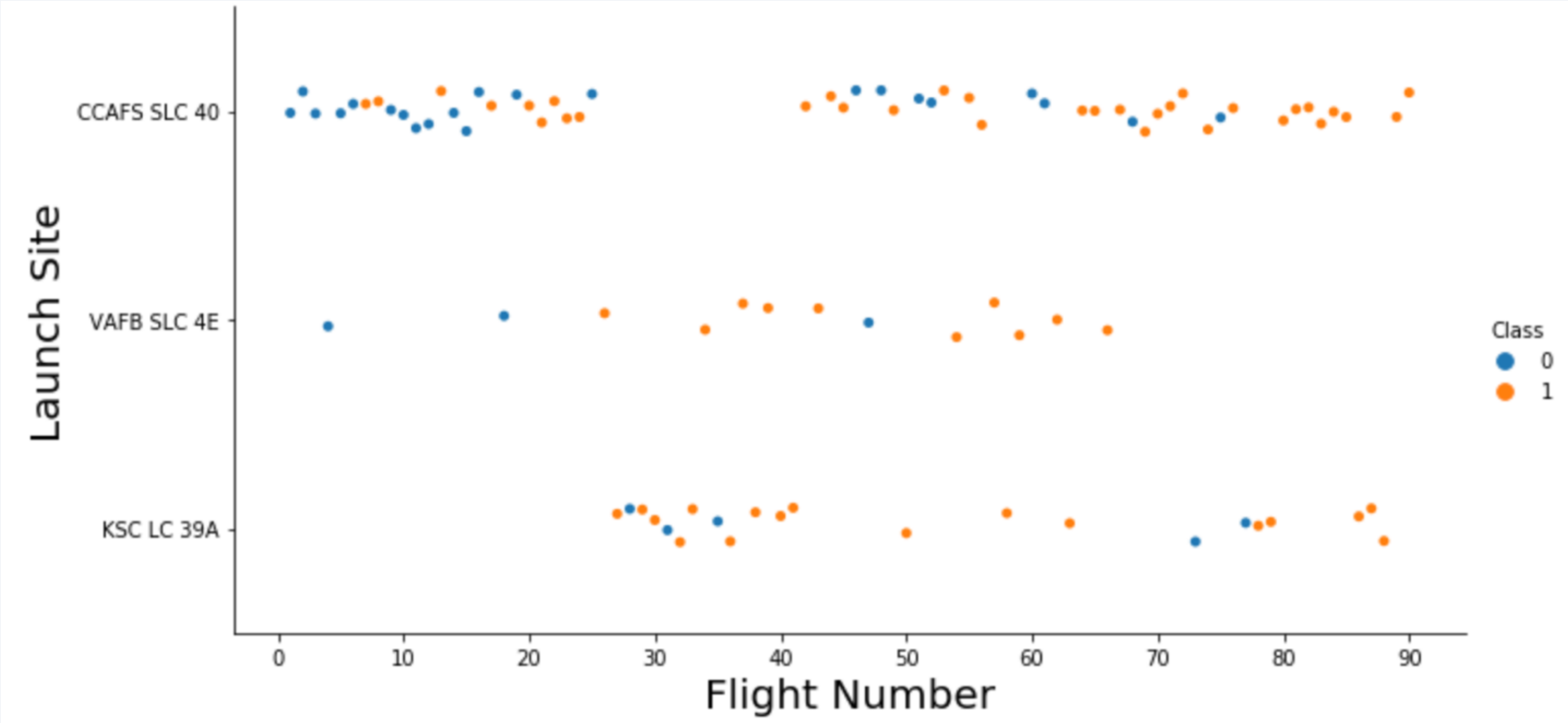
Results

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

Section 2

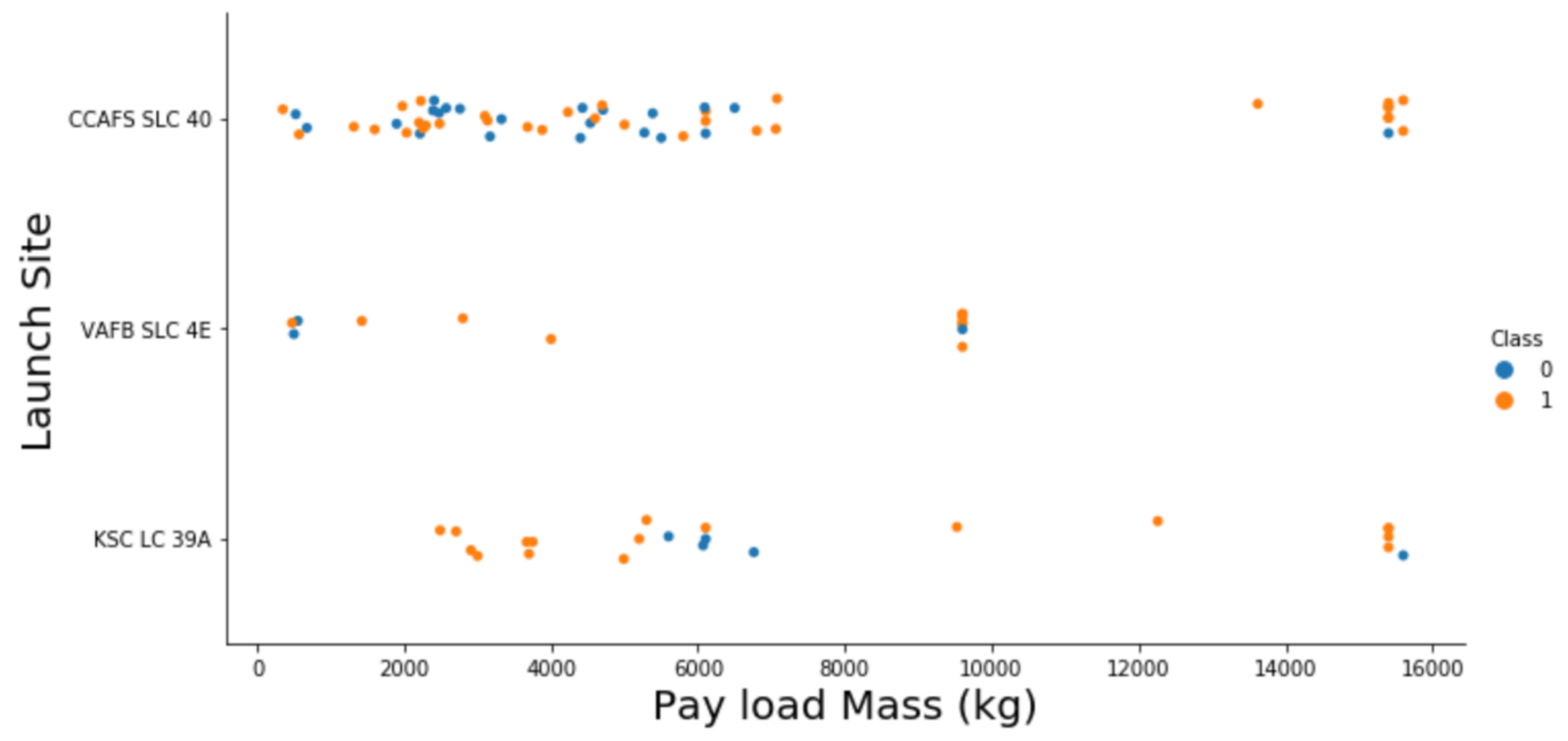
Insights drawn from EDA

Flight Number vs. Launch Site



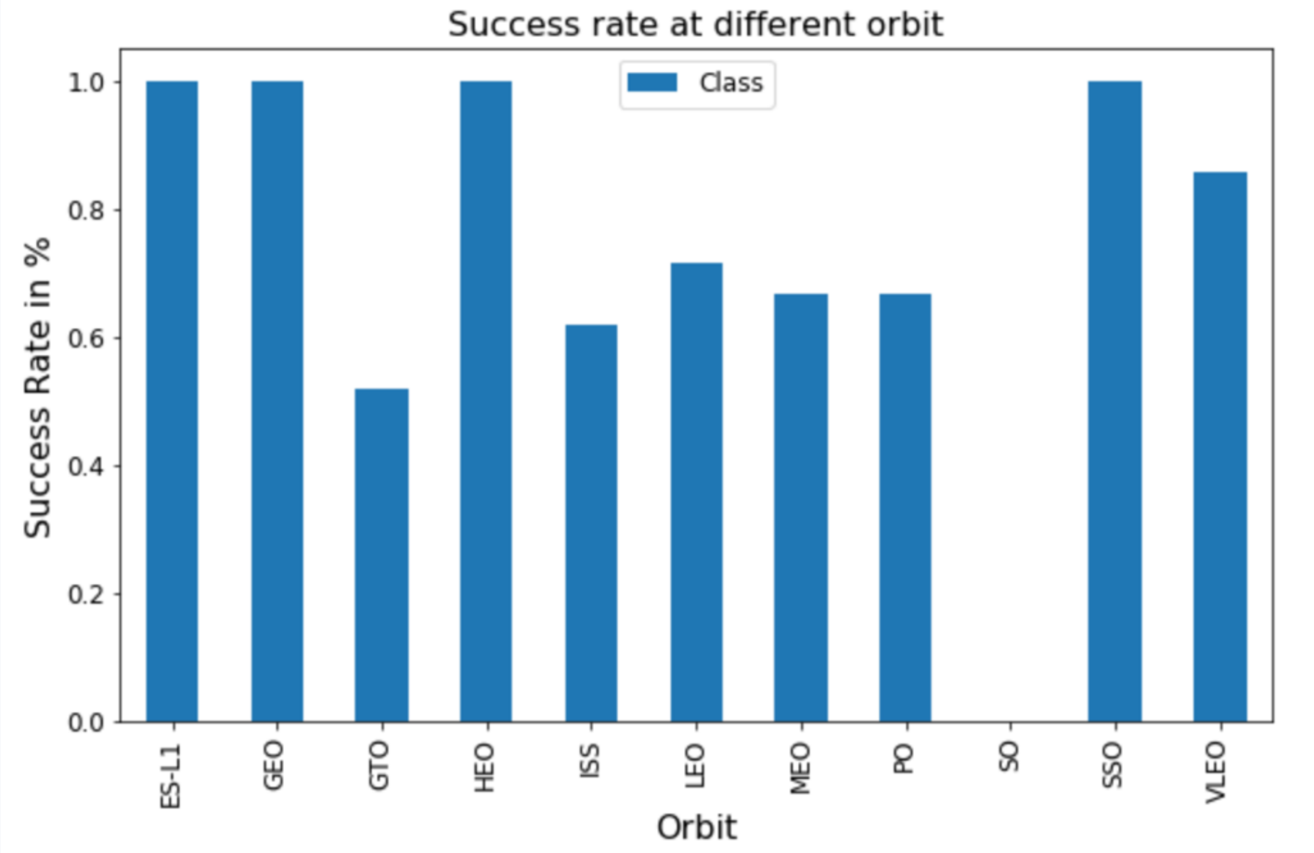
- Most number of flights were launched at CCAFS SLC 40
