IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeffrey Yeung
02/24/2022

Outline

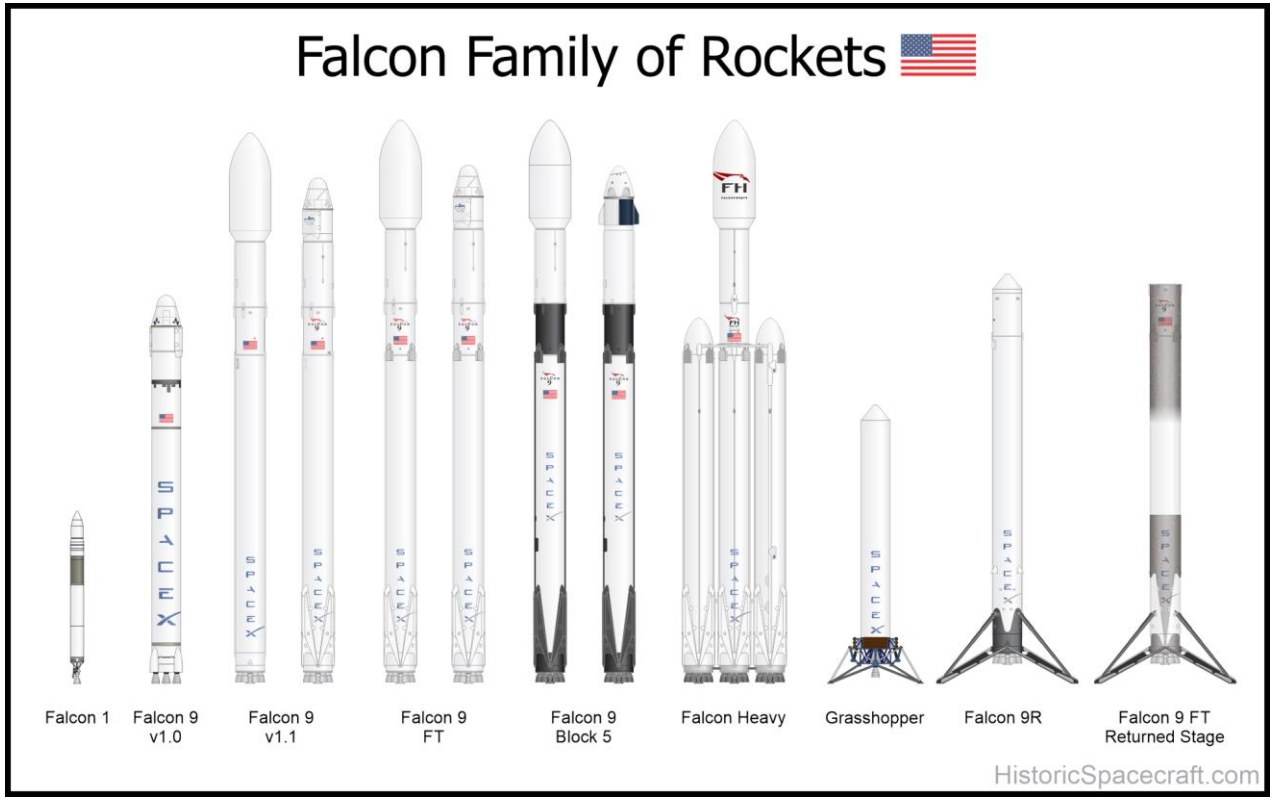
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The process is including data collection and wrangling methodology to make a prediction and convert results to visualization. The conclusion may help the company make a further plan or improve their current profit.
- Produced Four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K nearest Neighbors, and those results with accuracy rate of about 83.33%

Introduction

- With the developing of the commercial space, companies are making space travel affordable for everyone. SpaceX is a leading company in this field so a New company named SpaceY would like collect exist data from SpaceX to improve their business.
- How much cost of each launch in SpaceX?
- Will SpaceX reuse the first stage of Rocket?



Section 1

Methodology

Overview of Data Collection, Wrangling, visualization, dashboard, and Model Methods.

