IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Phuong Nguyen Ngoc Cuc>
<23/07/2022>

Outline

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

Executive Summary

- Summary of methodologies
  - ➢ Data collection
  - ➢ Data wrangling
  - ➢ EDA with SQL
  - ➢ EDA with data visualization
  - ➢ Create interactive map with Folium
  - ➢ Create a Dashboard with Plotly
  - ➢ Classification
- Summary of all results
  - ➢ EDA result
  - ➢ Interactive visual analytics
  - ➢ Predictive analysis

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers

  Predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant 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
  - LR, KNN, SVM, DT models have been built and evaluated for the best classifier

Data Collection

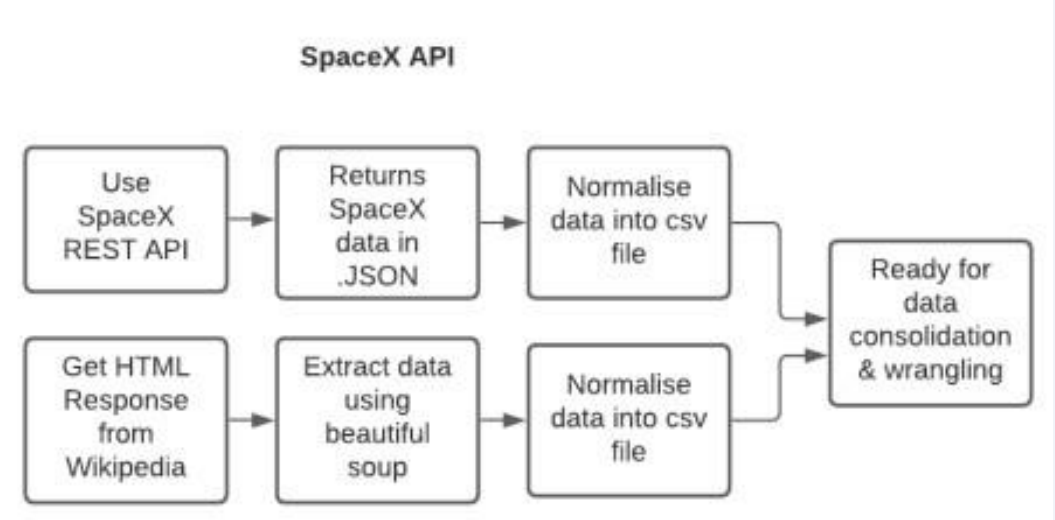
- Describe how data sets were collected.

➢ SpaceX launch data is gathered from the SpaceX REST API.

➢ This gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

➢ The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

➢ Another data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.

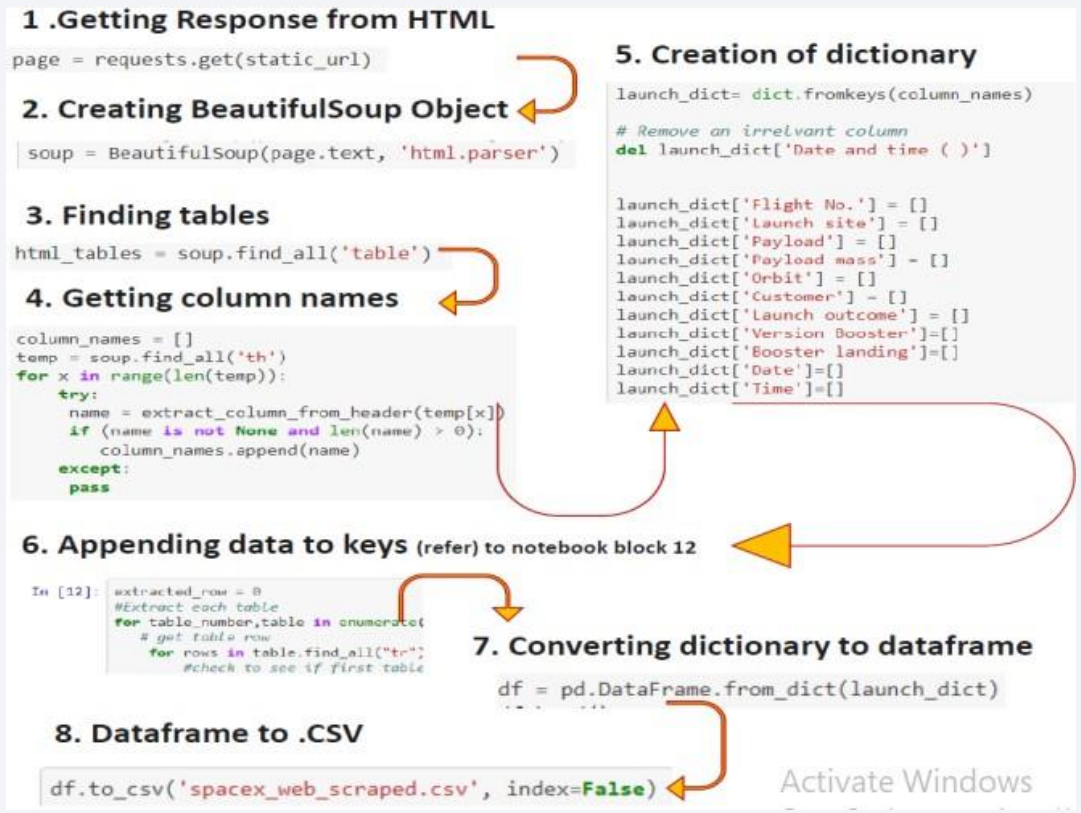


Data Collection – SpaceX API

- https://github.com/cucphuong/IBM_Data_science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