- Later flights (after number 78) were launched are likely to be successful

Payload vs. Launch Site



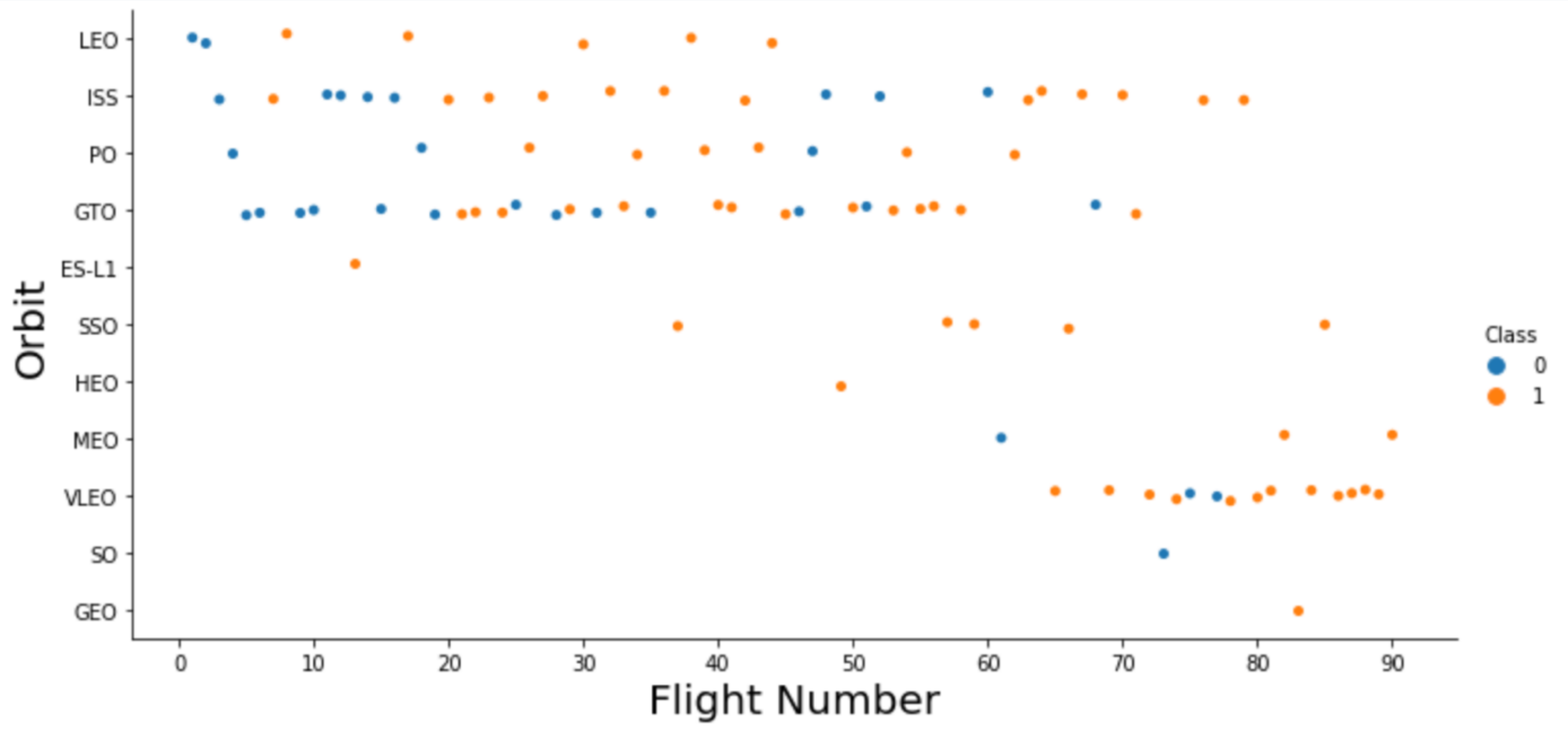
- No relationship between rockets and payload mass at CCAFS SLC 40
- No rockets launched for heavypayload mass(greater than 10000) at VAFB-SLC 4E

