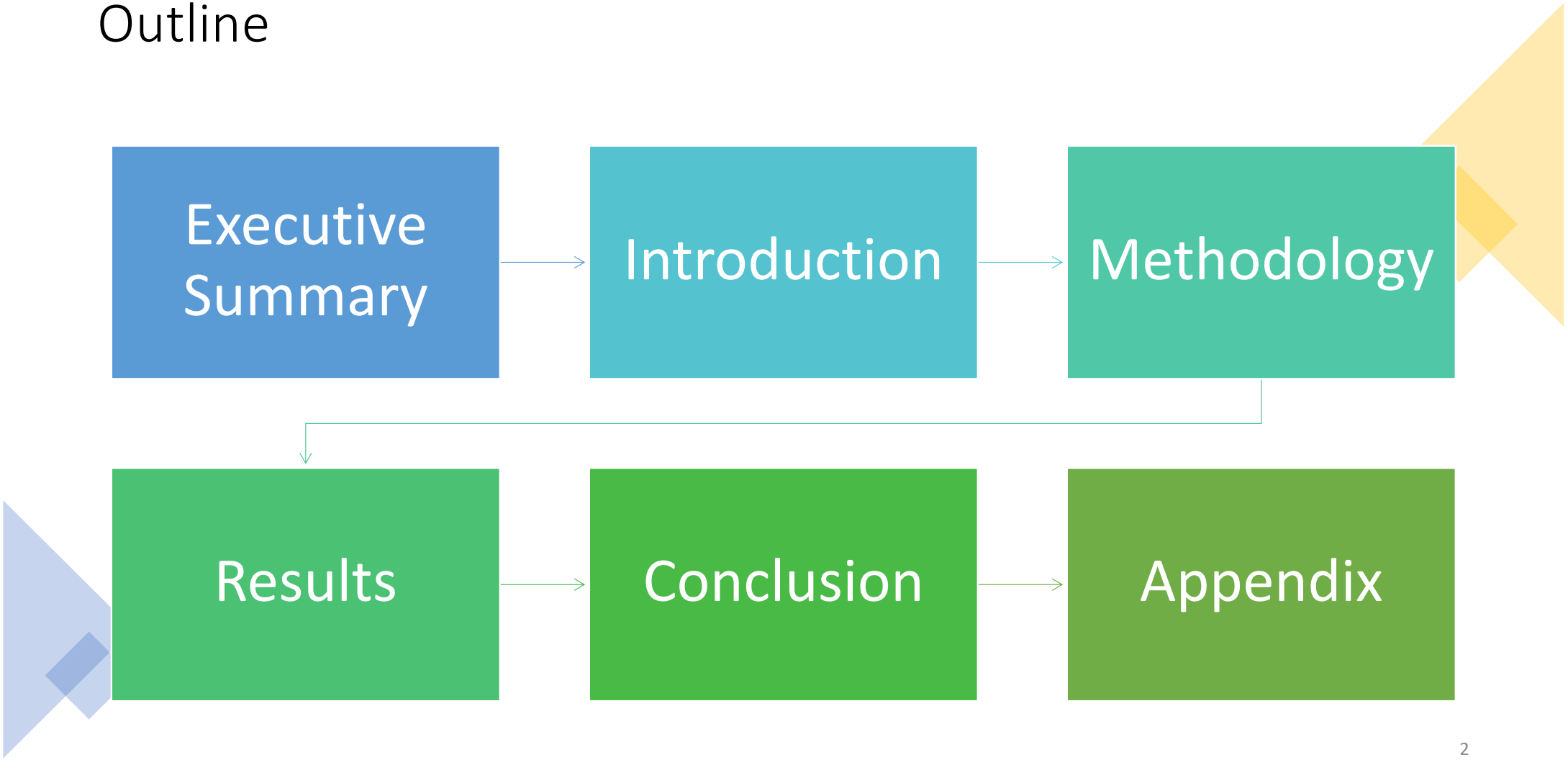
Winning Space Race with Data Science

Raghvendra Singh

30th November 2021

Outline


Executive Summary
Introduction
Methodology
Results
Conclusion
Appendix

Executive Summary

- **Summary of methodologies.**
- Data collection source used: Spacex API and Wikipedia page(web scrapping using get.response to get the information stored). Data wrangling Labels created for successful and failed landings. Further data analysis using several libraries of python such as Folium, visualization libraries.
- Further to identify the optimum parameters for machine learning models, used GridSearchCV. Visualization all of the models' accuracy scores. ML models such as logistic regression, SVM, KNN, classification tress.
- **Summary of all results.**
- With an accuracy percentage of around 83.33 percent, the models gave similar outcomes. All the models overestimated the likelihood of successful landings. For better model determination and accuracy, more data is required to arrive at a better predicting model.

Introduction

- **Project background and context.**
- We'll forecast if the Falcon 9's first stage will successfully land. On its website, SpaceX advertises Falcon 9 rocket flights for 62 million dollars; other suppliers charge upwards of 165 million dollars per launch; much of the savings comes from the fact that SpaceX can reuse the first stage. As a result, if we can figure out if the first stage will land, we can figure out how much a launch will cost. If another firm wishes to compete with SpaceX for a rocket launch, this information can be used.
- **Problems you want to find answers.**
- Space Y has given us the goal of developing a machine learning model that can predict whether a Stage 1 recovery will be effective.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
- Webscrapping SpaceX Wikipedia page and request to SpaceX API.
- Perform data wrangling
  - Successful landings are classified as true unsuccessful landings as False.
  - Further carrying out exploratory data 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
  - Evaluation of classification model is done with using GridsearchCV.

Data Collection

The data was gathered using a combination of API requests from the Space X public API and web scraping data from a table in the Wikipedia entry for Space X.

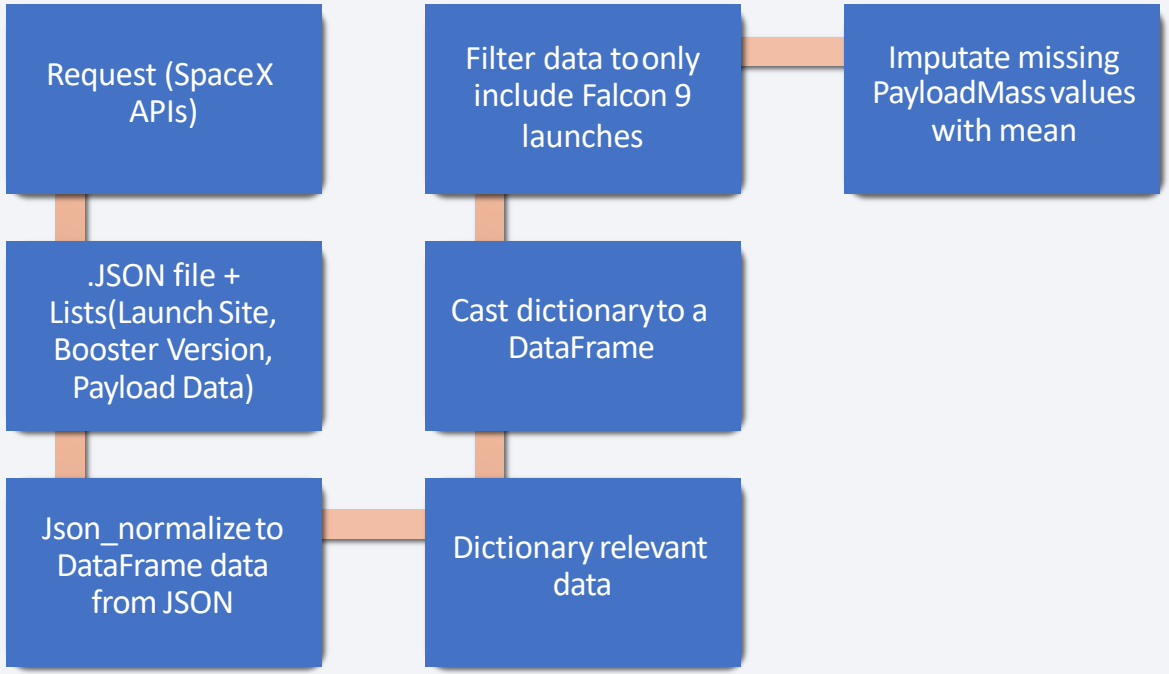
Web scrapping Falcon 9 and Falcon heavy launches from Wikipedia records.