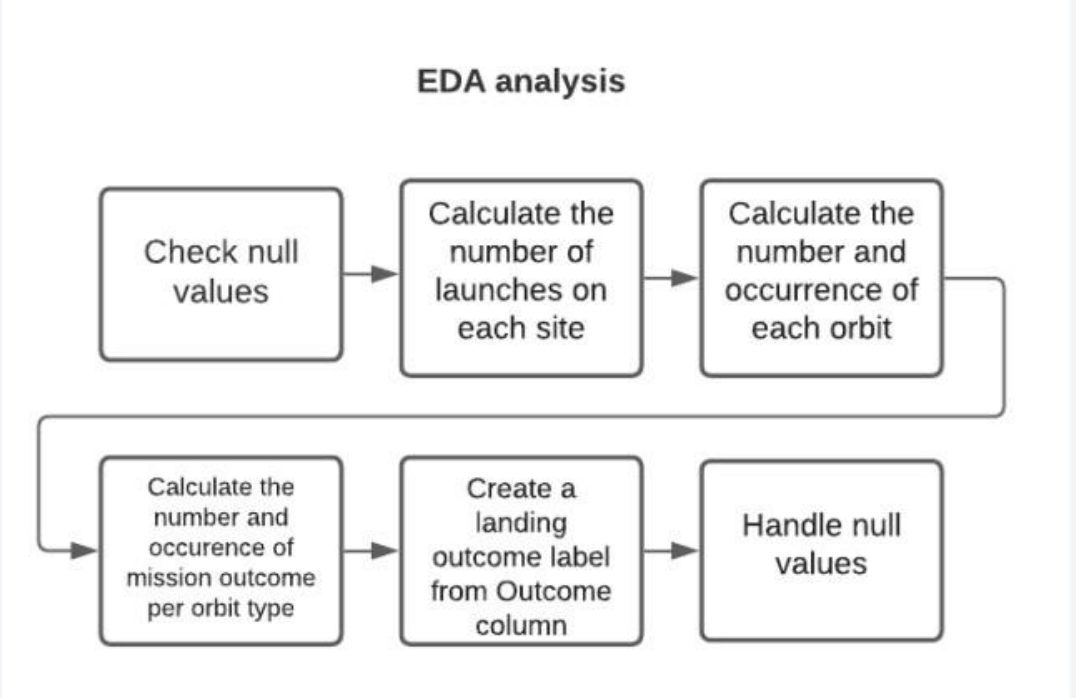
Data Collection - Scraping

- https://github.com/cucphuong/IBM_Data_science/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- https://github.com/cucphuong/IBM_Data_science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter charts are used to explore the relationship between two attributes
- Bar chart is used to visualize the relationship between success rate of each orbit type
- Line chart is used to visualize the launch success yearly trend

- https://github.com/cucphuong/IBM_Data_science/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

- Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'KSC'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster
  - versions, launch_site for the months in year 2017
  - Ranking the count of successful landing_outcomes between the date 2010 06 04 and 2017 03 20 in descendingorder.

- https://github.com/cucphuong/IBM_Data_science/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- I added markers, circles to locate the site's location of all launch sites and easily detect which sites have high success rates by eyes.
- https://github.com/cucphuong/IBM_Data_science/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- I added pie chart with site-dropdown to visualize the percent of launch success at each launch site.
- I also added a scatter chart with a range slider for selecting Payload to determine if the variable payload is correlated to the launch outcome.

- https://github.com/cucphuong/IBM_Data_science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

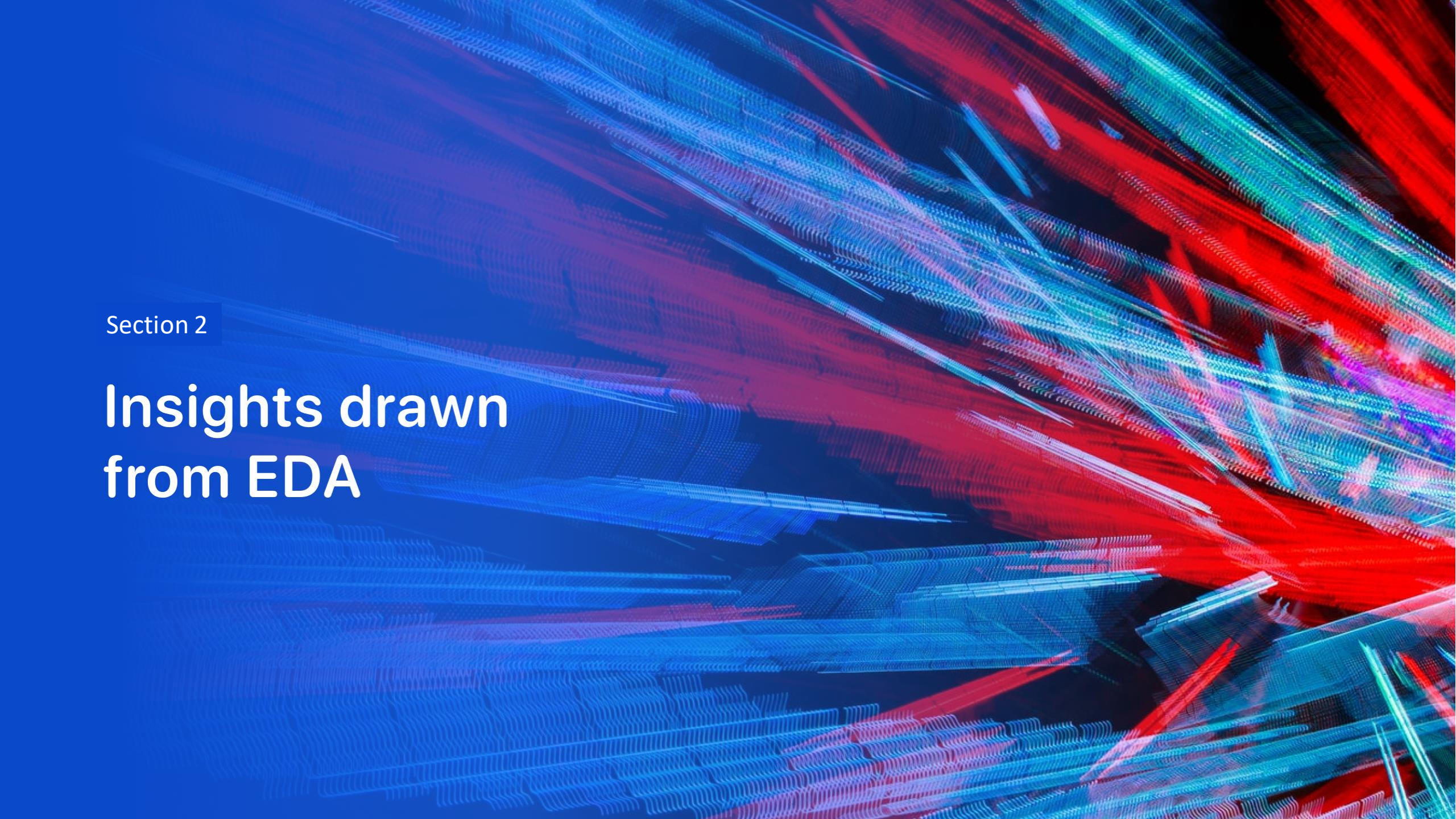
- SVM, KNN, and logistic regression gained the highest accuracy at 83.3%, while SVM performs the best in terms of Area under the curve at 0.958.
- Firstly, we standardized the data by the function StandardScaler()
- Then, split the dataset to train set (80%) and test set (20%)
- Use grid search to choose the best parameter values for logistic regression, support vector machine, k nearest neighbors and decision tree
- Finally, we used the confusion matrix, F1 score, Precision and Accuracy to evaluate the models and find the best model among them.
- https://github.com/cucphuong/IBM_Data_science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- SVM, KNN, and logistic regression gained the best result in terms of prediction accuracy for this dataset.
  - Low weighted payloads perform better than the heavier payloads.
  - The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
  - KSC LC 39A had the most successful launches from all the sites.
  - Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Section 2