Success Rate vs. Orbit Type



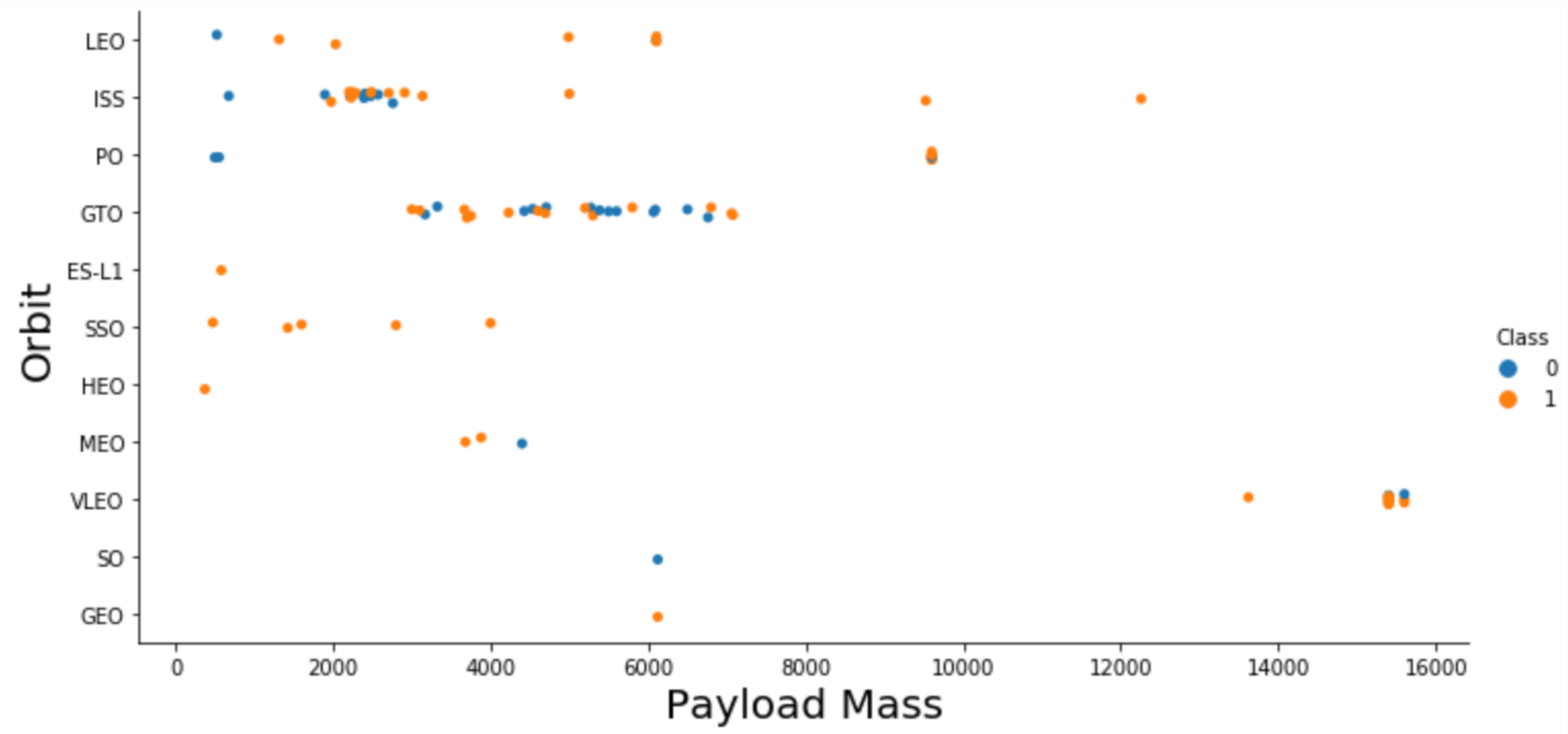
- 100% successful rate for flight launched to orbits at *ES-L1*, *GEO*, *HEO*, and *SSO*
- No successful launch to orbit *SO*