Making a request to SpaceX API for public data collection.

Converting the data into correct format with relevant columns for analysis.

Data Collection – SpaceX API

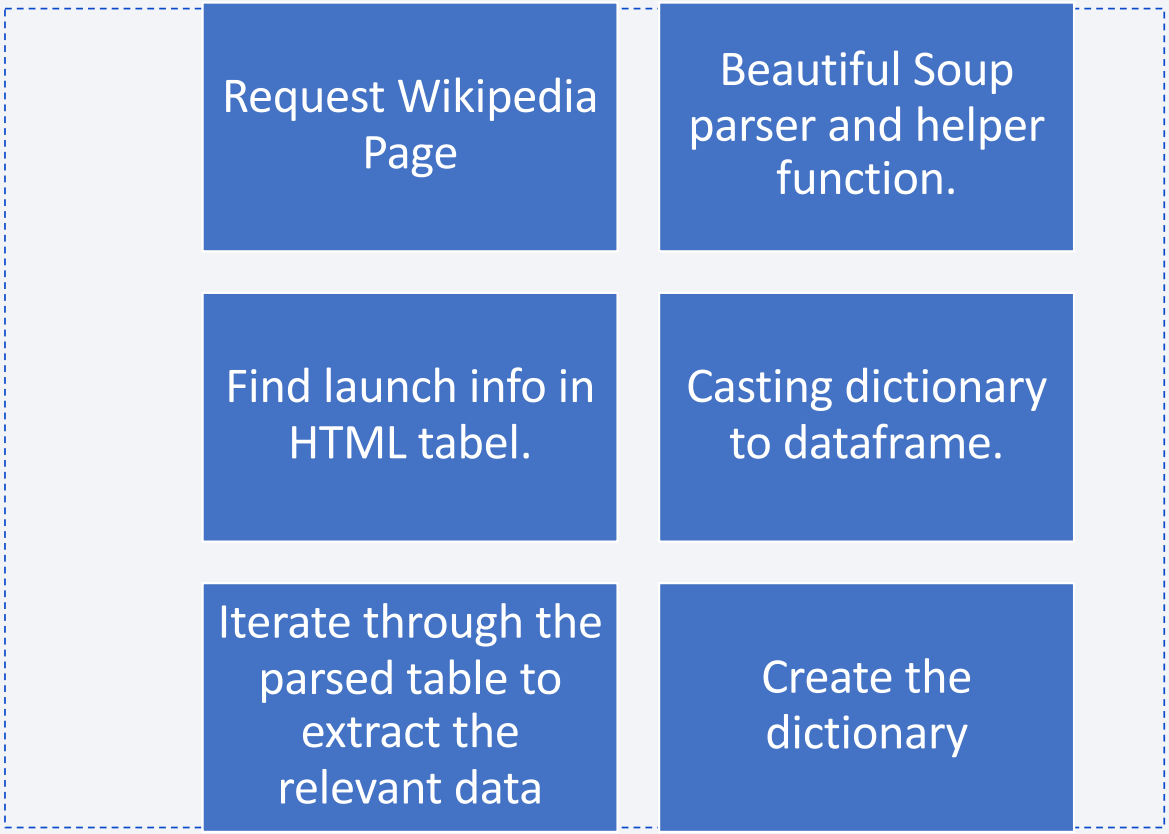
- GitHub URL: https://github.com/Raghhavv/Data_science/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



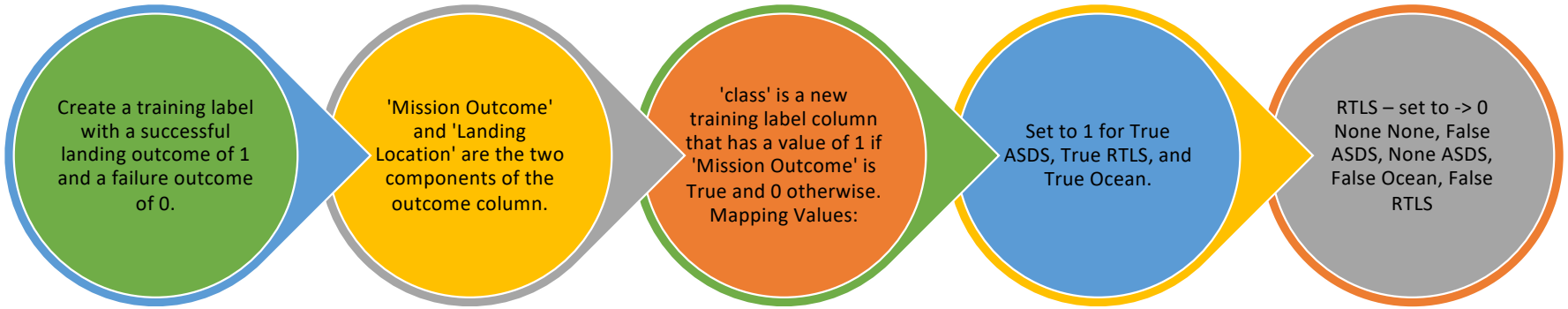
Data Collection - Scraping

- GitHub URL:

https://github.com/Raghhavv/Data_science/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling



EDA with Data Visualization

Charts plotted between variables	Charts used- Scatter plots, line charts, and bar plots
Payload Vs Flight Number	To understand the correlation between the variables.
Flight Number Vs Launch Site	To understand the correlation between the variables.
Payload Vs Launch Site	To understand the correlation between the variables.
Success Rate at each orbit type	Yearly trend
Flight number Vs Orbit Type	To understand the correlation between the variables.
Payload Vs Orbit Type	To understand the correlation between the variables.
Launch Success yearly trend	Yearly trend

GitHub URL: https://github.com/Raghhavv/Data_science/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

The data set has been loaded into the IBM DB2 database.

SQL Python integration was used to query the data.

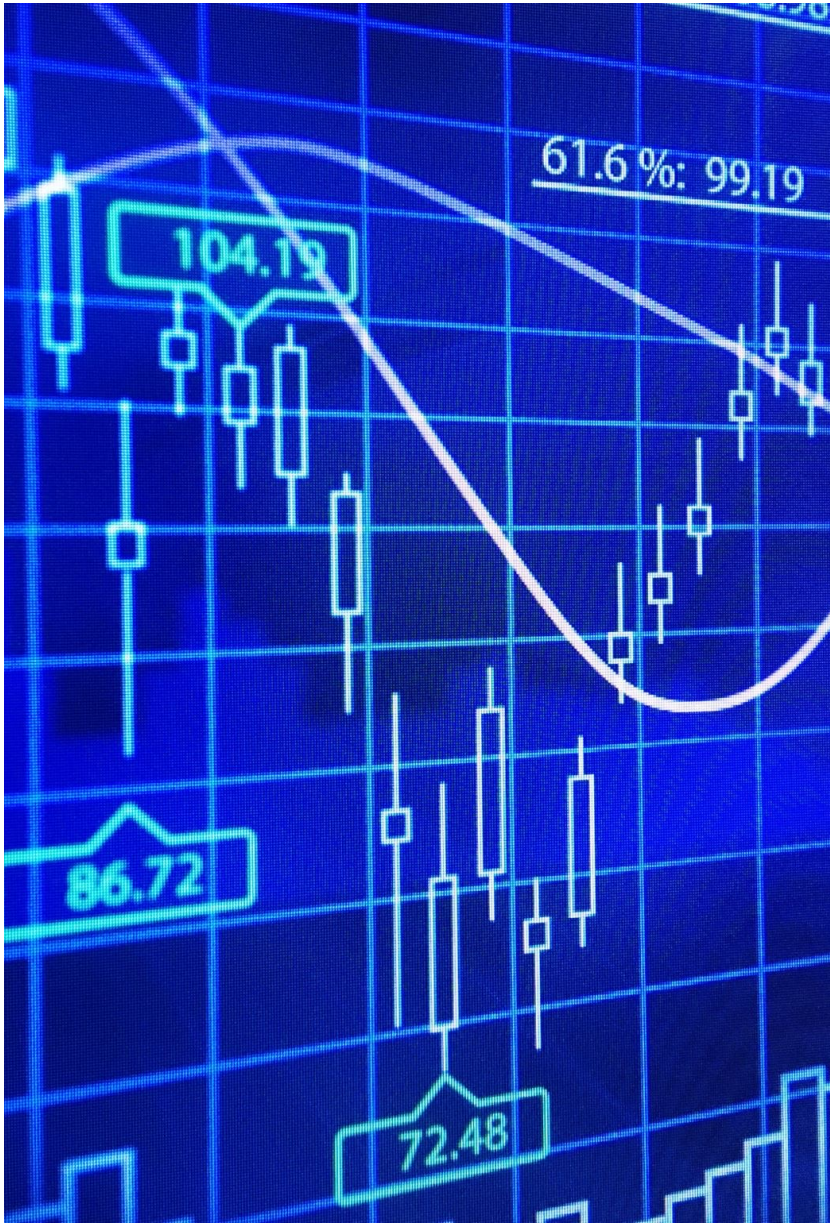
To gain a better grasp of the dataset, queries were run.