Methodology

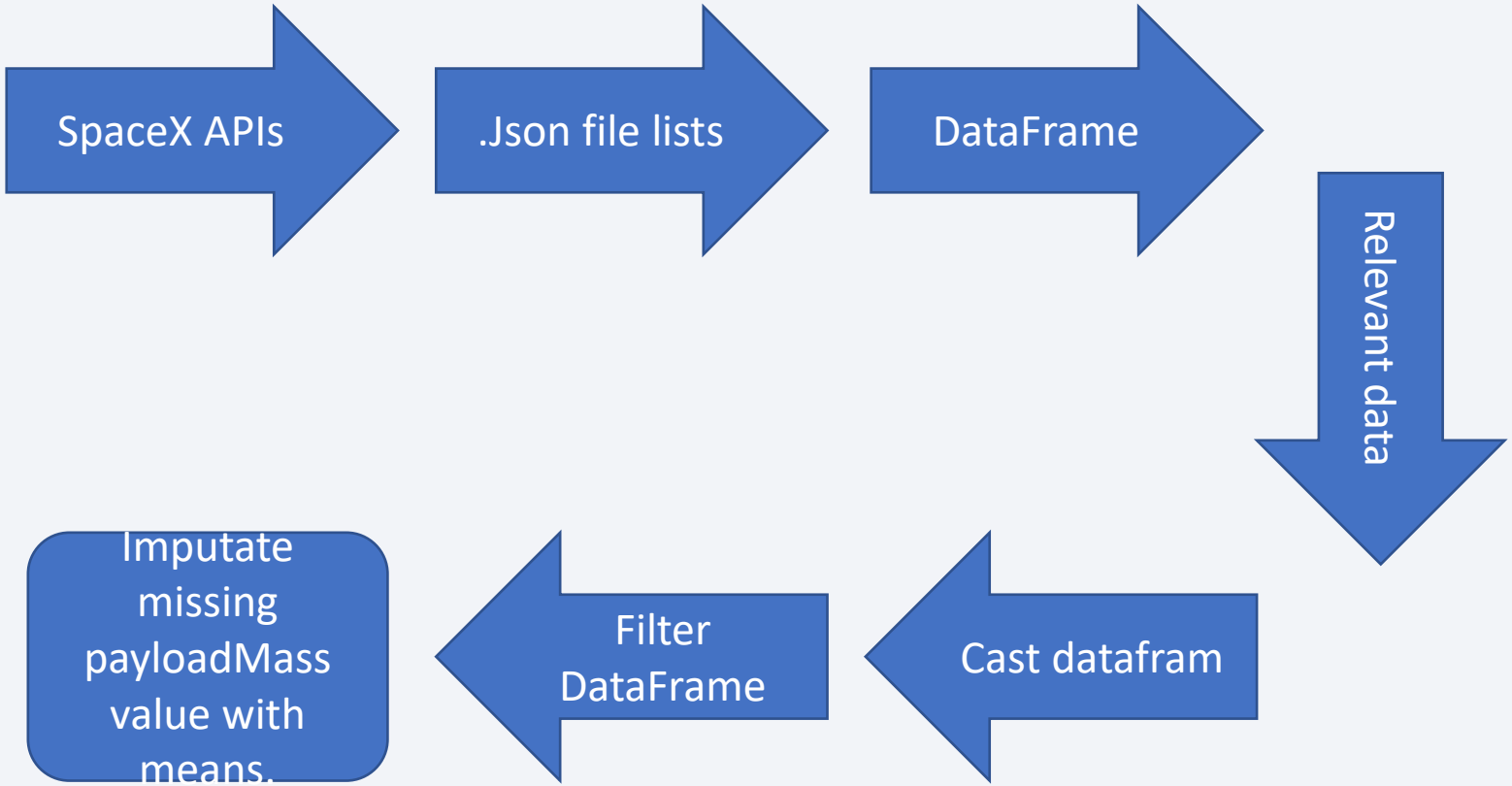
Executive Summary

- Data collection methodology:
  - Collecting SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. This API gave data about launches, including information about the rocket used, payload delivered, and etc.
- Perform data wrangling
  - The process of wrangling help us got detailed information that be mentioned above. In order to use those data to further analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Throughout build a machine learning pipeline to predict if the first stage of the Falcon 9 in SpaceX lands successfully. This model is able to training and testing data, then using the best hyperparameter values to predict a quite accuracy result.

Data Collection

- Data Collection process involved a combination of API requests from Space X public API and web scraping data.
- SpaceX API data columns are including: FlightNumber, Data, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, and etc.
- Web scraping data columns are including: Flight Number, Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Boostet and etc.

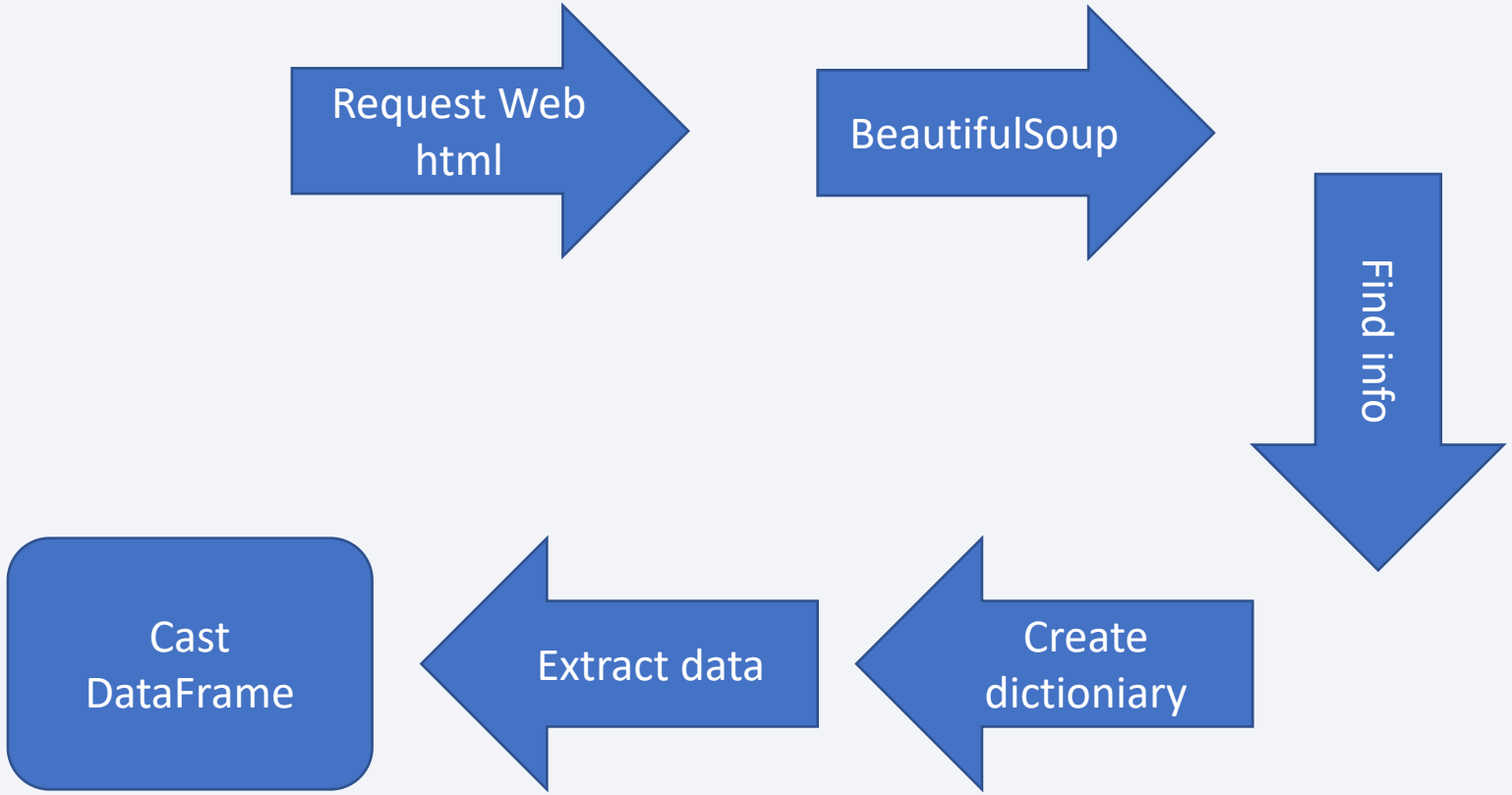
Data Collection – SpaceX API



Url link: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/f8986b67-cae7-40d3-bb6e-a52b7710b3ec/view?access_token=9e2c758ea7be25d1fbec06a212f8a926af45fff69901e559c7757f6e41288789