Flight Number vs. Orbit Type



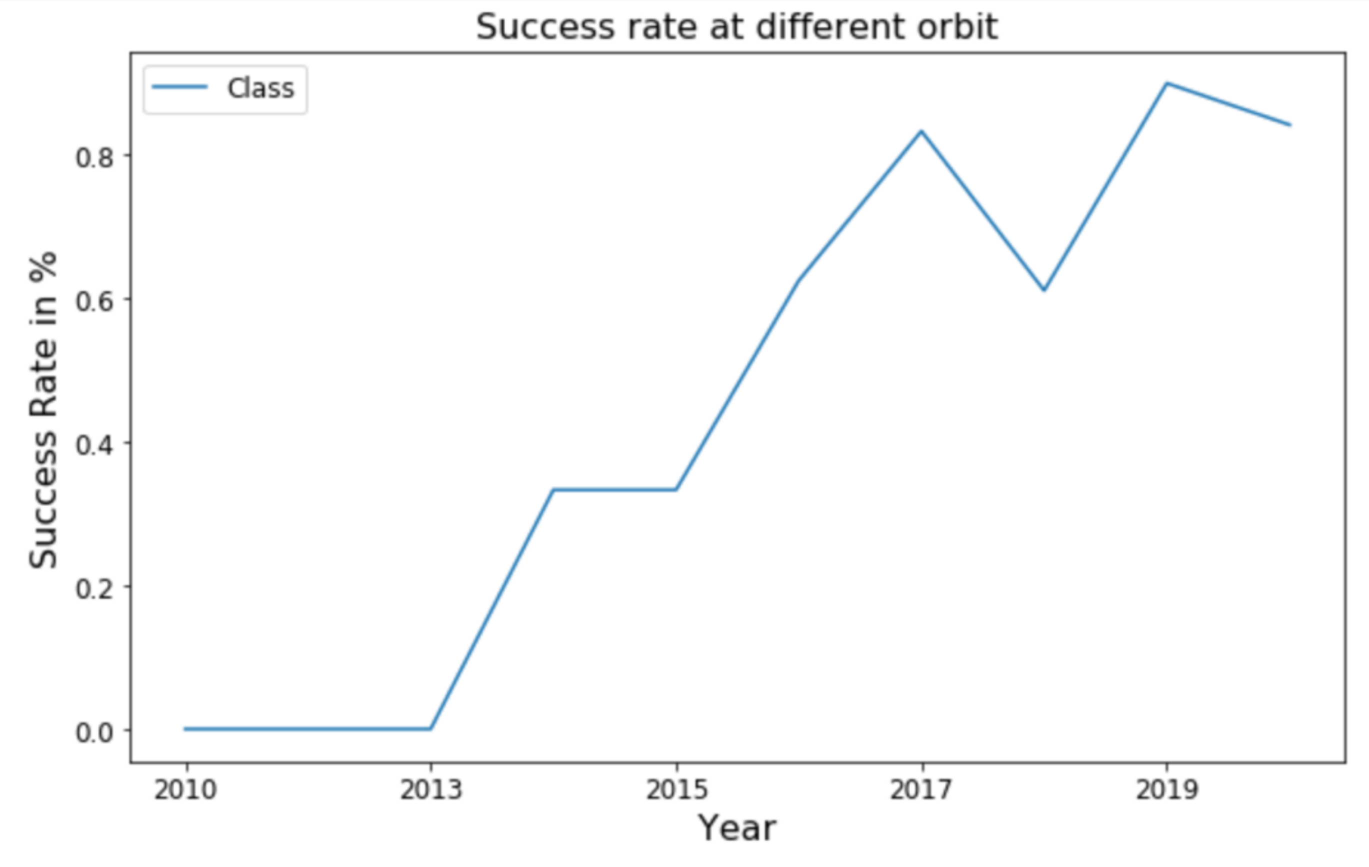
- There are only ONE launch recorded for *ES-L1*, *GEO*, *HEO*, and *SO* orbit types
- Only *SSO* has more than one launches with 100% successful rate

Payload vs. Orbit Type



- With heavy payloads the successful rate are more for *PO* ,*LEO* and *ISS*
- No impact from payload mass seems found on the launch to *GTO* orbit

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020, with a biggest drop over this period around 0.2% at 2018

All Launch Site Names

- There are FOUR unique launch site Names found
- However, there are only THREE unique launch sites, as CCAFS SLC-40 is previously CCAFS LC-40

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These 5 records found between 2010 to 2013, all with successful mission, but no successful landing
- All light payload mass less than 700 kg to Orbit LEO

Total Payload Mass

- Here list the top 10 customer from NASA by total payload mass, NASA (CRS) is the first

- NASA (CRS) launched highest total payload mass with over 45,000 kg

Customer	Total_payload_mass
NASA (CRS)	45596
NASA (CCDev)	12530
NASA (CCP)	12500
NASA (CCD)	12055
NASA (CTS)	12050
Iridium Communications GFZ , NASA	6460
NASA (CRS), Kacific 1	2617
NASA / NOAA / ESA / EUMETSAT	1192
U.S. Air Force NASA NOAA	570
NASA (LSP) NOAA CNES	553
NASA (COTS)	525
NASA (LSP)	362
NASA (COTS) NRO	0

Average Payload Mass by F9 v1.1

- Here list the top 10 *F9 v1.1* Booster Versions by average payload mass
- The booster version *F9 v1.1 B1016* launched the highest avg payload mass of 4707 kg
- *F9 v1.1* version launched an avg payload mass of 2928.4 kg

Booster_Version	avg_payload_mass
F9 v1.1 B1016	4707.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1014	4159.0
F9 v1.1	2928.4
F9 v1.1 B1012	2395.0
F9 v1.1 B1010	2216.0
F9 v1.1 B1018	1952.0
F9 v1.1 B1015	1898.0
F9 v1.1 B1013	570.0
F9 v1.1 B1017	553.0
F9 v1.1 B1003	500.0

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved is on 01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- FOUR boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- There are the total number of 100 successful and 1 failure mission outcomes

Boosters Carried Maximum Payload

- There are 12 difference versions of booster have carried the maximum payload mass

- All of them belongs to F9 B5 booster family

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

2015 Launch Records

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

- There are TWO failed landing_outcomes in drone ship in year 2015 Januray and April at launch site CCAFS LC-40
- Both launched flights had booster belongs to F9 v1.1 family

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

- There are EIGHT types of landing outcomes are found
- The most frequent type is *No attempt* and the frequency is up to 10 times, where
- Equal number of *Success (drone ship)* and *Failure (drone ship)* and ranked second on the list with frequency of 5 times