I wanted to know the identities of the launch sites, the mission outcomes, the varied payload sizes of customers and booster types, and the landing results.

Build an Interactive Map with Folium

Objects created.	Why the objects were used.
Folium.circle	To add a highlighted circle area with a text label on a specific coordinate
Folium.marker	To add a text information on a specific coordinate of relevance.
Marker Cluster	To cluster the coordinate containing many markers.

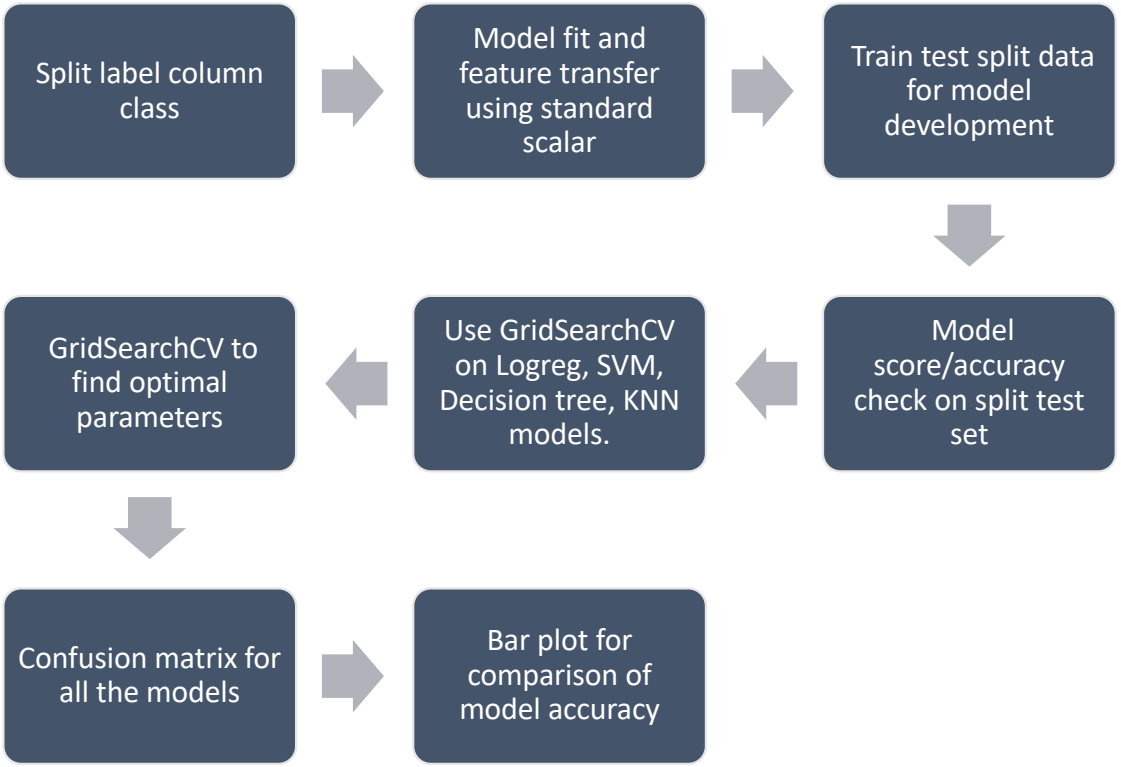
GitHub URL: https://github.com/Raghhavv/Data_science/blob/main/lab_jupyter_launch_site_location%20(1)%20(1).ipynb



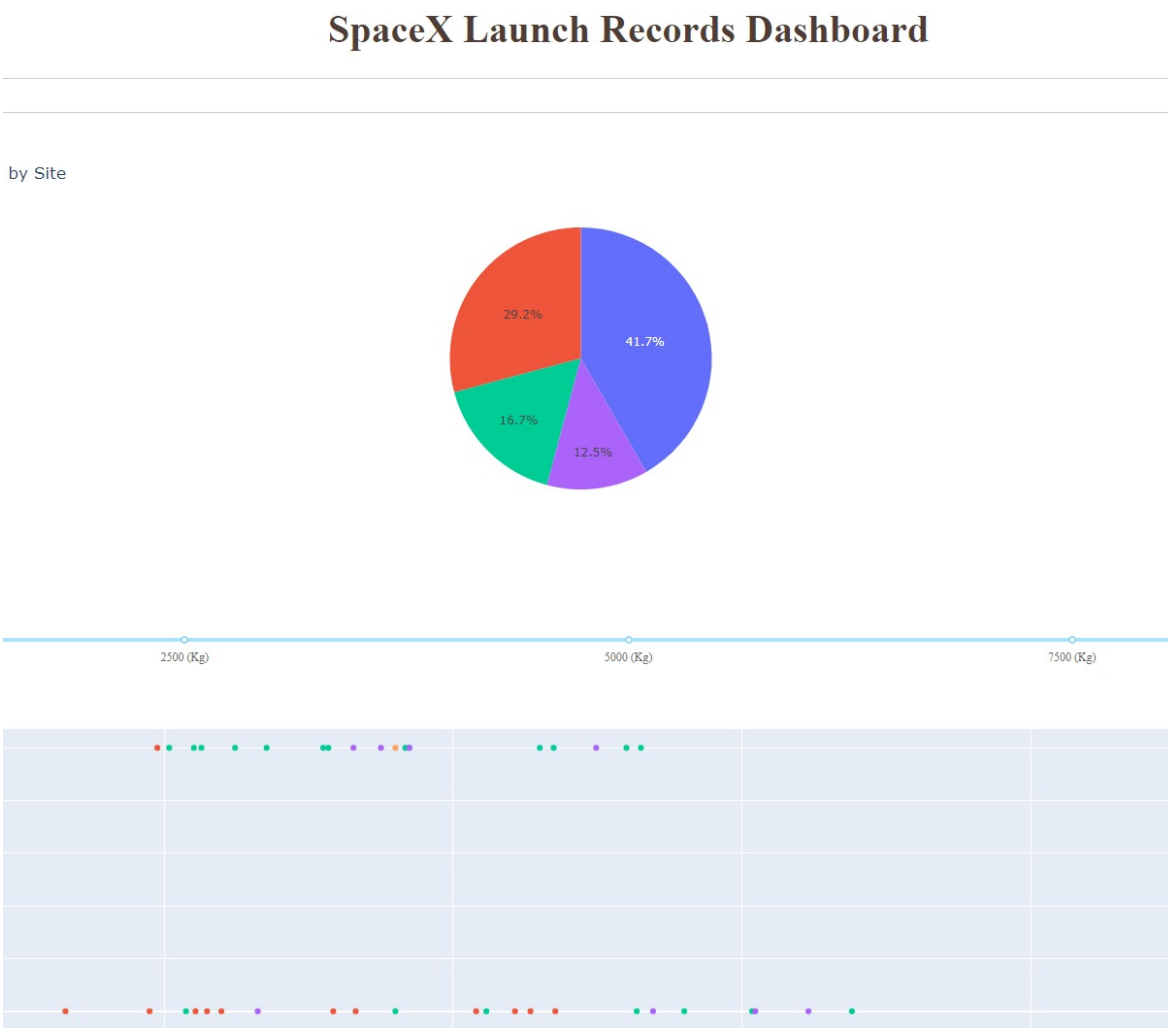
Build a Dashboard with Plotly Dash

- A pie chart and a scatter plot are included on the dashboard.
- The pie chart can be used to display the distribution of successful landings across all launch sites, or it can be used to represent the success rates of individual launch sites.
- Scatter plot requires two inputs: all sites or a single site, as well as a payload mass slider between 0 and 10000 kg.
- The pie chart is used to display the success rate of the launch site.
- The scatter plot can show how success differs depending on launch sites, payload mass, and booster version category.