Data Collection- Scraping

URL:https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/cabe66a0-919e-4ca6-8517-c9c6d57442a6/view?access_token=b29e7a622f7fcb20c25b7ea1aff5ab6b9268c61c8bd034e5ef1b99009ddd9c39



Data Wrangling

- Create a training label with landing outcomes where true= 1 and false= 0.
- Outcome column has two components named “Mission Outcome” and “Landing Location”
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. Value

Mapping:

- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- URL:https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/5aa0c45a-1127-44ce-8bbd-48adde50020e/view?access_token=e2899498726432d6f8554fd652d51e2ce8fca4035fa9d843fc26a60ae9ca0419

EDA with Data Visualization

- In data visualization, using Scatter plots, line chart, and bar plots to compare relationship between dependent and independent variables to verify if a relationship exists so that they are able to used in training the machine learning model.
- Plots are: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

- URL:https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/34a117f7-b11c-47ac-b335-7d6361e4c699/view?access_token=ddf5d06341f65963b0e7fa8e4c104d42e868b204b258fdde67a0f8ac62f10bd1

EDA with SQL

- Loaded Dataset into IBM DB2 Database
- Used SQL Python to querying integration.
- Extracted information about Rocket launch names, mission outcomes, various payload sizes of customers and booster versions, and landing outcomes.

- URL:https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/1d4de8f3-f3ba-4194-ab8d-8e2f66d0f3a2/view?access_token=69ec360498a8504a1546b7c09405c481822a60ea684dda1d53523b2d9daeb6f4

Build an Interactive Map with Folium

- Throughout adding Folium maps mark launch Sites, successful and unsuccessful landings, and key location such as railway, highway, coast, and city to
- This process allows client to understand why launch sites may be located where they are and relationship between successful landing and location.

- Url:https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/9663036f-a592-4622-b624-6ac711604e4a/view?access_token=2357fc487ecce7074acc6021574a0d63e7ab1b799f5d2728acc1e5d54e6a5424

Build a Dashboard with Plotly Dash

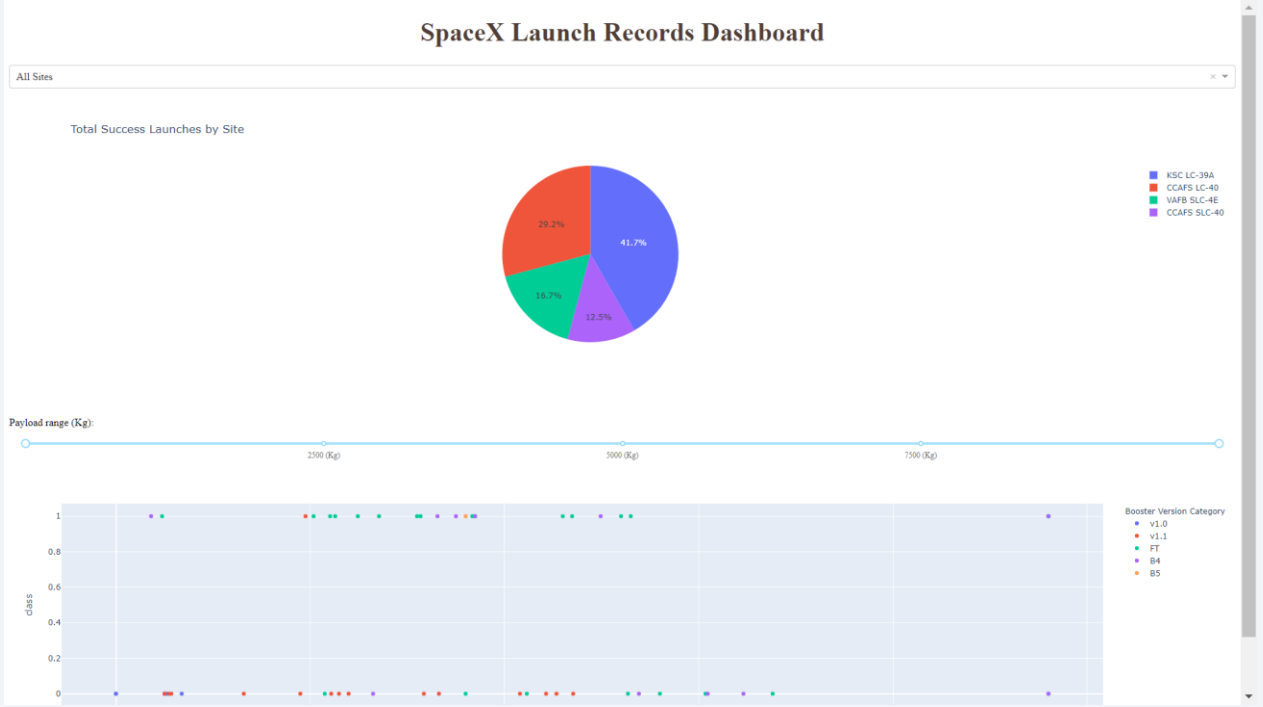
- Dashboard includes a pie chart and a scatter plot.
- Pie chart can clearly to show distribution of successful landings based on all launch sites and know what launch site success rates are.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize success rate
- The scartter plot can help client how success varies in launch sites, payload mass, and booster version category.