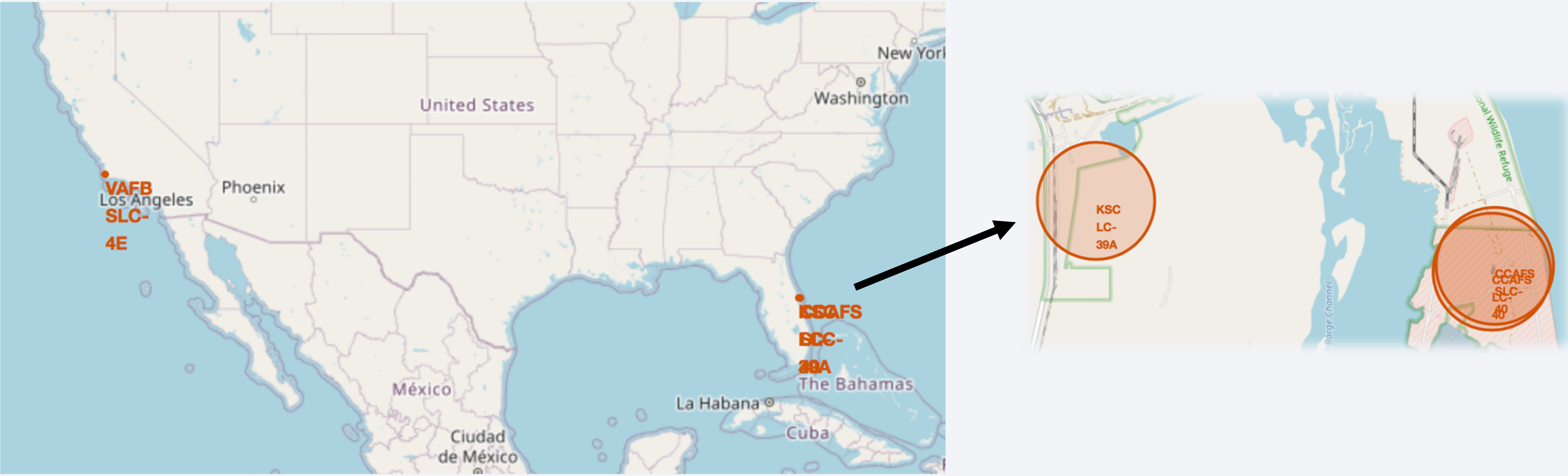
Landing _Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

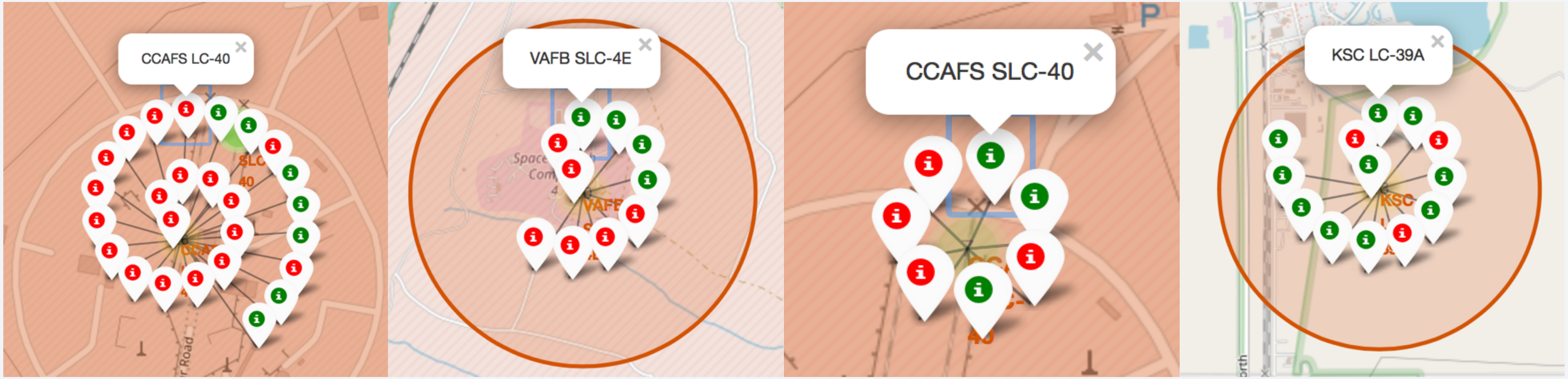
Launch Site Locations

- All Launch site are located along the coast, 3 of out 4 sites are closed to equator located on east coast