Insights drawn from EDA

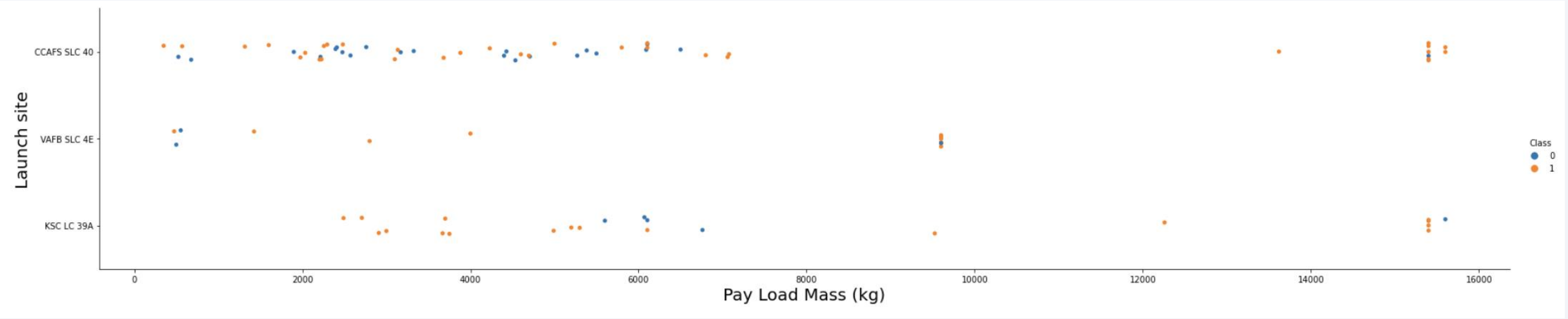


Flight Number vs. Launch Site

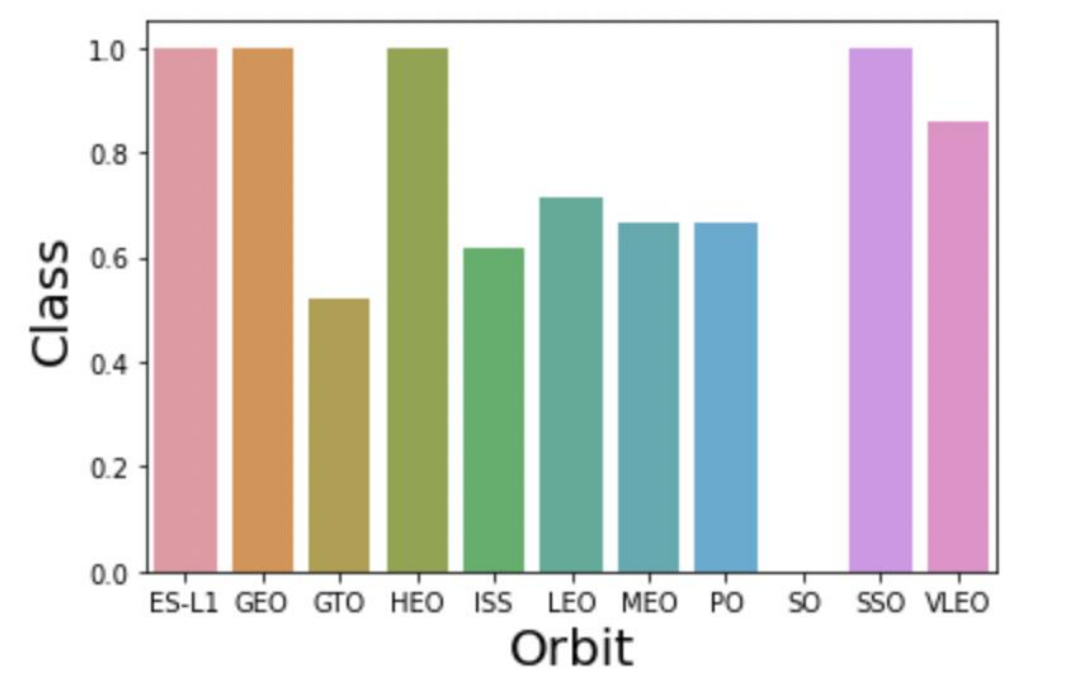
- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.

Payload vs. Launch Site

- The majority of lPay Loads with lower Mass have been launched from CCAFS SLC 40.



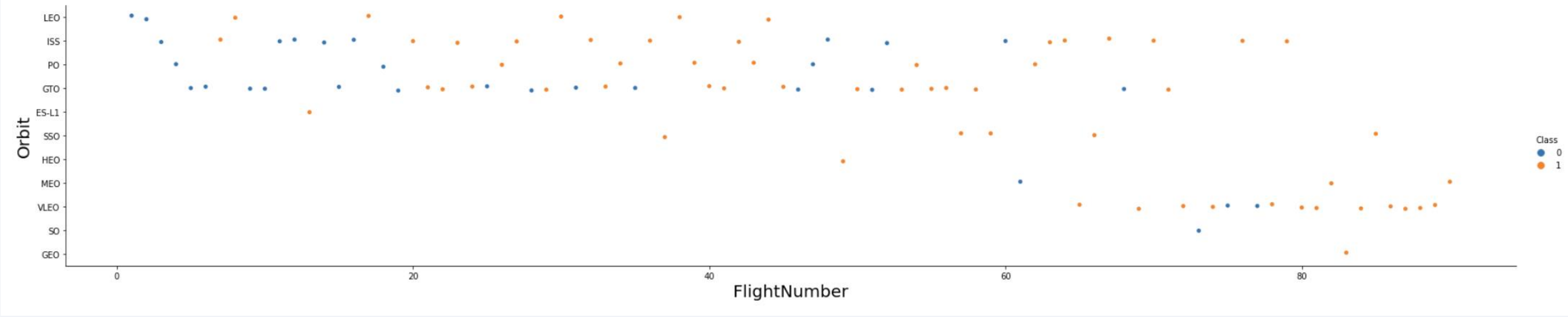
Success Rate vs. Orbit Type



Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

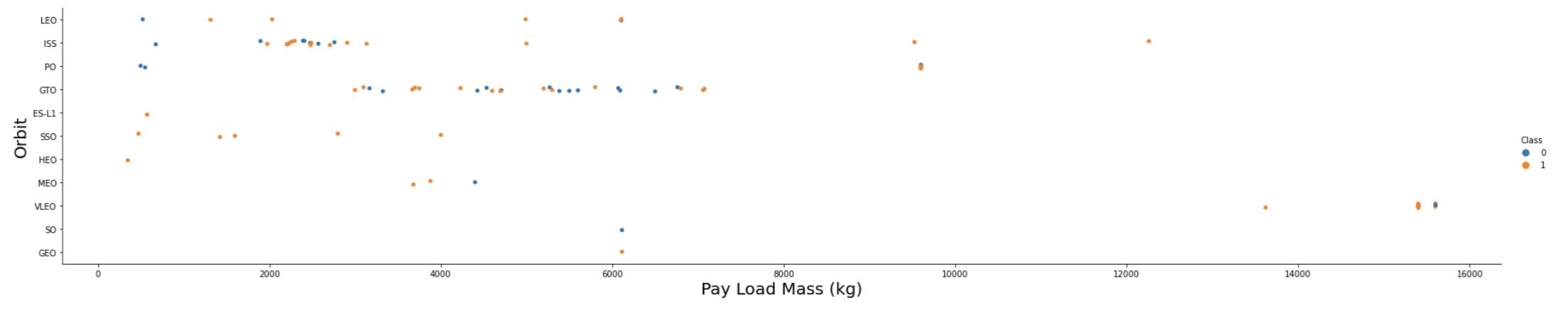
Flight Number vs. Orbit Type

- SA trend can be observed of shifting to VLEO launches in recent years.