Predictive Analysis (Classification)

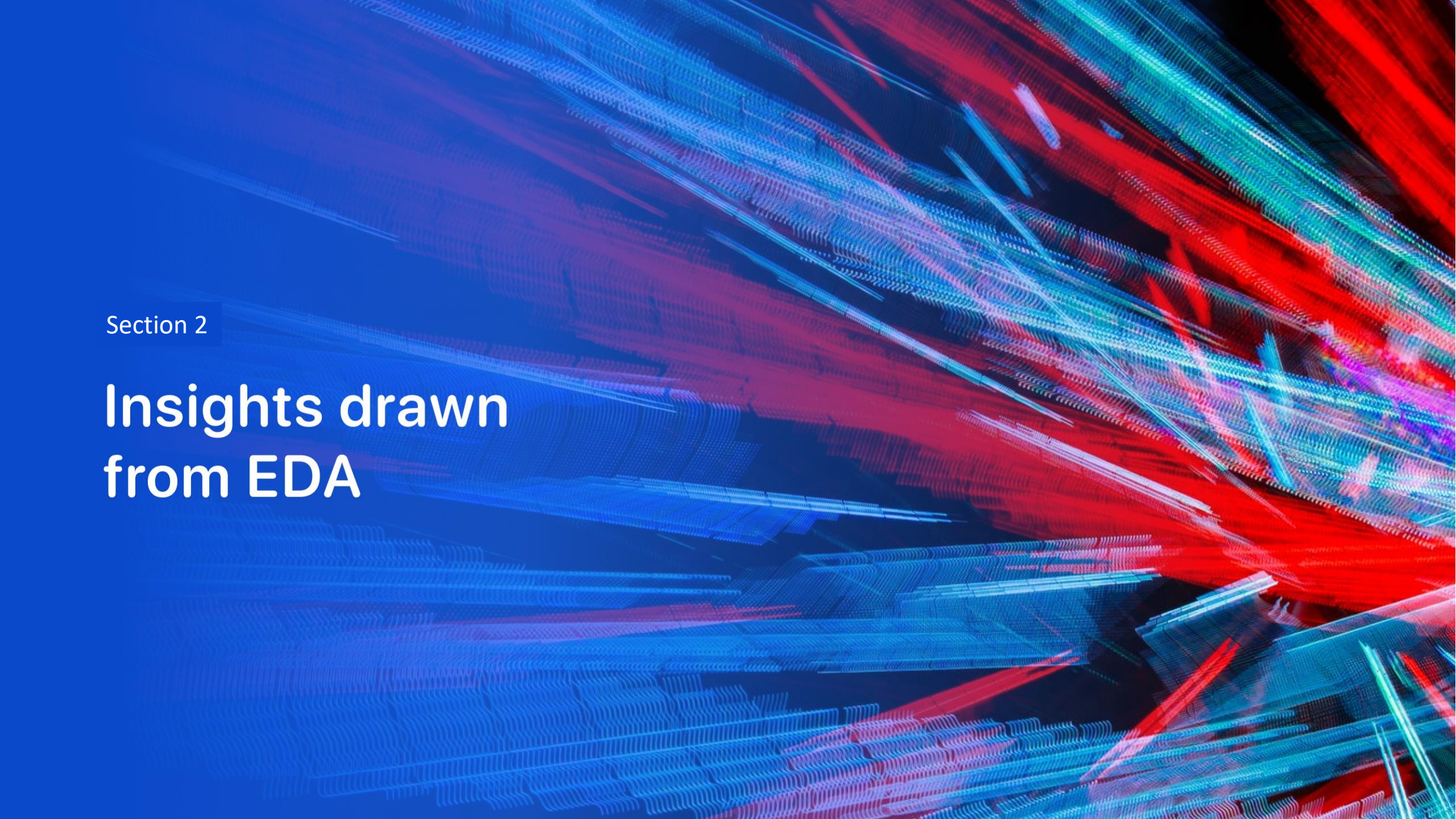
1. Split label column ‘class’ from dataset
2. Fit and transform features using standard scaler
3. Train and test split data
4. GridsearchCV to find optimal parameters
5. Use GridsearchCV on LogReg, SVM, Decision Tree, and KNN models
6. Score Models on split test set
7. Confusion Matrix for all model
8. Barplot to compare scores of models.

URLhttps://dataplatform.cloud.ibm.com/analytics/notebooks/v2/7ce774f5-21a9-44e7-a174-a51436c04f41/view?access_token=53efcdeccd17a07656529d4bae4f6689af63b9b2f4df31563c969c008ded6eb7

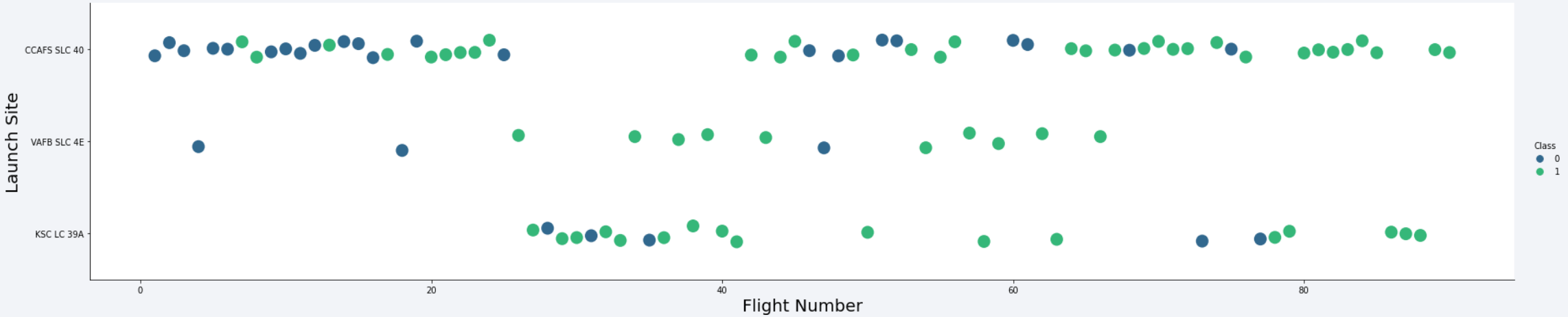
Results



- This preview of the plotly dashboard shows the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83.33%