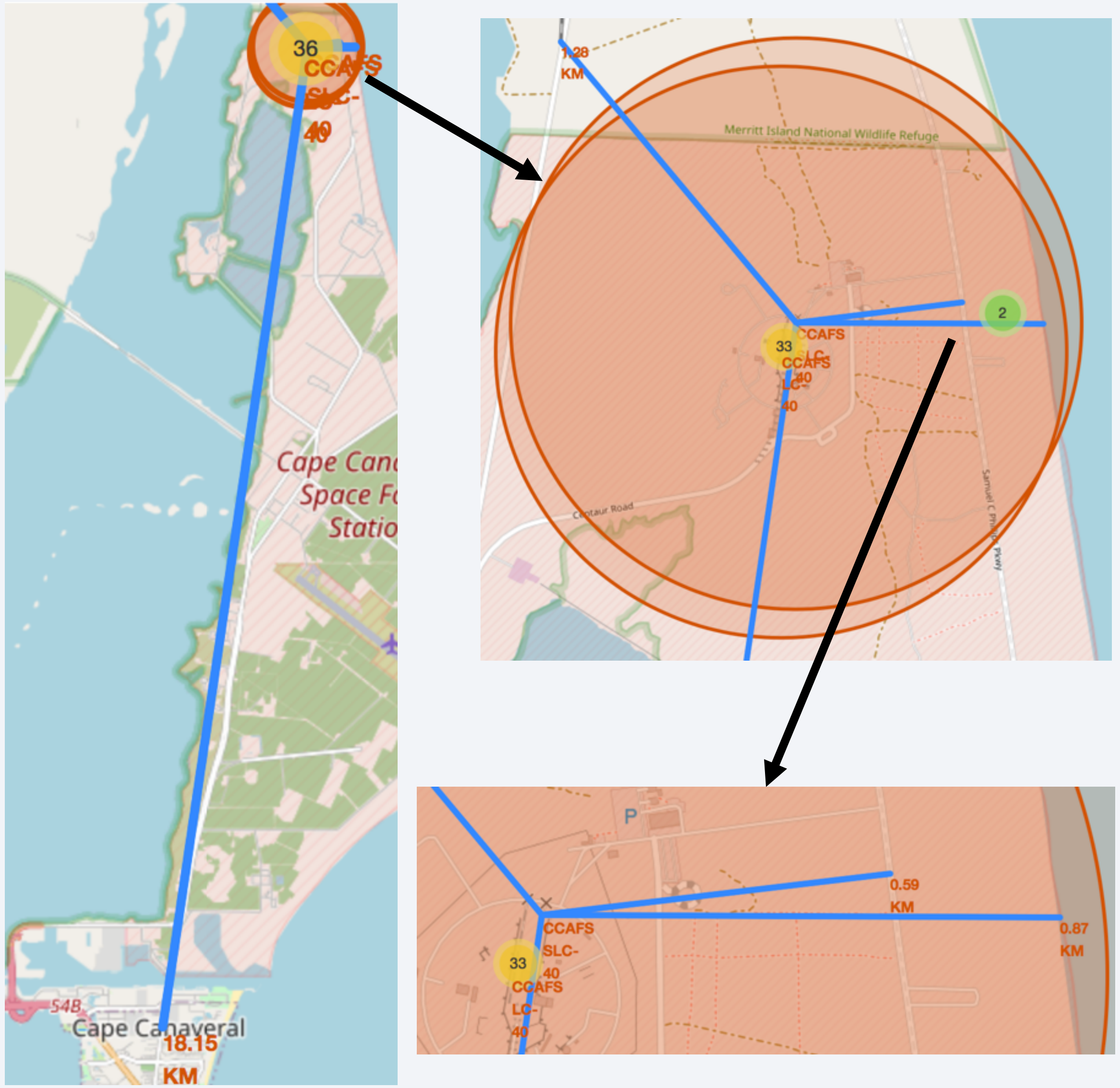
Launch Outcome for Each Site

- KSC LC-39A has highest launch successful rate



Launch Site – *CCAFS SLC-40* to its proximities

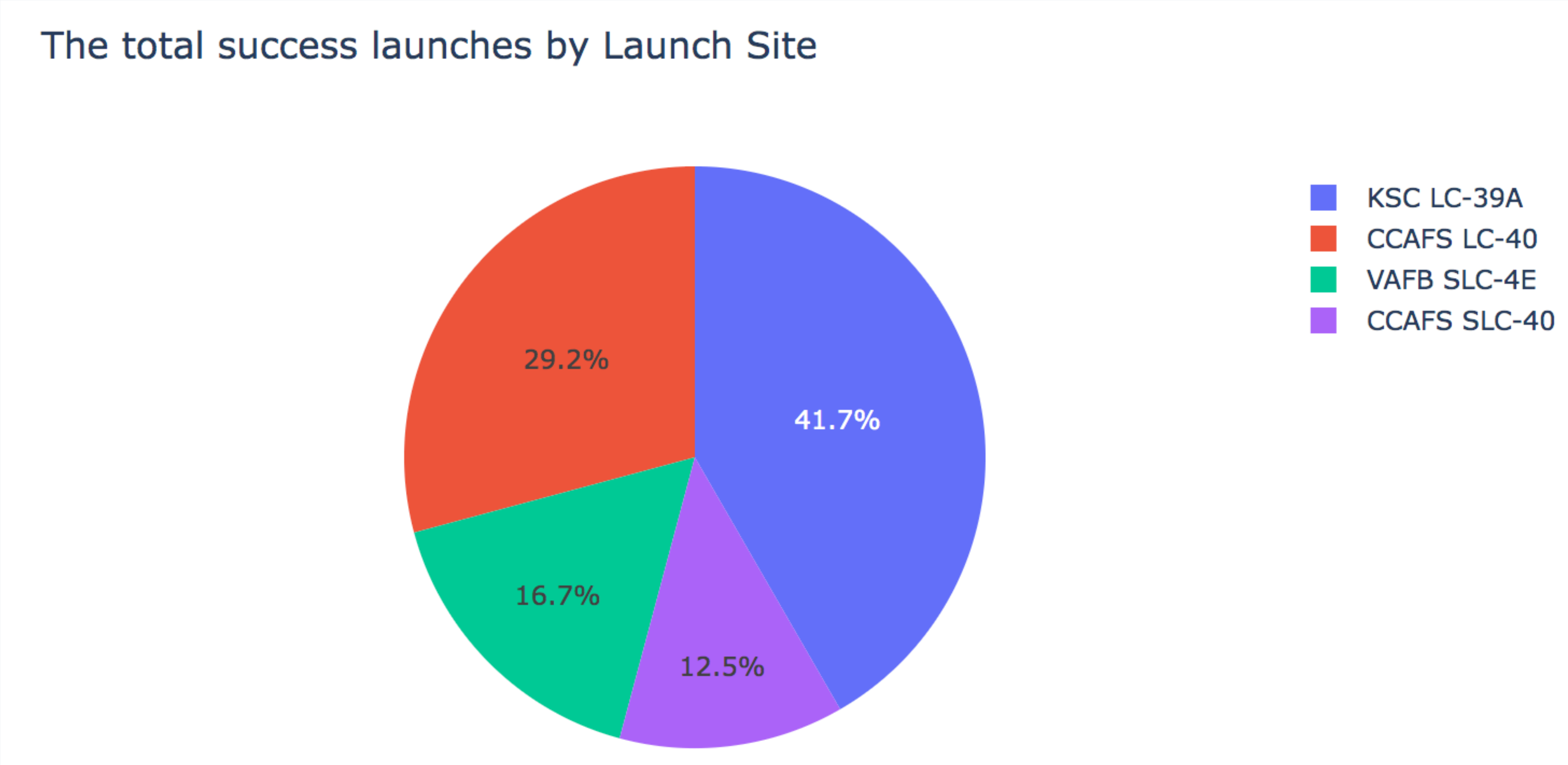
- The distance of the closest *Railway* to CCAFS SLC-40 is 1.28342 Km
- The distance of the closest *Highway* to *CCAFS SLC-40* is 0.58838 Km
- The distance of the closest *City* to CCAFS SLC-40 is 18.14695 Km