Predictive Analysis (Classification)

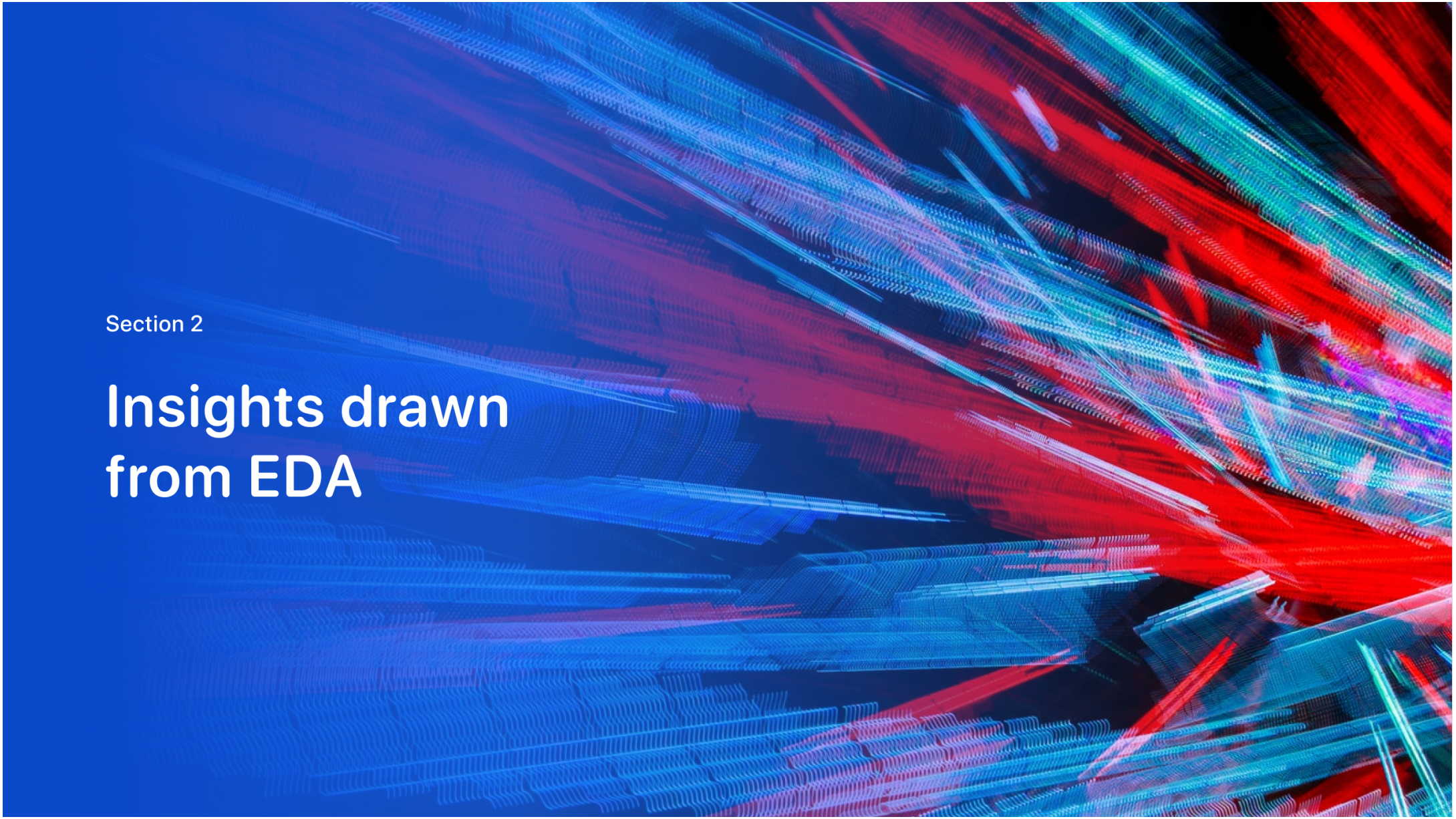


GitHub URL:
https://github.com/Raghhavv/Data_science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb



Results

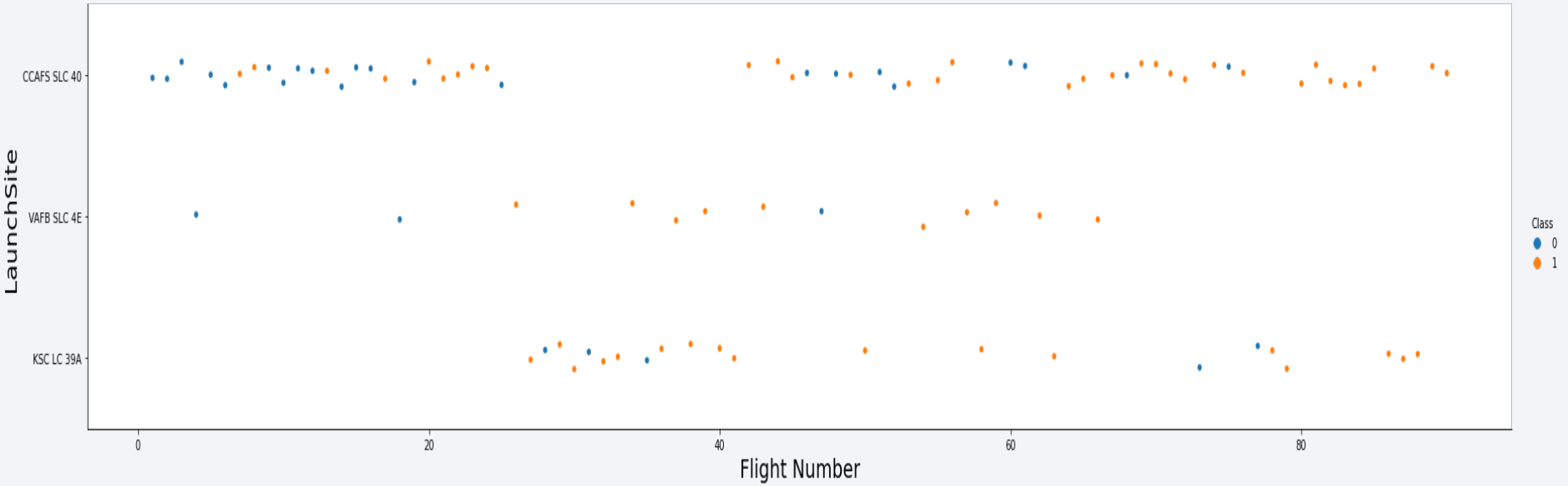
- This is a look at the Plotly dashboard in action. The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with roughly 83 percent accuracy will be displayed on the following sides.