Section 2
Insights drawn from EDA

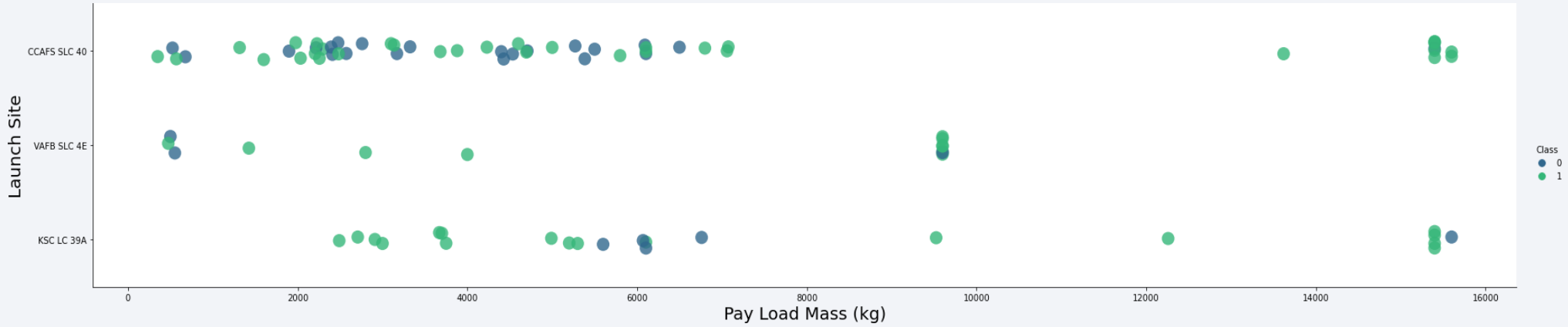
Flight Number vs. Launch Site



- Successful Launch
- Unsuccessful launch

The graphic shows an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased the success rate. CCAFS is considered the main launch site as it has the most volume.

Payload vs. Launch Site

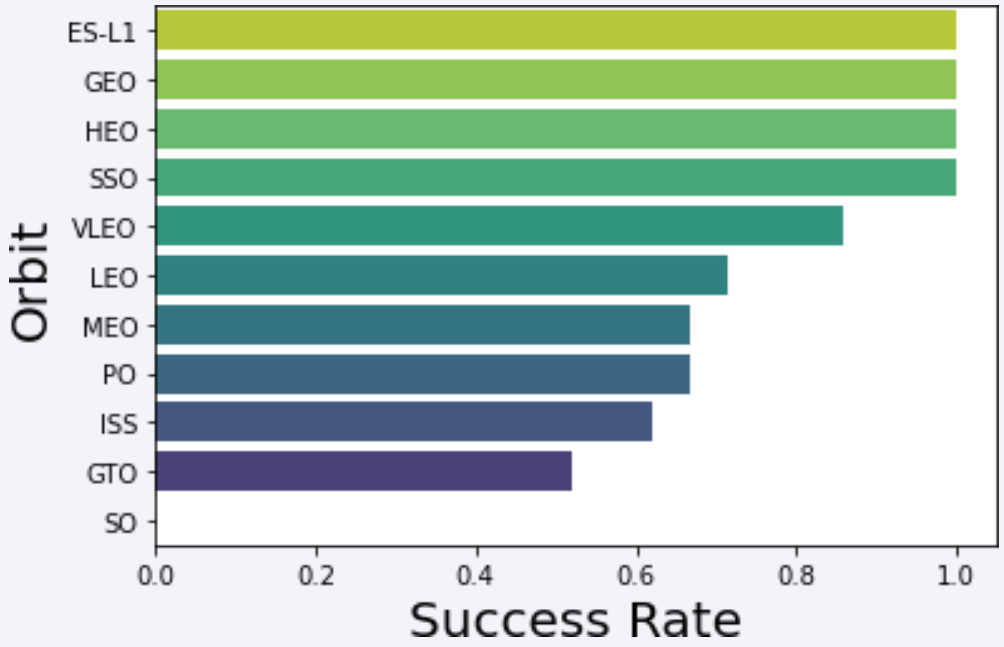


- Successful Launch
- Unsuccessful launch

- Payload mass mostly likes to fall between 0-6000 kg
- Different launch sites also seem to use different payload mass
- CCAFS SLC40 seem more likely fall under 6000kg