Section 4

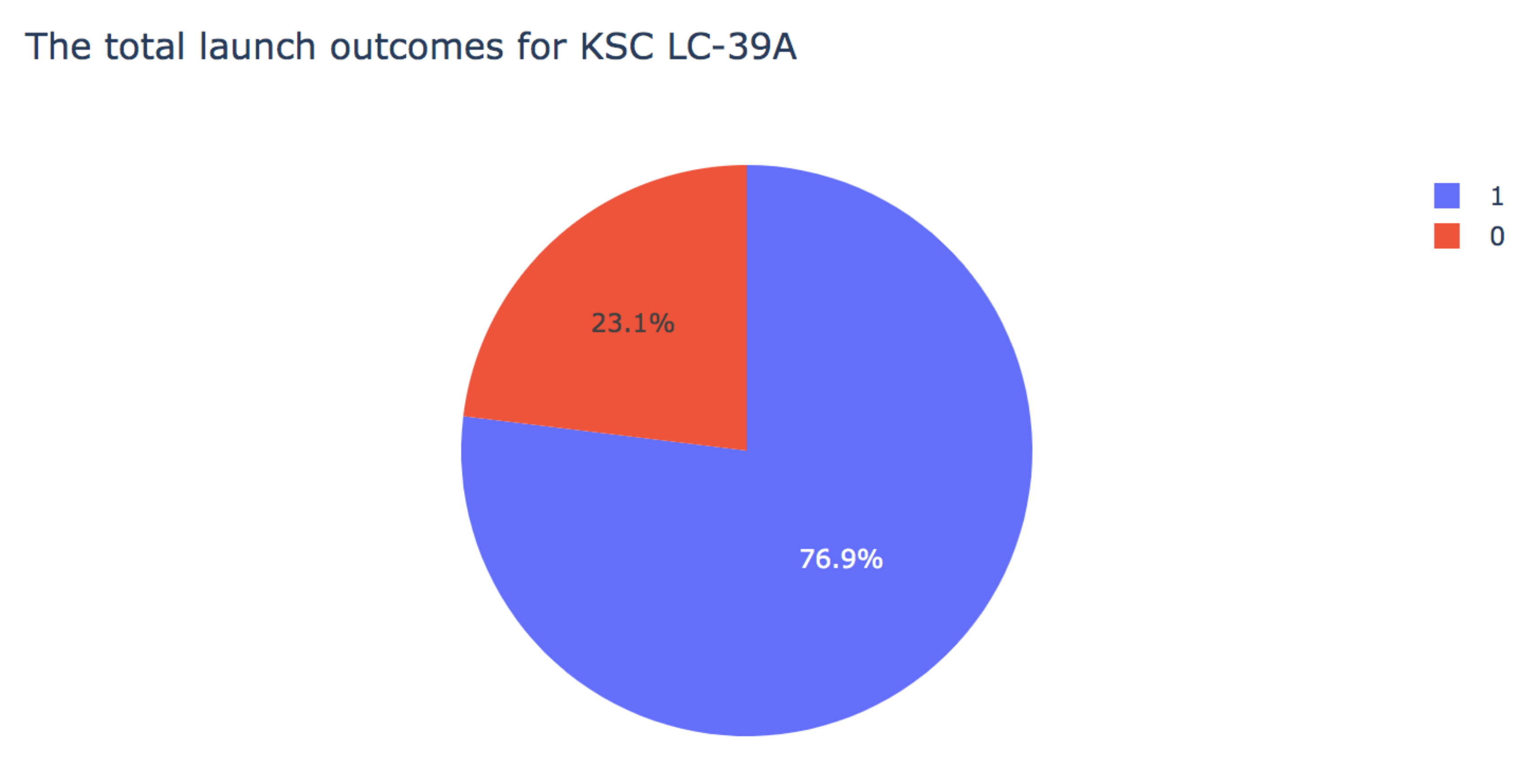
Build a Dashboard with Plotly Dash

Total Successful Launches Distribution by Launch Sites



- KSC LC-39A contributed the most to successful launches and up to 41.7%

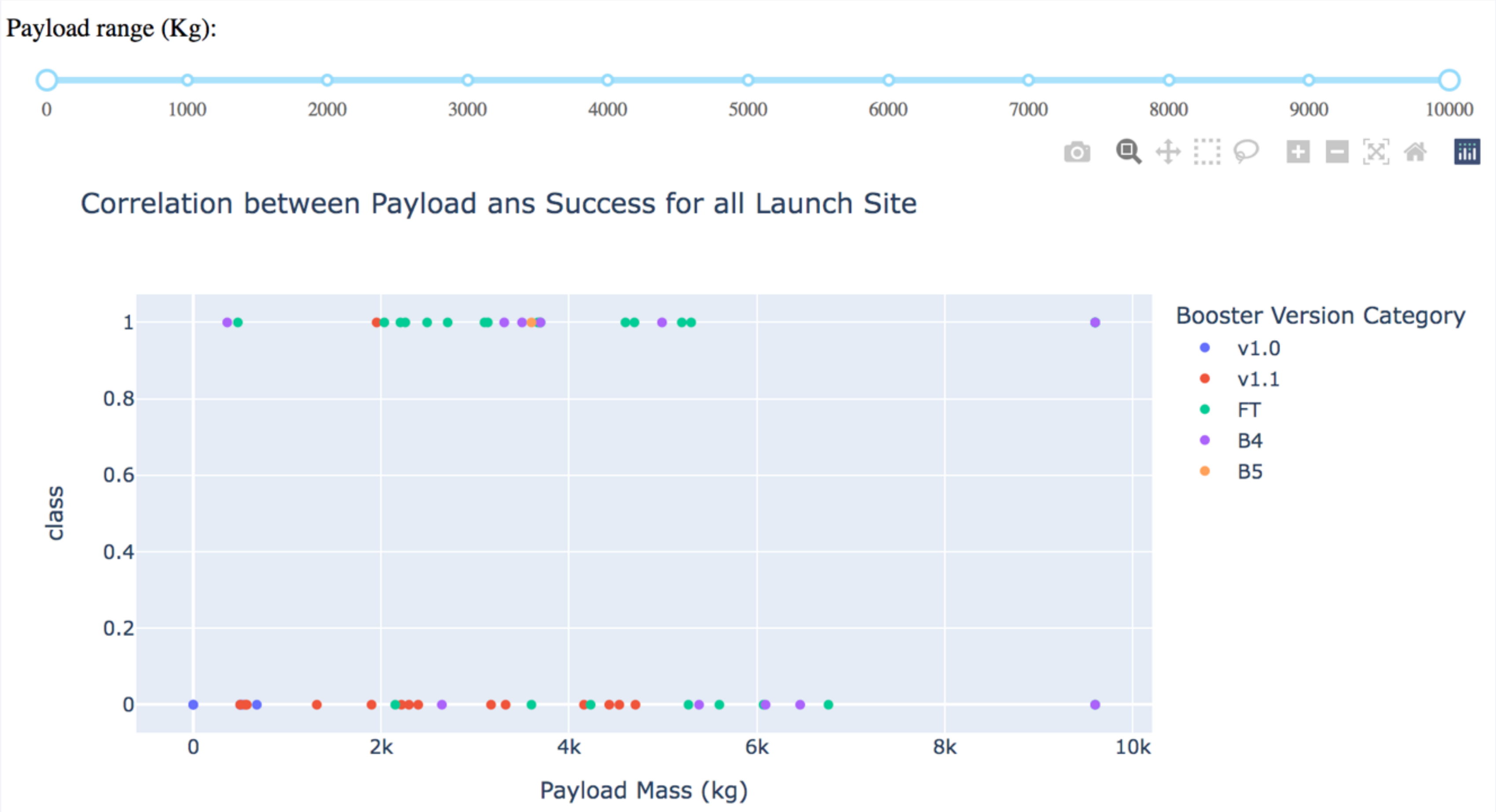
The Launch Site with Highest Launch Success Ratio