Section 2

Insights drawn from EDA

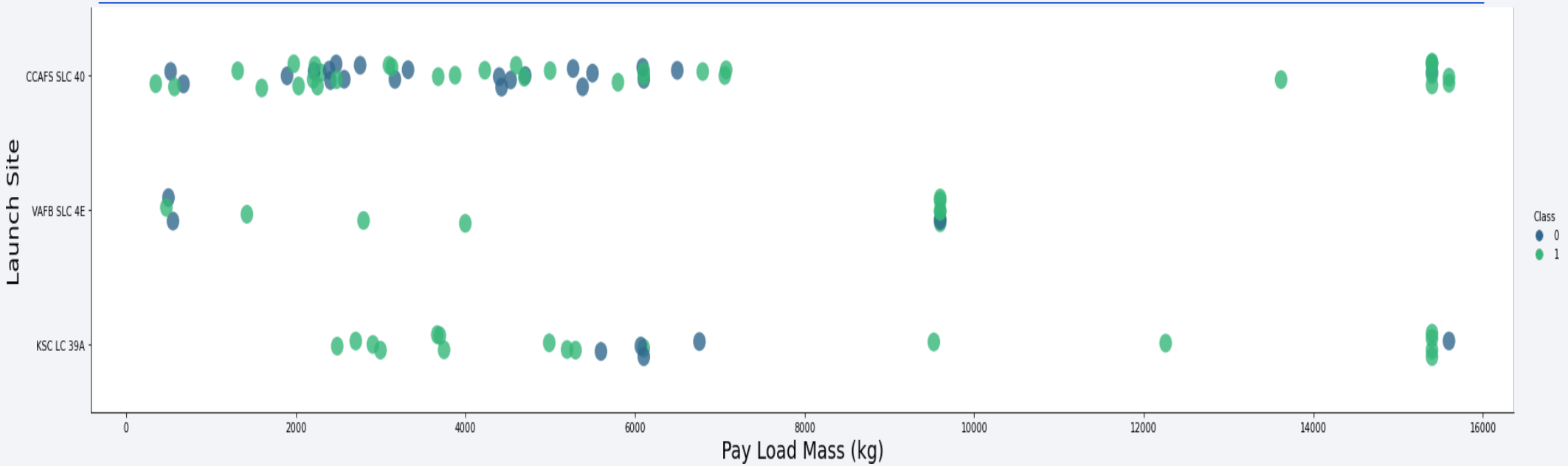
Flight Number vs. Launch Site



Blue indicates successful launch; Orange indicates unsuccessful launch.

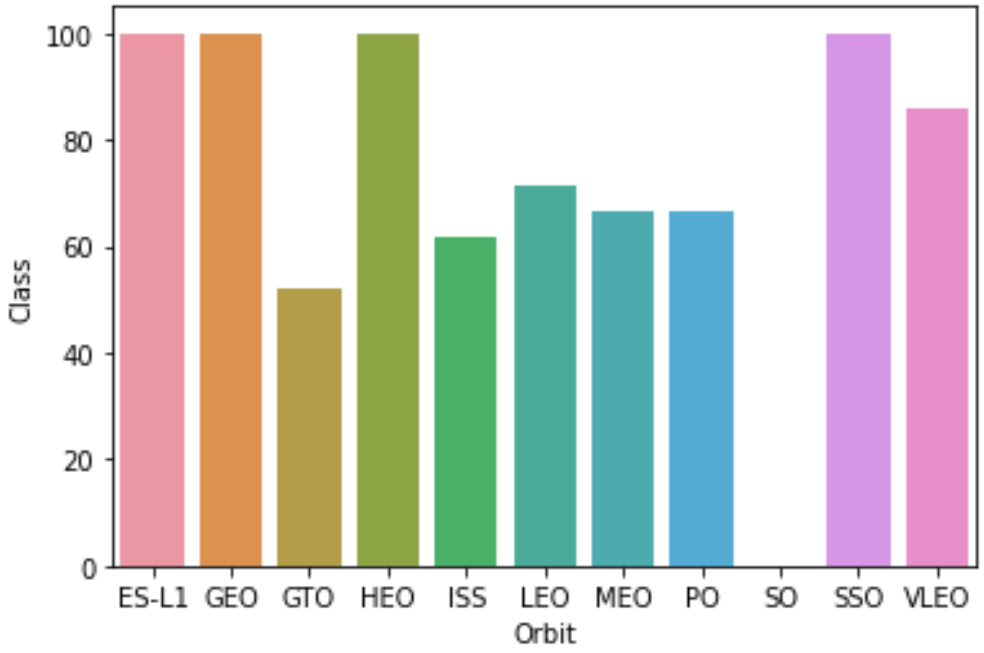
The graph depicts a rising success rate over time (indicated in Flight Number). Around flight 20, there was most likely a breakthrough that dramatically raised the success rate. Because it has the most traffic, CCAFS appears to be the primary launch point.

Payload vs. Launch Site



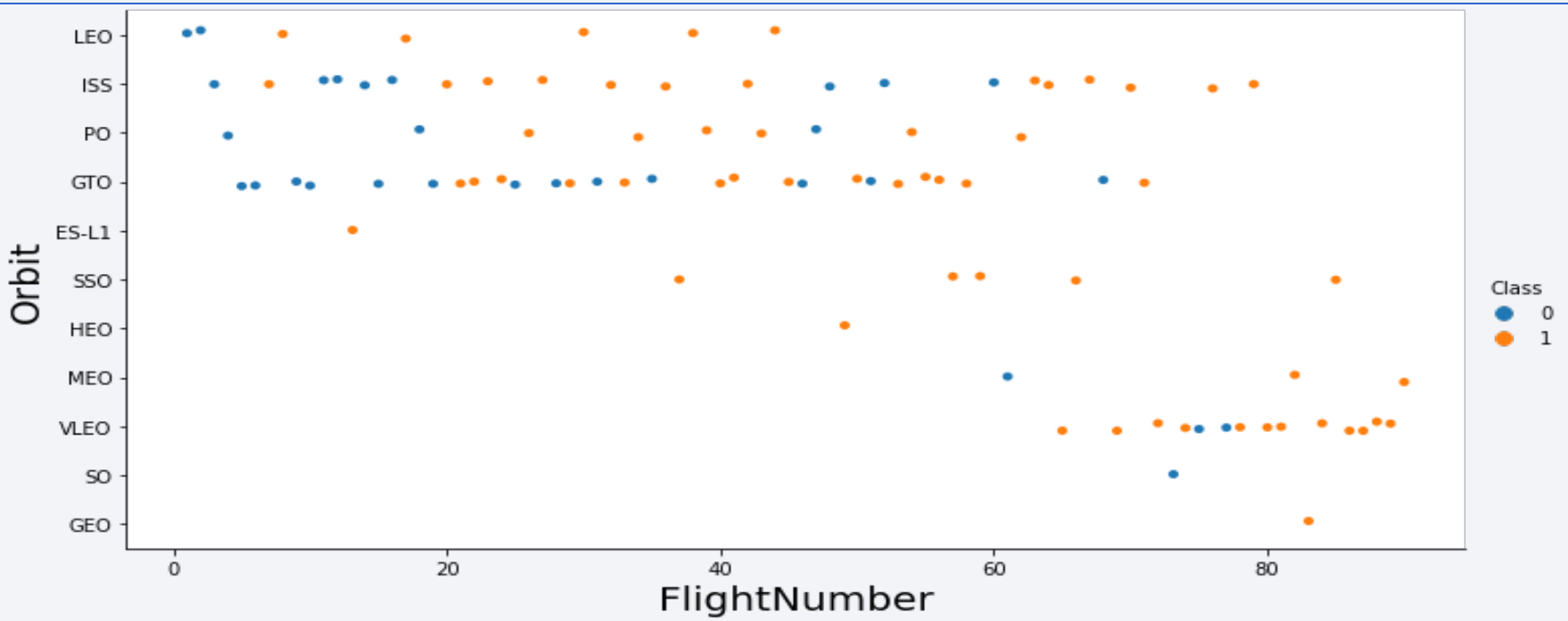
VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