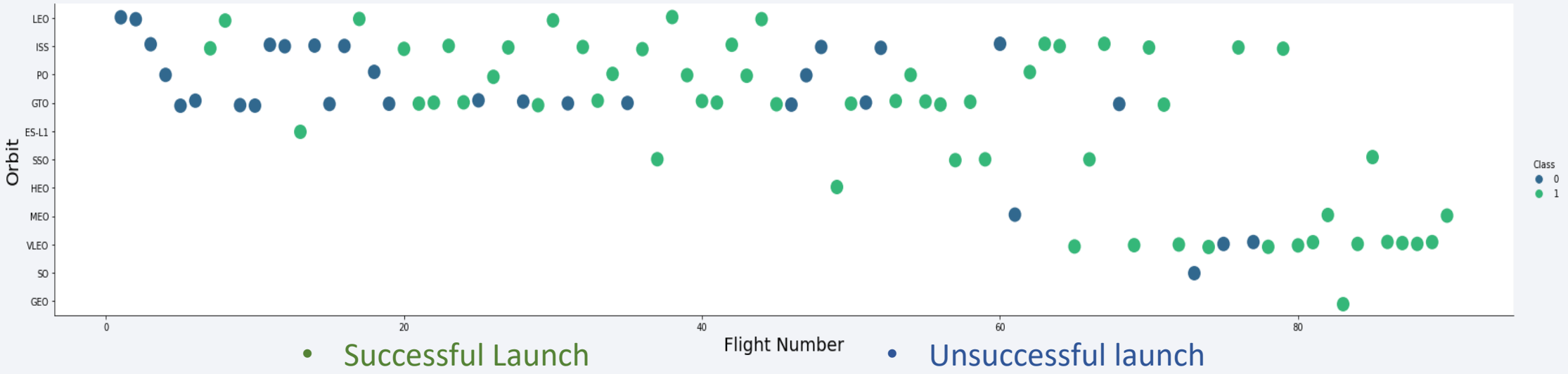
Success Rate vs. Orbit Type

Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%



- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

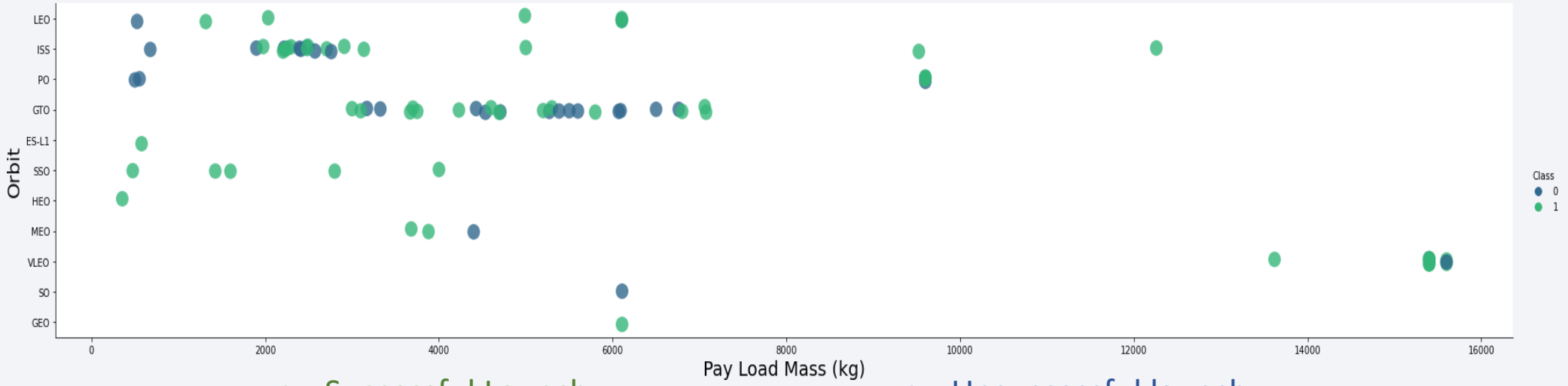
Flight Number vs. Orbit Type



- Successful Launch
- Unsuccessful launch

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches S
- paceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

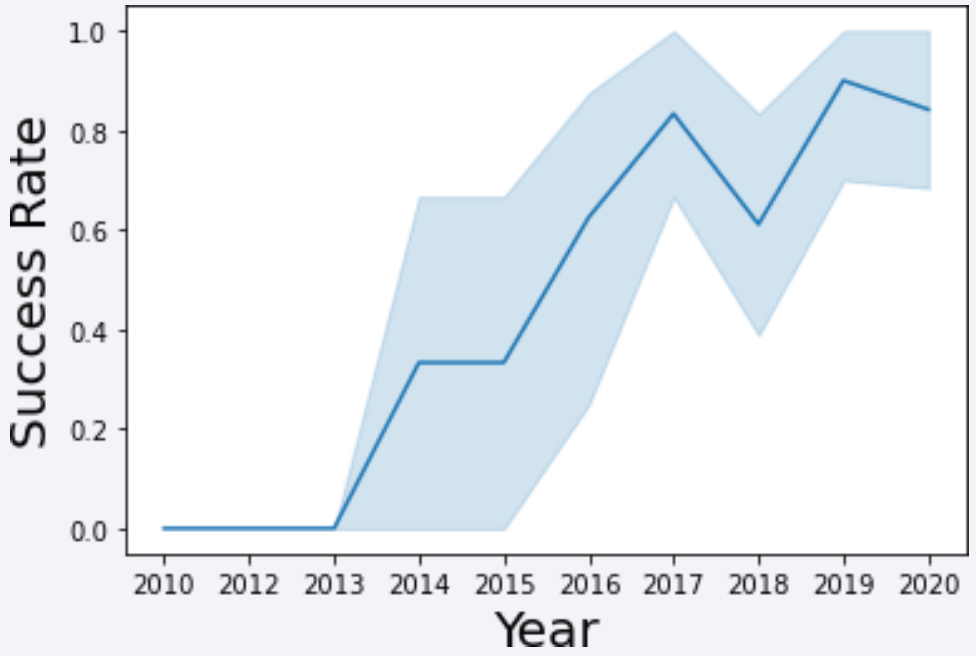


- Successful Launch
- Unsuccessful launch

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success generally increases rapidly over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



95% confidence interval
(light blue shading)

EDAwithSQL

All Launch Site Names

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
- launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/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

- This chart shows several Launch site names contain CCA and their lat, long, and class. It could help clint to understand the consulting clearly.

First Successful Ground Landing Date

- This query returns the first successful ground pad landing date.
- First ground pad landing wasn’t
- until the end of 2015.
- Successful landings in general
- appear starting 2014.

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- This query returns the 4 booster versions that had successful drone ship landings and a payload mass between 4k to 6k noninclusively

Total Number of Successful and Failure Mission Outcomes

- This query returns a count of each
- mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing
- failures are intended.
- Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