- KSC LC-39A has highest launch success ratio up to 76.9%

Highest and Lowest Success Rate by Payload Range

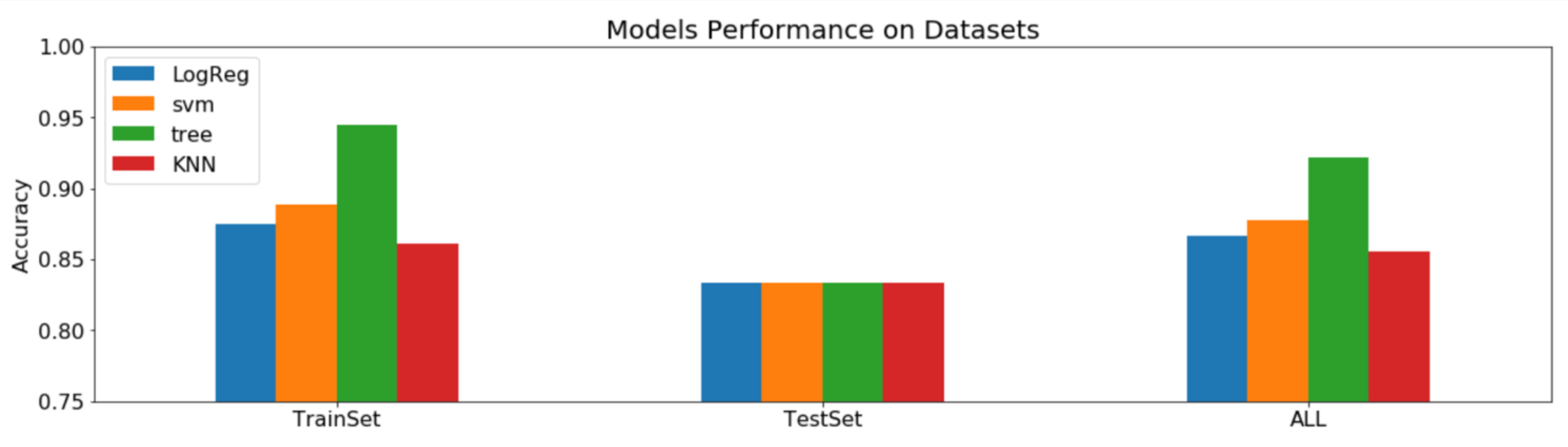
- <u>Highest</u> launch success find for payload range in between *3000 – 4000 kg*
- <u>No success</u> launch for payload in between *5500 – 7000 kg*
- *FT booster* seems have the <u>highest</u> launch success rate



Section 5

Predictive Analysis (Classification)

Classification Accuracy

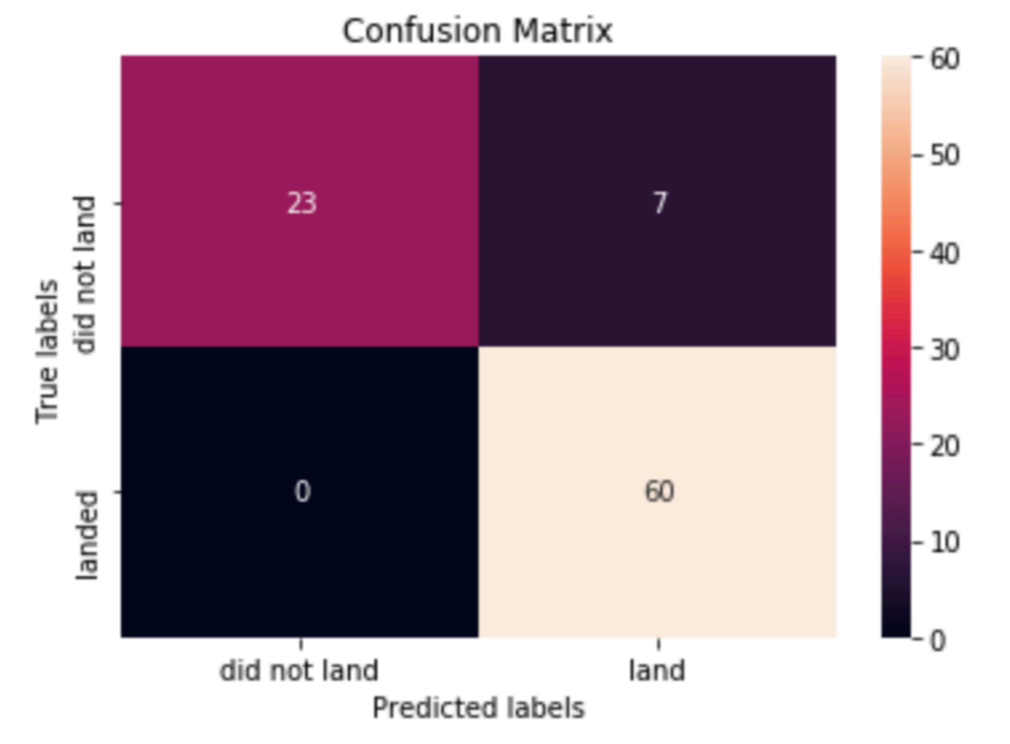


- Testing set show equal accuracy between all trained models
- However, Decision tree model has highest accuracy for both training set and whole dataset

Algorithm	TrainSet	TestSet	ALL
LogReg	0.875	0.833333	0.866667
svm	0.888889	0.833333	0.877778
tree	0.944444	0.833333	0.922222
KNN	0.861111	0.833333	0.855556

Confusion Matrix

- Here shows the confusion matrix of the overall dataset using trained decision tree model
- There are 7 false negative predictions
- The accuracy $= \frac{23+60}{23+7+60} = 92.2\%$
- The sensitivity is $\frac{60}{7+60} = 89.6\%$
- The specificity is $\frac{23}{23+0} = 100\%$



Conclusions

- Few important the rocket attributes related with success of first stage landing were identified
  - The attributes are: Launch Site, Orbit, Payload Mass, Booster_Version
  - Highest launch success find for payload mass in the range between **3000 – 4000 kg**
  - *F9 FT* Booster version seems have the highest launch success rate
  - The optimal location for rocket launch is at ***KSC LC-39A***
  - Rockets launched to **SSO** orbit has the highest successful rate
- Interactive plots are created using Folium and Dashboard
- Four machine learning models were trained and evaluated using the identified attributes to predict its first stage landing outcome
  - All trained model preformed equally on Testing Dataset, but Decision Tree shown the best performance up to 92.2% on the whole dataset

Appendix

Here is a summary of list contains all the source code for this project

- Data Collection API:
  https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/jupyter-labs-spacex-data-collection-api.ipynb
- Data Web Scraping:
  https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/jupyter-labs-webscraping.ipynb
- Data Wrangling:
  https://github.com/XiaoLIUau/Public_Folder/blob/2e5a487b47f295a3e9ce73eb7921b8e5f6c9d8ad/labs-jupyter-spacex-Data%20wrangling.ipynb
- EDA Visualisation:
  https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/jupyter-labs-eda-dataviz.ipynb
- EDA SQL:
  https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- Launch Site Location with Folium:
  https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/lab_jupyter_launch_site_location_folium.ipynb
- Interactive analytics Dashboard:
  https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/spacex_dash_app.py
- Machine Learning prediction:
  https://github.com/XiaoLIUau/Public_Folder/blob/455539ea817f34cbd3924402741b70cb20a65440/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!