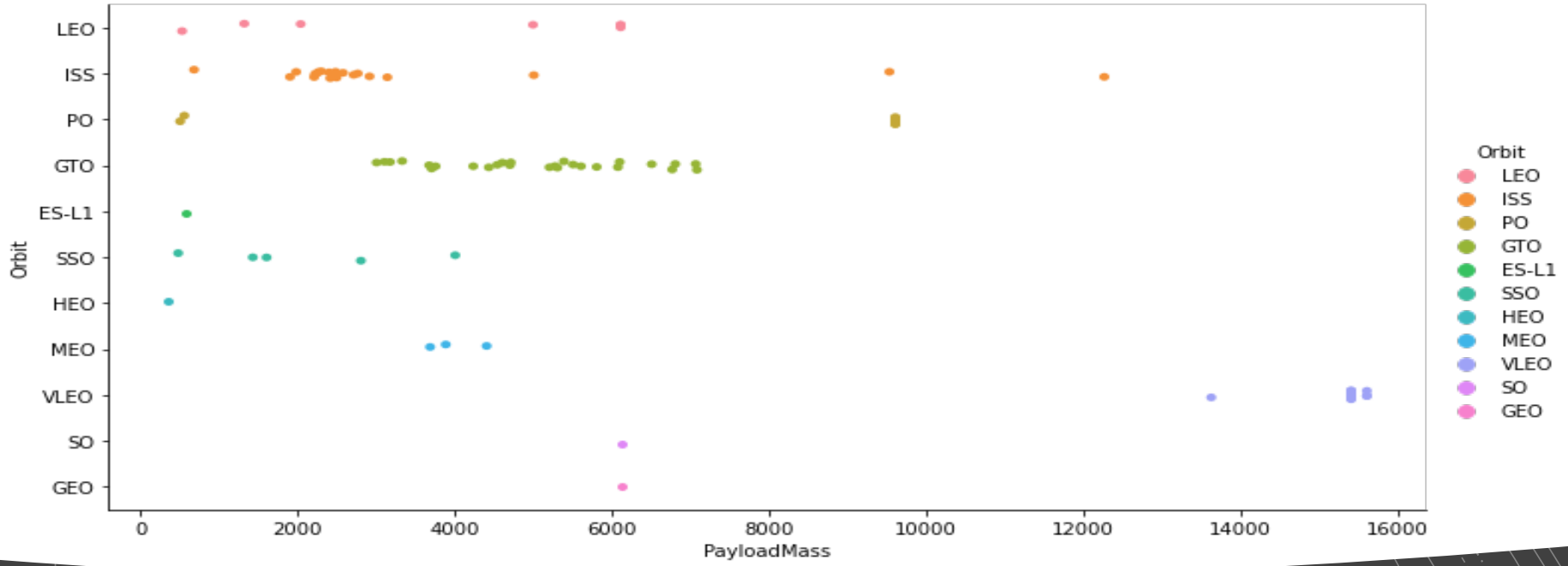


- ES-L1 (1), GEO (1), and HEO (1) all have a 100% success rate (sample sizes in parenthesis) SSO (5) has a 100% success rate.
- VLEO (14) has a good success rate and makes a lot of efforts.
- SO (1) has a failure rate of 0%.
- GTO (27) has a success rate of roughly 50%, but the greatest sample size.

Flight Number vs. Orbit Type



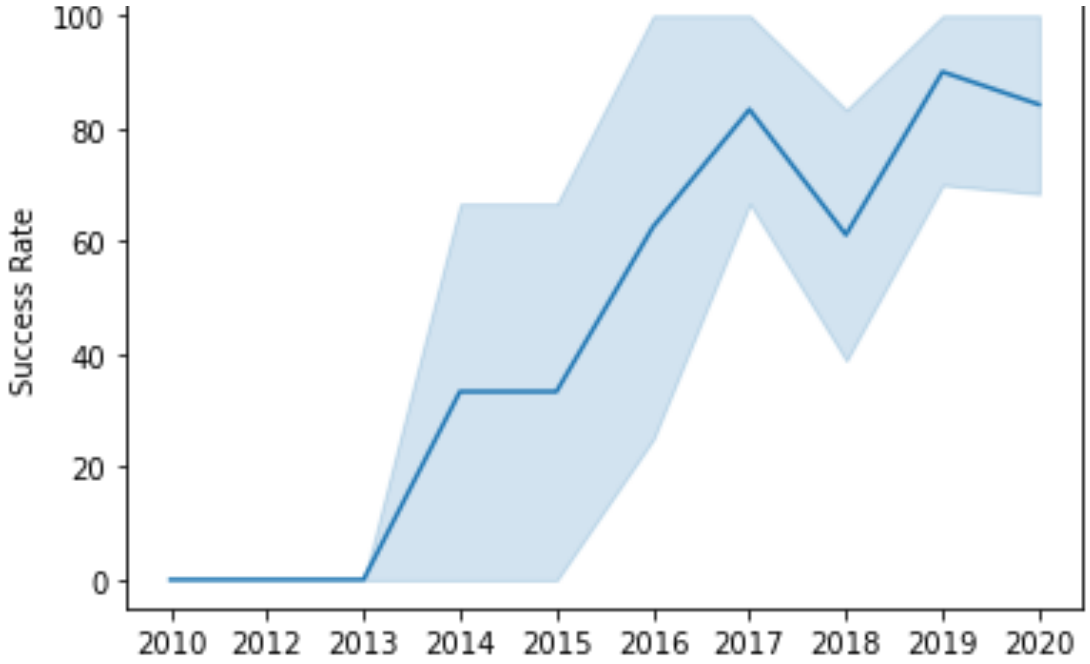
In LEO orbit, success appears to be linked to the number of flights.
In GTO orbit, there appears to be no link between flight number and success.



Payload vs. Orbit Type

- Payload mass appears to be related to orbit, with LEO having a relatively low payload mass and SSO having a comparatively high payload mass.
- VLEO, the other most successful orbit, only has payload mass values on the higher end of the spectrum.

Launch Success Yearly Trend



- *The trendline showcases the pickup of success rate in 2013, plateaued in year 2014 and again seeing an increase in success rate in year 2015 through the year ’16 seeing a dip in year ‘18. Finally reaching a success rate of 80%.*

All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

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

Query unique launch site names from database.

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'

In [5]:
```
%%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.
```

Out[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

Total Payload Mass

This query computes the average payload mass for launches that used the F9 v1.1 rocket.

The F9 1.1 average payload mass is on the bottom end of our payload mass range.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

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

avg_payload_mass_kg
2928

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

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

avg_payload_mass_kg
2928

This query computes the average payload mass for launches that used the F9 v1.1 rocket.

The F9 1.1 average payload mass is on the bottom end of our payload mass range.

First Successful Ground Landing Date

```
%%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

Total Number of Successful and Failure Mission Outcomes

```
%%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

The month, landing outcome, booster version, payload mass (kg), and launch site of 2015 missions where stage 1 failed to land on a drone ship are returned by this query.
There were two instances like this.

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.

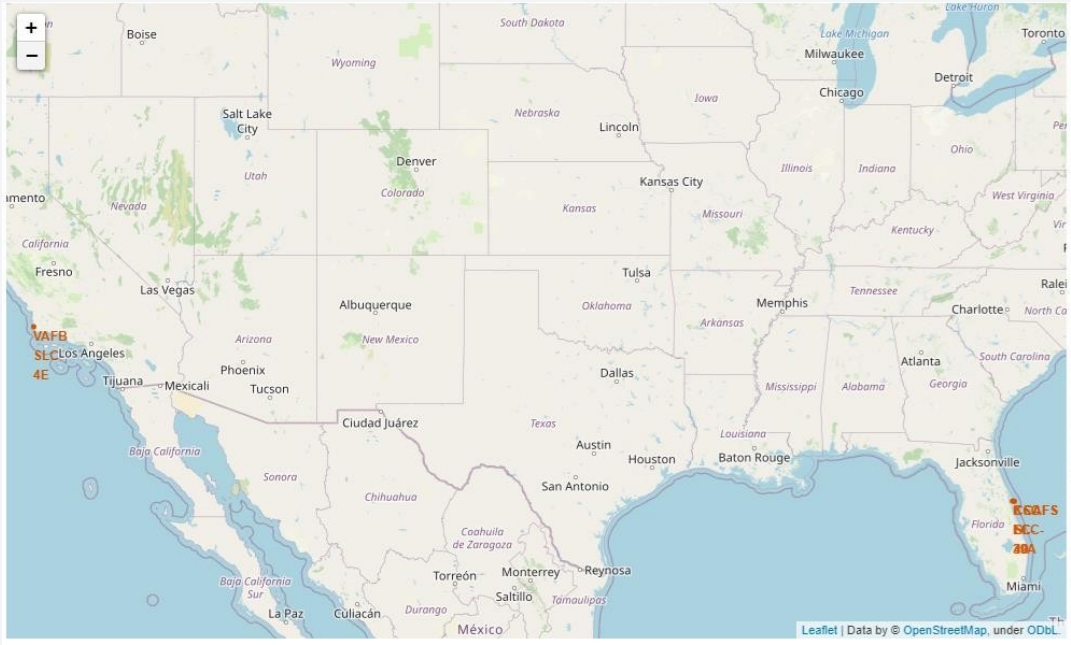
There are two types of successful landing outcomes: drone ship and ground pad landings.