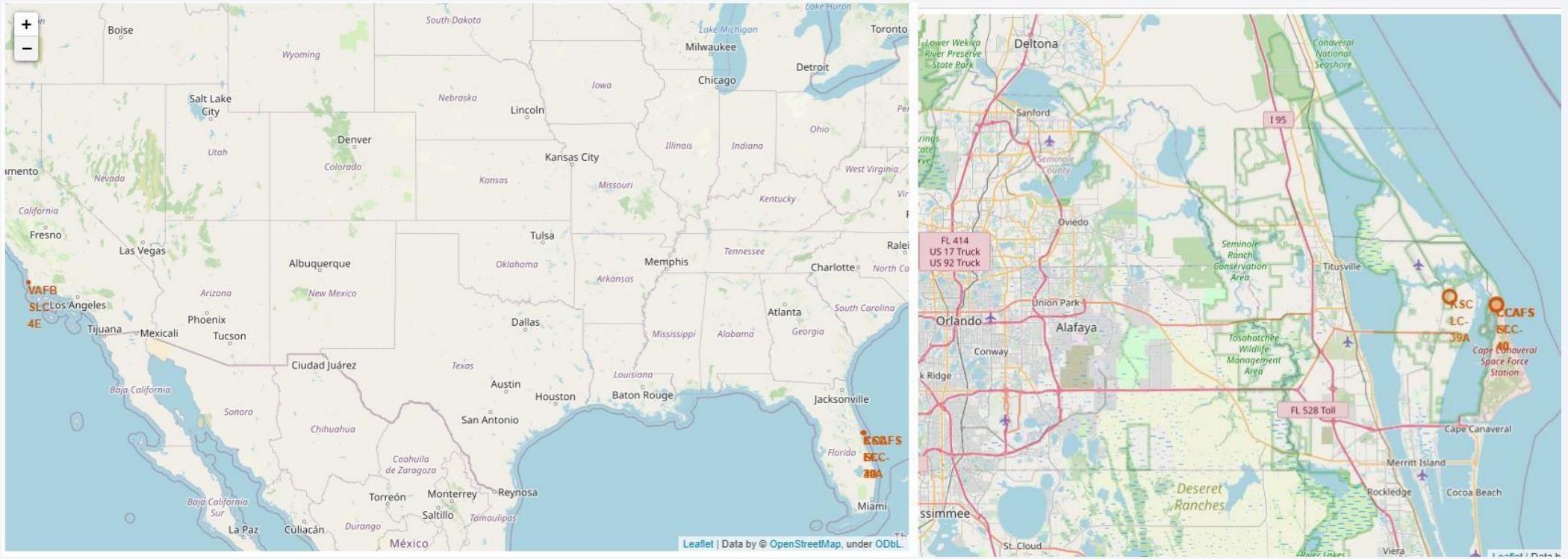
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

Section 3

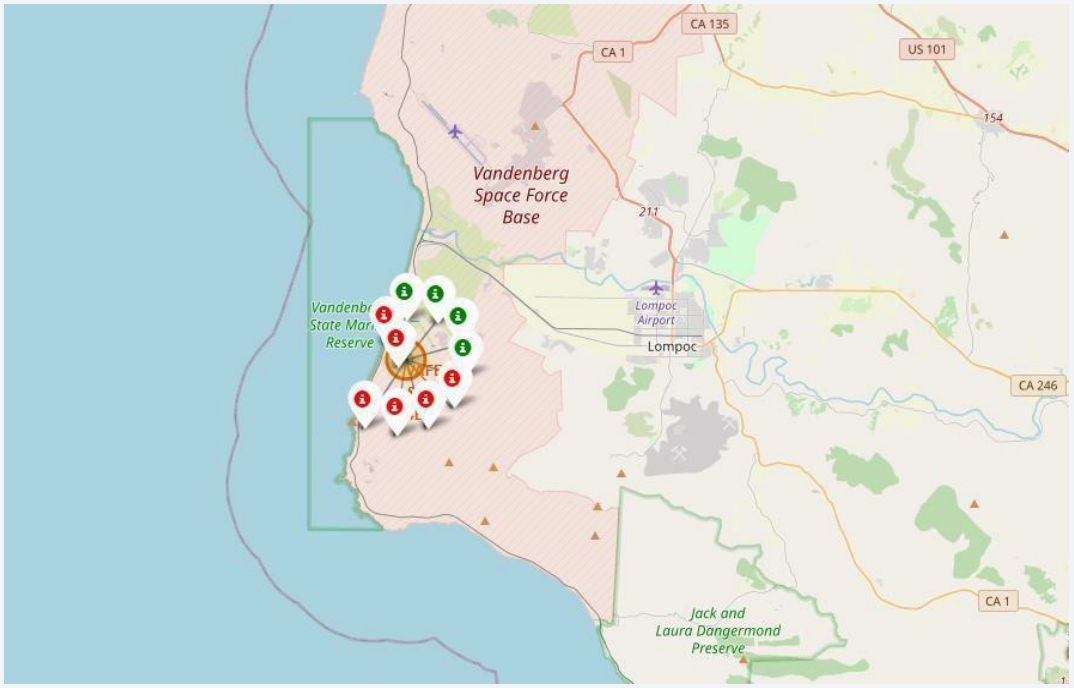
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