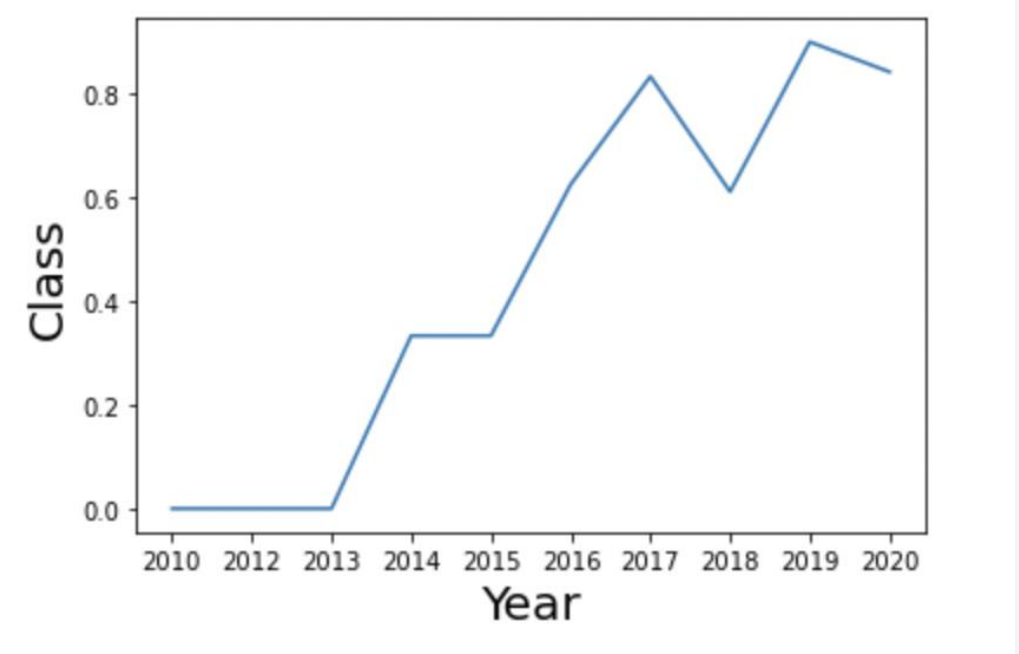
Payload vs. Orbit Type

- There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range of 4000-8000.



Launch Success Yearly Trend

- Launch success rate has increased significantly since 2013 and has stablised since 2019, potentially due to advance in technology and lessons learned.



All Launch Site Names

- %%sql select distinct LAUNCH_SITE from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- %%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5

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

Total Payload Mass

- %%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA(CRS)'

Total_payload_mass
45596

Average Payload Mass by F9 v1.1

- %%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL

where BOOSTER_VERSION = 'F9 v1.1'

aver
2534.6666666666665

First Successful Ground Landing Date

- %%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'

2015-12-22

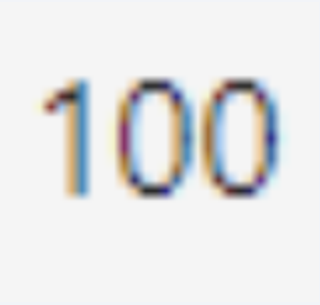
Successful Drone Ship Landing with Payload between 4000 and 6000

- %%sql select BOOSTER_VERSION from SPACEXTBL 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

Total Number of Successful and Failure Mission Outcomes

- %%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'



Boosters Carried Maximum Payload

- %%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

- %%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between ‘2015-01-01' and ‘2015-12-31') order by date desc

14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)

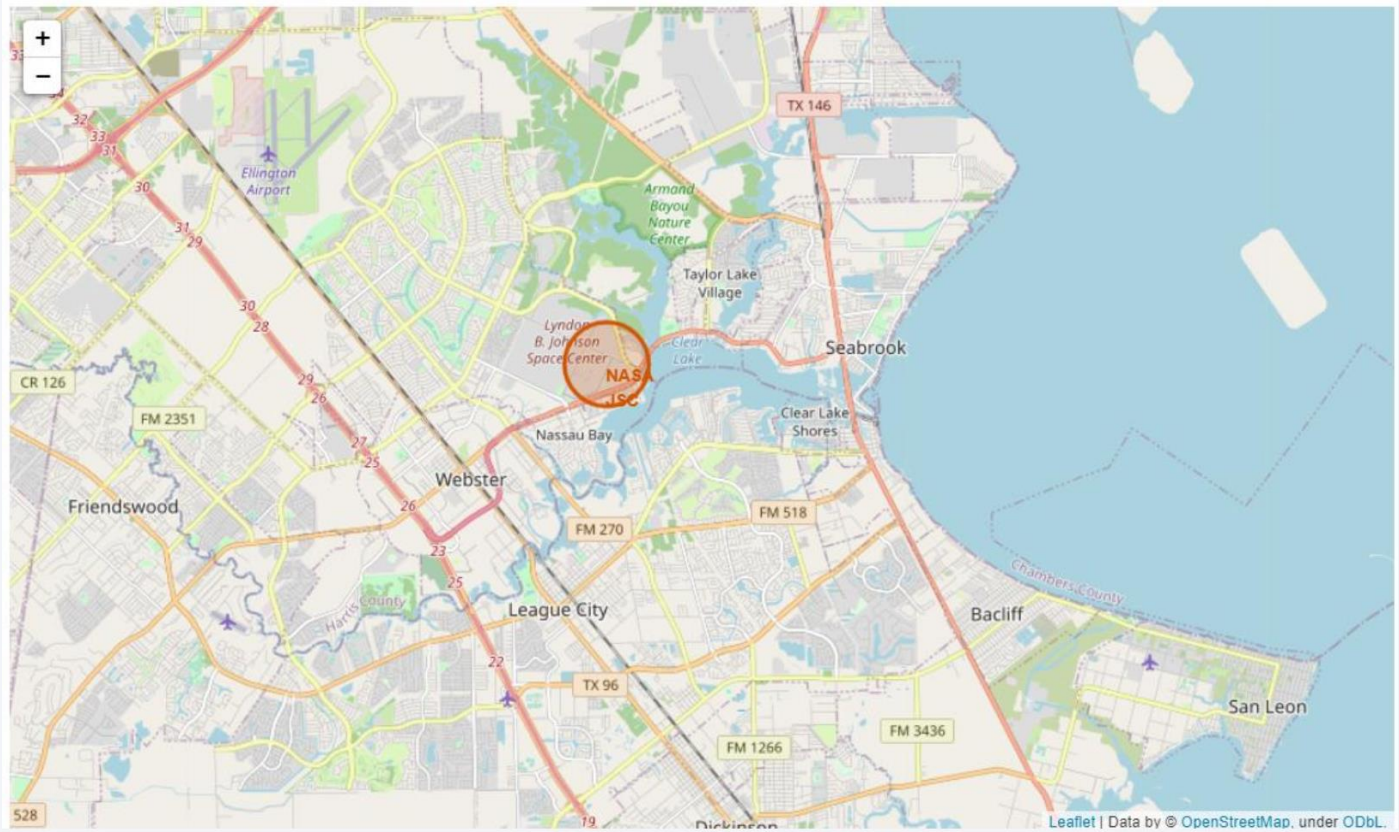
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- %%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc