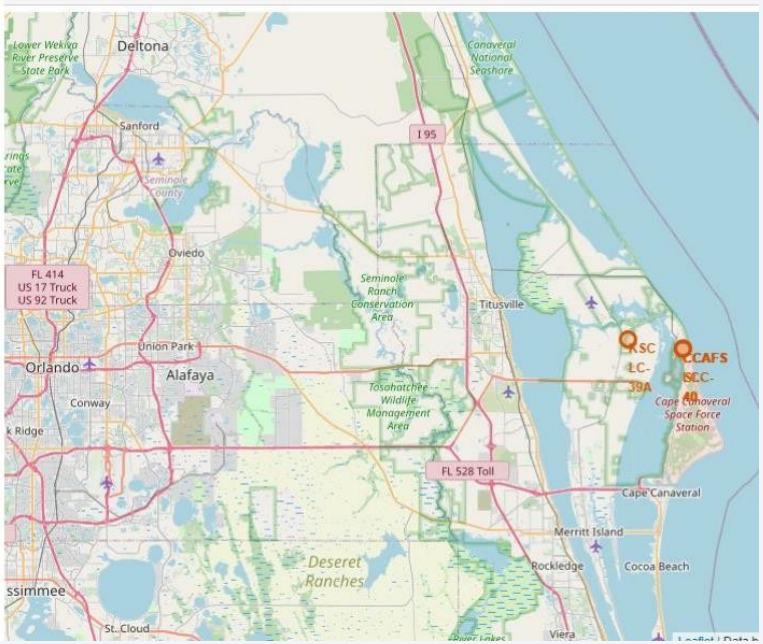
There were 8 successful landings in total during this time period

Section 4

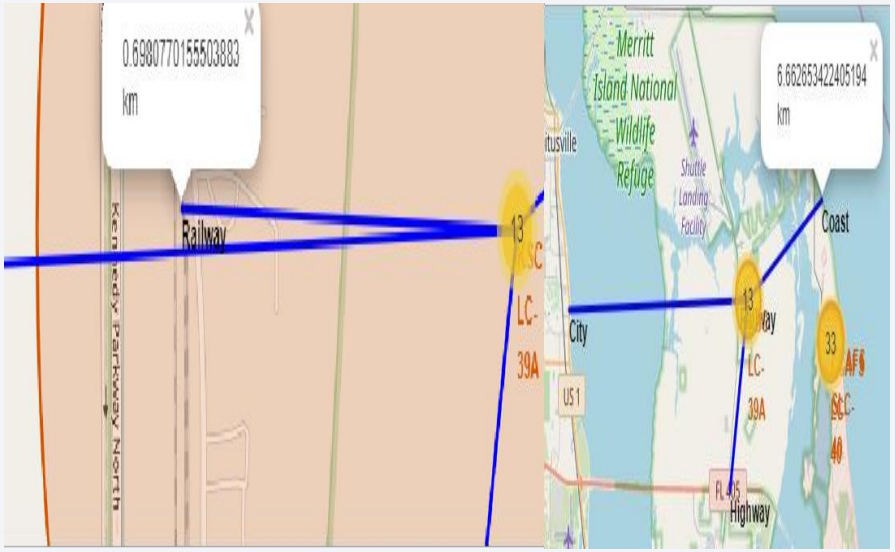
Launch Sites Proximities Analysis

Launch Site Location using Folium

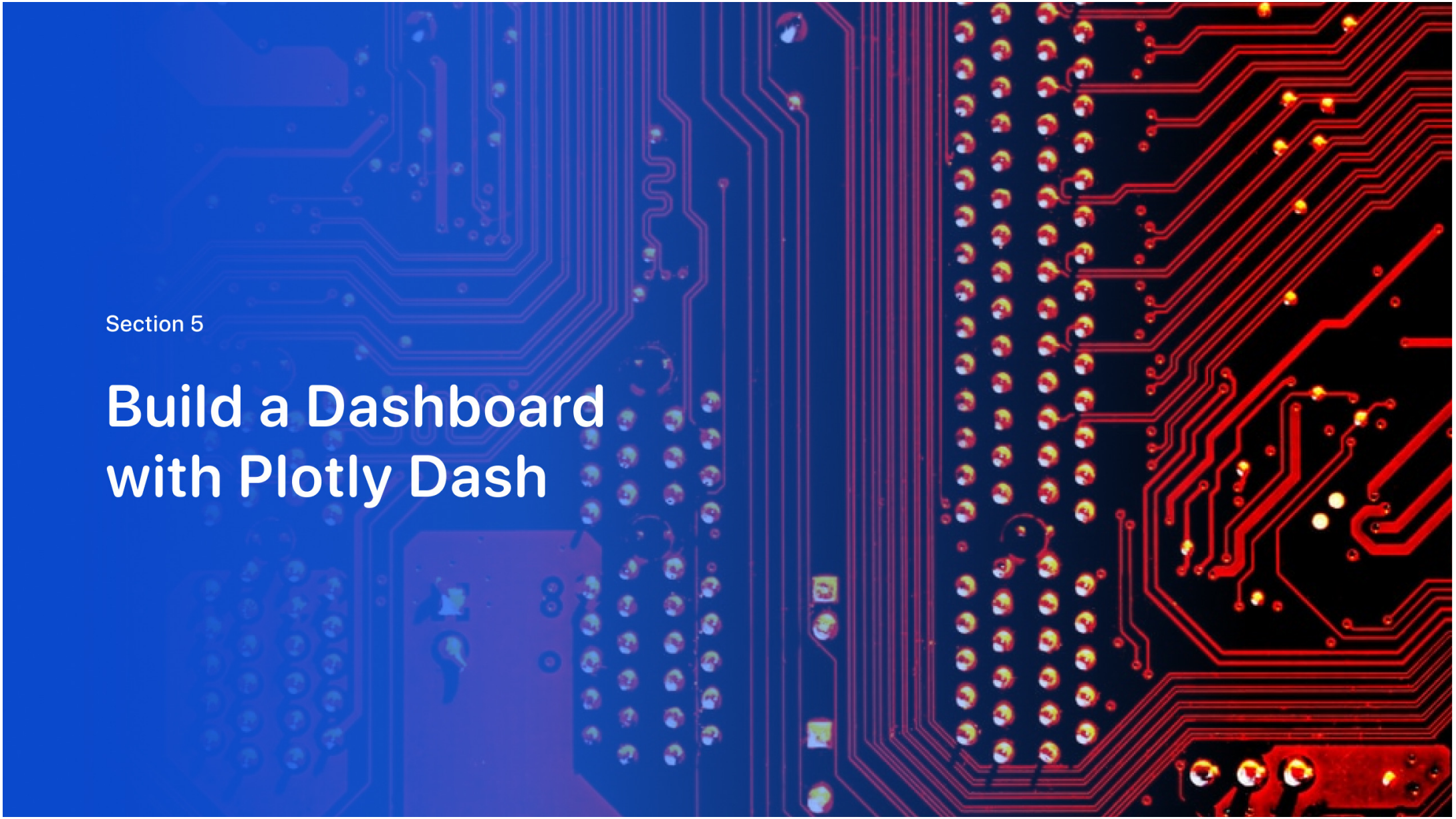




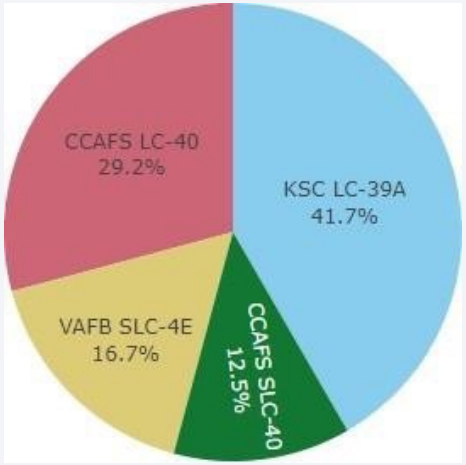
Launch sites adjacency



For the most part, launch sites are adjacent to trains and supply transports. Human and supply transportation are easily accessible from launch sites. Launch sites are also close to coasts and relatively far from towns, allowing launch failures to land in the sea rather than in densely populated areas.