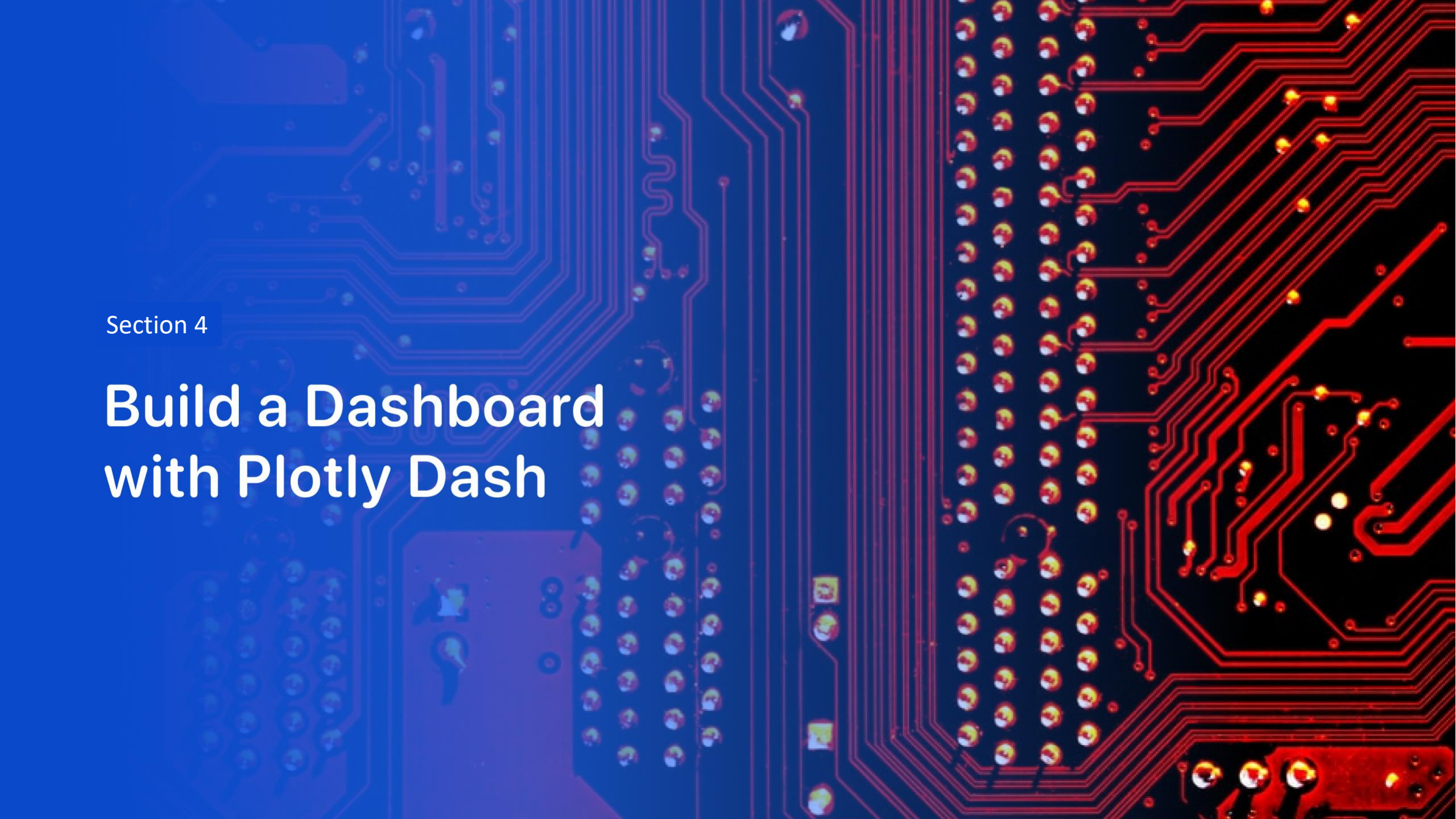


- The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

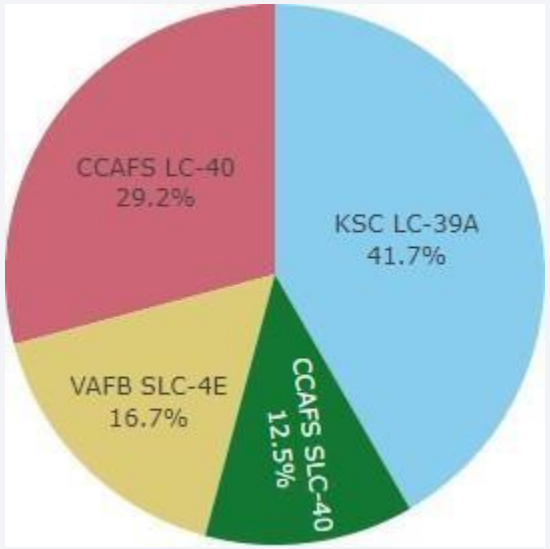
Color-coded launch markers



- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
- landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.


Section 4
Build a Dashboard with Plotly Dash

Successful Launches in Launch Sites

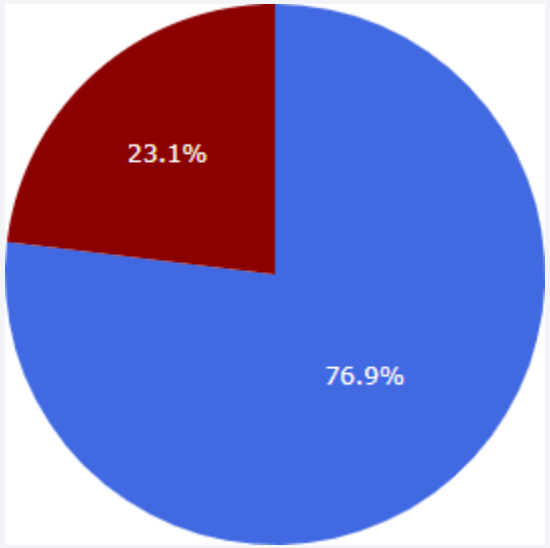


- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

Highest Success Rate Launch Site



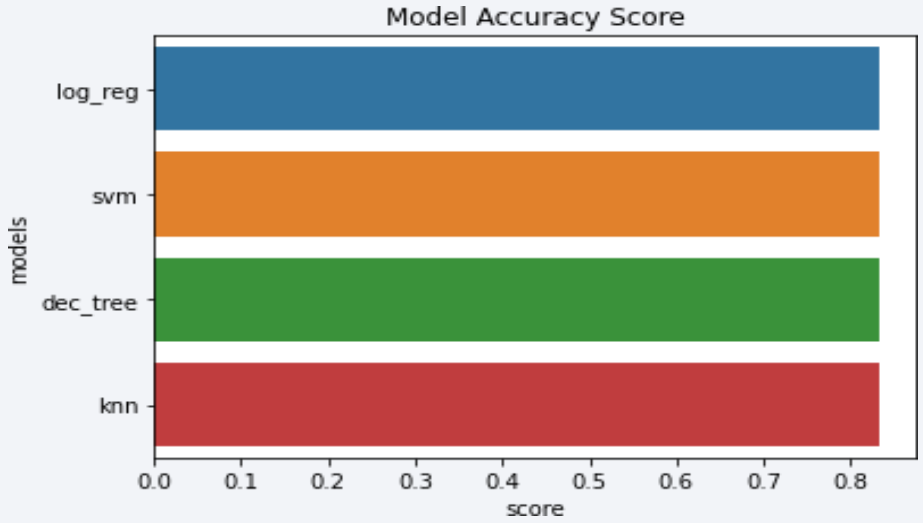
- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.



- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix

- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

- Task: To develop a machine learning model for SpaceY to against SpaceX and improve it further developing.
- The primary goal is finding the cost of each launch in SpaceX, details of SpaceX's first stagy and its information of successful and unsuccessful landing.
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- Created a machine learning model and fount an accuracy rate of 83.33%
- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not

Appendix

Githut repository url:

- https://github.com/Feifei0320/Final-Project

Thank you!