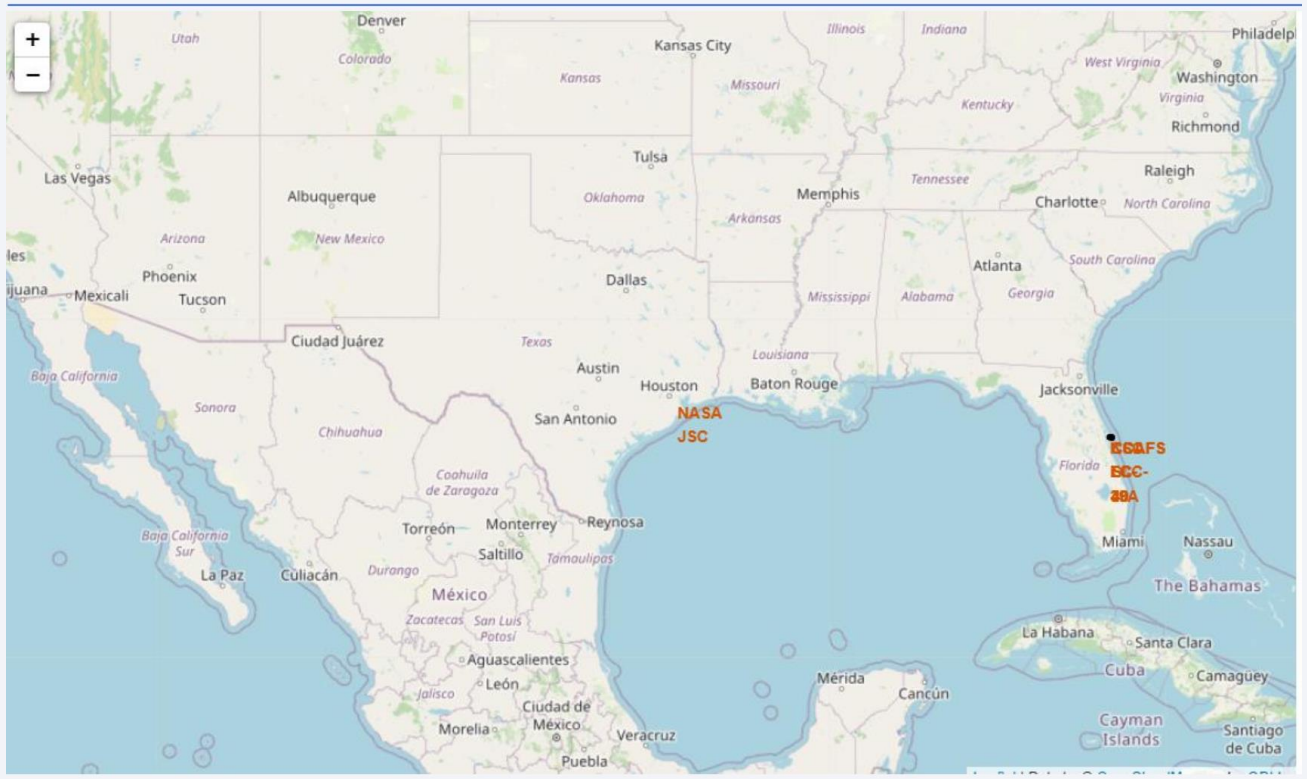
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)