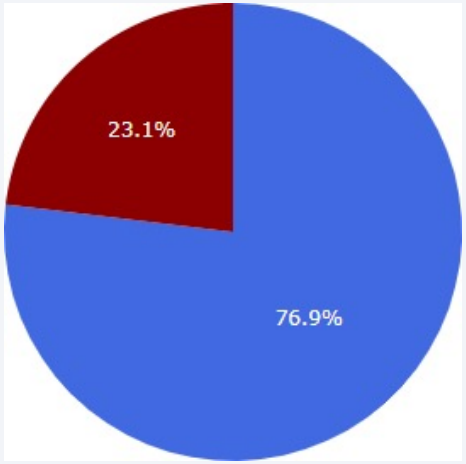

Section 5
Build a Dashboard with Plotly Dash

Launch Success Info.



- This is the percentage of landings that were successful across all launch sites. CCAFS LC-40 is the old name for CCAFS SLC-40,
- Hence CCAFS and KSC both have the same number of successful landings, although most of them happened before the name change.
- The number of successful landings at VAFB is the fewest. This could be attributed to a smaller sample size and increased launching difficulty on the west coast.

Most Successful Launch Site



KSC LC-39A Success Rate (blue=success)

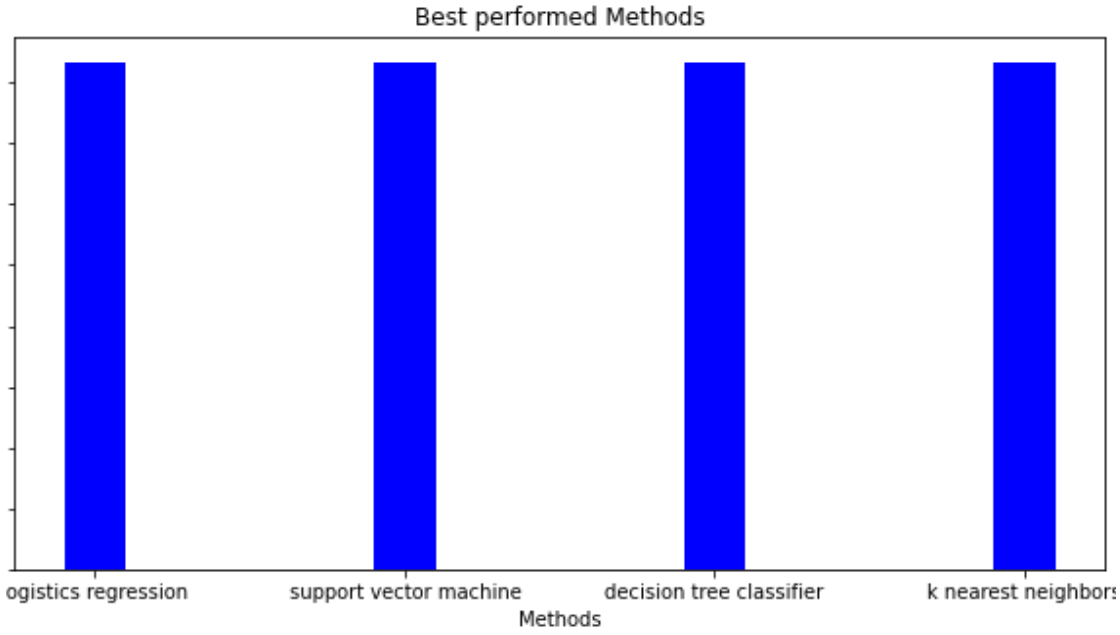
1
0

With ten successful landings and three failed landings, KSC LC-39A has the highest success rate.

Section 6

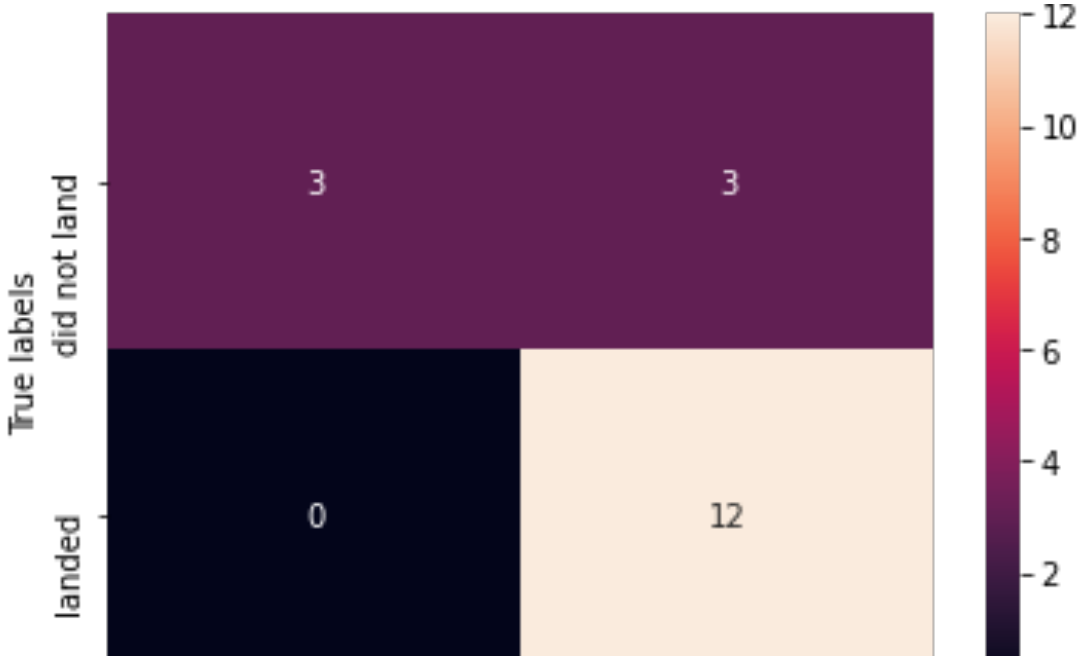
Predictive Analysis (Classification)

Classification Accuracy



- All the models have the same level of accuracy i.e., 83.33%.
- Hence, the data we have is not sufficient to quantify the best performing model.

Confusion Matrix



- Since all the models have the same accuracy levels.
- The confusion matrix for all of the models is same.
- When the true label was successful landing, the models predicted 12 successful landings.
- When the true label was failure landing, the models predicted three unsuccessful landings.
- When the true label was unsuccessful landings, the models projected three successful landings (false positives). Our models overestimate the likelihood of a successful landing.

Conclusions

- The primary data source for the analysis is Wikipedia web scraping and SpaceX API.
- Data labels created, and data was placed in a DB2 SQL database. For visualization, I created a dashboard. Developed a machine learning model that is 83% accurate.
- Models used for the predictive analysis are: KNN, SVM, Decision tree classifier, Logistic regression. All the models predicted the same level of accuracy as 83.33%. Hence no best performing model could be identified.
- Further, the data is inadequate to reach a conclusive best performing model and attain higher accuracy.

Appendix

- GitHub repository URL: https://github.com/Raghhavv/Data_science

- Course Instructor:
- Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Thank you!