Section 3
Launch Sites
Proximities Analysis

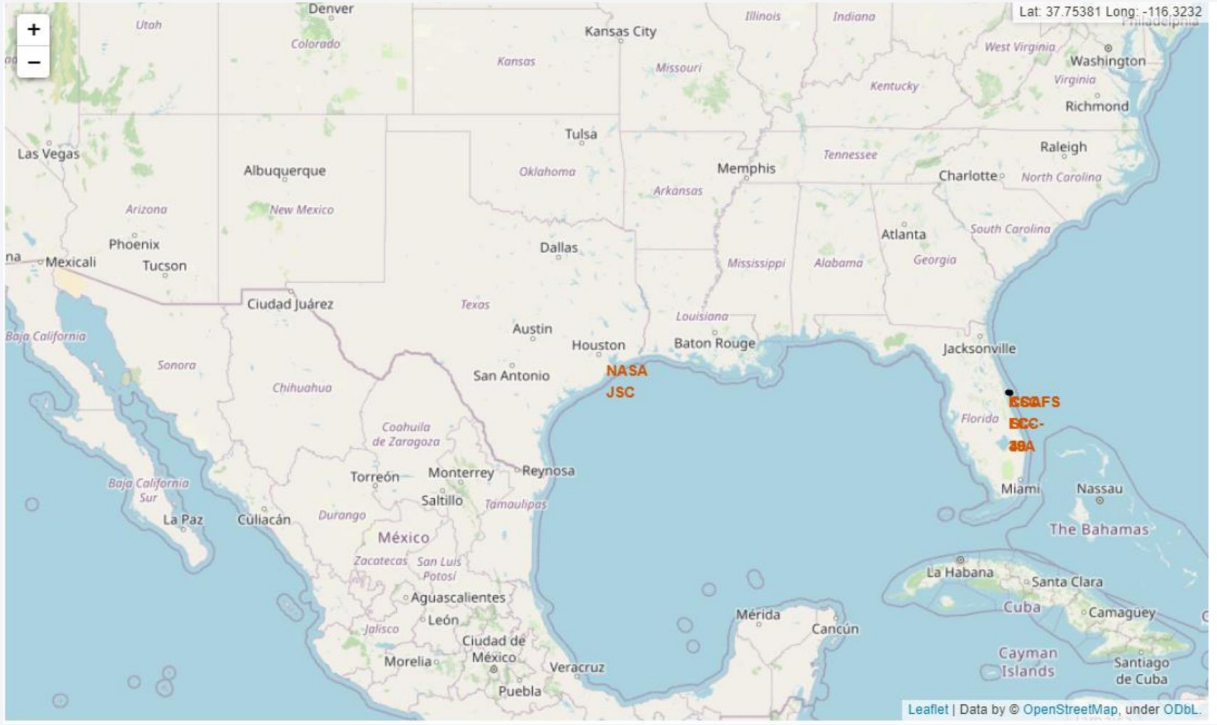
All launches marked on the map

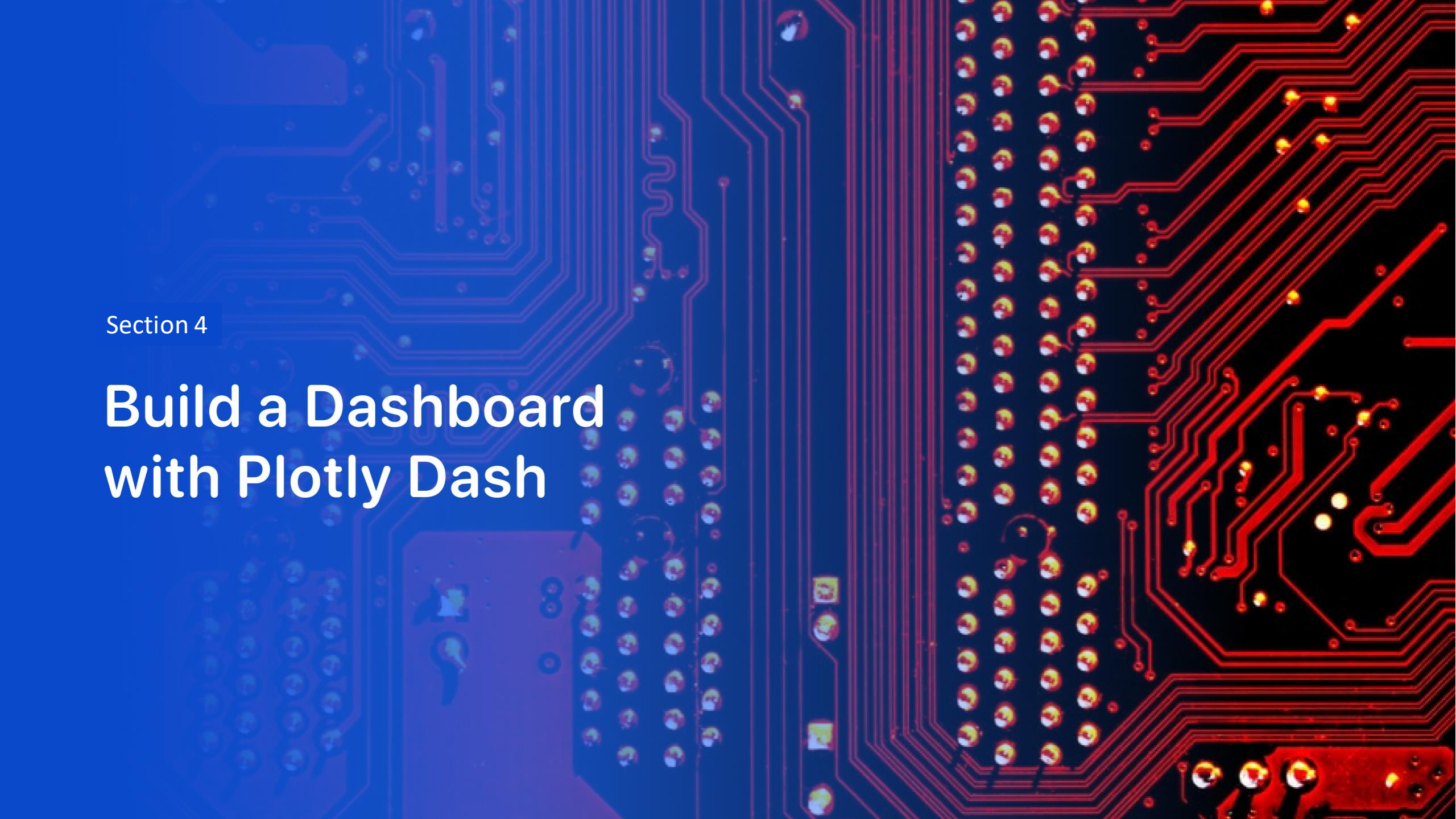


Success or fail launches marked on map

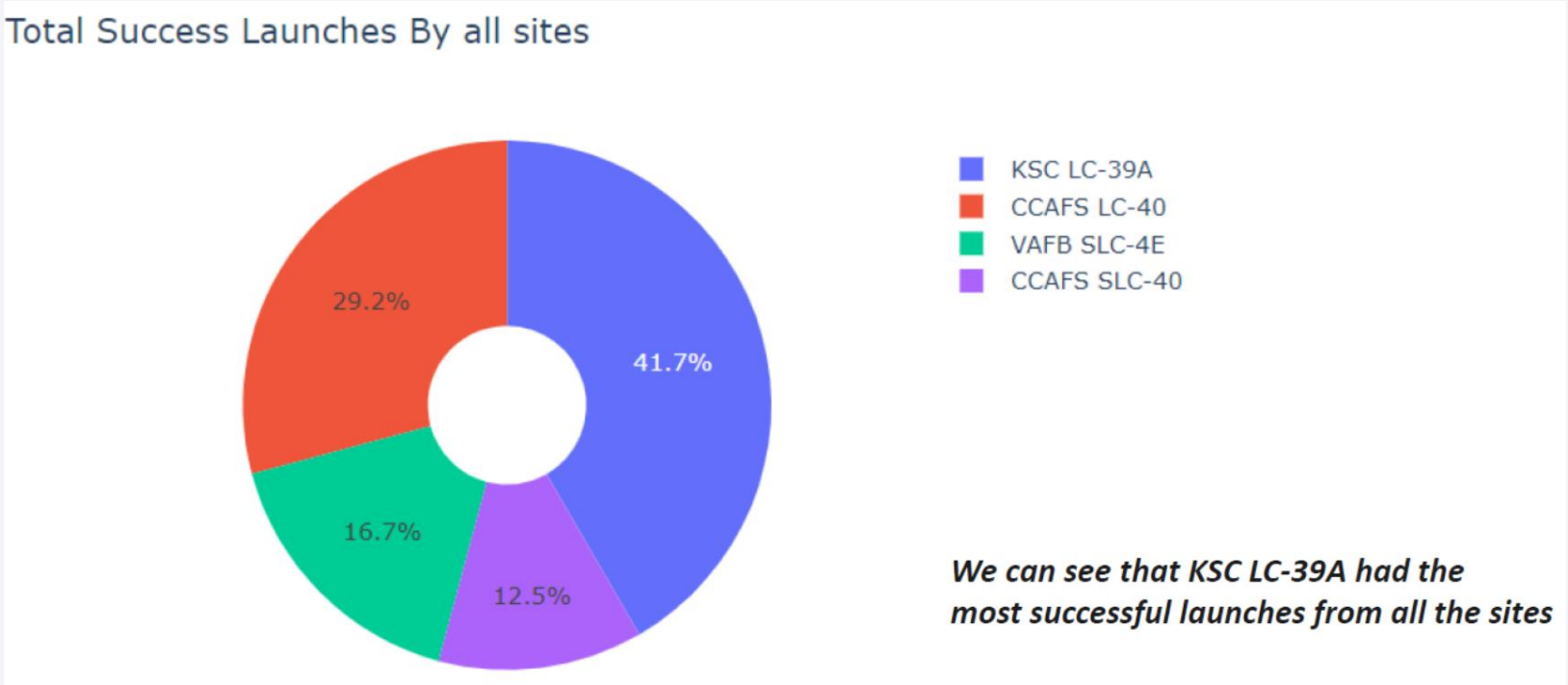


Distance from a launch site to its proximities

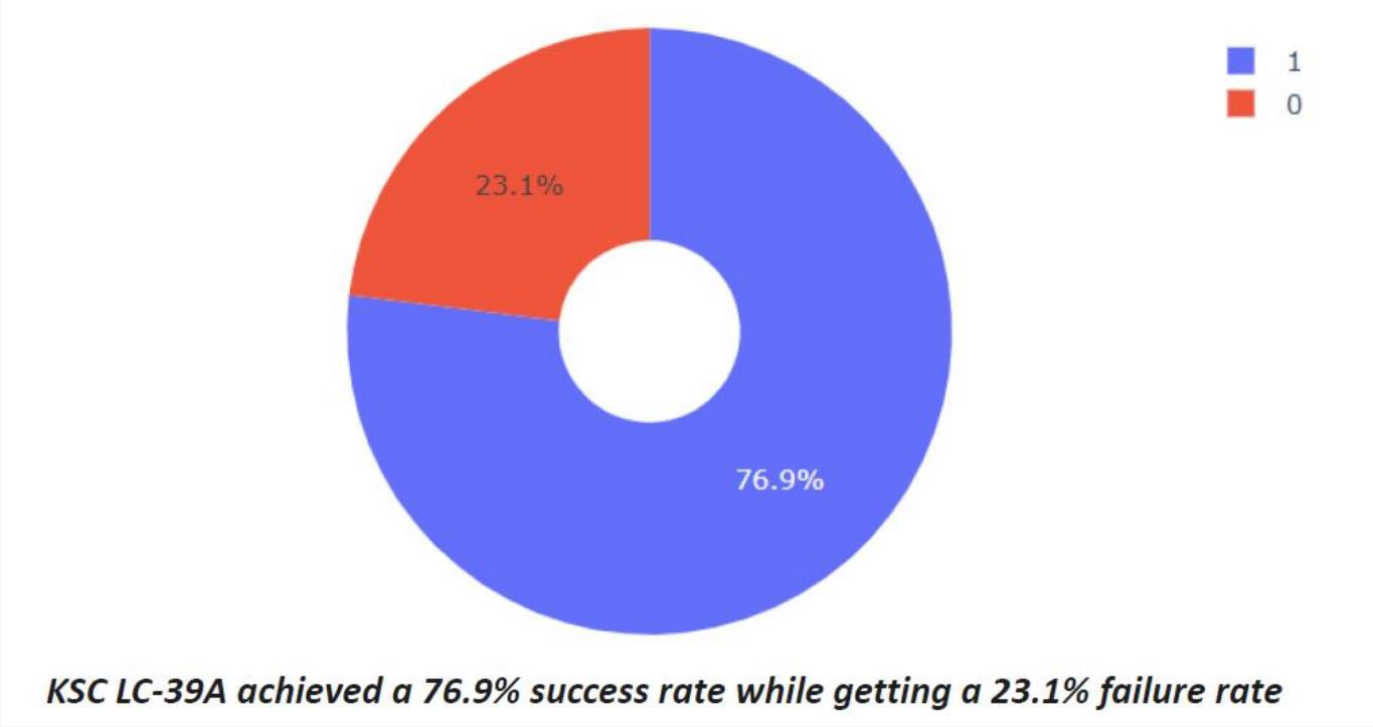



Section 4
Build a Dashboard with Plotly Dash

Total successful launches by all sites

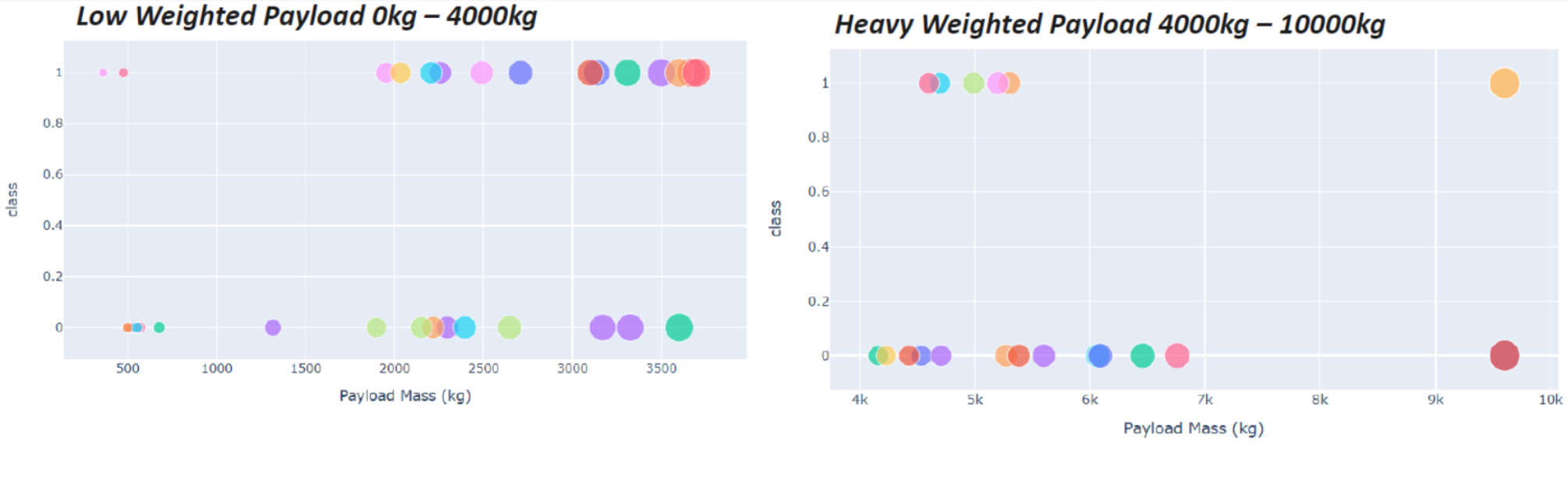


Successful rate by sites



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

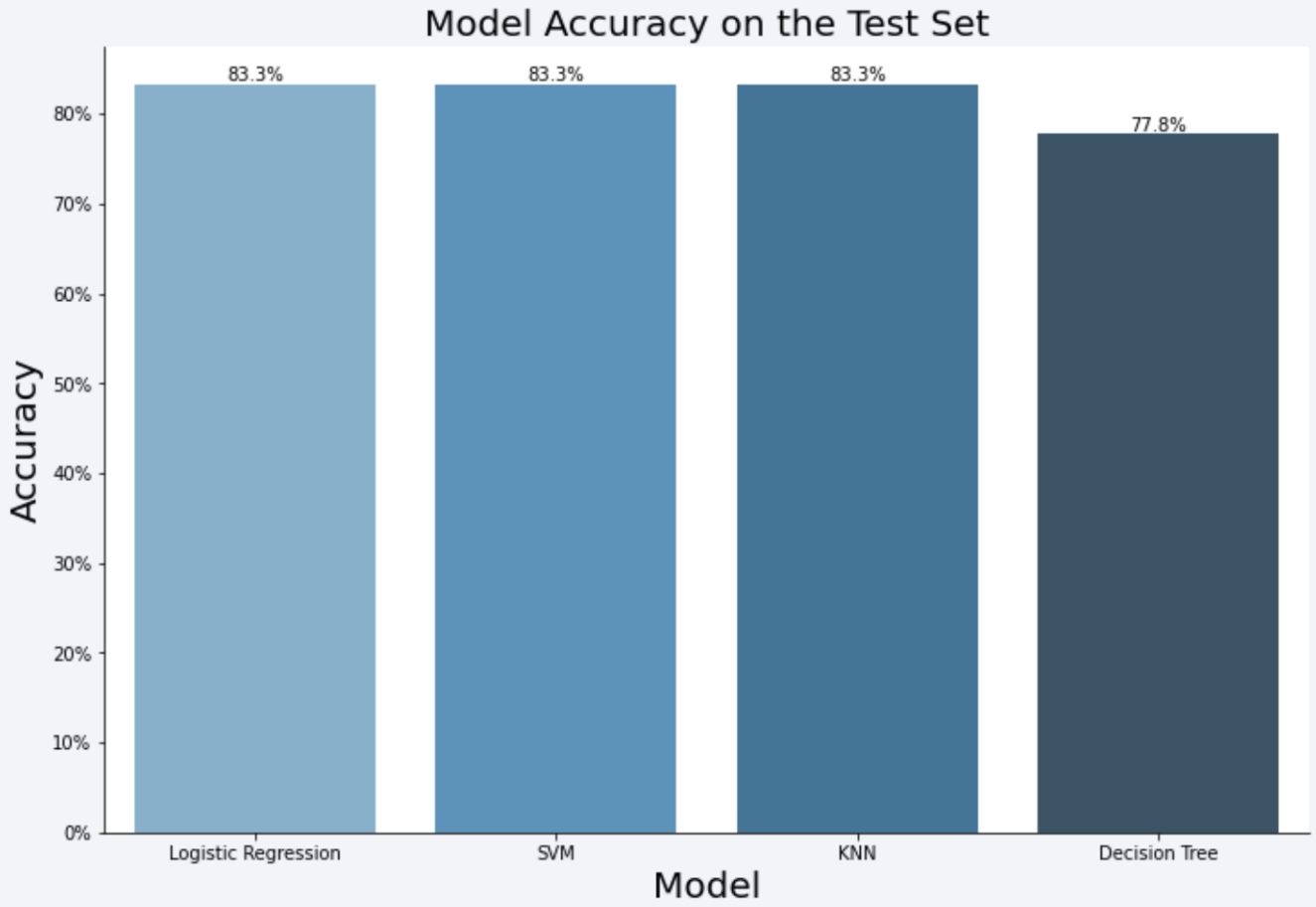
Payload and launch outcome



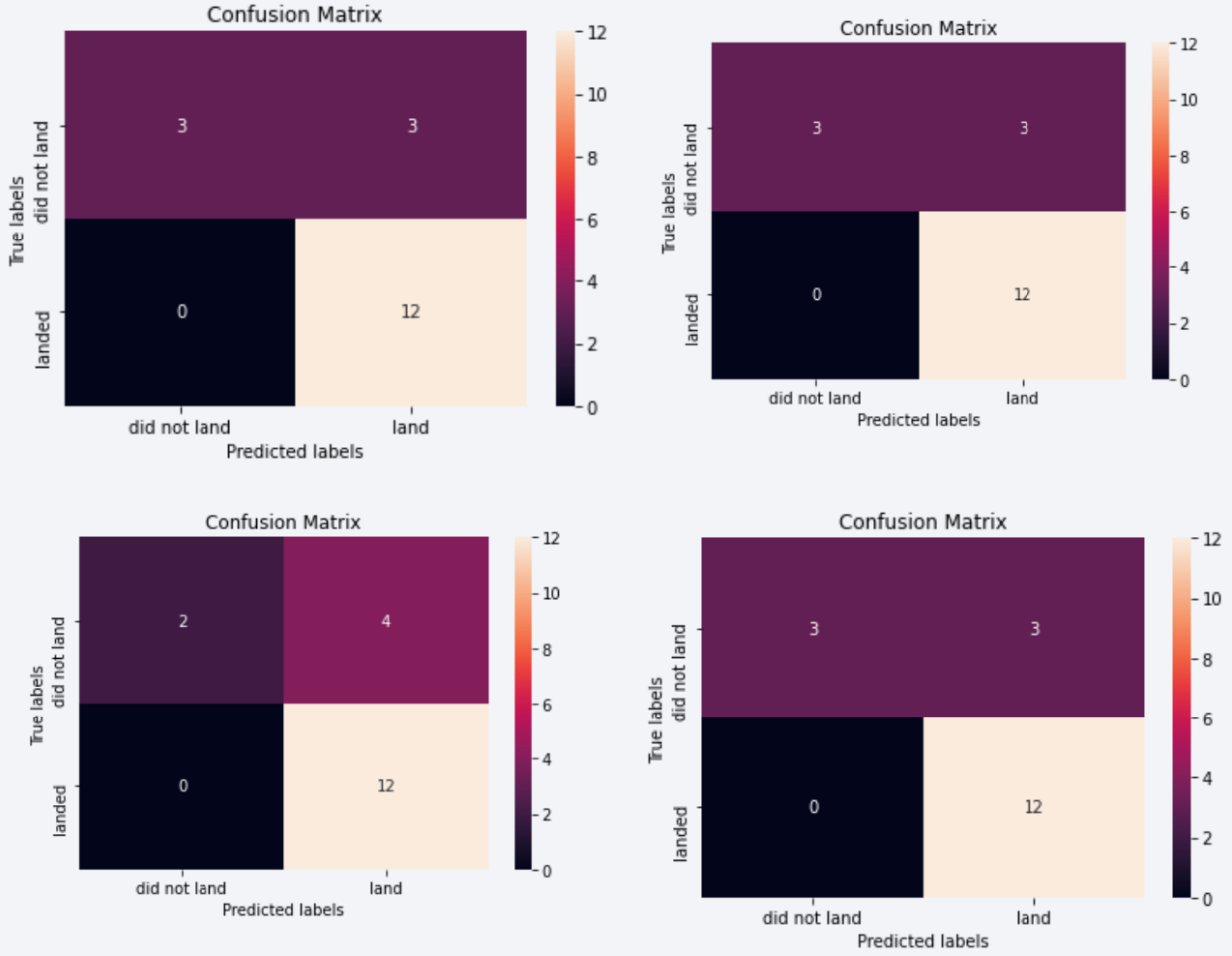
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- The SVM, KNN, and logistic regression models gave the best results in terms of prediction accuracy for this dataset.
  - Low weighted payloads perform better than the heavier payloads.
  - The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
  - KSC LC 39A had the most successful launches from all the sites.
  - Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!